HumulusSOLR

The module HumulusSOLR provides a SOLR storage handler for CumulusRDF. Bulk upload of RDF triples, individual triple insert and triple lookup operations are implemented. Processing of SPARQL queries is a responsibility of CumulsRDF which uses SESAME and operates also on multiple storage layers.

Index Structure

Each individual triple will be stored as a single document in SOLR. The elements `s`, `p`, and `o` are used to create a full index as in the CumulusRDF paper (1). Prefixes have to be resolved, we need full URLs for each part of the triple or a literal.

Triple Preparation

The flume based automatic triple import needs one triple at each input line. Multiple input formats can be converted into the N-Turtle format. The webservice is used in our scripts.

```
#!/bin/sh
curl -F "userid=1" -F "filecomment=This is an RDF XML file" -F "content=@$1/$2" \
http://rdf-translator.appspot.com/convert/xml/nt/content > $FLUME_SPOOL_DIR/$2.nt
```

Listing 1: **Webservice for online RDF file conversion.** This curl command uploads a local RDF-XML file and converts it into N-Turtle fomat.

Triple Import

An Apache-Flume source (see listing 2) observes a spooling-directory. All incoming files automatically loaded and renamed after processing. Flume maintains metadata to prevent redundant processing of the same file.

```
# Components on this agent
agent1.sources = spooldir-source
agent1.sinks = morphline-sink
agent1.channels = memory-channel

# Describe/configure the source
agent1.sources.spooldir-source.type = spooldir
agent1.sources.spooldir-source.spoolDir = /flume/triple_files_spooldir
agent1.sources.spooldir-source.batchSize= 100
agent1.sources.spooldir-source.channels = memory-channel

# Solr Sink Using Morphlines
agent1.sinks.morphline-sink.type=org.apache.flume.sink.solr.morphline.MorphlineSolrSink
agent1.sinks.morphline-sink.morphlineFile=triplestore-morphlines-v2.conf
agent1.sinks.morphline-sink.channel = memory-channel

# Use a channel which buffers events in memory
agent1.channels.memory-channel.type = memory
agent1.channels.memory-channel.capacity = 10000
agent1.channels.memory-channel.transactionCapacity = 10000
```

Listing 2: **Flume configuration.** This Flume agent observes a directory and writes all files into a MorphlineSink.

Each line is processed by a Kite-Morphline (see listing 3) and sent to SOLR.

```
SOLR_LOCATOR : {
    # Name of solr collection
    collection : triple_collection2
    # ZooKeeper ensemble
    zkHost : "dev.loudacre.com:2181/solr"
}
```

```
morphlines : [
{
    id : morphline1
    importCommands :  [
              "com.cloudera.**",
              "com.cloudera.cdk.morphline.stdlib.**",
              "org.apache.solr.**"]
    commands : [
           {   # Reads the incoming N-Triples, one on each line as plain text
               readLine {
                      charset : UTF-8
                      commentPrefix : "#"
               }
           },
           {   # Extracts the relevant 3 of the triple and creates
               # the fields for our index
               split {
                      inputField : message
                      outputFields : [s, p, o, d]
                      separator : " "
                      isRegex : false
                      addEmptyStrings : true
                      trim : true
               }
           },
           {   setValues {
                      triple : "@{message}"
                      spo : "@{s}@{p}@{o}"
                      spx : "@{s}@{p}"
                      xpo : "@{p}@{o}"
                      xpx : "@{p}"
                      sxo : "@{s}@{o}"
                      xxo : "@{o}"
                      sxx : "@{s}"
               }
           },
           {   generateUUID { field : id } },
           {   addCurrentTime {} },
           {   sanitizeUnknownSolrFields {
                   # Location from which to fetch Solr schema
                   solrLocator : ${SOLR_LOCATOR}
               }
           },
           {   loadSolr {
                   solrLocator : ${SOLR_LOCATOR}
               }
           }
    ]
}
]
```

Listing 3: **Triple index preprocessing Morphline.** This Morphline creates the index fields for triple pattern lookups and stores a triple with unique id and timestamp (procesing time) in an SOLR collection.

Before this Morphline can be used, a SOLR collection has to be created. The field definition for the SOLR schema is shown in listing 4.

```
<fields>
        <field name="id" type="string" indexed="true" stored="true" multiValued="false" />
        <field name="triple" type="string" indexed="true" stored="true" required="false" multiValued="false" />
        <field name="spo" type="string" indexed="true" stored="true" required="false" multiValued="false" />
        <field name="spx" type="string" indexed="true" stored="true" required="false" multiValued="false" />
        <field name="xpo" type="string" indexed="true" stored="true" required="false" multiValued="false" />
        <field name="xpx" type="string" indexed="true" stored="true" required="false" multiValued="false" />
        <field name="sxo" type="string" indexed="true" stored="true" required="false" multiValued="false" />
        <field name="sxx" type="string" indexed="true" stored="true" required="false" multiValued="false" />
        <field name="xxp" type="string" indexed="true" stored="true" required="false" multiValued="false" />
        <field name="_version_" type="long" indexed="true" stored="true"/>
 </fields>
 <uniqueKey>id</uniqueKey>
```

Listing 4: **Schema definition for a SOLR triple index.**

Download / Installation

All project files are on github. Clone the main project.

```
$> git clone https://github.com/kamir/Humulus
```

All index operations are done in the HumulusSOLR subproject. The files in src/main are used as templates. Please fork the project and work on your own version of the files. After you created great new morphlines and send a pull-request.

Triple Index Maintainance

The flume agent configuration is in: `Humulus/HumulusSOLR/src/main/FLUME`
The SOLR core configuration is in: `Humulus/HumulusSOLR/src/main/SOLR`
The data peparation tools are in: `Humulus/HumulusSOLR/src/main/PREP.TESTDATA`

The spool-directory is in: `/flume/triple-files-spooldir`.

Setup

```
$> export COLLECTION=triple_collection2
```

Prepare NT file from RDF-XML

The directory `PREP.TESTDATA` contains some converterscripts.

```
$> convertrdfxml2nt.sh . dc-2010-complete.rdf
$> bulkconvertrdfxml2nt.sh bulk
```

Create and Deploy Collection

```
$> solrctl --zk dev.loudacre.com:2181/solr instancedir \
   --create $COLLECTION triplestore_search_config
$> solrctl --zk dev.loudacre.com:2181/solr collection \
   --create $COLLECTION
```

Clear Index

```
$> solrctl --zk dev.loudacre.com:2181/solr collection \
   --deletedocs $COLLECTION
```

Start Flume Agent

```
$> flume-ng agent --conf /etc/flume-ng/conf \
   --conf-file triple-import-flume-v2.conf \
   --name agent1 -Dflume.root.logger=INFO,console
```