CIND719-DK0T
Midterm Report

Ramello Peralta 500519802

Question 1 (4 Points). Here is a setup of Data Centers, Racks, and Nodes of a HDFS processing facility:...give a potential minimum distance association list of processors to the data blocks.

Table 1.

Processors	Data Block Number	Data Block Location	Distance/cost
P1(D1/R1/N3)	B1	D1/R1/N3	0
P2(D1/R1/N7)	B2	D1/R2/N9	2
P3(D2/R4/N2)	B4	D2/R3/N5	4
P4(D1/R3/N8)	B3	D1/R3/N5	2

Total distance/cost 8

Table 2.

Processors	Data Block Number	Data Block Location	Distance/cost
P1(D1/R1/N3)	B2	D1/R2/N4	2
P2(D1/R1/N7)	B1	D1/R1/N8	2
P3(D2/R4/N2)	B4	D2/R3/N5	4
P4(D1/R3/N8)	B3	D1/R3/N5	2

Total distance/cost 10

Table 1 is better due to lower potential minimum distance.

Question 2 (6 Points). The following text is sent to a MapReduce to count the words (the text is provided in lowercase letters in purpose):

"the king was watching the knights fighting
for their king, their country and for their honor.
the fearless knights and the king never lost a battle"

The text is split into 3 Mapper tasks (each line goes to a separate mapper) and eventually processed by a single Reducer.

Q2.1 Provide the outputs of each mapper after processing the text.

MAP 1

The	1 1
King	1
Was	1
Watching	1
Knights	1
Fighting	1

MAP 2

For	1 1
Their	1 1 1
King	1
Country	1
And	1
Honor	1

MAP 3

The	1 1
Fearless	1
Knights	1
And	1
King	1
Never	1
Lost	1
A	1
Battle	1

Q2.2 Provide the output of the reducer after completing its task.

Reducer

The	4
King	3
Was	1
Watching	1
Knights	2
Fighting	1
For	2
Their	3
Country	1
And	1
Honor	1
Fearless	1
Never	1
Lost	1
A	1
Battle	1

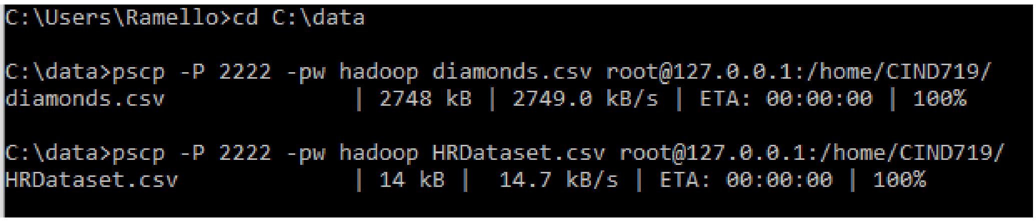
Question 3 (5 Points). An analyst is supposed to store the following information in the "diamonds" table that will be importing from a csv (comma separated value) file.

File schema:

- Id: row id for the data. The id contains only integer numbers
- Price: The sales price of the diamond with the properties given in US dollars ($326-$18,823)
- Carat: weight of the diamond (0.2--5.01)
- Cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- Color: diamond colour, from J (worst) to D (best)
- Clarity: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- X: length in mm (0 - 10.74)
- Y: width in mm (0 - 58.9)
- Z: depth in mm (0 - 31.8)
- Depth: total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
- Table Width: of top of diamond relative to widest point (43--95)

Note: The information between the last parenthesis represents the range of values or possible values of these database columns.

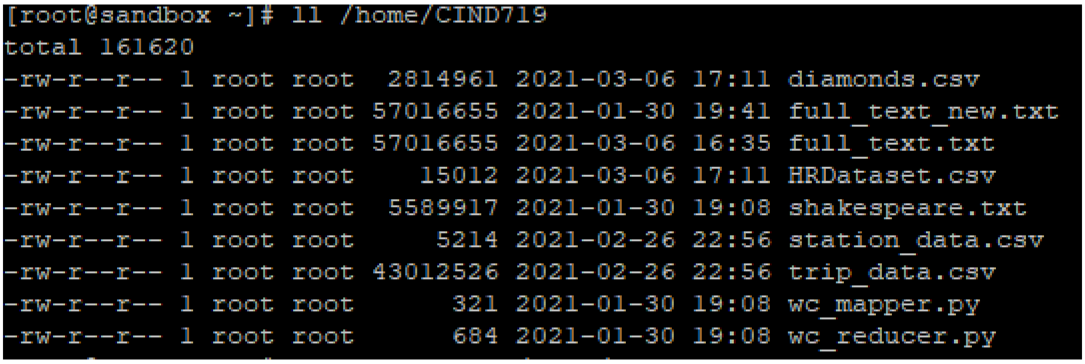
Q 3.1 Transfer the file to HDFS. Provide a screenshot of the console showing the file is in HDFS.



