

CSP-554 Big Data Technologies

BAIRI ROHITH REDDY – Assignment #4

Magic Number : 36481

```
[hadoop@ip-172-31-52-112 ~]$ [hadoop@ip-172-31-52-112 ~]$ [hadoop@ip-172-31-52-112 ~]$ [hadoop@ip-172-31-52-112 ~]$ [hadoop@ip-172-31-52-112 ~]$ [hadoop@ip-172-31-52-112 ~]$ java TestDataGen Magic Number = 36481 [hadoop@ip-172-31-52-112 ~]$ [hadoop@ip-172-31-52-112 ~]$ ls foodplaces36481.txt foodratings36481.txt hql.zip TestDataGen.class [hadoop@ip-172-31-52-112 ~]$
```

41°F Rain off and on

Search

File Explorer

Notepad

Google Chrome

OneDrive

PowerShell

FileZilla

Java

Calculator

Task View

12:30 AM 3/1/2023 19

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

CREATE DATABASE MyDb;

```

hadoop@ip-172-31-52-112:~/hql
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 cs595.salpart partition (jobtitle=null) from hdfs://ip-172-31-52-112.ec2.internal:8020/u
ser/hive/warehouse/cs595.db/salpart/.hive-staging_hive_2023-03-01_06-33-10_138_3363723758541480722-1/-ext-10000
INFO :
INFO : Time taken to load dynamic partitions: 82.672 seconds
INFO : Time taken for adding to write entity : 0.242 seconds
INFO : Starting task [Stage-3:STATS] in serial mode
INFO : Completed executing command(queryId=hive_20230301063310_d5d26b1b-237d-4885-8f80-a08725b9791c); Time taken: 168.825 seconds
INFO : OK
No rows affected (172.384 seconds)
0: jdbc:hive2://localhost:10000/ (cs595)>
0: jdbc:hive2://localhost:10000/ (cs595)> CREATE DATABASE MyDb;
INFO : Compiling command(queryId=hive_20230301063618_c56b5f8b-d6ac-4d62-ad4a-c2679e6d231f): CREATE DATABASE MyDb
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301063618_c56b5f8b-d6ac-4d62-ad4a-c2679e6d231f : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0

INFO : Completed compiling command(queryId=hive_20230301063618_c56b5f8b-d6ac-4d62-ad4a-c2679e6d231f); Time taken: 0.084 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301063618_c56b5f8b-d6ac-4d62-ad4a-c2679e6d231f): CREATE DATABASE MyDb
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301063618_c56b5f8b-d6ac-4d62-ad4a-c2679e6d231f); Time taken: 0.051 seconds
INFO : OK
No rows affected (0.175 seconds)
0: jdbc:hive2://localhost:10000/ (cs595)>

```

USE MyDb;

```

hadoop@ip-172-31-52-112:~/hql
INFO : Compiling command(queryId=hive_20230301063618_c56b5f8b-d6ac-4d62-ad4a-c2679e6d231f): CREATE DATABASE MyDb
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301063618_c56b5f8b-d6ac-4d62-ad4a-c2679e6d231f : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0

INFO : Completed compiling command(queryId=hive_20230301063618_c56b5f8b-d6ac-4d62-ad4a-c2679e6d231f); Time taken: 0.084 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301063618_c56b5f8b-d6ac-4d62-ad4a-c2679e6d231f): CREATE DATABASE MyDb
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301063618_c56b5f8b-d6ac-4d62-ad4a-c2679e6d231f); Time taken: 0.051 seconds
INFO : OK
No rows affected (0.175 seconds)
0: jdbc:hive2://localhost:10000/ (cs595)>
0: jdbc:hive2://localhost:10000/ (cs595)> USE MyDb;
INFO : Compiling command(queryId=hive_20230301063821_b0be0d80-798b-402f-8541-22a2c4e56557): USE MyDb
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301063821_b0be0d80-798b-402f-8541-22a2c4e56557 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0

INFO : Completed compiling command(queryId=hive_20230301063821_b0be0d80-798b-402f-8541-22a2c4e56557); Time taken: 0.019 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301063821_b0be0d80-798b-402f-8541-22a2c4e56557): USE MyDb
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301063821_b0be0d80-798b-402f-8541-22a2c4e56557); Time taken: 0.013 seconds
INFO : OK
No rows affected (0.042 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> |

```

CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) row format delimited fields terminated by ',' lines terminated by '\n' stored as TEXTFILE;

```

hadoop@ip-172-31-52-112:~/hql
INFO : Executing command(queryId=hive_20230301063821_b0be0d80-798b-402f-8541-22a2c4e56557): USE MyDb
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301063821_b0be0d80-798b-402f-8541-22a2c4e56557); Time taken: 0.013 seconds
INFO : OK
No rows affected (0.042 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
INFO : Compiling command(queryId=hive_20230301063859_9ef3fcf2-d135-48bc-990d-606399463415): CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE
INFO : Semantic Analysis completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301063859_9ef3fcf2-d135-48bc-990d-606399463415 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

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

INFO : Completed compiling command(queryId=hive_20230301063859_9ef3fcf2-d135-48bc-990d-606399463415); Time taken: 0.025 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301063859_9ef3fcf2-d135-48bc-990d-606399463415): CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301063859_9ef3fcf2-d135-48bc-990d-606399463415); Time taken: 0.08 seconds
INFO : OK
No rows affected (0.117 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

ALTER TABLE foodratings change food1 food1 int comment 'opinion on food1';

