Solr: Index pdf, word etc (Tika)

Goal: Index pdf documents to make them searble using solr.

The following technique can be used to index bulk of pdf docs in batches one or multiple times. Later, we will see how to use a request handler to index doc on demand.

Download some pdfs and upload to them HDFS. Below is an example.

```
$ wget http://www8.hp.com/h20195/v2/GetPDF.aspx/4AA6-1049EEP.pdf -O
sample1.pdf
$ hadoop fs -mkdir pdfs
$ hadoop fs -put sample1.pdf pdfs
```

Using solrctl command create a collection template - docs. Name of collection is docs.

```
$ export NAME=docs
$ export SOLR_ZK_ENSEMBLE=localhost:2181/solr
```

Above, since `SOLR_ZK_ENSEMBLE` is created as environment variable we can avoid mentioned --zk argument in solrctl command.

```
$ solrctl instancedir --generate $NAME
```

From the `$NAME/conf/schema.xml` file remove existing fields if required add the following.

```
    <field name="content" type="text_general" indexed="true"
stored="true" />
    <field name="title" type="text_general" indexed="true"
stored="true" multiValued="true"/>
    <field name="subject" type="text_general" indexed="true"
stored="true"/>
    <field name="description" type="text_general" indexed="true"
stored="true"/>
    <field name="comments" type="text_general" indexed="true"
stored="true"/>
    <field name="author" type="text_general" indexed="true"
stored="true"/>
    <field name="keywords" type="text_general" indexed="true"
stored="true"/>
    <field name="category" type="text_general" indexed="true"
stored="true"/>
    <field name="resourcename" type="text_general" indexed="true"
stored="true"/>
    <field name="url" type="text_general" indexed="true"
stored="true"/>
    <field name="content_type" type="string" indexed="true"
stored="true" multiValued="true"/>
    <field name="last_modified" type="date" indexed="true"
stored="true"/>
    <field name="links" type="string" indexed="true" stored="true"
```

```
multiValued="true"/>
```

Create morphlines conf file `$NAME/conf/morphlines.conf`. with the content below.

```
solrLocator : {
  collection: docs
  zkHost : "127.0.0.1:2181/solr"
  batchSize : 100
}
morphlines: [
    {
        id : morphlinepdfs
        importCommands : ["org.kitesdk.**", "org.apache.solr.**"]
        commands : [
            { detectMimeType { includeDefaultMimeTypes : true } }
            {
                solrCell {
                solrLocator : ${solrLocator}
                captureAttr : true
                lowernames : true
                capture : [id, title, author, content, content_type,
subject, description, keywords, category, resourcename, url,
last_modified, links]
                parsers : [ { parser :
org.apache.tika.parser.pdf.PDFParser } ]
                }
            }
            { generateUUID { field : id } }
            { sanitizeUnknownSolrFields { solrLocator :
${solrLocator} } }
            { loadSolr: { solrLocator : ${solrLocator} } }
        ]
    }]
```

Upload the configuration to Zookeeper.

```
$ solrctl instancedir --create $NAME $NAME
```

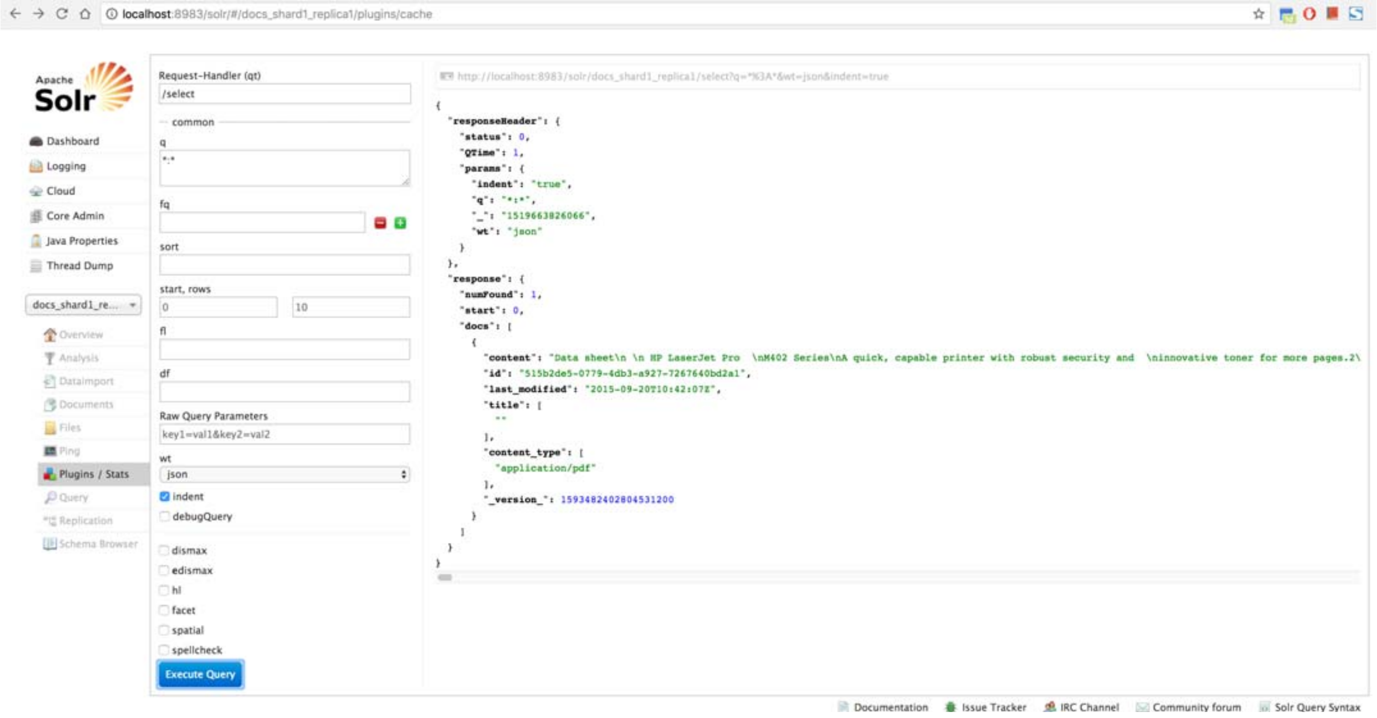
Create a solr collection

```
$ solrctl collection --create $NAME
```

