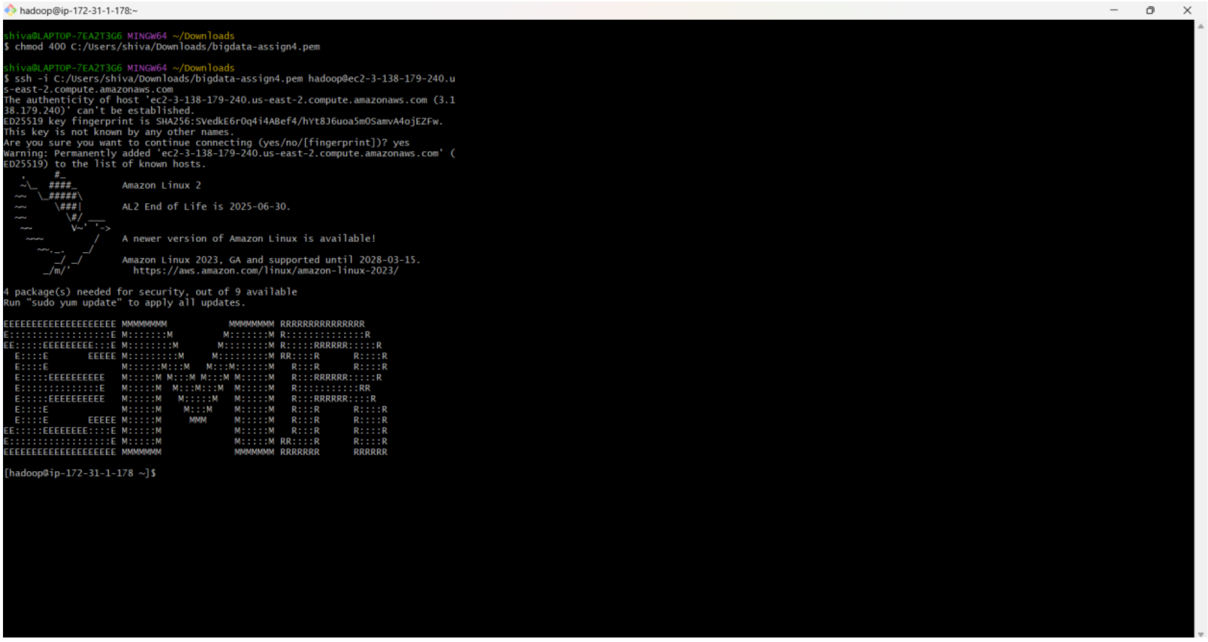
CSP 554 Big Data Technologies

Assignment – #4

Shiva Sankar Modala(A20517528)

Configure a Hadoop (AWS EMR) environment as you have done for previous assignments.

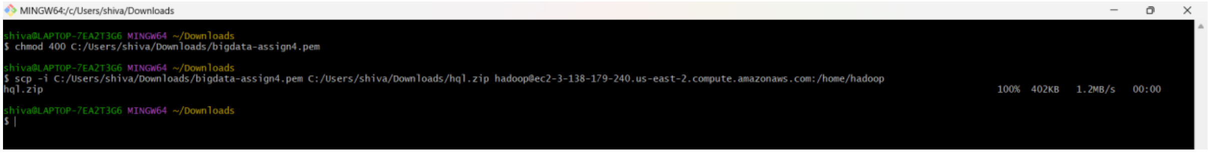


```
shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ chmod 400 C:/Users/shiva/Downloads/bigdata-assign4.pem

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ ssh -i C:/Users/shiva/Downloads/bigdata-assign4.pem hadoop@ec2-3-138-179-240.u
s-east-2.compute.amazonaws.com
The authenticity of host 'ec2-3-138-179-240.us-east-2.compute.amazonaws.com (3.1
38.179.240)' can't be established.
ED25519 key fingerprint is SHA256:SVedkE6rOq4i4ABef4/hYt8J6uoa5mOSamvA4ojEZFw.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-138-179-240.us-east-2.compute.amazonaws.com' (
ED25519) to the list of known hosts.
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

4 package(s) needed for security, out of 9 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-1-178 ~]$
```

Download the file TestDataGen.class from the Blackboard. It is one of the attachments to the assignment.

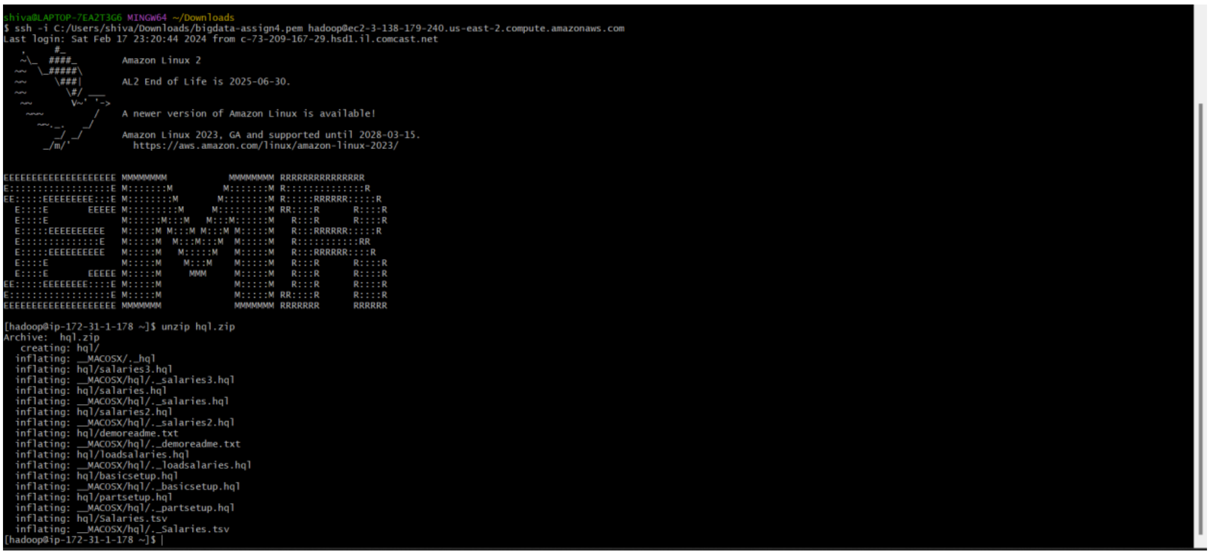
I have downloaded the required files from the blackboard

scp the file over to the home directory (/home/hadoop) on your Hadoop VM

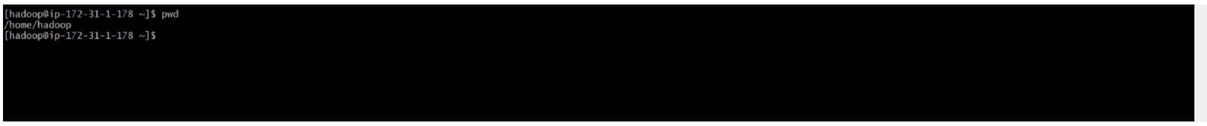


```
shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ chmod 400 C:/Users/shiva/Downloads/bigdata-assign4.pem

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ scp -i C:/Users/shiva/Downloads/bigdata-assign4.pem C:/Users/shiva/Downloads/hql.zip hadoop@ec2-3-138-179-240.us-east-2.compute.amazonaws.com:/home/hadoop
hql.zip                                                                                                    100%  402KB   1.2MB/s   00:00

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ |
```

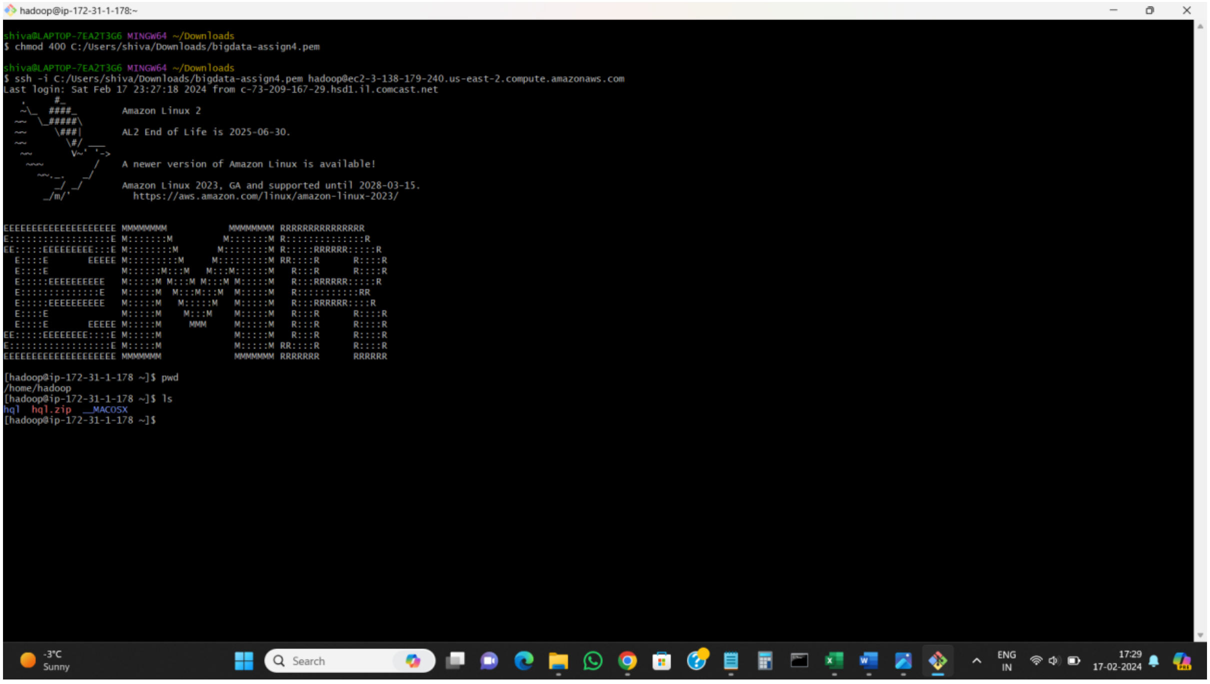
Log on to your VM using ssh and execute the file using “java TestDataGen”



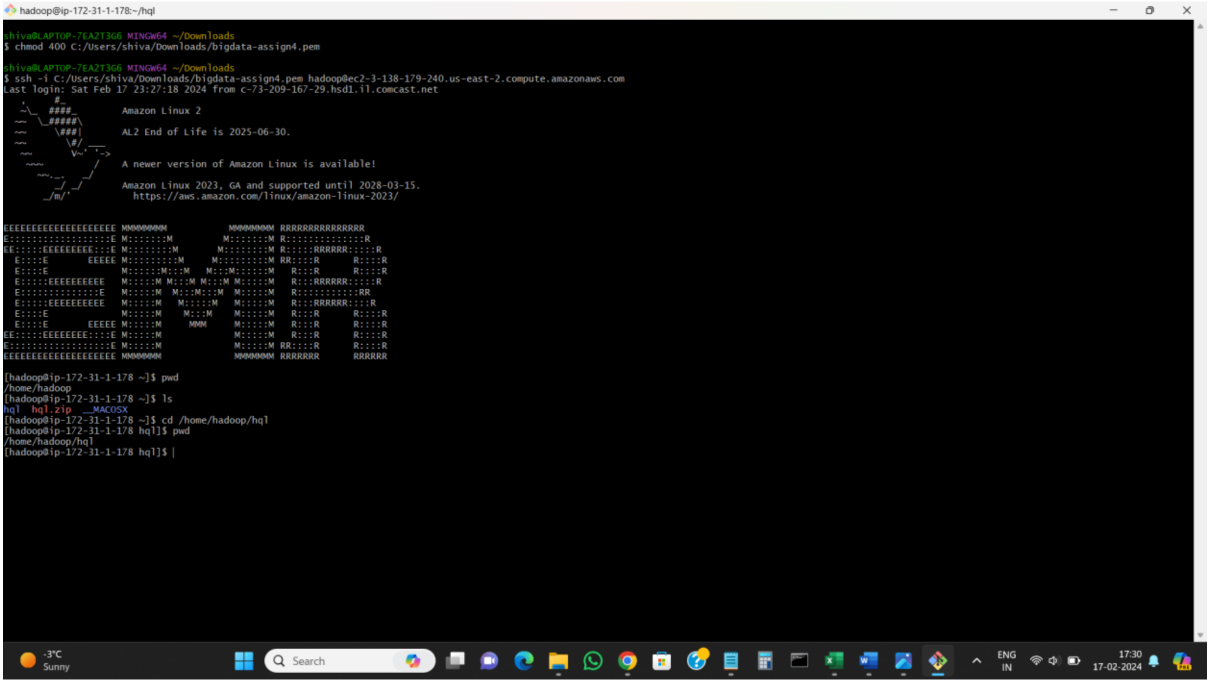
```
shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ ssh -i C:/Users/shiva/Downloads/bigdata-assign4.pem hadoop@ec2-3-138-179-240.us-east-2.compute.amazonaws.com
Last login: Sat Feb 17 23:20:44 2024 from c-73-209-167-29.hsd1.il.comcast.net
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/


EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-1-178 ~]$ unzip hql.zip
Archive:  hql.zip
   creating: hql/
  inflating: __MACOSX/._hql
  inflating: hql/salaries3.hql
  inflating: __MACOSX/hql/._salaries3.hql
  inflating: hql/salaries.hql
  inflating: __MACOSX/hql/._salaries.hql
  inflating: hql/salaries2.hql
  inflating: __MACOSX/hql/._salaries2.hql
  inflating: hql/demoreadme.txt
  inflating: __MACOSX/hql/._demoreadme.txt
  inflating: hql/loadsalaries.hql
  inflating: __MACOSX/hql/._loadsalaries.hql
  inflating: hql/basicsetup.hql
  inflating: __MACOSX/hql/._basicsetup.hql
  inflating: hql/partsetup.hql
  inflating: __MACOSX/hql/._partsetup.hql
  inflating: hql/Salaries.tsv
  inflating: __MACOSX/hql/._Salaries.tsv
[hadoop@ip-172-31-1-178 ~]$ |
```