```

hadoop@ip-172-31-52-112:~/hql
  serde name: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
  name: MyDb.foodratings

INFO : Completed compiling command(queryId=hive_20230301063859_9ef3fcf2-d135-48bc-990d-606399463415); Time taken: 0.025 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301063859_9ef3fcf2-d135-48bc-990d-606399463415): CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301063859_9ef3fcf2-d135-48bc-990d-606399463415); Time taken: 0.08 seconds
INFO : OK
No rows affected (0.117 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food1 food1 int comment 'opinion on food1';
INFO : Compiling command(queryId=hive_20230301063931_1dc75fc1-3d9f-49f4-92e4-b1b54374e9a2): ALTER TABLE foodratings change food1 food1 int comment 'opinion on food1'
INFO : Semantic Analysis completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301063931_1dc75fc1-3d9f-49f4-92e4-b1b54374e9a2 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

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

INFO : Completed compiling command(queryId=hive_20230301063931_1dc75fc1-3d9f-49f4-92e4-b1b54374e9a2); Time taken: 0.041 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301063931_1dc75fc1-3d9f-49f4-92e4-b1b54374e9a2): ALTER TABLE foodratings change food1 food1 int comment 'opinion on food1'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301063931_1dc75fc1-3d9f-49f4-92e4-b1b54374e9a2); Time taken: 0.045 seconds
INFO : OK
No rows affected (0.105 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

ALTER TABLE foodratings change food2 food2 int comment 'opinion on food2';

```

hadoop@ip-172-31-52-112:~/hql
    type: rename column
    old name: MyDb.foodratings

INFO : Completed compiling command(queryId=hive_20230301063931_1dc75fc1-3d9f-49f4-92e4-b1b54374e9a2); Time taken: 0.041 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301063931_1dc75fc1-3d9f-49f4-92e4-b1b54374e9a2): ALTER TABLE foodratings change food1 food1
int comment 'opinion on food1'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301063931_1dc75fc1-3d9f-49f4-92e4-b1b54374e9a2); Time taken: 0.045 seconds
INFO : OK
No rows affected (0.105 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food2 food2 int comment 'opinion on food2';
INFO : Compiling command(queryId=hive_20230301064005_d792dd09-2b72-4407-a1c4-0f27950fa808): ALTER TABLE foodratings change food2 food2
int comment 'opinion on food2'
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301064005_d792dd09-2b72-4407-a1c4-0f27950fa808 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

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

INFO : Completed compiling command(queryId=hive_20230301064005_d792dd09-2b72-4407-a1c4-0f27950fa808); Time taken: 0.035 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064005_d792dd09-2b72-4407-a1c4-0f27950fa808): ALTER TABLE foodratings change food2 food2
int comment 'opinion on food2'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064005_d792dd09-2b72-4407-a1c4-0f27950fa808); Time taken: 0.038 seconds
INFO : OK
No rows affected (0.089 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

ALTER TABLE foodratings change food3 food3 int comment 'opinion on food3';

```

hadoop@ip-172-31-52-112:~/hql
    type: rename column
    old name: MyDb.foodratings

INFO : Completed compiling command(queryId=hive_20230301064005_d792dd09-2b72-4407-a1c4-0f27950fa808); Time taken: 0.035 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064005_d792dd09-2b72-4407-a1c4-0f27950fa808): ALTER TABLE foodratings change food2 food2
int comment 'opinion on food2'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064005_d792dd09-2b72-4407-a1c4-0f27950fa808); Time taken: 0.038 seconds
INFO : OK
No rows affected (0.089 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food3 food3 int comment 'opinion on food3';
INFO : Compiling command(queryId=hive_20230301064034_5ef9720f-7c09-43d0-9641-158a0eac18a8): ALTER TABLE foodratings change food3 food3
int comment 'opinion on food3'
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301064034_5ef9720f-7c09-43d0-9641-158a0eac18a8 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

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

INFO : Completed compiling command(queryId=hive_20230301064034_5ef9720f-7c09-43d0-9641-158a0eac18a8); Time taken: 0.042 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064034_5ef9720f-7c09-43d0-9641-158a0eac18a8): ALTER TABLE foodratings change food3 food3
int comment 'opinion on food3'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064034_5ef9720f-7c09-43d0-9641-158a0eac18a8); Time taken: 0.039 seconds
INFO : OK
No rows affected (0.093 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

ALTER TABLE foodratings change food4 food4 int comment 'opinion on food4';

```

hadoop@ip-172-31-52-112:~/hql
    type: rename column
    old name: MyDb.foodratings

INFO : Completed compiling command(queryId=hive_20230301064034_5ef9720f-7c09-43d0-9641-158a0eac18a8); Time taken: 0.042 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064034_5ef9720f-7c09-43d0-9641-158a0eac18a8): ALTER TABLE foodratings change food3 food3
int comment 'opinion on food3'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064034_5ef9720f-7c09-43d0-9641-158a0eac18a8); Time taken: 0.039 seconds
INFO : OK
No rows affected (0.093 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food4 food4 int comment 'opinion on food4';
INFO : Compiling command(queryId=hive_20230301064101_000fee0c-426e-4dd8-a816-cffd28b6536d): ALTER TABLE foodratings change food4 food4
int comment 'opinion on food4'
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301064101_000fee0c-426e-4dd8-a816-cffd28b6536d : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

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

INFO : Completed compiling command(queryId=hive_20230301064101_000fee0c-426e-4dd8-a816-cffd28b6536d); Time taken: 0.041 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064101_000fee0c-426e-4dd8-a816-cffd28b6536d): ALTER TABLE foodratings change food4 food4
int comment 'opinion on food4'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064101_000fee0c-426e-4dd8-a816-cffd28b6536d); Time taken: 0.035 seconds
INFO : OK
No rows affected (0.098 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> |

