

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 HBase 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:

```
Ubuntu-12.04-MOOC [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help

FreqIndexBuilderClueWeb09.java (/root/MoocHomeworks/Project5/src/lu/pti/hbaseapp/clueweb09) - gedit

FreqIndexBuilderClueWeb09.java
public class FreqIndexBuilderClueWeb09 {
    /**
     * Internal Mapper to be run by Hadoop.
     */
    public static class FibMapper extends TableMapper<ImmutableBytesWritable, Writable> {
        @Override
        protected void map(ImmutableBytesWritable rowKey, Result result, Context context) throws IOException, InterruptedException {
            byte[] docIdBytes = rowKey.get();
            byte[] contentBytes = result.getValue(Constants.CF_DETAILS_BYTES, Constants.QUAL_CONTENT_BYTES);
            String content = Bytes.toString(contentBytes);

            // TODO: write your implementation for getting the term frequencies from each document, and generating Put objects for
            // clueWeb09IndexTable.
            // Hint: use the "getTermFreqs" function to count the frequencies of terms in content.
            // The schema of the clueWeb09IndexTable is:
            // row key: term, column family: "frequencies", qualifier: document Id, cell value: term frequency in the corresponding document
            // Check lu.pti.hbaseapp.Constants for useful constant values.
            HashMap<String, Integer> termFreqs = getTermFreqs(content);

            for (Map.Entry<String, Integer> t: termFreqs.entrySet())
            {
                Put p = new Put(Bytes.toBytes(t.getKey()));
                p.add(Constants.CF_FREQUENCIES_BYTES, docIdBytes, Bytes.toBytes(t.getValue()));
                context.write(null, p);
            }
        }
    }

    /**
     * get the terms, their frequencies and positions in a given string using a Lucene analyzer
     */
}
```

Output:

```
Ubuntu-12.04-MOOC [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help

project2.txt (/root/MoocHomeworks/Project5/output) - gedit

project2.txt
scanning table clueWeb09IndexTable on frequencies...
-----0'1-----
00000230265 : 1
-----0'23.08-----
00000235243 : 1
-----0,0.00,1,0.00-----
00000118373 : 1
-----0,0.00,1,0.00,2,0.00-----
00000118369 : 1
00000118370 : 1
00000118371 : 1
00000118372 : 1
-----0,0.00,1,0.00,2,0.00,3,0.00,4,0.00,5,0.00,6,0.00,7,0.00,8,0.00,9,0.00-----
00000118368 : 1
-----0,01euros-----
00000226930 : 1
-----0,1.7,5.0-----
00000231836 : 1
-----0,28804,1690753_1690758_1693514,00-----
00000121800 : 1
-----0,4458,360183_395924,00-----
00000121979 : 1
-----0,5px-----
00000200871 : 4
00000200872 : 4
00000200873 : 4
00000200874 : 4
-----0,8-----
00000230251 : 1
-----0,98mb-----
00000108663 : 1
```