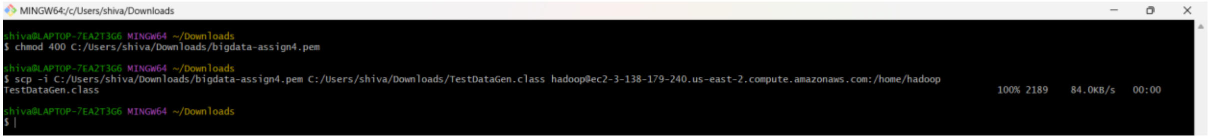
```
[hadoop@ip-172-31-1-178 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-1-178 ~]$
```



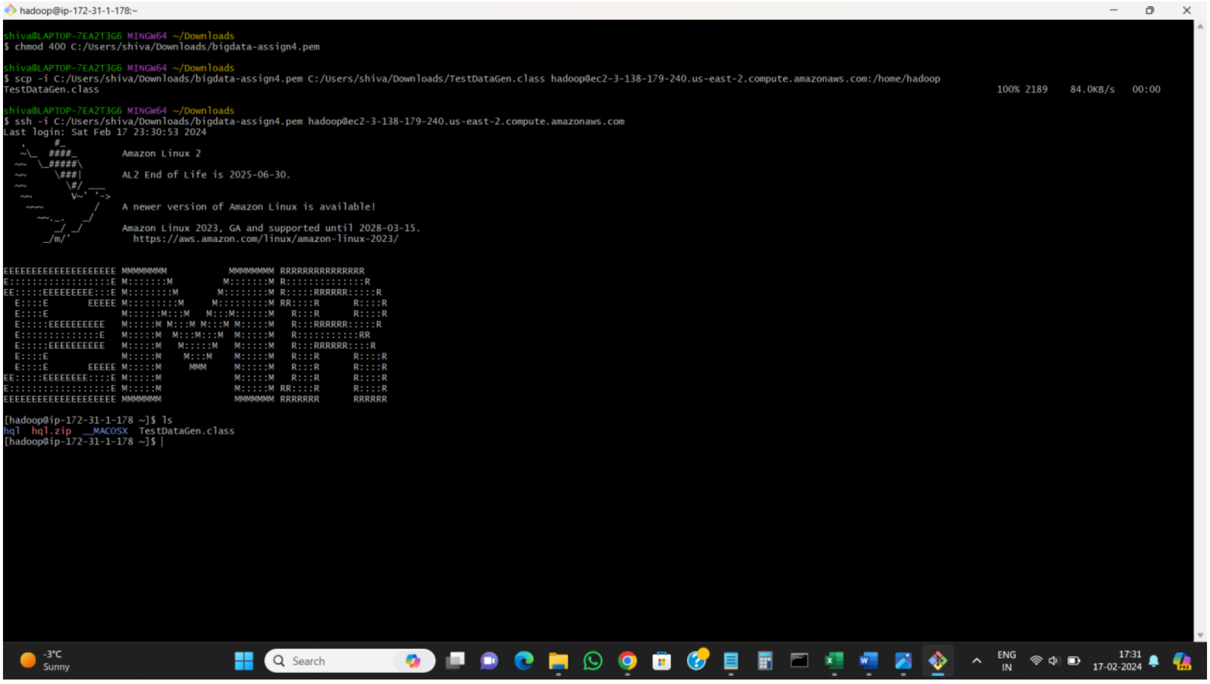
```
shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ chmod 400 C:/Users/shiva/Downloads/bigdata-assign4.pem

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ ssh -i C:/Users/shiva/Downloads/bigdata-assign4.pem hadoop@ec2-3-138-179-240.us-east-2.compute.amazonaws.com
Last login: Sat Feb 17 23:27:18 2024 from c-73-209-167-29.hsd1.il.comcast.net
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/


EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-1-178 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-1-178 ~]$ ls
hql  hql.zip  __MACOSX
[hadoop@ip-172-31-1-178 ~]$
```



```
shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ chmod 400 C:/Users/shiva/Downloads/bigdata-assign4.pem

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ ssh -i C:/Users/shiva/Downloads/bigdata-assign4.pem hadoop@ec2-3-138-179-240.us-east-2.compute.amazonaws.com
Last login: Sat Feb 17 23:27:18 2024 from c-73-209-167-29.hsd1.il.comcast.net
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-1-178 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-1-178 ~]$ ls
hql  hql.zip  __MACOSX
[hadoop@ip-172-31-1-178 ~]$ cd /home/hadoop/hql
[hadoop@ip-172-31-1-178 hql]$ pwd
/home/hadoop/hql
[hadoop@ip-172-31-1-178 hql]$ |
```

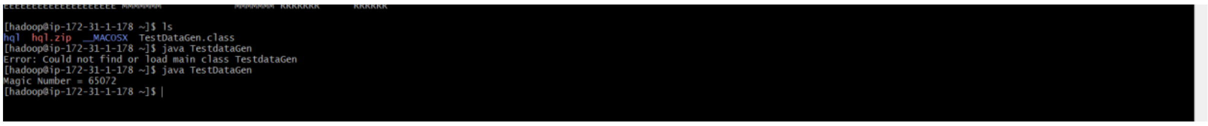


```
shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ chmod 400 C:/Users/shiva/Downloads/bigdata-assign4.pem

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ scp -i C:/Users/shiva/Downloads/bigdata-assign4.pem C:/Users/shiva/Downloads/TestDataGen.class hadoop@ec2-3-138-179-240.us-east-2.compute.amazonaws.com:/home/hadoop
TestDataGen.class                                                                                                                          100% 2189    84.0KB/s   00:00

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ |
```



```
shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ chmod 400 C:/Users/shiva/Downloads/bigdata-assign4.pem

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ scp -i C:/Users/shiva/Downloads/bigdata-assign4.pem C:/Users/shiva/Downloads/TestDataGen.class hadoop@ec2-3-138-179-240.us-east-2.compute.amazonaws.com:/home/hadoop
TestDataGen.class                                                                                                                          100% 2189    84.0KB/s   00:00

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ ssh -i C:/Users/shiva/Downloads/bigdata-assign4.pem hadoop@ec2-3-138-179-240.us-east-2.compute.amazonaws.com
Last login: Sat Feb 17 23:30:53 2024
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-1-178 ~]$ ls
hql  hql.zip  __MACOSX  TestDataGen.class
[hadoop@ip-172-31-1-178 ~]$ |
```

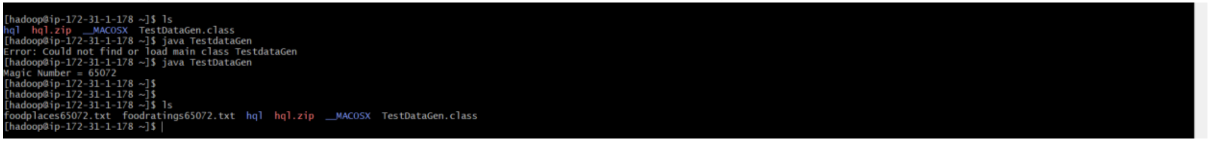
This will output a magic number which you should copy down and provide with the results of your assignment.



```
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-1-178 ~]$ ls
hql  hql.zip  __MACOSX  TestDataGen.class
[hadoop@ip-172-31-1-178 ~]$ java TestdataGen
Error: Could not find or load main class TestdataGen
[hadoop@ip-172-31-1-178 ~]$ java TestDataGen
Magic Number = 65072
[hadoop@ip-172-31-1-178 ~]$ |
```

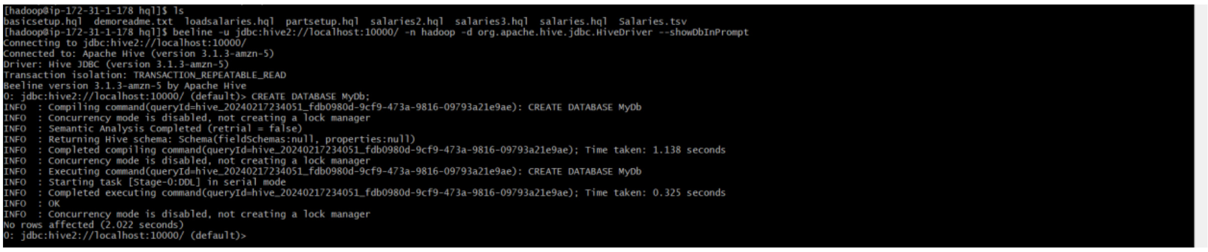
Magic Number = 65072

When we execute TestDataGen.class file, foodplace<magicnumber>.txt and foodratings<magicnumber>.txt are automatically created.



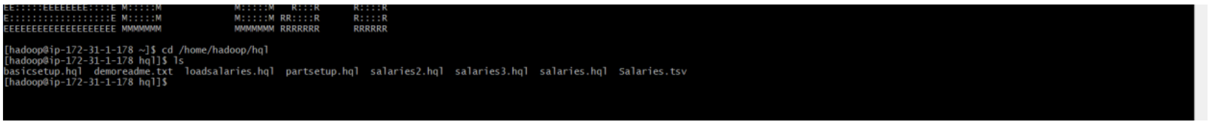
```
[hadoop@ip-172-31-1-178 ~]$ ls
hql  hql.zip  __MACOSX  TestDataGen.class
[hadoop@ip-172-31-1-178 ~]$ java TestdataGen
Error: Could not find or load main class TestdataGen
[hadoop@ip-172-31-1-178 ~]$ java TestDataGen
Magic Number = 65072
[hadoop@ip-172-31-1-178 ~]$
[hadoop@ip-172-31-1-178 ~]$
[hadoop@ip-172-31-1-178 ~]$ ls
foodplaces65072.txt  foodratings65072.txt  hql  hql.zip  __MACOSX  TestDataGen.class
[hadoop@ip-172-31-1-178 ~]$ |
```