Run the MapReduce indexer job.

```
$ hadoop jar /usr/lib/solr/contrib/mr/search-mr-job.jar \
org.apache.solr.hadoop.MapReduceIndexerTool \
--zk $SOLR_ZK_ENSEMBLE \
--collection $NAME \
--morphline-file $NAME/conf/morphlines.conf \
--go-live \
--output hdfs://localhost:8020/tmp/$NAME_out \
--verbose \
hdfs://localhost:8020/user/cloudera/pdfs
```

Open Solr UI, you should find the doc.



Request-Handler (qt): /select

common

q: *:*

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

□ debugQuery

□ dismax

□ edismax

□ hl

□ facet

□ spatial

□ spellcheck

http://localhost:8983/solr/docs_shard1_replica1/select?q=*%3A*&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "*:*",
      "_": "1519663826066",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 1,
    "start": 0,
    "docs": [
      {
        "content": "Data sheet\n \n HP LaserJet Pro  \nM402 Series\nA quick, capable printer with robust security and  \ninnovative toner for more pages.2\
        "id": "515b2de5-0779-4db3-a927-7267640bd2a1",
        "last_modified": "2015-09-20T10:42:07Z",
        "title": [
          ""
        ],
        "content_type": [
          "application/pdf"
        ],
        "_version_": 1593482402804531200
      }
    ]
  }
}
```

Using Custom Handler Extraction

This method can be used to index a pdf, word doc using on demand.

Add the following dynamic field to the schema - `$NAME/conf/schema.xml`.

```
<dynamicField name="ignored_*" type="text_general" indexed="true"
stored="true"/>
```

Add the following update handler in `$NAME/conf/solrconfig.xml`

```
<requestHandler name="/update/extract"
class="org.apache.solr.handler.extraction.ExtractingRequestHandler">
    <lst name="defaults">
      <str name="fmap.Last-Modified">last_modified</str>
      <str name="uprefix">ignored_</str>
    </lst>
</requestHandler>
```

Also at top of the `$NAME/conf/solrconfig.xml` file add the following library location. If the jars are not present, download them from internet.

