**Solr Chinese Segmentation and Cloudera Full Text Search**

**Author: Michelle Zhang**

**Company: Cloudera**

**Date: 10/26/2014**

# Environment

After installing and starting Solr + Flume + Lily HBase +Hue on CDH 5.2, you can deploy Chinese Segmentation on CDH Cloudera Search for both Solr Search UI and Hue Search UI, as well as CML.

# Download and Install Chinese Segmentation

## Download lucene-analyzers-smartcn

Download lucene-analyzers-smartcn-4.4.0-cdh5.2.0.jar from following repo which is built from CDH 5.2.

<https://repository.cloudera.com/artifactory/repo/org/apache/lucene/lucene-analyzers-smartcn/4.4.0-cdh5.2.0/>

The Chinese Segmentation lucene-analyzers-smartcn can be used by solr 4.4 and CDH 5.2, and its accuracy of Chinese word segmentation is good.

## Install Chinese Segmentation in CDH

### Copy the jar to every Cluster Solr node

Copy the jar to **EVERY** Solr node’ following path:

/opt/cloudera/parcels/CDH/lib/solr/webapps/solr/WEB-INF/lib/

Remember to **chmod** the permission of your new jars as same as other jars under the same dir.

Notice: copy the jar into MapReduce classpath or HBase classpath if you are using Chinese segmentation in your job, for example:

# scp /opt/cloudera/parcels/CDH/lib/solr/webapps/solr/WEB-INF/lib/lucene-analyzers-smartcn-4.4.0-cdh5.2.0.jar root@172.31.12.151:/opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/lib/

### Restart Solr Services to Enable Chinese Segmentation

Use CM UI to restart Solr services to enable Chinese Segmentation, the you will see the jar is updated into /var/lib/solr/tomcat-deployment/webapps/solr/WEB-INF/lib/ on your solr nodes.

# Configure Chinese Segmentation for Your Instance

## Add Chinese Segmentation Definition schema.xml

Append following line in your schema.xml to add Chinese Segmentation field type definition:

<fieldType name="**text\_smart**" class="solr.TextField" positionIncrementGap="100">

<analyzer type="index">

<tokenizer class="solr.SmartChineseSentenceTokenizerFactory"/>

<filter class="solr.SmartChineseWordTokenFilterFactory"/>

</analyzer>

<analyzer type="query">

<tokenizer class="solr.SmartChineseSentenceTokenizerFactory"/>

<filter class="solr.SmartChineseWordTokenFilterFactory"/>

</analyzer>

</fieldType>

## Use Chinese Segmentation Type as Field Type in schema.xml

Use Chinese Segmentation type as field type for your field definition in your schema.xml:

<field name="**sma**" type="**text\_smart**" indexed="true" stored="true" multiValued="true"/>

# Create Collection and Indexer

Create collection and indexer with the configuration and data.

Restart Solr services.

# Full Text Search Result with Chinese Segmentation

Chinese Segmentation with solr full text search in CDH can be by both Solr web UI and CDH Hue UI.

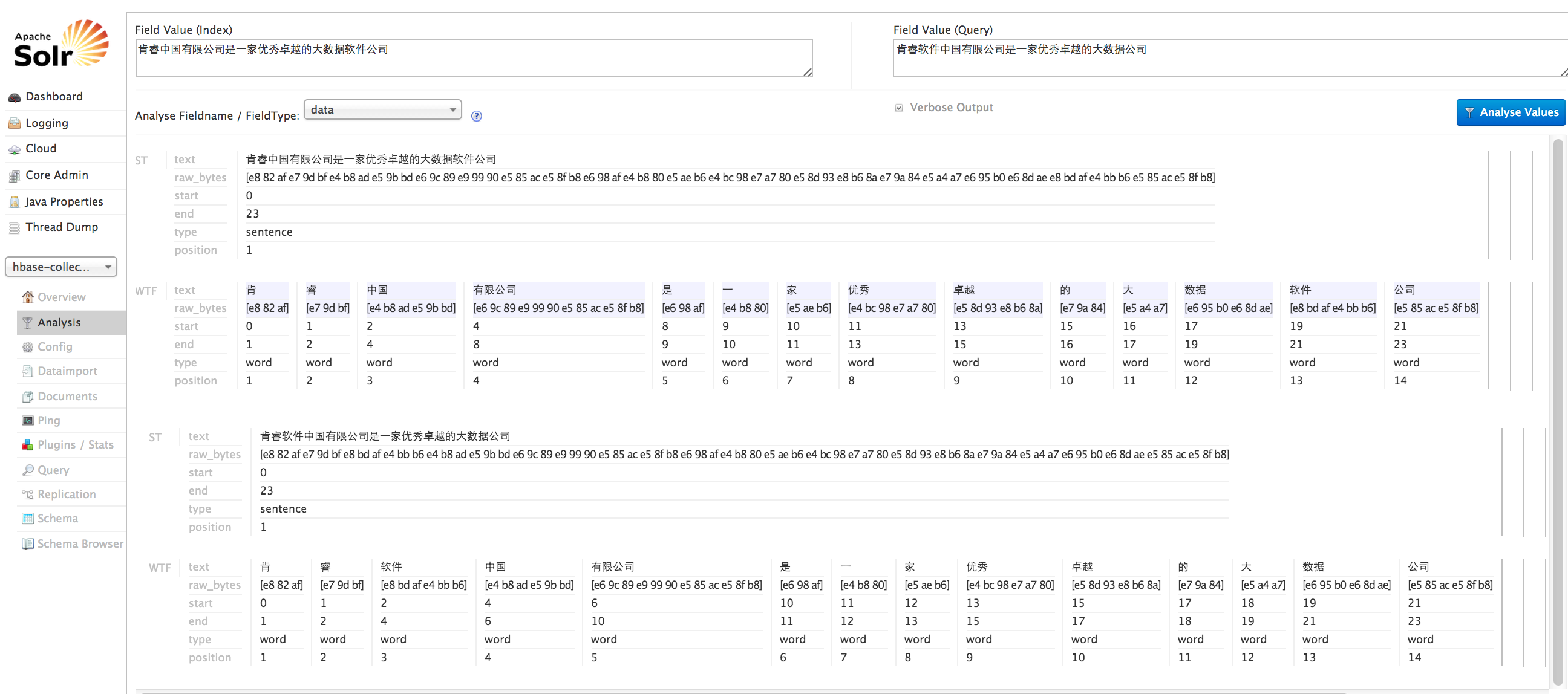
Following are search results.