Starting the Hive beeline



```
[hadoop@ip-172-31-1-178 hql]$ ls
basicsetup.hql  demoreadme.txt  loadsalaries.hql  partsetup.hql  salaries2.hql  salaries3.hql  salaries.hql  Salaries.tsv
[hadoop@ip-172-31-1-178 hql]$ beeline -u jdbc:hive2://localhost:10000/ -n hadoop -d org.apache.hive.jdbc.HiveDriver --showDbInPrompt
Connecting to jdbc:hive2://localhost:10000/
Connected to: Apache Hive (version 3.1.3-amzn-5)
Driver: Hive JDBC (version 3.1.3-amzn-5)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.3-amzn-5 by Apache Hive
0: jdbc:hive2://localhost:10000/ (default)> CREATE DATABASE MyDb;
INFO  : Compiling command(queryId=hive_20240217234051_fdb0980d-9cf9-473a-9816-09793a21e9ae): CREATE DATABASE MyDb
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20240217234051_fdb0980d-9cf9-473a-9816-09793a21e9ae); Time taken: 1.138 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240217234051_fdb0980d-9cf9-473a-9816-09793a21e9ae): CREATE DATABASE MyDb
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240217234051_fdb0980d-9cf9-473a-9816-09793a21e9ae); Time taken: 0.325 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (2.022 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

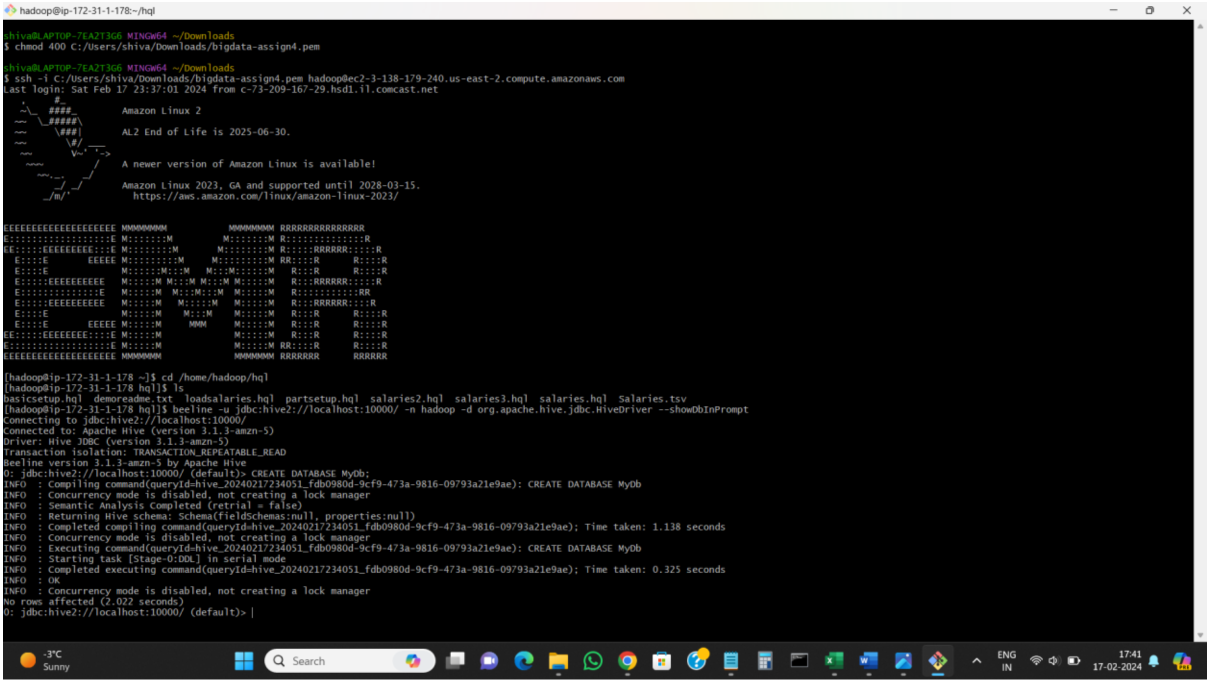
Files in Hadoop beeline



```
EE:::::EEEEEEEE::::E M:::::M            M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M            M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM            MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-1-178 ~]$ cd /home/hadoop/hql
[hadoop@ip-172-31-1-178 hql]$ ls
basicsetup.hql  demoreadme.txt  loadsalaries.hql  partsetup.hql  salaries2.hql  salaries3.hql  salaries.hql  Salaries.tsv
[hadoop@ip-172-31-1-178 hql]$
```

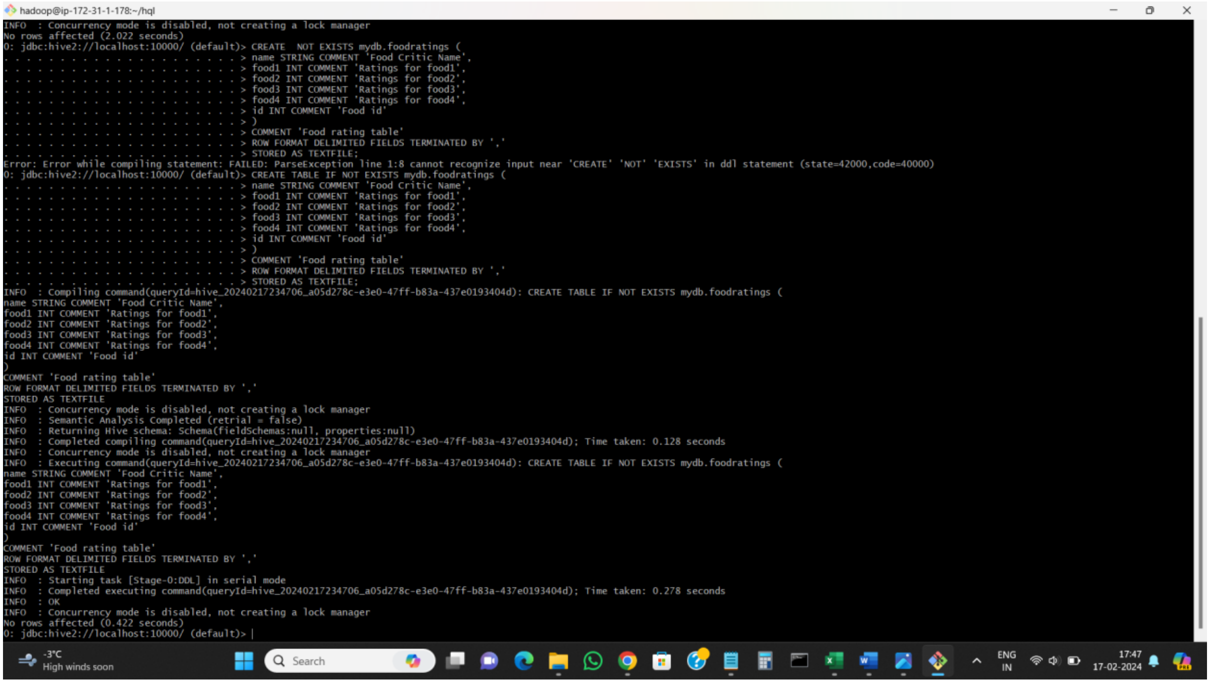
Exercise 1

Create a Hive database called “MyDb”.



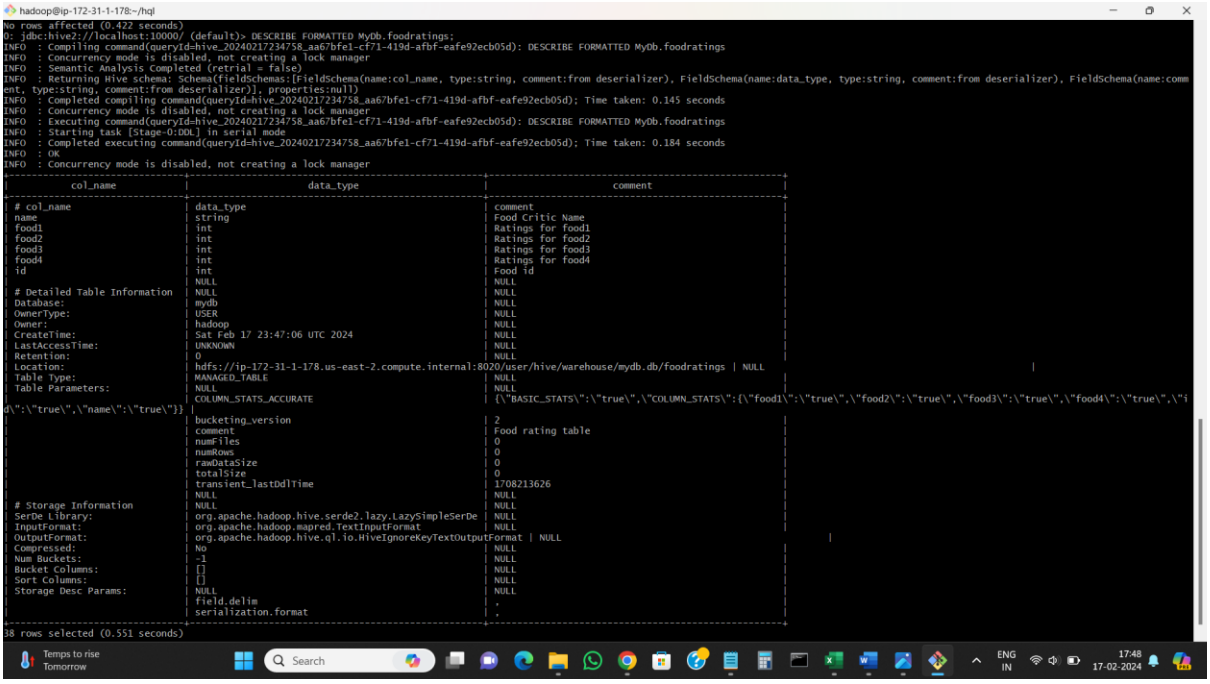
```
shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ chmod 400 C:/Users/shiva/Downloads/bigdata-assign4.pem

