BIG DATA ANALYTICS ASSIGNMENT-01

Hadoop Installation/setup and Designing Additional Queries

Submitted by:

Name: Khuram Shahzad

Roll No: p218742

Subject: Big Data Analytics (MS DS)

Submitted to:

Dr. Omar Usman Khan

JUNE 27, 2022

FAST - NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES – PESHAWAR CAMPUS
Jamrud Road 160 Industrial Estate Road, Peshawar - PK

Table of Contents

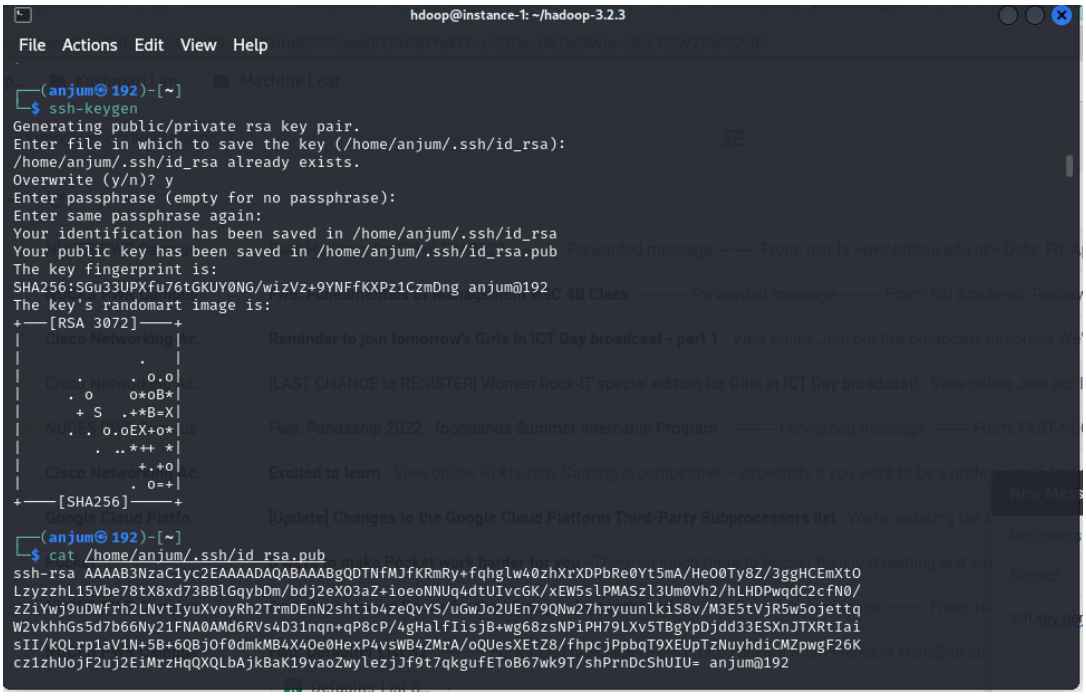
Prepared by: Engr. Khuram Shahzad (p21-8742)

Prepared by: Engr. Khuram Shahzad (p21-8742)

Prepared by: Engr. Khuram Shahzad (p21-8742)

0 TASK -0: HADOOP STEP UP AND INSTALLATION IN GOOGLE CLOUD:

0.1 SSH KEY GENERATING.



0.2 LOGIN WITH INSTANCE

Prepared by: Engr. Khuram Shahzad (p21-8742)

0.3 UPDATE

```
anjum@instance-1:~$ sudo apt update
Hit:1 http://us-west4.gce.archive.ubuntu.com/ubuntu focal InRelease
Get:2 http://us-west4.gce.archive.ubuntu.com/ubuntu focal-updates InRelease [114 kB]
Get:3 http://security.ubuntu.com/ubuntu focal-security InRelease [114 kB]
Get:4 http://us-west4.gce.archive.ubuntu.com/ubuntu focal-backports InRelease [108 kB]
Get:5 http://us-west4.gce.archive.ubuntu.com/ubuntu focal/universe amd64 Packages [8628 kB]
Get:6 http://us-west4.gce.archive.ubuntu.com/ubuntu focal/universe Translation-en [5124 kB]
Get:7 http://us-west4.gce.archive.ubuntu.com/ubuntu focal/universe amd64 c-n-f Metadata [265 kB
]
```

0.4 INSTALL OPEN JDK

```
anjum@instance-1:~$ sudo apt install openjdk-8-jdk
Reading package lists ... Done
Building dependency tree
Reading state information ... Done
The following packages were automatically installed and are no longer required:
  libatasmart4 libblockdev-fs2 libblockdev-loop2 libblockdev-part-err2 libblockdev-part2
  libblockdev-swap2 libblockdev-utils2 libblockdev2 libmm-glib0 libnuma1 libparted-fs-resize0
  libudisks2-0 usb-modeswitch usb-modeswitch-data
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  adwaita-icon-theme at-spi2-core ca-certificates-java fontconfig fontconfig-config
```

0.5 CREATING NEW USER FOR HADOOP

```
anjum@instance-1:~$ sudo adduser hdoop
Adding user `hdoop' ...
Adding new group `hdoop' (1003) ...
Adding new user `hdoop' (1002) with group `hdoop' ...
Creating home directory `/home/hdoop' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hdoop
Enter the new value, or press ENTER for the default
        Full Name []: BigData
        Room Number []:
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n]
anjum@instance-1:~$ su - hdoop
Password:
hdoop@instance-1:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
-bash: /home/hdoop/.ssh/authorized_keys: No such file or directory
hdoop@instance-1:~$ shh-keygen

Command 'shh-keygen' not found, did you mean:

  command 'ssh-keygen' from deb openssh-client (1:8.2p1-4ubuntu0.4)

Try: apt install <deb name>

hdoop@instance-1:~$ ssh-keygen
```


Prepared by: Engr. Khuram Shahzad (p21-8742)

0.6 DOWNLOAD HADOOP

```
hdoop@instance-1:~$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hdoop/.ssh/id_rsa):
Created directory '/home/hdoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hdoop/.ssh/id_rsa
Your public key has been saved in /home/hdoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:27tBzDRT5B0NRH7JoVivBdhHTZ+qIeEJwDJmAQladNc hdoop@instance-1
The key's randomart image is:
+---[RSA 3072]----+
|oo=.+...      .=**o|
|.o * o. E    +o+=oB|
|. o o   . .+ ...+*o|
|        o+oo  +. |
|        S++. o   |
|          +. o   |
|         . o.    |
|            o    |
|           o.    |
+----[SHA256]-----+
hdoop@instance-1:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hdoop@instance-1:~$ chmod 0600 ~/.ssh/authorized_keys
hdoop@instance-1:~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.2.3/hadoop-3.2.3.tar.g
z
--2022-04-29 06:47:37--  https://dlcdn.apache.org/hadoop/common/hadoop-3.2.3/hadoop-3.2.3.tar.g
z
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 492241961 (469M) [application/x-gzip]
Saving to: 'hadoop-3.2.3.tar.gz'
```

0.7 SETTING CONFIGURATION

```
hdoop@instance-1:~$ tar xzf hadoop-3.2.3.tar.gz
hdoop@instance-1:~$ nano ~/.bashrc
hdoop@instance-1:~$ readlink -f /usr/bin/javac
/usr/lib/jvm/java-8-openjdk-amd64/bin/javac
hdoop@instance-1:~$ nano ~/hadoop-3.2.3/etc/hadoop/hadoop-env.sh
hdoop@instance-1:~$ nano ~/hadoop-3.2.3/etc/hadoop/core-site.xml
hdoop@instance-1:~$ nano ~/hadoop-3.2.3/etc/hadoop/hdfs-site.xml
hdoop@instance-1:~$ nano ~/hadoop-3.2.3/etc/hadoop/mapred-site.xml
hdoop@instance-1:~$ nano ~/hadoop-3.2.3/etc/hadoop/yarn-site.xml
hdoop@instance-1:~$ hdfs namenode -format
```


Prepared by: Engr. Khuram Shahzad (p21-8742)

