INFSCI 2750: Cloud Computing
Mini Project 1
Group 4 – Harshitha Venkatesh
Radhika Purohit
Hithaishree Shankar

Part 1: Setting up Hadoop in Docker: (30 points)

We proceeded with constructing the Hadoop cluster by adhering to the prescribed sequence of preparatory actions for Hadoop installation and execution. Subsequently, we initiated the cluster and executed the default wordcount program, an integral component of the Hadoop package.

1. Created a small Ubuntu Docker image as a foundational step for virtualized environment setup.

2. Developed a Hadoop Docker image based on the Ubuntu image, implementing basic Hadoop functionalities for local task execution using provided bash files.

3. Tested the Hadoop Docker image by successfully running a Wordcount job locally, including Dockerfile, support files, and a bootstrap script for service startup in the submission.

```
# bin/hdfs dfs -cat output_pgmpart1/*
2024-02-25 20:43:31,946 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Hadoop  3
Hello   2
Wordcount       1
a       1
distributed     1
hello   1
in      1
is      1
job     1
system  1
world   2
#
```

Part 2: Developing a Hadoop program (N-Gram)

Developed a Hadoop program in Java to compute n-gram frequencies from a given text file. Implemented Mapper and Reducer classes to split the text into n-grams and aggregate the counts.

```
# bin/hdfs dfs -cat /ourtpu/*
2024-02-25 19:21:37,150 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
el      2
he      2
ld      2
ll      2
lo      2
or      2
rl      2
wo      2
#
```

Part 3: Developing a Hadoop program to analyze real logs (50 points)

Developed MapReduce programs in Hadoop to analyze real anonymous logs stored in an access_log file. Focused on answering specific questions based on the log data, disregarding the last two fields of the log entries. Implemented solutions without utilizing information from the ignored fields for the problems presented.

```
# bin/hdfs dfs -ls /output_saib
Found 4 items
-rw-r--r--   1 root supergroup         41 2024-02-26 13:45 /output_saib/part-00000
-rw-r--r--   1 root supergroup         24 2024-02-26 13:45 /output_saib/part-00001
-rw-r--r--   1 root supergroup         26 2024-02-26 13:46 /output_saib/part-00002
-rw-r--r--   1 root supergroup        150 2024-02-26 13:46 /output_saib/part-00003
# bin/hdfs dfs -cat /output_saib/part-00000
2024-02-26 14:24:14,644 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
Q5: 47.39.156.135 with 4305147 accesses
# bin/hdfs dfs -cat /output_saib/part-00001
2024-02-26 14:24:53,146 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
Q1: 330592
Q6: 471552
# bin/hdfs dfs -cat /output_saib/part-00002
2024-02-26 14:25:23,040 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
Q2: 1959647
Q7: 7209678
# bin/hdfs dfs -cat /output_saib/part-00003
2024-02-26 14:25:48,142 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
Q3: 16 methods used: get, VIMF, DEBUG, GET, DELETE, TRACK, INDEX, FLURP, ASDE, TRACE, PROPFIND, OPTIONS, POST, HEAD, PUT
, SEARCH
Q8: 10337151 bytes
#
```

```
# /opt/hadoop/bin/hadoop fs -cat ou/part-00000
2024-02-26 15:00:09,508 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
ostTrusted = false
Q4: /apache-log/access.log with 630006 hits
Q9: 14.139.110.137 (55384128894 bytes), 109.220.172.234 (54013537755 bytes), 51.210.183.78 (36837061299 bytes)
Q10: 2465568761 bytes
#
```

Problems:

1. How many hits were made to the website directory "/images/smilies/" (including subdirectories and files)?
   Sol:330592

2. How many hits were made from the IP: 96.32.128.5?
   Sol:1959647

3. How many HTTP request methods are used in this file? What are they?
   Sol:16 methods used: get, VIMF, DEBUG, GET, DELETE, TRACK, INDEX, FLURP, ASDE, TRACE, PROPFIND, OPTIONS, POST, HEAD, PUT, SEARCH

4. Which path in the website has been hit most? How many hits were made to the path?
   Sol: /apache-log/access.log with 630006 hits

5. Which IP accesses the website most? How many accesses were made by it?

Sol: 47.39.156.135 WITH 4305147 accesses

6. How many POST request were made?
   Sol:471552

7. How many requests received a 404 status code?
   Sol:7209678

8. How much data was requested on 19/Dec/2020?
   Sol:10337151 bytes

9. List 3 IPs that access the most, and what is the total data flow size of each IP?
   Sol:14.139.110.137 (55384128894 bytes), 109.220.172.234 (54013537755 bytes), 51.210.183.78

10. How much data(in bytes) was successfully(with status code 200) requested on 16/Jan/2022?
    Sol:2465568761 bytes