CSP554—BigData Technologies

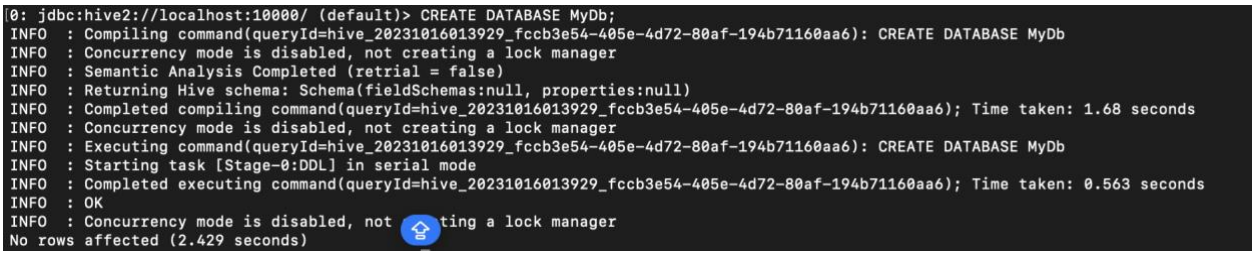
Assignment #4

Worth: 18 points

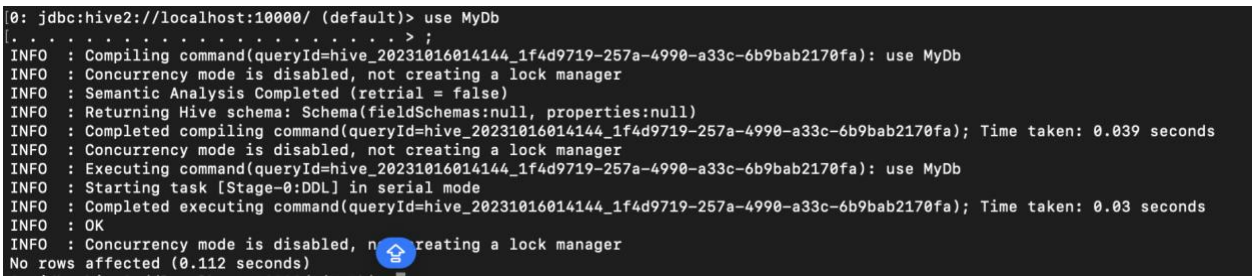
Magic Number Generation –

Magic Number: 42942

Exercise 1) Create a Hive database called “MyDb”.

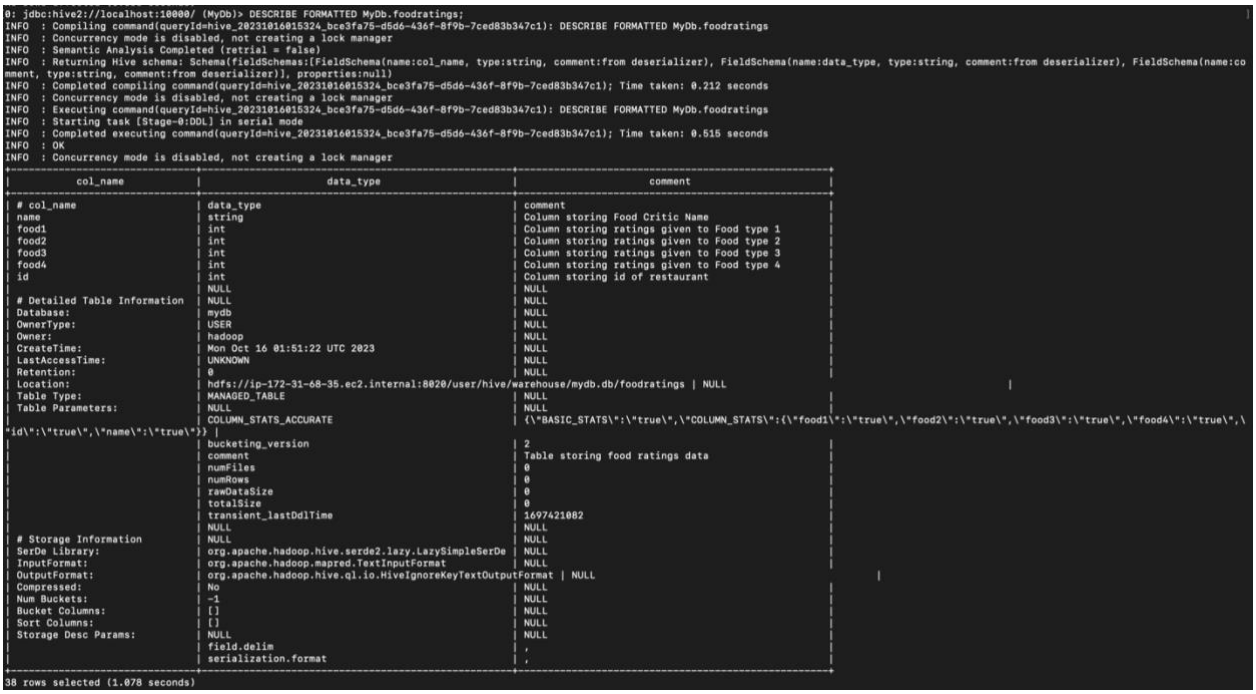


Use Database –



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

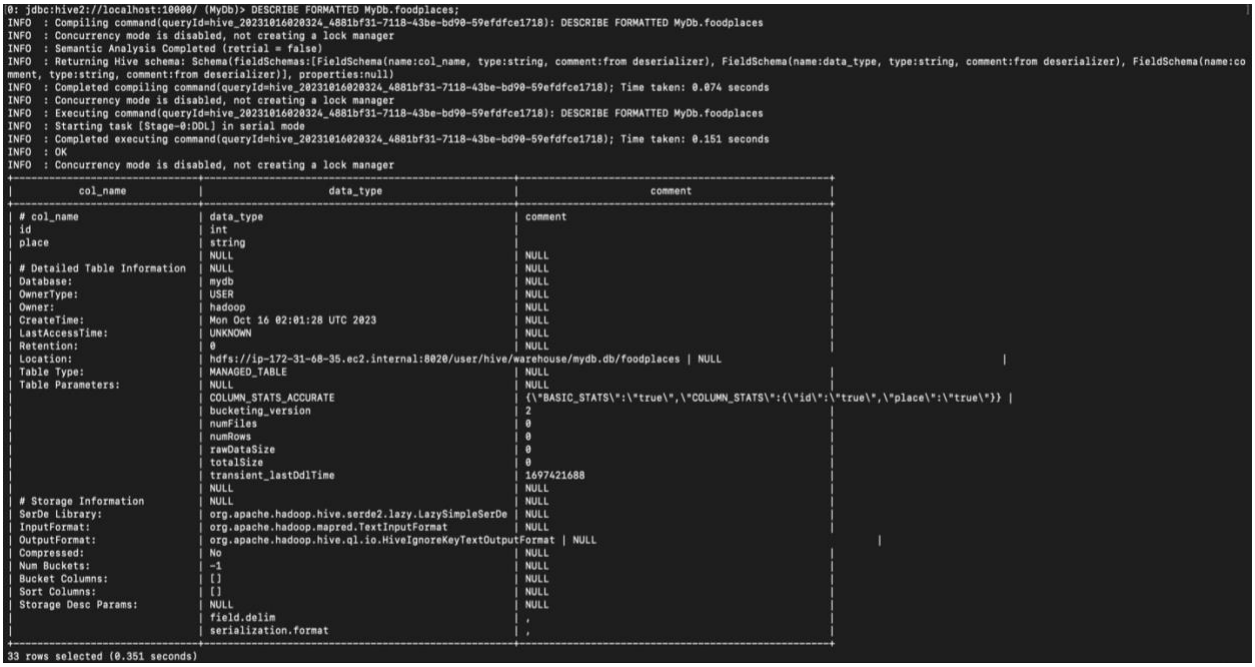
ANS:



```
0: jdbc:hive2://localhost:10000/ (MyDb)> DESCRIBE FORMATTED MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20231016015324_bce3fa75-d5d6-436f-8f9b-7ced83b347c1): 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:co
mment, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231016015324_bce3fa75-d5d6-436f-8f9b-7ced83b347c1); Time taken: 0.212 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231016015324_bce3fa75-d5d6-436f-8f9b-7ced83b347c1): DESCRIBE FORMATTED MyDb.foodratings
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20231016015324_bce3fa75-d5d6-436f-8f9b-7ced83b347c1); Time taken: 0.515 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                                             | Column storing Food Critic Name                    |
| food1                         | int                                                | Column storing ratings given to Food type 1        |
| food2                         | int                                                | Column storing ratings given to Food type 2        |
| food3                         | int                                                | Column storing ratings given to Food type 3        |
| food4                         | int                                                | Column storing ratings given to Food type 4        |
| id                            | int                                                | Column storing id of restaurant                    |
|                               | NULL                                               | NULL                                               |
| # Detailed Table Information  | NULL                                               | NULL                                               |
| Database:                     | mydb                                               | NULL                                               |
| OwnerType:                    | USER                                               | NULL                                               |
| Owner:                        | hadoop                                             | NULL                                               |
| CreateTime:                   | Mon Oct 16 01:51:22 UTC 2023                       | NULL                                               |
| LastAccessTime:               | UNKNOWN                                            | NULL                                               |
| Retention:                    | 0                                                  | NULL                                               |
| Location:                     | hdfs://ip-172-31-68-35.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\",\"COLUMN_STATS\":{\"food1\":\"true\",\"food2\":\"true\",\"food3\":\"true\",\"food4\":\"true\",\
"id\":\"true\",\"name\":\"true\"}} |
|                               | bucketing_version                                  | 2                                                  |
|                               | comment                                            | Table storing food ratings data                    |
|                               | numFiles                                           | 0                                                  |
|                               | numRows                                            | 0                                                  |
|                               | rawDataSize                                        | 0                                                  |
|                               | totalSize                                          | 0                                                  |
|                               | transient_lastDdlTime                              | 1697421082                                         |
|                               | 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 (1.078 seconds)
```

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

ANS:



```
0: jdbc:hive2://localhost:10000/ (MyDb)> DESCRIBE FORMATTED MyDb.foodplaces;
INFO  : Compiling command(queryId=hive_20231016020324_4881bf31-7118-43be-bd90-59efdfce1718): 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:co
mment, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231016020324_4881bf31-7118-43be-bd90-59efdfce1718); Time taken: 0.074 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231016020324_4881bf31-7118-43be-bd90-59efdfce1718): DESCRIBE FORMATTED MyDb.foodplaces
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20231016020324_4881bf31-7118-43be-bd90-59efdfce1718); Time taken: 0.151 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:                   | Mon Oct 16 02:01:28 UTC 2023                       | NULL                                               |
| LastAccessTime:               | UNKNOWN                                            | NULL                                               |
| Retention:                    | 0                                                  | NULL                                               |
| Location:                     | hdfs://ip-172-31-68-35.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\",\"COLUMN_STATS\":{\"id\":\"true\",\"place\":\"true\"}} |
|                               | bucketing_version                                  | 2                                                  |
|                               | numFiles                                           | 0                                                  |
|                               | numRows                                            | 0                                                  |
|                               | rawDataSize                                        | 0                                                  |
|                               | totalSize                                          | 0                                                  |
|                               | transient_lastDdlTime                              | 1697421688                                         |
|                               | 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 (0.351 seconds)
```

Exercise 2) 2 points

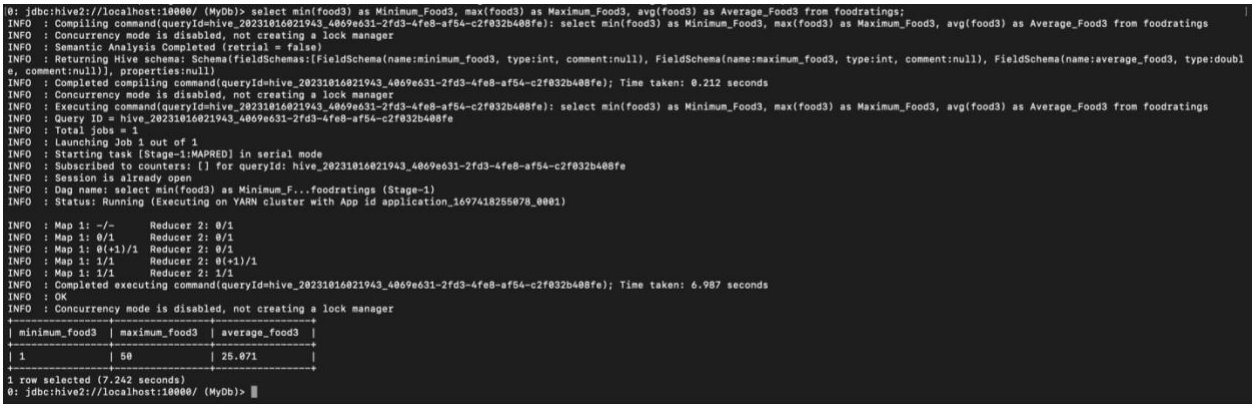
Load the foodratings<magic number>.txt file created using TestDataGen from your local file system into the foodratings table.

Execute a 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.

ANS:

select min(food3) as Minimum_Food3, max(food3) as Maximum_Food3, avg(food3) as Average_Food3
from foodratings;



Magic Number: 42942

Exercise 3) 2 points

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

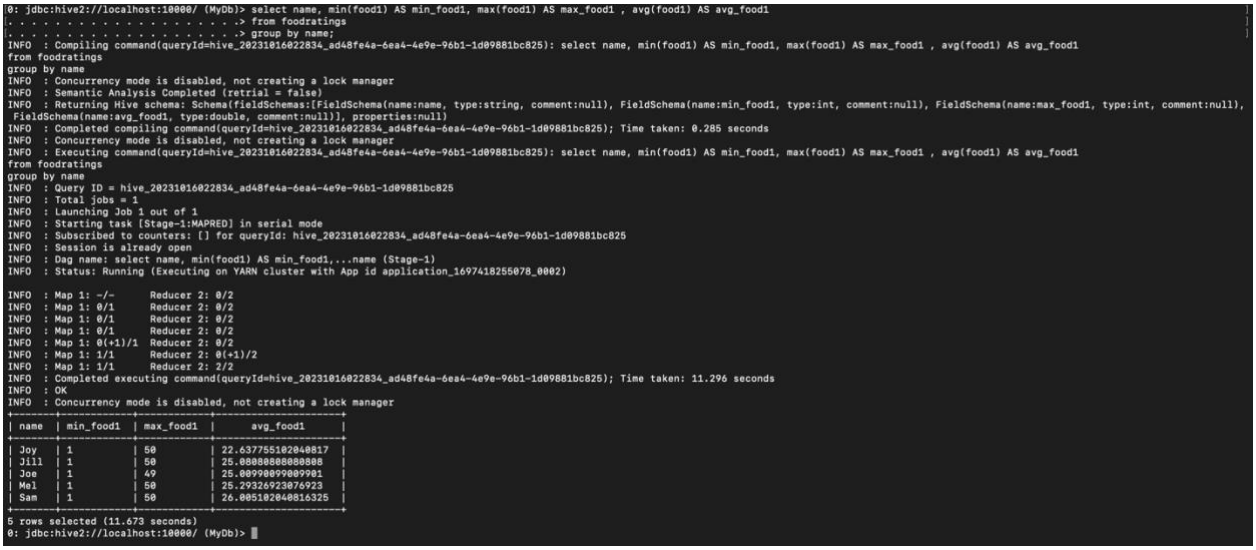
Bill 20, 30, 24

…

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

ANS:

select name, min(food1) AS min_food1, max(food1) AS max_food, avg(food1) AS avg_food1
from foodratings
group by name;



Magic Number: 42942

Exercise 4) 2 points

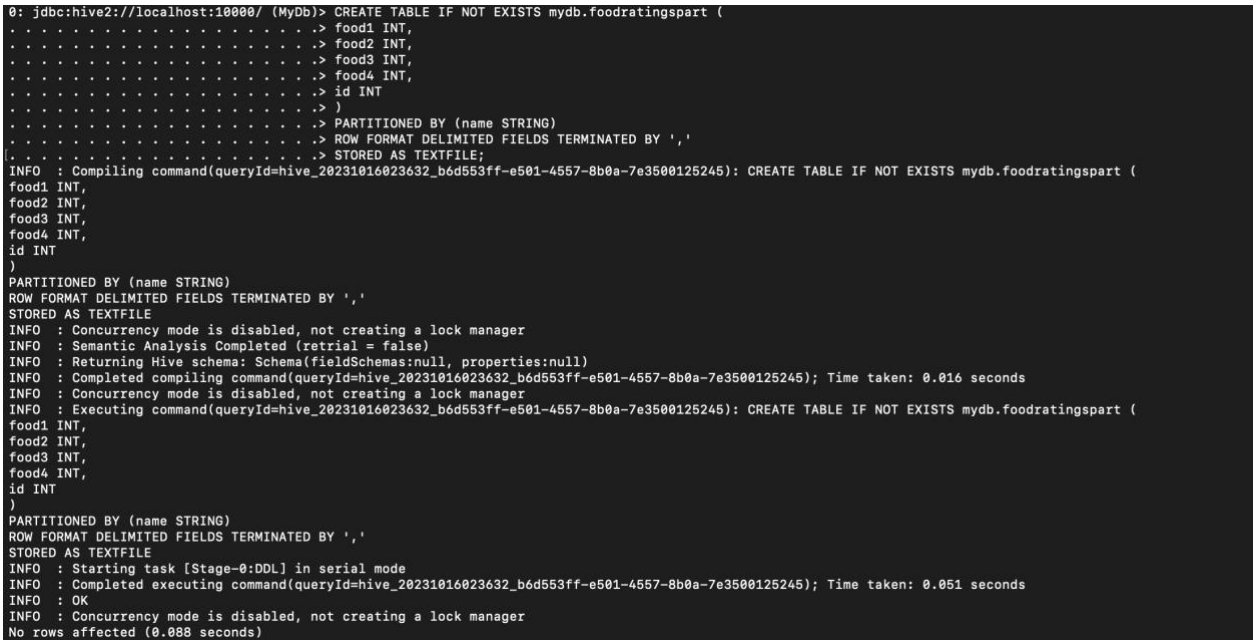
In MyDb create a partitioned table called ‘foodratingspart’

The partition field should be called ‘name’ and its type should be a string. The names of the nonpartition 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.

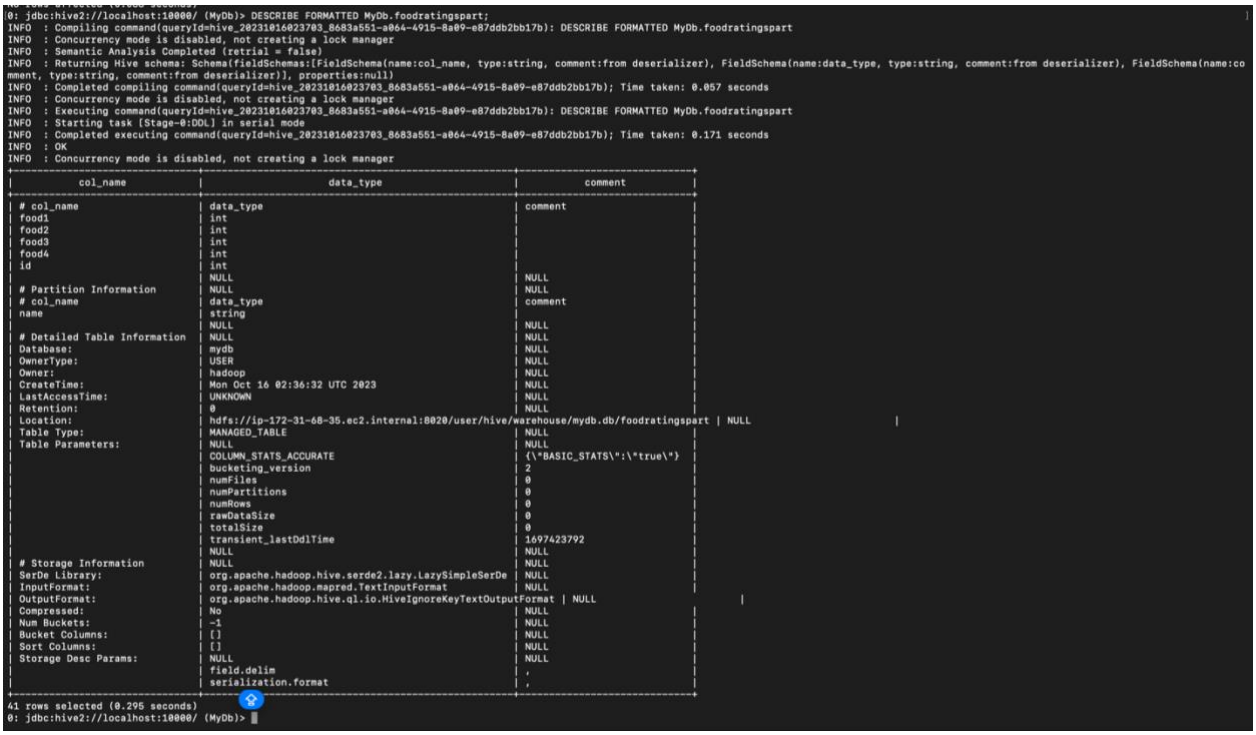
Execute a Hive command of ‘DESCRIBE FORMATTED MyDb.foodratingspart;’ and capture its output as the result of this exercise.

ANS:

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;



DESCRIBE FORMATTED MyDb.foodratingspart;



```
0: jdbc:hive2://localhost:10000/ (MyDb)> DESCRIBE FORMATTED MyDb.foodratingspart;
INFO  : Compiling command(queryId=hive_20231016023703_8683a551-a064-4915-8a09-e87ddb2bb17b): 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:co
mment, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231016023703_8683a551-a064-4915-8a09-e87ddb2bb17b); Time taken: 0.057 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231016023703_8683a551-a064-4915-8a09-e87ddb2bb17b): DESCRIBE FORMATTED MyDb.foodratingspart
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20231016023703_8683a551-a064-4915-8a09-e87ddb2bb17b); Time taken: 0.171 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:                   | Mon Oct 16 02:36:32 UTC 2023                       | NULL                        |
| LastAccessTime:               | UNKNOWN                                            | NULL                        |
| Retention:                    | 0                                                  | NULL                        |
| Location:                     | hdfs://ip-172-31-68-35.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\"}  |
|                               | bucketing_version                                  | 2                           |
|                               | numFiles                                           | 0                           |
|                               | numPartitions                                      | 0                           |
|                               | numRows                                            | 0                           |
|                               | rawDataSize                                        | 0                           |
|                               | totalSize                                          | 0                           |
|                               | transient_lastDdlTime                              | 1697423792                  |
|                               | 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 (0.295 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

Exercise 5) 2 points

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:

In a partitioned table, table is divided into segments, called partitions which makes it easier to manage and query data. Now since food critics name value is relatively small (less than 10) the table would be divided into as many partitions. So, when we want to retrieve the data, it will only investigate a particular partition, thereby reducing the retrieval time. It is not recommended to have a greater number of partitions as it creates number of files in HDFS, which may create overhead for NameNode since it maintains metadata.
With so many partitions created by partitioning by place id, organizing and accessing the data could become more difficult and time-consuming.

Exercise 6) 2 points

Configure Hive to allow dynamic partition creation. 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.

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. It should look something like

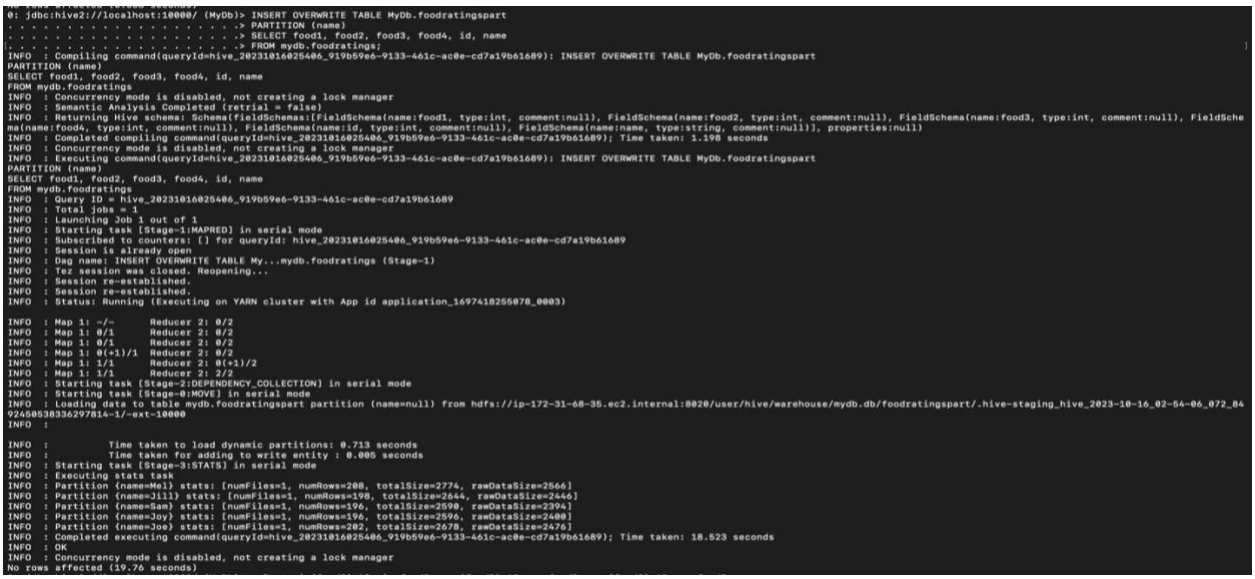
10 20 15

ANS:

set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrinct;