0.8 SETTING HDFS CONFIGURATION

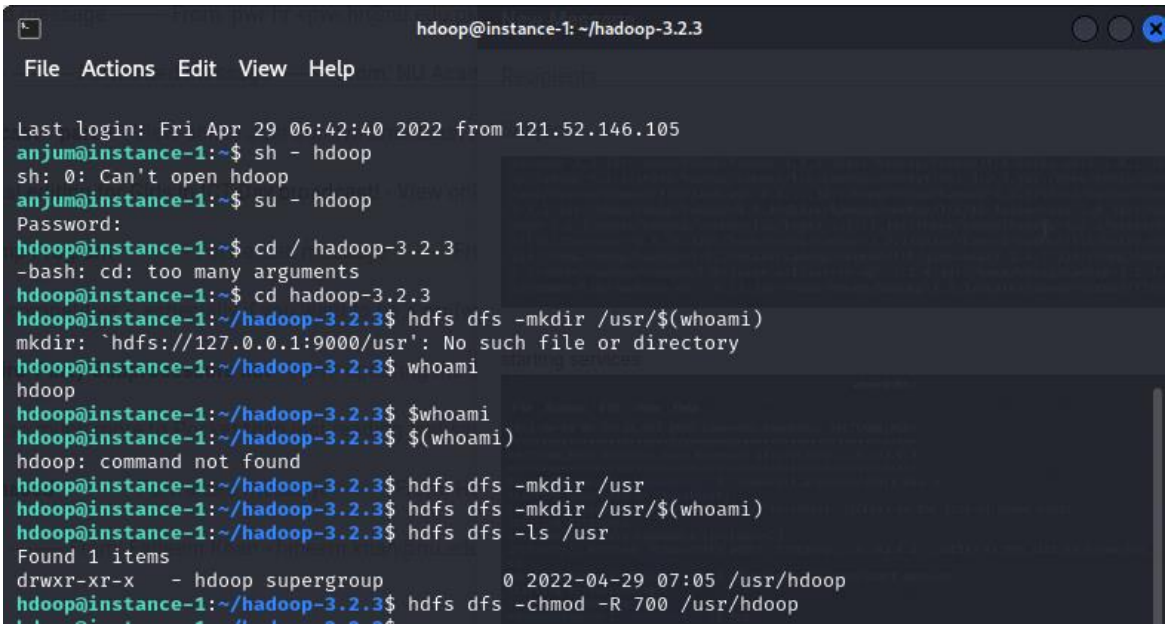
```
hdoop@instance-1:~$ cd hadoop-3.2.3/
hdoop@instance-1:~/hadoop-3.2.3$ hdfs namenode -format
WARNING: /home/hdoop/hadoop-3.2.3/logs does not exist. Creating.
2022-04-29 06:56:27,199 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = instance-1/10.182.0.2
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.2.3
STARTUP_MSG:   classpath = /home/hdoop/hadoop-3.2.3/etc/hadoop:/home/hdoop/hadoop-3.2.3/share/h
adoop/common/lib/commons-collections-3.2.2.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib
/kerb-client-1.0.1.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/animal-sniffer-annotati
ons-1.17.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/curator-client-2.13.0.jar:/home/h
doop/hadoop-3.2.3/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/home/hdoop/hadoop-3.2.3/
share/hadoop/common/lib/commons-configuration2-2.1.1.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/
common/lib/jetty-io-9.4.40.v20210413.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/jerse
y-servlet-1.19.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/commons-compress-1.21.jar:/
home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/error_prone_annotations-2.2.0.jar:/home/hdoop/h
adoop-3.2.3/share/hadoop/common/lib/listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.
jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/netty-3.10.6.Final.jar:/home/hdoop/hadoop-
3.2.3/share/hadoop/common/lib/jetty-webapp-9.4.40.v20210413.jar:/home/hdoop/hadoop-3.2.3/share/
hadoop/common/lib/token-provider-1.0.1.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/com
mons-io-2.8.0.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/nimbus-jose-jwt-9.8.1.jar:/h
ome/hdoop/hadoop-3.2.3/share/hadoop/common/lib/kerb-util-1.0.1.jar:/home/hdoop/hadoop-3.2.3/sha
re/hadoop/common/lib/commons-codec-1.11.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/pa
ranamer-2.3.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/kerby-pkix-1.0.1.jar:/home/hdo
op/hadoop-3.2.3/share/hadoop/common/lib/hadoop-annotations-3.2.3.jar:/home/hdoop/hadoop-3.2.3/s
hare/hadoop/common/lib/jaxb-api-2.2.11.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/re2
j-1.1.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/failureaccess-1.0.jar:/home/hdoop/ha
doop-3.2.3/share/hadoop/common/lib/log4j-1.2.17.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/commo
n/lib/zookeeper-3.4.14.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/kerby-config-1.0.1.
jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/json-smart-2.4.7.jar:/home/hdoop/hadoop-3.
2.3/share/hadoop/common/lib/javax.activation-api-1.2.0.jar:/home/hdoop/hadoop-3.2.3/share/hadoo
p/common/lib/jackson-xc-1.9.13.jar:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/jackson-ann
```

0.9 STARTING SERVICES

```
anjum@192: ~
File  Actions  Edit  View  Help
2022-04-29 06:56:28,953 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at instance-1/10.182.0.2
************************************************************/
hdoop@instance-1:~/hadoop-3.2.3$ ~/hadoop-3.2.3/sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Starting datanodes
Starting secondary namenodes [instance-1]
instance-1: Warning: Permanently added 'instance-1,10.182.0.2' (ECDSA) to the list of known hos
ts.
hdoop@instance-1:~/hadoop-3.2.3$ ~/hadoop-3.2.3/sbin/start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@instance-1:~/hadoop-3.2.3$ jps
6944 Jps
5905 NameNode
6293 SecondaryNameNode
6055 DataNode
6633 NodeManager
6474 ResourceManager
hdoop@instance-1:~/hadoop-3.2.3$ ls -lh
total 216K
-rw-rw-r-- 1 hdoop hdoop 148K Mar 10 05:39 LICENSE.txt
-rw-rw-r-- 1 hdoop hdoop  22K Mar 10 05:39 NOTICE.txt
-rw-rw-r-- 1 hdoop hdoop 1.4K Mar 10 05:39 README.txt
drwxr-xr-x 2 hdoop hdoop 4.0K Mar 20 01:58 bin
drwxrwxr-x 3 hdoop hdoop 4.0K Apr 29 06:56 dfsdata
drwxr-xr-x 3 hdoop hdoop 4.0K Mar 20 01:20 etc
drwxr-xr-x 2 hdoop hdoop 4.0K Mar 20 01:58 include
drwxr-xr-x 3 hdoop hdoop 4.0K Mar 20 01:58 lib
drwxr-xr-x 4 hdoop hdoop 4.0K Mar 20 01:58 libexec
drwxrwxr-x 3 hdoop hdoop 4.0K Apr 29 06:57 logs
drwxr-xr-x 3 hdoop hdoop 4.0K Mar 20 01:20 sbin
```


Prepared by: Engr. Khuram Shahzad (p21-8742)

1 TASK -01: CREATING YOUR DIRECTORY SPACE



2 TASK -02: UNDERSTANDING THE SYSTEM

It's to be noted that we have defined a function for each step and we have made the generalized code.

2.1 HOW MANY DATA NODES ARE PART OF THE HADOOP TOPOLOGY?

- Number of data-nodes: 1
- Live data nodes (1):

2.2 WHAT ARE THE IP ADDRESSES OF THESE DATANODES?

- Number of data-nodes: 1
- Name: 127.0.0.1:9866 (localhost)

2.3 WHAT IS THE CONFIGURED AND PRESENT CAPACITY OF THE HDFS?

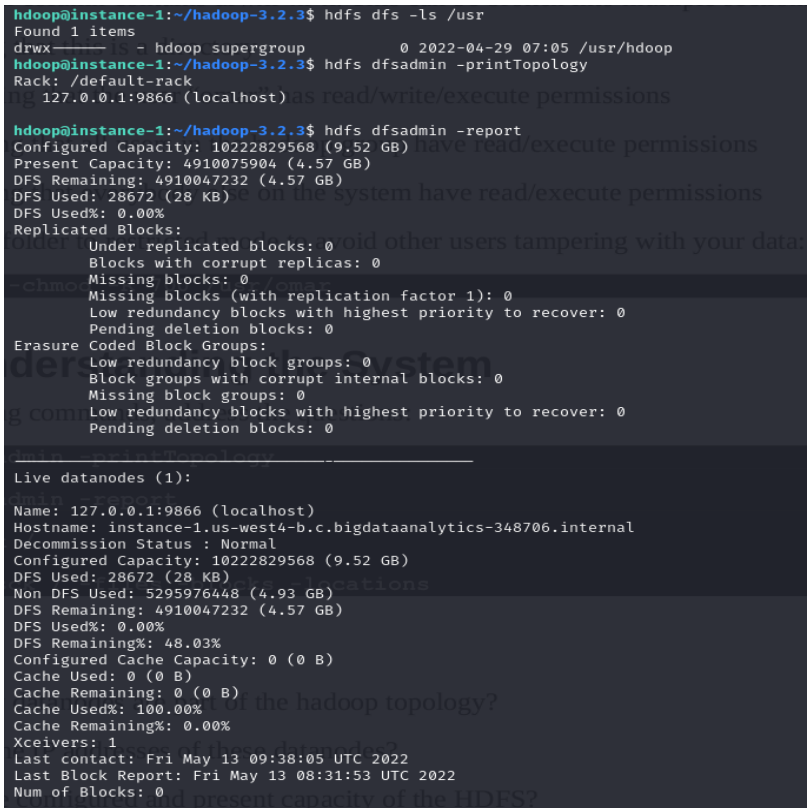
- Configured Capacity: 10222829568 (9.52 GB)
- Present Capacity: 4910075904 (4.57 GB)

