ID2221 DATA INTENSIVE COMPUTING
LAB ASSIGNMENT 1

Rithika Harish Kumar(rihk@kth.se)
Catherine Rakama(rakama@kth.se)

TASK

A list of users information is given, we have to get the information of the top ten users based on their reputation. Input is from Hdfs and the output should be in a table (Hbase) with two columns for id and reputation.

GIVEN

An input file users.xml and TopTen.java file with base code was given.

STEPS

1. Edit the Mapper, Reducer and driver classes.

- Mapper:

The input to the mapper is the xml file which is read from HDFS topten_input folder and the output is the list of top ten records.
The functionalities of the map method are,
- The transformXmlToMap helper function extracts the information from the file and returns as string.
- The Id and reputation is parsed from the string.
- The reputation and user information is stored in a TreeMap if Id is not null.
- Only the top ten records are stored.

The top ten records are output to the reducer in the cleanup method. The cleanup method gets called once after all key-value pairs have been through the map function.

- Reducer:

The input to reducer from the mapper would be the key with null and the value with the list of top ten user information. The reducer outputs the id and reputation of the top ten users to the hbase table “topten.”
The functionalities of the reduce function are,
- The output from the mapper is iterated to extract the id and reputation by parsing.
- After extracting it is stored in a TreeMap.
- An object is created to input values into the Hbase table named “topten” using Put.
- The value of Reputation is used as the key for the table.
- Using addColumn, two columns to put the id and the reputation is added.
- The result is stored to Hbase in sorted manner based on reputation.(small to big)

- Driver:

- A job instance is created for the Hbase configuration.
- Mapper, Reducer, output key, output value classes are set.
- Number of Reducer tasks is set to one.
- Input file path is added.

Rithika Harish Kumar(rihk@kth.se)
Catherine Rakama(rakama@kth.se)

- Output Hbase table is set.

2. Start the datanode and namenode.(start-dfs.sh)
3. Start the hbase.($HBASE_HOME/bin/start-hbase.sh)
4. Start the hbase shell and create a table ‘topten’.
   $HBASE_HOME/bin/hbase shell
   Create ‘topten’,’info’.
5. Exit from the shell.(exit)
6. Export the classpaths of hadoop and hbase.
   export HADOOP_CLASSPATH=$($HADOOP_HOME/bin/hadoop classpath)
   export HBASE_CLASSPATH=$($HBASE_HOME/bin/hbase classpath)
   export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$HBASE_CLASSPATH
7. Make a directory “ topten_classes”.
   mkdir topten_classes
8. Compile the TopTen.java file and store all the class files inside the directory.
   javac -cp $HADOOP_CLASSPATH -d topten_classes topten/TopTen.java
9. Obtain the jar file.
   jar -cvf topten.jar -C topten_classes/ .
10. Run the java program.
    $HADOOP_HOME/bin/hadoop jar topten.jar topten.TopTen topten_input.
11. Again to view the result start the hbase shell.
12. Scan ‘topten’.

RESULT

```
2018-09-17 15:07:47,097 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

hbase(main):001:0> scan 'topten'
ROW                             COLUMN+CELL
 \x00\x00\x076                  column=info:id, timestamp=1537187464976, value=\x00\x00\x03D
 \x00\x00\x076                  column=info:rep, timestamp=1537187464976, value=\x00\x00\x076
 \x00\x00\x07V                  column=info:id, timestamp=1537187464976, value=\x00\x00$\xCC
 \x00\x00\x07V                  column=info:rep, timestamp=1537187464976, value=\x00\x00\x07V
 \x00\x00\x08O                  column=info:id, timestamp=1537187464976, value=\x00\x00\x00l
 \x00\x00\x08O                  column=info:rep, timestamp=1537187464976, value=\x00\x00\x08O
 \x00\x00\x08S                  column=info:id, timestamp=1537187464976, value=\x00\x00\x01\xB2
 \x00\x00\x08S                  column=info:rep, timestamp=1537187464976, value=\x00\x00\x08S
 \x00\x00\x08\x83               column=info:id, timestamp=1537187464976, value=\x00\x00\x00T
 \x00\x00\x08\x83               column=info:rep, timestamp=1537187464976, value=\x00\x00\x08\x83
 \x00\x00\x08\xF1               column=info:id, timestamp=1537187464976, value=\x00\x00\x02$
 \x00\x00\x08\xF1               column=info:rep, timestamp=1537187464976, value=\x00\x00\x08\xF1
 \x00\x00\x0A\x1A               column=info:id, timestamp=1537187464976, value=\x00\x00\x00\x15
 \x00\x00\x0A\x1A               column=info:rep, timestamp=1537187464976, value=\x00\x00\x0A\x1A
 \x00\x00\x0B\x08               column=info:id, timestamp=1537187464976, value=\x00\x00+Y
 \x00\x00\x0B\x08               column=info:rep, timestamp=1537187464976, value=\x00\x00\x0B\x08
 \x00\x00\x0E6                  column=info:id, timestamp=1537187464976, value=\x00\x00\x01}
 \x00\x00\x0E6                  column=info:rep, timestamp=1537187464976, value=\x00\x00\x0E6
 \x00\x00\x11\x97               column=info:id, timestamp=1537187464976, value=\x00\x00\x09\x94
 \x00\x00\x11\x97               column=info:rep, timestamp=1537187464976, value=\x00\x00\x11\x97
10 row(s) in 3.7680 seconds

hbase(main):002:0> exit
hadoopusr@rithika-VirtualBox:~$ $HBASE_HOME/bin/stop-hbase.sh
```

All the java and class files can be found inside the /src folder.
The TopTen.java file is in the /src/topten path.
A output file (copypasted output) with the result is also present. The result is sorted from small value to big value among the top ten.