```

ALTER TABLE foodratings change id id int comment 'This is Restaurant ID';

```

hadoop@ip-172-31-52-112:~/hql
    type: rename column
    old name: MyDb.foodratings

INFO : Completed compiling command(queryId=hive_20230301064101_000fee0c-426e-4dd8-a816-cffd28b6536d); Time taken: 0.041 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064101_000fee0c-426e-4dd8-a816-cffd28b6536d): ALTER TABLE foodratings change food4 food4
int comment 'opinion on food4'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064101_000fee0c-426e-4dd8-a816-cffd28b6536d); Time taken: 0.035 seconds
INFO : OK
No rows affected (0.098 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change id id int comment 'This is Restaurant ID';
INFO : Compiling command(queryId=hive_20230301064126_02f5ff3a-07cd-4d78-a9f5-557245496986): ALTER TABLE foodratings change id id int c
omment 'This is Restaurant ID'
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301064126_02f5ff3a-07cd-4d78-a9f5-557245496986 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

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

INFO : Completed compiling command(queryId=hive_20230301064126_02f5ff3a-07cd-4d78-a9f5-557245496986); Time taken: 0.032 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064126_02f5ff3a-07cd-4d78-a9f5-557245496986): ALTER TABLE foodratings change id id int c
omment 'This is Restaurant ID'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064126_02f5ff3a-07cd-4d78-a9f5-557245496986); Time taken: 0.048 seconds
INFO : OK
No rows affected (0.091 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

Execute a Hive command of 'DESCRIBE FORMATTED MyDb.foodratings;'

```

hadoop@ip-172-31-52-112:~/hql
+-----+-----+-----+
| # col_name | data_type | comment |
+-----+-----+-----+
| name | NULL |          |
| food1 | string |          |
| food2 | int   | opinion on food1 |
| food3 | int   | opinion on food2 |
| food4 | int   | opinion on food3 |
| id   | int   | opinion on food4 |
|       | NULL  | This is Restaurant ID |
+-----+-----+-----+
| # Detailed Table Information |
| Database: mydb               |
| Owner: hadoop                |
| CreateTime: wed Mar 01 06:38:59 UTC 2023 |
| LastAccessTime: UNKNOWN      |
| Retention: 0                  |
| Location: hdfs://ip-172-31-52-112.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratings | NULL |
+-----+-----+-----+
| Table Type: MANAGED_TABLE    |
| Table Parameters:           |
| COLUMN_STATS_ACCURATE     |
| last_modified_by: hadoop    |
| last_modified_time: 1677652886 |
| numFiles: 0                 |
| numRows: 0                  |
| rawDataSize: 0              |
| totalSize: 0                |
| transient_lastDdlTime: 1677652886 |
+-----+-----+-----+
| # Storage Information |
| SerDe Library: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe |
| InputFormat: org.apache.hadoop.mapred.TextInputFormat |
| OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat | NULL |
| Compressed: No |
| Num Buckets: -1 |
| Bucket Columns: [] |
| Sort Columns: [] |
| Storage Desc Params: field.delim: , line.delim: \n serialization.format: , |
+-----+-----+-----+
39 rows selected (0.261 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

1 41°F Rain off and on Q Search 12:42 AM 3/1/2023

CREATE TABLE foodplaces(id int,places string) row format delimited fields terminated by ',' lines terminated by '\n' stored as TEXTFILE;

```

hadoop@ip-172-31-52-112:~/hql
+-----+-----+-----+
| InputFormat: org.apache.hadoop.mapred.TextInputFormat | NULL |
| OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat | NULL |
| Compressed: No |
| Num Buckets: -1 |
| Bucket Columns: [] |
| Sort Columns: [] |
| Storage Desc Params: field.delim: , line.delim: \n serialization.format: , |
+-----+-----+-----+
39 rows selected (0.261 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE TABLE foodplaces(id int,places string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
INFO : Compiling command(queryId=hive_20230301064316_76d47384-5cfa-4cb6-b4e9-11b5bf0ff451): CREATE TABLE foodplaces(id int,places string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301064316_76d47384-5cfa-4cb6-b4e9-11b5bf0ff451 : STAGE DEPENDENCIES: Stage-0 is a root stage [DDL]
STAGE PLANS:
Stage: Stage-0
Create Table Operator:
  Create Table
    columns: id int, places string
    field delimiter: ,
    input format: org.apache.hadoop.mapred.TextInputFormat
    line delimiter: ,
    output format: org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat
    serde name: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
    name: MyDb.foodplaces
INFO : Completed compiling command(queryId=hive_20230301064316_76d47384-5cfa-4cb6-b4e9-11b5bf0ff451); Time taken: 0.022 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064316_76d47384-5cfa-4cb6-b4e9-11b5bf0ff451): CREATE TABLE foodplaces(id int,places string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064316_76d47384-5cfa-4cb6-b4e9-11b5bf0ff451); Time taken: 0.08 seconds
INFO : OK
No rows affected (0.112 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

1 41°F Rain off and on Q Search 12:43 AM 3/1/2023

Execute a Hive command of 'DESCRIBE FORMATTED MyDb.foodplaces'