shiva@LAPTOP-7EA2T3G6 MINGW64 ~/Downloads
$ ssh -i C:/Users/shiva/Downloads/bigdata-assign4.pem hadoop@ec2-3-138-179-240.us-east-2.compute.amazonaws.com
Last login: Sat Feb 17 23:37:01 2024 from c-73-209-167-29.hsd1.il.comcast.net
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-1-178 ~]$ cd /home/hadoop/hql
[hadoop@ip-172-31-1-178 hql]$ ls
basicsetup.hql  demoreadme.txt  loadsalaries.hql  partsetup.hql  salaries2.hql  salaries3.hql  salaries.hql  Salaries.tsv
[hadoop@ip-172-31-1-178 hql]$ beeline -u jdbc:hive2://localhost:10000/ -n hadoop -d org.apache.hive.jdbc.HiveDriver --showDbInPrompt
Connecting to jdbc:hive2://localhost:10000/
Connected to: Apache Hive (version 3.1.3-amzn-5)
Driver: Hive JDBC (version 3.1.3-amzn-5)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.3-amzn-5 by Apache Hive
0: jdbc:hive2://localhost:10000/ (default)> CREATE DATABASE MyDb;
INFO  : Compiling command(queryId=hive_20240217234051_fdb0980d-9cf9-473a-9816-09793a21e9ae): CREATE DATABASE MyDb
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20240217234051_fdb0980d-9cf9-473a-9816-09793a21e9ae); Time taken: 1.138 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240217234051_fdb0980d-9cf9-473a-9816-09793a21e9ae): CREATE DATABASE MyDb
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240217234051_fdb0980d-9cf9-473a-9816-09793a21e9ae); Time taken: 0.325 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (2.022 seconds)
0: jdbc:hive2://localhost:10000/ (default)> |
```

Now in MyDb create a table with name foodratings having six columns with the name of the first 'name' and the type of the first a string and the names of the remaining columns food1, food2, food3, food4 and id and indicate their types each as an integer. The table should have storage format TEXTFILE and column separator a ",". That is the underlying format should be a CSV file. The table itself and each column should include a comment just to show me you know how to use comments (it does not matter what it says).



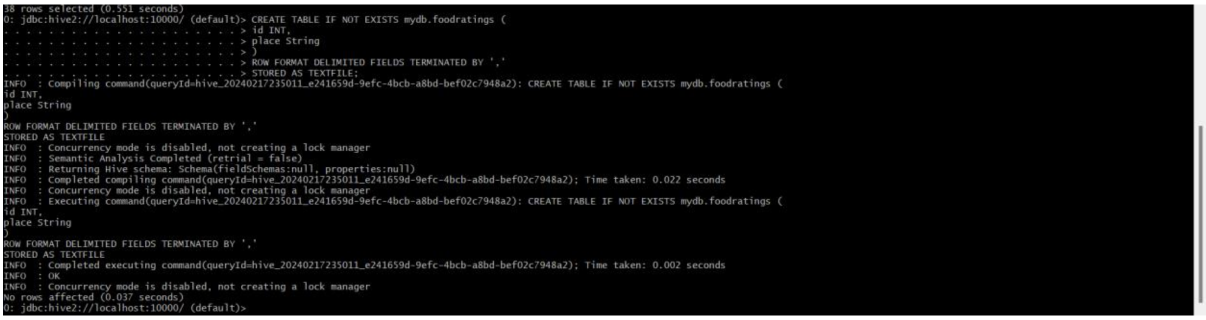
```
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (2.022 seconds)
0: jdbc:hive2://localhost:10000/ (default)> CREATE  NOT EXISTS mydb.foodratings (
. . . . . . . . . . . . . . . . . . . . . . > name STRING COMMENT 'Food Critic Name',
. . . . . . . . . . . . . . . . . . . . . . > food1 INT COMMENT 'Ratings for food1',
. . . . . . . . . . . . . . . . . . . . . . > food2 INT COMMENT 'Ratings for food2',
. . . . . . . . . . . . . . . . . . . . . . > food3 INT COMMENT 'Ratings for food3',
. . . . . . . . . . . . . . . . . . . . . . > food4 INT COMMENT 'Ratings for food4',
. . . . . . . . . . . . . . . . . . . . . . > id INT COMMENT 'Food id'
. . . . . . . . . . . . . . . . . . . . . . > )
. . . . . . . . . . . . . . . . . . . . . . > COMMENT 'Food rating table'
. . . . . . . . . . . . . . . . . . . . . . > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
. . . . . . . . . . . . . . . . . . . . . . > STORED AS TEXTFILE;
Error: Error while compiling statement: FAILED: ParseException line 1:8 cannot recognize input near 'CREATE' 'NOT' 'EXISTS' in ddl statement (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (default)> CREATE TABLE IF NOT EXISTS mydb.foodratings (
. . . . . . . . . . . . . . . . . . . . . . > name STRING COMMENT 'Food Critic Name',
. . . . . . . . . . . . . . . . . . . . . . > food1 INT COMMENT 'Ratings for food1',
. . . . . . . . . . . . . . . . . . . . . . > food2 INT COMMENT 'Ratings for food2',
. . . . . . . . . . . . . . . . . . . . . . > food3 INT COMMENT 'Ratings for food3',
. . . . . . . . . . . . . . . . . . . . . . > food4 INT COMMENT 'Ratings for food4',
. . . . . . . . . . . . . . . . . . . . . . > id INT COMMENT 'Food id'
. . . . . . . . . . . . . . . . . . . . . . > )
. . . . . . . . . . . . . . . . . . . . . . > COMMENT 'Food rating table'
. . . . . . . . . . . . . . . . . . . . . . > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
. . . . . . . . . . . . . . . . . . . . . . > STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20240217234706_a05d278c-e3e0-47ff-b83a-437e0193404d): CREATE TABLE IF NOT EXISTS mydb.foodratings (
name STRING COMMENT 'Food Critic Name',
food1 INT COMMENT 'Ratings for food1',
food2 INT COMMENT 'Ratings for food2',
food3 INT COMMENT 'Ratings for food3',
food4 INT COMMENT 'Ratings for food4',
id INT COMMENT 'Food id'
)
COMMENT 'Food rating table'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20240217234706_a05d278c-e3e0-47ff-b83a-437e0193404d); Time taken: 0.128 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240217234706_a05d278c-e3e0-47ff-b83a-437e0193404d): CREATE TABLE IF NOT EXISTS mydb.foodratings (
name STRING COMMENT 'Food Critic Name',
food1 INT COMMENT 'Ratings for food1',
food2 INT COMMENT 'Ratings for food2',
food3 INT COMMENT 'Ratings for food3',
food4 INT COMMENT 'Ratings for food4',
id INT COMMENT 'Food id'
)
COMMENT 'Food rating table'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240217234706_a05d278c-e3e0-47ff-b83a-437e0193404d); Time taken: 0.278 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.422 seconds)
0: jdbc:hive2://localhost:10000/ (default)> |
```

Execute a Hive command of 'DESCRIBE FORMATTED MyDb.foodratings;' and capture its output as one of the results of this exercise.



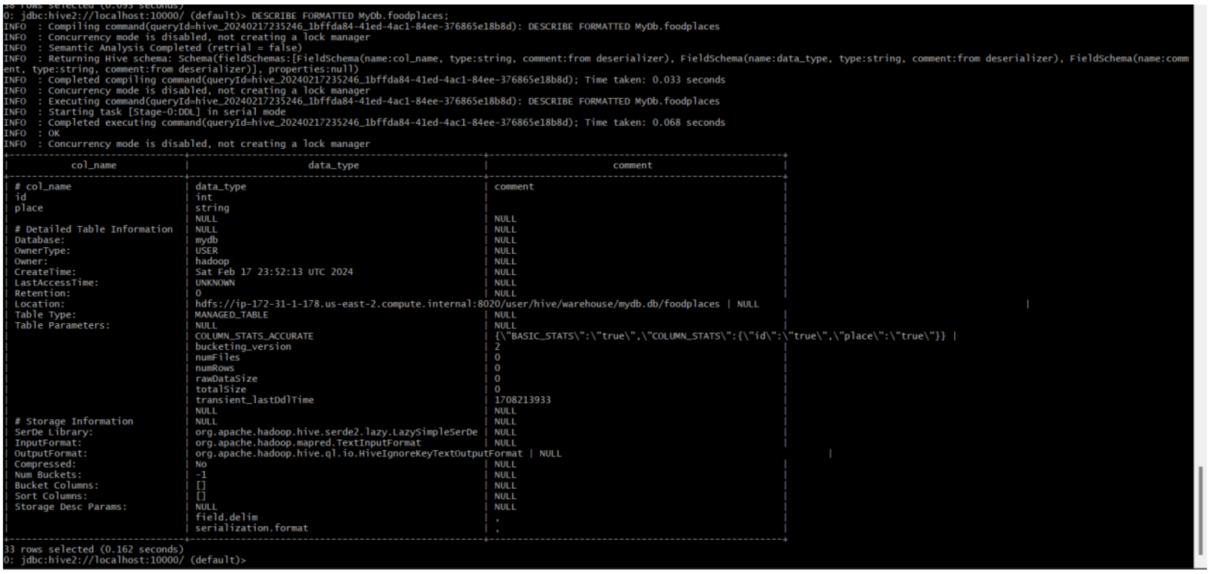
```
No rows affected (0.422 seconds)
0: jdbc:hive2://localhost:10000/ (default)> DESCRIBE FORMATTED MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20240217234758_aa67bfe1-cf71-419d-afbf-eafe92ecb05d): DESCRIBE FORMATTED MyDb.foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name:comm
ent, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240217234758_aa67bfe1-cf71-419d-afbf-eafe92ecb05d); Time taken: 0.145 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240217234758_aa67bfe1-cf71-419d-afbf-eafe92ecb05d): DESCRIBE FORMATTED MyDb.foodratings
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240217234758_aa67bfe1-cf71-419d-afbf-eafe92ecb05d); Time taken: 0.184 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------------------+----------------------------------------------------+----------------------------------------------------+
|           col_name            |                     data_type                      |                      comment                       |
+-------------------------------+----------------------------------------------------+----------------------------------------------------+
| # col_name                    | data_type                                          | comment                                            |
| name                          | string                                             | Food Critic Name                                   |
| food1                         | int                                                | Ratings for food1                                  |
| food2                         | int                                                | Ratings for food2                                  |
| food3                         | int                                                | Ratings for food3                                  |
| food4                         | int                                                | Ratings for food4                                  |
| id                            | int                                                | Food id                                            |
|                               | NULL                                               | NULL                                               |
| # Detailed Table Information  | NULL                                               | NULL                                               |
| Database:                     | mydb                                               | NULL                                               |
| OwnerType:                    | USER                                               | NULL                                               |
| Owner:                        | hadoop                                             | NULL                                               |
| CreateTime:                   | Sat Feb 17 23:47:06 UTC 2024                       | NULL                                               |
| LastAccessTime:               | UNKNOWN                                            | NULL                                               |
| Retention:                    | 0                                                  | NULL                                               |
| Location:                     | hdfs://ip-172-31-1-178.us-east-2.compute.internal:8020/user/hive/warehouse/mydb.db/foodratings | NULL                       |
| Table Type:                   | MANAGED_TABLE                                      | NULL                                               |
| Table Parameters:             | NULL                                               | NULL                                               |
|                               | COLUMN_STATS_ACCURATE                              | {\"BASIC_STATS\":\"true\",\"COLUMN_STATS\":{\"food1\":\"true\",\"food2\":\"true\",\"food3\":\"true\",\"food4\":\"true\",\"i
d\":\"true\",\"name\":\"true\"}} |
|                               | bucketing_version                                  | 2                                                  |
|                               | comment                                            | Food rating table                                  |
|                               | numFiles                                           | 0                                                  |
|                               | numRows                                            | 0                                                  |
|                               | rawDataSize                                        | 0                                                  |
|                               | totalSize                                          | 0                                                  |
|                               | transient_lastDdlTime                              | 1708213626                                         |
|                               | NULL                                               | NULL                                               |
| # Storage Information         | NULL                                               | NULL                                               |
| SerDe Library:                | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL                                               |
| InputFormat:                  | org.apache.hadoop.mapred.TextInputFormat           | NULL                                               |
| OutputFormat:                 | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat | NULL                                       |
| Compressed:                   | No                                                 | NULL                                               |
| Num Buckets:                  | -1                                                 | NULL                                               |
| Bucket Columns:               | []                                                 | NULL                                               |
| Sort Columns:                 | []                                                 | NULL                                               |
| Storage Desc Params:          | NULL                                               | NULL                                               |
|                               | field.delim                                        | ,                                                  |
|                               | serialization.format                               | ,                                                  |
+-------------------------------+----------------------------------------------------+----------------------------------------------------+
38 rows selected (0.551 seconds)
```

Then in MyDb create a table with name foodplaces having two columns with first called 'id' with the type of the first an integer, and the second column called 'place' with the type of the second a string. This table should also have storage format TEXTFILE and column separator a ",". That is the underlying format should be a CSV file. No comments are needed for this table.



```
38 rows selected (0.551 seconds)
0: jdbc:hive2://localhost:10000/ (default)> CREATE TABLE IF NOT EXISTS mydb.foodratings (
. . . . . . . . . . . . . . . . . . . . . . > id INT,
. . . . . . . . . . . . . . . . . . . . . . > place String
. . . . . . . . . . . . . . . . . . . . . . > )
. . . . . . . . . . . . . . . . . . . . . . > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
. . . . . . . . . . . . . . . . . . . . . . > STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20240217235011_e241659d-9efc-4bcb-a8bd-bef02c7948a2): CREATE TABLE IF NOT EXISTS mydb.foodratings (
id INT,
place String
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20240217235011_e241659d-9efc-4bcb-a8bd-bef02c7948a2); Time taken: 0.022 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240217235011_e241659d-9efc-4bcb-a8bd-bef02c7948a2): CREATE TABLE IF NOT EXISTS mydb.foodratings (
id INT,
place String
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Completed executing command(queryId=hive_20240217235011_e241659d-9efc-4bcb-a8bd-bef02c7948a2); Time taken: 0.002 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.037 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

