MSc Machine Learning

ID2221 Data Intensive Computing

Lab 1: MapReduce, HDFS and HBase

Author:

Alexandros Ferles
ferles@kth.se

George Zervakis
zervakis@kth.se

Professor:
Amir H. Payberah
payberah@kth.se

Contents

1 Top Ten Users **2**
1.1 Problem Description 2
1.2 Problem Solution 2
1.3 Execution 2
1.4 Results 3

1 Top Ten Users

1.1 Problem Description

In this assignment, we were instructed to implement a MapReduce job, that based on data located in the Hadoop file system 'users.xml' file, can compute the id of the top ten users in terms of their reputation levels, along with this corresponding reputation. The results should be stored at a table named 'topten' in HBase.

1.2 Problem Solution

The implementation is based on the Map-Reduce approach. During this setting, the classes **TopTenMapper** and **TopTenReducer** are driven by the main function:

- The **TopTenMapper** is responsible for the preparation and process of the data that belong in its node. During its execution, it isolates the *id* and *reputation* from each datum, and makes sure that its own top ten data are delivered to the **ToptenReducer**.
- The **ToptenReducer** accumulates the work of several Mappers from several nodes, estimates the top ten users from the combined data and writes the results to the HBase table 'topten' as instructed. In order to get only the top ten results, only one Reducer should be used, as several reducers would lead to several top ten users written in HBase.
- The **main** function serves as the Driver class of similar problems: it allocates data to the mappers and synchronizes the jobs executed by the mappers and the reducer. It also assigns the 'topten' table as the destination HBase table for the extracted results.

1.3 Execution

Make sure that you have cleared previous instances of the 'topten' table, if any. Then run the script 'run.sh' (bare in mind that this script also removes any previous folder or jar files, so there is a possibility that you get a harmless warning that there are no folders or jar files to delete) placed in the *src/main/java* folder via the command **sh run.sh** or **./run.sh**.

1.4 Results

You are expected to observe results similar to the following:

```
hbase(main):072:0> scan 'topten'
ROW                                                COLUMN+CELL
 \x00\x00\x00\x00                                  column=info:id, timestamp=1537733578725, value=2452
 \x00\x00\x00\x00                                  column=info:rep, timestamp=1537733578725, value=4503
 \x00\x00\x00\x01                                  column=info:id, timestamp=1537733578725, value=381
 \x00\x00\x00\x01                                  column=info:rep, timestamp=1537733578725, value=3638
 \x00\x00\x00\x02                                  column=info:id, timestamp=1537733578725, value=11097
 \x00\x00\x00\x02                                  column=info:rep, timestamp=1537733578725, value=2824
 \x00\x00\x00\x03                                  column=info:id, timestamp=1537733578725, value=21
 \x00\x00\x00\x03                                  column=info:rep, timestamp=1537733578725, value=2586
 \x00\x00\x00\x04                                  column=info:id, timestamp=1537733578725, value=548
 \x00\x00\x00\x04                                  column=info:rep, timestamp=1537733578725, value=2289
 \x00\x00\x00\x05                                  column=info:id, timestamp=1537733578725, value=84
 \x00\x00\x00\x05                                  column=info:rep, timestamp=1537733578725, value=2179
 \x00\x00\x00\x06                                  column=info:id, timestamp=1537733578725, value=434
 \x00\x00\x00\x06                                  column=info:rep, timestamp=1537733578725, value=2131
 \x00\x00\x00\x07                                  column=info:id, timestamp=1537733578725, value=108
 \x00\x00\x00\x07                                  column=info:rep, timestamp=1537733578725, value=2127
 \x00\x00\x00\x08                                  column=info:id, timestamp=1537733578725, value=9420
 \x00\x00\x00\x08                                  column=info:rep, timestamp=1537733578725, value=1878
 \x00\x00\x00\x09                                  column=info:id, timestamp=1537733578725, value=836
 \x00\x00\x00\x09                                  column=info:rep, timestamp=1537733578725, value=1846
10 row(s) in 0.0140 seconds
```

Figure 1: 'Topten' table after all sources are executed