```

hadoop@ip-172-31-52-112:~/hql
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064345_9a01c321-3a83-4205-9edd-6aa3449be681): DESCRIBE FORMATTED MyDb.foodplaces
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064345_9a01c321-3a83-4205-9edd-6aa3449be681); Time taken: 0.018 seconds
INFO : OK

+-----+-----+-----+
| col_name | data_type | comment |
+-----+-----+-----+
| # col_name | data_type | comment |
| id | int | NULL |
| places | string | NULL |
| # Detailed Table Information | | |
| Database: | mydb | NULL |
| Owner: | hadoop | NULL |
| CreateTime: | Wed Mar 01 06:43:16 UTC 2023 | NULL |
| LastAccessTime: | UNKNOWN | NULL |
| Retention: | 0 | NULL |
| Location: | hdfs://ip-172-31-52-112.ec2.internal:8020/user/hive/warehouse/mydb.db/foodplaces | NULL |
| Table Type: | MANAGED_TABLE | NULL |
| Table Parameters: | COLUMN_STATS_ACCURATE | {"BASIC_STATS": "true"} |
| numFiles | 0 | NULL |
| numRows | 0 | NULL |
| rawDataSize | 0 | NULL |
| totalSize | 0 | NULL |
| transient_lastDdlTime | 1677652996 | NULL |
| # Storage Information | | |
| 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 | , | NULL |
| line.delim | \n | NULL |
| serialization.format | , | NULL |
+-----+-----+-----+
33 rows selected (0.086 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> |

```

Exercise 2)

Load the foodratings.txt file created using TestDataGen from your local file system into the foodratings table.

LOAD DATA LOCAL INPATH '/home/hadoop/foodratings36481.txt' OVERWRITE INTO TABLE foodratings;

```

hadoop@ip-172-31-52-112:~/hql
| 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 | , | NULL |
| line.delim | \n | NULL |
| serialization.format | , | NULL |
+-----+-----+-----+
33 rows selected (0.086 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodratings36481.txt' OVERWRITE INTO TABLE foodratings;
INFO : Compiling command(queryId=hive_20230301064459_4bf7992b-1e01-415a-8f18-a20a970d9af9): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings36481.txt' OVERWRITE INTO TABLE foodratings
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301064459_4bf7992b-1e01-415a-8f18-a20a970d9af9 : 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/foodratings36481.txt
    table:
      input format: org.apache.hadoop.mapred.TextInputFormat
      output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
      properties:
        COLUMN_STATS_ACCURATE {"BASIC_STATS":"true"}
        bucket_count -1
        column.name delimiter ,
        columns name,food1,food2,food3,food4,id
        columns.comments ',opinion on food1','opinion on food2','opinion on food3','opinion on food4','This is Restaurant ID'
        columns.types string:int:int:int:int
        field.delim ,
        file.inputformat org.apache.hadoop.mapred.TextInputFormat
        file.outputformat org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
        last_modified_by hadoop
        last_modified_time 1677652886
        line.delim

```

SELECT MIN(food3) AS MINIMUM, MAX(food3) AS MAXIMUM, AVG(food3) AS AVERAGE FROM foodratings;

```

hadoop@ip-172-31-52-112:~/hql
    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_20230301064609_2612c94b-2d43-4802-8880-3dbfe28cbb03); Time taken: 0.713 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064609_2612c94b-2d43-4802-8880-3dbfe28cbb03): SELECT MIN(food3) AS MINIMUM, MAX(food3) AS MAXIMUM, AVG(food3) AS AVERAGE FROM foodratings
INFO : Query ID = hive_20230301064609_2612c94b-2d43-4802-8880-3dbfe28cbb03
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(food3) AS MINIMUM, ...foodratings(Stage-1)
INFO : Tez session was closed. Reopening...
INFO : Session re-established.
INFO : Status: Running (Executing on YARN cluster with App id application_1677651836098_0002)

INFO : Map 1: 0/1 Reducer 2: 0/1
INFO : Map 1: 0/1 Reducer 2: 0/1
INFO : Map 1: 0(+1)/1 Reducer 2: 0/1
INFO : Map 1: 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_20230301064609_2612c94b-2d43-4802-8880-3dbfe28cbb03); Time taken: 14.393 seconds
INFO : OK
+-----+
| minimum | maximum | average |
+-----+
| 1       | 50      | 25.154  |
+-----+
1 row selected (15.215 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```



Exercise 3)

Execute a hive command to output the min, max and average of the values of the food1 column grouped by the first column ‘name’.

SELECT name, MIN(food1), MAX(food1), AVG(food1) FROM foodratings GROUP BY name;

```

hadoop@ip-172-31-52-112:~/hql
GatherStats: false
MultiFilespray: false

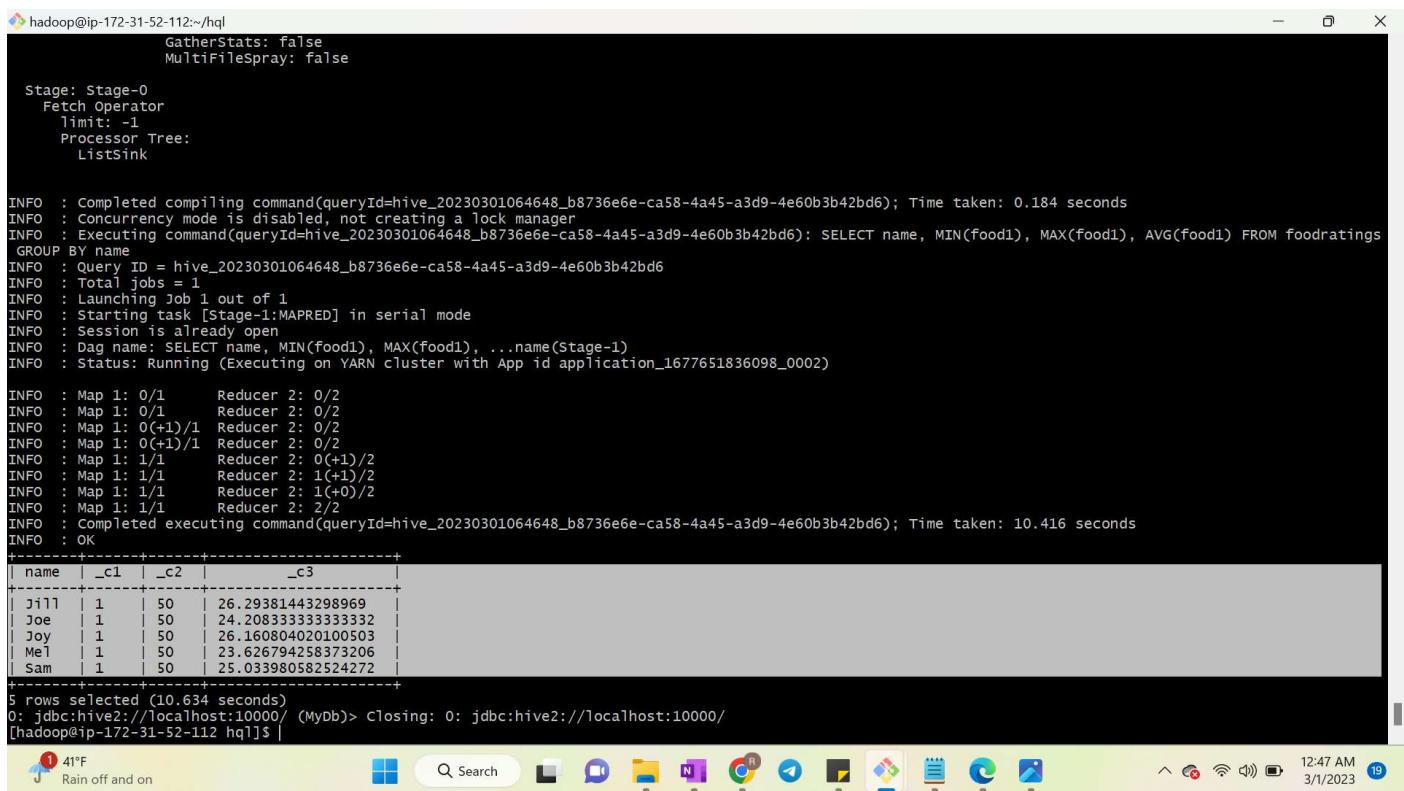
Stage: Stage-0
Fetch Operator
limit: -1
Processor Tree:
ListSink

INFO : Completed compiling command(queryId=hive_20230301064648_b8736e6e-ca58-4a45-a3d9-4e60b3b42bd6); Time taken: 0.184 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064648_b8736e6e-ca58-4a45-a3d9-4e60b3b42bd6): SELECT name, MIN(food1), MAX(food1), AVG(food1) FROM foodratings
GROUP BY name
INFO : Query ID = hive_20230301064648_b8736e6e-ca58-4a45-a3d9-4e60b3b42bd6
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), MAX(food1), ...name(Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1677651836098_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: 0(+1)/1 Reducer 2: 0/2
INFO : Map 1: 1/1      Reducer 2: 0(+1)/2
INFO : Map 1: 1/1      Reducer 2: 1(+1)/2
INFO : Map 1: 1/1      Reducer 2: 1(+0)/2
INFO : Map 1: 1/1      Reducer 2: 2/2
INFO : Completed executing command(queryId=hive_20230301064648_b8736e6e-ca58-4a45-a3d9-4e60b3b42bd6); Time taken: 10.416 seconds
INFO : OK

+-----+
| name | _c1 | _c2 |      _c3      |
+-----+
| Jill | 1  | 50 | 26.29381443298969 |
| Joe  | 1  | 50 | 24.20833333333332 |
| Joy  | 1  | 50 | 26.160804020100503 |
| Mel  | 1  | 50 | 23.626794258373206 |
| Sam  | 1  | 50 | 25.033980582524272 |
+-----+
5 rows selected (10.634 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> Closing: 0: jdbc:hive2://localhost:10000/
[hadoop@ip-172-31-52-112 hql]$ |

```



1 41°F
Rain off and on

Search



12:47 AM
3/1/2023 19

Exercise 4)

In MyDb create a partitioned table called ‘foodratingspart’.

CREATE TABLE foodratingspart (food1 int,food2 int,food3 int, food4 int,id int) partitioned by (name string) ROW FORMAT DELIMITED BY ‘,’ LINES DELIMITED BY ‘\n’ STORED AS TEXTFILE;

```

hadoop@ip-172-31-52-112:~/hql
INFO : Executing command(queryId=hive_20230301064744_d73e18f8-f165-4e6d-964b-c17a6a890e8b): use MyDb
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064744_d73e18f8-f165-4e6d-964b-c17a6a890e8b); Time taken: 0.013 seconds
INFO : OK
No rows affected (0.134 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE TABLE foodratingspart (food1 int,food2 int,food3 int, food4 int,id int) partitioned by (name string) ROW FORMA
T DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
INFO : Compiling command(queryId=hive_20230301064805_f56ea0bb-a7f8-4867-8467-61df690c68df): CREATE TABLE foodratingspart (food1 int,food2 int,food3 int,food
4 int,id int) partitioned by (name string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301064805_f56ea0bb-a7f8-4867-8467-61df690c68df : STAGE DEPENDENCIES:
Stage-0 is a root stage [DDL]

STAGE PLANS:
Stage: Stage-0
Create Table Operator:
  Create Table
    columns: food1 int, food2 int, food3 int, food4 int, id int
    field delimiter: ,
    input format: org.apache.hadoop.mapred.TextInputFormat
    line delimiter: 

    output format: org.apache.hadoop.hive.io.IgnoreKeyTextOutputFormat
    partition columns: name string
    serde name: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerde
    name: MyDb.foodratingspart

INFO : Completed compiling command(queryId=hive_20230301064805_f56ea0bb-a7f8-4867-8467-61df690c68df); Time taken: 0.032 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301064805_f56ea0bb-a7f8-4867-8467-61df690c68df): CREATE TABLE foodratingspart (food1 int,food2 int,food3 int, food
4 int,id int) partitioned by (name string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301064805_f56ea0bb-a7f8-4867-8467-61df690c68df); Time taken: 0.068 seconds
INFO : OK
No rows affected (0.123 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> |

```

Execute a Hive command of ‘DESCRIBE FORMATTED MyDb.foodratingspart;’

```

hadoop@ip-172-31-52-112:~/hql
+-----+-----+-----+
| col_name | data_type | comment |
+-----+-----+-----+
# col_name | data_type | comment |
# col_name | NULL | NULL |
# col_name | int | NULL |
# col_name | string | NULL |
# col_name | NULL | NULL |
# col_name | mydb | NULL |
# col_name | hadoop | NULL |
# col_name | UNKNOWN | NULL |
# col_name | 0 | NULL |
# col_name | hdfs://ip-172-31-52-112.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart | NULL |
# col_name | MANAGED_TABLE | NULL |
# col_name | COLUMN_STATS_ACCURATE | NULL |
# col_name | numFiles | {"BASIC_STATS":"true"} |
# col_name | numPartitions | 0 |
# col_name | numRows | 0 |
# col_name | rawDataSize | 0 |
# col_name | totalSize | 0 |
# col_name | transient_lastDdlTime | 1677653285 |
# col_name | NULL | NULL |
# col_name | NULL | NULL |
# col_name | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL |
# col_name | org.apache.hadoop.mapred.TextInputFormat | NULL |
# col_name | org.apache.hadoop.hive.io.IgnoreKeyTextOutputFormat | NULL |
# col_name | Compressed: No | NULL |
# col_name | Num Buckets: -1 | NULL |
# col_name | Bucket Columns: [] | NULL |
# col_name | Sort Columns: [] | NULL |
# col_name | Storage Desc Params: field.delim | ','
# col_name | line.delim | '\n'
# col_name | serialization.format | '' |
+-----+-----+-----+
42 rows selected (0.318 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> |

```

Exercise 5)

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

Partitioning is done to improve the query performance, so the column we choose to do partition with must be minimal. As stated the food critics are lesser in number and hence the number partitions are limited and are less than 10 (in this case as the number of critics are less than 10). But if we choose to partition on place id, then the number partitions will be huge and this will increase the expense of loading and retrieving data which results in performance degradation due to longer execution times.

Exercise 6)

Execute a hive command to output the min, max and average of the values of the food2 column of MyDB.foodratingspart where the food critic 'name' is either Mel or Jill.

SET hive.exec.dynamic.partition = true;

SET hive.exec.dynamic.partition.mode = non-strict;

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

```

hadoop@ip-172-31-52-112:~/hql
0: jdbc:hive2://localhost:10000/ (MyDb)
0: jdbc:hive2://localhost:10000/ (MyDb) > INSERT OVERWRITE TABLE MyDb.foodratingspart PARTITION (name) SELECT food1, food2, food3, food4, id, name FROM foodratings;
INFO : Compiling command(queryId=hive_20230301064933_c618608f-6b15-475c-a4fe-c05416717640): INSERT OVERWRITE TABLE MyDb.foodratingspart PARTITION (name) SELECT food1, food2, food3, food4, id, name FROM foodratings
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:food1, type:int, comment:null), FieldSchema(name:food2, type:int, comment:null), FieldSchema(name:food3, type:int, comment:null), 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_20230301064933_c618608f-6b15-475c-a4fe-c05416717640 : 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_20230301064933_c618608f-6b15-475c-a4fe-c05416717640:4
    DagName:
    Vertices:
      Map 1
        Map Operator Tree:
          TableScan
            alias: foodratings
            Statistics: Num rows: 145 Data size: 17432 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: 17432 Basic stats: COMPLETE Column stats: NONE
              File Output Operator
                compressed: false
                GlobalTableId: 1
                directory: hdfs://ip-172-31-52-112.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/.hive-staging_hive_2023-03-01_06-49-33_564_1459671489
134339503-5/-ext-10000
              NumFilesPerFileSink: 1
              Statistics: Num rows: 145 Data size: 17432 Basic stats: COMPLETE Column stats: NONE
              Stats Publishing Key Prefix: hdfs://ip-172-31-52-112.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart/.hive-staging_hive_2023-03-01_06-49-33_564_1459671489134339503-5/-ext-10000/
            table:
              input format: org.apache.hadoop.mapred.TextInputFormat
              output format: org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat
              properties:
                bucket_count -1
Rain off and on 41°F Search 12:50 AM 3/1/2023 19
  
