



Apache Solr Reference Guide

For Solr 8.1

Written by the Apache Lucene/Solr Project

Published 2019-06-11

Table of Contents

Apache Solr Reference Guide	2
About This Guide	4
Hosts and Port Examples	5
Directory Paths	6
API Examples	7
Special Inline Notes	8
Getting Started	9
Solr Tutorial	10
A Quick Overview	38
Solr System Requirements	40
Installing Solr	43
Deployment and Operations	48
Solr Control Script Reference	49
Solr Configuration Files	69
Taking Solr to Production	71
Making and Restoring Backups	79
Running Solr on HDFS	84
SolrCloud on AWS EC2	89
Upgrading a Solr Cluster	97
Solr Upgrade Notes	101
Using the Solr Administration User Interface	125
Overview of the Solr Admin UI	126
Logging	129
Cloud Screens	130
Collections / Core Admin	134
Java Properties	136
Thread Dump	137
Suggestions Screen	139
Collection-Specific Tools	142
Core-Specific Tools	153
Documents, Fields, and Schema Design	158
Overview of Documents, Fields, and Schema Design	159
Solr Field Types	161
Defining Fields	184
Copying Fields	187
Dynamic Fields	189
Other Schema Elements	190

Schema API	192
Putting the Pieces Together	219
DocValues.....	221
Schemaless Mode.....	224
Understanding Analyzers, Tokenizers, and Filters.....	232
Using Analyzers, Tokenizers, and Filters	233
Analyzers	234
About Tokenizers	237
About Filters.....	238
Tokenizers	239
Filter Descriptions.....	249
CharFilterFactories	289
Language Analysis	293
Phonetic Matching	332
Running Your Analyzer.....	335
Indexing and Basic Data Operations.....	338
Indexing Using Client APIs	339
Introduction to Solr Indexing	340
Post Tool.....	342
Uploading Data with Index Handlers.....	346
Indexing Nested Child Documents	377
Uploading Data with Solr Cell using Apache Tika	381
Uploading Structured Data Store Data with the Data Import Handler	393
Updating Parts of Documents.....	420
Detecting Languages During Indexing	429
De-Duplication.....	433
Content Streams	435
Reindexing	437
Searching	442
Overview of Searching in Solr.....	444
Velocity Search UI.....	447
Relevance	448
Query Syntax and Parsing.....	450
JSON Request API	514
JSON Facet API.....	529
Faceting	556
Highlighting	574
Spell Checking	584
Query Re-Ranking.....	593
Transforming Result Documents	612

Searching Nested Child Documents	619
Suggester	624
MoreLikeThis	637
Pagination of Results	640
Collapse and Expand Results	647
Result Grouping	650
Result Clustering	656
Spatial Search	666
The Terms Component	678
The Term Vector Component	685
The Stats Component	690
The Query Elevation Component	695
The Tagger Handler	699
Response Writers	705
Near Real Time Searching	716
RealTime Get	719
Exporting Result Sets	723
Parallel SQL Interface	725
Analytics Component	761
Streaming Expressions	792
Stream Language Basics	793
Types of Streaming Expressions	795
Stream Source Reference	796
Stream Decorator Reference	810
Stream Evaluator Reference	841
Math Expressions	892
Graph Traversal	1028
SolrCloud	1041
Getting Started with SolrCloud	1042
How SolrCloud Works	1046
SolrCloud Resilience	1061
SolrCloud Configuration and Parameters	1066
Rule-based Replica Placement	1149
Cross Data Center Replication (CDCR)	1153
SolrCloud Autoscaling	1176
Colocating Collections	1227
Legacy Scaling and Distribution	1229
Introduction to Scaling and Distribution	1230
Distributed Search with Index Sharding	1231
Index Replication	1235

Combining Distribution and Replication.....	1245
Merging Indexes.....	1247
The Well-Configured Solr Instance.....	1248
Configuring solrconfig.xml	1249
Solr Cores and solr.xml	1293
Configuration APIs	1310
Implicit RequestHandlers	1342
Solr Plugins	1349
JVM Settings.....	1354
v2 API	1356
Monitoring Solr	1360
Metrics Reporting	1361
Metrics History	1377
MBean Request Handler	1390
Configuring Logging	1391
Using JMX with Solr.....	1395
Monitoring Solr with Prometheus and Grafana.....	1397
Performance Statistics Reference	1407
Securing Solr	1413
Authentication and Authorization Plugins	1414
Enabling SSL	1447
Audit Logging	1456
Client APIs	1460
Introduction to Client APIs	1461
Choosing an Output Format.....	1462
Client API Lineup.....	1463
Using JavaScript.....	1464
Using Python	1465
Using SolrJ	1466
Using Solr From Ruby.....	1472
Further Assistance	1474
Solr Glossary	1475
Solr Terms	1476
Errata	1481
Errata For This Documentation.....	1482
How to Contribute to Solr Documentation	1483

Licenses

Licensed to the Apache Software Foundation (ASF) under one or more contributor license agreements. See the NOTICE file distributed with this work for additional information regarding copyright ownership. The ASF licenses this file to you under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.apache.org/licenses/LICENSE-2.0>.

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

Apache and the Apache feather logo are trademarks of The Apache Software Foundation. Apache Lucene, Apache Solr and their respective logos are trademarks of the Apache Software Foundation. Please see the [Apache Trademark Policy](#) for more information.

Apache Solr Reference Guide

Welcome to Apache Solr™, the open source solution for search and analytics.

Solr is the fast open source search platform built on Apache Lucene™ that provides scalable indexing and search, as well as faceting, hit highlighting and advanced analysis/tokenization capabilities. Solr and Lucene are managed by the [Apache Software Foundation](#).

This Reference Guide is the official Solr documentation, written and published by Lucene/Solr committers.

The Guide includes the following sections:

Getting Started with Solr

The [Getting Started](#) section guides you through the installation and setup of Solr. A detailed tutorial for first-time users shows many of Solr's features.

[Using the Solr Administration User Interface](#): This section introduces the Web-based interface for administering Solr. From your browser you can view configuration files, submit queries, view logfile settings and Java environment settings, and monitor and control distributed configurations.

Deploying Solr

[Deployment and Operations](#): Once you have Solr configured, you want to deploy it to production and keep it up to date. This section includes information about how to take Solr to production, run it in HDFS or AWS, and information about upgrades and managing Solr from the command line.

[Monitoring Solr](#): Solr includes options for keeping an eye on the performance of your Solr cluster with the web-based administration console, through the command line interface, or using REST APIs.

Indexing Documents

[Indexing and Basic Data Operations](#): This section describes the indexing process and basic index operations, such as commit, optimize, and rollback.

[Documents, Fields, and Schema Design](#): This section describes how Solr organizes data in the index. It explains how a Solr schema defines the fields and field types which Solr uses to organize data within the document files it indexes.

[Understanding Analyzers, Tokenizers, and Filters](#): This section explains how Solr prepares text for indexing and searching. Analyzers parse text and produce a stream of tokens, lexical units used for indexing and searching. Tokenizers break field data down into tokens. Filters perform other transformational or selective work on token streams.

Searching Documents

Searching: This section presents an overview of the search process in Solr. It describes the main components used in searches, including request handlers, query parsers, and response writers. It lists the query parameters that can be passed to Solr, and it describes features such as boosting and faceting, which can be used to fine-tune search results.

Streaming Expressions: A stream processing language for Solr, with a suite of functions to perform many types of queries and parallel execution tasks.

Client APIs: This section tells you how to access Solr through various client APIs, including JavaScript, JSON, and Ruby.

Scaling Solr

SolrCloud: This section describes SolrCloud, which provides comprehensive distributed capabilities.

Legacy Scaling and Distribution: This section tells you how to grow a Solr distribution by dividing a large index into sections called shards, which are then distributed across multiple servers, or by replicating a single index across multiple services.

Advanced Configuration

Securing Solr: When planning how to secure Solr, you should consider which of the available features or approaches are right for you.

The Well-Configured Solr Instance: This section discusses performance tuning for Solr. It begins with an overview of the `solrconfig.xml` file, then tells you how to configure cores with `solr.xml`, how to configure the Lucene index writer, and more.

About This Guide

This guide describes all of the important features and functions of Apache Solr.

Solr is free to download from <http://lucene.apache.org/solr/>.

Designed to provide high-level documentation, this guide is intended to be more encyclopedic and less of a cookbook. It is structured to address a broad spectrum of needs, ranging from new developers getting started to well-experienced developers extending their application or troubleshooting. It will be of use at any point in the application life cycle, for whenever you need authoritative information about Solr.

The material as presented assumes that you are familiar with some basic search concepts and that you can read XML. It does not assume that you are a Java programmer, although knowledge of Java is helpful when working directly with Lucene or when developing custom extensions to a Lucene/Solr installation.

Hosts and Port Examples

The default port when running Solr is 8983. The samples, URLs and screenshots in this guide may show different ports, because the port number that Solr uses is configurable.

If you have not customized your installation of Solr, please make sure that you use port 8983 when following the examples, or configure your own installation to use the port numbers shown in the examples. For information about configuring port numbers, see the section [Monitoring Solr](#).

Similarly, URL examples use localhost throughout; if you are accessing Solr from a location remote to the server hosting Solr, replace localhost with the proper domain or IP where Solr is running.

For example, we might provide a sample query like:

```
http://localhost:8983/solr/gettingstarted/select?q=brown+cow
```

There are several items in this URL you might need to change locally. First, if your server is running at "www.example.com", you'll replace "localhost" with the proper domain. If you aren't using port 8983, you'll replace that also. Finally, you'll want to replace "gettingstarted" (the collection or core name) with the proper one in use in your implementation. The URL would then become:

```
http://www.example.com/solr/mycollection/select?q=brown+cow
```

Directory Paths

Path information is given relative to `solr.home`, which is the location under the main Solr installation where Solr's collections and their `conf` and `data` directories are stored.

In many cases, this is in the `server/solr` directory of your installation. However, there can be exceptions, particularly if your installation has customized this.

In several cases of this Guide, our examples are built from the "techproducts" example (i.e., you have started Solr with the command `bin/solr -e techproducts`). In this case, `solr.home` will be a sub-directory of the example/ directory created for you automatically.

See also the section [Solr Home](#) for further details on what is contained in this directory.

API Examples

Solr has two styles of APIs that currently co-exist. The first has grown somewhat organically as Solr has developed over time, but the second, referred to as the "V2 API", redesigns many of the original APIs with a modernized and self-documenting API interface.

In many cases, but not all, the parameters and outputs of API calls are the same between the two styles. In all cases the paths and endpoints used are different.

Throughout this Guide, we have added examples of both styles with sections labeled "V1 API" and "V2 API". As of the 7.2 version of this Guide, these examples are not yet complete - more coverage will be added as future versions of the Guide are released.

The section [V2 API](#) provides more information about how to work with the new API structure, including how to disable it if you choose to do so.

All APIs return a response header that includes the status of the request and the time to process it. Some APIs will also include the parameters used for the request. Many of the examples in this Guide omit this header information, which you can do locally by adding the parameter `omitHeader=true` to any request.

Special Inline Notes

Special notes are included throughout these pages. There are several types of notes:



Information blocks provide additional information that's useful for you to know.



Important blocks provide information that we want to make sure you are aware of.



Tip blocks provide helpful tips.



Caution blocks provide details on scenarios or configurations you should be careful with.



Warning blocks are used to warn you from a possibly dangerous change or action.

Getting Started

Solr makes it easy for programmers to develop sophisticated, high-performance search applications with advanced features.

This section introduces you to the basic Solr architecture and features to help you get up and running quickly. It covers the following topics:

[Solr Tutorial](#): This tutorial covers getting Solr up and running

[A Quick Overview](#): A high-level overview of how Solr works.

[Solr System Requirements](#): Solr System Requirement

[Installing Solr](#): A walkthrough of the Solr installation process.

Solr Tutorial

This tutorial covers getting Solr up and running, ingesting a variety of data sources into Solr collections, and getting a feel for the Solr administrative and search interfaces.

The tutorial is organized into three sections that each build on the one before it. The [first exercise](#) will ask you to start Solr, create a collection, index some basic documents, and then perform some searches.

The [second exercise](#) works with a different set of data, and explores requesting facets with the dataset.

The [third exercise](#) encourages you to begin to work with your own data and start a plan for your implementation.

Finally, we'll introduce [spatial search](#) and show you how to get your Solr instance back into a clean state.

Before You Begin

To follow along with this tutorial, you will need...

1. To meet the [system requirements](#)
2. An Apache Solr release [download](#). This tutorial is designed for Apache Solr 8.1.

For best results, please run the browser showing this tutorial and the Solr server on the same machine so tutorial links will correctly point to your Solr server.

Unpack Solr

Begin by unzipping the Solr release and changing your working directory to the subdirectory where Solr was installed. For example, with a shell in UNIX, Cygwin, or MacOS:

```
~$ ls solr*
solr-8.1.0.zip

~$ unzip -q solr-8.1.0.zip

~$ cd solr-8.1.0/
```

If you'd like to know more about Solr's directory layout before moving to the first exercise, see the section [Directory Layout](#) for details.

Exercise 1: Index Techproducts Example Data

This exercise will walk you through how to start Solr as a two-node cluster (both nodes on the same machine) and create a collection during startup. Then you will index some sample data that ships with Solr and do some basic searches.

Launch Solr in SolrCloud Mode

To launch Solr, run: `bin/solr start -e cloud` on Unix or MacOS; `bin\solr.cmd start -e cloud` on Windows.

This will start an interactive session that will start two Solr "servers" on your machine. This command has an option to run without prompting you for input (`-noprompt`), but we want to modify two of the defaults so we won't use that option now.

```
solr-8.1.0:$ ./bin/solr start -e cloud
```

```
Welcome to the SolrCloud example!
```

```
This interactive session will help you launch a SolrCloud cluster on your local workstation.  
To begin, how many Solr nodes would you like to run in your local cluster? (specify 1-4 nodes)  
[2]:
```

The first prompt asks how many nodes we want to run. Note the [2] at the end of the last line; that is the default number of nodes. Two is what we want for this example, so you can simply press enter.

```
Ok, let's start up 2 Solr nodes for your example SolrCloud cluster.  
Please enter the port for node1 [8983]:
```

This will be the port that the first node runs on. Unless you know you have something else running on port 8983 on your machine, accept this default option also by pressing enter. If something is already using that port, you will be asked to choose another port.

```
Please enter the port for node2 [7574]:
```

This is the port the second node will run on. Again, unless you know you have something else running on port 8983 on your machine, accept this default option also by pressing enter. If something is already using that port, you will be asked to choose another port.

Solr will now initialize itself and start running on those two nodes. The script will print the commands it uses for your reference.

```
Starting up 2 Solr nodes for your example SolrCloud cluster.
```

```
Creating Solr home directory /solr-8.1.0/example/cloud/node1/solr
Cloning /solr-8.1.0/example/cloud/node1 into
/solr-8.1.0/example/cloud/node2
```

```
Starting up Solr on port 8983 using command:
"bin/solr" start -cloud -p 8983 -s "example/cloud/node1/solr"
```

```
Waiting up to 180 seconds to see Solr running on port 8983 []
Started Solr server on port 8983 (pid=34942). Happy searching!
```

```
Starting up Solr on port 7574 using command:
"bin/solr" start -cloud -p 7574 -s "example/cloud/node2/solr" -z localhost:9983
```

```
Waiting up to 180 seconds to see Solr running on port 7574 []
Started Solr server on port 7574 (pid=35036). Happy searching!
```

```
INFO - 2017-07-27 12:28:02.835; org.apache.solr.client.solrj.impl.ZkClientClusterStateProvider;
Cluster at localhost:9983 ready
```

Notice that two instances of Solr have started on two nodes. Because we are starting in SolrCloud mode, and did not define any details about an external ZooKeeper cluster, Solr launches its own ZooKeeper and connects both nodes to it.

After startup is complete, you'll be prompted to create a collection to use for indexing data.

```
Now let's create a new collection for indexing documents in your 2-node cluster.
Please provide a name for your new collection: [gettingstarted]
```

Here's the first place where we'll deviate from the default options. This tutorial will ask you to index some sample data included with Solr, called the "techproducts" data. Let's name our collection "techproducts" so it's easy to differentiate from other collections we'll create later. Enter techproducts at the prompt and hit enter.

```
How many shards would you like to split techproducts into? [2]
```

This is asking how many **shards** you want to split your index into across the two nodes. Choosing "2" (the default) means we will split the index relatively evenly across both nodes, which is a good way to start. Accept the default by hitting enter.

```
How many replicas per shard would you like to create? [2]
```

A replica is a copy of the index that's used for failover (see also the [Solr Glossary definition](#)). Again, the default of "2" is fine to start with here also, so accept the default by hitting enter.

Please choose a configuration for the techproducts collection, available options are:
_default or sample_techproducts_configs [_default]

We've reached another point where we will deviate from the default option. Solr has two sample sets of configuration files (called a configset) available out-of-the-box.

A collection must have a configset, which at a minimum includes the two main configuration files for Solr: the schema file (named either managed-schema or schema.xml), and solrconfig.xml. The question here is which configset you would like to start with. The _default is a bare-bones option, but note there's one whose name includes "techproducts", the same as we named our collection. This configset is specifically designed to support the sample data we want to use, so enter sample_techproducts_configs at the prompt and hit enter.

At this point, Solr will create the collection and again output to the screen the commands it issues.

```
Uploading /solr-8.1.0/server/solr/configsets/_default/conf for config techproducts to ZooKeeper  
at localhost:9983
```

```
Connecting to ZooKeeper at localhost:9983 ...
```

```
INFO - 2017-07-27 12:48:59.289; org.apache.solr.client.solrj.impl.ZkClientClusterStateProvider;  
Cluster at localhost:9983 ready  
Uploading /solr-8.1.0/server/solr/configsets/sample_techproducts_configs/conf for config  
techproducts to ZooKeeper at localhost:9983
```

```
Creating new collection 'techproducts' using command:
```

```
http://localhost:8983/solr/admin/collections?action=CREATE&name=techproducts&numShards=2&replicat  
ionFactor=2&maxShardsPerNode=2&collection.configName=techproducts
```

```
{  
  "responseHeader":{  
    "status":0,  
    "QTime":5460},  
  "success":{  
    "192.168.0.110:7574_solr":{  
      "responseHeader":{  
        "status":0,  
        "QTime":4056},  
      "core": "techproducts_shard1_replica_n1"},  
    "192.168.0.110:8983_solr":{  
      "responseHeader":{  
        "status":0,  
        "QTime":4056},  
      "core": "techproducts_shard2_replica_n2"}}}
```

```
Enabling auto soft-commits with maxTime 3 secs using the Config API
```

```
POSTing request to Config API: http://localhost:8983/solr/techproducts/config  
{"set-property": {"updateHandler.autoSoftCommit.maxTime": "3000"} }  
Successfully set-property updateHandler.autoSoftCommit.maxTime to 3000
```

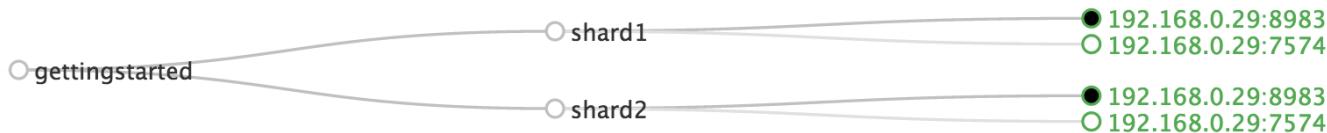
```
SolrCloud example running, please visit: http://localhost:8983/solr
```

Congratulations! Solr is ready for data!

You can see that Solr is running by launching the Solr Admin UI in your web browser: <http://localhost:8983/solr/>. This is the main starting point for administering Solr.

Solr will now be running two "nodes", one on port 7574 and one on port 8983. There is one collection created automatically, techproducts, a two shard collection, each with two replicas.

The [Cloud tab](#) in the Admin UI diagrams the collection nicely:



SolrCloud Diagram

Index the Techproducts Data

Your Solr server is up and running, but it doesn't contain any data yet, so we can't do any queries.

Solr includes the bin/post tool in order to facilitate indexing various types of documents easily. We'll use this tool for the indexing examples below.

You'll need a command shell to run some of the following examples, rooted in the Solr install directory; the shell from where you launched Solr works just fine.



Currently the bin/post tool does not have a comparable Windows script, but the underlying Java program invoked is available. We'll show examples below for Windows, but you can also see the [Windows section](#) of the Post Tool documentation for more details.

The data we will index is in the example/exampledocs directory. The documents are in a mix of document formats (JSON, CSV, etc.), and fortunately we can index them all at once:

Linux/Mac

```
solr-8.1.0:$ bin/post -c techproducts example/exampledocs/*
```

Windows

```
C:\solr-8.1.0> java -jar -Dc=techproducts -Dauto example\exampledocs\post.jar  
example\exampledocs\*
```

You should see output similar to the following:

```
SimplePostTool version 5.0.0
Posting files to [base] url http://localhost:8983/solr/techproducts/update...
Entering auto mode. File endings considered are
xml,json,jsonl,csv,pdf,doc,docx,ppt,pptx,xls,xlsx,odt,odp,ods,ott,otp,ots,rtf,htm,html,txt,log
POSTing file books.csv (text/csv) to [base]
POSTing file books.json (application/json) to [base]/json/docs
POSTing file gb18030-example.xml (application/xml) to [base]
POSTing file hd.xml (application/xml) to [base]
POSTing file ipod_other.xml (application/xml) to [base]
POSTing file ipod_video.xml (application/xml) to [base]
POSTing file manufacturers.xml (application/xml) to [base]
POSTing file mem.xml (application/xml) to [base]
POSTing file money.xml (application/xml) to [base]
POSTing file monitor.xml (application/xml) to [base]
POSTing file monitor2.xml (application/xml) to [base]
POSTing file more_books.jsonl (application/json) to [base]/json/docs
POSTing file mp500.xml (application/xml) to [base]
POSTing file post.jar (application/octet-stream) to [base]/extract
POSTing file sample.html (text/html) to [base]/extract
POSTing file sd500.xml (application/xml) to [base]
POSTing file solr-word.pdf (application/pdf) to [base]/extract
POSTing file solr.xml (application/xml) to [base]
POSTing file test_utf8.sh (application/octet-stream) to [base]/extract
POSTing file utf8-example.xml (application/xml) to [base]
POSTing file vidcard.xml (application/xml) to [base]
21 files indexed.
COMMITting Solr index changes to http://localhost:8983/solr/techproducts/update...
Time spent: 0:00:00.822
```

Congratulations again! You have data in your Solr!

Now we're ready to start searching.

Basic Searching

Solr can be queried via REST clients, curl, wget, Chrome POSTMAN, etc., as well as via native clients available for many programming languages.

The Solr Admin UI includes a query builder interface via the Query tab for the techproducts collection (at <http://localhost:8983/solr/#/techproducts/query>). If you click the **[Execute Query]** button without changing anything in the form, you'll get 10 documents in JSON format:

The URL sent by the Admin UI to Solr is shown in light grey near the top right of the above screenshot. If you click on it, your browser will show you the raw response.

```
http://localhost:8983/solr/techproducts/select?q=*&*:&*
```

Query Screen

The URL sent by the Admin UI to Solr is shown in light grey near the top right of the above screenshot. If you click on it, your browser will show you the raw response.

To use curl, give the same URL shown in your browser in quotes on the command line:

```
curl "http://localhost:8983/solr/techproducts/select?indent=on&q=*&*:"
```

What's happening here is that we are using Solr's query parameter (q) with a special syntax that requests all documents in the index (*: *). All of the documents are not returned to us, however, because of the default for a parameter called rows, which you can see in the form is 10. You can change the parameter in the UI or in the defaults if you wish.

Solr has very powerful search options, and this tutorial won't be able to cover all of them. But we can cover some of the most common types of queries.

Search for a Single Term

To search for a term, enter it as the q parameter value in the Solr Admin UI Query screen, replacing *: * with the term you want to find.

Enter "foundation" and hit [Execute Query] again.

If you prefer curl, enter something like this:

```
curl "http://localhost:8983/solr/techproducts/select?q=foundation"
```

You'll see something like this:

```
{  
  "responseHeader":{  
    "zkConnected":true,  
    "status":0,  
    "QTime":8,  
    "params":{  
      "q":"foundation"}},  
  "response":{ "numFound":4, "start":0, "maxScore":2.7879646, "docs": [  
    {  
      "id":"0553293354",  
      "cat":["book"],  
      "name":"Foundation",  
      "price":7.99,  
      "price_c":7.99,USD,  
      "inStock":true,  
      "author": "Isaac Asimov",  
      "author_s": "Isaac Asimov",  
      "series_t": "Foundation Novels",  
      "sequence_i":1,  
      "genre_s": "scifi",  
      "_version_":1574100232473411586,  
      "price_c_____l_ns":799} ]  
  } }
```

The response indicates that there are 4 hits ("numFound": 4). We've only included one document in the above sample output, but since 4 hits is lower than the rows parameter default of 10 to be returned, you should see all 4 of them.

Note the responseHeader before the documents. This header will include the parameters you have set for the search. By default it shows only the parameters you have set for this query, which in this case is only your query term.

The documents we got back include all the fields for each document that were indexed. This is, again, default behavior. If you want to restrict the fields in the response, you can use the fl parameter, which takes a comma-separated list of field names. This is one of the available fields on the query form in the Admin UI.

Put "id" (without quotes) in the "fl" box and hit [**Execute Query**] again. Or, to specify it with curl:

```
curl "http://localhost:8983/solr/techproducts/select?q=foundation&fl=id"
```

You should only see the IDs of the matching records returned.

Field Searches

All Solr queries look for documents using some field. Often you want to query across multiple fields at the same time, and this is what we've done so far with the "foundation" query. This is possible with the use of

copy fields, which are set up already with this set of configurations. We'll cover copy fields a little bit more in Exercise 2.

Sometimes, though, you want to limit your query to a single field. This can make your queries more efficient and the results more relevant for users.

Much of the data in our small sample data set is related to products. Let's say we want to find all the "electronics" products in the index. In the Query screen, enter "electronics" (without quotes) in the q box and hit [**Execute Query**]. You should get 14 results, such as:

```
{  
  "responseHeader":{  
    "zkConnected":true,  
    "status":0,  
    "QTime":6,  
    "params":{  
      "q":"electronics"}},  
  "response":{ "numFound":14, "start":0, "maxScore":1.5579545, "docs": [  
    {  
      "id":"IW-02",  
      "name":"iPod & iPod Mini USB 2.0 Cable",  
      "manu":"Belkin",  
      "manu_id_s":"belkin",  
      "cat":["electronics",  
        "connector"],  
      "features":["car power adapter for iPod, white"],  
      "weight":2.0,  
      "price":11.5,  
      "price_c":"11.50,USD",  
      "popularity":1,  
      "inStock":false,  
      "store":"37.7752,-122.4232",  
      "manufacturedate_dt":"2006-02-14T23:55:59Z",  
      "_version_":1574100232554151936,  
      "price_c___l_ns":1150} ]  
  } }
```

This search finds all documents that contain the term "electronics" anywhere in the indexed fields. However, we can see from the above there is a cat field (for "category"). If we limit our search for only documents with the category "electronics", the results will be more precise for our users.

Update your query in the q field of the Admin UI so it's cat:electronics. Now you get 12 results:

```
{  
  "responseHeader":{  
    "zkConnected":true,  
    "status":0,  
    "QTime":6,  
    "params":{  
      "q":"cat:electronics"}},  
  "response":{ "numFound":12, "start":0, "maxScore":0.9614112, "docs": [  
    {  
      "id": "SP2514N",  
      "name": "Samsung SpinPoint P120 SP2514N - hard drive - 250 GB - ATA-133",  
      "manu": "Samsung Electronics Co. Ltd.",  
      "manu_id_s": "samsung",  
      "cat": ["electronics",  
        "hard drive"],  
      "features": ["7200RPM, 8MB cache, IDE Ultra ATA-133",  
        "NoiseGuard, SilentSeek technology, Fluid Dynamic Bearing (FDB) motor"],  
      "price": 92.0,  
      "price_c": "92.0,USD",  
      "popularity": 6,  
      "inStock": true,  
      "manufacturedate_dt": "2006-02-13T15:26:37Z",  
      "store": "35.0752,-97.032",  
      "_version_": 1574100232511160320,  
      "price_c___l_ns": 9200} ]  
  ]}  
}
```

Using curl, this query would look like this:

```
curl "http://localhost:8983/solr/techproducts/select?q=cat:electronics"
```

Phrase Search

To search for a multi-term phrase, enclose it in double quotes: q="multiple terms here". For example, search for "CAS latency" by entering that phrase in quotes to the q box in the Admin UI.

If you're following along with curl, note that the space between terms must be converted to "+" in a URL, as so:

```
curl "http://localhost:8983/solr/techproducts/select?q=\"CAS+latency\""
```

We get 2 results:

```
{  
  "responseHeader":{  
    "zkConnected":true,  
    "status":0,  
    "QTime":7,  
    "params":{  
      "q":"\"CAS latency\""}},  
  "response":{ "numFound":2, "start":0, "maxScore":5.937691, "docs": [  
    {  
      "id": "VDBDB1A16",  
      "name": "A-DATA V-Series 1GB 184-Pin DDR SDRAM Unbuffered DDR 400 (PC 3200) System Memory - OEM",  
      "manu": "A-DATA Technology Inc.",  
      "manu_id_s": "corsair",  
      "cat": ["electronics",  
              "memory"],  
      "features": ["CAS latency 3, 2.7v"],  
      "popularity": 0,  
      "inStock": true,  
      "store": "45.18414,-93.88141",  
      "manufacturedate_dt": "2006-02-13T15:26:37Z",  
      "payloads": "electronics|0.9 memory|0.1",  
      "_version_": 1574100232590852096},  
    {  
      "id": "TWINX2048-3200PRO",  
      "name": "CORSAIR XMS 2GB (2 x 1GB) 184-Pin DDR SDRAM Unbuffered DDR 400 (PC 3200) Dual Channel Kit System Memory - Retail",  
      "manu": "Corsair Microsystems Inc.",  
      "manu_id_s": "corsair",  
      "cat": ["electronics",  
              "memory"],  
      "features": ["CAS latency 2, 2-3-3-6 timing, 2.75v, unbuffered, heat-spreader"],  
      "price": 185.0,  
      "price_c": "185.00,USD",  
      "popularity": 5,  
      "inStock": true,  
      "store": "37.7752,-122.4232",  
      "manufacturedate_dt": "2006-02-13T15:26:37Z",  
      "payloads": "electronics|6.0 memory|3.0",  
      "_version_": 1574100232584560640,  
      "price_c_l_ns": 18500]  
  ]}  
}
```

Combining Searches

By default, when you search for multiple terms and/or phrases in a single query, Solr will only require that one of them is present in order for a document to match. Documents containing more terms will be sorted higher in the results list.

You can require that a term or phrase is present by prefixing it with a +; conversely, to disallow the presence of a term or phrase, prefix it with a -.

To find documents that contain both terms "electronics" and "music", enter +electronics +music in the q box in the Admin UI Query tab.

If you're using curl, you must encode the + character because it has a reserved purpose in URLs (encoding the space character). The encoding for + is %2B as in:

```
curl "http://localhost:8983/solr/techproducts/select?q=%2Belectronics%20%2Bmusic"
```

You should only get a single result.

To search for documents that contain the term "electronics" but **don't** contain the term "music", enter +electronics -music in the q box in the Admin UI. For curl, again, URL encode + as %2B as in:

```
curl "http://localhost:8983/solr/techproducts/select?q=%2Belectronics+-music"
```

This time you get 13 results.

More Information on Searching

We have only scratched the surface of the search options available in Solr. For more Solr search options, see the section on [Searching](#).

Exercise 1 Wrap Up

At this point, you've seen how Solr can index data and have done some basic queries. You can choose now to continue to the next example which will introduce more Solr concepts, such as facetting results and managing your schema, or you can strike out on your own.

If you decide not to continue with this tutorial, the data we've indexed so far is likely of little value to you. You can delete your installation and start over, or you can use the bin/solr script we started out with to delete this collection:

```
bin/solr delete -c techproducts
```

And then create a new collection:

```
bin/solr create -c <yourCollection> -s 2 -rf 2
```

To stop both of the Solr nodes we started, issue the command:

```
bin/solr stop -all
```

For more information on start/stop and collection options with bin/solr, see [Solr Control Script Reference](#).

Exercise 2: Modify the Schema and Index Films Data

This exercise will build on the last one and introduce you to the index schema and Solr's powerful faceting features.

Restart Solr

Did you stop Solr after the last exercise? No? Then go ahead to the next section.

If you did, though, and need to restart Solr, issue these commands:

```
./bin/solr start -c -p 8983 -s example/cloud/node1/solr
```

This starts the first node. When it's done start the second node, and tell it how to connect to ZooKeeper:

```
./bin/solr start -c -p 7574 -s example/cloud/node2/solr -z localhost:9983
```



If you have defined ZK_HOST in solr.in.sh/solr.in.cmd (see [instructions](#)) you can omit -z <zookeeper host string> from the above command.

Create a New Collection

We're going to use a whole new data set in this exercise, so it would be better to have a new collection instead of trying to reuse the one we had before.

One reason for this is we're going to use a feature in Solr called "field guessing", where Solr attempts to guess what type of data is in a field while it's indexing it. It also automatically creates new fields in the schema for new fields that appear in incoming documents. This mode is called "Schemaless". We'll see the benefits and limitations of this approach to help you decide how and where to use it in your real application.

What is a "schema" and why do I need one?

Solr's schema is a single file (in XML) that stores the details about the fields and field types Solr is expected to understand. The schema defines not only the field or field type names, but also any modifications that should happen to a field before it is indexed. For example, if you want to ensure that a user who enters "abc" and a user who enters "ABC" can both find a document containing the term "ABC", you will want to normalize (lower-case it, in this case) "ABC" when it is indexed, and normalize the user query to be sure of a match. These rules are defined in your schema.

Earlier in the tutorial we mentioned copy fields, which are fields made up of data that originated from other fields. You can also define dynamic fields, which use wildcards (such as *_t or *_s) to dynamically create fields of a specific field type. These types of rules are also defined in the schema.

When you initially started Solr in the first exercise, we had a choice of a configset to use. The one we chose had a schema that was pre-defined for the data we later indexed. This time, we're going to use a configset that has a very minimal schema and let Solr figure out from the data what fields to add.

The data you're going to index is related to movies, so start by creating a collection named "films" that uses the _default configset:

```
bin/solr create -c films -s 2 -rf 2
```

Whoa, wait. We didn't specify a configset! That's fine, the _default is appropriately named, since it's the default and is used if you don't specify one at all.

We did, however, set two parameters -s and -rf. Those are the number of shards to split the collection across (2) and how many replicas to create (2). This is equivalent to the options we had during the interactive example from the first exercise.

You should see output like:

```
WARNING: Using _default configset. Data driven schema functionality is enabled by default, which  
is
```

```
NOT RECOMMENDED for production use.
```

```
To turn it off:
```

```
bin/solr config -c films -p 7574 -action set-user-property -property  
update.autoCreateFields -value false
```

```
Connecting to ZooKeeper at localhost:9983 ...
```

```
INFO - 2017-07-27 15:07:46.191; org.apache.solr.client.solrj.impl.ZkClientClusterStateProvider;
```

```
Cluster at localhost:9983 ready
```

```
Uploading /8.1.0/server/solr/configsets/_default/conf for config films to ZooKeeper at  
localhost:9983
```

```
Creating new collection 'films' using command:
```

```
http://localhost:7574/solr/admin/collections?action=CREATE&name=films&numShards=2&replicationFact  
or=2&maxShardsPerNode=2&collection.configName=films
```

```
{  
  "responseHeader":{  
    "status":0,  
    "QTime":3830},  
  "success":{  
    "192.168.0.110:8983_solr":{  
      "responseHeader":{  
        "status":0,  
        "QTime":2076},  
      "core":"films_shard2_replica_n1"},  
    "192.168.0.110:7574_solr":{  
      "responseHeader":{  
        "status":0,  
        "QTime":2494},  
      "core":"films_shard1_replica_n2"}}}
```

The first thing the command printed was a warning about not using this configset in production. That's due to some of the limitations we'll cover shortly.

Otherwise, though, the collection should be created. If we go to the Admin UI at <http://localhost:8983/solr/#/films/collection-overview> we should see the overview screen.

Preparing Schemaless for the Films Data

There are two parallel things happening with the schema that comes with the _default configset.

First, we are using a "managed schema", which is configured to only be modified by Solr's Schema API. That means we should not hand-edit it so there isn't confusion about which edits come from which source. Solr's Schema API allows us to make changes to fields, field types, and other types of schema rules.

Second, we are using "field guessing", which is configured in the solrconfig.xml file (and includes most of Solr's various configuration settings). Field guessing is designed to allow us to start using Solr without having to define all the fields we think will be in our documents before trying to index them. This is why we

call it "schemaless", because you can start quickly and let Solr create fields for you as it encounters them in documents.

Sounds great! Well, not really, there are limitations. It's a bit brute force, and if it guesses wrong, you can't change much about a field after data has been indexed without having to reindex. If we only have a few thousand documents that might not be bad, but if you have millions and millions of documents, or, worse, don't have access to the original data anymore, this can be a real problem.

For these reasons, the Solr community does not recommend going to production without a schema that you have defined yourself. By this we mean that the schemaless features are fine to start with, but you should still always make sure your schema matches your expectations for how you want your data indexed and how users are going to query it.

It is possible to mix schemaless features with a defined schema. Using the Schema API, you can define a few fields that you know you want to control, and let Solr guess others that are less important or which you are confident (through testing) will be guessed to your satisfaction. That's what we're going to do here.

Create the "names" Field

The films data we are going to index has a small number of fields for each movie: an ID, director name(s), film name, release date, and genre(s).

If you look at one of the files in example/films, you'll see the first film is named *.45*, released in 2006. As the first document in the dataset, Solr is going to guess the field type based on the data in the record. If we go ahead and index this data, that first film name is going to indicate to Solr that that field type is a "float" numeric field, and will create a "name" field with a type FloatPointField. All data after this record will be expected to be a float.

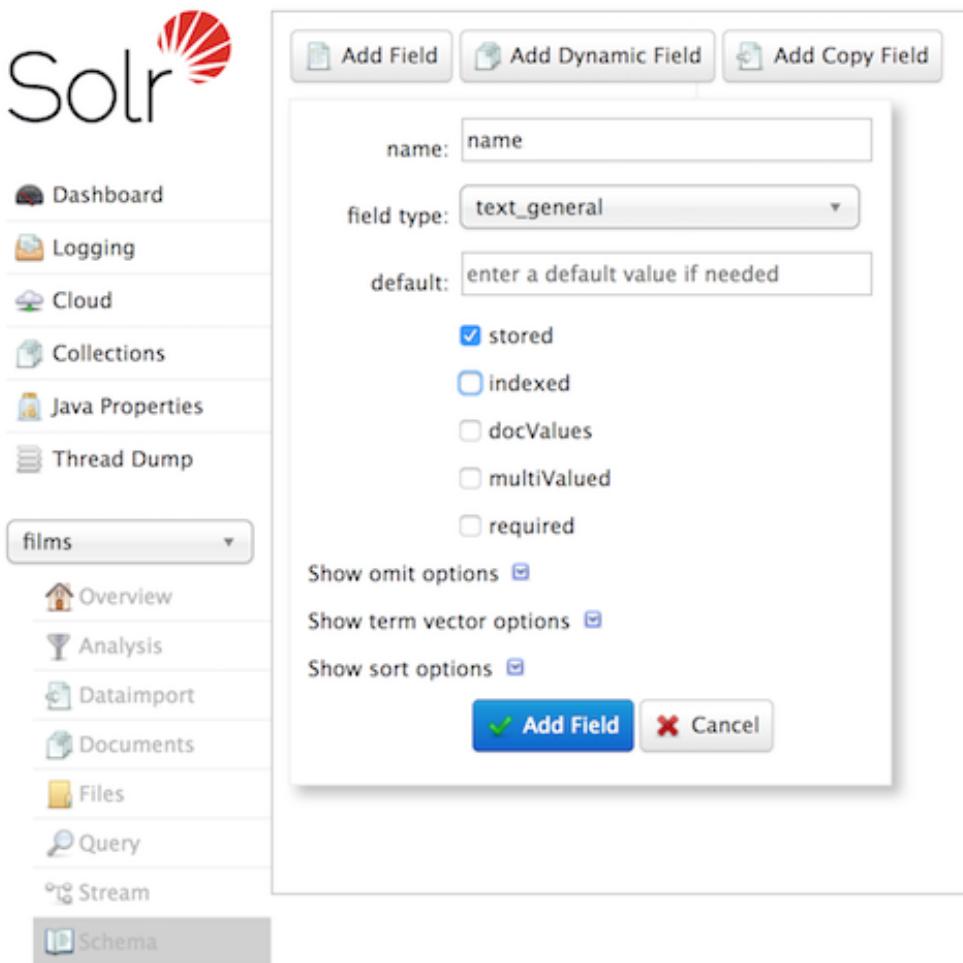
Well, that's not going to work. We have titles like *A Mighty Wind* and *Chicken Run*, which are strings - decidedly not numeric and not floats. If we let Solr guess the "name" field is a float, what will happen is later titles will cause an error and indexing will fail. That's not going to get us very far.

What we can do is set up the "name" field in Solr before we index the data to be sure Solr always interprets it as a string. At the command line, enter this curl command:

```
curl -X POST -H 'Content-type:application/json' --data-binary '{"add-field": {"name": "name", "type": "text_general", "multiValued":false, "stored":true}}'  
http://localhost:8983/solr/films/schema
```

This command uses the Schema API to explicitly define a field named "name" that has the field type "text_general" (a text field). It will not be permitted to have multiple values, but it will be stored (meaning it can be retrieved by queries).

You can also use the Admin UI to create fields, but it offers a bit less control over the properties of your field. It will work for our case, though:



Creating a field

Create a "catchall" Copy Field

There's one more change to make before we start indexing.

In the first exercise when we queried the documents we had indexed, we didn't have to specify a field to search because the configuration we used was set up to copy fields into a text field, and that field was the default when no other field was defined in the query.

The configuration we're using now doesn't have that rule. We would need to define a field to search for every query. We can, however, set up a "catchall field" by defining a copy field that will take all data from all fields and index it into a field named `_text_`. Let's do that now.

You can use either the Admin UI or the Schema API for this.

At the command line, use the Schema API again to define a copy field:

```
curl -X POST -H 'Content-type:application/json' --data-binary '{"add-copy-field" : {"source":"*","dest":"_text_"} }' http://localhost:8983/solr/films/schema
```

In the Admin UI, choose [**Add Copy Field**], then fill out the source and destination for your field, as in this screenshot.

source: *

destination: _text_

SchemaSimilarity. Default:
BM25(k1=1.2,b=0.75)

Add CopyField Cancel

Creating a copy field

What this does is make a copy of all fields and put the data into the "_text_" field.



It can be very expensive to do this with your production data because it tells Solr to effectively index everything twice. It will make indexing slower, and make your index larger. With your production data, you will want to be sure you only copy fields that really warrant it for your application.

OK, now we're ready to index the data and start playing around with it.

Index Sample Film Data

The films data we will index is located in the example/films directory of your installation. It comes in three formats: JSON, XML and CSV. Pick one of the formats and index it into the "films" collection (in each example, one command is for Unix/MacOS and the other is for Windows):

To Index JSON Format

```
bin/post -c films example/films/films.json  
C:\solr-8.1.0> java -jar -Dc=films -Dauto example\exampledocs\post.jar example\films\*.json
```

To Index XML Format

```
bin/post -c films example/films/films.xml  
C:\solr-8.1.0> java -jar -Dc=films -Dauto example\exampledocs\post.jar example\films\*.xml
```

To Index CSV Format

```
bin/post -c films example/films/films.csv -params  
"f.genre.split=true&f.directed_by.split=true&f.genre.separator=|&f.directed_by.separator=|"  
  
C:\solr-8.1.0> java -jar -Dc=films  
-Dparams=f.genre.split=true&f.directed_by.split=true&f.genre.separator=|&f.directed_by.separator=  
| -Dauto example\exampledocs\post.jar example\films\*.csv
```

Each command includes these main parameters:

- `-c films`: this is the Solr collection to index data to.
- `example/films/films.json` (or `films.xml` or `films.csv`): this is the path to the data file to index. You could simply supply the directory where this file resides, but since you know the format you want to index, specifying the exact file for that format is more efficient.

Note the CSV command includes extra parameters. This is to ensure multi-valued entries in the "genre" and "directed_by" columns are split by the pipe (|) character, used in this file as a separator. Telling Solr to split these columns this way will ensure proper indexing of the data.

Each command will produce output similar to the below seen while indexing JSON:

```
$ ./bin/post -c films example/films/films.json  
/bin/java -classpath /solr-8.1.0/dist/solr-core-8.1.0.jar -Dauto=yes -Dc=films -Ddata=files  
org.apache.solr.util.SimplePostTool example/films/films.json  
SimplePostTool version 5.0.0  
Posting files to [base] url http://localhost:8983/solr/films/update...  
Entering auto mode. File endings considered are  
xml,json,jsonl,csv,pdf,doc,docx,ppt,pptx,xls,xlsx,odt,odp,ods,ott,otp,ots,rtf,htm,html,txt,log  
POSTing file films.json (application/json) to [base]/json/docs  
1 files indexed.  
COMMITting Solr index changes to http://localhost:8983/solr/films/update...  
Time spent: 0:00:00.878
```

Hooray!

If you go to the Query screen in the Admin UI for films (<http://localhost:8983/solr/#/films/query>) and hit **[Execute Query]** you should see 1100 results, with the first 10 returned to the screen.

Let's do a query to see if the "catchall" field worked properly. Enter "comedy" in the q box and hit **[Execute Query]** again. You should see get 417 results. Feel free to play around with other searches before we move on to facetting.

Faceting

One of Solr's most popular features is facetting. Facetting allows the search results to be arranged into subsets (or buckets, or categories), providing a count for each subset. There are several types of facetting: field values, numeric and date ranges, pivots (decision tree), and arbitrary query facetting.

Field Facets

In addition to providing search results, a Solr query can return the number of documents that contain each unique value in the whole result set.

On the Admin UI Query tab, if you check the facet checkbox, you'll see a few facet-related options appear:

The screenshot shows the 'Query' tab of the Solr Admin UI. A checkbox labeled 'facet' is checked. Below it are three input fields: 'facet.query' (empty), 'facet.field' (empty), and 'facet.prefix' (empty).

Facet options in the Query screen

To see facet counts from all documents (`q=*:*`): turn on facetting (`facet=true`), and specify the field to facet on via the `facet.field` parameter. If you only want facets, and no document contents, specify `rows=0`. The curl command below will return facet counts for the `genre_str` field:

```
curl "http://localhost:8983/solr/films/select?q=*:*&rows=0&facet=true&facet.field=genre_str"
```

In your terminal, you'll see something like:

```
{
  "responseHeader": {
    "zkConnected": true,
    "status": 0,
    "QTime": 11,
    "params": {
      "q": "*:*",
      "facet.field": "genre_str",
      "rows": "0",
      "facet": "true"
    }
  },
  "response": {
    "numFound": 1100,
    "start": 0,
    "maxScore": 1.0,
    "docs": []
  },
  "facet_counts": {
    "facet_queries": {},
    "facet_fields": {
      "genre_str": [
        "Drama", 552,
        "Comedy", 389,
        "Romance Film", 270,
        "Thriller", 259,
        "Action Film", 196,
        "Crime Fiction", 170,
        "World cinema", 167
      ],
      "facet_ranges": {},
      "facet_intervals": {},
      "facet_heatmaps": {}
    }
  }
}
```

We've truncated the output here a little bit, but in the facet_counts section, you see by default you get a count of the number of documents using each genre for every genre in the index. Solr has a parameter facet.mincount that you could use to limit the facets to only those that contain a certain number of documents (this parameter is not shown in the UI). Or, perhaps you do want all the facets, and you'll let your application's front-end control how it's displayed to users.

If you wanted to control the number of items in a bucket, you could do something like this:

```
curl
"http://localhost:8983/solr/films/select?=&q=*:*&facet.field=genre_str&facet.mincount=200&facet=on&rows=0"
```

You should only see 4 facets returned.

There are a great deal of other parameters available to help you control how Solr constructs the facets and facet lists. We'll cover some of them in this exercise, but you can also see the section [Faceting](#) for more detail.

Range Facets

For numerics or dates, it's often desirable to partition the facet counts into ranges rather than discrete values. A prime example of numeric range facetting, using the example techproducts data from our previous exercise, is price. In the /browse UI, it looks like this:

price

[0.0 - 50.0](#) (19)
[50.0 - 100.0](#) (1)
[150.0 - 200.0](#) (2)
[250.0 - 300.0](#) (1)
[300.0 - 350.0](#) (1)
[350.0 - 400.0](#) (2)
[450.0 - 500.0](#) (1)
[More than 600.0](#) (2)

Range facets

The films data includes the release date for films, and we could use that to create date range facets, which are another common use for range facets.

The Solr Admin UI doesn't yet support range facet options, so you will need to use curl or similar command line tool for the following examples.

If we construct a query that looks like this:

```
curl 'http://localhost:8983/solr/films/select?q=*&rows=0' \
  '&facet=true' \
  '&facet.range=initial_release_date' \
  '&facet.range.start=NOW-20YEAR' \
  '&facet.range.end=NOW' \
  '&facet.range.gap=%2B1YEAR'
```

This will request all films and ask for them to be grouped by year starting with 20 years ago (our earliest release date is in 2000) and ending today. Note that this query again URL encodes a + as %2B.

In the terminal you will see:

```
{  
  "responseHeader":{  
    "zkConnected":true,  
    "status":0,  
    "QTime":8,  
    "params":{  
      "facet.range":"initial_release_date",  
      "facet.limit":"300",  
      "q":"*:*",  
      "facet.range.gap":"+1YEAR",  
      "rows":"0",  
      "facet":"on",  
      "facet.range.start":"NOW-20YEAR",  
      "facet.range.end":"NOW"}},  
  "response":{ "numFound":1100, "start":0, "maxScore":1.0, "docs":[]}  
},  
  "facet_counts":{  
    "facet_queries":{},  
    "facet_fields":{},  
    "facet_ranges":{  
      "initial_release_date":{  
        "counts": [  
          "1997-07-28T17:12:06.919Z", 0,  
          "1998-07-28T17:12:06.919Z", 0,  
          "1999-07-28T17:12:06.919Z", 48,  
          "2000-07-28T17:12:06.919Z", 82,  
          "2001-07-28T17:12:06.919Z", 103,  
          "2002-07-28T17:12:06.919Z", 131,  
          "2003-07-28T17:12:06.919Z", 137,  
          "2004-07-28T17:12:06.919Z", 163,  
          "2005-07-28T17:12:06.919Z", 189,  
          "2006-07-28T17:12:06.919Z", 92,  
          "2007-07-28T17:12:06.919Z", 26,  
          "2008-07-28T17:12:06.919Z", 7,  
          "2009-07-28T17:12:06.919Z", 3,  
          "2010-07-28T17:12:06.919Z", 0,  
          "2011-07-28T17:12:06.919Z", 0,  
          "2012-07-28T17:12:06.919Z", 1,  
          "2013-07-28T17:12:06.919Z", 1,  
          "2014-07-28T17:12:06.919Z", 1,  
          "2015-07-28T17:12:06.919Z", 0,  
          "2016-07-28T17:12:06.919Z", 0],  
        "gap": "+1YEAR",  
        "start": "1997-07-28T17:12:06.919Z",  
        "end": "2017-07-28T17:12:06.919Z"}},  
    "facet_intervals":{},  
    "facet_heatmaps":{}}}
```

Pivot Facets

Another faceting type is pivot facets, also known as "decision trees", allowing two or more fields to be

nested for all the various possible combinations. Using the films data, pivot facets can be used to see how many of the films in the "Drama" category (the genre_str field) are directed by a director. Here's how to get at the raw data for this scenario:

```
curl  
"http://localhost:8983/solr/films/select?q=*&rows=0&facet=on&facet.pivot=genre_str,directed_by_str"
```

This results in the following response, which shows a facet for each category and director combination:

```
{"responseHeader":{  
    "zkConnected":true,  
    "status":0,  
    "QTime":1147,  
    "params":{  
        "q":"*:*",  
        "facet.pivot":"genre_str,directed_by_str",  
        "rows":0,  
        "facet":on}},  
    "response":{ "numFound":1100, "start":0, "maxScore":1.0, "docs":[]},  
    "facet_counts":{  
        "facet_queries":{},  
        "facet_fields":{},  
        "facet_ranges":{},  
        "facet_intervals":{},  
        "facet_heatmaps":{},  
        "facet_pivot":{  
            "genre_str,directed_by_str":[]  
                "field":genre_str,  
                "value":Drama,  
                "count":552,  
                "pivot":[]  
                    "field":directed_by_str,  
                    "value":Ridley Scott,  
                    "count":5},  
                {  
                    "field":directed_by_str,  
                    "value":Steven Soderbergh,  
                    "count":5},  
                {  
                    "field":directed_by_str,  
                    "value":Michael Winterbottom,  
                    "count":4}}]}}}
```

We've truncated this output as well - you will see a lot of genres and directors in your screen.

Exercise 2 Wrap Up

In this exercise, we learned a little bit more about how Solr organizes data in the indexes, and how to work with the Schema API to manipulate the schema file. We also learned a bit about facets in Solr, including

range facets and pivot facets. In both of these things, we've only scratched the surface of the available options. If you can dream it, it might be possible!

Like our previous exercise, this data may not be relevant to your needs. We can clean up our work by deleting the collection. To do that, issue this command at the command line:

```
bin/solr delete -c films
```

Exercise 3: Index Your Own Data

For this last exercise, work with a dataset of your choice. This can be files on your local hard drive, a set of data you have worked with before, or maybe a sample of the data you intend to index to Solr for your production application.

This exercise is intended to get you thinking about what you will need to do for your application:

- What sorts of data do you need to index?
- What will you need to do to prepare Solr for your data (such as, create specific fields, set up copy fields, determine analysis rules, etc.)
- What kinds of search options do you want to provide to users?
- How much testing will you need to do to ensure everything works the way you expect?

Create Your Own Collection

Before you get started, create a new collection, named whatever you'd like. In this example, the collection will be named "localDocs"; replace that name with whatever name you choose if you want to.

```
./bin/solr create -c localDocs -s 2 -rf 2
```

Again, as we saw from Exercise 2 above, this will use the _default configset and all the schemaless features it provides. As we noted previously, this may cause problems when we index our data. You may need to iterate on indexing a few times before you get the schema right.

Indexing Ideas

Solr has lots of ways to index data. Choose one of the approaches below and try it out with your system:

Local Files with bin/post

If you have a local directory of files, the Post Tool (bin/post) can index a directory of files. We saw this in action in our first exercise.

We used only JSON, XML and CSV in our exercises, but the Post Tool can also handle HTML, PDF, Microsoft Office formats (such as MS Word), plain text, and more.

In this example, assume there is a directory named "Documents" locally. To index it, we would issue a command like this (correcting the collection name after the -c parameter as needed):

```
./bin/post -c localDocs ~/Documents
```

You may get errors as it works through your documents. These might be caused by the field guessing, or the file type may not be supported. Indexing content such as this demonstrates the need to plan Solr for

your data, which requires understanding it and perhaps also some trial and error.

DataImportHandler

Solr includes a tool called the [Data Import Handler \(DIH\)](#) which can connect to databases (if you have a jdbc driver), mail servers, or other structured data sources. There are several examples included for feeds, GMail, and a small HSQL database.

The README.txt file in example/example-DIH will give you details on how to start working with this tool.

SolrJ

SolrJ is a Java-based client for interacting with Solr. Use [SolrJ](#) for JVM-based languages or other Solr clients to programmatically create documents to send to Solr.

Documents Screen

Use the Admin UI [Documents tab](#) (at <http://localhost:8983/solr/#/localDocs/documents>) to paste in a document to be indexed, or select Document Builder from the Document Type dropdown to build a document one field at a time. Click on the **[Submit Document]** button below the form to index your document.

Updating Data

You may notice that even if you index content in this tutorial more than once, it does not duplicate the results found. This is because the example Solr schema (a file named either managed-schema or schema.xml) specifies a uniqueKey field called id. Whenever you POST commands to Solr to add a document with the same value for the uniqueKey as an existing document, it automatically replaces it for you.

You can see that that has happened by looking at the values for numDocs and maxDoc in the core-specific Overview section of the Solr Admin UI.

numDocs represents the number of searchable documents in the index (and will be larger than the number of XML, JSON, or CSV files since some files contained more than one document). The maxDoc value may be larger as the maxDoc count includes logically deleted documents that have not yet been physically removed from the index. You can re-post the sample files over and over again as much as you want and numDocs will never increase, because the new documents will constantly be replacing the old.

Go ahead and edit any of the existing example data files, change some of the data, and re-run the PostTool (bin/post). You'll see your changes reflected in subsequent searches.

Deleting Data

If you need to iterate a few times to get your schema right, you may want to delete documents to clear out the collection and try again. Note, however, that merely removing documents doesn't change the underlying field definitions. Essentially, this will allow you to reindex your data after making changes to fields for your needs.

You can delete data by POSTing a delete command to the update URL and specifying the value of the document's unique key field, or a query that matches multiple documents (be careful with that one!). We can use bin/post to delete documents also if we structure the request properly.

Execute the following command to delete a specific document:

```
bin/post -c localDocs -d "<delete><id>SP2514N</id></delete>"
```

To delete all documents, you can use "delete-by-query" command like:

```
bin/post -c localDocs -d "<delete><query>*:*</query></delete>"
```

You can also modify the above to only delete documents that match a specific query.

Exercise 3 Wrap Up

At this point, you're ready to start working on your own.

Jump ahead to the overall [wrap up](#) when you're ready to stop Solr and remove all the examples you worked with and start fresh.

Spatial Queries

Solr has sophisticated geospatial support, including searching within a specified distance range of a given location (or within a bounding box), sorting by distance, or even boosting results by the distance.

Some of the example techproducts documents we indexed in Exercise 1 have locations associated with them to illustrate the spatial capabilities. To reindex this data, see [Exercise 1](#).

Spatial queries can be combined with any other types of queries, such as in this example of querying for "ipod" within 10 kilometers from San Francisco:

The screenshot shows the Solr search interface with the 'Spatial' tab selected. The search bar contains 'ipod'. Below it, there are options for 'Boost by Price' (unchecked) and a 'Location Filter' set to 'San Francisco CA'. A 'Distance (KM)' input field is set to '10'. The search results page shows one result: 'iPod & iPod Mini USB 2.0 Cable' with a 'More Like This' link. The result details include: Id: IW-02, Price: 11.50, USD, Features: car power adapter for iPod, white, and In Stock: false. To the right of the result is a map of San Francisco with a red dot indicating the location of the result. A link 'Larger Map' is provided.

Spatial queries and results

This is from Solr's example search UI (called /browse), which has a nice feature to show a map for each item and allow easy selection of the location to search near. You can see this yourself by going to <http://localhost:8983/solr/techproducts/browse?q=ipod&pt=37.7752%2C-122.4232&d=10&sfield=store&fq=%7B%21bbox%7D&queryOpts=spatial&queryOpts=spatial> in a browser.

To learn more about Solr's spatial capabilities, see the section [Spatial Search](#).

Wrapping Up

If you've run the full set of commands in this quick start guide you have done the following:

- Launched Solr into SolrCloud mode, two nodes, two collections including shards and replicas
- Indexed several types of files
- Used the Schema API to modify your schema
- Opened the admin console, used its query interface to get results
- Opened the /browse interface to explore Solr's features in a more friendly and familiar interface

Nice work!

Cleanup

As you work through this tutorial, you may want to stop Solr and reset the environment back to the starting point. The following command line will stop Solr and remove the directories for each of the two nodes that were created all the way back in Exercise 1:

```
bin/solr stop -all ; rm -Rf example/cloud/
```

Where to next?

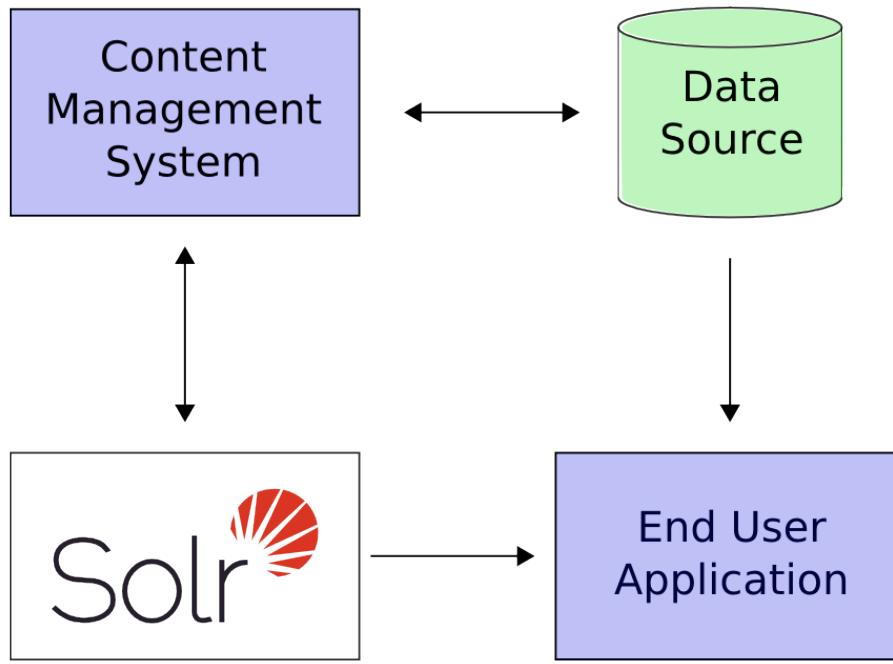
This Guide will be your best resource for learning more about Solr.

Solr also has a robust community made up of people happy to help you get started. For more information, check out the Solr website's [Resources page](#).

A Quick Overview

Solr is a search server built on top of Apache Lucene, an open source, Java-based, information retrieval library. It is designed to drive powerful document retrieval applications - wherever you need to serve data to users based on their queries, Solr can work for you.

Here is an example of how Solr could integrate with an application:



Solr integration with applications

In the scenario above, Solr runs alongside other server applications. For example, an online store application would provide a user interface, a shopping cart, and a way to make purchases for end users; while an inventory management application would allow store employees to edit product information. The product metadata would be kept in some kind of database, as well as in Solr.

Solr makes it easy to add the capability to search through the online store through the following steps:

1. Define a *schema*. The schema tells Solr about the contents of documents it will be indexing. In the online store example, the schema would define fields for the product name, description, price, manufacturer, and so on. Solr's schema is powerful and flexible and allows you to tailor Solr's behavior to your application. See [Documents, Fields, and Schema Design](#) for all the details.
2. Feed Solr documents for which your users will search.
3. Expose search functionality in your application.

Because Solr is based on open standards, it is highly extensible. Solr queries are simple HTTP request URLs and the response is a structured document: mainly JSON, but it could also be XML, CSV, or other formats. This means that a wide variety of clients will be able to use Solr, from other web applications to browser clients, rich client applications, and mobile devices. Any platform capable of HTTP can talk to Solr. See [Client APIs](#) for details on client APIs.

Solr offers support for the simplest keyword searching through to complex queries on multiple fields and faceted search results. [Searching](#) has more information about searching and queries.

If Solr's capabilities are not impressive enough, its ability to handle very high-volume applications should do the trick.

A relatively common scenario is that you have so much data, or so many queries, that a single Solr server is unable to handle your entire workload. In this case, you can scale up the capabilities of your application using [SolrCloud](#) to better distribute the data, and the processing of requests, across many servers. Multiple options can be mixed and matched depending on the scalability you need.

For example: "Sharding" is a scaling technique in which a collection is split into multiple logical pieces called "shards" in order to scale up the number of documents in a collection beyond what could physically fit on a single server. Incoming queries are distributed to every shard in the collection, which respond with merged results. Another technique available is to increase the "Replication Factor" of your collection, which allows you to add servers with additional copies of your collection to handle higher concurrent query load by spreading the requests around to multiple machines. Sharding and replication are not mutually exclusive, and together make Solr an extremely powerful and scalable platform.

Best of all, this talk about high-volume applications is not just hypothetical: some of the famous Internet sites that use Solr today are Macy's, EBay, and Zappo's. For more examples, take a look at <https://wiki.apache.org/solr/PublicServers>.

Solr System Requirements

You can install Solr in any system where a suitable Java Runtime Environment (JRE) is available.

Installation Requirements

Supported Operating Systems

Solr is tested on several versions of Linux, macOS and Windows.

Java Requirements

You will need the Java Runtime Environment (JRE) version 1.8 or higher. At a command line, check your Java version like this:

```
$ java -version
java version "1.8.0_60"
Java(TM) SE Runtime Environment (build 1.8.0_60-b27)
Java HotSpot(TM) 64-Bit Server VM (build 25.60-b23, mixed mode)
```

The exact output will vary, but you need to make sure you meet the minimum version requirement. We also recommend choosing a version that is not end-of-life from its vendor. Oracle/OpenJDK are the most tested JREs and are preferred. It's also preferred to use the latest available official release.

Some versions of Java VM have bugs that may impact your implementation. To be sure, check the page [Lucene Java Bugs](#).

Sources for Java

Java is available from a number of providers. Lucene and Solr regularly test with [OpenJDK](#) nad Oracle versions of Java. Some are free, others have a cost, some provide security patches and support, others do not. We recommend you read the article [Java is still free by Java Champions](#) to help you decide.

The Lucene project does not endorse any particular provider of Java.



While we reference the Java Development (JDK) on this page, any Java Runtime Environment (JRE) associated with the referenced JDKs is acceptable.

Java and Lucene/Solr Combinations

Each Lucene/Solr release has an extensively tested minimum Java version. For instance the minimum Java version for Solr 8 is Java 8. This section provides guidance when running Lucene/Solr with a more recent Java version than the minimum specified.

- OpenJDK and Oracle Java distributions are tested extensively and will continue to be tested going forward.
 - Distributions of Java from other sources are not regularly tested by our testing infrastructure,

therefore you must test Java from those sources in your environment.

- For the purposes of Lucene and Solr, Oracle's Java and OpenJDK are identical.
- Upgrading Java is not required with the understanding that no Java bugs will be addressed unless you are using a version of Java that provides LTS.
- Java 8 has been extensively tested by both automated tests and users through Solr 8. Long Term Support (LTS) for Java 8 is provided by some sources, see [Java is still free](#).
- The project's testing infrastructure continuously tests with the minimum and greater versions of Java for each development branch.
- Java 9 and 10 have no LTS. For this reason, Java 11 is preferred over 9 or 10 when upgrading Java.
- For specific questions the [Solr User's List](#) is a great resource.

Project Testing of Java-Solr Combinations

Solr and Lucene run a continuous integration model, running automated unit and integration tests using several versions of Java. In addition, some organizations also maintain their own test infrastructure and feed their results back to the community.

Our continuous testing is against the two code lines under active development, Solr 8x and the future Solr 9.0:

- Lucene/Solr 8.x is the current stable release line and will have "point releases", i.e., 8.1, 8.2, etc. until Lucene/Solr 9.0 is released.
 - Solr 8.x is currently tested against Java 8, 9, 10, 11, 12 and (pre-release) 13.
- There is also development and testing with the future Lucene/Solr 9.x release line, which will require Java 11 as a minimum version. This line is currently tested against Java 11, 12 and (pre-release) 13.
- Lucene/Solr 7.x and earlier release lines are not tested on a continuous basis.

Released Lucene/Solr and Java Versions

The success rate in our automated tests is similar with all the Java versions tested with the following caveats.

Lucene/Solr Prior to 7.0

- Lucene/Solr 7.0 was the first version that successfully passed our tests using Java 9 and higher. You should avoid Java 9 or later for Lucene/Solr 6.x or earlier.

Lucene/Solr 7.x

- Requires Java 8 or higher.
- This version had continuous testing with Java 9, 10, 11, 12 and the pre-release version of Java 13. Regular testing stopped when Lucene/Solr 8.0 was released.
- Hadoop with Java 9+ may not work in all situations, test in your environment.
- Kerberos with Java 9+ may not work in all situations, test in your environment.
- Be sure to test with SSL/TLS and/or authorization enabled in your environment if you require either when using Java 9+.

Lucene/Solr 8.x

- Requires Java 8 or higher.
- This version has continuous testing with Java 9, 10, 11, 12 and the pre-release version of Java 13.
- There were known issues with Kerberos with Java 9+ prior to Solr 8.1. If using 8.0, you should test in your environment.
- Be sure to test with SSL/TLS and/or authorization enabled in your environment if you require either when using Java 9+.

Installing Solr

Installation of Solr on Unix-compatible or Windows servers generally requires simply extracting (or, unzipping) the download package.

Please be sure to review the [Solr System Requirements](#) before starting Solr.

Available Solr Packages

Solr is available from the Solr website. Download the latest release <https://lucene.apache.org/solr/mirrors-solr-latest-redir.html>.

There are three separate packages:

- solr-8.1.0.tgz for Linux/Unix/OSX systems
- solr-8.1.0.zip for Microsoft Windows systems
- solr-8.1.0-src.tgz the package Solr source code. This is useful if you want to develop on Solr without using the official Git repository.

Preparing for Installation

When getting started with Solr, all you need to do is extract the Solr distribution archive to a directory of your choosing. This will suffice as an initial development environment, but take care not to overtax this "toy" installation before setting up your true development and production environments.

When you've progressed past initial evaluation of Solr, you'll want to take care to plan your implementation. You may need to reinstall Solr on another server or make a clustered SolrCloud environment.

When you're ready to setup Solr for a production environment, please refer to the instructions provided on the [Taking Solr to Production](#) page.

What Size Server Do I Need?

How to size your Solr installation is a complex question that relies on a number of factors, including the number and structure of documents, how many fields you intend to store, the number of users, etc.



It's highly recommended that you spend a bit of time thinking about the factors that will impact hardware sizing for your Solr implementation. A very good blog post that discusses the issues to consider is [Sizing Hardware in the Abstract: Why We Don't have a Definitive Answer](#).

One thing to note when planning your installation is that a hard limit exists in Lucene for the number of documents in a single index: approximately 2.14 billion documents (2,147,483,647 to be exact). In practice, it is highly unlikely that such a large number of documents would fit and perform well in a single index, and you will likely need to distribute your index across a cluster before you ever approach this number. If you know you will exceed this number of documents in total before you've even started indexing, it's best to plan your installation with [SolrCloud](#) as part of your design from the start.

Package Installation

To keep things simple for now, extract the Solr distribution archive to your local home directory, for instance on Linux, do:

```
cd ~/  
tar zxf solr-8.1.0.tgz
```

Once extracted, you are now ready to run Solr using the instructions provided in the [Starting Solr](#) section below.

Directory Layout

After installing Solr, you'll see the following directories and files within them:

bin/

This directory includes several important scripts that will make using Solr easier.

solr and solr.cmd

This is [Solr's Control Script](#), also known as bin/solr (*nix) / bin/solr.cmd (Windows). This script is the preferred tool to start and stop Solr. You can also create collections or cores, configure authentication, and work with configuration files when running in SolrCloud mode.

post

The [PostTool](#), which provides a simple command line interface for POSTing content to Solr.

solr.in.sh and solr.in.cmd

These are property files for *nix and Windows systems, respectively. System-level properties for Java, Jetty, and Solr are configured here. Many of these settings can be overridden when using bin/solr / bin/solr.cmd, but this allows you to set all the properties in one place.

install_solr_services.sh

This script is used on *nix systems to install Solr as a service. It is described in more detail in the section [Taking Solr to Production](#).

contrib/

Solr's contrib directory includes add-on plugins for specialized features of Solr.

dist/

The dist directory contains the main Solr .jar files.

docs/

The docs directory includes a link to online Javadocs for Solr.

example/

The example directory includes several types of examples that demonstrate various Solr capabilities. See the section [Solr Examples](#) below for more details on what is in this directory.

licenses/

The licenses directory includes all of the licenses for 3rd party libraries used by Solr.

server/

This directory is where the heart of the Solr application resides. A README in this directory provides a detailed overview, but here are some highlights:

- Solr's Admin UI (server/solr-webapp)
- Jetty libraries (server/lib)
- Log files (server/logs) and log configurations (server/resources). See the section [Configuring Logging](#) for more details on how to customize Solr's default logging.
- Sample configsets (server/solr/configsets)

Solr Examples

Solr includes a number of example documents and configurations to use when getting started. If you ran through the [Solr Tutorial](#), you have already interacted with some of these files.

Here are the examples included with Solr:

exampledocs

This is a small set of simple CSV, XML, and JSON files that can be used with bin/post when first getting started with Solr. For more information about using bin/post with these files, see [Post Tool](#).

example-DIH

This directory includes a few example DataImport Handler (DIH) configurations to help you get started with importing structured content in a database, an email server, or even an Atom feed. Each example will index a different set of data; see the README there for more details about these examples.

files

The files directory provides a basic search UI for documents such as Word or PDF that you may have stored locally. See the README there for details on how to use this example.

films

The films directory includes a robust set of data about movies in three formats: CSV, XML, and JSON. See the README there for details on how to use this dataset.

Starting Solr

Solr includes a command line interface tool called bin/solr (Linux/MacOS) or bin\solr.cmd (Windows). This tool allows you to start and stop Solr, create cores and collections, configure authentication, and check the status of your system.

To use it to start Solr you can simply enter:

```
bin/solr start
```

If you are running Windows, you can start Solr by running bin\solr.cmd instead.

```
bin\solr.cmd start
```

This will start Solr in the background, listening on port 8983.

When you start Solr in the background, the script will wait to make sure Solr starts correctly before returning to the command line prompt.



All of the options for the Solr CLI are described in the section [Solr Control Script Reference](#).

Start Solr with a Specific Bundled Example

Solr also provides a number of useful examples to help you learn about key features. You can launch the examples using the -e flag. For instance, to launch the "techproducts" example, you would do:

```
bin/solr -e techproducts
```

Currently, the available examples you can run are: techproducts, dih, schemaless, and cloud. See the section [Running with Example Configurations](#) for details on each example.



Getting Started with SolrCloud

Running the cloud example starts Solr in [SolrCloud](#) mode. For more information on starting Solr in SolrCloud mode, see the section [Getting Started with SolrCloud](#).

Check if Solr is Running

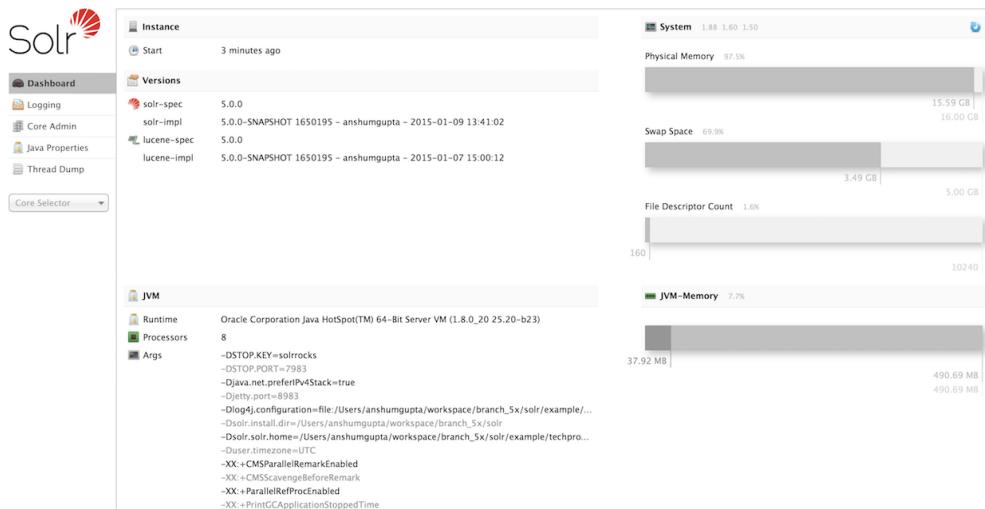
If you're not sure if Solr is running locally, you can use the status command:

```
bin/solr status
```

This will search for running Solr instances on your computer and then gather basic information about them, such as the version and memory usage.

That's it! Solr is running. If you need convincing, use a Web browser to see the Admin Console.

<http://localhost:8983/solr/>



The Solr Admin interface.

If Solr is not running, your browser will complain that it cannot connect to the server. Check your port number and try again.

Create a Core

If you did not start Solr with an example configuration, you would need to create a core in order to be able to index and search. You can do so by running:

```
bin/solr create -c <name>
```

This will create a core that uses a data-driven schema which tries to guess the correct field type when you add documents to the index.

To see all available options for creating a new core, execute:

```
bin/solr create -help
```

Deployment and Operations

An important aspect of Solr is that all operations and deployment can be done online, with minimal or no impact to running applications. This includes minor upgrades and provisioning and removing nodes, backing up and restoring indexes and editing configurations.

Common administrative tasks include:

[Solr Control Script Reference](#): This section provides information about all of the options available to the bin/solr / bin\solr.cmd scripts, which can start and stop Solr, configure authentication, and create or remove collections and cores.

[Solr Configuration Files](#): Overview of the installation layout and major configuration files.

[Taking Solr to Production](#): Detailed steps to help you install Solr as a service and take your application to production.

[Making and Restoring Backups](#): Describes backup strategies for your Solr indexes.

[Running Solr on HDFS](#): How to use HDFS to store your Solr indexes and transaction logs.

[SolrCloud on AWS EC2](#): A tutorial on deploying Solr in Amazon Web Services (AWS) using EC2 instances.

[Upgrading a Solr Cluster](#): Information for upgrading a production SolrCloud cluster.

[Solr Upgrade Notes](#): Information about changes made in Solr releases.

Solr Control Script Reference

Solr includes a script known as “bin/solr” that allows you to perform many common operations on your Solr installation or cluster.

You can start and stop Solr, create and delete collections or cores, perform operations on ZooKeeper and check the status of Solr and configured shards.

You can find the script in the bin/ directory of your Solr installation. The bin/solr script makes Solr easier to work with by providing simple commands and options to quickly accomplish common goals.

More examples of bin/solr in use are available throughout the Solr Reference Guide, but particularly in the sections [Starting Solr](#) and [Getting Started with SolrCloud](#).

Starting and Stopping

Start and Restart

The start command starts Solr. The restart command allows you to restart Solr while it is already running or if it has been stopped already.

The start and restart commands have several options to allow you to run in SolrCloud mode, use an example configuration set, start with a hostname or port that is not the default and point to a local ZooKeeper ensemble.

```
bin/solr start [options]
bin/solr start -help
bin/solr restart [options]
bin/solr restart -help
```

When using the restart command, you must pass all of the parameters you initially passed when you started Solr. Behind the scenes, a stop request is initiated, so Solr will be stopped before being started again. If no nodes are already running, restart will skip the step to stop and proceed to starting Solr.

Start Parameters

The bin/solr script provides many options to allow you to customize the server in common ways, such as changing the listening port. However, most of the defaults are adequate for most Solr installations, especially when just getting started.

-a "<string>"

Start Solr with additional JVM parameters, such as those starting with -X. If you are passing JVM parameters that begin with "-D", you can omit the -a option.

Example:

```
bin/solr start -a "-Xdebug -Xrunjdwp:transport=dt_socket, server=y, suspend=n,address=1044"
```

-cloud

Start Solr in SolrCloud mode, which will also launch the embedded ZooKeeper instance included with Solr.

This option can be shortened to simply -c.

If you are already running a ZooKeeper ensemble that you want to use instead of the embedded (single-node) ZooKeeper, you should also either specify ZK_HOST in solr.in.sh/solr.in.cmd (see [instructions](#)) or pass the -z parameter.

For more details, see the section [SolrCloud Mode](#) below.

Example: bin/solr start -c

-d <dir>

Define a server directory, defaults to server (as in, \$SOLR_HOME/server). It is uncommon to override this option. When running multiple instances of Solr on the same host, it is more common to use the same server directory for each instance and use a unique Solr home directory using the -s option.

Example: bin/solr start -d newServerDir

-e <name>

Start Solr with an example configuration. These examples are provided to help you get started faster with Solr generally, or just try a specific feature.

The available options are:

- cloud
- techproducts
- dih
- schemaless

See the section [Running with Example Configurations](#) below for more details on the example configurations.

Example: bin/solr start -e schemaless

-f

Start Solr in the foreground; you cannot use this option when running examples with the -e option.

Example: bin/solr start -f

-h <hostname>

Start Solr with the defined hostname. If this is not specified, 'localhost' will be assumed.

Example: bin/solr start -h search.mysolr.com

-m <memory>

Start Solr with the defined value as the min (-Xms) and max (-Xmx) heap size for the JVM.

Example: bin/solr start -m 1g

-noprompt

Start Solr and suppress any prompts that may be seen with another option. This would have the side effect of accepting all defaults implicitly.

For example, when using the "cloud" example, an interactive session guides you through several options for your SolrCloud cluster. If you want to accept all of the defaults, you can simply add the -noprompt option to your request.

Example: bin/solr start -e cloud -noprompt

-p <port>

Start Solr on the defined port. If this is not specified, '8983' will be used.

Example: bin/solr start -p 8655

-s <dir>

Sets the solr.solr.home system property; Solr will create core directories under this directory. This allows you to run multiple Solr instances on the same host while reusing the same server directory set using the -d parameter.

If set, the specified directory should contain a solr.xml file, unless solr.xml exists in ZooKeeper. The default value is server/solr.

This parameter is ignored when running examples (-e), as the solr.solr.home depends on which example is run.

Example: bin/solr start -s newHome

-v

Be more verbose. This changes the logging level of log4j from INFO to DEBUG, having the same effect as if you edited log4j2.xml accordingly.

Example: bin/solr start -f -v

-q

Be more quiet. This changes the logging level of log4j from INFO to WARN, having the same effect as if you edited log4j2.xml accordingly. This can be useful in a production setting where you want to limit logging to warnings and errors.

Example: bin/solr start -f -q

-V

Start Solr with verbose messages from the start script.

Example: bin/solr start -V

-z <zkHost>

Start Solr with the defined ZooKeeper connection string. This option is only used with the -c option, to start Solr in SolrCloud mode. If ZK_HOST is not specified in solr.in.sh/solr.in.cmd and this option is not provided, Solr will start the embedded ZooKeeper instance and use that instance for SolrCloud operations.

Example: bin/solr start -c -z server1:2181,server2:2181

-force

If attempting to start Solr as the root user, the script will exit with a warning that running Solr as "root" can cause problems. It is possible to override this warning with the -force parameter.

Example: sudo bin/solr start -force

To emphasize how the default settings work take a moment to understand that the following commands are equivalent:

bin/solr start

bin/solr start -h localhost -p 8983 -d server -s solr -m 512m

It is not necessary to define all of the options when starting if the defaults are fine for your needs.

Setting Java System Properties

The bin/solr script will pass any additional parameters that begin with -D to the JVM, which allows you to set arbitrary Java system properties.

For example, to set the auto soft-commit frequency to 3 seconds, you can do:

bin/solr start -Dsolr.autoSoftCommit.maxTime=3000

SolrCloud Mode

The -c and -cloud options are equivalent:

bin/solr start -c

bin/solr start -cloud

If you specify a ZooKeeper connection string, such as -z 192.168.1.4:2181, then Solr will connect to ZooKeeper and join the cluster.



If you have defined ZK_HOST in solr.in.sh/solr.in.cmd (see [instructions](#)) you can omit -z <zookeeper host string> from all bin/solr commands.

When starting Solr in SolrCloud mode, if you do not define 'ZK_HOST' in 'solr.in.sh'/'solr.in.cmd` nor specify the '-z' option, then Solr will launch an embedded ZooKeeper server listening on the Solr port + 1000, i.e., if Solr is running on port 8983, then the embedded ZooKeeper will be listening on port 9983.

If your ZooKeeper connection string uses a chroot, such as localhost:2181/solr, then you need to create the /solr znode before launching SolrCloud using the bin/solr script.

+ To do this use the mkroot command outlined below, for example: bin/solr zk mkroot /solr -z 192.168.1.4:2181

When starting in SolrCloud mode, the interactive script session will prompt you to choose a configset to use.

For more information about starting Solr in SolrCloud mode, see also the section [Getting Started with SolrCloud](#).

Running with Example Configurations

```
bin/solr start -e <name>
```

The example configurations allow you to get started quickly with a configuration that mirrors what you hope to accomplish with Solr.

Each example launches Solr with a managed schema, which allows use of the [Schema API](#) to make schema edits, but does not allow manual editing of a Schema file.

If you would prefer to manually modify a schema.xml file directly, you can change this default as described in the section [Schema Factory Definition in SolrConfig](#).

Unless otherwise noted in the descriptions below, the examples do not enable [SolrCloud](#) nor [schemaless mode](#).

The following examples are provided:

- **cloud**: This example starts a 1-4 node SolrCloud cluster on a single machine. When chosen, an interactive session will start to guide you through options to select the initial configset to use, the number of nodes for your example cluster, the ports to use, and name of the collection to be created.

When using this example, you can choose from any of the available configsets found in \$SOLR_HOME/server/solr/configsets.

- **techproducts**: This example starts Solr in standalone mode with a schema designed for the sample documents included in the \$SOLR_HOME/example/exampledocs directory.

The configset used can be found in \$SOLR_HOME/server/solr/configsets/sample_techproducts_configs.

- **dih**: This example starts Solr in standalone mode with the DataImportHandler (DIH) enabled and several example dataconfig.xml files pre-configured for different types of data supported with DIH (such as, database contents, email, RSS feeds, etc.).

The configset used is customized for DIH, and is found in \$SOLR_HOME/example/example-DIH/solr/conf.

For more information about DIH, see the section [Uploading Structured Data Store Data with the Data Import Handler](#).

- **schemaless**: This example starts Solr in standalone mode using a managed schema, as described in the section [Schema Factory Definition in SolrConfig](#), and provides a very minimal pre-defined schema. Solr will run in [Schemaless Mode](#) with this configuration, where Solr will create fields in the schema on the fly and will guess field types used in incoming documents.

The configset used can be found in \$SOLR_HOME/server/solr/configsets/_default.



The run in-foreground option (-f) is not compatible with the -e option since the script needs to perform additional tasks after starting the Solr server.

Stop

The stop command sends a STOP request to a running Solr node, which allows it to shutdown gracefully. The command will wait up to 180 seconds for Solr to stop gracefully and then will forcefully kill the process (kill -9).

```
bin/solr stop [options]
```

```
bin/solr stop --help
```

Stop Parameters

-p <port>

Stop Solr running on the given port. If you are running more than one instance, or are running in SolrCloud mode, you either need to specify the ports in separate requests or use the -all option.

Example: bin/solr stop -p 8983

-all

Stop all running Solr instances that have a valid PID.

Example: bin/solr stop -all

-k <key>

Stop key used to protect from stopping Solr inadvertently; default is "solrrocks".

Example: bin/solr stop -k solrrocks

System Information

Version

The version command simply returns the version of Solr currently installed and immediately exists.

```
$ bin/solr version
X.Y.0
```

Status

The status command displays basic JSON-formatted information for any Solr nodes found running on the local system.

The status command uses the SOLR_PID_DIR environment variable to locate Solr process ID files to find running Solr instances, which defaults to the bin directory.

```
bin/solr status
```

The output will include a status of each node of the cluster, as in this example:

```
Found 2 Solr nodes:
```

```
Solr process 39920 running on port 7574
{
  "solr_home":"/Applications/Solr/example/cloud/node2/solr/",
  "version":"X.Y.0",
  "startTime":"2015-02-10T17:19:54.739Z",
  "uptime":"1 days, 23 hours, 55 minutes, 48 seconds",
  "memory":"77.2 MB (%15.7) of 490.7 MB",
  "cloud":{
    "ZooKeeper":"localhost:9865",
    "liveNodes":"2",
    "collections":"2"}}
}
```

```
Solr process 39827 running on port 8865
{
  "solr_home":"/Applications/Solr/example/cloud/node1/solr/",
  "version":"X.Y.0",
  "startTime":"2015-02-10T17:19:49.057Z",
  "uptime":"1 days, 23 hours, 55 minutes, 54 seconds",
  "memory":"94.2 MB (%19.2) of 490.7 MB",
  "cloud":{
    "ZooKeeper":"localhost:9865",
    "liveNodes":"2",
    "collections":"2"}}
}
```

Assert

The `assert` command sanity checks common issues with Solr installations. These include checking the ownership/existence of particular directories, and ensuring Solr is available on the expected URL. The command can either output a specified error message, or change its exit code to indicate errors.

As an example:

```
bin/solr assert --exists /opt/bin/solr
```

Results in the output below:

```
ERROR: Directory /opt/bin/solr does not exist.
```

Use `bin/solr assert -help` for a full list of options.

Healthcheck

The `healthcheck` command generates a JSON-formatted health report for a collection when running in SolrCloud mode. The health report provides information about the state of every replica for all shards in a collection, including the number of committed documents and its current state.

```
bin/solr healthcheck [options]
```

```
bin/solr healthcheck -help
```

Healthcheck Parameters

-c <collection>

Name of the collection to run a healthcheck against (required).

Example: bin/solr healthcheck -c gettingstarted

-z <zkhost>

ZooKeeper connection string, defaults to localhost:9983. If you are running Solr on a port other than 8983, you will have to specify the ZooKeeper connection string. By default, this will be the Solr port + 1000. Unnecessary if ZK_HOST is defined in solr.in.sh or solr.in.cmd.

Example: bin/solr healthcheck -z localhost:2181

Below is an example healthcheck request and response using a non-standard ZooKeeper connect string, with 2 nodes running:

```
$ bin/solr healthcheck -c gettingstarted -z localhost:9865
```

```
{  
  "collection": "gettingstarted",  
  "status": "healthy",  
  "numDocs": 0,  
  "numShards": 2,  
  "shards": [  
    {  
      "shard": "shard1",  
      "status": "healthy",  
      "replicas": [  
        {  
          "name": "core_node1",  
          "url": "http://10.0.1.10:8865/solr/gettingstarted_shard1_replica2/",  
          "numDocs": 0,  
          "status": "active",  
          "uptime": "2 days, 1 hours, 18 minutes, 48 seconds",  
          "memory": "25.6 MB (%5.2) of 490.7 MB",  
          "leader": true},  
        {  
          "name": "core_node4",  
          "url": "http://10.0.1.10:7574/solr/gettingstarted_shard1_replica1/",  
          "numDocs": 0,  
          "status": "active",  
          "uptime": "2 days, 1 hours, 18 minutes, 42 seconds",  
          "memory": "95.3 MB (%19.4) of 490.7 MB"}],  
    },  
    {  
      "shard": "shard2",  
      "status": "healthy",  
      "replicas": [  
        {  
          "name": "core_node2",  
          "url": "http://10.0.1.10:8865/solr/gettingstarted_shard2_replica2/",  
          "numDocs": 0,  
          "status": "active",  
          "uptime": "2 days, 1 hours, 18 minutes, 48 seconds",  
          "memory": "25.8 MB (%5.3) of 490.7 MB"},  
        {  
          "name": "core_node3",  
          "url": "http://10.0.1.10:7574/solr/gettingstarted_shard2_replica1/",  
          "numDocs": 0,  
          "status": "active",  
          "uptime": "2 days, 1 hours, 18 minutes, 42 seconds",  
          "memory": "95.4 MB (%19.4) of 490.7 MB",  
          "leader": true}]]}
```

Collections and Cores

The `bin/solr` script can also help you create new collections (in SolrCloud mode) or cores (in standalone mode), or delete collections.

Create a Core or Collection

The create command detects the mode that Solr is running in (standalone or SolrCloud) and then creates a core or collection depending on the mode.

```
bin/solr create [options]
```

```
bin/solr create -help
```

Create Core or Collection Parameters

-c <name>

Name of the core or collection to create (required).

Example: bin/solr create -c mycollection

-d <confdir>

The configuration directory. This defaults to _default.

See the section [Configuration Directories and SolrCloud](#) below for more details about this option when running in SolrCloud mode.

Example: bin/solr create -d _default

-n <configName>

The configuration name. This defaults to the same name as the core or collection.

Example: bin/solr create -n basic

-p <port>

Port of a local Solr instance to send the create command to; by default the script tries to detect the port by looking for running Solr instances.

This option is useful if you are running multiple standalone Solr instances on the same host, thus requiring you to be specific about which instance to create the core in.

Example: bin/solr create -p 8983

-s <shards> or -shards

Number of shards to split a collection into, default is 1; only applies when Solr is running in SolrCloud mode.

Example: bin/solr create -s 2

-rf <replicas> or -replicationFactor

Number of copies of each document in the collection. The default is 1 (no replication).

Example: bin/solr create -rf 2

-force

If attempting to run create as "root" user, the script will exit with a warning that running Solr or actions against Solr as "root" can cause problems. It is possible to override this warning with the -force parameter.

Example: bin/solr create -c foo -force

Configuration Directories and SolrCloud

Before creating a collection in SolrCloud, the configuration directory used by the collection must be uploaded to ZooKeeper. The create command supports several use cases for how collections and configuration directories work. The main decision you need to make is whether a configuration directory in ZooKeeper should be shared across multiple collections.

Let's work through a few examples to illustrate how configuration directories work in SolrCloud.

First, if you don't provide the -d or -n options, then the default configuration (`$/SOLR_HOME/server/solr/configsets/_default/conf`) is uploaded to ZooKeeper using the same name as the collection.

For example, the following command will result in the _default configuration being uploaded to /configs/contacts in ZooKeeper: `bin/solr create -c contacts`.

If you create another collection with `bin/solr create -c contacts2`, then another copy of the _default directory will be uploaded to ZooKeeper under /configs/contacts2.

Any changes you make to the configuration for the contacts collection will not affect the contacts2 collection. Put simply, the default behavior creates a unique copy of the configuration directory for each collection you create.

You can override the name given to the configuration directory in ZooKeeper by using the -n option. For instance, the command `bin/solr create -c logs -d _default -n basic` will upload the `server/solr/configsets/_default/conf` directory to ZooKeeper as /configs/basic.

Notice that we used the -d option to specify a different configuration than the default. Solr provides several built-in configurations under `server/solr/configsets`. However you can also provide the path to your own configuration directory using the -d option. For instance, the command `bin/solr create -c mycoll -d /tmp/myconfigs`, will upload /tmp/myconfigs into ZooKeeper under /configs/mycoll.

To reiterate, the configuration directory is named after the collection unless you override it using the -n option.

Other collections can share the same configuration by specifying the name of the shared configuration using the -n option. For instance, the following command will create a new collection that shares the basic configuration created previously: `bin/solr create -c logs2 -n basic`.

Data-driven Schema and Shared Configurations

The _default schema can mutate as data is indexed, since it has schemaless functionality (i.e., data-driven changes to the schema). Consequently, we recommend that you do not share data-driven configurations between collections unless you are certain that all collections should inherit the changes made when indexing data into one of the collections. You can turn off schemaless functionality (i.e., data-driven changes to the schema) for a collection by the following, assuming the collection name is mycollection - see [Set or Unset Configuration Properties](#):

```
bin/solr config -c mycollection -p 8983 -action set-user-property -property update.autoCreateFields -value false
```

Delete Core or Collection

The delete command detects the mode that Solr is running in (standalone or SolrCloud) and then deletes the specified core (standalone) or collection (SolrCloud) as appropriate.

```
bin/solr delete [options]
```

```
bin/solr delete -help
```

If running in SolrCloud mode, the delete command checks if the configuration directory used by the collection you are deleting is being used by other collections. If not, then the configuration directory is also deleted from ZooKeeper.

For example, if you created a collection with `bin/solr create -c contacts`, then the delete command `bin/solr delete -c contacts` will check to see if the `/configs/contacts` configuration directory is being used by any other collections. If not, then the `/configs/contacts` directory is removed from ZooKeeper.

Delete Core or Collection Parameters

`-c <name>`

Name of the core / collection to delete (required).

Example: `bin/solr delete -c mycoll`

`-deleteConfig`

Whether or not the configuration directory should also be deleted from ZooKeeper. The default is true.

If the configuration directory is being used by another collection, then it will not be deleted even if you pass `-deleteConfig` as true.

Example: `bin/solr delete -deleteConfig false`

`-p <port>`

The port of a local Solr instance to send the delete command to. By default the script tries to detect the port by looking for running Solr instances.

This option is useful if you are running multiple standalone Solr instances on the same host, thus requiring you to be specific about which instance to delete the core from.

Example: `bin/solr delete -p 8983`

Authentication

The `bin/solr` script allows enabling or disabling Basic Authentication, allowing you to configure authentication from the command line.

Currently, this script only enables Basic Authentication, and is only available when using SolrCloud mode.

Enabling Basic Authentication

The command `bin/solr auth enable` configures Solr to use Basic Authentication when accessing the User Interface, using `bin/solr` and any API requests.



For more information about Solr's authentication plugins, see the section [Securing Solr](#). For more information on Basic Authentication support specifically, see the section [Basic Authentication Plugin](#).

The `bin/solr auth enable` command makes several changes to enable Basic Authentication:

- Creates a `security.json` file and uploads it to ZooKeeper. The `security.json` file will look similar to:

```
{  
  "authentication":{  
    "blockUnknown": false,  
    "class":"solr.BasicAuthPlugin",  
    "credentials": {"user": "vgGVo69YJeUg/06AcFioWsdy0UdqfQv0LsrpIPMCzk=  
7iTnaKOWe+Uj5ZfGoKKK2G6hrcF10h6xezMqK+LBvpI="}  
  },  
  "authorization":{  
    "class":"solr.RuleBasedAuthorizationPlugin",  
    "permissions": [  
      {"name": "security-edit", "role": "admin"},  
      {"name": "collection-admin-edit", "role": "admin"},  
      {"name": "core-admin-edit", "role": "admin"}  
    ],  
    "user-role": {"user": "admin"}  
  }  
}
```

- Adds two lines to `bin/solr.in.sh` or `bin\solr.in.cmd` to set the authentication type, and the path to `basicAuth.conf`:

```
# The following lines added by ./solr for enabling BasicAuth  
SOLR_AUTH_TYPE="basic"  
SOLR_AUTHENTICATION_OPTS="-Dsolr.httpClient.config=/path/to/solr-  
8.1.0/server/solr/basicAuth.conf"
```

- Creates the file `server/solr/basicAuth.conf` to store the credential information that is used with `bin/solr` commands.

The command takes the following parameters:

`-credentials`

The username and password in the format of `username:password` of the initial user.

If you prefer not to pass the username and password as an argument to the script, you can choose the `-prompt` option. Either `-credentials` or `-prompt` **must** be specified.

-prompt

If prompt is preferred, pass **true** as a parameter to request the script to prompt the user to enter a username and password.

Either **-credentials** or **-prompt** **must** be specified.

-blockUnknown

When **true**, blocks all unauthenticated users from accessing Solr. This defaults to **false**, which means unauthenticated users will still be able to access Solr.

-updateIncludeFileOnly

When **true**, only the settings in bin/solr.in.sh or bin\solr.in.cmd will be updated, and security.json will not be created.

-z

Defines the ZooKeeper connect string. This is useful if you want to enable authentication before all your Solr nodes have come up. Unnecessary if ZK_HOST is defined in solr.in.sh or solr.in.cmd.

-d

Defines the Solr server directory, by default \$SOLR_HOME/server. It is not common to need to override the default, and is only needed if you have customized the \$SOLR_HOME directory path.

-s

Defines the location of solr.solr.home, which by default is server/solr. If you have multiple instances of Solr on the same host, or if you have customized the \$SOLR_HOME directory path, you likely need to define this.

Disabling Basic Authentication

You can disable Basic Authentication with bin/solr auth disable.

If the **-updateIncludeFileOnly** option is set to **true**, then only the settings in bin/solr.in.sh or bin\solr.in.cmd will be updated, and security.json will not be removed.

If the **-updateIncludeFileOnly** option is set to **false**, then the settings in bin/solr.in.sh or bin\solr.in.cmd will be updated, and security.json will be removed. However, the basicAuth.conf file is not removed with either option.

Set or Unset Configuration Properties

The bin/solr script enables a subset of the Config API: [\(un\)setting common properties](#) and [\(un\)setting user-defined properties](#).

```
bin/solr config [options]
```

```
bin/solr config -help
```

Set or Unset Common Properties

To set the common property updateHandler.autoCommit.maxDocs to 100 on collection mycollection:

```
bin/solr config -c mycollection -p 8983 -action set-property -property
```

```
updateHandler.autoCommit.maxDocs -value 100
```

The default `-action` is `set-property`, so the above can be shortened by not mentioning it:

```
bin/solr config -c mycollection -p 8983 -property updateHandler.autoCommit.maxDocs -value 100
```

To unset a previously set common property, specify `-action unset-property` with no `-value`:

```
bin/solr config -c mycollection -p 8983 -action unset-property -property updateHandler.autoCommit.maxDocs
```

Set or Unset User-defined Properties

To set the user-defined property `update.autoCreateFields` to `false` (to disable [Schemaless Mode](#)):

```
bin/solr config -c mycollection -p 8983 -action set-user-property -property update.autoCreateFields -value false
```

To unset a previously set user-defined property, specify `-action unset-user-property` with no `-value`:

```
bin/solr config -c mycollection -p 8983 -action unset-user-property -property update.autoCreateFields
```

Config Parameters

`-c <name>`

Name of the core or collection on which to change configuration (required).

`-action <name>`

Config API action, one of: `set-property`, `unset-property`, `set-user-property`, `unset-user-property`; defaults to `set-property`.

`-property <name>`

Name of the property to change (required).

`-value <new-value>`

Set the property to this value.

`-z <zkHost>`

The ZooKeeper connection string, usable in SolrCloud mode. Unnecessary if `ZK_HOST` is defined in `solr.in.sh` or `solr.in.cmd`.

`-p <port>`

localhost port of the Solr node to use when applying the configuration change.

`-solrUrl <url>`

Base Solr URL, which can be used in SolrCloud mode to determine the ZooKeeper connection string if that's not known.

ZooKeeper Operations

The `bin/solr` script allows certain operations affecting ZooKeeper. These operations are for SolrCloud mode

only. The operations are available as sub-commands, which each have their own set of options.

```
bin/solr zk [sub-command] [options]
```

```
bin/solr zk -help
```



Solr should have been started at least once before issuing these commands to initialize ZooKeeper with the znodes Solr expects. Once ZooKeeper is initialized, Solr doesn't need to be running on any node to use these commands.

Upload a Configuration Set

Use the `zk upconfig` command to upload one of the pre-configured configuration set or a customized configuration set to ZooKeeper.

ZK Upload Parameters

All parameters below are required.

`-n <name>`

Name of the configuration set in ZooKeeper. This command will upload the configuration set to the "configs" ZooKeeper node giving it the name specified.

You can see all uploaded configuration sets in the Admin UI via the Cloud screens. Choose Cloud -> Tree -> configs to see them.

If a pre-existing configuration set is specified, it will be overwritten in ZooKeeper.

Example: `-n myconfig`

`-d <configset dir>`

The path of the configuration set to upload. It should have a `conf` directory immediately below it that in turn contains `solrconfig.xml` etc.

If just a name is supplied, `$SOLR_HOME/server/solr/configsets` will be checked for this name. An absolute path may be supplied instead.

Examples:

- `-d directory_under_configsets`
- `-d /path/to/configset/source`

`-z <zkHost>`

The ZooKeeper connection string. Unnecessary if `ZK_HOST` is defined in `solr.in.sh` or `solr.in.cmd`.

Example: `-z 123.321.23.43:2181`

An example of this command with all of the parameters is:

```
bin/solr zk upconfig -z 111.222.333.444:2181 -n mynewconfig -d /path/to/configset
```

Reload Collections When Changing Configurations



This command does **not** automatically make changes effective! It simply uploads the configuration sets to ZooKeeper. You can use the Collection API's **RELOAD** command to reload any collections that uses this configuration set.

Download a Configuration Set

Use the `zk downconfig` command to download a configuration set from ZooKeeper to the local filesystem.

ZK Download Parameters

All parameters listed below are required.

`-n <name>`

Name of the configset in ZooKeeper to download. The Admin UI Cloud -> Tree -> configs node lists all available configuration sets.

Example: `-n myconfig`

`-d <configset dir>`

The path to write the downloaded configuration set into. If just a name is supplied, `$SOLR_HOME/server/solr/configsets` will be the parent. An absolute path may be supplied as well.

In either case, *pre-existing configurations at the destination will be overwritten!*

Examples:

- `-d directory_under_configsets`
- `-d /path/to/configset/destination`

`-z <zkHost>`

The ZooKeeper connection string. Unnecessary if `ZK_HOST` is defined in `solr.in.sh` or `solr.in.cmd`.

Example: `-z 123.321.23.43:2181`

An example of this command with all parameters is:

```
bin/solr zk downconfig -z 111.222.333.444:2181 -n mynewconfig -d /path/to/configset
```

A "best practice" is to keep your configuration sets in some form of version control as the system-of-record. In that scenario, `downconfig` should rarely be used.

Copy between Local Files and ZooKeeper znodes

Use the `zk cp` command for transferring files and directories between ZooKeeper znodes and your local drive. This command will copy from the local drive to ZooKeeper, from ZooKeeper to the local drive or from ZooKeeper to ZooKeeper.

ZK Copy Parameters

-r

Optional. Do a recursive copy. The command will fail if the <src> has children unless '-r' is specified.

Example: -r

<src>

The file or path to copy from. If prefixed with zk: then the source is presumed to be ZooKeeper. If no prefix or the prefix is 'file:' this is the local drive. At least one of <src> or <dest> must be prefixed by 'zk:' or the command will fail.

Examples:

- zk:/configs/myconfigs/solrconfig.xml
- file:/Users/apache/configs/src

<dest>

The file or path to copy to. If prefixed with zk: then the source is presumed to be ZooKeeper. If no prefix or the prefix is file: this is the local drive.

At least one of <src> or <dest> must be prefixed by zk: or the command will fail. If <dest> ends in a slash character it names a directory.

Examples:

- zk:/configs/myconfigs/solrconfig.xml
- file:/Users/apache/configs/src

-z <zkHost>

The ZooKeeper connection string. Unnecessary if ZK_HOST is defined in solr.in.sh or solr.in.cmd.

Example: -z 123.321.23.43:2181

An example of this command with the parameters is:

Recursively copy a directory from local to ZooKeeper.

```
bin/solr zk cp -r file:/apache/confgs/whatever/conf zk:/configs/myconf -z  
111.222.333.444:2181
```

Copy a single file from ZooKeeper to local.

```
bin/solr zk cp zk:/configs/myconf/managed_schema /configs/myconf/managed_schema -z  
111.222.333.444:2181
```

Remove a znode from ZooKeeper

Use the zk rm command to remove a znode (and optionally all child nodes) from ZooKeeper.

ZK Remove Parameters

-r

Optional. Do a recursive removal. The command will fail if the <path> has children unless '-r' is specified.

Example: -r

<path>

The path to remove from ZooKeeper, either a parent or leaf node.

There are limited safety checks, you cannot remove '/' or '/zookeeper' nodes.

The path is assumed to be a ZooKeeper node, no zk: prefix is necessary.

Examples:

- /configs
- /configs/myconfigset
- /configs/myconfigset/solrconfig.xml

-z <zkHost>

The ZooKeeper connection string. Unnecessary if ZK_HOST is defined in solr.in.sh or solr.in.cmd.

Example: -z 123.321.23.43:2181

Examples of this command with the parameters are:

bin/solr zk rm -r /configs

bin/solr zk rm /configs/myconfigset/schema.xml

Move One ZooKeeper znode to Another (Rename)

Use the zk mv command to move (rename) a ZooKeeper znode.

ZK Move Parameters

<src>

The znode to rename. The zk: prefix is assumed.

Example: /configs/oldconfigset

<dest>

The new name of the znode. The zk: prefix is assumed.

Example: /configs/newconfigset

-z <zkHost>

The ZooKeeper connection string. Unnecessary if ZK_HOST is defined in solr.in.sh or solr.in.cmd.

Example: -z 123.321.23.43:2181

An example of this command is:

bin/solr zk mv /configs/oldconfigset /configs/newconfigset

List a ZooKeeper znode's Children

Use the zk ls command to see the children of a znode.

ZK List Parameters

-r Optional. Recursively list all descendants of a znode.

+ Example: -r

<path>

The path on ZooKeeper to list.

Example: /collections/mycollection

-z <zkHost>

The ZooKeeper connection string. Unnecessary if ZK_HOST is defined in solr.in.sh or solr.in.cmd.

Example: -z 123.321.23.43:2181

An example of this command with the parameters is:

```
bin/solr zk ls -r /collections/mycollection
```

```
bin/solr zk ls /collections
```

Create a znode (supports chroot)

Use the zk mkroot command to create a znode. The primary use-case for this command to support ZooKeeper's "chroot" concept. However, it can also be used to create arbitrary paths.

Create znode Parameters

<path>

The path on ZooKeeper to create. Intermediate znodes will be created if necessary. A leading slash is assumed even if not specified.

Example: /solr

-z <zkHost>

The ZooKeeper connection string. Unnecessary if ZK_HOST is defined in solr.in.sh or solr.in.cmd.

Example: -z 123.321.23.43:2181

Examples of this command:

```
bin/solr zk mkroot /solr -z 123.321.23.43:2181
```

```
bin/solr zk mkroot /solr/production
```

Solr Configuration Files

Solr has several configuration files that you will interact with during your implementation.

Many of these files are in XML format, although APIs that interact with configuration settings tend to accept JSON for programmatic access as needed.

Solr Home

When Solr runs, it needs access to a home directory.

When you first install Solr, your home directory is `server/solr`. However, some examples may change this location (such as, if you run `bin/solr start -e cloud`, your home directory will be `example/cloud`).

The home directory contains important configuration information and is the place where Solr will store its index. The layout of the home directory will look a little different when you are running Solr in standalone mode vs. when you are running in SolrCloud mode.

The crucial parts of the Solr home directory are shown in these examples:

Standalone Mode

```
<solr-home-directory>/  
  solr.xml  
  core_name1/  
    core.properties  
    conf/  
      solrconfig.xml  
      managed-schema  
    data/  
  core_name2/  
    core.properties  
    conf/  
      solrconfig.xml  
      managed-schema  
    data/
```

SolrCloud Mode

```
<solr-home-directory>/  
  solr.xml  
  core_name1/  
    core.properties  
    data/  
  core_name2/  
    core.properties  
    data/
```

You may see other files, but the main ones you need to know are discussed in the next section.

Configuration Files

Inside Solr's Home, you'll find these files:

- `solr.xml` specifies configuration options for your Solr server instance. For more information on `solr.xml` see [Solr Cores and solr.xml](#).
- Per Solr Core:
 - `core.properties` defines specific properties for each core such as its name, the collection the core belongs to, the location of the schema, and other parameters. For more details on `core.properties`, see the section [Defining core.properties](#).
 - `solrconfig.xml` controls high-level behavior. You can, for example, specify an alternate location for the data directory. For more information on `solrconfig.xml`, see [Configuring solrconfig.xml](#).
 - `managed-schema` (or `schema.xml` instead) describes the documents you will ask Solr to index. The Schema define a document as a collection of fields. You get to define both the field types and the fields themselves. Field type definitions are powerful and include information about how Solr processes incoming field values and query values. For more information on Solr Schemas, see [Documents, Fields, and Schema Design](#) and the [Schema API](#).
 - `data/` The directory containing the low level index files.

Note that the SolrCloud example does not include a `conf` directory for each Solr Core (so there is no `solrconfig.xml` or Schema file). This is because the configuration files usually found in the `conf` directory are stored in ZooKeeper so they can be propagated across the cluster.

If you are using SolrCloud with the embedded ZooKeeper instance, you may also see `zoo.cfg` and `zoo.data` which are ZooKeeper configuration and data files. However, if you are running your own ZooKeeper ensemble, you would supply your own ZooKeeper configuration file when you start it and the copies in Solr would be unused. For more information about SolrCloud, see the section [SolrCloud](#).

Taking Solr to Production

This section provides guidance on how to setup Solr to run in production on *nix platforms, such as Ubuntu. Specifically, we'll walk through the process of setting up to run a single Solr instance on a Linux host and then provide tips on how to support multiple Solr nodes running on the same host.

Service Installation Script

Solr includes a service installation script (`bin/install_solr_service.sh`) to help you install Solr as a service on Linux. Currently, the script only supports CentOS, Debian, Red Hat, SUSE and Ubuntu Linux distributions. Before running the script, you need to determine a few parameters about your setup. Specifically, you need to decide where to install Solr and which system user should be the owner of the Solr files and process.

Planning Your Directory Structure

We recommend separating your live Solr files, such as logs and index files, from the files included in the Solr distribution bundle, as that makes it easier to upgrade Solr and is considered a good practice to follow as a system administrator.

Solr Installation Directory

By default, the service installation script will extract the distribution archive into `/opt`. You can change this location using the `-i` option when running the installation script. The script will also create a symbolic link to the versioned directory of Solr. For instance, if you run the installation script for Solr 8.1.0, then the following directory structure will be used:

```
/opt/solr-8.1.0  
/opt/solr -> /opt/solr-8.1.0
```

Using a symbolic link insulates any scripts from being dependent on the specific Solr version. If, down the road, you need to upgrade to a later version of Solr, you can just update the symbolic link to point to the upgraded version of Solr. We'll use `/opt/solr` to refer to the Solr installation directory in the remaining sections of this page.

Separate Directory for Writable Files

You should also separate writable Solr files into a different directory; by default, the installation script uses `/var/solr`, but you can override this location using the `-d` option. With this approach, the files in `/opt/solr` will remain untouched and all files that change while Solr is running will live under `/var/solr`.

Create the Solr User

Running Solr as root is not recommended for security reasons, and the [control script](#) start command will refuse to do so. Consequently, you should determine the username of a system user that will own all of the Solr files and the running Solr process. By default, the installation script will create the `solr` user, but you can override this setting using the `-u` option. If your organization has specific requirements for creating new user accounts, then you should create the user before running the script. The installation script will make the Solr user the owner of the `/opt/solr` and `/var/solr` directories.

You are now ready to run the installation script.

Run the Solr Installation Script

To run the script, you'll need to download the latest Solr distribution archive and then do the following:

```
tar xzf solr-8.1.0.tgz solr-8.1.0/bin/install_solr_service.sh --strip-components=2
```

The previous command extracts the `install_solr_service.sh` script from the archive into the current directory. If installing on Red Hat, please make sure `lsof` is installed before running the Solr installation script (`sudo yum install lsof`). The installation script must be run as root:

```
sudo bash ./install_solr_service.sh solr-8.1.0.tgz
```

By default, the script extracts the distribution archive into `/opt`, configures Solr to write files into `/var/solr`, and runs Solr as the `solr` user. Consequently, the following command produces the same result as the previous command:

```
sudo bash ./install_solr_service.sh solr-8.1.0.tgz -i /opt -d /var/solr -u solr -s solr -p 8983
```

You can customize the service name, installation directories, port, and owner using options passed to the installation script. To see available options, simply do:

```
sudo bash ./install_solr_service.sh -help
```

Once the script completes, Solr will be installed as a service and running in the background on your server (on port 8983). To verify, you can do:

```
sudo service solr status
```

If you do not want to start the service immediately, pass the `-n` option. You can then start the service manually later, e.g., after completing the configuration setup.

We'll cover some additional configuration settings you can make to fine-tune your Solr setup in a moment. Before moving on, let's take a closer look at the steps performed by the installation script. This gives you a better overview and will help you understand important details about your Solr installation when reading other pages in this guide; such as when a page refers to Solr home, you'll know exactly where that is on your system.

Solr Home Directory

The Solr home directory (not to be confused with the Solr installation directory) is where Solr manages core directories with index files. By default, the installation script uses `/var/solr/data`. If the `-d` option is used on the install script, then this will change to the data subdirectory in the location given to the `-d` option. Take a moment to inspect the contents of the Solr home directory on your system. If you do not [store solr.xml in ZooKeeper](#), the home directory must contain a `solr.xml` file. When Solr starts up, the Solr Control Script

passes the location of the home directory using the `-Dsolr.solr.home=...` system property.

Environment Overrides Include File

The service installation script creates an environment specific include file that overrides defaults used by the `bin/solr` script. The main advantage of using an include file is that it provides a single location where all of your environment-specific overrides are defined. Take a moment to inspect the contents of the `/etc/default/solr.in.sh` file, which is the default path setup by the installation script. If you used the `-s` option on the install script to change the name of the service, then the first part of the filename will be different. For a service named `solr-demo`, the file will be named `/etc/default/solr-demo.in.sh`. There are many settings that you can override using this file. However, at a minimum, this script needs to define the `SOLR_PID_DIR` and `SOLR_HOME` variables, such as:

```
SOLR_PID_DIR=/var/solr  
SOLR_HOME=/var/solr/data
```

The `SOLR_PID_DIR` variable sets the directory where the [control script](#) will write out a file containing the Solr server's process ID.

Log Settings

Solr uses Apache Log4J for logging. The installation script copies `/opt/solr/server/resources/log4j2.xml` to `/var/solr/log4j2.xml`. Take a moment to verify that the Solr include file is configured to send logs to the correct location by checking the following settings in `/etc/default/solr.in.sh`:

```
LOG4J_PROPS=/var/solr/log4j2.xml  
SOLR_LOGS_DIR=/var/solr/logs
```

For more information about Log4J configuration, please see: [Configuring Logging](#)

init.d Script

When running a service like Solr on Linux, it's common to setup an `init.d` script so that system administrators can control Solr using the service tool, such as: `service solr start`. The installation script creates a very basic `init.d` script to help you get started. Take a moment to inspect the `/etc/init.d/solr` file, which is the default script name setup by the installation script. If you used the `-s` option on the install script to change the name of the service, then the filename will be different. Notice that the following variables are setup for your environment based on the parameters passed to the installation script:

```
SOLR_INSTALL_DIR=/opt/solr  
SOLR_ENV=/etc/default/solr.in.sh  
RUNAS=solr
```

The `SOLR_INSTALL_DIR` and `SOLR_ENV` variables should be self-explanatory. The `RUNAS` variable sets the owner of the Solr process, such as `solr`; if you don't set this value, the script will run Solr as `root`, which is not recommended for production. You can use the `/etc/init.d/solr` script to start Solr by doing the following as root:

```
service solr start
```

The /etc/init.d/solr script also supports the **stop**, **restart**, and **status** commands. Please keep in mind that the init script that ships with Solr is very basic and is intended to show you how to setup Solr as a service. However, it's also common to use more advanced tools like **supervisord** or **upstart** to control Solr as a service on Linux. While showing how to integrate Solr with tools like supervisord is beyond the scope of this guide, the init.d/solr script should provide enough guidance to help you get started. Also, the installation script sets the Solr service to start automatically when the host machine initializes.

Progress Check

In the next section, we cover some additional environment settings to help you fine-tune your production setup. However, before we move on, let's review what we've achieved thus far. Specifically, you should be able to control Solr using /etc/init.d/solr. Please verify the following commands work with your setup:

```
sudo service solr restart  
sudo service solr status
```

The status command should give some basic information about the running Solr node that looks similar to:

```
Solr process PID running on port 8983  
{  
  "version": "5.0.0 - ubuntu - 2014-12-17 19:36:58",  
  "startTime": "2014-12-19T19:25:46.853Z",  
  "uptime": "0 days, 0 hours, 0 minutes, 8 seconds",  
  "memory": "85.4 MB (%17.4) of 490.7 MB"}
```

If the status command is not successful, look for error messages in /var/solr/logs/solr.log.

Fine-Tune Your Production Setup

Dynamic Defaults for ConcurrentMergeScheduler

The Merge Scheduler is configured in solrconfig.xml and defaults to ConcurrentMergeScheduler. This scheduler uses multiple threads to merge Lucene segments in the background.

By default, the ConcurrentMergeScheduler auto-detects whether the underlying disk drive is rotational or a SSD and sets defaults for maxThreadCount and maxMergeCount accordingly. If the disk drive is determined to be rotational then the maxThreadCount is set to 1 and maxMergeCount is set to 6. Otherwise, maxThreadCount is set to 4 or half the number of processors available to the JVM whichever is greater and maxMergeCount is set to maxThreadCount+5.

This auto-detection works only on Linux and even then it is not guaranteed to be correct. On all other platforms, the disk is assumed to be rotational. Therefore, if the auto-detection fails or is incorrect then indexing performance can suffer badly due to the wrong defaults.

The auto-detected value is exposed by the [Metrics API](#) with the key

`solr.node:CONTAINER.fs.coreRoot.spins`. A value of true denotes that the disk is detected to be a rotational or spinning disk.

It is safer to explicitly set values for `maxThreadCount` and `maxMergeCount` in the [IndexConfig section of SolrConfig.xml](#) so that values appropriate to your hardware are used.

Alternatively, the boolean system property `lucene.cms.override_spins` can be set in the `SOLR_OPTS` variable in the include file to override the auto-detected value. Similarly, the system property `lucene.cms.override_core_count` can be set to the number of CPU cores to override the auto-detected processor count.

Memory and GC Settings

By default, the `bin/solr` script sets the maximum Java heap size to 512M (`-Xmx512m`), which is fine for getting started with Solr. For production, you'll want to increase the maximum heap size based on the memory requirements of your search application; values between 10 and 20 gigabytes are not uncommon for production servers. When you need to change the memory settings for your Solr server, use the `SOLR_JAVA_MEM` variable in the include file, such as:

```
SOLR_JAVA_MEM="-Xms10g -Xmx10g"
```

Also, the [Solr Control Script](#) comes with a set of pre-configured Garbage First Garbage Collection settings that have shown to work well with Solr for a number of different workloads. However, these settings may not work well for your specific use of Solr. Consequently, you may need to change the GC settings, which should also be done with the `GC_TUNE` variable in the `/etc/default/solr.in.sh` include file. For more information about garbage collection settings refer to folowing articles: 1. <https://wiki.apache.org/solr/ShawnHeisey> 2. <https://www.oracle.com/technetwork/articles/java/g1gc-1984535.html> You can also refer to [JVM Settings](#) for tuning your memory and garbage collection settings.

Out-of-Memory Shutdown Hook

The `bin/solr` script registers the `bin/oom_solr.sh` script to be called by the JVM if an `OutOfMemoryError` occurs. The `oom_solr.sh` script will issue a `kill -9` to the Solr process that experiences the `OutOfMemoryError`. This behavior is recommended when running in SolrCloud mode so that ZooKeeper is immediately notified that a node has experienced a non-recoverable error. Take a moment to inspect the contents of the `/opt/solr/bin/oom_solr.sh` script so that you are familiar with the actions the script will perform if it is invoked by the JVM.

Going to Production with SolrCloud

To run Solr in SolrCloud mode, you need to set the `ZK_HOST` variable in the include file to point to your ZooKeeper ensemble. Running the embedded ZooKeeper is not supported in production environments. For instance, if you have a ZooKeeper ensemble hosted on the following three hosts on the default client port 2181 (zk1, zk2, and zk3), then you would set:

```
ZK_HOST=zk1,zk2,zk3
```

When the `ZK_HOST` variable is set, Solr will launch in "cloud" mode.

ZooKeeper chroot

If you're using a ZooKeeper instance that is shared by other systems, it's recommended to isolate the SolrCloud znode tree using ZooKeeper's chroot support. For instance, to ensure all znodes created by SolrCloud are stored under /solr, you can put /solr on the end of your ZK_HOST connection string, such as:

```
ZK_HOST=zk1,zk2,zk3/solr
```

Before using a chroot for the first time, you need to create the root path (znode) in ZooKeeper by using the [Solr Control Script](#). We can use the mkroot command for that:

```
bin/solr zk mkroot /solr -z <ZK_node>:<ZK_PORT>
```



If you also want to bootstrap ZooKeeper with existing solr_home, you can instead use the zkcli.sh / zkcli.bat bootstrap command, which will also create the chroot path if it does not exist. See [Command Line Utilities](#) for more info.

Solr Hostname

Use the SOLR_HOST variable in the include file to set the hostname of the Solr server.

```
SOLR_HOST=solr1.example.com
```

Setting the hostname of the Solr server is recommended, especially when running in SolrCloud mode, as this determines the address of the node when it registers with ZooKeeper.

Override Settings in solrconfig.xml

Solr allows configuration properties to be overridden using Java system properties passed at startup using the -Dproperty=value syntax. For instance, in solrconfig.xml, the default auto soft commit settings are set to:

```
<autoSoftCommit>
  <maxTime>${solr.autoSoftCommit.maxTime:-1}</maxTime>
</autoSoftCommit>
```

In general, whenever you see a property in a Solr configuration file that uses the \${solr.PROPERTY:DEFAULT_VALUE} syntax, then you know it can be overridden using a Java system property. For instance, to set the maxTime for soft-commits to be 10 seconds, then you can start Solr with -Dsolr.autoSoftCommit.maxTime=10000, such as:

```
bin/solr start -Dsolr.autoSoftCommit.maxTime=10000
```

The bin/solr script simply passes options starting with -D on to the JVM during startup. For running in production, we recommend setting these properties in the SOLR_OPTS variable defined in the include file.

Keeping with our soft-commit example, in `/etc/default/solr.in.sh`, you would do:

```
SOLR_OPTS="$SOLR_OPTS -Dsolr.autoSoftCommit.maxTime=10000"
```

File Handles and Processes (ulimit settings)

Two common settings that result in errors on *nix systems are file handles and user processes.

It is common for the default limits for number of processes and file handles to default to values that are too low for a large Solr installation. The required number of each of these will increase based on a combination of the number of replicas hosted per node and the number of segments in the index for each replica.

The usual recommendation is to make processes and file handles at least 65,000 each, unlimited if possible. On most *nix systems, this command will show the currently-defined limits:

```
ulimit -a
```

It is strongly recommended that file handle and process limits be permanently raised as above. The exact form of the command will vary per operating system, and some systems require editing configuration files and restarting your server. Consult your system administrators for guidance in your particular environment.

If these limits are exceeded, the problems reported by Solr vary depending on the specific operation responsible for exceeding the limit. Errors such as "too many open files", "connection error", and "max processes exceeded" have been reported, as well as SolrCloud recovery failures.



Since exceeding these limits can result in such varied symptoms it is *strongly* recommended that these limits be permanently raised as recommended above.

Running Multiple Solr Nodes per Host

The `bin/solr` script is capable of running multiple instances on one machine, but for a **typical** installation, this is not a recommended setup. Extra CPU and memory resources are required for each additional instance. A single instance is easily capable of handling multiple indexes.

When to ignore the recommendation

For every recommendation, there are exceptions. For the recommendation above, that exception is mostly applicable when discussing extreme scalability. The best reason for running multiple Solr nodes on one host is decreasing the need for extremely large heaps.

When the Java heap gets very large, it can result in extremely long garbage collection pauses, even with the GC tuning that the startup script provides by default. The exact point at which the heap is considered "very large" will vary depending on how Solr is used. This means that there is no hard number that can be given as a threshold, but if your heap is reaching the neighborhood of 16 to 32 gigabytes, it might be time to consider splitting nodes. Ideally this would mean more machines, but budget constraints might make that impossible.

There is another issue once the heap reaches 32GB. Below 32GB, Java is able to use compressed pointers, but above that point, larger pointers are required, which uses more memory and slows down the JVM.

Because of the potential garbage collection issues and the particular issues that happen at 32GB, if a single instance would require a 64GB heap, performance is likely to improve greatly if the machine is set up with two nodes that each have a 31GB heap.

If your use case requires multiple instances, at a minimum you will need unique Solr home directories for each node you want to run; ideally, each home should be on a different physical disk so that multiple Solr nodes don't have to compete with each other when accessing files on disk. Having different Solr home directories implies that you'll need a different include file for each node. Moreover, if using the /etc/init.d/solr script to control Solr as a service, then you'll need a separate script for each node. The easiest approach is to use the service installation script to add multiple services on the same host, such as:

```
sudo bash ./install_solr_service.sh solr-8.1.0.tgz -s solr2 -p 8984
```

The command shown above will add a service named solr2 running on port 8984 using /var/solr2 for writable (aka "live") files; the second server will still be owned and run by the solr user and will use the Solr distribution files in /opt. After installing the solr2 service, verify it works correctly by doing:

```
sudo service solr2 restart  
sudo service solr2 status
```

Making and Restoring Backups

If you are worried about data loss, and of course you *should* be, you need a way to back up your Solr indexes so that you can recover quickly in case of catastrophic failure.

Solr provides two approaches to backing up and restoring Solr cores or collections, depending on how you are running Solr. If you run in SolrCloud mode, you will use the Collections API. If you run Solr in standalone mode, you will use the replication handler.

SolrCloud Backups

Support for backups when running SolrCloud is provided with the [Collections API](#). This allows the backups to be generated across multiple shards, and restored to the same number of shards and replicas as the original collection.



SolrCloud Backup/Restore requires a shared file system mounted at the same path on all nodes, or HDFS.

Two commands are available:

- `action=BACKUP`: This command backs up Solr indexes and configurations. More information is available in the section [Backup Collection](#).
- `action=RESTORE`: This command restores Solr indexes and configurations. More information is available in the section [Restore Collection](#).

Standalone Mode Backups

Backups and restoration uses Solr's replication handler. Out of the box, Solr includes implicit support for replication so this API can be used. Configuration of the replication handler can, however, be customized by defining your own replication handler in `solrconfig.xml`. For details on configuring the replication handler, see the section [Configuring the ReplicationHandler](#).

Backup API

The backup API requires sending a command to the `/replication` handler to back up the system.

You can trigger a back-up with an HTTP command like this (replace "gettingstarted" with the name of the core you are working with):

Backup API Example

```
http://localhost:8983/solr/gettingstarted/replication?command=backup
```

The backup command is an asynchronous call, and it will represent data from the latest index commit point. All indexing and search operations will continue to be executed against the index as usual.

Only one backup call can be made against a core at any point in time. While an ongoing backup operation is happening subsequent calls for restoring will throw an exception.

The backup request can also take the following additional parameters:

location

The path where the backup will be created. If the path is not absolute then the backup path will be relative to Solr's instance directory. |name| The snapshot will be created in a directory called `snapshot.<name>`. If a name is not specified then the directory name would have the following format: `snapshot.<yyyyMMddHHmmssSSS>`.

numberToKeep

The number of backups to keep. If `maxNumberOfBackups` has been specified on the replication handler in `solrconfig.xml`, `maxNumberOfBackups` is always used and attempts to use `numberToKeep` will cause an error. Also, this parameter is not taken into consideration if the backup name is specified. More information about `maxNumberOfBackups` can be found in the section [Configuring the ReplicationHandler](#).

repository

The name of the repository to be used for the backup. If no repository is specified then the local filesystem repository will be used automatically.

commitName

The name of the commit which was used while taking a snapshot using the `CREATESNAPSHOT` command.

Backup Status

The backup operation can be monitored to see if it has completed by sending the `details` command to the `/replication` handler, as in this example:

Status API Example

```
http://localhost:8983/solr/gettingstarted/replication?command=details&wt=xml
```

Output Snippet

```
<lst name="backup">
  <str name="startTime">Sun Apr 12 16:22:50 DAVT 2015</str>
  <int name="fileCount">10</int>
  <str name="status">success</str>
  <str name="snapshotCompletedAt">Sun Apr 12 16:22:50 DAVT 2015</str>
  <str name="snapshotName">my_backup</str>
</lst>
```

If it failed then a `snapShootException` will be sent in the response.

Restore API

Restoring the backup requires sending the `restore` command to the `/replication` handler, followed by the name of the backup to restore.

You can restore from a backup with a command like this:

Example Usage

```
http://localhost:8983/solr/gettingstarted/replication?command=restore&name=backup_name
```

This will restore the named index snapshot into the current core. Searches will start reflecting the snapshot data once the restore is complete.

The restore request can take these additional parameters:

location

The location of the backup snapshot file. If not specified, it looks for backups in Solr's data directory.

name

The name of the backed up index snapshot to be restored. If the name is not provided it looks for backups with snapshot.<timestamp> format in the location directory. It picks the latest timestamp backup in that case.

repository

The name of the repository to be used for the backup. If no repository is specified then the local filesystem repository will be used automatically.

The restore command is an asynchronous call. Once the restore is complete the data reflected will be of the backed up index which was restored.

Only one restore call can be made against a core at one point in time. While an ongoing restore operation is happening subsequent calls for restoring will throw an exception.

Restore Status API

You can also check the status of a restore operation by sending the restorestatus command to the /replication handler, as in this example:

Status API Example

```
http://localhost:8983/solr/gettingstarted/replication?command=restorestatus&wt=xml
```

Status API Output

```
<response>
  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">0</int>
  </lst>
  <lst name="restorestatus">
    <str name="snapshotName">snapshot.<name></str>
    <str name="status">success</str>
  </lst>
</response>
```

The status value can be "In Progress", "success" or "failed". If it failed then an "exception" will also be sent

in the response.

Create Snapshot API

The snapshot functionality is different from the backup functionality as the index files aren't copied anywhere. The index files are snapshotted in the same index directory and can be referenced while taking backups.

You can trigger a snapshot command with an HTTP command like this (replace "techproducts" with the name of the core you are working with):

Create Snapshot API Example

```
http://localhost:8983/solr/admin/cores?action=CREATESNAPSHOT&core=techproducts&commitName=commit1
```

The CREATESNAPSHOT request parameters are:

commitName

The name to store the snapshot as.

core

The name of the core to perform the snapshot on.

async

Request ID to track this action which will be processed asynchronously.

List Snapshot API

The LISTSNAPSHOTS command lists all the taken snapshots for a particular core.

You can trigger a list snapshot command with an HTTP command like this (replace "techproducts" with the name of the core you are working with):

List Snapshot API

```
http://localhost:8983/solr/admin/cores?action=LISTSNAPSHOTS&core=techproducts&commitName=commit1
```

The list snapshot request parameters are:

core

The name of the core to whose snapshots we want to list.

async

Request ID to track this action which will be processed asynchronously.

Delete Snapshot API

The DELETESNAPSHOT command deletes a snapshot for a particular core.

You can trigger a delete snapshot with an HTTP command like this (replace "techproducts" with the name of the core you are working with):

Delete Snapshot API Example

```
http://localhost:8983/solr/admin/cores?action=DELETESNAPSHOT&core=techproducts&commitName=commit1
```

The delete snapshot request parameters are:

commitName

Specify the commit name to be deleted.

core

The name of the core whose snapshot we want to delete.

async

Request ID to track this action which will be processed asynchronously.

Backup/Restore Storage Repositories

Solr provides interfaces to plug different storage systems for backing up and restoring. For example, you can have a Solr cluster running on a local filesystem like EXT3 but you can backup the indexes to a HDFS filesystem or vice versa.

The repository interfaces needs to be configured in the solr.xml file. While running backup/restore commands we can specify the repository to be used.

If no repository is configured then the local filesystem repository will be used automatically.

Example solr.xml section to configure a repository like [HDFS](#):

```
<backup>
  <repository name="hdfs" class="org.apache.solr.core.backup.repository.HdfsBackupRepository"
  default="false">
    <str name="location">${solr.hdfs.default.backup.path}</str>
    <str name="solr.hdfs.home">${solr.hdfs.home:}</str>
    <str name="solr.hdfs.confdir">${solr.hdfs.confdir:}</str>
  </repository>
</backup>
```

Better throughput might be achieved by increasing buffer size with `<int name="solr.hdfs.buffer.size">262144</int>`. Buffer size is specified in bytes, by default it's 4096 bytes (4KB).

Running Solr on HDFS

Solr has support for writing and reading its index and transaction log files to the HDFS distributed filesystem.

This does not use Hadoop MapReduce to process Solr data, rather it only uses the HDFS filesystem for index and transaction log file storage.

To use HDFS rather than a local filesystem, you must be using Hadoop 2.x and you will need to instruct Solr to use the `HdfsDirectoryFactory`. There are also several additional parameters to define. These can be set in one of three ways:

- Pass JVM arguments to the `bin/solr` script. These would need to be passed every time you start Solr with `bin/solr`.
- Modify `solr.in.sh` (or `solr.in.cmd` on Windows) to pass the JVM arguments automatically when using `bin/solr` without having to set them manually.
- Define the properties in `solrconfig.xml`. These configuration changes would need to be repeated for every collection, so is a good option if you only want some of your collections stored in HDFS.

Starting Solr on HDFS

Standalone Solr Instances

For standalone Solr instances, there are a few parameters you should modify before starting Solr. These can be set in `solrconfig.xml` (more on that [below](#)), or passed to the `bin/solr` script at startup.

- You need to use an `HdfsDirectoryFactory` and a data directory in the form `hdfs://host:port/path`
- You need to specify an `updateLog` location in the form `hdfs://host:port/path`
- You should specify a lock factory type of '`hdfs`' or `none`.

If you do not modify `solrconfig.xml`, you can instead start Solr on HDFS with the following command:

```
bin/solr start -Dsolr.directoryFactory=HdfsDirectoryFactory  
  -Dsolr.lock.type=hdfs  
  -Dsolr.data.dir=hdfs://host:port/path  
  -Dsolr.updatelog=hdfs://host:port/path
```

This example will start Solr in standalone mode, using the defined JVM properties (explained in more detail [below](#)).

SolrCloud Instances

In SolrCloud mode, it's best to leave the data and update log directories as the defaults Solr comes with and simply specify the `solr.hdfs.home`. All dynamically created collections will create the appropriate directories automatically under the `solr.hdfs.home` root directory.

- Set `solr.hdfs.home` in the form `hdfs://host:port/path`

- You should specify a lock factory type of 'hdfs' or none.

```
bin/solr start -c -Dsolr.directoryFactory=HdfsDirectoryFactory  
-Dsolr.lock.type=hdfs  
-Dsolr.hdfs.home=hdfs://host:port/path
```

This command starts Solr in SolrCloud mode, using the defined JVM properties.

Modifying solr.in.sh (*nix) or solr.in.cmd (Windows)

The examples above assume you will pass JVM arguments as part of the start command every time you use bin/solr to start Solr. However, bin/solr looks for an include file named solr.in.sh (solr.in.cmd on Windows) to set environment variables. By default, this file is found in the bin directory, and you can modify it to permanently add the HdfsDirectoryFactory settings and ensure they are used every time Solr is started.

For example, to set JVM arguments to always use HDFS when running in SolrCloud mode (as shown above), you would add a section such as this:

```
# Set HDFS DirectoryFactory & Settings  
-Dsolr.directoryFactory=HdfsDirectoryFactory \  
-Dsolr.lock.type=hdfs \  
-Dsolr.hdfs.home=hdfs://host:port/path \  
 
```

The Block Cache

For performance, the HdfsDirectoryFactory uses a Directory that will cache HDFS blocks. This caching mechanism replaces the standard file system cache that Solr utilizes. By default, this cache is allocated off-heap. This cache will often need to be quite large and you may need to raise the off-heap memory limit for the specific JVM you are running Solr in. For the Oracle/OpenJDK JVMs, the following is an example command-line parameter that you can use to raise the limit when starting Solr:

```
-XX:MaxDirectMemorySize=20g
```

HdfsDirectoryFactory Parameters

The HdfsDirectoryFactory has a number of settings defined as part of the directoryFactory configuration.

Solr HDFS Settings

`solr.hdfs.home`

A root location in HDFS for Solr to write collection data to. Rather than specifying an HDFS location for the data directory or update log directory, use this to specify one root location and have everything automatically created within this HDFS location. The structure of this parameter is `hdfs://host:port/path/solr`.

Block Cache Settings

`solr.hdfs.blockcache.enabled`

Enable the blockcache. The default is true.

`solr.hdfs.blockcache.read.enabled`

Enable the read cache. The default is true.

`solr.hdfs.blockcache.direct.memory.allocation`

Enable direct memory allocation. If this is false, heap is used. The default is true.

`solr.hdfs.blockcache.slab.count`

Number of memory slabs to allocate. Each slab is 128 MB in size. The default is 1.

`solr.hdfs.blockcache.global`

Enable/Disable using one global cache for all SolrCores. The settings used will be from the first HdfsDirectoryFactory created. The default is true.

NRTCachingDirectory Settings

`solr.hdfs.nrtcachingdirectory.enable`

true | Enable the use of NRTCachingDirectory. The default is true.

`solr.hdfs.nrtcachingdirectory.maxmergesizemb`

NRTCachingDirectory max segment size for merges. The default is 16.

`solr.hdfs.nrtcachingdirectory.maxcachedmb`

NRTCachingDirectory max cache size. The default is 192.

HDFS Client Configuration Settings

`solr.hdfs.confdir`

Pass the location of HDFS client configuration files - needed for HDFS HA for example.

Kerberos Authentication Settings

Hadoop can be configured to use the Kerberos protocol to verify user identity when trying to access core services like HDFS. If your HDFS directories are protected using Kerberos, then you need to configure Solr's HdfsDirectoryFactory to authenticate using Kerberos in order to read and write to HDFS. To enable Kerberos authentication from Solr, you need to set the following parameters:

`solr.hdfs.security.kerberos.enabled`

Set to true to enable Kerberos authentication. The default is false.

`solr.hdfs.security.kerberos.keytabfile`

A keytab file contains pairs of Kerberos principals and encrypted keys which allows for password-less authentication when Solr attempts to authenticate with secure Hadoop.

This file will need to be present on all Solr servers at the same path provided in this parameter.

`solr.hdfs.security.kerberos.principal`

The Kerberos principal that Solr should use to authenticate to secure Hadoop; the format of a typical Kerberos V5 principal is: primary/instance@realm.

Example solrconfig.xml for HDFS

Here is a sample solrconfig.xml configuration for storing Solr indexes on HDFS:

```
<directoryFactory name="DirectoryFactory" class="solr.HdfsDirectoryFactory">
  <str name="solr.hdfs.home">hdfs://host:port/solr</str>
  <bool name="solr.hdfs.blockcache.enabled">true</bool>
  <int name="solr.hdfs.blockcache.slab.count">1</int>
  <bool name="solr.hdfs.blockcache.direct.memory.allocation">true</bool>
  <int name="solr.hdfs.blockcache.blocksperbank">16384</int>
  <bool name="solr.hdfs.blockcache.read.enabled">true</bool>
  <bool name="solr.hdfs.nrtcachingdirectory.enable">true</bool>
  <int name="solr.hdfs.nrtcachingdirectory.maxmergesizemb">16</int>
  <int name="solr.hdfs.nrtcachingdirectory.maxcachedmb">192</int>
</directoryFactory>
```

If using Kerberos, you will need to add the three Kerberos related properties to the <directoryFactory> element in solrconfig.xml, such as:

```
<directoryFactory name="DirectoryFactory" class="solr.HdfsDirectoryFactory">
  ...
  <bool name="solr.hdfs.security.kerberos.enabled">true</bool>
  <str name="solr.hdfs.security.kerberos.keytabfile">/etc/krb5.keytab</str>
  <str name="solr.hdfs.security.kerberos.principal">solr/admin@KERBEROS.COM</str>
</directoryFactory>
```

Automatically Add Replicas in SolrCloud

The ability to automatically add new replicas when the Overseer notices that a shard has gone down was previously only available to users running Solr in HDFS, but it is now available to all users via Solr's autoscaling framework. See the section [SolrCloud Autoscaling Automatically Adding Replicas](#) for details on how to enable and disable this feature.

The ability to enable or disable the autoAddReplicas feature with cluster properties has been deprecated and will be removed in a future version. All users of this feature who have previously used that approach are encouraged to change their configurations to use the autoscaling framework to ensure continued operation of this feature in their Solr installations.

For users using this feature with the deprecated configuration, you can temporarily disable it cluster-wide by setting the cluster property autoAddReplicas to false, as in these examples:

V1 API

```
http://localhost:8983/solr/admin/collections?action=CLUSTERPROP&name=autoAddReplicas&val=false
```

V2 API



```
curl -X POST -H 'Content-type: application/json' -d '{"set-property": {"name": "autoAddReplicas", "val":false}}' http://localhost:8983/api/cluster
```

Re-enable the feature by unsetting the autoAddReplicas cluster property. When no val parameter is provided, the cluster property is unset:

V1 API

```
http://localhost:8983/solr/admin/collections?action=CLUSTERPROP&name=autoAddReplicas
```

V2 API

```
curl -X POST -H 'Content-type: application/json' -d '{"set-property": {"name": "autoAddReplicas"}}' http://localhost:8983/api/cluster
```

SolrCloud on AWS EC2

This guide is a tutorial on how to set up a multi-node SolrCloud cluster on [Amazon Web Services \(AWS\) EC2](#) instances for early development and design.

This tutorial is not meant for production systems. For one, it uses Solr's embedded ZooKeeper instance, and for production you should have at least 3 ZooKeeper nodes in an ensemble. There are additional steps you should take for a production installation; refer to [Taking Solr to Production](#) for how to deploy Solr in production.

In this guide we are going to:

1. Launch multiple AWS EC2 instances
 - Create new *Security Group*
 - Configure instances and launch
2. Install, configure and start Solr on newly launched EC2 instances
 - Install system prerequisites: Java 1.8 and later
 - Download latest version of Solr
 - Start the Solr nodes in SolrCloud mode
3. Create a collection, index documents and query the system
 - Create collection with multiple shards and replicas
 - Index documents to the newly created collection
 - Verify documents presence by querying the collection

Before You Start

To use this guide, you must have the following:

- An [AWS](#) account.
- Familiarity with setting up a single-node SolrCloud on local machine. Refer to the [Solr Tutorial](#) if you have never used Solr before.

Launch EC2 instances

Create new Security Group

1. Navigate to the [AWS EC2 console](#) and to the region of your choice.
2. Configure an [AWS security group](#) which will limit access to the installation and allow our launched EC2 instances to talk to each other without restrictions.
 - a. From the EC2 Dashboard, click [**Security Groups**] from the left-hand menu, under "Network & Security".
 - b. Click [**Create Security Group**] under the *Security Groups* section. Give your security group a descriptive name.

- c. You can select one of the existing [VPCs](#) or create a new one.
- d. We need two ports open for our cloud here:
 - i. Solr port. In this example we will use Solr's default port 8983.
 - ii. ZooKeeper Port: We'll use Solr's embedded ZooKeeper, so we'll use the default port 9983 (see the [Deploying with External ZooKeeper](#) to configure external ZooKeeper).
- e. Click [**Inbound**] to set inbound network rules, then select [**Add Rule**]. Select "Custom TCP" as the type. Enter 8983 for the "Port Range" and choose "My IP for the Source, then enter your public IP. Create a second rule with the same type and source, but enter 9983 for the port.

This will limit access to your current machine. If you want wider access to the instance in order to collaborate with others, you can specify that, but make sure you only allow as much access as needed. A Solr instance should not be exposed to general Internet traffic.

- f. Add another rule for SSH access. Choose "SSH" as the type, and again "My IP" for the source and again enter your public IP. You need SSH access on all instances to install and configure Solr.
- g. Review the details, your group configuration should look like this:

Type	Protocol	Port Range	Source	Description
SSH	TCP	22	My IP	103.228.221.249/32 e.g. SSH for Admin De
Custom TCP	TCP	8983	My IP	103.228.221.249/32 e.g. SSH for Admin De
Custom TCP	TCP	9983	My IP	103.228.221.249/32 e.g. SSH for Admin De

- h. Click [**Create**] when finished.
- i. We need to modify the rules so that instances that are part of the group can talk to all other instances that are part of the same group. We could not do this while creating the group, so we need to edit the group after creating it to add this.
 - i. Select the newly created group in the Security Group overview table. Under the "Inbound" tab, click [**Edit**].
 - ii. Click [**Add rule**]. Choose All TCP from the pulldown list for the type, and enter 0-65535 for the port range. Specify the name of the current Security Group as the solr-sample.
- j. Review the details, your group configuration should now look like this:

Edit inbound rules

Type	Protocol	Port Range	Source	Description
Custom TCP	TCP	9983	Custom	103.228.221.249/32 e.g. SSH for Admin Desktop
SSH	TCP	22	Custom	103.228.221.249/32 e.g. SSH for Admin Desktop
Custom TCP	TCP	8983	Custom	103.228.221.249/32 e.g. SSH for Admin Desktop
All TCP	TCP	0 - 65535	Custom	sg-acf285d1 e.g. SSH for Admin Desktop

Add Rule

NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.

Cancel **Save**

k. Click [Save] when finished.

Configure Instances and Launch

Once the security group is in place, you can choose [Instances] from the left-hand navigation menu.

Under Instances, click [Launch Instance] button and follow the wizard steps:

1. Choose your Amazon Machine Image (AMI): Choose **Amazon Linux AMI, SSD Volume Type** as the AMI. There are both commercial AMIs and Community based AMIs available, e.g., Amazon Linux AMI (HVM), SSD Volume Type, but this is a nice AMI to use for our purposes. Click [Select] next to the image you choose.
2. The next screen asks you to choose the instance type, **t2.medium** is sufficient. Choose it from the list, then click [Configure Instance Details].
3. Configure the instance. Enter **2** in the "Number of instances" field. Make sure the setting for "Auto-assign Public IP" is "Enabled".
4. When finished, click [Add Storage]. The default of **8 GB** for size and **General Purpose SSD** for the volume type is sufficient for running this quick start. Optionally select "Delete on termination" if you know you won't need the data stored in Solr indexes after you terminate the instances.
5. When finished, click [Add Tags]. You do not have to add any tags for this quick start, but you can add them if you want.
6. Click [Configure Security Group]. Choose **Select an existing security group** and select the security group you created earlier: solr-sample. You should see the expected inbound rules at the bottom of the page.
7. Click [Review].
8. If everything looks correct, click [Launch].
9. Select an existing "private key file" or create a new one and download to your local machine so you will be able to login into the instances via SSH.

Select an existing key pair or create a new key pair X

A key pair consists of a **public key** that AWS stores, and a **private key file** that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about [removing existing key pairs from a public AMI](#).

Create a new key pair ▼

Key pair name | Download Key Pair

Tip: You have to download the **private key file (*.pem file)** before you can continue. **Store it in a secure and accessible location.** You will not be able to download the file again after it's created.

Cancel Launch Instances

10. On the instances list, you can watch the states change. You cannot use the instances until they become “running”.

Install, Configure and Start

1. Locate the Public DNS record for the instance by selecting the instance from the list of instances, and log on to each machine one by one.

Using SSH, if your AWS identity key file is aws-key.pem and the AMI uses ec2-user as login user, on each AWS instance, do the following:

```
$ ssh-add aws-key.pem  
$ ssh -A ec2-user@<instance-public-dns>
```

2. While logged in to each of the AWS EC2 instances, configure Java 1.8 and download Solr:

```
# verify default java version packaged with AWS instances is 1.7  
$ java -version  
$ sudo yum install java-1.8.0  
$ sudo /usr/sbin/alternatives --config java  
# select jdk-1.8  
# verify default java version to java-1.8  
$ java -version
```

```
# download desired version of Solr
$ wget http://archive.apache.org/dist/lucene/solr/8.1.0/solr-8.1.0.tgz
# untar
$ tar -zxf solr-8.1.0.tgz
# set SOLR_HOME
$ export SOLR_HOME=$PWD/solr-8.1.0
# put the env variable in .bashrc
# vim ~/.bashrc
export SOLR_HOME=/home/ec2-user/solr-8.1.0
```

3. Resolve the Public DNS to simpler hostnames.

Let's assume AWS instances public DNS with IPv4 Public IP are as follows:

- ec2-54-1-2-3.us-east-2.compute.amazonaws.com: 54.1.2.3
- ec2-54-4-5-6.us-east-2.compute.amazonaws.com: 54.4.5.6

Edit /etc/hosts, and add entries for the above machines:

```
$ sudo vim /etc/hosts
54.1.2.3 solr-node-1
54.4.5.6 solr-node-2
```

4. Configure Solr in running EC2 instances.

In this case, one of the machines will host ZooKeeper embedded along with Solr node, say, ec2-101-1-2-3.us-east-2.compute.amazonaws.com (aka, solr-node-1)

See [Deploying with External ZooKeeper](#) for configure external ZooKeeper.

Inside the ec2-101-1-2-3.us-east-2.compute.amazonaws.com (solr-node-1)

```
$ cd $SOLR_HOME
# start Solr node on 8983 and ZooKeeper will start on 8983+1000 9983
$ bin/solr start -c -p 8983 -h solr-node-1
```

On the other node, ec2-101-4-5-6.us-east-2.compute.amazonaws.com (solr-node-2)

```
$ cd $SOLR_HOME
# start Solr node on 8983 and connect to ZooKeeper running on first node
$ bin/solr start -c -p 8983 -h solr-node-2 -z solr-node-1:9983
```

5. Inspect and Verify. Inspect the Solr nodes state from browser on local machine:

Go to:

```
http://ec2-101-1-2-3.us-east-2.compute.amazonaws.com:8983/solr (solr-node-1:8983/solr)
```

```
http://ec2-101-4-5-6.us-east-2.compute.amazonaws.com:8983/solr (solr-node-2:8983/solr)
```

You should able to see Solr UI dashboard for both nodes.

Create Collection, Index and Query

You can refer [Solr Tutorial](#) for an extensive walkthrough on creating collections with multiple shards and replicas, indexing data via different methods and querying documents accordingly.

Deploying with External ZooKeeper

If you want to configure an external ZooKeeper ensemble to avoid using the embedded single-instance ZooKeeper that runs in the same JVM as the Solr node, you need to make few tweaks in the above listed steps as follows.

- When creating the security group, instead of opening port 9983 for ZooKeeper, you'll open 2181 (or whatever port you are using for ZooKeeper: it's default is 2181).
- When configuring the number of instances to launch, choose to open 3 instances instead of 2.
- When modifying the /etc/hosts on each machine, add a third line for the 3rd instance and give it a recognizable name:

```
$ sudo vim /etc/hosts
54.1.2.3 solr-node-1
54.4.5.6 solr-node-2
54.7.8.9 zookeeper-node
```

- You'll need to install ZooKeeper manually, described in the next section.

Install ZooKeeper

These steps will help you install and configure a single instance of ZooKeeper on AWS. This is not sufficient for a production, use, however, where a ZooKeeper ensemble of at least three nodes is recommended. See the section [Setting Up an External ZooKeeper Ensemble](#) for information about how to change this single-instance into an ensemble.

1. Download a stable version of ZooKeeper. In this example we're using ZooKeeper v3.4.14. On the node you're using to host ZooKeeper (zookeeper-node), download the package and untar it:

```
# download stable version of ZooKeeper, here 3.4.14
$ wget https://archive.apache.org/dist/zookeeper/zookeeper-3.4.14/zookeeper-3.4.14.tar.gz
# untar
$ tar -zxvf zookeeper-3.4.14.tar.gz
```

Add an environment variable for ZooKeeper's home directory (ZOO_HOME) to the .bashrc for the user

who will be running the process. The rest of the instructions assume you have set this variable. Correct the path to the ZooKeeper installation as appropriate if where you put it does not match the below.

```
$ export ZOO_HOME=$PWD/zookeeper-3.4.14  
# put the env variable in .bashrc  
# vim ~/.bashrc  
export ZOO_HOME=/home/ec2-user/zookeeper-3.4.14
```

2. Change directories to ZOO_HOME, and create the ZooKeeper configuration by using the template provided by ZooKeeper.

```
$ cd $ZOO_HOME  
# create ZooKeeper config by using zoo_sample.cfg  
$ cp conf/zoo_sample.cfg conf/zoo.cfg
```

3. Create the ZooKeeper data directory in the filesystem, and edit the zoo.cfg file to uncomment the autopurge parameters and define the location of the data directory.

```
# create data dir for ZooKeeper, edit zoo.cfg, uncomment autopurge parameters  
$ mkdir data  
$ vim conf/zoo.cfg  
# -- uncomment --  
autopurge.snapRetainCount=3  
autopurge.purgeInterval=1  
# -- edit --  
dataDir=data
```

4. Start ZooKeeper.

```
$ cd $ZOO_HOME  
# start ZooKeeper, default port: 2181  
$ bin/zkServer.sh start
```

5. On the first node being used for Solr (solr-node-1), start Solr and tell it where to find ZooKeeper.

```
$ cd $SOLR_HOME  
# start Solr node on 8983 and connect to ZooKeeper running on ZooKeeper node  
$ bin/solr start -c -p 8983 -h solr-node-1 -z zookeeper-node:2181
```

6. On the second Solr node (solr-node-2), again start Solr and tell it where to find ZooKeeper.

```
$ cd $SOLR_HOME  
# start Solr node on 8983 and connect to ZooKeeper running on ZooKeeper node  
$ bin/solr start -c -p 8983 -h solr-node-1 -z zookeeper-node:2181
```

As noted earlier, a single ZooKeeper node is not sufficient for a production installation. See these additional resources for more information about deploying Solr in production, which can be used once you have the EC2 instances up and running:



- [Taking Solr to Production](#)
- [Setting Up an External ZooKeeper Ensemble](#)

Upgrading a Solr Cluster

This page covers how to upgrade an existing Solr cluster that was installed using the [service installation scripts](#).



The steps outlined on this page assume you use the default service name of solr. If you use an alternate service name or Solr installation directory, some of the paths and commands mentioned below will have to be modified accordingly.

Planning Your Upgrade

Here is a checklist of things you need to prepare before starting the upgrade process:

1. Examine the [Solr Upgrade Notes](#) to determine if any behavior changes in the new version of Solr will affect your installation.
2. If not using replication (i.e., collections with replicationFactor less than 1), then you should make a backup of each collection. If all of your collections use replication, then you don't technically need to make a backup since you will be upgrading and verifying each node individually.
3. Determine which Solr node is currently hosting the Overseer leader process in SolrCloud, as you should upgrade this node last. To determine the Overseer, use the Overseer Status API, see: [Collections API](#).
4. Plan to perform your upgrade during a system maintenance window if possible. You'll be doing a rolling restart of your cluster (each node, one-by-one), but we still recommend doing the upgrade when system usage is minimal.
5. Verify the cluster is currently healthy and all replicas are active, as you should not perform an upgrade on a degraded cluster.
6. Re-build and test all custom server-side components against the new Solr JAR files.
7. Determine the values of the following variables that are used by the Solr Control Scripts:
 - ZK_HOST: The ZooKeeper connection string your current SolrCloud nodes use to connect to ZooKeeper; this value will be the same for all nodes in the cluster.
 - SOLR_HOST: The hostname each Solr node used to register with ZooKeeper when joining the SolrCloud cluster; this value will be used to set the **host** Java system property when starting the new Solr process.
 - SOLR_PORT: The port each Solr node is listening on, such as 8983.
 - SOLR_HOME: The absolute path to the Solr home directory for each Solr node; this directory must contain a solr.xml file. This value will be passed to the new Solr process using the solr.solr.home system property, see: [Solr Cores and solr.xml](#).

If you are upgrading from an installation of Solr 5.x or later, these values can typically be found in either /var/solr/solr.in.sh or /etc/default/solr.in.sh.

You should now be ready to upgrade your cluster. Please verify this process in a test or staging cluster before doing it in production.

Upgrade Process

The approach we recommend is to perform the upgrade of each Solr node, one-by-one. In other words, you will need to stop a node, upgrade it to the new version of Solr, and restart it before moving on to the next node. This means that for a short period of time, there will be a mix of "Old Solr" and "New Solr" nodes running in your cluster. We also assume that you will point the new Solr node to your existing Solr home directory where the Lucene index files are managed for each collection on the node. This means that you won't need to move any index files around to perform the upgrade.

Step 1: Stop Solr

Begin by stopping the Solr node you want to upgrade. After stopping the node, if using a replication (i.e., collections with replicationFactor less than 1), verify that all leaders hosted on the downed node have successfully migrated to other replicas; you can do this by visiting the [Cloud panel in the Solr Admin UI](#). If not using replication, then any collections with shards hosted on the downed node will be temporarily off-line.

Step 2: Install Solr as a Service

Please follow the instructions to install Solr as a Service on Linux documented at [Taking Solr to Production](#). Use the -n parameter to avoid automatic start of Solr by the installer script. You need to update the /etc/default/solr.in.sh include file in the next step to complete the upgrade process.



If you have a /var/solr/solr.in.sh file for your existing Solr install, running the install_solr_service.sh script will move this file to its new location: /etc/default/solr.in.sh (see [SOLR-8101](#) for more details)

Step 3: Set Environment Variable Overrides

Open /etc/default/solr.in.sh with a text editor and verify that the following variables are set correctly, or add them bottom of the include file as needed:

```
ZK_HOST=
SOLR_HOST=
SOLR_PORT=
SOLR_HOME=
```

Make sure the user you plan to own the Solr process is the owner of the SOLR_HOME directory. For instance, if you plan to run Solr as the "solr" user and SOLR_HOME is /var/solr/data, then you would do: sudo chown -R solr: /var/solr/data

Step 4: Start Solr

You are now ready to start the upgraded Solr node by doing: sudo service solr start. The upgraded instance will join the existing cluster because you're using the same SOLR_HOME, SOLR_PORT, and SOLR_HOST settings used by the old Solr node; thus, the new server will look like the old node to the running cluster. Be sure to look in /var/solr/logs/solr.log for errors during startup.

Step 5: Run Healthcheck

You should run the Solr **healthcheck** command for all collections that are hosted on the upgraded node before proceeding to upgrade the next node in your cluster. For instance, if the newly upgraded node hosts a replica for the **MyDocuments** collection, then you can run the following command (replace ZK_HOST with the ZooKeeper connection string):

```
/opt/solr/bin/solr healthcheck -c MyDocuments -z ZK_HOST
```

Look for any problems reported about any of the replicas for the collection.

Lastly, repeat Steps 1-5 for all nodes in your cluster.

IndexUpgraderTool

The Lucene distribution includes [a tool that upgrades](#) an index from the previous Lucene version to the current file format.

The tool can be used from command line, or it can be instantiated and executed in Java.

Indexes can **only** be upgraded from the previous major release version to the current major release version.

This means that the IndexUpgraderTool in any Solr 8.x release, for example, can only work with indexes from 7.x releases, but cannot work with indexes from Solr 6.x or earlier.

If you are currently using a release two or more major versions older, such as moving from Solr 6x to Solr 8x, you will need to reindex your content.

The IndexUpgraderTool performs a forceMerge (optimize) down to one segment, which may be undesirable.

In a Solr distribution, the Lucene files are located in `./server/solr-webapp/webapp/WEB-INF/lib`. You will need to include the `lucene-core-<version>.jar` and `lucene-backwards-codecs-<version>.jar` on the classpath when running the tool.

```
java -cp lucene-core-8.1.0.jar:lucene-backward-codecs-8.1.0.jar  
org.apache.lucene.index.IndexUpgrader [-delete-prior-commits] [-verbose] /path/to/index
```

This tool keeps only the last commit in an index. For this reason, if the incoming index has more than one commit, the tool refuses to run by default. Specify `-delete-prior-commits` to override this, allowing the tool to delete all but the last commit.

Upgrading large indexes may take a long time. As a rule of thumb, the upgrade processes about 1 GB per minute.



This tool may reorder documents if the index was partially upgraded before execution (e.g., documents were added). If your application relies on monotonicity of document IDs (i.e., the order in which the documents were added to the index is preserved), do a full optimize instead.

Solr Upgrade Notes

The following notes describe changes to Solr in recent releases that you should be aware of before upgrading.

These notes highlight the biggest changes that may impact the largest number of implementations. It is not a comprehensive list of all changes to Solr in any release.

When planning your Solr upgrade, consider the customizations you have made to your system and review the CHANGES.txt file found in your Solr package. That file includes all the changes and updates that may effect your existing implementation.

Detailed steps for upgrading a Solr cluster are in the section [Upgrading a Solr Cluster](#).

Upgrading to 8.x Releases

If you are upgrading from 7.x, see the section [Upgrading from 7.x Releases](#) below.

Solr 8.1

See the [8.1 Release Notes](#) for an overview of the main new features of Solr 8.1.

When upgrading to 8.1.x, users should be aware of the following major changes from v8.0.

Global maxBooleanClauses Parameter

- The behavior of the maxBooleanClauses parameter has changed to reduce the risk of exponential query expansion when dealing with pathological query strings.

A default upper limit of 1024 clauses is now enforced at the node level. This was the default prior to 7.0, and it can be overridden with a new global parameter in solr.xml. This limit will be enforced for all queries whether explicitly defined by the user (or client), or created by Solr and Lucene internals.

An identical parameter is available in solrconfig.xml for limiting the size of queries explicitly defined by the user (or client), but this per-collection limit will still be restricted by the global limit set in solr.xml.

If your use case demands that you a lot of OR or AND clauses in your queries, upon upgrade to 8.1 you may need to adjust the global maxBooleanClauses parameter since between 7.0 and 8.1 the limit was effectively unbounded.

For more information about the new parameter, see the section [Format of solr.xml: maxBooleanClauses](#).

Security

- JSON Web Tokens (JWT) are now supported for authentication. These allow Solr to assert a user is already authenticated via an external identity provider, such as an OpenID Connect-enabled IdP. For more information, see the section [JWT Authentication Plugin](#).
- A new security plugin for audit logging has been added. A default class SolrLogAuditLoggerPlugin is available and configurable in security.json. The base class is also extendable for adding custom audit plugins if needed. See the section [Audit Logging](#) for more information.

Collections API

- The output of the REQUESTSTATUS command in the Collections API will now include internal asynchronous requests (if any) in the "success" or "failed" keys.
- The CREATE command will now return the appropriate status code (4xx, 5xx, etc.) when the command has failed. Previously, it always returned 0, even in failure.
- The MODIFYCOLLECTION command now accepts an attribute to set a collection as read-only. This can be used to block a collection from receiving any updates while still allowing queries to be served. See the section [MODIFYCOLLECTION](#) for details on how to use it.
- A new command RENAME allows renaming a collection by setting up a one-to-one alias using the new name. For more information, see the section [RENAME](#).
- A new command REINDEXCOLLECTION allows indexing existing stored fields from a source collection into a new collection. For more information, please see the section [REINDEXCOLLECTION](#).

Logging

- The default Log4j2 logging mode has been changed from synchronous to asynchronous. This will improve logging throughput and reduce system contention at the cost of a *slight* chance that some logging messages may be missed in the event of abnormal Solr termination.

If even this slight risk is unacceptable, the Log4j configuration file found in `server/resources/log4j2.xml` has the synchronous logging configuration in a commented section and can be edited to re-enable synchronous logging.

Metrics

- The SolrGangliaReporter has been removed from Solr. The metrics library used by Solr, Dropwizard Metrics, was updated to version 4, and Ganglia support was removed from it due to a dependency on the LGPL license.

Browse UI (Velocity)

- Velocity and Velocity Tools were both upgraded as part of this release. Velocity upgraded from 1.7 to 2.0. Please see <https://velocity.apache.org/engine/2.0/upgrading.html> about upgrading. Velocity Tools upgraded from 2.0 to 3.0. For more details, please see <https://velocity.apache.org/tools/3.0/upgrading.html> for details about the upgrade.

Default Garbage Collector (GC)

- Solr's default GC has been changed from CMS to G1. If you prefer to use CMS or any other GC method, you can modify the GC_TUNE section of `solr.in.sh` (*nix) or `solr.in.cmd` (Windows).

Upgrading from 7.x Releases

The upgrade from 7.x to Solr 8.0 introduces several major changes that you should be aware of before upgrading. These changes are described in the section [Major Changes in Solr 8](#). It's strongly recommended that you do a thorough review of that section before starting your upgrade.



If you run in SolrCloud mode, you must be on Solr version 7.3 or higher in order to upgrade to 8.x.

Upgrading from Pre-7.x Versions

Users upgrading from versions of Solr prior to 7.x are strongly encouraged to consult CHANGES.txt for the details of *all* changes since the version they are upgrading from.

The upgrade from Solr 6.x to Solr 7.0 introduced several **major** changes that you should be aware of before upgrading. Please do a thorough review of the section [Major Changes in Solr 7](#) before starting your upgrade.

A summary of the significant changes between Solr 5.x and Solr 6.0 is in the section [Major Changes from Solr 5 to Solr 6](#).

Major Changes in Solr 8

Solr 8.0 is a major new release of Solr.

This page highlights the biggest changes, including new features you may want to be aware of, and changes in default behavior and deprecated features that have been removed.

Solr 8 Upgrade Planning

Before starting an upgrade to Solr 8, please take the time to review all information about changes from the version you are currently on up to Solr 8.

You should also consider all changes that have been made to Solr in any version you have not upgraded to already. For example, if you are currently using Solr 7.4, you should review changes made in all subsequent 7.x releases in addition to changes for 8.0.

A thorough review of the list in [Major Changes in Earlier 7.x Versions](#), below, as well as the [CHANGES.txt](#) in your Solr instance will help you plan your migration to Solr 8.

Upgrade Prerequisites

If using SolrCloud, you must be on Solr 7.3.0 or higher. Solr's LeaderInRecovery (LIR) functionality [changed significantly](#) in Solr 7.3. While these changes were back-compatible for all subsequent 7.x releases, that compatibility has been removed in 8.0. In order to upgrade to Solr 8.x, all nodes of your cluster must be running Solr 7.3 or higher. If an upgrade is attempted with nodes running versions earlier than 7.3, documents could be lost.

If you are not using Solr in SolrCloud mode (you use Standalone Mode instead), we expect you can upgrade to Solr 8 from any 7.x version without major issues.

Rolling Upgrades with Solr 8

If you are planning to upgrade your cluster using a rolling upgrade model (upgrade each node in succession, as opposed to standing up a brand new 8.x cluster), please read the following carefully.

Solr nodes can listen and serve HTTP/2 or HTTP/1 requests. By default, most internal requests are sent using

HTTP/2. This means, though, that by default Solr 8.0 nodes **cannot** communicate with nodes running pre-8.0 versions of Solr.

However you can start Solr 8.0 with a parameter to force HTTP/1.1 communication until all nodes of the cluster have been upgraded. These are the steps to do rolling updates:

1. Do rolling updates as normally, but start the Solr 8.0 nodes with `-Dsolr.http1=true` as startup parameter. When using this parameter internal requests are sent by using HTTP/1.1.

```
./bin/solr start -c -Dsolr.http1=true -z localhost:2481/solr -s /path/to/solr/home
```

Note the above command **must** be customized for your environment. The section [Solr Control Script Reference](#) has all the possible options. If you are running Solr as a service, you may prefer to review the section [Upgrading a Solr Cluster](#).

2. When all nodes have been upgraded to 8.0, restart each one without the `-Dsolr.http1` parameter.

Reindexing After Upgrades

It is always strongly recommended that you fully reindex your documents after a major version upgrade.

Solr has a new section of the Reference Guide, [Reindexing](#) which covers several strategies for how to reindex.

New Features & Enhancements

HTTP/2 Support

As of Solr 8, Solr nodes support HTTP/2 requests.

Until now, Solr was limited to HTTP/1.1 only. HTTP/1.1 practically allows only one outstanding request per TCP connection which means that for sending multiple requests at the same time multiple TCP connections must be established. This leads to a waste of resources on both-sides and long garbage collection (GC) pauses.

Solr 8 with HTTP/2 support overcomes that problem by allowing multiple requests to be sent in parallel using a same TCP connection.

SSL Support with HTTP/2

In order to support SSL over HTTP/2 connections, Solr uses ALPN.

Java 8 does not include an implementation of ALPN, therefore Solr will start with HTTP/1 only when SSL is enabled and Java 8 is in use.

Client Changes for HTTP/2

`Http2SolrClient` with HTTP/2 and async capabilities based on Jetty Client is introduced. This client replaced `HttpSolrClient` and `ConcurrentUpdateSolrClient` for sending most internal requests (sent by `UpdateShardHandler` and `HttpShardHandler`).

However this causes the following changes in configuration and authentication setup:

- The updateShardHandler parameter maxConnections is no longer used and has been removed.
- The HttpShardHandler parameter maxConnections parameter is no longer being used and has been removed.
- Custom AuthenticationPlugin implementations must provide their own setup for Http2SolrClient through implementing HttpClientBuilderPlugin.setup, or internal requests will not be able to be authenticated.

Metrics Changes for HTTP/2

The Http2SolrClient does not support exposing connection-related metrics. For this reason, the following metrics are no longer available:

- Metrics from QUERY.httpShardHandler:
 - availableConnections
 - leasedConnections
 - maxConnections
 - pendingConnections
- Metrics from UPDATE.updateShardHandler
 - availableConnections
 - leasedConnections
 - maxConnections
 - pendingConnections

Nested Documents

Several improvements have been made for nested document support.

Solr now has the ability to store information about document relationships in the index. This stored information can be used for queries.

The Child Document Transformer can also now return children in nested form if the relationships have been properly stored in the index.

There are a few important changes to highlight in the context of upgrading to Solr 8:

- When JSON data is sent to Solr with nested child documents split using the split parameter, the child documents will now be associated to their parents by the field/label string used in the JSON instead of anonymously.

Most users probably won't notice the distinction since the label is lost unless special fields are in the schema. This choice used to be toggleable with an internal/expert anonChildDocs parameter flag, which has been removed.

- Deleting (or updating) documents by their uniqueKey is now scoped to only consider root documents, not child/nested documents. Thus a delete-by-id won't work on a child document (it will fail silently), and an attempt to update a child document by providing a new document with the same ID would add a new document (which will probably be erroneous).

Both these actions were and still are problematic. In-place-updates are safe though. If you want to delete

certain child documents and if you know they don't themselves have nested children then you must do so with a delete-by-query technique.

- Solr has a new field in the _default configset, called _nest_path_. This field stores the path of the document in the hierarchy for non-root documents.

See the sections [Indexing Nested Documents](#) and [Searching Nested Documents](#) for more information and configuration details.

Configuration and Default Parameter Changes

Schema Changes in 8.0

The following changes impact how fields behave.

Default Scoring (SimilarityFactory)

- If you explicitly use BM25SimilarityFactory in your schema, the absolute scoring will be lower since Lucene changed the calculation of BM25 to remove a multiplication factor (for technical details, see [LUCENE-8563](#) or [SOLR-13025](#)). Ordering of documents will not change in the normal case. Use LegacyBM25SimilarityFactory if you need to force the old 6.x/7.x scoring.

Note that if you have not specified any similarityFactory in the schema, or use the default SchemaSimilarityFactory, then LegacyBM25Similarity is automatically selected when the value for luceneMatchVersion is lower than 8.0.0.

See also the section [Similarity](#) for more information.

Memory Codecs Removed

- Memory codecs have been removed from Lucene (MemoryPostings, MemoryDocValues) and are no longer available in Solr. If you used postingsFormat="Memory" or docValuesFormat="Memory" on any field or field type configuration then either remove that setting to use the default or experiment with one of the other options.

For more information on defining a codec, see the section [Codec Factory](#); for more information on field properties, see the section [Field Type Definitions and Properties](#).

LowerCaseTokenizer

- The LowerCaseTokenizer has been deprecated and is likely to be removed in Solr 9. Users are encouraged to use the LetterTokenizer and the LowerCaseFilter instead.

Default Configset

- The _default configset now includes a ignored_* dynamic field rule.

Indexing Changes in 8.0

The following changes impact how documents are indexed.

Index-time Boosts

- Index-time boosts were removed from [Lucene in version 7.0](#), and in Solr 7.x the syntax was still allowed (although it logged a warning in the logs). The syntax was similar to:

```
{"id": "1", "val_s": {"value": "foo", "boost": 2.0}}
```

This syntax has been removed entirely and if sent to Solr it will now produce an error. This was done in conjunction with the improvements for nested document support.

ParseDateFieldUpdateProcessorFactory

- The date format patterns used by ParseDateFieldUpdateProcessorFactory (used by default in "schemaless mode") are now interpreted by Java 8's `java.time.DateTimeFormatter` instead of Joda Time. The pattern language is very similar but not the same. Typically, simply update the pattern by changing an uppercase 'Z' to lowercase 'z' and that's it.

For the current recommended set of patterns in schemaless mode, see the section [Schemaless Mode](#), or simply examine the `_default configSet` (found in `server/solr/configsets`).

Also note that the default set of date patterns (formats) have expanded from previous releases to subsume those patterns previously handled by the "extract" contrib (Solr Cell / Tika).

Solr Cell

- The extraction contrib ([Solr Cell](#)) no longer does any date parsing, and thus no longer supports the `date.formats` parameter. To ensure date strings are properly parsed, use the ParseDateFieldUpdateProcessorFactory in your update chain. This update request processor is found by default with the "parse-date" update processor when running Solr in "schemaless mode".

Langid Contrib

- The LanguageIdentifierUpdateProcessor base class in the langid contrib (found in `contrib/langid`) changed some method signatures. If you have a custom language identifier implementation you will need to adapt your code. See the Jira issue [SOLR-11774](#) for details of the changes.

Query Changes in 8.0

The following changes impact query behavior.

Highlighting

- The Unified Highlighter parameter `hl.weightMatches` now defaults to true. See the section [Highlighting](#) for more information about Highlighter parameters.

eDisMax Query Parser

- The eDisMax query parser will now throw an error when the `qf` parameter refers to a nonexistent field.

Function Query Parser

- The [Function Query Parser](#) now returns scores that are equal to zero (0) when a negative value is produced. This change is due to the fact that Lucene now requires scores to be positive.

Authentication & Security Changes in 8.0

- Authentication plugins can now intercept internode requests on a per-request basis.
- The Basic Authentication plugin now has an option `forwardCredentials` to let Basic Auth headers be forwarded on inter-node requests in case of distributed search, instead of falling back to PKI.
- Metrics are now reported for authentication requests.

UI Changes in 8.0

- The Radial Graph view of a Solr cluster when running in SolrCloud mode has been removed.
- The Nodes view introduced in Solr 7.5 is now the default when choosing the "Cloud" tab in the left navigation menu.

Autoscaling Changes in 8.0

- The default replica placement strategy used in Solr has been reverted to the "legacy" policy used by Solr 7.4 and previous versions. This is due to multiple bugs in the autoscaling based replica placement strategy that was made default in Solr 7.5 which causes multiple replicas of the same shard to be placed on the same node in addition to the `maxShardsPerNode` and `createNodeSet` parameters being ignored.

Although the default has changed, autoscaling will continue to be used if a cluster policy or preference is specified or a collection level policy is in use.

The default replica placement strategy can be changed to use autoscaling again by setting a cluster property:

```
curl -X POST -H 'Content-type:application/json' --data-binary '  
{  
    "set-obj-property": {  
        "defaults" : {  
            "cluster": {  
                "useLegacyReplicaAssignment":false  
            }  
        }  
    }  
}' http://$SOLR_HOST:$SOLR_PORT/api/cluster
```

- A new command-line option is available via `bin/solr autoscaling` to calculate autoscaling policy suggestions and diagnostic information outside of the running Solr cluster. This option can use the existing autoscaling policy, or test the impact of a new one from a file located on the server filesystem.

These options have been documented in the section [Testing Autoscaling Configuration and Suggestions](#).

Dependency Updates in 8.0

- All Hadoop dependencies have been upgraded to Hadoop 3.2.0 (from 2.7.2).

Major Changes in Earlier 7.x Versions

The following is a list of major changes released between Solr 7.1 and 7.7.

Please be sure to review this list so you understand what may have changed between the version of Solr you are currently running and Solr 8.0.

Solr 7.7

See the [7.7 Release Notes](#) for an overview of the main new features in Solr 7.7.

When upgrading to Solr 7.7.x, users should be aware of the following major changes from v7.6:

Admin UI

- The Admin UI now presents a login screen for any users with authentication enabled on their cluster. Clusters with [Basic Authentication](#) will prompt users to enter a username and password. On clusters configured to use [Kerberos Authentication](#), authentication is handled transparently by the browser as before, but if authentication fails, users will be directed to configure their browser to provide an appropriate Kerberos ticket.

The login screen's purpose is cosmetic only - Admin UI-triggered Solr requests were subject to authentication prior to 7.7 and still are today. The login screen changes only the user experience of providing this authentication.

Distributed Requests

- The shards parameter, used to manually select the shards and replicas that receive distributed requests, now checks nodes against a whitelist of acceptable values for security reasons.

In SolrCloud mode this whitelist is automatically configured to contain all live nodes. In standalone mode the whitelist is empty by default. Upgrading users who use the shards parameter in standalone mode can correct this value by setting the shardsWhitelist property in any shardHandler configurations in their solrconfig.xml file.

For more information, see the [Distributed Request](#) documentation.

Solr 7.6

See the [7.6 Release Notes](#) for an overview of the main new features in Solr 7.6.

When upgrading to Solr 7.6, users should be aware of the following major changes from v7.5:

Collections

- The JSON parameter to set cluster-wide default cluster properties with the [CLUSTERPROP](#) command has changed.

The old syntax nested the defaults into a property named clusterDefaults. The new syntax uses only defaults. The command to use is still set-obj-property.

An example of the new syntax is:

```
{  
  "set-obj-property": {  
    "defaults" : {  
      "collection": {  
        "numShards": 2,  
        "nrtReplicas": 1,  
        "tlogReplicas": 1,  
        "pullReplicas": 1  
      }  
    }  
  }  
}
```

The old syntax will be supported until at least Solr 9, but users are advised to begin using the new syntax as soon as possible.

- The parameter `min_rf` has been deprecated and no longer needs to be provided in order to see the achieved replication factor. This information will now always be returned to the client with the response.

Autoscaling

- An autoscaling policy is now used as the default strategy for selecting nodes on which new replicas or replicas of new collections are created.

A default policy is now in place for all users, which will sort nodes by the number of cores and available freedisk, which means by default a node with the fewest number of cores already on it and the highest available freedisk will be selected for new core creation.

- The change described above has two additional impacts on the `maxShardsPerNode` parameter:
 1. It removes the restriction against using `maxShardsPerNode` when an autoscaling policy is in place. This parameter can now always be set when creating a collection.
 2. It removes the default setting of `maxShardsPerNode=1` when an autoscaling policy is in place. It will be set correctly (if required) regardless of whether an autoscaling policy is in place or not.

The default value of `maxShardsPerNode` is still 1. It can be set to -1 if the old behavior of unlimited `maxShardsPerNode` is desired.

DirectoryFactory

- Lucene has introduced the `ByteBuffersDirectoryFactory` as a replacement for the `RAMDirectoryFactory`, which will be removed in Solr 9.

While most users are still encouraged to use the `NRTCachingDirectoryFactory`, which allows Lucene to select the best directory factory to use, if you have explicitly configured Solr to use the `RAMDirectoryFactory`, you are encouraged to switch to the new implementation as soon as possible before Solr 9 is released.

For more information about the new directory factory, see the Jira issue [LUCENE-8438](#).

For more information about the directory factory configuration in Solr, see the section [DataDir and](#)

[DirectoryFactory in SolrConfig](#).

Solr 7.5

See the [7.5 Release Notes](#) for an overview of the main new features in Solr 7.5.

When upgrading to Solr 7.5, users should be aware of the following major changes from v7.4:

Schema Changes

- Since Solr 7.0, Solr's schema field-guessing has created `_str` fields for all `_txt` fields, and returned those by default with queries. As of 7.5, `_str` fields will no longer be returned by default. They will still be available and can be requested with the `f1` parameter on queries. See also the section on [field guessing](#) for more information about how schema field guessing works.
- The Standard Filter, which has been non-operational since at least Solr v4, has been removed.

Index Merge Policy

- When using the `TieredMergePolicy`, the default merge policy for Solr, `optimize` and `expungeDeletes` now respect the `maxMergedSegmentMB` configuration parameter, which defaults to 5000 (5GB).

If it is absolutely necessary to control the number of segments present after `optimize`, specify `maxSegments` as a positive integer. Setting `maxSegments` higher than 1 are honored on a "best effort" basis.

The `TieredMergePolicy` will also reclaim resources from segments that exceed `maxMergedSegmentMB` more aggressively than earlier.

UIMA Removed

- The UIMA contrib has been removed from Solr and is no longer available.

Logging

- Solr's logging configuration file is now located in `server/resources/log4j2.xml` by default.
- A bug for Windows users has been corrected. When using Solr's examples (`bin/solr start -e`) log files will now be put in the correct location (`example/` instead of `server`). See also [Solr Examples](#) and [Solr Control Script Reference](#) for more information.

Solr 7.4

See the [7.4 Release Notes](#) for an overview of the main new features in Solr 7.4.

When upgrading to Solr 7.4, users should be aware of the following major changes from v7.3:

Logging

- Solr now uses Log4j v2.11. The Log4j configuration is now in `log4j2.xml` rather than `log4j.properties` files. This is a server side change only and clients using SolrJ won't need any changes. Clients can still use any logging implementation which is compatible with SLF4J. We now let Log4j handle rotation of Solr logs at startup, and `bin/solr start` scripts will no longer attempt this nor move existing console or garbage collection logs into `logs/archived` either. See [Configuring Logging](#) for more details about Solr logging.

- Configuring `slowQueryThresholdMillis` now logs slow requests to a separate file named `solr_slow_requests.log`. Previously they would get logged in the `solr.log` file.

Legacy Scaling (non-SolrCloud)

- In the [master-slave model](#) of scaling Solr, a slave no longer commits an empty index when a completely new index is detected on master during replication. To return to the previous behavior pass `false` to `skipCommitOnMasterVersionZero` in the slave section of replication handler configuration, or pass it to the `fetchindex` command.

If you are upgrading from a version earlier than Solr 7.3, please see previous version notes below.

Solr 7.3

See the [7.3 Release Notes](#) for an overview of the main new features in Solr 7.3.

When upgrading to Solr 7.3, users should be aware of the following major changes from v7.2:

ConfigSets

- Collections created without specifying a configset name have used a copy of the `_default` configset since Solr 7.0. Before 7.3, the copied configset was named the same as the collection name, but from 7.3 onwards it will be named with a new `".AUTOCREATED"` suffix. This is to prevent overwriting custom configset names.

Learning to Rank

- The `rq` parameter used with Learning to Rank rerank query parsing no longer considers the `defType` parameter. See [Running a Rerank Query](#) for more information about this parameter.

Autoscaling & AutoAddReplicas

- The behaviour of the autoscaling system will now pause all triggers from execution between the start of actions and the end of a cool down period. The triggers will resume after the cool down period expires. Previously, the cool down period was a fixed period started after actions for a trigger event completed and during this time all triggers continued to run but any events were rejected and tried later.
- The throttling mechanism used to limit the rate of autoscaling events processed has been removed. This deprecates the `actionThrottlePeriodSeconds` setting in the set-properties [Autoscaling API](#) which is now non-operational. Use the `triggerCooldownPeriodSeconds` parameter instead to pause event processing.
- The default value of `autoReplicaFailoverWaitAfterExpiration`, used with the AutoAddReplicas feature, has increased to 120 seconds from the previous default of 30 seconds. This affects how soon Solr adds new replicas to replace the replicas on nodes which have either crashed or shutdown.

Logging

- The default Solr log file size and number of backups have been raised to 32MB and 10 respectively. See the section [Configuring Logging](#) for more information about how to configure logging.

SolrCloud

- The old Leader-In-Recovery implementation (implemented in Solr 4.9) is now deprecated and replaced.

Solr will support rolling upgrades from old 7.x versions of Solr to future 7.x releases until the last release of the 7.x major version.

This means to upgrade to Solr 8 in the future, you will need to be on Solr 7.3 or higher.

- Replicas which are not up-to-date are no longer allowed to become leader. Use the [FORCELEADER command](#) of the Collections API to allow these replicas become leader.

Spatial

- If you are using the spatial JTS library with Solr, you must upgrade to 1.15.0. This new version of JTS is now dual-licensed to include a BSD style license. See the section on [Spatial Search](#) for more information.

Highlighting

- The top-level <highlighting> element in solrconfig.xml is now officially deprecated in favour of the equivalent <searchComponent> syntax. This element has been out of use in default Solr installations for several releases already.

If you are upgrading from a version earlier than Solr 7.2, please see previous version notes below.

Solr 7.2

See the [7.2 Release Notes](#) for an overview of the main new features in Solr 7.2.

When upgrading to Solr 7.2, users should be aware of the following major changes from v7.1:

Local Parameters

- Starting a query string with [local parameters](#) {!myparser ...} is used to switch from one query parser to another, and is intended for use by Solr system developers, not end users doing searches. To reduce negative side-effects of unintended hackability, Solr now limits the cases when local parameters will be parsed to only contexts in which the default parser is "lucene" or "func".

So, if defType=edismax then q={!myparser ...} won't work. In that example, put the desired query parser into the defType parameter.

Another example is if defType=edismax then hl.q={!myparser ...} won't work for the same reason. In this example, either put the desired query parser into the hl.qparser parameter or set hl.qparser=lucene. Most users won't run into these cases but some will need to change.

If you must have full backwards compatibility, use luceneMatchVersion=7.1.0 or an earlier version.

eDisMax Query Parser

- The eDisMax parser by default no longer allows subqueries that specify a Solr parser using either local parameters, or the older _query_ magic field trick.

For example, {!prefix f=myfield v=enterp} or _query_:"{!prefix f=myfield v=enterp}" are not supported by default any longer. If you want to allow power-users to do this, set uf=* query or some other value that includes _query_.

If you need full backwards compatibility for the time being, use luceneMatchVersion=7.1.0 or

something earlier.

If you are upgrading from a version earlier than Solr 7.1, please see previous version notes below.

Solr 7.1

See the [7.1 Release Notes](#) for an overview of the main new features of Solr 7.1.

When upgrading to Solr 7.1, users should be aware of the following major changes from v7.0:

AutoAddReplicas

- The feature to automatically add replicas if a replica goes down, previously available only when storing indexes in HDFS, has been ported to the autoscaling framework. Due to this, autoAddReplicas is now available to all users even if their indexes are on local disks.

Existing users of this feature should not have to change anything. However, they should note these changes:

- Behavior: Changing the autoAddReplicas property from disabled (false) to enabled (true) using [MODIFYCOLLECTION API](#) no longer replaces down replicas for the collection immediately. Instead, replicas are only added if a node containing them went down while autoAddReplicas was enabled. The parameters autoReplicaFailoverBadNodeExpiration and autoReplicaFailoverWorkLoopDelay are no longer used.
- Deprecations: Enabling/disabling autoAddReplicas cluster-wide with the API will be deprecated; use suspend/resume trigger APIs with name=".auto_add_replicas" instead.

More information about the changes to this feature can be found in the section [SolrCloud Automatically Adding Replicas](#).

Metrics Reporters

- Shard and cluster metric reporter configuration now require a class attribute.
 - If a reporter configures the group="shard" attribute then please also configure the class="org.apache.solr.metrics.reporters.solr.SolrShardReporter" attribute.
 - If a reporter configures the group="cluster" attribute then please also configure the class="org.apache.solr.metrics.reporters.solr.SolrClusterReporter" attribute.

See the section [Shard and Cluster Reporters](#) for more information.

Streaming Expressions

- All Stream Evaluators in solrj.io.eval have been refactored to have a simpler and more robust structure. This simplifies and condenses the code required to implement a new Evaluator and makes it much easier for evaluators to handle differing data types (primitives, objects, arrays, lists, and so forth).

ReplicationHandler

- In the ReplicationHandler, the master.commitReserveDuration sub-element is deprecated. Instead please configure a direct commitReserveDuration element for use in all modes (master, slave, cloud).

RunExecutableListener

- The RunExecutableListener was removed for security reasons. If you want to listen to events caused by updates, commits, or optimize, write your own listener as native Java class as part of a Solr plugin.

XML Query Parser

- In the XML query parser (defType=xmlparser or {!xmlparser ... }) the resolving of external entities is now disallowed by default.

If you are upgrading from a version earlier than Solr 7.0, please see [Major Changes in Solr 7](#) before starting your upgrade.

Major Changes in Solr 7

Solr 7 is a major new release of Solr which introduces new features and a number of other changes that may impact your existing installation.

Upgrade Planning

There are major changes in Solr 7 to consider before starting to migrate your configurations and indexes. This page is designed to highlight the biggest changes - new features you may want to be aware of, but also changes in default behavior and deprecated features that have been removed.

There are many hundreds of changes in Solr 7, however, so a thorough review of the [Solr Upgrade Notes](#) as well as the [CHANGES.txt](#) file in your Solr instance will help you plan your migration to Solr 7. This section attempts to highlight some of the major changes you should be aware of.

You should also consider all changes that have been made to Solr in any version you have not upgraded to already. For example, if you are currently using Solr 6.2, you should review changes made in all subsequent 6.x releases in addition to changes for 7.0.

[Reindexing](#) your data is considered the best practice and you should try to do so if possible. However, if reindexing is not feasible, keep in mind you can only upgrade one major version at a time. Thus, Solr 6.x indexes will be compatible with Solr 7 but Solr 5.x indexes will not be.

If you do not reindex now, keep in mind that you will need to either reindex your data or upgrade your indexes before you will be able to move to Solr 8 when it is released in the future. See the section [IndexUpgrader Tool](#) for more details on how to upgrade your indexes.

See also the section [Upgrading a Solr Cluster](#) for details on how to upgrade a SolrCloud cluster.

New Features & Enhancements

Replication Modes

Until Solr 7, the SolrCloud model for replicas has been to allow any replica to become a leader when a leader is lost. This is highly effective for most users, providing reliable failover in case of issues in the cluster. However, it comes at a cost in large clusters because all replicas must be in sync at all times.

To provide additional flexibility, two new types of replicas have been added, named TLOG & PULL. These new

types provide options to have replicas which only sync with the leader by copying index segments from the leader. The TLOG type has an additional benefit of maintaining a transaction log (the "tlog" of its name), which would allow it to recover and become a leader if necessary; the PULL type does not maintain a transaction log, so cannot become a leader.

As part of this change, the traditional type of replica is now named NRT. If you do not explicitly define a number of TLOG or PULL replicas, Solr defaults to creating NRT replicas. If this model is working for you, you will not have to change anything.

See the section [Types of Replicas](#) for more details on the new replica modes, and how define the replica type in your cluster.

Autoscaling

Solr autoscaling is a new suite of features in Solr to make managing a SolrCloud cluster easier and more automated.

At its core, Solr autoscaling provides users with a rule syntax to define preferences and policies for how to distribute nodes and shards in a cluster, with the goal of maintaining a balance in the cluster. As of Solr 7, Solr will take any policy or preference rules into account when determining where to place new shards and replicas created or moved with various Collections API commands.

See the section [SolrCloud Autoscaling](#) for details on the options available in 7.0. Expect more features to be released in subsequent 7.x releases in this area.

Other Features & Enhancements

- The Analytics Component has been refactored.
 - The documentation for this component is in progress; until it is available, please refer to [SOLR-11144](#) for more details.
- There were several other new features released in earlier 6.x releases, which you may have missed:
 - [Learning to Rank](#)
 - [Unified Highlighter](#)
 - [Metrics API](#). See also information about related deprecations in the section [JMX Support and MBeans](#) below.
 - [Payload queries](#)
 - [Streaming Evaluators](#)
 - [/v2 API](#)
 - [Graph streaming expressions](#)

Configuration and Default Changes

New Default ConfigSet

Several changes have been made to configsets that ship with Solr; not only their content but how Solr behaves in regard to them:

- The `data_driven_configset` and `basic_configset` have been removed, and replaced by the `_default`

configset. The sample_techproducts_configset also remains, and is designed for use with the example documents shipped with Solr in the example/exampledocs directory.

- When creating a new collection, if you do not specify a configset, the _default will be used.
 - If you use SolrCloud, the _default configset will be automatically uploaded to ZooKeeper.
 - If you use standalone mode, the instanceDir will be created automatically, using the _default configset as its basis.

Schemaless Improvements

To improve the functionality of Schemaless Mode, Solr now behaves differently when it detects that data in an incoming field should have a text-based field type.

- Incoming fields will be indexed as text_general by default (you can change this). The name of the field will be the same as the field name defined in the document.
- A copy field rule will be inserted into your schema to copy the new text_general field to a new field with the name <name>_str. This field's type will be a strings field (to allow for multiple values). The first 256 characters of the text field will be inserted to the new strings field.

This behavior can be customized if you wish to remove the copy field rule, or to change the number of characters inserted to the string field, or the field type used. See the section [Schemaless Mode](#) for details.



Because copy field rules can slow indexing and increase index size, it's recommended you only use copy fields when you need to. If you do not need to sort or facet on a field, you should remove the automatically-generated copy field rule.

Automatic field creation can be disabled with the update.autoCreateFields property. To do this, you can use the Config API with a command such as:

V1 API

```
curl http://host:8983/solr/mycollection/config -d '{"set-user-property": {"update.autoCreateFields": "false"}}'
```

V2 API

```
curl http://host:8983/api/collections/mycollection/config -d '{"set-user-property": {"update.autoCreateFields": "false"}}'
```

Changes to Default Behaviors

- JSON is now the default response format. If you rely on XML responses, you must now define wt=xml in your request. In addition, line indentation is enabled by default (indent=on).
- The sow parameter (short for "Split on Whitespace") now defaults to false, which allows support for multi-word synonyms out of the box. This parameter is used with the eDismax and standard/"lucene"

query parsers. If this parameter is not explicitly specified as true, query text will not be split on whitespace before analysis.

- The legacyCloud parameter now defaults to false. If an entry for a replica does not exist in state.json, that replica will not get registered.

This may affect users who bring up replicas and they are automatically registered as a part of a shard. It is possible to fall back to the old behavior by setting the property legacyCloud=true, in the cluster properties using the following command:

```
./server/scripts/cloud-scripts/zkcli.sh -zkhost 127.0.0.1:2181 -cmd clusterprop -name legacyCloud -val true
```

- The eDismax query parser parameter lowercaseOperators now defaults to false if the luceneMatchVersion in solrconfig.xml is 7.0.0 or above. Behavior for luceneMatchVersion lower than 7.0.0 is unchanged (so, true). This means that clients must send boolean operators (such as AND, OR and NOT) in upper case in order to be recognized, or you must explicitly set this parameter to true.
- The handleSelect parameter in solrconfig.xml now defaults to false if the luceneMatchVersion is 7.0.0 or above. This causes Solr to ignore the qt parameter if it is present in a request. If you have request handlers without a leading '/', you can set handleSelect="true" or consider migrating your configuration.

The qt parameter is still used as a SolrJ special parameter that specifies the request handler (tail URL path) to use.

- The lucenePlusSort query parser (aka the "Old Lucene Query Parser") has been deprecated and is no longer implicitly defined. If you wish to continue using this parser until Solr 8 (when it will be removed), you must register it in your solrconfig.xml, as in: <queryParser name="lucenePlusSort" class="solr.OldLuceneQParserPlugin"/>.
- The name of TemplateUpdateRequestProcessorFactory is changed to template from Template and the name of AtomicUpdateProcessorFactory is changed to atomic from Atomic
 - Also, TemplateUpdateRequestProcessorFactory now uses {} instead of \${} for template.

Deprecations and Removed Features

Point Fields Are Default Numeric Types

Solr has implemented *PointField types across the board, to replace Trie* based numeric fields. All Trie* fields are now considered deprecated, and will be removed in Solr 8.

If you are using Trie* fields in your schema, you should consider moving to PointFields as soon as feasible. Changing to the new PointField types will require you to reindex your data.

Spatial Fields

The following spatial-related fields have been deprecated:

- LatLonType
- GeoHashField
- SpatialVectorFieldType

- SpatialTermQueryPrefixTreeFieldType

Choose one of these field types instead:

- LatLonPointSpatialField
- SpatialRecursivePrefixTreeField
- RptWithGeometrySpatialField

See the section [Spatial Search](#) for more information.

JMX Support and MBeans

- The <jmx> element in solrconfig.xml has been removed in favor of <metrics><reporter> elements defined in solr.xml.

Limited back-compatibility is offered by automatically adding a default instance of SolrJmxReporter if it's missing AND when a local MBean server is found. A local MBean server can be activated either via ENABLE_REMOTE_JMX_OPTS in solr.in.sh or via system properties, e.g., -Dcom.sun.management.jmxremote. This default instance exports all Solr metrics from all registries as hierarchical MBeans.

This behavior can be also disabled by specifying a SolrJmxReporter configuration with a boolean init argument enabled set to false. For a more fine-grained control users should explicitly specify at least one SolrJmxReporter configuration.

See also the section [The <metrics><reporters> Element](#), which describes how to set up Metrics Reporters in solr.xml. Note that back-compatibility support may be removed in Solr 8.

- MBean names and attributes now follow the hierarchical names used in metrics. This is reflected also in /admin/mbeans and /admin/plugins output, and can be observed in the UI Plugins tab, because now all these APIs get their data from the metrics API. The old (mostly flat) JMX view has been removed.

SolrJ

The following changes were made in SolrJ.

- HttpClientInterceptorPlugin is now HttpClientBuilderPlugin and must work with a SolrHttpClientBuilder rather than an HttpClientConfigurer.
- HttpClientUtil now allows configuring HttpClient instances via SolrHttpClientBuilder rather than an HttpClientConfigurer. Use of env variable SOLR_AUTHENTICATION_CLIENT_CONFIGURER no longer works, please use SOLR_AUTHENTICATION_CLIENT_BUILDER
- SolrClient implementations now use their own internal configuration for socket timeouts, connect timeouts, and allowing redirects rather than what is set as the default when building the HttpClient instance. Use the appropriate setters on the SolrClient instance.
- HttpSolrClient#setAllowCompression has been removed and compression must be enabled as a constructor parameter.
- HttpSolrClient#setDefaultMaxConnectionsPerHost and HttpSolrClient#setMaxTotalConnections have been removed. These now default very high and can only be changed via parameter when creating an HttpClient instance.

Other Deprecations and Removals

- The defaultOperator parameter in the schema is no longer supported. Use the q.op parameter instead. This option had been deprecated for several releases. See the section [Standard Query Parser Parameters](#) for more information.
- The defaultSearchField parameter in the schema is no longer supported. Use the df parameter instead. This option had been deprecated for several releases. See the section [Standard Query Parser Parameters](#) for more information.
- The mergePolicy, mergeFactor and maxMergeDocs parameters have been removed and are no longer supported. You should define a mergePolicyFactory instead. See the section [the mergePolicyFactory](#) for more information.
- The PostingsSolrHighlighter has been deprecated. It's recommended that you move to using the UnifiedHighlighter instead. See the section [Unified Highlighter](#) for more information about this highlighter.
- Index-time boosts have been removed from Lucene, and are no longer available from Solr. If any boosts are provided, they will be ignored by the indexing chain. As a replacement, index-time scoring factors should be indexed in a separate field and combined with the query score using a function query. See the section [Function Queries](#) for more information.
- The StandardRequestHandler is deprecated. Use SearchHandler instead.
- To improve parameter consistency in the Collections API, the parameter names fromNode for the MOVEREPLICA command and source, target for the REPLACENODE command have been deprecated and replaced with sourceNode and targetNode instead. The old names will continue to work for back-compatibility but they will be removed in Solr 8.
- The unused valType option has been removed from ExternalFileField, if you have this in your schema you can safely remove it.

Major Changes in Earlier 6.x Versions

The following summary of changes in earlier 6.x releases highlights significant changes released between Solr 6.0 and 6.6 that were listed in earlier versions of this Guide. Mentions of deprecations are likely superseded by removal in Solr 7, as noted in the above sections.

Note again that this is not a complete list of all changes that may impact your installation, so a thorough review of CHANGES.txt is highly recommended if upgrading from any version earlier than 6.6.

- The Solr contribs map-reduce, morphlines-core and morphlines-cell have been removed.
- JSON Facet API now uses hyper-log-log for numBuckets cardinality calculation and calculates cardinality before filtering buckets by any mincount greater than 1.
- If you use historical dates, specifically on or before the year 1582, you should reindex for better date handling.
- If you use the JSON Facet API (json.facet) with method=stream, you must now set sort='index asc' to get the streaming behavior; otherwise it won't stream. Reminder: method is a hint that doesn't change defaults of other parameters.
- If you use the JSON Facet API (json.facet) to facet on a numeric field and if you use mincount=0 or if you set the prefix, you will now get an error as these options are incompatible with numeric faceting.

- Solr's logging verbosity at the INFO level has been greatly reduced, and you may need to update the log configs to use the DEBUG level to see all the logging messages you used to see at INFO level before.
- We are no longer backing up `solr.log` and `solr_gc.log` files in date-stamped copies forever. If you relied on the `solr_log_<date>` or `solr_gc_log_<date>` being in the logs folder that will no longer be the case. See the section [Configuring Logging](#) for details on how log rotation works as of Solr 6.3.
- The `create/deleteCollection` methods on `MiniSolrCloudCluster` have been deprecated. Clients should instead use the `CollectionAdminRequest API`. In addition, `MiniSolrCloudCluster#uploadConfigDir(File, String)` has been deprecated in favour of `#uploadConfigSet(Path, String)`.
- The `bin/solr.in.sh` (`bin/solr.in.cmd` on Windows) is now completely commented by default. Previously, this wasn't so, which had the effect of masking existing environment variables.
- The `_version_` field is no longer indexed and is now defined with `indexed=false` by default, because the field has DocValues enabled.
- The `/export` handler has been changed so it no longer returns zero (0) for numeric fields that are not in the original document. One consequence of this change is that you must be aware that some tuples will not have values if there were none in the original document.
- Metrics-related classes in `org.apache.solr.util.stats` have been removed in favor of the [Dropwizard metrics library](#). Any custom plugins using these classes should be changed to use the equivalent classes from the metrics library. As part of this, the following changes were made to the output of Overseer Status API:
 - The "totalTime" metric has been removed because it is no longer supported.
 - The metrics "75thPctlRequestTime", "95thPctlRequestTime", "99thPctlRequestTime" and "999thPctlRequestTime" in Overseer Status API have been renamed to "75thPcRequestTime", "95thPcRequestTime" and so on for consistency with stats output in other parts of Solr.
 - The metrics "avgRequestsPerMinute", "5minRateRequestsPerMinute" and "15minRateRequestsPerMinute" have been replaced by corresponding per-second rates viz. "avgRequestsPerSecond", "5minRateRequestsPerSecond" and "15minRateRequestsPerSecond" for consistency with stats output in other parts of Solr.
- A new highlighter named `UnifiedHighlighter` has been added. You are encouraged to try out the `UnifiedHighlighter` by setting `hl.method=unified` and report feedback. It's more efficient/faster than the other highlighters, especially compared to the original `Highlighter`. See `HighlightParams.java` for a listing of highlight parameters annotated with which highlighters use them.
`hl.useFastVectorHighlighter` is now considered deprecated in lieu of `hl.method=fastVector`.
- The `maxWarmingSearchers` [parameter](#) now defaults to 1, and more importantly commits will now block if this limit is exceeded instead of throwing an exception (a good thing). Consequently there is no longer a risk in overlapping commits. Nonetheless users should continue to avoid excessive committing. Users are advised to remove any pre-existing `maxWarmingSearchers` entries from their `solrconfig.xml` files.
- The [Complex Phrase query parser](#) now supports leading wildcards. Beware of its possible heaviness, users are encouraged to use `ReversedWildcardFilter` in index time analysis.
- The JMX metric "avgTimePerRequest" (and the corresponding metric in the metrics API for each handler) used to be a simple non-decaying average based on total cumulative time and the number of requests. The Codahale Metrics implementation applies exponential decay to this value, which heavily biases the average towards the last 5 minutes.

- Parallel SQL now uses Apache Calcite as its SQL framework. As part of this change the default aggregation mode has been changed to facet rather than map_reduce. There have also been changes to the SQL aggregate response and some SQL syntax changes. Consult the [Parallel SQL Interface](#) documentation for full details.

Major Changes from Solr 5 to Solr 6

There are some major changes in Solr 6 to consider before starting to migrate your configurations and indexes.

There are many hundreds of changes, so a thorough review of the [Solr Upgrade Notes](#) section as well as the [CHANGES.txt](#) file in your Solr instance will help you plan your migration to Solr 6. This section attempts to highlight some of the major changes you should be aware of.

Highlights of New Features in Solr 6

Some of the major improvements in Solr 6 include:

Streaming Expressions

Introduced in Solr 5, [Streaming Expressions](#) allow querying Solr and getting results as a stream of data, sorted and aggregated as requested.

Several new expression types have been added in Solr 6:

- Parallel expressions using a MapReduce-like shuffling for faster throughput of high-cardinality fields.
- Daemon expressions to support continuous push or pull streaming.
- Advanced parallel relational algebra like distributed joins, intersections, unions and complements.
- Publish/Subscribe messaging.
- JDBC connections to pull data from other systems and join with documents in the Solr index.

Parallel SQL Interface

Built on streaming expressions, new in Solr 6 is a [Parallel SQL interface](#) to be able to send SQL queries to Solr. SQL statements are compiled to streaming expressions on the fly, providing the full range of aggregations available to streaming expression requests. A JDBC driver is included, which allows using SQL clients and database visualization tools to query your Solr index and import data to other systems.

Cross Data Center Replication

Replication across data centers is now possible with [Cross Data Center Replication](#). Using an active-passive model, a SolrCloud cluster can be replicated to another data center, and monitored with a new API.

Graph QueryParser

A new graph [query parser](#) makes it possible to to graph traversal queries of Directed (Cyclic) Graphs modelled using Solr documents.

DocValues

Most non-text field types in the Solr sample configsets now default to using [DocValues](#).

Java 8 Required

The minimum supported version of Java for Solr 6 (and the [SolrJ client libraries](#)) is now Java 8.

Index Format Changes

Solr 6 has no support for reading Lucene/Solr 4.x and earlier indexes. Be sure to run the Lucene IndexUpgrader included with Solr 5.5 if you might still have old 4x formatted segments in your index. Alternatively: fully optimize your index with Solr 5.5 to make sure it consists only of one up-to-date index segment.

Managed Schema is now the Default

Solr's default behavior when a `solrconfig.xml` does not explicitly define a `<schemaFactory/>` is now dependent on the `luceneMatchVersion` specified in that `solrconfig.xml`. When `luceneMatchVersion < 6.0`, `ClassicIndexSchemaFactory` will continue to be used for back compatibility, otherwise an instance of `ManagedIndexSchemaFactory` will be used.

The most notable impacts of this change are:

- Existing `solrconfig.xml` files that are modified to use `luceneMatchVersion >= 6.0`, but do *not* have an explicitly configured `ClassicIndexSchemaFactory`, will have their `schema.xml` file automatically upgraded to a managed-schema file.
- Schema modifications via the [Schema API](#) will now be enabled by default.

Please review the [Schema Factory Definition in SolrConfig](#) section for more details.

Default Similarity Changes

Solr's default behavior when a Schema does not explicitly define a global `<similarity/>` is now dependent on the `luceneMatchVersion` specified in the `solrconfig.xml`. When `luceneMatchVersion < 6.0`, an instance of `ClassicSimilarityFactory` will be used, otherwise an instance of `SchemaSimilarityFactory` will be used. Most notably this change means that users can take advantage of per Field Type similarity declarations, without needing to also explicitly declare a global usage of `SchemaSimilarityFactory`.

Regardless of whether it is explicitly declared, or used as an implicit global default, `SchemaSimilarityFactory`'s implicit behavior when a Field Types do not declare an explicit `<similarity />` has also been changed to depend on the the `luceneMatchVersion`. When `luceneMatchVersion < 6.0`, an instance of `ClassicSimilarity` will be used, otherwise an instance of `BM25Similarity` will be used. A `defaultSimFromFieldType` init option may be specified on the `SchemaSimilarityFactory` declaration to change this behavior. Please review the `SchemaSimilarityFactory` javadocs for more details.

Replica & Shard Delete Command Changes

`DELETESHARD` and `DELETEREPLICA` now default to deleting the instance directory, data directory, and index directory for any replica they delete. Please review the [Collection API](#) documentation for details on new

request parameters to prevent this behavior if you wish to keep all data on disk when using these commands.

facet.date.* Parameters Removed

The facet.date parameter (and associated facet.date.* parameters) that were deprecated in Solr 3.x have been removed completely. If you have not yet switched to using the equivalent facet.range functionality you must do so now before upgrading.

Using the Solr Administration User Interface

This section discusses the Solr Administration User Interface ("Admin UI").

The [Overview of the Solr Admin UI](#) explains the basic features of the user interface, what's on the initial Admin UI page, and how to configure the interface. In addition, there are pages describing each screen of the Admin UI:

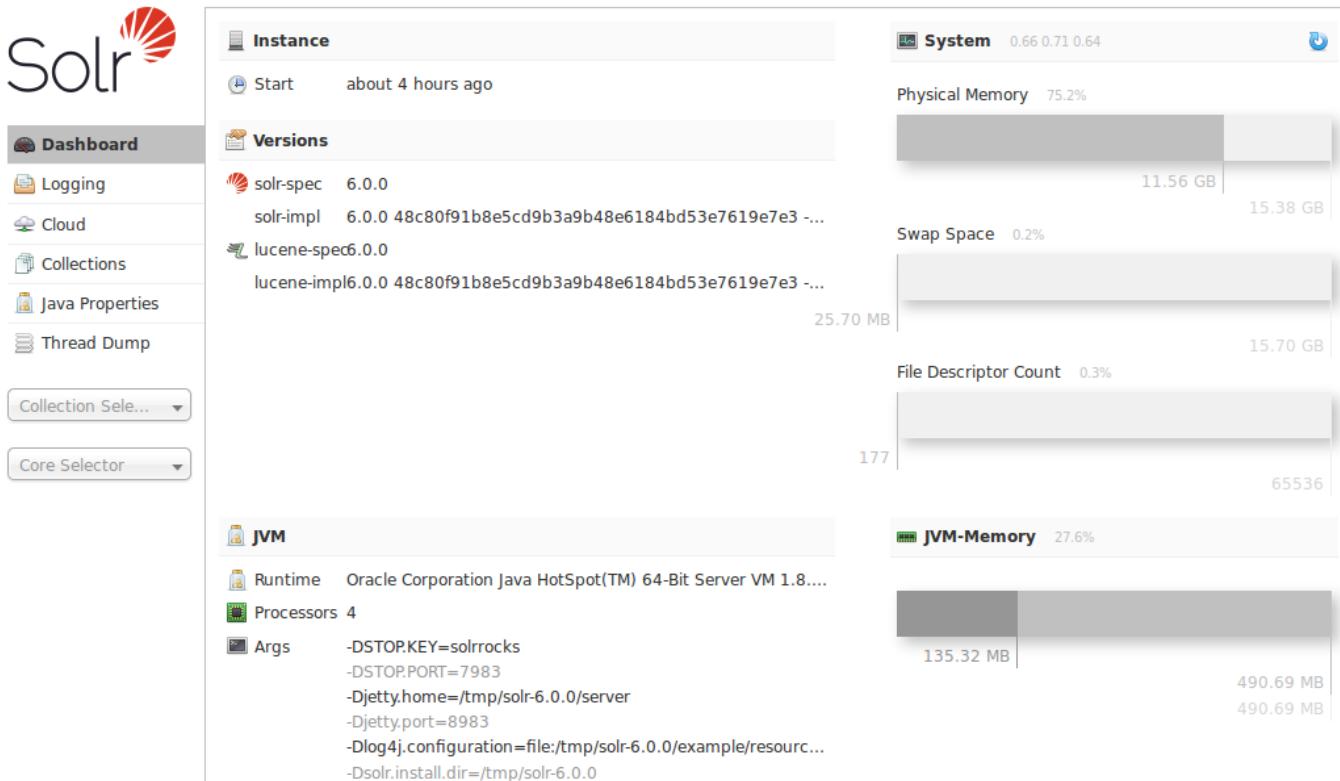
- [Logging](#) shows recent messages logged by this Solr node and provides a way to change logging levels for specific classes.
- [Cloud Screens](#) display information about nodes when running in SolrCloud mode.
- [Collections / Core Admin](#) explains how to get management information about each core.
- [Java Properties](#) shows the Java information about each core.
- [Thread Dump](#) lets you see detailed information about each thread, along with state information.
- [Suggestions Screen](#) displays the state of the system with regard to the autoscaling policies that are in place.
- [Collection-Specific Tools](#) is a section explaining additional screens available for each collection.
 - [Analysis](#) - lets you analyze the data found in specific fields.
 - [Dataimport](#) - shows you information about the current status of the Data Import Handler.
 - [Documents](#) - provides a simple form allowing you to execute various Solr indexing commands directly from the browser.
 - [Files](#) - shows the current core configuration files such as solrconfig.xml.
 - [Query](#) - lets you submit a structured query about various elements of a core.
 - [Stream](#) - allows you to submit streaming expressions and see results and parsing explanations.
 - [Schema Browser](#) - displays schema data in a browser window.
- [Core-Specific Tools](#) is a section explaining additional screens available for each named core.
 - [Ping](#) - lets you ping a named core and determine whether the core is active.
 - [Plugins/Stats](#) - shows statistics for plugins and other installed components.
 - [Replication](#) - shows you the current replication status for the core, and lets you enable/disable replication.
 - [Segments Info](#) - Provides a visualization of the underlying Lucene index segments.

Overview of the Solr Admin UI

Solr features a Web interface that makes it easy for Solr administrators and programmers to view Solr configuration details, run queries and analyze document fields in order to fine-tune a Solr configuration and access online documentation and other help.

Dashboard

Accessing the URL `http://hostname:8983/solr/` will show the main dashboard, which is divided into two parts.



Solr Dashboard

The left-side of the screen is a menu under the Solr logo that provides the navigation through the screens of the UI.

The first set of links are for system-level information and configuration and provide access to [Logging](#), [Collection/Core Administration](#), and [Java Properties](#), among other things.

At the end of this information is at least one pulldown listing Solr cores configured for this instance. On [SolrCloud](#) nodes, an additional pulldown list shows all collections in this cluster. Clicking on a collection or core name shows secondary menus of information for the specified collection or core, such as a [Schema Browser](#), [Config Files](#), [Plugins & Statistics](#), and an ability to perform [Queries](#) on indexed data.

The center of the screen shows the detail of the option selected. This may include a sub-navigation for the option or text or graphical representation of the requested data. See the sections in this guide for each screen for more details.

Under the covers, the Solr Admin UI re-uses the same HTTP APIs available to all clients to access Solr-related

data to drive an external interface.



The path to the Solr Admin UI given above is `http://hostname:port/solr`, which redirects to `http://hostname:port/solr/#/` in the current version. A convenience redirect is also supported, so simply accessing the Admin UI at `http://hostname:port/` will also redirect to `http://hostname:port/solr/#/`.

Login Screen

If authentication has been enabled, Solr will present a login screen to unauthenticated users before allowing them further access to the Admin UI.

Basic Authentication

require authentication

Solr requires authentication for resource Dashboard.
Please log in with your username and password for realm solr.

Username

Password

Login

Documentation Issue Tracker IRC Channel Community forum Solr Query Syntax

Login Screen

This login screen currently only works with Basic Authentication. See the section [Basic Authentication Plugin](#) for details on how to configure Solr to use this method of authentication.

If Kerberos is enabled and the user has a valid ticket, the login screen will be skipped. However, if the user does not have a valid ticket, they will see a message that they need to obtain a valid ticket before continuing.

Getting Assistance

At the bottom of each screen of the Admin UI is a set of links that can be used to get more assistance with configuring and using Solr.

[Documentation](#) [Issue Tracker](#) [IRC Channel](#) [Community forum](#) [Solr Query Syntax](#)

Assistance icons

These icons include the following links.

Link	Description
Documentation	Navigates to the Apache Solr documentation hosted on https://lucene.apache.org/solr/ .
Issue Tracker	Navigates to the JIRA issue tracking server for the Apache Solr project. This server resides at https://issues.apache.org/jira/browse/SOLR .
IRC Channel	Navigates to Solr's IRC live-chat room: http://webchat.freenode.net/?channels=#solr .
Community forum	Navigates to the Apache Wiki page which has further information about ways to engage in the Solr User community mailing lists: https://wiki.apache.org/solr/UsingMailingLists .
Solr Query Syntax	Navigates to the section Query Syntax and Parsing in this Reference Guide.

These links cannot be modified without editing the index.html in the server/solr/solr-webapp directory that contains the Admin UI files.

Logging

The Logging page shows recent messages logged by this Solr node.

When you click the link for "Logging", a page similar to the one below will be displayed:

The screenshot shows the Solr interface with the "Logging" link selected in the sidebar. The main content area displays a table titled "Log4j (org.slf4j.impl.Log4jLoggerFactory)". The table has columns: Time (Local), Level, Core, Logger, and Message. A single row is highlighted in orange, representing an error message. The details are as follows:

Time (Local)	Level	Core	Logger	Message
4/13/2016, 2:46:57 PM	ERROR	false	RequestHandlerBase	org.apache.solr.common.SolrException: ERROR: [doc=ea2675db-c491-4340-af57-c2591524d393] Error adding field 'foo_i'='bar' msg=For input string: "bar"

Below the table, there is a note: "Last Check: 4/13/2016, 2:48:00 PM" and a checkbox for "Show dates in UTC".

The Main Logging Screen, including an example of an error due to a bad document sent by a client

While this example shows logged messages for only one core, if you have multiple cores in a single instance, they will each be listed, with the level for each.

Selecting a Logging Level

When you select the **Level** link on the left, you see the hierarchy of classpaths andclassnames for your instance. A row highlighted in yellow indicates that the class has logging capabilities. Click on a highlighted row, and a menu will appear to allow you to change the log level for that class. Characters in boldface indicate that the class will not be affected by level changes to root.

The screenshot shows the Solr interface with the "Level" link selected in the sidebar. The main content area displays a hierarchical tree view of log levels. The root node "root" is set to "INFO". Under "org.apache.hadoop.http", the "hadoop" node is highlighted in green, indicating it has logging capabilities. A dropdown menu is open over this node, showing the following options: ALL, TRACE, DEBUG, INFO, WARN, ERROR, FATAL, OFF, and UNSET. Other nodes in the tree include "/solr", "org", "apache", "http", "client", "protocol", "conn", and "ssl".

Log level selection

For an explanation of the various logging levels, see [Configuring Logging](#).

Cloud Screens

When running in [SolrCloud](#) mode, a "Cloud" option will appear in the Admin UI between [Logging](#) and [Collections](#).

This screen provides status information about each collection & node in your cluster, as well as access to the low level data being stored in [ZooKeeper](#).

Only Visible When using SolrCloud



The "Cloud" menu option is only available on Solr instances running in [SolrCloud mode](#). Single node or master/slave replication instances of Solr will not display this option.

Click on the "Cloud" option in the left-hand navigation, and a small sub-menu appears with options called "Nodes", "Tree", "ZK Status" and "Graph". The sub-view selected by default is "Nodes".

Nodes View

The "Nodes" view shows a list of the hosts and nodes in the cluster along with key information for each: "CPU", "Heap", "Disk usage", "Requests", "Collections" and "Replicas".

The example below shows the default "cloud" example with some documents added to the "gettingstarted" collection. Details are expanded for node on port 7574, showing more metadata and more metrics details. The screen provides links to navigate to nodes, collections and replicas. The table supports paging and filtering on host/node names and collection names.

Host	Node	CPU	Heap	Disk usage	Requests	Collections	Replicas
192.168.127.248 Mac OS X 16.0Gb Java 1.8 Load: 4.61 show details...	7574_solr Uptime: 7m Java 1.8.0_102 Solr 8.0.0-SNAPSHOT hide details...	1%	26% Max: 490.7Mb Used: 129.8Mb	941.0Kb Total #docs: 62 Avg size/doc: 15.2Kb gettingstarted_s2r6 : 574.0Kb gettingstarted_s1r2 : 367.0Kb	RPM: 0.27 p95: 44ms	gettingstarted	gettingstarted_s2r6 (29 docs) deleted: 0 warmupTime: 0 avg size/doc: 19.8Kb gettingstarted_s1r2 (33 docs) deleted: 0 warmupTime: 0 avg size/doc: 11.1Kb
	8983_solr Uptime: 7m show details...	1%	24%	961.4Kb	RPM: 0.79 p95: 2ms	gettingstarted	gettingstarted_s2r4 (29 docs) gettingstarted_s1r1 (33 docs)

Tree View

The "Tree" view shows a directory structure of the data in ZooKeeper, including cluster wide information regarding the live_nodes and overseer status, as well as collection specific information such as the state.json, current shard leaders, and configuration files in use. In this example, we see part of the state.json definition for the "tlog" collection:

The screenshot shows the Apache Solr Cloud interface. On the left, there's a sidebar with various navigation options: Dashboard, Logging, Cloud (selected), Tree (selected), Graph, Graph (Radial), Collections, Java Properties, Thread Dump, and Suggestions. Below these are two dropdown menus: 'Collection Select...' and 'Core Selector'. The main area has a tree view under the 'Cloud' section, with 'gettingstarted' expanded to show 'tlog' and its sub-directories: counter, leader_elect, leaders, state.json, and terms. To the right of the tree view is a large panel titled 'Metadata' which displays a JSON object representing the configuration of the 'tlog' collection.

```
{"tlog":{ "pullReplicas":2, "replicationFactor":1, "router":{ "name": "compositeId" }, "maxShardsPerNode":4, "autoAddReplicas":false, "nrtReplicas":1, "tlogReplicas":2, "shards":{ "shard1":{ "range": "80000000-ffffffff", "state": "active", "replicas":{ "core_node3":{ "core": "tlog_shard1_replica_t1", "base_url": "http://192.168.1.201:8981/solr", "node_name": "192.168.1.201:8981_solr", "state": "active", "type": "TLOG", "force_set_state": "false"}, "core_node5":{ "core": "tlog_shard1_replica_t2", "base_url": "http://192.168.1.201:8982/solr", "node_name": "192.168.1.201:8982_solr", "state": "active", "type": "TLOG", "force_set_state": "false", "leader": true}, "core_node7":{ "core": "tlog_shard1_replica_p4", "base_url": "http://192.168.1.201:8983/solr", "node_name": "192.168.1.201:8983_solr", "state": "active", "type": "PLOG", "force_set_state": "false"} } }}
```

As an aid to debugging, the data shown in the "Tree" view can be exported locally using the following command bin/solr zk ls -r /

ZK Status View

The "ZK Status" view gives an overview over the ZooKeeper servers or ensemble used by Solr. It lists whether running in standalone or ensemble mode, shows how many ZooKeeper nodes are configured, and then displays a table listing detailed monitoring status for each node, including who is the leader, configuration parameters, and more.

Status: green
ZK connection string: localhost:2181,localhost:2182,localhost:2183
Ensemble size: 3
Ensemble mode: ensemble

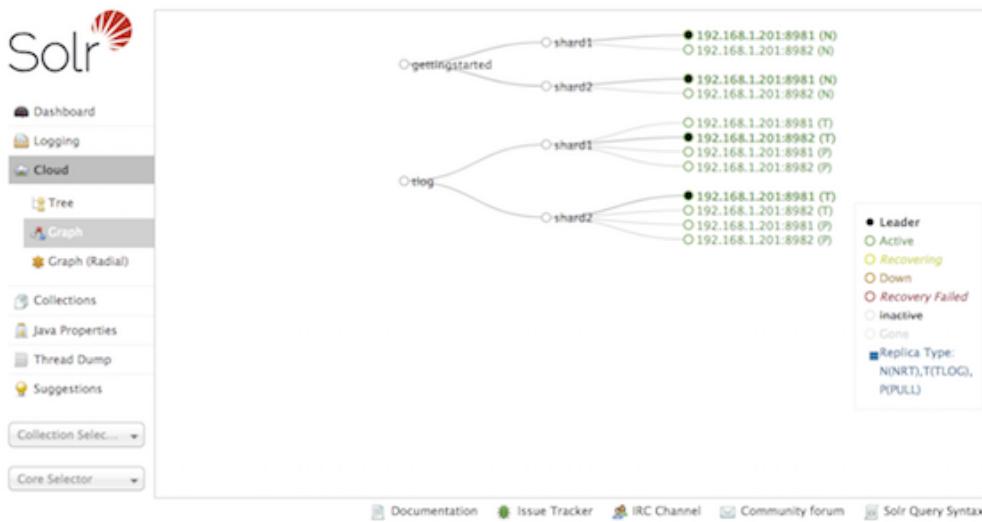
	localhost:2181	localhost:2182	localhost:2183
ok	true	true	true
clientPort	2181	2181	2181
zk_server_state	follower	follower	leader
zk_version	3.4.12	3.4.12	3.4.12
zk_approximate_data_size	210600	210600	210600
zk_znode_count	79	79	79
zk_num_alive_connections	1	3	1
serverId	1	2	3
electionPort	3888	3888	3888
quorumPort	2888	2888	2888
dataDir	/data/version-2	/data/version-2	/data/version-2
dataLogDir	/datalog/version-2	/datalog/version-2	/datalog/version-2
zk_avg_latency	1	0	3
zk_max_file_descriptor_count	1048576	1048576	1048576
zk_watch_count	0	17	0
zk_packets_sent	630	3417	105
zk_packets_received	626	3415	106
tickTime	2000	2000	2000
maxClientCnxns	60	60	60
minSessionTimeout	4000	4000	4000
maxSessionTimeout	40000	40000	40000
peerType	0	0	0
electionAlg	3	3	3
initLimit	5	5	5
syncLimit	2	2	2
zk_followers			2
zk_synced_followers			2
zk_pending_syncs			0

[Documentation](#) [Issue Tracker](#) [IRC Channel](#) [Community forum](#) [Solr Query Syntax](#)

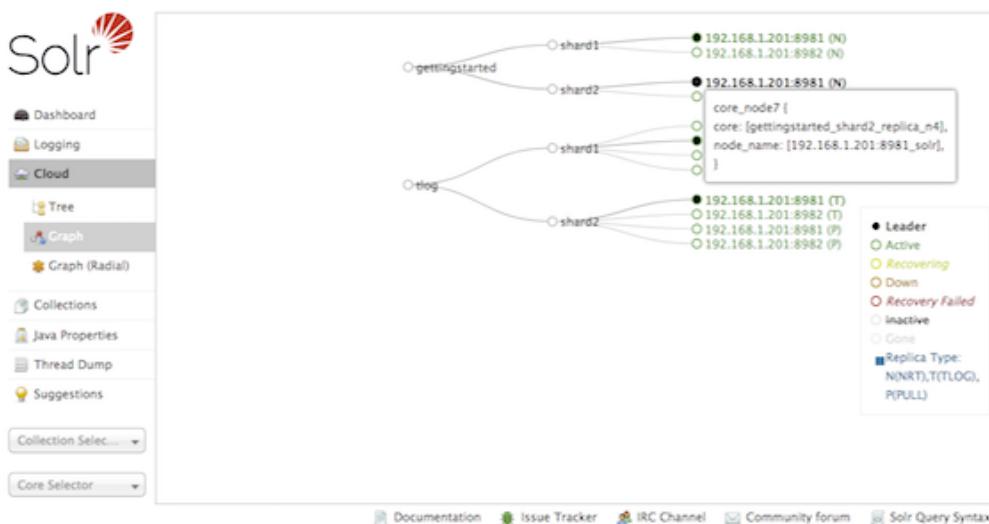
Graph View

The "Graph" view shows a graph of each collection, the shards that make up those collections, and the addresses and type ("NRT", "TLOG" or "PULL") of each replica for each shard.

This example shows a simple cluster. In addition to the 2 shard, 2 replica "gettingstarted" collection, there is an additional "tlog" collection consisting of mixed TLOG and PULL replica types.



Tooltips appear when hovering over each replica giving additional information.



Collections / Core Admin

The Collections screen provides some basic functionality for managing your Collections, powered by the [Collections API](#).

If you are running a single node Solr instance, you will not see a Collections option in the left nav menu of the Admin UI.



You will instead see a "Core Admin" screen that supports some comparable Core level information & manipulation via the [CoreAdmin API](#) instead.

The main display of this page provides a list of collections that exist in your cluster. Clicking on a collection name provides some basic metadata about how the collection is defined, and its current shards & replicas, with options for adding and deleting individual replicas.

The buttons at the top of the screen let you make various collection level changes to your cluster, from add new collections or aliases to reloading or deleting a single collection.

Shard	Replica	State
shard1	core_node2	active
shard1	core_node3	active
shard2		

Replicas can be deleted by clicking the red "X" next to the replica name.

If the shard is inactive, for example after a [SPLITSHARD](#) action, an option to delete the shard will appear as a red "X" next to the shard name.

The screenshot shows the Apache Solr Admin UI interface. On the left, there's a sidebar with links like Dashboard, Logging, Cloud, Collections (which is selected), Java Properties, and Thread Dump. Below the sidebar is a 'Collection Selector' dropdown and a 'Core Selector' dropdown.

The main area displays the configuration for the 'gettingstarted' collection. At the top, there are buttons for Add Collection, Delete, Reload, Create Alias, and Delete Alias. A link 'Use original UI' is also present.

The collection details are as follows:

- Shard count: 4
- configName: gettingstarted
- replicationFactor: 2
- maxShardsPerNode: 2
- router: compositesd
- autoAddReplicas: false

There are three shards listed:

- Shard: shard1**: State: inactive, Range: 80000000-ffffffff. It contains two replicas: Replica: core_node2 and Replica: core_node3.
- Shard: shard2**: State: active, Range: 0-7fffffff. It contains two replicas: Replica: core_node1 and Replica: core_node4.
- Shard: shard1_0** and **Shard: shard1_1**: These are empty shards.

At the bottom of the main area, there are links to Documentation, Issue Tracker, IRC Channel, Community forum, and Solr Query Syntax.

Java Properties

The Java Properties screen provides easy access to one of the most essential components of a top-performing Solr systems. With the Java Properties screen, you can see all the properties of the JVM running Solr, including the class paths, file encodings, JVM memory settings, operating system, and more.

The screenshot shows the Solr Java Properties screen. On the left is a navigation sidebar with links: Dashboard, Logging, Cloud, Collections, Java Properties (which is selected and highlighted in grey), Thread Dump, Collection Sele..., and Core Selector. The main content area displays a table of system properties:

STOP.KEY	solrrocks
STOP.PORT	7983
awt.toolkit	sun.awt.X11.XToolkit
file.encoding	UTF-8
file.encoding.pkg	sun.io
file.separator	/
java.awt.graphicsenv	sun.awt.X11GraphicsEnvironment
java.awt.printerjob	sun.print.PSPrinterJob
java.class.path	<pre>/tmp/solr-6.0.0/server/lib/javax.servlet-api-3.1.0.jar /tmp/solr-6.0.0/server/lib/jetty-continuation-9.3.8.v20160314.jar /tmp/solr-6.0.0/server/lib/jetty-deploy-9.3.8.v20160314.jar /tmp/solr-6.0.0/server/lib/jetty-http-9.3.8.v20160314.jar /tmp/solr-6.0.0/server/lib/jetty-io-9.3.8.v20160314.jar /tmp/solr-6.0.0/server/lib/jetty-jmx-9.3.8.v20160314.jar /tmp/solr-6.0.0/server/lib/jetty-rewrite-9.3.8.v20160314.jar /tmp/solr-6.0.0/server/lib/jetty-security-9.3.8.v20160314.jar /tmp/solr-6.0.0/server/lib/jetty-server-9.3.8.v20160314.jar /tmp/solr-6.0.0/server/lib/jetty-servlet-9.3.8.v20160314.jar /tmp/solr-6.0.0/server/lib/jetty-servlets-9.3.8.v20160314.jar</pre>

Java Properties Screen

Thread Dump

The Thread Dump screen lets you inspect the currently active threads on your server.

Each thread is listed and access to the stacktraces is available where applicable. Icons to the left indicate the state of the thread: for example, threads with a green check-mark in a green circle are in a "RUNNABLE" state. On the right of the thread name, a down-arrow means you can expand to see the stacktrace for that thread.

The screenshot shows the Apache Solr Thread Dump interface. On the left is a sidebar with links: Dashboard, Logging, Cloud, Collections, Java Properties, and Thread Dump (which is selected). Below these are Collection Selector and Core Selector dropdowns. The main area displays a table of threads:

name	cpuTime / userTime
commitScheduler-14-thread-1 (61) ↴	35.3845ms 30.0000ms
commitScheduler-13-thread-1 (60) ↴	5.9529ms 0.0000ms
qtp804611486-57 (57) ↴	32.5933ms 30.0000ms
qtp804611486-56 (56) ↴	31.9840ms 20.0000ms
Scheduler-2047526627 (49) ↴	5.0285ms 0.0000ms
searcherExecutor-9-thread-1 (44) ↴	2.1973ms 0.0000ms
searcherExecutor-8-thread-1 (43) ↴	3.8888ms 0.0000ms
DestroyJavaVM (40) ↴	1257.9428ms 1200.0000ms
Thread-14 (37) ↴	0.0864ms 0.0000ms

List of Threads

When you move your cursor over a thread name, a box floats over the name with the state for that thread. Thread states can be:

State	Meaning
NEW	A thread that has not yet started.
RUNNABLE	A thread executing in the Java virtual machine.
BLOCKED	A thread that is blocked waiting for a monitor lock.
WAITING	A thread that is waiting indefinitely for another thread to perform a particular action.
TIMED_WAITING	A thread that is waiting for another thread to perform an action for up to a specified waiting time.
TERMINATED	A thread that has exited.

When you click on one of the threads that can be expanded, you'll see the stacktrace, as in the example below:

The screenshot shows the Apache Solr dashboard with the 'Thread Dump' section selected. A table lists various threads with their names, stack traces, and performance metrics (cpuTime / userTime). The table has columns for 'name', 'cpuTime / userTime', and a small icon. Most threads have a yellow warning icon, except for one which has a green checkmark icon. The stack traces show calls to methods like 'take', 'getTask', 'runWorker', and 'run'.

name	cpuTime / userTime
commitScheduler-14-thread-1 (61) ↗ ↳ java.util.concurrent.locks.AbstractQueuedSynchronizer\$ConditionObject@637a2a51	35.3845ms 30.0000ms
commitScheduler-13-thread-1 (60) ↗ ↳ java.util.concurrent.locks.AbstractQueuedSynchronizer\$ConditionObject@77d598d8	5.9529ms 0.0000ms
qtp804611486-57 (57) ↗ ↳ java.util.concurrent.locks.AbstractQueuedSynchronizer\$ConditionObject@3926bd31	32.5933ms 30.0000ms
qtp804611486-56 (56) ↗ ↳ sun.misc.Unsafe.park(Native Method)	31.9840ms 20.0000ms
Scheduler-2047526627 (49) ↗ ↳ java.util.concurrent.locks.LockSupport.park(LockSupport.java:175)	5.0285ms 0.0000ms
searcherExecutor-9-thread-1 (44) ↗ ↳ java.util.concurrent.locks.AbstractQueuedSynchronizer\$ConditionObject.await(AbstractQueuedSynchronizer.java:2039)	2.1973ms 0.0000ms
java.util.concurrent.LinkedBlockingQueue.take(LinkedBlockingQueue.java:442) java.util.concurrent.ThreadPoolExecutor.getTask(ThreadPoolExecutor.java:1067) java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1127) java.util.concurrent.ThreadPoolExecutor\$Worker.run(ThreadPoolExecutor.java:617) java.lang.Thread.run(Thread.java:745)	

Inspecting a Thread

You can also check the **Show all Stacktraces** button to automatically enable expansion for all threads.

Suggestions Screen

The Suggestions screen shows violations to an [autoscaling policy](#) that exist in the system, and allows you to take action to correct the violations.

This screen is a visual representation of the output of the [Suggestions API](#).

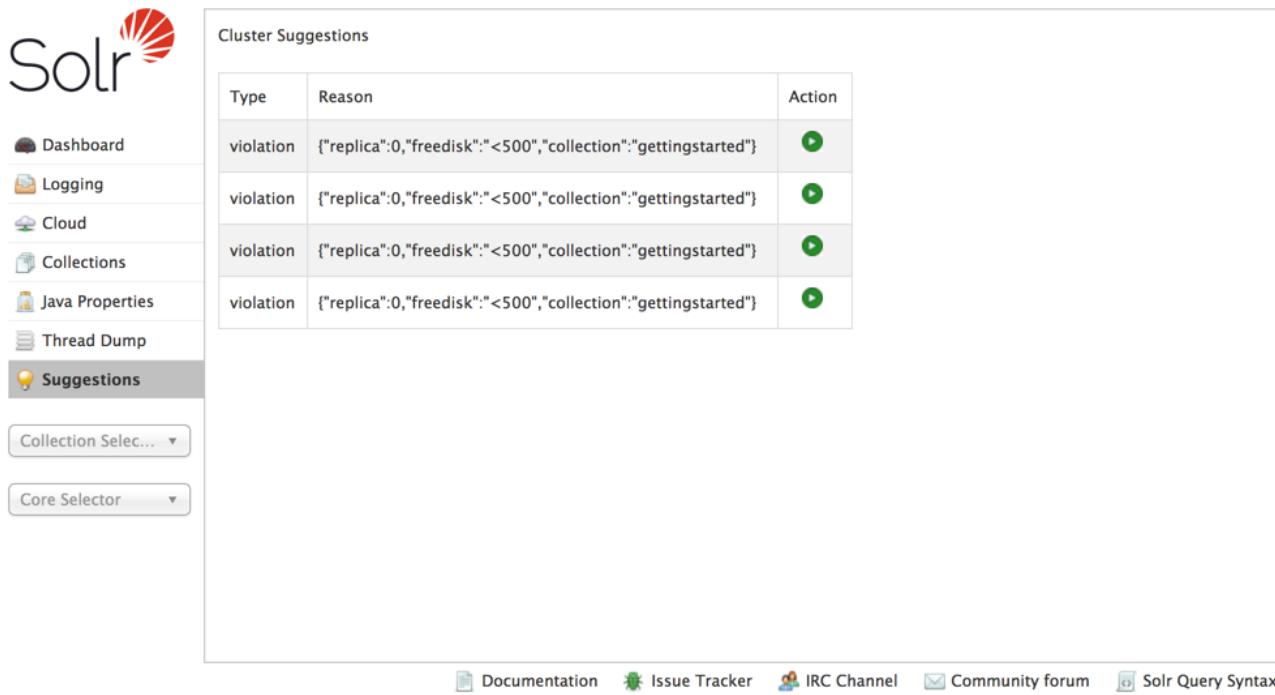
When there are no violations or other suggestions, the screen will appear somewhat blank:

The screenshot shows the Solr admin interface with the 'Suggestions' tab selected in the sidebar. The main content area is titled 'Cluster Suggestions' and contains a table with three columns: 'Type', 'Reason', and 'Action'. A single row is present in the table, with all three columns containing the text 'NA'. At the bottom of the page, there are links to Documentation, Issue Tracker, IRC Channel, Community forum, and Solr Query Syntax.

Type	Reason	Action
NA	NA	NA

Documentation Issue Tracker IRC Channel Community forum Solr Query Syntax

When the system is in violation of an aspect of a policy, each violation will be shown, as in this screenshot:



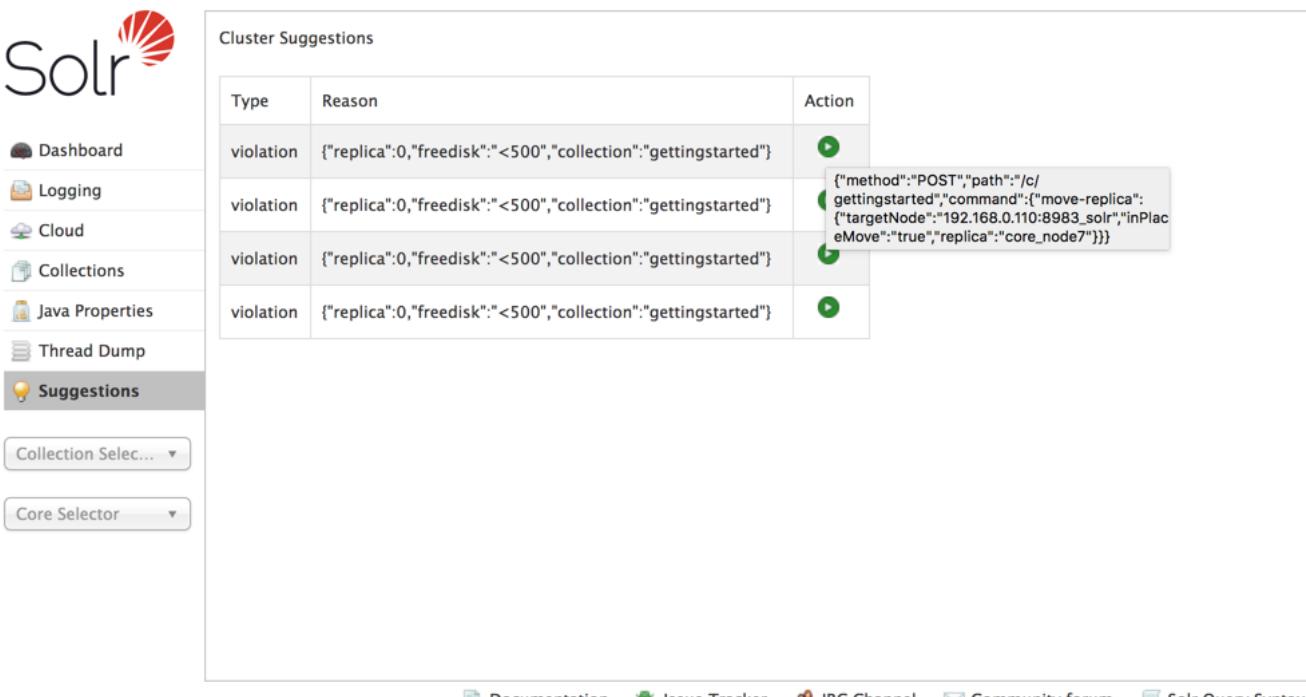
The screenshot shows the Apache Solr interface with the "Suggestions" tab selected. The main content area is titled "Cluster Suggestions" and displays a table of violations:

Type	Reason	Action
violation	{"replica":0,"freedisk":"<500","collection":"gettingstarted"}	

Below the table are two dropdown menus: "Collection Select..." and "Core Selector". At the bottom of the page are links to Documentation, Issue Tracker, IRC Channel, Community forum, and Solr Query Syntax.

A line is shown for each violation. In this case, we have defined a policy where no replica can exist on a node that has less than 500Gb of available disk space. In this example, 4 replicas in our sample cluster violates this rule.

In the "Action" column, the green button allows you to execute the recommended change to allow the system to return to compliance with the policy. If you hover your mouse over this button, you will see the recommended Collections API command:



The screenshot shows the Apache Solr interface with the "Suggestions" tab selected. The main content area is titled "Cluster Suggestions" and displays a table of violations. A tooltip is visible over the "Action" button for the first violation, showing the recommended Collections API command:

```
{"method":"POST","path":"/c/gettingstarted","command":{"move-replica":{"targetNode":"192.168.0.110:8983_solr","inPlaceMove":"true","replica":"core_node7"}}}
```

Below the table are two dropdown menus: "Collection Select..." and "Core Selector". At the bottom of the page are links to Documentation, Issue Tracker, IRC Channel, Community forum, and Solr Query Syntax.

In this case, the recommendation is to issue a MOVEREPLICA command to move this replica to a node with more available disk space.



Since autoscaling features are only available in SolrCloud mode, this screen will only appear when running Solr in SolrCloud mode.

Collection-Specific Tools

In the left-hand navigation bar, you will see a pull-down menu titled "Collection Selector" that can be used to access collection specific administration screens.

Only Visible When Using SolrCloud

The "Collection Selector" pull-down menu is only available on Solr instances running in [SolrCloud mode](#).



Single node or master/slave replication instances of Solr will not display this menu, instead the Collection specific UI pages described in this section will be available in the [Core Selector pull-down menu](#).

Clicking on the Collection Selector pull-down menu will show a list of the collections in your Solr cluster, with a search box that can be used to find a specific collection by name. When you select a collection from the pull-down, the main display of the page will display some basic metadata about the collection, and a secondary menu will appear in the left nav with links to additional collection specific administration screens.

The collection-specific UI screens are listed below, with a link to the section of this guide to find out more:

- [Analysis](#) - lets you analyze the data found in specific fields.
- [Dataimport](#) - shows you information about the current status of the Data Import Handler.
- [Documents](#) - provides a simple form allowing you to execute various Solr indexing commands directly from the browser.
- [Files](#) - shows the current core configuration files such as `solrconfig.xml`.
- [Query](#) - lets you submit a structured query about various elements of a core.
- [Stream](#) - allows you to submit streaming expressions and see results and parsing explanations.
- [Schema Browser](#) - displays schema data in a browser window.

Analysis Screen

The Analysis screen lets you inspect how data will be handled according to the field, field type and dynamic

field configurations found in your Schema. You can analyze how content would be handled during indexing or during query processing and view the results separately or at the same time. Ideally, you would want content to be handled consistently, and this screen allows you to validate the settings in the field type or field analysis chains.

Enter content in one or both boxes at the top of the screen, and then choose the field or field type definitions to use for analysis.

The screenshot shows the Apache Solr Analysis screen. On the left, there's a sidebar with icons for Dashboard, Logging, Cloud, Collections, Java Properties, Thread Dump, and several sections under 'gettingstarted'. The 'Analysis' section is currently selected and highlighted in grey. The main area has two text input fields: 'Field Value (Index)' containing 'Running is a Sport' and 'Field Value (Query)' containing 'run sport'. Below these is a dropdown for 'Analyse Fieldname / FieldType' set to 'text_en', a 'Schema Browser' link, and a checked 'Verbose Output' checkbox. A large table below shows the tokenization process for each input. The table has two columns of five rows each, with abbreviations like ST, SF, LCF, EPF, SKMF, PSF on the left and their corresponding tokens on the right.

	ST	run	sport	
SF	run	run	sport	
LCF	run	run	sport	
EPF	run	run	sport	
SKMF	run	run	sport	
PSF	run	run	sport	
ST	Running	is	a	Sport
SF	Running			Sport
LCF	running			sport
EPF	running			sport
SKMF	running			sport
PSF	run			sport

If you click the **Verbose Output** check box, you see more information, including more details on the transformations to the input (such as, convert to lower case, strip extra characters, etc.) including the raw bytes, type and detailed position information at each stage. The information displayed will vary depending on the settings of the field or field type. Each step of the process is displayed in a separate section, with an abbreviation for the tokenizer or filter that is applied in that step. Hover or click on the abbreviation, and you'll see the name and path of the tokenizer or filter.

Field Value (Index)				Field Value (Query)			
Running is a Sport				run sport			
Analyse Fieldname / FieldType: text_en				Schema Browser			
<input checked="" type="checkbox"/> Verbose Output				Analyze			
ST	text	Running	is	ST	text	run	sport
	raw_bytes	[52 75 6e 6e 69 6e 67]	[69 73]		raw_bytes	[72 75 6e]	[73 70 6f 72 74]
	start	0	8		start	0	4
	end	7	10		end	3	9
	positionLength	1	1		positionLength	1	1
	type	<ALPHANUM>	<ALPHANUM>		type	<ALPHANUM>	<ALPHANUM>
	position	1	2		position	1	2
SF	text	Running		SF	text	run	sport
	raw_bytes	[52 75 6e 6e 69 6e 67]			raw_bytes	[72 75 6e]	[73 70 6f 72 74]
	start	0			start	0	4
	end	7			end	3	9
	positionLength	1			positionLength	1	1
	type	<ALPHANUM>			type	<ALPHANUM>	<ALPHANUM>
	position	1			position	1	2
LCF	text	running		SF	text	run	sport
	raw_bytes	[72 75 6e 6e 69 6e 67]			raw_bytes	[72 75 6e]	[73 70 6f 72 74]
	start	0			start	0	4
	end	7			end	3	9
	positionLength	1			positionLength	1	1
	type	<ALPHANUM>			type	<ALPHANUM>	<ALPHANUM>
	position	1			position	1	2
PPF	text	running		PPF	text	run	sport
	raw_bytes	[72 75 6e 6e 69 6e 67]			raw_bytes	[72 75 6e]	[73 70 6f 72 74]
	start	0			start	0	4
	end	7			end	3	9
	positionLength	1			positionLength	1	1
	type	<ALPHANUM>			type	<ALPHANUM>	<ALPHANUM>
	position	1			position	1	2

In the example screenshot above, several transformations are applied to the input "Running is a sport." The words "is" and "a" have been removed and the word "running" has been changed to its basic form, "run". This is because we are using the field type `text_en` in this scenario, which is configured to remove stop words (small words that usually do not provide a great deal of context) and "stem" terms when possible to find more possible matches (this is particularly helpful with plural forms of words). If you click the question mark next to the **Analyze Fieldname/Field Type** pull-down menu, the [Schema Browser window](#) will open, showing you the settings for the field specified.

The section [Understanding Analyzers, Tokenizers, and Filters](#) describes in detail what each option is and how it may transform your data and the section [Running Your Analyzer](#) has specific examples for using the Analysis screen.

Dataimport Screen

The Dataimport screen shows the configuration of the DataImportHandler (DIH) and allows you start, and monitor the status of, import commands as defined by the options selected on the screen and defined in the configuration file.

The screenshot shows the Apache Solr Dataimport screen. On the left, there's a sidebar with links like Dashboard, Logging, Cloud, Collections, Java Properties, Thread Dump, RSS Overview, Analysis, Dataimport (which is selected), Documents, Files, Query, and Schema. The main area has a title bar with 'Last Update: 14:30:49' and 'No information available (idle)'. It includes sections for Command (set to 'full-import'), Raw Status-Output, Configuration, and a large text area displaying XML configuration code for importing from an RSS feed.

```

<dataConfig>
  <dataSource type="URLDataSource" />
  <document>
    <entity name="slashdot"
      pk="link"
      url="http://rss.slashdot.org/Slashdot/slashdot"
      processor="XPathEntityProcessor"
      forEach="/rss/channel/item"
      transformer="DateFormatTransformer">

      <field column="source" xpath="/rss/channel/title" commonField="true" />
      <field column="source-link" xpath="/rss/channel/link" commonField="true" />
      <field column="subject" xpath="/rss/channel/subject" commonField="true" />

      <field column="title" xpath="/rss/channel/item/title" />
      <field column="link" xpath="/rss/channel/item/link" />
      <field column="description" xpath="/rss/channel/item/description" />
      <field column="creator" xpath="/rss/channel/item/creator" />
      <field column="item-subject" xpath="/rss/channel/item/subject" />

    </entity>
  </document>
</dataConfig>

```

The Dataimport Screen

This screen also lets you adjust various options to control how the data is imported to Solr, and view the data import configuration file that controls the import.

For more information about data importing with DIH, see the section on [Uploading Structured Data Store Data with the Data Import Handler](#).

Documents Screen

The Documents screen provides a simple form allowing you to execute various Solr indexing commands in a variety of formats directly from the browser.

The screenshot shows the Apache Solr Documents screen. The sidebar on the left includes links for Dashboard, Logging, Cloud, Collections, Java Properties, Thread Dump, and 'gettingstarted' (which is selected). The main form has fields for Request-Handler (qt) set to '/update', Document Type (set to JSON), Document(s) containing the JSON object '{ "id": "my_doc_id", "title": "My Document Title" }', Commit Within (set to 1000), Overwrite (set to true), and a 'Submit Document' button. To the right, the status is shown as 'success' with a JSON response message.

```

Status: success
Response:
{
  "responseHeader": {
    "status": 0,
    "QTime": 152
  }
}

```

The Documents Screen

The screen allows you to:

- Submit JSON, CSV or XML documents in Solr-specific format for indexing
- Upload documents (in JSON, CSV or XML) for indexing
- Construct documents by selecting fields and field values

There are other ways to load data, see also these sections:



- [Uploading Data with Index Handlers](#)
- [Uploading Data with Solr Cell using Apache Tika](#)

Common Fields

- Request-Handler: The first step is to define the RequestHandler. By default /update will be defined. Change the request handler to /update/extract to use Solr Cell.
- Document Type: Select the Document Type to define the format of document to load. The remaining parameters may change depending on the document type selected.
- Document(s): Enter a properly-formatted Solr document corresponding to the Document Type selected. XML and JSON documents must be formatted in a Solr-specific format, a small illustrative document will be shown. CSV files should have headers corresponding to fields defined in the schema. More details can be found at: [Uploading Data with Index Handlers](#).
- Commit Within: Specify the number of milliseconds between the time the document is submitted and when it is available for searching.
- Overwrite: If true the new document will replace an existing document with the same value in the id field. If false multiple documents with the same id can be added.



Setting Overwrite to false is very rare in production situations, the default is true.

CSV, JSON and XML Documents

When using these document types the functionality is similar to submitting documents via curl or similar. The document structure must be in a Solr-specific format appropriate for the document type. Examples are illustrated in the Document(s) text box when you select the various types.

These options will only add or overwrite documents; for other update tasks, see the [Solr Command](#) option.

Document Builder

The Document Builder provides a wizard-like interface to enter fields of a document.

File Upload

The File Upload option allows choosing a prepared file and uploading it. If using /update for the Request-Handler option, you will be limited to XML, CSV, and JSON.

Other document types (e.g., Word, PDF, etc.) can be indexed using the ExtractingRequestHandler (aka, Solr Cell). You must modify the RequestHandler to /update/extract, which must be defined in your solrconfig.xml file with your desired defaults. You should also add &literal.id shown in the "Extracting Request Handler Params" field so the file chosen is given a unique id. More information can be found at:

Uploading Data with Solr Cell using Apache Tika

Solr Command

The Solr Command option allows you use the /update request handler with XML or JSON formatted commands to perform specific actions. A few examples are:

- Deleting documents
- Updating only certain fields of documents
- Issuing commit commands on the index

Files Screen

The Files screen lets you browse & view the various configuration files (such solrconfig.xml and the schema file) for the collection you selected.

```
http://127.0.1.1:7574/solr/gettingstarted/admin/file?wt=json&_=1460651618527&fil

<?xml version="1.0" encoding="UTF-8" ?>
<!--
Licensed to the Apache Software Foundation (ASF) under one or more
contributor license agreements. See the NOTICE file distributed with
this work for additional information regarding copyright ownership.
The ASF licenses this file to You under the Apache License, Version 2.0
(the "License"); you may not use this file except in compliance with
the License. You may obtain a copy of the License at
http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License.
-->

<!--
For more details about configurations options that may appear in
this file, see http://wiki.apache.org/solr/SolrConfigXml.
-->
<config>
<!-- In all configuration below, a prefix of "solr." for class names
is an alias that causes solr to search appropriate packages,
including org.apache.solr.common.cloud.Replica and org.apache.solr.common
```

The Files Screen

If you are using [SolrCloud](#), the files displayed are the configuration files for this collection stored in ZooKeeper. In a standalone Solr installations, all files in the conf directory are displayed.

While solrconfig.xml defines the behavior of Solr as it indexes content and responds to queries, the Schema allows you to define the types of data in your content (field types), the fields your documents will be broken into, and any dynamic fields that should be generated based on patterns of field names in the incoming documents. Any other configuration files are used depending on how they are referenced in either solrconfig.xml or your schema.

Configuration files cannot be edited with this screen, so a text editor of some kind must be used.

This screen is related to the [Schema Browser Screen](#), in that they both can display information from the schema, but the Schema Browser provides a way to drill into the analysis chain and displays linkages between field types, fields, and dynamic field rules.

Many of the options defined in these configuration files are described throughout the rest of this Guide. In particular, you will want to review these sections:

- [Indexing and Basic Data Operations](#)
- [Searching](#)
- [The Well-Configured Solr Instance](#)
- [Documents, Fields, and Schema Design](#)

Query Screen

You can use the **Query** screen to submit a search query to a Solr collection and analyze the results.

In the example in the screenshot, a query has been submitted, and the screen shows the query results sent to the browser as JSON.

The screenshot shows the Solr interface with the 'Query' screen selected. On the left, there's a sidebar with links like Dashboard, Logging, Cloud, Collections, Java Properties, Thread Dump, Overview, Analysis, Dataimport, Documents, Files, and Schema. Below that is a 'Core Selector' dropdown set to 'films'. The main area has a 'Request-Handler (qt)' dropdown set to '/select'. It contains several input fields: 'q' (set to 'genre:Fantasy'), 'fq' (empty), 'sort' (empty), 'start, rows' (0, 10), 'fl' (empty), 'df' (empty), and 'Raw Query Parameters' (set to 'key1=val1&key2=val2'). Below these are dropdowns for 'wt' (set to 'json') and checkboxes for 'indent' (checked), 'debugQuery', 'dismax', and 'edismax'. To the right, a browser window shows the URL <http://127.0.0.1:7574/solr/films/select?indent=on&q=genre:Fantasy&wt=json>. The response is a JSON object containing the query parameters and a list of documents. Each document includes fields like id, directed_by, initial_release_date, genre, name, and version.

```
http://127.0.0.1:7574/solr/films/select?indent=on&q=genre:Fantasy&wt=json

{
  "responseHeader": {
    "zkConnected": true,
    "status": 0,
    "QTime": 3,
    "params": {
      "q": "genre:Fantasy",
      "indent": "on",
      "wt": "json",
      "_": "1460656348479"
    }
  },
  "response": {
    "numFound": 82,
    "start": 0,
    "docs": [
      {
        "id": "/en/9_2005",
        "directed_by": ["Shane Acker"],
        "initial_release_date": "2005-04-21T00:00:00Z",
        "genre": ["Computer Animation", "Animation", "Apocalyptic and post-apocalyptic fiction", "Science Fiction", "Short Film", "Thriller", "Fantasy"],
        "name": ["9"],
        "version": 1531518174029152256
      },
      {
        "id": "/en/300_2007",
        "directed_by": ["Zack Snyder"],
        "initial_release_date": "2006-12-09T00:00:00Z",
        "genre": ["Epic film", "Adventure Film", "Fantasy"]
      }
    ]
  }
}
```

JSON Results of a Query

In this example, a query for genre:Fantasy was sent to a "films" collection. Defaults were used for all other options in the form, which are explained briefly in the table below, and covered in detail in later parts of this Guide.

The response is shown to the right of the form. Requests to Solr are simply HTTP requests, and the query submitted is shown in light type above the results; if you click on this it will open a new browser window with just this request and response (without the rest of the Solr Admin UI). The rest of the response is shown in JSON, which is the default output format.

The response has at least two sections, but may have several more depending on the options chosen. The two sections it always has are the responseHeader and the response. The responseHeader includes the status of the search (status), the processing time (QTime), and the parameters (params) that were used to process the query.

The response includes the documents that matched the query, in doc sub-sections. The fields return depend on the parameters of the query (and the defaults of the request handler used). The number of results is also included in this section.

This screen allows you to experiment with different query options, and inspect how your documents were indexed. The query parameters available on the form are some basic options that most users want to have available, but there are dozens more available which could be simply added to the basic request by hand (if opened in a browser). The following parameters are available:

Request-handler (qt)

Specifies the query handler for the request. If a query handler is not specified, Solr processes the response with the standard query handler.

q

The query event. See [Searching](#) for an explanation of this parameter.

fq

The filter queries. See [Common Query Parameters](#) for more information on this parameter.

sort

Sorts the response to a query in either ascending or descending order based on the response's score or another specified characteristic.

start, rows

start is the offset into the query result starting at which documents should be returned. The default value is 0, meaning that the query should return results starting with the first document that matches. This field accepts the same syntax as the start query parameter, which is described in [Searching](#). rows is the number of rows to return.

fl

Defines the fields to return for each document. You can explicitly list the stored fields, [functions](#), and [doc transformers](#) you want to have returned by separating them with either a comma or a space.

wt

Specifies the Response Writer to be used to format the query response. Defaults to JSON if not specified.

indent

Click this button to request that the Response Writer use indentation to make the responses more readable.

debugQuery

Click this button to augment the query response with debugging information, including "explain info" for each document returned. This debugging information is intended to be intelligible to the administrator or programmer.

dismax

Click this button to enable the Dismax query parser. See [The DisMax Query Parser](#) for further information.

edismax

Click this button to enable the Extended query parser. See [The Extended DisMax Query Parser](#) for further information.

hl

Click this button to enable highlighting in the query response. See [Highlighting](#) for more information.

facet

Enables facetting, the arrangement of search results into categories based on indexed terms. See [Faceting](#) for more information.

spatial

Click to enable using location data for use in spatial or geospatial searches. See [Spatial Search](#) for more information.

spellcheck

Click this button to enable the Spellchecker, which provides inline query suggestions based on other, similar, terms. See [Spell Checking](#) for more information.

Stream Screen

The Stream screen allows you to enter a [streaming expression](#) and see the results. It is very similar to the [Query Screen](#), except the input box is at the top and all options must be declared in the expression.

The screen will insert everything up to the streaming expression itself, so you do not need to enter the full URI with the hostname, port, collection, etc. Simply input the expression after the expr= part, and the URL will be constructed dynamically as appropriate.

Under the input box, the Execute button will run the expression. An option "with explanation" will show the parts of the streaming expression that were executed. Under this, the streamed results are shown. A URL to be able to view the output in a browser is also available.

The screenshot shows the Apache Solr Stream screen. At the top, there is a search bar with the text "search(gettingstarted,q="ipod",fl="id, name, price",sort="id asc")". Below the search bar are two buttons: "Execute" and "with explanation". A URL "http://192.168.0.6:8983/solr/gettingstarted/stream?explain=true&expr=search(gettingstarted,q="ipod",fl="id, name, price",sort="id asc")" is displayed. Below the URL, there is a legend: Stream Decorator (pink), Stream Source (blue), Graph Source (green), and Datastore (purple). The Stream Source icon is selected. The results section shows a JSON response with the following data:

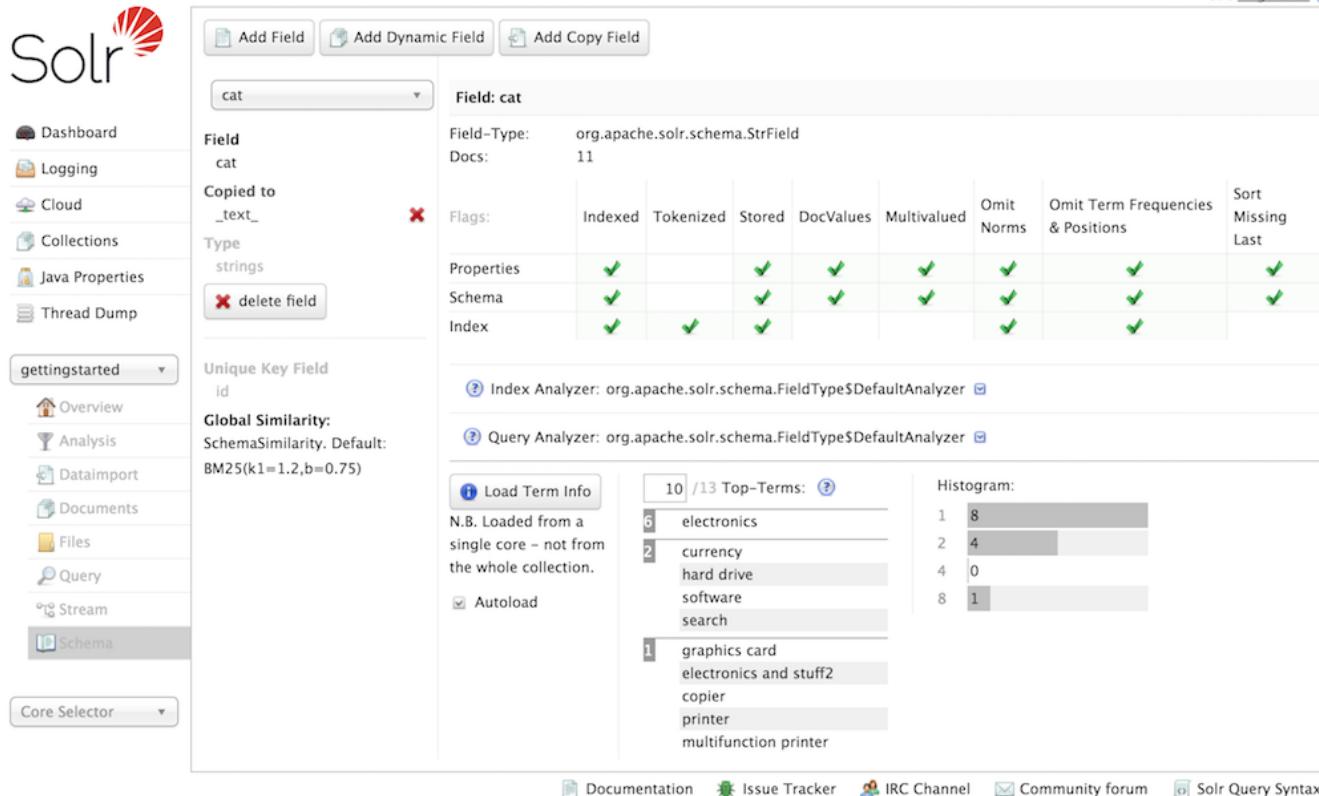
```
{"result-set": { "docs": [ { "name": [ "Belkin Mobile Power Cord for iPod w/ Dock" ], "id": "F8V7067-APL-KIT", "price": [ 19.95 ] }, { "name": [ "iPod & iPod Mini USB 2.0 Cable" ] } ] }}
```

Stream Screen with query and results

Schema Browser Screen

The Schema Browser screen lets you review schema data in a browser window.

If you have accessed this window from the Analysis screen, it will be opened to a specific field, dynamic field rule or field type. If there is nothing chosen, use the pull-down menu to choose the field or field type.

Use original UI 


The screenshot shows the Apache Solr Schema Browser interface. On the left, there's a sidebar with links like Dashboard, Logging, Cloud, Collections, Java Properties, Thread Dump, gettingstarted (which is selected), Overview, Analysis, Dataimport, Documents, Files, Query, Stream, and Schema. Below that is a Core Selector. The main area has a header with 'Add Field', 'Add Dynamic Field', and 'Add Copy Field'. A search bar at the top says 'Field: cat'. The 'Field' section shows 'Field: cat' with 'Field-Type: org.apache.solr.schema.StrField', 'Docs: 11', and 'Copied to: _text_'. It also shows 'Type: strings' and a 'delete field' button. Below this is a 'Unique Key Field' section with 'id'. Under 'Global Similarity:', it says 'SchemaSimilarity. Default: BM25(k1=1.2,b=0.75)'. The right side displays detailed properties for the 'cat' field, including flags (Indexed, Tokenized, Stored, DocValues, Multivalued, Omit Norms, Omit Term Frequencies & Positions) and sorting options (Sort Missing Last). Below these are sections for 'Index Analyzer' and 'Query Analyzer', both set to 'org.apache.solr.schema.FieldType\$DefaultAnalyzer'. There's also a 'Load Term Info' button with a note about loading from a single core, an 'Autoload' checkbox, a 'Top-Terms' list (electronics, currency, hard drive, software, search, graphics card, electronics and stuff2, copier, printer, multifunction printer), and a histogram showing term frequencies.

Schema Browser Screen

The screen provides a great deal of useful information about each particular field and fieldtype in the Schema, and provides a quick UI for adding fields or fieldtypes using the [Schema API](#) (if enabled). In the example above, we have chosen the `cat` field. On the left side of the main view window, we see the field name, that it is copied to the `_text_` (because of a `copyField` rule) and that it use the `strings` fieldtype. Click on one of those field or fieldtype names, and you can see the corresponding definitions.

In the right part of the main view, we see the specific properties of how the `cat` field is defined – either explicitly or implicitly via its fieldtype, as well as how many documents have populated this field. Then we see the analyzer used for indexing and query processing. Click the icon to the left of either of those, and you'll see the definitions for the tokenizers and/or filters that are used. The output of these processes is the information you see when testing how content is handled for a particular field with the [Analysis Screen](#).

Under the analyzer information is a button to **Load Term Info**. Clicking that button will show the top N terms that are in a sample shard for that field, as well as a histogram showing the number of terms with various frequencies. Click on a term, and you will be taken to the [Query Screen](#) to see the results of a query of that term in that field. If you want to always see the term information for a field, choose **Autoload** and it will always appear when there are terms for a field. A histogram shows the number of terms with a given frequency in the field.



Term Information is loaded from single arbitrarily selected core from the collection, to provide a representative sample for the collection. Full [Field Facet](#) query results are needed to see precise term counts across the entire collection.

Core-Specific Tools

The Core-Specific tools are a group of UI screens that allow you to see core-level information.

In the left-hand navigation bar, you will see a pull-down menu titled "Core Selector". Clicking on the menu will show a list of Solr cores hosted on this Solr node, with a search box that can be used to find a specific core by name.

When you select a core from the pull-down, the main display of the page will show some basic metadata about the core, and a secondary menu will appear in the left nav with links to additional core specific administration screens.

The screenshot shows the Apache Solr Core overview screen. On the left, there is a sidebar with a 'Collection Selector' dropdown set to 'films_shard1_r...', and a list of navigation items: Dashboard, Logging, Cloud, Collections, Java Properties, Thread Dump, Collection Selector, Ping, Plugins / Stats, Replication, and Segments info. The 'Segments info' item is highlighted with a blue border.

Statistics

- Last Modified: about 7 hours ago
- Num Docs: 1100
- Max Doc: 1100
- Heap Memory -1
- Usage:
- Deleted Docs: 0
- Version: 6
- Segment 1
- Count:
- Optimized: ✓
- Current: ✓

Instance

- CWD: /tmp/solr-6.0.0/server
- Instance: /tmp/solr-6.0.0/example/cloud/node2
- /solr/films_shard1_replica1
- Data: /tmp/solr-6.0.0/example/cloud/node2
- /solr/films_shard1_replica1/data
- Index: /tmp/solr-6.0.0/example/cloud/node2
- /solr/films_shard1_replica1/data/index
- Impl: org.apache.solr.core.NRTCachingDirectoryFactory

Replication (Master)

	Version	Gen	Size
Master (Searching)	1460569548390	2	375.33 KB
Master (Replicable)	-	-	-

Healthcheck

Ping request handler is not configured with a healthcheck file.

Admin Extra

Core overview screen

The core-specific UI screens are listed below, with a link to the section of this guide to find out more:

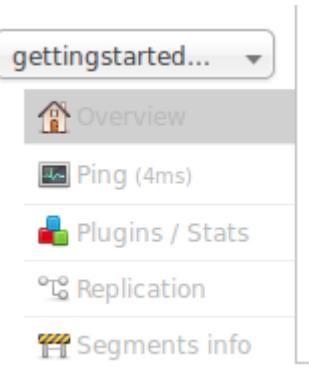
- [Ping](#) - lets you ping a named core and determine whether the core is active.
- [Plugins/Stats](#) - shows statistics for plugins and other installed components.
- [Replication](#) - shows you the current replication status for the core, and lets you enable/disable replication.
- [Segments Info](#) - Provides a visualization of the underlying Lucene index segments.

If you are running a single node instance of Solr, additional UI screens normally displayed on a per-collection bases will also be listed:

- [Analysis](#) - lets you analyze the data found in specific fields.
- [Dataimport](#) - shows you information about the current status of the Data Import Handler.
- [Documents](#) - provides a simple form allowing you to execute various Solr indexing commands directly from the browser.
- [Files](#) - shows the current core configuration files such as solrconfig.xml.
- [Query](#) - lets you submit a structured query about various elements of a core.
- [Stream](#) - allows you to submit streaming expressions and see results and parsing explanations.
- [Schema Browser](#) - displays schema data in a browser window.

Ping

Choosing Ping under a core name issues a ping request to check whether the core is up and responding to requests.



Ping Option in Core Dropdown

The search executed by a Ping is configured with the [Request Parameters API](#). See [Implicit RequestHandlers](#) for the paramset to use for the /admin/ping endpoint.

The Ping option doesn't open a page, but the status of the request can be seen on the core overview page shown when clicking on a collection name. The length of time the request has taken is displayed next to the Ping option, in milliseconds.

Ping API Examples

While the UI screen makes it easy to see the ping response time, the underlying ping command can be more useful when executed by remote monitoring tools:

Input

```
http://localhost:8983/solr/<core-name>/admin/ping
```

This command will ping the core name for a response.

Input

```
http://localhost:8983/solr/<collection-name>/admin/ping?distrib=true&wt=xml
```

This command will ping all replicas of the given collection name for a response:

Sample Output

```

<response>
  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">13</int>
    <lst name="params">
      <str name="q">{!lucene}*>*</str>
      <str name="distrib">false</str>
      <str name="df">_text_</str>
      <str name="rows">10</str>
      <str name="echoParams">all</str>
    </lst>
  </lst>
  <str name="status">OK</str>
</response>

```

Both API calls have the same output. A status=OK indicates that the nodes are responding.

SolrJ Example

```

SolrPing ping = new SolrPing();
ping.getParams().add("distrib", "true"); //To make it a distributed request against a collection
rsp = ping.process(solrClient, collectionName);
int status = rsp.getStatus();

```

Plugins & Stats Screen

The Plugins screen shows information and statistics about the status and performance of various plugins running in each Solr core. You can find information about the performance of the Solr caches, the state of Solr's searchers, and the configuration of Request Handlers and Search Components.

Choose an area of interest on the right, and then drill down into more specifics by clicking on one of the names that appear in the central part of the window. In this example, we've chosen to look at the Searcher stats, from the Core area:

The screenshot shows the Apache Solr interface with the 'Plugins / Stats' tab selected. On the left, there's a sidebar with links like Dashboard, Logging, Cloud, Collections, Java Properties, Thread Dump, Collection Selector, Overview, Ping, and Plugins / Stats. The Plugins / Stats link is highlighted. The main content area shows a tree view under 'core'. The 'searcher' node is expanded, showing its configuration and stats. The configuration includes class: org.apache.solr.search.SolrIndexSearcher, description: index searcher, src:, version: 1.0. The stats section provides detailed information about the searcher's performance, including caching (true), deletedDocs (0), indexVersion (6), maxDoc (1100), numDocs (1100), openedAt (2016-04-13T21:44:50.475Z), reader (ExitableDirectoryReader), readerDir (org.apache.lucene.store.NRTCachingDirectory), lockFactory (org.apache.lucene.store.NativeSLockFactory), registeredAt (2016-04-13T21:44:50.505Z), searcherName (Searcher@#8203;53ee53b3[films_shard1_replica1] main), and warmupTime (0).

Parameter	Value
class:	org.apache.solr.search.SolrIndexSearcher
description:	index searcher
src:	
version:	1.0
stats:	<ul style="list-style-type: none"> caching: true deletedDocs: 0 indexVersion: 6 maxDoc: 1100 numDocs: 1100 openedAt: 2016-04-13T21:44:50.475Z reader: ExitableDirectoryReader readerDir: org.apache.lucene.store.NRTCachingDirectory lockFactory: org.apache.lucene.store.NativeSLockFactory
registeredAt:	2016-04-13T21:44:50.505Z
searcherName:	Searcher@#8203;53ee53b3[films_shard1_replica1] main
warmupTime:	0

Searcher Statistics

The display is a snapshot taken when the page is loaded. You can get updated status by choosing to either **Watch Changes** or **Refresh Values**. Watching the changes will highlight those areas that have changed, while refreshing the values will reload the page with updated information.

Replication Screen

The Replication screen shows you the current replication state for the core you have specified. [SolrCloud](#) has supplanted much of this functionality, but if you are still using Master-Slave index replication, you can use this screen to:

1. View the replicatable index state. (on a master node)
2. View the current replication status (on a slave node)
3. Disable replication. (on a master node)

Caution When Using SolrCloud

When using [SolrCloud](#), do not attempt to disable replication via this screen.

The screenshot shows the Solr dashboard with the 'Replication' core selected. The left sidebar includes links for Dashboard, Logging, Cloud, Collections, Java Properties, Thread Dump, Collection Selection (set to 'gettingstarted...'), and Segments info. The main panel has a 'Refresh Status' button and a 'Disable Replication' button. It displays two tables: one for Index statistics (Master (Searching) and Master (Replicable)) and another for Settings (replication enable: checked, replicateAfter: commit). A yellow lightning bolt icon is visible on the left side of the dashboard.

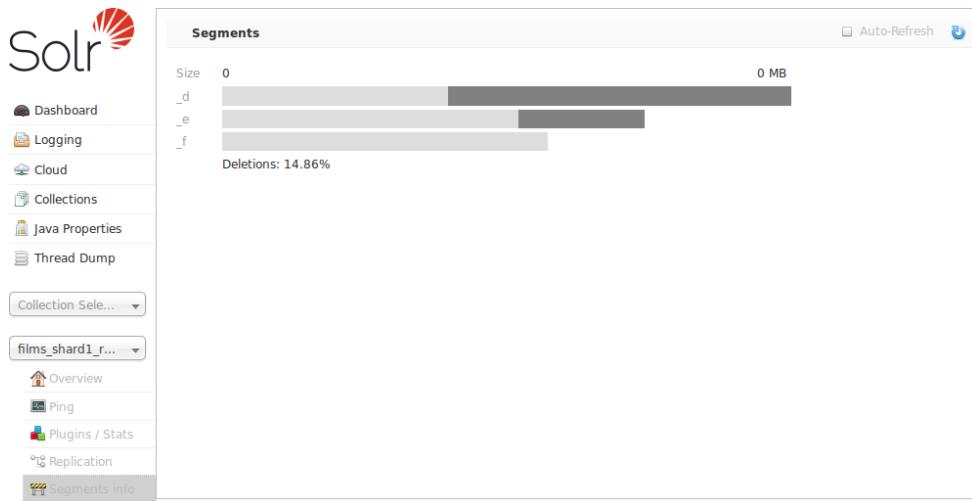
Index	Version	Gen	Size
Master (Searching)	1460583983039	2	2.8 KB
Master (Replicable)	1460583983039	2	-

Settings	replication enable:	✓
(Master):	replicateAfter:	commit

More details on how to configure replication is available in the section called [Index Replication](#).

Segments Info

The Segments Info screen lets you see a visualization of the various segments in the underlying Lucene index for this core, with information about the size of each segment – both bytes and in number of documents – as well as other basic metadata about those segments. Most visible is the the number of deleted documents, but you can hover your mouse over the segments to see additional numeric details.



This information may be useful for people to help make decisions about the optimal [merge settings](#) for their data.

Documents, Fields, and Schema Design

This section discusses how Solr organizes its data into documents and fields, as well as how to work with a schema in Solr.

This section includes the following topics:

[Overview of Documents, Fields, and Schema Design](#): An introduction to the concepts covered in this section.

[Solr Field Types](#): Detailed information about field types in Solr, including the field types in the default Solr schema.

[Defining Fields](#): Describes how to define fields in Solr.

[Copying Fields](#): Describes how to populate fields with data copied from another field.

[Dynamic Fields](#): Information about using dynamic fields in order to catch and index fields that do not exactly conform to other field definitions in your schema.

[Schema API](#): Use curl commands to read various parts of a schema or create new fields and copyField rules.

[Other Schema Elements](#): Describes other important elements in the Solr schema.

[Putting the Pieces Together](#): A higher-level view of the Solr schema and how its elements work together.

[DocValues](#): Describes how to create a docValues index for faster lookups.

[Schemaless Mode](#): Automatically add previously unknown schema fields using value-based field type guessing.

Overview of Documents, Fields, and Schema Design

The fundamental premise of Solr is simple. You give it a lot of information, then later you can ask it questions and find the piece of information you want. The part where you feed in all the information is called *indexing* or *updating*. When you ask a question, it's called a *query*.

One way to understand how Solr works is to think of a loose-leaf book of recipes. Every time you add a recipe to the book, you update the index at the back. You list each ingredient and the page number of the recipe you just added. Suppose you add one hundred recipes. Using the index, you can very quickly find all the recipes that use garbanzo beans, or artichokes, or coffee, as an ingredient. Using the index is much faster than looking through each recipe one by one. Imagine a book of one thousand recipes, or one million.

Solr allows you to build an index with many different fields, or types of entries. The example above shows how to build an index with just one field, `ingredients`. You could have other fields in the index for the recipe's cooking style, like Asian, Cajun, or vegan, and you could have an index field for preparation times. Solr can answer questions like "What Cajun-style recipes that have blood oranges as an ingredient can be prepared in fewer than 30 minutes?"

The schema is the place where you tell Solr how it should build indexes from input documents.

How Solr Sees the World

Solr's basic unit of information is a *document*, which is a set of data that describes something. A recipe document would contain the ingredients, the instructions, the preparation time, the cooking time, the tools needed, and so on. A document about a person, for example, might contain the person's name, biography, favorite color, and shoe size. A document about a book could contain the title, author, year of publication, number of pages, and so on.

In the Solr universe, documents are composed of *fields*, which are more specific pieces of information. Shoe size could be a field. First name and last name could be fields.

Fields can contain different kinds of data. A name field, for example, is text (character data). A shoe size field might be a floating point number so that it could contain values like 6 and 9.5. Obviously, the definition of fields is flexible (you could define a shoe size field as a text field rather than a floating point number, for example), but if you define your fields correctly, Solr will be able to interpret them correctly and your users will get better results when they perform a query.

You can tell Solr about the kind of data a field contains by specifying its *field type*. The field type tells Solr how to interpret the field and how it can be queried.

When you add a document, Solr takes the information in the document's fields and adds that information to an index. When you perform a query, Solr can quickly consult the index and return the matching documents.

Field Analysis

Field analysis tells Solr what to do with incoming data when building an index. A more accurate name for this process would be *processing* or even *digestion*, but the official name is *analysis*.

Consider, for example, a biography field in a person document. Every word of the biography must be

indexed so that you can quickly find people whose lives have had anything to do with ketchup, or dragonflies, or cryptography.

However, a biography will likely contain lots of words you don't care about and don't want clogging up your index—words like "the", "a", "to", and so forth. Furthermore, suppose the biography contains the word "Ketchup", capitalized at the beginning of a sentence. If a user makes a query for "ketchup", you want Solr to tell you about the person even though the biography contains the capitalized word.

The solution to both these problems is field analysis. For the biography field, you can tell Solr how to break apart the biography into words. You can tell Solr that you want to make all the words lower case, and you can tell Solr to remove accents marks.

Field analysis is an important part of a field type. [Understanding Analyzers, Tokenizers, and Filters](#) is a detailed description of field analysis.

Solr's Schema File

Solr stores details about the field types and fields it is expected to understand in a schema file. The name and location of this file may vary depending on how you initially configured Solr or if you modified it later.

- managed-schema is the name for the schema file Solr uses by default to support making Schema changes at runtime via the [Schema API](#), or [Schemaless Mode](#) features. You may [explicitly configure the managed schema features](#) to use an alternative filename if you choose, but the contents of the files are still updated automatically by Solr.
- schema.xml is the traditional name for a schema file which can be edited manually by users who use the [ClassicIndexSchemaFactory](#).
- If you are using SolrCloud you may not be able to find any file by these names on the local filesystem. You will only be able to see the schema through the Schema API (if enabled) or through the Solr Admin UI's [Cloud Screens](#).

Whichever name of the file in use in your installation, the structure of the file is not changed. However, the way you interact with the file will change. If you are using the managed schema, it is expected that you only interact with the file with the Schema API, and never make manual edits. If you do not use the managed schema, you will only be able to make manual edits to the file, the Schema API will not support any modifications.

Note that if you are not using the Schema API yet you do use SolrCloud, you will need to interact with schema.xml through ZooKeeper using upconfig and downconfig commands to make a local copy and upload your changes. The options for doing this are described in [Solr Control Script Reference](#) and [Using ZooKeeper to Manage Configuration Files](#).

Solr Field Types

The field type defines how Solr should interpret data in a field and how the field can be queried. There are many field types included with Solr by default, and they can also be defined locally.

Topics covered in this section:

- [Field Type Definitions and Properties](#)
- [Field Types Included with Solr](#)
- [Working with Currencies and Exchange Rates](#)
- [Working with Dates](#)
- [Working with Enum Fields](#)
- [Working with External Files and Processes](#)
- [Field Properties by Use Case](#)



See also the [FieldType Javadoc](#).

Field Type Definitions and Properties

A field type defines the analysis that will occur on a field when documents are indexed or queries are sent to the index.

A field type definition can include four types of information:

- The name of the field type (mandatory).
- An implementation class name (mandatory).
- If the field type is `TextField`, a description of the field analysis for the field type.
- Field type properties - depending on the implementation class, some properties may be mandatory.

Field Type Definitions in schema.xml

Field types are defined in `schema.xml`. Each field type is defined between `fieldType` elements. They can optionally be grouped within a `types` element. Here is an example of a field type definition for a type called `text_general`:

```

<fieldType name="text_general" class="solr.TextField" positionIncrementGap="100"> ①
  <analyzer type="index"> ②
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
    <!-- in this example, we will only use synonyms at query time
    <filter class="solr.SynonymFilterFactory" synonyms="index_synonyms.txt" ignoreCase="true"
expand="false"/>
    -->
    <filter class="solr.LowerCaseFilterFactory"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
    <filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true" expand=
"true"/>
    <filter class="solr.LowerCaseFilterFactory"/>
  </analyzer>
</fieldType>

```

- ① The first line in the example above contains the field type name, `text_general`, and the name of the implementing class, `solr.TextField`.
- ② The rest of the definition is about field analysis, described in [Understanding Analyzers, Tokenizers, and Filters](#).

The implementing class is responsible for making sure the field is handled correctly. In the class names in `schema.xml`, the string `solr` is shorthand for `org.apache.solr.schema` or `org.apache.solr.analysis`. Therefore, `solr.TextField` is really `org.apache.solr.schema.TextField`.

Field Type Properties

The field type class determines most of the behavior of a field type, but optional properties can also be defined. For example, the following definition of a date field type defines two properties, `sortMissingLast` and `omitNorms`.

```

<fieldType name="date" class="solr.DatePointField"
  sortMissingLast="true" omitNorms="true"/>

```

The properties that can be specified for a given field type fall into three major categories:

- Properties specific to the field type's class.
- [General Properties](#) Solr supports for any field type.
- [Field Default Properties](#) that can be specified on the field type that will be inherited by fields that use this type instead of the default behavior.

General Properties

These are the general properties for fields:

name

The name of the fieldType. This value gets used in field definitions, in the "type" attribute. It is strongly recommended that names consist of alphanumeric or underscore characters only and not start with a digit. This is not currently strictly enforced.

class

The class name that gets used to store and index the data for this type. Note that you may prefix included class names with "solr." and Solr will automatically figure out which packages to search for the class - so solr.TextField will work.

If you are using a third-party class, you will probably need to have a fully qualified class name. The fully qualified equivalent for solr.TextField is org.apache.solr.schema.TextField.

positionIncrementGap

For multivalued fields, specifies a distance between multiple values, which prevents spurious phrase matches.

autoGeneratePhraseQueries

For text fields. If true, Solr automatically generates phrase queries for adjacent terms. If false, terms must be enclosed in double-quotes to be treated as phrases.

synonymQueryStyle

Query used to combine scores of overlapping query terms (i.e., synonyms). Consider a search for "blue tee" with query-time synonyms tshirt, tee.

Use as_same_term (default) to blend terms, i.e., SynonymQuery(tshirt, tee) where each term will be treated as equally important. Use pick_best to select the most significant synonym when scoring Dismax(tee, tshirt). Use as_distinct_terms to bias scoring towards the most significant synonym (pants OR slacks).

as_same_term is appropriate when terms are true synonyms (television, tv). Use pick_best or as_distinct_terms when synonyms are expanding to hyponyms (q=jeans w/ jeans=>jeans, pants) and you want exact to come before parent and sibling concepts. See this [blog article](#).

enableGraphQueries

For text fields, applicable when querying with sow=false (which is the default for the sow parameter). Use true, the default, for field types with query analyzers including graph-aware filters, e.g., [Synonym Graph Filter](#) and [Word Delimiter Graph Filter](#).

Use false for field types with query analyzers including filters that can match docs when some tokens are missing, e.g., [Shingle Filter](#).

docValuesFormat

Defines a custom DocValuesFormat to use for fields of this type. This requires that a schema-aware codec, such as the SchemaCodecFactory has been configured in solrconfig.xml.

postingsFormat

Defines a custom PostingsFormat to use for fields of this type. This requires that a schema-aware codec, such as the SchemaCodecFactory has been configured in solrconfig.xml.



Lucene index back-compatibility is only supported for the default codec. If you choose to customize the postingsFormat or docValuesFormat in your schema.xml, upgrading to a future version of Solr may require you to either switch back to the default codec and optimize your index to rewrite it into the default codec before upgrading, or re-build your entire index from scratch after upgrading.

Field Default Properties

These are properties that can be specified either on the field types, or on individual fields to override the values provided by the field types.

The default values for each property depend on the underlying FieldType class, which in turn may depend on the version attribute of the <schema/>. The table below includes the default value for most FieldType implementations provided by Solr, assuming a schema.xml that declares version="1.6".

Property	Description	Values	Implicit Default
indexed	If true, the value of the field can be used in queries to retrieve matching documents.	true or false	true
stored	If true, the actual value of the field can be retrieved by queries.	true or false	true
docValues	If true, the value of the field will be put in a column-oriented DocValues structure.	true or false	false
sortMissingFirst sortMissingLast	Control the placement of documents when a sort field is not present.	true or false	false
multiValued	If true, indicates that a single document might contain multiple values for this field type.	true or false	false
uninvertible	If true, indicates that an indexed="true" docValues="false" field can be "un-inverted" at query time to build up large in memory data structure to serve in place of DocValues . Defaults to true for historical reasons, but users are strongly encouraged to set this to false for stability and use docValues="true" as needed.	true or false	true
omitNorms	If true, omits the norms associated with this field (this disables length normalization for the field, and saves some memory). Defaults to true for all primitive (non-analyzed) field types, such as int, float, data, bool, and string. Only full-text fields or fields need norms.	true or false	*

Property	Description	Values	Implicit Default
omitTermFreqAndPositions	If true, omits term frequency, positions, and payloads from postings for this field. This can be a performance boost for fields that don't require that information. It also reduces the storage space required for the index. Queries that rely on position that are issued on a field with this option will silently fail to find documents. This property defaults to true for all field types that are not text fields.	true or false	*
omitPositions	Similar to omitTermFreqAndPositions but preserves term frequency information.	true or false	*
termVectors termPositions termOffsets termPayloads	These options instruct Solr to maintain full term vectors for each document, optionally including position, offset and payload information for each term occurrence in those vectors. These can be used to accelerate highlighting and other ancillary functionality, but impose a substantial cost in terms of index size. They are not necessary for typical uses of Solr.	true or false	false
required	Instructs Solr to reject any attempts to add a document which does not have a value for this field. This property defaults to false.	true or false	false
useDocValuesAsStored	If the field has docValues enabled, setting this to true would allow the field to be returned as if it were a stored field (even if it has stored=false) when matching "*" in an fl parameter .	true or false	true
large	Large fields are always lazy loaded and will only take up space in the document cache if the actual value is < 512KB. This option requires stored="true" and multiValued="false". It's intended for fields that might have very large values so that they don't get cached in memory.	true or false	false

Field Type Similarity

A field type may optionally specify a `<similarity/>` that will be used when scoring documents that refer to fields with this type, as long as the "global" similarity for the collection allows it.

By default, any field type which does not define a similarity, uses BM25Similarity. For more details, and examples of configuring both global & per-type Similarities, please see [Other Schema Elements](#).

Field Types Included with Solr

The following table lists the field types that are available in Solr. The `org.apache.solr.schema` package includes all the classes listed in this table.

Class	Description
BinaryField	Binary data.
BoolField	Contains either true or false. Values of 1, t, or T in the first character are interpreted as true. Any other values in the first character are interpreted as false.
CollationField	Supports Unicode collation for sorting and range queries. The <code>ICUCollationField</code> is a better choice if you can use ICU4J. See the section Unicode Collation for more information.
CurrencyField	Deprecated. Use <code>CurrencyFieldType</code> instead.
CurrencyFieldType	Supports currencies and exchange rates. See the section Working with Currencies and Exchange Rates for more information.
DateRangeField	Supports indexing date ranges, to include point in time date instances as well (single-millisecond durations). See the section Working with Dates for more detail on using this field type. Consider using this field type even if it's just for date instances, particularly when the queries typically fall on UTC year/month/day/hour, etc., boundaries.
DatePointField	Date field. Represents a point in time with millisecond precision, encoded using a "Dimensional Points" based data structure that allows for very efficient searches for specific values, or ranges of values. See the section Working with Dates for more details on the supported syntax. For single valued fields, <code>docValues="true"</code> must be used to enable sorting.
DoublePointField	Double field (64-bit IEEE floating point). This class encodes double values using a "Dimensional Points" based data structure that allows for very efficient searches for specific values, or ranges of values. For single valued fields, <code>docValues="true"</code> must be used to enable sorting.
ExternalFileField	Pulls values from a file on disk. See the section Working with External Files and Processes for more information.
EnumField	Deprecated. Use <code>EnumFieldType</code> instead.
EnumFieldType	Allows defining an enumerated set of values which may not be easily sorted by either alphabetic or numeric order (such as a list of severities, for example). This field type takes a configuration file, which lists the proper order of the field values. See the section Working with Enum Fields for more information.
FloatPointField	Floating point field (32-bit IEEE floating point). This class encodes float values using a "Dimensional Points" based data structure that allows for very efficient searches for specific values, or ranges of values. For single valued fields, <code>docValues="true"</code> must be used to enable sorting.

Class	Description
ICUCollationField	Supports Unicode collation for sorting and range queries. See the section Unicode Collation for more information.
IntPointField	Integer field (32-bit signed integer). This class encodes int values using a "Dimensional Points" based data structure that allows for very efficient searches for specific values, or ranges of values. For single valued fields, docValues="true" must be used to enable sorting.
LatLonPointSpatialField	A latitude/longitude coordinate pair; possibly multi-valued for multiple points. Usually it's specified as "lat,lon" order with a comma. See the section Spatial Search for more information.
LatLonType	Deprecated. Consider using the LatLonPointSpatialField instead. A single-valued latitude/longitude coordinate pair. Usually it's specified as "lat,lon" order with a comma. See the section Spatial Search for more information.
LongPointField	Long field (64-bit signed integer). This class encodes foo values using a "Dimensional Points" based data structure that allows for very efficient searches for specific values, or ranges of values. For single valued fields, docValues="true" must be used to enable sorting.
PointType	A single-valued n-dimensional point. It's both for sorting spatial data that is <i>not</i> lat-lon, and for some more rare use-cases. (NOTE: this is <i>not</i> related to the "Point" based numeric fields). See Spatial Search for more information.
PreAnalyzedField	<p>Provides a way to send to Solr serialized token streams, optionally with independent stored values of a field, and have this information stored and indexed without any additional text processing.</p> <p>Configuration and usage of PreAnalyzedField is documented in the section Working with External Files and Processes.</p>
RandomSortField	Does not contain a value. Queries that sort on this field type will return results in random order. Use a dynamic field to use this feature.
SpatialRecursivePrefixTreeFieldType	(RPT for short) Accepts latitude comma longitude strings or other shapes in WKT format. See Spatial Search for more information.
StrField	String (UTF-8 encoded string or Unicode). Strings are intended for small fields and are <i>not</i> tokenized or analyzed in any way. They have a hard limit of slightly less than 32K.
SortableTextField	A specialized version of TextField that allows (and defaults to) docValues="true" for sorting on the first 1024 characters of the original string prior to analysis. The number of characters used for sorting can be overridden with the maxCharsForDocValues attribute.
TextField	Text, usually multiple words or tokens.
TrieDateField	Deprecated. Use DatePointField instead.
TrieDoubleField	Deprecated. Use DoublePointField instead.
TrieFloatField	Deprecated. Use FloatPointField instead.

Class	Description
TrieIntField	Deprecated. Use IntPointField instead.
TrieLongField	Deprecated. Use LongPointField instead.
TrieField	Deprecated. This field takes a type parameter to define the specific class of Trie* field to use; Use an appropriate Point Field type instead.
UUIDField	<p>Universally Unique Identifier (UUID). Pass in a value of NEW and Solr will create a new UUID.</p> <p>Note: configuring a UUIDField instance with a default value of NEW is not advisable for most users when using SolrCloud (and not possible if the UUID value is configured as the unique key field) since the result will be that each replica of each document will get a unique UUID value. Using UUIDUpdateProcessorFactory to generate UUID values when documents are added is recommended instead.</p>

All Trie* numeric and date field types have been deprecated in favor of *Point field types.



Point field types are better at range queries (speed, memory, disk), however simple field:value queries underperform relative to Trie. Either accept this, or continue to use Trie fields. This shortcoming may be addressed in a future release.

Working with Currencies and Exchange Rates

The currency FieldType provides support for monetary values to Solr/Lucene with query-time currency conversion and exchange rates. The following features are supported:

- Point queries
- Range queries
- Function range queries
- Sorting
- Currency parsing by either currency code or symbol
- Symmetric & asymmetric exchange rates (asymmetric exchange rates are useful if there are fees associated with exchanging the currency)
- Range faceting (using either facet.range or type:range in json.facet) as long as the start and end values are specified in the same Currency.

Configuring Currencies



CurrencyField has been Deprecated

CurrencyField has been deprecated in favor of CurrencyFieldType; all configuration examples below use CurrencyFieldType.

The currency field type is defined in schema.xml. This is the default configuration of this type.

```
<fieldType name="currency" class="solr.CurrencyFieldType"
    amountLongSuffix="_l_ns" codeStrSuffix="_s_ns"
    defaultCurrency="USD" currencyConfig="currency.xml" />
```

In this example, we have defined the name and class of the field type, and defined the defaultCurrency as "USD", for U.S. Dollars. We have also defined a currencyConfig to use a file called "currency.xml". This is a file of exchange rates between our default currency to other currencies. There is an alternate implementation that would allow regular downloading of currency data. See [Exchange Rates](#) below for more.

Many of the example schemas that ship with Solr include a [dynamic field](#) that uses this type, such as this example:

```
<dynamicField name="*_c" type="currency" indexed="true" stored="true"/>
```

This dynamic field would match any field that ends in _c and make it a currency typed field.

At indexing time, money fields can be indexed in a native currency. For example, if a product on an e-commerce site is listed in Euros, indexing the price field as "1000,EUR" will index it appropriately. The price should be separated from the currency by a comma, and the price must be encoded with a floating point value (a decimal point).

During query processing, range and point queries are both supported.

Sub-field Suffixes

You must specify parameters amountLongSuffix and codeStrSuffix, corresponding to dynamic fields to be used for the raw amount and the currency dynamic sub-fields, for example:

```
<fieldType name="currency" class="solr.CurrencyFieldType"
    amountLongSuffix="_l_ns" codeStrSuffix="_s_ns"
    defaultCurrency="USD" currencyConfig="currency.xml" />
```

In the above example, the raw amount field will use the "*_l_ns" dynamic field, which must exist in the schema and use a long field type, i.e., one that extends LongValueFieldType. The currency code field will use the "*_s_ns" dynamic field, which must exist in the schema and use a string field type, i.e., one that is or extends StrField.

Atomic Updates won't work if dynamic sub-fields are stored



As noted on [Updating Parts of Documents](#), stored dynamic sub-fields will cause indexing to fail when you use Atomic Updates. To avoid this problem, specify stored="false" on those dynamic fields.

Exchange Rates

You configure exchange rates by specifying a provider. Natively, two provider types are supported: FileExchangeRateProvider or OpenExchangeRatesOrgProvider.

FileExchangeRateProvider

This provider requires you to provide a file of exchange rates. It is the default, meaning that to use this provider you only need to specify the file path and name as a value for currencyConfig in the definition for this type.

There is a sample currency.xml file included with Solr, found in the same directory as the schema.xml file. Here is a small snippet from this file:

```
<currencyConfig version="1.0">
  <rates>
    <!-- Updated from http://www.exchangerate.com/ at 2011-09-27 -->
    <rate from="USD" to="ARS" rate="4.333871" comment="ARGENTINA Peso" />
    <rate from="USD" to="AUD" rate="1.025768" comment="AUSTRALIA Dollar" />
    <rate from="USD" to="EUR" rate="0.743676" comment="European Euro" />
    <rate from="USD" to="CAD" rate="1.030815" comment="CANADA Dollar" />

    <!-- Cross-rates for some common currencies -->
    <rate from="EUR" to="GBP" rate="0.869914" />
    <rate from="EUR" to="NOK" rate="7.800095" />
    <rate from="GBP" to="NOK" rate="8.966508" />

    <!-- Asymmetrical rates -->
    <rate from="EUR" to="USD" rate="0.5" />
  </rates>
</currencyConfig>
```

OpenExchangeRatesOrgProvider

You can configure Solr to download exchange rates from [OpenExchangeRates.Org](#), with updates rates between USD and 170 currencies hourly. These rates are symmetrical.

In this case, you need to specify the providerClass in the definitions for the field type and sign up for an API key. Here is an example:

```
<fieldType name="currency" class="solr.CurrencyFieldType"
  amountLongSuffix="_l_ns" codeStrSuffix="_s_ns"
  providerClass="solr.OpenExchangeRatesOrgProvider"
  refreshInterval="60"
  ratesFileLocation=
  "http://www.openexchangerates.org/api/latest.json?app_id=yourPersonalAppIdKey"/>
```

The refreshInterval is minutes, so the above example will download the newest rates every 60 minutes. The refresh interval may be increased, but not decreased.

Working with Dates

Date Formatting

Solr's date fields (DatePointField, DateRangeField and the deprecated TrieDateField) represent "dates" as a point in time with millisecond precision. The format used is a restricted form of the canonical representation of dateTime in the [XML Schema specification](#) – a restricted subset of [ISO-8601](#). For those familiar with Java date handling, Solr uses `DateTimeFormatter.ISO_INSTANT` for formatting, and parsing too with "leniency".

YYYY-MM-DDThh:mm:ssZ

- YYYY is the year.
- MM is the month.
- DD is the day of the month.
- hh is the hour of the day as on a 24-hour clock.
- mm is minutes.
- ss is seconds.
- Z is a literal 'Z' character indicating that this string representation of the date is in UTC

Note that no time zone can be specified; the String representations of dates is always expressed in Coordinated Universal Time (UTC). Here is an example value:

1972-05-20T17:33:18Z

You can optionally include fractional seconds if you wish, although any precision beyond milliseconds will be ignored. Here are example values with sub-seconds:

- 1972-05-20T17:33:18.772Z
- 1972-05-20T17:33:18.77Z
- 1972-05-20T17:33:18.7Z

There must be a leading '-' for dates prior to year 0000, and Solr will format dates with a leading '+' for years after 9999. Year 0000 is considered year 1 BC; there is no such thing as year 0 AD or BC.

Query escaping may be required

As you can see, the date format includes colon characters separating the hours, minutes, and seconds. Because the colon is a special character to Solr's most common query parsers, escaping is sometimes required, depending on exactly what you are trying to do.



This is normally an invalid query: `datefield:1972-05-20T17:33:18.772Z`

These are valid queries:

```
datefield:1972-05-20T17\:\:33\:\:18.772Z  
datefield:"1972-05-20T17:33:18.772Z"  
datefield:[1972-05-20T17:33:18.772Z TO *]
```

Date Range Formatting

Solr's DateRangeField supports the same point in time date syntax described above (with *date math* described below) and more to express date ranges. One class of examples is truncated dates, which

represent the entire date span to the precision indicated. The other class uses the range syntax ([TO]). Here are some examples:

- 2000-11 – The entire month of November, 2000.
- 1605-11-05 – The Fifth of November.
- 2000-11-05T13 – Likewise but for an hour of the day (1300 to before 1400, i.e., 1pm to 2pm).
- -0009 – The year 10 BC. A 0 in the year position is 0 AD, and is also considered 1 BC.
- [2000-11-01 TO 2014-12-01] – The specified date range at a day resolution.
- [2014 TO 2014-12-01] – From the start of 2014 till the end of the first day of December.
- [* TO 2014-12-01] – From the earliest representable time thru till the end of the day on 2014-12-01.

Limitations: The range syntax doesn't support embedded date math. If you specify a date instance supported by DatePointField with date math truncating it, like NOW/DAY, you still get the first millisecond of that day, not the entire day's range. Exclusive ranges (using { & }) work in *queries* but not for *indexing* ranges.

Date Math

Solr's date field types also supports *date math* expressions, which makes it easy to create times relative to fixed moments in time, include the current time which can be represented using the special value of "NOW".

Date Math Syntax

Date math expressions consist either adding some quantity of time in a specified unit, or rounding the current time by a specified unit. Expressions can be chained and are evaluated left to right.

For example: this represents a point in time two months from now:

NOW+2MONTHS

This is one day ago:

NOW-1DAY

A slash is used to indicate rounding. This represents the beginning of the current hour:

NOW/HOUR

The following example computes (with millisecond precision) the point in time six months and three days into the future and then rounds that time to the beginning of that day:

NOW+6MONTHS+3DAYS/DAY

Note that while date math is most commonly used relative to NOW it can be applied to any fixed moment in time as well:

1972-05-20T17:33:18.772Z+6MONTHS+3DAYS/DAY

Request Parameters That Affect Date Math

NOW

The NOW parameter is used internally by Solr to ensure consistent date math expression parsing across multiple nodes in a distributed request. But it can be specified to instruct Solr to use an arbitrary moment in time (past or future) to override for all situations where the the special value of "NOW" would impact date math expressions.

It must be specified as a (long valued) milliseconds since epoch.

Example:

```
q=solr&fq=start_date:[* TO NOW]&NOW=1384387200000
```

TZ

By default, all date math expressions are evaluated relative to the UTC TimeZone, but the TZ parameter can be specified to override this behaviour, by forcing all date based addition and rounding to be relative to the specified [time zone](#).

For example, the following request will use range facetting to facet over the current month, "per day" relative UTC:

```
http://localhost:8983/solr/my_collection/select?q=*&facet.range=my_date_field&facet=true&facet.range.start=NOW/MONTH&facet.range.end=NOW/MONTH%2B1MONTH&facet.range.gap=%2B1DAY&wt=xml
```

```
<int name="2013-11-01T00:00:00Z">0</int>
<int name="2013-11-02T00:00:00Z">0</int>
<int name="2013-11-03T00:00:00Z">0</int>
<int name="2013-11-04T00:00:00Z">0</int>
<int name="2013-11-05T00:00:00Z">0</int>
<int name="2013-11-06T00:00:00Z">0</int>
<int name="2013-11-07T00:00:00Z">0</int>
...
...
```

While in this example, the "days" will be computed relative to the specified time zone - including any applicable Daylight Savings Time adjustments:

```
http://localhost:8983/solr/my_collection/select?q=*&facet.range=my_date_field&facet=true&facet.range.start=NOW/MONTH&facet.range.end=NOW/MONTH%2B1MONTH&facet.range.gap=%2B1DAY&TZ=America/Los_Angeles&wt=xml
```

```
<int name="2013-11-01T07:00:00Z">0</int>
<int name="2013-11-02T07:00:00Z">0</int>
<int name="2013-11-03T07:00:00Z">0</int>
<int name="2013-11-04T08:00:00Z">0</int>
<int name="2013-11-05T08:00:00Z">0</int>
<int name="2013-11-06T08:00:00Z">0</int>
<int name="2013-11-07T08:00:00Z">0</int>
...
...
```

More DateRangeField Details

DateRangeField is almost a drop-in replacement for places where DatePointField is used. The only difference is that Solr's XML or SolrJ response formats will expose the stored data as a String instead of a Date. The underlying index data for this field will be a bit larger. Queries that align to units of time a second on up should be faster than TrieDateField, especially if it's in UTC.

The main point of DateRangeField, as its name suggests, is to allow indexing date ranges. To do that, simply supply strings in the format shown above. It also supports specifying 3 different relational predicates between the indexed data, and the query range:

- Intersects (default)
- Contains
- Within

You can specify the predicate by querying using the op local-params parameter like so:

```
fq={!field f=dateRange op=Contains}[2013 TO 2018]
```

Unlike most local parameters, op is actually *not* defined by any query parser (field), it is defined by the field type, in this case DateRangeField. In the above example, it would find documents with indexed ranges that *contain* (or equals) the range 2013 thru 2018. Multi-valued overlapping indexed ranges in a document are effectively coalesced.

For a DateRangeField example use-case, see [see Solr's community wiki](#).

Working with Enum Fields

EnumFieldType allows defining a field whose values are a closed set, and the sort order is pre-determined but is not alphabetic nor numeric. Examples of this are severity lists, or risk definitions.

EnumField has been Deprecated



EnumField has been deprecated in favor of EnumFieldType; all configuration examples below use EnumFieldType.

Defining an EnumFieldType in schema.xml

The EnumFieldType type definition is quite simple, as in this example defining field types for "priorityLevel"

and "riskLevel" enumerations:

```
<fieldType name="priorityLevel" class="solr.EnumFieldType" docValues="true" enumsConfig="enumsConfig.xml" enumName="priority"/>
<fieldType name="riskLevel" class="solr.EnumFieldType" docValues="true" enumsConfig="enumsConfig.xml" enumName="risk" />
```

Besides the name and the class, which are common to all field types, this type also takes two additional parameters:

enumsConfig

the name of a configuration file that contains the <enum/> list of field values and their order that you wish to use with this field type. If a path to the file is not defined specified, the file should be in the conf directory for the collection.

enumName

the name of the specific enumeration in the enumsConfig file to use for this type.

Note that docValues="true" must be specified either in the EnumFieldType fieldType or field specification.

Defining the EnumFieldType Configuration File

The file named with the enumsConfig parameter can contain multiple enumeration value lists with different names if there are multiple uses for enumerations in your Solr schema.

In this example, there are two value lists defined. Each list is between enum opening and closing tags:

```
<?xml version="1.0" ?>
<enumsConfig>
  <enum name="priority">
    <value>Not Available</value>
    <value>Low</value>
    <value>Medium</value>
    <value>High</value>
    <value>Urgent</value>
  </enum>
  <enum name="risk">
    <value>Unknown</value>
    <value>Very Low</value>
    <value>Low</value>
    <value>Medium</value>
    <value>High</value>
    <value>Critical</value>
  </enum>
</enumsConfig>
```

Changing Values



You cannot change the order, or remove, existing values in an <enum/> without reindexing.

You can however add new values to the end.

Working with External Files and Processes

The ExternalFileField Type

The ExternalFileField type makes it possible to specify the values for a field in a file outside the Solr index. For such a field, the file contains mappings from a key field to the field value. Another way to think of this is that, instead of specifying the field in documents as they are indexed, Solr finds values for this field in the external file.



External fields are not searchable. They can be used only for function queries or display.

For more information on function queries, see the section on [Function Queries](#).

The ExternalFileField type is handy for cases where you want to update a particular field in many documents more often than you want to update the rest of the documents. For example, suppose you have implemented a document rank based on the number of views. You might want to update the rank of all the documents daily or hourly, while the rest of the contents of the documents might be updated much less frequently. Without ExternalFileField, you would need to update each document just to change the rank. Using ExternalFileField is much more efficient because all document values for a particular field are stored in an external file that can be updated as frequently as you wish.

In schema.xml, the definition of this field type might look like this:

```
<fieldType name="entryRankFile" keyField="pkId" defVal="0" stored="false" indexed="false" class="solr.ExternalFileField"/>
```

The keyField attribute defines the key that will be defined in the external file. It is usually the unique key for the index, but it doesn't need to be as long as the keyField can be used to identify documents in the index. A defVal defines a default value that will be used if there is no entry in the external file for a particular document.

Format of the External File

The file itself is located in Solr's index directory, which by default is \$SOLR_HOME/data. The name of the file should be external_fieldname_ or external_fieldname_.*. For the example above, then, the file could be named external_entryRankFile or external_entryRankFile.txt.



If any files using the name pattern .* (such as .txt) appear, the last (after being sorted by name) will be used and previous versions will be deleted. This behavior supports implementations on systems where one may not be able to overwrite a file (for example, on Windows, if the file is in use).

The file contains entries that map a key field, on the left of the equals sign, to a value, on the right. Here are a few example entries:

```
doc33=1.414  
doc34=3.14159  
doc40=42
```

The keys listed in this file do not need to be unique. The file does not need to be sorted, but Solr will be able to perform the lookup faster if it is.

Reloading an External File

It's possible to define an event listener to reload an external file when either a searcher is reloaded or when a new searcher is started. See the section [Query-Related Listeners](#) for more information, but a sample definition in `solrconfig.xml` might look like this:

```
<listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloaded"/>  
<listener event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloaded"/>
```

The PreAnalyzedField Type

The PreAnalyzedField type provides a way to send to Solr serialized token streams, optionally with independent stored values of a field, and have this information stored and indexed without any additional text processing applied in Solr. This is useful if user wants to submit field content that was already processed by some existing external text processing pipeline (e.g., it has been tokenized, annotated, stemmed, synonyms inserted, etc.), while using all the rich attributes that Lucene's TokenStream provides (per-token attributes).

The serialization format is pluggable using implementations of PreAnalyzedParser interface. There are two out-of-the-box implementations:

- [JsonPreAnalyzedParser](#): as the name suggests, it parses content that uses JSON to represent field's content. This is the default parser to use if the field type is not configured otherwise.
- [SimplePreAnalyzedParser](#): uses a simple strict plain text format, which in some situations may be easier to create than JSON.

There is only one configuration parameter, `parserImpl`. The value of this parameter should be a fully qualified class name of a class that implements PreAnalyzedParser interface. The default value of this parameter is `org.apache.solr.schema.JsonPreAnalyzedParser`.

By default, the query-time analyzer for fields of this type will be the same as the index-time analyzer, which expects serialized pre-analyzed text. You must add a query type analyzer to your fieldType in order to perform analysis on non-pre-analyzed queries. In the example below, the index-time analyzer expects the default JSON serialization format, and the query-time analyzer will employ StandardTokenizer/LowerCaseFilter:

```
<fieldType name="pre_with_query_analyzer" class="solr.PreAnalyzedField">
  <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
  </analyzer>
</fieldType>
```

JsonPreAnalyzedParser

This is the default serialization format used by PreAnalyzedField type. It uses a top-level JSON map with the following keys:

Key	Description	Required
v	Version key. Currently the supported version is 1.	required
str	Stored string value of a field. You can use at most one of str or bin.	optional
bin	Stored binary value of a field. The binary value has to be Base64 encoded.	optional
tokens	serialized token stream. This is a JSON list.	optional

Any other top-level key is silently ignored.

Token Stream Serialization

The token stream is expressed as a JSON list of JSON maps. The map for each token consists of the following keys and values:

Key	Description	Lucene Attribute	Value	Required?
t	token	CharTermAttribute	UTF-8 string representing the current token	required
s	start offset	OffsetAttribute	Non-negative integer	optional
e	end offset	OffsetAttribute	Non-negative integer	optional
i	position increment	PositionIncrementAttribute	Non-negative integer - default is 1	optional
p	payload	PayloadAttribute	Base64 encoded payload	optional
y	lexical type	TypeAttribute	UTF-8 string	optional
f	flags	FlagsAttribute	String representing an integer value in hexadecimal format	optional

Any other key is silently ignored.

JsonPreAnalyzedParser Example

```
{  
  "v": "1",  
  "str": "test ąćęłńóśżż",  
  "tokens": [  
    {"t": "two", "s": 5, "e": 8, "i": 1, "y": "word"},  
    {"t": "three", "s": 20, "e": 22, "i": 1, "y": "foobar"},  
    {"t": "one", "s": 123, "e": 128, "i": 22, "p": "DQ4KDQsODg8=", "y": "word"}  
  ]  
}
```

SimplePreAnalyzedParser

The fully qualified class name to use when specifying this format via the parserImpl configuration parameter is org.apache.solr.schema.SimplePreAnalyzedParser.

SimplePreAnalyzedParser Syntax

The serialization format supported by this parser is as follows:

Serialization format

```
content ::= version (stored)? tokens  
version ::= digit+ "  
; stored field value - any "=" inside must be escaped!  
stored ::= "=" text "="  
tokens ::= (token (( " ") + token)*)*  
token ::= text ("," attrib)*  
attrib ::= name '=' value  
name ::= text  
value ::= text
```

Special characters in "text" values can be escaped using the escape character \. The following escape sequences are recognized:

EscapeSequence	Description
\	literal space character
\,	literal , character
\=	literal = character
\\	literal \ character
\n	newline
\r	carriage return
\t	horizontal tab

Please note that Unicode sequences (e.g., \u0001) are not supported.

Supported Attributes

The following token attributes are supported, and identified with short symbolic names:

Name	Description	Lucene attribute	Value format
i	position increment	PositionIncrementAttribute	integer
s	start offset	OffsetAttribute	integer
e	end offset	OffsetAttribute	integer
y	lexical type	TypeAttribute	string
f	flags	FlagsAttribute	hexadecimal integer
p	payload	PayloadAttribute	bytes in hexadecimal format; whitespace is ignored

Token positions are tracked and implicitly added to the token stream - the start and end offsets consider only the term text and whitespace, and exclude the space taken by token attributes.

Example Token Streams

```
1 one two three
```

- version: 1
- stored: null
- token: (term=one,startOffset=0,endOffset=3)
- token: (term=two,startOffset=4,endOffset=7)
- token: (term=three,startOffset=8,endOffset=13)

```
1 one two three
```

- version: 1
- stored: null
- token: (term=one,startOffset=0,endOffset=3)
- token: (term=two,startOffset=5,endOffset=8)
- token: (term=three,startOffset=11,endOffset=16)

```
1 one,s=123,e=128,i=22 two three,s=20,e=22
```

- version: 1
- stored: null
- token: (term=one,positionIncrement=22,startOffset=123,endOffset=128)
- token: (term=two,positionIncrement=1,startOffset=5,endOffset=8)

- token: (term=three,positionIncrement=1,startOffset=20,endOffset=22)

```
1 \ one\ \,,i=22,a=\_, two\=\\n,\ =\\ \\
```

- version: 1
- stored: null
- token: (term=one ,,positionIncrement=22,startOffset=0,endOffset=6)
- token: (term=two= ,positionIncrement=1,startOffset=7,endOffset=15)
- token: (term=_,positionIncrement=1,startOffset=17,endOffset=18)

Note that unknown attributes and their values are ignored, so in this example, the "a" attribute on the first token and the " " (escaped space) attribute on the second token are ignored, along with their values, because they are not among the supported attribute names.

```
1 ,i=22 ,i=33,s=2,e=20 ,
```

- version: 1
- stored: null
- token: (term=,positionIncrement=22,startOffset=0,endOffset=0)
- token: (term=,positionIncrement=33,startOffset=2,endOffset=20)
- token: (term=,positionIncrement=1,startOffset=2,endOffset=2)

```
1 =This is the stored part with \=\\n \t escapes.=one two three
```

- version: 1
- stored: This is the stored part with = \t escapes.
- token: (term=one,startOffset=0,endOffset=3)
- token: (term=two,startOffset=4,endOffset=7)
- token: (term=three,startOffset=8,endOffset=13)

Note that the \t in the above stored value is not literal; it's shown that way to visually indicate the actual tab char that is in the stored value.

```
1 ==
```

- version: 1
- stored: ""
- (no tokens)

```
1 =this is a test.=
```

- version: 1
- stored: this is a test.
- (no tokens)

Field Properties by Use Case

Here is a summary of common use cases, and the attributes the fields or field types should have to support the case. An entry of true or false in the table indicates that the option must be set to the given value for the use case to function correctly. If no entry is provided, the setting of that attribute has no impact on the case.

Use Case	indexed	stored	multiValue d	omitNorm s	termVecto rs	termPositi ons	docValues
search within field	true						
retrieve contents		true ⁸					true ⁸
use as unique key	true		false				
sort on field	true ⁷		false ⁹	true ¹			true ⁷
highlighting	true ⁴	true			true ²	true ³	
faceting ⁵	true ⁷						true ⁷
add multiple values, maintaining order			true				
field length affects doc score				false			
MoreLikeThis ⁵					true ⁶		

Notes:

1. Recommended but not necessary.
2. Will be used if present, but not necessary.
3. (if termVectors=true)
4. A tokenizer must be defined for the field, but it doesn't need to be indexed.
5. Described in [Understanding Analyzers, Tokenizers, and Filters](#).
6. Term vectors are not mandatory here. If not true, then a stored field is analyzed. So term vectors are

recommended, but only required if stored=false.

7. For most field types, either indexed or docValues must be true, but both are not required. [DocValues](#) can be more efficient in many cases. For [Int/Long/Float/Double/Date]PointFields, docValues=true is required.
8. Stored content will be used by default, but docValues can alternatively be used. See [DocValues](#).
9. Multi-valued sorting may be performed on docValues-enabled fields using the two-argument field() function, e.g., field(myfield,min); see the [field\(\) function in Function Queries](#).

Defining Fields

Fields are defined in the fields element of schema.xml. Once you have the field types set up, defining the fields themselves is simple.

Example Field Definition

The following example defines a field named price with a type named float and a default value of 0.0; the indexed and stored properties are explicitly set to true, while any other properties specified on the float field type are inherited.

```
<field name="price" type="float" default="0.0" indexed="true" stored="true"/>
```

Field Properties

Field definitions can have the following properties:

name

The name of the field. Field names should consist of alphanumeric or underscore characters only and not start with a digit. This is not currently strictly enforced, but other field names will not have first class support from all components and back compatibility is not guaranteed. Names with both leading and trailing underscores (e.g., _version_) are reserved. Every field must have a name.

type

The name of the fieldType for this field. This will be found in the name attribute on the fieldType definition. Every field must have a type.

default

A default value that will be added automatically to any document that does not have a value in this field when it is indexed. If this property is not specified, there is no default.

Optional Field Type Override Properties

Fields can have many of the same properties as field types. Properties from the table below which are specified on an individual field will override any explicit value for that property specified on the the fieldType of the field, or any implicit default property value provided by the underlying fieldType implementation. The table below is reproduced from [Field Type Definitions and Properties](#), which has more details:

Property	Description	Values	Implicit Default
indexed	If true, the value of the field can be used in queries to retrieve matching documents.	true or false	true
stored	If true, the actual value of the field can be retrieved by queries.	true or false	true

Property	Description	Values	Implicit Default
docValues	If true, the value of the field will be put in a column-oriented DocValues structure.	true or false	false
sortMissingFirst sortMissingLast	Control the placement of documents when a sort field is not present.	true or false	false
multiValued	If true, indicates that a single document might contain multiple values for this field type.	true or false	false
uninvertible	If true, indicates that an indexed="true" docValues="false" field can be "un-inverted" at query time to build up large in memory data structure to serve in place of DocValues . Defaults to true for historical reasons, but users are strongly encouraged to set this to false for stability and use docValues="true" as needed.	true or false	true
omitNorms	If true, omits the norms associated with this field (this disables length normalization for the field, and saves some memory). Defaults to true for all primitive (non-analyzed) field types, such as int, float, data, bool, and string. Only full-text fields or fields need norms.	true or false	*
omitTermFreqAndPositions	If true, omits term frequency, positions, and payloads from postings for this field. This can be a performance boost for fields that don't require that information. It also reduces the storage space required for the index. Queries that rely on position that are issued on a field with this option will silently fail to find documents. This property defaults to true for all field types that are not text fields.	true or false	*
omitPositions	Similar to omitTermFreqAndPositions but preserves term frequency information.	true or false	*

Property	Description	Values	Implicit Default
termVectors termPositions termOffsets termPayloads	These options instruct Solr to maintain full term vectors for each document, optionally including position, offset and payload information for each term occurrence in those vectors. These can be used to accelerate highlighting and other ancillary functionality, but impose a substantial cost in terms of index size. They are not necessary for typical uses of Solr.	true or false	false
required	Instructs Solr to reject any attempts to add a document which does not have a value for this field. This property defaults to false.	true or false	false
useDocValuesAsStored	If the field has docValues enabled, setting this to true would allow the field to be returned as if it were a stored field (even if it has stored=false) when matching "*" in an fl parameter .	true or false	true
large	Large fields are always lazy loaded and will only take up space in the document cache if the actual value is < 512KB. This option requires stored="true" and multiValued="false". It's intended for fields that might have very large values so that they don't get cached in memory.	true or false	false

Copying Fields

You might want to interpret some document fields in more than one way. Solr has a mechanism for making copies of fields so that you can apply several distinct field types to a single piece of incoming information.

The name of the field you want to copy is the *source*, and the name of the copy is the *destination*. In schema.xml, it's very simple to make copies of fields:

```
<copyField source="cat" dest="text" maxChars="30000" />
```

In this example, we want Solr to copy the cat field to a field named text. Fields are copied before analysis is done, meaning you can have two fields with identical original content, but which use different analysis chains and are stored in the index differently.

In the example above, if the text destination field has data of its own in the input documents, the contents of the cat field will be added as additional values – just as if all of the values had originally been specified by the client. Remember to configure your fields as multivalued="true" if they will ultimately get multiple values (either from a multivalued source or from multiple copyField directives).

A common usage for this functionality is to create a single "search" field that will serve as the default query field when users or clients do not specify a field to query. For example, title, author, keywords, and body may all be fields that should be searched by default, with copy field rules for each field to copy to a catchall field (for example, it could be named anything). Later you can set a rule in solrconfig.xml to search the catchall field by default. One caveat to this is your index will grow when using copy fields. However, whether this becomes problematic for you and the final size will depend on the number of fields being copied, the number of destination fields being copied to, the analysis in use, and the available disk space.

The maxChars parameter, an int parameter, establishes an upper limit for the number of characters to be copied from the source value when constructing the value added to the destination field. This limit is useful for situations in which you want to copy some data from the source field, but also control the size of index files.

Both the source and the destination of copyField can contain either leading or trailing asterisks, which will match anything. For example, the following line will copy the contents of all incoming fields that match the wildcard pattern *_t to the text field.:

```
<copyField source="*_t" dest="text" maxChars="25000" />
```



The copyField command can use a wildcard (*) character in the dest parameter only if the source parameter contains one as well. copyField uses the matching glob from the source field for the dest field name into which the source content is copied.

Copying is done at the stream source level and no copy feeds into another copy. This means that copy fields cannot be chained i.e., you cannot copy from here to there and then from there to elsewhere. However, the same source field can be copied to multiple destination fields:

```
<copyField source="here" dest="there"/>
<copyField source="here" dest="elsewhere"/>
```

Dynamic Fields

Dynamic fields allow Solr to index fields that you did not explicitly define in your schema.

This is useful if you discover you have forgotten to define one or more fields. Dynamic fields can make your application less brittle by providing some flexibility in the documents you can add to Solr.

A dynamic field is just like a regular field except it has a name with a wildcard in it. When you are indexing documents, a field that does not match any explicitly defined fields can be matched with a dynamic field.

For example, suppose your schema includes a dynamic field with a name of `*_i`. If you attempt to index a document with a `cost_i` field, but no explicit `cost_i` field is defined in the schema, then the `cost_i` field will have the field type and analysis defined for `*_i`.

Like regular fields, dynamic fields have a name, a field type, and options.

```
<dynamicField name="*_i" type="int" indexed="true" stored="true"/>
```

It is recommended that you include basic dynamic field mappings (like that shown above) in your `schema.xml`. The mappings can be very useful.

Other Schema Elements

This section describes several other important elements of schema.xml not covered in earlier sections.

Unique Key

The uniqueKey element specifies which field is a unique identifier for documents. Although uniqueKey is not required, it is nearly always warranted by your application design. For example, uniqueKey should be used if you will ever update a document in the index.

You can define the unique key field by naming it:

```
<uniqueKey>id</uniqueKey>
```

Schema defaults and copyFields cannot be used to populate the uniqueKey field. The fieldType of uniqueKey must not be analyzed and must not be any of the *PointField types. You can use UUIDUpdateProcessorFactory to have uniqueKey values generated automatically.

Further, the operation will fail if the uniqueKey field is used, but is multivalued (or inherits the multivalence-ness from the fieldtype). However, uniqueKey will continue to work, as long as the field is properly used.

Similarity

Similarity is a Lucene class used to score a document in searching.

Each collection has one "global" Similarity, and by default Solr uses an implicit SchemaSimilarityFactory which allows individual field types to be configured with a "per-type" specific Similarity and implicitly uses BM25Similarity for any field type which does not have an explicit Similarity.

This default behavior can be overridden by declaring a top level <similarity/> element in your schema.xml, outside of any single field type. This similarity declaration can either refer directly to the name of a class with a no-argument constructor, such as in this example showing BM25Similarity:

```
<similarity class="solr.BM25SimilarityFactory"/>
```

or by referencing a SimilarityFactory implementation, which may take optional initialization parameters:

```
<similarity class="solr.DFRSimilarityFactory">
  <str name="basicModel">P</str>
  <str name="afterEffect">L</str>
  <str name="normalization">H2</str>
  <float name="c">7</float>
</similarity>
```

In most cases, specifying global level similarity like this will cause an error if your schema.xml also includes field type specific <similarity/> declarations. One key exception to this is that you may explicitly declare a SchemaSimilarityFactory and specify what that default behavior will be for all field types that do not

declare an explicit Similarity using the name of field type (specified by defaultSimFromFieldType) that is configured with a specific similarity:

```
<similarity class="solr.SchemaSimilarityFactory">
  <str name="defaultSimFromFieldType">text_dfr</str>
</similarity>
<fieldType name="text_dfr" class="solr.TextField">
  <analyzer ... />
  <similarity class="solr.DFRSimilarityFactory">
    <str name="basicModel">I(F)</str>
    <str name="afterEffect">B</str>
    <str name="normalization">H3</str>
    <float name="mu">900</float>
  </similarity>
</fieldType>
<fieldType name="text_ib" class="solr.TextField">
  <analyzer ... />
  <similarity class="solr.IBSimilarityFactory">
    <str name="distribution">SPL</str>
    <str name="lambda">DF</str>
    <str name="normalization">H2</str>
  </similarity>
</fieldType>
<fieldType name="text_other" class="solr.TextField">
  <analyzer ... />
</fieldType>
```

In the example above IBSimilarityFactory (using the Information-Based model) will be used for any fields of type text_ib, while DFRSimilarityFactory (divergence from random) will be used for any fields of type text_dfr, as well as any fields using a type that does not explicitly specify a `<similarity/>`.

If SchemaSimilarityFactory is explicitly declared without configuring a defaultSimFromFieldType, then BM25Similarity is implicitly used as the default for luceneMatchVersion $\geq 8.0.0$ and otherwise LegacyBM25Similarity is used to mimic the same BM25 formula that was the default in those versions.

In addition to the various factories mentioned on this page, there are several other similarity implementations that can be used such as the SweetSpotSimilarityFactory, ClassicSimilarityFactory, LegacyBM25SimilarityFactory etc. For details, see the Solr Javadocs for the [similarity factories](#).

Schema API

The Schema API allows you to use an HTTP API to manage many of the elements of your schema.

The Schema API utilizes the ManagedIndexSchemaFactory class, which is the default schema factory in modern Solr versions. See the section [Schema Factory Definition in SolrConfig](#) for more information about choosing a schema factory for your index.

This API provides read and write access to the Solr schema for each collection (or core, when using standalone Solr). Read access to all schema elements is supported. Fields, dynamic fields, field types and copyField rules may be added, removed or replaced. Future Solr releases will extend write access to allow more schema elements to be modified.

Why is hand editing of the managed schema discouraged?

The file named "managed-schema" in the example configurations may include a note that recommends never hand-editing the file. Before the Schema API existed, such edits were the only way to make changes to the schema, and users may have a strong desire to continue making changes this way.



The reason that this is discouraged is because hand-edits of the schema may be lost if the Schema API described here is later used to make a change, unless the core or collection is reloaded or Solr is restarted before using the Schema API. If care is taken to always reload or restart after a manual edit, then there is no problem at all with doing those edits.

The API allows two output modes for all calls: JSON or XML. When requesting the complete schema, there is another output mode which is XML modeled after the managed-schema file itself, which is in XML format.

When modifying the schema with the API, a core reload will automatically occur in order for the changes to be available immediately for documents indexed thereafter. Previously indexed documents will **not** be automatically updated - they **must** be reindexed if existing index data uses schema elements that you changed.

Reindex after schema modifications!

If you modify your schema, you will likely need to reindex all documents. If you do not, you may lose access to documents, or not be able to interpret them properly, e.g., after replacing a field type.



Modifying your schema will never modify any documents that are already indexed. You must reindex documents in order to apply schema changes to them. Queries and updates made after the change may encounter errors that were not present before the change. Completely deleting the index and rebuilding it is usually the only option to fix such errors.

See the section [Reindexing](#) for more information about reindexing.

Modify the Schema

To add, remove or replace fields, dynamic field rules, copy field rules, or new field types, you can send a POST request to the /collection/schema/ endpoint with a sequence of commands in JSON format to

perform the requested actions. The following commands are supported:

- add-field: add a new field with parameters you provide.
- delete-field: delete a field.
- replace-field: replace an existing field with one that is differently configured.
- add-dynamic-field: add a new dynamic field rule with parameters you provide.
- delete-dynamic-field: delete a dynamic field rule.
- replace-dynamic-field: replace an existing dynamic field rule with one that is differently configured.
- add-field-type: add a new field type with parameters you provide.
- delete-field-type: delete a field type.
- replace-field-type: replace an existing field type with one that is differently configured.
- add-copy-field: add a new copy field rule.
- delete-copy-field: delete a copy field rule.

These commands can be issued in separate POST requests or in the same POST request. Commands are executed in the order in which they are specified.

In each case, the response will include the status and the time to process the request, but will not include the entire schema.

When modifying the schema with the API, a core reload will automatically occur in order for the changes to be available immediately for documents indexed thereafter. Previously indexed documents will **not** be automatically handled - they **must** be reindexed if they used schema elements that you changed.

Add a New Field

The add-field command adds a new field definition to your schema. If a field with the same name exists an error is thrown.

All of the properties available when defining a field with manual schema.xml edits can be passed via the API. These request attributes are described in detail in the section [Defining Fields](#).

For example, to define a new stored field named "sell_by", of type "pdate", you would POST the following request:

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "add-field":{  
    "name":"sell_by",  
    "type":"pdate",  
    "stored":true  
  }' http://localhost:8983/solr/gettingstarted/schema
```

V2 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "add-field":{  
    "name":"sell_by",  
    "type":"pdate",  
    "stored":true }  
' http://localhost:8983/api/cores/gettingstarted/schema
```

Delete a Field

The delete-field command removes a field definition from your schema. If the field does not exist in the schema, or if the field is the source or destination of a copy field rule, an error is thrown.

For example, to delete a field named "sell_by", you would POST the following request:

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "delete-field" : { "name":"sell_by" }  
' http://localhost:8983/solr/gettingstarted/schema
```

V2 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "delete-field" : { "name":"sell_by" }  
' http://localhost:8983/api/cores/gettingstarted/schema
```

Replace a Field

The replace-field command replaces a field's definition. Note that you must supply the full definition for a field - this command will **not** partially modify a field's definition. If the field does not exist in the schema an error is thrown.

All of the properties available when defining a field with manual schema.xml edits can be passed via the API. These request attributes are described in detail in the section [Defining Fields](#).

For example, to replace the definition of an existing field "sell_by", to make it be of type "date" and to not be stored, you would POST the following request:

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "replace-field":{  
    "name":"sell_by",  
    "type":"date",  
    "stored":false }  
}' http://localhost:8983/solr/gettingstarted/schema
```

V2 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "replace-field":{  
    "name":"sell_by",  
    "type":"date",  
    "stored":false }  
}' http://localhost:8983/api/cores/gettingstarted/schema
```

Add a Dynamic Field Rule

The add-dynamic-field command adds a new dynamic field rule to your schema.

All of the properties available when editing schema.xml can be passed with the POST request. The section [Dynamic Fields](#) has details on all of the attributes that can be defined for a dynamic field rule.

For example, to create a new dynamic field rule where all incoming fields ending with "_s" would be stored and have field type "string", you can POST a request like this:

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "add-dynamic-field":{  
    "name": "*_s",  
    "type": "string",  
    "stored": true }  
}' http://localhost:8983/solr/gettingstarted/schema
```

V2 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{
  "add-dynamic-field":{
    "name": "*_s",
    "type": "string",
    "stored": true
  }
}' http://localhost:8983/api/cores/gettingstarted/schema
```

Delete a Dynamic Field Rule

The delete-dynamic-field command deletes a dynamic field rule from your schema. If the dynamic field rule does not exist in the schema, or if the schema contains a copy field rule with a target or destination that matches only this dynamic field rule, an error is thrown.

For example, to delete a dynamic field rule matching "`*_s`", you can POST a request like this:

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{
  "delete-dynamic-field":{ "name": "*_s" }
}' http://localhost:8983/solr/gettingstarted/schema
```

V2 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{
  "delete-dynamic-field":{ "name": "*_s" }
}' http://localhost:8983/api/cores/gettingstarted/schema
```

Replace a Dynamic Field Rule

The replace-dynamic-field command replaces a dynamic field rule in your schema. Note that you must supply the full definition for a dynamic field rule - this command will **not** partially modify a dynamic field rule's definition. If the dynamic field rule does not exist in the schema an error is thrown.

All of the properties available when editing `schema.xml` can be passed with the POST request. The section [Dynamic Fields](#) has details on all of the attributes that can be defined for a dynamic field rule.

For example, to replace the definition of the "`*_s`" dynamic field rule with one where the field type is "`text_general`" and it's not stored, you can POST a request like this:

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "replace-dynamic-field":{  
    "name": "*_s",  
    "type": "text_general",  
    "stored":false }  
}' http://localhost:8983/solr/gettingstarted/schema
```

V2 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "replace-dynamic-field":{  
    "name": "*_s",  
    "type": "text_general",  
    "stored":false }  
}' http://localhost:8983/solr/gettingstarted/schema
```

Add a New Field Type

The add-field-type command adds a new field type to your schema.

All of the field type properties available when editing schema.xml by hand are available for use in a POST request. The structure of the command is a JSON mapping of the standard field type definition, including the name, class, index and query analyzer definitions, etc. Details of all of the available options are described in the section [Solr Field Types](#).

For example, to create a new field type named "myNewTxtField", you can POST a request as follows:

V1 API with Single Analysis

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "add-field-type" : {  
    "name": "myNewTxtField",  
    "class": "solr.TextField",  
    "positionIncrementGap": "100",  
    "analyzer" : {  
      "charFilters": [ {  
          "class": "solr.PatternReplaceCharFilterFactory",  
          "replacement": "$1$1",  
          "pattern": "([a-zA-Z])\\\\\\1+" } ],  
      "tokenizer": {  
        "class": "solrWhitespaceTokenizerFactory" },  
      "filters": [ {  
          "class": "solr.WordDelimiterFilterFactory",  
          "preserveOriginal": "0" } ] } } }  
' http://localhost:8983/solr/gettingstarted/schema
```

Note in this example that we have only defined a single analyzer section that will apply to index analysis and query analysis.

V1 API with Two Analyzers

If we wanted to define separate analysis, we would replace the analyzer section in the above example with separate sections for `indexAnalyzer` and `queryAnalyzer`. As in this example:

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "add-field-type": {  
    "name": "myNewTextField",  
    "class": "solr.TextField",  
    "indexAnalyzer": {  
      "tokenizer": {  
        "class": "solr.PathHierarchyTokenizerFactory",  
        "delimiter": "/" } },  
    "queryAnalyzer": {  
      "tokenizer": {  
        "class": "solr.KeywordTokenizerFactory" } } } }  
' http://localhost:8983/solr/gettingstarted/schema
```

V2 API with Two Analyzers

To define two analyzers with the V2 API, we just use a different endpoint:

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
    "add-field-type":{  
        "name":"myNewTextField",  
        "class":"solr.TextField",  
        "indexAnalyzer":{  
            "tokenizer":{  
                "class":"solr.PathHierarchyTokenizerFactory",  
                "delimiter":"/" }},  
        "queryAnalyzer":{  
            "tokenizer":{  
                "class":"solr.KeywordTokenizerFactory" }}}}  
' http://localhost:8983/api/cores/gettingstarted/schema
```

Delete a Field Type

The delete-field-type command removes a field type from your schema. If the field type does not exist in the schema, or if any field or dynamic field rule in the schema uses the field type, an error is thrown.

For example, to delete the field type named "myNewTxtField", you can make a POST request as follows:

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
    "delete-field-type":{ "name":"myNewTxtField" }  
' http://localhost:8983/solr/gettingstarted/schema
```

V2 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
    "delete-field-type":{ "name":"myNewTxtField" }  
' http://localhost:8983/api/cores/gettingstarted/schema
```

Replace a Field Type

The replace-field-type command replaces a field type in your schema. Note that you must supply the full definition for a field type - this command will **not** partially modify a field type's definition. If the field type does not exist in the schema an error is thrown.

All of the field type properties available when editing schema.xml by hand are available for use in a POST request. The structure of the command is a JSON mapping of the standard field type definition, including the

name, class, index and query analyzer definitions, etc. Details of all of the available options are described in the section [Solr Field Types](#).

For example, to replace the definition of a field type named "myNewTxtField", you can make a POST request as follows:

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
    "replace-field-type":{  
        "name":"myNewTxtField",  
        "class":"solr.TextField",  
        "positionIncrementGap":"100",  
        "analyzer":{  
            "tokenizer":{  
                "class":"solr.StandardTokenizerFactory" }}}}  
' http://localhost:8983/solr/gettingstarted/schema
```

V2 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
    "replace-field-type":{  
        "name":"myNewTxtField",  
        "class":"solr.TextField",  
        "positionIncrementGap":"100",  
        "analyzer":{  
            "tokenizer":{  
                "class":"solr.StandardTokenizerFactory" }}}}  
' http://localhost:8983/api/cores/gettingstarted/schema
```

Add a New Copy Field Rule

The add-copy-field command adds a new copy field rule to your schema.

The attributes supported by the command are the same as when creating copy field rules by manually editing the schema.xml, as below:

source

The source field. This parameter is required.

dest

A field or an array of fields to which the source field will be copied. This parameter is required.

maxChars

The upper limit for the number of characters to be copied. The section [Copying Fields](#) has more details.

For example, to define a rule to copy the field "shelf" to the "location" and "catchall" fields, you would POST

the following request:

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
    "add-copy-field":{  
        "source":"shelf",  
        "dest":[ "location", "catchall" ]}  
}' http://localhost:8983/solr/gettingstarted/schema
```

V2 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
    "add-copy-field":{  
        "source":"shelf",  
        "dest":[ "location", "catchall" ]}  
}' http://localhost:8983/api/cores/gettingstarted/schema
```

Delete a Copy Field Rule

The delete-copy-field command deletes a copy field rule from your schema. If the copy field rule does not exist in the schema an error is thrown.

The source and dest attributes are required by this command.

For example, to delete a rule to copy the field "shelf" to the "location" field, you would POST the following request:

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
    "delete-copy-field":{ "source":"shelf", "dest":"location" }  
}' http://localhost:8983/solr/gettingstarted/schema
```

V1 API

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
    "delete-copy-field":{ "source":"shelf", "dest":"location" }  
}' http://localhost:8983/api/cores/gettingstarted/schema
```

Multiple Commands in a Single POST

It is possible to perform one or more add requests in a single command. The API is transactional and all commands in a single call either succeed or fail together.

The commands are executed in the order in which they are specified. This means that if you want to create a new field type and in the same request use the field type on a new field, the section of the request that creates the field type must come before the section that creates the new field. Similarly, since a field must exist for it to be used in a copy field rule, a request to add a field must come before a request for the field to be used as either the source or the destination for a copy field rule.

The syntax for making multiple requests supports several approaches. First, the commands can simply be made serially, as in this request to create a new field type and then a field that uses that type:

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "add-field-type":{  
    "name":"myNewTxtField",  
    "class":"solr.TextField",  
    "positionIncrementGap":"100",  
    "analyzer":{  
      "charFilters":[]  
        "class":"solr.PatternReplaceCharFilterFactory",  
        "replacement":"$1$1",  
        "pattern":"(\\w+\\W*)\\1+" }],  
      "tokenizer":{  
        "class":"solr.WhitespaceTokenizerFactory" },  
      "filters":[]  
        "class":"solr.WordDelimiterFilterFactory",  
        "preserveOriginal":"0" }]}},  
  "add-field" : {  
    "name":"sell_by",  
    "type":"myNewTxtField",  
    "stored":true }  
' http://localhost:8983/solr/gettingstarted/schema
```

Or, the same command can be repeated, as in this example:

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "add-field":{  
    "name":"shelf",  
    "type":"myNewTxtField",  
    "stored":true },  
  "add-field":{  
    "name":"location",  
    "type":"myNewTxtField",  
    "stored":true },  
  "add-copy-field":{  
    "source":"shelf",  
    "dest":[ "location", "catchall" ]}  
}' http://localhost:8983/solr/gettingstarted/schema
```

Finally, repeated commands can be sent as an array:

```
curl -X POST -H 'Content-type:application/json' --data-binary '{  
  "add-field": [  
    { "name":"shelf",  
      "type":"myNewTxtField",  
      "stored":true },  
    { "name":"location",  
      "type":"myNewTxtField",  
      "stored":true }]  
}' http://localhost:8983/solr/gettingstarted/schema
```

Schema Changes among Replicas

When running in SolrCloud mode, changes made to the schema on one node will propagate to all replicas in the collection.

You can pass the updateTimeoutSecs parameter with your request to set the number of seconds to wait until all replicas confirm they applied the schema updates. This helps your client application be more robust in that you can be sure that all replicas have a given schema change within a defined amount of time.

If agreement is not reached by all replicas in the specified time, then the request fails and the error message will include information about which replicas had trouble. In most cases, the only option is to re-try the change after waiting a brief amount of time. If the problem persists, then you'll likely need to investigate the server logs on the replicas that had trouble applying the changes.

If you do not supply an updateTimeoutSecs parameter, the default behavior is for the receiving node to return immediately after persisting the updates to ZooKeeper. All other replicas will apply the updates asynchronously. Consequently, without supplying a timeout, your client application cannot be sure that all replicas have applied the changes.

Retrieve Schema Information

The following endpoints allow you to read how your schema has been defined. You can GET the entire schema, or only portions of it as needed.

To modify the schema, see the previous section [Modify the Schema](#).

Retrieve the Entire Schema

GET /collection/schema

Retrieve Schema Parameters

Path Parameters

collection

The collection (or core) name.

Query Parameters

The query parameters should be added to the API request after '?'.
wt

Defines the format of the response. The options are **json**, **xml** or **schema.xml**. If not specified, JSON will be returned by default.

Retrieve Schema Response

Output Content

The output will include all fields, field types, dynamic rules and copy field rules, in the format requested (JSON or XML). The schema name and version are also included.

Retrieve Schema Examples

Get the entire schema in JSON.

```
curl http://localhost:8983/solr/gettingstarted/schema
```

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 5
  },
  "schema": {
    "name": "example",
    "version": 1.5,
    "uniqueKey": "id",
    "fieldTypes": [
      {
        "name": "alphaOnlySort",
        "class": "solr.TextField",
        "sortMissingLast": true,
        "omitNorms": true,
        "analyzer": {
          "tokenizer": {
            "class": "solr.KeywordTokenizerFactory"
          },
          "filters": []
        }
      }
    ]
  }
}
```

```
    "class": "solr.LowerCaseFilterFactory"},  
    {  
        "class": "solr.TrimFilterFactory"},  
        {  
            "class": "solr.PatternReplaceFilterFactory",  
            "replace": "all",  
            "replacement": "",  
            "pattern": "([^\u00e1-\u00f1])"}]}],  
"fields": [{"  
    "name": "_version_",  
    "type": "long",  
    "indexed": true,  
    "stored": true},  
    {  
        "name": "author",  
        "type": "text_general",  
        "indexed": true,  
        "stored": true},  
        {  
            "name": "cat",  
            "type": "string",  
            "multiValued": true,  
            "indexed": true,  
            "stored": true}],  
"copyFields": [{"  
    "source": "author",  
    "dest": "text"},  
    {  
        "source": "cat",  
        "dest": "text"},  
        {  
            "source": "content",  
            "dest": "text"},  
            {  
                "source": "author",  
                "dest": "author_s"}]}]
```

Get the entire schema in XML.

```
curl http://localhost:8983/solr/gettingstarted/schema?wt=xml
```

```
<response>
<lst name="responseHeader">
  <int name="status">0</int>
  <int name="QTime">5</int>
</lst>
<lst name="schema">
  <str name="name">example</str>
  <float name="version">1.5</float>
  <str name="uniqueKey">id</str>
  <arr name="fieldTypes">
    <lst>
      <str name="name">alphaOnlySort</str>
      <str name="class">solr.TextField</str>
      <bool name="sortMissingLast">true</bool>
      <bool name="omitNorms">true</bool>
      <lst name="analyzer">
        <lst name="tokenizer">
          <str name="class">solr.KeywordTokenizerFactory</str>
        </lst>
        <arr name="filters">
          <lst>
            <str name="class">solr.LowerCaseFilterFactory</str>
          </lst>
          <lst>
            <str name="class">solr.TrimFilterFactory</str>
          </lst>
          <lst>
            <str name="class">solr.PatternReplaceFilterFactory</str>
            <str name="replace">all</str>
            <str name="replacement"/>
            <str name="pattern">([^\w])</str>
          </lst>
        </arr>
      </lst>
    </lst>
  ...
  <lst>
    <str name="source">author</str>
    <str name="dest">author_s</str>
  </lst>
</arr>
</lst>
</response>
```

Get the entire schema in "schema.xml" format.

```
curl http://localhost:8983/solr/gettingstarted/schema?wt=schema.xml
```

```
<schema name="example" version="1.5">
  <uniqueKey>id</uniqueKey>
  <types>
    <fieldType name="alphaOnlySort" class="solr.TextField" sortMissingLast="true" omitNorms="true">
      <analyzer>
        <tokenizer class="solr.KeywordTokenizerFactory"/>
        <filter class="solr.LowerCaseFilterFactory"/>
        <filter class="solr.TrimFilterFactory"/>
        <filter class="solr.PatternReplaceFilterFactory" replace="all" replacement="" pattern="([^\w])"/>
      </analyzer>
    </fieldType>
  ...
  <copyField source="url" dest="text"/>
  <copyField source="price" dest="price_c"/>
  <copyField source="author" dest="author_s"/>
</schema>
```

List Fields

GET /collection/schema/fields

GET /collection/schema/fields/fieldname

List Fields Parameters

Path Parameters

collection

The collection (or core) name.

fieldname

The specific fieldname (if limiting the request to a single field).

Query Parameters

The query parameters can be added to the API request after a '?'.

wt

Defines the format of the response. The options are json or xml. If not specified, JSON will be returned by default.

f1

Comma- or space-separated list of one or more fields to return. If not specified, all fields will be returned by default.

includeDynamic

If true, and if the f1 query parameter is specified or the fieldname path parameter is used, matching dynamic fields are included in the response and identified with the dynamicBase property.

If neither the f1 query parameter nor the fieldname path parameter is specified, the includeDynamic

query parameter is ignored.

If false, the default, matching dynamic fields will not be returned.

showDefaults

If true, all default field properties from each field's field type will be included in the response (e.g., tokenized for solr.TextField). If false, the default, only explicitly specified field properties will be included.

List Fields Response

The output will include each field and any defined configuration for each field. The defined configuration can vary for each field, but will minimally include the field name, the type, if it is indexed and if it is stored.

If multiValued is defined as either true or false (most likely true), that will also be shown. See the section [Defining Fields](#) for more information about each parameter.

List Fields Examples

Get a list of all fields.

```
curl http://localhost:8983/solr/gettingstarted/schema/fields
```

The sample output below has been truncated to only show a few fields.

```
{  
  "fields": [  
    {  
      "indexed": true,  
      "name": "_version_",  
      "stored": true,  
      "type": "long"  
    },  
    {  
      "indexed": true,  
      "name": "author",  
      "stored": true,  
      "type": "text_general"  
    },  
    {  
      "indexed": true,  
      "multiValued": true,  
      "name": "cat",  
      "stored": true,  
      "type": "string"  
    },  
    ...  
  ],  
  "responseHeader": {  
    "QTime": 1,  
    "status": 0  
  }  
}
```

List Dynamic Fields

GET /collection/schema/dynamicfields
GET /collection/schema/dynamicfields/name

List Dynamic Field Parameters

Path Parameters

collection

The collection (or core) name.

name

The name of the dynamic field rule (if limiting request to a single dynamic field rule).

Query Parameters

The query parameters can be added to the API request after a '?'.

wt

Defines the format of the response. The options are json or xml. If not specified, JSON will be returned by

default.

showDefaults

If true, all default field properties from each dynamic field's field type will be included in the response (e.g., tokenized for solr.TextField). If false, the default, only explicitly specified field properties will be included.

List Dynamic Field Response

The output will include each dynamic field rule and the defined configuration for each rule. The defined configuration can vary for each rule, but will minimally include the dynamic field name, the type, if it is indexed and if it is stored. See the section [Dynamic Fields](#) for more information about each parameter.

List Dynamic Field Examples

Get a list of all dynamic field declarations:

```
curl http://localhost:8983/solr/gettingstarted/schema/dynamicfields
```

The sample output below has been truncated.

```
{  
  "dynamicFields": [  
    {  
      "indexed": true,  
      "name": "*_coordinate",  
      "stored": false,  
      "type": "tdouble"  
    },  
    {  
      "multiValued": true,  
      "name": "ignored_*",  
      "type": "ignored"  
    },  
    {  
      "name": "random_*",  
      "type": "random"  
    },  
    {  
      "indexed": true,  
      "multiValued": true,  
      "name": "attr_*",  
      "stored": true,  
      "type": "text_general"  
    },  
    {  
      "indexed": true,  
      "multiValued": true,  
      "name": "*_txt",  
      "stored": true,  
      "type": "text_general"  
    }  
    ....  
  ],  
  "responseHeader": {  
    "QTime": 1,  
    "status": 0  
  }  
}
```

List Field Types

GET /collection/schema/fieldtypes

GET /collection/schema/fieldtypes/name

List Field Type Parameters

Path Parameters

collection

The collection (or core) name.

name

The name of the field type (if limiting request to a single field type).

Query Parameters

The query parameters can be added to the API request after a '?'.

wt

Defines the format of the response. The options are json or xml. If not specified, JSON will be returned by default.

showDefaults

If true, all default field properties from each dynamic field's field type will be included in the response (e.g., tokenized for solr.TextField). If false, the default, only explicitly specified field properties will be included.

List Field Type Response

The output will include each field type and any defined configuration for the type. The defined configuration can vary for each type, but will minimally include the field type name and the class. If query or index analyzers, tokenizers, or filters are defined, those will also be shown with other defined parameters. See the section [Solr Field Types](#) for more information about how to configure various types of fields.

List Field Type Examples

Get a list of all field types.

```
curl http://localhost:8983/solr/gettingstarted/schema/fieldtypes
```

The sample output below has been truncated to show a few different field types from different parts of the list.

```
{  
  "fieldTypes": [  
    {  
      "analyzer": {  
        "class": "solr.TokenizerChain",  
        "filters": [  
          {  
            "class": "solr.LowerCaseFilterFactory"  
          },  
          {  
            "class": "solr.TrimFilterFactory"  
          },  
          {  
            "class": "solr.PatternReplaceFilterFactory",  
            "pattern": "([^\w\W])",  
            "replace": "all",  
            "replacement": ""  
          }  
        ],  
        "tokenizer": {  
          "class": "solr.KeywordTokenizerFactory"  
        }  
      },  
      "class": "solr.TextField",  
      "dynamicFields": [],  
      "fields": [],  
      "name": "alphaOnlySort",  
      "omitNorms": true,  
      "sortMissingLast": true  
    },  
    {  
      "class": "solr.FloatPointField",  
      "dynamicFields": [  
        "*_fs",  
        "*_f"  
      ],  
      "fields": [  
        "price",  
        "weight"  
      ],  
      "name": "float",  
      "positionIncrementGap": "0",  
    }]  
  }  
}
```

List Copy Fields

GET /collection/schema/copyfields

List Copy Field Parameters

Path Parameters

collection

The collection (or core) name.

Query Parameters

The query parameters can be added to the API request after a '?'.

wt

Defines the format of the response. The options are json or xml. If not specified, JSON will be returned by default.

source.fl

Comma- or space-separated list of one or more copyField source fields to include in the response - copyField directives with all other source fields will be excluded from the response. If not specified, all copyField-s will be included in the response.

dest.fl

Comma- or space-separated list of one or more copyField destination fields to include in the response. copyField directives with all other dest fields will be excluded. If not specified, all copyField-s will be included in the response.

List Copy Field Response

The output will include the source and dest (destination) of each copy field rule defined in schema.xml. For more information about copying fields, see the section [Copying Fields](#).

List Copy Field Examples

Get a list of all copyFields.

```
curl http://localhost:8983/solr/gettingstarted/schema/copyfields
```

The sample output below has been truncated to the first few copy definitions.

```
{  
  "copyFields": [  
    {  
      "dest": "text",  
      "source": "author"  
    },  
    {  
      "dest": "text",  
      "source": "cat"  
    },  
    {  
      "dest": "text",  
      "source": "content"  
    },  
    {  
      "dest": "text",  
      "source": "content_type"  
    },  
  ],  
  "responseHeader": {  
    "QTime": 3,  
    "status": 0  
  }  
}
```

Show Schema Name

GET /collection/schema/name

Show Schema Parameters

Path Parameters

collection

The collection (or core) name.

Query Parameters

The query parameters can be added to the API request after a '?'.

wt

Defines the format of the response. The options are json or xml. If not specified, JSON will be returned by default.

Show Schema Response

The output will be simply the name given to the schema.

Show Schema Examples

Get the schema name.

```
curl http://localhost:8983/solr/gettingstarted/schema/name
```

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "name": "example"
}
```

Show the Schema Version

GET /collection/schema/version

Show Schema Version Parameters

Path Parameters

collection

The collection (or core) name.

Query Parameters

The query parameters can be added to the API request after a '?'.

wt

Defines the format of the response. The options are json or xml. If not specified, JSON will be returned by default.

Show Schema Version Response

The output will simply be the schema version in use.

Show Schema Version Example

Get the schema version:

```
curl http://localhost:8983/solr/gettingstarted/schema/version
```

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 2,
    "version": 1.5
}
```

List UniqueKey

GET /collection/schema/uniquekey

List UniqueKey Parameters

Path Parameters

|collection

The collection (or core) name.

Query Parameters

The query parameters can be added to the API request after a '?'.

|wt

Defines the format of the response. The options are json or xml. If not specified, JSON will be returned by default.

List UniqueKey Response

The output will include simply the field name that is defined as the uniqueKey for the index.

List UniqueKey Example

List the uniqueKey.

```
curl http://localhost:8983/solr/gettingstarted/schema/uniquekey
```

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 2,
    "uniqueKey": "id"
}
```

Show Global Similarity

GET /collection/schema/similarity

Show Global Similarity Parameters

Path Parameters

collection

The collection (or core) name.

Query Parameters

The query parameters can be added to the API request after a '?'.

wt

Defines the format of the response. The options are json or xml. If not specified, JSON will be returned by default.

Show Global Similary Response

The output will include the class name of the global similarity defined (if any).

Show Global Similarity Example

Get the similarity implementation.

```
curl http://localhost:8983/solr/gettingstarted/schema/similarity
```

```
{
  "responseHeader":{
    "status":0,
    "QTime":1},
  "similarity":{
    "class":"org.apache.solr.search.similarities.DefaultSimilarityFactory"}}
```

Manage Resource Data

The [Managed Resources](#) REST API provides a mechanism for any Solr plugin to expose resources that should support CRUD (Create, Read, Update, Delete) operations. Depending on what Field Types and Analyzers are configured in your Schema, additional /schema/ REST API paths may exist. See the [Managed Resources](#) section for more information and examples.

Putting the Pieces Together

At the highest level, schema.xml is structured as follows.

This example is not real XML, but it gives you an idea of the structure of the file.

```
<schema>
  <types>
    <fields>
      <uniqueKey>
        <copyField>
      </uniqueKey>
    </fields>
  </types>
</schema>
```

Obviously, most of the excitement is in types and fields, where the field types and the actual field definitions live.

These are supplemented by copyFields.

The uniqueKey must always be defined.

Types and fields are optional tags



Note that the types and fields sections are optional, meaning you are free to mix field, dynamicField, copyField and fieldType definitions on the top level. This allows for a more logical grouping of related tags in your schema.

Choosing Appropriate Numeric Types

For general numeric needs, consider using one of the IntPointField, LongPointField, FloatPointField, or DoublePointField classes, depending on the specific values you expect. These "Dimensional Point" based numeric classes use specially encoded data structures to support efficient range queries regardless of the size of the ranges used. Enable [DocValues](#) on these fields as needed for sorting and/or faceting.

Some Solr features may not yet work with "Dimensional Points", in which case you may want to consider the equivalent TrieIntField, TrieLongField, TrieFloatField, and TrieDoubleField classes. These field types are deprecated and are likely to be removed in a future major Solr release, but they can still be used if necessary. Configure a precisionStep="0" if you wish to minimize index size, but if you expect users to make frequent range queries on numeric types, use the default precisionStep (by not specifying it) or specify it as precisionStep="8" (which is the default). This offers faster speed for range queries at the expense of increasing index size.

Working With Text

Handling text properly will make your users happy by providing them with the best possible results for text searches.

One technique is using a text field as a catch-all for keyword searching. Most users are not sophisticated about their searches and the most common search is likely to be a simple keyword search. You can use copyField to take a variety of fields and funnel them all into a single text field for keyword searches.

In the schema.xml file for the “techproducts” example included with Solr, copyField declarations are used to dump the contents of cat, name, manu, features, and includes into a single field, text. In addition, it could be a good idea to copy ID into text in case users wanted to search for a particular product by passing its product number to a keyword search.

Another technique is using copyField to use the same field in different ways. Suppose you have a field that is a list of authors, like this:

Schildt, Herbert; Wolpert, Lewis; Davies, P.

For searching by author, you could tokenize the field, convert to lower case, and strip out punctuation:

schildt / herbert / wolpert / lewis / davies / p

For sorting, just use an untokenized field, converted to lower case, with punctuation stripped:

schildt herbert wolpert lewis davies p

Finally, for faceting, use the primary author only via a StringField:

Schildt, Herbert

DocValues

DocValues are a way of recording field values internally that is more efficient for some purposes, such as sorting and faceting, than traditional indexing.

Why DocValues?

The standard way that Solr builds the index is with an *inverted index*. This style builds a list of terms found in all the documents in the index and next to each term is a list of documents that the term appears in (as well as how many times the term appears in that document). This makes search very fast - since users search by terms, having a ready list of term-to-document values makes the query process faster.

For other features that we now commonly associate with search, such as sorting, faceting, and highlighting, this approach is not very efficient. The faceting engine, for example, must look up each term that appears in each document that will make up the result set and pull the document IDs in order to build the facet list. In Solr, this is maintained in memory, and can be slow to load (depending on the number of documents, terms, etc.).

In Lucene 4.0, a new approach was introduced. DocValue fields are now column-oriented fields with a document-to-value mapping built at index time. This approach promises to relieve some of the memory requirements of the fieldCache and make lookups for faceting, sorting, and grouping much faster.

Enabling DocValues

To use docValues, you only need to enable it for a field that you will use it with. As with all schema design, you need to define a field type and then define fields of that type with docValues enabled. All of these actions are done in schema.xml.

Enabling a field for docValues only requires adding docValues="true" to the field (or field type) definition, as in this example from the schema.xml of Solr's sample_techproducts_configs configset:

```
<field name="manu_exact" type="string" indexed="false" stored="false" docValues="true" />
```



If you have already indexed data into your Solr index, you will need to completely reindex your content after changing your field definitions in schema.xml in order to successfully use docValues.

DocValues are only available for specific field types. The types chosen determine the underlying Lucene docValue type that will be used. The available Solr field types are:

- StrField, and UUIDField:
 - If the field is single-valued (i.e., multi-valued is false), Lucene will use the SORTED type.
 - If the field is multi-valued, Lucene will use the SORTED_SET type. Entries are kept in sorted order and duplicates are removed.
- BoolField:
 - If the field is single-valued (i.e., multi-valued is false), Lucene will use the SORTED type.

- If the field is multi-valued, Lucene will use the SORTED_SET type. Entries are kept in sorted order and duplicates are removed.
- Any *PointField Numeric or Date fields, EnumFieldType, and CurrencyFieldType:
 - If the field is single-valued (i.e., multi-valued is false), Lucene will use the NUMERIC type.
 - If the field is multi-valued, Lucene will use the SORTED_NUMERIC type. Entries are kept in sorted order and duplicates are kept.
- Any of the deprecated Trie* Numeric or Date fields, EnumField and CurrencyField:
 - If the field is single-valued (i.e., multi-valued is false), Lucene will use the NUMERIC type.
 - If the field is multi-valued, Lucene will use the SORTED_SET type. Entries are kept in sorted order and duplicates are removed.

These Lucene types are related to how the [values are sorted and stored](#).

There is an additional configuration option available, which is to modify the docValuesFormat [used by the field type](#). The default implementation employs a mixture of loading some things into memory and keeping some on disk. In some cases, however, you may choose to specify an alternative [DocValuesFormat implementation](#). For example, you could choose to keep everything in memory by specifying docValuesFormat="Direct" on a field type:

```
<fieldType name="string_in_mem_dv" class="solr.StrField" docValues="true" docValuesFormat="Direct" />
```

Please note that the docValuesFormat option may change in future releases.



Lucene index back-compatibility is only supported for the default codec. If you choose to customize the docValuesFormat in your schema.xml, upgrading to a future version of Solr may require you to either switch back to the default codec and optimize your index to rewrite it into the default codec before upgrading, or re-build your entire index from scratch after upgrading.

Using DocValues

Sorting, Faceting & Functions

If docValues="true" for a field, then DocValues will automatically be used any time the field is used for [sorting](#), [faceting](#) or [function queries](#).

Retrieving DocValues During Search

Field values retrieved during search queries are typically returned from stored values. However, non-stored docValues fields will be also returned along with other stored fields when all fields (or pattern matching globs) are specified to be returned (e.g., "fl=*") for search queries depending on the effective value of the useDocValuesAsStored parameter for each field. For schema versions >= 1.6, the implicit default is useDocValuesAsStored="true". See [Field Type Definitions and Properties & Defining Fields](#) for more details.

When useDocValuesAsStored="false", non-stored DocValues fields can still be explicitly requested by name

in the [fl param](#), but will not match glob patterns ("*"). Note that returning DocValues along with "regular" stored fields at query time has performance implications that stored fields may not because DocValues are column-oriented and may therefore incur additional cost to retrieve for each returned document. Also note that while returning non-stored fields from DocValues, the values of a multi-valued field are returned in sorted order rather than insertion order and may have duplicates removed, see above. If you require the multi-valued fields to be returned in the original insertion order, then make your multi-valued field as stored (such a change requires reindexing).

In cases where the query is returning *only* docValues fields performance may improve since returning stored fields requires disk reads and decompression whereas returning docValues fields in the fl list only requires memory access.

When retrieving fields from their docValues form (such as when using the [/export handler](#), [streaming expressions](#) or if the field is requested in the f1 parameter), two important differences between regular stored fields and docValues fields must be understood:

1. Order is *not* preserved. When retrieving stored fields, the insertion order is the return order. For docValues, it is the *sorted* order.
2. For field types using SORTED_SET (see above), multiple identical entries are collapsed into a single value. Thus if values 4, 5, 2, 4, 1 are inserted, the values returned will be 1, 2, 4, 5.

Schemaless Mode

Schemaless Mode is a set of Solr features that, when used together, allow users to rapidly construct an effective schema by simply indexing sample data, without having to manually edit the schema.

These Solr features, all controlled via `solrconfig.xml`, are:

1. Managed schema: Schema modifications are made at runtime through Solr APIs, which requires the use of a `schemaFactory` that supports these changes. See the section [Schema Factory Definition in SolrConfig](#) for more details.
2. Field value class guessing: Previously unseen fields are run through a cascading set of value-based parsers, which guess the Java class of field values - parsers for Boolean, Integer, Long, Float, Double, and Date are currently available.
3. Automatic schema field addition, based on field value class(es): Previously unseen fields are added to the schema, based on field value Java classes, which are mapped to schema field types - see [Solr Field Types](#).

Using the Schemaless Example

The three features of schemaless mode are pre-configured in the `_default configset` in the Solr distribution. To start an example instance of Solr using these configs, run the following command:

```
bin/solr start -e schemaless
```

This will launch a single Solr server, and automatically create a collection (named “gettingstarted”) that contains only three fields in the initial schema: `id`, `_version_`, and `_text_`.

You can use the `/schema/fields` [Schema API](#) to confirm this: `curl http://localhost:8983/solr/gettingstarted/schema/fields` will output:

```
{  
  "responseHeader":{  
    "status":0,  
    "QTime":1},  
  "fields":[]  
  {  
    "name": "_text_",  
    "type": "text_general",  
    "multiValued":true,  
    "indexed":true,  
    "stored":false},  
  {  
    "name": "_version_",  
    "type": "long",  
    "indexed":true,  
    "stored":true},  
  {  
    "name": "id",  
    "type": "string",  
    "multiValued":false,  
    "indexed":true,  
    "required":true,  
    "stored":true,  
    "uniqueKey":true}]}  
}
```

Configuring Schemaless Mode

As described above, there are three configuration elements that need to be in place to use Solr in schemaless mode. In the `_default` configset included with Solr these are already configured. If, however, you would like to implement schemaless on your own, you should make the following changes.

Enable Managed Schema

As described in the section [Schema Factory Definition in SolrConfig](#), Managed Schema support is enabled by default, unless your configuration specifies that `ClassicIndexSchemaFactory` should be used.

You can configure the `ManagedIndexSchemaFactory` (and control the resource file used, or disable future modifications) by adding an explicit `<schemaFactory/>` like the one below, please see [Schema Factory Definition in SolrConfig](#) for more details on the options available.

```
<schemaFactory class="ManagedIndexSchemaFactory">  
  <bool name="mutable">true</bool>  
  <str name="managedSchemaResourceName">managed-schema</str>  
</schemaFactory>
```

Enable Field Class Guessing

In Solr, an `UpdateRequestProcessorChain` defines a chain of plugins that are applied to documents before or while they are indexed.

The field guessing aspect of Solr's schemaless mode uses a specially-defined UpdateRequestProcessorChain that allows Solr to guess field types. You can also define the default field type classes to use.

To start, you should define it as follows (see the javadoc links below for update processor factory documentation):

```
<updateProcessor class="solr.UUIDUpdateProcessorFactory" name="uuid"/>
<updateProcessor class="solr.RemoveBlankFieldUpdateProcessorFactory" name="remove-blank"/>
<updateProcessor class="solr.FieldNameMutatingUpdateProcessorFactory" name="field-name-
mutating"> ①
  <str name="pattern">[^w-\. ]</str>
  <str name="replacement">_</str>
</updateProcessor>
<updateProcessor class="solr.ParseBooleanFieldUpdateProcessorFactory" name="parse-boolean"/> ②
<updateProcessor class="solr.ParseLongFieldUpdateProcessorFactory" name="parse-long"/>
<updateProcessor class="solr.ParseDoubleFieldUpdateProcessorFactory" name="parse-double"/>
<updateProcessor class="solr.ParseDateFieldUpdateProcessorFactory" name="parse-date">
  <arr name="format">
    <str>yyyy-MM-dd['T'][HH:mm[:ss[.SSS]]][z]</str>
    <str>yyyy-MM-dd['T'][HH:mm[:ss[,SSS]]][z]</str>
    <str>yyyy-MM-dd HH:mm[:ss[.SSS]]][z]</str>
    <str>yyyy-MM-dd HH:mm[:ss[,SSS]][z]</str>
    <str>[EEE, ]dd MMM yyyy HH:mm[:ss] z</str>
    <str>EEE, dd-MMM-yy HH:mm:ss z</str>
    <str>EEE MMM ppd HH:mm:ss [z ]yyyy</str>
  </arr>
</updateProcessor>
<updateProcessor class="solr.AddSchemaFieldsUpdateProcessorFactory" name="add-schema-fields">
③
  <lst name="typeMapping">
    <str name="valueClass">java.lang.String</str> ④
    <str name="fieldType">text_general</str>
    <lst name="copyField"> ⑤
      <str name="dest">*_str</str>
      <int name="maxChars">256</int>
    </lst>
    <!-- Use as default mapping instead of defaultFieldType -->
    <bool name="default">true</bool>
  </lst>
  <lst name="typeMapping">
    <str name="valueClass">java.lang.Boolean</str>
    <str name="fieldType">booleans</str>
  </lst>
  <lst name="typeMapping">
    <str name="valueClass">java.util.Date</str>
    <str name="fieldType">pdates</str>
  </lst>
  <lst name="typeMapping">
    <str name="valueClass">java.lang.Long</str> ⑥
    <str name="valueClass">java.lang.Integer</str>
    <str name="fieldType">plongs</str>
  </lst>
```

```
<lst name="typeMapping">
  <str name="valueClass">java.lang.Number</str>
  <str name="fieldType">pdoubles</str>
</lst>
</updateProcessor>

<!-- The update.autoCreateFields property can be turned to false to disable schemaless mode --&gt;
&lt;updateRequestProcessorChain name="add-unknown-fields-to-the-schema" default=
"${update.autoCreateFields:true}"
  processor="uuid,remove-blank,fieldName-mutating,parse-boolean,parse-long,parse-
double,parse-date,add-schema-fields"&gt; ⑦
  &lt;processor class="solr.LogUpdateProcessorFactory"/&gt;
  &lt;processor class="solr.DistributedUpdateProcessorFactory"/&gt;
  &lt;processor class="solr.RunUpdateProcessorFactory"/&gt;
&lt;/updateRequestProcessorChain&gt;</pre>
```

There are many things defined in this chain. Let's step through a few of them.

- ① First, we're using the FieldNameMutatingUpdateProcessorFactory to lower-case all field names. Note that this and every following <processor> element include a name. These names will be used in the final chain definition at the end of this example.
- ② Next we add several update request processors to parse different field types. Note the ParseDateFieldUpdateProcessorFactory includes a long list of possible date formations that would be parsed into valid Solr dates. If you have a custom date, you could add it to this list (see the link to the Javadocs below to get information on how).
- ③ Once the fields have been parsed, we define the field types that will be assigned to those fields. You can modify any of these that you would like to change.
- ④ In this definition, if the parsing step decides the incoming data in a field is a string, we will put this into a field in Solr with the field type text_general. This field type by default allows Solr to query on this field.
- ⑤ After we've added the text_general field, we have also defined a copy field rule that will copy all data from the new text_general field to a field with the same name suffixed with _str. This is done by Solr's dynamic fields feature. By defining the target of the copy field rule as a dynamic field in this way, you can control the field type used in your schema. The default selection allows Solr to facet, highlight, and sort on these fields.
- ⑥ This is another example of a mapping rule. In this case we define that when either of the Long or Integer field parsers identify a field, they should both map their fields to the plongs field type.
- ⑦ Finally, we add a chain definition that calls the list of plugins. These plugins are each called by the names we gave to them when we defined them. We can also add other processors to the chain, as shown here. Note we have also given the entire chain a name ("add-unknown-fields-to-the-schema"). We'll use this name in the next section to specify that our update request handler should use this chain definition.

This chain definition will make a number of copy field rules for string fields to be created from corresponding text fields. If your data causes you to end up with a lot of copy field rules, indexing may be slowed down noticeably, and your index size will be larger. To control for these issues, it's recommended that you review the copy field rules that are created, and remove any which you do not need for faceting, sorting, highlighting, etc.



If you're interested in more information about the classes used in this chain, here are links to the Javadocs for update processor factories mentioned above:

- [UUIDUpdateProcessorFactory](#)
- [RemoveBlankFieldUpdateProcessorFactory](#)
- [FieldNameMutatingUpdateProcessorFactory](#)
- [ParseBooleanFieldUpdateProcessorFactory](#)
- [ParseLongFieldUpdateProcessorFactory](#)
- [ParseDoubleFieldUpdateProcessorFactory](#)
- [ParseDateFieldUpdateProcessorFactory](#)
- [AddSchemaFieldsUpdateProcessorFactory](#)

Set the Default UpdateRequestProcessorChain

Once the UpdateRequestProcessorChain has been defined, you must instruct your UpdateRequestHandlers to use it when working with index updates (i.e., adding, removing, replacing documents).

There are two ways to do this. The update chain shown above has a default=true attribute which will use it for any update handler.

An alternative, more explicit way is to use [InitParams](#) to set the defaults on all /update request handlers:

```
<initParams path="/update/**">
  <lst name="defaults">
    <str name="update.chain">add-unknown-fields-to-the-schema</str>
  </lst>
</initParams>
```



After all of these changes have been made, Solr should be restarted or the cores reloaded.

Disabling Automatic Field Guessing

Automatic field creation can be disabled with the update.autoCreateFields property. To do this, you can use bin/solr config with a command such as:

```
bin/solr config -c mycollection -p 8983 -action set-user-property -property
update.autoCreateFields -value false
```

Examples of Indexed Documents

Once the schemaless mode has been enabled (whether you configured it manually or are using the _default configset), documents that include fields that are not defined in your schema will be indexed, using the guessed field types which are automatically added to the schema.

For example, adding a CSV document will cause unknown fields to be added, with fieldType based on

values:

```
curl "http://localhost:8983/solr/gettingstarted/update?commit=true&wt=xml" -H "Content-type:application/csv" -d '  
id,Artist,Album,Released,Rating,FromDistributor,Sold  
44C,Old Shews,Mead for Walking,1988-08-13,0.01,14,0'
```

Output indicating success:

```
<response>  
<lst name="responseHeader"><int name="status">0</int><int name="QTime">106</int></lst>  
</response>
```

The fields now in the schema (output from curl

<http://localhost:8983/solr/gettingstarted/schema/fields>):

```
{  
  "responseHeader":{  
    "status":0,  
    "QTime":2},  
  "fields":[{  
    "name": "Album",  
    "type": "text_general"},  
    {  
      "name": "Artist",  
      "type": "text_general"},  
    {  
      "name": "FromDistributor",  
      "type": "plongs"},  
    {  
      "name": "Rating",  
      "type": "pdoubles"},  
    {  
      "name": "Released",  
      "type": "pdates"},  
    {  
      "name": "Sold",  
      "type": "plongs"},  
    {  
      "name": "_root_", ...},  
    {  
      "name": "_text_", ...},  
    {  
      "name": "_version_", ...},  
    {  
      "name": "id", ...}  
  }]
```

In addition string versions of the text fields are indexed, using copyFields to a *_str dynamic field: (output

from curl `http://localhost:8983/solr/gettingstarted/schema/copyfields`):

```
{  
  "responseHeader":{  
    "status":0,  
    "QTime":0},  
  "copyFields":[]  
  {  
    "source":"Artist",  
    "dest":"Artist_str",  
    "maxChars":256},  
  {  
    "source":"Album",  
    "dest":"Album_str",  
    "maxChars":256}]}}
```

You Can Still Be Explicit

Even if you want to use schemaless mode for most fields, you can still use the [Schema API](#) to pre-emptively create some fields, with explicit types, before you index documents that use them.



Internally, the Schema API and the Schemaless Update Processors both use the same [Managed Schema](#) functionality.

Also, if you do not need the `*_str` version of a text field, you can simply remove the `copyField` definition from the auto-generated schema and it will not be re-added since the original field is now defined.

Once a field has been added to the schema, its field type is fixed. As a consequence, adding documents with field value(s) that conflict with the previously guessed field type will fail. For example, after adding the above document, the “Sold” field has the `fieldType plongs`, but the document below has a non-integral decimal value in this field:

```
curl "http://localhost:8983/solr/gettingstarted/update?commit=true&wt=xml" -H "Content-Type:application/csv" -d '  
id,Description,Sold  
19F,Cassettes by the pound,4.93'
```

This document will fail, as shown in this output:

```
<response>
  <lst name="responseHeader">
    <int name="status">400</int>
    <int name="QTime">7</int>
  </lst>
  <lst name="error">
    <str name="msg">ERROR: [doc=19F] Error adding field 'Sold'='4.93' msg=For input string:
    "4.93"</str>
    <int name="code">400</int>
  </lst>
</response>
```

Understanding Analyzers, Tokenizers, and Filters

The following sections describe how Solr breaks down and works with textual data. There are three main concepts to understand: analyzers, tokenizers, and filters.

- [Field analyzers](#) are used both during ingestion, when a document is indexed, and at query time. An analyzer examines the text of fields and generates a token stream. Analyzers may be a single class or they may be composed of a series of tokenizer and filter classes.
- [Tokenizers](#) break field data into lexical units, or *tokens*.
- [Filters](#) examine a stream of tokens and keep them, transform or discard them, or create new ones. Tokenizers and filters may be combined to form pipelines, or *chains*, where the output of one is input to the next. Such a sequence of tokenizers and filters is called an *analyzer* and the resulting output of an analyzer is used to match query results or build indices.

Using Analyzers, Tokenizers, and Filters

Although the analysis process is used for both indexing and querying, the same analysis process need not be used for both operations. For indexing, you often want to simplify, or normalize, words. For example, setting all letters to lowercase, eliminating punctuation and accents, mapping words to their stems, and so on. Doing so can increase recall because, for example, "ram", "Ram" and "RAM" would all match a query for "ram". To increase query-time precision, a filter could be employed to narrow the matches by, for example, ignoring all-cap acronyms if you're interested in male sheep, but not Random Access Memory.

The tokens output by the analysis process define the values, or *terms*, of that field and are used either to build an index of those terms when a new document is added, or to identify which documents contain the terms you are querying for.

For More Information

These sections will show you how to configure field analyzers and also serves as a reference for the details of configuring each of the available tokenizer and filter classes. It also serves as a guide so that you can configure your own analysis classes if you have special needs that cannot be met with the included filters or tokenizers.

For Analyzers, see:

- [Analyzers](#): Detailed conceptual information about Solr analyzers.
- [Running Your Analyzer](#): Detailed information about testing and running your Solr analyzer.

For Tokenizers, see:

- [About Tokenizers](#): Detailed conceptual information about Solr tokenizers.
- [Tokenizers](#): Information about configuring tokenizers, and about the tokenizer factory classes included in this distribution of Solr.

For Filters, see:

- [About Filters](#): Detailed conceptual information about Solr filters.
- [Filter Descriptions](#): Information about configuring filters, and about the filter factory classes included in this distribution of Solr.
- [CharFilterFactories](#): Information about filters for pre-processing input characters.

To find out how to use Tokenizers and Filters with various languages, see:

- [Language Analysis](#): Information about tokenizers and filters for character set conversion or for use with specific languages.

Analyzers

An analyzer examines the text of fields and generates a token stream.

Analyzers are specified as a child of the <fieldType> element in the schema.xml configuration file (in the same conf/ directory as solrconfig.xml).

In normal usage, only fields of type solr.TextField or solr.SortableTextField will specify an analyzer. The simplest way to configure an analyzer is with a single <analyzer> element whose class attribute is a fully qualified Java class name. The named class must derive from org.apache.lucene.analysis.Analyzer. For example:

```
<fieldType name="nametext" class="solr.TextField">
  <analyzer class="org.apache.lucene.analysis.coreWhitespaceAnalyzer"/>
</fieldType>
```

In this case a single class, WhitespaceAnalyzer, is responsible for analyzing the content of the named text field and emitting the corresponding tokens. For simple cases, such as plain English prose, a single analyzer class like this may be sufficient. But it's often necessary to do more complex analysis of the field content.

Even the most complex analysis requirements can usually be decomposed into a series of discrete, relatively simple processing steps. As you will soon discover, the Solr distribution comes with a large selection of tokenizers and filters that covers most scenarios you are likely to encounter. Setting up an analyzer chain is very straightforward; you specify a simple <analyzer> element (no class attribute) with child elements that name factory classes for the tokenizer and filters to use, in the order you want them to run.

For example:

```
<fieldType name="nametext" class="solr.TextField">
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.StopFilterFactory"/>
    <filter class="solr.EnglishPorterFilterFactory"/>
  </analyzer>
</fieldType>
```

Note that classes in the org.apache.solr.analysis package may be referred to here with the shorthand solr. prefix.

In this case, no Analyzer class was specified on the <analyzer> element. Rather, a sequence of more specialized classes are wired together and collectively act as the Analyzer for the field. The text of the field is passed to the first item in the list (solr.StandardTokenizerFactory), and the tokens that emerge from the last one (solr.EnglishPorterFilterFactory) are the terms that are used for indexing or querying any fields that use the "nametext" fieldType.

Field Values versus Indexed Terms



The output of an Analyzer affects the *terms* indexed in a given field (and the terms used when parsing queries against those fields) but it has no impact on the *stored* value for the fields. For example: an analyzer might split "Brown Cow" into two indexed terms "brown" and "cow", but the stored value will still be a single String: "Brown Cow"

Analysis Phases

Analysis takes place in two contexts. At index time, when a field is being created, the token stream that results from analysis is added to an index and defines the set of terms (including positions, sizes, and so on) for the field. At query time, the values being searched for are analyzed and the terms that result are matched against those that are stored in the field's index.

In many cases, the same analysis should be applied to both phases. This is desirable when you want to query for exact string matches, possibly with case-insensitivity, for example. In other cases, you may want to apply slightly different analysis steps during indexing than those used at query time.

If you provide a simple <analyzer> definition for a field type, as in the examples above, then it will be used for both indexing and queries. If you want distinct analyzers for each phase, you may include two <analyzer> definitions distinguished with a type attribute. For example:

```
<fieldType name="nametext" class="solr.TextField">
  <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.KeepWordFilterFactory" words="keepwords.txt"/>
    <filter class="solr.SynonymFilterFactory" synonyms="syns.txt"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
  </analyzer>
</fieldType>
```

In this theoretical example, at index time the text is tokenized, the tokens are set to lowercase, any that are not listed in keepwords.txt are discarded and those that remain are mapped to alternate values as defined by the synonym rules in the file syns.txt. This essentially builds an index from a restricted set of possible values and then normalizes them to values that may not even occur in the original text.

At query time, the only normalization that happens is to convert the query terms to lowercase. The filtering and mapping steps that occur at index time are not applied to the query terms. Queries must then, in this example, be very precise, using only the normalized terms that were stored at index time.

Analysis for Multi-Term Expansion

In some types of queries (i.e., Prefix, Wildcard, Regex, etc.) the input provided by the user is not natural language intended for Analysis. Things like Synonyms or Stop word filtering do not work in a logical way in these types of Queries.

When Solr needs to perform analysis for a query that results in multi-term expansion, then the normalize method is called for each factory in the filter chain. Factories that provide filters that do not make sense in this context will return their inputs unchanged. Normalization applies to both CharFilters and TokenFilters.

For most use cases, this provides the best possible behavior, but if you wish for absolute control over the analysis performed on these types of queries, you may explicitly define a multiterm analyzer to use, such as in the following example:

```
<fieldType name="nametext" class="solr.TextField">
  <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.KeepWordFilterFactory" words="keepwords.txt"/>
    <filter class="solr.SynonymFilterFactory" synonyms="syns.txt"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
  </analyzer>
  <!-- No analysis at all when doing queries that involved Multi-Term expansion -->
  <analyzer type="multiterm">
    <tokenizer class="solr.KeywordTokenizerFactory" />
  </analyzer>
</fieldType>
```

About Tokenizers

The job of a [tokenizer](#) is to break up a stream of text into tokens, where each token is (usually) a subsequence of the characters in the text. An analyzer is aware of the field it is configured for, but a tokenizer is not. Tokenizers read from a character stream (a Reader) and produce a sequence of Token objects (a TokenStream).

Characters in the input stream may be discarded, such as whitespace or other delimiters. They may also be added to or replaced, such as mapping aliases or abbreviations to normalized forms. A token contains various metadata in addition to its text value, such as the location at which the token occurs in the field. Because a tokenizer may produce tokens that diverge from the input text, you should not assume that the text of the token is the same text that occurs in the field, or that its length is the same as the original text. It's also possible for more than one token to have the same position or refer to the same offset in the original text. Keep this in mind if you use token metadata for things like highlighting search results in the field text.

```
<fieldType name="text" class="solr.TextField">
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
  </analyzer>
</fieldType>
```

The class named in the tokenizer element is not the actual tokenizer, but rather a class that implements the TokenizerFactory API. This factory class will be called upon to create new tokenizer instances as needed. Objects created by the factory must derive from Tokenizer, which indicates that they produce sequences of tokens. If the tokenizer produces tokens that are usable as is, it may be the only component of the analyzer. Otherwise, the tokenizer's output tokens will serve as input to the first filter stage in the pipeline.

A TypeTokenFilterFactory is available that creates a TypeTokenFilter that filters tokens based on their TypeAttribute, which is set in factory.getStopTypes.

For a complete list of the available TokenFilters, see the section [Tokenizers](#).

When to Use a CharFilter vs. a TokenFilter

There are several pairs of CharFilters and TokenFilters that have related (i.e., MappingCharFilter and ASCIIFoldingFilter) or nearly identical (i.e., PatternReplaceCharFilterFactory and PatternReplaceFilterFactory) functionality and it may not always be obvious which is the best choice.

The decision about which to use depends largely on which Tokenizer you are using, and whether you need to preprocess the stream of characters.

For example, suppose you have a tokenizer such as StandardTokenizer and although you are pretty happy with how it works overall, you want to customize how some specific characters behave. You could modify the rules and re-build your own tokenizer with JFlex, but it might be easier to simply map some of the characters before tokenization with a CharFilter.

About Filters

Like [tokenizers](#), [filters](#) consume input and produce a stream of tokens. Filters also derive from `org.apache.lucene.analysis.TokenStream`. Unlike tokenizers, a filter's input is another `TokenStream`. The job of a filter is usually easier than that of a tokenizer since in most cases a filter looks at each token in the stream sequentially and decides whether to pass it along, replace it or discard it.

A filter may also do more complex analysis by looking ahead to consider multiple tokens at once, although this is less common. One hypothetical use for such a filter might be to normalize state names that would be tokenized as two words. For example, the single token "california" would be replaced with "CA", while the token pair "rhode" followed by "island" would become the single token "RI".

Because filters consume one `TokenStream` and produce a new `TokenStream`, they can be chained one after another indefinitely. Each filter in the chain in turn processes the tokens produced by its predecessor. The order in which you specify the filters is therefore significant. Typically, the most general filtering is done first, and later filtering stages are more specialized.

```
<fieldType name="text" class="solr.TextField">
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.EnglishPorterFilterFactory"/>
  </analyzer>
</fieldType>
```

This example starts with Solr's standard tokenizer, which breaks the field's text into tokens. All the tokens are then set to lowercase, which will facilitate case-insensitive matching at query time.

The last filter in the above example is a stemmer filter that uses the Porter stemming algorithm. A stemmer is basically a set of mapping rules that maps the various forms of a word back to the base, or *stem*, word from which they derive. For example, in English the words "hugs", "hugging" and "hugged" are all forms of the stem word "hug". The stemmer will replace all of these terms with "hug", which is what will be indexed. This means that a query for "hug" will match the term "hugged", but not "huge".

Conversely, applying a stemmer to your query terms will allow queries containing non stem terms, like "hugging", to match documents with different variations of the same stem word, such as "hugged". This works because both the indexer and the query will map to the same stem ("hug").

Word stemming is, obviously, very language specific. Solr includes several language-specific stemmers created by the [Snowball](#) generator that are based on the Porter stemming algorithm. The generic Snowball Porter Stemmer Filter can be used to configure any of these language stemmers. Solr also includes a convenience wrapper for the English Snowball stemmer. There are also several purpose-built stemmers for non-English languages. These stemmers are described in [Language Analysis](#).

Tokenizers

Tokenizers are responsible for breaking field data into lexical units, or *tokens*.

You configure the tokenizer for a text field type in schema.xml with a <tokenizer> element, as a child of <analyzer>:

```
<fieldType name="text" class="solr.TextField">
  <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
  </analyzer>
</fieldType>
```

The class attribute names a factory class that will instantiate a tokenizer object when needed. Tokenizer factory classes implement the org.apache.solr.analysis.TokenizerFactory. A TokenizerFactory's create() method accepts a Reader and returns a TokenStream. When Solr creates the tokenizer it passes a Reader object that provides the content of the text field.

Arguments may be passed to tokenizer factories by setting attributes on the <tokenizer> element.

```
<fieldType name="semicolonDelimited" class="solr.TextField">
  <analyzer type="query">
    <tokenizer class="solr.PatternTokenizerFactory" pattern="; "/>
  </analyzer>
</fieldType>
```

The following sections describe the tokenizer factory classes included in this release of Solr.

For user tips about Solr's tokenizers, see <http://wiki.apache.org/solr/AnalyzersTokenizersTokenFilters>.

Standard Tokenizer

This tokenizer splits the text field into tokens, treating whitespace and punctuation as delimiters. Delimiter characters are discarded, with the following exceptions:

- Periods (dots) that are not followed by whitespace are kept as part of the token, including Internet domain names.
- The "@" character is among the set of token-splitting punctuation, so email addresses are **not** preserved as single tokens.

Note that words are split at hyphens.

The Standard Tokenizer supports [Unicode standard annex UAX#29](#) word boundaries with the following token types: <ALPHANUM>, <NUM>, <SOUTHEAST_ASIAN>, <IDEOGRAPHIC>, and <HIRAGANA>.

Factory class: solr.StandardTokenizerFactory

Arguments:

maxTokenLength: (integer, default 255) Solr ignores tokens that exceed the number of characters specified by maxTokenLength.

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
</analyzer>
```

In: "Please, email john.doe@foo.com by 03-09, re: m37-xq."

Out: "Please", "email", "john.doe", "foo.com", "by", "03", "09", "re", "m37", "xq"

Classic Tokenizer

The Classic Tokenizer preserves the same behavior as the Standard Tokenizer of Solr versions 3.1 and previous. It does not use the [Unicode standard annex UAX#29](#) word boundary rules that the Standard Tokenizer uses. This tokenizer splits the text field into tokens, treating whitespace and punctuation as delimiters. Delimiter characters are discarded, with the following exceptions:

- Periods (dots) that are not followed by whitespace are kept as part of the token.
- Words are split at hyphens, unless there is a number in the word, in which case the token is not split and the numbers and hyphen(s) are preserved.
- Recognizes Internet domain names and email addresses and preserves them as a single token.

Factory class: solr.ClassicTokenizerFactory

Arguments:

maxTokenLength: (integer, default 255) Solr ignores tokens that exceed the number of characters specified by maxTokenLength.

Example:

```
<analyzer>
  <tokenizer class="solr.ClassicTokenizerFactory"/>
</analyzer>
```

In: "Please, email john.doe@foo.com by 03-09, re: m37-xq."

Out: "Please", "email", "john.doe@foo.com", "by", "03-09", "re", "m37-xq"

Keyword Tokenizer

This tokenizer treats the entire text field as a single token.

Factory class: solr.KeywordTokenizerFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.KeywordTokenizerFactory"/>
</analyzer>
```

In: "Please, email john.doe@foo.com by 03-09, re: m37-xq."

Out: "Please, email john.doe@foo.com by 03-09, re: m37-xq."

Letter Tokenizer

This tokenizer creates tokens from strings of contiguous letters, discarding all non-letter characters.

Factory class: solr.LetterTokenizerFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.LetterTokenizerFactory"/>
</analyzer>
```

In: "I can't."

Out: "I", "can", "t"

Lower Case Tokenizer

Tokenizes the input stream by delimiting at non-letters and then converting all letters to lowercase. Whitespace and non-letters are discarded.

Factory class: solr.LowerCaseTokenizerFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.LowerCaseTokenizerFactory"/>
</analyzer>
```

In: "I just *LOVE* my iPhone!"

Out: "i", "just", "love", "my", "iphone"

N-Gram Tokenizer

Reads the field text and generates n-gram tokens of sizes in the given range.

Factory class: solr.NGramTokenizerFactory

Arguments:

minGramSize: (integer, default 1) The minimum n-gram size, must be > 0.

maxGramSize: (integer, default 2) The maximum n-gram size, must be \geq minGramSize.

Example:

Default behavior. Note that this tokenizer operates over the whole field. It does not break the field at whitespace. As a result, the space character is included in the encoding.

```
<analyzer>
  <tokenizer class="solr.NGramTokenizerFactory" />
</analyzer>
```

In: "hey man"

Out: "h", "e", "y", " ", "m", "a", "n", "he", "ey", "y ", " m", "ma", "an"

Example:

With an n-gram size range of 4 to 5:

```
<analyzer>
  <tokenizer class="solr.NGramTokenizerFactory" minGramSize="4" maxGramSize="5" />
</analyzer>
```

In: "bicycle"

Out: "bicy", "bicyc", "icyc", "icycl", "cycl", "cycle", "ycle"

Edge N-Gram Tokenizer

Reads the field text and generates edge n-gram tokens of sizes in the given range.

Factory class: solr.EdgeNGramTokenizerFactory

Arguments:

minGramSize: (integer, default is 1) The minimum n-gram size, must be > 0.

maxGramSize: (integer, default is 1) The maximum n-gram size, must be \geq minGramSize.

Example:

Default behavior (min and max default to 1):

```
<analyzer>
  <tokenizer class="solr.EdgeNGramTokenizerFactory"/>
</analyzer>
```

In: "babaloo"

Out: "b"

Example:

Edge n-gram range of 2 to 5

```
<analyzer>
  <tokenizer class="solr.EdgeNGramTokenizerFactory" minGramSize="2" maxGramSize="5"/>
</analyzer>
```

In: "babaloo"

Out: "ba", "bab", "baba", "babal"

ICU Tokenizer

This tokenizer processes multilingual text and tokenizes it appropriately based on its script attribute.

You can customize this tokenizer's behavior by specifying [per-script rule files](#). To add per-script rules, add a rulefiles argument, which should contain a comma-separated list of code:rulefile pairs in the following format: four-letter ISO 15924 script code, followed by a colon, then a resource path. For example, to specify rules for Latin (script code "Latn") and Cyrillic (script code "Cyrl"), you would enter Latn:my.Latin.rules.rb, Cyrl:my.Cyrillic.rules.rb.

The default configuration for solr.ICUTokenizerFactory provides UAX#29 word break rules tokenization (like solr.StandardTokenizer), but also includes custom tailorings for Hebrew (specializing handling of double and single quotation marks), for syllable tokenization for Khmer, Lao, and Myanmar, and dictionary-based word segmentation for CJK characters.

Factory class: solr.ICUTokenizerFactory

Arguments:

rulefile: a comma-separated list of code:rulefile pairs in the following format: four-letter ISO 15924 script code, followed by a colon, then a resource path.

Example:

```
<analyzer>
  <!-- no customization -->
  <tokenizer class="solr.ICUTokenizerFactory"/>
</analyzer>
```

```
<analyzer>
  <tokenizer class="solr.ICUTokenizerFactory"
    rulefiles="Latn:my.Latin.rules.rbki,Cyrl:my.Cyrillic.rules.rbki"/>
</analyzer>
```



To use this tokenizer, you must add additional jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See the solr/contrib/analysis-extras/README.txt for information on which jars you need to add.

Path Hierarchy Tokenizer

This tokenizer creates synonyms from file path hierarchies.

Factory class: solr.PathHierarchyTokenizerFactory

Arguments:

delimiter: (character, no default) You can specify the file path delimiter and replace it with a delimiter you provide. This can be useful for working with backslash delimiters.

replace: (character, no default) Specifies the delimiter character Solr uses in the tokenized output.

Example:

```
<fieldType name="text_path" class="solr.TextField" positionIncrementGap="100">
  <analyzer>
    <tokenizer class="solr.PathHierarchyTokenizerFactory" delimiter="\\" replace="/" />
  </analyzer>
</fieldType>
```

In: "c:\usr\local\apache"

Out: "c:", "c:/usr", "c:/usr/local", "c:/usr/local/apache"

Regular Expression Pattern Tokenizer

This tokenizer uses a Java regular expression to break the input text stream into tokens. The expression provided by the pattern argument can be interpreted either as a delimiter that separates tokens, or to match patterns that should be extracted from the text as tokens.

See [the Javadocs for](#) java.util.regex.Pattern for more information on Java regular expression syntax.

Factory class: solr.PatternTokenizerFactory

Arguments:

pattern: (Required) The regular expression, as defined by in `java.util.regex.Pattern`.

group: (Optional, default -1) Specifies which regex group to extract as the token(s). The value -1 means the regex should be treated as a delimiter that separates tokens. Non-negative group numbers (≥ 0) indicate that character sequences matching that regex group should be converted to tokens. Group zero refers to the entire regex, groups greater than zero refer to parenthesized sub-expressions of the regex, counted from left to right.

Example:

A comma separated list. Tokens are separated by a sequence of zero or more spaces, a comma, and zero or more spaces.

```
<analyzer>
  <tokenizer class="solr.PatternTokenizerFactory" pattern="\s*,\s*"/>
</analyzer>
```

In: "fee,fie, foe , fum, foo"

Out: "fee", "fie", "foe", "fum", "foo"

Example:

Extract simple, capitalized words. A sequence of at least one capital letter followed by zero or more letters of either case is extracted as a token.

```
<analyzer>
  <tokenizer class="solr.PatternTokenizerFactory" pattern="[A-Z][A-Za-z]*" group="0"/>
</analyzer>
```

In: "Hello. My name is Inigo Montoya. You killed my father. Prepare to die."

Out: "Hello", "My", "Inigo", "Montoya", "You", "Prepare"

Example:

Extract part numbers which are preceded by "SKU", "Part" or "Part Number", case sensitive, with an optional semi-colon separator. Part numbers must be all numeric digits, with an optional hyphen. Regex capture groups are numbered by counting left parenthesis from left to right. Group 3 is the subexpression "[0-9-]+", which matches one or more digits or hyphens.

```
<analyzer>
  <tokenizer class="solr.PatternTokenizerFactory" pattern="(SKU|Part(\sNumber)?):?\s([\d-\d]+)" group="3"/>
</analyzer>
```

In: "SKU: 1234, Part Number 5678, Part: 126-987"

Out: "1234", "5678", "126-987"

Simplified Regular Expression Pattern Tokenizer

This tokenizer is similar to the PatternTokenizerFactory described above, but uses Lucene RegExp pattern matching to construct distinct tokens for the input stream. The syntax is more limited than PatternTokenizerFactory, but the tokenization is quite a bit faster.

Factory class: solr.SimplePatternTokenizerFactory

Arguments:

pattern: (Required) The regular expression, as defined by in the RegExp javadocs, identifying the characters to include in tokens. The matching is greedy such that the longest token matching at a given point is created. Empty tokens are never created.

maxDeterminizedStates: (Optional, default 10000) the limit on total state count for the determined automaton computed from the regexp.

Example:

To match tokens delimited by simple whitespace characters:

```
<analyzer>
  <tokenizer class="solr.SimplePatternTokenizerFactory" pattern="[^ \t\r\n]+"/>
</analyzer>
```

Simplified Regular Expression Pattern Splitting Tokenizer

This tokenizer is similar to the SimplePatternTokenizerFactory described above, but uses Lucene RegExp pattern matching to identify sequences of characters that should be used to split tokens. The syntax is more limited than PatternTokenizerFactory, but the tokenization is quite a bit faster.

Factory class: solr.SimplePatternSplitTokenizerFactory

Arguments:

pattern: (Required) The regular expression, as defined by in the RegExp javadocs, identifying the characters that should split tokens. The matching is greedy such that the longest token separator matching at a given point is matched. Empty tokens are never created.

maxDeterminizedStates: (Optional, default 10000) the limit on total state count for the determined automaton computed from the regexp.

Example:

To match tokens delimited by simple whitespace characters:

```
<analyzer>
  <tokenizer class="solr.SimplePatternSplitTokenizerFactory" pattern="[ \t\r\n]+"/>
</analyzer>
```

UAX29 URL Email Tokenizer

This tokenizer splits the text field into tokens, treating whitespace and punctuation as delimiters. Delimiter characters are discarded, with the following exceptions:

- Periods (dots) that are not followed by whitespace are kept as part of the token.
- Words are split at hyphens, unless there is a number in the word, in which case the token is not split and the numbers and hyphen(s) are preserved.
- Recognizes and preserves as single tokens the following:
 - Internet domain names containing top-level domains validated against the white list in the [IANA Root Zone Database](#) when the tokenizer was generated
 - email addresses
 - file://, http(s)://, and ftp:// URLs
 - IPv4 and IPv6 addresses

The UAX29 URL Email Tokenizer supports [Unicode standard annex UAX#29](#) word boundaries with the following token types: <ALPHANUM>, <NUM>, <URL>, <EMAIL>, <SOUTHEAST_ASIAN>, <IDEOGRAPHIC>, and <HIRAGANA>.

Factory class: solr.UAX29URLEmailTokenizerFactory

Arguments:

maxTokenLength: (integer, default 255) Solr ignores tokens that exceed the number of characters specified by maxTokenLength.

Example:

```
<analyzer>
  <tokenizer class="solr.UAX29URLEmailTokenizerFactory"/>
</analyzer>
```

In: "Visit <http://accarol.com/contact.htm?from=external&a=10> or e-mail bob.cratchet@accarol.com"

Out: "Visit", "http://accarol.com/contact.htm?from=external&a=10", "or", "e", "mail", "bob.cratchet@accarol.com"

White Space Tokenizer

Simple tokenizer that splits the text stream on whitespace and returns sequences of non-whitespace characters as tokens. Note that any punctuation *will* be included in the tokens.

Factory class: solrWhitespaceTokenizerFactory

Arguments:

rule

Specifies how to define whitespace for the purpose of tokenization. Valid values:

- java: (Default) Uses `Character.isWhitespace(int)`
- unicode: Uses Unicode's WHITESPACE property

Example:

```
<analyzer>
  <tokenizer class="solrWhitespaceTokenizerFactory" rule="java" />
</analyzer>
```

In: "To be, or what?"

Out: "To", "be,", "or", "what?"

OpenNLP Tokenizer and OpenNLP Filters

See [OpenNLP Integration](#) for information about using the OpenNLP Tokenizer, along with information about available OpenNLP token filters.

Filter Descriptions

Filters examine a stream of tokens and keep them, transform them or discard them, depending on the filter type being used.

You configure each filter with a `<filter>` element in `schema.xml` as a child of `<analyzer>`, following the `<tokenizer>` element. Filter definitions should follow a tokenizer or another filter definition because they take a `TokenStream` as input. For example:

```
<fieldType name="text" class="solr.TextField">
  <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>...
  </analyzer>
</fieldType>
```

The class attribute names a factory class that will instantiate a filter object as needed. Filter factory classes must implement the `org.apache.solr.analysis.TokenFilterFactory` interface. Like tokenizers, filters are also instances of `TokenStream` and thus are producers of tokens. Unlike tokenizers, filters also consume tokens from a `TokenStream`. This allows you to mix and match filters, in any order you prefer, downstream of a tokenizer.

Arguments may be passed to tokenizer factories to modify their behavior by setting attributes on the `<filter>` element. For example:

```
<fieldType name="semicolonDelimited" class="solr.TextField">
  <analyzer type="query">
    <tokenizer class="solr.PatternTokenizerFactory" pattern=";" />
    <filter class="solr.LengthFilterFactory" min="2" max="7"/>
  </analyzer>
</fieldType>
```

The following sections describe the filter factories that are included in this release of Solr.

For user tips about Solr's filters, see <http://wiki.apache.org/solr/AnalyzersTokenizersTokenFilters>.

ASCII Folding Filter

This filter converts alphabetic, numeric, and symbolic Unicode characters which are not in the Basic Latin Unicode block (the first 127 ASCII characters) to their ASCII equivalents, if one exists. This filter converts characters from the following Unicode blocks:

- [C1 Controls and Latin-1 Supplement \(PDF\)](#)
- [Latin Extended-A \(PDF\)](#)
- [Latin Extended-B \(PDF\)](#)
- [Latin Extended Additional \(PDF\)](#)

- Latin Extended-C (PDF)
- Latin Extended-D (PDF)
- IPA Extensions (PDF)
- Phonetic Extensions (PDF)
- Phonetic Extensions Supplement (PDF)
- General Punctuation (PDF)
- Superscripts and Subscripts (PDF)
- Enclosed Alphanumerics (PDF)
- Dingbats (PDF)
- Supplemental Punctuation (PDF)
- Alphabetic Presentation Forms (PDF)
- Halfwidth and Fullwidth Forms (PDF)

Factory class: solr.ASCIIFoldingFilterFactory

Arguments:

preserveOriginal

(boolean, default false) If true, the original token is preserved: "thé" -> "the", "thé"

Example:

```
<analyzer>
  <tokenizer class="solrWhitespaceTokenizer"/>
  <filter class="solrASCIIFoldingFilterFactory" preserveOriginal="false" />
</analyzer>
```

In: "á" (Unicode character 00E1)

Out: "a" (ASCII character 97)

Beider-Morse Filter

Implements the Beider-Morse Phonetic Matching (BMPM) algorithm, which allows identification of similar names, even if they are spelled differently or in different languages. More information about how this works is available in the section on [Phonetic Matching](#).



BeiderMorseFilter changed its behavior in Solr 5.0 due to an update to version 3.04 of the BMPM algorithm. Older version of Solr implemented BMPM version 3.00 (see <http://stevemorse.org/phoneticinfo.htm>). Any index built using this filter with earlier versions of Solr will need to be rebuilt.

Factory class: solr.BeiderMorseFilterFactory

Arguments:

nameType

Types of names. Valid values are GENERIC, ASHKENAZI, or SEPHARDIC. If not processing Ashkenazi or Sephardic names, use GENERIC.

ruleType

Types of rules to apply. Valid values are APPROX or EXACT.

concat

Defines if multiple possible matches should be combined with a pipe ("|").

languageSet

The language set to use. The value "auto" will allow the Filter to identify the language, or a comma-separated list can be supplied.

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.BeiderMorseFilterFactory" nameType="GENERIC" ruleType="APPROX" concat="true" languageSet="auto">
    </filter>
</analyzer>
```

Classic Filter

This filter takes the output of the [Classic Tokenizer](#) and strips periods from acronyms and "'s" from possessives.

Factory class: solr.ClassicFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.ClassicTokenizerFactory"/>
  <filter class="solr.ClassicFilterFactory"/>
</analyzer>
```

In: "I.B.M. cat's can't"

Tokenizer to Filter: "I.B.M", "cat's", "can't"

Out: "IBM", "cat", "can't"

Common Grams Filter

This filter creates word shingles by combining common tokens such as stop words with regular tokens. This is useful for creating phrase queries containing common words, such as "the cat." Solr normally ignores

stop words in queried phrases, so searching for "the cat" would return all matches for the word "cat."

Factory class: solr.CommonGramsFilterFactory

Arguments:

words

(a common word file in .txt format) Provide the name of a common word file, such as stopwords.txt.

format

(optional) If the stopwords list has been formatted for Snowball, you can specify format="snowball" so Solr can read the stopwords file.

ignoreCase

(boolean) If true, the filter ignores the case of words when comparing them to the common word file. The default is false.

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.CommonGramsFilterFactory" words="stopwords.txt" ignoreCase="true"/>
</analyzer>
```

In: "the Cat"

Tokenizer to Filter: "the", "Cat"

Out: "the_cat"

Collation Key Filter

Collation allows sorting of text in a language-sensitive way. It is usually used for sorting, but can also be used with advanced searches. We've covered this in much more detail in the section on [Unicode Collation](#).

Daitch-Mokotooff Soundex Filter

Implements the Daitch-Mokotooff Soundex algorithm, which allows identification of similar names, even if they are spelled differently. More information about how this works is available in the section on [Phonetic Matching](#).

Factory class: solr.DaitchMokotooffSoundexFilterFactory

Arguments:

inject

(true/false) If true (the default), then new phonetic tokens are added to the stream. Otherwise, tokens are replaced with the phonetic equivalent. Setting this to false will enable phonetic matching, but the exact spelling of the target word may not match.

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.DaitchMokotoffSoundexFilterFactory" inject="true"/>
</analyzer>
```

Double Metaphone Filter

This filter creates tokens using the DoubleMetaphone encoding algorithm from commons-codec. For more information, see the [Phonetic Matching](#) section.

Factory class: solr.DoubleMetaphoneFilterFactory

Arguments:

inject

(true/false) If true (the default), then new phonetic tokens are added to the stream. Otherwise, tokens are replaced with the phonetic equivalent. Setting this to false will enable phonetic matching, but the exact spelling of the target word may not match.

maxCodeLength

(integer) The maximum length of the code to be generated.

Example:

Default behavior for inject (true): keep the original token and add phonetic token(s) at the same position.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.DoubleMetaphoneFilterFactory"/>
</analyzer>
```

In: "four score and Kuczewski"

Tokenizer to Filter: "four"(1), "score"(2), "and"(3), "Kuczewski"(4)

Out: "four"(1), "FR"(1), "score"(2), "SKR"(2), "and"(3), "ANT"(3), "Kuczewski"(4), "KSSK"(4), "KXFS"(4)

The phonetic tokens have a position increment of 0, which indicates that they are at the same position as the token they were derived from (immediately preceding). Note that "Kuczewski" has two encodings, which are added at the same position.

Example:

Discard original token (inject="false").

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.DoubleMetaphoneFilterFactory" inject="false"/>
</analyzer>
```

In: "four score and Kuczewski"

Tokenizer to Filter: "four"(1), "score"(2), "and"(3), "Kuczewski"(4)

Out: "FR"(1), "SKR"(2), "ANT"(3), "KSSK"(4), "KXFS"(4)

Note that "Kuczewski" has two encodings, which are added at the same position.

Edge N-Gram Filter

This filter generates edge n-gram tokens of sizes within the given range.

Factory class: solr.EdgeNGramFilterFactory

Arguments:

minGramSize

(integer, default 1) The minimum gram size.

maxGramSize

(integer, default 1) The maximum gram size.

Example:

Default behavior.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.EdgeNGramFilterFactory"/>
</analyzer>
```

In: "four score and twenty"

Tokenizer to Filter: "four", "score", "and", "twenty"

Out: "f", "s", "a", "t"

Example:

A range of 1 to 4.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.EdgeNGramFilterFactory" minGramSize="1" maxGramSize="4"/>
</analyzer>
```

In: "four score"

Tokenizer to Filter: "four", "score"

Out: "f", "fo", "fou", "four", "s", "sc", "sco", "scor"

Example:

A range of 4 to 6.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.EdgeNGramFilterFactory" minGramSize="4" maxGramSize="6"/>
</analyzer>
```

In: "four score and twenty"

Tokenizer to Filter: "four", "score", "and", "twenty"

Out: "four", "scor", "score", "twen", "twent", "twenty"

English Minimal Stem Filter

This filter stems plural English words to their singular form.

Factory class: solr.EnglishMinimalStemFilterFactory

Arguments: None

Example:

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.EnglishMinimalStemFilterFactory"/>
</analyzer>
```

In: "dogs cats"

Tokenizer to Filter: "dogs", "cats"

Out: "dog", "cat"

English Possessive Filter

This filter removes singular possessives (trailing '**s**') from words. Note that plural possessives, e.g., the **s'** in "divers' snorkels", are not removed by this filter.

Factory class: solr.EnglishPossessiveFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrEnglishPossessiveFilterFactory"/>
</analyzer>
```

In: "Man's dog bites dogs' man"

Tokenizer to Filter: "Man's", "dog", "bites", "dogs'", "man"

Out: "Man", "dog", "bites", "dogs'", "man"

Fingerprint Filter

This filter outputs a single token which is a concatenation of the sorted and de-duplicated set of input tokens. This can be useful for clustering/linking use cases.

Factory class: solr.FingerprintFilterFactory

Arguments:

separator

The character used to separate tokens combined into the single output token. Defaults to " " (a space character).

maxOutputTokenSize

The maximum length of the summarized output token. If exceeded, no output token is emitted. Defaults to 1024.

Example:

```
<analyzer type="index">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrFingerprintFilterFactory" separator="_" />
</analyzer>
```

In: "the quick brown fox jumped over the lazy dog"

Tokenizer to Filter: "the", "quick", "brown", "fox", "jumped", "over", "the", "lazy", "dog"

Out: "brown_dog_fox_jumped_lazy_over_quick_the"

Flatten Graph Filter

This filter must be included on index-time analyzer specifications that include at least one graph-aware filter, including Synonym Graph Filter and Word Delimiter Graph Filter.

Factory class: solr.FlattenGraphFilterFactory

Arguments:

See the examples below for [Synonym Graph Filter](#) and [Word Delimiter Graph Filter](#).

Hunspell Stem Filter

The Hunspell Stem Filter provides support for several languages. You must provide the dictionary (.dic) and rules (.aff) files for each language you wish to use with the Hunspell Stem Filter. You can download those language files [here](#).

Be aware that your results will vary widely based on the quality of the provided dictionary and rules files. For example, some languages have only a minimal word list with no morphological information. On the other hand, for languages that have no stemmer but do have an extensive dictionary file, the Hunspell stemmer may be a good choice.

Factory class: solr.HunspellStemFilterFactory

Arguments:

dictionary

(required) The path of a dictionary file.

affix

(required) The path of a rules file.

ignoreCase

(boolean) controls whether matching is case sensitive or not. The default is false.

strictAffixParsing

(boolean) controls whether the affix parsing is strict or not. If true, an error while reading an affix rule causes a ParseException, otherwise is ignored. The default is true.

Example:

```
<analyzer type="index">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrHunspellStemFilterFactory"
    dictionary="en_GB.dic"
    affix="en_GB.aff"
    ignoreCase="true"
    strictAffixParsing="true" />
</analyzer>
```

In: "jump jumping jumped"

Tokenizer to Filter: "jump", "jumping", "jumped"

Out: "jump", "jump", "jump"

Hyphenated Words Filter

This filter reconstructs hyphenated words that have been tokenized as two tokens because of a line break or

other intervening whitespace in the field test. If a token ends with a hyphen, it is joined with the following token and the hyphen is discarded.

Note that for this filter to work properly, the upstream tokenizer must not remove trailing hyphen characters. This filter is generally only useful at index time.

Factory class: solr.HyphenatedWordsFilterFactory

Arguments: None

Example:

```
<analyzer type="index">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solr.HyphenatedWordsFilterFactory"/>
</analyzer>
```

In: "A hyphen- ated word"

Tokenizer to Filter: "A", "hyphen-", "ated", "word"

Out: "A", "hyphenated", "word"

ICU Folding Filter

This filter is a custom Unicode normalization form that applies the foldings specified in [Unicode TR #30: Character Foldings](#) in addition to the NFKC_Casefold normalization form as described in [ICU Normalizer 2 Filter](#). This filter is a better substitute for the combined behavior of the [ASCII Folding Filter](#), [Lower Case Filter](#), and [ICU Normalizer 2 Filter](#).

To use this filter, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See `solr/contrib/analysis-extras/README.txt` for instructions on which jars you need to add.

Factory class: solr.ICUFoldingFilterFactory

Arguments:

filter

(string, optional) A Unicode set filter that can be used to e.g., exclude a set of characters from being processed. See the [UnicodeSet javadocs](#) for more information.

Example without a filter:

```
<analyzer>
  <tokenizer class="solrStandardTokenizerFactory"/>
  <filter class="solrICUFoldingFilterFactory"/>
</analyzer>
```

Example with a filter to exclude Swedish/Finnish characters:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ICUFoldingFilterFactory" filter="[^åäöÅÄÖ]" />
</analyzer>
```

For detailed information on this normalization form, see [Unicode TR #30: Character Foldings](#).

ICU Normalizer 2 Filter

This filter factory normalizes text according to one of five Unicode Normalization Forms as described in [Unicode Standard Annex #15](#):

- NFC: (name="nfc" mode="compose") Normalization Form C, canonical decomposition
- NFD: (name="nfc" mode="decompose") Normalization Form D, canonical decomposition, followed by canonical composition
- NFKC: (name="nfkc" mode="compose") Normalization Form KC, compatibility decomposition
- NFKD: (name="nfkc" mode="decompose") Normalization Form KD, compatibility decomposition, followed by canonical composition
- NFKC_Casefold: (name="nfkc_cf" mode="compose") Normalization Form KC, with additional Unicode case folding. Using the ICU Normalizer 2 Filter is a better-performing substitution for the [Lower Case Filter](#) and NFKC normalization.

Factory class: solr.ICUNormalizer2FilterFactory

Arguments:

name

The name of the normalization form. Valid options are nfc, nfd, nfkc, nfkd, or nfkc_cf (the default). Required.

mode

The mode of Unicode character composition and decomposition. Valid options are: compose (the default) or decompose. Required.

filter

A Unicode set filter that can be used to e.g., exclude a set of characters from being processed. See the [UnicodeSet javadocs](#) for more information. Optional.

Example with NFKC_Casefold:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ICUNormalizer2FilterFactory" name="nfkc_cf" mode="compose" />
</analyzer>
```

Example with a filter to exclude Swedish/Finnish characters:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ICUNormalizer2FilterFactory" name="nfkc_cf" mode="compose" filter=
  "[^åäöÅÄÖ]"/>
</analyzer>
```

For detailed information about these normalization forms, see [Unicode Normalization Forms](#).

To use this filter, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See solr/contrib/analysis-extras/README.txt for instructions on which jars you need to add.

ICU Transform Filter

This filter applies [ICU Transforms](#) to text. This filter supports only ICU System Transforms. Custom rule sets are not supported.

Factory class: solr.ICYTransformFilterFactory

Arguments:

id

(string) The identifier for the ICU System Transform you wish to apply with this filter. For a full list of ICU System Transforms, see http://demo.icu-project.org/icu-bin/translit?TEMPLATE_FILE=data/translit_rule_main.html.

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ICYTransformFilterFactory" id="Traditional-Simplified"/>
</analyzer>
```

For detailed information about ICU Transforms, see <http://userguide.icu-project.org/transforms/general>.

To use this filter, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See solr/contrib/analysis-extras/README.txt for instructions on which jars you need to add.

Keep Word Filter

This filter discards all tokens except those that are listed in the given word list. This is the inverse of the Stop Words Filter. This filter can be useful for building specialized indices for a constrained set of terms.

Factory class: solr.KeepWordFilterFactory

Arguments:

words

(required) Path of a text file containing the list of keep words, one per line. Blank lines and lines that begin with "#" are ignored. This may be an absolute path, or a simple filename in the Solr conf directory.

ignoreCase

(true/false) If **true** then comparisons are done case-insensitively. If this argument is true, then the words file is assumed to contain only lowercase words. The default is **false**.

enablePositionIncrements

if luceneMatchVersion is 4.3 or earlier and enablePositionIncrements="false", no position holes will be left by this filter when it removes tokens. **This argument is invalid if luceneMatchVersion is 5.0 or later.**

Example:

Where keepwords.txt contains:

happy funny silly

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.KeepWordFilterFactory" words="keepwords.txt"/>
</analyzer>
```

In: "Happy, sad or funny"

Tokenizer to Filter: "Happy", "sad", "or", "funny"

Out: "funny"

Example:

Same keepwords.txt, case insensitive:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.KeepWordFilterFactory" words="keepwords.txt" ignoreCase="true"/>
</analyzer>
```

In: "Happy, sad or funny"

Tokenizer to Filter: "Happy", "sad", "or", "funny"

Out: "Happy", "funny"

Example:

Using LowerCaseFilterFactory before filtering for keep words, no ignoreCase flag.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.KeepWordFilterFactory" words="keepwords.txt"/>
</analyzer>
```

In: "Happy, sad or funny"

Tokenizer to Filter: "Happy", "sad", "or", "funny"

Filter to Filter: "happy", "sad", "or", "funny"

Out: "happy", "funny"

KStem Filter

KStem is an alternative to the Porter Stem Filter for developers looking for a less aggressive stemmer. KStem was written by Bob Krovetz, ported to Lucene by Sergio Guzman-Lara (UMASS Amherst). This stemmer is only appropriate for English language text.

Factory class: solr.KStemFilterFactory

Arguments: None

Example:

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory "/>
  <filter class="solr.KStemFilterFactory"/>
</analyzer>
```

In: "jump jumping jumped"

Tokenizer to Filter: "jump", "jumping", "jumped"

Out: "jump", "jump", "jump"

Length Filter

This filter passes tokens whose length falls within the min/max limit specified. All other tokens are discarded.

Factory class: solr.LengthFilterFactory

Arguments:

min

(integer, required) Minimum token length. Tokens shorter than this are discarded.

max

(integer, required, must be \geq min) Maximum token length. Tokens longer than this are discarded.

enablePositionIncrements

if luceneMatchVersion is 4.3 or earlier and enablePositionIncrements="false", no position holes will be left by this filter when it removes tokens. **This argument is invalid if luceneMatchVersion is 5.0 or later.**

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LengthFilterFactory" min="3" max="7"/>
</analyzer>
```

In: "turn right at Albuquerque"

Tokenizer to Filter: "turn", "right", "at", "Albuquerque"

Out: "turn", "right"

Limit Token Count Filter

This filter limits the number of accepted tokens, typically useful for index analysis.

By default, this filter ignores any tokens in the wrapped TokenStream once the limit has been reached, which can result in reset() being called prior to incrementToken() returning false. For most TokenStream implementations this should be acceptable, and faster than consuming the full stream. If you are wrapping a TokenStream which requires that the full stream of tokens be exhausted in order to function properly, use the consumeAllTokens="true" option.

Factory class: solr.LimitTokenCountFilterFactory

Arguments:

maxTokenCount

(integer, required) Maximum token count. After this limit has been reached, tokens are discarded.

consumeAllTokens

(boolean, defaults to false) Whether to consume (and discard) previous token filters' tokens after the maximum token count has been reached. See description above.

Example:

```
<analyzer type="index">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solr.LimitTokenCountFilterFactory" maxTokenCount="10"
         consumeAllTokens="false" />
</analyzer>
```

In: "1 2 3 4 5 6 7 8 9 10 11 12"

Tokenizer to Filter: "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12"

Out: "1", "2", "3", "4", "5", "6", "7", "8", "9", "10"

Limit Token Offset Filter

This filter limits tokens to those before a configured maximum start character offset. This can be useful to limit highlighting, for example.

By default, this filter ignores any tokens in the wrapped TokenStream once the limit has been reached, which can result in `reset()` being called prior to `incrementToken()` returning false. For most TokenStream implementations this should be acceptable, and faster than consuming the full stream. If you are wrapping a TokenStream which requires that the full stream of tokens be exhausted in order to function properly, use the `consumeAllTokens="true"` option.

Factory class: `solr.LimitTokenOffsetFilterFactory`

Arguments:

`maxStartOffset`

(integer, required) Maximum token start character offset. After this limit has been reached, tokens are discarded.

`consumeAllTokens`

(boolean, defaults to false) Whether to consume (and discard) previous token filters' tokens after the maximum start offset has been reached. See description above.

Example:

```
<analyzer>
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solr.LimitTokenOffsetFilterFactory" maxStartOffset="10"
         consumeAllTokens="false" />
</analyzer>
```

In: "0 2 4 6 8 A C E"

Tokenizer to Filter: "0", "2", "4", "6", "8", "A", "C", "E"

Out: "0", "2", "4", "6", "8", "A"

Limit Token Position Filter

This filter limits tokens to those before a configured maximum token position.

By default, this filter ignores any tokens in the wrapped TokenStream once the limit has been reached, which can result in `reset()` being called prior to `incrementToken()` returning false. For most TokenStream implementations this should be acceptable, and faster than consuming the full stream. If you are wrapping a TokenStream which requires that the full stream of tokens be exhausted in order to function properly, use the `consumeAllTokens="true"` option.

Factory class: solr.LimitTokenPositionFilterFactory

Arguments:

maxTokenPosition

(integer, required) Maximum token position. After this limit has been reached, tokens are discarded.

consumeAllTokens

(boolean, defaults to false) Whether to consume (and discard) previous token filters' tokens after the maximum start offset has been reached. See description above.

Example:

```
<analyzer>
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrLimitTokenPositionFilterFactory" maxTokenPosition="3"
         consumeAllTokens="false" />
</analyzer>
```

In: "1 2 3 4 5"

Tokenizer to Filter: "1", "2", "3", "4", "5"

Out: "1", "2", "3"

Lower Case Filter

Converts any uppercase letters in a token to the equivalent lowercase token. All other characters are left unchanged.

Factory class: solrLowerCaseFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solrStandardTokenizerFactory"/>
  <filter class="solrLowerCaseFilterFactory" />
</analyzer>
```

In: "Down With CamelCase"

Tokenizer to Filter: "Down", "With", "CamelCase"

Out: "down", "with", "camelcase"

Managed Stop Filter

This is specialized version of the [Stop Words Filter Factory](#) that uses a set of stop words that are [managed](#)

from a REST API.

Arguments:

managed

The name that should be used for this set of stop words in the managed REST API.

Example: With this configuration the set of words is named "english" and can be managed via /solr/collection_name/schema/analysis/stopwords/english

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ManagedStopFilterFactory" managed="english"/>
</analyzer>
```

See [Stop Filter](#) for example input/output.

Managed Synonym Filter

This is specialized version of the [Synonym Filter](#) that uses a mapping on synonyms that is [managed from a REST API](#).



Managed Synonym Filter has been Deprecated

Managed Synonym Filter has been deprecated in favor of Managed Synonym Graph Filter, which is required for multi-term synonym support.

Factory class: solr.ManagedSynonymFilterFactory

For arguments and examples, see the [Synonym Graph Filter](#) below.

Managed Synonym Graph Filter

This is specialized version of the [Synonym Graph Filter](#) that uses a mapping on synonyms that is [managed from a REST API](#).

This filter maps single- or multi-token synonyms, producing a fully correct graph output. This filter is a replacement for the Managed Synonym Filter, which produces incorrect graphs for multi-token synonyms.



Although this filter produces correct token graphs, it cannot consume an input token graph correctly.

Arguments:

managed

The name that should be used for this mapping on synonyms in the managed REST API.

Example: With this configuration the set of mappings is named "english" and can be managed via /solr/collection_name/schema/analysis/synonyms/english

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ManagedSynonymGraphFilterFactory" managed="english"/>
  <filter class="solr.FlattenGraphFilterFactory" /> <!-- required on index analyzers after graph
filters -->
</analyzer>
<analyzer type="query">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ManagedSynonymGraphFilterFactory" managed="english"/>
</analyzer>
```

See [Synonym Graph Filter](#) below for example input/output.

N-Gram Filter

Generates n-gram tokens of sizes in the given range. Note that tokens are ordered by position and then by gram size.

Factory class: solr.NGramFilterFactory

Arguments:

minGramSize

(integer, default 1) The minimum gram size.

maxGramSize

(integer, default 2) The maximum gram size.

Example:

Default behavior.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.NGramFilterFactory"/>
</analyzer>
```

In: "four score"

Tokenizer to Filter: "four", "score"

Out: "f", "o", "u", "r", "fo", "ou", "ur", "s", "c", "o", "r", "e", "sc", "co", "or", "re"

Example:

A range of 1 to 4.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.NGramFilterFactory" minGramSize="1" maxGramSize="4"/>
</analyzer>
```

In: "four score"

Tokenizer to Filter: "four", "score"

Out: "f", "fo", "fou", "four", "o", "ou", "our", "u", "ur", "r", "s", "sc", "sco", "scor", "c", "co", "cor", "core", "o", "or", "ore", "r", "re", "e"

Example:

A range of 3 to 5.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.NGramFilterFactory" minGramSize="3" maxGramSize="5"/>
</analyzer>
```

In: "four score"

Tokenizer to Filter: "four", "score"

Out: "fou", "four", "our", "sco", "scor", "score", "cor", "core", "ore"

Numeric Payload Token Filter

This filter adds a numeric floating point payload value to tokens that match a given type. Refer to the Javadoc for the org.apache.lucene.analysis.Token class for more information about token types and payloads.

Factory class: solr.NumericPayloadTokenFilterFactory

Arguments:

payload

(required) A floating point value that will be added to all matching tokens.

typeMatch

(required) A token type name string. Tokens with a matching type name will have their payload set to the above floating point value.

Example:

```
<analyzer>
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrNumericPayloadTokenFilterFactory" payload="0.75" typeMatch="word"/>
</analyzer>
```

In: "bing bang boom"

Tokenizer to Filter: "bing", "bang", "boom"

Out: "bing"[0.75], "bang"[0.75], "boom"[0.75]

Pattern Replace Filter

This filter applies a regular expression to each token and, for those that match, substitutes the given replacement string in place of the matched pattern. Tokens which do not match are passed through unchanged.

Factory class: solr.PatternReplaceFilterFactory

Arguments:

pattern

(required) The regular expression to test against each token, as per java.util.regex.Pattern.

replacement

(required) A string to substitute in place of the matched pattern. This string may contain references to capture groups in the regex pattern. See the Javadoc for java.util.regex.Matcher.

replace

("all" or "first", default "all") Indicates whether all occurrences of the pattern in the token should be replaced, or only the first.

Example:

Simple string replace:

```
<analyzer>
  <tokenizer class="solrStandardTokenizerFactory"/>
  <filter class="solrPatternReplaceFilterFactory" pattern="cat" replacement="dog"/>
</analyzer>
```

In: "cat concatenate catcat"

Tokenizer to Filter: "cat", "concatenate", "catcat"

Out: "dog", "condogenate", "dogydog"

Example:

String replacement, first occurrence only:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.PatternReplaceFilterFactory" pattern="cat" replacement="dog" replace="first"/>
</analyzer>
```

In: "cat concatenate catycat"

Tokenizer to Filter: "cat", "concatenate", "catycat"

Out: "dog", "condogenate", "dogycat"

Example:

More complex pattern with capture group reference in the replacement. Tokens that start with non-numeric characters and end with digits will have an underscore inserted before the numbers. Otherwise the token is passed through.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.PatternReplaceFilterFactory" pattern="(\D+)(\d+)$" replacement="$1_$2"/>
</analyzer>
```

In: "cat foo1234 9987 blah1234foo"

Tokenizer to Filter: "cat", "foo1234", "9987", "blah1234foo"

Out: "cat", "foo_1234", "9987", "blah1234foo"

Phonetic Filter

This filter creates tokens using one of the phonetic encoding algorithms in the org.apache.commons.codec.language package. For more information, see the section on [Phonetic Matching](#).

Factory class: solr.PhoneticFilterFactory

Arguments:

encoder

(required) The name of the encoder to use. The encoder name must be one of the following (case insensitive): [DoubleMetaphone](#), [Metaphone](#), [Soundex](#), [RefinedSoundex](#), [Caverphone](#) (v2.0), [ColognePhonetic](#), or [Nysiis](#).

inject

(true/false) If true (the default), then new phonetic tokens are added to the stream. Otherwise, tokens are replaced with the phonetic equivalent. Setting this to false will enable phonetic matching, but the exact spelling of the target word may not match.

maxCodeLength

(integer) The maximum length of the code to be generated by the Metaphone or Double Metaphone encoders.

Example:

Default behavior for DoubleMetaphone encoding.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.PhoneticFilterFactory" encoder="DoubleMetaphone"/>
</analyzer>
```

In: "four score and twenty"

Tokenizer to Filter: "four"(1), "score"(2), "and"(3), "twenty"(4)

Out: "four"(1), "FR"(1), "score"(2), "SKR"(2), "and"(3), "ANT"(3), "twenty"(4), "TNT"(4)

The phonetic tokens have a position increment of 0, which indicates that they are at the same position as the token they were derived from (immediately preceding).

Example:

Discard original token.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.PhoneticFilterFactory" encoder="DoubleMetaphone" inject="false"/>
</analyzer>
```

In: "four score and twenty"

Tokenizer to Filter: "four"(1), "score"(2), "and"(3), "twenty"(4)

Out: "FR"(1), "SKR"(2), "ANT"(3), "TWNT"(4)

Example:

Default Soundex encoder.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.PhoneticFilterFactory" encoder="Soundex"/>
</analyzer>
```

In: "four score and twenty"

Tokenizer to Filter: "four"(1), "score"(2), "and"(3), "twenty"(4)

Out: "four"(1), "F600"(1), "score"(2), "S600"(2), "and"(3), "A530"(3), "twenty"(4), "T530"(4)

Porter Stem Filter

This filter applies the Porter Stemming Algorithm for English. The results are similar to using the Snowball Porter Stemmer with the language="English" argument. But this stemmer is coded directly in Java and is not based on Snowball. It does not accept a list of protected words and is only appropriate for English language text. However, it has been benchmarked as [four times faster](#) than the English Snowball stemmer, so can provide a performance enhancement.

Factory class: solr.PorterStemFilterFactory

Arguments: None

Example:

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory" />
  <filter class="solr.PorterStemFilterFactory" />
</analyzer>
```

In: "jump jumping jumped"

Tokenizer to Filter: "jump", "jumping", "jumped"

Out: "jump", "jump", "jump"

Protected Term Filter

This filter enables a form of conditional filtering: it only applies its wrapped filters to terms that are **not contained** in a protected set.

Factory class: solr.ProtectedTermFilterFactory

Arguments:

protected

(required) Comma-separated list of files containing protected terms, one per line.

wrappedFilters

(required) Case-insensitive comma-separated list of TokenFilterFactory SPI names (strip trailing (Token)FilterFactory from the factory name - see the [java.util.ServiceLoader interface](#)). Each filter name must be unique, so if you need to specify the same filter more than once, you must add case-insensitive unique -id suffixes to each same-SPI-named filter (note that the -id suffix is stripped prior to SPI lookup).

ignoreCase

(true/false, default false) Ignore case when testing for protected words. If true, the protected list should contain lowercase words.

Example:

All terms except those in protectedTerms.txt are truncated at 4 characters and lowercased:

```
<analyzer>
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrProtectedTermFilterFactory"
    ignoreCase="true" protected="protectedTerms.txt"
    wrappedFilters="truncate,lowercase"
    truncate.prefixLength="4"/>
</analyzer>
```

Example:

This example includes multiple same-named wrapped filters with unique -id suffixes. Note that both the filter SPI names and -id suffixes are treated case-insensitively.

For all terms except those in protectedTerms.txt, synonyms are added, terms are reversed, and then synonyms are added for the reversed terms:

```
<analyzer type="query">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrProtectedTermFilterFactory"
    ignoreCase="true" protected="protectedTerms.txt"
    wrappedFilters="SynonymGraph-fwd,ReverseString,SynonymGraph-rev"
    synonymgraph-FWD.synonyms="fwd-syns.txt"
    synonymgraph-FWD.synonyms="rev-syns.txt"/>
</analyzer>
```

Remove Duplicates Token Filter

The filter removes duplicate tokens in the stream. Tokens are considered to be duplicates ONLY if they have the same text and position values.

Because positions must be the same, this filter might not do what a user expects it to do based on its name. It is a very specialized filter that is only useful in very specific circumstances. It has been so named for brevity, even though it is potentially misleading.

Factory class: solr.RemoveDuplicatesTokenFilterFactory

Arguments: None

Example:

One example of where RemoveDuplicatesTokenFilterFactory is useful in situations where a synonym file is being used in conjunction with a stemmer. In these situations, both the stemmer and the synonym filter can cause completely identical terms with the same positions to end up in the stream, increasing index size with no benefit.

Consider the following entry from a synonyms.txt file:

```
Television, Televisions, TV, TVs
```

When used in the following configuration:

```
<analyzer type="query">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.SynonymGraphFilterFactory" synonyms="synonyms.txt"/>
  <filter class="solr.EnglishMinimalStemFilterFactory"/>
  <filter class="solr.RemoveDuplicatesTokenFilterFactory"/>
</analyzer>
```

In: "Watch TV"

Tokenizer to Synonym Filter: "Watch"(1) "TV"(2)

Synonym Filter to Stem Filter: "Watch"(1) "Television"(2) "Televisions"(2) "TV"(2) "TVs"(2)

Stem Filter to Remove Dups Filter: "Watch"(1) "Television"(2) "Television"(2) "TV"(2) "TV"(2)

Out: "Watch"(1) "Television"(2) "TV"(2)

Reversed Wildcard Filter

This filter reverses tokens to provide faster leading wildcard and prefix queries. Tokens without wildcards are not reversed.

Factory class: solr.ReversedWildcardFilterFactory

Arguments:

withOriginal

(boolean) If true, the filter produces both original and reversed tokens at the same positions. If false, produces only reversed tokens.

maxPosAsterisk

(integer, default = 2) The maximum position of the asterisk wildcard ('*') that triggers the reversal of the query term. Terms with asterisks at positions above this value are not reversed.

maxPosQuestion

(integer, default = 1) The maximum position of the question mark wildcard ('?') that triggers the reversal of query term. To reverse only pure suffix queries (queries with a single leading asterisk), set this to 0 and maxPosAsterisk to 1.

maxFractionAsterisk

(float, default = 0.0) An additional parameter that triggers the reversal if asterisk ('*') position is less than this fraction of the query token length.

minTrailing

(integer, default = 2) The minimum number of trailing characters in a query token after the last wildcard character. For good performance this should be set to a value larger than 1.

Example:

```
<analyzer type="index">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrReversedWildcardFilterFactory" withOriginal="true"
    maxPosAsterisk="2" maxPosQuestion="1" minTrailing="2" maxFractionAsterisk="0"/>
</analyzer>
```

In: "*foo *bar"

Tokenizer to Filter: "*foo", "*bar"

Out: "oof*", "rab*"

Shingle Filter

This filter constructs shingles, which are token n-grams, from the token stream. It combines runs of tokens into a single token.

Factory class: solr.ShingleFilterFactory

Arguments:

minShingleSize

(integer, must be >= 2, default 2) The minimum number of tokens per shingle.

maxShingleSize

(integer, must be >= minShingleSize, default 2) The maximum number of tokens per shingle.

outputUnigrams

(boolean, default true) If true, then each individual token is also included at its original position.

outputUnigramsIfNoShingles

(boolean, default false) If true, then individual tokens will be output if no shingles are possible.

tokenSeparator

(string, default is " ") The string to use when joining adjacent tokens to form a shingle.

Example:

Default behavior.

```
<analyzer>
  <tokenizer class="solrStandardTokenizerFactory"/>
  <filter class="solrShingleFilterFactory"/>
</analyzer>
```

In: "To be, or what?"

Tokenizer to Filter: "To"(1), "be"(2), "or"(3), "what"(4)

Out: "To"(1), "To be"(1), "be"(2), "be or"(2), "or"(3), "or what"(3), "what"(4)

Example:

A shingle size of four, do not include original token.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ShingleFilterFactory" maxShingleSize="4" outputUnigrams="false"/>
</analyzer>
```

In: "To be, or not to be."

Tokenizer to Filter: "To"(1), "be"(2), "or"(3), "not"(4), "to"(5), "be"(6)

Out: "To be"(1), "To be or"(1), "To be or not"(1), "be or"(2), "be or not"(2), "be or not to"(2), "or not"(3), "or not to"(3), "or not to be"(3), "not to"(4), "not to be"(4), "to be"(5)

Snowball Porter Stemmer Filter

This filter factory instantiates a language-specific stemmer generated by Snowball. Snowball is a software package that generates pattern-based word stemmers. This type of stemmer is not as accurate as a table-based stemmer, but is faster and less complex. Table-driven stemmers are labor intensive to create and maintain and so are typically commercial products.

Solr contains Snowball stemmers for Armenian, Basque, Catalan, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, Swedish and Turkish. For more information on Snowball, visit <http://snowball.tartarus.org/>.

StopFilterFactory, CommonGramsFilterFactory, and CommonGramsQueryFilterFactory can optionally read stopwords in Snowball format (specify format="snowball" in the configuration of those FilterFactories).

Factory class: solr.SnowballPorterFilterFactory

Arguments:

language

(default "English") The name of a language, used to select the appropriate Porter stemmer to use. Case is significant. This string is used to select a package name in the org.tartarus.snowball.ext class hierarchy.

protected

Path of a text file containing a list of protected words, one per line. Protected words will not be stemmed. Blank lines and lines that begin with "#" are ignored. This may be an absolute path, or a simple file name in the Solr conf directory.

Example:

Default behavior:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.SnowballPorterFilterFactory"/>
</analyzer>
```

In: "flip flipped flipping"

Tokenizer to Filter: "flip", "flipped", "flipping"

Out: "flip", "flip", "flip"

Example:

French stemmer, English words:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.SnowballPorterFilterFactory" language="French"/>
</analyzer>
```

In: "flip flipped flipping"

Tokenizer to Filter: "flip", "flipped", "flipping"

Out: "flip", "flipped", "flipping"

Example:

Spanish stemmer, Spanish words:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.SnowballPorterFilterFactory" language="Spanish"/>
</analyzer>
```

In: "cante canta"

Tokenizer to Filter: "cante", "canta"

Out: "cant", "cant"

Stop Filter

This filter discards, or *stops* analysis of, tokens that are on the given stop words list. A standard stop words list is included in the Solr conf directory, named stopwords.txt, which is appropriate for typical English language text.

Factory class: solr.StopFilterFactory

Arguments:

words

(optional) The path to a file that contains a list of stop words, one per line. Blank lines and lines that begin with "#" are ignored. This may be an absolute path, or path relative to the Solr conf directory.

format

(optional) If the stopwords list has been formatted for Snowball, you can specify format="snowball" so Solr can read the stopwords file.

ignoreCase

(true/false, default false) Ignore case when testing for stop words. If true, the stop list should contain lowercase words.

enablePositionIncrements

if luceneMatchVersion is 4.4 or earlier and enablePositionIncrements="false", no position holes will be left by this filter when it removes tokens. **This argument is invalid if luceneMatchVersion is 5.0 or later.**

Example:

Case-sensitive matching, capitalized words not stopped. Token positions skip stopped words.

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.StopFilterFactory" words="stopwords.txt"/>
</analyzer>
```

In: "To be or what?"

Tokenizer to Filter: "To"(1), "be"(2), "or"(3), "what"(4)

Out: "To"(1), "what"(4)

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>
</analyzer>
```

In: "To be or what?"

Tokenizer to Filter: "To"(1), "be"(2), "or"(3), "what"(4)

Out: "what"(4)

Suggest Stop Filter

Like [Stop Filter](#), this filter discards, or *stops* analysis of, tokens that are on the given stop words list.

Suggest Stop Filter differs from Stop Filter in that it will not remove the last token unless it is followed by a

token separator. For example, a query "find the" would preserve the 'the' since it was not followed by a space, punctuation, etc., and mark it as a KEYWORD so that following filters will not change or remove it.

By contrast, a query like "find the popsicle" would remove 'the' as a stopword, since it's followed by a space. When using one of the analyzing suggesters, you would normally use the ordinary StopFilterFactory in your index analyzer and then SuggestStopFilter in your query analyzer.

Factory class: solr.SuggestStopFilterFactory

Arguments:

words

(optional; default: StopAnalyzer#ENGLISH_STOP_WORDS_SET) The name of a stopwords file to parse.

format

(optional; default: wordset) Defines how the words file will be parsed. If words is not specified, then format must not be specified. The valid values for the format option are:

wordset

This is the default format, which supports one word per line (including any intra-word whitespace) and allows whole line comments beginning with the # character. Blank lines are ignored.

snowball

This format allows for multiple words specified on each line, and trailing comments may be specified using the vertical line (|). Blank lines are ignored.

ignoreCase

(optional; default: **false**) If **true**, matching is case-insensitive.

Example:

```
<analyzer type="query">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrLowerCaseFilterFactory"/>
  <filter class="solrSuggestStopFilterFactory" ignoreCase="true"
         words="stopwords.txt" format="wordset"/>
</analyzer>
```

In: "The The"

Tokenizer to Filter: "the"(1), "the"(2)

Out: "the"(2)

Synonym Filter

This filter does synonym mapping. Each token is looked up in the list of synonyms and if a match is found, then the synonym is emitted in place of the token. The position value of the new tokens are set such they all occur at the same position as the original token.



Synonym Filter has been Deprecated

Synonym Filter has been deprecated in favor of Synonym Graph Filter, which is required for multi-term synonym support.

Factory class: solr.SynonymFilterFactory

For arguments and examples, see the Synonym Graph Filter below.

Synonym Graph Filter

This filter maps single- or multi-token synonyms, producing a fully correct graph output. This filter is a replacement for the Synonym Filter, which produces incorrect graphs for multi-token synonyms.

If you use this filter during indexing, you must follow it with a Flatten Graph Filter to squash tokens on top of one another like the Synonym Filter, because the indexer can't directly consume a graph. To get fully correct positional queries when your synonym replacements are multiple tokens, you should instead apply synonyms using this filter at query time.



Although this filter produces correct token graphs, it cannot consume an input token graph correctly.

Factory class: solr.SynonymGraphFilterFactory

Arguments:

synonyms

(required) The path of a file that contains a list of synonyms, one per line. In the (default) solr format - see the format argument below for alternatives - blank lines and lines that begin with "#" are ignored. This may be a comma-separated list of paths. See [Resource and Plugin Loading](#) for more information.

There are two ways to specify synonym mappings:

- A comma-separated list of words. If the token matches any of the words, then all the words in the list are substituted, which will include the original token.
- Two comma-separated lists of words with the symbol ">" between them. If the token matches any word on the left, then the list on the right is substituted. The original token will not be included unless it is also in the list on the right.

ignoreCase

(optional; default: false) If true, synonyms will be matched case-insensitively.

expand

(optional; default: true) If true, a synonym will be expanded to all equivalent synonyms. If false, all equivalent synonyms will be reduced to the first in the list.

format

(optional; default: solr) Controls how the synonyms will be parsed. The short names solr (for SolrSynonymParser) and wordnet (for WordnetSynonymParser) are supported, or you may alternatively supply the name of your own SynonymMap.Builder subclass.

tokenizerFactory

(optional; default: WhitespaceTokenizerFactory) The name of the tokenizer factory to use when parsing the synonyms file. Arguments with the name prefix tokenizerFactory.* will be supplied as init params to the specified tokenizer factory.

Any arguments not consumed by the synonym filter factory, including those without the tokenizerFactory.* prefix, will also be supplied as init params to the tokenizer factory.

If tokenizerFactory is specified, then analyzer may not be, and vice versa.

analyzer

(optional; default: WhitespaceTokenizerFactory) The name of the analyzer class to use when parsing the synonyms file. If analyzer is specified, then tokenizerFactory may not be, and vice versa.

For the following examples, assume a synonyms file named mysynonyms.txt:

```
couch,sofa,divan  
teh => the  
huge,ginormous,humungous => large  
small => tiny,teeny,weeny
```

Example:

```
<analyzer type="index">  
  <tokenizer class="solr.StandardTokenizerFactory"/>  
  <filter class="solr.SynonymGraphFilterFactory" synonyms="mysynonyms.txt"/>  
  <filter class="solr.FlattenGraphFilterFactory"/> <!-- required on index analyzers after graph  
filters -->  
</analyzer>  
<analyzer type="query">  
  <tokenizer class="solr.StandardTokenizerFactory"/>  
  <filter class="solr.SynonymGraphFilterFactory" synonyms="mysynonyms.txt"/>  
</analyzer>
```

In: "teh small couch"

Tokenizer to Filter: "teh"(1), "small"(2), "couch"(3)

Out: "the"(1), "tiny"(2), "teeny"(2), "weeny"(2), "couch"(3), "sofa"(3), "divan"(3)

Example:

In: "teh ginormous, humungous sofa"

Tokenizer to Filter: "teh"(1), "ginormous"(2), "humungous"(3), "sofa"(4)

Out: "the"(1), "large"(2), "large"(3), "couch"(4), "sofa"(4), "divan"(4)

Token Offset Payload Filter

This filter adds the numeric character offsets of the token as a payload value for that token.

Factory class: solr.TokenOffsetPayloadTokenFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solr.TokenOffsetPayloadTokenFilterFactory"/>
</analyzer>
```

In: "bing bang boom"

Tokenizer to Filter: "bing", "bang", "boom"

Out: "bing"[0,4], "bang"[5,9], "boom"[10,14]

Trim Filter

This filter trims leading and/or trailing whitespace from tokens. Most tokenizers break tokens at whitespace, so this filter is most often used for special situations.

Factory class: solr.TrimFilterFactory

Arguments:

updateOffsets

if luceneMatchVersion is 4.3 or earlier and updateOffsets="true", trimmed tokens' start and end offsets will be updated to those of the first and last characters (plus one) remaining in the token. **This argument is invalid if luceneMatchVersion is 5.0 or later.**

Example:

The PatternTokenizerFactory configuration used here splits the input on simple commas, it does not remove whitespace.

```
<analyzer>
  <tokenizer class="solr.PatternTokenizerFactory" pattern=", "/>
  <filter class="solr.TrimFilterFactory"/>
</analyzer>
```

In: "one, two , three ,four "

Tokenizer to Filter: "one", "two", "three", "four"

Out: "one", "two", "three", "four"

Type As Payload Filter

This filter adds the token's type, as an encoded byte sequence, as its payload.

Factory class: solr.TypeAsPayloadTokenFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solr.TypeAsPayloadTokenFilterFactory"/>
</analyzer>
```

In: "Pay Bob's I.O.U."

Tokenizer to Filter: "Pay", "Bob's", "I.O.U."

Out: "Pay"[<ALPHANUM>], "Bob's"[<APOSTROPHE>], "I.O.U." [<ACRONYM>]

Type As Synonym Filter

This filter adds the token's type, as a token at the same position as the token, optionally with a configurable prefix prepended.

Factory class: solr.TypeAsSynonymFilterFactory

Arguments:

prefix

(optional) The prefix to prepend to the token's type.

Examples:

With the example below, each token's type will be emitted verbatim at the same position:

```
<analyzer>
  <tokenizer class="solrStandardTokenizerFactory"/>
  <filter class="solr.TypeAsSynonymFilterFactory"/>
</analyzer>
```

With the example below, for a token "example.com" with type <URL>, the token emitted at the same position will be "_type_<URL>":

```
<analyzer>
  <tokenizer class="solrUAX29URLEmailTokenizerFactory"/>
  <filter class="solr.TypeAsSynonymFilterFactory" prefix="_type_"/>
</analyzer>
```

Type Token Filter

This filter blacklists or whitelists a specified list of token types, assuming the tokens have type metadata associated with them. For example, the [UAX29 URL Email Tokenizer](#) emits "<URL>" and "<EMAIL>" typed tokens, as well as other types. This filter would allow you to pull out only e-mail addresses from text as tokens, if you wish.

Factory class: solr.TypeTokenFilterFactory

Arguments:

types

Defines the location of a file of types to filter.

useWhitelist

If **true**, the file defined in types should be used as include list. If **false**, or undefined, the file defined in types is used as a blacklist.

enablePositionIncrements

if luceneMatchVersion is 4.3 or earlier and enablePositionIncrements="false", no position holes will be left by this filter when it removes tokens. **This argument is invalid if luceneMatchVersion is 5.0 or later.**

Example:

```
<analyzer>
  <filter class="solr.TypeTokenFilterFactory" types="stotypes.txt" useWhitelist="true"/>
</analyzer>
```

Word Delimiter Filter

This filter splits tokens at word delimiters.



Word Delimiter Filter has been Deprecated

Word Delimiter Filter has been deprecated in favor of Word Delimiter Graph Filter, which is required to produce a correct token graph so that e.g., phrase queries can work correctly.

Factory class: solr.WordDelimiterFilterFactory

For a full description, including arguments and examples, see the Word Delimiter Graph Filter below.

Word Delimiter Graph Filter

This filter splits tokens at word delimiters.

If you use this filter during indexing, you must follow it with a Flatten Graph Filter to squash tokens on top of one another like the Word Delimiter Filter, because the indexer can't directly consume a graph. To get fully correct positional queries when tokens are split, you should instead use this filter at query time.

Note: although this filter produces correct token graphs, it cannot consume an input token graph correctly.

The rules for determining delimiters are determined as follows:

- A change in case within a word: "CamelCase" -> "Camel", "Case". This can be disabled by setting splitOnCaseChange="0".
- A transition from alpha to numeric characters or vice versa: "Gonzo5000" -> "Gonzo", "5000" "4500XL" -> "4500", "XL". This can be disabled by setting splitOnNumerics="0".
- Non-alphanumeric characters (discarded): "hot-spot" -> "hot", "spot"
- A trailing "'s" is removed: "O'Reilly's" -> "O", "Reilly"
- Any leading or trailing delimiters are discarded: "--hot-spot--" -> "hot", "spot"

Factory class: solr.WordDelimiterGraphFilterFactory

Arguments:

generateWordParts

(integer, default 1) If non-zero, splits words at delimiters. For example: "CamelCase", "hot-spot" -> "Camel", "Case", "hot", "spot"

generateNumberParts

(integer, default 1) If non-zero, splits numeric strings at delimiters: "1947-32" ->* "1947", "32"

splitOnCaseChange

(integer, default 1) If 0, words are not split on camel-case changes: "BugBlaster-XL" -> "BugBlaster", "XL". Example 1 below illustrates the default (non-zero) splitting behavior.

splitOnNumerics

(integer, default 1) If 0, don't split words on transitions from alpha to numeric: "FemBot3000" -> "Fem", "Bot3000"

catenateWords

(integer, default 0) If non-zero, maximal runs of word parts will be joined: "hot-spot-sensor's" -> "hotspotsensor"

catenateNumbers

(integer, default 0) If non-zero, maximal runs of number parts will be joined: "1947-32" -> "194732"

catenateAll

(0/1, default 0) If non-zero, runs of word and number parts will be joined: "Zap-Master-9000" -> "ZapMaster9000"

preserveOriginal

(integer, default 0) If non-zero, the original token is preserved: "Zap-Master-9000" -> "Zap-Master-9000", "Zap", "Master", "9000"

protected

(optional) The pathname of a file that contains a list of protected words that should be passed through without splitting.

stemEnglishPossessive

(integer, default 1) If 1, strips the possessive 's from each subword.

types

(optional) The pathname of a file that contains **character => type** mappings, which enable customization of this filter's splitting behavior. Recognized character types: LOWER, UPPER, ALPHA, DIGIT, ALPHANUM, and SUBWORD_DELIM.

The default for any character without a customized mapping is computed from Unicode character properties. Blank lines and comment lines starting with '#' are ignored. An example file:

```
# Don't split numbers at '$', '.' or ','
$ => DIGIT
. => DIGIT
\u002C => DIGIT

# Don't split on ZWJ: http://en.wikipedia.org/wiki/Zero-width_joiner
\u200D => ALPHANUM
```

Example:

Default behavior. The whitespace tokenizer is used here to preserve non-alphanumeric characters.

```
<analyzer type="index">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrWordDelimiterGraphFilterFactory"/>
  <filter class="solrFlattenGraphFilterFactory"/> <!-- required on index analyzers after graph
filters -->
</analyzer>

<analyzer type="query">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrWordDelimiterGraphFilterFactory"/>
</analyzer>
```

In: "hot-spot RoboBlaster/9000 100XL"

Tokenizer to Filter: "hot-spot", "RoboBlaster/9000", "100XL"

Out: "hot", "spot", "Robo", "Blaster", "9000", "100", "XL"

Example:

Do not split on case changes, and do not generate number parts. Note that by not generating number parts, tokens containing only numeric parts are ultimately discarded.

```
<analyzer type="query">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrWordDelimiterGraphFilterFactory" generateNumberParts="0" splitOnCaseChange="0"/>
</analyzer>
```

In: "hot-spot RoboBlaster/9000 100-42"

Tokenizer to Filter: "hot-spot", "RoboBlaster/9000", "100-42"

Out: "hot", "spot", "RoboBlaster", "9000"

Example:

Concatenate word parts and number parts, but not word and number parts that occur in the same token.

```
<analyzer type="query">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrWordDelimiterGraphFilterFactory" catenateWords="1" catenateNumbers="1"/>
</analyzer>
```

In: "hot-spot 100+42 XL40"

Tokenizer to Filter: "hot-spot"(1), "100+42"(2), "XL40"(3)

Out: "hot"(1), "spot"(2), "hotspot"(2), "100"(3), "42"(4), "10042"(4), "XL"(5), "40"(6)

Example:

Concatenate all. Word and/or number parts are joined together.

```
<analyzer type="query">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrWordDelimiterGraphFilterFactory" catenateAll="1"/>
</analyzer>
```

In: "XL-4000/ES"

Tokenizer to Filter: "XL-4000/ES"(1)

Out: "XL"(1), "4000"(2), "ES"(3), "XL4000ES"(3)

Example:

Using a protected words list that contains "AstroBlaster" and "XL-5000" (among others).

```
<analyzer type="query">
  <tokenizer class="solrWhitespaceTokenizerFactory"/>
  <filter class="solrWordDelimiterGraphFilterFactory" protected="protwords.txt"/>
</analyzer>
```

In: "FooBar AstroBlaster XL-5000 ==ES-34-

Tokenizer to Filter: "FooBar", "AstroBlaster", "XL-5000", "==ES-34-

Out: "FooBar", "FooBar", "AstroBlaster", "XL-5000", "ES", "34"

CharFilterFactories

CharFilter is a component that pre-processes input characters.

CharFilters can be chained like Token Filters and placed in front of a Tokenizer. CharFilters can add, change, or remove characters while preserving the original character offsets to support features like highlighting.

solr.MappingCharFilterFactory

This filter creates org.apache.lucene.analysis.MappingCharFilter, which can be used for changing one string to another (for example, for normalizing é to e.).

This filter requires specifying a mapping argument, which is the path and name of a file containing the mappings to perform.

Example:

```
<analyzer>
  <charFilter class="solr.MappingCharFilterFactory" mapping="mapping-FoldToASCII.txt"/>
  <tokenizer ...>
  [...]
</analyzer>
```

Mapping file syntax:

- Comment lines beginning with a hash mark (#), as well as blank lines, are ignored.
- Each non-comment, non-blank line consists of a mapping of the form: "source" => "target"
 - Double-quoted source string, optional whitespace, an arrow (=>), optional whitespace, double-quoted target string.
- Trailing comments on mapping lines are not allowed.
- The source string must contain at least one character, but the target string may be empty.
- The following character escape sequences are recognized within source and target strings:

Escape Sequence	Resulting Character (ECMA-48 alias)	Unicode Character	Example Mapping Line
\\"	\\"	U+005C	"\\\" => "/"
\\"	"	U+0022	"\\\"and\\\" => '\"and\"'
\\b	backspace (BS)	U+0008	"\\b" => " "
\\t	tab (HT)	U+0009	"\\t" => " , "
\\n	newline (LF)	U+000A	"\\n" => " "
\\f	form feed (FF)	U+000C	"\\f" => "\\n"
\\r	carriage return (CR)	U+000D	"\\r" => "/carriage-return/"

Escape Sequence	Resulting Character (ECMA-48 alias)	Unicode Character	Example Mapping Line
\uXXXX	Unicode char referenced by the 4 hex digits	U+XXXX	"\uFEFF" => " "

- A backslash followed by any other character is interpreted as if the character were present without the backslash.

solr.HTMLStripCharFilterFactory

This filter creates org.apache.solr.analysis.HTMLStripCharFilter. This CharFilter strips HTML from the input stream and passes the result to another CharFilter or a Tokenizer.

This filter:

- Removes HTML/XML tags while preserving other content.
- Removes attributes within tags and supports optional attribute quoting.
- Removes XML processing instructions, such as: <?foo bar?>
- Removes XML comments.
- Removes XML elements starting with <!>.
- Removes contents of <script> and <style> elements.
- Handles XML comments inside these elements (normal comment processing will not always work).
- Replaces numeric character entities references like A or with the corresponding character.
- The terminating ';' is optional if the entity reference at the end of the input; otherwise the terminating ';' is mandatory, to avoid false matches on something like "Alpha&Omega Corp".
- Replaces all named character entity references with the corresponding character.
- s; is replaced with a space instead of the 0xa0 character.
- Newlines are substituted for block-level elements.
- <CDATA> sections are recognized.
- Inline tags, such as , <i>, or will be removed.
- Uppercase character entities like quot, gt, lt and amp are recognized and handled as lowercase.



The input need not be an HTML document. The filter removes only constructs that look like HTML. If the input doesn't include anything that looks like HTML, the filter won't remove any input.

The table below presents examples of HTML stripping.

Input	Output
my link	my link
 hello<!--comment-->	hello

Input	Output
hello<script><!-- f('!--internal--></script>'); --></script>	hello
if a<b then print a;	if a<b then print a;
hello <td height=22 nowrap align="left">	hello
a<b A Alpha&Omega Ω	a<b A Alpha&Omega Ω

Example:

```
<analyzer>
  <charFilter class="solr.HTMLStripCharFilterFactory"/>
  <tokenizer ...>
  [...]
</analyzer>
```

solr.ICUNormalizer2CharFilterFactory

This filter performs pre-tokenization Unicode normalization using [ICU4J](#).

Arguments:

name

A [Unicode Normalization Form](#), one of nfc, nfkc, nfkc_cf. Default is nfkc_cf.

mode

Either compose or decompose. Default is compose. Use decompose with name="nfc" or name="nfkc" to get NFD or NFKD, respectively.

filter

A [UnicodeSet](#) pattern. Codepoints outside the set are always left unchanged. Default is [] (the null set, no filtering - all codepoints are subject to normalization).

Example:

```
<analyzer>
  <charFilter class="solr.ICUNormalizer2CharFilterFactory"/>
  <tokenizer ...>
  [...]
</analyzer>
```

solr.PatternReplaceCharFilterFactory

This filter uses [regular expressions](#) to replace or change character patterns.

Arguments:

pattern

the regular expression pattern to apply to the incoming text.

replacement

the text to use to replace matching patterns.

You can configure this filter in schema.xml like this:

```
<analyzer>
  <charFilter class="solr.PatternReplaceCharFilterFactory"
    pattern="([nN][oO]).)\s*(\d+)" replacement="$1$2"/>
  <tokenizer ...>
  [...]
</analyzer>
```

The table below presents examples of regex-based pattern replacement:

Input	Pattern	Replacement	Output	Description
see-ing looking	(\w+)(ing)	\$1	see-ing look	Removes "ing" from the end of word.
see-ing looking	(\w+)ing	\$1	see-ing look	Same as above. 2nd parentheses can be omitted.
No.1 NO. no. 543	[nN][oO]\.\s*(\d+)	#\$1	#1 NO. #543	Replace some string literals
abc=1234=5678	(\w+)=(\d+)=(\d+)	\$3=\$1=\$2	5678=abc=1234	Change the order of the groups.

Language Analysis

This section contains information about tokenizers and filters related to character set conversion or for use with specific languages.

For the European languages, tokenization is fairly straightforward. Tokens are delimited by white space and/or a relatively small set of punctuation characters.

In other languages the tokenization rules are often not so simple. Some European languages may also require special tokenization rules, such as rules for decompounding German words.

For information about language detection at index time, see [Detecting Languages During Indexing](#).

KeywordMarkerFilterFactory

Protects words from being modified by stemmers. A customized protected word list may be specified with the "protected" attribute in the schema. Any words in the protected word list will not be modified by any stemmer in Solr.

A sample Solr protwords.txt with comments can be found in the sample_techproducts_configs configset directory:

```
<fieldtype name="myfieldtype" class="solr.TextField">
  <analyzer>
    <tokenizer class="solrWhitespaceTokenizerFactory"/>
    <filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt" />
    <filter class="solr.PorterStemFilterFactory" />
  </analyzer>
</fieldtype>
```

KeywordRepeatFilterFactory

Emits each token twice, one with the KEYWORD attribute and once without.

If placed before a stemmer, the result will be that you will get the unstemmed token preserved on the same position as the stemmed one. Queries matching the original exact term will get a better score while still maintaining the recall benefit of stemming. Another advantage of keeping the original token is that wildcard truncation will work as expected.

To configure, add the KeywordRepeatFilterFactory early in the analysis chain. It is recommended to also include RemoveDuplicatesTokenFilterFactory to avoid duplicates when tokens are not stemmed.

A sample fieldType configuration could look like this:

```
<fieldtype name="english_stem_preserve_original" class="solr.TextField">
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.KeywordRepeatFilterFactory" />
    <filter class="solr.PorterStemFilterFactory" />
    <filter class="solr.RemoveDuplicatesTokenFilterFactory" />
  </analyzer>
</fieldtype>
```



When adding the same token twice, it will also score twice (double), so you may have to re-tune your ranking rules.

StemmerOverrideFilterFactory

Overrides stemming algorithms by applying a custom mapping, then protecting these terms from being modified by stemmers.

A customized mapping of words to stems, in a tab-separated file, can be specified to the dictionary attribute in the schema. Words in this mapping will be stemmed to the stems from the file, and will not be further changed by any stemmer.

```
<fieldtype name="myfieldtype" class="solr.TextField">
  <analyzer>
    <tokenizer class="solrWhitespaceTokenizerFactory"/>
    <filter class="solr.StemmerOverrideFilterFactory" dictionary="stemdict.txt" />
    <filter class="solr.PorterStemFilterFactory" />
  </analyzer>
</fieldtype>
```

A sample `stemdict.txt` file is shown below:

```
# these must be tab-separated
monkeys monkey
otters otter
# some crazy ones that a stemmer would never do
dogs cat
```

If you have a checkout of Solr's source code locally, you can also find this example in Solr's test resources at `solr/core/src/test-files/solr/collection1/conf/stemdict.txt`.

Dictionary Compound Word Token Filter

This filter splits, or *decompounds*, compound words into individual words using a dictionary of the component words. Each input token is passed through unchanged. If it can also be decompounded into subwords, each subword is also added to the stream at the same logical position.

Compound words are most commonly found in Germanic languages.

Factory class: solr.DictionaryCompoundWordTokenFilterFactory

Arguments:

dictionary

(required) The path of a file that contains a list of simple words, one per line. Blank lines and lines that begin with "#" are ignored. See [Resource and Plugin Loading](#) for more information.

minWordSize

(integer, default 5) Any token shorter than this is not decompounded.

minSubwordSize

(integer, default 2) Subwords shorter than this are not emitted as tokens.

maxSubwordSize

(integer, default 15) Subwords longer than this are not emitted as tokens.

onlyLongestMatch

(true/false) If true (the default), only the longest matching subwords will generate new tokens.

Example:

Assume that germanwords.txt contains at least the following words: dummm kopf donau dampf schiff

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.DictionaryCompoundWordTokenFilterFactory" dictionary="germanwords.txt"/>
</analyzer>
```

In: "Donaudampfschiff dummkopf"

Tokenizer to Filter: "Donaudampfschiff"(1), "dummkopf"(2),

Out: "Donaudampfschiff"(1), "Donau"(1), "dampf"(1), "schiff"(1), "dummkopf"(2), "dumm"(2), "kopf"(2)

Unicode Collation

Unicode Collation is a language-sensitive method of sorting text that can also be used for advanced search purposes.

Unicode Collation in Solr is fast, because all the work is done at index time.

Rather than specifying an analyzer within `<fieldtype ... class="solr.TextField">`, the `solr.CollationField` and `solr.ICUCollationField` field type classes provide this functionality. `solr.ICUCollationField`, which is backed by [the ICU4J library](#), provides more flexible configuration, has more locales, is significantly faster, and requires less memory and less index space, since its keys are smaller than those produced by the JDK implementation that backs `solr.CollationField`.

To use `solr.ICUCollationField`, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See `solr/contrib/analysis-extras/README.txt` for instructions on which jars you need to add.

`solr.ICUCollationField` and `solr.CollationField` fields can be created in two ways:

- Based upon a system collator associated with a Locale.
- Based upon a tailored RuleBasedCollator ruleset.

Arguments for `solr.ICUCollationField`, specified as attributes within the `<fieldtype>` element:

Using a System collator:

`locale`

(required) [RFC 3066](#) locale ID. See [the ICU locale explorer](#) for a list of supported locales.

`strength`

Valid values are primary, secondary, tertiary, quaternary, or identical. See [Comparison Levels in ICU Collation Concepts](#) for more information.

`decomposition`

Valid values are no or canonical. See [Normalization in ICU Collation Concepts](#) for more information.

Using a Tailored ruleset:

`custom`

(required) Path to a UTF-8 text file containing rules supported by the ICU RuleBasedCollator

`strength`

Valid values are primary, secondary, tertiary, quaternary, or identical. See [Comparison Levels in ICU Collation Concepts](#) for more information.

`decomposition`

Valid values are no or canonical. See [Normalization in ICU Collation Concepts](#) for more information.

Expert options:

`alternate`

Valid values are shifted or non-ignorable. Can be used to ignore punctuation/whitespace.

`caseLevel`

(true/false) If true, in combination with `strength="primary"`, accents are ignored but case is taken into account. The default is false. See [CaseLevel in ICU Collation Concepts](#) for more information.

`caseFirst`

Valid values are lower or upper. Useful to control which is sorted first when case is not ignored.

`numeric`

(true/false) If true, digits are sorted according to numeric value, e.g., foobar-9 sorts before foobar-10. The default is false.

`variableTop`

Single character or contraction. Controls what is variable for alternate.

Sorting Text for a Specific Language

In this example, text is sorted according to the default German rules provided by ICU4J.

Locales are typically defined as a combination of language and country, but you can specify just the language if you want. For example, if you specify "de" as the language, you will get sorting that works well for the German language. If you specify "de" as the language and "CH" as the country, you will get German sorting specifically tailored for Switzerland.

```
<!-- Define a field type for German collation -->
<fieldType name="collatedGERMAN" class="solr.ICUCollationField"
    locale="de"
    strength="primary" />
...
<!-- Define a field to store the German collated manufacturer names. -->
<field name="manuGERMAN" type="collatedGERMAN" indexed="false" stored="false" docValues="true"/>
...
<!-- Copy the text to this field. We could create French, English, Spanish versions too,
     and sort differently for different users! -->
<copyField source="manu" dest="manuGERMAN"/>
```

In the example above, we defined the strength as "primary". The strength of the collation determines how strict the sort order will be, but it also depends upon the language. For example, in English, "primary" strength ignores differences in case and accents.

Another example:

```
<fieldType name="polishCaseInsensitive" class="solr.ICUCollationField"
    locale="pl_PL"
    strength="secondary" />
...
<field name="city" type="text_general" indexed="true" stored="true"/>
...
<field name="city_sort" type="polishCaseInsensitive" indexed="true" stored="false"/>
...
<copyField source="city" dest="city_sort"/>
```

The type will be used for the fields where the data contains Polish text. The "secondary" strength will ignore case differences, but, unlike "primary" strength, a letter with diacritic(s) will be sorted differently from the same base letter without diacritics.

An example using the "city_sort" field to sort:

```
q=*&fl=city&sort=city_sort+asc
```

Sorting Text for Multiple Languages

There are two approaches to supporting multiple languages: if there is a small list of languages you wish to

support, consider defining collated fields for each language and using copyField. However, adding a large number of sort fields can increase disk and indexing costs. An alternative approach is to use the Unicode default collator.

The Unicode default or ROOT locale has rules that are designed to work well for most languages. To use the default locale, simply define the locale as the empty string. This Unicode default sort is still significantly more advanced than the standard Solr sort.

```
<fieldType name="collatedROOT" class="solr.ICUCollationField"
    locale=""
    strength="primary" />
```

Sorting Text with Custom Rules

You can define your own set of sorting rules. It's easiest to take existing rules that are close to what you want and customize them.

In the example below, we create a custom rule set for German called DIN 5007-2. This rule set treats umlauts in German differently: it treats ö as equivalent to oe, ä as equivalent to ae, and ü as equivalent to ue. For more information, see the [ICU RuleBasedCollator javadocs](#).

This example shows how to create a custom rule set for solr.ICUCollationField and dump it to a file:

```
// get the default rules for Germany
// these are called DIN 5007-1 sorting
RuleBasedCollator baseCollator = (RuleBasedCollator) Collator.getInstance(new ULocale("de", "DE"));

// define some tailorings, to make it DIN 5007-2 sorting.
// For example, this makes ö equivalent to oe
String DIN5007_2_tailorings =
    "& ae , a\u0308 & AE , A\u0308"+
    "& oe , o\u0308 & OE , O\u0308"+
    "& ue , u\u0308 & UE , U\u0308";

// concatenate the default rules to the tailorings, and dump it to a String
RuleBasedCollator tailoredCollator = new RuleBasedCollator(baseCollator.getRules() +
DIN5007_2_tailorings);
String tailoredRules = tailoredCollator.getRules();

// write these to a file, be sure to use UTF-8 encoding!!!
FileOutputStream os = new FileOutputStream(new File("/solr_home/conf/customRules.dat"));
IOUtils.write(tailoredRules, os, "UTF-8");
```

This rule set can now be used for custom collation in Solr:

```
<fieldType name="collatedCUSTOM" class="solr.ICUCollationField"
    custom="customRules.dat"
    strength="primary" />
```

JDK Collation

As mentioned above, ICU Unicode Collation is better in several ways than JDK Collation, but if you cannot use ICU4J for some reason, you can use solr.CollationField.

The principles of JDK Collation are the same as those of ICU Collation; you just specify language, country and variant arguments instead of the combined locale argument.

Arguments for solr.CollationField, specified as attributes within the <fieldtype> element:

Using a System collator (see [Oracle's list of locales supported in Java](#)):

language

(required) [ISO-639](#) language code

country

[ISO-3166](#) country code

variant

Vendor or browser-specific code

strength

Valid values are primary, secondary, tertiary or identical. See [Java Collator javadocs](#) for more information.

decomposition

Valid values are no, canonical, or full. See [Java Collator javadocs](#) for more information.

Using a Tailored ruleset:

custom

(required) Path to a UTF-8 text file containing rules supported by the JDK RuleBasedCollator

strength

Valid values are primary, secondary, tertiary or identical. See [Java Collator javadocs](#) for more information.

decomposition

Valid values are no, canonical, or full. See [Java Collator javadocs](#) for more information.

A solr.CollationField example:

```
<fieldType name="collatedGERMAN" class="solr.CollationField"  
    language="de"  
    country="DE"  
    strength="primary" /> <!-- ignore Umlauts and letter case when sorting -->  
...  
<field name="manuGERMAN" type="collatedGERMAN" indexed="false" stored="false" docValues="true" />  
...  
<copyField source="manu" dest="manuGERMAN"/>
```

ASCII & Decimal Folding Filters

ASCII Folding

This filter converts alphabetic, numeric, and symbolic Unicode characters which are not in the first 127 ASCII characters (the "Basic Latin" Unicode block) into their ASCII equivalents, if one exists. Only those characters with reasonable ASCII alternatives are converted.

This can increase recall by causing more matches. On the other hand, it can reduce precision because language-specific character differences may be lost.

Factory class: solr.ASCIIFFoldingFilterFactory

Arguments: None

Example:

```
<analyzer>  
    <tokenizer class="solr.StandardTokenizerFactory"/>  
    <filter class="solr.ASCIIFFoldingFilterFactory"/>  
</analyzer>
```

In: "Björn Ångström"

Tokenizer to Filter: "Björn", "Ångström"

Out: "Bjorn", "Angstrom"

Decimal Digit Folding

This filter converts any character in the Unicode "Decimal Number" general category (Nd) into their equivalent Basic Latin digits (0-9).

This can increase recall by causing more matches. On the other hand, it can reduce precision because language-specific character differences may be lost.

Factory class: solr.DecimalDigitFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.DecimalDigitFilterFactory"/>
</analyzer>
```

OpenNLP Integration

The lucene/analysis/opennlp module provides OpenNLP integration via several analysis components: a tokenizer, a part-of-speech tagging filter, a phrase chunking filter, and a lemmatization filter. In addition to these analysis components, Solr also provides an update request processor to extract named entities - see [Update Processor Factories That Can Be Loaded as Plugins](#).



The [OpenNLP Tokenizer](#) must be used with all other OpenNLP analysis components, for two reasons: first, the OpenNLP Tokenizer detects and marks the sentence boundaries required by all the OpenNLP filters; and second, since the pre-trained OpenNLP models used by these filters were trained using the corresponding language-specific sentence-detection/tokenization models, the same tokenization, using the same models, must be used at runtime for optimal performance.

To use the OpenNLP components, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See solr/contrib/analysis-extras/README.txt for instructions on which jars you need to add.

OpenNLP Tokenizer

The OpenNLP Tokenizer takes two language-specific binary model files as parameters: a sentence detector model and a tokenizer model. The last token in each sentence is flagged, so that following OpenNLP-based filters can use this information to apply operations to tokens one sentence at a time. See the [OpenNLP website](#) for information on downloading pre-trained models.

Factory class: solr.OpenNLPTokenizerFactory

Arguments:

sentenceModel

(required) The path of a language-specific OpenNLP sentence detection model file. See [Resource and Plugin Loading](#) for more information.

tokenizerModel

(required) The path of a language-specific OpenNLP tokenization model file. See [Resource and Plugin Loading](#) for more information.

Example:

```
<analyzer>
  <tokenizer class="solr.OpenNLPTokenizerFactory"
    sentenceModel="en-sent.bin"
    tokenizerModel="en-tokenizer.bin"/>
</analyzer>
```

OpenNLP Part-Of-Speech Filter

This filter sets each token's type attribute to the part of speech (POS) assigned by the configured model. See the [OpenNLP website](#) for information on downloading pre-trained models.



Lucene currently does not index token types, so if you want to keep this information, you have to preserve it either in a payload or as a synonym; see the examples below.

Factory class: solr.OpenNLPPosFilterFactory

Arguments:

posTaggerModel

(required) The path of a language-specific OpenNLP POS tagger model file. See [Resource and Plugin Loading](#) for more information.

Examples:

The OpenNLP tokenizer will tokenize punctuation, which is useful for following token filters, but ordinarily you don't want to include punctuation in your index, so the TypeTokenFilter ([described here](#)) is included in the examples below, with `stop.pos.txt` containing the following:

`stop.pos.txt`

```
#  
$  
'  
'  
,  
-LRB-  
-RRB-  
:  
. 
```

Index the POS for each token as a payload:

```
<analyzer>
  <tokenizer class="solr.OpenNLPTokenizerFactory"
    sentenceModel="en-sent.bin"
    tokenizerModel="en-tokenizer.bin"/>
  <filter class="solr.OpenNLPOSFilterFactory" posTaggerModel="en-pos-maxent.bin"/>
  <filter class="solr.TypeAsPayloadFilterFactory"/>
  <filter class="solr.TypeTokenFilterFactory" types="stop.pos.txt"/>
</analyzer>
```

Index the POS for each token as a synonym, after prefixing the POS with "@" (see the [TypeAsSynonymFilter description](#)):

```
<analyzer>
  <tokenizer class="solr.OpenNLPTokenizerFactory"
    sentenceModel="en-sent.bin"
    tokenizerModel="en-tokenizer.bin"/>
  <filter class="solr.OpenNLPOSFilterFactory" posTaggerModel="en-pos-maxent.bin"/>
  <filter class="solr.TypeAsSynonymFilterFactory" prefix="@"/>
  <filter class="solr.TypeTokenFilterFactory" types="stop.pos.txt"/>
</analyzer>
```

Only index nouns - the keep.pos.txt file contains lines NN, NNS, NNP and NNPS:

```
<analyzer>
  <tokenizer class="solr.OpenNLPTokenizerFactory"
    sentenceModel="en-sent.bin"
    tokenizerModel="en-tokenizer.bin"/>
  <filter class="solr.OpenNLPOSFilterFactory" posTaggerModel="en-pos-maxent.bin"/>
  <filter class="solr.TypeTokenFilterFactory" types="keep.pos.txt" useWhitelist="true"/>
</analyzer>
```

OpenNLP Phrase Chunking Filter

This filter sets each token's type attribute based on the output of an OpenNLP phrase chunking model. The chunk labels replace the POS tags that previously were in each token's type attribute. See the [OpenNLP website](#) for information on downloading pre-trained models.

Prerequisite: the [OpenNLP Tokenizer](#) and the [OpenNLP Part-Of-Speech Filter](#) must precede this filter.



Lucene currently does not index token types, so if you want to keep this information, you have to preserve it either in a payload or as a synonym; see the examples below.

Factory class: solr.OpenNLPChunkerFilter

Arguments:

chunkerModel

(required) The path of a language-specific OpenNLP phrase chunker model file. See [Resource and Plugin](#)

[Loading](#) for more information.

Examples:

Index the phrase chunk label for each token as a payload:

```
<analyzer>
  <tokenizer class="solr.OpenNLPTokenizerFactory"
    sentenceModel="en-sent.bin"
    tokenizerModel="en-tokenizer.bin"/>
  <filter class="solr.OpenNLPOSFilterFactory" posTaggerModel="en-pos-maxent.bin"/>
  <filter class="solr.OpenNLPChunkerFactory" chunkerModel="en-chunker.bin"/>
  <filter class="solr.TypeAsPayloadFilterFactory"/>
</analyzer>
```

Index the phrase chunk label for each token as a synonym, after prefixing it with "#" (see the [TypeAsSynonymFilter description](#)):

```
<analyzer>
  <tokenizer class="solr.OpenNLPTokenizerFactory"
    sentenceModel="en-sent.bin"
    tokenizerModel="en-tokenizer.bin"/>
  <filter class="solr.OpenNLPOSFilterFactory" posTaggerModel="en-pos-maxent.bin"/>
  <filter class="solr.OpenNLPChunkerFactory" chunkerModel="en-chunker.bin"/>
  <filter class="solr.TypeAsSynonymFilterFactory" prefix="#" />
</analyzer>
```

OpenNLP Lemmatizer Filter

This filter replaces the text of each token with its lemma. Both a dictionary-based lemmatizer and a model-based lemmatizer are supported. If both are configured, the dictionary-based lemmatizer is tried first, and then the model-based lemmatizer is consulted for out-of-vocabulary tokens. See the [OpenNLP website](#) for information on downloading pre-trained models.

Factory class: solr.OpenNLPLemmatizerFilter

Arguments:

Either dictionary or lemmatizerModel must be provided, and both may be provided - see the examples below:

dictionary

(optional) The path of a lemmatization dictionary file. See [Resource and Plugin Loading](#) for more information. The dictionary file must be encoded as UTF-8, with one entry per line, in the form word[tab]lemma[tab]part-of-speech, e.g., wrote[tab]write[tab]VBD.

lemmatizerModel

(optional) The path of a language-specific OpenNLP lemmatizer model file. See [Resource and Plugin Loading](#) for more information.

Examples:

Perform dictionary-based lemmatization, and fall back to model-based lemmatization for out-of-vocabulary tokens (see the [OpenNLP Part-Of-Speech Filter](#) section above for information about using TypeTokenFilter to avoid indexing punctuation):

```
<analyzer>
  <tokenizer class="solr.OpenNLPTokenizerFactory"
    sentenceModel="en-sent.bin"
    tokenizerModel="en-tokenizer.bin"/>
  <filter class="solr.OpenNLPOSFilterFactory" posTaggerModel="en-pos-maxent.bin"/>
  <filter class="solr.OpenNLPLemmatizerFilterFactory"
    dictionary="lemmas.txt"
    lemmatizerModel="en-lemmatizer.bin"/>
  <filter class="solr.TypeTokenFilterFactory" types="stop.pos.txt"/>
</analyzer>
```

Perform dictionary-based lemmatization only:

```
<analyzer>
  <tokenizer class="solr.OpenNLPTokenizerFactory"
    sentenceModel="en-sent.bin"
    tokenizerModel="en-tokenizer.bin"/>
  <filter class="solr.OpenNLPOSFilterFactory" posTaggerModel="en-pos-maxent.bin"/>
  <filter class="solr.OpenNLPLemmatizerFilterFactory" dictionary="lemmas.txt"/>
  <filter class="solr.TypeTokenFilterFactory" types="stop.pos.txt"/>
</analyzer>
```

Perform model-based lemmatization only, preserving the original token and emitting the lemma as a synonym (see the [KeywordRepeatFilterFactory description](#)):

```
<analyzer>
  <tokenizer class="solr.OpenNLPTokenizerFactory"
    sentenceModel="en-sent.bin"
    tokenizerModel="en-tokenizer.bin"/>
  <filter class="solr.OpenNLPOSFilterFactory" posTaggerModel="en-pos-maxent.bin"/>
  <filter class="solr.KeywordRepeatFilterFactory"/>
  <filter class="solr.OpenNLPLemmatizerFilterFactory" lemmatizerModel="en-lemmatizer.bin"/>
  <filter class="solr.RemoveDuplicatesTokenFilterFactory"/>
  <filter class="solr.TypeTokenFilterFactory" types="stop.pos.txt"/>
</analyzer>
```

Language-Specific Factories

These factories are each designed to work with specific languages. The languages covered here are:

- [Arabic](#)
- [Bengali](#)

- [Brazilian Portuguese](#)
- [Bulgarian](#)
- [Catalan](#)
- [Traditional Chinese](#)
- [Simplified Chinese](#)
- [Czech](#)
- [Danish](#)
- [Dutch](#)
- [Finnish](#)
- [French](#)
- [Galician](#)
- [German](#)
- [Greek](#)
- [Hebrew, Lao, Myanmar, Khmer](#)
- [Hindi](#)
- [Indonesian](#)
- [Italian](#)
- [Irish](#)
- [Japanese](#)
- [Latvian](#)
- [Norwegian](#)
- [Persian](#)
- [Polish](#)
- [Portuguese](#)
- [Romanian](#)
- [Russian](#)
- [Scandinavian](#)
- [Serbian](#)
- [Spanish](#)
- [Swedish](#)
- [Thai](#)
- [Turkish](#)
- [Ukrainian](#)

Arabic

Solr provides support for the [Light-10](#) (PDF) stemming algorithm, and Lucene includes an example stopword

list.

This algorithm defines both character normalization and stemming, so these are split into two filters to provide more flexibility.

Factory classes: solr.ArabicStemFilterFactory, solr.ArabicNormalizationFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ArabicNormalizationFilterFactory"/>
  <filter class="solr.ArabicStemFilterFactory"/>
</analyzer>
```

Bengali

There are two filters written specifically for dealing with Bengali language. They use the Lucene classes org.apache.lucene.analysis.bn.BengaliNormalizationFilter and org.apache.lucene.analysis.bn.BengaliStemFilter.

Factory classes: solr.BengaliStemFilterFactory, solr.BengaliNormalizationFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.BengaliNormalizationFilterFactory"/>
  <filter class="solr.BengaliStemFilterFactory"/>
</analyzer>
```

Normalisation - →

Stemming - →

Brazilian Portuguese

This is a Java filter written specifically for stemming the Brazilian dialect of the Portuguese language. It uses the Lucene class org.apache.lucene.analysis.br.BrazilianStemmer. Although that stemmer can be configured to use a list of protected words (which should not be stemmed), this factory does not accept any arguments to specify such a list.

Factory class: solr.BrazilianStemFilterFactory

Arguments: None

Example:

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.BrazilianStemFilterFactory"/>
</analyzer>
```

In: "praia praias"

Tokenizer to Filter: "praia", "praias"

Out: "pra", "pra"

Bulgarian

Solr includes a light stemmer for Bulgarian, following [this algorithm](#) (PDF), and Lucene includes an example stopword list.

Factory class: solr.BulgarianStemFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.BulgarianStemFilterFactory"/>
</analyzer>
```

Catalan

Solr can stem Catalan using the Snowball Porter Stemmer with an argument of language="Catalan". Solr includes a set of contractions for Catalan, which can be stripped using solr.ElisionFilterFactory.

Factory class: solr.SnowballPorterFilterFactory

Arguments:

language

(required) stemmer language, "Catalan" in this case

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.ElisionFilterFactory"
    articles="lang/contractions_ca.txt"/>
  <filter class="solr.SnowballPorterFilterFactory" language="Catalan" />
</analyzer>
```

In: "llengües llengua"

Tokenizer to Filter: "llengües"(1) "llengua"(2),

Out: "llengu"(1), "llengu"(2)

Traditional Chinese

The default configuration of the [ICU Tokenizer](#) is suitable for Traditional Chinese text. It follows the Word Break rules from the Unicode Text Segmentation algorithm for non-Chinese text, and uses a dictionary to segment Chinese words. To use this tokenizer, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See the solr/contrib/analysis-extras/README.txt for information on which jars you need to add.

[Standard Tokenizer](#) can also be used to tokenize Traditional Chinese text. Following the Word Break rules from the Unicode Text Segmentation algorithm, it produces one token per Chinese character. When combined with [CJK Bigram Filter](#), overlapping bigrams of Chinese characters are formed.

[CJK Width Filter](#) folds fullwidth ASCII variants into the equivalent Basic Latin forms.

Examples:

```
<analyzer>
  <tokenizer class="solr.ICUTokenizerFactory"/>
  <filter class="solr.CJKWidthFilterFactory"/>
  <filter class="solrLowerCaseFilterFactory"/>
</analyzer>
```

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.CJKBigramFilterFactory"/>
  <filter class="solr.CJKWidthFilterFactory"/>
  <filter class="solrLowerCaseFilterFactory"/>
</analyzer>
```

CJK Bigram Filter

Forms bigrams (overlapping 2-character sequences) of CJK characters that are generated from [Standard Tokenizer](#) or [ICU Tokenizer](#).

By default, all CJK characters produce bigrams, but finer grained control is available by specifying orthographic type arguments han, hiragana, katakana, and hangul. When set to false, characters of the corresponding type will be passed through as unigrams, and will not be included in any bigrams.

When a CJK character has no adjacent characters to form a bigram, it is output in unigram form. If you want to always output both unigrams and bigrams, set the outputUnigrams argument to true.

In all cases, all non-CJK input is passed through unmodified.

Arguments:

han

(true/false) If false, Han (Chinese) characters will not form bigrams. Default is true.

hiragana

(true/false) If false, Hiragana (Japanese) characters will not form bigrams. Default is true.

katakana

(true/false) If false, Katakana (Japanese) characters will not form bigrams. Default is true.

hangul

(true/false) If false, Hangul (Korean) characters will not form bigrams. Default is true.

outputUnigrams

(true/false) If true, in addition to forming bigrams, all characters are also passed through as unigrams.

Default is false.

See the example under [Traditional Chinese](#).

Simplified Chinese

For Simplified Chinese, Solr provides support for Chinese sentence and word segmentation with the [HMM Chinese Tokenizer](#). This component includes a large dictionary and segments Chinese text into words with the Hidden Markov Model. To use this tokenizer, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See the solr/contrib/analysis-extras/README.txt for information on which jars you need to add.

The default configuration of the [ICU Tokenizer](#) is also suitable for Simplified Chinese text. It follows the Word Break rules from the Unicode Text Segmentation algorithm for non-Chinese text, and uses a dictionary to segment Chinese words. To use this tokenizer, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See the solr/contrib/analysis-extras/README.txt for information on which jars you need to add.

Also useful for Chinese analysis:

[CJK Width Filter](#) folds fullwidth ASCII variants into the equivalent Basic Latin forms, and folds halfwidth Katakana variants into their equivalent fullwidth forms.

Examples:

```
<analyzer>
  <tokenizer class="solr.HMMChineseTokenizerFactory"/>
  <filter class="solr.CJKWidthFilterFactory"/>
  <filter class="solr.StopFilterFactory"
    words="org/apache/lucene/analysis/cn/smart/stopwords.txt"/>
  <filter class="solr.PorterStemFilterFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
</analyzer>
```

```
<analyzer>
  <tokenizer class="solr.ICUTokenizerFactory"/>
  <filter class="solr.CJKWidthFilterFactory"/>
  <filter class="solr.StopFilterFactory"
    words="org/apache/lucene/analysis/cn/smart/stopwords.txt"/>
  <filter class="solr.LowerCaseFilterFactory"/>
</analyzer>
```

HMM Chinese Tokenizer

For Simplified Chinese, Solr provides support for Chinese sentence and word segmentation with the `solr.HMMChineseTokenizerFactory` in the `analysis-extras` contrib module. This component includes a large dictionary and segments Chinese text into words with the Hidden Markov Model. To use this tokenizer, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See `solr/contrib/analysis-extras/README.txt` for instructions on which jars you need to add.

Factory class: `solr.HMMChineseTokenizerFactory`

Arguments: None

Examples:

To use the default setup with fallback to English Porter stemmer for English words, use:

```
<analyzer class="org.apache.lucene.analysis.cn.smart.SmartChineseAnalyzer"/>
```

Or to configure your own analysis setup, use the `solr.HMMChineseTokenizerFactory` along with your custom filter setup. See an example of this in the [Simplified Chinese](#) section.

Czech

Solr includes a light stemmer for Czech, following [this algorithm](#), and Lucene includes an example stopword list.

Factory class: `solr.CzechStemFilterFactory`

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.CzechStemFilterFactory"/>
</analyzer>
```

In: "prezidenští, prezidenta, prezidentského"

Tokenizer to Filter: "prezidenští", "prezidenta", "prezidentského"

Out: "preziden", "preziden", "preziden"

Danish

Solr can stem Danish using the Snowball Porter Stemmer with an argument of language="Danish".

Also relevant are the [Scandinavian normalization filters](#).

Factory class: solr.SnowballPorterFilterFactory

Arguments:

language
(required) stemmer language, "Danish" in this case

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.SnowballPorterFilterFactory" language="Danish" />
</analyzer>
```

In: "undersøg undersøgelse"

Tokenizer to Filter: "undersøg"(1) "undersøgelse"(2),

Out: "undersøg"(1), "undersøg"(2)

Dutch

Solr can stem Dutch using the Snowball Porter Stemmer with an argument of language="Dutch".

Factory class: solr.SnowballPorterFilterFactory

Arguments:

language
(required) stemmer language, "Dutch" in this case

Example:

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.SnowballPorterFilterFactory" language="Dutch"/>
</analyzer>
```

In: "kanaal kanalen"

Tokenizer to Filter: "kanaal", "kanalen"

Out: "kanal", "kanal"

Finnish

Solr includes support for stemming Finnish, and Lucene includes an example stopword list.

Factory class: solr.FinnishLightStemFilterFactory

Arguments: None

Example:

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.FinnishLightStemFilterFactory"/>
</analyzer>
```

In: "kala kalat"

Tokenizer to Filter: "kala", "kalat"

Out: "kala", "kala"

French

Elision Filter

Removes article elisions from a token stream. This filter can be useful for languages such as French, Catalan, Italian, and Irish.

Factory class: solr.ElisionFilterFactory

Arguments:

articles

The pathname of a file that contains a list of articles, one per line, to be stripped. Articles are words such as "le", which are commonly abbreviated, such as in *l'avion* (the plane). This file should include the abbreviated form, which precedes the apostrophe. In this case, simply "/". If no articles attribute is specified, a default set of French articles is used.

ignoreCase

(boolean) If true, the filter ignores the case of words when comparing them to the common word file.
Defaults to false

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ElisionFilterFactory"
    ignoreCase="true"
    articles="lang/contractions_fr.txt"/>
</analyzer>
```

In: "L'histoire d'art"

Tokenizer to Filter: "L'histoire", "d'art"

Out: "histoire", "art"

French Light Stem Filter

Solr includes three stemmers for French: one in the solr.SnowballPorterFilterFactory, a lighter stemmer called solr.FrenchLightStemFilterFactory, and an even less aggressive stemmer called solr.FrenchMinimalStemFilterFactory. Lucene includes an example stopword list.

Factory classes: solr.FrenchLightStemFilterFactory, solr.FrenchMinimalStemFilterFactory

Arguments: None

Examples:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solrLowerCaseFilterFactory"/>
  <filter class="solr.ElisionFilterFactory"
    articles="lang/contractions_fr.txt"/>
  <filter class="solr.FrenchLightStemFilterFactory"/>
</analyzer>
```

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solrLowerCaseFilterFactory"/>
  <filter class="solr.ElisionFilterFactory"
    articles="lang/contractions_fr.txt"/>
  <filter class="solr.FrenchMinimalStemFilterFactory"/>
</analyzer>
```

In: "le chat, les chats"

Tokenizer to Filter: "le", "chat", "les", "chats"

Out: "le", "chat", "le", "chat"

Galician

Solr includes a stemmer for Galician following [this algorithm](#), and Lucene includes an example stopword list.

Factory class: solr.GalicianStemFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.GalicianStemFilterFactory"/>
</analyzer>
```

In: "felizmente Luzes"

Tokenizer to Filter: "felizmente", "luzes"

Out: "feliz", "luz"

German

Solr includes four stemmers for German: one in the solr.SnowballPorterFilterFactory language="German", a stemmer called solr.GermanStemFilterFactory, a lighter stemmer called solr.GermanLightStemFilterFactory, and an even less aggressive stemmer called solr.GermanMinimalStemFilterFactory. Lucene includes an example stopword list.

Factory classes: solr.GermanStemFilterFactory, solr.LightGermanStemFilterFactory, solr.MinimalGermanStemFilterFactory

Arguments: None

Examples:

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory "/>
  <filter class="solr.GermanStemFilterFactory"/>
</analyzer>
```

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.GermanLightStemFilterFactory"/>
</analyzer>
```

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory "/>
  <filter class="solr.GermanMinimalStemFilterFactory"/>
</analyzer>
```

In: "haus häuser"

Tokenizer to Filter: "haus", "häuser"

Out: "haus", "haus"

Greek

This filter converts uppercase letters in the Greek character set to the equivalent lowercase character.

Factory class: solr.GreekLowerCaseFilterFactory

Arguments: None



Use of custom charsets is no longer supported as of Solr 3.1. If you need to index text in these encodings, please use Java's character set conversion facilities (InputStreamReader, etc.) during I/O, so that Lucene can analyze this text as Unicode instead.

Example:

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.GreekLowerCaseFilterFactory"/>
</analyzer>
```

Hindi

Solr includes support for stemming Hindi following [this algorithm](#) (PDF), support for common spelling differences through the solr.HindiNormalizationFilterFactory, support for encoding differences through the solr.IndicNormalizationFilterFactory following [this algorithm](#), and Lucene includes an example stopword list.

Factory classes: solr.IndicNormalizationFilterFactory, solr.HindiNormalizationFilterFactory, solr.HindiStemFilterFactory

Arguments: None

Example:

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.IndicNormalizationFilterFactory"/>
  <filter class="solr.HindiNormalizationFilterFactory"/>
  <filter class="solr.HindiStemFilterFactory"/>
</analyzer>
```

Indonesian

Solr includes support for stemming Indonesian (Bahasa Indonesia) following [this algorithm](#) (PDF), and Lucene includes an example stopword list.

Factory class: solr.IndonesianStemFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solrLowerCaseFilterFactory"/>
  <filter class="solr.IndonesianStemFilterFactory" stemDerivational="true" />
</analyzer>
```

In: "sebagai sebagainya"

Tokenizer to Filter: "sebagai", "sebagainya"

Out: "bagai", "bagai"

Italian

Solr includes two stemmers for Italian: one in the `solr.SnowballPorterFilterFactory` `language="Italian"`, and a lighter stemmer called `solr.ItalianLightStemFilterFactory`. Lucene includes an example stopword list.

Factory class: `solr.ItalianStemFilterFactory`

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solrLowerCaseFilterFactory"/>
  <filter class="solr.ElisionFilterFactory"
    articles="lang/contractions_it.txt"/>
  <filter class="solr.ItalianLightStemFilterFactory"/>
</analyzer>
```

In: "propaga propagare propagamento"

Tokenizer to Filter: "propaga", "propagare", "propagamento"

Out: "propag", "propag", "propag"

Irish

Solr can stem Irish using the Snowball Porter Stemmer with an argument of `language="Irish"`. Solr includes `solr.IrishLowerCaseFilterFactory`, which can handle Irish-specific constructs. Solr also includes a set of contractions for Irish which can be stripped using `solr.ElisionFilterFactory`.

Factory class: `solr.SnowballPorterFilterFactory`

Arguments:

language

(required) stemmer language, "Irish" in this case

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ElisionFilterFactory"
    articles="lang/contractions_ga.txt"/>
  <filter class="solr.IrishLowerCaseFilterFactory"/>
  <filter class="solr.SnowballPorterFilterFactory" language="Irish" />
</analyzer>
```

In: "siopadóireacht síceapatacha b'fhearr m'athair"

Tokenizer to Filter: "siopadóireacht", "síceapatacha", "b'fhearr", "m'athair"

Out: "siopadóir", "síceapaite", "fearr", "athair"

Japanese

Solr includes support for analyzing Japanese, via the Lucene Kuromoji morphological analyzer, which includes several analysis components - more details on each below:

- JapaneseIterationMarkCharFilter normalizes Japanese horizontal iteration marks (odoriji) to their expanded form.
- JapaneseTokenizer tokenizes Japanese using morphological analysis, and annotates each term with part-of-speech, base form (a.k.a. lemma), reading and pronunciation.
- JapaneseBaseFormFilter replaces original terms with their base forms (a.k.a. lemmas).
- JapanesePartOfSpeechStopFilter removes terms that have one of the configured parts-of-speech.
- JapaneseKatakanaStemFilter normalizes common katakana spelling variations ending in a long sound character (U+30FC) by removing the long sound character.

Also useful for Japanese analysis, from lucene-analyzers-common:

- CJKWidthFilter folds fullwidth ASCII variants into the equivalent Basic Latin forms, and folds halfwidth Katakana variants into their equivalent fullwidth forms.

Japanese Iteration Mark CharFilter

Normalizes horizontal Japanese iteration marks (odoriji) to their expanded form. Vertical iteration marks are not supported.

Factory class: JapaneseIterationMarkCharFilterFactory

Arguments:

normalizeKanji

set to false to not normalize kanji iteration marks (default is true)

normalizeKana

set to false to not normalize kana iteration marks (default is true)

Japanese Tokenizer

Tokenizer for Japanese that uses morphological analysis, and annotates each term with part-of-speech, base form (a.k.a. lemma), reading and pronunciation.

JapaneseTokenizer has a search mode (the default) that does segmentation useful for search: a heuristic is used to segment compound terms into their constituent parts while also keeping the original compound terms as synonyms.

Factory class: solr.JapaneseTokenizerFactory

Arguments:

mode

Use search mode to get a noun-decompounding effect useful for search. search mode improves segmentation for search at the expense of part-of-speech accuracy. Valid values for mode are:

- normal: default segmentation
- search: segmentation useful for search (extra compound splitting)
- extended: search mode plus unigramming of unknown words (experimental)

For some applications it might be good to use search mode for indexing and normal mode for queries to increase precision and prevent parts of compounds from being matched and highlighted.

userDictionary

filename for a user dictionary, which allows overriding the statistical model with your own entries for segmentation, part-of-speech tags and readings without a need to specify weights. See lang/userdict_ja.txt for a sample user dictionary file.

userDictionaryEncoding

user dictionary encoding (default is UTF-8)

discardPunctuation

set to false to keep punctuation, true to discard (the default)

Japanese Base Form Filter

Replaces original terms' text with the corresponding base form (lemma). (JapaneseTokenizer annotates each term with its base form.)

Factory class: JapaneseBaseFormFilterFactory

(no arguments)

Japanese Part Of Speech Stop Filter

Removes terms with one of the configured parts-of-speech. JapaneseTokenizer annotates terms with parts-of-speech.

Factory class : JapanesePartOfSpeechStopFilterFactory

Arguments:

tags

filename for a list of parts-of-speech for which to remove terms; see conf/lang/stoptags_ja.txt in the sample_techproducts_config configset for an example.

enablePositionIncrements

if luceneMatchVersion is 4.3 or earlier and enablePositionIncrements="false", no position holes will be left by this filter when it removes tokens. **This argument is invalid if luceneMatchVersion is 5.0 or later.**

Japanese Katakana Stem Filter

Normalizes common katakana spelling variations ending in a long sound character (U+30FC) by removing the long sound character.

solr.CJKWidthFilterFactory should be specified prior to this filter to normalize half-width katakana to full-width.

Factory class: JapaneseKatakanaStemFilterFactory

Arguments:

minimumLength

terms below this length will not be stemmed. Default is 4, value must be 2 or more.

CJK Width Filter

Folds fullwidth ASCII variants into the equivalent Basic Latin forms, and folds halfwidth Katakana variants into their equivalent fullwidth forms.

Factory class: CJKWidthFilterFactory

(no arguments)

Example:

```
<fieldType name="text_ja" positionIncrementGap="100" autoGeneratePhraseQueries="false">
  <analyzer>
    <!-- Uncomment if you need to handle iteration marks: -->
    <!-- <charFilter class="solr.JapaneseIterationMarkCharFilterFactory" /> -->
    <tokenizer class="solr.JapaneseTokenizerFactory" mode="search" userDictionary=
      "lang/userdict_ja.txt"/>
    <filter class="solr.JapaneseBaseFormFilterFactory"/>
    <filter class="solr.JapanesePartOfSpeechStopFilterFactory" tags="lang/stoptags_ja.txt"/>
    <filter class="solr.CJKWidthFilterFactory"/>
    <filter class="solr.StopFilterFactory" ignoreCase="true" words="lang/stopwords_ja.txt"/>
    <filter class="solr.JapaneseKatakanaStemFilterFactory" minimumLength="4"/>
    <filter class="solrLowerCaseFilterFactory"/>
  </analyzer>
</fieldType>
```

Hebrew, Lao, Myanmar, Khmer

Lucene provides support, in addition to UAX#29 word break rules, for Hebrew's use of the double and single quote characters, and for segmenting Lao, Myanmar, and Khmer into syllables with the solr.ICUTokenizerFactory in the analysis-extras contrib module. To use this tokenizer, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See solr/contrib/analysis-extras/README.txt for instructions on which jars you need to add.

See [the ICUTokenizer](#) for more information.

Latvian

Solr includes support for stemming Latvian, and Lucene includes an example stopword list.

Factory class: solr.LatvianStemFilterFactory

Arguments: None

Example:

```
<fieldType name="text_lvstem" class="solr.TextField" positionIncrementGap="100">
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solrLowerCaseFilterFactory"/>
    <filter class="solr.LatvianStemFilterFactory"/>
  </analyzer>
</fieldType>
```

In: "tirgiem tergus"

Tokenizer to Filter: "tirgiem", "tergus"

Out: "tīg", "tīg"

Norwegian

Solr includes two classes for stemming Norwegian, NorwegianLightStemFilterFactory and NorwegianMinimalStemFilterFactory. Lucene includes an example stopword list.

Another option is to use the Snowball Porter Stemmer with an argument of language="Norwegian".

Also relevant are the [Scandinavian normalization filters](#).

Norwegian Light Stemmer

The NorwegianLightStemFilterFactory requires a "two-pass" sort for the -dom and -het endings. This means that in the first pass the word "kristendom" is stemmed to "kristen", and then all the general rules apply so it will be further stemmed to "krist". The effect of this is that "kristen," "kristendom," "kristendommen," and "kristendommens" will all be stemmed to "krist."

The second pass is to pick up -dom and -het endings. Consider this example:

One pass		Two passes	
Before	After	Before	After
forlegen	forleg	forlegen	forleg
forlegenhet	forlegen	forlegenhet	forleg
forlegenheten	forlegen	forlegenheten	forleg
forlegenhetens	forlegen	forlegenhetens	forleg
firkantet	firkant	firkantet	firkant
firkantethet	firkantet	firkantethet	firkant
firkantetheten	firkantet	firkantetheten	firkant

Factory class: solr.NorwegianLightStemFilterFactory

Arguments:

variant

Choose the Norwegian language variant to use. Valid values are:

- nb: Bokmål (default)
- nn: Nynorsk
- no: both

Example:

```
<fieldType name="text_no" class="solr.TextField" positionIncrementGap="100">
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.StopFilterFactory" ignoreCase="true" words="lang/stopwords_no.txt"
format="snowball"/>
    <filter class="solr.NorwegianLightStemFilterFactory"/>
  </analyzer>
</fieldType>
```

In: "Forelskelsen"

Tokenizer to Filter: "forelskelsen"

Out: "forelske"

Norwegian Minimal Stemmer

The NorwegianMinimalStemFilterFactory stems plural forms of Norwegian nouns only.

Factory class: solr.NorwegianMinimalStemFilterFactory

Arguments:

variant

Choose the Norwegian language variant to use. Valid values are:

- nb: Bokmål (default)
- nn: Nynorsk
- no: both

Example:

```
<fieldType name="text_no" class="solr.TextField" positionIncrementGap="100">
  <analyzer>
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.StopFilterFactory" ignoreCase="true" words="lang/stopwords_no.txt"
format="snowball"/>
    <filter class="solr.NorwegianMinimalStemFilterFactory"/>
  </analyzer>
</fieldType>
```

In: "Bilens"

Tokenizer to Filter: "bilens"

Out: "bil"

Persian

Persian Filter Factories

Solr includes support for normalizing Persian, and Lucene includes an example stopword list.

Factory class: solr.PersianNormalizationFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ArabicNormalizationFilterFactory"/>
  <filter class="solr.PersianNormalizationFilterFactory">
</analyzer>
```

Polish

Solr provides support for Polish stemming with the solr.StempelPolishStemFilterFactory, and solr.MorphologikFilterFactory for lemmatization, in the contrib/analysis-extras module. The solr.StempelPolishStemFilterFactory component includes an algorithmic stemmer with tables for Polish. To use either of these filters, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See solr/contrib/analysis-extras/README.txt for instructions on which jars you need to add.

Factory class: solr.StempelPolishStemFilterFactory and solr.MorfologikFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solrLowerCaseFilterFactory"/>
  <filter class="solrStempelPolishStemFilterFactory"/>
</analyzer>
```

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.MorfologikFilterFactory" dictionary="morfologik/stemming/polish/polish.dict"/>
  <filter class="solrLowerCaseFilterFactory"/>
</analyzer>
```

In: "studenta studenci"

Tokenizer to Filter: "studenta", "studenci"

Out: "student", "student"

More information about the Stempel stemmer is available in [the Lucene javadocs](#).

Note the lower case filter is applied *after* the Morfologik stemmer; this is because the Polish dictionary contains proper names and then proper term case may be important to resolve disambiguiities (or even lookup the correct lemma at all).

The Morfologik dictionary parameter value is a constant specifying which dictionary to choose. The dictionary resource must be named path/to/language.dict and have an associated .info metadata file. See [the Morfologik project](#) for details. If the dictionary attribute is not provided, the Polish dictionary is loaded and used by default.

Portuguese

Solr includes four stemmers for Portuguese: one in the solr.SnowballPorterFilterFactory, an alternative stemmer called solr.PortugueseStemFilterFactory, a lighter stemmer called solr.PortugueseLightStemFilterFactory, and an even less aggressive stemmer called solr.PortugueseMinimalStemFilterFactory. Lucene includes an example stopword list.

Factory classes: solr.PortugueseStemFilterFactory, solr.PortugueseLightStemFilterFactory, solr.PortugueseMinimalStemFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.PortugueseStemFilterFactory"/>
</analyzer>
```

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.PortugueseLightStemFilterFactory"/>
</analyzer>
```

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.PortugueseMinimalStemFilterFactory"/>
</analyzer>
```

In: "praia praias"

Tokenizer to Filter: "praia", "praias"

Out: "pra", "pra"

Romanian

Solr can stem Romanian using the Snowball Porter Stemmer with an argument of language="Romanian".

Factory class: solr.SnowballPorterFilterFactory

Arguments:

language
(required) stemmer language, "Romanian" in this case

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.SnowballPorterFilterFactory" language="Romanian" />
</analyzer>
```

Russian

Russian Stem Filter

Solr includes two stemmers for Russian: one in the solr.SnowballPorterFilterFactory language="Russian", and a lighter stemmer called solr.RussianLightStemFilterFactory. Lucene includes an example stopword list.

Factory class: solr.RussianLightStemFilterFactory

Arguments: None

Example:

```
<analyzer type="index">
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.RussianLightStemFilterFactory"/>
</analyzer>
```

Scandinavian

Scandinavian is a language group spanning three languages [Norwegian](#), [Swedish](#) and [Danish](#) which are very similar.

Swedish å, ä, ö are in fact the same letters as Norwegian and Danish å, æ, ø and thus interchangeable when used between these languages. They are however folded differently when people type them on a keyboard lacking these characters.

In that situation almost all Swedish people use a, a, o instead of å, ä, ö. Norwegians and Danes on the other hand usually type aa, ae and oe instead of å, æ and ø. Some do however use a, a, o, oo, ao and sometimes permutations of everything above.

There are two filters for helping with normalization between Scandinavian languages: one is solr.ScandinavianNormalizationFilterFactory trying to preserve the special characters (æääöå) and another solr.ScandinavianFoldingFilterFactory which folds these to the more broad ø/ö->o, etc.

See also each language section for other relevant filters.

Scandinavian Normalization Filter

This filter normalize use of the interchangeable Scandinavian characters æääööøø and folded variants (aa, ao, ae, oe and oo) by transforming them to åÅæÆøØ.

It's a semantically less destructive solution than ScandinavianFoldingFilter, most useful when a person with a Norwegian or Danish keyboard queries a Swedish index and vice versa. This filter does **not** perform the common Swedish folds of å and ä to a nor ö to o.

Factory class: solr.ScandinavianNormalizationFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.ScandinavianNormalizationFilterFactory"/>
</analyzer>
```

In: "blåbærsyltetøj blåbærsvytetøj blaabaarsyltetoej blabarsyltetoj"

Tokenizer to Filter: "blåbærsyltetøj", "blåbærsvytetøj", "blaabaersyltetoej", "blabarsyltetoj"

Out: "blåbærsyltetøj", "blåbærsvytetøj", "blåbærsvytetøj", "blabarsyltetoj"

Scandinavian Folding Filter

This filter folds Scandinavian characters åÅæÆ->a and öÖøØ->o. It also discriminate against use of double vowels aa, ae, ao, oe and oo, leaving just the first one.

It's a semantically more destructive solution than ScandinavianNormalizationFilter, but can in addition help with matching raksmorgas as räksmörgås.

Factory class: solr.ScandinavianFoldingFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solrLowerCaseFilterFactory"/>
  <filter class="solr.ScandinavianFoldingFilterFactory"/>
</analyzer>
```

In: "blåbærsyltetøj blåbærsyltetøj blaabaarsyltetoej blabarsyltetoj"

Tokenizer to Filter: "blåbærsyltetøj", "blåbærsyltetøj", "blaabaersyltetoej", "blabarsyltetoj"

Out: "blabarsyltetoj", "blabarsyltetoj", "blabarsyltetoj", "blabarsyltetoj"

Serbian

Serbian Normalization Filter

Solr includes a filter that normalizes Serbian Cyrillic and Latin characters. Note that this filter only works with lowercased input.

See the Solr wiki for tips & advice on using this filter: <https://wiki.apache.org/solr/SerbianLanguageSupport>

Factory class: solr.SerbianNormalizationFilterFactory

Arguments:

haircut

Select the extend of normalization. Valid values are:

- bald: (Default behavior) Cyrillic characters are first converted to Latin; then, Latin characters have their diacritics removed, with the exception of **LATIN SMALL LETTER D WITH STROKE** (U+0111) which is converted to "dj"
- regular: Only Cyrillic to Latin normalization will be applied, preserving the Latin diacritics

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solrLowerCaseFilterFactory"/>
  <filter class="solr.SerbianNormalizationFilterFactory" haircut="bald"/>
</analyzer>
```

Spanish

Solr includes two stemmers for Spanish: one in the `solr.SnowballPorterFilterFactory` language="Spanish", and a lighter stemmer called `solr.SpanishLightStemFilterFactory`. Lucene includes an example stopword list.

Factory class: solr.SpanishStemFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.SpanishLightStemFilterFactory"/>
</analyzer>
```

In: "torear toreada toreado"

Tokenizer to Filter: "torear", "toreada", "toreado"

Out: "tor", "tor", "tor"

Swedish

Swedish Stem Filter

Solr includes two stemmers for Swedish: one in the `solr.SnowballPorterFilterFactory` language="Swedish", and a lighter stemmer called `solr.SwedishLightStemFilterFactory`. Lucene includes an example stopword list.

Also relevant are the [Scandinavian normalization filters](#).

Factory class: `solr.SwedishStemFilterFactory`

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.SwedishLightStemFilterFactory"/>
</analyzer>
```

In: "kloke klokhet klokheten"

Tokenizer to Filter: "kloke", "klokhet", "klokheten"

Out: "klok", "klok", "klok"

Thai

This filter converts sequences of Thai characters into individual Thai words. Unlike European languages, Thai does not use whitespace to delimit words.

Factory class: `solr.ThaiTokenizerFactory`

Arguments: None

Example:

```
<analyzer type="index">
  <tokenizer class="solr.ThaiTokenizerFactory"/>
  <filter class="solr.LowerCaseFilterFactory"/>
</analyzer>
```

Turkish

Solr includes support for stemming Turkish with the solr.SnowballPorterFilterFactory; support for case-insensitive search with the solr.TurkishLowerCaseFilterFactory; support for stripping apostrophes and following suffixes with solr.ApostropheFilterFactory (see [Role of Apostrophes in Turkish Information Retrieval](#)); support for a form of stemming that truncating tokens at a configurable maximum length through the solr.TruncateTokenFilterFactory (see [Information Retrieval on Turkish Texts](#)); and Lucene includes an example stopword list.

Factory class: solr.TurkishLowerCaseFilterFactory

Arguments: None

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ApostropheFilterFactory"/>
  <filter class="solr.TurkishLowerCaseFilterFactory"/>
  <filter class="solr.SnowballPorterFilterFactory" language="Turkish"/>
</analyzer>
```

Another example, illustrating diacritics-insensitive search:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ApostropheFilterFactory"/>
  <filter class="solr.TurkishLowerCaseFilterFactory"/>
  <filter class="solr.ASCIIFFoldingFilterFactory" preserveOriginal="true"/>
  <filter class="solr.KeywordRepeatFilterFactory"/>
  <filter class="solr.TruncateTokenFilterFactory" prefixLength="5"/>
  <filter class="solr.RemoveDuplicatesTokenFilterFactory"/>
</analyzer>
```

Ukrainian

Solr provides support for Ukrainian lemmatization with the solr.MorphologikFilterFactory, in the contrib/analysis-extras module. To use this filter, you must add additional .jars to Solr's classpath (as described in the section [Resources and Plugins on the Filesystem](#)). See solr/contrib/analysis-extras/README.txt for instructions on which jars you need to add.

Lucene also includes an example Ukrainian stopword list, in the lucene-analyzers-morfologik jar.

Factory class: solr.MorfologikFilterFactory

Arguments:

dictionary

(required) lemmatizer dictionary - the lucene-analyzers-morfologik jar contains a Ukrainian dictionary at org/apache/lucene/analysis/uk/ukrainian.dict.

Example:

```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.StopFilterFactory" words="org/apache/lucene/analysis/uk/stopwords.txt"/>
  <filter class="solr.LowerCaseFilterFactory"/>
  <filter class="solr.MorfologikFilterFactory" dictionary=
"org/apache/lucene/analysis/uk/ukrainian.dict"/>
</analyzer>
```

The Morfologik dictionary parameter value is a constant specifying which dictionary to choose. The dictionary resource must be named path/to/language.dict and have an associated .info metadata file. See [the Morfologik project](#) for details. If the dictionary attribute is not provided, the Polish dictionary is loaded and used by default.

Phonetic Matching

Phonetic matching algorithms may be used to encode tokens so that two different spellings that are pronounced similarly will match.

For overviews of and comparisons between algorithms, see http://en.wikipedia.org/wiki/Phonetic_algorithm and <http://ntz-develop.blogspot.com/2011/03/phonetic-algorithms.html>

Beider-Morse Phonetic Matching (BMPM)

For examples of how to use this encoding in your analyzer, see [Beider Morse Filter](#) in the Filter Descriptions section.

Beider-Morse Phonetic Matching (BMPM) is a "soundalike" tool that lets you search using a new phonetic matching system. BMPM helps you search for personal names (or just surnames) in a Solr/Lucene index, and is far superior to the existing phonetic codecs, such as regular soundex, metaphone, caverphone, etc.

In general, phonetic matching lets you search a name list for names that are phonetically equivalent to the desired name. BMPM is similar to a soundex search in that an exact spelling is not required. Unlike soundex, it does not generate a large quantity of false hits.

From the spelling of the name, BMPM attempts to determine the language. It then applies phonetic rules for that particular language to transliterate the name into a phonetic alphabet. If it is not possible to determine the language with a fair degree of certainty, it uses generic phonetic instead. Finally, it applies language-independent rules regarding such things as voiced and unvoiced consonants and vowels to further insure the reliability of the matches.

For example, assume that the matches found when searching for Stephen in a database are "Stefan", "Steph", "Stephen", "Steve", "Steven", "Stove", and "Stuffin". "Stefan", "Stephen", and "Steven" are probably relevant, and are names that you want to see. "Stuffin", however, is probably not relevant. Also rejected were "Steph", "Steve", and "Stove". Of those, "Stove" is probably not one that we would have wanted. But "Steph" and "Steve" are possibly ones that you might be interested in.

For Solr, BMPM searching is available for the following languages:

- English
- French
- German
- Greek
- Hebrew written in Hebrew letters
- Hungarian
- Italian
- Polish
- Romanian
- Russian written in Cyrillic letters
- Russian transliterated into English letters

- Spanish
- Turkish

The name matching is also applicable to non-Jewish surnames from the countries in which those languages are spoken.

For more information, see here: <http://stevemorse.org/phoneticinfo.htm> and <http://stevemorse.org/phonetics/bmpm.htm>.

Daitch-Mokotoff Soundex

To use this encoding in your analyzer, see [Daitch-Mokotoff Soundex Filter](#) in the Filter Descriptions section.

The Daitch-Mokotoff Soundex algorithm is a refinement of the Russel and American Soundex algorithms, yielding greater accuracy in matching especially Slavic and Yiddish surnames with similar pronunciation but differences in spelling.

The main differences compared to the other soundex variants are:

- coded names are 6 digits long
- initial character of the name is coded
- rules to encoded multi-character n-grams
- multiple possible encodings for the same name (branching)

Note: the implementation used by Solr (commons-codec's DaitchMokotoffSoundex) has additional branching rules compared to the original description of the algorithm.

For more information, see http://en.wikipedia.org/wiki/Daitch%E2%80%93Mokotoff_Soundex and <http://www.avotaynu.com/soundex.htm>

Double Metaphone

To use this encoding in your analyzer, see [Double Metaphone Filter](#) in the Filter Descriptions section.

Alternatively, you may specify encoder="DoubleMetaphone" with the [Phonetic Filter](#), but note that the Phonetic Filter version will **not** provide the second ("alternate") encoding that is generated by the Double Metaphone Filter for some tokens.

Encodes tokens using the double metaphone algorithm by Lawrence Philips. See the original article at <http://www.drdobbs.com/the-double-metaphone-search-algorithm/184401251?pgno=2>

Metaphone

To use this encoding in your analyzer, specify encoder="Metaphone" with the [Phonetic Filter](#).

Encodes tokens using the Metaphone algorithm by Lawrence Philips, described in "Hanging on the Metaphone" in Computer Language, Dec. 1990.

Another reference for more information is [Double Metaphone Search Algorithm](#), by Lawrence Philips.

Soundex

To use this encoding in your analyzer, specify encoder="Soundex" with the [Phonetic Filter](#).

Encodes tokens using the Soundex algorithm, which is used to relate similar names, but can also be used as a general purpose scheme to find words with similar phonemes.

See also <http://en.wikipedia.org/wiki/Soundex>.

Refined Soundex

To use this encoding in your analyzer, specify encoder="RefinedSoundex" with the [Phonetic Filter](#).

Encodes tokens using an improved version of the Soundex algorithm.

See <http://en.wikipedia.org/wiki/Soundex>.

Caverphone

To use this encoding in your analyzer, specify encoder="Caverphone" with the [Phonetic Filter](#).

Caverphone is an algorithm created by the Caversham Project at the University of Otago. The algorithm is optimised for accents present in the southern part of the city of Dunedin, New Zealand.

See <http://en.wikipedia.org/wiki/Caverphone> and the Caverphone 2.0 specification at <http://caversham.otago.ac.nz/files/working/ctp150804.pdf>

Kölner Phonetik a.k.a. Cologne Phonetic

To use this encoding in your analyzer, specify encoder="ColognePhonetic" with the [Phonetic Filter](#).

The Kölner Phonetik, an algorithm published by Hans Joachim Postel in 1969, is optimized for the German language.

See http://de.wikipedia.org/wiki/K%C3%B6lner_Phonetik

NYSIIS

To use this encoding in your analyzer, specify encoder="Nysiis" with the [Phonetic Filter](#).

NYSIIS is an encoding used to relate similar names, but can also be used as a general purpose scheme to find words with similar phonemes.

See <http://en.wikipedia.org/wiki/NYSIIS> and <http://www.dropby.com/NYSIIS.html>

Running Your Analyzer

Once you've [defined a field type in your Schema](#), and specified the analysis steps that you want applied to it, you should test it out to make sure that it behaves the way you expect it to.

Luckily, there is a very handy page in the Solr [admin interface](#) that lets you do just that. You can invoke the analyzer for any text field, provide sample input, and display the resulting token stream.

For example, let's look at some of the "Text" field types available in the bin/solr -e techproducts example configuration, and use the [Analysis Screen](#)

(<http://localhost:8983/solr/#/techproducts/analysis>) to compare how the tokens produced at index time for the sentence "Running an Analyzer" match up with a slightly different query text of "run my analyzer"

We can begin with "text_ws" - one of the most simplified Text field types available:

WT	text	Running	an	Analyzer
raw_bytes	[52 75 6e 6e 69 6e 67]	[61 6e]	[41 6e 61 6c 79 7a 65 72]	
start	0	8	11	
end	7	10	19	
positionLength	1	1	1	
type	word	word	word	
position	1	2	3	

WT	text	run	my	analyzer
raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]	
start	0	4	7	
end	3	6	15	
positionLength	1	1	1	
type	word	word	word	
position	1	2	3	

By looking at the start and end positions for each term, we can see that the only thing this field type does is tokenize text on whitespace. Notice in this image that the term "Running" has a start position of 0 and an end position of 7, while "an" has a start position of 8 and an end position of 10, and "Analyzer" starts at 11 and ends at 19. If the whitespace between the terms was also included, the count would be 21; since it is 19, we know that whitespace has been removed from this query.

Note also that the indexed terms and the query terms are still very different. "Running" doesn't match "run", "Analyzer" doesn't match "analyzer" (to a computer), and obviously "an" and "my" are totally different words. If our objective is to allow queries like "run my analyzer" to match indexed text like "Running an Analyzer" then we will evidently need to pick a different field type with index & query time text analysis that does more processing of the inputs.

In particular we will want:

- Case insensitivity, so "Analyzer" and "analyzer" match.
- Stemming, so words like "Run" and "Running" are considered equivalent terms.
- Stop Word Pruning, so small words like "an" and "my" don't affect the query.

For our next attempt, let's try the "text_general" field type:

Solr

Dashboard Logging Core Admin Java Properties Thread Dump techproducts Overview Analysis Dataimport Documents Files Ping Plugins / Stats Query Replication Schema Browser

Analyse Fieldname / FieldType: text_general

Field Value (Index)				
Running an Analyzer				
ST	text	Running	an	Analyzer
	raw_bytes	[52 75 6e 6e 69 6e 67]	[61 6e]	[41 6e 61 6c 79 7a 65 72]
	start	0	8	11
	end	7	10	19
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
SF	text	Running	an	Analyzer
	raw_bytes	[52 75 6e 6e 69 6e 67]	[61 6e]	[41 6e 61 6c 79 7a 65 72]
	start	0	8	11
	end	7	10	19
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
LCF	text	running	an	analyzer
	raw_bytes	[72 75 6e 6e 69 6e 67]	[61 6e]	[61 6e 61 6c 79 7a 65 72]
	start	0	8	11
	end	7	10	19
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3

Field Value (Query)

run my analyzer

Verbose Output

▼ Analyse Values

Field Value (Query)				
run my analyzer				
ST	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
SF	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
SF	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
LCF	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3

With the verbose output enabled, we can see how each stage of our new analyzers modify the tokens they receive before passing them on to the next stage. As we scroll down to the final output, we can see that we do start to get a match on "analyzer" from each input string, thanks to the "LCF" stage — which if you hover over with your mouse, you'll see is the "LowerCaseFilter":

Solr

Dashboard Logging Core Admin Java Properties Thread Dump techproducts Overview Analysis Dataimport Documents Files Ping Plugins / Stats Query Replication Schema Browser

Field Value (Index)				
end	7	10	19	
positionLength	1	1	1	
type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	
position	1	2	3	
SF	text	Running	an	Analyzer
	raw_bytes	[52 75 6e 6e 69 6e 67]	[61 6e]	[41 6e 61 6c 79 7a 65 72]
	start	0	8	11
	end	7	10	19
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
LCF	text	running	an	analyzer
	raw_bytes	[72 75 6e 6e 69 6e 67]	[61 6e]	[61 6e 61 6c 79 7a 65 72]
	start	0	8	11
	end	7	10	19
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3

Field Value (Query)				
end	3	6	15	
positionLength	1	1	1	
type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	
position	1	2	3	
SF	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
SF	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
LCF	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3

The "text_general" field type is designed to be generally useful for any language, and it has definitely gotten us closer to our objective than "text_ws" from our first example by solving the problem of case sensitivity. It's still not quite what we are looking for because we don't see stemming or stopword rules being applied. So now let us try the "text_en" field type:

Field Value (Index)				
Running an Analyzer				
ST	text	Running	an	Analyzer
	raw_bytes	[52 75 6e 6e 69 6e 67]	[61 6e]	[41 6e 61 6c 79 7a 65 72]
	start	0	8	11
	end	7	10	19
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
SF	text	Running		Analyzer
	raw_bytes	[52 75 6e 6e 69 6e 67]		[41 6e 61 6c 79 7a 65 72]
	start	0	11	
	end	7	19	
	positionLength	1	1	
	type	<ALPHANUM>	<ALPHANUM>	
	position	1	3	
LCF	text	running		analyzer
	raw_bytes	[72 75 6e 6e 69 6e 67]		[61 6e 61 6c 79 7a 65 72]
	start	0	11	
	end	7	19	
	positionLength	1	1	

Field Value (Query)				
run my analyzer				
<input checked="" type="checkbox"/> Verbose Output				
				Analyse Values
ST	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
SF	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1

Now we can see the "SF" (StopFilter) stage of the analyzers solving the problem of removing Stop Words ("an"), and as we scroll down, we also see the "PSF" (PorterStemFilter) stage apply stemming rules suitable for our English language input, such that the terms produced by our "index analyzer" and the terms produced by our "query analyzer" match the way we expect.

Field Value (Index)				
SKMF	type	<ALPHANUM>		
	position	1	3	
	text	running	analyzer	
	raw_bytes	[72 75 6e 6e 69 6e 67]	[61 6e 61 6c 79 7a 65 72]	
	keyword	false	false	
	start	0	11	
	end	7	19	
	positionLength	1	1	
	type	<ALPHANUM>	<ALPHANUM>	
	position	1	3	
PSF	text	run	analyz	
	raw_bytes	[72 75 6e]	[61 6e 61 6c 79 7a]	
	start	0	11	
	end	7	19	
	positionLength	1	1	
	type	<ALPHANUM>	<ALPHANUM>	
	keyword	false	false	
	position	1	3	
EPF	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
SKMF	text	run	my	analyzer
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a 65 72]
	keyword	false	false	
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	position	1	2	3
PSF	text	run	my	analyz
	raw_bytes	[72 75 6e]	[6d 79]	[61 6e 61 6c 79 7a]
	start	0	4	7
	end	3	6	15
	positionLength	1	1	1
	type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
	keyword	false	false	
	position	1	2	3

At this point, we can continue to experiment with additional inputs, verifying that our analyzers produce matching tokens when we expect them to match, and disparate tokens when we do not expect them to match, as we iterate and tweak our field type configuration.

Indexing and Basic Data Operations

This section describes how Solr adds data to its index. It covers the following topics:

- [Introduction to Solr Indexing](#): An overview of Solr's indexing process.
- [Post Tool](#): Information about using post.jar to quickly upload some content to your system.
- [Uploading Data with Index Handlers](#): Information about using Solr's Index Handlers to upload XML/XSLT, JSON and CSV data.
- [Transforming and Indexing Custom JSON](#): Index any JSON of your choice
- [Indexing Nested Documents](#): Detailed information about indexing and schema configuration for nested documents.
- [Uploading Data with Solr Cell using Apache Tika](#): Information about using the Solr Cell framework to upload data for indexing.
- [Uploading Structured Data Store Data with the Data Import Handler](#): Information about uploading and indexing data from a structured data store.
- [Updating Parts of Documents](#): Information about how to use atomic updates and optimistic concurrency with Solr.
- [Detecting Languages During Indexing](#): Information about using language identification during the indexing process.
- [De-Duplication](#): Information about configuring Solr to mark duplicate documents as they are indexed.
- [Content Streams](#): Information about streaming content to Solr Request Handlers.
- [Reindexing](#): Details about when reindexing is required or recommended, and some strategies for completely reindexing your documents.

Indexing Using Client APIs

Using client APIs, such as [SolrJ](#), from your applications is an important option for updating Solr indexes. See the [Client APIs](#) section for more information.

Introduction to Solr Indexing

This section describes the process of indexing: adding content to a Solr index and, if necessary, modifying that content or deleting it.

By adding content to an index, we make it searchable by Solr.

A Solr index can accept data from many different sources, including XML files, comma-separated value (CSV) files, data extracted from tables in a database, and files in common file formats such as Microsoft Word or PDF.

Here are the three most common ways of loading data into a Solr index:

- Using the [Solr Cell](#) framework built on Apache Tika for ingesting binary files or structured files such as Office, Word, PDF, and other proprietary formats.
- Uploading XML files by sending HTTP requests to the Solr server from any environment where such requests can be generated.
- Writing a custom Java application to ingest data through Solr's Java Client API (which is described in more detail in [Client APIs](#)). Using the Java API may be the best choice if you're working with an application, such as a Content Management System (CMS), that offers a Java API.

Regardless of the method used to ingest data, there is a common basic data structure for data being fed into a Solr index: a *document* containing multiple *fields*, each with a *name* and containing *content*, which may be empty. One of the fields is usually designated as a unique ID field (analogous to a primary key in a database), although the use of a unique ID field is not strictly required by Solr.

If the field name is defined in the Schema that is associated with the index, then the analysis steps associated with that field will be applied to its content when the content is tokenized. Fields that are not explicitly defined in the Schema will either be ignored or mapped to a dynamic field definition (see [Documents, Fields, and Schema Design](#)), if one matching the field name exists.

For more information on indexing in Solr, see the [Solr Wiki](#).

The Solr Example Directory

When starting Solr with the "-e" option, the example/ directory will be used as base directory for the example Solr instances that are created. This directory also includes an example/exemplardocs/ subdirectory containing sample documents in a variety of formats that you can use to experiment with indexing into the various examples.

The curl Utility for Transferring Files

Many of the instructions and examples in this section make use of the curl utility for transferring content through a URL. curl posts and retrieves data over HTTP, FTP, and many other protocols. Most Linux distributions include a copy of curl. You'll find curl downloads for Linux, Windows, and many other operating systems at <http://curl.haxx.se/download.html>. Documentation for curl is available here: <http://curl.haxx.se/docs/manpage.html>.

Using curl or other command line tools for posting data is just fine for examples or tests, but it's not the recommended method for achieving the best performance for updates in production environments. You will achieve better performance with Solr Cell or the other methods described in this section.

Instead of curl, you can use utilities such as GNU wget (<http://www.gnu.org/software/wget/>) or manage GETs and POSTS with Perl, although the command line options will differ.



Post Tool

Solr includes a simple command line tool for POSTing various types of content to a Solr server.

The tool is bin/post. The bin/post tool is a Unix shell script; for Windows (non-Cygwin) usage, see the section [Post Tool Windows Support](#) below.

To run it, open a window and enter:

```
bin/post -c gettingstarted example/films/films.json
```

This will contact the server at localhost:8983. Specifying the collection/core name is **mandatory**. The -help (or simply -h) option will output information on its usage (i.e., bin/post -help).

Using the bin/post Tool

Specifying either the collection/core name or the full update url is **mandatory** when using bin/post.

The basic usage of bin/post is:

```
$ bin/post -h
Usage: post -c <collection> [OPTIONS] <files|directories|urls|-d ["...",...]>
      or post -help

      collection name defaults to DEFAULT_SOLR_COLLECTION if not specified

OPTIONS
=====
Solr options:
  -url <base Solr update URL> (overrides collection, host, and port)
  -host <host> (default: localhost)
  -p or -port <port> (default: 8983)
  -commit yes|no (default: yes)
  -u or -user <user:pass> (sets BasicAuth credentials)

Web crawl options:
  -recursive <depth> (default: 1)
  -delay <seconds> (default: 10)

Directory crawl options:
  -delay <seconds> (default: 0)

stdin/args options:
  -type <content/type> (default: application/xml)

Other options:
  -filetypes <type>[,<type>,...] (default:
    xml,json,csv,pdf,doc,docx,ppt,pptx,xls,xlsx,odt,odp,ods,ott,otp,ots,rtf,htm,html,txt,log)
  -params "<key>=<value>[&<key>=<value>...]" (values must be URL-encoded; these pass through to
    Solr update request)
  -out yes|no (default: no; yes outputs Solr response to console)
  ...
```

Examples Using bin/post

There are several ways to use bin/post. This section presents several examples.

Indexing XML

Add all documents with file extension .xml to collection or core named gettingstarted.

```
bin/post -c gettingstarted *.xml
```

Add all documents with file extension .xml to the gettingstarted collection/core on Solr running on port 8984.

```
bin/post -c gettingstarted -p 8984 *.xml
```

Send XML arguments to delete a document from `gettingstarted`.

```
bin/post -c gettingstarted -d '<delete><id>42</id></delete>'
```

Indexing CSV

Index all CSV files into `gettingstarted`:

```
bin/post -c gettingstarted *.csv
```

Index a tab-separated file into `gettingstarted`:

```
bin/post -c signals -params "separator=%09" -type text/csv data.tsv
```

The content type (`-type`) parameter is required to treat the file as the proper type, otherwise it will be ignored and a WARNING logged as it does not know what type of content a `.tsv` file is. The [CSV handler](#) supports the separator parameter, and is passed through using the `-params` setting.

Indexing JSON

Index all JSON files into `gettingstarted`.

```
bin/post -c gettingstarted *.json
```

Indexing Rich Documents (PDF, Word, HTML, etc.)

Index a PDF file into `gettingstarted`.

```
bin/post -c gettingstarted a.pdf
```

Automatically detect content types in a folder, and recursively scan it for documents for indexing into `gettingstarted`.

```
bin/post -c gettingstarted afolder/
```

Automatically detect content types in a folder, but limit it to PPT and HTML files and index into `gettingstarted`.

```
bin/post -c gettingstarted -filetypes ppt,html afolder/
```

Indexing to a Password Protected Solr (Basic Auth)

Index a PDF as the user "solr" with password "SolrRocks":

```
bin/post -u solr:SolrRocks -c gettingstarted a.pdf
```

Post Tool Windows Support

bin/post exists currently only as a Unix shell script, however it delegates its work to a cross-platform capable Java program. The [SimplePostTool](#) can be run directly in supported environments, including Windows.

SimplePostTool

The bin/post script currently delegates to a standalone Java program called SimplePostTool.

This tool, bundled into a executable JAR, can be run directly using java -jar example/exampledocs/post.jar. See the help output and take it from there to post files, recurse a website or file system folder, or send direct commands to a Solr server.

```
$ java -jar example/exampledocs/post.jar -h
SimplePostTool version 5.0.0
Usage: java [SystemProperties] -jar post.jar [-h|-] [<file|folder|url|arg>
[<file|folder|url|arg>...]]]
.
.
.
```

Uploading Data with Index Handlers

Index Handlers are Request Handlers designed to add, delete and update documents to the index. In addition to having plugins for importing rich documents [using Tika](#) or from structured data sources using the [Data Import Handler](#), Solr natively supports indexing structured documents in XML, CSV and JSON.

The recommended way to configure and use request handlers is with path based names that map to paths in the request url. However, request handlers can also be specified with the qt (query type) parameter if the requestDispatcher is appropriately configured. It is possible to access the same handler using more than one name, which can be useful if you wish to specify different sets of default options.

A single unified update request handler supports XML, CSV, JSON, and javabin update requests, delegating to the appropriate ContentStreamLoader based on the Content-Type of the [ContentStream](#).

UpdateRequestHandler Configuration

The default configuration file has the update request handler configured by default.

```
<requestHandler name="/update" class="solr.UpdateRequestHandler" />
```

XML Formatted Index Updates

Index update commands can be sent as XML message to the update handler using Content-type: application/xml or Content-type: text/xml.

Adding Documents

The XML schema recognized by the update handler for adding documents is very straightforward:

- The <add> element introduces one more documents to be added.
- The <doc> element introduces the fields making up a document.
- The <field> element presents the content for a specific field.

For example:

```
<add>
  <doc>
    <field name="authors">Patrick Eagar</field>
    <field name="subject">Sports</field>
    <field name="dd">796.35</field>
    <field name="numpages">128</field>
    <field name="desc"></field>
    <field name="price">12.40</field>
    <field name="title">Summer of the all-rounder: Test and championship cricket in England
1982</field>
    <field name="isbn">0002166313</field>
    <field name="yearpub">1982</field>
    <field name="publisher">Collins</field>
  </doc>
  <doc>
    ...
  </doc>
</add>
```

The add command supports some optional attributes which may be specified.

commitWithin

Add the document within the specified number of milliseconds.

overwrite

Default is true. Indicates if the unique key constraints should be checked to overwrite previous versions of the same document (see below).

If the document schema defines a unique key, then by default an /update operation to add a document will overwrite (i.e., replace) any document in the index with the same unique key. If no unique key has been defined, indexing performance is somewhat faster, as no check has to be made for an existing documents to replace.

If you have a unique key field, but you feel confident that you can safely bypass the uniqueness check (e.g., you build your indexes in batch, and your indexing code guarantees it never adds the same document more than once) you can specify the `overwrite="false"` option when adding your documents.

XML Update Commands

Commit and Optimize During Updates

The `<commit>` operation writes all documents loaded since the last commit to one or more segment files on the disk. Before a commit has been issued, newly indexed content is not visible to searches. The commit operation opens a new searcher, and triggers any event listeners that have been configured.

Commits may be issued explicitly with a `<commit/>` message, and can also be triggered from `<autocommit>` parameters in `solrconfig.xml`.

The `<optimize>` operation requests Solr to merge internal data structures. For a large index, optimization will take some time to complete, but by merging many small segment files into larger segments, search

performance may improve. If you are using Solr's replication mechanism to distribute searches across many systems, be aware that after an optimize, a complete index will need to be transferred.



You should only consider using optimize on static indexes, i.e., indexes that can be optimized as part of the regular update process (say once-a-day updates). Applications requiring NRT functionality should not use optimize.

The <commit> and <optimize> elements accept these optional attributes:

waitSearcher

Default is true. Blocks until a new searcher is opened and registered as the main query searcher, making the changes visible.

expungeDeletes

(commit only) Default is false. Merges segments that have more than 10% deleted docs, expunging the deleted documents in the process. Resulting segments will respect maxMergedSegmentMB.



expungeDeletes is "less expensive" than optimize, but the same warnings apply.

maxSegments

(optimize only) Default is unlimited, resulting segments respect the maxMergedSegmentMB setting. Makes a best effort attempt to merge the segments down to no more than this number of segments but does not guarantee that the goal will be achieved. Unless there is tangible evidence that optimizing to a small number of segments is beneficial, this parameter should be omitted and the default behavior accepted.

Here are examples of <commit> and <optimize> using optional attributes:

```
<commit waitSearcher="false"/>
<commit waitSearcher="false" expungeDeletes="true"/>
<optimize waitSearcher="false"/>
```

Delete Operations

Documents can be deleted from the index in two ways. "Delete by ID" deletes the document with the specified ID, and can be used only if a UniqueID field has been defined in the schema. It doesn't work for child/nested docs. "Delete by Query" deletes all documents matching a specified query, although commitWithin is ignored for a Delete by Query. A single delete message can contain multiple delete operations.

```
<delete>
  <id>0002166313</id>
  <id>0031745983</id>
  <query>subject:sport</query>
  <query>publisher:penguin</query>
</delete>
```



When using the Join query parser in a Delete By Query, you should use the score parameter with a value of "none" to avoid a ClassCastException. See the section on the [Join Query Parser](#) for more details on the score parameter.

Rollback Operations

The rollback command rolls back all add and deletes made to the index since the last commit. It neither calls any event listeners nor creates a new searcher. Its syntax is simple: <rollback/>.

Grouping Operations

You can post several commands in a single XML file by grouping them with the surrounding <update> element.

```
<update>
  <add>
    <doc><!-- doc 1 content --></doc>
  </add>
  <add>
    <doc><!-- doc 2 content --></doc>
  </add>
  <delete>
    <id>0002166313</id>
  </delete>
</update>
```

Using curl to Perform Updates

You can use the curl utility to perform any of the above commands, using its --data-binary option to append the XML message to the curl command, and generating a HTTP POST request. For example:

```
curl http://localhost:8983/solr/my_collection/update -H "Content-Type: text/xml" --data-binary '
<add>
  <doc>
    <field name="authors">Patrick Eagar</field>
    <field name="subject">Sports</field>
    <field name="dd">796.35</field>
    <field name="isbn">0002166313</field>
    <field name="yearpub">1982</field>
    <field name="publisher">Collins</field>
  </doc>
</add>'
```

For posting XML messages contained in a file, you can use the alternative form:

```
curl http://localhost:8983/solr/my_collection/update -H "Content-Type: text/xml" --data-binary
@myfile.xml
```

The approach above works well, but using the --data-binary option causes curl to load the whole myfile.xml into memory before posting it to server. This may be problematic when dealing with multi-gigabyte files. This alternative curl command performs equivalent operations but with minimal curl memory usage:

```
curl http://localhost:8983/solr/my_collection/update -H "Content-Type: text/xml" -T "myfile.xml"  
-X POST
```

Short requests can also be sent using a HTTP GET command, if enabled in [RequestDispatcher](#) in [SolrConfig](#) element, URL-encoding the request, as in the following. Note the escaping of "<" and ">":

```
curl http://localhost:8983/solr/my_collection/update?stream.body=%3Ccommit%3E&wt=xml
```

Responses from Solr take the form shown here:

```
<response>  
  <lst name="responseHeader">  
    <int name="status">0</int>  
    <int name="QTime">127</int>  
  </lst>  
</response>
```

The status field will be non-zero in case of failure.

Using XSLT to Transform XML Index Updates

The UpdateRequestHandler allows you to index any arbitrary XML using the `<tr>` parameter to apply an [XSL transformation](#). You must have an XSLT stylesheet in the `conf/xslt` directory of your [configset](#) that can transform the incoming data to the expected `<add><doc/></add>` format, and use the `tr` parameter to specify the name of that stylesheet.

Here is an example XSLT stylesheet:

```
<xsl:stylesheet version='1.0' xmlns:xsl='http://www.w3.org/1999/XSL/Transform'>
  <xsl:output media-type="text/xml" method="xml" indent="yes"/>
  <xsl:template match='/'>
    <add>
      <xsl:apply-templates select="response/result/doc"/>
    </add>
  </xsl:template>
  <!-- Ignore score (makes no sense to index) -->
  <xsl:template match="doc/*[@name='score']" priority="100"></xsl:template>
  <xsl:template match="doc">
    <xsl:variable name="pos" select="position()"/>
    <doc>
      <xsl:apply-templates>
        <xsl:with-param name="pos"><xsl:value-of select="$pos" /></xsl:with-param>
      </xsl:apply-templates>
    </doc>
  </xsl:template>
  <!-- Flatten arrays to duplicate field lines -->
  <xsl:template match="doc/arr" priority="100">
    <xsl:variable name="fn" select="@name"/>
    <xsl:for-each select="*>
      <xsl:element name="field">
        <xsl:attribute name="name"><xsl:value-of select="$fn" /></xsl:attribute>
        <xsl:value-of select=". />
      </xsl:element>
    </xsl:for-each>
  </xsl:template>
  <xsl:template match="doc/*">
    <xsl:variable name="fn" select="@name"/>
    <xsl:element name="field">
      <xsl:attribute name="name"><xsl:value-of select="$fn" /></xsl:attribute>
      <xsl:value-of select=". />
    </xsl:element>
  </xsl:template>
  <xsl:template match="*"/>
</xsl:stylesheet>
```

This stylesheet transforms Solr's XML search result format into Solr's Update XML syntax. One example usage would be to copy a Solr 1.3 index (which does not have CSV response writer) into a format which can be indexed into another Solr file (provided that all fields are stored):

```
http://localhost:8983/solr/my_collection/select?q=*&wt=xslt&tr=updateXml.xsl&rows=1000
```

You can also use the stylesheet in XsltUpdateRequestHandler to transform an index when updating:

```
curl "http://localhost:8983/solr/my_collection/update?commit=true&tr=updateXml.xsl" -H "Content-Type: text/xml" --data-binary @myexporteddata.xml
```

JSON Formatted Index Updates

Solr can accept JSON that conforms to a defined structure, or can accept arbitrary JSON-formatted documents. If sending arbitrarily formatted JSON, there are some additional parameters that need to be sent with the update request, described below in the section [Transforming and Indexing Custom JSON](#).

Solr-Style JSON

JSON formatted update requests may be sent to Solr's /update handler using Content-Type: application/json or Content-Type: text/json.

JSON formatted updates can take 3 basic forms, described in depth below:

- [A single document to add](#), expressed as a top level JSON Object. To differentiate this from a set of commands, the json.command=false request parameter is required.
- [A list of documents to add](#), expressed as a top level JSON Array containing a JSON Object per document.
- [A sequence of update commands](#), expressed as a top level JSON Object (aka: Map).

Adding a Single JSON Document

The simplest way to add Documents via JSON is to send each document individually as a JSON Object, using the /update/json/docs path:

```
curl -X POST -H 'Content-Type: application/json'
'http://localhost:8983/solr/my_collection/update/json/docs' --data-binary '
{
  "id": "1",
  "title": "Doc 1"
}'
```

Adding Multiple JSON Documents

Adding multiple documents at one time via JSON can be done via a JSON Array of JSON Objects, where each object represents a document:

```
curl -X POST -H 'Content-Type: application/json'
'http://localhost:8983/solr/my_collection/update' --data-binary '
[
  {
    "id": "1",
    "title": "Doc 1"
  },
  {
    "id": "2",
    "title": "Doc 2"
  }
]'
```

A sample JSON file is provided at example/exampledocs/books.json and contains an array of objects that you can add to the Solr techproducts example:

```
curl 'http://localhost:8983/solr/techproducts/update?commit=true' --data-binary  
@example/exampledocs/books.json -H 'Content-type:application/json'
```

Sending JSON Update Commands

In general, the JSON update syntax supports all of the update commands that the XML update handler supports, through a straightforward mapping. Multiple commands, adding and deleting documents, may be contained in one message:

```
curl -X POST -H 'Content-Type: application/json'  
'http://localhost:8983/solr/my_collection/update' --data-binary '  
{  
  "add": {  
    "doc": {  
      "id": "DOC1",  
      "my_field": 2.3,  
      "my_multivalued_field": [ "aaa", "bbb" ] ①  
    }  
  },  
  "add": {  
    "commitWithin": 5000, ②  
    "overwrite": false, ③  
    "doc": {  
      "f1": "v1", ④  
      "f1": "v2"  
    }  
  },  
  
  "commit": {},  
  "optimize": { "waitSearcher":false },  
  
  "delete": { "id":"ID" }, ⑤  
  "delete": { "query":"QUERY" } ⑥  
}'
```

- ① Can use an array for a multi-valued field
- ② Commit this document within 5 seconds
- ③ Don't check for existing documents with the same uniqueKey
- ④ Can use repeated keys for a multi-valued field
- ⑤ Delete by ID (uniqueKey field)
- ⑥ Delete by Query

As with other update handlers, parameters such as commit, commitWithin, optimize, and overwrite may be specified in the URL instead of in the body of the message.

The JSON update format allows for a simple delete-by-id. The value of a delete can be an array which contains a list of zero or more specific document id's (not a range) to be deleted. For example, a single document:

```
{ "delete": "myid" }
```

Or a list of document IDs:

```
{ "delete": [ "id1", "id2" ] }
```

Note: Delete-by-id doesn't work for child/nested docs.

You can also specify `_version_` with each "delete":

```
{
  "delete": "id": 50,
  "_version_": 12345
}
```

You can specify the version of deletes in the body of the update request as well.

JSON Update Convenience Paths

In addition to the /update handler, there are a few additional JSON specific request handler paths available by default in Solr, that implicitly override the behavior of some request parameters:

Path	Default Parameters
/update/json	stream.contentType=application/json
/update/json/docs	stream.contentType=application/json json.command=false

The /update/json path may be useful for clients sending in JSON formatted update commands from applications where setting the Content-Type proves difficult, while the /update/json/docs path can be particularly convenient for clients that always want to send in documents – either individually or as a list – without needing to worry about the full JSON command syntax.

Custom JSON Documents

Solr can support custom JSON. This is covered in the section [Transforming and Indexing Custom JSON](#).

CSV Formatted Index Updates

CSV formatted update requests may be sent to Solr's /update handler using Content-Type: application/csv or Content-Type: text/csv.

A sample CSV file is provided at example/exemplodedocs/books.csv that you can use to add some documents to the Solr techproducts example:

```
curl 'http://localhost:8983/solr/my_collection/update?commit=true' --data-binary  
@example/exemplodedocs/books.csv -H 'Content-type:application/csv'
```

CSV Update Parameters

The CSV handler allows the specification of many parameters in the URL in the form: f.parameter.optional_fieldname=value.

The table below describes the parameters for the update handler.

separator

Character used as field separator; default is ",". This parameter is global; for per-field usage, see the split parameter.

Example: separator=%09

trim

If true, remove leading and trailing whitespace from values. The default is false. This parameter can be either global or per-field.

Examples: f.isbn.trim=true or trim=false

header

Set to true if first line of input contains field names. These will be used if the fieldnames parameter is absent. This parameter is global.

fieldnames

Comma-separated list of field names to use when adding documents. This parameter is global.

Example: fieldnames=isbn,price,title

literal.field_name

A literal value for a specified field name. This parameter is global.

Example: literal.color=red

skip

Comma separated list of field names to skip. This parameter is global.

Example: skip=uninteresting,shoesize

skipLines

Number of lines to discard in the input stream before the CSV data starts, including the header, if present. Default=0. This parameter is global.

Example: skipLines=5

encapsulator

The character optionally used to surround values to preserve characters such as the CSV separator or

whitespace. This standard CSV format handles the encapsulator itself appearing in an encapsulated value by doubling the encapsulator.

This parameter is global; for per-field usage, see split.

Example: `encapsulator=""`

escape

The character used for escaping CSV separators or other reserved characters. If an escape is specified, the encapsulator is not used unless also explicitly specified since most formats use either encapsulation or escaping, not both. `|g|`

Example: `escape=\`

keepEmpty

Keep and index zero length (empty) fields. The default is false. This parameter can be global or per-field.

Example: `f.price.keepEmpty=true`

map

Map one value to another. Format is `value:replacement` (which can be empty). This parameter can be global or per-field.

Example: `map=left:right` or `f.subject.map=history:bunk`

split

If true, split a field into multiple values by a separate parser. This parameter is used on a per-field basis.

overwrite

If true (the default), check for and overwrite duplicate documents, based on the uniqueKey field declared in the Solr schema. If you know the documents you are indexing do not contain any duplicates then you may see a considerable speed up setting this to false.

This parameter is global.

commit

Issues a commit after the data has been ingested. This parameter is global.

commitWithin

Add the document within the specified number of milliseconds. This parameter is global.

Example: `commitWithin=10000`

rowid

Map the rowid (line number) to a field specified by the value of the parameter, for instance if your CSV doesn't have a unique key and you want to use the row id as such. This parameter is global.

Example: `rowid=id`

rowidOffset

Add the given offset (as an integer) to the rowid before adding it to the document. Default is 0. This parameter is global.

Example: rowidOffset=10

Indexing Tab-Delimited files

The same feature used to index CSV documents can also be easily used to index tab-delimited files (TSV files) and even handle backslash escaping rather than CSV encapsulation.

For example, one can dump a MySQL table to a tab delimited file with:

```
SELECT * INTO OUTFILE '/tmp/result.txt' FROM mytable;
```

This file could then be imported into Solr by setting the separator to tab (%09) and the escape to backslash (%5c).

