Assignment #4 (Big Data Technologies) Solutions

Uploaded the files hql.zip and TestDataGen.class to the /home/hadoop folder using "scp" Command.

```
(base) akshayjain@dhcp45 Assignment4 % scp -i emr-key-pair.pem hql.zip hadoop@ec
2-54-205-44-5.compute-1.amazonaws.com:/home/hadoop
hql.zip                                       100%  402KB   2.7MB/s   00:00
```

Execute TestDataGen.class using "java TestDataGen"

```
Warning: Permanently added 'ec2-54-205-44-5.compute-1.amazonaws.com' (ED25519) t
o the list of known hosts.

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEE 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
EEEEEEEEEEEEEEEEEEE MMMMMMMM             MMMMMMMM RRRRRRR      RRRRRR
```

Magic Number = 168337
Data of foodratings168337.txt

```
[hadoop@ip-172-31-31-40 ~]$ ls
hql.zip  TestDataGen.class
[hadoop@ip-172-31-31-40 ~]$ java TestDataGen
Magic Number = 168337
[hadoop@ip-172-31-31-40 ~]$ ls
foodplaces168337.txt  foodratings168337.txt  hql.zip  TestDataGen.class
[hadoop@ip-172-31-31-40 ~]$ cat foodratings168337.txt
Joy,15,36,45,49,1
Sam,1,6,23,43,3
Sam,24,32,37,18,3
Jill,44,4,26,22,2
Mel,35,14,43,49,1
Jill,28,22,25,10,2
Sam,14,31,48,27,5
Mel,19,18,34,15,2
Jill,23,16,5,30,1
Jill,20,41,9,8,1
Sam,38,6,23,17,2
Mel,17,44,32,36,4
Jill,5,41,4,16,5
Jill,42,17,25,13,2
Sam,31,17,49,29,3
Sam,12,49,44,37,3
Mel,3,50,41,5,3
```

Data of foodplaces168337.txt

```
[hadoop@ip-172-31-31-40 ~]$ ls
foodplaces168337.txt  foodratings168337.txt  hql.zip  TestDataGen.class
[hadoop@ip-172-31-31-40 ~]$ cat foodplaces168337.txt
1,China Bistro
2,Atlantic
3,Food Town
4,Jake's
5,Soup Bowl
```

Exercise 1. Create Database MyDb

Command used —> create database MyDb;

```
0: jdbc:hive2://localhost:10000/ (cs595)> create database MyDb;
INFO  : Compiling command(queryId=hive_20220929224122_9777c5c6-45ed-44e5-bb0e-9db1b30d4156): create database MyDb
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220929224122_9777c5c6-45ed-44e5-bb0e-9db1b30d4156 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
```

Select Database

Command Used → use MyDb;

```
0: jdbc:hive2://localhost:10000/ (cs595)> use MyDb;
INFO  : Compiling command(queryId=hive_20220929224146_cf016e44-be78-406d-a264-2b9bb251ffcf): use MyDb
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220929224146_cf016e44-be78-406d-a264-2b9bb251ffcf : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
```

Making sure if database is working.

Command Used: - set hive.cli.print.current.db=true;

```
0: jdbc:hive2://localhost:10000/ (MyDb)> set hive.cli.print.current.db=true;
No rows affected (0.007 seconds)
```

Creating foodratings table: -

Command Used: - CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) row format delimited fields terminated by ',';

```
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) row format delimited fields terminated by ',';
INFO  : Compiling command(queryId=hive_20220929224413_76a53bcf-fbfe-4114-bc09-88560fec5074): CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) row format delimited field
s terminated by ','
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220929224413_76a53bcf-fbfe-4114-bc09-88560fec5074 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
```

Showing the created table: -

Command Used: - show tables;

```
0: jdbc:hive2://localhost:10000/ (MyDb)> show tables;
INFO  : Compiling command(queryId=hive_20220212174450_ca021991-b679-4259-aa36-f320a747756b): show tables
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220212174450_ca021991-b679-4259-aa36-f320a747756b : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]
  Stage-1 depends on stages: Stage-0 [FETCH]

STAGE PLANS:
  Stage: Stage-0
      Show Table Operator:
        Show Tables
          database name: MyDb
          result file: file:/mnt/tmp/hive/30d5f57c-0436-48c3-a211-0731b90ce2a4/hive_2022-02-12_17-44-50_176_7931896969328643915-1/-local-10000

  Stage: Stage-1
    Fetch Operator
      limit: -1
      Processor Tree:
        ListSink


INFO  : Completed compiling command(queryId=hive_20220212174450_ca021991-b679-4259-aa36-f320a747756b); Time taken: 0.042 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220212174450_ca021991-b679-4259-aa36-f320a747756b): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220212174450_ca021991-b679-4259-aa36-f320a747756b); Time taken: 0.021 seconds
INFO  : OK
+--------------+
|   tab_name   |
+--------------+
| foodratings  |
+--------------+
1 row selected (0.185 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

Adding comments to the table: -

Command Used: - ALTER TABLE <TableName> CHANGE <ColumnName> <ColumnName> <ColumnDataType> comment '<comments>';

```
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change name food1 food1 int comment 'name comments';
Error: Error while compiling statement: FAILED: ParseException line 1:42 cannot recognize input near 'food1' 'int' 'comment' in column type (state=42000,cod
e=40000)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food1 food1 int comment 'food1 comments';
INFO  : Compiling command(queryId=hive_20220212175804_53f71563-3ecd-496d-83be-7a37346a1aff): ALTER TABLE foodratings change food1 food1 int comment 'food1 c
omments'
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220212175804_53f71563-3ecd-496d-83be-7a37346a1aff : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
      Alter Table Operator:
        Alter Table
          type: rename column
          old name: MyDb.foodratings


INFO  : Completed compiling command(queryId=hive_20220212175804_53f71563-3ecd-496d-83be-7a37346a1aff); Time taken: 0.091 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220212175804_53f71563-3ecd-496d-83be-7a37346a1aff): ALTER TABLE foodratings change food1 food1 int comment 'food1 c
omments'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220212175804_53f71563-3ecd-496d-83be-7a37346a1aff); Time taken: 0.227 seconds
INFO  : OK
No rows affected (0.356 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food2 food2 int comment 'food2 comments';
INFO  : Compiling command(queryId=hive_20220212175821_efc7e410-77c4-400b-b3e9-12b301633aaf): ALTER TABLE foodratings change food2 food2 int comment 'food2 c
omments'
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220212175821_efc7e410-77c4-400b-b3e9-12b301633aaf : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food2 food2 int comment 'food2 comments';
INFO  : Compiling command(queryId=hive_20220212175821_efc7e410-77c4-400b-b3e9-12b301633aaf): ALTER TABLE foodratings change food2 food2 int comment 'food2 c
omments'
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220212175821_efc7e410-77c4-400b-b3e9-12b301633aaf : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
      Alter Table Operator:
        Alter Table
          type: rename column
          old name: MyDb.foodratings


INFO  : Completed compiling command(queryId=hive_20220212175821_efc7e410-77c4-400b-b3e9-12b301633aaf); Time taken: 0.049 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220212175821_efc7e410-77c4-400b-b3e9-12b301633aaf): ALTER TABLE foodratings change food2 food2 int comment 'food2 c
omments'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220212175821_efc7e410-77c4-400b-b3e9-12b301633aaf); Time taken: 0.07 seconds
INFO  : OK
No rows affected (0.136 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food3 food3 int comment 'food3 comments';
INFO  : Compiling command(queryId=hive_20220212175844_ef8378ad-e1dd-4c14-8da1-5951830b259e): ALTER TABLE foodratings change food3 food3 int comment 'food3 c
omments'
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220212175844_ef8378ad-e1dd-4c14-8da1-5951830b259e : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
      Alter Table Operator:
        Alter Table
          type: rename column
          old name: MyDb.foodratings
```

Showing descriptions of the table.
Command used: - describe formatted foodratings;

```
INFO  : Executing command(queryId=hive_20220929224654_150a9ea6-6b30-429a-a3f9-8d5fd443b304): describe formatted foodratings
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220929224654_150a9ea6-6b30-429a-a3f9-8d5fd443b304); Time taken: 0.089 seconds
INFO  : OK
+-------------------------------+----------------------------------------------------+-------------------------------+
|           col_name            |                     data_type                      |            comment            |
+-------------------------------+----------------------------------------------------+-------------------------------+
| # col_name                    | data_type                                          | comment                       |
|                               | NULL                                               | NULL                          |
| name                          | string                                             |                               |
| food1                         | int                                                |                               |
| food2                         | int                                                |                               |
| food3                         | int                                                |                               |
| food4                         | int                                                |                               |
| id                            | int                                                |                               |
|                               | NULL                                               | NULL                          |
| # Detailed Table Information  | NULL                                               | NULL                          |
| Database:                     | mydb                                               | NULL                          |
| Owner:                        | hadoop                                             | NULL                          |
| CreateTime:                   | Thu Sep 29 22:44:13 UTC 2022                       | NULL                          |
| LastAccessTime:               | UNKNOWN                                            | NULL                          |
| Retention:                    | 0                                                  | NULL                          |
| Location:                     | hdfs://ip-172-31-31-40.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratings | NULL                          |
| Table Type:                   | MANAGED_TABLE                                      | NULL                          |
| Table Parameters:             | NULL                                               | NULL                          |
|                               | COLUMN_STATS_ACCURATE                              | {\"BASIC_STATS\":\"true\"}    |
|                               | numFiles                                           | 0                             |
|                               | numRows                                            | 0                             |
|                               | rawDataSize                                        | 0                             |
|                               | totalSize                                          | 0                             |
|                               | transient_lastDdlTime                              | 1664491453                    |
|                               | 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                               | ,                             |
+-------------------------------+----------------------------------------------------+-------------------------------+
36 rows selected (0.163 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE TABLE foodplaces(id int,places string) row format delimited fields terminated by ',';
INFO  : Compiling command(queryId=hive_20220929224902_9fe180d6-71c7-48f4-8c99-d9b88b763b7c): CREATE TABLE foodplaces(id int,places string) row format delimited fields terminated by ','
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220929224902_9fe180d6-71c7-48f4-8c99-d9b88b763b7c : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
      Create Table Operator:
        Create Table
          columns: id int, places string
          field delimiter: ,
          input format: org.apache.hadoop.mapred.TextInputFormat
          output format: org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat
          serde name: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
          name: MyDb.foodplaces
```

Creating table foodplaces: -
Command Used: - CREATE TABLE foodplaces(id int,places string) row format delimited fields terminated by ',';

```
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE TABLE foodplaces(id int,places string) row format delimited fields terminated by ',';
INFO  : Compiling command(queryId=hive_20220212180224_b0c06521-f2ea-40de-a687-765078222e28): CREATE TABLE foodplaces(id int,places string) row format delimi
ted fields terminated by ','
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220212180224_b0c06521-f2ea-40de-a687-765078222e28 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
      Create Table Operator:
        Create Table
          columns: id int, places string
          field delimiter: ,
          input format: org.apache.hadoop.mapred.TextInputFormat
          output format: org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat
          serde name: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
          name: MyDb.foodplaces


INFO  : Completed compiling command(queryId=hive_20220212180224_b0c06521-f2ea-40de-a687-765078222e28); Time taken: 0.051 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220212180224_b0c06521-f2ea-40de-a687-765078222e28): CREATE TABLE foodplaces(id int,places string) row format delimi
ted fields terminated by ','
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220212180224_b0c06521-f2ea-40de-a687-765078222e28); Time taken: 0.142 seconds
INFO  : OK
No rows affected (0.22 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> show tables;
```

Showing the created table: -
Command Used: - show tables;

```
No rows affected (0.22 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> show tables;
INFO  : Compiling command(queryId=hive_20220212180233_6bcf3d39-d22f-4b78-998d-675f953080d1): show tables
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220212180233_6bcf3d39-d22f-4b78-998d-675f953080d1 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]
  Stage-1 depends on stages: Stage-0 [FETCH]

STAGE PLANS:
  Stage: Stage-0
      Show Table Operator:
        Show Tables
          database name: MyDb
          result file: file:/mnt/tmp/hive/30d5f57c-0436-48c3-a211-0731b90ce2a4/hive_2022-02-12_18-02-33_363_3671963478344190709-1/-local-10000

  Stage: Stage-1
    Fetch Operator
      limit: -1
      Processor Tree:
        ListSink


INFO  : Completed compiling command(queryId=hive_20220212180233_6bcf3d39-d22f-4b78-998d-675f953080d1); Time taken: 0.084 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220212180233_6bcf3d39-d22f-4b78-998d-675f953080d1): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220212180233_6bcf3d39-d22f-4b78-998d-675f953080d1); Time taken: 0.02 seconds
INFO  : OK
+--------------+
|   tab_name   |
+--------------+
| foodplaces   |
| foodratings  |
+--------------+
2 rows selected (0.161 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

Showing descriptions of the table.
Command used: - describe formatted foodplaces;

```
INFO  : Executing command(queryId=hive_20220929225126_38c53725-ddbd-498f-bc3d-d5b90de972a4): describe formatted foodplaces
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220929225126_38c53725-ddbd-498f-bc3d-d5b90de972a4); Time taken: 0.027 seconds
INFO  : OK
+-------------------------------+----------------------------------------------------+-----------------------------+
|           col_name            |                     data_type                      |           comment           |
+-------------------------------+----------------------------------------------------+-----------------------------+
| # col_name                    | data_type                                          | comment                     |
|                               | NULL                                               | NULL                        |
| id                            | int                                                |                             |
| places                        | string                                             |                             |
|                               | NULL                                               | NULL                        |
| # Detailed Table Information  | NULL                                               | NULL                        |
| Database:                     | mydb                                               | NULL                        |
| Owner:                        | hadoop                                             | NULL                        |
| CreateTime:                   | Thu Sep 29 22:49:02 UTC 2022                       | NULL                        |
| LastAccessTime:               | UNKNOWN                                            | NULL                        |
| Retention:                    | 0                                                  | NULL                        |
| Location:                     | hdfs://ip-172-31-31-40.ec2.internal:8020/user/hive/warehouse/mydb.db/foodplaces | NULL                        |
| Table Type:                   | MANAGED_TABLE                                      | NULL                        |
| Table Parameters:             | NULL                                               | NULL                        |
|                               | COLUMN_STATS_ACCURATE                              | {\"BASIC_STATS\":\"true\"}  |
|                               | numFiles                                           | 0                           |
|                               | numRows                                            | 0                           |
|                               | rawDataSize                                        | 0                           |
|                               | totalSize                                          | 0                           |
|                               | transient_lastDdlTime                              | 1664491742                  |
|                               | 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                               | ,                           |
+-------------------------------+----------------------------------------------------+-----------------------------+
32 rows selected (0.104 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces168337.txt' OWERWRITE INTO TABLE foodplaces;
Error: Error while compiling statement: FAILED: ParseException line 1:59 extraneous input 'OWERWRITE' expecting INTO near '<EOF>' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces168337.txt' OWERWRITE INTO TABLE foodplaces;
Error: Error while compiling statement: FAILED: ParseException line 1:59 extraneous input 'OWERWRITE' expecting INTO near '<EOF>' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (MyDb)> ls
. . . . . > . . . . . > . . . . . > . . . . . > LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces168337.txt' OWERWRITE INTO TABLE foodplaces;
Error: Error while compiling statement: FAILED: ParseException line 1:0 cannot recognize input near 'ls' 'LOAD' 'DATA' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces168337.txt' OWERWRITE INTO TABLE foodplaces;
Error: Error while compiling statement: FAILED: ParseException line 1:59 extraneous input 'OWERWRITE' expecting INTO near '<EOF>' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces168337.txt' OWERWRITE INTO TABLE foodplaces;
Error: Error while compiling statement: FAILED: ParseException line 1:59 extraneous input 'OWERWRITE' expecting INTO near '<EOF>' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodratings168337.txt'
. . . . . > . . . . . > . . . . . > OWERWRITE INTO TABLE foodratings;
Error: Error while compiling statement: FAILED: ParseException line 2:0 extraneous input 'OWERWRITE' expecting INTO near '<EOF>' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (MyDb)> describe formatted foodplaces;
INFO  : Compiling command(queryId=hive_20220929231427_be6ad47d-ac41-4a40-92ad-00dda2af8c00): describe formatted foodplaces
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(name:data_type, type:string, comment:from deserialize
e:comment, type:string, comment:from deserializer)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220929231427_be6ad47d-ac41-4a40-92ad-00dda2af8c00 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]
  Stage-1 depends on stages: Stage-0 [FETCH]
```

Exercise 2. Load Data of foodratings168337.txt into the table foodratings
Command used → LOAD DATA LOCAL INPATH '/home/hadoop/foodratings168337.txt' OVERWRITE INTO TABLE foodratings;

```
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodratings168337.txt'
. . . . . .> OVERWRITE INTO TABLE foodratings;
INFO  : Compiling command(queryId=hive_20220929231841_cadd013c-bfb2-4b0d-9c58-8c35c3f7f111): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings168337.txt'
OVERWRITE INTO TABLE foodratings
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220929231841_cadd013c-bfb2-4b0d-9c58-8c35c3f7f111 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [MOVE]
  Stage-1 depends on stages: Stage-0 [STATS]

STAGE PLANS:
  Stage: Stage-0
    Move Operator
      tables:
          replace: true
          source: file:/home/hadoop/foodratings168337.txt
          table:
              input format: org.apache.hadoop.mapred.TextInputFormat
              output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
              properties:
                COLUMN_STATS_ACCURATE {"BASIC_STATS":"true"}
                bucket_count -1
                column.name.delimiter ,
                columns name,food1,food2,food3,food4,id
                columns.comments
                columns.types string:int:int:int:int:int
                field.delim ,
                file.inputformat org.apache.hadoop.mapred.TextInputFormat
                file.outputformat org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
                location hdfs://ip-172-31-31-40.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratings
                name mydb.foodratings
                numFiles 0
                numRows 0
                rawDataSize 0
                serialization.ddl struct foodratings { string name, i32 food1, i32 food2, i32 food3, i32 food4, i32 id}
                serialization.format ,
                serialization.lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                totalSize 0
                transient_lastDdlTime 1664491453
              serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
              name: mydb.foodratings

  Stage: Stage-1
    Stats-Aggr Operator
```

Print min, max and average of the values of the food3 column of the foodratings table
Command Used → select min(food3) as min,max(food3) as max,avg(food3) as avg from foodratings;

```
No rows affected (0.345 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> select min(food3) as min,max(food3) as max,avg(food3) as avg from foodratings;
INFO  : Compiling command(queryId=hive_20220929232036_c583ce3b-12db-48b0-8a79-f2fc943d678c): select min(food3) as min,max(food3) as max,avg(food3) as avg from foodratings
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:min, type:int, comment:null), FieldSchema(name:max, type:int, comment:null), FieldSchema(name:avg, type:d
ies:null)
INFO  : EXPLAIN output for queryid hive_20220929232036_c583ce3b-12db-48b0-8a79-f2fc943d678c : STAGE DEPENDENCIES:
  Stage-1 is a root stage [MAPRED]
  Stage-0 depends on stages: Stage-1 [FETCH]

STAGE PLANS:
  Stage: Stage-1
    Tez
      DagId: hive_20220929232036_c583ce3b-12db-48b0-8a79-f2fc943d678c:2
      Edges:
        Reducer 2 <- Map 1 (CUSTOM_SIMPLE_EDGE)
      DagName:
      Vertices:
        Map 1
            Map Operator Tree:
                TableScan
                  alias: foodratings
                  Statistics: Num rows: 4371 Data size: 17485 Basic stats: COMPLETE Column stats: NONE
                  GatherStats: false
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: select min(food3) as min,max(f...foodratings(Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1664489100889_0002)

INFO  : Map 1: 0/1      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: 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_20220929232036_c583ce3b-12db-48b0-8a79-f2fc943d678c); Time taken: 17.765 seconds
INFO  : OK
+------+------+----------+
| min  | max  |   avg    |
+------+------+----------+
| 1    | 50   | 26.607   |
+------+------+----------+
1 row selected (18.516 seconds)
```

Exercise 3.

Print min, max and average of the values of the food1 column grouped by the first column 'name'
Command Used → SELECT NAME, min(food1) as min,max(food1) as max, avg(food1) as avg from foodratings group by name;

```
0: jdbc:hive2://localhost:10000/ (MyDb)> SELECT NAME, min(food1) as min,max(food1) as max, avg(food1) as avg from foodratings group by name;
INFO  : Compiling command(queryId=hive_20220929232155_50383efa-891b-422f-ac4c-e2abb0e4b11e): SELECT NAME, min(food1) as min,max(food1) as max, avg(food1) as avg from foodratings group by name
INFO  : Semantic Analysis Completed
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), FieldSc
hema(name:avg, type:double, comment:null)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220929232155_50383efa-891b-422f-ac4c-e2abb0e4b11e : STAGE DEPENDENCIES:
  Stage-1 is a root stage [MAPRED]
  Stage-0 depends on stages: Stage-1 [FETCH]

STAGE PLANS:
  Stage: Stage-1
    Tez
      DagId: hive_20220929232155_50383efa-891b-422f-ac4c-e2abb0e4b11e:3
      Edges:
        Reducer 2 <- Map 1 (SIMPLE_EDGE)
      DagName:
      Vertices:
        Map 1
            Map Operator Tree:
                TableScan
INFO  : Dag name: SELECT NAME, min(food1) as min,max(fo...name(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1664489100889_0002)

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: 1(+1)/2
INFO  : Map 1: 1/1      Reducer 2: 2/2
INFO  : Completed executing command(queryId=hive_20220929232155_50383efa-891b-422f-ac4c-e2abb0e4b11e); Time taken: 5.9 seconds
INFO  : OK
+-------+------+------+---------------------+
| name  | min  | max  |         avg         |
+-------+------+------+---------------------+
| Jill  | 1    | 50   | 27.02185792349727   |
| Joe   | 1    | 50   | 25.1968085106383    |
| Joy   | 1    | 50   | 26.935323383084576  |
| Mel   | 1    | 50   | 24.805825242718445  |
| Sam   | 1    | 50   | 25.60810810810811   |
+-------+------+------+---------------------+
```

Exercise 4.

Create Table foodratingspart
Command Used: - CREATE TABLE foodratingspart (food1 int,food2 int,food3 int,food4 int,id int) partitioned by (name string) row format delimited fields terminated by ',';

```
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE TABLE foodratingspart (food1 int,food2 int,food3 int,food4 int,id int) partitioned by (name string) row format delimited fields terminated by ',';
INFO  : Compiling command(queryId=hive_20220929232241_688538ec-1798-42b8-aa8a-b4a3e31ae607): CREATE TABLE foodratingspart (food1 int,food2 int,food3 int,food4 int,id int) partitioned by (name string) row
format delimited fields terminated by ','
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220929232241_688538ec-1798-42b8-aa8a-b4a3e31ae607 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
      Create Table Operator:
        Create Table
          columns: food1 int, food2 int, food3 int, food4 int, id int
          field delimiter: ,
          input format: org.apache.hadoop.mapred.TextInputFormat
          output format: org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat
          partition columns: name string
          serde name: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
          name: MyDb.foodratingspart


INFO  : Completed compiling command(queryId=hive_20220929232241_688538ec-1798-42b8-aa8a-b4a3e31ae607); Time taken: 0.028 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220929232241_688538ec-1798-42b8-aa8a-b4a3e31ae607): CREATE TABLE foodratingspart (food1 int,food2 int,food3 int,food4 int,id int) partitioned by (name string) row
format delimited fields terminated by ','
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220929232241_688538ec-1798-42b8-aa8a-b4a3e31ae607); Time taken: 0.052 seconds
INFO  : OK
No rows affected (0.108 seconds)
```

Showing descriptions of the table.

