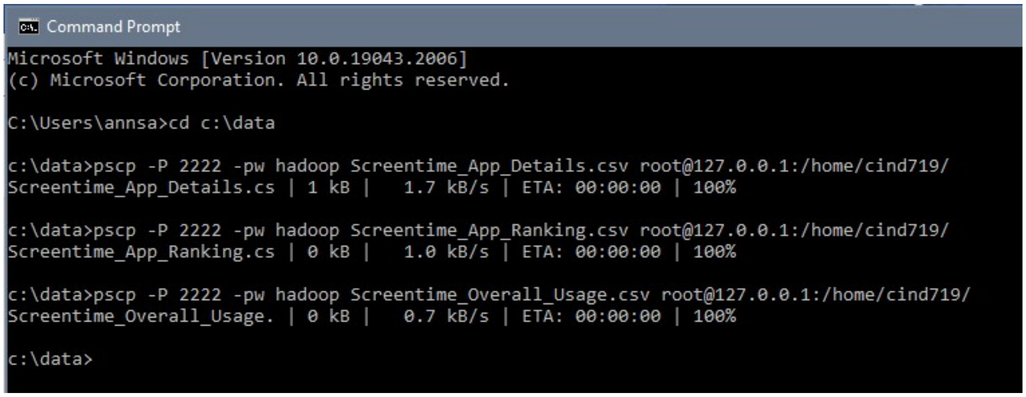
Question 1

- Import instructions
  - Transferring the files from host OS to HDP access node using the pspc utility in Windows Command Prompt:

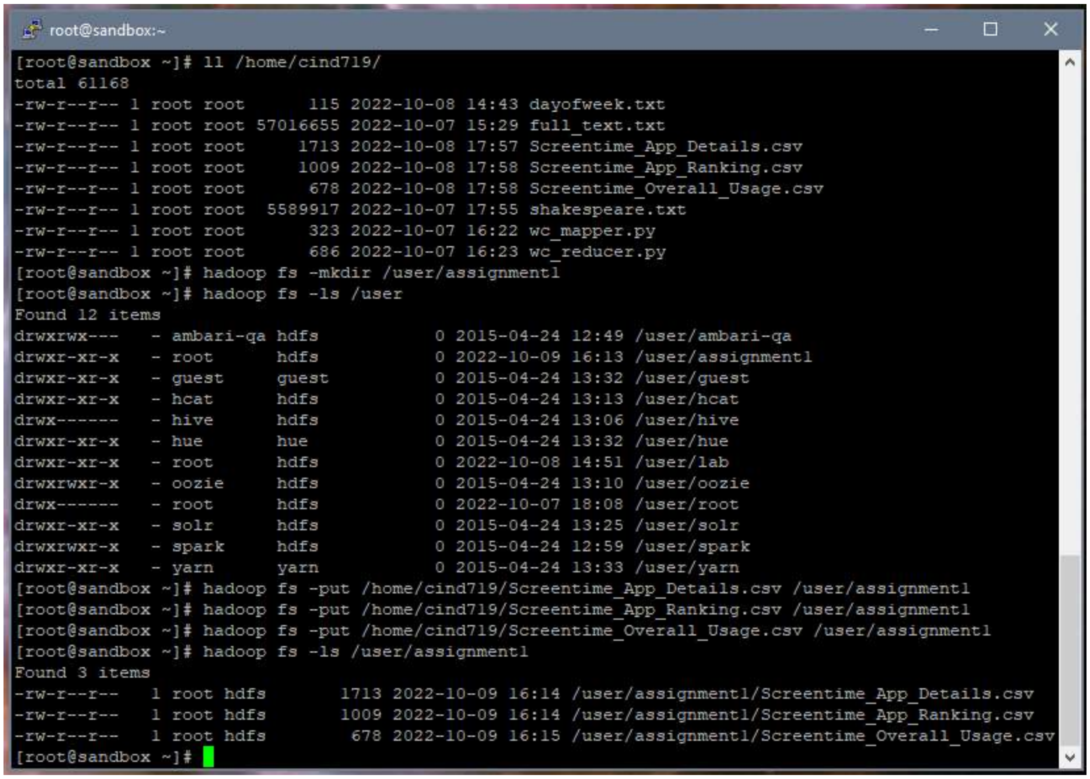
```
pscp -P 2222 -pw hadoop Screentime_App_Details.csv
root@127.0.0.1:/home/cind719/
pscp -P 2222 -pw hadoop Screentime_App_Ranking.csv
root@127.0.0.1:/home/cind719/
pscp -P 2222 -pw hadoop Screentime_Overall_Usage.csv
root@127.0.0.1:/home/cind719/
```



```
Microsoft Windows [Version 10.0.19043.2006]
(c) Microsoft Corporation. All rights reserved.

C:\Users\annsa>cd c:\data

c:\data>pscp -P 2222 -pw hadoop Screentime_App_Details.csv root@127.0.0.1:/home/cind719/
Screentime_App_Details.cs | 1 kB |   1.7 kB/s | ETA: 00:00:00 | 100%

c:\data>pscp -P 2222 -pw hadoop Screentime_App_Ranking.csv root@127.0.0.1:/home/cind719/
Screentime_App_Ranking.cs | 0 kB |   1.0 kB/s | ETA: 00:00:00 | 100%

c:\data>pscp -P 2222 -pw hadoop Screentime_Overall_Usage.csv root@127.0.0.1:/home/cind719/
Screentime_Overall_Usage. | 0 kB |   0.7 kB/s | ETA: 00:00:00 | 100%

c:\data>
```