```
curl 'http://localhost:8983/solr/my_collection/update/csv?commit=true&separator=%09&escape=%5c'  
--data-binary @/tmp/result.txt
```

CSV Update Convenience Paths

In addition to the /update handler, there is an additional CSV specific request handler path available by default in Solr, that implicitly override the behavior of some request parameters:

Path	Default Parameters
/update/csv	stream.contentType=application/csv

The /update/csv path may be useful for clients sending in CSV formatted update commands from applications where setting the Content-Type proves difficult.

Transforming and Indexing Custom JSON

If you have JSON documents that you would like to index without transforming them into Solr's structure, you can add them to Solr by including some parameters with the update request.

These parameters provide information on how to split a single JSON file into multiple Solr documents and how to map fields to Solr's schema. One or more valid JSON documents can be sent to the /update/json/docs path with the configuration params.

Mapping Parameters

These parameters allow you to define how a JSON file should be read for multiple Solr documents.

split

Defines the path at which to split the input JSON into multiple Solr documents and is required if you have multiple documents in a single JSON file. If the entire JSON makes a single Solr document, the path must be "/".

It is possible to pass multiple split paths by separating them with a pipe (|), for example:

split=/|/foo|/foo/bar. If one path is a child of another, they automatically become a child document.

f

Provides multivalued mapping to map document field names to Solr field names. The format of the parameter is target-field-name:json-path, as in f=first:/first. The json-path is required. The target-field-name is the Solr document field name, and is optional. If not specified, it is automatically derived from the input JSON. The default target field name is the fully qualified name of the field.

Wildcards can be used here, see [Using Wildcards for Field Names](#) below for more information.

mapUniqueKeyOnly

(boolean) This parameter is particularly convenient when the fields in the input JSON are not available in the schema and [schemaless mode](#) is not enabled. This will index all the fields into the default search field (using the df parameter, below) and only the uniqueKey field is mapped to the corresponding field in the schema. If the input JSON does not have a value for the uniqueKey field then a UUID is generated for the same.

df

If the mapUniqueKeyOnly flag is used, the update handler needs a field where the data should be indexed to. This is the same field that other handlers use as a default search field.

srcField

This is the name of the field to which the JSON source will be stored into. This can only be used if split=/ (i.e., you want your JSON input file to be indexed as a single Solr document). Note that atomic updates will cause the field to be out-of-sync with the document.

echo

This is for debugging purpose only. Set it to true if you want the docs to be returned as a response. Nothing will be indexed.

For example, if we have a JSON file that includes two documents, we could define an update request like this:

V1 API

```
curl 'http://localhost:8983/solr/techproducts/update/json/docs'\
'?split=/exams'\
'&f=first:/first'\
'&f=last:/last'\
'&f=grade:/grade'\
'&f=subject:/exams/subject'\
'&f=test:/exams/test'\
'&f=marks:/exams/marks'\
-H 'Content-type:application/json' -d \
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

V2 API Standalone Solr

```
curl 'http://localhost:8983/api/cores/techproducts/update/json/docs' \
'?split=/exams' \
'&f=first:/first' \
'&f=last:/last' \
'&f=grade:/grade' \
'&f=subject:/exams/subject' \
'&f=test:/exams/test' \
'&f=marks:/exams/marks' \
-H 'Content-type:application/json' -d ' \
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

V2 API SolrCloud

```
curl 'http://localhost:8983/api/collections/techproducts/update/json/docs'\
'?split=/exams'\n'&f=first:/first'\n'&f=last:/last'\n'&f=grade:/grade'\n'&f=subject:/exams/subject'\n'&f=test:/exams/test'\n'&f=marks:/exams/marks'\n-H 'Content-type:application/json' -d '\n{\n    "first": "John",\n    "last": "Doe",\n    "grade": 8,\n    "exams": [\n        {\n            "subject": "Maths",\n            "test": "term1",\n            "marks": 90},\n        {\n            "subject": "Biology",\n            "test": "term1",\n            "marks": 86}\n    ]\n}'
```

With this request, we have defined that "exams" contains multiple documents. In addition, we have mapped several fields from the input document to Solr fields.

When the update request is complete, the following two documents will be added to the index:

```
{\n    "first": "John",\n    "last": "Doe",\n    "marks": 90,\n    "test": "term1",\n    "subject": "Maths",\n    "grade": 8\n}\n{\n    "first": "John",\n    "last": "Doe",\n    "marks": 86,\n    "test": "term1",\n    "subject": "Biology",\n    "grade": 8\n}
```

In the prior example, all of the fields we wanted to use in Solr had the same names as they did in the input JSON. When that is the case, we can simplify the request by only specifying the json-path portion of the f parameter, as in this example:

V1 API

```
curl 'http://localhost:8983/solr/techproducts/update/json/docs'\
'?split=/exams'\
'&f=/first'\
'&f=/last'\
'&f=/grade'\
'&f=/exams/subject'\
'&f=/exams/test'\
'&f=/exams/marks'\
-H 'Content-type:application/json' -d \
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

V2 API Standalone Solr

```
curl 'http://localhost:8983/api/cores/techproducts/update/json/docs' \
'?split=/exams' \
'&f=/first' \
'&f=/last' \
'&f=/grade' \
'&f=/exams/subject' \
'&f=/exams/test' \
'&f=/exams/marks' \
-H 'Content-type:application/json' -d ' \
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

V2 API SolrCloud

```
curl 'http://localhost:8983/api/collections/techproducts/update/json/docs'\
'?split=/exams'\
'&f=/first'\
'&f=/last'\
'&f=/grade'\
'&f=/exams/subject'\
'&f=/exams/test'\
'&f=/exams/marks'\
-H 'Content-type:application/json' -d \
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test" : "term1",
      "marks" : 90},
    {
      "subject": "Biology",
      "test" : "term1",
      "marks" : 86}
  ]
}'
```

In this example, we simply named the field paths (such as /exams/test). Solr will automatically attempt to add the content of the field from the JSON input to the index in a field with the same name.



Documents will be rejected during indexing if the fields do not exist in the schema before indexing. So, if you are NOT using schemaless mode, you must pre-create all fields. If you are working in [Schemaless Mode](#), however, fields that don't exist will be created on the fly with Solr's best guess for the field type.

Reusing Parameters in Multiple Requests

You can store and re-use parameters with Solr's [Request Parameters API](#).

Say we wanted to define parameters to split documents at the exams field, and map several other fields. We could make an API request such as:

V1 API

```
curl http://localhost:8983/solr/techproducts/config/params -H 'Content-type:application/json' -d '{
  "set": {
    "my_params": {
      "split": "/exams",
      "f": [
        "first:/first", "last:/last", "grade:/grade", "subject:/exams/subject", "test:/exams/test"
      ]}}'
```

V2 API Standalone Solr

```
curl http://localhost:8983/api/cores/techproducts/config/params -H 'Content-type:application/json' -d '{
  "set": {
    "my_params": {
      "split": "/exams",
      "f": [
        "first:/first", "last:/last", "grade:/grade", "subject:/exams/subject", "test:/exams/test"
      ]}}'
```

V2 API SolrCloud

```
curl http://localhost:8983/api/collections/techproducts/config/params -H 'Content-type:application/json' -d '{
  "set": {
    "my_params": {
      "split": "/exams",
      "f": [
        "first:/first", "last:/last", "grade:/grade", "subject:/exams/subject", "test:/exams/test"
      ]}}'
```

When we send the documents, we'd use the `useParams` parameter with the name of the parameter set we defined:

V1 API

```
curl 'http://localhost:8983/solr/techproducts/update/json/docs?useParams=my_params' -H  
'Content-type:application/json' -d '{  
    "first": "John",  
    "last": "Doe",  
    "grade": 8,  
    "exams": [ {  
        "subject": "Maths",  
        "test": "term1",  
        "marks": 90  
    },  
    {  
        "subject": "Biology",  
        "test": "term1",  
        "marks": 86  
    }  
]  
'
```

V2 API Standalone Solr

```
curl 'http://localhost:8983/api/cores/techproducts/update/json?useParams=my_params' -H  
'Content-type:application/json' -d '{  
    "first": "John",  
    "last": "Doe",  
    "grade": 8,  
    "exams": [ {  
        "subject": "Maths",  
        "test": "term1",  
        "marks": 90  
    },  
    {  
        "subject": "Biology",  
        "test": "term1",  
        "marks": 86  
    }  
]  
'
```

V2 API SolrCloud

```
curl 'http://localhost:8983/api/collections/techproducts/update/json?useParams=my_params' -H  
'Content-type:application/json' -d '{  
    "first": "John",  
    "last": "Doe",  
    "grade": 8,  
    "exams": [ {  
        "subject": "Maths",  
        "test": "term1",  
        "marks": 90  
    },  
    {  
        "subject": "Biology",  
        "test": "term1",  
        "marks": 86  
    }  
]  
'
```

Using Wildcards for Field Names

Instead of specifying all the field names explicitly, it is possible to specify wildcards to map fields automatically.

There are two restrictions: wildcards can only be used at the end of the json-path, and the split path cannot use wildcards.

A single asterisk `*` maps only to direct children, and a double asterisk `**` maps recursively to all descendants. The following are example wildcard path mappings:

- `f=$FQN:/**:` maps all fields to the fully qualified name (`$FQN`) of the JSON field. The fully qualified name is obtained by concatenating all the keys in the hierarchy with a period (.) as a delimiter. This is the default behavior if no f path mappings are specified.
- `f=/docs/*:` maps all the fields under docs and in the name as given in JSON
- `f=/docs/**:` maps all the fields under docs and its children in the name as given in JSON
- `f=searchField:/docs/*:` maps all fields under /docs to a single field called 'searchField'
- `f=searchField:/docs/**:` maps all fields under /docs and its children to searchField

With wildcards we can further simplify our previous example as follows:

V1 API

```
curl 'http://localhost:8983/solr/techproducts/update/json/docs'\
'?split=/exams'\
'&f=/**'\n
-H 'Content-type:application/json' -d '
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

V2 API Standalone Solr

```
curl 'http://localhost:8983/api/cores/techproducts/update/json'\
'?split=/exams'\
'&f=/**'\n
-H 'Content-type:application/json' -d '
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

V2 API SolrCloud

```
curl 'http://localhost:8983/api/collections/techproducts/update/json' \
'?split=/exams' \
'&f=/**' \
-H 'Content-type:application/json' -d '
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

Because we want the fields to be indexed with the field names as they are found in the JSON input, the double wildcard in `f=/**` will map all fields and their descendants to the same fields in Solr.

It is also possible to send all the values to a single field and do a full text search on that. This is a good option to blindly index and query JSON documents without worrying about fields and schema.

V1 API

```
curl 'http://localhost:8983/solr/techproducts/update/json/docs'\
'?split=/'\>
'&f=txt:/**'\>
-H 'Content-type:application/json' -d '
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

V2 API Standalone Solr

```
curl 'http://localhost:8983/api/cores/techproducts/update/json'\
'?split=/'\>
'&f=txt:/**'\>
-H 'Content-type:application/json' -d '
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

V2 API SolrCloud

```
curl 'http://localhost:8983/api/collections/techproducts/update/json' \
'?split=/\' \
'&f=txt:/**' \
-H 'Content-type:application/json' -d '
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

In the above example, we've said all of the fields should be added to a field in Solr named 'txt'. This will add multiple fields to a single field, so whatever field you choose should be multi-valued.

The default behavior is to use the fully qualified name (FQN) of the node. So, if we don't define any field mappings, like this:

V1 API

```
curl 'http://localhost:8983/solr/techproducts/update/json/docs?split=/exams' \
-H 'Content-type:application/json' -d '
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

V2 API Standalone Solr

```
curl 'http://localhost:8983/api/cores/techproducts/update/json?split=/exams'\
-H 'Content-type:application/json' -d \
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

V2 API SolrCloud

```
curl 'http://localhost:8983/api/collections/techproducts/update/json?split=/exams'\
-H 'Content-type:application/json' -d \
{
  "first": "John",
  "last": "Doe",
  "grade": 8,
  "exams": [
    {
      "subject": "Maths",
      "test": "term1",
      "marks": 90},
    {
      "subject": "Biology",
      "test": "term1",
      "marks": 86}
  ]
}'
```

The indexed documents would be added to the index with fields that look like this:

```
{  
  "first": "John",  
  "last": "Doe",  
  "grade": 8,  
  "exams.subject": "Maths",  
  "exams.test": "term1",  
  "exams.marks": 90},  
  
{  
  "first": "John",  
  "last": "Doe",  
  "grade": 8,  
  "exams.subject": "Biology",  
  "exams.test": "term1",  
  "exams.marks": 86}
```

Multiple Documents in a Single Payload

This functionality supports documents in the [JSON Lines](#) format (.jsonl), which specifies one document per line.

For example:

V1 API

```
curl 'http://localhost:8983/solr/techproducts/update/json/docs' -H 'Content-Type:application/json' -d '  
{ "first": "Steve", "last": "Jobs", "grade": 1, "subject": "Social Science", "test": "term1", "marks": 90}  
{ "first": "Steve", "last": "Woz", "grade": 1, "subject": "Political Science", "test": "term1", "marks": 86}'
```

V2 API Standalone Solr

```
curl 'http://localhost:8983/api/collections/techproducts/update/json' -H 'Content-Type:application/json' -d '  
{ "first": "Steve", "last": "Jobs", "grade": 1, "subject": "Social Science", "test": "term1", "marks": 90}  
{ "first": "Steve", "last": "Woz", "grade": 1, "subject": "Political Science", "test": "term1", "marks": 86}'
```

V2 API SolrCloud

```
curl 'http://localhost:8983/api/collections/techproducts/update/json' -H 'Content-type:application/json' -d '[{"first":"Steve", "last":"Jobs", "grade":1, "subject":"Social Science", "test":"term1", "marks":90}, {"first":"Steve", "last":"Woz", "grade":1, "subject":"Political Science", "test":"term1", "marks":86}]'
```

Or even an array of documents, as in this example:

V1 API

```
curl 'http://localhost:8983/solr/techproducts/update/json/docs' -H 'Content-type:application/json' -d '[{"first":"Steve", "last":"Jobs", "grade":1, "subject":"Computer Science", "test":"term1", "marks":90}, {"first":"Steve", "last":"Woz", "grade":1, "subject":"Calculus", "test":"term1", "marks":86}]'
```

V2 API Standalone Solr

```
curl 'http://localhost:8983/api/cores/techproducts/update/json' -H 'Content-type:application/json' -d '[{"first":"Steve", "last":"Jobs", "grade":1, "subject":"Computer Science", "test":"term1", "marks":90}, {"first":"Steve", "last":"Woz", "grade":1, "subject":"Calculus", "test":"term1", "marks":86}]'
```

V2 API SolrCloud

```
curl 'http://localhost:8983/api/collections/techproducts/update/json' -H 'Content-type:application/json' -d '[{"first":"Steve", "last":"Jobs", "grade":1, "subject":"Computer Science", "test":"term1", "marks":90}, {"first":"Steve", "last":"Woz", "grade":1, "subject":"Calculus", "test":"term1", "marks":86}]'
```

Tips for Custom JSON Indexing

1. Schemaless mode: This handles field creation automatically. The field guessing may not be exactly as you expect, but it works. The best thing to do is to setup a local server in schemaless mode, index a few

sample docs and create those fields in your real setup with proper field types before indexing

2. Pre-created Schema: Post your docs to the /update/json/docs endpoint with echo=true. This gives you the list of field names you need to create. Create the fields before you actually index
3. No schema, only full-text search: All you need to do is to do full-text search on your JSON. Set the configuration as given in the Setting JSON Defaults section.

Setting JSON Defaults

It is possible to send any JSON to the /update/json/docs endpoint and the default configuration of the component is as follows:

```
<initParams path="/update/json/docs">
  <lst name="defaults">
    <!-- this ensures that the entire JSON doc will be stored verbatim into one field -->
    <str name="srcField">_src_</str>
    <!-- This means a the uniqueKeyField will be extracted from the fields and
        all fields go into the 'df' field. In this config df is already configured to be 'text'
        -->
    <str name="mapUniqueKeyOnly">true</str>
    <!-- The default search field where all the values are indexed to -->
    <str name="df">text</str>
  </lst>
</initParams>
```

So, if no parameters are passed, the entire JSON file would get indexed to the _src_ field and all the values in the input JSON would go to a field named text. If there is a value for the uniqueKey it is stored and if no value could be obtained from the input JSON, a UUID is created and used as the uniqueKey field value.

Alternately, use the Request Parameters feature to set these parameters, as shown earlier in the section [Reusing Parameters in Multiple Requests](#).

V1 API

```
curl http://localhost:8983/solr/techproducts/config/params -H 'Content-Type:application/json' -d '{
  "set": {
    "full_txt": {
      "srcField": "_src_",
      "mapUniqueKeyOnly" : true,
      "df": "text"
    }}}'
```

V2 API Standalone Solr

```
curl http://localhost:8983/api/cores/techproducts/config/params -H 'Content-type:application/json' -d '{  
"set": {  
"full_txt": {  
"srcField": "_src_",  
"mapUniqueKeyOnly" : true,  
"df": "text"  
}}}'
```

V2 API SolrCloud

```
curl http://localhost:8983/api/collections/techproducts/config/params -H 'Content-type:application/json' -d '{  
"set": {  
"full_txt": {  
"srcField": "_src_",  
"mapUniqueKeyOnly" : true,  
"df": "text"  
}}}'
```

To use these parameters, send the parameter `useParams=full_txt` with each request.

Indexing Nested Child Documents

Solr supports indexing nested documents, described here, and ways to [search and retrieve](#) them very efficiently. By way of example, nested documents in Solr can be used to bind a blog post (parent document) with comments (child documents) — or products as parent documents and sizes, colors, or other variations as child documents.

The parent with all children is referred to as a nested document or "block" and it explains some of the nomenclature of related features. At query time, the [Block Join Query Parsers](#) can search these relationships, and the [\[child\] Document Transformer](#) can attach child documents to the result documents. In terms of performance, indexing the relationships between documents usually yields much faster queries than an equivalent "query time join", since the relationships are already stored in the index and do not need to be computed. However, nested documents are less flexible than query time joins as it imposes rules that some applications may not be able to accept. Nested documents may be indexed via either the XML or JSON data syntax, and is also supported by [SolrJ](#) with javabin.

Limitation



With the exception of in-place updates, the whole block must be updated or deleted together, not separately. For some applications this may result in tons of extra indexing and thus may be a deal-breaker.

Schema Configuration

- The schema must include an indexed field `_root_`. Solr automatically populates this with the value of the top/parent ID.
`<field name="_root_" type="string" indexed="true" stored="false" docValues="false" />`
 - `_root_` must be set either as stored (`stored="true"`) or doc values (`docValues="true"`) to enable [atomic updates of nested documents](#).
- `_nest_path_` is populated by Solr automatically with the path of the document in the hierarchy for non-root documents. This field is optional.
`<fieldType name="_nest_path_" class="solr.NestPathField" /> <field name="_nest_path_" type="nest_path" />`
- `_nest_parent_` is populated by Solr automatically to store the ID of each document's parent document (if there is one). This field is optional.
`<field name="_nest_parent_" type="string" indexed="true" stored="true"/>`
- Nested documents are very much documents in their own right even if certain nested documents hold different information from the parent. Therefore:
 - a field can only be configured one way no matter what sort of document uses it
 - it may be infeasible to use required
 - even child documents need a unique ID
- Even though child documents are provided as field values syntactically and with SolrJ, it's a matter of syntax and it isn't an actual field in the schema. Consequently, the field need not be defined in the schema and probably shouldn't be as it would be confusing. There is no child document field type, at least not yet.

Rudimentary Root-only Schemas

These schemas do not contain any other nested related fields apart from `_root_`. Many schemas in existence are this way simply because default configsets are this way, even if the application isn't using nested documents. If an application uses nested documents with such a schema, keep in mind that some related features aren't as effective since there is less information. Mainly the [\[child\]](#) transformer returns matching children in a flat list (not nested) and it's attached to the parent using the special field name `_childDocuments_`.

With such a schema, typically you should have a field that differentiates a root doc from any nested children. However this isn't strictly necessary; so long as it's possible to write a query that can select only root documents somehow. Such a query is needed for the [block join query parsers](#) and [\[child\]](#) doc transformer to function.

XML Examples

Here are two documents and their child documents. It illustrates two styles of adding child documents: the first is associated via a field "comment" (preferred), and the second is done in the classic way now referred to as an "anonymous" or "unlabelled" child document. This field label relationship is available to the URP chain in Solr but is ultimately discarded unless the special fields are defined.

```
<add>
  <doc>
    <field name="ID">1</field>
    <field name="title">Solr adds block join support</field>
    <field name="content_type">parentDocument</field>
    <field name="content">
      <doc>
        <field name="ID">2</field>
        <field name="comments">SolrCloud supports it too!</field>
      </doc>
    </field>
  </doc>
  <doc>
    <field name="ID">3</field>
    <field name="title">New Lucene and Solr release is out</field>
    <field name="content_type">parentDocument</field>
    <doc>
      <field name="ID">4</field>
      <field name="comments">Lots of new features</field>
    </doc>
  </doc>
</add>
```

In this example, we have indexed the parent documents with the field `content_type`, which has the value "parentDocument". We could have also used a boolean field, such as `isParent`, with a value of "true", or any other similar approach.

JSON Examples

This example is equivalent to the XML example above. Again, the field labelled relationship is preferred. The labelled relationship here is one child document but could have been wrapped in array brackets. For the anonymous relationship, note the special `_childDocuments_` key whose contents must be an array of child documents.

```
[  
  {  
    "ID": "1",  
    "title": "Solr adds block join support",  
    "content_type": "parentDocument",  
    "comments": [{  
      "ID": "2",  
      "content": "SolrCloud supports it too!"  
    },  
    {  
      "ID": "3",  
      "content": "New filter syntax"  
    }  
  ],  
  {  
    "ID": "4",  
    "title": "New Lucene and Solr release is out",  
    "content_type": "parentDocument",  
    "_childDocuments_": [  
      {  
        "ID": "5",  
        "comments": "Lots of new features"  
      }  
    ]  
  }  
]
```

Root-Only Mode



In Root-only schemas, these two documents will result in the same docs being indexed (Root-only schemas do not honor nested relationships). When queried, child docs will be appended to the `childDocuments` field/key.

Important: Maintaining Integrity with Updates and Deletes

Nested documents (children and all) can simply be replaced by adding a new document with more or fewer documents as an application desires. This aspect isn't different than updating any normal document except that Solr takes care to ensure that all related child documents of the existing version get deleted.

Do **not** add a root document that has the same ID of a child document. *This will violate integrity assumptions that Solr expects.*

To delete a nested document, you can delete it by the ID of the root document. If you try to use an ID of a child document, nothing will happen since only root document IDs are considered. If you use Solr's delete-by-query APIs, you **have to be careful** to ensure that no children remain of any documents that are being deleted. *Doing otherwise will violate integrity assumptions that Solr expects.*

Uploading Data with Solr Cell using Apache Tika

If the documents you need to index are in a binary format, such as Word, Excel, PDFs, etc., Solr includes a request handler which uses [Apache Tika](#) to extract text for indexing to Solr.

Solr uses code from the Tika project to provide a framework for incorporating many different file-format parsers such as [Apache PDFBox](#) and [Apache POI](#) into Solr itself.

Working with this framework, Solr's ExtractingRequestHandler uses Tika internally to support uploading binary files for data extraction and indexing. Downloading Tika is not required to use Solr Cell.

When this framework was under development, it was called the Solr *Content Extraction Library*, or *CEL*; from that abbreviation came this framework's name: Solr Cell. The names Solr Cell and ExtractingRequestHandler are used interchangeably for this feature.

Key Solr Cell Concepts

When using the Solr Cell framework, it is helpful to keep the following in mind:

- Tika will automatically attempt to determine the input document type (e.g., Word, PDF, HTML) and extract the content appropriately. If you like, you can explicitly specify a MIME type for Tika with the stream.type parameter. See <http://tika.apache.org/1.19.1/formats.html> for the file types supported.
- Briefly, Tika internally works by synthesizing an XHTML document from the core content of the parsed document which is passed to a configured SAX ContentHandler provided by Solr Cell. Solr responds to Tika's SAX events to create one or more text fields from the content. Tika exposes document metadata as well (apart from the XHTML).
- Tika produces metadata such as Title, Subject, and Author according to specifications such as the DublinCore. The metadata available is highly dependent on the file types and what they in turn contain. Some of the general metadata created is described in the section [Metadata Created by Tika](#) below. Solr Cell supplies some metadata of its own too.
- Solr Cell concatenates text from the internal XHTML into a content field. You can configure which elements should be included/ignored, and which should map to another field.
- Solr Cell maps each piece of metadata onto a field. By default it maps to the same name but several parameters control how this is done.
- When Solr Cell finishes creating the internal SolrInputDocument, the rest of the Lucene/Solr indexing stack takes over. The next step after any update handler is the [Update Request Processor](#) chain.

Solr Cell is a contrib, which means it's not automatically included with Solr but must be configured. The example configsets have Solr Cell configured, but if you are not using those, you will want to pay attention to the section [Configuring the ExtractingRequestHandler in solrconfig.xml](#) below.

Solr Cell Performance Implications

Rich document formats are frequently not well documented, and even in cases where there is documentation for the format, not everyone who creates documents will follow the specifications faithfully.

This creates a situation where Tika may encounter something that it is simply not able to handle gracefully,

despite taking great pains to support as many formats as possible. PDF files are particularly problematic, mostly due to the PDF format itself.

In case of a failure processing any file, the ExtractingRequestHandler does not have a secondary mechanism to try to extract some text from the file; it will throw an exception and fail.

If any exceptions cause the ExtractingRequestHandler and/or Tika to crash, Solr as a whole will also crash because the request handler is running in the same JVM that Solr uses for other operations.

Indexing can also consume all available Solr resources, particularly with large PDFs, presentations, or other files that have a lot of rich media embedded in them.

For these reasons, Solr Cell is not recommended for use in a production system.

It is a best practice to use Solr Cell as a proof-of-concept tool during development and then run Tika as an external process that sends the extracted documents to Solr (via [SolrJ](#)) for indexing. This way, any extraction failures that occur are isolated from Solr itself and can be handled gracefully.

For a few examples of how this could be done, see this blog post by Erick Erickson, [Indexing with SolrJ](#).

Trying out Solr Cell

You can try out the Tika framework using the schemaless example included in Solr.

This command will simply start Solr and create a core/collection named "gettingstarted" with the _default configset.

```
bin/solr -e schemaless
```

Once Solr is started, you can use curl to send a sample PDF included with Solr via HTTP POST:

```
curl  
'http://localhost:8983/solr/gettingstarted/update/extract?literal.id=doc1&uprefix=ignored_&commit=true' -F "myfile=@example/exampledocs/solr-word.pdf"
```

The URL above calls the ExtractingRequestHandler, uploads the file solr-word.pdf, and assigns it the unique ID doc1. Here's a closer look at the components of this command:

- The literal.id=doc1 parameter provides a unique ID for the document being indexed. Without this, the ID would be set to the absolute path to the file.

There are alternatives to this, such as mapping a metadata field to the ID, generating a new UUID, or generating an ID from a signature (hash) of the content.

- The commit=true parameter causes Solr to perform a commit after indexing the document, making it immediately searchable. For optimum performance when loading many documents, don't call the commit command until you are done.
- The -F flag instructs curl to POST data using the Content-Type multipart/form-data and supports the uploading of binary files. The @ symbol instructs curl to upload the attached file.

- The argument myfile=@example/exampledocs/solr-word.pdf uploads the sample file. Note this includes the path, so if you upload a different file, always be sure to include either the relative or absolute path to the file.

You can also use bin/post to do the same thing:

```
bin/post -c gettingstarted example/exampledocs/solr-word.pdf -params "literal.id=doc1"
```

Now you can execute a query and find that document with a request like

<http://localhost:8983/solr/gettingstarted/select?q=pdf>. The document will look something like this:

```
http://localhost:8983/solr/gettingstarted/select?omitHeader=true&q=pdf

{
  "response": { "numFound": 1, "start": 0, "maxScore": 0.5678674, "docs": [
    {
      "id": "doc1",
      "date": ["2008-11-13T13:35:51Z"],
      "pdf_docinfo_custom_aapl_keywords": ["solr, word, pdf"],
      "pdf_pdfversion": [1.3],
      "pdf_docinfo_title": ["solr-word"],
      "xmp_creatortool": ["Microsoft Word"],
      "stream_content_type": ["application/pdf"],
      "access_permission_can_print_degraded": [true],
      "subject": ["solr word"],
      "dc_format": ["application/pdf; version=1.3"],
      "pdf_docinfo_creator_tool": ["Microsoft Word"],
      "access_permission_fill_in_form": [true],
      "pdf_encrypted": [false],
      "dc_title": ["solr-word"],
      "modified": ["2008-11-13T13:35:51Z"],
      "cp_subject": ["solr word"],
      "pdf_docinfo_subject": ["solr word"],
      "pdf_docinfo_creator": ["Grant Ingersoll"],
      "meta_author": ["Grant Ingersoll"],
      "meta_creation_date": ["2008-11-13T13:35:51Z"],
      "created": ["Thu Nov 13 13:35:51 UTC 2008"],
      "access_permission_extract_for_accessibility": [true],
      "creation_date": ["2008-11-13T13:35:51Z"],
      "resourcename": ["/Applications/Solr/solr-7.5.0/example/exampledocs/solr-word.pdf"],
      "author": ["Grant Ingersoll"],
      "producer": ["Mac OS X 10.5.5 Quartz PDFContext"],
      "pdf_docinfo_producer": ["Mac OS X 10.5.5 Quartz PDFContext"],
      "keywords": ["solr, word, pdf"],
      "access_permission_modify_annotations": [true],
      "aapl_keywords": ["solr, word, pdf"],
      "dc_creator": ["Grant Ingersoll"],
      "dcterms_created": ["2008-11-13T13:35:51Z"]
    }
  ]
}
```

You may notice there are many metadata fields associated with this document. Solr's configuration is by default in "schemaless" (data driven) mode, and thus all metadata fields extracted get their own field.

You might instead want to ignore them generally except for a few you specify. To do that, use the uprefix parameter to map unknown (to the schema) metadata field names to a schema field name that is effectively ignored. The dynamic field ignored_* is good for this purpose.

For the fields you do want to map, explicitly set them using fmap.IN=OUT and/or ensure the field is defined in the schema. Here's an example:

```
bin/post -c gettingstarted example/exemplledocs/solr-word.pdf -params  
"literal.id=doc1&uprefix=ignored_&fmap.last_modified=last_modified_dt"
```

The above example won't work as expected if you run it after you've already indexed the document one or more times.



Previously we added the document without these parameters so all fields were added to the index at that time. The uprefix parameter only applies to fields that are *undefined*, so these won't be prefixed if the document is reindexed later. However, you would see the new last_modified_dt field.

The easiest way to try this parameter is to start over with a fresh collection.

ExtractingRequestHandler Parameters and Configuration

Solr Cell Parameters

The following parameters are accepted by the ExtractingRequestHandler.

These parameters can be set for each indexing request (as request parameters), or they can be set for all requests to the request handler generally by defining them in solrconfig.xml, as described in [Configuring the ExtractingRequestHandler in solrconfig.xml](#).

capture

Captures XHTML elements with the specified name for a supplementary addition to the Solr document.

This parameter can be useful for copying chunks of the XHTML into a separate field. For instance, it could be used to grab paragraphs (<p>) and index them into a separate field. Note that content is still also captured into the content field.

Example: capture=p (in a request) or <str name="capture">p</str> (in solrconfig.xml)

Output: "p": {"This is a paragraph from my document."}

This parameter can also be used with the fmap.source_field parameter to map content from attributes to a new field.

captureAttr

Indexes attributes of the Tika XHTML elements into separate fields, named after the element. If set to true, when extracting from HTML, Tika can return the href attributes in <a> tags as fields named "a".

Example: captureAttr=true

Output: "div": {"classname1", "classname2"}

commitWithin

Add the document within the specified number of milliseconds.

Example: commitWithin=10000 (10 seconds)

defaultField

A default field to use if the uprefix parameter is not specified and a field cannot otherwise be determined.

Example: `defaultField=_text_`

`extractOnly`

Default is `false`. If `true`, returns the extracted content from Tika without indexing the document. This returns the extracted XHTML as a string in the response. When viewing on a screen, it may be useful to set the `extractFormat` parameter for a response format other than XML to aid in viewing the embedded XHTML tags.

Example: `extractOnly=true`

`extractFormat`

The default is `xml`, but the other option is `text`. Controls the serialization format of the extract content.

The `xml` format is actually XHTML, the same format that results from passing the `-x` command to the Tika command line application, while the `text` format is like that produced by Tika's `-t` command.

This parameter is valid only if `extractOnly` is set to `true`.

Example: `extractFormat=text`

Output: For an example output (in XML), see <http://wiki.apache.org/solr/TikaExtractOnlyExampleOutput>

`fmap.source_field`

Maps (moves) one field name to another. The `source_field` must be a field in incoming documents, and the value is the Solr field to map to.

Example: `fmap.content=text` causes the data in the `content` field generated by Tika to be moved to the Solr's `text` field.

`ignoreTikaException`

If `true`, exceptions found during processing will be skipped. Any metadata available, however, will be indexed.

Example: `ignoreTikaException=true`

`literal.fieldname`

Populates a field with the name supplied with the specified value for each document. The data can be multivalued if the field is multivalued.

Example: `literal.doc_status=published`

Output: `"doc_status": "published"`

`literalsOverride`

If `true` (the default), literal field values will override other values with the same field name.

If `false`, literal values defined with `literal.fieldname` will be appended to data already in the fields extracted from Tika. When setting `literalsOverride` to `false`, the field must be multivalued.

Example: `literalsOverride=false`

lowernames

If true, all field names will be mapped to lowercase with underscores, if needed.

Example: `lowernames=true`

Output: Assuming input of "Content-Type", the result in documents would be a field `content_type`

multipartUploadLimitInKB

Defines the size in kilobytes of documents to allow. The default is 2048 (2Mb). If you have very large documents, you should increase this or they will be rejected.

Example: `multipartUploadLimitInKB=2048000`

parseContext.config

If a Tika parser being used allows parameters, you can pass them to Tika by creating a parser configuration file and pointing Solr to it. See the section [Parser-Specific Properties](#) for more information about how to use this parameter.

Example: `parseContext.config=pdf-config.xml`

passwordsFile

Defines a file path and name for a file of file name to password mappings. See the section [Indexing Encrypted Documents](#) for more information about using a password file.

Example: `passwordsFile=/path/to/passwords.txt`

resource.name

Specifies the name of the file to index. This is optional, but Tika can use it as a hint for detecting a file's MIME type.

Example: `resource.name=mydoc.doc`

resource.password

Defines a password to use for a password-protected PDF or OOXML file. See the section [Indexing Encrypted Documents](#) for more information about using this parameter.

Example: `resource.password=secret`

tika.config

Defines a file path and name to a custom Tika configuration file. This is only required if you have customized your Tika implementation.

Example: `tika.config=/path/to/tika.config`

uprefix

PREFIXES all fields *that are undefined in the schema* with the given prefix. This is very useful when combined with dynamic field definitions.

Example: `uprefix=ignored_` would add `ignored_` as a prefix to all unknown fields. In this case, you could additionally define a rule in the Schema to not index these fields:

```
<dynamicField name="ignored_*" type="ignored" />
```

xpath

When extracting, only return Tika XHTML content that satisfies the given XPath expression. See <http://tika.apache.org/1.19.1/> for details on the format of Tika XHTML, it varies with the format being parsed. Also see the section [Defining XPath Expressions](#) for an example.

Configuring the ExtractingRequestHandler in solrconfig.xml

If you have started Solr with one of the supplied [example configsets](#), you already have the ExtractingRequestHandler configured by default and you only need to customize it for your content.

If you are not working with an example configset, the jars required to use Solr Cell will not be loaded automatically. You will need to configure your `solrconfig.xml` to find the ExtractingRequestHandler and its dependencies:

```
<lib dir="${solr.install.dir:../../..}/contrib/extraction/lib" regex=".*/\.jar" />
<lib dir="${solr.install.dir:../../..}/dist/" regex="solr-cell-\d.*\.jar" />
```

You can then configure the ExtractingRequestHandler in `solrconfig.xml`. The following is the default configuration found in Solr's `_default` configset, which you can modify as needed:

```
<requestHandler name="/update/extract"
                 startup="lazy"
                 class="solr.extraction.ExtractingRequestHandler" >
  <lst name="defaults">
    <str name="lowernames">true</str>
    <str name="fmap.content">_text_</str>
  </lst>
</requestHandler>
```

In this setup, all field names are lower-cased (with the `lowernames` parameter), and Tika's content field is mapped to Solr's `text` field.

You may need to configure [Update Request Processors](#) (URPs) that parse numbers and dates and do other manipulations on the metadata fields generated by Solr Cell.

In Solr's default configsets, "[schemaless](#)" (aka data driven, or field guessing) mode is enabled, which does a variety of such processing already.

If you instead explicitly define the fields for your schema, you can selectively specify the desired URPs. An easy way to specify this is to configure the parameter processor (under defaults) to `uuid,remove-blank,field-name-mutating,parse-boolean,parse-long,parse-double,parse-date`. For example:



```
<requestHandler name="/update/extract"
    startup="lazy"
    class="solr.extraction.ExtractingRequestHandler" >
  <lst name="defaults">
    <str name="lowerNames">true</str>
    <str name="fmap.content">_text_</str>
    <str name="processor">uuid,remove-blank,field-name-mutating,parse-
boolean,parse-long,parse-double,parse-date</processor>
  </lst>
</requestHandler>
```

The above suggested list was taken from the list of URPs that run as a part of schemaless mode and provide much of its functionality. However, one major part of the schemaless functionality is missing from the suggested list, `add-unknown-fields-to-the-schema`, which is the part that adds fields to the schema. So you can use the other URPs without worrying about unexpected field additions.

Parser-Specific Properties

Parsers used by Tika may have specific properties to govern how data is extracted. These can be passed through Solr for special parsing situations.

For instance, when using the Tika library from a Java program, the `PDFParserConfig` class has a method `setSortByPosition(boolean)` that can extract vertically oriented text. To access that method via configuration with the `ExtractingRequestHandler`, one can add the `parseContext.config` property to `solrconfig.xml` and then set properties in Tika's `PDFParserConfig` as in the example below.

```
<entries>
  <entry class="org.apache.tika.parser.pdf.PDFParserConfig" impl=
"org.apache.tika.parser.pdf.PDFParserConfig">
    <property name="extractInlineImages" value="true"/>
    <property name="sortByPosition" value="true"/>
  </entry>
  <entry>...</entry>
</entries>
```

Consult the Tika Java API documentation for configuration parameters that can be set for any particular parsers that require this level of control.

Indexing Encrypted Documents

The ExtractingRequestHandler will decrypt encrypted files and index their content if you supply a password in either resource.password on the request, or in a passwordsFile file.

In the case of passwordsFile, the file supplied must be formatted so there is one line per rule. Each rule contains a file name regular expression, followed by "=", then the password in clear-text. Because the passwords are in clear-text, the file should have strict access restrictions.

```
# This is a comment
myFileName = myPassword
.*\.docx$ = myWordPassword
.*\.pdf$ = myPdfPassword
```

Multi-Core Configuration

For a multi-core configuration, you can specify sharedLib='lib' in the <solr/> section of solr.xml and place the necessary jar files there.

For more information about Solr cores, see [The Well-Configured Solr Instance](#).

Extending the ExtractingRequestHandler

If you want to supply your own ContentHandler for Solr to use, you can extend the ExtractingRequestHandler and override the createFactory() method. This factory is responsible for constructing the SolrContentHandler that interacts with Tika, and allows literals to override Tika-parsed values. Set the parameter literalsOverride, which normally defaults to true, to false to append Tika-parsed values to literal values.

Solr Cell Internals

Metadata Created by Tika

As mentioned before, Tika produces metadata about the document. Metadata describes different aspects of a document, such as the author's name, the number of pages, the file size, and so on. The metadata produced depends on the type of document submitted. For instance, PDFs have different metadata than Word documents do.

Metadata Added by Solr

In addition to the metadata added by Tika's parsers, Solr adds the following metadata:

stream_name

The name of the Content Stream as uploaded to Solr. Depending on how the file is uploaded, this may or may not be set.

stream_source_info

Any source info about the stream.

stream_size

The size of the stream in bytes.

stream_content_type

The content type of the stream, if available.



It's recommended to use the extractOnly option before indexing to discover the values Solr will set for these metadata elements on your content.

Order of Input Processing

Here is the order in which the Solr Cell framework processes its input:

1. Tika generates fields or passes them in as literals specified by literal.<fieldname>=<value>. If literalsOverride=false, literals will be appended as multi-value to the Tika-generated field.
2. If lowerNames=true, Tika maps fields to lowercase.
3. Tika applies the mapping rules specified by fmap.source=target parameters.
4. If uprefix is specified, any unknown field names are prefixed with that value, else if defaultField is specified, any unknown fields are copied to the default field.

Solr Cell Examples

Using capture and Mapping Fields

The command below captures <div> tags separately (capture=div), and then maps all the instances of that field to a dynamic field named foo_t (fmap.div=foo_t).

```
bin/post -c gettingstarted example/exemplar/docs/sample.html -params  
"literal.id=doc2&captureAttr=true&defaultField=_text_&fmap.div=foo_t&capture=div"
```

Using Literals to Define Custom Metadata

To add in your own metadata, pass in the literal parameter along with the file:

```
bin/post -c gettingstarted -params  
"literal.id=doc4&captureAttr=true&defaultField=text&capture=div&fmap.div=foo_t&literal.blah_s=Bah  
" example/exemplar/docs/sample.html
```

The parameter literal.blah_s=Bah will insert a field blah_s into every document. Every instance of the text will be "Bah".

Defining XPath Expressions

The example below passes in an XPath expression to restrict the XHTML returned by Tika:

```
bin/post -c gettingstarted -params  
"literal.id=doc5&captureAttr=true&defaultField=text&capture=div&fmap.div=foo_t&xpath=/xhtml:html/  
xhtml:body/xhtml:div//node()" example/exapeddocs/sample.html
```

Extracting Data without Indexing

Solr allows you to extract data without indexing. You might want to do this if you're using Solr solely as an extraction server or if you're interested in testing Solr extraction.

The example below sets the `extractOnly=true` parameter to extract data without indexing it.

```
curl "http://localhost:8983/solr/gettingstarted/update/extract?&extractOnly=true" --data-binary  
@example/exapeddocs/sample.html -H 'Content-type:text/html'
```

The output includes XML generated by Tika (and further escaped by Solr's XML) using a different output format to make it more readable (`-out yes` instructs the tool to echo Solr's output to the console):

```
bin/post -c gettingstarted -params "extractOnly=true&wt=ruby&indent=true" -out yes  
example/exapeddocs/sample.html
```

Using Solr Cell with a POST Request

The example below streams the file as the body of the POST, which does not, then, provide information to Solr about the name of the file.

```
curl  
"http://localhost:8983/solr/gettingstarted/update/extract?literal.id=doc6&defaultField=text&commi  
t=true" --data-binary @example/exapeddocs/sample.html -H 'Content-type:text/html'
```

Using Solr Cell with SolrJ

SolrJ is a Java client that you can use to add documents to the index, update the index, or query the index. You'll find more information on SolrJ in [Using SolrJ](#).

Here's an example of using Solr Cell and SolrJ to add documents to a Solr index.

First, let's use SolrJ to create a new SolrClient, then we'll construct a request containing a ContentStream (essentially a wrapper around a file) and sent it to Solr:

```
public class SolrCellRequestDemo {  
    public static void main (String[] args) throws IOException, SolrServerException {  
        SolrClient client = new HttpSolrClient.Builder("http://localhost:8983/solr/my_collection")  
.build();  
        ContentStreamUpdateRequest req = new ContentStreamUpdateRequest("/update/extract");  
        req.addFile(new File("my-file.pdf"));  
        req.setParam(ExtractingParams.EXTRACT_ONLY, "true");  
        NamedList<Object> result = client.request(req);  
        System.out.println("Result: " + result);  
    }  
}
```

This operation streams the file `my-file.pdf` into the Solr index for `my_collection`.

The sample code above calls the extract command, but you can easily substitute other commands that are supported by Solr Cell. The key class to use is the `ContentStreamUpdateRequest`, which makes sure the ContentStreams are set properly. SolrJ takes care of the rest.

Note that the `ContentStreamUpdateRequest` is not just specific to Solr Cell. You can send CSV to the CSV Update handler and to any other Request Handler that works with Content Streams for updates.

Uploading Structured Data Store Data with the Data Import Handler

Many search applications store the content to be indexed in a structured data store, such as a relational database. The Data Import Handler (DIH) provides a mechanism for importing content from a data store and indexing it.

In addition to relational databases, DIH can index content from HTTP based data sources such as RSS and ATOM feeds, e-mail repositories, and structured XML where an XPath processor is used to generate fields.

DIH Concepts and Terminology

Descriptions of the Data Import Handler use several familiar terms, such as entity and processor, in specific ways, as explained in the table below.

Datasource

As its name suggests, a datasource defines the location of the data of interest. For a database, it's a DSN. For an HTTP datasource, it's the base URL.

Entity

Conceptually, an entity is processed to generate a set of documents, containing multiple fields, which (after optionally being transformed in various ways) are sent to Solr for indexing. For a RDBMS data source, an entity is a view or table, which would be processed by one or more SQL statements to generate a set of rows (documents) with one or more columns (fields).

Processor

An entity processor does the work of extracting content from a data source, transforming it, and adding it to the index. Custom entity processors can be written to extend or replace the ones supplied.

Transformer

Each set of fields fetched by the entity may optionally be transformed. This process can modify the fields, create new fields, or generate multiple rows/documents from a single row. There are several built-in transformers in the DIH, which perform functions such as modifying dates and stripping HTML. It is possible to write custom transformers using the publicly available interface.

Solr's DIH Examples

The example/example-DIH directory contains several collections to demonstrate many of the features of the data import handler. These are available with the dih example from the [Solr Control Script](#):

```
bin/solr -e dih
```

This launches a standalone Solr instance with several collections that correspond to detailed examples. The available examples are atom, db, mail, solr, and tika.

All examples in this section assume you are running the DIH example server.

Configuring DIH

Configuring solrconfig.xml for DIH

The Data Import Handler has to be registered in `solrconfig.xml`. For example:

```
<requestHandler name="/dataimport" class="org.apache.solr.handler.dataimport.DataImportHandler">
  <lst name="defaults">
    <str name="config">/path/to/my/DIHconfigfile.xml</str>
  </lst>
</requestHandler>
```

The only required parameter is the `config` parameter, which specifies the location of the DIH configuration file that contains specifications for the data source, how to fetch data, what data to fetch, and how to process it to generate the Solr documents to be posted to the index.

You can have multiple DIH configuration files. Each file would require a separate definition in the `solrconfig.xml` file, specifying a path to the file.

Configuring the DIH Configuration File

An annotated configuration file, based on the `db` collection in the `dih` example server, is shown below (this file is located in `example/example-DIH/solr/db/conf/db-data-config.xml`).

This example shows how to extract fields from four tables defining a simple product database. More information about the parameters and options shown here will be described in the sections following.

```

<dataConfig>

    <dataSource driver="org.hsqldb.jdbcDriver" url="jdbc:hsqldb:./example-DIH/hsqldb/ex" ①
        user="sa" password="secret"/> ②

    <document> ③
        <entity name="item" query="select * from item"
            deltaQuery="select id from item where last_modified >
'${dataimporter.last_index_time}'"> ④
            <field column="NAME" name="name" />

            <entity name="feature"
                query="select DESCRIPTION from FEATURE where ITEM_ID='${item.ID}'"
                deltaQuery="select ITEM_ID from FEATURE where last_modified >
'${dataimporter.last_index_time}'"
                parentDeltaQuery="select ID from item where ID=${feature.ITEM_ID}"> ⑤
                <field name="features" column="DESCRIPTION" />
            </entity>

            <entity name="item_category"
                query="select CATEGORY_ID from item_category where ITEM_ID='${item.ID}'"
                deltaQuery="select ITEM_ID, CATEGORY_ID from item_category where last_modified >
'${dataimporter.last_index_time}'"
                parentDeltaQuery="select ID from item where ID=${item_category.ITEM_ID}">
                <entity name="category"
                    query="select DESCRIPTION from category where ID =
'${item_category.CATEGORY_ID}'"
                    deltaQuery="select ID from category where last_modified >
'${dataimporter.last_index_time}'"
                    parentDeltaQuery="select ITEM_ID, CATEGORY_ID from item_category where
CATEGORY_ID=${category.ID}">
                    <field column="description" name="cat" />
                </entity>
            </entity>
        </entity>
    </document>
</dataConfig>

```

- ① The first element is the dataSource, in this case an HSQLDB database. The path to the JDBC driver and the JDBC URL and login credentials are all specified here. Other permissible attributes include whether or not to autocommit to Solr, the batchsize used in the JDBC connection, and a readOnly flag.
- ② The password attribute is optional if there is no password set for the DB. Alternately, the password can be encrypted; the section [Encrypting a Database Password](#) below describes how to do this.
- ③ A document element follows, containing multiple entity elements. Note that entity elements can be nested, and this allows the entity relationships in the sample database to be mirrored here, so that we can generate a denormalized Solr record which may include multiple features for one item, for instance.
- ④ The possible attributes for the entity element are described in later sections. Entity elements may contain one or more field elements, which map the data source field names to Solr fields, and optionally specify per-field transformations. This entity is the root entity.
- ⑤ This entity is nested and reflects the one-to-many relationship between an item and its multiple

features. Note the use of variables; \${item.ID} is the value of the column 'ID' for the current item (item referring to the entity name).

Datasources can still be specified in solrconfig.xml. These must be specified in the defaults section of the handler in solrconfig.xml. However, these are not parsed until the main configuration is loaded.

The entire configuration itself can be passed as a request parameter using the dataConfig parameter rather than using a file. When configuration errors are encountered, the error message is returned in XML format.

A reload-config command is also supported, which is useful for validating a new configuration file, or if you want to specify a file, load it, and not have it reloaded again on import. If there is an xml mistake in the configuration a user-friendly message is returned in xml format. You can then fix the problem and do a reload-config.



You can also view the DIH configuration in the Solr Admin UI from the [Dataimport Screen](#). It includes an interface to import content.

DIH Request Parameters

Request parameters can be substituted in configuration with placeholder \${dataimporter.request.paramname}, as in this example:

```
<dataSource driver="org.hsqldb.jdbcDriver"
            url="${dataimporter.request.jdbcurl}"
            user="${dataimporter.request.jdbcuser}"
            password="${dataimporter.request.jdbcpassword}" />
```

These parameters can then be passed to the full-import command or defined in the <defaults> section in solrconfig.xml. This example shows the parameters with the full-import command:

```
http://localhost:8983/solr/dih/dataimport?command=full-import&jdbcurl=jdbc:hsqldb:/example-DIH/hsqldb/ex&jdbcuser=sa&jdbcpassword=secret
```

Encrypting a Database Password

The database password can be encrypted if necessary to avoid plaintext passwords being exposed in unsecured files. To do this, we will replace the password in data-config.xml with an encrypted password. We will use the openssl tool for the encryption, and the encryption key will be stored in a file which is only readable to the solr process. Please follow these steps:

1. Create a strong encryption password and store it in a file. Then make sure it is readable only for the solr user. Example commands:

```
echo -n "a-secret" > /var/solr/data/dih-encryptionkey
chown solr:solr /var/solr/data/dih-encryptionkey
chmod 600 /var/solr/data/dih-encryptionkey
```



Note that we use the `-n` argument to echo to avoid including a newline character at the end of the password. If you use another method to generate the encrypted password, make sure to avoid newlines as well.

2. Encrypt the JDBC database password using `openssl` as follows:

```
echo -n "my-jdbc-password" | openssl enc -aes-128-cbc -a -salt -md md5 -pass  
file:/var/solr/data/dih-encryptionkey
```

The output of the command will be a long string such as
`U2FsdGVkX18QMjY0yfCq1fBMvAB4d3XkwY96L7gf02o=`. You will use this as password in your `data-config.xml` file.

3. In your `data-config.xml`, you'll add the `password` and `encryptKeyFile` parameters to the `<datasource>` configuration, as in this example:

```
<dataSource driver="org.hsqldb.jdbcDriver"  
    url="jdbc:hsqldb:./example-DIH/hsqldb/ex"  
    user="sa"  
    password="U2FsdGVkX18QMjY0yfCq1fBMvAB4d3XkwY96L7gf02o="  
    encryptKeyFile="/var/solr/data/dih-encryptionkey" />
```

DataImportHandler Commands

DIH commands are sent to Solr via an HTTP request. The following operations are supported.

abort

Aborts an ongoing operation. For example:

```
http://localhost:8983/solr/dih/dataimport?command=abort.
```

delta-import

For incremental imports and change detection. Only the [SqlEntityProcessor](#) supports delta imports.

For example: `http://localhost:8983/solr/dih/dataimport?command=delta-import`.

This command supports the same `clean`, `commit`, `optimize` and `debug` parameters as `full-import` command described below.

full-import

A Full Import operation can be started with a URL such as

```
http://localhost:8983/solr/dih/dataimport?command=full-import. The command returns  
immediately.
```

The operation will be started in a new thread and the `status` attribute in the response should be shown as `busy`. The operation may take some time depending on the size of dataset. Queries to Solr are not blocked during full-imports.

When a `full-import` command is executed, it stores the start time of the operation in a file located at

`conf/dataimport.properties`. This stored timestamp is used when a delta-import operation is executed.

Commands available to full-import are:

clean

Default is true. Tells whether to clean up the index before the indexing is started.

commit

Default is true. Tells whether to commit after the operation.

debug

Default is false. Runs the command in debug mode and is used by the interactive development mode.

Note that in debug mode, documents are never committed automatically. If you want to run debug mode and commit the results too, add `commit=true` as a request parameter.

entity

The name of an entity directly under the `<document>` tag in the configuration file. Use this to execute one or more entities selectively.

Multiple "entity" parameters can be passed on to run multiple entities at once. If nothing is passed, all entities are executed.

optimize

Default is true. Tells Solr whether to optimize after the operation.

synchronous

Blocks request until import is completed. Default is false.

reload-config

If the configuration file has been changed and you wish to reload it without restarting Solr, run the command `http://localhost:8983/solr/dih/dataimport?command=reload-config`

status

This command returns statistics on the number of documents created, deleted, queries run, rows fetched, status, and so on. For example:

`http://localhost:8983/solr/dih/dataimport?command=status`.

show-config

This command responds with configuration:

`http://localhost:8983/solr/dih/dataimport?command=show-config`.

Property Writer

The `propertyWriter` element defines the date format and locale for use with delta queries. It is an optional configuration. Add the element to the DIH configuration file, directly under the `dataConfig` element.

```
<propertyWriter dateFormat="yyyy-MM-dd HH:mm:ss" type="SimplePropertiesWriter"  
directory="data" filename="my_dih.properties" locale="en-US" />
```

The parameters available are:

dateFormat

A `java.text.SimpleDateFormat` to use when converting the date to text. The default is `yyyy-MM-dd HH:mm:ss`.

type

The implementation class. Use `SimplePropertiesWriter` for non-SolrCloud installations. If using SolrCloud, use `ZKPropertiesWriter`.

If this is not specified, it will default to the appropriate class depending on if SolrCloud mode is enabled.

directory

Used with the `SimplePropertiesWriter` only. The directory for the properties file. If not specified, the default is `conf`.

filename

Used with the `SimplePropertiesWriter` only. The name of the properties file.

If not specified, the default is the `requestHandler` name (as defined in `solrconfig.xml`, appended by `".properties"` (such as, `dataimport.properties`).

locale

The locale. If not defined, the ROOT locale is used. It must be specified as language-country ([BCP 47 language tag](#)). For example, `en-US`.

Data Sources

A data source specifies the origin of data and its type. Somewhat confusingly, some data sources are configured within the associated entity processor. Data sources can also be specified in `solrconfig.xml`, which is useful when you have multiple environments (for example, development, QA, and production) differing only in their data sources.

You can create a custom data source by writing a class that extends `org.apache.solr.handler.dataimport.DataSource`.

The mandatory attributes for a data source definition are its name and type. The name identifies the data source to an Entity element.

The types of data sources available are described below.

ContentStreamDataSource

This takes the POST data as the data source. This can be used with any EntityProcessor that uses a `DataSource<Reader>`.

FieldReaderDataSource

This can be used where a database field contains XML which you wish to process using the `XPathEntityProcessor`. You would set up a configuration with both JDBC and FieldReader data sources, and two entities, as follows:

```
<dataSource name="a1" driver="org.hsqldb.jdbcDriver" ... />
<dataSource name="a2" type="FieldReaderDataSource" />
<document>

    <!-- processor for database -->
    <entity name ="e1" dataSource="a1" processor="SqlEntityProcessor" pk="docid"
           query="select * from t1 ..." >

        <!-- nested XpathEntity; the field in the parent which is to be used for
            XPath is set in the "datafield" attribute in place of the "url" attribute -->
        <entity name="e2" dataSource="a2" processor="XPathEntityProcessor"
               dataField="e1.fieldToUseForXPath">

            <!-- XPath configuration follows -->
            ...
        </entity>
    </entity>
</document>
```

The `FieldReaderDataSource` can take an encoding parameter, which will default to "UTF-8" if not specified. It must be specified as language-country. For example, en-US.

FileDataSource

This can be used like a [URLDataSource](#), but is used to fetch content from files on disk. The only difference from `URLDataSource`, when accessing disk files, is how a pathname is specified.

This data source accepts these optional attributes.

basePath

The base path relative to which the value is evaluated if it is not absolute.

encoding

Defines the character encoding to use. If not defined, UTF-8 is used.

JdbcDataSource

This is the default datasource. It's used with the [SqlEntityProcessor](#). See the example in the [FieldReaderDataSource](#) section for details on configuration. JdbcDatasource supports at least the following attributes:

driver, url, user, password, encryptKeyFile

Usual JDBC connection properties.

batchSize

Passed to Statement#setFetchSize, default value 500.

For MySQL driver, which doesn't honor batchSize and pulls whole resultSet, which often lead to OutOfMemoryError.

In this case, set batchSize=-1 that pass setFetchSize(Integer.MIN_VALUE), and switch result set to pull row by row

All of them substitute properties via \${placeholders}.

URLDataSource

This data source is often used with [XPathEntityProcessor](#) to fetch content from an underlying file:// or http:// location. Here's an example:

```
<dataSource name="a"
    type="URLDataSource"
    baseUrl="http://host:port/"
    encoding="UTF-8"
    connectionTimeout="5000"
    readTimeout="10000"/>
```

The URLDataSource type accepts these optional parameters:

baseUrl

Specifies a new baseURL for pathnames. You can use this to specify host/port changes between Dev/QA/Prod environments. Using this attribute isolates the changes to be made to the solrconfig.xml

connectionTimeout

Specifies the length of time in milliseconds after which the connection should time out. The default value is 5000ms.

encoding

By default the encoding in the response header is used. You can use this property to override the default encoding.

readTimeout

Specifies the length of time in milliseconds after which a read operation should time out. The default value is 10000ms.

Entity Processors

Entity processors extract data, transform it, and add it to a Solr index. Examples of entities include views or tables in a data store.

Each processor has its own set of attributes, described in its own section below. In addition, there are several attributes common to all entities which may be specified:

dataSource

The name of a data source. If there are multiple data sources defined, use this attribute with the name of the data source for this entity.

name

Required. The unique name used to identify an entity.

pk

The primary key for the entity. It is optional, and required only when using delta-imports. It has no relation to the uniqueKey defined in schema.xml but they can both be the same.

This attribute is mandatory if you do delta-imports and then refer to the column name in \${dataimporter.delta.<column-name>} which is used as the primary key.

processor

Default is [SqlEntityProcessor](#). Required only if the datasource is not RDBMS.

onError

Defines what to do if an error is encountered.

Permissible values are:

abort

Stops the import.

skip

Skips the current document.

continue

Ignores the error and processing continues.

preImportDeleteQuery

Before a full-import command, use this query this to cleanup the index instead of using *.*. This is honored only on an entity that is an immediate sub-child of <document>.

postImportDeleteQuery

Similar to preImportDeleteQuery, but it executes after the import has completed.

rootEntity

By default the entities immediately under <document> are root entities. If this attribute is set to false, the entity directly falling under that entity will be treated as the root entity (and so on). For every row returned by the root entity, a document is created in Solr.

transformer

Optional. One or more transformers to be applied on this entity.

cacheImpl

Optional. A class (which must implement DIHCache) to use for caching this entity when doing lookups from an entity which wraps it. Provided implementation is SortedMapBackedCache.

cacheKey

The name of a property of this entity to use as a cache key if cacheImpl is specified.

cacheLookup

An entity + property name that will be used to lookup cached instances of this entity if cacheImpl is specified.

where

An alternative way to specify cacheKey and cacheLookup concatenated with '='.

For example, where="CODE=People.COUNTRY_CODE" is equivalent to cacheKey="CODE"
cacheLookup="People.COUNTRY_CODE"

child="true"

Enables indexing document blocks aka [Nested Child Documents](#) for searching with [Block Join Query Parsers](#). It can be only specified on the <entity> element under another root entity. It switches from default behavior (merging field values) to nesting documents as children documents.

Note: parent <entity> should add a field which is used as a parent filter in query time.

join="zipper"

Enables merge join, aka "zipper" algorithm, for joining parent and child entities without cache. It should be specified at child (nested) <entity>. It implies that parent and child queries return results ordered by keys, otherwise it throws an exception. Keys should be specified either with where attribute or with cacheKey and cacheLookup.

Entity Caching

Caching of entities in DIH is provided to avoid repeated lookups for same entities again and again. The default SortedMapBackedCache is a HashMap where a key is a field in the row and the value is a bunch of rows for that same key.

In the example below, each manufacturer entity is cached using the id property as a cache key. Cache lookups will be performed for each product entity based on the product's manu property. When the cache has no data for a particular key, the query is run and the cache is populated

```
<entity name="product" query="select description,sku, manu from product" >
  <entity name="manufacturer" query="select id, name from manufacturer"
    cacheKey="id" cacheLookup="product.manu" cacheImpl="SortedMapBackedCache" />
</entity>
```

The SQL Entity Processor

The SqlEntityProcessor is the default processor. The associated [JdbcDataSource](#) should be a JDBC URL.

The entity attributes specific to this processor are shown in the table below. These are in addition to the attributes common to all entity processors described above.

query

Required. The SQL query used to select rows.

deltaQuery

SQL query used if the operation is delta-import. This query selects the primary keys of the rows which will be parts of the delta-update. The pks will be available to the deltaImportQuery through the variable \${dataimporter.delta.<column-name>}.

parentDeltaQuery

SQL query used if the operation is delta-import.

deletedPkQuery

SQL query used if the operation is delta-import.

deltaImportQuery

SQL query used if the operation is delta-import. If this is not present, DIH tries to construct the import query by (after identifying the delta) modifying the 'query' (this is error prone).

There is a namespace \${dataimporter.delta.<column-name>} which can be used in this query. For example, select * from tbl where id=\${dataimporter.delta.id}.

The **XPathEntityProcessor**

This processor is used when indexing XML formatted data. The data source is typically [URLDataSource](#) or [FileDataSource](#). XPath can also be used with the [FileListEntityProcessor](#) described below, to generate a document from each file.

The entity attributes unique to this processor are shown below. These are in addition to the attributes common to all entity processors described above.

Processor

Required. Must be set to XpathEntityProcessor.

url

Required. The HTTP URL or file location.

stream

Optional: Set to true for a large file or download.

forEach

Required unless you define useSolrAddSchema. The XPath expression which demarcates each record. This will be used to set up the processing loop.

xsl

Optional: Its value (a URL or filesystem path) is the name of a resource used as a preprocessor for applying the XSL transformation.

useSolrAddSchema

Set this to true if the content is in the form of the standard Solr update XML schema.

Each <field> element in the entity can have the following attributes as well as the default ones.

xpath

Required. The XPath expression which will extract the content from the record for this field. Only a subset of XPath syntax is supported.

commonField

Optional. If true, then when this field is encountered in a record it will be copied to future records when creating a Solr document.

flatten

Optional. If set to true, then any children text nodes are collected to form the value of a field.



The default value is false, meaning that if there are any sub-elements of the node pointed to by the XPath expression, they will be quietly omitted.

Here is an example from the atom collection in the dih example (data-config file found at example/example-DIH/solr/atom/conf/atom-data-config.xml):

```
<dataConfig>
  <dataSource type="URLDataSource"/>
  <document>

    <entity name="stackoverflow"
      url="https://stackoverflow.com/feeds/tag/solr"
      processor="XPathEntityProcessor"
      forEach="/feed/feed/entry"
      transformer="HTMLStripTransformer,RegexTransformer">

      <!-- Pick this value up from the feed level and apply to all documents -->
      <field column="lastchecked_dt" xpath="/feed/updated" commonField="true"/>

      <!-- Keep only the final numeric part of the URL -->
      <field column="id" xpath="/feed/entry/id" regex=".*/" replaceWith="" />

      <field column="title" xpath="/feed/entry/title"/>
      <field column="author" xpath="/feed/entry/author/name"/>
      <field column="category" xpath="/feed/entry/category/@term"/>
      <field column="link" xpath="/feed/entry/link[@rel='alternate']/@href"/>

      <!-- Use transformers to convert HTML into plain text.
          There is also an UpdateRequestProcess to trim remaining spaces.
          -->
      <field column="summary" xpath="/feed/entry/summary" stripHTML="true" regex="( |\n)+"
        replaceWith=" "/>

      <!-- Ignore namespaces when matching XPath -->
      <field column="rank" xpath="/feed/entry/rank"/>

      <field column="published_dt" xpath="/feed/entry/published"/>
      <field column="updated_dt" xpath="/feed/entry/updated"/>
    </entity>

  </document>
</dataConfig>
```

The MailEntityProcessor

The MailEntityProcessor uses the Java Mail API to index email messages using the IMAP protocol.

The MailEntityProcessor works by connecting to a specified mailbox using a username and password, fetching the email headers for each message, and then fetching the full email contents to construct a document (one document for each mail message).

The entity attributes unique to the MailEntityProcessor are shown below. These are in addition to the attributes common to all entity processors described above.

processor

Required. Must be set to MailEntityProcessor.

user

Required. Username for authenticating to the IMAP server; this is typically the email address of the mailbox owner.

password

Required. Password for authenticating to the IMAP server.

host

Required. The IMAP server to connect to.

protocol

Required. The IMAP protocol to use, valid values are: imap, imaps, gimap, and gimaps.

fetchMailsSince

Optional. Date/time used to set a filter to import messages that occur after the specified date; expected format is: yyyy-MM-dd HH:mm:ss.

folders

Required. Comma-delimited list of folder names to pull messages from, such as "inbox".

recurse

Optional. Default is true. Flag to indicate if the processor should recurse all child folders when looking for messages to import.

include

Optional. Comma-delimited list of folder patterns to include when processing folders (can be a literal value or regular expression).

exclude

Optional. Comma-delimited list of folder patterns to exclude when processing folders (can be a literal value or regular expression). Excluded folder patterns take precedence over include folder patterns.

processAttachment or processAttachments

Optional. Default is true. Use Tika to process message attachments.

includeContent

Optional. Default is true. Include the message body when constructing Solr documents for indexing.

Here is an example from the `mail` collection of the dih example (data-config file found at `example/example-DIH/mail/conf/mail-data-config.xml`):

```
<dataConfig>
  <document>
    <entity processor="MailEntityProcessor"
      user="email@gmail.com"
      password="password"
      host="imap.gmail.com"
      protocol="imaps"
      fetchMailsSince="2014-06-30 00:00:00"
      batchSize="20"
      folders="inbox"
      processAttachment="false"
      name="mail_entity"/>
  </document>
</dataConfig>
```

Importing New Emails Only

After running a full import, the MailEntityProcessor keeps track of the timestamp of the previous import so that subsequent imports can use the fetchMailsSince filter to only pull new messages from the mail server. This occurs automatically using the DataImportHandler dataimport.properties file (stored in conf).

For instance, if you set fetchMailsSince="2014-08-22 00:00:00" in your mail-data-config.xml, then all mail messages that occur after this date will be imported on the first run of the importer. Subsequent imports will use the date of the previous import as the fetchMailsSince filter, so that only new emails since the last import are indexed each time.

GMail Extensions

When connecting to a GMail account, you can improve the efficiency of the MailEntityProcessor by setting the protocol to **gimap** or **gimaps**.

This allows the processor to send the fetchMailsSince filter to the GMail server to have the date filter applied on the server, which means the processor only receives new messages from the server. However, GMail only supports date granularity, so the server-side filter may return previously seen messages if run more than once a day.

The TikaEntityProcessor

The TikaEntityProcessor uses Apache Tika to process incoming documents. This is similar to [Uploading Data with Solr Cell using Apache Tika](#), but using DataImportHandler options instead.

The parameters for this processor are described in the table below. These are in addition to the attributes common to all entity processors described above.

dataSource

This parameter defines the data source and an optional name which can be referred to in later parts of the configuration if needed. This is the same dataSource explained in the description of general entity processor attributes above.

The available data source types for this processor are:

- BinURLDataSource: used for HTTP resources, but can also be used for files.
- BinContentStreamDataSource: used for uploading content as a stream.
- BinFileDataSource: used for content on the local filesystem.

url

Required. The path to the source file(s), as a file path or a traditional internet URL.

htmlMapper

Optional. Allows control of how Tika parses HTML. If this parameter is defined, it must be either **default** or **identity**; if it is absent, "default" is assumed.

The "default" mapper strips much of the HTML from documents while the "identity" mapper passes all HTML as-is with no modifications.

format

The output format. The options are **text**, **xml**, **html** or **none**. The default is "text" if not defined. The format "none" can be used if metadata only should be indexed and not the body of the documents.

parser

Optional. The default parser is org.apache.tika.parser.AutoDetectParser. If a custom or other parser should be used, it should be entered as a fully-qualified name of the class and path.

fields

The list of fields from the input documents and how they should be mapped to Solr fields. If the attribute meta is defined as "true", the field will be obtained from the metadata of the document and not parsed from the body of the main text.

extractEmbedded

Instructs the TikaEntityProcessor to extract embedded documents or attachments when **true**. If false, embedded documents and attachments will be ignored.

onError

By default, the TikaEntityProcessor will stop processing documents if it finds one that generates an error. If you define onError to "skip", the TikaEntityProcessor will instead skip documents that fail processing and log a message that the document was skipped.

Here is an example from the tika collection of the dih example (data-config file found in example/example-DIH/tika/conf/tika-data-config.xml):

```
<dataConfig>
  <dataSource type="BinFileDataSource"/>
  <document>
    <entity name="file" processor="FileListEntityProcessor" dataSource="null"
      baseDir="${solr.install.dir}/example/exampledocs" fileName=".pdf"
      rootEntity="false">

      <field column="file" name="id"/>

      <entity name="pdf" processor="TikaEntityProcessor"
        url="${file.fileAbsolutePath}" format="text">

        <field column="Author" name="author" meta="true"/>
        <!-- in the original PDF, the Author meta-field name is upper-cased,
            but in Solr schema it is lower-cased
        -->

        <field column="title" name="title" meta="true"/>
        <field column="dc:format" name="format" meta="true"/>

        <field column="text" name="text"/>

      </entity>
    </entity>
  </document>
</dataConfig>
```

The FileListEntityProcessor

This processor is basically a wrapper, and is designed to generate a set of files satisfying conditions specified in the attributes which can then be passed to another processor, such as the [XPathEntityProcessor](#).

The entity information for this processor would be nested within the FileListEntity entry. It generates five implicit fields: fileAbsolutePath, fileDir, fileSize, fileLastModified, and file, which can be used in the nested processor. This processor does not use a data source.

The attributes specific to this processor are described in the table below:

fileName

Required. A regular expression pattern to identify files to be included.

basedir

Required. The base directory (absolute path).

recursive

Whether to search directories recursively. Default is 'false'.

excludes

A regular expression pattern to identify files which will be excluded.

newerThan

A date in the format yyyy-MM-ddHH:mm:ss or a date math expression (NOW - 2YEARS).

olderThan

A date, using the same formats as newerThan.

rootEntity

This should be set to false. This ensures that each row (filepath) emitted by this processor is considered to be a document.

dataSource

Must be set to null.

The example below shows the combination of the FileListEntityProcessor with another processor which will generate a set of fields from each file found.

```
<dataConfig>
  <dataSource type="FileDataSource"/>
  <document>
    <!-- this outer processor generates a list of files satisfying the conditions
        specified in the attributes -->
    <entity name="f" processor="FileListEntityProcessor"
      fileName=".xml"
      newerThan="'NOW-30DAYS'"
      recursive="true"
      rootEntity="false"
      dataSource="null"
      baseDir="/my/document/directory">

      <!-- this processor extracts content using XPath from each file found -->

      <entity name="nested" processor="XPathEntityProcessor"
        forEach="/rootelement" url="${f.fileAbsolutePath}" >
        <field column="name" xpath="/rootelement/name"/>
        <field column="number" xpath="/rootelement/number"/>
      </entity>
    </entity>
  </document>
</dataConfig>
```

LineEntityProcessor

This EntityProcessor reads all content from the data source on a line by line basis and returns a field called rawLine for each line read. The content is not parsed in any way; however, you may add transformers to manipulate the data within the rawLine field, or to create other additional fields.

The lines read can be filtered by two regular expressions specified with the acceptLineRegex and omitLineRegex attributes.

The LineEntityProcessor has the following attributes:

url

A required attribute that specifies the location of the input file in a way that is compatible with the configured data source. If this value is relative and you are using FileDataSource or URLDataSource, it assumed to be relative to baseLoc.

acceptLineRegex

An optional attribute that if present discards any line which does not match the regular expression.

omitLineRegex

An optional attribute that is applied after any acceptLineRegex and that discards any line which matches this regular expression.

For example:

```
<entity name="jc"
    processor="LineEntityProcessor"
    acceptLineRegex="^.*\.xml$"
    omitLineRegex="/obsolete"
    url="file:///Volumes/ts/files.lis"
    rootEntity="false"
    dataSource="myURIreader1"
    transformer="RegexTransformer,DateFormatTransformer">
</entity>
```

While there are use cases where you might need to create a Solr document for each line read from a file, it is expected that in most cases that the lines read by this processor will consist of a pathname, which in turn will be consumed by another entity processor, such as the XPathEntityProcessor.

PlainTextEntityProcessor

This EntityProcessor reads all content from the data source into a single implicit field called plainText. The content is not parsed in any way, however you may add [transformers](#) to manipulate the data within the plainText as needed, or to create other additional fields.

For example:

```
<entity processor="PlainTextEntityProcessor" name="x" url="http://abc.com/a.txt" dataSource=
"data-source-name">
  <!-- copies the text to a field called 'text' in Solr-->
  <field column="plainText" name="text"/>
</entity>
```

Ensure that the dataSource is of type [DataSource<Reader>](#) (FileDataSource, URLDataSource).

SolrEntityProcessor

This EntityProcessor imports data from different Solr instances and cores. The data is retrieved based on a specified filter query. This EntityProcessor is useful in cases you want to copy your Solr index and want to modify the data in the target index.

The SolrEntityProcessor can only copy fields that are stored in the source index.

The SolrEntityProcessor supports the following parameters:

url

Required. The URL of the source Solr instance and/or core.

query

Required. The main query to execute on the source index.

fq

Any filter queries to execute on the source index. If more than one filter query is defined, they must be separated by a comma.

rows

The number of rows to return for each iteration. The default is 50 rows.

fl

A comma-separated list of fields to fetch from the source index. Note, these fields must be stored in the source Solr instance.

qt

The search handler to use, if not the default.

wt

The response format to use, either **javabin** or **xml**.

timeout

The query timeout in seconds. The default is 5 minutes (300 seconds).

cursorMark="true"

Use this to enable cursor for efficient result set scrolling.

sort="id asc"

This should be used to specify a sort parameter referencing the uniqueKey field of the source Solr instance. See [Pagination of Results](#) for details.

Here is a simple example of a SolrEntityProcessor:

```
<dataConfig>
  <document>
    <entity name="sep" processor="SolrEntityProcessor"
      url="http://127.0.0.1:8983/solr/db "
      query="*:*"
      fl="*,orig_version_l:_version_,ignored_price_c:price_c"/>
  </document>
</dataConfig>
```

Transformers

Transformers manipulate the fields in a document returned by an entity. A transformer can create new fields or modify existing ones. You must tell the entity which transformers your import operation will be using, by adding an attribute containing a comma separated list to the <entity> element.

```
<entity name="abcde" transformer="org.apache.solr....,my.own.transformer,..." />
```

Specific transformation rules are then added to the attributes of a <field> element, as shown in the examples below. The transformers are applied in the order in which they are specified in the transformer attribute.

The DataImportHandler contains several built-in transformers. You can also write your own custom transformers, as described in the [DIHCustomTransformer](#) section of the Solr Wiki. The ScriptTransformer (described below) offers an alternative method for writing your own transformers.

ClobTransformer

You can use the ClobTransformer to create a string out of a CLOB in a database. A **CLOB** is a character large object: a collection of character data typically stored in a separate location that is referenced in the database.

The ClobTransformer accepts these attributes:

clob

Boolean value to signal if ClobTransformer should process this field or not. If this attribute is omitted, then the corresponding field is not transformed.

sourceColumnName

The source column to be used as input. If this is absent source and target are same

Here's an example of invoking the ClobTransformer.

```
<entity name="example" transformer="ClobTransformer" ...>
  <field column="hugeTextField" clob="true" />
  ...
</entity>
```

The DateFormatTransformer

This transformer converts dates from one format to another. This would be useful, for example, in a situation where you wanted to convert a field with a fully specified date/time into a less precise date format, for use in faceting.

DateFormatTransformer applies only on the fields with an attribute dateFormat. Other fields are not modified.

This transformer recognizes the following attributes:

dateTimeFormat

The format used for parsing this field. This must comply with the syntax of the [Java SimpleDateFormat class](#).

sourceColName

The column on which the dateFormat is to be applied. If this is absent source and target are same.

locale

The locale to use for date transformations. If not defined, the ROOT locale is used. It must be specified as language-country ([BCP 47 language tag](#)). For example, en-US.

Here is example code that returns the date rounded up to the month "2007-JUL":

```
<entity name="en" pk="id" transformer="DateFormatTransformer" ... >
...
<field column="date" sourceColName="fulldate" dateTimeFormat="yyyy-MMM"/>
</entity>
```

The HTMLStripTransformer

You can use this transformer to strip HTML out of a field.

There is one attribute for this transformer, stripHTML, which is a boolean value (true or false) to signal if the HTMLStripTransformer should process the field or not.

For example:

```
<entity name="e" transformer="HTMLStripTransformer" ... >
<field column="htmlText" stripHTML="true" />
...
</entity>
```

The LogTransformer

You can use this transformer to log data to the console or log files. For example:

```
<entity ...
    transformer="LogTransformer"
    logTemplate="The name is ${e.name}" LogLevel="debug">
    ....
</entity>
```

Unlike other transformers, the LogTransformer does not apply to any field, so the attributes are applied on the entity itself.

The NumberFormatTransformer

Use this transformer to parse a number from a string, converting it into the specified format, and optionally

using a different locale.

NumberFormatTransformer will be applied only to fields with an attribute `formatStyle`.

This transformer recognizes the following attributes:

formatStyle

The format used for parsing this field. The value of the attribute must be one of number, percent, integer, or currency. This uses the semantics of the Java NumberFormat class.

sourceColName

The column on which the NumberFormat is to be applied. This attribute is absent. The source column and the target column are the same.

locale

The locale to be used for parsing the strings. The locale. If not defined, the ROOT locale is used. It must be specified as language-country ([BCP 47 language tag](#)). For example, en-US.

For example:

```
<entity name="en" pk="id" transformer="NumberFormatTransformer" ...>
  ...
  <!-- treat this field as UK pounds -->
  <field name="price_uk" column="price" formatStyle="currency" locale="en-UK"/>
</entity>
```

The RegexTransformer

The regex transformer helps in extracting or manipulating values from fields (from the source) using Regular Expressions. The actual class name is `org.apache.solr.handler.dataimport.RegexTransformer`. But as it belongs to the default package the package-name can be omitted.

The table below describes the attributes recognized by the regex transformer.

regex

The regular expression that is used to match against the column or `sourceColName`'s value(s). If `replaceWith` is absent, each regex *group* is taken as a value and a list of values is returned.

sourceColName

The column on which the regex is to be applied. If not present, then the source and target are identical.

splitBy

Used to split a string. It returns a list of values. Note, this is a regular expression so it may need to be escaped (e.g., via back-slashes).

groupNames

A comma separated list of field column names, used where the regex contains groups and each group is to be saved to a different field. If some groups are not to be named leave a space between commas.

replaceWith

Used along with regex. It is equivalent to the method new
String(<sourceColVal>).replaceAll(<regex>, <replaceWith>).

Here is an example of configuring the regex transformer:

```
<entity name="foo" transformer="RegexTransformer"
    query="select full_name, emailids from foo"> ①
    <field column="full_name"/> ②
    <field column="firstName" regex="Mr(\w*)\b.*" sourceColName="full_name"/>
    <field column="lastName" regex="Mr.*?\b(\w*)" sourceColName="full_name"/>

    <!-- another way of doing the same -->

    <field column="fullName" regex="Mr(\w*)\b(.*)" groupNames="firstName,lastName"/>
    <field column="mailId" splitBy="," sourceColName="emailids"/> ③
</entity>
```

- ① In this example, regex and sourceColName are custom attributes used by the transformer.
- ② The transformer reads the field full_name from the result set and transforms it to two new target fields, firstName and lastName. Even though the query returned only one column, full_name, in the result set, the Solr document gets two extra fields firstName and lastName which are "derived" fields. These new fields are only created if the regexp matches.
- ③ The emailids field in the table can be a comma-separated value. It ends up producing one or more email IDs, and we expect the mailId to be a multivalued field in Solr.

Note that this transformer can be used to either split a string into tokens based on a splitBy pattern, or to perform a string substitution as per replaceWith, or it can assign groups within a pattern to a list of groupNames. It decides what it is to do based upon the above attributes splitBy, replaceWith and groupNames which are looked for in order. This first one found is acted upon and other unrelated attributes are ignored.

The ScriptTransformer

The script transformer allows arbitrary transformer functions to be written in any scripting language supported by Java, such as Javascript, JRuby, Jython, Groovy, or BeanShell. Javascript is integrated into Java by default; you'll need to integrate other languages yourself.

Each function you write must accept a row variable (which corresponds to a Java Map<String, Object>, thus permitting get, put, remove operations). Thus you can modify the value of an existing field or add new fields. The return value of the function is the returned object.

The script is inserted into the DIH configuration file at the top level and is called once for each row.

Here is a simple example.

```

<dataconfig>

    <!-- simple script to generate a new row, converting a temperature from Fahrenheit to
        Centigrade -->

    <script><![CDATA[
        function f2c(row) {
            var tempf, tempc;
            tempf = row.get('temp_f');
            if (tempf != null) {
                tempc = (tempf - 32.0)*5.0/9.0;
                row.put('temp_c', temp_c);
            }
            return row;
        }
    ]]>
</script>
<document>

    <!-- the function is specified as an entity attribute -->

    <entity name="e1" pk="id" transformer="script:f2c" query="select * from X">
        ....
    </entity>
</document>
</dataConfig>

```

The TemplateTransformer

You can use the template transformer to construct or modify a field value, perhaps using the value of other fields. You can insert extra text into the template.

```

<entity name="en" pk="id" transformer="TemplateTransformer" ...>
    ...
    <!-- generate a full address from fields containing the component parts -->
    <field column="full_address" template="${en.street},${en.city},${en.zip}" />
</entity>

```

Special Commands for DIH

You can pass special commands to the DIH by adding any of the variables listed below to any row returned by any component:

\$skipDoc

Skip the current document; that is, do not add it to Solr. The value can be the string true or false.

\$skipRow

Skip the current row. The document will be added with rows from other entities. The value can be the

string true or false.

\$deleteDocById

Delete a document from Solr with this ID. The value has to be the uniqueKey value of the document.

\$deleteDocByQuery

Delete documents from Solr using this query. The value must be a Solr Query.

Updating Parts of Documents

Once you have indexed the content you need in your Solr index, you will want to start thinking about your strategy for dealing with changes to those documents. Solr supports three approaches to updating documents that have only partially changed.

The first is [atomic updates](#). This approach allows changing only one or more fields of a document without having to reindex the entire document.

The second approach is known as [in-place updates](#). This approach is similar to atomic updates (is a subset of atomic updates in some sense), but can be used only for updating single valued non-indexed and non-stored docValue-based numeric fields.

The third approach is known as [optimistic concurrency](#) or [optimistic locking](#). It is a feature of many NoSQL databases, and allows conditional updating a document based on its version. This approach includes semantics and rules for how to deal with version matches or mis-matches.

Atomic Updates (and in-place updates) and Optimistic Concurrency may be used as independent strategies for managing changes to documents, or they may be combined: you can use optimistic concurrency to conditionally apply an atomic update.

Atomic Updates

Solr supports several modifiers that atomically update values of a document. This allows updating only specific fields, which can help speed indexing processes in an environment where speed of index additions is critical to the application.

To use atomic updates, add a modifier to the field that needs to be updated. The content can be updated, added to, or incrementally increased if the field has a numeric type.

set

Set or replace the field value(s) with the specified value(s), or remove the values if 'null' or empty list is specified as the new value.

May be specified as a single value, or as a list for multiValued fields.

add

Adds the specified values to a multiValued field. May be specified as a single value, or as a list.

add-distinct

Adds the specified values to a multiValued field, only if not already present. May be specified as a single value, or as a list.

remove

Removes (all occurrences of) the specified values from a multiValued field. May be specified as a single value, or as a list.

removeregex

Removes all occurrences of the specified regex from a multiValued field. May be specified as a single value, or as a list.

```
inc
```

Increments a numeric value by a specific amount. Must be specified as a single numeric value.

Field Storage

The core functionality of atomically updating a document requires that all fields in your schema must be configured as stored (`stored="true"`) or docValues (`docValues="true"`) except for fields which are `<copyField/>` destinations, which must be configured as `stored="false"`. Atomic updates are applied to the document represented by the existing stored field values. All data in `copyField` destinations fields must originate from ONLY `copyField` sources.

If `<copyField/>` destinations are configured as stored, then Solr will attempt to index both the current value of the field as well as an additional copy from any source fields. If such fields contain some information that comes from the indexing program and some information that comes from `copyField`, then the information which originally came from the indexing program will be lost when an atomic update is made.

There are other kinds of derived fields that must also be set so they aren't stored. Some spatial field types, such as `BBoxField` and `LatLonType`, use derived fields. `CurrencyFieldType` also uses derived fields. These types create additional fields which are normally specified by a dynamic field definition. That dynamic field definition must be not stored, or indexing will fail.

Example Updating Part of a Document

If the following document exists in our collection:

```
{"id":"mydoc",
 "price":10,
 "popularity":42,
 "categories":["kids"],
 "sub_categories":["under_5", "under_10"],
 "promo_ids":["a123x"],
 "tags":["free_to_try", "buy_now", "clearance", "on_sale"]}
```

And we apply the following update command:

```
{"id":"mydoc",
 "price":{"set":99},
 "popularity":{"inc":20},
 "categories":{"add":["toys", "games"]},
 "sub_categories":{"add-distinct":"under_10"},
 "promo_ids":{"remove":"a123x"},
 "tags":{"remove":["free_to_try", "on_sale"]}}
```

The resulting document in our collection will be:

```
{"id": "mydoc",
  "price": 99,
  "popularity": 62,
  "categories": ["kids", "toys", "games"],
  "sub_categories": ["under_5", "under_10"],
  "tags": ["buy_now", "clearance"]
}
```

Updating Child Documents

Solr supports modifying, adding and removing child documents as part of atomic updates.

Schema and configuration requirements are detailed in [Field Storage](#) and [Indexing Nested Documents](#).

Under the hood, Solr retrieves the whole nested structure, deletes the old documents, and re-indexes the structure after applying the atomic update.

Syntactically, nested/partial updates are very similar to a regular atomic update, as demonstrated by the examples below.



route Parameter

To ensure each nested update is routed to its respective shard, `_route_` parameter must be set to the root document's ID when the update does not have that root document.

If the following document exists in our collection:

```
{
  "id": "mydoc",
  "product": "T-Shirt",
  "stock": {
    "id": "mydoc2",
    "color": "red",
    "size": ["L"]
  }
}
```

And we apply the following update command:

```
{
  "id": "mydoc",
  "stock": {
    "add": [
      {
        "id": "mydoc3",
        "color": "blue",
        "size": ["M"]
      }
    ]
  }
}
```

The resulting document in our collection will be:

```
{  
  "id": "mydoc",  
  "product": "T-Shirt",  
  "stock": [{  
    "id": "mydoc2",  
    "color": "red",  
    "size": ["L"]  
  },  
  {  
    "id": "mydoc3",  
    "color": "blue",  
    "size": ["M"]  
  }]  
}
```

Documents inside nested structures can also be updated. These type of updates require setting the `_route_` set to the root document's ID

If we send this update, setting `_route_=mydoc`

```
{  
  "id": "mydoc2",  
  "size": {"add": ["S"]}  
}
```

The resulting document in our collection will be:

```
{  
  "id": "mydoc",  
  "product": "T-Shirt",  
  "stock": [{  
    "id": "mydoc2",  
    "color": "red",  
    "size": ["L", "S"]  
  },  
  {  
    "id": "mydoc3",  
    "color": "blue",  
    "size": ["M"]  
  }]  
}
```

In-Place Updates

In-place updates are very similar to atomic updates; in some sense, this is a subset of atomic updates. In regular atomic updates, the entire document is reindexed internally during the application of the update.

However, in this approach, only the fields to be updated are affected and the rest of the documents are not reindexed internally. Hence, the efficiency of updating in-place is unaffected by the size of the documents that are updated (i.e., number of fields, size of fields, etc.). Apart from these internal differences, there is no functional difference between atomic updates and in-place updates.

An atomic update operation is performed using this approach only when the fields to be updated meet these three conditions:

- are non-indexed (`indexed="false"`), non-stored (`stored="false"`), single valued (`multiValued="false"`) numeric docValues (`docValues="true"`) fields;
- the `_version_` field is also a non-indexed, non-stored single valued docValues field; and,
- copy targets of updated fields, if any, are also non-indexed, non-stored single valued numeric docValues fields.

To use in-place updates, add a modifier to the field that needs to be updated. The content can be updated or incrementally increased.

`set`

Set or replace the field value(s) with the specified value(s). May be specified as a single value.

`inc`

Increments a numeric value by a specific amount. Must be specified as a single numeric value.

In-Place Update Example

If the price and popularity fields are defined in the schema as:

```
<field name="price" type="float" indexed="false" stored="false" docValues="true"/>  
<field name="popularity" type="float" indexed="false" stored="false" docValues="true"/>
```

If the following document exists in our collection:

```
{  
  "id": "mydoc",  
  "price": 10,  
  "popularity": 42,  
  "categories": ["kids"],  
  "promo_ids": ["a123x"],  
  "tags": ["free_to_try", "buy_now", "clearance", "on_sale"]  
}
```

And we apply the following update command:

```
{  
  "id": "mydoc",  
  "price": {"set": 99},  
  "popularity": {"inc": 20}  
}
```

The resulting document in our collection will be:

```
{  
  "id": "mydoc",  
  "price": 99,  
  "popularity": 62,  
  "categories": ["kids"],  
  "promo_ids": ["a123x"],  
  "tags": ["free_to_try", "buy_now", "clearance", "on_sale"]  
}
```

Optimistic Concurrency

Optimistic Concurrency is a feature of Solr that can be used by client applications which update/replace documents to ensure that the document they are replacing/updating has not been concurrently modified by another client application. This feature works by requiring a `_version_` field on all documents in the index, and comparing that to a `_version_` specified as part of the update command. By default, Solr's Schema includes a `_version_` field, and this field is automatically added to each new document.

In general, using optimistic concurrency involves the following work flow:

1. A client reads a document. In Solr, one might retrieve the document with the `/get` handler to be sure to have the latest version.
2. A client changes the document locally.
3. The client resubmits the changed document to Solr, for example, perhaps with the `/update` handler.
4. If there is a version conflict (HTTP error code 409), the client starts the process over.

When the client resubmits a changed document to Solr, the `_version_` can be included with the update to invoke optimistic concurrency control. Specific semantics are used to define when the document should be updated or when to report a conflict.

- If the content in the `_version_` field is greater than '1' (i.e., '12345'), then the `_version_` in the document must match the `_version_` in the index.
- If the content in the `_version_` field is equal to '1', then the document must simply exist. In this case, no version matching occurs, but if the document does not exist, the updates will be rejected.
- If the content in the `_version_` field is less than '0' (i.e., '-1'), then the document must **not** exist. In this case, no version matching occurs, but if the document exists, the updates will be rejected.
- If the content in the `_version_` field is equal to '0', then it doesn't matter if the versions match or if the document exists or not. If it exists, it will be overwritten; if it does not exist, it will be added.

If the document being updated does not include the `_version_` field, and atomic updates are not being used, the document will be treated by normal Solr rules, which is usually to discard the previous version.

When using Optimistic Concurrency, clients can include an optional `versions=true` request parameter to indicate that the *new* versions of the documents being added should be included in the response. This allows clients to immediately know what the `_version_` is of every document added without needing to make a redundant `/get` request.

Following are some examples using versions=true in queries:

```
$ curl -X POST -H 'Content-Type: application/json'  
'http://localhost:8983/solr/techproducts/update?versions=true' --data-binary '  
[ { "id" : "aaa" },  
  { "id" : "bbb" } ]'
```

```
{"responseHeader": {"status": 0, "QTime": 6},  
 "adds": [ "aaa", 1498562471222312960,  
           "bbb", 1498562471225458688 ]}
```

In this example, we have added 2 documents "aaa" and "bbb". Because we added versions=true to the request, the response shows the document version for each document.

```
$ curl -X POST -H 'Content-Type: application/json'  
'http://localhost:8983/solr/techproducts/update?_version_=999999&versions=true' --data-binary '  
[ { "id" : "aaa",  
    "foo_s" : "update attempt with wrong existing version" } ]'
```

```
{"responseHeader": {"status": 409, "QTime": 3},  
 "error": {"msg": "version conflict for aaa expected=999999 actual=1498562471222312960",  
           "code": 409}}
```

In this example, we've attempted to update document "aaa" but specified the wrong version in the request: version=999999 doesn't match the document version we just got when we added the document. We get an error in response.

```
$ curl -X POST -H 'Content-Type: application/json'  
'http://localhost:8983/solr/techproducts/update?_version_=1498562471222312960&versions=true&committ=true' --data-binary '  
[ { "id" : "aaa",  
    "foo_s" : "update attempt with correct existing version" } ]'
```

```
{"responseHeader": {"status": 0, "QTime": 5},  
 "adds": [ "aaa", 1498562624496861184 ]}
```

Now we've sent an update with a value for _version_ that matches the value in the index, and it succeeds. Because we included versions=true to the update request, the response includes a different value for the _version_ field.

```
$ curl 'http://localhost:8983/solr/techproducts/query?q=*&fl=id,_version_'
```

```
{  
  "responseHeader":{  
    "status":0,  
    "QTime":5,  
    "params":{  
      "fl":"id,_version_",  
      "q":"*:*"}}},  
  "response":{ "numFound":2, "start":0, "docs": [  
    {  
      "id": "bbb",  
      "_version_":1498562471225458688},  
    {  
      "id": "aaa",  
      "_version_":1498562624496861184}]  
  ]}  
}
```

Finally, we can issue a query that requests the `_version_` field be included in the response, and we can see that for the two documents in our example index.

For more information, please also see Yonik Seeley's presentation on [NoSQL features in Solr 4](#) from Apache Lucene EuroCon 2012.

Document Centric Versioning Constraints

Optimistic Concurrency is extremely powerful, and works very efficiently because it uses an internally assigned, globally unique values for the `_version_` field. However, in some situations users may want to configure their own document specific version field, where the version values are assigned on a per-document basis by an external system, and have Solr reject updates that attempt to replace a document with an "older" version. In situations like this the `DocBasedVersionConstraintsProcessorFactory` can be useful.

The basic usage of `DocBasedVersionConstraintsProcessorFactory` is to configure it in `solrconfig.xml` as part of the `UpdateRequestProcessorChain` and specify the name of your custom `versionField` in your schema that should be checked when validating updates:

```
<processor class="solr.DocBasedVersionConstraintsProcessorFactory">  
  <str name="versionField">my_version_1</str>  
</processor>
```

Note that `versionField` is a comma delimited list of fields to check for version numbers. Once configured, this update processor will reject (HTTP error code 409) any attempt to update an existing document where the value of the `my_version_1` field in the "new" document is not greater than the value of that field in the existing document.

versionField vs _version_



The `_version_` field used by Solr for its normal optimistic concurrency also has important semantics in how updates are distributed to replicas in SolrCloud, and **MUST** be assigned internally by Solr. Users can not re-purpose that field and specify it as the `versionField` for use in the `DocBasedVersionConstraintsProcessorFactory` configuration.

`DocBasedVersionConstraintsProcessorFactory` supports the following additional configuration parameters, which are all optional:

`ignoreOldUpdates`

A boolean option which defaults to false. If set to true, the update will be silently ignored (and return a status 200 to the client) instead of rejecting updates where the `versionField` is too low.

`deleteVersionParam`

A String parameter that can be specified to indicate that this processor should also inspect Delete By Id commands.

The value of this option should be the name of a request parameter that the processor will consider mandatory for all attempts to Delete By Id, and must be used by clients to specify a value for the `versionField` which is greater than the existing value of the document to be deleted.

When using this request parameter, any Delete By Id command with a high enough document version number to succeed will be internally converted into an Add Document command that replaces the existing document with a new one which is empty except for the Unique Key and `versionField` to keep a record of the deleted version so future Add Document commands will fail if their "new" version is not high enough.

If `versionField` is specified as a list, then this parameter too must be specified as a comma delimited list of the same size so that the parameters correspond with the fields.

`supportMissingVersionOnOldDocs`

This boolean parameter defaults to false, but if set to true allows any documents written **before** this feature is enabled, and which are missing the `versionField`, to be overwritten.

Please consult the [DocBasedVersionConstraintsProcessorFactory javadocs](#) and [test solrconfig.xml file](#) for additional information and example usages.

Detecting Languages During Indexing

Solr can identify languages and map text to language-specific fields during indexing using the langid UpdateRequestProcessor.

Solr supports three implementations of this feature:

- Tika's language detection feature: <https://tika.apache.org/1.19.1/detection.html>
- LangDetect language detection: <https://github.com/shuyo/language-detection>
- OpenNLP language detection: <http://opennlp.apache.org/docs/1.9.1/manual/opennlp.html#tools.langdetect>

You can see a comparison between the Tika and LangDetect implementations here:

<http://blog.mikemccandless.com/2011/10/accuracy-and-performance-of-gogles.html>. In general, the LangDetect implementation supports more languages with higher performance.

For specific information on each of these language identification implementations, including a list of supported languages for each, see the relevant project websites.

For more information about language analysis in Solr, see [Language Analysis](#).

Configuring Language Detection

You can configure the langid UpdateRequestProcessor in solrconfig.xml. Both implementations take the same parameters, which are described in the following section. At a minimum, you must specify the fields for language identification and a field for the resulting language code.

Configuring Tika Language Detection

Here is an example of a minimal Tika langid configuration in solrconfig.xml:

```
<processor class="org.apache.solr.update.processor.TikaLanguageIdentifierUpdateProcessorFactory">
  <lst name="defaults">
    <str name="langid.fl">title,subject,text,keywords</str>
    <str name="langid.langField">language_s</str>
  </lst>
</processor>
```

Configuring LangDetect Language Detection

Here is an example of a minimal LangDetect langid configuration in solrconfig.xml:

```
<processor class="org.apache.solr.update.processor.LangDetectLanguageIdentifierUpdateProcessorFactory">
  <lst name="defaults">
    <str name="langid.fl">title,subject,text,keywords</str>
    <str name="langid.langField">language_s</str>
  </lst>
</processor>
```

Configuring OpenNLP Language Detection

Here is an example of a minimal OpenNLP langid configuration in solrconfig.xml:

```
<processor class="org.apache.solr.update.processor.OpenNLPLangDetectUpdateProcessorFactory">
  <lst name="defaults">
    <str name="langid.fl">title,subject,text,keywords</str>
    <str name="langid.langField">language_s</str>
    <str name="langid.model">langdetect-183.bin</str>
  </lst>
</processor>
```

OpenNLP-specific Parameters

langid.model

An OpenNLP language detection model. The OpenNLP project provides a pre-trained 103 language model on the [OpenNLP site's model download page](#). Model training instructions are provided on the [OpenNLP website](#). This parameter is required. See [Resource and Plugin Loading](#) for information on where to put the model.

OpenNLP Language Codes

OpenNLPLangDetectUpdateProcessor automatically converts the 3-letter ISO 639-3 codes detected by the OpenNLP model into 2-letter ISO 639-1 codes.

langid Parameters

As previously mentioned, both implementations of the langid UpdateRequestProcessor take the same parameters.

langid

When true, the default, enables language detection.

langid.fl

A comma- or space-delimited list of fields to be processed by langid. This parameter is required.

langid.langField

Specifies the field for the returned language code. This parameter is required.

langid.langsField

Specifies the field for a list of returned language codes. If you use langid.map.individual, each detected

language will be added to this field.

langid.overwrite

Specifies whether the content of the langField and langsField fields will be overwritten if they already contain values. The default is false.

langid.lcmap

A space-separated list specifying colon delimited language code mappings to apply to the detected languages.

For example, you might use this to map Chinese, Japanese, and Korean to a common cjk code, and map both American and British English to a single en code by using langid.lcmap=ja:cjk zh:cjk ko:cjk en_GB:en en_US:en.

This affects both the values put into the langField and langsField fields, as well as the field suffixes when using langid.map, unless overridden by langid.map.lcmap.

langid.threshold

Specifies a threshold value between 0 and 1 that the language identification score must reach before langid accepts it.

With longer text fields, a high threshold such as 0.8 will give good results. For shorter text fields, you may need to lower the threshold for language identification, though you will be risking somewhat lower quality results. We recommend experimenting with your data to tune your results.

The default is 0.5.

langid.whitelist

Specifies a list of allowed language identification codes. Use this in combination with langid.map to ensure that you only index documents into fields that are in your schema.

langid.map

Enables field name mapping. If true, Solr will map field names for all fields listed in langid.fl. The default is false.

langid.map.fl

A comma-separated list of fields for langid.map that is different than the fields specified in langid.fl.

langid.map.keepOrig

If true, Solr will copy the field during the field name mapping process, leaving the original field in place. The default is false.

langid.map.individual

If true, Solr will detect and map languages for each field individually. The default is false.

langid.map.individual.fl

A comma-separated list of fields for use with langid.map.individual that is different than the fields specified in langid.fl.

langid.fallback

Specifies a language code to use if no language is detected or specified in langid.fallbackFields.

langid.fallbackFields

If no language is detected that meets the langid.threshold score, or if the detected language is not on the langid.whitelist, this field specifies language codes to be used as fallback values.

If no appropriate fallback languages are found, Solr will use the language code specified in langid.fallback.

langid.map.lcmap

A space-separated list specifying colon-delimited language code mappings to use when mapping field names.

For example, you might use this to make Chinese, Japanese, and Korean language fields use a common *_cjk suffix, and map both American and British English fields to a single *_en by using langid.map.lcmap=ja:cjk zh:cjk ko:cjk en_GB:en en_US:en.

A list defined with this parameter will override any configuration set with langid.lcmap.

langid.map.pattern

By default, fields are mapped as <field>_<language>. To change this pattern, you can specify a Java regular expression in this parameter.

langid.map.replace

By default, fields are mapped as <field>_<language>. To change this pattern, you can specify a Java replace in this parameter.

langid.enforceSchema

If false, the langid processor does not validate field names against your schema. This may be useful if you plan to rename or delete fields later in the UpdateChain.

The default is true.

De-Duplication

If duplicate, or near-duplicate documents are a concern in your index, de-duplication may be worth implementing.

Preventing duplicate or near duplicate documents from entering an index or tagging documents with a signature/fingerprint for duplicate field collapsing can be efficiently achieved with a low collision or fuzzy hash algorithm. Solr natively supports de-duplication techniques of this type via the `Signature` class and allows for the easy addition of new hash/signature implementations. A `Signature` can be implemented in a few ways:

- `MD5Signature`: 128-bit hash used for exact duplicate detection.
- `Lookup3Signature`: 64-bit hash used for exact duplicate detection. This is much faster than MD5 and smaller to index.
- `TextProfileSignature`: Fuzzy hashing implementation from Apache Nutch for near duplicate detection. It's tunable but works best on longer text.

Other, more sophisticated algorithms for fuzzy/near hashing can be added later.

Adding in the de-duplication process will change the `allowDups` setting so that it applies to an update term (with `signatureField` in this case) rather than the unique field `Term`.

 Of course the `signatureField` could be the unique field, but generally you want the unique field to be unique. When a document is added, a signature will automatically be generated and attached to the document in the specified `signatureField`.

Configuration Options

There are two places in Solr to configure de-duplication: in `solrconfig.xml` and in `schema.xml`.

In `solrconfig.xml`

The `SignatureUpdateProcessorFactory` has to be registered in `solrconfig.xml` as part of an [Update Request Processor Chain](#), as in this example:

```
<updateRequestProcessorChain name="dedupe">
  <processor class="solr.processor.SignatureUpdateProcessorFactory">
    <bool name="enabled">true</bool>
    <str name="signatureField">id</str>
    <bool name="overwriteDups">false</bool>
    <str name="fields">name,features,cat</str>
    <str name="signatureClass">solr.processor.Lookup3Signature</str>
  </processor>
  <processor class="solr.LogUpdateProcessorFactory" />
  <processor class="solr.RunUpdateProcessorFactory" />
</updateRequestProcessorChain>
```

The `SignatureUpdateProcessorFactory` takes several properties:

signatureClass

A Signature implementation for generating a signature hash. The default is org.apache.solr.update.processor.Lookup3Signature.

The full classpath of the implementation must be specified. The available options are described above, the associated classpaths to use are:

- org.apache.solr.update.processor.Lookup3Signature
- org.apache.solr.update.processor.MD5Signature
- org.apache.solr.update.processor.TextProfileSignature

fields

The fields to use to generate the signature hash in a comma separated list. By default, all fields on the document will be used.

signatureField

The name of the field used to hold the fingerprint/signature. The field should be defined in schema.xml. The default is signatureField.

enabled

Set to **false** to disable de-duplication processing. The default is **true**.

overwriteDuplicates

If true, the default, when a document exists that already matches this signature, it will be overwritten.

In schema.xml

If you are using a separate field for storing the signature, you must have it indexed:

```
<field name="signatureField" type="string" stored="true" indexed="true" multiValued="false" />
```

Be sure to change your update handlers to use the defined chain, as below:

```
<requestHandler name="/update" class="solr.UpdateRequestHandler" >
  <lst name="defaults">
    <str name="update.chain">dedupe</str>
  </lst>
  ...
</requestHandler>
```

This example assumes you have other sections of your request handler defined.



The update processor can also be specified per request with a parameter of update.chain=dedupe.

Content Streams

Content streams are bulk data passed with a request to Solr.

When Solr RequestHandlers are accessed using path based URLs, the SolrQueryRequest object containing the parameters of the request may also contain a list of ContentStreams containing bulk data for the request. (The name SolrQueryRequest is a bit misleading: it is involved in all requests, regardless of whether it is a query request or an update request.)

Content Stream Sources

Currently request handlers can get content streams in a variety of ways:

- For multipart file uploads, each file is passed as a stream.
- For POST requests where the content-type is not application/x-www-form-urlencoded, the raw POST body is passed as a stream. The full POST body is parsed as parameters and included in the Solr parameters.
- The contents of parameter stream.body is passed as a stream.
- If remote streaming is enabled and URL content is called for during request handling, the contents of each stream.url and stream.file parameters are fetched and passed as a stream.

By default, curl sends a contentType="application/x-www-form-urlencoded" header. If you need to test a SolrContentHeader content stream, you will need to set the content type with curl's -H flag.

Remote Streaming

Remote streaming lets you send the contents of a URL as a stream to a given Solr RequestHandler. You could use remote streaming to send a remote or local file to an update plugin.

Remote streaming is disabled by default. Enabling it is not recommended in a production situation without additional security between you and untrusted remote clients.

In solrconfig.xml, you can enable it by changing the following enableRemoteStreaming parameter to true:

```
*** WARNING ***
Before enabling remote streaming, you should make sure your
system has authentication enabled.
```

```
<requestParsers enableRemoteStreaming="false" />
```

When enableRemoteStreaming is not specified in solrconfig.xml, the default behavior is to *not* allow remote streaming (i.e., enableRemoteStreaming="false").

Remote streaming can also be enabled through the [Config API](#) as follows:

V1 API

```
curl -H 'Content-type:application/json' -d '{"set-property": {"requestDispatcher.requestParsers.enableRemoteStreaming":true}}' 'http://localhost:8983/solr/techproducts/config'
```

V2 API

```
curl -X POST -H 'Content-type: application/json' -d '{"set-property": {"requestDispatcher.requestParsers.enableRemoteStreaming":true}}' 'http://localhost:8983/api/collections/techproducts/config'
```



If `enableRemoteStreaming="true"` is used, be aware that this allows *anyone* to send a request to any URL or local file. If the [DumpRequestHandler](#) is enabled, it will allow anyone to view any file on your system.

The source of the data can be compressed using gzip, and Solr will generally detect this. The detection is based on either the presence of a Content-Encoding: gzip HTTP header or the file ending with .gz or .gzip. Gzip doesn't apply to stream.body.

Debugging Requests

The implicit "dump" RequestHandler (see [Implicit RequestHandlers](#)) simply outputs the contents of the Solr QueryRequest using the specified writer type wt. This is a useful tool to help understand what streams are available to the RequestHandlers.

Reindexing

There are several types of changes to Solr configuration that require you to reindex your data.

These changes include editing properties of fields or field types; adding fields, field types, or copy field rules; upgrading Solr; and some system configuration properties.

It's important to be aware that many changes require reindexing, because there are times when not reindexing can have negative consequences for Solr as a system, or for the ability of your users to find what they are looking for.

There is no process in Solr for programmatically reindexing data. When we say "reindex", we mean, literally, "index it again". However you got the data into the index the first time, you will run that process again. It is strongly recommended that Solr users index their data in a repeatable, consistent way, so that the process can be easily repeated when the need for reindexing arises.

Reindexing is recommended during major upgrades, so in addition to covering what types of configuration changes should trigger a reindex, this section will also cover strategies for reindexing.

Changes that Require Reindex

Schema Changes

All changes to a collection's schema require reindexing. This is because many of the available options are only applied during the indexing process. Solr simply has no way to implement the desired change without reindexing the data.

To understand the general reason why reindexing is ever required, it's helpful to understand the relationship between Solr's schema and the underlying Lucene index. Lucene does not use a schema, it is a Solr-only concept. When you delete a field from Solr's schema, it does not modify Lucene's index in any way. When you add a field to Solr's schema, the field does not exist in Lucene's index until a document that contains the field is indexed.

This means that there are many types of schema changes that cannot be reflected in the index simply by modifying Solr's schema. This is different from most database models where schemas are used. With regard to indexing, Solr's schema acts like a rulebook for indexing documents by telling Lucene how to interpret the data being sent. Once the documents are in Lucene, Solr's schema has no control over the underlying data structure.

In addition to the types of schema changes described in the following sections, changing the schema version property is equivalent to changing field type properties. This type of change is usually only made during or because of a major upgrade.

Adding or Deleting Fields

If you add or delete a field from Solr's schema, it's strongly recommended to reindex.

When you add a field, you generally do so with the intent to use the field in some way. Since documents were indexed before the field was added, the index will not hold any references to the field for earlier documents. If you want to use the new field for facetting, for example, the new field facet will not include any

documents that were not indexed with the new field.

There is a slightly different situation when deleting a field. In this case, since simply removing the field from the schema doesn't change anything about the index, the field will still be in the index until the documents are reindexed. In fact, Lucene may keep a reference to a deleted field *forever* (see also [LUCENE-1761](#)). This may only be an issue for your environment if you try to add a field that has the same name as a deleted field, but it can also be an issue for dynamic field rules that are later removed.

Changing Field and Field Type Properties

Solr has two ways of defining field properties.

The first is to define properties on a field type. These properties are then applied to all fields of that type unless they are explicitly overridden.

The second is an override to a property inherited from the field type defined on the field itself.

If a property has been defined for a field type but the property is not overridden by defining a different value for the property for a field, then changing the property on the field type is equivalent to changing it on the field itself.

Changes to **any** field/field type property described in [Field Type Properties](#) must be reindexed in order for the change to be reflected in all documents. The list of changes that require reindexing includes (but is not limited to):

- Changing a field from stored to not stored, and vice versa.
- Changing a field from indexed to not indexed, and vice versa.
- Changing a field from multi-valued to single-valued, and vice versa.
- [Changing Field Analysis](#).
- Changing the type of field, or the class for a field type.
- Enabling or disabling [docValues](#).

Be sure to reference the Field Type Properties section linked above for the complete list of properties that would require a reindex.

In some cases, it can be possible to change a field/field type property value and it will only apply to documents indexed *after* the change.

For example, you could change a field from being indexed (`indexed="true"`) to no longer indexed (`indexed="false"`) and over time, as documents are updated, the index will be purged of the fields that shouldn't be indexed anymore.

You could also change a field from not being stored (`stored="false"`) to being stored (`stored="true"`). In this case, if you want to use the field immediately, only documents indexed after the change will contain data in the field. However, you would need to ensure that your client is able to handle fields missing from documents that have not yet been reindexed.

It's important to note this is not possible for all field/field type properties. If you change whether or not `docValues` are enabled, for example, you absolutely must reindex. This is due to the way `docValues` have been implemented in Lucene, and how Lucene handles `docValue` segments.

Changing any field properties without reindexing is *never* recommended to ensure consistent behavior, and should only be attempted when you have tested thoroughly and feel confident that you understand the ramifications on your documents and front-end clients.

Changing Field Analysis

Beyond specific field-level properties, [analysis chains](#) are also configured on field types, and are applied at index and/or query time.

It's possible to define separate analysis chains for indexing and query events, or you can define a single chain that is applied to both event types.

If you change the analysis chain that applies to indexing events, it is strongly recommended that you reindex. This is because all of the changes that occur due to the chain configuration are applied to documents as they are being indexed, and only reindexing will allow your changes to take effect on documents.

While reindexing after analyzer changes is not required, be aware that not reindexing can cause unexpected query results in many cases.

For example, if you indexed a number of documents and then decide you'd like to use the `LowerCaseTokenizerFactory` to ensure all text is converted to lower case, you will have a mix of entries in the field: some in their original case ("iPhone"), and newer documents in all lower-case ("iphone"). If you do not reindex the original set of documents, a query such as "iphone" will not match documents with "iPhone", because the schema rules enforce lower case on the query, but that's not what is in the index.

The only time you do not have to reindex when changing a field type's analysis chain is when the changes impact queries **only** (and you know that you do not need to make corresponding changes to the index analysis).

Solrconfig Changes

Only one parameter change to Solr's solrconfig.xml requires reindexing. That parameter is the luceneMatchVersion, which controls the compatibility of Solr with Lucene changes. Since this parameter can change the rules for analysis behind the scenes, it's always recommended to reindex when changing this value. Usually, however, this is only changed in conjunction with a major upgrade.

However, if you make a change to Solr's [Update Request Processors](#), it's generally because you want to change something about how *update requests* (documents) are *processed* (indexed). In this case, you can decide based on the change if you want to reindex your documents to implement the changes you've made.

Similarly, if you change the codecFactory parameter in solrconfig.xml, it is again strongly recommended that you plan to reindex your documents to avoid unintended behavior.

Upgrades

When upgrading between major versions (for example, from a 7.x release to 8.0 or 8.x), a best practice is to always reindex your data. The reason for this is that subtle changes may occur in default field type definitions or the underlying code.



If you have **not** changed your schema as part of an upgrade from one minor release to another (such as, from 7.x to a later 7.x release), you can often skip reindexing your documents. However, when upgrading to a major release, you should plan to reindex your documents because of the likelihood of changes that break back-compatibility.

Reindexing Strategies

There are a few approaches available to perform the reindex.

The strategies described below ensure that the Lucene index is completely dropped so you can recreate it to accommodate your changes. They allow you to recreate the Lucene index without having Lucene segments lingering with stale data.

Delete All Documents

The best approach is to first delete everything from the index, and then index your data again. You can delete all documents with a "delete-by-query", such as this:

```
curl -X POST -H 'Content-Type: application/json' --data-binary '{"delete": {"query": "*:*" }}'  
http://localhost:8983/solr/my_collection/update
```

It's important to verify that **all** documents have been deleted, as that ensures the Lucene index segments have been deleted as well.

To verify that there are no segments in your index, look in the data directory and confirm it is empty. Since the data directory can be customized, see the section [Specifying a Location for Index Data with the dataDir Parameter](#) for where to look to find the index files.

Note you will need to verify the indexes have been removed in every shard and every replica on every node

of a cluster. It is not sufficient to only query for the number of documents because you may have no documents but still have index segments.

Once the indexes have been cleared, you can start reindexing by re-running the original index process.

Index to Another Collection

In cases where you cannot take a production collection offline to delete all the documents, one option is to use Solr's [collection alias](#) feature.

This option is only available for Solr installations running in SolrCloud mode.

With this approach, you will index your documents into a newly created collection and once everything is completed, create an alias for the collection and point your front-end at the collection alias. Queries will be routed to the new collection seamlessly.

Here is an example of creating an alias that points to a single collection:

```
http://localhost:8983/solr/admin/collections?action=CREATEALIAS&name=myData&collections=newCollection
```

Once the alias is in place and you are satisfied you no longer need the old data, you can delete the old collection with the [DELETE command](#) of the Collections API:

```
http://localhost:8983/solr/admin/collections?action=DELETE&name=oldCollection
```

Changes that Do Not Require Reindex

The types of changes that do not require or strongly indicate reindexing are changes that do not impact the index.

Creating or modifying request handlers, search components, and other elements of `solrconfig.xml` don't require reindexing.

Cluster and core management actions, such as adding nodes, replicas, or new cores, or splitting shards, also don't require reindexing.

Searching

This section describes how Solr works with search requests. It covers the following topics:

- [Overview of Searching in Solr](#): An introduction to searching with Solr.
- [Velocity Search UI](#): A simple search UI using the VelocityResponseWriter.
- [Relevance](#): Conceptual information about understanding relevance in search results.
- [Query Syntax and Parsing](#): A brief conceptual overview of query syntax and parsing. It also contains the following sub-sections:
 - [Common Query Parameters](#): No matter the query parser, there are several parameters that are common to all of them.
 - [The Standard Query Parser](#): Detailed information about the standard Lucene query parser.
 - [The DisMax Query Parser](#): Detailed information about Solr's DisMax query parser.
 - [The Extended DisMax Query Parser](#): Detailed information about Solr's Extended DisMax (eDisMax) Query Parser.
 - [Function Queries](#): Detailed information about parameters for generating relevancy scores using values from one or more numeric fields.
 - [Local Parameters in Queries](#): How to add local arguments to queries.
 - [Other Parsers](#): More parsers designed for use in specific situations.
- [JSON Request API](#): Overview of Solr's JSON Request API.
 - [JSON Query DSL](#): Detailed information about a simple yet powerful query language for JSON Request API.
- [JSON Facet API](#): Overview of Solr's JSON Facet API.
- [Faceting](#): Detailed information about categorizing search results based on indexed terms.
- [Highlighting](#): Detailed information about Solr's highlighting capabilities, including multiple underlying highlighter implementations.
- [Spell Checking](#): Detailed information about Solr's spelling checker.
- [Query Re-Ranking](#): Detailed information about re-ranking top scoring documents from simple queries using more complex scores.
 - [Learning To Rank](#): How to use LTR to run machine learned ranking models in Solr.
- [Transforming Result Documents](#): Detailed information about using DocTransformers to add computed information to individual documents
- [Searching Nested Documents](#): Detailed information about constructing nested and hierarchical queries.
- [Suggester](#): Detailed information about Solr's powerful autosuggest component.
- [MoreLikeThis](#): Detailed information about Solr's similar results query component.
- [Pagination of Results](#): Detailed information about fetching paginated results for display in a UI, or for fetching all documents matching a query.
- [Result Grouping](#): Detailed information about grouping results based on common field values.
- [Result Clustering](#): Detailed information about grouping search results based on cluster analysis applied

to text fields. A bit like "unsupervised" faceting.

- [Spatial Search](#): How to use Solr's spatial search capabilities.
- [The Terms Component](#): Detailed information about accessing indexed terms and the documents that include them.
- [The Term Vector Component](#): How to get term information about specific documents.
- [The Stats Component](#): How to return information from numeric fields within a document set.
- [The Query Elevation Component](#): How to force documents to the top of the results for certain queries.
- [The Tagger Handler](#): The SolrTextTagger, for basic named entity tagging in text.
- [Response Writers](#): Detailed information about configuring and using Solr's response writers.
- [Near Real Time Searching](#): How to include documents in search results nearly immediately after they are indexed.
- [RealTime Get](#): How to get the latest version of a document without opening a searcher.
- [Exporting Result Sets](#): Functionality to export large result sets out of Solr.
- [Parallel SQL Interface](#): An interface for sending SQL statements to Solr, and using advanced parallel query processing and relational algebra for complex data analysis.
- [The Analytics Component](#): A framework to compute complex analytics over a result set.

Overview of Searching in Solr

Solr offers a rich, flexible set of features for search. To understand the extent of this flexibility, it's helpful to begin with an overview of the steps and components involved in a Solr search.

When a user runs a search in Solr, the search query is processed by a **request handler**. A request handler is a Solr plug-in that defines the logic to be used when Solr processes a request. Solr supports a variety of request handlers. Some are designed for processing search queries, while others manage tasks such as index replication.

Search applications select a particular request handler by default. In addition, applications can be configured to allow users to override the default selection in preference of a different request handler.

To process a search query, a request handler calls a **query parser**, which interprets the terms and parameters of a query. Different query parsers support different syntax. Solr's default query parser is known as the [Standard Query Parser](#), or more commonly just the "lucene" query parser. Solr also includes the [DisMaxquery](#) parser, and the [Extended DisMax](#) (eDisMax) query parser. The [standard](#) query parser's syntax allows for greater precision in searches, but the DisMax query parser is much more tolerant of errors. The DisMax query parser is designed to provide an experience similar to that of popular search engines such as Google, which rarely display syntax errors to users. The Extended DisMax query parser is an improved version of DisMax that handles the full Lucene query syntax while still tolerating syntax errors. It also includes several additional features.

In addition, there are [common query parameters](#) that are accepted by all query parsers.

Input to a query parser can include:

- search strings---that is, *terms* to search for in the index
- *parameters for fine-tuning the query* by increasing the importance of particular strings or fields, by applying Boolean logic among the search terms, or by excluding content from the search results
- *parameters for controlling the presentation of the query response*, such as specifying the order in which results are to be presented or limiting the response to particular fields of the search application's schema.

Search parameters may also specify a **filter query**. As part of a search response, a filter query runs a query against the entire index and caches the results. Because Solr allocates a separate cache for filter queries, the strategic use of filter queries can improve search performance. (Despite their similar names, query filters are not related to analysis filters. Filter queries perform queries at search time against data already in the index, while analysis filters, such as Tokenizers, parse content for indexing, following specified rules).

A search query can request that certain terms be highlighted in the search response; that is, the selected terms will be displayed in colored boxes so that they "jump out" on the screen of search results.

[Highlighting](#) can make it easier to find relevant passages in long documents returned in a search. Solr supports multi-term highlighting. Solr includes a rich set of search parameters for controlling how terms are highlighted.

Search responses can also be configured to include **snippets** (document excerpts) featuring highlighted text. Popular search engines such as Google and Yahoo! return snippets in their search results: 3-4 lines of text offering a description of a search result.

To help users zero in on the content they're looking for, Solr supports two special ways of grouping search results to aid further exploration: faceting and clustering.

Faceting is the arrangement of search results into categories (which are based on indexed terms). Within each category, Solr reports on the number of hits for relevant term, which is called a facet constraint. Faceting makes it easy for users to explore search results on sites such as movie sites and product review sites, where there are many categories and many items within a category.

The screen shot below shows an example of faceting from the CNET Web site (CBS Interactive Inc.), which was the first site to use Solr.

The screenshot shows a search interface for 'Digital cameras'. At the top, a header says 'Digital cameras'. Below it, a section titled 'Refine your results' contains three facets: 'Manufacturer', 'Resolution', and 'Zoom range'. Each facet has a list of values and their counts. To the right of these facets is a 'More' section with additional filter options like 'LCD size', 'Image stabilizer', etc. A callout box points to the 'Manufacturer' facet with the text: 'Manufacturer is a facet, a way of categorizing the results'. Another callout box points to the 'More' section with the text: 'The facet count or constraint count shows how many results match each value'. At the bottom of the refine section, there's a 'you selected:' summary showing filters applied: '\$400 - \$500', 'SLR', and 'remove all'. Below this is a 'Regular search results list' showing 17 results. The first result is 'Canon EOS Rebel XS (silver, with 18-55mm lens)'. Callouts point to the breadcrumb trail ('you selected:'), the 'Regular search results list' heading, and the result card itself. The breadcrumb trail shows constraints applied: '\$400 - \$500', 'SLR'. The result card includes a camera icon, the product name, price (\$459 to \$699), and availability at 15 stores. Navigation links '1 2 next' are also visible.

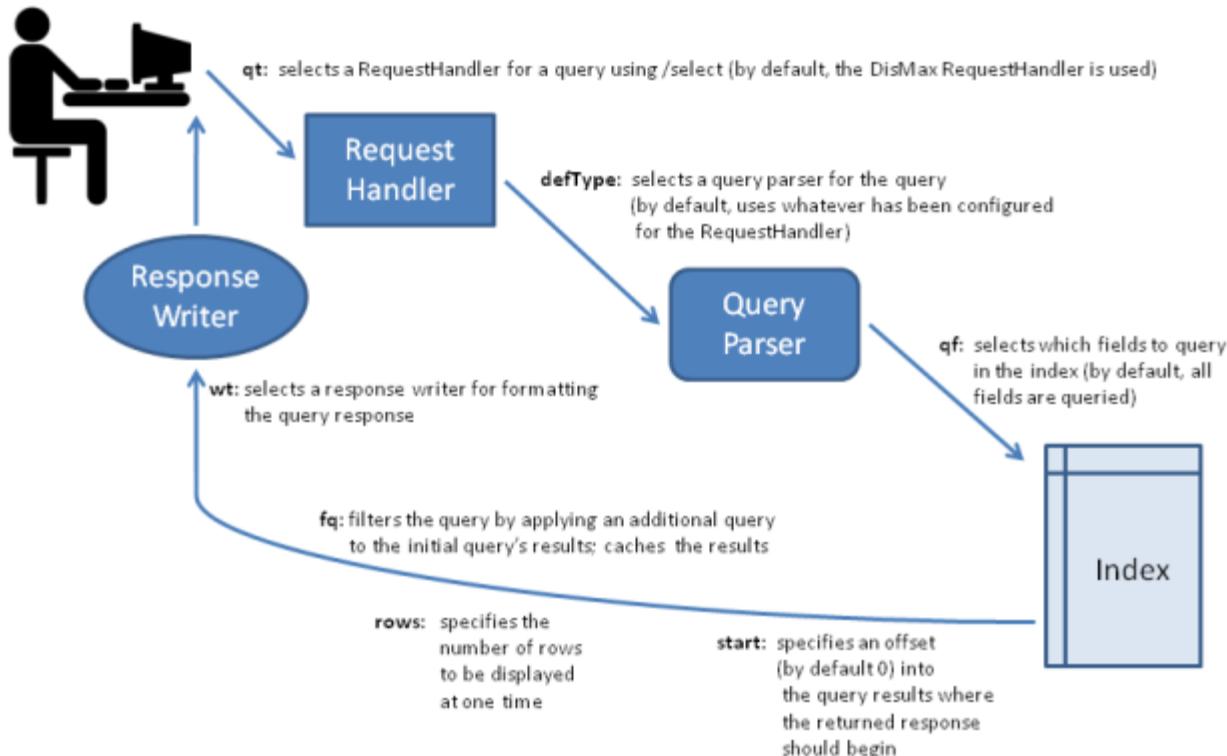
Faceting makes use of fields defined when the search applications were indexed. In the example above, these fields include categories of information that are useful for describing digital cameras: manufacturer, resolution, and zoom range.

Clustering groups search results by similarities discovered when a search is executed, rather than when content is indexed. The results of clustering often lack the neat hierarchical organization found in faceted search results, but clustering can be useful nonetheless. It can reveal unexpected commonalities among search results, and it can help users rule out content that isn't pertinent to what they're really searching for.

Solr also supports a feature called **MoreLikeThis**, which enables users to submit new queries that focus on particular terms returned in an earlier query. MoreLikeThis queries can make use of faceting or clustering to provide additional aid to users.

A Solr component called a **response writer** manages the final presentation of the query response. Solr includes a variety of response writers, including an [XML Response Writer](#) and a [JSON Response Writer](#).

The diagram below summarizes some key elements of the search process.



Velocity Search UI

Solr includes a sample search UI based on the `VelocityResponseWriter` (also known as Solritas) that demonstrates several useful features, such as searching, facetting, highlighting, autocomplete, and geospatial searching.

When using the `sample_techproducts_configs` configset, you can access the Velocity sample Search UI: <http://localhost:8983/solr/techproducts/browse>

The screenshot shows the Velocity Search UI interface. At the top, there's a navigation bar with the Solr logo and tabs for Simple, Spatial, and Group By. Below that is a search bar with a 'Find:' input field, a 'Submit' button, and a 'Reset' button. There's also a checkbox for 'Boost by Price'. On the left, a 'Field Facets' sidebar lists categories like cat, manu_exact, and manu. The main area displays search results for three products:

- Test with some GB18030 encoded characters**
More Like This
Id: GB18030TEST
Price: 0.0,USD
Features: No accents here ... 这是一个功能 ... This is a feature (translated) ... 这份文件是很有光泽 ... This document is very shiny (translated)
In Stock: true
- Samsung SpinPoint P120 SP2514N - hard drive - 250 GB - ATA-133**
More Like This
Id: SP2514N
Price: 92.0,USD
Features: 7200RPM, 8MB cache, IDE Ultra ATA-133 ... NoiseGuard, SilentSeek technology, Fluid Dynamic Bearing (FDB) motor
In Stock: true
- Maxtor DiamondMax 11 - hard drive - 500 GB - SATA-300**
More Like This
Id: 6H500F0
Price: 350.0,USD
Features: SATA 3.0Gb/s, NCQ ... 8.5ms seek ... 16MB cache
In Stock: true

Each result card includes a 'Google Map' link and a 'Larger Map' link.

The Velocity Search UI

For more information about the Velocity Response Writer, see the [Response Writer page](#).

Relevance

Relevance is the degree to which a query response satisfies a user who is searching for information.

The relevance of a query response depends on the context in which the query was performed. A single search application may be used in different contexts by users with different needs and expectations. For example, a search engine of climate data might be used by a university researcher studying long-term climate trends, a farmer interested in calculating the likely date of the last frost of spring, a civil engineer interested in rainfall patterns and the frequency of floods, and a college student planning a vacation to a region and wondering what to pack. Because the motivations of these users vary, the relevance of any particular response to a query will vary as well.

How comprehensive should query responses be? Like relevance in general, the answer to this question depends on the context of a search. The cost of *not* finding a particular document in response to a query is high in some contexts, such as a legal e-discovery search in response to a subpoena, and quite low in others, such as a search for a cake recipe on a Web site with dozens or hundreds of cake recipes. When configuring Solr, you should weigh comprehensiveness against other factors such as timeliness and ease-of-use.

The e-discovery and recipe examples demonstrate the importance of two concepts related to relevance:

- **Precision** is the percentage of documents in the returned results that are relevant.
- **Recall** is the percentage of relevant results returned out of all relevant results in the system. Obtaining perfect recall is trivial: simply return every document in the collection for every query.

Returning to the examples above, it's important for an e-discovery search application to have 100% recall returning all the documents that are relevant to a subpoena. It's far less important that a recipe application offer this degree of precision, however. In some cases, returning too many results in casual contexts could overwhelm users. In some contexts, returning fewer results that have a higher likelihood of relevance may be the best approach.

Using the concepts of precision and recall, it's possible to quantify relevance across users and queries for a collection of documents. A perfect system would have 100% precision and 100% recall for every user and every query. In other words, it would retrieve all the relevant documents and nothing else. In practical terms, when talking about precision and recall in real systems, it is common to focus on precision and recall at a certain number of results, the most common (and useful) being ten results.

Through faceting, query filters, and other search components, a Solr application can be configured with the flexibility to help users fine-tune their searches in order to return the most relevant results for users. That is, Solr can be configured to balance precision and recall to meet the needs of a particular user community.

The configuration of a Solr application should take into account:

- the needs of the application's various users (which can include ease of use and speed of response, in addition to strictly informational needs)
- the categories that are meaningful to these users in their various contexts (e.g., dates, product categories, or regions)
- any inherent relevance of documents (e.g., it might make sense to ensure that an official product description or FAQ is always returned near the top of the search results)
- whether or not the age of documents matters significantly (in some contexts, the most recent

documents might always be the most important)

Keeping all these factors in mind, it's often helpful in the planning stages of a Solr deployment to sketch out the types of responses you think the search application should return for sample queries. Once the application is up and running, you can employ a series of testing methodologies, such as focus groups, in-house testing, [TREC](#) tests and A/B testing to fine tune the configuration of the application to best meet the needs of its users.

For more information about relevance, see Grant Ingersoll's tech article [Debugging Search Application Relevance Issues](#) which is available on SearchHub.org.

Query Syntax and Parsing

Solr supports several query parsers, offering search application designers great flexibility in controlling how queries are parsed.

This section explains how to specify the query parser to be used. It also describes the syntax and features supported by the main query parsers included with Solr and describes some other parsers that may be useful for particular situations. There are some query parameters common to all Solr parsers; these are discussed in the section [Common Query Parameters](#).

The parsers discussed in this Guide are:

- [The Standard Query Parser](#)
- [The DisMax Query Parser](#)
- [The Extended DisMax Query Parser](#)
- [Other Parsers](#)

The query parser plugins are all subclasses of [QParserPlugin](#). If you have custom parsing needs, you may want to extend that class to create your own query parser.

Common Query Parameters

Several query parsers share supported query parameters.

The following sections describe Solr's common query parameters, which are supported by the [Search RequestHandlers](#).

defType Parameter

The defType parameter selects the query parser that Solr should use to process the main query parameter (q) in the request. For example:

```
defType=dismax
```

If no defType parameter is specified, then by default, the [The Standard Query Parser](#) is used. (e.g., defType=lucene)

sort Parameter

The sort parameter arranges search results in either ascending (asc) or descending (desc) order. The parameter can be used with either numerical or alphabetical content. The directions can be entered in either all lowercase or all uppercase letters (i.e., both asc and ASC are accepted).

Solr can sort query responses according to:

- Document scores
- [Function results](#)
- The value of any primitive field (numerics, string, boolean, dates, etc.) which has docValues="true" (or multiValued="false" and indexed="true", in which case the indexed terms will be used to build DocValue

like structures on the fly at runtime)

- A SortableTextField which implicitly uses docValues="true" by default to allow sorting on the original input string regardless of the analyzers used for Searching.
- A single-valued TextField that uses an analyzer (such as the KeywordTokenizer) that produces only a single term per document. TextField does not support docValues="true", but a DocValue-like structure will be built on the fly at runtime.
 - **NOTE:** If you want to be able to sort on a field whose contents you want to tokenize to facilitate searching, [use a copyField directive](#) in the Schema to clone the field. Then search on the field and sort on its clone.

In the case of primitive fields, or SortableTextFields, that are multiValued="true" the representative value used for each doc when sorting depends on the sort direction: The minimum value in each document is used for ascending (asc) sorting, while the maximal value in each document is used for descending (desc) sorting. This default behavior is equivalent to explicitly sorting using the 2 argument [field\(\)](#) function:
`sort=field(name,min) asc` and `sort=field(name,max) desc`

The table below explains how Solr responds to various settings of the sort parameter.

Example	Result
	If the sort parameter is omitted, sorting is performed as though the parameter were set to score desc.
score desc	Sorts in descending order from the highest score to the lowest score.
price asc	Sorts in ascending order of the price field
div(popularity,price) desc	Sorts in descending order of the result of the function popularity / price
inStock desc, price asc	Sorts by the contents of the inStock field in descending order, then when multiple documents have the same value for the inStock field, those results are sorted in ascending order by the contents of the price field.
categories asc, price asc	Sorts by the lowest value of the (multivalued) categories field in ascending order, then when multiple documents have the same lowest categories value, those results are sorted in ascending order by the contents of the price field.

Regarding the sort parameter's arguments:

- A sort ordering must include a field name (or score as a pseudo field), followed by whitespace (escaped as + or %20 in URL strings), followed by a sort direction (asc or desc).
- Multiple sort orderings can be separated by a comma, using this syntax: `sort=<field name><direction>,<field name><direction>],...`
 - When more than one sort criteria is provided, the second entry will only be used if the first entry results in a tie. If there is a third entry, it will only be used if the first AND second entries are tied. This pattern continues with further entries.

start Parameter

When specified, the start parameter specifies an offset into a query's result set and instructs Solr to begin displaying results from this offset.

The default value is 0. In other words, by default, Solr returns results without an offset, beginning where the results themselves begin.

Setting the start parameter to some other number, such as 3, causes Solr to skip over the preceding records and start at the document identified by the offset.

You can use the start parameter this way for paging. For example, if the rows parameter is set to 10, you could display three successive pages of results by setting start to 0, then re-issuing the same query and setting start to 10, then issuing the query again and setting start to 20.

rows Parameter

You can use the rows parameter to paginate results from a query. The parameter specifies the maximum number of documents from the complete result set that Solr should return to the client at one time.

The default value is 10. That is, by default, Solr returns 10 documents at a time in response to a query.

fq (Filter Query) Parameter

The fq parameter defines a query that can be used to restrict the superset of documents that can be returned, without influencing score. It can be very useful for speeding up complex queries, since the queries specified with fq are cached independently of the main query. When a later query uses the same filter, there's a cache hit, and filter results are returned quickly from the cache.

When using the fq parameter, keep in mind the following:

- The fq parameter can be specified multiple times in a query. Documents will only be included in the result if they are in the intersection of the document sets resulting from each instance of the parameter. In the example below, only documents which have a popularity greater than 10 and have a section of 0 will match.

```
fq=popularity:[10 TO *]&fq=section:0
```

- Filter queries can involve complicated Boolean queries. The above example could also be written as a single fq with two mandatory clauses like so:

```
fq+=popularity:[10 TO *] +section:0
```

- The document sets from each filter query are cached independently. Thus, concerning the previous examples: use a single fq containing two mandatory clauses if those clauses appear together often, and use two separate fq parameters if they are relatively independent. (To learn about tuning cache sizes and making sure a filter cache actually exists, see [The Well-Configured Solr Instance](#).)
- It is also possible to use [filter\(condition\) syntax](#) inside the fq to cache clauses individually and - among

other things - to achieve union of cached filter queries.

- As with all parameters: special characters in an URL need to be properly escaped and encoded as hex values. Online tools are available to help you with URL-encoding. For example: <http://meyerweb.com/eric/tools/dencoder/>.

fl (Field List) Parameter

The fl parameter limits the information included in a query response to a specified list of fields. The fields must be either stored="true" or docValues="true".

The field list can be specified as a space-separated or comma-separated list of field names. The string "score" can be used to indicate that the score of each document for the particular query should be returned as a field. The wildcard character * selects all the fields in the document which are either stored="true" or docValues="true" and useDocValuesAsStored="true" (which is the default when docValues are enabled). You can also add pseudo-fields, functions and transformers to the field list request.

This table shows some basic examples of how to use fl:

Field List	Result
id name price	Return only the id, name, and price fields.
id,name,price	Return only the id, name, and price fields.
id name, price	Return only the id, name, and price fields.
id score	Return the id field and the score.
*	Return all the stored fields in each document, as well as any docValues fields that have useDocValuesAsStored="true". This is the default value of the fl parameter.
* score	Return all the fields in each document, along with each field's score.
* ,dv_field_name	Return all the stored fields in each document, and any docValues fields that have useDocValuesAsStored="true" and the docValues from dv_field_name even if it has useDocValuesAsStored="false"

Functions with fl

Functions can be computed for each document in the result and returned as a pseudo-field:

```
fl=id,title,product(price,popularity)
```

Document Transformers with fl

Document Transformers can be used to modify the information returned about each documents in the results of a query:

```
fl=id,title,[explain]
```

Field Name Aliases

You can change the key used to in the response for a field, function, or transformer by prefixing it with a ` "displayName: ` ". For example:

```
fl=id,sales_price:price,secret_sauce:prod(price,popularity),why_score:[explain style=nl]
```

```
{
  "response": {
    "numFound": 2,
    "start": 0,
    "docs": [
      {
        "id": "6H500F0",
        "secret_sauce": 2100.0,
        "sales_price": 350.0,
        "why_score": {
          "match": true,
          "value": 1.052226,
          "description": "weight(features:cache in 2) [DefaultSimilarity], result of:",
          "details": [
            ...
          ]
        }
      }
    ]
  }
}
```

debug Parameter

The debug parameter can be specified multiple times and supports the following arguments:

- debug=query: return debug information about the query only.
- debug=timing: return debug information about how long the query took to process.
- debug=results: return debug information about the score results (also known as "explain").
 - By default, score explanations are returned as large string values, using newlines and tab indenting for structure & readability, but an additional debug.explain.structured=true parameter may be specified to return this information as nested data structures native to the response format requested by wt.
- debug=all: return all available debug information about the request request. (alternatively usage: debug=true)

For backwards compatibility with older versions of Solr, debugQuery=true may instead be specified as an alternative way to indicate debug=all

The default behavior is not to include debugging information.

explainOther Parameter

The explainOther parameter specifies a Lucene query in order to identify a set of documents. If this parameter is included and is set to a non-blank value, the query will return debugging information, along with the "explain info" of each document that matches the Lucene query, relative to the main query (which

is specified by the `q` parameter). For example:

```
q=supervillians&debugQuery=on&explainOther=id:juggernaut
```

The query above allows you to examine the scoring explain info of the top matching documents, compare it to the explain info for documents matching `id:juggernaut`, and determine why the rankings are not as you expect.

The default value of this parameter is blank, which causes no extra "explain info" to be returned.

timeAllowed Parameter

This parameter specifies the amount of time, in milliseconds, allowed for a search to complete. If this time expires before the search is complete, any partial results will be returned, but values such as `numFound`, `facet` counts, and result `stats` may not be accurate for the entire result set.

This value is only checked at the time of:

1. Query Expansion, and
2. Document collection

As this check is periodically performed, the actual time for which a request can be processed before it is aborted would be marginally greater than or equal to the value of `timeAllowed`. If the request consumes more time in other stages, custom components, etc., this parameter is not expected to abort the request.

segmentTerminateEarly Parameter

This parameter may be set to either true or false.

If set to true, and if `the mergePolicyFactory` for this collection is a `SortingMergePolicyFactory` which uses a sort option compatible with `the sort parameter` specified for this query, then Solr will be able to skip documents on a per-segment basis that are definitively not candidates for the current page of results.

If early termination is used, a `segmentTerminatedEarly` header will be included in the `responseHeader`.

Similar to using `the timeAllowed Parameter`, when early segment termination happens values such as `numFound`, `Facet` counts, and result `Stats` may not be accurate for the entire result set.

The default value of this parameter is false.

omitHeader Parameter

This parameter may be set to either true or false.

If set to true, this parameter excludes the header from the returned results. The header contains information about the request, such as the time it took to complete. The default value for this parameter is false.

wt Parameter

The wt parameter selects the Response Writer that Solr should use to format the query's response. For detailed descriptions of Response Writers, see [Response Writers](#).

If you do not define the wt parameter in your queries, JSON will be returned as the format of the response.

cache Parameter

Solr caches the results of all queries and filter queries by default. To disable result caching, set the cache=false parameter.

You can also use the cost option to control the order in which non-cached filter queries are evaluated. This allows you to order less expensive non-cached filters before expensive non-cached filters.

For very high cost filters, if cache=false and cost \geq 100 and the query implements the PostFilter interface, a Collector will be requested from that query and used to filter documents after they have matched the main query and all other filter queries. There can be multiple post filters; they are also ordered by cost.

For most queries the default behavior is cost=0 — but some types of queries such as {!frange} default to cost=100, because they are most efficient when used as a PostFilter.

For example:

This is an example of 3 regular filters, where all matching documents generated by each are computed up front and cached independently:

```
q=some keywords
fq=quantity_in_stock:[5 TO *]
fq={!frange l=10 u=100}mul(popularity,price)
fq={!frange cost=200 l=0}pow(mul(sum(1, query('tag:smartphone'))), div(1,avg_rating)), 2.3)
```

These are the same filters run w/o caching. The simple range query on the quantity_in_stock field will be run in parallel with the main query like a traditional lucene filter, while the 2 frange filters will only be checked against each document has already matched the main query and the quantity_in_stock range query — first the simpler mul(popularity,price) will be checked (because of its implicit cost=100) and only if it matches will the final very complex filter (with its higher cost=200) be checked.

```
q=some keywords
fq={!cache=false}quantity_in_stock:[5 TO *]
fq={!frange cache=false l=10 u=100}mul(popularity,price)
fq={!frange cache=false cost=200 l=0}pow(mul(sum(1, query('tag:smartphone'))), div(1,avg_rating)), 2.3)
```

logParamsList Parameter

By default, Solr logs all parameters of requests. Set this parameter to restrict which parameters of a request are logged. This may help control logging to only those parameters considered important to your organization.

For example, you could define this like:

```
logParamsList=q,fq
```

And only the 'q' and 'fq' parameters will be logged.

If no parameters should be logged, you can send logParamsList as empty (i.e., logParamsList=).



This parameter not only applies to query requests, but to any kind of request to Solr.

echoParams Parameter

The echoParams parameter controls what information about request parameters is included in the response header.

The echoParams parameter accepts the following values:

- explicit: This is the default value. Only parameters included in the actual request, plus the _ parameter (which is a 64-bit numeric timestamp) will be added to the params section of the response header.
- all: Include all request parameters that contributed to the query. This will include everything defined in the request handler definition found in solrconfig.xml as well as parameters included with the request, plus the _ parameter. If a parameter is included in the request handler definition AND the request, it will appear multiple times in the response header.
- none: Entirely removes the params section of the response header. No information about the request parameters will be available in the response.

Here is an example of a JSON response where the echoParams parameter was not included, so the default of explicit is active. The request URL that created this response included three parameters - q, wt, and indent:

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 0,
    "params": {
      "q": "solr",
      "indent": "true",
      "wt": "json",
      "_": "1458227751857"
    }
  },
  "response": {
    "numFound": 0,
    "start": 0,
    "docs": []
  }
}
```

This is what happens if a similar request is sent that adds echoParams=all to the three parameters used in the previous example:

```
{  
  "responseHeader": {  
    "status": 0,  
    "QTime": 0,  
    "params": {  
      "q": "solr",  
      "df": "text",  
      "preferLocalShards": "false",  
      "indent": "true",  
      "echoParams": "all",  
      "rows": "10",  
      "wt": "json",  
      "_": "1458228887287"  
    }  
  },  
  "response": {  
    "numFound": 0,  
    "start": 0,  
    "docs": []  
  }  
}
```

The Standard Query Parser

Solr's default Query Parser is also known as the "lucene" parser.

The key advantage of the standard query parser is that it supports a robust and fairly intuitive syntax allowing you to create a variety of structured queries. The largest disadvantage is that it's very intolerant of syntax errors, as compared with something like the [DisMax](#) query parser which is designed to throw as few errors as possible.

Standard Query Parser Parameters

In addition to the [Common Query Parameters](#), [Faceting Parameters](#), [Highlighting Parameters](#), and [MoreLikeThis Parameters](#), the standard query parser supports the parameters described in the table below.

q

Defines a query using standard query syntax. This parameter is mandatory.

q.op

Specifies the default operator for query expressions, overriding the default operator specified in the Schema. Possible values are "AND" or "OR".

df

Specifies a default field, overriding the definition of a default field in the Schema.

sow

Split on whitespace. If set to true, text analysis is invoked separately for each individual whitespace-separated term. The default is false; whitespace-separated term sequences will be provided to text analysis in one shot, enabling proper function of analysis filters that operate over term sequences, e.g.,

multi-word synonyms and shingles.

Default parameter values are specified in `solrconfig.xml`, or overridden by query-time values in the request.

Standard Query Parser Response

By default, the response from the standard query parser contains one `<result>` block, which is unnamed. If the debug `parameter` is used, then an additional `<lst>` block will be returned, using the name "debug". This will contain useful debugging info, including the original query string, the parsed query string, and explain info for each document in the `<result>` block. If the `explainOther` `parameter` is also used, then additional explain info will be provided for all the documents matching that query.

Sample Responses

This section presents examples of responses from the standard query parser.

The URL below submits a simple query and requests the XML Response Writer to use indentation to make the XML response more readable.

`http://localhost:8983/solr/techproducts/select?q=id:SP2514N&wt=xml`

Results:

```
<response>
<responseHeader><status>0</status><QTime>1</QTime></responseHeader>
<result numFound="1" start="0">
  <doc>
    <arr name="cat"><str>electronics</str><str>hard drive</str></arr>
    <arr name="features"><str>7200RPM, 8MB cache, IDE Ultra ATA-133</str>
      <str>NoiseGuard, SilentSeek technology, Fluid Dynamic Bearing (FDB) motor</str></arr>
    <str name="id">SP2514N</str>
    <bool name="inStock">true</bool>
    <str name="manu">Samsung Electronics Co. Ltd.</str>
    <str name="name">Samsung SpinPoint P120 SP2514N - hard drive - 250 GB - ATA-133</str>
    <int name="popularity">6</int>
    <float name="price">92.0</float>
    <str name="sku">SP2514N</str>
  </doc>
</result>
</response>
```

Here's an example of a query with a limited field list.

`http://localhost:8983/solr/techproducts/select?q=id:SP2514N&fl=id+name&wt=xml`

Results:

```
<response>
<responseHeader><status>0</status><QTime>2</QTime></responseHeader>
<result numFound="1" start="0">
<doc>
<str name="id">SP2514N</str>
<str name="name">Samsung SpinPoint P120 SP2514N - hard drive - 250 GB - ATA-133</str>
</doc>
</result>
</response>
```

Specifying Terms for the Standard Query Parser

A query to the standard query parser is broken up into terms and operators. There are two types of terms: single terms and phrases.

- A single term is a single word such as "test" or "hello"
- A phrase is a group of words surrounded by double quotes such as "hello dolly"

Multiple terms can be combined together with Boolean operators to form more complex queries (as described below).



It is important that the analyzer used for queries parses terms and phrases in a way that is consistent with the way the analyzer used for indexing parses terms and phrases; otherwise, searches may produce unexpected results.

Term Modifiers

Solr supports a variety of term modifiers that add flexibility or precision, as needed, to searches. These modifiers include wildcard characters, characters for making a search "fuzzy" or more general, and so on. The sections below describe these modifiers in detail.

Wildcard Searches

Solr's standard query parser supports single and multiple character wildcard searches within single terms. Wildcard characters can be applied to single terms, but not to search phrases.

Wildcard Search Type	Special Character	Example
Single character (matches a single character)	?	The search string te?t would match both test and text.
Multiple characters (matches zero or more sequential characters)	*	The wildcard search: tes* would match test, testing, and tester. You can also use wildcard characters in the middle of a term. For example: te*t would match test and text. *est would match pest and test.

Fuzzy Searches

Solr's standard query parser supports fuzzy searches based on the Damerau-Levenshtein Distance or Edit Distance algorithm. Fuzzy searches discover terms that are similar to a specified term without necessarily

being an exact match. To perform a fuzzy search, use the tilde ~ symbol at the end of a single-word term. For example, to search for a term similar in spelling to "roam," use the fuzzy search:

```
roam~
```

This search will match terms like roams, foam, & foams. It will also match the word "roam" itself.

An optional distance parameter specifies the maximum number of edits allowed, between 0 and 2, defaulting to 2. For example:

```
roam~1
```

This will match terms like roams & foam - but not foams since it has an edit distance of "2".



In many cases, stemming (reducing terms to a common stem) can produce similar effects to fuzzy searches and wildcard searches.

Proximity Searches

A proximity search looks for terms that are within a specific distance from one another.

To perform a proximity search, add the tilde character ~ and a numeric value to the end of a search phrase. For example, to search for a "apache" and "jakarta" within 10 words of each other in a document, use the search:

```
"jakarta apache"~10
```

The distance referred to here is the number of term movements needed to match the specified phrase. In the example above, if "apache" and "jakarta" were 10 spaces apart in a field, but "apache" appeared before "jakarta", more than 10 term movements would be required to move the terms together and position "apache" to the right of "jakarta" with a space in between.

Range Searches

A range search specifies a range of values for a field (a range with an upper bound and a lower bound). The query matches documents whose values for the specified field or fields fall within the range. Range queries can be inclusive or exclusive of the upper and lower bounds. Sorting is done lexicographically, except on numeric fields. For example, the range query below matches all documents whose popularity field has a value between 52 and 10,000, inclusive.

```
popularity:[52 TO 10000]
```

Range queries are not limited to date fields or even numerical fields. You could also use range queries with non-date fields:

```
title:{Aida TO Carmen}
```

This will find all documents whose titles are between Aida and Carmen, but not including Aida and Carmen.

The brackets around a query determine its inclusiveness.

- Square brackets [&] denote an inclusive range query that matches values including the upper and lower bound.

- Curly brackets { & } denote an exclusive range query that matches values between the upper and lower bounds, but excluding the upper and lower bounds themselves.
- You can mix these types so one end of the range is inclusive and the other is exclusive. Here's an example: count:{1 TO 10]

Boosting a Term with "[^]"

Lucene/Solr provides the relevance level of matching documents based on the terms found. To boost a term use the caret symbol [^] with a boost factor (a number) at the end of the term you are searching. The higher the boost factor, the more relevant the term will be.

Boosting allows you to control the relevance of a document by boosting its term. For example, if you are searching for

"jakarta apache" and you want the term "jakarta" to be more relevant, you can boost it by adding the [^] symbol along with the boost factor immediately after the term. For example, you could type:

```
jakarta^4 apache
```

This will make documents with the term jakarta appear more relevant. You can also boost Phrase Terms as in the example:

```
"jakarta apache"^4 "Apache Lucene"
```

By default, the boost factor is 1. Although the boost factor must be positive, it can be less than 1 (for example, it could be 0.2).

Constant Score with "^{^=}"

Constant score queries are created with <query_clause>^{^=}<score>, which sets the entire clause to the specified score for any documents matching that clause. This is desirable when you only care about matches for a particular clause and don't want other relevancy factors such as term frequency (the number of times the term appears in the field) or inverse document frequency (a measure across the whole index for how rare a term is in a field).

Example:

```
(description:blue OR color:blue)^=1.0 text:shoes
```

Querying Specific Fields

Data indexed in Solr is organized in fields, which are [defined in the Solr Schema](#). Searches can take advantage of fields to add precision to queries. For example, you can search for a term only in a specific field, such as a title field.

The Schema defines one field as a default field. If you do not specify a field in a query, Solr searches only the default field. Alternatively, you can specify a different field or a combination of fields in a query.

To specify a field, type the field name followed by a colon ":" and then the term you are searching for within the field.

For example, suppose an index contains two fields, title and text, and that text is the default field. If you want to find a document called "The Right Way" which contains the text "don't go this way," you could include either of the following terms in your search query:

```
title:"The Right Way" AND text:go
```

```
title:"Do it right" AND go
```

Since text is the default field, the field indicator is not required; hence the second query above omits it.

The field is only valid for the term that it directly precedes, so the query title:Do it right will find only "Do" in the title field. It will find "it" and "right" in the default field (in this case the text field).

Boolean Operators Supported by the Standard Query Parser

Boolean operators allow you to apply Boolean logic to queries, requiring the presence or absence of specific terms or conditions in fields in order to match documents. The table below summarizes the Boolean operators supported by the standard query parser.

Boolean Operator	Alternative Symbol	Description
AND	&&	Requires both terms on either side of the Boolean operator to be present for a match.
NOT	!	Requires that the following term not be present.
OR		Requires that either term (or both terms) be present for a match.
	+	Requires that the following term be present.
	-	Prohibits the following term (that is, matches on fields or documents that do not include that term). The - operator is functionally similar to the Boolean operator !. Because it's used by popular search engines such as Google, it may be more familiar to some user communities.

Boolean operators allow terms to be combined through logic operators. Lucene supports AND, "+", OR, NOT and "-" as Boolean operators.



When specifying Boolean operators with keywords such as AND or NOT, the keywords must appear in all uppercase.



The standard query parser supports all the Boolean operators listed in the table above. The DisMax query parser supports only + and -.

The OR operator is the default conjunction operator. This means that if there is no Boolean operator between two terms, the OR operator is used. The OR operator links two terms and finds a matching

document if either of the terms exist in a document. This is equivalent to a union using sets. The symbol || can be used in place of the word OR.

To search for documents that contain either "jakarta apache" or just "jakarta," use the query:

"jakarta apache" jakarta

or

"jakarta apache" OR jakarta

The Boolean Operator "+"

The + symbol (also known as the "required" operator) requires that the term after the + symbol exist somewhere in a field in at least one document in order for the query to return a match.

For example, to search for documents that must contain "jakarta" and that may or may not contain "lucene," use the following query:

+jakarta lucene



This operator is supported by both the standard query parser and the DisMax query parser.

The Boolean Operator AND ("&&")

The AND operator matches documents where both terms exist anywhere in the text of a single document. This is equivalent to an intersection using sets. The symbol && can be used in place of the word AND.

To search for documents that contain "jakarta apache" and "Apache Lucene," use either of the following queries:

"jakarta apache" AND "Apache Lucene"

"jakarta apache" && "Apache Lucene"

The Boolean Operator NOT ("!")

The NOT operator excludes documents that contain the term after NOT. This is equivalent to a difference using sets. The symbol ! can be used in place of the word NOT.

The following queries search for documents that contain the phrase "jakarta apache" but do not contain the phrase "Apache Lucene":

"jakarta apache" NOT "Apache Lucene"

"jakarta apache" ! "Apache Lucene"

The Boolean Operator "-"

The - symbol or "prohibit" operator excludes documents that contain the term after the - symbol.

For example, to search for documents that contain "jakarta apache" but not "Apache Lucene," use the following query:

"jakarta apache" -"Apache Lucene"

Escaping Special Characters

Solr gives the following characters special meaning when they appear in a query:

```
+ - && || ! ( ) { } [ ] ^ " ~ * ? : /
```

To make Solr interpret any of these characters literally, rather than as a special character, precede the character with a backslash character \. For example, to search for (1+1):2 without having Solr interpret the plus sign and parentheses as special characters for formulating a sub-query with two terms, escape the characters by preceding each one with a backslash:

```
\(1\+1\)\:2
```

Grouping Terms to Form Sub-Queries

Lucene/Solr supports using parentheses to group clauses to form sub-queries. This can be very useful if you want to control the Boolean logic for a query.

The query below searches for either "jakarta" or "apache" and "website":

```
(jakarta OR apache) AND website
```

This adds precision to the query, requiring that the term "website" exist, along with either term "jakarta" and "apache."

Grouping Clauses within a Field

To apply two or more Boolean operators to a single field in a search, group the Boolean clauses within parentheses. For example, the query below searches for a title field that contains both the word "return" and the phrase "pink panther":

```
title:(+return +"pink panther")
```

Comments in Queries

C-Style comments are supported in query strings.

Example:

```
"jakarta apache" /* this is a comment in the middle of a normal query string */ OR jakarta
```

Comments may be nested.

Differences between Lucene's Classic Query Parser and Solr's Standard Query Parser

Solr's standard query parser originated as a variation of Lucene's "classic" QueryParser. It diverges in the following ways:

- A * may be used for either or both endpoints to specify an open-ended range query
 - field:[* TO 100] finds all field values less than or equal to 100

- field:[100 TO *] finds all field values greater than or equal to 100
- field:[* TO *} matches all documents with the field
- Pure negative queries (all clauses prohibited) are allowed (only as a top-level clause)
 - -inStock:false finds all field values where inStock is not false
 - -field:[* TO *} finds all documents without a value for field
- Support for embedded Solr queries (sub-queries) using any type of query parser as a nested clause using the local-params syntax.
 - inStock:true OR {!dismax qf='name manu' v='ipod'}

Gotcha: Be careful not to start your query with {! at the very beginning, which changes the parsing of the entire query string, which may not be what you want if there are additional clauses. So flipping the example above so the sub-query comes first would fail to work as expected without a leading space.

Sub-queries can also be done with the magic field `_query_` and for function queries with the magic field `_val_` but it should be considered deprecated since it is less clear. Example:
`_val_:"recip(rord(myfield),1,2,3)"`

- Support for a special filter(...) syntax to indicate that some query clauses should be cached in the filter cache (as a constant score boolean query). This allows sub-queries to be cached and re-used in other queries. For example inStock:true will be cached and re-used in all three of the queries below:

- q=features:songs OR filter(inStock:true)
- q=+manu:Apple +filter(inStock:true)
- q=+manu:Apple & fq=inStock:true

This can even be used to cache individual clauses of complex filter queries. In the first query below, 3 items will be added to the filter cache (the top level fq and both filter(...) clauses) and in the second query, there will be 2 cache hits, and one new cache insertion (for the new top level fq):

- q=features:songs & fq=+filter(inStock:true) +filter(price:[* TO 100])
- q=manu:Apple & fq=-filter(inStock:true) -filter(price:[* TO 100])

- Range queries ("[a TO z]"), prefix queries ("a*"), and wildcard queries ("a*b") are constant-scoring (all matching documents get an equal score). The scoring factors TF, IDF, index boost, and "coord" are not used. There is no limitation on the number of terms that match (as there was in past versions of Lucene).
- Constant score queries are created with `<query_clause>^=<score>`, which sets the entire clause to the specified score for any documents matching that clause:
 - q=(description:blue color:blue)^=1.0 title:blue^=5.0

Specifying Dates and Times

Queries against date based fields must use the [appropriate date formating](#). Queries for exact date values will require quoting or escaping since : is the parser syntax used to denote a field query:

- createdate:1976-03-06T23\:\:59\:59.999Z
- createdate:"1976-03-06T23:59:59.999Z"
- createdate:[1976-03-06T23:59:59.999Z TO *]

- createdate:[1995-12-31T23:59:59.999Z TO 2007-03-06T00:00:00Z]
- timestamp:[* TO NOW]
- pubdate:[NOW-1YEAR/DAY TO NOW/DAY+1DAY]
- createdate:[1976-03-06T23:59:59.999Z TO 1976-03-06T23:59:59.999Z+1YEAR]
- createdate:[1976-03-06T23:59:59.999Z/YEAR TO 1976-03-06T23:59:59.999Z]

The DisMax Query Parser

The DisMax query parser is designed to process simple phrases (without complex syntax) entered by users and to search for individual terms across several fields using different weighting (boosts) based on the significance of each field. Additional options enable users to influence the score based on rules specific to each use case (independent of user input).

In general, the DisMax query parser's interface is more like that of Google than the interface of the 'lucene' Solr query parser. This similarity makes DisMax the appropriate query parser for many consumer applications. It accepts a simple syntax, and it rarely produces error messages.

The DisMax query parser supports an extremely simplified subset of the Lucene QueryParser syntax. As in Lucene, quotes can be used to group phrases, and +/- can be used to denote mandatory and optional clauses. All other Lucene query parser special characters (except AND and OR) are escaped to simplify the user experience. The DisMax query parser takes responsibility for building a good query from the user's input using Boolean clauses containing DisMax queries across fields and boosts specified by the user. It also lets the Solr administrator provide additional boosting queries, boosting functions, and filtering queries to artificially affect the outcome of all searches. These options can all be specified as default parameters for the request handler in the `solrconfig.xml` file or overridden in the Solr query URL.

Interested in the technical concept behind the DisMax name? DisMax stands for Maximum Disjunction. Here's a definition of a Maximum Disjunction or "DisMax" query:

A query that generates the union of documents produced by its subqueries, and that scores each document with the maximum score for that document as produced by any subquery, plus a tie breaking increment for any additional matching subqueries.

Whether or not you remember this explanation, do remember that the DisMax Query Parser was primarily designed to be easy to use and to accept almost any input without returning an error.

DisMax Query Parser Parameters

In addition to the common request parameters, highlighting parameters, and simple facet parameters, the DisMax query parser supports the parameters described below. Like the standard query parser, the DisMax query parser allows default parameter values to be specified in `solrconfig.xml`, or overridden by query-time values in the request.

The sections below explain these parameters in detail.

q Parameter

The `q` parameter defines the main "query" constituting the essence of the search. The parameter supports raw input strings provided by users with no special escaping. The + and - characters are treated as

"mandatory" and "prohibited" modifiers for terms. Text wrapped in balanced quote characters (for example, "San Jose") is treated as a phrase. Any query containing an odd number of quote characters is evaluated as if there were no quote characters at all.



The q parameter does not support wildcard characters such as *.

q.alt Parameter

If specified, the q.alt parameter defines a query (which by default will be parsed using standard query parsing syntax) when the main q parameter is not specified or is blank. The q.alt parameter comes in handy when you need something like a query to match all documents (don't forget &rows=0 for that one!) in order to get collection-wide facetting counts.

qf (Query Fields) Parameter

The qf parameter introduces a list of fields, each of which is assigned a boost factor to increase or decrease that particular field's importance in the query. For example, the query below:

```
qf="fieldOne^2.3 fieldTwo fieldThree^0.4"
```

assigns fieldOne a boost of 2.3, leaves fieldTwo with the default boost (because no boost factor is specified), and fieldThree a boost of 0.4. These boost factors make matches in fieldOne much more significant than matches in fieldTwo, which in turn are much more significant than matches in fieldThree.

mm (Minimum Should Match) Parameter

When processing queries, Lucene/Solr recognizes three types of clauses: mandatory, prohibited, and "optional" (also known as "should" clauses). By default, all words or phrases specified in the q parameter are treated as "optional" clauses unless they are preceded by a "+" or a "-". When dealing with these "optional" clauses, the mm parameter makes it possible to say that a certain minimum number of those clauses must match. The DisMax query parser offers great flexibility in how the minimum number can be specified.

The table below explains the various ways that mm values can be specified.

Syntax	Example	Description
Positive integer	3	Defines the minimum number of clauses that must match, regardless of how many clauses there are in total.
Negative integer	-2	Sets the minimum number of matching clauses to the total number of optional clauses, minus this value.
Percentage	75%	Sets the minimum number of matching clauses to this percentage of the total number of optional clauses. The number computed from the percentage is rounded down and used as the minimum.
Negative percentage	-25%	Indicates that this percent of the total number of optional clauses can be missing. The number computed from the percentage is rounded down, before being subtracted from the total to determine the minimum number.

Syntax	Example	Description
An expression beginning with a positive integer followed by a > or < sign and another value	3<90%	Defines a conditional expression indicating that if the number of optional clauses is equal to (or less than) the integer, they are all required, but if it's greater than the integer, the specification applies. In this example: if there are 1 to 3 clauses they are all required, but for 4 or more clauses only 90% are required.
Multiple conditional expressions involving > or < signs	2<-25% 9<-3	Defines multiple conditions, each one being valid only for numbers greater than the one before it. In the example at left, if there are 1 or 2 clauses, then both are required. If there are 3-9 clauses all but 25% are required. If there are more than 9 clauses, all but three are required.

When specifying `mm` values, keep in mind the following:

- When dealing with percentages, negative values can be used to get different behavior in edge cases. 75% and -25% mean the same thing when dealing with 4 clauses, but when dealing with 5 clauses 75% means 3 are required, but -25% means 4 are required.
- If the calculations based on the parameter arguments determine that no optional clauses are needed, the usual rules about Boolean queries still apply at search time. (That is, a Boolean query containing no required clauses must still match at least one optional clause).
- No matter what number the calculation arrives at, Solr will never use a value greater than the number of optional clauses, or a value less than 1. In other words, no matter how low or how high the calculated result, the minimum number of required matches will never be less than 1 or greater than the number of clauses.
- When searching across multiple fields that are configured with different query analyzers, the number of optional clauses may differ between the fields. In such a case, the value specified by `mm` applies to the maximum number of optional clauses. For example, if a query clause is treated as stopword for one of the fields, the number of optional clauses for that field will be smaller than for the other fields. A query with such a stopword clause would not return a match in that field if `mm` is set to 100% because the removed clause does not count as matched.

The default value of `mm` is 0% (all clauses optional), unless `q.op` is specified as "AND", in which case `mm` defaults to 100% (all clauses required).

pf (Phrase Fields) Parameter

Once the list of matching documents has been identified using the `fq` and `qf` parameters, the `pf` parameter can be used to "boost" the score of documents in cases where all of the terms in the `q` parameter appear in close proximity.

The format is the same as that used by the `qf` parameter: a list of fields and "boosts" to associate with each of them when making phrase queries out of the entire `q` parameter.

ps (Phrase Slop) Parameter

The `ps` parameter specifies the amount of "phrase slop" to apply to queries specified with the `pf` parameter. Phrase slop is the number of positions one token needs to be moved in relation to another token in order to

match a phrase specified in a query.

qs (Query Phrase Slop) Parameter

The qs parameter specifies the amount of slop permitted on phrase queries explicitly included in the user's query string with the qf parameter. As explained above, slop refers to the number of positions one token needs to be moved in relation to another token in order to match a phrase specified in a query.

The tie (Tie Breaker) Parameter

The tie parameter specifies a float value (which should be something much less than 1) to use as tiebreaker in DisMax queries.

When a term from the user's input is tested against multiple fields, more than one field may match. If so, each field will generate a different score based on how common that word is in that field (for each document relative to all other documents). The tie parameter lets you control how much the final score of the query will be influenced by the scores of the lower scoring fields compared to the highest scoring field.

A value of "0.0" - the default - makes the query a pure "disjunction max query": that is, only the maximum scoring subquery contributes to the final score. A value of "1.0" makes the query a pure "disjunction sum query" where it doesn't matter what the maximum scoring sub query is, because the final score will be the sum of the subquery scores. Typically a low value, such as 0.1, is useful.

bq (Boost Query) Parameter

The bq parameter specifies an additional, optional, query clause that will be added to the user's main query to influence the score. For example, if you wanted to add a relevancy boost for recent documents:

```
q=cheese  
bq=date:[NOW/DAY-1YEAR TO NOW/DAY]
```

You can specify multiple bq parameters. If you want your query to be parsed as separate clauses with separate boosts, use multiple bq parameters.

bf (Boost Functions) Parameter

The bf parameter specifies functions (with optional boosts) that will be used to construct FunctionQueries which will be added to the user's main query as optional clauses that will influence the score. Any function supported natively by Solr can be used, along with a boost value. For example:

```
recip(rord(myfield),1,2,3)^1.5
```

Specifying functions with the bf parameter is essentially just shorthand for using the bq parameter combined with the {!func} parser.

For example, if you want to show the most recent documents first, you could use either of the following:

```
bf=recip(rord(creationDate),1,1000,1000)
...or...
bq={!func}recip(rord(creationDate),1,1000,1000)
```

Examples of Queries Submitted to the DisMax Query Parser

All of the sample URLs in this section assume you are running Solr's "techproducts" example:

```
bin/solr -e techproducts
```

Results for the word "video" using the standard query parser, and we assume "df" is pointing to a field to search:

```
http://localhost:8983/solr/techproducts/select?q=video&fl=name+score
```

The "dismax" parser is configured to search across the text, features, name, sku, id, manu, and cat fields all with varying boosts designed to ensure that "better" matches appear first, specifically: documents which match on the name and cat fields get higher scores.

```
http://localhost:8983/solr/techproducts/select?defType=dismax&q=video
```

Note that this instance is also configured with a default field list, which can be overridden in the URL.

```
http://localhost:8983/solr/techproducts/select?defType=dismax&q=video&fl=*,score
```

You can also override which fields are searched on and how much boost each field gets.

```
http://localhost:8983/solr/techproducts/select?defType=dismax&q=video&qf=features^20.0+text^0.3
```

You can boost results that have a field that matches a specific value.

```
http://localhost:8983/solr/techproducts/select?defType=dismax&q=video&bq=cat:electronics^5.0
```

Another request handler is registered at "/instock" and has slightly different configuration options, notably: a filter for (you guessed it) inStock:true).

```
http://localhost:8983/solr/techproducts/select?defType=dismax&q=video&fl=name, score, inStock
```

```
http://localhost:8983/solr/techproducts/instock?defType=dismax&q=video&fl=name, score, inStock
```

One of the other really cool features in this parser is robust support for specifying the "BooleanQuery.minimumNumberShouldMatch" you want to be used based on how many terms are in your user's query. These allows flexibility for typos and partial matches. For the dismax parser, one and two word queries require that all of the optional clauses match, but for three to five word queries one missing word is allowed.

```
http://localhost:8983/solr/techproducts/select?defType=dismax&q=belkin+ipod
```

```
http://localhost:8983/solr/techproducts/select?defType=dismax&q=belkin+ipod+gibberish
```

```
http://localhost:8983/solr/techproducts/select?defType=dismax&q=belkin+ipod+apple
```

Use the `debugQuery` option to see the parsed query, and the score explanations for each document.

`http://localhost:8983/solr/techproducts/select?defType=dismax&q=belkin+iPod+gibberish&debugQuery=true`

`http://localhost:8983/solr/techproducts/select?defType=dismax&q=video+card&debugQuery=true`

The Extended DisMax (eDismax) Query Parser

The Extended DisMax (eDisMax) query parser is an improved version of the [DisMax query parser](#).

In addition to supporting all the DisMax query parser parameters, Extended Dismax:

- supports [Solr's standard query parser](#) syntax such as (non-exhaustive list):
 - boolean operators such as AND (+, &&), OR (||), NOT (-).
 - optionally treats lowercase "and" and "or" as "AND" and "OR" in Lucene syntax mode
 - optionally allows embedded queries using other query parsers or functions
- includes improved smart partial escaping in the case of syntax errors; fielded queries, +/-, and phrase queries are still supported in this mode.
- improves proximity boosting by using word shingles; you do not need the query to match all words in the document before proximity boosting is applied.
- includes advanced stopword handling: stopwords are not required in the mandatory part of the query but are still used in the proximity boosting part. If a query consists of all stopwords, such as "to be or not to be", then all words are required.
- includes improved boost function: in Extended DisMax, the boost function is a multiplier rather than an addend, improving your boost results; the additive boost functions of DisMax (bf and bq) are also supported.
- supports pure negative nested queries: queries such as +foo (-foo) will match all documents.
- lets you specify which fields the end user is allowed to query, and to disallow direct fielded searches.

Extended DisMax Parameters

In addition to all the [DisMax parameters](#), Extended DisMax includes these query parameters:

`sow`

Split on whitespace. If set to true, text analysis is invoked separately for each individual whitespace-separated term. The default is false; whitespace-separated term sequences will be provided to text analysis in one shot, enabling proper function of analysis filters that operate over term sequences, e.g., multi-word synonyms and shingles.

`mm`

Minimum should match. See the [DisMax mm parameter](#) for a description of `mm`. The default eDisMax `mm` value differs from that of DisMax:

- The default `mm` value is 0%:
 - if the query contains an explicit operator other than "AND" ("-", "+", "OR", "NOT"); or
 - if `q.op` is "OR" or is not specified.

- The default mm value is 100% if q.op is "AND" and the query does not contain any explicit operators other than "AND".

mm.autoRelax

If true, the number of clauses required ([minimum should match](#)) will automatically be relaxed if a clause is removed (by e.g., stopwords filter) from some but not all qf fields. Use this parameter as a workaround if you experience that queries return zero hits due to uneven stopword removal between the qf fields.

Note that relaxing mm may cause undesired side effects, such as hurting the precision of the search, depending on the nature of your index content.

boost

A multivalued list of strings parsed as queries with scores multiplied by the score from the main query for all matching documents. This parameter is shorthand for wrapping the query produced by eDisMax using the BoostQParserPlugin.

lowercaseOperators

A Boolean parameter indicating if lowercase "and" and "or" should be treated the same as operators "AND" and "OR". Defaults to false.

ps

Phrase Slop. The default amount of slop - distance between terms - on phrase queries built with pf, pf2 and/or pf3 fields (affects boosting). See also the section [Using 'Slop'](#) below.

pf2

A multivalued list of fields with optional weights. Similar to pf, but based on word *pair* shingles.

ps2

This is similar to ps but overrides the slop factor used for pf2. If not specified, ps is used.

pf3

A multivalued list of fields with optional weights, based on triplets of word shingles. Similar to pf, except that instead of building a phrase per field out of all the words in the input, it builds a set of phrases for each field out of word *triplet* shingles.

ps3

This is similar to ps but overrides the slop factor used for pf3. If not specified, ps is used.

stopwords

A Boolean parameter indicating if the StopFilterFactory configured in the query analyzer should be respected when parsing the query. If this is set to false, then the StopFilterFactory in the query analyzer is ignored.

uf

Specifies which schema fields the end user is allowed to explicitly query and to toggle whether embedded Solr queries are supported. This parameter supports wildcards. Multiple fields must be separated by a space.

The default is to allow all fields and no embedded Solr queries, equivalent to uf=*_query_.

- To allow only title field, use uf=title.

- To allow title and all fields ending with '_s', use `uf=title *_s`.
- To allow all fields except title, use `uf=* -title`.
- To disallow all fielded searches, use `uf=-*`.
- To allow embedded Solr queries (e.g., `_query_:"..."` or `_val_:"..."` or `{!lucene ...}`), you *must* expressly enable this by referring to the magic field `_query_` in `uf`.

Field Aliasing using Per-Field qf Overrides

Per-field overrides of the `qf` parameter may be specified to provide 1-to-many aliasing from field names specified in the query string, to field names used in the underlying query. By default, no aliasing is used and field names specified in the query string are treated as literal field names in the index.

Examples of eDismax Queries

All of the sample URLs in this section assume you are running Solr's "techproducts" example:

```
bin/solr -e techproducts
```

Boost the result of the query term "hello" based on the document's popularity:

```
http://localhost:8983/solr/techproducts/select?defType=edismax&q=hello&pf=text&qf=text&boost=popularity
```

Search for iPods OR video:

```
http://localhost:8983/solr/techproducts/select?defType=edismax&q=ipod+OR+video
```

Search across multiple fields, specifying (via boosts) how important each field is relative each other:

```
http://localhost:8983/solr/techproducts/select?q=video&defType=edismax&qf=features^20.0+text^0.3
```

You can boost results that have a field that matches a specific value:

```
http://localhost:8983/solr/techproducts/select?q=video&defType=edismax&qf=features^20.0+text^0.3&bq=cat:electronics^5.0
```

Using the `mm` parameter, 1 and 2 word queries require that all of the optional clauses match, but for queries with three or more clauses one missing clause is allowed:

```
http://localhost:8983/solr/techproducts/select?q=belkin+ipod&defType=edismax&mm=2  
http://localhost:8983/solr/techproducts/select?q=belkin+ipod+gibberish&defType=edismax&mm=2  
http://localhost:8983/solr/techproducts/select?q=belkin+ipod+apple&defType=edismax&mm=2
```

In the example below, we see a per-field override of the `qf` parameter being used to alias "name" in the

query string to either the "last_name" and "first_name" fields:

```
defType=edismax
q=sysadmin name:Mike
qf=title text last_name first_name
f.name.qf=last_name first_name
```

Using Negative Boost

Negative query boosts have been supported at the "Query" object level for a long time (resulting in negative scores for matching documents). Now the QueryParsers have been updated to handle this too.

Using 'Slop'

Dismax and Edismax can run queries against all query fields, and also run a query in the form of a phrase against the phrase fields. (This will work only for boosting documents, not actually for matching.) However, that phrase query can have a 'slop,' which is the distance between the terms of the query while still considering it a phrase match. For example:

```
q=foo bar
qf=field1^5 field2^10
pf=field1^50 field2^20
defType=dismax
```

With these parameters, the Dismax Query Parser generates a query that looks something like this:

```
(+(field1:foo^5 OR field2:foo^10) AND (field1:bar^5 OR field2:bar^10))
```

But it also generates another query that will only be used for boosting results:

```
field1:"foo bar"^50 OR field2:"foo bar"^20
```

Thus, any document that has the terms "foo" and "bar" will match; however if some of those documents have both of the terms as a phrase, it will score much higher because it's more relevant.

If you add the parameter ps (phrase slop), the second query will instead be:

```
ps=10 field1:"foo bar"~10^50 OR field2:"foo bar"~10^20
```

This means that if the terms "foo" and "bar" appear in the document with less than 10 terms between each other, the phrase will match. For example the doc that says:

```
*Foo* term1 term2 term3 *bar*
```

will match the phrase query.

How does one use phrase slop? Usually it is configured in the request handler (in solrconfig).

With query slop (qs) the concept is similar, but it applies to explicit phrase queries from the user. For example, if you want to search for a name, you could enter:

```
q="Hans Anderson"
```

A document that contains "Hans Anderson" will match, but a document that contains the middle name "Christian" or where the name is written with the last name first ("Anderson, Hans") won't. For those cases one could configure the query field qs, so that even if the user searches for an explicit phrase query, a slop is applied.

Finally, in addition to the phrase fields (pf) parameter, edismax also supports the pf2 and pf3 parameters, for fields over which to create bigram and trigram phrase queries. The phrase slop for these parameters' queries can be specified using the ps2 and ps3 parameters, respectively. If you use pf2/pf3 but not ps2/ps3, then the phrase slop for these parameters' queries will be taken from the ps parameter, if any.

Synonyms Expansion in Phrase Queries with Slop

When a phrase query with slop (e.g., pf with ps) triggers synonym expansions, a separate clause will be generated for each combination of synonyms. For example, with configured synonyms dog, canine and cat, feline, the query "dog chased cat" will generate the following phrase query clauses:

- "dog chased cat"
- "canine chased cat"
- "dog chased feline"
- "canine chased feline"

Function Queries

Function queries enable you to generate a relevancy score using the actual value of one or more numeric fields.

Function queries are supported by the [DisMax](#), [Extended DisMax](#), and [standard](#) query parsers.

Function queries use *functions*. The functions can be a constant (numeric or string literal), a field, another function or a parameter substitution argument. You can use these functions to modify the ranking of results for users. These could be used to change the ranking of results based on a user's location, or some other calculation.

Using Function Query

Functions must be expressed as function calls (for example, sum(a,b) instead of simply a+b).

There are several ways of using function queries in a Solr query:

- Via an explicit query parser that expects function arguments, such func or frange. For example:

```
q={!func}div(popularity,price)&fq={!frange l=1000}customer_ratings
```

- In a Sort expression. For example:

```
sort=div(popularity,price) desc, score desc
```

- Add the results of functions as pseudo-fields to documents in query results. For instance, for:

```
&fl=sum(x, y),id,a,b,c,score&wt=xml
```

the output would be:

```
...
<str name="id">foo</str>
<float name="sum(x,y)">40</float>
<float name="score">0.343</float>
...
```

- Use in a parameter that is explicitly for specifying functions, such as the eDisMax query parser's [boost parameter](#), or the DisMax query parser's [bf \(boost function\) parameter](#). (Note that the bf parameter actually takes a list of function queries separated by white space and each with an optional boost. Make sure you eliminate any internal white space in single function queries when using bf). For example:

```
q=dismax&bf="ord(popularity)^0.5 recip(rord(price),1,1000,1000)^0.3"
```

- Introduce a function query inline in the Lucene query parser with the `_val_` keyword. For example:

```
q=_val_:*mynumericfield _val_:"recip(rord(myfield),1,2,3)"
```

Only functions with fast random access are recommended.

Available Functions

The table below summarizes the functions available for function queries.

abs Function

Returns the absolute value of the specified value or function.

Syntax Examples

- `abs(x)`
- `abs(-5)`

childfield(field) Function

Returns the value of the given field for one of the matched child docs when searching by `{!parent}`. It can be used only in sort parameter.

Syntax Examples

- `sort=childfield(name) asc` implies `$q` as a second argument and therefore it assumes `q={!parent ...}..;`
- `sort=childfield(field,$bjq) asc` refers to a separate parameter `bj=q={!parent ...}..;`
- `sort=childfield(field,{!parent of=...}...) desc` allows to inline block join parent query

concat Function

Concatenates the given string fields, literals and other functions.

Syntax Example

- `concat(name, " ", $param, def(opt, "-"))`

"constant" Function

Specifies a floating point constant.

Syntax Example

- `1.5`

def Function

`def` is short for `default`. Returns the value of field "field", or if the field does not exist, returns the default value specified. Yields the first value where `exists() == true`.

Syntax Examples

- `def(rating,5):` This `def()` function returns the rating, or if no rating specified in the doc, returns 5
- `def(myfield, 1.0):` equivalent to `if(exists(myfield),myfield,1.0)`

div Function

Divides one value or function by another. `div(x,y)` divides `x` by `y`.

Syntax Examples

- `div(1,y)`
- `div(sum(x,100),max(y,1))`

dist Function

Returns the distance between two vectors (points) in an n-dimensional space. Takes in the power, plus two or more ValueSource instances and calculates the distances between the two vectors. Each ValueSource must be a number.

There must be an even number of ValueSource instances passed in and the method assumes that the first half represent the first vector and the second half represent the second vector.

Syntax Examples

- `dist(2, x, y, 0, 0)`: calculates the Euclidean distance between (0,0) and (x,y) for each document.
- `dist(1, x, y, 0, 0)`: calculates the Manhattan (taxicab) distance between (0,0) and (x,y) for each document.
- `dist(2, x,y,z,0,0,0)`: Euclidean distance between (0,0,0) and (x,y,z) for each document.
- `dist(1,x,y,z,e,f,g)`: Manhattan distance between (x,y,z) and (e,f,g) where each letter is a field name.

docfreq(field,val) Function

Returns the number of documents that contain the term in the field. This is a constant (the same value for all documents in the index).

You can quote the term if it's more complex, or do parameter substitution for the term value.

Syntax Examples

- `docfreq(text,'solr')`
- `...&defType=func &q=docfreq(text,$myterm)&myterm=solr`

field Function

Returns the numeric docValues or indexed value of the field with the specified name. In its simplest (single argument) form, this function can only be used on single valued fields, and can be called using the name of the field as a string, or for most conventional field names simply use the field name by itself without using the `field(...)` syntax.

When using docValues, an optional 2nd argument can be specified to select the `min` or `max` value of multivalued fields.

0 is returned for documents without a value in the field.

Syntax Examples These 3 examples are all equivalent:

- `myFloatFieldName`
- `field(myFloatFieldName)`
- `field("myFloatFieldName")`

The last form is convenient when your field name is atypical:

- `field("my complex float fieldName")`

For multivalued docValues fields:

- `field(myMultiValuedFloatField,min)`
- `field(myMultiValuedFloatField,max)`

hsin Function

The Haversine distance calculates the distance between two points on a sphere when traveling along the sphere. The values must be in radians. hsin also take a Boolean argument to specify whether the function should convert its output to radians.

Syntax Example

- `hsin(2, true, x, y, 0, 0)`

idf Function

Inverse document frequency; a measure of whether the term is common or rare across all documents. Obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. See also tf.

Syntax Example

- `idf(fieldName, 'solr')`: measures the inverse of the frequency of the occurrence of the term 'solr' in fieldName.

if Function

Enables conditional function queries. In `if(test, value1, value2)`:

- test is or refers to a logical value or expression that returns a logical value (TRUE or FALSE).
- value1 is the value that is returned by the function if test yields TRUE.
- value2 is the value that is returned by the function if test yields FALSE.

An expression can be any function which outputs boolean values, or even functions returning numeric values, in which case value 0 will be interpreted as false, or strings, in which case empty string is interpreted as false.

Syntax Example

- `if(termfreq(cat, 'electronics'), popularity, 42)`: This function checks each document for to see if it contains the term "electronics" in the cat field. If it does, then the value of the popularity field is returned, otherwise the value of 42 is returned.

linear Function

Implements $m \cdot x + c$ where m and c are constants and x is an arbitrary function. This is equivalent to `sum(product(m, x), c)`, but slightly more efficient as it is implemented as a single function.

Syntax Examples

- `linear(x, m, c)`
- `linear(x, 2, 4)`: returns $2 \cdot x + 4$

log Function

Returns the log base 10 of the specified function.

Syntax Examples

- `log(x)`
- `log(sum(x,100))`

map Function

Maps any values of an input function `x` that fall within `min` and `max` inclusive to the specified target. The arguments `min` and `max` must be constants. The arguments `target` and `default` can be constants or functions.

If the value of `x` does not fall between `min` and `max`, then either the value of `x` is returned, or a default value is returned if specified as a 5th argument.

Syntax Examples

- `map(x,min,max,target)`
 - `map(x,0,0,1)`: Changes any values of 0 to 1. This can be useful in handling default 0 values.
- `map(x,min,max,target,default)`
 - `map(x,0,100,1,-1)`: Changes any values between 0 and 100 to 1, and all other values to ` -1 `.
 - `map(x,0,100,sum(x,599),docfreq(text,solr))`: Changes any values between 0 and 100 to `x+599`, and all other values to frequency of the term 'solr' in the field text.

max Function

Returns the maximum numeric value of multiple nested functions or constants, which are specified as arguments: `max(x,y,...)`. The `max` function can also be useful for "bottoming out" another function or field at some specified constant.

Use the `field(myfield,max)` syntax for [selecting the maximum value of a single multivalued field](#).

Syntax Example

- `max(myfield,myotherfield,0)`

maxdoc Function

Returns the number of documents in the index, including those that are marked as deleted but have not yet been purged. This is a constant (the same value for all documents in the index).

Syntax Example

- `maxdoc()`

min Function

Returns the minimum numeric value of multiple nested functions or constants, which are specified as arguments: `min(x,y,...)`. The `min` function can also be useful for providing an "upper bound" on a function using a constant.

Use the `field(myfield,min)` [syntax for selecting the minimum value of a single multivalued field](#).

Syntax Example

- `min(myfield,myotherfield,0)`

ms Function

Returns milliseconds of difference between its arguments. Dates are relative to the Unix or POSIX time epoch, midnight, January 1, 1970 UTC.

Arguments may be the name of a DatePointField, TrieDateField, or date math based on a [constant date](#) or NOW.

- `ms():` Equivalent to `ms(NOW)`, number of milliseconds since the epoch.
- `ms(a):` Returns the number of milliseconds since the epoch that the argument represents.
- `ms(a,b):` Returns the number of milliseconds that b occurs before a (that is, $a - b$)

Syntax Examples

- `ms(NOW/DAY)`
- `ms(2000-01-01T00:00:00Z)`
- `ms(mydatefield)`
- `ms(NOW,mydatefield)`
- `ms(mydatefield, 2000-01-01T00:00:00Z)`
- `ms(datefield1, datefield2)`

norm(*field*) Function

Returns the "norm" stored in the index for the specified field. This is the product of the index time boost and the length normalization factor, according to the [Similarity](#) for the field.

Syntax Example

- `norm(fieldName)`

numdocs Function

Returns the number of documents in the index, not including those that are marked as deleted but have not yet been purged. This is a constant (the same value for all documents in the index).

Syntax Example

- `numdocs()`

ord Function

Returns the ordinal of the indexed field value within the indexed list of terms for that field in Lucene index order (lexicographically ordered by unicode value), starting at 1.

In other words, for a given field, all values are ordered lexicographically; this function then returns the offset of a particular value in that ordering. The field must have a maximum of one value per document (not multi-valued). 0 is returned for documents without a value in the field.



ord() depends on the position in an index and can change when other documents are inserted or deleted.

See also rord below.

Syntax Example

- `ord(myIndexedField)`
- If there were only three values ("apple","banana","pear") for a particular field X, then `ord(X)` would be 1 for documents containing "apple", 2 for documents containing "banana", etc.

payload Function

Returns the float value computed from the decoded payloads of the term specified.

The return value is computed using the `min`, `max`, or average of the decoded payloads. A special `first` function can be used instead of the others, to short-circuit term enumeration and return only the decoded payload of the first term.

The field specified must have float or integer payload encoding capability (via `DelimitedPayloadTokenFilter` or `NumericPayloadTokenFilter`). If no payload is found for the term, the default value is returned.

- `payload(field_name,term)`: default value is 0.0, average function is used.
- `payload(field_name,term,default_value)`: default value can be a constant, field name, or another float returning function. average function used.
- `payload(field_name,term,default_value,function)`: function values can be `min`, `max`, `average`, or `first`.

Syntax Example

- `payload(payloaded_field_dpf,term,0.0,first)`

pow Function

Raises the specified base to the specified power. `pow(x,y)` raises x to the power of y.

Syntax Examples

- `pow(x,y)`
- `pow(x,log(y))`
- `pow(x,0.5)`: the same as `sqrt`

product Function

Returns the product of multiple values or functions, which are specified in a comma-separated list. `mul(...)` may also be used as an alias for this function.

Syntax Examples

- `product(x,y,...)`

- product(x,2)
- mul(x,y)

query Function

Returns the score for the given subquery, or the default value for documents not matching the query. Any type of subquery is supported through either parameter de-referencing \${otherparam} or direct specification of the query string in the [Local Parameters](#) through the v key.

Syntax Examples

- query(subquery, default)
- q=product (popularity,query({!dismax v='solr rocks'})): returns the product of the popularity and the score of the DisMax query.
- q=product (popularity,query(\$qq))&qq={!dismax}solr rocks: equivalent to the previous query, using parameter de-referencing.
- q=product (popularity,query(\$qq,0.1))&qq={!dismax}solr rocks: specifies a default score of 0.1 for documents that don't match the DisMax query.

recip Function

Performs a reciprocal function with $\text{recip}(x,m,a,b)$ implementing $a/(m*x+b)$ where m, a, b are constants, and x is any arbitrarily complex function.

When a and b are equal, and $x \geq 0$, this function has a maximum value of 1 that drops as x increases. Increasing the value of a and b together results in a movement of the entire function to a flatter part of the curve. These properties can make this an ideal function for boosting more recent documents when x is rord(datefield).

Syntax Examples

- recip(myfield,m,a,b)
- recip(rord (creationDate), 1,1000,1000)

rord Function

Returns the reverse ordering of that returned by ord.

Syntax Example

- rord(myDateField)

scale Function

Scales values of the function x such that they fall between the specified minTarget and maxTarget inclusive. The current implementation traverses all of the function values to obtain the min and max, so it can pick the correct scale.

The current implementation cannot distinguish when documents have been deleted or documents that have no value. It uses 0.0 values for these cases. This means that if values are normally all greater than 0.0, one can still end up with 0.0 as the min value to map from. In these cases, an appropriate map() function could

be used as a workaround to change 0..0 to a value in the real range, as shown here:

```
scale(map(x,0,0,5),1,2)
```

Syntax Examples

- `scale(x, minTarget, maxTarget)`
- `scale(x,1,2)`: scales the values of x such that all values will be between 1 and 2 inclusive.

sqedist Function

The Square Euclidean distance calculates the 2-norm (Euclidean distance) but does not take the square root, thus saving a fairly expensive operation. It is often the case that applications that care about Euclidean distance do not need the actual distance, but instead can use the square of the distance. There must be an even number of ValueSource instances passed in and the method assumes that the first half represent the first vector and the second half represent the second vector.

Syntax Example

- `sqedist(x_td, y_td, 0, 0)`

sqrt Function

Returns the square root of the specified value or function.

Syntax Examples

- `sqrt(x)`
- `sqrt(100)`
- `sqrt(sum(x,100))`

strdist Function

Calculate the distance between two strings. Uses the Lucene spell checker StringDistance interface and supports all of the implementations available in that package, plus allows applications to plug in their own via Solr's resource loading capabilities. `strdist` takes (string1, string2, distance measure).

Possible values for distance measure are:

- `jw`: Jaro-Winkler
- `edit`: Levenshtein or Edit distance
- `ngram`: The NGramDistance, if specified, can optionally pass in the ngram size too. Default is 2.
- `FQN`: Fully Qualified class Name for an implementation of the StringDistance interface. Must have a no-arg constructor.

Syntax Example

- `strdist("SOLR",id,edit)`

sub Function

Returns $x-y$ from `sub(x,y)`.

Syntax Examples

- `sub(myfield,myfield2)`
- `sub(100, sqrt(myfield))`

sum Function

Returns the sum of multiple values or functions, which are specified in a comma-separated list. `add(...)` may be used as an alias for this function.

Syntax Examples

- `sum(x,y,...)`
- `sum(x,1)`
- `sum(sqrt(x),log(y),z,0.5)`
- `add(x,y)`

sumtotaltermfreq Function

Returns the sum of `totalTermFreq` values for all terms in the field in the entire index (i.e., the number of indexed tokens for that field). (Aliases `sumtotaltermfreq` to `sttf`.)

Syntax Example If `doc1:(fieldX:A B C)` and `doc2:(fieldX:A A A A)`:

- `docFreq(fieldX:A) = 2` (A appears in 2 docs)
- `freq(doc1, fieldX:A) = 4` (A appears 4 times in doc 2)
- `totalTermFreq(fieldX:A) = 5` (A appears 5 times across all docs)
- `sumTotalTermFreq(fieldX) = 7` in fieldX, there are 5 As, 1 B, 1 C

termfreq Function

Returns the number of times the term appears in the field for that document.

Syntax Example

- `termfreq(text,'memory')`

tf Function

Term frequency; returns the term frequency factor for the given term, using the [Similarity](#) for the field. The `tf-idf` value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the document, which helps to control for the fact that some words are generally more common than others. See also `idf`.

Syntax Examples

- `tf(text,'solr')`

top Function

Causes the function query argument to derive its values from the top-level `IndexReader` containing all parts of an index. For example, the ordinal of a value in a single segment will be different from the ordinal of that

same value in the complete index.

The `ord()` and `rord()` functions implicitly use `top()`, and hence `ord(foo)` is equivalent to `top(ord(foo))`.

totaltermfreq Function

Returns the number of times the term appears in the field in the entire index. (Aliases `totaltermfreq` to `ttf`.)

Syntax Example

- `ttf(text, 'memory')`

Boolean Functions

The following functions are boolean – they return true or false. They are mostly useful as the first argument of the `if` function, and some of these can be combined. If used somewhere else, it will yield a '1' or '0'.

and Function

Returns a value of true if and only if all of its operands evaluate to true.

Syntax Example

- `and(not(exists(popularity)), exists(price))`: returns true for any document which has a value in the `price` field, but does not have a value in the `popularity` field.

or Function

A logical disjunction.

Syntax Example

- `or(value1,value2)`: true if either `value1` or `value2` is true.

xor Function

Logical exclusive disjunction, or one or the other but not both.

Syntax Example

- `xor(field1,field2)` returns true if either `field1` or `field2` is true; FALSE if both are true.

not Function

The logically negated value of the wrapped function.

Syntax Example

- `not(exists(author))`: true only when `exists(author)` is false.

exists Function

Returns true if any member of the field exists.

Syntax Example

- exists(author): returns true for any document has a value in the "author" field.
- exists(query(price:5.00)): returns true if "price" matches "5.00".

Comparison Functions

gt, gte, lt, lte, eq

5 comparison functions: Greater Than, Greater Than or Equal, Less Than, Less Than or Equal, Equal. eq works on not just numbers but essentially any value like a string field.

Syntax Example

- if(lt(ms(mydatefield),315569259747),0.8,1) translates to this pseudocode: if mydatefield < 315569259747 then 0.8 else 1

Example Function Queries

To give you a better understanding of how function queries can be used in Solr, suppose an index stores the dimensions in meters x,y,z of some hypothetical boxes with arbitrary names stored in field boxname.

Suppose we want to search for box matching name findbox but ranked according to volumes of boxes. The query parameters would be:

```
q=boxname:findbox _val_:"product(x,y,z)"
```

This query will rank the results based on volumes. In order to get the computed volume, you will need to request the score, which will contain the resultant volume:

```
&fl=*, score
```

Suppose that you also have a field storing the weight of the box as weight. To sort by the density of the box and return the value of the density in score, you would submit the following query:

```
http://localhost:8983/solr/collection_name/select?q=boxname:findbox
_val_:"div(weight,product(x,y,z))"&fl=boxname x y z weight score
```

Sort By Function

You can sort your query results by the output of a function. For example, to sort results by distance, you could enter:

```
http://localhost:8983/solr/collection_name/select?q=*:*&sort=dist(2, point1, point2) desc
```

Sort by function also supports pseudo-fields: fields can be generated dynamically and return results as though it was normal field in the index. For example,

```
&fl=id,sum(x, y),score&wt=xml
```

would return:

```
<str name="id">foo</str>
<float name="sum(x,y)">40</float>
<float name="score">0.343</float>
```

Local Parameters in Queries

Local parameters are arguments in a Solr request that are specific to a query parameter.

Local parameters provide a way to add meta-data to certain argument types such as query strings. (In Solr documentation, local parameters are sometimes referred to as LocalParams.)

Local parameters are specified as prefixes to arguments. Take the following query argument, for example:

```
q=solr rocks
```

We can prefix this query string with local parameters to provide more information to the Standard Query Parser. For example, we can change the default operator type to "AND" and the default field to "title":

```
q={!q.op=AND df=title}solr rocks
```

These local parameters would change the query to require a match on both "solr" and "rocks" while searching the "title" field by default.

Basic Syntax of Local Parameters

To specify a local parameter, insert the following before the argument to be modified:

- Begin with { !
- Insert any number of key=value pairs separated by white space
- End with } and immediately follow with the query argument

You may specify only one local parameters prefix per argument. Values in the key-value pairs may be quoted via single or double quotes, and backslash escaping works within quoted strings.

Query Type Short Form

If a local parameter value appears without a name, it is given the implicit name of "type". This allows short-form representation for the type of query parser to use when parsing a query string. Thus

```
q={!dismax qf=myfield}solr rocks
```

is equivalent to:

```
q={!type=dismax qf=myfield}solr rocks
```

If no "type" is specified (either explicitly or implicitly) then the [lucene parser](#) is used by default. Thus

```
fq={!df=summary}solr rocks
```

is equivalent to:

```
fq={!type=lucene df=summary}solr rocks
```

Specifying the Parameter Value with the v Key

A special key of `v` within local parameters is an alternate way to specify the value of that parameter.

```
q={!dismax qf=myfield}solr rocks
```

is equivalent to

```
q={!type=dismax qf=myfield v='solr rocks'}
```

Parameter Dereferencing

Parameter dereferencing, or indirection, lets you use the value of another argument rather than specifying it directly. This can be used to simplify queries, decouple user input from query parameters, or decouple front-end GUI parameters from defaults set in `solrconfig.xml`.

```
q={!dismax qf=myfield}solr rocks
```

is equivalent to:

```
q={!type=dismax qf=myfield v=$qq}&qq=solr rocks
```

Other Parsers

In addition to the main query parsers discussed earlier, there are several other query parsers that can be used instead of or in conjunction with the main parsers for specific purposes.

This section details the other parsers, and gives examples for how they might be used.

Many of these parsers are expressed the same way as [Local Parameters in Queries](#).

Block Join Query Parsers

There are two query parsers that support block joins. These parsers allow indexing and searching for relational content that has been [indexed as Nested Documents](#).

The example usage of the query parsers below assumes these two documents and each of their child documents have been indexed:

```
<add>
  <doc>
    <field name="id">1</field>
    <field name="title">Solr has block join support</field>
    <field name="content_type">parentDocument</field>
    <doc>
      <field name="id">2</field>
      <field name="comments">SolrCloud supports it too!</field>
    </doc>
  </doc>
  <doc>
    <field name="id">3</field>
    <field name="title">New Lucene and Solr release</field>
    <field name="content_type">parentDocument</field>
    <doc>
      <field name="id">4</field>
      <field name="comments">Lots of new features</field>
    </doc>
  </doc>
</add>
```

Block Join Children Query Parser

This parser takes a query that matches some parent documents and returns their children.

The syntax for this parser is: q={!child of=<allParents>}<someParents>.

The parameter allParents is a filter that matches **only parent documents**; here you would define the field and value that you used to identify **all parent documents**.

The parameter someParents identifies a query that will match some of the parent documents. The output is the children.

Using the example documents above, we can construct a query such as q={!child of="content_type:parentDocument"}title:lucene&wt=xml. We only get one document in response:

```
<result name="response" numFound="1" start="0">
  <doc>
    <str name="id">4</str>
    <str name="comments">Lots of new features</str>
  </doc>
</result>
```

The query for someParents should match only parent documents passed by allParents or you may get an exception:



Parent query must not match any docs besides parent filter. Combine them as must (+) and must-not (-) clauses to find a problem doc.

You can search for `q=+(someParents) -(allParents)` to find a cause if you encounter this error.

Filtering and Tagging

`{!child}` also supports filters and excludeTags local parameters like the following:

```
{!child of=<allParents> filters=$parentfq  
excludeTags=certain}<someParents>&parentfq=BRAND:Foo&parentfq=NAME:Bar&parentfq={!tag=certain}CAT  
EGORY:Baz
```

This is equivalent to:

```
{!child of=<allParents>}+<someParents> +BRAND:Foo +NAME:Bar
```

Notice "\$" syntax in filters for referencing queries; comma-separated tags excludeTags allows to exclude certain queries by tagging. Overall the idea is similar to [excluding fq in facets](#). Note, that filtering is applied to the subordinate clause (`<someParents>`), and the intersection result is joined to the children.

All Children Syntax

When subordinate clause (`<someParents>`) is omitted, it's parsed as a *segmented* and *cached* filter for children documents. More precisely, `q={!child of=<allParents>}` is equivalent to `q=*: * -<allParents>`.

Block Join Parent Query Parser

This parser takes a query that matches child documents and returns their parents.

The syntax for this parser is similar: `q={!parent which=<allParents>}<someChildren>`.

The parameter `allParents` is a filter that matches **only parent documents**; here you would define the field and value that you used to identify **all parent documents**.

The parameter `someChildren` is a query that matches some or all of the child documents.

The query for `someChildren` should match only child documents or you may get an exception:



Child query must not match same docs with parent filter. Combine them as must clauses (+) to find a problem doc.

You can search for `q=+(parentFilter) +(someChildren)` to find a cause.

Again using the example documents above, we can construct a query such as q={!parent which="content_type:parentDocument"}comments:SolrCloud&wt=xml. We get this document in response:

```
<result name="response" numFound="1" start="0">
<doc>
<str name="id">1</str>
<arr name="title"><str>Solr has block join support</str></arr>
<arr name="content_type"><str>parentDocument</str></arr>
</doc>
</result>
```

Using which

A common mistake is to try to filter parents with a which filter, as in this bad example:



q={!parent which="title:join"}comments:SolrCloud

Instead, you should use a sibling mandatory clause as a filter:

q= +title:join +{!parent which="content_type:parentDocument"}comments:SolrCloud

Filtering and Tagging

The {!parent} query supports filters and excludeTags local parameters like the following:

```
{!parent which=<allParents> filters=$childfq excludeTags=certain}<someChildren>&
childfq=COLOR:Red&
childfq=SIZE:XL&
childfq={!tag=certain}PRINT:Hatched
```

This is equivalent to:

```
{!parent which=<allParents>}+<someChildren> +COLOR:Red +SIZE:XL
```

Notice the "\$" syntax in filters for referencing queries. Comma-separated tags in excludeTags allow excluding certain queries by tagging. Overall the idea is similar to [excluding fq in facets](#). Note that filtering is applied to the subordinate clause (<someChildren>) first, and the intersection result is joined to the parents.

Scoring with the Block Join Parent Query Parser

You can optionally use the score local parameter to return scores of the subordinate query. The values to use for this parameter define the type of aggregation, which are avg (average), max (maximum), min (minimum), total (sum). Implicit default is none which returns 0.0.

All Parents Syntax

When subordinate clause (<someChildren>) is omitted, it's parsed as a *segmented* and *cached* filter for all parent documents, or more precisely q={!parent which=<allParents>} is equivalent to q=<allParents>.

Boolean Query Parser

The BoolQParser creates a Lucene BooleanQuery which is a boolean combination of other queries. Sub-queries along with their typed occurrences indicate how documents will be matched and scored.

Parameters

must

A list of queries that **must** appear in matching documents and contribute to the score.

must_not

A list of queries that **must not** appear in matching documents.

should

A list of queries **should** appear in matching documents. For a BooleanQuery with no must queries, one or more should queries must match a document for the BooleanQuery to match.

filter

A list of queries that **must** appear in matching documents. However, unlike must, the score of filter queries is ignored.

Examples

```
{!bool must=foo must=bar}
```

```
{!bool filter=foo should=bar}
```

Boost Query Parser

BoostQParser extends the QParserPlugin and creates a boosted query from the input value. The main value is the query to be boosted. Parameter b is the function query to use as the boost. The query to be boosted may be of any type.

Boost Query Parser Examples

Creates a query "foo" which is boosted (scores are multiplied) by the function query log(popularity):

```
{!boost b=log(popularity)}foo
```

Creates a query "foo" which is boosted by the date boosting function referenced in ReciprocalFloatFunction:

```
{!boost b=recip(ms(NOW,mydatefield),3.16e-11,1,1)}foo
```

Collapsing Query Parser

The CollapsingQParser is really a *post filter* that provides more performant field collapsing than Solr's

standard approach when the number of distinct groups in the result set is high.

This parser collapses the result set to a single document per group before it forwards the result set to the rest of the search components. So all downstream components (faceting, highlighting, etc.) will work with the collapsed result set.

Details about using the CollapsingQParser can be found in the section [Collapse and Expand Results](#).

Complex Phrase Query Parser

The ComplexPhraseQParser provides support for wildcards, ORs, etc., inside phrase queries using Lucene's ComplexPhraseQueryParser.

Under the covers, this query parser makes use of the Span group of queries, e.g., spanNear, spanOr, etc., and is subject to the same limitations as that family of parsers.

Parameters

inOrder

Set to true to force phrase queries to match terms in the order specified. The default is true.

df

The default search field.

Examples

```
{!complexphrase inOrder=true}name:"Jo* Smith"
```

```
{!complexphrase inOrder=false}name:"(john jon jonathan~) peters*"
```

A mix of ordered and unordered complex phrase queries:

```
+_query_:"{!complexphrase inOrder=true}manu:\\"a* c*\\" +_query_:"{!complexphrase inOrder=false df=name}\\"bla* pla*\\""
```

Complex Phrase Parser Limitations

Performance is sensitive to the number of unique terms that are associated with a pattern. For instance, searching for "a*" will form a large OR clause (technically a SpanOr with many terms) for all of the terms in your index for the indicated field that start with the single letter 'a'. It may be prudent to restrict wildcards to at least two or preferably three letters as a prefix. Allowing very short prefixes may result in many low-quality documents being returned.

Notice that it also supports leading wildcards "*a" as well with consequent performance implications. Applying [ReversedWildcardFilterFactory](#) in index-time analysis is usually a good idea.

MaxBooleanClauses with Complex Phrase Parser

You may need to increase MaxBooleanClauses in solrconfig.xml as a result of the term expansion above:

```
<maxBooleanClauses>4096</maxBooleanClauses>
```

This property is described in more detail in the section [Query Sizing and Warming](#).

Stopwords with Complex Phrase Parser

It is recommended not to use stopword elimination with this query parser.

Lets say we add the terms **the**, **up**, and **to** to stopwords.txt for your collection, and index a document containing the text "*Stores up to 15,000 songs, 25,00 photos, or 150 yours of video*" in a field named "features".

While the query below does not use this parser:

```
q=features:"Stores up to 15,000"
```

the document is returned. The next query that *does* use the Complex Phrase Query Parser, as in this query:

```
q=features:"sto* up to 15*">&defType=complexphrase
```

does *not* return that document because SpanNearQuery has no good way to handle stopwords in a way analogous to PhraseQuery. If you must remove stopwords for your use case, use a custom filter factory or perhaps a customized synonyms filter that reduces given stopwords to some impossible token.

Escaping with Complex Phrase Parser

Special care has to be given when escaping: clauses between double quotes (usually whole query) is parsed twice, these parts have to be escaped as twice, e.g., "foo\\\: bar\\^".

Field Query Parser

The FieldQParser extends the QParserPlugin and creates a field query from the input value, applying text analysis and constructing a phrase query if appropriate. The parameter f is the field to be queried.

Example:

```
{!field f=myfield}Foo Bar
```

This example creates a phrase query with "foo" followed by "bar" (assuming the analyzer for myfield is a text field with an analyzer that splits on whitespace and lowercase terms). This is generally equivalent to the Lucene query parser expression myfield: "Foo Bar".

Filters Query Parser

The syntax is:

```
q={!filters param=$fqs excludeTags=sample}field:text&
fqs=COLOR:Red&
fqs=SIZE:XL&
fqs={!tag=sample}BRAND:Foo
```

which is equivalent to:

```
q=+field:text +COLOR:Red +SIZE:XL
```

param local parameter uses "\$" syntax to refer to a few queries, where excludeTags may omit some of them.

Function Query Parser

The FunctionQParser extends the QParserPlugin and creates a function query from the input value. This is only one way to use function queries in Solr; for another, more integrated, approach, see the section on [Function Queries](#).

Example:

```
{!func}log(foo)
```

Function Range Query Parser

The FunctionRangeQParser extends the QParserPlugin and creates a range query over a function. This is also referred to as frange, as seen in the examples below.

Parameters

l

The lower bound. This parameter is optional.

u

The upper bound. This parameter is optional.

incl

Include the lower bound. This parameter is optional. The default is true.

incu

Include the upper bound. This parameter is optional. The default is true.

Examples

```
{!frange l=1000 u=50000}myfield
```

```
fq={!frange l=0 u=2.2} sum(user_ranking,editor_ranking)
```

Both of these examples restrict the results by a range of values found in a declared field or a function query. In the second example, we're doing a sum calculation, and then defining only values between 0 and 2.2 should be returned to the user.

For more information about range queries over functions, see Yonik Seeley's introductory blog post [Ranges over Functions in Solr 1.4](#).

Graph Query Parser

The graph query parser does a breadth first, cyclic aware, graph traversal of all documents that are "reachable" from a starting set of root documents identified by a wrapped query.

The graph is built according to linkages between documents based on the terms found in `from` and `to` fields that you specify as part of the query.

Supported field types are point fields with `docValues` enabled, or string fields with `indexed=true` or `docValues=true`.



For string fields which are `indexed=false` and `docValues=true`, please refer to the javadocs for `DocValuesTermsQuery` for its performance characteristics so `indexed=true` will perform better for most use-cases.

Graph Query Parameters

`to`

The field name of matching documents to inspect to identify outgoing edges for graph traversal. Defaults to `edge_ids`.

`from`

The field name `to` of candidate documents to inspect to identify incoming graph edges. Defaults to `node_id`.

`traversalFilter`

An optional query that can be supplied to limit the scope of documents that are traversed.

`maxDepth`

Integer specifying how deep the breadth first search of the graph should go beginning with the initial query. Defaults to -1 (unlimited).

`returnRoot`

Boolean to indicate if the documents that matched the original query (to define the starting points for graph) should be included in the final results. Defaults to true.

`returnOnlyLeaf`

Boolean that indicates if the results of the query should be filtered so that only documents with no outgoing edges are returned. Defaults to false.

`useAutn`

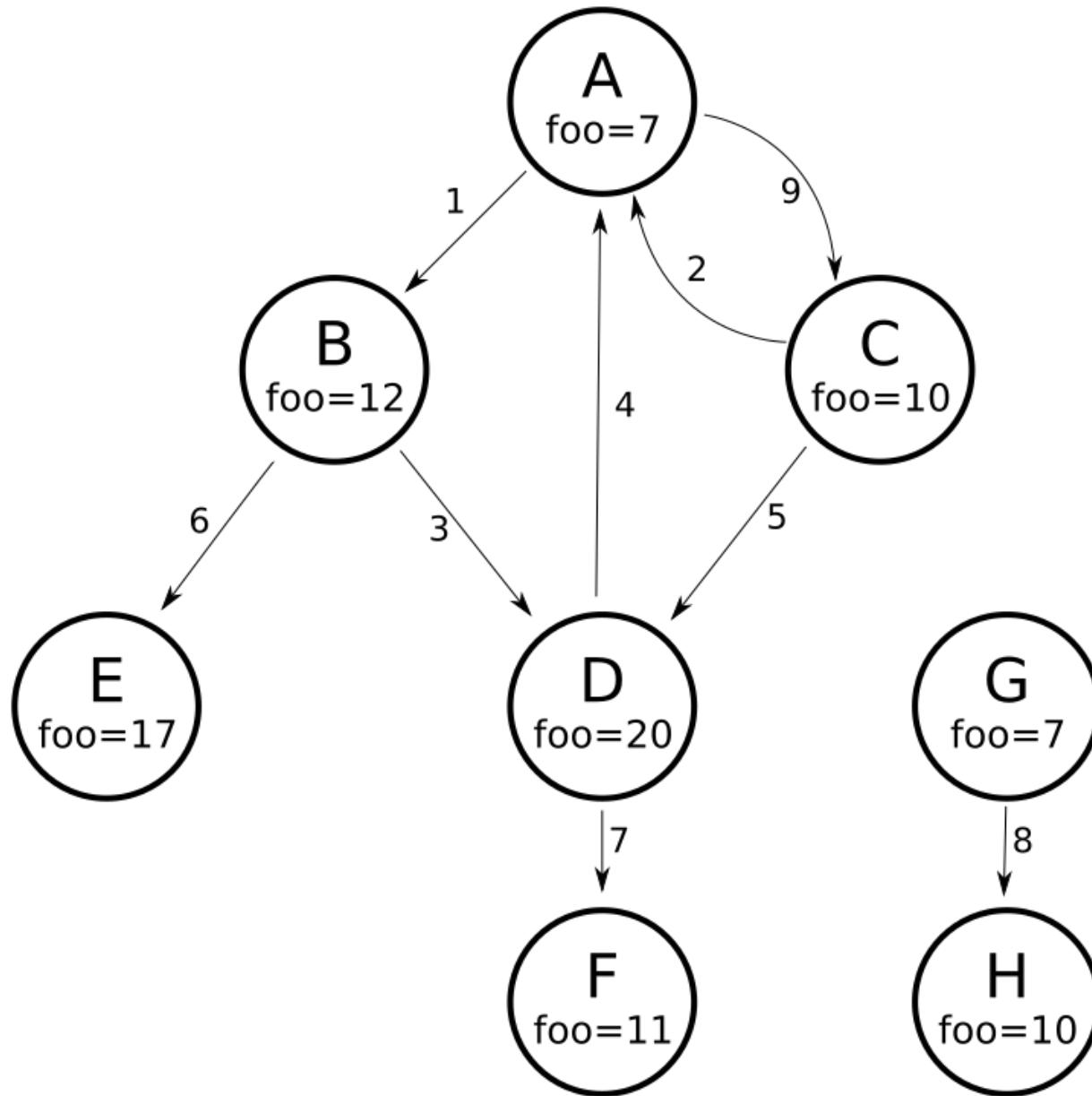
Boolean that indicates if an Automatons should be compiled for each iteration of the breadth first search, which may be faster for some graphs. Defaults to false.

Graph Query Limitations

The graph parser only works in single node Solr installations, or with [SolrCloud](#) collections that use exactly 1 shard.

Graph Query Examples

To understand how the graph parser works, consider the following Directed Cyclic Graph, containing 8 nodes (A to H) and 9 edges (1 to 9):



One way to model this graph as Solr documents, would be to create one document per node, with multivalued fields identifying the incoming and outgoing edges for each node:

```
curl -H 'Content-Type: application/json' 'http://localhost:8983/solr/my_graph/update?commit=true'
--data-binary '[{"id": "A", "foo": 7, "out_edge": ["1", "9"], "in_edge": ["4", "2"]}, {"id": "B", "foo": 12, "out_edge": ["3", "6"], "in_edge": ["1"]}, {"id": "C", "foo": 10, "out_edge": ["5", "2"], "in_edge": ["9"]}, {"id": "D", "foo": 20, "out_edge": ["4", "7"], "in_edge": ["3", "5"]}, {"id": "E", "foo": 17, "out_edge": [], "in_edge": ["6"]}, {"id": "F", "foo": 11, "out_edge": [], "in_edge": ["7"]}, {"id": "G", "foo": 7, "out_edge": ["8"], "in_edge": []}, {"id": "H", "foo": 10, "out_edge": [], "in_edge": ["8"]}]'
```

With the model shown above, the following query demonstrates a simple traversal of all nodes reachable from node A:

```
http://localhost:8983/solr/my_graph/query?fl=id&q={!graph+from=in_edge+to=out_edge}id:A
```

```
"response": {"numFound": 6, "start": 0, "docs": [{"id": "A"}, {"id": "B"}, {"id": "C"}, {"id": "D"}, {"id": "E"}, {"id": "F"}]}
```

We can also use the traversalFilter to limit the graph traversal to only nodes with maximum value of 15 in the foo field. In this case that means D, E, and F are excluded – F has a value of foo=11, but it is unreachable because the traversal skipped D:

```
http://localhost:8983/solr/my_graph/query?fl=id&q={!graph+from=in_edge+to=out_edge+traversalFilter=foo:[*+TO+15]}id:A
```

```
...
"response": {"numFound": 3, "start": 0, "docs": [{"id": "A"}, {"id": "B"}, {"id": "C"}]}
```

The examples shown so far have all used a query for a single document ("id:A") as the root node for the graph traversal, but any query can be used to identify multiple documents to use as root nodes. The next example demonstrates using the maxDepth parameter to find all nodes that are at most one edge away from an root node with a value in the foo field less than or equal to 10:

```
http://localhost:8983/solr/my_graph/query?fl=id&q={!graph+from=in_edge+to=out_edge+maxDepth=1}foo:[*+T0+10]
```

```
...
"response":{ "numFound":6, "start":0, "docs":[
  { "id":"A" },
  { "id":"B" },
  { "id":"C" },
  { "id":"D" },
  { "id":"G" },
  { "id":"H" } ]
}
```

Simplified Models

The Document & Field modeling used in the above examples enumerated all of the outgoing and income edges for each node explicitly, to help demonstrate exactly how the "from" and "to" parameters work, and to give you an idea of what is possible. With multiple sets of fields like these for identifying incoming and outgoing edges, it's possible to model many independent Directed Graphs that contain some or all of the documents in your collection.

But in many cases it can also be possible to drastically simplify the model used.

For example, the same graph shown in the diagram above can be modeled by Solr Documents that represent each node and know only the ids of the nodes they link to, without knowing anything about the incoming links:

```
curl -H 'Content-Type: application/json'
'http://localhost:8983/solr/alt_graph/update?commit=true' --data-binary '[
  {"id":"A", "foo": 7, "out_edge":["B", "C"] },
  {"id":"B", "foo": 12, "out_edge":["E", "D"] },
  {"id":"C", "foo": 10, "out_edge":["A", "D"] },
  {"id":"D", "foo": 20, "out_edge":["A", "F"] },
  {"id":"E", "foo": 17, "out_edge":[] },
  {"id":"F", "foo": 11, "out_edge":[] },
  {"id":"G", "foo": 7, "out_edge":["H"] },
  {"id":"H", "foo": 10, "out_edge":[] }
]'
```

With this alternative document model, all of the same queries demonstrated above can still be executed, simply by changing the "from" parameter to replace the "in_edge" field with the "id" field:

```
http://localhost:8983/solr/alt_graph/query?fl=id&q={!graph+from=id+to=out_edge+maxDepth=1}foo:[*+T0+10]
```

```
...
"response": {"numFound": 6, "start": 0, "docs": [
    { "id": "A" },
    { "id": "B" },
    { "id": "C" },
    { "id": "D" },
    { "id": "G" },
    { "id": "H" } ]
}
```

Join Query Parser

JoinQParser extends the QParserPlugin. It allows normalizing relationships between documents with a join operation. This is different from the concept of a join in a relational database because no information is being truly joined. An appropriate SQL analogy would be an "inner query".

Examples:

Find all products containing the word "ipod", join them against manufacturer docs and return the list of manufacturers:

```
{!join from=manu_id_s to=id}ipod
```

Find all manufacturer docs named "belkin", join them against product docs, and filter the list to only products with a price less than \$12:

```
q = {!join from=id to=manu_id_s}compName_s:Belkin
fq = price:[* TO 12]
```

The join operation is done on a term basis, so the "from" and "to" fields must use compatible field types. For example: joining between a StrField and a IntPointField will not work, likewise joining between a StrField and a TextField that uses LowerCaseFilterFactory will only work for values that are already lower cased in the string field.

Join Parser Scoring

You can optionally use the score parameter to return scores of the subordinate query. The values to use for this parameter define the type of aggregation, which are avg (average), max (maximum), min (minimum) total, or none.

Score parameter and single value numerics



Specifying score local parameter switches the join algorithm. This might have performance implication on large indices, but it's more important that this algorithm won't work for single value numeric field starting from 7.0. Users are encouraged to change field types to string and rebuild indexes during migration.

Joining Across Collections

You can also specify a `fromIndex` parameter to join with a field from another core or collection. If running in SolrCloud mode, then the collection specified in the `fromIndex` parameter must have a single shard and a replica on all Solr nodes where the collection you're joining to has a replica.

Let's consider an example where you want to use a Solr join query to filter movies by directors that have won an Oscar. Specifically, imagine we have two collections with the following fields:

movies: id, title, director_id, ...

movie_directors: id, name, has_oscar, ...

To filter movies by directors that have won an Oscar using a Solr join on the **movie_directors** collection, you can send the following filter query to the **movies** collection:

```
fq={!join from=id fromIndex=movie_directors to=director_id}has_oscar:true
```

Notice that the query criteria of the filter (`has_oscar : true`) is based on a field in the collection specified using `fromIndex`. Keep in mind that you cannot return fields from the `fromIndex` collection using join queries, you can only use the fields for filtering results in the "to" collection (**movies**).

Next, let's understand how these collections need to be deployed in your cluster. Imagine the **movies** collection is deployed to a four node SolrCloud cluster and has two shards with a replication factor of two. Specifically, the **movies** collection has replicas on the following four nodes:

node 1: movies_shard1_replica1

node 2: movies_shard1_replica2

node 3: movies_shard2_replica1

node 4: movies_shard2_replica2

To use the **movie_directors** collection in Solr join queries with the **movies** collection, it needs to have a replica on each of the four nodes. In other words, **movie_directors** must have one shard and replication factor of four:

node 1: movie_directors_shard1_replica1

node 2: movie_directors_shard1_replica2

node 3: movie_directors_shard1_replica3

node 4: movie_directors_shard1_replica4

At query time, the JoinQParser will access the local replica of the **movie_directors** collection to perform the join. If a local replica is not available or active, then the query will fail. At this point, it should be clear that since you're limited to a single shard and the data must be replicated across all nodes where it is needed, this approach works better with smaller data sets where there is a one-to-many relationship between the from collection and the to collection. Moreover, if you add a replica to the to collection, then you also need to add a replica for the from collection.

For more information about join queries, see the Solr Wiki page on [Joins](#). Erick Erickson has also written a blog post about join performance titled [Solr and Joins](#).

Lucene Query Parser

The LuceneQParser extends the QParserPlugin by parsing Solr's variant on the Lucene QueryParser syntax. This is effectively the same query parser that is used in Lucene. It uses the operators q.op, the default operator ("OR" or "AND") and df, the default field name.

Example:

```
{!lucene q.op=AND df=text}myfield:foo +bar -baz
```

For more information about the syntax for the Lucene Query Parser, see the [Classic QueryParser javadocs](#).

Learning To Rank Query Parser

The LTRQParserPlugin is a special purpose parser for reranking the top results of a simple query using a more complex ranking query which is based on a machine learnt model.

Example:

```
{!ltr model=myModel reRankDocs=100}
```

Details about using the LTRQParserPlugin can be found in the [Learning To Rank](#) section.

Max Score Query Parser

The MaxScoreQParser extends the LuceneQParser but returns the Max score from the clauses. It does this by wrapping all SHOULD clauses in a DisjunctionMaxQuery with tie=1.0. Any MUST or PROHIBITED clauses are passed through as-is. Non-boolean queries, e.g., NumericRange falls-through to the LuceneQParser parser behavior.

Example:

```
{!maxscore tie=0.01}C OR (D AND E)
```

More Like This Query Parser

MLTQParser enables retrieving documents that are similar to a given document. It uses Lucene's existing MoreLikeThis logic and also works in SolrCloud mode. The document identifier used here is the unique id value and not the Lucene internal document id. The list of returned documents excludes the queried document.

This query parser takes the following parameters:

qf

Specifies the fields to use for similarity.

`mintf`

Specifies the Minimum Term Frequency, the frequency below which terms will be ignored in the source document.

`mindf`

Specifies the Minimum Document Frequency, the frequency at which words will be ignored when they do not occur in at least this many documents.

`maxdf`

Specifies the Maximum Document Frequency, the frequency at which words will be ignored when they occur in more than this many documents.

`minwl`

Sets the minimum word length below which words will be ignored.

`maxwl`

Sets the maximum word length above which words will be ignored.

`maxqt`

Sets the maximum number of query terms that will be included in any generated query.

`maxntp`

Sets the maximum number of tokens to parse in each example document field that is not stored with TermVector support.

`boost`

Specifies if the query will be boosted by the interesting term relevance. It can be either "true" or "false".

Examples

Find documents like the document with id=1 and using the name field for similarity.

```
{!mlt qf=name}1
```

Adding more constraints to what qualifies as similar using mintf and mindf.

```
{!mlt qf=name mintf=2 mindf=3}1
```

Nested Query Parser

The NestedParser extends the QParserPlugin and creates a nested query, with the ability for that query to redefine its type via local parameters. This is useful in specifying defaults in configuration and letting clients indirectly reference them.

Example:

```
{!query defType=func v=$q1}
```

If the q1 parameter is price, then the query would be a function query on the price field. If the q1 parameter

is `\{!lucene}inStock:true}` then a term query is created from the Lucene syntax string that matches documents with `inStock=true`. These parameters would be defined in `solrconfig.xml`, in the `defaults` section:

```
<lst name="defaults">
  <str name="q1">\{!lucene}inStock:true</str>
</lst>
```

For more information about the possibilities of nested queries, see Yonik Seeley's blog post [Nested Queries in Solr](#).

Payload Query Parsers

These query parsers utilize payloads encoded on terms during indexing.

The main query, for both of these parsers, is parsed straightforwardly from the field type's query analysis into a `SpanQuery`. The generated `SpanQuery` will be either a `SpanTermQuery` or an ordered, zero slop `SpanNearQuery`, depending on how many tokens are emitted. Payloads can be encoded on terms using either the `DelimitedPayloadTokenFilter` or the `NumericPayloadTokenFilter`. The payload using parsers are:

- `PayloadScoreQParser`
- `PayloadCheckQParser`

Payload Score Parser

`PayloadScoreQParser` incorporates each matching term's numeric (integer or float) payloads into the scores.

This parser accepts the following parameters:

`f`

The field to use. This parameter is required.

`func`

The payload function. The options are: `min`, `max`, `average`, or `sum`. This parameter is required.

`operator`

A search operator. The options are `or` and `phrase`, which is the default. This defines if the search query should be an OR query or a phrase query.

`includeSpanScore`

If true, multiples the computed payload factor by the score of the original query. If false, the default, the computed payload factor is the score.

Examples

```
{!payload_score f=my_field_dpf v=some_term func=max}
```

```
{!payload_score f=payload_field func=sum operator=or}A B C
```

Payload Check Parser

PayloadCheckQParser only matches when the matching terms also have the specified payloads.

This parser accepts the following parameters:

f

The field to use (required).

payloads

A space-separated list of payloads that must match the query terms (required)

Each specified payload will be encoded using the encoder determined from the field type and encoded accordingly for matching.

DelimitedPayloadTokenFilter 'identity' encoded payloads also work here, as well as float and integer encoded ones.

Example

```
{!payload_check f=words_dps payloads="VERB NOUN"}searching stuff
```

Prefix Query Parser

PrefixQParser extends the QParserPlugin by creating a prefix query from the input value. Currently no analysis or value transformation is done to create this prefix query.

The parameter is f, the field. The string after the prefix declaration is treated as a wildcard query.

Example:

```
{!prefix f=myfield}foo
```

This would be generally equivalent to the Lucene query parser expression myfield:foo*.

Raw Query Parser

RawQParser extends the QParserPlugin by creating a term query from the input value without any text analysis or transformation. This is useful in debugging, or when raw terms are returned from the terms component (this is not the default).

The only parameter is f, which defines the field to search.

Example:

```
{!raw f=myfield}Foo Bar
```

This example constructs the query: `TermQuery(Term("myfield", "Foo Bar"))`.

For easy filter construction to drill down in facetting, the [TermQParserPlugin](#) is recommended.

For full analysis on all fields, including text fields, you may want to use the [FieldQParserPlugin](#).

Re-Ranking Query Parser

The ReRankQParserPlugin is a special purpose parser for Re-Ranking the top results of a simple query using a more complex ranking query.

Details about using the ReRankQParserPlugin can be found in the [Query Re-Ranking](#) section.

Simple Query Parser

The Simple query parser in Solr is based on Lucene's SimpleQueryParser. This query parser is designed to allow users to enter queries however they want, and it will do its best to interpret the query and return results.

This parser takes the following parameters:

q.operators

Comma-separated list of names of parsing operators to enable. By default, all operations are enabled, and this parameter can be used to effectively disable specific operators as needed, by excluding them from the list. Passing an empty string with this parameter disables all operators.

Name	Operator	Description	Example query
AND	+	Specifies AND	token1+token2
OR		Specifies OR	token1 token2
NOT	-	Specifies NOT	-token3
PREFIX	*	Specifies a prefix query	term*
PHRASE	"	Creates a phrase	"term1 term2"
PRECEDENCE	()	Specifies precedence; tokens inside the parenthesis will be analyzed first. Otherwise, normal order is left to right.	token1 + (token2 token3)
ESCAPE	\	Put it in front of operators to match them literally	C\\

Name	Operator	Description	Example query
WHITESPACE	space or [\r\t\n]	<p>Delimits tokens on whitespace. If not enabled, whitespace splitting will not be performed prior to analysis – usually most desirable.</p> <p>Not splitting whitespace is a unique feature of this parser that enables multi-word synonyms to work. However, it probably actually won't unless synonyms are configured to normalize instead of expand to all that match a given synonym. Such a configuration requires normalizing synonyms at both index time and query time. Solr's analysis screen can help here.</p>	term1 term2
FUZZY	~ ~N	<p>At the end of terms, specifies a fuzzy query.</p> <p>"N" is optional and may be either "1" or "2" (the default)</p>	term~1
NEAR	~N	At the end of phrases, specifies a NEAR query	"term1 term2"~5

q.op

Defines the default operator to use if none is defined by the user. Allowed values are AND and OR. OR is used if none is specified.

qf

A list of query fields and boosts to use when building the query.

df

Defines the default field if none is defined in the Schema, or overrides the default field if it is already defined.

Any errors in syntax are ignored and the query parser will interpret queries as best it can. However, this can lead to odd results in some cases.

Spatial Query Parsers

There are two spatial QParsers in Solr: geofilt and bbox. But there are other ways to query spatially: using the range parser with a distance function, using the standard (lucene) query parser with the range syntax to pick the corners of a rectangle, or with RPT and BBoxField you can use the standard query parser but use a special syntax within quotes that allows you to pick the spatial predicate.

All these options are documented further in the section [Spatial Search](#).

Surround Query Parser

The SurroundQParser enables the Surround query syntax, which provides proximity search functionality.

There are two positional operators: w creates an ordered span query and n creates an unordered one. Both operators take a numeric value to indicate distance between two terms. The default is 1, and the maximum is 99.

Note that the query string is not analyzed in any way.

Example:

```
{!surround} 3w(foo, bar)
```

This example finds documents where the terms "foo" and "bar" are no more than 3 terms away from each other (i.e., no more than 2 terms between them).

This query parser will also accept boolean operators (AND, OR, and NOT, in either upper- or lowercase), wildcards, quoting for phrase searches, and boosting. The w and n operators can also be expressed in upper- or lowercase.

The non-unary operators (everything but NOT) support both infix (a AND b AND c) and prefix AND(a, b, c) notation.

Switch Query Parser

SwitchQParser is a QParserPlugin that acts like a "switch" or "case" statement.

The primary input string is trimmed and then prefixed with case. for use as a key to lookup a "switch case" in the parser's local params. If a matching local param is found the resulting param value will then be parsed as a subquery, and returned as the parse result.

The case local param can be optionally be specified as a switch case to match missing (or blank) input strings. The default local param can optionally be specified as a default case to use if the input string does not match any other switch case local params. If default is not specified, then any input which does not match a switch case local param will result in a syntax error.

In the examples below, the result of each query is "XXX":

```
{!switch case.foo=XXX case.bar=zzz case.yak=qqq}foo
```

The extra whitespace between } and bar is trimmed automatically.

```
{!switch case.foo=qqq case.bar=XXX case.yak=zzz} bar
```

The result will fallback to the default.

```
{!switch case.foo=qqq case.bar=zzz default=XXX}asdf
```

No input uses the value for case instead.

```
{!switch case=XXX case.bar=zzz case.yak=qqq}
```

A practical usage of this parser, is in specifying appends filter query (fq) parameters in the configuration of a SearchHandler, to provide a fixed set of filter options for clients using custom parameter names.

Using the example configuration below, clients can optionally specify the custom parameters in_stock and shipping to override the default filtering behavior, but are limited to the specific set of legal values (shipping=any|free, in_stock=yes|no|all).

```
<requestHandler name="/select" class="solr.SearchHandler">
  <lst name="defaults">
    <str name="in_stock">yes</str>
    <str name="shipping">any</str>
  </lst>
  <lst name="appends">
    <str name="fq">{!switch case.all='*:*'
      case.yes='inStock:true'
      case.no='inStock:false'
      v=$in_stock}</str>
    <str name="fq">{!switch case.any='*:*'
      case.free='shipping_cost:0.0'
      v=$shipping}</str>
  </lst>
</requestHandler>
```

Term Query Parser

TermQParser extends the QParserPlugin by creating a single term query from the input value equivalent to readableToIndexed(). This is useful for generating filter queries from the external human readable terms returned by the faceting or terms components. The only parameter is f, for the field.

Example:

```
{!term f=weight}1.5
```

For text fields, no analysis is done since raw terms are already returned from the faceting and terms components. To apply analysis to text fields as well, see the [Field Query Parser](#), above.

If no analysis or transformation is desired for any type of field, see the [Raw Query Parser](#), above.

Terms Query Parser

TermsQParser functions similarly to the [Term Query Parser](#) but takes in multiple values separated by commas and returns documents matching any of the specified values.

This can be useful for generating filter queries from the external human readable terms returned by the faceting or terms components, and may be more efficient in some cases than using the [Standard Query Parser](#) to generate a boolean query since the default implementation method avoids scoring.

This query parser takes the following parameters:

f

The field on which to search. This parameter is required.

separator

Separator to use when parsing the input. If set to " " (a single blank space), will trim additional white space from the input terms. Defaults to a comma (,).

method

The internal query-building implementation: termsFilter, booleanQuery, automaton, or docValuesTermsFilter. Defaults to termsFilter.

Examples

```
{!terms f=tags}software,apache,solr,lucene
```

```
{!terms f=categoryId method=booleanQuery separator=" "}8 6 7 5309
```

XML Query Parser

The [XmlQParserPlugin](#) extends the [QParserPlugin](#) and supports the creation of queries from XML. Example:

Parameter	Value
defType	xmlparser
q	<pre><BooleanQuery fieldName="description"> <Clause occurs="must"> <TermQuery>shirt</TermQuery> </Clause> <Clause occurs="mustnot"> <TermQuery>plain</TermQuery> </Clause> <Clause occurs="should"> <TermQuery>cotton</TermQuery> </Clause> <Clause occurs="must"> <BooleanQuery fieldName="size"> <Clause occurs="should"> <TermsQuery>S M L</TermsQuery> </Clause> </BooleanQuery> </Clause> </BooleanQuery></pre>

The XmlQParser implementation uses the [SolrCoreParser](#) class which extends Lucene's [CoreParser](#) class. XML elements are mapped to [QueryBuilder](#) classes as follows:

XML element	QueryBuilder class
<BooleanQuery>	BooleanQueryBuilder
<BoostingTermQuery>	BoostingTermBuilder
<ConstantScoreQuery>	ConstantScoreQueryBuilder
<DisjunctionMaxQuery>	DisjunctionMaxQueryBuilder
<MatchAllDocsQuery>	MatchAllDocsQueryBuilder
<RangeQuery>	RangeQueryBuilder
<SpanFirst>	SpanFirstBuilder
<SpanNear>	SpanNearBuilder
<SpanNot>	SpanNotBuilder
<SpanOr>	SpanOrBuilder
<SpanOrTerms>	SpanOrTermsBuilder
<SpanTerm>	SpanTermBuilder
<TermQuery>	TermQueryBuilder
<TermsQuery>	TermsQueryBuilder
<UserQuery>	UserInputQueryBuilder
<LegacyNumericRangeQuery>	LegacyNumericRangeQuery(Builder) is deprecated

Customizing XML Query Parser

You can configure your own custom query builders for additional XML elements. The custom builders need to extend the [SolrQueryBuilder](#) or the [SolrSpanQueryBuilder](#) class. Example solrconfig.xml snippet:

```
<queryParser name="xmlparser" class="XmlQParserPlugin">
  <str name="MyCustomQuery">com.mycompany.solr.search.MyCustomQueryBuilder</str>
</queryParser>
```

JSON Request API

Solr supports an alternate request API which accepts requests composed in part or entirely of JSON objects. This alternate API can be preferable in some situations, where its increased readability and flexibility make it easier to use than the entirely query-param driven alternative. There is also some functionality which can only be accessed through this JSON request API, such as much of the analytics capabilities of [JSON Faceting](#)

Building JSON Requests

The core of the JSON Request API is its ability to specify request parameters as JSON in the request body, as shown by the example below:

curl

```
curl http://localhost:8983/solr/techproducts/query -d '  
{  
    "query" : "memory",  
    "filter" : "inStock:true"  
}'
```

Solrj

```
final JsonQueryRequest simpleQuery = new JsonQueryRequest()  
    .setQuery("memory")  
    .withFilter("inStock:true");  
QueryResponse queryResponse = simpleQuery.process(solrClient, COLLECTION_NAME);
```

JSON objects are typically sent in the request body, but they can also be sent as values for json-prefixed query parameters. This can be used to override or supplement values specified in the request body. For example the query parameter `json.limit=5` will override any `limit` value provided in the JSON request body. You can also specify the entire JSON body in a single `json` query parameter, as shown in the example below:

```
curl http://localhost:8983/solr/techproducts/query -d 'json={"query": "memory"}'
```

JSON Parameter Merging

If multiple json parameters are provided in a single request, Solr attempts to merge the parameter values together before processing the request.

The JSON Request API has several properties (`filter`, `fields`, etc) which accept multiple values. During the merging process, all values for these "multivalued" properties are retained. Many properties though (`query`, `limit`, etc.) can have only a single value. When multiple parameter values conflict with one another a single value is chosen based on the following precedence rules:

- Traditional query parameters (q, rows, etc.) take first precedence and are used over any other specified values.
- json-prefix query parameters are considered next.
- Values specified in the JSON request body have the lowest precedence and are only used if specified nowhere else.

This layered merging gives the best of both worlds. Requests can be specified using readable, structured JSON. But users also have the flexibility to separate out parts of the request that change often. This can be seen at work in the following example, which combines json.-style parameters to override and supplement values found in the main JSON body.

curl

```
curl  
'http://localhost:8983/solr/techproducts/query?json.limit=5&json.filter="cat:electronics"' -d  
'  
{  
  query: "memory",  
  limit: 10,  
  filter: "inStock:true"  
'
```

SolrJ

```
final ModifiableSolrParams overrideParams = new ModifiableSolrParams();  
final JsonQueryRequest queryWithParamOverrides = new JsonQueryRequest(overrideParams)  
    .setQuery("memory")  
    .setLimit(10)  
    .withFilter("inStock:true");  
overrideParams.set("json.limit", 5);  
overrideParams.add("json.filter", "\"cat:electronics\"");  
QueryResponse queryResponse = queryWithParamOverrides.process(solrClient, COLLECTION_NAME);
```

This is equivalent to:

curl

```
curl http://localhost:8983/solr/techproducts/query -d '  
{  
    "query": "memory",  
    "limit": 5,      // this single-valued parameter was overwritten.  
    "filter": ["inStock:true","cat:electronics"]    // this multi-valued parameter was appended  
    to.  
}'
```

SolrJ

```
final JsonQueryRequest query = new JsonQueryRequest()  
    .setQuery("memory")  
    .setLimit(5)  
    .withFilter("inStock:true")  
    .withFilter("cat:electronics");  
QueryResponse queryResponse = query.process(solrClient, COLLECTION_NAME);
```

Similarly, smart merging can be used to create JSON API requests which have no proper request body at all, such as the example below:

curl

```
curl http://localhost:8983/solr/techproducts/query -d 'q=*&rows=1&  
json.facet.avg_price="avg(price)"&  
json.facet.top_cats={type:terms,field:"cat",limit:3}'
```

SolrJ

```
final ModifiableSolrParams params = new ModifiableSolrParams();  
final SolrQuery query = new SolrQuery("*:*");  
query.setRows(1);  
query.setParam("json.facet.avg_price", "\\"avg(price)\\\"");  
query.setParam("json.facet.top_cats", "{type:terms,field:\\\"cat\\\",limit:3}");  
QueryResponse queryResponse = solrClient.query(COLLECTION_NAME, query);
```

That is equivalent to the following request:

curl

```
curl http://localhost:8983/solr/techproducts/query -d '  
{  
  "query": "*:*",  
  "limit": 1,  
  "facet": {  
    "avg_price": "avg(price)",  
    "top_cats": {  
      "type": "terms",  
      "field": "cat",  
      "limit": 5  
    }  
  }  
}'
```

SolrJ

```
final JsonQueryRequest jsonQueryRequest = new JsonQueryRequest()  
  .setQuery("*:*")  
  .setLimit(1)  
  .withStatFacet("avg_price", "avg(price)")  
  .withFacet("top_cats", new TermsFacetMap("cat").setLimit(3));  
QueryResponse queryResponse = jsonQueryRequest.process(solrClient, COLLECTION_NAME);
```

See the [JSON Facet API](#) for more on faceting and analytics commands.

Supported Properties and Syntax

Right now, only some of Solr's traditional query parameters have first class JSON equivalents. Those that do are shown in the table below:

Table 1. Standard query parameters to JSON field

Query parameters	JSON field equivalent
q	query
fq	filter
start	offset
rows	limit
f1	fields
sort	sort
json.facet	facet

Query parameters	JSON field equivalent
json.<param_name>	<param_name>

Parameters not specified in the table above can still be used in the main body of JSON API requests, but they must be put within a `params` block as shown in the example below.

curl

```
curl "http://localhost:8983/solr/techproducts/query?fl=name,price"-d '
{
  params: {
    q: "memory",
    rows: 1
  }
}'
```

SolrJ

```
final ModifiableSolrParams params = new ModifiableSolrParams();
params.set("fl", "name", "price");
final JsonQueryRequest simpleQuery = new JsonQueryRequest(params)
  .withParam("q", "memory")
  .withParam("rows", 1);
QueryResponse queryResponse = simpleQuery.process(solrClient, COLLECTION_NAME);
```

Parameters placed in a `params` block act as if they were added verbatim to the `query-parameters` of the request. The request above is equivalent to:

```
curl "http://localhost:8983/solr/techproducts/query?fl=name,price&q=memory&rows=1"
```

Parameter Substitution / Macro Expansion

Of course request templating via parameter substitution works fully with JSON request bodies or parameters as well. For example:

curl

```
curl "http://localhost:8983/solr/techproducts/query?FIELD=text&TERM=memory" -d '
{
  query:"${FIELD}:${TERM}",
}'
```

SolrJ

```
final ModifiableSolrParams params = new ModifiableSolrParams();
params.set("FIELD", "text");
params.set("TERM", "memory");
final JsonQueryRequest simpleQuery = new JsonQueryRequest(params)
    .setQuery("${FIELD}:${TERM}");
QueryResponse queryResponse = simpleQuery.process(solrClient, COLLECTION_NAME);
```

JSON Extensions

Solr uses the **Noggit** JSON parser in its request API. Noggit is capable of more relaxed JSON parsing, and allows a number of deviations from the JSON standard:

- bare words can be left unquoted
- single line comments can be inserted using either // or #
- Multi-line ("C style") comments can be inserted using /* and */
- strings can be single-quoted
- special characters can be backslash-escaped
- trailing (extra) commas are silently ignored (e.g., [9, 4, 3,])
- nbsp (non-break space, \u00a0) is treated as whitespace.

Debugging

If you want to see what your merged/parsed JSON looks like, you can turn on debugging (debug=timing), and it will come back under the "json" key along with the other debugging information. Note that debug=true and debugQuery=true can often have significant performance implications, and should be reserved for debugging. Also note that debug=query has no effect on JSON facets in SolrCloud.

JSON Query DSL

Queries and filters provided in JSON requests can be specified using a rich, powerful query DSL.

Query DSL Structure

The JSON Request API accepts query values in three different formats:

- A valid **query string** that uses the default deftype (lucene, in most cases). e.g., title:solr.
- A valid **local parameters query string** that specifies its deftype explicitly. e.g., {!dismax qf=title}solr.
- A valid JSON object with the name of the query parser and any relevant parameters. e.g., { "lucene": {"df": "title", "query": "solr"} }.
 - The top level "query" JSON block generally only has a single property representing the name of the query parser to use. The value for the query parser property is a child block containing any relevant parameters as JSON properties. The whole structure is analogous to a "local-params" query string.

The query itself (often represented in local params using the name v) is specified with the key query instead.

All of these syntaxes can be used to specify queries for either the JSON Request API's query or filter properties.

Query DSL Examples

The examples below show how to use each of the syntaxes discussed above to represent a query. Each snippet represents the same basic search: the term iPod in a field called name:

1. Using the standard query API, with a simple query string

curl

```
curl -X GET "http://localhost:8983/solr/techproducts/query?q=name:iPod"
```

SolrJ

```
final SolrQuery query = new SolrQuery("name:iPod");
final QueryResponse response = solrClient.query(COLLECTION_NAME, query);
```

2. Using the JSON Request API, with a simple query string

curl

```
curl -X POST http://localhost:8983/solr/techproducts/query -d '
{
  "query" : "name:iPod"
}'
```

SolrJ

```
final JsonQueryRequest query = new JsonQueryRequest()
  .setQuery("name:iPod");
final QueryResponse response = query.process(solrClient, COLLECTION_NAME);
```

3. Using the JSON Request API, with a local-params string

curl

```
curl -X POST http://localhost:8983/solr/techproducts/query -d '  
{  
  "query": "{!lucene df=name v=iPod}"  
}'
```

SolrJ

```
final JsonQueryRequest query = new JsonQueryRequest()  
    .setQuery("{!lucene df=name}iPod");  
final QueryResponse response = query.process(solrClient, COLLECTION_NAME);
```

4. Using the JSON Request API, with a fully expanded JSON object

curl

```
curl -X POST http://localhost:8983/solr/techproducts/query -d '  
{  
  "query": {  
    "lucene": {  
      "df": "name",  
      "query": "iPod"  
    }  
  }  
}'
```

SolrJ

```
final Map<String, Object> queryTopLevel = new HashMap<>();  
final Map<String, Object> luceneQueryProperties = new HashMap<>();  
queryTopLevel.put("lucene", luceneQueryProperties);  
luceneQueryProperties.put("df", "name");  
luceneQueryProperties.put("query", "iPod");  
final JsonQueryRequest query = new JsonQueryRequest()  
    .setQuery(queryTopLevel);  
final QueryResponse response = query.process(solrClient, COLLECTION_NAME);
```

Nested Queries

Many of Solr's query parsers allow queries to be nested within one another. When these are used, requests

using the standard query API quickly become hard to write, read, and understand. These sorts of queries are often much easier to work with in the JSON Request API.

Nested Boost Query Example

As an example, consider the three requests below, which wrap a simple query (the term iPod in the field name) within a boost query:

1. Using the standard query API.

curl

```
curl -X GET "http://localhost:8983/solr/techproducts/query?q={!boost b=log(popularity) v='!{!lucene df=name}iPod'}"
```

SolrJ

```
final SolrQuery query = new SolrQuery("{!boost b=log(popularity) v='!{!lucene df=name}iPod'}");
final QueryResponse response = solrClient.query(LOCATION_NAME, query);
```

2. Using the JSON Request API, with a mix of fully-expanded and local-params queries. As you can see, the special key v is replaced with the key query.

curl

```
curl -X POST http://localhost:8983/solr/techproducts/query -d '
{
  "query" : {
    "boost": {
      "query": {!lucene df=name}iPod,
      "b": "log(popularity)"
    }
  }
}'
```

SolrJ

```
final Map<String, Object> queryTopLevel = new HashMap<>();
final Map<String, Object> boostQuery = new HashMap<>();
queryTopLevel.put("boost", boostQuery);
boostQuery.put("b", "log(popularity)");
boostQuery.put("query", "{!lucene df=name}iPod");
final JsonQueryRequest query = new JsonQueryRequest()
    .setQuery(queryTopLevel);
final QueryResponse response = query.process(solrClient, COLLECTION_NAME);
```

3. Using the JSON Request API, with all queries fully expanded as JSON

curl

```
curl -X POST http://localhost:8983/solr/techproducts/query -d '
{
  "query": {
    "boost": {
      "query": {
        "lucene": {
          "df": "name",
          "query": "iPod"
        }
      },
      "b": "log(popularity)"
    }
  }
}'
```

SolrJ

```
final Map<String, Object> queryTopLevel = new HashMap<>();
final Map<String, Object> boostProperties = new HashMap<>();
final Map<String, Object> luceneTopLevel = new HashMap<>();
final Map<String, Object> luceneProperties = new HashMap<>();
queryTopLevel.put("boost", boostProperties);
boostProperties.put("b", "log(popularity)");
boostProperties.put("query", luceneTopLevel);
luceneTopLevel.put("lucene", luceneProperties);
luceneProperties.put("df", "name");
luceneProperties.put("query", "iPod");
final JsonQueryRequest query = new JsonQueryRequest()
    .setQuery(queryTopLevel);
final QueryResponse response = query.process(solrClient, COLLECTION_NAME);
```

Nested Boolean Query Example

Query nesting is commonly seen when combining multiple query clauses together using pseudo-boolean logic with the [BoolQParser](#).

The example below shows how the [BoolQParser](#) can be used to create powerful nested queries. In this example, a user searches for results with iPod in the field name which are *not* in the bottom half of the popularity rankings.

curl

```
curl -X POST http://localhost:8983/solr/techproducts/query -d '
{
  "query": {
    "bool": {
      "must": [
        {"lucene": {"df": "name", "query": "iPod"}}
      ],
      "must_not": [
        {"frange": {"l": "0", "u": "5", "query": "popularity"}}
      ]
    }
  }
}'
```

SolrJ

```
final Map<String, Object> queryTopLevel = new HashMap<>();
final Map<String, Object> boolProperties = new HashMap<>();
final List<Object> mustClauses = new ArrayList<>();
final List<Object> mustNotClauses = new ArrayList<>();
final Map<String, Object> frangeTopLevel = new HashMap<>();
final Map<String, Object> frangeProperties = new HashMap<>();

queryTopLevel.put("bool", boolProperties);
boolProperties.put("must", mustClauses);
mustClauses.add("name:iPod");

boolProperties.put("must_not", mustNotClauses);
frangeTopLevel.put("frange", frangeProperties);
frangeProperties.put("l", 0);
frangeProperties.put("u", 5);
frangeProperties.put("query", "popularity");
mustNotClauses.add(frangeTopLevel);

final JsonQueryRequest query = new JsonQueryRequest()
    .setQuery(queryTopLevel);
final QueryResponse response = query.process(solrClient, COLLECTION_NAME);
```

If lucene is the default query parser, the example above can be simplified to:

curl

```
curl -X POST http://localhost:8983/solr/techproducts/query -d '
{
  "query": {
    "bool": {
      "must": [
        "name:iPod"
      ],
      "must_not": "{!frange l=0 u=5}popularity"
    }
  }
}'
```

SolrJ

```
final Map<String, Object> queryTopLevel = new HashMap<>();
final Map<String, Object> boolProperties = new HashMap<>();
final List<Object> mustClauses = new ArrayList<>();
final List<Object> mustNotClauses = new ArrayList<>();
queryTopLevel.put("bool", boolProperties);
boolProperties.put("must", "name:iPod");
boolProperties.put("must_not", "{!frange l=0 u=5}popularity");

final JsonQueryRequest query = new JsonQueryRequest()
    .setQuery(queryTopLevel);
final QueryResponse response = query.process(solrClient, COLLECTION_NAME);
```

Filter Queries

The syntaxes discussed above can also be used to specify query filters (under the filter key) in addition to the main query itself.

For example, the above query can be rewritten using a filter clause as:

curl

```
curl -X POST http://localhost:8983/solr/techproducts/query -d '
{
    "query": {
        "bool": {
            "must_not": "{!frange l=0 u=5}popularity"
        }
    },
    "filter": [
        "name:iPod"
    ]
}'
```

SolrJ

```
final Map<String, Object> queryTopLevel = new HashMap<>();
final Map<String, Object> boolProperties = new HashMap<>();
queryTopLevel.put("bool", boolProperties);
boolProperties.put("must_not", "{!frange l=0 u=5}popularity");

final JsonQueryRequest query = new JsonQueryRequest()
    .setQuery(queryTopLevel)
    .withFilter("name:iPod");
final QueryResponse response = query.process(solrClient, COLLECTION_NAME);
```

Tagging in JSON Query DSL

Query and filter clauses can also be individually "tagged". Tags serve as handles for query clauses, allowing them to be referenced from elsewhere in the request. This is most commonly used by the filter-exclusion functionality offered by both [traditional](#) and [JSON](#) faceting.

Queries and filters are tagged by wrapping them in a surrounding JSON object. The name of the tag is specified as a JSON key, with the query string (or object) becoming the value associated with that key. Tag name properties are prefixed with a hash, and may include multiple tags, separated by commas. For example: {"#title,tag2,tag3": "title:solr"}. Note that unlike the rest of the JSON request API which uses lax JSON parsing rules, tags must be surrounded by double-quotes because of the leading # character. The example below creates two tagged clauses: titleTag and inStockTag.

curl

```
curl -X POST http://localhost:8983/solr/techproducts/select -d '
{
  "query": "*:*",
  "filter": [
    {
      "#titleTag": "name:Solr"
    },
    {
      "#inStockTag": "inStock:true"
    }
  ]
}'
```

SolrJ

```
final Map<String, Object> titleTaggedQuery = new HashMap<>();
titleTaggedQuery.put("#titleTag", "name:Solr");
final Map<String, Object> inStockTaggedQuery = new HashMap<>();
inStockTaggedQuery.put("#inStockTag", "inStock:true");
final JsonQueryRequest query = new JsonQueryRequest()
    .setQuery("*:*")
    .withFilter(titleTaggedQuery)
    .withFilter(inStockTaggedQuery);
final QueryResponse response = query.process(solrClient, COLLECTION_NAME);
```

Note that the tags created in the example above have no impact in how the search is executed. Tags will not affect a query unless they are referenced by some other part of the request that uses them.

JSON Facet API

Facet & Analytics Module

The JSON Faceting module exposes similar functionality to Solr's traditional faceting module but with a stronger emphasis on usability. It has several benefits over traditional faceting:

- easier programmatic construction of complex or nested facets
- the nesting and structure offered by JSON makes facets easier to read and understand than the flat namespace of the traditional faceting API.
- first class support for metrics and analytics
- more standardized response format makes responses easier for clients to parse and use

Faceted Search

Faceted search is about aggregating data and calculating metrics about that data.

There are two main types of facets:

- Facets that partition or categorize data (the domain) into multiple **buckets**
- Facets that calculate data for a given bucket (normally a metric, statistic or analytic function)

Bucketing Facet Example

Here's an example of a bucketing facet, that partitions documents into bucket based on the `cat` field (short for category), and returns the top 3 buckets:

curl

```
curl http://localhost:8983/solr/techproducts/query -d '  
{  
  "query": "*:*",  
  "facet": {  
    "categories" : {  
      "type": "terms",  
      "field": "cat",  
      "limit": 3  
    }  
  }  
}'
```

SolrJ

```
final TermsFacetMap categoryFacet = new TermsFacetMap("cat").setLimit(3);
final JsonQueryRequest request = new JsonQueryRequest()
    .setQuery("*:*")
    .withFacet("categories", categoryFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

The response below shows us that 32 documents match the default root domain. and 12 documents have cat:electronics, 4 documents have cat:currency, etc.

```
[...]
"facets": {
  "count": 32,
  "categories": {
    "buckets": [
      {
        "val": "electronics",
        "count": 12},
      {
        "val": "currency",
        "count": 4},
      {
        "val": "memory",
        "count": 3},
    ]
  }
}
```

Stat Facet Example

Stat (also called aggregation or analytic) facets are useful for displaying information derived from query results, in addition to those results themselves. For example, stat facets can be used to provide context to users on an e-commerce site looking for memory. The example below computes the average price (and other statistics) and would allow a user to gauge whether the memory stick in their cart is a good price.

curl

```
curl http://localhost:8983/solr/techproducts/query -d '
q=memory&
fq=inStock:true&
json.facet={
  "avg_price" : "avg(price)",
  "num_suppliers" : "unique(manu_exact)",
  "median_weight" : "percentile(weight,50)"
}'
```

SolrJ

```
final JsonQueryRequest request = new JsonQueryRequest()
    .setQuery("memory")
    .withFilter("inStock:true")
    .withStatFacet("avg_price", "avg(price)")
    .withStatFacet("num_suppliers", "unique(manu_exact)")
    .withStatFacet("median_weight", "percentile(weight,50)");
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

The response to the facet request above will start with documents matching the root domain (docs containing "memory" with inStock:true) followed by the requested statistics in a facets block:

```
"facets" : {
    "count" : 4,
    "avg_price" : 109.9950008392334,
    "num_suppliers" : 3,
    "median_weight" : 352.0
}
```

Types of Facets

There are 4 different types of bucketing facets, which behave in two different ways:

- "terms" and "range" facets produce multiple buckets and assign each document in the domain into one (or more) of these buckets
- "query" and "heatmap" facets always produce a single bucket which all documents in the domain belong to

Each of these facet-types are covered in detail below.

Terms Facet

A terms facet buckets the domain based on the unique values in a field.

curl

```
curl http://localhost:8983/solr/techproducts/query -d '
{
  "query": "*:*",
  "facet": {
    "categories":{
      "type": "terms",
      "field" : "cat",
      "limit" : 5
    }
  }
}'
```

SolrJ

```
final TermsFacetMap categoryFacet = new TermsFacetMap("cat").setLimit(5);
final JsonQueryRequest request = new JsonQueryRequest()
  .setQuery("*:*")
  .withFacet("categories", categoryFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

Parameter	Description
field	The field name to facet over.
offset	Used for paging, this skips the first N buckets. Defaults to 0.
limit	Limits the number of buckets returned. Defaults to 10.
sort	Specifies how to sort the buckets produced. “count” specifies document count, “index” sorts by the index (natural) order of the bucket value. One can also sort by any facet function / statistic that occurs in the bucket. The default is “count desc”. This parameter may also be specified in JSON like sort:{count:desc}. The sort order may either be “asc” or “desc”
overrequest	Number of buckets beyond the limit to internally request from shards during a distributed search. Larger values can increase the accuracy of the final “Top Terms” returned when the individual shards have very diff top terms. The default of -1 causes a heuristic to be applied based on the other options specified.

Parameter	Description
refine	If true, turns on distributed facet refining. This uses a second phase to retrieve any buckets needed for the final result from shards that did not include those buckets in their initial internal results, so that every shard contributes to every returned bucket in this facet and any sub-facets. This makes counts & stats for returned buckets exact.
overrefine	<p>Number of buckets beyond the limit to consider internally during a distributed search when determining which buckets to refine.</p> <p>Larger values can increase the accuracy of the final "Top Terms" returned when the individual shards have very diff top terms, and the current sort option can result in refinement pushing terms lower down the sorted list (ex: sort: "count asc")</p> <p>The default of -1 causes a heuristic to be applied based on other options specified.</p>
mincount	Only return buckets with a count of at least this number. Defaults to 1.
missing	A boolean that specifies if a special "missing" bucket should be returned that is defined by documents without a value in the field. Defaults to false.
numBuckets	A boolean. If true, adds "numBuckets" to the response, an integer representing the number of buckets for the facet (as opposed to the number of buckets returned). Defaults to false.
allBuckets	A boolean. If true, adds an "allBuckets" bucket to the response, representing the union of all of the buckets. For multi-valued fields, this is different than a bucket for all of the documents in the domain since a single document can belong to multiple buckets. Defaults to false.
prefix	Only produce buckets for terms starting with the specified prefix.
facet	Aggregations, metrics or nested facets that will be calculated for every returned bucket
method	<p>This parameter indicates the facet algorithm to use:</p> <ul style="list-style-type: none"> • "dv" DocValues, collect into ordinal array • "uif" UnInvertedField, collect into ordinal array • "dhash" DocValues, collect into hash - improves efficiency over high cardinality fields • "enum" TermsEnum then intersect DocSet (stream-able) • "stream" Presently equivalent to "enum" • "smart" Pick the best method for the field type (this is the default)
prelim_sort	An optional parameter for specifying an approximation of the final sort to use during initial collection of top buckets when the sort parameter is very costly .

Query Facet

The query facet produces a single bucket of documents that match the domain as well as the specified

query.

curl

```
curl http://localhost:8983/solr/techproducts/query -d '  
{  
  "query": "*:*",  
  "facet": {  
    "high_popularity": {  
      "type": "query",  
      "q": "popularity:[8 TO 10]"  
    }  
  }  
}'
```

SolrJ

```
QueryFacetMap queryFacet = new QueryFacetMap("popularity:[8 TO 10]");  
final JsonQueryRequest request = new JsonQueryRequest()  
    .setQuery("*:*")  
    .withFacet("high_popularity", queryFacet);  
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

Users may also specify sub-facets (either "bucketing" facets or metrics):

curl

```
curl http://localhost:8983/solr/techproducts/query -d '  
{  
  "query": "*:*",  
  "facet": {  
    "high_popularity": {  
      "type": "query",  
      "q": "popularity:[8 TO 10]",  
      "facet" : {  
        "average_price" : "avg(price)"  
      }  
    }  
  }  
}'
```

SolrJ

```
QueryFacetMap queryFacet = new QueryFacetMap("popularity:[8 TO 10]")
    .withStatSubFacet("average_price", "avg(price)");
final JsonQueryRequest request = new JsonQueryRequest()
    .setQuery("*:*")
    .withFacet("high_popularity", queryFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

Example response:

```
"high_popularity" : {
  "count" : 36,
  "average_price" : 36.75
}
```

Range Facet

The range facet produces multiple buckets over a date or numeric field.

curl

```
curl http://localhost:8983/solr/techproducts/query -d '
{
  "query": "*:*",
  "facet": {
    "prices": {
      "type": "range",
      "field": "price",
      "start": 0,
      "end": 100,
      "gap": 20
    }
  }
}'
```

SolrJ

```
RangeFacetMap rangeFacet = new RangeFacetMap("price", 0.0, 100.0, 20.0);
final JsonQueryRequest request = new JsonQueryRequest()
    .setQuery("*:*")
    .withFacet("prices", rangeFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

The output from the range facet above would look a bit like:

```
"prices":{  
  "buckets": [  
    {  
      "val": 0.0, // the bucket value represents the start of each range. This bucket covers 0-  
20  
      "count": 5},  
    {  
      "val": 20.0,  
      "count": 0},  
    {  
      "val": 40.0,  
      "count": 0},  
    {  
      "val": 60.0,  
      "count": 1},  
    {  
      "val": 80.0,  
      "count": 1}  
  ]  
}
```

Range Facet Parameters

Range facet parameter names and semantics largely mirror facet.range query-parameter style facetting. For example "start" here corresponds to "facet.range.start" in a facet.range command.

Parameter	Description
field	The numeric field or date field to produce range buckets from.
start	Lower bound of the ranges.
end	Upper bound of the ranges.
gap	Size of each range bucket produced.
hardend	A boolean, which if true means that the last bucket will end at "end" even if it is less than "gap" wide. If false, the last bucket will be "gap" wide, which may extend past "end".
other	This parameter indicates that in addition to the counts for each range constraint between start and end, counts should also be computed for... <ul style="list-style-type: none">• "before" all records with field values lower than lower bound of the first range• "after" all records with field values greater than the upper bound of the last range• "between" all records with field values between the start and end bounds of all ranges• "none" compute none of this information• "all" shortcut for before, between, and after

Parameter	Description
include	By default, the ranges used to compute range faceting between start and end are inclusive of their lower bounds and exclusive of the upper bounds. The “before” range is exclusive and the “after” range is inclusive. This default, equivalent to “lower” below, will not result in double counting at the boundaries. The include parameter may be any combination of the following options: <ul style="list-style-type: none"> • “lower” all gap based ranges include their lower bound • “upper” all gap based ranges include their upper bound • “edge” the first and last gap ranges include their edge bounds (i.e., lower for the first one, upper for the last one) even if the corresponding upper/lower option is not specified • “outer” the “before” and “after” ranges will be inclusive of their bounds, even if the first or last ranges already include those boundaries. • “all” shorthand for lower, upper, edge, outer
facet	Aggregations, metrics, or nested facets that will be calculated for every returned bucket

Heatmap Facet

The heatmap facet generates a 2D grid of facet counts for documents having spatial data in each grid cell.

This feature is primarily documented in the [spatial](#) section of the reference guide. The key parameters are type to specify heatmap and field to indicate a spatial RPT field. The rest of the parameter names use the same names and semantics mirroring facet.heatmap query-parameter style faceting, albeit without the “facet.heatmap.” prefix. For example geom here corresponds to facet.heatmap.geom in a facet.heatmap command.



Unlike other facets that partition the domain into buckets, heatmap facets do not currently support [Nested Facets](#).

curl

```
curl http://localhost:8983/solr/spatialdata/query -d '
{
  "query": "*:*",
  "facet": {
    "locations": {
      "type": "heatmap",
      "field": "location_srpt",
      "geom": "[\"50 20\" TO \"180 90\"]",
      "gridLevel": 4
    }
  }
}'
```

SolrJ

```
final JsonQueryRequest request = new JsonQueryRequest()
    .setQuery("*:*")
    .setLimit(0)
    .withFacet("locations", new HeatmapFacetMap("location_srpt")
        .setHeatmapFormat(HeatmapFacetMap.HeatmapFormat.INTS2D)
        .setRegionQuery("[\"50 20\" TO \"180 90\"]")
        .setGridLevel(4)
    );

```

And the facet response will look like:

```
{
  "facets": {
    "locations":{
      "gridLevel":1,
      "columns":6,
      "rows":4,
      "minX":-180.0,
      "maxX":90.0,
      "minY":-90.0,
      "maxY":90.0,
      "counts_ints2D":[[68,1270,459,5359,39456,1713],[123,10472,13620,7777,18376,6239],[88,6,3898,989,1314,255],[0,0,30,1,0,1]]
    }
  }
}
```

Stat Facet Functions

Unlike all the facets discussed so far, Aggregation functions (also called **facet functions**, **analytic functions**, or **metrics**) do not partition data into buckets. Instead, they calculate something over all the documents in the domain.

Aggregation	Example	Description
sum	sum(sales)	summation of numeric values
avg	avg(popularity)	average of numeric values
min	min(salary)	minimum value
max	max(mul(price,popularity))	maximum value
unique	unique(author)	number of unique values of the given field. Beyond 100 values it yields not exact estimate

Aggregation	Example	Description
uniqueBlock	uniqueBlock(_root_)	same as above with smaller footprint strictly for counting the number of Block Join blocks . The given field must be unique across blocks, and only singlevalued string fields are supported, docValues are recommended.
hll	hll(author)	distributed cardinality estimate via hyper-log-log algorithm
percentile	percentile(salary,50,75,99,99.9)	Percentile estimates via t-digest algorithm. When sorting by this metric, the first percentile listed is used as the sort value.
sumsq	sumsq(rent)	sum of squares of field or function
variance	variance(rent)	variance of numeric field or function
stddev	stddev(rent)	standard deviation of field or function
relatedness	relatedness('popularity:[100 TO *]', 'inStock:true')	A function for computing a relatedness score of the documents in the domain to a Foreground set, relative to a Background set (both defined as queries). This is primarily for use when building Semantic Knowledge Graphs .

Numeric aggregation functions such as avg can be on any numeric field, or on a [nested function](#) of multiple numeric fields such as avg(div(popularity,price)).

The most common way of requesting an aggregation function is as a simple String containing the expression you wish to compute:

curl

```
curl http://localhost:8983/solr/techproducts/query -d '
{
  "query": "*:*",
  "filter": [
    "price:[1.0 TO *]",
    "popularity:[0 TO 10]"
  ],
  "facet": {
    "avg_value": "avg(div(popularity,price))"
  }
}'
```

SolrJ

```
final JsonQueryRequest request = new JsonQueryRequest()
    .setQuery("*:*")
    .withFilter("price:[1.0 TO *]")
    .withFilter("popularity:[0 TO 10]")
    .withStatFacet("avg_value", "avg(div(popularity,price))");
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

An expanded form allows for [Local Parameters](#) to be specified. These may be used explicitly by some specialized aggregations such as [relatedness\(\)](#), but can also be used as parameter references to make aggregation expressions more readable, without needing to use (global) request parameters:

curl

```
curl http://localhost:8983/solr/techproducts/query -d '
{
  "query": "*:*",
  "filter": [
    "price:[1.0 TO *]",
    "popularity:[0 TO 10]"
  ],
  "facet": {
    "avg_value" : {
      "type": "func",
      "func": "avg(div($numer,$denom))",
      "numer": "mul(popularity,3.0)",
      "denom": "price"
    }
  }
}'
```

SolrJ

```
final Map<String, Object> expandedStatFacet = new HashMap<>();
expandedStatFacet.put("type", "func");
expandedStatFacet.put("func", "avg(div($numer,$denom))");
expandedStatFacet.put("numer", "mul(popularity,3.0)");
expandedStatFacet.put("denom", "price");
final JsonQueryRequest request = new JsonQueryRequest()
    .setQuery("*:*")
    .withFilter("price:[1.0 TO *]")
    .withFilter("popularity:[0 TO 10]")
    .withFacet("avg_value", expandedStatFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

Nested Facets

Nested facets, or **sub-facets**, allow one to nest facet commands under any facet command that partitions the domain into buckets (i.e., terms, range, query). These sub-facets are then evaluated against the domains defined by the set of all documents in each bucket of their parent.

The syntax is identical to top-level facets - just add a facet command to the facet command block of the parent facet. Technically, every facet command is actually a sub-facet since we start off with a single facet bucket with a domain defined by the main query and filters.

Nested Facet Example

Let's start off with a simple non-nested terms facet on the category field cat:

curl

```
curl http://localhost:8983/solr/techproducts/query -d '
{
  "query": "*:*",
  "facet": {
    "categories": {
      "type": "terms",
      "field": "cat",
      "limit": 3
    }
  }
}'
```

SolrJ

```
final TermsFacetMap categoryFacet = new TermsFacetMap("cat").setLimit(3);
final JsonQueryRequest request = new JsonQueryRequest()
    .setQuery("*:*")
    .withFacet("categories", categoryFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

The response for the facet above will show the top category and the number of documents that falls into each category bucket. Nested facets can be used to gather additional information about each bucket of documents. For example, using the nested facet below, we can find the top categories as well as who the leading manufacturer is in each category:

curl

```
curl http://localhost:8983/solr/techproducts/query -d '
{
  "query": "*:*",
  "facet": {
    "categories": {
      "type": "terms",
      "field": "cat",
      "limit": 3,
      "facet": {
        "top_manufacturer": {
          "type": "terms",
          "field": "manu_id_s",
          "limit": 1
        }
      }
    }
  }
}'
```

SolrJ

```
final TermsFacetMap topCategoriesFacet = new TermsFacetMap("cat").setLimit(3);
final TermsFacetMap topManufacturerFacet = new TermsFacetMap("manu_id_s").setLimit(1);
topCategoriesFacet.withSubFacet("top_manufacturers", topManufacturerFacet);
final JsonQueryRequest request = new JsonQueryRequest()
    .setQuery("*:*")
    .withFacet("categories", topCategoriesFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

And the response will look something like:

```
"facets":{  
    "count":32,  
    "categories":{  
        "buckets":[]  
            "val":"electronics",  
            "count":12,  
            "top_manufacturer":{  
                "buckets":[]  
                    "val":"corsair",  
                    "count":3]}},  
    {  
        "val":"currency",  
        "count":4,  
        "top_manufacturer":{  
            "buckets":[]  
                "val":"boa",  
                "count":1}}},
```

Sorting Facets By Nested Functions

The default sort for a field or terms facet is by bucket count descending. We can optionally sort ascending or descending by any facet function that appears in each bucket.

curl

```
curl http://localhost:8983/solr/techproducts/query -d '  
{  
    "query": "*:*",  
    "facet": {  
        "categories":{  
            "type" : "terms",      // terms facet creates a bucket for each indexed term in the  
            "field"  
            "field" : "cat",  
            "limit": 3,  
            "sort" : "avg_price desc",  
            "facet" : {  
                "avg_price" : "avg(price)",  
            }  
        }  
    }'  
}'
```

SolrJ

```

final TermsFacetMap topCategoriesFacet = new TermsFacetMap("cat")
    .setLimit(3)
    .withStatSubFacet("avg_price", "avg(price)")
    .setSort("avg_price desc");
final JsonQueryRequest request = new JsonQueryRequest()
    .setQuery("*:*")
    .withFacet("categories", topCategoriesFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);

```

In some situations the desired sort may be an aggregation function that is very costly to compute for every bucket. A `prelim_sort` option can be used to specify an approximation of the sort, for initially ranking the buckets to determine the top candidates (based on the limit and overrequest). Only after the top candidate buckets have been refined, will the actual sort be used.

```
{
  categories:{
    type : terms,
    field : cat,
    refine: true,
    limit: 10,
    overrequest: 100,
    prelim_sort: "sales_rank desc",
    sort : "prod_quality desc",
    facet : {
      prod_quality : "avg(div(prod(rating,sales_rank),prod(num_returns,price)))"
      sales_rank : "sum(sales_rank)"
    }
  }
}
```

Changing the Domain

As discussed above, facets compute buckets or statistics based on their "domain" of documents.

- By default, top-level facets use the set of all documents matching the main query as their domain.
- Nested "sub-facets" are computed for every bucket of their parent facet, using a domain containing all documents in that bucket.

In addition to this default behavior, domains can be also be widened, narrowed, or changed entirely. The JSON Faceting API supports modifying domains through its `domain` property. This is discussed in more detail [here](#)

Special Stat Facet Functions

Most stat facet functions (avg, sumsq, etc.) allow users to perform math computations on groups of

documents. A few functions are more involved though, and deserve an explanation of their own. These are described in more detail in the sections below.

uniqueBlock() and Block Join Counts

When a collection contains [Nested Documents](#), the `blockChildren` and `blockParent` domain changes can be useful when searching for parent documents and you want to compute stats against all of the affected children documents (or vice versa). But if you only need to know the *count* of all the blocks that exist in the current domain, a more efficient option is the `uniqueBlock()` aggregate function.

Suppose we have products with multiple SKUs, and we want to count products for each color.

```
{  
  "id": "1", "type": "product", "name": "Solr T-Shirt",  
  "_childDocuments_": [  
    { "id": "11", "type": "SKU", "color": "Red", "size": "L" },  
    { "id": "12", "type": "SKU", "color": "Blue", "size": "L" },  
    { "id": "13", "type": "SKU", "color": "Red", "size": "M" }  
  ]  
,  
{  
  "id": "2", "type": "product", "name": "Solr T-Shirt",  
  "_childDocuments_": [  
    { "id": "21", "type": "SKU", "color": "Blue", "size": "S" }  
  ]  
}
```

When searching against a set of SKU documents, we can ask for a facet on color, with a nested statistic counting all the "blocks"—aka: products:

```
color: {  
  type: terms,  
  field: color,  
  limit: -1,  
  facet: {  
    productsCount: "uniqueBlock(_root_)"  
  }  
}
```

and get:

```
color:{  
  buckets:[  
    { val:Blue, count:2, productsCount:2 },  
    { val:Red, count:2, productsCount:1 }  
  ]  
}
```

Please notice that `_root_` is an internal field added by Lucene to each child document to reference on parent one. Aggregation `uniqueBlock(_root_)` is functionally equivalent to `unique(_root_)`, but is optimized for nested documents block structure. It's recommended to define limit: -1 for `uniqueBlock` calculation, like in above example, since default value of limit parameter is 10, while `uniqueBlock` is supposed to be much faster with -1.

relatedness() and Semantic Knowledge Graphs

The `relatedness(...)` stat function allows for sets of documents to be scored relative to Foreground and Background sets of documents, for the purposes of finding ad-hoc relationships that make up a "Semantic Knowledge Graph":

At its heart, the Semantic Knowledge Graph leverages an inverted index, along with a complementary uninverted index, to represent nodes (terms) and edges (the documents within intersecting postings lists for multiple terms/nodes). This provides a layer of indirection between each pair of nodes and their corresponding edge, enabling edges to materialize dynamically from underlying corpus statistics. As a result, any combination of nodes can have edges to any other nodes materialize and be scored to reveal latent relationships between the nodes.

— Grainger et al., [The Semantic Knowledge Graph](#)

The `relatedness(...)` function is used to "score" these relationships, relative to "Foreground" and "Background" sets of documents, specified in the function params as queries.

Unlike most aggregation functions, the `relatedness(...)` function is aware of whether and how it's used in [Nested Facets](#). It evaluates the query defining the current bucket *independently* from it's parent/ancestor buckets, and intersects those documents with a "Foreground Set" defined by the foreground query *combined with the ancestor buckets*. The result is then compared to a similar intersection done against the "Background Set" (defined exclusively by background query) to see if there is a positive, or negative, correlation between the current bucket and the Foreground Set, relative to the Background Set.



While it's very common to define the Background Set as `*:*`, or some other super-set of the Foreground Query, it is not strictly required. The `relatedness(...)` function can be used to compare the statistical relatedness of sets of documents to orthogonal foreground/background queries.

relatedness() Options

When using the extended type:func syntax for specifying a `relatedness()` aggregation, an optional `min_popularity` (float) option can be used to specify a lower bound on the `foreground_popularity` and `background_popularity` values, that must be met in order for the `relatedness` score to be valid — If this `min_popularity` is not met, then the `relatedness` score will be `-Infinity`.

```
{ "type": "func",
  "func": "relatedness($fore,$back)",
  "min_popularity": 0.001,
}
```

This can be particularly useful when using a descending sorting on `relatedness()` with foreground and

background queries that are disjoint, to ensure the "top buckets" are all relevant to both sets.



When sorting on relatedness(...) requests can be processed much more quickly by adding a `prelim_sort: "count desc"` option. Increasing the overrequest can help improve the accuracy of the top buckets.