2.4 WHAT IS THE DEFAULT FILE REPLICATION COUNT?

- Default replication factor: 1

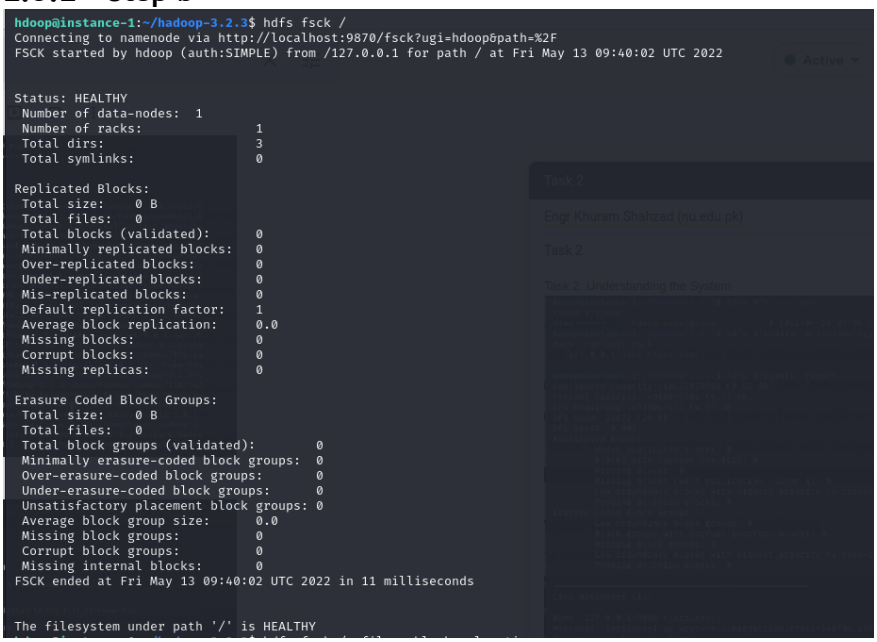
2.5 TASK 02 STEPS:

2.5.1 Step a

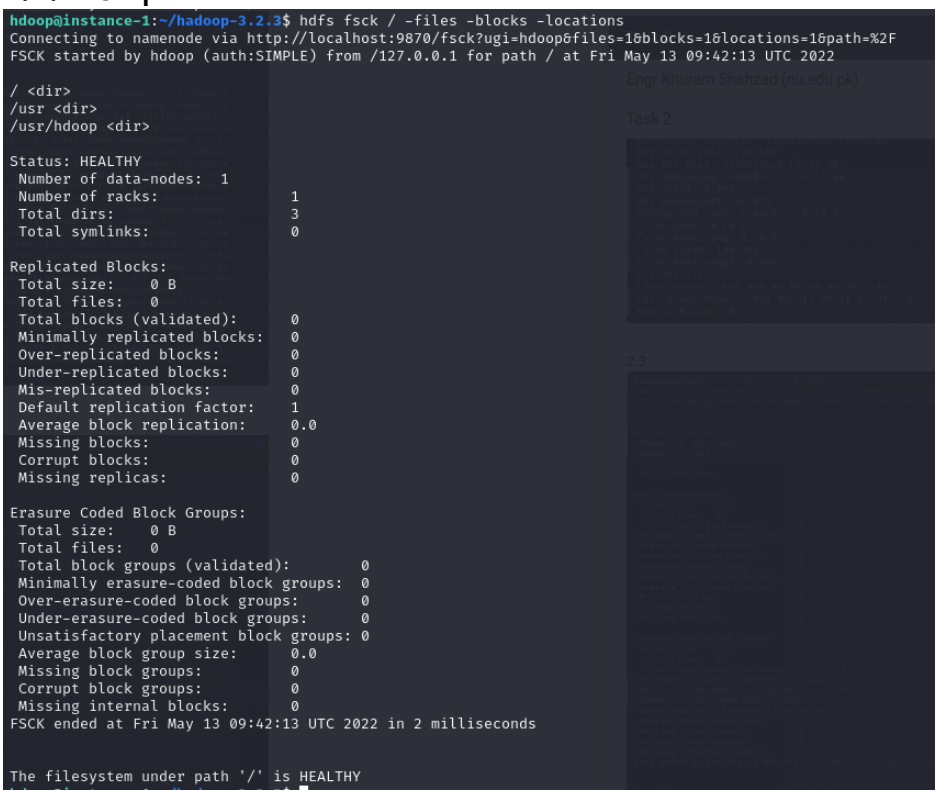


Prepared by: Engr. Khuram Shahzad (p21-8742)

2.5.2 Step b



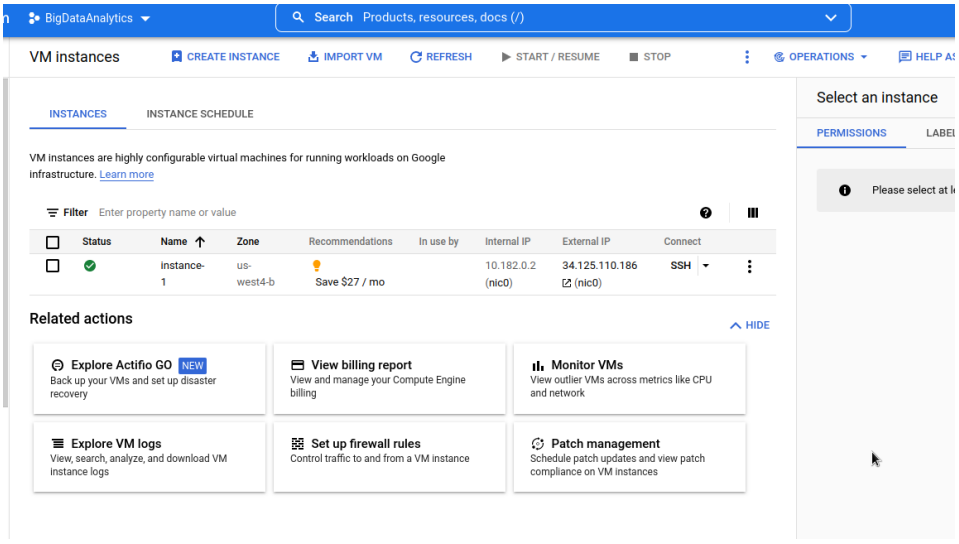
2.5.3 Step c



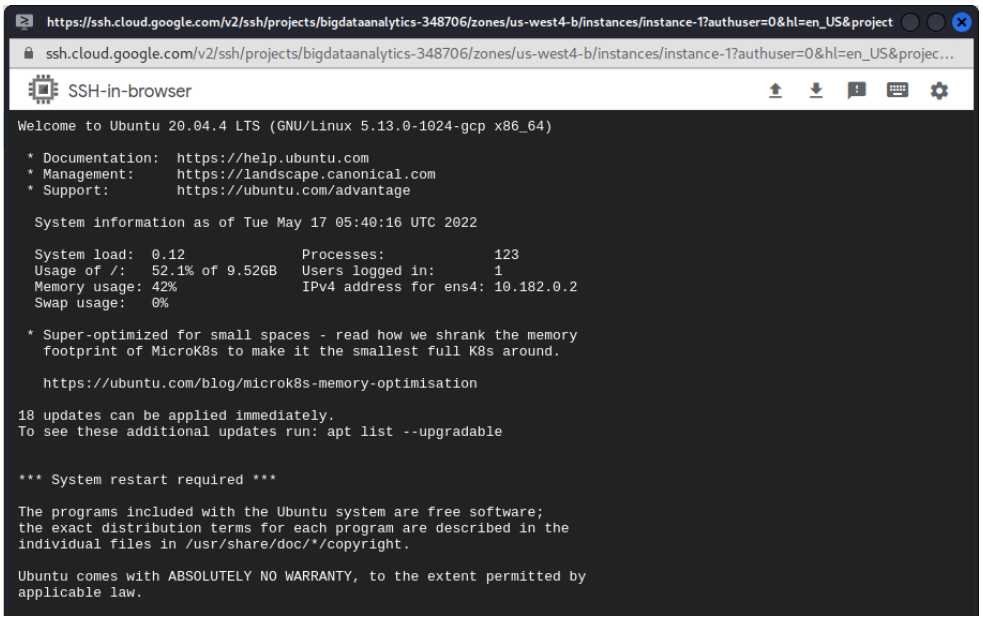
3 TASK -03: GETTING SAMPLE DATA

3.1 HOW TO UPLOAD FILE IN GOOGLE INSTANCES

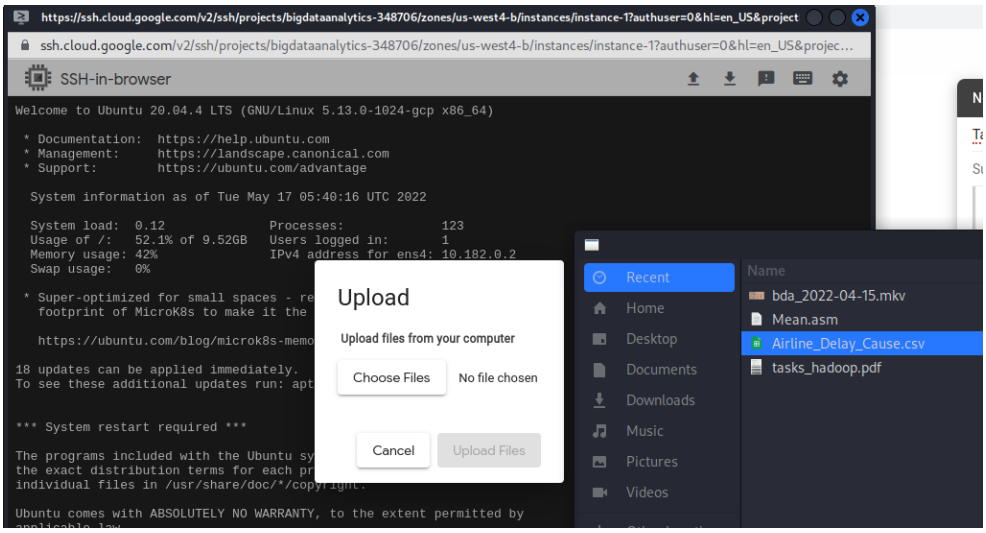
3.1.1 Step1: In instances Click on SSH



3.1.2 This screen will be shown



3.1.3 Click on upload icon



3.1.4 We can see that file uploaded successfully

```
itskhuramshahzad@instance-1:~$ ls /
bin   dev  home  lib32  libx32      media  opt   root  sbin  srv  tmp  var
boot  etc  lib   lib64  lost+found  mnt    proc  run   snap  sys  usr
itskhuramshahzad@instance-1:~$
itskhuramshahzad@instance-1:~$ ls /home/
anjum  hdoop  itskhuramshahzad  ubuntu
itskhuramshahzad@instance-1:~$ ls /home/anjum
itskhuramshahzad@instance-1:~$ ls /home/anjum/
itskhuramshahzad@instance-1:~$ ls /itskhuramshahzad/
ls: cannot access '/itskhuramshahzad/': No such file or directory
itskhuramshahzad@instance-1:~$ ls /home/itskhuramshahzad/
Airline_Delay_Cause.csv
itskhuramshahzad@instance-1:~$ cls
```

Now came to local system in Hadoop

3.1.5 Upload file from Instances to Hadoop user directory.

```
hdoop@instance-1:~$ hdfs dfs -put  /home/itskhuramshahzad/Airline_Delay_Cause.csv  /usr/hdoop/Airline_Delay_Cause.csv
hdoop@instance-1:~$ hdfs dfs -ls /usr/hdoop
Found 1 items
-rw-r--r--   1 hdoop supergroup     277233 2022-05-17 05:47 /usr/hdoop/Airline_Delay_Cause.csv
```

3.1.6 Now run hdfs fsck command to see detail about file

```
hdoop@instance-1:~$ hdfs fsck /usr/hdoop/Airline_Delay_Cause.csv
Connecting to namenode via http://localhost:9870/fsck?ugi=hdoop&path=%2Fusr%2Fhdoop%2FAirline_Delay_Cause.csv
FSCK started by hdoop (auth:SIMPLE) from /127.0.0.1 for path /usr/hdoop/Airline_Delay_Cause.csv at Tue May 17 06:08:21 UTC 20

Status: HEALTHY
 Number of data-nodes:  1
 Number of racks:               1
 Total dirs:                    0
 Total symlinks:                0

Replicated Blocks:
 Total size:    277233 B
 Total files:   1
 Total blocks (validated):      1 (avg. block size 277233 B)
 Minimally replicated blocks:   1 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Missing blocks:                0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)

Erasure Coded Block Groups:
 Total size:    0 B
 Total files:   0
 Total block groups (validated):        0
 Minimally erasure-coded block groups:  0
 Over-erasure-coded block groups:       0
 Under-erasure-coded block groups:      0
 Unsatisfactory placement block groups: 0
 Average block group size:      0.0
 Missing block groups:          0
 Corrupt block groups:          0
 Missing internal blocks:       0
FSCK ended at Tue May 17 06:08:22 UTC 2022 in 13 milliseconds


The filesystem under path '/usr/hdoop/Airline_Delay_Cause.csv' is HEALTHY
```

Questions

3.1.7 What is the default block size (in Mb) of the airline_data.csv file?

- 277233 Bytes.

3.1.8 Is there any missing replicas for the file airline_data.csv?

- NO

3.1.9 What command will you use to change this block size to 6 Mb (remember to convert into bytes)

- To change for single file.

```
hdoop@instance-1:~$ hdfs dfs  -D dfs.blocksize=6291456 -put /home/itskhuramshahzad/Airline_Delay_Cause.csv  /usr/hdoop/Airline_Delay_Cause3.csv
hdoop@instance-1:~$ hadoop fs -stat %o   /usr/hdoop/Airline_Delay_Cause3.csv
6291456
hdoop@instance-1:~$ hdfs getconf -confKey dfs.blocksize
6291456
```

- hdfs dfs -D dfs.bloksize=6291456 -put /home/itskhuramshahzad/airline_data.csv /usr/hdoop/airline_data.csv

- To change in configuration
  vim ~/hadoop-3.2.3/etc/hadoop/hdfs-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>dfs.data.dir</name>
<value>/home/hdoop/hadoop-3.2.3/dfsdata/namenode</value>
</property>
<property>
<name>dfs.data.dir</name>
<value>/home/hdoop/hadoop-3.2.3/dfsdata/datanode</value>
</property>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>


<property>
<name>dfs.block.size</name>
<value> 750000  </value>
<description> Updated Block Size </decription>
</property>
</configuration>
```

```
<property>
<name>dfs.block.size</name>
<value> 750000  </value>
</property>
```

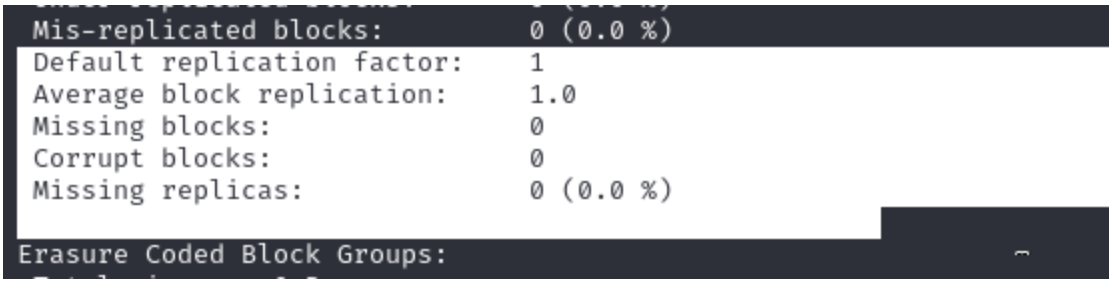
3.1.10 How many blocks are used by airline_data.csv after changing block size in Question 2?

```
hdoop@instance-1:~$ hdfs getconf -confKey dfs.blocksize
134217728
hdoop@instance-1:~$ hadoop fs -stat %o /usr/hdoop/Airline_Delay_Cause2.csv
6291456
hdoop@instance-1:~$
```

5 blocks are used.

3.1.11 How many missing replicas are there for file airline_data.csv after block change?

Zero.



3.1.12

Why are there missing replicas?

Missing blocks can happen when all replicas of a block in the file are corrupted or all replicas go missing

4 TASK -04: SETTING UP FIRST MAP REDUCE JOB

4.1 MAPPER CODE:

reducer.py

```
#!/usr/bin/python
import sys
for line in sys.stdin:
        data = line.strip().split(",");
        key= data[0]
        value= 1
        print("{0}\t{1}".format(key, value))
```

4.2 REDUCER CODE:

```
#!/usr/bin/python
import sys
total =0
oldkey= None

for line in sys.stdin:
        data = line.strip().split("\t")
        thiskey= data[0]
        value = data[1]
        if thiskey≠ oldkey and oldkey≠ None:
                print("{0}\t{1}".format(oldkey, total))
                oldkey= thiskey
                total=0
        oldkey= thiskey
        total+= float(value)
if oldkey ≠ None:
        print("{0}\t{1}".format(oldkey, total))
```

4.3 WRITE AND RUN LOCALLY

```
File  Actions  Edit  View  Help
┌──(anjum㉿192)-[~]
└─$ cat airline_data.csv | ./mapper.py | sort | ./reducer.py
2010    17575.0
2011    15585.0
2012    14387.0
2013    16089.0
2014    13980.0
2015    13528.0
2016    12217.0
2017    12518.0
2018    20231.0
2019    20946.0
year    1.0

┌──(anjum㉿192)-[~]
```

4.4 UPLOADED FILES VM INSTANCES AND ITS USER.

```
hdoop@instance-1:~$ hdfs dfs -put /home/itskhuramshahzad/mapper.py /usr/hdoop/mapper.py
hdoop@instance-1:~$ hdfs dfs -put /home/itskhuramshahzad/reducer.py /usr/hdoop/reducer.py
hdoop@instance-1:~$
```

4.5 TO RUN JOB IN HADOOP:

```
hdoop@instance-1:~$ hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.2.3.jar -file /home/itskhuramshahzad/mapperX.py -file /home
/itskhuramshahzad/reducerX.py -mapper mapperX.py -reducer ./reducerX.py -input /usr/hdoop/airline_data.csv  -output /usr/hdoop/_output14
2022-05-27 16:07:02,735 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/itskhuramshahzad/mapperX.py, /home/itskhuramshahzad/reducerX.py, /tmp/hadoop-unjar4584759314613397 51/] [] /tmp/streamjob640
6172208683873126.jar tmpDir=null
2022-05-27 16:07:03,970 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:5353
2022-05-27 16:07:04,222 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:5353
2022-05-27 16:07:04,561 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_165366
6094282_0005
2022-05-27 16:07:04,913 INFO mapred.FileInputFormat: Total input files to process : 1
2022-05-27 16:07:04,995 INFO mapreduce.JobSubmitter: number of splits:5
2022-05-27 16:07:05,199 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1653666094282_0005
2022-05-27 16:07:05,201 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-05-27 16:07:05,441 INFO conf.Configuration: resource-types.xml not found
2022-05-27 16:07:05,442 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-05-27 16:07:05,532 INFO impl.YarnClientImpl: Submitted application application_1653666094282_0005
2022-05-27 16:07:05,570 INFO mapreduce.Job: The url to track the job: http://instance-1:5349/proxy/application_1653666094282_0005/
2022-05-27 16:07:05,571 INFO mapreduce.Job: Running job: job_1653666094282_0005
2022-05-27 16:07:12,778 INFO mapreduce.Job: Job job_1653666094282_0005 running in uber mode : false
2022-05-27 16:07:12,779 INFO mapreduce.Job:  map 0% reduce 0%
2022-05-27 16:07:31,028 INFO mapreduce.Job:  map 20% reduce 0%
2022-05-27 16:07:32,081 INFO mapreduce.Job:  map 60% reduce 0%
2022-05-27 16:07:33,088 INFO mapreduce.Job:  map 100% reduce 0%
2022-05-27 16:07:38,131 INFO mapreduce.Job:  map 100% reduce 100%
2022-05-27 16:07:39,146 INFO mapreduce.Job: Job job_1653666094282_0005 completed successfully
2022-05-27 16:07:39,265 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=1413519
                FILE: Number of bytes written=4266269
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=26661848
                HDFS: Number of bytes written=139
                HDFS: Number of read operations=20
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
```

```
                Launched map tasks=5
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=86216
                Total time spent by all reduces in occupied slots (ms)=4100
                Total time spent by all map tasks (ms)=86216
                Total time spent by all reduce tasks (ms)=4100
                Total vcore-milliseconds taken by all map tasks=86216
                Total vcore-milliseconds taken by all reduce tasks=4100
                Total megabyte-milliseconds taken by all map tasks=88285184
                Total megabyte-milliseconds taken by all reduce tasks=4198400
        Map-Reduce Framework
                Map input records=157057
                Map output records=157057
                Map output bytes=1099399
                Map output materialized bytes=1413543
                Input split bytes=500
                Combine input records=0
                Combine output records=0
                Reduce input groups=11
                Reduce shuffle bytes=1413543
                Reduce input records=157057
                Reduce output records=11
                Spilled Records=314114
                Shuffled Maps =5
                Failed Shuffles=0
                Merged Map outputs=5
                GC time elapsed (ms)=1468
                CPU time spent (ms)=8450
                Physical memory (bytes) snapshot=1554722816
                Virtual memory (bytes) snapshot=15194402816
                Total committed heap usage (bytes)=1188036608
                Peak Map Physical memory (bytes)=277102592
                Peak Map Virtual memory (bytes)=2534268928
                Peak Reduce Physical memory (bytes)=179793920
                Peak Reduce Virtual memory (bytes)=2535227392
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=26661348
        File Output Format Counters
                Bytes Written=139
22-05-27 16:07:39,271 INFO streaming.StreamJob: Output directory: /usr/hdoop/_output14
```

- hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.2.3.jar -file /home/itskhuramshahzad/mapperX.py -file /home/itskhuramshahzad/reducerX.py -mapper mapperX.py -reducer ./reducerX.py -input /usr/hdoop/airline_data.csv -output /usr/hdoop/_output14
  or
- hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.2.3.jar -files /home/itskhuramshahzad/mapperX.py -file /home/itskhuramshahzad/reducerX.py -mapper mapperX.py -reducer ./reducerX.py -input /usr/hdoop/airline_data.csv -output /usr/hdoop/_output14

Questions

4.6 WHAT WAS THE <KEY,VALUE> PAIR USED IN THIS QUERY?

Key was the Year

And value?

4.7 HOW MANY MAPPER THREADS WERE USED?

- Launched map tasks=5

4.8 HOW MANY MAPPER THREADS WERE USED?

- Launched reduce tasks=1

4.9 WHAT WAS THE TIME SPENT BY ALL MAPPER THREADS?

- Total time spent by all maps in occupied slots (ms)=86216
- Total time spent by all map tasks (ms)=86216
- Total vcore-milliseconds taken by all map tasks=86216
- Total megabyte-milliseconds taken by all map tasks=88285184

4.10 WHAT WAS THE TIME SPENT BY ALL REDUCER THREADS?

- Total time spent by all reduces in occupied slots (ms)=4100
- Total time spent by all reduce tasks (ms)=4100
- Total vcore-milliseconds taken by all reduce tasks=4100
- Total megabyte-milliseconds taken by all reduce tasks=4198400

4.11 WHAT IS THE FILE NAME IN WHICH YOUR OUTPUT IS LOCATED?

- Output directory: /usr/hdoop/_output14

4.12 VARIATION 1:

# of Reducer Tasks	airline_data.csv block size variation (Mb)				
	2 MB	4 MB	8 MB	16 MB	Deafult
2	514893	476332	59484	16296	18438
4	249571	132127	60183	17220	18104
8	250578	139141	57556	16322	18742
16	261119	136873	57444	16482	18706

- hdfs dfs -D mapred.reduce.tasks=8 -put /home/itskhuramshahzad/airline_data.csv /usr/hdoop/airline_dataD8.csv
- hdfs dfs -D dfs.blocksize=8388608 -D mapred.reduce.tasks=2 -put /home/itskhuramshahzad/airline_data.csv /usr/hdoop/airline_data24.csv

4.13 VARIATION 2:

# of Mapper Tasks	airline_data.csv block size variation (Mb)				
	2 MB	4 MB	8 MB	16 MB	Deafult
	2097152b	4194304b	8388608b	16777216	
2	251854s	136445	57065	16065s	17839
4	249292s	129286	57049	16484s	17700
8	259378s	136540	56928	16171s	18515
16	250639s	138447	59177	16046s	18238

4.14 VARIATION 3:

# of Maper Tasks	# of Reducer Tasks			
	2	4	8	16
2	17218s	16662s	16820s	15892s
4	16049s	15582s	17067s	17236s
8	16371s	16160s	16018s	16728s
16	16097s	16579s	17044s	15997s

- hdfs dfs -D dfs.blocksize=16777216 -D mapred.map.tasks=2 -D mapred.reduce.tasks=2 -put /home/itskhuramshahzad/airline_data.csv /usr/hdoop/airline_data1622.csv

5 TASK -05: DESIGNING ADDITIONAL QUERIES

- **Note: we have used best result input file for: no of mapper , no of reducer and block size.**
  - **Command upload:**

hdfs dfs -D dfs.blocksize=16777216 -D mapred.map.tasks=4 -D mapred.reduce.tasks=2 -put /home/itskhuramshahzad/airline_data.csv /usr/hdoop/airline_data_best_results.csv

```
hdoop@instance-1:~$ hdfs dfs -ls /usr/hdoop/years_counts_output
Found 2 items
-rw-r--r--   1 hdoop supergroup          0 2022-06-11 06:23 /usr/hdoop/years_counts_output/_SUCCESS
-rw-r--r--   1 hdoop supergroup        139 2022-06-11 06:23 /usr/hdoop/years_counts_output/part-00000
hdoop@instance-1:~$ hdfs dfs -cat /usr/hdoop/years_counts_output/part-00000
2010    17575.0
2011    15585.0
2012    14387.0
2013    16089.0
2014    13980.0
2015    13528.0
2016    12217.0
2017    12518.0
2018    20231.0
2019    20946.0
year    1.0
hdoop@instance-1:~$
```

so we have used this file for all 5 tasks.

Prepared by: Engr. Khuram Shahzad (p21-8742)

5.1 TASK 5A: PRESENT THE TOTAL FLIGHTS PER YEAR AS A PERCENTAGE

<key, value>=	< key_year, value_flights >
Output Folder:	/usr/hdoop/Task_5a/output3/
Sample Output	

```
2022-06-21 04:23:48,831 INFO streaming.StreamJob: Output directory: /usr/hdoop/Task_5a/output3
hdoop@instance-1:~$ hdfs dfs -cat /usr/hdoop/Task_5a/output3/part-00000
Year No: 2010    Total flight of year : 6450117.0        Total flight Percentage per Year:  10 %
Year No: 2011    Total flight of year : 6085281.0        Total flight Percentage per Year:  9 %
Year No: 2012    Total flight of year : 6096762.0        Total flight Percentage per Year:  9 %
Year No: 2013    Total flight of year : 6369482.0        Total flight Percentage per Year:  10 %
Year No: 2014    Total flight of year : 5819811.0        Total flight Percentage per Year:  9 %
Year No: 2015    Total flight of year : 5819079.0        Total flight Percentage per Year:  9 %
Year No: 2016    Total flight of year : 5617658.0        Total flight Percentage per Year:  8 %
Year No: 2017    Total flight of year : 5674621.0        Total flight Percentage per Year:  9 %
Year No: 2018    Total flight of year : 7213446.0        Total flight Percentage per Year:  11 %
Year No: 2019    Total flight of year : 7422037.0        Total flight Percentage per Year:  11 %
hdoop@instance-1:~$
```

5.1.1 Command used for this task

- **hdoop@instance-1:~$ hdfs dfs -put /home/itskhuramshahzad/5a_mapperX.py /usr/hdoop/Task_5a/5a_mapperX.py**
- **hdoop@instance-1:~$ hdfs dfs -put /home/itskhuramshahzad/5a_reducerX.py /usr/hdoop/Task_5a/5a_reducerX.py**
- **hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.2.3.jar -file /home/itskhuramshahzad/5a_mapperX.py -file /home/itskhuramshahzad/5a_reducerX.py -mapper 5a_mapperX.py -reducer ./5a_reducerX.py -input /usr/hdoop/airline_data_best_results.csv -output /usr/hdoop/Task_5a/output**

5.1.2 Mapper Code:

```python
#!/usr/bin/python3
import sys
import csv
for line in sys.stdin:
        data = line.strip().split(",");
        key_year= data[0]
        if data[5].strip().isdigit():
                value_flights= float (data[8].strip())
        else:
                value_flights=0
        print("{0}\t{1}".format(key_year,value_flights))

```

5.1.3 Reducer Code:

5a_mapperX.py

```
#!/usr/bin/python3
import sys
import csv
total_year =0
total_flights =0
oldkey= None
f = open("data.txt", "w")
for line in sys.stdin:
        data = line.strip().split("\t")
        thiskey= data[0]
        value_flights =  data[1]
        if thiskey≠ oldkey and oldkey≠ None:
                with open('data.txt', 'a') as file:
                        file.write(str(oldkey))
                        file.write("\t")
                        file.write(str(total_flights))
                        file.write("\n")
                oldkey= thiskey
                total_flights=0
        oldkey= thiskey
        total_flights+= float(value_flights)
totalflights=0
with open('data.txt', 'r') as file:
        csv_reader = csv.reader(file, delimiter='\t')
        for data in csv_reader:
                totalflights+= float (data[1])
with open('data.txt', 'r') as file:
        csv_reader = csv.reader(file, delimiter='\t')
        for data in csv_reader:
                key_year= data[0]
                flightcount= float (data[1])
                percentage= int ((flightcount/totalflights)*100)
                print("Year No: {0}\t Total flight of year : {1}\t Total flight Percentage per Year:  {2} %"
                .format(key_year, flightcount, percentage))

```

5.2 TASK 5B: PRESENT THE TOTAL FLIGHTS PER YEAR AS A PERCENTAGE

<key, value>=	<month, value_flights >
Output Folder:	/usr/hdoop/Task_5b/output/
Sample Output	

```
hdoop@instance-1:~$ hdfs dfs -cat /usr/hdoop/Task_5b/output2/part-00000

The busiest Month is : 7        Total flight of this month : 5592214
hdoop@instance-1:~$
```

5.2.1 Mapper Code:

5b_mapperX.py

```
#!/usr/bin/python3
import sys
f = open("data.txt", "w")
for line in sys.stdin:
        data = line.strip().split(",");
        key_month= data[1]
        if data[5].strip().isdigit():
                value_flights= float (data[5].strip())
        else:
                value_flights=0
        print("{0}\t{1}".format(key_month,value_flights))

```

5.2.2 Reducer Code:

5b_reducerX.py

```
#!/usr/bin/python3
import sys
Dict = {1: 0, 2: 0, 3: 0, 4: 0, 5: 0, 6: 0, 7: 0, 8: 0, 9: 0, 10: 0, 11: 0, 12: 0}
for line in sys.stdin:
        data = line.strip().split("\t")
        thiskey= data[0]
        flights = data[1]
        if thiskey.strip().isdigit():
                val= int(Dict[int (thiskey)])
                val+= float (flights)
                Dict.update({int (thiskey): int (val)})
busiest_month=  1
busiest_month_flights =Dict[1];
for x in Dict:
        if int (Dict[x]) > busiest_month_flights:
                busiest_month=x
                busiest_month_flights= Dict[x]

print("\nThe busiest Month is : {0}\tTotal flight of this month : {1}"
.format(busiest_month, busiest_month_flights))
```

5.2.3 Command:

- hdfs dfs -put /home/itskhuramshahzad/5b_reducerX.py /usr/hdoop/Task_5b/5b_reducerX.py
- hdfs dfs -put /home/itskhuramshahzad/5b_reducerX.py /usr/hdoop/Task_5b/5b_reducerX.py
- hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.2.3.jar -file /home/itskhuramshahzad/5b_mapperX.py -file /home/itskhuramshahzad/5b_reducerX.py -mapper 5b_mapperX.py -reducer ./reducerX.py -iducerX.py -mapper 5b_mapperX.py -reducer ./5b_reducerX.py -input /usr/hdoop/airline_data_best_results.csv -output /usr/hdoop/Task_5b/output1
- hdfs dfs -cat /usr/hdoop/Task_5b/output2/part-00000

5.3 TASK 5C: WHICH AIRLINE CARRIER HAS FLOWN THE MOST FLIGHTS OVER THE 10 YEAR PERIOD?

<key, value>=	<Airline, value_flights >
Output Folder:	/usr/hdoop/Task_5c/output/
Sample Output	

```
2022-06-22 18:58:10,396 INFO streaming.StreamJob: Output directory: /usr/hdoop/Task_5c/output1
hdoop@instance-1:~$ hdfs dfs -cat /usr/hdoop/Task_5c/output1/part-00000

Airline Carrier : WN     has flown most flights over the 10 year period are: 12333317.0
hdoop@instance-1:~$
```

5.3.1 Commands used for this task:

- hdfs dfs -mkdir /usr/hdoop/Task_5c
- hdfs dfs -put /home/itskhuramshahzad/5c_reducerX.py /usr/hdoop/Task_5c/5c_reducerX.py
- hdfs dfs -put /home/itskhuramshahzad/5c_mapperX.py /usr/hdoop/Task_5c/5c_mapperX.py

5.3.2 Mapper Code:

```python
#!/usr/bin/python3
import sys
import csv
for line in sys.stdin:
        data = line.strip().split(",");
        key_airline= data[2]
        if data[5].strip().isdigit():
                value_flights= float (data[5].strip())
        else:
                value_flights=0
        print("{0}\t{1}".format(key_airline,value_flights))
```

5.3.3 Reducer Code:

```python
#!/usr/bin/python3
import sys
Dict = {}
for line in sys.stdin:
        data = line.strip().split("\t")
        thiskey= data[0]
        carrier = data[1]
        if thiskey in Dict.keys():
                val= float(Dict[thiskey])
                val+= float (carrier)
                Dict.update({thiskey: float (val)})
        else:
                Dict[thiskey]= float (carrier)
busiest_carrier=  ""
busiest_carrier_flights =0;
flag= True
for x in Dict:
        if flag:
                busiest_carrier= x
                busiest_carrier_flights= usiest_carrier_flights= Dict[x]
                flag= False
        elif float (Dict[x]) > busiest_carrier_flights:
                busiest_carrier=x
                busiest_carrier_flights= Dict[x]
print("\nAirline Carrier : {0}\t has flown most flights over the 10 year period are: {1}"
.format(busiest_carrier, busiest_carrier_flights))
```


Prepared by: Engr. Khuram Shahzad (p21-8742)

5.4 TASK 5D: WHICH AIRPORT HAS BEEN THE MOST BUSIEST OVER THE 10 YEAR?

<key, value>=	<ariport, flights>
Output Folder:	/usr/hdoop/Task_5d/output/
Sample Output	

```
hdoop@instance-1:~$ hdfs dfs -cat /usr/hdoop/Task_5d/output/part-00000

 The Airport: ATL  has been the most busiest over the 10 year period , flights: 3881775
hdoop@instance-1:~$
```

5.4.1 Command used in this task:

- hdfs dfs -mkdir /usr/hdoop/Task_5d
- hdfs dfs -put /home/itskhuramshahzad/5d_reducerX.py /usr/hdoop/Task_5d/5d_reducerX.py
- hdfs dfs -put /home/itskhuramshahzad/5d_mapperX.py /usr/hdoop/Task_5d/5d_mapperX.py
- hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.2.3.jar -file /home/itskhuramshahzad/5d_mapperX.py -file /home/itskhuramshahzad/5d_reducerX.py -mapper 5d_mapperX.py -reducer ./5d_reducerX.py -input /usr/hdoop/airline_data_best_results.csv -output /usr/hdoop/Task_5d/output

5.4.2 Mapper Code:

```python
#!/usr/bin/python3
import sys
import csv
for line in sys.stdin:
        data = line.strip().split(",");
        key_airport= data[4]
        if data[5].strip().isdigit():
                value_flights=  int (data[5].strip())
        else:
                value_flights=0
        value_percentage=0
        print("{0}\t{1}".format(key_airport,value_flights))

```

5.4.3 Reducer Code:

```python
#!/usr/bin/python3
import sys
Dict = {}
for line in sys.stdin:
        data = line.strip().split("\t")
        thiskey= data[0]
        flights = data[1]
        if thiskey in Dict.keys():
                val= int(Dict[thiskey])
                val+= int (flights)
                Dict.update({thiskey: int (val)})
        else:
                Dict[thiskey]= int (flights)
busiest_airport=  ""
busiest_airport_flights =0;
flag= True
for x in Dict:
        if flag:
                busiest_airport= x
                busiest_airport_flights= Dict[x]
                flag= False
        elif int (Dict[x]) > busiest_airport_flights:
                busiest_airport=x
                busiest_airport_flights= Dict[x]
print("\n The Airport: {0}  has been the most busiest over the 10 year period , flights: {1}"
.format(busiest_airport, busiest_airport_flights))
```


Prepared by: Engr. Khuram Shahzad (p21-8742)

5.5 TASK 5E: WHICH AIRPORT HAS THE LARGEST FLIGHTS TO CANCELLATION RATIO?

<key, value>=	<key_airport,value_flights, value_cancel>
Output Folder:	/usr/hdoop/Task_5e/output/
Sample Output	

```
hdoop@instance-1:~$ hdfs dfs -cat /usr/hdoop/Task_5e/output/part-00000

 The Airport: ORD  has the Largest Flights to Cancellation Ratio [ 3028013 : 88288.0 ] = 34.296993928959765
```

5.5.1 Mapper Code:

5e_mapperX.py

```
#!/usr/bin/python3
import sys
value_flights=0
value_cancel=0

for line in sys.stdin:
        data = line.strip().split(",");
        key_airport= data[4]
        if data[5].strip().isdigit() and data[15].strip().isdigit():
                value_flights=  int (data[5].strip())
                value_cancel=  int (data[15].strip())
        else:
                value_flights=0
                value_cancel=0
        print("{0}\t{1}\t{2}".format(key_airport,value_flights, value_cancel))

```

5.5.2 Reducer Code:

Prepared by: Engr. Khuram Shahzad (p21-8742)

5e_mapperX.py

```
#!/usr/bin/python3
import sys
Dict = {}
for line in sys.stdin:
        data = line.strip().split("\t")
        thiskey= data[0]
        flights = data[1]
        cancellation = data[2]
        if thiskey in Dict.keys():
                val= Dict[thiskey]
                val_1 = val['0'];
                val_2 = val['1'];
                val_1+= int (flights)
                val_2+= float (cancellation)
                new_entry = {"0": val_1, "1": val_2}
                Dict[thiskey].update(new_entry)
                #print("airport: {0}\t{1}\t {2}".format(thiskey, val_1, val_2))
        else:
                val_1= int (flights)
                val_2= float (cancellation)
                new_entry = {"0": val_1, "1": val_2}
                Dict[thiskey]= new_entry
```

5e_reducerX.py

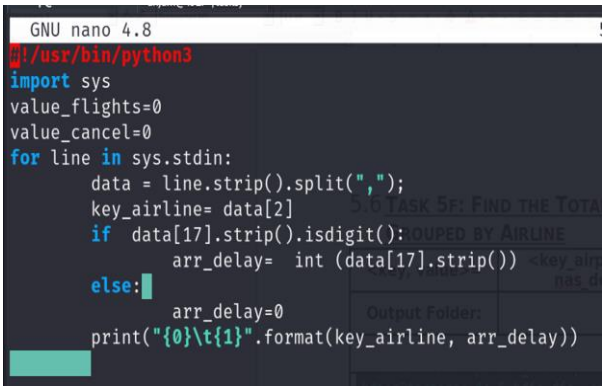
```
                val_1= int (flights)
                val_2= float (cancellation)
                new_entry = {"0": val_1, "1": val_2}
                Dict[thiskey]= new_entry
airport=  ""
airport_flights =0;
airport_cancellation=0
flag= True
for x in Dict:
        val= Dict[x]
        flights= val['0']
        cancellation= val['1']
        if flag:
                airport= x
                airport_flights= val['0']
                airport_cancellation= val['1']
                flag= False
        elif float (cancellation) > airport_cancellation:
                val= Dict[x]
                airport= x
                airport_flights= val['0']
                airport_cancellation= val['1']
print("\n The Airport: {0}  has the Largest Flights to Cancellation Ratio [ {1} : {2} ] = {3}"
.format(airport, airport_flights, airport_cancellation,airport_flights/airport_cancellation ))
```

5.6 TASK 5F: FIND THE TOTAL AMOUNT OF DELAY MINUTES GROUPED BY AIRLINE

<key, value>=	<airline, arr_delay>
Output Folder:	/usr/hdoop/Task_5f/output4/
Sample Output (Sorted in descending Order)	

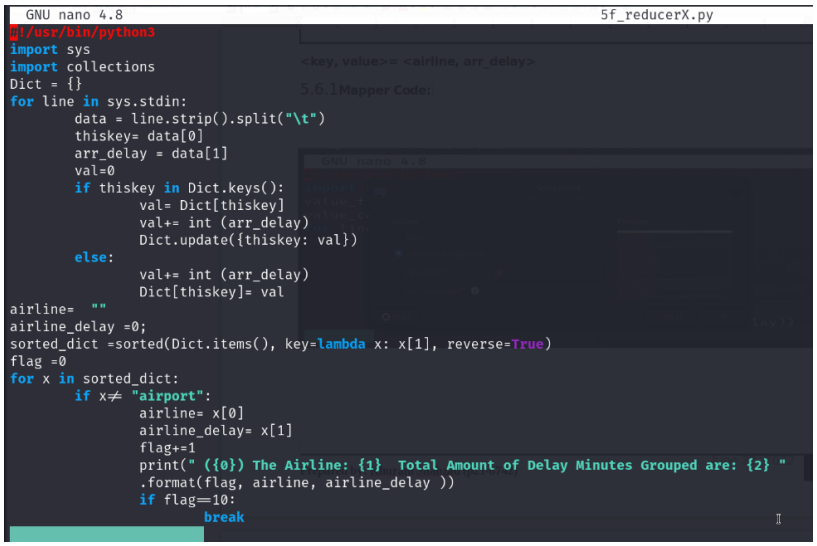
```
hdoop@instance-1:~$ hdfs dfs -cat /usr/hdoop/Task_5f/output4/part-00000
 (1) The Airline: WN  Total Amount of Delay Minutes Grouped are: 118025352
 (2) The Airline: AA  Total Amount of Delay Minutes Grouped are: 83920254
 (3) The Airline: OO  Total Amount of Delay Minutes Grouped are: 78760906
 (4) The Airline: DL  Total Amount of Delay Minutes Grouped are: 75761785
 (5) The Airline: EV  Total Amount of Delay Minutes Grouped are: 66347717
 (6) The Airline: UA  Total Amount of Delay Minutes Grouped are: 62923308
 (7) The Airline: B6  Total Amount of Delay Minutes Grouped are: 40695644
 (8) The Airline: MQ  Total Amount of Delay Minutes Grouped are: 37970355
 (9) The Airline: US  Total Amount of Delay Minutes Grouped are: 17821781
 (10) The Airline: F9  Total Amount of Delay Minutes Grouped are: 14194561
```

5.6.1 Mapper Code:



```
GNU nano 4.8
#!/usr/bin/python3
import sys
value_flights=0
value_cancel=0
for line in sys.stdin:
        data = line.strip().split(",");
        key_airline= data[2]
        if  data[17].strip().isdigit():
                arr_delay=  int (data[17].strip())
        else:
                arr_delay=0
        print("{0}\t{1}".format(key_airline, arr_delay))
```

5.6.2 Reducer Code:



5.7 TASK 5G: FIND THE AIRPORT WITH MOST CANCELED FLIGHTS IN 2016.

<key, value>=	<key_airport, arr_cancelled>
Output Folder:	/usr/hdoop/Task_5g/output/
Sample Output	

```
hdoop@instance-1:~$ hdfs dfs -cat /usr/hdoop/Task_5g/output/part-00000

The Airport: WN  with most Cancelled Flights in 2016 are: 15866.0
hdoop@instance-1:~$
```

5.7.1 Mapper Code:

```
#!/usr/bin/python3
import sys
value_flights=0
value_cancel=0
for line in sys.stdin:
        data = line.strip().split(",");
        key_year= data[0]
        key_airport= data[2]
        if key_year == "2016":
                if  data[15].strip().isdigit():
                        arr_delay=  float (data[15].strip())
                else:
                        arr_delay=0
                print("{0}\t{1}".format(key_airport,arr_cancelled))
```

5.7.2 Reducer Code:

```
#!/usr/bin/python3
import sys
import collections
Dict = {}
for line in sys.stdin:
        data = line.strip().split("\t")
        thiskey= data[0]
        cancellation = data[1]
        if thiskey in Dict.keys():
                val= Dict[thiskey]
                val+= int (cancellation)
                Dict.update({thiskey: val})
        else:
                val_1= int (cancellation)
                Dict[thiskey]= val_1
airline=  ""
airline_cancellation =0;
sorted_dict =sorted(Dict.items(), key=lambda x: x[1], reverse=True)
flag =0
for x in sorted_dict:
        if x!= "arr_cancelled":
                airline= x[0]
                airline_cancellation= x[1]
                flag+=1
                print("The Airport: {0}  with most Cancelled Flights in 2016 are: {1} "
                .format(airline, airline_cancellation ))
                if flag==1:
                        break
```

5.8 TASK 5H: FIND THE AVERAGE DELAY TIME FOR AN AIRPORT THAT IS THE MOST BUSIEST OF ALL OTHER AIRPORTS.

<key, value>	<key_airport, value_flights, value_delay>
Output Folder:	/usr/hdoop/Task_5h/output/
Sample Output	

```
hdoop@instance-1:~$ hdfs dfs -cat /usr/hdoop/Task_5h/output/part-00000

 The Airport: ATL  the most busiest of all other airports flights: 3881775, the average delay time: 5.692734664422602
hdoop@instance-1:~$
```

5.8.1 Mapper Code

5h_mapperX.py

```
#!/usr/bin/python3
import sys
value_flights=0
value_delay=0

for line in sys.stdin:
        data = line.strip().split(",");
        key_airport= data[4]
        if data[5].strip().isdigit() and data[17].strip().isdigit():
                value_flights=  int (data[5].strip())
                value_delay=  int (data[17].strip())
        else:
                value_flights=0
                value_delay=0
        print("{0}\t{1}\t{2}".format(key_airport,value_flights, value_delay))
```

5.8.2 Reducer Code:

5h_mapperX.py

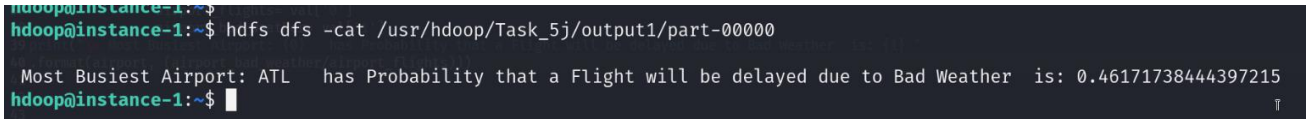
```
#!/usr/bin/python3
import sys
Dict = {}
for line in sys.stdin:
        data = line.strip().split("\t")
        thiskey= data[0]
        flights = data[1]
        delay = data[2]
        if thiskey in Dict.keys():
                val= Dict[thiskey]
                val_1 = val['0'];
                val_2 = val['1'];
                val_1+= int (flights)
                val_2+= float (delay)
                new_entry = {"0": val_1, "1": val_2}
                Dict[thiskey].update(new_entry)
                #print("airport: {0}\t{1}\t {2}".format
        else:
                val_1= int (flights)
                val_2= float (delay)
                new_entry = {"0": val_1, "1": val_2}
                Dict[thiskey]= new_entry
airport=  ""
airport_flights =0;
airport_delay=0
total_delay=0

flag= True
```


Prepared by: Engr. Khuram Shahzad (p21-8742)

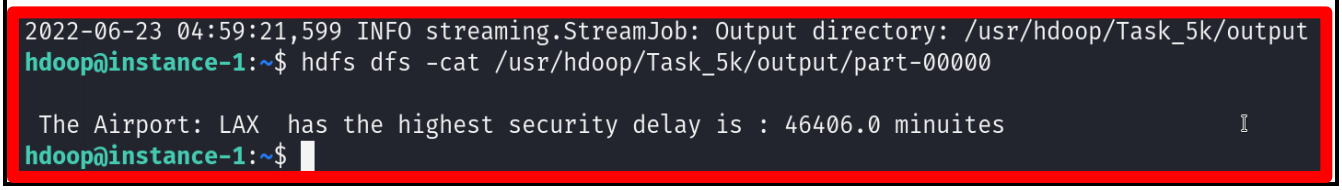
5.9 TASK 5J: WHAT IS THE PROBABILITY THAT A FLIGHT WILL BE CANCELLED DUE TO BAD WEATHER AT THE MOST BUSIEST AIRPORT OF ALL OTHER AIRPORTS?

<key, value>	<key_airport,value_flights, value_bad_weather)>
5.9.1 Output Folder:	/usr/hdoop/Task_5j/output/
Sample Output	



5.10 TASK 5K: FIND OUT THE AIRPORT THAT HAVE HIGHEST SECURITY DELAY THAN OF ALL OTHER AIRPORTS?

<key, value>	<key_airport,security_delay>
Output Folder:	/usr/hdoop/Task_5k/output/
Sample Output	



5.10.1 Mapper code:

5k_reducerX.py

```
#!/usr/bin/python3
import sys
for line in sys.stdin:
        data = line.strip().split(",");
        key_airport= data[4]
        if data[21].strip().isdigit():
                security_delay=  int (data[21].strip())
        else:
                security_delay=0
        print("{0}\t{1}".format(key_airport,security_delay))
```

5.10.2 Reducer Code:

```
#!/usr/bin/python3
import sys
Dict = {}
for line in sys.stdin:
        data = line.strip().split("\t")
        thiskey= data[0]
        security_delay = data[1]
        if thiskey in Dict.keys():
                val= Dict[thiskey]
                val+= float (security_delay)
                Dict.update({thiskey: val})
        else:
                val_1= float (security_delay)
                Dict[thiskey]= val_1
airport=  ""
security_delay=0
flag= True
for x in Dict:
        val= Dict[x]
        if flag:
                airport= x
                security_delay= val
                flag= False
        elif float (val) > security_delay:
                val= Dict[x]
                airport= x
                security_delay= val
print("\n The Airport: {0}  has the highest security delay is : {1} minuite"
.format(airport, security_delay))
```


Prepared by: Engr. Khuram Shahzad (p21-8742)