HBASE WORDCOUNT

Sohil Jain Ashutosh Bhargave

[**sohjain@iu.edu**](mailto:sohjain@iu.edu)[**ashubhar@indiana.edu**](mailto:ashubhar@indiana.edu)

**Goal**

Write an HBase WordCount program to count all unique terms’ occurrences from the clueWeb09 dataset. Each row record of columnfamily “frequencies” is unique; the rowkey is the unique term stored in byte format, column name is “count” and value is the term frequency shown in all documents. Load the result to HBase WordCountTable.

**Technical Report:**

**Flow:**

Start Hadoop, HBASE and Insert Data files into wordclueWeb09 table in Hbase

Read The Hbase Table in the Mapper and send the count to the Reducer which calculates overall frequency.

Read The Hbase Table in the Mapper and send the count to the Reducer which calculates overall frequency.

Insert the word and Frequency to the wordcount table in Hbase.

In the Map task we followed these steps,

1] Get key and value as URI and Content. Stored the content in the string named “content”

2] Call “getWordFreq(content)” function which returns the key value pair of word and the count of that word in that row,

So, consider I have 100 rows as,

|  |  |
| --- | --- |
| URI | Content |
| “http://www.asd.html” | Indiana University is a good university. |
| “http://www.indiana.html” | Indiana is developing state. |
| …. | ….. |
| “http://www.q.html” | Indiana University ranks 6th in MIS. |

We will iterate over all these rows and getWordFreq will first give count of all the words in first row as,

Indiana 1

University 2

Is 1

A 1

Good 1

This iterates over all the rows, and we store it in a hashmap. We send the counts of words **from different mappers** to the reducer which calculates the total count of the words.

3] When reducer calculates all the frewuencies, we put these values in the wordcounttable in HBASE.

For this we use the constants file which has the column family and column names initialized to a value. Reducer inserts the values in to the “frequencies” **column family** and “count” **column.**

Put put = new Put(rowKey);

put.add(Constants.CF\_FREQUENCIES\_BYTES, Constants.QUAL\_COUNT\_BYTES, contentBytes);

**Compile and run your code**

$ cd /root/MoocHomeworks/HBaseWordCount

$ ./compileAndExecWordCount.sh

**Result**

The result is generated into a project1.txt file. The output is as follows,

scanning table WordCountTable on frequencies...

------------0'1------------

count : 1

------------0'23.08------------

count : 1

------------0,0.00,1,0.00------------

count : 1

------------0,0.00,1,0.00,2,0.00------------

count : 4

------------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------------

count : 1

------------0,01euros------------

count : 1

------------0,1.7,5.0------------

count : 1

------------0,28804,1690753\_1690758\_1693514,00------------

count : 1

------------0,4458,360183\_395924,00------------

count : 1

------------0,5px------------

count : 16