  - Creating a new directory using the HDP terminal and moving the files to the HDFS:

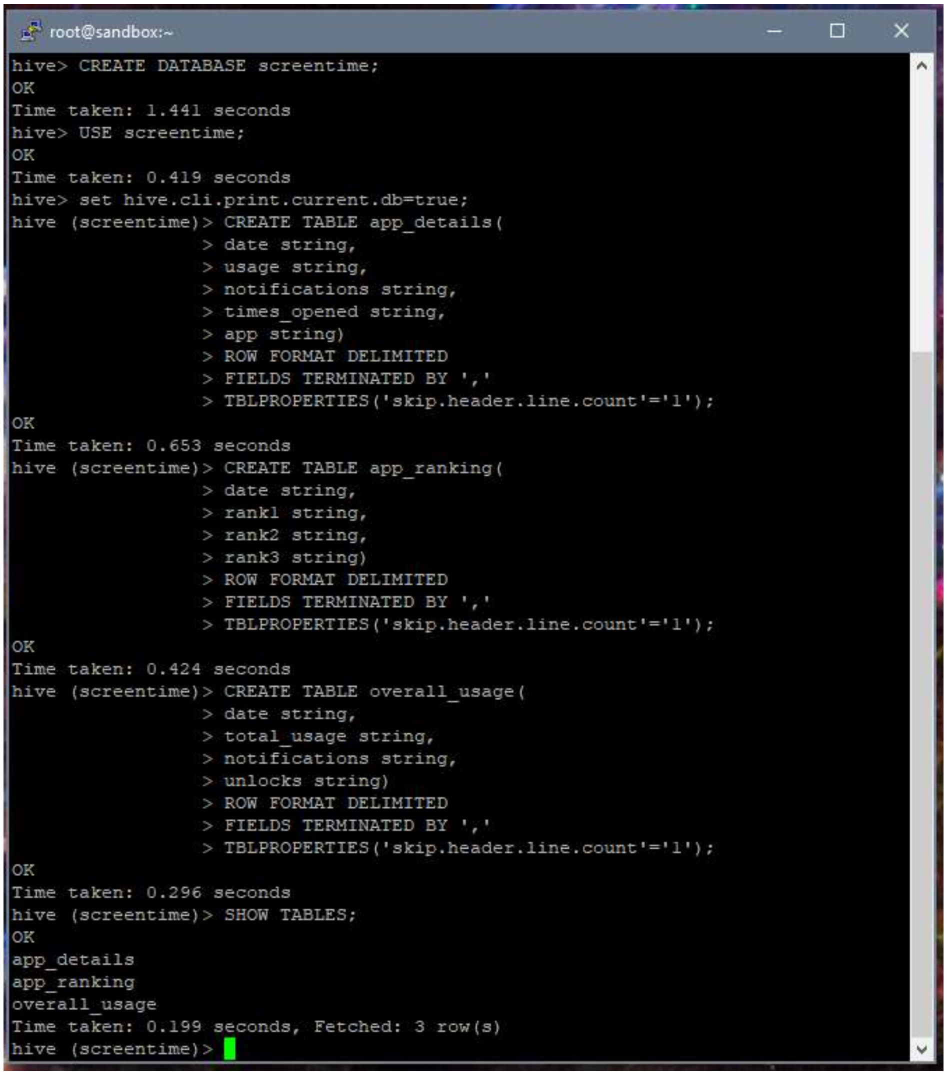
```
hadoop fs -mkdir /user/assignment1
hadoop fs -put /home/cind719/Screentime_App_Details.csv
/user/assignment1
hadoop fs -put /home/cind719/Screentime_App_Ranking.csv
/user/assignment1
hadoop fs -put /home/cind719/Screentime_Overall_Usage.csv
user/assignment1
```



```
[root@sandbox ~]# ll /home/cind719/
total 61168
-rw-r--r-- 1 root root      115 2022-10-08 14:43 dayofweek.txt
-rw-r--r-- 1 root root 57016655 2022-10-07 15:29 full_text.txt
-rw-r--r-- 1 root root     1713 2022-10-08 17:57 Screentime_App_Details.csv
-rw-r--r-- 1 root root     1009 2022-10-08 17:58 Screentime_App_Ranking.csv
-rw-r--r-- 1 root root      678 2022-10-08 17:58 Screentime_Overall_Usage.csv
-rw-r--r-- 1 root root  5589917 2022-10-07 17:55 shakespeare.txt
-rw-r--r-- 1 root root      323 2022-10-07 16:22 wc_mapper.py
-rw-r--r-- 1 root root      686 2022-10-07 16:23 wc_reducer.py
[root@sandbox ~]# hadoop fs -mkdir /user/assignment1
[root@sandbox ~]# hadoop fs -ls /user
Found 12 items
drwxrwx---   - ambari-qa hdfs          0 2015-04-24 12:49 /user/ambari-qa
drwxr-xr-x   - root      hdfs          0 2022-10-09 16:13 /user/assignment1
drwxr-xr-x   - guest     guest         0 2015-04-24 13:32 /user/guest
drwxr-xr-x   - hcat      hdfs          0 2015-04-24 13:13 /user/hcat
drwx------   - hive      hdfs          0 2015-04-24 13:06 /user/hive
drwxr-xr-x   - hue       hue           0 2015-04-24 13:32 /user/hue
drwxr-xr-x   - root      hdfs          0 2022-10-08 14:51 /user/lab
drwxrwxr-x   - oozie     hdfs          0 2015-04-24 13:10 /user/oozie
drwx------   - root      hdfs          0 2022-10-07 18:08 /user/root
drwxr-xr-x   - solr      hdfs          0 2015-04-24 13:25 /user/solr
drwxrwxr-x   - spark     hdfs          0 2015-04-24 12:59 /user/spark
drwxr-xr-x   - yarn      yarn          0 2015-04-24 13:33 /user/yarn
[root@sandbox ~]# hadoop fs -put /home/cind719/Screentime_App_Details.csv /user/assignment1
[root@sandbox ~]# hadoop fs -put /home/cind719/Screentime_App_Ranking.csv /user/assignment1
[root@sandbox ~]# hadoop fs -put /home/cind719/Screentime_Overall_Usage.csv /user/assignment1
[root@sandbox ~]# hadoop fs -ls /user/assignment1
Found 3 items
-rw-r--r--   1 root hdfs       1713 2022-10-09 16:14 /user/assignment1/Screentime_App_Details.csv
-rw-r--r--   1 root hdfs       1009 2022-10-09 16:14 /user/assignment1/Screentime_App_Ranking.csv
-rw-r--r--   1 root hdfs        678 2022-10-09 16:15 /user/assignment1/Screentime_Overall_Usage.csv
[root@sandbox ~]#
```

- Database and table creation scripts:

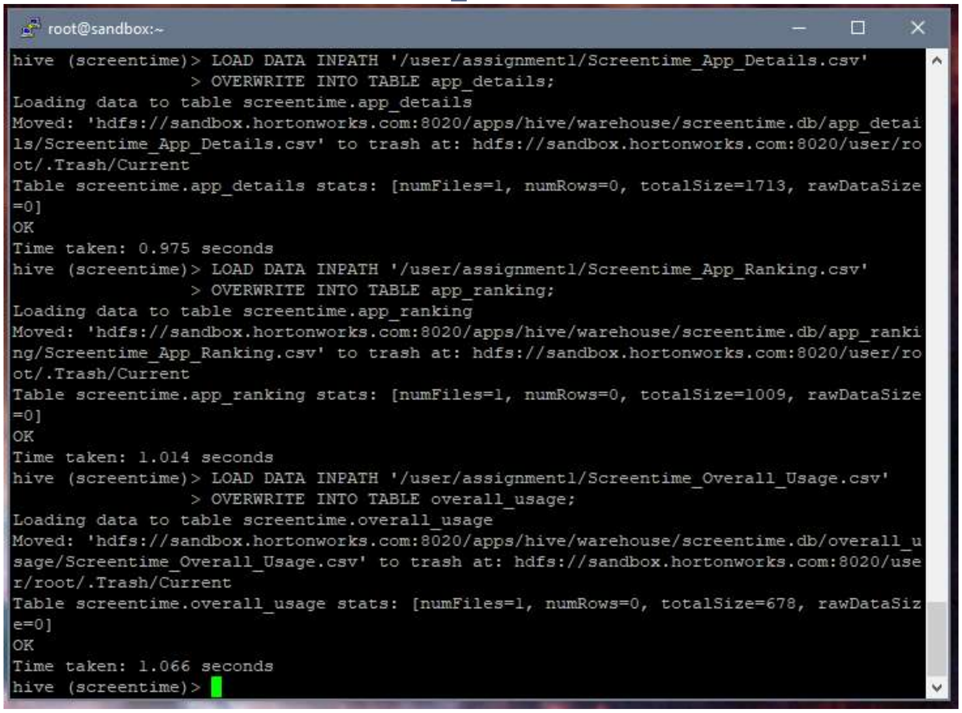
```
CREATE DATABASE screentime;
USE screentime;
set hive.cli.print.current.db=true;

CREATE TABLE detail_app(date string, usage string, notifications
string, time_opened string, app string) ROW FORMAT DELIMITED FIELDS
TERMINATED BY ‘,’ TBLPROPERTIES(‘skip.header.line.count’=’1’);

CREATE TABLE detail_ranking(date string, rank1 string, rank2 string,
rank3 string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’
TBLPROPERTIES(‘skip.header.line.count’=’1’);

CREATE TABLE overall_usage(date string, total_usage string,
notifications string, unlocks string) ROW FORMAT DELIMITED FIELDS
TERMINATED BY ‘,’ TBLPROPERTIES(‘skip.header.line.count’=’1’);
```



```
hive> CREATE DATABASE screentime;
OK
Time taken: 1.441 seconds
hive> USE screentime;
OK
Time taken: 0.419 seconds
hive> set hive.cli.print.current.db=true;
hive (screentime)> CREATE TABLE app_details(
                 > date string,
                 > usage string,
                 > notifications string,
                 > times_opened string,
                 > app string)
                 > ROW FORMAT DELIMITED
                 > FIELDS TERMINATED BY ','
                 > TBLPROPERTIES('skip.header.line.count'='1');
OK
Time taken: 0.653 seconds
hive (screentime)> CREATE TABLE app_ranking(
                 > date string,
                 > rank1 string,
                 > rank2 string,
                 > rank3 string)
                 > ROW FORMAT DELIMITED
                 > FIELDS TERMINATED BY ','
                 > TBLPROPERTIES('skip.header.line.count'='1');
OK
Time taken: 0.424 seconds
hive (screentime)> CREATE TABLE overall_usage(
                 > date string,
                 > total_usage string,
                 > notifications string,
                 > unlocks string)
                 > ROW FORMAT DELIMITED
                 > FIELDS TERMINATED BY ','
                 > TBLPROPERTIES('skip.header.line.count'='1');
OK
Time taken: 0.296 seconds
hive (screentime)> SHOW TABLES;
OK
app_details
app_ranking
overall_usage
Time taken: 0.199 seconds, Fetched: 3 row(s)
hive (screentime)>
```