Semantic Knowledge Graph Example

Sample Documents

```
curl -sS -X POST 'http://localhost:8983/solr/gettingstarted/update?commit=true' -d '[{"id":"01", "age":15, "state":"AZ", "hobbies":["soccer", "painting", "cycling"]}, {"id":"02", "age":22, "state":"AZ", "hobbies":["swimming", "darts", "cycling"]}, {"id":"03", "age":27, "state":"AZ", "hobbies":["swimming", "frisbee", "painting"]}, {"id":"04", "age":33, "state":"AZ", "hobbies":["darts"]}, {"id":"05", "age":42, "state":"AZ", "hobbies":["swimming", "golf", "painting"]}, {"id":"06", "age":54, "state":"AZ", "hobbies":["swimming", "golf"]}, {"id":"07", "age":67, "state":"AZ", "hobbies":["golf", "painting"]}, {"id":"08", "age":71, "state":"AZ", "hobbies":["painting"]}, {"id":"09", "age":14, "state":"CO", "hobbies":["soccer", "frisbee", "skiing", "swimming", "skating"]}, {"id":"10", "age":23, "state":"CO", "hobbies":["skiing", "darts", "cycling", "swimming"]}, {"id":"11", "age":26, "state":"CO", "hobbies":["skiing", "golf"]}, {"id":"12", "age":35, "state":"CO", "hobbies":["golf", "frisbee", "painting", "skiing"]}, {"id":"13", "age":47, "state":"CO", "hobbies":["skiing", "darts", "painting", "skating"]}, {"id":"14", "age":51, "state":"CO", "hobbies":["skiing", "golf"]}, {"id":"15", "age":64, "state":"CO", "hobbies":["skating", "cycling"]}, {"id":"16", "age":73, "state":"CO", "hobbies":["painting"]}]'
```

Example Query

```
curl -sS -X POST http://localhost:8983/solr/gettingstarted/query -d 'rows=0&q=*&*:*
&back=*&*:*
&fore=age:[35 TO *]                                     ①
&json.facet={                                         ②
  hobby : {
    type : terms,
    field : hobbies,
    limit : 5,
    sort : { r1: desc },                                ③
    facet : {
      r1 : "relatedness($fore,$back)",                ④
      location : {
        type : terms,
        field : state,
        limit : 2,
        sort : { r2: desc },                                ③
        facet : {
          r2 : "relatedness($fore,$back)" ④
        }
      }
    }
  }
}'
```

- ① Use the entire collection as our "Background Set"
- ② Use a query for "age >= 35" to define our (initial) "Foreground Set"
- ③ For both the top level hobbies facet & the sub-facet on state we will be sorting on the relatedness(...) values
- ④ In both calls to the relatedness(...) function, we use [Parameter Variables](#) to refer to the previously defined fore and back queries.

The Facet Response

```

"facets":{
  "count":16,
  "hobby":{

    "buckets":[]{

      "val":"golf",
      "count":6,                                     ①
      "r1":{

        "relatedness":0.01225,
        "foreground_popularity":0.3125,            ②
        "background_popularity":0.375},             ③

      "location":{

        "buckets":[]{

          "val":"az",
          "count":3,
          "r2":{

            "relatedness":0.00496,                  ④
            "foreground_popularity":0.1875,          ⑥
            "background_popularity":0.5}},           ⑦

        {

          "val":"co",
          "count":3,
          "r2":{

            "relatedness":-0.00496,                 ⑤
            "foreground_popularity":0.125,
            "background_popularity":0.5}}]}]},           ⑧

      {
        "val":"painting",
        "count":8,                                     ①
        "r1":{

          "relatedness":0.01097,
          "foreground_popularity":0.375,
          "background_popularity":0.5},             ⑨

        "location":{

          "buckets":[]{

            ...
```

```

- ① Even though hobbies:golf has a lower total facet count than hobbies:painting, it has a higher relatedness score, indicating that relative to the Background Set (the entire collection) Golf has a stronger correlation to our Foreground Set (people age 35+) than Painting.
- ② The number of documents matching age:[35 TO \*] and hobbies:golf is 31.25% of the total number of documents in the Background Set
- ③ 37.5% of the documents in the Background Set match hobbies:golf
- ④ The state of Arizona (AZ) has a *positive* relatedness correlation with the *nested* Foreground Set (people ages 35+ who play Golf) compared to the Background Set — i.e., "People in Arizona are statistically more likely to be '35+ year old Golfers' than the country as a whole."
- ⑤ The state of Colorado (CO) has a *negative* correlation with the nested Foreground Set — i.e., "People in Colorado are statistically less likely to be '35+ year old Golfers' than the country as a whole."

- ⑥ The number documents matching `age:[35 TO *] and hobbies:golf and state:AZ` is 18.75% of the total number of documents in the Background Set
- ⑦ 50% of the documents in the Background Set match `state:AZ`

## JSON Faceting Domain Changes

Facet computation operates on a "domain" of documents. By default, this domain consists of the documents matched by the main query. For sub-facets, the domain consists of all documents placed in their bucket by the parent facet.

Users can also override the "domain" of a facet that partitions data, using an explicit `domain` attribute whose value is a JSON object that can support various options for restricting, expanding, or completely changing the original domain before the buckets are computed for the associated facet.

domain changes can only be specified on individual facets that do data partitioning — not statistical/metric facets, or groups of facets.



A `*:*` query facet with a domain change can be used to group multiple sub-facets of any type, for the purpose of applying a common domain change.

## Adding Domain Filters

The simplest example of a domain change is to specify an additional filter which will be applied to the existing domain. This can be done via the `filter` keyword in the `domain` block of the facet.

### curl

```
curl http://localhost:8983/solr/techproducts/query -d '
{
 "query": "*:*",
 "facet": {
 "categories": {
 "type": "terms",
 "field": "cat",
 "limit": 3,
 "domain": {
 "filter": "popularity:[5 TO 10]"
 }
 }
 }
}'
```

## SolrJ

```
final TermsFacetMap categoryFacet = new TermsFacetMap("cat")
 .setLimit(3)
 .withDomain(new DomainMap().withFilter("popularity:[5 TO 10]"));
final JsonQueryRequest request = new JsonQueryRequest()
 .setQuery("*:*")
 .withFacet("categories", categoryFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

The value of filter can be a single query to treat as a filter, or a JSON list of filter queries. Each query can be:

- a string containing a query in Solr query syntax.
- a reference to a request parameter containing Solr query syntax, of the form: {param: <request\_param\_name>}. The referred parameter might have 0 (absent) or many values.
  - When no values are specified, no filter is applied and no error is thrown.
  - When many values are specified, each value is parsed and used as filters.

When a filter option is combined with other domain changing options, the filtering is applied *after* the other domain changes take place.

## Filter Exclusions

Domains can also be expanded by using the excludeTags keyword to discard or ignore particular tagged query filters.

This is used in the example below to show the top two manufacturers matching a search. The search results match the filter manu\_id\_s:apple, but the computed facet discards this filter and operates a domain widened by discarding the manu\_id\_s filter.

## curl

```
curl http://localhost:8983/solr/techproducts/query -d '
{
 "query": "cat:electronics",
 "filter": "{!tag=MANU}manu_id_s:apple",
 "facet": {
 "stock": {"type": "terms", "field": "inStock", "limit": 2},
 "manufacturers": {
 "type": "terms",
 "field": "manu_id_s",
 "limit": 2,
 "domain": { "excludeTags": "MANU" }
 }
 }
}'
```

## SolrJ

```
final TermsFacetMap inStockFacet = new TermsFacetMap("inStock").setLimit(2);
final TermsFacetMap allManufacturersFacet = new TermsFacetMap("manu_id_s")
 .setLimit(2)
 .withDomain(new DomainMap().withTagsToExclude("MANU"));
final JsonQueryRequest request = new JsonQueryRequest()
 .setQuery("cat:electronics")
 .withFilter("{!tag=MANU}manu_id_s:apple")
 .withFacet("stock", inStockFacet)
 .withFacet("manufacturers", allManufacturersFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

The value of `excludeTags` can be a single string tag, an array of string tags, or comma-separated tags in the single string.

When an `excludeTags` option is combined with other domain changing options, it expands the domain *before* any other domain changes take place.

See also the section on [multi-select facetting](#).

## Arbitrary Domain Query

A query domain can be specified when you wish to compute a facet against an arbitrary set of documents, regardless of the original domain. The most common use case would be to compute a top level facet against a specific subset of the collection, regardless of the main query. But it can also be useful on nested facets when building [Semantic Knowledge Graphs](#).

Example:

## curl

```
curl http://localhost:8983/solr/techproducts/query -d '
{
 "query": "apple",
 "facet": {
 "popular_categories": {
 "type": "terms",
 "field": "cat",
 "domain": { "query": "popularity:[8 TO 10]" },
 "limit": 3
 }
 }
}'
```

## SolrJ

```
final TermsFacetMap inStockFacet = new TermsFacetMap("inStock").setLimit(2);
final TermsFacetMap popularCategoriesFacet = new TermsFacetMap("cat")
 .withDomain(new DomainMap().withQuery("popularity:[8 TO 10]"))
 .setLimit(3);
final JsonQueryRequest request = new JsonQueryRequest()
 .setQuery("apple")
 .withFacet("popular_categories", popularCategoriesFacet);
QueryResponse queryResponse = request.process(solrClient, COLLECTION_NAME);
```

The value of query can be a single query, or a JSON list of queries. Each query can be:

- a string containing a query in Solr query syntax.
- a reference to a request parameter containing Solr query syntax, of the form: {param: <request\_param\_name>}. The referred parameter might have 0 (absent) or many values.
  - When no values are specified, no error is thrown.
  - When many values are specified, each value is parsed and used as queries.



While a query domain can be combined with an additional domain filter, It is not possible to also use excludeTags, because the tags would be meaningless: The query domain already completely ignores the top-level query and all previous filters.

## Block Join Domain Changes

When a collection contains [Nested Documents](#), the blockChildren or blockParent domain options can be used transform an existing domain containing one type of document, into a domain containing the documents with the specified relationship (child or parent of) to the documents from the original domain.

Both of these options work similarly to the corresponding [Block Join Query Parsers](#) by taking in a single

String query that exclusively matches all parent documents in the collection. If blockParent is used, then the resulting domain will contain all parent documents of the children from the original domain. If blockChildren is used, then the resulting domain will contain all child documents of the parents from the original domain.

```
{
 "colors": {
 "type": "terms",
 "field": "sku_color",
 "facet": {
 "brands": {
 "type": "terms",
 "field": "product_brand",
 "domain": {
 "blockParent": "doc_type:product"
 }
 }
 }
 }
}
```

- ① This example assumes we parent documents corresponding to Products, with child documents corresponding to individual SKUs with unique colors, and that our original query was against SKU documents.
- ② The colors facet will be computed against all of the original SKU documents matching our search.
- ③ For each bucket in the colors facet, the set of all matching SKU documents will be transformed into the set of corresponding parent Product documents. The resulting brands sub-facet will count how many Product documents (that have SKUs with the associated color) exist for each Brand.

## Join Query Domain Changes

A join domain change option can be used to specify arbitrary from and to fields to use in transforming from the existing domain to a related set of documents.

This works very similar to the [Join Query Parser](#), and has the same limitations when dealing with multi-shard collections.

Example:

```
{
 "colors": {
 "type": "terms",
 "field": "sku_color",
 "facet": {
 "brands": {
 "type": "terms",
 "field": "product_brand",
 "domain": {
 "join": {
 "from": "product_id_of_this_sku",
 "to": "id"
 },
 "filter": "doc_type:product"
 }
 }
 }
 }
}
```

## Graph Traversal Domain Changes

A graph domain change option works similarly to the join domain option, but can do traversal multiple hops from the existing domain to other documents.

This works very similar to the [Graph Query Parser](#), supporting all of it's optional parameters, and has the same limitations when dealing with multi-shard collections.

Example:

```
{
 "related_brands": {
 "type": "terms",
 "field": "brand",
 "domain": {
 "graph": {
 "from": "related_product_ids",
 "to": "id",
 "maxDepth": 3
 }
 }
 }
}
```

# Faceting

Faceting is the arrangement of search results into categories based on indexed terms.

Searchers are presented with the indexed terms, along with numerical counts of how many matching documents were found for each term. Faceting makes it easy for users to explore search results, narrowing in on exactly the results they are looking for.

## General Facet Parameters

There are two general parameters for controlling facetting.

`facet`

If set to true, this parameter enables facet counts in the query response. If set to false, a blank or missing value, this parameter disables facetting. None of the other parameters listed below will have any effect unless this parameter is set to true. The default value is blank (false).

`facet.query`

This parameter allows you to specify an arbitrary query in the Lucene default syntax to generate a facet count.

By default, Solr's facetting feature automatically determines the unique terms for a field and returns a count for each of those terms. Using `facet.query`, you can override this default behavior and select exactly which terms or expressions you would like to see counted. In a typical implementation of facetting, you will specify a number of `facet.query` parameters. This parameter can be particularly useful for numeric-range-based facets or prefix-based facets.

You can set the `facet.query` parameter multiple times to indicate that multiple queries should be used as separate facet constraints.

To use facet queries in a syntax other than the default syntax, prefix the facet query with the name of the query notation. For example, to use the hypothetical `myfunc` query parser, you could set the `facet.query` parameter like so:

```
facet.query={!myfunc}name~fred
```

## Field-Value Faceting Parameters

Several parameters can be used to trigger facetting based on the indexed terms in a field.

When using these parameters, it is important to remember that "term" is a very specific concept in Lucene: it relates to the literal field/value pairs that are indexed after any analysis occurs. For text fields that include stemming, lowercasing, or word splitting, the resulting terms may not be what you expect.

If you want Solr to perform both analysis (for searching) and facetting on the full literal strings, use the `copyField` directive in your Schema to create two versions of the field: one Text and one String. Make sure both are `indexed="true"`. (For more information about the `copyField` directive, see [Documents, Fields, and Schema Design](#).)

Unless otherwise specified, all of the parameters below can be specified on a per-field basis with the syntax

of `f.<fieldname>.facet.<parameter>`

#### facet.field

The `facet.field` parameter identifies a field that should be treated as a facet. It iterates over each Term in the field and generate a facet count using that Term as the constraint. This parameter can be specified multiple times in a query to select multiple facet fields.



If you do not set this parameter to at least one field in the schema, none of the other parameters described in this section will have any effect.

#### facet.prefix

The `facet.prefix` parameter limits the terms on which to facet to those starting with the given string prefix. This does not limit the query in any way, only the facets that would be returned in response to the query.

#### facet.contains

The `facet.contains` parameter limits the terms on which to facet to those containing the given substring. This does not limit the query in any way, only the facets that would be returned in response to the query.

#### facet.contains.ignoreCase

If `facet.contains` is used, the `facet.contains.ignoreCase` parameter causes case to be ignored when matching the given substring against candidate facet terms.

#### facet.matches

If you want to only return facet buckets for the terms that match a regular expression.

#### facet.sort

This parameter determines the ordering of the facet field constraints.

There are two options for this parameter.

##### count

Sort the constraints by count (highest count first).

##### index

Return the constraints sorted in their index order (lexicographic by indexed term). For terms in the ASCII range, this will be alphabetically sorted.

The default is `count` if `facet.limit` is greater than 0, otherwise, the default is `index`. Note that the default logic is changed when [Limiting Facet with Certain Terms](#)

#### facet.limit

This parameter specifies the maximum number of constraint counts (essentially, the number of facets for a field that are returned) that should be returned for the facet fields. A negative value means that Solr will return unlimited number of constraint counts.

The default value is 100.

#### facet.offset

The `facet.offset` parameter indicates an offset into the list of constraints to allow paging.

The default value is 0.

#### facet.mincount

The facet.mincount parameter specifies the minimum counts required for a facet field to be included in the response. If a field's counts are below the minimum, the field's facet is not returned.

The default value is 0.

#### facet.missing

If set to true, this parameter indicates that, in addition to the Term-based constraints of a facet field, a count of all results that match the query but which have no facet value for the field should be computed and returned in the response.

The default value is false.

#### facet.method

The facet.method parameter selects the type of algorithm or method Solr should use when faceting a field.

The following methods are available.

##### enum

Enumerates all terms in a field, calculating the set intersection of documents that match the term with documents that match the query.

This method is recommended for faceting multi-valued fields that have only a few distinct values. The average number of values per document does not matter.

For example, faceting on a field with U.S. States such as Alabama, Alaska, ... Wyoming would lead to fifty cached filters which would be used over and over again. The filterCache should be large enough to hold all the cached filters.

##### fc

Calculates facet counts by iterating over documents that match the query and summing the terms that appear in each document.

This is currently implemented using an UnInvertedField cache if the field either is multi-valued or is tokenized (according to FieldType.isTokened()). Each document is looked up in the cache to see what terms/values it contains, and a tally is incremented for each value.

This method is excellent for situations where the number of indexed values for the field is high, but the number of values per document is low. For multi-valued fields, a hybrid approach is used that uses term filters from the filterCache for terms that match many documents. The letters fc stand for field cache.

##### fcs

Per-segment field faceting for single-valued string fields. Enable with facet.method=fcs and control the number of threads used with the threads local parameter. This parameter allows faceting to be faster in the presence of rapid index changes.

The default value is fc (except for fields using the BoolField field type and when facet.exists=true is requested) since it tends to use less memory and is faster when a field has many unique terms in the

index.

#### facet.enum.cache.minDf

This parameter indicates the minimum document frequency (the number of documents matching a term) for which the filterCache should be used when determining the constraint count for that term. This is only used with the facet.method=enum method of faceting.

A value greater than zero decreases the filterCache's memory usage, but increases the time required for the query to be processed. If you are faceting on a field with a very large number of terms, and you wish to decrease memory usage, try setting this parameter to a value between 25 and 50, and run a few tests. Then, optimize the parameter setting as necessary.

The default value is 0, causing the filterCache to be used for all terms in the field.

#### facet.exists

To cap facet counts by 1, specify facet.exists=true. This parameter can be used with facet.method=enum or when it's omitted. It can be used only on non-trie fields (such as strings). It may speed up facet counting on large indices and/or high-cardinality facet values.

#### facet.excludeTerms

If you want to remove terms from facet counts but keep them in the index, the facet.excludeTerms parameter allows you to do that.

#### facet.overrequest.count **and** facet.overrequest.ratio

In some situations, the accuracy in selecting the "top" constraints returned for a facet in a distributed Solr query can be improved by "over requesting" the number of desired constraints (i.e., facet.limit) from each of the individual shards. In these situations, each shard is by default asked for the top  $10 + (1.5 * \text{facet.limit})$  constraints.

In some situations, depending on how your docs are partitioned across your shards and what facet.limit value you used, you may find it advantageous to increase or decrease the amount of over-requesting Solr does. This can be achieved by setting the facet.overrequest.count (defaults to 10) and facet.overrequest.ratio (defaults to 1.5) parameters.

#### facet.threads

This parameter will cause loading the underlying fields used in faceting to be executed in parallel with the number of threads specified. Specify as facet.threads=N where N is the maximum number of threads used.

Omitting this parameter or specifying the thread count as 0 will not spawn any threads, and only the main request thread will be used. Specifying a negative number of threads will create up to Integer.MAX\_VALUE threads.

## Range Faceting

You can use Range Faceting on any date field or any numeric field that supports range queries. This is particularly useful for stitching together a series of range queries (as facet by query) for things like prices.

#### facet.range

The facet.range parameter defines the field for which Solr should create range facets. For example:

```
facet.range=price&facet.range=age
facet.range=lastModified_dt
facet.range.start
```

The `facet.range.start` parameter specifies the lower bound of the ranges. You can specify this parameter on a per field basis with the syntax of `f.<fieldname>.facet.range.start`. For example:

```
f.price.facet.range.start=0.0&f.age.facet.range.start=10
f.lastModified_dt.facet.range.start=NOW/DAY-30DAYS
```

```
facet.range.end
```

The `facet.range.end` specifies the upper bound of the ranges. You can specify this parameter on a per field basis with the syntax of `f.<fieldname>.facet.range.end`. For example:

```
f.price.facet.range.end=1000.0&f.age.facet.range.start=99
f.lastModified_dt.facet.range.end=NOW/DAY+30DAYS
```

```
facet.range.gap
```

The span of each range expressed as a value to be added to the lower bound. For date fields, this should be expressed using the DateMathParser [syntax](#) (such as, `facet.range.gap=%2B1DAY ... '+1DAY'`). You can specify this parameter on a per-field basis with the syntax of `f.<fieldname>.facet.range.gap`. For example:

```
f.price.facet.range.gap=100&f.age.facet.range.gap=10
f.lastModified_dt.facet.range.gap=+1DAY
```

```
facet.range.hardend
```

The `facet.range.hardend` parameter is a Boolean parameter that specifies how Solr should handle cases where the `facet.range.gap` does not divide evenly between `facet.range.start` and `facet.range.end`.

If true, the last range constraint will have the `facet.range.end` value as an upper bound. If false, the last range will have the smallest possible upper bound greater than `facet.range.end` such that the range is the exact width of the specified range gap. The default value for this parameter is false.

This parameter can be specified on a per field basis with the syntax  
`f.<fieldname>.facet.range.hardend`.

```
facet.range.include
```

By default, the ranges used to compute range facetting between `facet.range.start` and `facet.range.end` are inclusive of their lower bounds and exclusive of the upper bounds. The "before" range defined with the `facet.range.other` parameter is exclusive and the "after" range is inclusive. This default, equivalent to "lower" below, will not result in double counting at the boundaries. You can use the `facet.range.include` parameter to modify this behavior using the following options:

- `lower`: All gap-based ranges include their lower bound.
- `upper`: All gap-based ranges include their upper bound.
- `edge`: The first and last gap ranges include their edge bounds (lower for the first one, upper for the last one) even if the corresponding upper/lower option is not specified.

- outer: The "before" and "after" ranges will be inclusive of their bounds, even if the first or last ranges already include those boundaries.
- all: Includes all options: lower, upper, edge, and outer.

You can specify this parameter on a per field basis with the syntax of `f.<fieldname>.facet.range.include`, and you can specify it multiple times to indicate multiple choices.



To ensure you avoid double-counting, do not choose both lower and upper, do not choose outer, and do not choose all.

#### facet.range.other

The `facet.range.other` parameter specifies that in addition to the counts for each range constraint between `facet.range.start` and `facet.range.end`, counts should also be computed for these options:

- before: All records with field values lower than lower bound of the first range.
- after: All records with field values greater than the upper bound of the last range.
- between: All records with field values between the start and end bounds of all ranges.
- none: Do not compute any counts.
- all: Compute counts for before, between, and after.

This parameter can be specified on a per field basis with the syntax of `f.<fieldname>.facet.range.other`. In addition to the `all` option, this parameter can be specified multiple times to indicate multiple choices, but `none` will override all other options.

#### facet.range.method

The `facet.range.method` parameter selects the type of algorithm or method Solr should use for range faceting. Both methods produce the same results, but performance may vary.

#### filter

This method generates the ranges based on other `facet.range` parameters, and for each of them executes a filter that later intersects with the main query resultset to get the count. It will make use of the `filterCache`, so it will benefit of a cache large enough to contain all ranges.

#### dv

This method iterates the documents that match the main query, and for each of them finds the correct range for the value. This method will make use of `docValues` (if enabled for the field) or `fieldCache`. The `dv` method is not supported for field type `DateRangeField` or when using `group.facets`.

The default value for this parameter is `filter`.

#### Date Ranges & Time Zones

Range facetting on date fields is a common situation where the `TZ` parameter can be useful to ensure that the "facet counts per day" or "facet counts per month" are based on a meaningful definition of when a given day/month "starts" relative to a particular `TimeZone`.

For more information, see the examples in the [Working with Dates](#) section.

## facet.mincount in Range Faceting

The facet.mincount parameter, the same one as used in field faceting is also applied to range faceting. When used, no ranges with a count below the minimum will be included in the response.

## Pivot (Decision Tree) Faceting

Pivoting is a summarization tool that lets you automatically sort, count, total or average data stored in a table. The results are typically displayed in a second table showing the summarized data. Pivot faceting lets you create a summary table of the results from a faceting documents by multiple fields.

Another way to look at it is that the query produces a Decision Tree, in that Solr tells you "for facet A, the constraints/counts are X/N, Y/M, etc. If you were to constrain A by X, then the constraint counts for B would be S/P, T/Q, etc.". In other words, it tells you in advance what the "next" set of facet results would be for a field if you apply a constraint from the current facet results.

### facet.pivot

The facet.pivot parameter defines the fields to use for the pivot. Multiple facet.pivot values will create multiple "facet\_pivot" sections in the response. Separate each list of fields with a comma.

### facet.pivot.mincount

The facet.pivot.mincount parameter defines the minimum number of documents that need to match in order for the facet to be included in results. The default is 1.

Using the "bin/solr -e techproducts" example, A query URL like this one will return the data below, with the pivot faceting results found in the section "facet\_pivot":

```
http://localhost:8983/solr/techproducts/select?q=*&facet.pivot=cat,popularity,inStock
&facet.pivot=popularity,cat&facet=true&facet.field=cat&facet.limit=5&rows=0&facet.pivot.mincount=
2
```

```
{
 "facet_counts":{},
 "facet_queries":{},
 "facet_fields":{},
 "cat": [
 "electronics", 14,
 "currency", 4,
 "memory", 3,
 "connector", 2,
 "graphics card", 2
],
 "facet_dates":{},
 "facet_ranges":{},
 "facet_pivot": {
 "cat,popularity,inStock": [
 {
 "field": "cat",
 "value": "electronics",
 "count": 14,
 "pivot": [
 {
 "field": "popularity",
 "value": 6,
 "count": 5,
 "pivot": [
 {
 "field": "inStock",
 "value": true,
 "count": 5
 }
]
 }
]
 }
]
 }
}
```

## Combining Stats Component With Pivots

In addition to some of the [general local parameters](#) supported by other types of faceting, a stats local parameters can be used with facet.pivot to refer to stats.field instances (by tag) that you would like to have computed for each Pivot Constraint.

In the example below, two different (overlapping) sets of statistics are computed for each of the facet.pivot result hierarchies:

```
stats=true
stats.field={!tag=piv1,piv2 min=true max=true}price
stats.field={!tag=piv2 mean=true}popularity
facet=true
facet.pivot={!stats=piv1}cat,inStock
facet.pivot={!stats=piv2}manu,inStock
```

Results:

```
{"facet_pivot": {
 "cat,inStock": [
 {
 "field": "cat",
 "value": "electronics",
 "count": 12,
 "pivot": [
 {
 "field": "popularity",
 "value": 6,
 "count": 5
 }
]
 }
]
}}
```

```
"pivot": [{"
 "field": "inStock",
 "value": true,
 "count": 8,
 "stats": {
 "stats_fields": {
 "price": {
 "min": 74.98999786376953,
 "max": 399.0}}},
 {
 "field": "inStock",
 "value": false,
 "count": 4,
 "stats": {
 "stats_fields": {
 "price": {
 "min": 11.5,
 "max": 649.989990234375}}}},
 "stats": {
 "stats_fields": {
 "price": {
 "min": 11.5,
 "max": 649.989990234375}}},
 },
 {"
 "field": "cat",
 "value": "currency",
 "count": 4,
 "pivot": [{"
 "field": "inStock",
 "value": true,
 "count": 4,
 "stats": {
 "stats_fields": {
 "price": {
 "..."
 }
 }
 }
 }]
 },
 "manu,inStock": [{"
 "field": "manu",
 "value": "inc",
 "count": 8,
 "pivot": [{"
 "field": "inStock",
 "value": true,
 "count": 7,
 "stats": {
 "stats_fields": {
 "price": {
 "min": 74.98999786376953,
 "max": 2199.0},
 "popularity": {
 "mean": 5.857142857142857}}}
 }]
 },
 {"
 "field": "inStock",
 }]
```

```
"value":false,
"count":1,
"stats":{
 "stats_fields":{
 "price":{
 "min":479.95001220703125,
 "max":479.95001220703125},
 "popularity":{
 "mean":7.0}}}}}],
"..."}]}]}]}]}}}
```

## Combining Facet Queries And Facet Ranges With Pivot Facets

A query local parameter can be used with facet.pivot to refer to facet.query instances (by tag) that should be computed for each pivot constraint. Similarly, a range local parameter can be used with facet.pivot to refer to facet.range instances.

In the example below, two query facets are computed for h of the facet.pivot result hierarchies:

```
facet=true
facet.query={!tag=q1}manufacturedate_dt:[2006-01-01T00:00:00Z TO NOW]
facet.query={!tag=q1}price:[0 TO 100]
facet.pivot={!query=q1}cat,inStock
```

```
{"facet_counts": {
 "facet_queries": {
 "{!tag=q1}manufacturedate_dt:[2006-01-01T00:00:00Z TO NOW)": 9,
 "{!tag=q1}price:[0 TO 100)": 7
 },
 "facet_fields": {},
 "facet_dates": {},
 "facet_ranges": {},
 "facet_intervals": {},
 "facet_heatmaps": {},
 "facet_pivot": {
 "cat,inStock": [
 {
 "field": "cat",
 "value": "electronics",
 "count": 12,
 "queries": {
 "{!tag=q1}manufacturedate_dt:[2006-01-01T00:00:00Z TO NOW)": 9,
 "{!tag=q1}price:[0 TO 100)": 4
 },
 "pivot": [
 {
 "field": "inStock",
 "value": true,
 "count": 8,
 "queries": {
 "{!tag=q1}manufacturedate_dt:[2006-01-01T00:00:00Z TO NOW)": 6,
 "{!tag=q1}price:[0 TO 100)": 2
 }
 },
 ...
]
 }
]
 }
}}
```

In a similar way, in the example below, two range facets are computed for each of the facet.pivot result hierarchies:

```
facet=true
facet.range={!tag=r1}manufacturedate_dt
facet.range.start=2006-01-01T00:00:00Z
facet.range.end=NOW/YEAR
facet.range.gap=+1YEAR
facet.pivot={!range=r1}cat,inStock
```

```
{"facet_counts":{
 "facet_queries":{},
 "facet_fields":{},
 "facet_dates":{},
 "facet_ranges":{
 "manufacturedate_dt":{
 "counts":[

```

```
"2006-01-01T00:00:00Z", 9,
"2007-01-01T00:00:00Z", 0,
"2008-01-01T00:00:00Z", 0,
"2009-01-01T00:00:00Z", 0,
"2010-01-01T00:00:00Z", 0,
"2011-01-01T00:00:00Z", 0,
"2012-01-01T00:00:00Z", 0,
"2013-01-01T00:00:00Z", 0,
"2014-01-01T00:00:00Z", 0],
"gap": "+1YEAR",
"start": "2006-01-01T00:00:00Z",
"end": "2015-01-01T00:00:00Z"}},
"facet_intervals":{},
"facet_heatmaps":{},
"facet_pivot":{
"cat,inStock": [
{"field": "cat",
"value": "electronics",
"count": 12,
"ranges": {
"manufacturedate_dt": {
"counts": [
"2006-01-01T00:00:00Z", 9,
"2007-01-01T00:00:00Z", 0,
"2008-01-01T00:00:00Z", 0,
"2009-01-01T00:00:00Z", 0,
"2010-01-01T00:00:00Z", 0,
"2011-01-01T00:00:00Z", 0,
"2012-01-01T00:00:00Z", 0,
"2013-01-01T00:00:00Z", 0,
"2014-01-01T00:00:00Z", 0],
"gap": "+1YEAR",
"start": "2006-01-01T00:00:00Z",
"end": "2015-01-01T00:00:00Z"}},
"pivot": [
{"field": "inStock",
"value": true,
"count": 8,
"ranges": {
"manufacturedate_dt": {
"counts": [
"2006-01-01T00:00:00Z", 6,
"2007-01-01T00:00:00Z", 0,
"2008-01-01T00:00:00Z", 0,
"2009-01-01T00:00:00Z", 0,
"2010-01-01T00:00:00Z", 0,
"2011-01-01T00:00:00Z", 0,
"2012-01-01T00:00:00Z", 0,
"2013-01-01T00:00:00Z", 0,
"2014-01-01T00:00:00Z", 0],
"gap": "+1YEAR",
"start": "2006-01-01T00:00:00Z",
```

```
"end": "2015-01-01T00:00:00Z"}],
"..."]}]}}
```

## Additional Pivot Parameters

Although facet.pivot.mincount deviates in name from the facet.mincount parameter used by field faceting, many of the faceting parameters described above can also be used with pivot faceting:

- facet.limit
- facet.offset
- facet.sort
- facet.overrequest.count
- facet.overrequest.ratio

## Interval Faceting

Another supported form of faceting is interval faceting. This sounds similar to range faceting, but the functionality is really closer to doing facet queries with range queries. Interval faceting allows you to set variable intervals and count the number of documents that have values within those intervals in the specified field.

Even though the same functionality can be achieved by using a facet query with range queries, the implementation of these two methods is very different and will provide different performance depending on the context.

If you are concerned about the performance of your searches you should test with both options. Interval faceting tends to be better with multiple intervals for the same fields, while facet query tend to be better in environments where filter cache is more effective (static indexes for example).

This method will use [docValues](#) if they are enabled for the field, will use fieldCache otherwise.

Use these parameters for interval faceting:

`facet.interval`

This parameter Indicates the field where interval faceting must be applied. It can be used multiple times in the same request to indicate multiple fields.

`facet.interval=price&facet.interval=size`

`facet.interval.set`

This parameter is used to set the intervals for the field, it can be specified multiple times to indicate multiple intervals. This parameter is global, which means that it will be used for all fields indicated with `facet.interval` unless there is an override for a specific field. To override this parameter on a specific field you can use: `f.<fieldname>.facet.interval.set`, for example:

```
f.price.facet.interval.set=[0,10]&f.price.facet.interval.set=(10,100]
```

## Interval Syntax

Intervals must begin with either '(' or '[', be followed by the start value, then a comma (','), the end value, and finally a closing ')' or ']'.

For example:

- (1,10) -> will include values greater than 1 and lower than 10
- [1,10) -> will include values greater or equal to 1 and lower than 10
- [1,10] -> will include values greater or equal to 1 and lower or equal to 10

The initial and end values cannot be empty.

If the interval needs to be unbounded, the special character \* can be used for both, start and end, limits. When using this special character, the start syntax options (( and [), and end syntax options () and ]) will be treated the same. [\* , \*] will include all documents with a value in the field.

The interval limits may be strings but there is no need to add quotes. All the text until the comma will be treated as the start limit, and the text after that will be the end limit. For example: [Buenos Aires, New York]. Keep in mind that a string-like comparison will be done to match documents in string intervals (case-sensitive). The comparator can't be changed.

Commas, brackets and square brackets can be escaped by using \ in front of them. Whitespaces before and after the values will be omitted.

The start limit can't be grater than the end limit. Equal limits are allowed, this allows you to indicate the specific values that you want to count, like [A,A], [B,B] and [C,Z].

Interval faceting supports output key replacement described below. Output keys can be replaced in both the facet.interval parameter and in the facet.interval.set parameter. For example:

```
&facet.interval={!key=popularity}some_field
&facet.interval.set={!key=bad}[0,5]
&facet.interval.set={!key=good}[5,*]
&facet=true
```

## Local Parameters for Faceting

The [LocalParams syntax](#) allows overriding global settings. It can also provide a method of adding metadata to other parameter values, much like XML attributes.

## Tagging and Excluding Filters

You can tag specific filters and exclude those filters when faceting. This is useful when doing multi-select faceting.

Consider the following example query with faceting:

```
q=mainquery&fq=status:public&fqdoctype:pdf&facet=true&facet.field=doctype
```

Because everything is already constrained by the filter `doctype:pdf`, the `facet.field=doctype` facet command is currently redundant and will return 0 counts for everything except `doctype:pdf`.

To implement a multi-select facet for `doctype`, a GUI may want to still display the other `doctype` values and their associated counts, as if the `doctype:pdf` constraint had not yet been applied. For example:

```
==== Document Type ====
 Word (42)
 PDF (96)
 Excel(11)
 HTML (63)
```

To return counts for `doctype` values that are currently not selected, tag filters that directly constrain `doctype`, and exclude those filters when faceting on `doctype`.

```
q=mainquery&fq=status:public&fq={!tag=dt}doctype:pdf&facet=true&facet.field={!ex=dt}doctype
```

Filter exclusion is supported for all types of facets. Both the `tag` and `ex` local parameters may specify multiple values by separating them with commas.

## Changing the Output Key

To change the output key for a faceting command, specify a new name with the `key` local parameter. For example:

```
facet.field={!ex=dt key=mylabel}doctype
```

The parameter setting above causes the field facet results for the "doctype" field to be returned using the key "mylabel" rather than "doctype" in the response. This can be helpful when faceting on the same field multiple times with different exclusions.

## Limiting Facet with Certain Terms

To limit field facet with certain terms specify them comma separated with `terms` local parameter. Commas and quotes in terms can be escaped with backslash, as in `\,,`. In this case facet is calculated on a way similar to `facet.method=enum`, but ignores `facet.enum.cache.minDf`. For example:

```
facet.field={!terms='alfa,betta,with\,,with\' ,with space'}symbol
```

This local parameter overrides default logic for `facet.sort`. If `facet.sort` is omitted, facets are returned in the given terms order that might be changed with `index` and `count` values. Note: other parameters might not be fully supported when this parameter is supplied.

## Related Topics

See also [Heatmap Faceting \(Spatial\)](#).

## BlockJoin Faceting

BlockJoin facets allow you to aggregate children facet counts by their parents.

It is a common requirement that if a parent document has several children documents, all of them need to increment facet value count only once. This functionality is provided by `BlockJoinDocSetFacetComponent`, and `BlockJoinFacetComponent` just an alias for compatibility.

This functionality is considered deprecated. Users are encouraged to use `uniqueBlock(_root_)` aggregation under a terms facet in the [JSON Facet API](#). If this component is used, it must be explicitly enabled for a request handler in `solrconfig.xml`, in the same way as any other [search component](#).

This example shows how you could add this search components to `solrconfig.xml` and define it in request handler:

```
<searchComponent name="bjqFacetComponent" class=
"org.apache.solr.search.join.BlockJoinDocSetFacetComponent"/>

<requestHandler name="/bjqfacet" class="org.apache.solr.handler.component.SearchHandler">
<lst name="defaults">
<str name="shards.qt">/bjqfacet</str>
</lst>
<arr name="last-components">
<str>bjqlist</str>
</arr>
</requestHandler>
```

This component can be added into any search request handler. This component work with distributed search in SolrCloud mode.

Documents should be added in children-parent blocks as described in [indexing nested child documents](#). Examples:

### Sample document

```
<add>
 <doc>
 <field name="id">1</field>
 <field name="type_s">parent</field>
 <doc>
 <field name="id">11</field>
 <field name="COLOR_s">Red</field>
 <field name="SIZE_s">XL</field>
 <field name="PRICE_i">6</field>
 </doc>
 <doc>
 <field name="id">12</field>
 <field name="COLOR_s">Red</field>
 <field name="SIZE_s">XL</field>
 <field name="PRICE_i">7</field>
 </doc>
 <doc>
 <field name="id">13</field>
 <field name="COLOR_s">Blue</field>
 <field name="SIZE_s">L</field>
 <field name="PRICE_i">5</field>
 </doc>
 </doc>
 <doc>
 <field name="id">2</field>
 <field name="type_s">parent</field>
 <doc>
 <field name="id">21</field>
 <field name="COLOR_s">Blue</field>
 <field name="SIZE_s">XL</field>
 <field name="PRICE_i">6</field>
 </doc>
 <doc>
 <field name="id">22</field>
 <field name="COLOR_s">Blue</field>
 <field name="SIZE_s">XL</field>
 <field name="PRICE_i">7</field>
 </doc>
 <doc>
 <field name="id">23</field>
 <field name="COLOR_s">Red</field>
 <field name="SIZE_s">L</field>
 <field name="PRICE_i">5</field>
 </doc>
 </doc>
</add>
```

Queries are constructed the same way as for a [Parent Block Join query](#). For example:

```
http://localhost:8983/solr/bjqfacet?q={!parent
which=type_s:parent}SIZE_s:XL&child.facet.field=COLOR_s
```

As a result we should have facets for Red(1) and Blue(1), because matches on children id=11 and id=12 are aggregated into single hit into parent with id=1.

The key components of the request shown above are:

/bjqfacet?

The name of the request handler that has been defined with a block join facet component enabled.

q={!parent which=type\_s:parent}SIZE\_s:XL

The mandatory parent query as a main query. The parent query could also be a subordinate clause in a more complex query.

&child.facet.field=COLOR\_s

The child document field, which might be repeated many times with several fields, as necessary.

# Highlighting

Highlighting in Solr allows fragments of documents that match the user's query to be included with the query response.

The fragments are included in a special section of the query response (the highlighting section), and the client uses the formatting clues also included to determine how to present the snippets to users. Fragments are a portion of a document field that contains matches from the query and are sometimes also referred to as "snippets" or "passages".

Highlighting is extremely configurable, perhaps more than any other part of Solr. There are many parameters each for fragment sizing, formatting, ordering, backup/alternate behavior, and more options that are hard to categorize. Nonetheless, highlighting is very simple to use.

## Usage

### Common Highlighter Parameters

You only need to set the `hl` and often `hl.fl` parameters to get results. The following table documents these and some other supported parameters. Note that many highlighting parameters support per-field overrides, such as: `f.title_txt.hl.snippets`

#### `hl`

Use this parameter to enable or disable highlighting. The default is `false`. If you want to use highlighting, you must set this to `true`.

#### `hl.method`

The highlighting implementation to use. Acceptable values are: `unified`, `original`, `fastVector`. The default is `original`.

See the [Choosing a Highlighter](#) section below for more details on the differences between the available highlighters.

#### `hl.fl`

Specifies a list of fields to highlight, either comma- or space-delimited. These must be "stored". A wildcard of `*` (asterisk) can be used to match field globs, such as `text_*` or even `*` to highlight on all fields where highlighting is possible. When using `*`, consider adding `hl.requireFieldMatch=true`.

Note that the field(s) listed here ought to have compatible text-analysis (defined in the schema) with field(s) referenced in the query to be highlighted. It may be necessary to modify `hl.q` and `hl.qparser` and/or modify the text analysis. The following example uses the [local-params](#) syntax and [the edismax parser](#) to highlight fields in `hl.fl: &hl.fl=field1 field2&hl.q={!edismax qf=$hl.fl v=$q}&hl.qparser=lucene&hl.requireFieldMatch=true` (along with other applicable parameters, of course).

The default is the value of the `df` parameter which in turn has no default.

#### `hl.q`

A query to use for highlighting. This parameter allows you to highlight different terms or fields than those being used to search for documents. When setting this, you might also need to set `hl.qparser`.

The default is the value of the q parameter (already parsed).

#### h1.qparser

The [query parser](#) to use for the h1.q query. It only applies when h1.q is set.

The default is the value of the defType parameter which in turn defaults to lucene.

#### h1.requireFieldMatch

By default, false, all query terms will be highlighted for each field to be highlighted (h1.f1) no matter what fields the parsed query refer to. If set to true, only query terms aligning with the field being highlighted will in turn be highlighted.

If the query references fields different from the field being highlighted and they have different text analysis, the query may not highlight query terms it should have and vice versa. The analysis used is that of the field being highlighted (h1.f1), not the query fields.

#### h1.usePhraseHighlighter

If set to true, the default, Solr will highlight phrase queries (and other advanced position-sensitive queries) accurately – as phrases. If false, the parts of the phrase will be highlighted everywhere instead of only when it forms the given phrase.

#### h1.highlightMultiTerm

If set to true, the default, Solr will highlight wildcard queries (and other MultiTermQuery subclasses). If false, they won't be highlighted at all.

#### h1.snippets

Specifies maximum number of highlighted snippets to generate per field. It is possible for any number of snippets from zero to this value to be generated. The default is 1.

#### h1 fragsize

Specifies the approximate size, in characters, of fragments to consider for highlighting. The default is 100. Using 0 indicates that no fragmenting should be considered and the whole field value should be used.

#### h1.tag.pre

(h1.simple.pre for the Original Highlighter) Specifies the “tag” to use before a highlighted term. This can be any string, but is most often an HTML or XML tag.

The default is <em>.

#### h1.tag.post

(h1.simple.post for the Original Highlighter) Specifies the “tag” to use after a highlighted term. This can be any string, but is most often an HTML or XML tag.

The default is </em>.

#### h1.encoder

If blank, the default, then the stored text will be returned without any escaping/encoding performed by the highlighter. If set to html then special HTML/XML characters will be encoded (e.g., & becomes &amp;). The pre/post snippet characters are never encoded.

#### h1.maxAnalyzedChars

The character limit to look for highlights, after which no highlighting will be done. This is mostly only a

performance concern for an *analysis* based offset source since it's the slowest. See [Schema Options and Performance Considerations](#).

The default is 51200 characters.

There are more parameters supported as well depending on the highlighter (via `hl.method`) chosen.

## Highlighting in the Query Response

In the response to a query, Solr includes highlighting data in a section separate from the documents. It is up to a client to determine how to process this response and display the highlights to users.

Using the example documents included with Solr, we can see how this might work:

In response to a query such as:

```
http://localhost:8983/solr/gettingstarted/select?hl=on&q=apple&hl.fl=manu&fl=id,name,manu,cat
```

we get a response such as this (truncated slightly for space):

```
{
 "response": {
 "numFound": 1,
 "start": 0,
 "docs": [
 {
 "id": "MA147LL/A",
 "name": "Apple 60 GB iPod with Video Playback Black",
 "manu": "Apple Computer Inc.",
 "cat": [
 "electronics",
 "music"
]
 }
],
 "highlighting": {
 "MA147LL/A": {
 "manu": [
 "Apple Computer Inc."
]
 }
 }
}
```

Note the two sections `docs` and `highlighting`. The `docs` section contains the fields of the document requested with the `f1` parameter of the query (only `"id"`, `"name"`, `"manu"`, and `"cat"`).

The `highlighting` section includes the ID of each document, and the field that contains the highlighted portion. In this example, we used the `hl.fl` parameter to say we wanted query terms highlighted in the `"manu"` field. When there is a match to the query term in that field, it will be included for each document ID in the list.

## Choosing a Highlighter

Solr provides a `HighlightComponent` (a `SearchComponent`) and it's in the default list of components for search handlers. It offers a somewhat unified API over multiple actual highlighting implementations (or simply "highlighters") that do the business of highlighting.

There are many parameters supported by more than one highlighter, and sometimes the implementation details and semantics will be a bit different, so don't expect identical results when switching highlighters. You should use the `hl.method` parameter to choose a highlighter but it's also possible to explicitly configure an implementation by class name in `solrconfig.xml`.

There are four highlighters available that can be chosen at runtime with the `hl.method` parameter, in order of general recommendation:

### Unified Highlighter

(`hl.method=unified`)

The Unified Highlighter is the newest highlighter (as of Solr 6.4), which stands out as the most performant and accurate of the options. It can handle typical requirements and others possibly via plugins/extension. We recommend that you try this highlighter even though it isn't the default (yet).

The UH highlights a query very *accurately* and thus is true to what the underlying Lucene query actually matches. Other highlighters highlight terms more liberally (over-highlight). A strong benefit to this highlighter is that you can opt to configure Solr to put more information in the underlying index to speed up highlighting of large documents; multiple configurations are supported, even on a per-field basis. There is little or no such flexibility of offset sources for the other highlighters. More on this below.

There are some reasons not to choose this highlighter: The surround query parser doesn't yet work here — SOLR-12895. Passage scoring does not consider boosts in the query. Some people want more/better passage breaking flexibility.

### Original Highlighter

(`hl.method=original`, the default)

The Original Highlighter, sometimes called the "Standard Highlighter" or "Default Highlighter", is Lucene's original highlighter – a venerable option with a high degree of customization options. Its query accuracy is good enough for most needs, although it's not quite as good/perfect as the Unified Highlighter.

The Original Highlighter will normally analyze stored text on the fly in order to highlight. It will use full term vectors if available.

Where this highlighter falls short is performance; it's often twice as slow as the Unified Highlighter. And despite being the most customizable, it doesn't have a `BreakIterator` based fragmenter (all the others do), which could pose a challenge for some languages.

### FastVector Highlighter

(`hl.method=fastVector`)

The FastVector Highlighter *requires* full term vector options (`termVectors`, `termPositions`, and `termOffsets`) on the field, and is optimized with that in mind. It is nearly as configurable as the Original

Highlighter with some variability.

This highlighter notably supports multi-colored highlighting such that different query words can be denoted in the fragment with different marking, usually expressed as an HTML tag with a unique color.

This highlighter's query-representation is less advanced than the Original or Unified Highlighters: for example it will not work well with the surround parser, and there are multiple reported bugs pertaining to queries with stop-words.

Note that both the FastVector and Original Highlighters can be used in conjunction in a search request to highlight some fields with one and some the other. In contrast, the other highlighters can only be chosen exclusively.

The Unified Highlighter is exclusively configured via search parameters. In contrast, some settings for the Original and FastVector Highlighters are set in `solrconfig.xml`. There's a robust example of the latter in the "techproducts" configset.

In addition to further information below, more information can be found in the [Solr javadocs](#).

## Schema Options and Performance Considerations

Fundamental to the internals of highlighting are detecting the *offsets* of the individual words that match the query. Some of the highlighters can run the stored text through the analysis chain defined in the schema, some can look them up from *postings*, and some can look them up from *term vectors*. These choices have different trade-offs:

- **Analysis:** Supported by the Unified and Original Highlighters. If you don't go out of your way to configure the other options below, the highlighter will analyze the stored text on the fly (during highlighting) to calculate offsets.

The benefit of this approach is that your index won't grow larger with any extra data that isn't strictly necessary for highlighting.

The down side is that highlighting speed is roughly linear with the amount of text to process, with a large factor being the complexity of your analysis chain.

For "short" text, this is a good choice. Or maybe it's not short but you're prioritizing a smaller index and indexing speed over highlighting performance.

- **Postings:** Supported by the Unified Highlighter. Set `storeOffsetsWithPositions` to true. This adds a moderate amount of extra data to the index but it speeds up highlighting tremendously, especially compared to analysis with longer text fields.

However, wildcard queries will fall back to analysis unless "light" term vectors are added.

- **with Term Vectors (light):** Supported only by the Unified Highlighter. To enable this mode set `termVectors` to true but no other term vector related options on the field being highlighted.

This adds even more data to the index than just `storeOffsetsWithPositions` but not as much as enabling all the extra term vector options. Term Vectors are only accessed by the highlighter when a wildcard query is used and will prevent a fall back to analysis of the stored text.

This is definitely the fastest option for highlighting wildcard queries on large text fields.

- **Term Vectors (full)**: Supported by the Unified, FastVector, and Original Highlighters. Set termVectors, termPositions, and termOffsets to true, and potentially termPayloads for advanced use cases.

This adds substantial weight to the index – similar in size to the compressed stored text. If you are using the Unified Highlighter then this is not a recommended configuration since it's slower and heavier than postings with light term vectors. However, this could make sense if full term vectors are already needed for another use-case.

## The Unified Highlighter

The Unified Highlighter supports these following additional parameters to the ones listed earlier:

`hl.offsetSource`

By default, the Unified Highlighter will usually pick the right offset source (see above). However it may be ambiguous such as during a migration from one offset source to another that hasn't completed.

The offset source can be explicitly configured to one of: ANALYSIS, POSTINGS, POSTINGS\_WITH\_TERM\_VECTORS, or TERM\_VECTORS.

`hl.tag.ellipsis`

By default, each snippet is returned as a separate value (as is done with the other highlighters). Set this parameter to instead return one string with this text as the delimiter. *Note: this is likely to be removed in the future.*

`hl.defaultSummary`

If true, use the leading portion of the text as a snippet if a proper highlighted snippet can't otherwise be generated. The default is false.

`hl.score.k1`

Specifies BM25 term frequency normalization parameter 'k1'. For example, it can be set to 0 to rank passages solely based on the number of query terms that match. The default is 1.2.

`hl.score.b`

Specifies BM25 length normalization parameter 'b'. For example, it can be set to "0" to ignore the length of passages entirely when ranking. The default is 0.75.

`hl.score.pivot`

Specifies BM25 average passage length in characters. The default is 87.

`hl.bs.language`

Specifies the breakiterator language for dividing the document into passages.

`hl.bs.country`

Specifies the breakiterator country for dividing the document into passages.

`hl.bs.variant`

Specifies the breakiterator variant for dividing the document into passages.

`hl.bs.type`

Specifies the breakiterator type for dividing the document into passages. Can be SEPARATOR, SENTENCE, WORD\*, CHARACTER, LINE, or WHOLE. SEPARATOR is special value that splits text on a user-provided character in `hl.bs.separator`.

The default is SENTENCE.

#### `hl.bs.separator`

Indicates which character to break the text on. Use only if you have defined `hl.bs.type=SEPARATOR`.

This is useful when the text has already been manipulated in advance to have a special delineation character at desired highlight passage boundaries. This character will still appear in the text as the last character of a passage.

#### `hl.weightMatches`

Tells the UH to use Lucene's new "Weight Matches" API instead of doing SpanQuery conversion. This is the most accurate highlighting mode reflecting the query. Furthermore, phrases will be highlighted as a whole instead of word by word.

The default is true. However if either `hl.usePhraseHighlighter` or `hl.multiTermQuery` are set to false, then this setting is effectively false no matter what you set it to.

## The Original Highlighter

The Original Highlighter supports these following additional parameters to the ones listed earlier:

#### `hl.mergeContiguous`

Instructs Solr to collapse contiguous fragments into a single fragment. A value of true indicates contiguous fragments will be collapsed into single fragment. The default value, false, is also the backward-compatible setting.

#### `hl.maxMultiValuedToExamine`

Specifies the maximum number of entries in a multi-valued field to examine before stopping. This can potentially return zero results if the limit is reached before any matches are found.

If used with the `maxMultiValuedToMatch`, whichever limit is reached first will determine when to stop looking.

The default is `Integer.MAX_VALUE`.

#### `hl.maxMultiValuedToMatch`

Specifies the maximum number of matches in a multi-valued field that are found before stopping.

If `hl.maxMultiValuedToExamine` is also defined, whichever limit is reached first will determine when to stop looking.

The default is `Integer.MAX_VALUE`.

#### `hl.alternateField`

Specifies a field to be used as a backup default summary if Solr cannot generate a snippet (i.e., because no terms match).

#### hl.maxAlternateFieldLength

Specifies the maximum number of characters of the field to return. Any value less than or equal to 0 means the field's length is unlimited (the default behavior).

This parameter is only used in conjunction with the hl.alternateField parameter.

#### hl.highlightAlternate

If set to true, the default, and hl.alternateFieldName is active, Solr will show the entire alternate field, with highlighting of occurrences. If hl.maxAlternateFieldLength=N is used, Solr returns max N characters surrounding the best matching fragment.

If set to false, or if there is no match in the alternate field either, the alternate field will be shown without highlighting.

#### hl.formatter

Selects a formatter for the highlighted output. Currently the only legal value is simple, which surrounds a highlighted term with a customizable pre- and post-text snippet.

#### hl.simple.pre, hl.simple.post

Specifies the text that should appear before (hl.simple.pre) and after (hl.simple.post) a highlighted term, when using the simple formatter. The default is <em> and </em>.

#### hl.fragmenter

Specifies a text snippet generator for highlighted text. The standard (default) fragmenter is gap, which creates fixed-sized fragments with gaps for multi-valued fields.

Another option is regex, which tries to create fragments that resemble a specified regular expression.

#### hl.regex.slop

When using the regex fragmenter (hl.fragmenter=regex), this parameter defines the factor by which the fragmenter can stray from the ideal fragment size (given by hl fragsize) to accommodate a regular expression.

For instance, a slop of 0.2 with hl fragsize=100 should yield fragments between 80 and 120 characters in length. It is usually good to provide a slightly smaller hl fragsize value when using the regex fragmenter.

The default is 0.6.

#### hl.regex.pattern

Specifies the regular expression for fragmenting. This could be used to extract sentences.

#### hl.regex.maxAnalyzedChars

Instructs Solr to analyze only this many characters from a field when using the regex fragmenter (after which, the fragmenter produces fixed-sized fragments). The default is 10000.

Note, applying a complicated regex to a huge field is computationally expensive.

#### hl.preserveMulti

If true, multi-valued fields will return all values in the order they were saved in the index. If false, the default, only values that match the highlight request will be returned.

## h1.payloads

When h1.usePhraseHighlighter is true and the indexed field has payloads but not term vectors (generally quite rare), the index's payloads will be read into the highlighter's memory index along with the postings.

If this may happen and you know you don't need them for highlighting (i.e., your queries don't filter by payload) then you can save a little memory by setting this to false.

The Original Highlighter has a plugin architecture that enables new functionality to be registered in solrconfig.xml. The "techproducts" configset shows most of these settings explicitly. You can use it as a guide to provide your own components to include a SolrFormatter, SolrEncoder, and SolrFragmenter.

## The FastVector Highlighter

The FastVector Highlighter (F VH) can be used in conjunction with the Original Highlighter if not all fields should be highlighted with the F VH. In such a mode, set h1.method=original and f.yourTermVecField.h1.method=fastVector for all fields that should use the F VH. One annoyance to keep in mind is that the Original Highlighter uses h1.simple.pre whereas the F VH (and other highlighters) use h1.tag.pre.

In addition to the initial listed parameters, the following parameters documented for the Original Highlighter above are also supported by the F VH:

- h1.alternateField
- h1.maxAlternateFieldLength
- h1.highlightAlternate

And here are additional parameters supported by the F VH:

### h1.fragListBuilder

The snippet fragmenting algorithm. The weighted fragListBuilder uses IDF-weights to order fragments. This fragListBuilder is the default.

Other options are single, which returns the entire field contents as one snippet, or simple. You can select a fragListBuilder with this parameter, or modify an existing implementation in solrconfig.xml to be the default by adding "default=true".

### h1.fragmentsBuilder

The fragments builder is responsible for formatting the fragments, which uses <em> and </em> markup by default (if h1.tag.pre and h1.tag.post are not defined).

Another pre-configured choice is colored, which is an example of how to use the fragments builder to insert HTML into the snippets for colored highlights if you choose. You can also implement your own if you'd like. You can select a fragments builder with this parameter, or modify an existing implementation in solrconfig.xml to be the default by adding "default=true".

### h1.boundaryScanner

See [Using Boundary Scanners with the FastVector Highlighter](#) below.

### h1.bs.\*

See [Using Boundary Scanners with the FastVector Highlighter](#) below.

## h1.phraseLimit

The maximum number of phrases to analyze when searching for the highest-scoring phrase. The default is 5000.

## h1.multiValuedSeparatorChar

Text to use to separate one value from the next for a multi-valued field. The default is " " (a space).

# Using Boundary Scanners with the FastVector Highlighter

The FastVector Highlighter will occasionally truncate highlighted words. To prevent this, implement a boundary scanner in `solrconfig.xml`, then use the `hl.boundaryScanner` parameter to specify the boundary scanner for highlighting.

Solr supports two boundary scanners: `breakIterator` and `simple`.

## The `breakIterator` Boundary Scanner

The `breakIterator` boundary scanner offers excellent performance right out of the box by taking locale and boundary type into account. In most cases you will want to use the `breakIterator` boundary scanner. To implement the `breakIterator` boundary scanner, add this code to the highlighting section of your `solrconfig.xml` file, adjusting the type, language, and country values as appropriate to your application:

```
<boundaryScanner name="breakIterator" class="solr.highlight.BreakIteratorBoundaryScanner">
 <lst name="defaults">
 <str name="hl.bs.type">WORD</str>
 <str name="hl.bs.language">en</str>
 <str name="hl.bs.country">US</str>
 </lst>
</boundaryScanner>
```

Possible values for the `hl.bs.type` parameter are WORD, LINE, SENTENCE, and CHARACTER.

## The `simple` Boundary Scanner

The `simple` boundary scanner scans term boundaries for a specified maximum character value (`hl.bs.maxScan`) and for common delimiters such as punctuation marks (`hl.bs.chars`). To implement the `simple` boundary scanner, add this code to the highlighting section of your `solrconfig.xml` file, adjusting the values as appropriate to your application:

```
<boundaryScanner name="simple" class="solr.highlight.SimpleBoundaryScanner" default="true">
 <lst name="defaults">
 <str name="hl.bs.maxScan">10</str>
 <str name="hl.bs.chars">.,!?\t\n</str>
 </lst>
</boundaryScanner>
```

# Spell Checking

The SpellCheck component is designed to provide inline query suggestions based on other, similar, terms.

The basis for these suggestions can be terms in a field in Solr, externally created text files, or fields in other Lucene indexes.

## Configuring the SpellCheckComponent

### Define Spell Check in solrconfig.xml

The first step is to specify the source of terms in `solrconfig.xml`. There are three approaches to spell checking in Solr, discussed below.

#### IndexBasedSpellChecker

The `IndexBasedSpellChecker` uses a Solr index as the basis for a parallel index used for spell checking. It requires defining a field as the basis for the index terms; a common practice is to copy terms from some fields (such as title, body, etc.) to another field created for spell checking. Here is a simple example of configuring `solrconfig.xml` with the `IndexBasedSpellChecker`:

```
<searchComponent name="spellcheck" class="solr.SpellCheckComponent">
 <lst name="spellchecker">
 <str name="classname">solr.IndexBasedSpellChecker</str>
 <str name="spellcheckIndexDir">./spellchecker</str>
 <str name="field">content</str>
 <str name="buildOnCommit">true</str>
 <!-- optional elements with defaults
 <str name="distanceMeasure">org.apache.lucene.search.spell.LevenshteinDistance</str>
 <str name="accuracy">0.5</str>
 -->
 </lst>
</searchComponent>
```

The first element defines the `searchComponent` to use the `solr.SpellCheckComponent`. The `classname` is the specific implementation of the `SpellCheckComponent`, in this case `solr.IndexBasedSpellChecker`. Defining the `classname` is optional; if not defined, it will default to `IndexBasedSpellChecker`.

The `spellcheckIndexDir` defines the location of the directory that holds the spellcheck index, while the `field` defines the source field (defined in the Schema) for spell check terms. When choosing a field for the spellcheck index, it's best to avoid a heavily processed field to get more accurate results. If the field has many word variations from processing synonyms and/or stemming, the dictionary will be created with those variations in addition to more valid spelling data.

Finally, `buildOnCommit` defines whether to build the spell check index at every commit (that is, every time new documents are added to the index). It is optional, and can be omitted if you would rather set it to false.

## DirectSolrSpellChecker

The DirectSolrSpellChecker uses terms from the Solr index without building a parallel index like the IndexBasedSpellChecker. This spell checker has the benefit of not having to be built regularly, meaning that the terms are always up-to-date with terms in the index. Here is how this might be configured in solrconfig.xml

```
<searchComponent name="spellcheck" class="solr.SpellCheckComponent">
 <lst name="spellchecker">
 <str name="name">default</str>
 <str name="field">name</str>
 <str name="classname">solr.DirectSolrSpellChecker</str>
 <str name="distanceMeasure">internal</str>
 <float name="accuracy">0.5</float>
 <int name="maxEdits">2</int>
 <int name="minPrefix">1</int>
 <int name="maxInspections">5</int>
 <int name="minQueryLength">4</int>
 <float name="maxQueryFrequency">0.01</float>
 <float name="thresholdTokenFrequency">.01</float>
 </lst>
</searchComponent>
```

When choosing a field to query for this spell checker, you want one which has relatively little analysis performed on it (particularly analysis such as stemming). Note that you need to specify a field to use for the suggestions, so like the IndexBasedSpellChecker, you may want to copy data from fields like title, body, etc., to a field dedicated to providing spelling suggestions.

Many of the parameters relate to how this spell checker should query the index for term suggestions. The distanceMeasure defines the metric to use during the spell check query. The value "internal" uses the default Levenshtein metric, which is the same metric used with the other spell checker implementations.

Because this spell checker is querying the main index, you may want to limit how often it queries the index to be sure to avoid any performance conflicts with user queries. The accuracy setting defines the threshold for a valid suggestion, while maxEdits defines the number of changes to the term to allow. Since most spelling mistakes are only 1 letter off, setting this to 1 will reduce the number of possible suggestions (the default, however, is 2); the value can only be 1 or 2. minPrefix defines the minimum number of characters the terms should share. Setting this to 1 means that the spelling suggestions will all start with the same letter, for example.

The maxInspections parameter defines the maximum number of possible matches to review before returning results; the default is 5. minQueryLength defines how many characters must be in the query before suggestions are provided; the default is 4.

At first, spellchecker analyses incoming query words by looking up them in the index. Only query words, which are absent in index or too rare ones (below maxQueryFrequency) are considered as misspelled and used for finding suggestions. Words which are frequent than maxQueryFrequency bypass spellchecker unchanged. After suggestions for every misspelled word are found they are filtered for enough frequency with thresholdTokenFrequency as boundary value. These parameters (maxQueryFrequency and thresholdTokenFrequency) can be a percentage (such as .01, or 1%) or an absolute value (such as 4).

## FileBasedSpellChecker

The FileBasedSpellChecker uses an external file as a spelling dictionary. This can be useful if using Solr as a spelling server, or if spelling suggestions don't need to be based on actual terms in the index. In solrconfig.xml, you would define the searchComponent as so:

```
<searchComponent name="spellcheck" class="solr.SpellCheckComponent">
 <lst name="spellchecker">
 <str name="classname">solr.FileBasedSpellChecker</str>
 <str name="name">file</str>
 <str name="sourceLocation">spellings.txt</str>
 <str name="characterEncoding">UTF-8</str>
 <str name="spellcheckIndexDir">./spellcheckerFile</str>
 <!-- optional elements with defaults
 <str name="distanceMeasure">org.apache.lucene.search.spell.LevenshteinDistance</str>
 <str name="accuracy">0.5</str>
 -->
 </lst>
</searchComponent>
```

The differences here are the use of the sourceLocation to define the location of the file of terms and the use of characterEncoding to define the encoding of the terms file.



In the previous example, *name* is used to name this specific definition of the spellchecker. Multiple definitions can co-exist in a single solrconfig.xml, and the *name* helps to differentiate them. If only defining one spellchecker, no name is required.

## WordBreakSolrSpellChecker

WordBreakSolrSpellChecker offers suggestions by combining adjacent query terms and/or breaking terms into multiple words. It is a SpellCheckComponent enhancement, leveraging Lucene's WordBreakSpellChecker. It can detect spelling errors resulting from misplaced whitespace without the use of shingle-based dictionaries and provides collation support for word-break errors, including cases where the user has a mix of single-word spelling errors and word-break errors in the same query. It also provides shard support.

Here is how it might be configured in solrconfig.xml:

```
<searchComponent name="spellcheck" class="solr.SpellCheckComponent">
 <lst name="spellchecker">
 <str name="name">wordbreak</str>
 <str name="classname">solr.WordBreakSolrSpellChecker</str>
 <str name="field">lowerfilt</str>
 <str name="combineWords">true</str>
 <str name="breakWords">true</str>
 <int name="maxChanges">10</int>
 </lst>
</searchComponent>
```

Some of the parameters will be familiar from the discussion of the other spell checkers, such as name, classname, and field. New for this spell checker is combineWords, which defines whether words should be combined in a dictionary search (default is true); breakWords, which defines if words should be broken during a dictionary search (default is true); and maxChanges, an integer which defines how many times the spell checker should check collation possibilities against the index (default is 10).

The spellchecker can be configured with a traditional checker (i.e., DirectSolrSpellChecker). The results are combined and collations can contain a mix of corrections from both spellcheckers.

## Add It to a Request Handler

Queries will be sent to a [RequestHandler](#). If every request should generate a suggestion, then you would add the following to the requestHandler that you are using:

```
<str name="spellcheck">true</str>
```

One of the possible parameters is the spellcheck.dictionary to use, and multiples can be defined. With multiple dictionaries, all specified dictionaries are consulted and results are interleaved. Collations are created with combinations from the different spellcheckers, with care taken that multiple overlapping corrections do not occur in the same collation.

Here is an example with multiple dictionaries:

```
<requestHandler name="spellCheckWithWordbreak" class="org.apache.solr.handler.component.SearchHandler">
 <lst name="defaults">
 <str name="spellcheck.dictionary">default</str>
 <str name="spellcheck.dictionary">wordbreak</str>
 <str name="spellcheck.count">20</str>
 </lst>
 <arr name="last-components">
 <str>spellcheck</str>
 </arr>
</requestHandler>
```

## Spell Check Parameters

The SpellCheck component accepts the parameters described below.

### spellcheck

This parameter turns on SpellCheck suggestions for the request. If true, then spelling suggestions will be generated. This is required if spell checking is desired.

### spellcheck.q or q

This parameter specifies the query to spellcheck.

If spellcheck.q is defined, then it is used; otherwise the original input query is used. The spellcheck.q parameter is intended to be the original query, minus any extra markup like field names, boosts, and so

on. If the `q` parameter is specified, then the `SpellingQueryConverter` class is used to parse it into tokens; otherwise the `WhitespaceTokenizer` is used.

The choice of which one to use is up to the application. Essentially, if you have a spelling "ready" version in your application, then it is probably better to use `spellcheck.q`. Otherwise, if you just want Solr to do the job, use the `q` parameter.



The `SpellingQueryConverter` class does not deal properly with non-ASCII characters. In this case, you have either to use `spellcheck.q`, or implement your own `QueryConverter`.

#### `spellcheck.build`

If set to `true`, this parameter creates the dictionary to be used for spell-checking. In a typical search application, you will need to build the dictionary before using the spell check. However, it's not always necessary to build a dictionary first. For example, you can configure the spellchecker to use a dictionary that already exists.

The dictionary will take some time to build, so this parameter should not be sent with every request.

#### `spellcheck.reload`

If set to `true`, this parameter reloads the spellchecker. The results depend on the implementation of `SolrSpellChecker.reload()`. In a typical implementation, reloading the spellchecker means reloading the dictionary.

#### `spellcheck.count`

This parameter specifies the maximum number of suggestions that the spellchecker should return for a term. If this parameter isn't set, the value defaults to 1. If the parameter is set but not assigned a number, the value defaults to 5. If the parameter is set to a positive integer, that number becomes the maximum number of suggestions returned by the spellchecker.

#### `spellcheck.queryAnalyzerFieldType`

A field type from Solr's schema. The analyzer configured for the provided field type is used by the `QueryConverter` to tokenize the value for "`q`" parameter. The field type specified by this parameter should do minimal transformations. It's usually a best practice to avoid types that aggressively stem or NGram, for instance, since those types of analysis can throw off spell checking.

#### `spellcheck.onlyMorePopular`

If true, Solr will return suggestions that result in more hits for the query than the existing query. Note that this will return more popular suggestions even when the given query term is present in the index and considered "correct".

#### `spellcheck.maxResultsForSuggest`

If, for example, this is set to 5 and the user's query returns 5 or fewer results, the spellchecker will report "`correctlySpelled=false`" and also offer suggestions (and collations if requested). Setting this greater than zero is useful for creating "did-you-mean?" suggestions for queries that return a low number of hits.

#### `spellcheck.alternativeTermCount`

Defines the number of suggestions to return for each query term existing in the index and/or dictionary. Presumably, users will want fewer suggestions for words with `docFrequency>0`. Also, setting this value enables context-sensitive spell suggestions.

#### spellcheck.extendedResults

If true, this parameter causes to Solr to return additional information about spellcheck results, such as the frequency of each original term in the index (origFreq) as well as the frequency of each suggestion in the index (frequency). Note that this result format differs from the non-extended one as the returned suggestion for a word is actually an array of lists, where each list holds the suggested term and its frequency.

#### spellcheck.collate

If true, this parameter directs Solr to take the best suggestion for each token (if one exists) and construct a new query from the suggestions.

For example, if the input query was "jawa class lording" and the best suggestion for "jawa" was "java" and "lording" was "loading", then the resulting collation would be "java class loading".

The spellcheck.collate parameter only returns collations that are guaranteed to result in hits if re-queried, even when applying original fq parameters. This is especially helpful when there is more than one correction per query.



This only returns a query to be used. It does not actually run the suggested query.

#### spellcheck.maxCollations

The maximum number of collations to return. The default is 1. This parameter is ignored if spellcheck.collate is false.

#### spellcheck.maxCollationTries

This parameter specifies the number of collation possibilities for Solr to try before giving up. Lower values ensure better performance. Higher values may be necessary to find a collation that can return results. The default value is 0, which is equivalent to not checking collations. This parameter is ignored if spellcheck.collate is false.

#### spellcheck.maxCollationEvaluations

This parameter specifies the maximum number of word correction combinations to rank and evaluate prior to deciding which collation candidates to test against the index. This is a performance safety-net in case a user enters a query with many misspelled words. The default is 10000 combinations, which should work well in most situations.

#### spellcheck.collateExtendedResults

If true, this parameter returns an expanded response format detailing the collations Solr found. The default value is false and this is ignored if spellcheck.collate is false.

#### spellcheck.collateMaxCollectDocs

This parameter specifies the maximum number of documents that should be collected when testing potential collations against the index. A value of 0 indicates that all documents should be collected, resulting in exact hit-counts. Otherwise an estimation is provided as a performance optimization in cases where exact hit-counts are unnecessary – the higher the value specified, the more precise the estimation.

The default value for this parameter is 0, but when spellcheck.collateExtendedResults is false, the optimization is always used as if 1 had been specified.

#### spellcheck.collateParam.\* **Prefix**

This parameter prefix can be used to specify any additional parameters that you wish to the Spellchecker to use when internally validating collation queries. For example, even if your regular search results allow for loose matching of one or more query terms via parameters like `q.op=OR` and `mm=20%` you can specify override parameters such as `spellcheck.collateParam.q.op=AND&spellcheck.collateParam.mm=100%` to require that only collations consisting of words that are all found in at least one document may be returned.

#### spellcheck.dictionary

This parameter causes Solr to use the dictionary named in the parameter's argument. The default setting is `default`. This parameter can be used to invoke a specific spellchecker on a per request basis.

#### spellcheck.accuracy

Specifies an accuracy value to be used by the spell checking implementation to decide whether a result is worthwhile or not. The value is a float between 0 and 1. Defaults to `Float.MIN_VALUE`.

#### spellcheck.<DICT\_NAME>.key

Specifies a key/value pair for the implementation handling a given dictionary. The value that is passed through is just `key=value` (`spellcheck.<DICT_NAME>.` is stripped off).

For example, given a dictionary called `foo`, `spellcheck.foo.myKey=myValue` would result in `myKey=myValue` being passed through to the implementation handling the dictionary `foo`.

## Spell Check Example

Using Solr's `bin/solr -e techproducts` example, this query shows the results of a simple request that defines a query using the `spellcheck.q` parameter, and forces the collations to require all input terms must match:

```
http://localhost:8983/solr/techproducts/spell?df=text&spellcheck.q=delll+ultra+sharp&spellcheck=true&spellcheck.collateParam.q.op=AND&wt=xml
```

Results:

```
<lst name="spellcheck">
 <lst name="suggestions">
 <lst name="delll">
 <int name="numFound">1</int>
 <int name="startOffset">0</int>
 <int name="endOffset">5</int>
 <int name="origFreq">0</int>
 <arr name="suggestion">
 <lst>
 <str name="word">dell</str>
 <int name="freq">1</int>
 </lst>
 </arr>
 </lst>
 <lst name="ultra sharp">
 <int name="numFound">1</int>
 <int name="startOffset">6</int>
 <int name="endOffset">17</int>
 <int name="origFreq">0</int>
 <arr name="suggestion">
 <lst>
 <str name="word">ultrasharp</str>
 <int name="freq">1</int>
 </lst>
 </arr>
 </lst>
 </lst>
 <bool name="correctlySpelled">false</bool>
 <lst name="collations">
 <lst name="collation">
 <str name="collationQuery">dell ultrasharp</str>
 <int name="hits">1</int>
 <lst name="misspellingsAndCorrections">
 <str name="delll">dell</str>
 <str name="ultra sharp">ultrasharp</str>
 </lst>
 </lst>
 </lst>
</lst>
```

## Distributed SpellCheck

The SpellCheckComponent also supports spellchecking on distributed indexes. If you are using the SpellCheckComponent on a request handler other than "/select", you must provide the following two parameters:

shards

Specifies the shards in your distributed indexing configuration. For more information about distributed indexing, see [Distributed Search with Index Sharding](#)

shards.qt

Specifies the request handler Solr uses for requests to shards. This parameter is not required for the /select request handler.

For example:

```
http://localhost:8983/solr/techproducts/spell?spellcheck=true&spellcheck.build=true&spellcheck.q=toyata&shards.qt=/spell&shards=solr-shard1:8983/solr/techproducts,solr-shard2:8983/solr/techproducts
```

In case of a distributed request to the SpellCheckComponent, the shards are requested for at least five suggestions even if the spellcheck.count parameter value is less than five. Once the suggestions are collected, they are ranked by the configured distance measure (Levenshtein Distance by default) and then by aggregate frequency.

# Query Re-Ranking

Query Re-Ranking allows you to run a simple query (A) for matching documents and then re-rank the top N documents using the scores from a more complex query (B).

Since the more costly ranking from query B is only applied to the top *N* documents, it will have less impact on performance than just using the complex query B by itself. The trade off is that documents which score very low using the simple query A may not be considered during the re-ranking phase, even if they would score very highly using query B.

## Specifying a Ranking Query

A Ranking query can be specified using the `rq` request parameter. The `rq` parameter must specify a query string that when parsed, produces a [RankQuery](#).

Three rank queries are currently included in the Solr distribution. You can also configure a custom [QParserPlugin](#) you have written, but most users can just use a parser provided with Solr.

Parser	QParserPlugin class
rerank	<a href="#">ReRankQParserPlugin</a>
xport	<a href="#">ExportQParserPlugin</a>
ltr	<a href="#">LTRQParserPlugin</a>

## ReRank Query Parser

The rerank parser wraps a query specified by an local parameter, along with additional parameters indicating how many documents should be re-ranked, and how the final scores should be computed:

### reRankQuery

The query string for your complex ranking query - in most cases a [variable](#) will be used to refer to another request parameter. This parameter is required.

### reRankDocs

The number of top N documents from the original query that should be re-ranked. This number will be treated as a minimum, and may be increased internally automatically in order to rank enough documents to satisfy the query (i.e., `start+rows`). The default is 200.

### reRankWeight

A multiplicative factor that will be applied to the score from the `reRankQuery` for each of the top matching documents, before that score is added to the original score. The default is 2.0.

In the example below, the top 1000 documents matching the query "greetings" will be re-ranked using the query "(hi hello hey hiya)". The resulting scores for each of those 1000 documents will be 3 times their score from the "(hi hello hey hiya)", plus the score from the original "greetings" query:

```
q=greetings&rq={!rerank reRankQuery=$rq reRankDocs=1000 reRankWeight=3}&rqq=(hi+hello+hey+hiya)
```

If a document matches the original query, but does not match the re-ranking query, the document's original score will remain.

## LTR Query Parser

The ltr stands for Learning To Rank, please see [Learning To Rank](#) for more detailed information.

# Combining Ranking Queries with Other Solr Features

The rq parameter and the re-ranking feature in general works well with other Solr features. For example, it can be used in conjunction with the [collapse parser](#) to re-rank the group heads after they've been collapsed. It also preserves the order of documents elevated by the [elevation component](#). And it even has its own custom explain so you can see how the re-ranking scores were derived when looking at [debug information](#).

## Learning To Rank

With the **Learning To Rank** (or **LTR** for short) contrib module you can configure and run machine learned ranking models in Solr.

The module also supports feature extraction inside Solr. The only thing you need to do outside Solr is train your own ranking model.

## Learning to Rank Concepts

### Re-Ranking

Re-Ranking allows you to run a simple query for matching documents and then re-rank the top N documents using the scores from a different, more complex query. This page describes the use of **LTR** complex queries, information on other rank queries included in the Solr distribution can be found on the [Query Re-Ranking](#) page.

### Learning To Rank Models

In information retrieval systems, [Learning to Rank](#) is used to re-rank the top N retrieved documents using trained machine learning models. The hope is that such sophisticated models can make more nuanced ranking decisions than standard ranking functions like [TF-IDF](#) or [BM25](#).

### Ranking Model

A ranking model computes the scores used to rerank documents. Irrespective of any particular algorithm or implementation, a ranking model's computation can use three types of inputs:

- parameters that represent the scoring algorithm
- features that represent the document being scored
- features that represent the query for which the document is being scored

### Feature

A feature is a value, a number, that represents some quantity or quality of the document being scored or of the query for which documents are being scored. For example documents often have a 'recency' quality and

'number of past purchases' might be a quantity that is passed to Solr as part of the search query.

## Normalizer

Some ranking models expect features on a particular scale. A normalizer can be used to translate arbitrary feature values into normalized values e.g., on a 0..1 or 0..100 scale.

## Training Models

### Feature Engineering

The LTR contrib module includes several feature classes as well as support for custom features. Each feature class's javadocs contain an example to illustrate use of that class. The process of [feature engineering](#) itself is then entirely up to your domain expertise and creativity.

Feature	Class	Example parameters	External Feature Information
field length	<a href="#">FieldLengthFeature</a>	{"field": "title"}	not (yet) supported
field value	<a href="#">FieldValueFeature</a>	{"field": "hits"}	not (yet) supported
original score	<a href="#">OriginalScoreFeature</a>	{}	not applicable
solr query	<a href="#">SolrFeature</a>	{"q": "{!func} recip(ms(NOW, last_modified), 3.16e-11, 1, 1)"}	supported
solr filter query	<a href="#">SolrFeature</a>	{"fq": ["{!terms f=category}book"]}	supported
solr query + filter query	<a href="#">SolrFeature</a>	{"q": "{!func} recip(ms(NOW, last_modified), 3.16e-11, 1, 1)", "fq": ["{!terms f=category}book"]}	supported
value	<a href="#">ValueFeature</a>	{"value": "\${userFromMobile}", "required": true}	supported
(custom)	(custom class extending <a href="#">Feature</a> )		

Normalizer	Class	Example parameters
Identity	<a href="#">IdentityNormalizer</a>	{}
MinMax	<a href="#">MinMaxNormalizer</a>	{"min": "0", "max": "50" }
Standard	<a href="#">StandardNormalizer</a>	{"avg": "42", "std": "6"}
(custom)	(custom class extending <a href="#">Normalizer</a> )	

## Feature Extraction

The ltr contrib module includes a [features] transformer to support the calculation and return of feature values for [feature extraction](#) purposes including and especially when you do not yet have an actual reranking model.

## Feature Selection and Model Training

Feature selection and model training take place offline and outside Solr. The ltr contrib module supports two generalized forms of models as well as custom models. Each model class's javadocs contain an example to illustrate configuration of that class. In the form of JSON files your trained model or models (e.g., different models for different customer geographies) can then be directly uploaded into Solr using provided REST APIs.

General form	Class	Specific examples
Linear	<a href="#">LinearModel</a>	RankSVM, Pranking
Multiple Additive Trees	<a href="#">MultipleAdditiveTreesModel</a>	LambdaMART, Gradient Boosted Regression Trees (GBRT)
Neural Network	<a href="#">NeuralNetworkModel</a>	RankNet
(wrapper)	<a href="#">DefaultWrapperModel</a>	(not applicable)
(custom)	(custom class extending <a href="#">AdapterModel</a> )	(not applicable)
(custom)	(custom class extending <a href="#">LTRScoringModel</a> )	(not applicable)

## Quick Start with LTR

The "techproducts" example included with Solr is pre-configured with the plugins required for learning-to-rank, but they are disabled by default.

To enable the plugins, please specify the `solr.ltr.enabled` JVM System Property when running the example:

```
bin/solr start -e techproducts -Dsolr.ltr.enabled=true
```

## Uploading Features

To upload features in a /path/myFeatures.json file, please run:

```
curl -XPUT 'http://localhost:8983/solr/techproducts/schema/feature-store' --data-binary
"@/path/myFeatures.json" -H 'Content-type:application/json'
```

To view the features you just uploaded please open the following URL in a browser:

```
http://localhost:8983/solr/techproducts/schema/feature-store/_DEFAULT_
```

Example: /path/myFeatures.json

```
[
 {
 "name" : "documentRecency",
 "class" : "org.apache.solr.ltr.feature.SolrFeature",
 "params" : {
 "q" : "{!func}recip(ms(NOW,last_modified), 3.16e-11, 1, 1)"
 }
 },
 {
 "name" : "isBook",
 "class" : "org.apache.solr.ltr.feature.SolrFeature",
 "params" : {
 "fq": ["{!terms f=cat}book"]
 }
 },
 {
 "name" : "originalScore",
 "class" : "org.apache.solr.ltr.feature.OriginalScoreFeature",
 "params" : {}
 }
]
```

## Extracting Features

To extract features as part of a query, add [features] to the fl parameter, for example:

```
http://localhost:8983/solr/techproducts/query?q=test&fl=id,score,[features]
```

The output XML will include feature values as a comma-separated list, resembling the output shown here:

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 0,
 "params": {
 "q": "test",
 "f1": "id, score, [features]"}
 },
 "response": {
 "numFound": 2,
 "start": 0,
 "maxScore": 1.959392,
 "docs": [
 {
 "id": "GB18030TEST",
 "score": 1.959392,
 "[features]": "documentRecency=0.020893794, isBook=0.0, originalScore=1.959392"
 },
 {
 "id": "UTF8TEST",
 "score": 1.5513437,
 "[features]": "documentRecency=0.020893794, isBook=0.0, originalScore=1.5513437"
 }
]
 }
}
```

## Uploading a Model

To upload the model in a /path/myModel.json file, please run:

```
curl -XPUT 'http://localhost:8983/solr/techproducts/schema/model-store' --data-binary
"@/path/myModel.json" -H 'Content-type:application/json'
```

To view the model you just uploaded please open the following URL in a browser:

```
http://localhost:8983/solr/techproducts/schema/model-store
```

*Example: /path/myModel.json*

```
{
 "class" : "org.apache.solr.ltr.model.LinearModel",
 "name" : "myModel",
 "features" : [
 { "name" : "documentRecency" },
 { "name" : "isBook" },
 { "name" : "originalScore" }
],
 "params" : {
 "weights" : {
 "documentRecency" : 1.0,
 "isBook" : 0.1,
 "originalScore" : 0.5
 }
 }
}
```

## Running a Rerank Query

To rerank the results of a query, add the `rq` parameter to your search, for example:

```
http://localhost:8983/solr/techproducts/query?q=test&rq={!ltr model=myModel
reRankDocs=100}&fl=id,score
```

The addition of the `rq` parameter will not change the output XML of the search.

To obtain the feature values computed during reranking, add `[features]` to the `f1` parameter, for example:

```
http://localhost:8983/solr/techproducts/query?q=test&rq={!ltr model=myModel
reRankDocs=100}&f1=id,score,[features]
```

The output XML will include feature values as a comma-separated list, resembling the output shown here:

```
{
 "responseHeader":{
 "status":0,
 "QTime":0,
 "params":{
 "q":"test",
 "f1":"id,score,[features]",
 "rq":"={!ltr model=myModel reRankDocs=100}"}},
 "response":{ "numFound":2, "start":0, "maxScore":1.0005897, "docs": [
 {
 "id": "GB18030TEST",
 "score": 1.0005897,
 "[features]": "documentRecency=0.020893792,isBook=0.0,originalScore=1.959392"},
 {
 "id": "UTF8TEST",
 "score": 0.79656565,
 "[features]": "documentRecency=0.020893792,isBook=0.0,originalScore=1.5513437"}]
]}
}
```

## External Feature Information

The `ValueFeature` and `SolrFeature` classes support the use of external feature information, `efi` for short.

### Uploading Features

To upload features in a `/path/myEfiFeatures.json` file, please run:

```
curl -XPUT 'http://localhost:8983/solr/techproducts/schema/feature-store' --data-binary
"@/path/myEfiFeatures.json" -H 'Content-type:application/json'
```

To view the features you just uploaded please open the following URL in a browser:

```
http://localhost:8983/solr/techproducts/schema/feature-store/myEfiFeatureStore
```

Example: /path/myEfiFeatures.json

```
[
 {
 "store" : "myEfiFeatureStore",
 "name" : "isPreferredManufacturer",
 "class" : "org.apache.solr.ltr.feature.SolrFeature",
 "params" : { "fq" : ["{!field f=manu}${preferredManufacturer}"] }
 },
 {
 "store" : "myEfiFeatureStore",
 "name" : "userAnswerValue",
 "class" : "org.apache.solr.ltr.feature.ValueFeature",
 "params" : { "value" : "${answer:42}" }
 },
 {
 "store" : "myEfiFeatureStore",
 "name" : "userFromMobileValue",
 "class" : "org.apache.solr.ltr.feature.ValueFeature",
 "params" : { "value" : "${fromMobile}" , "required" : true }
 },
 {
 "store" : "myEfiFeatureStore",
 "name" : "userTextCat",
 "class" : "org.apache.solr.ltr.feature.SolrFeature",
 "params" : { "q" : "{!field f=cat}${text}" }
 }
]
```

As an aside, you may have noticed that the myEfiFeatures.json example uses "store": "myEfiFeatureStore" attributes: read more about feature store in the [LTR Lifecycle](#) section of this page.

### Extracting Features

To extract myEfiFeatureStore features as part of a query, add efi.\* parameters to the [features] part of the fl parameter, for example:

```
http://localhost:8983/solr/techproducts/query?q=test&fl=id,cat,manu,score,[features
store=myEfiFeatureStore efi.text=test efi.preferredManufacturer=Apache efi.fromMobile=1]
```

```
http://localhost:8983/solr/techproducts/query?q=test&fl=id,cat,manu,score,[features
store=myEfiFeatureStore efi.text=test efi.preferredManufacturer=Apache efi.fromMobile=0
efi.answer=13]
```

## Uploading a Model

To upload the model in a /path/myEfiModel.json file, please run:

```
curl -XPUT 'http://localhost:8983/solr/techproducts/schema/model-store' --data-binary
"@/path/myEfiModel.json" -H 'Content-type:application/json'
```

To view the model you just uploaded please open the following URL in a browser:

```
http://localhost:8983/solr/techproducts/schema/model-store
```

*Example: /path/myEfiModel.json*

```
{
 "store" : "myEfiFeatureStore",
 "name" : "myEfiModel",
 "class" : "org.apache.solr.ltr.model.LinearModel",
 "features" : [
 { "name" : "isPreferredManufacturer" },
 { "name" : "userAnswerValue" },
 { "name" : "userFromMobileValue" },
 { "name" : "userTextCat" }
,
 "params" : {
 "weights" : {
 "isPreferredManufacturer" : 0.2,
 "userAnswerValue" : 1.0,
 "userFromMobileValue" : 1.0,
 "userTextCat" : 0.1
 }
 }
}
```

## Running a Rerank Query

To obtain the feature values computed during reranking, add [features] to the fl parameter and efi.\* parameters to the rq parameter, for example:

```
http://localhost:8983/solr/techproducts/query?q=test&rq={!ltr model=myEfiModel efi.text=test
efi.preferredManufacturer=Apache efi.fromMobile=1}&fl=id,cat,manu,score,[features]
```

```
http://localhost:8983/solr/techproducts/query?q=test&rq={!ltr model=myEfiModel efi.text=test
efi.preferredManufacturer=Apache efi.fromMobile=0 efi.answer=13}&fl=id,cat,manu,score,[features]
```

Notice the absence of efi.\* parameters in the [features] part of the fl parameter.

## Extracting Features While Reranking

To extract features for myEfiFeatureStore features while still reranking with myModel:

```
http://localhost:8983/solr/techproducts/query?q=test&rq={!ltr
model=myModel}&fl=id,cat,manu,score,[features store=myEfiFeatureStore efi.text=test
efi.preferredManufacturer=Apache efi.fromMobile=1]
```

Notice the absence of efi.\* parameters in the rq parameter (because myModel does not use efi feature) and the presence of efi.\* parameters in the [features] part of the fl parameter (because myEfiFeatureStore contains efi features).

Read more about model evolution in the [LTR Lifecycle](#) section of this page.

## Training Example

Example training data and a demo train\_and\_upload\_demo\_model.py script can be found in the solr/contrib/ltr/example folder in the [Apache lucene-solr Git repository](#) (mirrored on [github.com](#)). This example folder is not shipped in the Solr binary release.

## Installation of LTR

The ltr contrib module requires the dist/solr-ltr-\*.jar JARs.

## LTR Configuration

Learning-To-Rank is a contrib module and therefore its plugins must be configured in solrconfig.xml.

### Minimum Requirements

- Include the required contrib JARs. Note that by default paths are relative to the Solr core so they may need adjustments to your configuration, or an explicit specification of the \$solr.install.dir.

```
<lib dir="${solr.install.dir:../../../../../}dist/" regex="solr-ltr-\d.*\.jar" />
```

- Declaration of the ltr query parser.

```
<queryParser name="ltr" class="org.apache.solr.ltr.search.LTRQParserPlugin"/>
```

- Configuration of the feature values cache.

```
<cache name="QUERY_DOC_FV"
 class="solr.search.LRUCache"
 size="4096"
 initialSize="2048"
 autowarmCount="4096"
 regenerator="solr.search.NoOpRegenerator" />
```

- Declaration of the [features] transformer.

```
<transformer name="features" class="org.apache.solr.ltr.response.transformLTRFeatureLoggerTransformerFactory">
<str name="fvCacheName">QUERY_DOC_FV</str>
</transformer>
```

## Advanced Options

### LTRThreadModule

A thread module can be configured for the query parser and/or the transformer to parallelize the creation of feature weights. For details, please refer to the [LTRThreadModule](#) javadocs.

### Feature Vector Customization

The features transformer returns dense CSV values such as  
featureA=0.1, featureB=0.2, featureC=0.3, featureD=0.0.

For sparse CSV output such as featureA:0.1 featureB:0.2 featureC:0.3 you can customize the [feature logger transformer](#) declaration in solrconfig.xml as follows:

```
<transformer name="features" class="org.apache.solr.ltr.response.transformLTRFeatureLoggerTransformerFactory">
<str name="fvCacheName">QUERY_DOC_FV</str>
<str name="defaultFormat">sparse</str>
<str name="csvKeyValueDelimiter">:</str>
<str name="csvFeatureSeparator"> </str>
</transformer>
```

## Implementation and Contributions

### How does Solr Learning-To-Rank work under the hood?

Please refer to the [ltr](#) javadocs for an implementation overview.

### How could I write additional models and/or features?

Contributions for further models, features and normalizers are welcome. Related links:

- [LTRScoringModel](#) javadocs
- [Feature](#) javadocs
- [Normalizer](#) javadocs
- <http://wiki.apache.org/solr/HowToContribute>
- <http://wiki.apache.org/lucene-java/HowToContribute>

## LTR Lifecycle

## Feature Stores

It is recommended that you organise all your features into stores which are akin to namespaces:

- Features within a store must be named uniquely.
- Across stores identical or similar features can share the same name.
- If no store name is specified then the default \_DEFAULT\_ feature store will be used.

To discover the names of all your feature stores:

```
http://localhost:8983/solr/techproducts/schema/feature-store
```

To inspect the content of the commonFeatureStore feature store:

```
http://localhost:8983/solr/techproducts/schema/feature-store/commonFeatureStore
```

## Models

- A model uses features from exactly one feature store.
- If no store is specified then the default \_DEFAULT\_ feature store will be used.
- A model need not use all the features defined in a feature store.
- Multiple models can use the same feature store.

To extract features for currentFeatureStore 's features:

```
http://localhost:8983/solr/techproducts/query?q=test&fl=id,score,[features store=currentFeatureStore]
```

To extract features for nextFeatureStore features whilst reranking with currentModel based on currentFeatureStore:

```
http://localhost:8983/solr/techproducts/query?q=test&rq={!ltr model=currentModel reRankDocs=100}&fl=id,score,[features store=nextFeatureStore]
```

To view all models:

```
http://localhost:8983/solr/techproducts/schema/model-store
```

To delete the currentModel model:

```
curl -XDELETE 'http://localhost:8983/solr/techproducts/schema/model-store/currentModel'
```



A feature store may be deleted only when there are no models using it.

To delete the currentFeatureStore feature store:

```
curl -XDELETE 'http://localhost:8983/solr/techproducts/schema/feature-store/currentFeatureStore'
```

## Using Large Models

With SolrCloud, large models may fail to upload due to the limitation of ZooKeeper's buffer. In this case, DefaultWrapperModel may help you to separate the model definition from uploaded file.

Assuming that you consider to use a large model placed at /path/to/models/myModel.json through DefaultWrapperModel.

```
{
 "store" : "largeModelsFeatureStore",
 "name" : "myModel",
 "class" : ...,
 "features" : [
 ...
],
 "params" : {
 ...
 }
}
```

First, add the directory to Solr's resource paths with a <lib/> directive:

```
<lib dir="/path/to" regex="models" />
```

Then, configure DefaultWrapperModel to wrap myModel.json:

```
{
 "store" : "largeModelsFeatureStore",
 "name" : "myWrapperModel",
 "class" : "org.apache.solr.ltr.model.DefaultWrapperModel",
 "params" : {
 "resource" : "myModel.json"
 }
}
```

myModel.json will be loaded during the initialization and be able to use by specifying model=myWrapperModel.



No "features" are configured in myWrapperModel because the features of the wrapped model (myModel) will be used; also note that the "store" configured for the wrapper model must match that of the wrapped model i.e., in this example the feature store called largeModelsFeatureStore is used.



<lib dir="/path/to/models" regex=".\*\.\.json" /> doesn't work as expected in this case, because SolrResourceLoader considers given resources as JAR if <lib /> indicates files.

As an alternative to the above-described DefaultWrapperModel, it is possible to [increase ZooKeeper's file size limit](#).

## Applying Changes

The feature store and the model store are both [Managed Resources](#). Changes made to managed resources are not applied to the active Solr components until the Solr collection (or Solr core in single server mode) is reloaded.

## LTR Examples

### One Feature Store, Multiple Ranking Models

- leftModel and rightModel both use features from commonFeatureStore and the only different between the two models is the weights attached to each feature.
- Conventions used:
  - commonFeatureStore.json file contains features for the commonFeatureStore feature store
  - leftModel.json file contains model named leftModel
  - rightModel.json file contains model named rightModel
  - The model's features and weights are sorted alphabetically by name, this makes it easy to see what the commonalities and differences between the two models are.
  - The stores features are sorted alphabetically by name, this makes it easy to lookup features used in the models

Example: /path/commonFeatureStore.json

```
[
 {
 "store" : "commonFeatureStore",
 "name" : "documentRecency",
 "class" : "org.apache.solr.ltr.feature.SolrFeature",
 "params" : {
 "q" : "{!func}recip(ms(NOW,last_modified), 3.16e-11, 1, 1)"
 }
 },
 {
 "store" : "commonFeatureStore",
 "name" : "isBook",
 "class" : "org.apache.solr.ltr.feature.SolrFeature",
 "params" : {
 "fq": ["{!terms f=category}book"]
 }
 },
 {
 "store" : "commonFeatureStore",
 "name" : "originalScore",
 "class" : "org.apache.solr.ltr.feature.OriginalScoreFeature",
 "params" : {}
 }
]
```

Example: /path/leftModel.json

```
{
 "store" : "commonFeatureStore",
 "name" : "leftModel",
 "class" : "org.apache.solr.ltr.model.LinearModel",
 "features" : [
 { "name" : "documentRecency" },
 { "name" : "isBook" },
 { "name" : "originalScore" }
],
 "params" : {
 "weights" : {
 "documentRecency" : 0.1,
 "isBook" : 1.0,
 "originalScore" : 0.5
 }
 }
}
```

Example: /path/rightModel.json

```
{
 "store" : "commonFeatureStore",
 "name" : "rightModel",
 "class" : "org.apache.solr.ltr.model.LinearModel",
 "features" : [
 { "name" : "documentRecency" },
 { "name" : "isBook" },
 { "name" : "originalScore" }
,
 "params" : {
 "weights" : {
 "documentRecency" : 1.0,
 "isBook" : 0.1,
 "originalScore" : 0.5
 }
 }
}
```

## Model Evolution

- linearModel201701 uses features from featureStore201701
- treesModel201702 uses features from featureStore201702
- linearModel201701 and treesModel201702 and their feature stores can co-exist whilst both are needed.
- When linearModel201701 has been deleted then featureStore201701 can also be deleted.
- Conventions used:
  - <store>.json file contains features for the <store> feature store
  - <model>.json file contains model name <model>
  - a 'generation' id (e.g., YYYYMM year-month) is part of the feature store and model names
  - The model's features and weights are sorted alphabetically by name, this makes it easy to see what the commonalities and differences between the two models are.
  - The stores features are sorted alphabetically by name, this makes it easy to see what the commonalities and differences between the two feature stores are.

Example: /path/featureStore201701.json

```
[
 {
 "store" : "featureStore201701",
 "name" : "documentRecency",
 "class" : "org.apache.solr.ltr.feature.SolrFeature",
 "params" : {
 "q" : "{!func}recip(ms(NOW,last_modified), 3.16e-11, 1, 1)"
 }
 },
 {
 "store" : "featureStore201701",
 "name" : "isBook",
 "class" : "org.apache.solr.ltr.feature.SolrFeature",
 "params" : {
 "fq": ["{!terms f=category}book"]
 }
 },
 {
 "store" : "featureStore201701",
 "name" : "originalScore",
 "class" : "org.apache.solr.ltr.feature.OriginalScoreFeature",
 "params" : {}
 }
]
```

Example: /path/linearModel201701.json

```
{
 "store" : "featureStore201701",
 "name" : "linearModel201701",
 "class" : "org.apache.solr.ltr.model.LinearModel",
 "features" : [
 { "name" : "documentRecency" },
 { "name" : "isBook" },
 { "name" : "originalScore" }
],
 "params" : {
 "weights" : {
 "documentRecency" : 0.1,
 "isBook" : 1.0,
 "originalScore" : 0.5
 }
 }
}
```

Example: /path/featureStore201702.json

```
[
 {
 "store" : "featureStore201702",
 "name" : "isBook",
 "class" : "org.apache.solr.ltr.feature.SolrFeature",
 "params" : {
 "fq": ["{!terms f=category}book"]
 }
 },
 {
 "store" : "featureStore201702",
 "name" : "originalScore",
 "class" : "org.apache.solr.ltr.feature.OriginalScoreFeature",
 "params" : {}
 }
]
```

Example: /path/treesModel201702.json

```
{
 "store" : "featureStore201702",
 "name" : "treesModel201702",
 "class" : "org.apache.solr.ltr.model.MultipleAdditiveTreesModel",
 "features" : [
 { "name" : "isBook" },
 { "name" : "originalScore" }
,
 "params" : {
 "trees" : [
 {
 "weight" : "1",
 "root" : {
 "feature" : "isBook",
 "threshold" : "0.5",
 "left" : { "value" : "-100" },
 "right" : {
 "feature" : "originalScore",
 "threshold" : "10.0",
 "left" : { "value" : "50" },
 "right" : { "value" : "75" }
 }
 }
 }
 },
 {
 "weight" : "2",
 "root" : {
 "value" : "-10"
 }
 }
]
}
```

## Additional LTR Resources

- "Learning to Rank in Solr" presentation at Lucene/Solr Revolution 2015 in Austin:
  - Slides: <http://www.slideshare.net/lucidworks/learning-to-rank-in-solr-presented-by-michael-nilsson-diego-ceccarelli-bloomberg-lp>
  - Video: <https://www.youtube.com/watch?v=M7BKwJoh96s>

# Transforming Result Documents

Document Transformers modify the information returned about documents in the results of a query.

## Using Document Transformers

When executing a request, a document transformer can be used by including it in the f1 parameter using square brackets, for example:

```
f1=id,name,score,[shard]
```

Some transformers allow, or require, local parameters which can be specified as key value pairs inside the brackets:

```
f1=id,name,score,[explain style=n1]
```

As with regular fields, you can change the key used when a Transformer adds a field to a document via a prefix:

```
f1=id,name,score,my_val_a:[value v=42 t=int],my_val_b:[value v=7 t=float]
```

The sections below discuss exactly what these various transformers do.

## Available Transformers

### [value] - ValueAugmenterFactory

Modifies every document to include the exact same value, as if it were a stored field in every document:

```
q=*&f1=id,greeting:[value v='hello']&wt=xml
```

The above query would produce results like the following:

```
<result name="response" numFound="32" start="0">
<doc>
 <str name="id">1</str>
 <str name="greeting">hello</str></doc>
</doc>
...
```

By default, values are returned as a String, but a t parameter can be specified using a value of int, float, double, or date to force a specific return type:

```
q=*&fl=id,my_number:[value v=42 t=int],my_string:[value v=42]
```

In addition to using these request parameters, you can configure additional named instances of ValueAugmenterFactory, or override the default behavior of the existing [value] transformer in your solrconfig.xml file:

```
<transformer name="mytrans2" class="org.apache.solr.response.transform.ValueAugmenterFactory" >
 <int name="value">5</int>
</transformer>
<transformer name="value" class="org.apache.solr.response.transform.ValueAugmenterFactory" >
 <double name="defaultValue">5</double>
</transformer>
```

The value option forces an explicit value to always be used, while the defaultValue option provides a default that can still be overridden using the v and t local parameters.

## [explain] - ExplainAugmenterFactory

Augments each document with an inline explanation of its score exactly like the information available about each document in the debug section:

```
q=features:cache&fl=id,[explain style=n1]
```

Supported values for style are text, html, and n1 which returns the information as structured data. Here is the output of the above request using style=n1:

```
{ "response":{ "numFound":2, "start":0, "docs":[
 {
 "id":"6H500F0",
 "[explain]:{
 "match":true,
 "value":1.052226,
 "description":"weight(features:cache in 2) [DefaultSimilarity], result of:",
 "details":[]}}]}}
```

A default style can be configured by specifying an args parameter in your solrconfig.xml configuration:

```
<transformer name="explain" class="org.apache.solr.response.transform.ExplainAugmenterFactory" >
 <str name="args">n1</str>
</transformer>
```

## [child] - ChildDocTransformerFactory

This transformer returns all [descendant documents](#) of each parent document matching your query in a flat list nested inside the matching parent document. This is useful when you have indexed nested child

documents and want to retrieve the child documents for the relevant parent documents for any type of search query.

```
f1=id,[child parentFilter=doc_type:book childFilter=doc_type:chapter limit=100]
```

Note that this transformer can be used even though the query itself is not a [Block Join query](#).

When using this transformer, the parentFilter parameter must be specified *unless* the schema declares `_nest_path_`. It works the same as in all Block Join Queries. Additional optional parameters are:

#### childFilter

A query to filter which child documents should be included. This can be particularly useful when you have multiple levels of hierarchical documents. The default is all children. This query supports a special syntax to match nested doc patterns so long as `_nest_path_` is defined in the schema and the query contains a / preceding the first :. Example: `childFilter=/comments/content:recipe` Further details of this are experimental.

#### limit

The maximum number of child documents to be returned per parent document. The default is 10.

#### f1

The field list which the transformer is to return. The default is the top level f1).

There is a further limitation in which the fields here should be a subset of those specified by the top level f1 parameter.

## [shard] - ShardAugmenterFactory

This transformer adds information about what shard each individual document came from in a distributed request.

ShardAugmenterFactory does not support any request parameters, or configuration options.

## [docid] - DocIdAugmenterFactory

This transformer adds the internal Lucene document id to each document – this is primarily only useful for debugging purposes.

DocIdAugmenterFactory does not support any request parameters, or configuration options.

## [elevated] and [excluded]

These transformers are available only when using the [Query Elevation Component](#).

- [elevated] annotates each document to indicate if it was elevated or not.
- [excluded] annotates each document to indicate if it would have been excluded - this is only supported if you also use the markExcludes parameter.

```
f1=id,[elevated],[excluded]&excludeIds=GB18030TEST&elevateIds=6H500F0&markExcludes=true
```

```
{ "response":{ "numFound":32, "start":0, "docs":[
 {
 "id":"6H500F0",
 "[elevated]":true,
 "[excluded]":false},
 {
 "id":"GB18030TEST",
 "[elevated]":false,
 "[excluded]":true},
 {
 "id":"SP2514N",
 "[elevated]":false,
 "[excluded]":false},
]]}
```

## [json] / [xml]

These transformers replace a field value containing a string representation of a valid XML or JSON structure with the actual raw XML or JSON structure instead of just the string value. Each applies only to the specific writer, such that [json] only applies to `wt=json` and [xml] only applies to `wt=xml`.

```
fl=id,source_s:[json]&wt=json
```

## [subquery]

This transformer executes a separate query per transforming document passing document fields as an input for subquery parameters. It's usually used with `{!join}` and `{!parent}` query parsers, and is intended to be an improvement for [child].

- It must be given an unique name: `fl=*,children:[subquery]`
- There might be a few of them, e.g., `fl=*,sons:[subquery],daughters:[subquery]`.
- Every [subquery] occurrence adds a field into a result document with the given name, the value of this field is a document list, which is a result of executing subquery using document fields as an input.
- Subquery will use the /select search handler by default, and will return an error if /select is not configured. This can be changed by supplying `foo.qt` parameter.

Here is how it looks like using various formats:

**XML**

```

<result name="response" numFound="2" start="0">
 <doc>
 <int name="id">1</int>
 <arr name="title">
 <str>vdczoypirs</str>
 </arr>
 <result name="children" numFound="1" start="0">
 <doc>
 <int name="id">2</int>
 <arr name="title">
 <str>vdczoypirs</str>
 </arr>
 </doc>
 </result>
 </doc>
 ...

```

**JSON**

```

{
 "response": {
 "numFound": 2, "start": 0,
 "docs": [
 {
 "id": 1,
 "subject": ["parentDocument"],
 "title": ["xrxvomgu"],
 "children": {
 "numFound": 1, "start": 0,
 "docs": [
 { "id": 2,
 "cat": ["childDocument"]
 }
]
 }
 }
]
 }
}

```

**SolrJ**

```
SolrDocumentList subResults = (SolrDocumentList) doc.getFieldValue("children");
```

**Subquery Result Fields**

To appear in subquery document list, a field should be specified in both f1 parameters: in the main f1 (despite the main result documents have no this field), and in subquery's f1 (e.g., foo.f1).

Wildcards can be used in one or both of these parameters. For example, if field title should appear in categories subquery, it can be done via one of these ways:

```
fl=...title, categories:[subquery]&categories.fl=title&categories.q=...
fl=...title, categories:[subquery]&categories.fl=*&categories.q=...
fl=...*, categories:[subquery]&categories.fl=title&categories.q=...
fl=...*, categories:[subquery]&categories.fl=*&categories.q=...
```

## Subquery Parameters Shift

If a subquery is declared as `fl=*`, `foo:[subquery]`, subquery parameters are prefixed with the given name and period. For example:

```
q=*:*&fl=*,**foo**:[subquery]&**foo.**q=to be continued&**foo.**rows=10&**foo.**sort=id desc
```

## Document Field as an Input for Subquery Parameters

It's necessary to pass some document field values as a parameter for subquery. It's supported via an implicit `row.fieldname` parameter, and can be (but might not only) referred via local parameters syntax:

```
q=name:john&fl=name,id,depts:[subquery]&depts.q={!terms f=id
v=$row.dept_id}&depts.rows=10
```

Here departments are retrieved per every employee in search result. We can say that it's like SQL join ON `emp.dept_id=dept.id`.

Note, when a document field has multiple values they are concatenated with a comma by default. This can be changed with the local parameter `foo:[subquery separator=' ']`, this mimics `{!terms}` to work smoothly with it.

To log substituted subquery request parameters, add the corresponding parameter names, as in:  
`depts.logParamsList=q,f1,rows, row.dept_id`

## Cores and Collections in SolrCloud

Use `foo:[subquery fromIndex=departments]` to invoke subquery on another core on the same node. This is what `{!join}` does for non-SolrCloud mode. But with SolrCloud, just (and only) explicitly specify its native parameters like collection, shards for subquery, for example:

```
q=*:*&fl=*,foo:[subquery]&foo.q=cloud&foo.collection=departments
```

If subquery collection has a different unique key field name (such as `foo_id` instead of `id` in the primary collection), add the following parameters to accommodate this difference:



```
foo.fl=id:foo_id&foo.distrib.singlePass=true
```

Otherwise you'll get `NullPointerException` from `QueryComponent.mergeIds`.

## [geo] - Geospatial formatter

Formats spatial data from a spatial field using a designated format type name. Two inner parameters are required: f for the field name, and w for the format name. Example: geojson:[geo f=mySpatialField w=GeoJSON].

Normally you'll simply be consistent in choosing the format type you want by setting the format attribute on the spatial field type to WKT or GeoJSON – see the section [Spatial Search](#) for more information. If you are consistent, it'll come out the way you stored it. This transformer offers a convenience to transform the spatial format to something different on retrieval.

In addition, this feature is very useful with the RptWithGeometrySpatialField to avoid double-storage of the potentially large vector geometry. This transformer will detect that field type and fetch the geometry from an internal compact binary representation on disk (in docValues), and then format it as desired. As such, you needn't mark the field as stored, which would be redundant. In a sense this double-storage between docValues and stored-value storage isn't unique to spatial but with polygonal geometry it can be a lot of data, and furthermore you'd like to avoid storing it in a verbose format (like GeoJSON or WKT).

## [features] - LTRFeatureLoggerTransformerFactory

The "LTR" prefix stands for [Learning To Rank](#). This transformer returns the values of features and it can be used for feature extraction and feature logging.

```
fl=id,[features store=yourFeatureStore]
```

This will return the values of the features in the yourFeatureStore store.

```
fl=id,[features]&rq={!ltr model=yourModel}
```

If you use [features] together with an Learning-To-Rank reranking query then the values of the features in the reranking model (yourModel) will be returned.

# Searching Nested Child Documents

This section exposes potential techniques which can be used for searching deeply nested documents, showing how more complex queries can be constructed using some of Solr's query parsers and Doc Transformers. These features require `_root_` and `_nest_path_` to be declared in the schema. Please refer to the [Indexing Nested Documents](#) section for more details about schema and index configuration.



This section does not show case facetting on nested documents. For nested document facetting, please refer to the [Block Join Faceting](#) section.

## Query Examples

For the upcoming examples, assume the following documents have been indexed:

```
[
 {
 "ID": "1",
 "title": "Cooking Recommendations",
 "tags": ["cooking", "meetup"],
 "posts": [{
 "ID": "2",
 "title": "Cookies",
 "comments": [{
 "ID": "3",
 "content": "Lovely recipe"
 }],
 {
 "ID": "4",
 "content": "A-"
 }
 }],
 {
 "ID": "5",
 "title": "Cakes"
 }
 },
],
 {
 "ID": "6",
 "title": "For Hire",
 "tags": ["professional", "jobs"],
 "posts": [{
 "ID": "7",
 "title": "Search Engineer",
 "comments": [{
 "ID": "8",
 "content": "I am interested"
 }],
 {
 "ID": "9",
 "content": "How large is the team?"
 }
 }],
 {
 "ID": "10",
 "title": "Low level Engineer"
 }
 }]
]
```

## Child Doc Transformer

Can be used enrich query results with the documents' descendants.

For a detailed explanation of this transformer, see the section [\[child\] - ChildDocTransformerFactory](#).

For example, let us examine this query: q=ID:1, fl=ID,[child childFilter=/comments/content:recipe]. The Child Doc Transformer can be used to enrich matching docs with comments that match a particular filter.

In this particular query, the child Filter will only match the first comment of doc(ID:1), therefore only that particular comment will be appended to the result. This is a special syntax feature.

```
{ "response":{ "numFound":1, "start":0, "docs":[
 {
 "ID": "1",
 "title": "Cooking Recommendations",
 "tags": ["cooking", "meetup"],
 "posts": [
 {
 "ID": "2",
 "title": "Cookies",
 "comments": [
 {
 "ID": "3",
 "content": "Lovely recipe"
 }
]
 }
]
 }
]}
```

## Children Query Parser

Can be used to retrieve children of a matching document.

For a detailed explanation of this parser, see the section [Block Join Children Query Parser](#).

For example, let us examine this query: q={!child of='nest\_path:/posts}content:"Search Engineer". The 'of' filter returns all posts. This is used to filter out all documents in a particular path of the hierarchy(all parents). The second part of the query is a filter for some parents, which we wish to return their children.

In this example, all comments of posts which had "Search Engineer in their content field will be returned.

```
{ "response":{ "numFound":2, "start":0, "docs":[
 {
 "ID": "8",
 "content": "I am interested"
 },
 {
 "ID": "9",
 "content": "How large is the team?"
 }
]}
```

## Parents Query Parser

Can be used to retrieve parents of a child document.

For a detailed explanation of this parser, see the section [Block Join Parent Query Parser](#).

For example, let us examine this query: q={!parent which='-\_nest\_path\_:\* \*:\*'}title:"Search Engineer".

The 'which' filter returns all root documents. The second part of this query is a filter to match some child documents. This query returns the parent at the root(since all parents filter returns root documents) of each matching child document. In this case, all child documents which had Search Engineer in their title field.

```
{ "response":{ "numFound":1, "start":0, "docs":[
 {
 "ID": "6",
 "title": "For Hire",
 "tags": ["professional", "jobs"]
 }
]}
```

## Combining Block Join Query Parsers with Child Doc Transformer

The combination of these two features enable seamless creation of powerful queries.

For example, querying posts which are under a page tagged as a job, contain the words "Search Engineer".

The comments for matching posts can also be fetched, all done in a single Solr Query.

For example, let us examine this query: q=+{!child of='-\_nest\_path\_:\* \*:\*'}+tags:"jobs" &fl=\*,[child] &fq=\_nest\_path\_:/posts.

This query returns all posts and their comments, which had "Search Engineer" in their title, and are indexed under a page tagged with "jobs". The comments are appended to the matching posts, since the ChildDocTransformer is specified under the fl parameter.

```
{ "response":{ "numFound":1, "start":0, "docs":[
 {
 "ID": "7",
 "title": "Search Engineer",
 "comments": [{

 "ID": "8",
 "content": "I am interested"
 },
 {
 "ID": "9",
 "content": "How large is the team?"
 }
],
 {
 "ID": "10",
 "title": "Low level Engineer"
 }]
}]}
}
```

# Suggester

The SuggestComponent in Solr provides users with automatic suggestions for query terms.

You can use this to implement a powerful auto-suggest feature in your search application.

Although it is possible to use the [Spell Checking](#) functionality to power autosuggest behavior, Solr has a dedicated [SuggestComponent](#) designed for this functionality.

This approach utilizes Lucene's Suggester implementation and supports all of the lookup implementations available in Lucene.

The main features of this Suggester are:

- Lookup implementation pluggability
- Term dictionary pluggability, giving you the flexibility to choose the dictionary implementation
- Distributed support

The solrconfig.xml found in Solr's "techproducts" example has a Suggester implementation configured already. For more on search components, see the section [RequestHandlers and SearchComponents in SolrConfig](#).

The "techproducts" example solrconfig.xml has a suggest search component and a /suggest request handler already configured. You can use that as the basis for your configuration, or create it from scratch, as detailed below.

## Adding the Suggest Search Component

The first step is to add a search component to solrconfig.xml and tell it to use the SuggestComponent. Here is some sample code that could be used.

```
<searchComponent name="suggest" class="solr.SuggestComponent">
 <lst name="suggester">
 <str name="name">mySuggester</str>
 <str name="lookupImpl">FuzzyLookupFactory</str>
 <str name="dictionaryImpl">DocumentDictionaryFactory</str>
 <str name="field">cat</str>
 <str name="weightField">price</str>
 <str name="suggestAnalyzerFieldType">string</str>
 <str name="buildOnStartup">false</str>
 </lst>
</searchComponent>
```

## Suggester Search Component Parameters

The Suggester search component takes several configuration parameters.

The choice of the lookup implementation (lookupImpl, how terms are found in the suggestion dictionary) and the dictionary implementation (dictionaryImpl, how terms are stored in the suggestion dictionary) will

dictate some of the parameters required.

Below are the main parameters that can be used no matter what lookup or dictionary implementation is used. In the following sections additional parameters are provided for each implementation.

#### searchComponent name

Arbitrary name for the search component.

#### name

A symbolic name for this suggester. You can refer to this name in the URL parameters and in the SearchHandler configuration. It is possible to have multiples of these in one solrconfig.xml file.

#### lookupImpl

Lookup implementation. There are several possible implementations, described below in the section [Lookup Implementations](#). If not set, the default lookup is JaspellLookupFactory.

#### dictionaryImpl

The dictionary implementation to use. There are several possible implementations, described below in the section [Dictionary Implementations](#).

If not set, the default dictionary implementation is HighFrequencyDictionaryFactory. However, if a sourceLocation is used, the dictionary implementation will be FileDictionaryFactory.

#### field

A field from the index to use as the basis of suggestion terms. If sourceLocation is empty (meaning any dictionary implementation other than FileDictionaryFactory), then terms from this field in the index will be used.

To be used as the basis for a suggestion, the field must be stored. You may want to [use copyField rules](#) to create a special 'suggest' field comprised of terms from other fields in documents. In any event, you very likely want a minimal amount of analysis on the field, so an additional option is to create a field type in your schema that only uses basic tokenizers or filters. One option for such a field type is shown here:

```
<fieldType class="solr.TextField" name="textSuggest" positionIncrementGap="100">
 <analyzer>
 <tokenizer class="solr.StandardTokenizerFactory"/>
 <filter class="solr.LowerCaseFilterFactory"/>
 </analyzer>
</fieldType>
```

However, this minimal analysis is not required if you want more analysis to occur on terms. If using the AnalyzingLookupFactory as your lookupImpl, however, you have the option of defining the field type rules to use for index and query time analysis.

#### sourceLocation

The path to the dictionary file if using the FileDictionaryFactory. If this value is empty then the main index will be used as a source of terms and weights.

#### storeDir

The location to store the dictionary file.

**buildOnCommit and buildOnOptimize**

If true, the lookup data structure will be rebuilt after soft-commit. If false, the default, then the lookup data will be built only when requested by URL parameter suggest.build=true. Use buildOnCommit to rebuild the dictionary with every soft-commit, or buildOnOptimize to build the dictionary only when the index is optimized.

Some lookup implementations may take a long time to build, especially with large indexes. In such cases, using buildOnCommit or buildOnOptimize, particularly with a high frequency of softCommits is not recommended; it's recommended instead to build the suggester at a lower frequency by manually issuing requests with suggest.build=true.

**buildOnStartup**

If true, then the lookup data structure will be built when Solr starts or when the core is reloaded. If this parameter is not specified, the suggester will check if the lookup data structure is present on disk and build it if not found.

Enabling this to true could lead to the core talking longer to load (or reload) as the suggester data structure needs to be built, which can sometimes take a long time. It's usually preferred to have this setting set to false, the default, and build suggesters manually issuing requests with suggest.build=true.

## Lookup Implementations

The lookupImpl parameter defines the algorithms used to look up terms in the suggest index. There are several possible implementations to choose from, and some require additional parameters to be configured.

### AnalyzingLookupFactory

A lookup that first analyzes the incoming text and adds the analyzed form to a weighted FST, and then does the same thing at lookup time.

This implementation uses the following additional properties:

**suggestAnalyzerFieldType**

The field type to use for the query-time and build-time term suggestion analysis.

**exactMatchFirst**

If true, the default, exact suggestions are returned first, even if they are prefixes or other strings in the FST have larger weights.

**preserveSep**

If true, the default, then a separator between tokens is preserved. This means that suggestions are sensitive to tokenization (e.g., baseball is different from base ball).

**preservePositionIncrements**

If true, the suggester will preserve position increments. This means that token filters which leave gaps (for example, when StopFilter matches a stopword) the position would be respected when building the suggester. The default is false.

## FuzzyLookupFactory

This is a suggester which is an extension of the AnalyzingSuggester but is fuzzy in nature. The similarity is measured by the Levenshtein algorithm.

This implementation uses the following additional properties:

### exactMatchFirst

If true, the default, exact suggestions are returned first, even if they are prefixes or other strings in the FST have larger weights.

### preserveSep

If true, the default, then a separator between tokens is preserved. This means that suggestions are sensitive to tokenization (e.g., baseball is different from base ball).

### maxSurfaceFormsPerAnalyzedForm

The maximum number of surface forms to keep for a single analyzed form. When there are too many surface forms we discard the lowest weighted ones.

### maxGraphExpansions

When building the FST ("index-time"), we add each path through the tokenstream graph as an individual entry. This places an upper-bound on how many expansions will be added for a single suggestion. The default is -1 which means there is no limit.

### preservePositionIncrements

If true, the suggester will preserve position increments. This means that token filters which leave gaps (for example, when StopFilter matches a stopword) the position would be respected when building the suggester. The default is false.

### maxEdits

The maximum number of string edits allowed. The system's hard limit is 2. The default is 1.

### transpositions

If true, the default, transpositions should be treated as a primitive edit operation.

### nonFuzzyPrefix

The length of the common non fuzzy prefix match which must match a suggestion. The default is 1.

### minFuzzyLength

The minimum length of query before which any string edits will be allowed. The default is 3.

### unicodeAware

If true, the maxEdits, minFuzzyLength, transpositions and nonFuzzyPrefix parameters will be measured in unicode code points (actual letters) instead of bytes. The default is false.

## AnalyzingInfixLookupFactory

Analyzes the input text and then suggests matches based on prefix matches to any tokens in the indexed text. This uses a Lucene index for its dictionary.

This implementation uses the following additional properties.

**indexPath**

When using AnalyzingInfixSuggester you can provide your own path where the index will get built. The default is `analyzingInfixSuggesterIndexDir` and will be created in your collection's data/ directory.

**minPrefixChars**

Minimum number of leading characters before PrefixQuery is used (default is 4). Prefixes shorter than this are indexed as character ngrams (increasing index size but making lookups faster).

**allTermsRequired**

Boolean option for multiple terms. The default is true, all terms will be required.

**highlight**

Highlight suggest terms. Default is true.

This implementation supports [Context Filtering](#).

## BlendedInfixLookupFactory

An extension of the AnalyzingInfixSuggester which provides additional functionality to weight prefix matches across the matched documents. It scores higher if a hit is closer to the start of the suggestion.

This implementation uses the following additional properties:

**blenderType**

Used to calculate weight coefficient using the position of the first matching word. Available options are:

**position\_linear**

`weightFieldValue * (1 - 0.10*position)`: Matches to the start will be given a higher score. This is the default.

**position\_reciprocal**

`weightFieldValue / (1 + position)`: Matches to the start will be given a higher score. The score of matches positioned far from the start of the suggestion decays faster than linear.

**position\_exponential\_reciprocal**

`weightFieldValue / pow(1 + position, exponent)`: Matches to the start will be given a higher score. The score of matches positioned far from the start of the suggestion decays faster than reciprocal.

**exponent**

An optional configuration variable for `position_exponential_reciprocal` to control how fast the score will decrease. Default 2.0.

**numFactor**

The factor to multiply the number of searched elements from which results will be pruned. Default is 10.

**indexPath**

When using BlendedInfixSuggester you can provide your own path where the index will get built. The default directory name is `blendedInfixSuggesterIndexDir` and will be created in your collection's data directory.

**minPrefixChars**

Minimum number of leading characters before PrefixQuery is used (the default is 4). Prefixes shorter

than this are indexed as character ngrams, which increases index size but makes lookups faster.

This implementation supports [Context Filtering](#).

### FreeTextLookupFactory

It looks at the last tokens plus the prefix of whatever final token the user is typing, if present, to predict the most likely next token. The number of previous tokens that need to be considered can also be specified. This suggester would only be used as a fallback, when the primary suggester fails to find any suggestions.

This implementation uses the following additional properties:

suggestFreeTextAnalyzerFieldType

The analyzer used at "query-time" and "build-time" to analyze suggestions. This parameter is required.

ngrams

The max number of tokens out of which singles will be made the dictionary. The default value is 2.

Increasing this would mean you want more than the previous 2 tokens to be taken into consideration when making the suggestions.

### FSTLookupFactory

An automaton-based lookup. This implementation is slower to build, but provides the lowest memory cost. We recommend using this implementation unless you need more sophisticated matching results, in which case you should use the Jaspell implementation.

This implementation uses the following additional properties:

exactMatchFirst

If true, the default, exact suggestions are returned first, even if they are prefixes or other strings in the FST have larger weights.

weightBuckets

The number of separate buckets for weights which the suggester will use while building its dictionary.

### TSTLookupFactory

A simple compact ternary trie based lookup.

### WFSTLookupFactory

A weighted automaton representation which is an alternative to FSTLookup for more fine-grained ranking. WFSTLookup does not use buckets, but instead a shortest path algorithm.

Note that it expects weights to be whole numbers. If weight is missing it's assumed to be 1.0. Weights affect the sorting of matching suggestions when spellcheck.onlyMorePopular=true is selected: weights are treated as "popularity" score, with higher weights preferred over suggestions with lower weights.

### JaspellLookupFactory

A more complex lookup based on a ternary trie from the [JaSpell](#) project. Use this implementation if you need more sophisticated matching results.

## Dictionary Implementations

The dictionary implementations define how terms are stored. There are several options, and multiple dictionaries can be used in a single request if necessary.

### DocumentDictionaryFactory

A dictionary with terms, weights, and an optional payload taken from the index.

This dictionary implementation takes the following parameters in addition to parameters described for the Suggester generally and for the lookup implementation:

`weightField`

A field that is stored or a numeric DocValue field. This parameter is optional.

`payloadField`

The payloadField should be a field that is stored. This parameter is optional.

`contextField`

Field to be used for context filtering. Note that only some lookup implementations support filtering.

### DocumentExpressionDictionaryFactory

This dictionary implementation is the same as the DocumentDictionaryFactory but allows users to specify an arbitrary expression into the weightExpression tag.

This dictionary implementation takes the following parameters in addition to parameters described for the Suggester generally and for the lookup implementation:

`payloadField`

The payloadField should be a field that is stored. This parameter is optional.

`weightExpression`

An arbitrary expression used for scoring the suggestions. The fields used must be numeric fields. This parameter is required.

`contextField`

Field to be used for context filtering. Note that only some lookup implementations support filtering.

### HighFrequencyDictionaryFactory

This dictionary implementation allows adding a threshold to prune out less frequent terms in cases where very common terms may overwhelm other terms.

This dictionary implementation takes one parameter in addition to parameters described for the Suggester generally and for the lookup implementation:

`threshold`

A value between zero and one representing the minimum fraction of the total documents where a term should appear in order to be added to the lookup dictionary.

## FileDictionaryFactory

This dictionary implementation allows using an external file that contains suggest entries. Weights and payloads can also be used.

If using a dictionary file, it should be a plain text file in UTF-8 encoding. You can use both single terms and phrases in the dictionary file. If adding weights or payloads, those should be separated from terms using the delimiter defined with the `fieldDelimiter` property (the default is '\t', the tab representation). If using payloads, the first line in the file **must** specify a payload.

This dictionary implementation takes one parameter in addition to parameters described for the Suggester generally and for the lookup implementation:

### fieldDelimiter

Specifies the delimiter to be used separating the entries, weights and payloads. The default is tab (\t).

#### Example File

```
acquire
accidentally 2.0
accommodate 3.0
```

## Multiple Dictionaries

It is possible to include multiple `dictionaryImpl` definitions in a single `SuggestComponent` definition.

To do this, simply define separate suggesters, as in this example:

```
<searchComponent name="suggest" class="solr.SuggestComponent">
 <lst name="suggester">
 <str name="name">mySuggester</str>
 <str name="lookupImpl">FuzzyLookupFactory</str>
 <str name="dictionaryImpl">DocumentDictionaryFactory</str>
 <str name="field">cat</str>
 <str name="weightField">price</str>
 <str name="suggestAnalyzerFieldType">string</str>
 </lst>
 <lst name="suggester">
 <str name="name">altSuggester</str>
 <str name="dictionaryImpl">DocumentExpressionDictionaryFactory</str>
 <str name="lookupImpl">FuzzyLookupFactory</str>
 <str name="field">product_name</str>
 <str name="weightExpression">((price * 2) + ln(popularity))</str>
 <str name="sortField">weight</str>
 <str name="sortField">price</str>
 <str name="storeDir">suggest_fuzzy_doc_expr_dict</str>
 <str name="suggestAnalyzerFieldType">text_en</str>
 </lst>
</searchComponent>
```

When using these Suggesters in a query, you would define multiple `suggest.dictionary` parameters in the

request, referring to the names given for each Suggester in the search component definition. The response will include the terms in sections for each Suggester. See the [Example Usages](#) section below for an example request and response.

## Adding the Suggest Request Handler

After adding the search component, a request handler must be added to `solrconfig.xml`. This request handler works the [same as any other request handler](#), and allows you to configure default parameters for serving suggestion requests. The request handler definition must incorporate the "suggest" search component defined previously.

```
<requestHandler name="/suggest" class="solr.SearchHandler" startup="lazy">
 <lst name="defaults">
 <str name="suggest">true</str>
 <str name="suggest.count">10</str>
 </lst>
 <arr name="components">
 <str>suggest</str>
 </arr>
</requestHandler>
```

## Suggest Request Handler Parameters

The following parameters allow you to set defaults for the Suggest request handler:

`suggest=true`

This parameter should always be true, because we always want to run the Suggester for queries submitted to this handler.

`suggest.dictionary`

The name of the dictionary component configured in the search component. This is a mandatory parameter. It can be set in the request handler, or sent as a parameter at query time.

`suggest.q`

The query to use for suggestion lookups.

`suggest.count`

Specifies the number of suggestions for Solr to return.

`suggest.cfg`

A Context Filter Query used to filter suggestions based on the context field, if supported by the suggester.

`suggest.build`

If true, it will build the suggester index. This is likely useful only for initial requests; you would probably not want to build the dictionary on every request, particularly in a production system. If you would like to keep your dictionary up to date, you should use the `buildOnCommit` or `buildOnOptimize` parameter for the search component.

`suggest.reload`

If true, it will reload the suggester index.

suggest.buildAll

If true, it will build all suggester indexes.

suggest.reloadAll

If true, it will reload all suggester indexes.

These properties can also be overridden at query time, or not set in the request handler at all and always sent at query time.

#### *Context Filtering*



Context filtering (`suggest.cfq`) is currently only supported by `AnalyzingInfixLookupFactory` and `BlendedInfixLookupFactory`, and only when backed by a `Document*Dictionary`. All other implementations will return unfiltered matches as if filtering was not requested.

## Example Usages

### Get Suggestions with Weights

This is a basic suggestion using a single dictionary and a single Solr core.

Example query:

```
http://localhost:8983/solr/techproducts/suggest?suggest=true&suggest.build=true&suggest.dictionary=mySuggester&suggest.q=elec
```

In this example, we've simply requested the string 'elec' with the `suggest.q` parameter and requested that the suggestion dictionary be built with `suggest.build` (note, however, that you would likely not want to build the index on every query - instead you should use `buildOnCommit` or `buildOnOptimize` if you have regularly changing documents).

Example response:

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 35
 },
 "command": "build",
 "suggest": {
 "mySuggester": {
 "elec": {
 "numFound": 3,
 "suggestions": [
 {
 "term": "electronics and computer1",
 "weight": 2199,
 "payload": ""
 },
 {
 "term": "electronics",
 "weight": 649,
 "payload": ""
 },
 {
 "term": "electronics and stuff2",
 "weight": 279,
 "payload": ""
 }
]
 }
 }
 }
}
```

## Using Multiple Dictionaries

If you have defined multiple dictionaries, you can use them in queries.

Example query:

```
http://localhost:8983/solr/techproducts/suggest?suggest=true&suggest.dictionary=mySuggester&sugge
st.dictionary=altSuggester&suggest.q=elec
```

In this example we have sent the string 'elec' as the suggest.q parameter and named two suggest.dictionary definitions to be used.

Example response:

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 3
 },
 "suggest": {
 "mySuggester": {
 "elec": {
 "numFound": 1,
 "suggestions": [
 {
 "term": "electronics and computer1",
 "weight": 100,
 "payload": ""
 }
]
 }
 },
 "altSuggester": {
 "elec": {
 "numFound": 1,
 "suggestions": [
 {
 "term": "electronics and computer1",
 "weight": 10,
 "payload": ""
 }
]
 }
 }
 }
}
```

## Context Filtering

Context filtering lets you filter suggestions by a separate context field, such as category, department or any other token. The `AnalyzingInfixLookupFactory` and `BlendedInfixLookupFactory` currently support this feature, when backed by `DocumentDictionaryFactory`.

Add `contextField` to your suggester configuration. This example will suggest names and allow to filter by category:

## *solrconfig.xml*

```
<searchComponent name="suggest" class="solr.SuggestComponent">
 <lst name="suggester">
 <str name="name">mySuggester</str>
 <str name="lookupImpl">AnalyzingInfixLookupFactory</str>
 <str name="dictionaryImpl">DocumentDictionaryFactory</str>
 <str name="field">name</str>
 <str name="weightField">price</str>
 <str name="contextField">cat</str>
 <str name="suggestAnalyzerFieldType">string</str>
 <str name="buildOnStartup">false</str>
 </lst>
</searchComponent>
```

Example context filtering suggest query:

```
http://localhost:8983/solr/techproducts/suggest?suggest=true&suggest.build=true&suggest.dictionary=mySuggester&suggest.q=c&suggest.cfq=memory
```

The suggester will only bring back suggestions for products tagged with 'cat=memory'.

## MoreLikeThis

The MoreLikeThis search component enables users to query for documents similar to a document in their result list.

It does this by using terms from the original document to find similar documents in the index.

There are three ways to use MoreLikeThis. The first, and most common, is to use it as a request handler. In this case, you would send text to the MoreLikeThis request handler as needed (as in when a user clicked on a "similar documents" link).

The second is to use it as a search component. This is less desirable since it performs the MoreLikeThis analysis on every document returned. This may slow search results.

The final approach is to use it as a request handler but with externally supplied text. This case, also referred to as the MoreLikeThisHandler, will supply information about similar documents in the index based on the text of the input document.

## How MoreLikeThis Works

MoreLikeThis constructs a Lucene query based on terms in a document. It does this by pulling terms from the defined list of fields ( see the `mlt.fl` parameter, below). For best results, the fields should have stored term vectors in `schema.xml`. For example:

```
<field name="cat" ... termVectors="true" />
```

If term vectors are not stored, MoreLikeThis will generate terms from stored fields. A `uniqueKey` must also be stored in order for MoreLikeThis to work properly.

The next phase filters terms from the original document using thresholds defined with the MoreLikeThis parameters. Finally, a query is run with these terms, and any other query parameters that have been defined (see the `mlt.qf` parameter, below) and a new document set is returned.

## Common Parameters for MoreLikeThis

The table below summarizes the MoreLikeThis parameters supported by Lucene/Solr. These parameters can be used with any of the three possible MoreLikeThis approaches.

### `mlt.fl`

Specifies the fields to use for similarity. If possible, these should have stored `termVectors`.

### `mlt.mintf`

Specifies the Minimum Term Frequency, the frequency below which terms will be ignored in the source document.

### `mlt.mindf`

Specifies the Minimum Document Frequency, the frequency at which words will be ignored which do not occur in at least this many documents.

**mlt.maxdf**

Specifies the Maximum Document Frequency, the frequency at which words will be ignored which occur in more than this many documents.

**mlt.maxdfpct**

Specifies the Maximum Document Frequency using a relative ratio to the number of documents in the index. The argument must be an integer between 0 and 100. For example 75 means the word will be ignored if it occurs in more than 75 percent of the documents in the index.

**mlt.minwl**

Sets the minimum word length below which words will be ignored.

**mlt.maxwl**

Sets the maximum word length above which words will be ignored.

**mlt.maxqt**

Sets the maximum number of query terms that will be included in any generated query.

**mlt.maxntp**

Sets the maximum number of tokens to parse in each example document field that is not stored with TermVector support.

**mlt.boost**

Specifies if the query will be boosted by the interesting term relevance. It can be either "true" or "false".

**mlt.qf**

Query fields and their boosts using the same format as that used by the [DisMax Query Parser](#). These fields must also be specified in mlt.fl.

## Parameters for the MoreLikeThisComponent

Using MoreLikeThis as a search component returns similar documents for each document in the response set. In addition to the common parameters, these additional options are available:

**mlt**

If set to true, activates the MoreLikeThis component and enables Solr to return MoreLikeThis results.

**mlt.count**

Specifies the number of similar documents to be returned for each result. The default value is 5.

## Parameters for the MoreLikeThisHandler

The table below summarizes parameters accessible through the MoreLikeThisHandler. It supports facetting, paging, and filtering using common query parameters, but does not work well with alternate query parsers.

**mlt.match.include**

Specifies whether or not the response should include the matched document. If set to false, the response will look like a normal select response.

**mlt.match.offset**

Specifies an offset into the main query search results to locate the document on which the MoreLikeThis

query should operate. By default, the query operates on the first result for the q parameter.

#### mlt.interestingTerms

Controls how the MoreLikeThis component presents the "interesting" terms (the top TF/IDF terms) for the query. Supports three settings. The setting list lists the terms. The setting none lists no terms. The setting details lists the terms along with the boost value used for each term. Unless mlt.boost=true, all terms will have boost=1.0.

## MoreLikeThis Query Parser

The mlt query parser provides a mechanism to retrieve documents similar to a given document, like the handler. More information on the usage of the mlt query parser can be found in the section [Other Parsers](#).

# Pagination of Results

In most search applications, the "top" matching results (sorted by score, or some other criteria) are displayed to some human user.

In many applications the UI for these sorted results are displayed to the user in "pages" containing a fixed number of matching results, and users don't typically look at results past the first few pages worth of results.

## Basic Pagination

In Solr, this basic paginated searching is supported using the `start` and `rows` parameters, and performance of this common behaviour can be tuned by utilizing the `queryResultCache` and adjusting the `queryResultWindowSize` configuration options based on your expected page sizes.

### Basic Pagination Examples

The easiest way to think about simple pagination, is to simply multiply the page number you want (treating the "first" page number as "0") by the number of rows per page; such as in the following pseudo-code:

```
function fetch_solr_page($page_number, $rows_per_page) {
 $start = $page_number * $rows_per_page
 $params = [q = $some_query, rows = $rows_per_page, start = $start]
 return fetch_solr($params)
}
```

## How Basic Pagination is Affected by Index Updates

The `start` parameter specified in a request to Solr indicates an **absolute** "offset" in the complete sorted list of matches that the client wants Solr to use as the beginning of the current "page".

If an index modification (such as adding or removing documents) which affects the sequence of ordered documents matching a query occurs in between two requests from a client for subsequent pages of results, then it is possible that these modifications can result in the same document being returned on multiple pages, or documents being "skipped" as the result set shrinks or grows.

For example, consider an index containing 26 documents like so:

id	name
1	A
2	B
...	
26	Z

Followed by the following requests & index modifications interleaved:

- A client requests `q=:&rows=5&start=0&sort=name asc`
  - documents with the ids 1–5 will be returned to the client
- Document id 3 is deleted
- The client requests "page #2" using `q=:&rows=5&start=5&sort=name asc`
  - Documents 7–11 will be returned
  - Document 6 has been skipped, since it is now the 5th document in the sorted set of all matching results – it would be returned on a new request for "page #1"
- 3 new documents are now added with the ids 90, 91, and 92; All three documents have a name of A
- The client requests "page #3" using `q=:&rows=5&start=10&sort=name asc`
  - Documents 9–13 will be returned
  - Documents 9, 10, and 11 have now been returned on both page #2 and page #3 since they moved farther back in the list of sorted results

In typical situations these impacts from index changes on paginated searching don't significantly affect user experience — either because they happen extremely infrequently in fairly static collections, or because the users recognize that the collection of data is constantly evolving and expect to see documents shift up and down in the result sets.

## Performance Problems with "Deep Paging"

In some situations, the results of a Solr search are not destined for a simple paginated user interface.

When you wish to fetch a very large number of sorted results from Solr to feed into an external system, using very large values for the start or rows parameters can be very inefficient. Pagination using start and rows not only require Solr to compute (and sort) in memory all of the matching documents that should be fetched for the current page, but also all of the documents that would have appeared on previous pages.

While a request for `start=0&rows=1000000` may be obviously inefficient because it requires Solr to maintain & sort in memory a set of 1 million documents, likewise a request for `start=999000&rows=1000` is equally inefficient for the same reasons. Solr can't compute which matching document is the 999001st result in sorted order, without first determining what the first 999000 matching sorted results are.

If the index is distributed, which is common when running in SolrCloud mode, then 1 million documents are retrieved from **each shard**. For a ten shard index, ten million entries must be retrieved and sorted to figure out the 1000 documents that match those query parameters.

## Fetching A Large Number of Sorted Results: Cursors

As an alternative to increasing the "start" parameter to request subsequent pages of sorted results, Solr supports using a "Cursor" to scan through results.

Cursors in Solr are a logical concept that doesn't involve caching any state information on the server. Instead the sort values of the last document returned to the client are used to compute a "mark" representing a logical point in the ordered space of sort values. That "mark" can be specified in the parameters of subsequent requests to tell Solr where to continue.

## Using Cursors

To use a cursor with Solr, specify a cursorMark parameter with the value of \*. You can think of this being analogous to start=0 as a way to tell Solr "start at the beginning of my sorted results" except that it also informs Solr that you want to use a Cursor.

In addition to returning the top N sorted results (where you can control N using the rows parameter) the Solr response will also include an encoded String named nextCursorMark. You then take the nextCursorMark String value from the response, and pass it back to Solr as the cursorMark parameter for your next request. You can repeat this process until you've fetched as many docs as you want, or until the nextCursorMark returned matches the cursorMark you've already specified — indicating that there are no more results.

## Constraints when using Cursors

There are a few important constraints to be aware of when using cursorMark parameter in a Solr request.

1. cursorMark and start are mutually exclusive parameters.
  - Your requests must either not include a start parameter, or it must be specified with a value of "0".
2. sort clauses must include the uniqueKey field (either asc or desc).
  - If id is your uniqueKey field, then sort parameters like id asc and name asc, id desc would both work fine, but name asc by itself would not
3. Sorts including [Date Math](#) based functions that involve calculations relative to NOW will cause confusing results, since every document will get a new sort value on every subsequent request. This can easily result in cursors that never end, and constantly return the same documents over and over – even if the documents are never updated.

In this situation, choose & re-use a fixed value for the NOW [request param](#) in all of your cursor requests.

Cursor mark values are computed based on the sort values of each document in the result, which means multiple documents with identical sort values will produce identical Cursor mark values if one of them is the last document on a page of results. In that situation, the subsequent request using that cursorMark would not know which of the documents with the identical mark values should be skipped. Requiring that the uniqueKey field be used as a clause in the sort criteria guarantees that a deterministic ordering will be returned, and that every cursorMark value will identify a unique point in the sequence of documents.

## Cursor Examples

### Fetch All Docs

The pseudo-code shown here shows the basic logic involved in fetching all documents matching a query using a cursor:

```
// when fetching all docs, you might as well use a simple id sort
// unless you really need the docs to come back in a specific order
$params = [q => $some_query, sort => 'id asc', rows => $r, cursorMark => '*']
$done = false
while (not $done) {
 $results = fetch_solr($params)
 // do something with $results
 if ($params[cursorMark] == $results[nextCursorMark]) {
 $done = true
 }
 $params[cursorMark] = $results[nextCursorMark]
}
```

Using SolrJ, this pseudo-code would be:

```
SolrQuery q = (new SolrQuery(some_query)).setRows(r).setSort(SortClause.asc("id"));
String cursorMark = CursorMarkParams.CURSOR_MARK_START;
boolean done = false;
while (! done) {
 q.set(CursorMarkParams.CURSOR_MARK_PARAM, cursorMark);
 QueryResponse rsp = solrServer.query(q);
 String nextCursorMark = rsp.getNextCursorMark();
 doCustomProcessingOfResults(rsp);
 if (cursorMark.equals(nextCursorMark)) {
 done = true;
 }
 cursorMark = nextCursorMark;
}
```

If you wanted to do this by hand using curl, the sequence of requests would look something like this:

```
$ curl '...&rows=10&sort=id+asc&cursorMark=*'
{
 "response": {"numFound": 32, "start": 0, "docs": [
 // ... 10 docs here ...
]},
 "nextCursorMark": "AoEjR0JQ"
$ curl '...&rows=10&sort=id+asc&cursorMark=AoEjR0JQ'
{
 "response": {"numFound": 32, "start": 0, "docs": [
 // ... 10 more docs here ...
]},
 "nextCursorMark": "AoEpVkJCREIxQTE2"
$ curl '...&rows=10&sort=id+asc&cursorMark=AoEpVkJCREIxQTE2'
{
 "response": {"numFound": 32, "start": 0, "docs": [
 // ... 10 more docs here ...
]},
 "nextCursorMark": "AoEmbWF4dG9y"
$ curl '...&rows=10&sort=id+asc&cursorMark=AoEmbWF4dG9y'
{
 "response": {"numFound": 32, "start": 0, "docs": [
 // ... 2 docs here because we've reached the end.
]},
 "nextCursorMark": "AoEpdm1ld3Nvbmlj"
$ curl '...&rows=10&sort=id+asc&cursorMark=AoEpdm1ld3Nvbmlj'
{
 "response": {"numFound": 32, "start": 0, "docs": [
 // no more docs here, and note that the nextCursorMark
 // matches the cursorMark param we used
]},
 "nextCursorMark": "AoEpdm1ld3Nvbmlj"}
```

## Fetch First N docs, Based on Post Processing

Since the cursor is stateless from Solr's perspective, your client code can stop fetching additional results as soon as you have decided you have enough information:

```
while (! done) {
 q.set(CursorMarkParams.CURSOR_MARK_PARAM, cursorMark);
 QueryResponse rsp = solrServer.query(q);
 String nextCursorMark = rsp.getNextCursorMark();
 boolean hadEnough = doCustomProcessingOfResults(rsp);
 if (hadEnough || cursorMark.equals(nextCursorMark)) {
 done = true;
 }
 cursorMark = nextCursorMark;
}
```

## How Cursors are Affected by Index Updates

Unlike basic pagination, Cursor pagination does not rely on using an absolute "offset" into the completed sorted list of matching documents. Instead, the cursorMark specified in a request encapsulates information about the **relative** position of the last document returned, based on the **absolute** sort values of that document. This means that the impact of index modifications is much smaller when using a cursor compared to basic pagination. Consider the same example index described when discussing basic pagination:

id	name
1	A
2	B
...	
26	Z

- A client requests `q=:&rows=5&start=0&sort=name asc, id asc&cursorMark=*`
  - Documents with the ids 1–5 will be returned to the client in order
- Document id 3 is deleted
- The client requests 5 more documents using the nextCursorMark from the previous response
  - Documents 6–10 will be returned — the deletion of a document that's already been returned doesn't affect the relative position of the cursor
- 3 new documents are now added with the ids 90, 91, and 92; All three documents have a name of A
- The client requests 5 more documents using the nextCursorMark from the previous response
  - Documents 11–15 will be returned — the addition of new documents with sort values already past does not affect the relative position of the cursor
- Document id 1 is updated to change its 'name' to Q
- Document id 17 is updated to change its 'name' to A
- The client requests 5 more documents using the nextCursorMark from the previous response
  - The resulting documents are 16, 1, 18, 19, 20 in that order
  - Because the sort value of document 1 changed so that it is *after* the cursor position, the document is returned to the client twice
  - Because the sort value of document 17 changed so that it is *before* the cursor position, the document has been "skipped" and will not be returned to the client as the cursor continues to progress

In a nutshell: When fetching all results matching a query using cursorMark, the only way index modifications can result in a document being skipped, or returned twice, is if the sort value of the document changes.

One way to ensure that a document will never be returned more than once, is to use the uniqueKey field as the primary (and therefore: only significant) sort criterion.

In this situation, you will be guaranteed that each document is only returned once, no matter how it may be modified during the use of the cursor.



## "Tailing" a Cursor

Because Cursor requests are stateless, and the cursorMark values encapsulate the absolute sort values of the last document returned from a search, it's possible to "continue" fetching additional results from a cursor that has already reached its end. If new documents are added (or existing documents are updated) to the end of the results.

You can think of this as similar to using something like "tail -f" in Unix. The most common examples of how this can be useful is when you have a "timestamp" field recording when a document has been added/updated in your index. Client applications can continuously poll a cursor using a `sort=timestamp asc, id asc` for documents matching a query, and always be notified when a document is added or updated matching the request criteria.

Another common example is when you have uniqueKey values that always increase as new documents are created, and you can continuously poll a cursor using `sort=id asc` to be notified about new documents.

The pseudo-code for tailing a cursor is only a slight modification from our early example for processing all docs matching a query:

```
while (true) {
 $doneForNow = false
 while (not $doneForNow) {
 $results = fetch_solr($params)
 // do something with $results
 if ($params[cursorMark] == $results[nextCursorMark]) {
 $doneForNow = true
 }
 $params[cursorMark] = $results[nextCursorMark]
 }
 sleep($some_configured_delay)
}
```



For certain specialized cases, the [/export handler](#) may be an option.

# Collapse and Expand Results

The Collapsing query parser and the Expand component combine to form an approach to grouping documents for field collapsing in search results.

The Collapsing query parser groups documents (collapsing the result set) according to your parameters, while the Expand component provides access to documents in the collapsed group for use in results display or other processing by a client application. Collapse & Expand can together do what the older [Result Grouping](#) (group=true) does for *most* use-cases but not all. Generally, you should prefer Collapse & Expand.



In order to use these features with SolrCloud, the documents must be located on the same shard. To ensure document co-location, you can define the router.name parameter as compositeId when creating the collection. For more information on this option, see the section [Document Routing](#).

## Collapsing Query Parser

The CollapsingQParser is really a *post filter* that provides more performant field collapsing than Solr's standard approach when the number of distinct groups in the result set is high. This parser collapses the result set to a single document per group before it forwards the result set to the rest of the search components. So all downstream components (faceting, highlighting, etc.) will work with the collapsed result set.

The CollapsingQParser accepts the following local parameters:

### field

The field that is being collapsed on. The field must be a single valued String, Int or Float-type of field.

### min or max

Selects the group head document for each group based on which document has the min or max value of the specified numeric field or [function query](#).

At most only one of the min, max, or sort (see below) parameters may be specified.

If none are specified, the group head document of each group will be selected based on the highest scoring document in that group. The default is none.

### sort

Selects the group head document for each group based on which document comes first according to the specified [sort string](#).

At most only one of the min, max, (see above) or sort parameters may be specified.

If none are specified, the group head document of each group will be selected based on the highest scoring document in that group. The default is none.

### nullPolicy

There are three available null policies:

- ignore: removes documents with a null value in the collapse field. This is the default.

- expand: treats each document with a null value in the collapse field as a separate group.
- collapse: collapses all documents with a null value into a single group using either highest score, or minimum/maximum.

The default is ignore.

#### hint

Currently there is only one hint available: top\_fc, which stands for top level FieldCache.

The top\_fc hint is only available when collapsing on String fields. top\_fc usually provides the best query time speed but takes the longest to warm on startup or following a commit. top\_fc will also result in having the collapsed field cached in memory twice if it's used for faceting or sorting. For very high cardinality (high distinct count) fields, top\_fc may not fare so well.

The default is none.

#### size

Sets the initial size of the collapse data structures when collapsing on a **numeric field only**.

The data structures used for collapsing grow dynamically when collapsing on numeric fields. Setting the size above the number of results expected in the result set will eliminate the resizing cost.

The default is 100,000.

## Sample Usage Syntax

Collapse on group\_field selecting the document in each group with the highest scoring document:

```
fq={!collapse field=group_field}
```

Collapse on group\_field selecting the document in each group with the minimum value of numeric\_field:

```
fq={!collapse field=group_field min=numeric_field}
```

Collapse on group\_field selecting the document in each group with the maximum value of numeric\_field:

```
fq={!collapse field=group_field max=numeric_field}
```

Collapse on group\_field selecting the document in each group with the maximum value of a function. Note that the **cscore()** function can be used with the min/max options to use the score of the current document being collapsed.

```
fq={!collapse field=group_field max=sum(cscore(),numeric_field)}
```

Collapse on group\_field with a null policy so that all docs that do not have a value in the group\_field will be treated as a single group. For each group, the selected document will be based first on a numeric\_field, but ties will be broken by score:

```
fq={!collapse field=group_field nullPolicy=collapse sort='numeric_field asc, score desc'}
```

Collapse on group\_field with a hint to use the top level field cache:

```
fq={!collapse field=group_field hint=top_fc}
```

The CollapsingQParserPlugin fully supports the QueryElevationComponent.

## Expand Component

The ExpandComponent can be used to expand the groups that were collapsed by the CollapsingQParserPlugin.

Example usage with the CollapsingQParserPlugin:

```
q=foo&fq={!collapse field=ISBN}
```

In the query above, the CollapsingQParserPlugin will collapse the search results on the *ISBN* field. The main search results will contain the highest ranking document from each book.

The ExpandComponent can now be used to expand the results so you can see the documents grouped by ISBN. For example:

```
q=foo&fq={!collapse field=ISBN}&expand=true
```

The “expand=true” parameter turns on the ExpandComponent. The ExpandComponent adds a new section to the search output labeled “expanded”.

Inside the expanded section there is a *map* with each group head pointing to the expanded documents that are within the group. As applications iterate the main collapsed result set, they can access the *expanded* map to retrieve the expanded groups.

The ExpandComponent has the following parameters:

**expand.sort**

Orders the documents within the expanded groups. The default is `score desc`.

**expand.rows**

The number of rows to display in each group. The default is 5 rows.

**expand.q**

Overrides the main query (`q`), determines which documents to include in the main group. The default is to use the main query.

**expand.fq**

Overrides main filter queries (`fq`), determines which documents to include in the main group. The default is to use the main filter queries.

# Result Grouping

Result Grouping groups documents with a common field value into groups and returns the top documents for each group.

For example, if you searched for "DVD" on an electronic retailer's e-commerce site, you might be returned three categories such as "TV and Video", "Movies", and "Computers" with three results per category. In this case, the query term "DVD" appeared in all three categories, so Solr groups them together in order to increase relevancy for the user.

## *Prefer Collapse & Expand instead*



Solr's [Collapse and Expand](#) feature is newer and mostly overlaps with Result Grouping. There are features unique to both, and they have different performance characteristics. That said, in most cases Collapse and Expand is preferable to Result Grouping.

Result Grouping is separate from [Faceting](#). Though it is conceptually similar, facetting returns all relevant results and allows the user to refine the results based on the facet category. For example, if you search for "shoes" on a footwear retailer's e-commerce site, Solr would return all results for that query term, along with selectable facets such as "size," "color," "brand," and so on.

You can however combine grouping with facetting. Grouped facetting supports `facet.field` and `facet.range` but currently doesn't support date and pivot facetting. The facet counts are computed based on the first `group.field` parameter, and other `group.field` parameters are ignored.

Grouped facetting differs from non grouped facets ( $\text{sum of all facets} == (\text{total of products with that property})$ ) as shown in the following example:

### Object 1

- name: Phaser 4620a
- ppm: 62
- product\_range: 6

### Object 2

- name: Phaser 4620i
- ppm: 65
- product\_range: 6

### Object 3

- name: ML6512
- ppm: 62
- product\_range: 7

If you ask Solr to group these documents by "product\_range", then the total amount of groups is 2, but the facets for ppm are 2 for 62 and 1 for 65.

## Grouping Parameters

Result Grouping takes the following request parameters. Any number of these request parameters can be included in a single request:

group

If true, query results will be grouped.

group.field

The name of the field by which to group results. The field must be single-valued, and either be indexed or a field type that has a value source and works in a function query, such as ExternalFileField. It must also be a string-based field, such as StrField or TextField

group.func

Group based on the unique values of a function query.



This option does not work with [distributed searches](#).

group.query

Return a single group of documents that match the given query.

rows

The number of groups to return. The default value is 10.

start

Specifies an initial offset for the list of groups.

group.limit

Specifies the number of results to return for each group. The default value is 1.

group.offset

Specifies an initial offset for the document list of each group.

sort

Specifies how Solr sorts the groups relative to each other. For example, sort=popularity desc will cause the groups to be sorted according to the highest popularity document in each group. The default value is score desc.

group.sort

Specifies how Solr sorts documents within each group. The default behavior if group.sort is not specified is to use the same effective value as the sort parameter.

group.format

If this parameter is set to simple, the grouped documents are presented in a single flat list, and the start and rows parameters affect the numbers of documents instead of groups. An alternate value for this parameter is grouped.

group.main

If true, the result of the first field grouping command is used as the main result list in the response, using group.format=simple.

## group.ngroups

If true, Solr includes the number of groups that have matched the query in the results. The default value is false.

See below for [Distributed Result Grouping Caveats](#) when using sharded indexes.

## group.truncate

If true, facet counts are based on the most relevant document of each group matching the query. The default value is false.

## group.facet

Determines whether to compute grouped facets for the field facets specified in facet.field parameters. Grouped facets are computed based on the first specified group. As with normal field faceting, fields shouldn't be tokenized (otherwise counts are computed for each token). Grouped facetting supports single and multivalued fields. Default is false.



There can be a heavy performance cost to this option.

See below for [Distributed Result Grouping Caveats](#) when using sharded indexes.

## group.cache.percent

Setting this parameter to a number greater than 0 enables caching for result grouping. Result Grouping executes two searches; this option caches the second search. The default value is 0. The maximum value is 100.

Testing has shown that group caching only improves search time with Boolean, wildcard, and fuzzy queries. For simple queries like term or "match all" queries, group caching degrades performance.

Any number of group commands (e.g., group.field, group.func, group.query, etc.) may be specified in a single request.

## Grouping Examples

All of the following sample queries work with Solr's "bin/solr -e techproducts" example.

### Grouping Results by Field

In this example, we will group results based on the manu\_exact field, which specifies the manufacturer of the items in the sample dataset.

```
http://localhost:8983/solr/techproducts/select?fl=id,name&q=solr+memory&group=true&group.field=manu_exact
```

```
{
"..."
"grouped":{
 "manu_exact":{
 "matches":6,
 "groups":[{
 "groupValue":"Apache Software Foundation",
 "doclist":{ "numFound":1,"start":0,"docs": [
 {
 "id":"SOLR1000",
 "name":"Solr, the Enterprise Search Server"}]
 }},
 {
 "groupValue":"Corsair Microsystems Inc.",
 "doclist":{ "numFound":2,"start":0,"docs": [
 {
 "id":"VS1GB400C3",
 "name":"CORSAIR ValueSelect 1GB 184-Pin DDR SDRAM Unbuffered DDR 400 (PC 3200)
System Memory - Retail"}]
 }},
 {
 "groupValue":"A-DATA Technology Inc.",
 "doclist":{ "numFound":1,"start":0,"docs": [
 {
 "id":"VDBDB1A16",
 "name":"A-DATA V-Series 1GB 184-Pin DDR SDRAM Unbuffered DDR 400 (PC 3200) System
Memory - OEM"}]
 }},
 {
 "groupValue":"Canon Inc.",
 "doclist":{ "numFound":1,"start":0,"docs": [
 {
 "id":"0579B002",
 "name":"Canon PIXMA MP500 All-In-One Photo Printer"}]
 }},
 {
 "groupValue":"ASUS Computer Inc.",
 "doclist":{ "numFound":1,"start":0,"docs": [
 {
 "id":"EN7800GTX/2DHTV/256M",
 "name":"ASUS Extreme N7800GTX/2DHTV (256 MB)"}]
 }
 }]}}}
```

The response indicates that there are six total matches for our query. For each of the five unique values of group.field, Solr returns a docList for that groupValue such that the numFound indicates the total number of documents in that group, and the top documents are returned according to the implicit default group.limit=1 and group.sort=score desc parameters. The resulting groups are then sorted by the score of the top document within each group based on the implicit sort=score desc, and the number of groups returned is limited to the implicit rows=10.

We can run the same query with the request parameter group.main=true. This will format the results as a single flat document list. This flat format does not include as much information as the normal result grouping query results – notably the numFound in each group – but it may be easier for existing Solr clients to parse.

```
http://localhost:8983/solr/techproducts/select?fl=id,name,manufacturer&q=solr+memory&group=true&group.field=manu_exact&group.main=true
```

```
{
 "responseHeader":{
 "status":0,
 "QTime":1,
 "params":{
 "fl":"id,name,manufacturer",
 "indent":"true",
 "q":"solr memory",
 "group.field":"manu_exact",
 "group.main":"true",
 "group":"true"}},
 "grouped":{},
 "response":{ "numFound":6, "start":0, "docs": [
 {
 "id":"SOLR1000",
 "name":"Solr, the Enterprise Search Server"},
 {
 "id":"VS1GB400C3",
 "name":"CORSAIR ValueSelect 1GB 184-Pin DDR SDRAM Unbuffered DDR 400 (PC 3200) System Memory - Retail"},
 {
 "id":"VDBDB1A16",
 "name":"A-DATA V-Series 1GB 184-Pin DDR SDRAM Unbuffered DDR 400 (PC 3200) System Memory - OEM"},
 {
 "id":"0579B002",
 "name":"Canon PIXMA MP500 All-In-One Photo Printer"},
 {
 "id":"EN7800GTX/2DHTV/256M",
 "name":"ASUS Extreme N7800GTX/2DHTV (256 MB)"}]
 }
}
```

## Grouping by Query

In this example, we will use the group.query parameter to find the top three results for "memory" in two different price ranges: 0.00 to 99.99, and over 100.

```
http://localhost:8983/solr/techproducts/select?indent=true&fl=name,price&q=memory&group=true&group.query=price:[0+TO+99.99]&group.query=price:[100+TO+*]&group.limit=3
```

```
{
 "responseHeader":{
 "status":0,
 "QTime":42,
 "params":{
 "f1":"name,price",
 "indent":"true",
 "q":"memory",
 "group.limit":"3",
 "group.query":["price:[0 TO 99.99]",
 "price:[100 TO *]"],
 "group":"true"}},
 "grouped":{
 "price:[0 TO 99.99]":{
 "matches":5,
 "doclist":{ "numFound":1, "start":0, "docs": [
 {
 "name": "CORSAIR ValueSelect 1GB 184-Pin DDR SDRAM Unbuffered DDR 400 (PC 3200) System
Memory - Retail",
 "price":74.99}]
 }},
 "price:[100 TO *]":{
 "matches":5,
 "doclist":{ "numFound":3, "start":0, "docs": [
 {
 "name": "CORSAIR XMS 2GB (2 x 1GB) 184-Pin DDR SDRAM Unbuffered DDR 400 (PC 3200)
Dual Channel Kit System Memory - Retail",
 "price":185.0},
 {
 "name": "Canon PIXMA MP500 All-In-One Photo Printer",
 "price":179.99},
 {
 "name": "ASUS Extreme N7800GTX/2DHTV (256 MB)",
 "price":479.95}]
 }},
 }},
 }}
```

In this case, Solr found five matches for "memory," but only returns four results grouped by price. This is because one result for "memory" did not have a price assigned to it.

## Distributed Result Grouping Caveats

Grouping is supported for [distributed searches](#), with some caveats:

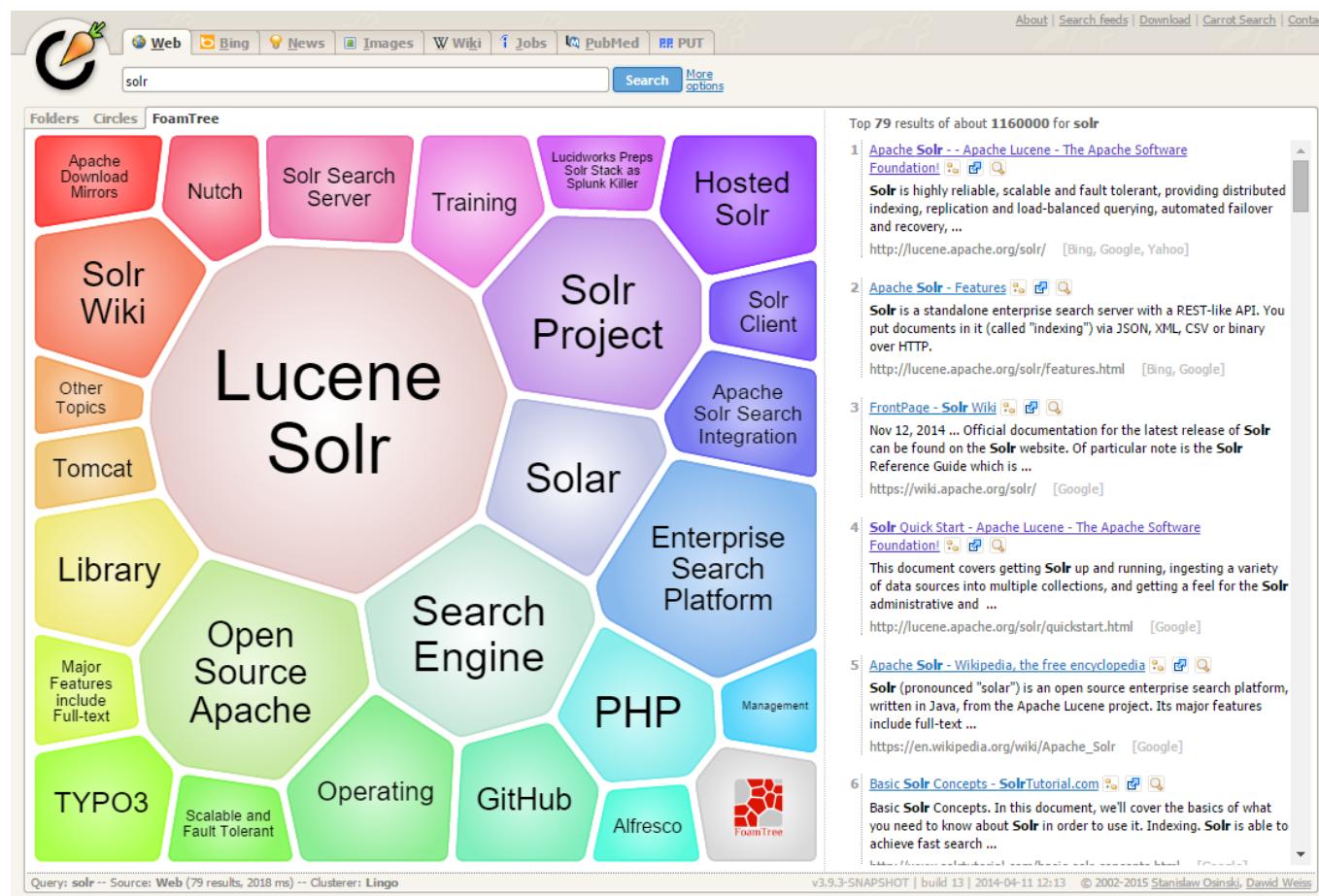
- Currently group.func is not supported in any distributed searches
- group.ngroups and group.facet require that all documents in each group must be co-located on the same shard in order for accurate counts to be returned. [Document routing via composite keys](#) can be a useful solution in many situations.

# Result Clustering

The **clustering** (or **cluster analysis**) plugin attempts to automatically discover groups of related search hits (documents) and assign human-readable labels to these groups.

By default in Solr, the clustering algorithm is applied to the search result of each single query — this is called an *on-line* clustering. While Solr contains an extension for full-index clustering (*off-line* clustering) this section will focus on discussing on-line clustering only.

Clusters discovered for a given query can be perceived as *dynamic facets*. This is beneficial when regular faceting is difficult (field values are not known in advance) or when the queries are exploratory in nature. Take a look at the [Carrot2](#) project's demo page to see an example of search results clustering in action (the groups in the visualization have been discovered automatically in search results to the right, there is no external information involved).



The query issued to the system was *Solr*. It seems clear that faceting could not yield a similar set of groups, although the goals of both techniques are similar—to let the user explore the set of search results and either rephrase the query or narrow the focus to a subset of current documents. Clustering is also similar to [Result Grouping](#) in that it can help to look deeper into search results, beyond the top few hits.

## Clustering Concepts

Each **document** passed to the clustering component is composed of several logical parts:

- a unique identifier,
- origin URL,
- the title,
- the main content,
- a language code of the title and content.

The identifier part is mandatory, everything else is optional but at least one of the text fields (title or content) will be required to make the clustering process reasonable. It is important to remember that logical document parts must be mapped to a particular schema and its fields. The content (text) for clustering can be sourced from either a stored text field or context-filtered using a highlighter, all these options are explained below in the [configuration](#) section.

A **clustering algorithm** is the actual logic (implementation) that discovers relationships among the documents in the search result and forms human-readable cluster labels. Depending on the choice of the algorithm the clusters may (and probably will) vary. Solr comes with several algorithms implemented in the open source [Carrot2](#) project, commercial alternatives also exist.

## Clustering Quick Start Example

The “techproducts” example included with Solr is pre-configured with all the necessary components for result clustering — but they are disabled by default.

To enable the clustering component contrib and a dedicated search handler configured to use it, specify a JVM System Property when running the example:

```
bin/solr start -e techproducts -Dsolr.clustering.enabled=true
```

You can now try out the clustering handler by opening the following URL in a browser:

```
http://localhost:8983/solr/techproducts/clustering?q=*&rows=100&wt=xml
```

The output XML should include search hits and an array of automatically discovered clusters at the end, resembling the output shown here:

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">299</int>
 </lst>
 <result name="response" numFound="32" start="0" maxScore="1.0">
 <doc>
 <str name="id">GB18030TEST</str>
 <str name="name">Test with some GB18030 encoded characters</str>
 <arr name="features">
 <str>No accents here</str>
 <str> </str>
 <str>This is a feature (translated)</str>
 <str> </str>
```

```
<str>This document is very shiny (translated)</str>
</arr>
<float name="price">0.0</float>
<str name="price_c">0,USD</str>
<bool name="inStock">true</bool>
<long name="_version_">1448955395025403904</long>
<float name="score">1.0</float>
</doc>

<!-- more search hits, omitted -->
</result>

<arr name="clusters">
<lst>
<arr name="labels">
<str>DDR</str>
</arr>
<double name="score">3.9599865057283354</double>
<arr name="docs">
<str>TWINX2048-3200PRO</str>
<str>VS1GB400C3</str>
<str>VDBDB1A16</str>
</arr>
</lst>
<lst>
<arr name="labels">
<str>iPod</str>
</arr>
<double name="score">11.959228467119022</double>
<arr name="docs">
<str>F8V7067-APL-KIT</str>
<str>IW-02</str>
<str>MA147LL/A</str>
</arr>
</lst>

<!-- More clusters here, omitted. -->

<lst>
<arr name="labels">
<str>Other Topics</str>
</arr>
<double name="score">0.0</double>
<bool name="other-topics">true</bool>
<arr name="docs">
<str>adata</str>
<str>apple</str>
<str>asus</str>
<str>ati</str>
<!-- other unassigned document IDs here -->
</arr>
</lst>
```

```
</arr>
</response>
```

There were a few clusters discovered for this query (\*:\*), separating search hits into various categories: DDR, iPod, Hard Drive, etc. Each cluster has a label and score that indicates the "goodness" of the cluster. The score is algorithm-specific and is meaningful only in relation to the scores of other clusters in the same set. In other words, if cluster A has a higher score than cluster B, cluster A should be of better quality (have a better label and/or more coherent document set). Each cluster has an array of identifiers of documents belonging to it. These identifiers correspond to the uniqueKey field declared in the schema.

Depending on the quality of input documents, some clusters may not make much sense. Some documents may be left out and not be clustered at all; these will be assigned to the synthetic *Other Topics* group, marked with the other-topics property set to true (see the XML dump above for an example). The score of the other topics group is zero.

## Installing the Clustering Contrib

The clustering contrib extension requires dist/solr-clustering-\*.jar and all JARs under contrib/clustering/lib.

## Clustering Configuration

### Declaration of the Clustering Search Component and Request Handler

Clustering extension is a search component and must be declared in solrconfig.xml. Such a component can be then appended to a request handler as the last component in the chain (because it requires search results which must be previously fetched by the search component).

An example configuration could look as shown below.

1. Include the required contrib JARs. Note that by default paths are relative to the Solr core so they may need adjustments to your configuration, or an explicit specification of the \$solr.install.dir.

```
<lib dir="${solr.install.dir:../../..}/contrib/clustering/lib/" regex=".*\.\.jar" />
<lib dir="${solr.install.dir:../../..}/dist/" regex="solr-clustering-\d.*\.\.jar" />
```

2. Declaration of the search component. Each component can also declare multiple clustering pipelines ("engines"), which can be selected at runtime by passing clustering.engine=(engine name) URL parameter.

```
<searchComponent name="clustering" class="solr.clustering.ClusteringComponent">
 <!-- Lingo clustering algorithm -->
 <lst name="engine">
 <str name="name">lingo</str>
 <str name="carrot.algorithm">org.carrot2.clustering.lingo.LingoClusteringAlgorithm</str>
 </lst>

 <!-- An example definition for the STC clustering algorithm. -->
 <lst name="engine">
 <str name="name">stc</str>
 <str name="carrot.algorithm">org.carrot2.clustering.stc.STCCLusteringAlgorithm</str>
 </lst>
</searchComponent>
```

3. A request handler to which we append the clustering component declared above.

```
<requestHandler name="/clustering"
 class="solr.SearchHandler">
 <lst name="defaults">
 <bool name="clustering">true</bool>
 <bool name="clustering.results">true</bool>

 <!-- Logical field to physical field mapping. -->
 <str name="carrot.url">id</str>
 <str name="carrot.title">doctitle</str>
 <str name="carrot.snippet">content</str>

 <!-- Configure any other request handler parameters. We will cluster the
 top 100 search results so bump up the 'rows' parameter. -->
 <str name="rows">100</str>
 <str name="f1">*,score</str>
 </lst>

 <!-- Append clustering at the end of the list of search components. -->
 <arr name="last-components">
 <str>clustering</str>
 </arr>
</requestHandler>
```

## Configuration Parameters of the Clustering Component

The following parameters of each clustering engine or the entire clustering component (depending where they are declared) are available.

`clustering`

When true, clustering component is enabled.

`clustering.engine`

Declares which clustering engine to use. If not present, the first declared engine will become the default

one.

#### clustering.results

When true, the component will perform clustering of search results (this should be enabled).

#### clustering.collection

When true, the component will perform clustering of the whole document index (this section does not cover full-index clustering).

At the engine declaration level, the following parameters are supported.

#### carrot.algorithm

The algorithm class.

#### carrot.resourcesDir

Algorithm-specific resources and configuration files (stop words, other lexical resources, default settings).

By default points to conf/clustering/carrot2/

#### carrot.outputSubClusters

If true and the algorithm supports hierarchical clustering, sub-clusters will also be emitted. Default value: true.

#### carrot.numDescriptions

Maximum number of per-cluster labels to return (if the algorithm assigns more than one label to a cluster).

The carrot.algorithm parameter should contain a fully qualified class name of an algorithm supported by the [Carrot2](#) framework. Currently, the following algorithms are available:

- org.carrot2.clustering.lingo.LingoClusteringAlgorithm (open source)
- org.carrot2.clustering.stc.STCCLusteringAlgorithm (open source)
- org.carrot2.clustering.kmeans.BisectingKMeansClusteringAlgorithm (open source)
- com.carrotsearch.lingo3g.Lingo3GClusteringAlgorithm (commercial)

For a comparison of characteristics of these algorithms see the following links:

- <http://doc.carrot2.org/#section.advanced-topics.fine-tuning.choosing-algorithm>
- <http://project.carrot2.org/algorithms.html>
- <http://carrotsearch.com/lingo3g-comparison.html>

The question of which algorithm to choose depends on the amount of traffic (STC is faster than Lingo, but arguably produces less intuitive clusters, Lingo3G is the fastest algorithm but is not free or open source), expected result (Lingo3G provides hierarchical clusters, Lingo and STC provide flat clusters), and the input data (each algorithm will cluster the input slightly differently). There is no one answer which algorithm is "the best".

## Contextual and Full Field Clustering

The clustering engine can apply clustering to the full content of (stored) fields or it can run an internal highlighter pass to extract context-snippets before clustering. Highlighting is recommended when the

logical snippet field contains a lot of content (this would affect clustering performance). Highlighting can also increase the quality of clustering because the content passed to the algorithm will be more focused around the query (it will be query-specific context). The following parameters control the internal highlighter.

#### carrot.produceSummary

When true the clustering component will run a highlighter pass on the content of logical fields pointed to by `carrot.title` and `carrot.snippet`. Otherwise full content of those fields will be clustered.

#### carrot.fragSize

The size, in characters, of the snippets (aka fragments) created by the highlighter. If not specified, the default highlighting fragsize (`hl.fragsize`) will be used.

#### carrot.summarySnippets

The number of summary snippets to generate for clustering. If not specified, the default highlighting snippet count (`hl.snippets`) will be used.

## Logical to Document Field Mapping

As already mentioned in [Clustering Concepts](#), the clustering component clusters "documents" consisting of logical parts that need to be mapped onto physical schema of data stored in Solr. The field mapping attributes provide a connection between fields and logical document parts. Note that the content of title and snippet fields must be **stored** so that it can be retrieved at search time.

#### carrot.title

The field (alternatively comma- or space-separated list of fields) that should be mapped to the logical document's title. The clustering algorithms typically give more weight to the content of the title field compared to the content (snippet). For best results, the field should contain concise, noise-free content. If there is no clear title in your data, you can leave this parameter blank.

#### carrot.snippet

The field (alternatively comma- or space-separated list of fields) that should be mapped to the logical document's main content. If this mapping points to very large content fields the performance of clustering may drop significantly. An alternative then is to use query-context snippets for clustering instead of full field content. See the description of the `carrot.produceSummary` parameter for details.

#### carrot.url

The field that should be mapped to the logical document's content URL. Leave blank if not required.

## Clustering Multilingual Content

The field mapping specification can include a `carrot.lang` parameter, which defines the field that stores [ISO 639-1](#) code of the language in which the title and content of the document are written. This information can be stored in the index based on apriori knowledge of the documents' source or a language detection filter applied at indexing time. All algorithms inside the Carrot2 framework will accept ISO codes of languages defined in [LanguageCode enum](#).

The language hint makes it easier for clustering algorithms to separate documents from different languages on input and to pick the right language resources for clustering. If you do have multi-lingual query results (or query results in a language different than English), it is strongly advised to map the language field

appropriately.

`carrot.lang`

The field that stores ISO 639-1 code of the language of the document's text fields.

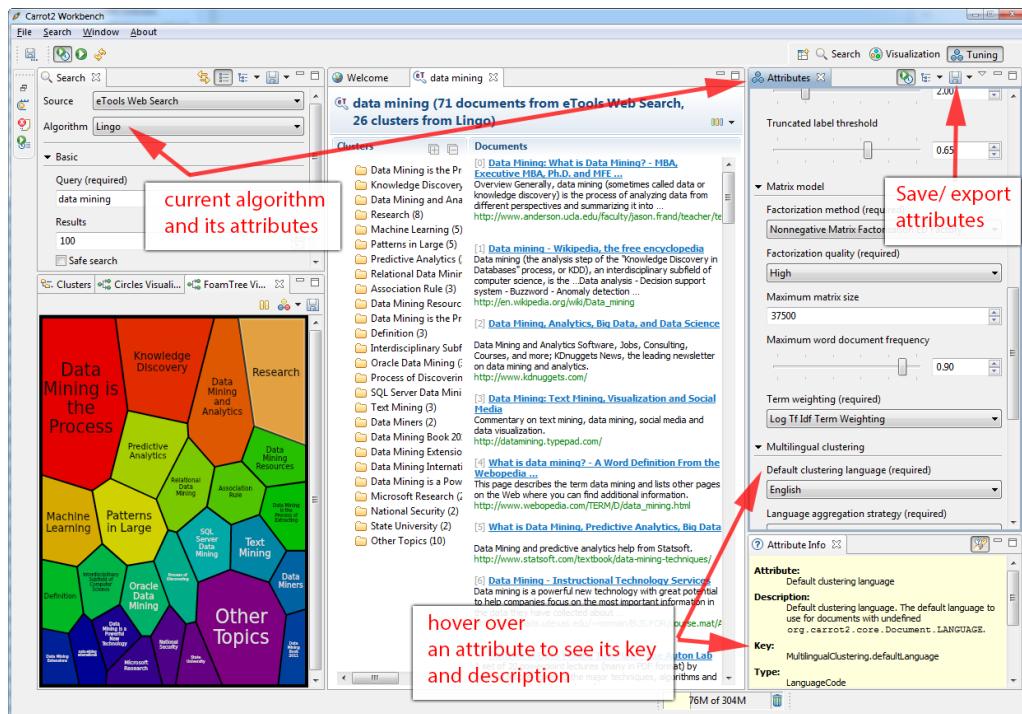
`carrot.lcmap`

A mapping of arbitrary strings into ISO 639 two-letter codes used by `carrot.lang`. The syntax of this parameter is the same as `langid.map.lcmap`, for example: `langid.map.lcmap=japanese:ja polish:pl english:en`

The default language can also be set using Carrot2-specific algorithm attributes (in this case the `MultilingualClustering.defaultLanguage` attribute).

## Tweaking Algorithm Settings

The algorithms that come with Solr are using their default settings which may be inadequate for all data sets. All algorithms have lexical resources and resources (stop words, stemmers, parameters) that may require tweaking to get better clusters (and cluster labels). For Carrot2-based algorithms it is probably best to refer to a dedicated tuning application called Carrot2 Workbench (screenshot below). From this application one can export a set of algorithm attributes as an XML file, which can be then placed under the location pointed to by `carrot.resourcesDir`.



## Providing Defaults for Clustering

The default attributes for all engines (algorithms) declared in the clustering component are placed under `carrot.resourcesDir` and with an expected file name of `engineName-attributes.xml`. So for an engine named `lingo` and the default value of `carrot.resourcesDir`, the attributes would be read from a file in `conf/clustering/carrot2/lingo-attributes.xml`.

An example XML file changing the default language of documents to Polish is shown below.

```
<attribute-sets default="attributes">
 <attribute-set id="attributes">
 <value-set>
 <label>attributes</label>
 <attribute key="MultilingualClustering.defaultLanguage">
 <value type="org.carrot2.core.LanguageCode" value="POLISH"/>
 </attribute>
 </value-set>
 </attribute-set>
</attribute-sets>
```

## Tweaking Algorithms at Query-Time

The clustering component and Carrot2 clustering algorithms can accept query-time attribute overrides. Note that certain things (for example lexical resources) can only be initialized once (at startup, via the XML configuration files).

An example query that changes the LingoClusteringAlgorithm.desiredClusterCountBase parameter for the Lingo algorithm:

```
http://localhost:8983/solr/techproducts/clustering?q=*&rows=100&LingoClusteringAlgorithm.desire
dClusterCountBase=20
```

The clustering engine (the algorithm declared in `solrconfig.xml`) can also be changed at runtime by passing `clustering.engine=name` request attribute:

```
http://localhost:8983/solr/techproducts/clustering?q=*&rows=100&clustering.engine=kmeans
```

## Performance Considerations with Dynamic Clustering

Dynamic clustering of search results comes with two major performance penalties:

- Increased cost of fetching a larger-than-usual number of search results (50, 100 or more documents),
- Additional computational cost of the clustering itself.

For simple queries, the clustering time will usually dominate the fetch time. If the document content is very long the retrieval of stored content can become a bottleneck. The performance impact of clustering can be lowered in several ways:

- feed less content to the clustering algorithm by enabling `carrot.produceSummary` attribute,
- perform clustering on selected fields (titles only) to make the input smaller,
- use a faster algorithm (STC instead of Lingo, Lingo3G instead of STC),
- tune the performance attributes related directly to a specific algorithm.

Some of these techniques are described in *Apache SOLR and Carrot2 integration strategies* document, available at <http://carrot2.github.io/solr-integration-strategies>. The topic of improving performance is also

included in the Carrot2 manual at <http://doc.carrot2.org/#section.advanced-topics.fine-tuning.performance>.

## Additional Resources

The following resources provide additional information about the clustering component in Solr and its potential applications.

- Apache Solr and Carrot2 integration strategies: <http://carrot2.github.io/solr-integration-strategies>
- Clustering and Visualization of Solr search results (video from Berlin BuzzWords conference, 2011):  
<http://vimeo.com/26616444>

# Spatial Search

Solr supports location data for use in spatial/geospatial searches.

Using spatial search, you can:

- Index points or other shapes
- Filter search results by a bounding box or circle or by other shapes
- Sort or boost scoring by distance between points, or relative area between rectangles
- Generate a 2D grid of facet count numbers for heatmap generation or point-plotting.

There are four main field types available for spatial search:

- LatLonPointSpatialField
- LatLonType (now deprecated) and its non-geodetic twin PointType
- SpatialRecursivePrefixTreeFieldType (RPT for short), including RptWithGeometrySpatialField, a derivative
- BBoxField

LatLonPointSpatialField is the ideal field type for the most common use-cases for lat-lon point data. It replaces LatLonType which still exists for backwards compatibility. RPT offers some more features for more advanced/custom use cases and options like polygons and heatmaps.

RptWithGeometrySpatialField is for indexing and searching non-point data though it can do points too. It can't do sorting/boosting.

BBoxField is for indexing bounding boxes, querying by a box, specifying a search predicate (Intersects,Within,Contains,Disjoint,Equals), and a relevancy sort/boost like overlapRatio or simply the area.

Some esoteric details that are not in this guide can be found at <http://wiki.apache.org/solr/SpatialSearch>.

## LatLonPointSpatialField

Here's how LatLonPointSpatialField (LLPSF) should usually be configured in the schema:

```
<fieldType name="location" class="solr.LatLonPointSpatialField" docValues="true"/>
```

LLPSF supports toggling indexed, stored, docValues, and multiValued. LLPSF internally uses a 2-dimensional Lucene "Points" (BDK tree) index when "indexed" is enabled (the default). When "docValues" is enabled, a latitude and longitudes pair are bit-interleaved into 64 bits and put into Lucene DocValues. The accuracy of the docValues data is about a centimeter.

## Indexing Points

For indexing geodetic points (latitude and longitude), supply it in "lat,lon" order (comma separated).

For indexing non-geodetic points, it depends. Use x y (a space) if RPT. For PointType however, use x,y (a

comma).

If you'd rather use a standard industry format, Solr supports [WKT](#) and [GeoJSON](#). However it's much bulkier than the raw coordinates for such simple data. (Not supported by the deprecated LatLonType or PointType)

## Indexing GeoJSON and WKT

Using the bin/post tool:

```
bin/post -type "application/json" -url
"http://localhost:8983/solr/mycollection/update?format=geojson" /path/to/geojson.file
```

The key parameter to pass in with your request is:

**format**

The format of the file to pass in. Accepted values: WKT or geojson.

## Searching with Query Parsers

There are two spatial Solr "query parsers" for geospatial search: geofilt and bbox. They take the following parameters:

**d**

The radial distance, usually in kilometers. RPT & BBoxField can set other units via the setting `distanceUnits`.

**pt**

The center point using the format "lat,lon" if latitude & longitude. Otherwise, "x,y" for PointType or "x y" for RPT field types.

**sfield**

A spatial indexed field.

**score**

(Advanced option; not supported by LatLonType (deprecated) or PointType) If the query is used in a scoring context (e.g., as the main query in `q`), this [local parameter](#) determines what scores will be produced. Valid values are:

- none: A fixed score of 1.0. (the default)
- kilometers: distance in kilometers between the field value and the specified center point
- miles: distance in miles between the field value and the specified center point
- degrees: distance in degrees between the field value and the specified center point
- distance: distance between the field value and the specified center point in the `distanceUnits` configured for this field
- recipDistance: 1 / the distance



Don't use this for indexed non-point shapes (e.g., polygons). The results will be erroneous. And with RPT, it's only recommended for multi-valued point data, as the implementation doesn't scale very well and for single-valued fields, you should instead use a separate non-RPT field purely for distance sorting.

When used with BBoxField, additional options are supported:

- overlapRatio: The relative overlap between the indexed shape & query shape.
- area: haversine based area of the overlapping shapes expressed in terms of the distanceUnits configured for this field
- area2D: cartesian coordinates based area of the overlapping shapes expressed in terms of the distanceUnits configured for this field

#### filter

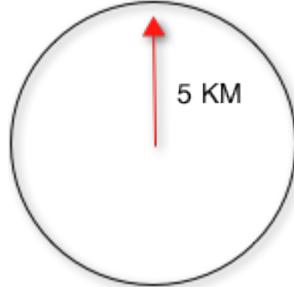
(Advanced option; not supported by LatLonType (deprecated) or PointType). If you only want the query to score (with the above score local parameter), not filter, then set this local parameter to false.

## geofilt

The geofilt filter allows you to retrieve results based on the geospatial distance (AKA the "great circle distance") from a given point. Another way of looking at it is that it creates a circular shape filter. For example, to find all documents within five kilometers of a given lat/lon point, you could enter:

```
&q=*:&fq={!geofilt sfield=store}&pt=45.15,-93.85&d=5
```

This filter returns all results within a circle of the given radius around the initial point:



## bbox

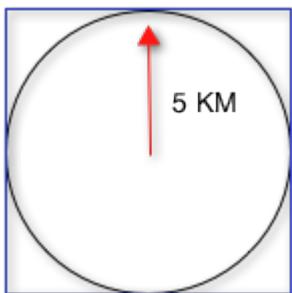
The bbox filter is very similar to geofilt except it uses the *bounding box* of the calculated circle. See the blue box in the diagram below. It takes the same parameters as geofilt.

Here's a sample query:

```
&q=*:&fq={!bbox sfield=store}&pt=45.15,-93.85&d=5
```

The rectangular shape is faster to compute and so it's sometimes used as an alternative to geofilt when it's acceptable to return points outside of the radius. However, if the ideal goal is a circle but you want it to run faster, then instead consider using the RPT field and try a large distErrPct value like 0.1 (10% radius).

This will return results outside the radius but it will do so somewhat uniformly around the shape.



When a bounding box includes a pole, the bounding box ends up being a "bounding bowl" (a *spherical cap*) that includes all values north of the lowest latitude of the circle if it touches the north pole (or south of the highest latitude if it touches the south pole).

## Filtering by an Arbitrary Rectangle

Sometimes the spatial search requirement calls for finding everything in a rectangular area, such as the area covered by a map the user is looking at. For this case, geofilt and bbox won't cut it. This is somewhat of a trick, but you can use Solr's range query syntax for this by supplying the lower-left corner as the start of the range and the upper-right corner as the end of the range.

Here's an example:

```
&q=*&fq=store:[45, -94 TO 46, -93]
```

LatLonType (deprecated) does **not** support rectangles that cross the dateline. For RPT and BBoxField, if you are non-geospatial coordinates (geo="false") then you must quote the points due to the space, e.g., "x y".

## Optimizing: Cache or Not

It's most common to put a spatial query into an "fq" parameter – a filter query. By default, Solr will cache the query in the filter cache.

If you know the filter query (be it spatial or not) is fairly unique and not likely to get a cache hit then specify cache="false" as a local-param as seen in the following example. The only spatial types which stand to benefit from this technique are LatLonPointSpatialField and LatLonType (deprecated). Enable docValues on the field (if it isn't already). LatLonType (deprecated) additionally requires a cost="100" (or more) local-param.

```
&q=...mykeywords...&fq=...someotherfilters...&fq={!geofilt cache=false}&sfield=store&pt=45.15,-93.85&d=5
```

LLPSF does not support Solr's "PostFilter".

## Distance Sorting or Boosting (Function Queries)

There are four distance function queries:

- geodist, see below, usually the most appropriate;
- dist, to calculate the p-norm distance between multi-dimensional vectors;
- hsin, to calculate the distance between two points on a sphere;
- sqedist, to calculate the squared Euclidean distance between two points.

For more information about these function queries, see the section on [Function Queries](#).

## geodist

geodist is a distance function that takes three optional parameters: (sfield, latitude, longitude). You can use the geodist function to sort results by distance or score return results.

For example, to sort your results by ascending distance, use a request like:

```
&q=*:&fq={!geofilt}&sfield=store&pt=45.15,-93.85&d=50&sort=geodist() asc
```

To return the distance as the document score, use a request like:

```
&q={!func}geodist()&sfield=store&pt=45.15,-93.85&sort=score+asc&f1=*,score
```

## More Spatial Search Examples

Here are a few more useful examples of what you can do with spatial search in Solr.

### Use as a Sub-Query to Expand Search Results

Here we will query for results in Jacksonville, Florida, or within 50 kilometers of 45.15,-93.85 (near Buffalo, Minnesota):

```
&q=*:&fq=(state:"FL" AND city:"Jacksonville") OR {!geofilt}&sfield=store&pt=45.15,-93.85&d=50&sort=geodist() asc
```

### Facet by Distance

To facet by distance, you can use the frange query parser:

```
&q=*:&sfield=store&pt=45.15,-93.85&facet.query={!frange l=0 u=5}geodist()&facet.query={!frange l=5.001 u=3000}geodist()
```

There are other ways to do it too, like using a `{!geofilt}` in each facet.query.

### Boost Nearest Results

Using the [DisMax](#) or [Extended DisMax](#), you can combine spatial search with the boost function to boost the nearest results:

```
&q.alt=*&fq={!geofilt}&sfield=store&pt=45.15,-
93.85&d=50&bf=recip(geodist(),2,200,20)&sort=score desc
```

## RPT

RPT refers to either SpatialRecursivePrefixTreeFieldType (aka simply RPT) and an extended version: RptWithGeometrySpatialField (aka RPT with Geometry). RPT offers several functional improvements over LatLonPointSpatialField:

- Non-geodetic – geo=false general x & y (*not* latitude and longitude) — if desired
- Query by polygons and other complex shapes, in addition to circles & rectangles
- Ability to index non-point shapes (e.g., polygons) as well as points – see RptWithGeometrySpatialField
- Heatmap grid faceting

RPT *shares* various features in common with LatLonPointSpatialField. Some are listed here:

- Latitude/Longitude indexed point data; possibly multi-valued
- Fast filtering with geofilt, bbox filters, and range query syntax (dateline crossing is supported)
- Well-Known-Text (WKT) shape syntax (required for specifying polygons & other complex shapes), and GeoJSON too. In addition to indexing and searching, this works with the wt=geojson (GeoJSON Solr response-writer) and [geo f=myfield] (geo Solr document-transformer).
- Sort/boost via geodist — *although not recommended*



Although RPT supports distance sorting/boosting, it is so inefficient at doing this that it might be removed in the future. Fortunately, you can use LatLonPointSpatialField *as well* as RPT. Use LLPSF for the distance sorting/boosting; it only needs to have docValues for this; the index attribute can be disabled as it won't be used.

## Schema Configuration for RPT

To use RPT, the field type must be registered and configured in schema.xml. There are many options for this field type.

### name

The name of the field type.

### class

This should be solr.SpatialRecursivePrefixTreeFieldType. But be aware that the Lucene spatial module includes some other so-called "spatial strategies" other than RPT, notably TermQueryPT\*, BBox, PointVector\*, and SerializedDV. Solr requires a field type to parallel these in order to use them. The asterisked ones have them.

### spatialContextFactory

This is a Java class name to an internal extension point governing support for shape definitions & parsing. If you require polygon support, set this to JTS – an alias for org.locationtech.spatial4j.context.jts.JtsSpatialContextFactory; otherwise it can be omitted.

See important info below about JTS. (note: prior to Solr 6, the "org.locationtech.spatial4j" part was "com.spatial4j.core" and there used to be no convenience JTS alias)

#### geo

If true, the default, latitude and longitude coordinates will be used and the mathematical model will generally be a sphere. If false, the coordinates will be generic X & Y on a 2D plane using Euclidean/Cartesian geometry.

#### format

Defines the shape syntax/format to be used. Defaults to WKT but GeoJSON is another popular format. Spatial4j governs this feature and supports [other formats](#). If a given shape is parseable as "lat,lon" or "x y" then that is always supported.

#### distanceUnits

This is used to specify the units for distance measurements used throughout the use of this field. This can be degrees, kilometers or miles. It is applied to nearly all distance measurements involving the field: maxDistErr, distErr, d, geodist and the score when score is distance, area, or area2d. However, it doesn't affect distances embedded in WKT strings, (e.g., BUFFER(POINT(200 10), 0.2)), which are still in degrees.

distanceUnits defaults to either kilometers if geo is true, or degrees if geo is false.

distanceUnits replaces the units attribute; which is now deprecated and mutually exclusive with this attribute.

#### distErrPct

Defines the default precision of non-point shapes (both index & query), as a fraction between 0.0 (fully precise) to 0.5. The closer this number is to zero, the more accurate the shape will be. However, more precise indexed shapes use more disk space and take longer to index.

Bigger distErrPct values will make queries faster but less accurate. At query time this can be overridden in the query syntax, such as to 0.0 so as to not approximate the search shape. The default for the RPT field is 0.025.



For RPTWithGeometrySpatialField (see below), there's always complete accuracy with the serialized geometry and so this doesn't control accuracy so much as it controls the trade-off of how big the index should be. distErrPct defaults to 0.15 for that field.

#### maxDistErr

Defines the highest level of detail required for indexed data. If left blank, the default is one meter – just a bit less than 0.000009 degrees. This setting is used internally to compute an appropriate maxLevels (see below).

#### worldBounds

Defines the valid numerical ranges for x and y, in the format of ENVELOPE(minX, maxX, maxY, minY). If geo="true", the standard lat-lon world boundaries are assumed. If geo=false, you should define your boundaries.

#### distCalculator

Defines the distance calculation algorithm. If geo=true, haversine is the default. If geo=false, cartesian

will be the default. Other possible values are lawOfCosines, vincentySphere and cartesian<sup>^2</sup>.

#### prefixTree

Defines the spatial grid implementation. Since a PrefixTree (such as RecursivePrefixTree) maps the world as a grid, each grid cell is decomposed to another set of grid cells at the next level.

If geo=true then the default prefix tree is geohash, otherwise it's quad. Geohash has 32 children at each level, quad has 4. Geohash can only be used for geo=true as it's strictly geospatial.

A third choice is packedQuad, which is generally more efficient than quad, provided there are many levels — perhaps 20 or more.

#### maxLevels

Sets the maximum grid depth for indexed data. Instead, it's usually more intuitive to compute an appropriate maxLevels by specifying maxDistErr.

**And there are others:** normWrapLongitude, datelineRule, validationRule, autoIndex, allowMultiOverlap, precisionModel. For further info, see notes below about spatialContextFactory implementations referenced above, especially the link to the JTS based one.

## Standard Shapes

The RPT field types support a set of standard shapes: points, circles (aka buffered points), envelopes (aka rectangles or bounding boxes), line strings, polygons, and "multi" variants of these. The envelopes and line strings are Euclidean/cartesian (flat 2D) shapes. Underlying Solr is the Spatial4j library which implements them. To support other shapes, you can configure the spatialContextFactory attribute on the field type to reference other options. Two are available: JTS and Geo3D.

## JTS and Polygons (flat)

The [JTS Topology Suite](#) is a popular computational geometry library with a Euclidean/cartesian (flat 2D) model. It supports a variety of shapes including polygons, buffering shapes, and some invalid polygon repair fall-backs. With the help of Spatial4j, included with Solr, the polygons support dateline (anti-meridian) crossing. You must download it (a JAR file) and put that in a special location internal to Solr: SOLR\_INSTALL/server/solr-webapp/webapp/WEB-INF/lib/. You can readily download it here: <http://central.maven.org/maven2/org/locationtech/jts/jts-core/1.15.0/>. *It will not work if placed in other more typical Solr lib directories, unfortunately.*

Set the spatialContextFactory attribute on the field type to JTS.

When activated, there are additional configuration attributes available; see [org.locationtech.spatial4j.context.jts.JtsSpatialContextFactory](#) for the Javadocs, and remember to look at the superclass's options as well. One option in particular you should most likely enable is autoIndex (i.e., use JTS's PreparedGeometry) as it's been shown to be a major performance boost for non-trivial polygons.

```
<fieldType name="location_rpt" class="solr.SpatialRecursivePrefixTreeFieldType"
 spatialContextFactory="JTS"
 autoIndex="true"
 validationRule="repairBuffer0"
 distErrPct="0.025"
 maxDistErr="0.001"
 distanceUnits="kilometers" />
```

Once the field type has been defined, define a field that uses it.

Here's an example polygon query for a field "geo" that can be either solr.SpatialRecursivePrefixTreeFieldType or RptWithGeometrySpatialField:

```
&q=*&fq={!field f=geo}Intersects(POLYGON((-10 30, -40 40, -10 -20, 40 20, 0 0, -10 30)))
```

Inside the parenthesis following the search predicate is the shape definition. The format of that shape is governed by the format attribute on the field type, defaulting to WKT. If you prefer GeoJSON, you can specify that instead.

Beyond this Reference Guide and Spatila4j's docs, there are some details that remain at the Solr Wiki at <http://wiki.apache.org/solr/SolrAdaptersForLuceneSpatial4>.

## Geo3D and Polygons (on the ellipsoid)

Geo3D is the colloquial name of the Lucene spatial-3d module, included with Solr. It's a computational geometry library implementing a variety of shapes (including polygons) on a sphere or WGS84 ellipsoid. Geo3D is particularly suited for spatial applications where the geometries cover large distances across the globe. Geo3D is named as-such due to its internal implementation that uses geocentric coordinates (X,Y,Z), **not** for 3-dimensional geometry, which it does not support. Despite these internal details, you still supply latitude and longitude as you would normally in Solr.

Set the spatialContextFactory attribute on the field type to Geo3D.

```
<fieldType name="geom" class="solr.SpatialRecursivePrefixTreeFieldType"
 spatialContextFactory="Geo3D" planetModel="WGS84"/><!-- or "sphere" -->
```

Once the field type has been defined, define a field that uses it.

## RptWithGeometrySpatialField

The RptWithGeometrySpatialField field type is a derivative of SpatialRecursivePrefixTreeFieldType that also stores the original geometry internally in Lucene DocValues, which it uses to achieve accurate search. It can also be used for indexed point fields. The Intersects predicate (the default) is particularly fast, since many search results can be returned as an accurate hit without requiring a geometry check. This field type is configured just like RPT except that the default distErrPct is 0.15 (higher than 0.025) because the grid squares are purely for performance and not to fundamentally represent the shape.

An optional in-memory cache can be defined in solrconfig.xml, which should be done when the data tends

to have shapes with many vertices. Assuming you name your field "geom", you can configure an optional cache in `solrconfig.xml` by adding the following – notice the suffix of the cache name:

```
<cache name="perSegSpatialFieldCache_geom"
 class="solr.LRUCache"
 size="256"
 initialSize="0"
 autowarmCount="100%"
 regenerator="solr.NoOpRegenerator"/>
```

When using this field type, you will likely *not* want to mark the field as stored because it's redundant with the DocValues data and surely larger because of the formatting (be it WKT or GeoJSON). To retrieve the spatial data in search results from DocValues, use the [geo] transformer — [Transforming Result Documents](#).

## Heatmap Faceting

The RPT field supports generating a 2D grid of facet counts for documents having spatial data in each grid cell. For high-detail grids, this can be used to plot points, and for lesser detail it can be used for heatmap generation. The grid cells are determined at index-time based on RPT's configuration. At facet counting time, the indexed cells in the region of interest are traversed and a grid of counters corresponding to each cell are incremented. Solr can return the data in a straight-forward 2D array of integers or in a PNG which compresses better for larger data sets but must be decoded.

The heatmap feature is accessible both from Solr's standard faceting feature, plus the newer more flexible [JSON Facet API](#). We'll proceed now with standard faceting. As a part of faceting, it supports the key local parameter as well as excluding tagged filter queries, just like other types of faceting do. This allows multiple heatmaps to be returned on the same field with different filters.

`facet`

Set to true to enable standard faceting.

`facet.heatmap`

The field name of type RPT.

`facet.heatmap.geom`

The region to compute the heatmap on, specified using the rectangle-range syntax or WKT. It defaults to the world. ex: `[-180 -90" TO "180 90"]`.

`facet.heatmap.gridLevel`

A specific grid level, which determines how big each grid cell is. Defaults to being computed via `distErrPct` (or `distErr`).

`facet.heatmap.distErrPct`

A fraction of the size of geom used to compute gridLevel. Defaults to 0.15. It's computed the same as a similarly named parameter for RPT.

`facet.heatmap.distErr`

A cell error distance used to pick the grid level indirectly. It's computed the same as a similarly named parameter for RPT.

facet.heatmap.format

The format, either ints2D (default) or png.

You'll experiment with different distErrPct values (probably 0.10 - 0.20) with various input geometries till the default size is what you're looking for. The specific details of how it's computed isn't important. For high-detail grids used in point-plotting (loosely one cell per pixel), set distErr to be the number of decimal-degrees of several pixels or so of the map being displayed. Also, you probably don't want to use a geohash-based grid because the cell orientation between grid levels flip-flops between being square and rectangle. Quad is consistent and has more levels, albeit at the expense of a larger index.

Here's some sample output in JSON (with "..." inserted for brevity):

```
{gridLevel=6,columns=64,rows=64,minX=-180.0,maxX=180.0,minY=-90.0,maxY=90.0,
counts_ints2D=[[0, 0, 2, 1, ...],[1, 1, 3, 2, ...],...]}
```

The output shows the gridLevel which is interesting since it's often computed from other parameters. If an interface being developed allows an explicit resolution increase/decrease feature then subsequent requests can specify the gridLevel explicitly.

The minX, maxX, minY, maxY reports the region where the counts are. This is the minimally enclosing bounding rectangle of the input geom at the target grid level. This may wrap the dateline. The columns and rows values are how many columns and rows that the output rectangle is to be divided by evenly. Note: Don't divide an on-screen projected map rectangle evenly to plot these rectangles/points since the cell data is in the coordinate space of decimal degrees if geo=true or whatever units were given if geo=false. This could be arranged to be the same as an on-screen map but won't necessarily be.

The counts\_ints2D key has a 2D array of integers. The initial outer level is in row order (top-down), then the inner arrays are the columns (left-right). If any array would be all zeros, a null is returned instead for efficiency reasons. The entire value is null if there is no matching spatial data.

If format=png then the output key is counts\_png. It's a base-64 encoded string of a 4-byte PNG. The PNG logically holds exactly the same data that the ints2D format does. Note that the alpha channel byte is flipped to make it easier to view the PNG for diagnostic purposes, since otherwise counts would have to exceed  $2^{24}$  before it becomes non-opaque. Thus counts greater than this value will become opaque.

## BBoxField

The BBoxField field type indexes a single rectangle (bounding box) per document field and supports searching via a bounding box. It supports most spatial search predicates, it has enhanced relevancy modes based on the overlap or area between the search rectangle and the indexed rectangle. It's particularly useful for its relevancy modes. To configure it in the schema, use a configuration like this:

```
<field name="bbox" type="bbox" />

<fieldType name="bbox" class="solr.BBoxField"
 geo="true" distanceUnits="kilometers" numberType="pdouble" />
<fieldType name="pdouble" class="solr.DoublePointField" docValues="true"/>
```

BBoxField is actually based off of 4 instances of another field type referred to by numberType. It also uses a boolean to flag a dateline cross. Assuming you want to use the relevancy feature, docValues is required. Some of the attributes are in common with the RPT field like geo, units, worldBounds, and spatialContextFactory because they share some of the same spatial infrastructure.

To index a box, add a field value to a bbox field that's a string in the WKT/CQL ENVELOPE syntax. Example: ENVELOPE(-10, 20, 15, 10) which is minX, maxX, maxY, minY order. The parameter ordering is unintuitive but that's what the spec calls for. Alternatively, you could provide a rectangular polygon in WKT (or GeoJSON if you set set format="GeoJSON").

To search, you can use the {!bbox} query parser, or the range syntax e.g., [10,-10 TO 15,20], or the ENVELOPE syntax wrapped in parenthesis with a leading search predicate. The latter is the only way to choose a predicate other than Intersects. For example:

```
&q={!field f=bbox}Contains(ENVELOPE(-10, 20, 15, 10))
```

Now to sort the results by one of the relevancy modes, use it like this:

```
&q={!field f=bbox score=overlapRatio}Intersects(ENVELOPE(-10, 20, 15, 10))
```

The score local parameter can be one of overlapRatio, area, and area2D. area scores by the document area using surface-of-a-sphere (assuming geo=true) math, while area2D uses simple width \* height. overlapRatio computes a [0-1] ranged score based on how much overlap exists relative to the document's area and the query area. The javadocs of [BBoxOverlapRatioValueSource](#) have more info on the formula. There is an additional parameter queryTargetProportion that allows you to weight the query side of the formula to the index (target) side of the formula. You can also use &debug=results to see useful score computation info.

# The Terms Component

The Terms Component provides access to the indexed terms in a field and the number of documents that match each term. This can be useful for building an auto-suggest feature or any other feature that operates at the term level instead of the search or document level. Retrieving terms in index order is very fast since the implementation directly uses Lucene's TermEnum to iterate over the term dictionary.

In a sense, this search component provides fast field-faceting over the whole index, not restricted by the base query or any filters. The document frequencies returned are the number of documents that match the term, including any documents that have been marked for deletion but not yet removed from the index.

## Configuring the Terms Component

By default, the Terms Component is already configured in `solrconfig.xml` for each collection.

### Defining the Terms Component

Defining the Terms search component is straightforward: simply give it a name and use the class `solr.TermsComponent`.

```
<searchComponent name="terms" class="solr.TermsComponent"/>
```

This makes the component available for use, but by itself will not be useable until included with a request handler.

### Using the Terms Component in a Request Handler

The terms component is included with the `/terms` request handler, which is among Solr's out-of-the-box request handlers - see [Implicit RequestHandlers](#).

Note that the defaults for this request handler set the parameter "terms" to true, which allows terms to be returned on request. The parameter "distrib" is set to false, which allows this handler to be used only on a single Solr core.

You could add this component to another handler if you wanted to, and pass "terms=true" in the HTTP request in order to get terms back. If it is only defined in a separate handler, you must use that handler when querying in order to get terms and not regular documents as results.

### Terms Component Parameters

The parameters below allow you to control what terms are returned. You can also configure any of these with the request handler if you'd like to set them permanently. Or, you can add them to the query request. These parameters are:

`terms`

If set to true, enables the Terms Component. By default, the Terms Component is off (false).

Example: `terms=true`

**terms.fl**

Specifies the field from which to retrieve terms. This parameter is required if `terms=true`.

Example: `terms.fl=title`

**terms.list**

Fetches the document frequency for a comma delimited list of terms. Terms are always returned in index order. If `terms.ttf` is set to true, also returns their total term frequency. If multiple `terms.fl` are defined, these statistics will be returned for each term in each requested field.

Example: `terms.list=termA,termB,termC`

**terms.limit**

Specifies the maximum number of terms to return. The default is 10. If the limit is set to a number less than 0, then no maximum limit is enforced. Although this is not required, either this parameter or `terms.upper` must be defined.

Example: `terms.limit=20`

**terms.lower**

Specifies the term at which to start. If not specified, the empty string is used, causing Solr to start at the beginning of the field.

Example: `terms.lower=orange`

**terms.lower.incl**

If set to true, includes the lower-bound term (specified with `terms.lower` in the result set).

Example: `terms.lower.incl=false`

**terms.mincount**

Specifies the minimum document frequency to return in order for a term to be included in a query response. Results are inclusive of the mincount (that is,  $\geq$  mincount).

Example: `terms.mincount=5`

**terms.maxcount**

Specifies the maximum document frequency a term must have in order to be included in a query response. The default setting is -1, which sets no upper bound. Results are inclusive of the maxcount (that is,  $\leq$  maxcount).

Example: `terms.maxcount=25`

**terms.prefix**

Restricts matches to terms that begin with the specified string.

Example: `terms.prefix=inter`

**terms.raw**

If set to true, returns the raw characters of the indexed term, regardless of whether it is human-readable. For instance, the indexed form of numeric numbers is not human-readable.

Example: `terms.raw=true`

## terms.regex

Restricts matches to terms that match the regular expression.

Example: `terms.regex=.*pedist`

## terms.regex.flag

Defines a Java regex flag to use when evaluating the regular expression defined with `terms.regex`. See <http://docs.oracle.com/javase/tutorial/essential/regex/pattern.html> for details of each flag. Valid options are:

- `case_insensitive`
- `comments`
- `multiline`
- `literal`
- `dotall`
- `unicode_case`
- `canon_eq`
- `unix_lines`

Example: `terms.regex.flag=case_insensitive`

## terms.stats

Include index statistics in the results. Currently returns only the **numDocs** for a collection. When combined with `terms.list` it provides enough information to compute inverse document frequency (IDF) for a list of terms.

## terms.sort

Defines how to sort the terms returned. Valid options are `count`, which sorts by the term frequency, with the highest term frequency first, or `index`, which sorts in index order.

Example: `terms.sort=index`

## terms.ttf

If set to true, returns both df (docFreq) and ttf (totalTermFreq) statistics for each requested term in `terms.list`. In this case, the response format is:

```
<lst name="terms">
 <lst name="field">
 <lst name="termA">
 <long name="df">22</long>
 <long name="ttf">73</long>
 </lst>
 </lst>
</lst>
```

## terms.upper

Specifies the term to stop at. Although this parameter is not required, either this parameter or `terms.limit` must be defined.

Example: terms.upper=plum

terms.upper.incl

If set to true, the upper bound term is included in the result set. The default is false.

Example: terms.upper.incl=true

The response to a terms request is a list of the terms and their document frequency values.

You may also be interested in the [TermsComponent javadoc](#).

## Terms Component Examples

All of the following sample queries work with Solr's "bin/solr -e techproducts" example.

### Get Top 10 Terms

This query requests the first ten terms in the name field:

```
http://localhost:8983/solr/techproducts/terms?terms.fl=name&wt=xml
```

Results:

```
<response>
<lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">2</int>
</lst>
<lst name="terms">
 <lst name="name">
 <int name="one">5</int>
 <int name="184">3</int>
 <int name="1gb">3</int>
 <int name="3200">3</int>
 <int name="400">3</int>
 <int name="ddr">3</int>
 <int name="gb">3</int>
 <int name="ipod">3</int>
 <int name="memory">3</int>
 <int name="pc">3</int>
 </lst>
</lst>
</response>
```

### Get First 10 Terms Starting with Letter 'a'

This query requests the first ten terms in the name field, in index order (instead of the top 10 results by document count):

```
http://localhost:8983/solr/techproducts/terms?terms.fl=name&terms.lower=a&terms.sort=index&wt=xml
```

Results:

```
<response>
<lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">0</int>
</lst>
<lst name="terms">
 <lst name="name">
 <int name="a">1</int>
 <int name="all">1</int>
 <int name="apple">1</int>
 <int name="asus">1</int>
 <int name="ata">1</int>
 <int name="ati">1</int>
 <int name="belkin">1</int>
 <int name="black">1</int>
 <int name="british">1</int>
 <int name="cable">1</int>
 </lst>
</lst>
</response>
```

## SolrJ Invocation

```
SolrQuery query = new SolrQuery();
query.setRequestHandler("/terms");
query.setTerms(true);
query.setTermsLimit(5);
query.setTermsLower("s");
query.setTermsPrefix("s");
query.addTermsField("terms_s");
query.setTermsMinCount(1);

QueryRequest request = new QueryRequest(query);
List<Term> terms = request.process(getSolrClient()).getTermsResponse().getTerms("terms_s");
```

## Using the Terms Component for an Auto-Suggest Feature

If the [Suggester](#) doesn't suit your needs, you can use the Terms component in Solr to build a similar feature for your own search application. Simply submit a query specifying whatever characters the user has typed so far as a prefix. For example, if the user has typed "at", the search engine's interface would submit the following query:

```
http://localhost:8983/solr/techproducts/terms?terms.fl=name&terms.prefix=at&wt=xml
```

Result:

```
<response>
<lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1</int>
</lst>
<lst name="terms">
 <lst name="name">
 <int name="ata">1</int>
 <int name="ati">1</int>
 </lst>
</lst>
</response>
```

You can use the parameter `omitHeader=true` to omit the response header from the query response, like in this example, which also returns the response in JSON format:

```
http://localhost:8983/solr/techproducts/terms?terms.fl=name&terms.prefix=at&omitHeader=true
```

Result:

```
{
 "terms": {
 "name": [
 "ata",
 1,
 "ati",
 1
]
 }
}
```

## Distributed Search Support

The TermsComponent also supports distributed indexes. For the `/terms` request handler, you must provide the following two parameters:

`shards`

Specifies the shards in your distributed indexing configuration. For more information about distributed indexing, see [Distributed Search with Index Sharding](#).

The `shards` parameter is subject to a host whitelist that has to be configured in the component's parameters using the configuration key `shardsWhitelist` and the list of hosts as values.

By default the whitelist will be populated with all live nodes when running in SolrCloud mode. If you need to disable this feature for backwards compatibility, you can set the system property `solr.disable.shardsWhitelist=true`.

See the section [Configuring the ShardHandlerFactory](#) for more information about how the whitelist works.

#### shards.qt

Specifies the request handler Solr uses for requests to shards.

# The Term Vector Component

The TermVectorComponent is a search component designed to return additional information about documents matching your search.

For each document in the response, the TermVectorComponent can return the term vector, the term frequency, inverse document frequency, position, and offset information.

## Term Vector Component Configuration

The TermVectorComponent is not enabled implicitly in Solr - it must be explicitly configured in your solrconfig.xml file. The examples on this page show how it is configured in Solr's "techproducts" example:

```
bin/solr -e techproducts
```

To enable this component, you need to configure it using a `searchComponent` element:

```
<searchComponent name="tvComponent" class="org.apache.solr.handler.component.TermVectorComponent" />
```

A request handler must then be configured to use this component name. In the techproducts example, the component is associated with a special request handler named /tvrh, that enables term vectors by default using the `tv=true` parameter; but you can associate it with any request handler:

```
<requestHandler name="/tvrh" class="org.apache.solr.handler.component.SearchHandler">
 <lst name="defaults">
 <bool name="tv">true</bool>
 </lst>
 <arr name="last-components">
 <str>tvComponent</str>
 </arr>
</requestHandler>
```

Once your handler is defined, you may use in conjunction with any schema (that has a `uniqueKeyField`) to fetch term vectors for fields configured with the `termVector` attribute, such as in the techproducts sample schema. For example:

```
<field name="includes"
 type="text_general"
 indexed="true"
 stored="true"
 multiValued="true"
 termVectors="true"
 termPositions="true"
 termOffsets="true" />
```

## Invoking the Term Vector Component

The example below shows an invocation of this component using the above configuration:

```
http://localhost:8983/solr/techproducts/tvrh?q=*&start=0&rows=10&fl=id,includes&wt=xml
```

```
...
<lst name="termVectors">
 <lst name="GB18030TEST">
 <str name="uniqueKey">GB18030TEST</str>
 </lst>
 <lst name="EN7800GTX/2DHTV/256M">
 <str name="uniqueKey">EN7800GTX/2DHTV/256M</str>
 </lst>
 <lst name="100-435805">
 <str name="uniqueKey">100-435805</str>
 </lst>
 <lst name="3007WFP">
 <str name="uniqueKey">3007WFP</str>
 <lst name="includes">
 <lst name="cable"/>
 <lst name="usb"/>
 </lst>
 </lst>
 <lst name="SOLR1000">
 <str name="uniqueKey">SOLR1000</str>
 </lst>
 <lst name="0579B002">
 <str name="uniqueKey">0579B002</str>
 </lst>
 <lst name="UTF8TEST">
 <str name="uniqueKey">UTF8TEST</str>
 </lst>
 <lst name="9885A004">
 <str name="uniqueKey">9885A004</str>
 <lst name="includes">
 <lst name="32mb"/>
 <lst name="av"/>
 <lst name="battery"/>
 <lst name="cable"/>
 <lst name="card"/>
 <lst name="sd"/>
 <lst name="usb"/>
 </lst>
 </lst>
 <lst name="adata">
 <str name="uniqueKey">adata</str>
 </lst>
 <lst name="apple">
 <str name="uniqueKey">apple</str>
 </lst>
</lst>
```

## Term Vector Request Parameters

The example below shows some of the available request parameters for this component:

```
http://localhost:8983/solr/techproducts/tvrh?q=includes:[* TO
*]&rows=10&indent=true&tv=true&tv.tf=true&tv.df=true&tv.positions=true&tv.offsets=true&tv.payload
s=true&tv.fl=includes
```

## tv

If true, the Term Vector Component will run.

## tv.docIds

For a given comma-separated list of Lucene document IDs (**not** the Solr Unique Key), term vectors will be returned.

## tv.fl

For a given comma-separated list of fields, term vectors will be returned. If not specified, the fl parameter is used.

## tv.all

If true, all the boolean parameters listed below (tv.df, tv.offsets, tv.positions, tv.payloads, tv.tf and tv.tf\_idf) will be enabled.

## tv.df

If true, returns the Document Frequency (DF) of the term in the collection. This can be computationally expensive.

## tv.offsets

If true, returns offset information for each term in the document.

## tv.positions

If true, returns position information.

## tv.payloads

If true, returns payload information.

## tv.tf

If true, returns document term frequency info for each term in the document.

## tv.tf\_idf

If true, calculates TF / DF (i.e.,: TF \* IDF) for each term. Please note that this is a *literal* calculation of "Term Frequency multiplied by Inverse Document Frequency" and **not** a classical TF-IDF similarity measure.

This parameter requires both tv.tf and tv.df to be "true". This can be computationally expensive. (The results are not shown in example output)

To see an example of TermVector component output, see the Wiki page: <http://wiki.apache.org/solr/TermVectorComponentExampleOptions>

For schema requirements, see also the section [Field Properties by Use Case](#).

## SolrJ and the Term Vector Component

Neither the SolrQuery class nor the QueryResponse class offer specific method calls to set Term Vector Component parameters or get the "termVectors" output. However, there is a patch for it: [SOLR-949](#).

# The Stats Component

The Stats component returns simple statistics for numeric, string, and date fields within the document set.

The sample queries in this section assume you are running the “techproducts” example included with Solr:

```
bin/solr -e techproducts
```

## Stats Component Parameters

The Stats Component accepts the following parameters:

`stats`

If true, then invokes the Stats component.

`stats.field`

Specifies a field for which statistics should be generated. This parameter may be invoked multiple times in a query in order to request statistics on multiple fields.

[Local Parameters](#) may be used to indicate which subset of the supported statistics should be computed, and/or that statistics should be computed over the results of an arbitrary numeric function (or query) instead of a simple field name. See the examples below.

## Stats Component Example

The query below demonstrates computing stats against two different fields numeric fields, as well as stats over the results of a termfreq() function call using the text field:

```
http://localhost:8983/solr/techproducts/select?q=*&wt=xml&stats=true&stats.field={!func}termfre
q('text','memory')&stats.field=price&stats.field=popularity&rows=0&indent=true
```

```
<lst name="stats">
 <lst name="stats_fields">
 <lst name="termfreq(text, memory)">
 <double name="min">0.0</double>
 <double name="max">3.0</double>
 <long name="count">32</long>
 <long name="missing">0</long>
 <double name="sum">10.0</double>
 <double name="sumOfSquares">22.0</double>
 <double name="mean">0.3125</double>
 <double name="stddev">0.7803018439949604</double>
 <lst name="facets" />
 </lst>
 <lst name="price">
 <double name="min">0.0</double>
 <double name="max">2199.0</double>
 <long name="count">16</long>
 <long name="missing">16</long>
 <double name="sum">5251.270030975342</double>
 <double name="sumOfSquares">6038619.175900028</double>
 <double name="mean">328.20437693595886</double>
 <double name="stddev">536.3536996709846</double>
 <lst name="facets" />
 </lst>
 <lst name="popularity">
 <double name="min">0.0</double>
 <double name="max">10.0</double>
 <long name="count">15</long>
 <long name="missing">17</long>
 <double name="sum">85.0</double>
 <double name="sumOfSquares">603.0</double>
 <double name="mean">5.6666666666666667</double>
 <double name="stddev">2.943920288775949</double>
 <lst name="facets" />
 </lst>
 </lst>
</lst>
```

## Statistics Supported

The table below explains the statistics supported by the Stats component. Not all statistics are supported for all field types, and not all statistics are computed by default (see [Local Parameters with the Stats Component](#) below for details)

### min

The minimum value of the field/function in all documents in the set. This statistic is computed for all field types and is computed by default.

### max

The maximum value of the field/function in all documents in the set. This statistic is computed for all field

types and is computed by default.

#### sum

The sum of all values of the field/function in all documents in the set. This statistic is computed for numeric and date field types and is computed by default.

#### count

The number of values found in all documents in the set for this field/function. This statistic is computed for all field types and is computed by default.

#### missing

The number of documents in the set which do not have a value for this field/function. This statistic is computed for all field types and is computed by default.

#### sumOfSquares

Sum of all values squared (a by product of computing stddev). This statistic is computed for numeric and date field types and is computed by default.

#### mean

The average ( $v_1 + v_2 \dots + v_N / N$ ). This statistic is computed for numeric and date field types and is computed by default.

#### stddev

Standard deviation, measuring how widely spread the values in the data set are. This statistic is computed for numeric and date field types and is computed by default.

#### percentiles

A list of percentile values based on cut-off points specified by the parameter value, such as 1, 99, 99.9.

These values are an approximation, using the [t-digest algorithm](#). This statistic is computed for numeric field types and is not computed by default.

#### distinctValues

The set of all distinct values for the field/function in all of the documents in the set. This calculation can be very expensive for fields that do not have a tiny cardinality. This statistic is computed for all field types but is not computed by default.

#### countDistinct

The exact number of distinct values in the field/function in all of the documents in the set. This calculation can be very expensive for fields that do not have a tiny cardinality. This statistic is computed for all field types but is not computed by default.

#### cardinality

A statistical approximation (currently using the [HyperLogLog](#) algorithm) of the number of distinct values in the field/function in all of the documents in the set. This calculation is much more efficient than using the countDistinct option, but may not be 100% accurate.

Input for this option can be floating point number between 0.0 and 1.0 indicating how aggressively the algorithm should try to be accurate: 0.0 means use as little memory as possible; 1.0 means use as much memory as needed to be as accurate as possible. true is supported as an alias for 0.3.

This statistic is computed for all field types but is not computed by default.

## Local Parameters with the Stats Component

Similar to the [Facet Component](#), the stats.field parameter supports local parameters for:

- Tagging & Excluding Filters: stats.field={!ex=filterA}price
- Changing the Output Key: stats.field={!key=my\_price\_stats}price
- Tagging stats for [use with facet.pivot](#): stats.field={!tag=my\_pivot\_stats}price

Local parameters can also be used to specify individual statistics by name, overriding the set of statistics computed by default, e.g., stats.field={!min=true max=true percentiles='99,99.9,99.99'}price.



If any supported statistics are specified via local parameters, then the entire set of default statistics is overridden and only the requested statistics are computed.

Additional "Expert" local params are supported in some cases for affecting the behavior of some statistics:

- percentiles
  - tdigestCompression - a positive numeric value defaulting to 100.0 controlling the compression factor of the T-Digest. Larger values means more accuracy, but also uses more memory.
- cardinality
  - hllPreHashed - a boolean option indicating that the statistics are being computed over a "long" field that has already been hashed at index time – allowing the HLL computation to skip this step.
  - hllLog2m - an integer value specifying an explicit "log2m" value to use, overriding the heuristic value determined by the cardinality local param and the field type – see the [java-hll](#) documentation for more details
  - hllRegwidth - an integer value specifying an explicit "regwidth" value to use, overriding the heuristic value determined by the cardinality local param and the field type – see the [java-hll](#) documentation for more details

## Examples with Local Parameters

Here we compute some statistics for the price field. The min, max, mean, 90th, and 99th percentile price values are computed against all products that are in stock (q=: and fq=inStock:true), and independently all of the default statistics are computed against all products regardless of whether they are in stock or not (by excluding that filter).

```
http://localhost:8983/solr/techproducts/select?q=*&fq={!tag=stock_check}inStock:true&stats=true&stats.field={!ex=stock_check+key=inStock_prices+min=true+max=true+mean=true+percentiles='90,99'}price&stats.field={!key=all_prices}price&rows=0&indent=true&wt=xml
```

```
<lst name="stats">
 <lst name="stats_fields">
 <lst name="instock_prices">
 <double name="min">0.0</double>
 <double name="max">2199.0</double>
 <double name="mean">328.20437693595886</double>
 <lst name="percentiles">
 <double name="90.0">564.9700012207031</double>
 <double name="99.0">1966.6484985351556</double>
 </lst>
 </lst>
 <lst name="all_prices">
 <double name="min">0.0</double>
 <double name="max">2199.0</double>
 <long name="count">12</long>
 <long name="missing">5</long>
 <double name="sum">4089.880027770996</double>
 <double name="sumOfSquares">5385249.921747174</double>
 <double name="mean">340.823335647583</double>
 <double name="stddev">602.3683083752779</double>
 </lst>
 </lst>
</lst>
```

## The Stats Component and Faceting

Sets of stats.field parameters can be referenced by 'tag' when using Pivot Faceting to compute multiple statistics at every level (i.e.: field) in the tree of pivot constraints.

For more information and a detailed example, please see [Combining Stats Component With Pivots](#).

# The Query Elevation Component

The Query Elevation Component lets you configure the top results for a given query regardless of the normal Lucene scoring.

This is sometimes called "sponsored search", "editorial boosting", or "best bets." This component matches the user query text to a configured map of top results. The text can be any string or non-string IDs, as long as it's indexed. Although this component will work with any QueryParser, it makes the most sense to use with [DisMax](#) or [eDisMax](#).

The Query Elevation Component also supports distributed searching.

All of the sample configuration and queries used in this section assume you are running Solr's "techproducts" example:

```
bin/solr -e techproducts
```

## Configuring the Query Elevation Component

You can configure the Query Elevation Component in the `solrconfig.xml` file. Search components like `QueryElevationComponent` may be added to any request handler; a dedicated request handler is used here for brevity.

```
<searchComponent name="elevator" class="solr.QueryElevationComponent" >
 <!-- pick a fieldType to analyze queries -->
 <str name="queryFieldType">string</str>
 <str name="config-file">elevate.xml</str>
</searchComponent>

<requestHandler name="/elevate" class="solr.SearchHandler" startup="lazy">
 <lst name="defaults">
 <str name="echoParams">explicit</str>
 </lst>
 <arr name="last-components">
 <str>elevator</str>
 </arr>
</requestHandler>
```

Optionally, in the Query Elevation Component configuration you can also specify the following to distinguish editorial results from "normal" results:

```
<str name="editorialMarkerFieldName">foo</str>
```

The Query Elevation Search Component takes the following parameters:

`queryFieldType`

Specifies which fieldType should be used to analyze the incoming text. For example, it may be

appropriate to use a fieldType with a LowerCaseFilter.

#### config-file

Path to the file that defines query elevation. This file must exist in <instanceDir>/conf/<config-file> or <dataDir>/<config-file>. If the file exists in the conf/ directory it will be loaded once at startup. If it exists in the data/ directory, it will be reloaded for each IndexReader.

#### forceElevation

By default, this component respects the requested sort parameter: if the request asks to sort by date, it will order the results by date. If forceElevation=true (the default), results will first return the boosted docs, then order by date. This is also a request parameter, which will override the config.

#### useConfiguredElevatedOrder

When multiple docs are elevated, should their relative order be the order in the configuration file or should they be subject to whatever the sort criteria is? True by default. This is also a request parameter, which will override the config. The effect is most apparent when forceElevation is true and there is sorting on fields.

## The elevate.xml File

Elevated query results can be configured in an external XML file specified in the config-file argument. An elevate.xml file might look like this:

```
<elevate>
 <query text="foo bar">
 <doc id="1" />
 <doc id="2" />
 <doc id="3" />
 </query>

 <query text="ipod">
 <doc id="MA147LL/A" /> <!-- put the actual ipod at the top -->
 <doc id="IW-02" exclude="true" /> <!-- exclude this cable -->
 </query>
</elevate>
```

In this example, the query "foo bar" would first return documents 1, 2 and 3, then whatever normally appears for the same query. For the query "ipod", it would first return "MA147LL/A", and would make sure that "IW-02" is not in the result set.

If documents to be elevated are not defined in the elevate.xml file, they should be passed in at query time with the elevateIds [parameter](#).

## Using the Query Elevation Component

### The enableElevation Parameter

For debugging it may be useful to see results with and without the elevated docs. To hide results, use enableElevation=false:

```
http://localhost:8983/solr/techproducts/elevate?q=ipod&df=text&debugQuery=true&enableElevation=true
```

```
http://localhost:8983/solr/techproducts/elevate?q=ipod&df=text&debugQuery=true&enableElevation=false
```

## The forceElevation Parameter

You can force elevation during runtime by adding `forceElevation=true` to the query URL:

```
http://localhost:8983/solr/techproducts/elevate?q=ipod&df=text&debugQuery=true&enableElevation=true&forceElevation=true
```

## The exclusive Parameter

You can force Solr to return only the results specified in the elevation file by adding `exclusive=true` to the URL:

```
http://localhost:8983/solr/techproducts/elevate?q=ipod&df=text&debugQuery=true&exclusive=true
```

## The useConfiguredElevatedOrder Parameter

You can force set `useConfiguredElevatedOrder` during runtime by supplying it as a request parameter.

## Document Transformers and the markExcludes Parameter

The [elevated] Document Transformer can be used to annotate each document with information about whether or not it was elevated:

```
http://localhost:8983/solr/techproducts/elevate?q=ipod&df=text&fl=id,[elevated]
```

Likewise, it can be helpful when troubleshooting to see all matching documents – including documents that the elevation configuration would normally exclude. This is possible by using the `markExcludes=true` parameter, and then using the [excluded] transformer:

```
http://localhost:8983/solr/techproducts/elevate?q=ipod&df=text&markExcludes=true&fl=id,[elevated],[excluded]
```

## The elevateIds and excludeIds Parameters

When the elevation component is in use, the pre-configured list of elevations for a query can be overridden at request time to use the unique keys specified in these request parameters.

For example, in the request below documents 3007WFP and 9885A004 will be elevated, and document IW-02

will be excluded — regardless of what elevations or exclusions are configured for the query "cable" in `elevate.xml`:

```
http://localhost:8983/solr/techproducts/elevate?q=cable&df=text&excludeIds=IW-02&elevateIds=3007WFP,9885A004
```

If either one of these parameters is specified at request time, the entire elevation configuration for the query is ignored.

For example, in the request below documents `IW-02` and `F8V7067-APL-KIT` will be elevated, and no documents will be excluded — regardless of what elevations or exclusions are configured for the query "ipod" in `elevate.xml`:

```
http://localhost:8983/solr/techproducts/elevate?q=ipod&df=text&elevateIds=IW-02,F8V7067-APL-KIT
```

## The `fq` Parameter with Elevation

Query elevation respects the standard filter query (`fq`) parameter. That is, if the query contains the `fq` parameter, all results will be within that filter even if `elevate.xml` adds other documents to the result set.

# The Tagger Handler

The "Tagger" Request Handler, AKA the "SolrTextTagger" is a "text tagger".

Given a dictionary (a Solr index) with a name-like field, you can post text to this request handler and it will return every occurrence of one of those names with offsets and other document metadata desired. It's used for named entity recognition (NER).

The tagger doesn't do any natural language processing (NLP) (outside of Lucene text analysis) so it's considered a "naive tagger", but it's definitely useful as-is and a more complete NER or ERD (entity recognition and disambiguation) system can be built with this as a key component. The SolrTextTagger might be used on queries for query-understanding or large documents as well.

To get a sense of how to use it, jump to the [tutorial](#) below.

The tagger does not yet support a sharded index. Tens, perhaps hundreds of millions of names (documents) are supported, mostly limited by memory.

## Tagger Configuration

To configure the tagger, your Solr schema needs 2 fields:

- A unique key field (see [Unique Key](#) for how to define a unique key in your schema). Recommended field settings: set docValues=true.
- A tag field, which must be a TextField, with ConcatenateGraphFilterFactory at the end of the index chain (not the query chain): Set preservePositionIncrements=false on that filter. Recommended field settings: omitNorms=true, omitTermFreqAndPositions=true and postingsFormat=FST50.

The text field's *index analysis chain*, aside from needing ConcatenateGraphFilterFactory at the end, can otherwise have whatever tokenizer and filters suit your matching preferences. It can have multi-word synonyms and use WordDelimiterGraphFilterFactory for example. However, do *not* use FlattenGraphFilterFactory as it will interfere with ConcatenateGraphFilterFactory. Position gaps (e.g., stop words) get ignored; it's not (yet) supported for the gap to be significant.

The text field's *query analysis chain*, on the other hand, is more limited. There should not be tokens at the same position, thus no synonym expansion — do that at index time instead. Stop words (or any other filter introducing a position gap) are supported. At runtime the tagger can be configured to either treat it as a tag break or to ignore it.

Your solrconfig.xml needs the solr.TagRequestHandler defined, which supports defaults, invariants, and appends sections just like the search handler.

For configuration examples, jump to the [tutorial](#) below.

## Tagger Parameters

The tagger's execution is completely configurable with request parameters. Only field is required.

field

The tag field that serves as the dictionary. This is required; you'll probably specify it in the request

handler.

#### fq

You can specify some number of *filter queries* to limit the dictionary used for tagging. This parameter is the same one used by the solr.SearchHandler.

#### rows

The maximum number of documents to return, but defaulting to 10000 for a tag request. This parameter is the same as is used by the solr.SearchHandler.

#### f1

Solr's standard parameter for listing the fields to return. This parameter is the same one used by the solr.SearchHandler.

#### overlaps

Choose the algorithm to determine which tags in an overlapping set should be retained, versus being pruned away. Options are:

- ALL: Emit all tags.
- NO\_SUB: Don't emit a tag that is completely within another tag (i.e., no subtag).
- LONGEST\_DOMINANT\_RIGHT: Given a cluster of overlapping tags, emit the longest one (by character length). If there is a tie, pick the right-most. Remove any tags overlapping with this tag then repeat the algorithm to potentially find other tags that can be emitted in the cluster.

#### matchText

A boolean indicating whether to return the matched text in the tag response. This will trigger the tagger to fully buffer the input before tagging.

#### tagsLimit

The maximum number of tags to return in the response. Tagging effectively stops after this point. By default this is 1000.

#### skipAltTokens

A boolean flag used to suppress errors that can occur if, for example, you enable synonym expansion at query time in the analyzer, which you normally shouldn't do. Let this default to false unless you know that such tokens can't be avoided.

#### ignoreStopwords

A boolean flag that causes stopwords (or any condition causing positions to skip like >255 char words) to be ignored as if they aren't there. Otherwise, the behavior is to treat them as breaks in tagging on the presumption your indexed text-analysis configuration doesn't have a StopWordFilter defined. By default the indexed analysis chain is checked for the presence of a StopWordFilter and if found then ignoreStopWords is true if unspecified. You probably shouldn't have a StopWordFilter configured and probably won't need to set this parameter either.

#### xmlOffsetAdjust

A boolean indicating that the input is XML and furthermore that the offsets of returned tags should be adjusted as necessary to allow for the client to insert an opening and closing element at the tag offset pair. If it isn't possible to do so then the tag will be omitted. You are expected to configure HTMLStripCharFilterFactory in the schema when using this option. This will trigger the tagger to fully

buffer the input before tagging.

Solr's parameters for controlling the response format are also supported, such as echoParams, wt, indent, etc.

## Tutorial with Geonames

This is a tutorial that demonstrates how to configure and use the text tagger with the popular [Geonames](#) data set. It's more than a tutorial; it's a how-to with information that wasn't described above.

### Create and Configure a Solr Collection

Create a Solr collection named "geonames". For the tutorial, we'll assume the default "data-driven" configuration. It's good for experimentation and getting going fast but not for production or being optimal.

```
bin/solr create -c geonames
```

### Configuring the Tagger

We need to configure the schema first. The "data driven" mode we're using allows us to keep this step fairly minimal — we just need to declare a field type, 2 fields, and a copy-field.

The critical part up-front is to define the "tag" field type. There are many many ways to configure text analysis; and we're not going to get into those choices here. But an important bit is the `ConcatenateGraphFilterFactory` at the end of the index analyzer chain. Another important bit for performance is `postingsFormat=FST50` resulting in a compact FST based in-memory data structure that is especially beneficial for the text tagger.

Schema configuration:

```
curl -X POST -H 'Content-type:application/json' http://localhost:8983/solr/geonames/schema -d '{
 "add-field-type": {
 "name": "tag",
 "class": "solr.TextField",
 "postingsFormat": "FST50",
 "omitNorms": true,
 "omitTermFreqAndPositions": true,
 "indexAnalyzer": {
 "tokenizer": {
 "class": "solr.StandardTokenizerFactory" },
 "filters": [
 {"class": "solr.EnglishPossessiveFilterFactory"}, {"class": "solr.ASCII FoldingFilterFactory"}, {"class": "solr.LowerCaseFilterFactory"}, {"class": "solr.ConcatenateGraphFilterFactory", "preservePositionIncrements": false }
]},
 "queryAnalyzer": {
 "tokenizer": {
 "class": "solr.StandardTokenizerFactory" },
 "filters": [
 {"class": "solr.EnglishPossessiveFilterFactory"}, {"class": "solr.ASCII FoldingFilterFactory"}, {"class": "solr.LowerCaseFilterFactory"}
]}
 },
 "add-field": {"name": "name", "type": "text_general"},

 "add-field": {"name": "name_tag", "type": "tag", "stored": false },

 "add-copy-field": {"source": "name", "dest": ["name_tag"]}
 }
}'
```

Configure a custom Solr Request Handler:

```
curl -X POST -H 'Content-type:application/json' http://localhost:8983/solr/geonames/config -d '{
 "add-requesthandler" : {
 "name": "/tag",
 "class": "solr.TaggerRequestHandler",
 "defaults": {"field": "name_tag"}
 }
}'
```

## Load Some Sample Data

We'll go with some Geonames.org data in CSV format. Solr is quite flexible in loading data in a variety of formats. This [cities1000.zip](#) should be almost 7MB file expanding to a cities1000.txt file around 22.2MB containing 145k lines, each a city in the world of at least 1000 population.

Using bin/post:

```
bin/post -c geonames -type text/csv \
 -params
'optimize=true&maxSegments=1&separator=%09&encapsulator=%00&fieldnames=id,name,,alternative_names
,latitude,longitude,,,countrycode,,,,,population,elevation,,timezone,lastupdate' \
/tmp/cities1000.txt
```

or using curl:

```
curl -X POST --data-binary @/path/to/cities1000.txt -H 'Content-type:application/csv' \
'http://localhost:8983/solr/geonames/update?commit=true&optimize=true&maxSegments=1&separator=%09
&encapsulator=%00&fieldnames=id,name,,alternative_names,latitude,longitude,,,countrycode,,,,,
population,elevation,,timezone,lastupdate'
```

That might take around 35 seconds; it depends. It can be a lot faster if the schema were tuned to only have what we truly need (no text search if not needed).

In that command we said `optimize=true&maxSegments=1` to put the index in a state that will make tagging faster. The `encapsulator=%00` is a bit of a hack to disable the default double-quote.

## Tag Time!

This is a trivial example tagging a small piece of text. For more options, see the earlier documentation.

```
curl -X POST \
'http://localhost:8983/solr/geonames/tag?overlaps=NO_SUB&tagsLimit=5000&fl=id,name,countrycode&wt
=json&indent=on' \
-H 'Content-Type:text/plain' -d 'Hello New York City'
```

The response should be this (the QTime may vary):

```
{
 "responseHeader":{
 "status":0,
 "QTime":1},
 "tagsCount":1,
 "tags":[[
 "startOffset",6,
 "endOffset",19,
 "ids",["5128581"]]],
 "response":{ "numFound":1, "start":0, "docs": [
 {
 "id": "5128581",
 "name": ["New York City"],
 "countrycode": ["US"] }]
]}
}
```

## Tagger Tips

Performance Tips:

- Follow the recommended configuration field settings, especially `postingsFormat=FST50`.
- "optimize" after loading your dictionary down to 1 Lucene segment, or at least to as few as possible.
- For bulk tagging lots of documents, there are some strategies, not mutually exclusive:
  - Batch them. The tagger doesn't directly support batching but as a hack you can send a bunch of documents concatenated with a nonsense word that is not in the dictionary like "ZZYYXXAABBCC" between them. You'll need to keep track of the character offsets of these so you can subtract them from the results.
  - For reducing tagging latency even further, consider embedding Solr with `EmbeddedSolrServer`. See `EmbeddedSolrNoSerializeTest`.
  - Use more than one thread — perhaps as many as there are CPU cores available to Solr.

# Response Writers

A Response Writer generates the formatted response of a search.

Solr supports a variety of Response Writers to ensure that query responses can be parsed by the appropriate language or application.

The `wt` parameter selects the Response Writer to be used. The list below describe shows the most common settings for the `wt` parameter, with links to further sections that discuss them in more detail.

- [csv](#)
- [geojson](#)
- [javabin](#)
- [json](#)
- [php](#)
- [phps](#)
- [python](#)
- [ruby](#)
- [smile](#)
- [velocity](#)
- [xlsx](#)
- [xml](#)
- [xslt](#)

## JSON Response Writer

The default Solr Response Writer is the `JsonResponseWriter`, which formats output in JavaScript Object Notation (JSON), a lightweight data interchange format specified in RFC 4627. The default response writer is used when:

- the `wt` parameter is not specified in the request, or
- a non-existent response writer is specified.

Here is a sample response for a simple query like `q=id:VS1GB400C3`:

```
{
 "responseHeader": {
 "zkConnected": true,
 "status": 0,
 "QTime": 7,
 "params": {
 "q": "id:VS1GB400C3" }
 },
 "response": {
 "numFound": 1,
 "start": 0,
 "maxScore": 2.3025851,
 "docs": [
 {
 "id": "VS1GB400C3",
 "name": ["CORSAIR ValueSelect 1GB 184-Pin DDR SDRAM Unbuffered DDR 400 (PC 3200) System Memory - Retail"],
 "manu": ["Corsair Microsystems Inc."],
 "manu_id_s": "corsair",
 "cat": ["electronics", "memory"],
 "price": [74.99],
 "popularity": [7],
 "inStock": [true],
 "store": ["37.7752,-100.0232"],
 "manufacturedate_dt": "2006-02-13T15:26:37Z",
 "payloads": ["electronics|4.0 memory|2.0"],
 "_version_": 1549728120626479104
 }
]
 }
}
```

The default mime type for the JSON writer is application/json, however this can be overridden in the solrconfig.xml - such as in this example from the “techproducts” configuration:

```
<queryResponseWriter name="json" class="solr.JSONResponseWriter">
 <!-- For the purposes of the tutorial, JSON response are written as
 plain text so that it's easy to read in *any* browser.
 If you are building applications that consume JSON, just remove
 this override to get the default "application/json" mime type.
 -->
 <str name="content-type">text/plain</str>
</queryResponseWriter>
```

## JSON-Specific Parameters

### json.nl

This parameter controls the output format of NamedLists, where order is more important than access by name. NamedList is currently used for field faceting data.

The json.nl parameter takes the following values:

`flat`

The default. NamedList is represented as a flat array, alternating names and values.

With input of `NamedList("a"=1, "bar"="foo", null=3, null=null)`, the output would be `["a", 1,`

```
"bar", "foo", null, 3, null, null].
```

#### map

NamedList is represented as a JSON object. Although this is the simplest mapping, a NamedList can have optional keys, repeated keys, and preserves order. Using a JSON object (essentially a map or hash) for a NamedList results in the loss of some information.

With input of `NamedList("a"=1, "bar"="foo", null=3, null=null)`, the output would be `{"a":1, "bar":"foo", "":3, "":null}`.

#### arrarr

NamedList is represented as an array of two element arrays.

With input of `NamedList("a"=1, "bar"="foo", null=3, null=null)`, the output would be `[["a",1], ["bar", "foo"], [null,3], [null,null]]`.

#### arrmap

NamedList is represented as an array of JSON objects.

With input of `NamedList("a"=1, "bar"="foo", null=3, null=null)`, the output would be `[{"a":1}, {"b":2}, 3, null]`.

#### arrntv

NamedList is represented as an array of Name Type Value JSON objects.

With input of `NamedList("a"=1, "bar"="foo", null=3, null=null)`, the output would be `[{"name": "a", "type": "int", "value": 1}, {"name": "bar", "type": "str", "value": "foo"}, {"name": null, "type": "int", "value": 3}, {"name": null, "type": "null", "value": null}]`.

## json.wrf

`json.wrf=function` adds a wrapper-function around the JSON response, useful in AJAX with dynamic script tags for specifying a JavaScript callback function.

- <http://www.xml.com/pub/a/2005/12/21/json-dynamic-script-tag.html>
- <http://www.theurer.cc/blog/2005/12/15/web-services-json-dump-your-proxy/>

## Standard XML Response Writer

The XML Response Writer is the most general purpose and reusable Response Writer currently included with Solr. It is the format used in most discussions and documentation about the response of Solr queries.

Note that the XSLT Response Writer can be used to convert the XML produced by this writer to other vocabularies or text-based formats.

The behavior of the XML Response Writer can be driven by the following query parameters.

#### version

The version parameter determines the XML protocol used in the response. Clients are strongly encouraged to *always* specify the protocol version, so as to ensure that the format of the response they receive does not change unexpectedly if the Solr server is upgraded and a new default format is introduced.

The only currently supported version value is 2.2. The format of the responseHeader changed to use the same <lst> structure as the rest of the response.

The default value is the latest supported.

#### stylesheet

The stylesheet parameter can be used to direct Solr to include a <?xml-stylesheet type="text/xsl" href="..."?> declaration in the XML response it returns.

The default behavior is not to return any stylesheet declaration at all.



Use of the stylesheet parameter is discouraged, as there is currently no way to specify external stylesheets, and no stylesheets are provided in the Solr distributions. This is a legacy parameter, which may be developed further in a future release.

#### indent

If the indent parameter is used, and has a non-blank value, then Solr will make some attempts at indenting its XML response to make it more readable by humans.

The default behavior is not to indent.

## XSLT Response Writer

The XSLT Response Writer applies an XML stylesheet to output. It can be used for tasks such as formatting results for an RSS feed.

### tr Parameter

The XSLT Response Writer accepts one parameter: the tr parameter, which identifies the XML transformation to use. The transformation must be found in the Solr conf/xslt directory.

The Content-Type of the response is set according to the <xsl:output> statement in the XSLT transform, for example: <xsl:output media-type="text/html"/>

### XSLT Configuration

The example below, from the sample\_techproducts\_configs configset in the Solr distribution, shows how the XSLT Response Writer is configured.

```
<!--
Changes to XSLT transforms are taken into account
every xsltCacheLifetimeSeconds at most.

-->
<queryResponseWriter name="xslt"
 class="org.apache.solr.request.XSLTResponseWriter">
 <int name="xsltCacheLifetimeSeconds">5</int>
</queryResponseWriter>
```

A value of 5 for xsltCacheLifetimeSeconds is good for development, to see XSLT changes quickly. For production you probably want a much higher value.

## Binary Response Writer

This is a custom binary format used by Solr for inter-node communication as well as client-server communication. SolrJ uses this as the default for indexing as well as querying. See [Client APIs](#) for more details.

## GeoJSON Response Writer

Returns Solr results in [GeoJSON](#) augmented with Solr-specific JSON. To use this, set `wt=geojson` and `geojson.field` to the name of a spatial Solr field. Not all spatial fields types are supported, and you'll get an error if you use an unsupported one.

## Python Response Writer

Solr has an optional Python response format that extends its JSON output in the following ways to allow the response to be safely evaluated by the python interpreter:

- true and false changed to True and False
- Python unicode strings are used where needed
- ASCII output (with unicode escapes) is used for less error-prone interoperability
- newlines are escaped
- null changed to None

## PHP Response Writer and PHP Serialized Response Writer

Solr has a PHP response format that outputs an array (as PHP code) which can be evaluated. Setting the `wt` parameter to `php` invokes the PHP Response Writer.

Example usage:

```
$code = file_get_contents('http://localhost:8983/solr/techproducts/select?q=iPod&wt=php');
eval("$result = " . $code . ";");
print_r($result);
```

Solr also includes a PHP Serialized Response Writer that formats output in a serialized array. Setting the `wt` parameter to `phps` invokes the PHP Serialized Response Writer.

Example usage:

```
$serializedResult = file_get_contents(
'http://localhost:8983/solr/techproducts/select?q=iPod&wt=phps');
$result = unserialize($serializedResult);
print_r($result);
```

## Ruby Response Writer

Solr has an optional Ruby response format that extends its JSON output in the following ways to allow the response to be safely evaluated by Ruby's interpreter:

- Ruby's single quoted strings are used to prevent possible string exploits.
- \ and ' are the only two characters escaped.
- Unicode escapes are not used. Data is written as raw UTF-8.
- nil used for null.
- => is used as the key/value separator in maps.

Here is a simple example of how one may query Solr using the Ruby response format:

```
require 'net/http'
h = Net::HTTP.new('localhost', 8983)
hresp, data = h.get('/solr/techproducts/select?q=iPod&wt=ruby', nil)
rsp = eval(data)
puts 'number of matches = ' + rsp['response']['numFound'].to_s
#print out the name field for each returned document
rsp['response']['docs'].each { |doc| puts 'name field = ' + doc['name'] }
```

## CSV Response Writer

The CSV response writer returns a list of documents in comma-separated values (CSV) format. Other information that would normally be included in a response, such as facet information, is excluded.

The CSV response writer supports multi-valued fields, as well as [pseudo-fields](#), and the output of this CSV format is compatible with Solr's [CSV update format](#).

### CSV Parameters

These parameters specify the CSV format that will be returned. You can accept the default values or specify your own.

Parameter	Default Value
csv.encapsulator	"
csv.escape	None
csv.separator	,
csv.header	Defaults to true. If false, Solr does not print the column headers.
csv.newline	\n

Parameter	Default Value
csv.null	Defaults to a zero length string. Use this parameter when a document has no value for a particular field.

## Multi-Valued Field CSV Parameters

These parameters specify how multi-valued fields are encoded. Per-field overrides for these values can be done using f.<fieldname>.csv.separator=|.

Parameter	Default Value
csv.mv.encapsulator	None
csv.mv.escape	\
csv.mv.separator	Defaults to the csv.separator value.

## CSV Writer Example

<http://localhost:8983/solr/techproducts/select?q=ipod&fl=id,cat,name,popularity,price,score&wt=csv> returns:

```
id,cat,name,popularity,price,score
IW-02,"electronics,connector",iPod & iPod Mini USB 2.0 Cable,1,11.5,0.98867977
F8V7067-APL-KIT,"electronics,connector",Belkin Mobile Power Cord for iPod w/
Dock,1,19.95,0.6523595
MA147LL/A,"electronics,music",Apple 60 GB iPod with Video Playback Black,10,399.0,0.2446348
```

## Velocity Response Writer

The VelocityResponseWriter processes the Solr response and request context through Apache Velocity templating.

See the [Velocity Response Writer](#) section for details.

## Smile Response Writer

The Smile format is a JSON-compatible binary format, described in detail here: <http://wiki.fasterxml.com/SmileFormat>.

## XLSX Response Writer

Use this to get the response as a spreadsheet in the .xlsx (Microsoft Excel) format. It accepts parameters in the form colwidth.<field-name> and colname.<field-name> which helps you customize the column

widths and column names.

This response writer has been added as part of the extraction library, and will only work if the extraction contrib is present in the server classpath. Defining the classpath with the lib directive is not sufficient. Instead, you will need to copy the necessary .jars to the Solr webapp's lib directory manually. You can run these commands from your \$SOLR\_INSTALL directory:

```
cp contrib/extraction/lib/*.jar server/solr-webapp/webapp/WEB-INF/lib/
cp dist/solr-cell-6.3.0.jar server/solr-webapp/webapp/WEB-INF/lib/
```

Once the libraries are in place, you can add wt=xlsx to your request, and results will be returned as an XLSX sheet.

## Velocity Response Writer

The VelocityResponseWriter is an optional plugin available in the contrib/velocity directory. It powers the /browse user interfaces when using configurations such as "\_default", "techproducts", and "example/files".

Its JAR and dependencies must be added (via <lib> or solr/home lib inclusion), and must be registered in solrconfig.xml like this:

```
<queryResponseWriter name="velocity" class="solr.VelocityResponseWriter">
 <str name="template.base.dir">${velocity.template.base.dir:}</str>

<!--
 <str name="init.properties.file">velocity-init.properties</str>
 <bool name="params.resource.loader.enabled">true</bool>
 <bool name="solr.resource.loader.enabled">false</bool>
 <lst name="tools">
 <str name="mytool">com.example.MyCustomTool</str>
 </lst>
-->
</queryResponseWriter>
```

The above example shows the optional initialization and custom tool parameters used by VelocityResponseWriter; these are detailed in the following table. These initialization parameters are only specified in the writer registration in solrconfig.xml, not as request-time parameters. See further below for request-time parameters.

## Configuration & Usage

### VelocityResponseWriter Initialization Parameters

#### template.base.dir

If specified and exists as a file system directory, a file resource loader will be added for this directory. Templates in this directory will override "solr" resource loader templates.

#### init.properties.file

Specifies a properties file name which must exist in the Solr conf/ directory (**not** under a velocity/

subdirectory) or root of a JAR file in a <lib>.

`params.resource.loader.enabled`

The "params" resource loader allows templates to be specified in Solr request parameters. For example:

```
http://localhost:8983/solr/gettingstarted/select?q=*&wt=velocity&v.template=custom&v.template.custom=CUSTOM%3A%20%23core_name
```

where `v.template=custom` says to render a template called "custom" and the value of `v.template.custom` is the custom template. This is false by default; it'd be a niche, unusual, use case to need this enabled.

`solr.resource.loader.enabled`

The "solr" resource loader is the only template loader registered by default. Templates are served from resources visible to the SolrResourceLoader under a `velocity`/ subdirectory. The VelocityResponseWriter itself has some built-in templates (in its JAR file, under `velocity/`) that are available automatically through this loader. These built-in templates can be overridden when the same template name is in `conf/velocity/` or by using the `template.base.dir` option.

`tools`

External "tools" can be specified as list of string name/value (tool name / class name) pairs. Tools, in the Velocity context, are simply Java objects. Tool classes are constructed using a no-arg constructor (or a single-SolrCore-arg constructor if it exists) and added to the Velocity context with the specified name.

A custom registered tool can override the built-in context objects with the same name, except for `$request`, `$response`, `$page`, and `$debug` (these tools are designed to not be overridden).

## VelocityResponseWriter Request Parameters

`v.template`

Specifies the name of the template to render.

`v.layout`

Specifies a template name to use as the layout around the main, `v.template`, specified template.

The main template is rendered into a string value included into the layout rendering as `$content`.

`v.layout.enabled`

Determines if the main template should have a layout wrapped around it. The default is `true`, but requires `v.layout` to be specified as well.

`v.contentType`

Specifies the content type used in the HTTP response. If not specified, the default will depend on whether `v.json` is specified or not.

The default without `v.json=wrf`: `text/html;charset=UTF-8`.

The default with `v.json=wrf`: `application/json;charset=UTF-8`.

`v.json`

Specifies a function name to wrap around the response rendered as JSON. If specified, the content type

used in the response will be "application/json; charset=UTF-8", unless overridden by v.contentType.

Output will be in this format (with v.json=wrf):

```
wrf("result": "<Velocity generated response string, with quotes and backslashes escaped>")
```

#### v.locale

Locale to use with the \$resource tool and other LocaleConfig implementing tools. The default locale is Locale.ROOT. Localized resources are loaded from standard Java resource bundles named resources[\_locale-code].properties.

Resource bundles can be added by providing a JAR file visible by the SolrResourceLoader with resource bundles under a velocity sub-directory. Resource bundles are not loadable under conf/, as only the class loader aspect of SolrResourceLoader can be used here.

#### v.template.template\_name

When the "params" resource loader is enabled, templates can be specified as part of the Solr request.

## VelocityResponseWriter Context Objects

Context Reference	Description
request	<a href="#">SolrQueryRequest</a> javadocs
response	<a href="#">QueryResponse</a> most of the time, but in some cases where QueryResponse doesn't like the request handler's output ( <a href="#">AnalysisRequestHandler</a> , for example, causes a ClassCastException parsing "response"), the response will be a SolrResponseBase object.
esc	A Velocity <a href="#">EscapeTool</a> instance
date	A Velocity <a href="#">ComparisonDateTool</a> instance
math	A Velocity <a href="#">MathTool</a> instance
number	A Velocity <a href="#">NumberTool</a> instance
sort	A Velocity <a href="#">SortTool</a> instance
display	A Velocity <a href="#">DisplayTool</a> instance
resource	A Velocity <a href="#">ResourceTool</a> instance
engine	The current VelocityEngine instance
page	An instance of Solr's PageTool (only included if the response is a QueryResponse where paging makes sense)
debug	A shortcut to the debug part of the response, or null if debug is not on. This is handy for having debug-only sections in a template using #if(\$debug)...#end
content	The rendered output of the main template, when rendering the layout (v.layout.enabled=true and v.layout=<template>).

Context Reference	Description
[custom tool(s)]	Tools provided by the optional "tools" list of the VelocityResponseWriter registration are available by their specified name.

# Near Real Time Searching

Near Real Time (NRT) search means that documents are available for search soon after being indexed. NRT searching is one of the main features of SolrCloud and is rarely attempted in master/slave configurations.

Document durability and searchability are controlled by commits. The "Near" in "Near Real Time" is configurable to meet the needs of your application. Commits are either "hard" or "soft" and can be issued by a client (say SolrJ), via a REST call or configured to occur automatically in `solrconfig.xml`. The recommendation usually gives is to configure your commit strategy in `solrconfig.xml` (see below) and avoid issuing commits externally.

Typically in NRT applications, hard commits are configured with `openSearcher=false`, and soft commits are configured to make documents visible for search.

When a commit occurs, various background tasks are initiated, segment merging for example. These background tasks do not block additional updates to the index nor do they delay the availability of the documents for search.

When configuring for NRT, pay special attention to cache and autowarm settings as they can have a significant impact on NRT performance. For extremely short autoCommit intervals, consider disabling caching and autowarming completely.

## Commits and Searching

A **hard commit** calls `fsync` on the index files to ensure they have been flushed to stable storage. The current transaction log is closed and a new one is opened. See the "transaction log" discussion below for how data is recovered in the absence of a hard commit. Optionally a hard commit can also make documents visible for search, but this is not recommended for NRT searching as it is more expensive than a soft commit.

A **soft commit** is faster since it only makes index changes visible and does not `fsync` index files, start a new segment or start a new transaction log. Search collections that have NRT requirements will want to soft commit often enough to satisfy the visibility requirements of the application. A softCommit may be "less expensive" than a hard commit (`openSearcher=true`), but it is not free. It is recommended that this be set for as long as is reasonable given the application requirements.

Both hard and soft commits have two primary configuration parameters: `maxDocs` and `maxTime`.

`maxDocs`

Integer. Defines the number of updates to process before activating.

`maxTime`

Integer. The number of milliseconds to wait before activating.

If both of these parameters are specified, the first one to expire is honored. Generally, it is preferred to use `maxTime` rather than `maxDocs`, especially when indexing large numbers of documents in batches. Use `maxDocs` and `maxTime` judiciously to fine-tune your commit strategies.

Hard commit has an additional parameter `openSearcher`

`openSearcher`

true|false, whether to make documents visible for search. For NRT applications this is usually set to false and soft commit is configured to control when documents are visible for search.

## Transaction Logs (tlogs)

Transaction logs are a "rolling window" of updates since the last hard commit. The current transaction log is closed and a new one opened each time any variety of hard commit occurs. Soft commits have no effect on the transaction log.

When tlogs are enabled, documents being added to the index are written to the tlog before the indexing call returns to the client. In the event of an un-graceful shutdown (power loss, JVM crash, kill -9, etc.) any documents written to the tlog but not yet committed with a hard commit when Solr was stopped are replayed on startup. Therefore the data is not lost.

When Solr is shut down gracefully (using the bin/solr stop command) Solr will close the tlog file and index segments so no replay will be necessary on startup.

One point of confusion is how much data is contained in a transaction log. A tlog does not contain all documents, only the ones since the last hard commit. Older transaction log files are deleted when no longer needed.



Implicit in the above is that transaction logs will grow forever if hard commits are disabled. Therefore it is important that hard commits be enabled when indexing.

## Configuring Commits

As mentioned above, it is usually preferable to configure your commits (both hard and soft) in solrconfig.xml and avoid sending commits from an external source. Check your solrconfig.xml file since the defaults are likely not tuned to your needs. Here is an example NRT configuration for the two flavors of commit, a hard commit every 60 seconds and a soft commit every 30 seconds. Note that these are *not* the values in some of the examples!

```
<autoCommit>
 <maxTime>${solr.autoCommit.maxTime:60000}</maxTime>
 <openSearcher>false</openSearcher>
</autoCommit>

<autoSoftCommit>
 <maxTime>${solr.autoSoftCommit.maxTime:30000}</maxTime>
</autoSoftCommit>
```



These parameters can be overridden at run time by defining Java "system variables", for example specifying '-Dsolr.autoCommit.maxTime=15000' would override the hard commit interval with a value of 15 seconds.

The choices for autoCommit (with openSearcher=false) and autoSoftCommit have different consequences. In the event of un-graceful shutdown, it can take up to the time specified in autoCommit for Solr to replay the uncommitted documents from the transaction log.

The time chosen for autoSoftCommit determines the maximum time after a document is sent to Solr before it becomes searchable and does not affect the transaction log. Choose as long an interval as your application can tolerate for this value, often 15-60 seconds is reasonable, or even longer depending on the requirements. In situations where the the time is set to a very short interval (say 1 second), consider disabling your caches (queryResultCache and filterCache especially) as they will have little utility.



For extremely high bulk indexing, especially for the initial load if there is no searching, consider turning off autoSoftCommit by specifying a value of -1 for the maxTime parameter.

## Advanced Commit Options

All varieties of commits can be invoked from a SolrJ client or via a URL. The usual recommendation is to *not* call commits externally. For those cases where it is desirable, see [Update Commands](#). These options are listed for XML update commands that can be issued from a browser or curl, etc., and the equivalents are available from a SolrJ client.

## RealTime Get

For index updates to be visible (searchable), some kind of commit must reopen a searcher to a new point-in-time view of the index.

The **realtime get** feature allows retrieval (by unique-key) of the latest version of any documents without the associated cost of reopening a searcher. This is primarily useful when using Solr as a NoSQL data store and not just a search index.

Real Time Get relies on the update log feature, which is enabled by default and can be configured in `solrconfig.xml`:

```
<updateLog>
 <str name="dir">${solr.ulog.dir:}</str>
</updateLog>
```

Real Time Get requests can be performed using the `/get` handler which exists implicitly in Solr - see [Implicit RequestHandlers](#) - it's equivalent to the following configuration:

```
<requestHandler name="/get" class="solr.RealTimeGetHandler">
 <lst name="defaults">
 <str name="omitHeader">true</str>
 </lst>
</requestHandler>
```

For example, if you started Solr using the `bin/solr -e techproducts` example command, you could then index a new document without committing it, like so:

```
curl 'http://localhost:8983/solr/techproducts/update/json?commitWithin=10000000' \
-H 'Content-type:application/json' -d '[{"id":"mydoc","name":"realtime-get test!"}]'
```

If you search for this document, it should not be found yet:

```
http://localhost:8983/solr/techproducts/query?q=id:mydoc
```

```
{"response":
 {"numFound":0, "start":0, "docs":[]}
}
```

However if you use the Real Time Get handler exposed at `/get`, you can retrieve the document:

## V1 API

```
http://localhost:8983/solr/techproducts/get?id=mydoc
```

```
{"doc": {
 "id": "mydoc",
 "name": "realtime-get test!",
 "_version_": 1487137811571146752
}
}
```

## V2 API

```
http://localhost:8983/api/collections/techproducts/get?id=mydoc
```

```
{"doc": {
 "id": "mydoc",
 "name": "realtime-get test!",
 "_version_": 1487137811571146752
}
}
```

You can also specify multiple documents at once via the `ids` parameter and a comma separated list of ids, or by using multiple `id` parameters. If you specify multiple ids, or use the `ids` parameter, the response will mimic a normal query response to make it easier for existing clients to parse.

For example:

## V1 API

```
http://localhost:8983/solr/techproducts/get?ids=mydoc,IW-02
http://localhost:8983/solr/techproducts/get?id=mydoc&id=IW-02
```

```
{"response":
 {"numFound":2,"start":0,"docs":
 [{ "id":"mydoc",
 "name":"realtime-get test!",
 "_version_":1487137811571146752},
 {
 "id":"IW-02",
 "name":"iPod & iPod Mini USB 2.0 Cable"
 }
]
 }
}
```

## V2 API

```
http://localhost:8983/api/collections/techproducts/get?ids=mydoc,IW-02
http://localhost:8983/api/collections/techproducts/get?id=mydoc&id=IW-02
```

```
{"response":
 {"numFound":2,"start":0,"docs":
 [{ "id":"mydoc",
 "name":"realtime-get test!",
 "_version_":1487137811571146752},
 {
 "id":"IW-02",
 "name":"iPod & iPod Mini USB 2.0 Cable"
 }
]
 }
}
```

Real Time Get requests can also be combined with filter queries, specified with an `fq` parameter:

## V1 API

```
http://localhost:8983/solr/techproducts/get?id=mydoc&id=IW-02&fq=name:realtime-get
```

```
{"response":
 {"numFound":1,"start":0,"docs":
 [{ "id": "mydoc",
 "name": "realtime-get test!",
 "_version_":1487137811571146752}
]
 }
}
```

## V2 API

```
http://localhost:8983/api/collections/techproducts/get?id=mydoc&id=IW-02&fq=name:realtime-get
```

```
{"response":
 {"numFound":1,"start":0,"docs":
 [{ "id": "mydoc",
 "name": "realtime-get test!",
 "_version_":1487137811571146752}
]
 }
}
```

Do **NOT** disable the realtime get handler at /get if you are using SolrCloud. Doing so will result in any leader election to cause a full sync in **ALL** replicas for the shard in question.



Similarly, a replica recovery will also always fetch the complete index from the leader because a partial sync will not be possible in the absence of this handler.

# Exporting Result Sets

It's possible to export fully sorted result sets using a special [rank query parser](#) and [response writer](#) specifically designed to work together to handle scenarios that involve sorting and exporting millions of records.

This feature uses a stream sorting technique that begins to send records within milliseconds and continues to stream results until the entire result set has been sorted and exported.

The cases where this functionality may be useful include: session analysis, distributed merge joins, time series roll-ups, aggregations on high cardinality fields, fully distributed field collapsing, and sort based stats.

## Field Requirements

All the fields being sorted and exported must have docValues set to true. For more information, see the section on [DocValues](#).

## The /export RequestHandler

The /export request handler with the appropriate configuration is one of Solr's out-of-the-box request handlers - see [Implicit RequestHandlers](#) for more information.

Note that this request handler's properties are defined as "invariants", which means they cannot be overridden by other properties passed at another time (such as at query time).

## Requesting Results Export

You can use /export to make requests to export the result set of a query.

All queries must include sort and fl parameters, or the query will return an error. Filter queries are also supported.

The supported response writers are json and javabin. For backward compatibility reasons wt=xsort is also supported as input, but wt=xsort behaves same as wt=json. The default output format is json.

Here is an example of an export request of some indexed log data:

```
http://localhost:8983/solr/core_name/export?q=my-
query&sort=severity+desc,timestamp+desc&fl=severity,timestamp,msg
```

## Specifying the Sort Criteria

The sort property defines how documents will be sorted in the exported result set. Results can be sorted by any field that has a field type of int, long, float, double, string. The sort fields must be single valued fields.

The export performance will get slower as you add more sort fields. If there is enough physical memory available outside of the JVM to load up the sort fields then the performance will be linearly slower with addition of sort fields. It can get worse otherwise.

## Specifying the Field List

The `f1` property defines the fields that will be exported with the result set. Any of the field types that can be sorted (i.e., int, long, float, double, string, date, boolean) can be used in the field list. The fields can be single or multi-valued. However, returning scores and wildcards are not supported at this time.

## Distributed Support

See the section [Streaming Expressions](#) for distributed support.

# Parallel SQL Interface

Solr's Parallel SQL Interface brings the power of SQL to SolrCloud.

The SQL interface seamlessly combines SQL with Solr's full-text search capabilities. Both MapReduce style and JSON Facet API aggregations are supported, which means the SQL interface can be used to support both **high query volume** and **high cardinality** use cases.

## SQL Architecture

The SQL interface allows sending a SQL query to Solr and getting documents streamed back in response. Under the covers, Solr's SQL interface uses the Apache Calcite SQL engine to translate SQL queries to physical query plans implemented as [Streaming Expressions](#).

## Solr Collections and DB Tables

In a standard SELECT statement such as `SELECT <expressions> FROM <table>`, the table names correspond to Solr collection names. Table names are case insensitive.

Column names in the SQL query map directly to fields in the Solr index for the collection being queried. These identifiers are case sensitive. Aliases are supported, and can be referenced in the ORDER BY clause.

The `*` syntax to indicate all fields is not supported in either limited or unlimited queries. The score field can be used only with queries that contain a LIMIT clause.

For example, we could index Solr's sample documents and then construct an SQL query like this:

```
SELECT manu as mfr, price as retail FROM techproducts
```

The collection in Solr we are using is "techproducts", and we've asked for the "manu" and "price" fields to be returned and aliased with new names. While this example does not use those aliases, we could build on this to ORDER BY one or more of those fields.

More information about how to structure SQL queries for Solr is included in the section [Solr SQL Syntax](#) below.

## Aggregation Modes

The SQL feature of Solr can work with aggregations (grouping of results) in two ways:

- facet: This is the **default** aggregation mode, which uses the JSON Facet API or StatsComponent for aggregations. In this scenario the aggregations logic is pushed down into the search engine and only the aggregates are sent across the network. This is Solr's normal mode of operation. This is fast when the cardinality of GROUP BY fields is low to moderate. But it breaks down when you have high cardinality fields in the GROUP BY field.
- map\_reduce: This implementation shuffles tuples to worker nodes and performs the aggregation on the worker nodes. It involves sorting and partitioning the entire result set and sending it to worker nodes. In this approach the tuples arrive at the worker nodes sorted by the GROUP BY fields. The worker nodes

can then rollup the aggregates one group at a time. This allows for unlimited cardinality aggregation, but you pay the price of sending the entire result set across the network to worker nodes.

These modes are defined with the aggregationMode property when sending the request to Solr.

As noted, the choice between aggregation modes depends on the cardinality of the fields you are working with. If you have low-to-moderate cardinality in the fields you are grouping by, the 'facet' aggregation mode will give you a higher performance because only the final groups are returned, very similar to how facets work today. If, however, you have high cardinality in the fields, the "map\_reduce" aggregation mode with worker nodes provide a much more performant option.

## Configuration

The request handlers used for the SQL interface are configured to load implicitly, meaning there is little to do to start using this feature.

### /sql Request Handler

The /sql handler is the front end of the Parallel SQL interface. All SQL queries are sent to the /sql handler to be processed. The handler also coordinates the distributed MapReduce jobs when running GROUP\_BY and SELECT DISTINCT queries in map\_reduce mode. By default the /sql handler will choose worker nodes from its own collection to handle the distributed operations. In this default scenario the collection where the /sql handler resides acts as the default worker collection for MapReduce queries.

By default, the /sql request handler is configured as an implicit handler, meaning that it is always enabled in every Solr installation and no further configuration is required.



As described below in the section [Best Practices](#), you may want to set up a separate collection for parallelized SQL queries. If you have high cardinality fields and a large amount of data, please be sure to review that section and consider using a separate collection.

### /stream and /export Request Handlers

The Streaming API is an extensible parallel computing framework for SolrCloud. [Streaming Expressions](#) provide a query language and a serialization format for the Streaming API.

The Streaming API provides support for fast MapReduce allowing it to perform parallel relational algebra on extremely large data sets. Under the covers the SQL interface parses SQL queries using the Apache Calcite SQL Parser. It then translates the queries to the parallel query plan. The parallel query plan is expressed using the Streaming API and Streaming Expressions.

Like the /sql request handler, the /stream and /export request handlers are configured as implicit handlers, and no further configuration is required.

## Fields

In some cases, fields used in SQL queries must be configured as DocValue fields. If queries are unlimited, all fields must be DocValue fields. If queries are limited (with the limit clause) then fields do not have to be have DocValues enabled.

## Sending Queries

The SQL Interface provides a basic JDBC driver and an HTTP interface to perform queries.

### JDBC Driver

The JDBC Driver ships with SolrJ. Below is sample code for creating a connection and executing a query with the JDBC driver:

```
Connection con = null;
try {
 con = DriverManager.getConnection("jdbc:solr://" + zkHost +
"?collection=collection1&aggregationMode=map_reduce&numWorkers=2");
 stmt = con.createStatement();
 rs = stmt.executeQuery("SELECT a_s, sum(a_f) as sum FROM collection1 GROUP BY a_s ORDER BY
sum desc");

 while(rs.next()) {
 String a_s = rs.getString("a_s");
 double s = rs.getDouble("sum");
 }
} finally {
 rs.close();
 stmt.close();
 con.close();
}
```

The connection URL must contain the zkHost and the collection parameters. The collection must be a valid SolrCloud collection at the specified ZooKeeper host. The collection must also be configured with the /sql handler. The aggregationMode and numWorkers parameters are optional.

### HTTP Interface

Solr accepts parallel SQL queries through the /sql handler.

Below is a sample curl command performing a SQL aggregate query in facet mode:

```
curl --data-urlencode 'stmt=SELECT to, count(*) FROM collection4 GROUP BY to ORDER BY count(*) desc LIMIT 10' http://localhost:8983/solr/collection4/sql?aggregationMode=facet
```

Below is sample result set:

```
{"result-set": {"docs": [
 {"count(*)": 9158, "to": "pete.davis@enron.com"},
 {"count(*)": 6244, "to": "tana.jones@enron.com"},
 {"count(*)": 5874, "to": "jeff.dasovich@enron.com"},
 {"count(*)": 5867, "to": "sara.shackleton@enron.com"},
 {"count(*)": 5595, "to": "steven.kean@enron.com"},
 {"count(*)": 4904, "to": "vkaminski@aol.com"},
 {"count(*)": 4622, "to": "mark.taylor@enron.com"},
 {"count(*)": 3819, "to": "kay.mann@enron.com"},
 {"count(*)": 3678, "to": "richard.shapiro@enron.com"},
 {"count(*)": 3653, "to": "kate.symes@enron.com"},
 {"EOF": "true", "RESPONSE_TIME": 10}]}
}
```

Notice that the result set is an array of tuples with key/value pairs that match the SQL column list. The final tuple contains the EOF flag which signals the end of the stream.

## Solr SQL Syntax

Solr supports a broad range of SQL syntax.

*SQL Parser is Case Insensitive*



The SQL parser being used by Solr to translate the SQL statements is case insensitive. However, for ease of reading, all examples on this page use capitalized keywords.

## Escaping Reserved Words

The SQL parser will return an error if a reserved word is used in the SQL query. Reserved words can be escaped and included in the query using the back tick. For example:

```
select `from` from emails
```

## SELECT Statements

Solr supports limited and unlimited select queries. The syntax between the two types of queries are identical except for the LIMIT clause in the SQL statement. However, they have very different execution plans and different requirements for how the data is stored. The sections below explores both types of queries.

### Basic SELECT statement with LIMIT

A limited select query follows this basic syntax:

```
SELECT fieldA as fa, fieldB as fb, fieldC as fc FROM tableA WHERE fieldC = 'term1 term2' ORDER BY fa desc LIMIT 100
```

We've covered many syntax options with this example, so let's walk through what's possible below.

## WHERE Clause and Boolean Predicates



The WHERE clause must have a field on one side of the predicate. Two constants (`5 < 10`) or two fields (`fielda > fieldb`) is not supported. Subqueries are also not supported.

The WHERE clause allows Solr's search syntax to be injected into the SQL query. In the example:

```
WHERE fieldC = 'term1 term2'
```

The predicate above will execute a full text search for the phrase 'term1 term2' in fieldC.

To execute a non-phrase query, simply add parens inside of the single quotes. For example:

```
WHERE fieldC = '(term1 term2)'
```

The predicate above searches for term1 OR term2 in fieldC.

The Solr range query syntax can be used as follows:

```
WHERE fieldC = '[0 TO 100]'
```

Complex boolean queries can be specified as follows:

```
WHERE ((fieldC = 'term1' AND fieldA = 'term2') OR (fieldB = 'term3'))
```

To specify NOT queries, you use the AND NOT syntax as follows:

```
WHERE (fieldA = 'term1') AND NOT (fieldB = 'term2')
```

### Supported WHERE Operators

The parallel SQL interface supports and pushes down most common SQL operators, specifically:

Operator	Description	Example	Solr Query
=	Equals	fielda = 10	fielda:10
<>	Does not equal	fielda <> 10	-fielda:10
!=	Does not equal	fielda != 10	-fielda:10
>	Greater than	fielda > 10	fielda:{10 TO *}
>=	Greater than or equals	fielda >= 10	fielda:[10 TO *]
<	Less than	fielda < 10	fielda:[* TO 10]
<=	Less than or equals	fielda <= 10	fielda:[* TO 10]

Some operators that are not supported are BETWEEN, LIKE and IN. However, there are workarounds for BETWEEN and LIKE.

- BETWEEN can be supported with a range query, such as field = [50 TO 100].
- A simplistic LIKE can be used with a wildcard, such as field = 'sam\*'.

## ORDER BY Clause

The ORDER BY clause maps directly to Solr fields. Multiple ORDER BY fields and directions are supported.

The score field is accepted in the ORDER BY clause in queries where a limit is specified.

If the ORDER BY clause contains the exact fields in the GROUP BY clause, then there is no-limit placed on the returned results. If the ORDER BY clause contains different fields than the GROUP BY clause, a limit of 100 is automatically applied. To increase this limit you must specify a value in the LIMIT clause.

Order by fields are case sensitive.

## LIMIT Clause

Limits the result set to the specified size. In the example above the clause LIMIT 100 will limit the result set to 100 records.

There are a few differences to note between limited and unlimited queries:

- Limited queries support score in the field list and ORDER BY. Unlimited queries do not.
- Limited queries allow any stored field in the field list. Unlimited queries require the fields to be stored as a DocValues field.
- Limited queries allow any indexed field in the ORDER BY list. Unlimited queries require the fields to be stored as a DocValues field.

## SELECT DISTINCT Queries

The SQL interface supports both MapReduce and Facet implementations for SELECT DISTINCT queries.

The MapReduce implementation shuffles tuples to worker nodes where the Distinct operation is performed. This implementation can perform the Distinct operation over extremely high cardinality fields.

The Facet implementation pushes down the Distinct operation into the search engine using the JSON Facet API. This implementation is designed for high performance, high QPS scenarios on low-to-moderate cardinality fields.

The aggregationMode parameter is available in the both the JDBC driver and HTTP interface to choose the underlying implementation (map\_reduce or facet). The SQL syntax is identical for both implementations:

```
SELECT distinct fieldA as fa, fieldB as fb FROM tableA ORDER BY fa desc, fb desc
```

## Statistical Functions

The SQL interface supports simple statistics calculated on numeric fields. The supported functions are count(\*), min, max, sum, and avg.

Because these functions never require data to be shuffled, the aggregations are pushed down into the search engine and are generated by the [StatsComponent](#).