## Solr UI Analysis Result

Select Core in the left “Core Selector” DropDownList. For example “collection1\_shard1\_replica1”

* Select “Analysis” buttom
* Fill your words in Chinese in the “Field Value (Index)” or “Field Value (Query)”
* Choose “**smc**” or “**text\_smart**” which you defined in schema.xml in “Analyse Fieldname / FieldType” DropDownList.
* Click “Analyse Value”

Then it will show as following:

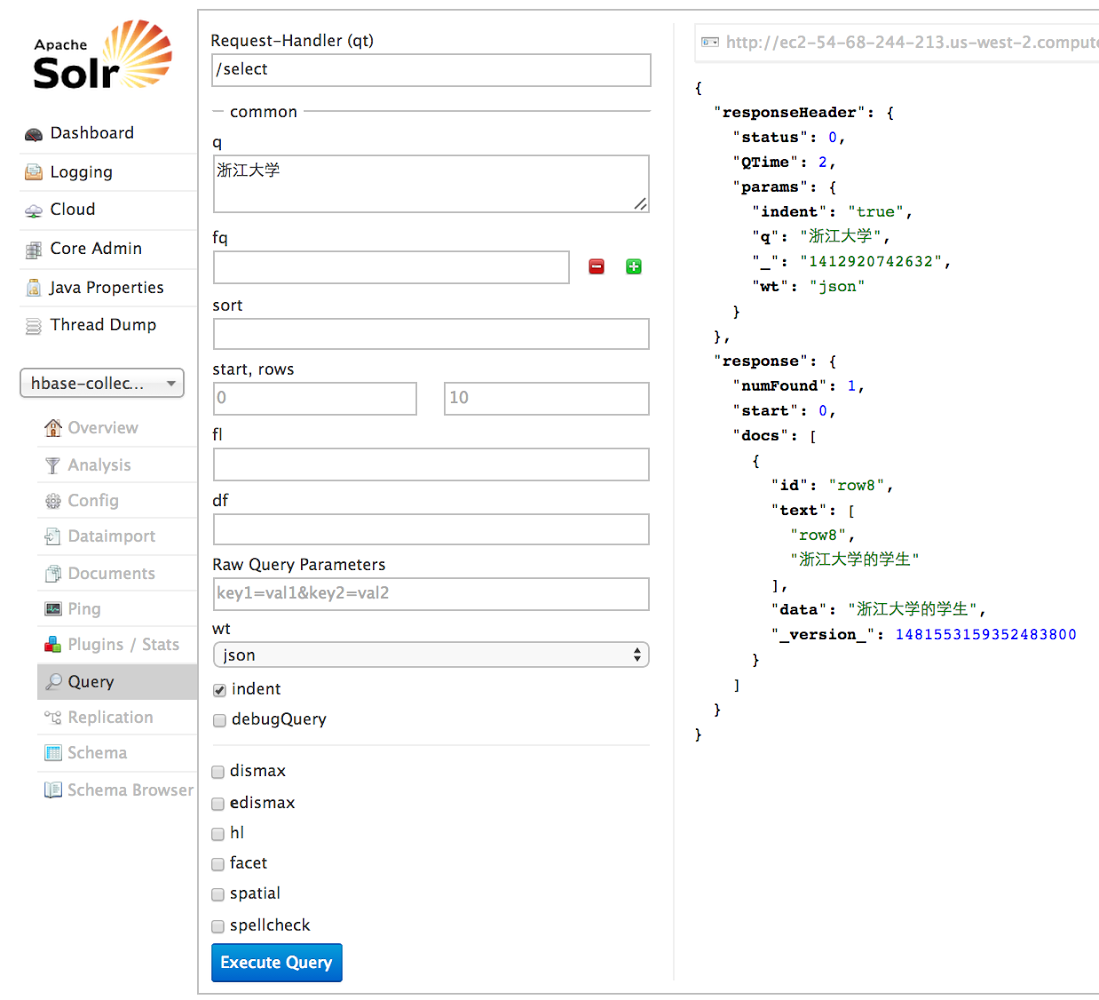


## Solr Query UI Result

Select Core in the left “Core Selector” DropDownList. For example “collection1\_shard1\_replica1”

* Select “Query” buttom
* Fill your query conditions
* Click “Execute Query”

Then it will show as following:



## Cloudera Search Hue UI Result

Enter CM “Hue Web UI”.

* Click in sequence as “Search” -> “Indexes” -> select your collection, for example “hbase-collection1” -> “ACTIONS” -> “Search”.
* Fill what keyword that you want to search in the Search bar

Then it will show as following:

