Foster Mosden
Big Data - Assignment 6

Question 1: (10 points, 1 page) Briefly discuss the role of the following Hadoop ecosystem technologies:

Yarn: Yet Another Resource Negotiator (YARN) serves as the Hadoop resource management system. The primary draw of YARN is that it allows users to run diverse workloads on the same Hadoop cluster; for instance, running a disk-heavy command at the same time as a memory and CPU heavy command at the same time. In short, it maximizes the available compute and storage abilities by distributing different types of workloads across nodes.

Zookeeper: Zookeeper is a coordination service that provides a hierarchical and reliable configuration management system. It is used for distributed synchronization, configuration maintenance, and naming registry for large-scale distributed systems. Essentially, it helps make sure all of the hardware and software plays nicely together.

Oozie: Oozie is a "workflow engine" which runs outside of a cluster to schedule and trigger the running of various Hadoop jobs (such as Pig, Hive, and Sqoop). Users can schedule jobs to be run at a specific time, or trigger them to occur based on the presence or change in data. As well, users can add additional actions such as sending an email when the jobs are complete, or if there are any errors.

Sqoop: SQL to Hadoop, or Sqoop, imports tables from an RDBMS into HDFS. It uses MapReduce jobs to import the database, and a JDBC interface. Sqoop's purpose is to simplify the integration of Hadoop with existing data infrastructure.

Hue: Hue is the web based interface for Hadoop. Simply, it makes Hadoop easier to use by providing a familiar GUI interface in the browser.

2. Pig commands output:

```
Output(s):
Successfully stored 2 records (70 bytes) in: "hdfs://ip-172-31-39-171.ec2.internal:8020/tmp/temp-1255997575/tmp-1542517162"

2023-11-14 23:19:20,997 INFO input.FileInputFormat: Total input files to process : 2
2215151 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil  - Total input paths to process : 2
2023-11-14 23:19:20,997 INFO util.MapRedUtil: Total input paths to process : 2
(NYSE,CVE,2009-12-09,1.0021276)
(NYSE,CHT,2009-03-03,0.0)
```

3.

Get distinct rate_code_id from the table

```
75
84
100
120
144
240
300
480
576
800
Time taken: 35.013 seconds, Fetched: 76 row(s)
```

Show all rows/columns where rate_code_id = 1

```
1       1/1/17 5:52     1/1/17 5:57     N       1       166     151     1       1.3     6.0     0.5     0.5     0.0     0.0             0.3     7.3     2       1
1       1/1/17 5:52     1/1/17 5:56     N       1       157     36      1       0.7     5.0     0.5     0.5     0.0     0.0             0.3     6.3     2       1
1       1/1/17 5:52     1/1/17 6:38     N       1       228     223     2       21.7    61.0    0.0     0.5     18.5    0.0             0.3     80.3    1       1
1       1/1/17 5:52     1/1/17 6:05     N       1       7       262     1       4.2     14.5    0.0     0.5     2.0     0.0             0.3     17.3    1       1
1       1/1/17 5:52     1/1/17 6:02     N       1       255     17      1       2.5     10.0    0.0     0.5     2.0     0.0             0.3     12.8    1       1
2       1/1/17 5:21     1/1/17 5:36     N       1       106     80      1       6.46    20.0    0.5     0.5     5.32    0.0             0.3     26.62   1       1
2       1/1/17 5:38     1/1/17 5:50     N       1       80      61      1       2.97    11.5    0.5     0.5     0.0     0.0             0.3     12.8    2       1
Time taken: 0.167 seconds, Fetched: 20777 row(s)
```

```
Time taken: 0.167 seconds, Fetched: 20777 row(s)
```

4.

Write a Pig script (logs.pig) to parse the data to have "ipaddress, timestamp, request, status code, and data size" and store it in a hdfs folder.

```
Input(s):
Successfully read 91714 records from: "s3://fm-ga2/access_log_Jul95.txt"

Output(s):
Successfully stored 91714 records (9377354 bytes) in: "hdfs:/user/hadoop/weblog_clean"

30936 [main] INFO  org.apache.pig.Main  - Pig script completed in 30 seconds and 976 milliseconds (30976 ms)
2023-11-15 00:47:15,475 INFO pig.Main: Pig script completed in 30 seconds and 976 milliseconds (30976 ms)
```

Write a HiveQL script to query this stored data to show the number of times (count) each ipaddress received a 404 error.

```
statl1.hacom.nl 1
sunday.isltd.insignia.com       1
svasu.extern.ucsd.edu   1
syseng55.sunnyvale.telebit.com  1
tiber.gsfc.nasa.gov     2
training-macl8.caltech.edu      1
troll.vestnett.no       1
tsl-and-l5.iquest.net   1
unix.ccsnet.com 1
vector.wantree.com.au   1
viking.cris.com 2
vortex.pc.ingr.com      1
winnie.fit.edu  1
www-b2.proxy.aol.com    5
www-b3.proxy.aol.com    5
www-b4.proxy.aol.com    5
www-d3.proxy.aol.com    1
www-d4.proxy.aol.com    4
wwwproxy.ac.il  1
Time taken: 111.225 seconds, Fetched: 236 row(s)
```

Write a shell script (logs.sh) that calls the above two scripts. You can run the shell script as "bash logs.sh" to run both Pig and Hive scripts.

```
www-b2.proxy.aol.com    5
www-b3.proxy.aol.com    5
www-b4.proxy.aol.com    5
www-d3.proxy.aol.com    1
www-d4.proxy.aol.com    4
wwwproxy.ac.il  1
Time taken: 20.798 seconds, Fetched: 236 row
Script execution completed.
```