```
SELECT count(*) as count, sum(fieldB) as sum FROM tableA WHERE fieldC = 'Hello'
```

## GROUP BY Aggregations

The SQL interface also supports GROUP BY aggregate queries.

As with SELECT DISTINCT queries, the SQL interface supports both a MapReduce implementation and a Facet implementation. The MapReduce implementation can build aggregations over extremely high cardinality fields. The Facet implementations provides high performance aggregation over fields with moderate levels of cardinality.

### Basic GROUP BY with Aggregates

Here is a basic example of a GROUP BY query that requests aggregations:

```
SELECT fieldA as fa, fieldB as fb, count(*) as count, sum(fieldC) as sum, avg(fieldY) as avg FROM tableA WHERE fieldC = 'term1 term2'
GROUP BY fa, fb HAVING sum > 1000 ORDER BY sum asc LIMIT 100
```

Let's break this down into pieces:

### Column Identifiers and Aliases

The Column Identifiers can contain both fields in the Solr index and aggregate functions. The supported aggregate functions are:

- count(\*) : Counts the number of records over a set of buckets.
- sum(field) : Sums a numeric field over over a set of buckets.
- avg(field) : Averages a numeric field over a set of buckets.
- min(field) : Returns the min value of a numeric field over a set of buckets.
- max:(field) : Returns the max value of a numerics over a set of buckets.

The non-function fields in the field list determine the fields to calculate the aggregations over.

### HAVING Clause

The HAVING clause may contain any function listed in the field list. Complex HAVING clauses such as this are supported:

```
SELECT fieldA, fieldB, count(*), sum(fieldC), avg(fieldY)
FROM tableA
WHERE fieldC = 'term1 term2'
GROUP BY fieldA, fieldB
HAVING ((sum(fieldC) > 1000) AND (avg(fieldY) <= 10))
ORDER BY sum(fieldC) asc
LIMIT 100
```

## Best Practices

### Separate Collections

It makes sense to create a separate SolrCloud collection just for the /sql handler. This collection can be created using SolrCloud's standard collection API.

Since this collection only exists to handle /sql requests and provide a pool of worker nodes, this collection does not need to hold any data. Worker nodes are selected randomly from the entire pool of available nodes in the /sql handler's collection. So to grow this collection dynamically replicas can be added to existing shards. New replicas will automatically be put to work after they've been added.

### Parallel SQL Queries

An earlier section describes how the SQL interface translates the SQL statement to a streaming expression. One of the parameters of the request is the aggregationMode, which defines if the query should use a MapReduce-like shuffling technique or push the operation down into the search engine.

### Parallelized Queries

The Parallel SQL architecture consists of three logical tiers: a **SQL** tier, a **Worker** tier, and a **Data Table** tier. By default the SQL and Worker tiers are collapsed into the same physical SolrCloud collection.

#### SQL Tier

The SQL tier is where the /sql handler resides. The /sql handler takes the SQL query and translates it to a parallel query plan. It then selects worker nodes to execute the plan and sends the query plan to each worker node to be run in parallel.

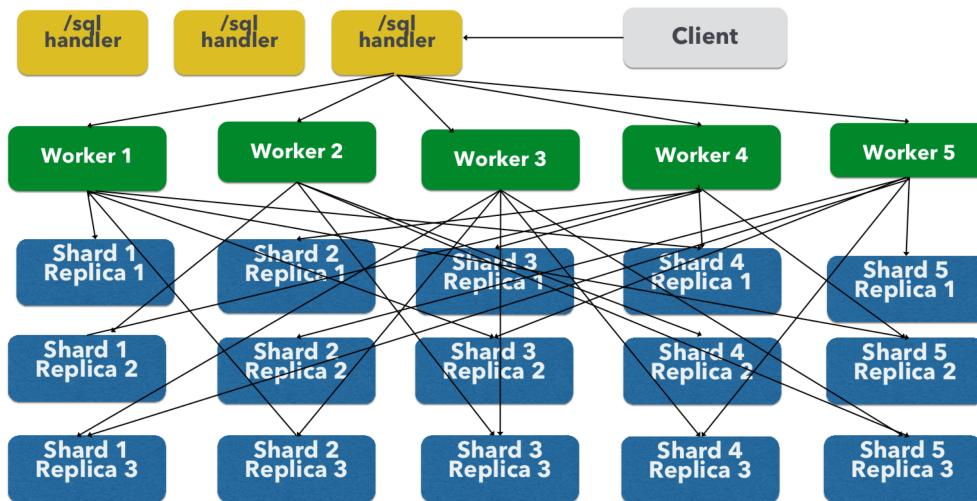
Once the query plan has been executed by the worker nodes, the /sql handler then performs the final merge of the tuples returned by the worker nodes.

#### Worker Tier

The workers in the worker tier receive the query plan from the /sql handler and execute the parallel query plan. The parallel execution plan includes the queries that need to be made on the Data Table tier and the relational algebra needed to satisfy the query. Each worker node assigned to the query is shuffled 1/N of the tuples from the Data Tables. The worker nodes execute the query plan and stream tuples back to the worker nodes.

## Data Table Tier

The Data Table tier is where the tables reside. Each table is its own SolrCloud collection. The Data Table layer receives queries from the worker nodes and emits tuples (search results). The Data Table tier also handles the initial sorting and partitioning of tuples sent to the workers. This means the tuples are always sorted and partitioned before they hit the network. The partitioned tuples are sent directly to the correct worker nodes in the proper sort order, ready to be reduced.



### How Parallel SQL Queries are Distributed

The image above shows the three tiers broken out into different SolrCloud collections for clarity. In practice the /sql handler and worker collection by default share the same collection.

**Note:** The image shows the network flow for a single Parallel SQL Query (SQL over MapReduce). This network flow is used when `map_reduce` aggregation mode is used for GROUP BY aggregations or the SELECT DISTINCT query. The traditional SolrCloud network flow (without workers) is used when the facet aggregation mode is used.

Below is a description of the flow:

1. The client sends a SQL query to the /sql handler. The request is handled by a single /sql handler instance.
2. The /sql handler parses the SQL query and creates the parallel query plan.
3. The query plan is sent to worker nodes (in green).
4. The worker nodes execute the plan in parallel. The diagram shows each worker node contacting a collection in the Data Table tier (in blue).
5. The collection in the Data Table tier is the table from the SQL query. Notice that the collection has five shards each with 3 replicas.
6. Notice that each worker contacts one replica from each shard. Because there are 5 workers, each worker is returned 1/5 of the search results from each shard. The partitioning is done inside of the Data Table tier so there is no duplication of data across the network.
7. Also notice with this design ALL replicas in the data layer are shuffling (sorting & partitioning) data simultaneously. As the number of shards, replicas and workers grows this design allows for a massive amount of computing power to be applied to a single query.

8. The worker nodes process the tuples returned from the Data Table tier in parallel. The worker nodes perform the relational algebra needed to satisfy the query plan.
9. The worker nodes stream tuples back to the /sql handler where the final merge is done, and finally the tuples are streamed back to the client.

## SQL Clients and Database Visualization Tools

The SQL interface supports queries sent from SQL clients and database visualization tools such as DbVisualizer and Apache Zeppelin.

### Generic Clients

For most Java based clients, the following jars will need to be placed on the client classpath:

- all .jars found in \$SOLR\_HOME/dist/solrj-libs
- the SolrJ .jar found at \$SOLR\_HOME/dist/solr-solrj-<version>.jar

If you are using Maven, the org.apache.solr.solrj artifact contains the required jars.

Once the jars are available on the classpath, the Solr JDBC driver name is org.apache.solr.client.solrj.io.sql.DriverImpl and a connection can be made with the following connection string format:

```
jdbc:solr://$SOLR_ZK_CONNECTION_STRING?collection=$COLLECTION_NAME
```

There are other parameters that can be optionally added to the connection string like aggregationMode and numWorkers.

### DbVisualizer

A step-by-step guide for setting up [DbVisualizer](#) is in the section [Solr JDBC - DbVisualizer](#).

### SQuirreL SQL

A step-by-step guide for setting up [SQuirreL SQL](#) is in the section [Solr JDBC - SQuirreL SQL](#).

### Apache Zeppelin (incubating)

A step-by-step guide for setting up [Apache Zeppelin](#) is in the section [Solr JDBC - Apache Zeppelin](#).

### Python/Jython

Examples of using Python and Jython for connecting to Solr with the Solr JDBC driver are available in the section [Solr JDBC - Python/Jython](#).

### R

Examples of using R for connecting to Solr with the Solr JDBC driver are available in the section [Solr JDBC - R](#).

## Solr JDBC - DbVisualizer

Solr's JDBC driver supports DbVisualizer for querying Solr.

For [DbVisualizer](#), you will need to create a new driver for Solr using the DbVisualizer Driver Manager. This will add several SolrJ client .jars to the DbVisualizer classpath. The files required are:

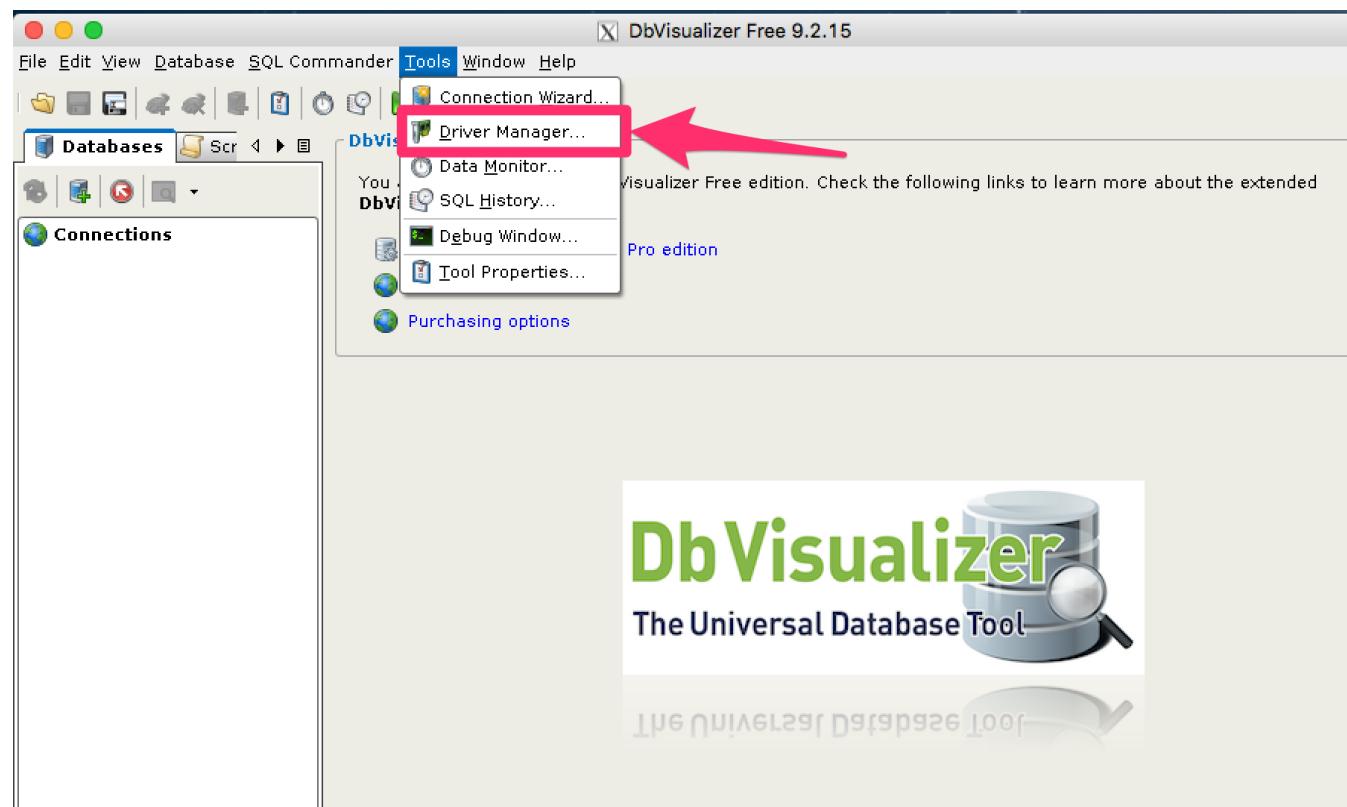
- all .jars found in \$SOLR\_HOME/dist/solrj-lib
- the SolrJ .jar found at \$SOLR\_HOME/dist/solr-solrj-<version>.jar

Once the driver has been created, you can create a connection to Solr with the connection string format outlined in the generic section and use the SQL Commander to issue queries.

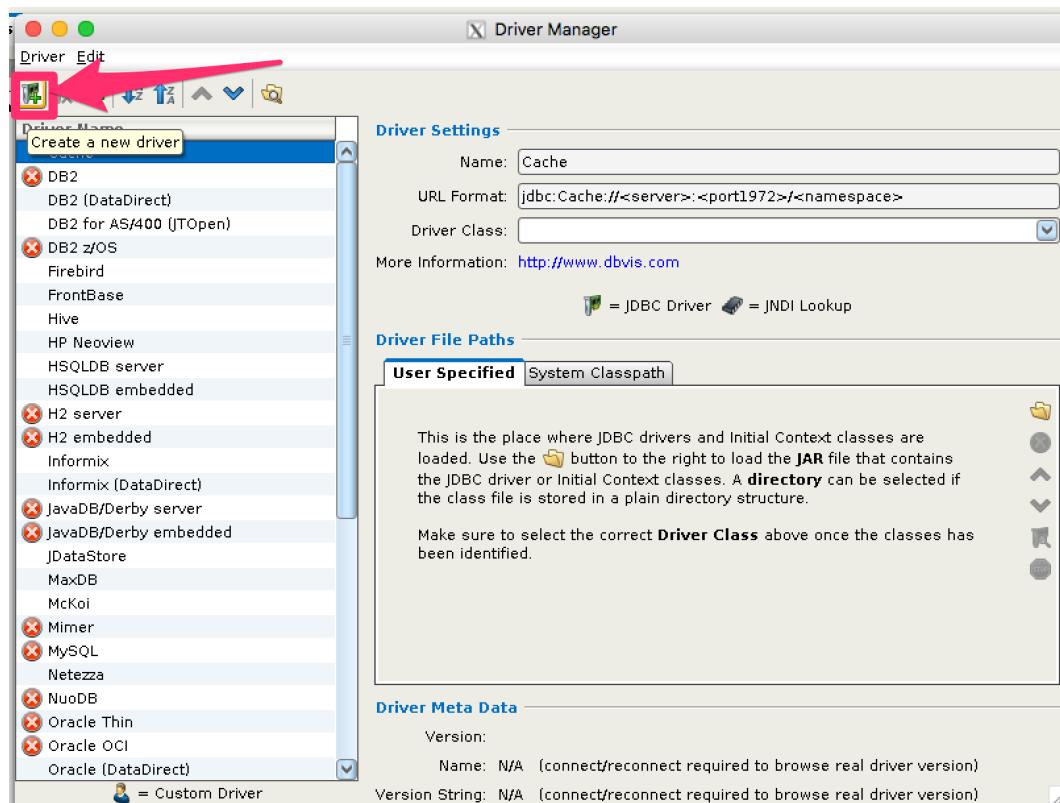
### Setup Driver

#### Open Driver Manager

From the Tools menu, choose Driver Manager to add a driver.



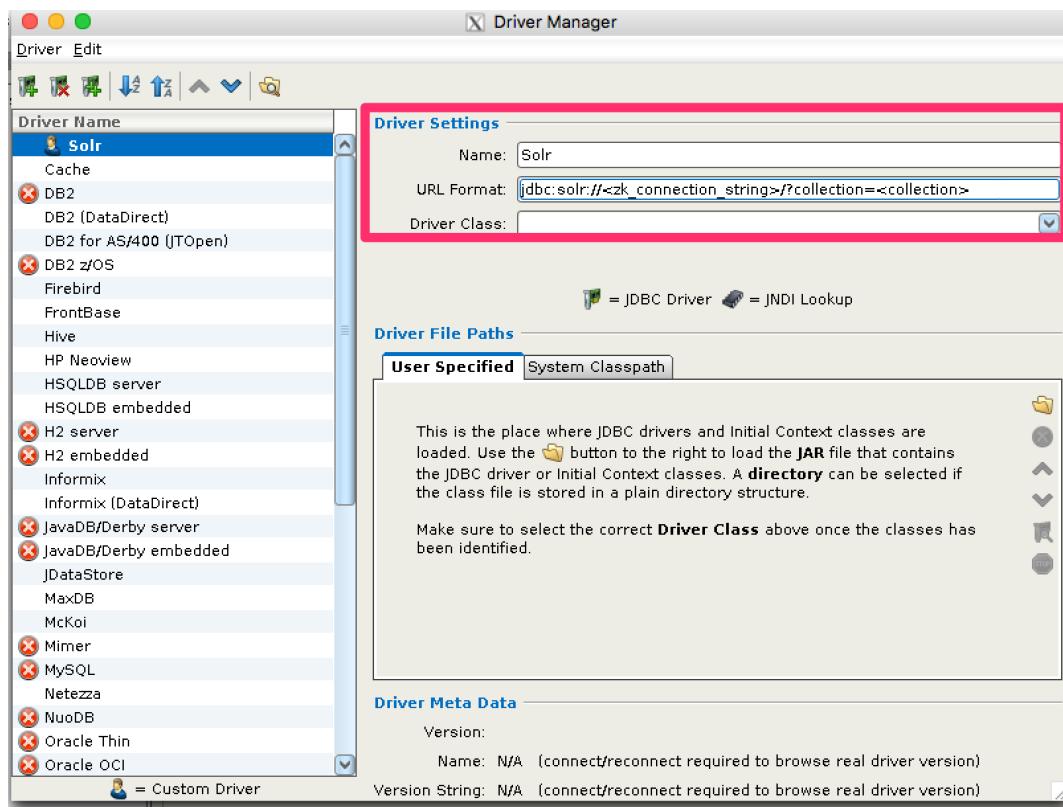
#### Create a New Driver



## Name the Driver in Driver Manager

Provide a name for the driver, and provide the URL format:

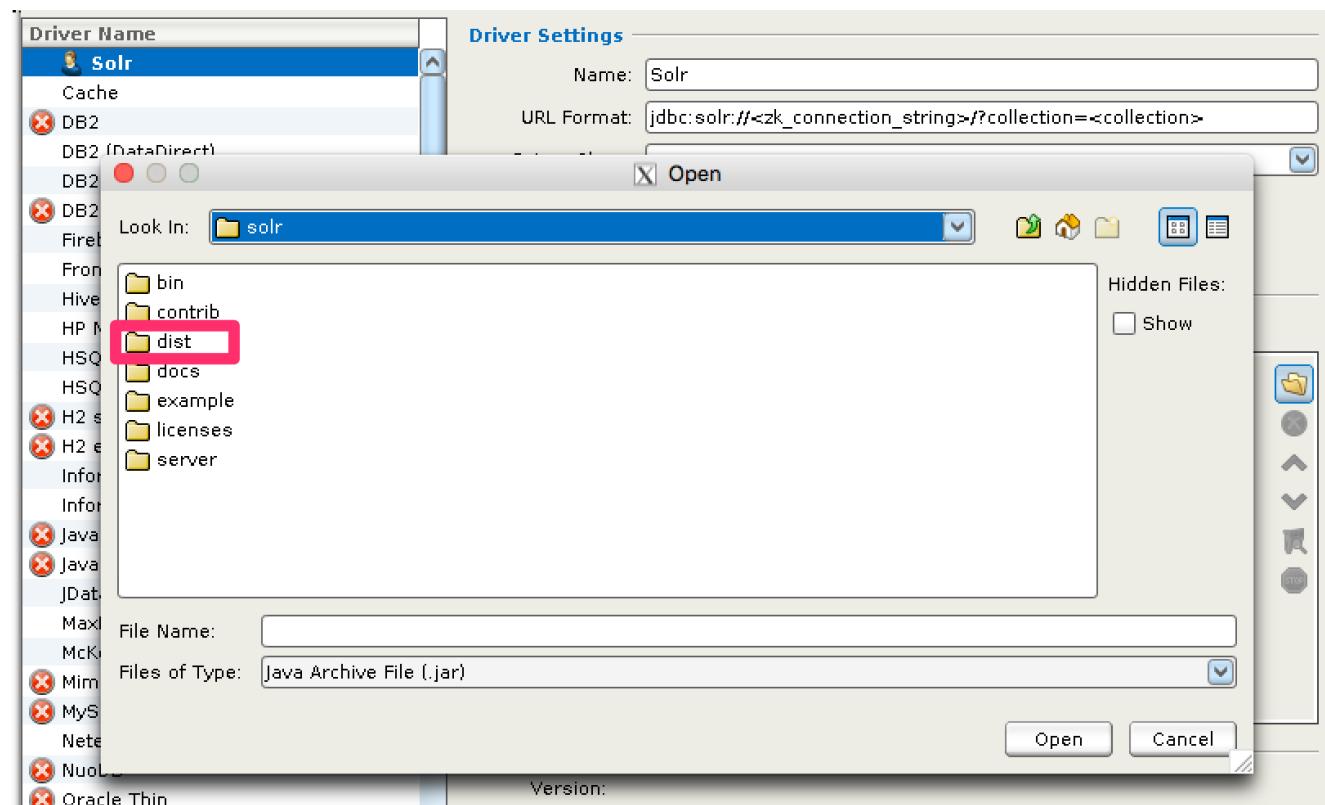
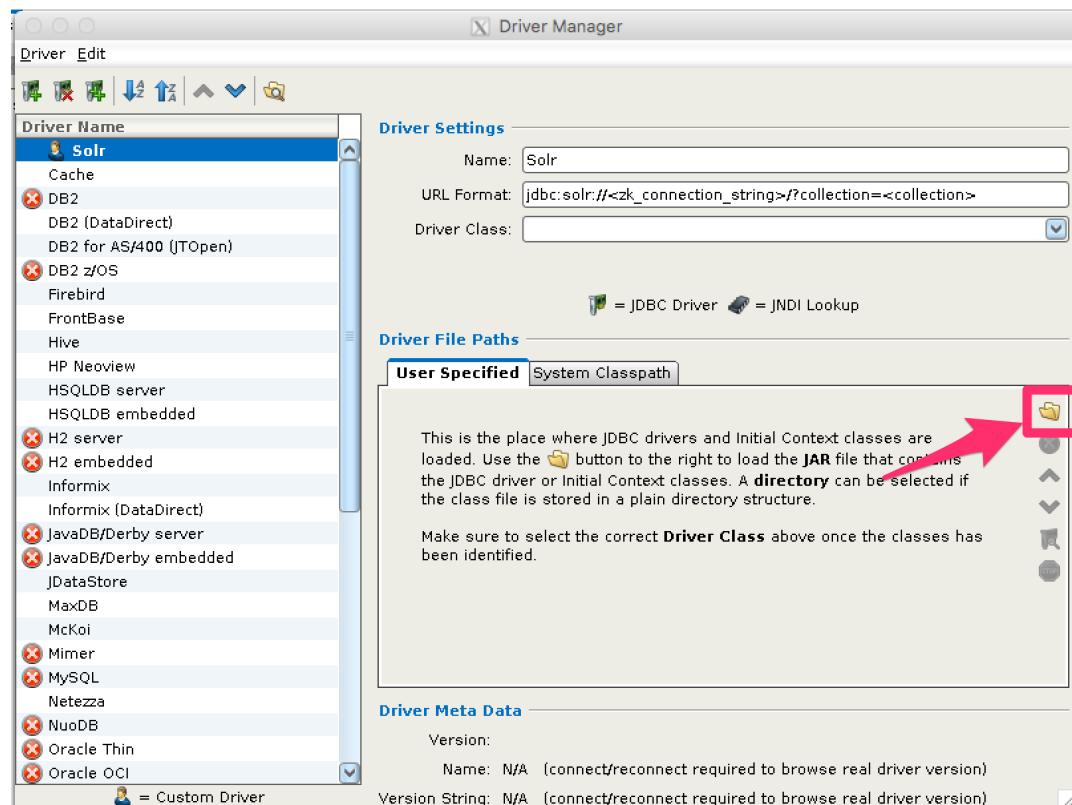
`jdbc:solr://<zk_connection_string>/?collection=<collection>`. Do not fill in values for the variables "zk\_connection\_string" and "collection", those will be provided later when the connection to Solr is configured. The Driver Class will also be automatically added when the driver .jars are added.

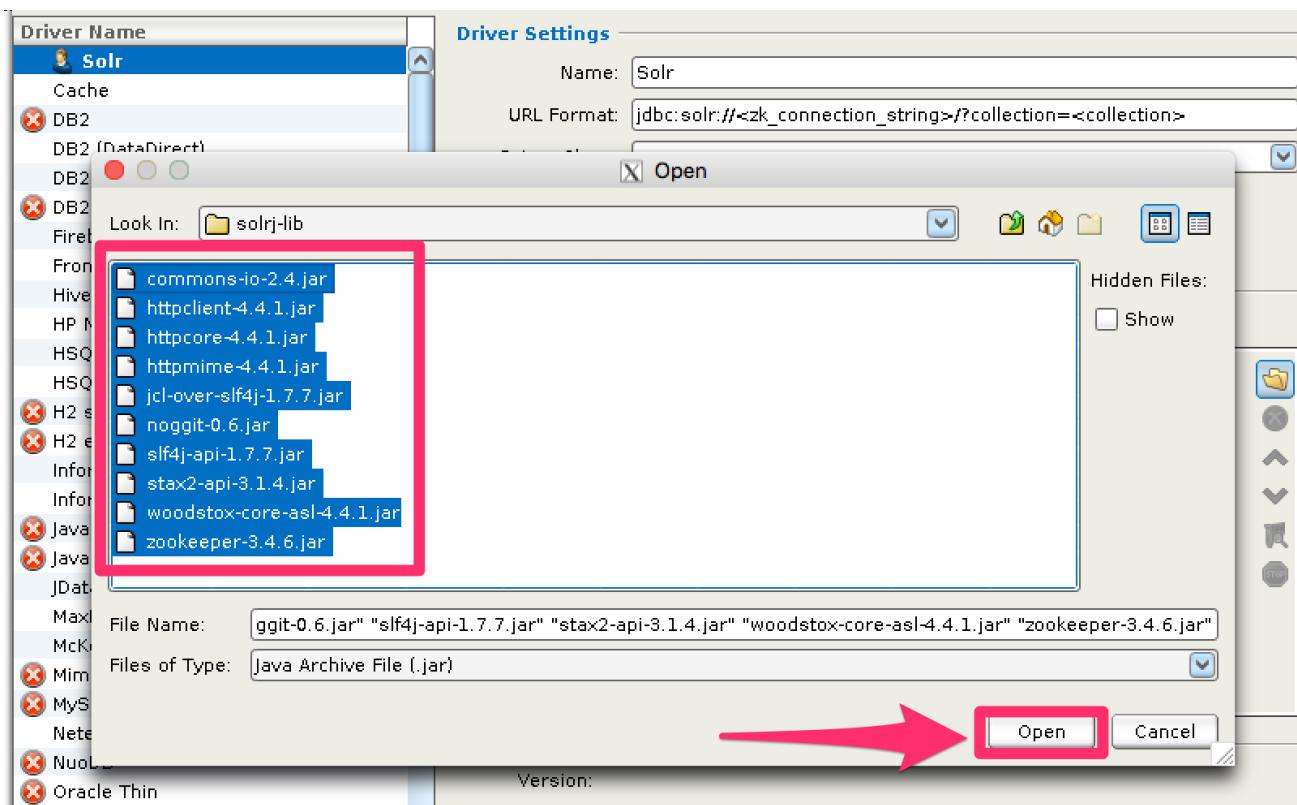
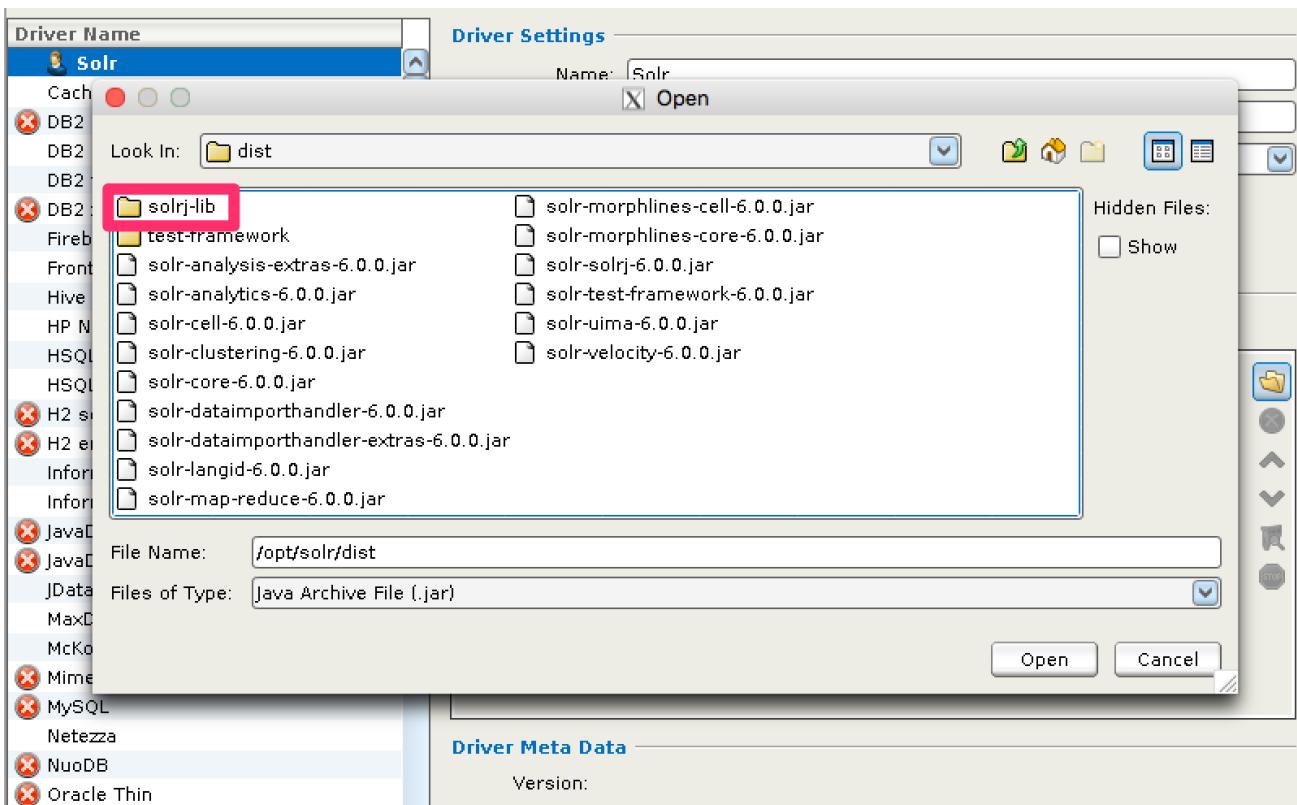


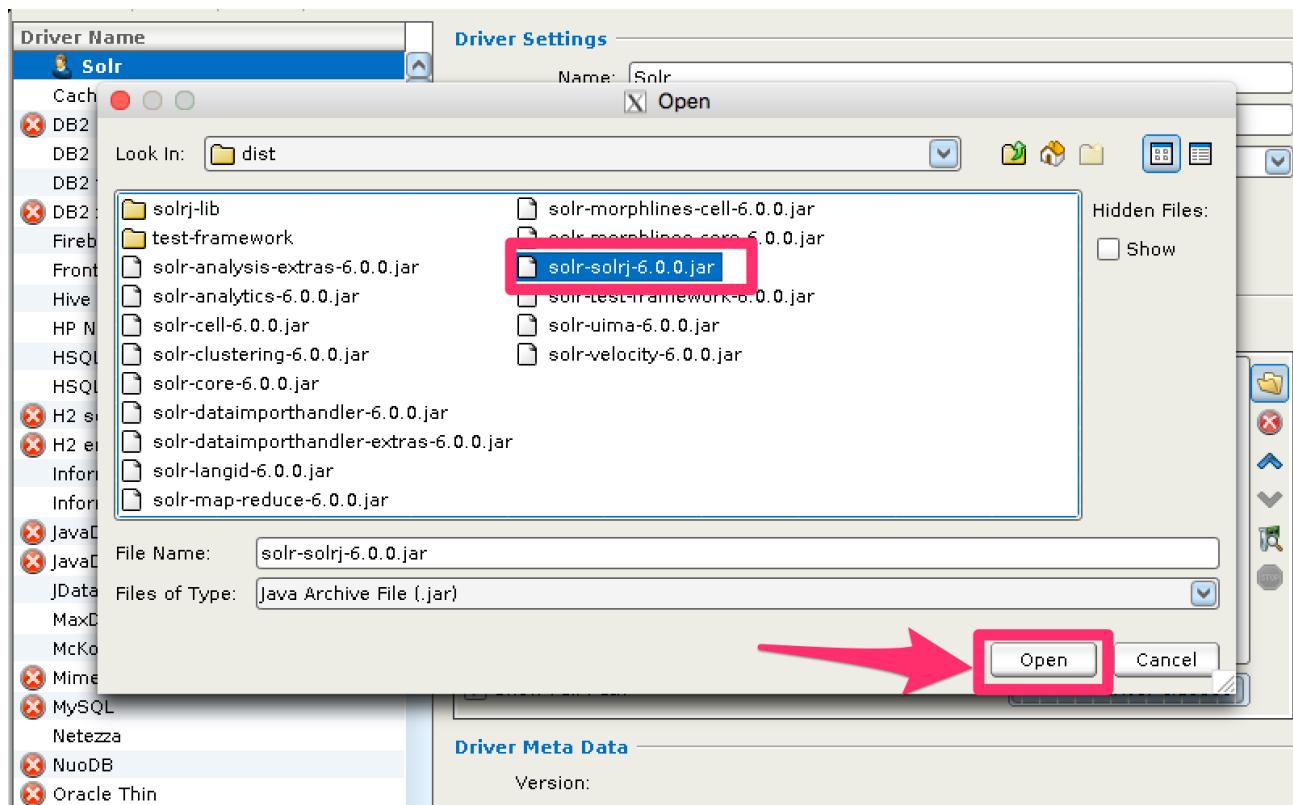
## Add Driver Files to Classpath

The driver files to be added are:

- all .jars in \$SOLR\_HOME/dist/solrj-lib
- the SolrJ .jar found in \$SOLR\_HOME/dist/solr-solrj-<version>.jar







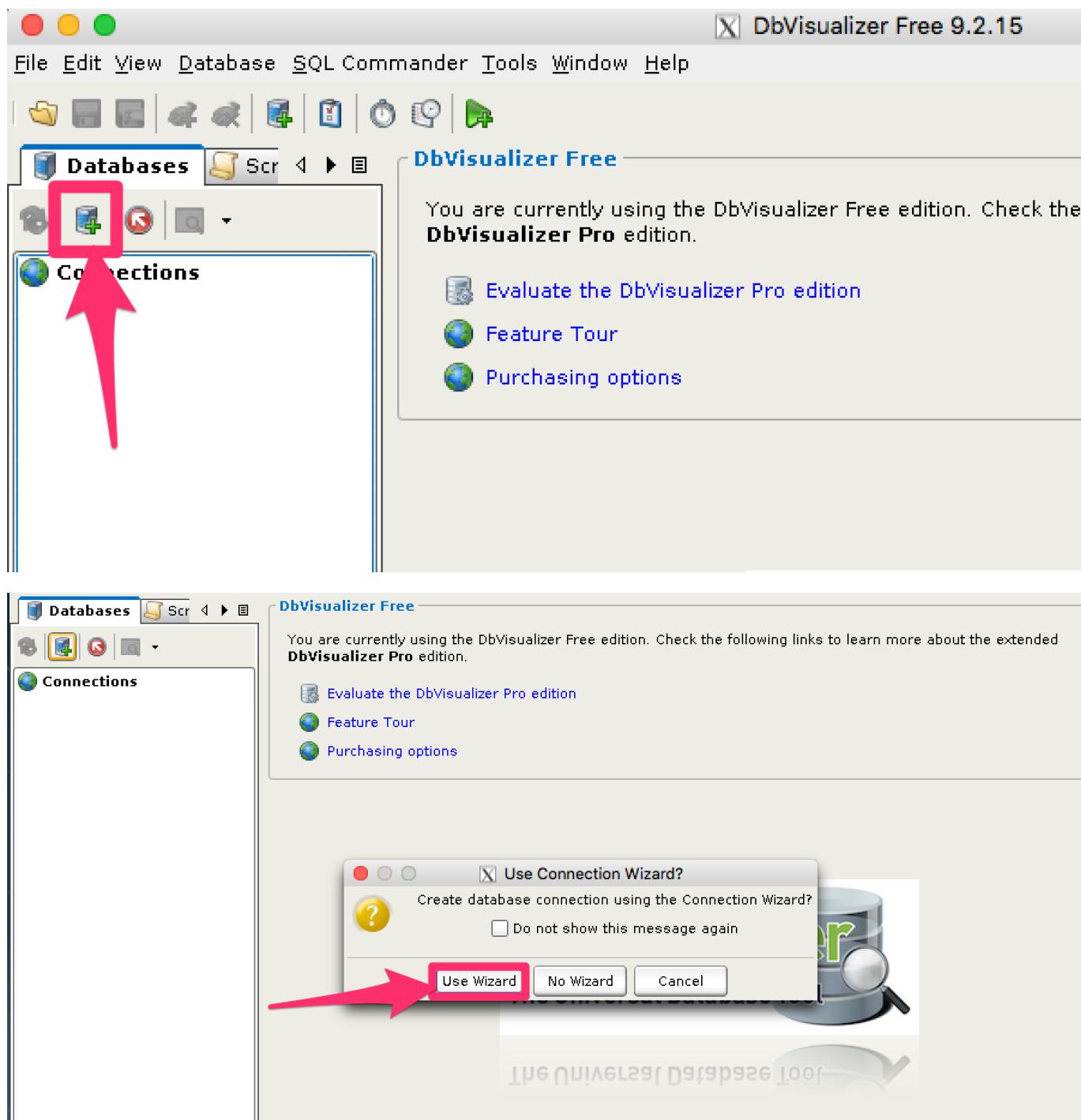
## Review and Close Driver Manager

Once the driver files have been added, you can close the Driver Manager.

## Create a Connection

Next, create a connection to Solr using the driver just created.

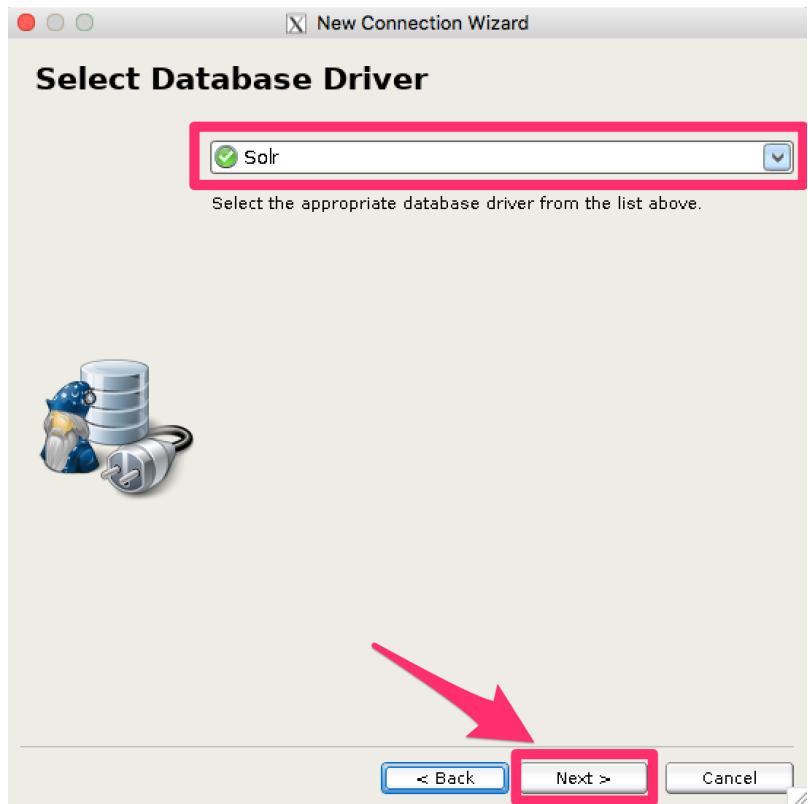
## Use the Connection Wizard



### Name the Connection

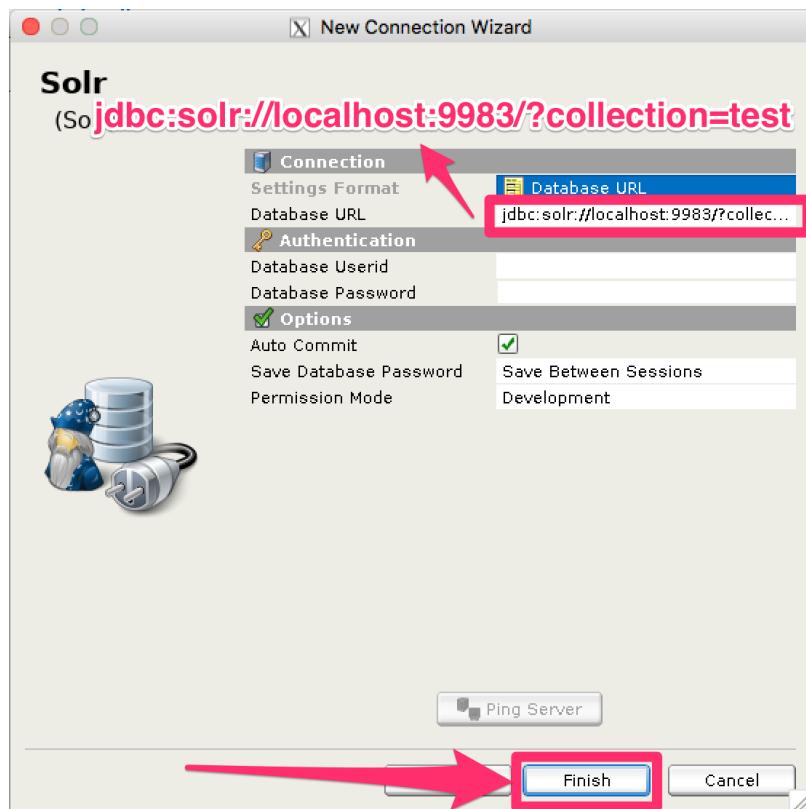


### Select the Solr driver



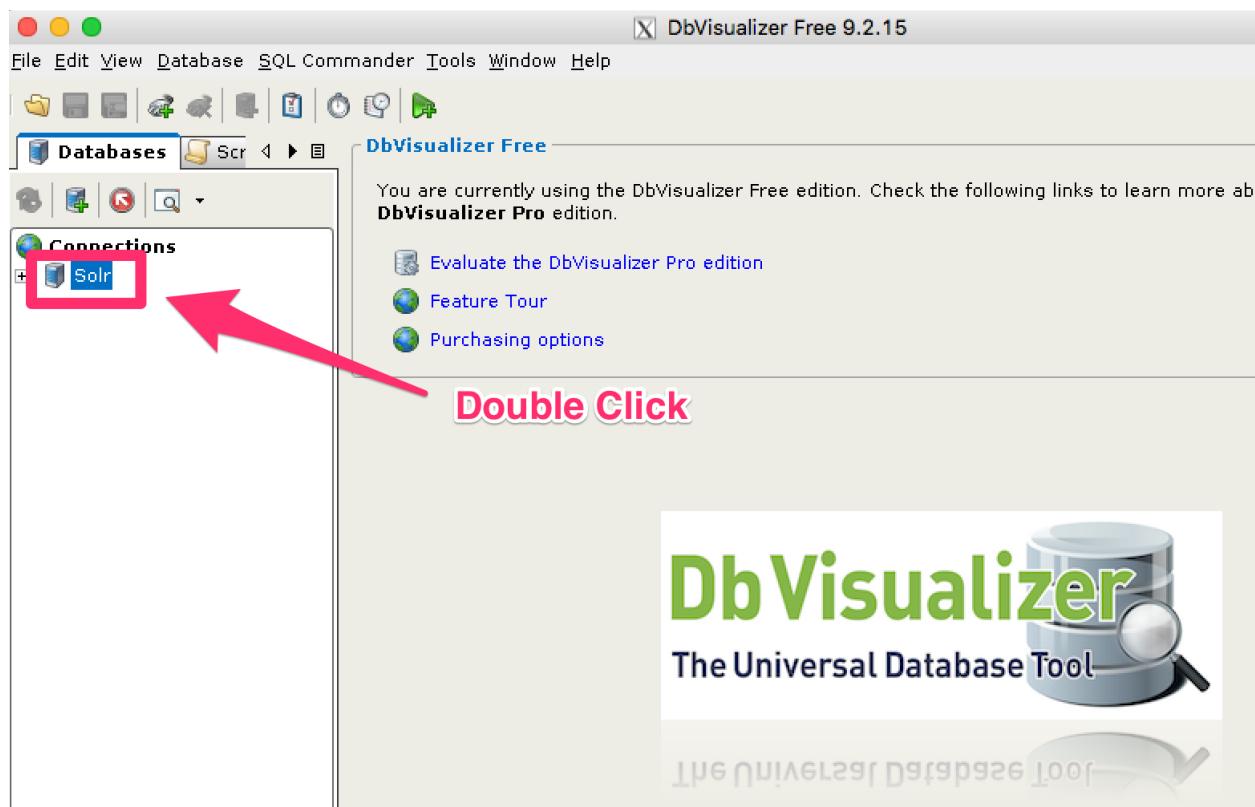
### Specify the Solr URL

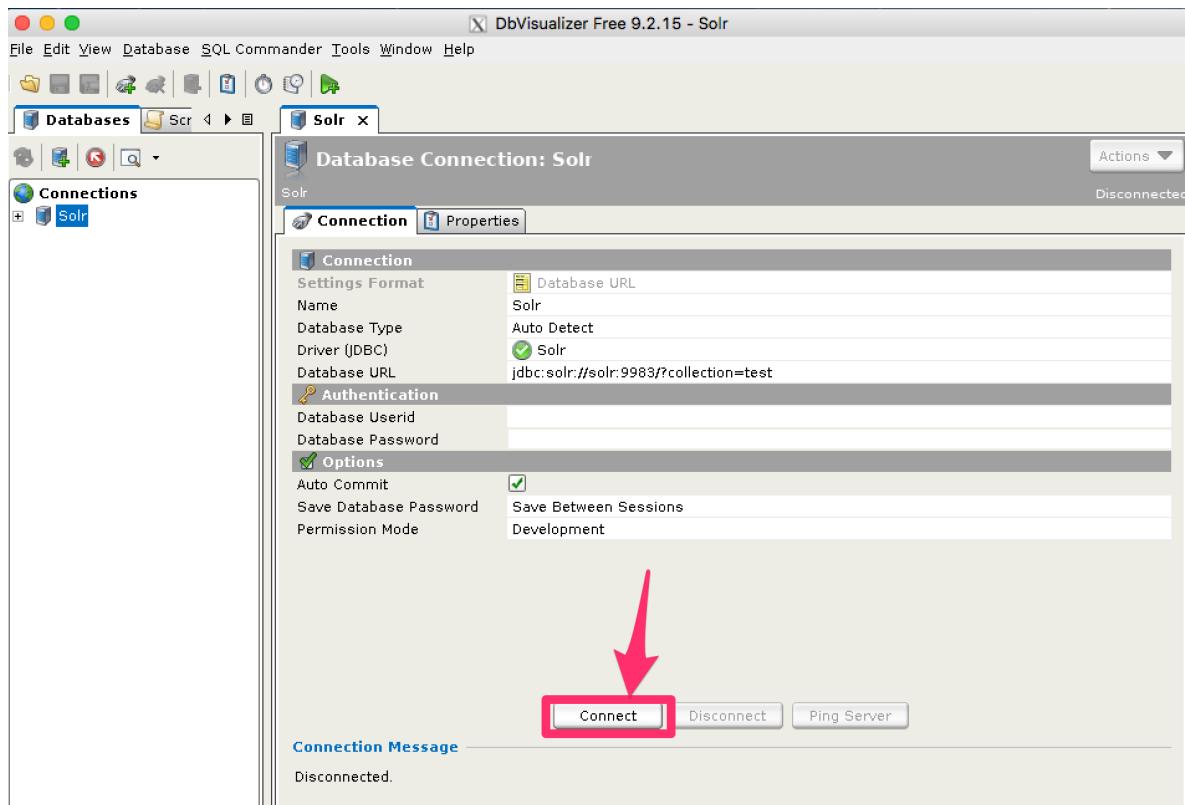
Provide the Solr URL, using the ZooKeeper host and port and the collection. For example,  
jdbc:solr://localhost:9983?collection=test



## Open and Connect to Solr

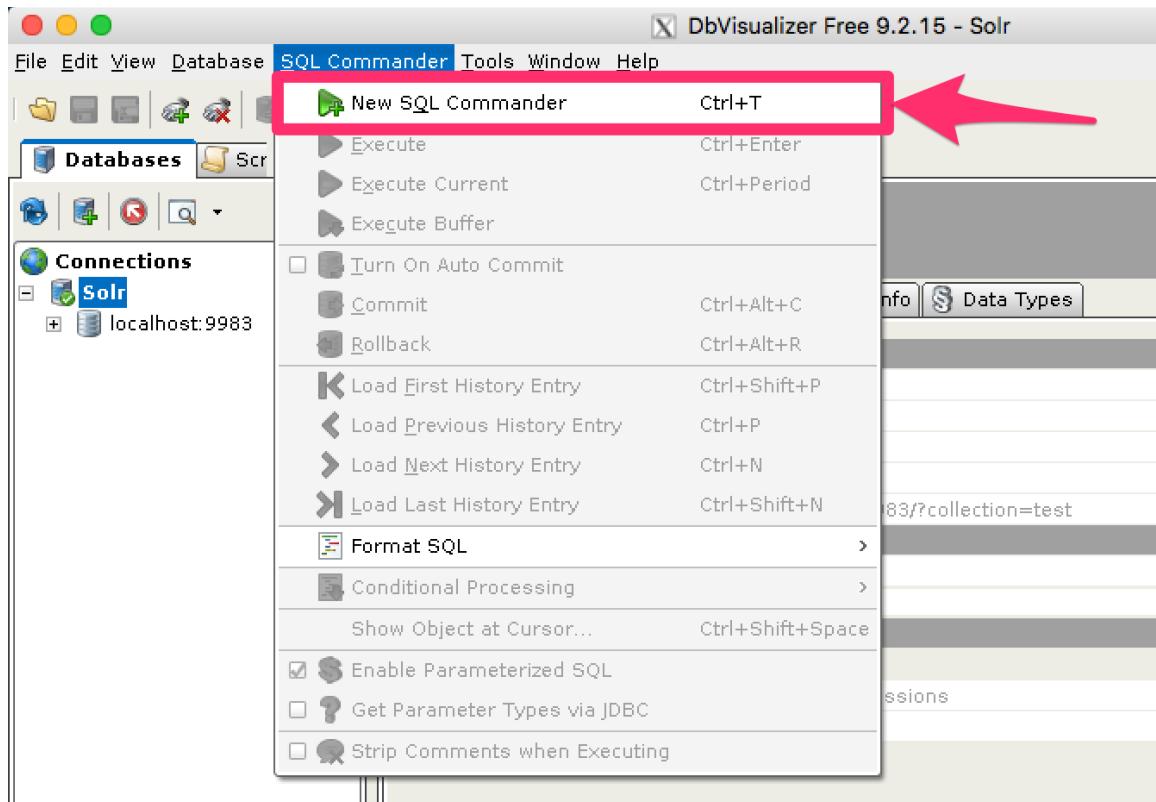
Once the connection has been created, double-click on it to open the connection details screen and connect to Solr.

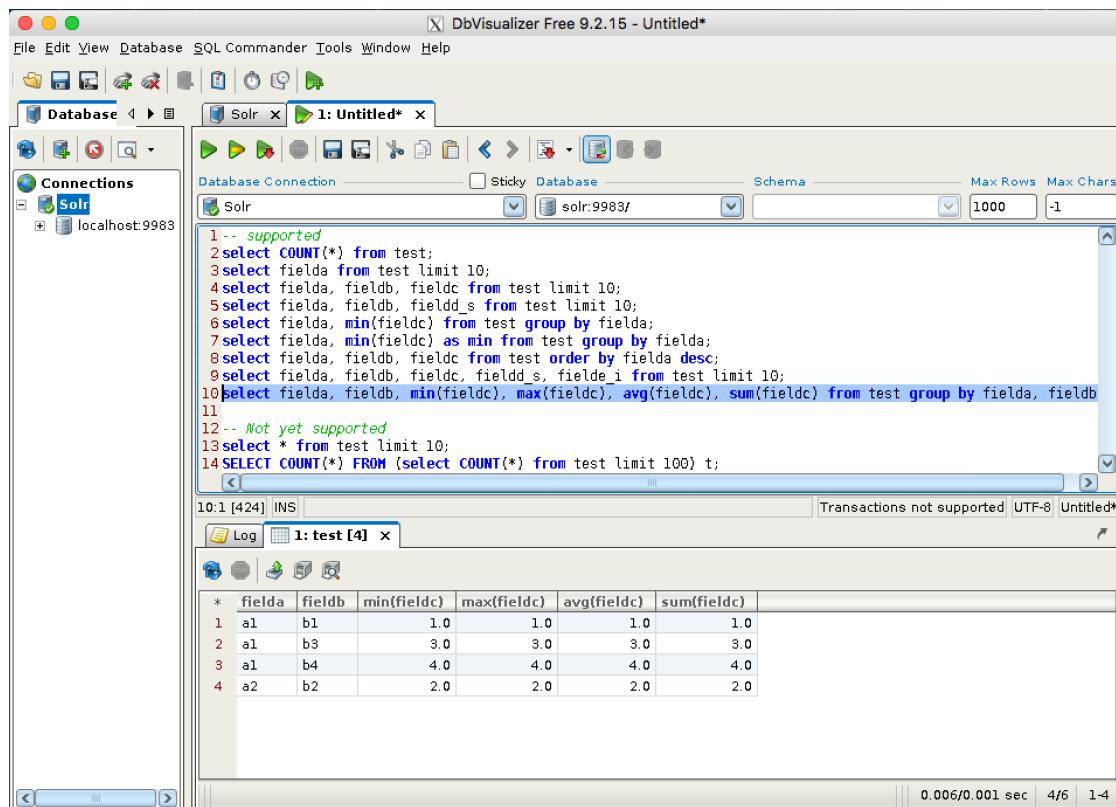




## Open SQL Commander to Enter Queries

When the connection is established, you can use the SQL Commander to issue queries and view data.





## Solr JDBC - SQuirreL SQL

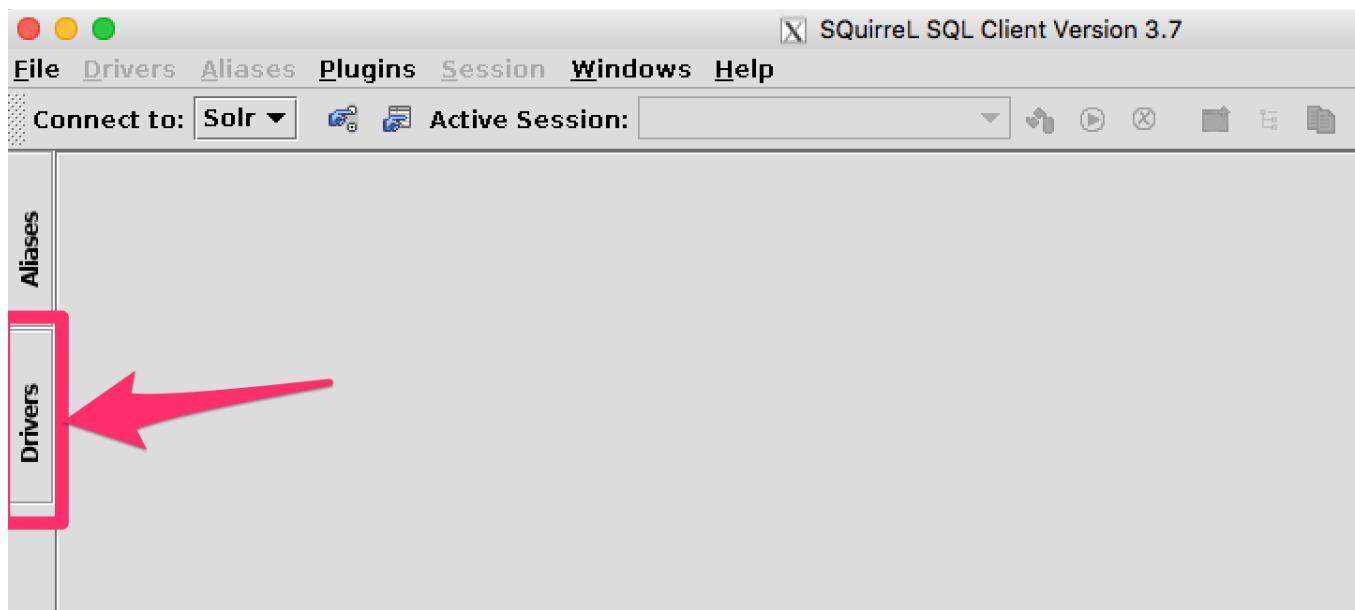
For [SQuirreL SQL](#), you will need to create a new driver for Solr. This will add several SolrJ client .jars to the SQuirreL SQL classpath. The files required are:

- all .jars found in \$SOLR\_HOME/dist/solrj-libs
- the SolrJ .jar found at \$SOLR\_HOME/dist/solr-solrj-<version>.jar

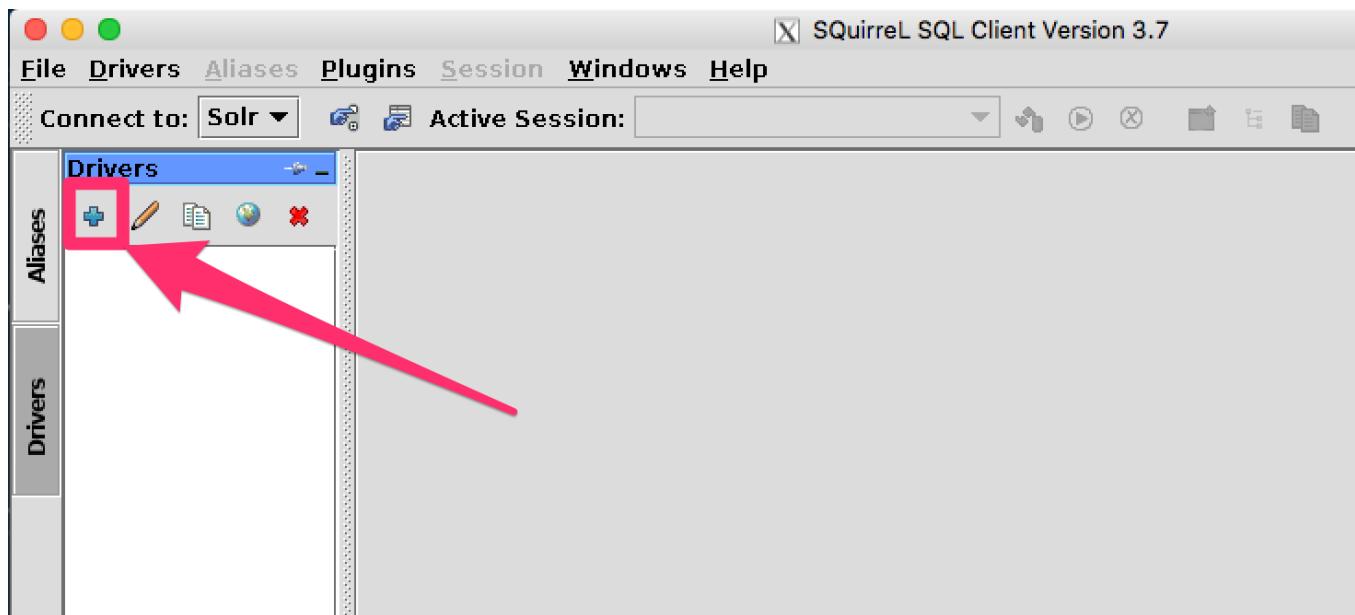
Once the driver has been created, you can create a connection to Solr with the connection string format outlined in the generic section and use the editor to issue queries.

## Add Solr JDBC Driver

### Open Drivers



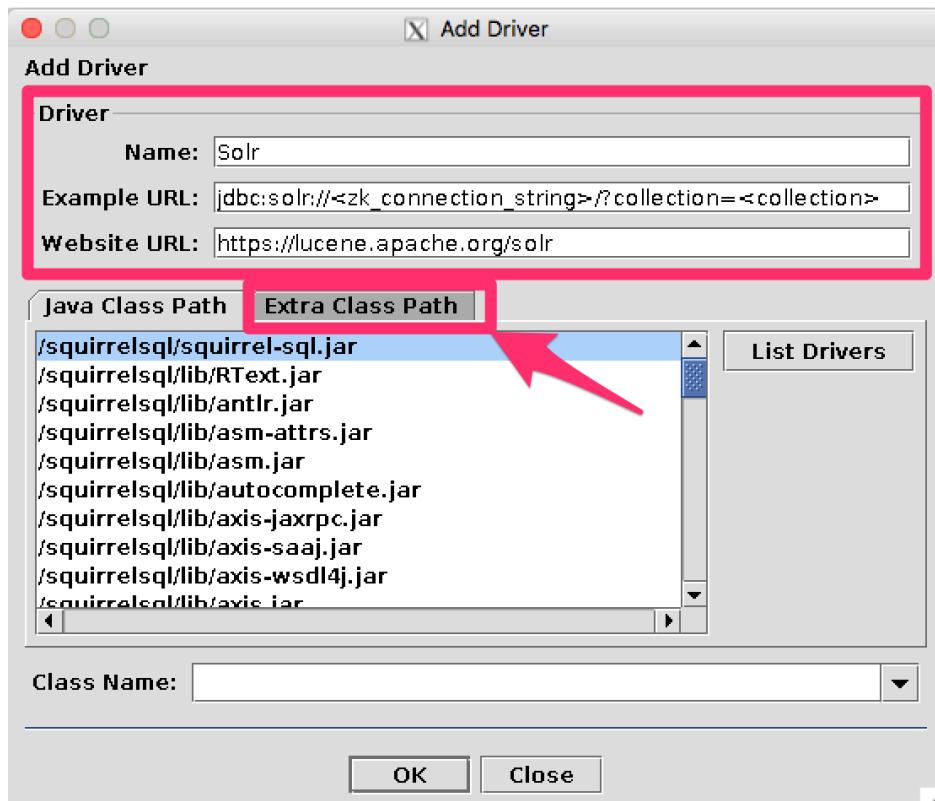
### Add Driver



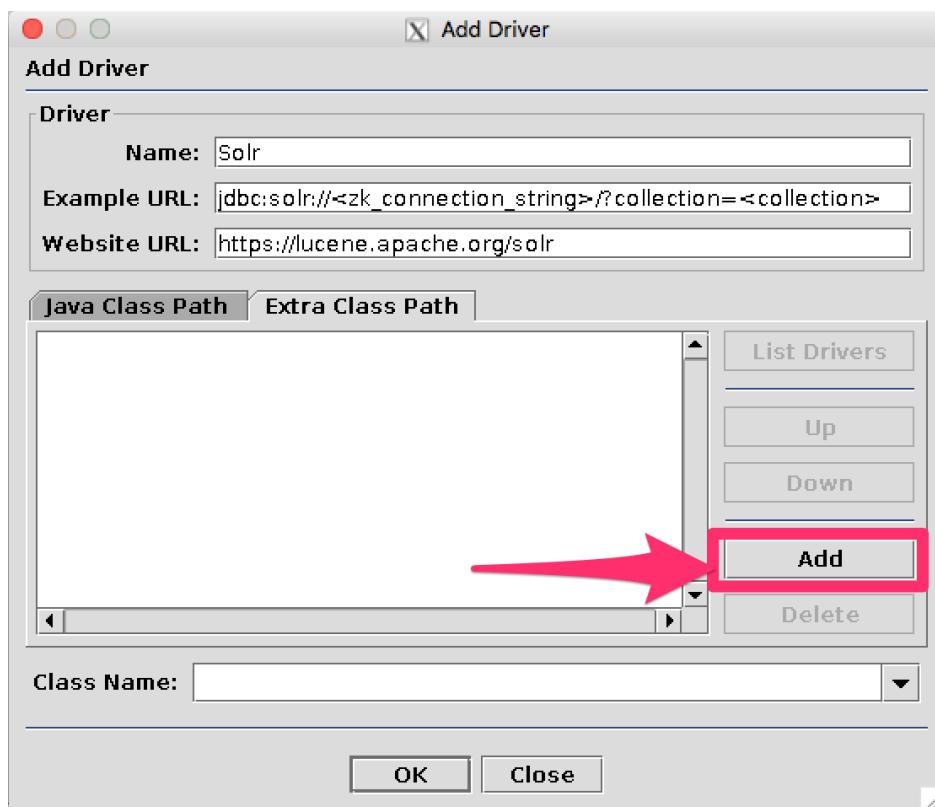
### Name the Driver

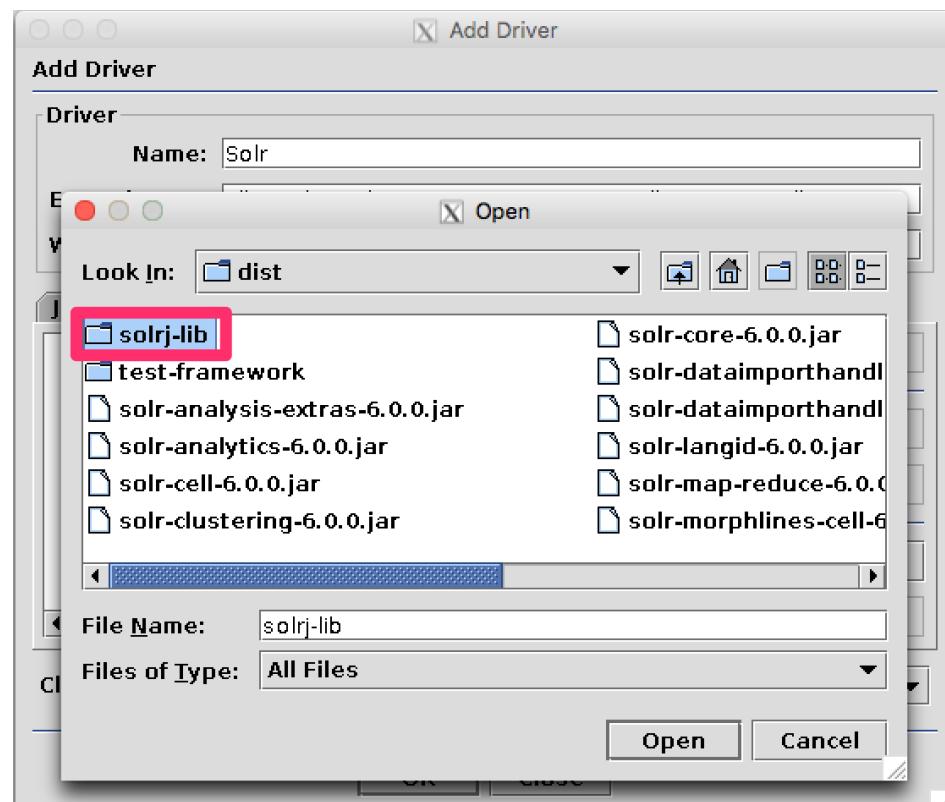
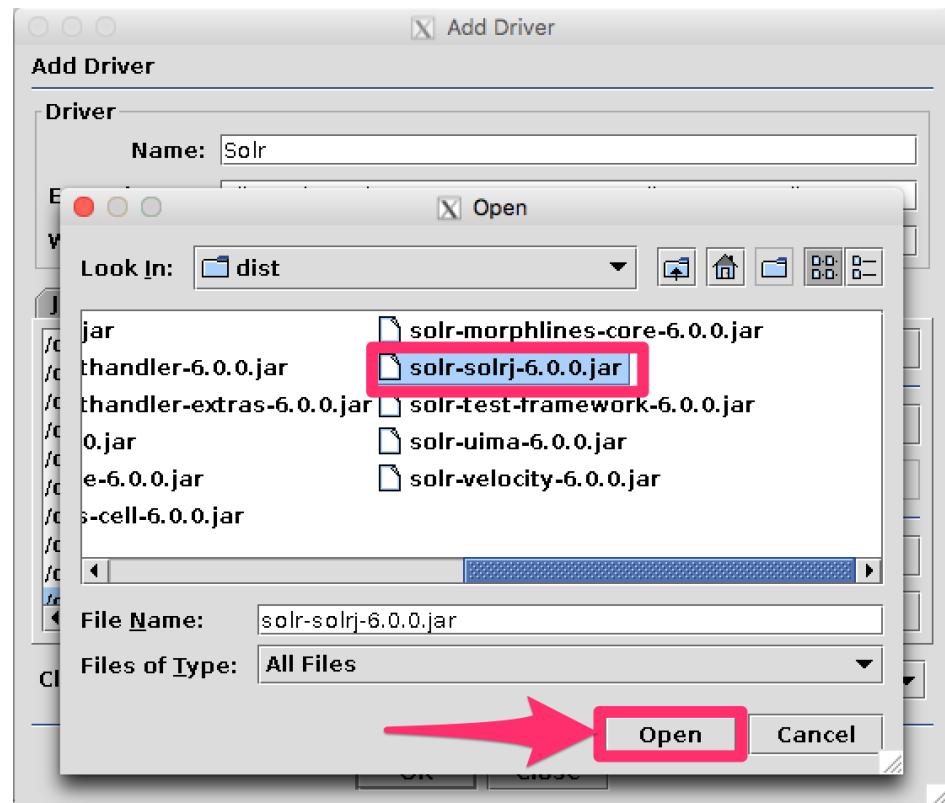
Provide a name for the driver, and provide the URL format:

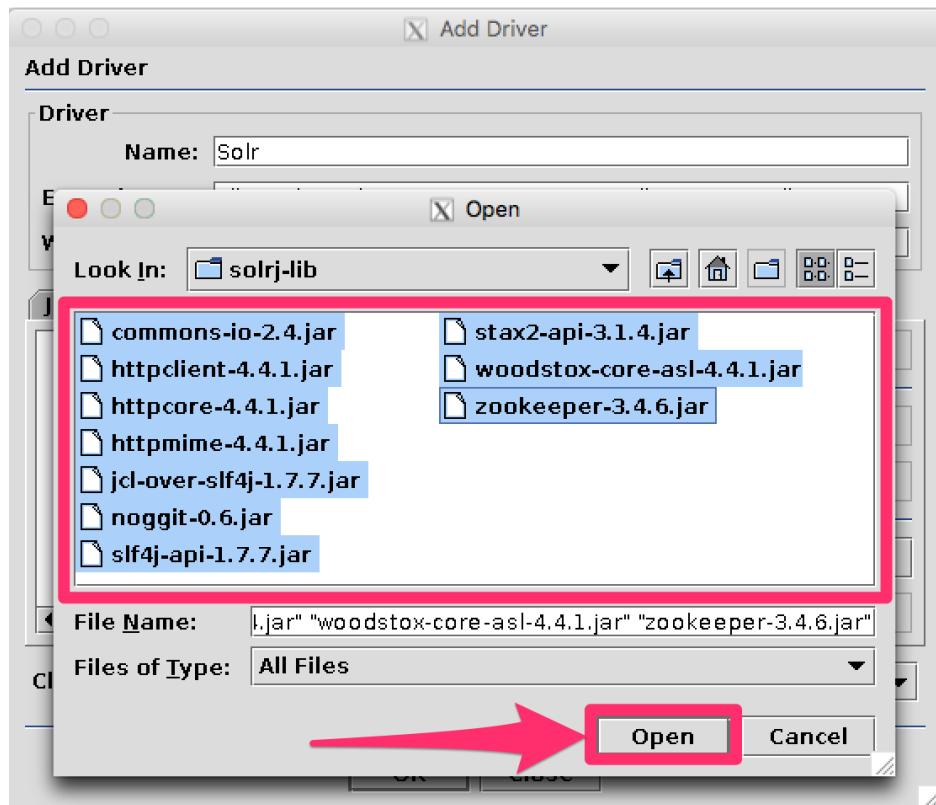
`jdbc:solr://<zk_connection_string>/?collection=<collection>`. Do not fill in values for the variables "zk\_connection\_string" and "collection", those will be defined later when the connection to Solr is configured.



### Add Solr JDBC jars to Classpath

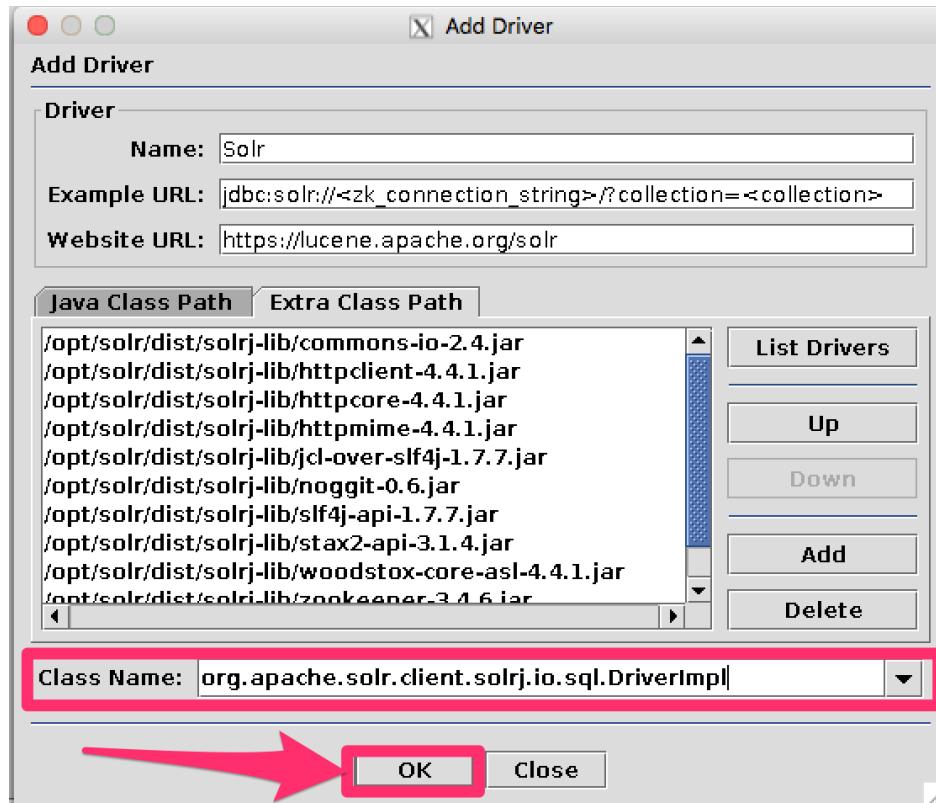






### Add the Solr JDBC driver class name

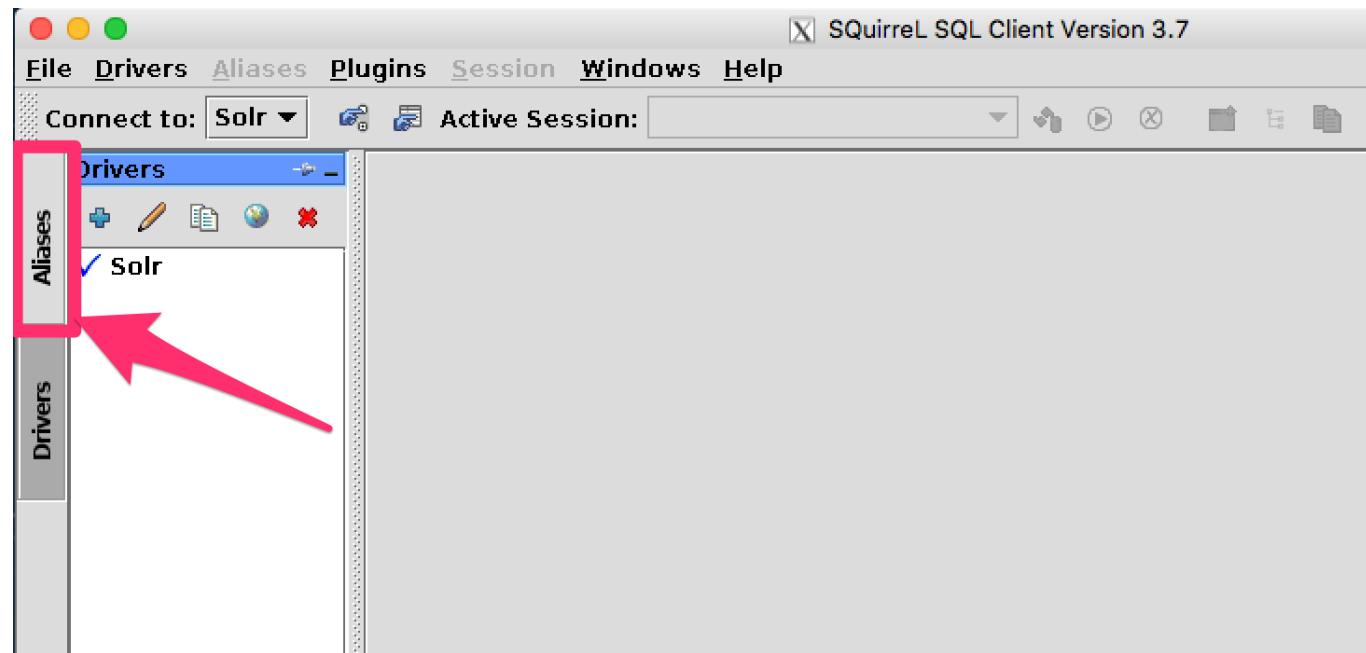
After adding the .jars, you will need to additionally define the Class Name  
org.apache.solr.client.solrj.io.sql.DriverImpl.



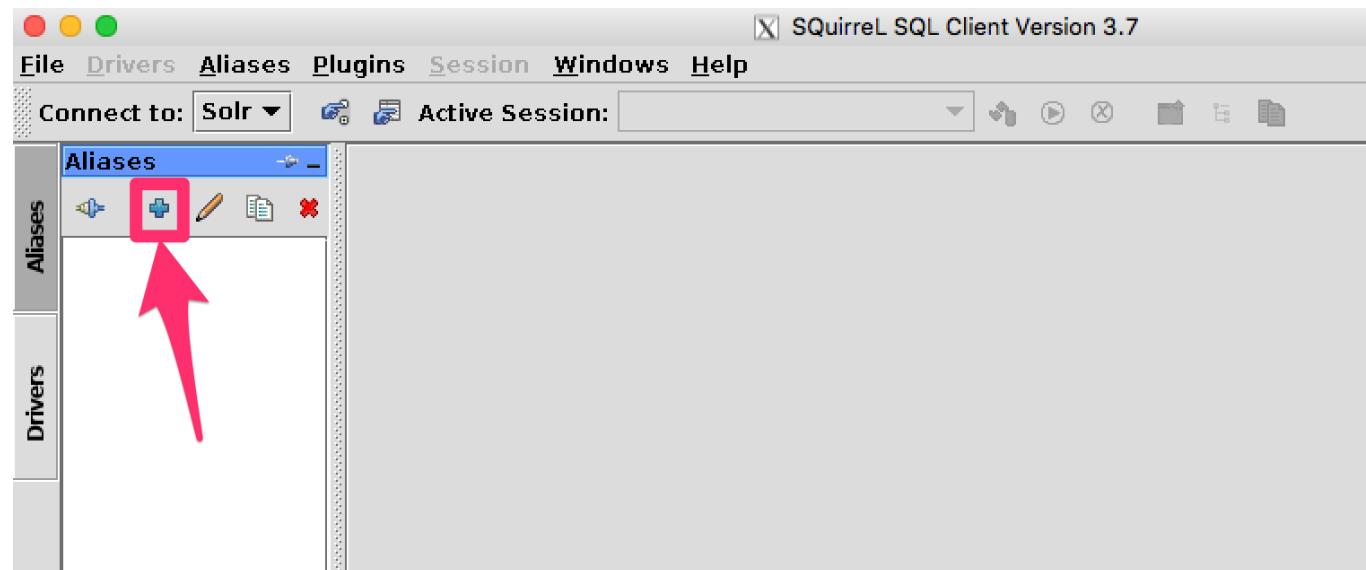
## Create an Alias

To define a JDBC connection, you must define an alias.

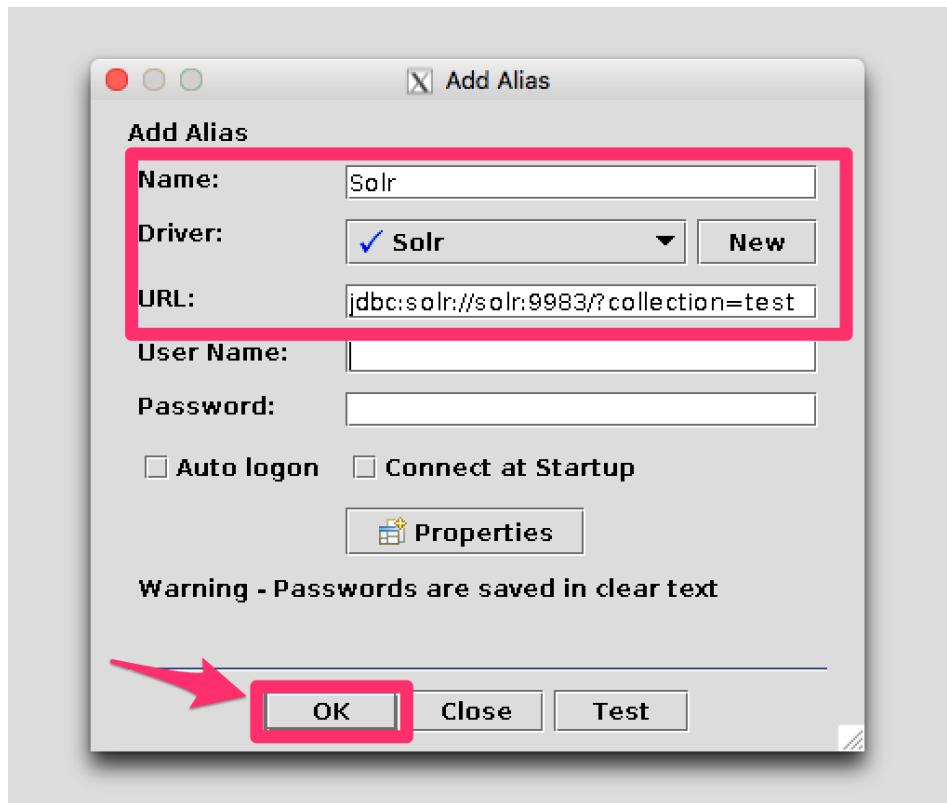
### Open Aliases



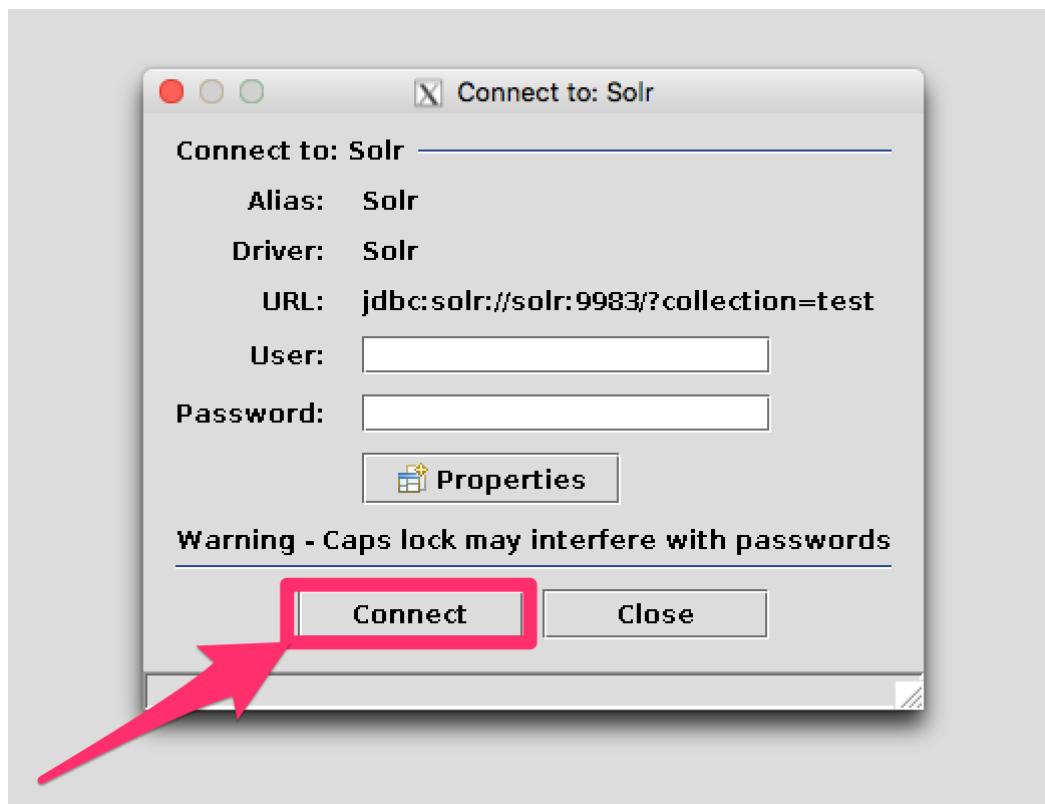
### Add an Alias



### Configure the Alias



## Connect to the Alias



## Querying

Once you've successfully connected to Solr, you can use the SQL interface to enter queries and work with

data.

The screenshot shows the SQuirreL SQL Client interface. The top menu bar includes File, Drivers, Aliases, Plugins, Session, Windows, and Help. The session dropdown shows "2 - Solr (solr:9983)". The main area has tabs for Objects and SQL. A SQL query is entered: "select fielda, fieldb, min(fieldc), max(fieldc), avg(fieldc), sum(fieldc) from test group by fielda, fieldb". The results tab displays the following table:

fielda	fieldb	min(fieldc)	max(fieldc)	avg(fieldc)	sum(fieldc)
a1	b1	1	1	1	1
a1	b3	3	3	3	3
a1	b4	4	4	4	4
a2	b2	2	2	2	2

The status bar at the bottom shows "/Solr" and "1,108 / 108". A message in the status bar says: "Driver class org.apache.solr.client.solrj.io.sql.DriverImpl successfully registered for driver definition: Solr. Please try out the Tools popup by hitting ctrl+t in the SQL Editor. Do it three times to stop this message." Below the status bar, there are logs: Errors 83, Warnings 0, Infos 12, and a timestamp of 5:09:38 PM GMT.

## Solr JDBC - Apache Zeppelin

The Solr JDBC driver can support Apache Zeppelin.



This requires Apache Zeppelin 0.6.0 or greater which contains the JDBC interpreter.

To use [Apache Zeppelin](#) with Solr, you will need to create a JDBC interpreter for Solr. This will add SolrJ to the interpreter classpath. Once the interpreter has been created, you can create a notebook to issue queries. The [Apache Zeppelin JDBC interpreter documentation](#) provides additional information about JDBC prefixes and other features.

## Create the Apache Solr JDBC Interpreter

Welcome to Zeppelin! (0.6.0-incubating-SNAPSHOT)

Zeppelin is web-based notebook that enables interactive data analytics.  
You can make beautiful data-driven, interactive, collaborative document with SQL, code and even more!

Notebook

- Import note
- Create new note

Filter

Solr JDBC

Help

Get started with [Zeppelin documentation](#)

Community

Please feel free to help us to improve Zeppelin,  
Any contribution are welcome!

Mailing list  
 Issues tracking  
 Github

Click "Interpreter" in the top navigation

## Interpreters

Manage interpreters settings. You can create create / remove settings. Note can bind/unbind these interpreter settings.

md	%md (default)	edit	restart	remove
<b>Option</b>				
<input type="checkbox"/> Separate Interpreter for each note				
<b>Properties</b>				
name	value			
zeppelin.interpreter.localRepo	/zeppelin/local-repo/2BDZY2GGQ			

Click "Create"

**Create new interpreter**

**Name**: Solr

**Interpreter**: jdbc

**Option**:  Separate interpreter for each note

**Properties**

name	value	description	action
default.password		The JDBC user password	x
default.url	jdbc:solr://solr:9983?collection=test	The URL, for JDBC.	x
default.driver	org.apache.solr.client.solrj.io.sql.DriverImpl	JDBC Driver Name	x
default.user	solr	The JDBC user name	x
common.max_count	1000	Max number of SQL result to display.	x
			+

**Dependencies**

artifact	exclude	action
org.apache.solr:solr-solrj:6.0.0	(Optional) comma separated groupId:artifactId list	+

**Save**

Enter information about your Solr installation



For most installations, Apache Zeppelin configures PostgreSQL as the JDBC interpreter default driver. The default driver can either be replaced by the Solr driver as outlined above or you can add a separate JDBC interpreter prefix as outlined in the [Apache Zeppelin JDBC interpreter documentation](#).

## Create a Notebook

**Interpreters**

Manage interpreters settings. You can move settings. Note can bind/unbind these interpreter settings.

**md** %md (default)

**Properties**

name	value
zeppelin.interpreter.localRepo	/zeppelin/local-repo/2BDZY2GGQ

**Solr** %jdbc

**Properties**

name	value
default.password	
default.user	solr
default.driver	org.apache.solr.client.solrj.io.sql.DriverImpl
default.url	jdbc:solr://solr:9983?collection=test
common.max_count	1000

Click Notebook -> Create new note

**Notebook** - Interpreter Configuration

**Create new note**

**Note Name**

Solr JDBC

**Create Note**

**Interpreters**

Manage interpreters settings. You can move settings. Note can bind/unbind these interpreter settings.

**md** %md (default)

**Properties**

name	value
zeppelin.interpreter.localRepo	/zeppelin/local-repo/2BDZY2GGQ

**Solr** %jdbc

**Properties**

name	value
default.password	
default.user	solr
default.driver	org.apache.solr.client.solrj.io.sql.DriverImpl
default.url	jdbc:solr://solr:9983?collection=test
common.max_count	1000

Provide a name and click "Create Note"

## JDBC Interpreter Copy Sheet

To facilitate easy copying the parameters mentioned in the screenshots, here is a consolidated list of the parameters:

```
Name : Solr
Interpreter : jdbc
default.url : jdbc:solr://SOLR_ZK_CONNECTION_STRING?collection=<collection_name>
default.driver : org.apache.solr.client.solrj.io.sql.DriverImpl
default.user : solr
dependency : org.apache.solr:solr-solrj:8.1.0
```

## Query with the Notebook



For some notebooks, the JDBC interpreter will not be bound to the notebook by default. Instructions on how to bind the JDBC interpreter to a notebook are available [here](#).

The screenshot shows the Zeppelin notebook interface with the following sections:

- Solr JDBC**: A pie chart with two segments labeled "a1" and "a2".
- Results of Solr query**: A table with three columns: fielda, fieldb, and fieldd\_s.

fielda	fieldb	fieldd_s
a1	b1	d1
a2	b2	d1
a1	b3	null
a1	b4	d2
a2	b2	d2

### Results of Solr query

The below code block assumes that the Apache Solr driver is setup as the default JDBC interpreter driver. If that is not the case, instructions for using a different prefix is available [here](#).

```
%jdbc
select fielda, fieldb, from test limit 10
```

## Solr JDBC - Python/Jython

Solr's JDBC driver supports Python and Jython.

## Python

Python supports accessing JDBC using the [JayDeBeApi](#) library. The CLASSPATH variable must be configured to contain the solr-solrj jar and the supporting solrj-lib jars.

### JayDeBeApi

#### *run.sh*

```
#!/usr/bin/env bash
Java must already be installed

pip install JayDeBeApi

export CLASSPATH=$(echo $(ls /opt/solr/dist/solr-solrj* /opt/solr/dist/solrj-lib/*) | tr ' '
':')

python solr_jaydebeapi.py
```

#### *solr\_jaydebeapi.py*

```
#!/usr/bin/env python

https://pypi.python.org/pypi/JayDeBeApi/

import jaydebeapi
import sys
if __name__ == '__main__':
 jdbc_url = "jdbc:solr://localhost:9983?collection=test"
 driverName = "org.apache.solr.client.solrj.io.sql.DriverImpl"
 statement = "select fielda, fieldb, fieldc, fieldd_s, fielde_i from test limit 10"

 conn = jaydebeapi.connect(driverName, jdbc_url)
 curs = conn.cursor()
 curs.execute(statement)
 print(curs.fetchall())

 conn.close()

 sys.exit(0)
```

## Jython

Jython supports accessing JDBC natively with Java interfaces or with the zxJDBC library. The CLASSPATH variable must be configured to contain the solr-solrj jar and the supporting solrj-lib jars.

### run.sh

```
#!/usr/bin/env bash
Java and Jython must already be installed

export CLASSPATH=$(echo $(ls /opt/solr/dist/solr-solrj* /opt/solr/dist/solrj-lib/*) | tr ' '
':')

python [solr_java_native.py | solr_zxjdbc.py]
```

## Java Native

### *solr\_java\_native.py*

```
#!/usr/bin/env jython

http://www.jython.org/jythonbook/en/1.0/DatabasesAndJython.html
https://wiki.python.org/jython/DatabaseExamples#SQLite_using_JDBC

import sys

from java.lang import Class
from java.sql import DriverManager, SQLException

if __name__ == '__main__':
 jdbc_url = "jdbc:solr://localhost:9983?collection=test"
 driverName = "org.apache.solr.client.solrj.io.sql.DriverImpl"
 statement = "select fielda, fieldb, fieldc, fieldd_s, field_e_i from test limit 10"

 dbConn = DriverManager.getConnection(jdbc_url)
 stmt = dbConn.createStatement()

 resultSet = stmt.executeQuery(statement)
 while resultSet.next():
 print(resultSet.getString("fielda"))

 resultSet.close()
 stmt.close()
 dbConn.close()

 sys.exit(0)
```

## zxJDBC

### *solr\_zxjdbc.py*

```
#!/usr/bin/env python

http://www.jython.org/jythonbook/en/1.0/DatabasesAndJython.html
https://wiki.python.org/jython/DatabaseExamples#SQLite_using_ziclix

import sys

from com.ziclix.python.sql import zxJDBC

if __name__ == '__main__':
 jdbc_url = "jdbc:solr://localhost:9983?collection=test"
 driverName = "org.apache.solr.client.solrj.io.sql.DriverImpl"
 statement = "select fielda, fieldb, fieldc, fieldd_s, fielde_i from test limit 10"

 with zxJDBC.connect(jdbc_url, None, None, driverName) as conn:
 with conn:
 with conn.cursor() as c:
 c.execute(statement)
 print(c.fetchall())

 sys.exit(0)
```

## Solr JDBC - R

R supports accessing JDBC using the [RJDBC](#) library.

### RJDBC

#### *run.sh*

```
#!/usr/bin/env bash

Java must already be installed and R configured with `R CMD javareconf`

Rscript -e 'install.packages("RJDBC", dep=TRUE)'
Rscript solr_rjdbc.R
```

### *solr\_rjabc.R*

```
https://www.rforge.net/RJDBC/

library("RJDBC")

solrCP <- c(list.files('/opt/solr/dist/solrj-lib', full.names=TRUE), list.files('/opt/solr/dist', pattern='solrj', full.names=TRUE, recursive = TRUE))

drv <- JDBC("org.apache.solr.client.solrj.io.sql.DriverImpl",
 solrCP,
 identifier.quote="`")
conn <- dbConnect(drv, "jdbc:solr://localhost:9983?collection=test", "user", "pwd")

dbGetQuery(conn, "select fielda, fieldb, fieldc, fieldd_s, fielde_i from test limit 10")

dbDisconnect(conn)
```

# Analytics Component

The Analytics Component allows users to calculate complex statistical aggregations over result sets.

The component enables interacting with data in a variety of ways, both through a diverse set of analytics functions as well as powerful faceting functionality. The standard facets are supported within the analytics component with additions that leverage its analytical capabilities.

## Analytics Configuration

The Analytics component is in a contrib module, therefore it will need to be enabled in the `solrconfig.xml` for each collection where you would like to use it.

Since the Analytics framework is a *search component*, it must be declared as such and added to the search handler.

For distributed analytics requests over cloud collections, the component uses the `AnalyticsHandler` strictly for inter-shard communication. The Analytics Handler should not be used by users to submit analytics requests.

To configure Solr to use the Analytics Component, the first step is to add a `<lib/>` directive so Solr loads the Analytic Component classes (for more about the `<lib/>` directive, see [Lib Directives in SolrConfig](#)). In the section of `solrconfig.xml` where the default `<lib/>` directives are, add a line:

```
<lib dir="${solr.install.dir:../../../../../}dist/" regex="solr-analytics-\d.*\.jar" />
```

Next you need to enable the request handler and search component. Add the following lines to `solrconfig.xml`, near the definitions for other request handlers:

### `solrconfig.xml`

```
<!-- To handle user requests -->
<searchComponent name="analytics" class="org.apache.solr.handler.component.AnalyticsComponent" />

<requestHandler name="/select" class="solr.SearchHandler">
 <arr name="last_components">
 <str>analytics</str>
 </arr>
</requestHandler>

<!-- For inter-shard communication during distributed requests -->
<requestHandler name="/analytics" class="org.apache.solr.handler.AnalyticsHandler" />
```

For these changes to take effect, restart Solr or reload the core or collection.

## Request Syntax

An Analytics request is passed to Solr with the parameter `analytics` in a request sent to the [Search Handler](#). Since the analytics request is sent inside of a search handler request, it will compute results based on the

result set determined by the search handler.

For example, this curl command encodes and POSTs a simple analytics request to the the search handler:

```
curl --data-urlencode 'analytics={
 "expressions" : {
 "revenue" : "sum(mult(price,quantity))"
 }
}'
http://localhost:8983/solr/sales/select?q=*&wt=json&rows=0
```

There are 3 main parts of any analytics request:

## Expressions

A list of calculations to perform over the entire result set. Expressions aggregate the search results into a single value to return. This list is entirely independent of the expressions defined in each of the groupings. Find out more about them in the section [Expressions](#).

## Functions

One or more [Variable Functions](#) to be used throughout the rest of the request. These are essentially lambda functions and can be combined in a number of ways. These functions for the expressions defined in expressions as well as groupings.

## Groupings

The list of [Groupings](#) to calculate in addition to the expressions. Groupings hold a set of facets and a list of expressions to compute over those facets. The expressions defined in a grouping are only calculated over the facets defined in that grouping.

### *Optional Parameters*



Either the expressions or the groupings parameter must be present in the request, or else there will be no analytics to compute. The functions parameter is always optional.

### Example Analytics Request

```
{
 "functions": {
 "sale()": "mult(price,quantity)"
 },
 "expressions" : {
 "max_sale" : "max(sale())",
 "med_sale" : "median(sale())"
 },
 "groupings" : {
 "sales" : {
 "expressions" : {
 "stddev_sale" : "stddev(sale())",
 "min_price" : "min(price)",
 "max_quantity" : "max(quantity)"
 },
 "facets" : {
 "category" : {
 "type" : "value",
 "expression" : "fill_missing(category, 'No Category')",
 "sort" : {
 "criteria" : [
 {
 "type" : "expression",
 "expression" : "min_price",
 "direction" : "ascending"
 },
 {
 "type" : "facetvalue",
 "direction" : "descending"
 }
],
 "limit" : 10
 }
 }
 },
 "temps" : {
 "type" : "query",
 "queries" : {
 "hot" : "temp:[90 TO *]",
 "cold" : "temp:[* TO 50]"
 }
 }
 }
 }
}
```

## Expressions

Expressions are the way to request pieces of information from the analytics component. These are the

statistical expressions that you want computed and returned in your response.

## Constructing an Expression

### Expression Components

An expression is built using fields, constants, mapping functions and reduction functions. The ways that these can be defined are described below.

#### Sources

- Constants: The values defined in the expression. The supported constant types are described in the [Analytics Expression Source Reference](#).
- Fields: Solr fields that are read from the index. The supported fields are listed in the [Analytics Expression Source Reference](#).

#### Mapping Functions

Mapping functions map values for each Solr Document or Reduction. The provided mapping functions are detailed in the [Analytics Mapping Function Reference](#).

- Unreduced Mapping: Mapping a Field with another Field or Constant returns a value for every Solr Document. Unreduced mapping functions can take fields, constants as well as other unreduced mapping functions as input.
- Reduced Mapping: Mapping a Reduction Function with another Reduction Function or Constant returns a single value.

#### Reduction Functions

Functions that reduce the values of sources and/or unreduced mapping functions for every Solr Document to a single value. The provided reduction functions are detailed in the [Analytics Reduction Function Reference](#).

### Component Ordering

The expression components must be used in the following order to create valid expressions.

1. Reduced Mapping Function
  - a. Constants
  - b. Reduction Function
    - i. Sources
    - ii. Unreduced Mapping Function
      - A. Sources
      - B. Unreduced Mapping Function
  - c. Reduced Mapping Function
2. Reduction Function

This ordering is based on the following rules:

- No reduction function can be an argument of another reduction function. Since all reduction is done

together in one step, one reduction function cannot rely on the result of another.

- No fields can be left unreduced, since the analytics component cannot return a list of values for an expression (one for every document). Every expression must be reduced to a single value.
- Mapping functions are not necessary when creating functions, however as many nested mappings as needed can be used.
- Nested mapping functions must be the same type, so either both must be unreduced or both must be reduced. A reduced mapping function cannot take an unreduced mapping function as a parameter and vice versa.

## Example Construction

With the above definitions and ordering, an example expression can be broken up into its components:

```
div(sum(a,fill_missing(b,0)),add(10.5,count(mult(a,c))))
```

As a whole, this is a reduced mapping function. The div function is a reduced mapping function since it is a [provided mapping function](#) and has reduced arguments.

If we break down the expression further:

- sum(a,fill\_missing(b,0)): Reduction Function  
sum is a [provided reduction function](#).
  - a: Field
  - fill\_missing(b,0): Unreduced Mapping Function  
fill\_missing is an unreduced mapping function since it is a [provided mapping function](#) and has a field argument.
    - b: Field
    - 0: Constant
- add(10.5,count(mult(a,c))): Reduced Mapping Function  
add is a reduced mapping function since it is a [provided mapping function](#) and has a reduction function argument.
  - 10.5: Constant
  - count(mult(a,c))): Reduction Function  
count is a [provided reduction function](#)
    - mult(a,c): Unreduced Mapping Function  
mult is an unreduced mapping function since it is a [provided mapping function](#) and has two field arguments.
      - a: Field
      - c: Field

## Expression Cardinality (Multi-Valued and Single-Valued)

The root of all multi-valued expressions are multi-valued fields. Single-valued expressions can be started with constants or single-valued fields. All single-valued expressions can be treated as multi-valued

expressions that contain one value.

Single-valued expressions and multi-valued expressions can be used together in many mapping functions, as well as multi-valued expressions being used alone, and many single-valued expressions being used together. For example:

`add(<single-valued double>, <single-valued double>, ...)`

Returns a single-valued double expression where the value of the values of each expression are added.

`add(<single-valued double>, <multi-valued double>)`

Returns a multi-valued double expression where each value of the second expression is added to the single value of the first expression.

`add(<multi-valued double>, <single-valued double>)`

Acts the same as the above function.

`add(<multi-valued double>)`

Returns a single-valued double expression which is the sum of the multiple values of the parameter expression.

## Types and Implicit Casting

The new analytics component currently supports the types listed in the below table. These types have one-way implicit casting enabled for the following relationships:

Type	Implicitly Casts To
<b>Boolean</b>	String
<b>Date</b>	Long, String
<b>Integer</b>	Long, Float, Double, String
<b>Long</b>	Double, String
<b>Float</b>	Double, String
<b>Double</b>	String
<b>String</b>	<i>none</i>

An implicit cast means that if a function requires a certain type of value as a parameter, arguments will be automatically converted to that type if it is possible.

For example, `concat()` only accepts string parameters and since all types can be implicitly cast to strings, any type is accepted as an argument.

This also goes for dynamically typed functions. `fill_missing()` requires two arguments of the same type. However, two types that implicitly cast to the same type can also be used.

For example, `fill_missing(<long>,<float>)` will be cast to `fill_missing(<double>,<double>)` since long cannot be cast to float and float cannot be cast to long implicitly.

There is an ordering to implicit casts, where the more specialized type is ordered ahead of the more general

type. Therefore even though both long and float can be implicitly cast to double and string, they will be cast to double. This is because double is a more specialized type than string, which every type can be cast to.

The ordering is the same as their order in the above table.

Cardinality can also be implicitly cast. Single-valued expressions can always be implicitly cast to multi-valued expressions, since all single-valued expressions are multi-valued expressions with one value.

Implicit casting will only occur when an expression will not "compile" without it. If an expression follows all typing rules initially, no implicit casting will occur. Certain functions such as `string()`, `date()`, `round()`, `floor()`, and `ceil()` act as explicit casts, declaring the type that is desired. However `round()`, `floor()` and `ceil()` can return either int or long, depending on the argument type.

## Variable Functions

Variable functions are a way to shorten your expressions and make writing analytics queries easier. They are essentially lambda functions defined in a request.

### *Example Basic Function*

```
{
 "functions" : {
 "sale()" : "mult(price,quantity)"
 },
 "expressions" : {
 "max_sale" : "max(sale())",
 "med_sale" : "median(sale())"
 }
}
```

In the above request, instead of writing `mult(price,quantity)` twice, a function `sale()` was defined to abstract this idea. Then that function was used in the multiple expressions.

Suppose that we want to look at the sales of specific categories:

```
{
 "functions" : {
 "clothing_sale()" : "filter(mult(price,quantity),equal(category,'Clothing'))",
 "kitchen_sale()" : "filter(mult(price,quantity),equal(category,\\"Kitchen\\"))"
 },
 "expressions" : {
 "max_clothing_sale" : "max(clothing_sale())"
 , "med_clothing_sale" : "median(clothing_sale())"
 , "max_kitchen_sale" : "max(kitchen_sale())"
 , "med_kitchen_sale" : "median(kitchen_sale())"
 }
}
```

## Arguments

Instead of making a function for each category, it would be much easier to use category as an input to the sale() function. An example of this functionality is shown below:

### *Example Function with Arguments*

```
{
 "functions" : {
 "sale(cat)" : "filter(mult(price,quantity),equal(category,cat))"
 },
 "expressions" : {
 "max_clothing_sale" : "max(sale('Clothing'))"
 , "med_clothing_sale" : "median(sale('Clothing'))"
 , "max_kitchen_sale" : "max(sale('Kitchen'))"
 , "med_kitchen_sale" : "median(sale('Kitchen'))"
 }
}
```

Variable Functions can take any number of arguments and use them in the function expression as if they were a field or constant.

## Variable Length Arguments

There are analytics functions that take a variable amount of parameters. Therefore there are use cases where variable functions would need to take a variable amount of parameters.

For example, maybe there are multiple, yet undetermined, number of components to the price of a product. Functions can take a variable length of parameters if the last parameter is followed by ..

### *Example Function with a Variable Length Argument*

```
{
 "functions" : {
 "sale(cat, costs..)" : "filter(mult(add(costs),quantity),equal(category,cat))"
 },
 "expressions" : {
 "max_clothing_sale" : "max(sale('Clothing', material, tariff, tax))"
 , "med_clothing_sale" : "median(sale('Clothing', material, tariff, tax))"
 , "max_kitchen_sale" : "max(sale('Kitchen', material, construction))"
 , "med_kitchen_sale" : "median(sale('Kitchen', material, construction))"
 }
}
```

In the above example a variable length argument is used to encapsulate all of the costs to use for a product. There is no definite number of arguments requested for the variable length parameter, therefore the clothing expressions can use 3 and the kitchen expressions can use 2. When the sale() function is called, costs is expanded to the arguments given.

Therefore in the above request, inside of the sale function:

- add(costs)

is expanded to both of the following:

- add(material, tariff, tax)
- add(material, construction)

## For-Each Functions



### Advanced Functionality

The following function details are for advanced requests.

Although the above functionality allows for an undefined number of arguments to be passed to a function, it does not allow for interacting with those arguments.

Many times we might want to wrap each argument in additional functions. For example maybe we want to be able to look at multiple categories at the same time. So we want to see if category EQUALS x OR category EQUALS y and so on.

In order to do this we need to use for-each lambda functions, which transform each value of the variable length parameter. The for-each is started with the : character after the variable length parameter.

#### Example Function with a For-Each

```
{
 "functions" : {
 "sale(cats..)" : "filter(mult(price,quantity),or(cats:equal(category,_)))"
 },
 "expressions" : {
 "max_sale_1" : "max(sale('Clothing', 'Kitchen'))"
 , "med_sale_1" : "median(sale('Clothing', 'Kitchen'))"
 , "max_sale_2" : "max(sale('Electronics', 'Entertainment', 'Travel'))"
 , "med_sale_2" : "median(sale('Electronics', 'Entertainment', 'Travel'))"
 }
}
```

In this example, cats: is the syntax that starts a for-each lambda function over every parameter cats, and the \_ character is used to refer to the value of cats in each iteration in the for-each. When sale("Clothing", "Kitchen") is called, the lambda function equal(category,\_) is applied to both Clothing and Kitchen inside of the or() function.

Using all of these rules, the expression:

```
`sale("Clothing", "Kitchen")`
```

is expanded to:

```
`filter(mult(price,quantity),or(equal(category,"Kitchen"),equal(category,"Clothing")))`
```

by the expression parser.

## Groupings And Facets

Facets, much like in other parts of Solr, allow analytics results to be broken up and grouped by attributes of the data that the expressions are being calculated over.

The currently available facets for use in the analytics component are Value Facets, Pivot Facets, Range Facets and Query Facets. Each facet is required to have a unique name within the grouping it is defined in, and no facet can be defined outside of a grouping.

Groupings allow users to calculate the same grouping of expressions over a set of facets. Groupings must have both expressions and facets given.

### Example Base Facet Request

```
{
 "functions" : {
 "sale()" : "mult(price,quantity)"
 },
 "groupings" : {
 "sales_numbers" : {
 "expressions" : {
 "max_sale" : "max(sale())",
 "med_sale" : "median(sale())"
 },
 "facets" : {
 "<name>" : "< facet request >"
 }
 }
 }
}
```

### Example Base Facet Response

```
{
 "analytics_response" : {
 "groupings" : {
 "sales_numbers" : {
 "<name>" : "< facet response >"
 }
 }
 }
}
```

## Facet Sorting

Some Analytics facets allow for complex sorting of their results. The two current sortable facets are [Analytic Value Facets](#) and [Analytic Pivot Facets](#).

## Parameters

### criteria

The list of criteria to sort the facet by.

It takes the following parameters:

#### type

The type of sort. There are two possible values:

- expression: Sort by the value of an expression defined in the same grouping.
- facetvalue: Sort by the string-representation of the facet value.

#### Direction

(Optional) The direction to sort.

- ascending (Default)
- descending

#### expression

When type = expression, the name of an expression defined in the same grouping.

### limit

Limit the number of returned facet values to the top N. (Optional)

### offset

When a limit is set, skip the top N facet values. (Optional)

## Example Sort Request

```
{
 "criteria" : [
 {
 "type" : "expression",
 "expression" : "max_sale",
 "direction" : "ascending"
 },
 {
 "type" : "facetvalue",
 "direction" : "descending"
 }
],
 "limit" : 10,
 "offset" : 5
}
```

## Value Facets

Value Facets are used to group documents by the value of a mapping expression applied to each document. Mapping expressions are expressions that do not include a reduction function.

For more information, refer to the [Expressions section](#).

- `mult(quantity, sum(price, tax))`: breakup documents by the revenue generated
- `fillmissing(state, "N/A")`: breakup documents by state, where N/A is used when the document doesn't contain a state

Value Facets can be sorted.

## Parameters

`expression`

The expression to choose a facet bucket for each document.

`sort`

A `sort` for the results of the pivot.



### *Optional Parameters*

The sort parameter is optional.

## Example Value Facet Request

```
{
 "type" : "value",
 "expression" : "fillmissing(category,'No Category')",
 "sort" : {}
}
```

## Example Value Facet Response

```
[
 { "...": "..."},
 {
 "value" : "Electronics",
 "results" : {
 "max_sale" : 103.75,
 "med_sale" : 15.5
 }
 },
 {
 "value" : "Kitchen",
 "results" : {
 "max_sale" : 88.25,
 "med_sale" : 11.37
 }
 },
 { "...": "..."}
]
```

### Field Facets



This is a replacement for Field Facets in the original Analytics Component. Field Facet functionality is maintained in Value Facets by using the name of a field as the expression.

## Analytic Pivot Facets

Pivot Facets are used to group documents by the value of multiple mapping expressions applied to each document.

Pivot Facets work much like layers of [Analytic Value Facets](#). A list of pivots is required, and the order of the list directly impacts the results returned. The first pivot given will be treated like a normal value facet. The second pivot given will be treated like one value facet for each value of the first pivot. Each of these second-level value facets will be limited to the documents in their first-level facet bucket. This continues for however many pivots are provided.

Sorting is enabled on a per-pivot basis. This means that if your top pivot has a sort with `limit:1`, then only that first value of the facet will be drilled down into. Sorting in each pivot is independent of the other pivots.

### Parameters

#### pivots

The list of pivots to calculate a drill-down facet for. The list is ordered by top-most to bottom-most level.

#### name

The name of the pivot.

#### expression

The expression to choose a facet bucket for each document.

#### sort

A [sort](#) for the results of the pivot.

### Optional Parameters



The sort parameter within the pivot object is optional, and can be given in any, none or all of the provided pivots.

### Example Pivot Facet Request

```
{
 "type" : "pivot",
 "pivots" : [
 {
 "name" : "country",
 "expression" : "country",
 "sort" : {}
 },
 {
 "name" : "state",
 "expression" : "fillmissing(state, fillmissing(providence, territory))"
 },
 {
 "name" : "city",
 "expression" : "fillmissing(city, 'N/A')",
 "sort" : {}
 }
]
}
```

### Example Pivot Facet Response

```
[
 { "...": "..."},
 {
 "pivot" : "Country",
 "value" : "USA",
 "results" : {
 "max_sale" : 103.75,
 "med_sale" : 15.5
 },
 "children" : [
 { "...": "..."},
 {
 "pivot" : "State",
 "value" : "Texas",
 "results" : {
 "max_sale" : 99.2,
 "med_sale" : 20.35
 },
 "children" : [
 { "...": "..."},
 {
 "pivot" : "City",
 "value" : "Austin",
 "results" : {
 "max_sale" : 94.34,
 "med_sale" : 17.60
 }
 },
 { "...": "..."}
]
 },
 { "...": "..."}
]
 },
 { "...": "..."}
]
```

## Analytics Range Facets

Range Facets are used to group documents by the value of a field into a given set of ranges. The inputs for analytics range facets are identical to those used for Solr range facets. Refer to the [Range Facet documentation](#) for additional questions regarding use.

### Parameters

#### field

Field to be faceted over

#### start

The bottom end of the range

end

The top end of the range

gap

A list of range gaps to generate facet buckets. If the buckets do not add up to fit the start to end range, then the last gap value will be repeated as many times as needed to fill any unused range.

hardend

Whether to cutoff the last facet bucket range at the end value if it spills over. Defaults to false.

include

The boundaries to include in the facet buckets. Defaults to lower.

- lower - All gap-based ranges include their lower bound.
- upper - All gap-based ranges include their upper bound.
- edge - The first and last gap ranges include their edge bounds (lower for the first one, upper for the last one) even if the corresponding upper/lower option is not specified.
- outer - The before and after ranges will be inclusive of their bounds, even if the first or last ranges already include those boundaries.
- all - Includes all options: lower, upper, edge, and outer

others

Additional ranges to include in the facet. Defaults to none.

- before - All records with field values lower than the lower bound of the first range.
- after - All records with field values greater than the upper bound of the last range.
- between - All records with field values between the lower bound of the first range and the upper bound of the last range.
- none - Include facet buckets for none of the above.
- all - Include facet buckets for before, after and between.



#### *Optional Parameters*

The hardend, include and others parameters are all optional.

### Example Range Facet Request

```
{
 "type" : "range",
 "field" : "price",
 "start" : "0",
 "end" : "100",
 "gap" : [
 "5",
 "10",
 "10",
 "25"
,
 "hardend" : true,
 "include" : [
 "lower",
 "upper"
,
 "others" : [
 "after",
 "between"
,
]
}
```

### Example Range Facet Response

```
[
 {
 "value" : "[0 TO 5]",
 "results" : {
 "max_sale" : 4.75,
 "med_sale" : 3.45
 }
 },
 {
 "value" : "[5 TO 15]",
 "results" : {
 "max_sale" : 13.25,
 "med_sale" : 10.20
 }
 },
 {
 "value" : "[15 TO 25]",
 "results" : {
 "max_sale" : 22.75,
 "med_sale" : 18.50
 }
 },
 {
 "value" : "[25 TO 50]",
 "results" : {
 "max_sale" : 47.55,
 "med_sale" : 30.33
 }
 },
 {
 "value" : "[50 TO 75]",
 "results" : {
 "max_sale" : 70.25,
 "med_sale" : 64.54
 }
 },
 { "..." : "..." }
]
```

## Query Facets

Query Facets are used to group documents by given set of queries.

### Parameters

#### queries

The list of queries to facet by.

### Example Query Facet Request

```
{
 "type" : "query",
 "queries" : {
 "high_quantity" : "quantity:[5 TO 14] AND price:[100 TO *]",
 "low_quantity" : "quantity:[1 TO 4] AND price:[100 TO *]"
 }
}
```

### Example Query Facet Response

```
[
 {
 "value" : "high_quantity",
 "results" : {
 "max_sale" : 4.75,
 "med_sale" : 3.45
 }
 },
 {
 "value" : "low_quantity",
 "results" : {
 "max_sale" : 13.25,
 "med_sale" : 10.20
 }
 }
]
```

## Analytics Expression Sources

Expression sources are the source of the data being aggregated in [analytics expressions](#).

These sources can be either Solr fields indexed with docValues, or constants.

### Supported Field Types

The following [Solr field types](#) are supported. Fields of these types can be either multi-valued and single-valued.

All fields used in analytics expressions **must** have [docValues](#) enabled.

#### String

StrField

#### Boolean

BoolField

#### Integer

TrieIntField

IntPointField

## Long

TrieLongField  
LongPointField

## Float

TrieFloatField  
FloatPointField

## Double

TrieDoubleField  
DoublePointField

## Date

TrieDateField  
DatePointField

### *Multi-valued Field De-duplication*



All multi-valued field types, except for PointFields, are de-duplicated, meaning duplicate values for the same field are removed during indexing. In order to save duplicates, you must use PointField types.

## Constants

Constants can be included in expressions to use along side fields and functions. The available constants are shown below. Constants do not need to be surrounded by any function to define them, they can be used exactly like fields in an expression.

## Strings

There are two possible ways of specifying constant strings, as shown below.

- Surrounded by double quotes, inside the quotes both " and \ must be escaped with a \ character.

"Inside of 'double' \\ \"quotes\" => Inside of 'double' \\ \"quotes"

- Surrounded by single quotes, inside the quotes both ' and \ must be escaped with a \ character.

'Inside of "single" \\ \'quotes\' => Inside of "double" \\ 'quotes'

## Dates

Dates can be specified in the same way as they are in Solr queries. Just use ISO-8601 format. For more information, refer to the [Working with Dates](#) section.

- 2017-07-17T19:35:08Z

## Numeric

Any non-decimal number will be read as an integer, or as a long if it is too large for an integer. All decimal

numbers will be read as doubles.

- -123421: Integer
- 800000000000: Long
- 230.34: Double

## Analytics Mapping Functions

Mapping functions map values for each Solr Document or Reduction.

Below is a list of all mapping functions provided by the Analytics Component. These mappings can be chained together to implement more complex functionality.

### Numeric Functions

#### Negation

Negates the result of a numeric expression.

```
neg(<_Numeric_ T>) => <T>
 • neg(10.53) => -10.53
 • neg([1, -4]) => [-1, 4]
```

#### Absolute Value

Returns the absolute value of the numeric expression.

```
abs(< Numeric T >) => < T >
 • abs(-10.53) => 10.53
 • abs([1, -4]) => [1, 4]
```

#### Round

Rounds the numeric expression to the nearest Integer or Long value.

```
round(< Float >) => < Int >
round(< Double >) => < Long >
 • round(-1.5) => -1
 • round([1.75, 100.34]) => [2, 100]
```

#### Ceiling

Rounds the numeric expression to the nearest Integer or Long value that is greater than or equal to the original value.

```
ceil(< Float >) => < Int >
ceil(< Double >) => < Long >
 • ceil(5.01) => 5
 • ceil([-4.999, 6.99]) => [-4, 7]
```

## Floor

Rounds the numeric expression to the nearest Integer or Long value that is less than or equal to the original value.

```
floor(< Float >) => < Int >
floor(< Double >) => < Long >
 • floor(5.75) => 5
 • floor([-4.001, 6.01]) => [-5, 6]
```

## Addition

Adds the values of the numeric expressions.

```
add(< Multi Double >) => < Single Double >
 • add([1, -4]) => -3.0

add(< Single Double >, < Multi Double >) => < Multi Double >
 • add(3.5, [1, -4]) => [4.5, -0.5]

add(< Multi Double >, < Single Double >) => < Multi Double >
 • add([1, -4], 3.5) => [4.5, -0.5]

add(< Single Double >, ...) => < Single Double >
 • add(3.5, 100, -27.6) => 75.9
```

## Subtraction

Subtracts the values of the numeric expressions.

```
sub(< Single Double >, < Single Double >) => < Single Double >
 • sub(3.5, 100) => -76.5

sub(< Single Double >, < Multi Double >) => < Multi Double >
 • sub(3.5, [1, -4]) => [2.5, 7.5]

sub(< Multi Double >, < Single Double >) => < Multi Double >
 • sub([1, -4], 3.5) => [-2.5, -7.5]
```

## Multiplication

Multiplies the values of the numeric expressions.

```
mult(< Multi Double >) => < Single Double >
 • mult([1, -4]) => -4.0

mult(< Single Double >, < Multi Double >) => < Multi Double >
 • mult(3.5, [1, -4]) => [3.5, -16.0]

mult(< Multi Double >, < Single Double >) => < Multi Double >
 • mult([1, -4], 3.5) => [3.5, 16.0]
```

`mult(< Single Double >, ...)` => < Single Double >

- `mult(3.5, 100, -27.6)` => -9660

## Division

Divides the values of the numeric expressions.

`div(< Single Double >, < Single Double >)` => < Single Double >

- `div(3.5, 100)` => .035

`div(< Single Double >, < Multi Double >)` => < Multi Double >

- `div(3.5, [1, -4])` => [3.5, -0.875]

`div(< Multi Double >, < Single Double >)` => < Multi Double >

- `div([1, -4], 25)` => [0.04, -0.16]

## Power

Takes one numeric expression to the power of another.

**NOTE:** The square root function `sqrt(< Double >)` can be used as shorthand for `pow(< Double >, .5)`

`pow(< Single Double >, < Single Double >)` => < Single Double >

- `pow(2, 4)` => 16.0

`pow(< Single Double >, < Multi Double >)` => < Multi Double >

- `pow(16, [-1, 0])` => [0.0625, 1]

`pow(< Multi Double >, < Single Double >)` => < Multi Double >

- `pow([1, 16], .25)` => [1.0, 2.0]

## Logarithm

Takes one logarithm of numeric expressions, with an optional second numeric expression as the base. If only one expression is given, the natural log is used.

`log(< Double >)` => < Double >

- `log(5)` => 1.6094...
- `log([1.0, 100.34])` => [0.0, 4.6085...]

`log(< Single Double >, < Single Double >)` => < Single Double >

- `log(2, 4)` => 0.5

`log(< Single Double >, < Multi Double >)` => < Multi Double >

- `log(16, [2, 4])` => [4, 2]

`log(< Multi Double >, < Single Double >)` => < Multi Double >

- `log([81, 3], 9)` => [2.0, 0.5]

## Logic

### Negation

Negates the result of a boolean expression.

```
neg(< Bool >) => < Bool>
```

- neg(F) => T
- neg([F, T]) => [T, F]

### And

ANDs the values of the boolean expressions.

```
and(< Multi Bool >) => < Single Bool >
```

- and([T, F, T]) => F

```
and(< Single Bool >, < Multi Bool >) => < Multi Bool >
```

- and(F, [T, T]) => [F, F]

```
and(< Multi Bool >, < Single Bool >) => < Multi Bool >
```

- and([F, T], T) => [F, T]

```
and(< Single Bool >, ...) => < Single Bool >
```

- and(T, T, T) => T

### Or

ORs the values of the boolean expressions.

```
or(< Multi Bool >) => < Single Bool >
```

- or([T, F, T]) => T

```
or(< Single Bool >, < Multi Bool >) => < Multi Bool >
```

- or(F, [F, T]) => [F, T]

```
or(< Multi Bool >, < Single Bool >) => < Multi Bool >
```

- or([F, T], T) => [T, T]

```
or(< Single Bool >, ...) => < Single Bool >
```

- or(F, F, F) => F

### Exists

Checks whether any value(s) exist for the expression.

```
exists(T) => < Single Bool >
```

- exists([1, 2, 3]) => T
- exists([]) => F
- exists(empty) => F

- exists('abc') => T

## Comparison

### Equality

Checks whether two expressions' values are equal. The parameters must be the same type, after implicit casting.

`equal(< Single T >, < Single T >) => < Single Bool >`

- `equal(F, F) => T`

`equal(< Single T >, < Multi T >) => < Multi Bool >`

- `equal("a", ["a", "ab"]) => [T, F]`

`equal(< Multi T >, < Single T >) => < Multi Bool >`

- `equal([1.5, -3.0], -3) => [F, T]`

### Greater Than

Checks whether a numeric or Date expression's values are greater than another expression's values. The parameters must be the same type, after implicit casting.

`gt(< Single Numeric/Date T >, < Single T >) => < Single Bool >`

- `gt(1800-01-02, 1799-12-20) => F`

`gt(< Single Numeric/Date T >, < Multi T >) => < Multi Bool >`

- `gt(30.756, [30, 100]) => [F, T]`

`gt(< Multi Numeric/Date T >, < Single T >) => < Multi Bool >`

- `gt([30, 75.6], 30) => [F, T]`

### Greater Than or Equals

Checks whether a numeric or Date expression's values are greater than or equal to another expression's values. The parameters must be the same type, after implicit casting.

`gte(< Single Numeric/Date T >, < Single T >) => < Single Bool >`

- `gte(1800-01-02, 1799-12-20) => F`

`gte(< Single Numeric/Date T >, < Multi T >) => < Multi Bool >`

- `gte(30.756, [30, 100]) => [F, T]`

`gte(< Multi Numeric/Date T >, < Single T >) => < Multi Bool >`

- `gte([30, 75.6], 30) => [T, T]`

### Less Than

Checks whether a numeric or Date expression's values are less than another expression's values. The parameters must be the same type, after implicit casting.

```
lt(< Single Numeric/Date T >, < Single T >) => < Single Bool >
 • lt(1800-01-02, 1799-12-20) => T

lt(< Single Numeric/Date T >, < Multi T >) => < Multi Bool >
 • lt(30.756, [30, 100]) => [T, F]

lt(< Multi Numeric/Date T >, < Single T >) => < Multi Bool >
 • lt([30, 75.6], 30) => [F, F]
```

## Less Than or Equals

Checks whether a numeric or Date expression's values are less than or equal to another expression's values. The parameters must be the same type, after implicit casting.

```
lte(< Single Numeric/Date T >, < Single T >) => < Single Bool >
 • lte(1800-01-02, 1799-12-20) => T

lte(< Single Numeric/Date T >, < Multi T >) => < Multi Bool >
 • lte(30.756, [30, 100]) => [T, F]

lte(< Multi Numeric/Date T >, < Single T >) => < Multi Bool >
 • lte([30, 75.6], 30) => [T, F]
```

## Top

Returns the maximum of the numeric, Date or String expression(s)' values. The parameters must be the same type, after implicit casting. (Currently the only type not compatible is Boolean, which will be converted to a String implicitly in order to compile the expression)

```
top(< Multi T >) => < Single T >
 • top([30, 400, -10, 0]) => 400

top(< Single T >, ...) => < Single T >
 • top("a", 1, "d") => "d"
```

## Bottom

Returns the minimum of the numeric, Date or String expression(s)' values. The parameters must be the same type, after implicit casting. (Currently the only type not compatible is Boolean, which will be converted to a String implicitly in order to compile the expression)

```
bottom(< Multi T >) => < Single T >
 • bottom([30, 400, -10, 0]) => -10

bottom(< Single T >, ...) => < Single T >
 • bottom("a", 1, "d") => "1"
```

## Conditional

## If

Returns the value(s) of the THEN or ELSE expressions depending on whether the boolean conditional expression's value is true or false. The THEN and ELSE expressions must be of the same type and cardinality after implicit casting is done.

```
if(< Single Bool>, < T >, < T >) => < T >
 • if(true, "abc", [1,2]) => ["abc"]
 • if(false, "abc", 123) => "123"
```

## Replace

Replace all values from the 1<sup>st</sup> expression that are equal to the value of the 2<sup>nd</sup> expression with the value of the 3<sup>rd</sup> expression. All parameters must be the same type after implicit casting is done.

```
replace(< T >, < Single T >, < Single T >) => < T >
 • replace([1,3], 3, "4") => ["1", "4"]
 • replace("abc", "abc", 18) => "18"
 • replace("abc", 1, "def") => "abc"
```

## Fill Missing

If the 1<sup>st</sup> expression does not have values, fill it with the values for the 2<sup>nd</sup> expression. Both expressions must be of the same type and cardinality after implicit casting is done.

```
fill_missing(< T >, < T >) => < T >
 • fill_missing([], 3) => [3]
 • fill_missing(empty, "abc") => "abc"
 • fill_missing("abc", [1]) => ["abc"]
```

## Remove

Remove all occurrences of the 2<sup>nd</sup> expression's value from the values of the 1<sup>st</sup> expression. Both expressions must be of the same type after implicit casting is done.

```
remove(< T >, < Single T >) => < T >
 • remove([1,2,3,2], 2) => [1, 3]
 • remove("1", 1) => empty
 • remove(1, "abc") => "1"
```

## Filter

Return the values of the 1<sup>st</sup> expression if the value of the 2<sup>nd</sup> expression is true, otherwise return no values.

```
filter(< T >, < Single Boolean >) => < T >
 • filter([1,2,3], true) => [1,2,3]
 • filter([1,2,3], false) => []
 • filter("abc", false) => empty
```

- filter("abc", true) => 1

## Date

### Date Parse

Explicitly converts the values of a String or Long expression into Dates.

date(< String >) => < Date >

- date('1800-01-02') => 1800-01-02T00:00:00Z
- date(['1800-01-02', '2016-05-23']) => [1800-01-02T..., 2016-05-23T...]

date(< Long >) => < Date >

- date(1232343246648) => 2009-01-19T05:34:06Z
- date([1232343246648, 223234324664]) => [2009-01-19T..., 1977-01-27T...]

### Date Math

Compute the given date math strings for the values of a Date expression. The date math strings **must** be constant.

date\_math(< Date >, < Constant String >...) => < Date >

- date\_math(1800-04-15, '+1DAY', '-1MONTH') => 1800-03-16
- date\_math([1800-04-15, 2016-05-24], '+1DAY', '-1MONTH') => [1800-03-16, 2016-04-25]

## String

### Explicit Casting

Explicitly casts the expression to a String expression.

string(< String >) => < String >

- string(1) => '1'
- string([1.5, -2.0]) => ['1.5', '-2.0']

### Concatenation

Concatenates the values of the String expression(s) together.

concat(< Multi String >) => < Single String >

- concat(['a', 'b', 'c']) => 'abc'

concat(< Single String >, < Multi String >) => < Multi String >

- concat(1, ['a', 'b', 'c']) => ['1a', '1b', '1c']

concat(< Multi String >, < Single String >) => < Multi String >

- concat(['a', 'b', 'c'], 1) => ['a1', 'b1', 'c1']

concat(< Single String >...) => < Single String >

- concat('a','b','c') => 'abc'
- concat('a',empty,'c') => 'ac'  
*Empty values are ignored*

## Separated Concatenation

Concatenates the values of the String expression(s) together using the given [constant string](#) value as a separator.

concat\_sep(< Constant String >, < Multi String >) => < Single String >

- concat\_sep('-', ['a','b']) => 'a-b'

concat\_sep(< Constant String >, < Single String >, < Multi String >) => < Multi String >

- concat\_sep(2,1,['a','b']) => ['12a','12b']

concat\_sep(< Constant String >, < Multi String >, < Single String >) => < Multi String >

- concat\_sep(2,['a','b'],1) => ['a21','b21']

- concat\_sep('-', 'a', 2, 3) => 'a-2-3'

- concat\_sep(';', 'a', empty, 'c') => 'a;c'

*Empty values are ignored*

## Analytics Reduction Functions

Reduction functions reduce the values of [sources](#) and/or unreduced [mapping functions](#) for every Solr Document to a single value.

Below is a list of all reduction functions provided by the Analytics Component. These can be combined using mapping functions to implement more complex functionality.

### Counting Reductions

#### Count

The number of existing values for an expression. For single-valued expressions, this is equivalent to docCount. If no expression is given, the number of matching documents is returned.

count() => < Single Long > count(< T >) => < Single Long >

#### Doc Count

The number of documents for which an expression has existing values. For single-valued expressions, this is equivalent to count. If no expression is given, the number of matching documents is returned.

doc\_count() => < Single Long >

doc\_count(< T >) => < Single Long >

#### Missing

The number of documents for which an expression has no existing value.

```
missing(< T >) => < Single Long >
```

## Unique

The number of unique values for an expression. This function accepts Numeric, Date and String expressions.

```
unique(< T >) => < Single Long >
```

## Math Reductions

### Sum

Returns the sum of all values for the expression.

```
sum(< Double >) => < Single Double >
```

### Variance

Returns the variance of all values for the expression.

```
variance(< Double >) => < Single Double >
```

### Standard Deviation

Returns the standard deviation of all values for the expression.

```
stddev(< Double >) => < Single Double >
```

### Mean

Returns the arithmetic mean of all values for the expression.

```
mean(< Double >) => < Single Double >
```

### Weighted Mean

Returns the arithmetic mean of all values for the second expression weighted by the values of the first expression.

```
wmean(< Double >, < Double >) => < Single Double >
```



The expressions must satisfy the rules for `mult` function parameters.

## Ordering Reductions

### Minimum

Returns the minimum value for the expression. This function accepts Numeric, Date and String expressions.

```
min(< T >) => < Single T >
```

## Maximum

Returns the maximum value for the expression. This function accepts Numeric, Date and String expressions.

```
max(< T >) => < Single T >
```

## Median

Returns the median of all values for the expression. This function accepts Numeric and Date expressions.

```
median(< T >) => < Single T >
```

## Percentile

Calculates the given percentile of all values for the expression. This function accepts Numeric, Date and String expressions for the 2<sup>nd</sup> parameter.

The percentile, given as the 1<sup>st</sup> parameter, must be a [constant double](#) between [0, 100].

```
percentile(<Constant Double>, < T >) => < Single T >
```

## Ordinal

Calculates the given ordinal of all values for the expression. This function accepts Numeric, Date and String expressions for the 2<sup>nd</sup> parameter. The ordinal, given as the 1<sup>st</sup> parameter, must be a [constant integer](#). **0 is not accepted as an ordinal value.**

If the ordinal is positive, the returned value will be the  $n^{\text{th}}$  smallest value.

If the ordinal is negative, the returned value will be the  $n^{\text{th}}$  largest value.

```
ordinal(<Constant Int>, < T >) => < Single T >
```

# Streaming Expressions

Streaming Expressions provide a simple yet powerful stream processing language for Solr Cloud.

Streaming expressions are a suite of functions that can be combined to perform many different parallel computing tasks. These functions are the basis for the [Parallel SQL Interface](#).

There is a growing library of functions that can be combined to implement:

- Request/response stream processing
- Batch stream processing
- Fast interactive MapReduce
- Aggregations (Both pushed down faceted and shuffling MapReduce)
- Parallel relational algebra (distributed joins, intersections, unions, complements)
- Publish/subscribe messaging
- Distributed graph traversal
- Machine learning and parallel iterative model training
- Anomaly detection
- Recommendation systems
- Retrieve and rank services
- Text classification and feature extraction
- Streaming NLP
- Statistical Programming

Streams from outside systems can be joined with streams originating from Solr and users can add their own stream functions by following Solr's [Java streaming API](#).



Both streaming expressions and the streaming API are considered experimental, and the APIs are subject to change.

## Stream Language Basics

Streaming Expressions are comprised of streaming functions which work with a Solr collection. They emit a stream of tuples (key/value Maps).

Many of the provided streaming functions are designed to work with entire result sets rather than the top N results like normal search. This is supported by the [/export handler](#).

Some streaming functions act as stream sources to originate the stream flow. Other streaming functions act as stream decorators to wrap other stream functions and perform operations on the stream of tuples. Many streams functions can be parallelized across a worker collection. This can be particularly powerful for relational algebra functions.

## Streaming Requests and Responses

Solr has a /stream request handler that takes streaming expression requests and returns the tuples as a JSON stream. This request handler is implicitly defined, meaning there is nothing that has to be defined in solrconfig.xml - see [Implicit RequestHandlers](#).

The /stream request handler takes one parameter, expr, which is used to specify the streaming expression. For example, this curl command encodes and POSTs a simple search() expression to the /stream handler:

```
curl --data-urlencode 'expr=search(enron_emails,
 q="from:1800flowers*",
 fl="from, to",
 sort="from asc",
 qt="/export")' http://localhost:8983/solr/enron_emails/stream
```

Details of the parameters for each function are included below.

For the above example the /stream handler responded with the following JSON response:

```
{"result-set": {"docs": [
 {"from": "1800flowers.133139412@s2u2.com", "to": "lcampbel@enron.com"},
 {"from": "1800flowers.93690065@s2u2.com", "to": "jtholt@ect.enron.com"},
 {"from": "1800flowers.96749439@s2u2.com", "to": "alewis@enron.com"},
 {"from": "1800flowers@1800flowers.flonetwork.com", "to": "lcampbel@enron.com"},
 {"from": "1800flowers@1800flowers.flonetwork.com", "to": "ebass@enron.com"},
 {"from": "1800flowers@shop2u.com", "to": "lcampbel@enron.com"},
 {"from": "1800flowers@shop2u.com", "to": "lcampbel@enron.com"},
 {"from": "1800flowers@shop2u.com", "to": "lcampbel@enron.com"},
 {"from": "1800flowers@shop2u.com", "to": "ebass@enron.com"},
 {"from": "1800flowers@shop2u.com", "to": "ebass@enron.com"},
 {"from": "1800flowers@shop2u.com", "to": "ebass@enron.com"},
 {"EOF": true, "RESPONSE_TIME": 33}]}
}
```

Note the last tuple in the above example stream is `{"EOF": true, "RESPONSE_TIME": 33}`. The EOF indicates the end of the stream. To process the JSON response, you'll need to use a streaming JSON implementation because streaming expressions are designed to return the entire result set which may have millions of records. In your JSON client you'll need to iterate each doc (tuple) and check for the EOF tuple to determine the end of stream.

The `org.apache.solr.client.solrj.io` package provides Java classes that compile streaming expressions into streaming API objects. These classes can be used to execute streaming expressions from inside a Java application. For example:

```
StreamFactory streamFactory = new DefaultStreamFactory().withCollectionZkHost("collection1",
zkServer.getZkAddress());
InjectionDefense defense = new InjectionDefense("parallel(collection1,
group(search(collection1, q="*:*", fl="id,a_s,a_i,a_f", sort="a_s asc,a_f asc"),
partitionKeys="a_s"), by="a_s asc"), workers="2", zkHost="?/??", sort="a_s asc"));
defense.addParameter(zkhost);
ParallelStream pstream = (ParallelStream)streamFactory.constructStream(defense
.safeExpressionString());
```

Note that `InjectionDefense` need only be used if the string being inserted could contain user supplied data. See the javadoc for `InjectionDefense` for usage details and SOLR-12891 for an example of the potential risks. Also note that for security reasons normal parameter substitution no longer applies to the `expr` parameter unless the jvm has been started with `-DstreamingExpressionMacros=true` (usually via `solr.in.sh`)

## Data Requirements

Because streaming expressions relies on the `/export` handler, many of the field and field type requirements to use `/export` are also requirements for `/stream`, particularly for `sort` and `f1` parameters. Please see the section [Exporting Result Sets](#) for details.

# Types of Streaming Expressions

## About Stream Sources

Stream sources originate streams. The most commonly used one of these is search, which does a query.

A full reference to all available source expressions is available in [Stream Source Reference](#).

## About Stream Decorators

Stream decorators wrap other stream functions or perform operations on a stream.

A full reference to all available decorator expressions is available in [Stream Decorator Reference](#).

## About Stream Evaluators

Stream Evaluators can be used to evaluate (calculate) new values based on other values in a tuple. That newly evaluated value can be put into the tuple (as part of a select(...) clause), used to filter streams (as part of a having(...) clause), and for other things. Evaluators can contain field names, raw values, or other evaluators, giving you the ability to create complex evaluation logic, including conditional if/then choices.

In cases where you want to use raw values as part of an evaluation you will need to consider the order of how evaluators are parsed.

1. If the parameter can be parsed into a valid number, then it is considered a number. For example, add(3,4.5)
2. If the parameter can be parsed into a valid boolean, then it is considered a boolean. For example, eq(true, false)
3. If the parameter can be parsed into a valid evaluator, then it is considered an evaluator. For example, eq(add(10,4),add(7,7))
4. The parameter is considered a field name, even if it quoted. For example, eq(fieldA,"fieldB")

If you wish to use a raw string as part of an evaluation, you will want to consider using the raw(string) evaluator. This will always return the raw value, no matter what is entered.

A full reference to all available evaluator expressions is available in [Stream Evaluator Reference](#).

# Stream Source Reference

## search

The search function searches a SolrCloud collection and emits a stream of tuples that match the query. This is very similar to a standard Solr query, and uses many of the same parameters.

This expression allows you to specify a request handler using the qt parameter. By default, the /select handler is used. The /select handler can be used for simple rapid prototyping of expressions. For production, however, you will most likely want to use the /export handler which is designed to sort and export entire result sets. The /export handler is not used by default because it has stricter requirements than the /select handler so it's not as easy to get started working with. To read more about the /export handler requirements review the section [Exporting Result Sets](#).

### search Parameters

- collection: (Mandatory) the collection being searched.
- q: (Mandatory) The query to perform on the Solr index.
- f1: (Mandatory) The list of fields to return.
- sort: (Mandatory) The sort criteria.
- zkHost: Only needs to be defined if the collection being searched is found in a different zkHost than the local stream handler.
- qt: Specifies the query type, or request handler, to use. Set this to /export to work with large result sets. The default is /select.
- rows: (Mandatory with the /select handler) The rows parameter specifies how many rows to return. This parameter is only needed with the /select handler (which is the default) since the /export handler always returns all rows.
- partitionKeys: Comma delimited list of keys to partition the search results by. To be used with the parallel function for parallelizing operations across worker nodes. See the [parallel](#) function for details.

### search Syntax

```
expr=search(collection1,
 zkHost="localhost:9983",
 qt="/export",
 q="*:*",
 f1="id,a_s,a_i,a_f",
 sort="a_f asc, a_i asc")
```

## jdbc

The jdbc function searches a JDBC datasource and emits a stream of tuples representing the JDBC result set. Each row in the result set is translated into a tuple and each tuple contains all the cell values for that row.

## jdbc Parameters

- connection: (Mandatory) JDBC formatted connection string to whatever driver you are using.
- sql: (Mandatory) query to pass off to the JDBC endpoint
- sort: (Mandatory) The sort criteria indicating how the data coming out of the JDBC stream is sorted
- driver: The name of the JDBC driver used for the connection. If provided then the driver class will attempt to be loaded into the JVM. If not provided then it is assumed that the driver is already loaded into the JVM. Some drivers require explicit loading so this option is provided.
- [driverProperty]: One or more properties to pass to the JDBC driver during connection. The format is `propertyName="PropertyValue"`. You can provide as many of these properties as you'd like and they will all be passed to the connection.

## Connections and Drivers

Because some JDBC drivers require explicit loading the driver parameter can be used to provide the driver class name. If provided, then during stream construction the driver will be loaded. If the driver cannot be loaded because the class is not found on the classpath, then stream construction will fail.

When the JDBC stream is opened it will validate that a driver can be found for the provided connection string. If a driver cannot be found (because it hasn't been loaded) then the open will fail.

## Datatypes

Due to the inherent differences in datatypes across JDBC sources the following datatypes are supported. The table indicates what Java type will be used for a given JDBC type. Types marked as requiring conversion will go through a conversion for each value of that type. For performance reasons the cell data types are only considered when the stream is opened as this is when the converters are created.

JDBC Type	Java Type	Requires Conversion
String	String	No
Short	Long	Yes
Integer	Long	Yes
Long	Long	No
Float	Double	Yes
Double	Double	No
Boolean	Boolean	No

## jdbc Syntax

A basic jdbc expression:

```
jdbc(
 connection="jdbc:hsqldb:mem: .",
 sql="select NAME, ADDRESS, EMAIL, AGE from PEOPLE where AGE > 25 order by AGE, NAME DESC",
 sort="AGE asc, NAME desc",
 driver="org.hsqldb.jdbcDriver"
)
```

A jdbc expression that passes a property to the driver:

```
// get_column_name is a property to pass to the hsqldb driver
jdbc(
 connection="jdbc:hsqldb:mem: .",
 sql="select NAME as FIRST_NAME, ADDRESS, EMAIL, AGE from PEOPLE where AGE > 25 order by AGE,
NAME DESC",
 sort="AGE asc, NAME desc",
 driver="org.hsqldb.jdbcDriver",
 get_column_name="false"
)
```

## echo

The echo function returns a single Tuple echoing its text parameter. Echo is the simplest stream source designed to provide text to a text analyzing stream decorator.

### echo Syntax

```
echo("Hello world")
```

## facet

The facet function provides aggregations that are rolled up over buckets. Under the covers the facet function pushes down the aggregation into the search engine using Solr's JSON Facet API. This provides sub-second performance for many use cases. The facet function is appropriate for use with a low to moderate number of distinct values in the bucket fields. To support high cardinality aggregations see the rollup function.

### facet Parameters

- collection: (Mandatory) Collection the facets will be aggregated from.
- q: (Mandatory) The query to build the aggregations from.
- buckets: (Mandatory) Comma separated list of fields to rollup over. The comma separated list represents the dimensions in a multi-dimensional rollup.
- bucketSorts: (Mandatory) Comma separated list of sorts to apply to each dimension in the buckets parameters. Sorts can be on the computed metrics or on the bucket values.
- rows: (Default 10) The number of rows to return. '-1' will return all rows.

- offset:(Default 0) The offset in the result set to start from.
- overfetch: (Default 150) Over-fetching is used to provide accurate aggregations over high cardinality fields.
- method: The JSON facet API aggregation method.
- bucketSizeLimit: Sets the absolute number of rows to fetch. This is incompatible with rows, offset and overfetch. This value is applied to each dimension. '-1' will fetch all the buckets.
- metrics: List of metrics to compute for the buckets. Currently supported metrics are sum(col), avg(col), min(col), max(col), count(\*) .

## facet Syntax

Example 1:

```
facet(collection1,
 q="*:*",
 buckets="a_s",
 bucketSorts="sum(a_i) desc",
 rows=100,
 sum(a_i),
 sum(a_f),
 min(a_i),
 min(a_f),
 max(a_i),
 max(a_f),
 avg(a_i),
 avg(a_f),
 count(*))
```

The example above shows a facet function with rollups over a single bucket, where the buckets are returned in descending order by the calculated value of the sum(a\_i) metric.

Example 2:

```
facet(collection1,
 q="*:*",
 buckets="year_i, month_i, day_i",
 bucketSorts="year_i desc, month_i desc, day_i desc",
 rows=10,
 offset=20,
 sum(a_i),
 sum(a_f),
 min(a_i),
 min(a_f),
 max(a_i),
 max(a_f),
 avg(a_i),
 avg(a_f),
 count(*))
```

The example above shows a facet function with rollups over three buckets, where the buckets are returned in descending order by bucket value. The rows parameter returns 10 rows and the offset parameter starts returning rows from the 20th row.

## features

The features function extracts the key terms from a text field in a classification training set stored in a SolrCloud collection. It uses an algorithm known as **Information Gain**, to select the important terms from the training set. The features function was designed to work specifically with the [train](#) function, which uses the extracted features to train a text classifier.

The features function is designed to work with a training set that provides both positive and negative examples of a class. It emits a tuple for each feature term that is extracted along with the inverse document frequency (IDF) for the term in the training set.

The features function uses a query to select the training set from a collection. The IDF for each selected feature is calculated relative to the training set matching the query. This allows multiple training sets to be stored in the same SolrCloud collection without polluting the IDF across training sets.

## features Parameters

- **collection:** (Mandatory) The collection that holds the training set
- **q:** (Mandatory) The query that defines the training set. The IDF for the features will be generated specific to the result set matching the query.
- **featureSet:** (Mandatory) The name of the feature set. This can be used to retrieve the features if they are stored in a SolrCloud collection.
- **field:** (Mandatory) The text field to extract the features from.
- **outcome:** (Mandatory) The field that defines the class, positive or negative
- **numTerms:** (Mandatory) How many feature terms to extract.
- **positiveLabel:** (defaults to 1) The value in the outcome field that defines a positive outcome.

## features Syntax

```
features(collection1,
 q="*:*",
 featureSet="features1",
 field="body",
 outcome="out_i",
 numTerms=250)
```

## nodes

The nodes function provides breadth-first graph traversal. For details, see the section [Graph Traversal](#).

## knnSearch

The knnSearch function returns the k-nearest neighbors for a document based on text similarity. Under the covers the knnSearch function uses the More Like This query parser plugin.

### knnSearch Parameters

- collection: (Mandatory) The collection to perform the search in.
- id: (Mandatory) The id of the source document to begin the knn search from.
- qf: (Mandatory) The query field used to compare documents.
- k: (Mandatory) The number of nearest neighbors to return.
- fl: (Mandatory) The field list to return.
- mindf: (Optional) The minimum number of occurrences in the corpus to be included in the search.
- maxdf: (Optional) The maximum number of occurrences in the corpus to be included in the search.
- minwl: (Optional) The minimum word length of to be included in the search.
- maxwl: (Optional) The maximum word length of to be included in the search.

## knnSearch Syntax

```
knnSearch(collection1,
 id="doc1",
 qf="text_field",
 k="10",
 fl="id, title",
 mintf="3",
 maxdf="10000000")
```

## model

The model function retrieves and caches logistic regression text classification models that are stored in a SolrCloud collection. The model function is designed to work with models that are created by the [train](#)

[function](#), but can also be used to retrieve text classification models trained outside of Solr, as long as they conform to the specified format. After the model is retrieved it can be used by the [classify function](#) to classify documents.

A single model tuple is fetched and returned based on the **id** parameter. The model is retrieved by matching the **id** parameter with a model name in the index. If more than one iteration of the named model is stored in the index, the highest iteration is selected.

## Caching with model

The `model` function has an internal LRU (least-recently-used) cache so models do not have to be retrieved with each invocation of the `model` function. The time to cache for each model ID can be passed as a parameter to the function call. Retrieving a cached model does not reset the time for expiring the model ID in the cache.

## Model Storage

The storage format of the models in Solr is below. The `train` function outputs the format below so you only need to know schema details if you plan to use the `model` function with logistic regression models trained outside of Solr.

- `name_s` (Single value, String, Stored): The name of the model.
- `iteration_i` (Single value, Integer, Stored): The iteration number of the model. Solr can store all iterations of the models generated by the `train` function.
- `terms_ss` (Multi value, String, Stored): The array of terms/features of the model.
- `weights_ds` (Multi value, double, Stored): The array of term weights. Each weight corresponds by array index to a term.
- `idfs_ds` (Multi value, double, Stored): The array of term IDFs (Inverse document frequency). Each IDF corresponds by array index to a term.

## model Parameters

- `collection:` (Mandatory) The collection where the model is stored.
- `id:` (Mandatory) The id/name of the model. The `model` function always returns one model. If there are multiple iterations of the name, the highest iteration is returned.
- `cacheMillis:` (Optional) The amount of time to cache the model in the LRU cache.

## model Syntax

```
model(modelCollection,
 id="myModel"
 cacheMillis="200000")
```

## random

The `random` function searches a SolrCloud collection and emits a pseudo-random set of results that match

the query. Each invocation of random will return a different pseudo-random result set.

## random Parameters

- collection: (Mandatory) The collection the stats will be aggregated from.
- q: (Mandatory) The query to build the aggregations from.
- rows: (Mandatory) The number of pseudo-random results to return.
- fl: (Mandatory) The field list to return.
- fq: (Optional) Filter query

## random Syntax

```
random(baskets,
 q="productID:productX",
 rows="100",
 fl="basketID")
```

In the example above the random function is searching the baskets collections for all rows where "productID:productX". It will return 100 pseudo-random results. The field list returned is the basketID.

## significantTerms

The significantTerms function queries a SolrCloud collection, but instead of returning documents, it returns significant terms found in documents in the result set. The significantTerms function scores terms based on how frequently they appear in the result set and how rarely they appear in the entire corpus. The significantTerms function emits a tuple for each term which contains the term, the score, the foreground count and the background count. The foreground count is how many documents the term appears in in the result set. The background count is how many documents the term appears in in the entire corpus. The foreground and background counts are global for the collection.

## significantTerms Parameters

- collection: (Mandatory) The collection that the function is run on.
- q: (Mandatory) The query that describes the foreground document set.
- field: (Mandatory) The field to extract the terms from.
- limit: (Optional, Default 20) The max number of terms to return.
- minDocFreq: (Optional, Defaults to 5 documents) The minimum number of documents the term must appear in on a shard. This is a float value. If greater than 1.0 then it's considered the absolute number of documents. If less than 1.0 it's treated as a percentage of documents.
- maxDocFreq: (Optional, Defaults to 30% of documents) The maximum number of documents the term can appear in on a shard. This is a float value. If greater than 1.0 then it's considered the absolute number of documents. If less than 1.0 it's treated as a percentage of documents.
- minTermLength: (Optional, Default 4) The minimum length of the term to be considered significant.

## significantTerms Syntax

```
significantTerms(collection1,
 q="body:Solr",
 field="author",
 limit="50",
 minDocFreq="10",
 maxDocFreq=".20",
 minTermLength="5")
```

In the example above the `significantTerms` function is querying `collection1` and returning at most 50 significant terms from the `authors` field that appear in 10 or more documents but not more than 20% of the corpus.

## shortestPath

The `shortestPath` function is an implementation of a shortest path graph traversal. The `shortestPath` function performs an iterative breadth-first search through an unweighted graph to find the shortest paths between two nodes in a graph. The `shortestPath` function emits a tuple for each path found. Each tuple emitted will contain a `path` key which points to a `List` of `nodeIDs` comprising the path.

### shortestPath Parameters

- `collection`: (Mandatory) The collection that the topic query will be run on.
- `from`: (Mandatory) The `nodeID` to start the search from
- `to`: (Mandatory) The `nodeID` to end the search at
- `edge`: (Mandatory) Syntax: `from_field=to_field`. The `from_field` defines which field to search from. The `to_field` defines which field to search to. See example below for a detailed explanation.
- `threads`: (Optional: Default 6) The number of threads used to perform the partitioned join in the traversal.
- `partitionSize`: (Optional: Default 250) The number of nodes in each partition of the join.
- `fq`: (Optional) Filter query
- `maxDepth`: (Mandatory) Limits to the search to a maximum depth in the graph.

## shortestPath Syntax

```
shortestPath(collection,
 from="john@company.com",
 to="jane@company.com",
 edge="from_address=to_address",
 threads="6",
 partitionSize="300",
 fq="limiting query",
 maxDepth="4")
```

The expression above performs a breadth-first search to find the shortest paths in an unweighted, directed graph.

The search starts from the nodeID "john@company.com" in the `from_address` field and searches for the nodeID "jane@company.com" in the `to_address` field. This search is performed iteratively until the `maxDepth` has been reached. Each level in the traversal is implemented as a parallel partitioned nested loop join across the entire collection. The `threads` parameter controls the number of threads performing the join at each level, while the `partitionSize` parameter controls the of number of nodes in each join partition. The `maxDepth` parameter controls the number of levels to traverse. `fq` is a limiting query applied to each level in the traversal.

## shuffle

The `shuffle` expression sorts and exports entire result sets. The `shuffle` expression is similar to the `search` expression except that under the covers `shuffle` always uses the `/export` handler. The `shuffle` expression is designed to be combined with the relational algebra decorators that require complete, sorted result sets. Shuffled result sets can be partitioned across worker nodes with the parallel stream decorator to perform parallel relational algebra. When used in parallel mode the `partitionKeys` parameter must be provided.

### shuffle Parameters

- `collection`: (Mandatory) the collection being searched.
- `q`: (Mandatory) The query to perform on the Solr index.
- `f1`: (Mandatory) The list of fields to return.
- `sort`: (Mandatory) The sort criteria.
- `zkHost`: Only needs to be defined if the collection being searched is found in a different zkHost than the local stream handler.
- `partitionKeys`: Comma delimited list of keys to partition the search results by. To be used with the `parallel` function for parallelizing operations across worker nodes. See the [parallel](#) function for details.

### shuffle Syntax

```
shuffle(collection1,
 q="*:*",
 f1="id,a_s,a_i,a_f",
 sort="a_f asc, a_i asc")
```

## stats

The `stats` function gathers simple aggregations for a search result set. The `stats` function does not support rollups over buckets, so the `stats` stream always returns a single tuple with the rolled up stats. Under the covers the `stats` function pushes down the generation of the stats into the search engine using the `StatsComponent`. The `stats` function currently supports the following metrics: `count(*)`, `sum()`, `avg()`, `min()`, and `max()`.

## stats Parameters

- collection: (Mandatory) Collection the stats will be aggregated from.
- q: (Mandatory) The query to build the aggregations from.
- metrics: (Mandatory) The metrics to include in the result tuple. Current supported metrics are sum(col), avg(col), min(col), max(col) and count(\*)

## stats Syntax

```
stats(collection1,
 q=*:*,
 sum(a_i),
 sum(a_f),
 min(a_i),
 min(a_f),
 max(a_i),
 max(a_f),
 avg(a_i),
 avg(a_f),
 count(*))
```

## timeseries

The timeseries function builds a time series aggregation. Under the covers the timeseries function uses the JSON Facet API as its high performance aggregation engine.

## timeseries Parameters

- collection: (Mandatory) Collection the stats will be aggregated from.
- q: (Mandatory) The query to build the aggregations from.
- field: (Mandatory) The date field for the time series.
- start: (Mandatory) The start of the time series expressed in Solr date or date math syntax.
- end: (Mandatory) The end of the time series expressed in Solr date or date math syntax.
- gap: (Mandatory) The time gap between time series aggregation points expressed in Solr date math syntax.
- format: (Optional) Date template to format the date field in the output tuples. Formatting is performed by Java's SimpleDateFormat class.
- metrics: (Mandatory) The metrics to include in the result tuple. Current supported metrics are sum(col), avg(col), min(col), max(col) and count(\*)

## timeseries Syntax

```
timeseries(collection1,
 q=*:*,
 field="rec_dt"
 start="NOW-30DAYS",
 end="NOW",
 gap="+1DAY",
 format="YYYY-MM-dd",
 sum(a_i),
 max(a_i),
 max(a_f),
 avg(a_i),
 avg(a_f),
 count(*))
```

## train

The `train` function trains a Logistic Regression text classifier on a training set stored in a SolrCloud collection. It uses a parallel iterative, batch Gradient Descent approach to train the model. The training algorithm is embedded inside Solr so with each iteration only the model is streamed across the network.

The `train` function wraps a [features](#) function which provides the terms and inverse document frequency (IDF) used to train the model. The `train` function operates over the same training set as the `features` function, which includes both positive and negative examples of the class.

With each iteration the `train` function emits a tuple with the model. The model contains the feature terms, weights, and the confusion matrix for the model. The optimized model can then be used to classify documents based on their feature terms.

### train Parameters

- `collection`: (Mandatory) Collection that holds the training set
- `q`: (Mandatory) The query that defines the training set. The IDF for the features will be generated on the
- `name`: (Mandatory) The name of model. This can be used to retrieve the model if they stored in a Solr Cloud collection.
- `field`: (Mandatory) The text field to extract the features from.
- `outcome`: (Mandatory) The field that defines the class, positive or negative
- `maxIterations`: (Mandatory) How many training iterations to perform.
- `positiveLabel`: (defaults to 1) The value in the outcome field that defines a positive outcome.

### train Syntax

```
train(collection1,
 features(collection1, q="*:*", featureSet="first", field="body", outcome="out_i",
 numTerms=250),
 q="*:*",
 name="model1",
 field="body",
 outcome="out_i",
 maxIterations=100)
```

## topic

The topic function provides publish/subscribe messaging capabilities built on top of SolrCloud. The topic function allows users to subscribe to a query. The function then provides one-time delivery of new or updated documents that match the topic query. The initial call to the topic function establishes the checkpoints for the specific topic ID. Subsequent calls to the same topic ID will return documents added or updated after the initial checkpoint. Each run of the topic query updates the checkpoints for the topic ID. Setting the initialCheckpoint parameter to 0 will cause the topic to process all documents in the index that match the topic query.



The topic function should be considered in beta until [SOLR-8709](#) is committed and released.

### topic Parameters

- `checkpointCollection`: (Mandatory) The collection where the topic checkpoints are stored.
- `collection`: (Mandatory) The collection that the topic query will be run on.
- `id`: (Mandatory) The unique ID for the topic. The checkpoints will be saved under this id.
- `q`: (Mandatory) The topic query.
- `f1`: (Mandatory) The field list returned by the topic function.
- `initialCheckpoint`: (Optional) Sets the initial Solr \_version\_ number to start reading from the queue. If not set, it defaults to the highest version in the index. Setting to 0 will process all records that match query in the index.

### topic Syntax

```
topic(checkpointCollection,
 collection,
 id="uniqueId",
 q="topic query",
 f1="id, name, country")
```

## tuple

The tuple function emits a single Tuple with name/value pairs. The values can be set to variables assigned in a let expression, literals, Stream Evaluators or Stream Expressions. In the case of Stream Evaluators the tuple will output the return value from the evaluator. This could be a numeric, list or map. If a value is set to

a Stream Expression, the tuple function will flatten the tuple stream from the Stream Expression into a list of Tuples.

## tuple Parameters

- name=value pairs

## tuple Syntax

```
tuple(a=add(1,1),
 b=search(collection1, q="cat:a", fl="a, b, c", sort="a desc"))
```

# Stream Decorator Reference

## cartesianProduct

The cartesianProduct function turns a single tuple with a multi-valued field (i.e., an array) into multiple tuples, one for each value in the array field. That is, given a single tuple containing an array of N values for fieldA, the cartesianProduct function will output N tuples, each with one value from the original tuple's array. In essence, you can flatten arrays for further processing.

For example, using cartesianProduct you can turn this tuple:

```
{
 "fieldA": "foo",
 "fieldB": ["bar", "baz", "bat"]
}
```

into the following 3 tuples:

```
{
 "fieldA": "foo",
 "fieldB": "bar"
}
{
 "fieldA": "foo",
 "fieldB": "baz"
}
{
 "fieldA": "foo",
 "fieldB": "bat"
}
```

## cartesianProduct Parameters

- incoming stream: (Mandatory) A single incoming stream.
- fieldName or evaluator: (Mandatory) Name of field to flatten values for, or evaluator whose result should be flattened.
- productSort='fieldName ASC|DESC': (Optional) Sort order of the newly generated tuples.

## cartesianProduct Syntax

```
cartesianProduct(
 <stream>,
 <fieldName | evaluator> [as newFieldName],
 productSort='fieldName ASC|DESC'
)
```

## cartesianProduct Examples

The following examples show different outputs for this source tuple:

```
{
 "fieldA": "valueA",
 "fieldB": ["valueB1", "valueB2"],
 "fieldC": [1, 2, 3]
}
```

### Single Field, No Sorting

```
cartesianProduct(
 search(collection1, q="*:*", qt="/export", fl="fieldA, fieldB, fieldC", sort="fieldA asc"),
 fieldB
)

{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": [1, 2, 3]
}
{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": [1, 2, 3]
}
```

### Single Evaluator, No Sorting

```
cartesianProduct(
 search(collection1, q="*:*", qt="/export", fl="fieldA, fieldB, fieldC", sort="fieldA asc"),
 sequence(3,4,5) as fieldE
)

{
 "fieldA": "valueA",
 "fieldB": ["valueB1","valueB2"],
 "fieldC": [1,2,3],
 "fieldE": 4
}
{
 "fieldA": "valueA",
 "fieldB": ["valueB1","valueB2"],
 "fieldC": [1,2,3],
 "fieldE": 9
}
{
 "fieldA": "valueA",
 "fieldB": ["valueB1","valueB2"],
 "fieldC": [1,2,3],
 "fieldE": 14
}
```

## Single Field, Sorted by Value

```
cartesianProduct(
 search(collection1, q="*:*", qt="/export", fl="fieldA, fieldB, fieldC", sort="fieldA asc"),
 fieldB,
 productSort="fieldB desc"
)

{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": [1,2,3]
}
{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": [1,2,3]
}
```

## Single Evaluator, Sorted by Evaluator Values

```
cartesianProduct(
 search(collection1, q="*:*", qt="/export", fl="fieldA, fieldB, fieldC", sort="fieldA asc"),
 sequence(3,4,5) as fieldE,
 productSort="newFieldE desc"
)
{
 "fieldA": "valueA",
 "fieldB": ["valueB1","valueB2"],
 "fieldC": [1,2,3],
 "fieldE": 14
}
{
 "fieldA": "valueA",
 "fieldB": ["valueB1","valueB2"],
 "fieldC": [1,2,3],
 "fieldE": 9
}
{
 "fieldA": "valueA",
 "fieldB": ["valueB1","valueB2"],
 "fieldC": [1,2,3],
 "fieldE": 4
}
```

## Renamed Single Field, Sorted by Value

```
cartesianProduct(
 search(collection1, q="*:*", qt="/export", fl="fieldA, fieldB, fieldC", sort="fieldA asc"),
 fieldB as newFieldB,
 productSort="fieldB desc"
)
{
 "fieldA": "valueA",
 "fieldB": ["valueB1","valueB2"],
 "fieldC": [1,2,3]
 "newFieldB": "valueB2",
}
{
 "fieldA": "valueA",
 "fieldB": ["valueB1","valueB2"],
 "fieldC": [1,2,3]
 "newFieldB": "valueB1",
}
```

## Multiple Fields, No Sorting

```
cartesianProduct(
 search(collection1, q="*:*", qt="/export", fl="fieldA, fieldB, fieldC", sort="fieldA asc"),
 fieldB,
 fieldC
)

{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": 1
}
{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": 2
}
{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": 3
}
{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": 1
}
{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": 2
}
{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": 3
}
```

## Multiple Fields, Sorted by Single Field

```
cartesianProduct(
 search(collection1, qt="/export", q="*:*", fl="fieldA, fieldB, fieldC", sort="fieldA asc"),
 fieldB,
 fieldC,
 productSort="fieldC asc"
)

{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": 1
}
{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": 1
}
{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": 2
}
{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": 2
}
{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": 3
}
{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": 3
}
```

## Multiple Fields, Sorted by Multiple Fields

```
cartesianProduct(
 search(collection1, q="*:*", qt="/export", fl="fieldA, fieldB, fieldC", sort="fieldA asc"),
 fieldB,
 fieldC,
 productSort="fieldC asc, fieldB desc"
)

{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": 1
}
{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": 1
}
{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": 2
}
{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": 2
}
{
 "fieldA": "valueA",
 "fieldB": "valueB2",
 "fieldC": 3
}
{
 "fieldA": "valueA",
 "fieldB": "valueB1",
 "fieldC": 3
}
```

## Field and Evaluator, No Sorting

```
cartesianProduct(
 search(collection1, q="*:*", qt="/export", fl="fieldA, fieldB, fieldC", sort="fieldA asc"),
 sequence(3,4,5) as fieldE,
 fieldB
)

{
 "fieldA": "valueA",
 "fieldB": valueB1,
 "fieldC": [1,2,3],
 "fieldE": 4
}
}
{
 "fieldA": "valueA",
 "fieldB": valueB2,
 "fieldC": [1,2,3],
 "fieldE": 4
}
}
{
 "fieldA": "valueA",
 "fieldB": valueB1,
 "fieldC": [1,2,3],
 "fieldE": 9
}
}
{
 "fieldA": "valueA",
 "fieldB": valueB2,
 "fieldC": [1,2,3],
 "fieldE": 9
}
}
{
 "fieldA": "valueA",
 "fieldB": valueB1,
 "fieldC": [1,2,3],
 "fieldE": 14
}
}
{
 "fieldA": "valueA",
 "fieldB": valueB2,
 "fieldC": [1,2,3],
 "fieldE": 14
}
}
```

As you can see in the examples above, the cartesianProduct function does support flattening tuples across multiple fields and/or evaluators.

## classify

The classify function classifies tuples using a logistic regression text classification model. It was designed specifically to work with models trained using the [train function](#). The classify function uses the [model](#)

`function` to retrieve a stored model and then scores a stream of tuples using the model. The tuples read by the classifier must contain a text field that can be used for classification. The `classify` function uses a Lucene analyzer to extract the features from the text so the model can be applied. By default the `classify` function looks for the analyzer using the name of text field in the tuple. If the Solr schema on the worker node does not contain this field, the analyzer can be looked up in another field by specifying the `analyzerField` parameter.

Each tuple that is classified is assigned two scores:

- `probability_d*`: A float between 0 and 1 which describes the probability that the tuple belongs to the class. This is useful in the classification use case.
- `score_d*`: The score of the document that has not been squashed between 0 and 1. The score may be positive or negative. The higher the score the better the document fits the class. This un-squashed score will be useful in query re-ranking and recommendation use cases. This score is particularly useful when multiple high ranking documents have a `probability_d` score of 1, which won't provide a meaningful ranking between documents.

## classify Parameters

- `model expression`: (Mandatory) Retrieves the stored logistic regression model.
- `field`: (Mandatory) The field in the tuples to apply the classifier to. By default the analyzer for this field in the schema will be used to extract the features.
- `analyzerField`: (Optional) Specifies a different field to find the analyzer from in the schema.

## classify Syntax

```
classify(model(modelCollection,
 id="model1",
 cacheMillis=5000),
 search(contentCollection,
 q="id:(a b c)",
 qt="/export",
 fl="text_t, id",
 sort="id asc"),
 field="text_t")
```

In the example above the `classify` expression is retrieving the model using the `model` function. It is then classifying tuples returned by the `search` function. The `text_t` field is used for the text classification and the analyzer for the `text_t` field in the Solr schema is used to analyze the text and extract the features.

## commit

The `commit` function wraps a single stream (A) and given a collection and batch size will send commit messages to the collection when the batch size is fulfilled or the end of stream is reached. A commit stream is used most frequently with an update stream and as such the commit will take into account possible summary tuples coming from the update stream. All tuples coming into the commit stream will be returned out of the commit stream - no tuples will be dropped and no tuples will be added.

## commit Parameters

- collection: The collection to send commit messages to (required)
- batchSize: The commit batch size, sends commit message when batch size is hit. If not provided (or provided as value 0) then a commit is only sent at the end of the incoming stream.
- waitFlush: The value passed directly to the commit handler (true/false, default: false)
- waitSearcher: The value passed directly to the commit handler (true/false, default: false)
- softCommit: The value passed directly to the commit handler (true/false, default: false)
- StreamExpression for StreamA (required)

## commit Syntax

```
commit(
 destinationCollection,
 batchSize=2,
 update(
 destinationCollection,
 batchSize=5,
 search(collection1, q="*:*", qt="/export", fl="id,a_s,a_i,a_f,s_multi,i_multi", sort="a_f
asc, a_i asc")
)
)
```

## complement

The complement function wraps two streams (A and B) and emits tuples from A which do not exist in B. The tuples are emitted in the order in which they appear in stream A. Both streams must be sorted by the fields being used to determine equality (using the on parameter).

## complement Parameters

- StreamExpression for StreamA
- StreamExpression for StreamB
- on: Fields to be used for checking equality of tuples between A and B. Can be of the format on="fieldName", on="fieldNameInLeft=fieldNameInRight", or on="fieldName, otherFieldName=rightOtherFieldName".

## complement Syntax

```
complement(
 search(collection1, q="a_s:(setA || setAB)", qt="/export", fl="id,a_s,a_i", sort="a_i asc, a_s
asc"),
 search(collection1, q="a_s:(setB || setAB)", qt="/export", fl="id,a_s,a_i", sort="a_i asc"),
 on="a_i"
)

complement(
 search(collection1, q="a_s:(setA || setAB)", qt="/export", fl="id,a_s,a_i", sort="a_i asc, a_s
asc"),
 search(collection1, q="a_s:(setB || setAB)", qt="/export", fl="id,a_s,a_i", sort="a_i asc, a_s
asc"),
 on="a_i,a_s"
)
```

## daemon

The daemon function wraps another function and runs it at intervals using an internal thread. The daemon function can be used to provide both continuous push and pull streaming.

### Continuous Push Streaming

With continuous push streaming the daemon function wraps another function and is then sent to the /stream handler for execution. The /stream handler recognizes the daemon function and keeps it resident in memory, so it can run its internal function at intervals.

In order to facilitate the pushing of tuples, the daemon function must wrap another stream decorator that pushes the tuples somewhere. One example of this is the update function, which wraps a stream and sends the tuples to another SolrCloud collection for indexing.

### daemon Syntax

```
daemon(id="uniqueId",
 runInterval="1000",
 terminate="true",
 update(destinationCollection,
 batchSize=100,
 topic(checkpointCollection,
 topicCollection,
 q="topic query",
 fl="id, title, abstract, text",
 id="topicId",
 initialCheckpoint=0)
)
)
```

The sample code above shows a daemon function wrapping an update function, which is wrapping a topic function. When this expression is sent to the /stream handler, the /stream hander sees the daemon function

and keeps it in memory where it will run at intervals. In this particular example, the daemon function will run the update function every second. The update function is wrapping a topic [function](#), which will stream tuples that match the topic function query in batches. Each subsequent call to the topic will return the next batch of tuples for the topic. The update function will send all the tuples matching the topic to another collection to be indexed. The terminate parameter tells the daemon to terminate when the topic function stops sending tuples.

The effect of this is to push documents that match a specific query into another collection. Custom push functions can be plugged in that push documents out of Solr and into other systems, such as Kafka or an email system.

Push streaming can also be used for continuous background aggregation scenarios where aggregates are rolled up in the background at intervals and pushed to other Solr collections. Another use case is continuous background machine learning model optimization, where the optimized model is pushed to another Solr collection where it can be integrated into queries.

The /stream handler supports a small set commands for listing and controlling daemon functions:

```
http://localhost:8983/collection/stream?action=list
```

This command will provide a listing of the current daemon's running on the specific node along with there current state.

```
http://localhost:8983/collection/stream?action=stop&id=daemonId
```

This command will stop a specific daemon function but leave it resident in memory.

```
http://localhost:8983/collection/stream?action=start&id=daemonId
```

This command will start a specific daemon function that has been stopped.

```
http://localhost:8983/collection/stream?action=kill&id=daemonId
```

This command will stop a specific daemon function and remove it from memory.

## Continuous Pull Streaming

The [DaemonStream](#) java class (part of the SolrJ libraries) can also be embedded in a java application to provide continuous pull streaming. Sample code:

```
StreamContext context = new StreamContext()
SolrClientCache cache = new SolrClientCache();
context.setSolrClientCache(cache);

Map topicQueryParams = new HashMap();
topicQueryParams.put("q", "hello"); // The query for the topic
topicQueryparams.put("rows", "500"); // How many rows to fetch during each run
```

```
topicQueryparams.put("fl", "id", "title"); // The field list to return with the documents

TopicStream topicStream = new TopicStream(zkHost, // Host address for the ZooKeeper
service housing the collections
 "checkpoints", // The collection to store the topic
checkpoints
 "topicData", // The collection to query for the topic
records
 "topicId", // The id of the topic
-1, // checkpoint every X tuples, if set -1
it will checkpoint after each run.
 topicQueryParams); // The query parameters for the
TopicStream

DaemonStream daemonStream = new DaemonStream(topicStream, // The underlying stream to
run.
 "daemonId", // The id of the daemon
1000, // The interval at which to
run the internal stream
 500); // The internal queue size
for the daemon stream. Tuples will be placed in the queue
 // as they are read by the
internal internal thread.
 // Calling read() on the
daemon stream reads records from the internal queue.

daemonStream.setStreamContext(context);

daemonStream.open();

//Read until it's time to shutdown the DaemonStream. You can define the shutdown criteria.
while(!shutdown()) {
 Tuple tuple = daemonStream.read() // This will block until tuples become available from the
underlying stream (TopicStream)
 // The EOF tuple (signaling the end of the stream) will
never occur until the DaemonStream has been shutdown.
 //Do something with the tuples
}

// Shutdown the DaemonStream.
daemonStream.shutdown();

//Read the DaemonStream until the EOF Tuple is found.
//This allows the underlying stream to perform an orderly shutdown.

while(true) {
 Tuple tuple = daemonStream.read();
 if(tuple.EOF) {
 break;
 } else {
 //Do something with the tuples.
 }
}
```

```
}
```

```
//Finally close the stream
```

```
daemonStream.close();
```

## eval

The eval function allows for use cases where new streaming expressions are generated on the fly and then evaluated. The eval function wraps a streaming expression and reads a single tuple from the underlying stream. The eval function then retrieves a string Streaming Expressions from the expr\_s field of the tuple. The eval function then compiles the string Streaming Expression and emits the tuples.

### eval Parameters

- StreamExpression: (Mandatory) The stream which provides the streaming expression to be evaluated.

### eval Syntax

```
eval(expr)
```

In the example above the eval expression reads the first tuple from the underlying expression. It then compiles and executes the string Streaming Expression in the expr\_s field.

## executor

The executor function wraps a stream source that contains streaming expressions, and executes the expressions in parallel. The executor function looks for the expression in the expr\_s field in each tuple. The executor function has an internal thread pool that runs tasks that compile and run expressions in parallel on the same worker node. This function can also be parallelized across worker nodes by wrapping it in the parallel function to provide parallel execution of expressions across a cluster.

The executor function does not do anything specific with the output of the expressions that it runs. Therefore the expressions that are executed must contain the logic for pushing tuples to their destination. The [update function](#) can be included in the expression being executed to send the tuples to a SolrCloud collection for storage.

This model allows for asynchronous execution of jobs where the output is stored in a SolrCloud collection where it can be accessed as the job progresses.

### executor Parameters

- threads: (Optional) The number of threads in the executors thread pool for executing expressions.
- StreamExpression: (Mandatory) The stream source which contains the Streaming Expressions to execute.

### executor Syntax

```
daemon(id="myDaemon",
 terminate="true",
 executor(threads=10,
 topic(checkpointCollection
 storedExpressions,
 q="*:*",
 fl="id, expr_s",
 initialCheckPoint=0,
 id="myTopic")))
```

In the example above a `daemon` wraps an executor, which wraps a `topic` that is returning tuples with expressions to execute. When sent to the stream handler, the daemon will call the executor at intervals which will cause the executor to read from the topic and execute the expressions found in the `expr_s` field. The daemon will repeatedly call the executor until all the tuples that match the topic have been iterated, then it will terminate. This is the approach for executing batches of streaming expressions from a topic queue.

## fetch

The `fetch` function iterates a stream and fetches additional fields and adds them to the tuples. The `fetch` function fetches in batches to limit the number of calls back to Solr. Tuples streamed from the `fetch` function will contain the original fields and the additional fields that were fetched. The `fetch` function supports one-to-one fetches. Many-to-one fetches, where the stream source contains duplicate keys, will also work, but one-to-many fetches are currently not supported by this function.

### fetch Parameters

- `Collection`: (Mandatory) The collection to fetch the fields from.
- `StreamExpression`: (Mandatory) The stream source for the `fetch` function.
- `f1`: (Mandatory) The fields to be fetched.
- `on`: Fields to be used for checking equality of tuples between stream source and fetched records. Formatted as `on="fieldNameInTuple=fieldNameInCollection"`.
- `batchSize`: (Optional) The batch fetch size.

### fetch Syntax

```
fetch(addresses,
 search(people, q="*:*", qt="/export", fl="username, firstName, lastName", sort="username
asc"),
 fl="streetAddress, city, state, country, zip",
 on="username=userId")
```

The example above fetches addresses for users by matching the `username` in the tuple with the `userId` field in the `addresses` collection.

## having

The having expression wraps a stream and applies a boolean operation to each tuple. It emits only tuples for which the boolean operation returns **true**.

### having Parameters

- StreamExpression: (Mandatory) The stream source for the having function.
- booleanEvaluator: (Mandatory) The following boolean operations are supported: eq (equals), gt (greater than), lt (less than), gteq (greater than or equal to), lteq (less than or equal to), and, or, eor (exclusive or), and not. Boolean evaluators can be nested with other evaluators to form complex boolean logic.

The comparison evaluators compare the value in a specific field with a value, whether a string, number, or boolean. For example: eq(field1, 10), returns true if field1 is equal to 10.

### having Syntax

```
having(rollup(over=a_s,
 sum(a_i),
 search(collection1,
 q="*:*",
 qt="/export",
 fl="id,a_s,a_i,a_f",
 sort="a_s asc")),
 and(gt(sum(a_i), 100), lt(sum(a_i), 110)))
```

In this example, the having expression iterates the aggregated tuples from the rollup expression and emits all tuples where the field sum(a\_i) is greater than 100 and less than 110.

## leftOuterJoin

The leftOuterJoin function wraps two streams, Left and Right, and emits tuples from Left. If there is a tuple in Right equal (as defined by on) then the values in that tuple will be included in the emitted tuple. An equal tuple in Right **need not** exist for the Left tuple to be emitted. This supports one-to-one, one-to-many, many-to-one, and many-to-many left outer join scenarios. The tuples are emitted in the order in which they appear in the Left stream. Both streams must be sorted by the fields being used to determine equality (using the on parameter). If both tuples contain a field of the same name then the value from the Right stream will be used in the emitted tuple.

You can wrap the incoming streams with a select function to be specific about which field values are included in the emitted tuple.

### leftOuterJoin Parameters

- StreamExpression for StreamLeft
- StreamExpression for StreamRight
- on: Fields to be used for checking equality of tuples between Left and Right. Can be of the format

on="fieldName", on="fieldNameInLeft=fieldNameInRight", or on="fieldName, otherFieldName=rightOtherFieldName".

## leftOuterJoin Syntax

```
leftOuterJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 search(pets, q="type:cat", qt="/export", fl="personId,petName", sort="personId asc"),
 on="personId"
)

leftOuterJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 search(pets, q="type:cat", qt="/export", fl="ownerId,petName", sort="ownerId asc"),
 on="personId=ownerId"
)

leftOuterJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 select(
 search(pets, q="type:cat", qt="/export", fl="ownerId,name", sort="ownerId asc"),
 ownerId,
 name as petName
),
 on="personId=ownerId"
)
```

## hashJoin

The hashJoin function wraps two streams, Left and Right, and for every tuple in Left which exists in Right will emit a tuple containing the fields of both tuples. This supports one-to-one, one-to-many, many-to-one, and many-to-many inner join scenarios. The tuples are emitted in the order in which they appear in the Left stream. The order of the streams does not matter. If both tuples contain a field of the same name then the value from the Right stream will be used in the emitted tuple.

You can wrap the incoming streams with a select function to be specific about which field values are included in the emitted tuple.

The hashJoin function can be used when the tuples of Left and Right cannot be put in the same order. Because the tuples are out of order this stream functions by reading all values from the Right stream during the open operation and will store all tuples in memory. The result of this is a memory footprint equal to the size of the Right stream.

## hashJoin Parameters

- StreamExpression for StreamLeft
- hashed=StreamExpression for StreamRight
- on: Fields to be used for checking equality of tuples between Left and Right. Can be of the format on="fieldName", on="fieldNameInLeft=fieldNameInRight", or on="fieldName,

otherFieldName=rightOtherFieldName".

## hashJoin Syntax

```
hashJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 hashed=search(pets, q="type:cat", qt="/export", fl="personId,petName", sort="personId asc"),
 on="personId"
)

hashJoin(
 search(people, q="*:*", fl="personId,name", sort="personId asc"),
 hashed=search(pets, q="type:cat", qt="/export", fl="ownerId,petName", sort="ownerId asc"),
 on="personId=ownerId"
)

hashJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 hashed=select(
 search(pets, q="type:cat", qt="/export", fl="ownerId,name", sort="ownerId asc"),
 ownerId,
 name as petName
),
 on="personId=ownerId"
)
```

## innerJoin

Wraps two streams, Left and Right. For every tuple in Left which exists in Right a tuple containing the fields of both tuples will be emitted. This supports one-to-one, one-to-many, many-to-one, and many-to-many inner join scenarios. The tuples are emitted in the order in which they appear in the Left stream. Both streams must be sorted by the fields being used to determine equality (the 'on' parameter). If both tuples contain a field of the same name then the value from the Right stream will be used in the emitted tuple. You can wrap the incoming streams with a select(...) expression to be specific about which field values are included in the emitted tuple.

### innerJoin Parameters

- StreamExpression for StreamLeft
- StreamExpression for StreamRight
- on: Fields to be used for checking equality of tuples between Left and Right. Can be of the format on="fieldName", on="fieldNameInLeft=fieldNameInRight", or on="fieldName, otherFieldName=rightOtherFieldName".

## innerJoin Syntax

```
innerJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 search(pets, q="type:cat", qt="/export", fl="personId,petName", sort="personId asc"),
 on="personId"
)

innerJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 search(pets, q="type:cat", qt="/export", fl="ownerId,petName", sort="ownerId asc"),
 on="personId=ownerId"
)

innerJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 select(
 search(pets, q="type:cat", qt="/export", fl="ownerId,name", sort="ownerId asc"),
 ownerId,
 name as petName
,
 on="personId=ownerId"
)
```

## intersect

The `intersect` function wraps two streams, A and B, and emits tuples from A which **DO** exist in B. The tuples are emitted in the order in which they appear in stream A. Both streams must be sorted by the fields being used to determine equality (the `on` parameter). Only tuples from A are emitted.

### intersect Parameters

- StreamExpression for StreamA
- StreamExpression for StreamB
- `on`: Fields to be used for checking equality of tuples between A and B. Can be of the format `on="fieldName", on="fieldNameInLeft=fieldNameInRight"`, or `on="fieldName, otherFieldName=rightOtherFieldName"`.

### intersect Syntax

```
intersect(
 search(collection1, q="a_s:(setA || setAB)", qt="/export", fl="id,a_s,a_i", sort="a_i asc, a_s
asc"),
 search(collection1, q="a_s:(setB || setAB)", qt="/export", fl="id,a_s,a_i", sort="a_i asc"),
 on="a_i"
)

intersect(
 search(collection1, q="a_s:(setA || setAB)", qt="/export", fl="id,a_s,a_i", sort="a_i asc, a_s
asc"),
 search(collection1, q="a_s:(setB || setAB)", qt="/export", fl="id,a_s,a_i", sort="a_i asc, a_s
asc"),
 on="a_i,a_s"
)
```

## list

The `list` function wraps N Stream Expressions and opens and iterates each stream sequentially. This has the effect of concatenating the results of multiple Streaming Expressions.

### list Parameters

- StreamExpressions ...: N Streaming Expressions

### list Syntax

```
list(tuple(a="hello world"), tuple(a="HELLO WORLD"))

list(search(collection1, q="*:*", fl="id, prod_ss", sort="id asc"),
 search(collection2, q="*:*", fl="id, prod_ss", sort="id asc"))

list(tuple(a=search(collection1, q="*:*", fl="id, prod_ss", sort="id asc")),
 tuple(a=search(collection2, q="*:*", fl="id, prod_ss", sort="id asc")))
```

## merge

The `merge` function merges two or more streaming expressions and maintains the ordering of the underlying streams. Because the order is maintained, the sorts of the underlying streams must line up with the `on` parameter provided to the `merge` function.

### merge Parameters

- StreamExpression A
- StreamExpression B
- Optional StreamExpression C,D,...Z
- `on`: Sort criteria for performing the merge. Of the form `fieldName order` where `order` is `asc` or `desc`. Multiple fields can be provided in the form `fieldA order, fieldB order`.

## merge Syntax

```
Merging two stream expressions together
merge(
 search(collection1,
 q="id:(0 3 4)",
 qt="/export",
 fl="id,a_s,a_i,a_f",
 sort="a_f asc"),
 search(collection1,
 q="id:(1)",
 qt="/export",
 fl="id,a_s,a_i,a_f",
 sort="a_f asc"),
 on="a_f asc")
```

```
Merging four stream expressions together. Notice that while the sorts of each stream are not
identical they are
comparable. That is to say the first N fields in each stream's sort matches the N fields in the
merge's on clause.
merge(
 search(collection1,
 q="id:(0 3 4)",
 qt="/export",
 fl="id,fieldA,fieldB,fieldC",
 sort="fieldA asc, fieldB desc"),
 search(collection1,
 q="id:(1)",
 qt="/export",
 fl="id,fieldA",
 sort="fieldA asc"),
 search(collection2,
 q="id:(10 11 13)",
 qt="/export",
 fl="id,fieldA,fieldC",
 sort="fieldA asc"),
 search(collection3,
 q="id:(987)",
 qt="/export",
 fl="id,fieldA,fieldC",
 sort="fieldA asc"),
 on="fieldA asc")
```

## null

The null expression is a useful utility function for understanding bottlenecks when performing parallel relational algebra (joins, intersections, rollups etc.). The null function reads all the tuples from an underlying stream and returns a single tuple with the count and processing time. Because the null stream adds minimal overhead of its own, it can be used to isolate the performance of Solr's /export handler. If the /export

handlers performance is not the bottleneck, then the bottleneck is likely occurring in the workers where the stream decorators are running.

The null expression can be wrapped by the parallel function and sent to worker nodes. In this scenario each worker will return one tuple with the count of tuples processed on the worker and the timing information for that worker. This gives valuable information such as:

1. As more workers are added does the performance of the /export handler improve or not.
2. Are tuples being evenly distributed across the workers, or is the hash partitioning sending more documents to a single worker.
3. Are all workers processing data at the same speed, or is one of the workers the source of the bottleneck.

## null Parameters

- StreamExpression: (Mandatory) The expression read by the null function.

## null Syntax

```
parallel(workerCollection,
 null(search(collection1, q="*:*", fl="id,a_s,a_i,a_f", sort="a_s desc", qt="/export",
partitionKeys="a_s")),
 workers="20",
 zkHost="localhost:9983",
 sort="a_s desc")
```

The expression above shows a parallel function wrapping a null function. This will cause the null function to be run in parallel across 20 worker nodes. Each worker will return a single tuple with number of tuples processed and time it took to iterate the tuples.

## outerHashJoin

The outerHashJoin function wraps two streams, Left and Right, and emits tuples from Left. If there is a tuple in Right equal (as defined by the on parameter) then the values in that tuple will be included in the emitted tuple. An equal tuple in Right **need not** exist for the Left tuple to be emitted. This supports one-to-one, one-to-many, many-to-one, and many-to-many left outer join scenarios. The tuples are emitted in the order in which they appear in the Left stream. The order of the streams does not matter. If both tuples contain a field of the same name then the value from the Right stream will be used in the emitted tuple.

You can wrap the incoming streams with a select function to be specific about which field values are included in the emitted tuple.

The outerHashJoin stream can be used when the tuples of Left and Right cannot be put in the same order. Because the tuples are out of order, this stream functions by reading all values from the Right stream during the open operation and will store all tuples in memory. The result of this is a memory footprint equal to the size of the Right stream.

## outerHashJoin Parameters

- StreamExpression for StreamLeft
- hashed=StreamExpression for StreamRight
- on: Fields to be used for checking equality of tuples between Left and Right. Can be of the format on="fieldName", on="fieldNameInLeft=fieldNameInRight", or on="fieldName, otherFieldName=rightOtherFieldName".

## outerHashJoin Syntax

```
outerHashJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 hashed=search(pets, q="type:cat", qt="/export", fl="personId,petName", sort="personId asc"),
 on="personId"
)

outerHashJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 hashed=search(pets, q="type:cat", qt="/export", fl="ownerId,petName", sort="ownerId asc"),
 on="personId=ownerId"
)

outerHashJoin(
 search(people, q="*:*", qt="/export", fl="personId,name", sort="personId asc"),
 hashed=select(
 search(pets, q="type:cat", qt="/export", fl="ownerId,name", sort="ownerId asc"),
 ownerId,
 name as petName
,
 on="personId=ownerId"
)
```

## parallel

The parallel function wraps a streaming expression and sends it to N worker nodes to be processed in parallel.

The parallel function requires that the partitionKeys parameter be provided to the underlying searches. The partitionKeys parameter will partition the search results (tuples) across the worker nodes. Tuples with the same values as partitionKeys will be shuffled to the same worker nodes.

The parallel function maintains the sort order of the tuples returned by the worker nodes, so the sort criteria must incorporate the sort order of the tuples returned by the workers.

For example if you sort on year, month and day you could partition on year only as long as there are enough different years to spread the tuples around the worker nodes.

Solr allows sorting on more than 4 fields, but you cannot specify more than 4 partitionKeys for speed considerations. Also it's overkill to specify many partitionKeys when we one or two keys could be enough

to spread the tuples.

Parallel stream was designed when the underlying search stream will emit a lot of tuples from the collection. If the search stream only emits a small subset of the data from the collection using parallel could potentially be slower.

#### *Worker Collections*



The worker nodes can be from the same collection as the data, or they can be a different collection entirely, even one that only exists for parallel streaming expressions. A worker collection can be any SolrCloud collection that has the /stream handler configured. Unlike normal SolrCloud collections, worker collections don't have to hold any data. Worker collections can be empty collections that exist only to execute streaming expressions.

## parallel Parameters

- collection: Name of the worker collection to send the StreamExpression to.
- StreamExpression: Expression to send to the worker collection.
- workers: Number of workers in the worker collection to send the expression to.
- zkHost: (Optional) The ZooKeeper connect string where the worker collection resides.
- sort: The sort criteria for ordering tuples returned by the worker nodes.

## parallel Syntax

```
parallel(workerCollection,
 rollup(search(collection1, q="*:*", fl="id,year_i,month_i,day_i", qt="/export",
sort="year_i desc,month_i desc,day_i asc", partitionKeys="year_i"),
 over="year_i", count(*)),
 workers="20",
 zkHost="localhost:9983",
 sort="year_i desc")
```

The expression above shows a parallel function wrapping a rollup function. This will cause the rollup function to be run in parallel across 20 worker nodes.

### Warmup

The parallel function uses the hash query parser to split the data amongst the workers. It executes on all the documents and the result bitset is cached in the filterCache.

- + For a parallel stream with the same number of workers and partitionKeys the first query would be slower than subsequent queries. A trick to not pay the penalty for the first slow query would be to use a warmup query for every new searcher. The following is a solrconfig.xml snippet for 2 workers and "year\_i" as the partitionKeys.



```
<listener event="newSearcher" class="solr.QuerySenderListener">
<arr name="queries">
 <lst><str name="q">:</str><str name="fq">{!hash workers=2 worker=0}</str><str
name="partitionKeys">year_i</str></lst>
 <lst><str name="q">:</str><str name="fq">{!hash workers=2 worker=1}</str><str
name="partitionKeys">year_i</str></lst>
</arr>
</listener>
```

## plist

The plist function wraps N Stream Expressions and opens the streams in parallel and iterates each stream sequentially. The difference between the list and plist is that the streams are opened in parallel. Since many streams such as facet, stats and significantTerms push down heavy operations to Solr when they are opened, the plist function can dramatically improve performance by doing these operations in parallel.

### plist Parameters

- StreamExpressions ...: N Streaming Expressions

### plist Syntax

```
plist(tuple(a="hello world"), tuple(a="HELLO WORLD"))

plist(search(collection1, q="*:*", fl="id, prod_ss", sort="id asc"),
 search(collection2, q="*:*", fl="id, prod_ss", sort="id asc"))

plist(tuple(a=search(collection1, q="*:*", fl="id, prod_ss", sort="id asc")),
 tuple(a=search(collection2, q="*:*", fl="id, prod_ss", sort="id asc")))
```

## priority

The priority function is a simple priority scheduler for the [executor](#) function. The executor function doesn't directly have a concept of task prioritization; instead it simply executes tasks in the order that they are read from its underlying stream. The priority function provides the ability to schedule a higher priority task ahead of lower priority tasks that were submitted earlier.

The priority function wraps two [topics](#) that are both emitting tuples that contain streaming expressions to

execute. The first topic is considered the higher priority task queue.

Each time the priority function is called, it checks the higher priority task queue to see if there are any tasks to execute. If tasks are waiting in the higher priority queue then the priority function will emit the higher priority tasks. If there are no high priority tasks to run, the lower priority queue tasks are emitted.

The priority function will only emit a batch of tasks from one of the queues each time it is called. This ensures that no lower priority tasks are executed until the higher priority queue has no tasks to run.

## priority Parameters

- `topic expression`: (Mandatory) the high priority task queue
- `topic expression`: (Mandatory) the lower priority task queue

## priority Syntax

```
daemon(id="myDaemon",
 executor(threads=10,
 priority(topic(checkpointCollection, storedExpressions, q="priority:high",
 fl="id, expr_s", initialCheckPoint=0,id="highPriorityTasks"),
 topic(checkpointCollection, storedExpressions, q="priority:low", fl="id,
 expr_s", initialCheckPoint=0,id="lowPriorityTasks"))))
```

In the example above the daemon function is calling the executor iteratively. Each time it's called, the executor function will execute the tasks emitted by the priority function. The priority function wraps two topics. The first topic is the higher priority task queue, the second topics is the lower priority topic.

## reduce

The reduce function wraps an internal stream and groups tuples by common fields.

Each tuple group is operated on as a single block by a pluggable reduce operation. The group operation provided with Solr implements distributed grouping functionality. The group operation also serves as an example reduce operation that can be referred to when building custom reduce operations.



The reduce function relies on the sort order of the underlying stream. Accordingly the sort order of the underlying stream must be aligned with the group by field.

## reduce Parameters

- `StreamExpression`: (Mandatory)
- `by`: (Mandatory) A comma separated list of fields to group by.
- `Reduce Operation`: (Mandatory)

## reduce Syntax

```
reduce(search(collection1, q="*:*", qt="/export", fl="id,a_s,a_i,a_f", sort="a_s asc, a_f asc"),
 by="a_s",
 group(sort="a_f desc", n="4")
)
```

## rollup

The rollup function wraps another stream function and rolls up aggregates over bucket fields. The rollup function relies on the sort order of the underlying stream to rollup aggregates one grouping at a time. Accordingly, the sort order of the underlying stream must match the fields in the over parameter of the rollup function.

The rollup function also needs to process entire result sets in order to perform its aggregations. When the underlying stream is the search function, the /export handler can be used to provide full sorted result sets to the rollup function. This sorted approach allows the rollup function to perform aggregations over very high cardinality fields. The disadvantage of this approach is that the tuples must be sorted and streamed across the network to a worker node to be aggregated. For faster aggregation over low to moderate cardinality fields, the facet function can be used.

### rollup Parameters

- StreamExpression (Mandatory)
- over: (Mandatory) A list of fields to group by.
- metrics: (Mandatory) The list of metrics to compute. Currently supported metrics are sum(col), avg(col), min(col), max(col), count(\*).

### rollup Syntax

```
rollup(
 search(collection1, q="*:*", qt="/export", fl="a_s,a_i,a_f", qt="/export", sort="a_s asc"),
 over="a_s",
 sum(a_i),
 sum(a_f),
 min(a_i),
 min(a_f),
 max(a_i),
 max(a_f),
 avg(a_i),
 avg(a_f),
 count(*)
)
```

The example above shows the rollup function wrapping the search function. Notice that search function is using the /export handler to provide the entire result set to the rollup stream. Also notice that the search function's sort parameter matches up with the rollup's over parameter. This allows the rollup function to rollup the over the a\_s field, one group at a time.

## scoreNodes

See section in [graph traversal](#).

## select

The select function wraps a streaming expression and outputs tuples containing a subset or modified set of fields from the incoming tuples. The list of fields included in the output tuple can contain aliases to effectively rename fields. The select stream supports both operations and evaluators. One can provide a list of operations and evaluators to perform on any fields, such as replace, add, if, etc.

### select Parameters

- StreamExpression
- fieldName: name of field to include in the output tuple (can include multiple of these), such as `outputTuple[fieldName] = inputTuple[fieldName]`
- fieldName as aliasFieldName: aliased field name to include in the output tuple (can include multiple of these), such as `outputTuple[aliasFieldName] = incomingTuple[fieldName]`
- `replace(fieldName, value, withValue=replacementValue)`: if `incomingTuple[fieldName] == value` then `outgoingTuple[fieldName]` will be set to `replacementValue`. `value` can be the string "null" to replace a null value with some other value.
- `replace(fieldName, value, withField=otherFieldName)`: if `incomingTuple[fieldName] == value` then `outgoingTuple[fieldName]` will be set to the value of `incomingTuple[otherFieldName]`. `value` can be the string "null" to replace a null value with some other value.

### select Syntax

```
// output tuples with fields teamName, wins, losses, and winPercentages where a null value for
// wins or losses is translated to the value of 0
select(
 search(collection1, fl="id,teamName_s,wins,losses", q="*:*", qt="/export", sort="id asc"),
 teamName_s as teamName,
 wins,
 losses,
 replace(wins,null,value=0),
 replace(losses,null,value=0),
 if(eq(0,wins), 0, div(add(wins,losses), wins)) as winPercentage
)
```

## sort

The sort function wraps a streaming expression and re-orders the tuples. The sort function emits all incoming tuples in the new sort order. The sort function reads all tuples from the incoming stream, re-orders them using an algorithm with  $O(n \log(n))$  performance characteristics, where  $n$  is the total number of tuples in the incoming stream, and then outputs the tuples in the new sort order. Because all tuples are read into memory, the memory consumption of this function grows linearly with the number of tuples in the

incoming stream.

## sort Parameters

- StreamExpression
- by: Sort criteria for re-ordering the tuples

## sort Syntax

The expression below finds dog owners and orders the results by owner and pet name. Notice that it uses an efficient innerJoin by first ordering by the person/owner id and then re-orders the final output by the owner and pet names.

```
sort(
 innerJoin(
 search(people, q="*:*", qt="/export", fl="id,name", sort="id asc"),
 search(pets, q="type:dog", qt="/export", fl="owner,petName", sort="owner asc"),
 on="id=owner"
),
 by="name asc, petName asc"
)
```

## top

The top function wraps a streaming expression and re-orders the tuples. The top function emits only the top N tuples in the new sort order. The top function re-orders the underlying stream so the sort criteria **does not** have to match up with the underlying stream.

## top Parameters

- n: Number of top tuples to return.
- StreamExpression
- sort: Sort criteria for selecting the top N tuples.

## top Syntax

The expression below finds the top 3 results of the underlying search. Notice that it reverses the sort order. The top function re-orders the results of the underlying stream.

```
top(n=3,
 search(collection1,
 q="*:*",
 qt="/export",
 fl="id,a_s,a_i,a_f",
 sort="a_f desc, a_i desc"),
 sort="a_f asc, a_i asc")
```

## unique

The unique function wraps a streaming expression and emits a unique stream of tuples based on the over parameter. The unique function relies on the sort order of the underlying stream. The over parameter must match up with the sort order of the underlying stream.

The unique function implements a non-co-located unique algorithm. This means that records with the same unique over field do not need to be co-located on the same shard. When executed in the parallel, the partitionKeys parameter must be the same as the unique over field so that records with the same keys will be shuffled to the same worker.

### unique Parameters

- StreamExpression
- over: The unique criteria.

### unique Syntax

```
unique(
 search(collection1,
 q="*:*",
 qt="/export",
 fl="id,a_s,a_i,a_f",
 sort="a_f asc, a_i asc"),
 over="a_f")
```

## update

The update function wraps another functions and sends the tuples to a SolrCloud collection for indexing.

### update Parameters

- destinationCollection: (Mandatory) The collection where the tuples will indexed.
- batchSize: (Mandatory) The indexing batch size.
- StreamExpression: (Mandatory)

### update Syntax

```
update(destinationCollection,
 batchSize=500,
 search(collection1,
 q="*:*",
 qt="/export",
 fl="id,a_s,a_i,a_f,s_multi,i_multi",
 sort="a_f asc, a_i asc"))
```

The example above sends the tuples returned by the search function to the destinationCollection to be

indexed.

# Stream Evaluator Reference

Stream evaluators are different than stream sources or stream decorators. Both stream sources and stream decorators return streams of tuples. Stream evaluators are more like a traditional function that evaluates its parameters and returns a result. That result can be a single value, array, map or other structure.

Stream evaluators can be nested so that the output of an evaluator becomes the input for another evaluator.

Stream evaluators can be called in different contexts. For example a stream evaluator can be called on its own or it can be called within the context of a streaming expression.

## abs

The abs function will return the absolute value of the provided single parameter. The abs function will fail to execute if the value is non-numeric. If a null value is found then null will be returned as the result.

### abs Parameters

- Field Name | Raw Number | Number Evaluator

### abs Syntax

The expressions below show the various ways in which you can use the abs evaluator. Only one parameter is accepted. Returns a numeric value.

```
abs(1) // 1, not really a good use case for it
abs(-1) // 1, not really a good use case for it
abs(add(fieldA,fieldB)) // absolute value of fieldA + fieldB
abs(fieldA) // absolute value of fieldA
```

## acos

The acos function returns the trigonometric arccosine of a number.

### acos Parameters

- Field Name | Raw Number | Number Evaluator: The value to return the arccosine of.

### acos Syntax

```
acos(100.4) // returns the arccosine of 100.4
acos(fieldA) // returns the arccosine for fieldA.
if(gt(fieldA,fieldB),sin(fieldA),sin(fieldB)) // if fieldA > fieldB then return the arccosine of
fieldA, else return the arccosine of fieldB
```

## add

The add function will take 2 or more numeric values and add them together. The add function will fail to execute if any of the values are non-numeric. If a null value is found then null will be returned as the result.

### add Parameters

- Field Name | Raw Number | Number Evaluator
- Field Name | Raw Number | Number Evaluator
- .....
- Field Name | Raw Number | Number Evaluator

### add Syntax

The expressions below show the various ways in which you can use the add evaluator. The number and order of these parameters do not matter and is not limited except that at least two parameters are required. Returns a numeric value.

```
add(1,2,3,4) // 1 + 2 + 3 + 4 == 10
add(1,fieldA) // 1 + value of fieldA
add(fieldA,1.4) // value of fieldA + 1.4
add(fieldA,fieldB,fieldC) // value of fieldA + value of fieldB + value of fieldC
add(fieldA,div(fieldA,fieldB)) // value of fieldA + (value of fieldA / value of fieldB)
add(fieldA,if(gt(fieldA,fieldB),fieldA,fieldB)) // if fieldA > fieldB then fieldA + fieldA, else
fieldA + fieldB
```

## analyze

The analyze function analyzes text using a Lucene/Solr analyzer and returns a list of tokens emitted by the analyzer. The analyze function can be called on its own or within the [select](#) and [cartesianProduct](#) streaming expressions.

### analyze Parameters

- Field Name | Raw Text: Either the field in a tuple or the raw text to be analyzed.
- Analyzer Field Name: The field name of the analyzer to use to analyze the text.

### analyze Syntax

The expressions below show the various ways in which you can use the analyze evaluator.

- Analyze the raw text: `analyze("hello world", analyzerField)`
- Analyze a text field within a select expression. This will annotate tuples with the output of the analyzer:  
`select(expr, analyze(textField, analyzerField) as outField)`
- Analyze a text field with a cartesianProduct expression. This will stream each token emitted by the analyzer in its own tuple:  
`cartesianProduct(expr, analyze(textField, analyzer) as outField)`

## and

The and function will return the logical AND of at least 2 boolean parameters. The function will fail to execute if any parameters are non-boolean or null. Returns a boolean value.

### and Parameters

- Field Name | Raw Boolean | Boolean Evaluator
- Field Name | Raw Boolean | Boolean Evaluator
- .....
- Field Name | Raw Boolean | Boolean Evaluator

### and Syntax

The expressions below show the various ways in which you can use the and evaluator. At least two parameters are required, but there is no limit to how many you can use.

```
and(true,fieldA) // true && fieldA
and(fieldA,fieldB) // fieldA && fieldB
and(or(fieldA,fieldB),fieldC) // (fieldA || fieldB) && fieldC
and(fieldA,fieldB,fieldC,or(fieldD,fieldE),fieldF)
```

## anova

The anova function calculates the [analysis of variance](#) for two or more numeric arrays.

### anova Parameters

- numeric array ... (two or more)

### anova Syntax

```
anova(numericArray1, numericArray2) // calculates ANOVA for two numeric arrays
anova(numericArray1, numericArray2, numericArray3) // calculates ANOVA for three numeric arrays
```

## array

The array function returns an array of numerics or other objects including other arrays.

### array Parameters

- numeric | array ...

### array Syntax

```
array(1, 2, 3) // Array of numerics
array(array(1,2,3), array(4,5,6)) // Array of arrays
```

## asin

The asin function returns the trigonometric arcsine of a number.

### asin Parameters

- Field Name | Raw Number | Number Evaluator: The value to return the arcsine of.

### asin Syntax

```
asin(100.4) // returns the sine of 100.4
asine(fieldA) // returns the sine for fieldA.
if(gt(fieldA,fieldB),asin(fieldA),asin(fieldB)) // if fieldA > fieldB then return the asine of
fieldA, else return the asine of fieldB
```

## atan

The atan function returns the trigonometric arctangent of a number.

### atan Parameters

- Field Name | Raw Number | Number Evaluator: The value to return the arctangent of.

### atan Syntax

```
atan(100.4) // returns the arctangent of 100.4
atan(fieldA) // returns the arctangent for fieldA.
if(gt(fieldA,fieldB),atan(fieldA),atan(fieldB)) // if fieldA > fieldB then return the arctangent
of fieldA, else return the arctangent of fieldB
```

## betaDistribution

The betaDistribution function returns a [beta probability distribution](#) based on its parameters. This function is part of the probability distribution framework and is designed to work with the [sample](#), [kolmogorovSmirnov](#) and [cumulativeProbability](#) functions.

### betaDistribution Parameters

- double: shape1
- double: shape2

## betaDistribution Returns

A probability distribution function.

## betaDistribution Syntax

```
betaDistribution(1, 5)
```

## binomialCoefficient

The binomialCoefficient function returns a [Binomial Coefficient](#), the number of k-element subsets that can be selected from an n-element set.

## binomialCoefficient Parameters

- integer: [n] set
- integer: [k] subset

## binomialCoefficient Returns

A long value: The number of k-element subsets that can be selected from an n-element set.

## binomialCoefficient Syntax

```
binomialCoefficient(8, 3) // Returns the number of 3 element subsets from an 8 element set.
```

## binomialDistribution

The binomialDistribution function returns a [binomial probability distribution](#) based on its parameters. This function is part of the probability distribution framework and is designed to work with the [sample](#), [probability](#) and [cumulativeProbability](#) functions.

## binomialDistribution Parameters

- integer: number of trials
- double: probability of success

## binomialDistribution Returns

A probability distribution function.

## binomialDistribution Syntax

```
binomialDistribution(1000, .5)
```

## cbrt

The cbrt function returns the trigonometric cube root of a number.

### cbrt Parameters

- Field Name | Raw Number | Number Evaluator: The value to return the cube root of.

### cbrt Syntax

```
cbrt(100.4) // returns the square root of 100.4
cbrt(fieldA) // returns the square root for fieldA.
if(gt(fieldA,fieldB),cbrt(fieldA),cbrt(fieldB)) // if fieldA > fieldB then return the cbrt of
fieldA, else return the cbrt of fieldB
```

## ceil

The ceil function rounds a decimal value to the next highest whole number.

### ceil Parameters

- Field Name | Raw Number | Number Evaluator: The decimal to round up.

### ceil Syntax

The expressions below show the various ways in which you can use the ceil evaluator.

```
ceil(100.4) // returns 101.
ceil(fieldA) // returns the next highest whole number for fieldA.
if(gt(fieldA,fieldB),ceil(fieldA),ceil(fieldB)) // if fieldA > fieldB then return the ceil of
fieldA, else return the ceil of fieldB.
```

## col

The col function returns a numeric array from a list of Tuples. The col function is used to create numeric arrays from stream sources.

### col Parameters

- list of Tuples
- field name: The field to create the array from.

### col Syntax

```
col(tupleList, fieldName)
```

## colAt

The colAt function returns the column of a matrix at a specific index as a numeric array.

### colAt Parameters

- matrix: the matrix to operate on
- integer: the index of the column to return

### colAt Syntax

```
colAt(matrix, 10)
```

### colAt Returns

numeric array: the column of the matrix

## columnCount

The columnCount function returns the number of columns in a matrix.

### columnCount Parameters

- matrix: the matrix to operate on

### columnCount Syntax

```
columnCount(matrix)
```

### columnCount Returns

integer: number columns in the matrix.

## constantDistribution

The constantDistribution function returns a constant probability distribution based on its parameter. This function is part of the probability distribution framework and is designed to work with the [sample](#) and [cumulativeProbability](#) functions.

When sampled the constant distribution always returns its constant value.

### constantDistribution Parameters

- double: constant value

## constantDistribution Returns

A probability distribution function.

## constantDistribution Syntax

```
constantDistribution(constantValue)
```

## conv

The conv function returns the [convolution](#) of two numeric arrays.

## conv Parameters

- numeric array
- numeric array

## conv Syntax

```
conv(numericArray1, numericArray2)
```

## copyOf

The copyOf function creates a copy of a numeric array.

## copyOf Parameters

- numeric array
- length: The length of the copied array. The returned array will be right padded with zeros if the length parameter exceeds the size of the original array.

## copyOf Syntax

```
copyOf(numericArray, length)
```

## copyOfRange

The copyOfRange function creates a copy of a range of a numeric array.

## copyOfRange Parameters

- numeric array
- start index
- end index

## copyOfRange Syntax

```
copyOfRange(numericArray, startIndex, endIndex)
```

## corr

The corr function returns the correlation of two numeric arrays or the correlation matrix for a matrix.

The corr function support Pearson's, Kendall's and Spearman's correlations.

### corr Positional Parameters

- numeric array: The first numeric array
- numeric array: The second numeric array

OR

- matrix: The matrix to compute the correlation matrix for. Note that correlation is computed between the columns in the matrix.

### corr Named Parameters

- type: (Optional) The type of correlation. Possible values are pearsons, kendalls, or spearmans. The default is pearsons.

## corr Syntax

```
corr(numericArray1, numericArray2) // Compute the Pearsons correlation for two numeric arrays
corr(numericArray1, numericArray2, type=kendalls) // Compute the Kendalls correlation for two numeric arrays
corr(matrix) // Compute the Pearsons correlation matrix for a matrix
corr(matrix, type=spearmans) // Compute the Spearmans correlation matrix for a matrix
```

## corr Returns

number | matrix: Either the correlation or correlation matrix.

## cos

The cos function returns the trigonometric cosine of a number.

### cos Parameters

- Field Name | Raw Number | Number Evaluator: The value to return the hyperbolic cosine of.

## cos Syntax

```
cos(100.4) // returns the arccosine of 100.4
cos(fieldA) // returns the arccosine for fieldA.
if(gt(fieldA,fieldB),cos(fieldA),cos(fieldB)) // if fieldA > fieldB then return the arccosine of
fieldA, else return the cosine of fieldB
```

## cosineSimilarity

The cosineSimilarity function returns the [cosine similarity](#) of two numeric arrays.

### cosineSimilarity Parameters

- numeric array
- numeric array

### cosineSimilarity Returns

A numeric.

## cosineSimilarity Syntax

```
cosineSimilarity(numericArray, numericArray)
```

## cov

The cov function returns the covariance of two numeric array or the covariance matrix for matrix.

### cov Parameters

- numeric array: The first numeric array
- numeric array: The second numeric array

OR

- matrix: The matrix to compute the covariance matrix from. Note that covariance is computed between the columns in the matrix.

## cov Syntax

```
cov(numericArray, numericArray) // Computes the covariance of a two numeric arrays
cov(matrix) // Computes the covariance matrix for the matrix.
```

## cov Returns

number | matrix: Either the covariance or covariance matrix.

## cumulativeProbability

The `cumulativeProbability` function returns the cumulative probability of a random variable within a probability distribution. The cumulative probability is the total probability of all random variables less than or equal to a random variable.

### cumulativeProbability Parameters

- probability distribution
- number: Value to compute the probability for.

### cumulativeProbability Returns

A double: the cumulative probability.

### cumulativeProbability Syntax

```
cumulativeProbability(normalDistribution(500, 25), 502) // Returns the cumulative probability of
the random sample 502 in a normal distribution with a mean of 500 and standard deviation of 25.
```

## derivative

The derivative function returns the [derivative](#) of a function. The derivative function can compute the derivative of the [spline](#) function and the [loess](#) function. The derivative can also take the derivative of a derivative.

### derivative Parameters

- spline | loess | akima | lerp | derivative: The functions to compute the derivative for.

### derivative Syntax

```
derivative(spline(...))
derivative(loess(...))
derivative(derivative(...))
```

### derivative Returns

function: The function can be treated as both a numeric array and function.

## describe

The describe function returns a tuple containing the descriptive statistics for an array.

## describe Parameters

- numeric array

## describe Syntax

```
describe(numericArray)
```

## diff

The diff function performs [time series differencing](#).

Time series differencing is often used to make a time series stationary before further analysis.

## diff Parameters

- numeric array: The time series data.
- integer: (Optional) The lag. Defaults to 1.

## diff Syntax

```
diff(numericArray1) // Perform time series differencing with a default lag of 1.
diff(numericArray1, 30) // Perform time series differencing with a lag of 30.
```

## diff Returns

numeric array: The differenced time series data. The size of the array will be equal to (original array size - lag).

## distance

The distance function computes the distance of two numeric arrays or the distance matrix for a matrix.

## distance Positional Parameters

- numeric array: The first numeric array
- numeric array: The second numeric array

OR

- matrix: The matrix to compute the distance matrix for. Note that distance is computed between the columns in the matrix.

## distance Named Parameters

- type: (Optional) The distance type. Possible values are euclidean, manhattan, canberra, or earthMovers. The default is euclidean.

## distance Syntax

```
distance(numericArray1, numericArray2) // Computes the euclidean distance for two numeric arrays.
distance(numericArray1, numericArray2, type=manhattan) // Computes the manhattan distance for two
numeric arrays.
distance(matrix) // Computes the euclidean distance matrix for a matrix.
distance(matrix, type=canberra) // Computes the canberra distance matrix for a matrix.
```

## distance Returns

number | matrix: Either the distance or distance matrix.

## div

The div function will take two numeric values and divide them. The function will fail to execute if any of the values are non-numeric or null, or the 2nd value is 0. Returns a numeric value.

## div Parameters

- Field Name | Raw Number | Number Evaluator
- Field Name | Raw Number | Number Evaluator

## div Syntax

The expressions below show the various ways in which you can use the div evaluator. The first value will be divided by the second and as such the second cannot be 0.

```
div(1,2) // 1 / 2
div(1,fieldA) // 1 / fieldA
div(fieldA,1.4) // fieldA / 1.4
div(fieldA,add(fieldA,fieldB)) // fieldA / (fieldA + fieldB)
```

## dotProduct

The dotProduct function returns the [dotproduct](#) of two numeric arrays.

## dotProduct Parameters

- numeric array
- numeric array

## dotProduct Returns

A number.

## dotProduct Syntax

```
dotProduct(numericArray, numericArray)
```

## ebeAdd

The ebeAdd function performs an element-by-element addition of two numeric arrays.

### ebeAdd Parameters

- numeric array
- numeric array

### ebeAdd Returns

A numeric array.

## ebeAdd Syntax

```
ebeAdd(numericArray, numericArray)
```

## ebeDivide

The ebeDivide function performs an element-by-element division of two numeric arrays.

### ebeDivide Parameters

- numeric array
- numeric array

### ebeDivide Returns

A numeric array.

## ebeDivide Syntax

```
ebeDivide(numericArray, numericArray)
```

## ebeMultiple

The ebeMultiply function performs an element-by-element multiplication of two numeric arrays.

### ebeMultiply Parameters

- numeric array

- numeric array

## ebeMultiply Returns

A numeric array.

## ebeMultiply Syntax

```
ebeMultiply(numericArray, numericArray)
```

## ebeSubtract

The ebeSubtract function performs an element-by-element subtraction of two numeric arrays.

## ebeSubtract Parameters

- numeric array
- numeric array

## ebeSubtract Returns

A numeric array.

## ebeSubtract Syntax

```
ebeSubtract(numericArray, numericArray)
```

## empiricalDistribution

The empiricalDistribution function returns [empirical distribution function](#), a continuous probability distribution function based on an actual data set. This function is part of the probability distribution framework and is designed to work with the [sample](#), [kolmogorovSmirnov](#) and [cumulativeProbability](#) functions.

This function is designed to work with continuous data. To build a distribution from a discrete data set use the [enumeratedDistribution](#).

## empiricalDistribution Parameters

- numeric array: empirical observations

## empiricalDistribution Returns

A probability distribution function.

## empiricalDistribution Syntax

```
empiricalDistribution(numericArray)
```

## enumeratedDistribution

The enumeratedDistribution function returns a discrete probability distribution function based on an actual data set or a pre-defined set of data and probabilities. This function is part of the probability distribution framework and is designed to work with the [sample](#), [probability](#) and [cumulativeProbability](#) functions.

The enumeratedDistribution can be called in two different scenarios:

- 1) Single array of discrete values. This works like an empirical distribution for discrete data.
- 2) An array of singleton discrete values and an array of double values representing the probabilities of the discrete values.

This function is designed to work with discrete data. To build a distribution from a continuous data set use the [empiricalDistribution](#).

## enumeratedDistribution Parameters

- integer array: discrete observations or singleton discrete values.
- double array: (Optional) values representing the probabilities of the singleton discrete values.

## enumeratedDistribution Returns

A probability distribution function.

## enumeratedDistribution Syntax

```
enumeratedDistribution(integerArray) // This creates an enumerated distribution from the
observations in the numeric array.
enumeratedDistribution(array(1,2,3,4), array(.25,.25,.25,.25)) // This creates an enumerated
distribution with four discrete values (1,2,3,4) each with a probability of .25.
```

## eor

The eor function will return the logical exclusive or of at least two boolean parameters. The function will fail to execute if any parameters are non-boolean or null. Returns a boolean value.

## eor Parameters

- Field Name | Raw Boolean | Boolean Evaluator
- Field Name | Raw Boolean | Boolean Evaluator
- .....

- Field Name | Raw Boolean | Boolean Evaluator

## eor Syntax

The expressions below show the various ways in which you can use the eor evaluator. At least two parameters are required, but there is no limit to how many you can use.

```
eor(true,fieldA) // true iff fieldA is false
eor(fieldA,fieldB) // true iff either fieldA or fieldB is true but not both
eor(eq(fieldA,fieldB),eq(fieldC,fieldD)) // true iff either fieldA == fieldB or fieldC == fieldD
but not both
```

## eq

The eq function will return whether all the parameters are equal, as per Java's standard equals(...) function. The function accepts parameters of any type, but will fail to execute if all the parameters are not of the same type. That is, all are Boolean, all are String, or all are Numeric. If any any parameters are null and there is at least one parameter that is not null then false will be returned. Returns a boolean value.

## eq Parameters

- Field Name | Raw Value | Evaluator
- Field Name | Raw Value | Evaluator
- .....
- Field Name | Raw Value | Evaluator

## eq Syntax

The expressions below show the various ways in which you can use the eq evaluator.

```
eq(1,2) // 1 == 2
eq(1,fieldA) // 1 == fieldA
eq(fieldA,val(foo)) fieldA == "foo"
eq(add(fieldA,fieldB),6) // fieldA + fieldB == 6
```

## expMovingAge

The expMovingAverage function computes an [exponential moving average](#) for a numeric array.

## expMovingAge Parameters

- numeric array: The array to compute the exponential moving average from.
- integer: window size

## expMovingAvg Returns

A numeric array. The first element of the returned array will start from the windowSize-1 index of the

original array.

## expMovingAvg Syntax

```
expMovingAvg(numericArray, 5) //Computes an exponential moving average with a window size of 5.
```

## factorial

The factorial function returns the [factorial](#) of its parameter.

### factorial Parameters

- integer: The value to compute the factorial for. The largest supported value of this parameter is 170.

### factorial Returns

A double.

## factorial Syntax

```
factorial(100) //Computes the factorial of 100
```

## finddelay

The finddelay function performs a cross-correlation between two numeric arrays and returns the delay.

### finddelay Parameters

- numeric array
- numeric array

## finddelay Syntax

```
finddelay(numericArray1, numericArray2)
```

## floor

The floor function rounds a decimal value to the next lowest whole number.

### floor Parameters

- Field Name | Raw Number | Number Evaluator: The decimal to round down.

## floor Syntax

The expressions below show the various ways in which you can use the floor evaluator.

```
floor(100.4) // returns 100.
ceil(fieldA) // returns the next lowestt whole number for fieldA.
if(gt(fieldA,fieldB),floor(fieldA),floor(fieldB)) // if fieldA > fieldB then return the floor of
fieldA, else return the floor of fieldB.
```

## freqTable

The freqTable function returns a [frequency distribution](#) from an array of discrete values.

This function is designed to work with discrete values. To work with continuous data use the [hist](#) function.

### freqTable Parameters

- integer array: The values to build the frequency distribution from.

### freqTable Returns

A list of tuples containing the frequency information for each discrete value.

### freqTable Syntax

```
freqTable(integerArray)
```

## gammaDistribution

The gammaDistribution function returns a [gamma probability distribution](#) based on its parameters. This function is part of the probability distribution framework and is designed to work with the [sample](#), [kolmogorovSmirnov](#) and [cumulativeProbability](#) functions.

### gammaDistribution Parameters

- double: shape
- double: scale

### gammaDistribution Returns

A probability distribution function,

### gammaDistribution Syntax

```
gammaDistribution(1, 10)
```

## geometricDistribution

The geometricDistribution function returns a [geometric probability distribution](#) based on its parameters.

This function is part of the probability distribution framework and is designed to work with the [sample](#), [probability](#) and [cumulativeProbability](#) functions.

## geometricDistribution Parameters

- double: probability

## geometricDistribution Syntax

```
geometricDistribution(.5) // Creates a geometric distribution with probability of .5
```

## geometricDistribution Returns

A probability distribution function

## getAttribute

The `getAttribute` function returns an attribute from a matrix by its key. Any function that returns a matrix can also set attributes on the matrix with additional information. The `setAttribute` function can also be used to set attributes on a matrix. The key to an attribute is always a string. The value of attribute can be any object including numerics, arrays, maps, matrixes, etc.

## getAttribute Parameters

- matrix: The matrix to set the attribute on
- string: The key for the attribute

## getAttribute Syntax

```
getAttribute(matrix, key)
```

## getAttribute Returns

object: any object

## getAttributes

The `getAttributes` function returns the attribute map from matrix. See the `getAttribute` function for more details on attributes.

## getAttributes Parameters

- matrix: The matrix to retrieve the attribute map from.

## getAttributes Syntax:

```
getAttributes(matrix)
```

## getAttributes Returns

map: The map of attributes.

## getColumnLabels

The getColumnLabels function returns the columns labels of a matrix. The column labels can be optionally set by any function that returns a matrix. The column labels can also be set via the setColumnLabels function.

### getColumnLabels Parameters

- matrix: The matrix to return the column labels of.

### getColumnLabels Syntax

```
getColumnLabels(matrix)
```

## getColumnLabels Returns

string array: The labels for each column in the matrix

## getRowLabels

The getRowLabels function returns the row labels of a matrix. The row labels can be optionally set by any function that returns a matrix. The row labels can also be set via the setRowLabels function.

### getRowLabels Parameters

- matrix: The matrix to return the row labels from.

### getRowLabels Syntax

```
getRowLabels(matrix)
```

## getRowLabels Returns

string array: The labels for each row in the matrix

## getValue

The getValue function returns the value of a single Tuple entry by key.

### getValue Parameters

- tuple: The Tuple to return the entry from.

- key: The key of the entry to return the value for.

## getValue Syntax

getValue(tuple, key)

## getValue Returns

object: Returns an object of the same type as the Tuple entry.

## grandSum

The grandSum function sums all the values in a matrix.

## grandSum Parameters

- matrix: The matrix to operate on.

## grandSum Syntax

```
grandSum(matrix)
```

## grandSum Returns

number: the sum of all the values in the matrix.

## gt

The gt function will return whether the first parameter is greater than the second parameter. The function accepts numeric or string parameters, but will fail to execute if all the parameters are not of the same type. That is, all are String or all are Numeric. If any any parameters are null then an error will be raised. Returns a boolean value.

## gt Parameters

- Field Name | Raw Value | Evaluator
- Field Name | Raw Value | Evaluator

## gt Syntax

The expressions below show the various ways in which you can use the gt evaluator.

```
gt(1,2) // 1 > 2
gt(1,fieldA) // 1 > fieldA
gt(fieldA,val(foo)) // fieldA > "foo"
gt(add(fieldA,fieldB),6) // fieldA + fieldB > 6
```

## gteq

The gteq function will return whether the first parameter is greater than or equal to the second parameter. The function accepts numeric and string parameters, but will fail to execute if all the parameters are not of the same type. That is, all are String or all are Numeric. If any any parameters are null then an error will be raised. Returns a boolean value.

### gteq Parameters

- Field Name | Raw Value | Evaluator
- Field Name | Raw Value | Evaluator

### gteq Syntax

The expressions below show the various ways in which you can use the gteq evaluator.

```
gteq(1,2) // 1 >= 2
gteq(1,fieldA) // 1 >= fieldA
gteq(fieldA,val(foo)) fieldA >= "foo"
gteq(add(fieldA,fieldB),6) // fieldA + fieldB >= 6
```

## hist

The hist function creates a histogram from a numeric array. The hist function is designed to work with continuous variables.

### hist Parameters

- numeric array
- bins: The number of bins in the histogram. Each returned tuple contains summary statistics for the observations that were within the bin.

### hist Syntax

```
hist(numericArray, bins)
```

## hsin

The hsin function returns the trigonometric hyperbolic sine of a number.

### hsin Parameters

- Field Name | Raw Number | Number Evaluator: The value to return the hyperbolic sine of.

## hsin Syntax

```
hsin(100.4) // returns the hsine of 100.4
hsin(fieldA) // returns the hsine for fieldA.
if(gt(fieldA,fieldB),sin(fieldA),sin(fieldB)) // if fieldA > fieldB then return the hsine of
fieldA, else return the hsine of fieldB
```

## if

The if function works like a standard conditional if/then statement. If the first parameter is true, then the second parameter will be returned, else the third parameter will be returned. The function accepts a boolean as the first parameter and anything as the second and third parameters. An error will occur if the first parameter is not a boolean or is null.

### if Parameters

- Field Name | Raw Value | Boolean Evaluator
- Field Name | Raw Value | Evaluator
- Field Name | Raw Value | Evaluator

### if Syntax

The expressions below show the various ways in which you can use the if evaluator.

```
if(fieldA,fieldB,fieldC) // if fieldA is true then fieldB else fieldC
if(gt(fieldA,5), fieldA, 5) // if fieldA > 5 then fieldA else 5
if(eq(fieldB,null), null, div(fieldA,fieldB)) // if fieldB is null then null else fieldA / fieldB
```

## indexOf

The indexOf function returns the index of a string in an array of strings.

### indexOf Parameters

- string array: The array to operate on.
- string: The string to search for in the array.

### indexOf Syntax

```
indexOf(stringArray, string)
```

### indexOf Returns

integer: The index of the string in the array or -1 if the string is not found.

## integrate

The integrate function computes the integral of an interpolation function for a specific range of the curve.

### integrate Parameters

- spline | akima | lerp | loess: The interpolation function to compute the integral for.
- numeric: start of integral range
- numeric: end of integral range

### integrate Syntax

```
integrate(function, start, end)
```

### integrate Returns

numeric: The integral

## length

The length function returns the length of a numeric array.

### length Parameters

- numeric array

### length Syntax

```
length(numericArray)
```

## lerp (TODO)

## loess

The loess function is a smoothing curve fitter which uses a [local regression](#) algorithm. Unlike the [spline](#) function which touches each control point, the loess function puts a smooth curve through the control points without having to touch the control points. The loess result can be used by the [derivative](#) function to produce smooth derivatives from data that is not smooth.

### loess Positional Parameters

- numeric array: (Optional) x values. If omitted a sequence will be created for the x values.
- numeric array: y values

## loess Named Parameters

- bandwidth: (Optional) The percent of the data points to use when drawing the local regression line, defaults to .25. Decreasing the bandwidth increases the number of curves that loess can fit.
- robustIterations: (Optional) The number of iterations used to smooth outliers, defaults to 2.

## loess Syntax

```
loess(yValues) // This creates the xValues automatically and fits a smooth curve through the data points.
loess(xValues, yValues) // This will fit a smooth curve through the data points.
loess(xValues, yValues, bandwidth=.15) // This will fit a smooth curve through the data points using 15 percent of the data points for each local regression line.
```

## loess Returns

function: The function can be treated as both a numeric array of the smoothed data points and function.

## log

The log function will return the natural log of the provided single parameter. The log function will fail to execute if the value is non-numeric. If a null value is found, then null will be returned as the result.

## log Parameters

- Field Name | Raw Number | Number Evaluator

## log Syntax

The expressions below show the various ways in which you can use the log evaluator. Only one parameter is accepted. Returns a numeric value.

```
log(100)
log(add(fieldA,fieldB))
log(fieldA)
```

## logNormalDistribution

The logNormalDistribution function returns a [log normal probability distribution](#) based on its parameters. This function is part of the probability distribution framework and is designed to work with the [sample](#), [kolmogorovSmirnov](#) and [cumulativeProbability](#) functions.

## logNormalDistribution Parameters

- double: shape
- double: scale

## logNormalDistribution Returns

A probability distribution function.

## logNormalDistribution Syntax

```
logNormalDistribution(.3, .0)
```

## kolmogorovSmirnov

The kolmogorovSmirnov function performs a [Kolmogorov Smirnov test](#), between a reference continuous probability distribution and a sample set.

The supported distribution functions are: [empiricalDistribution](#), [normalDistribution](#), [logNormalDistribution](#), [weibullDistribution](#), [gammaDistribution](#), and [betaDistribution](#).

## kolmogorovSmirnov Parameters

- continuous probability distribution: Reference distribution
- numeric array: sample set

## kolmogorovSmirnov Returns

A result tuple: A tuple containing the p-value and d-statistic for the test result.

## kolmogorovSmirnov Syntax

```
kolmogorovSmirnov(normalDistribution(10, 2), sampleSet)
```

## lt

The lt function will return whether the first parameter is less than the second parameter. The function accepts numeric or string parameters, but will fail to execute if all the parameters are not of the same type. That is, all are String or all are Numeric. If any any parameters are null then an error will be raised. Returns a boolean value.

## lt Parameters

- Field Name | Raw Value | Evaluator
- Field Name | Raw Value | Evaluator

## lt Syntax

The expressions below show the various ways in which you can use the lt evaluator.

```
lt(1,2) // 1 < 2
lt(1,fieldA) // 1 < fieldA
lt(fieldA,val(foo)) fieldA < "foo"
lt(add(fieldA,fieldB),6) // fieldA + fieldB < 6
```

## Iteq

The `lteq` function will return whether the first parameter is less than or equal to the second parameter. The function accepts numeric and string parameters, but will fail to execute if all the parameters are not of the same type. That is, all are String or all are Numeric. If any any parameters are null then an error will be raised. Returns a boolean value.

### Iteq Parameters

- Field Name | Raw Value | Evaluator
- Field Name | Raw Value | Evaluator

### Iteq Syntax

The expressions below show the various ways in which you can use the `lteq` evaluator.

```
lteq(1,2) // 1 <= 2
lteq(1,fieldA) // 1 <= fieldA
lteq(fieldA,val(foo)) fieldA <= "foo"
lteq(add(fieldA,fieldB),6) // fieldA + fieldB <= 6
```

## markovChain

The `markovChain` function can be used to perform [Markov Chain](#) simulations. The `markovChain` function takes as its parameter a [transition matrix](#) and returns a mathematical model that can be sampled using the [sample](#) function. Each sample taken from the Markov Chain represents the current state of system.

### markovChain Parameters

- `matrix`: Transition matrix

### markovChain Syntax

```
sample(markovChain(transitionMatrix), 5) // This creates a Markov Chain given a specific
 transition matrix. The sample function takes 5 samples from the Markov Chain, representing the
 next five states of the system.
```

### markovChain Returns

Markov Chain model: The Markoff Chain model can be used with [sample](#) function.

## matrix

The matrix function returns a [matrix](#) which can be operated on by functions that support matrix operations.

### matrix Parameters

- numeric array ...: One or more numeric arrays that will be the rows of the matrix.

### matrix Syntax

```
matrix(numericArray1, numericArray2, numericArray3) // Returns a matrix with three rows of data:
numericaArray1, numericArray2, numericArray3
```

### matrix Returns

matrix

## meanDifference

The meanDifference function calculates the mean of the differences following the element-by-element subtraction between two numeric arrays.

### meanDifference Parameters

- numeric array
- numeric array

### meanDifference Returns

A numeric.

### meanDifference Syntax

```
meanDifference(numericArray, numericArray)
```

## minMaxScale

The minMaxScale function scales numeric arrays within a minimum and maximum value. By default minMaxScale scales between 0 and 1. The minMaxScale function can operate on both numeric arrays and matrices.

When operating on a matrix the minMaxScale function operates on each row of the matrix.

### minMaxScale Parameters

- numeric array | matrix: The array or matrix to scale

- double: (Optional) The min value. Defaults to 0.
- double: (Optional) The max value. Defaults to 1.

## minMaxScale Syntax

```
minMaxScale(numericArray) // scale a numeric array between 0 and 1
minMaxScale(numericArray, 0, 100) // scale a numeric array between 1 and 100
minMaxScale(matrix) // Scale each row in a matrix between 0 and 1
minMaxScale(matrix, 0, 100) // Scale each row in a matrix between 0 and 100
```

## minMaxScale Returns

A numeric array or matrix

## mod

The mod function returns the remainder (modulo) of the first parameter divided by the second parameter.

## mod Parameters

- Field Name | Raw Number | Number Evaluator: Parameter 1
- Field Name | Raw Number | Number Evaluator: Parameter 2

## mod Syntax

The expressions below show the various ways in which you can use the mod evaluator.

```
mod(100,3) // returns the remainder of 100 / 3 .
mod(100,fieldA) // returns the remainder of 100 divided by the value of fieldA.
mod(fieldA,1.4) // returns the remainder of fieldA divided by 1.4.
if(gt(fieldA,fieldB),mod(fieldA,fieldB),mod(fieldB,fieldA)) // if fieldA > fieldB then return the
remainder of fieldA/fieldB, else return the remainder of fieldB/fieldA.
```

## monteCarlo

The monteCarlo function performs a [Monte Carlo simulation](#) based on its parameters. The monteCarlo function runs another function a specified number of times and returns the results. The function being run typically has one or more variables that are drawn from probability distributions on each run. The [sample](#) function is used in the function to draw the samples.

The simulation's result array can then be treated as an empirical distribution to understand the probabilities of the simulation results.

## monteCarlo Parameters

- numeric function: The function being run by the simulation, which must return a numeric value.

- integer: The number of times to run the function.

## monteCarlo Returns

A numeric array: The results of simulation runs.

## monteCarlo Syntax

```
let(a=uniformIntegerDistribution(1, 6),
 b=uniformIntegerDistribution(1, 6),
 c=monteCarlo(add(sample(a), sample(b)), 1000))
```

In the expression above the monteCarlo function is running the function add(sample(a), sample(b)) 1000 times and returning the result. Each time the function is run samples are drawn from the probability distributions stored in variables a and b.

## movingAvg

The movingAvg function calculates a [moving average](#) over an array of numbers.

## movingAvg Parameters

- numeric array
- window size

## movingAvg Returns

A numeric array. The first element of the returned array will start from the windowSize-1 index of the original array.

## movingAvg Syntax

```
movingAverage(numericArray, 30)
```

## movingMedian

The movingMedian function calculates a moving median over an array of numbers.

## movingMedian Parameters

- numeric array
- window size

## movingMedian Returns

A numeric array. The first element of the returned array will start from the windowSize-1 index of the original array.

## movingMedian Syntax

```
movingMedian(numericArray, 30)
```

## mult

The `mult` function will take two or more numeric values and multiply them together. The `mult` function will fail to execute if any of the values are non-numeric. If a null value is found then null will be returned as the result.

### mult Parameters

- Field Name | Raw Number | Number Evaluator
- Field Name | Raw Number | Number Evaluator
- .....
- Field Name | Raw Number | Number Evaluator

### mult Syntax

The expressions below show the various ways in which you can use the `mult` evaluator. The number and order of these parameters do not matter and is not limited except that at least two parameters are required. Returns a numeric value.

```
mult(1,2,3,4) // 1 * 2 * 3 * 4
mult(1,fieldA) // 1 * value of fieldA
mult(fieldA,1.4) // value of fieldA * 1.4
mult(fieldA,fieldB,fieldC) // value of fieldA * value of fieldB * value of fieldC
mult(fieldA,div(fieldA,fieldB)) // value of fieldA * (value of fieldA / value of fieldB)
mult(fieldA,if(gt(fieldA,fieldB),fieldA,fieldB)) // if fieldA > fieldB then fieldA * fieldA, else
fieldA * fieldB
```

## normalDistribution

The `normalDistribution` function returns a [normal probability distribution](#) based on its parameters. This function is part of the probability distribution framework and is designed to work with the [sample](#), [kolmogorovSmirnov](#) and [cumulativeProbability](#) functions.

### normalDistribution Parameters

- double: mean
- double: standard deviation

### normalDistribution Returns

A probability distribution function.

## normalDistribution Syntax

```
normalDistribution(mean, stddev)
```

## normalizeSum

The `normalizeSum` function scales numeric arrays so that they sum to 1. The `normalizeSum` function can operate on both numeric arrays and matrices.

When operating on a matrix the `normalizeSum` function operates on each row of the matrix.

## normalizeSum Parameters

- numeric array | matrix

## normalizeSum Syntax

```
normalizeSum(numericArray)
normalizeSum(matrix)
```

## normalizeSum Returns

numeric array | matrix

## not

The `not` function will return the logical NOT of a single boolean parameter. The function will fail to execute if the parameter is non-boolean or null. Returns a boolean value.

## not Parameters

- Field Name | Raw Boolean | Boolean Evaluator

## not Syntax

The expressions below show the various ways in which you can use the `not` evaluator. Only one parameter is allowed.

```
not(true) // false
not(fieldA) // true if fieldA is false else false
not(eq(fieldA,fieldB)) // true if fieldA != fieldB
```

## olsRegress

The `olsRegress` function performs [ordinary least squares](#), multivariate, linear regression.

The olsRegress function returns a single Tuple containing the regression model with estimated regression parameters, RSquared and regression diagnostics.

The output of olsRegress can be used with the [predict](#) function to predict values based on the regression model.

## olsRegress Parameters

- `matrix`: The regressor observation matrix. Each row in the matrix represents a single multi-variate regressor observation. Note that there is no need to add an initial unitary column (column of 1's) when specifying a model including an intercept term, this column will be added automatically.
- `numeric array`: The outcomes array which matches up with each row in the regressor observation matrix.

## olsRegress Syntax

```
olsRegress(matrix, numericArray) // This performs the olsRegression analysis on given regressor matrix and outcome array.
```

## olsRegress Returns

Tuple: The regression model including the estimated regression parameters and diagnostics.

## or

The or function will return the logical OR of at least 2 boolean parameters. The function will fail to execute if any parameters are non-boolean or null. Returns a boolean value.

## or Parameters

- Field Name | Raw Boolean | Boolean Evaluator
- Field Name | Raw Boolean | Boolean Evaluator
- .....
- Field Name | Raw Boolean | Boolean Evaluator

## or Syntax

The expressions below show the various ways in which you can use the or evaluator. At least two parameters are required, but there is no limit to how many you can use.

```
or(true,fieldA) // true || fieldA
or(fieldA,fieldB) // fieldA || fieldB
or(and(fieldA,fieldB),fieldC) // (fieldA && fieldB) || fieldC
or(fieldA,fieldB,fieldC, and(fieldD,fieldE),fieldF)
```

## poissonDistribution

The poissonDistribution function returns a [poisson probability distribution](#) based on its parameter. This function is part of the probability distribution framework and is designed to work with the [sample](#), [probability](#) and [cumulativeProbability](#) functions.

### poissonDistribution Parameters

- double: mean

### poissonDistribution Returns

A probability distribution function.

### poissonDistribution Syntax

```
poissonDistribution(mean)
```

## polyFit

The polyFit function performs [polynomial curve fitting](#).

### polyFit Parameters

- numeric array: (Optional) x values. If omitted a sequence will be created for the x values.
- numeric array: y values
- integer: (Optional) polynomial degree. Defaults to 3.

### polyFit Returns

A numeric array: curve that was fit to the data points.

### polyFit Syntax

```
polyFit(yValues) // This creates the xValues automatically and fits a curve through the data
points using the default 3 degree polynomial.
polyFit(yValues, 5) // This creates the xValues automatically and fits a curve through the data
points using a 5 degree polynomial.
polyFit(xValues, yValues, 5) // This will fit a curve through the data points using a 5 degree
polynomial.
```

## pow

The pow function returns the value of its first parameter raised to the power of its second parameter.

## pow Parameters

- Field Name | Raw Number | Number Evaluator: Parameter 1
- Field Name | Raw Number | Number Evaluator: Parameter 2

## pow Syntax

The expressions below show the various ways in which you can use the pow evaluator.

```
pow(2,3) // returns 2 raised to the 3rd power.
pow(4,fieldA) // returns 4 raised by the value of fieldA.
pow(fieldA,1.4) // returns the value of fieldA raised by 1.4.
if(gt(fieldA,fieldB),pow(fieldA,fieldB),pow(fieldB,fieldA)) // if fieldA > fieldB then raise
fieldA by fieldB, else raise fieldB by fieldA.
```

## predict

The predict function predicts the value of dependent variables based on regression models or functions.

The predict function can predict values based on the output of the following functions: [spline](#), [loess](#), [regress](#), [olsRegress](#).

## predict Parameters

- regression model | function: The model or function used for the prediction
- number | numeric array | matrix: Depending on the regression model or function used, the predictor variable can be a number, numeric array or matrix.

## predict Syntax

```
predict(regressModel, number) // predict using the output of the <><regress>> function and single numeric predictor. This will return a single numeric prediction.
```

```
predict(regressModel, numericArray) // predict using the output of the <><regress>> function and a numeric array of predictors. This will return a numeric array of predictions.
```

```
predict(splineFunc, number) // predict using the output of the <><spline>> function and single numeric predictor. This will return a single numeric prediction.
```

```
predict(splineFunc, numericArray) // predict using the output of the <><spline>> function and a numeric array of predictors. This will return a numeric array of predictions.
```

```
predict(olsRegressModel, numericArray) // predict using the output of the <><olsRegress>> function and a numeric array containing one multi-variate predictor. This will return a single numeric prediction.
```

```
predict(olsRegressModel, matrix) // predict using the output of the <><olsRegress>> function and a matrix containing rows of multi-variate predictor arrays. This will return a numeric array of predictions.
```

## primes

The primes function returns an array of prime numbers starting from a specified number.

### primes Parameters

- integer: The number of primes to return in the list
- integer: The starting point for returning the primes

### primes Returns

A numeric array.

### primes Syntax

```
primes(100, 2000) // returns 100 primes starting from 2000
```

## probability

The probability function returns the probability of a random variable within a probability distribution.

The probability function computes the probability between random variable ranges for both [continuous](#) and [discrete](#) probability distributions.

The probability function can compute probabilities for a specific random variable for discrete probability distributions only.

The supported continuous distribution functions are: [normalDistribution](#), [logNormalDistribution](#),

[betaDistribution](#), [gammaDistribution](#), [empiricalDistribution](#), [triangularDistribution](#), [weibullDistribution](#), [uniformDistribution](#), [constantDistribution](#)

The supported discrete distributions are: [poissonDistribution](#), [binomialDistribution](#), [enumeratedDistribution](#), [zipFDistribution](#), [geometricDistribution](#), [uniformIntegerDistribution](#)

## probability Parameters

- probability distribution: the probability distribution to compute the probability from.
- number: low value of the range.
- number: (Optional for discrete probability distributions) high value of the range. If the high range is omitted then the probability function will compute a probability for the low range value.

## probability Syntax

```
probability(poissonDistribution(10), 7) // Returns the probability of a random sample of 7 in a poisson distribution with a mean of 10.
```

```
probability(normalDistribution(10, 2), 7.5, 8.5) // Returns the probability between the range of 7.5 to 8.5 for a normal distribution with a mean of 10 and standard deviation of 2.
```

## probability Returns

double: probability

## rank

The rank performs a rank transformation on a numeric array.

## rank Parameters

- numeric array

## rank Syntax

```
rank(numericArray)
```

## raw

The raw function will return whatever raw value is the parameter. This is useful for cases where you want to use a string as part of another evaluator.

## raw Parameters

- Raw Value

## raw Syntax

The expressions below show the various ways in which you can use the raw evaluator. Whatever is inside will be returned as-is. Internal evaluators are considered strings and are not evaluated.

```
raw(foo) // "foo"
raw(count(*)) // "count(*)"
raw(45) // 45
raw(true) // "true" (note: this returns the string "true" and not the boolean true)
eq(raw(fieldA), fieldA) // true if the value of fieldA equals the string "fieldA"
```

## regress

The regress function performs a simple regression of two numeric arrays.

The result of this expression is also used by the [predict](#) function.

### regress Parameters

- numeric array
- numeric array

## regress Syntax

```
regress(numericArray1, numericArray2)
```

## rev

The rev function reverses the order of a numeric array.

### rev Parameters

- numeric array

## rev Syntax

```
rev(numericArray)
```

## round

The round function returns the closest whole number to the argument.

### round Parameters

- Field Name | Raw Number | Number Evaluator: The value to return the square root of.

## round Syntax

```
round(100.4)
round(fieldA)
if(gt(fieldA,fieldB),sqrt(fieldA),sqrt(fieldB)) // if fieldA > fieldB then return the round of
fieldA, else return the round of fieldB
```

## rowAt

The rowAt function returns the row of a matrix at a specific index as a numeric array.

### rowAt Parameters

- matrix: the matrix to operate on
- integer: the index of the row to return

## rowAt Syntax

```
rowAt(matrix, 10)
```

## rowAt Returns

numeric array: the row of the matrix

## rowCount

The rowCount function returns the number of rows in a matrix.

### rowCount Parameters

- matrix: the matrix to operate on

## rowCount Syntax

```
rowCount(matrix)
```

## rowCount Returns

integer: number rows in the matrix.

## sample

The sample function can be used to draw random samples from a probability distribution or Markov Chain.

## sample Parameters

- probability distribution | Markov Chain: The distribution or Markov Chain to sample.
- integer: (Optional) Sample size. Defaults to 1.

## sample Returns

Either a single numeric random sample, or a numeric array depending on the sample size parameter.

## sample Syntax

```
sample(poissonDistribution(5)) // Returns a single random sample from a poissonDistribution with
mean of 5.
sample(poissonDistribution(5), 1000) // Returns 1000 random samples from poissonDistribution with
a mean of 5.
sample(markovChain(transitionMatrix), 1000) // Returns 1000 random samples from a Markov Chain.
```

## scalarAdd

The scalarAdd function adds a scalar value to every value in a numeric array or matrix. When working with numeric arrays, scalarAdd returns a new array with the new values. When working with a matrix, scalarAdd returns a new matrix with new values.

## scalarAdd Parameters

number: value to add numeric array | matrix: the numeric array or matrix to add the value to.

## scalarAdd Syntax

```
scalarAdd(number, numericArray) // Adds the number to each element in the number in the array.
scalarAdd(number, matrix) // Adds the number to each value in a matrix
```

## scalarAdd Returns

numericArray | matrix: Depending on what is being operated on.

## scalarDivide

The scalarDivide function divides each number in numeric array or matrix by a scalar value. When working with numeric arrays, scalarDivide returns a new array with the new values. When working with a matrix, scalarDivide returns a new matrix with new values.

## scalarDivide Parameters

number: value to divide by numeric array | matrix: the numeric array or matrix to divide by the value to.

## scalarDivide Syntax

```
scalarDivide(number, numericArray) // Divides each element in the numeric array by the number.
scalarDivide(number, matrix) // Divides each element in the matrix by the number.
```

## scalarDivide Returns

numericArray | matrix: depending on what is being operated on.

## scalarMultiply

The scalarMultiply function multiplies each element in a numeric array or matrix by a scalar value. When working with numeric arrays, scalarMultiply returns a new array with the new values. When working with a matrix, scalarMultiply returns a new matrix with new values.

## scalarMultiply Parameters

number: value to divide by numeric array | matrix: the numeric array or matrix to divide by the value to.

## scalarMultiply Syntax

```
scalarMultiply(number, numericArray) // Multiplies each element in the numeric array by the
number.
scalarMultiply(number, matrix) // Multiplies each element in the matrix by the number.
```

## scalarMultiply Returns

numericArray | matrix: depending on what is being operated on

## scalarSubtract

The scalarSubtract function subtracts a scalar value from every value in a numeric array or matrix. When working with numeric arrays, scalarSubtract returns a new array with the new values. When working with a matrix, scalarSubtract returns a new matrix with new values.

## scalarSubtract Parameters

number: value to add numeric array | matrix: the numeric array or matrix to subtract the value from.

## scalarSubtract Syntax

```
scalarSubtract(number, numericArray) // Subtracts the number from each element in the number in
the array.
scalarSubtract(number, matrix) // Subtracts the number from each value in a matrix
```

## scalarSubtract Returns

numericArray | matrix: depending on what is being operated on.

## scale

The scale function multiplies all the elements of an array by a number.

### scale Parameters

- number
- numeric array

### scale Syntax

```
scale(number, numericArray)
```

## sequence

The sequence function returns an array of numbers based on its parameters.

### sequence Parameters

- length
- start
- stride

### sequence Syntax

```
sequence(100, 0, 1) // Returns a sequence of length 100, starting from 0 with a stride of 1.
```

## setAttributes

The setAttributes function sets an attributes map of a matrix.

### setAttributes Parameters

- matrix: The matrix to set the attributes map to.
- map: The map of attributes to set on the matrix.

### setAttributes Syntax

```
setAttributes(matrix, map)
```

## setAttributes Returns

matrix: The matrix with the attributes set.

## setColumnLabels

The setColumnLabels function sets the columns labels of a matrix.

### setColumnLabels Parameters

- matrix: The matrix to set the column labels to.
- string array: The column labels to set the matrix

### setColumnLabels Syntax

setColumnLabels(matrix, labels)

## setColumnLabels Returns

matrix: The matrix with the labels set.

## setRowLabels

The setRowLabels function sets the row labels of a matrix.

### setRowLabels Parameters

- matrix: The matrix to set the row labels to.
- string array: The row labels to set to the matrix

### setRowLabels Syntax

setRowLabels(matrix, labels)

## setRowLabels Returns

matrix: The matrix with the labels set.

## setValue

The setValue function sets a new value for a Tuple entry.

### setValue Parameters

- tuple: The Tuple to return the entry from.
- key: The key of the entry to set.
- value: The value to set.

## setValue Syntax

setValue(tuple, key, value)

## setValue Returns

tuple: Returns the new modified tuple

## sin

The sin function returns the trigonometric sine of a number.

## sin Parameters

- Field Name | Raw Number | Number Evaluator: The value to return the sine of.

## sin Syntax

```
sin(100.4) // returns the sine of 100.4
sine(fieldA) // returns the sine for fieldA.
if(gt(fieldA,fieldB),sin(fieldA),sin(fieldB)) // if fieldA > fieldB then return the sine of
fieldA, else return the sine of fieldB
```

## spline

The spline function performs a cubic spline interpolation ([https://en.wikiversity.org/wiki/Cubic\\_Spline\\_Interpolation](https://en.wikiversity.org/wiki/Cubic_Spline_Interpolation)) of a curve given a set of x,y coordinates. The return value of the spline function is an interpolation function which can be used to **predict** values along the curve and generate a **derivative** of the curve.

## spline Parameters

- numeric array: (Optional) x values. If omitted a sequence will be created for the x values.
- numeric array: y values

## spline Syntax

```
spline(yValues) // This creates the xValues automatically and fits a spline through the data
points.
spline(xValues, yValues) // This will fit a spline through the data points.
```

## spline Returns

function: the function can be treated as both a numeric array and function.

## sqrt

The sqrt function returns the trigonometric square root of a number.

### sqrt Parameters

- Field Name | Raw Number | Number Evaluator: The value to return the square root of.

### sqrt Syntax

```
sqrt(100.4) // returns the square root of 100.4
sqrt(fieldA) // returns the square root for fieldA.
if(gt(fieldA,fieldB),sqrt(fieldA),sqrt(fieldB)) // if fieldA > fieldB then return the sqrt of
fieldA, else return the sqrt of fieldB
```

## standardize

The standardize function standardizes a numeric array so that values within the array have a mean of 0 and standard deviation of 1.

### standardize Parameters

- numeric array: the array to standardize

### standardize Syntax

```
standardize(numericArray)
```

## standardize Returns

numeric array: the standardized values

## sub

The sub function will take 2 or more numeric values and subtract them, from left to right. The sub function will fail to execute if any of the values are non-numeric. If a null value is found then null will be returned as the result.

### sub Parameters

- Field Name | Raw Number | Number Evaluator
- Field Name | Raw Number | Number Evaluator
- .....
- Field Name | Raw Number | Number Evaluator

## sub Syntax

The expressions below show the various ways in which you can use the sub evaluator. The number of these parameters does not matter and is not limited except that at least two parameters are required. Returns a numeric value.

```
sub(1,2,3,4) // 1 - 2 - 3 - 4
sub(1,fieldA) // 1 - value of fieldA
sub(fieldA,1.4) // value of fieldA - 1.4
sub(fieldA,fieldB,fieldC) // value of fieldA - value of fieldB - value of fieldC
sub(fieldA,div(fieldA,fieldB)) // value of fieldA - (value of fieldA / value of fieldB)
if(gt(fieldA,fieldB),sub(fieldA,fieldB),sub(fieldB,fieldA)) // if fieldA > fieldB then fieldA - fieldB, else fieldB - field
```

## sumDifference

The sumDifference function calculates the sum of the differences following an element-by-element subtraction between two numeric arrays.

### sumDifference Parameters

- numeric array
- numeric array

### sumDifference Returns

A numeric.

### sumDifference Syntax

```
sumDifference(numericArray, numericArray)
```

## sumColumns

The sumColumns function sums the columns in a matrix and returns a numeric array with the result.

### sumColumns Parameters

- matrix: the matrix to operate on

### sumColumns Syntax

```
sumColumns(matrix)
```

## sumColumns Returns

numeric array: the sum of the columns

## sumRows

The sumRows function sums the rows in a matrix and returns a numeric array with the result.

### sumRows Parameters

- matrix: the matrix to operate on

### sumRows Syntax

```
sumRows(matrix)
```

## sumRows Returns

numeric array: sum of the rows.

## sumSq

The sumSq function returns the sum-of-squares of the values in a numeric array.

### sumSq Parameters

- numeric array: The numeric array to compute the sumSq of.

### sumSq Syntax

```
sumSq(numericArray)
```

## sumSq Returns

numeric: result of the sumSq calculation

## transpose

The transpose function [transposes](#) a matrix .

### transpose Parameters

- matrix: the matrix to transpose

### transpose Syntax

```
transpose(matrix)
```

## transpose Returns

matrix: the transposed matrix

## triangularDistribution

The triangularDistribution function returns a [triangular probability distribution](#) based on its parameters. This function is part of the probability distribution framework and is designed to work with the [sample](#), [probability](#) and [cumulativeProbability](#) functions.

### triangularDistribution Parameters

- double: low value
- double: most likely value
- double: high value

### triangularDistribution Syntax

```
triangularDistribution(10, 15, 20) // A triangular distribution with a low value of 10, most likely value of 15 and high value of 20.
```

## triangularDistribution Returns

Probability distribution function

## uniformDistribution

The uniformDistribution function returns a [continuous uniform probability distribution](#) based on its parameters. See the [uniformIntegerDistribution](#) to work with discrete uniform distributions. This function is part of the probability distribution framework and is designed to work with the [sample](#) and [cumulativeProbability](#) functions.

### uniformDistribution Parameters

- double: start
- double: end

### uniformDistribution Returns

Probability distribution function.

### uniformDistribution Syntax

```
uniformDistribution(0.0, 100.0)
```

## uniformIntegerDistribution

The `uniformIntegerDistribution` function returns a [discrete uniform probability distribution](#) based on its parameters. See the [uniformDistribution](#) to work with continuous uniform distributions. This function is part of the probability distribution framework and is designed to work with the [sample](#), [probability](#) and [cumulativeProbability](#) functions.

### uniformIntegerDistribution Parameters

- `integer`: start
- `integer`: end

### uniformIntegerDistribution Returns

A probability distribution function.

### uniformIntegerDistribution Syntax

```
uniformDistribution(1, 6)
```

## unitize

The `unitize` function scales numeric arrays to a magnitude of 1, often called [unit vectors](#). The `unitize` function can operate on both numeric arrays and matrices.

When operating on a matrix the `unitize` function unitizes each row of the matrix.

### unitize Parameters

- `numeric array | matrix`: The array or matrix to unitize

### unitize Syntax

```
unitize(numericArray) // Unitize a numeric array
unitize(matrix) // Unitize each row in a matrix
```

### unitize Returns

`numeric array | matrix`

## weibullDistribution

The `weibullDistribution` function returns a [Weibull probability distribution](#) based on its parameters. This

function is part of the probability distribution framework and is designed to work with the [sample](#), [kolmogorovSmirnov](#) and [cumulativeProbability](#) functions.

## weibullDistribution Parameters

- double: shape
- double: scale

## weibullDistribution Returns

A probability distribution function.

## weibullDistribution Syntax

```
weibullDistribution(.5, 10)
```

## zipFDistribution

The zipFDistribution function returns a [ZipF distribution](#) based on its parameters. This function is part of the probability distribution framework and is designed to work with the [sample](#), [probability](#) and [cumulativeProbability](#) functions.

## zipFDistribution Parameters

- integer: size
- double: exponent

## zipFDistribution Returns

A probability distribution function.

## zipFDistribution Syntax

```
zipFDistribution(5000, 1.0)
```

# Math Expressions

The Streaming Expression library includes a powerful mathematical programming syntax with many of the features of a functional programming language. The syntax includes variables, data structures and a growing set of mathematical functions.

This user guide provides an overview of the different areas of mathematical coverage starting with basic scalar math and ending with machine learning. Along the way the guide covers variables and data structures and techniques for combining Solr's powerful streams with mathematical functions to make every record in your Solr Cloud cluster computable.

**Scalar Math:** The functions that apply to scalar numbers.

**Vector Math:** Vector math expressions and vector manipulation.

**Variables and Caching:** Assigning and caching variables.

**Matrix Math:** Matrix creation, manipulation, and matrix math.

**Streams and Vectorization:** Retrieving streams and vectorizing numeric and lat/lon location fields.

**Text Analysis and Term Vectors:** Using math expressions for text analysis and TF-IDF term vectors.

**Statistics:** Statistical functions in math expressions.

**Probability:** Mathematical models of probability.

**Monte Carlo Simulations:** Performing uncorrelated and correlated Monte Carlo simulations.

**Linear Regression:** Simple and multivariate linear regression.

**Interpolation, Derivatives and Integrals:** Numerical analysis math expressions.

**Digital Signal Processing:** Functions commonly used with digital signal processing.

**Curve Fitting:** Polynomial, Harmonic and Gaussian curve fitting.

**Time Series:** Aggregation, smoothing and differencing of time series.

**Machine Learning:** Functions used in machine learning.

**Computational Geometry:** Convex Hulls and Enclosing Disks.

## Scalar Math

The most basic math expressions are scalar expressions. Scalar expressions perform mathematical operations on numbers.

For example the expression below adds two numbers together:

```
add(1, 1)
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": 2
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 2
 }
]
 }
}
```

Math expressions can be nested. For example in the expression below the output of the add function is the second parameter of the pow function:

```
pow(10, add(1,1))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": 100
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Streaming Scalar Math

Scalar math expressions can also be applied to each tuple in a stream through use of the select stream decorator. The select function wraps a stream of tuples and selects fields to include in each tuple. The select function can also use math expressions to compute new values and add them to the outgoing tuples.

In the example below the select expression is wrapping a search expression. The select function is selecting the **price\_f** field and computing a new field called **newPrice** using the **mult** math expression.

The first parameter of the mult expression is the **price\_f** field. The second parameter is the scalar value 10.

This multiplies the value of the `price_f` field in each tuple by 10.

```
select(search(collection2, q="*:*", fl="price_f", sort="price_f desc", rows="3"),
 price_f,
 mult(price_f, 10) as newPrice)
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "price_f": 0.99999994,
 "newPrice": 9.9999994
 },
 {
 "price_f": 0.99999994,
 "newPrice": 9.9999994
 },
 {
 "price_f": 0.99999992,
 "newPrice": 9.9999992
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 3
 }
]
 }
}
```

## More Scalar Math Functions

The following scalar math functions are available in the math expressions library:

abs, add, div, mult, sub, log, log10, pow, mod, ceil, floor, sin, asin, sinh, cos, acos, cosh, tan, atan, tanh, round, precision, recip, sqrt, cbrt

## Vector Math

This section covers vector math and vector manipulation functions.

## Arrays

Arrays can be created with the array function.

For example, the expression below creates a numeric array with three elements:

```
array(1, 2, 3)
```

When this expression is sent to the /stream handler it responds with a JSON array:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": [
 1,
 2,
 3
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Array Operations

Arrays can be passed as parameters to functions that operate on arrays.

For example, an array can be reversed with the rev function:

```
rev(array(1, 2, 3))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": [
 3,
 2,
 1
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

Another example is the `length` function, which returns the length of an array:

```
length(array(1, 2, 3))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": 3
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

A slice of an array can be taken with the `copyOfRange` function, which copies elements of an array from a start and end range.

```
copyOfRange(array(1,2,3,4,5,6), 1, 4)
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": [
 2,
 3,
 4
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

Elements of an array can be trimmed using the `ltrim` (left trim) and `rtrim` (right trim) functions. The `ltrim` and `rtrim` functions remove a specific number of elements from the left or right of an array.

The example below shows the `ltrim` function trimming the first 2 elements of an array:

```
ltrim(array(0,1,2,3,4,5,6), 2)
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": [
 2,
 3,
 4,
 5,
 6,
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 1
 }
]
 }
}
```

## Vector Sorting

An array can be sorted in natural ascending order with the asc function.

The example below shows the asc function sorting an array:

```
asc(array(10,1,2,3,4,5,6))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": [
 1,
 2,
 3,
 4,
 5,
 6,
 10
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 1
 }
]
 }
}
```

## Vector Summarizations and Norms

There are a set of functions that perform summarizations and return norms of arrays. These functions operate over an array and return a single value. The following vector summarizations and norm functions are available: mult, add, sumSq, mean, l1norm, l2norm, linfnorm.

The example below shows the mult function, which multiples all the values of an array.

```
mult(array(2,4,8))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": 64
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

The vector norm functions provide different formulas for calculating vector magnitude.

The example below calculates the  $\text{l2norm}$  of an array.

```
l2norm(array(2,4,8))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": 9.16515138991168
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Scalar Vector Math

Scalar vector math functions add, subtract, multiply or divide a scalar value with every value in a vector. The following functions perform these operations: `scalarAdd`, `scalarSubtract`, `scalarMultiply` and `scalarDivide`.

Below is an example of the `scalarMultiply` function, which multiplies the scalar value 3 with every value of an array.

```
scalarMultiply(3, array(1,2,3))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": [
 3,
 6,
 9
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Element-By-Element Vector Math

Two vectors can be added, subtracted, multiplied and divided using element-by-element vector math functions. The available element-by-element vector math functions are: ebeAdd, ebeSubtract, ebeMultiply, ebeDivide.

The expression below performs the element-by-element subtraction of two arrays.

```
ebeSubtract(array(10, 15, 20), array(1,2,3))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": [
 9,
 13,
 17
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 5
 }
]
 }
}
```

## Dot Product and Cosine Similarity

The dotProduct and cosineSimilarity functions are often used as similarity measures between two sparse vectors. The dotProduct is a measure of both angle and magnitude while cosineSimilarity is a measure only of angle.

Below is an example of the dotProduct function:

```
dotProduct(array(2,3,0,0,0,1), array(2,0,1,0,0,3))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": 7
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 15
 }
]
 }
}
```

Below is an example of the cosineSimilarity function:

```
cosineSimilarity(array(2,3,0,0,0,1), array(2,0,1,0,0,3))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": 0.5
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 7
 }
]
 }
}
```

## Variables

### The Let Expression

The let expression sets variables and returns the value of the last variable by default. The output of any streaming expression or math expression can be set to a variable.

Below is a simple example setting three variables **a**, **b** and **c**. Variables **a** and **b** are set to arrays. The variable **c** is set to the output of the ebeAdd function which performs element-by-element addition of the two arrays.

```
let(a=array(1, 2, 3),
 b=array(10, 20, 30),
 c=ebeAdd(a, b))
```

In the response, notice that the last variable, **c**, is returned:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 11,
 22,
 33
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 4
 }
]
 }
}
```

## Echoing Variables

All variables can be output by setting the echo variable to true.

```
let(echo=true,
 a=array(1, 2, 3),
 b=array(10, 20, 30),
 c=ebeAdd(a, b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "a": [
 1,
 2,
 3
],
 "b": [
 10,
 20,
 30
],
 "c": [
 11,
 22,
 33
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

A specific set of variables can be echoed by providing a comma delimited list of variables to the echo parameter. Because variables have been provided, the true value is assumed.

```
let(echo="a,b",
 a=array(1, 2, 3),
 b=array(10, 20, 30),
 c=ebeAdd(a, b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "a": [
 1,
 2,
 3
],
 "b": [
 10,
 20,
 30
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Caching Variables

Variables can be cached in-memory on the Solr node where the math expression was run. A cached variable can then be used in future expressions. Any object that can be set to a variable, including data structures and mathematical models, can be cached in-memory for future use.

The `putCache` function adds a variable to the cache.

In the example below an array is cached in the workspace "workspace1" and bound to the key "key1". The workspace allows different users to cache objects in their own workspace. The `putCache` function returns the variable that was added to the cache.

```
let(a=array(1, 2, 3),
 b=array(10, 20, 30),
 c=ebeAdd(a, b),
 d=putCache(workspace1, key1, c))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "d": [
 11,
 22,
 33
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 11
 }
]
 }
}
```

The `getCache` function retrieves an object from the cache by its workspace and key.

In the example below the `getCache` function retrieves the array that was cached above and assigns it to variable `a`.

```
let(a=getCache(workspace1, key1))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "a": [
 11,
 22,
 33
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 11
 }
]
 }
}
```

The `listCache` function can be used to list the workspaces or the keys in a specific workspace.

In the example below `listCache` returns all the workspaces in the cache as an array of strings.

```
let(a=listCache())
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "a": [
 "workspace1"
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

In the example below all the keys in a specific workspace are listed:

```
let(a=listCache(workspace1))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "a": [
 "key1"
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

The removeCache function can be used to remove a key from a specific workspace. The removeCache function removes the key from the cache and returns the object that was removed.

In the example below the array that was cached above is removed from the cache.

```
let(a=removeCache(workspace1, key1))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "a": [
 11,
 22,
 33
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Matrices and Matrix Math

This section of the user guide covers the basics of matrix creation, manipulation and matrix math. Other sections of the user guide demonstrate how matrices are used by the statistics, probability and machine learning functions.

### Matrix Creation

A matrix can be created with the `matrix` function. The `matrix` function is passed a list of arrays with each array representing a **row** in the matrix.

The example below creates a two-by-two matrix.

```
matrix(array(1, 2),
 array(4, 5))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": [
 [
 1,
 2
],
 [
 4,
 5
]
],
 {"EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
 }
}
```

## Accessing Rows and Columns

The rows and columns of a matrix can be accessed using the `rowAt` and `colAt` functions.

The example below creates a 2 by 2 matrix and returns the second column of the matrix. Notice that the matrix is passed variables in this example rather than directly passed a list of arrays.

```
let(a=array(1, 2),
 b=array(4, 5),
 c=matrix(a, b),
 d=colAt(c, 1))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "d": [
 2,
 5
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Pair Sorting Vectors

The pairSort function can be used to sort two vectors based on the values in the first vector. The sorting operation maintains the pairing between the two vectors during the sort.

The pairSort function returns a matrix containing the pair sorted vectors. The first row in the matrix is the first vector, the second row in the matrix is the second vector.

The individual vectors can then be accessed using the rowAt function.

The example below performs a pair sort of two vectors and returns the matrix containing the sorted vectors.

```
let(a=array(10, 2, 1),
 b=array(100, 200, 300),
 c=pairSort(a, b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 [
 1,
 2,
 10
],
 [
 300,
 200,
 100
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 1
 }
]
 }
}
```

## Row and Column Labels

A matrix can have column and rows and labels. The functions `setRowLabels`, `setColumnLabels`, `getRowLabels` and `getColumnLabels` can be used to set and get the labels. The label values are set using string arrays.

The example below sets the row and column labels. In other sections of the user guide examples are shown where functions return matrices with the labels already set.

Below is a simple example of setting and getting row and column labels on a matrix.

```
let(echo="d, e",
 a=matrix(array(1, 2),
 array(4, 5)),
 b=setRowLabels(a, array("row0", "row1")),
 c=setColumnLabels(b, array("col0", "col1")),
 d=getRowLabels(c),
 e=getColumnLabels(c))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "d": [
 "row0",
 "row1"
],
 "e": [
 "col0",
 "col1"
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Matrix Attributes

A matrix can also have an arbitrary set of named attributes associated with it. Certain functions, such as the termVectors function, return matrices that contain attributes that describe data in the matrix.

Attributes can be retrieved by name using the getAttribute function and the entire attribute map can be returned using the getAttributes function.

## Matrix Dimensions

The dimensions of a matrix can be determined using the rowCount and columnCount functions.

The example below retrieves the dimensions of a matrix.

```
let(echo="b,c",
 a=matrix(array(1, 2, 3),
 array(4, 5, 6)),
 b=rowCount(a),
 c=columnCount(a))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": 2,
 "c": 3
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Matrix Transposition

A matrix can be [transposed](#) using the transpose function.

An example of matrix transposition is shown below:

```
let(a=matrix(array(1, 2),
 array(4, 5)),
 b=transpose(a))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 [
 1,
 4
],
 [
 2,
 5
]
],
 {"EOF": true,
 "RESPONSE_TIME": 24
 }
]
 }
 }
}
```

## Matrix Summations

The rows and columns of a matrix can be summed with the sumRows and sumColumns functions. Below is an example of the sumRows function which returns an array with the sum of each row.

```
let(a=matrix(array(1, 2, 3),
 array(4, 5, 6)),
 b=sumRows(a))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 6,
 15
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 2
 }
]
 }
}
```

The grandSum function returns the sum of all values in the matrix. Below is an example of the grandSum function:

```
let(a=matrix(array(1, 2, 3),
 array(4, 5, 6)),
 b=grandSum(a))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": 21
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Scalar Matrix Math

The same scalar math functions that apply to vectors can also be applied to matrices: scalarAdd, scalarSubtract, scalarMultiply, scalarDivide. Below is an example of the scalarAdd function which adds a scalar value to each element in a matrix.

```
let(a=matrix(array(1, 2),
 array(4, 5)),
 b=scalarAdd(10, a))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 [
 11,
 12
],
 [
 14,
 15
]
],
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
 }
}
```

## Matrix Addition and Subtraction

Two matrices can be added and subtracted using the ebeAdd and ebeSubtract functions, which perform element-by-element addition and subtraction of matrices.

Below is a simple example of an element-by-element addition of a matrix by itself:

```
let(a=matrix(array(1, 2),
 array(4, 5)),
 b=ebeAdd(a, a))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 [
 2,
 4
],
 [
 8,
 10
]
],
 {"EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
 }
}
```

## Matrix Multiplication

Matrix multiplication can be accomplished using the `matrixMult` function. Below is a simple example of matrix multiplication:

```
let(a=matrix(array(1, 2),
 array(4, 5)),
 b=matrix(array(11, 12),
 array(14, 15)),
 c=matrixMult(a, b))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 [
 39,
 42
],
 [
 114,
 123
]
],
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
 }
}
```

## Streams and Vectorization

This section of the user guide explores techniques for retrieving streams of data from Solr and vectorizing the numeric fields.

See the section [Text Analysis and Term Vectors](#) which describes how to vectorize text fields.

### Streams

Streaming Expressions has a wide range of stream sources that can be used to retrieve data from Solr Cloud collections. Math expressions can be used to vectorize and analyze the results sets.

Below are some of the key stream sources:

- **facet**: Multi-dimensional aggregations are a powerful tool for generating co-occurrence counts for categorical data. The facet function uses the JSON facet API under the covers to provide fast, distributed, multi-dimension aggregations. With math expressions the aggregated results can be pivoted into a co-occurrence matrix which can be mined for correlations and hidden similarities within the data.
- **random**: Random sampling is widely used in statistics, probability and machine learning. The random function returns a random sample of search results that match a query. The random samples can be vectorized and operated on by math expressions and the results can be used to describe and make inferences about the entire population.
- **timeseries**: The timeseries expression provides fast distributed time series aggregations, which can be vectorized and analyzed with math expressions.
- **knnSearch**: K-nearest neighbor is a core machine learning algorithm. The knnSearch function is a

specialized knn algorithm optimized to find the k-nearest neighbors of a document in a distributed index. Once the nearest neighbors are retrieved they can be vectorized and operated on by machine learning and text mining algorithms.