- Importing csv files into tables:

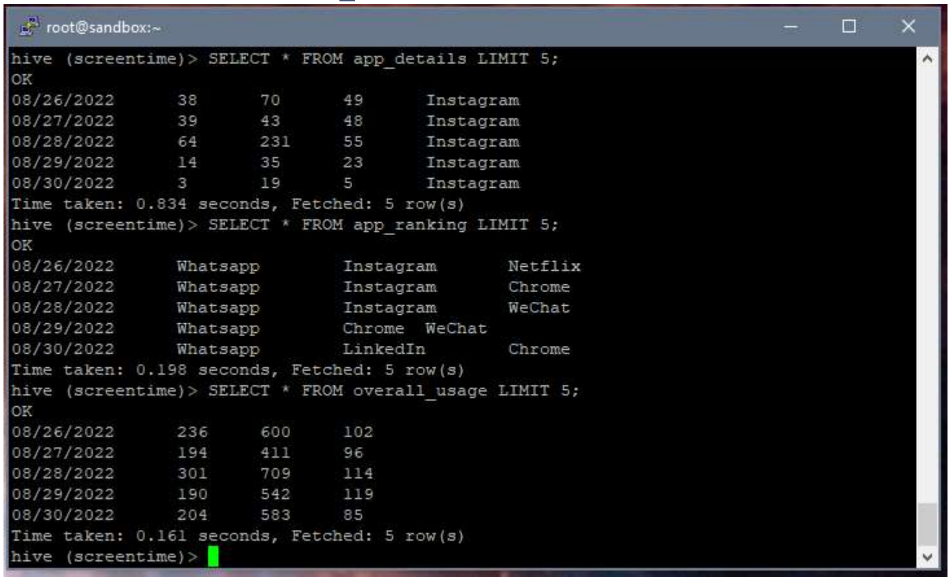
```
LOAD DATA INPATH ‘/user/assignment1/Screentime_App_Details.csv’
OVERWRITE INTO TABLE app_details;

LOAD DATA INPATH ‘/user/assignment1/Screentime_App_Ranking.csv’
OVERWRITE INTO TABLE app_ranking;

LOAD DATA INPATH ‘/user/assignment1/Screentime_Overall_Usage.csv’
OVERWRITE INTO TABLE overall_usage;
```



```
hive (screentime)> LOAD DATA INPATH '/user/assignment1/Screentime_App_Details.csv'
                 > OVERWRITE INTO TABLE app_details;
Loading data to table screentime.app_details
Moved: 'hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/screentime.db/app_detai
ls/Screentime_App_Details.csv' to trash at: hdfs://sandbox.hortonworks.com:8020/user/ro
ot/.Trash/Current
Table screentime.app_details stats: [numFiles=1, numRows=0, totalSize=1713, rawDataSize
=0]
OK
Time taken: 0.975 seconds
hive (screentime)> LOAD DATA INPATH '/user/assignment1/Screentime_App_Ranking.csv'
                 > OVERWRITE INTO TABLE app_ranking;
Loading data to table screentime.app_ranking
Moved: 'hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/screentime.db/app_ranki
ng/Screentime_App_Ranking.csv' to trash at: hdfs://sandbox.hortonworks.com:8020/user/ro
ot/.Trash/Current
Table screentime.app_ranking stats: [numFiles=1, numRows=0, totalSize=1009, rawDataSize
=0]
OK
Time taken: 1.014 seconds
hive (screentime)> LOAD DATA INPATH '/user/assignment1/Screentime_Overall_Usage.csv'
                 > OVERWRITE INTO TABLE overall_usage;
Loading data to table screentime.overall_usage
Moved: 'hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/screentime.db/overall_u
sage/Screentime_Overall_Usage.csv' to trash at: hdfs://sandbox.hortonworks.com:8020/use
r/root/.Trash/Current
Table screentime.overall_usage stats: [numFiles=1, numRows=0, totalSize=678, rawDataSiz
e=0]
OK
Time taken: 1.066 seconds
hive (screentime)>
```

- Displaying first 5 rows from each table:

```
SELECT * FROM app_details LIMIT 5;
SELECT * FROM app_ranking LIMIT 5;
SELECT * FROM overall_usage LIMIT 5;
```



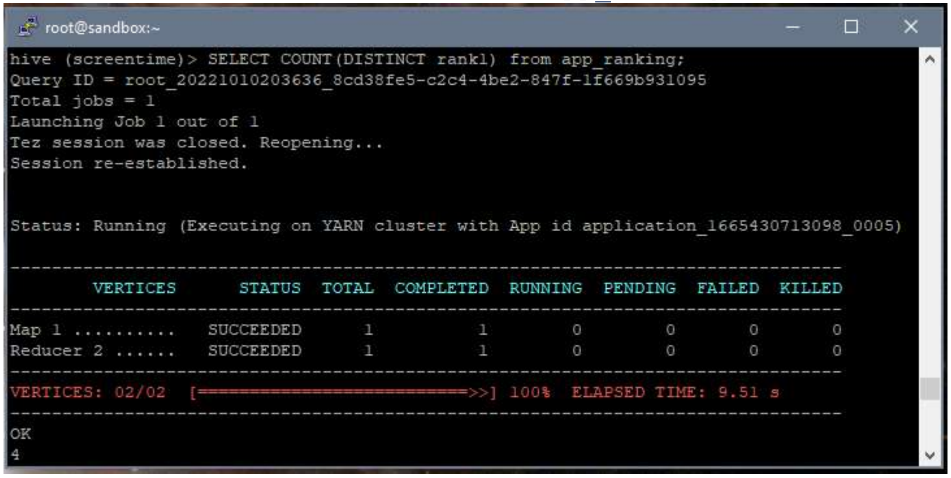
```
hive (screentime)> SELECT * FROM app_details LIMIT 5;
OK
08/26/2022      38      70      49      Instagram
08/27/2022      39      43      48      Instagram
08/28/2022      64      231     55      Instagram
08/29/2022      14      35      23      Instagram
08/30/2022      3       19      5       Instagram
Time taken: 0.834 seconds, Fetched: 5 row(s)
hive (screentime)> SELECT * FROM app_ranking LIMIT 5;
OK
08/26/2022      Whatsapp        Instagram       Netflix
08/27/2022      Whatsapp        Instagram       Chrome
08/28/2022      Whatsapp        Instagram       WeChat
08/29/2022      Whatsapp        Chrome  WeChat
08/30/2022      Whatsapp        LinkedIn        Chrome
Time taken: 0.198 seconds, Fetched: 5 row(s)
hive (screentime)> SELECT * FROM overall_usage LIMIT 5;
OK
08/26/2022      236     600     102
08/27/2022      194     411     96
08/28/2022      301     709     114
08/29/2022      190     542     119
08/30/2022      204     583     85
Time taken: 0.161 seconds, Fetched: 5 row(s)
hive (screentime)>
```

Question 2

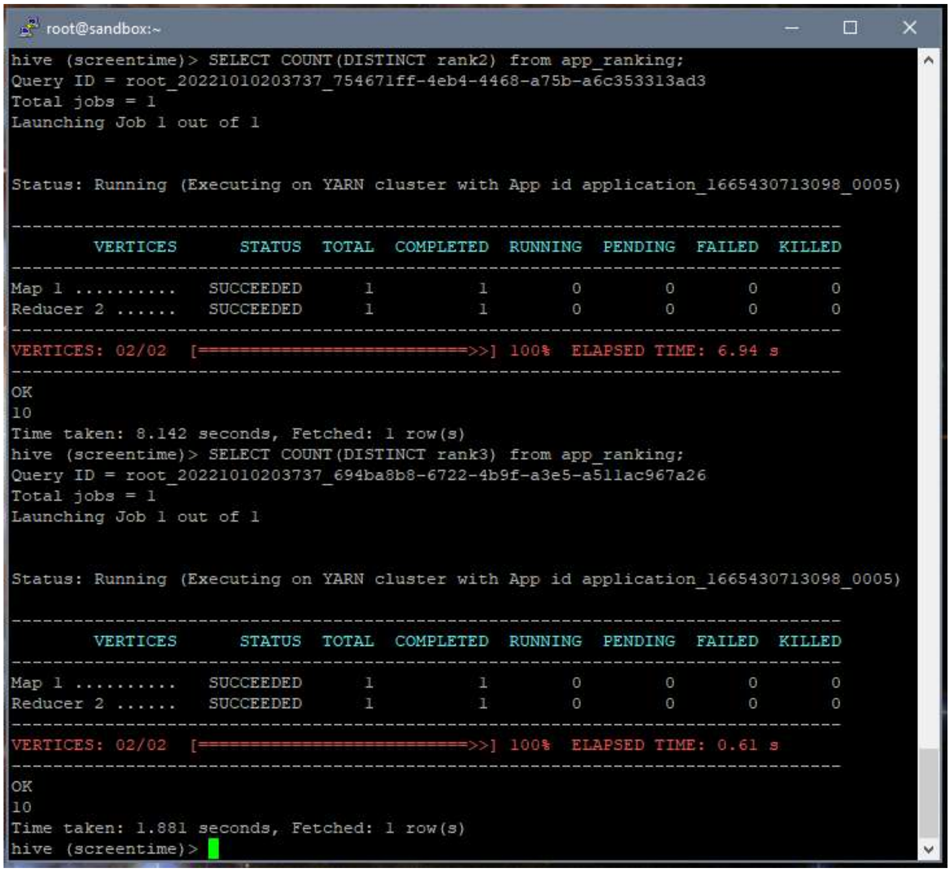
Find the unique number of applications in Rank 1, Rank 2 and Rank 3 columns. You can write 3 different HIVE commands for each Rank column.

- Unique number of applications in Rank 1: 4
  Unique number of applications in Rank 2: 10
  Unique number of applications in Rank 3: 10

```
SELECT COUNT(DISTINCT rank1) FROM app_ranking;
SELECT COUNT(DISTINCT rank2) FROM app_ranking;
SELECT COUNT(DISTINCT rank3) FROM app_ranking;
```



```
hive (screentime)> SELECT COUNT(DISTINCT rank1) from app_ranking;
Query ID = root_20221010203636_8cd38fe5-c2c4-4be2-847f-1f669b931095
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1665430713098_0005)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.51 s
----------------------------------------------------------------------------------------------
OK
4
```



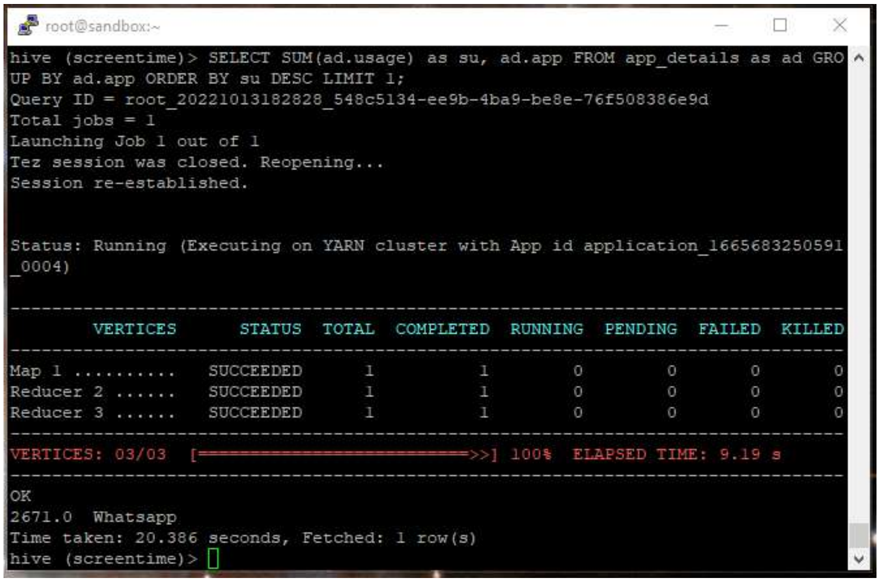
```
hive (screentime)> SELECT COUNT(DISTINCT rank2) from app_ranking;
Query ID = root_20221010203737_754671ff-4eb4-4468-a75b-a6c353313ad3
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1665430713098_0005)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.94 s
----------------------------------------------------------------------------------------------
OK
10
Time taken: 8.142 seconds, Fetched: 1 row(s)
hive (screentime)> SELECT COUNT(DISTINCT rank3) from app_ranking;
Query ID = root_20221010203737_694ba8b8-6722-4b9f-a3e5-a511ac967a26
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1665430713098_0005)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.61 s
----------------------------------------------------------------------------------------------
OK
10
Time taken: 1.881 seconds, Fetched: 1 row(s)
hive (screentime)>
```

Question 3

Which application has maximum usage?

- Application with the maximum usage: Whatsapp

```
SELECT SUM(ad.usage) as su, ad.app FROM app_details as ad GROUP BY ad.app
ORDER BY su DESC LIMIT 1;
```



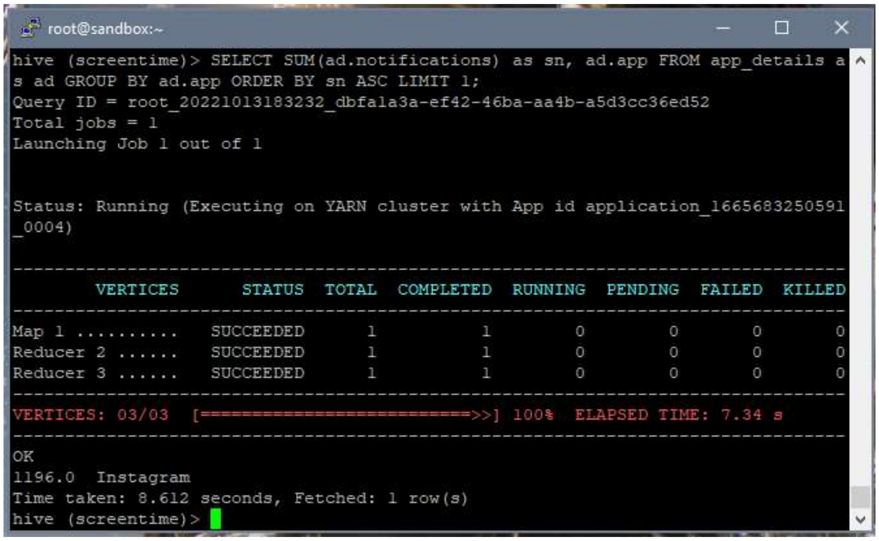
```
hive (screentime)> SELECT SUM(ad.usage) as su, ad.app FROM app_details as ad GRO
UP BY ad.app ORDER BY su DESC LIMIT 1;
Query ID = root_20221013182828_548c5134-ee9b-4ba9-be8e-76f508386e9d
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1665683250591
_0004)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.19 s
----------------------------------------------------------------------------------
OK
2671.0  Whatsapp
Time taken: 20.386 seconds, Fetched: 1 row(s)
hive (screentime)>
```

Question 4

Which application has least no of Notifications?

- Application with the least number of Notification: Instagram

```
SELECT SUM(ad.notifications) as sn, ad.app FROM app_details as as GROUP BY
ad.app ORDER BY sn ASC LIMIT 1;
```



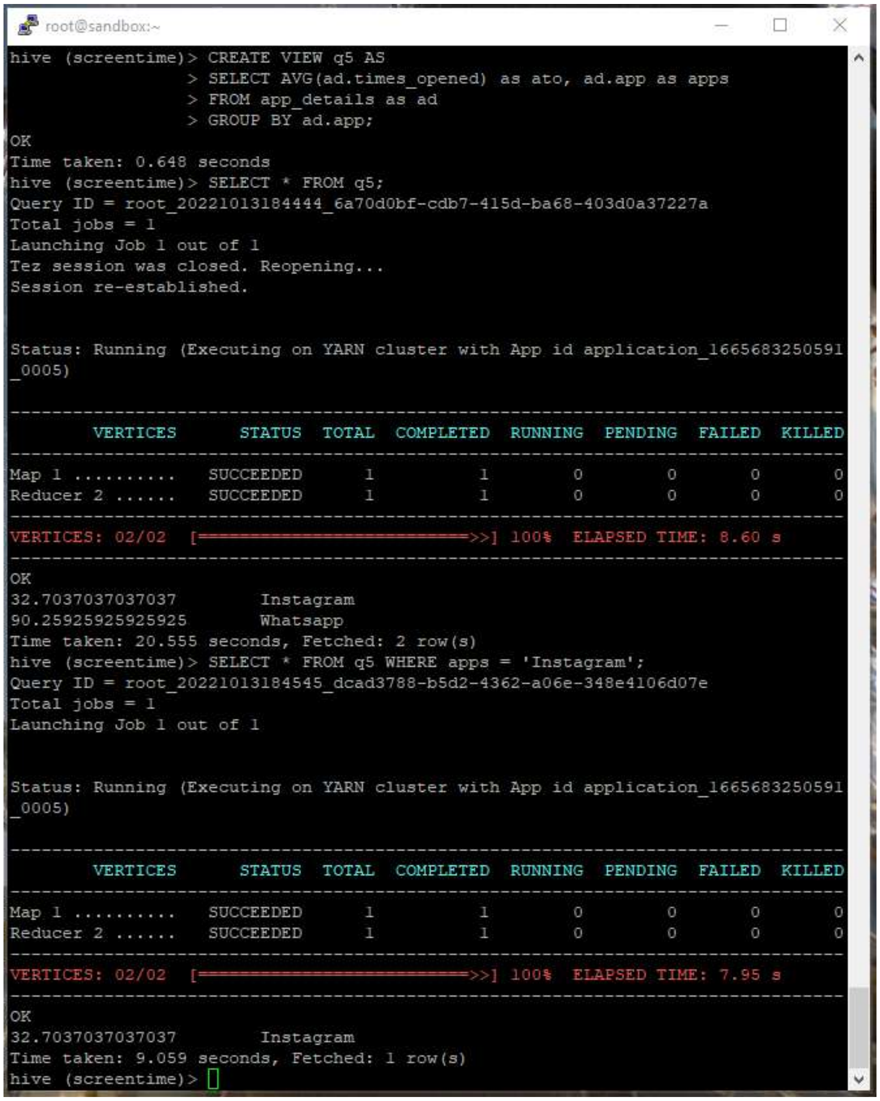
```
hive (screentime)> SELECT SUM(ad.notifications) as sn, ad.app FROM app_details a
s ad GROUP BY ad.app ORDER BY sn ASC LIMIT 1;
Query ID = root_20221013183232_dbfala3a-ef42-46ba-aa4b-a5d3cc36ed52
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1665683250591
_0004)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 7.34 s
----------------------------------------------------------------------------------
OK
1196.0  Instagram
Time taken: 8.612 seconds, Fetched: 1 row(s)
hive (screentime)>
```

Question 5

Find the average number of times an Instagram app has been opened up?