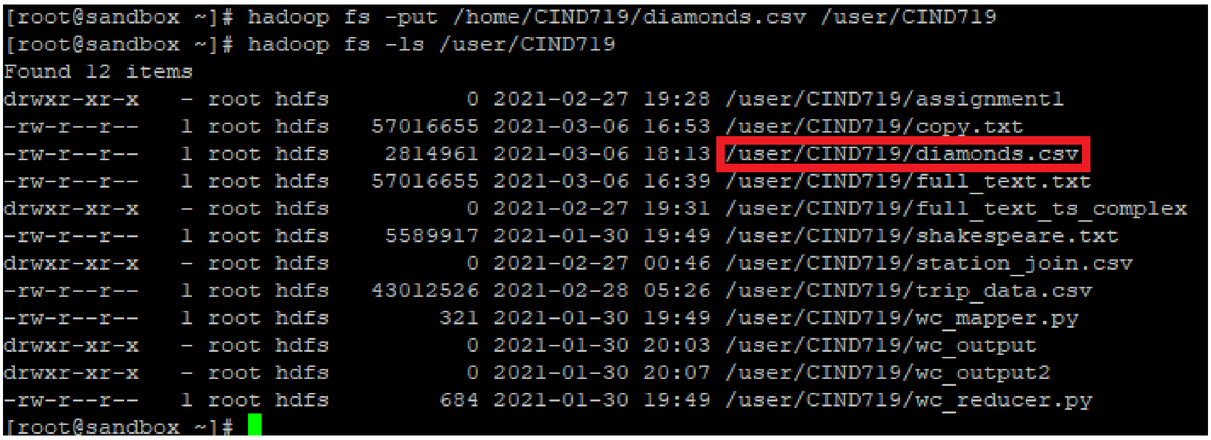
```
C:\Users\Ramello>cd C:\data

C:\data>pscp -P 2222 -pw hadoop diamonds.csv root@127.0.0.1:/home/CIND719/
diamonds.csv              | 2748 kB | 2749.0 kB/s | ETA: 00:00:00 | 100%

C:\data>pscp -P 2222 -pw hadoop HRDataset.csv root@127.0.0.1:/home/CIND719/
HRDataset.csv             | 14 kB |  14.7 kB/s | ETA: 00:00:00 | 100%
```

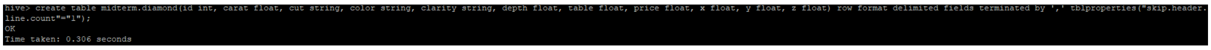


```
[root@sandbox ~]# ll /home/CIND719
total 161620
-rw-r--r-- 1 root root  2814961 2021-03-06 17:11 diamonds.csv
-rw-r--r-- 1 root root 57016655 2021-01-30 19:41 full_text_new.txt
-rw-r--r-- 1 root root 57016655 2021-03-06 16:35 full_text.txt
-rw-r--r-- 1 root root    15012 2021-03-06 17:11 HRDataset.csv
-rw-r--r-- 1 root root  5589917 2021-01-30 19:08 shakespeare.txt
-rw-r--r-- 1 root root     5214 2021-02-26 22:56 station_data.csv
-rw-r--r-- 1 root root 43012526 2021-02-26 22:56 trip_data.csv
-rw-r--r-- 1 root root      321 2021-01-30 19:08 wc_mapper.py
-rw-r--r-- 1 root root      684 2021-01-30 19:08 wc_reducer.py
```



```
[root@sandbox ~]# hadoop fs -put /home/CIND719/diamonds.csv /user/CIND719
[root@sandbox ~]# hadoop fs -ls /user/CIND719
Found 12 items
drwxr-xr-x   - root hdfs          0 2021-02-27 19:28 /user/CIND719/assignment1
-rw-r--r--   1 root hdfs   57016655 2021-03-06 16:53 /user/CIND719/copy.txt
-rw-r--r--   1 root hdfs    2814961 2021-03-06 18:13 /user/CIND719/diamonds.csv
-rw-r--r--   1 root hdfs   57016655 2021-03-06 16:39 /user/CIND719/full_text.txt
drwxr-xr-x   - root hdfs          0 2021-02-27 19:31 /user/CIND719/full_text_ts_complex
-rw-r--r--   1 root hdfs    5589917 2021-01-30 19:49 /user/CIND719/shakespeare.txt
drwxr-xr-x   - root hdfs          0 2021-02-27 00:46 /user/CIND719/station_join.csv
-rw-r--r--   1 root hdfs   43012526 2021-02-28 05:26 /user/CIND719/trip_data.csv
-rw-r--r--   1 root hdfs        321 2021-01-30 19:49 /user/CIND719/wc_mapper.py
drwxr-xr-x   - root hdfs          0 2021-01-30 20:03 /user/CIND719/wc_output
drwxr-xr-x   - root hdfs          0 2021-01-30 20:07 /user/CIND719/wc_output2
-rw-r--r--   1 root hdfs        684 2021-01-30 19:49 /user/CIND719/wc_reducer.py
[root@sandbox ~]#
```