- **sql:** SQL is the primary query language used by data scientists. The sql function supports data retrieval using a subset of SQL which includes both full text search and fast distributed aggregations. The result sets can then be vectorized and operated on by math expressions.
- **jdbc:** The jdbc function allows data from any JDBC compliant data source to be combined with streams originating from Solr. Result sets from outside data sources can be vectorized and operated on by math expressions in the same manner as result sets originating from Solr.
- **topic:** Messaging is an important foundational technology for large scale computing. The topic function provides publish/subscribe messaging capabilities by treating Solr Cloud as a distributed message queue. Topics are extremely powerful because they allow subscription by query. Topics can be used to support a broad set of use cases including bulk text mining operations and AI alerting.
- **nodes:** Graph queries are frequently used by recommendation engines and are an important machine learning tool. The nodes function provides fast, distributed, breadth first graph traversal over documents in a Solr Cloud collection. The node sets collected by the nodes function can be operated on by statistical and machine learning expressions to gain more insight into the graph.
- **search:** Ranked search results are a powerful tool for finding the most relevant documents from a large document corpus. The search expression returns the top N ranked search results that match any Solr query, including geo-spatial queries. The smaller set of relevant documents can then be explored with statistical, machine learning and text mining expressions to gather insights about the data set.

## Assigning Streams to Variables

The output of any streaming expression can be set to a variable. Below is a very simple example using the random function to fetch three random samples from collection1. The random samples are returned as tuples which contain name/value pairs.

```
let(a=random(collection1, q="*:*", rows="3", f1="price_f"))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "a": [
 {
 "price_f": 0.7927976
 },
 {
 "price_f": 0.060795486
 },
 {
 "price_f": 0.55128294
 }
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 11
 }
]
 }
}
```

## Creating a Vector with the col Function

The col function iterates over a list of tuples and copies the values from a specific column into an array.

The output of the col function is an numeric array that can be set to a variable and operated on by math expressions.

Below is an example of the col function:

```
let(a=random(collection1, q="*:*", rows="3", fl="price_f"),
 b=col(a, price_f))
```

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 0.42105234,
 0.85237443,
 0.7566981
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 9
 }
]
 }
}
```

## Applying Math Expressions to the Vector

Once a vector has been created any math expression that operates on vectors can be applied. In the example below the mean function is applied to the vector assigned to variable **b**.

```
let(a=random(collection1, q="*:*", rows="15000", fl="price_f"),
 b=col(a, price_f),
 c=mean(b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": 0.5016035594638814
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 306
 }
]
 }
}
```

## Creating Matrices

Matrices can be created by vectorizing multiple numeric fields and adding them to a matrix. The matrices can then be operated on by any math expression that operates on matrices.



Note that this section deals with the creation of matrices from numeric data. The section [Text Analysis and Term Vectors](#) describes how to build TF-IDF term vector matrices from text fields.

Below is a simple example where four random samples are taken from different sub-populations in the data. The `price_f` field of each random sample is vectorized and the vectors are added as rows to a matrix. Then the `sumRows` function is applied to the matrix to return a vector containing the sum of each row.

```
let(a=random(collection1, q="market:A", rows="5000", fl="price_f"),
 b=random(collection1, q="market:B", rows="5000", fl="price_f"),
 c=random(collection1, q="market:C", rows="5000", fl="price_f"),
 d=random(collection1, q="market:D", rows="5000", fl="price_f"),
 e=col(a, price_f),
 f=col(b, price_f),
 g=col(c, price_f),
 h=col(d, price_f),
 i=matrix(e, f, g, h),
 j=sumRows(i))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "j": [
 154390.1293375,
 167434.89453,
 159293.258493,
 149773.42769,
],
 {
 "EOF": true,
 "RESPONSE_TIME": 9
 }
]
 }
 }
}
```

## Facet Co-occurrence Matrices

The `facet` function can be used to quickly perform multi-dimension aggregations of categorical data from records stored in a Solr Cloud collection. These multi-dimension aggregations can represent co-occurrence counts for the values in the dimensions. The `pivot` function can be used to move two dimensional aggregations into a co-occurrence matrix. The co-occurrence matrix can then be clustered or analyzed for correlations to learn about the hidden connections within the data.

In the example below the `facet` expression is used to generate a two dimensional faceted aggregation. The

first dimension is the US State that a car was purchased in and the second dimension is the car model. This two dimensional facet generates the co-occurrence counts for the number of times a particular car model was purchased in a particular state.

```
facet(collection1, q="*:*", buckets="state, model", bucketSorts="count(*) desc", rows=5,
count(*))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "state": "NY",
 "model": "camry",
 "count(*)": 13342
 },
 {
 "state": "NJ",
 "model": "accord",
 "count(*)": 13002
 },
 {
 "state": "NY",
 "model": "civic",
 "count(*)": 12901
 },
 {
 "state": "CA",
 "model": "focus",
 "count(*)": 12892
 },
 {
 "state": "TX",
 "model": "f150",
 "count(*)": 12871
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 171
 }
]
 }
}
```

The pivot function can be used to move the facet results into a co-occurrence matrix. In the example below The pivot function is used to create a matrix where the rows of the matrix are the US States (state) and the columns of the matrix are the car models (model). The values in the matrix are the co-occurrence counts (count(\*)) from the facet results. Once the co-occurrence matrix has been created the US States can be

clustered by car model, or the matrix can be transposed and car models can be clustered by the US States where they were bought.

```
let(a=facet(collection1, q="*:*", buckets="state, model", bucketSorts="count(*) desc", rows="-1",
count(*)),
b=pivot(a, state, model, count(*)),
c=kmeans(b, 7))
```

## Latitude / Longitude Vectors

The latlonVectors function wraps a list of tuples and parses a lat/lion location field into a matrix of lat/long vectors. Each row in the matrix is a vector that contains the lat/long pair for the corresponding tuple in the list. The row labels for the matrix are automatically set to the id field in the tuples. The lat/lion matrix can then be operated on by distance-based machine learning functions using the haversineMeters distance measure.

The latlonVectors function takes two parameters: a list of tuples and a named parameter called field, which tells the latlonVectors function which field to parse the lat/lion vectors from.

Below is an example of the latlonVectors.

```
let(a=random(collection1, q="*:*", fl="id, loc_p", rows="5"),
b=latlonVectors(a, field="loc_p"))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 [
 42.87183530723629,
 76.74102353397778
],
 [
 42.91372904094898,
 76.72874889228416
],
 [
 42.911528804897564,
 76.70537292977619
],
 [
 42.91143870500213,
 76.74749913047408
],
 [
 42.904666267479705,
 76.73933236046092
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 21
 }
]
 }
}
```

## Text Analysis and Term Vectors

Term frequency-inverse document frequency (TF-IDF) term vectors are often used to represent text documents when performing text mining and machine learning operations. The math expressions library can be used to perform text analysis and create TF-IDF term vectors.

### Text Analysis

The `analyze` function applies a Solr analyzer to a text field and returns the tokens emitted by the analyzer in an array. Any analyzer chain that is attached to a field in Solr's schema can be used with the `analyze` function.

In the example below, the text "hello world" is analyzed using the analyzer chain attached to the `subject` field in the schema. The `subject` field is defined as the field type `text_general` and the text is analyzed using the analysis chain configured for the `text_general` field type.

```
analyze("hello world", subject)
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "return-value": [
 "hello",
 "world"
],
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
 }
}
```

## Annotating Documents

The analyze function can be used inside of a select function to annotate documents with the tokens generated by the analysis.

The example below performs a search in "collection1". Each tuple returned by the search function contains an id and subject. For each tuple, the select function selects the id field and calls the analyze function on the subject field. The analyzer chain specified by the subject\_bigram field is configured to perform a bigram analysis. The tokens generated by the analyze function are added to each tuple in a field called terms.

```
select(search(collection1, q="*:*", fl="id, subject", sort="id asc"),
 id,
 analyze(subject, subject_bigram) as terms)
```

Notice in the output that an array of bigram terms have been added to the tuples:

```
{
 "result-set": {
 "docs": [
 {
 "terms": [
 "text analysis",
 "analysis example"
],
 "id": "1"
 },
 {
 "terms": [
 "example number",
 "number two"
],
 "id": "2"
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 4
 }
]
 }
}
```

## TF-IDF Term Vectors

The termVectors function can be used to build TF-IDF term vectors from the terms generated by the analyze function.

The termVectors function operates over a list of tuples that contain a field called id and a field called terms. Notice that this is the exact output structure of the document annotation example above.

The termVectors function builds a matrix from the list of tuples. There is a row in the matrix for each tuple in the list. There is a column in the matrix for each term in the terms field.

```
let(echo="c, d", ①
 a=select(search(collection3, q="*:*", fl="id, subject", sort="id asc"), ②
 id,
 analyze(subject, subject_bigram) as terms),
 b=termVectors(a, minTermLength=4, minDocFreq=0, maxDocFreq=1), ③
 c=getRowLabels(b), ④
 d=getColumnLabels(b))
```

The example below builds on the document annotation example.

- ① The echo parameter will echo variables **c** and **d**, so the output includes the row and column labels, which will be defined later in the expression.
- ② The list of tuples are stored in variable **a**. The termVectors function operates over variable **a** and builds a

matrix with 2 rows and 4 columns.

- ③ The termVectors function sets the row and column labels of the term vectors matrix as variable **b**. The row labels are the document ids and the column labels are the terms.
- ④ The getRowLabels and getColumnLabels functions return the row and column labels which are then stored in variables **c** and **d**.

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 "1",
 "2"
],
 "d": [
 "analysis example",
 "example number",
 "number two",
 "text analysis"
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 5
 }
]
 }
}
```

## TF-IDF Values

The values within the term vectors matrix are the TF-IDF values for each term in each document. The example below shows the values of the matrix.

```
let(a=select(search(collection3, q="*:*", fl="id, subject", sort="id asc"),
 id,
 analyze(subject, subject_bigram) as terms),
 b=termVectors(a, minTermLength=4, minDocFreq=0, maxDocFreq=1))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 [
 1.4054651081081644,
 0,
 0,
 1.4054651081081644
],
 [
 0,
 1.4054651081081644,
 1.4054651081081644,
 0
]
],
 {
 "EOF": true,
 "RESPONSE_TIME": 5
 }
]
]
 }
}
```

## Limiting the Noise

One of the key challenges when working with term vectors is that text often has a significant amount of noise which can obscure the important terms in the data. The `termVectors` function has several parameters designed to filter out the less meaningful terms. This is also important because eliminating the noisy terms helps keep the term vector matrix small enough to fit comfortably in memory.

There are four parameters designed to filter noisy terms from the term vector matrix:

### `minTermLength`

The minimum term length required to include the term in the matrix.

### `minDocFreq`

The minimum percentage, expressed as a number between 0 and 1, of documents the term must appear in to be included in the index.

### `maxDocFreq`

The maximum percentage, expressed as a number between 0 and 1, of documents the term can appear in to be included in the index.

### `exclude`

A comma delimited list of strings used to exclude terms. If a term contains any of the exclude strings that term will be excluded from the term vector.

## Statistics

This section of the user guide covers the core statistical functions available in math expressions.

### Descriptive Statistics

The describe function can be used to return descriptive statistics about a numeric array. The describe function returns a single **tuple** with name/value pairs containing descriptive statistics.

Below is a simple example that selects a random sample of documents, vectorizes the **price\_f** field in the result set and uses the describe function to return descriptive statistics about the vector:

```
let(a=random(collection1, q="*:*", rows="1500", fl="price_f"),
 b=col(a, price_f),
 c=describe(b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": {
 "sumsq": 4999.041975263254,
 "max": 0.99995726,
 "var": 0.08344429493940454,
 "geometricMean": 0.36696588922559575,
 "sum": 7497.460565552007,
 "kurtosis": -1.2000739963006035,
 "N": 15000,
 "min": 0.00012338161,
 "mean": 0.49983070437013266,
 "popVar": 0.08343873198640858,
 "skewness": -0.001735537500095477,
 "stdev": 0.28886726179926403
 }
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 305
 }
]
 }
}
```

### Histograms and Frequency Tables

Histograms and frequency tables are tools for understanding the distribution of a random variable.

The `hist` function creates a histogram designed for usage with continuous data. The `freqTable` function creates a frequency table for use with discrete data.

## histograms

Below is an example that selects a random sample, creates a vector from the result set and uses the `hist` function to return a histogram with 5 bins. The `hist` function returns a list of tuples with summary statistics for each bin.

```
let(a=random(collection1, q="*:*", rows="15000", fl="price_f"),
 b=col(a, price_f),
 c=hist(b, 5))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 {
 "prob": 0.2057939717603699,
 "min": 0.000010371208,
 "max": 0.19996578,
 "mean": 0.10010319358402578,
 "var": 0.003366805016271609,
 "cumProb": 0.10293732468049072,
 "sum": 309.0185585938884,
 "stdev": 0.058024176136086666,
 "N": 3087
 },
 {
 "prob": 0.19381868629885585,
 "min": 0.20007741,
 "max": 0.3999073,
 "mean": 0.2993590803885827,
 "var": 0.003401644034068929,
 "cumProb": 0.3025295802728267,
 "sum": 870.536205770005,
 "stdev": 0.0583236147205309,
 "N": 2908
 },
 {
 "prob": 0.20565789836690007,
 "min": 0.39995712,
 "max": 0.5999038,
 "mean": 0.4993620963792545,
 "var": 0.0033158364923609046,
 "cumProb": 0.5023006239697967,
 "sum": 1540.5320673300018,
 }
]
 }
]
 }
}
```

```

 "stdev": 0.05758330046429177,
 "N": 3085
 },
 {
 "prob": 0.19437108496008693,
 "min": 0.6000449,
 "max": 0.79973197,
 "mean": 0.7001752711861512,
 "var": 0.0033895105082360185,
 "cumProb": 0.7026537198687285,
 "sum": 2042.4112660500066,
 "stdev": 0.058219502816805456,
 "N": 2917
 },
 {
 "prob": 0.20019582213899467,
 "min": 0.7999126,
 "max": 0.99987316,
 "mean": 0.8985428275824184,
 "var": 0.003312360017780078,
 "cumProb": 0.899450457219298,
 "sum": 2698.3241112299997,
 "stdev": 0.05755310606544253,
 "N": 3003
 }
]
},
{
 "EOF": true,
 "RESPONSE_TIME": 322
}
]
}
}

```

The `col` function can be used to **vectorize** a column of data from the list of tuples returned by the `hist` function.

In the example below, the `N` field, which is the number of observations in the each bin, is returned as a vector.

```

let(a=random(collection1, q="*:*", rows="15000", fl="price_f"),
 b=col(a, price_f),
 c=hist(b, 11),
 d=col(c, N))

```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "d": [
 1387,
 1396,
 1391,
 1357,
 1384,
 1360,
 1367,
 1375,
 1307,
 1310,
 1366
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 307
 }
]
 }
}
```

## Frequency Tables

The freqTable function returns a frequency distribution for a discrete data set. The freqTable function doesn't create bins like the histogram. Instead it counts the occurrence of each discrete data value and returns a list of tuples with the frequency statistics for each value. Fields from a frequency table can be vectorized using the col function in the same manner as a histogram.

Below is a simple example of a frequency table built from a random sample of a discrete variable.

```
let(a=random(collection1, q="*:*", rows="15000", fl="day_i"),
 b=col(a, day_i),
 c=freqTable(b))
```

When this expression is sent to the /stream handler it responds with:

```
"result-set": {
 "docs": [
 {
 "c": [
 {
 "pct": 0.0318,
 "count": 477,
 "cumFreq": 477,
 }
]
 }
]
}
```

```
 "cumPct": 0.0318,
 "value": 0
 },
 {
 "pct": 0.0331333333333334,
 "count": 497,
 "cumFreq": 974,
 "cumPct": 0.0649333333333333,
 "value": 1
 },
 {
 "pct": 0.0342666666666667,
 "count": 514,
 "cumFreq": 1488,
 "cumPct": 0.0992,
 "value": 2
 },
 {
 "pct": 0.0346,
 "count": 519,
 "cumFreq": 2007,
 "cumPct": 0.1338,
 "value": 3
 },
 {
 "pct": 0.0313333333333333,
 "count": 470,
 "cumFreq": 2477,
 "cumPct": 0.1651333333333333,
 "value": 4
 },
 {
 "pct": 0.0333333333333333,
 "count": 500,
 "cumFreq": 2977,
 "cumPct": 0.1984666666666666,
 "value": 5
 }
],
},
{
 "EOF": true,
 "RESPONSE_TIME": 281
}
]
```

## Percentiles

The percentile function returns the estimated value for a specific percentile in a sample set. The example

below returns the estimation for the 95th percentile of the **price\_f** field.

```
let(a=random(collection1, q="*:*", rows="15000", fl="price_f"),
 b=col(a, price_f),
 c=percentile(b, 95))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": 312.94
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 286
 }
]
 }
}
```

The percentile function also operates on an array of percentile values. The example below is computing the 20th, 40th, 60th and 80th percentiles for a random sample of the **response\_d** field:

```
let(a=random(collection2, q="*:*", rows="15000", fl="response_d"),
 b=col(a, response_d),
 c=percentile(b, array(20,40,60,80)))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 818.0835543394625,
 843.5590348165282,
 866.1789509894824,
 892.5033386599067
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 291
 }
]
 }
}
```

## Covariance and Correlation

Covariance and Correlation measure how random variables move together.

### Covariance and Covariance Matrices

The cov function calculates the covariance of two sample sets of data.

In the example below covariance is calculated for two numeric arrays.

The example below uses arrays created by the array function. Its important to note that vectorized data from Solr Cloud collections can be used with any function that operates on arrays.

```
let(a=array(1, 2, 3, 4, 5),
 b=array(100, 200, 300, 400, 500),
 c=cov(a, b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": 0.9484775349999998
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 286
 }
]
 }
}
```

If a matrix is passed to the cov function it will automatically compute a covariance matrix for the columns of the matrix.

Notice in the example three numeric arrays are added as rows in a matrix. The matrix is then transposed to turn the rows into columns, and the covariance matrix is computed for the columns of the matrix.

```
let(a=array(1, 2, 3, 4, 5),
 b=array(100, 200, 300, 400, 500),
 c=array(30, 40, 80, 90, 110),
 d=transpose(matrix(a, b, c)),
 e=cov(d))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "e": [
 [
 2.5,
 250,
 52.5
],
 [
 250,
 25000,
 5250
],
 [
 52.5,
 5250,
 1150
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 2
 }
]
 }
}
```

## Correlation and Correlation Matrices

Correlation is measure of covariance that has been scaled between -1 and 1.

Three correlation types are supported:

- **pearsons** (default)
- **kendalls**
- **spearmans**

The type of correlation is specified by adding the **type** named parameter in the function call. The example below demonstrates the use of the **type** named parameter.

```
let(a=array(1, 2, 3, 4, 5),
 b=array(100, 200, 300, 400, 5000),
 c=corr(a, b, type=spearmans))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": 0.7432941462471664
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

Like the cov function, the corr function automatically builds a correlation matrix if a matrix is passed as a parameter. The correlation matrix is built by correlating the columns of the matrix passed in.

## Statistical Inference Tests

Statistical inference tests test a hypothesis on **random samples** and return p-values which can be used to infer the reliability of the test for the entire population.

The following statistical inference tests are available:

- anova: One-Way-Anova tests if there is a statistically significant difference in the means of two or more random samples.
- ttest: The T-test tests if there is a statistically significant difference in the means of two random samples.
- pairedTtest: The paired t-test tests if there is a statistically significant difference in the means of two random samples with paired data.
- gTestDataSet: The G-test tests if two samples of binned discrete data were drawn from the same population.
- chiSquareDataset: The Chi-Squared test tests if two samples of binned discrete data were drawn from the same population.
- mannWhitney: The Mann-Whitney test is a non-parametric test that tests if two samples of continuous were pulled from the same population. The Mann-Whitney test is often used instead of the T-test when the underlying assumptions of the T-test are not met.
- ks: The Kolmogorov-Smirnov test tests if two samples of continuous data were drawn from the same distribution.

Below is a simple example of a T-test performed on two random samples. The returned p-value of .93 means we can accept the null hypothesis that the two samples do not have statistically significantly differences in the means.

```
let(a=random(collection1, q="*:*", rows="1500", fl="price_f"),
 b=random(collection1, q="*:*", rows="1500", fl="price_f"),
 c=col(a, price_f),
 d=col(b, price_f),
 e=ttest(c, d))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "e": {
 "p-value": 0.9350135639249795,
 "t-statistic": 0.081545541074817
 }
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 48
 }
]
 }
}
```

## Transformations

In statistical analysis it's often useful to transform data sets before performing statistical calculations. The statistical function library includes the following commonly used transformations:

- rank: Returns a numeric array with the rank-transformed value of each element of the original array.
- log: Returns a numeric array with the natural log of each element of the original array.
- log10: Returns a numeric array with the base 10 log of each element of the original array.
- sqrt: Returns a numeric array with the square root of each element of the original array.
- cbrt: Returns a numeric array with the cube root of each element of the original array.
- recip: Returns a numeric array with the reciprocal of each element of the original array.

Below is an example of a ttest performed on log transformed data sets:

```
let(a=random(collection1, q="*:*", rows="1500", fl="price_f"),
 b=random(collection1, q="*:*", rows="1500", fl="price_f"),
 c=log(col(a, price_f)),
 d=log(col(b, price_f)),
 e=ttest(c, d))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "e": {
 "p-value": 0.9655110070265056,
 "t-statistic": -0.04324265449471238
 }
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 58
 }
]
 }
}
```

## Back Transformations

Vectors that have been transformed with the log, log10, sqrt and cbrt functions can be back transformed using the pow function.

The example below shows how to back transform data that has been transformed by the sqrt function.

```
let(echo="b,c",
 a=array(100, 200, 300),
 b=sqrt(a),
 c=pow(b, 2))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 10,
 14.142135623730951,
 17.320508075688775
],
 "c": [
 100,
 200.0000000000003,
 300.0000000000006
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

The example below shows how to back transform data that has been transformed by the log10 function.

```
let(echo="b,c",
 a=array(100, 200, 300),
 b=log10(a),
 c=pow(10, b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 2,
 2.3010299956639813,
 2.4771212547196626
],
 "c": [
 100,
 200.0000000000003,
 300.0000000000001
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

Vectors that have been transformed with the `recip` function can be back-transformed by taking the reciprocal of the reciprocal.

The example below shows an example of the back-transformation of the `recip` function.

```
let(echo="b,c",
 a=array(100, 200, 300),
 b=recip(a),
 c=recip(b))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 0.01,
 0.005,
 0.003333333333333335
],
 "c": [
 100,
 200,
 300
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Z-scores

The `zscores` function converts a numeric array to an array of z-scores. The z-score is the number of standard deviations a number is from the mean.

The example below computes the z-scores for the values in an array.

```
let(a=array(1,2,3),
 b=zscores(a))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 -1,
 0,
 1
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 27
 }
]
 }
}
```

## Probability Distributions

This section of the user guide covers the probability distribution framework included in the math expressions library.

### Probability Distribution Framework

The probability distribution framework includes many commonly used [real](#) and [discrete](#) probability distributions, including support for [empirical](#) and [enumerated](#) distributions that model real world data.

The probability distribution framework also includes a set of functions that use the probability distributions to support probability calculations and sampling.

#### Real Distributions

The probability distribution framework has the following functions which support well known real probability distributions:

- `normalDistribution`: Creates a normal distribution function.
- `logNormalDistribution`: Creates a log normal distribution function.
- `gammaDistribution`: Creates a gamma distribution function.
- `betaDistribution`: Creates a beta distribution function.
- `uniformDistribution`: Creates a uniform real distribution function.
- `weibullDistribution`: Creates a Weibull distribution function.
- `triangularDistribution`: Creates a triangular distribution function.
- `constantDistribution`: Creates constant real distribution function.

## Empirical Distribution

The `empiricalDistribution` function creates a real probability distribution from actual data. An empirical distribution can be used interchangeably with any of the theoretical real distributions.

### Discrete

The probability distribution framework has the following functions which support well known discrete probability distributions:

- `poissonDistribution`: Creates a Poisson distribution function.
- `binomialDistribution`: Creates a binomial distribution function.
- `uniformIntegerDistribution`: Creates a uniform integer distribution function.
- `geometricDistribution`: Creates a geometric distribution function.
- `zipFDistribution`: Creates a Zipf distribution function.

### Enumerated Distributions

The `enumeratedDistribution` function creates a discrete distribution function from a data set of discrete values, or from an enumerated list of values and probabilities.

Enumerated distribution functions can be used interchangeably with any of the theoretical discrete distributions.

### Cumulative Probability

The `cumulativeProbability` function can be used with all probability distributions to calculate the cumulative probability of encountering a specific random variable within a specific distribution.

Below is example of calculating the cumulative probability of a random variable within a normal distribution.

```
let(a=normalDistribution(10, 5),
 b=cumulativeProbability(a, 12))
```

In this example a normal distribution function is created with a mean of 10 and a standard deviation of 5. Then the cumulative probability of the value 12 is calculated for this specific distribution.

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": 0.6554217416103242
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

Below is an example of a cumulative probability calculation using an empirical distribution.

In the example an empirical distribution is created from a random sample taken from the price\_f field.

The cumulative probability of the value .75 is then calculated. The price\_f field in this example was generated using a uniform real distribution between 0 and 1, so the output of the cumulativeProbability function is very close to .75.

```
let(a=random(collection1, q="*:*", rows="30000", fl="price_f"),
 b=col(a, price_f),
 c=empiricalDistribution(b),
 d=cumulativeProbability(c, .75))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": 0.7554217416103242
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Discrete Probability

The probability function can be used with any discrete distribution function to compute the probability of a discrete value.

Below is an example which calculates the probability of a discrete value within a Poisson distribution.

In the example a Poisson distribution function is created with a mean of 100. Then the probability of encountering a sample of the discrete value 101 is calculated for this specific distribution.

```
let(a=poissonDistribution(100),
 b=probability(a, 101))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": 0.039466333474403106
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

Below is an example of a probability calculation using an enumerated distribution.

In the example an enumerated distribution is created from a random sample taken from the day\_i field, which was created using a uniform integer distribution between 0 and 30.

The probability of the discrete value 10 is then calculated.

```
let(a=random(collection1, q="*:*", rows="30000", fl="day_i"),
 b=col(a, day_i),
 c=enumeratedDistribution(b),
 d=probability(c, 10))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "d": 0.03356666666666666
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 488
 }
]
 }
}
```

## Sampling

All probability distributions support sampling. The sample function returns 1 or more random samples from a probability distribution.

Below is an example drawing a single sample from a normal distribution.

```
let(a=normalDistribution(10, 5),
 b=sample(a))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": 11.24578055004963
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

Below is an example drawing 10 samples from a normal distribution.

```
let(a=normalDistribution(10, 5),
 b=sample(a, 10))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 10.18444709339441,
 9.466947971749377,
 1.2420697166234458,
 11.074501226984806,
 7.659629052136225,
 0.4440887839190708,
 13.710925254778786,
 2.089566359480239,
 0.7907293097654424,
 2.8184587681006734
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 3
 }
]
 }
}
```

## Multivariate Normal Distribution

The multivariate normal distribution is a generalization of the univariate normal distribution to higher dimensions.

The multivariate normal distribution models two or more random variables that are normally distributed. The relationship between the variables is defined by a covariance matrix.

### Sampling

The `sample` function can be used to draw samples from a multivariate normal distribution in much the same way as a univariate normal distribution.

The difference is that each sample will be an array containing a sample drawn from each of the underlying normal distributions. If multiple samples are drawn, the `sample` function returns a matrix with a sample in each row. Over the long term the columns of the sample matrix will conform to the covariance matrix used to parametrize the multivariate normal distribution.

The example below demonstrates how to initialize and draw samples from a multivariate normal distribution.

In this example 5000 random samples are selected from a collection of log records. Each sample contains the fields `filesize_d` and `response_d`. The values of both fields conform to a normal distribution.

Both fields are then vectorized. The `filesize_d` vector is stored in variable `b` and the `response_d` variable is stored in variable `c`.

An array is created that contains the means of the two vectorized fields.

Then both vectors are added to a matrix which is transposed. This creates an observation matrix where each row contains one observation of `filesize_d` and `response_d`. A covariance matrix is then created from the columns of the observation matrix with the `cov` function. The covariance matrix describes the covariance between `filesize_d` and `response_d`.

The `multivariateNormalDistribution` function is then called with the array of means for the two fields and the covariance matrix. The model for the multivariate normal distribution is assigned to variable `g`.

Finally five samples are drawn from the multivariate normal distribution.

```
let(a=random(collection2, q="*:*", rows="5000", fl="filesize_d, response_d"),
 b=col(a, filesize_d),
 c=col(a, response_d),
 d=array(mean(b), mean(c)),
 e=transpose(matrix(b, c)),
 f=cov(e),
 g=multiVariateNormalDistribution(d, f),
 h=sample(g, 5))
```

The samples are returned as a matrix, with each row representing one sample. There are two columns in the matrix. The first column contains samples for `filesize_d` and the second column contains samples for `response_d`. Over the long term the covariance between the columns will conform to the covariance matrix used to instantiate the multivariate normal distribution.

```
{
 "result-set": {
 "docs": [
 {
 "h": [
 [
 41974.85669321393,
 779.4097049705296
],
 [
 42869.19876441414,
 834.2599296790783
],
 [
 38556.30444839889,
 720.3683470060988
],
 [
 37689.31290928216,
 686.5549428100018
],
 [
 40564.74398214547,
 769.9328090774
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 162
 }
]
 }
}
```

## Monte Carlo Simulations

Monte Carlo simulations are commonly used to model the behavior of stochastic systems. This section describes how to perform both uncorrelated and correlated Monte Carlo simulations using the sampling capabilities of the probability distribution framework.

### Uncorrelated Simulations

Uncorrelated Monte Carlo simulations model stochastic systems with the assumption that the underlying random variables move independently of each other. A simple example of a Monte Carlo simulation using two independently changing random variables is described below.

In this example a Monte Carlo simulation is used to determine the probability that a simple hinge assembly will fall within a required length specification.

The hinge has two components A and B. The combined length of the two components must be less than 5 centimeters to fall within specification.

A random sampling of lengths for component A has shown that its length conforms to a normal distribution with a mean of 2.2 centimeters and a standard deviation of .0195 centimeters.

A random sampling of lengths for component B has shown that its length conforms to a normal distribution with a mean of 2.71 centimeters and a standard deviation of .0198 centimeters.

```
let(componentA=normalDistribution(2.2, .0195), ①
 componentB=normalDistribution(2.71, .0198), ②
 simresults=monteCarlo(sampleA=sample(componentA), ③
 sampleB=sample(componentB),
 add(sampleA, sampleB), ④
 100000), ⑤
 simmodel=empiricalDistribution(simresults), ⑥
 prob=cumulativeProbability(simmodel, 5)) ⑦
```

The Monte Carlo simulation below performs the following steps:

- ① A normal distribution with a mean of 2.2 and a standard deviation of .0195 is created to model the length of componentA.
- ② A normal distribution with a mean of 2.71 and a standard deviation of .0198 is created to model the length of componentB.
- ③ The monteCarlo function samples from the componentA and componentB distributions and sets the values to variables sampleA and sampleB.
- ④ It then calls the add(sampleA, sampleB)\* function to find the combined lengths of the samples.
- ⑤ The monteCarlo function runs a set number of times, 100000, and collects the results in an array. Each time the function is called new samples are drawn from the componentA and componentB distributions. On each run, the add function adds the two samples to calculate the combined length. The result of each run is collected in an array and assigned to the simresults variable.
- ⑥ An empiricalDistribution function is then created from the simresults array to model the distribution of the simulation results.
- ⑦ Finally, the cumulativeProbability function is called on the simmodel to determine the cumulative probability that the combined length of the components is 5 or less.

Based on the simulation there is .9994371944629039 probability that the combined length of a component pair will be 5 or less:

```
{
 "result-set": {
 "docs": [
 {
 "prob": 0.9994371944629039
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 660
 }
]
 }
}
```

## Correlated Simulations

The simulation above assumes that the lengths of componentA and componentB vary independently. What would happen to the probability model if there was a correlation between the lengths of componentA and componentB?

In the example below a database containing assembled pairs of components is used to determine if there is a correlation between the lengths of the components, and how the correlation effects the model.

Before performing a simulation of the effects of correlation on the probability model its useful to understand what the correlation is between the lengths of componentA and componentB.

```
let(a=random(collection5, q="*:*", rows="5000", fl="componentA_d, componentB_d"), ①
 b=col(a, componentA_d)), ②
 c=col(a, componentB_d)),
 d=corr(b, c)) ③
```

- ① In the example, 5000 random samples are selected from a collection of assembled hinges. Each sample contains lengths of the components in the fields componentA\_d and componentB\_d.
- ② Both fields are then vectorized. The **componentA\_d** vector is stored in variable **b** and the **componentB\_d** variable is stored in variable **c**.
- ③ Then the correlation of the two vectors is calculated using the corr function.

Note from the result that the outcome from corr is 0.9996931313216989. This means that componentA\_d and \*componentB\_d are almost perfectly correlated.

```
{
 "result-set": {
 "docs": [
 {
 "d": 0.9996931313216989
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 309
 }
]
 }
}
```

## Correlation Effects on the Probability Model

The example below explores how to use a multivariate normal distribution function to model how correlation effects the probability of hinge defects.

In this example 5000 random samples are selected from a collection containing length data for assembled hinges. Each sample contains the fields componentA\_d and componentB\_d.

Both fields are then vectorized. The componentA\_d vector is stored in variable **b** and the componentB\_d variable is stored in variable **c**.

An array is created that contains the means of the two vectorized fields.

Then both vectors are added to a matrix which is transposed. This creates an observation matrix where each row contains one observation of componentA\_d and componentB\_d. A covariance matrix is then created from the columns of the observation matrix with the cov function. The covariance matrix describes the covariance between componentA\_d and componentB\_d.

The multivariateNormalDistribution function is then called with the array of means for the two fields and the covariance matrix. The model for the multivariate normal distribution is stored in variable **g**.

The monteCarlo function then calls the function add(sample(g)) 50000 times and collects the results in a vector. Each time the function is called a single sample is drawn from the multivariate normal distribution. Each sample is a vector containing one componentA and componentB pair. The add function adds the values in the vector to calculate the length of the pair. Over the long term the samples drawn from the multivariate normal distribution will conform to the covariance matrix used to construct it.

Just as in the non-correlated example an empirical distribution is used to model probabilities of the simulation vector and the cumulativeProbability function is used to compute the cumulative probability that the combined component length will be 5 centimeters or less.

Notice that the probability of a hinge meeting specification has dropped to 0.9889517439980468. This is because the strong correlation between the lengths of components means that their lengths rise together causing more hinges to fall out of the 5 centimeter specification.

```
let(a=random(hinges, q="*:*", rows="5000", fl="componentA_d, componentB_d"),
 b=col(a, componentA_d),
 c=col(a, componentB_d),
 cor=corr(b,c),
 d=array(mean(b), mean(c)),
 e=transpose(matrix(b, c)),
 f=cov(e),
 g=multiVariateNormalDistribution(d, f),
 h=monteCarlo(add(sample(g)), 50000),
 i=empiricalDistribution(h),
 j=cumulativeProbability(i, 5))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "j": 0.9889517439980468
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 599
 }
]
 }
}
```

## Time Series

This section of the user guide provides an overview of time series **aggregation**, **smoothing** and **differencing**.

### Time Series Aggregation

The timeseries function performs fast, distributed time series aggregation leveraging Solr's builtin faceting and date math capabilities.

The example below performs a monthly time series aggregation:

```
timeseries(collection1,
 q="*:*",
 field="recdate_dt",
 start="2012-01-20T17:33:18Z",
 end="2012-12-20T17:33:18Z",
 gap="+1MONTH",
 format="YYYY-MM",
 count(*))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "recdate_dt": "2012-01",
 "count(*)": 8703
 },
 {
 "recdate_dt": "2012-02",
 "count(*)": 8648
 },
 {
 "recdate_dt": "2012-03",
 "count(*)": 8621
 },
 {
 "recdate_dt": "2012-04",
 "count(*)": 8533
 },
 {
 "recdate_dt": "2012-05",
 "count(*)": 8792
 },
 {
 "recdate_dt": "2012-06",
 "count(*)": 8598
 },
 {
 "recdate_dt": "2012-07",
 "count(*)": 8679
 },
 {
 "recdate_dt": "2012-08",
 "count(*)": 8469
 },
 {
 "recdate_dt": "2012-09",
 "count(*)": 8637
 },
 {
 "recdate_dt": "2012-10",
 "count(*)": 8536
 },
 {
 "recdate_dt": "2012-11",
 "count(*)": 8785
 },
 {
 "EOF": true,
]
 }
}
```

```
 "RESPONSE_TIME": 16
 }
]
}
}
```

## Vectorizing the Time Series

Before a time series result can be operated on by math expressions the data will need to be vectorized. Specifically in the example above, the aggregation field count(\*) will need to be moved into an array. As described in the Streams and Vectorization section of the user guide, the col function can be used to copy a numeric column from a list of tuples into an array.

The expression below demonstrates the vectorization of the count(\*) field.

```
let(a=timeseries(collection1,
 q=*:*,
 field="test_dt",
 start="2012-01-20T17:33:18Z",
 end="2012-12-20T17:33:18Z",
 gap="+1MONTH",
 format="YYYY-MM",
 count(*)),
 b=col(a, count*))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 8703,
 8648,
 8621,
 8533,
 8792,
 8598,
 8679,
 8469,
 8637,
 8536,
 8785
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 5
 }
]
 }
}
```

## Smoothing

Time series smoothing is often used to remove the noise from a time series and help spot the underlying trends. The math expressions library has three **sliding window** approaches for time series smoothing. The **sliding window** approaches use a summary value from a sliding window of the data to calculate a new set of smoothed data points.

The three **sliding window** functions are lagging indicators, which means they don't start to move in the direction of the trend until the trend effects the summary value of the sliding window. Because of this lagging quality these smoothing functions are often used to confirm the direction of the trend.

### Moving Average

The `movingAvg` function computes a simple moving average over a sliding window of data. The example below generates a time series, vectorizes the `count(*)` field and computes the moving average with a window size of 3.

The moving average function returns an array that is of shorter length than the original data set. This is because results are generated only when a full window of data is available for computing the average. With a window size of three the moving average will begin generating results at the 3rd value. The prior values are not included in the result.

This is true for all the sliding window functions.

```
let(a=timeseries(collection1,
 q=*:*,
 field="test_dt",
 start="2012-01-20T17:33:18Z",
 end="2012-12-20T17:33:18Z",
 gap="+1MONTH",
 format="YYYY-MM",
 count(*)),
b=col(a, count(*)),
c=movingAvg(b, 3))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 8657.333333333334,
 8600.666666666666,
 8648.666666666666,
 8641,
 8689.666666666666,
 8582,
 8595,
 8547.333333333334,
 8652.666666666666
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 7
 }
]
 }
}
```

## Exponential Moving Average

The expMovingAvg function uses a different formula for computing the moving average that responds faster to changes in the underlying data. This means that it is less of a lagging indicator than the simple moving average.

Below is an example that computes an exponential moving average:

```
let(a=timeseries(collection1, q=*:*,
 field="test_dt",
 start="2012-01-20T17:33:18Z",
 end="2012-12-20T17:33:18Z",
 gap="+1MONTH",
 format="YYYY-MM",
 count(*)),
 b=col(a, count(*)),
 c=expMovingAvg(b, 3))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 8657.33333333334,
 8595.166666666668,
 8693.58333333334,
 8645.791666666668,
 8662.39583333334,
 8565.697916666668,
 8601.34895833334,
 8568.674479166668,
 8676.83723958334
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 5
 }
]
 }
}
```

## Moving Median

The `movingMedian` function uses the median of the sliding window rather than the average. In many cases the moving median will be more **robust** to outliers than moving averages.

Below is an example computing the moving median:

```
let(a=timeseries(collection1,
 q=*:*,
 field="test_dt",
 start="2012-01-20T17:33:18Z",
 end="2012-12-20T17:33:18Z",
 gap="+1MONTH",
 format="YYYY-MM",
 count(*)),
b=col(a, count(*)),
c=movingMedian(b, 3))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 8648,
 8621,
 8621,
 8598,
 8679,
 8598,
 8637,
 8536,
 8637
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 7
 }
]
 }
}
```

## Differencing

Differencing is often used to remove the trend or seasonality from a time series. This is known as making a time series **stationary**.

### First Difference

The actual technique of differencing is to use the difference between values rather than the original values. The **first difference** takes the difference between a value and the value that came directly before it. The first difference is often used to remove the trend from a time series.

In the example below, the diff function computes the first difference of a time series. The result array

length is one value smaller then the original array. This is because the diff function only returns a result for values where the prior value has been subtracted.

```
let(a=timeseries(collection1,
 q=*:*,
 field="test_dt",
 start="2012-01-20T17:33:18Z",
 end="2012-12-20T17:33:18Z",
 gap="+1MONTH",
 format="YYYY-MM",
 count(*)),
b=col(a, count(*)),
c=diff(b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 -55,
 -27,
 -88,
 259,
 -194,
 81,
 -210,
 168,
 -101,
 249
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 11
 }
]
 }
}
```

## Lagged Differences

The diff function has an optional second parameter to specify a lag in the difference. If a lag is specified the difference is taken between a value and the value at a specified lag in the past. Lagged differences are often used to remove seasonality from a time series.

The simple example below demonstrates how lagged differencing works. Notice that the array in the example follows a simple repeated pattern. This type of pattern is often displayed with seasonality. In this example we can remove this pattern using the diff function with a lag of 4. This will subtract the value

lagging four indexes behind the current index. Notice that result set size is the original array size minus the lag. This is because the diff function only returns results for values where the lag of 4 is possible to compute.

```
let(a=array(1,2,5,2,1,2,5,2,1,2,5),
 b=diff(a, 4))
```

Expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 0,
 0,
 0,
 0,
 0,
 0,
 0
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Linear Regression

The math expressions library supports simple and multivariate linear regression.

### Simple Linear Regression

The regress function is used to build a linear regression model between two random variables. Sample observations are provided with two numeric arrays. The first numeric array is the independent variable and the second array is the dependent variable.

In the example below the random function selects 5000 random samples each containing the fields filesize\_d and response\_d. The two fields are vectorized and stored in variables **b** and **c**. Then the regress function performs a regression analysis on the two numeric arrays.

The regress function returns a single tuple with the results of the regression analysis.

```
let(a=random(collection2, q="*:*", rows="5000", fl="filesize_d, response_d"),
 b=col(a, filesize_d),
 c=col(a, response_d),
 d=regress(b, c))
```

Note that in this regression analysis the value of RSquared is .75. This means that changes in filesize\_d explain 75% of the variability of the response\_d variable:

```
{
 "result-set": {
 "docs": [
 {
 "d": {
 "significance": 0,
 "totalSumSquares": 10564812.895147054,
 "R": 0.8674822407146515,
 "RSquared": 0.7525254379553127,
 "meanSquareError": 523.1137343558588,
 "intercept": -49.528134913099095,
 "slopeConfidenceInterval": 0.0003171801710329995,
 "regressionSumSquares": 7950290.450836472,
 "slope": 0.019945557923159506,
 "interceptStdErr": 6.489732340389941,
 "N": 5000
 }
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 98
 }
]
 }
}
```

## Prediction

The predict function uses the regression model to make predictions. Using the example above the regression model can be used to predict the value of response\_d given a value for filesize\_d.

In the example below the predict function uses the regression analysis to predict the value of response\_d for the filesize\_d value of 40000.

```
let(a=random(collection2, q="*:*", rows="5000", fl="filesize_d, response_d"),
 b=col(a, filesize_d),
 c=col(a, response_d),
 d=regress(b, c),
 e=predict(d, 40000))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "e": 748.079241022975
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 95
 }
]
 }
}
```

The predict function can also make predictions for an array of values. In this case it returns an array of predictions.

In the example below the predict function uses the regression analysis to predict values for each of the 5000 samples of filesize\_d used to generate the model. In this case 5000 predictions are returned.

```
let(a=random(collection2, q="*:*", rows="5000", fl="filesize_d, response_d"),
 b=col(a, filesize_d),
 c=col(a, response_d),
 d=regress(b, c),
 e=predict(d, b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "e": [
 742.2525322514165,
 709.6972488729955,
 687.8382568904871,
 820.2511324266264,
 720.4006432289061,
 761.1578181053039,
 759.1304101159126,
 699.5597256337142,
 742.4738911248204,
 769.0342605881644,
 746.6740473150268,
 ...
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 113
 }
]
 }
}
```

## Residuals

The difference between the observed value and the predicted value is known as the residual. There isn't a specific function to calculate the residuals but vector math can be used to perform the calculation.

In the example below the predictions are stored in variable **e**. The `ebeSubtract` function is then used to subtract the predictions from the actual `response_d` values stored in variable **c**. Variable **f** contains the array of residuals.

```
let(a=random(collection2, q="*:*", rows="5000", fl="filesize_d, response_d"),
 b=col(a, filesize_d),
 c=col(a, response_d),
 d=regress(b, c),
 e=predict(d, b),
 f=ebeSubtract(c, e))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "e": [
 31.30678554491226,
 -30.292830927953446,
 -30.49508862647258,
 -30.499884780783532,
 -9.696458959319784,
 -30.521563961535094,
 -30.28380938033081,
 -9.890289849359306,
 30.819723560583157,
 -30.213178859683012,
 -30.609943619066826,
 10.527700442607625,
 10.68046928406568,
 ...
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 113
 }
]
 }
}
```

## Multivariate Linear Regression

The `olsRegress` function performs a multivariate linear regression analysis. Multivariate linear regression models the linear relationship between two or more independent variables and a dependent variable.

The example below extends the simple linear regression example by introducing a new independent variable called `service_d`. The `service_d` variable is the service level of the request and it can range from 1 to 4 in the data-set. The higher the service level, the higher the bandwidth available for the request.

Notice that the two independent variables `filesize_d` and `service_d` are vectorized and stored in the variables `b` and `c`. The variables `b` and `c` are then added as rows to a matrix. The matrix is then transposed so that each row in the matrix represents one observation with `filesize_d` and `service_d`. The `olsRegress` function then performs the multivariate regression analysis using the observation matrix as the independent variables and the `response_d` values, stored in variable `d`, as the dependent variable.

```
let(a=random(collection2, q="*:*", rows="30000", fl="filesize_d, service_d, response_d"),
 b=col(a, filesize_d),
 c=col(a, service_d),
 d=col(a, response_d),
 e=transpose(matrix(b, c)),
 f=olsRegress(e, d))
```

Notice in the response that the RSquared of the regression analysis is 1. This means that linear relationship between filesize\_d and service\_d describe 100% of the variability of the response\_d variable:

```
{
 "result-set": {
 "docs": [
 {
 "f": {
 "regressionParametersStandardErrors": [
 2.0660690430026933e-13,
 5.1212982077663434e-18,
 9.10920932555875e-15
],
 "RSquared": 1,
 "regressionParameters": [
 6.553210695971329e-12,
 0.019999999999999858,
 -20.49999999999968
],
 "regressandVariance": 2124.130825172683,
 "regressionParametersVariance": [
 [
 0.013660174897582315,
 -3.361258014840509e-7,
 -0.00006893737578369605
],
 [
 -3.361258014840509e-7,
 8.393183709503206e-12,
 6.430253229589981e-11
],
 [
 -0.00006893737578369605,
 6.430253229589981e-11,
 0.000026553878455570856
]
],
 "adjustedRSquared": 1,
 "residualSumSquares": 9.373703759269822e-20
 }
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 690
 }
]
 }
}
```

## Prediction

The predict function can also be used to make predictions for multivariate linear regression.

Below is an example of a single prediction using the multivariate linear regression model and a single

observation. The observation is an array that matches the structure of the observation matrix used to build the model. In this case the first value represents a `filesize_d` of 40000 and the second value represents a `service_d` of 4.

```
let(a=random(collection2, q="*:*", rows="5000", fl="filesize_d, service_d, response_d"),
 b=col(a, filesize_d),
 c=col(a, service_d),
 d=col(a, response_d),
 e=transpose(matrix(b, c)),
 f=olsRegress(e, d),
 g=predict(f, array(40000, 4)))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "g": 718.000000000005
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 117
 }
]
 }
}
```

The predict function can also make predictions for more than one multivariate observation. In this scenario an observation matrix used.

In the example below the observation matrix used to build the multivariate regression model is passed to the predict function and it returns an array of predictions.

```
let(a=random(collection2, q="*:*", rows="5000", fl="filesize_d, service_d, response_d"),
 b=col(a, filesize_d),
 c=col(a, service_d),
 d=col(a, response_d),
 e=transpose(matrix(b, c)),
 f=olsRegress(e, d),
 g=predict(f, e))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "e": [
 685.498283591961,
 801.2175699959365,
 776.7638245911025,
 610.3559852681935,
 751.0925865965207,
 787.2914663381897,
 744.3632053810668,
 688.3729301599697,
 765.367783417171,
 724.9309687628346,
 834.4350712384264,
 ...
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 113
 }
]
 }
}
```

## Residuals

Once the predictions are generated the residuals can be calculated using the same approach used with simple linear regression.

Below is an example of the residuals calculation following a multivariate linear regression. In the example the predictions stored variable **g** are subtracted from observed values stored in variable **d**.

```
let(a=random(collection2, q="*:*", rows="5000", fl="filesize_d, service_d, response_d"),
 b=col(a, filesize_d),
 c=col(a, service_d),
 d=col(a, response_d),
 e=transpose(matrix(b, c)),
 f=olsRegress(e, d),
 g=predict(f, e),
 h=ebeSubtract(d, g))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "e": [
 1.1368683772161603e-13,
 1.1368683772161603e-13,
 0,
 1.1368683772161603e-13,
 0,
 1.1368683772161603e-13,
 0,
 2.2737367544323206e-13,
 1.1368683772161603e-13,
 2.2737367544323206e-13,
 1.1368683772161603e-13,
 ...
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 113
 }
]
 }
}
```

## Interpolation, Derivatives and Integrals

Interpolation, derivatives and integrals are three interrelated topics which are part of the field of mathematics called numerical analysis. This section explores the math expressions available for numerical analysis.

### Interpolation

Interpolation is used to construct new data points between a set of known control of points. The ability to predict new data points allows for sampling along the curve defined by the control points.

The interpolation functions described below all return an *interpolation model* that can be passed to other functions which make use of the sampling capability.

If returned directly the interpolation model returns an array containing predictions for each of the control points. This is useful in the case of loess interpolation which first smooths the control points and then interpolates the smoothed points. All other interpolation functions simply return the original control points because interpolation predicts a curve that passes through the original control points.

There are different algorithms for interpolation that will result in different predictions along the curve. The math expressions library currently supports the following interpolation functions:

- `lerp`: Linear interpolation predicts points that pass through each control point and form straight lines

between control points.

- **spline:** Spline interpolation predicts points that pass through each control point and form a smooth curve between control points.
- **akima:** Akima spline interpolation is similar to spline interpolation but is stable to outliers.
- **loess:** Loess interpolation first performs a non-linear local regression to smooth the original control points. Then a spline is used to interpolate the smoothed control points.

## Upsampling

Interpolation can be used to increase the sampling rate along a curve. One example of this would be to take a time series with samples every minute and create a data set with samples every second. In order to do this the data points between the minutes must be created.

The `predict` function can be used to predict values anywhere within the bounds of the interpolation range. The example below shows a very simple example of upsampling.

```
let(x=array(0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20), ①
 y=array(5, 10, 60, 190, 100, 130, 100, 20, 30, 10, 5), ②
 l=lerp(x, y), ③
 u=array(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20), ④
 p=predict(l, u)) ⑤
```

- ① In the example linear interpolation is performed on the arrays in variables `x` and `y`. The `x` variable, which is the x-axis, is a sequence from 0 to 20 with a stride of 2.
- ② The `y` variable defines the curve along the x-axis.
- ③ The `lerp` function performs the interpolation and returns the interpolation model.
- ④ The `u` value is an array from 0 to 20 with a stride of 1. This fills in the gaps of the original `x` axis. The `predict` function then uses the interpolation function in variable `l` to predict values for every point in the array assigned to variable `u`.
- ⑤ The variable `p` is the array of predictions, which is the upsampled set of `y` values.

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "g": [
 5,
 7.5,
 10,
 35,
 60,
 125,
 190,
 145,
 100,
 115,
 130,
 115,
 100,
 60,
 20,
 25,
 30,
 20,
 10,
 7.5,
 5
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Smoothing Interpolation

The loess function is a smoothing interpolator which means it doesn't derive a function that passes through the original control points. Instead the loess function returns a function that smooths the original control points.

A technique known as local regression is used to compute the smoothed curve. The size of the neighborhood of the local regression can be adjusted to control how close the new curve conforms to the original control points.

The loess function is passed x- and y-axes and fits a smooth curve to the data. If only a single array is provided it is treated as the y-axis and a sequence is generated for the x-axis.

The example below uses the loess function to fit a curve to a set of y values in an array. The bandwidth parameter defines the percent of data to use for the local regression. The lower the percent the smaller the

neighborhood used for the local regression and the closer the curve will be to the original data.

```
let(echo="residuals, sumSqError",
y=array(0, 1, 2, 3, 4, 5.7, 6, 7, 7, 7, 6, 7, 7, 6, 5, 5, 3, 2, 1, 0),
curve=loess(y, bandwidth=.3),
residuals=ebeSubtract(y, curve),
sumSqError=sumSq(residuals))
```

In the example the fitted curve is subtracted from the original curve using the ebeSubtract function. The output shows the error between the fitted curve and the original curve, known as the residuals. The output also includes the sum-of-squares of the residuals which provides a measure of how large the error is:

```
{
 "result-set": {
 "docs": [
 {
 "residuals": [
 0,
 0,
 0,
 -0.040524802275866634,
 -0.10531988096456502,
 0.5906115002526198,
 0.004215074334896762,
 0.4201374330912433,
 0.09618315578013803,
 0.012107948556718817,
 -0.9892939034492398,
 0.012014364143757561,
 0.1093830927709325,
 0.523166271893805,
 0.09658362075164639,
 -0.011433819306139625,
 0.9899403519886416,
 -0.011707983372932773,
 -0.004223284004140737,
 -0.00021462867928434548,
 0.0018723112875456138
],
 "sumSqError": 2.8016013870800616
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

In the next example the curve is fit using a bandwidth of .25:

```
let(echo="residuals, sumSqError",
 y=array(0, 1, 2, 3, 4, 5.7, 6, 7, 6, 5, 5, 3, 2, 1, 0),
 curve=loess(y, .25),
 residuals=ebeSubtract(y, curve),
 sumSqError=sumSq(residuals))
```

Notice that the curve is a closer fit, shown by the smaller residuals and lower value for the sum-of-squares of the residuals:

```
{
 "result-set": {
 "docs": [
 {
 "residuals": [
 0,
 0,
 0,
 0,
 -0.19117650587715396,
 0.442863451538809,
 -0.18553845993358564,
 0.29990769020356645,
 0,
 0.23761890236245709,
 -0.7344358765888117,
 0.2376189023624491,
 0,
 0.30373119215254984,
 -3.552713678800501e-15,
 -0.23761890236245264,
 0.7344358765888046,
 -0.2376189023625095,
 0,
 2.842170943040401e-14,
 -2.4868995751603507e-14
],
 "sumSqError": 1.7539413576337557
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Derivatives

The derivative of a function measures the rate of change of the  $y$  value in respects to the rate of change of

the `x` value.

The derivative function can compute the derivative of any interpolation function. It can also compute the derivative of a derivative.

The example below computes the derivative for a loess interpolation function.

```
let(x=array(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20),
 y=array(0, 1, 2, 3, 4, 5.7, 6, 7, 7, 7.6, 7, 7, 7, 6, 5, 5, 3, 2, 1, 0),
 curve=loess(x, y, bandwidth=.3),
 derivative=derivative(curve))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "derivative": [
 1.0022002675659012,
 0.9955994648681976,
 1.0154018729613081,
 1.0643674501141696,
 1.0430879694757085,
 0.9698717643975381,
 0.7488201070357539,
 0.44627000894357516,
 0.19019561285422165,
 0.01703599324311178,
 -0.001908408138535126,
 -0.009121607450087499,
 -0.2576361507216319,
 -0.49378951291352746,
 -0.7288073815664,
 -0.9871806872210384,
 -1.0025400632604322,
 -1.001836567536853,
 -1.0076227586138085,
 -1.0021524620888589,
 -1.0020541789058157
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Integrals

An integral is a measure of the volume underneath a curve. The `integrate` function computes an integral for a specific range of an interpolated curve.

In the example below the `integrate` function computes an integral for the entire range of the curve, 0 through 20.

```
let(x=array(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20),
 y=array(0, 1, 2, 3, 4, 5.7, 6, 7, 7, 7, 6, 7, 7, 6, 5, 5, 3, 2, 1, 0),
 curve=loess(x, y, bandwidth=.3),
 integral=integrate(curve, 0, 20))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "integral": 90.17446104846645
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

In the next example an integral is computed for the range of 0 through 10.

```
let(x=array(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20),
 y=array(0, 1, 2, 3, 4, 5.7, 6, 7, 7, 7, 6, 7, 7, 6, 5, 5, 3, 2, 1, 0),
 curve=loess(x, y, bandwidth=.3),
 integral=integrate(curve, 0, 10))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "integral": 45.300912584519914
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Bicubic Spline

The `bicubicSpline` function can be used to interpolate and predict values anywhere within a grid of data.

A simple example will make this more clear:

```
let(years=array(1998, 2000, 2002, 2004, 2006),
 floors=array(1, 5, 9, 13, 17, 19),
 prices = matrix(array(300000, 320000, 330000, 350000, 360000, 370000),
 array(320000, 330000, 340000, 350000, 365000, 380000),
 array(400000, 410000, 415000, 425000, 430000, 440000),
 array(410000, 420000, 425000, 435000, 445000, 450000),
 array(420000, 430000, 435000, 445000, 450000, 470000)),
 bspline=bicubicSpline(years, floors, prices),
 prediction=predict(bspline, 2003, 8))
```

In this example a bicubic spline is used to interpolate a matrix of real estate data. Each row of the matrix represent specific years. Each column of the matrix represents floors of the building. The grid of numbers is the average selling price of an apartment for each year and floor. For example in 2002 the average selling price for the 9th floor was 415000 (row 3, column 3).

The `bicubicSpline` function is then used to interpolate the grid, and the `predict` function is used to predict a value for year 2003, floor 8. Notice that the matrix does not include a data point for year 2003, floor 8. The `bicubicSpline` function creates that data point based on the surrounding data in the matrix:

```
{
 "result-set": {
 "docs": [
 {
 "prediction": 418279.5009328358
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Curve Fitting

These functions support constructing a curve.

### Polynomial Curve Fitting

The `polyfit` function is a general purpose curve fitter used to model the non-linear relationship between two random variables.

The `polyfit` function is passed x- and y-axes and fits a smooth curve to the data. If only a single array is provided it is treated as the y-axis and a sequence is generated for the x-axis.

The `polyfit` function also has a parameter that specifies the degree of the polynomial. The higher the degree the more curves that can be modeled.

The example below uses the `polyfit` function to fit a curve to an array using a 3 degree polynomial. The fitted curve is then subtracted from the original curve. The output shows the error between the fitted curve and the original curve, known as the residuals. The output also includes the sum-of-squares of the residuals which provides a measure of how large the error is.

```
let(echo="residuals, sumSqError",
 y=array(0, 1, 2, 3, 4, 5.7, 6, 7, 6, 5, 5, 3, 2, 1, 0),
 curve=polyfit(y, 3),
 residuals=ebeSubtract(y, curve),
 sumSqError=sumSq(residuals))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "residuals": [
 0.5886274509803899,
 -0.0746078431372561,
 -0.49492135315664765,
 -0.6689571213100631,
 -0.5933591898297781,
 0.4352283990519288,
 0.32016160310277897,
 1.1647963800904968,
 0.272488687782805,
 -0.3534055160525744,
 0.2904697263520779,
 -0.7925296272355089,
 -0.5990476190476182,
 -0.12572829131652274,
 0.6307843137254909
],
 "sumSqError": 4.7294282482223595
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

In the next example the curve is fit using a 5 degree polynomial. Notice that the curve is fit closer, shown by the smaller residuals and lower value for the sum-of-squares of the residuals. This is because the higher polynomial produced a closer fit.

```
let(echo="residuals, sumSqError",
y=array(0, 1, 2, 3, 4, 5.7, 6, 7, 6, 5, 5, 3, 2, 1, 0),
curve=polyfit(y, 5),
residuals=ebeSubtract(y, curve),
sumSqError=sumSq(residuals))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "residuals": [
 -0.12337461300309674,
 0.22708978328173413,
 0.12266015718028167,
 -0.16502738747320755,
 -0.41142804563857105,
 0.2603044014808713,
 -0.12128970101106162,
 0.6234168308471704,
 -0.1754692675745293,
 -0.5379689969473249,
 0.4651616185671843,
 -0.288175756132409,
 0.027970945463215102,
 0.18699690402476687,
 -0.09086687306501587
],
 "sumSqError": 1.413089480179252
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Prediction, Derivatives and Integrals

The polyfit function returns a function that can be used with the predict function.

In the example below the x-axis is included for clarity. The polyfit function returns a function for the fitted curve. The predict function is then used to predict a value along the curve, in this case the prediction is made for the x value of 5.

```
let(x=array(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14),
 y=array(0, 1, 2, 3, 4, 5.7, 6, 7, 6, 5, 5, 3, 2, 1, 0),
 curve=polyfit(x, y, 5),
 p=predict(curve, 5))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "p": 5.439695598519129
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

The derivative and integrate functions can be used to compute the derivative and integrals for the fitted curve. The example below demonstrates how to compute a derivative for the fitted curve.

```
let(x=array(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14),
 y=array(0, 1, 2, 3, 4, 5.7, 6, 7, 6, 5, 5, 3, 2, 1, 0),
 curve=polyfit(x, y, 5),
 d=derivative(curve))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "d": [
 0.3198918573686361,
 0.9261492094077225,
 1.2374272373653175,
 1.30051359631081,
 1.1628032287629813,
 0.8722983646900058,
 0.47760852150945,
 0.02795050408827482,
 -0.42685159525716865,
 -0.8363663967611356,
 -1.1495552332084857,
 -1.3147721499346892,
 -1.2797639048258267,
 -0.9916699683185771,
 -0.3970225234002308
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Harmonic Curve Fitting

The harmonicFit function (or harmfit, for short) fits a smooth line through control points of a sine wave. The harmfit function is passed x- and y-axes and fits a smooth curve to the data. If a single array is provided it is treated as the y-axis and a sequence is generated for the x-axis.

The example below shows harmfit fitting a single oscillation of a sine wave. The harmfit function returns the smoothed values at each control point. The return value is also a model which can be used by the predict, derivative and integrate functions.

There are also three helper functions that can be used to retrieve the estimated parameters of the fitted model:

- getAmplitude: Returns the amplitude of the sine wave.
- getAngularFrequency: Returns the angular frequency of the sine wave.
- getPhase: Returns the phase of the sine wave.



The `harmfit` function works best when run on a single oscillation rather than a long sequence of oscillations. This is particularly true if the sine wave has noise. After the curve has been fit it can be extrapolated to any point in time in the past or future.

In the example below the `harmfit` function fits control points, provided as `x` and `y` axes, and then the angular frequency, phase and amplitude are retrieved from the fitted model.

```
let(echo="freq, phase, amp",
 x=array(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19),
 y=array(-0.7441113653915925,-0.8997532112139415, -0.9853140681578838, -0.9941296760805463,
 -0.9255133950087844, -0.7848096869247675, -0.5829778403072583, -0.33573836075915076,
 -0.06234851460699166, 0.215897602691855, 0.47732764497752245, 0.701579055431586,
 0.8711850882773975, 0.9729352782968976, 0.9989043923858761, 0.9470697190130273,
 0.8214686154479715, 0.631884041542757, 0.39308257356494, 0.12366424851680227),
 model=harmfit(x, y),
 freq=getAngularFrequency(model),
 phase=getPhase(model),
 amp=getAmplitude(model))
```

```
{
 "result-set": {
 "docs": [
 {
 "freq": 0.28,
 "phase": 2.4100000000000006,
 "amp": 0.999999999999999
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Interpolation and Extrapolation

The `harmfit` function returns a fitted model of the sine wave that can be used by the `predict` function to interpolate or extrapolate the sine wave.

The example below uses the fitted model to extrapolate the sine wave beyond the control points to the `x`-axis points 20, 21, 22, 23.

```
let(x=array(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19),
 y=array(-0.7441113653915925,-0.8997532112139415, -0.9853140681578838, -0.9941296760805463,
 -0.9255133950087844, -0.7848096869247675, -0.5829778403072583, -0.33573836075915076,
 -0.06234851460699166, 0.215897602691855, 0.47732764497752245, 0.701579055431586,
 0.8711850882773975, 0.9729352782968976, 0.9989043923858761, 0.9470697190130273,
 0.8214686154479715, 0.631884041542757, 0.39308257356494, 0.12366424851680227),
 model=harmfit(x, y),
 extrapolation=predict(model, array(20, 21, 22, 23)))
```

```
{
 "result-set": {
 "docs": [
 {
 "extrapolation": [
 -0.1553861764415666,
 -0.42233370833176975,
 -0.656386037906838,
 -0.8393130343914845
],
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Gaussian Curve Fitting

The gaussfit function fits a smooth curve through a Gaussian peak. This is shown in the example below.

```
let(x=array(0,1,2,3,4,5,6,7,8,9, 10),
 y=array(4,55,1200,3028,12000,18422,13328,6426,1696,239,20),
 f=gaussfit(x, y))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "f": [
 2.81764431935644,
 61.157417979413424,
 684.2328985468831,
 3945.9411154167447,
 11729.758936952656,
 17972.951897338007,
 14195.201949425435,
 5779.03836032222,
 1212.7224502169634,
 131.17742331530349,
 7.3138931735866946
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

Like the polyfit function, the gaussfit function returns a function that can be used directly by the predict, derivative and integrate functions.

The example below demonstrates how to compute an integral for a fitted Gaussian curve.

```
let(x=array(0,1,2,3,4,5,6,7,8,9, 10),
 y=array(4,55,1200,3028,12000,18422,13328,6426,1696,239,20),
 f=gaussfit(x, y),
 i=integrate(f, 0, 5))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "i": 25261.666789766092
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 3
 }
]
 }
}
```

## Digital Signal Processing

This section of the user guide explores functions that are commonly used in the field of Digital Signal Processing (DSP).

### Dot Product

The dotProduct function is used to calculate the dot product of two numeric arrays. The dot product is a fundamental calculation for the DSP functions discussed in this section. Before diving into the more advanced DSP functions its useful to develop a deeper intuition of the dot product.

The dot product operation is performed in two steps:

1. Element-by-element multiplication of two vectors which produces a vector of products.
2. Sum the vector of products to produce a scalar result.

This simple bit of math has a number of important applications.

### Representing Linear Combinations

The dotProduct performs the math of a *linear combination*. A linear combination has the following form:

$$(a1*v1)+(a2*v2) \dots$$

In the above example a1 and a2 are random variables that change. v1 and v2 are constant values.

When computing the dot product the elements of two vectors are multiplied together and the results are added. If the first vector contains random variables and the second vector contains constant values then the dot product is performing a linear combination.

This scenario comes up again and again in machine learning. For example both linear and logistic regression solve for a vector of constant weights. In order to perform a prediction, a dot product is calculated between a random observation vector and the constant weight vector. That dot product is a linear combination because one of the vectors holds constant weights.

Lets look at simple example of how a linear combination can be used to find the mean of a vector of numbers.

In the example below two arrays are set to variables **a** and **b** and then operated on by the dotProduct function. The output of the dotProduct function is set to variable **c**.

The mean function is then used to compute the mean of the first array which is set to the variable **d**.

Both the dot product and the mean are included in the output.

When we look at the output of this expression we see that the dot product and the mean of the first array are both 30.

The dotProduct function calculated the mean of the first array.

```
let(echo="c, d",
 a=array(10, 20, 30, 40, 50),
 b=array(.2, .2, .2, .2, .2),
 c=dotProduct(a, b),
 d=mean(a))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": 30,
 "d": 30
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

To get a better understanding of how the dot product calculated the mean we can perform the steps of the calculation using vector math and look at the output of each step.

In the example below the ebeMultiply function performs an element-by-element multiplication of two arrays. This is the first step of the dot product calculation. The result of the element-by-element multiplication is assigned to variable **c**.

In the next step the add function adds all the elements of the array in variable **c**.

Notice that multiplying each element of the first array by .2 and then adding the results is equivalent to the formula for computing the mean of the first array. The formula for computing the mean of an array is to add all the elements and divide by the number of elements.

The output includes the output of both the ebeMultiply function and the add function.

```
let(echo="c, d",
 a=array(10, 20, 30, 40, 50),
 b=array(.2, .2, .2, .2, .2),
 c=ebeMultiply(a, b),
 d=add(c))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 2,
 4,
 6,
 8,
 10
],
 "d": 30
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

In the example above two arrays were combined in a way that produced the mean of the first. In the second array each value was set to .2. Another way of looking at this is that each value in the second array is applying the same weight to the values in the first array. By varying the weights in the second array we can produce a different result. For example if the first array represents a time series, the weights in the second array can be set to add more weight to a particular element in the first array.

The example below creates a weighted average with the weight decreasing from right to left. Notice that the weighted mean of 36.666 is larger than the previous mean which was 30. This is because more weight was given to last element in the array.

```
let(echo="c, d",
 a=array(10, 20, 30, 40, 50),
 b=array(.0666666666666666, .1333333333333333, .2, .2666666666666666, .3333333333333333),
 c=ebeMultiply(a, b),
 d=add(c))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 0.666666666666666,
 2.666666666666666,
 6,
 10.66666666666664,
 16.66666666666665
],
 "d": 36.66666666666646
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Representing Correlation

Often when we think of correlation, we are thinking of *Pearson correlation* in the field of statistics. But the definition of correlation is actually more general: a mutual relationship or connection between two or more things. In the field of digital signal processing the dot product is used to represent correlation. The examples below demonstrates how the dot product can be used to represent correlation.

In the example below the dot product is computed for two vectors. Notice that the vectors have different values that fluctuate together. The output of the dot product is 190, which is hard to reason about because it's not scaled.

```
let(echo="c, d",
 a=array(10, 20, 30, 20, 10),
 b=array(1, 2, 3, 2, 1),
 c=dotProduct(a, b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": 190
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

One approach to scaling the dot product is to first scale the vectors so that both vectors have a magnitude of 1. Vectors with a magnitude of 1, also called unit vectors, are used when comparing only the angle between vectors rather than the magnitude. The `unitize` function can be used to unitize the vectors before calculating the dot product.

Notice in the example below the dot product result, set to variable `e`, is effectively 1. When applied to unit vectors the dot product will be scaled between 1 and -1. Also notice in the example `cosineSimilarity` is calculated on the unscaled vectors and the answer is also effectively 1. This is because cosine similarity is a scaled dot product.

```
let(echo="e, f",
 a=array(10, 20, 30, 20, 10),
 b=array(1, 2, 3, 2, 1),
 c=unitize(a),
 d=unitize(b),
 e=dotProduct(c, d),
 f=cosineSimilarity(a, b))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "e": 0.9999999999999998,
 "f": 0.9999999999999999
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

If we transpose the first two numbers in the first array, so that the vectors are not perfectly correlated, we see that the cosine similarity drops. This illustrates how the dot product represents correlation.

```
let(echo="c, d",
 a=array(20, 10, 30, 20, 10),
 b=array(1, 2, 3, 2, 1),
 c=cosineSimilarity(a, b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": 0.9473684210526314
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Convolution

The conv function calculates the convolution of two vectors. The convolution is calculated by reversing the second vector and sliding it across the first vector. The dot product of the two vectors is calculated at each point as the second vector is slid across the first vector. The dot products are collected in a third vector which is the convolution of the two vectors.

### Moving Average Function

Before looking at an example of convolution its useful to review the movingAvg function. The moving average function computes a moving average by sliding a window across a vector and computing the average of the window at each shift. If that sounds similar to convolution, that's because the movingAvg function is syntactic sugar for convolution.

Below is an example of a moving average with a window size of 5. Notice that original vector has 13 elements but the result of the moving average has only 9 elements. This is because the movingAvg function only begins generating results when it has a full window. In this case because the window size is 5 so the moving average starts generating results from the 4<sup>th</sup> index of the original array.

```
let(a=array(1, 2, 3, 4, 5, 6, 7, 6, 5, 4, 3, 2, 1),
 b=movingAvg(a, 5))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "b": [
 3,
 4,
 5,
 5.6,
 5.8,
 5.6,
 5,
 4,
 3
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Convolutional Smoothing

The moving average can also be computed using convolution. In the example below the conv function is used to compute the moving average of the first array by applying the second array as the filter.

Looking at the result, we see that it is not exactly the same as the result of the movingAvg function. That is because the conv pads zeros to the front and back of the first vector so that the window size is always full.

```
let(a=array(1, 2, 3, 4, 5, 6, 7, 6, 5, 4, 3, 2, 1),
 b=array(.2, .2, .2, .2, .2),
 c=conv(a, b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 0.2,
 0.6000000000000001,
 1.2,
 2.0000000000000004,
 3.0000000000000004,
 4,
 5,
 5.6000000000000005,
 5.8000000000000001,
 5.6000000000000005,
 5.0000000000000001,
 4,
 3,
 2,
 1.2000000000000002,
 0.6000000000000001,
 0.2
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

We achieve the same result as the `movingAvg` function by using the `copyOfRange` function to copy a range of the result that drops the first and last 4 values of the convolution result. In the example below the `precision` function is also used to remove floating point errors from the convolution result. When this is added the output is exactly the same as the `movingAvg` function.

```
let(a=array(1, 2, 3, 4, 5, 6, 7, 6, 5, 4, 3, 2, 1),
 b=array(.2, .2, .2, .2, .2),
 c=conv(a, b),
 d=copyOfRange(c, 4, 13),
 e=precision(d, 2))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "e": [
 3,
 4,
 5,
 5.6,
 5.8,
 5.6,
 5,
 4,
 3
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Cross-Correlation

Cross-correlation is used to determine the delay between two signals. This is accomplished by sliding one signal across another and calculating the dot product at each shift. The dot products are collected into a vector which represents the correlation at each shift. The highest dot product in the cross-correlation vector is the point where the two signals are most closely correlated.

The sliding dot product used in convolution can also be used to represent cross-correlation between two vectors. The only difference in the formula when representing correlation is that the second vector is **not reversed**.

Notice in the example below that the second vector is reversed by the `rev` function before it is operated on by the `conv` function. The `conv` function reverses the second vector so it will be flipped back to its original order to perform the correlation calculation rather than the convolution calculation.

Notice in the result the highest value is 217. This is the point where the two vectors have the highest correlation.

```
let(a=array(1, 2, 3, 4, 5, 6, 7, 6, 5, 4, 3, 2, 1),
 b=array(4, 5, 6, 7, 6, 5, 4, 3, 2, 1),
 c=conv(a, rev(b)))
```

When this expression is sent to the `/stream` handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": [
 1,
 4,
 10,
 20,
 35,
 56,
 84,
 116,
 149,
 180,
 203,
 216,
 217,
 204,
 180,
 148,
 111,
 78,
 50,
 28,
 13,
 4
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Find Delay

It is fairly simple to compute the delay from the cross-correlation result, but a convenience function called `finddelay` can be used to find the delay directly. Under the covers `finddelay` uses convolutional math to compute the cross-correlation vector and then computes the delay between the two signals.

Below is an example of the `finddelay` function. Notice that the `finddelay` function reports a 3 period delay between the first and second signal.

```
let(a=array(1, 2, 3, 4, 5, 6, 7, 6, 5, 4, 3, 2, 1),
 b=array(4, 5, 6, 7, 6, 5, 4, 3, 2, 1),
 c=finddelay(a, b))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "c": 3
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

## Oscillate (Sine Wave)

The oscillate function generates a periodic oscillating signal which can be used to model and study sine waves.

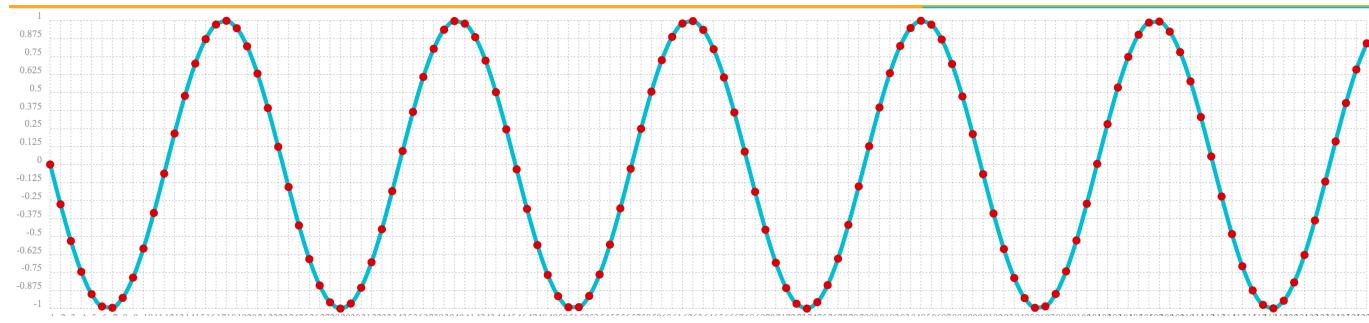
The oscillate function takes three parameters: **amplitude**, **angular frequency** and **phase** and returns a vector containing the y-axis points of a sine wave.

The y-axis points were generated from an x-axis sequence of 0-127.

Below is an example of the oscillate function called with an amplitude of 1, and angular frequency of .28 and phase of 1.57.

```
oscillate(1, 0.28, 1.57)
```

The result of the oscillate function is plotted below:

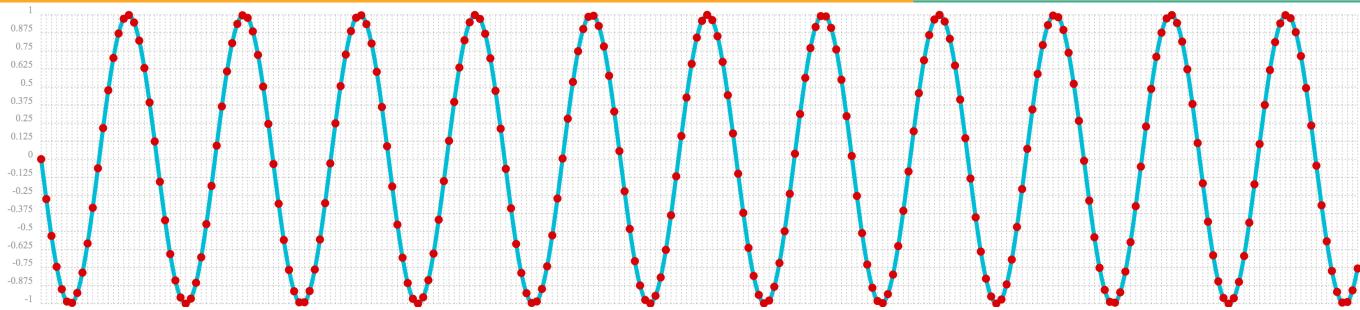


## Sine Wave Interpolation, Extrapolation

The oscillate function returns a function which can be used by the predict function to interpolate or extrapolate a sine wave. The example below extrapolates the sine wave to an x-axis sequence of 0-256.

```
let(a=oscillate(1, 0.28, 1.57),
 b=predict(a, sequence(256, 0, 1)))
```

The extrapolated sine wave is plotted below:



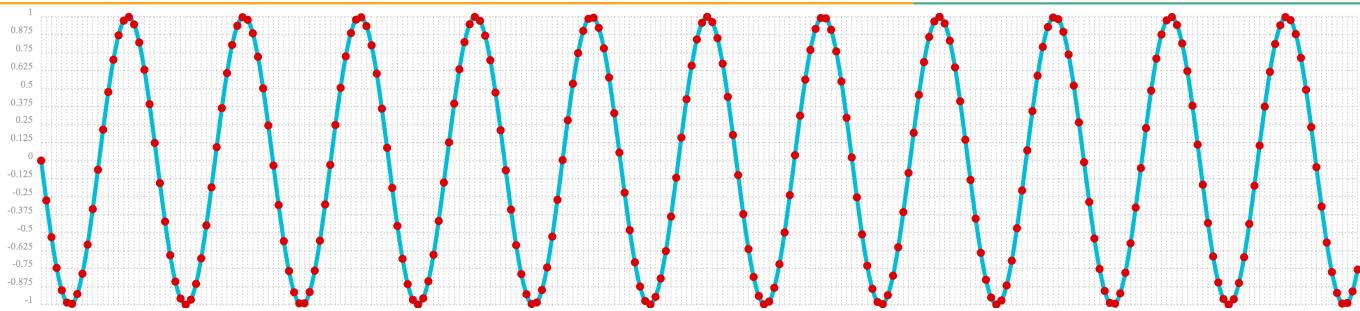
## Autocorrelation

Autocorrelation measures the degree to which a signal is correlated with itself. Autocorrelation is used to determine if a vector contains a signal or is purely random.

A few examples, with plots, will help to understand the concepts.

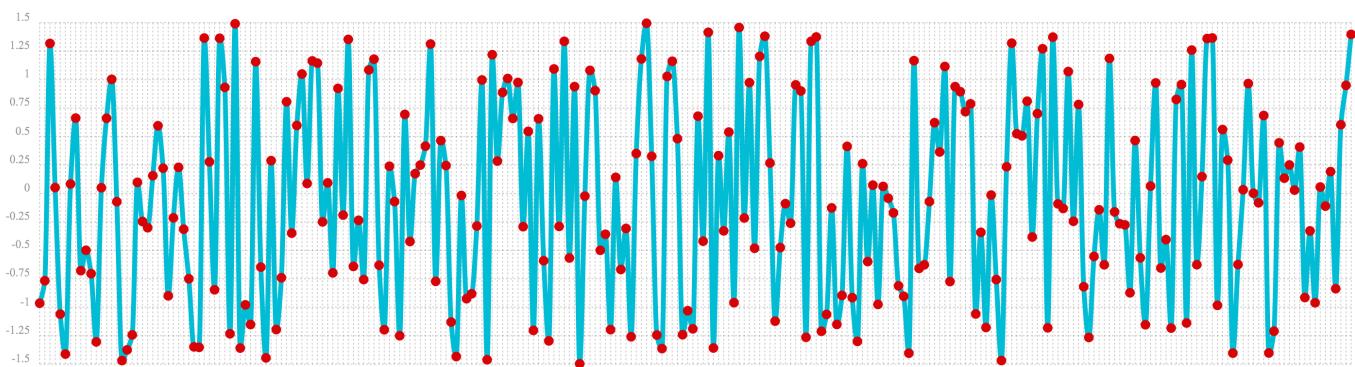
The first example simply revisits the example above of an extrapolated sine wave. The result of this is plotted in the image below. Notice that there is a structure to the plot that is clearly not random.

```
let(a=oscillate(1, 0.28, 1.57),
 b=predict(a, sequence(256, 0, 1)))
```



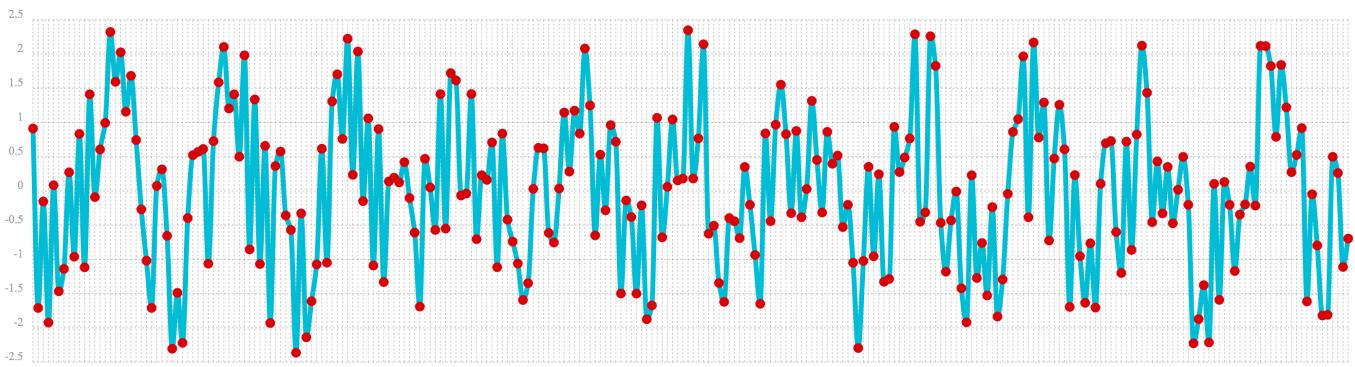
In the next example the sample function is used to draw 256 samples from a uniformDistribution to create a vector of random data. The result of this is plotted in the image below. Notice that there is no clear structure to the data and the data appears to be random.

```
sample(uniformDistribution(-1.5, 1.5), 256)
```



In the next example the random noise is added to the sine wave using the ebeAdd function. The result of this is plotted in the image below. Notice that the sine wave has been hidden somewhat within the noise. Its difficult to say for sure if there is structure. As plots becomes more dense it can become harder to see a pattern hidden within noise.

```
let(a=oscillate(1, 0.28, 1.57),
 b=predict(a, sequence(256, 0, 1)),
 c=sample(uniformDistribution(-1.5, 1.5), 256),
 d=ebeAdd(b,c))
```



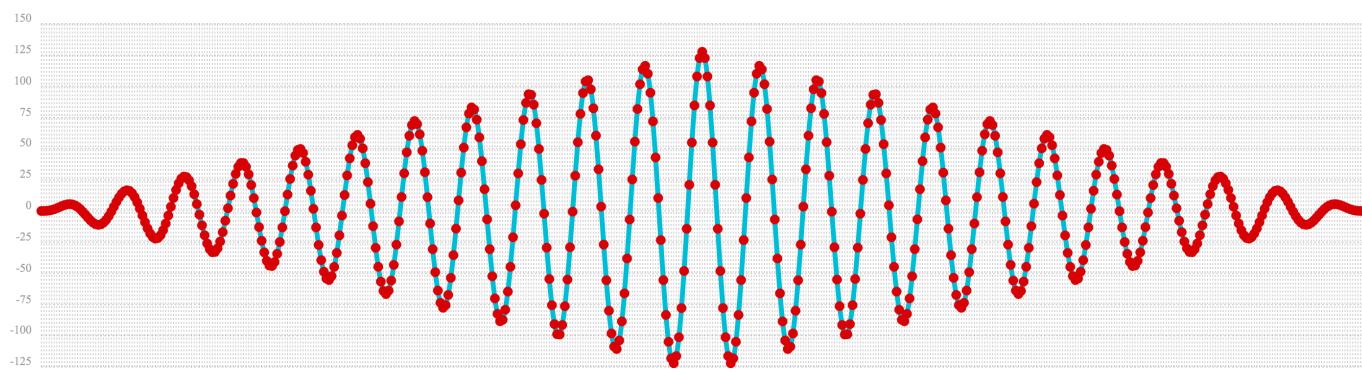
In the next examples autocorrelation is performed with each of the vectors shown above to see what the autocorrelation plots look like.

In the example below the conv function is used to autocorrelate the first vector which is the sine wave. Notice that the conv function is simply correlating the sine wave with itself.

The plot has a very distinct structure to it. As the sine wave is slid across a copy of itself the correlation moves up and down in increasing intensity until it reaches a peak. This peak is directly in the center and is the point where the sine waves are directly lined up. Following the peak the correlation moves up and down in decreasing intensity as the sine wave slides farther away from being directly lined up.

This is the autocorrelation plot of a pure signal.

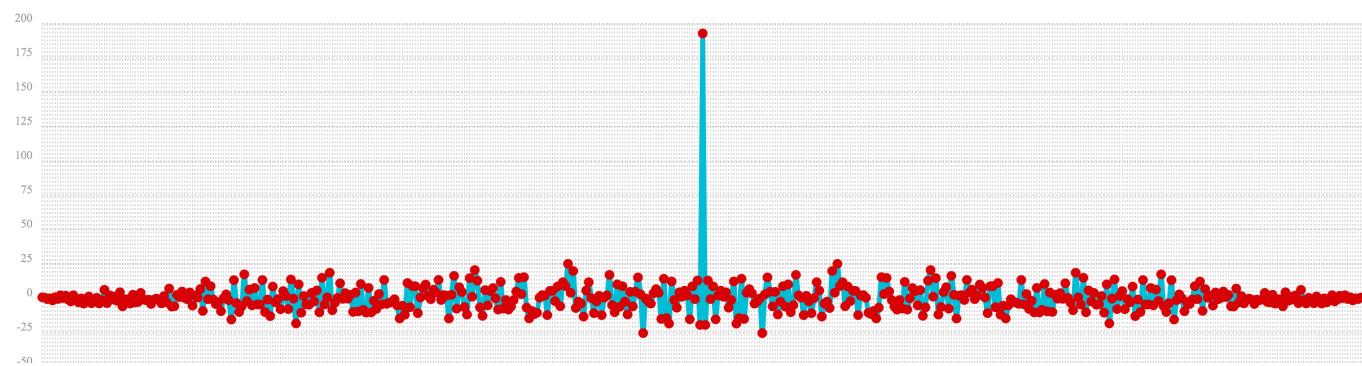
```
let(a=oscillate(1, 0.28, 1.57),
 b=predict(a, sequence(256, 0, 1)),
 c=conv(b, rev(b)))
```



In the example below autocorrelation is performed with the vector of pure noise. Notice that the autocorrelation plot has a very different plot then the sine wave. In this plot there is long period of low intensity correlation that appears to be random. Then in the center a peak of high intensity correlation where the vectors are directly lined up. This is followed by another long period of low intensity correlation.

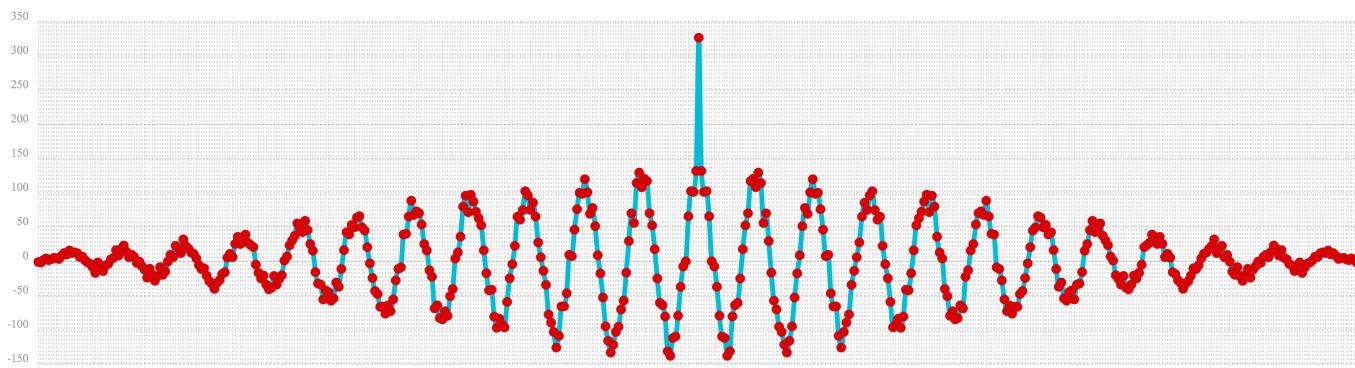
This is the autocorrelation plot of pure noise.

```
let(a=sample(uniformDistribution(-1.5, 1.5), 256),
 b=conv(a, rev(a)),
```



In the example below autocorrelation is performed on the vector with the sine wave hidden within the noise. Notice that this plot shows very clear signs of structure which is similar to autocorrelation plot of the pure signal. The correlation is less intense due to noise but the shape of the correlation plot suggests strongly that there is an underlying signal hidden within the noise.

```
let(a=oscillate(1, 0.28, 1.57),
 b=predict(a, sequence(256, 0, 1)),
 c=sample(uniformDistribution(-1.5, 1.5), 256),
 d=ebeAdd(b, c),
 e=conv(d, rev(d)))
```



## Discrete Fourier Transform

The convolution based functions described above are operating on signals in the time domain. In the time domain the X axis is time and the Y axis is the quantity of some value at a specific point in time.

The discrete Fourier Transform translates a time domain signal into the frequency domain. In the frequency domain the X axis is frequency, and Y axis is the accumulated power at a specific frequency.

The basic principle is that every time domain signal is composed of one or more signals (sine waves) at different frequencies. The discrete Fourier transform decomposes a time domain signal into its component frequencies and measures the power at each frequency.

The discrete Fourier transform has many important uses. In the example below, the discrete Fourier transform is used to determine if a signal has structure or if it is purely random.

### Complex Result

The `fft` function performs the discrete Fourier Transform on a vector of **real** data. The result of the `fft` function is returned as **complex** numbers. A complex number has two parts, **real** and **imaginary**. The imaginary part of the complex number is ignored in the examples below, but there are many tutorials on the FFT and that include complex numbers available online.

But before diving into the examples it is important to understand how the `fft` function formats the complex numbers in the result.

The `fft` function returns a matrix with two rows. The first row in the matrix is the **real** part of the complex result. The second row in the matrix is the **imaginary** part of the complex result.

The `rowAt` function can be used to access the rows so they can be processed as vectors. This approach was taken because all of the vector math functions operate on vectors of real numbers. Rather than introducing a complex number abstraction into the expression language, the `fft` result is represented as two vectors of real numbers.

### Fast Fourier Transform Examples

In the first example the `fft` function is called on the sine wave used in the autocorrelation example.

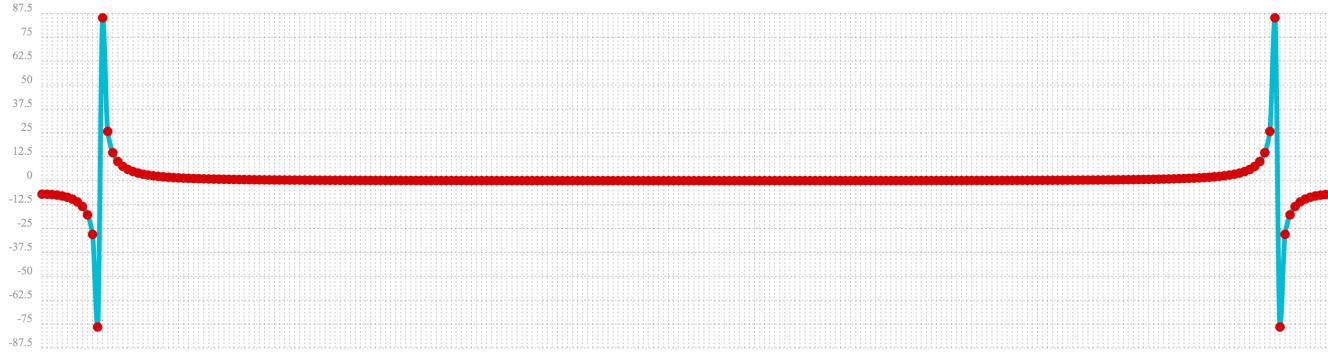
The results of the `fft` function is a matrix. The `rowAt` function is used to return the first row of the matrix which is a vector containing the real values of the `fft` response.

The plot of the real values of the `fft` response is shown below. Notice there are two peaks on opposite sides

of the plot. The plot is actually showing a mirrored response. The right side of the plot is an exact mirror of the left side. This is expected when the fft is run on real rather than complex data.

Also notice that the fft has accumulated significant power in a single peak. This is the power associated with the specific frequency of the sine wave. The vast majority of frequencies in the plot have close to 0 power associated with them. This fft shows a clear signal with very low levels of noise.

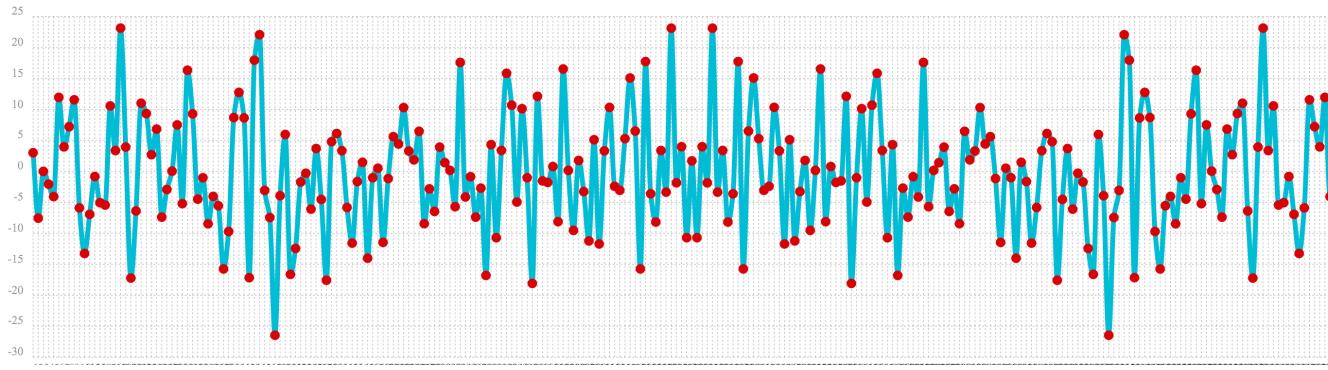
```
let(a=oscillate(1, 0.28, 1.57),
 b=predict(a, sequence(256, 0, 1)),
 c=fft(b),
 d=rowAt(c, 0))
```



In the second example the fft function is called on a vector of random data similar to one used in the autocorrelation example. The plot of the real values of the fft response is shown below.

Notice that in this response there is no clear peak. Instead all frequencies have accumulated a random level of power. This fft shows no clear sign of signal and appears to be noise.

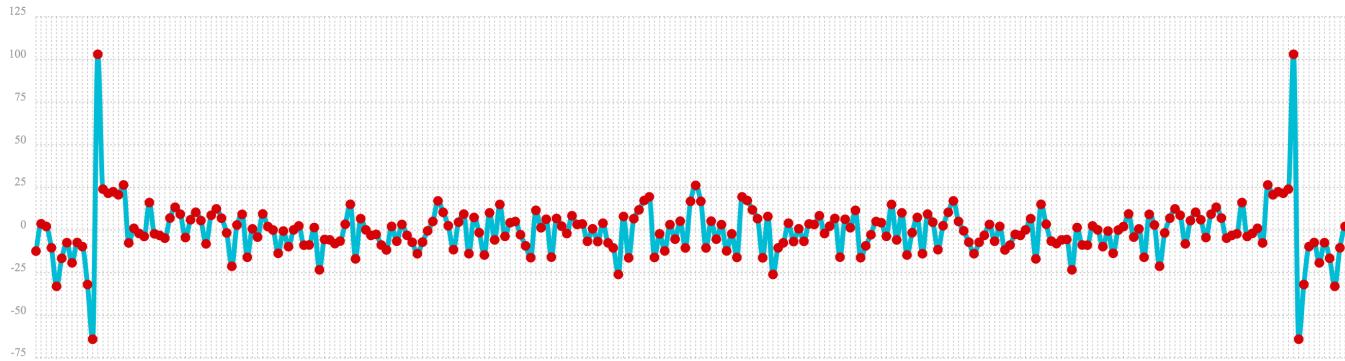
```
let(a=sample(uniformDistribution(-1.5, 1.5), 256),
 b=fft(a),
 c=rowAt(b, 0))
```



In the third example the fft function is called on the same signal hidden within noise that was used for the autocorrelation example. The plot of the real values of the fft response is shown below.

Notice that there are two clear mirrored peaks, at the same locations as the fft of the pure signal. But there is also now considerable noise on the frequencies. The fft has found the signal and but also shows that there is considerable noise along with the signal.

```
let(a=oscillate(1, 0.28, 1.57),
 b=predict(a, sequence(256, 0, 1)),
 c=sample(uniformDistribution(-1.5, 1.5), 256),
 d=ebeAdd(b, c),
 e=fft(d),
 f=rowAt(e, 0))
```



## Machine Learning

This section of the math expressions user guide covers machine learning functions.

### Feature Scaling

Before performing machine learning operations its often necessary to scale the feature vectors so they can be compared at the same scale.

All the scaling function operate on vectors and matrices. When operating on a matrix the rows of the matrix are scaled.

#### Min/Max Scaling

The `minMaxScale` function scales a vector or matrix between a minimum and maximum value. By default it will scale between 0 and 1 if min/max values are not provided.

Below is a simple example of min/max scaling between 0 and 1. Notice that once brought into the same scale the vectors are the same.

```
let(a=array(20, 30, 40, 50),
 b=array(200, 300, 400, 500),
 c=matrix(a, b),
 d=minMaxScale(c))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "d": [
 [
 0,
 0.3333333333333333,
 0.6666666666666666,
 1
],
 [
 0,
 0.3333333333333333,
 0.6666666666666666,
 1
]
],
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
 }
}
```

## Standardization

The `standardize` function scales a vector so that it has a mean of 0 and a standard deviation of 1. Standardization can be used with machine learning algorithms, such as [Support Vector Machine \(SVM\)](#), that perform better when the data has a normal distribution.

```
let(a=array(20, 30, 40, 50),
 b=array(200, 300, 400, 500),
 c=matrix(a, b),
 d=standardize(c))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "d": [
 [
 -1.161895003862225,
 -0.3872983346207417,
 0.3872983346207417,
 1.161895003862225
],
 [
 -1.1618950038622249,
 -0.38729833462074165,
 0.38729833462074165,
 1.1618950038622249
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 17
 }
]
 }
}
```

## Unit Vectors

The `unitize` function scales vectors to a magnitude of 1. A vector with a magnitude of 1 is known as a unit vector. Unit vectors are preferred when the vector math deals with vector direction rather than magnitude.

```
let(a=array(20, 30, 40, 50),
 b=array(200, 300, 400, 500),
 c=matrix(a, b),
 d=unitize(c))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "d": [
 [
 0.2721655269759087,
 0.40824829046386296,
 0.5443310539518174,
 0.6804138174397716
],
 [
 0.2721655269759087,
 0.4082482904638631,
 0.5443310539518174,
 0.6804138174397717
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 6
 }
]
 }
}
```

## Distance and Distance Measures

The distance function computes the distance for two numeric arrays or a distance matrix for the columns of a matrix.

There are five distance measure functions that return a function that performs the actual distance calculation:

- euclidean (default)
- manhattan
- canberra
- earthMovers
- haversineMeters (Geospatial distance measure)

The distance measure functions can be used with all machine learning functions that support distance measures.

Below is an example for computing Euclidean distance for two numeric arrays:

```
let(a=array(20, 30, 40, 50),
 b=array(21, 29, 41, 49),
 c=distance(a, b))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "c": 2
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 0
 }
]
 }
}
```

Below the distance is calculated using **Manhattan** distance.

```
let(a=array(20, 30, 40, 50),
 b=array(21, 29, 41, 49),
 c=distance(a, b, manhattan()))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "c": 4
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 1
 }
]
 }
}
```

Below is an example for computing a distance matrix for columns of a matrix:

```
let(a=array(20, 30, 40),
 b=array(21, 29, 41),
 c=array(31, 40, 50),
 d=matrix(a, b, c),
 c=distance(d))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "e": [
 [
 0,
 15.652475842498529,
 34.07345007480164
],
 [
 15.652475842498529,
 0,
 18.547236990991408
],
 [
 34.07345007480164,
 18.547236990991408,
 0
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 24
 }
]
 }
}
```

## K-Means Clustering

The `kmeans` function performs k-means clustering of the rows of a matrix. Once the clustering has been completed there are a number of useful functions available for examining the clusters and centroids.

The examples below cluster *term vectors*. The section [Text Analysis and Term Vectors](#) offers a full explanation of these features.

### Centroid Features

In the example below the `kmeans` function is used to cluster a result set from the Enron email data-set and then the top features are extracted from the cluster centroids.

```
let(a=select(random(enron, q="body:oil", rows="500", fl="id, body"), ①
 id,
 analyze(body, body_bigram) as terms),
 b=termVectors(a, maxDocFreq=.10, minDocFreq=.05, minTermLength=14, exclude="_,copyright"), ②
 c=kmeans(b, 5), ③
 d=getCentroids(c), ④
 e=topFeatures(d, 5)) ⑤
```

Let's look at what data is assigned to each variable:

- ① **a**: The random function returns a sample of 500 documents from the "enron" collection that match the query "body:oil". The select function selects the id and and annotates each tuple with the analyzed bigram terms from the body field.
- ② **b**: The termVectors function creates a TF-IDF term vector matrix from the tuples stored in variable **a**. Each row in the matrix represents a document. The columns of the matrix are the bigram terms that were attached to each tuple.
- ③ **c**: The kmeans function clusters the rows of the matrix into 5 clusters. The k-means clustering is performed using the Euclidean distance measure.
- ④ **d**: The getCentroids function returns a matrix of cluster centroids. Each row in the matrix is a centroid from one of the 5 clusters. The columns of the matrix are the same bigrams terms of the term vector matrix.
- ⑤ **e**: The topFeatures function returns the column labels for the top 5 features of each centroid in the matrix. This returns the top 5 bigram terms for each centroid.

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "e": [
 [
 "enron enronxgate",
 "north american",
 "energy services",
 "conference call",
 "power generation"
],
 [
 "financial times",
 "chief financial",
 "financial officer",
 "exchange commission",
 "houston chronicle"
],
 [
 "southern california",
 "california edison",
 "public utilities",
 "utilities commission",
 "rate increases"
],
 [
 "rolling blackouts",
 "public utilities",
 "electricity prices",
 "federal energy",
 "price controls"
],
 [
 "california edison",
 "regulatory commission",
 "southern california",
 "federal energy",
 "power generators"
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 982
 }
]
 }
}
```

## Cluster Features

The example below examines the top features of a specific cluster. This example uses the same techniques as the centroids example but the top features are extracted from a cluster rather than the centroids.

```
let(a=select(random(collection3, q="body:oil", rows="500", fl="id, body"),
 id,
 analyze(body, body_bigram) as terms),
 b=termVectors(a, maxDocFreq=.09, minDocFreq=.03, minTermLength=14, exclude="_,copyright"),
 c=kmeans(b, 25),
 d=getCluster(c, 0), ①
 e=topFeatures(d, 4)) ②
```

- ① The getCluster function returns a cluster by its index. Each cluster is a matrix containing term vectors that have been clustered together based on their features.
- ② The topFeatures function is used to extract the top 4 features from each term vector in the cluster.

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "e": [
 [
 "electricity board",
 "maharashtra state",
 "power purchase",
 "state electricity",
 "reserved enron"
],
 [
 "electricity board",
 "maharashtra state",
 "state electricity",
 "purchase agreement",
 "independent power"
],
 [
 "maharashtra state",
 "reserved enron",
 "federal government",
 "state government",
 "dabhol project"
],
 [
 "purchase agreement",
 "power purchase",
 "electricity board",
 "maharashtra state",
 "state government"
]
]
 }
]
 }
}
```

```
],
 [
 "investment grade",
 "portland general",
 "general electric",
 "holding company",
 "transmission lines"
],
 [
 "state government",
 "state electricity",
 "purchase agreement",
 "electricity board",
 "maharashtra state"
],
 [
 "electricity board",
 "state electricity",
 "energy management",
 "maharashtra state",
 "energy markets"
],
 [
 "electricity board",
 "maharashtra state",
 "state electricity",
 "state government",
 "second quarter"
]
],
 {
 "EOF": true,
 "RESPONSE_TIME": 978
 }
]
}
```

## Multi K-Means Clustering

K-means clustering will produce different results depending on the initial placement of the centroids. K-means is fast enough that multiple trials can be performed and the best outcome selected.

The `multiKmeans` function runs the k-means clustering algorithm for a given number of trials and selects the best result based on which trial produces the lowest intra-cluster variance.

The example below is identical to `centroids` example except that it uses `multiKmeans` with 100 trials, rather than a single trial of the `kmeans` function.

```
let(a=select(random(collection3, q="body:oil", rows="500", fl="id, body"),
 id,
 analyze(body, body_bigram) as terms),
 b=termVectors(a, maxDocFreq=.09, minDocFreq=.03, minTermLength=14, exclude="_,copyright"),
 c=multiKmeans(b, 5, 100),
 d=getCentroids(c),
 e=topFeatures(d, 5))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "e": [
 [
 "enron enronxgate",
 "energy trading",
 "energy markets",
 "energy services",
 "unleaded gasoline"
],
 [
 "maharashtra state",
 "electricity board",
 "state electricity",
 "energy trading",
 "chief financial"
],
 [
 "price controls",
 "electricity prices",
 "francisco chronicle",
 "wholesale electricity",
 "power generators"
],
 [
 "southern california",
 "california edison",
 "public utilities",
 "francisco chronicle",
 "utilities commission"
],
 [
 "california edison",
 "power purchases",
 "system operator",
 "term contracts",
 "independent system"
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 1182
 }
]
 }
}
```

## Fuzzy K-Means Clustering

The fuzzyKmeans function is a soft clustering algorithm which allows vectors to be assigned to more than one cluster. The fuzziness parameter is a value between 1 and 2 that determines how fuzzy to make the cluster assignment.

After the clustering has been performed the getMembershipMatrix function can be called on the clustering result to return a matrix describing which clusters each vector belongs to. There is a row in the matrix for each vector that was clustered. There is a column in the matrix for each cluster. The values in the columns are the probability that the vector belonged to the specific cluster.

A simple example will make this more clear. In the example below 300 documents are analyzed and then turned into a term vector matrix. Then the fuzzyKmeans function clusters the term vectors into 12 clusters with a fuzziness factor of 1.25.

```
let(a=select(random(collection3, q="body:oil", rows="300", fl="id, body"),
 id,
 analyze(body, body_bigram) as terms),
 b=termVectors(a, maxDocFreq=.09, minDocFreq=.03, minTermLength=14, exclude="_copyright"),
 c=fuzzyKmeans(b, 12, fuzziness=1.25),
 d=getMembershipMatrix(c), ①
 e=rowAt(d, 0), ②
 f=precision(e, 5)) ③
```

- ① The getMembershipMatrix function is used to return the membership matrix;
- ② and the first row of membership matrix is retrieved with the rowAt function.
- ③ The precision function is then applied to the first row of the matrix to make it easier to read.

This expression returns a single vector representing the cluster membership probabilities for the first term vector. Notice that the term vector has the highest association with the 12<sup>th</sup> cluster, but also has significant associations with the 3<sup>rd</sup>, 5<sup>th</sup>, 6<sup>th</sup> and 7<sup>th</sup> clusters:

```
{
 "result-set": {
 "docs": [
 {
 "f": [
 0,
 0,
 0.178,
 0,
 0.17707,
 0.17775,
 0.16214,
 0,
 0,
 0,
 0,
 0.30504
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 2157
 }
]
 }
}
```

## K-Nearest Neighbor (KNN)

The knn function searches the rows of a matrix for the k-nearest neighbors of a search vector. The knn function returns a matrix of the k-nearest neighbors.

The knn function supports changing of the distance measure by providing one of these distance measure functions as the fourth parameter:

- euclidean (Default)
- manhattan
- canberra
- earthMovers

The example below builds on the clustering examples to demonstrate the knn function.

```
let(a=select(random(collection3, q="body:oil", rows="500", fl="id, body"),
 id,
 analyze(body, body_bigram) as terms),
 b=termVectors(a, maxDocFreq=.09, minDocFreq=.03, minTermLength=14, exclude="_,copyright"),
 c=multiKmeans(b, 5, 100),
 d=getCentroids(c), ①
 e=rowAt(d, 0), ②
 g=knn(b, e, 3), ③
 h=topFeatures(g, 4)) ④
```

- ① In the example, the centroids matrix is set to variable **d**.
- ② The first centroid vector is selected from the matrix with the `rowAt` function.
- ③ Then the `knn` function is used to find the 3 nearest neighbors to the centroid vector in the term vector matrix (variable **b**).
- ④ The `topFeatures` function is used to request the top 4 features of the term vectors in the `knn` matrix.

The `knn` function returns a matrix with the 3 nearest neighbors based on the default distance measure which is euclidean. Finally, the top 4 features of the term vectors in the nearest neighbor matrix are returned:

```
{
 "result-set": {
 "docs": [
 {
 "h": [
 [
 "california power",
 "electricity supply",
 "concerned about",
 "companies like"
],
 [
 "maharashtra state",
 "california power",
 "electricity board",
 "alternative energy"
],
 [
 "electricity board",
 "maharashtra state",
 "state electricity",
 "houston chronicle"
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 1243
 }
]
 }
}
```

## K-Nearest Neighbor Regression

K-nearest neighbor regression is a non-linear, multi-variate regression method. Knn regression is a lazy learning technique which means it does not fit a model to the training set in advance. Instead the entire training set of observations and outcomes are held in memory and predictions are made by averaging the outcomes of the k-nearest neighbors.

The knnRegress function prepares the training set for use with the predict function.

Below is an example of the knnRegress function. In this example 10,000 random samples are taken, each containing the variables filesize\_d, service\_d and response\_d. The pairs of filesize\_d and service\_d will be used to predict the value of response\_d.

```
let(samples=random(collection1, q="*:*", rows="10000", fl="filesize_d, service_d, response_d"),
 filesizes=col(samples, filesize_d),
 serviceLevels=col(samples, service_d),
 outcomes=col(samples, response_d),
 observations=transpose(matrix(filesizes, serviceLevels)),
 lazyModel=knnRegress(observations, outcomes , 5))
```

This expression returns the following response. Notice that knnRegress returns a tuple describing the regression inputs:

```
{
 "result-set": {
 "docs": [
 {
 "lazyModel": {
 "features": 2,
 "robust": false,
 "distance": "EuclideanDistance",
 "observations": 10000,
 "scale": false,
 "k": 5
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 170
 }
]
 }
 }
}
```

## Prediction and Residuals

The output of knnRegress can be used with the predict function like other regression models.

In the example below the predict function is used to predict results for the original training data. The sumSq of the residuals is then calculated.

```
let(samples=random(collection1, q="*:*", rows="10000", fl="filesize_d, service_d, response_d"),
 filesizes=col(samples, filesize_d),
 serviceLevels=col(samples, service_d),
 outcomes=col(samples, response_d),
 observations=transpose(matrix(filesizes, serviceLevels)),
 lazyModel=knnRegress(observations, outcomes , 5),
 predictions=predict(lazyModel, observations),
 residuals=ebeSubtract(outcomes, predictions),
 sumSqErr=sumSq(residuals))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "sumSqErr": 1920290.1204126712
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 3796
 }
]
 }
}
```

## Setting Feature Scaling

If the features in the observation matrix are not in the same scale then the larger features will carry more weight in the distance calculation than the smaller features. This can greatly impact the accuracy of the prediction. The `knnRegress` function has a `scale` parameter which can be set to `true` to automatically scale the features in the same range.

The example below shows `knnRegress` with feature scaling turned on.

Notice that when feature scaling is turned on the `sumSqErr` in the output is much lower. This shows how much more accurate the predictions are when feature scaling is turned on in this particular example. This is because the `filesize_d` feature is significantly larger than the `service_d` feature.

```
let(samples=random(collection1, q="*:*", rows="10000", fl="filesize_d, service_d, response_d"),
 filesizes=col(samples, filesize_d),
 serviceLevels=col(samples, service_d),
 outcomes=col(samples, response_d),
 observations=transpose(matrix(filesizes, serviceLevels)),
 lazyModel=knnRegress(observations, outcomes , 5, scale=true),
 predictions=predict(lazyModel, observations),
 residuals=ebeSubtract(outcomes, predictions),
 sumSqErr=sumSq(residuals))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "sumSqErr": 4076.794951120683
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 3790
 }
]
 }
}
```

## Setting Robust Regression

The default prediction approach is to take the mean of the outcomes of the k-nearest neighbors. If the outcomes contain outliers the mean value can be skewed. Setting the robust parameter to true will take the median outcome of the k-nearest neighbors. This provides a regression prediction that is robust to outliers.

## Setting the Distance Measure

The distance measure can be changed for the k-nearest neighbor search by adding a distance measure function to the knnRegress parameters. Below is an example using manhattan distance.

```
let(samples=random(collection1, q="*:*", rows="10000", fl="filesize_d, service_d, response_d"),
 filesizes=col(samples, filesize_d),
 serviceLevels=col(samples, service_d),
 outcomes=col(samples, response_d),
 observations=transpose(matrix(filesizes, serviceLevels)),
 lazyModel=knnRegress(observations, outcomes, 5, manhattan(), scale=true),
 predictions=predict(lazyModel, observations),
 residuals=ebeSubtract(outcomes, predictions),
 sumSqErr=sumSq(residuals))
```

This expression returns the following response:

```
{
 "result-set": {
 "docs": [
 {
 "sumSqErr": 4761.221942288098
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 3571
 }
]
 }
}
```

## Computational Geometry

This section of the math expressions user guide covers computational geometry functions.

### Convex Hull

A convex hull is the smallest convex set of points that encloses a data set. Math expressions has support for computing the convex hull of a 2D data set. Once a convex hull has been calculated, a set of math expression functions can be applied to geometrically describe the convex hull.

The `convexHull` function finds the convex hull of an observation matrix of 2D vectors. Each row of the matrix is a 2D observation.

In the example below a convex hull is calculated for a randomly generated set of 100 2D observations.

Then the following functions are called on the convex hull:

- `getBaryCenter`: Returns the 2D point that is the bary center of the convex hull.
- `getArea`: Returns the area of the convex hull.
- `getBoundarySize`: Returns the boundary size of the convex hull.
- `getVertices`: Returns a set of 2D points that are the vertices of the convex hull.

```
let(echo="baryCenter, area, boundarySize, vertices",
 x=sample(normalDistribution(0, 20), 100),
 y=sample(normalDistribution(0, 10), 100),
 observations=transpose(matrix(x,y)),
 chull=convexHull(observations),
 baryCenter=getBaryCenter(chull),
 area=getArea(chull),
 boundarySize=getBoundarySize(chull),
 vertices=getVertices(chull))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "baryCenter": [
 -3.0969292101230343,
 1.2160948182691975
],
 "area": 3477.480599967595,
 "boundarySize": 267.52419019533664,
 "vertices": [
 [
 -66.17632818958485,
 -8.394931552315256
],
 [
 -47.556667594765216,
 -16.940434013651263
],
 [
 -33.13582183446102,
 -17.30914425443977
],
 [
 -9.97459859015698,
 -17.795012801599654
],
 [
 27.7705917246824,
 -14.487224686587767
],
 [
 54.689432954170236,
 -1.3333371984299605
],
 [
 35.97568654458672,
 23.054169251772556
],
 [
 -15.539456215337585,
 19.811330468093704
],
 [
 -17.05125031092752,
 19.53581741341663
],
 [
 -35.92010024412891,
 15.126430698395572
]
]
 }
]
 }
}
```

```
],
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 3
 }
]
}
```

## Enclosing Disk

The enclosingDisk function finds the smallest enclosing circle the encloses a 2D data set. Once an enclosing disk has been calculated, a set of math expression functions can be applied to geometrically describe the enclosing disk.

In the example below an enclosing disk is calculated for a randomly generated set of 1000 2D observations.

Then the following functions are called on the enclosing disk:

- getCenter: Returns the 2D point that is the center of the disk.
- getRadius: Returns the radius of the disk.
- getSupportPoints: Returns the support points of the disk.

```
let(echo="center, radius, support",
 x=sample(normalDistribution(0, 20), 1000),
 y=sample(normalDistribution(0, 20), 1000),
 observations=transpose(matrix(x,y)),
 disk=enclosingDisk(observations),
 center=getCenter(disk),
 radius=getRadius(disk),
 support=getSupportPoints(disk))
```

When this expression is sent to the /stream handler it responds with:

```
{
 "result-set": {
 "docs": [
 {
 "center": [
 -6.668825009733749,
 -2.9825450908240025
],
 "radius": 72.66109546907208,
 "support": [
 [
 20.350992271739464,
 64.46791279377014
],
 [
 33.02079953093981,
 57.880978456420365
],
 [
 -44.7273247899923,
 -64.87911518353323
]
]
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 8
 }
]
 }
}
```

# Graph Traversal

Graph traversal with streaming expressions uses the nodes function to perform a breadth-first graph traversal.

The nodes function can be combined with the scoreNodes function to provide recommendations. nodes can also be combined with the wider streaming expression library to perform complex operations on gathered node sets.

nodes traversals are distributed within a SolrCloud collection and can span collections.

nodes is designed for use cases that involve zooming into a neighborhood in the graph and performing precise traversals to gather node sets and aggregations. In these types of use cases nodes will often provide sub-second performance. Some sample use cases are provided later in the document.



This document assumes a basic understanding of graph terminology and streaming expressions. You can begin exploring graph traversal concepts with this [Wikipedia article](#). More details about streaming expressions are available in this Guide, in the section [Streaming Expressions](#).

## Basic Syntax

We'll start with the most basic syntax and slowly build up more complexity. The most basic syntax for nodes is:

```
nodes(emails,
 walk="johndoe@apache.org->from",
 gather="to")
```

Let's break down this simple expression.

The first parameter, emails, is the collection being traversed. The second parameter, walk, maps a hard-coded node ID ("johndoe@apache.org") to a field in the index (from). This will return all the **edges** in the index that have johndoe@apache.org in the from field.

The gather parameter tells the function to gather the values in the `to` field. The values that are gathered are the node IDs emitted by the function.

In the example above the nodes emitted will be all of the people that "[johndoe@apache.org](#)" has emailed.

The walk parameter also accepts a list of root node IDs:

```
nodes(emails,
 walk="johndoe@apache.org, janesmith@apache.org->from",
 gather="to")
```

The nodes function above finds all the edges with "[johndoe@apache.org](#)" or "[janesmith@apache.org](#)" in the from field and gathers the to field.

Like all [Streaming Expressions](#), you can execute a nodes expression by sending it to the /stream handler. For example:

```
curl --data-urlencode 'expr=nodes(emails,
 walk="johndoe@apache.org, janesmith@apache.org->from",
 gather="to")' http://localhost:8983/solr/emails/stream
```

The output of this expression would look like this:

```
{
 "result-set": {
 "docs": [
 {
 "node": "slist@campbell.com",
 "collection": "emails",
 "field": "to",
 "level": 1
 },
 {
 "node": "catherine.pernot@enron.com",
 "collection": "emails",
 "field": "to",
 "level": 1
 },
 {
 "node": "airam.arteaga@enron.com",
 "collection": "emails",
 "field": "to",
 "level": 1
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 44
 }
]
 }
}
```

All of the tuples returned have the node field. The node field contains the node IDs gathered by the function. The collection, field, and level of the traversal are also included in the output.

Notice that the level is "1" for each tuple in the example. The root nodes are level 0 (in the example above, the root nodes are "[johndoe@apache.org, janesmith@apache.org](#)") By default the nodes function emits only the *\*leaf nodes\** of the traversal, which is the outer-most node set. To emit the root nodes you can specify the scatter parameter:

```
nodes(emails,
 walk="johndoe@apache.org->from",
 gather="to",
 scatter="branches, leaves")
```

The scatter parameter controls whether to emit the *branches* with the *leaves*. The root nodes are considered "branches" because they are not the outer-most level of the traversal.

When scattering both branches and leaves the output would like this:

```
{
 "result-set": {
 "docs": [
 {
 "node": "johndoe@apache.org",
 "collection": "emails",
 "field": "node",
 "level": 0
 },
 {
 "node": "slist@campbell.com",
 "collection": "emails",
 "field": "to",
 "level": 1
 },
 {
 "node": "catherine.pernot@enron.com",
 "collection": "emails",
 "field": "to",
 "level": 1
 },
 {
 "node": "airam.arteaga@enron.com",
 "collection": "emails",
 "field": "to",
 "level": 1
 },
 {
 "EOF": true,
 "RESPONSE_TIME": 44
 }
]
 }
}
```

Now the level 0 root node is included in the output.

## Aggregations

nodes also supports aggregations. For example:

```
nodes(emails,
 walk="johndoe@apache.org, janesmith@apache.org->from",
 gather="to",
 count(*))
```

The expression above finds the edges with "johndoe@apache.org" or "janesmith@apache.org" in the from field and gathers the values from the to field. It also aggregates the count for each node ID gathered.

A gathered node could have a count of 2 if both "johndoe@apache.org" and "janesmith@apache.org" have emailed the same person. Node sets contain a unique set of nodes, so the same person won't appear twice in the node set, but the count will reflect that it appeared twice during the traversal.

Edges are uniqued as part of the traversal so the count will **not** reflect the number of times "johndoe@apache.org" emailed the same person. For example, personA might have emailed personB 100 times. These edges would get uniqued and only be counted once. But if person personC also emailed personB this would increment the count for personB.

The aggregation functions supported are count(\*), sum(field), min(field), max(field), and avg(field). The fields being aggregated should be present in the edges collected during the traversal. Later examples (below) will show aggregations can be a powerful tool for providing recommendations and limiting the scope of traversals.

## Nesting nodes Functions

The nodes function can be nested to traverse deeper into the graph. For example:

```
nodes(emails,
 nodes(emails,
 walk="johndoe@apache.org->from",
 gather="to"),
 walk="node->from",
 gather="to")
```

In the example above the outer nodes function operates on the node set collected from the inner nodes function.

Notice that the inner nodes function behaves exactly as the examples already discussed. But the walk parameter of the outer nodes function behaves differently.

In the outer nodes function the walk parameter works with tuples coming from an internal streaming expression. In this scenario the walk parameter maps the node field to the from field. Remember that the node IDs collected from the inner nodes expression are placed in the node field.

Put more simply, the inner expression gathers all the people that "johndoe@apache.org" has emailed. We can call this group the "friends of johndoe@apache.org". The outer expression gathers all the people that

the "friends of johndoe@apache.org" have emailed. This is a basic friends-of-friends traversal.

This construct of nesting nodes functions is the basic technique for doing a controlled traversal through the graph.

## Cycle Detection

The nodes function performs cycle detection across the entire traversal. This ensures that nodes that have already been visited are not traversed again. Cycle detection is important for both limiting the size of traversals and gathering accurate aggregations. Without cycle detection the size of the traversal could grow exponentially with each hop in the traversal. With cycle detection only new nodes encountered are traversed.

Cycle detection **does not** cross collection boundaries. This is because internally the collection name is part of the node ID. For example the node ID "johndoe@apache.org", is really emails/johndoe@apache.org. When traversing to another collection "johndoe@apache.org" will be traversed.

## Filtering the Traversal

Each level in the traversal can be filtered with a filter query. For example:

```
nodes(emails,
 walk="johndoe@apache.org->from",
 fq="body:(solr rocks)",
 gather="to")
```

In the example above only emails that match the filter query will be included in the traversal. Any Solr query can be included here. So you can do fun things like [geospatial queries](#), apply any of the available [query parsers](#), or even write custom query parsers to limit the traversal.

## Root Streams

Any streaming expression can be used to provide the root nodes for a traversal. For example:

```
nodes(emails,
 search(emails, q="body:(solr rocks)", fl="to", sort="score desc", rows="20")
 walk="to->from",
 gather="to")
```

The example above provides the root nodes through a search expression. You can also provide arbitrarily complex, nested streaming expressions with joins, etc., to specify the root nodes.

Notice that the walk parameter maps a field from the tuples generated by the inner stream. In this case it maps the to field from the inner stream to the from field.

## Skipping High Frequency Nodes

It's often desirable to skip traversing high frequency nodes in the graph. This is similar in nature to a search

term stop list. The best way to describe this is through an example use case.

Let's say that you want to recommend content for a user based on a collaborative filter. Below is one approach for a simple collaborative filter:

1. Find all content userA has read.
2. Find users whose reading list is closest to userA. These are users with similar tastes as userA.
3. Recommend content based on what the users in step 2 have read, that userA has not yet read.

Look closely at step 2. In large graphs, step 2 can lead to a very large traversal. This is because userA may have viewed content that has been viewed by millions of other people. We may want to skip these high frequency nodes for two reasons:

1. A large traversal that visit millions of unique nodes is slow and takes a lot of memory because cycle detection is tracked in memory.
2. High frequency nodes are also not useful in determining users with similar tastes. The content that fewer people have viewed provides a more precise recommendation.

The nodes function has the maxDocFreq parameter to allow for filtering out high frequency nodes. The sample code below shows steps 1 and 2 of the recommendation:

```
nodes(logs,
 search(logs, q="userID:user1", fl="articleID", sort="articleID asc", fq="action:view",
 qt="/export"),
 walk="articleID->articleID",
 gather="userID",
 fq="action:view",
 maxDocFreq="10000",
 count(*)))
```

In the example above, the inner search expression searches the logs collection and returning all the articles viewed by "user1". The outer nodes expression takes all the articles emitted from the inner search expression and finds all the records in the logs collection for those articles. It then gathers and aggregates the users that have read the articles. The maxDocFreq parameter limits the articles returned to those that appear in no more than 10,000 log records (per shard). This guards against returning articles that have been viewed by millions of users.

## Tracking the Traversal

By default the nodes function only tracks enough information to do cycle detection. This provides enough information to output the nodes and aggregations in the graph.

For some use cases, such as graph visualization, we also need to output the edges. Setting trackTraversal="true" tells nodes to track the connections between nodes, so the edges can be constructed. When trackTraversal is enabled a new ancestors property will appear with each node. The ancestors property contains a list of node IDs that pointed to the node.

Below is a sample nodes expression with trackTraversal set to true:

```
nodes(emails,
 nodes(emails,
 walk="johndoe@apache.org->from",
 gather="to",
 trackTraversal="true"),
 walk="node->from",
 trackTraversal="true",
 gather="to")
```

## Cross-Collection Traversals

Nested nodes functions can operate on different SolrCloud collections. This allows traversals to "walk" from one collection to another to gather nodes. Cycle detection does not cross collection boundaries, so nodes collected in one collection will be traversed in a different collection. This was done deliberately to support cross-collection traversals. Note that the output from a cross-collection traversal will likely contain duplicate nodes with different collection attributes.

Below is a sample nodes expression that traverses from the "emails" collection to the "logs" collection:

```
nodes(logs,
 nodes(emails,
 search(emails, q="body:(solr rocks)", fl="from", sort="score desc", rows="20")
 walk="from->from",
 gather="to",
 scatter="leaves, branches"),
 walk="node->user",
 fq="action:edit",
 gather="contentID")
```

The example above finds all people who sent emails with a body that contains "solr rocks". It then finds all the people these people have emailed. Then it traverses to the logs collection and gathers all the content IDs that these people have edited.

## Combining nodes With Other Streaming Expressions

The nodes function can act as both a stream source and a stream decorator. The connection with the wider stream expression library provides tremendous power and flexibility when performing graph traversals. Here is an example of using the streaming expression library to intersect two friend networks:

```
intersect(on="node",
 sort(by="node asc",
 nodes(emails,
 nodes(emails,
 walk="johndoe@apache.org->from",
 gather="to"),
 walk="node->from",
 gather="to",
 scatter="branches,leaves")),
 sort(by="node asc",
 nodes(emails,
 nodes(emails,
 walk="janedoe@apache.org->from",
 gather="to"),
 walk="node->from",
 gather="to",
 scatter="branches,leaves")))
```

The example above gathers two separate friend networks, one rooted with "johndoe@apache.org" and another rooted with "janedoe@apache.org". The friend networks are then sorted by the node field, and intersected. The resulting node set will be the intersection of the two friend networks.

## Sample Use Cases for Graph Traversal

### Calculate Market Basket Co-occurrence

It is often useful to know which products are most frequently purchased with a particular product. This example uses a simple market basket table (indexed in Solr) to store past shopping baskets. The schema for the table is very simple with each row containing a basketID and a productID. This can be seen as a graph with each row in the table representing an edge. And it can be traversed very quickly to calculate basket co-occurrence, even when the graph contains billions of edges.

Here is the sample syntax:

```
top(n="5",
 sort="count(*) desc",
 nodes(baskets,
 random(baskets, q="productID:ABC", fl="basketID", rows="500"),
 walk="basketID->basketID",
 fq="-productID:ABC",
 gather="productID",
 count(*)))
```

Let's break down exactly what this traversal is doing.

1. The first expression evaluated is the inner random expression, which returns 500 random basketIDs, from the baskets collection, that have the productID "ABC". The random expression is very useful for recommendations because it limits the traversal to a fixed set of baskets, and because it adds the

element of surprise into the recommendation. Using the random function you can provide fast sample sets from very large graphs.

2. The outer nodes expression finds all the records in the baskets collection for the basketIDs generated in step 1. It also filters out productID "ABC" so it doesn't show up in the results. It then gathers and counts the productID's across these baskets.
3. The outer top expression ranks the productIDs emitted in step 2 by the count and selects the top 5.

In a nutshell this expression finds the products that most frequently co-occur with product "ABC" in past shopping baskets.

## Using the scoreNodes Function to Make a Recommendation

This use case builds on the market basket example [above](#) that calculates which products co-occur most frequently with productID:ABC. The ranked co-occurrence counts provide candidates for a recommendation. The scoreNodes function can be used to score the candidates to find the best recommendation.

Before diving into the syntax of the scoreNodes function it's useful to understand why the raw co-occurrence counts may not produce the best recommendation. The reason is that raw co-occurrence counts favor items that occur frequently across all baskets. A better recommendation would find the product that has the most significant relationship with productID ABC. The scoreNodes function uses a term frequency-inverse document frequency (TF-IDF) algorithm to find the most significant relationship.

### How scoreNodes Works

The scoreNodes function assigns a score to each node emitted by the nodes expression. By default the scoreNodes function uses the count(\*) aggregation, which is the co-occurrence count, as the TF value. The IDF value for each node is fetched from the collection where the node was gathered. Each node is then scored using the TF\*IDF formula, which provides a boost to nodes with a lower frequency across all market baskets.

Combining the co-occurrence count with the IDF provides a score that shows how important the relationship is between productID ABC and the recommendation candidates.

The scoreNodes function adds the score to each node in the nodeScore field.

### Example scoreNodes Syntax

```
top(n="1",
 sort="nodeScore desc",
 scoreNodes(top(n="50",
 sort="count(*) desc",
 nodes(baskets,
 random(baskets, q="productID:ABC", fl="basketID", rows="500"),
 walk="basketID->basketID",
 fq="-productID:ABC",
 gather="productID",
 count()))))
```

This example builds on the earlier example "Calculate market basket co-occurrence".

1. Notice that the inner-most top function is taking the top 50 products that co-occur most frequently with productID ABC. This provides 50 candidate recommendations.
2. The scoreNodes function then assigns a score to the candidates based on the TF\*IDF of each node.
3. The outer top expression selects the highest scoring node. This is the recommendation.

## Recommend Content Based on Collaborative Filter

In this example we'll recommend content for a user based on a collaborative filter. This recommendation is made using log records that contain the userID and articleID and the action performed. In this scenario each log record can be viewed as an edge in a graph. The userID and articleID are the nodes and the action is an edge property used to filter the traversal.

Here is the sample syntax:

```
top(n="5",
 sort="count(*) desc",
 nodes(logs,
 top(n="30",
 sort="count(*) desc",
 nodes(logs,
 search(logs, q="userID:user1", fl="articleID", sort="articleID asc",
 fq="action:read", qt="/export"),
 walk="articleID->articleID",
 gather="userID",
 fq="action:read",
 maxDocFreq="10000",
 count()))),
 walk="node->userID",
 gather="articleID",
 fq="action:read",
 count())))
 count(*)))
```

Let's break down the expression above step-by-step.

1. The first expression evaluated is the inner search expression. This expression searches the logs collection for all records matching "user1". This is the user we are making the recommendation for.

There is a filter applied to pull back only records where the "action:read". It returns the articleID for each record found. In other words, this expression returns all the articles "user1" has read.

2. The inner nodes expression operates over the articleIDs returned from step 1. It takes each articleID found and searches them against the articleID field.

Note that it skips high frequency nodes using the maxDocFreq parameter to filter out articles that appear over 10,000 times in the logs. It gathers userIDs and aggregates the counts for each user. This step finds the users that have read the same articles that "user1" has read and counts how many of the same articles they have read.

3. The inner top expression ranks the users emitted from step 2. It will emit the top 30 users who have the most overlap with user1's reading list.

4. The outer nodes expression gathers the reading list for the users emitted from step 3. It counts the articleIDs that are gathered.

Any article selected in step 1 (user1 reading list), will not appear in this step due to cycle detection. So this step returns the articles read by the users with the most similar readings habits to "user1" that "user1" has not read yet. It also counts the number of times each article has been read across this user group.

5. The outer top expression takes the top articles emitted from step 4. This is the recommendation.

## Protein Pathway Traversal

In recent years, scientists have become increasingly able to rationally design drugs that target the mutated proteins, called oncogenes, responsible for some cancers. Proteins typically act through long chains of chemical interactions between multiple proteins, called pathways, and, while the oncogene in the pathway may not have a corresponding drug, another protein in the pathway may. Graph traversal on a protein collection that records protein interactions and drugs may yield possible candidates. (Thanks to Lewis Geer of the NCBI, for providing this example).

The example below illustrates a protein pathway traversal:

```
nodes(proteins,
 nodes(proteins,
 walk="NRAS->name",
 gather="interacts"),
 walk="node->name",
 gather="drug")
```

Let's break down exactly what this traversal is doing.

1. The inner nodes expression traverses in the proteins collection. It finds all the edges in the graph where the name of the protein is "NRAS". Then it gathers the proteins in the interacts field. This gathers all the proteins that "NRAS" interacts with.
2. The outer nodes expression also works with the proteins collection. It gathers all the drugs that correspond to proteins emitted from step 1.
3. Using this stepwise approach you can gather the drugs along the pathway of interactions any number of steps away from the root protein.

## Exporting GraphML to Support Graph Visualization

In the examples above, the nodes expression was sent to Solr's /stream handler like any other streaming expression. This approach outputs the nodes in the same JSON tuple format as other streaming expressions so that it can be treated like any other streaming expression. You can use the /stream handler when you need to operate directly on the tuples, such as in the recommendation use cases above.

There are other graph traversal use cases that involve graph visualization. Solr supports these use cases with the introduction of the /graph request handler, which takes a nodes expression and outputs the results in GraphML.

GraphML is an XML format supported by graph visualization tools such as [Gephi](#), which is a sophisticated open source tool for statistically analyzing and visualizing graphs. Using a nodes expression, parts of a larger graph can be exported in GraphML and then imported into tools like Gephi.

There are a few things to keep mind when exporting a graph in GraphML:

1. The /graph handler can export both the nodes and edges in the graph. By default, it only exports the nodes. To export the edges you must set trackTraversal="true" in the nodes expression.
2. The /graph handler currently accepts an arbitrarily complex streaming expression which includes a nodes expression. If the streaming expression doesn't include a nodes expression, the /graph handler will not properly output GraphML.
3. The /graph handler currently accepts a single arbitrarily complex, nested nodes expression per request. This means you cannot send in a streaming expression that joins or intersects the node sets from multiple nodes expressions. The /graph handler does support any level of nesting within a single nodes expression. The /stream handler does support joining and intersecting node sets, but the /graph handler currently does not.

## Sample GraphML Request

```
curl --data-urlencode 'expr=nodes(enron_emails,
 nodes(enron_emails,
 walk="kayne.coulter@enron.com->from",
 trackTraversal="true",
 gather="to"),
 walk="node->from",
 scatter="leaves,branches",
 trackTraversal="true",
 gather="to")' http://localhost:8983/solr/enron_emails/graph
```

## Sample GraphML Output

```
<graphml xmlns="http://graphml.graphdrawing.org/xmlns"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
 http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">
<graph id="G" edgedefault="directed">
 <node id="kayne.coulter@enron.com">
 <data key="field">node</data>
 <data key="level">0</data>
 <data key="count(*)">0.0</data>
 </node>
 <node id="don.baughman@enron.com">
 <data key="field">to</data>
 <data key="level">1</data>
 <data key="count(*)">1.0</data>
 </node>
 <edge id="1" source="kayne.coulter@enron.com" target="don.baughman@enron.com"/>
 <node id="john.kinser@enron.com">
 <data key="field">to</data>
 <data key="level">1</data>
 <data key="count(*)">1.0</data>
 </node>
 <edge id="2" source="kayne.coulter@enron.com" target="john.kinser@enron.com"/>
 <node id="jay.wills@enron.com">
 <data key="field">to</data>
 <data key="level">1</data>
 <data key="count(*)">1.0</data>
 </node>
 <edge id="3" source="kayne.coulter@enron.com" target="jay.wills@enron.com"/>
</graph></graphml>
```

# SolrCloud

Apache Solr includes the ability to set up a cluster of Solr servers that combines fault tolerance and high availability. Called **SolrCloud**, these capabilities provide distributed indexing and search capabilities, supporting the following features:

- Central configuration for the entire cluster
- Automatic load balancing and fail-over for queries
- ZooKeeper integration for cluster coordination and configuration.

SolrCloud is flexible distributed search and indexing, without a master node to allocate nodes, shards and replicas. Instead, Solr uses ZooKeeper to manage these locations, depending on configuration files and schemas. Queries and updates can be sent to any server. Solr will use the information in the ZooKeeper database to figure out which servers need to handle the request.

In this section, we'll cover everything you need to know about using Solr in SolrCloud mode. We've split up the details into the following topics:

- [Getting Started with SolrCloud](#)
- [How SolrCloud Works](#)
  - [Shards and Indexing Data in SolrCloud](#)
  - [Distributed Requests](#)
- [SolrCloud Resilience](#)
  - [SolrCloud Recoveries and Write Tolerance](#)
  - [SolrCloud Query Routing And Read Tolerance](#)
- [SolrCloud Configuration and Parameters](#)
  - [Setting Up an External ZooKeeper Ensemble](#)
  - [Using ZooKeeper to Manage Configuration Files](#)
  - [ZooKeeper Access Control](#)
  - [Collections API](#)
  - [Parameter Reference](#)
  - [Command Line Utilities](#)
  - [SolrCloud with Legacy Configuration Files](#)
  - [ConfigSets API](#)
- [Rule-based Replica Placement](#)
- [Cross Data Center Replication \(CDCR\)](#)
- [SolrCloud Autoscaling](#)
- [Colocating collections together](#)

# Getting Started with SolrCloud

SolrCloud is designed to provide a highly available, fault tolerant environment for distributing your indexed content and query requests across multiple servers.

It's a system in which data is organized into multiple pieces, or shards, that can be hosted on multiple machines, with replicas providing redundancy for both scalability and fault tolerance, and a ZooKeeper server that helps manage the overall structure so that both indexing and search requests can be routed properly.

This section explains SolrCloud and its inner workings in detail, but before you dive in, it's best to have an idea of what it is you're trying to accomplish.

This page provides a simple tutorial to start Solr in SolrCloud mode, so you can begin to get a sense for how shards interact with each other during indexing and when serving queries. To that end, we'll use simple examples of configuring SolrCloud on a single machine, which is obviously not a real production environment, which would include several servers or virtual machines. In a real production environment, you'll also use the real machine names instead of "localhost" which we've used here.

In this section you will learn how to start a SolrCloud cluster using startup scripts and a specific configset.



This tutorial assumes that you're already familiar with the basics of using Solr. If you need a refresher, please see the [Getting Started section](#) to get a grounding in Solr concepts. If you load documents as part of that exercise, you should start over with a fresh Solr installation for these SolrCloud tutorials.

## SolrCloud Example

### Interactive Startup

The bin/solr script makes it easy to get started with SolrCloud as it walks you through the process of launching Solr nodes in SolrCloud mode and adding a collection. To get started, simply do:

```
bin/solr -e cloud
```

This starts an interactive session to walk you through the steps of setting up a simple SolrCloud cluster with embedded ZooKeeper.

The script starts by asking you how many Solr nodes you want to run in your local cluster, with the default being 2.

```
Welcome to the SolrCloud example!
```

```
This interactive session will help you launch a SolrCloud cluster on your local workstation.
To begin, how many Solr nodes would you like to run in your local cluster? (specify 1-4 nodes)
[2]
```

The script supports starting up to 4 nodes, but we recommend using the default of 2 when starting out. These nodes will each exist on a single machine, but will use different ports to mimic operation on different servers.

Next, the script will prompt you for the port to bind each of the Solr nodes to, such as:

```
Please enter the port for node1 [8983]
```

Choose any available port for each node; the default for the first node is 8983 and 7574 for the second node. The script will start each node in order and show you the command it uses to start the server, such as:

```
solr start -cloud -s example/cloud/node1/solr -p 8983
```

The first node will also start an embedded ZooKeeper server bound to port 9983. The Solr home for the first node is in example/cloud/node1/solr as indicated by the -s option.

After starting up all nodes in the cluster, the script prompts you for the name of the collection to create:

```
Please provide a name for your new collection: [gettingstarted]
```

The suggested default is "gettingstarted" but you might want to choose a name more appropriate for your specific search application.

Next, the script prompts you for the number of shards to distribute the collection across. [Sharding](#) is covered in more detail later on, so if you're unsure, we suggest using the default of 2 so that you can see how a collection is distributed across multiple nodes in a SolrCloud cluster.

Next, the script will prompt you for the number of replicas to create for each shard. [Replication](#) is covered in more detail later in the guide, so if you're unsure, then use the default of 2 so that you can see how replication is handled in SolrCloud.

Lastly, the script will prompt you for the name of a configuration directory for your collection. You can choose **\_default**, or **sample\_techproducts\_configs**. The configuration directories are pulled from server/solr/configsets/ so you can review them beforehand if you wish. The **\_default** configuration is useful when you're still designing a schema for your documents and need some flexibility as you experiment with Solr, since it has schemaless functionality. However, after creating your collection, the schemaless functionality can be disabled in order to lock down the schema (so that documents indexed after doing so will not alter the schema) or to configure the schema by yourself. This can be done as follows (assuming your collection name is mycollection):

## V1 API

```
curl http://host:8983/solr/mycollection/config -d '{"set-user-property": {"update.autoCreateFields":"false"}}'
```

## V2 API SolrCloud

```
curl http://host:8983/api/collections/mycollection/config -d '{"set-user-property": {"update.autoCreateFields":"false"}}'
```

At this point, you should have a new collection created in your local SolrCloud cluster. To verify this, you can run the status command:

```
bin/solr status
```

If you encounter any errors during this process, check the Solr log files in example/cloud/node1/logs and example/cloud/node2/logs.

You can see how your collection is deployed across the cluster by visiting the cloud panel in the Solr Admin UI: <http://localhost:8983/solr/#/~cloud>. Solr also provides a way to perform basic diagnostics for a collection using the healthcheck command:

```
bin/solr healthcheck -c gettingstarted
```

The healthcheck command gathers basic information about each replica in a collection, such as number of docs, current status (active, down, etc.), and address (where the replica lives in the cluster).

Documents can now be added to SolrCloud using the [Post Tool](#).

To stop Solr in SolrCloud mode, you would use the bin/solr script and issue the stop command, as in:

```
bin/solr stop -all
```

## Starting with -noprompt

You can also get SolrCloud started with all the defaults instead of the interactive session using the following command:

```
bin/solr -e cloud -noprompt
```

## Restarting Nodes

You can restart your SolrCloud nodes using the bin/solr script. For instance, to restart node1 running on port 8983 (with an embedded ZooKeeper server), you would do:

```
bin/solr restart -c -p 8983 -s example/cloud/node1/solr
```

To restart node2 running on port 7574, you can do:

```
bin/solr restart -c -p 7574 -z localhost:9983 -s example/cloud/node2/solr
```

Notice that you need to specify the ZooKeeper address (-z localhost:9983) when starting node2 so that it can join the cluster with node1.

## Adding a Node to a Cluster

Adding a node to an existing cluster is a bit advanced and involves a little more understanding of Solr. Once you startup a SolrCloud cluster using the startup scripts, you can add a new node to it by:

```
mkdir <solr.home for new Solr node>
cp <existing solr.xml path> <new solr.home>
bin/solr start -cloud -s solr.home/solr -p <port num> -z <zk hosts string>
```

Notice that the above requires you to create a Solr home directory. You either need to copy solr.xml to the solr\_home directory, or keep it centrally in ZooKeeper /solr.xml.

Example (with directory structure) that adds a node to an example started with "bin/solr -e cloud":

```
mkdir -p example/cloud/node3/solr
cp server/solr/solr.xml example/cloud/node3/solr
bin/solr start -cloud -s example/cloud/node3/solr -p 8987 -z localhost:9983
```

The previous command will start another Solr node on port 8987 with Solr home set to example/cloud/node3/solr. The new node will write its log files to example/cloud/node3/logs.

Once you're comfortable with how the SolrCloud example works, we recommend using the process described in [Taking Solr to Production](#) for setting up SolrCloud nodes in production.

# How SolrCloud Works

The following sections cover provide general information about how various SolrCloud features work. To understand these features, it's important to first understand a few key concepts that relate to SolrCloud.

- [Shards and Indexing Data in SolrCloud](#)
- [Distributed Requests](#)
- [Standard and Routed Aliases](#)

If you are already familiar with SolrCloud concepts and basic functionality, you can skip to the section covering [SolrCloud Configuration and Parameters](#).

## Key SolrCloud Concepts

A SolrCloud cluster consists of some "logical" concepts layered on top of some "physical" concepts.

### Logical Concepts

- A Cluster can host multiple Collections of Solr Documents.
- A collection can be partitioned into multiple Shards, which contain a subset of the Documents in the Collection.
- The number of Shards that a Collection has determines:
  - The theoretical limit to the number of Documents that Collection can reasonably contain.
  - The amount of parallelization that is possible for an individual search request.

### Physical Concepts

- A Cluster is made up of one or more Solr Nodes, which are running instances of the Solr server process.
- Each Node can host multiple Cores.
- Each Core in a Cluster is a physical Replica for a logical Shard.
- Every Replica uses the same configuration specified for the Collection that it is a part of.
- The number of Replicas that each Shard has determines:
  - The level of redundancy built into the Collection and how fault tolerant the Cluster can be in the event that some Nodes become unavailable.
  - The theoretical limit in the number concurrent search requests that can be processed under heavy load.

## Shards and Indexing Data in SolrCloud

When your collection is too large for one node, you can break it up and store it in sections by creating multiple **shards**.

A Shard is a logical partition of the collection, containing a subset of documents from the collection, such that every document in a collection is contained in exactly one Shard. Which shard contains each document

in a collection depends on the overall "Sharding" strategy for that collection.

For example, you might have a collection where the "country" field of each document determines which shard it is part of, so documents from the same country are co-located. A different collection might simply use a "hash" on the uniqueKey of each document to determine its Shard.

Before SolrCloud, Solr supported Distributed Search, which allowed one query to be executed across multiple shards, so the query was executed against the entire Solr index and no documents would be missed from the search results. So splitting an index across shards is not exclusively a SolrCloud concept. There were, however, several problems with the distributed approach that necessitated improvement with SolrCloud:

1. Splitting an index into shards was somewhat manual.
2. There was no support for distributed indexing, which meant that you needed to explicitly send documents to a specific shard; Solr couldn't figure out on its own what shards to send documents to.
3. There was no load balancing or failover, so if you got a high number of queries, you needed to figure out where to send them and if one shard died it was just gone.

SolrCloud addresses those limitations. There is support for distributing both the index process and the queries automatically, and ZooKeeper provides failover and load balancing. Additionally, every shard can have multiple replicas for additional robustness.

## Leaders and Replicas

In SolrCloud there are no masters or slaves. Instead, every shard consists of at least one physical **replica**, exactly one of which is a **leader**. Leaders are automatically elected, initially on a first-come-first-served basis, and then based on the ZooKeeper process described at [http://zookeeper.apache.org/doc/trunk/recipes.html#sc\\_leaderElection..](http://zookeeper.apache.org/doc/trunk/recipes.html#sc_leaderElection..)

If a leader goes down, one of the other replicas is automatically elected as the new leader.

When a document is sent to a Solr node for indexing, the system first determines which Shard that document belongs to, and then which node is currently hosting the leader for that shard. The document is then forwarded to the current leader for indexing, and the leader forwards the update to all of the other replicas.

### Types of Replicas

By default, all replicas are eligible to become leaders if their leader goes down. However, this comes at a cost: if all replicas could become a leader at any time, every replica must be in sync with its leader at all times. New documents added to the leader must be routed to the replicas, and each replica must do a commit. If a replica goes down, or is temporarily unavailable, and then rejoins the cluster, recovery may be slow if it has missed a large number of updates.

These issues are not a problem for most users. However, some use cases would perform better if the replicas behaved a bit more like the former model, either by not syncing in real-time or by not being eligible to become leaders at all.

Solr accomplishes this by allowing you to set the replica type when creating a new collection or when adding a replica. The available types are:

- **NRT:** This is the default. A NRT replica (NRT = NearRealTime) maintains a transaction log and writes new documents to its indexes locally. Any replica of this type is eligible to become a leader. Traditionally, this was the only type supported by Solr.
- **TLOG:** This type of replica maintains a transaction log but does not index document changes locally. This type helps speed up indexing since no commits need to occur in the replicas. When this type of replica needs to update its index, it does so by replicating the index from the leader. This type of replica is also eligible to become a shard leader; it would do so by first processing its transaction log. If it does become a leader, it will behave the same as if it was a NRT type of replica.
- **PULL:** This type of replica does not maintain a transaction log nor index document changes locally. It only replicates the index from the shard leader. It is not eligible to become a shard leader and doesn't participate in shard leader election at all.

If you do not specify the type of replica when it is created, it will be NRT type.

## Combining Replica Types in a Cluster

There are three combinations of replica types that are recommended:

- All NRT replicas
- All TLOG replicas
- TLOG replicas with PULL replicas

### All NRT Replicas

Use this for small to medium clusters, or even big clusters where the update (index) throughput is not too high. NRT is the only type of replica that supports soft-commits, so also use this combination when NearRealTime is needed.

### All TLOG Replicas

Use this combination if NearRealTime is not needed and the number of replicas per shard is high, but you still want all replicas to be able to handle update requests.

### TLOG replicas plus PULL replicas

Use this combination if NearRealTime is not needed, the number of replicas per shard is high, and you want to increase availability of search queries over document updates even if that means temporarily serving outdated results.

### Other Combinations of Replica Types

Other combinations of replica types are not recommended. If more than one replica in the shard is writing its own index instead of replicating from an NRT replica, a leader election can cause all replicas of the shard to become out of sync with the leader, and all would have to replicate the full index.

### Recovery with PULL Replicas

If a PULL replica goes down or leaves the cluster, there are a few scenarios to consider.

If the PULL replica cannot sync to the leader because the leader is down, replication would not occur. However, it would continue to serve queries. Once it can connect to the leader again, replication would

resume.

If the PULL replica cannot connect to ZooKeeper, it would be removed from the cluster and queries would not be routed to it from the cluster.

If the PULL replica dies or is unreachable for any other reason, it won't be query-able. When it rejoins the cluster, it would replicate from the leader and when that is complete, it would be ready to serve queries again.

## Queries with Preferred Replica Types

By default all replicas serve queries. See the section [shards.preference Parameter](#) for details on how to indicate preferred replica types for queries.

## Document Routing

Solr offers the ability to specify the router implementation used by a collection by specifying the router.name parameter when [creating your collection](#).

If you use the compositeId router (the default), you can send documents with a prefix in the document ID which will be used to calculate the hash Solr uses to determine the shard a document is sent to for indexing. The prefix can be anything you'd like it to be (it doesn't have to be the shard name, for example), but it must be consistent so Solr behaves consistently.

For example, if you want to co-locate documents for a customer, you could use the customer name or ID as the prefix. If your customer is "IBM", for example, with a document with the ID "12345", you would insert the prefix into the document id field: "IBM!12345". The exclamation mark (!) is critical here, as it distinguishes the prefix used to determine which shard to direct the document to.

Then at query time, you include the prefix(es) into your query with the \_route\_ parameter (i.e., q=solr&\_route\_=IBM!) to direct queries to specific shards. In some situations, this may improve query performance because it overcomes network latency when querying all the shards.

The compositeId router supports prefixes containing up to 2 levels of routing. For example: a prefix routing first by region, then by customer: "USA!IBM!12345"

Another use case could be if the customer "IBM" has a lot of documents and you want to spread it across multiple shards. The syntax for such a use case would be: shard\_key/num!document\_id where the /num is the number of bits from the shard key to use in the composite hash.

So IBM/3!12345 will take 3 bits from the shard key and 29 bits from the unique doc id, spreading the tenant over 1/8th of the shards in the collection. Likewise if the num value was 2 it would spread the documents across 1/4th the number of shards. At query time, you include the prefix(es) along with the number of bits into your query with the \_route\_ parameter (i.e., q=solr&\_route\_=IBM/3!) to direct queries to specific shards.

If you do not want to influence how documents are stored, you don't need to specify a prefix in your document ID.

If you created the collection and defined the "implicit" router at the time of creation, you can additionally define a router.field parameter to use a field from each document to identify a shard where the document belongs. If the field specified is missing in the document, however, the document will be rejected.

You could also use the `_route_` parameter to name a specific shard.

## Shard Splitting

When you create a collection in SolrCloud, you decide on the initial number shards to be used. But it can be difficult to know in advance the number of shards that you need, particularly when organizational requirements can change at a moment's notice, and the cost of finding out later that you chose wrong can be high, involving creating new cores and reindexing all of your data.

The ability to split shards is in the Collections API. It currently allows splitting a shard into two pieces. The existing shard is left as-is, so the split action effectively makes two copies of the data as new shards. You can delete the old shard at a later time when you're ready.

More details on how to use shard splitting is in the section on the Collection API's [SPLITSHARD command](#).

## Ignoring Commits from Client Applications in SolrCloud

In most cases, when running in SolrCloud mode, indexing client applications should not send explicit commit requests. Rather, you should configure auto commits with `openSearcher=false` and auto soft-commits to make recent updates visible in search requests. This ensures that auto commits occur on a regular schedule in the cluster.

To enforce a policy where client applications should not send explicit commits, you should update all client applications that index data into SolrCloud. However, that is not always feasible, so Solr provides the `IgnoreCommitOptimizeUpdateProcessorFactory`, which allows you to ignore explicit commits and/or optimize requests from client applications without having refactor your client application code.

To activate this request processor you'll need to add the following to your `solrconfig.xml`:

```
<updateRequestProcessorChain name="ignore-commit-from-client" default="true">
 <processor class="solr.IgnoreCommitOptimizeUpdateProcessorFactory">
 <int name="statusCode">200</int>
 </processor>
 <processor class="solr.LogUpdateProcessorFactory" />
 <processor class="solr.DistributedUpdateProcessorFactory" />
 <processor class="solr.RunUpdateProcessorFactory" />
</updateRequestProcessorChain>
```

As shown in the example above, the processor will return 200 to the client but will ignore the commit / optimize request. Notice that you need to wire-in the implicit processors needed by SolrCloud as well, since this custom chain is taking the place of the default chain.

In the following example, the processor will raise an exception with a 403 code with a customized error message:

```
<updateRequestProcessorChain name="ignore-commit-from-client" default="true">
 <processor class="solr.IgnoreCommitOptimizeUpdateProcessorFactory">
 <int name="statusCode">403</int>
 <str name="responseMessage">Thou shall not issue a commit!</str>
 </processor>
 <processor class="solr.LogUpdateProcessorFactory" />
 <processor class="solr.DistributedUpdateProcessorFactory" />
 <processor class="solr.RunUpdateProcessorFactory" />
</updateRequestProcessorChain>
```

Lastly, you can also configure it to just ignore optimize and let commits pass thru by doing:

```
<updateRequestProcessorChain name="ignore-optimize-only-from-client-403">
 <processor class="solr.IgnoreCommitOptimizeUpdateProcessorFactory">
 <str name="responseMessage">Thou shall not issue an optimize, but commits are OK!</str>
 <bool name="ignoreOptimizeOnly">true</bool>
 </processor>
 <processor class="solr.RunUpdateProcessorFactory" />
</updateRequestProcessorChain>
```

## Distributed Requests

When a Solr node receives a search request, the request is routed behind the scenes to a replica of a shard that is part of the collection being searched.

The chosen replica acts as an aggregator: it creates internal requests to randomly chosen replicas of every shard in the collection, coordinates the responses, issues any subsequent internal requests as needed (for example, to refine facets values, or request additional stored fields), and constructs the final response for the client.

### Limiting Which Shards are Queried

While one of the advantages of using SolrCloud is the ability to query very large collections distributed across various shards, in some cases you may have configured Solr so you know [you are only interested in results from a specific subset of shards](#). You have the option of searching over all of your data or just parts of it.

A query across all shards for a collection is simply a query that does not define a shards parameter:

```
http://localhost:8983/solr/gettingstarted/select?q=*:*
```

If you want to search just one shard, use the shards parameter to specify the shard by its logical ID, as in:

```
http://localhost:8983/solr/gettingstarted/select?q=*:*&shards=shard1
```

If you want to search a group of shards, you can specify each shard separated by a comma in one request:

```
http://localhost:8983/solr/gettingstarted/select?q=*&shards=shard1,shard2
```

In both of the above examples, while only the specific shards are queried, any random replica of the shard will get the request.

Alternatively, you can specify a list of replicas you wish to use in place of a shard IDs by separating the replica IDs with commas:

```
http://localhost:8983/solr/gettingstarted/select?q=*&shards=localhost:7574/solr/gettingstarted,
localhost:8983/solr/gettingstarted
```

Or you can specify a list of replicas to choose from for a single shard (for load balancing purposes) by using the pipe symbol (|) between different replica IDs:

```
http://localhost:8983/solr/gettingstarted/select?q=*&shards=localhost:7574/solr/gettingstarted|
localhost:7500/solr/gettingstarted
```

Finally, you can specify a list of shards (separated by commas) each defined by a list of replicas (seperated by pipes).

In the following example, 2 shards are queried, the first being a random replica from shard1, the second being a random replica from the explicit pipe delimited list:

```
http://localhost:8983/solr/gettingstarted/select?q=*&shards=shard1,localhost:7574/solr/gettings
tarted|localhost:7500/solr/gettingstarted
```

## Configuring the ShardHandlerFactory

For finer-grained control, you can directly configure and tune aspects of the concurrency and thread-pooling used within distributed search in Solr. The default configuration favors throughput over latency.

This is done by defining a shardHandler in the configuration for your search handler.

To add a shardHandler to the standard search handler, provide a configuration in `solrconfig.xml`, as in this example:

```
<requestHandler name="/select" class="solr.SearchHandler">
 <!-- other params go here -->
 <shardHandler class="HttpShardHandlerFactory">
 <int name="socketTimeOut">1000</int>
 <int name="connTimeOut">5000</int>
 </shardHandler>
</requestHandler>
```

`HttpShardHandlerFactory` is the only `ShardHandlerFactory` implementation included out of the box with Solr. It accepts the following parameters:

**socketTimeout**

The amount of time in ms that a socket is allowed to wait. The default is 0, where the operating system's default will be used.

**connTimeout**

The amount of time in ms that is accepted for binding / connecting a socket. The default is 0, where the operating system's default will be used.

**maxConnectionsPerHost**

The maximum number of concurrent connections that is made to each individual shard in a distributed search. The default is 100000.

**corePoolSize**

The retained lowest limit on the number of threads used in coordinating distributed search. The default is 0.

**maximumPoolSize**

The maximum number of threads used for coordinating distributed search. The default is Integer.MAX\_VALUE.

**maxThreadIdleTime**

The amount of time in seconds to wait for before threads are scaled back in response to a reduction in load. The default is 5.

**sizeOfQueue**

If specified, the thread pool will use a backing queue instead of a direct handoff buffer. High throughput systems will want to configure this to be a direct hand off (with -1). Systems that desire better latency will want to configure a reasonable size of queue to handle variations in requests. The default is -1.

**fairnessPolicy**

Chooses the JVM specifics dealing with fair policy queuing, if enabled distributed searches will be handled in a First in First out fashion at a cost to throughput. If disabled throughput will be favored over latency. The default is false.

**shardsWhitelist**

If specified, this lists limits what nodes can be requested in the shards request parameter.

In SolrCloud mode this whitelist is automatically configured to include all live nodes in the cluster.

In standalone mode the whitelist defaults to empty (sharding not allowed).

If you need to disable this feature for backwards compatibility, you can set the system property solr.disable.shardsWhitelist=true. The value of this parameter is a comma separated list of the nodes that will be whitelisted, i.e.: 10.0.0.1:8983/solr,10.0.0.1:8984/solr.



In SolrCloud mode, if at least one node is included in the whitelist, then the live\_nodes will no longer be used as source for the list. This means that if you need to do a cross-cluster request using the shards parameter in SolrCloud mode (in addition to regular within-cluster requests), you'll need to add all nodes (local cluster + remote nodes) to the whitelist.

## Configuring statsCache (Distributed IDF)

Document and term statistics are needed in order to calculate relevancy. Solr provides four implementations out of the box when it comes to document stats calculation:

- LocalStatsCache: This only uses local term and document statistics to compute relevance. In cases with uniform term distribution across shards, this works reasonably well. This option is the default if no <statsCache> is configured.
- ExactStatsCache: This implementation uses global values (across the collection) for document frequency.
- ExactSharedStatsCache: This is exactly like the exact stats cache in its functionality but the global stats are reused for subsequent requests with the same terms.
- LRUStatsCache: This implementation uses an LRU cache to hold global stats, which are shared between requests.

The implementation can be selected by setting <statsCache> in solrconfig.xml. For example, the following line makes Solr use the ExactStatsCache implementation:

```
<statsCache class="org.apache.solr.search.stats.ExactStatsCache"/>
```

## Avoiding Distributed Deadlock

Each shard serves top-level query requests and then makes sub-requests to all of the other shards. Care should be taken to ensure that the max number of threads serving HTTP requests is greater than the possible number of requests from both top-level clients and other shards. If this is not the case, the configuration may result in a distributed deadlock.

For example, a deadlock might occur in the case of two shards, each with just a single thread to service HTTP requests. Both threads could receive a top-level request concurrently, and make sub-requests to each other. Because there are no more remaining threads to service requests, the incoming requests will be blocked until the other pending requests are finished, but they will not finish since they are waiting for the sub-requests. By ensuring that Solr is configured to handle a sufficient number of threads, you can avoid deadlock situations like this.

## preferLocalShards Parameter

Deprecated, use shards.preference=replica.location:local instead (see below).

## shards.preference Parameter

Solr allows you to pass an optional string parameter named shards.preference to indicate that a distributed query should sort the available replicas in the given order of precedence within each shard.

The syntax is: shards.preference=property:value. The order of the properties and the values are significant: the first one is the primary sort, the second is secondary, etc.



shards.preference only works for distributed queries, i.e., queries targeting multiple shards. Single shard scenarios are not supported.

The properties that can be specified are as follows:

#### replica.type

One or more replica types that are preferred. Any combination of PULL, TLOG and NRT is allowed.

#### replica.location

One or more replica locations that are preferred.

A location starts with <http://hostname:port>. Matching is done for the given string as a prefix, so it's possible to e.g., leave out the port.

A special value local may be used to denote any local replica running on the same Solr instance as the one handling the query. This is useful when a query requests many fields or large fields to be returned per document because it avoids moving large amounts of data over the network when it is available locally. In addition, this feature can be useful for minimizing the impact of a problematic replica with degraded performance, as it reduces the likelihood that the degraded replica will be hit by other healthy replicas.

The value of replica.location:local diminishes as the number of shards (that have no locally-available replicas) in a collection increases because the query controller will have to direct the query to non-local replicas for most of the shards.

In other words, this feature is mostly useful for optimizing queries directed towards collections with a small number of shards and many replicas.

Also, this option should only be used if you are load balancing requests across all nodes that host replicas for the collection you are querying, as Solr's CloudSolrClient will do. If not load-balancing, this feature can introduce a hotspot in the cluster since queries won't be evenly distributed across the cluster.

Examples:

- Prefer PULL replicas: shards.preference=replica.type:PULL
- Prefer PULL replicas, or TLOG replicas if PULL replicas not available: shards.preference=replica.type:PULL,replica.type:TLOG
- Prefer any local replicas: shards.preference=replica.location:local
- Prefer any replicas on a host called "server1" with "server2" as the secondary option: shards.preference=replica.location:http://server1,replica.location:http://server2
- Prefer PULL replicas if available, otherwise TLOG replicas, and local ones among those: shards.preference=replica.type:PULL,replica.type:TLOG,replica.location:local
- Prefer local replicas, and among them PULL replicas when available TLOG otherwise: shards.preference=replica.location:local,replica.type:PULL,replica.type:TLOG

Note that if you provide the settings in a query string, they need to be properly URL-encoded.

## Aliases

SolrCloud has the ability to query one or more collections via an alternative name. These alternative names for collections are known as aliases, and are useful when you want to:

1. Atomically switch to using a newly (re)indexed collection with zero down time (by re-defining the alias)

2. Insulate the client programming versus changes in collection names
3. Issue a single query against several collections with identical schemas

There are two types of aliases: standard aliases and routed aliases. Within routed aliases, there are two types: category-routed aliases and time-routed aliases. These types are discussed in this section.

It's possible to send collection update commands to aliases, but only to those that either resolve to a single collection or those that define the routing between multiple collections ([Routed Aliases](#)). In other cases update commands are rejected with an error since there is no logic by which to distribute documents among the multiple collections.

## Standard Aliases

Standard aliases are created and updated using the [CREATEALIAS](#) command.

The current list of collections that are members of an alias can be verified via the [CLUSTERSTATUS](#) command.

The full definition of all aliases including metadata about that alias (in the case of routed aliases, see below) can be verified via the [LISTALIASES](#) command.

Alternatively this information is available by checking `/aliases.json` in ZooKeeper with either the native ZooKeeper client or in the [tree page](#) of the cloud menu in the admin UI.

Aliases may be deleted via the [DELETEALIAS](#) command. When deleting an alias, underlying collections are **unaffected**.



Any alias (standard or routed) that references multiple collections may complicate relevancy. By default, SolrCloud scores documents on a per-shard basis.

+ With multiple collections in an alias this is always a problem, so if you have a use case for which BM25 or TF/IDF relevancy is important you will want to turn on one of the [ExactStatsCache](#) implementations.

+ However, for analytical use cases where results are sorted on numeric, date, or alphanumeric field values, rather than relevancy calculations, this is not a problem.

## Routed Aliases

To address the update limitations associated with standard aliases and provide additional useful features, the concept of routed aliases has been developed. There are presently two types of routed alias: time routed and category routed. These are described in detail below, but share some common behavior.

When processing an update for a routed alias, Solr initializes its [UpdateRequestProcessor](#) chain as usual, but when [DistributedUpdateProcessor](#) (DUP) initializes, it detects that the update targets a routed alias and injects [RoutedAliasUpdateProcessor](#) (RAUP) in front of itself. RAUP, in coordination with the Overseer, is the main part of a routed alias, and must immediately precede DUP. It is not possible to configure custom chains with other types of UpdateRequestProcessors between RAUP and DUP.

Ideally, as a user of a routed alias, you needn't concern yourself with the particulars of the collection naming pattern since both queries and updates may be done via the alias. When adding data, you should usually direct documents to the alias (e.g., reference the alias name instead of any collection). The Solr server and

CloudSolrClient will direct an update request to the first collection that an alias points to. Once the server receives the data it will perform the necessary routing.



It is possible to update the collections directly, but there is no safeguard against putting data in the incorrect collection if the alias is circumvented in this manner.



It is a bad idea to use "data driven" mode (aka [schemaless-mode](#)) with routed aliases, as duplicate schema mutations might happen concurrently leading to errors.

## Time Routed Aliases

Time Routed Aliases (TRAs) are a SolrCloud feature that manages an alias and a time sequential series of collections.

It automatically creates new collections and (optionally) deletes old ones as it routes documents to the correct collection based on its timestamp. This approach allows for indefinite indexing of data without degradation of performance otherwise experienced due to the continuous growth of a single index.

If you need to store a lot of timestamped data in Solr, such as logs or IoT sensor data, then this feature probably makes more sense than creating one sharded hash-routed collection.

### How It Works

First you create a time routed aliases using the [CREATEALIAS](#) command with the desired router settings. Most of the settings are editable at a later time using the [ALIASPROP](#) command.

The first collection will be created automatically, along with an alias pointing to it. Each underlying Solr "core" in a collection that is a member of a TRA has a special core property referencing the alias. The name of each collection is comprised of the TRA name and the start timestamp (UTC), with trailing zeros and symbols truncated.

The collections list for a TRA is always reverse sorted, and thus the connection path of the request will route to the lead collection. Using CloudSolrClient is preferable as it can reduce the number of underlying physical HTTP requests by one. If you know that a particular set of documents to be delivered is going to a particular older collection then you could direct it there from the client side as an optimization but it's not necessary. CloudSolrClient does not (yet) do this.

RAUP first reads TRA configuration from the alias properties when it is initialized. As it sees each document, it checks for changes to TRA properties, updates its cached configuration if needed, and then determines which collection the document belongs to:

- If RAUP needs to send it to a time segment represented by a collection other than the one that the client chose to communicate with, then it will do so using mechanisms shared with DUP. Once the document is forwarded to the correct collection (i.e., the correct TRA time segment), it skips directly to DUP on the target collection and continues normally, potentially being routed again to the correct shard & replica within the target collection.
- If it belongs in the current collection (which is usually the case if processing events as they occur), the document passes through to DUP. DUP does its normal collection-level processing that may involve routing the document to another shard & replica.
- If the timestamp on the document is more recent than the most recent TRA segment, then a new

collection needs to be added at the front of the TRA. RAUP will create this collection, add it to the alias, and then forward the document to the collection it just created. This can happen recursively if more than one collection needs to be created.

Each time a new collection is added, the oldest collections in the TRA are examined for possible deletion, if that has been configured. All this happens synchronously, potentially adding seconds to the update request and indexing latency.

If `router.preemptiveCreateMath` is configured and if the document arrives within this window then it will occur asynchronously. See [Time Routed Alias Parameters](#) for more information.

Any other type of update like a commit or delete is routed by RAUP to all collections. Generally speaking, this is not a performance concern. When Solr receives a delete or commit wherein nothing is deleted or nothing needs to be committed, then it's pretty cheap.

### Limitations & Assumptions

- Only **time** routed aliases are supported. If you instead have some other sequential number, you could fake it as a time (e.g., convert to a timestamp assuming some epoch and increment).

The smallest possible interval is one second. No other routing scheme is supported, although this feature was developed with considerations that it could be extended/improved to other schemes.

- The underlying collections form a contiguous sequence without gaps. This will not be suitable when there are large gaps in the underlying data, as Solr will insist that there be a collection for each increment. This is due in part to Solr calculating the end time of each interval collection based on the timestamp of the next collection, since it is otherwise not stored in any way.
- Avoid sending updates to the oldest collection if you have also configured that old collections should be automatically deleted. It could lead to exceptions bubbling back to the indexing client.

### Category Routed Aliases

Category Routed Aliases (CRAs) are a feature to manage aliases and a set of dependent collections based on the value of a single field.

CRAs automatically create new collections but because the partitioning is on categorical information rather than continuous numerically based values there's no logic for automatic deletion. This approach allows for simplified indexing of data that must be segregated into collections for cluster management or security reasons.

### How It Works

First you create a category routed alias using the [CREATEALIAS](#) command with the desired router settings. Most of the settings are editable at a later time using the [ALIASPROP](#) command.

The alias will be created with a special place-holder collection which will always be named `myAlias__CRA__NEW_CATEGORY_ROUTED_ALIAS_WAITING_FOR_DATA__TEMP`. The first document indexed into the CRA will create a second collection named `myAlias__CRA__foo` (for a routed field value of `foo`). The second document indexed will cause the temporary place holder collection to be deleted. Thereafter collections will be created whenever a new value for the field is encountered.



To guard against runaway collection creation options for limiting the total number of categories, and for rejecting values that don't match, a regular expression parameter is provided (see [Category Routed Alias Parameters](#) for details).

- + Note that by providing very large or very permissive values for these options you are accepting the risk that garbled data could potentially create thousands of collections and bring your cluster to a grinding halt.

Field values (and thus the collection names) are case sensitive.

As elsewhere in Solr, manipulation and cleaning of the data is expected to be done by external processes before data is sent to Solr, with one exception. Throughout Solr there are limitations on the allowable characters in collection names. Any characters other than ASCII alphanumeric characters (A-Za-z0-9), hyphen (-) or underscore (\_) are replaced with an underscore when calculating the collection name for a category. For a CRA named myAlias the following table shows how collection names would be calculated:

Value	CRA Collection Name
foo	myAlias_CRA_foo
Foo	myAlias_CRA_Foo
foo bar	myAlias_CRA_foo_bar
FOÓB&R	myAlias_CRA_FO_B_R
□□□□□	myAlias_CRA_____
foo_CRA_bar	<b>Causes 400 Bad Request</b>
<null>	<b>Causes 400 Bad Request</b>

Since collection creation can take upwards of 1-3 seconds, systems inserting data in a CRA should be constructed to handle such pauses whenever a new collection is created. Unlike time routed aliases, there is no way to predict the next value so such pauses are unavoidable.

There is no automated means of removing a category. If a category needs to be removed from a CRA the following procedure is recommended:

1. Ensure that no documents with the value corresponding to the category to be removed will be sent either by stopping indexing or by fixing the incoming data stream
2. Modify the alias definition in ZooKeeper, removing the collection corresponding to the category.
3. Delete the collection corresponding to the category. Note that if the collection is not removed from the alias first, this step will fail.

### Limitations & Assumptions

- CRAs are presently unsuitable for non-English data values due to the limits on collection names. This can be worked around by duplicating the route value to a **url safe** Base64-encoded field and routing on that value instead.
- The check for the CRA infix is independent of the regular expression validation and occurs after the name of the collection to be created has been calculated. It may not be avoided and is necessary to support future features.

## Improvement Possibilities

Routed aliases are a relatively new feature of SolrCloud that can be expected to be improved. Some *potential* areas for improvement that *are not implemented yet* are:

- **TRAs:** Searches with time filters should only go to applicable collections.
- **TRAs:** Ways to automatically optimize (or reduce the resources of) older collections that aren't expected to receive more updates, and might have less search demand.
- **CRAs:** Intrinsic support for non-English text via Base64 encoding.
- **CRAs:** Supply an initial list of values for cases where these are known before hand to reduce pauses during indexing.
- CloudSolrClient could route documents to the correct collection based on the route value instead always picking the latest/first.
- Presently only updates are routed and queries are distributed to all collections in the alias, but future features might enable routing of the query to the single appropriate collection based on a special parameter or perhaps a filter on the routed field.
- Collections might be constrained by their size instead of or in addition to time or category value. This might be implemented as another type of routed alias, or possibly as an option on the existing routed aliases
- Compatibility with CDCR.
- Option for deletion of aliases that also deletes the underlying collections in one step. Routed Aliases may quickly create more collections than expected during initial testing. Removing them after such events is overly tedious.

As always, patches and pull requests are welcome!

## Collection Commands and Aliases

Starting with version 8.1 SolrCloud supports using alias names in collection commands where normally a collection name is expected. This works only when the following criteria are satisfied:

- an alias must not refer to more than one collection
- an alias must not refer to a [Routed Alias](#) (see below)

If all criteria are satisfied then the command will resolve alias names and operate on the collections the aliases refer to as if it was invoked with the collection names instead. Otherwise the command will not be executed and an exception will be thrown.

# SolrCloud Resilience

In this section, we'll cover how does Solr handle reads and writes when all the nodes in the cluster are not healthy.

The following sections cover these topics:

- [SolrCloud Recoveries and Write Tolerance](#)
- [SolrCloud Query Routing And Read Tolerance](#)

## SolrCloud Recoveries and Write Tolerance

SolrCloud is highly available and fault tolerant in reads and writes.

### Write Side Fault Tolerance

SolrCloud is designed to replicate documents to ensure redundancy for your data, and enable you to send update requests to any node in the cluster. That node will determine if it hosts the leader for the appropriate shard, and if not it will forward the request to the leader, which will then forward it to all existing replicas, using versioning to make sure every replica has the most up-to-date version. If the leader goes down, another replica can take its place. This architecture enables you to be certain that your data can be recovered in the event of a disaster, even if you are using [Near Real Time Searching](#).

#### Recovery

A Transaction Log is created for each node so that every change to content or organization is noted. The log is used to determine which content in the node should be included in a replica. When a new replica is created, it refers to the Leader and the Transaction Log to know which content to include. If it fails, it retries.

Since the Transaction Log consists of a record of updates, it allows for more robust indexing because it includes redoing the uncommitted updates if indexing is interrupted.

If a leader goes down, it may have sent requests to some replicas and not others. So when a new potential leader is identified, it runs a sync process against the other replicas. If this is successful, everything should be consistent, the leader registers as active, and normal actions proceed. If a replica is too far out of sync, the system asks for a full replication/replay-based recovery.

If an update fails because cores are reloading schemas and some have finished but others have not, the leader tells the nodes that the update failed and starts the recovery procedure.

#### Achieved Replication Factor

When using a replication factor greater than one, an update request may succeed on the shard leader but fail on one or more of the replicas. For instance, consider a collection with one shard and a replication factor of three. In this case, you have a shard leader and two additional replicas. If an update request succeeds on the leader but fails on both replicas, for whatever reason, the update request is still considered successful from the perspective of the client. The replicas that missed the update will sync with the leader when they recover.

Behind the scenes, this means that Solr has accepted updates that are only on one of the nodes (the current

leader). To make the client aware of this, Solr includes in the response header the "Achieved Replication Factor" (rf). The achieved replication factor is the number of replicas of the shard that actually received the update request (including the leader), in the above example, 1. In the case of multi-shard update requests, the achieved replication factor is the minimum achieved replication factor across all shards.

On the client side, if the achieved replication factor is less than the acceptable level, then the client application can take additional measures to handle the degraded state. For instance, a client application may want to keep a log of which update requests were sent while the state of the collection was degraded and then resend the updates once the problem has been resolved.



In previous version of Solr, the `min_rf` parameter had to be specified to ask Solr for the achieved replication factor. Now it is always included in the response.

## SolrCloud Query Routing And Read Tolerance

SolrCloud is highly available and fault tolerant in reads and writes.

### Read Side Fault Tolerance

In a SolrCloud cluster each individual node load balances read requests across all the replicas in a collection. You still need a load balancer on the 'outside' that talks to the cluster, or you need a smart client which understands how to read and interact with Solr's metadata in ZooKeeper and only requests the ZooKeeper ensemble's address to start discovering to which nodes it should send requests. (Solr provides a smart Java SolrJ client called [CloudSolrClient](#).)

Even if some nodes in the cluster are offline or unreachable, a Solr node will be able to correctly respond to a search request as long as it can communicate with at least one replica of every shard, or one replica of every *relevant* shard if the user limited the search via the shards or `_route_` parameters. The more replicas there are of every shard, the more likely that the Solr cluster will be able to handle search results in the event of node failures.

### Query Parameters for Query Routing

#### zkConnected Parameter

A Solr node will return the results of a search request as long as it can communicate with at least one replica of every shard that it knows about, even if it can *not* communicate with ZooKeeper at the time it receives the request. This is normally the preferred behavior from a fault tolerance standpoint, but may result in stale or incorrect results if there have been major changes to the collection structure that the node has not been informed of via ZooKeeper (i.e., shards may have been added or removed, or split into sub-shards).

A `zkConnected` header is included in every search response indicating if the node that processed the request was connected with ZooKeeper at the time:

## Solr Response with zkConnected

```
{
 "responseHeader": {
 "status": 0,
 "zkConnected": true,
 "QTime": 20,
 "params": {
 "q": "*:*"
 }
 },
 "response": {
 "numFound": 107,
 "start": 0,
 "docs": ["..."]
 }
}
```

To prevent stale or incorrect results in the event that the request-serving node can't communicate with ZooKeeper, set the shards.tolerant parameter to requireZkConnected. This will cause requests to fail rather than setting a zkConnected header to false.

### shards Parameter

By default, SolrCloud will run searches on all shards and combine the results if the shards parameter is not specified. You can specify one or more shard names as the value of the shards parameter to limit the shards that you want to search against.

```
http://localhost:8983/solr/collection1/select?q=*&shards=shard1
http://localhost:8983/solr/collection1/select?q=*&shards=shard2,shard3
```

### shards.tolerant Parameter

In the event that one or more shards queried are unavailable, then Solr's default behavior is to fail the request. However, there are many use-cases where partial results are acceptable and so Solr provides a boolean shards.tolerant parameter (default false). In addition to true and false, shards.tolerant may also be set to requireZkConnected - see below.

If shards.tolerant=true then partial results may be returned. If the returned response does not contain results from all the appropriate shards then the response header contains a special flag called partialResults.

If shards.tolerant=requireZkConnected and the node serving the search request cannot communicate with ZooKeeper, the request will fail, rather than returning potentially stale or incorrect results. This will also cause requests to fail when one or more queried shards are completely unavailable, just like when shards.tolerant=false.

The client can specify 'shards.info' along with the shards.tolerant parameter to retrieve more fine-grained details.

Example response with partialResults flag set to true:

#### Solr Response with partialResults

```
{
 "responseHeader": {
 "status": 0,
 "zkConnected": true,
 "partialResults": true,
 "QTime": 20,
 "params": {
 "q": "*:*"
 }
 },
 "response": {
 "numFound": 77,
 "start": 0,
 "docs": ["..."]
 }
}
```

#### collection Parameter

The collection parameter allows you to specify a collection or a number of collections on which the query should be executed. This allows you to query multiple collections at once and all the feature of Solr which work in a distributed manner can work across collections.

```
http://localhost:8983/solr/collection1/select?collection=collection1,collection2,collection3
```

#### \_route\_ Parameter

The \_route\_ parameter can be used to specify a route key which is used to figure out the corresponding shards. For example, if you have a document with a unique key "user1!123", then specifying the route key as "route=user1!" (notice the trailing '!' character) will route the request to the shard which hosts that user. You can specify multiple route keys separated by comma. This parameter can be leveraged when we have shard data by users. See 'Document Routing' for more information

```
http://localhost:8983/solr/collection1/select?q=*&*_&_route_=user1!
http://localhost:8983/solr/collection1/select?q=*&*_&_route_=user1!,user2!
```

## Distributed Tracing and Debugging

The debug parameter with a value of track can be used to trace the request as well as find timing information for each phase of a distributed request.

## Optimization

## distrib.singlePass Parameter

If set to true, the `distrib.singlePass` parameter changes the distributed search algorithm to fetch all requested stored fields from each shard in the first phase itself. This eliminates the need for making a second request to fetch the stored fields.

This can be faster when requesting a very small number of fields containing small values. However, if large fields are requested or if a lot of fields are requested then the overhead of fetching them over the network from all shards can make the request slower as compared to the normal distributed search path.

Note that this optimization only applies to distributed search. Certain features such as facetting may make additional network requests for refinements, etc.

# SolrCloud Configuration and Parameters

In this section, we'll cover the various configuration options for SolrCloud.

The following sections cover these topics:

- [Setting Up an External ZooKeeper Ensemble](#)
- [Using ZooKeeper to Manage Configuration Files](#)
- [ZooKeeper Access Control](#)
- [Collections API](#)
- [Parameter Reference](#)
- [Command Line Utilities](#)
- [SolrCloud with Legacy Configuration Files](#)
- [ConfigSets API](#)

## Setting Up an External ZooKeeper Ensemble

Although Solr comes bundled with [Apache ZooKeeper](#), you are strongly encouraged to use an external ZooKeeper setup in production.

While using Solr's embedded ZooKeeper instance is fine for getting started, you shouldn't use this in production because it does not provide any failover: if the Solr instance that hosts ZooKeeper shuts down, ZooKeeper is also shut down. Any shards or Solr instances that rely on it will not be able to communicate with it or each other.

The solution to this problem is to set up an external ZooKeeper *ensemble*, which is a number of servers running ZooKeeper that communicate with each other to coordinate the activities of the cluster.

### How Many ZooKeeper Nodes?

The first question to answer is the number of ZooKeeper nodes you will run in your ensemble.

When planning how many ZooKeeper nodes to configure, keep in mind that the main principle for a ZooKeeper ensemble is maintaining a majority of servers to serve requests. This majority is called a *quorum*.

*"For a ZooKeeper service to be active, there must be a majority of non-failing machines that can communicate with each other. To create a deployment that can tolerate the failure of F machines, you should count on deploying  $2xF+1$  machines.*

— ZooKeeper Administrator's Guide, <http://zookeeper.apache.org/doc/r3.4.14/zookeeperAdmin.html>

To properly maintain a quorum, it's highly recommended to have an odd number of ZooKeeper servers in your ensemble, so a majority is maintained.

To explain why, think about this scenario: If you have two ZooKeeper nodes and one goes down, this means only 50% of available servers are available. Since this is not a majority, ZooKeeper will no longer serve requests.

However, if you have three ZooKeeper nodes and one goes down, you have 66% of your servers available and ZooKeeper will continue normally while you repair the one down node. If you have 5 nodes, you could continue operating with two down nodes if necessary.

It's not generally recommended to go above 5 nodes. While it may seem that more nodes provide greater fault-tolerance and availability, in practice it becomes less efficient because of the amount of inter-node coordination that occurs. Unless you have a truly massive Solr cluster (on the scale of 1,000s of nodes), try to stay to 3 as a general rule, or maybe 5 if you have a larger cluster.

More information on ZooKeeper clusters is available from the ZooKeeper documentation at [http://zookeeper.apache.org/doc/r3.4.14/zookeeperAdmin.html#sc\\_zkMulitServerSetup](http://zookeeper.apache.org/doc/r3.4.14/zookeeperAdmin.html#sc_zkMulitServerSetup).

## Download Apache ZooKeeper

The first step in setting up Apache ZooKeeper is, of course, to download the software. It's available from <http://zookeeper.apache.org/releases.html>.

Solr currently uses Apache ZooKeeper v3.4.14.



When using an external ZooKeeper ensemble, you will need to keep your local installation up-to-date with the latest version distributed with Solr. Since it is a stand-alone application in this scenario, it does not get upgraded as part of a standard Solr upgrade.

## ZooKeeper Installation

Installation consists of extracting the files into a specific target directory where you'd like to have ZooKeeper store its internal data. The actual directory itself doesn't matter, as long as you know where it is.

The command to unpack the ZooKeeper package is:

```
tar xvf zookeeper-3.4.14.tar.gz
```

This location is the <ZOOKEEPER\_HOME> for ZooKeeper on this server.

Installing and unpacking ZooKeeper must be repeated on each server where ZooKeeper will be run.

## Configuration for a ZooKeeper Ensemble

After installation, we'll first take a look at the basic configuration for ZooKeeper, then specific parameters for configuring each node to be part of an ensemble.

### Initial Configuration

To configure your ZooKeeper instance, create a file named <ZOOKEEPER\_HOME>/conf/zoo.cfg. A sample configuration file is included in your ZooKeeper installation, as conf/zoo\_sample.cfg. You can edit and rename that file instead of creating it new if you prefer.

The file should have the following information to start:

```
tickTime=2000
dataDir=/var/lib/zookeeper
clientPort=2181
```

The parameters are as follows:

#### tickTime

Part of what ZooKeeper does is determine which servers are up and running at any given time, and the minimum session time out is defined as two "ticks". The tickTime parameter specifies in milliseconds how long each tick should be.

#### dataDir

This is the directory in which ZooKeeper will store data about the cluster. This directory must be empty before starting ZooKeeper for the first time.

#### clientPort

This is the port on which Solr will access ZooKeeper.

These are the basic parameters that need to be in use on each ZooKeeper node, so this file must be copied to or created on each node.

Next we'll customize this configuration to work within an ensemble.

## Ensemble Configuration

To complete configuration for an ensemble, we need to set additional parameters so each node knows who it is in the ensemble and where every other node is.

Each of the examples below assume you are installing ZooKeeper on different servers with different hostnames.

Once complete, your zoo.cfg file might look like this:

```
tickTime=2000
dataDir=/var/lib/zookeeper
clientPort=2181

initLimit=5
syncLimit=2
server.1=zoo1:2888:3888
server.2=zoo2:2888:3888
server.3=zoo3:2888:3888

autopurge.snapRetainCount=3
autopurge.purgeInterval=1
```

We've added these parameters to the three we had already:

#### initLimit

Amount of time, in ticks, to allow followers to connect and sync to a leader. In this case, you have 5 ticks,

each of which is 2000 milliseconds long, so the server will wait as long as 10 seconds to connect and sync with the leader.

#### syncLimit

Amount of time, in ticks, to allow followers to sync with ZooKeeper. If followers fall too far behind a leader, they will be dropped.

#### server.X

These are the server IDs (the X part), hostnames (or IP addresses) and ports for all servers in the ensemble. The IDs differentiate each node of the ensemble, and allow each node to know where each of the other node is located. The ports can be any ports you choose; ZooKeeper's default ports are 2888:3888.

Since we've assigned server IDs to specific hosts/ports, we must also define which server in the list this node is. We do this with a `myid` file stored in the data directory (defined by the `dataDir` parameter). The contents of the `myid` file is only the server ID.

In the case of the configuration example above, you would create the file `/var/lib/zookeeper/1/myid` with the content "1" (without quotes), as in this example:

```
1
```

#### autopurge.snapRetainCount

The number of snapshots and corresponding transaction logs to retain when purging old snapshots and transaction logs.

ZooKeeper automatically keeps a transaction log and writes to it as changes are made. A snapshot of the current state is taken periodically, and this snapshot supersedes transaction logs older than the snapshot. However, ZooKeeper never cleans up either the old snapshots or the old transaction logs; over time they will silently fill available disk space on each server.

To avoid this, set the `autopurge.snapRetainCount` and `autopurge.purgeInterval` parameters to enable an automatic clean up (purge) to occur at regular intervals. The `autopurge.snapRetainCount` parameter will keep the set number of snapshots and transaction logs when a clean up occurs. This parameter can be configured higher than 3, but cannot be set lower than 3.

#### autopurge.purgeInterval

The time in hours between purge tasks. The default for this parameter is 0, so must be set to 1 or higher to enable automatic clean up of snapshots and transaction logs. Setting it as high as 24, for once a day, is acceptable if preferred.

We'll repeat this configuration on each node.

On the second node, update `<ZOOKEEPER_HOME>/conf/zoo.cfg` file so it matches the content on node 1 (particularly the server hosts and ports):

```
tickTime=2000
dataDir=/var/lib/zookeeper
clientPort=2181

initLimit=5
syncLimit=2
server.1=zoo1:2888:3888
server.2=zoo2:2888:3888
server.3=zoo3:2888:3888

autopurge.snapRetainCount=3
autopurge.purgeInterval=1
```

On the second node, create a `myid` file with the contents "2", and put it in the `/var/lib/zookeeper` directory.

2

On the third node, update `<ZOOKEEPER_HOME>/conf/zoo.cfg` file so it matches the content on nodes 1 and 2 (particularly the server hosts and ports):

```
tickTime=2000
dataDir=/var/lib/zookeeper
clientPort=2181

initLimit=5
syncLimit=2
server.1=zoo1:2888:3888
server.2=zoo2:2888:3888
server.3=zoo3:2888:3888

autopurge.snapRetainCount=3
autopurge.purgeInterval=1
```

And create the `myid` file in the `/var/lib/zookeeper` directory:

3

Repeat this for servers 4 and 5 if you are creating a 5-node ensemble (a rare case).

## ZooKeeper Environment Configuration

To ease troubleshooting in case of problems with the ensemble later, it's recommended to run ZooKeeper with logging enabled and with proper JVM garbage collection (GC) settings.

1. Create a file named `zookeeper-env.sh` and put it in the `<ZOOKEEPER_HOME>/conf` directory (the same place you put `zoo.cfg`). This file will need to exist on each server of the ensemble.
2. Add the following settings to the file:

```
ZOO_LOG_DIR="/path/for/log/files"
ZOO_LOG4J_PROP="INFO,ROLLINGFILE"

SERVER_JVMFLAGS="-Xms2048m -Xmx2048m -verbose:gc -XX:+PrintHeapAtGC -XX:+PrintGCDetails
-XX:+PrintGCDateStamps -XX:+PrintGCTimeStamps -XX:+PrintTenuringDistribution
-XX:+PrintGCApplicationStoppedTime -Xloggc:$ZOO_LOG_DIR/zookeeper_gc.log
-XX:+UseGCLogFileRotation -XX:NumberOfGCLogFiles=9 -XX:GCLogFileSize=20M"
```

The property ZOO\_LOG\_DIR defines the location on the server where ZooKeeper will print its logs. ZOO\_LOG4J\_PROP sets the logging level and log appenders.

With SERVER\_JVMFLAGS, we've defined several parameters for garbage collection and logging GC-related events. One of the system parameters is -Xloggc:\$ZOO\_LOG\_DIR/zookeeper\_gc.log, which will put the garbage collection logs in the same directory we've defined for ZooKeeper logs, in a file named zookeeper\_gc.log.

3. Review the default settings in <ZOOKEEPER\_HOME>/conf/log4j.properties, especially the log4j.appenders.ROLLINGFILE.MaxFileSize parameter. This sets the size at which log files will be rolled over, and by default it is 10MB.
4. Copy zookeeper-env.sh and any changes to log4j.properties to each server in the ensemble.



The above instructions are for Linux servers only. The default zkServer.sh script includes support for a zookeeper-env.sh file but the Windows version of the script, zkServer.cmd, does not. To make the same configuration on a Windows server, the changes would need to be made directly in the zkServer.cmd.

At this point, you are ready to start your ZooKeeper ensemble.

## More Information about ZooKeeper

ZooKeeper provides a great deal of power through additional configurations, but delving into them is beyond the scope of Solr's documentation. For more information, see the [ZooKeeper documentation](#).

## Starting and Stopping ZooKeeper

### Start ZooKeeper

To start the ensemble, use the <ZOOKEEPER\_HOME>/bin/zkServer.sh or zkServer.cmd script, as with this command:

#### Linux OS

```
zkServer.sh start
```

#### Windows OS

```
zkServer.cmd start
```

This command needs to be run on each server that will run ZooKeeper.



You should see the ZooKeeper logs in the directory where you defined to store them. However, immediately after startup, you may not see the `zookeeper_gc.log` yet, as it likely will not appear until garbage collection has happened the first time.

## Shut Down ZooKeeper

To shut down ZooKeeper, use the same `zkServer.sh` or `zkServer.cmd` script on each server with the "stop" command:

### Linux OS

```
zkServer.sh stop
```

### Windows OS

```
zkServer.cmd stop
```

## Solr Configuration

When starting Solr, you must provide an address for ZooKeeper or Solr won't know how to use it. This can be done in two ways: by defining the *connect string*, a list of servers where ZooKeeper is running, at every startup on every node of the Solr cluster, or by editing Solr's include file as a permanent system parameter. Both approaches are described below.

When referring to the location of ZooKeeper within Solr, it's best to use the addresses of all the servers in the ensemble. If one happens to be down, Solr will automatically be able to send its request to another server in the list.

### Using a chroot

If your ensemble is or will be shared among other systems besides Solr, you should consider defining application-specific *znodes*, or a hierarchical namespace that will only include Solr's files.

Once you create a *znode* for each application, you add its name, also called a *chroot*, to the end of your connect string whenever you tell Solr where to access ZooKeeper.

Creating a chroot is done with a `bin/solr` command:

```
bin/solr zk mkroot /solr -z zk1:2181,zk2:2181,zk3:2181
```

See the section [Create a znode](#) for more examples of this command.

Once the *znode* is created, it behaves in a similar way to a directory on a filesystem: the data stored by Solr in ZooKeeper is nested beneath the main data directory and won't be mixed with data from another system or process that uses the same ZooKeeper ensemble.

### Using the `-z` Parameter with `bin/solr`

Pointing Solr at the ZooKeeper ensemble you've created is a simple matter of using the `-z` parameter when

using the bin/solr script.

For example, to point the Solr instance to the ZooKeeper you've started on port 2181 on three servers with chroot /solr (see [Using a chroot](#) above), this is what you'd need to do:

```
bin/solr start -e cloud -z zk1:2181,zk2:2181,zk3:2181/solr
```

## Updating Solr's Include Files

If you update Solr's include file (solr.in.sh or solr.in.cmd), which overrides defaults used with bin/solr, you will not have to use the -z parameter with bin/solr commands.

### Linux: solr.in.sh

The section to look for will be commented out:

```
Set the ZooKeeper connection string if using an external ZooKeeper ensemble
e.g. host1:2181,host2:2181/chroot
Leave empty if not using SolrCloud
#ZK_HOST=""
```

Remove the comment marks at the start of the line and enter the ZooKeeper connect string:

```
Set the ZooKeeper connection string if using an external ZooKeeper ensemble
e.g. host1:2181,host2:2181/chroot
Leave empty if not using SolrCloud
ZK_HOST="zk1:2181,zk2:2181,zk3:2181/solr"
```

### Windows: solr.in.cmd

The section to look for will be commented out:

```
REM Set the ZooKeeper connection string if using an external ZooKeeper ensemble
REM e.g. host1:2181,host2:2181/chroot
REM Leave empty if not using SolrCloud
REM set ZK_HOST=
```

Remove the comment marks at the start of the line and enter the ZooKeeper connect string:

```
REM Set the ZooKeeper connection string if using an external ZooKeeper ensemble
REM e.g. host1:2181,host2:2181/chroot
REM Leave empty if not using SolrCloud
set ZK_HOST=zk1:2181,zk2:2181,zk3:2181/solr
```

Now you will not have to enter the connection string when starting Solr.

## Increasing the File Size Limit

ZooKeeper is designed to hold small files, on the order of kilobytes. By default, ZooKeeper's file size limit is 1MB. Attempting to write or read files larger than this will cause errors.

Some Solr features, e.g., text analysis synonyms, LTR, and OpenNLP named entity recognition, require configuration resources that can be larger than the default limit. ZooKeeper can be configured, via Java system property `jute.maxbuffer`, to increase this limit. Note that this configuration, which is required both for ZooKeeper server(s) and for all clients that connect to the server(s), must be the same everywhere it is specified.

### Configuring `jute.maxbuffer` on ZooKeeper Nodes

`jute.maxbuffer` must be configured on each external ZooKeeper node. This can be achieved in any of the following ways; note though that only the first option works on Windows:

1. In `<ZOOKEEPER_HOME>/conf/zoo.cfg`, e.g., to increase the file size limit to one byte less than 10MB, add this line:

```
jute.maxbuffer=0x9fffff
```

2. In `<ZOOKEEPER_HOME>/conf/zookeeper-env.sh`, e.g., to increase the file size limit to 50MiB, add this line:

```
JVMFLAGS="$JVMFLAGS -Djute.maxbuffer=50000000"
```

3. In `<ZOOKEEPER_HOME>/bin/zkServer.sh`, add a `JVMFLAGS` environment variable assignment near the top of the script, e.g., to increase the file size limit to 5MiB:

```
JVMFLAGS="$JVMFLAGS -Djute.maxbuffer=5000000"
```

### Configuring `jute.maxbuffer` for ZooKeeper Clients

The `bin/solr` script invokes Java programs that act as ZooKeeper clients. When you use Solr's bundled ZooKeeper server instead of setting up an external ZooKeeper ensemble, the configuration described below will also configure the ZooKeeper server.

Add the setting to the `SOLR_OPTS` environment variable in Solr's include file (`bin/solr.in.sh` or `solr.in.cmd`):

## Linux: solr.in.sh

The section to look for will start:

```
Anything you add to the SOLR_OPTS variable will be included in the java
start command line as-is, in ADDITION to other options. If you specify the
-a option on start script, those options will be appended as well. Examples:
```

Add the following line to increase the file size limit to 2MB:

```
SOLR_OPTS="$SOLR_OPTS -Djute.maxbuffer=0x200000"
```

## Windows: solr.in.cmd

The section to look for will start:

```
REM Anything you add to the SOLR_OPTS variable will be included in the java
REM start command line as-is, in ADDITION to other options. If you specify the
REM -a option on start script, those options will be appended as well. Examples:
```

Add the following line to increase the file size limit to 2MB:

```
set SOLR_OPTS=%SOLR_OPTS% -Djute.maxbuffer=0x200000
```

## Securing the ZooKeeper Connection

You may also want to secure the communication between ZooKeeper and Solr.

To setup ACL protection of znodes, see the section [ZooKeeper Access Control](#).

## Using ZooKeeper to Manage Configuration Files

With SolrCloud your configuration files are kept in ZooKeeper.

These files are uploaded in either of the following cases:

- When you start a SolrCloud example using the bin/solr script.
- When you create a collection using the bin/solr script.
- Explicitly upload a configuration set to ZooKeeper.

## Startup Bootstrap

When you try SolrCloud for the first time using the bin/solr -e cloud, the related configset gets uploaded to ZooKeeper automatically and is linked with the newly created collection.

The below command would start SolrCloud with the default collection name (gettingstarted) and default configset (\_default) uploaded and linked to it.

```
bin/solr -e cloud -noprompt
```

You can also explicitly upload a configuration directory when creating a collection using the bin/solr script with the -d option, such as:

```
bin/solr create -c mycollection -d _default
```

The create command will upload a copy of the \_default configuration directory to ZooKeeper under /configs/mycollection. Refer to the [Solr Control Script Reference](#) page for more details about the create command for creating collections.

Once a configuration directory has been uploaded to ZooKeeper, you can update them using the [Solr Control Script](#)



It's a good idea to keep these files under version control.

## Uploading Configuration Files using bin/solr or SolrJ

In production situations, [Config Sets](#) can also be uploaded to ZooKeeper independent of collection creation using either Solr's [Solr Control Script](#) or SolrJ.

The below command can be used to upload a new configset using the bin/solr script.

```
bin/solr zk upconfig -n <name for configset> -d <path to directory with configset>
```

The following code shows how this can also be achieved using SolrJ:

```
try (SolrZkClient zkClient = new SolrZkClient(zkConnectionString, ZK_TIMEOUT_MILLIS)) {
 ZkConfigManager manager = new ZkConfigManager(zkClient);
 manager.uploadConfigDir(Paths.get(localConfigSetDirectory), "nameForConfigset");
}
```

It is strongly recommended that the configurations be kept in a version control system, Git, SVN or similar.

## Managing Your SolrCloud Configuration Files

To update or change your SolrCloud configuration files:

1. Download the latest configuration files from ZooKeeper, using the source control checkout process.
2. Make your changes.
3. Commit your changed file to source control.
4. Push the changes back to ZooKeeper.

5. Reload the collection so that the changes will be in effect.

## Preparing ZooKeeper before First Cluster Start

If you will share the same ZooKeeper instance with other applications you should use a *chroot* in ZooKeeper. Please see [ZooKeeper chroot](#) for instructions.

There are certain configuration files containing cluster wide configuration. Since some of these are crucial for the cluster to function properly, you may need to upload such files to ZooKeeper before starting your Solr cluster for the first time. Examples of such configuration files (not exhaustive) are solr.xml, security.json and clusterprops.json.

If you for example would like to keep your solr.xml in ZooKeeper to avoid having to copy it to every node's solr\_home directory, you can push it to ZooKeeper with the bin/solr utility (Unix example):

```
bin/solr zk cp file:local/file/path/to/solr.xml zk:/solr.xml -z localhost:2181
```



If you have defined ZK\_HOST in solr.in.sh/solr.in.cmd (see [instructions](#)) you can omit -z <zk host string> from the above command.

## ZooKeeper Access Control

This section describes using ZooKeeper access control lists (ACLs) with Solr. For information about ZooKeeper ACLs, see the ZooKeeper documentation at [http://zookeeper.apache.org/doc/r3.4.14/zookeeperProgrammers.html#sc\\_ZooKeeperAccessControl](http://zookeeper.apache.org/doc/r3.4.14/zookeeperProgrammers.html#sc_ZooKeeperAccessControl).

### About ZooKeeper ACLs

SolrCloud uses ZooKeeper for shared information and for coordination.

This section describes how to configure Solr to add more restrictive ACLs to the ZooKeeper content it creates, and how to tell Solr about the credentials required to access the content in ZooKeeper. If you want to use ACLs in your ZooKeeper nodes, you will have to activate this functionality; by default, Solr behavior is open-unsafe ACL everywhere and uses no credentials.

Content stored in ZooKeeper is critical to the operation of a SolrCloud cluster. Open access to SolrCloud content on ZooKeeper could lead to a variety of problems. For example:

- Changing configuration might cause Solr to fail or behave in an unintended way.
- Changing cluster state information into something wrong or inconsistent might very well make a SolrCloud cluster behave strangely.
- Adding a delete-collection job to be carried out by the Overseer will cause data to be deleted from the cluster.

You may want to enable ZooKeeper ACLs with Solr if you grant access to your ZooKeeper ensemble to entities you do not trust, or if you want to reduce risk of bad actions resulting from, for example:

- Malware that found its way into your system.

- Other systems using the same ZooKeeper ensemble (a "bad thing" might be done by accident).

You might even want to limit read-access, if you think there is stuff in ZooKeeper that not everyone should know about. Or you might just in general work on a need-to-know basis.

Protecting ZooKeeper itself could mean many different things. **This section is about protecting Solr content in ZooKeeper.** ZooKeeper content basically lives persisted on disk and (partly) in memory of the ZooKeeper processes. **This section is not about protecting ZooKeeper data at storage or ZooKeeper process levels** - that's for ZooKeeper to deal with.

But this content is also available to "the outside" via the ZooKeeper API. Outside processes can connect to ZooKeeper and create/update/delete/read content; for example, a Solr node in a SolrCloud cluster wants to create/update/delete/read, and a SolrJ client wants to read from the cluster. It is the responsibility of the outside processes that create/update content to setup ACLs on the content. ACLs describe who is allowed to read, update, delete, create, etc. Each piece of information (znode/content) in ZooKeeper has its own set of ACLs, and inheritance or sharing is not possible. The default behavior in Solr is to add one ACL on all the content it creates - one ACL that gives anyone the permission to do anything (in ZooKeeper terms this is called "the open-unsafe ACL").

## How to Enable ACLs

We want to be able to:

1. Control the credentials Solr uses for its ZooKeeper connections. The credentials are used to get permission to perform operations in ZooKeeper.
2. Control which ACLs Solr will add to znodes (ZooKeeper files/folders) it creates in ZooKeeper.
3. Control it "from the outside", so that you do not have to modify and/or recompile Solr code to turn this on.

Solr nodes, clients and tools (e.g., ZkCLI) always use a java class called SolrZkClient to deal with their ZooKeeper stuff. The implementation of the solution described here is all about changing SolrZkClient. If you use SolrZkClient in your application, the descriptions below will be true for your application too.

### Controlling Credentials

You control which credentials provider will be used by configuring the zkCredentialsProvider property in solr.xml 's <solrcloud> section to the name of a class (on the classpath) implementing the ZkCredentialsProvider interface. server/solr/solr.xml in the Solr distribution defines the zkCredentialsProvider such that it will take on the value of the same-named zkCredentialsProvider system property if it is defined (e.g., by uncommenting the SOLR\_ZK\_CREDS\_AND\_ACLS environment variable definition in solr.in.sh/.cmd - see below), or if not, default to the DefaultZkCredentialsProvider implementation.

### Out of the Box Credential Implementations

You can always make your own implementation, but Solr comes with two implementations:

- org.apache.solr.common.cloud.DefaultZkCredentialsProvider: Its getCredentials() returns a list of length zero, or "no credentials used". This is the default.
- org.apache.solr.common.cloud.VMParamsSingleSetCredentialsDigestZkCredentialsProvider: This

lets you define your credentials using system properties. It supports at most one set of credentials.

- The schema is "digest". The username and password are defined by system properties zkDigestUsername and zkDigestPassword. This set of credentials will be added to the list of credentials returned by getCredentials() if both username and password are provided.
- If the one set of credentials above is not added to the list, this implementation will fall back to default behavior and use the (empty) credentials list from DefaultZkCredentialsProvider.

## Controlling ACLs

You control which ACLs will be added by configuring zkACLProvider property in solr.xml's <solrcloud> section to the name of a class (on the classpath) implementing the ZkACLProvider interface.

server/solr/solr.xml in the Solr distribution defines the zkACLProvider such that it will take on the value of the same-named zkACLProvider system property if it is defined (e.g., by uncommenting the SOLR\_ZK\_CREDS\_AND\_ACLS environment variable definition in solr.in.sh/.cmd - see below), or if not, default to the DefaultZkACLProvider implementation.

## Out of the Box ACL Implementations

You can always make your own implementation, but Solr comes with:

- org.apache.solr.common.cloud.DefaultZkACLProvider: It returns a list of length one for all zNodePath -s. The single ACL entry in the list is "open-unsafe". This is the default.
- org.apache.solr.common.cloud.VMParamsAllAndReadonlyDigestZkACLProvider: This lets you define your ACLs using system properties. Its getACLsToAdd() implementation does not use zNodePath for anything, so all znodes will get the same set of ACLs. It supports adding one or both of these options:
  - A user that is allowed to do everything.
    - The permission is ALL (corresponding to all of CREATE, READ, WRITE, DELETE, and ADMIN), and the schema is "digest".
    - The username and password are defined by system properties zkDigestUsername and zkDigestPassword, respectively.
    - This ACL will not be added to the list of ACLs unless both username and password are provided.
  - A user that is only allowed to perform read operations.
    - The permission is READ and the schema is digest.
    - The username and password are defined by system properties zkDigestReadOnlyUsername and zkDigestReadOnlyPassword, respectively.
    - This ACL will not be added to the list of ACLs unless both username and password are provided.
- org.apache.solr.common.cloud.SaslZkACLProvider: Requires SASL authentication. Gives all permissions for the user specified in system property solr.authorization.superuser (default: solr) when using SASL, and gives read permissions for anyone else. Designed for a setup where configurations have already been set up and will not be modified, or where configuration changes are controlled via Solr APIs. This provider will be useful for administration in a kerberos environment. In such an environment, the administrator wants Solr to authenticate to ZooKeeper using SASL, since this is the only way to authenticate with ZooKeeper via Kerberos.

If none of the above ACLs is added to the list, the (empty) ACL list of DefaultZkACLProvider will be used by

default.

Notice the overlap in system property names with credentials provider `VMParamsSingleSetCredentialsDigestZkCredentialsProvider` (described above). This is to let the two providers collaborate in a nice and perhaps common way: we always protect access to content by limiting to two users - an admin-user and a readonly-user - AND we always connect with credentials corresponding to this same admin-user, basically so that we can do anything to the content/znodes we create ourselves.

You can give the readonly credentials to "clients" of your SolrCloud cluster - e.g., to be used by SolrJ clients. They will be able to read whatever is necessary to run a functioning SolrJ client, but they will not be able to modify any content in ZooKeeper.

## ZooKeeper ACLs in Solr Scripts

There are two scripts that impact ZooKeeper ACLs:

- For \*nix systems: `bin/solr & server/scripts/cloud-scripts/zkcli.sh`
- For Windows systems: `bin/solr.cmd & server/scripts/cloud-scripts/zkcli.bat`

These Solr scripts can enable use of ZooKeeper ACLs by setting the appropriate system properties: uncomment the following and replace the passwords with ones you choose to enable the above-described VM parameters ACL and credentials providers in the following files:

### *solr.in.sh*

```
Settings for ZK ACL
#SOLR_ZK_CREDS_AND_ACLS=-
DzkACLProvider=org.apache.solr.common.cloud.VMParamsAllAndReadOnlyDigestZkACLProvider \
#
-DzkCredentialsProvider=org.apache.solr.common.cloud.VMParamsSingleSetCredentialsDigestZkCredenti
alsProvider \
-DzkDigestUsername=admin-user -DzkDigestPassword=CHANGEME-ADMIN-PASSWORD \
-DzkDigestReadOnlyUsername=readonly-user -DzkDigestReadOnlyPassword=CHANGEME-READONLY
-PASSWORD"
#SOLR_OPTS="$SOLR_OPTS $SOLR_ZK_CREDS_AND_ACLS"
```

### *solr.in.cmd*

```
REM Settings for ZK ACL
REM set SOLR_ZK_CREDS_AND_ACLS=-
DzkACLProvider=org.apache.solr.common.cloud.VMParamsAllAndReadOnlyDigestZkACLProvider ^
REM
-DzkCredentialsProvider=org.apache.solr.common.cloud.VMParamsSingleSetCredentialsDigestZkCredenti
alsProvider ^
REM -DzkDigestUsername=admin-user -DzkDigestPassword=CHANGEME-ADMIN-PASSWORD ^
REM -DzkDigestReadOnlyUsername=readonly-user -DzkDigestReadOnlyPassword=CHANGEME-READONLY
-PASSWORD
REM set SOLR_OPTS=%SOLR_OPTS% %SOLR_ZK_CREDS_AND_ACLS%
```

***zkcli.sh***

```
Settings for ZK ACL
#SOLR_ZK_CREDS_AND_ACLS=-
DzkACLProvider=org.apache.solr.common.cloud.VMParamsAllAndReadOnlyDigestZkACLProvider \
#
-DzkCredentialsProvider=org.apache.solr.common.cloud.VMParamsSingleSetCredentialsDigestZkCredenti
alsProvider \
-DzkDigestUsername=admin-user -DzkDigestPassword=CHANGEME-ADMIN-PASSWORD \
-DzkDigestReadOnlyUsername=readonly-user -DzkDigestReadOnlyPassword=CHANGEME-READONLY
-PASSWORD"
```

***zkcli.bat***

```
REM Settings for ZK ACL
REM set SOLR_ZK_CREDS_AND_ACLS=-
DzkACLProvider=org.apache.solr.common.cloud.VMParamsAllAndReadOnlyDigestZkACLProvider ^
REM
-DzkCredentialsProvider=org.apache.solr.common.cloud.VMParamsSingleSetCredentialsDigestZkCredenti
alsProvider ^
REM -DzkDigestUsername=admin-user -DzkDigestPassword=CHANGEME-ADMIN-PASSWORD ^
REM -DzkDigestReadOnlyUsername=readonly-user -DzkDigestReadOnlyPassword=CHANGEME-READONLY
-PASSWORD
```

## Changing ACL Schemes

Over the lifetime of operating your Solr cluster, you may decide to move from an unsecured ZooKeeper to a secured instance. Changing the configured zkACLProvider in solr.xml will ensure that newly created nodes are secure, but will not protect the already existing data. To modify all existing ACLs, you can use the updateacls command with Solr's ZkCLI. First uncomment the SOLR\_ZK\_CREDS\_AND\_ACLS environment variable definition in server/scripts/cloud-scripts/zkcli.sh (or zkcli.bat on Windows) and fill in the passwords for the admin-user and the readonly-user - see above - then run server/scripts/cloud-scripts/zkcli.sh -cmd updateacls /zk-path, or on Windows run server\scripts\cloud-scripts\zkcli.bat cmd updateacls /zk-path.

Changing ACLs in ZK should only be done while your SolrCloud cluster is stopped. Attempting to do so while Solr is running may result in inconsistent state and some nodes becoming inaccessible.

The VM properties zkACLProvider and zkCredentialsProvider, included in the SOLR\_ZK\_CREDS\_AND\_ACLS environment variable in zkcli.sh/.bat, control the conversion:

- The Credentials Provider must be one that has current admin privileges on the nodes. When omitted, the process will use no credentials (suitable for an unsecure configuration).
- The ACL Provider will be used to compute the new ACLs. When omitted, the process will set all permissions to all users, removing any security present.

The uncommented SOLR\_ZK\_CREDS\_AND\_ACLS environment variable in zkcli.sh/.bat sets the credentials and ACL providers to the VMParamsSingleSetCredentialsDigestZkCredentialsProvider and VMParamsAllAndReadOnlyDigestZkACLProvider implementations, described earlier in the page.

## Collections API

The Collections API is used to create, remove, or reload collections.

In the context of SolrCloud you can use it to create collections with a specific number of shards and replicas, move replicas or shards, and create or delete collection aliases.

### CREATE: Create a Collection

/admin/collections?action=CREATE&name=name

#### CREATE Parameters

The CREATE action allows the following parameters:

name

The name of the collection to be created. This parameter is required.

router.name

The router name that will be used. The router defines how documents will be distributed among the shards. Possible values are `implicit` or `compositeId`, which is the default.

The `implicit` router does not automatically route documents to different shards. Whichever shard you indicate on the indexing request (or within each document) will be used as the destination for those documents.

The `compositeId` router hashes the value in the `uniqueKey` field and looks up that hash in the collection's clusterstate to determine which shard will receive the document, with the additional ability to manually direct the routing.

When using the `implicit` router, the `shards` parameter is required. When using the `compositeId` router, the `numShards` parameter is required.

For more information, see also the section [Document Routing](#).

numShards

The number of shards to be created as part of the collection. This is a required parameter when the `router.name` is `compositeId`.

shards

A comma separated list of shard names, e.g., `shard-x,shard-y,shard-z`. This is a required parameter when the `router.name` is `implicit`.

replicationFactor

The number of replicas to be created for each shard. The default is 1.

This will create a NRT type of replica. If you want another type of replica, see the `tlogReplicas` and `pullReplica` parameters below. See the section [Types of Replicas](#) for more information about replica types.

nrtReplicas

The number of NRT (Near-Real-Time) replicas to create for this collection. This type of replica maintains a

transaction log and updates its index locally. If you want all of your replicas to be of this type, you can simply use replicationFactor instead.

#### tlogReplicas

The number of TLOG replicas to create for this collection. This type of replica maintains a transaction log but only updates its index via replication from a leader. See the section [Types of Replicas](#) for more information about replica types.

#### pullReplicas

The number of PULL replicas to create for this collection. This type of replica does not maintain a transaction log and only updates its index via replication from a leader. This type is not eligible to become a leader and should not be the only type of replicas in the collection. See the section [Types of Replicas](#) for more information about replica types.

#### maxShardsPerNode

When creating collections, the shards and/or replicas are spread across all available (i.e., live) nodes, and two replicas of the same shard will never be on the same node.

If a node is not live when the CREATE action is called, it will not get any parts of the new collection, which could lead to too many replicas being created on a single live node. Defining maxShardsPerNode sets a limit on the number of replicas the CREATE action will spread to each node.

If the entire collection can not be fit into the live nodes, no collection will be created at all. The default maxShardsPerNode value is 1. A value of -1 means unlimited. If a policy is also specified then the stricter of maxShardsPerNode and policy rules apply.

#### createNodeSet

Allows defining the nodes to spread the new collection across. The format is a comma-separated list of node\_names, such as localhost:8983\_solr,localhost:8984\_solr,localhost:8985\_solr.

If not provided, the CREATE operation will create shard-replicas spread across all live Solr nodes.

Alternatively, use the special value of EMPTY to initially create no shard-replica within the new collection and then later use the [ADDREPLICA](#) operation to add shard-replicas when and where required.

#### createNodeSet.shuffle

Controls whether or not the shard-replicas created for this collection will be assigned to the nodes specified by the createNodeSet in a sequential manner, or if the list of nodes should be shuffled prior to creating individual replicas.

A false value makes the results of a collection creation predictable and gives more exact control over the location of the individual shard-replicas, but true can be a better choice for ensuring replicas are distributed evenly across nodes. The default is true.

This parameter is ignored if createNodeSet is not also specified.

#### collection.configName

Defines the name of the configuration (which **must already be stored in ZooKeeper**) to use for this collection. If not provided, Solr will use the configuration of \_default configset to create a new (and mutable) configset named <collectionName>.AUTOCREATED and will use it for the new collection. When such a collection (that uses a copy of the \_default configset) is deleted, the autogenerated configset is not

deleted by default.

#### router.field

If this parameter is specified, the router will look at the value of the field in an input document to compute the hash and identify a shard instead of looking at the uniqueKey field. If the field specified is null in the document, the document will be rejected.

Please note that [RealTime Get](#) or retrieval by document ID would also require the parameter `_route_` (or `shard.keys`) to avoid a distributed search.

#### property.name=value

Set core property *name* to *value*. See the section [Defining core.properties](#) for details on supported properties and values.

#### autoAddReplicas

When set to true, enables automatic addition of replicas when the number of active replicas falls below the value set for `replicationFactor`. This may occur if a replica goes down, for example. The default is false, which means new replicas will not be added.

While this parameter is provided as part of Solr's set of features to provide autoscaling of clusters, it is available even when you have not implemented any other part of autoscaling (such as a policy). See the section [SolrCloud Autoscaling Automatically Adding Replicas](#) for more details about this option and how it can be used.

#### async

Request ID to track this action which will be [processed asynchronously](#).

#### rule

Replica placement rules. See the section [Rule-based Replica Placement](#) for details.

#### snitch

Details of the snitch provider. See the section [Rule-based Replica Placement](#) for details.

#### policy

Name of the collection-level policy. See [Defining Collection-Specific Policies](#) for details.

#### waitForFinalState

If true, the request will complete only when all affected replicas become active. The default is false, which means that the API will return the status of the single action, which may be before the new replica is online and active.

#### withCollection

The name of the collection with which all replicas of this collection must be co-located. The collection must already exist and must have a single shard named `shard1`. See [Colocating collections](#) for more details.

#### alias

Starting with version 8.1 when a collection is created additionally an alias can be created that points to this collection. This parameter allows specifying the name of this alias, effectively combining this operation with [CREATEALIAS](#)

Collections are first created in read-write mode but can be put in readOnly mode using the [MODIFYCOLLECTION](#) action.

## CREATE Response

The response will include the status of the request and the new core names. If the status is anything other than "success", an error message will explain why the request failed.

### Examples using CREATE

#### Input

```
http://localhost:8983/solr/admin/collections?action=CREATE&name=newCollection&numShards=2&replicaFactor=1&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">3764</int>
 </lst>
 <lst name="success">
 <lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">3450</int>
 </lst>
 <str name="core">newCollection_shard1_replica1</str>
 </lst>
 <lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">3597</int>
 </lst>
 <str name="core">newCollection_shard2_replica1</str>
 </lst>
 </lst>
</response>
```

## MODIFYCOLLECTION: Modify Attributes of a Collection

```
/admin/collections?action=MODIFYCOLLECTION&collection=<collection-name>&<attribute-name>=<attribute-value>&<another-attribute-name>=<another-value>&<yet_another_attribute_name>=
```

It's possible to edit multiple attributes at a time. Changing these values only updates the z-node on ZooKeeper, they do not change the topology of the collection. For instance, increasing replicationFactor will *not* automatically add more replicas to the collection but *will* allow more ADDREPLICA commands to succeed.

An attribute can be deleted by passing an empty value. For example, `yet_another_attribute_name=` (with no value) will delete the `yet_another_attribute_name` parameter from the collection.

## MODIFYCOLLECTION Parameters

`collection`

The name of the collection to be modified. This parameter is required.

`attribute=value`

Key-value pairs of attribute names and attribute values.

At least one attribute parameter is required.

The attributes that can be modified are:

- `maxShardsPerNode`
- `replicationFactor`
- `autoAddReplicas`
- `collection.configName`
- `rule`
- `snitch`
- `policy`
- `withCollection`
- `readOnly`
- other custom properties that use a property. prefix

See the [CREATE action](#) section above for details on these attributes.

### Read-Only Mode

Setting the `readOnly` attribute to true puts the collection in read-only mode, in which any index update requests are rejected. Other collection-level actions (e.g., adding / removing / moving replicas) are still available in this mode.

The transition from the (default) read-write to read-only mode consists of the following steps:

- the `readOnly` flag is changed in collection state,
- any new update requests are rejected with 403 FORBIDDEN error code (ongoing long-running requests are aborted, too),
- a forced commit is performed to flush and commit any in-flight updates.



This may potentially take a long time if there are still major segment merges running in the background.

- a collection [RELOAD action](#) is executed.

Removing the `readOnly` property or setting it to false enables the processing of updates and reloads the collection.

## REINDEXCOLLECTION: Re-Index a Collection

/admin/collections?action=REINDEXCOLLECTION&name=name

The REINDEXCOLLECTION command re-indexes a collection using existing data from the source collection.



Re-indexing is potentially a lossy operation - some of the existing indexed data that is not available as stored fields may be lost, so users should use this command with caution, evaluating the potential impact by using different source and target collection names first, and preserving the source collection until the evaluation is complete.

The target collection must not exist (and may not be an alias). If the target collection name is the same as the source collection then first a unique sequential name will be generated for the target collection, and then after re-indexing is done an alias will be created that points from the source name to the actual sequentially-named target collection.

When re-indexing is started the source collection is put in [read-only mode](#) to ensure that all source documents are properly processed.

Using optional parameters a different index schema, collection shape (number of shards and replicas) or routing parameters can be requested for the target collection.

Re-indexing is executed as a streaming expression daemon, which runs on one of the source collection's replicas. It is usually a time-consuming operation so it's recommended to execute it as an asynchronous request in order to avoid request time outs. Only one re-indexing operation may execute concurrently for a given source collection. Long-running, erroneous or crashed re-indexing operations may be terminated by using the abort option, which also removes partial results.

### REINDEXCOLLECTION Parameters

name

Source collection name, may be an alias. This parameter is required.

cmd

Optional command. Default command is start. Currently supported commands are:

- start - default, starts processing if not already running,
- abort - aborts an already running re-indexing (or clears a left-over status after a crash), and deletes partial results,
- status - returns detailed status of a running re-indexing command.

target

Target collection name, optional. If not specified a unique name will be generated and after all documents have been copied an alias will be created that points from the source collection name to the unique sequentially-named collection, effectively "hiding" the original source collection from regular update and search operations.

q

Optional query to select documents for re-indexing. Default value is \*:\*.

f1

Optional list of fields to re-index. Default value is \*.

rows

Documents are transferred in batches. Depending on the average size of the document large batch sizes may cause memory issues. Default value is 100.

configName

collection.configName

Optional name of the configset for the target collection. Default is the same as the source collection.

There's a number of optional parameters that determine the target collection layout. If they are not specified in the request then their values are copied from the source collection. The following parameters are currently supported (described in details in the [CREATE collection](#) section): numShards, replicationFactor, nrtReplicas, tlogReplicas, pullReplicas, maxShardsPerNode, autoAddReplicas, shards, policy, createNodeSet, createNodeSet.shuffle, router.\*.

removeSource

Optional boolean. If true then after the processing is successfully finished the source collection will be deleted.

async

Optional request ID to track this action which will be [processed asynchronously](#).

When the re-indexing process has completed the target collection is marked using property .rx: "finished", and the source collection state is updated to become read-write. On any errors the command will delete any temporary and target collections and also reset the state of the source collection's read-only flag.

## Examples using REINDEXCOLLECTION

### Input

```
http://localhost:8983/solr/admin/collections?action=REINDEXCOLLECTION&name=newCollection&numShard
s=3&configName=conf2&q=id:aa*&f1=id,string_s
```

This request specifies a different schema for the target collection, copies only some of the fields, selects only the documents matching a query, and also potentially re-shapes the collection by explicitly specifying 3 shards. Since the target collection hasn't been specified in the parameters, a collection with a unique name, e.g., .rx\_newCollection\_2, will be created and on success an alias pointing from newCollection to .rx\_newCollection\_2 will be created, effectively replacing the source collection for the purpose of indexing and searching. The source collection is assumed to be small so a synchronous request was made.

### Output

```
{
 "responseHeader":{
 "status":0,
 "QTime":10757},
 "reindexStatus":{
 "phase":"done",
 "inputDocs":13416,
 "processedDocs":376,
 "actualSourceCollection":".rx_newCollection_1",
 "state":"finished",
 "actualTargetCollection":".rx_newCollection_2",
 "checkpointCollection":".rx_ck_newCollection"
 }
}
```

As a result a new collection .rx\_newCollection\_2 has been created, with selected documents re-indexed to 3 shards, and with an alias pointing from newCollection to this one. The status also shows that the source collection was already an alias to .rx\_newCollection\_1, which was likely a result of a previous re-indexing.

## RELOAD: Reload a Collection

/admin/collections?action=RELOAD&name=name

The RELOAD action is used when you have changed a configuration in ZooKeeper.

### RELOAD Parameters

name

The name of the collection to reload. This parameter is required.

async

Request ID to track this action which will be [processed asynchronously](#).

### RELOAD Response

The response will include the status of the request and the cores that were reloaded. If the status is anything other than "success", an error message will explain why the request failed.

### Examples using RELOAD

#### Input

```
http://localhost:8983/solr/admin/collections?action=RELOAD&name=newCollection&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1551</int>
 </lst>
 <lst name="success">
 <lst name="10.0.1.6:8983_solr">
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">761</int>
 </lst>
 </lst>
 <lst name="10.0.1.4:8983_solr">
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1527</int>
 </lst>
 </lst>
 </lst>
</response>
```

## RENAME: Rename a Collection

/admin/collections?action=RENAME&name=existingName&target=targetName

Renaming a collection sets up a standard alias that points to the underlying collection, so that the same (unmodified) collection can now be referred to in query, index and admin operations using the new name.

This command does NOT actually rename the underlying Solr collection - it sets up a new one-to-one alias using the new name, or renames the existing alias so that it uses the new name, while still referring to the same underlying Solr collection. However, from the user's point of view the collection can now be accessed using the new name, and the new name can be also referred to in other aliases.

The following limitations apply:

- the existing name must be either a SolrCloud collection or a standard alias referring to a single collection. Aliases that refer to more than 1 collection are not supported.
- the existing name must not be a Routed Alias.
- the target name must not be an existing alias.

### RENAME Command Parameters

#### name

Name of the existing SolrCloud collection or an alias that refers to exactly one collection and is not a Routed Alias.

#### target

Target name of the collection. This will be the new alias that refers to the underlying SolrCloud collection. The original name (or alias) of the collection will be replaced also in the existing aliases so that they also refer to the new name. Target name must not be an existing alias.

## Examples using RENAME

Assuming there are two actual SolrCloud collections named collection1 and collection2, and the following aliases already exist:

- col1 -> collection1: this resolves to collection1.
- col2 -> collection2: this resolves to collection2.
- simpleAlias -> col1: this resolves to collection1.
- compoundAlias -> col1,col2: this resolves to collection1,collection2

The RENAME of col1 to foo will change the aliases to the following:

- foo -> collection1: this resolves to collection1.
- col2 -> collection2: this resolves to collection2.
- simpleAlias -> foo: this resolves to collection1.
- compoundAlias -> foo,col2: this resolves to collection1,collection2.

If we then rename collection1 (which is an actual collection name) to collection2 (which is also an actual collection name) the following aliases will exist now:

- foo -> collection2: this resolves to collection2.
- col2 -> collection2: this resolves to collection2.
- simpleAlias -> foo: this resolves to collection2.
- compoundAlias -> foo,col2: this would resolve now to collection2,collection2 so it's reduced to simply collection2.
- collection1 -> collection2: this newly created alias effectively hides collection1 from regular query and update commands, which are directed now to collection2.

## SPLITSHARD: Split a Shard

/admin/collections?action=SPLITSHARD&collection=name&shard=shardID

Splitting a shard will take an existing shard and break it into two pieces which are written to disk as two (new) shards. The original shard will continue to contain the same data as-is but it will start re-routing requests to the new shards. The new shards will have as many replicas as the original shard. A soft commit is automatically issued after splitting a shard so that documents are made visible on sub-shards. An explicit commit (hard or soft) is not necessary after a split operation because the index is automatically persisted to disk during the split operation.

This command allows for seamless splitting and requires no downtime. A shard being split will continue to accept query and indexing requests and will automatically start routing requests to the new shards once this operation is complete. This command can only be used for SolrCloud collections created with numShards parameter, meaning collections which rely on Solr's hash-based routing mechanism.

The split is performed by dividing the original shard's hash range into two equal partitions and dividing up the documents in the original shard according to the new sub-ranges. Two parameters discussed below, ranges and split.key provide further control over how the split occurs.

The newly created shards will have as many replicas as the parent shard, of the same replica types.

When using `splitMethod=rewrite` (default) you must ensure that the node running the leader of the parent shard has enough free disk space i.e., more than twice the index size, for the split to succeed. The API uses the Autoscaling framework to find nodes that can satisfy the disk requirements for the new replicas but only when an Autoscaling policy is configured. Refer to [Autoscaling Policy and Preferences](#) section for more details.

Also, the first replicas of resulting sub-shards will always be placed on the shard leader node, which may cause Autoscaling policy violations that need to be resolved either automatically (when appropriate triggers are in use) or manually.

Shard splitting can be a long running process. In order to avoid timeouts, you should run this as an [asynchronous call](#).

## SPLITSHARD Parameters

### collection

The name of the collection that includes the shard to be split. This parameter is required.

### shard

The name of the shard to be split. This parameter is required when `split.key` is not specified.

### ranges

A comma-separated list of hash ranges in hexadecimal, such as `ranges=0-1f4,1f5-3e8,3e9-5dc`.

This parameter can be used to divide the original shard's hash range into arbitrary hash range intervals specified in hexadecimal. For example, if the original hash range is 0-1500 then adding the parameter: `ranges=0-1f4,1f5-3e8,3e9-5dc` will divide the original shard into three shards with hash range 0-500, 501-1000, and 1001-1500 respectively.

### split.key

The key to use for splitting the index.

This parameter can be used to split a shard using a route key such that all documents of the specified route key end up in a single dedicated sub-shard. Providing the shard parameter is not required in this case because the route key is enough to figure out the right shard. A route key which spans more than one shard is not supported.

For example, suppose `split.key=A!` hashes to the range 12-15 and belongs to shard 'shard1' with range 0-20. Splitting by this route key would yield three sub-shards with ranges 0-11, 12-15 and 16-20. Note that the sub-shard with the hash range of the route key may also contain documents for other route keys whose hash ranges overlap.

### numSubShards

The number of sub-shards to split the parent shard into. Allowed values for this are in the range of 2-8 and defaults to 2.

This parameter can only be used when `ranges` or `split.key` are not specified.

### splitMethod

Currently two methods of shard splitting are supported:

- splitMethod=rewrite (default) after selecting documents to retain in each partition this method creates sub-indexes from scratch, which is a lengthy CPU- and I/O-intensive process but results in optimally-sized sub-indexes that don't contain any data from documents not belonging to each partition.
- splitMethod=link uses file system-level hard links for creating copies of the original index files and then only modifies the file that contains the list of deleted documents in each partition. This method is many times quicker and lighter on resources than the rewrite method but the resulting sub-indexes are still as large as the original index because they still contain data from documents not belonging to the partition. This slows down the replication process and consumes more disk space on replica nodes (the multiple hard-linked copies don't occupy additional disk space on the leader node, unless hard-linking is not supported).

#### splitFuzz

A float value (default is 0.0f, must be smaller than 0.5f) that allows to vary the sub-shard ranges by this percentage of total shard range, odd shards being larger and even shards being smaller.

#### property.name=value

Set core property *name* to *value*. See the section [Defining core.properties](#) for details on supported properties and values.

#### waitForFinalState

If true, the request will complete only when all affected replicas become active. The default is false, which means that the API will return the status of the single action, which may be before the new replica is online and active.

#### timing

If true then each stage of processing will be timed and a timing section will be included in response.

#### async

Request ID to track this action which will be [processed asynchronously](#)

## SPLITSHARD Response

The output will include the status of the request and the new shard names, which will use the original shard as their basis, adding an underscore and a number. For example, "shard1" will become "shard1\_0" and "shard1\_1". If the status is anything other than "success", an error message will explain why the request failed.

## Examples using SPLITSHARD

### Input

Split shard1 of the "anotherCollection" collection.

```
http://localhost:8983/solr/admin/collections?action=SPLITSHARD&collection=anotherCollection&shard=shard1&wt=xml
```

### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">6120</int>
 </lst>
 <lst name="success">
 <lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">3673</int>
 </lst>
 <str name="core">anotherCollection_shard1_1_replica1</str>
 </lst>
 <lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">3681</int>
 </lst>
 <str name="core">anotherCollection_shard1_0_replica1</str>
 </lst>
 <lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">6008</int>
 </lst>
 </lst>
 <lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">6007</int>
 </lst>
 </lst>
 <lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">71</int>
 </lst>
 </lst>
 <lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">0</int>
 </lst>
 <str name="core">anotherCollection_shard1_1_replica1</str>
 <str name="status">EMPTY_BUFFER</str>
 </lst>
 <lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">0</int>
 </lst>
 <str name="core">anotherCollection_shard1_0_replica1</str>
 </lst>
```

```
<str name="status">EMPTY_BUFFER</str>
</lst>
</lst>
</response>
```

## CREATESHARD: Create a Shard

Shards can only be created with this API for collections that use the 'implicit' router (i.e., when the collection was created, `router.name=implicit`). A new shard with a name can be created for an existing 'implicit' collection.

Use `SPLITSHARD` for collections created with the 'compositeId' router (`router.key=compositeId`).

```
/admin/collections?action=CREATESHARD&shard=shardName&collection=name
```

The default values for `replicationFactor` or `nrtReplicas`, `tlogReplicas`, `pullReplicas` from the collection is used to determine the number of replicas to be created for the new shard. This can be customized by explicitly passing the corresponding parameters to the request.

The API uses the Autoscaling framework to find the best possible nodes in the cluster when an Autoscaling preferences or policy is configured. Refer to [Autoscaling Policy and Preferences](#) section for more details.

### CREATESHARD Parameters

#### collection

The name of the collection that includes the shard to be split. This parameter is required.

#### shard

The name of the shard to be created. This parameter is required.

#### createNodeSet

Allows defining the nodes to spread the new collection across. If not provided, the CREATESHARD operation will create shard-replica spread across all live Solr nodes.

The format is a comma-separated list of `node_names`, such as

`localhost:8983_solr,localhost:8984_solr,localhost:8985_solr`.

#### nrtReplicas

The number of nrt replicas that should be created for the new shard (optional, the defaults for the collection is used if omitted)

#### tlogReplicas

The number of tlog replicas that should be created for the new shard (optional, the defaults for the collection is used if omitted)

#### pullReplicas

The number of pull replicas that should be created for the new shard (optional, the defaults for the collection is used if omitted)

#### property.name=value

Set core property `name` to `value`. See the section [Defining core.properties](#) for details on supported properties and values.

#### waitForFinalState

If true, the request will complete only when all affected replicas become active. The default is false, which means that the API will return the status of the single action, which may be before the new replica is online and active.

#### async

Request ID to track this action which will be [processed asynchronously](#).

## CREATESHARD Response

The output will include the status of the request. If the status is anything other than "success", an error message will explain why the request failed.

## Examples using CREATESHARD

### Input

Create 'shard-z' for the "anImplicitCollection" collection.

```
http://localhost:8983/solr/admin/collections?action=CREATESHARD&collection=anImplicitCollection&shard=shard-z&wt=xml
```

### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">558</int>
 </lst>
</response>
```

## DELETESHARD: Delete a Shard

Deleting a shard will unload all replicas of the shard, remove them from clusterstate.json, and (by default) delete the instanceDir and dataDir for each replica. It will only remove shards that are inactive, or which have no range given for custom sharding.

```
/admin/collections?action=DELETESHARD&shard=shardID&collection=name
```

## DELETESHARD Parameters

#### collection

The name of the collection that includes the shard to be deleted. This parameter is required.

#### shard

The name of the shard to be deleted. This parameter is required.

#### deleteInstanceDir

By default Solr will delete the entire instanceDir of each replica that is deleted. Set this to false to prevent the instance directory from being deleted.

#### deleteDataDir

By default Solr will delete the dataDir of each replica that is deleted. Set this to false to prevent the data directory from being deleted.

#### deleteIndex

By default Solr will delete the index of each replica that is deleted. Set this to false to prevent the index directory from being deleted.

#### async

Request ID to track this action which will be [processed asynchronously](#).

## DELETESHARD Response

The output will include the status of the request. If the status is anything other than "success", an error message will explain why the request failed.

## Examples using DELETESHARD

### Input

Delete 'shard1' of the "anotherCollection" collection.

```
http://localhost:8983/solr/admin/collections?action=DELETESHARD&collection=anotherCollection&shard=d=shard1&wt=xml
```

### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">558</int>
 </lst>
 <lst name="success">
 <lst name="10.0.1.4:8983_solr">
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">27</int>
 </lst>
 </lst>
 </lst>
</response>
```

## CREATEALIAS: Create or Modify an Alias for a Collection

The CREATEALIAS action will create a new alias pointing to one or more collections. Aliases come in 2 flavors: standard and routed.

**Standard aliases** are simple: CREATEALIAS registers the alias name with the names of one or more collections provided by the command. If an existing alias exists, it is replaced/updated. A standard alias can serve as a means to rename a collection, and can be used to atomically swap which backing/underlying

collection is "live" for various purposes. When Solr searches an alias pointing to multiple collections, Solr will search all shards of all the collections as an aggregated whole. While it is possible to send updates to an alias spanning multiple collections, standard aliases have no logic for distributing documents among the referenced collections so all updates will go to the first collection in the list.

```
/admin/collections?action=CREATEALIAS&name=name&collections=collectionlist
```

**Routed aliases** are aliases with additional capabilities to act as a kind of super-collection that route updates to the correct collection. Routing is data driven and may be based on a temporal field or on categories specified in a field (normally string based). See [Routed Aliases](#) for some important high-level information before getting started.

```
localhost:8983/solr/admin/collections?action=CREATEALIAS&name=timedata&router.start=NOW/DAY&route.r.field=evt_dt&router.name=time&router.interval=%2B1DAY&router.maxFutureMs=3600000&create-collection.collection.configName=myConfig&create-collection.numShards=2
```

If run on Jan 15, 2018, the above will create an time routed alias named `timedata`, that contains collections with names prefixed with `timedata` and an initial collection named `timedata_2018_01_15` will be created immediately. Updates sent to this alias with a (required) value in `evt_dt` that is before or after 2018-01-15 will be rejected, until the last 60 minutes of 2018-01-15. After 2018-01-15T23:00:00 documents for either 2018-01-15 or 2018-01-16 will be accepted. As soon as the system receives a document for an allowable time window for which there is no collection it will automatically create the next required collection (and potentially any intervening collections if `router.interval` is smaller than `router.maxFutureMs`). Both the initial collection and any subsequent collections will be created using the specified configset. All collection creation parameters other than name are allowed, prefixed by `create-collection`.

This means that one could, for example, partition their collections by day, and within each daily collection route the data to shards based on customer id. Such shards can be of any type (NRT, PULL or TLOG), and rule-based replica placement strategies may also be used.

The values supplied in this command for collection creation will be retained in alias properties, and can be verified by inspecting `aliases.json` in ZooKeeper.



Presently only updates are routed and queries are distributed to all collections in the alias, but future features may enable routing of the query to the single appropriate collection based on a special parameter or perhaps a filter on the routed field.

## CREATEALIAS Parameters

### name

The alias name to be created. This parameter is required. If the alias is to be routed it also functions as a prefix for the names of the dependent collections that will be created. It must therefore adhere to normal requirements for collection naming.

### async

Request ID to track this action which will be [processed asynchronously](#).

## Standard Alias Parameters

collections

A comma-separated list of collections to be aliased. The collections must already exist in the cluster. This parameter signals the creation of a standard alias. If it is present all routing parameters are prohibited. If routing parameters are present this parameter is prohibited.

### Routed Alias Parameters

Most routed alias parameters become *alias properties* that can subsequently be inspected and [modified](#).

router.name

The type of routing to use. Presently only time and category are valid. This parameter is required.

router.field

The field to inspect to determine which underlying collection an incoming document should be routed to. This field is required on all incoming documents.

create-collection.\*

The \* wildcard can be replaced with any parameter from the [CREATE](#) command except name. All other fields are identical in requirements and naming except that we insist that the configset be explicitly specified. The configset must be created beforehand, either uploaded or copied and modified. It's probably a bad idea to use "data driven" mode as schema mutations might happen concurrently leading to errors.

### Time Routed Alias Parameters

router.start

The start date/time of data for this time routed alias in Solr's standard date/time format (i.e., ISO-8601 or "NOW" optionally with [date math](#)).

The first collection created for the alias will be internally named after this value. If a document is submitted with an earlier value for router.field then the earliest collection the alias points to then it will yield an error since it can't be routed. This date/time MUST NOT have a milliseconds component other than 0. Particularly, this means NOW will fail 999 times out of 1000, though NOW/SECOND, NOW/MINUTE, etc. will work just fine. This parameter is required.

TZ

The timezone to be used when evaluating any date math in router.start or router.interval. This is equivalent to the same parameter supplied to search queries, but understand in this case it's persisted with most of the other parameters as an alias property.

If GMT-4 is supplied for this value then a document dated 2018-01-14T21:00:00:01.2345Z would be stored in the myAlias\_2018-01-15\_01 collection (assuming an interval of +1HOUR).

The default timezone is UTC.

router.interval

A date math expression that will be appended to a timestamp to determine the next collection in the series. Any date math expression that can be evaluated if appended to a timestamp of the form 2018-01-15T16:17:18 will work here.

This parameter is required.

**router.maxFutureMs**

The maximum milliseconds into the future that a document is allowed to have in router.field for it to be accepted without error. If there was no limit, than an erroneous value could trigger many collections to be created.

The default is 10 minutes.

**router.preemptiveCreateMath**

A date math expression that results in early creation of new collections.

If a document arrives with a timestamp that is after the end time of the most recent collection minus this interval, then the next (and only the next) collection will be created asynchronously. Without this setting, collections are created synchronously when required by the document time stamp and thus block the flow of documents until the collection is created (possibly several seconds). Preemptive creation reduces these hiccups. If set to enough time (perhaps an hour or more) then if there are problems creating a collection, this window of time might be enough to take corrective action. However after a successful preemptive creation, the collection is consuming resources without being used, and new documents will tend to be routed through it only to be routed elsewhere. Also, note that router.autoDeleteAge is currently evaluated relative to the date of a newly created collection, and so you may want to increase the delete age by the preemptive window amount so that the oldest collection isn't deleted too soon. Note that it has to be possible to subtract the interval specified from a date, so if prepending a minus sign creates invalid date math, this will cause an error. Also note that a document that is itself destined for a collection that does not exist will still trigger synchronous creation up to that destination collection but will not trigger additional async preemptive creation. Only one type of collection creation can happen per document. Example: 90MINUTES.

This property is blank by default indicating just-in-time, synchronous creation of new collections.

**router.autoDeleteAge**

A date math expression that results in the oldest collections getting deleted automatically.

The date math is relative to the timestamp of a newly created collection (typically close to the current time), and thus this must produce an earlier time via rounding and/or subtracting. Collections to be deleted must have a time range that is entirely before the computed age. Collections are considered for deletion immediately prior to new collections getting created. Example: /DAY-90DAYS.

The default is not to delete.

**Category Routed Alias Parameters****router.maxCardinality**

The maximum number of categories allowed for this alias. This setting safeguards against the inadvertent creation of an infinite number of collections in the event of bad data.

**router.mustMatch**

A regular expression that the value of the field specified by router.field must match before a corresponding collection will be created. Note that changing this setting after data has been added will not alter the data already indexed. Any valid Java regular expression pattern may be specified. This expression is pre-compiled at the start of each request so batching of updates is strongly recommended. Overly complex patterns will produce cpu or garbage collecting overhead during indexing as determined by the JVM's implementation of regular expressions.

## CREATEALIAS Response

The output will simply be a responseHeader with details of the time it took to process the request. To confirm the creation of the alias, you can look in the Solr Admin UI, under the Cloud section and find the aliases.json file. The initial collection for routed aliases should also be visible in various parts of the admin UI.

### Examples using CREATEALIAS

#### Input

Create an alias named "testalias" and link it to the collections named "anotherCollection" and "testCollection".

```
http://localhost:8983/solr/admin/collections?action=CREATEALIAS&name=testalias&collections=anotherCollection,testCollection&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">122</int>
 </lst>
</response>
```

#### Input

Create an alias named "myTimeData" for data beginning on 2018-01-15 in the UTC time zone and partitioning daily based on the evt\_dt field in the incoming documents. Data more than one hour beyond the latest (most recent) partition is to be rejected and collections are created using a configset named "myConfig".

```
http://localhost:8983/solr/admin/collections?action=CREATEALIAS&name=myTimeData&router.start=NOW/DAY&router.field=evt_dt&router.name=time&router.interval=%2B1DAY&router.maxFutureMs=3600000&create-collection.collection.configName=myConfig&create-collection.numShards=2
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1234</int>
 </lst>
</response>
```

#### Input

A somewhat contrived example demonstrating the [V2 API](#) usage and additional collection creation options. Notice that the collection creation parameters follow the v2 API naming convention, not the v1 naming conventions.

```
POST /api/c
{
 "create-routed-alias" : {
 "name": "somethingTemporalThisWayComes",
 "router" : {
 "name": "time",
 "field": "evt_dt",
 "start": "NOW/MINUTE",
 "interval": "+2HOUR",
 "maxFutureMs": "14400000"
 },
 "create-collection" : {
 "config": "_default",
 "router": {
 "name": "implicit",
 "field": "foo_s"
 },
 "shards": "foo,bar,baz",
 "numShards": 3,
 "tlogReplicas": 1,
 "pullReplicas": 1,
 "maxShardsPerNode": 2,
 "properties" : {
 "foobar": "bazbam"
 }
 }
 }
}
```

## Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 1234
 }
}
```

## LISTALIASES: List of all aliases in the cluster

/admin/collections?action=LISTALIASES

The LISTALIASES action does not take any parameters.

## LISTALIASES Response

The output will contain a list of aliases with the corresponding collection names.

### Examples using LISTALIASES

#### Input

List the existing aliases, requesting information as XML from Solr:

```
http://localhost:8983/solr/admin/collections?action=LISTALIASES&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">0</int>
 </lst>
 <lst name="aliases">
 <str name="testalias1">collection1</str>
 <str name="testalias2">collection1,collection2</str>
 </lst>
 <lst name="properties">
 <lst name="testalias1"/>
 <lst name="testalias2">
 <str name="someKey">someValue</str>
 </lst>
 </lst>
</response>
```

## ALIASPROP: Modify Alias Properties for a Collection

The ALIASPROP action modifies the properties (metadata) on an alias. If a key is set with a value that is empty it will be removed.

```
/admin/collections?action=ALIASPROP&name=name&property .someKey=somevalue
```



This command allows you to revise any property. No alias specific validation is performed. Routed aliases may cease to function, function incorrectly or cause errors if property values are set carelessly.

### ALIASPROP Parameters

#### name

The alias name on which to set properties. This parameter is required.

#### property.\*

The name of the property to be modified replaces '\*', the value for the parameter is passed as the value for the property.

async

Request ID to track this action which will be [processed asynchronously](#).

## ALIASPROP Response

The output will simply be a responseHeader with details of the time it took to process the request. To confirm the creation of the property or properties, you can look in the Solr Admin UI, under the Cloud section and find the aliases.json file or use the LISTALIASES api command.

### Examples using ALIASPROP

#### Input

For an alias named "testalias2" and set the value "someValue" for a property of "someKey" and "otherValue" for "otherKey".

```
http://localhost:8983/solr/admin/collections?action=ALIASPROP&name=testalias2&property.someKey=someValue&property.otherKey=otherValue&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">122</int>
 </lst>
</response>
```

## DELETEALIAS: Delete a Collection Alias

/admin/collections?action=DELETEALIAS&name=name

### DELETEALIAS Parameters

name

The name of the alias to delete. This parameter is required.

async

Request ID to track this action which will be [processed asynchronously](#).

## DELETEALIAS Response

The output will simply be a responseHeader with details of the time it took to process the request. To confirm the removal of the alias, you can look in the Solr Admin UI, under the Cloud section, and find the aliases.json file.

### Examples using DELETEALIAS

#### Input

Remove the alias named "testalias".

```
http://localhost:8983/solr/admin/collections?action=DELETEALIAS&name=testalias&wt=xml
```

## Output

```
<response>
<lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">117</int>
</lst>
</response>
```

## DELETE: Delete a Collection

/admin/collections?action=DELETE&name=collection

### DELETE Parameters

name

The name of the collection to delete. This parameter is required.

async

Request ID to track this action which will be [processed asynchronously](#).

### DELETE Response

The response will include the status of the request and the cores that were deleted. If the status is anything other than "success", an error message will explain why the request failed.

### Examples using DELETE

#### Input

Delete the collection named "newCollection".

```
http://localhost:8983/solr/admin/collections?action=DELETE&name=newCollection&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">603</int>
 </lst>
 <lst name="success">
 <lst name="10.0.1.6:8983_solr">
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">19</int>
 </lst>
 </lst>
 <lst name="10.0.1.4:8983_solr">
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">67</int>
 </lst>
 </lst>
 </lst>
</response>
```

## DELETEREPLICA: Delete a Replica

Deletes a named replica from the specified collection and shard.

If the corresponding core is up and running the core is unloaded, the entry is removed from the clusterstate, and (by default) delete the instanceDir and dataDir. If the node/core is down, the entry is taken off the clusterstate and if the core comes up later it is automatically unregistered.

/admin/collections?action=DELETEREPLICA&collection=collection&shard=shard&replica=replica

### DELETEREPLICA Parameters

collection

The name of the collection. This parameter is required.

shard

The name of the shard that includes the replica to be removed. This parameter is required.

replica

The name of the replica to remove.

If count is used instead, this parameter is not required. Otherwise, this parameter must be supplied.

count

The number of replicas to remove. If the requested number exceeds the number of replicas, no replicas will be deleted. If there is only one replica, it will not be removed.

If replica is used instead, this parameter is not required. Otherwise, this parameter must be supplied.

deleteInstanceDir

By default Solr will delete the entire instanceDir of the replica that is deleted. Set this to false to prevent the instance directory from being deleted.

#### deleteDataDir

By default Solr will delete the dataDir of the replica that is deleted. Set this to false to prevent the data directory from being deleted.

#### deleteIndex

By default Solr will delete the index of the replica that is deleted. Set this to false to prevent the index directory from being deleted.

#### onlyIfDown

When set to true, no action will be taken if the replica is active. Default false.

#### async

Request ID to track this action which will be [processed asynchronously](#).

## Examples using DELETEREPLICA

### Input

```
http://localhost:8983/solr/admin/collections?action=DELETEREPLICA&collection=test2&shard=shard2&r
eplica=core_node3&wt=xml
```

### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">110</int>
 </lst>
</response>
```

## ADDREPLICA: Add Replica

Add one or more replicas to a shard in a collection. The node name can be specified if the replica is to be created in a specific node. Otherwise, a set of nodes can be specified and the most suitable ones among them will be chosen to create the replica(s).

The API uses the Autoscaling framework to find nodes that can satisfy the disk requirements for the new replica(s) but only when an Autoscaling preferences or policy is configured. Refer to [Autoscaling Policy and Preferences](#) section for more details.

```
/admin/collections?action=ADDREPLICA&collection=collection&shard=shard&node=nodeName
```

## ADDREPLICA Parameters

#### collection

The name of the collection where the replica should be created. This parameter is required.

**shard**

The name of the shard to which replica is to be added.

If shard is not specified, then `_route_` must be.

**\_route\_**

If the exact shard name is not known, users may pass the `_route_` value and the system would identify the name of the shard.

Ignored if the shard parameter is also specified.

**node**

The name of the node where the replica should be created (optional).

**createNodeSet**

A comma-separated list of nodes among which the best ones will be chosen to place the replicas (optional)

The format is a comma-separated list of `node_names`, such as

`localhost:8983_solr,localhost:8984_solr,localhost:8985_solr`.



If neither `node` nor `createNodeSet` are specified then the best node(s) from among all the live nodes in the cluster are chosen.

**instanceDir**

The `instanceDir` for the core that will be created.

**dataDir**

The directory in which the core should be created.

**type**

The type of replica to create. These possible values are allowed:

- `nrt`: The NRT type maintains a transaction log and updates its index locally. This is the default and the most commonly used.
- `tlog`: The TLOG type maintains a transaction log but only updates its index via replication.
- `pull`: The PULL type does not maintain a transaction log and only updates its index via replication. This type is not eligible to become a leader.

See the section [Types of Replicas](#) for more information about replica type options.

**nrtReplicas**

The number of `nrt` replicas that should be created (optional, defaults to 1 if type is `nrt` otherwise 0).

**tlogReplicas**

The number of `tlog` replicas that should be created (optional, defaults to 1 if type is `tlog` otherwise 0).

**pullReplicas**

The number of `pull` replicas that should be created (optional, defaults to 1 if type is `pull` otherwise 0).

property .name=value

Set core property *name* to *value*. See [Defining core.properties](#) for details about supported properties and values.

waitForFinalState

If true, the request will complete only when all affected replicas become active. The default is false, which means that the API will return the status of the single action, which may be before the new replica is online and active.

async

Request ID to track this action which will be [processed asynchronously](#)

## Examples using ADDREPLICA

### Input

Create a replica for the "test" collection on the node "192.167.1.2:8983\_solr".

```
http://localhost:8983/solr/admin/collections?action=ADDREPLICA&collection=test2&shard=shard2&node=192.167.1.2:8983_solr&wt=xml
```

### Output

```
<response>
<lst name="responseHeader">
<int name="status">0</int>
<int name="QTime">3764</int>
</lst>
<lst name="success">
<lst>
<lst name="responseHeader">
<int name="status">0</int>
<int name="QTime">3450</int>
</lst>
<str name="core">test2_shard2_replica4</str>
</lst>
</lst>
</response>
```

### Input

Create a replica for the "gettingstarted" collection with one PULL replica and one TLOG replica.

```
http://localhost:8983/solr/admin/collections?action=addreplica&collection=gettingstarted&shard=shard1&tlogReplicas=1&pullReplicas=1
```

### Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 784
 },
 "success": {
 "127.0.1.1:7574_solr": {
 "responseHeader": {
 "status": 0,
 "QTime": 257
 },
 "core": "gettingstarted_shard1_replica_p11"
 },
 "127.0.1.1:8983_solr": {
 "responseHeader": {
 "status": 0,
 "QTime": 295
 },
 "core": "gettingstarted_shard1_replica_t10"
 }
 }
}
```

## CLUSTERPROP: Cluster Properties

Add, edit or delete a cluster-wide property.

`/admin/collections?action=CLUSTERPROP&name=propertyName&val=propertyValue`

### CLUSTERPROP Parameters

**name**

The name of the property. Supported properties names are `autoAddReplicas`, `legacyCloud`, `location`, `maxCoresPerNode` and `urlScheme`. Other properties can be set (for example, if you need them for custom plugins) but they must begin with the prefix `ext..`. Unknown properties that don't begin with `ext.` will be rejected.

**val**

The value of the property. If the value is empty or null, the property is unset.

### CLUSTERPROP Response

The response will include the status of the request and the properties that were updated or removed. If the status is anything other than "0", an error message will explain why the request failed.

### Examples using CLUSTERPROP

#### Input

```
http://localhost:8983/solr/admin/collections?action=CLUSTERPROP&name=urlScheme&val=https&wt=xml
```

## Output

```
<response>
<lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">0</int>
</lst>
</response>
```

## Setting Cluster-Wide Defaults

It is possible to set cluster-wide default values for certain attributes of a collection, using the `defaults` parameter.

### Set/update default values

```
curl -X POST -H 'Content-type:application/json' --data-binary '
{
 "set-obj-property": {
 "defaults" : {
 "collection": {
 "numShards": 2,
 "nrtReplicas": 1,
 "tlogReplicas": 1,
 "pullReplicas": 1
 }
 }
 }
}' http://localhost:8983/api/cluster
```

### Unset the only value of nrtReplicas

```
curl -X POST -H 'Content-type:application/json' --data-binary '
{
 "set-obj-property": {
 "defaults" : {
 "collection": {
 "nrtReplicas": null
 }
 }
 }
}' http://localhost:8983/api/cluster
```

### Unset all values in defaults

```
curl -X POST -H 'Content-type:application/json' --data-binary '
{ "set-obj-property" : {
 "defaults" : null
}}' http://localhost:8983/api/cluster
```

 Until Solr 7.5, cluster properties supported a collectionDefaults key which is now deprecated and replaced with defaults. Using the collectionDefaults parameter in Solr 7.4 or 7.5 will continue to work but the format of the properties will automatically be converted to the new nested structure. Support for the "collectionDefaults" key will be removed in Solr 9.

## COLLECTIONPROP: Collection Properties

Add, edit or delete a collection property.

```
/admin/collections?action=COLLECTIONPROP&name=collectionName&propertyName=propertyName&pro
pertyValue=propertyValue
```

### COLLECTIONPROP Parameters

name

The name of the collection for which the property would be set.

propertyName

The name of the property.

propertyValue

The value of the property. When not provided, the property is deleted.

### COLLECTIONPROP Response

The response will include the status of the request and the properties that were updated or removed. If the status is anything other than "0", an error message will explain why the request failed.

### Examples using COLLECTIONPROP

#### Input

```
http://localhost:8983/solr/admin/collections?action=COLLECTIONPROP&name=coll&propertyName=foo&pro
pertyValue=bar&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">0</int>
 </lst>
</response>
```

## COLSTATUS: Detailed Status of a Collection's Indexes

The COLSTATUS command provides a detailed description of the collection status, including low-level index information about segments and field data.

This command also checks the compliance of Lucene index field types with the current Solr collection schema and indicates the names of non-compliant fields, i.e., Lucene fields with field types incompatible (or different) from the corresponding Solr field types declared in the current schema. Such incompatibilities may result from incompatible schema changes or after migration of data to a different major Solr release.

```
/admin/collections?action=COLSTATUS&collection=coll&coreInfo=true&segments=true&fieldInfo=true&sizeInfo=true
```

### COLSTATUS Parameters

#### collection

Collection name (optional). If missing then it means all collections.

#### coreInfo

Optional boolean. If true then additional information will be provided about SolrCore of shard leaders.

#### segments

Optional boolean. If true then segment information will be provided.

#### fieldInfo

Optional boolean. If true then detailed Lucene field information will be provided and their corresponding Solr schema types.

#### sizeInfo

Optional boolean. If true then additional information about the index files size and their RAM usage will be provided.

### COLSTATUS Response

The response will include an overview of the collection status, the number of active or inactive shards and replicas, and additional index information of shard leaders.

### Examples using COLSTATUS

#### Input

```
http://localhost:8983/solr/admin/collections?action=COLSTATUS&collection=gettingstarted&fieldInfo=true&sizeInfo=true
```

## Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 50
 },
 "gettingstarted": {
 "stateFormat": 2,
 "znodeVersion": 16,
 "properties": {
 "autoAddReplicas": "false",
 "maxShardsPerNode": "-1",
 "nrtReplicas": "2",
 "pullReplicas": "0",
 "replicationFactor": "2",
 "router": {
 "name": "compositeId"
 },
 "tlogReplicas": "0"
 },
 "activeShards": 2,
 "inactiveShards": 0,
 "schemaNonCompliant": [
 "(NONE)"
],
 "shards": {
 "shard1": {
 "state": "active",
 "range": "80000000-ffffffff",
 "replicas": {
 "total": 2,
 "active": 2,
 "down": 0,
 "recovering": 0,
 "recovery_failed": 0
 },
 "leader": {
 "coreNode": "core_node4",
 "core": "gettingstarted_shard1_replica_n1",
 "base_url": "http://192.168.0.80:8983/solr",
 "node_name": "192.168.0.80:8983_solr",
 "state": "active",
 "type": "NRT",
 "force_set_state": "false",
 "leader": "true",
 "segInfos": {
 "info": {
 "minSegmentLuceneVersion": "9.0.0",
 "commitLuceneVersion": "9.0.0",
 "numSegments": 40,
 "lastCommit": "2019-06-11T10:45:00Z",
 "lastMerge": "2019-06-11T10:45:00Z",
 "lastOptimize": "2019-06-11T10:45:00Z",
 "lastRecover": "2019-06-11T10:45:00Z",
 "lastUpdate": "2019-06-11T10:45:00Z",
 "lastVersion": 16
 }
 }
 }
 }
 }
 }
}
```

```
"segmentsFileName": "segments_w",
"totalMaxDoc": 686953,
"userData": {
 "commitCommandVer": "1627350608019193856",
 "commitTimeMSec": "1551962478819"
},
"fieldInfoLegend": [
 "I - Indexed",
 "D - DocValues",
 "xxx - DocValues type",
 "V - TermVector Stored",
 "O - Omit Norms",
 "F - Omit Term Frequencies & Positions",
 "P - Omit Positions",
 "H - Store Offsets with Positions",
 "p - field has payloads",
 "s - field uses soft deletes",
 ":x:x:x - point data dim : index dim : num bytes"
],
"segments": {
 "_i": {
 "name": "_i",
 "delCount": 738,
 "softDelCount": 0,
 "hasFieldUpdates": false,
 "sizeInBytes": 109398213,
 "size": 70958,
 "age": "2019-03-07T12:34:24.761Z",
 "source": "merge",
 "version": "9.0.0",
 "createdVersionMajor": 9,
 "minVersion": "9.0.0",
 "diagnostics": {
 "os": "Mac OS X",
 "java.vendor": "Oracle Corporation",
 "java.version": "1.8.0_191",
 "java.vm.version": "25.191-b12",
 "lucene.version": "9.0.0",
 "mergeMaxNumSegments": "-1",
 "os.arch": "x86_64",
 "java.runtime.version": "1.8.0_191-b12",
 "source": "merge",
 "mergeFactor": "10",
 "os.version": "10.14.3",
 "timestamp": "1551962064761"
 },
 "attributes": {
 "Lucene50StoredFieldsFormat.mode": "BEST_SPEED"
 },
 "largestFiles": {
 "_i.fdt": "42.5 MB",

```

```
_i_Lucene80_0.dvd": "35.3 MB",
"_i_Lucene50_0.pos": "11.1 MB",
"_i_Lucene50_0.doc": "10 MB",
"_i_Lucene50_0.tim": "4.3 MB"
},
"ramBytesUsed": {
 "total": 49153,
 "postings [PerFieldPostings(segment=_i formats=1)": {
 "total": 31023,
 "fields": {
 "dc": {
 "flags": "I-----",
 "schemaType": "text_general"
 },
 "dc_str": {
 "flags": "-Dsrs-----",
 "schemaType": "strings"
 },
 "dc.title": {
 "flags": "I-----",
 "docCount": 70958,
 "sumDocFreq": 646756,
 "sumTotalTermFreq": 671817,
 "schemaType": "text_general"
 },
 "dc.date": {
 "flags": "-Dsrn-----:1:1:8",
 "schemaType": "pdates"
 }
 }}}}}}}}}}}}
```

## MIGRATE: Migrate Documents to Another Collection

/admin/collections?action=MIGRATE&collection=name&split.key=key1!&target.collection=target\_collection&forward.timeout=60

The MIGRATE command is used to migrate all documents having a given routing key to another collection. The source collection will continue to have the same data as-is but it will start re-routing write requests to the target collection for the number of seconds specified by the forward.timeout parameter. It is the responsibility of the user to switch to the target collection for reads and writes after the MIGRATE action completes.

The routing key specified by the split.key parameter may span multiple shards on both the source and the target collections. The migration is performed shard-by-shard in a single thread. One or more temporary collections may be created by this command during the ‘migrate’ process but they are cleaned up at the end automatically.

This is a long running operation and therefore using the async parameter is highly recommended. If the async parameter is not specified then the operation is synchronous by default and keeping a large read timeout on the invocation is advised. Even with a large read timeout, the request may still timeout but that doesn’t necessarily mean that the operation has failed. Users should check logs, cluster state, source and

target collections before invoking the operation again.

This command works only with collections using the compositeId router. The target collection must not receive any writes during the time the MIGRATE command is running otherwise some writes may be lost.

Please note that the MIGRATE API does not perform any de-duplication on the documents so if the target collection contains documents with the same uniqueKey as the documents being migrated then the target collection will end up with duplicate documents.

## MIGRATE Parameters

### collection

The name of the source collection from which documents will be split. This parameter is required.

### target.collection

The name of the target collection to which documents will be migrated. This parameter is required.

### split.key

The routing key prefix. For example, if the uniqueKey of a document is "a!123", then you would use split.key=a!. This parameter is required.

### forward.timeout

The timeout, in seconds, until which write requests made to the source collection for the given split.key will be forwarded to the target shard. The default is 60 seconds.

### property.name=value

Set core property *name* to *value*. See the section [Defining core.properties](#) for details on supported properties and values.

### async

Request ID to track this action which will be [processed asynchronously](#).

## MIGRATE Response

The response will include the status of the request.

## Examples using MIGRATE

### Input

```
http://localhost:8983/solr/admin/collections?action=MIGRATE&collection=test1&split.key=a!&target.collection=test2&wt=xml
```

### Output

```
<response>
<lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">19014</int>
</lst>
<lst name="success">
```

```
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1</int>
 </lst>
 <str name="core">test2_shard1_0_replica1</str>
 <str name="status">BUFFERING</str>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">2479</int>
 </lst>
 <str name="core">split_shard1_0_temp_shard1_0_shard1_replica1</str>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1002</int>
 </lst>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">21</int>
 </lst>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1655</int>
 </lst>
 <str name="core">split_shard1_0_temp_shard1_0_shard1_replica2</str>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">4006</int>
 </lst>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">17</int>
 </lst>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1</int>
 </lst>
 <str name="core">test2_shard1_0_replica1</str>
```

```
<str name="status">EMPTY_BUFFER</str>
</lst>
<lst name="192.168.43.52:8983_solr">
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">31</int>
 </lst>
</lst>
<lst name="192.168.43.52:8983_solr">
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">31</int>
 </lst>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1</int>
 </lst>
</lst>
<str name="core">test2_shard1_1_replica1</str>
<str name="status">BUFFERING</str>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1742</int>
 </lst>
 <str name="core">split_shard1_1_temp_shard1_1_shard1_replica1</str>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1002</int>
 </lst>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">15</int>
 </lst>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1917</int>
 </lst>
 <str name="core">split_shard1_1_temp_shard1_1_shard1_replica2</str>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">5007</int>
```

```
</lst>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">8</int>
 </lst>
</lst>
<lst>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1</int>
 </lst>
 <str name="core">test2_shard1_1_replica1</str>
 <str name="status">EMPTY_BUFFER</str>
</lst>
<lst name="192.168.43.52:8983_solr">
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">30</int>
 </lst>
</lst>
<lst name="192.168.43.52:8983_solr">
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">30</int>
 </lst>
</lst>
</lst>
</response>
```

## ADDRole: Add a Role

/admin/collections?action=ADDRole&role=roleName&node=nodeName

Assigns a role to a given node in the cluster. The only supported role is overseer.

Use this command to dedicate a particular node as Overseer. Invoke it multiple times to add more nodes. This is useful in large clusters where an Overseer is likely to get overloaded. If available, one among the list of nodes which are assigned the 'overseer' role would become the overseer. The system would assign the role to any other node if none of the designated nodes are up and running.

### ADDRole Parameters

#### role

The name of the role. The only supported role as of now is overseer. This parameter is required.

#### node

The name of the node that will be assigned the role. It is possible to assign a role even before that node is started. This parameter is started.

## ADDOROLE Response

The response will include the status of the request and the properties that were updated or removed. If the status is anything other than "0", an error message will explain why the request failed.

### Examples using ADDROLE

#### Input

```
http://localhost:8983/solr/admin/collections?action=ADDOROLE&role=overseer&node=192.167.1.2:8983_s
olr&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">0</int>
 </lst>
</response>
```

## REMOVEROLE: Remove Role

Remove an assigned role. This API is used to undo the roles assigned using ADDROLE operation

```
/admin/collections?action=REMOVEROLE&role=roleName&node=nodeName
```

### REMOVEROLE Parameters

#### role

The name of the role. The only supported role as of now is overseer. This parameter is required.

#### node

The name of the node where the role should be removed.

### REMOVEROLE Response

The response will include the status of the request and the properties that were updated or removed. If the status is anything other than "0", an error message will explain why the request failed.

### Examples using REMOVEROLE

#### Input

```
http://localhost:8983/solr/admin/collections?action=REMOVEROLE&role=overseer&node=192.167.1.2:898
3_solr&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">0</int>
 </lst>
</response>
```

## OVERSEERSTATUS: Overseer Status and Statistics

Returns the current status of the overseer, performance statistics of various overseer APIs, and the last 10 failures per operation type.

/admin/collections?action=OVERSEERSTATUS

### Examples using OVERSEERSTATUS

#### Input:

```
http://localhost:8983/solr/admin/collections?action=OVERSEERSTATUS
```

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 33,
 "leader": "127.0.1.1:8983_solr",
 "overseer_queue_size": 0,
 "overseer_work_queue_size": 0,
 "overseer_collection_queue_size": 2,
 "overseer_operations": [
 "createcollection", {
 "requests": 2,
 "errors": 0,
 "avgRequestsPerSecond": 0.7467088842794136,
 "5minRateRequestsPerSecond": 7.525069023276674,
 "15minRateRequestsPerSecond": 10.271274280947182,
 "avgTimePerRequest": 0.5050685,
 "medianRequestTime": 0.5050685,
 "75thPcRequestTime": 0.519016,
 "95thPcRequestTime": 0.519016,
 "99thPcRequestTime": 0.519016,
 "999thPcRequestTime": 0.519016},
 "removeshard", {
 "..."},
],
 "collection_operations": [
 "splitshard", {
 "requests": 1,
 "errors": 1,
 "recent_failures": [{"
```

```
"request":{
 "operation":"splitshard",
 "shard":"shard2",
 "collection":"example1"},
 "response": [
 "Operation splitshard caused exception:",
 "org.apache.solr.common.SolrException:org.apache.solr.common.SolrException: No shard with the
 specified name exists: shard2",
 "exception", {
 "msg": "No shard with the specified name exists: shard2",
 "rspCode": 400}],
 "avgRequestsPerSecond": 0.8198143044809885,
 "5minRateRequestsPerSecond": 8.043840552427673,
 "15minRateRequestsPerSecond": 10.502079828515368,
 "avgTimePerRequest": 2952.7164175,
 "medianRequestTime": 2952.7164175000003,
 "75thPcRequestTime": 5904.384052,
 "95thPcRequestTime": 5904.384052,
 "99thPcRequestTime": 5904.384052,
 "999thPcRequestTime": 5904.384052},
 "..."
],
 "overseer_queue": [
 "..."
],
 "..."
}
```

## CLUSTERSTATUS: Cluster Status

Fetch the cluster status including collections, shards, replicas, configuration name as well as collection aliases and cluster properties.

/admin/collections?action=CLUSTERSTATUS

### CLUSTERSTATUS Parameters

collection

The collection or alias name for which information is requested. If omitted, information on all collections in the cluster will be returned. If an alias is supplied, information on the collections in the alias will be returned.

shard

The shard(s) for which information is requested. Multiple shard names can be specified as a comma-separated list.

\_route\_

This can be used if you need the details of the shard where a particular document belongs to and you don't know which shard it falls under.

## CLUSTERSTATUS Response

The response will include the status of the request and the status of the cluster.

### Examples using CLUSTERSTATUS

#### Input

```
http://localhost:8983/solr/admin/collections?action=CLUSTERSTATUS
```

#### Output

```
{
 "responseHeader":{
 "status":0,
 "QTime":333},
 "cluster":{
 "collections":{
 "collection1":{
 "shards":{
 "shard1":{
 "range": "80000000-ffffffff",
 "state": "active",
 "replicas":{
 "core_node1":{
 "state": "active",
 "core": "collection1",
 "node_name": "127.0.1.1:8983_solr",
 "base_url": "http://127.0.1.1:8983/solr",
 "leader": "true"},
 "core_node3":{
 "state": "active",
 "core": "collection1",
 "node_name": "127.0.1.1:8900_solr",
 "base_url": "http://127.0.1.1:8900/solr"}}},
 "shard2":{
 "range": "0-7fffffff",
 "state": "active",
 "replicas":{
 "core_node2":{
 "state": "active",
 "core": "collection1",
 "node_name": "127.0.1.1:7574_solr",
 "base_url": "http://127.0.1.1:7574/solr",
 "leader": "true"},
 "core_node4":{
 "state": "active",
 "core": "collection1",
 "node_name": "127.0.1.1:7500_solr",
 "base_url": "http://127.0.1.1:7500/solr"}}},
 "maxShardsPerNode": "1",
 }
 }
 }
}
```

```
"router": {"name": "compositeId"},
"replicationFactor": "1",
"znodeVersion": 11,
"autoCreated": "true",
"configName": "my_config",
"aliases": ["both_collections"]
},
"collection2": {
 "..."
}
},
"aliases": { "both_collections": "collection1,collection2" },
"roles": {
 "overseer": [
 "127.0.1.1:8983_solr",
 "127.0.1.1:7574_solr"]
},
"live_nodes": [
 "127.0.1.1:7574_solr",
 "127.0.1.1:7500_solr",
 "127.0.1.1:8983_solr",
 "127.0.1.1:8900_solr"]
}
}
}
```

## REQUESTSTATUS: Request Status of an Async Call

Request the status and response of an already submitted [Asynchronous Collection API](#) (below) call. This call is also used to clear up the stored statuses.

/admin/collections?action=REQUESTSTATUS&requestid=request-id

### REQUESTSTATUS Parameters

requestid

The user defined request ID for the request. This can be used to track the status of the submitted asynchronous task. This parameter is required.

### Examples using REQUESTSTATUS

#### Input: Valid Request ID

```
http://localhost:8983/solr/admin/collections?action=REQUESTSTATUS&requestid=1000&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1</int>
 </lst>
 <lst name="status">
 <str name="state">completed</str>
 <str name="msg">found 1000 in completed tasks</str>
 </lst>
</response>
```

### Input: Invalid Request ID

```
http://localhost:8983/solr/admin/collections?action=REQUESTSTATUS&requestid=1004&wt=xml
```

### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1</int>
 </lst>
 <lst name="status">
 <str name="state">notfound</str>
 <str name="msg">Did not find taskid [1004] in any tasks queue</str>
 </lst>
</response>
```

## DELETESTATUS: Delete Status

Deletes the stored response of an already failed or completed [Asynchronous Collection API](#) call.

```
/admin/collections?action=DELETESTATUS&requestid=request-id
```

### DELETESTATUS Parameters

requestid

The request ID of the asynchronous call whose stored response should be cleared.

flush

Set to true to clear all stored completed and failed async request responses.

### Examples using DELETESTATUS

#### Input: Valid Request ID

```
http://localhost:8983/solr/admin/collections?action=DELETESTATUS&requestid=foo&wt=xml
```

## Output

```
<response>
<lst name="responseHeader">
<int name="status">0</int>
<int name="QTime">1</int>
</lst>
<str name="status">successfully removed stored response for [foo]</str>
</response>
```

## Input: Invalid Request ID

```
http://localhost:8983/solr/admin/collections?action=DELETESTATUS&requestid=bar&wt=xml
```

## Output

```
<response>
<lst name="responseHeader">
<int name="status">0</int>
<int name="QTime">1</int>
</lst>
<str name="status">[bar] not found in stored responses</str>
</response>
```

## Input: Clear All Stored Statuses

```
http://localhost:8983/solr/admin/collections?action=DELETESTATUS&flush=true&wt=xml
```

## Output

```
<response>
<lst name="responseHeader">
<int name="status">0</int>
<int name="QTime">1</int>
</lst>
<str name="status"> successfully cleared stored collection api responses </str>
</response>
```

## LIST: List Collections

Fetch the names of the collections in the cluster.

```
/admin/collections?action=LIST
```

### Examples using LIST

#### Input

```
http://localhost:8983/solr/admin/collections?action=LIST
```

## Output

```
{
 "responseHeader":{
 "status":0,
 "QTime":2011},
 "collections":["collection1",
 "example1",
 "example2"]}
```

## ADDREPLICAPROP: Add Replica Property

Assign an arbitrary property to a particular replica and give it the value specified. If the property already exists, it will be overwritten with the new value.

```
/admin/collections?action=ADDREPLICAPROP&collection=collectionName&shard=shardName&replica=replicaName&property=propertyName&property.value=value
```

### ADDREPLICAPROP Parameters

**collection**

The name of the collection the replica belongs to. This parameter is required.