```
0: jdbc:hive2://localhost:10000/ (MyDb)> set hive.exec.dynamic.partition=true;
No rows affected (0.033 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> set hive.exec.dynamic.partition.mode=nonstrinct;
No rows affected (0.008 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

INSERT OVERWRITE TABLE MyDb.foodratingspart
PARTITION (name)
SELECT food1, food2, food3, food4, id, name
FROM mydb.foodratings;



select min(food2) AS min_food2, max(food2) AS max_food2, avg(food2) AS avg_food2
from MyDb.foodratingspart
where name = 'Mel' or name = 'Jill';

```
0: jdbc:hive2://localhost:10000/ (MyDb)> select min(food2) AS min_food2, max(food2) AS max_food2, avg(food2) AS avg_food2
. . . . . .> from MyDb.foodratingspart
. . . . . .> where name = 'Mel' or name = 'Jill';
INFO  : Compiling command(queryId=hive_20231016025705_9c95c199-245b-472a-938e-30dc1eed801c): select min(food2) AS min_food2, max(food2) AS max_food2, avg(food2) AS avg_food2
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_food2, type:int, comment:null), FieldSchema(name:max_food2, type:int, comment:null), FieldSchema(name:avg_food2, type:double, comment:n
ull)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231016025705_9c95c199-245b-472a-938e-30dc1eed801c); Time taken: 1.081 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231016025705_9c95c199-245b-472a-938e-30dc1eed801c): select min(food2) AS min_food2, max(food2) AS max_food2, avg(food2) AS avg_food2
from MyDb.foodratingspart
where name = 'Mel' or name = 'Jill'
INFO  : Query ID = hive_20231016025705_9c95c199-245b-472a-938e-30dc1eed801c
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_20231016025705_9c95c199-245b-472a-938e-30dc1eed801c
INFO  : Session is already open
INFO  : Dag name: select min(food2) AS min_food2, max...'Jill' (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1697418255078_0003)

INFO  : Map 1: -/-      Reducer 2: 0/1
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: 1/1      Reducer 2: 0(+1)/1
INFO  : Map 1: 1/1      Reducer 2: 1/1
INFO  : Completed executing command(queryId=hive_20231016025705_9c95c199-245b-472a-938e-30dc1eed801c); Time taken: 7.368 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------+------------+--------------------+
| min_food2  | max_food2  |     avg_food2      |
+------------+------------+--------------------+
| 1          | 50         | 26.58128078817734  |
+------------+------------+--------------------+
1 row selected (8.508 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

Exercise 7) 2 points

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

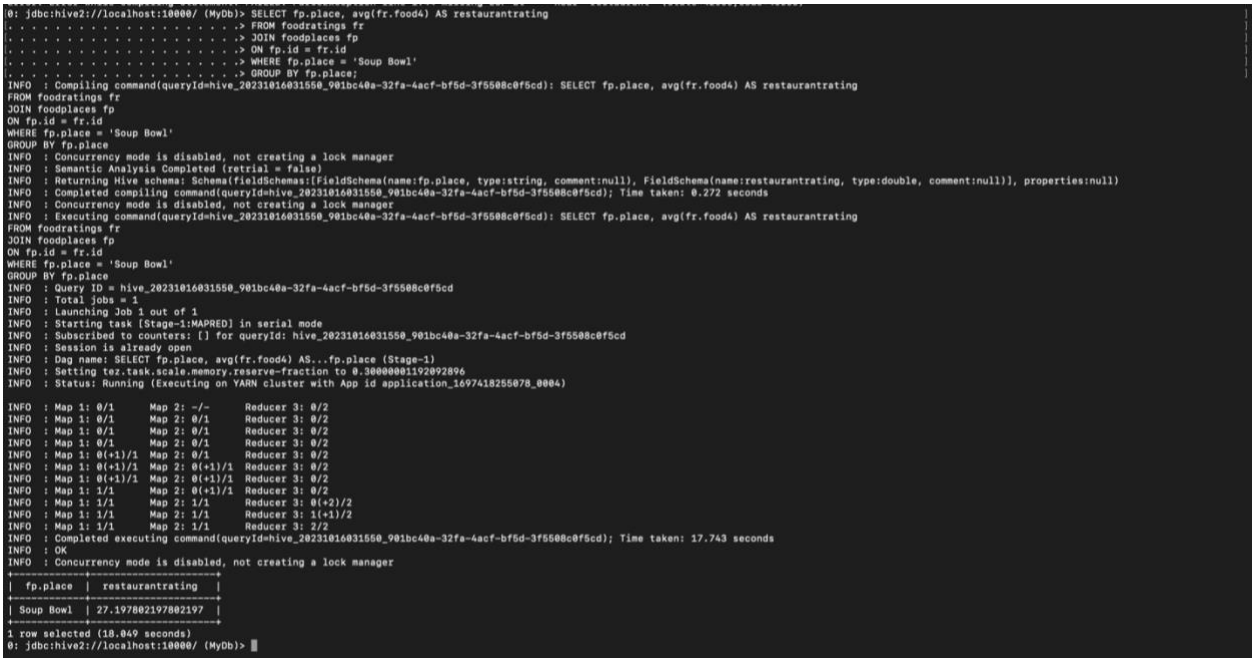
Use a join operation between the two tables (foodratings and foodplaces) to provide the average rating for field food4 for the restaurant ‘Soup Bowl’

The output of this query is the result of this exercise. It should look something like

Soup Bowl 20

ANS:

```
SELECT fp.place, avg(fr.food4) AS restaurantrating
FROM foodratings fr
JOIN foodplaces fp
ON fp.id = fr.id
WHERE fp.place = 'Soup Bowl'
GROUP BY fp.place;
```



Exercise 8)

a) **When is the most important consideration when choosing a row format and when a column format for your big data file?**
ANS:

When using a column format, the data is kept in columns, top to bottom. The column format is typically chosen when running analytics queries on extremely large data sets that only need a subset of the column. Because we can concentrate on a specific column of data thanks to a column format, the scan is more effective because we just read the pertinent values and avoid needless overhead associated with fetching columns that are not important for the outcome.
Row format is the most basic type of data table and is employed in numerous applications. The first row is read first as the data is read from left to right. The best time to use this format is when we need to access one or more entries and many or all columns.

b) **What is "splittability" for a column file format and why is it important when processing large volumes of data?**
ANS:

The term "splittability" describes a file's capacity to be divided up into smaller records or processes. These records are then processed parallel which in turn increases the performance. When the intended results of the query can be found in a single column, column-based formats are easier to split into separate jobs. A batch of rows is stored in a column format in a row-column file. The boundaries for data splitting are these batches.

c) **What can files stored in column format achieve better compression than those stored in row format?**
ANS:

Data that is based on columns can be compressed more effectively than data based on rows. When we store values in a column based format, we store same type of values next to each other. This allows/enables us apply better compression on the data than that of row based. E.G – Storing all dates (same data type) together in memory gives us efficient compression output as opposed to storing various data types next to each other like string, number, date, string etc.

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 analytical database. Parquet file allows for fast, one-pass writing and possess high compression & splitting benefits. Parquet file is proficient at analysing wide datasets with many columns. Parquet is a perfect choice when we are working with read-heavy workload data.