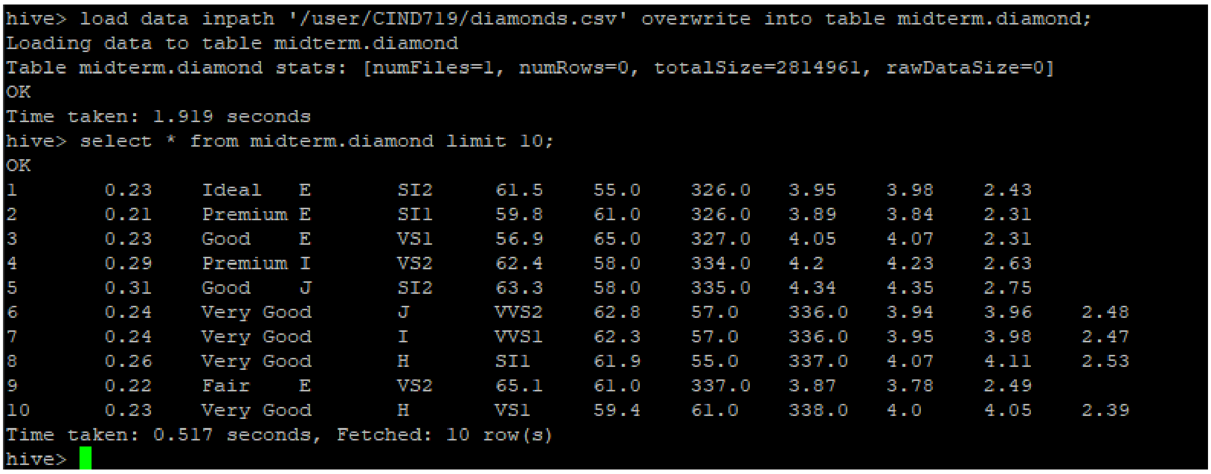
Q 3.2 Write the script to create the diamonds table in Hive. Provide screenshot of the console showing the table creation command and first ten rows of the hive table.



```
hive> create table midterm.diamond(id int, carat float, cut string, color string, clarity string, depth float, table float, price float, x float, y float, z float) row format delimited fields terminated by ',' tblproperties("skip.header.
line.count"="1");
OK
Time taken: 0.306 seconds
```

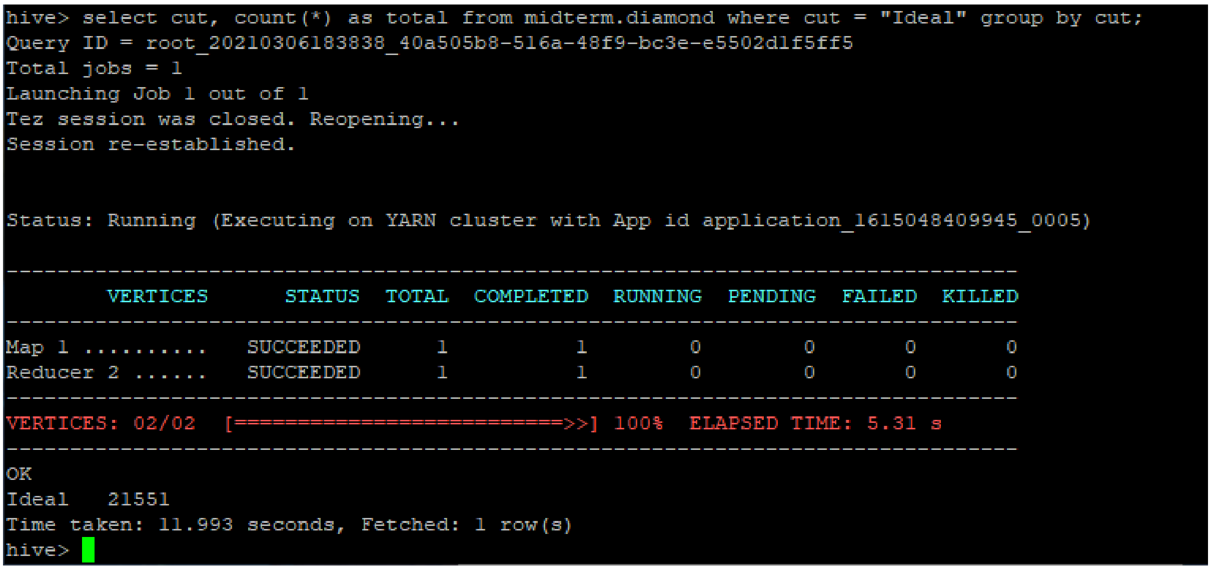
- For readability:

Create table midterm.diamond(id int, carat float, cut string, color string, clarity string, depth float, table float, price float, x float, y float, z float) row format delimited fields terminated by ‘,’ tblproperties(“skip.header.line.count”=”1”);
OK
Time taken: 0.306 seconds



```
hive> load data inpath '/user/CIND719/diamonds.csv' overwrite into table midterm.diamond;
Loading data to table midterm.diamond
Table midterm.diamond stats: [numFiles=1, numRows=0, totalSize=2814961, rawDataSize=0]
OK
Time taken: 1.919 seconds
hive> select * from midterm.diamond limit 10;
OK
1       0.23    Ideal   E       SI2     61.5    55.0    326.0   3.95    3.98    2.43
2       0.21    Premium E       SI1     59.8    61.0    326.0   3.89    3.84    2.31
3       0.23    Good    E       VS1     56.9    65.0    327.0   4.05    4.07    2.31
4       0.29    Premium I       VS2     62.4    58.0    334.0   4.2     4.23    2.63
5       0.31    Good    J       SI2     63.3    58.0    335.0   4.34    4.35    2.75
6       0.24    Very Good       J       VVS2    62.8    57.0    336.0   3.94    3.96    2.48
7       0.24    Very Good       I       VVS1    62.3    57.0    336.0   3.95    3.98    2.47
8       0.26    Very Good       H       SI1     61.9    55.0    337.0   4.07    4.11    2.53
9       0.22    Fair    E       VS2     65.1    61.0    337.0   3.87    3.78    2.49
10      0.23    Very Good       H       VS1     59.4    61.0    338.0   4.0     4.05    2.39
Time taken: 0.517 seconds, Fetched: 10 row(s)
hive>
```

Q 3.3 Write a query to get the total number of diamond rows where the Cut information is "Ideal". Provide screenshot of the console showing the query and the output.

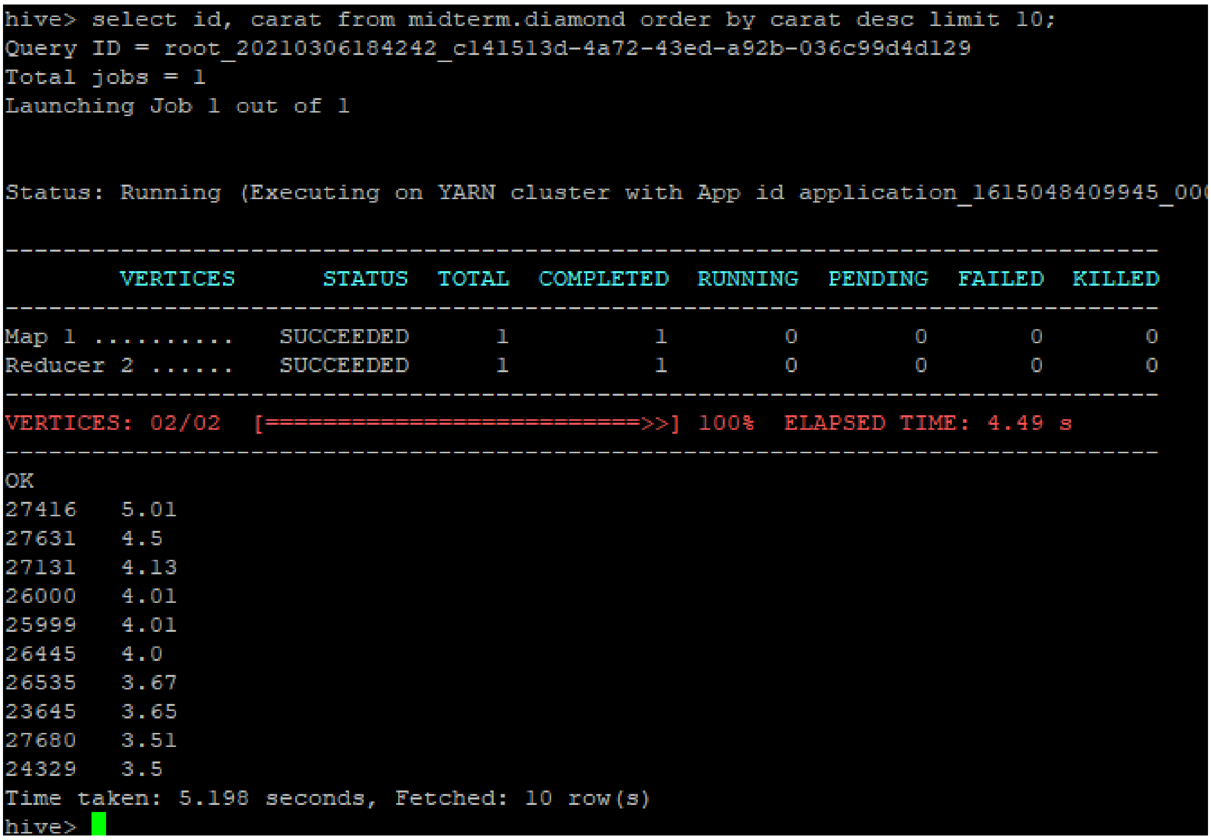


```
hive> select cut, count(*) as total from midterm.diamond where cut = "Ideal" group by cut;
Query ID = root_20210306183838_40a505b8-516a-48f9-bc3e-e5502d1f5ff5
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1615048409945_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: 5.31 s
--------------------------------------------------------------------------------
OK
Ideal   21551
Time taken: 11.993 seconds, Fetched: 1 row(s)
hive>
```

Q 3.4 Write a query to get the top 10 diamonds with the biggest weights (Carats). Provide screenshot of the console showing the query and the output.



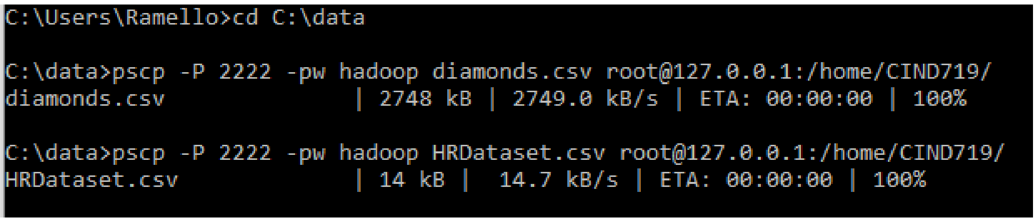
```
hive> select id, carat from midterm.diamond order by carat desc limit 10;
Query ID = root_20210306184242_c141513d-4a72-43ed-a92b-036c99d4d129
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1615048409945_00

--------------------------------------------------------------------------------
        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: 4.49 s
--------------------------------------------------------------------------------
OK
27416   5.01
27631   4.5
27131   4.13
26000   4.01
25999   4.01
26445   4.0
26535   3.67
23645   3.65
27680   3.51
24329   3.5
Time taken: 5.198 seconds, Fetched: 10 row(s)
hive>
```

Question 4 (5 Points). A schema of a simplified version of HR Dataset is given below.

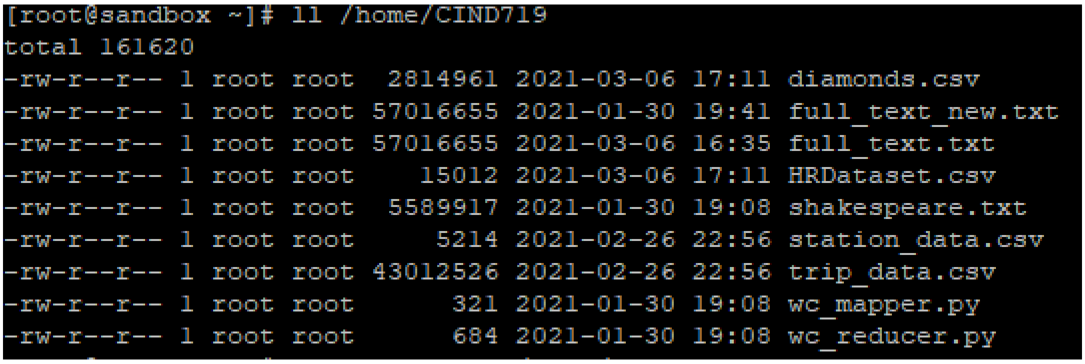
Q 4.1 Transfer the file to HDFS. Provide screenshot of the console showing the file is in HDFS.



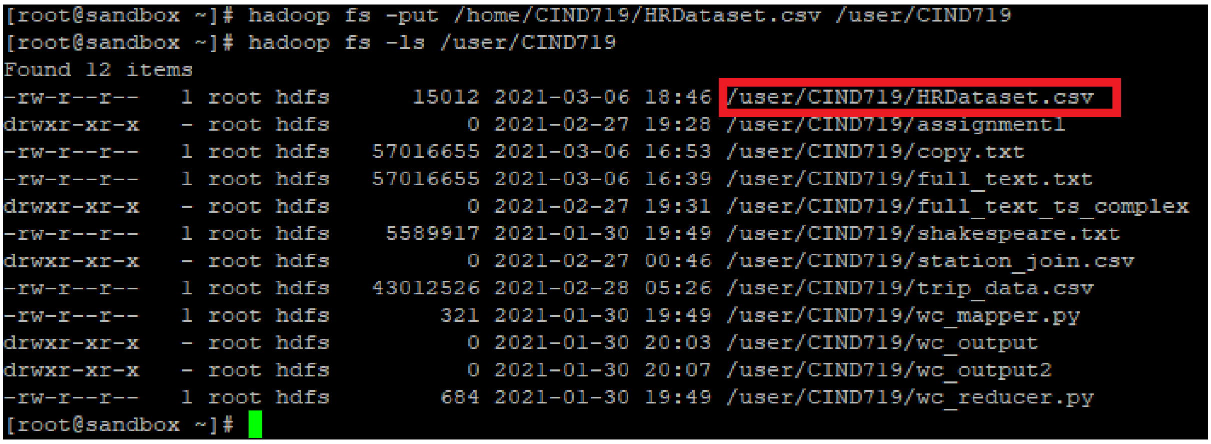
```
C:\Users\Ramello>cd C:\data

C:\data>pscp -P 2222 -pw hadoop diamonds.csv root@127.0.0.1:/home/CIND719/
diamonds.csv              | 2748 kB | 2749.0 kB/s | ETA: 00:00:00 | 100%

C:\data>pscp -P 2222 -pw hadoop HRDataset.csv root@127.0.0.1:/home/CIND719/
HRDataset.csv             | 14 kB |  14.7 kB/s | ETA: 00:00:00 | 100%
```



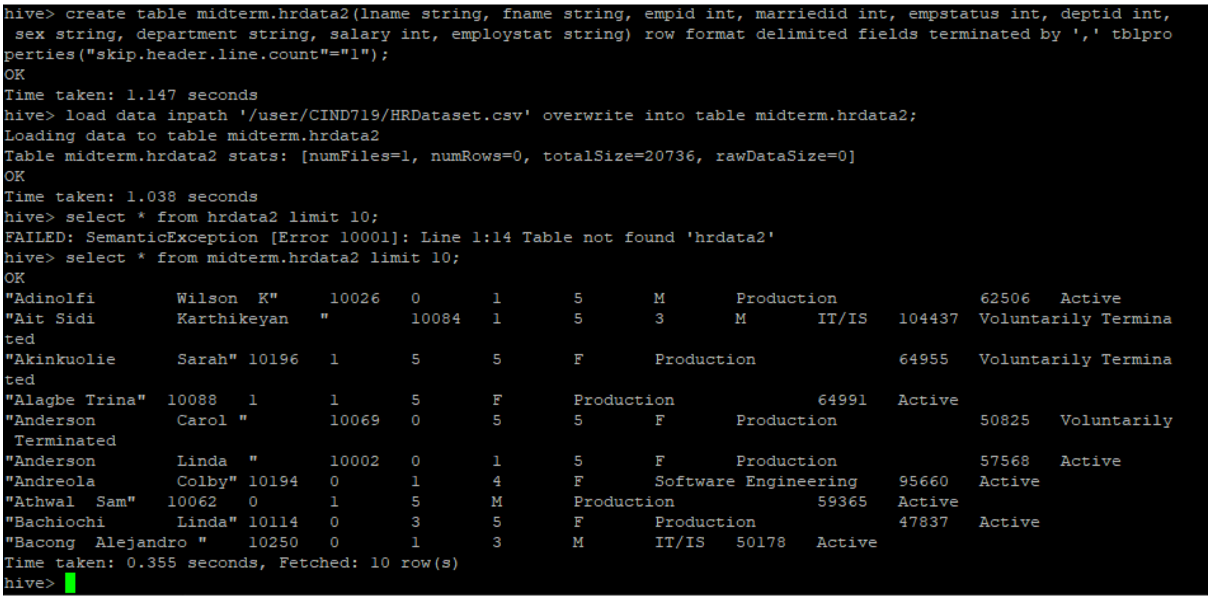
```
[root@sandbox ~]# ll /home/CIND719
total 161620
-rw-r--r-- 1 root root  2814961 2021-03-06 17:11 diamonds.csv
-rw-r--r-- 1 root root 57016655 2021-01-30 19:41 full_text_new.txt
-rw-r--r-- 1 root root 57016655 2021-03-06 16:35 full_text.txt
-rw-r--r-- 1 root root    15012 2021-03-06 17:11 HRDataset.csv
-rw-r--r-- 1 root root  5589917 2021-01-30 19:08 shakespeare.txt
-rw-r--r-- 1 root root     5214 2021-02-26 22:56 station_data.csv
-rw-r--r-- 1 root root 43012526 2021-02-26 22:56 trip_data.csv
-rw-r--r-- 1 root root      321 2021-01-30 19:08 wc_mapper.py
-rw-r--r-- 1 root root      684 2021-01-30 19:08 wc_reducer.py
```



```
[root@sandbox ~]# hadoop fs -put /home/CIND719/HRDataset.csv /user/CIND719
[root@sandbox ~]# hadoop fs -ls /user/CIND719
Found 12 items
-rw-r--r--   1 root hdfs      15012 2021-03-06 18:46 /user/CIND719/HRDataset.csv
drwxr-xr-x   - root hdfs          0 2021-02-27 19:28 /user/CIND719/assignment1
-rw-r--r--   1 root hdfs   57016655 2021-03-06 16:53 /user/CIND719/copy.txt
-rw-r--r--   1 root hdfs   57016655 2021-03-06 16:39 /user/CIND719/full_text.txt
drwxr-xr-x   - root hdfs          0 2021-02-27 19:31 /user/CIND719/full_text_ts_complex
-rw-r--r--   1 root hdfs    5589917 2021-01-30 19:49 /user/CIND719/shakespeare.txt
drwxr-xr-x   - root hdfs          0 2021-02-27 00:46 /user/CIND719/station_join.csv
-rw-r--r--   1 root hdfs   43012526 2021-02-28 05:26 /user/CIND719/trip_data.csv
-rw-r--r--   1 root hdfs        321 2021-01-30 19:49 /user/CIND719/wc_mapper.py
drwxr-xr-x   - root hdfs          0 2021-01-30 20:03 /user/CIND719/wc_output
drwxr-xr-x   - root hdfs          0 2021-01-30 20:07 /user/CIND719/wc_output2
-rw-r--r--   1 root hdfs        684 2021-01-30 19:49 /user/CIND719/wc_reducer.py
[root@sandbox ~]#
```