**shard**

The name of the shard the replica belongs to. This parameter is required.

**replica**

The replica, e.g., core\_node1. This parameter is required.

**property**

The name of the property to add. This property is required.

This will have the literal property . prepended to distinguish it from system-maintained properties. So these two forms are equivalent:

property=special

and

property=property.special

**property.value**

The value to assign to the property. This parameter is required.

**shardUnique**

If true, then setting this property in one replica will remove the property from all other replicas in that shard. The default is false.

There is one pre-defined property preferredLeader for which shardUnique is forced to true and an error

returned if shardUnique is explicitly set to false.

PreferredLeader is a boolean property. Any value assigned that is not equal (case insensitive) to true will be interpreted as false for preferredLeader.

## ADDREPLICAPROP Response

The response will include the status of the request. If the status is anything other than "0", an error message will explain why the request failed.

### Examples using ADDREPLICAPROP

#### Input

This command would set the "preferredLeader" property (property.preferredLeader) to "true" on "core\_node1", and remove that property from any other replica in the shard.

```
http://localhost:8983/solr/admin/collections?action=ADDREPLICAPROP&shard=shard1&collection=collection1&replica=core_node1&property=preferredLeader&property.value=true&wt=xml
```

#### Output

```
<response>
<lst name="responseHeader">
<int name="status">0</int>
<int name="QTime">46</int>
</lst>
</response>
```

#### Input

This pair of commands will set the "testprop" property (property.testprop) to 'value1' and 'value2' respectively for two nodes in the same shard.

```
http://localhost:8983/solr/admin/collections?action=ADDREPLICAPROP&shard=shard1&collection=collection1&replica=core_node1&property=testprop&property.value=value1
```

```
http://localhost:8983/solr/admin/collections?action=ADDREPLICAPROP&shard=shard1&collection=collection1&replica=core_node3&property=property.testprop&property.value=value2
```

#### Input

This pair of commands would result in "core\_node\_3" having the "testprop" property (property.testprop) value set because the second command specifies shardUnique=true, which would cause the property to be removed from "core\_node\_1".

```
http://localhost:8983/solr/admin/collections?action=ADDREPLICAPROP&shard=shard1&collection=collection1&replica=core_node1&property=testprop&property.value=value1
```

```
http://localhost:8983/solr/admin/collections?action=ADDREPLICAPROP&shard=shard1&collection=collection1&replica=core_node3&property=testprop&property.value=value2&shardUnique=true
```

## DELETEREPLICAPROP: Delete Replica Property

Deletes an arbitrary property from a particular replica.

```
/admin/collections?action=DELETEREPLICAPROP&collection=collectionName&shard=shardName&replica=replicaName&property=propertyName
```

### DELETEREPLICAPROP Parameters

**collection**

The name of the collection the replica belongs to. This parameter is required.

**shard**

The name of the shard the replica belongs to. This parameter is required.

**replica**

The replica, e.g., core\_node1. This parameter is required.

**property**

The property to add. This will have the literal property . prepended to distinguish it from system-maintained properties. So these two forms are equivalent:

property=special

and

property=property.special

### DELETEREPLICAPROP Response

The response will include the status of the request. If the status is anything other than "0", an error message will explain why the request failed.

### Examples using DELETEREPLICAPROP

#### Input

This command would delete the preferredLeader (property.preferredLeader) from core\_node1.

```
http://localhost:8983/solr/admin/collections?action=DELETEREPLICAPROP&shard=shard1&collection=collection1&replica=core_node1&property=preferredLeader&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">9</int>
 </lst>
</response>
```

## BALANCESHARDUNIQUE: Balance a Property Across Nodes

/admin/collections?action=BALANCESHARDUNIQUE&collection=collectionName&property=propertyName

Insures that a particular property is distributed evenly amongst the physical nodes that make up a collection. If the property already exists on a replica, every effort is made to leave it there. If the property is **not** on any replica on a shard, one is chosen and the property is added.

### BALANCESHARDUNIQUE Parameters

**collection**

The name of the collection to balance the property in. This parameter is required.

**property**

The property to balance. The literal property . is prepended to this property if not specified explicitly. This parameter is required.

**onlyactivenodes**

Defaults to true. Normally, the property is instantiated on active nodes only. If this parameter is specified as false, then inactive nodes are also included for distribution.

**shardUnique**

Something of a safety valve. There is one pre-defined property (preferredLeader) that defaults this value to true. For all other properties that are balanced, this must be set to true or an error message will be returned.

### BALANCESHARDUNIQUE Response

The response will include the status of the request. If the status is anything other than "0", an error message will explain why the request failed.

### Examples using BALANCESHARDUNIQUE

#### Input

Either of these commands would put the "preferredLeader" property on one replica in every shard in the "collection1" collection.

```
http://localhost:8983/solr/admin/collections?action=BALANCESHARDUNIQUE&collection=collection1&property=preferredLeader&wt=xml
```

```
http://localhost:8983/solr/admin/collections?action=BALANCESHARDUNIQUE&collection=collection1&property=property.preferredLeader&wt=xml
```

## Output

```
<response>
<lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">9</int>
</lst>
</response>
```

Examining the clusterstate after issuing this call should show exactly one replica in each shard that has this property.

## REBALANCELEADERS: Rebalance Leaders

Reassigns leaders in a collection according to the preferredLeader property across active nodes.

/admin/collections?action=REBALANCELEADERS&collection=collectionName

Leaders are assigned in a collection according to the preferredLeader property on active nodes. This command should be run after the preferredLeader property has been assigned via the BALANCESHARDUNIQUE or ADDREPLICAPROP commands.



It is not *required* that all shards in a collection have a preferredLeader property. Rebalancing will only attempt to reassign leadership to those replicas that have the preferredLeader property set to true *and* are not currently the shard leader *and* are currently active.

### REBALANCELEADERS Parameters

collection

The name of the collection to rebalance preferredLeaders on. This parameter is required.

maxAtOnce

The maximum number of reassignments to have queue up at once. Values <=0 are use the default value Integer.MAX\_VALUE.

When this number is reached, the process waits for one or more leaders to be successfully assigned before adding more to the queue.

maxWaitSeconds

Defaults to 60. This is the timeout value when waiting for leaders to be reassigned. If maxAtOnce is less than the number of reassignments that will take place, this is the maximum interval that any *single* wait for at least one reassignment.

For example, if 10 reassignments are to take place and maxAtOnce is 1 and maxWaitSeconds is 60, the upper bound on the time that the command may wait is 10 minutes.

### REBALANCELEADERS Response

The response will include the status of the request. A status of "0" indicates the request was *processed*, not that all assignments were successful. Examine the "Summary" section for that information.

## Examples using REBALANCELEADERS

### Input

Either of these commands would cause all the active replicas that had the preferredLeader property set and were *not* already the preferred leader to become leaders.

```
http://localhost:8983/solr/admin/collections?action=REBALANCELEADERS&collection=collection1&wt=json
http://localhost:8983/solr/admin/collections?action=REBALANCELEADERS&collection=collection1&maxAtOnce=5&maxWaitSeconds=30&wt=json
```

### Output

In this example:

- In the "alreadyLeaders" section, core\_node5 was already the leader, so there were no changes in leadership for shard1.
- In the "inactivePreferreds" section, core\_node57 had the preferredLeader property set, but the node was not active, the leader for shard7 was not changed. This is considered successful.
- In the "successes" section, core\_node23 was *not* the leader for shard3, so leadership was assigned to that replica.

The "Summary" section with the "Success" tag indicates that the command rebalanced all *active* replicas with the preferredLeader property set as required. If a replica cannot be made leader due to not being healthy (for example, it is on a Solr instance that is not running), it's also considered success.

```
{
 "responseHeader":{
 "status":0,
 "QTime":3054},
 "Summary":{
 "Success":"All active replicas with the preferredLeader property set are leaders"},
 "alreadyLeaders":{
 "core_node5":{
 "status":"skipped",
 "msg":"Replica core_node5 is already the leader for shard shard1. No change necessary"}},
 "inactivePreferreds":{
 "core_node57":{
 "status":"skipped",
 "msg":"Replica core_node57 is a referredLeader for shard shard7, but is inactive. No change necessary"}},
 "successes":{
 "shard3":{
 "status":"success",
 "msg":"Successfully changed leader of slice shard3 to core_node23"}}}
```

Examining the clusterstate after issuing this call should show that every active replica that has the

preferredLeader property should also have the "leader" property set to *true*.



The added work done by an NRT leader is quite small and only present when indexing. The primary use-case is to redistribute the leader role if there are a large number of leaders concentrated on a small number of nodes. Rebalancing will likely not improve performance unless the imbalance of leadership roles is measured in multiples of 10.



The BALANCESHARDUNIQUE command that distributes the preferredLeader property does not guarantee perfect distribution and in some collection topologies it is impossible to make that guarantee.

## FORCELEADER: Force Shard Leader

In the unlikely event of a shard losing its leader, this command can be invoked to force the election of a new leader.

```
/admin/collections?action=FORCELEADER&collection=<collectionName>&shard=<shardName>
```

### FORCELEADER Parameters

#### collection

The name of the collection. This parameter is required.

#### shard

The name of the shard where leader election should occur. This parameter is required.



This is an expert level command, and should be invoked only when regular leader election is not working. This may potentially lead to loss of data in the event that the new leader doesn't have certain updates, possibly recent ones, which were acknowledged by the old leader before going down.

## MIGRATESTATEFORMAT: Migrate Cluster State

A expert level utility API to move a collection from shared clusterstate.json ZooKeeper node (created with stateFormat=1, the default in all Solr releases prior to 5.0) to the per-collection state.json stored in ZooKeeper (created with stateFormat=2, the current default) seamlessly without any application down-time.

```
/admin/collections?action=MIGRATESTATEFORMAT&collection=<collection_name>
```

### MIGRATESTATEFORMAT Parameters

#### collection

The name of the collection to be migrated from clusterstate.json to its own state.json ZooKeeper node. This parameter is required.

#### async

Request ID to track this action which will be [processed asynchronously](#).

This API is useful in migrating any collections created prior to Solr 5.0 to the more scalable cluster state format now used by default. If a collection was created in any Solr 5.x version or higher, then executing this command is not necessary.

## BACKUP: Backup Collection

Backs up Solr collections and associated configurations to a shared filesystem - for example a Network File System.

```
/admin/collections?action=BACKUP&name=myBackupName&collection=myCollectionName&location=/path/to/my/shared/drive
```

The BACKUP command will backup Solr indexes and configurations for a specified collection. The BACKUP command takes one copy from each shard for the indexes. For configurations, it backs up the configset that was associated with the collection and metadata.

### BACKUP Parameters

**collection**

The name of the collection to be backed up. This parameter is required.

**location**

The location on a shared drive for the backup command to write to. Alternately it can be set as a [cluster property](#).

**async**

Request ID to track this action which will be [processed asynchronously](#).

**repository**

The name of a repository to be used for the backup. If no repository is specified then the local filesystem repository will be used automatically.

## RESTORE: Restore Collection

Restores Solr indexes and associated configurations.

```
/admin/collections?action=RESTORE&name=myBackupName&location=/path/to/my/shared/drive&collection=myRestoredCollectionName
```

The RESTORE operation will create a collection with the specified name in the collection parameter. You cannot restore into the same collection the backup was taken from. Also the target collection should not be present at the time the API is called as Solr will create it for you.

The collection created will have the same number of shards and replicas as the original collection, preserving routing information, etc. Optionally, you can override some parameters documented below.

While restoring, if a configset with the same name exists in ZooKeeper then Solr will reuse that, or else it will upload the backed up configset in ZooKeeper and use that.

You can use the collection [CREATEALIAS](#) command to make sure clients don't need to change the endpoint to query or index against the newly restored collection.

### RESTORE Parameters

**collection**

The collection where the indexes will be restored into. This parameter is required.

## location

The location on a shared drive for the RESTORE command to read from. Alternately it can be set as a [cluster property](#).

## async

Request ID to track this action which will be [processed asynchronously](#).

## repository

The name of a repository to be used for the backup. If no repository is specified then the local filesystem repository will be used automatically.

## Override Parameters

Additionally, there are several parameters that may have been set on the original collection that can be overridden when restoring the backup:

### collection.configName

Defines the name of the configurations to use for this collection. These must already be stored in ZooKeeper. If not provided, Solr will default to the collection name as the configuration name.

### replicationFactor

The number of replicas to be created for each shard.

### nrtReplicas

The number of NRT (Near-Real-Time) replicas to create for this collection. This type of replica maintains a transaction log and updates its index locally. This parameter behaves the same way as setting replicationFactor parameter.

### tlogReplicas

The number of TLOG replicas to create for this collection. This type of replica maintains a transaction log but only updates its index via replication from a leader. See the section [Types of Replicas](#) for more information about replica types.

### pullReplicas

The number of PULL replicas to create for this collection. This type of replica does not maintain a transaction log and only updates its index via replication from a leader. This type is not eligible to become a leader and should not be the only type of replicas in the collection. See the section [Types of Replicas](#) for more information about replica types.

### maxShardsPerNode

When creating collections, the shards and/or replicas are spread across all available (i.e., live) nodes, and two replicas of the same shard will never be on the same node.

If a node is not live when the CREATE operation is called, it will not get any parts of the new collection, which could lead to too many replicas being created on a single live node. Defining maxShardsPerNode sets a limit on the number of replicas CREATE will spread to each node. If the entire collection can not be fit into the live nodes, no collection will be created at all.

### autoAddReplicas

When set to true, enables auto addition of replicas on shared file systems. See the section [Automatically Add Replicas in SolrCloud](#) for more details on settings and overrides.

property.name=value

Set core property *name* to *value*. See the section [Defining core.properties](#) for details on supported properties and values.

## DELETENODE: Delete Replicas in a Node

Deletes all replicas of all collections in that node. Please note that the node itself will remain as a live node after this operation.

/admin/collections?action=DELETENODE&node=nodeName

### DELETENODE Parameters

node

The node to be removed. This parameter is required.

async

Request ID to track this action which will be [processed asynchronously](#).

## REPLACENODE: Move All Replicas in a Node to Another

This command recreates replicas in one node (the source) to another node(s) (the target). After each replica is copied, the replicas in the source node are deleted.

For source replicas that are also shard leaders the operation will wait for the number of seconds set with the timeout parameter to make sure there's an active replica that can become a leader (either an existing replica becoming a leader or the new replica completing recovery and becoming a leader).

The API uses the Autoscaling framework to find nodes that can satisfy the disk requirements for the new replicas but only when an Autoscaling policy is configured. Refer to [Autoscaling Policy and Preferences](#) section for more details.

/admin/collections?action=REPLACENODE&sourceNode=source-node&targetNode=target-node

### REPLACENODE Parameters

sourceNode

The source node from which the replicas need to be copied from. This parameter is required.

targetNode

The target node where replicas will be copied. If this parameter is not provided, Solr will identify nodes automatically based on policies or number of cores in each node.

parallel

If this flag is set to true, all replicas are created in separate threads. Keep in mind that this can lead to very high network and disk I/O if the replicas have very large indices. The default is false.

async

Request ID to track this action which will be [processed asynchronously](#).

timeout

Time in seconds to wait until new replicas are created, and until leader replicas are fully recovered. The

default is 300, or 5 minutes.



This operation does not hold necessary locks on the replicas that belong to on the source node. So don't perform other collection operations in this period.

## MOVEREPLICA: Move a Replica to a New Node

This command moves a replica from one node to a new node. In case of shared filesystems the dataDir will be reused.

The API uses the Autoscaling framework to find nodes that can satisfy the disk requirements for the replica to be moved but only when an Autoscaling policy is configured. Refer to [Autoscaling Policy and Preferences](#) section for more details.

```
/admin/collections?action=MOVEREPLICA&collection=collection&shard=shard&replica=replica&sourceNode=nodeName&targetNode=nodeName
```

### MOVEREPLICA Parameters

**collection**

The name of the collection. This parameter is required.

**shard**

The name of the shard that the replica belongs to. This parameter is required.

**replica**

The name of the replica. This parameter is required.

**sourceNode**

The name of the node that contains the replica. This parameter is required.

**targetNode**

The name of the destination node. This parameter is required.

**async**

Request ID to track this action which will be [processed asynchronously](#).

## UTILIZENODE: Utilize a New Node

This command can be used to move some replicas from the existing nodes to either a new node or a less loaded node to reduce the load on the existing node.

This uses your autoscaling policies and preferences to identify which replica needs to be moved. It tries to fix any policy violations first and then it tries to move some load off of the most loaded nodes according to the preferences.

```
/admin/collections?action=UTILIZENODE&node=nodeName
```

### UTILIZENODE Parameters

**node**

The name of the node that needs to be utilized. This parameter is required.

## Asynchronous Calls

Since some collection API calls can be long running tasks (such as SPLITSHARD), you can optionally have the calls run asynchronously. Specifying `async=<request-id>` enables you to make an asynchronous call, the status of which can be requested using the [REQUESTSTATUS](#) call at any time.

As of now, REQUESTSTATUS does not automatically clean up the tracking data structures, meaning the status of completed or failed tasks stays stored in ZooKeeper unless cleared manually. DELETSTATUS can be used to clear the stored statuses. However, there is a limit of 10,000 on the number of async call responses stored in a cluster.

### Examples of Async Requests

#### Input

```
http://localhost:8983/solr/admin/collections?action=SPLITSHARD&collection=collection1&shard=shard1&async=1000&wt=xml
```

#### Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">99</int>
 </lst>
 <str name="requestid">1000</str>
</response>
```

## Parameter Reference

### Cluster Parameters

#### numShards

Defaults to 1. The number of shards to hash documents to. There must be one leader per shard and each leader can have  $N$  replicas.

### SolrCloud Instance Parameters

These are set in `solr.xml`, but by default the `host` and `hostContext` parameters are set up to also work with system properties.

#### host

Defaults to the first local host address found. If the wrong host address is found automatically, you can override the host address with this parameter.

#### hostPort

Defaults to the port specified via `bin/solr -p <port>`, or 8983 if not specified. The port that Solr is running on. This value is only used when `-DzkRun` is specified without a value (see below), to calculate the

default port on which embedded ZooKeeper will run. In the `solr.xml` shipped with Solr, the `hostPort` system property is not referenced, and so is ignored. If you want to run Solr on a non-default port, use `bin/solr -p <port>` rather than specifying `-DhostPort`.

#### hostContext

Defaults to `solr`. The context path for the Solr web application.

## SolrCloud Instance ZooKeeper Parameters

#### zkRun

Defaults to `localhost:<hostPort+1000>`. Causes Solr to run an embedded version of ZooKeeper. Set to the address of ZooKeeper on this node; this allows us to know who you are in the list of addresses in the `zkHost` connect string. Use `-DzkRun` (with no value) to get the default value.

#### zkHost

The host address for ZooKeeper. Usually this is a comma-separated list of addresses to each node in your ZooKeeper ensemble.

#### zkClientTimeout

Defaults to 15000. The time a client is allowed to not talk to ZooKeeper before its session expires.

`zkRun` and `zkHost` are set up using system properties. `zkClientTimeout` is set up in `solr.xml` by default, but can also be set using a system property.

## SolrCloud Core Parameters

#### shard

Defaults to being automatically assigned based on `numShards`. Specifies which shard this core acts as a replica of. `shard` can be specified in the `core.properties` for each core.

Additional cloud related parameters are discussed in [Format of solr.xml](#)

## Command Line Utilities

A ZooKeeper Command Line Interface (CLI) script is available to allow you to interact directly with Solr configuration files stored in ZooKeeper.

While Solr's Administration UI includes pages dedicated to the state of your SolrCloud cluster, it does not allow you to download or modify related configuration files.



See the section [Cloud Screens](#) for more information about using the Admin UI screens.

The ZooKeeper CLI scripts found in `server/scripts/cloud-scripts` let you upload configuration information to ZooKeeper, in the same ways shown in the examples in [Parameter Reference](#). It also provides a few other commands that let you link collection sets to collections, make ZooKeeper paths or clear them, and download configurations from ZooKeeper to the local filesystem.

Many of the functions provided by the `zkCli.sh` script are also provided by the [Solr Control Script](#), which may be more familiar as the start script ZooKeeper maintenance commands are very similar to Unix commands.

### Solr's zkcli.sh vs ZooKeeper's zkCli.sh

The zkcli.sh provided by Solr is not the same as the zkCli.sh [included in ZooKeeper distributions](#).

ZooKeeper's zkCli.sh provides a completely general, application-agnostic shell for manipulating data in ZooKeeper. Solr's zkcli.sh – discussed in this section – is specific to Solr, and has command line arguments specific to dealing with Solr data in ZooKeeper.

## Using Solr's ZooKeeper CLI

Use the help option to get a list of available commands from the script itself, as in  
./server/scripts/cloud-scripts/zkcli.sh help.

Both zkcli.sh (for Unix environments) and zkcli.bat (for Windows environments) support the following command line options:

**-cmd <arg>**

The CLI Command to be executed. This parameter is **mandatory**. The following commands are supported:

- bootstrap
- upconfig
- downconfig
- linkconfig
- makepath
- get and getFile
- put and putFile
- clear
- list
- ls
- clusterprop

**-z or -zkhost <locations>**

ZooKeeper host address. This parameter is **mandatory** for all CLI commands.

**-c or -collection <name>**

For linkconfig: name of the collection.

**-d or -confdir <path>**

For upconfig: a directory of configuration files. For downconfig: the destination of files pulled from ZooKeeper

**-h or -help**

Display help text.

**-n or -confname <arg>**

For upconfig, linkconfig, downconfig: name of the configuration set.

-r **or** -runzk <port>

Run ZooKeeper internally by passing the Solr run port; only for clusters on one machine.

-s **or** -solrhome <path>

For bootstrap or when using -runzk: the **mandatory** solrhome location.

-name <name>

For clusterprop: the **mandatory** cluster property name.

-val <value>

For clusterprop: the cluster property value. If not specified, **null** will be used as value.

The short form parameter options may be specified with a single dash (e.g., -c mycollection).

The long form parameter options may be specified using either a single dash (e.g., -collection mycollection) or a double dash (e.g., --collection mycollection)



## ZooKeeper CLI Examples

Below are some examples of using the zkcli.sh CLI which assume you have already started the SolrCloud example (bin/solr -e cloud -noprompt)

If you are on Windows machine, simply replace zkcli.sh with zkcli.bat in these examples.

### Upload a Configuration Directory

```
./server/scripts/cloud-scripts/zkcli.sh -zkhost 127.0.0.1:9983 -cmd upconfig -confname my_new_config -confdir server/solr/configsets/_default/conf
```

### Bootstrap ZooKeeper from an Existing solr.home

```
./server/scripts/cloud-scripts/zkcli.sh -zkhost 127.0.0.1:2181 -cmd bootstrap -solrhome /var/solr/data
```

#### Bootstrap with chroot

Using the bootstrap command with a ZooKeeper chroot in the -zkhost parameter, e.g., -zkhost 127.0.0.1:2181/solr, will automatically create the chroot path before uploading the configs.



### Put Arbitrary Data into a New ZooKeeper file

```
./server/scripts/cloud-scripts/zkcli.sh -zkhost 127.0.0.1:9983 -cmd put /my_zk_file.txt 'some data'
```

## Put a Local File into a New ZooKeeper File

```
./server/scripts/cloud-scripts/zkcli.sh -zkhost 127.0.0.1:9983 -cmd putfile /my_zk_file.txt
/tmp/my_local_file.txt
```

## Link a Collection to a ConfigSet

```
./server/scripts/cloud-scripts/zkcli.sh -zkhost 127.0.0.1:9983 -cmd linkconfig -collection
gettingstarted -confname my_new_config
```

## Create a New ZooKeeper Path

This can be useful to create a chroot path in ZooKeeper before first cluster start.

```
./server/scripts/cloud-scripts/zkcli.sh -zkhost 127.0.0.1:2181 -cmd makepath /solr
```

## Set a Cluster Property

This command will add or modify a single cluster property in `clusterprops.json`. Use this command instead of the usual `getfile -> edit -> putfile` cycle.

Unlike the `CLUSTERPROP` command on the [Collections API](#), this command does **not** require a running Solr cluster.

```
./server/scripts/cloud-scripts/zkcli.sh -zkhost 127.0.0.1:2181 -cmd clusterprop -name urlScheme
-val https
```

## SolrCloud with Legacy Configuration Files

If you are migrating from a non-SolrCloud environment to SolrCloud, this information may be helpful.

All of the required configuration is already set up in the sample configurations shipped with Solr. You only need to add the following if you are migrating old configuration files. Do not remove these files and parameters from a new Solr instance if you intend to use Solr in SolrCloud mode.

These properties exist in 3 files: `schema.xml`, `solrconfig.xml`, and `solr.xml`.

1. In `schema.xml`, you must have a `_version_` field defined:

```
<field name="_version_" type="long" indexed="true" stored="true" multiValued="false"/>
```

2. In `solrconfig.xml`, you must have an `UpdateLog` defined. This should be defined in the `updateHandler` section.

```
<updateHandler>
...
<updateLog>
 <str name="dir">${solr.data.dir:}</str>
</updateLog>
...
</updateHandler>
```

3. The [DistributedUpdateProcessor](#) is part of the default update chain and is automatically injected into any of your custom update chains, so you don't actually need to make any changes for this capability. However, should you wish to add it explicitly, you can still add it to the `solrconfig.xml` file as part of an `updateRequestProcessorChain`. For example:

```
<updateRequestProcessorChain name="sample">
 <processor class="solr.LogUpdateProcessorFactory" />
 <processor class="solr.DistributedUpdateProcessorFactory"/>
 <processor class="my.package.UpdateFactory"/>
 <processor class="solr.RunUpdateProcessorFactory" />
</updateRequestProcessorChain>
```

If you do not want the `DistributedUpdateProcessorFactory` auto-injected into your chain (for example, if you want to use SolrCloud functionality, but you want to distribute updates yourself) then specify the `NoOpDistributingUpdateProcessorFactory` update processor factory in your chain:

```
<updateRequestProcessorChain name="sample">
 <processor class="solr.LogUpdateProcessorFactory" />
 <processor class="solr.NoOpDistributingUpdateProcessorFactory"/>
 <processor class="my.package.MyDistributedUpdateFactory"/>
 <processor class="solr.RunUpdateProcessorFactory" />
</updateRequestProcessorChain>
```

In the update process, Solr skips updating processors that have already been run on other nodes.

## Configsets API

The Configsets API enables you to upload new configsets to ZooKeeper, create, and delete configsets when Solr is running SolrCloud mode.

Configsets are a collection of configuration files such as `solrconfig.xml`, `synonyms.txt`, the schema, language-specific files, DIH-related configuration, and other collection-level configuration files (everything that normally lives in the `conf` directory). Solr ships with two example configsets (`_default` and `sample_techproducts_configs`) which can be used when creating collections. Using the same concept, you can create your own configsets and make them available when creating collections.

This API provides a way to upload configuration files to ZooKeeper and share the same set of configuration files between two or more collections.

Once a configset has been uploaded to ZooKeeper, use the configset name when creating the collection with the [Collections API](#) and the collection will use your configuration files.

Configsets do not have to be shared between collections if they are uploaded with this API, but this API makes it easier to do so if you wish. An alternative to uploading your configsets in advance would be to put the configuration files into a directory under `server/solr/configsets` and using the directory name as the `-d` parameter when using `bin/solr create` to create a collection.



This API can only be used with Solr running in SolrCloud mode. If you are not running Solr in SolrCloud mode but would still like to use shared configurations, please see the section [Config Sets](#).

The API works by passing commands to the `configs` endpoint. The path to the endpoint varies depending on the API being used: the v1 API uses `solr/admin/configs`, while the v2 API uses `api/cluster/configs`. Examples of both types are provided below.

## List Configsets

The list command fetches the names of the configsets that are available for use during collection creation.

### V1 API

With the v1 API, the list command must be capitalized as LIST:

```
http://localhost:8983/solr/admin/configs?action=LIST&omitHeader=true
```

### V2 API

With the v2 API, the list command is implied when there is no data sent with the request.

```
http://localhost:8983/api/cluster/configs?omitHeader=true
```

The output will look like:

```
{
 "configSets": [
 "_default",
 "techproducts",
 "gettingstarted"
]
}
```

## Upload a Configset

Upload a configset, which is sent as a zipped file.

This functionality is enabled by default, but can be disabled via a runtime parameter `-Dconfigset.upload.enabled=false`. Disabling this feature is advisable if you want to expose Solr installation to untrusted users (even though you should never do that!).

A configset is uploaded in a "trusted" mode if authentication is enabled and the upload operation is performed as an authenticated request. Without authentication, a configset is uploaded in an "untrusted" mode. Upon creation of a collection using an "untrusted" configset, the following functionality will not work:

- If specified in the configset, the DataImportHandler's ScriptTransformer will not initialize.
- The XSLT transformer (`tr` parameter) cannot be used at request processing time.
- If specified in the configset, the StatelessScriptUpdateProcessor will not initialize.

If you use any of these parameters or features, you must have enabled security features in your Solr installation and you must upload the configset as an authenticated user.

The upload command takes one parameter:

#### **name**

The configset to be created when the upload is complete. This parameter is required.

The body of the request should be a zip file that contains the configset. The zip file must be created from within the conf directory (i.e., `solrconfig.xml` must be the top level entry in the zip file).

Here is an example on how to create the zip file named "myconfig.zip" and upload it as a configset named "myConfigSet":

```
$ (cd solr/server/solr/configsets/sample_techproducts_configs/conf && zip -r - *) >
myconfigset.zip

$ curl -X POST --header "Content-Type:application/octet-stream" --data-binary @myconfigset.zip
"http://localhost:8983/solr/admin/configs?action=UPLOAD&name=myConfigSet"
```

The same can be achieved using a Unix pipe with a single request as follows:

```
$ (cd server/solr/configsets/sample_techproducts_configs/conf && zip -r - *) | curl -X POST
--header "Content-Type:application/octet-stream" --data-binary @-
"http://localhost:8983/solr/admin/configs?action=UPLOAD&name=myConfigSet"
```



The UPLOAD command does not yet have a v2 equivalent API.

## Create a Configset

The create command creates a new configset based on a configset that has been previously uploaded.

If you have not yet uploaded any configsets, see the [Upload a Configset](#) command above.

The following parameters are supported when creating a configset.

#### **name**

The configset to be created. This parameter is required.

### baseConfigSet

The name of the configset to copy as a base. This defaults to \_default

### configSetProp.property=value

A configset property from the base configset to override in the copied configset.

For example, to create a configset named "myConfigset" based on a previously defined "predefinedTemplate" configset, overriding the immutable property to false.

### V1 API

With the v1 API, the create command must be capitalized as CREATE:

```
http://localhost:8983/solr/admin/configs?action=CREATE&name=myConfigSet&baseConfigSet=predefinedTemplate&configSetPropimmutable=false&wt=xml&omitHeader=true
```

### V2 API

With the v2 API, the create command is provided as part of the JSON data that contains the required parameters:

```
curl -X POST -H 'Content-type: application/json' -d '{\n \"create\":{\n \"name\": \"myConfigSet\",\\n \"baseConfigSet\": \"predefinedTemplate\",\\n \"configSetPropimmutable\": \"false\"}}'\nhttp://localhost:8983/api/cluster/configs?omitHeader=true
```

## Output

```
<response>\n<lst name="responseHeader">\n <int name="status">0</int>\n <int name="QTime">323</int>\n</lst>\n</response>
```

## Delete a Configset

The delete command removes a configset. It does not remove any collections that were created with the configset.

### name

The configset to be deleted. This parameter is required.

To delete a configset named "myConfigSet":

### V1 API

With the v1 API, the delete command must be capitalized as DELETE. The name of the configset to delete is provided with the name parameter:

```
http://localhost:8983/solr/admin/configs?action=DELETE&name=myConfigSet&omitHeader=true
```

### V2 API

With the v2 API, the delete command is provided as the request method, as in -X DELETE. The name of the configset to delete is provided as a path parameter:

```
curl -X DELETE http://localhost:8983/api/cluster/configs/myConfigSet?omitHeader=true
```

## Output

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">170</int>
 </lst>
</response>
```

# Rule-based Replica Placement

When Solr needs to assign nodes to collections, it can either automatically assign them randomly or the user can specify a set of nodes where it should create the replicas.

With very large clusters, it is hard to specify exact node names and it still does not give you fine grained control over how nodes are chosen for a shard. The user should be in complete control of where the nodes are allocated for each collection, shard and replica. This helps to optimally allocate hardware resources across the cluster.

Rule-based replica assignment allows the creation of rules to determine the placement of replicas in the cluster. In the future, this feature will help to automatically add or remove replicas when systems go down, or when higher throughput is required. This enables a more hands-off approach to administration of the cluster.

This feature is used in the following instances:

- Collection creation
- Shard creation
- Replica creation
- Shard splitting

## Common Use Cases

There are several situations where this functionality may be used. A few of the rules that could be implemented are listed below:

- Don't assign more than 1 replica of this collection to a host.
- Assign all replicas to nodes with more than 100GB of free disk space or, assign replicas where there is more disk space.
- Do not assign any replica on a given host because I want to run an overseer there.
- Assign only one replica of a shard in a rack.
- Assign replica in nodes hosting less than 5 cores.
- Assign replicas in nodes hosting the least number of cores.

## Rule Conditions

A rule is a set of conditions that a node must satisfy before a replica core can be created there.

There are three possible conditions.

- **shard**: this is the name of a shard or a wild card (\* means for all shards). If shard is not specified, then the rule applies to the entire collection.
- **replica**: this can be a number or a wild-card (\* means any number zero to infinity).
- **tag**: this is an attribute of a node in the cluster that can be used in a rule, e.g., "freedisk", "cores", "rack", "dc", etc. The tag name can be a custom string. If creating a custom tag, a snitch is responsible for

providing tags and values. The section [Snitches](#) below describes how to add a custom tag, and defines six pre-defined tags (cores, freedisk, host, port, node, and sysprop).

## Rule Operators

A condition can have one of the following operators to set the parameters for the rule.

- **equals (no operator required)**: tag:x means tag value must be equal to 'x'
- **greater than (>)**: tag:>x means tag value greater than 'x'. x must be a number
- **less than (<)**: tag:<x means tag value less than 'x'. x must be a number
- **not equal (!)**: tag: !x means tag value MUST NOT be equal to 'x'. The equals check is performed on String value

## Fuzzy Operator (~)

This can be used as a suffix to any condition. This would first try to satisfy the rule strictly. If Solr can't find enough nodes to match the criterion, it tries to find the next best match which may not satisfy the criterion. For example, if we have a rule such as, freedisk:>200~, Solr will try to assign replicas of this collection on nodes with more than 200GB of free disk space. If that is not possible, the node which has the most free disk space will be chosen instead.

## Choosing Among Equals

The nodes are sorted first and the rules are used to sort them. This ensures that even if many nodes match the rules, the best nodes are picked up for node assignment. For example, if there is a rule such as freedisk:>20, nodes are sorted first on disk space descending and the node with the most disk space is picked up first. Or, if the rule is cores:<5, nodes are sorted with number of cores ascending and the node with the least number of cores is picked up first.

## Rules for New Shards

The rules are persisted along with collection state. So, when a new replica is created, the system will assign replicas satisfying the rules. When a new shard is created as a result of using the Collection API's [CREATESHARD command](#), ensure that you have created rules specific for that shard name. Rules can be altered using the [MODIFYCOLLECTION command](#). However, it is not required to do so if the rules do not specify explicit shard names. For example, a rule such as shard:shard1,replica:\*,ip\_3:168:, will not apply to any new shard created. But, if your rule is replica:\*,ip\_3:168, then it will apply to any new shard created.

The same is applicable to shard splitting. Shard splitting is treated exactly the same way as shard creation. Even though shard1\_1 and shard1\_2 may be created from shard1, the rules treat them as distinct, unrelated shards.

## Snitches

Tag values come from a plugin called Snitch. If there is a tag named 'rack' in a rule, there must be Snitch which provides the value for 'rack' for each node in the cluster. A snitch implements the Snitch interface. Solr, by default, provides a default snitch which provides the following tags:

- **cores**: Number of cores in the node
- **freedisk**: Disk space available in the node
- **host**: host name of the node
- **port**: port of the node
- **node**: node name
- **role**: The role of the node. The only supported role is 'overseer'
- **ip\_1, ip\_2, ip\_3, ip\_4**: These are ip fragments for each node. For example, in a host with ip 192.168.1.2, ip\_1 = 2, ip\_2 = 1, ip\_3 = 168 and ` ip\_4 = 192`
- **sysprop.{PROPERTY\_NAME}**: These are values available from system properties. sysprop.key means a value that is passed to the node as -Dkey=keyValue during the node startup. It is possible to use rules like sysprop.key:expectedVal, shard:\*

## How Snitches are Configured

It is possible to use one or more snitches for a set of rules. If the rules only need tags from default snitch it need not be explicitly configured. For example:

```
snitch=class:fqn.ClassName,key1:val1,key2:val2,key3:val3
```

## How Tag Values are Collected

1. Identify the set of tags in the rules
2. Create instances of Snitches specified. The default snitch is always created.
3. Ask each Snitch if it can provide values for the any of the tags. If even one tag does not have a snitch, the assignment fails.
4. After identifying the Snitches, they provide the tag values for each node in the cluster.
5. If the value for a tag is not obtained for a given node, it cannot participate in the assignment.

## Replica Placement Examples

### Keep less than 2 replicas (at most 1 replica) of this collection on any node

For this rule, we define the replica condition with operators for "less than 2", and use a pre-defined tag named node to define nodes with any name.

```
replica:<2,node:*
// this is equivalent to replica:<2,node:*,shard:**. We can omit shard:** because ** is the
default value of shard
```

### For a given shard, keep less than 2 replicas on any node

For this rule, we use the shard condition to define any shard, the replica condition with operators for "less than 2", and finally a pre-defined tag named node to define nodes with any name.

```
shard:*,replica:<2,node:*
```

## Assign all replicas in shard1 to rack 730

This rule limits the shard condition to 'shard1', but any number of replicas. We're also referencing a custom tag named rack. Before defining this rule, we will need to configure a custom Snitch which provides values for the tag rack.

```
shard:shard1,replica:*,rack:730
```

In this case, the default value of replica is \*, or all replicas. It can be omitted and the rule will be reduced to:

```
shard:shard1,rack:730
```

## Create replicas in nodes with less than 5 cores only

This rule uses the replica condition to define any number of replicas, but adds a pre-defined tag named core and uses operators for "less than 5".

```
replica:*,cores:<5
```

Again, we can simplify this to use the default value for replica, like so:

```
cores:<5
```

## Do not create any replicas in host 192.45.67.3

This rule uses only the pre-defined tag host to define an IP address where replicas should not be placed.

```
host:!192.45.67.3
```

## Defining Rules

Rules are specified per collection during collection creation as request parameters. It is possible to specify multiple 'rule' and 'snitch' parameters as in this example:

```
snitch=class:EC2Snitch&rule=shard:*,replica:1,dc:dc1&rule=shard:*,replica:<2,dc:dc3
```

These rules are persisted in `clusterstate.json` in ZooKeeper and are available throughout the lifetime of the collection. This enables the system to perform any future node allocation without direct user interaction. The rules added during collection creation can be modified later using the [MODIFYCOLLECTION](#) API.

# Cross Data Center Replication (CDCR)

Cross Data Center Replication (CDCR) allows you to create multiple SolrCloud data centers and keep them in sync.

## What is CDCR?

The [SolrCloud](#) architecture is designed to support [Near Real Time \(NRT\)](#) searches on a Solr collection that usually consists of multiple nodes in a single data center. CDCR augments this model by forwarding updates from a Solr collection in one data center to a parallel Solr collection in another data center where the network latencies are greater than the SolrCloud model was designed to accommodate.

For more information about CDCR, see the following sections:

- [CDCR Architecture](#): A detailed overview of how CDCR works.
- [CDCR Configuration](#): How to set up and initialize CDCR for your cluster.
- [CDCR Operations](#): Information on monitoring CDCR and upgrading your cluster when using CDCR.
- [CDCR API](#): Reference for the CDCR API.

## CDCR Glossary

For the purposes of discussing CDCR, the following terminology is used. If you are already familiar with SolrCloud, many of these terms will already be familiar to you.

### Node

A JVM instance running Solr; a server.

### Cluster

A set of Solr nodes managed as a single unit by a ZooKeeper ensemble hosting one or more Collections.

### Data Center

A group of networked servers hosting a Solr cluster. For CDCR, the terms *Cluster* and *Data Center* are interchangeable as we assume that each Solr cluster is hosted in a different group of networked servers.

### Shard

A sub-index of a single logical collection. This may be spread across multiple nodes of the cluster. Each shard can have 1-N replicas.

### Leader

Each shard has replica identified as its leader. All the writes for documents belonging to a shard are routed through the leader.

### Replica

A copy of a shard for use in failover or load balancing. Replicas comprising a shard can either be leaders or non-leaders.

### Follower

A convenience term for a replica that is *not* the leader of a shard.

## Collection

A logical index, consisting of one or more shards. A cluster can have multiple collections.

## Update

An operation that changes the collection's index in any way. This could be adding a new document, deleting documents or changing a document.

## Update Log(s)

An append-only log of write operations maintained by each node.

# CDCR Architecture

## CDCR Architecture Overview

With CDCR, Source and Target data centers can each serve search queries when CDCR is operating. The Target data center will lag somewhat behind the Source cluster due to propagation delays.

Data changes on the Source data center are replicated to the Target data center only after they are persisted to disk. The data changes can be replicated in near real-time (with a small delay) or could be scheduled to be sent at longer intervals to the Target data center. CDCR can "bootstrap" the collection to the Target data center. Since this is a full copy of the entire index, network bandwidth should be considered. Of course both Source and Target collections may be empty to start.

Each shard leader in the Source data center will be responsible for replicating its updates to the corresponding leader in the Target data center. When receiving updates from the Source data center, shard leaders in the Target data center will replicate the changes to their own replicas as normal SolrCloud updates.

This replication model is designed to tolerate some degradation in connectivity, accommodate limited bandwidth, and support batch updates to optimize communication.

Replication supports both a new empty index and pre-built indexes. In the scenario where the replication is set up on a pre-built index in the Source cluster and nothing on the Target cluster, CDCR will replicate the *entire* index from the Source to Target.

The directional nature of the implementation implies a "push" model from the Source collection to the Target collection. Therefore, the Source configuration must be able to "see" the ZooKeeper ensemble in the Target cluster. The ZooKeeper ensemble is provided configured in the Source's solrconfig.xml file.

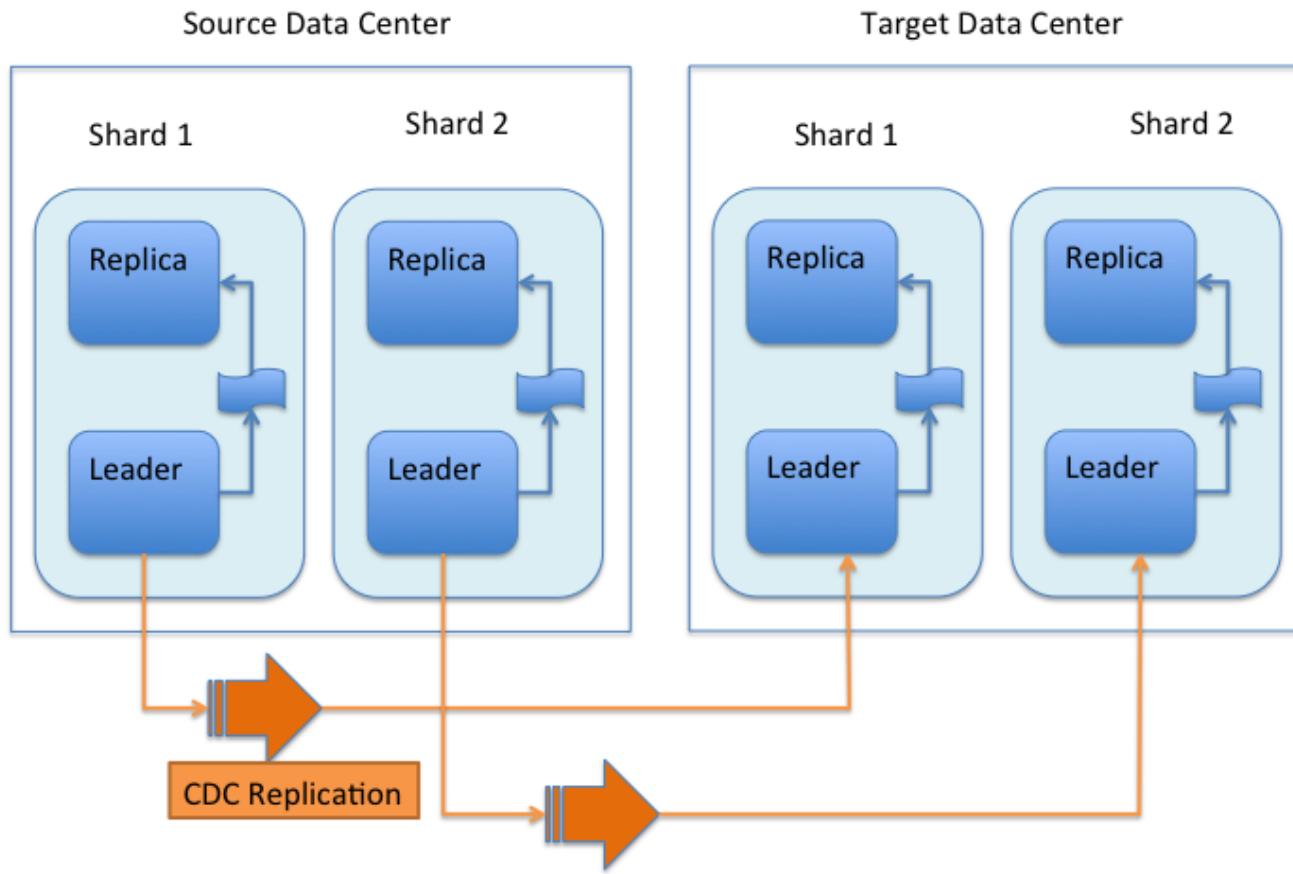
CDCR is configured to replicate from collections in the Source cluster to collections in the Target cluster on a collection-by-collection basis. Since CDCR is configured in solrconfig.xml (on both Source and Target clusters), the settings can be tailored for the needs of each collection.

CDCR can be configured to replicate from one collection to a second collection *within the same cluster*. That is a specialized scenario not covered in this Guide.

## Uni-Directional Architecture

When uni-directional updates are configured, updates and deletes are first written to the Source cluster,

then forwarded to one or more Target data centers, as illustrated in this graphic:



#### *Uni-Directional Data Flow*

With uni-directional updates, the Target data center(s) will not propagate updates such as adds, updates, or deletes to the Source data center and updates should not be sent to any of the Target data center(s).

The data flow sequence is:

1. A shard leader receives a new update that is processed by its update processor chain.
2. The data update is first applied to the local index.
3. Upon successful application of the data update on the local index, the data update is added to CDCR's Update Logs queue.
4. After the data update is persisted to disk, the data update is sent to the replicas within the data center.
5. After Step 4 is successful, CDCR reads the data update from the Update Logs and pushes it to the corresponding collection in the Target data center. This is necessary in order to ensure consistency between the Source and Target data centers.
6. The leader on the Target data center writes the data locally and forwards it to all its followers.

Steps 1, 2, 3 and 4 are performed synchronously by SolrCloud; Step 5 is performed asynchronously by a background thread. Given that CDCR replication is performed asynchronously, it becomes possible to push

batch updates in order to minimize network communication overhead. Also, if CDCR is unable to push the update at a given time, for example, due to a degradation in connectivity, it can retry later without any impact on the Source data center.

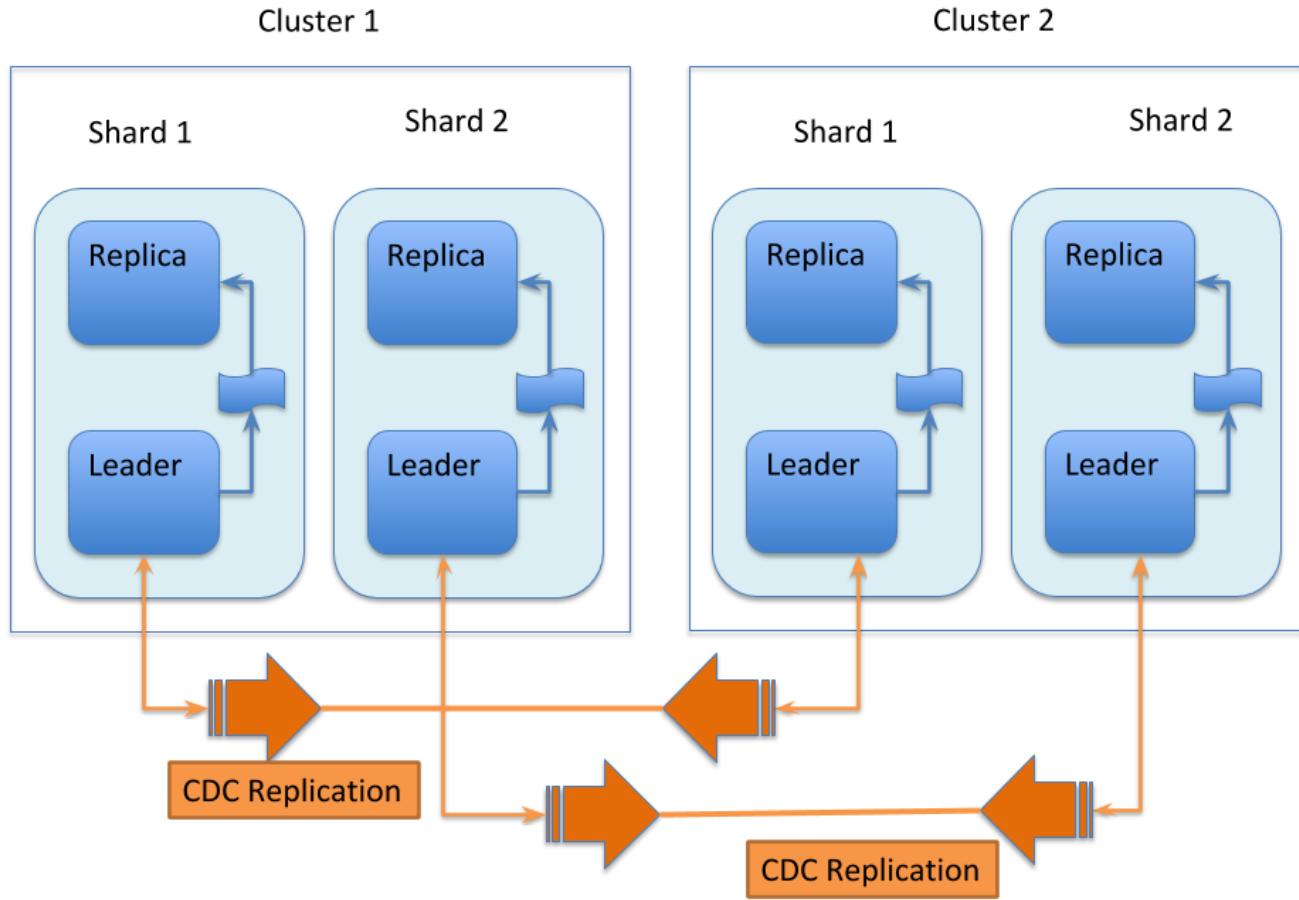
One implication of the architecture is that the leaders in the Source cluster must be able to "see" the leaders in the Target cluster. Since leaders may change in both Source and Target collections, all nodes in the Source cluster must be able to "see" all Solr nodes in the Target cluster. Firewalls, ACL rules, etc., must be configured to allow this.

This design works most robustly if both the Source and Target clusters have the same number of shards. There is no requirement that the shards in the Source and Target collection have the same number of replicas.

Having different numbers of shards on the Source and Target cluster is possible, but is also an "expert" configuration as that option imposes certain constraints and is not generally recommended. Most of the scenarios where having differing numbers of shards are contemplated are better accomplished by hosting multiple shards on each Solr instance.

### Bi-Directional Architecture

When bi-directional updates are configured, either cluster can act as a Source or a Target, and that role can shift between the clusters, as illustrated in this graphic:



### *Bi-Directional Data Flow*

With bi-directional updates, indexing and querying must be done on a single cluster at a time to maintain consistency. The second cluster is used when the first cluster is down. Simplifying, one cluster can act as Source and other as Target but both roles, Source and Target, cannot be assigned to any single cluster at the same time. Failover is handled smoothly without any configuration changes. Updates sent from Source data center to Target is not propagated back to Source when bi-directional updates are configured.

The data flow sequence is similar from Step 1 to 6 above, with an additional step:

7. When bi-directional updates are configured, the updates received from Source are flagged on Target and not forwarded further.

All the behavior(s) and constraint(s) explained in uni-directional data flow are applicable to the respective Source and Target clusters in this scenario.

## Major Components of CDCR

What follows is a discussion of the key features and components in CDCR's architecture:

### CDCR Configuration

In order to configure CDCR, the Source data center requires the host address of the ZooKeeper cluster

associated with the Target data center. The ZooKeeper host address is the only information needed by CDCR to instantiate the communication with the Target Solr cluster. The CDCR configuration section of `solrconfig.xml` file on the Source cluster will therefore contain a list of ZooKeeper hosts. The CDCR configuration section of `solrconfig.xml` might also contain secondary/optional configuration, such as the number of CDC Replicator threads, batch updates related settings, etc.

## CDCR Initialization

CDCR supports incremental updates to either new or existing collections. CDCR may not be able to keep up with very high volume updates, especially if there are significant communications latencies due to a slow "pipe" between the data centers. Some scenarios:

- There is an initial bulk load of a corpus followed by lower volume incremental updates. In this case, one can do the initial bulk load and then enable CDCR. See the section [Initial Startup](#) for more information.
- The index is being built up from scratch, without a significant initial bulk load. CDCR can be set up on empty collections and keep them synchronized from the start.
- The index is always being updated at a volume too high for CDCR to keep up. This is especially possible in situations where the connection between the Source and Target data centers is poor. This scenario is unsuitable for CDCR in its current form.

## Inter-Data Center Communication

The CDCR REST API is the primary form of end-user communication for admin commands.

A SolrJ client is used internally for CDCR operations. The SolrJ client gets its configuration information from the `solrconfig.xml` file. Users of CDCR will not interact directly with the internal SolrJ implementation and will interact with CDCR exclusively through the REST API.

## Updates Tracking & Pushing

CDCR replicates data updates from the Source to the Target data center by leveraging Update Logs. These logs will replace SolrCloud's transaction log.

A background thread regularly checks the Update Logs for new entries, and then forwards them to the Target data center. The thread therefore needs to keep a checkpoint in the form of a pointer to the last update successfully processed in the Update Logs. Upon acknowledgement from the Target data center that updates have been successfully processed, the Update Logs pointer is updated to reflect the current checkpoint.

This pointer must be synchronized across all the replicas. In the case where the leader goes down and a new leader is elected, the new leader will be able to resume replication from the last update by using this synchronized pointer. The strategy to synchronize such a pointer across replicas will be explained next.

If for some reason, the Target data center is offline or fails to process the updates, the thread will periodically try to contact the Target data center and push the updates while buffering updates on the Source cluster. One implication of this is that the Source Update Logs directory should be periodically monitored as the updates will continue to accumulate and will not be purged until the connection to the Target data center is restored.

## Synchronization of Update Checkpoints

A reliable synchronization of the update checkpoints between the shard leader and shard replicas is critical to avoid introducing inconsistency between the Source and Target data centers. Another important requirement is that the synchronization must be performed with minimal network traffic to maximize scalability.

In order to achieve this, the strategy is to:

- Uniquely identify each update operation. This unique identifier will serve as pointer.
- Rely on two storages: an ephemeral storage on the Source shard leader, and a persistent storage on the Target cluster.

The shard leader in the Source cluster will be in charge of generating a unique identifier for each update operation, and will keep a copy of the identifier of the last processed updates in memory. The identifier will be sent to the Target cluster as part of the update request. On the Target data center side, the shard leader will receive the update request, store it along with the unique identifier in the Update Logs, and replicate it to the other shards.

SolrCloud already provides a unique identifier for each update operation, i.e., a “version” number. This version number is generated using a time-based Import clock which is incremented for each update operation sent. This provides a “happened-before” ordering of the update operations that will be leveraged in (1) the initialization of the update checkpoint on the Source cluster, and in (2) the maintenance strategy of the Update Logs.

The persistent storage on the Target cluster is used only during the election of a new shard leader on the Source cluster. If a shard leader goes down on the Source cluster and a new leader is elected, the new leader will contact the Target cluster to retrieve the last update checkpoint and instantiate its ephemeral pointer. On such a request, the Target cluster will retrieve the latest identifier received across all the shards, and send it back to the Source cluster. To retrieve the latest identifier, every shard leader will look up the identifier of the first entry in its Update Logs and send it back to a coordinator. The coordinator will have to select the highest among them.

This strategy does not require any additional network traffic and ensures reliable pointer synchronization. Consistency is principally achieved by leveraging SolrCloud. The update workflow of SolrCloud ensures that every update is applied to the leader and also to any of the replicas. If the leader goes down, a new leader is elected. During the leader election, a synchronization is performed between the new leader and the other replicas. This ensures that the new leader has a consistent Update Logs with the previous leader. Having a consistent Update Logs means that:

- On the Source cluster, the update checkpoint can be reused by the new leader.
- On the Target cluster, the update checkpoint will be consistent between the previous and new leader. This ensures the correctness of the update checkpoint sent by a newly elected leader from the Target cluster.

## Maintenance of Update Logs

The CDCR replication logic requires modification to the maintenance logic of Update Logs on the Source data center. Initially, the Update Logs acts as a fixed size queue, limited to 100 update entries by default. In CDCR, the Update Logs must act as a queue of variable size as they need to keep track of all the updates up

through the last processed update by the Target data center. Entries in the Update Logs are removed only when all pointers (one pointer per Target data center) are after them.

If the communication with one of the Target data center is slow, the Update Logs on the Source data center can grow to a substantial size. In such a scenario, it is necessary for the Update Logs to be able to efficiently find a given update operation given its identifier. Given that its identifier is an incremental number, it is possible to implement an efficient search strategy. Each transaction log file contains as part of its filename the version number of the first element. This is used to quickly traverse all the transaction log files and find the transaction log file containing one specific version number.

## Monitoring Operations

CDCR provides the following monitoring capabilities over the replication operations:

- Monitoring of the outgoing and incoming replications, with information such as the Source and Target nodes, their status, etc.
- Statistics about the replication, with information such as operations (add/delete) per second, number of documents in the queue, etc.

Information about the lifecycle and statistics will be provided on a per-shard basis by the CDC Replicator thread. The CDCR API can then aggregate this information on a collection level.

## CDC Replicator

The CDC Replicator is a background thread that is responsible for replicating updates from a Source data center to one or more Target data centers. It is responsible for providing monitoring information on a per-shard basis. As there can be a large number of collections and shards in a cluster, we will use a fixed-size pool of CDC Replicator threads that will be shared across shards.

## CDCR Limitations

The current design of CDCR has some limitations. CDCR will continue to evolve over time and many of these limitations will be addressed. Among them are:

- CDCR is unlikely to be satisfactory for bulk-load situations where the update rate is high, especially if the bandwidth between the Source and Target clusters is restricted. In this scenario, the initial bulk load should be performed, the Source and Target data centers synchronized and CDCR be utilized for incremental updates.
- CDCR works most robustly with the same number of shards in the Source and Target collection. The shards in the two collections may have different numbers of replicas.
- Running CDCR with the indexes on HDFS is not currently supported, see the [Solr CDCR over HDFS](#) JIRA issue.
- Configuration files (`solrconfig.xml`, `managed-schema`, etc.) are not automatically synchronized between the Source and Target clusters. This means that when the Source schema or `solrconfig.xml` files are changed, those changes must be replicated manually to the Target cluster. This includes adding fields by the [Schema API](#) or [Managed Resources](#) as well as hand editing those files.
- CDCR doesn't support [Basic Authentication features](#) across clusters.
- CDCR does not yet support TLOG or PULL replica types.

## CDCR Configuration

The Source and Target configurations differ in the case of the data centers being in separate clusters. "Cluster" here means separate ZooKeeper ensembles controlling disjoint Solr instances. Whether these data centers are physically separated or not is immaterial for this discussion.

As described in the section [CDCR Architecture](#), two approaches are supported: uni-directional updates and bi-directional updates.

All CDCR configuration is done in the `solrconfig.xml` file. Because this is a per-collection configuration file, all CDCR configuration is done for each collection.

### Uni-Directional Updates

#### Source Configuration

Here is a sample of a Source configuration file, a section in `solrconfig.xml`. The presence of the `<replica>` section causes CDCR to use this cluster as the Source and it should not be present in the Target collections. Details about each setting are after the two examples. The source example has buffering disabled, the default is enabled:

```
<requestHandler name="/cdcr" class="solr.CdcrRequestHandler">
 <lst name="replica">
 <str name="zkHost">10.240.18.211:2181,10.240.18.212:2181</str>
 <!--
 If you have chrooted your Solr information at the target you must include the chroot, for
 example:
 <str name="zkHost">10.240.18.211:2181,10.240.18.212:2181/solr</str>
 -->
 <str name="source">collection1</str>
 <str name="target">collection1</str>
 </lst>

 <lst name="replicator">
 <str name="threadPoolSize">8</str>
 <str name="schedule">1000</str>
 <str name="batchSize">128</str>
 </lst>

 <lst name="updateLogSynchronizer">
 <str name="schedule">1000</str>
 </lst>

</requestHandler>

<!-- Modify the <updateLog> section of your existing <updateHandler>
 in your config as below -->
<updateHandler class="solr.DirectUpdateHandler2">
 <updateLog class="solr.CdcrUpdateLog">
 <str name="dir">${solr.ulog.dir:}</str>
 <!--Any parameters from the original <updateLog> section -->
 </updateLog>

 <!-- Other configuration options such as autoCommit should still be present -->
</updateHandler>
```

## Target Configuration

Here is a typical Target configuration.

Target instance must configure an update processor chain that is specific to CDCR. The update processor chain must include the CdcrUpdateProcessorFactory. The task of this processor is to ensure that the version numbers attached to update requests coming from a CDCR Source SolrCloud are reused and not overwritten by the Target. A properly configured Target configuration looks similar to this:

```
<requestHandler name="/cdcr" class="solr.CdcrRequestHandler">
 <!-- recommended for Target clusters -->
 <lst name="buffer">
 <str name="defaultState">disabled</str>
 </lst>
</requestHandler>

<requestHandler name="/update" class="solr.UpdateRequestHandler">
 <lst name="defaults">
 <str name="update.chain">cdcr-processor-chain</str>
 </lst>
</requestHandler>

<updateRequestProcessorChain name="cdcr-processor-chain">
 <processor class="solr.CdcrUpdateProcessorFactory"/>
 <processor class="solr.RunUpdateProcessorFactory"/>
</updateRequestProcessorChain>

<!-- Modify the <updateLog> section of your existing <updateHandler> in your
 config as below -->
<updateHandler class="solr.DirectUpdateHandler2">
 <updateLog class="solr.CdcrUpdateLog">
 <str name="dir">${solr.ulog.dir:}</str>
 <!--Any parameters from the original <updateLog> section -->
 </updateLog>

 <!-- Other configuration options such as autoCommit should still be present -->
</updateHandler>
```

## Bi-Directional Updates

The configurations in both Cluster 1 and 2 are identical with respective zkHost string specified in each cluster's solrconfig.xml.



Both Cluster 1 and Cluster 2 can act as Source and Target at any given point of time but a cluster cannot be both Source and Target at the same time.

### Cluster 1 Configuration

Here is a sample of a Cluster 1 configuration file, a section in solrconfig.xml. Cluster 2 zkhost string is specified in a CdcrRequestHandler declaration:

```
<requestHandler name="/update" class="solr.UpdateRequestHandler">
 <lst name="defaults">
 <str name="update.chain">cdcr-processor-chain</str>
 </lst>
</requestHandler>

<updateRequestProcessorChain name="cdcr-processor-chain">
 <processor class="solr.CdcrUpdateProcessorFactory"/>
 <processor class="solr.RunUpdateProcessorFactory"/>
</updateRequestProcessorChain>

<requestHandler name="/cdcr" class="solr.CdcrRequestHandler">
 <lst name="replica">
 <str name="zkHost">10.240.19.241:2181,10.240.19.242:2181</str>
 <!--
 If you have chrooted your Solr information at the target you must include the chroot, for
 example:
 <str name="zkHost">10.240.19.241:2181,10.240.19.242:2181/solr</str>
 -->
 <str name="source">collection1</str>
 <str name="target">collection1</str>
 </lst>

 <lst name="replicator">
 <str name="threadPoolSize">8</str>
 <str name="schedule">1000</str>
 <str name="batchSize">128</str>
 </lst>

 <lst name="updateLogSynchronizer">
 <str name="schedule">1000</str>
 </lst>
</requestHandler>

<!-- Modify the <updateLog> section of your existing <updateHandler>
 in your config as below -->
<updateHandler class="solr.DirectUpdateHandler2">
 <updateLog class="solr.CdcrUpdateLog">
 <str name="dir">${solr.ulog.dir:}</str>
 <!--Any parameters from the original <updateLog> section -->
 </updateLog>
</updateHandler>
```

## Cluster 2 Configuration

The configuration of the 2nd cluster is identical to the configuration of Cluster 1, with the Cluster 1 zkHost string specified in CdcrRequestHandler definition:

```
<requestHandler name="/update" class="solr.UpdateRequestHandler">
 <lst name="defaults">
 <str name="update.chain">cdcr-processor-chain</str>
 </lst>
</requestHandler>

<updateRequestProcessorChain name="cdcr-processor-chain">
 <processor class="solr.CdcrUpdateProcessorFactory"/>
 <processor class="solr.RunUpdateProcessorFactory"/>
</updateRequestProcessorChain>

<requestHandler name="/cdcr" class="solr.CdcrRequestHandler">
 <lst name="replica">
 <str name="zkHost">10.250.18.211:2181,10.250.18.212:2181</str>
 <!--
 If you have chrooted your Solr information at the target you must include the chroot, for
 example:
 <str name="zkHost">10.250.18.211:2181,10.250.18.212:2181/solr</str>
 -->
 <str name="source">collection1</str>
 <str name="target">collection1</str>
 </lst>

 <lst name="replicator">
 <str name="threadPoolSize">8</str>
 <str name="schedule">1000</str>
 <str name="batchSize">128</str>
 </lst>

 <lst name="updateLogSynchronizer">
 <str name="schedule">1000</str>
 </lst>
</requestHandler>

<!-- Modify the <updateLog> section of your existing <updateHandler>
 in your config as below -->
<updateHandler class="solr.DirectUpdateHandler2">
 <updateLog class="solr.CdcrUpdateLog">
 <str name="dir">${solr.ulog.dir:}</str>
 <!--Any parameters from the original <updateLog> section -->
 </updateLog>
</updateHandler>
```

## CDCR Configuration Parameters

The configuration details, defaults and options are as follows:

### The Replica Element

CDCR can be configured to forward update requests to one or more Target collections. A Target collection is

defined with a “replica” list as follows:

#### zkHost

The host address for ZooKeeper of the Target SolrCloud. Usually this is a comma-separated list of addresses to each node in the Target ZooKeeper ensemble. This parameter is required.

#### Source

The name of the collection on the Source SolrCloud to be replicated. This parameter is required.

#### Target

The name of the collection on the Target SolrCloud to which updates will be forwarded. This parameter is required.

## The Replicator Element

The CDC Replicator is the component in charge of forwarding updates to the replicas. The replicator will monitor the update logs of the Source collection and will forward any new updates to the Target collection.

The replicator uses a fixed thread pool to forward updates to multiple replicas in parallel. If more than one replica is configured, one thread will forward a batch of updates from one replica at a time in a round-robin fashion. The replicator can be configured with a “replicator” list as follows:

#### threadPoolSize

The number of threads to use for forwarding updates. One thread per replica is recommended. The default is 2.

#### schedule

The delay in milliseconds for monitoring the update log(s). The default is 10.

#### batchSize

The number of updates to send in one batch. The optimal size depends on the size of the documents. Large batches of large documents can increase your memory usage significantly. The default is 128.

## The updateLogSynchronizer Element

Expert: Non-leader nodes need to synchronize their update logs with their leader node from time to time in order to clean deprecated transaction log files. By default, such a synchronization process is performed every minute. The schedule of the synchronization can be modified with a “updateLogSynchronizer” list as follows:



If the updateLogSynchronizer element is omitted from the Source cluster, transaction logs may accumulate on non-leaders.

#### schedule

The delay in milliseconds for synchronizing the update logs. The default is 60000.

## The Buffer Element

When buffering updates, the update logs will store all the updates indefinitely. It is best to disable buffering on both the Source and Target clusters during normal operation as when buffering is enabled the Update Logs will grow without limit. Enabling buffering is intended for special maintenance periods. Buffering can be disabled at startup with a “buffer” list and the parameter “defaultState” as follows:

## defaultState

The state of the buffer at startup. The default is enabled.

### *Buffering should be enabled only for maintenance windows*

Buffering is designed to augment maintenance windows. The following points should be kept in mind:

- When buffering is enabled, the Update Logs will grow without limit; they will never be purged.
- During normal operation, the Update Logs will automatically accrue on the Source data center if the Target data center is unavailable; It is not necessary to enable buffering for CDCR to handle routine network disruptions.
  - For this reason, monitoring disk usage on the Source data center is recommended as an additional check that the Target data center is receiving updates.
- For uni-directional updates, buffering should *not* be enabled on the Target data center as Update Logs would accrue without limit.
- If buffering is enabled and then disabled, the Update Logs will be removed when their contents have been sent to the Target data center. This process may take some time and is triggered by additional updates the Source cluster.
  - Update Log cleanup is not triggered until a new update is sent to the Source data center.



## Initial Startup

### Uni-Directional Approach

This is a general approach for initializing CDCR in a production environment. It's based upon an approach taken by the initial working installation of CDCR and generously contributed to illustrate a "real world" scenario.

- CDCR is used to keep a remote disaster-recovery instance available for production backup.
- This example uses 26 clouds with 200 million assets per cloud (15GB indexes). Total document count is over 4.8 billion.
  - Source and Target clouds were synched in 2-3 hour maintenance windows to establish the base index for the Targets.

As usual, it is good to start small. Sync a single cloud and monitor for a period of time before doing the others. You may need to adjust your settings several times before finding the right balance.

- Before starting, stop or pause the indexers. This is best done during a small maintenance window.
- Stop the SolrCloud instances at the Source.
- Upload the modified `solrconfig.xml` to ZooKeeper on both Source and Target as appropriate, see the examples above.
- Sync the index directories from the Source collection to Target collection across to the corresponding shard nodes. `rsync` works well for this.

For example, if there are two shards on collection1 with 2 replicas for each shard, copy the corresponding index directories from:

shard1replica1Source	to	shard1replica1Target
shard1replica2Source	to	shard1replica2Target
shard2replica1Source	to	shard2replica1Target
shard2replica2Source	to	shard2replica2Target

- Start ZooKeeper on the Target (DR).
- Start SolrCloud on the Target (DR).
- Start ZooKeeper on the Source.
- Start SolrCloud on the Source. As a general rule, the Target (DR) should be started before the Source.
- Activate CDCR on Source instance using the CDCR API:

```
http://host:port/solr/<collection_name>/cdcr?action=START
```

There is no need to run the /cdcr?action=START command on the Target.

- Disable the buffer on the Target and Source:

```
http://host:port/solr/collection_name/cdcr?action=DISABLEBUFFER
```

- Re-enable indexing.

## Bi-Directional Approach



When using the bi-directional approach, it is highly recommended to enable CDCR on both cluster-collections before any indexing has taken place.

Based on the same example from uni-directional solution, let's walk through the necessary steps:

- Before you begin, stop or pause any indexing processes. This is best done during a small maintenance window.
- Stop the SolrCloud instances in both Cluster 1 and Cluster 2.
- Upload the modified solrconfig.xml to ZooKeeper on both Cluster 1 and Cluster 2 as appropriate, see the examples above in the section [Bi-Directional Updates](#).
- If documents were indexed prior to this exercise, sync the index directories from the Cluster 1 collection to the Cluster 2 collection to the corresponding shard nodes or vice versa. The rsync utility works well for this if it's available on your server. Check to be sure the updated index is copied across.

For example, if there are 2 shards on collection 'cluster1' (the updated collection) with 2 replicas for each shard, copy the corresponding index directories from:

shard1replica1cluster1	to	shard1replica1cluster2
------------------------	----	------------------------

shard1replica2cluster1	to	shard1replica2cluster2
shard2replica1cluster1	to	shard2replica1cluster2
shard2replica2cluster1	to	shard2replica2cluster2

- Start ZooKeeper on Cluster 1.
- Start ZooKeeper on Cluster 2.
- Start SolrCloud on Cluster 1.
- Start SolrCloud on Cluster 2.
- If not present, create respective collections in both Cluster 1 and Cluster 2.
- Activate the CDCR on Cluster 1 and Cluster 2 instance using the CDCR API:

```
http://host:port/solr/<collection_name>/cdcr?action=START
```

- Disable the buffer on Cluster 1 and Cluster 2:

```
http://host:port/solr/collection_name/cdcr?action=DISABLEBUFFER
```

- Re-enable indexing.

## ZooKeeper Settings

With CDCR, the Target ZooKeepers will have connections from the Target clouds and the Source clouds. You may need to increase the `maxClientCnxns` setting in `zoo.cfg`.

```
set numbers of connection to 800 from client
is maxClientCnxns=0 that means no limit
maxClientCnxns=800
```

## Cross Data Center Replication Operations

### Monitoring

1. Network and disk space monitoring are essential. Ensure that the system has plenty of available storage to queue up changes if there is a disconnect between the Source and Target. A network outage between the two data centers can cause your disk usage to grow. Some tips:
  - a. Set a monitor for your disks to send alerts when the disk gets over a certain percentage (e.g., 70%).
  - b. Run a test. With moderate indexing, how long can the system queue changes before you run out of disk space?
2. Create a simple way to check the counts between the Source and the Target.
  - a. Keep in mind that if indexing is running, the Source and Target may not match document for document. Set an alert to fire if the difference is greater than some percentage of the overall cloud

size.

## Upgrading and Patching Production

When rolling in upgrades to your indexer or application, you should shutdown the Source and the Target. Depending on your setup, you may want to pause/stop indexing, deploy the release or patch, then re-enable indexing. Then start the Target last.

- There is no need to reissue the DISABLEBUFFERS or START commands. These are persisted.
- After starting the Target, run a simple test. Add a test document to each of the Source clouds. Then check for it on the Target.

```
#send to the Source
curl http://<Source>/solr/cloud1/update -H 'Content-type:application/json' -d '[{"SKU": "ABC"}]'

#check the Target
curl "http://<Target>:8983/solr/<collection_name>/select?q=SKU:ABC&indent=true"
```

## CDCR API

The CDCR API is used to control and monitor the replication process. Control actions are performed at a collection level, i.e., by using the following base URL for API calls:

`http://localhost:8983/solr/<collection>/cdcr`.

Monitor actions are performed at a core level, i.e., by using the following base URL for API calls:  
`http://localhost:8983/solr/<core>/cdcr`.

Currently, none of the CDCR API calls have parameters.

## API Entry Points

### Control

- `<collection>/cdcr?action=STATUS`: Returns the [current state](#) of CDCR.
- `<collection>/cdcr?action=START`: Starts [CDCR](#) replication
- `<collection>/cdcr?action=STOP`: Stops [CDCR](#) replication.
- `<collection>/cdcr?action=ENABLEBUFFER`: Enables the [buffering](#) of updates.
- `<collection>/cdcr?action=DISABLEBUFFER`: Disables the [buffering](#) of updates.

### Monitoring

- `core/cdcr?action=QUEUES`: Fetches statistics about the [queue](#) for each replica and about the update logs.
- `core/cdcr?action=OPS`: Fetches statistics about the [replication performance](#) (operations per second) for each replica.
- `core/cdcr?action=ERRORS`: Fetches statistics and other information about [replication errors](#) for each replica.

## Control Commands

### CDCR STATUS

`solr/<collection>/cdcr?action=STATUS`

#### CDCR Status Example

##### Input

```
http://localhost:8983/solr/techproducts/cdcr?action=STATUS
```

##### Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 0
 },
 "status": {
 "process": "stopped",
 "buffer": "enabled"
 }
}
```

### ENABLEBUFFER

`solr/<collection>/cdcr?action=ENABLEBUFFER`

#### Enable Buffer Example

##### Input

```
http://localhost:8983/solr/techproducts/cdcr?action=ENABLEBUFFER
```

##### Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 0
 },
 "status": {
 "process": "started",
 "buffer": "enabled"
 }
}
```

## DISABLEBUFFER

`solr/<collection>/cdcr?action=DISABLEBUFFER`

### Disable Buffer Example

#### Input

```
http://localhost:8983/solr/techproducts/cdcr?action=DISABLEBUFFER
```

#### Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 0
 },
 "status": {
 "process": "started",
 "buffer": "disabled"
 }
}
```

## CDCR START

`solr/<collection>/cdcr?action=START`

### CDCR Start Examples

#### Input

```
http://localhost:8983/solr/techproducts/cdcr?action=START
```

#### Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 0
 },
 "status": {
 "process": "started",
 "buffer": "enabled"
 }
}
```

## CDCR STOP

`solr/<collection>/cdcr?action=STOP`

## CDCR Stop Examples

### Input

```
http://localhost:8983/solr/techproducts/cdcr?action=STOP
```

### Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 0
 },
 "status": {
 "process": "stopped",
 "buffer": "enabled"
 }
}
```

## CDCR Monitoring Commands

### QUEUES

```
solr/<core>/cdcr?action=QUEUES
```

#### QUEUES Response

The output is composed of a list “queues” which contains a list of (ZooKeeper) Target hosts, themselves containing a list of Target collections. For each collection, the current size of the queue and the timestamp of the last update operation successfully processed is provided. The timestamp of the update operation is the original timestamp, i.e., the time this operation was processed on the Source SolrCloud. This allows an estimate the latency of the replication process.

The “queues” object also contains information about the update logs, such as the size (in bytes) of the update logs on disk (tlogTotalSize), the number of transaction log files (tlogTotalCount) and the status of the update logs synchronizer (updateLogSynchronizer).

#### QUEUES Examples

### Input

```
http://localhost:8983/solr/<replica_name>/cdcr?action=QUEUES
```

### Output

```
{
 "responseHeader":{
 "status": 0,
 "QTime": 1
 },
 "queues":{
 "127.0.0.1: 40342/solr":{
 "Target_collection":{
 "queueSize": 104,
 "lastTimestamp": "2014-12-02T10:32:15.879Z"
 }
 }
 },
 "tlogTotalSize":3817,
 "tlogTotalCount":1,
 "updateLogSynchronizer": "stopped"
}
```

## OPS

`solr/<core>/cdcr?action=OPS`

### OPS Response

Provides the average number of operations as a sum and broken down by adds/deletes.

### OPS Examples

#### Input

```
http://localhost:8983/solr/<replica_name>/cdcr?action=OPS
```

#### Output

```
{
 "responseHeader":{
 "status":0,
 "QTime":1
 },
 "operationsPerSecond":{
 "127.0.0.1: 59661/solr":{
 "Target_collection":{
 "all": 297.102944952749052,
 "adds": 297.102944952749052,
 "deletes": 0.0
 }
 }
 }
}
```

## ERRORS

`solr/<core>/cdcr?action=ERRORS`

### ERRORS Response

Provides the number of consecutive errors encountered by the replicator thread, the number of bad requests or internal errors since the start of the replication process, and a list of the last errors encountered ordered by timestamp.

### ERRORS Examples

#### Input

```
http://localhost:8983/solr/<replica_name>/cdcr?action=ERRORS
```

#### Output

```
{
 "responseHeader":{
 "status":0,
 "QTime":2
 },
 "errors": {
 "127.0.0.1: 36872/solr":{
 "Target_collection":{
 "consecutiveErrors":3,
 "bad_request":0,
 "internal":3,
 "last":{
 "2014-12-02T11:04:42.523Z":"internal",
 "2014-12-02T11:04:39.223Z":"internal",
 "2014-12-02T11:04:38.22Z":"internal"
 }
 }
 }
 }
}
```

# SolrCloud Autoscaling

The goal of autoscaling is to make SolrCloud cluster management easier by providing a way for changes to the cluster to be more automatic and more intelligent.

Autoscaling includes an API to manage cluster-wide and collection-specific policies and preferences and a rules syntax to define the guidelines for your cluster. Also included are features to utilize the policies and preferences so they perform actions automatically when certain conditions are met.

The following sections describe the autoscaling features of SolrCloud:

- [Overview of Autoscaling in SolrCloud](#)
- [Autoscaling Policy and Preferences](#)
- [Autoscaling Triggers](#)
- [Autoscaling Trigger Actions](#)
- [Autoscaling Listeners](#)
- [Automatically Adding Replicas](#)
- [Autoscaling Fault Tolerance](#)
- [Autoscaling API](#)
- [Migrating Rule-Based Replica Rules to Autoscaling Policies](#)

## Overview of SolrCloud Autoscaling

Autoscaling in Solr aims to provide good defaults so a SolrCloud cluster remains balanced and stable in the face of various cluster change events. This balance is achieved by satisfying a set of rules and sorting preferences to select the target of cluster management operations automatically on cluster events.

A simple example is automatically adding a replica for a SolrCloud collection when a node containing an existing replica goes down.

The goal of autoscaling in SolrCloud is to make cluster management easier, more automatic, and more intelligent. It aims to provide good defaults such that the cluster remains balanced and stable in the face of various events such as a node joining the cluster or leaving the cluster. This is achieved by satisfying a set of rules and sorting preferences that help Solr select the target of cluster management operations.

There are three distinct problems that this feature solves:

- When to run cluster management tasks? For example, we might want to add a replica when an existing replica is no longer alive.
- Which cluster management task to run? For example, do we add a new replica or should we move an existing one to a new node?
- How do we run the cluster management tasks so the cluster remains balanced and stable?

Before we get into the details of how each of these problems are solved, let's take a quick look at the easiest

way to setup autoscaling for your cluster.

## Quick Start: Automatically Adding Replicas

Say that we want to create a collection which always requires us to have three replicas available for each shard all the time. We can set the replicationFactor=3 while creating the collection, but what happens if a node containing one or more of the replicas either crashed or was shutdown for maintenance? In such a case, we'd like to create additional replicas to replace the ones that are no longer available to preserve the original number of replicas.

We have an easy way to enable this behavior without needing to understand the autoscaling features in depth. We can create a collection with such behavior by adding an additional parameter autoAddReplicas=true with the CREATE command of the Collection API. For example:

```
/admin/collections?action=CREATE&name=_name_of_collection_&numShards=1&replicationFactor=3&autoAddReplicas=true
```

A collection created with autoAddReplicas=true will be monitored by Solr such that if a node containing a replica of this collection goes down, Solr will add new replicas on other nodes after waiting for up to thirty seconds for the node to come back.

You can see the section [Autoscaling Automatically Adding Replicas](#) to learn more about how to enable or disable this feature as well as other details.

The selection of the node that will host the new replica is made according to the default cluster preferences that we will learn more about in the next sections.

## Cluster Preferences

Cluster preferences allow you to tell Solr how to assess system load on each node. This information is used to guide selection of the node(s) on which cluster management operations will be performed.

In general, when an operation increases replica counts, the **least loaded** qualified node will be chosen, and when the operation reduces replica counts, the **most loaded** qualified node will be chosen.

The default cluster preferences are [{minimize:cores}, {maximize:freedisk}], which tells Solr to minimize the number of cores on all nodes and if number of cores are equal, maximize the free disk space available. In this case, the least loaded node is the one with the fewest cores or if two nodes have an equal number of cores, the node with the most free disk space.

You can learn more about preferences in the section on [Cluster Preferences Specification](#).

## Cluster Policy

A cluster policy is a set of rules that a node, shard, or collection must satisfy before it can be chosen as the target of a cluster management operation. These rules are applied across the cluster regardless of the collection being managed. For example, the rule {"cores": "<10", "node": "#ANY"} means that any node must have less than 10 Solr cores in total, regardless of which collection they belong to.

There are many metrics on which the rule can be based, e.g., system load average, heap usage, free disk

space, etc. The full list of supported metrics can be found in the section describing [Autoscaling Policy Rule Attributes](#).

When a node, shard, or collection does not satisfy a policy rule, we call it a **violation**. By default, cluster management operations will fail if there is even one violation. You can allow operations to succeed in the face of a violation by marking the corresponding rule with "strict":false. When you do this, Solr ensures that cluster management operations minimize the number of violations.

Solr also supports [collection-specific policies](#), which operate in tandem with the cluster policy.

## Triggers

Now that we have an idea about how cluster management operations use policies and preferences help Solr keep the cluster balanced and stable, we can talk about when to invoke such operations.

Triggers are used to watch for events such as a node joining or leaving the cluster. When the event happens, the trigger executes a set of actions that compute and execute a **plan**, i.e., a set of operations to change the cluster so that the policy and preferences are respected.

The autoAddReplicas parameter passed with the CREATE Collection API command in the [Quick Start](#) section above automatically creates a trigger that watches for a node going away. When the trigger fires, it executes a set of actions that compute and execute a plan to move all replicas hosted by the lost node to new nodes in the cluster. The target nodes are chosen based on the policy and preferences.

You can learn more about Triggers in the section [Autoscaling Triggers](#).

## Trigger Actions

A trigger executes **actions** that tell Solr what to do in response to the trigger. Solr ships with two actions that are added to every trigger by default. The first is called the **ComputePlanAction** and the other is **ExecutePlanAction**. The former computes the cluster management operations necessary to stabilize the cluster and the latter executes them on the cluster.

You can learn more about Trigger Actions in the section [Autoscaling Trigger Actions](#).

## Listeners

An Autoscaling **listener** can be attached to a trigger. Solr calls the listener each time the trigger fires as well as before and after the actions performed by the trigger. Listeners are useful as a call back mechanism to perform tasks such as logging or informing external systems about events. For example, a listener is automatically added by Solr to each trigger to log details of the trigger fire and actions to the .system collection.

You can learn more about Listeners in the section [Autoscaling Listeners](#).

## Autoscaling APIs

The autoscaling APIs available at /admin/autoscaling can be used to read and modify each of the components discussed above.

You can learn more about these APIs in the section [Autoscaling API](#).

## Autoscaling Policy and Preferences

The autoscaling policy and preferences are a set of rules and sorting preferences that help Solr select the target of cluster management operations so the overall load on the cluster remains balanced.

The configured autoscaling policy and preferences are used by [Collections API commands](#) in all contexts: manual, for example using bin/solr to create a collection; semi-automatic, via the [Suggestions API](#) or the Admin UI's [Suggestions Screen](#); or fully automatic, via configured [Triggers](#).

See the section [Example: Manual Collection Creation with a Policy](#) for an example of how policy and preferences affect replica placement.

### Cluster Preferences Specification

A preference is a hint to Solr on how to sort nodes based on their utilization.

The default cluster preference is to sort by the total number of Solr cores (or replicas) hosted by a node, with a precision of 1. Therefore, by default, when selecting a node to which to add a replica, Solr can apply the preferences and choose the node with the fewest cores. In the case of a tie in the number of cores, available freedisk will be used to further sort nodes.

More than one preference can be added to break ties. For example, we may choose to use free disk space to break ties if the number of cores on two nodes is the same. The node with the higher free disk space can be chosen as the target of the cluster operation.

Each preference takes the following form:

```
{"<sort_order>": "<sort_param>", "precision": "<precision_val>"}
```

#### sort\_order

The value can be either `maximize` or `minimize`. Choose `minimize` to sort the nodes with least value as the least loaded. For example, `{"minimize": "cores"}` sorts the nodes with the least number of cores as the least loaded node. A sort order such as `{"maximize": "freedisk"}` sorts the nodes with maximum free disk space as the least loaded node.

The objective of the system is to make every node the least loaded. So, in case of a `MOVEREPLICA` operation, it usually targets the *most loaded* node and takes load off of it. In a sort of more loaded to less loaded, `minimize` is akin to sorting in descending order and `maximize` is akin to sorting in ascending order.

This is a required parameter.

#### sort\_param

One and only one of the following supported parameters must be specified:

1. `cores`: The number of total Solr cores on a node.
2. `freedisk`: The amount of free disk space for Solr's data home directory. This is always in gigabytes.
3. `sysLoadAvg`: The system load average on a node as reported by the Metrics API under the key `solr.jvm/os.systemLoadAverage`. This is always a double value between 0 and 1 and the higher the

value, the more loaded the node is.

4. heapUsage: The heap usage of a node as reported by the Metrics API under the key solr.jvm/memory.heap.usage. This is always a double value between 0 and 1 and the higher the value, the more loaded the node is.

#### precision

Precision tells the system the minimum (absolute) difference between 2 values to treat them as distinct values.

For example, a precision of 10 for freedisk means that two nodes whose free disk space is within 10GB of each other should be treated as equal for the purpose of sorting. This helps create ties without which specifying multiple preferences is not useful. This is an optional parameter whose value must be a positive integer. The maximum value of precision must be less than the maximum value of the sort\_value, if any.

See the section [Create and Modify Cluster Preferences](#) for details on how to manage cluster preferences with the API.

## Examples of Cluster Preferences

### Default Preferences

The following shows the default cluster preferences. This is applied automatically by Solr when no explicit cluster preferences have been set using the [Autoscaling API](#).

```
[
 {"minimize": "cores"}
]
```

### Minimize Cores; Maximize Free Disk

In this example, we want to minimize the number of Solr cores and in case of a tie, maximize the amount of free disk space on each node.

```
[
 {"minimize" : "cores"},
 {"maximize" : "freedisk"}
]
```

### Add Precision to Free Disk; Minimize System Load

In this example, we add a precision to the freedisk parameter so that nodes with free disk space within 10GB of each other are considered equal. In such a case, the tie is broken by minimizing sysLoadAvg.

```
[
 {"minimize" : "cores"},
 {"maximize" : "freedisk", "precision" : 10},
 {"minimize" : "sysLoadAvg"}
]
```

## Policy Specification

A policy is a hard rule to be satisfied by each node. If a node does not satisfy the rule then it is called a **violation**. Solr ensures that the number of violations are minimized while invoking any cluster management operations.

### Policy Rule Structure

#### Rule Types

Policy rules can be either global or per-collection:

- **Global rules** constrain the number of cores per node or node group. This type of rule applies to cores from all collections hosted on the specified node(s). As a result, [collection-specific policies](#), which are associated with individual collections, may not contain global rules.
- **Per-collection rules** constrain the number of replicas per node or node group.

Global rules have three parts:

- [Node Selector](#)
- [Core Count Constraint \("cores": "..."\)](#)
- [Rule Strictness \(optional\)](#)

Per-collection rules have four parts:

- [Node Selector](#)
- [Replica Selector and Rule Evaluation Context](#)
- [Replica Count Constraint \("replica": "..."\)](#)
- [Rule Strictness \(optional\)](#)

#### Node Selector

Rule evaluation is restricted to node(s) matching the value of one of the following attributes: `node`, `port`, `ip_*`, `sysprop.*`, or `diskType`. For replica/core count constraints other than `#EQUAL`, a condition specified in one of the following attributes may instead be used to select nodes: `freedisk`, `host`, `sysLoadAvg`, `heapUsage`, `nodeRole`, or `metrics.*`.

Except for `node`, the attributes above cause selected nodes to be partitioned into node groups. A node group is referred to as a "bucket". Those attributes usable with the `#EQUAL` directive may define buckets either via the special function `#EACH` or an [array](#) `["value1", ...]` (a subset of all possible values); in both cases, each node is placed in the bucket corresponding to the matching attribute value.

The `node` attribute always places each selected node into its own bucket, regardless of the attribute value's

form (#ANY, node-name, or ["node1-name", ...]).

Replica and core count constraints, described below, are evaluated against the total number in each bucket.

### Core Count Constraint

The cores attribute value can be specified in one of the following forms:

- #EQUAL: distribute all cores equally across all the [selected nodes](#).
- a constraint on the core count on each [selected node](#); see [Specifying Replica and Core Count Constraints](#).

### Replica Selector and Rule Evaluation Context

Rule evaluation can be restricted to replicas that meet any combination of conditions specified with the following attributes:

- collection: The replica is of a shard belonging to the collection specified in the attribute value. (Not usable with [collection-specific policies](#).)
- shard: The replica is of the shard named in the attribute value.
- type: The replica has the specified replica type (NRT, TLOG, or PULL).

If none of the above attributes is specified, then the rule is evaluated separately for each collection against all types of replicas of all shards.

Specifying #EACH as the shard attribute value causes the rule to be evaluated separately for each shard of each collection.

### Replica Count Constraint

The replica attribute value can be specified in one of the following forms:

- #ALL: All [selected replicas](#) will be placed on the [selected nodes](#).
- #EQUAL: Distribute [selected replicas](#) equally across all the [selected nodes](#).
- a constraint on the replica count on each [selected node](#); see [Specifying Replica and Core Count Constraints](#).

### Specifying Replica and Core Count Constraints

[Replica count constraints](#) ("replica": "...") and [core count constraints](#) ("cores": "...") allow specification of acceptable counts for replicas (cores tied to a collection) and cores (regardless of the collection to which they belong), respectively.

You can specify one of the following as the value of a replica and cores policy rule attribute:

- an exact integer (e.g., 2)
- an exclusive lower integer bound (e.g., >0)
- an exclusive upper integer bound (e.g., <3)
- a decimal value, interpreted as an acceptable range of core counts, from the floor of the value to the ceiling of the value, with the system preferring the rounded value (e.g., 1.6: 1 or 2 is acceptable, and 2 is preferred)

- a [range](#) of acceptable replica/core counts, as inclusive lower and upper integer bounds separated by a hyphen (e.g., 3-5)
- a percentage (e.g., 33%), which is multiplied at runtime either by the number of [selected replicas](#) (for a replica constraint) or the number of cores in the cluster (for a cores constraint). This value is then interpreted as described above for a literal decimal value.



Using an exact integer value for count constraints is of limited utility, since collection or cluster changes could quickly invalidate them. For example, attempting to add a third replica to each shard of a collection on a two-node cluster with policy rule `{"replica":1, "shard": "#EACH", "node": "#ANY"}` would cause a violation, since at least one node would have to host more than one replica. Percentage rules are less brittle. Rewriting the rule as `{"replica": "50%", "shard": "#EACH", "node": "#ANY"}` eliminates the violation: 50% of 3 replicas = 1.5 replicas per node, meaning that it's acceptable for a node to host either one or two replicas of each shard.

## Policy Rule Attributes

### Rule Strictness

This attribute is usable in all rules:

`strict`

An optional boolean value. The default is true. If true, the rule must be satisfied; if the rule is not satisfied, the resulting violation will cause the cluster management operation to fail. If false, Solr tries to satisfy the rule on a best effort basis, but if no node can satisfy the rule, the cluster management operation will not fail, and any node may be chosen. If multiple rules declared to be `strict:false` can not be satisfied by some nodes, then a node will be chosen such that the number of such violations is minimized.

### Global Rule Attributes

`cores`

The number of cores that must exist to satisfy the rule. This is a required attribute for [global policy rules](#). The node [attribute](#) must also be specified, and the only other allowed attribute is the optional `strict` attribute. See [Core Count Constraint](#) for possible attribute values.

### Per-collection Rule Attributes

The following attributes are usable with [per-collection policy rules](#), in addition to the attributes in the [Node Selection Attributes](#) section below:

`collection`

The name of the collection to which the policy rule should apply. If omitted, the rule applies to all collections. This attribute is optional.

`shard`

The name of the shard to which the policy rule should apply. If omitted, the rule is applied for all shards in the collection. It supports the special function `#EACH` which means that the rule is applied for each shard in the collection.

**type**

The type of the replica to which the policy rule should apply. If omitted, the rule is applied for all replica types of this collection/shard. The allowed values are NRT, TLOG and PULL

**replica**

The number of replicas that must exist to satisfy the rule. This is a required attribute for [per-collection rules](#). See [Replica Count Constraint](#) for possible attribute values.

**Node Selection Attributes**

One and only one of the following attributes can be specified in addition to the above attributes. See the [Node Selector](#) section for more information:

**node**

The name of the node to which the rule should apply. The [!\(not\) operator](#) or the [array operator](#) or the [#ANY function](#) may be used in this attribute's value.

**port**

The port of the node to which the rule should apply. The [!\(not\) operator](#) or the [array operator](#) may be used in this attribute's value.

**freedisk**

The free disk space in gigabytes of the node. This must be a positive 64-bit integer value, or a [percentage](#). If a percentage is specified, either an upper or lower bound may also be specified using the < or > operators, respectively, e.g., >50%, <25%.

**host**

The host name of the node.

**sysLoadAvg**

The system load average of the node as reported by the Metrics API under the key `solr.jvm/os.systemLoadAverage`. This is floating point value between 0 and 1.

**heapUsage**

The heap usage of the node as reported by the Metrics API under the key `solr.jvm/memory.heap.usage`. This is floating point value between 0 and 1.

**nodeRole**

The role of the node. The only supported value currently is `overseer`.

**ip\_1, ip\_2, ip\_3, ip\_4**

The least significant to most significant segments of IP address. For example, for an IP address `192.168.1.2`, `"ip_1": "2"`, `"ip_2": "1"`, `"ip_3": "168"`, `"ip_4": "192"`. The [array operator](#) may be used in any of these attributes' values.

**sysprop.<system\_property\_name>**

Any arbitrary system property set on the node on startup. The [!\(not\) operator](#) or the [array operator](#) may be used in this attribute's value.

**metrics:<full-path-to-the metric>**

Any arbitrary metric. For example, `metrics:solr.node:CONTAINER.fs.totalSpace`. Refer to the key

parameter in the [Metrics API](#) section.

#### diskType

The type of disk drive being used for Solr's coreRootDirectory. The only two supported values are rotational and ssd. Refer to coreRootDirectory parameter in the [Solr.xml Parameters](#) section. The ! (not) operator or the array operator may be used in this attribute's value.

Its value is fetched from the Metrics API with the key named solr.node:CONTAINER.fs.coreRoot.spins. The disk type is auto-detected by Lucene using various heuristics and it is not guaranteed to be correct across all platforms or operating systems. Refer to the [Dynamic defaults for ConcurrentMergeScheduler](#) section for more details.

### Policy Operators

Each attribute in the policy may specify one of the following operators along with the value.

- No operator means equality
- <: Less than
- >: Greater than
- !: Not
- Range operator (-): a value such as "3-5" means a value between 3 to 5 (inclusive). This is only supported in the replica and cores attributes.
- Array operator []: e.g., sysprop.zone= ["east", "west", "apac"]. This is equivalent to having multiple rules with each of these values. This can be used in the following attributes:
  - node
  - sysprop.\*
  - port
  - ip\_\*
  - diskType

### Special Functions

This supports values calculated at the time of execution.

- % : A certain percentage of the value. This is supported by the following attributes:
  - replica
  - cores
  - freedisk
- #ANY: Applies to the node [attribute](#) only. This means the rule applies to any node.
- #ALL: Applies to the replica [attribute](#) only. This means all replicas that meet the rule condition.
- #EACH: Applies to the shard [attribute](#) (meaning the rule should be evaluated separately for each shard), and to the attributes used to define the buckets for the [#EQUAL function](#) (meaning all possible values for the bucket-defining attribute).
- #EQUAL: Applies to the replica and cores attributes only. This means an equal number of replicas/cores in each bucket. The buckets can be defined using the below attributes with a value that can either be

#EACH or a list specified with the [array operator](#) ([]):

- node <- [global rules](#), i.e., those with the cores attribute, may only specify this attribute
- sysprop.\*
- port
- diskType
- ip\_\*

## Examples of Policy Rules

### Limit Replica Placement

Do not place more than one replica of the same shard on the same node. The rule is evaluated separately for [each](#) shard in each collection. The rule is applied to [any](#) node.

```
{"replica": "<2", "shard": "#EACH", "node": "#ANY"}
```

### Limit Cores per Node

Do not place more than 10 cores in [any](#) node. This rule can only be added to the cluster policy because it is a [global rule](#).

```
{"cores": "<10", "node": "#ANY"}
```

### Place Replicas Based on Port

Place exactly 1 replica of [each](#) shard of collection xyz on a node running on port 8983.

```
{"replica": 1, "shard": "#EACH", "collection": "xyz", "port": "8983"}
```

### Place Replicas Based on a System Property

Place [all](#) replicas on nodes with system property availability\_zone=us-east-1a.

```
{"replica": "#ALL", "sysprop.availability_zone": "us-east-1a"}
```

### Use Percentage

Place a maximum of (roughly) a third of the replicas of [each](#) shard in [any](#) node. In the following example, the value of replica is computed in real time as a percentage of the replicas of [each](#) shard of each collection:

```
{"replica": "33%", "shard": "#EACH", "node": "#ANY"}
```

If the number of replicas in a shard is 2, 33% of 2 = 0.66. This means a node may have a maximum of 1 and a minimum of 0 replicas of each shard.

It is possible to get the same effect by hard coding the value of replica as a decimal value:

```
{"replica": 0.66, "shard": "#EACH", "node": "#ANY"}
```

or using the [range operator](#):

```
{"replica": "0-1", "shard": "#EACH", "node": "#ANY"}
```

## Multiple Percentage Rules

Distribute replicas of [each](#) shard of each collection across datacenters east and west at a 1:2 ratio:

```
{"replica": "33%", "shard": "#EACH", "sysprop.zone": "east"}
 {"replica": "66%", "shard": "#EACH", "sysprop.zone": "west"}
```

For the above rules to work, all nodes must start with a system property called "zone"

## Distribute Replicas Equally in Each Zone

For [each](#) shard of each collection, distribute replicas equally across the east and west zones.

```
{"replica": "#EQUAL", "shard": "#EACH", "sysprop.zone": ["east", "west"]}
```

Distribute replicas equally across [each](#) zone.

```
{"replica": "#EQUAL", "shard": "#EACH", "sysprop.zone": "#EACH"}
```

## Place Replicas Based on Node Role

Do not place any replica on any node that has the overseer role. Note that the role is added by the `addRole` collection API. It is **not** automatically the node which is currently the overseer.

```
{"replica": 0, "nodeRole": "overseer"}
```

## Place Replicas Based on Free Disk

Place [all](#) replicas in nodes where [freedisk](#) is greater than 500GB.

```
{"replica": "#ALL", "freedisk": ">500"}
```

Keep all replicas in nodes where [freedisk](#) percentage is greater than 50%.

```
{"replica": "#ALL", "freedisk": ">50%"}
```

## Try to Place Replicas Based on Free Disk

When possible, place `all` replicas in nodes where `freedisk` is greater than 500GB. Here we use the `strict` attribute to signal that this rule is to be honored on a best effort basis.

```
{"replica": "#ALL", "freedisk": ">500", "strict": false}
```

## Place All Replicas of Type TLOG on Nodes with SSD Drives

```
{"replica": "#ALL", "type": "TLOG", "diskType": "ssd"}
```

## Place All Replicas of Type PULL on Nodes with Rotational Disk Drives

```
{"replica": "#ALL", "type": "PULL", "diskType": "rotational"}
```

## Defining Collection-Specific Policies

By default, the cluster policy, if it exists, is used automatically for all collections in the cluster. However, we can create named policies that can be attached to a collection at the time of its creation by specifying the policy name along with a `policy` parameter.

When a collection-specific policy is used, the rules in that policy are **appended** to the rules in the cluster policy and the combination of both are used. Therefore, it is recommended that you do not add rules to collection-specific policy that conflict with the ones in the cluster policy. Doing so will disqualify all nodes in the cluster from matching all criteria and make the policy useless.

It is possible to override rules specified in the cluster policy using collection-specific policy. For example, if a rule `{replica: '<3', node: '#ANY'}` is present in the cluster policy and the collection-specific policy has a rule `{replica: '<4', node: '#ANY'}`, the cluster policy is ignored in favor of the collection policy.

Also, if `maxShardsPerNode` is specified during the time of collection creation, then both `maxShardsPerNode` and the policy rules must be satisfied.

Some attributes such as `cores` can only be used in the cluster policy. See the section [Policy Rule Attributes](#) for details.

To create a new named policy, use the [set-policy API](#). Once you have a named policy, you can specify the `policy=<policy_name>` parameter to the `CREATE` command of the Collection API:

```
/admin/collections?action=CREATE&name=coll1&numShards=1&replicationFactor=2&policy=policy1
```

The above `CREATE` collection command will associate a policy named `policy1` with the collection named `coll1`. Only a single policy may be associated with a collection.

## Example: Manual Collection Creation with a Policy

The starting state for this example is a Solr cluster with 3 nodes: "nodeA", "nodeB", and "nodeC". An

existing 2-shard FirstCollection with a replicationFactor of 1 has one replica on "nodeB" and one on "nodeC". The default Autoscaling preferences are in effect:

```
[{"minimize": "cores"}]
```

The configured policy rule allows at most 1 core per node:

```
[{"cores": "<2", "node": "#ANY"}]
```

We now issue a CREATE command for a SecondCollection with two shards and a replicationFactor of 1:

```
http://localhost:8983/solr/admin/collections?action=CREATE&name=SecondCollection&numShards=2&replicationFactor=1
```

For each of the two replicas to be created, each Solr node is tested, in order from least to most loaded: would all policy rules be satisfied if a replica were placed there using an ADDREPLICA sub-command?

- ADDREPLICA for shard1: According to the Autoscaling preferences, the least loaded node is the one with the fewest cores: "nodeA", because it hosts no cores, while the other two nodes each host one core. The test to place a replica here succeeds, because doing so causes no policy violations, since the core count after adding the replica would not exceed the configured maximum of 1. Because "nodeA" can host the first shard's replica, Solr skips testing of the other two nodes.
- ADDREPLICA for shard2: After placing the shard1 replica, all nodes would be equally loaded, since each would have one core. The test to place the shard2 replica fails on each node, because placement would push the node over its maximum core count. This causes a policy violation.

Since there is no node that can host a replica for shard2 without causing a violation, the overall CREATE command fails. Let's try again after increasing the maximum core count on all nodes to 2:

```
[{"cores": "<3", "node": "#ANY"}]
```

After re-issuing the SecondCollection CREATE command, the replica for shard1 will be placed on "nodeA": it's least loaded, so is tested first, and no policy violation will result from placement there. The shard2 replica could be placed on any of the 3 nodes, since they're all equally loaded, and the chosen node will remain below its maximum core count after placement. The CREATE command succeeds.

## Testing Autoscaling Configuration and Suggestions

It's not always easy to predict the impact of autoscaling configuration changes on the cluster layout. Starting with release 8.1 Solr provides a tool for assessing the impact of such changes without affecting the state of the target cluster.

This testing tool is a part of bin/solr autoscaling command. In addition to other options that provide detailed status of the current cluster layout the following options specifically allow users to test new autoscaling configurations and run "what if" scenarios:

-a <CONFIG>

JSON file containing autoscaling configuration to test. This file needs to be in the same format as the result of the /solr/admin/autoscaling call. If this parameter is missing then the currently deployed autoscaling configuration is used.

-simulate

Simulate the effects of applying all autoscaling suggestions on the cluster layout. NOTE: this does not affect in any way the actual cluster - this option uses the simulation framework to calculate the new layout without actually making the changes. Calculations are performed in the tool's JVM so they don't affect the performance of the running cluster either. This process is repeated several times until a limit is reached or there are no more suggestions left to apply (although unresolved violations may still remain!)

-i <NUMBER>

Number of iterations of the simulation loop. Default is 10.

Results of the simulation contain the initial suggestions, suggestions at each step of the simulation and the final simulated state of the cluster.

## SolrCloud Autoscaling Triggers

Triggers are used in autoscaling to watch for cluster events such as nodes joining or leaving, search rate, index rate, on a schedule, or any other metric breaching a threshold.

Trigger implementations verify the state of resources that they monitor. When they detect a change that merits attention they generate *events*, which are then queued and processed by configured TriggerAction implementations. This usually involves computing and executing a plan to do something (e.g., move replicas). Solr provides predefined implementations of triggers for [specific event types](#).

Triggers execute on the node that runs Overseer. They are scheduled to run periodically, at a default interval of 1 second between each execution (although it's important to note that not every execution of a trigger produces events).

## Event Types

Currently the following event types (and corresponding trigger implementations) are defined:

- nodeAdded: generated when a node joins the cluster. See [Node Added Trigger](#).
- nodeLost: generated when a node leaves the cluster. See [Node Lost Trigger](#) and [Auto Add Replicas Trigger](#).
- metric: generated when the configured metric crosses a configured lower or upper threshold value. See [Metric Trigger](#).
- indexSize: generated when a shard size (defined as index size in bytes or number of documents) exceeds upper or lower threshold values. See [Index Size Trigger](#).
- searchRate: generated when the search rate exceeds configured upper or lower thresholds. See [Search Rate Trigger](#).
- scheduled: generated according to a scheduled time period such as every 24 hours, etc. See [Scheduled Trigger](#).

Events are not necessarily generated immediately after the corresponding state change occurred; the maximum rate of events is controlled by the `waitFor` configuration parameter (see [Trigger Configuration](#) below for more explanation).

The following properties are common to all event types:

`id`

(string) A unique time-based event id.

`eventType`

(string) The type of event.

`source`

(string) The name of the trigger that produced this event.

`eventTime`

(long) Unix time when the condition that caused this event occurred. For example, for a `nodeAdded` event this will be the time when the node was added and not when the event was actually generated, which may significantly differ due to the rate limits set by `waitFor`.

`properties`

(map, optional) Any additional properties. Currently includes e.g., `nodeNames` property that indicates the nodes that were lost or added.

## Trigger Configuration

Trigger configurations are managed using the [Autoscaling Write API](#) with the commands `set-trigger`, `remove-trigger`, `suspend-trigger`, and `resume-trigger`.

### Trigger Properties

Trigger configuration consists of the following properties:

`name`

(string, required) A unique trigger configuration name.

`event`

(string, required) One of the predefined event types (`nodeAdded` or `nodeLost`).

`actions`

(list of action configs, optional) An ordered list of actions to execute when event is fired.

`waitFor`

(string, optional) The time to wait between generating new events, as an integer number immediately followed by unit symbol, one of s (seconds), m (minutes), or h (hours). Default is 0s. A condition must persist at least for the `waitFor` period to generate an event.

`enabled`

(boolean, optional) When true the trigger is enabled. Default is true.

Additional implementation-specific properties may be provided, as described in the sections for individual triggers below.

## Action Properties

Action configuration consists of the following properties:

name

(string, required) A unique name of the action configuration.

class

(string, required) The action implementation class.

Additional implementation-specific properties may be provided, as described in the sections for individual triggers below.

If the actions configuration is omitted, then by default, the ComputePlanAction and the ExecutePlanAction are automatically added to the trigger configuration.

## Example Trigger Configuration

This simple example shows the configuration for adding (or updating) a trigger for nodeAdded events.

```
{
 "set-trigger": {
 "name" : "node_added_trigger",
 "event" : "nodeAdded",
 "waitFor" : "1s",
 "enabled" : true,
 "actions" : [
 {
 "name" : "compute_plan",
 "class": "solr.ComputePlanAction"
 },
 {
 "name" : "custom_action",
 "class": "com.example.CustomAction"
 },
 {
 "name" : "execute_plan",
 "class": "solr.ExecutePlanAction"
 }
]
 }
}
```

This trigger configuration will compute and execute a plan to allocate the resources available on the new node. A custom action could also be used to possibly modify the plan.

## Available Triggers

As described earlier, there are several triggers available to watch for events.

## Node Added Trigger

The NodeAddedTrigger generates nodeAdded events when a node joins the cluster. It can be used to either move replicas from other nodes to the new node or to add new replicas.

In addition to the parameters described at [Trigger Configuration](#), this trigger supports one more parameter: preferredOperation

(string, optional, defaults to movereplica) The operation to be performed in response to an event generated by this trigger. By default, replicas will be moved from other nodes to the added node. The only other supported value is addreplica which adds more replicas of the existing collections on the new node.

*Example: Node Added Trigger to move replicas to new node*

```
{
 "set-trigger": {
 "name": "node_added_trigger",
 "event": "nodeAdded",
 "waitFor": "5s"
 }
}
```

*Example: Node Added Trigger to add replicas on new node*

```
{
 "set-trigger": {
 "name": "node_added_trigger",
 "event": "nodeAdded",
 "waitFor": "5s",
 "preferredOperation": "ADDREPLICA"
 }
}
```

## Node Lost Trigger

The NodeLostTrigger generates nodeLost events when a node leaves the cluster. It can be used to either move replicas that were hosted by the lost node to other nodes or to delete them from the cluster.

In addition to the parameters described at [Trigger Configuration](#), this trigger supports the one more parameter:

preferredOperation

(string, optional, defaults to MOVEREPLICA) The operation to be performed in response to an event generated by this trigger. By default, replicas will be moved from the lost nodes to the other nodes in the cluster. The only other supported value is DELETENODE which deletes all information about replicas that were hosted by the lost node.

*Example: Node Lost Trigger to move replicas to new node*

```
{
 "set-trigger": {
 "name": "node_lost_trigger",
 "event": "nodeLost",
 "waitFor": "120s"
 }
}
```

*Example: Node Lost Trigger to delete replicas*

```
{
 "set-trigger": {
 "name": "node_lost_trigger",
 "event": "nodeLost",
 "waitFor": "120s",
 "preferredOperation": "DELETENODE"
 }
}
```



It is recommended that the value of waitFor configuration for the node lost trigger be larger than 1 minute so that large full garbage collection pauses do not cause this trigger to generate events and needlessly move or delete replicas in the cluster.

## Auto Add Replicas Trigger

When a collection has the parameter autoAddReplicas set to true then a trigger configuration named .auto\_add\_replicas is automatically created to watch for nodes going away. This trigger produces nodeLost events, which are then processed by configured actions (usually resulting in computing and executing a plan to add replicas on the live nodes to maintain the expected replication factor).

Refer to the section [Autoscaling Automatically Adding Replicas](#) to learn more about how the .autoAddReplicas trigger works.

In addition to the parameters described at [Trigger Configuration](#), this trigger supports one parameter, which is defined in the <solrcloud> section of solr.xml:

### autoReplicaFailoverWaitAfterExpiration

The minimum time in milliseconds to wait for initiating replacement of a replica after first noticing it not being live. This is important to prevent false positives while stopping or starting the cluster. The default is 120000 (2 minutes). The value provided for this parameter is used as the value for the waitFor parameter in the .auto\_add\_replicas trigger.



See [The <solrcloud> Element](#) for more details about how to work with solr.xml.

## Metric Trigger

The metric trigger can be used to monitor any metric exposed by the [Metrics API](#). It supports lower and upper threshold configurations as well as optional filters to limit operation to specific collection, shards, and

nodes.

In addition to the parameters described at [Trigger Configuration](#), this trigger supports the following parameters:

**metric**

(string, required) The metric property name to be watched in the format metric:*group:prefix*, e.g., metric:solr.node:CONTAINER.fs.coreRoot.usableSpace.

**below**

(double, optional) The lower threshold for the metric value. The trigger produces a metric breached event if the metric's value falls below this value.

**above**

(double, optional) The upper threshold for the metric value. The trigger produces a metric breached event if the metric's value crosses above this value.

**collection**

(string, optional) The collection used to limit the nodes on which the given metric is watched. When the metric is breached, trigger actions will limit operations to this collection only.

**shard**

(string, optional) The shard used to limit the nodes on which the given metric is watched. When the metric is breached, trigger actions will limit operations to this shard only.

**node**

(string, optional) The node on which the given metric is watched. Trigger actions will operate on this node only.

**preferredOperation**

(string, optional, defaults to MOVEREPLICA) The operation to be performed in response to an event generated by this trigger. By default, replicas will be moved from the hot node to others. The only other supported value is ADDREPLICA which adds more replicas if the metric is breached.

*Example: a metric trigger that fires when total usable space on a node having replicas of "mycollection" falls below 100GB*

```
{
 "set-trigger": {
 "name": "metric_trigger",
 "event": "metric",
 "waitFor": "5s",
 "metric": "metric:solr.node:CONTAINER.fs.coreRoot.usableSpace",
 "below": 107374182400,
 "collection": "mycollection"
 }
}
```

## Index Size Trigger

This trigger can be used for monitoring the size of collection shards, measured either by the number of

documents in a shard or the physical size of the shard's index in bytes.

When either of the upper thresholds is exceeded the trigger will generate an event with a (configurable) requested operation to perform on the offending shards - by default this is a SPLITSHARD operation.

Similarly, when either of the lower thresholds is exceeded the trigger will generate an event with a (configurable) requested operation to perform on two of the smallest shards. By default this is a MERGESHARDS operation, and is currently ignored because that operation is not yet implemented (see [SOLR-9407](#)).

Additionally, monitoring can be restricted to a list of collections; by default all collections are monitored.

In addition to the parameters described at [Trigger Configuration](#), this trigger supports the following configuration parameters (all thresholds are exclusive):

#### aboveBytes

A upper threshold in bytes. This value is compared to the INDEX.sizeInBytes metric.

#### belowBytes

A lower threshold in bytes. Note that this value should be at least 2x smaller than aboveBytes

#### aboveDocs

An upper threshold expressed as the number of documents. This value is compared with SEARCHER.searcher.numDocs metric.



Due to the way Lucene indexes work, a shard may exceed the aboveBytes threshold even if the number of documents is relatively small, because replaced and deleted documents keep occupying disk space until they are actually removed during Lucene index merging.

#### belowDocs

A lower threshold expressed as the number of documents.

#### aboveOp

The operation to request when an upper threshold is exceeded. If not specified the default value is SPLITSHARD.

#### belowOp

The operation to request when a lower threshold is exceeded. If not specified the default value is MERGESHARDS (but see the note above).

#### collections

A comma-separated list of collection names that this trigger should monitor. If not specified or empty all collections are monitored.

#### maxOps

Maximum number of operations requested in a single event. This property limits the speed of changes in a highly dynamic situation, which may lead to more serious threshold violations, but it also limits the maximum load on the cluster that the large number of requested operations may cause. The default value is 10.

### splitMethod

One of the supported methods for index splitting to use. Default value is rewrite, which is slow and puts a high CPU load on the shard leader but results in optimized sub-shard indexes. The link method is much faster and puts very little load on the shard leader but results in indexes that are initially as large as the parent shard's index, which slows down replication and may lead to excessive initial disk space consumption on replicas.

### splitFuzz

A float value (default is 0.0f, must be smaller than 0.5f) that allows to vary the sub-shard ranges by this percentage of total shard range, odd shards being larger and even shards being smaller. Non-zero values are useful for large indexes with aggressively growing size, as they help to prevent avalanches of split shard requests when the total size of the index reaches even multiples of the maximum shard size thresholds.

Events generated by this trigger contain additional details about the shards that exceeded thresholds and the types of violations (upper / lower bounds, bytes / docs metrics).

#### *Example: Index Size Trigger*

This configuration specifies an index size trigger that monitors collections "test1" and "test2", with both bytes (1GB) and number of docs (1 million) upper limits, and a custom belowOp operation NONE (which still can be monitored and acted upon by an appropriate trigger listener):

```
{
 "set-trigger": {
 "name" : "index_size_trigger",
 "event" : "indexSize",
 "collections" : "test1,test2",
 "aboveBytes" : 1000000000,
 "aboveDocs" : 1000000000,
 "belowBytes" : 200000,
 "belowDocs" : 200000,
 "belowOp" : "NONE",
 "waitFor" : "1m",
 "enabled" : true,
 "actions" : [
 {
 "name" : "compute_plan",
 "class": "solr.ComputePlanAction"
 },
 {
 "name" : "execute_plan",
 "class": "solr.ExecutePlanAction"
 }
]
 }
}
```

### Search Rate Trigger

The search rate trigger can be used for monitoring search rates in a selected collection (1-min average rate

by default), and request that either replicas be moved from "hot nodes" to different nodes, or new replicas be added to "hot shards" to reduce the per-replica search rate for a collection or shard with hot spots.

Similarly, if the search rate falls below a threshold then the trigger may request that some replicas are deleted from "cold" shards. It can also optionally issue node-level action requests when a cumulative node-level rate falls below a threshold.

Per-shard rates are calculated as arithmetic average of rates of all searchable replicas in a given shard. This method was chosen to avoid generating false events when a simple client keeps sending requests to a single specific replica (because adding or removing other replicas can't solve this situation, only proper load balancing can - either by using CloudSolrClient or another load-balancing client).

This trigger calculates node-level cumulative rates using per-replica rates reported by replicas that are part of monitored collections / shards on each node. This means that it may report some nodes as "cold" (underutilized) because it ignores other, perhaps more active, replicas belonging to other collections. Also, nodes that don't host any of the monitored replicas or those that are explicitly excluded by node configuration property won't be reported at all.

#### *Calculating* waitFor

Special care should be taken when configuring the waitFor property. By default the trigger monitors a 1-minute average search rate of a replica. Changes to the number of replicas that should in turn change per-replica search rates may be requested and executed relatively quickly if the waitFor is set to comparable values of 1 min or shorter.

However, the metric value, being a moving average, will always lag behind the new "momentary" rate after the changes. This in turn means that the monitored metric may not change sufficiently enough to prevent the trigger from firing again, because it will continue to measure the average rate as still violating the threshold for some time after the change was executed. As a result the trigger may keep requesting that even more replicas be added (or removed) and thus it may "overshoot" the optimal number of replicas.

For this reason it's recommended to always set waitFor to values several times longer than the time constant of the used metric. For example, with the default 1-minute average the waitFor should be set to at least 2m (2 minutes) or more.

In addition to the parameters described at [Trigger Configuration](#), this trigger supports the following configuration properties:

#### collections

(string, optional) A comma-separated list of collection names to monitor, or any collection if empty or not set.

#### shard

(string, optional) A shard name within the collection (requires collections to be set to exactly one name), or any shard if empty.

#### node

(string, optional) A node name to monitor, or any if empty.

#### metric

(string, optional) A metric name that represents the search rate. The default is QUERY./select.requestTimes:1minRate. This name has to identify a single numeric metric value, and it may use the colon syntax for selecting one property of a complex metric. This value is collected from all replicas for a shard, and then an arithmetic average is calculated per shard to determine shard-level violations.

#### maxOps

(integer, optional) The maximum number of ADDREPLICA or DELETEREPLICA operations requested in a single autoscaling event. The default value is 3 and it helps to smooth out the changes to the number of replicas during periods of large search rate fluctuations.

#### minReplicas

(integer, optional) The minimum acceptable number of searchable replicas (i.e., replicas other than PULL type). The trigger will not generate any DELETEREPLICA requests when the number of searchable replicas in a shard reaches this threshold.

When this value is not set (the default) the replicationFactor property of the collection is used, and if that property is not set then the value is set to 1. Note also that shard leaders are never deleted.

#### aboveRate

(float) The upper bound for the request rate metric value. At least one of aboveRate or belowRate must be set.

#### belowRate

(float) The lower bound for the request rate metric value. At least one of aboveRate or belowRate must be set.

#### aboveNodeRate

(float) The upper bound for the total request rate metric value per node. If not set then cumulative per-node rates will be ignored.

#### belowNodeRate

(float) The lower bound for the total request rate metric value per node. If not set then cumulative per-node rates will be ignored.

#### aboveOp

(string, optional) A collection action to request when the upper threshold for a shard is exceeded. Default action is ADDREPLICA and the trigger will request from 1 up to maxOps operations per shard per event, proportionally to how much the rate is exceeded. This property can be set to 'NONE' to effectively disable the action but still report it to the listeners.

#### aboveNodeOp

(string, optional) The collection action to request when the upper threshold for a node (aboveNodeRate) is exceeded. Default action is MOVEREPLICA, and the trigger will request 1 replica operation per hot node per event. If both aboveOp and aboveNodeOp operations are to be requested then aboveNodeOp operations are always requested first, and only if no aboveOp (shard level) operations are to be requested (because aboveOp operations will change node-level rates anyway). This property can be set to 'NONE' to effectively disable the action but still report it to the listeners.

#### belowOp

(string, optional) The collection action to request when the lower threshold for a shard is exceeded.

Default action is DELETEREPLICA, and the trigger will request at most maxOps replicas to be deleted from eligible cold shards. This property can be set to 'NONE' to effectively disable the action but still report it to the listeners.

#### belowNodeOp

(string, optional) The action to request when the lower threshold for a node (belowNodeRate) is exceeded. Default action is null (not set) and the condition is ignored, because in many cases the trigger will monitor only some selected resources (replicas from selected collections or shards) so setting this by default to e.g., DELETENODE could interfere with these non-monitored resources. The trigger will request 1 operation per cold node per event. If both belowOp and belowNodeOp operations are requested then belowOp operations are always requested first.

#### Example:

A search rate trigger that monitors collection "test" and adds new replicas if 5-minute average request rate of "/select" handler exceeds 100 requests/sec, and the condition persists for over 20 minutes. If the rate falls below 0.01 and persists for 20 min the trigger will request not only replica deletions (leaving at most 1 replica per shard) but also it may request node deletion.

```
{
 "set-trigger": {
 "name" : "search_rate_trigger",
 "event" : "searchRate",
 "collections" : "test",
 "metric" : "QUERY./select.requestTimes:5minRate",
 "aboveRate" : 100.0,
 "belowRate" : 0.01,
 "belowNodeRate" : 0.01,
 "belowNodeOp" : "DELETENODE",
 "minReplicas" : 1,
 "waitFor" : "20m",
 "enabled" : true,
 "actions" : [
 {
 "name" : "compute_plan",
 "class": "solr.ComputePlanAction"
 },
 {
 "name" : "execute_plan",
 "class": "solr.ExecutePlanAction"
 }
]
 }
}
```

## Scheduled Trigger

The Scheduled trigger generates events according to a fixed rate schedule.

In addition to the parameters described at [Trigger Configuration](#), this trigger supports the following configuration:

**startTime**

(string, required) The start date/time of the schedule. This should either be a DateMath string e.g., 'NOW', or be an ISO-8601 date time string (the same standard used during search and indexing in Solr, which defaults to UTC), or be specified without the trailing 'Z' accompanied with the timeZone parameter. For example, each of the following values are acceptable:

- 2018-01-31T15:30:00Z: ISO-8601 date time string. The trailing Z signals that the time is in UTC
- NOW+5MINUTES: Solr's date math string
- 2018-01-31T15:30:00: No trailing 'Z' signals that the timeZone parameter must be specified to avoid ambiguity

**every**

(string, required) A positive Solr date math string which is added to the startTime or the last run time to arrive at the next scheduled time.

**graceTime**

(string, optional) A positive Solr date math string. This is the additional grace time over the scheduled time within which the trigger is allowed to generate an event.

**timeZone**

(string, optional) A time zone string which is used for calculating the scheduled times.

**preferredOperation**

(string, optional, defaults to MOVEREPLICA) The preferred operation to perform in response to an event generated by this trigger. The only supported values are MOVEREPLICA or ADDREPLICA.

This trigger applies the every date math expression on the startTime or the last event time to derive the next scheduled time and if current time is greater than next scheduled time but within graceTime then an event is generated.

Apart from the common event properties described in the [Event Types](#) section, the trigger adds an additional actualEventTime event property which has the actual event time as opposed to the scheduled time.

For example, if the scheduled time was 2018-01-31T15:30:00Z and grace time was +15MINUTES then an event may be fired at 2018-01-31T15:45:00Z. Such an event will have eventTime as 2018-01-31T15:30:00Z, the scheduled time, but the actualEventTime property will have a value of 2018-01-31T15:45:00Z, the actual time.

### *Frequently scheduled events and trigger starvation*

Be cautious with scheduled triggers that are set to run as or more frequently than the trigger cooldown period (defaults to 5 seconds).

 Solr pauses all triggers for a cooldown period after a trigger fires so that the system has some time to stabilize. An aggressive scheduled trigger can starve all other triggers from ever executing if a new scheduled event is ready as soon as the cooldown period is over. The same starvation scenario can happen to the scheduled trigger as well.

Solr randomizes the order in which the triggers are resumed after the cooldown period to mitigate this problem. However, it is recommended that scheduled triggers are not used with low every values and an external scheduling process such as cron be used for such cases instead.

## SolrCloud Autoscaling Trigger Actions

TriggerAction implementations process events generated by triggers in order to ensure the cluster's health and good use of resources.

Currently two implementations are provided: ComputePlanAction and ExecutePlanAction.

### Compute Plan Action

The ComputePlanAction uses the policy and preferences to calculate the optimal set of Collection API commands which can re-balance the cluster in response to trigger events.

The following parameters are configurable:

#### `collections`

A comma-separated list of collection names. If this list is not empty then the computed operations will only calculate collection operations that affect listed collections and ignore any other collection operations for collections not listed here. Note that non-collection operations are not affected by this.

Example configuration:

```
{
 "set-trigger" : {
 "name" : "node_added_trigger",
 "event" : "nodeAdded",
 "waitFor" : "1s",
 "enabled" : true,
 "actions" : [
 {
 "name" : "compute_plan",
 "class" : "solr.ComputePlanAction",
 "collections" : "test1,test2",
 },
 {
 "name" : "execute_plan",
 "class" : "solr.ExecutePlanAction",
 }
]
 }
}
```

In this example only collections test1 and test2 will be potentially replicated / moved to an added node, other collections will be ignored even if they cause policy violations.

## Execute Plan Action

The ExecutePlanAction executes the Collection API commands emitted by the ComputePlanAction against the cluster using SolrJ. It executes the commands serially, waiting for each of them to succeed before continuing with the next one.

Currently, it has no configurable parameters.

If any one of the commands fail, then the complete chain of actions are executed again at the next run of the trigger. If the Overseer node fails while ExecutePlanAction is running, then the new Overseer node will run the chain of actions for the same event again after waiting for any running Collection API operations belonging to the event to complete.

Please see [SolrCloud Autoscaling Fault Tolerance](#) for more details on fault tolerance within the autoscaling framework.

## SolrCloud Autoscaling Listeners

Trigger Listeners allow users to configure additional behavior related to trigger events as they are being processed.

For example, users may want to record autoscaling events to an external system, or notify an administrator when a particular type of event occurs or when its processing reaches certain stage (e.g., failed).

Listener configuration always refers to a specific trigger configuration because a listener is notified of events generated by that specific trigger. Several (or none) named listeners can be registered for a trigger, and they will be notified in the order in which they were defined.

Listener configuration can specify what processing stages are of interest, and when an event enters this processing stage the listener will be notified. Currently the following stages are recognized:

- STARTED - when an event has been generated by a trigger and its processing is starting.
- ABORTED - when event was being processed while the source trigger closed.
- BEFORE\_ACTION - when a TriggerAction is about to be invoked. Action name and the current ActionContext are passed to the listener.
- AFTER\_ACTION - after a TriggerAction has been successfully invoked. Action name, ActionContext and the list of action names invoked so far are passed to the listener.
- FAILED - when event processing failed (or when a TriggerAction failed)
- SUCCEEDED - when event processing completes successfully

Listener configuration can also specify what particular actions are of interest, both before and/or after they are invoked.

## Listener Configuration

Currently the following listener configuration properties are supported:

`name`

(string, required) A unique listener configuration name.

`trigger`

(string, required) The name of an existing trigger configuration.

`class`

(string, required) A listener implementation class name.

`stage`

(list of strings, optional, ignored case) A list of processing stages that this listener should be notified. Default is empty list.

`beforeAction`

(list of strings, optional) A list of action names (as defined in trigger configuration) before which the listener will be notified. Default is empty list.

`afterAction`

(list of strings, optional) A list of action names after which the listener will be notified. Default is empty list.



Additional implementation-specific properties may be provided, depending on the listener implementation.

Note: when both stage and beforeAction / afterAction lists are non-empty then the listener will be notified both when a specified stage is entered and before / after specified actions.

## Managing Listener Configurations

Listener configurations are managed using the Autoscaling Write API, and using set-listener and remove-

listener commands.

For example:

```
{
 "set-listener": {
 "name": "foo",
 "trigger": "node_lost_trigger",
 "stage": ["STARTED", "ABORTED", "SUCCEEDED", "FAILED"],
 "class": "solr.SystemLogListener"
 }
}
```

```
{
 "remove-listener": {
 "name": "foo"
 }
}
```

## Listener Implementations

Trigger listeners must implement the TriggerListener interface. Solr provides some implementations of trigger listeners, which cover common use cases. These implementations are described below, together with their configuration parameters.

### SystemLogListener

This trigger listener sends trigger events and processing context as documents for indexing in SolrCloud .system collection.

When a trigger configuration is first created, a corresponding trigger listener configuration that uses SystemLogListener is also automatically created, to make sure that all events and actions related to the autoscaling framework are logged to the .system collection.

Supported configuration properties:

collection

(string, optional) Specifies the target collection where documents are sent. Default value is .system. If the target collection is missing the listener will silently discard events.



In rare situations when the target collection is in an unstable state (e.g., when some leader replicas were just lost and the leader election hasn't finished running yet), the listener may not be able to index some events. In such cases a WARN message with the details of the event(s) will be added to the regular logs.

enabled

(boolean, optional) Enables the listener when true. Default value is true.

Documents created by this listener have several predefined fields:

- id - time-based random id
- type - always set to autoscaling\_event
- source\_s - always set to SystemLogListener
- timestamp - current time when document was created
- stage\_s - current stage of event processing
- action\_s - current action name, if available
- message\_t - optional additional message
- error.message\_t - message from Throwable, if available
- error.details\_t - stacktrace from Throwable, if available
- before.actions\_ss - list of action names to be invoked so far
- after.actions\_ss - list of action names that have been successfully invoked so far
- event\_str - JSON representation of all event properties
- context\_str - JSON representation of all ActionContext properties, if available

The following fields are created using the information from trigger event:

- event.id\_s - event id
- event.type\_s - event type
- event.source\_s - event source (trigger name)
- event.time\_l - Unix time when the event was created (may significantly differ from the time when it was actually processed)
- event.property.\* - additional fields that represent other arbitrary event properties. These fields use either \_s or \_ss suffix depending on whether the property value is a collection (values inside collection are treated as strings, there's no recursive flattening)

The following configuration is used for the automatically created listener (in this case for a trigger named foo):

```
{
 "name" : "foo.system",
 "trigger" : "solr.SystemLogListener",
 "stage" : ["STARTED", "ABORTED", "SUCCEEDED", "FAILED", "BEFORE_ACTION", "AFTER_ACTION"]
}
```

## HttpTriggerListener

This listener uses HTTP POST to send a representation of the event and context to a specified URL. The URL, payload, and headers may contain property substitution patterns, which are then replaced with values taken from the current event or context properties.

Templates use the same syntax as property substitution in Solr configuration files, e.g., \${foo.bar:baz} means that the value of foo.bar property should be taken, and baz should be used if the value is absent.

Supported configuration properties:

url

(string, required) A URL template.

payload

(string, optional) A payload template. If absent, a JSON map of all properties listed above will be used.

contentType

(string, optional) A payload content type. If absent then application/json will be used.

header .\*

(string, optional) A header template(s). The name of the property without "header." prefix defines the literal header name.

timeout

(int, optional) Connection and socket timeout in milliseconds. Default is 60000 milliseconds (60 seconds).

followRedirects

(boolean, optional) Allows following redirects. Default is false.

The following properties are available in context and can be referenced from templates:

- config.\* - listener configuration properties
- event.\* - current event properties
- stage - current stage of event processing
- actionPerformed - optional current action name
- context.\* - optional ActionContext properties
- error - optional error string (from Throwable.toString())
- message - optional message

#### *Example HttpTriggerListener*

```
{
 "name": "foo",
 "trigger": "node_added_trigger",
 "class": "solr.HttpTriggerListener",
 "url":
 "http://foo.com/${config.name:invalidName}/${config.properties.xyz:invalidXyz}/${event.eventType}
 ",
 "xyz": "foobar",
 "header.X-Trigger": "${config.trigger}",
 "payload": "actionName=${actionName}, source=${event.source}, type=${event.eventType}",
 "contentType": "text/plain",
 "stage": ["STARTED", "ABORTED", "SUCCEEDED", "FAILED"],
 "beforeAction": ["compute_plan", "execute_plan"],
 "afterAction": ["compute_plan", "execute_plan"]
}
```

This configuration specifies that each time one of the listed stages is reached, or before and after each of the listed actions is executed, the listener will send the templated payload to a URL that also depends on the

configuration and the current event, and with a custom header that indicates the trigger name.

## SolrCloud Autoscaling Automatically Adding Replicas

Solr provides a way to automatically add replicas for a collection when the number of active replicas drops below the replication factor specified at the time of the creation of the collection.

### The `autoAddReplicas` Parameter

The boolean `autoAddReplicas` parameter can be passed to the `CREATE` command of the Collection API to enable this feature for a given collection.

*Create a collection with `autoAddReplicas` enabled*

```
http://localhost:8983/solr/admin/collections?action=CREATE&name=my_collection&numShards=1&replica
tionFactor=5&autoAddReplicas=true
```

The `MODIFYCOLLECTION` command can be used to enable or disable this feature for any collection.

*Modify collection to disable `autoAddReplicas`*

```
http://localhost:8983/solr/admin/collections?action=MODIFYCOLLECTION&collection=my_collection&aut
oAddReplicas=false
```

### Implementation Using `.autoAddReplicas` Trigger

A Trigger named `.autoAddReplicas` is automatically created whenever any collection has the `autoAddReplicas` feature enabled.

Only one trigger is sufficient to serve all collections having this feature enabled. The `.autoAddReplicas` trigger watches for nodes that are lost from the cluster and uses the default `TriggerActions` to create new replicas to replace the ones which were hosted by the lost node. If the old node comes back online, it unloads the moved replicas and the node is free to host other replicas as and when required.

Since the trigger provides the `autoAddReplicas` feature for all collections, the `suspend-trigger` and `resume-trigger` Autoscaling API commands can be used to disable and enable this feature for all collections in one API call.

*Suspending `autoAddReplicas` for all collections*

```
{
 "suspend-trigger": {
 "name" : ".autoAddReplicas"
 }
}
```

## Resuming autoAddReplicas for all collections

```
{
 "resume-trigger": {
 "name" : ".autoAddReplicas"
 }
}
```

## Using Cluster Property to Enable autoAddReplicas

A cluster property, also named autoAddReplicas, can be set to false to disable this feature for all collections. If this cluster property is missing or set to true, the autoAddReplicas is enabled for all collections.

### *Deprecation Warning*



Using a cluster property to enable or disable autoAddReplicas is deprecated and only supported for back compatibility. Please use the suspend-trigger and resume-trigger API commands instead.

## SolrCloud Autoscaling Fault Tolerance

The autoscaling framework uses a few strategies to ensure it's able to still trigger actions in the event of unexpected changes to the system.

### Node Added or Lost Markers

Since triggers execute on the node that runs the Overseer, should the Overseer node go down the nodeLost event would be lost because there would be no mechanism to generate it. Similarly, if a node has been added before the Overseer leader change was completed, the nodeAdded event would not be generated.

For this reason Solr implements additional mechanisms to ensure that these events are generated reliably.

With standard SolrCloud behavior, when a node joins a cluster its presence is marked as an ephemeral ZooKeeper path in the /live\_nodes/<nodeName> ZooKeeper directory. Now an ephemeral path is also created under /autoscaling/nodeAdded/<nodeName>. When a new instance of Overseer leader is started it will run the nodeAdded trigger (if it's configured) and discover the presence of this ZooKeeper path, at which point it will remove it and generate a nodeAdded event.

When a node leaves the cluster, up to three remaining nodes will try to create a persistent ZooKeeper path /autoscaling/nodeLost/<nodeName> and eventually one of them succeeds. When a new instance of Overseer leader is started it will run the nodeLost trigger (if it's configured) and discover the presence of this ZooKeeper path, at which point it will remove it and generate a nodeLost event.

### Trigger State Checkpointing

Triggers generate events based on their internal state. If the Overseer leader goes down while the trigger is about to generate a new event, it's likely that the event would be lost because a new trigger instance running on the new Overseer leader would start from a clean slate.

For this reason, after each time a trigger is executed its internal state is persisted to ZooKeeper, and on Overseer start its internal state is restored.

## Trigger Event Queues

Autoscaling framework limits the rate at which events are processed using several different mechanisms. One is the locking mechanism that prevents concurrent processing of events, and another is a single-threaded executor that runs trigger actions.

This means that the processing of an event may take significant time, and during this time it's possible that the Overseer may go down. In order to avoid losing events that were already generated but not yet fully processed, events are queued before processing is started.

Separate ZooKeeper queues are created for each trigger, and events produced by triggers are put on these per-trigger queues. When a new Overseer leader is started it will first check these queues and process events accumulated there, and only then it will continue to run triggers normally. Queued events that fail processing during this "replay" stage are discarded.

## Autoscaling API

The Autoscaling API is used to manage autoscaling policies, preferences, triggers, listeners and to get diagnostics on the state of the cluster.

### Read API

The autoscaling Read API is available at `/solr/admin/autoscaling` or `/api/cluster/autoscaling` (v2 API style). It returns information about the configured cluster preferences, cluster policy, collection-specific policies triggers and listeners.

This API does not take any parameters.

### Read API Response

The output will contain cluster preferences, cluster policy and collection specific policies.

### Examples using Read API

#### Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 2
 },
 "cluster-policy": [
 {
 "replica": "<2",
 "shard": "#EACH",
 "node": "#ANY"
 }
],
 "WARNING": "This response format is experimental. It is likely to change in the future."
}
```

## Diagnostics API

The diagnostics API shows the violations, if any, of all conditions in the cluster and, if applicable, the collection-specific policy. It is available at the /admin/autoscaling/diagnostics path.

This API does not take any parameters.

### Diagnostics API Response

The output will contain sortedNodes which is a list of nodes in the cluster sorted according to overall load in descending order (as determined by the preferences) and violations which is a list of nodes along with the conditions that they violate.

### Examples Using Diagnostics API

Here is an example with no violations but in the sortedNodes section, we can see that the first node is most loaded (according to number of cores):

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 65
 },
 "diagnostics": {
 "sortedNodes": [
 {
 "node": "127.0.0.1:8983_solr",
 "cores": 3
 },
 {
 "node": "127.0.0.1:7574_solr",
 "cores": 2
 }

```

Suppose we added a condition to the cluster policy as follows:

```
{"replica": "<2", "shard": "#EACH", "node": "#ANY"}
```

However, since the first node in the first example had more than 1 replica for a shard already, then the diagnostics API will return:

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 45
 },
 "diagnostics": {
 "sortedNodes": [
 {
 "node": "127.0.0.1:8983_solr",
 "cores": 3
 },
 {
 "node": "127.0.0.1:7574_solr",
 "cores": 2
 }
],
 "violations": [
 {
 "collection": "gettingstarted",
 "shard": "shard1",
 "node": "127.0.0.1:8983_solr",
 "tagKey": "127.0.0.1:8983_solr",
 "violation": {
 "replica": "2",
 "delta": 0
 },
 "clause": {
 "replica": "<2",
 "shard": "#EACH",
 "node": "#ANY",
 "collection": "gettingstarted"
 }
 }
]
 },
 "WARNING": "This response format is experimental. It is likely to change in the future."
}
```

In the above example the node with port 8983 has two replicas for shard1 in violation of our policy.

## Suggestions API

Suggestions are operations recommended by the system according to the policies and preferences the user has set.

Suggestions are made only if there are violations to active policies. The operation section of the response uses the defined preferences to identify the target node.

The API is available at /admin/autoscaling/suggestions. Here is an example output from a suggestion request:

```
{
 "responseHeader":{
 "status":0,
 "QTime":101},
 "suggestions":[]
 }
 [{"
 "type": "violation",
 "violation":{
 "collection": "mycoll",
 "shard": "shard2",
 "tagKey": "7574",
 "violation":{ "delta": -1},
 "clause":{
 "replica": "0",
 "shard": "#EACH",
 "port": 7574,
 "collection": "mycoll"}},
 "operation":{
 "method": "POST",
 "path": "/c/mycoll",
 "command":{ "move-replica":{
 "targetNode": "192.168.43.37:8983_solr",
 "replica": "core_node7"}}}},
 {}
 [{"
 "type": "violation",
 "violation":{
 "collection": "mycoll",
 "shard": "shard2",
 "tagKey": "7574",
 "violation":{ "delta": -1},
 "clause":{
 "replica": "0",
 "shard": "#EACH",
 "port": 7574,
 "collection": "mycoll"}},
 "operation":{
 "method": "POST",
 "path": "/c/mycoll",
 "command":{ "move-replica":{
 "targetNode": "192.168.43.37:7575_solr",
 "replica": "core_node15"}}}]},
 "WARNING": "This response format is experimental. It is likely to change in the future."}
```

The suggested operation is an API call that can be invoked to remedy the current violation.

## History API

The history of autoscaling events is available at /admin/autoscaling/history. It returns information about past autoscaling events and details about their processing. This history is kept in the .system collection, and is populated by a trigger listener SystemLogListener. By default this listener is added to all new triggers.

History events are regular Solr documents so they can be also accessed directly by searching on the .system

collection. The history handler acts as a regular search handler, so all query parameters supported by /select handler for that collection are supported here too. However, the history handler makes this process easier by offering a simpler syntax and knowledge of field names used by SystemLogListener for serialization of event data.

History documents contain the action context, if it was available, which gives further insight into e.g., exact operations that were computed and/or executed.

Specifically, the following query parameters can be used (they are turned into filter queries, so an implicit AND is applied):

`trigger`

The name of the trigger.

`eventType`

The event type or trigger type (e.g., nodeAdded).

`collection`

The name of the collection involved in event processing.

`stage`

An event processing stage.

`action`

A trigger action.

`node`

A node name that the event refers to.

`beforeAction`

A beforeAction stage.

`afterAction`

An afterAction stage.

### *Example output*

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 64
 },
 "response": {
 "numFound": 2,
 "start": 0,
 "docs": [
 {
 "type": "autoscaling_event",
 "source_s": "SystemLogListener",
 "id": "15f53efdf4bT2qlmj80580yuu997vktddfob3",
 "event.id_s": "14f0d67fe7b97d80T2qlmj80580yuu997vktddfob2",
 "event.type_s": "NODELOST",
 "score": 1.0
 }
]
 }
}
```

```
"event.source_s": ".auto_add_replicas",
"event.time_l": 1508941720006000000,
"timestamp": "2017-10-25T14:29:10.091Z",
"event.property.eventTimes_ss": [
 "1508941720006000000"
],
"event.property._enqueue_time__ss": [
 "1508941750088000000"
],
"event.property.nodeName_ss": [
 "192.168.1.104:7574_solr"
],
"stage_s": "STARTED",
"event_str": "{\n \"id\": \"14f0d67fe7b97d80T2qlmj80580yuu997vktddfob2\", \n \"source\": \".auto_add_replicas\", \n \"eventTime\": 1508941720006000000, \n \"eventType\": \"NODELOST\", \n \"properties\": {\n \"eventTimes\": [1508941720006000000], \n \"_enqueue_time_\": 1508941750088000000, \n \"nodeNames\": [\"192.168.1.104:7574_solr\"]\n }\n},\n{\n \"type\": \"autoscaling_event\",\n \"source_s\": \"SystemLogListener\",\n \"id\": \"15f53eff316T2qlmj80580yuu997vktddfob6\",\n \"event.id_s\": \"14f0d67fe7b97d80T2qlmj80580yuu997vktddfob2\",\n \"event.type_s\": \"NODELOST\",\n \"event.source_s\": \".auto_add_replicas\",\n \"event.time_l\": 1508941720006000000,\n \"timestamp\": \"2017-10-25T14:29:15.158Z\",\n \"event.property.eventTimes_ss\": [\n \"1508941720006000000\"\n],\n \"event.property._enqueue_time__ss\": [\n \"1508941750088000000\"\n],\n \"event.property.nodeName_ss\": [\n \"192.168.1.104:7574_solr\"\n],\n \"stage_s\": \"SUCCEEDED\",\n \"event_str\": \"{\\n \"id\": \"14f0d67fe7b97d80T2qlmj80580yuu997vktddfob2\", \\n \"source\": \".auto_add_replicas\", \\n \"eventTime\": 1508941720006000000, \\n \"eventType\": \"NODELOST\", \\n \"properties\": {\\n \"eventTimes\": [1508941720006000000], \\n \"_enqueue_time_\": 1508941750088000000, \\n \"nodeNames\": [\"192.168.1.104:7574_solr\"]\\n }\\n}\\n}\n]\n}\n}
```

## Write API

The Write API is available at the same /admin/autoscaling and /api/cluster/autoscaling endpoints as

the Read API but can only be used with the **POST** HTTP verb.

The payload of the POST request is a JSON message with commands to set and remove components. Multiple commands can be specified together in the payload. The commands are executed in the order specified and the changes are atomic, i.e., either all succeed or none.

## Create and Modify Cluster Preferences

Cluster preferences are specified as a list of sort preferences. Multiple sorting preferences can be specified and they are applied in the order they are set.

They are defined using the `set-cluster-preferences` command.

Each preference is a JSON map having the following syntax:

```
{'<sort_order>': '<sort_param>', 'precision': '<precision_val>'}
```

See the section [Cluster Preferences Specification](#) for details about the allowed values for the `sort_order`, `sort_param` and `precision` parameters.

Changing the cluster preferences after the cluster is already built doesn't automatically reconfigure the cluster. However, all future cluster management operations will use the changed preferences.

### Input

```
{
 "set-cluster-preferences" : [
 {"minimize": "cores"}
]
}
```

### Output

The output has a key named `result` which will return either `success` or `failure` depending on whether the command succeeded or failed.

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 138
 },
 "result": "success",
 "WARNING": "This response format is experimental. It is likely to change in the future."
}
```

## Example Setting Cluster Preferences

In this example we add cluster preferences that sort on three different parameters:

```
{
 "set-cluster-preferences": [
 {
 "minimize": "cores",
 "precision": 2
 },
 {
 "maximize": "freedisk",
 "precision": 100
 },
 {
 "minimize": "sysLoadAvg",
 "precision": 10
 }
]
}
```

We can remove all cluster preferences by setting preferences to an empty list.

```
{
 "set-cluster-preferences": []
}
```

## Create and Modify Cluster Policies

Cluster policies are set using the `set-cluster-policy` command.

Like `set-cluster-preferences`, the policy definition is a JSON map defining the desired attributes and values.

Refer to the [Policy Specification](#) section for details of the allowed values for each condition in the policy.

### Input:

```
{
 "set-cluster-policy": [
 {"replica": "<2", "shard": "#EACH", "node": "#ANY"}
]
}
```

### Output:

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 47
 },
 "result": "success",
 "WARNING": "This response format is experimental. It is likely to change in the future."
}
```

We can remove all cluster policy conditions by setting policy to an empty list.

```
{
 "set-cluster-policy": []
}
```

Changing the cluster policy after the cluster is already built doesn't automatically reconfigure the cluster. However, all future cluster management operations will use the changed cluster policy.

## Create and Modify Collection-Specific Policy

The set-policy command accepts a map of policy names to the list of conditions for that policy. Multiple named policies can be specified together. A named policy that does not exist already is created and if the named policy accepts already then it is replaced.

Refer to the [Policy Specification](#) section for details of the allowed values for each condition in the policy.

### Input

```
{
 "set-policy": {
 "policy1": [
 {"replica": "1", "shard": "#EACH", "port": "8983"}
]
 }
}
```

### Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 246
 },
 "result": "success",
 "WARNING": "This response format is experimental. It is likely to change in the future."
}
```

Changing the policy after the collection is already built doesn't automatically reconfigure the collection.

However, all future cluster management operations will use the changed policy.

## Remove a Collection-Specific Policy

The remove-policy command accepts a policy name to be removed from Solr. The policy being removed must not be attached to any collection otherwise the command will fail.

### Input

```
{"remove-policy": "policy1"}
```

### Output

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 42
 },
 "result": "success",
 "WARNING": "This response format is experimental. It is likely to change in the future."
}
```

If you attempt to remove a policy that is being used by a collection, this command will fail to delete the policy until the collection itself is deleted.

## Create/Update Trigger

The set-trigger command can be used to create a new trigger or overwrite an existing one.

You can see the section [Trigger Configuration](#) for a full list of configuration options.

### *Creating a nodeAdded Trigger*

```
{
 "set-trigger": {
 "name" : "node_added_trigger",
 "event" : "nodeAdded",
 "waitFor" : "1s"
 }
}
```

### Updating Trigger with waitFor set to 5 seconds

```
{
 "set-trigger": {
 "name" : "node_added_trigger",
 "event" : "nodeAdded",
 "waitFor" : "5s",
 }
}
```

### Creating a nodeLost Trigger

```
{
 "set-trigger": {
 "name" : "node_lost_trigger1",
 "event" : "nodeLost",
 "waitFor" : "60s",
 }
}
```

## Remove Trigger

The remove-trigger command can be used to remove a trigger. It accepts a single parameter: the name of the trigger.

### Removing the nodeLost Trigger

```
{
 "remove-trigger": {
 "name" : "node_lost_trigger1"
 }
}
```

## Create/Update Trigger Listener

The set-listener command can be used to create or modify a listener for a trigger.

You can see the section [Trigger Listener Configuration](#) for a full list of configuration options.

### Creating a listener for the nodeAdded Trigger

```
{
 "set-listener": {
 "name": "foo",
 "trigger": "node_added_trigger",
 "stage": ["STARTED", "ABORTED", "SUCCEEDED", "FAILED"],
 "class": "com.example.Listener"
 }
}
```

## Remove Trigger Listener

The remove-listener command can be used to remove an existing listener. It accepts a single parameter: the name of the listener.

### *Removing the foo listener*

```
{
 "remove-listener": {
 "name": "foo"
 }
}
```

## Change Autoscaling Properties

The set-properties command can be used to change the default properties used by the Autoscaling framework.

The following properties can be specified in the payload:

### triggerScheduleDelaySeconds

This is the delay in seconds between two executions of a trigger. Every trigger is scheduled using Java's ScheduledThreadPoolExecutor with this delay. The default is 1 second.

### triggerCooldownPeriodSeconds

Solr pauses all other triggers for this cool down period after a trigger fires so that the system can stabilize before running triggers again. The default is 5 seconds.

### triggerCorePoolSize

The core pool size of the ScheduledThreadPoolExecutor used to schedule triggers. The default is 4 threads.

The command allows setting arbitrary properties in addition to the above properties. Such arbitrary properties can be useful in custom TriggerAction instances.

### *Change default triggerScheduleDelaySeconds*

```
{
 "set-properties": {
 "triggerScheduleDelaySeconds": 8
 }
}
```

The set-properties command replaces older values if present. So using set-properties to set the same value twice will overwrite the old value. If a property is not specified then it retains the last set value or the default, if no change was made. A changed value can be unset by using a null value.

Revert changed value of triggerScheduleDelaySeconds to default

```
{
 "set-properties": {
 "triggerScheduleDelaySeconds": null
 }
}
```

The changed values of these properties, if any, can be read using the Autoscaling [Read API](#) in the properties section.

## Migrating Rule-Based Replica Rules to Autoscaling Policies

Creating rules for replica placement in a Solr cluster is now done with the [autoscaling framework](#).

This document outlines how to migrate from the legacy [rule-based replica placement](#) to an [autoscaling policy](#).

The autoscaling framework is designed to fully automate your cluster management. However, if you do not want actions taken on your cluster in an automatic way, you can still use the framework to set rules and preferences. With a set of rules and preferences in place, instead of taking action directly the system will suggest actions you can take manually.

The section [Cluster Preferences Specification](#) describes the capabilities of an autoscaling policy in detail. Below we'll walk through a few examples to show how you would express the your legacy rules in the autoscaling syntax. Every rule in the legacy rule-based replica framework can be expressed in the new syntax.

### How Rules are Defined

One key difference between the frameworks is the way rules are defined.

With the rule-based replica placement framework, rules are defined with the Collections API at the time of collection creation.

The autoscaling framework, however, has its own [API](#). Policies can be configured for the entire cluster or for individual collections depending on your needs.

The following is the legacy syntax for a rule that limits the cluster to one replica for each shard in any Solr node:

```
replica:<2,node:*,shard:**
```

The equivalent rule in the autoscaling policy is:

```
{"replica": "<2", "node": "#ANY", "shard": "#EACH"}
```

## Differences in Rule Syntaxes

Many elements of defining rules are similar in both frameworks, but some elements are different.

### Rule Operators

All of the following operators can be directly used in the new policy syntax and they mean the same in both frameworks.

- **equals (no operator required)**: tag:x means the value for a tag must be equal to 'x'.
- **greater than (>)**: tag:>x means the tag value must be greater than 'x'. In this case, 'x' must be a number.
- **less than (<)**: tag:<x means the tag value must be less than 'x'. In this case also, 'x' must be a number.
- **not equal (!)**: tag: !x means tag value MUST NOT be equal to 'x'. The equals check is performed on a String value.

### Fuzzy Operator (~)

There is no ~ operator in the autoscaling policy syntax. Instead, it uses the strict parameter, which can be true or false. To replace the ~ operator, use the attribute "strict":false instead.

For example:

#### *Rule-based replica placement framework:*

```
replica:<2~,node:*,shard:**
```

#### *Autoscaling framework:*

```
{"replica": "<2", "node": "#ANY", "shard": "#EACH", "strict": false}
```

### Attributes

Attributes were known as "tags" in the rule-based replica placement framework. In the autoscaling framework, they are attributes that are used for node selection or to set global cluster-wide rules.

The available attributes in the autoscaling framework are similar to the tags that were available with rule-based replica placement. Attributes with the same name mean the same in both frameworks.

- **cores**: Number of cores in the node
- **freedisk**: Disk space available in the node
- **host**: host name of the node
- **port**: port of the node
- **node**: node name
- **role**: The role of the node. The only supported role is 'overseer'
- **ip\_1, ip\_2, ip\_3, ip\_4**: These are ip fragments for each node. For example, in a host with ip 192.168.1.2,

ip\_1 = 2, ip\_2 =1, ip\_3 = 168 and ` ip\_4 = 192`

- **sysprop.{PROPERTY\_NAME}**: These are values available from system properties. sysprop.key means a value that is passed to the node as -Dkey=keyValue during the node startup. It is possible to use rules like sysprop.key:expectedVal, shard:\*

## Snitches

There is no equivalent for a snitch in the autoscaling policy framework.

## Porting Existing Replica Placement Rules

There is no automatic way to move from using rule-based replica placement rules to an autoscaling policy. Instead you will need to remove your replica rules from each collection and institute a policy using the [autoscaling API](#).

The following examples are intended to help you translate your existing rules into new rules that fit the autoscaling framework.

### Keep less than 2 replicas (at most 1 replica) of this collection on any node

For this rule, we define the replica condition with operators for "less than 2", and use a pre-defined tag named node to define nodes with any name.

#### *Rule-based replica placement framework:*

```
replica:<2, node:*
```

#### *Autoscaling framework:*

```
{"replica": "<2", "node": "#ANY"}
```

### For a given shard, keep less than 2 replicas on any node

For this rule, we use the shard condition to define any shard, the replica condition with operators for "less than 2", and finally a pre-defined tag named node to define nodes with any name.

#### *Rule-based replica placement framework:*

```
shard:*, replica:<2, node:*
```

#### *Autoscaling framework:*

```
{"replica": "<2", "shard": "#EACH", "node": "#ANY"}
```

### Assign all replicas in shard1 to rack 730

This rule limits the shard condition to 'shard1', but any number of replicas. We're also referencing a custom tag named rack.

*Rule-based replica placement framework:*

```
shard:shard1,replica:*,rack:730
```

*Autoscaling framework:*

```
{"replica": "#ALL", "shard": "shard1", "sysprop.rack": "730"}
```

In the rule-based replica placement framework, we needed to configure a custom Snitch which provides values for the tag rack.

With the autoscaling framework, however, we need to start all nodes with a system property to define the rack values. For example, bin/solr start -c -Drack=<rack-number>.

**Create replicas in nodes with less than 5 cores only**

This rule uses the replica condition to define any number of replicas, but adds a pre-defined tag named core and uses operators for "less than 5".

*Rule-based replica placement framework:*

```
cores:<5
```

*Autoscaling framework:*

```
{"cores": "<5", "node": "#ANY"}
```

**Do not create any replicas in host 192.45.67.3***legacy syntax:*

```
host:!192.45.67.3
```

*autoscaling framework:*

```
{"replica": 0, "host": "192.45.67.3"}
```

# Colocating Collections

Solr provides a way to colocate a collection with another so that cross-collection joins are always possible.

The colocation guarantee applies to all future Collection operations made either via Collections API or by Autoscaling actions.

A collection may only be colocated with exactly one `withCollection`. However, arbitrarily many collections may be *linked* to the same `withCollection`.

## Create a Colocated Collection

The Create Collection API supports a parameter named `withCollection` which can be used to specify a collection with which the replicas of the newly created collection should be colocated. See [Create Collection API](#).

```
/admin/collections?action=CREATE&name=techproducts&numShards=1&replicationFactor=2&withCollection=tech_categories
```

In the above example, all replicas of the `techproducts` collection will be colocated on a node with at least one replica of the `tech_categories` collection.

## Colocating Existing Collections

When collections already exist beforehand, the [Modify Collection API](#) can be used to set the `withCollection` parameter so that the two collections can be linked. This will **not** trigger changes to the cluster automatically because moving a large number of replicas immediately might de-stabilize the system. Instead, it is recommended that the Suggestions UI page should be consulted on the operations that can be performed to change the cluster manually.

Example:

```
/admin/collections?action=MODIFYCOLLECTION&collection=techproducts&withCollection=tech_categories
```

## Deleting Colocated Collections

Deleting a collection which has been linked to another will fail unless the link itself is deleted first by using the [Modify Collection API](#) to un-set the `withCollection` attribute.

Example: `/admin/collections?action=MODIFYCOLLECTION&collection=techproducts&withCollection=`

## Limitations and Caveats

The collection being used as the `withCollection` must have one shard only and that shard should be named `shard1`. Note that when using the default router, the shard name is always set to `shard1` but special care must be taken to name the shard as `shard1` when using the implicit router.

In case new replicas of the `withCollection` have to be added to maintain the colocation guarantees then the new replicas will be of type NRT only. Automatically creating replicas of TLOG or PULL types is not supported.

In case, replicas have to be moved from one node to another, perhaps in response to a node lost trigger, then the target nodes will be chosen by preferring nodes that already have a replica of the withCollection so that the number of moves is minimized. However, this also means that unless there are Autoscaling policy violations, Solr will continue to move such replicas to already loaded nodes instead of preferring empty nodes. Therefore, it is advised to have policy rules which can prevent such overloading by e.g., setting the maximum number of cores per node to a fixed value.

Example: `{'cores' : '<8', 'node' : '#ANY'}`

The colocation guarantee is one-way only i.e., a collection 'X' colocated with 'Y' will always have one or more replicas of 'Y' on any node that has a replica of 'X' but the reverse is not true. There may be nodes which have one or more replicas of 'Y' but no replicas of 'X'. Such replicas of 'Y' will not be considered a violation of colocation rules and will not be cleaned up automatically.

# Legacy Scaling and Distribution

This section describes how to set up distribution and replication in Solr. It is considered "legacy" behavior, since while it is still supported in Solr, the SolrCloud functionality described in the previous chapter is where the current development is headed. However, if you don't need all that SolrCloud delivers, search distribution and index replication may be sufficient.

This section covers the following topics:

[Introduction to Scaling and Distribution](#): Conceptual information about distribution and replication in Solr.

[Distributed Search with Index Sharding](#): Detailed information about implementing distributed searching in Solr.

[Index Replication](#): Detailed information about replicating your Solr indexes.

[Combining Distribution and Replication](#): Detailed information about replicating shards in a distributed index.

[Merging Indexes](#): Information about combining separate indexes in Solr.

# Introduction to Scaling and Distribution

Both Lucene and Solr were designed to scale to support large implementations with minimal custom coding.

This section covers:

- [distributing](#) an index across multiple servers
- [replicating](#) an index on multiple servers
- [merging indexes](#)

If you need full scale distribution of indexes and queries, as well as replication, load balancing and failover, you may want to use SolrCloud. Full details on configuring and using SolrCloud is available in the section [SolrCloud](#).

## What Problem Does Distribution Solve?

If searches are taking too long or the index is approaching the physical limitations of its machine, you should consider distributing the index across two or more Solr servers.

To distribute an index, you divide the index into partitions called shards, each of which runs on a separate machine. Solr then partitions searches into sub-searches, which run on the individual shards, reporting results collectively.

The architectural details underlying index sharding are invisible to end users, who simply experience faster performance on queries against very large indexes.

## What Problem Does Replication Solve?

Replicating an index is useful when:

- You have a large search volume which one machine cannot handle, so you need to distribute searches across multiple read-only copies of the index.
- There is a high volume/high rate of indexing which consumes machine resources and reduces search performance on the indexing machine, so you need to separate indexing and searching.
- You want to make a backup of the index (see [Making and Restoring Backups](#)).

# Distributed Search with Index Sharding

When using traditional index sharding, you will need to consider how to query your documents.

It is highly recommended that you use [SolrCloud](#) when needing to scale up or scale out. The setup described below is legacy and was used prior to the existence of SolrCloud. SolrCloud provides for a truly distributed set of features with support for things like automatic routing, leader election, optimistic concurrency and other sanity checks that are expected out of a distributed system.

Everything on this page is specific to legacy setup of distributed search. Users trying out SolrCloud should not follow any of the steps or information below.

Update reorders (i.e., replica A may see update X then Y, and replica B may see update Y then X).

**deleteByQuery** also handles reorders the same way, to ensure replicas are consistent. All replicas of a shard are consistent, even if the updates arrive in a different order on different replicas.

## Distributing Documents across Shards

When not using SolrCloud, it is up to you to get all your documents indexed on each shard of your server farm. Solr supports distributed indexing (routing) in its true form only in the SolrCloud mode.

In the legacy distributed mode, Solr does not calculate universal term/doc frequencies. For most large-scale implementations, it is not likely to matter that Solr calculates TF/IDF at the shard level. However, if your collection is heavily skewed in its distribution across servers, you may find misleading relevancy results in your searches. In general, it is probably best to randomly distribute documents to your shards.

## Executing Distributed Searches with the shards Parameter

If a query request includes the shards parameter, the Solr server distributes the request across all the shards listed as arguments to the parameter. The shards parameter uses this syntax:

host:port/base\_url,host:port/base\_url\*

For example, the shards parameter below causes the search to be distributed across two Solr servers: **solr1** and **solr2**, both of which are running on port 8983:

`http://localhost:8983/solr/core1/select?shards=solr1:8983/solr/core1,solr2:8983/solr/core1&indent=true&q=ipod+solr`

Rather than require users to include the shards parameter explicitly, it is usually preferred to configure this parameter as a default in the RequestHandler section of solrconfig.xml.



Do not add the shards parameter to the standard request handler; doing so may cause search queries to enter an infinite loop. Instead, define a new request handler that uses the shards parameter, and pass distributed search requests to that handler.

With Legacy mode, only query requests are distributed. This includes requests to the SearchHandler (or any handler extending from org.apache.solr.handler.component.SearchHandler) using standard components that support distributed search.

As in SolrCloud mode, when `shards.info=true`, distributed responses will include information about the shard (where each shard represents a logically different index or physical location)

The following components support distributed search:

- The **Query** component, which returns documents matching a query
- The **Facet** component, which processes `facet.query` and `facet.field` requests where facets are sorted by count (the default).
- The **Highlighting** component, which enables Solr to include "highlighted" matches in field values.
- The **Stats** component, which returns simple statistics for numeric fields within the DocSet.
- The **Debug** component, which helps with debugging.

## Shards Whitelist

The nodes allowed in the `shards` parameter is configurable through the `shardsWhitelist` property in `solr.xml`. This whitelist is automatically configured for SolrCloud but needs explicit configuration for master/slave mode. Read more details in the section [Configuring the ShardHandlerFactory](#).

## Limitations to Distributed Search

Distributed searching in Solr has the following limitations:

- Each document indexed must have a unique key.
- If Solr discovers duplicate document IDs, Solr selects the first document and discards subsequent documents.
- The index for distributed searching may become momentarily out of sync if a commit happens between the first and second phase of the distributed search. This might cause a situation where a document that once matched a query and was subsequently changed may no longer match the query but will still be retrieved. This situation is expected to be quite rare, however, and is only possible for a single query request.
- The number of shards is limited by number of characters allowed for GET method's URI; most Web servers generally support at least 4000 characters, but many servers limit URI length to reduce their vulnerability to Denial of Service (DoS) attacks.
- Shard information can be returned with each document in a distributed search by including `f1=id, [shard]` in the search request. This returns the shard URL.
- In a distributed search, the data directory from the core descriptor overrides any data directory in `solrconfig.xml`.
- Update commands may be sent to any server with distributed indexing configured correctly. Document adds and deletes are forwarded to the appropriate server/shard based on a hash of the unique document id. **commit** commands and **deleteByQuery** commands are sent to every server in shards.

Formerly a limitation was that TF/IDF relevancy computations only used shard-local statistics. This is still the case by default. If your data isn't randomly distributed, or if you want more exact statistics, then remember to configure the `ExactStatsCache`.

## Avoiding Distributed Deadlock with Distributed Search

Like in SolrCloud mode, inter-shard requests could lead to a distributed deadlock. It can be avoided by following the instructions in the section [Distributed Requests](#).

## Testing Index Sharding on Two Local Servers

For simple functional testing, it's easiest to just set up two local Solr servers on different ports. (In a production environment, of course, these servers would be deployed on separate machines.)

1. Make two Solr home directories and copy solr.xml into the new directories:

```
mkdir example/nodes
mkdir example/nodes/node1
Copy solr.xml into this solr.home
cp server/solr/solr.xml example/nodes/node1/.
Repeat the above steps for the second node
mkdir example/nodes/node2
cp server/solr/solr.xml example/nodes/node2/.
```

2. Start the two Solr instances

```
Start first node on port 8983
bin/solr start -s example/nodes/node1 -p 8983

Start second node on port 8984
bin/solr start -s example/nodes/node2 -p 8984
```

3. Create a core on both the nodes with the sample\_techproducts\_configs.

```
bin/solr create_core -c core1 -p 8983 -d sample_techproducts_configs
Create a core on the Solr node running on port 8984
bin/solr create_core -c core1 -p 8984 -d sample_techproducts_configs
```

4. In a third window, index an example document to each of the server:

```
bin/post -c core1 example/exapedocs/monitor.xml -port 8983

bin/post -c core1 example/exapedocs/monitor2.xml -port 8984
```

5. Search on the node on port 8983:

```
curl http://localhost:8983/solr/core1/select?q=*&wt=xml&indent=true
```

This should bring back one document.

Search on the node on port 8984:

```
curl http://localhost:8984/solr/core1/select?q=*:*&wt=xml&indent=true
```

This should also bring back a single document.

Now do a distributed search across both servers with your browser or curl. In the example below, an extra parameter 'fl' is passed to restrict the returned fields to id and name.

```
curl
http://localhost:8983/solr/core1/select?q=*:*&indent=true&shards=localhost:8983/solr/core1,loc
alhost:8984/solr/core1&fl=id,name&wt=xml
```

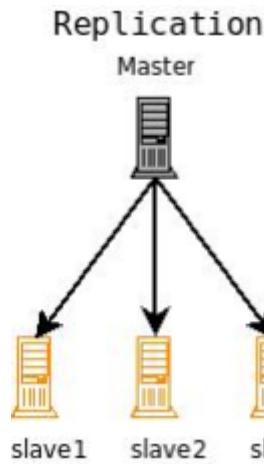
This should contain both the documents as shown below:

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">8</int>
 <lst name="params">
 <str name="q">*:*</str>
 <str name="shards">localhost:8983/solr/core1,localhost:8984/solr/core1</str>
 <str name="indent">true</str>
 <str name="fl">id,name</str>
 <str name="wt">xml</str>
 </lst>
 </lst>
 <result name="response" numFound="2" start="0" maxScore="1.0">
 <doc>
 <str name="id">3007WFP</str>
 <str name="name">Dell Widescreen UltraSharp 3007WFP</str>
 </doc>
 <doc>
 <str name="id">VA902B</str>
 <str name="name">ViewSonic VA902B - flat panel display - TFT - 19"</str>
 </doc>
 </result>
</response>
```

# Index Replication

Index Replication distributes complete copies of a master index to one or more slave servers. The master server continues to manage updates to the index. All querying is handled by the slaves. This division of labor enables Solr to scale to provide adequate responsiveness to queries against large search volumes.

The figure below shows a Solr configuration using index replication. The master server's index is replicated on the slaves.



A Solr index can be replicated across multiple slave servers, which then process requests.

## Index Replication in Solr

Solr includes a Java implementation of index replication that works over HTTP:

- The configuration affecting replication is controlled by a single file, `solrconfig.xml`
- Supports the replication of configuration files as well as index files
- Works across platforms with same configuration
- No reliance on OS-dependent file system features (e.g., hard links)
- Tightly integrated with Solr; an admin page offers fine-grained control of each aspect of replication
- The Java-based replication feature is implemented as a request handler. Configuring replication is therefore similar to any normal request handler.

### Replication In SolrCloud

Although there is no explicit concept of "master/slave" nodes in a [SolrCloud](#) cluster, the `ReplicationHandler` discussed on this page is still used by SolrCloud as needed to support "shard recovery" – but this is done in a peer to peer manner.



When using SolrCloud, the `ReplicationHandler` must be available via the `/replication` path. Solr does this implicitly unless overridden explicitly in your `solrconfig.xml`, but if you wish to override the default behavior, make certain that you do not explicitly set any of the "master" or "slave" configuration options mentioned below, or they will interfere with normal SolrCloud operation.

# Replication Terminology

The table below defines the key terms associated with Solr replication.

## Index

A Lucene index is a directory of files. These files make up the searchable and returnable data of a Solr Core.

## Distribution

The copying of an index from the master server to all slaves. The distribution process takes advantage of Lucene's index file structure.

## Inserts and Deletes

As inserts and deletes occur in the index, the directory remains unchanged. Documents are always inserted into newly created segment files. Documents that are deleted are not removed from the segment files. They are flagged in the file, deletable, and are not removed from the segments until the segment is merged as part of normal index updates.

## Master and Slave

A Solr replication master is a single node which receives all updates initially and keeps everything organized. Solr replication slave nodes receive no updates directly, instead all changes (such as inserts, updates, deletes, etc.) are made against the single master node. Changes made on the master are distributed to all the slave nodes which service all query requests from the clients.

## Update

An update is a single change request against a single Solr instance. It may be a request to delete a document, add a new document, change a document, delete all documents matching a query, etc. Updates are handled synchronously within an individual Solr instance.

## Optimization

A process that compacts the index and merges segments in order to improve query performance. Optimization should only be run on the master nodes. An optimized index may give query performance gains compared to an index that has become fragmented over a period of time with many updates. Distributing an optimized index requires a much longer time than the distribution of new segments to an un-optimized index.



optimizing is not recommended unless it can be performed regularly as it may lead to a significantly larger portion of the index consisting of deleted documents than would normally be the case.

## Segments

A self contained subset of an index consisting of some documents and data structures related to the inverted index of terms in those documents.

## mergeFactor

A parameter that controls the number of segments in an index. For example, when mergeFactor is set to 3, Solr will fill one segment with documents until the limit maxBufferedDocs is met, then it will start a new segment. When the number of segments specified by mergeFactor is reached (in this example, 3) then

Solr will merge all the segments into a single index file, then begin writing new documents to a new segment.

### Snapshot

A directory containing hard links to the data files of an index. Snapshots are distributed from the master nodes when the slaves pull them, "smart copying" any segments the slave node does not have in snapshot directory that contains the hard links to the most recent index data files.

## Configuring the ReplicationHandler

In addition to ReplicationHandler configuration options specific to the master/slave roles, there are a few special configuration options that are generally supported (even when using SolrCloud).

- `maxNumberofBackups` an integer value dictating the maximum number of backups this node will keep on disk as it receives backup commands.
- Similar to most other request handlers in Solr you may configure a set of [defaults, invariants, and/or appends](#) parameters corresponding with any request parameters supported by the ReplicationHandler when [processing commands](#).

## Configuring the Replication RequestHandler on a Master Server

Before running a replication, you should set the following parameters on initialization of the handler:

#### replicateAfter

String specifying action after which replication should occur. Valid values are commit, optimize, or startup. There can be multiple values for this parameter. If you use "startup", you need to have a "commit" and/or "optimize" entry also if you want to trigger replication on future commits or optimizes.

#### backupAfter

String specifying action after which a backup should occur. Valid values are commit, optimize, or startup. There can be multiple values for this parameter. It is not required for replication, it just makes a backup.

#### maxNumberofBackups

Integer specifying how many backups to keep. This can be used to delete all but the most recent N backups.

#### confFiles

The configuration files to replicate, separated by a comma.

#### commitReserveDuration

If your commits are very frequent and your network is slow, you can tweak this parameter to increase the amount of time expected to be required to transfer data. The default is `00:00:10` i.e., 10 seconds.

The example below shows a possible 'master' configuration for the ReplicationHandler, including a fixed number of backups and an invariant setting for the `maxWriteMBPerSec` request parameter to prevent slaves from saturating its network interface

```
<requestHandler name="/replication" class="solr.ReplicationHandler">
 <lst name="master">
 <str name="replicateAfter">optimize</str>
 <str name="backupAfter">optimize</str>
 <str name="confFiles">schema.xml,stopwords.txt,elevate.xml</str>
 </lst>
 <int name="maxNumberOfBackups">2</int>
 <str name="commitReserveDuration">00:00:10</str>
 <lst name="invariants">
 <str name="maxWriteMBPerSec">16</str>
 </lst>
</requestHandler>
```

## Replicating solrconfig.xml

In the configuration file on the master server, include a line like the following:

```
<str name="confFiles">solrconfig_slave.xml:solrconfig.xml,x.xml,y.xml</str>
```

This ensures that the local configuration `solrconfig_slave.xml` will be saved as `solrconfig.xml` on the slave. All other files will be saved with their original names.

On the master server, the file name of the slave configuration file can be anything, as long as the name is correctly identified in the `confFiles` string; then it will be saved as whatever file name appears after the colon ':'.

## Configuring the Replication RequestHandler on a Slave Server

The code below shows how to configure a ReplicationHandler on a slave.

```
<requestHandler name="/replication" class="solr.ReplicationHandler">
 <lst name="slave">

 <!-- fully qualified url for the replication handler of master. It is
 possible to pass on this as a request param for the fetchindex command -->
 <str name="masterUrl">http://remote_host:port/solr/core_name/replication</str>

 <!-- Interval in which the slave should poll master. Format is HH:mm:ss .
 If this is absent slave does not poll automatically.

 But a fetchindex can be triggered from the admin or the http API -->

 <str name="pollInterval">00:00:20</str>

 <!-- THE FOLLOWING PARAMETERS ARE USUALLY NOT REQUIRED-->

 <!-- To use compression while transferring the index files. The possible
 values are internal|external. If the value is 'external' make sure
 that your master Solr has the settings to honor the accept-encoding header.
 See here for details: http://wiki.apache.org/solr/SolrHttpCompression
 If it is 'internal' everything will be taken care of automatically.
 USE THIS ONLY IF YOUR BANDWIDTH IS LOW.
 THIS CAN ACTUALLY SLOWDOWN REPLICATION IN A LAN -->
 <str name="compression">internal</str>

 <!-- The following values are used when the slave connects to the master to
 download the index files. Default values implicitly set as 5000ms and
 10000ms respectively. The user DOES NOT need to specify these unless the
 bandwidth is extremely low or if there is an extremely high latency -->

 <str name="httpConnTimeout">5000</str>
 <str name="httpReadTimeout">10000</str>

 <!-- If HTTP Basic authentication is enabled on the master, then the slave
 can be configured with the following -->

 <str name="httpBasicAuthUser">username</str>
 <str name="httpBasicAuthPassword">password</str>
 </lst>
</requestHandler>
```

## Setting Up a Repeater with the ReplicationHandler

A master may be able to serve only so many slaves without affecting performance. Some organizations have deployed slave servers across multiple data centers. If each slave downloads the index from a remote data center, the resulting download may consume too much network bandwidth. To avoid performance degradation in cases like this, you can configure one or more slaves as repeaters. A repeater is simply a node that acts as both a master and a slave.

- To configure a server as a repeater, the definition of the Replication requestHandler in the

solrconfig.xml file must include file lists of use for both masters and slaves.

- Be sure to set the replicateAfter parameter to commit, even if replicateAfter is set to optimize on the main master. This is because on a repeater (or any slave), a commit is called only after the index is downloaded. The optimize command is never called on slaves.
- Optionally, one can configure the repeater to fetch compressed files from the master through the compression parameter to reduce the index download time.

Here is an example of a ReplicationHandler configuration for a repeater:

```
<requestHandler name="/replication" class="solr.ReplicationHandler">
 <lst name="master">
 <str name="replicateAfter">commit</str>
 <str name="confFiles">schema.xml,stopwords.txt,synonyms.txt</str>
 </lst>
 <lst name="slave">
 <str name="masterUrl">http://master.solr.company.com:8983/solr/core_name/replication</str>
 <str name="pollInterval">00:00:60</str>
 </lst>
</requestHandler>
```

## Commit and Optimize Operations

When a commit or optimize operation is performed on the master, the RequestHandler reads the list of file names which are associated with each commit point. This relies on the replicateAfter parameter in the configuration to decide which types of events should trigger replication.

These operations are supported:

- commit: Triggers replication whenever a commit is performed on the master index.
- optimize: Triggers replication whenever the master index is optimized.
- startup: Triggers replication whenever the master index starts up.

The replicateAfter parameter can accept multiple arguments. For example:

```
<str name="replicateAfter">startup</str>
<str name="replicateAfter">commit</str>
<str name="replicateAfter">optimize</str>
```

## Slave Replication

The master is totally unaware of the slaves.

The slave continuously keeps polling the master (depending on the pollInterval parameter) to check the current index version of the master. If the slave finds out that the master has a newer version of the index it initiates a replication process. The steps are as follows:

- The slave issues a filelist command to get the list of the files. This command returns the names of the

files as well as some metadata (for example, size, a lastmodified timestamp, an alias if any).

- The slave checks with its own index if it has any of those files in the local index. It then runs the filecontent command to download the missing files. This uses a custom format (akin to the HTTP chunked encoding) to download the full content or a part of each file. If the connection breaks in between, the download resumes from the point it failed. At any point, the slave tries 5 times before giving up a replication altogether.
- The files are downloaded into a temp directory, so that if either the slave or the master crashes during the download process, no files will be corrupted. Instead, the current replication will simply abort.
- After the download completes, all the new files are moved to the live index directory and the file's timestamp is same as its counterpart on the master.
- A commit command is issued on the slave by the Slave's ReplicationHandler and the new index is loaded.

## Replicating Configuration Files

To replicate configuration files, list them using the confFiles parameter. Only files found in the conf directory of the master's Solr instance will be replicated.

Solr replicates configuration files only when the index itself is replicated. That means even if a configuration file is changed on the master, that file will be replicated only after there is a new commit/optimize on master's index.

Unlike the index files, where the timestamp is good enough to figure out if they are identical, configuration files are compared against their checksum. The schema.xml files (on master and slave) are judged to be identical if their checksums are identical.

As a precaution when replicating configuration files, Solr copies configuration files to a temporary directory before moving them into their ultimate location in the conf directory. The old configuration files are then renamed and kept in the same conf/ directory. The ReplicationHandler does not automatically clean up these old files.

If a replication involved downloading of at least one configuration file, the ReplicationHandler issues a core-reload command instead of a commit command.

## Resolving Corruption Issues on Slave Servers

If documents are added to the slave, then the slave is no longer in sync with its master. However, the slave will not undertake any action to put itself in sync, until the master has new index data.

When a commit operation takes place on the master, the index version of the master becomes different from that of the slave. The slave then fetches the list of files and finds that some of the files present on the master are also present in the local index but with different sizes and timestamps. This means that the master and slave have incompatible indexes.

To correct this problem, the slave then copies all the index files from master to a new index directory and asks the core to load the fresh index from the new directory.

## HTTP API Commands for the ReplicationHandler

You can use the HTTP commands below to control the ReplicationHandler's operations.

**enablereplication**

Enable replication on the "master" for all its slaves.

```
http://_master_host:port/_solr/_core_name_/replication?command=enablereplication
```

**disablereplication**

Disable replication on the master for all its slaves.

```
http://_master_host:port/_solr/_core_name_/replication?command=disablereplication
```

**indexversion**

Return the version of the latest replicatable index on the specified master or slave.

```
http://_host:port/_solr/_core_name_/replication?command=indexversion
```

**fetchindex**

Force the specified slave to fetch a copy of the index from its master.

```
http://_slave_host:port/_solr/_core_name_/replication?command=fetchindex
```

If you like, you can pass an extra attribute such as `masterUrl` or `compression` (or any other parameter which is specified in the `<lst name="slave">` tag) to do a one time replication from a master. This obviates the need for hard-coding the master in the slave.

**abortfetch**

Abort copying an index from a master to the specified slave.

```
http://_slave_host:port/_solr/_core_name_/replication?command=abortfetch
```

**enablepoll**

Enable the specified slave to poll for changes on the master.

```
http://_slave_host:port/_solr/_core_name_/replication?command=enablepoll
```

**disablepoll**

Disable the specified slave from polling for changes on the master.

```
http://_slave_host:port/_solr/_core_name_/replication?command=disablepoll
```

**details**

Retrieve configuration details and current status.

```
http://_slave_host:port/_solr/_core_name_/replication?command=details
```

#### filelist

Retrieve a list of Lucene files present in the specified host's index.

```
http://_host:port/_solr/_core_name_/replication?command=filelist&generation=<_generation-number_>
```

You can discover the generation number of the index by running the `indexversion` command.

#### backup

Create a backup on master if there are committed index data in the server; otherwise, does nothing.

```
http://_master_host:port/_solr/_core_name_/replication?command=backup
```

This command is useful for making periodic backups. There are several supported request parameters:

- `numberToKeep`:: This can be used with the `backup` command unless the `maxNumberOfBackups` initialization parameter has been specified on the handler – in which case `maxNumberOfBackups` is always used and attempts to use the `numberToKeep` request parameter will cause an error.
- `name`: (optional) Backup name. The snapshot will be created in a directory called `snapshot.<name>` within the data directory of the core. By default the name is generated using date in `yyyyMMddHHmmssSSS` format. If `location` parameter is passed, that would be used instead of the data directory
- `location`: Backup location.

#### deletebackup

Delete any backup created using the `backup` command.

```
http://_master_host:port/_solr/_core_name_/replication?command=deletebackup
```

There are two supported parameters:

- `name`: The name of the snapshot. A snapshot with the name `snapshot.name` must exist. If not, an error is thrown.
- `location`: Location where the snapshot is created.

## Distribution and Optimization

Optimizing an index is not something most users should generally worry about - but in particular users should be aware of the impacts of optimizing an index when using the `ReplicationHandler`.

The time required to optimize a master index can vary dramatically. A small index may be optimized in minutes. A very large index may take hours. The variables include the size of the index and the speed of the hardware.

Distributing a newly optimized index may take only a few minutes or up to an hour or more, again depending on the size of the index and the performance capabilities of network connections and disks. During optimization the machine is under load and does not process queries very well. Given a schedule of updates being driven a few times an hour to the slaves, we cannot run an optimize with every committed snapshot.

Copying an optimized index means that the **entire** index will need to be transferred during the next snappull. This is a large expense, but not nearly as huge as running the optimize everywhere.

Consider this example: on a three-slave one-master configuration, distributing a newly-optimized index takes approximately 80 seconds *total*. Rolling the change across a tier would require approximately ten minutes per machine (or machine group). If this optimize were rolled across the query tier, and if each slave node being optimized were disabled and not receiving queries, a rollout would take at least twenty minutes and potentially as long as an hour and a half. Additionally, the files would need to be synchronized so that the *following* the optimize, snappull would not think that the independently optimized files were different in any way. This would also leave the door open to independent corruption of indexes instead of each being a perfect copy of the master.

Optimizing on the master allows for a straight-forward optimization operation. No query slaves need to be taken out of service. The optimized index can be distributed in the background as queries are being normally serviced. The optimization can occur at any time convenient to the application providing index updates.

While optimizing may have some benefits in some situations, a rapidly changing index will not retain those benefits for long, and since optimization is an intensive process, it may be better to consider other options, such as lowering the merge factor (discussed in the section on [Index Configuration](#)).



Do not elect to optimize your index unless you have tangible evidence that it will significantly improve your search performance. Recent changes in Solr/Lucene have dramatically lessened the need to optimize as discussed at the above link.

# Combining Distribution and Replication

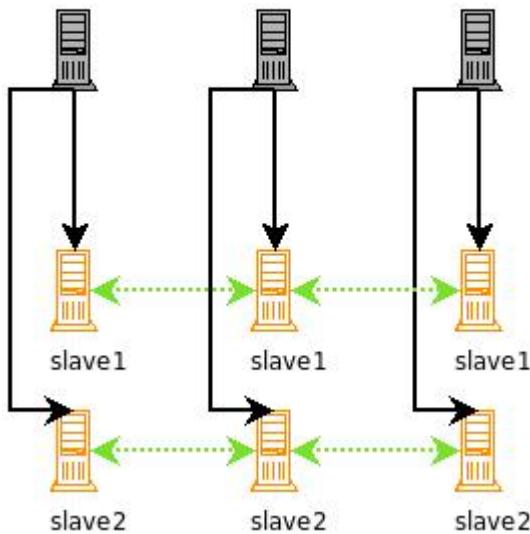
When your index is too large for a single machine and you have a query volume that single shards cannot keep up with, it's time to replicate each shard in your distributed search setup.

The idea is to combine distributed search with replication. As shown in the figure below, a combined distributed-replication configuration features a master server for each shard and then 1-n slaves that are replicated from the master. As in a standard replicated configuration, the master server handles updates and optimizations without adversely affecting query handling performance.

Query requests should be load balanced across each of the shard slaves. This gives you both increased query handling capacity and fail-over backup if a server goes down.

## Distributed + Replication

Shard 1 Master    Shard 2 Master    Shard 3 Master



A Solr configuration combining both replication and master-slave distribution.

None of the master shards in this configuration know about each other. You index to each master, the index is replicated to each slave, and then searches are distributed across the slaves, using one slave from each master/slave shard.

For high availability you can use a load balancer to set up a virtual IP for each shard's set of slaves. If you are new to load balancing, HAProxy (<http://haproxy.1wt.eu/>) is a good open source software load-balancer. If a slave server goes down, a good load-balancer will detect the failure using some technique (generally a heartbeat system), and forward all requests to the remaining live slaves that served with the failed slave. A single virtual IP should then be set up so that requests can hit a single IP, and get load balanced to each of the virtual IPs for the search slaves.

With this configuration you will have a fully load balanced, search-side fault-tolerant system (Solr does not yet support fault-tolerant indexing). Incoming searches will be handed off to one of the functioning slaves, then the slave will distribute the search request across a slave for each of the shards in your configuration. The slave will issue a request to each of the virtual IPs for each shard, and the load balancer will choose one of the available slaves. Finally, the results will be combined into a single results set and returned. If any of

the slaves go down, they will be taken out of rotation and the remaining slaves will be used. If a shard master goes down, searches can still be served from the slaves until you have corrected the problem and put the master back into production.

# Merging Indexes

If you need to combine indexes from two different projects or from multiple servers previously used in a distributed configuration, you can use either the `IndexMergeTool` included in `lucene-misc` or the `CoreAdminHandler`.

To merge indexes, they must meet these requirements:

- The two indexes must be compatible: their schemas should include the same fields and they should analyze fields the same way.
- The indexes must not include duplicate data.

Optimally, the two indexes should be built using the same schema.

## Using IndexMergeTool

To merge the indexes, do the following:

1. Make sure that both indexes you want to merge are closed.
2. Issue this command:

```
java -cp $SOLR/server/solr-webapp/webapp/WEB-INF/lib/lucene-core-
VERSION.jar:$SOLR/server/solr-webapp/webapp/WEB-INF/lib/lucene-misc-VERSION.jar
org/apache/lucene/misc/IndexMergeTool /path/to/newindex /path/to/old/index1
/path/to/old/index2
```

This will create a new index at `/path/to/newindex` that contains both `index1` and `index2`.

3. Copy this new directory to the location of your application's Solr index (move the old one aside first, of course) and start Solr.

## Using CoreAdmin

The `MERGEINDEXES` command of the `CoreAdminHandler` can be used to merge indexes into a new core – either from one or more arbitrary `indexDir` directories or by merging from one or more existing `srcCore` core names.

See the [CoreAdminHandler](#) section for details.

# The Well-Configured Solr Instance

This section tells you how to fine-tune your Solr instance for optimum performance.

This section covers the following topics:

[Configuring solrconfig.xml](#): Describes how to work with the main configuration file for Solr, `solrconfig.xml`, covering the major sections of the file.

[Sol Cores and solr.xml](#): Describes how to work with `solr.xml` and `core.properties` to configure your Solr core, or multiple Solr cores within a single instance.

[Configuration APIs](#): Describes several APIs used to configure Solr: Blob Store, Config, Request Parameters and Managed Resources.

[Implicit RequestHandlers](#): Describes various end-points automatically provided by Solr and how to configure them.

[Solr Plugins](#): Introduces Solr plugins with pointers to more information.

[JVM Settings](#): Gives some guidance on best practices for working with Java Virtual Machines.

[V2 API](#): Describes how to use the new V2 APIs, a redesigned API framework covering most Solr APIs.



The focus of this section is generally on configuring a single Solr instance, but for those interested in scaling a Solr implementation in a cluster environment, see also the section [SolrCloud](#). There are also options to scale through sharding or replication, described in the section [Legacy Scaling and Distribution](#).

# Configuring solrconfig.xml

The `solrconfig.xml` file is the configuration file with the most parameters affecting Solr itself.

While configuring Solr, you'll work with `solrconfig.xml` often, either directly or via the [Config API](#) to create "configuration overlays" (`configoverlay.json`) to override the values in `solrconfig.xml`.

In `solrconfig.xml`, you configure important features such as:

- request handlers, which process the requests to Solr, such as requests to add documents to the index or requests to return results for a query
- listeners, processes that "listen" for particular query-related events; listeners can be used to trigger the execution of special code, such as invoking some common queries to warm-up caches
- the Request Dispatcher for managing HTTP communications
- the Admin Web interface
- parameters related to replication and duplication (these parameters are covered in detail in [Legacy Scaling and Distribution](#))

The `solrconfig.xml` file is located in the `conf/` directory for each collection. Several well-commented example files can be found in the `server/solr/configsets/` directories demonstrating best practices for many different types of installations.

We've covered the options in the following sections:

- [DataDir and DirectoryFactory in SolrConfig](#)
- [Lib Directives in SolrConfig](#)
- [Schema Factory Definition in SolrConfig](#)
- [IndexConfig in SolrConfig](#)
- [RequestHandlers and SearchComponents in SolrConfig](#)
- [InitParams in SolrConfig](#)
- [UpdateHandlers in SolrConfig](#)
- [Query Settings in SolrConfig](#)
- [RequestDispatcher in SolrConfig](#)
- [Update Request Processors](#)
- [Codec Factory](#)

## Substituting Properties in Solr Config Files

Solr supports variable substitution of property values in configuration files, which allows runtime specification of various configuration options in `solrconfig.xml`. The syntax is  `${propertyname[:option default value]}` . This allows defining a default that can be overridden when Solr is launched. If a default value is not specified, then the property *must* be specified at runtime or the configuration file will generate an error when parsed.

There are multiple methods for specifying properties that can be used in configuration files. Of those below, strongly consider "config overlay" as the preferred approach, as it stays local to the configset and is easy to modify.

## JVM System Properties

Any JVM System properties, usually specified using the -D flag when starting the JVM, can be used as variables in any XML configuration file in Solr.

For example, in the sample solrconfig.xml files, you will see this value which defines the locking type to use:

```
<lockType>${solr.lock.type:native}</lockType>
```

Which means the lock type defaults to "native" but when starting Solr, you could override this using a JVM system property by launching the Solr it with:

```
bin/solr start -Dsolr.lock.type=none
```

In general, any Java system property that you want to set can be passed through the bin/solr script using the standard -Dproperty=value syntax. Alternatively, you can add common system properties to the SOLR\_OPTS environment variable defined in the Solr include file (bin/solr.in.sh or bin/solr.in.cmd). For more information about how the Solr include file works, refer to: [Taking Solr to Production](#).

## Config API to Override solrconfig.xml

The [Config API](#) allows you to use an API to modify Solr's configuration, specifically user defined properties. Changes made with this API are stored in a file named configoverlay.json. This file should only be edited with the API, but will look like this example:

```
{"userProps": {
 "dih.db.url": "jdbc:oracle:thin:@localhost:1521",
 "dih.db.user": "username",
 "dih.db.pass": "password"}}
```

For more details, see the section [Config API](#).

## solrcore.properties

If the configuration directory for a Solr core contains a file named solrcore.properties that file can contain any arbitrary user-defined property names and values using the Java [properties file format](#). Those properties can then be used as variables in other configuration files for that Solr core.

For example, the following solrcore.properties file could be created in the conf/ directory of a collection using one of the example configurations, to override the lockType used.

```
#conf/solrcore.properties
solr.lock.type=none
```

#### *Deprecation*



`solrcore.properties` won't work in SolrCloud mode (it is not read from ZooKeeper). This feature is likely to be removed in the future. Instead, use another mechanism like a config overlay.



The path and name of the `solrcore.properties` file can be overridden using the `properties` property in `core.properties`.

## User-Defined Properties in `core.properties`

Every Solr core has a `core.properties` file, automatically created when using the APIs. When you create a SolrCloud collection, you can pass through custom parameters by prefixing the parameter name with `property.name` as a parameter.

For example, to add a property named "my.custom.prop":

### V1 API

```
http://localhost:8983/solr/admin/collections?action=CREATE&name=gettingstarted&numShards=1&property.my.custom.prop=edismax
```

### V2 API

```
curl -X POST -H 'Content-type: application/json' -d '{"create": {"name": "gettingstarted", "numShards": "1", "property.my.custom.prop": "edismax"}}'
http://localhost:8983/api/collections
```

This will create a `core.properties` file that has at least the following properties (others omitted for brevity):

```
#core.properties
name=gettingstarted
my.custom.prop=edismax
```

The `my.custom.prop` property can then be used as a variable, such as in `solrconfig.xml`:

```
<requestHandler name="/select">
 <lst name="defaults">
 <str name="defType">${my.custom.prop}</str>
 </lst>
</requestHandler>
```

## Implicit Core Properties

Several attributes of a Solr core are available as "implicit" properties that can be used in variable substitution, independent of where or how the underlying value is initialized.

For example, regardless of whether the name for a particular Solr core is explicitly configured in core.properties or inferred from the name of the instance directory, the implicit property solr.core.name is available for use as a variable in that core's configuration file:

```
<requestHandler name="/select">
 <lst name="defaults">
 <str name="collection_name">${solr.core.name}</str>
 </lst>
</requestHandler>
```

All implicit properties use the solr.core. name prefix, and reflect the runtime value of the equivalent core.properties property:

- solr.core.name
- solr.core.config
- solr.core.schema
- solr.core.dataDir
- solr.core.transient
- solr.core.loadOnStartup

## DataDir and DirectoryFactory in SolrConfig

Where and how Solr stores its indexes are configurable options.

### Specifying a Location for Index Data with the dataDir Parameter

By default, Solr stores its index data in a directory called /data under the core's instance directory (instanceDir). If you would like to specify a different directory for storing index data, you can configure the dataDir in the core.properties file for the core, or use the <dataDir> parameter in the solrconfig.xml file. You can specify another directory either with an absolute path or a pathname relative to the instanceDir of the SolrCore. For example:

```
<dataDir>/solr/data/${solr.core.name}</dataDir>
```

The \${solr.core.name} substitution will cause the name of the current core to be substituted, which results

in each core's data being kept in a separate subdirectory.

If you are using replication to replicate the Solr index (as described in [Legacy Scaling and Distribution](#)), then the <dataDir> directory should correspond to the index directory used in the replication configuration.



If the environment variable SOLR\_DATA\_HOME is defined, or if solr.data.home is configured for your DirectoryFactory, or if solr.xml contains an element <solrDataHome> then the location of data directory will be <SOLR\_DATA\_HOME>/<instance\_name>/data.

## Specifying the DirectoryFactory For Your Index

The default solr.NRTCachingDirectoryFactory is filesystem based, and tries to pick the best implementation for the current JVM and platform. You can force a particular implementation and/or configuration options by specifying solr.MMapDirectoryFactory, solr.NIOFSDirectoryFactory, or solr.SimpleFSDirectoryFactory.

```
<directoryFactory name="DirectoryFactory"
 class="solr.MMapDirectoryFactory">
 <bool name="preload">true</bool>
</directoryFactory>
```

The solr.RAMDirectoryFactory is memory based, not persistent, and does not work with replication. Use this DirectoryFactory to store your index in RAM.

```
<directoryFactory class="org.apache.solr.core.RAMDirectoryFactory" />
```



If you are using Hadoop and would like to store your indexes in HDFS, you should use the solr.HdfsDirectoryFactory instead of either of the above implementations. For more details, see the section [Running Solr on HDFS](#).

## Resource and Plugin Loading

Solr components can be configured using **resources**: data stored in external files that may be referred to in a location-independent fashion. Examples include: files needed by schema components, e.g., a stopword list for [Stop Filter](#); and machine-learned models for [Learning to Rank](#).

Solr **plugins**, which can be configured in solrconfig.xml, are Java classes that are normally packaged in .jar files along with supporting classes and data. Solr ships with a number of built-in plugins, and can also be configured to use custom plugins. Example plugins are the [Data Import Handler](#) and custom search components.

Resources and plugins may be stored:

- in ZooKeeper under a collection's configset node (SolrCloud only);
- on a filesystem accessible to Solr nodes; or
- in Solr's [Blob Store](#) (SolrCloud only).



Schema components may not be stored as plugins in the Blob Store, and cannot access resources stored in the Blob Store.

## Resource and Plugin Loading Sequence

Under SolrCloud, resources and plugins to be loaded are first looked up in ZooKeeper under the collection's configset znode. If the resource or plugin is not found there, Solr will fall back to loading [from the filesystem](#).

Note that by default, Solr will not attempt to load resources and plugins from the Blob Store. To enable this, see the section [Use a Blob in a Handler or Component](#). When loading from the Blob Store is enabled for a component, lookups occur only in the Blob Store, and never in ZooKeeper or on the filesystem.

## Resources and Plugins in ConfigSets on ZooKeeper

Resources and plugins may be uploaded to ZooKeeper as part of a configset, either via the [Configsets API](#) or `bin/solr zk upload`.

To upload a plugin or resource to a configset already stored on ZooKeeper, you can use `bin/solr zk cp`.



By default, ZooKeeper's file size limit is 1MB. If your files are larger than this, you'll need to either [increase the ZooKeeper file size limit](#) or store them instead [on the filesystem](#).

## Resources and Plugins on the Filesystem

Under standalone Solr, when looking up a plugin or resource to be loaded, Solr's resource loader will first look under the `<instanceDir>/conf/` directory. If the plugin or resource is not found, the configured plugin and resource file paths are searched - see the section [Lib Directives in SolrConfig](#) below.

On core load, Solr's resource loader constructs a list of paths (subdirectories and jars), first under `solr_home/lib`, and then under directories pointed to by `<lib/>` [directives in SolrConfig](#).

When looking up a resource or plugin to be loaded, the paths on the list are searched in the order they were added.



Under SolrCloud, each node hosting a collection replica will need its own copy of plugins and resources to be loaded.

To get Solr's resource loader to find resources either under subdirectories or in jar files that were created after Solr's resource path list was constructed, reload the collection (SolrCloud) or the core (standalone Solr). Restarting all affected Solr nodes also works.



Resource files **will not be loaded** if they are located directly under either `solr_home/lib` or a directory given by the `dir` attribute on a `<lib/>` directive in SolrConfig. Resources are only searched for under subdirectories or in jar files found in those locations.

### `solr_home/lib`

Each Solr node can have a directory named `lib/` under the [Solr home directory](#). In order to use this directory to host resources or plugins, it must first be manually created.

## Lib Directives in SolrConfig

Plugin and resource file paths are configurable via <lib/> directives in solrconfig.xml.

Loading occurs in the order <lib/> directives appear in solrconfig.xml. If there are dependencies, list the lowest level dependency jar first.

A regular expression supplied in the <lib/> element's regex attribute value can be used to restrict which subdirectories and/or jar files are added to the Solr resource loader's list of search locations. If no regular expression is given, all direct subdirectory and jar children are included in the resource path list. All directories are resolved as relative to the Solr core's instanceDir.

From an example SolrConfig:

```
<lib dir="../../../contrib/extraction/lib" regex=".*\.\.jar" />
<lib dir="../../../dist/" regex="solr-cell-\d.*\.\.jar" />

<lib dir="../../../contrib/clustering/lib/" regex=".*\.\.jar" />
<lib dir="../../../dist/" regex="solr-clustering-\d.*\.\.jar" />

<lib dir="../../../contrib/langid/lib/" regex=".*\.\.jar" />
<lib dir="../../../dist/" regex="solr-langid-\d.*\.\.jar" />

<lib dir="../../../contrib/velocity/lib" regex=".*\.\.jar" />
<lib dir="../../../dist/" regex="solr-velocity-\d.*\.\.jar" />
```

## Schema Factory Definition in SolrConfig

Solr's [Schema API](#) enables remote clients to access [schema](#) information, and make schema modifications, through a REST interface.

Other features such as Solr's [Schemaless Mode](#) also work via schema modifications made programmatically at run time.

Using the Managed Schema is required to be able to use the Schema API to modify your schema. However, using Managed Schema does not by itself mean you are also using Solr in Schemaless Mode (or "schema guessing" mode).

 Schemaless mode requires enabling the Managed Schema if it is not already, but full schema guessing requires additional configuration as described in the section [Schemaless Mode](#).

While the "read" features of the Schema API are supported for all schema types, support for making schema modifications programmatically depends on the <schemaFactory/> in use.

## Solr Uses Managed Schema by Default

When a <schemaFactory/> is not explicitly declared in a solrconfig.xml file, Solr implicitly uses a ManagedIndexSchemaFactory, which is by default "mutable" and keeps schema information in a managed-schema file.

```
<!-- An example of Solr's implicit default behavior if no
 no schemaFactory is explicitly defined.
-->
<schemaFactory class="ManagedIndexSchemaFactory">
 <bool name="mutable">true</bool>
 <str name="managedSchemaResourceName">managed-schema</str>
</schemaFactory>
```

If you wish to explicitly configure ManagedIndexSchemaFactory the following options are available:

- `mutable` - controls whether changes may be made to the Schema data. This must be set to `true` to allow edits to be made with the Schema API.
- `managedSchemaResourceName` is an optional parameter that defaults to "managed-schema", and defines a new name for the schema file that can be anything other than "schema.xml".

With the default configuration shown above, you can use the [Schema API](#) to modify the schema as much as you want, and then later change the value of `mutable` to `false` if you wish to "lock" the schema in place and prevent future changes.

## Classic schema.xml

An alternative to using a managed schema is to explicitly configure a ClassicIndexSchemaFactory. ClassicIndexSchemaFactory requires the use of a `schema.xml` configuration file, and disallows any programmatic changes to the Schema at run time. The `schema.xml` file must be edited manually and is only loaded only when the collection is loaded.

```
<schemaFactory class="ClassicIndexSchemaFactory"/>
```

## Switching from schema.xml to Managed Schema

If you have an existing Solr collection that uses ClassicIndexSchemaFactory, and you wish to convert to use a managed schema, you can simply modify the `solrconfig.xml` to specify the use of the ManagedIndexSchemaFactory.

Once Solr is restarted and it detects that a `schema.xml` file exists, but the `managedSchemaResourceName` file (i.e., "managed-schema") does not exist, the existing `schema.xml` file will be renamed to `schema.xml.bak` and the contents are re-written to the managed schema file. If you look at the resulting file, you'll see this at the top of the page:

```
<!-- Solr managed schema - automatically generated - DO NOT EDIT -->
```

You are now free to use the [Schema API](#) as much as you want to make changes, and remove the `schema.xml.bak`.

## Switching from Managed Schema to Manually Edited schema.xml

If you have started Solr with managed schema enabled and you would like to switch to manually editing a `schema.xml` file, you should take the following steps:

1. Rename the managed-schema file to schema.xml.
2. Modify solrconfig.xml to replace the schemaFactory class.
  - a. Remove any ManagedIndexSchemaFactory definition if it exists.
  - b. Add a ClassicIndexSchemaFactory definition as shown above
3. Reload the core(s).

If you are using SolrCloud, you may need to modify the files via ZooKeeper. The bin/solr script provides an easy way to download the files from ZooKeeper and upload them back after edits. See the section [ZooKeeper Operations](#) for more information.



To have full control over your schema.xml file, you may also want to disable schema guessing, which allows unknown fields to be added to the schema during indexing. The properties that enable this feature are discussed in the section [Schemaless Mode](#).

## IndexConfig in SolrConfig

The <indexConfig> section of solrconfig.xml defines low-level behavior of the Lucene index writers.

By default, the settings are commented out in the sample solrconfig.xml included with Solr, which means the defaults are used. In most cases, the defaults are fine.

```
<indexConfig>
 ...
</indexConfig>
```

## Writing New Segments

### ramBufferSizeMB

Once accumulated document updates exceed this much memory space (defined in megabytes), then the pending updates are flushed. This can also create new segments or trigger a merge. Using this setting is generally preferable to maxBufferedDocs. If both maxBufferedDocs and ramBufferSizeMB are set in solrconfig.xml, then a flush will occur when either limit is reached. The default is 100Mb.

```
<ramBufferSizeMB>100</ramBufferSizeMB>
```

### maxBufferedDocs

Sets the number of document updates to buffer in memory before they are flushed as a new segment. This may also trigger a merge. The default Solr configuration sets to flush by RAM usage (ramBufferSizeMB).

```
<maxBufferedDocs>1000</maxBufferedDocs>
```

### useCompoundFile

Controls whether newly written (and not yet merged) index segments should use the [Compound File](#)

[Segments](#) format. The default is false.

```
<useCompoundFile>false</useCompoundFile>
```

## Merging Index Segments

### mergePolicyFactory

Defines how merging segments is done.

The default in Solr is to use a `TieredMergePolicy`, which merges segments of approximately equal size, subject to an allowed number of segments per tier.

Other policies available are the `LogByteSizeMergePolicy`, `LogDocMergePolicy`, and `UninvertDocValuesMergePolicy`. For more information on these policies, please see [the MergePolicy javadocs](#).

```
<mergePolicyFactory class="org.apache.solr.index.TieredMergePolicyFactory">
 <int name="maxMergeAtOnce">10</int>
 <int name="segmentsPerTier">10</int>
</mergePolicyFactory>
```

### Controlling Segment Sizes

The most common adjustment users make to the configuration of `TieredMergePolicy` (or `LogByteSizeMergePolicy`) are the "merge factors" to change how many segments should be merged at one time and, in the `TieredMergePolicy` case, the maximum size of an merged segment.

For `TieredMergePolicy`, this is controlled by setting the `maxMergeAtOnce` (default 10), `segmentsPerTier` (default 10) and `maxMergedSegmentMB` (default 5000) options.

`LogByteSizeMergePolicy` has a single `mergeFactor` option (default 10).

To understand why these options are important, consider what happens when an update is made to an index using `LogByteSizeMergePolicy`: Documents are always added to the most recently opened segment. When a segment fills up, a new segment is created and subsequent updates are placed there.

If creating a new segment would cause the number of lowest-level segments to exceed the `mergeFactor` value, then all those segments are merged together to form a single large segment. Thus, if the merge factor is 10, each merge results in the creation of a single segment that is roughly ten times larger than each of its ten constituents. When there are 10 of these larger segments, then they in turn are merged into an even larger single segment. This process can continue indefinitely.

When using `TieredMergePolicy`, the process is the same, but instead of a single `mergeFactor` value, the `segmentsPerTier` setting is used as the threshold to decide if a merge should happen, and the `maxMergeAtOnce` setting determines how many segments should be included in the merge.

Choosing the best merge factors is generally a trade-off of indexing speed vs. searching speed. Having fewer segments in the index generally accelerates searches, because there are fewer places to look. It also can also result in fewer physical files on disk. But to keep the number of segments low, merges will occur

more often, which can add load to the system and slow down updates to the index.

Conversely, keeping more segments can accelerate indexing, because merges happen less often, making an update is less likely to trigger a merge. But searches become more computationally expensive and will likely be slower, because search terms must be looked up in more index segments. Faster index updates also means shorter commit turnaround times, which means more timely search results.

## Customizing Merge Policies

If the configuration options for the built-in merge policies do not fully suit your use case, you can customize them: either by creating a custom merge policy factory that you specify in your configuration, or by configuring a [merge policy wrapper](#) which uses a wrapped.prefix configuration option to control how the factory it wraps will be configured:

```
<mergePolicyFactory class="org.apache.solr.index.SortingMergePolicyFactory">
 <str name="sort">timestamp desc</str>
 <str name="wrapped.prefix">inner</str>
 <str name="inner.class">org.apache.solr.index.TieredMergePolicyFactory</str>
 <int name="inner.maxMergeAtOnce">10</int>
 <int name="inner.segmentsPerTier">10</int>
</mergePolicyFactory>
```

The example above shows Solr's SortingMergePolicyFactory being configured to sort documents in merged segments by "timestamp desc", and wrapped around a TieredMergePolicyFactory configured to use the values maxMergeAtOnce=10 and segmentsPerTier=10 via the inner prefix defined by SortingMergePolicyFactory 's wrapped.prefix option. For more information on using SortingMergePolicyFactory, see [the segmentTerminateEarly parameter](#).

## mergeScheduler

The merge scheduler controls how merges are performed. The default ConcurrentMergeScheduler performs merges in the background using separate threads. The alternative, SerialMergeScheduler, does not perform merges with separate threads.

The ConcurrentMergeScheduler has the following configurable attributes:

### maxMergeCount

The maximum number of simultaneous merges that are allowed. If a merge is necessary yet we already have this many threads running, the indexing thread will block until a merge thread has completed. Note that Solr will only run the smallest maxThreadCount merges at a time.

### maxThreadCount

The maximum number of simultaneous merge threads that should be running at once. This must be less than maxMergeCount.

### ioThrottle

A Boolean value (true/ false) to explicitly control I/O throttling. By default throttling is enabled and the CMS will limit I/O throughput when merging to leave other (search, indexing) some room.

The defaults for the above attributes are dynamically set based on whether the underlying disk drive is

rotational disk or not. Refer to the [Dynamic defaults for ConcurrentMergeScheduler](#) section for more details.

#### Example: Dynamic defaults

```
<mergeScheduler class="org.apache.lucene.index.ConcurrentMergeScheduler"/>
```

#### Example: Explicit defaults

```
<mergeScheduler class="org.apache.lucene.index.ConcurrentMergeScheduler">
 <int name="maxMergeCount">9</int>
 <int name="maxThreadCount">4</int>
</mergeScheduler>
```

### mergedSegmentWarmer

When using Solr in for [Near Real Time Searching](#) a merged segment warmer can be configured to warm the reader on the newly merged segment, before the merge commits. This is not required for near real-time search, but will reduce search latency on opening a new near real-time reader after a merge completes.

```
<mergedSegmentWarmer class="org.apache.lucene.index.SimpleMergedSegmentWarmer"/>
```

## Compound File Segments

Each Lucene segment is typically comprised of a dozen or so files. Lucene can be configured to bundle all of the files for a segment into a single compound file using a file extension of .cfs; it's an abbreviation for Compound File Segment.

CFS segments may incur a minor performance hit for various reasons, depending on the runtime environment. For example, filesystem buffers are typically associated with open file descriptors, which may limit the total cache space available to each index.

On systems where the number of open files allowed per process is limited, CFS may avoid hitting that limit. The open files limit might also be tunable for your OS with the Linux/Unix ulimit command, or something similar for other operating systems.

#### CFS: New Segments vs Merged Segments

To configure whether *newly written segments* should use CFS, see the `useCompoundFile` setting described above. To configure whether *merged segments* use CFS, review the Javadocs for your `mergePolicyFactory`.

Many [Merge Policy](#) implementations support `noCFSRatio` and `maxCFSSegmentSizeMB` settings with default values that prevent compound files from being used for large segments, but do use compound files for small segments.



## Index Locks

## lockType

The LockFactory options specify the locking implementation to use.

The set of valid lock type options depends on the [DirectoryFactory](#) you have configured. The values listed below are supported by StandardDirectoryFactory (the default):

- native (default) uses NativeFSLockFactory to specify native OS file locking. If a second Solr process attempts to access the directory, it will fail. Do not use when multiple Solr web applications are attempting to share a single index.
- simple uses SimpleFSLockFactory to specify a plain file for locking.
- single (expert) uses SingleInstanceLockFactory. Use for special situations of a read-only index directory, or when there is no possibility of more than one process trying to modify the index (even sequentially). This type will protect against multiple cores within the *same* JVM attempting to access the same index. **WARNING!** If multiple Solr instances in different JVMs modify an index, this type will *not* protect against index corruption.
- hdfs uses HdfsLockFactory to support reading and writing index and transaction log files to a HDFS filesystem. See the section [Running Solr on HDFS](#) for more details on using this feature.

For more information on the nuances of each LockFactory, see <http://wiki.apache.org/lucene-java/AvailableLockFactories>.

```
<lockType>native</lockType>
```

## writeLockTimeout

The maximum time to wait for a write lock on an IndexWriter. The default is 1000, expressed in milliseconds.

```
<writeLockTimeout>1000</writeLockTimeout>
```

## Other Indexing Settings

There are a few other parameters that may be important to configure for your implementation. These settings affect how or when updates are made to an index.

### deletionPolicy

Controls how commits are retained in case of rollback. The default is SolrDeletionPolicy, which has sub-parameters for the maximum number of commits to keep (`maxCommitsToKeep`), the maximum number of optimized commits to keep (`maxOptimizedCommitsToKeep`), and the maximum age of any commit to keep (`maxCommitAge`), which supports DateMathParser syntax.

### infoStream

The InfoStream setting instructs the underlying Lucene classes to write detailed debug information from the indexing process as Solr log messages.

```
<deletionPolicy class="solr.SolrDeletionPolicy">
 <str name="maxCommitsToKeep">1</str>
 <str name="maxOptimizedCommitsToKeep">0</str>
 <str name="maxCommitAge">1DAY</str>
</deletionPolicy>
<infoStream>false</infoStream>
```

## RequestHandlers and SearchComponents in SolrConfig

After the `<query>` section of `solrconfig.xml`, request handlers and search components are configured.

A *request handler* processes requests coming to Solr. These might be query requests or index update requests. You will likely need several of these defined, depending on how you want Solr to handle the various requests you will make.

A *search component* is a feature of search, such as highlighting or facetting. The search component is defined in `solrconfig.xml` separate from the request handlers, and then registered with a request handler as needed.

These are often referred to as "requestHandler" and "searchComponent", which is how they are defined in `solrconfig.xml`.

### Request Handlers

Every request handler is defined with a name and a class. The name of the request handler is referenced with the request to Solr, typically as a path. For example, if Solr is installed at `http://localhost:8983/solr/` and you have a collection named "gettingstarted", you can make a request that looks like this:

```
http://localhost:8983/solr/gettingstarted/select?q=solr
```

This query will be processed by the request handler with the name `/select`. We've only used the "q" parameter here, which includes our query term, a simple keyword of "solr". If the request handler has more parameters defined, those will be used with any query we send to this request handler unless they are overridden by the client (or user) in the query itself.

If you have another request handler defined, you would send your request with that name. For example, `/update` is a request handler that handles index updates (i.e., sending new documents to the index). By default, `/select` is a request handler that handles query requests.

Request handlers can also process requests for nested paths of their names, for example, a request using `/myhandler/extrapath` may be processed by a request handler registered with the name `/myhandler`. If a request handler is explicitly defined by the name `/myhandler/extrapath`, that would take precedence over the nested path. This assumes you are using the request handler classes included with Solr; if you create your own request handler, you should make sure it includes the ability to handle nested paths if you want to use them with your custom request handler.

It is also possible to configure defaults for request handlers with a section called `initParams`. These defaults can be used when you want to have common properties that will be used by each separate handler. For

example, if you intend to create several request handlers that will all request the same list of fields in the response, you can configure an `initParams` section with your list of fields. For more information about `initParams`, see the section [InitParams in SolrConfig](#).

## SearchHandlers

The primary request handler defined with Solr by default is the "SearchHandler", which handles search queries. The request handler is defined, and then a list of defaults for the handler are defined with a `defaults` list.

For example, in the default `solrconfig.xml`, the first request handler defined looks like this:

```
<requestHandler name="/select" class="solr.SearchHandler">
 <lst name="defaults">
 <str name="echoParams">explicit</str>
 <int name="rows">10</int>
 </lst>
</requestHandler>
```

This example defines the `rows` parameter, which defines how many search results to return, to "10". The `echoParams` parameter defines that the parameters defined in the query should be returned when debug information is returned. Note also that the way the defaults are defined in the list varies if the parameter is a string, an integer, or another type.

All of the parameters described in the section [Searching](#) can be defined as defaults for any of the SearchHandlers.

Besides `defaults`, there are other options for the `SearchHandler`, which are:

- `appends`: This allows definition of parameters that are added to the user query. These might be [filter queries](#), or other query rules that should be added to each query. There is no mechanism in Solr to allow a client to override these additions, so you should be absolutely sure you always want these parameters applied to queries.

```
<lst name="appends">
 <str name="fq">inStock:true</str>
</lst>
```

In this example, the filter query "inStock:true" will always be added to every query.

- `invariants`: This allows definition of parameters that cannot be overridden by a client. The values defined in an `invariants` section will always be used regardless of the values specified by the user, by the client, in `defaults` or in `appends`.

```
<lst name="invariants">
 <str name="facet.field">cat</str>
 <str name="facet.field">manu_exact</str>
 <str name="facet.query">price:[* TO 500]</str>
 <str name="facet.query">price:[500 TO *]</str>
</lst>
```

In this example, facet fields have been defined which limits the facets that will be returned by Solr. If the client requests facets, the facets defined with a configuration like this are the only facets they will see.

The final section of a request handler definition is components, which defines a list of search components that can be used with a request handler. They are only registered with the request handler. How to define a search component is discussed further on in the section on [Search Components](#) below. The components element can only be used with a request handler that is a SearchHandler.

The solrconfig.xml file includes many other examples of SearchHandlers that can be used or modified as needed.

## UpdateRequestHandlers

The UpdateRequestHandlers are request handlers which process updates to the index.

In this guide, we've covered these handlers in detail in the section [Uploading Data with Index Handlers](#).

## ShardHandlers

It is possible to configure a request handler to search across shards of a cluster, used with distributed search. More information about distributed search and how to configure the shardHandler is in the section [Distributed Search with Index Sharding](#).

## Implicit Request Handlers

Solr includes many out-of-the-box request handlers that are not configured in solrconfig.xml, and so are referred to as "implicit" - see [Implicit RequestHandlers](#).

## Search Components

Search components define the logic that is used by the SearchHandler to perform queries for users.

## Default Components

There are several default search components that work with all SearchHandlers without any additional configuration. If no components are defined (with the exception of first-components and last-components - see below), these are executed by default, in the following order:

Component Name	Class Name	More Information
query	solr.QueryComponent	Described in the section <a href="#">Query Syntax and Parsing</a> .
facet	solr.FacetComponent	Described in the section <a href="#">Faceting</a> .

Component Name	Class Name	More Information
mlt	solr.MoreLikeThisComponent	Described in the section <a href="#">MoreLikeThis</a> .
highlight	solr.HighlightComponent	Described in the section <a href="#">Highlighting</a> .
stats	solr.StatsComponent	Described in the section <a href="#">The Stats Component</a> .
debug	solr.DebugComponent	Described in the section on <a href="#">Common Query Parameters</a> .
expand	solr.ExpandComponent	Described in the section <a href="#">Collapse and Expand Results</a> .

If you register a new search component with one of these default names, the newly defined component will be used instead of the default.

## First-Components and Last-Components

It's possible to define some components as being used before (with `first-components`) or after (with `last-components`) the default components listed above.



`first-components` and/or `last-components` may only be used in conjunction with the default components. If you define your own components, the default components will not be executed, and `first-components` and `last-components` are disallowed.

```
<arr name="first-components">
 <str>mycomponent</str>
</arr>
<arr name="last-components">
 <str>spellcheck</str>
</arr>
```

## Components

If you define components, the default components (see above) will not be executed, and `first-components` and `last-components` are disallowed:

```
<arr name="components">
 <str>mycomponent</str>
 <str>query</str>
 <str>debug</str>
</arr>
```

## Other Useful Components

Many of the other useful components are described in sections of this Guide for the features they support. These are:

- SpellCheckComponent, described in the section [Spell Checking](#).

- TermVectorComponent, described in the section [The Term Vector Component](#).
- QueryElevationComponent, described in the section [The Query Elevation Component](#).
- TermsComponent, described in the section [The Terms Component](#).

## InitParams in SolrConfig

The <initParams> section of solrconfig.xml allows you to define request handler parameters outside of the handler configuration.

There are a couple of use cases where this might be desired:

- Some handlers are implicitly defined in code - see [Implicit RequestHandlers](#) - and there should be a way to add/append/override some of the implicitly defined properties.
- There are a few properties that are used across handlers. This helps you keep only a single definition of those properties and apply them over multiple handlers.

For example, if you want several of your search handlers to return the same list of fields, you can create an <initParams> section without having to define the same set of parameters in each request handler definition. If you have a single request handler that should return different fields, you can define the overriding parameters in individual <requestHandler> sections as usual.

The properties and configuration of an <initParams> section mirror the properties and configuration of a request handler. It can include sections for defaults, appends, and invariants, the same as any request handler.

For example, here is one of the <initParams> sections defined by default in the \_default example:

```
<initParams path="/update/**,/query,/select,/tvrh,/elevate,/spell,/browse">
 <lst name="defaults">
 <str name="df">_text_-</str>
 </lst>
</initParams>
```

This sets the default search field ("df") to be "text" for all of the request handlers named in the path section. If we later want to change the /query request handler to search a different field by default, we could override the <initParams> by defining the parameter in the <requestHandler> section for /query.

The syntax and semantics are similar to that of a <requestHandler>. The following are the attributes:

path

A comma-separated list of paths which will use the parameters. Wildcards can be used in paths to define nested paths, as described below.

name

The name of this set of parameters. The name can be used directly in a requestHandler definition if a path is not explicitly named. If you give your <initParams> a name, you can refer to the parameters in a <requestHandler> that is not defined as a path.

For example, if an <initParams> section has the name "myParams", you can call the name when defining

your request handler:

```
<requestHandler name="/dump1" class="DumpRequestHandler" initParams="myParams"/>
```

## Wildcards in initParams

An `<initParams>` section can support wildcards to define nested paths that should use the parameters defined. A single asterisk (\*) denotes that a nested path one level deeper should use the parameters. Double asterisks (\*\*\*) denote all nested paths no matter how deep should use the parameters.

For example, if we have an `<initParams>` that looks like this:

```
<initParams name="myParams" path="/myhandler,/root/*,/root1/**">
 <lst name="defaults">
 <str name="f1">_text_</str>
 </lst>
 <lst name="invariants">
 <str name="rows">10</str>
 </lst>
 <lst name="appends">
 <str name="df">title</str>
 </lst>
</initParams>
```

We've defined three paths with this section:

- `/myhandler` declared as a direct path.
- `/root/*` with a single asterisk to indicate the parameters should apply to paths that are one level deep.
- `/root1/**` with double asterisks to indicate the parameters should apply to all nested paths, no matter how deep.

When we define the request handlers, the wildcards will work in the following ways:

```
<requestHandler name="/myhandler" class="SearchHandler"/>
```

The `/myhandler` class was named as a path in the `<initParams>` so this will use those parameters.

Next we have a request handler named `/root/search5`:

```
<requestHandler name="/root/search5" class="SearchHandler"/>
```

We defined a wildcard for nested paths that are one level deeper than `/root`, so this request handler will use the parameters. This one, however, will not, because `/root/search5/test` is more than one level deep from `/root`:

```
<requestHandler name="/root/search5/test" class="SearchHandler"/>
```

If we want to define all levels of nested paths, we should use double asterisks, as in the example path /root1/\*\*:

```
<requestHandler name="/root1/search/tests" class="SearchHandler"/>
```

Any path under /root1, whether explicitly defined in a request handler or not, will use the parameters defined in the matching initParams section.

## UpdateHandlers in SolrConfig

The settings in this section are configured in the <updateHandler> element in solrconfig.xml and may affect the performance of index updates. These settings affect how updates are done internally.

<updateHandler> configurations do not affect the higher level configuration of [RequestHandlers](#) that process client update requests.

```
<updateHandler class="solr.DirectUpdateHandler2">
 ...
</updateHandler>
```

## Commits

Data sent to Solr is not searchable until it has been *committed* to the index. The reason for this is that in some cases commits can be slow and they should be done in isolation from other possible commit requests to avoid overwriting data. So, it's preferable to provide control over when data is committed. Several options are available to control the timing of commits.

### commit and softCommit

In Solr, a commit is an action which asks Solr to "commit" those changes to the Lucene index files. By default commit actions result in a "hard commit" of all the Lucene index files to stable storage (disk). When a client includes a `commit=true` parameter with an update request, this ensures that all index segments affected by the adds & deletes on an update are written to disk as soon as index updates are completed.

If an additional flag `softCommit=true` is specified, then Solr performs a 'soft commit', meaning that Solr will commit your changes to the Lucene data structures quickly but not guarantee that the Lucene index files are written to stable storage. This is an implementation of Near Real Time storage, a feature that boosts document visibility, since you don't have to wait for background merges and storage (to ZooKeeper, if using [SolrCloud](#)) to finish before moving on to something else. A full commit means that, if a server crashes, Solr will know exactly where your data was stored; a soft commit means that the data is stored, but the location information isn't yet stored. The tradeoff is that a soft commit gives you faster visibility because it's not waiting for background merges to finish.

For more information about Near Real Time operations, see [Near Real Time Searching](#).

### autoCommit

These settings control how often pending updates will be automatically pushed to the index. An alternative to autoCommit is to use `commitWithin`, which can be defined when making the update request to Solr (i.e.,

when pushing documents), or in an update RequestHandler.

#### maxDocs

The number of updates that have occurred since the last commit.

#### maxTime

The number of milliseconds since the oldest uncommitted update.

#### maxSize

The maximum size of the transaction log (tlog) on disk, after which a hard commit is triggered. This is useful when the size of documents is unknown and the intention is to restrict the size of the transaction log to reasonable size. Valid values can be bytes (default with no suffix), kilobytes (if defined with a k suffix, as in 25k), megabytes (m) or gigabytes (g).

#### openSearcher

Whether to open a new searcher when performing a commit. If this is false, the commit will flush recent index changes to stable storage, but does not cause a new searcher to be opened to make those changes visible. The default is true.

If any of the maxDocs, maxTime, or maxSize limits are reached, Solr automatically performs a commit operation. If the autoCommit tag is missing, then only explicit commits will update the index. The decision whether to use auto-commit or not depends on the needs of your application.

Determining the best auto-commit settings is a tradeoff between performance and accuracy. Settings that cause frequent updates will improve the accuracy of searches because new content will be searchable more quickly, but performance may suffer because of the frequent updates. Less frequent updates may improve performance but it will take longer for updates to show up in queries.

```
<autoCommit>
 <maxDocs>10000</maxDocs>
 <maxTime>30000</maxTime>
 <maxSize>512m</maxSize>
 <openSearcher>false</openSearcher>
</autoCommit>
```

You can also specify 'soft' autoCommits in the same way that you can specify 'soft' commits, except that instead of using autoCommit you set the autoSoftCommit tag.

```
<autoSoftCommit>
 <maxTime>60000</maxTime>
</autoSoftCommit>
```

#### commitWithin

The commitWithin settings allow forcing document commits to happen in a defined time period. This is used most frequently with [Near Real Time Searching](#), and for that reason the default is to perform a soft commit. This does not, however, replicate new documents to slave servers in a master/slave environment. If that's a requirement for your implementation, you can force a hard commit by adding a parameter, as in this example:

```
<commitWithin>
 <softCommit>false</softCommit>
</commitWithin>
```

With this configuration, when you call commitWithin as part of your update message, it will automatically perform a hard commit every time.

## Event Listeners

The UpdateHandler section is also where update-related event listeners can be configured. These can be triggered to occur after any commit (event="postCommit") or only after optimize commands (event="postOptimize").

Users can write custom update event listener classes in Solr plugins. As of Solr 7.1, RunExecutableListener was removed for security reasons.

## Transaction Log

As described in the section [RealTime Get](#), a transaction log is required for that feature. It is configured in the updateHandler section of solrconfig.xml.

Realtime Get currently relies on the update log feature, which is enabled by default. It relies on an update log, which is configured in solrconfig.xml, in a section like:

```
<updateLog>
 <str name="dir">${solr.ulog.dir:}</str>
</updateLog>
```

Three additional expert-level configuration settings affect indexing performance and how far a replica can fall behind on updates before it must enter into full recovery - see the section on [write side fault tolerance](#) for more information:

numRecordsToKeep

The number of update records to keep per log. The default is 100.

maxNumLogsToKeep

The maximum number of logs keep. The default is 10.

numVersionBuckets

The number of buckets used to keep track of max version values when checking for re-ordered updates; increase this value to reduce the cost of synchronizing access to version buckets during high-volume indexing, this requires (8 bytes (long) \* numVersionBuckets) of heap space per Solr core. The default is 65536.

An example, to be included under <config><updateHandler> in solrconfig.xml, employing the above advanced settings:

```
<updateLog>
 <str name="dir">${solr.ulog.dir:}</str>
 <int name="numRecordsToKeep">500</int>
 <int name="maxNumLogsToKeep">20</int>
 <int name="numVersionBuckets">65536</int>
</updateLog>
```

## Other Options

In some cases complex updates (such as spatial/shape) may take very long time to complete. In the default configuration other updates that fall into the same internal version bucket will wait indefinitely and eventually these outstanding requests may pile up and lead to thread exhaustion and eventually to OutOfMemory errors.

The option `versionBucketLockTimeoutMs` in the `updateHandler` section helps to prevent that by specifying a limited timeout for such extremely long running update requests. If this limit is reached this update will fail but it won't block forever all other updates. See SOLR-12833 for more details.

There's a memory cost associated with this setting. Values greater than the default 0 (meaning unlimited timeout) cause Solr to use a different internal implementation of the version bucket, which increases memory consumption from ~1.5MB to ~6.8MB per Solr core.

An example of specifying this option under `<config>` section of `solrconfig.xml`:

```
<updateHandler class="solr.DirectUpdateHandler2">
 ...
 <int name="versionBucketLockTimeoutMs">10000</int>
</updateHandler>
```

## Query Settings in SolrConfig

The settings in this section affect the way that Solr will process and respond to queries.

These settings are all configured in child elements of the `<query>` element in `solrconfig.xml`.

```
<query>
 ...
</query>
```

## Caches

Solr caches are associated with a specific instance of an Index Searcher, a specific view of an index that doesn't change during the lifetime of that searcher. As long as that Index Searcher is being used, any items in its cache will be valid and available for reuse. Caching in Solr differs from caching in many other applications in that cached Solr objects do not expire after a time interval; instead, they remain valid for the lifetime of the Index Searcher.

When a new searcher is opened, the current searcher continues servicing requests while the new one auto-warms its cache. The new searcher uses the current searcher's cache to pre-populate its own. When the new searcher is ready, it is registered as the current searcher and begins handling all new search requests. The old searcher will be closed once it has finished servicing all its requests.

In Solr, there are three cache implementations: `solr.search.LRUCache`, `solr.search.FastLRUCache`, and `solr.search.LFUCache`.

The acronym LRU stands for Least Recently Used. When an LRU cache fills up, the entry with the oldest last-accessed timestamp is evicted to make room for the new entry. The net effect is that entries that are accessed frequently tend to stay in the cache, while those that are not accessed frequently tend to drop out and will be re-fetched from the index if needed again.

The FastLRUCache, which was introduced in Solr 1.4, is designed to be lock-free, so it is well suited for caches which are hit several times in a request.

Both LRUCache and FastLRUCache use an auto-warm count that supports both integers and percentages which get evaluated relative to the current size of the cache when warming happens.

The LFUCache refers to the Least Frequently Used cache. This works in a way similar to the LRU cache, except that when the cache fills up, the entry that has been used the least is evicted.

The Statistics page in the Solr Admin UI will display information about the performance of all the active caches. This information can help you fine-tune the sizes of the various caches appropriately for your particular application. When a Searcher terminates, a summary of its cache usage is also written to the log.

Each cache has settings to define its initial size (`initialSize`), maximum size (`size`) and number of items to use for during warming (`autowarmCount`). The LRU and FastLRU cache implementations can take a percentage instead of an absolute value for `autowarmCount`.

FastLRUCache and LFUCache support `showItems` attribute. This is the number of cache items to display in the stats page for the cache. It is for debugging.

Details of each cache are described below.

## filterCache

This cache is used by `SolrIndexSearcher` for filters (DocSets) for unordered sets of all documents that match a query. The numeric attributes control the number of entries in the cache.

The most typical way Solr uses the `filterCache` is to cache results of each `fq` search parameter, though there are some other cases as well. Subsequent queries using the same parameter filter query result in cache hits and rapid returns of results. See [Searching](#) for a detailed discussion of the `fq` parameter. Another Solr feature using this cache is the `filter(...)` syntax in the default Lucene query parser.

Solr also uses this cache for faceting when the configuration parameter `facet.method` is set to `fc`. For a discussion of faceting, see [Searching](#).

The filter cache uses a specialized cache named as `FastLRUCache` which is optimized for fast concurrent access with the trade-off that writes and evictions are costlier than the `LRUCache` used for query result cache and document cache.

```
<filterCache class="solr.FastLRUCache"
 size="512"
 initialSize="512"
 autowarmCount="128"/>
```

The FastLRUCache used for filter cache also supports a maxRamMB parameter which restricts the maximum amount of heap used by this cache. The FastLRUCache only supports evictions by either heap usage or size but not both. Therefore, the size parameter is ignored if maxRamMB is specified.

```
<filterCache class="solr.FastLRUCache"
 maxRamMB="1000"
 autowarmCount="128"/>
```

## queryResultCache

This cache holds the results of previous searches: ordered lists of document IDs (DocList) based on a query, a sort, and the range of documents requested.

The queryResultCache has an additional (optional) setting to limit the maximum amount of RAM used (maxRamMB). This lets you specify the maximum heap size, in megabytes, used by the contents of this cache. When the cache grows beyond this size, oldest accessed queries will be evicted until the heap usage of the cache decreases below the specified limit. If a size is specified in addition to maxRamMB then both heap usage and maximum size limits are respected.

```
<queryResultCache class="solr.LRUCache"
 size="512"
 initialSize="512"
 autowarmCount="128"
 maxRamMB="1000"/>
```

## documentCache

This cache holds Lucene Document objects (the stored fields for each document). Since Lucene internal document IDs are transient, this cache is not auto-warmed. The size for the documentCache should always be greater than max\_results times the max\_concurrent\_queries, to ensure that Solr does not need to refetch a document during a request. The more fields you store in your documents, the higher the memory usage of this cache will be.

```
<documentCache class="solr.LRUCache"
 size="512"
 initialSize="512"
 autowarmCount="0"/>
```

## User Defined Caches

You can also define named caches for your own application code to use. You can locate and use your cache object by name by calling the SolrIndexSearcher methods getCache(), cacheLookup() and cacheInsert().

```
<cache name="myUserCache" class="solr.LRUCache"
 size="4096"
 initialSize="1024"
 autowarmCount="1024"
 regenerator="org.mycompany.mypackage.MyRegenerator" />
```

If you want auto-warming of your cache, include a regenerator attribute with the fully qualified name of a class that implements `solr.search.CacheRegenerator`. You can also use the `NoOpRegenerator`, which simply repopulates the cache with old items. Define it with the regenerator parameter as`:  
`regenerator="solr.NoOpRegenerator"``.

## Query Sizing and Warming

### maxBooleanClauses

Sets the maximum number of clauses allowed when parsing a boolean query string.

This limit only impacts boolean queries specified by a user as part of a query string, and provides per-collection controls on how complex user specified boolean queries can be. Query strings that specify more clauses than this will result in an error.

If this per-collection limit is greater than [the global maxBooleanClauses limit specified in solr.xml](#) it will have no effect, as that setting also limits the size of user specified boolean queries.

In default configurations this property uses the value of the `solr.max.booleanClauses` system property if specified. This is the same system property used in the [global maxBooleanClauses setting in the default solr.xml](#) making it easy for Solr administrators to increase both values (in all collections) without needing to search through and update all of their configs.

```
<maxBooleanClauses>${solr.max.booleanClauses:1024}</maxBooleanClauses>
```

### enableLazyFieldLoading

If this parameter is set to true, then fields that are not directly requested will be loaded lazily as needed. This can boost performance if the most common queries only need a small subset of fields, especially if infrequently accessed fields are large in size.

```
<enableLazyFieldLoading>true</enableLazyFieldLoading>
```

### useFilterForSortedQuery

This parameter configures Solr to use a filter to satisfy a search. If the requested sort does not include "score", the filterCache will be checked for a filter matching the query. For most situations, this is only useful if the same search is requested often with different sort options and none of them ever use "score".

```
<useFilterForSortedQuery>true</useFilterForSortedQuery>
```

## queryResultWindowSize

Used with the queryResultCache, this will cache a superset of the requested number of document IDs. For example, if the a search in response to a particular query requests documents 10 through 19, and queryWindowSize is 50, documents 0 through 49 will be cached.

```
<queryResultWindowSize>20</queryResultWindowSize>
```

## queryResultMaxDocsCached

This parameter sets the maximum number of documents to cache for any entry in the queryResultCache.

```
<queryResultMaxDocsCached>200</queryResultMaxDocsCached>
```

## useColdSearcher

This setting controls whether search requests for which there is not a currently registered searcher should wait for a new searcher to warm up (false) or proceed immediately (true). When set to "false", requests will block until the searcher has warmed its caches.

```
<useColdSearcher>false</useColdSearcher>
```

## maxWarmingSearchers

This parameter sets the maximum number of searchers that may be warming up in the background at any given time. Exceeding this limit will raise an error. For read-only slaves, a value of two is reasonable. Masters should probably be set a little higher.

```
<maxWarmingSearchers>2</maxWarmingSearchers>
```

## Query-Related Listeners

As described in the section on [Caches](#), new Index Searchers are cached. It's possible to use the triggers for listeners to perform query-related tasks. The most common use of this is to define queries to further "warm" the Index Searchers while they are starting. One benefit of this approach is that field caches are pre-populated for faster sorting.

Good query selection is key with this type of listener. It's best to choose your most common and/or heaviest queries and include not just the keywords used, but any other parameters such as sorting or filtering requests.

There are two types of events that can trigger a listener. A firstSearcher event occurs when a new searcher is being prepared but there is no current registered searcher to handle requests or to gain auto-warming data from (i.e., on Solr startup). A newSearcher event is fired whenever a new searcher is being prepared and there is a current searcher handling requests.

The (commented out) examples below can be found in the solrconfig.xml file of the

sample\_techproducts\_configs [configset](#) included with Solr, and demonstrate using the `solr.QuerySenderListener` class to warm a set of explicit queries:

```
<listener event="newSearcher" class="solr.QuerySenderListener">
 <arr name="queries">
 <!--
 <lst><str name="q">solr</str><str name="sort">price asc</str></lst>
 <lst><str name="q">rocks</str><str name="sort">weight asc</str></lst>
 -->
 </arr>
</listener>

<listener event="firstSearcher" class="solr.QuerySenderListener">
 <arr name="queries">
 <lst><str name="q">static firstSearcher warming in solrconfig.xml</str></lst>
 </arr>
</listener>
```

The above code comes from a *sample* `solrconfig.xml`.

A key best practice is to modify these defaults before taking your application to production, but please note: while the sample queries are commented out in the section for the "newSearcher", the sample query is not commented out for the "firstSearcher" event.

 There is no point in auto-warming your Index Searcher with the query string "static firstSearcher warming in solrconfig.xml" if that is not relevant to your search application.

## RequestDispatcher in SolrConfig

The `requestDispatcher` element of `solrconfig.xml` controls the way the Solr HTTP RequestDispatcher implementation responds to requests.

Included are parameters for defining if it should handle `/select` urls (for Solr 1.1 compatibility), if it will support remote streaming, the maximum size of file uploads and how it will respond to HTTP cache headers in requests.

### handleSelect Element



`handleSelect` is for legacy back-compatibility; those new to Solr do not need to change anything about the way this is configured by default.

The first configurable item is the `handleSelect` attribute on the `<requestDispatcher>` element itself. This attribute can be set to one of two values, either "true" or "false". It governs how Solr responds to requests such as `/select?qt=XXX`. The default value "false" will ignore requests to `/select` if a `requestHandler` is not explicitly registered with the name `/select`. A value of "true" will route query requests to the parser defined with the `qt` value.

In recent versions of Solr, a `/select` `requestHandler` is defined by default, so a value of "false" will work fine. See the section [RequestHandlers and SearchComponents in SolrConfig](#) for more information.

```
<requestDispatcher handleSelect="true" >
...
</requestDispatcher>
```

## requestParsers Element

The `<requestParsers>` sub-element controls values related to parsing requests. This is an empty XML element that doesn't have any content, only attributes.

### enableRemoteStreaming

This attribute controls whether remote streaming of content is allowed. If omitted or set to `false` (the default), streaming will not be allowed. Setting it to `true` lets you specify the location of content to be streamed using `stream.file` or `stream.url` parameters.

### enableStreamBody

This attribute controls whether streaming content from the HTTP parameter `stream.body` is allowed. If omitted or set to `false` (the default), streaming will not be allowed. Setting it to `true` lets you pass data in the `stream.body` parameter.

If you enable remote streaming, be sure that you have authentication enabled. Otherwise, someone could potentially gain access to your content by accessing arbitrary URLs. It's also a good idea to place Solr behind a firewall to prevent it from being accessed from untrusted clients.

### multipartUploadLimitInKB

This attribute sets an upper limit in kilobytes on the size of a document that may be submitted in a multi-part HTTP POST request. The value specified is multiplied by 1024 to determine the size in bytes. A value of `-1` means `MAX_INT`, which is also the system default if omitted.

### formdataUploadLimitInKB

This attribute sets a limit in kilobytes on the size of form data (`application/x-www-form-urlencoded`) submitted in a HTTP POST request, which can be used to pass request parameters that will not fit in a URL. A value of `-1` means `MAX_INT`, which is also the system default if omitted.

### addHttpRequestToContext

This attribute can be used to indicate that the original `HttpServletRequest` object should be included in the context map of the `SolrQueryRequest` using the key `httpRequest`. This `HttpServletRequest` is not used by any Solr component, but may be useful when developing custom plugins.

```
<requestParsers enableRemoteStreaming="false"
 enableStreamBody="false"
 multipartUploadLimitInKB="2048"
 formdataUploadLimitInKB="2048"
 addHttpRequestToContext="false" />
```

The below command is an example of how to enable RemoteStreaming and BodyStreaming through the [Config API](#):

## V1 API

```
curl -H 'Content-type:application/json' -d '{"set-property": {"requestDispatcher.requestParsers.enableRemoteStreaming": true}, "set-property": {"requestDispatcher.requestParsers.enableStreamBody": true}}' http://localhost:8983/solr/gettingstarted/config
```

## V2 API Standalone Solr

```
curl -H 'Content-type:application/json' -d '{"set-property": {"requestDispatcher.requestParsers.enableRemoteStreaming": true}, "set-property": {"requestDispatcher.requestParsers.enableStreamBody": true}}' http://localhost:8983/api/cores/gettingstarted/config
```

## V2 API SolrCloud

```
curl -H 'Content-type:application/json' -d '{"set-property": {"requestDispatcher.requestParsers.enableRemoteStreaming": true}, "set-property": {"requestDispatcher.requestParsers.enableStreamBody": true}}' http://localhost:8983/api/collections/gettingstarted/config
```

## httpCaching Element

The `<httpCaching>` element controls HTTP cache control headers. Do not confuse these settings with Solr's internal cache configuration. This element controls caching of HTTP responses as defined by the W3C HTTP specifications.

This element allows for three attributes and one sub-element. The attributes of the `<httpCaching>` element control whether a 304 response to a GET request is allowed, and if so, what sort of response it should be. When an HTTP client application issues a GET, it may optionally specify that a 304 response is acceptable if the resource has not been modified since the last time it was fetched.

never304

If present with the value `true`, then a GET request will never respond with a 304 code, even if the requested resource has not been modified. When this attribute is set to `true`, the next two attributes are ignored. Setting this to `true` is handy for development, as the 304 response can be confusing when tinkering with Solr responses through a web browser or other client that supports cache headers.

lastModFrom

This attribute may be set to either `openTime` (the default) or `dirLastMod`. The value `openTime` indicates that last modification times, as compared to the `If-Modified-Since` header sent by the client, should be calculated relative to the time the Searcher started. Use `dirLastMod` if you want times to exactly correspond to when the index was last updated on disk.

## etagSeed

This value of this attribute is sent as the value of the ETag header. Changing this value can be helpful to force clients to re-fetch content even when the indexes have not changed---for example, when you've made some changes to the configuration.

```
<httpCaching never304="false"
 lastModFrom="openTime"
 etagSeed="Solr">
 <cacheControl>max-age=30, public</cacheControl>
</httpCaching>
```

## cacheControl Element

In addition to these attributes, `<httpCaching>` accepts one child element: `<cacheControl>`. The content of this element will be sent as the value of the Cache-Control header on HTTP responses. This header is used to modify the default caching behavior of the requesting client. The possible values for the Cache-Control header are defined by the HTTP 1.1 specification in [Section 14.9](#).

Setting the max-age field controls how long a client may re-use a cached response before requesting it again from the server. This time interval should be set according to how often you update your index and whether or not it is acceptable for your application to use content that is somewhat out of date. Setting must-revalidate will tell the client to validate with the server that its cached copy is still good before re-using it. This will ensure that the most timely result is used, while avoiding a second fetch of the content if it isn't needed, at the cost of a request to the server to do the check.

# Update Request Processors

Every update request received by Solr is run through a chain of plugins known as Update Request Processors, or *URPs*.

This can be useful, for example, to add a field to the document being indexed; to change the value of a particular field; or to drop an update if the incoming document doesn't fulfill certain criteria. In fact, a surprisingly large number of features in Solr are implemented as Update Processors and therefore it is necessary to understand how such plugins work and where are they configured.

## URP Anatomy and Lifecycle

An Update Request Processor is created as part of a [chain](#) of one or more update processors. Solr creates a default update request processor chain comprising of a few update request processors which enable essential Solr features. This default chain is used to process every update request unless a user chooses to configure and specify a different custom update request processor chain.

The easiest way to describe an Update Request Processor is to look at the Javadocs of the abstract class [UpdateRequestProcessor](#). Every `UpdateRequestProcessor` must have a corresponding factory class which extends [UpdateRequestProcessorFactory](#). This factory class is used by Solr to create a new instance of this plugin. Such a design provides two benefits:

1. An update request processor need not be thread safe because it is used by one and only one request thread and destroyed once the request is complete.

2. The factory class can accept configuration parameters and maintain any state that may be required between requests. The factory class must be thread-safe.

Every update request processor chain is constructed during loading of a Solr core and cached until the core is unloaded. Each `UpdateRequestProcessorFactory` specified in the chain is also instantiated and initialized with configuration that may have been specified in `solrconfig.xml`.

When an update request is received by Solr, it looks up the update chain to be used for this request. A new instance of each `UpdateRequestProcessor` specified in the chain is created using the corresponding factory. The update request is parsed into corresponding `UpdateCommand` objects which are run through the chain. Each `UpdateRequestProcessor` instance is responsible for invoking the next plugin in the chain. It can choose to short circuit the chain by not invoking the next processor and even abort further processing by throwing an exception.



A single update request may contain a batch of multiple new documents or deletes and therefore the corresponding `processXXX` methods of an `UpdateRequestProcessor` will be invoked multiple times for every individual update. However, it is guaranteed that a single thread will serially invoke these methods.

## Update Request Processor Configuration

Update request processors chains can be created by either creating the whole chain directly in `solrconfig.xml` or by creating individual update processors in `solrconfig.xml` and then dynamically creating the chain at run-time by specifying all processors via request parameters.

However, before we understand how to configure update processor chains, we must learn about the default update processor chain because it provides essential features which are needed in most custom request processor chains as well.

### Default Update Request Processor Chain

In case no update processor chains are configured in `solrconfig.xml`, Solr will automatically create a default update processor chain which will be used for all update requests. This default update processor chain consists of the following processors (in order):

1. `LogUpdateProcessorFactory` - Tracks the commands processed during this request and logs them
2. `DistributedUpdateProcessorFactory` - Responsible for distributing update requests to the right node e.g., routing requests to the leader of the right shard and distributing updates from the leader to each replica. This processor is activated only in SolrCloud mode.
3. `RunUpdateProcessorFactory` - Executes the update using internal Solr APIs.

Each of these perform an essential function and as such any custom chain usually contain all of these processors. The `RunUpdateProcessorFactory` is usually the last update processor in any custom chain.

### Custom Update Request Processor Chain

The following example demonstrates how a custom chain can be configured inside `solrconfig.xml`.

### Example dedupe updateRequestProcessorChain

```
<updateRequestProcessorChain name="dedupe">
 <processor class="solr.processor.SignatureUpdateProcessorFactory">
 <bool name="enabled">true</bool>
 <str name="signatureField">id</str>
 <bool name="overwriteDuplicates">false</bool>
 <str name="fields">name,features,cat</str>
 <str name="signatureClass">solr.processor.Lookup3Signature</str>
 </processor>
 <processor class="solr.LogUpdateProcessorFactory" />
 <processor class="solr.RunUpdateProcessorFactory" />
</updateRequestProcessorChain>
```

In the above example, a new update processor chain named "dedupe" is created with SignatureUpdateProcessorFactory, LogUpdateProcessorFactory and RunUpdateProcessorFactory in the chain. The SignatureUpdateProcessorFactory is further configured with different parameters such as "signatureField", "overwriteDuplicates", etc. This chain is an example of how Solr can be configured to perform de-duplication of documents by calculating a signature using the value of name, features, cat fields which is then used as the "id" field. As you may have noticed, this chain does not specify the DistributedUpdateProcessorFactory. Because this processor is critical for Solr to operate properly, Solr will automatically insert DistributedUpdateProcessorFactory in any chain that does not include it just prior to the RunUpdateProcessorFactory.

#### *RunUpdateProcessorFactory*



Do not forget to add RunUpdateProcessorFactory at the end of any chains you define in solrconfig.xml. Otherwise update requests processed by that chain will not actually affect the indexed data.

## Configuring Individual Processors as Top-Level Plugins

Update request processors can also be configured independent of a chain in solrconfig.xml.

### *updateProcessor Configuration*

```
<updateProcessor class="solr.processor.SignatureUpdateProcessorFactory" name="signature">
 <bool name="enabled">true</bool>
 <str name="signatureField">id</str>
 <bool name="overwriteDuplicates">false</bool>
 <str name="fields">name,features,cat</str>
 <str name="signatureClass">solr.processor.Lookup3Signature</str>
</updateProcessor>
<updateProcessor class="solr.RemoveBlankFieldUpdateProcessorFactory" name="remove_blanks" />
```

In this case, an instance of SignatureUpdateProcessorFactory is configured with the name "signature" and a RemoveBlankFieldUpdateProcessorFactory is defined with the name "remove\_blanks". Once the above has been specified in solrconfig.xml, we can refer to them in update request processor chains in solrconfig.xml as follows:

## updateRequestProcessorChain Configuration

```
<updateProcessorChain name="custom" processor="remove_blanks,signature">
 <processor class="solr.RunUpdateProcessorFactory" />
</updateProcessorChain>
```

## Update Processors in SolrCloud

In a single node, stand-alone Solr, each update is run through all the update processors in a chain exactly once. But the behavior of update request processors in SolrCloud deserves special consideration.

A critical SolrCloud functionality is the routing and distributing of requests. For update requests this routing is implemented by the `DistributedUpdateRequestProcessor`, and this processor is given a special status by Solr due to its important function.

In SolrCloud mode, all processors in the chain *before* the `DistributedUpdateProcessor` are run on the first node that receives an update from the client, regardless of this node's status as a leader or replica. The `DistributedUpdateProcessor` then forwards the update to the appropriate shard leader for the update (or to multiple leaders in the event of an update that affects multiple documents, such as a delete by query or commit). The shard leader uses a transaction log to apply [Atomic Updates & Optimistic Concurrency](#) and then forwards the update to all of the shard replicas. The leader and each replica run all of the processors in the chain that are listed *after* the `DistributedUpdateProcessor`.

For example, consider the "dedupe" chain which we saw in a section above. Assume that a 3-node SolrCloud cluster exists where node A hosts the leader of shard1, node B hosts the leader of shard2 and node C hosts the replica of shard2. Assume that an update request is sent to node A which forwards the update to node B (because the update belongs to shard2) which then distributes the update to its replica node C. Let's see what happens at each node:

- **Node A:** Runs the update through the `SignatureUpdateProcessor` (which computes the signature and puts it in the "id" field), then `LogUpdateProcessor` and then `DistributedUpdateProcessor`. This processor determines that the update actually belongs to node B and is forwarded to node B. The update is not processed further. This is required because the next processor, `RunUpdateProcessor`, will execute the update against the local shard1 index which would lead to duplicate data on shard1 and shard2.
- **Node B:** Receives the update and sees that it was forwarded by another node. The update is directly sent to `DistributedUpdateProcessor` because it has already been through the `SignatureUpdateProcessor` on node A and doing the same signature computation again would be redundant. The `DistributedUpdateProcessor` determines that the update indeed belongs to this node, distributes it to its replica on Node C and then forwards the update further in the chain to `RunUpdateProcessor`.
- **Node C:** Receives the update and sees that it was distributed by its leader. The update is directly sent to `DistributedUpdateProcessor` which performs some consistency checks and forwards the update further in the chain to `RunUpdateProcessor`.

In summary:

1. All processors before `DistributedUpdateProcessor` are only run on the first node that receives an update request whether it be a forwarding node (e.g., node A in the above example) or a leader (e.g., node B). We call these "pre-processors" or just "processors".

2. All processors after DistributedUpdateProcessor run only on the leader and the replica nodes. They are not executed on forwarding nodes. Such processors are called "post-processors".

In the previous section, we saw that the updateRequestProcessorChain was configured with processor="remove\_blanks, signature". This means that such processors are of the #1 kind and are run only on the forwarding nodes. Similarly, we can configure them as the #2 kind by specifying with the attribute "post-processor" as follows:

#### *post-processor Configuration*

```
<updateProcessorChain name="custom" processor="signature" post-processor="remove_blanks">
 <processor class="solr.RunUpdateProcessorFactory" />
</updateProcessorChain>
```

However executing a processor only on the forwarding nodes is a great way of distributing an expensive computation such as de-duplication across a SolrCloud cluster by sending requests randomly via a load balancer. Otherwise the expensive computation is repeated on both the leader and replica nodes.

*Custom update chain post-processors may never be invoked on a recovering replica*



While a replica is in [recovery](#), inbound update requests are buffered to the transaction log. After recovery has completed successfully, those buffered update requests are replayed. As of this writing, however, custom update chain post-processors are never invoked for buffered update requests. See [SOLR-8030](#). To work around this problem until SOLR-8030 has been fixed, **avoid specifying post-processors in custom update chains**.

## Atomic Update Processor Factory

If the AtomicUpdateProcessorFactory is in the update chain before the DistributedUpdateProcessor, the incoming document to the chain will be a partial document.

Because DistributedUpdateProcessor is responsible for processing [Atomic Updates](#) into full documents on the leader node, this means that pre-processors which are executed only on the forwarding nodes can only operate on the partial document. If you have a processor which must process a full document then the only choice is to specify it as a post-processor.

## Using Custom Chains

### [update.chain Request Parameter](#)

The update .chain parameter can be used in any update request to choose a custom chain which has been configured in solrconfig.xml. For example, in order to choose the "dedupe" chain described in a previous section, one can issue the following request:

## Using update.chain

```
curl "http://localhost:8983/solr/gettingstarted/update/json?update.chain=dedupe&commit=true" -H
'Content-type: application/json' -d '
[
 {
 "name" : "The Lightning Thief",
 "features" : "This is just a test",
 "cat" : ["book", "hardcover"]
 },
 {
 "name" : "The Lightning Thief",
 "features" : "This is just a test",
 "cat" : ["book", "hardcover"]
 }
'
```

The above should dedupe the two identical documents and index only one of them.

## Processor & Post-Processor Request Parameters

We can dynamically construct a custom update request processor chain using the processor and post-processor request parameters. Multiple processors can be specified as a comma-separated value for these two parameters. For example:

### *Executing processors configured in solrconfig.xml as (pre)-processors*

```
curl
"http://localhost:8983/solr/gettingstarted/update/json?processor=remove_blanks,signature&commit=true" -H
'Content-type: application/json' -d '
[
 {
 "name" : "The Lightning Thief",
 "features" : "This is just a test",
 "cat" : ["book", "hardcover"]
 },
 {
 "name" : "The Lightning Thief",
 "features" : "This is just a test",
 "cat" : ["book", "hardcover"]
 }
'
```

## Executing processors configured in solrconfig.xml as pre- and post-processors

```
curl "http://localhost:8983/solr/gettingstarted/update/json?processor=remove_blanks&post-processor=signature&commit=true" -H 'Content-type: application/json' -d '
[
 {
 "name" : "The Lightning Thief",
 "features" : "This is just a test",
 "cat" : ["book", "hardcover"]
 },
 {
 "name" : "The Lightning Thief",
 "features" : "This is just a test",
 "cat" : ["book", "hardcover"]
 }
'
```

In the first example, Solr will dynamically create a chain which has "signature" and "remove\_blanks" as pre-processors to be executed only on the forwarding node where as in the second example, "remove\_blanks" will be executed as a pre-processor and "signature" will be executed on the leader and replicas as a post-processor.

## Configuring a Custom Chain as a Default

We can also specify a custom chain to be used by default for all requests sent to specific update handlers instead of specifying the names in request parameters for each request.

This can be done by adding either "update.chain" or "processor" and "post-processor" as default parameter for a given path which can be done either via [initParams](#) or by adding them in a ["defaults"](#) section which is supported by all request handlers.

The following is an initParam defined in the [schemaless configuration](#) which applies a custom update chain to all request handlers starting with "/update/".

### Example initParams

```
<initParams path="/update/**">
 <lst name="defaults">
 <str name="update.chain">add-unknown-fields-to-the-schema</str>
 </lst>
</initParams>
```

Alternately, one can achieve a similar effect using the "defaults" as shown in the example below:

## Example defaults

```
<requestHandler name="/update/extract" startup="lazy" class=
"solr.extraction.ExtractingRequestHandler" >
 <lst name="defaults">
 <str name="update.chain">add-unknown-fields-to-the-schema</str>
 </lst>
</requestHandler>
```

# Update Request Processor Factories

What follows are brief descriptions of the currently available update request processors. An `UpdateRequestProcessorFactory` can be integrated into an update chain in `solrconfig.xml` as necessary. You are strongly urged to examine the Javadocs for these classes; these descriptions are abridged snippets taken for the most part from the Javadocs.

## General Use UpdateProcessorFactories

### AddSchemaFieldsUpdateProcessorFactory

This processor will dynamically add fields to the schema if an input document contains one or more fields that don't match any field or dynamic field in the schema.

### AtomicUpdateProcessorFactory

This processor will convert conventional field-value documents to atomic update documents. This processor can be used at runtime (without defining it in `solrconfig.xml`), see the section [AtomicUpdateProcessorFactory](#) below.

### ClassificationUpdateProcessorFactory

This processor uses Lucene's classification module to provide simple document classification. See <https://wiki.apache.org/solr/SolrClassification> for more details on how to use this processor.

### CloneFieldUpdateProcessorFactory

Clones the values found in any matching `source` field into the configured `dest` field.

### DefaultValueUpdateProcessorFactory

A simple processor that adds a default value to any document which does not already have a value in `fieldName`.

### DocBasedVersionConstraintsProcessorFactory

This Factory generates an `UpdateProcessor` that helps to enforce version constraints on documents based on per-document version numbers using a configured name of a `versionField`.

### DocExpirationUpdateProcessorFactory

Update Processor Factory for managing automatic "expiration" of documents.

### FieldNameMutatingUpdateProcessorFactory

Modifies field names by replacing all matches to the configured pattern with the configured replacement.

## IgnoreCommitOptimizeUpdateProcessorFactory

Allows you to ignore commit and/or optimize requests from client applications when running in SolrCloud mode, for more information, see: Shards and Indexing Data in SolrCloud

## IgnoreLargeDocumentProcessorFactory

Allows you to prevent large documents with size more than `limit` (in KB) from getting indexed. It can help to prevent unexpected problems on indexing as well as on recovering because of very large documents.

## CloneFieldUpdateProcessorFactory

Clones the values found in any matching `source` field into the configured `dest` field.

## RegexpBoostProcessorFactory

A processor which will match content of "inputField" against regular expressions found in "boostFilename", and if it matches will return the corresponding boost value from the file and output this to "boostField" as a double value.

## SignatureUpdateProcessorFactory

Uses a defined set of fields to generate a hash "signature" for the document. Useful for only indexing one copy of "similar" documents.

## StatelessScriptUpdateProcessorFactory

An update request processor factory that enables the use of update processors implemented as scripts.

## TemplateUpdateProcessorFactory

Allows adding new fields to documents based on a template pattern. This update processor can also be used at runtime (without defining it in `solrconfig.xml`), see the section [TemplateUpdateProcessorFactory](#) below.

## TimestampUpdateProcessorFactory

An update processor that adds a newly generated date value of "NOW" to any document being added that does not already have a value in the specified field.

## URLClassifyProcessorFactory

Update processor which examines a URL and outputs to various other fields with characteristics of that URL, including length, number of path levels, whether it is a top level URL (`levels==0`), whether it looks like a landing/index page, a canonical representation of the URL (e.g., stripping `index.html`), the domain and path parts of the URL, etc.

## UUIDUpdateProcessorFactory

An update processor that adds a newly generated UUID value to any document being added that does not already have a value in the specified field. This processor can also be used at runtime (without defining it in `solrconfig.xml`), see the section [UUIDUpdateProcessorFactory](#) below.

## FieldMutatingUpdateProcessorFactory Derived Factories

These factories all provide functionality to *modify* fields in a document as they're being indexed. When using any of these factories, please consult the [FieldMutatingUpdateProcessorFactory javadocs](#) for details on the common options they all support for configuring which fields are modified.

## [ConcatFieldUpdateProcessorFactory](#)

Concatenates multiple values for fields matching the specified conditions using a configurable delimiter.

## [CountFieldValuesUpdateProcessorFactory](#)

Replaces any list of values for a field matching the specified conditions with the count of the number of values for that field.

## [FieldLengthUpdateProcessorFactory](#)

Replaces any CharSequence values found in fields matching the specified conditions with the lengths of those CharSequences (as an Integer).

## [FirstFieldValueUpdateProcessorFactory](#)

Keeps only the first value of fields matching the specified conditions.

## [HTMLStripFieldUpdateProcessorFactory](#)

Strips all HTML Markup in any CharSequence values found in fields matching the specified conditions.

## [IgnoreFieldUpdateProcessorFactory](#)

Ignores and removes fields matching the specified conditions from any document being added to the index.

## [LastFieldValueUpdateProcessorFactory](#)

Keeps only the last value of fields matching the specified conditions.

## [MaxFieldValueUpdateProcessorFactory](#)

An update processor that keeps only the maximum value from any selected fields where multiple values are found.

## [MinFieldValueUpdateProcessorFactory](#)

An update processor that keeps only the minimum value from any selected fields where multiple values are found.

## [ParseBooleanFieldUpdateProcessorFactory](#)

Attempts to mutate selected fields that have only CharSequence-typed values into Boolean values.

## [ParseDateFieldUpdateProcessorFactory](#)

Attempts to mutate selected fields that have only CharSequence-typed values into Date values.

## [ParseNumericFieldUpdateProcessorFactory derived classes](#)

### [ParseDoubleFieldUpdateProcessorFactory](#)

Attempts to mutate selected fields that have only CharSequence-typed values into Double values.

### [ParseFloatFieldUpdateProcessorFactory](#)

Attempts to mutate selected fields that have only CharSequence-typed values into Float values.

### [ParseIntFieldUpdateProcessorFactory](#)

Attempts to mutate selected fields that have only CharSequence-typed values into Integer values.

### [ParseLongFieldUpdateProcessorFactory](#)

Attempts to mutate selected fields that have only CharSequence-typed values into Long values.

### PreAnalyzedUpdateProcessorFactory

An update processor that parses configured fields of any document being added using *PreAnalyzedField* with the configured format parser.

### RegexReplaceProcessorFactory

An updated processor that applies a configured regex to any CharSequence values found in the selected fields, and replaces any matches with the configured replacement string.

### RemoveBlankFieldUpdateProcessorFactory

Removes any values found which are CharSequence with a length of 0 (i.e., empty strings).

### TrimFieldUpdateProcessorFactory

Trims leading and trailing whitespace from any CharSequence values found in fields matching the specified conditions.

### TruncateFieldUpdateProcessorFactory

Truncates any CharSequence values found in fields matching the specified conditions to a maximum character length.

### UniqFieldsUpdateProcessorFactory

Removes duplicate values found in fields matching the specified conditions.

## Update Processor Factories That Can Be Loaded as Plugins

These processors are included in Solr releases as "contribs", and require additional jars loaded at runtime. See the README files associated with each contrib for details:

### The langid contrib provides

#### LangDetectLanguageIdentifierUpdateProcessorFactory

Identifies the language of a set of input fields using <http://code.google.com/p/language-detection>.

#### TikaLanguageIdentifierUpdateProcessorFactory

Identifies the language of a set of input fields using Tika's LanguageIdentifier.

### The analysis-extras contrib provides

#### OpenNLPExtractNamedEntitiesUpdateProcessorFactory

Update document(s) to be indexed with named entities extracted using an OpenNLP NER model. Note that in order to use model files larger than 1MB on SolrCloud, you must either [configure both ZooKeeper server and clients](#) or [store the model files on the filesystem](#) on each node hosting a collection replica.

## Update Processor Factories You Should *Not* Modify or Remove

These are listed for completeness, but are part of the Solr infrastructure, particularly SolrCloud. Other than insuring you do *not* remove them when modifying the update request handlers (or any copies you make), you will rarely, if ever, need to change these.

## DistributedUpdateProcessorFactory

Used to distribute updates to all necessary nodes.

## NoOpDistributingUpdateProcessorFactory

An alternative No-Op implementation of DistributingUpdateProcessorFactory that always returns null. Designed for experts who want to bypass distributed updates and use their own custom update logic.

## LogUpdateProcessorFactory

A logging processor. This keeps track of all commands that have passed through the chain and prints them on finish().

## RunUpdateProcessorFactory

Executes the update commands using the underlying UpdateHandler. Almost all processor chains should end with an instance of RunUpdateProcessorFactory unless the user is explicitly executing the update commands in an alternative custom UpdateRequestProcessorFactory.

## Update Processors That Can Be Used at Runtime

These Update processors do not need any configuration in solrconfig.xml. They are automatically initialized when their name is added to the processor parameter sent with an update request. Multiple processors can be used by appending multiple processor names separated by commas.

### AtomicUpdateProcessorFactory

The AtomicUpdateProcessorFactory is used to atomically update documents.

Use the parameter processor=atomic to invoke it. Use it to convert your normal update operations to atomic update operations. This is particularly useful when you use endpoints such as /update/csv or /update/json/docs which does not otherwise have a syntax for atomic operations.

For example:

```
processor=atomic&atomic.field1=add&atomic.field2=set&atomic.field3=inc&atomic.field4=remove&atomic.field4=remove
```

The above parameters convert a normal update operation in the following ways:

- field1 to an atomic add operation
- field2 to an atomic set operation
- field3 to an atomic inc operation
- field4 to an atomic remove operation

### TemplateUpdateProcessorFactory

The TemplateUpdateProcessorFactory can be used to add new fields to documents based on a template pattern.

Use the parameter processor=template to use it. The template parameter template.field (multivalued)

defines the field to add and the pattern. Templates may contain placeholders which refer to other fields in the document. You can have multiple `Template.field` parameters in a single request.

For example:

```
processor=template&template.field=fullName:Mr. {firstName} {lastName}
```

The above example would add a new field to the document called `fullName`. The fields `firstName` and `lastName` are supplied from the document fields. If either of them is missing, that part is replaced with an empty string. If those fields are multi-valued, only the first value is used.

### UUIDUpdateProcessorFactory

The `UUIDUpdateProcessorFactory` is used to add generated UUIDs to documents.

Use the parameter `processor=uuid` to invoke it. You will also need to specify the field where the UUID will be added with the `uuid.fieldName` parameter.

For example:

```
processor=uuid&uuid.fieldName=somefield_name
```

## Codec Factory

A `codecFactory` can be specified in `solrconfig.xml` to determine which Lucene Codec is used when writing the index to disk.

If not specified, Lucene's default codec is implicitly used.

### Alternatives to the Default Codec

There are two alternatives to Lucene's default codec.

#### solr.SchemaCodecFactory

The `solr.SchemaCodecFactory` supports 2 key features:

- Schema based per-fieldtype configuration for `docValuesFormat` and `postingsFormat` - see the [Field Type Properties](#) section for more details.
- A `compressionMode` option:
  - `BEST_SPEED` (default) is optimized for search speed performance
  - `BEST_COMPRESSION` is optimized for disk space usage

Example:

```
<codecFactory class="solr.SchemaCodecFactory">
 <str name="compressionMode">BEST_COMPRESSION</str>
</codecFactory>
```

## solr.SimpleTextCodecFactory

This factory for Lucene's SimpleTextCodecFactory produces a plain text human-readable index format.



**FOR RECREATIONAL USE ONLY.** This codec should never be used in production.  
SimpleTextCodec is relatively slow and takes up a large amount of disk space. Its use  
should be limited to educational and debugging purposes.

Example:

```
<codecFactory class="solr.SimpleTextCodecFactory"/>
```

# Solr Cores and solr.xml

In Solr, the term *core* is used to refer to a single index and associated transaction log and configuration files (including the `solrconfig.xml` and Schema files, among others). Your Solr installation can have multiple cores if needed, which allows you to index data with different structures in the same server, and maintain more control over how your data is presented to different audiences. In SolrCloud mode you will be more familiar with the term *collection*. Behind the scenes a collection consists of one or more cores.

Cores can be created using `bin/solr` script or as part of SolrCloud collection creation using the APIs. Core-specific properties (such as the directories to use for the indexes or configuration files, the core name, and other options) are defined in a `core.properties` file. Any `core.properties` file in any directory of your Solr installation (or in a directory under where `solr_home` is defined) will be found by Solr and the defined properties will be used for the core named in the file.

In standalone mode, `solr.xml` must reside in `solr_home`. In SolrCloud mode, `solr.xml` will be loaded from ZooKeeper if it exists, with fallback to `solr_home`.



In older versions of Solr, cores had to be predefined as `<core>` tags in `solr.xml` in order for Solr to know about them. Now, however, Solr supports automatic discovery of cores and they no longer need to be explicitly defined. The recommended way is to dynamically create cores/collections using the APIs.

The following sections describe these options in more detail.

- [Format of solr.xml](#): Details on how to define `solr.xml`, including the acceptable parameters for the `solr.xml` file
- [Defining core.properties](#): Details on placement of `core.properties` and available property options.
- [CoreAdmin API](#): Tools and commands for core administration using a REST API.
- [Config Sets](#): How to use configsets to avoid duplicating effort when defining a new core.

## Format of solr.xml

The `solr.xml` file defines some global configuration options that apply to all or many cores.

This section will describe the default `solr.xml` file included with Solr and how to modify it for your needs. For details on how to configure `core.properties`, see the section [Defining core.properties](#).

## Defining solr.xml

You can find `solr.xml` in your `$SOLR_HOME` directory (usually `server/solr`) in standalone mode or in ZooKeeper when using SolrCloud. The default `solr.xml` file looks like this:

```
<solr>

 <int name="maxBooleanClauses">${solr.max.booleanClauses:1024}</int>

 <solrcloud>
 <str name="host">${host:}</str>
 <int name="hostPort">${jetty.port:8983}</int>
 <str name="hostContext">${hostContext:solr}</str>
 <int name="zkClientTimeout">${zkClientTimeout:15000}</int>
 <bool name="genericCoreNodeNames">${genericCoreNodeNames:true}</bool>
 </solrcloud>

 <shardHandlerFactory name="shardHandlerFactory"
 class="HttpShardHandlerFactory">
 <int name="socketTimeout">${socketTimeout:0}</int>
 <int name="connTimeout">${connTimeout:0}</int>
 </shardHandlerFactory>

</solr>
```

As you can see, the discovery Solr configuration is "SolrCloud friendly". However, the presence of the `<solrcloud>` element does *not* mean that the Solr instance is running in SolrCloud mode. Unless the `-DzkHost` or `-DzkRun` are specified at startup time, this section is ignored.

## Solr.xml Parameters

### The `<solr>` Element

There are no attributes that you can specify in the `<solr>` tag, which is the root element of `solr.xml`. The tables below list the child nodes of each XML element in `solr.xml`.

#### adminHandler

This attribute does not need to be set.

If used, this attribute should be set to the FQN (Fully qualified name) of a class that inherits from `CoreAdminHandler`. For example, `<str name="adminHandler">com.myorg.MyAdminHandler</str>` would configure the custom admin handler (`MyAdminHandler`) to handle admin requests.

If this attribute isn't set, Solr uses the default admin handler, `org.apache.solr.handler.admin.CoreAdminHandler`.

#### collectionsHandler

As above, for custom CollectionsHandler implementations.

#### infoHandler

As above, for custom InfoHandler implementations.

#### coreLoadThreads

Specifies the number of threads that will be assigned to load cores in parallel.

**replayUpdatesThreads**

Specifies the number of threads that will be assigned to replay updates in parallel. This pool is shared for all cores of the node. The default value is equal to the number of processors.

**coreRootDirectory**

The root of the core discovery tree, defaults to \$SOLR\_HOME (by default, server/solr).

**managementPath**

Currently non-operational.

**sharedLib**

Specifies the path to a common library directory that will be shared across all cores. Any JAR files in this directory will be added to the search path for Solr plugins. This path is relative to \$SOLR\_HOME. Custom handlers may be placed in this directory.

**shareSchema**

This attribute, when set to true, ensures that the multiple cores pointing to the same Schema resource file will be referring to the same IndexSchema Object. Sharing the IndexSchema Object makes loading the core faster. If you use this feature, make sure that no core-specific property is used in your Schema file.

**transientCacheSize**

Defines how many cores with transient=true that can be loaded before swapping the least recently used core for a new core.

**configSetBaseDir**

The directory under which configsets for Solr cores can be found. Defaults to \$SOLR\_HOME/configsets.

**maxBooleanClauses**

Sets the maximum number of clauses allowed in any boolean query.

This global limit provides a safety constraint on the number of clauses allowed in any boolean queries against any collection — regardless of whether those clauses were explicitly specified in a query string, or were the result of query expansion/re-writing from a more complex type of query based on the terms in the index.

In default configurations this property uses the value of the solr.max.booleanClauses system property if specified. This is the same system property used in the default configset for the `<maxBooleanClauses>` setting of solrconfig.xml making it easy for Solr administrators to increase both values (in all collections) without needing to search through and update all of their configs.

```
<maxBooleanClauses>${solr.max.booleanClauses:1024}</maxBooleanClauses>
```

## The `<solrcloud>` Element

This element defines several parameters that relate to SolrCloud. This section is ignored unless theSolr instance is started with either -DzkRun or -DzkHost

**distribUpdateConnTimeout**

Used to set the underlying connTimeout for intra-cluster updates.

distribUpdateSoTimeout

Used to set the underlying socketTimeout for intra-cluster updates.

host

The hostname Solr uses to access cores.

hostContext

The url context path.

hostPort

The port Solr uses to access cores.

In the default solr.xml file, this is set to \${jetty.port:8983}, which will use the Solr port defined in Jetty, and otherwise fall back to 8983.

leaderVoteWait

When SolrCloud is starting up, how long each Solr node will wait for all known replicas for that shard to be found before assuming that any nodes that haven't reported are down.

leaderConflictResolveWait

When trying to elect a leader for a shard, this property sets the maximum time a replica will wait to see conflicting state information to be resolved; temporary conflicts in state information can occur when doing rolling restarts, especially when the node hosting the Overseer is restarted.

Typically, the default value of 180000 (ms) is sufficient for conflicts to be resolved; you may need to increase this value if you have hundreds or thousands of small collections in SolrCloud.

zkClientTimeout

A timeout for connection to a ZooKeeper server. It is used with SolrCloud.

zkHost

In SolrCloud mode, the URL of the ZooKeeper host that Solr should use for cluster state information.

genericCoreNodeNames

If TRUE, node names are not based on the address of the node, but on a generic name that identifies the core. When a different machine takes over serving that core things will be much easier to understand.

zkCredentialsProvider & zkACLProvider

Optional parameters that can be specified if you are using [ZooKeeper Access Control](#).

## The <logging> Element

class

The class to use for logging. The corresponding JAR file must be available to Solr, perhaps through a <lib> directive in solrconfig.xml.

enabled

true/false - whether to enable logging or not.

## The <logging><watcher> Element

size

The number of log events that are buffered.

#### threshold

The logging level above which your particular logging implementation will record. For example when using log4j one might specify DEBUG, WARN, INFO, etc.

## The <shardHandlerFactory> Element

Custom shard handlers can be defined in solr.xml if you wish to create a custom shard handler.

```
<shardHandlerFactory name="ShardHandlerFactory" class="qualified.class.name">
```

Since this is a custom shard handler, sub-elements are specific to the implementation. The default and only shard handler provided by Solr is the HttpShardHandlerFactory in which case, the following sub-elements can be specified:

#### socketTimeout

The read timeout for intra-cluster query and administrative requests. The default is the same as the distribUpdateSoTimeout specified in the <solrcloud> section.

