B649 Cloud Computing Project 5 Report

HBase Inverted Indexing

Team Members: Yash Ketkar (yketkar@indiana.edu) | Neelam Tikone (ntikone@indiana.edu)

The main **steps** in running this program are:

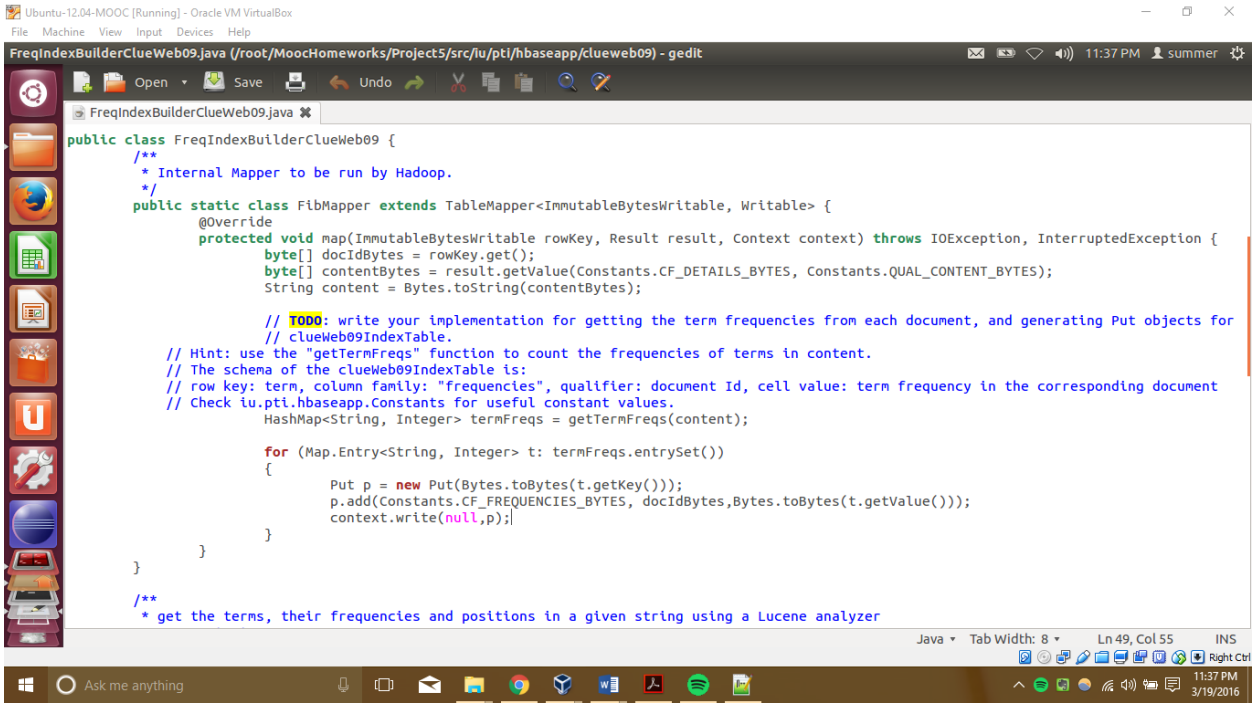
1) We first need to configure the working environment, start Hadoop and HBase, create HBase tables and load data into HBase.

```
# start hadoop
$ cd /root/software/hadoop-1.1.2/
$ . ./MultiNodesOneClickStartUp.sh /root/software/jdk1.6.0_33/ nodes

# start hbase
$ cd /root/software/hbase-0.94.7/
$ ./bin/start-hbase.sh

# prepare for hadoop and hbase environment
$ cp /root/software/hbase-0.94.7/conf/hbase-site.xml /root/software/hadoop-
1.1.2/conf/
$ cd /root/software/hadoop-1.1.2/
$ export HADOOP_CLASSPATH=`/root/software/hbase-0.94.7/bin/hbase classpath`

# create hbase tables
$ ./bin/hadoop jar lib/cglHBaseMooc.jar
iu.pti.hbaseapp.clueweb09.TableCreatorClueWeb09

# create one directory for mapreduce data input
$ mkdir -p /root/MoocHomeworks/HBaseWordCount/data/clueweb09/mrInput

# create input's metadata for HBbase data loader
$ ./bin/hadoop jar lib/cglHBaseMooc.jar iu.pti.hbaseapp.clueweb09.Helpers
create-mr-input /root/MoocHomeworks/HBaseWordCount/data/clueweb09/files/
/root/MoocHomeworks/HBaseWordCount/data/clueweb09/mrInput/ 1

# copy metadata to Hadoop HDFS
$ ./bin/hadoop dfs -copyFromLocal
/root/MoocHomeworks/HBaseWordCount/data/clueweb09/mrInput/ /cw09LoadInput
$ ./bin/hadoop dfs -ls /cw09LoadInput

# load data into HBase (takes 10-20 minutes to finish)
$ ./bin/hadoop jar lib/cglHBaseMooc.jar
iu.pti.hbaseapp.clueweb09.DataLoaderClueWeb09 /cw09LoadInput
```

2) Thus we have uploaded records with the iu.pti.hbaseapp.clueweb09.DataLoaderClueWeb09 from the clueWeb09DataTable. Then to run the HBaseInvertedIndexing program we use the following commands.

```
summer@ubuntu:~$ cd /root/MoocHomeworks/Project5/
summer@ubuntu:/root/MoocHomeworks/Project5$
compileAndExecFreqIndexBuilderClueWeb.sh
```

Updated Code:



Output: