Assignment 4

Atharva Tanaji Kadam (A20467229)

Magic Number - 80416

Question 1:

```
0: jdbc:hive2://localhost:10000/ (default)> DESCRIBE FORMATTED MyDb.foodratings;
Error: Error while compiling statement: FAILED: SemanticException [Error 10072]: Database does not exist: MyDb (state=42000,code=10072)
0: jdbc:hive2://localhost:10000/ (default)> hive
. . . . . . . . . . . . . . . . . . . . . . >
. . . . . . . . . . . . . . . . . . . . . . > create database MyDb
. . . . . . . . . . . . . . . . . . . . . . > ;
Error: Error while compiling statement: FAILED: ParseException line 1:0 cannot recognize input near 'hive' 'create' 'database' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000/ (default)> create database MyDb
. . . . . . . . . . . . . . . . . . . . . . > ;
INFO  : Compiling command(queryId=hive_20221002040857_4e9b1b3e-1aa2-431e-bce0-89a21dfd31ab): create database MyDb
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20221002040857_4e9b1b3e-1aa2-431e-bce0-89a21dfd31ab : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0


INFO  : Completed compiling command(queryId=hive_20221002040857_4e9b1b3e-1aa2-431e-bce0-89a21dfd31ab); Time taken: 0.083 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20221002040857_4e9b1b3e-1aa2-431e-bce0-89a21dfd31ab): create database MyDb
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221002040857_4e9b1b3e-1aa2-431e-bce0-89a21dfd31ab); Time taken: 0.413 seconds
INFO  : OK
No rows affected (0.962 seconds)
0: jdbc:hive2://localhost:10000/ (default)> describe formatted MyDb.foodratings;
Error: Error while compiling statement: FAILED: SemanticException [Error 10001]: Table not found MyDb.foodratings (state=42S02,code=10001)
0: jdbc:hive2://localhost:10000/ (default)> create table if not exists MyDb.foodratings(name STRING COMMENT 'Food Critic Name', food1 INT COMMENT 'Food Type 1', food2 INT COMMENT 'F
ood Type 2', food3 INT COMMENT 'Food Type 3', food4 INT COMMENT 'Food Type 4', ID INT COMMENT 'Restaurant ID') ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE Locat
ion '/home/hadoop';
INFO  : Compiling command(queryId=hive_20221002041422_84611aed-0fd3-4fff-bf37-84f6d861f694): create table if not exists MyDb.foodratings(name STRING COMMENT 'Food Critic Name', food
1 INT COMMENT 'Food Type 1', food2 INT COMMENT 'Food Type 2', food3 INT COMMENT 'Food Type 3', food4 INT COMMENT 'Food Type 4', ID INT COMMENT 'Restaurant ID') ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' STORED AS TEXTFILE Location '/home/hadoop'
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20221002041422_84611aed-0fd3-4fff-bf37-84f6d861f694 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0
      Create Table Operator:
```

```
0: jdbc:hive2://localhost:10000/ (default)> create table if not exists MyDb.foodratings(name STRING COMMENT 'Food Critic Name', food1 INT COMMENT 'Food Type 1', food2 INT COMMENT 'F
ood Type 2', food3 INT COMMENT 'Food Type 3', food4 INT COMMENT 'Food Type 4', ID INT COMMENT 'Restaurant ID') ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE Locat
ion '/home/hadoop';
INFO  : Compiling command(queryId=hive_20221002041422_84611aed-0fd3-4fff-bf37-84f6d861f694): create table if not exists MyDb.foodratings(name STRING COMMENT 'Food Critic Name', food
1 INT COMMENT 'Food Type 1', food2 INT COMMENT 'Food Type 2', food3 INT COMMENT 'Food Type 3', food4 INT COMMENT 'Food Type 4', ID INT COMMENT 'Restaurant ID') ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' STORED AS TEXTFILE Location '/home/hadoop'
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
```

```
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20221002041432_675a6a7d-b99a-4f6c-ab65-03be87f66f3d): describe formatted MyDb.foodratings
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221002041432_675a6a7d-b99a-4f6c-ab65-03be87f66f3d); Time taken: 0.131 seconds
INFO  : OK
```

col_name	data_type	comment
# col_name	data_type	comment
	NULL	NULL
name	string	Food Critic Name
food1	int	Food Type 1
food2	int	Food Type 2
food3	int	Food Type 3
food4	int	Food Type 4
id	int	Restaurant ID
	NULL	NULL
# Detailed Table Information	NULL	NULL
Database:	mydb	NULL
Owner:	hadoop	NULL
CreateTime:	Sun Oct 02 04:14:22 UTC 2022	NULL
LastAccessTime:	UNKNOWN	NULL
Retention:	0	NULL
Location:	hdfs://ip-172-31-87-179.ec2.internal:8020/home/hadoop	NULL
Table Type:	MANAGED_TABLE	NULL
Table Parameters:	NULL	NULL
	transient_lastDdlTime	1664684062
	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	,

```
31 rows selected (0.737 seconds)
0: jdbc:hive2://localhost:10000/ (default)>   sc
```

```
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20221002042145_8b12a805-d216-4481-bf2d-3e1b46dee3f9): CREATE TABLE IF NOT EXISTS MyDb.foodplaces(
id INt,
place STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
Location '/home/hadoop'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221002042145_8b12a805-d216-4481-bf2d-3e1b46dee3f9); Time taken: 0.066 seconds
INFO  : OK
No rows affected (0.103 seconds)
0: jdbc:hive2://localhost:10000/ (default)> DESCRIBE FORMATTED MyDb.foodplaces;
INFO  : Compiling command(queryId=hive_20221002042212_fbd4d131-2423-4790-9ff8-6fda49b71888): 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 deseria
lizer), FieldSchema(name:comment, type:string, comment:from deserializer)], properties:null)
INFO  : EXPLAIN output for queryid hive_20221002042212_fbd4d131-2423-4790-9ff8-6fda49b71888 : 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/d7bebee6-b56d-462b-b57c-74a7790fd88e/hive_2022-10-02_04-22-12_319_8010466869955766363-1/-local-10000
          table: MyDb.foodplaces

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


INFO  : Completed compiling command(queryId=hive_20221002042212_fbd4d131-2423-4790-9ff8-6fda49b71888); Time taken: 0.048 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20221002042212_fbd4d131-2423-4790-9ff8-6fda49b71888): DESCRIBE FORMATTED MyDb.foodplaces
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221002042212_fbd4d131-2423-4790-9ff8-6fda49b71888); Time taken: 0.031 seconds
INFO  : OK
```

col_name	data_type	comment
# col_name	data_type	comment
	NULL	NULL

```
+-------------------------------+----------------------------------------------------+-----------------------+
|           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:                   | Sun Oct 02 04:21:45 UTC 2022                       | NULL                  |
| LastAccessTime:               | UNKNOWN                                            | NULL                  |
| Retention:                    | 0                                                  | NULL                  |
| Location:                     | hdfs://ip-172-31-87-179.ec2.internal:8020/home/hadoop | NULL
| Table Type:                   | MANAGED_TABLE                                      | NULL                  |
| Table Parameters:             | NULL                                               | NULL                  |
|                               | numFiles                                           | 0                     |
|                               | totalSize                                          | 0                     |
|                               | transient_lastDdlTime                              | 1664684505            |
|                               | 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                               | ,                     |
+-------------------------------+----------------------------------------------------+-----------------------+
29 rows selected (0.131 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

Question 2:

```
0: jdbc:hive2://localhost:10000/ (default)> load data local inpath '/home/hadoop/foodratings80416.txt' overwrite into table MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20221002042924_f3b290b4-5f3f-4296-9935-b4f78aaef293): load data local inpath '/home/hadoop/foodratings80416.txt' overwrite into table MyDb.foo
dratings
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : EXPLAIN output for queryid hive_20221002042924_f3b290b4-5f3f-4296-9935-b4f78aaef293 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [MOVE]
  Stage-1 depends on stages: Stage-0 [STATS]
```

```
0: jdbc:hive2://localhost:10000/ (default)> select min(food3) as Minimum, max(food3) as Maximum, avg(food3) as Average from MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20221002043041_2304128c-d7b2-4048-8b1c-8424bbaffe42): select min(food3) as Minimum, max(food3) as Maximu
dratings
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:minimum, type:int, comment:null), FieldSchema(name:maximum, type:int, comm
ype:double, comment:null)], properties:null)
INFO  : EXPLAIN output for queryid hive_20221002043041_2304128c-d7b2-4048-8b1c-8424bbaffe42 : 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_20221002043041_2304128c-d7b2-4048-8b1c-8424bbaffe42:1
      Edges:
        Reducer 2 <- Map 1 (CUSTOM_SIMPLE_EDGE)
      DagName:
      Vertices:
        Map 1
            Map Operator Tree:
                TableScan
                  alias: foodratings
                  Statistics: Num rows: 4372 Data size: 17490 Basic stats: COMPLETE Column stats: NONE
                  GatherStats: false
                  Select Operator
                    expressions: food3 (type: int)
                    outputColumnNames: food3
                    Statistics: Num rows: 4372 Data size: 17490 Basic stats: COMPLETE Column stats: NONE
                    Group By Operator
                      aggregations: min(food3), max(food3), avg(food3)
                      mode: hash
      Processor Tree:
        ListSink


INFO  : Completed compiling command(queryId=hive_20221002043041_2304128c-d7b2-4048-8b1c-8424b
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20221002043041_2304128c-d7b2-4048-8b1c-8424bbaffe42):
dratings
INFO  : Query ID = hive_20221002043041_2304128c-d7b2-4048-8b1c-8424bbaffe42
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 Mini...MyDb.foodratings(Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1664682330707_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
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_20221002043041_2304128c-d7b2-4048-8b1c-8424b
INFO  : OK
+----------+----------+----------+
| minimum  | maximum  | average  |
+----------+----------+----------+
| 1        | 50       | 25.521   |
+----------+----------+----------+
1 row selected (24.243 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

Question 3:

select Name, min(food1) as Minimum, max(food1) as Maximum, avg(food1) as Average from MyDb.foodratings groupby Name;

```
INFO  : Completed compiling command(queryId=hive_20221002043658_7da5545d-9558-495f-b1fe-3c0c330ea454); Time taken: 0.246 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20221002043658_7da5545d-9558-495f-b1fe-3c0c330ea454): select name,  min(food1) as Minimum, max(food1) as Maximum, avg(food1) as Average from M
yDb.foodratings group by name
INFO  : Query ID = hive_20221002043658_7da5545d-9558-495f-b1fe-3c0c330ea454
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 Minimum, ...name(Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1664682330707_0002)

INFO  : Map 1: 0/1      Reducer 2: 0/2
INFO  : Map 1: 0/1      Reducer 2: 0/2
INFO  : Map 1: 0(+1)/1  Reducer 2: 0/2
INFO  : Map 1: 1/1      Reducer 2: 0/2
INFO  : Map 1: 1/1      Reducer 2: 0(+1)/2
INFO  : Map 1: 1/1      Reducer 2: 1(+1)/2
INFO  : Map 1: 1/1      Reducer 2: 2/2
INFO  : Completed executing command(queryId=hive_20221002043658_7da5545d-9558-495f-b1fe-3c0c330ea454); Time taken: 14.655 seconds
INFO  : OK
+-------+----------+----------+---------------------+
| name  | minimum  | maximum  |       average       |
+-------+----------+----------+---------------------+
| Jill  | 1        | 50       | 26.238317757009344  |
| Joe   | 1        | 50       | 25.505102040816325  |
| Joy   | 1        | 50       | 26.685714285714287  |
| Mel   | 1        | 50       | 24.700534759358288  |
| Sam   | 1        | 50       | 24.196891191709845  |
+-------+----------+----------+---------------------+
5 rows selected (14.944 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

Question 4:

```
INFO  : Executing command(queryId=hive_20221002044039_25264526-43cb-42da-b3e1-33b88c0f91c4): 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
Location '/home/hadoop'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221002044039_25264526-43cb-42da-b3e1-33b88c0f91c4); Time taken: 0.049 seconds
INFO  : OK
No rows affected (0.099 seconds)
0: jdbc:hive2://localhost:10000/ (default)> DESCRIBE FORMATTED MyDb.foodratingspart;
INFO  : Compiling command(queryId=hive_20221002044105_e1540498-980b-433e-87a7-356245e0ade8): 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_
lizer), FieldSchema(name:comment, type:string, comment:from deserializer)], properties:null)
INFO  : EXPLAIN output for queryid hive_20221002044105_e1540498-980b-433e-87a7-356245e0ade8 : 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/d7bebee6-b56d-462b-b57c-74a7790fd88e/hive_2022-10-02_04-41-05_457_2973151931750599108-1/-local-10000
          table: MyDb.foodratingspart

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

```
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20221002044105_e1540498-980b-433e-87a7-356245e0ade8): DESCRIBE FORMATTED MyDb.foodratingspart
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221002044105_e1540498-980b-433e-87a7-356245e0ade8); Time taken: 0.09 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:	Sun Oct 02 04:40:39 UTC 2022	NULL
LastAccessTime:	UNKNOWN	NULL
Retention:	0	NULL
Location:	hdfs://ip-172-31-87-179.ec2.internal:8020/home/hadoop	NULL
Table Type:	MANAGED_TABLE	NULL
Table Parameters:	NULL	NULL
	COLUMN_STATS_ACCURATE	{\"BASIC_STATS\":\"true\"}
	numFiles	0
	numPartitions	0
	numRows	0
	rawDataSize	0
	totalSize	0
	transient_lastDdlTime	1664685639
	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 (0.214 seconds)
```

Question 5:

The partition based on the name results in a tiny partition if the number of food critics (name) is comparatively lower than the number of locations (id). When partitions are created based on places(id), a huge number of partitions are created, which makes it challenging to load and retrieve data. Therefore, dividing according to (name of food critic) will be a preferable option.

Question 6:

```
0: jdbc:hive2://localhost:10000/ (default)> set hive.exec.dynamic.partition = true;
No rows affected (0.013 seconds)
0: jdbc:hive2://localhost:10000/ (default)> set hive.exec.dynamic.partition.mode = non-strict;
No rows affected (0.005 seconds)
0: jdbc:hive2://localhost:10000/ (default)> insert overwrite table MyDb.foodratingspart partition (name) select food1, food2, food3, food4, ID, name from MyDb.foodratings;
INFO  : Compiling command(queryId=hive_20221002044559_19fd89e8-6c2f-4375-934d-462b9ead88f1): 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:in
t, comment:null), FieldSchema(name:food4, type:int, comment:null), FieldSchema(name:id, type:int, comment:null), FieldSchema(name:name, type:string, comment:null)], properties:null)
INFO  : EXPLAIN output for queryid hive_20221002044559_19fd89e8-6c2f-4375-934d-462b9ead88f1 : 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_20221002044559_19fd89e8-6c2f-4375-934d-462b9ead88f1:3

INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1664682330707_0003)

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-172-31-87-179.ec2.internal:8020/home/hadoop
9915587050675-1/-ext-10000
INFO  :

INFO  :          Time taken to load dynamic partitions: 1.018 seconds
INFO  :          Time taken for adding to write entity : 0.003 seconds
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20221002044559_19fd89e8-6c2f-4375-934d-462b9ead88f1); Time taken: 17.172 seconds
INFO  : OK
No rows affected (17.428 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

Question 7:

```
INFO  : Executing command(queryId=hive_20221002045112_82006c84-919f-47c8-8f99-42464267d353): select x.place, avg(y.food4) from MyDb.foodratings y join MyDb.foodplaces x on y.ID = x.
ID where x.place = 'Soup Bowl' group by x.place
INFO  : Query ID = hive_20221002045112_82006c84-919f-47c8-8f99-42464267d353
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 x.place, avg(y.food4) from ...x.place(Stage-1)
INFO  : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO  : Status: Running (Executing on YARN cluster with App id application_1664682330707_0003)

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(+2)/2
INFO  : Map 1: 1/1      Map 3: 1/1      Reducer 2: 2/2
INFO  : Completed executing command(queryId=hive_20221002045112_82006c84-919f-47c8-8f99-42464267d353); Time taken: 12.546 seconds
INFO  : OK
+-----------+------+
| x.place   | _c1  |
+-----------+------+
+-----------+------+
No rows selected (13.125 seconds)
0: jdbc:hive2://localhost:10000/ (default)>
```

Question 8:

a. The row format is employed when a query needs access to all or the majority of the columns in a sizably large data source. Column format is advised for enormous data files when conducting analytics queries that only need a small subset of the dataset's columns.

b. In a column-based file format, "splittability" means the capacity to divide a single work into many jobs. It's important because jobs are divided into various jobs (parts) and parallelism is introduced when processing a large amount of data. Batches serve as the split boundaries for row-columnar data when it is gathered and stored in column format. It is essential for processing vast volumes of data because breaking the task up into smaller tasks would increase its efficiency.

c. When storing items of the same kind close to one another, the column format is employed since it offers greater compression than the row format. Values of the same kind should be kept together for more effective compression.

d. The Parquet column file format is used when there are more columns, as well as when analyzing massive datasets with large columns. The binary information in each column is arranged in a row group. The platforms that support the Parquet column file format are Impala, Arrow, Drill, and Spark.