```
<lib path="/usr/lib/solr/solr-cell.jar" />
<lib path="/usr/lib/solr/tika-core-1.9.jar" />
<lib path="/usr/lib/solr/apache-xml-xerces.jar" />
```

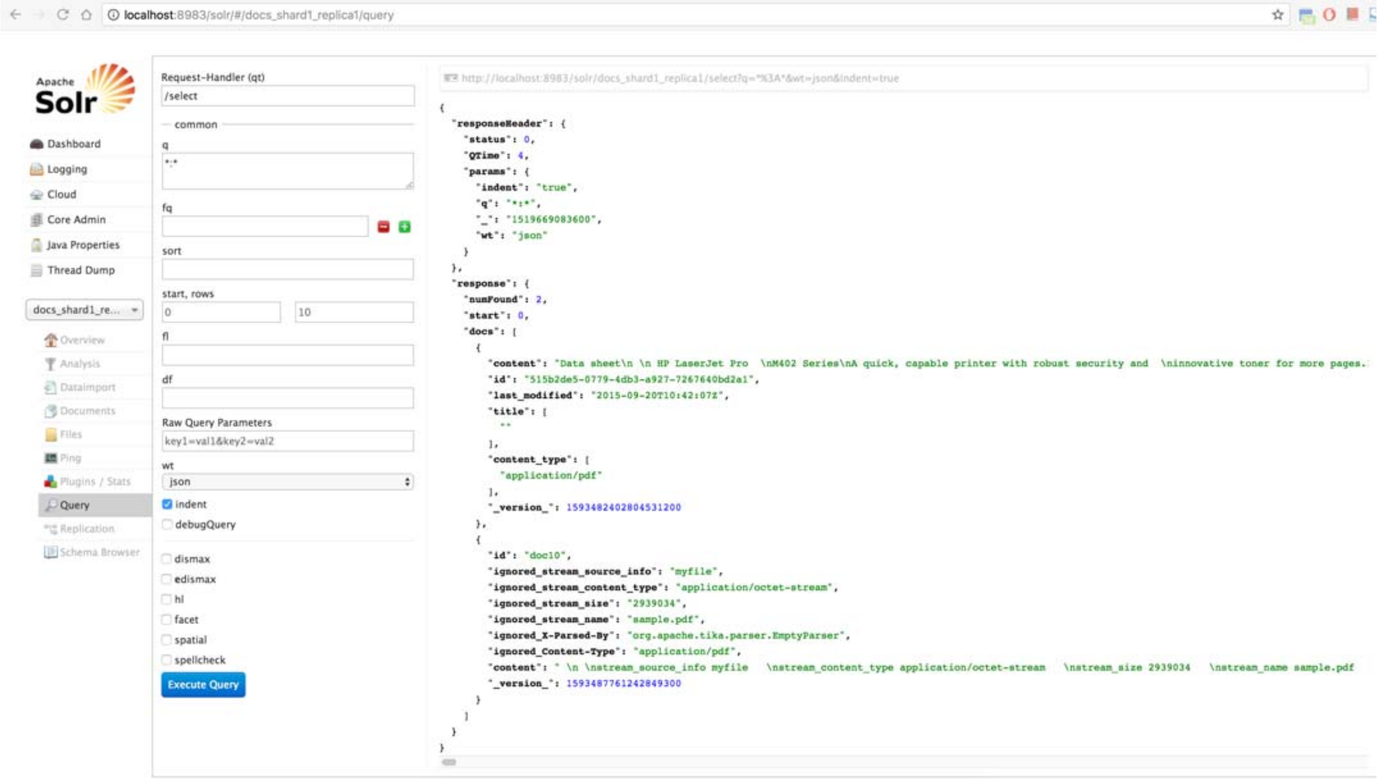
Update and reload the collection information.

```
$ solrctl instancedir --update $NAME $NAME
$ solrctl collection --reload $NAME
```

Using curl, send one doc for indexing.

```
$ curl 'http://localhost:8983/solr/docs/update
/extract?literal.id=doc10&commit=true' -F "myfile=@sample.pdf"
```

Open Solr UI, you should be able to find the doc.



Request-Handler (qt): /select
q: *:*
fq:
sort:
start, rows: 0 10
fl:
df:
Raw Query Parameters: key1=val1&key2=val2
wt: json
☑ indent
□ debugQuery
□ dismax
□ edismax
□ hl
□ facet
□ spatial
□ spellcheck

http://localhost:8983/solr/docs_shard1_replica1/select?q=*%3A*&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 4,
    "params": {
      "indent": "true",
      "q": "*:*",
      "_": "1519669083600",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 2,
    "start": 0,
    "docs": [
      {
        "content": "Data sheet\n \n HP LaserJet Pro  \nM402 Series\nA quick, capable printer with robust security and  \ninnovative toner for more pages.
        "id": "515b2de5-0779-4db3-a927-7267640bd2a1",
        "last_modified": "2015-09-20T10:42:07Z",
        "title": [
          ""
        ],
        "content_type": [
          "application/pdf"
        ],
        "_version_": 1593482402804531200
      },
      {
        "id": "doc10",
        "ignored_stream_source_info": "myfile",
        "ignored_stream_content_type": "application/octet-stream",
        "ignored_stream_size": "2939034",
        "ignored_stream_name": "sample.pdf",
        "ignored_X-Parsed-By": "org.apache.tika.parser.EmptyParser",
        "ignored_Content-Type": "application/pdf",
        "content": " \n \nstream_source_info myfile   \nstream_content_type application/octet-stream   \nstream_size 2939034   \nstream_name sample.pdf
        "_version_": 1593487761242849300
      }
    ]
  }
}
```