```

SELECT MIN(food2) as MIN, MAX(food2) as MAX, AVG(food2) as average FROM foodratingspart WHERE name = 'Mel' OR name = 'Jill';

```

hadoop@ip-172-31-52-112:~/hql
table:
  input format: org.apache.hadoop.mapred.SequenceFileInputFormat
  output format: org.apache.hadoop.hive.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_20230301065046_595f90e0-c1e8-482a-89aa-d182e6203e9e); Time taken: 1.198 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301065046_595f90e0-c1e8-482a-89aa-d182e6203e9e): SELECT MIN(food2) as MIN, MAX(food2) as MAX, AVG(food2) as average FROM foodratingspart WHERE name = 'Mel' OR name = 'Jill'
INFO : Query ID = hive_20230301065046_595f90e0-c1e8-482a-89aa-d182e6203e9e
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Session is already open
INFO : Dag name: SELECT MIN(food2) as MIN, MAX(food2...'Jill'(Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1677651836098_0003)

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_20230301065046_595f90e0-c1e8-482a-89aa-d182e6203e9e); Time taken: 6.07 seconds
INFO : OK
+-----+
| min | max | average |
+-----+
| 1   | 50  | 25.620347394540943 |
+-----+
1 row selected (7.334 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
  
```

Rain off and on 41°F Search 12:51 AM 3/1/2023 19

Exercise 7)

Load the foodplaces.txt file created using TestDataGen from your local file system into the foodplaces table.

LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces76011.txt' OVERWRITE INTO TABLE foodplaces;

```
hadoop@ip-172-31-52-112:~/hql
-----+
| 1 | 50 | 25.620347394540943 |
-----+
1 row selected (7.334 seconds)
0: jdbc:hive2://localhost:10000/ (Mydb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces36481.txt' OVERWRITE INTO TABLE foodplaces;
INFO : Compiling command(queryId=hive_20230301065122_cb216cf6-a30f-4ca0-a077-6b6fdf84265): LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces36481.txt' OVERWRITE INTO TABLE foodplaces
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301065122_cb216cf6-a30f-4ca0-a077-6b6fdf84265 : 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/foodplaces36481.txt
      table:
        input format: org.apache.hadoop.mapred.TextInputFormat
        output format: org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat
        properties:
          COLUMN_STATS_ACCURATE {"BASIC_STATS":"true"}
          bucket_count -1
          column.name.delimiter ,
          columns id,places
          columns.comments
          columns.types int:string
          field_delim ,
          file.inputformat org.apache.hadoop.mapred.TextInputFormat
          file.outputformat org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat
          line.delim

        location hdfs://ip-172-31-52-112.ec2.internal:8020/user/hive/warehouse/mydb.db/foodplaces
        name mydb.foodplaces
        numfiles 0
        numrows 0
        rawDataSize 0
        serialization.ddl struct foodplaces { i32 id, string places}
        serialization.format ,
        serialization.lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
        totalSize 0
        transient_lastDdlTime 1677652996
        serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
        name: mydb.foodplaces

Stage: Stage-1
```



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

SELECT foodplaces.places as Place, avg(foodratings.food4) as Average FROM foodplaces JOIN foodratings ON (foodratings.id=foodplaces.id) WHERE foodplaces.places='Soup Bowl' GROUP BY foodplaces.places;

```

hadoop@ip-172-31-52-112:~/hql
    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_20230301065153_97a7f5b1-0c3d-46e8-8450-c8c437d41b54); Time taken: 0.676 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301065153_97a7f5b1-0c3d-46e8-8450-c8c437d41b54): SELECT foodplaces.places as Place, avg(foodratings.food4) as Average FROM foodplaces JOIN foodratings ON (foodratings.id=foodplaces.id) WHERE foodplaces.places='Soup Bowl' GROUP BY foodplaces.places
INFO : Query ID = hive_20230301065153_97a7f5b1-0c3d-46e8-8450-c8c437d41b54
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 foodplaces.places...foodplaces.places(Stage-1)
INFO : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO : Status: Running (Executing on YARN cluster with App id application_1677651836098_0003)

INFO : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/2
INFO : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/2
INFO : Map 1: 0/1      Map 2: 0(+1)/1  Reducer 3: 0/2
INFO : Map 1: 0(+1)/1 Map 2: 0(+1)/1  Reducer 3: 0/2
INFO : Map 1: 0(+1)/1 Map 2: 0(+1)/1  Reducer 3: 0/2
INFO : Map 1: 0(+1)/1 Map 2: 0(+1)/1  Reducer 3: 0/2
INFO : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/2
INFO : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+2)/2
INFO : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2
INFO : Completed executing command(queryId=hive_20230301065153_97a7f5b1-0c3d-46e8-8450-c8c437d41b54); Time taken: 13.34 seconds
INFO : OK
+-----+
| place | average |
+-----+
| Soup Bowl | 25.588785046728972 |
+-----+
1 row selected (14.068 seconds)
0: jdbc:hive2://localhost:10000/ (<MyDb>)|
```

41°F Rain off and on 12:52 AM 3/1/2023 19

Exercise 8)

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

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

Row format is preferred when the query needs access to all the columns in the table.