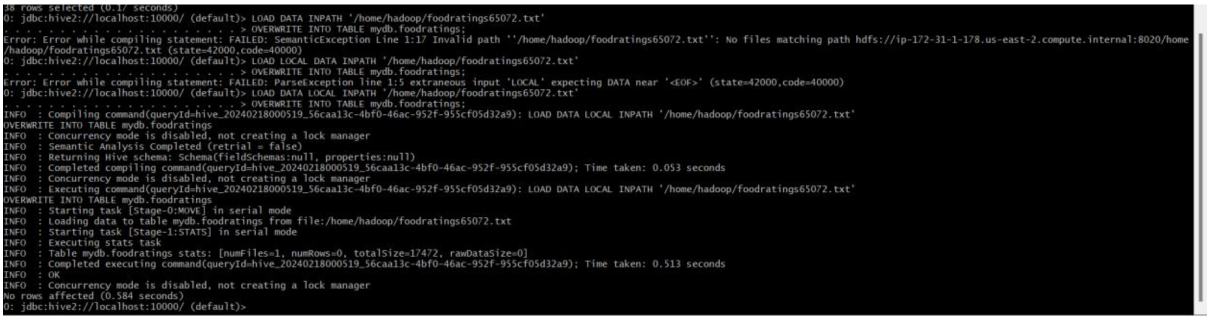
Execute a Hive command of 'DESCRIBE FORMATTED MyDb.foodplaces' and capture its output as another of the results of this exercise.



```
38 rows selected (0.093 seconds)
0: jdbc:hive2://localhost:10000/ (default)> DESCRIBE FORMATTED MyDb.foodplaces;
INFO  : Compiling command(queryId=hive_20240217235246_1bffda84-41ed-4ac1-84ee-376865e18b8d): DESCRIBE FORMATTED MyDb.foodplaces
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name:comm
ent, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240217235246_1bffda84-41ed-4ac1-84ee-376865e18b8d); Time taken: 0.033 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240217235246_1bffda84-41ed-4ac1-84ee-376865e18b8d): DESCRIBE FORMATTED MyDb.foodplaces
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240217235246_1bffda84-41ed-4ac1-84ee-376865e18b8d); Time taken: 0.068 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------------------+----------------------------------------------------+----------------------------------------------------+
|           col_name            |                     data_type                      |                      comment                       |
+-------------------------------+----------------------------------------------------+----------------------------------------------------+
| # col_name                    | data_type                                          | comment                                            |
| id                            | int                                                |                                                    |
| place                         | string                                             |                                                    |
|                               | NULL                                               | NULL                                               |
| # Detailed Table Information  | NULL                                               | NULL                                               |
| Database:                     | mydb                                               | NULL                                               |
| OwnerType:                    | USER                                               | NULL                                               |
| Owner:                        | hadoop                                             | NULL                                               |
| CreateTime:                   | Sat Feb 17 23:52:13 UTC 2024                       | NULL                                               |
| LastAccessTime:               | UNKNOWN                                            | NULL                                               |
| Retention:                    | 0                                                  | NULL                                               |
| Location:                     | hdfs://ip-172-31-1-178.us-east-2.compute.internal:8020/user/hive/warehouse/mydb.db/foodplaces | NULL                    |
| Table Type:                   | MANAGED_TABLE                                      | NULL                                               |
| Table Parameters:             | NULL                                               | NULL                                               |
|                               | COLUMN_STATS_ACCURATE                              | {\"BASIC_STATS\":\"true\",\"COLUMN_STATS\":{\"id\":\"true\",\"place\":\"true\"}} |
|                               | bucketing_version                                  | 2                                                  |
|                               | numFiles                                           | 0                                                  |
|                               | numRows                                            | 0                                                  |
|                               | rawDataSize                                        | 0                                                  |
|                               | totalSize                                          | 0                                                  |
|                               | transient_lastDdlTime                              | 1708213933                                         |
|                               | NULL                                               | NULL                                               |
| # Storage Information         | NULL                                               | NULL                                               |
| SerDe Library:                | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL                                               |
| InputFormat:                  | org.apache.hadoop.mapred.TextInputFormat           | NULL                                               |
| OutputFormat:                 | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat | NULL                                       |
| Compressed:                   | No                                                 | NULL                                               |
| Num Buckets:                  | -1                                                 | NULL                                               |
| Bucket Columns:               | []                                                 | NULL                                               |
| Sort Columns:                 | []                                                 | NULL                                               |
| Storage Desc Params:          | NULL                                               | NULL                                               |
|                               | field.delim                                        | ,                                                  |
|                               | serialization.format                               | ,                                                  |
+-------------------------------+----------------------------------------------------+----------------------------------------------------+
33 rows selected (0.162 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

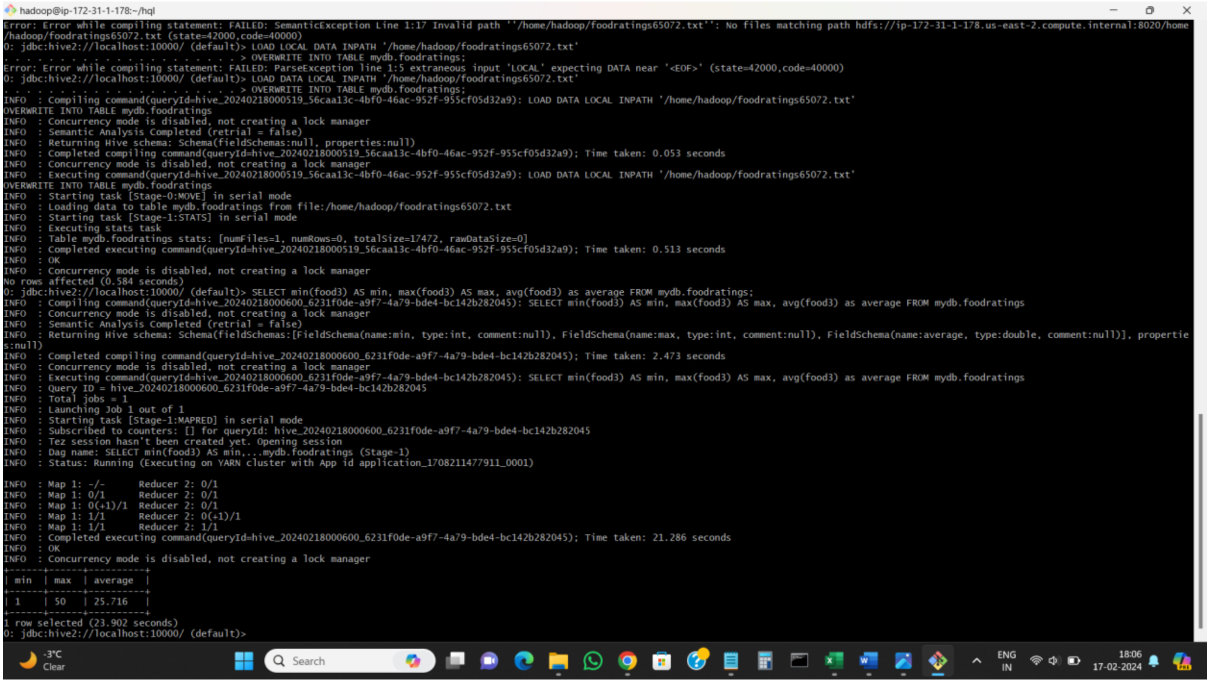
Exercise 2

Load the foodratings<magic number>.txt file created using TestDataGen from your local file system into the foodratings table.



```
38 rows selected (0.17 seconds)
0: jdbc:hive2://localhost:10000/ (default)> LOAD DATA INPATH '/home/hadoop/foodratings65072.txt'
. . . . . . . . . . . . . . . . . . . . . . > OVERWRITE INTO TABLE mydb.foodratings;
Error: Error while compiling statement: FAILED: SemanticException Line 1:17 Invalid path ''/home/hadoop/foodratings65072.txt'': No files matching path hdfs://ip-172-31-1-178.us-east-2.compute.internal:8020/home
/hadoop/foodratings65072.txt (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (default)> LOAD LOCAL DATA INPATH '/home/hadoop/foodratings65072.txt'
. . . . . . . . . . . . . . . . . . . . . . > OVERWRITE INTO TABLE mydb.foodratings;
Error: Error while compiling statement: FAILED: ParseException line 1:5 extraneous input 'LOCAL' expecting DATA near '<EOF>' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (default)> LOAD DATA LOCAL INPATH '/home/hadoop/foodratings65072.txt'
. . . . . . . . . . . . . . . . . . . . . . > OVERWRITE INTO TABLE mydb.foodratings;
INFO  : Compiling command(queryId=hive_20240218000519_56caa13c-4bf0-46ac-952f-955cf05d32a9): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings65072.txt'
OVERWRITE INTO TABLE mydb.foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20240218000519_56caa13c-4bf0-46ac-952f-955cf05d32a9); Time taken: 0.053 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218000519_56caa13c-4bf0-46ac-952f-955cf05d32a9): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings65072.txt'
OVERWRITE INTO TABLE mydb.foodratings
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodratings from file:/home/hadoop/foodratings65072.txt
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Executing stats task
INFO  : Table mydb.foodratings stats: [numFiles=1, numRows=0, totalSize=17472, rawDataSize=0]
INFO  : Completed executing command(queryId=hive_20240218000519_56caa13c-4bf0-46ac-952f-955cf05d32a9); Time taken: 0.513 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.584 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

Execute a single hive command to output the min, max and average of the values of the food3 column of the foodratings table. This should be one hive command, not three separate ones. A copy of the hive command you wrote, the output of this query and the magic number are the result of this exercise.



```
Error: Error while compiling statement: FAILED: SemanticException Line 1:17 Invalid path ''/home/hadoop/foodratings65072.txt'': No files matching path hdfs://ip-172-31-1-178.us-east-2.compute.internal:8020/home
/hadoop/foodratings65072.txt (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (default)> LOAD LOCAL DATA INPATH '/home/hadoop/foodratings65072.txt'
. . . . . . . . . . . . . . . . . . . . . . .> OVERWRITE INTO TABLE mydb.foodratings;
Error: Error while compiling statement: FAILED: ParseException line 1:5 extraneous input 'LOCAL' expecting DATA near '<EOF>' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (default)> LOAD DATA LOCAL INPATH '/home/hadoop/foodratings65072.txt'
. . . . . . . . . . . . . . . . . . . . . . .> OVERWRITE INTO TABLE mydb.foodratings;
INFO  : Compiling command(queryId=hive_20240218000519_56caa13c-4bf0-46ac-952f-955cf05d32a9): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings65072.txt'
OVERWRITE INTO TABLE mydb.foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20240218000519_56caa13c-4bf0-46ac-952f-955cf05d32a9); Time taken: 0.053 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218000519_56caa13c-4bf0-46ac-952f-955cf05d32a9): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings65072.txt'
OVERWRITE INTO TABLE mydb.foodratings
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodratings from file:/home/hadoop/foodratings65072.txt
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Executing stats task
INFO  : Table mydb.foodratings stats: [numFiles=1, numRows=0, totalSize=17472, rawDataSize=0]
INFO  : Completed executing command(queryId=hive_20240218000519_56caa13c-4bf0-46ac-952f-955cf05d32a9); Time taken: 0.513 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.584 seconds)
0: jdbc:hive2://localhost:10000/ (default)> SELECT min(food3) AS min, max(food3) AS max, avg(food3) as average FROM mydb.foodratings;
INFO  : Compiling command(queryId=hive_20240218000600_6231f0de-a9f7-4a79-bde4-bc142b282045): SELECT min(food3) AS min, max(food3) AS max, avg(food3) as average FROM mydb.foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:min, type:int, comment:null), FieldSchema(name:max, type:int, comment:null), FieldSchema(name:average, type:double, comment:null)], propertie
s:null)
INFO  : Completed compiling command(queryId=hive_20240218000600_6231f0de-a9f7-4a79-bde4-bc142b282045); Time taken: 2.473 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218000600_6231f0de-a9f7-4a79-bde4-bc142b282045): SELECT min(food3) AS min, max(food3) AS max, avg(food3) as average FROM mydb.foodratings
INFO  : Query ID = hive_20240218000600_6231f0de-a9f7-4a79-bde4-bc142b282045
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240218000600_6231f0de-a9f7-4a79-bde4-bc142b282045
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT min(food3) AS min,...mydb.foodratings (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1708211477911_0001)

INFO  : Map 1: -/-      Reducer 2: 0/1
INFO  : Map 1: 0/1      Reducer 2: 0/1
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/1
INFO  : Map 1: 1/1      Reducer 2: 0(+1)/1
INFO  : Map 1: 1/1      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20240218000600_6231f0de-a9f7-4a79-bde4-bc142b282045); Time taken: 21.286 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------+------+----------+
| min  | max  | average  |
+------+------+----------+
| 1    | 50   | 25.716   |
+------+------+----------+
1 row selected (23.902 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

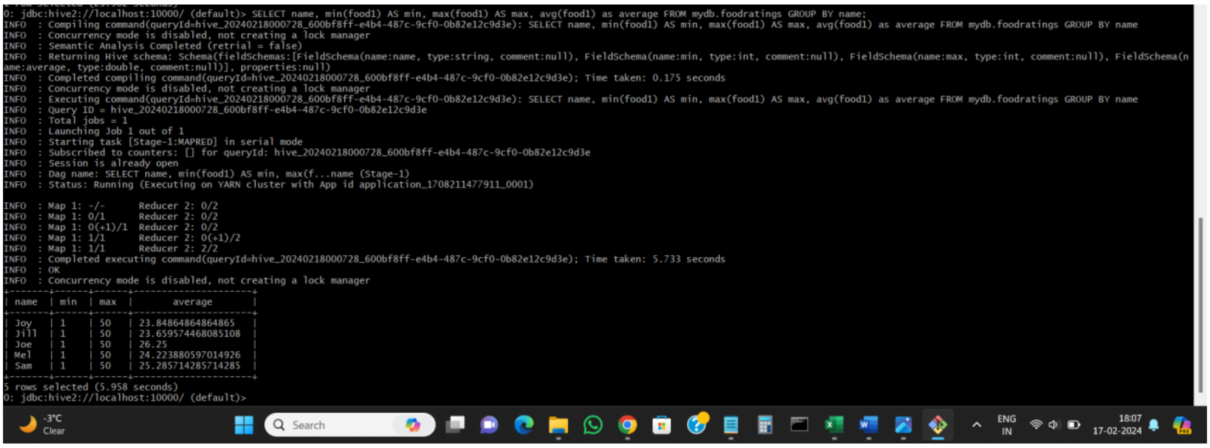
Exercise 3

Execute a hive command to output the min, max and average of the values of the food1 column grouped by the first column ‘name’. This should be one hive command, not three separate ones. The output should look something like:

Mel 10 20 15

Bill 20, 30, 24 …

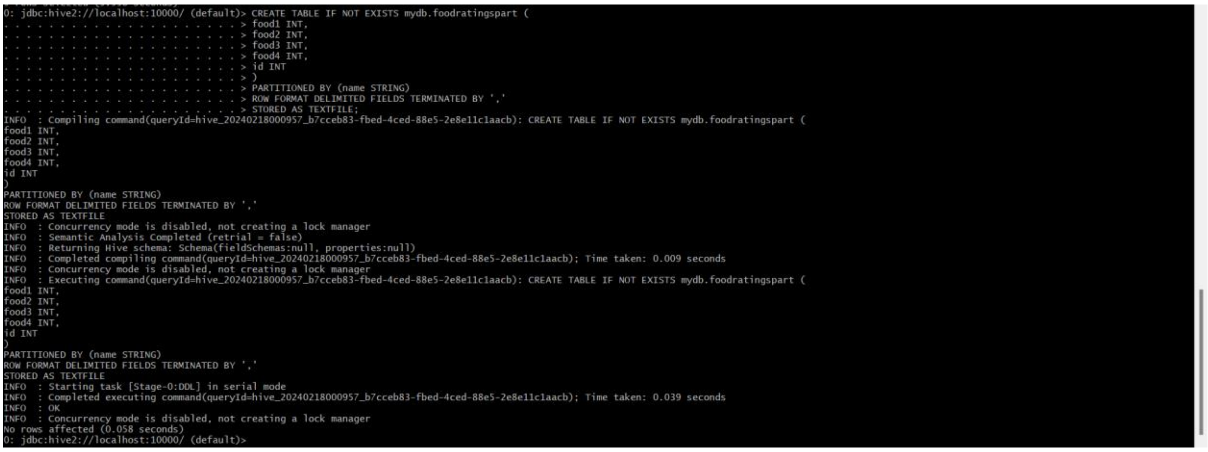
A copy of the hive command you wrote, the output of this query and the magic number are the result of this exercise.



```
0: jdbc:hive2://localhost:10000/ (default)> SELECT name, min(food1) AS min, max(food1) AS max, avg(food1) as average FROM mydb.foodratings GROUP BY name;
INFO  : Compiling command(queryId=hive_20240218000728_600bf8ff-e4b4-487c-9cf0-0b82e12c9d3e): SELECT name, min(food1) AS min, max(food1) AS max, avg(food1) as average FROM mydb.foodratings GROUP BY name
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:name, type:string, comment:null), FieldSchema(name:min, type:int, comment:null), FieldSchema(name:max, type:int, comment:null), FieldSchema(n
ame:average, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240218000728_600bf8ff-e4b4-487c-9cf0-0b82e12c9d3e); Time taken: 0.175 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218000728_600bf8ff-e4b4-487c-9cf0-0b82e12c9d3e): SELECT name, min(food1) AS min, max(food1) AS max, avg(food1) as average FROM mydb.foodratings GROUP BY name
INFO  : Query ID = hive_20240218000728_600bf8ff-e4b4-487c-9cf0-0b82e12c9d3e
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240218000728_600bf8ff-e4b4-487c-9cf0-0b82e12c9d3e
INFO  : Session is already open
INFO  : Dag name: SELECT name, min(food1) AS min, max(f...name (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1708211477911_0001)

INFO  : Map 1: -/-      Reducer 2: 0/2
INFO  : Map 1: 0/1      Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 1/1      Reducer 2: 0(+1)/2
INFO  : Map 1: 1/1      Reducer 2: 2/2
INFO  : Completed executing command(queryId=hive_20240218000728_600bf8ff-e4b4-487c-9cf0-0b82e12c9d3e); Time taken: 5.733 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------+------+------+---------------------+
| name  | min  | max  |       average       |
+-------+------+------+---------------------+
| Joy   | 1    | 50   | 23.84864864864865   |
| Jill  | 1    | 50   | 23.659574468085108  |
| Joe   | 1    | 50   | 26.25               |
| Mel   | 1    | 50   | 24.223880597014926  |
| Sam   | 1    | 50   | 25.285714285714285  |
+-------+------+------+---------------------+
5 rows selected (5.958 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

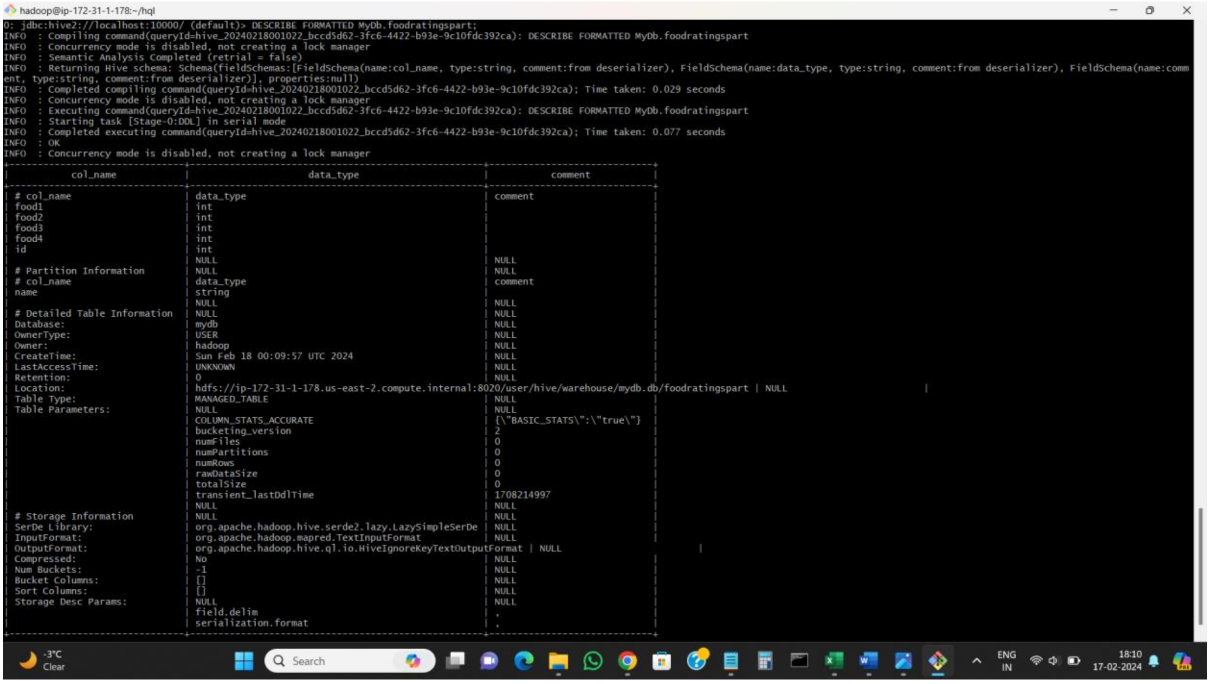
Exercise 4

In MyDb create a partitioned table called ‘foodratingspart’ The partition field should be called ‘name’ and its type should be a string. The names of the non-partition columns should be food1, food2, food3, food4 and id and their types each an integer. The table should have storage format TEXTFILE and column separator a “,”. That is the underlying format should be a CSV file. No comments are needed for this table. Provide the code you have used to create the table as a result of this exercise.



```
0: jdbc:hive2://localhost:10000/ (default)> CREATE TABLE IF NOT EXISTS mydb.foodratingspart (
. . . . . . . . . . . . . . . . . . . . . . > food1 INT,
. . . . . . . . . . . . . . . . . . . . . . > food2 INT,
. . . . . . . . . . . . . . . . . . . . . . > food3 INT,
. . . . . . . . . . . . . . . . . . . . . . > food4 INT,
. . . . . . . . . . . . . . . . . . . . . . > id INT
. . . . . . . . . . . . . . . . . . . . . . > )
. . . . . . . . . . . . . . . . . . . . . . > PARTITIONED BY (name STRING)
. . . . . . . . . . . . . . . . . . . . . . > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
. . . . . . . . . . . . . . . . . . . . . . > STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20240218000957_b7cceb83-fbed-4ced-88e5-2e8e11c1aacb): CREATE TABLE IF NOT EXISTS mydb.foodratingspart (
food1 INT,
food2 INT,
food3 INT,
food4 INT,
id INT
)
PARTITIONED BY (name STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20240218000957_b7cceb83-fbed-4ced-88e5-2e8e11c1aacb); Time taken: 0.009 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218000957_b7cceb83-fbed-4ced-88e5-2e8e11c1aacb): CREATE TABLE IF NOT EXISTS mydb.foodratingspart (
food1 INT,
food2 INT,
food3 INT,
food4 INT,
id INT
)
PARTITIONED BY (name STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240218000957_b7cceb83-fbed-4ced-88e5-2e8e11c1aacb); Time taken: 0.039 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.058 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

Execute a Hive command of ‘DESCRIBE FORMATTED MyDb.foodratingspart;’ and capture its output as the result of this exercise.



```
0: jdbc:hive2://localhost:10000/ (default)> DESCRIBE FORMATTED MyDb.foodratingspart;
INFO  : Compiling command(queryId=hive_20240218001022_bccd5d62-3fc6-4422-b93e-9c10fdc392ca): DESCRIBE FORMATTED MyDb.foodratingspart
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name:comm
ent, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240218001022_bccd5d62-3fc6-4422-b93e-9c10fdc392ca); Time taken: 0.029 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218001022_bccd5d62-3fc6-4422-b93e-9c10fdc392ca): DESCRIBE FORMATTED MyDb.foodratingspart
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20240218001022_bccd5d62-3fc6-4422-b93e-9c10fdc392ca); Time taken: 0.077 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------------------------+----------------------------------------------------+------------------------------+
|           col_name             |                     data_type                      |           comment            |
+--------------------------------+----------------------------------------------------+------------------------------+
| # col_name                     | data_type                                          | comment                      |
| food1                          | int                                                |                              |
| food2                          | int                                                |                              |
| food3                          | int                                                |                              |
| food4                          | int                                                |                              |
| id                             | int                                                |                              |
|                                | NULL                                               | NULL                         |
| # Partition Information        | NULL                                               | NULL                         |
| # col_name                     | data_type                                          | comment                      |
| name                           | string                                             |                              |
|                                | NULL                                               | NULL                         |
| # Detailed Table Information   | NULL                                               | NULL                         |
| Database:                      | mydb                                               | NULL                         |
| OwnerType:                     | USER                                               | NULL                         |
| Owner:                         | hadoop                                             | NULL                         |
| CreateTime:                    | Sun Feb 18 00:09:57 UTC 2024                       | NULL                         |
| LastAccessTime:                | UNKNOWN                                            | NULL                         |
| Retention:                     | 0                                                  | NULL                         |
| Location:                      | hdfs://ip-172-31-1-178.us-east-2.compute.internal:8020/user/hive/warehouse/mydb.db/foodratingspart | NULL                         |
| Table Type:                    | MANAGED_TABLE                                      | NULL                         |
| Table Parameters:              | NULL                                               | NULL                         |
|                                | COLUMN_STATS_ACCURATE                              | {\"BASIC_STATS\":\"true\"}   |
|                                | bucketing_version                                  | 2                            |
|                                | numFiles                                           | 0                            |
|                                | numPartitions                                      | 0                            |
|                                | numRows                                            | 0                            |
|                                | rawDataSize                                        | 0                            |
|                                | totalSize                                          | 0                            |
|                                | transient_lastDdlTime                              | 1708214997                   |
|                                | NULL                                               | NULL                         |
| # Storage Information          | NULL                                               | NULL                         |
| SerDe Library:                 | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL                         |
| InputFormat:                   | org.apache.hadoop.mapred.TextInputFormat           | NULL                         |
| OutputFormat:                  | org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat | NULL                         |
| Compressed:                    | No                                                 | NULL                         |
| Num Buckets:                   | -1                                                 | NULL                         |
| Bucket Columns:                | []                                                 | NULL                         |
| Sort Columns:                  | []                                                 | NULL                         |
| Storage Desc Params:           | NULL                                               | NULL                         |
|                                | field.delim                                        | ,                            |
|                                | serialization.format                               | ,                            |
+--------------------------------+----------------------------------------------------+------------------------------+
```

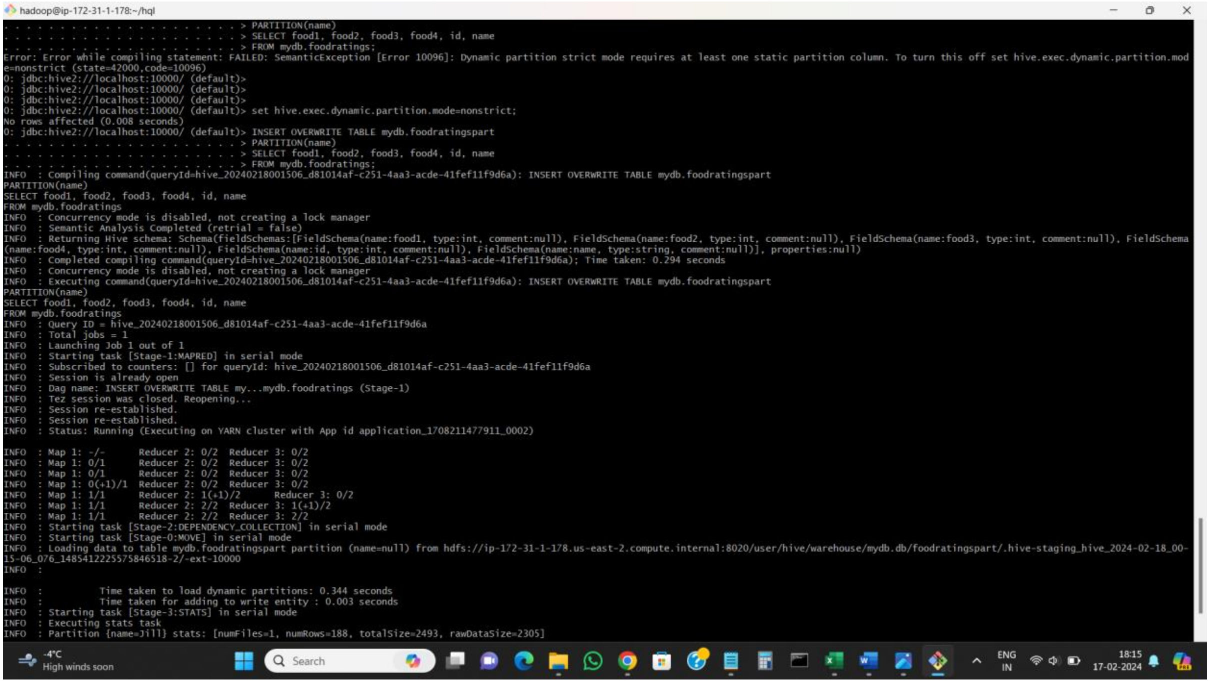
Exercise 5

Assume that the number of food critics is relatively small, say less than 10 and the number places to eat is very large, say more than 10,000. In a few short sentences explain why using the (critic) name is a good choice for a partition field while using the place id is not.

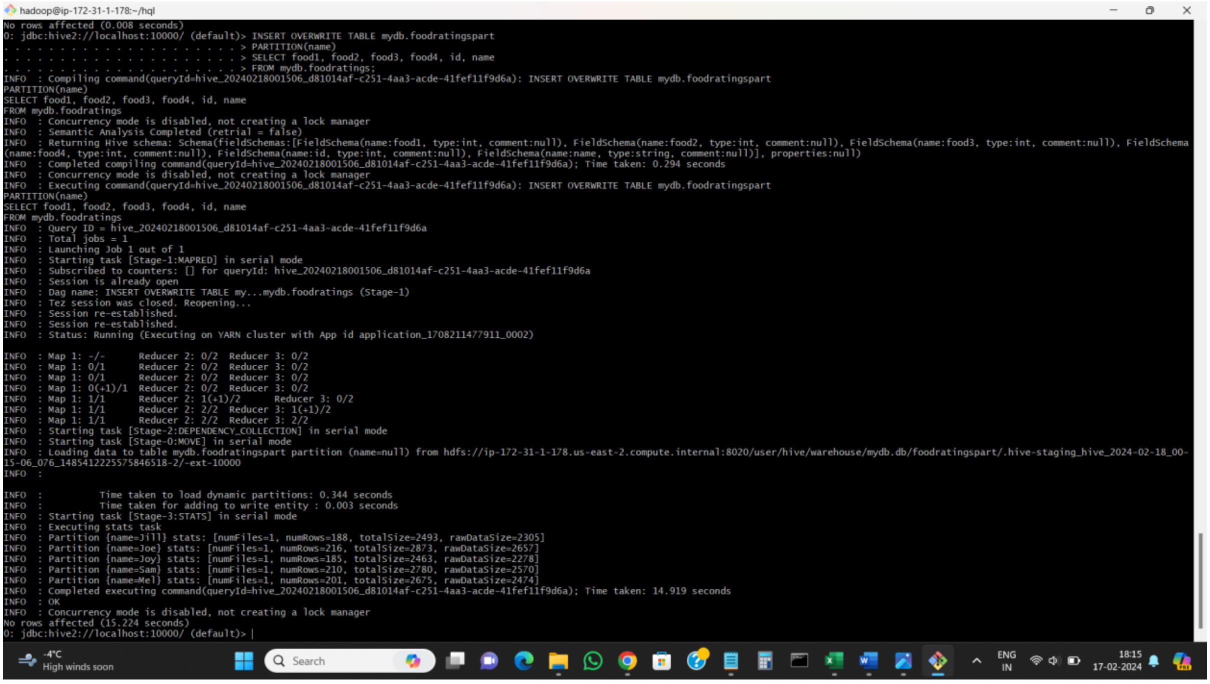
Ans: According to the given scenario, if we create a partition table for the name since we have only 10 places, then during the search operation it will only search the query for the particular partition rather than searching whole table with 10000 rows in the database. This dynamic partition will be time efficient and enhances the performances. This method will only work when the number of critics is very less than the number of places. If not, this will be a redundant job.

Exercise 6

Configure Hive to allow dynamic partition creation as described in the lecture. Now, use a hive command to copy from MyDB.foodratings into MyDB.foodratingspart to create a partitioned table from a non-partitioned one. Hint: The ‘name’ column from MyDB.foodratings should be mentioned last in this command (whatever it is). Provide a copy of the command you use to load the ‘foodratingspart’ table as a result of this exercise.



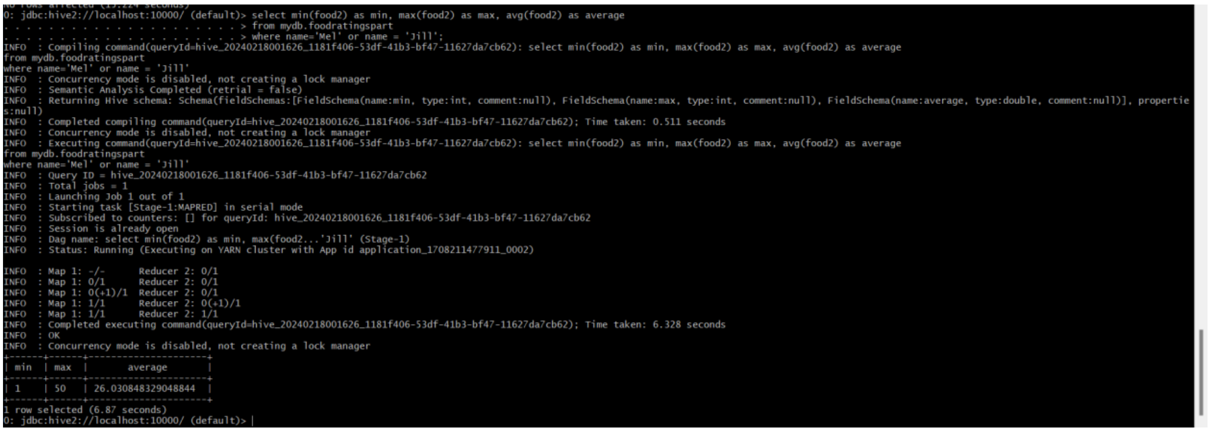
```
. . . . . . . . . . . . . . . . . . . . . . . > PARTITION(name)
. . . . . . . . . . . . . . . . . . . . . . . > SELECT food1, food2, food3, food4, id, name
. . . . . . . . . . . . . . . . . . . . . . . > FROM mydb.foodratings;
Error: Error while compiling statement: FAILED: SemanticException [Error 10096]: Dynamic partition strict mode requires at least one static partition column. To turn this off set hive.exec.dynamic.partition.mod
e=nonstrict (state=42000,code=10096)
0: jdbc:hive2://localhost:10000/ (default)>
0: jdbc:hive2://localhost:10000/ (default)>
0: jdbc:hive2://localhost:10000/ (default)>
0: jdbc:hive2://localhost:10000/ (default)> set hive.exec.dynamic.partition.mode=nonstrict;
No rows affected (0.008 seconds)
0: jdbc:hive2://localhost:10000/ (default)> INSERT OVERWRITE TABLE mydb.foodratingspart
. . . . . . . . . . . . . . . . . . . . . . . > PARTITION(name)
. . . . . . . . . . . . . . . . . . . . . . . > SELECT food1, food2, food3, food4, id, name
. . . . . . . . . . . . . . . . . . . . . . . > FROM mydb.foodratings;
INFO  : Compiling command(queryId=hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a): INSERT OVERWRITE TABLE mydb.foodratingspart
PARTITION(name)
SELECT food1, food2, food3, food4, id, name
FROM mydb.foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:food1, type:int, comment:null), FieldSchema(name:food2, type:int, comment:null), FieldSchema(name:food3, type:int, comment:null), FieldSchema
(name:food4, type:int, comment:null), FieldSchema(name:id, type:int, comment:null), FieldSchema(name:name, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a); Time taken: 0.294 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a): INSERT OVERWRITE TABLE mydb.foodratingspart
PARTITION(name)
SELECT food1, food2, food3, food4, id, name
FROM mydb.foodratings
INFO  : Query ID = hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a
INFO  : Session is already open
INFO  : Dag name: INSERT OVERWRITE TABLE my...mydb.foodratings (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1708211477911_0002)

INFO  : Map 1: -/-      Reducer 2: 0/2  Reducer 3: 0/2
INFO  : Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/2
INFO  : Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2  Reducer 3: 0/2
INFO  : Map 1: 1/1      Reducer 2: 1(+1)/2      Reducer 3: 0/2
INFO  : Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1(+1)/2
INFO  : Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 2/2
INFO  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodratingspart partition (name=null) from hdfs://ip-172-31-1-178.us-east-2.compute.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/.hive-staging_hive_2024-02-18_00-
15-06_076_1485412225575846518-2/-ext-10000
INFO  :

INFO  :          Time taken to load dynamic partitions: 0.344 seconds
INFO  :          Time taken for adding to write entity : 0.003 seconds
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Executing stats task
INFO  : Partition {name=Jill} stats: [numFiles=1, numRows=188, totalSize=2493, rawDataSize=2305]
```



```
No rows affected (0.008 seconds)
0: jdbc:hive2://localhost:10000/ (default)> INSERT OVERWRITE TABLE mydb.foodratingspart
. . . . . . . . . . . . . . . . . . . . . . . > PARTITION(name)
. . . . . . . . . . . . . . . . . . . . . . . > SELECT food1, food2, food3, food4, id, name
. . . . . . . . . . . . . . . . . . . . . . . > FROM mydb.foodratings;
INFO  : Compiling command(queryId=hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a): INSERT OVERWRITE TABLE mydb.foodratingspart
PARTITION(name)
SELECT food1, food2, food3, food4, id, name
FROM mydb.foodratings
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:food1, type:int, comment:null), FieldSchema(name:food2, type:int, comment:null), FieldSchema(name:food3, type:int, comment:null), FieldSchema
(name:food4, type:int, comment:null), FieldSchema(name:id, type:int, comment:null), FieldSchema(name:name, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a); Time taken: 0.294 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a): INSERT OVERWRITE TABLE mydb.foodratingspart
PARTITION(name)
SELECT food1, food2, food3, food4, id, name
FROM mydb.foodratings
INFO  : Query ID = hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a
INFO  : Session is already open
INFO  : Dag name: INSERT OVERWRITE TABLE my...mydb.foodratings (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1708211477911_0002)

INFO  : Map 1: -/-      Reducer 2: 0/2  Reducer 3: 0/2
INFO  : Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/2
INFO  : Map 1: 0/1      Reducer 2: 0/2  Reducer 3: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2  Reducer 3: 0/2
INFO  : Map 1: 1/1      Reducer 2: 1(+1)/2      Reducer 3: 0/2
INFO  : Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 1(+1)/2
INFO  : Map 1: 1/1      Reducer 2: 2/2  Reducer 3: 2/2
INFO  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodratingspart partition (name=null) from hdfs://ip-172-31-1-178.us-east-2.compute.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/.hive-staging_hive_2024-02-18_00-
15-06_076_1485412225575846518-2/-ext-10000
INFO  :

INFO  :          Time taken to load dynamic partitions: 0.344 seconds
INFO  :          Time taken for adding to write entity : 0.003 seconds
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Executing stats task
INFO  : Partition {name=Jill} stats: [numFiles=1, numRows=188, totalSize=2493, rawDataSize=2305]
INFO  : Partition {name=Joe} stats: [numFiles=1, numRows=216, totalSize=2873, rawDataSize=2657]
INFO  : Partition {name=Joy} stats: [numFiles=1, numRows=185, totalSize=2463, rawDataSize=2278]
INFO  : Partition {name=Sam} stats: [numFiles=1, numRows=210, totalSize=2780, rawDataSize=2570]
INFO  : Partition {name=Mel} stats: [numFiles=1, numRows=201, totalSize=2675, rawDataSize=2474]
INFO  : Completed executing command(queryId=hive_20240218001506_d81014af-c251-4aa3-acde-41fef11f9d6a); Time taken: 14.919 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (15.224 seconds)
0: jdbc:hive2://localhost:10000/ (default)> |
```

Execute a hive command to output the min, max and average of the values of the food2 column of MyDB.foodratingspart where the food critic ‘name’ is either Mel or Jill. The query and the output of this query are other results of this exercise, which you must also provide. It should look something like

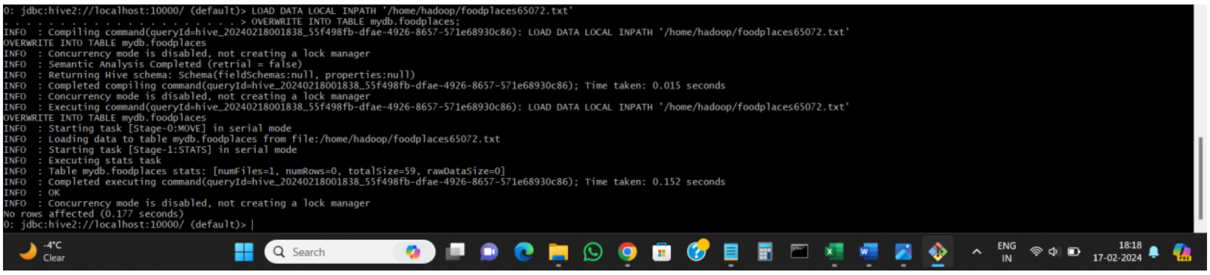
10 20 15



```
No rows affected (15.224 seconds)
0: jdbc:hive2://localhost:10000/ (default)> select min(food2) as min, max(food2) as max, avg(food2) as average
. . . . . . . . . . . . . . . . . . . . . . > from mydb.foodratingspart
. . . . . . . . . . . . . . . . . . . . . . > where name='Mel' or name = 'Jill';
INFO  : Compiling command(queryId=hive_20240218001626_1181f406-53df-41b3-bf47-11627da7cb62): select min(food2) as min, max(food2) as max, avg(food2) as average
from mydb.foodratingspart
where name='Mel' or name = 'Jill'
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:min, type:int, comment:null), FieldSchema(name:max, type:int, comment:null), FieldSchema(name:average, type:double, comment:null)], propertie
s:null)
INFO  : Completed compiling command(queryId=hive_20240218001626_1181f406-53df-41b3-bf47-11627da7cb62); Time taken: 0.511 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218001626_1181f406-53df-41b3-bf47-11627da7cb62): select min(food2) as min, max(food2) as max, avg(food2) as average
from mydb.foodratingspart
where name='Mel' or name = 'Jill'
INFO  : Query ID = hive_20240218001626_1181f406-53df-41b3-bf47-11627da7cb62
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240218001626_1181f406-53df-41b3-bf47-11627da7cb62
INFO  : Session is already open
INFO  : Dag name: select min(food2) as min, max(food2...'Jill' (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1708211477911_0002)

INFO  : Map 1: -/-      Reducer 2: 0/1
INFO  : Map 1: 0/1      Reducer 2: 0/1
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/1
INFO  : Map 1: 1/1      Reducer 2: 0(+1)/1
INFO  : Map 1: 1/1      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20240218001626_1181f406-53df-41b3-bf47-11627da7cb62); Time taken: 6.328 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------+------+--------------------+
| min  | max  |      average       |
+------+------+--------------------+
| 1    | 50   | 26.030848329048844 |
+------+------+--------------------+
1 row selected (6.87 seconds)
0: jdbc:hive2://localhost:10000/ (default)> |
```

Exercise 7

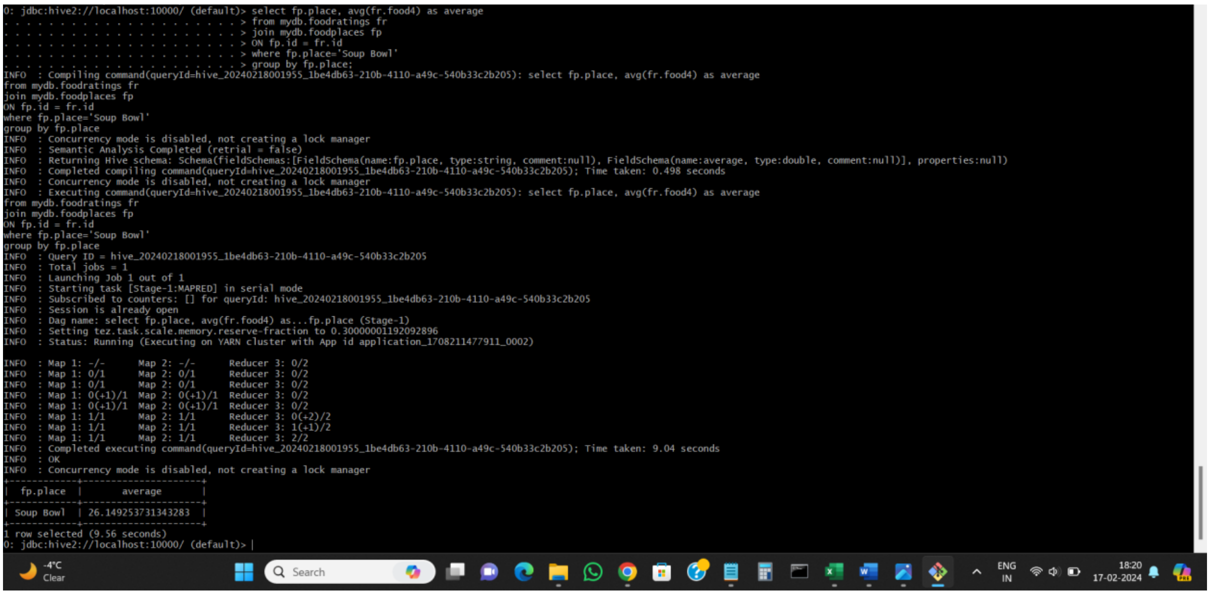
Load the foodplaces<.magic number>.txt file created using TestDataGen from your local file system into the foodplaces table.



```
0: jdbc:hive2://localhost:10000/ (default)> LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces65072.txt'
. . . . . . . . . . . . . . . . . . . . . . > OVERWRITE INTO TABLE mydb.foodplaces;
INFO  : Compiling command(queryId=hive_20240218001838_55f498fb-dfae-4926-8657-571e68930c86): LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces65072.txt'
OVERWRITE INTO TABLE mydb.foodplaces
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20240218001838_55f498fb-dfae-4926-8657-571e68930c86); Time taken: 0.015 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218001838_55f498fb-dfae-4926-8657-571e68930c86): LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces65072.txt'
OVERWRITE INTO TABLE mydb.foodplaces
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodplaces from file:/home/hadoop/foodplaces65072.txt
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Executing stats task
INFO  : Table mydb.foodplaces stats: [numFiles=1, numRows=0, totalSize=59, rawDataSize=0]
INFO  : Completed executing command(queryId=hive_20240218001838_55f498fb-dfae-4926-8657-571e68930c86); Time taken: 0.152 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.177 seconds)
0: jdbc:hive2://localhost:10000/ (default)> |
```

Use a join operation between the two tables (foodratings and foodplaces) to provide the average rating for field food4 for the restaurant 'Soup Bowl' The output of this query is the result of this exercise. It should look something like Soup Bowl 20

Provide the join command and the output as the result of this exercise.



```
0: jdbc:hive2://localhost:10000/ (default)> select fp.place, avg(fr.food4) as average
. . . . . . . . . . . . . . . . . . . . . . > from mydb.foodratings fr
. . . . . . . . . . . . . . . . . . . . . . > join mydb.foodplaces fp
. . . . . . . . . . . . . . . . . . . . . . > ON fp.id = fr.id
. . . . . . . . . . . . . . . . . . . . . . > where fp.place='Soup Bowl'
. . . . . . . . . . . . . . . . . . . . . . > group by fp.place;
INFO  : Compiling command(queryId=hive_20240218001955_1be4db63-210b-4110-a49c-540b33c2b205): select fp.place, avg(fr.food4) as average
from mydb.foodratings fr
join mydb.foodplaces fp
ON fp.id = fr.id
where fp.place='Soup Bowl'
group by fp.place
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:fp.place, type:string, comment:null), FieldSchema(name:average, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240218001955_1be4db63-210b-4110-a49c-540b33c2b205); Time taken: 0.498 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240218001955_1be4db63-210b-4110-a49c-540b33c2b205): select fp.place, avg(fr.food4) as average
from mydb.foodratings fr
join mydb.foodplaces fp
ON fp.id = fr.id
where fp.place='Soup Bowl'
group by fp.place
INFO  : Query ID = hive_20240218001955_1be4db63-210b-4110-a49c-540b33c2b205
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20240218001955_1be4db63-210b-4110-a49c-540b33c2b205
INFO  : Session is already open
INFO  : Dag name: select fp.place, avg(fr.food4) as...fp.place (Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Status: Running (Executing on YARN cluster with App id application_1708211477911_0002)

INFO  : Map 1: -/-      Map 2: -/-      Reducer 3: 0/2
INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/2
INFO  : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/2
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/2
INFO  : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/2
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+2)/2
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1(+1)/2
INFO  : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2
INFO  : Completed executing command(queryId=hive_20240218001955_1be4db63-210b-4110-a49c-540b33c2b205); Time taken: 9.04 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------+--------------------+
|  fp.place  |      average       |
+------------+--------------------+
| Soup Bowl  | 26.149253731343283 |
+------------+--------------------+
1 row selected (9.56 seconds)
0: jdbc:hive2://localhost:10000/ (default)> |
```

Exercise 8

Read the article "An Introduction to Big Data Formats" found on the blackboard in section "Articles" and provide short (2 to 4 sentence) answers to the following questions:

a) When is the most important consideration when choosing a row format and when a column format for your big data file? Column-based storage is most beneficial for running analytics queries that involve only a subset of columns to be analyzed across massive sets of data. Row-based storage is more appropriate if the majority or all of the columns in each row of data must be accessible for your queries.
Ans: At the highest level, column-based storage is most useful when performing analytics queries that require only a subset of columns examined over very large data sets. If your queries require access to all or most of the columns of each row of data, row-based storage will be better suited to your needs.

b) What is "splittability" for a column file format and why is it important when processing large volumes of data? Big data is huge. In order to handle huge datasets efficiently, it is typically necessary to divide the work into segments that can be assigned to different processors. If the query computation is only concerned with one column at a time, a column-based format will be easier to divide into distinct jobs. A batch of rows is taken and stored in columnar format using row-columnar columnar formats. Once split, these batches form boundaries.
Ans: Big data is HUGE. Processing such datasets efficiently usually requires breaking the job into parts that can be farmed out to separate processors. A column-based format will be more amenable to splitting into separate jobs if the query calculation is concerned with a single column at a time. The columnar formats are row-columnar, which means they take a batch of rows and store that batch in columnar format. These batches then become split boundaries.

c) What can files stored in column format achieve better compression than those stored in row format?
Ans: Compression uses encoding for frequently repeating data to achieve this reduction. Columnar data can achieve better compression rates than row-based data. Storing values by column, with the same type next to each other, allows you to do more efficient compression on them than if

you’re storing rows of data. For example, storing all dates together in memory allows for more efficient compression than storing data of various types next to each other—such as string, number, date, string, date.

d) Under what circumstances would it be the best choice to use the “Parquet” column file format?
Ans: Parquet is commonly used with Apache Impala, an analytics database for Hadoop. Parquet is especially adept at analysing wide datasets with many columns. Each Parquet file contains binary data organized by “row group.” For each row group, the data values are organized by column. This enables the compression benefits that we described above. Parquet is a good choice for read-heavy workloads.