HBASE WORD COUNT

Project Report – Hbase Word Count

Indiana University

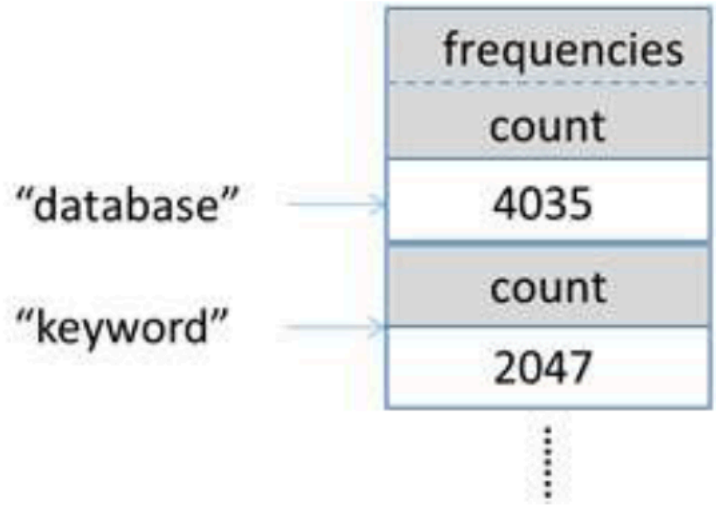
Revision	Date	Description	Author
1.0	04-Mar-2017	Coding	Karthik Vegi
1.1	20-Mar-2017	Testing and report	Melita Dsouza

HBASE WORD COUNT

PROJECT DESCRIPTION

In this project, we calculate the frequency of words appearing in the ClueWeb09 dataset stored in HBase.

The word count schema is as shown below:



The table schema for the Clueweb09 data set is shown below:

Data Flow and Logic

Mapper Code:

- The Mapper code counts the number of times a word appears in each input record of the HBase and then outputs it to the Reducer
- The map function receives as input the ***<key, value>*** pair in the form of ***<ImmutableBytesWritable row, Result result>*** where each row in the HBase record is related to a specified URI and result is the text stored that belongs to that URI. The content in Hbase table is accessed using the below code:

```
byte[] contentBytes = result.getValue(Constants.CF_DETAILS_BYTES,
Constants.QUAL_CONTENT_BYTES);
```

- where ***Constants.CF_DETAILS_BYTES*** indicate the column family and ***Constants.QUAL_CONTENT_BYTES*** is the column content

HBASE WORD COUNT

- The output ***<key, value>*** pair of this function is ***<Text word, LongWritable freqs>*** where word is the tokenized word extracted from the input row and freqs is the frequency of that word returned by the function getWordFreq()

Reducer Code:

- The Reducer code counts the final frequency of each particular word, adding all the partial results from the Map function
- The reducer function receives the ***<key, value>*** pair in the form ***<Text word, Iterable<LongWritable> freqs>,*** which is the output of the Mapper phase
- After counting the final values for each distinct word, the reducer outputs a Put object to add a row in the **WordCountTable** in the HBase with the count for a specific word
- The Put object is filled with the following command:

```
WordCountTable.add(Bytes.toBytes("frequencies"), Bytes.toBytes("count"), Bytes.toBytes(sum));
```

Output

project4_output.txt

```
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
------------0,5px8,360183_395924,00------------
count : 16
```