When, only the data from a specific columns is needed, Column format is employed. Most analytical queries in big data require only a subset of columns from the relations, so using column format here is appropriate.

- b) **What is “splittability” for a column file format and why is it important when processing large volumes of data?**

Splittability, refers the ability to split larger jobs into smaller jobs for parallel execution. Column file format stores the larger files in smaller chunks of columns for efficient querying and processing. Since, processing of large volumes of data could be very slow or impossible on a single machine. Therefore, splittability is a critical feature that makes column file formats an effective solution to this problem.

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

Column file format can achieve better compression because of the data is stored column by column and these columns are of single data type allowing for better compression whereas the row format has all the columns with different data types making it difficult for compression.

Files stored in column formats achieve better query performance as the query has less data to search from where as the row format has a lot of columns resulting in longer processing time.

- d) **Under what circumstances would it be the best choice to use the “Parquet” column file format**

Columnar file storage format of ‘Parquet’ allows for efficient compression, splittable data, faster and improved query performance.

Parquet is extremely useful for distributed computing frameworks such as Hadoop, which allows for parallel-processing across multiple nodes in a cluster. It is very helpful for processing of big data datasets which have numerous columns, where storage efficiency, query performance and data compression are important.

DOCUMENTATION:-**STEP BY STEP WORKFLOW OF HOMEWORK-4:-**

In bash shell-2/ normal bash shell:-

```
scp -i C:/Users/rohit/OneDrive/Desktop/all_files/BIGDATA/emrkey-pair.cer
C:/Users/rohit/OneDrive/Desktop/all_files/BIGDATA/Homework-Assignments/Homework-4/TestDataGen.class
hadoop@<MASTER-NODE-DNS>/home/hadoop
```

```
scp -i C:/Users/rohit/OneDrive/Desktop/all_files/BIGDATA/emrkey-pair.cer
C:/Users/rohit/OneDrive/Desktop/all_files/BIGDATA/Homework-Assignments/Homework-4/hql.zip
hadoop@<MASTER-NODE-DNS>/home/hadoop
```

In hadoop cmd/ hadoop shell:-

```
ssh -i C:\Users\rohit\OneDrive\Desktop\all_files\BIGDATA\emrkey-pair.cer hadoop@<MASTER-NODE-DNS>
```

```
ls
```

```
java TestDataGen
```

```
ls
```

```
unzip hql.zip
```

```
cd /home/hadoop/hql
```

```
beeline -u jdbc:hive2://localhost:10000/ -n hadoop -d org.apache.hive.jdbc.HiveDriver --showDbInPrompt
```

```
source ./basicsetup.hql;
```

```
source ./partsetup.hql;
```

```
source ./salaries.hql;
```

```
source ./loadsalaries.hql;
```

```
source ./salaries2.hql;
```

```
source ./salaries3.hql;
```

1. Question

```
CREATE DATABASE MyDb;
```

```
USE MyDb;
```

```
CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
```

```
ALTER TABLE foodratings change food1 food1 int comment 'opinion on food1';
```

```
ALTER TABLE foodratings change food2 food2 int comment 'opinion on food2';
```

```
ALTER TABLE foodratings change food3 food3 int comment 'opinion on food3';
```

```
ALTER TABLE foodratings change food4 food4 int comment 'opinion on food4';
```

```
ALTER TABLE foodratings change id id int comment 'This is Restaurant ID';
```

```
DESCRIBE FORMATTED MyDb.foodratings;
```

```
CREATE TABLE foodplaces(id int,places string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES
TERMINATED BY '\n' STORED AS TEXTFILE;
```

```
DESCRIBE FORMATTED MySql.foodplaces;
```

2. Question

```
LOAD DATA LOCAL INPATH '/home/hadoop/foodratings36481.txt' OVERWRITE INTO TABLE foodratings;
```

```
SELECT MIN(food3) AS MINIMUM, MAX(food3) AS MAXIMUM, AVG(food3) AS AVERAGE FROM foodratings;
```

3. Question

```
SELECT name, MIN(food1), MAX(food1), AVG(food1) FROM foodratings GROUP BY name;
```

4. question

```
CREATE TABLE foodratingspart (food1 int, food2 int, food3 int, food4 int, id int) partitioned by (name string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
```

```
DESCRIBE FORMATTED MySql.foodratingspart;
```

5. Question answer the theory question.

6. question

```
SET hive.exec.dynamic.partition = true;
```

```
SET hive.exec.dynamic.partition.mode = non-strict;
```

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

```
SELECT MIN(food2) as MIN, MAX(food2) as MAX, AVG(food2) as average FROM foodratingspart WHERE name = 'Mel'  
OR name = 'Jill';
```

7. Question

```
LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces36481.txt' OVERWRITE INTO TABLE foodplaces;
```

```
SELECT foodplaces.places as Place, avg(foodratings.food4) as Average FROM foodplaces JOIN foodratings ON  
(foodratings.id=foodplaces.id) WHERE foodplaces.places='Soup Bowl' GROUP BY foodplaces.places;
```

8. Question Answer the following questions about the data formats used in Big data