Shadman Ahmed
Amir Zamli

MapReduce, HDFS, HBase

Introduction

This lab introduced the use of a basic implementation of distributed file systems. The main parts of the assignment involved a Hadoop distributed file system which, for instance, provides a distributed storage for large amounts of data. Hadoop MapReduce framework, works as a pipeline for data-processing. And HBase, which serves as the structure of the data. For the assignment, the input data contained different users with id numbers, reputation and other types of information. Based on the reputation, the aim was to extract the top ten users with the help of the Hadoop distributed file system and the MapReduce framework.

Implementation

The input data contained various user data in xml format separated by rows. Each mapper unit processes the file blocks row by row. For each input value(row) we extracted the id, if the id is equal to -1 or not found at all, the row gets filtered away, otherwise the reputation is extracted and gets inserted into a TreeMap as the key, with the raw row data as value. When the mapper unit is completed the cleanup function gets executed, here we can utilize the sorting feature of tree map by using the descendingMap method and then iterated through the ten first items and wrote that to the context with null as intermediate key, as we are only using one reducer. This was served as the input to the reducer. We initialized one reducer which took all the potential top ten users from the mappers and iterated through all of them. With the help of TreeMap again extracted ten users with the highest reputation value. And stored them in an HBase table in regard to their id number. Inserted columns of id and reputation to see for which user id had the corresponding reputation and id number of the top ten users.

Result

```
=> Hbase::Table - topten
hbase(main):022:0> scan 'topten'
ROW                                   COLUMN+CELL
 108                                  column=info:id, timestamp=1537041290060, value=108
 108                                  column=info:rep, timestamp=1537041290060, value=2127
 11097                                column=info:id, timestamp=1537041290060, value=11097
 11097                                column=info:rep, timestamp=1537041290060, value=2824
 21                                   column=info:id, timestamp=1537041290060, value=21
 21                                   column=info:rep, timestamp=1537041290060, value=2586
 2452                                 column=info:id, timestamp=1537041290060, value=2452
 2452                                 column=info:rep, timestamp=1537041290060, value=4503
 381                                  column=info:id, timestamp=1537041290060, value=381
 381                                  column=info:rep, timestamp=1537041290060, value=3638
 434                                  column=info:id, timestamp=1537041290060, value=434
 434                                  column=info:rep, timestamp=1537041290060, value=2131
 548                                  column=info:id, timestamp=1537041290060, value=548
 548                                  column=info:rep, timestamp=1537041290060, value=2289
 836                                  column=info:id, timestamp=1537041290060, value=836
 836                                  column=info:rep, timestamp=1537041290060, value=1846
 84                                   column=info:id, timestamp=1537041290060, value=84
 84                                   column=info:rep, timestamp=1537041290060, value=2179
 9420                                 column=info:id, timestamp=1537041290060, value=9420
 9420                                 column=info:rep, timestamp=1537041290060, value=1878
10 row(s) in 0.1590 seconds

hbase(main):023:0>
```

Fig 1. Top ten users in descending order based on their row key.

In figure 1, the Id value for the user is presented as the first line and the reputation as the second line. In total there are ten rows and users with the first ten highest reputation values.