- Average number of times Instagram app is opened up: 32.7 times

```
CREATE VIEW q5 AS SELECT AVG(ad.times_opened), ad.app FROM app_details as ad
GROUP BY ad.app;

SELECT * FROM q5 WHERE apps = ‘Instagram’;
```

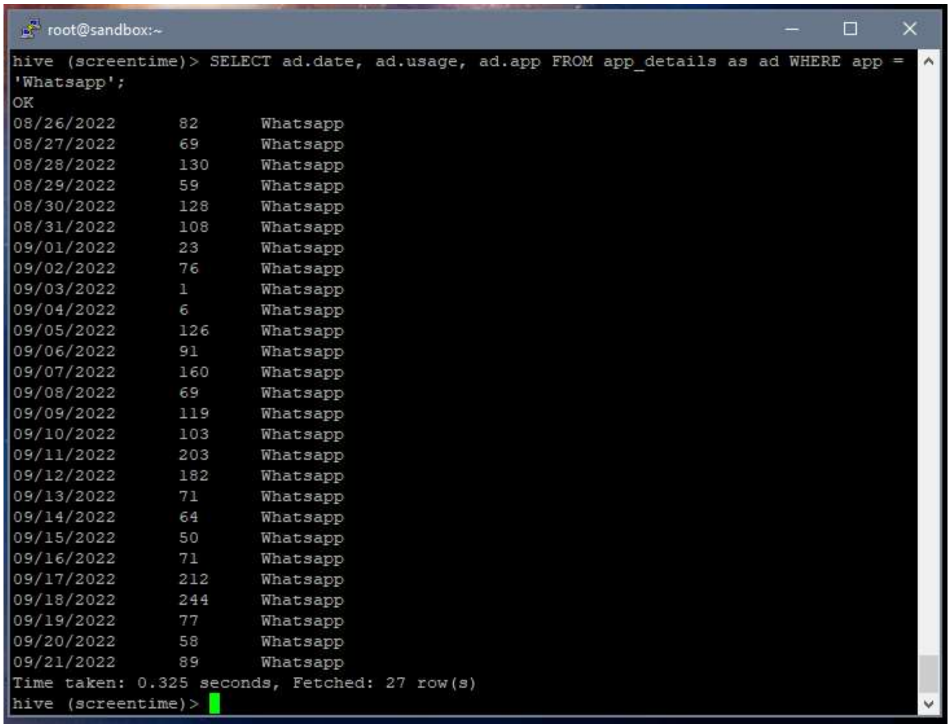


```
hive (screentime)> CREATE VIEW q5 AS
                 > SELECT AVG(ad.times_opened) as ato, ad.app as apps
                 > FROM app_details as ad
                 > GROUP BY ad.app;
OK
Time taken: 0.648 seconds
hive (screentime)> SELECT * FROM q5;
Query ID = root_20221013184444_6a70d0bf-cdb7-415d-ba68-403d0a37227a
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1665683250591
_0005)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.60 s
--------------------------------------------------------------------------------
OK
32.7037037037037        Instagram
90.25925925925925       Whatsapp
Time taken: 20.555 seconds, Fetched: 2 row(s)
hive (screentime)> SELECT * FROM q5 WHERE apps = 'Instagram';
Query ID = root_20221013184545_dcad3788-b5d2-4362-a06e-348e4106d07e
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1665683250591
_0005)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.95 s
--------------------------------------------------------------------------------
OK
32.7037037037037        Instagram
Time taken: 9.059 seconds, Fetched: 1 row(s)
hive (screentime)>
```

Question 6

Print total usage for WhatsApp application with dates?

```
SELECT ad.date, ad.usage, ad.app FROM app_details as ad WHERE app =
‘Whatsapp’;
```



```
hive (screentime)> SELECT ad.date, ad.usage, ad.app FROM app_details as ad WHERE app =
'Whatsapp';
OK
08/26/2022      82      Whatsapp
08/27/2022      69      Whatsapp
08/28/2022      130     Whatsapp
08/29/2022      59      Whatsapp
08/30/2022      128     Whatsapp
08/31/2022      108     Whatsapp
09/01/2022      23      Whatsapp
09/02/2022      76      Whatsapp
09/03/2022      1       Whatsapp
09/04/2022      6       Whatsapp
09/05/2022      126     Whatsapp
09/06/2022      91      Whatsapp
09/07/2022      160     Whatsapp
09/08/2022      69      Whatsapp
09/09/2022      119     Whatsapp
09/10/2022      103     Whatsapp
09/11/2022      203     Whatsapp
09/12/2022      182     Whatsapp
09/13/2022      71      Whatsapp
09/14/2022      64      Whatsapp
09/15/2022      50      Whatsapp
09/16/2022      71      Whatsapp
09/17/2022      212     Whatsapp
09/18/2022      244     Whatsapp
09/19/2022      77      Whatsapp
09/20/2022      58      Whatsapp
09/21/2022      89      Whatsapp
Time taken: 0.325 seconds, Fetched: 27 row(s)
hive (screentime)>
```