Command used: - describe formatted foodratingspart;

```
0: jdbc:hive2://localhost:10000/ (MyDb)> describe formatted foodratingspart;
INFO  : Compiling command(queryId=hive_20220929232254_ac4d58ca-a457-4246-83e1-eaf802f5ce72): describe formatted foodratingspart
INFO  : Semantic Analysis Completed
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(nam
e:comment, type:string, comment:from deserializer)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220929232254_ac4d58ca-a457-4246-83e1-eaf802f5ce72 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]
  Stage-1 depends on stages: Stage-0 [FETCH]

STAGE PLANS:
  Stage: Stage-0
      Describe Table Operator:
        Describe Table
          result file: file:/mnt/tmp/hive/e7e6b92d-ba67-427a-8f2c-8458b95f0f40/hive_2022-09-29_23-22-54_074_5674419839106407086-1/-local-10000
          table: foodratingspart

  Stage: Stage-1
    Fetch Operator
      limit: -1
      Processor Tree:
        ListSink


INFO  : Completed compiling command(queryId=hive_20220929232254_ac4d58ca-a457-4246-83e1-eaf802f5ce72); Time taken: 0.026 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220929232254_ac4d58ca-a457-4246-83e1-eaf802f5ce72): describe formatted foodratingspart
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220929232254_ac4d58ca-a457-4246-83e1-eaf802f5ce72); Time taken: 0.035 seconds
INFO  : OK
+-------------------------------+----------------------------------------------------+-----------------------------+
|           col_name            |                     data_type                      |           comment           |
+-------------------------------+----------------------------------------------------+-----------------------------+
| # col_name                    | data_type                                          | comment                     |
|                               | NULL                                               | NULL                        |
| food1                         | int                                                |                             |
| food2                         | int                                                |                             |
| food3                         | int                                                |                             |
| food4                         | int                                                |                             |
| id                            | int                                                |                             |
|                               | NULL                                               | NULL                        |
| # Partition Information       | NULL                                               | NULL                        |
| # col_name                    | data_type                                          | comment                     |
|                               | NULL                                               | NULL                        |
| name                          | string                                             |                             |
|                               | NULL                                               | NULL                        |
| # Detailed Table Information  | NULL                                               | NULL                        |
| Database:                     | mydb                                               | NULL                        |
| Owner:                        | hadoop                                             | NULL                        |
| CreateTime:                   | Thu Sep 29 23:22:41 UTC 2022                       | NULL                        |
| LastAccessTime:               | UNKNOWN                                            | NULL                        |
| Retention:                    | 0                                                  | NULL                        |
| Location:                     | hdfs://ip-172-31-31-40.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart | NULL                |
| Table Type:                   | MANAGED_TABLE                                      | NULL                        |
| Table Parameters:             | NULL                                               | NULL                        |
|                               | COLUMN_STATS_ACCURATE                              | {\"BASIC_STATS\":\"true\"}  |
|                               | numFiles                                           | 0                           |
|                               | numPartitions                                      | 0                           |
|                               | numRows                                            | 0                           |
|                               | rawDataSize                                        | 0                           |
|                               | totalSize                                          | 0                           |
|                               | transient_lastDdlTime                              | 1664493761                  |
|                               | NULL                                               | NULL                        |
| # Storage Information         | NULL                                               | NULL                        |


|                               | numRows                                            | 0                           |
|                               | rawDataSize                                        | 0                           |
|                               | totalSize                                          | 0                           |
|                               | transient_lastDdlTime                              | 1664493761                  |
|                               | 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                               | ,                           |
+-------------------------------+----------------------------------------------------+-----------------------------+
41 rows selected (0.094 seconds)
```

Exercise 5.

As number of critics less so when we partition by critic each partition has data for one single critic and we can perform all operation related critic easily. Example max critic review, number of places critic visited.

Exercise 6.

1. Allowed hive to do Dynamic partition
Command Used —> set hive.exec.dynamic.partition.mode=nonstrict;

```
0: jdbc:hive2://localhost:10000/ (MyDb)> set hive.cli.print.current.db=true;
No rows affected (0.007 seconds)
```

2. Copy Data from foodratings into foodratingspart to create a par--oned table from a non-par--oned one

Command Used → INSERT OVERWRITE TABLE foodratingspart PARTITION (name) SELECT food1,food2,food3,food4,id,name from foodratings

```
0: jdbc:hive2://localhost:10000/ (MyDb)> INSERT OVERWRITE TABLE foodratingspart PARTITION (name) SELECT food1,food2,food3,food4,id,name from foodratings;
INFO  : Compiling command(queryId=hive_20220929232339_d07a2c3b-bb5d-4985-b318-18ecce2be2b2): INSERT OVERWRITE TABLE foodratingspart PARTITION (name) SELECT food1,food2,food3,food4,id,name from foodratings
INFO  : Semantic Analysis Completed
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), Field
Schema(name:food4, type:int, comment:null), FieldSchema(name:id, type:int, comment:null), FieldSchema(name:name, type:string, comment:null)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220929232339_d07a2c3b-bb5d-4985-b318-18ecce2be2b2 : STAGE DEPENDENCIES:
  Stage-1 is a root stage [MAPRED]
  Stage-2 depends on stages: Stage-1 [DEPENDENCY_COLLECTION]
  Stage-0 depends on stages: Stage-2 [MOVE]
  Stage-3 depends on stages: Stage-0 [STATS]

STAGE PLANS:
  Stage: Stage-1
    Tez
      DagId: hive_20220929232339_d07a2c3b-bb5d-4985-b318-18ecce2be2b2:4
      Edges:
        Reducer 2 <- Map 1 (SIMPLE_EDGE)
      DagName:
      Vertices:
        Map 1
            Map Operator Tree:
                TableScan
                  alias: foodratings
                  Statistics: Num rows: 145 Data size: 17485 Basic stats: COMPLETE Column stats: NONE
                  GatherStats: false
                  Select Operator
                    expressions: food1 (type: int), food2 (type: int), food3 (type: int), food4 (type: int), id (type: int), name (type: string)
                    outputColumnNames: _col0, _col1, _col2, _col3, _col4, _col5
                    Statistics: Num rows: 145 Data size: 17485 Basic stats: COMPLETE Column stats: NONE
                    Reduce Output Operator
                      key expressions: _col5 (type: string)
                      null sort order: a
                      sort order: +
                      Map-reduce partition columns: _col5 (type: string)
                      Statistics: Num rows: 145 Data size: 17485 Basic stats: COMPLETE Column stats: NONE
                      tag: -1
                      value expressions: _col0 (type: int), _col1 (type: int), _col2 (type: int), _col3 (type: int), _col4 (type: int)
                      auto parallelism: true
            Path -> Alias:
              hdfs://ip-172-31-31-40.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratings [foodratings]
            Path -> Partition:
              hdfs://ip-172-31-31-40.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratings
                Partition
                  base file name: foodratings
                  input format: org.apache.hadoop.mapred.TextInputFormat
                  output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
                  properties:
                    bucket_count -1
                    column.name.delimiter ,
                    columns name,food1,food2,food3,food4,id
                    columns.comments
```

```
                  transient_lastDdlTime 1644690440
                serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                name: mydb.foodratingspart
              name: mydb.foodratingspart
        Processor Tree:
          TableScan
            alias: foodratingspart
            GatherStats: false
            Select Operator
              expressions: food1 (type: int), food2 (type: int), food3 (type: int), food4 (type: int), id (type: int), name (type: string)
              outputColumnNames: _col0, _col1, _col2, _col3, _col4, _col5
              ListSink

INFO  : Completed compiling command(queryId=hive_20220212183420_ac52319c-0f30-4e90-a117-fc5780fadbfe); Time taken: 0.273 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220212183420_ac52319c-0f30-4e90-a117-fc5780fadbfe): Select * from foodratingspart
INFO  : Completed executing command(queryId=hive_20220212183420_ac52319c-0f30-4e90-a117-fc5780fadbfe); Time taken: 0.001 seconds
INFO  : OK
+-----------------------------+-----------------------------+-----------------------------+-----------------------------+-----------------------------+-----------------------------+
| foodratingspart.food1       | foodratingspart.food2       | foodratingspart.food3       | foodratingspart.food4       | foodratingspart.id          | foodratingspart.name        |
+-----------------------------+-----------------------------+-----------------------------+-----------------------------+-----------------------------+-----------------------------+
| 4                           | 28                          | 37                          | 3                           | 5                           | Jill                        |
| 34                          | 3                           | 15                          | 20                          | 1                           | Jill                        |
| 40                          | 15                          | 47                          | 28                          | 5                           | Jill                        |
| 31                          | 50                          | 15                          | 21                          | 2                           | Jill                        |
| 16                          | 32                          | 9                           | 32                          | 5                           | Jill                        |
| 24                          | 20                          | 4                           | 23                          | 4                           | Jill                        |
| 25                          | 28                          | 23                          | 37                          | 4                           | Jill                        |
| 18                          | 27                          | 11                          | 42                          | 3                           | Jill                        |
| 38                          | 8                           | 11                          | 48                          | 3                           | Jill                        |
| 25                          | 17                          | 24                          | 30                          | 2                           | Jill                        |
| 1                           | 6                           | 21                          | 18                          | 5                           | Jill                        |
| 36                          | 36                          | 5                           | 45                          | 4                           | Jill                        |
| 4                           | 33                          | 9                           | 45                          | 5                           | Jill                        |
| 9                           | 20                          | 11                          | 21                          | 1                           | Jill                        |
| 47                          | 16                          | 22                          | 22                          | 5                           | Jill                        |
| 44                          | 37                          | 22                          | 26                          | 2                           | Jill                        |
```

3. Calculating min, max and average of the values of the food2 column of MyDB.foodratingspart where the food critic 'name' is either Mel or Jill.

Command Used → SELECT min(food2) as min,max(food2) as max,avg(food2) as avg,name from foodratingspart where name='Mel' or name='Jill' GROUP BY name;

```
0: jdbc:hive2://localhost:10000/ (MyDb)> SELECT min(food2) as min,max(food2) as max,avg(food2) as avg,name from foodratingspart where name='Mel' or name='Jill' GROUP BY name;
INFO  : Compiling command(queryId=hive_20220929232444_8cfbdcf0-c3b3-40e6-a82e-bf3e97367b78): SELECT min(food2) as min,max(food2) as max,avg(food2) as avg,name from foodratingspart where name='Mel' or name
='Jill' GROUP BY name
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:min, type:int, comment:null), FieldSchema(name:max, type:int, comment:null), FieldSchema(name:avg, type:double, comment:null), FieldSch
ema(name:name, type:string, comment:null)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220929232444_8cfbdcf0-c3b3-40e6-a82e-bf3e97367b78 : STAGE DEPENDENCIES:
  Stage-1 is a root stage [MAPRED]
  Stage-0 depends on stages: Stage-1 [FETCH]

STAGE PLANS:
  Stage: Stage-1
    Tez
      DagId: hive_20220929232444_8cfbdcf0-c3b3-40e6-a82e-bf3e97367b78:5
      Edges:
        Reducer 2 <- Map 1 (SIMPLE_EDGE)
      DagName:
      Vertices:
        Map 1
            Map Operator Tree:
                TableScan
                  alias: foodratingspart
                  Statistics: Num rows: 389 Data size: 76362 Basic stats: COMPLETE Column stats: PARTIAL
                  GatherStats: false
                  Select Operator
                    expressions: name (type: string), food2 (type: int)

INFO  : Completed compiling command(queryId=hive_20220929232444_8cfbdcf0-c3b3-40e6-a82e-bf3e97367b78); Time taken: 1.529 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220929232444_8cfbdcf0-c3b3-40e6-a82e-bf3e97367b78): SELECT min(food2) as min,max(food2) as max,avg(food2) as avg,name from foodratingspart where name='Mel' or name
='Jill' GROUP BY name
INFO  : Query ID = hive_20220929232444_8cfbdcf0-c3b3-40e6-a82e-bf3e97367b78
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: SELECT min(food2) as min,max(food2) a...name(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1664489100889_0002)

INFO  : Map 1: 0/1      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: 1(+1)/2
INFO  : Map 1: 1/1      Reducer 2: 2/2
INFO  : Completed executing command(queryId=hive_20220929232444_8cfbdcf0-c3b3-40e6-a82e-bf3e97367b78); Time taken: 6.661 seconds
INFO  : OK
+------+------+---------------------+-------+
| min  | max  |         avg         | name  |
+------+------+---------------------+-------+
| 1    | 50   | 25.207650273224044  | Jill  |
| 1    | 50   | 24.233009708737864  | Mel   |
+------+------+---------------------+-------+
2 rows selected (8.223 seconds)
```

Exercise 7.

Load the foodplaces<.magic number>.txt fifile from your local fifile system into the foodplaces table

Command Used → LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces168337.txt' OWERWRITE INTO TABLE foodplaces;
LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces168337.txt' OWERWRITE INTO TABLE foodplaces;

```
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces168337.txt'
. . . . . . . . . . . . . . . . . . . .> OVERWRITE INTO TABLE foodplaces;
INFO  : Compiling command(queryId=hive_20220929232544_e79ce5dc-b0bc-4fa1-bfbb-8f5c9aca9c9e): LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces168337.txt'
OVERWRITE INTO TABLE foodplaces
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220929232544_e79ce5dc-b0bc-4fa1-bfbb-8f5c9aca9c9e : STAGE DEPENDENCIES:
  Stage-0 is a root stage [MOVE]
  Stage-1 depends on stages: Stage-0 [STATS]

STAGE PLANS:
  Stage: Stage-0
    Move Operator
      tables:
          replace: true
          source: file:/home/hadoop/foodplaces168337.txt
          table:
              input format: org.apache.hadoop.mapred.TextInputFormat
              output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
              properties:
                COLUMN_STATS_ACCURATE {"BASIC_STATS":"true"}
                bucket_count -1
                column.name.delimiter ,
                columns id,places
                columns.comments
                columns.types int:string
                field.delim ,
                file.inputformat org.apache.hadoop.mapred.TextInputFormat
                file.outputformat org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
                location hdfs://ip-172-31-31-40.ec2.internal:8020/user/hive/warehouse/mydb.db/foodplaces
                name mydb.foodplaces
                numFiles 0
                numRows 0
                rawDataSize 0
                serialization.ddl struct foodplaces { i32 id, string places}
                serialization.format ,
                serialization.lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                totalSize 0
                transient_lastDdlTime 1664491742
              serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
              name: mydb.foodplaces
```

Use a join opera-on between foodratings and foodplaces to provide the average ra-ng for field food4 for the restaurant 'Soup Bowl'
Command Used → SELECT fp.places, avg(fr.food4) avg FROM foodplaces fp JOIN foodratings fr ON (fr.id=fp.id) WHERE fp.places='Soup Bowl' GROUP BY fp.places;

```
INFO  : Executing command(queryId=hive_20220929232603_661de6c0-ece1-4087-8eb3-2e6082c20586): SELECT fp.places, avg(fr.food4) avg FROM foodplaces fp JOIN foodratings fr ON (fr.id=fp.id) WHERE fp.places='So
up Bowl' GROUP BY fp.places
INFO  : Query ID = hive_20220929232603_661de6c0-ece1-4087-8eb3-2e6082c20586
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Session is already open
INFO  : Dag name: SELECT fp.places, avg(fr.food4) ...fp.places(Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Status: Running (Executing on YARN cluster with App id application_1664489100889_0002)

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: 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_20220929232603_661de6c0-ece1-4087-8eb3-2e6082c20586); Time taken: 11.918 seconds
INFO  : OK
+------------+--------------------+
| fp.places  |        avg         |
+------------+--------------------+
| Soup Bowl  | 24.46629213483146  |
+------------+--------------------+
1 row selected (12.378 seconds)
```

Exercise 8.

a) Row data format is chosen when it requires to access all the columns in the row. Column format is used when we require data from only some columns.
b) Breaking down the data into smaller chunks which can be handled independently is called as split ability. It is used to process large data in an efficient manner, and it is done by breaking the job into smaller jobs for multiple processors.
c) Files stored in column format can achieve better compression than those stored in row format because the same type of data is stored next to each other, so it allows the user to be more efficient in compressing the data.
d) When we have datasets with many numbers of columns then Parquet is the best choice. In this each Parquet file contains binary data organized in “row groups”. Parquet is better choice to read-heavy workloads.