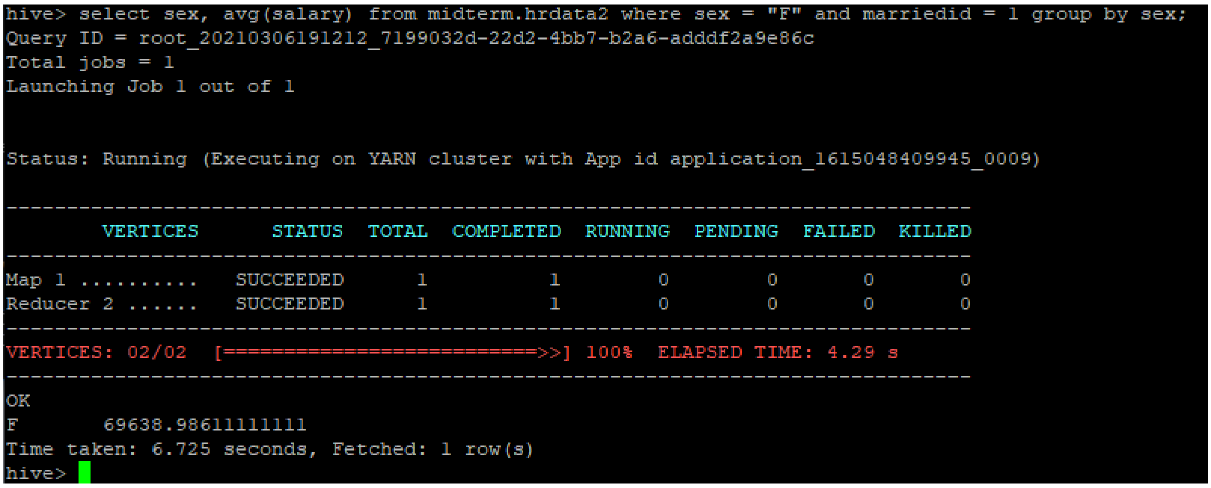
Q 4.2 Write the script to create hrdata table in Hive. Load the data into the table. Provide screenshot of the console showing the table creation command and first ten rows of the hive table.

