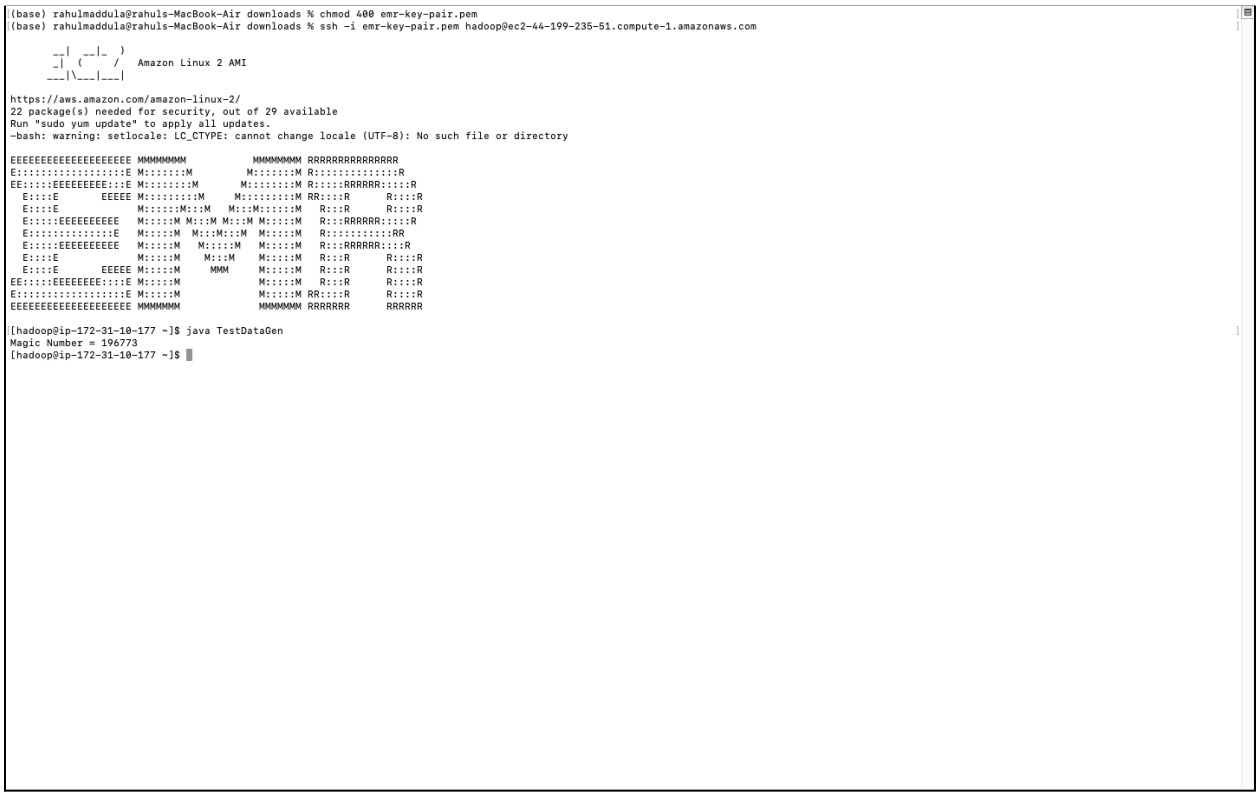
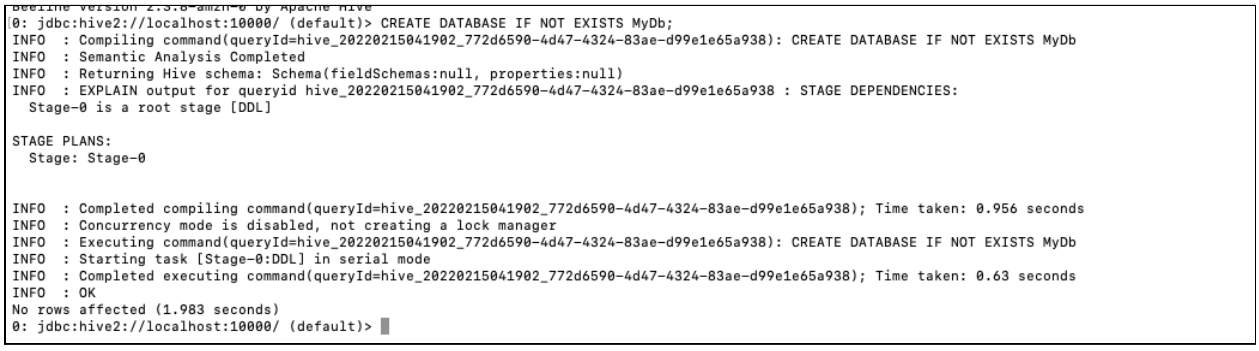
ASSIGNMENT 04



Magic number=196773

Exercise 1:



```
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE EXTERNAl TABLE IF NOT EXISTS MyDb.foodratings(name STRING COMMENT 'name',food1 INT COMMENT 'rating1',food2 INT COMMENT 'rating2',food3 INT COMMENT 'rating
3',food4 INT COMMENT 'rating4',id INT COMMENT 'ID') COMMENT 'THIS IS MY FOOD RATING TABLE' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/home/hadoop/foodratings196773.txt';

INFO  : Compiling command(queryId=hive_20220215043733_7644b398-260b-4ad1-8e08-ac5dc4dbad00): CREATE EXTERNAl TABLE IF NOT EXISTS MyDb.foodratings(name STRING COMMENT 'name',food1 INT COMMENT 'rating1',f
ood2 INT COMMENT 'rating2',food3 INT COMMENT 'rating3',food4 INT COMMENT 'rating4',id INT COMMENT 'ID') COMMENT 'THIS IS MY FOOD RATING TABLE' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEX
TFILE LOCATION '/home/hadoop/foodratings196773.txt'
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220215043733_7644b398-260b-4ad1-8e08-ac5dc4dbad00 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
      Create Table Operator:
        Create Table
          columns: name string name, food1 int rating1, food2 int rating2, food3 int rating3, food4 int rating4, id int ID
          comment: THIS IS MY FOOD RATING TABLE
          field delimiter: ,
          if not exists: true
          input format: org.apache.hadoop.mapred.TextInputFormat
          location: /home/hadoop/foodratings196773.txt
          output format: org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat
          serde name: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
          name: MyDb.foodratings
          isExternal: true


INFO  : Completed compiling command(queryId=hive_20220215043733_7644b398-260b-4ad1-8e08-ac5dc4dbad00); Time taken: 0.202 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215043733_7644b398-260b-4ad1-8e08-ac5dc4dbad00): CREATE EXTERNAl TABLE IF NOT EXISTS MyDb.foodratings(name STRING COMMENT 'name',food1 INT COMMENT 'rating1',f
ood2 INT COMMENT 'rating2',food3 INT COMMENT 'rating3',food4 INT COMMENT 'rating4',id INT COMMENT 'ID') COMMENT 'THIS IS MY FOOD RATING TABLE' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEX
TFILE LOCATION '/home/hadoop/foodratings196773.txt'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220215043733_7644b398-260b-4ad1-8e08-ac5dc4dbad00); Time taken: 0.288 seconds
INFO  : OK
No rows affected (0.543 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> DESCRIBE FORMATTED MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20220215043855_544296d6-5d9d-4bc3-9b5b-aaa986a848d0): DESCRIBE FORMATTED MyDb.foodratings
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(n
ame:comment, type:string, comment:from deserializer)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220215043855_544296d6-5d9d-4bc3-9b5b-aaa986a848d0 : 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/2bdead98-47ec-46ec-acae-2186b5e82908/hive_2022-02-15_04-38-55_735_1815178407136241554-1/-local-10000
          table: MyDb.foodratings

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


INFO  : Completed compiling command(queryId=hive_20220215043855_544296d6-5d9d-4bc3-9b5b-aaa986a848d0); Time taken: 0.303 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215043855_544296d6-5d9d-4bc3-9b5b-aaa986a848d0): DESCRIBE FORMATTED MyDb.foodratings
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220215043855_544296d6-5d9d-4bc3-9b5b-aaa986a848d0); Time taken: 0.159 seconds
INFO  : OK
+-------------------------------+----------------------------------------------------+-------------------------------+
|           col_name            |                     data_type                      |            comment            |
+-------------------------------+----------------------------------------------------+-------------------------------+
| # col_name                    | data_type                                          | comment                       |
|                               | NULL                                               | NULL                          |
| name                          | string                                             | name                          |
| food1                         | int                                                | rating1                       |
| food2                         | int                                                | rating2                       |
| food3                         | int                                                | rating3                       |
| food4                         | int                                                | rating4                       |
| id                            | int                                                | ID                            |
|                               | NULL                                               | NULL                          |
| # Detailed Table Information  | NULL                                               | NULL                          |
| Database:                     | mydb                                               | NULL                          |
| Owner:                        | hadoop                                             | NULL                          |
| CreateTime:                   | Tue Feb 15 04:37:33 UTC 2022                       | NULL                          |
| LastAccessTime:               | UNKNOWN                                            | NULL                          |
| Retention:                    | 0                                                  | NULL                          |
| Location:                     | hdfs://ip-172-31-10-177.ec2.internal:8020/home/hadoop/foodratings196773.txt | NULL                |
| Table Type:                   | EXTERNAL_TABLE                                     | NULL                          |
| Table Parameters:             | NULL                                               | NULL                          |
|                               | EXTERNAL                                           | TRUE                          |
|                               | comment                                            | THIS IS MY FOOD RATING TABLE  |
|                               | transient_lastDdlTime                              | 1644899853                    |
|                               | 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                          |
```

```
33 rows selected (0.82 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodplaces(id INT,place STRING) ROW FORMATTED DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/home/hadoop/foodplace196773.txt';
Error: Error while compiling statement: FAILED: ParseException line 1:77 cannot recognize input near 'ROW' 'FORMATTED' 'DELIMITED' in table row format specification (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodplaces(id INT,place STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/home/hadoop/foodplace196773.txt';
INFO  : Compiling command(queryId=hive_20220215044537_8d8d6bd8-f227-4d0e-9364-51b8bb1a15ca): CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodplaces(id INT,place STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/home/hadoop/foodplace196773.txt'
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220215044537_8d8d6bd8-f227-4d0e-9364-51b8bb1a15ca : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
      Create Table Operator:
        Create Table
          columns: id int, place string
          field delimiter: ,
          if not exists: true
          input format: org.apache.hadoop.mapred.TextInputFormat
          location: /home/hadoop/foodplace196773.txt
          output format: org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat
          serde name: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
          name: MyDb.foodplaces
          isExternal: true


INFO  : Completed compiling command(queryId=hive_20220215044537_8d8d6bd8-f227-4d0e-9364-51b8bb1a15ca); Time taken: 0.027 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215044537_8d8d6bd8-f227-4d0e-9364-51b8bb1a15ca): CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodplaces(id INT,place STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/home/hadoop/foodplace196773.txt'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220215044537_8d8d6bd8-f227-4d0e-9364-51b8bb1a15ca); Time taken: 0.078 seconds
INFO  : OK
No rows affected (0.136 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> DESCRIBE FORMATTED MyDb.foodpalaces;
Error: Error while compiling statement: FAILED: SemanticException [Error 10001]: Table not found MyDb.foodpalaces (state=42S02,code=10001)
0: jdbc:hive2://localhost:10000/ (MyDb)> DESCRIBE FORMATTED MyDb.foodplaces;
INFO  : Compiling command(queryId=hive_20220215044617_b1f04291-25e6-4a0f-bf75-cb28886697d4): DESCRIBE FORMATTED MyDb.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 deserializer), FieldSchema(name:comment, type:string, comment:from deserializer)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220215044617_b1f04291-25e6-4a0f-bf75-cb28886697d4 : 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/2bdead98-47ec-46ec-acae-2186b5e82908/hive_2022-02-15_04-46-17_767_5020893720628952765-1/-local-10000
          table: MyDb.foodplaces

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


INFO  : Completed compiling command(queryId=hive_20220215044617_b1f04291-25e6-4a0f-bf75-cb28886697d4); Time taken: 0.137 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> DESCRIBE FORMATTED MyDb.foodplaces;
INFO  : Compiling command(queryId=hive_20220215044617_b1f04291-25e6-4a0f-bf75-cb28886697d4): DESCRIBE FORMATTED MyDb.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 deserializer), FieldSchema(name:comment, type:string, comment:from deserializer)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220215044617_b1f04291-25e6-4a0f-bf75-cb28886697d4 : 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/2bdead98-47ec-46ec-acae-2186b5e82908/hive_2022-02-15_04-46-17_767_5020893720628952765-1/-local-10000
          table: MyDb.foodplaces

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


INFO  : Completed compiling command(queryId=hive_20220215044617_b1f04291-25e6-4a0f-bf75-cb28886697d4); Time taken: 0.137 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215044617_b1f04291-25e6-4a0f-bf75-cb28886697d4): DESCRIBE FORMATTED MyDb.foodplaces
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220215044617_b1f04291-25e6-4a0f-bf75-cb28886697d4); Time taken: 0.054 seconds
INFO  : OK
+-------------------------------+----------------------------------------------------+-------------------------+
|           col_name            |                     data_type                      |         comment         |
+-------------------------------+----------------------------------------------------+-------------------------+
| # col_name                    | data_type                                          | comment                 |
|                               | NULL                                               | NULL                    |
| id                            | int                                                |                         |
| place                         | string                                             |                         |
|                               | NULL                                               | NULL                    |
| # Detailed Table Information  | NULL                                               | NULL                    |
| Database:                     | mydb                                               | NULL                    |
| Owner:                        | hadoop                                             | NULL                    |
| CreateTime:                   | Tue Feb 15 04:45:37 UTC 2022                       | NULL                    |
| LastAccessTime:               | UNKNOWN                                            | NULL                    |
| Retention:                    | 0                                                  | NULL                    |
| Location:                     | hdfs://ip-172-31-10-177.ec2.internal:8020/home/hadoop/foodplace196773.txt | NULL                    |
| Table Type:                   | EXTERNAL_TABLE                                     | NULL                    |
| Table Parameters:             | NULL                                               | NULL                    |
|                               | EXTERNAL                                           | TRUE                    |
|                               | transient_lastDdlTime                              | 1644900337              |
|                               | 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                               | ,                       |
+-------------------------------+----------------------------------------------------+-------------------------+
28 rows selected (0.296 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

Commands Executed:

```
CREATE DATABASE IF NOT EXISTS MyDb;

USE MyDb;

CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodratings (
name STRING COMMENT 'name',
food1 INT COMMENT 'rating1',
food2 INT COMMENT 'rating2',
food3 INT COMMENT 'rating3',
food4 INT COMMENT 'rating4',
id INT  COMMENT 'ID')
COMMENT 'Ratings data'
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE;

DESCRIBE FORMATTED MyDb.foodratings;


CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodplaces (
id INT,
place STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE';

DESCRIBE FORMATTED MyDb.foodplaces;
```

Exercise 2:

```
STAGE PLANS:
  Stage: Stage-0


INFO  : Completed compiling command(queryId=hive_20220215050225_172d95ab-1e2f-4cb6-8d37-61364e92d568); Time taken: 0.07 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215050225_172d95ab-1e2f-4cb6-8d37-61364e92d568): USE MyDb
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220215050225_172d95ab-1e2f-4cb6-8d37-61364e92d568); Time taken: 0.038 seconds
INFO  : OK
No rows affected (0.321 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodratings196773.txt' OVERWRITE INTO TABLE MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20220215050320_dc81c106-1eca-48dd-8522-39bc7ce94447): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings196773.txt' OVERWRITE INTO TABLE MyDb.foodratings
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220215050320_dc81c106-1eca-48dd-8522-39bc7ce94447 : 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/foodratings196773.txt
          table:
              input format: org.apache.hadoop.mapred.TextInputFormat
              output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
              properties:
                EXTERNAL TRUE
                bucket_count -1
                column.name.delimiter ,
                columns name,food1,food2,food3,food4,id
                columns.comments 'name','rating1','rating2','rating3','rating4','ID'
                columns.types string:int:int:int:int:int
                comment THIS IS MY FOOD RATING TABLE
                field.delim ,
                file.inputformat org.apache.hadoop.mapred.TextInputFormat
                file.outputformat org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
                location hdfs://ip-172-31-10-177.ec2.internal:8020/home/hadoop/foodratings196773.txt
                name mydb.foodratings
                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
                transient_lastDdlTime 1644899853
              serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
              name: mydb.foodratings

  Stage: Stage-1
    Stats-Aggr Operator


INFO  : Completed compiling command(queryId=hive_20220215050320_dc81c106-1eca-48dd-8522-39bc7ce94447); Time taken: 0.056 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215050320_dc81c106-1eca-48dd-8522-39bc7ce94447): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings196773.txt' OVERWRITE INTO TABLE MyDb.foodratings
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodratings from file:/home/hadoop/foodratings196773.txt
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20220215050320_dc81c106-1eca-48dd-8522-39bc7ce94447); Time taken: 1.41 seconds
INFO  : OK
No rows affected (1.495 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> SELECT min(food3) AS MIN, max(food3) AS MAX, avg(food3) AS AVG from MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20220215050809_97bf40e0-6654-4376-a391-4aa8ebf41ade): SELECT min(food3) AS MIN, max(food3) AS MAX, avg(food3) AS AVG from MyDb.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:double, comment:null)], prope
rties:null)
INFO  : EXPLAIN output for queryid hive_20220215050809_97bf40e0-6654-4376-a391-4aa8ebf41ade : 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_20220215050809_97bf40e0-6654-4376-a391-4aa8ebf41ade:1
      Edges:
        Reducer 2 <- Map 1 (CUSTOM_SIMPLE_EDGE)
      DagName:
      Vertices:
        Map 1
            Map Operator Tree:
                TableScan
                  alias: foodratings
                  Statistics: Num rows: 4366 Data size: 17464 Basic stats: COMPLETE Column stats: NONE
                  GatherStats: false
                  Select Operator
                    expressions: food3 (type: int)
                    outputColumnNames: food3
                    Statistics: Num rows: 4366 Data size: 17464 Basic stats: COMPLETE Column stats: NONE
                    Group By Operator
                      aggregations: min(food3), max(food3), avg(food3)
                      mode: hash
                      outputColumnNames: _col0, _col1, _col2
                      Statistics: Num rows: 1 Data size: 84 Basic stats: COMPLETE Column stats: NONE
                      Reduce Output Operator
                        null sort order:
                        sort order:
                        Statistics: Num rows: 1 Data size: 84 Basic stats: COMPLETE Column stats: NONE
                        tag: -1
                        value expressions: _col0 (type: int), _col1 (type: int), _col2 (type: struct<count:bigint,sum:double,input:int>)
                        auto parallelism: false
            Path -> Alias:
              hdfs://ip-172-31-10-177.ec2.internal:8020/home/hadoop/foodratings196773.txt [foodratings]
            Path -> Partition:
              hdfs://ip-172-31-10-177.ec2.internal:8020/home/hadoop/foodratings196773.txt
                Partition
                  base file name: foodratings196773.txt
                  input format: org.apache.hadoop.mapred.TextInputFormat
                  output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
                  properties:
                    EXTERNAL TRUE
                    bucket_count -1
                    column.name.delimiter ,
                    columns name,food1,food2,food3,food4,id
                    columns.comments 'name','rating1','rating2','rating3','rating4','ID'
                    columns.types string:int:int:int:int:int
                    comment THIS IS MY FOOD RATING TABLE
                    field.delim ,
                    file.inputformat org.apache.hadoop.mapred.TextInputFormat
                    file.outputformat org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
                    location hdfs://ip-172-31-10-177.ec2.internal:8020/home/hadoop/foodratings196773.txt
                    name mydb.foodratings
                    numFiles 1
                    serialization.ddl struct foodratings { string name, i32 food1, i32 food2, i32 food3, i32 food4, i32 id}
                    serialization.format ,
```

```
        Reducer 2
            Needs Tagging: false
            Reduce Operator Tree:
              Group By Operator
                aggregations: min(VALUE._col0), max(VALUE._col1), avg(VALUE._col2)
                mode: mergepartial
                outputColumnNames: _col0, _col1, _col2
                Statistics: Num rows: 1 Data size: 84 Basic stats: COMPLETE Column stats: NONE
                File Output Operator
                  compressed: false
                  GlobalTableId: 0
                  directory: hdfs://ip-172-31-10-177.ec2.internal:8020/tmp/hive/hadoop/a406efdb-244a-45d1-8696-eba1ef351819/hive_2022-02-15_05-08-09_962_1169093120885675801-2/-mr-10001/.hive-staging_hiv
e_2022-02-15_05-08-09_962_1169093120885675801-2/-ext-10002
                  NumFilesPerFileSink: 1
                  Statistics: Num rows: 1 Data size: 84 Basic stats: COMPLETE Column stats: NONE
                  Stats Publishing Key Prefix: hdfs://ip-172-31-10-177.ec2.internal:8020/tmp/hive/hadoop/a406efdb-244a-45d1-8696-eba1ef351819/hive_2022-02-15_05-08-09_962_1169093120885675801-2/-mr-10001
/.hive-staging_hive_2022-02-15_05-08-09_962_1169093120885675801-2/-ext-10002/
                  table:
                      input format: org.apache.hadoop.mapred.SequenceFileInputFormat
                      output format: org.apache.hadoop.hive.ql.io.HiveSequenceFileOutputFormat
                      properties:
                        columns _col0,_col1,_col2
                        columns.types int:int:double
                        escape.delim \
                        hive.serialization.extend.additional.nesting.levels true
                        serialization.escape.crlf true
                        serialization.format 1
                        serialization.lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                      serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                TotalFiles: 1
                GatherStats: false
                MultiFileSpray: false

  Stage: Stage-0
    Fetch Operator
      limit: -1
      Processor Tree:
        ListSink


INFO  : Completed compiling command(queryId=hive_20220215050809_97bf40e0-6654-4376-a391-4aa8ebf41ade); Time taken: 3.208 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215050809_97bf40e0-6654-4376-a391-4aa8ebf41ade): SELECT min(food3) AS MIN, max(food3) AS MAX, avg(food3) AS AVG from MyDb.foodratings
INFO  : Query ID = hive_20220215050809_97bf40e0-6654-4376-a391-4aa8ebf41ade
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: SELECT min(food3) AS MIN,...MyDb.foodratings(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1644897119306_0001)

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_20220215050809_97bf40e0-6654-4376-a391-4aa8ebf41ade); Time taken: 23.256 seconds
INFO  : OK
+------+------+---------+
| min  | max  |   avg   |
+------+------+---------+
| 1    | 50   | 24.467  |
+------+------+---------+
```

Command Executed:

Magic number-196773

LOAD DATA LOCAL INPATH '/home/hadoop/foodratings196773.txt' OVERWRITE INTO TABLE MyDb.foodratings;

SELECT min(food3) AS MIN, max(food3) AS MAX, avg(food3) AS AVG
from MyDb.foodratings;

Exercise 3:

```
1 row selected (26.704 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> SELECT name, min(food1) AS MIN, max(food1) AS MAX ,avg(food1) AS AVG from MyDb.foodratings GROUP BY name;
INFO  : Compiling command(queryId=hive_20220215051206_c70fb1f3-75a5-4fd6-8213-f86cb603d1f7): SELECT name, min(food1) AS MIN, max(food1) AS MAX ,avg(food1) AS AVG from MyDb.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), Field
Schema(name:avg, type:double, comment:null)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220215051206_c70fb1f3-75a5-4fd6-8213-f86cb603d1f7 : 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_20220215051206_c70fb1f3-75a5-4fd6-8213-f86cb603d1f7:2
      Edges:
        Reducer 2 <- Map 1 (SIMPLE_EDGE)
      DagName:
      Vertices:
        Map 1
            Map Operator Tree:
                TableScan
                  alias: foodratings
                  Statistics: Num rows: 167 Data size: 17464 Basic stats: COMPLETE Column stats: NONE
                  GatherStats: false
                  Select Operator
                    expressions: name (type: string), food1 (type: int)
                    outputColumnNames: name, food1
                    Statistics: Num rows: 167 Data size: 17464 Basic stats: COMPLETE Column stats: NONE
                    Group By Operator
                      aggregations: min(food1), max(food1), avg(food1)
                      keys: name (type: string)
                      mode: hash
                      outputColumnNames: _col0, _col1, _col2, _col3
                      Statistics: Num rows: 167 Data size: 17464 Basic stats: COMPLETE Column stats: NONE
                      Reduce Output Operator
                        key expressions: _col0 (type: string)
                        null sort order: a
                        sort order: +
                        Map-reduce partition columns: _col0 (type: string)
                        Statistics: Num rows: 167 Data size: 17464 Basic stats: COMPLETE Column stats: NONE
                        tag: -1
                        value expressions: _col1 (type: int), _col2 (type: int), _col3 (type: struct<count:bigint,sum:double,input:int>)
                        auto parallelism: true
            Path -> Alias:
              hdfs://ip-172-31-10-177.ec2.internal:8020/home/hadoop/foodratings196773.txt [foodratings]
            Path -> Partition:
              hdfs://ip-172-31-10-177.ec2.internal:8020/home/hadoop/foodratings196773.txt
                Partition
                  base file name: foodratings196773.txt
                  input format: org.apache.hadoop.mapred.TextInputFormat
                  output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
                  properties:
                    EXTERNAL TRUE
                    bucket_count -1
                    column.name.delimiter ,
                    columns name,food1,food2,food3,food4,id
                    columns.comments 'name','rating1','rating2','rating3','rating4','ID'
                    columns.types string:int:int:int:int:int
                    comment THIS IS MY FOOD RATING TABLE
                    field.delim ,
                    file.inputformat org.apache.hadoop.mapred.TextInputFormat
                    file.outputformat org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
                    location hdfs://ip-172-31-10-177.ec2.internal:8020/home/hadoop/foodratings196773.txt
```

```
                outputColumnNames: _col0, _col1, _col2, _col3
                Statistics: Num rows: 83 Data size: 8679 Basic stats: COMPLETE Column stats: NONE
                File Output Operator
                  compressed: false
                  GlobalTableId: 0
                  directory: hdfs://ip-172-31-10-177.ec2.internal:8020/tmp/hive/hadoop/a406efdb-244a-45d1-8696-eba1ef351819/hive_2022-02-15_05-12-06_203_6786773188220802165-2/-mr-10001/.hive-staging_hiv
e_2022-02-15_05-12-06_203_6786773188220802165-2/-ext-10002
                  NumFilesPerFileSink: 1
                  Statistics: Num rows: 83 Data size: 8679 Basic stats: COMPLETE Column stats: NONE
                  Stats Publishing Key Prefix: hdfs://ip-172-31-10-177.ec2.internal:8020/tmp/hive/hadoop/a406efdb-244a-45d1-8696-eba1ef351819/hive_2022-02-15_05-12-06_203_6786773188220802165-2/-mr-10001
/.hive-staging_hive_2022-02-15_05-12-06_203_6786773188220802165-2/-ext-10002/
                  table:
                      input format: org.apache.hadoop.mapred.SequenceFileInputFormat
                      output format: org.apache.hadoop.hive.ql.io.HiveSequenceFileOutputFormat
                      properties:
                        columns _col0,_col1,_col2,_col3
                        columns.types string:int:int:double
                        escape.delim \
                        hive.serialization.extend.additional.nesting.levels true
                        serialization.escape.crlf true
                        serialization.format 1
                        serialization.lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                      serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                TotalFiles: 1
                GatherStats: false
                MultiFileSpray: false

  Stage: Stage-0
    Fetch Operator
      limit: -1
      Processor Tree:
        ListSink


INFO  : Completed compiling command(queryId=hive_20220215051206_c70fb1f3-75a5-4fd6-8213-f86cb603d1f7); Time taken: 0.34 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215051206_c70fb1f3-75a5-4fd6-8213-f86cb603d1f7): SELECT name, min(food1) AS MIN, max(food1) AS MAX ,avg(food1) AS AVG from MyDb.foodratings GROUP BY name
INFO  : Query ID = hive_20220215051206_c70fb1f3-75a5-4fd6-8213-f86cb603d1f7
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 name, min(food1) AS MIN, max(f...name(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1644897119306_0001)

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(+0)/2
INFO  : Map 1: 1/1      Reducer 2: 2/2
INFO  : Completed executing command(queryId=hive_20220215051206_c70fb1f3-75a5-4fd6-8213-f86cb603d1f7); Time taken: 6.613 seconds
INFO  : OK
+-------+------+------+---------------------+
| name  | min  | max  |         avg         |
+-------+------+------+---------------------+
| Jill  | 1    | 50   | 24.328205128205127  |
| Joe   | 1    | 49   | 24.402010050251256  |
| Joy   | 1    | 50   | 26.746641148325359  |
| Mel   | 1    | 50   | 25.265402843601894  |
| Sam   | 1    | 50   | 26.086021505376344  |
+-------+------+------+---------------------+
5 rows selected (7.042 seconds)
```

Executed Commands:

Magic number-196773

SELECT name, min(food1) AS MIN, max(food1) AS MAX, avg(food1) AS AVG
from MyDb.foodratings
GROUP BY name;

Exercise 4:

```
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodratingspart(food1 INT,food2 INT,food3 INT,food4 INT,id INT) PARTITIONED BY(name STRING) ROW FORMAT DELIMITED FIELDS
TERMINATED BY ',' STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20220215052949_6d90b7f7-c537-4a7d-82a3-b24cbeb49fb7): CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodratingspart(food1 INT,food2 INT,food3 INT,food4 INT,id INT) PART
ITIONED BY(name STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220215052949_6d90b7f7-c537-4a7d-82a3-b24cbeb49fb7 : 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: ,
          if not exists: true
          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
          isExternal: true


INFO  : Completed compiling command(queryId=hive_20220215052949_6d90b7f7-c537-4a7d-82a3-b24cbeb49fb7); Time taken: 0.027 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215052949_6d90b7f7-c537-4a7d-82a3-b24cbeb49fb7): CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodratingspart(food1 INT,food2 INT,food3 INT,food4 INT,id INT) PART
ITIONED BY(name STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220215052949_6d90b7f7-c537-4a7d-82a3-b24cbeb49fb7); Time taken: 0.064 seconds
INFO  : OK
No rows affected (0.112 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> DESCRIBE FORMATTED MyDb.foodratingspart;
INFO  : Compiling command(queryId=hive_20220215053255_bcd52f21-91c8-46bc-b2e5-908ec4351560): DESCRIBE FORMATTED MyDb.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(n
ame:comment, type:string, comment:from deserializer)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220215053255_bcd52f21-91c8-46bc-b2e5-908ec4351560 : 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/a406efdb-244a-45d1-8696-eba1ef351819/hive_2022-02-15_05-32-55_396_8105455780155810296-2/-local-10000
          table: MyDb.foodratingspart

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


INFO  : Completed compiling command(queryId=hive_20220215053255_bcd52f21-91c8-46bc-b2e5-908ec4351560); Time taken: 0.068 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215053255_bcd52f21-91c8-46bc-b2e5-908ec4351560): DESCRIBE FORMATTED MyDb.foodratingspart
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20220215053255_bcd52f21-91c8-46bc-b2e5-908ec4351560); Time taken: 0.105 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:                   | Tue Feb 15 05:29:49 UTC 2022                       | NULL                        |
| LastAccessTime:               | UNKNOWN                                            | NULL                        |
| Retention:                    | 0                                                  | NULL                        |
| Location:                     | hdfs://ip-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart | NULL                        |
| Table Type:                   | EXTERNAL_TABLE                                     | NULL                        |
| Table Parameters:             | NULL                                               | NULL                        |
|                               | COLUMN_STATS_ACCURATE                              | {\"BASIC_STATS\":\"true\"}  |
|                               | EXTERNAL                                           | TRUE                        |
|                               | numFiles                                           | 0                           |
|                               | numPartitions                                      | 0                           |
|                               | numRows                                            | 0                           |
|                               | rawDataSize                                        | 0                           |
|                               | totalSize                                          | 0                           |
|                               | transient_lastDdlTime                              | 1644902989                  |
|                               | NULL                                               | NULL                        |
```

Commands Executed:

CREATE EXTERNAL 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;

Exercise 5:

It is the best way to choose the critic name as the partition field rather than using place id, with some operations using the place id makes the operation more expensive than the critic name. A string will take a byte of space for a single character, ID on the other hand is, we have 10,000 places, Int takes around 4 bytes and since we have 10,000 ID's that is way less memory when we store ID and partition based on name instead of partitioning based on ID which stores critic's names 10,000 times which is 10,000 files. If we partitioned based on ID then , to perform any function every 10,000 files must be checked which creates a pointless option using the partitioning of tables.

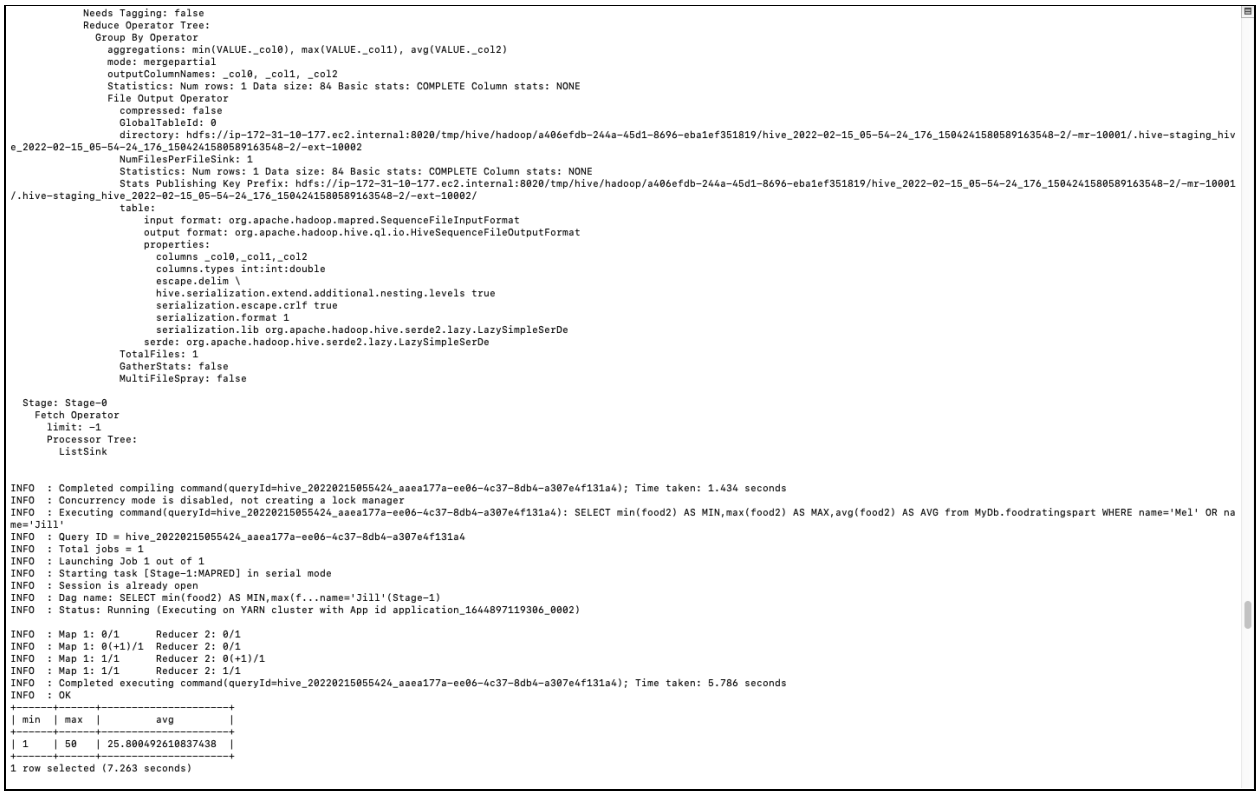
Exercise 6:

```
0: jdbc:hive2://localhost:10000/ (MyDb)> SET hive.exec.dynamic.partition=true;
No rows affected (0.01 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> SET hive.exec.dynamic.partition.mode=non-strict;
No rows affected (0.005 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> INSERT OVERWRITE TABLE MyDb.foodrratingspart PARTITION (name) SELECT food1,food2,food3,food4,id,name FROM MyDb.foodratings;
Error: Error while compiling statement: FAILED: SemanticException [Error 10001]: Line 1:23 Table not found 'foodrratingspart' (state=42S02,code=10001)
0: jdbc:hive2://localhost:10000/ (MyDb)> INSERT OVERWRITE TABLE MyDb.foodratingspart PARTITION (name) SELECT food1,food2,food3,food4,id,name FROM MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20220215055001_24bf3582-aada-4523-b7ad-59b8c33fc311): INSERT OVERWRITE TABLE MyDb.foodratingspart PARTITION (name) SELECT food1,food2,food3,food4,id,name FROM MyDb
.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), Fie
ldSchema(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_20220215055001_24bf3582-aada-4523-b7ad-59b8c33fc311 : 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_20220215055001_24bf3582-aada-4523-b7ad-59b8c33fc311:3
      DagName:
      Vertices:
        Map 1
            Map Operator Tree:
            TableScan
              alias: foodratings
              Statistics: Num rows: 145 Data size: 17464 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: 17464 Basic stats: COMPLETE Column stats: NONE
                File Output Operator
                  compressed: false
                  GlobalTableId: 1
                  directory: hdfs://ip-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/.hive-staging_hive_2022-02-15_05-50-01_340_6964987786179967154-2/-ext-10000
                  NumFilesPerFileSink: 1
                  Statistics: Num rows: 145 Data size: 17464 Basic stats: COMPLETE Column stats: NONE
                  Stats Publishing Key Prefix: hdfs://ip-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/.hive-staging_hive_2022-02-15_05-50-01_340_6964987786179967154-2/
-ext-10000/
                  table:
                      input format: org.apache.hadoop.mapred.TextInputFormat
                      output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
                      properties:
                        EXTERNAL TRUE
                        bucket_count -1
                        column.name.delimiter ,
                        columns food1,food2,food3,food4,id
                        columns.comments
                        columns.types 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-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart
                        name mydb.foodratingspart
                        partition_columns name
                        partition_columns.types string
                        serialization.ddl struct foodratingspart { i32 food1, i32 food2, i32 food3, i32 food4, i32 id}
                        serialization.format ,
                        serialization.lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                        transient_lastDdlTime 1644902989
                      serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                      name: mydb.foodratingspart
                  TotalFiles: 1
                  GatherStats: true
```

```
          partition:
            name
          replace: true
          source: hdfs://ip-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/.hive-staging_hive_2022-02-15_05-50-01_340_6964987786179967154-2/-ext-10000
          table:
              input format: org.apache.hadoop.mapred.TextInputFormat
              output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
              properties:
                EXTERNAL TRUE
                bucket_count -1
                column.name.delimiter ,
                columns food1,food2,food3,food4,id
                columns.comments
                columns.types 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-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart
                name mydb.foodratingspart
                partition_columns name
                partition_columns.types string
                serialization.ddl struct foodratingspart { i32 food1, i32 food2, i32 food3, i32 food4, i32 id}
                serialization.format ,
                serialization.lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                transient_lastDdlTime 1644902989
              serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
              name: mydb.foodratingspart

  Stage: Stage-3
    Stats-Aggr Operator
      Stats Aggregation Key Prefix: hdfs://ip-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/.hive-staging_hive_2022-02-15_05-50-01_340_6964987786179967154-2/-ext-10000/


INFO  : Completed compiling command(queryId=hive_20220215055001_24bf3582-aada-4523-b7ad-59b8c33fc311); Time taken: 0.362 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215055001_24bf3582-aada-4523-b7ad-59b8c33fc311): INSERT OVERWRITE TABLE MyDb.foodratingspart PARTITION (name) SELECT food1,food2,food3,food4,id,name FROM MyDb
.foodratings
INFO  : Query ID = hive_20220215055001_24bf3582-aada-4523-b7ad-59b8c33fc311
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: INSERT OVERWRITE TABLE My...MyDb.foodratings(Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1644897119306_0002)

INFO  : Map 1: 0/1
INFO  : Map 1: 0/1
INFO  : Map 1: 0(+1)/1
INFO  : Map 1: 1/1
INFO  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodratingspart partition (name=null) from hdfs://ip-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/.hive-staging_hive_2022-02-15_05-50-01
_340_6964987786179967154-2/-ext-10000
INFO  :

INFO  :          Time taken to load dynamic partitions: 0.537 seconds
INFO  :          Time taken for adding to write entity : 0.002 seconds
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20220215055001_24bf3582-aada-4523-b7ad-59b8c33fc311); Time taken: 14.827 seconds
INFO  : OK
No rows affected (15.202 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> SELECT min(food2) AS MIN,max(food2) AS MAX,avg(food2) AS AVG from MyDb.foodratingspart WHERE name='Mel' OR name='Jill';
INFO  : Compiling command(queryId=hive_20220215055424_aaea177a-ee06-4c37-8db4-a307e4f131a4): SELECT min(food2) AS MIN,max(food2) AS MAX,avg(food2) AS AVG from MyDb.foodratingspart WHERE name='Mel' OR na
me='Jill'
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)], prope
rties:null)
INFO  : EXPLAIN output for queryid hive_20220215055424_aaea177a-ee06-4c37-8db4-a307e4f131a4 : 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_20220215055424_aaea177a-ee06-4c37-8db4-a307e4f131a4:4
      Edges:
        Reducer 2 <- Map 1 (CUSTOM_SIMPLE_EDGE)
      DagName:
      Vertices:
        Map 1
            Map Operator Tree:
                TableScan
                  alias: foodratingspart
                  Statistics: Num rows: 406 Data size: 8212 Basic stats: COMPLETE Column stats: NONE
                  GatherStats: false
                  Select Operator
                    expressions: food2 (type: int)
                    outputColumnNames: food2
                    Statistics: Num rows: 406 Data size: 8212 Basic stats: COMPLETE Column stats: NONE
                    Group By Operator
                      aggregations: min(food2), max(food2), avg(food2)
                      mode: hash
                      outputColumnNames: _col0, _col1, _col2
                      Statistics: Num rows: 1 Data size: 84 Basic stats: COMPLETE Column stats: NONE
                      Reduce Output Operator
                        null sort order:
                        sort order:
                        Statistics: Num rows: 1 Data size: 84 Basic stats: COMPLETE Column stats: NONE
                        tag: -1
                        value expressions: _col0 (type: int), _col1 (type: int), _col2 (type: struct<count:bigint,sum:double,input:int>)
                        auto parallelism: false
            Path -> Alias:
              hdfs://ip-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/name=Jill [foodratingspart]
              hdfs://ip-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/name=Mel [foodratingspart]
            Path -> Partition:
              hdfs://ip-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/name=Jill
                Partition
                  base file name: name=Jill
                  input format: org.apache.hadoop.mapred.TextInputFormat
                  output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
                  partition values:
                    name Jill
                  properties:
                    COLUMN_STATS_ACCURATE {"BASIC_STATS":"true"}
                    bucket_count -1
                    column.name.delimiter ,
                    columns food1,food2,food3,food4,id
                    columns.comments
                    columns.types 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-172-31-10-177.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/name=Jill
                    name mydb.foodratingspart
```



```
           Needs Tagging: false
           Reduce Operator Tree:
             Group By Operator
                aggregations: min(VALUE._col0), max(VALUE._col1), avg(VALUE._col2)
                mode: mergepartial
                outputColumnNames: _col0, _col1, _col2
                Statistics: Num rows: 1 Data size: 84 Basic stats: COMPLETE Column stats: NONE
                File Output Operator
                  compressed: false
                  GlobalTableId: 0
                  directory: hdfs://ip-172-31-10-177.ec2.internal:8020/tmp/hive/hadoop/a406efdb-244a-45d1-8696-eba1ef351819/hive_2022-02-15_05-54-24_176_1504241580589163548-2/-mr-10001/.hive-staging_hiv
e_2022-02-15_05-54-24_176_1504241580589163548-2/-ext-10002
                  NumFilesPerFileSink: 1
                  Statistics: Num rows: 1 Data size: 84 Basic stats: COMPLETE Column stats: NONE
                  Stats Publishing Key Prefix: hdfs://ip-172-31-10-177.ec2.internal:8020/tmp/hive/hadoop/a406efdb-244a-45d1-8696-eba1ef351819/hive_2022-02-15_05-54-24_176_1504241580589163548-2/-mr-10001
/.hive-staging_hive_2022-02-15_05-54-24_176_1504241580589163548-2/-ext-10002/
                  table:
                      input format: org.apache.hadoop.mapred.SequenceFileInputFormat
                      output format: org.apache.hadoop.hive.ql.io.HiveSequenceFileOutputFormat
                      properties:
                        columns _col0,_col1,_col2
                        columns.types int:int:double
                        escape.delim \
                        hive.serialization.extend.additional.nesting.levels true
                        serialization.escape.crlf true
                        serialization.format 1
                        serialization.lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                      serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                TotalFiles: 1
                GatherStats: false
                MultiFileSpray: false

  Stage: Stage-0
    Fetch Operator
      limit: -1
      Processor Tree:
        ListSink


INFO  : Completed compiling command(queryId=hive_20220215055424_aaea177a-ee06-4c37-8db4-a307e4f131a4); Time taken: 1.434 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215055424_aaea177a-ee06-4c37-8db4-a307e4f131a4): SELECT min(food2) AS MIN,max(food2) AS MAX,avg(food2) AS AVG from MyDb.foodratingspart WHERE name='Mel' OR na
me='Jill'
INFO  : Query ID = hive_20220215055424_aaea177a-ee06-4c37-8db4-a307e4f131a4
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(f...name='Jill'(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1644897119306_0002)

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_20220215055424_aaea177a-ee06-4c37-8db4-a307e4f131a4); Time taken: 5.786 seconds
INFO  : OK
+------+------+---------------------+
| min  | max  |         avg         |
+------+------+---------------------+
| 1    | 50   | 25.800492610837438  |
+------+------+---------------------+
1 row selected (7.263 seconds)
```

Executed Commands:

SET hive.exec.dynamic.partition=true;
SET hive.exec.dynamic.partition.mode=nonstrict;

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

SELECT min(food2) AS MIN, max(food2) AS MAX, avg(food2) AS AVG
from MyDb.foodratingspart
WHERE name = ‘Mel’ OR name=’Jill’;

Exercise 7:

```
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces196773.txt' OVERWRITE INTO TABLE MyDb.foodplaces;
INFO  : Compiling command(queryId=hive_20220215055716_7b9dea27-e0c9-47e7-82b5-7b36fad17bab): LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces196773.txt' OVERWRITE INTO TABLE MyDb.foodplaces
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20220215055716_7b9dea27-e0c9-47e7-82b5-7b36fad17bab : 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/foodplaces196773.txt
          table:
              input format: org.apache.hadoop.mapred.TextInputFormat
              output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
              properties:
                EXTERNAL TRUE
                bucket_count -1
                column.name.delimiter ,
                columns id,place
                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-172-31-10-177.ec2.internal:8020/home/hadoop/foodplace196773.txt
                name mydb.foodplaces
                serialization.ddl struct foodplaces { i32 id, string place}
                serialization.format ,
                serialization.lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                transient_lastDdlTime 1644900337
              serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
              name: mydb.foodplaces

  Stage: Stage-1
    Stats-Aggr Operator


INFO  : Completed compiling command(queryId=hive_20220215055716_7b9dea27-e0c9-47e7-82b5-7b36fad17bab); Time taken: 0.036 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215055716_7b9dea27-e0c9-47e7-82b5-7b36fad17bab): LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces196773.txt' OVERWRITE INTO TABLE MyDb.foodplaces
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table mydb.foodplaces from file:/home/hadoop/foodplaces196773.txt
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20220215055716_7b9dea27-e0c9-47e7-82b5-7b36fad17bab); Time taken: 0.947 seconds
INFO  : OK
No rows affected (1.004 seconds)
```

```
0: jdbc:hive2://localhost:10000/ (MyDb)> SELECT fp.place AS Place,avg(fr.food4) AS FOOD4_AVG from MyDb.foodratings fr JOIN MyDb.foodplaces fp  ON fp.id=fr.id WHERE fp.place='Soup Bowl' GROUP BY fp.place
;
INFO  : Compiling command(queryId=hive_20220215060612_aa04c165-089a-4b26-80f6-eed27e4dbbf8): SELECT fp.place AS Place,avg(fr.food4) AS FOOD4_AVG from MyDb.foodratings fr JOIN MyDb.foodplaces fp  ON fp.i
d=fr.id WHERE fp.place='Soup Bowl' GROUP BY fp.place
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:place, type:string, comment:null), FieldSchema(name:food4_avg, type:double, comment:null)], properties:null)
INFO  : EXPLAIN output for queryid hive_20220215060612_aa04c165-089a-4b26-80f6-eed27e4dbbf8 : 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_20220215060612_aa04c165-089a-4b26-80f6-eed27e4dbbf8:5
      Edges:
        Map 1 <- Map 3 (BROADCAST_EDGE)
        Reducer 2 <- Map 1 (SIMPLE_EDGE)
      DagName:
      Vertices:
        Map 1
            Map Operator Tree:
                TableScan
                  alias: fr
                  Statistics: Num rows: 2183 Data size: 17464 Basic stats: COMPLETE Column stats: NONE
                  GatherStats: false
                  Filter Operator
                    isSamplingPred: false
                    predicate: id is not null (type: boolean)
                    Statistics: Num rows: 2183 Data size: 17464 Basic stats: COMPLETE Column stats: NONE
                    Select Operator
                      expressions: food4 (type: int), id (type: int)
                      outputColumnNames: _col0, _col1
                      Statistics: Num rows: 2183 Data size: 17464 Basic stats: COMPLETE Column stats: NONE
                      Map Join Operator
                        condition map:
                             Inner Join 0 to 1
                        Estimated key counts: Map 3 => 1
                        keys:
                          0 _col1 (type: int)
                          1 _col0 (type: int)
                        outputColumnNames: _col0
                        input vertices:
                          1 Map 3
                        Position of Big Table: 0
                        Statistics: Num rows: 2401 Data size: 19210 Basic stats: COMPLETE Column stats: NONE
                        Select Operator
                          expressions: _col0 (type: int)
                          outputColumnNames: _col1
                          Statistics: Num rows: 2401 Data size: 19210 Basic stats: COMPLETE Column stats: NONE
                          Group By Operator
                            aggregations: avg(_col1)
                            keys: 'Soup Bowl' (type: string)
                            mode: hash
                            outputColumnNames: _col0, _col1
                            Statistics: Num rows: 2401 Data size: 19210 Basic stats: COMPLETE Column stats: NONE
                            Reduce Output Operator
                              key expressions: _col0 (type: string)
                              null sort order: a
                              sort order: +
                              Map-reduce partition columns: _col0 (type: string)
                              Statistics: Num rows: 2401 Data size: 19210 Basic stats: COMPLETE Column stats: NONE
                              tag: -1
                              value expressions: _col1 (type: struct<count:bigint,sum:double,input:int>)
```

```
ive_2022-02-15_06-06-12_831_8738106101688149834-2/-ext-10002
                    NumFilesPerFileSink: 1
                    Statistics: Num rows: 1200 Data size: 9600 Basic stats: COMPLETE Column stats: NONE
                    Stats Publishing Key Prefix: hdfs://ip-172-31-10-177.ec2.internal:8020/tmp/hive/hadoop/a406efdb-244a-45d1-8696-eba1ef351819/hive_2022-02-15_06-06-12_831_8738106101688149834-2/-mr-100
01/.hive-staging_hive_2022-02-15_06-06-12_831_8738106101688149834-2/-ext-10002/
                    table:
                        input format: org.apache.hadoop.mapred.SequenceFileInputFormat
                        output format: org.apache.hadoop.hive.ql.io.HiveSequenceFileOutputFormat
                        properties:
                          columns _col0,_col1
                          columns.types string:double
                          escape.delim \
                          hive.serialization.extend.additional.nesting.levels true
                          serialization.escape.crlf true
                          serialization.format 1
                          serialization.lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                        serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
                    TotalFiles: 1
                    GatherStats: false
                    MultiFileSpray: false

  Stage: Stage-0
    Fetch Operator
      limit: -1
      Processor Tree:
        ListSink


INFO  : Completed compiling command(queryId=hive_20220215060612_aa04c165-089a-4b26-80f6-eed27e4dbbf8); Time taken: 0.508 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20220215060612_aa04c165-089a-4b26-80f6-eed27e4dbbf8): SELECT fp.place AS Place,avg(fr.food4) AS FOOD4_AVG from MyDb.foodratings fr JOIN MyDb.foodplaces fp  ON fp.i
d=fr.id WHERE fp.place='Soup Bowl' GROUP BY fp.place
INFO  : Query ID = hive_20220215060612_aa04c165-089a-4b26-80f6-eed27e4dbbf8
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.place AS Place,avg(fr.f...fp.place(Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1644897119306_0003)

INFO  : Map 1: -/-      Map 3: -/-      Reducer 2: 0/2
INFO  : Map 1: 0/1      Map 3: 0/1      Reducer 2: 0/2
INFO  : Map 1: 0/1      Map 3: 0/1      Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Map 3: 0/1      Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Map 3: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Map 3: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Map 3: 1/1      Reducer 2: 0/2
INFO  : Map 1: 1/1      Map 3: 1/1      Reducer 2: 0(+1)/2
INFO  : Map 1: 1/1      Map 3: 1/1      Reducer 2: 0(+2)/2
INFO  : Map 1: 1/1      Map 3: 1/1      Reducer 2: 1(+1)/2
INFO  : Map 1: 1/1      Map 3: 1/1      Reducer 2: 2/2
INFO  : Completed executing command(queryId=hive_20220215060612_aa04c165-089a-4b26-80f6-eed27e4dbbf8); Time taken: 20.711 seconds
INFO  : OK
+------------+---------------------+
|   place    |      food4_avg      |
+------------+---------------------+
| Soup Bowl  | 25.217142857142857  |
+------------+---------------------+
1 row selected (21.244 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

Commands Executed :

LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces196773.txt' OVERWRITE INTO TABLE MyDb.foodplaces;

SELECT fp.place AS place ,avg(fr.food4) as food4_avg

FROM foodratings as fr JOIN foodplaces as fp ON fr.id=fp.id

WHERE fp.place=’Soup Bowl’

GROUP BY fp.place;

Exercise 8:

a)If your queries require access to all or most of the columns of each row of data, row-based storage will be better suited to your needs. And also if the data has to be in sequentially we can use row-based format. And coming into column-based format is useful when performing analytics queries that require only a subset of columns examined over very large data sets and also column-based format are easy to read and work faster for particular operations such as aggregate.

b)Splitability helps achieve parallelism when conducting queries on huge volumes of data, which reduces computing costs dramatically as compared to processing the complete dataset in one system.MetadataThis allows columnar formats to skip over unnecessary data and process data rapidly. A column-based format will be more amenable to splitting into separate jobs if the query calculation is concerned with a single column at a time.

c) Storing values by column, with the same datatype next to each other, allows you to do more efficient compression on them than if you're storing rows of data. And also the performance of the CPU when visiting column values is better than when visiting row values.

d)Parquet is especially adept at analyzing wide datasets with many columns. Each Parquet file contains binary data organized by “row group.” For each row group, the data values are organized by column. This enables the compression benefits that we described above. Parquet is a good choice for read-heavy workloads and also Parquet can support evolution by storing file schema in file metadata.In certain circumstances, reading simply a single column boosts performance by a factor of ten.