#### connTimeout

The connection timeout for intra-cluster query and administrative requests. Defaults to the distribUpdateConnTimeout specified in the <solrcloud> section.

#### urlScheme

The URL scheme to be used in distributed search.

#### maxConnectionsPerHost

Maximum connections allowed per host. Defaults to 100000.

#### corePoolSize

The initial core size of the threadpool servicing requests. Default is 0.

#### maximumPoolSize

The maximum size of the threadpool servicing requests. Default is unlimited.

#### maxThreadIdleTime

The amount of time in seconds that idle threads persist for in the queue, before being killed. Default is 5 seconds.

#### sizeOfQueue

If the threadpool uses a backing queue, what is its maximum size to use direct handoff. Default is to use a SynchronousQueue.

#### fairnessPolicy

A boolean to configure if the threadpool favors fairness over throughput. Default is false to favor throughput.

## The <metrics> Element

The <metrics> element in solr.xml allows you to customize the metrics reported by Solr. You can define

system properties that should not be returned, or define custom suppliers and reporters.

In a default solr.xml you will not see any <metrics> configuration. If you would like to customize the metrics for your installation, see the section [Metrics Configuration](#).

## Substituting JVM System Properties in solr.xml

Solr supports variable substitution of JVM system property values in solr.xml, which allows runtime specification of various configuration options. The syntax is \${propertyname[:option default value]} . This allows defining a default that can be overridden when Solr is launched. If a default value is not specified, then the property must be specified at runtime or the solr.xml file will generate an error when parsed.

Any JVM system properties usually specified using the -D flag when starting the JVM, can be used as variables in the solr.xml file.

For example, in the solr.xml file shown below, the socketTimeout and connTimeout values are each set to "0". However, if you start Solr using bin/solr -DsocketTimeout=1000, the socketTimeout option of the HttpShardHandlerFactory to be overridden using a value of 1000ms, while the connTimeout option will continue to use the default property value of "0".

```
<solr>
 <shardHandlerFactory name="shardHandlerFactory"
 class="HttpShardHandlerFactory">
 <int name="socketTimeout">${socketTimeout:0}</int>
 <int name="connTimeout">${connTimeout:0}</int>
 </shardHandlerFactory>
</solr>
```

## Defining core.properties

Core discovery means that creating a core is as simple as a core.properties file located on disk.

The core.properties file is a simple Java Properties file where each line is just a key=value pair, e.g., name=core1. Notice that no quotes are required.

A minimal core.properties file looks like the example below. However, it can also be empty, see information on placement of core.properties below.

```
name=my_core_name
```

## Placement of core.properties

Solr cores are configured by placing a file named core.properties in a sub-directory under solr.home. There are no a-priori limits to the depth of the tree, nor are there limits to the number of cores that can be defined. Cores may be anywhere in the tree with the exception that cores may *not* be defined under an existing core. That is, the following is not allowed:

```
./cores/core1/core.properties
./cores/core1/coremore/core5/core.properties
```

In this example, the enumeration will stop at "core1".

The following is legal:

```
./cores/somecores/core1/core.properties
./cores/somecores/core2/core.properties
./cores/othercores/core3/core.properties
./cores/extracores/deeptree/core4/core.properties
```

It is possible to segment Solr into multiple cores, each with its own configuration and indices. Cores may be dedicated to a single application or to very different ones, but all are administered through a common administration interface. You can create new Solr cores on the fly, shutdown cores, even replace one running core with another, all without ever stopping or restarting Solr.

Your core.properties file can be empty if necessary. Suppose core.properties is located in ./cores/core1 (relative to solr\_home) but is empty. In this case, the core name is assumed to be "core1". The instanceDir will be the folder containing core.properties (i.e., ./cores/core1). The dataDir will be ../cores/core1/data, etc.



You can run Solr without configuring any cores.

## Defining core.properties Files

The minimal core.properties file is an empty file, in which case all of the properties are defaulted appropriately.

Java properties files allow the hash (#) or bang (!) characters to specify comment-to-end-of-line.

The following properties are available:

name

The name of the SolrCore. You'll use this name to reference the SolrCore when running commands with the CoreAdminHandler.

config

The configuration file name for a given core. The default is solrconfig.xml.

schema

The schema file name for a given core. The default is schema.xml but please note that if you are using a "managed schema" (the default behavior) then any value for this property which does not match the effective managedSchemaResourceName will be read once, backed up, and converted for managed schema use. See [Schema Factory Definition in SolrConfig](#) for more details.

dataDir

The core's data directory (where indexes are stored) as either an absolute pathname, or a path relative to the value of instanceDir. This is data by default.

#### configSet

The name of a defined configset, if desired, to use to configure the core (see the section [Config Sets](#) for more details).

#### properties

The name of the properties file for this core. The value can be an absolute pathname or a path relative to the value of `instanceDir`.

#### transient

If **true**, the core can be unloaded if Solr reaches the `transientCacheSize`. The default if not specified is **false**. Cores are unloaded in order of least recently used first. *Setting this to **true** is not recommended in SolrCloud mode.*

#### loadOnStartup

If **true**, the default if it is not specified, the core will load when Solr starts. *Setting this to **false** is not recommended in SolrCloud mode.*

#### coreNodeName

Used only in SolrCloud, this is a unique identifier for the node hosting this replica. By default a `coreNodeName` is generated automatically, but setting this attribute explicitly allows you to manually assign a new core to replace an existing replica. For example, this can be useful when replacing a machine that has had a hardware failure by restoring from backups on a new machine with a new hostname or port.

#### ulogDir

The absolute or relative directory for the update log for this core (SolrCloud).

#### shard

The shard to assign this core to (SolrCloud).

#### collection

The name of the collection this core is part of (SolrCloud).

#### roles

Future parameter for SolrCloud or a way for users to mark nodes for their own use.

Additional user-defined properties may be specified for use as variables. For more information on how to define local properties, see the section [Substituting Properties in Solr Config Files](#).

## CoreAdmin API

The Core Admin API is primarily used under the covers by the [Collections API](#) when running a [SolrCloud](#) cluster.

SolrCloud users should not typically use the CoreAdmin API directly, but the API may be useful for users of single-node or master/slave Solr installations for core maintenance operations.

The CoreAdmin API is implemented by the `CoreAdminHandler`, which is a special purpose [request handler](#) that is used to manage Solr cores. Unlike other request handlers, the `CoreAdminHandler` is not attached to a single core. Instead, there is a single instance of the `CoreAdminHandler` in each Solr node that manages all the cores running in that node and is accessible at the `/solr/admin/cores` path.

CoreAdmin actions can be executed by via HTTP requests that specify an action request parameter, with additional action specific arguments provided as additional parameters.

All action names are uppercase, and are defined in depth in the sections below.

## STATUS

The STATUS action returns the status of all running Solr cores, or status for only the named core.

```
admin/cores?action=STATUS&core=core-name
```

### STATUS Parameters

core

The name of a core, as listed in the "name" attribute of a <core> element in solr.xml.

indexInfo

If false, information about the index will not be returned with a core STATUS request. In Solr implementations with a large number of cores (i.e., more than hundreds), retrieving the index information for each core can take a lot of time and isn't always required. The default is true.

## CREATE

The CREATE action creates a new core and registers it.

If a Solr core with the given name already exists, it will continue to handle requests while the new core is initializing. When the new core is ready, it will take new requests and the old core will be unloaded.

```
admin/cores?action=CREATE&name=core-name&instanceDir= path/to/dir&config=solrconfig.xml&dataDir=data
```

Note that this command is the only one of the Core Admin API commands that **does not** support the core parameter. Instead, the name parameter is required, as shown below.

### *CREATE must be able to find a configuration!*

Your CREATE call must be able to find a configuration, or it will not succeed.

When you are running SolrCloud and create a new core for a collection, the configuration will be inherited from the collection. Each collection is linked to a configName, which is stored in ZooKeeper. This satisfies the configuration requirement. There is something to note, though: if you're running SolrCloud, you should **NOT** use the CoreAdmin API at all. Use the [Collections API](#).



When you are not running SolrCloud, if you have [Config Sets](#) defined, you can use the configSet parameter as documented below. If there are no configsets, then the instanceDir specified in the CREATE call must already exist, and it must contain a conf directory which in turn must contain solrconfig.xml, your schema (usually named either managed-schema or schema.xml), and any files referenced by those configs.

The config and schema filenames can be specified with the config and schema parameters, but these are expert options. One thing you could do to avoid creating the conf directory is use config and schema parameters that point at absolute paths, but this can lead to confusing configurations unless you fully understand what you are doing.

### *CREATE and the core.properties file*



The core.properties file is built as part of the CREATE command. If you create a core.properties file yourself in a core directory and then try to use CREATE to add that core to Solr, you will get an error telling you that another core is already defined there. The core.properties file must NOT exist before calling the CoreAdmin API with the CREATE command.

## CREATE Core Parameters

### name

The name of the new core. Same as name on the <core> element. This parameter is required.

### instanceDir

The directory where files for this core should be stored. Same as instanceDir on the <core> element.

The default is the value specified for the name parameter if not supplied.

### config

Name of the config file (i.e., solrconfig.xml) relative to instanceDir.

### schema

Name of the schema file to use for the core. Please note that if you are using a "managed schema" (the default behavior) then any value for this property which does not match the effective managedSchemaResourceName will be read once, backed up, and converted for managed schema use. See [Schema Factory Definition in SolrConfig](#) for details.

### dataDir

Name of the data directory relative to instanceDir.

### configSet

Name of the configset to use for this core. For more information, see the section [Config Sets](#).

#### collection

The name of the collection to which this core belongs. The default is the name of the core.

`collection.param=value` causes a property of `param=value` to be set if a new collection is being created.

Use `collection.configName=config-name` to point to the configuration for a new collection.



While it's possible to create a core for a non-existent collection, this approach is not supported and not recommended. Always create a collection using the [Collections API](#) before creating a core directly for it.

#### shard

The shard id this core represents. Normally you want to be auto-assigned a shard id.

#### property.name=value

Sets the core property `name` to `value`. See the section on defining [core.properties file contents](#).

#### async

Request ID to track this action which will be processed asynchronously.

Use `collection.configName=configname` to point to the config for a new collection.

### CREATE Example

```
http://localhost:8983/solr/admin/cores?action=CREATE&name=my_core&collection=my_collection&shard=shard2
```

### RELOAD

The RELOAD action loads a new core from the configuration of an existing, registered Solr core. While the new core is initializing, the existing one will continue to handle requests. When the new Solr core is ready, it takes over and the old core is unloaded.

#### admin/cores?action=RELOAD&core=core-name

This is useful when you've made changes to a Solr core's configuration on disk, such as adding new field definitions. Calling the RELOAD action lets you apply the new configuration without having to restart Solr.



RELOAD performs "live" reloads of SolrCore, reusing some existing objects. Some configuration options, such as the `dataDir` location and `IndexWriter`-related settings in `solrconfig.xml` can not be changed and made active with a simple RELOAD action.

### RELOAD Core Parameters

#### core

The name of the core, as listed in the "name" attribute of a `<core>` element in `solr.xml`. This parameter is required.

## RENAME

The RENAME action changes the name of a Solr core.

```
admin/cores?action=RENAME&core=core-name&other=other-core-name
```

### RENAME Parameters

core

The name of the Solr core to be renamed. This parameter is required.

other

The new name for the Solr core. If the persistent attribute of <solr> is true, the new name will be written to solr.xml as the name attribute of the <core> attribute. This parameter is required.

async

Request ID to track this action which will be processed asynchronously.

## SWAP

SWAP atomically swaps the names used to access two existing Solr cores. This can be used to swap new content into production. The prior core remains available and can be swapped back, if necessary. Each core will be known by the name of the other, after the swap.

```
admin/cores?action=SWAP&core=core-name&other=other-core-name
```



Do not use SWAP with a SolrCloud node. It is not supported and can result in the core being unusable.

### SWAP Parameters

core

The name of one of the cores to be swapped. This parameter is required.

other

The name of one of the cores to be swapped. This parameter is required.

async

Request ID to track this action which will be processed asynchronously.

## UNLOAD

The UNLOAD action removes a core from Solr. Active requests will continue to be processed, but no new requests will be sent to the named core. If a core is registered under more than one name, only the given name is removed.

```
admin/cores?action=UNLOAD&core=core-name
```

The UNLOAD action requires a parameter (core) identifying the core to be removed. If the persistent attribute of <solr> is set to true, the <core> element with this name attribute will be removed from solr.xml.



Unloading all cores in a SolrCloud collection causes the removal of that collection's metadata from ZooKeeper.

## UNLOAD Parameters

### core

The name of a core to be removed. This parameter is required.

### deleteIndex

If true, will remove the index when unloading the core. The default is false.

### deleteDataDir

If true, removes the data directory and all sub-directories. The default is false.

### deleteInstanceDir

If true, removes everything related to the core, including the index directory, configuration files and other related files. The default is false.

### async

Request ID to track this action which will be processed asynchronously.

## MERGEINDEXES

The MERGEINDEXES action merges one or more indexes to another index. The indexes must have completed commits, and should be locked against writes until the merge is complete or the resulting merged index may become corrupted. The target core index must already exist and have a compatible schema with the one or more indexes that will be merged to it. Another commit on the target core should also be performed after the merge is complete.

```
admin/cores?action=MERGEINDEXES&core=new-core-name&indexDir= path/to/core1/data/index&indexDir=path/to/core2/data/index
```

In this example, we use the `indexDir` parameter to define the index locations of the source cores. The `core` parameter defines the target index. A benefit of this approach is that we can merge any Lucene-based index that may not be associated with a Solr core.

Alternatively, we can instead use a `srcCore` parameter, as in this example:

```
admin/cores?action=mergeindexes&core=new-core-name&srcCore=core1-name&srcCore=core2-name
```

This approach allows us to define cores that may not have an index path that is on the same physical server as the target core. However, we can only use Solr cores as the source indexes. Another benefit of this approach is that we don't have as high a risk for corruption if writes occur in parallel with the source index.

We can make this call run asynchronously by specifying the `async` parameter and passing a request-id. This id can then be used to check the status of the already submitted task using the REQUESTSTATUS API.

## MERGEINDEXES Parameters

### core

The name of the target core/index. This parameter is required.

indexDir

Multi-valued, directories that would be merged.

srcCore

Multi-valued, source cores that would be merged.

async

Request ID to track this action which will be processed asynchronously.

## SPLIT

The SPLIT action splits an index into two or more indexes. The index being split can continue to handle requests. The split pieces can be placed into a specified directory on the server's filesystem or it can be merged into running Solr cores.

The SPLIT action supports five parameters, which are described in the table below.

### SPLIT Parameters

core

The name of the core to be split. This parameter is required.

path

Multi-valued, the directory path in which a piece of the index will be written. Either this parameter or targetCore must be specified. If this is specified, the targetCore parameter may not be used.

targetCore

Multi-valued, the target Solr core to which a piece of the index will be merged. Either this parameter or path must be specified. If this is specified, the path parameter may not be used.

ranges

A comma-separated list of hash ranges in hexadecimal format. If this parameter is used, split.key should not be. See the [SPLIT Examples](#) below for an example of how this parameter can be used.

split.key

The key to be used for splitting the index. If this parameter is used, ranges should not be. See the [SPLIT Examples](#) below for an example of how this parameter can be used.

async

Request ID to track this action which will be processed asynchronously.

### SPLIT Examples

The core index will be split into as many pieces as the number of path or targetCore parameters.

#### Usage with two targetCore parameters:

```
http://localhost:8983/solr/admin/cores?action=SPLIT&core=core0&targetCore=core1&targetCore=core2
```

Here the core index will be split into two pieces and merged into the two targetCore indexes.

## Usage with two path parameters:

```
http://localhost:8983/solr/admin/cores?action=SPLIT&core=core0&path=/path/to/index/1&path=/path/to/index/2
```

The core index will be split into two pieces and written into the two directory paths specified.

## Usage with the split.key parameter:

```
http://localhost:8983/solr/admin/cores?action=SPLIT&core=core0&targetCore=core1&split.key=A!
```

Here all documents having the same route key as the split.key i.e., 'A!' will be split from the core index and written to the targetCore.

## Usage with ranges parameter:

```
http://localhost:8983/solr/admin/cores?action=SPLIT&core=core0&targetCore=core1&targetCore=core2&targetCore=core3&ranges=0-1f4,1f5-3e8,3e9-5dc
```

This example uses the ranges parameter with hash ranges 0-500, 501-1000 and 1001-1500 specified in hexadecimal. Here the index will be split into three pieces with each targetCore receiving documents matching the hash ranges specified i.e., core1 will get documents with hash range 0-500, core2 will receive documents with hash range 501-1000 and finally, core3 will receive documents with hash range 1001-1500. At least one hash range must be specified. Please note that using a single hash range equal to a route key's hash range is NOT equivalent to using the split.key parameter because multiple route keys can hash to the same range.

The targetCore must already exist and must have a compatible schema with the core index. A commit is automatically called on the core index before it is split.

This command is used as part of the [SPLITSHARD](#) command but it can be used for non-cloud Solr cores as well. When used against a non-cloud core without split.key parameter, this action will split the source index and distribute its documents alternately so that each split piece contains an equal number of documents. If the split.key parameter is specified then only documents having the same route key will be split from the source index.

## REQUESTSTATUS

Request the status of an already submitted asynchronous CoreAdmin API call.

```
admin/cores?action=REQUESTSTATUS&requestid=id
```

### Core REQUESTSTATUS Parameters

The REQUESTSTATUS command has only one parameter.

requestid

The user defined request-id for the asynchronous request. This parameter is required.

The call below will return the status of an already submitted asynchronous CoreAdmin call.

```
http://localhost:8983/solr/admin/cores?action=REQUESTSTATUS&requestid=1
```

## REQUESTRECOVERY

The REQUESTRECOVERY action manually asks a core to recover by synching with the leader. This should be considered an "expert" level command and should be used in situations where the node (SolrCloud replica) is unable to become active automatically.

```
admin/cores?action=REQUESTRECOVERY&core=core-name
```

### REQUESTRECOVERY Parameters

core

The name of the core to re-sync. This parameter is required.

### REQUESTRECOVERY Examples

```
http://localhost:8981/solr/admin/cores?action=REQUESTRECOVERY&core=gettingstarted_shard1_replica1
```

The core to specify can be found by expanding the appropriate ZooKeeper node via the admin UI.

## Config Sets

On a multicore Solr instance, you may find that you want to share configuration between a number of different cores. You can achieve this using named configsets, which are essentially shared configuration directories stored under a configurable configset base directory.

Configsets are made up of the configuration files used in a Solr installation: including `solrconfig.xml`, the schema, language-files, `synonyms.txt`, DIH-related configuration, and others as needed for your implementation.

Solr ships with two example configsets located in `server/solr/configsets`, which can be used as a base for your own. These example configsets are named `_default` and `sample_techproducts_configs`.

### Configsets in Standalone Mode

If you are using Solr in standalone mode, configsets are created on the filesystem.

To create a configset, add a new directory under the configset base directory. The configset will be identified by the name of this directory. Then into this copy the configuration directory you want to share. The structure should look something like this:

```
<configSetBaseDir>
 /configset1
 /conf
 /managed-schema
 /solrconfig.xml
 /configset2
 /conf
 /managed-schema
 /solrconfig.xml
```

The default base directory is \$SOLR\_HOME/configsets. This path can be configured in solr.xml (see [Format of solr.xml](#) for details).

To create a new core using a configset, pass configSet as one of the core properties. For example, if you do this via the CoreAdmin API:

### V1 API

```
curl
http://localhost:8983/admin/cores?action=CREATE&name=mycore&instanceDir=path/to/instance&configSet=configset2
```

### V2 API

```
curl -v -X POST -H 'Content-type: application/json' -d '{
 "create": [
 {
 "name": "mycore",
 "instanceDir": "path/to/instance",
 "configSet": "configSet2"
 }
]
}' http://localhost:8983/api/cores
```

## Configsets in SolrCloud Mode

In SolrCloud mode, you can use the [Configsets API](#) to manage your configsets.

# Configuration APIs

Solr includes several APIs that can be used to modify settings in `solrconfig.xml`.

- [Blob Store API](#)
- [Config API](#)
- [Request Parameters API](#)
- [Managed Resources](#)

## Blob Store API

The Blob Store REST API provides REST methods to store, retrieve or list files in a Lucene index.

It can be used to upload a jar file which contains standard Solr components such as RequestHandlers, SearchComponents, or other custom code you have written for Solr. Schema components *do not* yet support the Blob Store.

When using the blob store, note that the API does not delete or overwrite a previous object if a new one is uploaded with the same name. It always adds a new version of the blob to the index. Because the `.system` collection is a standard Solr collection, deleting blobs is the same as deleting documents.

**The blob store is only available when running in SolrCloud mode.** Solr in standalone mode does not support use of a blob store.

The blob store API is implemented as a requestHandler. A special collection named "`.system`" is used to store the blobs. This collection can be created in advance, but if it does not exist it will be created automatically.

## About the `.system` Collection

Before uploading blobs to the blob store, a special collection must be created and it must be named `.system`. Solr will automatically create this collection if it does not already exist, but you can also create it manually if you choose.

The BlobHandler is automatically registered in the `.system` collection. The `solrconfig.xml`, Schema, and other configuration files for the collection are automatically provided by the system and don't need to be defined specifically.

If you do not use the `-shards` or `-replicationFactor` options, then defaults of `numShards=1` and `replicationFactor=3` (or maximum nodes in the cluster) will be used.

You can create the `.system` collection with the [CREATE command](#) of the Collections API, as in this example:

## V1 API

```
curl
http://localhost:8983/solr/admin/collections?action=CREATE&name=.system&replicationFactor=2&n
umShards=2
```

Note that this example will create the .system collection across 2 shards with a replication factor of 2; you may need to customize this for your Solr implementation.

## V2 API

```
curl -X POST -H 'Content-type: application/json' -d '{"create": {"name": ".system",
"numShards": "2", "replicationFactor": "2"}}' http://localhost:8983/api/collections
```

Note that this example will create the .system collection across 2 shards with a replication factor of 2; you may need to customize this for your Solr implementation.



The bin/solr script cannot be used to create the .system collection.

## Upload Files to Blob Store

After the .system collection has been created, files can be uploaded to the blob store with a request similar to the following:

### V1 API

```
curl -X POST -H 'Content-Type: application/octet-stream' --data-binary @{filename}
http://localhost:8983/solr/.system/blob/{blobname}
```

For example, to upload a file named "test1.jar" as a blob named "test", you would make a POST request like:

```
curl -X POST -H 'Content-Type: application/octet-stream' --data-binary @test1.jar
http://localhost:8983/solr/.system/blob/test
```

## V2 API

```
curl -X POST -H 'Content-Type: application/octet-stream' --data-binary @{filename}
http://localhost:8983/api/collections/.system/blob/{blobname}
```

For example, to upload a file named "test1.jar" as a blob named "test", you would make a POST request like:

```
curl -X POST -H 'Content-Type: application/octet-stream' --data-binary @test1.jar
http://localhost:8983/api/collections/.system/blob/test
```

Note that by default, the blob store has a limit of 5Mb for any blob. This can be increased if necessary by changing the value for the `maxSize` setting in `solrconfig.xml` for the `.system` collection. See the section [Configuring solrconfig.xml](#) for information about how to modify `solrconfig.xml` for any collection.

A GET request will return the list of blobs and other details:

## V1 API

For all blobs:

```
curl http://localhost:8983/solr/.system/blob?omitHeader=true
```

For a single blob:

```
curl http://localhost:8983/solr/.system/blob/test?omitHeader=true
```

Output:

```
{
 "response": { "numFound": 1, "start": 0, "docs": [
 {
 "id": "test/1",
 "md5": "20ff915fa3f5a5d66216081ae705c41b",
 "blobName": "test",
 "version": 1,
 "timestamp": "2015-02-04T16:45:48.374Z",
 "size": 13108
 }
]
}
```

## V2 API

For all blobs:

```
curl http://localhost:8983/api/collections/.system/blob?omitHeader=true
```

For a single blob:

```
curl http://localhost:8983/api/collections/.system/blob/test?omitHeader=true
```

Output:

```
{
 "response":{ "numFound":1, "start":0, "docs":[
 {
 "id":"test/1",
 "md5":"20ff915fa3f5a5d66216081ae705c41b",
 "blobName":"test",
 "version":1,
 "timestamp":"2015-02-04T16:45:48.374Z",
 "size":13108}
 }
]}
```

The filestream response writer can retrieve a blob for download, as in:

## V1 API

For a specific version of a blob, include the version to the request:

```
curl http://localhost:8983/solr/.system/blob/{blobname}/{version}?wt=filestream >
{outputfilename}
```

For the latest version of a blob, the {version} can be omitted:

```
curl http://localhost:8983/solr/.system/blob/{blobname}?wt=filestream > {outputfilename}
```

**V2 API** For a specific version of a blob, include the version to the request:

```
curl http://localhost:8983/api/collections/.system/blob/{blobname}/{version}?wt=filestream > {outputfilename}
```

For the latest version of a blob, the {version} can be omitted:

```
curl http://localhost:8983/api/collections/.system/blob/{blobname}?wt=filestream > {outputfilename}
```

## Use a Blob in a Handler or Component

To use the blob as the class for a request handler or search component, you create a request handler in solrconfig.xml as usual. You will need to define the following parameters:

**class**

the fully qualified class name. For example, if you created a new request handler class called CRUDHandler, you would enter org.apache.solr.core.CRUDHandler.

**runtimeLib**

Set to true to require that this component should be loaded from the classloader that loads the runtime jars.

For example, to use a blob named test, you would configure solrconfig.xml like this:

```
<requestHandler name="/myhandler" class="org.apache.solr.core.myHandler" runtimeLib="true" version="1">
</requestHandler>
```

If there are parameters available in the custom handler, you can define them in the same way as any other request handler definition.



Blob store can only be used to dynamically load components configured in solrconfig.xml. Components specified in schema.xml cannot be loaded from blob store.

## Deleting Blobs

Once loaded to the blob store, blobs are handled very similarly to usual indexed documents in Solr. To delete blobs, you can use the same approaches used to delete individual documents from the index, namely Delete By ID and Delete By Query.

For example, to delete a blob with the id test/1, you would issue a command like this:

```
curl -H 'Content-Type: application/json' -d '{"delete": {"id": "test/1"}}'
http://localhost:8983/solr/.system/update?commit=true
```

Be sure to tell Solr to perform a [commit](#) as part of the request (commit=true in the above example) to see the change immediately. If you do not instruct Solr to perform a commit, Solr will use the .system collection autoCommit settings, which may not be the expected behavior.

You can also use the delete by query syntax, as so:

```
curl -H 'Content-Type: application/json' -d '{"delete": {"query": "id:test/1"}}'
http://localhost:8983/solr/.system/update?commit=true
```

For more on deleting documents generally, see the section [Sending JSON Update Commands](#).

## Config API

The Config API enables manipulating various aspects of your solrconfig.xml using REST-like API calls.

This feature is enabled by default and works similarly in both SolrCloud and standalone mode. Many commonly edited properties (such as cache sizes and commit settings) and request handler definitions can be changed with this API.

When using this API, solrconfig.xml is not changed. Instead, all edited configuration is stored in a file called configoverlay.json. The values in configoverlay.json override the values in solrconfig.xml.

### Config API Endpoints

All Config API endpoints are collection-specific, meaning this API can inspect or modify the configuration for a single collection at a time.

- collection/config: retrieve the full effective config, or modify the config. Use GET to retrieve and POST for executing commands.
- collection/config/overlay: retrieve the details in the configoverlay.json only, removing any options defined in solrconfig.xml directly or implicitly through defaults.
- collection/config/params: create parameter sets that can override or take the place of parameters defined in solrconfig.xml. See [Request Parameters API](#) for more information about this endpoint.

### Retrieving the Config

All configuration items can be retrieved by sending a GET request to the /config endpoint:

#### V1 API

```
http://localhost:8983/solr/techproducts/config
```

**V2 API**

```
http://localhost:8983/api/collections/techproducts/config
```

The response will be the Solr configuration resulting from merging settings in configoverlay.json with those in solrconfig.xml.

It's possible to restrict the returned config to a top-level section, such as, query, requestHandler or updateHandler. To do this, append the name of the section to the config endpoint. For example, to retrieve configuration for all request handlers:

**V1 API**

```
http://localhost:8983/solr/techproducts/config/requestHandler
```

**V2 API**

```
http://localhost:8983/api/collections/techproducts/config/requestHandler
```

The output will be details of each request handler defined in solrconfig.xml, all [defined implicitly](#) by Solr, and all defined with this Config API stored in configoverlay.json. To see the configuration for implicit request handlers, add expandParams=true to the request. See the documentation for the implicit request handlers for examples using this command.

The available top-level sections that can be added as path parameters are: query, requestHandler, searchComponent, updateHandler, queryResponseWriter, initParams, znodeVersion, listener, directoryFactory, indexConfig, and codecFactory.

To further restrict the request to a single component within a top-level section, use the componentName request parameter.

For example, to return configuration for the /select request handler:

**V1 API**

```
http://localhost:8983/solr/techproducts/config/requestHandler?componentName=/select
```

## V2 API

```
http://localhost:8983/api/collections/techproducts/config/requestHandler?componentName=/select
```

The output of this command will look similar to:

```
{
 "config": {"requestHandler": {"/select": {
 "name": "/select",
 "class": "solr.SearchHandler",
 "defaults": {
 "echoParams": "explicit",
 "rows": 10,
 "preferLocalShards": false
 }
 }}}
}
```

The ability to restrict to objects within a top-level section is limited to request handlers (requestHandler), search components (searchComponent), and response writers (queryResponseWriter).

## Commands to Modify the Config

This API uses specific commands with POST requests to tell Solr what property or type of property to add to or modify in configoverlay.json. The commands are passed with the data to add or modify the property or component.

The Config API commands for modifications are categorized into 3 types, each of which manipulate specific data structures in solrconfig.xml. These types are:

- set-property and unset-property for [Common Properties](#)
- Component-specific add-, update-, and delete- commands for [Custom Handlers and Local Components](#)
- set-user-property and unset-user-property for [User-defined properties](#)

### Commands for Common Properties

The common properties are those that are frequently customized in a Solr instance. They are manipulated with two commands:

- set-property: Set a well known property. The names of the properties are predefined and fixed. If the property has already been set, this command will overwrite the previous setting.
- unset-property: Remove a property set using the set-property command.

The properties that can be configured with set-property and unset-property are predefined and listed below. The names of these properties are derived from their XML paths as found in solrconfig.xml.

### Update Handler Settings

See [UpdateHandlers in SolrConfig](#) for defaults and acceptable values for these settings.

- updateHandler.autoCommit.maxDocs
- updateHandler.autoCommit.maxTime
- updateHandler.autoCommit.openSearcher
- updateHandler.autoSoftCommit.maxDocs
- updateHandler.autoSoftCommit.maxTime
- updateHandler.commitWithin.softCommit
- updateHandler.indexWriter.closeWaitsForMerges

## Query Settings

See [Query Settings in SolrConfig](#) for defaults and acceptable values for these settings.

### Caches and Cache Sizes

- query.filterCache.class
- query.filterCache.size
- query.filterCache.initialSize
- query.filterCache.autowarmCount
- query.filterCache.maxRamMB
- query.filterCache.regenerator
- query.queryResultCache.class
- query.queryResultCache.size
- query.queryResultCache.initialSize
- query.queryResultCache.autowarmCount
- query.queryResultCache.maxRamMB
- query.queryResultCache.regenerator
- query.documentCache.class
- query.documentCache.size
- query.documentCache.initialSize
- query.documentCache.autowarmCount
- query.documentCache.regenerator
- query.fieldValueCache.class
- query.fieldValueCache.size
- query.fieldValueCache.initialSize
- query.fieldValueCache.autowarmCount
- query.fieldValueCache.regenerator

### Query Sizing and Warming

- query.maxBooleanClauses
- query.enableLazyFieldLoading
- query.useFilterForSortedQuery

- query.queryResultWindowSize
- query.queryResultMaxDocCached

## RequestDispatcher Settings

See [RequestDispatcher in SolrConfig](#) for defaults and acceptable values for these settings.

- requestDispatcher.handleSelect
- requestDispatcher.requestParsers.enableRemoteStreaming
- requestDispatcher.requestParsers.enableStreamBody
- requestDispatcher.requestParsers.multipartUploadLimitInKB
- requestDispatcher.requestParsers.formDataUploadLimitInKB
- requestDispatcher.requestParsers.addHttpRequestToContext

## Examples of Common Properties

Constructing a command to modify or add one of these properties follows this pattern:

```
{"set-property": {"<>property</>": "<>value</>"}}
```

A request to increase the updateHandler.autoCommit.maxTime would look like:

### V1 API

```
curl -X POST -H 'Content-type: application/json' -d '{"set-property":{"updateHandler.autoCommit.maxTime":15000}}'
http://localhost:8983/solr/techproducts/config
```

### V2 API

```
curl -X POST -H 'Content-type: application/json' -d '{"set-property":{"updateHandler.autoCommit.maxTime":15000}}'
http://localhost:8983/api/collections/techproducts/config
```

You can use the config/overlay endpoint to verify the property has been added to configoverlay.json:

### V1 API

```
curl http://localhost:8983/solr/techproducts/config/overlay?omitHeader=true
```

## V2 API

```
curl http://localhost:8983/api/collections/techproducts/config/overlay?omitHeader=true
```

Output:

```
{
 "overlay": {
 "znodeVersion": 1,
 "props": {
 "updateHandler": {
 "autoCommit": {"maxTime": 15000}
 }
 }
 }
}
```

To unset the property:

## V1 API

```
curl -X POST -H 'Content-type: application/json' -d '{"unset-property":
"updateHandler.autoCommit.maxTime"}' http://localhost:8983/solr/techproducts/config
```

## V2 API

```
curl -X POST -H 'Content-type: application/json' -d '{"unset-property":
"updateHandler.autoCommit.maxTime"}'
http://localhost:8983/api/collections/techproducts/config
```

## Commands for Handlers and Components

Request handlers, search components, and other types of localized Solr components (such as query parsers, update processors, etc.) can be added, updated and deleted with specific commands for the type of component being modified.

The syntax is similar in each case: add-<component-name>, update-<component-name>, and delete-<component-name>. The command name is not case sensitive, so Add-RequestHandler, ADD-REQUESTHANDLER and add-requesthandler are equivalent.

In each case, add- commands add a new configuration to configoverlay.json, which will override any other settings for the component in solrconfig.xml.

update- commands overwrite an existing setting in configoverlay.json.

delete- commands remove the setting from configoverlay.json.

Settings removed from configoverlay.json are not removed from solrconfig.xml if they happen to be duplicated there.

The full list of available commands follows below:

### Basic Commands for Components

These commands are the most commonly used:

- add-requesthandler
- update-requesthandler
- delete-requesthandler
- add-searchcomponent
- update-searchcomponent
- delete-searchcomponent
- add-initparams
- update-initparams
- delete-initparams
- add-queryresponsewriter
- update-queryresponsewriter
- delete-queryresponsewriter

### Advanced Commands for Components

These commands allow registering more advanced customizations to Solr:

- add-queryparser
- update-queryparser
- delete-queryparser
- add-valuesourceparser
- update-valuesourceparser
- delete-valuesourceparser
- add-transformer
- update-transformer
- delete-transformer
- add-updateprocessor
- update-updateprocessor
- delete-updateprocessor
- add-queryconverter
- update-queryconverter
- delete-queryconverter
- add-listener
- update-listener
- delete-listener

- add-runtimelib
- update-runtimelib
- delete-runtimelib

### Examples of Handler and Component Commands

To create a request handler, we can use the add-requesthandler command:

```
curl -X POST -H 'Content-type:application/json' -d '{
 "add-requesthandler": {
 "name": "/mypath",
 "class": "solr.DumpRequestHandler",
 "defaults": { "x": "y" , "a": "b" , "rows":10 },
 "useParams": "x"
 }
' http://localhost:8983/solr/techproducts/config
```

### V1 API

```
curl -X POST -H 'Content-type:application/json' -d '{
 "add-requesthandler": {
 "name": "/mypath",
 "class": "solr.DumpRequestHandler",
 "defaults": { "x": "y" , "a": "b" , "rows":10 },
 "useParams": "x"
 }
' http://localhost:8983/solr/techproducts/config
```

### V2 API

```
curl -X POST -H 'Content-type:application/json' -d '{
 "add-requesthandler": {
 "name": "/mypath",
 "class": "solr.DumpRequestHandler",
 "defaults": { "x": "y" , "a": "b" , "rows":10 },
 "useParams": "x"
 }
' http://localhost:8983/api/collections/techproducts/config
```

Make a call to the new request handler to check if it is registered:

```
curl http://localhost:8983/solr/techproducts/mypath?omitHeader=true
```

And you should see the following as output:

```
{
 "params": {
 "indent": "true",
 "a": "b",
 "x": "y",
 "rows": "10"},
 "context": {
 "webapp": "/solr",
 "path": "/mypath",
 "httpMethod": "GET"}}
```

To update a request handler, you should use the update-requesthandler command:

## V1 API

```
curl -X POST -H 'Content-type:application/json' -d '{
 "update-requesthandler": {
 "name": "/mypath",
 "class": "solr.DumpRequestHandler",
 "defaults": {"x": "new value for X", "rows": "20"},
 "useParams": "x"
 }
' http://localhost:8983/solr/techproducts/config
```

## V2 API

```
curl -X POST -H 'Content-type:application/json' -d '{
 "update-requesthandler": {
 "name": "/mypath",
 "class": "solr.DumpRequestHandler",
 "defaults": {"x": "new value for X", "rows": "20"},
 "useParams": "x"
 }
' http://localhost:8983/api/collections/techproducts/config
```

As a second example, we'll create another request handler, this time adding the 'terms' component as part of the definition:

## V1 API

```
curl -X POST -H 'Content-type:application/json' -d '{
 "add-requesthandler": {
 "name": "/myterms",
 "class": "solr.SearchHandler",
 "defaults": {"terms": true, "distrib":false},
 "components": ["terms"]
 }
' http://localhost:8983/solr/techproducts/config
```

## V2 API

```
curl -X POST -H 'Content-type:application/json' -d '{
 "add-requesthandler": {
 "name": "/myterms",
 "class": "solr.SearchHandler",
 "defaults": {"terms": true, "distrib":false},
 "components": ["terms"]
 }
' http://localhost:8983/api/collections/techproducts/config
```

## Commands for User-Defined Properties

Solr lets users templatize the solrconfig.xml using the place holder format \${variable\_name:default\_val}. You could set the values using system properties, for example, -Dvariable\_name= my\_customvalue. The same can be achieved during runtime using these commands:

- **set-user-property**: Set a user-defined property. If the property has already been set, this command will overwrite the previous setting.
- **unset-user-property**: Remove a user-defined property.

The structure of the request is similar to the structure of requests using other commands, in the format of "command": {"variable\_name": "property\_value"}. You can add more than one variable at a time if necessary.

For more information about user-defined properties, see the section [User defined properties in core.properties](#).

See also the section [Creating and Updating User-Defined Properties](#) below for examples of how to use this type of command.

### Creating and Updating User-Defined Properties

This command sets a user property.

**V1 API**

```
curl -X POST -H 'Content-type:application/json' -d '{"set-user-property": {"variable_name": "some_value"}}' http://localhost:8983/solr/techproducts/config
```

**V2 API**

```
curl -X POST -H 'Content-type:application/json' -d '{"set-user-property": {"variable_name": "some_value"}}' http://localhost:8983/api/collections/techproducts/config
```

Again, we can use the /config/overlay endpoint to verify the changes have been made:

**V1 API**

```
curl http://localhost:8983/solr/techproducts/config/overlay?omitHeader=true
```

**V2 API**

```
curl http://localhost:8983/api/collections/techproducts/config/overlay?omitHeader=true
```

And we would expect to see output like this:

```
{"overlay":{
 "znodeVersion":5,
 "userProps":{
 "variable_name": "some_value"}
}
```

To unset the variable, issue a command like this:

**V1 API**

```
curl -X POST -H 'Content-type:application/json' -d '{"unset-user-property": "variable_name"}' http://localhost:8983/solr/techproducts/config
```

## V2 API

```
curl -X POST -H 'Content-type:application/json' -d '{"unset-user-property": "variable_name"}'
http://localhost:8983/api/collections/techproducts/config
```

### What about updateRequestProcessorChain?

The Config API does not let you create or edit updateRequestProcessorChain elements. However, it is possible to create updateProcessor entries and use them by name to create a chain.

For example:

## V1 API

```
curl -X POST -H 'Content-type:application/json' -d '{"add-updateprocessor":
{"name": "firstFld",
"class": "solr.FirstFieldValueUpdateProcessorFactory",
"fieldName": "test_s"}
' http://localhost:8983/solr/techproducts/config
```

## V2 API

```
curl -X POST -H 'Content-type:application/json' -d '{"add-updateprocessor":
{"name": "firstFld",
"class": "solr.FirstFieldValueUpdateProcessorFactory",
"fieldName": "test_s"}
' http://localhost:8983/api/collections/techproducts/config
```

You can use this directly in your request by adding a parameter in the updateRequestProcessorChain for the specific update processor called processor=firstFld.

## How to Map solrconfig.xml Properties to JSON

By using this API, you will be generating JSON representations of properties defined in solrconfig.xml. To understand how properties should be represented with the API, let's take a look at a few examples.

Here is what a request handler looks like in solrconfig.xml:

```
<requestHandler name="/query" class="solr.SearchHandler">
 <lst name="defaults">
 <str name="echoParams">explicit</str>
 <int name="rows">10</int>
 </lst>
</requestHandler>
```

The same request handler defined with the Config API would look like this:

```
{
 "add-requesthandler":{
 "name": "/query",
 "class": "solr.SearchHandler",
 "defaults":{
 "echoParams": "explicit",
 "rows": 10
 }
 }
}
```

The QueryElevationComponent searchComponent in solrconfig.xml looks like this:

```
<searchComponent name="elevator" class="solr.QueryElevationComponent" >
 <str name="queryFieldType">string</str>
 <str name="config-file">elevate.xml</str>
</searchComponent>
```

And the same searchComponent with the Config API:

```
{
 "add-searchcomponent":{
 "name": "elevator",
 "class": "solr.QueryElevationComponent",
 "queryFieldType": "string",
 "config-file": "elevate.xml"
 }
}
```

Removing the searchComponent with the Config API:

```
{
 "delete-searchcomponent": "elevator"
}
```

A simple highlighter looks like this in solrconfig.xml (example has been truncated for space):

```
<searchComponent class="solr.HighlightComponent" name="highlight">
 <highlighting>
 <fragmenter name="gap"
 default="true"
 class="solr.highlight.GapFragmenter">
 <lst name="defaults">
 <int name="hl.fragsize">100</int>
 </lst>
 </fragmenter>

 <formatter name="html"
 default="true"
 class="solr.highlight.HtmlFormatter">
 <lst name="defaults">
 <str name="hl.simple.pre"><! [CDATA[]></str>
 <str name="hl.simple.post"><! [CDATA[]]></str>
 </lst>
 </formatter>

 <encoder name="html" class="solr.highlight.HtmlEncoder" />
 ...
</highlighting>
```

The same highlighter with the Config API:

```
{
 "add-searchcomponent": {
 "name": "highlight",
 "class": "solr.HighlightComponent",
 "": {
 "gap": {
 "default": "true",
 "name": "gap",
 "class": "solr.highlight.GapFrgmenter",
 "defaults": {
 "hl.fragsize": 100
 }
 }
 },
 "html": [
 {
 "default": "true",
 "name": "html",
 "class": "solr.highlight.HtmlFormatter",
 "defaults": {
 "hl.simple.pre": "before-",
 "hl.simple.post": "-after"
 }
 },
 {
 "name": "html",
 "class": "solr.highlight.HtmlEncoder"
 }
]
 }
}
```

Set autoCommit properties in solrconfig.xml:

```
<autoCommit>
 <maxTime>15000</maxTime>
 <openSearcher>false</openSearcher>
</autoCommit>
```

Define the same properties with the Config API:

```
{
 "set-property": {
 "updateHandler.autoCommit.maxTime":15000,
 "updateHandler.autoCommit.openSearcher":false
 }
}
```

## Name Components for the Config API

The Config API always allows changing the configuration of any component by name. However, some configurations such as listener or initParams do not require a name in solrconfig.xml. In order to be

able to update and delete of the same item in configoverlay.json, the name attribute becomes mandatory.

## How the Config API Works

Every core watches the ZooKeeper directory for the configset being used with that core. In standalone mode, however, there is no watch (because ZooKeeper is not running). If there are multiple cores in the same node using the same configset, only one ZooKeeper watch is used.

For instance, if the configset 'myconf' is used by a core, the node would watch /configs/myconf. Every write operation performed through the API would 'touch' the directory and all watchers are notified. Every core would check if the schema file, solrconfig.xml, or configoverlay.json has been modified by comparing the znode versions. If any have been modified, the core is reloaded.

If params.json is modified, the params object is just updated without a core reload (see [Request Parameters API](#) for more information about params.json).

### Empty Command

If an empty command is sent to the /config endpoint, the watch is triggered on all cores using this configset. For example:

#### V1 API

```
curl -X POST -H 'Content-type:application/json' -d '{}'
http://localhost:8983/solr/techproducts/config
```

#### V2 API

```
curl -X POST -H 'Content-type:application/json' -d '{}'
http://localhost:8983/api/collections/techproducts/config
```

Directly editing any files without 'touching' the directory **will not** make it visible to all nodes.

It is possible for components to watch for the configset 'touch' events by registering a listener using `SolrCore#registerConfListener()`.

### Listening to Config Changes

Any component can register a listener using:

```
SolrCore#addConfListener(Runnable listener)
```

to get notified for config changes. This is not very useful if the files modified result in core reloads (i.e., configoverlay.xml or the schema). Components can use this to reload the files they are interested in.

## Request Parameters API

The Request Parameters API allows creating parameter sets, a.k.a. paramsets, that can override or take the place of parameters defined in `solrconfig.xml`.

The parameter sets defined with this API can be used in requests to Solr, or referenced directly in `solrconfig.xml` request handler definitions.

It is really another endpoint of the [Config API](#) instead of a separate API, and has distinct commands. It does not replace or modify any sections of `solrconfig.xml`, but instead provides another approach to handling parameters used in requests. It behaves in the same way as the Config API, by storing parameters in another file that will be used at runtime. In this case, the parameters are stored in a file named `params.json`. This file is kept in ZooKeeper or in the `conf` directory of a standalone Solr instance.

The settings stored in `params.json` are used at query time to override settings defined in `solrconfig.xml` in some cases as described below.

When might you want to use this feature?

- To avoid frequently editing your `solrconfig.xml` to update request parameters that change often.
- To reuse parameters across various request handlers.
- To mix and match parameter sets at request time.
- To avoid a reload of your collection for small parameter changes.

## The Request Parameters Endpoint

All requests are sent to the `/config/params` endpoint of the Config API.

### Setting Request Parameters

The request to set, unset, or update request parameters is sent as a set of Maps with names. These objects can be directly used in a request or a request handler definition.

The available commands are:

- `set`: Create or overwrite a parameter set map.
- `unset`: delete a parameter set map.
- `update`: update a parameter set map. This is equivalent to a `map.putAll(newMap)`. Both the maps are merged and if the new map has same keys as old they are overwritten.

You can mix these commands into a single request if necessary.

Each map must include a name so it can be referenced later, either in a direct request to Solr or in a request handler definition.

In the following example, we are setting 2 sets of parameters named 'myFacets' and 'myQueries'.

```
curl http://localhost:8983/solr/techproducts/config/params -H 'Content-type:application/json' -d
'{
 "set": {
 "myFacets": {
 "facet": "true",
 "facet.limit": 5
 }
 },
 "set": {
 "myQueries": {
 "defType": "edismax",
 "rows": "5",
 "df": "text_all"
 }
 }
}'
```

In the above example all the parameters are equivalent to the "defaults" in solrconfig.xml. It is possible to add invariants and appends as follows:

```
curl http://localhost:8983/solr/techproducts/config/params -H 'Content-type:application/json' -d
'{
 "set": {
 "my_handler_params": {
 "facet.limit": 5,
 "_invariants_": {
 "facet": true
 },
 "_appends_": {
 "facet.field": ["field1", "field2"]
 }
 }
 }
}'
```

## Using Request Parameters with RequestHandlers

After creating the my\_handler\_params paramset in the above section, it is possible to define a request handler as follows:

```
<requestHandler name="/my_handler" class="solr.SearchHandler" useParams="my_handler_params"/>
```

It will be equivalent to a standard request handler definition such as this one:

```
<requestHandler name="/my_handler" class="solr.SearchHandler">
 <lst name="defaults">
 <int name="facet.limit">5</int>
 </lst>
 <lst name="invariants">
 <bool name="facet">true</bool>
 </lst>
 <lst name="appends">
 <arr name="facet.field">
 <str>field1</str>
 <str>field2</str>
 </arr>
 </lst>
</requestHandler>
```

## Implicit RequestHandlers with the Request Parameters API

Solr ships with many out-of-the-box request handlers that may only be configured via the Request Parameters API, because their configuration is not present in `solrconfig.xml`. See [Implicit RequestHandlers](#) for the paramset to use when configuring an implicit request handler.

## Viewing Expanded Paramsets and Effective Parameters with RequestHandlers

To see the expanded paramset and the resulting effective parameters for a RequestHandler defined with `useParams`, use the `expandParams` request parameter. As an example, for the `/export` request handler:

```
curl
http://localhost:8983/solr/techproducts/config/requestHandler?componentName=/export&expandParams=true
```

## Viewing Request Parameters

To see the paramsets that have been created, you can use the `/config/params` endpoint to read the contents of `params.json`, or use the name in the request:

```
curl http://localhost:8983/solr/techproducts/config/params

#Or use the paramset name
curl http://localhost:8983/solr/techproducts/config/params/myQueries
```

## The `useParams` Parameter

When making a request, the `useParams` parameter applies the request parameters sent to the request. This is translated at request time to the actual parameters.

For example (using the names we set up in the earlier examples, please replace with your own name):

```
http://localhost/solr/techproducts/select?useParams=myQueries
```

It is possible to pass more than one parameter set in the same request. For example:

```
http://localhost/solr/techproducts/select?useParams=myFacets,myQueries
```

In the above example the parameter set 'myQueries' is applied on top of 'myFacets'. So, values in 'myQueries' take precedence over values in 'myFacets'. Additionally, any values passed in the request take precedence over useParams parameters. This acts like the "defaults" specified in the <requestHandler> definition in solrconfig.xml.

The parameter sets can be used directly in a request handler definition as follows. Please note that the useParams specified is always applied even if the request contains useParams.

```
<requestHandler name="/terms" class="solr.SearchHandler" useParams="myQueries">
 <lst name="defaults">
 <bool name="terms">true</bool>
 <bool name="distrib">false</bool>
 </lst>
 <arr name="components">
 <str>terms</str>
 </arr>
</requestHandler>
```

To summarize, parameters are applied in this order:

- parameters defined in <invariants> in solrconfig.xml.
- parameters applied in invariants in params.json and that is specified in the requesthandler definition or even in request
- parameters defined in the request directly.
- parameter sets defined in the request, in the order they have been listed with useParams.
- parameter sets defined in params.json that have been defined in the request handler.
- parameters defined in <defaults> in solrconfig.xml.

## Public APIs

The RequestParams Object can be accessed using the method SolrConfig#getRequestParams(). Each paramset can be accessed by their name using the method RequestParams#getRequestParams(String name).

## Examples Using the Request Parameters API

The Solr "films" example demonstrates the use of the parameters API. You can use this example in your Solr installation (in the example/films directory) or view the files in the Apache GitHub mirror at <https://github.com/apache/lucene-solr/tree/master/solr/example/films>.

## Managed Resources

Managed resources expose a REST API endpoint for performing Create-Read-Update-Delete (CRUD) operations on a Solr object.

Any long-lived Solr object that has configuration settings and/or data is a good candidate to be a managed resource. Managed resources complement other programmatically manageable components in Solr, such as the RESTful schema API to add fields to a managed schema.

Consider a Web-based UI that offers Solr-as-a-Service where users need to configure a set of stop words and synonym mappings as part of an initial setup process for their search application. This type of use case can easily be supported using the Managed Stop Filter & Managed Synonym Graph Filter Factories provided by Solr, via the Managed resources REST API.

Users can also write their own custom plugins, that leverage the same internal hooks to make additional resources REST managed.

All of the examples in this section assume you are running the "techproducts" Solr example:

```
bin/solr -e techproducts
```

## Managed Resources Overview

Let's begin learning about managed resources by looking at a couple of examples provided by Solr for managing stop words and synonyms using a REST API. After reading this section, you'll be ready to dig into the details of how managed resources are implemented in Solr so you can start building your own implementation.

### Managing Stop Words

To begin, you need to define a field type that uses the [ManagedStopFilterFactory](#), such as:

```
<fieldType name="managed_en" positionIncrementGap="100">
 <analyzer>
 <tokenizer class="solr.StandardTokenizerFactory"/>
 <filter class="solr.ManagedStopFilterFactory" ①
 managed="english" /> ②
 </analyzer>
</fieldType>
```

There are two important things to notice about this field type definition:

- ① The filter implementation class is `solr.ManagedStopFilterFactory`. This is a special implementation of the [StopFilterFactory](#) that uses a set of stop words that are managed from a REST API.
- ② The `managed="english"` attribute gives a name to the set of managed stop words, in this case indicating the stop words are for English text.

The REST endpoint for managing the English stop words in the `techproducts` collection is:  
`/solr/techproducts/schema/analysis/stopwords/english`.

The example resource path should be mostly self-explanatory. It should be noted that the ManagedStopFilterFactory implementation determines the /schema/analysis/stopwords part of the path, which makes sense because this is an analysis component defined by the schema.

It follows that a field type that uses the following filter:

```
<filter class="solr.ManagedStopFilterFactory"
 managed="french" />
```

would resolve to path: /solr/techproducts/schema/analysis/stopwords/french.

So now let's see this API in action, starting with a simple GET request:

```
curl "http://localhost:8983/solr/techproducts/schema/analysis/stopwords/english"
```

Assuming you sent this request to Solr, the response body is a JSON document:

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 1
 },
 "wordSet": {
 "initArgs": {"ignoreCase": true},
 "initializedOn": "2014-03-28T20:53:53.058Z",
 "managedList": [
 "a",
 "an",
 "and",
 "are"
]
 }
}
```

The sample\_techproducts\_configs configset ships with a pre-built set of managed stop words, however you should only interact with this file using the API and not edit it directly.

One thing that should stand out to you in this response is that it contains a managedList of words as well as initArgs. This is an important concept in this framework — managed resources typically have configuration and data. For stop words, the only configuration parameter is a boolean that determines whether to ignore the case of tokens during stop word filtering (ignoreCase=true|false). The data is a list of words, which is represented as a JSON array named managedList in the response.

Now, let's add a new word to the English stop word list using an HTTP PUT:

```
curl -X PUT -H 'Content-type:application/json' --data-binary '["foo"]'
"http://localhost:8983/solr/techproducts/schema/analysis/stopwords/english"
```

Here we're using curl to PUT a JSON list containing a single word "foo" to the managed English stop words set. Solr will return 200 if the request was successful. You can also put multiple words in a single PUT request.

You can test to see if a specific word exists by sending a GET request for that word as a child resource of the set, such as:

```
curl "http://localhost:8983/solr/techproducts/schema/analysis/stopwords/english/foo"
```

This request will return a status code of 200 if the child resource (foo) exists or 404 if it does not exist the managed list.

To delete a stop word, you would do:

```
curl -X DELETE "http://localhost:8983/solr/techproducts/schema/analysis/stopwords/english/foo"
```



PUT/POST is used to add terms to an existing list instead of replacing the list entirely. This is because it is more common to add a term to an existing list than it is to replace a list altogether, so the API favors the more common approach of incrementally adding terms especially since deleting individual terms is also supported.

## Managing Synonyms

For the most part, the API for managing synonyms behaves similar to the API for stop words, except instead of working with a list of words, it uses a map, where the value for each entry in the map is a set of synonyms for a term. As with stop words, the sample\_techproducts\_configs configset includes a pre-built set of synonym mappings suitable for the sample data that is activated by the following field type definition in schema.xml:

```
<fieldType name="managed_en" positionIncrementGap="100">
 <analyzer type="index">
 <tokenizer class="solr.StandardTokenizerFactory"/>
 <filter class="solr.ManagedStopFilterFactory" managed="english" />
 <filter class="solr.ManagedSynonymGraphFilterFactory" managed="english" />
 <filter class="solr.FlattenGraphFilterFactory"/> <!-- required on index analyzers after graph filters -->
 </analyzer>
 <analyzer type="query">
 <tokenizer class="solr.StandardTokenizerFactory"/>
 <filter class="solr.ManagedStopFilterFactory" managed="english" />
 <filter class="solr.ManagedSynonymGraphFilterFactory" managed="english" />
 </analyzer>
</fieldType>
```

To get the map of managed synonyms, send a GET request to:

```
curl "http://localhost:8983/solr/techproducts/schema/analysis/synonyms/english"
```

This request will return a response that looks like:

```
{
 "responseHeader":{
 "status":0,
 "QTime":3},
 "synonymMappings":{
 "initArgs":{
 "ignoreCase":true,
 "format":"solr"},
 "initializedOn":"2014-12-16T22:44:05.33Z",
 "managedMap":{
 "GB":
 ["GiB",
 "Gigabyte"],
 "TV":
 ["Television"],
 "happy":
 ["glad",
 "joyful"]}}}
```

Managed synonyms are returned under the managedMap property which contains a JSON Map where the value of each entry is a set of synonyms for a term, such as "happy" has synonyms "glad" and "joyful" in the example above.

To add a new synonym mapping, you can PUT/POST a single mapping such as:

```
curl -X PUT -H 'Content-type:application/json' --data-binary '{"mad":["angry", "upset"]}'
"http://localhost:8983/solr/techproducts/schema/analysis/synonyms/english"
```

The API will return status code 200 if the PUT request was successful. To determine the synonyms for a specific term, you send a GET request for the child resource, such as /schema/analysis/synonyms/english/mad would return ["angry", "upset"].

You can also PUT a list of symmetric synonyms, which will be expanded into a mapping for each term in the list. For example, you could PUT the following list of symmetric synonyms using the JSON list syntax instead of a map:

```
curl -X PUT -H 'Content-type:application/json' --data-binary '["funny", "entertaining",
"whimsical", "jocular"]'
"http://localhost:8983/solr/techproducts/schema/analysis/synonyms/english"
```

Note that the expansion is performed when processing the PUT request so the underlying persistent state is still a managed map. Consequently, if after sending the previous PUT request, you did a GET for /schema/analysis/synonyms/english/jocular, then you would receive a list containing ["funny", "entertaining", "whimsical"]. Once you've created synonym mappings using a list, each term must be managed separately.

Lastly, you can delete a mapping by sending a DELETE request to the managed endpoint.

## Applying Managed Resource Changes

Changes made to managed resources via this REST API are not applied to the active Solr components until the Solr collection (or Solr core in single server mode) is reloaded.

For example: after adding or deleting a stop word, you must reload the core/collection before changes become active; related APIs: [CoreAdmin API](#) and [Collections API](#).

This approach is required when running in distributed mode so that we are assured changes are applied to all cores in a collection at the same time so that behavior is consistent and predictable. It goes without saying that you don't want one of your replicas working with a different set of stop words or synonyms than the others.

One subtle outcome of this *apply-changes-at-reload* approach is that the once you make changes with the API, there is no way to read the active data. In other words, the API returns the most up-to-date data from an API perspective, which could be different than what is currently being used by Solr components.

However, the intent of this API implementation is that changes will be applied using a reload within a short time frame after making them so the time in which the data returned by the API differs from what is active in the server is intended to be negligible.



Changing things like stop words and synonym mappings typically require reindexing existing documents if being used by index-time analyzers. The RestManager framework does not guard you from this, it simply makes it possible to programmatically build up a set of stop words, synonyms, etc. See the section [Reindexing](#) for more information about reindexing your documents.

## RestManager Endpoint

Metadata about registered ManagedResources is available using the /schema/managed endpoint for each collection.

Assuming you have the managed\_en field type shown above defined in your schema.xml, sending a GET request to the following resource will return metadata about which schema-related resources are being managed by the RestManager:

```
curl "http://localhost:8983/solr/techproducts/schema/managed"
```

The response body is a JSON document containing metadata about managed resources under the /schema root:

```
{
 "responseHeader":{
 "status":0,
 "QTime":3
 },
 "managedResources": [
 {
 "resourceId":"/schema/analysis/stopwords/english",
 "class":"org.apache.solr.rest.schema.analysis.ManagedWordSetResource",
 "numObservers": "1"
 },
 {
 "resourceId":"/schema/analysis/synonyms/english",
 "class":
 "org.apache.solr.rest.schema.analysis.ManagedSynonymGraphFilterFactory$SynonymManager",
 "numObservers": "1"
 }
]
}
```

You can also create new managed resource using PUT/POST to the appropriate URL – before ever configuring anything that uses these resources.

For example, imagine we want to build up a set of German stop words. Before we can start adding stop words, we need to create the endpoint:

```
/solr/techproducts/schema/analysis/stopwords/german
```

To create this endpoint, send the following PUT/POST request to the endpoint we wish to create:

```
curl -X PUT -H 'Content-type:application/json' --data-binary \
'{"class":"org.apache.solr.rest.schema.analysis.ManagedWordSetResource"}' \
"http://localhost:8983/solr/techproducts/schema/analysis/stopwords/german"
```

Solr will respond with status code 200 if the request is successful. Effectively, this action registers a new endpoint for a managed resource in the RestManager. From here you can start adding German stop words as we saw above:

```
curl -X PUT -H 'Content-type:application/json' --data-binary '["die"]' \
"http://localhost:8983/solr/techproducts/schema/analysis/stopwords/german"
```

For most users, creating resources in this way should never be necessary, since managed resources are created automatically when configured.

However, You may want to explicitly delete managed resources if they are no longer being used by a Solr component.

For instance, the managed resource for German that we created above can be deleted because there are no Solr components that are using it, whereas the managed resource for English stop words cannot be deleted

because there is a token filter declared in schema.xml that is using it.

```
curl -X DELETE "http://localhost:8983/solr/techproducts/schema/analysis/stopwords/german"
```

# Implicit RequestHandlers

Solr ships with many out-of-the-box RequestHandlers, which are called implicit because they do not need to be configured in `solrconfig.xml` before you are able to use them.

These handlers have pre-defined default parameters, known as *paramsets*, which can be modified if necessary.

## Available Implicit Endpoints



All endpoint paths listed below should be placed after Solr's host and port (if a port is used) to construct a URL.

### Admin Handlers

Many of these handlers are used throughout the Admin UI to show information about Solr.

#### File

Returns content of files in  `${solr.home}/conf/`. This handler must have a collection name in the path to the endpoint.

API Endpoint	Class & Javadocs	Paramset
<code>solr/&lt;collection&gt;/admin/file</code>	<a href="#">ShowFileRequestHandler</a>	_ADMIN_FILE

#### Logging

Retrieve and modify registered loggers.

API Endpoints	Class & Javadocs	Paramset
<code>v1: solr/admin/info/logging</code>	<a href="#">LoggingHandler</a>	_ADMIN_LOGGING
<code>v2: api/node/logging</code>		

#### Luke

Expose the internal lucene index. This handler must have a collection name in the path to the endpoint.

**Documentation:** <http://wiki.apache.org/solr/LukeRequestHandler>

API Endpoint	Class & Javadocs	Paramset
<code>solr/&lt;collection&gt;/admin/luke</code>	<a href="#">LukeRequestHandler</a>	_ADMIN_LUKE

#### MBeans

Provide info about all registered [SolrInfoMBeans](#). This handler must have a collection name in the path to the endpoint.

**Documentation:** [MBean Request Handler](#)

API Endpoint	Class & Javadocs	Paramset
solr/<collection>/admin/mbeans	<a href="#">SolrInfoMBeanHandler</a>	_ADMIN_MBEANS

## Ping

Health check. This handler must have a collection name in the path to the endpoint.

**Documentation:** [Ping](#)

API Endpoint	Class & Javadocs	Paramset
solr/<collection>/admin/ping	<a href="#">PingRequestHandler</a>	_ADMIN_PING

## Plugins

Return info about all registered plugins. This handler must have a collection name in the path to the endpoint.

API Endpoint	Class & Javadocs	Paramset
solr/<collection>/admin/plugins	<a href="#">PluginInfoHandler</a>	None.

## System Properties

Return JRE system properties.

API Endpoints	Class & Javadocs	Paramset
v1: solr/admin/info/properties	<a href="#">PropertiesRequestHandler</a>	_ADMIN_PROPERTIES
v2: api/node/properties		

## Segments

Return info on last commit generation Lucene index segments.

API Endpoint	Class & Javadocs	Paramset
solr/<collection>/admin/segments	<a href="#">SegmentsInfoRequestHandler</a>	_ADMIN_SEGMENTS

## System Settings

Return server statistics and settings.

**Documentation:** <https://wiki.apache.org/solr/SystemInformationRequestHandlers#SystemInfoHandler>

API Endpoints	Class & Javadocs	Paramset
v1: solr/admin/info/system	<a href="#">SystemInfoHandler</a>	_ADMIN_SYSTEM
v2: api/node/system		

This endpoint can also take the collection or core name in the path (solr/<collection>/admin/system or

`solr/<core>/admin/system`) which will include all of the system-level information and additional information about the specific core that served the request.

## Threads

Return info on all JVM threads.

API Endpoints	Class & Javadocs	Paramset
v1: solr/admin/info/threads	<a href="#">ThreadDumpHandler</a>	_ADMIN_THREADS
v2: api/node/threads		

## Health

Reporting the health of the node (*available only in SolrCloud mode*)

API Endpoints	Class & Javadocs	Paramset
v1: solr/admin/info/health	<a href="#">HealthCheckHandler</a>	_ADMIN_HEALTH
v2: api/node/health		

This endpoint can also take the collection or core name in the path (`solr/<collection>/admin/health` or `solr/<core>/admin/health`).

## Analysis Handlers

### Document Analysis

Return a breakdown of the analysis process of the given document.

**Documentation:** <https://wiki.apache.org/solr/AnalysisRequestHandler>

API Endpoint	Class & Javadocs	Paramset
<code>solr/&lt;collection&gt;/analysis/document</code>	<a href="#">DocumentAnalysisRequestHandler</a>	_ANALYSIS_DOCUMENT

### Field Analysis

Return index- and query-time analysis over the given field(s)/field type(s). This handler drives the [Analysis screen](#) in Solr's Admin UI.

API Endpoint	Class & Javadocs	Paramset
<code>solr/&lt;collection&gt;/analysis/field</code>	<a href="#">FieldAnalysisRequestHandler</a>	_ANALYSIS_FIELD

## Handlers for Configuration

### Config API

Retrieve and modify Solr configuration.

**Documentation:** [Config API](#)

API Endpoint	Class & Javadocs	Paramset
v1: solr/<collection>/config  v2: api/collections/<collection>/config	<a href="#">SolrConfigHandler</a>	_CONFIG

## Dump

Echo the request contents back to the client.

API Endpoint	Class & Javadocs	Paramset
solr/debug/dump	<a href="#">DumpRequestHandler</a>	_DEBUG_DUMP

## Replication

Replicate indexes for SolrCloud recovery and Master/Slave index distribution. This handler must have a core name in the path to the endpoint.

API Endpoint	Class & Javadocs	Paramset
solr/<core>/replication	<a href="#">ReplicationHandler</a>	_REPLICATION

## Schema API

Retrieve and modify the Solr schema.

**Documentation:** [Schema API](#)

API Endpoint	Class & Javadocs	Paramset
v1: solr/<collection>/schema, solr/<core>/schema	<a href="#">SchemaHandler</a>	_SCHEMA
v2: api/collections/<collection>/schema, api/cores/<core>/schema		

## Query Handlers

### Export

Export full sorted result sets.

**Documentation:** [Exporting Result Sets](#)

API Endpoint	Class & Javadocs	Paramset
solr/<collection>/export	<a href="#">ExportHandler</a>	_EXPORT

### RealTimeGet

Low-latency retrieval of the latest version of a document.

**Documentation:** [RealTime Get](#)

API Endpoint	Class & Javadocs	Paramset
<code>solr/&lt;collection&gt;/get</code>	<a href="#">RealTimeGetHandler</a>	_GET

**Graph Traversal**

Return [GraphML](#) formatted output from a gatherNodes streaming expression.

**Documentation:** [Graph Traversal](#)

API Endpoint	Class & Javadocs	Paramset
<code>solr/&lt;collection&gt;/graph</code>	<a href="#">GraphHandler</a>	_ADMIN_GRAPH

**SQL**

Front end of the Parallel SQL interface.

**Documentation:** [SQL Request Handler](#)

API Endpoint	Class & Javadocs	Paramset
<code>solr/&lt;collection&gt;/sql</code>	<a href="#">SQLHandler</a>	_SQL

**Streaming Expressions**

Distributed stream processing.

**Documentation:** [Streaming Requests and Responses](#)

API Endpoint	Class & Javadocs	Paramset
<code>solr/&lt;collection&gt;/stream</code>	<a href="#">StreamHandler</a>	_STREAM

**Terms**

Return a field's indexed terms and the number of documents containing each term.

**Documentation:** [Using the Terms Component in a Request Handler](#)

API Endpoint	Class & Javadocs	Paramset
<code>solr/&lt;collection&gt;/terms</code>	<a href="#">SearchHandler</a>	_TERMS

## Update Handlers

**Update**

Add, delete and update indexed documents formatted as SolrXML, CSV, SolrJSON or javabin.

**Documentation:** [Uploading Data with Index Handlers](#)

API Endpoint	Class & Javadocs	Paramset
solr/<collection>/update	<a href="#">UpdateRequestHandler</a>	_UPDATE

## CSV Updates

Add and update CSV-formatted documents.

**Documentation:** [CSV Update Convenience Paths](#)

API Endpoint	Class & Javadocs	Paramset
solr/<collection>/update/csv	<a href="#">UpdateRequestHandler</a>	_UPDATE_CSV

## JSON Updates

Add, delete and update SolrJSON-formatted documents.

**Documentation:** [JSON Update Convenience Paths](#)

API Endpoint	Class & Javadocs	Paramset
solr/<collection>/update/json	<a href="#">UpdateRequestHandler</a>	_UPDATE_JSON

## Custom JSON Updates

Add and update custom JSON-formatted documents.

**Documentation:** [Transforming and Indexing Custom JSON](#)

API Endpoint	Class & Javadocs	Paramset
solr/<collection>/update/json /docs	<a href="#">UpdateRequestHandler</a>	_UPDATE_JSON_DOCS

## How to View Implicit Handler Paramsets

You can see configuration for all request handlers, including the implicit request handlers, via the [Config API](#).

To include the expanded paramset in the response, as well as the effective parameters from merging the paramset parameters with the built-in parameters, use the `expandParams` request parameter. For the `/export` request handler, you can make a request like this:

### V1 API

```
http://localhost:8983/solr/gettingstarted/config/requestHandler?componentName=/export&expandParams=true
```

## V2 API

```
http://localhost:8983/api/collections/gettingstarted/config/requestHandler?componentName=/export&expandParams=true
```

The response will look similar to:

```
{
 "config": {
 "requestHandler": {
 "/export": {
 "class": "solr.ExportHandler",
 "useParams": "_EXPORT",
 "components": ["query"],
 "defaults": {
 "wt": "json"
 },
 "invariants": {
 "rq": "{!xport}",
 "distrib": false
 },
 "name": "/export",
 "_useParamsExpanded_": {
 "_EXPORT": "[NOT AVAILABLE]"
 },
 "_effectiveParams_": {
 "distrib": "false",
 "omitHeader": "true",
 "wt": "json",
 "rq": "{!xport}"
 }
 }
 }
 }
}
```

## How to Edit Implicit Handler Paramsets

Because implicit request handlers are not present in `solrconfig.xml`, configuration of their associated default, invariant and appends parameters may be edited via the [Request Parameters API](#) using the paramset listed in the above table. However, other parameters, including SearchHandler components, may not be modified. The invariants and appends specified in the implicit configuration cannot be overridden.

# Solr Plugins

Solr allows you to load custom code to perform a variety of tasks within Solr, from custom Request Handlers to process your searches, to custom Analyzers and Token Filters for your text field. You can even load custom Field Types. These pieces of custom code are called plugins.

Not everyone will need to create plugins for their Solr instances - what's provided is usually enough for most applications. However, if there's something that you need, you may want to review the Solr Wiki documentation on plugins at [SolrPlugins](#).

If you have a plugin you would like to use, and you are running in SolrCloud mode, you can use the Blob Store API and the Config API to load the jars to Solr. The commands to use are described in the section [Adding Custom Plugins in SolrCloud Mode](#).

## Adding Custom Plugins in SolrCloud Mode

In SolrCloud mode, custom plugins need to be shared across all nodes of the cluster.

When running Solr in SolrCloud mode and you want to use custom code (such as custom analyzers, tokenizers, query parsers, and other plugins), it can be cumbersome to add jars to the classpath on all nodes in your cluster. Using the [Blob Store API](#) and special commands with the [Config API](#), you can upload jars to a special system-level collection and dynamically load plugins from them at runtime without needing to restart any nodes.

*This Feature is Disabled By Default*

In addition to requiring that Solr is running in [SolrCloud](#) mode, this feature is also disabled by default unless all Solr nodes are run with the `-Denable.runtime.lib=true` option on startup.

Before enabling this feature, users should carefully consider the issues discussed in the [Securing Runtime Libraries](#) section below.



## Uploading Jar Files

The first step is to use the [Blob Store API](#) to upload your jar files. This will put your jars in the `.system` collection and distribute them across your SolrCloud nodes. These jars are added to a separate classloader and only accessible to components that are configured with the property `runtimeLib=true`. These components are loaded lazily because the `.system` collection may not be loaded when a particular core is loaded.

## Config API Commands to use Jars as Runtime Libraries

The runtime library feature uses a special set of commands for the [Config API](#) to add, update, or remove jar files currently available in the blob store to the list of runtime libraries.

The following commands are used to manage runtime libs:

- `add-runtimelib`
- `update-runtimelib`

- delete-runtimelib

## V1 API

```
curl http://localhost:8983/solr/techproducts/config -H 'Content-type:application/json' -d '{
 "add-runtimelib": { "name":"jarblobname", "version":2 },
 "update-runtimelib": { "name":"jarblobname", "version":3 },
 "delete-runtimelib": "jarblobname"
}'
```

## V2 API

```
curl http://localhost:8983/api/collections/techproducts/config -H 'Content-type:application/json' -d '{
 "add-runtimelib": { "name":"jarblobname", "version":2 },
 "update-runtimelib": { "name":"jarblobname", "version":3 },
 "delete-runtimelib": "jarblobname"
}'
```

The name to use is the name of the blob that you specified when you uploaded your jar to the blob store. You should also include the version of the jar found in the blob store that you want to use. These details are added to configoverlay.json.

The default SolrResourceLoader does not have visibility to the jars that have been defined as runtime libraries. There is a classloader that can access these jars which is made available only to those components which are specially annotated.

Every pluggable component can have an optional extra attribute called runtimeLib=true, which means that the components are not loaded at core load time. Instead, they will be loaded on demand. If all the dependent jars are not available when the component is loaded, an error is thrown.

This example shows creating a ValueSourceParser using a jar that has been loaded to the Blob store.

## V1 API

```
curl http://localhost:8983/solr/techproducts/config -H 'Content-type:application/json' -d '{
 "create-valuesourceparser": {
 "name": "nvl",
 "runtimeLib": true,
 "class": "solr.org.apache.solr.search.function.NvlValueSourceParser,
 "nvlFloatValue": 0.0 }
}'
```

## V2 API

```
curl http://localhost:8983/api/collections/techproducts/config -H 'Content-type:application/json' -d '{
 "create-valuesourceparser": {
 "name": "nvl",
 "runtimeLib": true,
 "class": "solr.org.apache.solr.search.function.NvlValueSourceParser",
 "nvlFloatValue": 0.0
 }'
```

## Securing Runtime Libraries

A drawback of this feature is that it could be used to load malicious executable code into the system. However, it is possible to restrict the system to load only trusted jars using [PKI](#) to verify that the executables loaded into the system are trustworthy.

The following steps will allow you enable security for this feature. The instructions assume you have started all your Solr nodes with the `-Denable.runtime.lib=true`.

### Step 1: Generate an RSA Private Key

The first step is to generate an RSA private key. The example below uses a 512-bit key, but you should use the strength appropriate to your needs.

```
$ openssl genrsa -out priv_key.pem 512
```

### Step 2: Output the Public Key

The public portion of the key should be output in DER format so Java can read it.

```
$ openssl rsa -in priv_key.pem -pubout -outform DER -out pub_key.der
```

### Step 3: Load the Key to ZooKeeper

The .der files that are output from Step 2 should then be loaded to ZooKeeper under a node `/keys/exe` so they are available throughout every node. You can load any number of public keys to that node and all are valid. If a key is removed from the directory, the signatures of that key will cease to be valid. So, before removing the a key, make sure to update your runtime library configurations with valid signatures with the `update-runtimelib` command.

At the current time, you can only use the ZooKeeper `zkCli.sh` (or `zkCli.cmd` on Windows) script to issue these commands (the Solr version has the same name, but is not the same). If you have your own ZooKeeper ensemble running already, you can find the script in `$ZK_INSTALL/bin/zkCli.sh` (or `zkCli.cmd` if you are using Windows).



If you are running the embedded ZooKeeper that is included with Solr, you **do not** have this script already; in order to use it, you will need to download a copy of ZooKeeper v3.4.14 from <http://zookeeper.apache.org/>. Don't worry about configuring the download, you're just trying to get the command line utility script. When you start the script, you will connect to the embedded ZooKeeper.

To load the keys, you will need to connect to ZooKeeper with zkCli.sh, create the directories, and then create the key file, as in the following example.

```
Connect to ZooKeeper
Replace the server location below with the correct ZooKeeper connect string for your
installation.
$.bin/zkCli.sh -server localhost:9983

After connection, you will interact with the ZK prompt.
Create the directories
[zk: localhost:9983(CONNECTED) 5] create /keys
[zk: localhost:9983(CONNECTED) 5] create /keys/exe

Now create the public key file in ZooKeeper
The second path is the path to the .der file on your local machine
[zk: localhost:9983(CONNECTED) 5] create /keys/exe/pub_key.der /myLocal/pathTo/pub_key.der
```

After this, any attempt to load a jar will fail. All your jars must be signed with one of your private keys for Solr to trust it. The process to sign your jars and use the signature is outlined in Steps 4-6.

#### Step 4: Sign the jar File

Next you need to sign the sha1 digest of your jar file and get the base64 string.

```
$ openssl dgst -sha1 -sign priv_key.pem myjar.jar | openssl enc -base64
```

The output of this step will be a string that you will need to add the jar to your classpath in Step 6 below.

#### Step 5: Load the jar to the Blob Store

Load your jar to the Blob store, using the [Blob Store API](#). This step does not require a signature; you will need the signature in Step 6 to add it to your classpath.

```
curl -X POST -H 'Content-Type: application/octet-stream' --data-binary @filename
http://localhost:8983/solr/.system/blob/{blobname}
```

The blob name that you give the jar file in this step will be used as the name in the next step.

#### Step 6: Add the jar to the Classpath

Finally, add the jar to the classpath using the Config API as detailed above. In this step, you will need to provide the signature of the jar that you got in Step 4.

## V1 API

```
curl http://localhost:8983/solr/techproducts/config -H 'Content-type:application/json' -d '{
 "add-runtimelib": {
 "name": "blobname",
 "version": 2,

 "sig": "mW1Gwtz2QazjfVdrLFHfbGwcr8xzFYgU0Lu68LhqWRDvLG0uLcy1McQ+AzVmeZFBf1yLPDEHBWJb5KXr8bdbHN
 /

 PYgUB1nsr9pk4EFyD9KfJ8TqeH/ijQ9waa/vjqyiKEI9U550EtSzruLVZ32wJ7smvV0fj2YYhrUaaPzOn9g0=" }'
}'
```

## V2 API

```
curl http://localhost:8983/api/collections/techproducts/config -H 'Content-type:application/json' -d '{
 "add-runtimelib": {
 "name": "blobname",
 "version": 2,

 "sig": "mW1Gwtz2QazjfVdrLFHfbGwcr8xzFYgU0Lu68LhqWRDvLG0uLcy1McQ+AzVmeZFBf1yLPDEHBWJb5KXr8bdbHN
 /

 PYgUB1nsr9pk4EFyD9KfJ8TqeH/ijQ9waa/vjqyiKEI9U550EtSzruLVZ32wJ7smvV0fj2YYhrUaaPzOn9g0=" }'
}'
```

# JVM Settings

Optimizing the JVM can be a key factor in getting the most from your Solr installation.

Configuring your JVM can be a complex topic and a full discussion is beyond the scope of this document. Luckily, most modern JVMs are quite good at making the best use of available resources with default settings. The following sections contain a few tips that may be helpful when the defaults are not optimal for your situation.

For more general information about improving Solr performance, see <https://wiki.apache.org/solr/SolrPerformanceFactors>.

## Choosing Memory Heap Settings

The most important JVM configuration settings are those that determine the amount of memory it is allowed to allocate. There are two primary command-line options that set memory limits for the JVM. These are `-Xms`, which sets the initial size of the JVM's memory heap, and `-Xmx`, which sets the maximum size to which the heap is allowed to grow.

If your Solr application requires more heap space than you specify with the `-Xms` option, the heap will grow automatically. It's quite reasonable to not specify an initial size and let the heap grow as needed. The only downside is a somewhat slower startup time since the application will take longer to initialize. Setting the initial heap size higher than the default may avoid a series of heap expansions, which often results in objects being shuffled around within the heap, as the application spins up.

The maximum heap size, set with `-Xmx`, is more critical. If the memory heap grows to this size, object creation may begin to fail and throw `OutOfMemoryException`. Setting this limit too low can cause spurious errors in your application, but setting it too high can be detrimental as well.

It doesn't always cause an error when the heap reaches the maximum size. Before an error is raised, the JVM will first try to reclaim any available space that already exists in the heap. Only if all garbage collection attempts fail will your application see an exception. As long as the maximum is big enough, your app will run without error, but it may run more slowly if forced garbage collection kicks in frequently.

The larger the heap the longer it takes to do garbage collection. This can mean minor, random pauses or, in extreme cases, "freeze the world" pauses of a minute or more. As a practical matter, this can become a serious problem for heap sizes that exceed about two gigabytes, even if far more physical memory is available. On robust hardware, you may get better results running multiple JVMs, rather than just one with a large memory heap. Some specialized JVM implementations may have customized garbage collection algorithms that do better with large heaps. Consult your JVM vendor's documentation.

When setting the maximum heap size, be careful not to let the JVM consume all available physical memory. If the JVM process space grows too large, the operating system will start swapping it, which will severely impact performance. In addition, the operating system uses memory space not allocated to processes for file system cache and other purposes. This is especially important for I/O-intensive applications, like Lucene/Solr. The larger your indexes, the more you will benefit from filesystem caching by the OS. It may require some experimentation to determine the optimal tradeoff between heap space for the JVM and memory space for the OS to use.

On systems with many CPUs/cores, it can also be beneficial to tune the layout of the heap and/or the behavior of the garbage collector. Adjusting the relative sizes of the generational pools in the heap can affect how often GC sweeps occur and whether they run concurrently. Configuring the various settings of how the garbage collector should behave can greatly reduce the overall performance impact when it does run. There is a lot of good information on this topic available on Sun's website. A good place to start is here: [Oracle's Java HotSpot Garbage Collection](#).

## Use the Server HotSpot VM

If you are using Sun's JVM, add the `-server` command-line option when you start Solr. This tells the JVM that it should optimize for a long running, server process. If the Java runtime on your system is a JRE, rather than a full JDK distribution (including `javac` and other development tools), then it is possible that it may not support the `-server` JVM option. Test this by running `java -help` and look for `-server` as an available option in the displayed usage message.

## Checking JVM Settings

A great way to see what JVM settings your server is using, along with other useful information, is to use the admin RequestHandler, `solr/admin/system`. This request handler will display a wealth of server statistics and settings.

You can also use any of the tools that are compatible with the Java Management Extensions (JMX). See the section [Using JMX with Solr](#) for more information.

## v2 API

The v2 API is a modernized self-documenting API interface covering most current Solr APIs. It is anticipated that once the v2 API reaches full coverage, and Solr-internal API usages like SolrJ and the Admin UI have been converted from the old API to the v2 API, the old API will eventually be retired.

For now the two API styles will coexist, and all the old APIs will continue to work without any change. You can disable all v2 API endpoints by starting your servers with this system property: `-Ddisable.v2.api=true`.

The old API and the v2 API differ in three principle ways:

1. Command format: The old API commands and associated parameters are provided through URL request parameters on HTTP GET requests, while in the v2 API most API commands are provided via a JSON body POST'ed to v2 API endpoints. The v2 API also supports HTTP methods GET and DELETE where appropriate.
2. Endpoint structure: The v2 API endpoint structure has been rationalized and regularized.
3. Documentation: The v2 APIs are self-documenting: append `/_introspect` to any valid v2 API path and the API specification will be returned in JSON format.

## v2 API Path Prefixes

Following are some v2 API URL paths and path prefixes, along with some of the operations that are supported at these paths and their sub-paths.

Path prefix	Some Supported Operations
<code>/api/collections</code> or equivalently: <code>/api/c</code>	Create, alias, backup, and restore a collection.
<code>/api/c/collection-name/update</code>	Update requests.
<code>/api/c/collection-name/config</code>	Configuration requests.
<code>/api/c/collection-name/schema</code>	Schema requests.
<code>/api/c/collection-name/handler-name</code>	Handler-specific requests.
<code>/api/c/collection-name/shards</code>	Split a shard, create a shard, add a replica.
<code>/api/c/collection-name/shards/shard-name</code>	Delete a shard, force leader election
<code>/api/c/collection-name/shards/shard-name/replica-name</code>	Delete a replica.
<code>/api/cores</code>	Create a core.
<code>/api/cores/core-name</code>	Reload, rename, delete, and unload a core.
<code>/api/node</code>	Perform overseer operation, rejoin leader election.
<code>/api/cluster</code>	Add role, remove role, set cluster property.
<code>/api/c/.system/blob</code>	Upload and download blobs and metadata.

## Introspect

Append /\_introspect to any valid v2 API path and the API specification will be returned in JSON format.

`http://localhost:8983/api/c/_introspect`

To limit the introspect output to include just one particular HTTP method, add the request parameter `method` with value GET, POST, or DELETE.

`http://localhost:8983/api/c/_introspect?method=POST`

Most endpoints support commands provided in a body sent via POST. To limit the introspect output to only one command, add the request parameter `command=command-name`.

`http://localhost:8983/api/c/gettingstarted/_introspect?method=POST&command=modify`

## Interpreting the Introspect Output

Example: `http://localhost:8983/api/c/gettingstarted/get/_introspect`

```
{
 "spec": [{}
 "documentation": "https://lucene.apache.org/solr/guide/real-time-get.html",
 "description": "RealTime Get allows retrieving documents by ID before the documents have
 been committed to the index. It is useful when you need access to documents as soon as they are
 indexed but your commit times are high for other reasons.",
 "methods": ["GET"],
 "url": {
 "paths": ["/c/gettingstarted/get"],
 "params": {
 "id": {
 "type": "string",
 "description": "A single document ID to retrieve."},
 "ids": {
 "type": "string",
 "description": "One or more document IDs to retrieve. Separate by commas if more than
 one ID is specified."},
 "fq": {
 "type": "string",
 "description": "An optional filter query to add to the query. One use case for this is
 security filtering, in case users or groups should not be able to retrieve the document ID
 requested."}}}],
 "WARNING": "This response format is experimental. It is likely to change in the future.",
 "availableSubPaths": {}
 }
}
```

Description of some of the keys in the above example:

- **documentation:** URL to the online Solr reference guide section for this API
- **description:** A text description of the feature/variable/command, etc.
- **spec/methods:** HTTP methods supported by this API

- **spec/url/paths**: URL paths supported by this API
- **spec/url/params**: List of supported URL request params
- **availableSubPaths**: List of valid URL subpaths and the HTTP method(s) each supports

Example of introspect for a POST API:

`http://localhost:8983/api/c/gettingstarted/_introspect?method=POST&command=modify`

```
{
 "spec": [{}
 "documentation": "https://lucene.apache.org/solr/guide/collections-api.html",
 "description": "Several collection-level operations are supported with this endpoint: modify collection attributes; reload a collection; migrate documents to a different collection; rebalance collection leaders; balance properties across shards; and add or delete a replica property.",
 "methods": ["POST"],
 "url": {"paths": ["/collections/{collection}",
 "/c/{collection}"]},
 "commands": {"modify": {
 "documentation": "https://lucene.apache.org/solr/guide/collections-api.html#modifycollection",
 "description": "Modifies specific attributes of a collection. Multiple attributes can be changed at one time.",
 "type": "object",
 "properties": {
 "rule": {
 "type": "array",
 "documentation": "https://lucene.apache.org/solr/guide/rule-based-replica-placement.html",
 "description": "Modifies the rules for where replicas should be located in a cluster.",
 "items": {"type": "string"}},
 "snitch": {
 "type": "array",
 "documentation": "https://lucene.apache.org/solr/guide/rule-based-replica-placement.html",
 "description": "Details of the snitch provider",
 "items": {"type": "string"}},
 "autoAddReplicas": {
 "type": "boolean",
 "description": "When set to true, enables auto addition of replicas on shared file systems (such as HDFS). See https://lucene.apache.org/solr/guide/running-solr-on-hdfs.html for more details on settings and overrides."},
 "replicationFactor": {
 "type": "string",
 "description": "The number of replicas to be created for each shard. Replicas are physical copies of each shard, acting as failover for the shard. Note that changing this value on an existing collection does not automatically add more replicas to the collection. However, it will allow add-replica commands to succeed."},
 "maxShardsPerNode": {
 "type": "integer",
 "description": "When creating collections, the shards and/or replicas are spread
```

```
across all available, live, nodes, and two replicas of the same shard will never be on the same node. If a node is not live when the collection is created, it will not get any parts of the new collection, which could lead to too many replicas being created on a single live node. Defining maxShardsPerNode sets a limit on the number of replicas can be spread to each node. If the entire collection can not be fit into the live nodes, no collection will be created at all."}}]}],
 "WARNING": "This response format is experimental. It is likely to change in the future.",
 "availableSubPaths": {
 "/c/gettingstarted/select": ["POST", "GET"],
 "/c/gettingstarted/config": ["POST", "GET"],
 "/c/gettingstarted/schema": ["POST", "GET"],
 "/c/gettingstarted/export": ["POST", "GET"],
 "/c/gettingstarted/admin/ping": ["POST", "GET"],
 "/c/gettingstarted/update": ["POST"]
 }
}
```

The "commands" section in the above example has one entry for each command supported at this endpoint. The key is the command name and the value is a JSON object describing the command structure using JSON schema (see <http://json-schema.org/> for a description).

## Invocation Examples

For the "gettingstarted" collection, set the replication factor and whether to automatically add replicas (see above for the introspect output for the "modify" command used here):

```
$ curl http://localhost:8983/api/c/gettingstarted -H 'Content-type:application/json' -d '
{ modify: { replicationFactor: "3", autoAddReplicas: false } }'

{"responseHeader": {"status":0,"QTime":842}}
```

See the state of the cluster:

```
$ curl http://localhost:8983/api/cluster

{"responseHeader": {"status":0,"QTime":0}, "collections": ["gettingstarted", ".system"]}
```

Set a cluster property:

```
$ curl http://localhost:8983/api/cluster -H 'Content-type: application/json' -d '
{ set-property: { name: autoAddReplicas, val: "false" } }'

{"responseHeader": {"status":0,"QTime":4}}
```

# Monitoring Solr

Administration and monitoring can be performed using the web-based administration console, through the command line interface, or using REST APIs.

Common administrative tasks include:

[Metrics Reporting](#): Details of Solr's metrics registries and Metrics API.

[Metrics History](#): Metrics history collection, configuration and API.

[MBean Request Handler](#): How to use Solr's MBeans for programmatic access to the system plugins and stats.

[Configuring Logging](#): Describes how to configure logging for Solr.

[Using JMX with Solr](#): Describes how to use Java Management Extensions with Solr.

[Monitoring Solr with Prometheus and Grafana](#): Describes how to monitor Solr with Prometheus and Grafana.

[Performance Statistics Reference](#): Additional information on statistics returned from JMX.

# Metrics Reporting

Solr includes a developer API and instrumentation for the collection of detailed performance-oriented metrics throughout the life-cycle of Solr service and its various components.

Internally this feature uses the [Dropwizard Metrics API](#), which uses the following classes of meters to measure events:

- **counters** - simply count events. They provide a single long value, e.g., the number of requests.
- **meters** - additionally compute rates of events. Provide a count (as above) and 1-, 5-, and 15-minute exponentially decaying rates, similar to the Unix system load average.
- **histograms** - calculate approximate distribution of events according to their values. Provide the following approximate statistics, with a similar exponential decay as above: mean (arithmetic average), median, maximum, minimum, standard deviation, and 75<sup>th</sup>, 95<sup>th</sup>, 98<sup>th</sup>, 99<sup>th</sup> and 999<sup>th</sup> percentiles.
- **timers** - measure the number and duration of events. They provide a count and histogram of timings.
- **gauges** - offer instantaneous reading of a current value, e.g., current queue depth, current number of active connections, free heap size.

Each group of related metrics with unique names is managed in a **metric registry**. Solr maintains several such registries, each corresponding to a high-level group such as: jvm, jetty, node, and core (see [Metric Registries](#) below).

For each group (and/or for each registry) there can be several **reporters**, which are components responsible for communication of metrics from selected registries to external systems. Currently implemented reporters support emitting metrics via JMX, Ganglia, Graphite and SLF4J.

There is also a dedicated /admin/metrics handler that can be queried to report all or a subset of the current metrics from multiple registries.

## Metric Registries

Solr includes multiple metric registries, which group related metrics.

Metrics are maintained and accumulated through all lifecycles of components from the start of the process until its shutdown - e.g., metrics for a particular SolrCore are tracked through possibly several load, unload and/or rename operations, and are deleted only when a core is explicitly deleted. However, metrics are not persisted across process restarts; restarting Solr will discard all collected metrics.

These are the major groups of metrics that are collected:

### JVM Registry

This registry is returned at solr.jvm and includes the following information. When making requests with the [Metrics API](#), you can specify &group=jvm to limit to only these metrics.

- direct and mapped buffer pools
- class loading / unloading
- OS memory, CPU time, file descriptors, swap, system load

- GC count and time
- heap, non-heap memory and GC pools
- number of threads, their states and deadlocks
- System properties such as Java information, various installation directory paths, ports, and similar information. You can control what appears here by modifying solr.xml.

## Node / CoreContainer Registry

This registry is returned at solr.node and includes the following information. When making requests with the [Metrics API](#), you can specify &group=node to limit to only these metrics.

- handler requests (count, timing): collections, info, admin, configsets, etc.
- number of cores (loaded, lazy, unloaded)

## Core (SolrCore) Registry

The [Core \(SolrCore\) Registry](#) includes solr.core.<collection>, one for each core. When making requests with the [Metrics API](#), you can specify &group=core to limit to only these metrics.

- all common RequestHandlers report: request timers / counters, timeouts, errors. Handlers that support process distributed shard requests also report shardRequests sub-counters for each type of distributed request.
- [index-level events](#): meters for minor / major merges, number of merged docs, number of deleted docs, gauges for currently running merges and their size.
- shard replication and transaction log replay on replicas,
- open / available / pending connections for shard handler and update handler.

## Jetty Registry

This registry is returned at solr.jetty and includes the following information. When making requests with the [Metrics API](#), you can specify &group=jetty to limit to only these metrics.

- threads and pools,
- connection and request timers,
- meters for responses by HTTP class (1xx, 2xx, etc.)

In the future, metrics will be added for shard leaders and cluster nodes, including aggregations from per-core metrics.

## Metrics Configuration

The metrics available in your system can be customized by modifying the <metrics> element in solr.xml.



See also the section [Format of Solr.xml](#) for more information about the solr.xml file, where to find it, and how to edit it.

## The <metrics><hiddenSysProps> Element

This section of solr.xml allows you to define the system properties which are considered system-sensitive and should not be exposed via the Metrics API.

If this section is not defined, the following default configuration is used which hides password and authentication information:

```
<metrics>
 <hiddenSysProps>
 <str>javax.net.ssl.keyStorePassword</str>
 <str>javax.net.ssl.trustStorePassword</str>
 <str>basicauth</str>
 <str>zkDigestPassword</str>
 <str>zkDigestReadOnlyPassword</str>
 </hiddenSysProps>
</metrics>
```

## The <metrics><reporters> Element

Reporters consume the metrics data generated by Solr. See the section [Reporters](#) below for more details on how to configure custom reporters.

## The <metrics><suppliers> Element

Suppliers help Solr generate metrics data. The <metrics><suppliers> section of solr.xml allows you to define your own implementations of metrics and configure parameters for them.

Implementation of a custom metrics supplier is beyond the scope of this guide, but there are other customizations possible with the default implementation, via the elements described below.

### <counter>

This element defines the implementation and configuration of a Counter supplier. The default implementation does not support any configuration.

### <meter>

This element defines the implementation of a Meter supplier. The default implementation supports an additional parameter:

```
<str name="clock">
```

The type of clock to use for calculating EWMA rates. The supported values are:

- user, the default, which uses System.nanoTime()
- cpu, which uses the current thread's CPU time

### <histogram>

This element defines the implementation of a Histogram supplier. This element also supports the clock parameter shown above with the meter element, and also:

```
<str name="reservoir">
```

The fully-qualified class name of the Reservoir implementation to use. The default is com.codahale.metrics.ExponentiallyDecayingReservoir but there are other options available with the [Codahale Metrics library](#) that Solr uses. The following parameters are supported, within the mentioned limitations:

- size, the reservoir size. The default is 1028.
- alpha, the decay parameter. The default is 0.015. This is only valid for the ExponentiallyDecayingReservoir.
- window, the window size, in seconds, and only valid for the SlidingTimeWindowReservoir. The default is 300 (5 minutes).

#### <timer>

This element defines an implementation of a Timer supplier. The default implementation supports the clock and reservoir parameters described above.

As an example of a section of solr.xml that defines some of these custom parameters, the following defines the default Meter supplier with a non-default clock and the default Timer is used with a non-default reservoir:

```
<metrics>
 <suppliers>
 <meter>
 <str name="clock">cpu</str>
 </meter>
 <timer>
 <str name="reservoir">com.codahale.metrics.SlidingTimeWindowReservoir</str>
 <long name="window">600</long>
 </timer>
 </suppliers>
</metrics>
```

## Reporters

Reporter configurations are specified in solr.xml file in <metrics><reporter> sections, for example:

```
<solr>
 <metrics>
 <reporter name="graphite" group="node, jvm" class=
"org.apache.solr.metrics.reporters.SolrGraphiteReporter">
 <str name="host">graphite-server</str>
 <int name="port">9999</int>
 <int name="period">60</int>
 </reporter>
 <reporter name="log_metrics" group="core" class=
"org.apache.solr.metrics.reporters.SolrSlf4jReporter">
 <int name="period">60</int>
 <str name="filter">QUERY./select.requestTimes</str>
 <str name="filter">QUERY./get.requestTimes</str>
 <str name="filter">UPDATE./update.requestTimes</str>
 <str name="filter">UPDATE./update.clientErrors</str>
 <str name="filter">UPDATE./update.errors</str>
 <str name="filter">SEARCHER.new.time</str>
 <str name="filter">SEARCHER.new.warmup</str>
 <str name="logger">org.apache.solr.metrics.reporters.SolrSlf4jReporter</str>
 </reporter>
 </metrics>
...
</solr>
```

This example configures two reporters: [Graphite](#) and [SLF4J](#). See below for more details on how to configure reporters.

## Reporter Arguments

Reporter plugins use the following arguments:

**name**

The unique name of the reporter plugin (required).

**class**

The fully-qualified implementation class of the plugin, which must extend `SolrMetricReporter` (required).

**group**

One or more of the predefined groups (see above).

**registry**

One or more of valid fully-qualified registry names.

If both `group` and `registry` attributes are specified only the `group` attribute is considered. If neither attribute is specified then the plugin will be used for all groups and registries. Multiple group or registry names can be specified, separated by comma and/or space.

Additionally, several implementation-specific initialization arguments can be specified in nested elements. There are some arguments that are common to SLF4J, Ganglia and Graphite reporters:

period

The period in seconds between reports. Default value is 60.

prefix

A prefix to be added to metric names, which may be helpful in logical grouping of related Solr instances, e.g., machine name or cluster name. Default is empty string, i.e., just the registry name and metric name will be used to form a fully-qualified metric name.

filter

If not empty then only metric names that start with this value will be reported. Default is no filtering, i.e., all metrics from the selected registry will be reported.

Reporters are instantiated for every group and registry that they were configured for, at the time when the respective components are initialized (e.g., on JVM startup or SolrCore load).

When reporters are created their configuration is validated (and e.g., necessary connections are established). Uncaught errors at this initialization stage cause the reporter to be discarded from the running configuration.

Reporters are closed when the corresponding component is being closed (e.g., on SolrCore close, or JVM shutdown) but metrics that they reported are still maintained in respective registries, as explained in the previous section.

The following sections provide information on implementation-specific arguments. All implementation classes provided with Solr can be found under `org.apache.solr.metrics.reporters`.

## JMX Reporter

The JMX Reporter uses the `org.apache.solr.metrics.reporters.SolrJmxReporter` class.

It takes the following arguments:

domain

The JMX domain name. If not specified then the registry name will be used.

serviceUrl

The service URL for a JMX server. If not specified, Solr will attempt to discover if the JVM has an MBean server and will use that address. See below for additional information on this.

agentId

The agent ID for a JMX server. Note either `serviceUrl` or `agentId` can be specified but not both - if both are specified then the default MBean server will be used.

Object names created by this reporter are hierarchical, dot-separated but also properly structured to form corresponding hierarchies in e.g., JConsole. This hierarchy consists of the following elements in the top-down order:

- registry name (e.g., `solr.core.collection1.shard1.replica1`). Dot-separated registry names are also split into ObjectName hierarchy levels, so that metrics for this registry will be shown under `/solr/core/collection1/shard1/replica1` in JConsole, with each domain part being assigned to `dom1`, `dom2`, ... `domN` property.

- reporter name (the value of reporter's name attribute)
- category, scope and name for request handlers
- or additional name1, name2, ... nameN elements for metrics from other components.

The JMX Reporter replaces the JMX functionality available in Solr versions before 7.0. If you have upgraded from an earlier version and have an MBean Server running when Solr starts, Solr will automatically discover the location of the local MBean server and use a default configuration for the SolrJmxReporter.

You can start a local MBean server with a system property at startup by adding -Dcom.sun.management.jmxremote to your start command. This will not add the reporter configuration to solr.xml, so if you enable it with a system property, you must always start Solr with the system property or JMX will not be enabled in subsequent starts.

## SLF4J Reporter

The SLF4J Reporter uses the org.apache.solr.metrics.reporters.SolrSlf4jReporter class.

It takes the following arguments, in addition to common arguments described [above](#).

### logger

The name of the logger to use. Default is empty, in which case the group (or the initial part of the registry name that identifies a metrics group) will be used if specified in the plugin configuration.

Users can specify logger name (and the corresponding logger configuration in e.g., Log4j configuration) to output metrics-related logging to separate file(s), which can then be processed by external applications. Here is an example for configuring the default log4j2.xml which ships in Solr. This can be used in conjunction with the solr.xml example provided earlier in this page to configure the SolrSlf4jReporter:

```

<Configuration>
 <Appenders>
 ...
 <RollingFile
 name="MetricsFile"
 fileName="${sys:solr.log.dir}/solr_metrics.log"
 filePattern="${sys:solr.log.dir}/solr_metrics.log.%i" >
 <PatternLayout>
 <Pattern>
 %d{yyyy-MM-dd HH:mm:ss.SSS} %-5p (%t) [%X{node_name} %X{collection} %X{shard}
 %X{replica} %X{core}] %m%n
 </Pattern>
 </PatternLayout>
 <Policies>
 <OnStartupTriggeringPolicy />
 <SizeBasedTriggeringPolicy size="32 MB"/>
 </Policies>
 <DefaultRolloverStrategy max="10"/>
 </RollingFile>
 ...
 </Appenders>

 <Loggers>
 ...
 <Logger name="org.apache.solr.metrics.reporters.SolrSlf4jReporter" level="info" additivity=
 "false">
 <AppenderRef ref="MetricsFile"/>
 </Logger>
 ...
 </Loggers>
</Configuration>

```

Each log line produced by this reporter consists of configuration-specific fields, and a message that follows this format:

```

type=COUNTER, name={}, count={}

type=GAUGE, name={}, value={}

type=TIMER, name={}, count={}, min={}, max={}, mean={}, stddev={}, median={}, p75={}, p95={},
p98={}, p99={}, p999={}, mean_rate={}, m1={}, m5={}, m15={}, rate_unit={}, duration_unit={}

type=METER, name={}, count={}, mean_rate={}, m1={}, m5={}, m15={}, rate_unit={}

type=HISTOGRAM, name={}, count={}, min={}, max={}, mean={}, stddev={}, median={}, p75={}, p95={},
p98={}, p99={}, p999={}

```

(curly braces added here only as placeholders for actual values).

Additionally, the following MDC context properties are passed to the logger and can be used in log formats:

node\_name

Solr node name (for SolrCloud deployments, otherwise null), prefixed with n::.

registry

Metric registry name, prefixed with m::.

For reporters that are specific to a SolrCore also the following properties are available:

collection

Collection name, prefixed with c::.

shard

Shard name, prefixed with s::.

replica

Replica name (core node name), prefixed with r::.

core

SolrCore name, prefixed with x::.

tag

Reporter instance tag, prefixed with t::.

## Graphite Reporter

The [Graphite](#) Reporter uses the org.apache.solr.metrics.reporters.SolrGraphiteReporter class.

It takes the following attributes, in addition to the common attributes [above](#).

host

The host name where Graphite server is running (required).

port

The port number for the server (required).

pickled

If true, use "pickled" Graphite protocol which may be more efficient. Default is false (use plain-text protocol).

When plain-text protocol is used (pickled==false) it's possible to use this reporter to integrate with systems other than Graphite, if they can accept space-separated and line-oriented input over network in the following format:

```
dot.separated.metric.name[.and.attribute] value epochTimestamp
```

For example:

```
example.solr.node.cores.lazy 0 1482932097
example.solr.node.cores.loaded 1 1482932097
example.solr.jetty.org.eclipse.jetty.server.handler.DefaultHandler.2xx-responses.count 21
1482932097
example.solr.jetty.org.eclipse.jetty.server.handler.DefaultHandler.2xx-responses.m1_rate
2.5474287707930614 1482932097
example.solr.jetty.org.eclipse.jetty.server.handler.DefaultHandler.2xx-responses.m5_rate
3.8003171557510305 1482932097
example.solr.jetty.org.eclipse.jetty.server.handler.DefaultHandler.2xx-responses.m15_rate
4.0623076220244245 1482932097
example.solr.jetty.org.eclipse.jetty.server.handler.DefaultHandler.2xx-responses.mean_rate
0.5698031798408144 1482932097
```

## Ganglia Reporter

The [Ganglia](#) reporter uses the `org.apache.solr.metrics.reporters.SolrGangliaReporter` class.

It takes the following arguments, in addition to the common arguments [above](#).

`host`

The host name where Ganglia server is running (required).

`port`

The port number for the server.

`multicast`

When true use multicast UDP communication, otherwise use UDP unicast. Default is false.

## Shard and Cluster Reporters

These two reporters can be used for aggregation of metrics reported from replicas to shard leader (the "shard" reporter), and from any local registry to the Overseer node.

Metric reports from these reporters are periodically sent as batches of regular `SolrInputDocuments`, so they can be processed by any Solr handler. By default they are sent to `/admin/metrics/collector` handler (an instance of `MetricsCollectorHandler`) on a target node, which aggregates these reports and keeps them in additional local metric registries so that they can be accessed using `/admin/metrics` handler, and re-reported elsewhere as necessary.

In case of shard reporter the target node is the shard leader, in case of cluster reporter the target node is the Overseer leader.

### Shard Reporter

This reporter uses predefined shard group, and the implementing class must be (a subclass of) `solr.SolrShardReporter`. It publishes selected metrics from replicas to the node where shard leader is located. Reports use a target registry name that is the replica's registry name with a `.leader` suffix, e.g., for a `SolrCore` name `collection1_shard1_replica_n3` the target registry name is `solr.core.collection1.shard1.replica_n3.leader`.

The following configuration properties are supported:

**handler**

The handler path where reports are sent. Default is /admin/metrics/collector.

**period**

How often reports are sent, in seconds. Default is 60. Setting this to 0 disables the reporter.

**filter**

An optional regular expression(s) matching selected metrics to be reported.

The following filter expressions are used by default:

```
TLOG.*
CORE\.\fs\.*
REPLICATION\.*
INDEX\.\flush\.*
INDEX\.\merge\.\major\.*
UPDATE\./update\.*requests
QUERY\./select\.*requests
```

Example configuration:

```
<reporter name="test" group="shard" class="solr.SolrShardReporter">
 <int name="period">11</int>
 <str name="filter">UPDATE\./update\.*requests</str>
 <str name="filter">QUERY\./select\.*requests</str>
</reporter>
```

## Cluster Reporter

This reporter uses predefined cluster group and the implementing class must be (a subclass of) solr.SolrClusterReporter. It publishes selected metrics from any local registry to the Overseer leader node.

The following configuration properties are supported:

**handler**

The handler path where reports are sent. Default is /admin/metrics/collector.

**period**

How often reports are sent, in seconds. Default is 60. Setting this to 0 disables the reporter.

**report**

report configuration(s), see below.

Each report configuration consists of the following properties:

**registry**

A regular expression pattern matching local source registries (see SolrMetricManager.registryNames(String...)), may contain regex capture groups (required).

**group**

The target registry name where metrics will be grouped. This can be a regular expression pattern that contains back-references to capture groups collected by registry pattern (required).

**label**

An optional prefix to prepend to metric names, may contain back-references to capture groups collected by registry pattern.

**filter**

An optional regular expression(s) matching selected metrics to be reported.

The following report specifications are used by default (their result is a single additional metric registry in Overseer, called solr.cluster):

```
<lst name="report">
 <str name="group">cluster</str>
 <str name="registry">solr\.jetty</str>
 <str name="label">jetty</str>
</lst>
<lst name="report">
 <str name="group">cluster</str>
 <str name="registry">solr\.node</str>
 <str name="label">node</str>
 <str name="filter">CONTAINER\cores\..*</str>
 <str name="filter">CONTAINER\fs\..*</str>
</lst>
<lst name="report">
 <str name="group">cluster</str>
 <str name="label">jvm</str>
 <str name="registry">solr\.jvm</str>
 <str name="filter">memory\total\..*</str>
 <str name="filter">memory\heap\..*</str>
 <str name="filter">os\SystemLoadAverage</str>
 <str name="filter">os\FreePhysicalMemorySize</str>
 <str name="filter">os\FreeSwapSpaceSize</str>
 <str name="filter">os\OpenFileDescriptorCount</str>
 <str name="filter">threads\count</str>
</lst>
<lst name="report">
 <str name="group">cluster</str>
 <str name="registry">solr\core\(.*)\leader</str>
 <str name="label">leader.\$1</str>
 <str name="filter">QUERY\./select/.*</str>
 <str name="filter">UPDATE\./update/.*</str>
 <str name="filter">INDEX\..*</str>
 <str name="filter">TLOG\..*</str>
</lst>
```

Example configuration:

```
<reporter name="test" group="cluster" class="solr.SolrClusterReporter">
 <str name="handler"/>/admin/metrics/collector</str>
 <int name="period">11</int>
 <lst name="report">
 <str name="group">aggregated_jvms</str>
 <str name="label">jvm</str>
 <str name="registry">solr\.jvm</str>
 <str name="filter">memory\.total\...*</str>
 <str name="filter">memory\.heap\...*</str>
 <str name="filter">os\.SystemLoadAverage</str>
 <str name="filter">threads\.count</str>
 </lst>
 <lst name="report">
 <str name="group">aggregated_shard_leaders</str>
 <str name="registry">solr\.core\.(.*)\.leader</str>
 <str name="label">leader.$1</str>
 <str name="filter">UPDATE\.update/.*</str>
 </lst>
</reporter>
```

## Core Level Metrics

These metrics are available only on a per-core basis. Metrics can be aggregated across cores using Shard and Cluster reporters.

## Index Merge Metrics

These metrics are collected in respective registries for each core (e.g., solr.core.collection1....), under the INDEX category.

Basic metrics are always collected - collection of additional metrics can be turned on using boolean parameters in the /config/indexConfig/metrics section of solrconfig.xml:

```
<config>
 ...
 <indexConfig>
 <metrics>
 <majorMergeDocs>524288</majorMergeDocs>
 <bool name="mergeDetails">true</bool>
 </metrics>
 ...
 </indexConfig>
 ...
</config>
```

The following metrics are collected:

- INDEX.merge.major - timer for merge operations that include at least "majorMergeDocs" (default value for this parameter is 512k documents).

- INDEX.merge.minor - timer for merge operations that include less than "majorMergeDocs".
- INDEX.merge.errors - counter for merge errors.
- INDEX.flush - meter for index flush operations.

Additionally, the following gauges are reported, which help to monitor the momentary state of index merge operations:

- INDEX.merge.major.running - number of running major merge operations (depending on the implementation of MergeScheduler that is used there can be several concurrently running merge operations).
- INDEX.merge.minor.running - as above, for minor merge operations.
- INDEX.merge.major.running.docs - total number of documents in the segments being currently merged in major merge operations.
- INDEX.merge.minor.running.docs - as above, for minor merge operations.
- INDEX.merge.major.running.segments - number of segments being currently merged in major merge operations.
- INDEX.merge.minor.running.segments - as above, for minor merge operations.

If the boolean flag `mergeDetails` is true then the following additional metrics are collected:

- INDEX.merge.major.docs - meter for the number of documents merged in major merge operations
- INDEX.merge.major.deletedDocs - meter for the number of deleted documents expunged in major merge operations

## Metrics API

The `admin/metrics` endpoint provides access to all the metrics for all metric groups.

A few query parameters are available to limit your request to only certain metrics:

`group`

The metric group to retrieve. The default is `all` to retrieve all metrics for all groups. Other possible values are: `jvm`, `jetty`, `node`, and `core`. More than one group can be specified in a request; multiple group names should be separated by a comma.

`type`

The type of metric to retrieve. The default is `all` to retrieve all metric types. Other possible values are `counter`, `gauge`, `histogram`, `meter`, and `timer`. More than one type can be specified in a request; multiple types should be separated by a comma.

`prefix`

The first characters of metric name that will filter the metrics returned to those starting with the provided string. It can be combined with `group` and/or `type` parameters. More than one prefix can be specified in a request; multiple prefixes should be separated by a comma. Prefix matching is also case-sensitive.

`regex`

A regular expression matching metric names. Note: dot separators in metric names must be escaped, e.g., `QUERY\./select\..*` is a valid regex that matches all metrics with the `QUERY./select.` prefix.

## property

Allows requesting only this metric from any compound metric. Multiple property parameters can be combined to act as an OR request. For example, to only get the 99th and 999th percentile values from all metric types and groups, you can add &property=p99\_ms&property=p999\_ms to your request. This can be combined with group, type, and prefix as necessary.

## key

fully-qualified metric name, which specifies one concrete metric instance (parameter can be specified multiple times to retrieve multiple concrete metrics). **NOTE: when this parameter is used, other selection methods listed above are ignored.** Fully-qualified name consists of registry name, colon and metric name, with optional colon and metric property. Colons in names can be escaped using back-slash \ character. Examples:

- key=solr.node:CONTAINER.fs.totalSpace
- key=solr.core.collection1:QUERY./select.requestTimes:max\_ms
- key=solr.jvm:system.properties:user.name

## compact

When false, a more verbose format of the response will be returned. Instead of a response like this:

```
{"metrics": [
 "solr.core.gettingstarted",
 {
 "CORE.aliases": {
 "value": ["gettingstarted"]
 },
 "CORE.coreName": {
 "value": "gettingstarted"
 },
 "CORE.indexDir": {
 "value": "/solr/example/schemaless/solr/gettingstarted/data/index/"
 },
 "CORE.instanceDir": {
 "value": "/solr/example/schemaless/solr/gettingstarted"
 },
 "CORE.refCount": {
 "value": 1
 },
 "CORE.startTime": {
 "value": "2017-03-14T11:43:23.822Z"
 }
 }
]}
```

The response will look like this:

```
{"metrics": [
 "solr.core.gettingstarted",
 {
 "CORE.aliases": [
 "gettingstarted"
],
 "CORE.coreName": "gettingstarted",
 "CORE.indexDir": "/solr/example/schemaless/solr/gettingstarted/data/index/",
 "CORE.instanceDir": "/solr/example/schemaless/solr/gettingstarted",
 "CORE.refCount": 1,
 "CORE.startTime": "2017-03-14T11:43:23.822Z"
 }
]}
```

Like other request handlers, the Metrics API can also take the `wt` parameter to define the output format.

## Examples

Request only "counter" type metrics in the "core" group, returned in JSON:

```
http://localhost:8983/solr/admin/metrics?type=counter&group=core
```

Request only "core" group metrics that start with "INDEX", returned in XML:

```
http://localhost:8983/solr/admin/metrics?wt=xml&prefix=INDEX&group=core
```

Request only "core" group metrics that end with ".requests":

```
http://localhost:8983/solr/admin/metrics?regex=.*\\.requests&group=core
```

Request only "user.name" property of "system.properties" metric from registry "solr.jvm":

```
http://localhost:8983/solr/admin/metrics?wt=xml?key=solr.jvm:system.properties:user.name
```

# Metrics History

Solr collects long-term history of certain key metrics both in SolrCloud and in standalone mode.

This information can be used for very simple monitoring and troubleshooting, but also some SolrCloud components (e.g., autoscaling) can use this data for making informed decisions based on long-term trends of selected metrics.



The `.system` collection must exist if metrics history should be persisted. If this collection is absent then metrics history will still be collected and kept in memory but it will be lost on node restart.

## Design

Before discussing how to configure metrics storage, a bit of explanation about how it works may be helpful.

### Round-Robin Databases

The metrics history data is maintained as multi-resolution time series, with a fixed total number of data points per metric history (a fixed size window). Multi-resolution refers to the fact that data from the most detailed time series is periodically resampled to create coarser-grained time series, which in turn are periodically resampled again to build even coarser-grained series.

In the default configuration selected metrics are sampled every 60 seconds, and the following time series are built:

- 240 samples, every 60 sec (4 hours)
- 288 samples, every 600 sec (48 hours)
- 336 samples, every 1h (2 weeks)
- 180 samples, every 4h (2 months)
- 365 samples, every 1 day (1 year)

This means that the total number of samples in all data series is constant, and consequently the size of this data structure is also constant (because the size of the moving window is fixed, and older samples are replaced by newer ones). This arrangement is referred to as a round-robin database, and Solr uses implementation of this concept provided by the [RRD4j](#) library.

## Storage

Databases created with RRD4j are compact - for the time series specified above the total size of data is around 11kB for each of the primary time series, including its resampled data. Each database may contain several primary time series ("datasources" in RRD4j parlance) and their re-sampled versions (called "archives").

This data is updated in memory and then periodically stored in the `.system` collection in the form of Solr documents with a binary `data_bin` field, each document containing data of one full database. This method of storage is much more compact and generates less update operations than storing each data point in a separate Solr document. The Metrics History API allows retrieving detailed data from each database,

including retrieval of all individual datapoints.

Databases are identified primarily by their corresponding metric registry name, so for databases that keep track of aggregated metrics this will be e.g., solr.jvm, solr.node, solr.collection.gettingstarted. For databases with non-aggregated metrics the name consists of the registry name, optionally with a node name to identify databases with the same name coming from different nodes. For example, per-node databases are named like this: solr.jvm.localhost:8983\_solr, solr.node.localhost:7574\_solr, but per-replica names are already unique across the cluster so they are named like this: solr.core.gettingstarted.shard1.replica\_n1.

## Collected Metrics

Currently the following selected metrics are tracked:

- Non-aggregated solr.core and aggregated solr.collection metrics:
  - QUERY./select.requests
  - UPDATE./update.requests
  - INDEX.sizeInBytes
  - numShards (aggregated, active shards)
  - numReplicas (aggregated, active replicas)
- solr.node metrics:
  - CONTAINER.fs.coreRoot.usableSpace
  - numNodes (aggregated, number of live nodes)
- solr.jvm metrics:
  - memory.heap.used
  - os.processCpuLoad
  - os.systemLoadAverage

Separate databases are created for each of these groups, and each database keeps data for all metrics listed in that group.



Currently this list is not configurable. Also, if you change this list in the code then all existing databases must be first removed from the .system collection because RRD4j doesn't allow adding new datasources once the database is created.

## SolrRrdBackendFactory

This component is responsible for managing in-memory databases and periodically saving them to the .system collection. If the .system collection is not available the updates to the databases will be kept in memory, until the time when .system collection becomes available.

If the .system collection is permanently unavailable then data will not be saved and it will be lost when the Solr node is shut down.

## MetricsHistoryHandler

This component provides a REST API for accessing the metrics history. It is also responsible for collecting and periodically updating the in-memory databases.

This handler also performs aggregation of metrics on per-collection level, and on a cluster level. By default only these aggregated metrics are tracked - historic data from each node and each replica in each collection is not collected separately. Aggregated databases are managed on the Overseer leader node but they are still accessible from other nodes even if they are not persisted - the handler redirects the call from originating node to the current Overseer leader.

The handler assumes that a simple aggregation (sum of partial metric values from each resource) is sufficient. This happens to make sense for the default built-in sets of metrics. Future extensions will provide other aggregation strategies (such as, average, max, min, etc.).

## Metrics History Configuration

There are two ways to configure this subsystem:

- `/clusterprops.json` - this is the primary mechanism. It uses the cluster properties JSON file in ZooKeeper. Configuration is stored in the `/metrics/history` element in a JSON map.
- `solr.xml` - this is the secondary mechanism, which is not recommended but provided for consistency with the existing metrics configuration section in this file. Configuration is stored in the `/solr/metrics/history` element of this file.

Currently the following configuration options are supported:

`enable`

boolean, default is `true`. If this is `false` then metrics history is not collected but can still be retrieved from existing databases. When this is `true` then metrics are periodically collected, aggregated and saved.

`enableReplicas`

boolean, default is `false`. When this is `true` non-aggregated history will be collected for each replica in each collection. When this is `false` then only aggregated history is collected for each collection.

`enableNodes`

boolean, default is `false`. When this is `true` then non-aggregated history will be collected separately for each node (for node and JVM metrics), with database names consisting of base registry name with appended node name, e.g., `solr.jvm.localhost:8983_solr`. When this is `false` then only aggregated history will be collected in a single `solr.jvm` and `solr.node` cluster-wide databases.

`collectPeriod`

integer, in seconds, default is 60. Metrics values will be collected and respective databases updated every `collectPeriod` seconds.



Value of `collectPeriod` must be at least 1, and if it's changed then all previously existing databases with their historic data must be manually removed (new databases will be created automatically).

## syncPeriod

integer, in seconds, default is 60. Data from modified databases will be saved to Solr every syncPeriod seconds. When accessing the databases via REST API in index mode the visibility of most recent data depends on this period, because requests accessing the data from other nodes see only the version of the data that is stored in the .system collection.

## Example Configuration

Example /clusterprops.json file with metrics history configuration that turns on the collection of per-node metrics history for node and JVM metrics. Typically this file will also contain other properties unrelated to Metrics History API.

```
{
 "metrics" : {
 "history" : {
 "enable" : true,
 "enableNodes" : true,
 "syncPeriod" : 300
 }
 }
}
```

## Metrics History API

Main entry point for accessing metrics history is /admin/metrics/history (or /api/cluster/metrics/history for v2 API).

The following sections describe actions available in this API. All calls have at least one required parameter action.

All responses contain a section named state, which reports the current internal state of the API:

### enableReplicas

boolean, corresponds to the enableReplicas configuration setting.

### enableNodes

boolean, corresponds to the enableNodes configuration setting.

### mode

one of the following values:

- inactive - when metrics collection is disabled (but access to existing metrics history is still available).
- memory - when metrics history is kept only in memory because .system collection doesn't exist. In this mode clients can access metrics history available on the node that received the request and on the Overseer leader.
- index - when metrics history is periodically stored in the .system collection. Data available in memory on the node that accepted the request is retrieved from memory, any other data is retrieved from the .system collection (so it's at least syncPeriod old).

Also, the response header section (`responseHeader`) contains `zkConnected` boolean property that indicates whether the current node is a part of SolrCloud cluster.

## List Databases

The query parameter `action=list` produces a list of available databases. It supports the following parameters:

`rows`

optional integer, default is 500. Maximum number of results to return.

Example: In this SolrCloud example the API is in memory mode, and the request was made to a node that is not Overseer leader. The API transparently forwarded the request to Overseer leader.

```
curl http://localhost:7574/solr/admin/metrics/history?action=list&rows=10
```

```
{
 "responseHeader": {
 "zkConnected": true,
 "status": 0,
 "QTime": 9
 },
 "metrics": {
 "solr.collection..system": {
 "lastModified": 1528360138,
 "node": "127.0.0.1:8983_solr"
 },
 "solr.collection.gettingstarted": {
 "lastModified": 1528360138,
 "node": "127.0.0.1:8983_solr"
 },
 "solr.jvm": {
 "lastModified": 1528360138,
 "node": "127.0.0.1:8983_solr"
 },
 "solr.node": {
 "lastModified": 1528360138,
 "node": "127.0.0.1:8983_solr"
 }
 },
 "state": {
 "enableReplicas": false,
 "enableNodes": false,
 "mode": "memory"
 }
}
```

Note the presence of the `node` element in each section, which shows where the information is coming from - when API is in memory mode this indicates which results are local and which ones are retrieved from the

Overseer leader node. When the API is in index mode this element always shows the node name that received the request (because the data is retrieved from the .system collection anyway).

Each section also contains a lastModified element, which contains the last modification time when the database was update. All timestamps returned from this API correspond to Unix epoch time in seconds.

## Database Status

The query parameter action=status provides detailed status of the selected database.

The following parameters are supported:

name

string, required: database name.

Example:

```
curl
http://localhost:7574/solr/admin/metrics/history?action=status&name=solr.collection.gettingstarte
d
```

```
{
 "responseHeader": {
 "zkConnected": true,
 "status": 0,
 "QTime": 46
 },
 "metrics": {
 "solr.collection.gettingstarted": {
 "status": {
 "lastModified": 1528318361,
 "step": 60,
 "datasourceCount": 5,
 "archiveCount": 5,
 "datasourceNames": [
 "numShards",
 "numReplicas",
 "QUERY./select.requests",
 "UPDATE./update.requests",
 "INDEX.sizeInBytes"
],
 "datasources": [
 {
 "datasource": "DS:numShards:GAUGE:120:U:U",
 "lastValue": 2
 },
 {
 "datasource": "DS:numReplicas:GAUGE:120:U:U",
 "lastValue": 4
 },
 "..."
]
 }
 }
 }
}
```

```
],
 "archives": [
 {
 "archive": "RRA:AVERAGE:0.5:1:240",
 "steps": 1,
 "consolFun": "AVERAGE",
 "xff": 0.5,
 "startTime": 1528303980,
 "endTime": 1528318320,
 "rows": 240
 },
 {
 "archive": "RRA:AVERAGE:0.5:10:288",
 "steps": 10,
 "consolFun": "AVERAGE",
 "xff": 0.5,
 "startTime": 1528146000,
 "endTime": 1528318200,
 "rows": 288
 },
 ...
]
 },
 "node": "127.0.0.1:7574_solr"
},
{
 "state": {
 "enableReplicas": false,
 "enableNodes": false,
 "mode": "index"
 }
}
```

## Get Database Data

The query parameter `action=get` retrieves all data collected in the specified database.

The following parameters are supported:

`name`

string, required: database name.

`format`

string, optional, default is `list`. Format of the data. Currently the following formats are supported:

- `list` - each datapoint is returned as separate JSON element. For efficiency, for each datasource in a database for each time series the timestamps are provided separately from values (because points from all datasources in a given time series share the same timestamps).
- `string` - all datapoint values and timestamps are returned as strings, with values separated by new line character.

- graph - data is returned as PNG images, Base64-encoded, containing graphs of each time series values over time.

In each case the response is structured in a similar way: archive identifiers are keys in a JSON map, all data is placed in a data element, with timestamps / datapoints / graphs as values in lists or maps.

## Examples

This is the output using the default list format:

```
curl
http://localhost:8983/solr/admin/metrics/history?action=get&name=solr.collection.gettingstarted
```

```
{
 "responseHeader": {
 "zkConnected": true,
 "status": 0,
 "QTime": 4
 },
 "metrics": {
 "solr.collection.gettingstarted": {
 "data": {
 "RRA:AVERAGE:0.5:1:240": {
 "timestamps": [
 1528304160,
 1528304220,
 "..."
],
 "values": {
 "numShards": [
 "NaN",
 2.0,
 "..."
],
 "numReplicas": [
 "NaN",
 4.0,
 "..."
],
 }
 }
 },
 "RRA:AVERAGE:0.5:10:288": {
 "timestamps": [
 1528145400,
 1528146000,
 ...
],
 "lastModified": 1528318606,
 "node": "127.0.0.1:8983_solr"
 }
 },
 "state": {
 "enableReplicas": false,
 "enableNodes": false,
 "mode": "index"
 }
 }
 }
}
```

This is the output when using the string format:

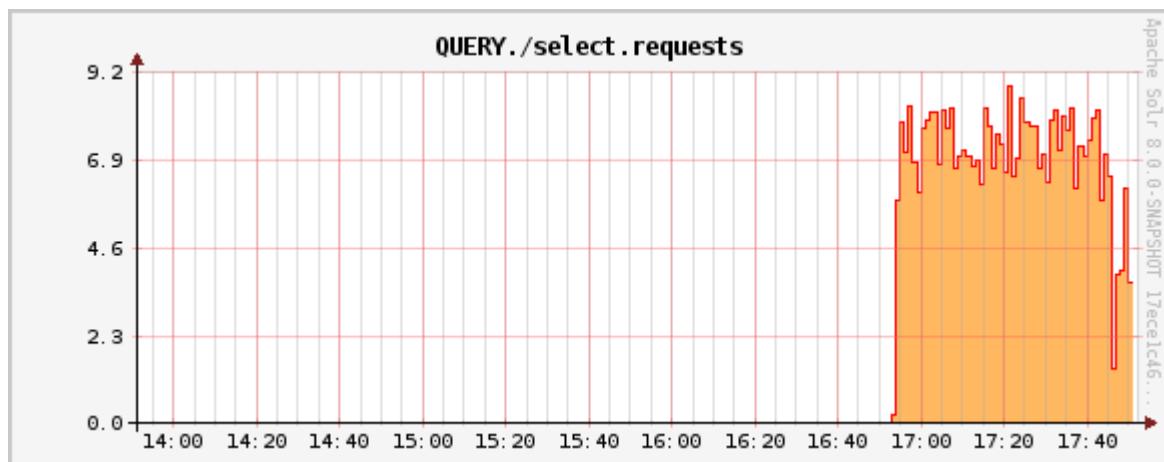
```
curl
http://localhost:8983/solr/admin/metrics/history?action=get&name=solr.collection.gettingstarted&f
ormat=string
```

```
{
 "responseHeader": {
 "zkConnected": true,
 "status": 0,
 "QTime": 2
 },
 "metrics": {
 "solr.collection.gettingstarted": {
 "data": {
 "RRA:AVERAGE:0.5:1:240": {
 "timestamps": "1527254820\n1527254880\n1527254940\n...",
 "values": {
 "numShards": "NaN\n2.0\n2.0\n2.0\n2.0\n2.0\n2.0\n...",
 "numReplicas": "NaN\n4.0\n4.0\n4.0\n4.0\n4.0\n4.0\n...",
 "QUERY./select.requests": "NaN\n123\n456\n789\n...",
 "..."
 }
 },
 "RRA:AVERAGE:0.5:10:288": {
 "..."
 }
 }
 }
 }
}
```

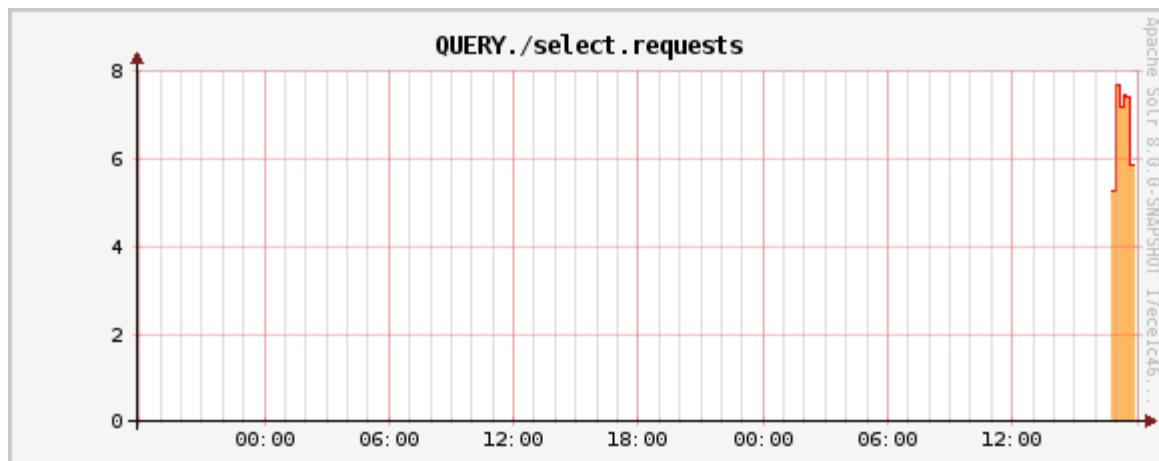
This is the output when using the graph format:

```
curl
http://localhost:8983/solr/admin/metrics/history?action=get&name=solr.collection.gettingstarted&f
ormat=graph
```

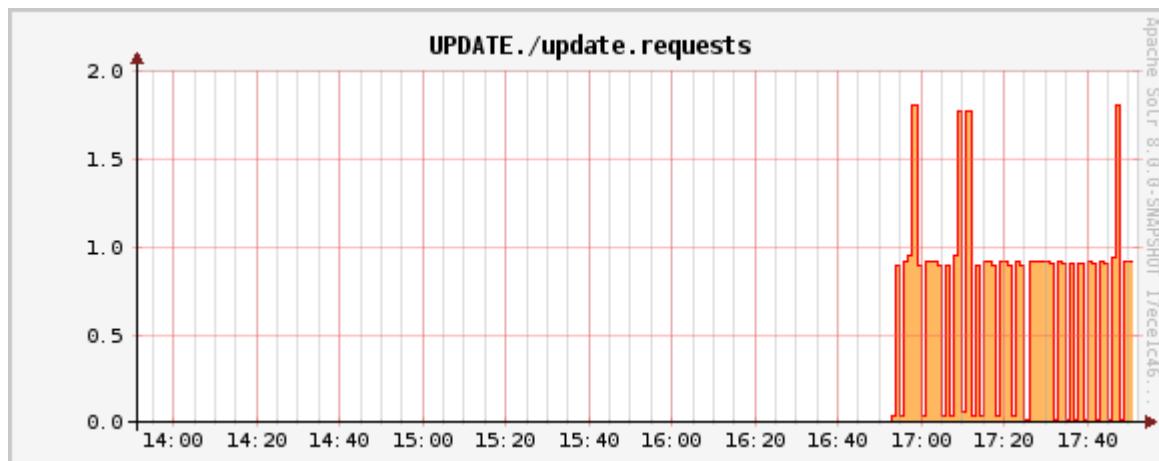
```
{
 "responseHeader": {
 "zkConnected": true,
 "status": 0,
 "QTime": 2
 },
 "metrics": {
 "solr.collection.gettingstarted": {
 "data": {
 "RRA:AVERAGE:0.5:1:240": {
 "values": {
 "numShards": "iVBORw0KGgoAAAANSUhEUgAAkQAA... ",
 "numReplicas": "iVBORw0KGgoAAAANSUhEUgAAkQA... ",
 "QUERY./select.requests": "iVBORw0KGgoAAAANS... ",
 "... "
 }
 },
 "RRA:AVERAGE:0.5:10:288": {
 "values": {
 "numShards": "iVBORw0KGgoAAAANSUhEUgAAkQAA... ",
 "... "
 }
 }
 }
 }
 }
}
```



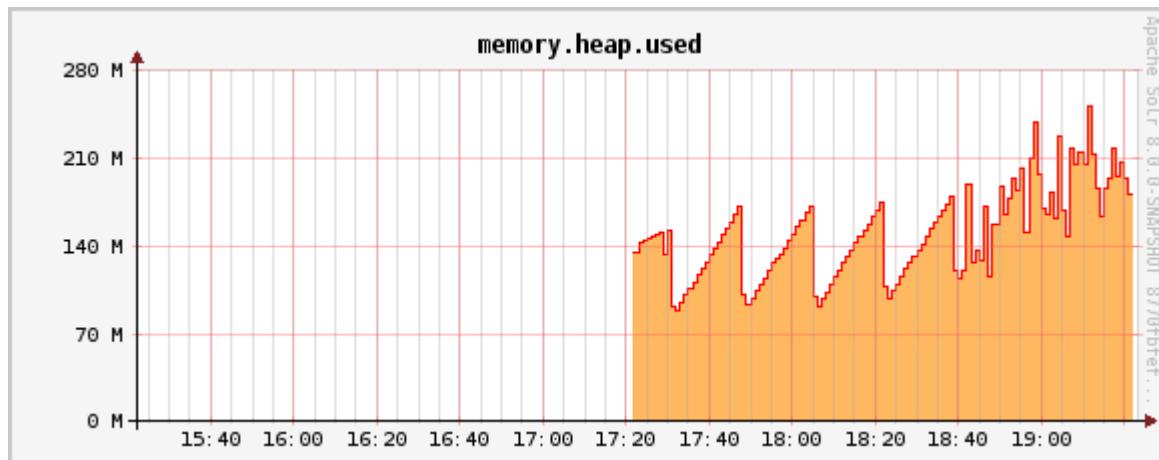
Example 60 sec resolution history graph for QUERY./select.requests metric



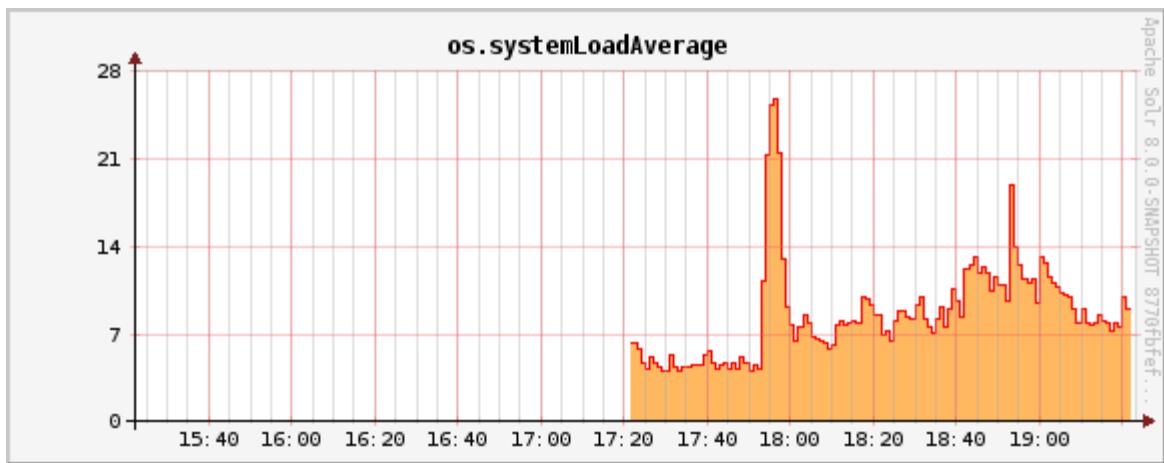
Example 10 min resolution history graph for `QUERY./select.requests` metric



Example 60 sec resolution history graph for `UPDATE./update.requests` metric



Example 60 sec resolution history graph for `memory.heap.used` metric



Example 60 sec resolution history graph for `os.systemLoadAverage` metric

# MBean Request Handler

The MBean Request Handler offers programmatic access to the information provided on the [Plugin/Stats](#) page of the Admin UI.

The MBean Request Handler accepts the following parameters:

key

Restricts results by object key.

cat

Restricts results by category name.

stats

Specifies whether statistics are returned with results. You can override the stats parameter on a per-field basis. The default is false.

wt

The output format. This operates the same as the wt [parameter in a query](#). The default is json.

## MBeanRequestHandler Examples

The following examples assume you are running Solr's techproducts example configuration:

```
bin/solr start -e techproducts
```

To return information about the CACHE category only:

```
http://localhost:8983/solr/techproducts/admin/mbeans?cat=CACHE
```

To return information and statistics about the CACHE category only, formatted in XML:

```
http://localhost:8983/solr/techproducts/admin/mbeans?stats=true&cat=CACHE&wt=xml
```

To return information for everything, and statistics for everything except the fieldCache:

```
http://localhost:8983/solr/techproducts/admin/mbeans?stats=true&f.fieldCache.stats=false
```

To return information and statistics for the fieldCache only:

```
http://localhost:8983/solr/techproducts/admin/mbeans?key=fieldCache&stats=true
```

# Configuring Logging

Solr logs are a key way to know what's happening in the system. There are several ways to adjust the default logging configuration.



In addition to the logging options described below, there is a way to configure which request parameters (such as parameters sent as part of queries) are logged with an additional request parameter called `logParamsList`. See the section on [Common Query Parameters](#) for more information.

## Temporary Logging Settings

You can control the amount of logging output in Solr by using the Admin Web interface. Select the **LOGGING** link. Note that this page only lets you change settings in the running system and is not saved for the next run. (For more information about the Admin Web interface, see [Using the Solr Administration User Interface](#).)

Time (Local)	Level	Core	Logger	Message
4/13/2016, 2:46:57 PM	ERROR	false	RequestHandlerBase	org.apache.solr.common.SolrException: ERROR: [doc=ea2675db-c491-4340-af57-c2591524d393] Error adding field 'foo_i'='bar' msg=For input string: "bar"

### The Logging Screen

This part of the Admin Web interface allows you to set the logging level for many different log categories. Fortunately, any categories that are **unset** will have the logging level of its parent. This makes it possible to change many categories at once by adjusting the logging level of their parent.

When you select **Level**, you see the following menu:

The screenshot shows the Solr admin interface with the 'Level' option selected in the sidebar. The main content area displays a hierarchical log configuration tree. A context menu is open over the 'hadoop' node, listing various log levels: ALL, TRACE, DEBUG, INFO, WARN, ERROR, FATAL, OFF, and UNSET. The 'INFO' level is currently highlighted.

### The Log Level Menu

Directories are shown with their current logging levels. The Log Level Menu floats over these. To set a log level for a particular directory, select it and click the appropriate log level button.

Log levels settings are as follows:

Level	Result
FINEST	Reports everything.
FINE	Reports everything but the least important messages.
CONFIG	Reports configuration errors.
INFO	Reports everything but normal status.
WARN	Reports all warnings.
SEVERE	Reports only the most severe warnings.
OFF	Turns off logging.
UNSET	Removes the previous log setting.

Multiple settings at one time are allowed.

## LogLevel API

There is also a way of sending REST commands to the logging endpoint to do the same. Example:

```
Set the root logger to level WARN
curl -s http://localhost:8983/solr/admin/info/logging --data-binary "set=root:WARN"
```

## Choosing Log Level at Startup

You can temporarily choose a different logging level as you start Solr. There are two ways:

The first way is to set the SOLR\_LOG\_LEVEL environment variable before you start Solr, or place the same variable in bin/solr.in.sh or bin/solr.in.cmd. The variable must contain an uppercase string with a supported log level (see above).

The second way is to start Solr with the -v or -q options, see [Solr Control Script Reference](#) for details.

Examples:

```
Start with verbose (DEBUG) logging
bin/solr start -f -v
Start with quiet (WARN) logging
bin/solr start -f -q
```

## Permanent Logging Settings

Solr uses [Log4J version \\${ivyversions.org.log4j.major.version}](#) for logging which is configured using server/resources/log4j2.xml. Take a moment to inspect the contents of the log4j2.xml file so that you are familiar with its structure. By default, Solr log messages will be written to SOLR\_LOGS\_DIR/solr.log.

When you're ready to deploy Solr in production, set the variable SOLR\_LOGS\_DIR to the location where you want Solr to write log files, such as /var/solr/logs. You may also want to tweak log4j2.xml. Note that if you installed Solr as a service using the instructions provided in [Taking Solr to Production](#), then see /var/solr/log4j2.xml instead of the default server/resources version.

When starting Solr in the foreground (-f option), all logs will be sent to the console, in addition to solr.log. When starting Solr in the background, it will write all stdout and stderr output to a log file in solr-<port>-console.log, and automatically disable the CONSOLE logger configured in log4j2.xml, having the same effect as if you removed the CONSOLE appender from the rootLogger manually.

Also, in log4j2.xml if the default log rotation size threshold of 32MB is too small for production servers then you should increase it to a larger value (such as 100MB or more).

```
<SizeBasedTriggeringPolicy size="100 MB"/>
```

Java Garbage Collection logs are rotated by the JVM when size hits 20M, for a max of 9 generations.

On every startup or restart of Solr, log4j2 performs log rotation. If you choose to use another log framework that does not support rotation on startup, you may enable SOLR\_LOG\_PRESTART\_ROTATION in bin/solr.in.sh or bin/solr.in.cmd to let the start script rotate the logs on startup.

## Logging Slow Queries

For high-volume search applications, logging every query can generate a large amount of logs and, depending on the volume, potentially impact performance. If you mine these logs for additional insights into your application, then logging every query request may be useful.

On the other hand, if you're only concerned about warnings and error messages related to requests, then you can set the log verbosity to WARN. However, this poses a potential problem in that you won't know if any queries are slow, as slow queries are still logged at the INFO level.

Solr provides a way to set your log verbosity threshold to WARN and be able to set a latency threshold above which a request is considered "slow" and log that request at the WARN level to help you identify slow queries in your application. To enable this behavior, configure the `<slowQueryThresholdMillis>` element in the **query** section of `solrconfig.xml`:

```
<slowQueryThresholdMillis>1000</slowQueryThresholdMillis>
```

Any queries that take longer than the specified threshold will be logged as "slow" queries at the WARN level. The log file under which you can find all these queries is called `solr_slow_requests.log` and will be found in your `SOLR_LOGS_DIR` (see [Permanent Logging Settings](#) for more about defining log locations).

# Using JMX with Solr

Java Management Extensions (JMX) is a technology that makes it possible for complex systems to be controlled by tools without the systems and tools having any previous knowledge of each other. In essence, it is a standard interface by which complex systems can be viewed and manipulated.

Solr, like any other good citizen of the Java universe, can be controlled via a JMX interface. Once enabled, you can use a JMX client, like jconsole, to connect with Solr.

If you are unfamiliar with JMX, you may find the following overview useful: <http://docs.oracle.com/javase/8/docs/technotes/guides/management/agent.html>.

## Configuring JMX

JMX support is configured by defining a metrics reporter, as described in the section the section [JMX Reporter](#).

If you have an existing MBean server running in Solr's JVM, or if you start Solr with the system property `-Dcom.sun.management.jmxremote`, Solr will automatically identify its location on startup even if you have not defined a reporter explicitly in `solr.xml`. You can also define the location of the MBean server with parameters defined in the reporter definition.

## Configuring MBean Servers

Versions of Solr prior to 7.0 defined JMX support in `solrconfig.xml`. This has been changed to the metrics reporter configuration defined above. Parameters for the reporter configuration allow defining the location or address of an existing MBean server.

An MBean server can be started at the time of Solr's startup by passing the system parameter `-Dcom.sun.management.jmxremote`. See Oracle's documentation for additional settings available to start and control an MBean server at <http://docs.oracle.com/javase/8/docs/technotes/guides/management/agent.html>.

## Configuring a Remote Connection to Solr JMX

If you need to attach a JMX-enabled Java profiling tool, such as JConsole or VisualVM, to a remote Solr server, then you need to enable remote JMX access when starting the Solr server. Simply change the `ENABLE_REMOTE_JMX_OPTS` property in the `solr.in.sh` or `solr.in.cmd` (for Windows) file to true. You'll also need to choose a port for the JMX RMI connector to bind to, such as 18983. For example, if your Solr include script sets:

```
ENABLE_REMOTE_JMX_OPTS=true
RMI_PORT=18983
```

The JMX RMI connector will allow Java profiling tools to attach to port 18983. When enabled, the following properties are passed to the JVM when starting Solr:

```
-Dcom.sun.management.jmxremote \
-Dcom.sun.management.jmxremote.local.only=false \
-Dcom.sun.management.jmxremote.ssl=false \
-Dcom.sun.management.jmxremote.authenticate=false \
-Dcom.sun.management.jmxremote.port=18983 \
-Dcom.sun.management.jmxremote.rmi.port=18983
```

We don't recommend enabling remote JMX access in production, but it can sometimes be useful when doing performance and user-acceptance testing prior to going into production.

For more information about these settings, see: <http://docs.oracle.com/javase/8/docs/technotes/guides/management/agent.html>.



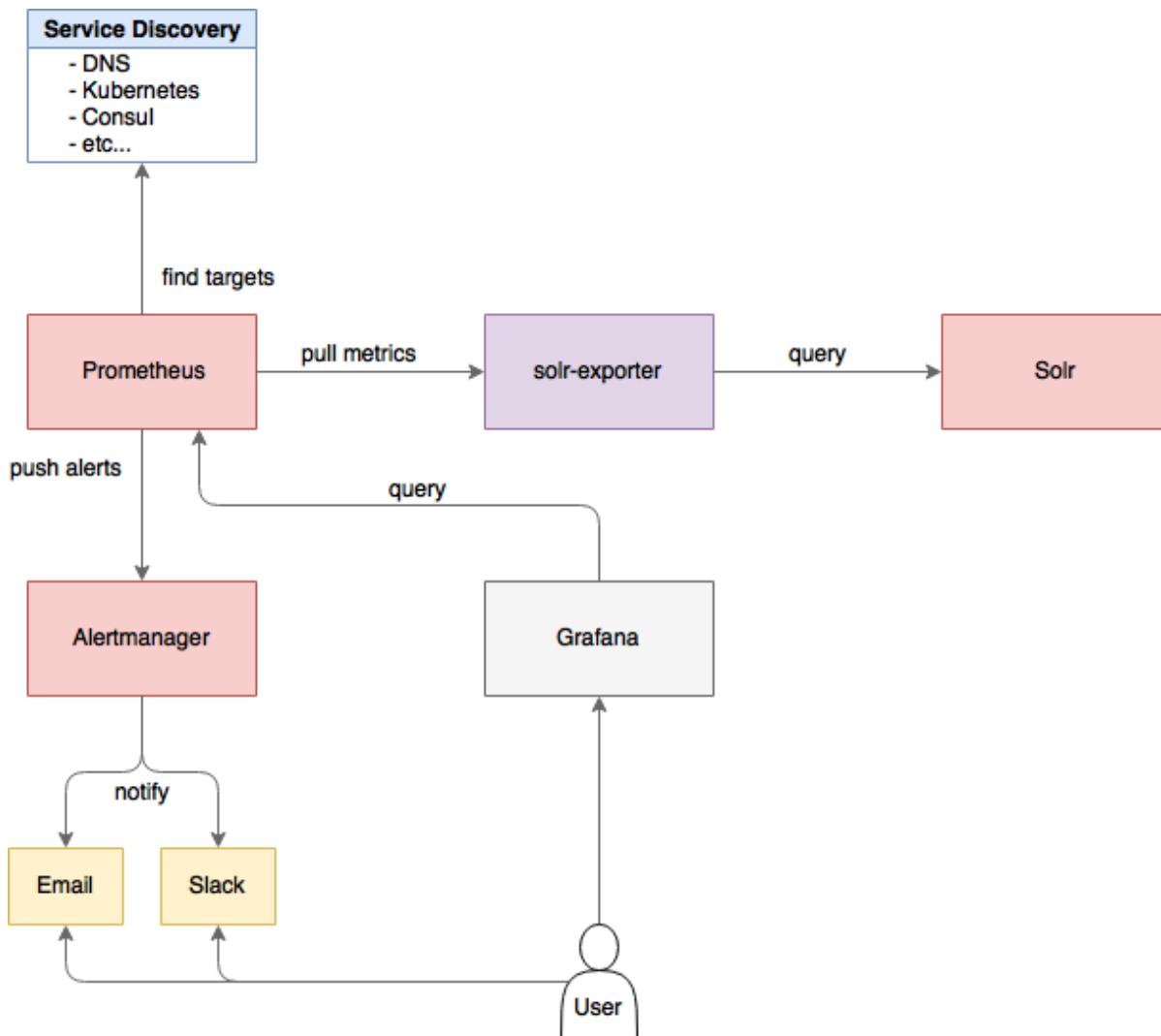
Making JMX connections into machines running behind NATs (e.g., Amazon's EC2 service) is not a simple task. The `java.rmi.server.hostname` system property may help, but running `jconsole` on the server itself and using a remote desktop is often the simplest solution. See <http://web.archive.org/web/20130525022506/http://jmsbrdy.com/monitoring-java-applications-running-on-ec2-i>.

# Monitoring Solr with Prometheus and Grafana

If you use [Prometheus](#) and [Grafana](#) for metrics storage and data visualization, Solr includes a Prometheus exporter to collect metrics and other data.

A Prometheus exporter (`solr-exporter`) allows users to monitor not only Solr metrics which come from [Metrics API](#), but also facet counts which come from [Searching](#) and responses to [Collections API](#) commands and [PingRequestHandler](#) requests.

This graphic provides a more detailed view:



*solr-exporter Diagram*



This feature is considered experimental, meaning future improvements may break compatibility.

The Prometheus exporter is included in Solr as a contrib, and is located in `contrib/prometheus-exporter` in your Solr instance.

There are three aspects to running `solr-exporter`:

- Modify the `solr-exporter-config.xml` to define the data to collect. Solr has a default configuration you can use, but if you would like to modify it before running the exporter the first time, see the section below [Exporter Configuration](#).
- Start the exporter from within Solr. See the section below [Starting the Exporter](#).
- Modify your Prometheus configuration to listen on the correct port. See the section below [Prometheus Configuration](#)

## Starting the Exporter

You can start `solr-exporter` by running `./bin/solr-exporter` (Linux) or `.\bin\solr-exporter.cmd` (Windows) from the `contrib/prometheus-exporter` directory.

See the commands below depending on your operating system and Solr operating mode:

### Linux

#### *Standalone mode*

```
$ cd contrib/prometheus-exporter
$./bin/solr-exporter -p 9854 -b http://localhost:8983/solr -f ./conf/solr-exporter-
config.xml -n 8
```

#### *SolrCloud mode*

```
$ cd contrib/prometheus-exporter
$./bin/solr-exporter -p 9854 -z localhost:2181/solr -f ./conf/solr-exporter-config.xml -n 16
```

### Windows

#### *Standalone mode*

```
> cd contrib\prometheus-exporter
> .\bin\solr-exporter.cmd -p 9854 -b http://localhost:8983/solr -f .\conf\solr-exporter-
config.xml -n 8
```

#### *SolrCloud mode*

```
> cd contrib\prometheus-exporter
> .\bin\solr-exporter -p 9854 -z localhost:2181/solr -f .\conf\solr-exporter-config.xml -n 16
```

## Command Line Parameters

The parameters in the example start commands shown above:

`h, --help`

Displays command line help and usage.

**-p, --port**

The port where Prometheus will listen for new data. This port will be used to configure Prometheus. It can be any port not already in use on your server. The default is 9983.

**-b, --baseurl**

The Solr base URL (such as `http://localhost:8983/solr`) when Solr is running in Standalone mode. If you are running Solr in SolrCloud mode, do not specify this parameter. If neither the `-b` parameter nor the `-z` parameter are defined, the default is `-b http://localhost:8983/solr`.

**-z, --zkhost**

The ZooKeeper connect string (such as `localhost:8983`, or `localhost:2181/solr`) when Solr is running in SolrCloud mode. If you are running Solr in Standalone mode, do not specify this parameter. If neither the `-b` parameter nor the `-z` parameter are defined, the `-b` parameter default is used.

**-f, --config-file**

The path to the configuration file that defines the Solr metrics to read. The default is `contrib/prometheus-exporter/conf/solr-exporter-config.xml`.

**-n, --num-threads**

The number of threads. The `solr-exporter` creates thread pools for requests to Solr. Request latency can be improved by increasing the number of threads. The default is 1.

**-s, --scrape-interval**

The number of seconds between collecting metrics from Solr. The `solr-exporter` collects metrics from Solr every few seconds controlled by this setting. These metrics are cached and returned regardless of how frequently Prometheus is configured to pull metrics from this tool. The freshness of the metrics can be improved by reducing the scrape interval but do not set it to a very low value because metrics collection can be expensive and can execute arbitrary searches to ping Solr. The default value is 60 seconds.

The Solr's metrics exposed by `solr-exporter` can be seen at:

`http://localhost:9983/solr/admin/metrics`.

## Getting metrics from a secure Solr(Cloud)

Your Solr(Cloud) might be secured by measures described in [Securing Solr](#). The security configuration can be injected into `solr-exporter` using environment variables in a fashion similar to other clients using [SolrJ](#). This is possible because the main script picks up two external environment variables and passes them on to the Java process:

- `JAVA_OPTS` allows to add extra JVM options
- `CLASSPATH_PREFIX` allows to add extra libraries

Example for a SolrCloud instance secured by [Basic Authentication](#), [SSL](#) and [ZooKeeper Access Control](#):

Suppose you have a file `basicauth.properties` with the Solr Basic-Auth credentials:

```
httpBasicAuthUser=myUser
httpBasicAuthPassword=myPassword
```

Then you can start the Exporter as follows (Linux).

```
$ cd contrib/prometheus-exporter
$ export JAVA_OPTS="-Djavax.net.ssl.trustStore=truststore.jks
-Djavax.net.ssl.trustStorePassword=truststorePassword
-Dsolr.httpclient.builder.factory=org.apache.solr.client.solrj.impl.PreemptiveBasicAuthClientBuilderFactory -Dsolr.httpclient.config=basicauth.properties
-DzkCredentialsProvider=org.apache.solr.common.cloud.VMParamsSingleSetCredentialsDigestZkCredentialsProvider -DzkDigestUsername=readonly-user -DzkDigestPassword=zkuUserPassword"
$ export CLASSPATH_PREFIX="../../server/solr-webapp/webapp/WEB-INF/lib/commons-codec-1.11.jar"
$./bin/solr-exporter -p 9854 -z zk1:2181,zk2:2181,zk3:2181 -f ./conf/solr-exporter-config.xml -n 16
```

Note: The Exporter needs the commons-codec library for SSL/BasicAuth, but does not bring it. Therefore the example reuses it from the Solr web app. Of course, you can use a different source.

## Exporter Configuration

The configuration for the solr-exporter defines the data to get from Solr. This includes the metrics, but can also include queries to the PingRequestHandler, the Collections API, and a query to any query request handler.

A default example configuration is in contrib/prometheus-exporter/config/solr-exporter-config.xml. Below is a slightly shortened version of it:

```
<config>

<rules>

<ping>
 <lst name="request">
 <lst name="query">
 <str name="path"/>/admin/ping</str>
 </lst>
 <arr name="jsonQueries">
 <str>
 . as $object | $object |
 (if $object.status == "OK" then 1.0 else 0.0 end) as $value |
 {
 name : "solr_ping",
 type : "GAUGE",
 help : "See following URL: https://lucene.apache.org/solr/guide/ping.html",
 label_names: [],
 label_values: [],
 value : $value
 }
 </str>
 </arr>
 </lst>
</ping>
```

```
</str>
</arr>
</lst>
</ping>

<metrics>
<lst name="request">
<lst name="query">
<str name="path"/>/admin/metrics</str>
<lst name="params">
<str name="group">all</str>
<str name="type">all</str>
<str name="prefix"></str>
<str name="property"></str>
</lst>
</lst>
<arr name="jsonQueries">
<!--
jetty metrics
-->
<str>
.metrics["solr.jetty"] | to_entries | .[] | select(.key |
startsWith("org.eclipse.jetty.server.handler.DefaultHandler")) | select(.key | endsWith("xx-
responses")) as $object |
$object.key | split(".") | last | split("-") | first as $status |
$value.count as $value |
{
 name : "solr_metrics_jetty_response_total",
 type : "COUNTER",
 help : "See following URL: https://lucene.apache.org/solr/guide/metrics-
reporting.html",
 label_names : ["status"],
 label_values : [$status],
 value : $value
}
</str>
...
</arr>
</lst>
</metrics>

<collections>
<lst name="request">
<lst name="query">
<str name="path"/>/admin/collections</str>
<lst name="params">
<str name="action">CLUSTERSTATUS</str>
</lst>
</lst>
<arr name="jsonQueries">
<str>
.cluster.live_nodes | length as $value|
```

```
{
 name : "solr_collections_live_nodes",
 type : "GAUGE",
 help : "See following URL:
https://lucene.apache.org/solr/guide/collections-api.html#clusterstatus",
 label_names : [],
 label_values : [],
 value : $value
}
</str>
...
 </arr>
</lst>
</collections>

<search>
 <lst name="request">
 <lst name="query">
 <str name="collection">collection1</str>
 <str name="path">/select</str>
 <lst name="params">
 <str name="q">*:*</str>
 <str name="start">0</str>
 <str name="rows">0</str>
 <str name="json.facet">
 {
 category: {
 type: terms,
 field: cat
 }
 }
 </str>
 </lst>
 </lst>
 </lst>
 <arr name="jsonQueries">
 <str>
 .facets.category.buckets[] as $object |
 $object.val as $term |
 $object.count as $value |
 {
 name : "solr_facets_category",
 type : "GAUGE",
 help : "Category facets",
 label_names : ["term"],
 label_values : [$term],
 value : $value
 }
 </str>
 </arr>
 </lst>
</search>
```

```
</rules>
```

```
</config>
```

## Configuration Tags and Elements

The solr-exporter works by making a request to Solr according to the definitions in the configuration file, scraping the response, and converting it to a JSON structure Prometheus can understand. The configuration file defines the elements to request, how to scrape them, and where to place the extracted data in the JSON template.

The solr-exporter configuration file always starts and closes with two simple elements:

```
<config>
 <rules>

 </rules>
 </config>
```

Between these elements, the data the solr-exporter should request is defined. There are several possible types of requests to make:

```
<ping>
```

Scrape the response to a [PingRequestHandler](#) request.

```
<metrics>
```

Scrape the response to a [Metrics API](#) request.

```
<collections>
```

Scrape the response to a [Collections API](#) request.

```
<search>
```

Scrape the response to a [search](#) request.

Within each of these types, we need to define the query and how to work with the response. To do this, we define two additional elements:

```
<query>
```

Defines the query parameter(s) used for the request. This section uses several additional properties to define your query:

**collection**

The collection to issue the query against. Only used in SolrCloud mode.

**core**

The core to issue the query against. Only used in Standalone mode.

**path**

The path to the query endpoint where the request will be sent. Examples include admin/metrics or /select or admin/collections.

## params

Additional query parameters. These will vary depending on the request type and the endpoint. For example, if using the Metrics endpoint, you can add parameters to limit the query to a certain group and/or prefix. If you're using the Collections API, the command you want to use would be a parameter.

### <jsonQueries>

This is an array that defines one or more JSON Queries in jq syntax. For more details about how to structure these queries, see [the jq user manual](#).

A jq query has to output JSON in the following format:

```
{
 "name": "solr_ping",
 "type": "GAUGE",
 "help": "See following URL: https://lucene.apache.org/solr/guide/ping.html",
 "label_names": ["base_url", "core"],
 "label_values": ["http://localhost:8983/solr", "collection1"],
 "value": 1.0
}
```

See the section [Exposition Format](#) below for information about what information should go into each property, and an example of how the above example is translated for Prometheus.

## Exposition Format

The solr-exporter converts the JSON to the following exposition format:

```
TYPE <name> <type>
HELP <name> <help>
<name>{<label_names[0]>=<label_values[0]>,<label_names[1]>=<labelvalues[1]>,...} <value>
```

The following parameters should be set:

name

The metric name to set. For more details, see [Prometheus naming best practices](#).

type

The type of the metric, can be COUNTER, GAUGE, SUMMARY, HISTOGRAM or UNTYPED. For more details, see [Prometheus metric types](#).

help

Help text for the metric.

label\_names

Label names for the metric. For more details, see [Prometheus naming best practices](#).

label\_values

Label values for the metric. For more details, see [Prometheus naming best practices](#).

## value

Value for the metric. Value must be set to Double type.

For example, solr-exporter converts the JSON in the previous section to the following:

```
TYPE solr_ping gauge
HELP solr_ping See following URL: https://lucene.apache.org/solr/guide/ping.html
solr_ping{base_url="http://localhost:8983/solr",core="collection1"} 1.0
```

## Prometheus Configuration

In order for Prometheus to know about the solr-exporter, the listen address must be added to `prometheus.yml`, as in this example:

```
scrape_configs:
- job_name: 'solr'
 static_configs:
 - targets: ['localhost:9854']
```

If you already have a section for `scrape_configs`, you can add the `job_name` and other values in the same section.

When you apply the settings to Prometheus, it will start to pull Solr's metrics from solr-exporter.

## Sample Grafana Dashboard

A Grafana sample dashboard is provided in the following JSON file: `contrib/prometheus-exporter/conf/grafana-solr-dashboard.json`. You can place this with your other Grafana dashboard configurations and modify it as necessary depending on any customization you've done for the solr-exporter configuration.

This screenshot shows what it might look like:



## Grafana Dashboard

# Performance Statistics Reference

This page explains some of the statistics that Solr exposes.

There are two approaches to retrieving metrics. First, you can use the [Metrics API](#), or you can enable JMX and get metrics from the [MBean Request Handler](#) or via an external tool such as JConsole. The below descriptions focus on retrieving the metrics using the Metrics API, but the metric names are the same if using the MBean Request Handler or an external tool.

These statistics are per core. When you are running in SolrCloud mode these statistics would co-relate to the performance of an individual replica.

## Request Handler Statistics

### Update Request Handler

The update request handler is an endpoint to send data to Solr. We can see how many update requests are being fired, how fast is it performing, and other valuable information regarding requests.

**Registry & Path:** solr.<core>:UPDATE./update

You can request update request handler statistics with an API request such as  
`http://localhost:8983/solr/admin/metrics?group=core&prefix=UPDATE.`

### Search Request Handler

Can be useful to measure and track number of search queries, response times, etc. If you are not using the "select" handler then the path needs to be changed appropriately. Similarly if you are using the "sql" handler or "export" handler, the realtime handler "get", or any other handler similar statistics can be found for that as well.

**Registry & Path:** solr.<core>:QUERY./select

You can request statistics for the /select request handler with an API request such as  
`http://localhost:8983/solr/admin/metrics?group=core&prefix=QUERY./select.`

### Commonly Used Stats for Request Handlers

All of the update and search request handlers will provide the following statistics.

#### Request Times

To get request times, specifically, you can send an API request such as:

- `http://localhost:8983/solr/admin/metrics?group=core&prefix=UPDATE./update.requestTimes`
- `http://localhost:8983/solr/admin/metrics?group=core&prefix=QUERY./select.requestTimes`

Attribute	Description
15minRate	Requests per second received over the past 15 minutes.

Attribute	Description
5minRate	Requests per second received over the past 5 minutes.
p75_ms	Request processing time for the request which belongs to the 75 <sup>th</sup> Percentile. E.g., if 100 requests are received, then the 75 <sup>th</sup> fastest request time will be reported by this statistic.
p95_ms	Request processing time in milliseconds for the request which belongs to the 95 <sup>th</sup> Percentile. E.g., if 100 requests are received, then the 95 <sup>th</sup> fastest request time will be reported in this statistic.
p999_ms	Request processing time in milliseconds for the request which belongs to the 99.9 <sup>th</sup> Percentile. E.g., if 1000 requests are received, then the 999 <sup>th</sup> fastest request time will be reported in this statistic.
p99_ms	Request processing time in milliseconds for the request which belongs to the 99 <sup>th</sup> Percentile. E.g., if 100 requests are received, then the 99 <sup>th</sup> fastest request time will be reported in this statistic.
count	Total number of requests made since the Solr process was started.
median_ms	Median of all the request processing time.
avgRequestsPerSecond	Average number of requests received per second.
avgTimePerRequest	Average time taken for processing the requests. This parameter will decay over time, with a bias toward activity in the last 5 minutes.

## Errors and Other Times

Other types of data such as errors and timeouts are also provided. These are available under different metric names. For example:

- `http://localhost:8983/solr/admin/metrics?group=core&prefix=UPDATE./update.errors`
- `http://localhost:8983/solr/admin/metrics?group=core&prefix=QUERY./select.errors`

The table below shows the metric names and attributes to request:

Metric name	Description
QUERY./select.errors UPDATE./update.errors	Number of errors encountered by handler. In addition to a count of errors, mean, 1 minute, 5 minute, and 15 minute rates are also available.
QUERY./select.clientErrors UPDATE./update.clientErrors	Number of syntax or parse errors made by a client while making requests. In addition to a count of errors, mean, 1 minute, 5 minute, and 15 minute rates are also available.
QUERY./select.serverErrors UPDATE./update.serverErrors	Number of errors thrown by the server while executing the request. In addition to a count of errors, mean, 1 minute, 5 minute, and 15 minute rates are also available.
QUERY./select.timeouts UPDATE./update.timeouts	Number of responses received with partial results. In addition to a count of timeout events, mean, 1 minute, 5 minute, and 15 minute rates are also available.

Metric name	Description
QUERY./select.totalTime UPDATE./update.totalTime	The sum of all request processing times since the Solr process was started.
QUERY./select.handlerStart UPDATE./update.handlerStart	Epoch time when the handler was registered.

## Update Handler

This section has information on the total number of adds and how many commits have been fired against a Solr core.

**Registry & Path:** solr.<core>:UPDATE.updateHandler

You can get all update handler statistics shown in the table below with an API request such as <http://localhost:8983/solr/admin/metrics?group=core&prefix=UPDATE.updateHandler>.

The following describes the specific statistics you can get:

Attribute	Description
UPDATE.updateHandler.adds	Total number of “add” requests since last commit.
UPDATE.updateHandler.autoCommitMaxTime	Maximum time between two auto-commits execution.
UPDATE.updateHandler.autoCommits	Total number of auto-commits executed.
UPDATE.updateHandler.commits	Number of total commits executed.  In addition to a count of commits, mean, 1 minute, 5 minute, and 15 minute rates are also available.
UPDATE.updateHandler.cumulativeAdds	Number of “effective” additions executed over the lifetime. The counter is incremented when “add” command is executed while decremented when “rollback” is executed.  In addition to a count of adds, mean, 1 minute, 5 minute, and 15 minute rates are also available.
UPDATE.updateHandler.cumulativeDeletesById	Number of document deletions executed by ID over the lifetime. The counter is incremented when “delete” command is executed and decremented when “rollback” is executed.  In addition to a count of deletes, mean, 1 minute, 5 minute, and 15 minute rates are also available.
UPDATE.updateHandler.cumulativeDeletesByQuery	Number of document deletions executed by query over the lifetime. The counter is incremented when “delete” command is executed and decremented when “rollback” is executed.  In addition to a count of deletes, mean, 1 minute, 5 minute, and 15 minute rates are also available.

Attribute	Description
UPDATE.updateHandler.cumulativeErrors	Number of error messages received while performing addition/deletion actions on documents over the lifetime.  In addition to a count of errors, mean, 1 minute, 5 minute, and 15 minute rates are also available.
UPDATE.updateHandler.deletesById	Currently uncommitted deletions by ID.
UPDATE.updateHandler.deletesByQuery	Currently uncommitted deletions by query.
UPDATE.updateHandler.docsPending	Number of documents which are pending commit.
UPDATE.updateHandler.errors	Number of error messages received while performing addition/deletion/commit/rollback actions on documents over the lifetime of the core.
UPDATE.updateHandler.expungeDeletes	Number of commit commands issued with expunge deletes.  In addition to a count of expunge deletes, mean, 1 minute, 5 minute, and 15 minute rates are also available.
UPDATE.updateHandler.merges	Number of index merges that have occurred.  In addition to a count of merges, mean, 1 minute, 5 minute, and 15 minute rates are also available.
UPDATE.updateHandler.optimizes	Number of explicit optimize commands issued.  In addition to a count of optimizations, mean, 1 minute, 5 minute, and 15 minute rates are also available.
UPDATE.updateHandler.rollbacks	Number of rollbacks executed.  In addition to a count of rollbacks, mean, 1 minute, 5 minute, and 15 minute rates are also available.
UPDATE.updateHandler.softAutoCommitMaxTime	Maximum document 'adds' between two soft auto-commits.
UPDATE.updateHandler.softAutoCommits	Number of soft commits executed.

## Cache Statistics

### Document Cache

This cache holds Lucene Document objects (the stored fields for each document). Since Lucene internal document IDs are transient, this cache cannot be auto-warmed.

**Registry and Path:** solr.<core>:CACHE.searcher.documentCache

You can get the statistics shown in the table below with an API request such as

<http://localhost:8983/solr/admin/metrics?group=core&prefix=CACHE.searcher.documentCache>.

## Query Result Cache

This cache holds the results of previous searches: ordered lists of document IDs based on a query, a sort, and the range of documents requested

**Registry and Path:** solr.<core>:CACHE.searcher.queryResultCache

You can get the statistics shown in the table below with an API request such as  
`http://localhost:8983/solr/admin/metrics?group=core&prefix=CACHE.searcher.queryResultCache.`

## Filter Cache

This cache is used for filters for unordered sets of all documents that match a query.

**Registry and Path:** solr.<core>:CACHE.searcher.filterCache

You can get the statistics shown in the table below with an API request such as  
`http://localhost:8983/solr/admin/metrics?group=core&prefix=CACHE.searcher.filterCache.`

## Statistics for Caches

The following statistics are available for each of the caches mentioned above:

Attribute	Description
cumulative_evictions	Number of cache evictions across all caches since this node has been running.
cumulative_hitratio	Ratio of cache hits to lookups across all the caches since this node has been running.
cumulative_hits	Number of cache hits across all the caches since this node has been running.
cumulative_inserts	Number of cache insertions across all the caches since this node has been running.
cumulative_lookups	Number of cache lookups across all the caches since this node has been running.
evictions	Number of cache evictions for the current index searcher.
hitratio	Ratio of cache hits to lookups for the current index searcher.
hits	Number of hits for the current index searcher.
inserts	Number of inserts into the cache.
lookups	Number of lookups against the cache.
size	Number of entries in the cache at that particular instance.
warmupTime	Warm-up time for the registered index searcher. This time is taken in account for the “auto-warming” of caches.

When eviction by heap usage is enabled, the following additional statistics are available for the Query Result Cache:

Attribute	Description
maxRamMB	Maximum heap that should be used by the cache beyond which keys will be evicted.
ramBytesUsed	Actual heap usage of the cache at that particular instance.
evictionsRamUsage	Number of cache evictions for the current index searcher because heap usage exceeded maxRamMB.

More information on Solr caches is available in the section [Query Settings in SolrConfig](#).

# Securing Solr

When planning how to secure Solr, you should consider which of the available features or approaches are right for you.

- Authentication or authorization of users using:
  - Kerberos Authentication Plugin
  - Basic Authentication Plugin
  - Rule-Based Authorization Plugin
  - Custom authentication or authorization plugin
- [Enabling SSL](#)
- If using SolrCloud, [ZooKeeper Access Control](#)
- [Audit logging](#) for recording an audit trail



No Solr API, including the Admin UI, is designed to be exposed to non-trusted parties. Tune your firewall so that only trusted computers and people are allowed access. Because of this, the project will not regard e.g., Admin UI XSS issues as security vulnerabilities. However, we still ask you to report such issues in JIRA.

# Authentication and Authorization Plugins

Solr has security frameworks for supporting authentication and authorization of users. This allows for verifying a user's identity and for restricting access to resources in a Solr cluster.

Solr includes some plugins out of the box, and additional plugins can be developed using the authentication and authorization frameworks described below.

All authentication and authorization plugins can work with Solr whether they are running in SolrCloud mode or standalone mode. All authentication and authorization configuration, including users and permission rules, are stored in a file named `security.json`. When using Solr in standalone mode, this file must be in the `$SOLR_HOME` directory (usually `server/solr`). When using SolrCloud, this file must be located in ZooKeeper.

The following section describes how to enable plugins with `security.json` and place them in the proper locations for your mode of operation.

## Enable Plugins with `security.json`

All of the information required to initialize either type of security plugin is stored in a `security.json` file. This file contains 2 sections, one each for authentication and authorization.

*Sample `security.json`*

```
{
 "authentication" : {
 "class": "class.thatimplements.authentication"
 },
 "authorization": {
 "class": "class.thatimplements.authorization"
 }
}
```

The `/security.json` file needs to be in the proper location before a Solr instance comes up so Solr starts with the security plugin enabled. See the section [Using `security.json` with Solr](#) below for information on how to do this.

Depending on the plugin(s) in use, other information will be stored in `security.json` such as user information or rules to create roles and permissions. This information is added through the APIs for each plugin provided by Solr, or, in the case of a custom plugin, the approach designed by you.

Here is a more detailed `security.json` example. In this, the Basic authentication and rule-based authorization plugins are enabled, and some data has been added:

```
{
 "authentication":{
 "class":"solr.BasicAuthPlugin",
 "credentials": {"solr": "IV0EHq1OnNrj6gvRCwvFwTrZ1+z1oBbnQdiVC3otuq0=
Ndd7LKvVBAaZIF0QAVi1ekCfAJXr1GGfLtrUXhgrF8c="}
 },
 "authorization":{
 "class": "solr.RuleBasedAuthorizationPlugin",
 "permissions": [{"name": "security-edit",
 "role": "admin"}],
 "user-role": {"solr": "admin"}
 }
}
```

## Using security.json with Solr

### In SolrCloud Mode

While configuring Solr to use an authentication or authorization plugin, you will need to upload a `security.json` file to ZooKeeper. The following command writes the file as it uploads it - you could also upload a file that you have already created locally.

```
>server/scripts/cloud-scripts/zkcli.sh -zkhost localhost:2181 -cmd put /security.json
'{"authentication": {"class": "org.apache.solr.security.KerberosPlugin"}}'
```

Note that this example defines the KerberosPlugin for authentication. You will want to modify this section as appropriate for the plugin you are using.

This example also defines `security.json` on the command line, but you can also define a file locally and upload it to ZooKeeper.



Depending on the authentication and authorization plugin that you use, you may have user information stored in `security.json`. If so, we highly recommend that you implement access control in your ZooKeeper nodes. Information about how to enable this is available in the section [ZooKeeper Access Control](#).

Once `security.json` has been uploaded to ZooKeeper, you should use the appropriate APIs for the plugins you're using to update it. You can edit it manually, but you must take care to remove any version data so it will be properly updated across all ZooKeeper nodes. The version data is found at the end of the `security.json` file, and will appear as the letter "v" followed by a number, such as `{"v":138}`.

### In Standalone Mode

When running Solr in standalone mode, you need to create the `security.json` file and put it in the `$SOLR_HOME` directory for your installation (this is the same place you have located `solr.xml` and is usually `server/solr`).

If you are using [Legacy Scaling and Distribution](#), you will need to place `security.json` on each node of the cluster.

You can use the authentication and authorization APIs, but if you are using the legacy scaling model, you will need to make the same API requests on each node separately. You can also edit security.json by hand if you prefer.

## Authentication Plugins

Authentication plugins help in securing the endpoints of Solr by authenticating incoming requests. A custom plugin can be implemented by extending the AuthenticationPlugin class.

An authentication plugin consists of two parts:

1. Server-side component, which intercepts and authenticates incoming requests to Solr using a mechanism defined in the plugin, such as Kerberos, Basic Auth or others.
2. Client-side component, i.e., an extension of HttpClientConfigurer, which enables a SolrJ client to make requests to a secure Solr instance using the authentication mechanism which the server understands.

## Enabling a Plugin

- Specify the authentication plugin in /security.json as in this example:

```
{
 "authentication": {
 "class": "class.that.implements.authentication",
 "other_data" : "..."
 }
}
```

- All of the content in the authentication block of security.json would be passed on as a map to the plugin during initialization.
- An authentication plugin can also be used with a standalone Solr instance by passing in -DauthenticationPlugin=<plugin class name> during startup.

## Available Authentication Plugins

Solr has the following implementations of authentication plugins:

- [Kerberos Authentication Plugin](#)
- [Basic Authentication Plugin](#)
- [Hadoop Authentication Plugin](#)
- [JWT Authentication Plugin](#)

## Authorization

An authorization plugin can be written for Solr by extending the [AuthorizationPlugin](#) interface.

## Loading a Custom Plugin

- Make sure that the plugin implementation is in the classpath.

- The plugin can then be initialized by specifying the same in security.json in the following manner:

```
{
 "authorization": {
 "class": "org.apache.solr.security.MockAuthorizationPlugin",
 "other_data" : "..."}
}
```

All of the content in the authorization block of security.json would be passed on as a map to the plugin during initialization.



The authorization plugin is only supported in SolrCloud mode. Also, reloading the plugin isn't yet supported and requires a restart of the Solr installation (meaning, the JVM should be restarted, not simply a core reload).

## Available Authorization Plugins

Solr has one implementation of an authorization plugin:

- [Rule-Based Authorization Plugin](#)

## Authenticating in the Admin UI

Whenever an authentication plugin is enabled, authentication is also required for all or some operations in the Admin UI. The Admin UI is an AngularJS application running inside your browser, and is treated as any other external client by Solr.

When authentication is required the Admin UI will present you with a login dialogue. The authentication plugins currently supported by the Admin UI are:

- [Basic Authentication Plugin](#)
- [JWT Authentication Plugin](#)

If your plugin of choice is not supported, the Admin UI will still let you perform unrestricted operations, while for restricted operations you will need to interact with Solr by sending HTTP requests instead of through the graphical user interface of the Admin UI. All operations supported by Admin UI can be performed through Solr's RESTful APIs.

## Securing Inter-Node Requests

There are a lot of requests that originate from the Solr nodes itself. For example, requests from overseer to nodes, recovery threads, etc. We call these 'inter-node' requests. Solr has a special built-in PKIAuthenticationPlugin (see below) that will always be available to secure inter-node traffic.

Each Authentication plugin may also decide to secure inter-node requests on its own. They may do this through the so-called HttpClientBuilder mechanism, or they may alternatively choose on a per-request basis whether to delegate to PKI or not by overriding a interceptInternodeRequest() method from the base class, where any HTTP headers can be set.

## PKIAuthenticationPlugin

The PKIAuthenticationPlugin provides a built-in authentication mechanism where each Solr node is a super user and is fully trusted by other Solr nodes through the use of Public Key Infrastructure (PKI). Each Authentication plugin may choose to delegate all or some inter-node traffic to the PKI plugin.

For each outgoing request PKIAuthenticationPlugin adds a special header 'SolrAuth' which carries the timestamp and principal encrypted using the private key of that node. The public key is exposed through an API so that any node can read it whenever it needs it. Any node who gets the request with that header, would get the public key from the sender and decrypt the information. If it is able to decrypt the data, the request is trusted. It is invalid if the timestamp is more than 5 secs old. This assumes that the clocks of different nodes in the cluster are synchronized. Only traffic from other Solr nodes registered with Zookeeper is trusted.

The timeout is configurable through a system property called `pkiauth.ttl`. For example, if you wish to bump up the time-to-live to 10 seconds (10000 milliseconds), start each node with a property '`-Dpkiauth.ttl=10000`'.

## Basic Authentication Plugin

Solr can support Basic authentication for users with the use of the BasicAuthPlugin.

An authorization plugin is also available to configure Solr with permissions to perform various activities in the system. The authorization plugin is described in the section [Rule-Based Authorization Plugin](#).

### Enable Basic Authentication

To use Basic authentication, you must first create a `security.json` file. This file and where to put it is described in detail in the section [Enable Plugins with security.json](#).

For Basic authentication, the `security.json` file must have an authentication part which defines the class being used for authentication. Usernames and passwords (as a sha256(password+salt) hash) could be added when the file is created, or can be added later with the Basic authentication API, described below.

The authorization part is not related to Basic authentication, but is a separate authorization plugin designed to support fine-grained user access control. For more information, see the section [Rule-Based Authorization Plugin](#).

An example `security.json` showing both sections is shown below to show how these plugins can work together:

```
{
 "authentication": { ①
 "blockUnknown": true, ②
 "class": "solr.BasicAuthPlugin",
 "credentials": {"solr": "IV0EHq1OnNrj6gvRCwvFwTrZ1+z1oBbnQdiVC3otuq0=
Ndd7LKvVBAaZIF0QAVi1ekCfAJXr1GGfLtTRUXhgrF8c="}, ③
 "realm": "My Solr users", ④
 "forwardCredentials": false ⑤
 },
 "authorization": {
 "class": "solr.RuleBasedAuthorizationPlugin",
 "permissions": [{"name": "security-edit",
 "role": "admin"}], ⑥
 "user-role": {"solr": "admin"} ⑦
 }
}
```

There are several things defined in this file:

- ① Basic authentication and rule-based authorization plugins are enabled.
- ② The parameter "blockUnknown": true means that unauthenticated requests are not allowed to pass through.
- ③ A user called 'solr', with a password 'SolrRocks' has been defined.
- ④ We override the realm property to display another text on the login prompt.
- ⑤ The parameter "forwardCredentials": false means we let Solr's PKI authenticaion handle distributed request instead of forwarding the Basic Auth header.
- ⑥ The 'admin' role has been defined, and it has permission to edit security settings.
- ⑦ The 'solr' user has been defined to the 'admin' role.

Save your settings to a file called `security.json` locally. If you are using Solr in standalone mode, you should put this file in `$SOLR_HOME`.

If `blockUnknown` does not appear in the `security.json` file, it will default to `false`. This has the effect of not requiring authentication at all. In some cases, you may want this; for example, if you want to have `security.json` in place but aren't ready to enable authentication. However, you will want to ensure that this parameter is set to `true` in order for authentication to be truly enabled in your system.

If `realm` is not defined, it will default to `solr`.

If you are using SolrCloud, you must upload `security.json` to ZooKeeper. You can use this example command, ensuring that the ZooKeeper port is correct:

```
bin/solr zk cp file:path_to_local_security.json zk:/security.json -z localhost:9983
```



If you have defined `ZK_HOST` in `solr.in.sh`/`solr.in.cmd` (see [instructions](#)) you can omit `-z <zookeeper host string>` from the above command.

## Caveats

There are a few things to keep in mind when using the Basic authentication plugin.

- Credentials are sent in plain text by default. It's recommended to use SSL for communication when Basic authentication is enabled, as described in the section [Enabling SSL](#).
- A user who has access to write permissions to security.json will be able to modify all the permissions and how users have been assigned permissions. Special care should be taken to only grant access to editing security to appropriate users.
- Your network should, of course, be secure. Even with Basic authentication enabled, you should not unnecessarily expose Solr to the outside world.

## Editing Basic Authentication Plugin Configuration

An Authentication API allows modifying user IDs and passwords. The API provides an endpoint with specific commands to set user details or delete a user.

### API Entry Point

- v1: `http://localhost:8983/solr/admin/authentication`
- v2: `http://localhost:8983/api/cluster/security/authentication`

This endpoint is not collection-specific, so users are created for the entire Solr cluster. If users need to be restricted to a specific collection, that can be done with the authorization rules.

### Add a User or Edit a Password

The `set-user` command allows you to add users and change their passwords. For example, the following defines two users and their passwords:

#### V1 API

```
curl --user solr:SolrRocks http://localhost:8983/solr/admin/authentication -H 'Content-type:application/json' -d '{"set-user": {"tom":"TomIsCool", "harry":"HarrysSecret"}}'
```

#### V2 API

```
curl --user solr:SolrRocks http://localhost:8983/api/cluster/security/authentication -H 'Content-type:application/json' -d '{"set-user": {"tom":"TomIsCool", "harry":"HarrysSecret"}}'
```

### Delete a User

The `delete-user` command allows you to remove a user. The user password does not need to be sent to remove a user. In the following example, we've asked that user IDs 'tom' and 'harry' be removed from the system.

**V1 API**

```
curl --user solr:SolrRocks http://localhost:8983/solr/admin/authentication -H 'Content-type:application/json' -d '{"delete-user": ["tom", "harry"]}'
```

**V2 API**

```
curl --user solr:SolrRocks http://localhost:8983/api/cluster/security/authentication -H 'Content-type:application/json' -d '{"delete-user": ["tom", "harry"]}'
```

## Set a Property

Set properties for the authentication plugin. The currently supported properties for the Basic Authentication plugin are `blockUnknown`, `realm` and `forwardCredentials`.

**V1 API**

```
curl --user solr:SolrRocks http://localhost:8983/solr/admin/authentication -H 'Content-type:application/json' -d '{"set-property": {"blockUnknown":false}}'
```

**V2 API**

```
curl --user solr:SolrRocks http://localhost:8983/api/cluster/security/authentication -H 'Content-type:application/json' -d '{"set-property": {"blockUnknown":false}}'
```

The authentication realm defaults to `solr` and is displayed in the `WWW-Authenticate` HTTP header and in the Admin UI login page. To change the realm, set the `realm` property:

**V1 API**

```
curl --user solr:SolrRocks http://localhost:8983/solr/admin/authentication -H 'Content-type:application/json' -d '{"set-property": {"realm":"My Solr users"}}'
```

**V2 API**

```
curl --user solr:SolrRocks http://localhost:8983/api/cluster/security/authentication -H 'Content-type:application/json' -d '{"set-property": {"realm":"My Solr users"}}'
```

## Using Basic Auth with SolrJ

There are two main ways to use SolrJ with Solr servers protected by basic authentication: either the permissions can be set on each individual request, or the underlying http client can be configured to add credentials to all requests that it sends.

### Per-Request Basic Auth Credentials

The simplest way to setup basic authentication in SolrJ is use the `setBasicAuthCredentials` method on each request as in this example:

```
SolrRequest req ;//create a new request object
req.setBasicAuthCredentials(userName, password);
solrClient.request(req);
```

Query example:

```
QueryRequest req = new QueryRequest(new SolrQuery("*:*"));
req.setBasicAuthCredentials(userName, password);
QueryResponse rsp = req.process(solrClient);
```

While this is method is simple, it can often be inconvenient to ensure the credentials are provided everywhere they're needed. It also doesn't work with the many `SolrClient` methods which don't consume `SolrRequest` objects.

### Global (JVM) Basic Auth Credentials

Alternatively, users can use SolrJ's `PreemptiveBasicAuthClientBuilderFactory` to add basic authentication credentials to *all* requests automatically. To enable this feature, users should set the following system property

`-Dsolr.httpclient.builder.factory=org.apache.solr.client.solrj.impl.PreemptiveBasicAuthClientBuilderFactory`. `PreemptiveBasicAuthClientBuilderFactory` allows applications to provide credentials in two different ways:

1. The `basicauth` system property can be passed, containing the credentials directly (e.g., `-Dbasicauth=username:password`). This option is straightforward, but may expose the credentials in the command line, depending on how they're set.
2. The `solr.httpclient.config` system property can be passed, containing a path to a properties file holding the credentials. Inside this file the username and password can be specified as `httpBasicAuthUser` and `httpBasicAuthPassword`, respectively.

```
httpBasicAuthUser=my_username
httpBasicAuthPassword=secretPassword
```

## Using the Solr Control Script with Basic Auth

Add the following line to the `solr.in.sh` or `solr.in.cmd` file. This example tells the `bin/solr` command line

to use "basic" as the type of authentication, and to pass credentials with the user-name "solr" and password "SolrRocks":

```
SOLR_AUTH_TYPE="basic"
SOLR_AUTHENTICATION_OPTS="-Dbasicauth=solr:SolrRocks"
```

## Hadoop Authentication Plugin

The Hadoop authentication plugin enables Solr to use the [Hadoop authentication library](#) for securing Solr nodes.

This authentication plugin is a thin wrapper that delegates all functionality to the Hadoop authentication library. All configuration parameters for the library are passed through the plugin.

This plugin can be particularly useful in leveraging an extended set of features or newly available features in the Hadoop authentication library.

Please note that the version of Hadoop library used by Solr is upgraded periodically. While Solr will ensure the stability and backwards compatibility of the structure of the plugin configuration (viz., the parameter names of this plugin), the values of these parameters may change based on the version of Hadoop library. Please review the Hadoop documentation for the version used by your Solr installation for more details.

For some of the authentication schemes (e.g., Kerberos), Solr provides a native implementation of authentication plugin. If you require a more stable setup, in terms of configuration, ability to perform rolling upgrades, backward compatibility, etc., you should consider using such plugin. Please review the section [Authentication and Authorization Plugins](#) for an overview of authentication plugin options in Solr.

There are two plugin classes:

- `HadoopAuthPlugin`: This can be used with standalone Solr as well as Solrcloud with [PKI authentication](#) for internode communication.
- `ConfigurableInternodeAuthHadoopPlugin`: This is an extension of `HadoopAuthPlugin` that allows you to configure the authentication scheme for internode communication.



For most SolrCloud or standalone Solr setups, the `HadoopAuthPlugin` should suffice.

## Plugin Configuration

`class`

Should be either `solr.HadoopAuthPlugin` or `solr.ConfigurableInternodeAuthHadoopPlugin`. This parameter is required.

`type`

The type of authentication scheme to be configured. See [configuration](#) options. This parameter is required.

`sysPropPrefix`

The prefix to be used to define the Java system property for configuring the authentication mechanism. This property is required.

The name of the Java system property is defined by appending the configuration parameter name to this prefix value. For example, if the prefix is `solr` then the Java system property `solr.kerberos.principal` defines the value of configuration parameter `kerberos.principal`.

#### authConfigs

Configuration parameters required by the authentication scheme defined by the `type` property. This property is required. For more details, see [Hadoop configuration](#) options.

#### defaultConfigs

Default values for the configuration parameters specified by the `authConfigs` property. The default values are specified as a collection of key-value pairs (i.e., "property-name": "default\_value").

#### enableDelegationToken

If true, the delegation tokens functionality will be enabled.

#### initKerberosZk

For enabling initialization of kerberos before connecting to ZooKeeper (if applicable).

#### proxyUserConfigs

Configures proxy users for the underlying Hadoop authentication mechanism. This configuration is expressed as a collection of key-value pairs (i.e., "property-name": "default\_value").

#### clientBuilderFactory

No | The `HttpClientBuilderFactory` implementation used for the Solr internal communication. Only applicable for `ConfigurableInternodeAuthHadoopPlugin`.

## Example Configurations

### Kerberos Authentication using Hadoop Authentication Plugin

This example lets you configure Solr to use Kerberos Authentication, similar to how you would use the [Kerberos Authentication Plugin](#).

After consulting the Hadoop authentication library's documentation, you can supply per host configuration parameters using the `solr.*` prefix. As an example, the Hadoop authentication library expects a parameter `kerberos.principal`, which can be supplied as a system property named `solr.kerberos.principal` when starting a Solr node. Refer to the section [Kerberos Authentication Plugin](#) for other typical configuration parameters.

Please note that this example uses `ConfigurableInternodeAuthHadoopPlugin`, and hence you must provide the `clientBuilderFactory` implementation. As a result, all internode communication will use the Kerberos mechanism, instead of PKI authentication.

To setup this plugin, use the following in your `security.json` file.

```
{
 "authentication": {
 "class": "solr.ConfigurableInternodeAuthHadoopPlugin",
 "sysPropPrefix": "solr.",
 "type": "kerberos",
 "clientBuilderFactory": "org.apache.solr.client.solrj.impl.Krb5HttpClientBuilder",
 "initKerberosZk": "true",
 "authConfigs": [
 "kerberos.principal",
 "kerberos.keytab",
 "kerberos.name.rules"
],
 "defaultConfigs": {
 }
 }
}
```

## Simple Authentication with Delegation Tokens

Similar to the previous example, this is an example of setting up a Solr cluster that uses delegation tokens. Refer to the parameters in the Hadoop authentication library's [documentation](#) or refer to the section [Kerberos Authentication Plugin](#) for further details. Please note that this example does not use Kerberos and the requests made to Solr must contain valid delegation tokens.

To setup this plugin, use the following in your security.json file.

```
{
 "authentication": {
 "class": "solr.HadoopAuthPlugin",
 "sysPropPrefix": "solr.",
 "type": "simple",
 "enableDelegationToken": "true",
 "authConfigs": [
 "delegation-token.token-kind",
 "delegation-token.update-interval.sec",
 "delegation-token.max-lifetime.sec",
 "delegation-token.renewal-interval.sec",
 "delegation-token.remove-scan-interval.sec",
 "cookie.domain",
 "signer.secret.provider",
 "zk-dt-secret-manager.enable",
 "zk-dt-secret-manager.znodeWorkingPath",
 "signer.secret.provider.zookeeper.path"
],
 "defaultConfigs": {
 "delegation-token.token-kind": "solr-dt",
 "signer.secret.provider": "zookeeper",
 "zk-dt-secret-manager.enable": "true",
 "token.validity": "36000",
 "zk-dt-secret-manager.znodeWorkingPath": "solr/security/zkdtsm",
 "signer.secret.provider.zookeeper.path": "/token",
 "cookie.domain": "127.0.0.1"
 }
 }
}
```

## Kerberos Authentication Plugin

If you are using Kerberos to secure your network environment, the Kerberos authentication plugin can be used to secure a Solr cluster.

This allows Solr to use a Kerberos service principal and keytab file to authenticate with ZooKeeper and between nodes of the Solr cluster (if applicable). Users of the Admin UI and all clients (such as [SolrJ](#)) would also need to have a valid ticket before being able to use the UI or send requests to Solr.

Support for the Kerberos authentication plugin is available in SolrCloud mode or standalone mode.

 If you are using Solr with a Hadoop cluster secured with Kerberos and intend to store your Solr indexes in HDFS, also see the section [Running Solr on HDFS](#) for additional steps to configure Solr for that purpose. The instructions on this page apply only to scenarios where Solr will be secured with Kerberos. If you only need to store your indexes in a Kerberized HDFS system, please see the other section referenced above.

## How Solr Works With Kerberos

When setting up Solr to use Kerberos, configurations are put in place for Solr to use a *service principal*, or a Kerberos username, which is registered with the Key Distribution Center (KDC) to authenticate requests. The configurations define the service principal name and the location of the keytab file that contains the credentials.

### security.json

The Solr authentication model uses a file called `security.json`. A description of this file and how it is created and maintained is covered in the section [Authentication and Authorization Plugins](#). If this file is created after an initial startup of Solr, a restart of each node of the system is required.

### Service Principals and Keytab Files

Each Solr node must have a service principal registered with the Key Distribution Center (KDC). The Kerberos plugin uses SPNego to negotiate authentication.

Using `HTTP/host1@YOUR-DOMAIN.ORG`, as an example of a service principal:

- `HTTP` indicates the type of requests which this service principal will be used to authenticate. The `HTTP/` in the service principal is a must for SPNego to work with requests to Solr over HTTP.
- `host1` is the host name of the machine hosting the Solr node.
- `YOUR-DOMAIN.ORG` is the organization wide Kerberos realm.

Multiple Solr nodes on the same host may have the same service principal, since the host name is common to them all.

Along with the service principal, each Solr node needs a keytab file which should contain the credentials of the service principal used. A keytab file contains encrypted credentials to support passwordless logins while obtaining Kerberos tickets from the KDC. For each Solr node, the keytab file should be kept in a secure location and not shared with users of the cluster.

Since a Solr cluster requires internode communication, each node must also be able to make Kerberos enabled requests to other nodes. By default, Solr uses the same service principal and keytab as a 'client principal' for internode communication. You may configure a distinct client principal explicitly, but doing so is not recommended and is not covered in the examples below.

### Kerberized ZooKeeper

When setting up a kerberized SolrCloud cluster, it is recommended to enable Kerberos security for ZooKeeper as well.

In such a setup, the client principal used to authenticate requests with ZooKeeper can be shared for internode communication as well. This has the benefit of not needing to renew the ticket granting tickets (TGTs) separately, since the ZooKeeper client used by Solr takes care of this. To achieve this, a single JAAS configuration (with the app name as Client) can be used for the Kerberos plugin as well as for the ZooKeeper client.

See the [ZooKeeper Configuration](#) section below for an example of starting ZooKeeper in Kerberos mode.

## Browser Configuration

In order for your browser to access the Solr Admin UI after enabling Kerberos authentication, it must be able to negotiate with the Kerberos authenticator service to allow you access. Each browser supports this differently, and some (like Chrome) do not support it at all. If you see 401 errors when trying to access the Solr Admin UI after enabling Kerberos authentication, it's likely your browser has not been configured properly to know how or where to negotiate the authentication request.

Detailed information on how to set up your browser is beyond the scope of this documentation; please see your system administrators for Kerberos for details on how to configure your browser.

## Kerberos Authentication Configuration

### *Consult Your Kerberos Admins!*



Before attempting to configure Solr to use Kerberos authentication, please review each step outlined below and consult with your local Kerberos administrators on each detail to be sure you know the correct values for each parameter. Small errors can cause Solr to not start or not function properly, and are notoriously difficult to diagnose.

Configuration of the Kerberos plugin has several parts:

- Create service principals and keytab files
- ZooKeeper configuration
- Create or update /security.json
- Define jaas-client.conf
- Solr startup parameters

We'll walk through each of these steps below.

### *Using Hostnames*



To use host names instead of IP addresses, use the SOLR\_HOST configuration in bin/solr.in.sh or pass a -Dhost=<hostname> system parameter during Solr startup. This guide uses IP addresses. If you specify a hostname, replace all the IP addresses in the guide with the Solr hostname as appropriate.

## Get Service Principals and Keytabs

Before configuring Solr, make sure you have a Kerberos service principal for each Solr host and ZooKeeper (if ZooKeeper has not already been configured) available in the KDC server, and generate a keytab file as shown below.

This example assumes the hostname is 192.168.0.107 and your home directory is /home/foo/. This example should be modified for your own environment.

```
root@kdc:/# kadmin.local
Authenticating as principal foo/admin@EXAMPLE.COM with password.

kadmin.local: addprinc HTTP/192.168.0.107
WARNING: no policy specified for HTTP/192.168.0.107@EXAMPLE.COM; defaulting to no policy
Enter password for principal "HTTP/192.168.0.107@EXAMPLE.COM":
Re-enter password for principal "HTTP/192.168.0.107@EXAMPLE.COM":
Principal "HTTP/192.168.0.107@EXAMPLE.COM" created.

kadmin.local: ktadd -k /tmp/107.keytab HTTP/192.168.0.107
Entry for principal HTTP/192.168.0.107 with kvno 2, encryption type aes256-cts-hmac-sha1-96 added
to keytab WRFILE:/tmp/107.keytab.
Entry for principal HTTP/192.168.0.107 with kvno 2, encryption type arcfour-hmac added to keytab
WRFILE:/tmp/107.keytab.
Entry for principal HTTP/192.168.0.107 with kvno 2, encryption type des3-cbc-sha1 added to keytab
WRFILE:/tmp/108.keytab.
Entry for principal HTTP/192.168.0.107 with kvno 2, encryption type des-cbc-crc added to keytab
WRFILE:/tmp/107.keytab.

kadmin.local: quit
```

Copy the keytab file from the KDC server's /tmp/107.keytab location to the Solr host at /keytabs/107.keytab. Repeat this step for each Solr node.

You might need to take similar steps to create a ZooKeeper service principal and keytab if it has not already been set up. In that case, the example below shows a different service principal for ZooKeeper, so the above might be repeated with zookeeper/host1 as the service principal for one of the nodes

## ZooKeeper Configuration

If you are using a ZooKeeper that has already been configured to use Kerberos, you can skip the ZooKeeper-related steps shown here.

Since ZooKeeper manages the communication between nodes in a SolrCloud cluster, it must also be able to authenticate with each node of the cluster. Configuration requires setting up a service principal for ZooKeeper, defining a JAAS configuration file and instructing ZooKeeper to use both of those items.

The first step is to create a file java.env in ZooKeeper's conf directory and add the following to it, as in this example:

```
export JVMFLAGS="-Djava.security.auth.login.config=/etc/zookeeper/conf/jaas-client.conf"
```

The JAAS configuration file should contain the following parameters. Be sure to change the principal and keyTab path as appropriate. The file must be located in the path defined in the step above, with the filename specified.

```
Server {
 com.sun.security.auth.module.Krb5LoginModule required
 useKeyTab=true
 keyTab="/keytabs/zkhost1.keytab"
 storeKey=true
 doNotPrompt=true
 useTicketCache=false
 debug=true
 principal="zookeeper/host1@EXAMPLE.COM";
};
```

Finally, add the following lines to the ZooKeeper configuration file `zoo.cfg`:

```
authProvider.1=org.apache.zookeeper.server.auth.SASLAuthenticationProvider
jaasLoginRenew=3600000
```

Once all of the pieces are in place, start ZooKeeper with the following parameter pointing to the JAAS configuration file:

```
bin/zkServer.sh start -Djava.security.auth.login.config=/etc/zookeeper/conf/jaas-client.conf
```

## Create security.json

Create the `security.json` file.

In SolrCloud mode, you can set up Solr to use the Kerberos plugin by uploading the `security.json` to ZooKeeper while you create it, as follows:

```
server/scripts/cloud-scripts/zkcli.sh -zkhost localhost:2181 -cmd put /security.json
'{"authentication": {"class": "org.apache.solr.security.KerberosPlugin"}}'
```

If you are using Solr in standalone mode, you need to create the `security.json` file and put it in your `$SOLR_HOME` directory.

More details on how to use a `/security.json` file in Solr are available in the section [Authentication and Authorization Plugins](#).



If you already have a `/security.json` file in ZooKeeper, download the file, add or modify the authentication section and upload it back to ZooKeeper using the [Command Line Utilities](#) available in Solr.

## Define a JAAS Configuration File

The JAAS configuration file defines the properties to use for authentication, such as the service principal and the location of the keytab file. Other properties can also be set to ensure ticket caching and other features.

The following example can be copied and modified slightly for your environment. The location of the file can

be anywhere on the server, but it will be referenced when starting Solr so it must be readable on the filesystem. The JAAS file may contain multiple sections for different users, but each section must have a unique name so it can be uniquely referenced in each application.

In the below example, we have created a JAAS configuration file with the name and path of /home/foo/jaas-client.conf. We will use this name and path when we define the Solr start parameters in the next section. Note that the client principal here is the same as the service principal. This will be used to authenticate internode requests and requests to ZooKeeper. Make sure to use the correct principal hostname and the keyTab file path.

```
Client {
 com.sun.security.auth.module.Krb5LoginModule required
 useKeyTab=true
 keyTab="/keytabs/107.keytab"
 storeKey=true
 useTicketCache=true
 debug=true
 principal="HTTP/192.168.0.107@EXAMPLE.COM";
};
```

The first line of this file defines the section name, which will be used with the `solr.kerberos.jaas.appname` parameter, defined below.

The main properties we are concerned with are the `keyTab` and `principal` properties, but there are others which may be required for your environment. The [javadocs for the Krb5LoginModule](#) (the class that's being used and is called in the second line above) provide a good outline of the available properties, but for reference the ones in use in the above example are explained here:

- `useKeyTab`: this boolean property defines if we should use a keytab file (true, in this case).
- `keyTab`: the location and name of the keytab file for the principal this section of the JAAS configuration file is for. The path should be enclosed in double-quotes.
- `storeKey`: this boolean property allows the key to be stored in the private credentials of the user.
- `useTicketCache`: this boolean property allows the ticket to be obtained from the ticket cache.
- `debug`: this boolean property will output debug messages for help in troubleshooting.
- `principal`: the name of the service principal to be used.

## Solr Startup Parameters

While starting up Solr, the following host-specific parameters need to be passed. These parameters can be passed at the command line with the `bin/solr start` command (see [Solr Control Script Reference](#) for details on how to pass system parameters) or defined in `bin/solr.in.sh` or `bin/solr.in.cmd` as appropriate for your operating system.

`solr.kerberos.name.rules`

Used to map Kerberos principals to short names. Default value is DEFAULT. Example of a name rule:

`RULE:[1:$1@$0](.*EXAMPLE.COM)s/@.*//`.

`solr.kerberos.cookie.domain`

Used to issue cookies and should have the hostname of the Solr node. This parameter is required.

#### `solr.kerberos.cookie.portaware`

When set to true, cookies are differentiated based on host and port, as opposed to standard cookies which are not port aware. This should be set if more than one Solr node is hosted on the same host. The default is false.

#### `solr.kerberos.principal`

The service principal. This parameter is required.

#### `solr.kerberos.keytab`

Keytab file path containing service principal credentials. This parameter is required.

#### `solr.kerberos.jaas.appname`

The app name (section name) within the JAAS configuration file which is required for internode communication. Default is Client, which is used for ZooKeeper authentication as well. If different users are used for ZooKeeper and Solr, they will need to have separate sections in the JAAS configuration file.

#### `java.security.auth.login.config`

Path to the JAAS configuration file for configuring a Solr client for internode communication. This parameter is required.

Here is an example that could be added to bin/solr.in.sh. Make sure to change this example to use the right hostname and the keytab file path.

```
SOLR_AUTH_TYPE="kerberos"
SOLR_AUTHENTICATION_OPTS="-Djava.security.auth.login.config=/home/foo/jaas-client.conf
-Dsolr.kerberos.cookie.domain=192.168.0.107 -Dsolr.kerberos.cookie.portaware=true
-Dsolr.kerberos.principal=HTTP/192.168.0.107@EXAMPLE.COM
-Dsolr.kerberos.keytab=/keytabs/107.keytab"
```

#### *KDC with AES-256 encryption*

If your KDC uses AES-256 encryption, you need to add the Java Cryptography Extension (JCE) Unlimited Strength Jurisdiction Policy Files to your JRE before a Kerberized Solr can interact with the KDC.



You will know this when you see an error like this in your Solr logs: "KrbException: Encryption type AES256 CTS mode with HMAC SHA1-96 is not supported/enabled".

For Java 1.8, this is available here: <http://www.oracle.com/technetwork/java/javase/downloads/jce8-download-2133166.html>.

Replace the local\_policy.jar present in JAVA\_HOME/jre/lib/security/ with the new local\_policy.jar from the downloaded package and restart the Solr node.

## Using Delegation Tokens

The Kerberos plugin can be configured to use delegation tokens, which allow an application to reuse the authentication of an end-user or another application.

There are a few use cases for Solr where this might be helpful:

- Using distributed clients (such as MapReduce) where each client may not have access to the user's credentials.
- When load on the Kerberos server is high. Delegation tokens can reduce the load because they do not access the server after the first request.
- If requests or permissions need to be delegated to another user.

To enable delegation tokens, several parameters must be defined. These parameters can be passed at the command line with the bin/solr start command (see [Solr Control Script Reference](#) for details on how to pass system parameters) or defined in bin/solr.in.sh or bin/solr.in.cmd as appropriate for your operating system.

#### solr.kerberos.delegation.token.enabled

This is false by default, set to true to enable delegation tokens. This parameter is required if you want to enable tokens.

#### solr.kerberos.delegation.token.kind

The type of delegation tokens. By default this is solr-dt. Likely this does not need to change. No other option is available at this time.

#### solr.kerberos.delegation.token.validity

Time, in seconds, for which delegation tokens are valid. The default is 36000 seconds.

#### solr.kerberos.delegation.token.signer.secret.provider

Where delegation token information is stored internally. The default is zookeeper which must be the location for delegation tokens to work across Solr servers (when running in SolrCloud mode). No other option is available at this time.

#### solr.kerberos.delegation.token.signer.secret.provider.zookeeper.path

The ZooKeeper path where the secret provider information is stored. This is in the form of the path + /security/token. The path can include the chroot or the chroot can be omitted if you are not using it. This example includes the chroot: server1:9983,server2:9983,server3:9983/solr/security/token.

#### solr.kerberos.delegation.token.secret.manager.znode.working.path

The ZooKeeper path where token information is stored. This is in the form of the path + /security/zkdtsm. The path can include the chroot or the chroot can be omitted if you are not using it. This example includes the chroot: server1:9983,server2:9983,server3:9983/solr/security/zkdtsm.

## Start Solr

Once the configuration is complete, you can start Solr with the bin/solr script, as in the example below, which is for users in SolrCloud mode only. This example assumes you modified bin/solr.in.sh or bin/solr.in.cmd, with the proper values, but if you did not, you would pass the system parameters along with the start command. Note you also need to customize the -z property as appropriate for the location of your ZooKeeper nodes.

```
bin/solr -c -z server1:2181,server2:2181,server3:2181/solr
```



If you have defined ZK\_HOST in solr.in.sh/solr.in.cmd (see [instructions](#)) you can omit -z <zk host string> from the above command.

## Test the Configuration

1. Do a kinit with your username. For example, kinit user@EXAMPLE.COM.
2. Try to access Solr using curl. You should get a successful response.

```
curl --negotiate -u : "http://192.168.0.107:8983/solr/"
```

## Using SolrJ with a Kerberized Solr

To use Kerberos authentication in a SolrJ application, you need the following two lines before you create a SolrClient:

```
System.setProperty("java.security.auth.login.config", "/home/foo/jaas-client.conf");
HttpClientUtil.setConfigurer(new Krb5HttpClientConfigurer());
```

You need to specify a Kerberos service principal for the client and a corresponding keytab in the JAAS client configuration file above. This principal should be different from the service principal we created for Solr.

Here's an example:

```
SolrJClient {
 com.sun.security.auth.module.Krb5LoginModule required
 useKeyTab=true
 keyTab="/keytabs/foo.keytab"
 storeKey=true
 useTicketCache=true
 debug=true
 principal="solrclient@EXAMPLE.COM";
};
```

## Delegation Tokens with SolrJ

Delegation tokens are also supported with SolrJ, in the following ways:

- DelegationTokenRequest and DelegationTokenResponse can be used to get, cancel, and renew delegation tokens.
- HttpSolrClient.Builder includes a withDelegationToken function for creating an HttpSolrClient that uses a delegation token to authenticate.

Sample code to get a delegation token:

```
private String getDelegationToken(final String renewer, final String user, HttpSolrClient solrClient) throws Exception {
 DelegationTokenRequest.Get get = new DelegationTokenRequest.Get(renewer) {
 @Override
 public SolrParams getParams() {
 ModifiableSolrParams params = new ModifiableSolrParams(super.getParams());
 params.set("user", user);
 return params;
 }
 };
 DelegationTokenResponse.Get getResponse = get.process(solrClient);
 return getResponse.getDelegationToken();
}
```

To create a `HttpSolrClient` that uses delegation tokens:

```
HttpSolrClient client = new HttpSolrClient.Builder("http://localhost:8983/solr")
 .withDelegationToken(token).build();
```

To create a `CloudSolrClient` that uses delegation tokens:

```
CloudSolrClient client = new CloudSolrClient.Builder()
 .withZkHost("localhost:2181")
 .withLBHttpSolrClientBuilder(new LBHttpSolrClient.Builder()
 .withResponseParser(client.getParser())
 .withHttpSolrClientBuilder(
 new HttpSolrClient.Builder()
 .withKerberosDelegationToken(token)
)
)
 .build();
```



Hadoop's delegation token responses are in JSON map format. A response parser for that is available in `DelegationTokenResponse`. Other response parsers may not work well with Hadoop responses.

## Rule-Based Authorization Plugin

Solr allows configuring roles to control user access to the system.

This is accomplished through rule-based permission definitions which are assigned to users. The roles are fully customizable, and provide the ability to limit access to specific collections, request handlers, request parameters, and request methods.

The roles can be used with any of the authentication plugins or with a custom authentication plugin if you have created one. You will only need to ensure that you configure the role-to-user mappings with the proper user IDs that your authentication system provides.

Once defined through the API, roles are stored in `security.json`.



Solr's Admin UI interacts with Solr using its regular APIs. When rule-based authorization is in use, logged-in users not authorized to access the full range of these APIs may see some sections of the UI that appear blank or "broken". For best results, the Admin UI should only be accessed by users with full API access.

## Enable the Authorization Plugin

The plugin must be enabled in `security.json`. This file and where to put it in your system is described in detail in the section [Enable Plugins with `security.json`](#).

This file has two parts, the authentication part and the authorization part. The authentication part stores information about the class being used for authentication.

The authorization part is not related to Basic authentication, but is a separate authorization plugin designed to support fine-grained user access control. When creating `security.json` you can add the permissions to the file, or you can use the Authorization API described below to add them as needed.

This example `security.json` shows how the [Basic authentication plugin](#) can work with this authorization plugin:

```
{
 "authentication": {
 "class": "solr.BasicAuthPlugin", ①
 "blockUnknown": true, ②
 "credentials": {"solr": "IV0EHq1OnNrj6gvRCwvFwTrZ1+z1oBbnQdiVC3otuq0=
Ndd7LKvVBAaZIF0QAVi1ekCfAJXr1GGfLtrUXhgrF8c="} ③
 },
 "authorization": {
 "class": "solr.RuleBasedAuthorizationPlugin", ④
 "permissions": [{"name": "security-edit",
 "role": "admin"}], ⑤
 "user-role": {"solr": "admin"} ⑥
 }
}
```

There are several things defined in this example:

- ① Basic authentication plugin is enabled.
- ② All requests w/o credentials will be rejected with a 401 error. Set 'blockUnknown' to false (or remove it altogether) if you wish to let unauthenticated requests to go through. However, if a particular resource is protected by a rule, they are rejected anyway with a 401 error.
- ③ A user named 'solr', with a password has been defined.
- ④ Rule-based authorization plugin is enabled.
- ⑤ The 'admin' role has been defined, and it has permission to edit security settings.
- ⑥ The 'solr' user has been defined to the 'admin' role.

## Permission Attributes

Each role is comprised of one or more permissions which define what the user is allowed to do. Each

permission is made up of several attributes that define the allowed activity. There are some pre-defined permissions which cannot be modified.

The permissions are consulted in order they appear in `security.json`. The first permission that matches is applied for each user, so the strictest permissions should be at the top of the list. Permissions order can be controlled with a parameter of the Authorization API, as described below.

## Predefined Permissions

There are several permissions that are pre-defined. These have fixed default values, which cannot be modified, and new attributes cannot be added. To use these attributes, simply define a role that includes this permission, and then assign a user to that role.

The pre-defined permissions are:

- **security-edit**: this permission is allowed to edit the security configuration, meaning any update action that modifies `security.json` through the APIs will be allowed.
- **security-read**: this permission is allowed to read the security configuration, meaning any action that reads `security.json` settings through the APIs will be allowed.
- **schema-edit**: this permission is allowed to edit a collection's schema using the [Schema API](#). Note that this allows schema edit permissions for *all* collections. If edit permissions should only be applied to specific collections, a custom permission would need to be created.
- **schema-read**: this permission is allowed to read a collection's schema using the [Schema API](#). Note that this allows schema read permissions for *all* collections. If read permissions should only be applied to specific collections, a custom permission would need to be created.
- **config-edit**: this permission is allowed to edit a collection's configuration using the [Config API](#), the [Request Parameters API](#), and other APIs which modify `configoverlay.json`. Note that this allows configuration edit permissions for *all* collections. If edit permissions should only be applied to specific collections, a custom permission would need to be created.
- **core-admin-read**: Read operations on the core admin API
- **core-admin-edit**: Core admin commands that can mutate the system state.
- **config-read**: this permission is allowed to read a collection's configuration using the [Config API](#), the [Request Parameters API](#), and other APIs which modify `configoverlay.json`. Note that this allows configuration read permissions for *all* collections. If read permissions should only be applied to specific collections, a custom permission would need to be created.
- **collection-admin-edit**: this permission is allowed to edit a collection's configuration using the [Collections API](#). Note that this allows configuration edit permissions for *all* collections. If edit permissions should only be applied to specific collections, a custom permission would need to be created. Specifically, the following actions of the Collections API would be allowed:
  - CREATE
  - RELOAD
  - SPLITSHARD
  - CREATESHARD
  - DELETESHARD

- CREATEALIAS
  - DELETEALIAS
  - DELETE
  - DELETEREPLICA
  - ADDREPLICA
  - CLUSTERPROP
  - MIGRATE
  - ADDROLE
  - REMOVEROLE
  - ADDREPLICAPROP
  - DELETEREPLICAPROP
  - BALANCESHARDUNIQUE
  - REBALANCELEADERS
- **collection-admin-read:** this permission is allowed to read a collection's configuration using the [Collections API](#). Note that this allows configuration read permissions for *all* collections. If read permissions should only be applied to specific collections, a custom permission would need to be created. Specifically, the following actions of the Collections API would be allowed:
    - LIST
    - OVERSEERSTATUS
    - CLUSTERSTATUS
    - REQUESTSTATUS
  - **update:** this permission is allowed to perform any update action on any collection. This includes sending documents for indexing (using an [update request handler](#)). This applies to all collections by default (collection: "\*").
  - **read:** this permission is allowed to perform any read action on any collection. This includes querying using search handlers (using [request handlers](#)) such as /select, /get, /browse, /tvrh, /terms, /clustering, /elevate, /export, /spell, /clustering, and /sql. This applies to all collections by default (collection: "\*").
  - **all:** Any requests coming to Solr.

## Authorization API

### Authorization API Endpoint

/admin/authorization: takes a set of commands to create permissions, map permissions to roles, and map roles to users.

### Manage Permissions

Three commands control managing permissions:

- **set-permission:** create a new permission, overwrite an existing permission definition, or assign a pre-

defined permission to a role.

- update-permission: update some attributes of an existing permission definition.
- delete-permission: remove a permission definition.

Permissions need to be created if they are not on the list of pre-defined permissions above.

Several properties can be used to define your custom permission.

#### name

The name of the permission. This is required only if it is a predefined permission.

#### collection

The collection or collections the permission will apply to.

When the path that will be allowed is collection-specific, such as when setting permissions to allow use of the Schema API, omitting the collection property will allow the defined path and/or method for all collections. However, when the path is one that is non-collection-specific, such as the Collections API, the collection value must be null. The default value is \*, or all collections.

#### path

A request handler name, such as /update or /select. A wild card is supported, to allow for all paths as appropriate (such as, /update/\*).

#### method

HTTP methods that are allowed for this permission. You could allow only GET requests, or have a role that allows PUT and POST requests. The method values that are allowed for this property are GET, POST, PUT,DELETE and HEAD.

#### params

The names and values of request parameters. This property can be omitted if all request parameters are to be matched, but will restrict access only to the values provided if defined.

For example, this property could be used to limit the actions a role is allowed to perform with the Collections API. If the role should only be allowed to perform the LIST or CLUSTERSTATUS requests, you would define this as follows:

```
{"params": {
 "action": ["LIST", "CLUSTERSTATUS"]
}
}
```

The value of the parameter can be a simple string or it could be a regular expression. Use the prefix REGEX: to use a regular expression match instead of a string identity match

If the commands LIST and CLUSTERSTATUS are case insensitive, the above example should be as follows

```
{"params": {
 "action": ["REGEX:(?i)LIST", "REGEX:(?i)CLUSTERSTATUS"]
}
}
```

## before

This property allows ordering of permissions. The value of this property is the index of the permission that this new permission should be placed before in security.json. The index is automatically assigned in the order they are created.

## role

The name of the role(s) to give this permission. This name will be used to map user IDs to the role to grant these permissions. The value can be wildcard such as (\*), which means that any user is OK, but no user is NOT OK.

The following creates a new permission named "collection-mgr" that is allowed to create and list collections. The permission will be placed before the "read" permission. Note also that we have defined "collection" as null, this is because requests to the Collections API are never collection-specific.

```
curl --user solr:SolrRocks -H 'Content-type:application/json' -d '{
 "set-permission": {"collection": null,
 "path": "/admin/collections",
 "params": {"action": ["LIST", "CREATE"]},
 "before": 3,
 "role": "admin"}
}' http://localhost:8983/solr/admin/authorization
```

Apply an update permission on all collections to a role called dev and read permissions to a role called guest:

```
curl --user solr:SolrRocks -H 'Content-type:application/json' -d '{
 "set-permission": {"name": "update", "role": "dev"},
 "set-permission": {"name": "read", "role": "guest"}
}' http://localhost:8983/solr/admin/authorization
```

## Update or Delete Permissions

Permissions can be accessed using their index in the list. Use the /admin/authorization API to see the existing permissions and their indices.

The following example updates the 'role' attribute of permission at index 3:

```
curl --user solr:SolrRocks -H 'Content-type:application/json' -d '{
 "update-permission": {"index": 3,
 "role": ["admin", "dev"]}
}' http://localhost:8983/solr/admin/authorization
```

The following example deletes permission at index 3:

```
curl --user solr:SolrRocks -H 'Content-type:application/json' -d '{
 "delete-permission": 3
' http://localhost:8983/solr/admin/authorization
```

## Map Roles to Users

A single command allows roles to be mapped to users:

- set-user-role: map a user to a permission.

To remove a user's permission, you should set the role to null. There is no command to delete a user role.

The values supplied to the command are simply a user ID and one or more roles the user should have.

For example, the following would grant a user "solr" the "admin" and "dev" roles, and remove all roles from the user ID "harry":

```
curl -u solr:SolrRocks -H 'Content-type:application/json' -d '{
 "set-user-role" : {"solr": ["admin", "dev"],
 "harry": null}
' http://localhost:8983/solr/admin/authorization
```

## JWT Authentication Plugin

Solr can support [JSON Web Token](#) (JWT) based Bearer authentication with the use of the `JWTAuthPlugin`. This allows Solr to assert that a user is already authenticated with an external [Identity Provider](#) by validating that the JWT formatted [access token](#) is digitally signed by the Identity Provider. The typical use case is to integrate Solr with an [OpenID Connect](#) enabled IdP.

## Enable JWT Authentication

To use JWT Bearer authentication, the `security.json` file must have an authentication part which defines the class being used for authentication along with configuration parameters.

The simplest possible `security.json` for registering the plugin without configuration is:

```
{
 "authentication": {
 "class": "solr.JWTAuthPlugin"
 }
}
```

The plugin will NOT block anonymous traffic in this mode, since the default for `blockUnknown` is false. It is then possible to start configuring the plugin using REST API calls, which is described below.

## Configuration Parameters

Key	Description	Default
blockUnknown	Set to true in order to block requests from users without a token	false
wellKnownUrl	URL to an <a href="#">OpenID Connect Discovery</a> endpoint	(no default)
clientId	Client identifier for use with OpenID Connect	(no default value) Required to authenticate with Admin UI
realm	Name of the authentication realm to echo back in HTTP 401 responses. Will also be displayed in Admin UI login page	'solr-jwt'
scope	Whitespace separated list of valid scopes. If configured, the JWT access token MUST contain a scope claim with at least one of the listed scopes. Example: solr:read solr:admin	
jwkUrl	An https URL to a <a href="#">JWK</a> keys file.	Auto configured if wellKnownUrl is provided
jwk	As an alternative to jwkUrl you may provide a JSON object here containing the public key(s) of the issuer.	
iss	Validates that the iss (issuer) claim equals this string	Auto configured if wellKnownUrl is provided
aud	Validates that the aud (audience) claim equals this string	If clientId is configured, require aud to match it
requireSub	Makes sub (subject) claim mandatory	true
requireExp	Makes exp (expiry time) claim mandatory	true
algWhitelist	JSON array with algorithms to accept: HS256, HS384, HS512, RS256, RS384, RS512, ES256, ES384, ES512, PS256, PS384, PS512, `none	Default is to allow all algorithms
jwkCacheDur	Duration of JWK cache in seconds	3600 (1 hour)
principalClaim	What claim id to pull principal from	sub

Key	Description	Default
claimsMatch	JSON object of claims (key) that must match a regular expression (value). Example: { "foo" : "A B" } will require the foo claim to be either "A" or "B".	(none)
adminUiScope	Define what scope is requested when logging in from Admin UI	If not defined, the first scope from scope parameter is used
authorizationEndpoint	The URL for the Id Provider's authorization endpoint	Auto configured if wellKnownUrl is provided
redirectUris	Valid location(s) for redirect after external authentication. Takes a string or array of strings. Must be the base URL of Solr, e.g., <a href="https://solr1.example.com:8983/solr/">https://solr1.example.com:8983/solr/</a> and must match the list of redirect URIs registered with the Identity Provider beforehand.	Defaults to empty list, i.e., any node is assumed to be a valid redirect target.

## More Configuration Examples

### With JWK URL

To start enforcing authentication for all users, requiring a valid JWT in the Authorization header, you need to configure the plugin with one or more [JSON Web Keys](#) (JWK). This is a JSON document containing the key used to sign/encrypt the JWT. It could be a symmetric or asymmetric key. The JWK can either be fetched (and cached) from an external HTTPS endpoint or specified directly in security.json. Below is an example of the former:

```
{
 "authentication": {
 "class": "solr.JWTAuthPlugin",
 "blockUnknown": true,
 "jwkUrl": "https://my.key.server/jwk.json"
 }
}
```

### With Admin UI Support

The next example shows configuring using [OpenID Connect Discovery](#) with a well-known URI for automatic configuration of many common settings, including ability to use the Admin UI with an OpenID Connect enabled Identity Provider.

```
{
 "authentication": {
 "class": "solr.JWTAuthPlugin",
 "blockUnknown": true,
 "wellKnownUrl": "https://idp.example.com/.well-known/openid-configuration",
 "clientId": "xyz",
 "redirectUri": "https://my.solr.server:8983/solr/"
 }
}
```

In this case, jwkUrl, iss and authorizationEndpoint will be automatically configured from the fetched configuration.

## Complex Example

Let's look at a more complex configuration, this time with a static embedded JWK:

```
{
 "authentication": {
 "class": "solr.JWTAuthPlugin", ①
 "blockUnknown": true, ②
 "jwk": { ③
 "e": "AQAB",
 "kid": "k1",
 "kty": "RSA",
 "n": "3ZF6wBGPMsLzsS1KLghxaVpZtXD3nTLzDm0c974i9-KNU_1rhhBeiVfs64VfEQmP8SA470jEy7yWcvnz9GvG-YAlm9i0wVF7jL12awdws0ocFjdSPT3sjPQKzOeM07G9XqNTkrvoFCn1YAi26fbhgcqwZDoeTcHQdRN32frzccuPhZrwImApI-edroKL1KWv2IvPDnz2Bpe2WWVc2HdoWYqEVD3p_BEy8f-RTSHK3_8kDDF9yAwI9jx7CK1_C-eYxXltm-6rpS5NGyFm0UNTZmxVU28T17LX8Vb6CiKyCQ9YRCtk_CvpKwmeuKEp9I28KHQNmGkDYT90nt3vjbCXxw"
 },
 "clientId": "solr-client-12345", ④
 "iss": "https://example.com/idp", ⑤
 "aud": "https://example.com/solr", ⑥
 "principalClaim": "solruid", ⑦
 "claimsMatch": { "foo" : "A|B", "dept" : "IT" }, ⑧
 "scope": "solr:read solr:write solr:admin", ⑨
 "algWhitelist" : ["RS256", "RS384", "RS512"] ⑩
 }
}
```

Let's comment on this config:

- ① Plugin class
- ② Make sure to block anyone without a valid token
- ③ Here we pass the JWK inline instead of referring to a URL with jwkUrl
- ④ Set the client id registered with Identity Provider
- ⑤ The issuer claim must match "https://example.com/idp"
- ⑥ The audience claim must match "https://example.com/solr"

- ⑦ Fetch the user id from another claim than the default sub
- ⑧ Require that the roles claim is one of "A" or "B" and that the dept claim is "IT"
- ⑨ Require one of the scopes solr:read, solr:write or solr:admin
- ⑩ Only accept RSA algorithms for signatures

## Editing JWT Authentication Plugin Configuration

All properties mentioned above can be set or changed using the Config Edit API. You can thus start with a simple configuration with only class configured and then configure the rest using the API.

### Set a Configuration Property

Set properties for the authentication plugin. Each of the configuration keys in the table above can be used as parameter keys for the set-property command.

Example:

#### V1 API

```
curl http://localhost:8983/solr/admin/authentication -H 'Content-type:application/json' -H
'Authorization: Bearer xxx.yyy.zzz' -d '{"set-property": {"blockUnknown":true,
"wellKnownUrl": "https://example.com/.well-known/openid-configuration", "scope": "solr:read
solr:write"}}'
```

#### V2 API

```
curl http://localhost:8983/api/cluster/security/authentication -H 'Content-
type:application/json' -H 'Authorization: Bearer xxx.yyy.zzz' -d -d '{"set-property":
{"blockUnknown":true, "wellKnownUrl": "https://example.com/.well-known/openid-configuration",
"scope": "solr:read solr:write"}}'
```

Insert a valid JWT access token in compact serialization format (xxx.yyy.zzz above) to authenticate with Solr once the plugin is active.

## Using Clients with JWT Auth

### SolrJ

SolrJ does not currently support supplying JWT tokens per request.

### cURL

To authenticate with Solr when using the cURL utility, supply a valid JWT access token in an Authorization header, as follows (replace xxxxxxx.xxxxxxx.xxxxxxx with your JWT compact token):

```
curl -H "Authorization: Bearer xxxxxxx.xxxxxxx.xxxxxxx" http://localhost:8983/solr/admin/info/system
```

## Admin UI

When this plugin is enabled, users will be redirected to a login page in the Admin UI once they attempt to do a restricted action. The page has a button that users will click and be redirected to the Identity Provider's login page. Once authenticated, the user will be redirected back to Solr Admin UI to the last known location. The session will last as long as the JWT token expiry time and is valid for one Solr server only. That means you have to login again when navigating to another Solr node. There is also a logout menu in the left column where user can explicitly log out.

## Using the Solr Control Script with JWT Auth

The control script (bin/solr) does not currently support JWT Auth.

# Enabling SSL

Solr can encrypt communications to and from clients, and between nodes in SolrCloud mode, with SSL.

This section describes enabling SSL using a self-signed certificate.

For background on SSL certificates and keys, see <http://www.tldp.org/HOWTO/SSL-Certificates-HOWTO/>.

## Basic SSL Setup

### Generate a Self-Signed Certificate and a Key

To generate a self-signed certificate and a single key that will be used to authenticate both the server and the client, we'll use the JDK keytool command and create a separate keystore. This keystore will also be used as a truststore below. It's possible to use the keystore that comes with the JDK for these purposes, and to use a separate truststore, but those options aren't covered here.

Run the commands below in the `server/etc/` directory in the binary Solr distribution. It's assumed that you have the JDK keytool utility on your PATH, and that openssl is also on your PATH. See <https://www.openssl.org/related/binaries.html> for OpenSSL binaries for Windows and Solaris.

The `-ext SAN=...` keytool option allows you to specify all the DNS names and/or IP addresses that will be allowed during hostname verification (but see below for how to skip hostname verification between Solr nodes so that you don't have to specify all hosts here).

In addition to `localhost` and `127.0.0.1`, this example includes a LAN IP address `192.168.1.3` for the machine the Solr nodes will be running on:

```
keytool -genkeypair -alias solr-ssl -keyalg RSA -keysize 2048 -keypass secret -storepass secret
-validity 9999 -keystore solr-ssl.keystore.jks -ext SAN=DNS:localhost,IP:192.168.1.3,IP:127.0.0.1
-dname "CN=localhost, OU=Organizational Unit, O=Organization, L=Location, ST=State, C=Country"
```

The above command will create a keystore file named `solr-ssl.keystore.jks` in the current directory.

### Convert the Certificate and Key to PEM Format for Use with curl

curl isn't capable of using JKS formatted keystores, so the JKS keystore needs to be converted to PEM format, which curl understands.

First convert the JKS keystore into PKCS12 format using keytool:

```
keytool -importkeystore -srckeystore solr-ssl.keystore.jks -destkeystore solr-ssl.keystore.p12
-srcstoretype jks -deststoretype pkcs12
```

The keytool application will prompt you to create a destination keystore password and for the source keystore password, which was set when creating the keystore ("secret" in the example shown above).

Next convert the PKCS12 format keystore, including both the certificate and the key, into PEM format using

the openssl command:

```
openssl pkcs12 -in solr-ssl.keystore.p12 -out solr-ssl.pem
```

If you want to use curl on OS X Yosemite (10.10), you'll need to create a certificate-only version of the PEM format, as follows:

```
openssl pkcs12 -nokeys -in solr-ssl.keystore.p12 -out solr-ssl.cacert.pem
```

## Set Common SSL-Related System Properties

The Solr Control Script is already setup to pass SSL-related Java system properties to the JVM. To activate the SSL settings, uncomment and update the set of properties beginning with SOLR\_SSL\_\* in bin/solr.in.sh. (or bin\solr.in.cmd on Windows).



If you setup Solr as a service on Linux using the steps outlined in [Taking Solr to Production](#), then make these changes in /var/solr/solr.in.sh instead.

### *bin/solr.in.sh example SOLR\_SSL\_\* configuration*

```
Enables HTTPS. It is implicitly true if you set SOLR_SSL_KEY_STORE. Use this config
to enable https module with custom jetty configuration.
SOLR_SSL_ENABLED=true
Uncomment to set SSL-related system properties
Be sure to update the paths to the correct keystore for your environment
SOLR_SSL_KEY_STORE=etc/solr-ssl.keystore.jks
SOLR_SSL_KEY_STORE_PASSWORD=secret
SOLR_SSL_TRUST_STORE=etc/solr-ssl.keystore.jks
SOLR_SSL_TRUST_STORE_PASSWORD=secret
Require clients to authenticate
SOLR_SSL_NEED_CLIENT_AUTH=false
Enable clients to authenticate (but not require)
SOLR_SSL_WANT_CLIENT_AUTH=false
SSL Certificates contain host/ip "peer name" information that is validated by default. Setting
this to false can be useful to disable these checks when re-using a certificate on many hosts
SOLR_SSL_CHECK_PEER_NAME=true
Override Key/Trust Store types if necessary
SOLR_SSL_KEY_STORE_TYPE=JKS
SOLR_SSL_TRUST_STORE_TYPE=JKS
```

When you start Solr, the bin/solr script includes the settings in bin/solr.in.sh and will pass these SSL-related system properties to the JVM.

### *Client Authentication Settings*



Enable either SOLR\_SSL\_NEED\_CLIENT\_AUTH or SOLR\_SSL\_WANT\_CLIENT\_AUTH but not both at the same time. They are mutually exclusive and Jetty will select one of them which may not be what you expect.

Similarly, when you start Solr on Windows, the bin\solr.cmd script includes the settings in bin\solr.in.cmd - uncomment and update the set of properties beginning with SOLR\_SSL\_\* to pass these SSL-related system properties to the JVM:

#### *bin\solr.in.cmd example SOLR\_SSL\_\* configuration*

```
REM Enables HTTPS. It is implicitly true if you set SOLR_SSL_KEY_STORE. Use this config
REM to enable https module with custom jetty configuration.
set SOLR_SSL_ENABLED=true
REM Uncomment to set SSL-related system properties
REM Be sure to update the paths to the correct keystore for your environment
set SOLR_SSL_KEY_STORE=etc/solr-ssl.keystore.jks
set SOLR_SSL_KEY_STORE_PASSWORD=secret
set SOLR_SSL_TRUST_STORE=etc/solr-ssl.keystore.jks
set SOLR_SSL_TRUST_STORE_PASSWORD=secret
REM Require clients to authenticate
set SOLR_SSL_NEED_CLIENT_AUTH=false
REM Enable clients to authenticate (but not require)
set SOLR_SSL_WANT_CLIENT_AUTH=false
REM SSL Certificates contain host/ip "peer name" information that is validated by default.
Setting
REM this to false can be useful to disable these checks when re-using a certificate on many hosts
set SOLR_SSL_CHECK_PEER_NAME=true
REM Override Key/Trust Store types if necessary
set SOLR_SSL_KEY_STORE_TYPE=JKS
set SOLR_SSL_TRUST_STORE_TYPE=JKS
```

## Run Single Node Solr using SSL

Start Solr using the command shown below; by default clients will not be required to authenticate:

### \*nix Command

```
bin/solr -p 8984
```

### Windows Command

```
bin\solr.cmd -p 8984
```

## Password Distribution via Hadoop Credential Store

Solr supports reading keystore and truststore passwords from Hadoop credential store. This approach can be beneficial if password rotation and distribution is already handled by credential stores.

Hadoop credential store can be used with Solr using the following two steps.

## Provide a Hadoop Credential Store

Create a Hadoop credstore file and define the entries below with the actual keystore passwords.

```
solr.jetty.keystore.password
solr.jetty.truststore.password
javax.net.ssl.keyStorePassword
javax.net.ssl.trustStorePassword
```

Note that if the javax.net.ssl.\* configurations are not set, they will fallback to the corresponding solr.jetty.\* configurations.

## Configure Solr to use Hadoop Credential Store

Solr requires three parameters to be configured in order to use the credential store file for keystore passwords.

`solr.ssl.credential.provider.chain`

The credential provider chain. This should be set to hadoop.

`SOLR_HADOOP_CREDENTIAL_PROVIDER_PATH`

The path to the credential store file.

`HADOOP_CREDSTORE_PASSWORD`

The password to the credential store.

### \*nix Example

```
SOLR_OPTS="-Dsolr.ssl.credential.provider.chain=hadoop"
SOLR_HADOOP_CREDENTIAL_PROVIDER_PATH=localjceks://file/home/solr/hadoop-credential-
provider.jceks
HADOOP_CREDSTORE_PASSWORD="credStorePass123"
```

### Windows Example

```
set SOLR_OPTS="-Dsolr.ssl.credential.provider.chain=hadoop"
set SOLR_HADOOP_CREDENTIAL_PROVIDER_PATH=localjceks://file/home/solr/hadoop-credential-
provider.jceks
set HADOOP_CREDSTORE_PASSWORD="credStorePass123"
```

## SSL with SolrCloud

This section describes how to run a two-node SolrCloud cluster with no initial collections and a single-node external ZooKeeper. The commands below assume you have already created the keystore described above.

## Configure ZooKeeper



ZooKeeper does not support encrypted communication with clients like Solr. There are several related JIRA tickets where SSL support is being planned/worked on: [ZOOKEEPER-235](#); [ZOOKEEPER-236](#); [ZOOKEEPER-1000](#); and [ZOOKEEPER-2120](#).

Before you start any SolrCloud nodes, you must configure your Solr cluster properties in ZooKeeper, so that Solr nodes know to communicate via SSL.

This section assumes you have created and started a single-node external ZooKeeper on port 2181 on localhost - see [Setting Up an External ZooKeeper Ensemble](#).

The urlScheme cluster-wide property needs to be set to https before any Solr node starts up. The example below uses the zkcli tool that comes with the binary Solr distribution to do this:

*\*nix command*

```
server/scripts/cloud-scripts/zkcli.sh -zkhost localhost:2181 -cmd clusterprop -name urlScheme
-val https
```

*Windows command*

```
server\scripts\cloud-scripts\zkcli.bat -zkhost localhost:2181 -cmd clusterprop -name urlScheme
-val https
```

If you have set up your ZooKeeper cluster to use a [chroot for Solr](#), make sure you use the correct zkhost string with zkcli, e.g., -zkhost localhost:2181/solr.

## Run SolrCloud with SSL



If you have defined ZK\_HOST in solr.in.sh/solr.in.cmd (see [instructions](#)) you can omit -z <zk host string> from all of the bin/solr/bin\solr.cmd commands below.

### Create Solr Home Directories for Two Nodes

Create two copies of the server/solr/ directory which will serve as the Solr home directories for each of your two SolrCloud nodes:

*\*nix commands*

```
mkdir cloud
cp -r server/solr cloud/node1
cp -r server/solr cloud/node2
```

*Windows commands*

```
mkdir cloud
xcopy /E server\solr cloud\NODE1\
xcopy /E server\solr cloud\NODE2\
xcopy /E server\solr cloud\NODE3\
```

## Start the First Solr Node

Next, start the first Solr node on port 8984. Be sure to stop the standalone server first if you started it when working through the previous section on this page.

### \*nix command

```
bin/solr -cloud -s cloud/node1 -z localhost:2181 -p 8984
```

### Windows command

```
bin\solr.cmd -cloud -s cloud\NODE1 -z localhost:2181 -p 8984
```

Notice the use of the `-s` option to set the location of the Solr home directory for node1.

If you created your SSL key without all DNS names/IP addresses on which Solr nodes will run, you can tell Solr to skip hostname verification for inter-Solr-node communications by setting the `solr.ssl.checkPeerName` system property to false:

### \*nix command

```
bin/solr -cloud -s cloud/node1 -z localhost:2181 -p 8984 -Dsolr.ssl.checkPeerName=false
```

### Windows command

```
bin\solr.cmd -cloud -s cloud\NODE1 -z localhost:2181 -p 8984 -Dsolr.ssl.checkPeerName=false
```

## Start the Second Solr Node

Finally, start the second Solr node on port 7574 - again, to skip hostname verification, add `-Dsolr.ssl.checkPeerName=false`:

### \*nix command

```
bin/solr -cloud -s cloud/node2 -z localhost:2181 -p 7574
```

### Windows command

```
bin\solr.cmd -cloud -s cloud\NODE2 -z localhost:2181 -p 7574
```

## Example Client Actions

curl on OS X Mavericks (10.9) has degraded SSL support. For more information and workarounds to allow one-way SSL, see <http://curl.haxx.se/mail/archive-2013-10/0036.html>. curl on OS X Yosemite (10.10) is improved - 2-way SSL is possible - see <http://curl.haxx.se/mail/archive-2014-10/0053.html>.



The curl commands in the following sections will not work with the system curl on OS X Yosemite (10.10). Instead, the certificate supplied with the -E parameter must be in PKCS12 format, and the file supplied with the --cacert parameter must contain only the CA certificate, and no key (see [above](#) for instructions on creating this file):

```
curl -E solr-ssl.keystore.p12:secret --cacert solr-ssl.cacert.pem ...
```



If your operating system does not include curl, you can download binaries here:  
<http://curl.haxx.se/download.html>

## Create a SolrCloud Collection using bin/solr

Create a 2-shard, replicationFactor=1 collection named mycollection using the default configset (\_default):

*\*nix command*

```
bin/solr create -c mycollection -shards 2
```

*Windows command*

```
bin\solr.cmd create -c mycollection -shards 2
```

The create action will pass the SOLR\_SSL\_\* properties set in your include file to the SolrJ code used to create the collection.

## Retrieve SolrCloud Cluster Status using curl

To get the resulting cluster status (again, if you have not enabled client authentication, remove the -E solr-ssl.pem:secret option):

```
curl -E solr-ssl.pem:secret --cacert solr-ssl.pem
"https://localhost:8984/solr/admin/collections?action=CLUSTERSTATUS&indent=on"
```

You should get a response that looks like this:

```
{
 "responseHeader": {
 "status": 0,
 "QTime": 2041},
 "cluster": {
 "collections": {
 "mycollection": {
 "shards": {
 "shard1": {
 "range": "80000000-ffffffff",
 "state": "active",
 "replicas": {
 "core_node1": {
 "state": "active",
 "base_url": "https://127.0.0.1:8984/solr",
 "core": "mycollection_shard1_replica1",
 "node_name": "127.0.0.1:8984_solr",
 "leader": "true"
 }
 }
 },
 "shard2": {
 "range": "0-7fffffff",
 "state": "active",
 "replicas": {
 "core_node2": {
 "state": "active",
 "base_url": "https://127.0.0.1:7574/solr",
 "core": "mycollection_shard2_replica1",
 "node_name": "127.0.0.1:7574_solr",
 "leader": "true"
 }
 }
 }
 },
 "maxShardsPerNode": "1",
 "router": {
 "name": "compositeId"
 },
 "replicationFactor": "1"
 }
 },
 "properties": {
 "urlScheme": "https"
 }
}
```

## Index Documents using post.jar

Use post.jar to index some example documents to the SolrCloud collection created above:

```
cd example/exampledocs

java -Djavax.net.ssl.keyStorePassword=secret -Djavax.net.ssl.keyStore=../../server/etc/solr
-ssl.keystore.jks -Djavax.net.ssl.trustStore=../../server/etc/solr-ssl.keystore.jks
-Djavax.net.ssl.trustStorePassword=secret -Durl=https://localhost:8984/solr/mycollection/update
-jar post.jar *.xml
```

## Query Using curl

Use curl to query the SolrCloud collection created above, from a directory containing the PEM formatted certificate and key created above (e.g., example/etc/) - if you have not enabled client authentication (system property -Djetty.ssl.clientAuth=true), then you can remove the -E solr-ssl.pem:secret option:

```
curl -E solr-ssl.pem:secret --cacert solr-ssl.pem
"https://localhost:8984/solr/mycollection/select?q=*:*"
```

## Index a Document using CloudSolrClient

From a java client using SolrJ, index a document. In the code below, the `javax.net.ssl.*` system properties are set programmatically, but you could instead specify them on the java command line, as in the `post.jar` example above:

```
System.setProperty("javax.net.ssl.keyStore", "/path/to/solr-ssl.keystore.jks");
System.setProperty("javax.net.ssl.keyStorePassword", "secret");
System.setProperty("javax.net.ssl.trustStore", "/path/to/solr-ssl.keystore.jks");
System.setProperty("javax.net.ssl.trustStorePassword", "secret");
String zkHost = "127.0.0.1:2181";
CloudSolrClient client = new CloudSolrClient.Builder().withZkHost(zkHost).build();
client.setDefaultCollection("mycollection");
SolrInputDocument doc = new SolrInputDocument();
doc.addField("id", "1234");
doc.addField("name", "A lovely summer holiday");
client.add(doc);
client.commit();
```

# Audit Logging

Solr has the ability to log an audit trail of all events in the system. Audit loggers are pluggable to suit any possible format or log destination.

*An audit trail (also called audit log) is a security-relevant chronological record, set of records, and/or destination and source of records that provide documentary evidence of the sequence of activities that have affected at any time a specific operation, procedure, or event. ([Wikipedia](#))*

## Configuration in security.json

Audit logging is configured in `security.json` under the `auditlogging` key.

The example `security.json` below configures synchronous audit logging to Solr default log file.

```
{
 "auditlogging": {
 "class": "solr.SolrLogAuditLoggerPlugin"
 }
}
```

By default any AuditLogger plugin configured will log asynchronously in the background to avoid slowing down the requests. To make audit logging happen synchronously, add the parameter `async: false`. For async logging, you may optionally also configure queue size, number of threads and whether it should block when the queue is full or discard events:

```
{
 "auditlogging": {
 "class": "solr.SolrLogAuditLoggerPlugin",
 "async": true,
 "blockAsync": false,
 "numThreads": 2,
 "queueSize": 4096,
 "eventTypes": ["REJECTED", "ANONYMOUS_REJECTED", "UNAUTHORIZED", "COMPLETED", "ERROR"]
 }
}
```

The defaults are `async: true`, `blockAsync: false`, `queueSize: 4096`. The default for `numThreads` is 2, or if the server has more than 4 CPU-cores then we use CPU-cores/2.

## Event Types

These are the event types triggered by the framework:

EventType	Usage
AUTHENTICATED	User successfully authenticated

EventType	Usage
REJECTED	Authentication request rejected
ANONYMOUS	Request proceeds with unknown user
ANONYMOUS_REJECTED	Request from unknown user rejected
AUTHORIZED	Authorization succeeded
UNAUTHORIZED	Authorization failed
COMPLETED	Request completed successfully
ERROR	Request was not executed due to an error

By default only the final event types REJECTED, ANONYMOUS\_REJECTED, UNAUTHORIZED, COMPLETED and ERROR are logged. What eventTypes are logged can be configured with the eventTypes configuration parameter.

## Muting Certain Events

The configuration parameter `muteRules` lets you mute logging for certain events. You may specify multiple rules and combination of rules that will cause muting. You can mute by request type, username, collection name, path, request parameters or IP address. We'll explain through examples:

The below example will mute logging for all SEARCH requests as well as all requests made my user johndoe or from IP address 192.168.0.10:

```
{
 "auditlogging": {
 "class": "solr.SolrLogAuditLoggerPlugin"
 "muteRules": ["type:SEARCH", "user:johndoe", "ip:192.168.0.10"]
 }
}
```

An mute rule may also be a list, in which case all must be true for muting to happen. The configuration below has three mute rules:

```
{
 "auditlogging": {
 "class": "solr.SolrLogAuditLoggerPlugin"
 "muteRules": [
 "ip:192.168.0.10", ①
 ["path:/admin/collections", "param:action=LIST"], ②
 ["path:/admin/collections", "param:collection=test"] ③
]
 }
}
```

① The first will mute all events from client IP 192.168.0.10

② The second rule will mute collection admin requests with `action=LIST`

③ The third rule will mute collection admin requests for the collection named test

Note how you can mix single string rules with lists of rules that must all match:

### Valid mute rules are:

- type:<request-type> (request-type by name: ADMIN, SEARCH, UPDATE, STREAMING, UNKNOWN)
- collection:<collection-name> (collection by name)
- user:<userid> (user by userid)
- path:</path/to/handler> (request path relative to /solr or for search/update requests relative to collection. Path is prefix matched, i.e., /admin will mute any sub path as well.)
- ip:<ip-address> (IPv4-address)
- param:<param>=<value> (request parameter)

## Chaining Multiple Loggers

Using the MultiDestinationAuditLogger you can configure multiple audit logger plugins in a chain, to log to multiple destinations, as follows:

```
{
 "auditlogging":{
 "class" : "solr.MultiDestinationAuditLogger",
 "plugins" : [
 { "class" : "solr.SolrLogAuditLoggerPlugin" },
 { "class" : "solr.MyOtherAuditPlugin",
 "customParam" : "value"
 }
]
 }
}
```

## Metrics

AuditLoggerPlugins record metrics about count and timing of log requests, as well as queue size for async loggers. The metrics keys are all recorded on the SECURITY category, and each metric name are prefixed with a scope of /auditlogging and the class name of the logger, e.g., SolrLogAuditLoggerPlugin. The individual metrics are:

- count (type: meter. Records number and rate of audit logs done)
- errors (type: meter. Records number and rate of errors)
- lost (type: meter. Records number and rate of events lost due to queue full and blockAsync=false)
- requestTimes (type: timer. Records latency and percentiles for logging performance)
- totalTime (type: counter. Records total time spent)
- queueCapacity (type: gauge. Records the max size of the async logging queue)
- queueSize (type: gauge. Records the number of events currently waiting in the queue)

- queuedTime (type: timer. Records the amount of time events waited in queue. Adding this with requestTimes you get total time from event to logging complete)
- async (type: gauge. Tells whether this logger is in async mode)



If you expect a very high request rate and have a slow audit logger plugin, you may see that the queueSize and queuedTime metrics increase, and in worst case start dropping events and see an increase in lost count. In this case you may want to increase the numThreads setting.

# Client APIs

This section discusses the available client APIs for Solr. It covers the following topics:

[Introduction to Client APIs](#): A conceptual overview of Solr client APIs.

[Choosing an Output Format](#): Information about choosing a response format in Solr.

[Using JavaScript](#): Explains why a client API is not needed for JavaScript responses.

[Using Python](#): Information about Python and JSON responses.

[Client API Lineup](#): A list of all Solr Client APIs, with links.

[Using SolrJ](#): Detailed information about SolrJ, an API for working with Java applications.

[Using Solr From Ruby](#): Detailed information about using Solr with Ruby applications.

[MBean Request Handler](#): Describes the MBean request handler for programmatic access to Solr server statistics and information.

# Introduction to Client APIs

At its heart, Solr is a Web application, but because it is built on open protocols, any type of client application can use Solr.

HTTP is the fundamental protocol used between client applications and Solr. The client makes a request and Solr does some work and provides a response. Clients use requests to ask Solr to do things like perform queries or index documents.

Client applications can reach Solr by creating HTTP requests and parsing the HTTP responses. Client APIs encapsulate much of the work of sending requests and parsing responses, which makes it much easier to write client applications.

Clients use Solr's five fundamental operations to work with Solr. The operations are query, index, delete, commit, and optimize.

Queries are executed by creating a URL that contains all the query parameters. Solr examines the request URL, performs the query, and returns the results. The other operations are similar, although in certain cases the HTTP request is a POST operation and contains information beyond whatever is included in the request URL. An index operation, for example, may contain a document in the body of the request.

Solr also features an `EmbeddedSolrServer` that offers a Java API without requiring an HTTP connection. For details, see [Using Solr](#).

# Choosing an Output Format

Many programming environments are able to send HTTP requests and retrieve responses. Parsing the responses is a slightly more thorny problem. Fortunately, Solr makes it easy to choose an output format that will be easy to handle on the client side.

Specify a response format using the `wt` parameter in a query. The available response formats are documented in [Response Writers](#).

Most client APIs hide this detail for you, so for many types of client applications, you won't ever have to specify a `wt` parameter. In JavaScript, however, the interface to Solr is a little closer to the metal, so you will need to add this parameter yourself.

# Client API Lineup

The Solr Wiki contains a list of client APIs at <http://wiki.apache.org/solr/IntegratingSolr>.

Here is the list of client APIs, current at this writing (November 2011):

Name	Environment	URL
SolRuby	Ruby	<a href="https://github.com/rsolr/rsolr">https://github.com/rsolr/rsolr</a>
DelSolr	Ruby	<a href="https://github.com/avvo/delsolr">https://github.com/avvo/delsolr</a>
acts_as_solr	Rails	<a href="http://acts-as-solr.rubyforge.org/">http://acts-as-solr.rubyforge.org/</a> , <a href="http://rubyforge.org/projects/background-solr/">http://rubyforge.org/projects/background-solr/</a>
Flare	Rails	<a href="http://wiki.apache.org/solr/Flare">http://wiki.apache.org/solr/Flare</a>
SolPHP	PHP	<a href="http://wiki.apache.org/solr/SolPHP">http://wiki.apache.org/solr/SolPHP</a>
SolrJ	Java	<a href="http://wiki.apache.org/solr/SolJava">http://wiki.apache.org/solr/SolJava</a>
Python API	Python	<a href="http://wiki.apache.org/solr/SolPython">http://wiki.apache.org/solr/SolPython</a>
PySolr	Python	<a href="http://code.google.com/p/pysolr/">http://code.google.com/p/pysolr/</a>
SolPerl	Perl	<a href="http://wiki.apache.org/solr/SolPerl">http://wiki.apache.org/solr/SolPerl</a>
Solr.pm	Perl	<a href="http://search.cpan.org/~garafola/Solr-0.03/lib/Solr.pm">http://search.cpan.org/~garafola/Solr-0.03/lib/Solr.pm</a>
SolrForrest	Forrest/Cocoon	<a href="http://wiki.apache.org/solr/SolrForrest">http://wiki.apache.org/solr/SolrForrest</a>
SolrSharp	C#	<a href="http://www.codeplex.com/solrsharp">http://www.codeplex.com/solrsharp</a>
SolColdfusion	ColdFusion	<a href="http://solcoldfusion.riaforge.org/">http://solcoldfusion.riaforge.org/</a>
SolrNet	.NET	<a href="https://github.com/mausch/SolrNet">https://github.com/mausch/SolrNet</a>
AJAX Solr	AJAX	<a href="http://github.com/evolvingweb/ajax-solr/wiki">http://github.com/evolvingweb/ajax-solr/wiki</a>

## Using JavaScript

Using Solr from JavaScript clients is so straightforward that it deserves a special mention. In fact, it is so straightforward that there is no client API. You don't need to install any packages or configure anything.

HTTP requests can be sent to Solr using the standard XMLHttpRequest mechanism.

By default, Solr sends [JavaScript Object Notation \(JSON\) responses](#), which are easily interpreted in JavaScript. You don't need to add anything to the request URL to have responses sent as JSON.

For more information and an excellent example, read the SolJSON page on the Solr Wiki:

<http://wiki.apache.org/solr/SolJSON>

# Using Python

Solr includes an output format specifically for [Python](#), but [JSON output](#) is a little more robust.

## Simple Python

Making a query is a simple matter. First, tell Python you will need to make HTTP connections.

```
from urllib2 import *
```

Now open a connection to the server and get a response. The `wt` query parameter tells Solr to return results in a format that Python can understand.

```
connection = urlopen('http://localhost:8983/solr/collection_name/select?q=cheese&wt=python')
response = eval(connection.read())
```

Now interpreting the response is just a matter of pulling out the information that you need.

```
print response['response']['numFound'], "documents found."
Print the name of each document.

for document in response['response']['docs']:
 print " Name =", document['name']
```

## Python with JSON

JSON is a more robust response format, and Python has support for it in its standard library since version 2.6.

Making a query is nearly the same as before. However, notice that the `wt` query parameter is now `json` (which is also the default if no `wt` parameter is specified), and the response is now digested by `json.load()`.

```
from urllib2 import *
import json
connection = urlopen('http://localhost:8983/solr/collection_name/select?q=cheese&wt=json')
response = json.load(connection)
print response['response']['numFound'], "documents found."
Print the name of each document.

for document in response['response']['docs']:
 print " Name =", document['name']
```

# Using SolrJ

SolrJ is an API that makes it easy for applications written in Java (or any language based on the JVM) to talk to Solr. SolrJ hides a lot of the details of connecting to Solr and allows your application to interact with Solr with simple high-level methods. SolrJ supports most Solr APIs, and is highly configurable.

## Building and Running SolrJ Applications

The SolrJ API ships with Solr, so you do not have to download or install anything else. But you will need to configure your build to include SolrJ and its dependencies.

### Common Build Systems

Most mainstream build systems greatly simplify dependency management, making it easy to add SolrJ to your project.

For projects built with Ant (using [Ivy](#)), place the following in your ivy.xml:

```
<dependency org="org.apache.solr" name="solr-solrj" rev="8.1.0"/>
```

For projects built with Maven, place the following in your pom.xml:

```
<dependency>
 <groupId>org.apache.solr</groupId>
 <artifactId>solr-solrj</artifactId>
 <version>8.1.0</version>
</dependency>
```

For projects built with Gradle, place the following in your build.gradle:

```
compile group: 'org.apache.solr', name: 'solr-solrj', version: '8.1.0'
```

### Adding SolrJ to the Classpath Manually

If you are not using one of the above build system, it's still easy to add SolrJ to your build.

At build time, all that is required is the SolrJ jar itself: solr-solrj-8.1.0.jar. To compile code manually that uses SolrJ, use a javac command similar to:

```
javac -cp .:$SOLR_HOME/dist/solr-solrj-8.1.0.jar ...
```

At runtime, you need a few of SolrJ's dependencies, in addition to SolrJ itself. For convenience, these dependencies are made available in the dist/solrj-lib directory. Run your project with a classpath like:

```
java -cp .:$SOLR_HOME/dist/solrj-lib/*:$SOLR_HOME/dist/solr-solrj-8.1.0.jar ...
```

If you are worried about the SolrJ libraries expanding the size of your client application, you can use a code obfuscator like [ProGuard](#) to remove APIs that you are not using.

## SolrJ Overview

For all its flexibility, SolrJ is built around a few simple interfaces.

All requests to Solr are sent by a SolrClient. SolrClient's are the main workhorses at the core of SolrJ. They handle the work of connecting to and communicating with Solr, and are where most of the user configuration happens.

Requests are sent in the form of SolrRequests, and are returned as SolrResponses.

## Types of SolrClients

SolrClient has a few concrete implementations, each geared towards a different usage-pattern or resiliency model:

- HttpSolrClient - geared towards query-centric workloads, though also a good general-purpose client. Communicates directly with a single Solr node.
- Http2SolrClient - async, non-blocking and general-purpose client that leverage HTTP/2. This class is experimental therefore its API's might change or be removed in minor versions of SolrJ.
- LBHttpSolrClient - balances request load across a list of Solr nodes. Adjusts the list of "in-service" nodes based on node health.
- LBHttp2SolrClient - just like LBHttpSolrClient but using Http2SolrClient instead. This class is experimental therefore its API's might change or be removed in minor versions of SolrJ.
- CloudSolrClient - geared towards communicating with SolrCloud deployments. Uses already-recorded ZooKeeper state to discover and route requests to healthy Solr nodes.
- ConcurrentUpdateSolrClient - geared towards indexing-centric workloads. Buffers documents internally before sending larger batches to Solr.
- ConcurrentUpdateSolrClient - just like ConcurrentUpdateSolrClient but using Http2SolrClient instead. This class is experimental therefore its API's might change or be removed in minor versions of SolrJ.

## Common Configuration Options

Most SolrJ configuration happens at the SolrClient level. The most common/important of these are discussed below. For comprehensive information on how to tweak your SolrClient, see the Javadocs for the involved client, and its corresponding builder object.

### Base URLs

Most SolrClient implementations (except for CloudSolrClient and Http2SolrClient) require users to specify one or more Solr base URLs, which the client then uses to send HTTP requests to Solr. The path users include on the base URL they provide has an effect on the behavior of the created client from that point on.

1. A URL with a path pointing to a specific core or collection (e.g., `http://hostname:8983/solr/core1`).  
When a core or collection is specified in the base URL, subsequent requests made with that client are not

required to re-specify the affected collection. However, the client is limited to sending requests to that core/collection, and can not send requests to any others.

2. A URL pointing to the root Solr path (e.g., `http://hostname:8983/solr`). When no core or collection is specified in the base URL, requests can be made to any core/collection, but the affected core/collection must be specified on all requests.

Generally speaking, if your SolrClient will only be used on a single core/collection, including that entity in the path is the most convenient. Where more flexibility is required, the collection/core should be excluded.

## Base URLs of Http2SolrClient

Since Http2SolrClient can manage connections to different nodes efficiently. Http2SolrClient does not require a baseUrl. In case of baseUrl is not provided, SolrRequest.basePath must be set, so Http2SolrClient can know which node it will request to. If not an IllegalArgumentException will be thrown.

## Timeouts

All SolrClient implementations allow users to specify the connection and read timeouts for communicating with Solr. These are provided at client creation time, as in the example below:

```
final String solrUrl = "http://localhost:8983/solr";
return new HttpSolrClient.Builder(solrUrl)
 .withConnectionTimeout(10000)
 .withSocketTimeout(60000)
 .build();
```

When these values are not explicitly provided, SolrJ falls back to using the defaults for the OS/environment it is running on.

## Querying in Solr

SolrClient has a number of query() methods for fetching results from Solr. Each of these methods takes in a SolrParams, an object encapsulating arbitrary query-parameters. And each method outputs a QueryResponse, a wrapper which can be used to access the result documents and other related metadata.

The following snippet uses a SolrClient to query Solr's "techproducts" example collection, and iterate over the results.

```
final SolrClient client = getSolrClient();

final Map<String, String> queryParamMap = new HashMap<String, String>();
queryParamMap.put("q", "*:*");
queryParamMap.put("fl", "id, name");
queryParamMap.put("sort", "id asc");
MapSolrParams queryParams = new MapSolrParams(queryParamMap);

final QueryResponse response = client.query("techproducts", queryParams);
final SolrDocumentList documents = response.getResults();

print("Found " + documents.getNumFound() + " documents");
for(SolrDocument document : documents) {
 final String id = (String) document.getFirstValue("id");
 final String name = (String) document.getFirstValue("name");

 print("id: " + id + "; name: " + name);
}
```

SolrParams has a SolrQuery subclass, which provides some convenience methods that greatly simplifies query creation. The following snippet shows how the query from the previous example can be built using some of the convenience methods in SolrQuery:

```
final SolrQuery query = new SolrQuery("*:*");
query.addField("id");
query.addField("name");
query.setSort("id", ORDER.asc);
query.setRows(numResultsToReturn);
```

## Indexing in Solr

Indexing is also simple using SolrJ. Users build the documents they want to index as instances of SolrInputDocument, and provide them as arguments to one of the add() methods on SolrClient.

The following example shows how to use SolrJ to add a document to Solr's "techproducts" example collection:

```
final SolrClient client = getSolrClient();

final SolrInputDocument doc = new SolrInputDocument();
doc.addField("id", UUID.randomUUID().toString());
doc.addField("name", "Amazon Kindle Paperwhite");

final UpdateResponse updateResponse = client.add("techproducts", doc);
// Indexed documents must be committed
client.commit("techproducts");
```



The indexing examples above are intended to show syntax. For brevity, they break several Solr indexing best-practices. Under normal circumstances, documents should be indexed in larger batches, instead of one at a time. It is also suggested that Solr administrators commit documents using Solr's autocommit settings, and not using explicit commit() invocations.

## Java Object Binding

While the UpdateResponse and QueryResponse interfaces that SolrJ provides are useful, it is often more convenient to work with domain-specific objects that can more easily be understood by your application. Thankfully, SolrJ supports this by implicitly converting documents to and from any class that has been specially marked with Field annotations.

Each instance variable in a Java object can be mapped to a corresponding Solr field, using the Field annotation. The Solr field shares the name of the annotated variable by default, however, this can be overridden by providing the annotation with an explicit field name.

The example snippet below shows an annotated TechProduct class that can be used to represent results from Solr's "techproducts" example collection.

```
public static class TechProduct {
 @Field public String id;
 @Field public String name;

 public TechProduct(String id, String name) {
 this.id = id; this.name = name;
 }

 public TechProduct() {}
}
```

Application code with access to the annotated TechProduct class above can index TechProduct objects directly without any conversion, as in the example snippet below:

```
final SolrClient client = getSolrClient();

final TechProduct kindle = new TechProduct("kindle-id-4", "Amazon Kindle Paperwhite");
final UpdateResponse response = client.addBean("techproducts", kindle);

client.commit("techproducts");
```

Similarly, search results can be converted directly into bean objects using the getBeans() method on QueryResponse:

```
final SolrClient client = getSolrClient();

final SolrQuery query = new SolrQuery("*:*");
query.addField("id");
query.addField("name");
query.setSort("id", ORDER.asc);

final QueryResponse response = client.query("techproducts", query);
final List<TechProduct> products = response.getBeans(TechProduct.class);
```

## Other APIs

SolrJ allows more than just querying and indexing. It supports all of Solr's APIs. Accessing Solr's other APIs is as easy as finding the appropriate request object, providing any necessary parameters, and passing it to the `request()` method of your `SolrClient`. `request()` will return a `NamedList`: a generic object which mirrors the hierarchical structure of the JSON or XML returned by their request.

The example below shows how SolrJ users can call the CLUSTERSTATUS API of SolrCloud deployments, and manipulate the returned `NamedList`:

```
final SolrClient client = getSolrClient();

final SolrRequest request = new CollectionAdminRequest.ClusterStatus();

final NamedList<Object> response = client.request(request);
final NamedList<Object> cluster = (NamedList<Object>) response.get("cluster");
final List<String> liveNodes = (List<String>) cluster.get("live_nodes");

print("Found " + liveNodes.size() + " live nodes");
```

# Using Solr From Ruby

Solr has an optional Ruby response format that extends the [JSON output](#) to allow the response to be safely eval'd by Ruby's interpreter

This Ruby response format differs from JSON in the following ways:

- Ruby's single quoted strings are used to prevent possible string exploits
  - \ and ' are the only two characters escaped...
  - unicode escapes not used... data is written as raw UTF-8
- nil used for null
- => used as the key/value separator in maps

Here's an example Ruby response from Solr, for a request like

`http://localhost:8983/solr/techproducts/select?q=iPod&wt=ruby&indent=on` (with Solr launching using `bin/solr start -e techproducts`):

```
{
 'responseHeader'=>{
 'status'=>0,
 'QTime'=>0,
 'params'=>{
 'q'=>'iPod',
 'indent'=>'on',
 'wt'=>'ruby'}}},
 'response'=>{ 'numFound'=>3, 'start'=>0, 'docs'=>[
 {
 'id'=>'IW-02',
 'name'=>'iPod & iPod Mini USB 2.0 Cable',
 'manu'=>'Belkin',
 'manu_id_s'=>'belkin',
 'cat'=>['electronics',
 'connector'],
 'features'=>['car power adapter for iPod, white'],
 'weight'=>2.0,
 'price'=>11.5,
 'price_c'=>'11.50,USD',
 'popularity'=>1,
 'inStock'=>false,
 'store'=>'37.7752,-122.4232',
 'manufacturedate_dt'=>'2006-02-14T23:55:59Z',
 '_version_'=>1491038048794705920},
 {
 'id'=>'F8V7067-APL-KIT',
 'name'=>'Belkin Mobile Power Cord for iPod w/ Dock',
 'manu'=>'Belkin',
 'manu_id_s'=>'belkin',
 'cat'=>['electronics',
 'connector']}
```

```

'features'=>['car power adapter, white'],
'weight'=>4.0,
'price'=>19.95,
'price_c'=>'19.95,USD',
'popularity'=>1,
'inStock'=>false,
'store'=>'45.18014,-93.87741',
'manufacturedate_dt'=>'2005-08-01T16:30:25Z',
'_version_=>1491038048792608768},
{
'id'=>'MA147LL/A',
'name'=>'Apple 60 GB iPod with Video Playback Black',
'manu'=>'Apple Computer Inc.',
'manu_id_s'=>'apple',
'cat'=>['electronics',
'music'],
'features'=>['iTunes, Podcasts, Audiobooks',
'Stores up to 15,000 songs, 25,000 photos, or 150 hours of video',
'2.5-inch, 320x240 color TFT LCD display with LED backlight',
'Up to 20 hours of battery life',
'Plays AAC, MP3, WAV, AIFF, Audible, Apple Lossless, H.264 video',
'Notes, Calendar, Phone book, Hold button, Date display, Photo wallet, Built-in games,
JPEG photo playback, Upgradeable firmware, USB 2.0 compatibility, Playback speed control,
Rechargeable capability, Battery level indication'],
'includes'=>'earbud headphones, USB cable',
'weight'=>5.5,
'price'=>399.0,
'price_c'=>'399.00,USD',
'popularity'=>10,
'inStock'=>true,
'store'=>'37.7752,-100.0232',
'manufacturedate_dt'=>'2005-10-12T08:00:00Z',
'_version_=>1491038048799948800}]
}}
```

Here is a simple example of how one may query Solr using the Ruby response format:

```

require 'net/http'

h = Net::HTTP.new('localhost', 8983)
http_response = h.get('/solr/techproducts/select?q=iPod&wt=ruby')
rsp = eval(http_response.body)

puts 'number of matches = ' + rsp['response']['numFound'].to_s
#print out the name field for each returned document
rsp['response']['docs'].each { |doc| puts 'name field = ' + doc['name'] }
```

For simple interactions with Solr, this may be all you need! If you are building complex interactions with Solr, then consider the libraries mentioned at <https://wiki.apache.org/solr/Ruby%20Response%20Format>

## Further Assistance

There is a very active user community around Solr and Lucene. The solr-user mailing list, and #solr IRC channel are both great resources for asking questions.

To view the mailing list archives, subscribe to the list, or join the IRC channel, please see  
<https://lucene.apache.org/solr/community.html>.

# Solr Glossary

These are common terms used with Solr.

# Solr Terms

Where possible, terms are linked to relevant parts of the Solr Reference Guide for more information.

**Jump to a letter:**

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

## A

### Atomic updates

An approach to updating only one or more fields of a document, instead of reindexing the entire document.

## B

### Boolean operators

These control the inclusion or exclusion of keywords in a query by using operators such as AND, OR, and NOT.

## C

### Cluster

In Solr, a cluster is a set of Solr nodes operating in coordination with each other via [ZooKeeper](#), and managed as a unit. A cluster may contain many collections. See also [SolrCloud](#).

### Collection

In Solr, one or more [Documents](#) grouped together in a single logical index using a single configuration and Schema.

In [SolrCloud](#) a collection may be divided up into multiple logical shards, which may in turn be distributed across many nodes, or in a Single node Solr installation, a collection may be a single [Core](#).

### Commit

To make document changes permanent in the index. In the case of added documents, they would be searchable after a *commit*.

### Core

An individual Solr instance (represents a logical index). Multiple cores can run on a single node. See also [SolrCloud](#).

### Core reload

To re-initialize a Solr core after changes to schema.xml, solrconfig.xml or other configuration files.

## D

### Distributed search

Distributed search is one where queries are processed across more than one [Shard](#).

## Document

A group of [fields](#) and their values. Documents are the basic unit of data in a [collection](#). Documents are assigned to [shards](#) using standard hashing, or by specifically assigning a shard within the document ID. Documents are versioned after each write operation.

## E

### Ensemble

A [ZooKeeper](#) term to indicate multiple ZooKeeper instances running simultaneously and in coordination with each other for fault tolerance.

## F

### Facet

The arrangement of search results into categories based on indexed terms.

### Field

The content to be indexed/searched along with metadata defining how the content should be processed by Solr.

## I

### Inverse document frequency (IDF)

A measure of the general importance of a term. It is calculated as the number of total Documents divided by the number of Documents that a particular word occurs in the collection. See <http://en.wikipedia.org/wiki/Tf-idf> and [the Lucene TFIDFSimilarity javadocs](#) for more info on TF-IDF based scoring and Lucene scoring in particular. See also [Term frequency](#).

### Inverted index

A way of creating a searchable index that lists every word and the documents that contain those words, similar to an index in the back of a book which lists words and the pages on which they can be found. When performing keyword searches, this method is considered more efficient than the alternative, which would be to create a list of documents paired with every word used in each document. Since users search using terms they expect to be in documents, finding the term before the document saves processing resources and time.

## L

### Leader

A single [Replica](#) for each [Shard](#) that takes charge of coordinating index updates (document additions or deletions) to other replicas in the same shard. This is a transient responsibility assigned to a node via an election, if the current Shard Leader goes down, a new node will automatically be elected to take its place. See also [SolrCloud](#).

## M

## Metadata

Literally, *data about data*. Metadata is information about a document, such as its title, author, or location.

## N

### Natural language query

A search that is entered as a user would normally speak or write, as in, "What is aspirin?"

## Node

A JVM instance running Solr. Also known as a Solr server.

## O

### Optimistic concurrency

Also known as "optimistic locking", this is an approach that allows for updates to documents currently in the index while retaining locking or version control.

## Overseer

A single node in [SolrCloud](#) that is responsible for processing and coordinating actions involving the entire cluster. It keeps track of the state of existing nodes, collections, shards, and replicas, and assigns new replicas to nodes. This is a transient responsibility assigned to a node via an election, if the current Overseer goes down, a new node will be automatically elected to take its place. See also [SolrCloud](#).

## Q

### Query parser

A query parser processes the terms entered by a user.

## R

## Recall

The ability of a search engine to retrieve *all* of the possible matches to a user's query.

## Relevance

The appropriateness of a document to the search conducted by the user.

## Replica

A [Core](#) that acts as a physical copy of a [Shard](#) in a [SolrCloud Collection](#).

## Replication

A method of copying a master index from one server to one or more "slave" or "child" servers.

## RequestHandler

Logic and configuration parameters that tell Solr how to handle incoming "requests", whether the requests are to return search results, to index documents, or to handle other custom situations.

## S

### SearchComponent

Logic and configuration parameters used by request handlers to process query requests. Examples of search components include facetting, highlighting, and "more like this" functionality.

### Shard

In SolrCloud, a logical partition of a single [Collection](#). Every shard consists of at least one physical [Replica](#), but there may be multiple Replicas distributed across multiple [Nodes](#) for fault tolerance. See also [SolrCloud](#).

### SolrCloud

Umbrella term for a suite of functionality in Solr which allows managing a [Cluster](#) of Solr [Nodes](#) for scalability, fault tolerance, and high availability.

### Solr Schema (managed-schema or schema.xml)

The Solr index Schema defines the fields to be indexed and the type for the field (text, integers, etc.) By default schema data can be "managed" at run time using the [Schema API](#) and is typically kept in a file named managed-schema which Solr modifies as needed, but a collection may be configured to use a static Schema, which is only loaded on startup from a human edited configuration file - typically named schema.xml. See [Schema Factory Definition in SolrConfig](#) for details.

### SolrConfig (solrconfig.xml)

The Apache Solr configuration file. Defines indexing options, RequestHandlers, highlighting, spellchecking and various other configurations. The file, solrconfig.xml, is located in the Solr home conf directory.

### Spell Check

The ability to suggest alternative spellings of search terms to a user, as a check against spelling errors causing few or zero results.

### Stopwords

Generally, words that have little meaning to a user's search but which may have been entered as part of a [natural language](#) query. Stopwords are generally very small pronouns, conjunctions and prepositions (such as, "the", "with", or "and")

### Suggester

Functionality in Solr that provides the ability to suggest possible query terms to users as they type.

### Synonyms

Synonyms generally are terms which are near to each other in meaning and may substitute for one another. In a search engine implementation, synonyms may be abbreviations as well as words, or terms that are not consistently hyphenated. Examples of synonyms in this context would be "Inc." and "Incorporated" or "iPod" and "i-pod".

## T

### Term frequency

The number of times a word occurs in a given document. See <http://en.wikipedia.org/wiki/Tf-idf> and [the Lucene TFIDFSimilarity javadocs](#) for more info on TF-IDF based scoring and Lucene scoring in particular. See also [Inverse document frequency \(IDF\)](#).

## Transaction log

An append-only log of write operations maintained by each [Replica](#). This log is required with SolrCloud implementations and is created and managed automatically by Solr.

## W

### Wildcard

A wildcard allows a substitution of one or more letters of a word to account for possible variations in spelling or tenses.

## Z

### ZooKeeper

Also known as [Apache ZooKeeper](#). The system used by SolrCloud to keep track of configuration files and node names for a cluster. A ZooKeeper cluster is used as the central configuration store for the cluster, a coordinator for operations requiring distributed synchronization, and the system of record for cluster topology. See also [SolrCloud](#).

## Errata

## Errata For This Documentation

Any mistakes found in this documentation after its release will be listed on the on-line version of this page:

<https://lucene.apache.org/solr/guide/8.1/errata.html>

## Errata For Past Versions of This Documentation

Any known mistakes in past releases of this documentation will be noted below.

# How to Contribute to Solr Documentation

The Lucene/Solr project has made it easy for anyone to contribute to the Solr Reference Guide with a patch.

The Guide is written in simple AsciiDoc-formatted files, and the source lives in the main Lucene/Solr source repository, right alongside the code.

Find information on how to contribute to documentation online at <https://lucene.apache.org/solr/guide/how-to-contribute.html>.