- Emp name was split into two columns because hive doesn’t process in-quotation values as a single value and instead read the comma as a delimiter



```
hive> create table midterm.hrdata2(lname string, fname string, empid int, marriedid int, empstatus int, deptid int,
 sex string, department string, salary int, employstat string) row format delimited fields terminated by ',' tblpro
perties("skip.header.line.count"="1");
OK
Time taken: 1.147 seconds
hive> load data inpath '/user/CIND719/HRDataset.csv' overwrite into table midterm.hrdata2;
Loading data to table midterm.hrdata2
Table midterm.hrdata2 stats: [numFiles=1, numRows=0, totalSize=20736, rawDataSize=0]
OK
Time taken: 1.038 seconds
hive> select * from hrdata2 limit 10;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'hrdata2'
hive> select * from midterm.hrdata2 limit 10;
OK
"Adinolfi       Wilson  K"      10026   0       1       5       M       Production              62506   Active
"Ait Sidi       Karthikeyan    "        10084   1       5       3       M       IT/IS   104437  Voluntarily Termina
ted
"Akinkuolie     Sarah" 10196    1       5       5       F       Production              64955   Voluntarily Termina
ted
"Alagbe Trina"  10088   1       1       5       F       Production              64991   Active
"Anderson       Carol "         10069   0       5       5       F       Production              50825   Voluntarily
 Terminated
"Anderson       Linda   "       10002   0       1       5       F       Production              57568   Active
"Andreola       Colby" 10194    0       1       4       F       Software Engineering    95660   Active
"Athwal  Sam"   10062   0       1       5       M       Production              59365   Active
"Bachiochi      Linda" 10114    0       3       5       F       Production              47837   Active
"Bacong  Alejandro "    10250   0       1       3       M       IT/IS   50178   Active
Time taken: 0.355 seconds, Fetched: 10 row(s)
hive>
```

Q 4.3 Write a query to return the average salary for the married women in the table. Provide screenshot of the console showing the query and the output.

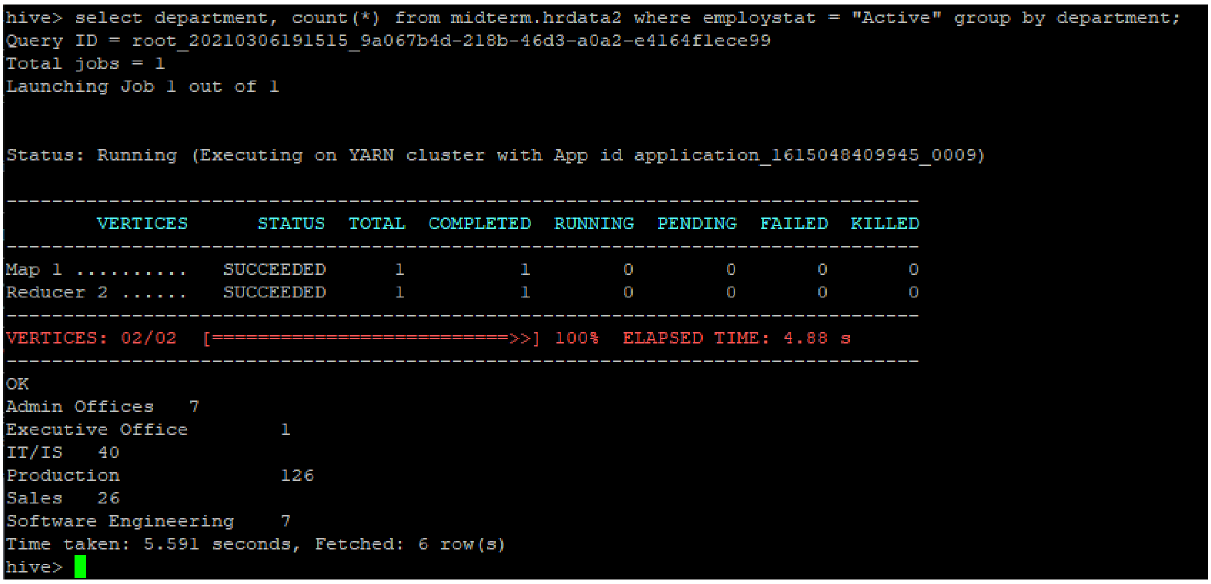


```
hive> select sex, avg(salary) from midterm.hrdata2 where sex = "F" and marriedid = 1 group by sex;
Query ID = root_20210306191212_7199032d-22d2-4bb7-b2a6-adddf2a9e86c
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1615048409945_0009)

--------------------------------------------------------------------------------
        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: 4.29 s
--------------------------------------------------------------------------------
OK
F       69638.98611111111
Time taken: 6.725 seconds, Fetched: 1 row(s)
hive>
```

Q 4.4 Write a query to return the number of working employees per departments. We do not want to count the employees that are not active. Provide screenshot of the console showing the query and the output.



```
hive> select department, count(*) from midterm.hrdata2 where employstat = "Active" group by department;
Query ID = root_20210306191515_9a067b4d-218b-46d3-a0a2-e4164f1ece99
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1615048409945_0009)

--------------------------------------------------------------------------------
        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: 4.88 s
--------------------------------------------------------------------------------
OK
Admin Offices   7
Executive Office        1
IT/IS   40
Production              126
Sales   26
Software Engineering    7
Time taken: 5.591 seconds, Fetched: 6 row(s)
hive> 
```