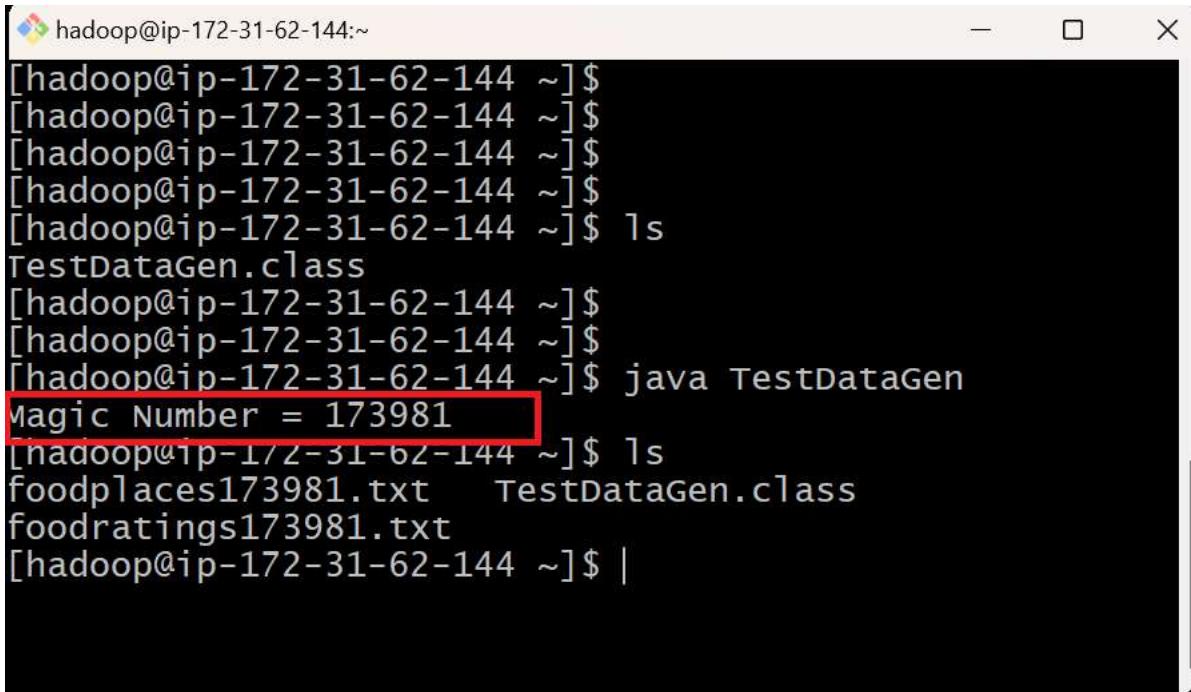


CSP-554 Big Data Technologies

BAIRI ROHITH REDDY – Assignment #4

Magic Number :



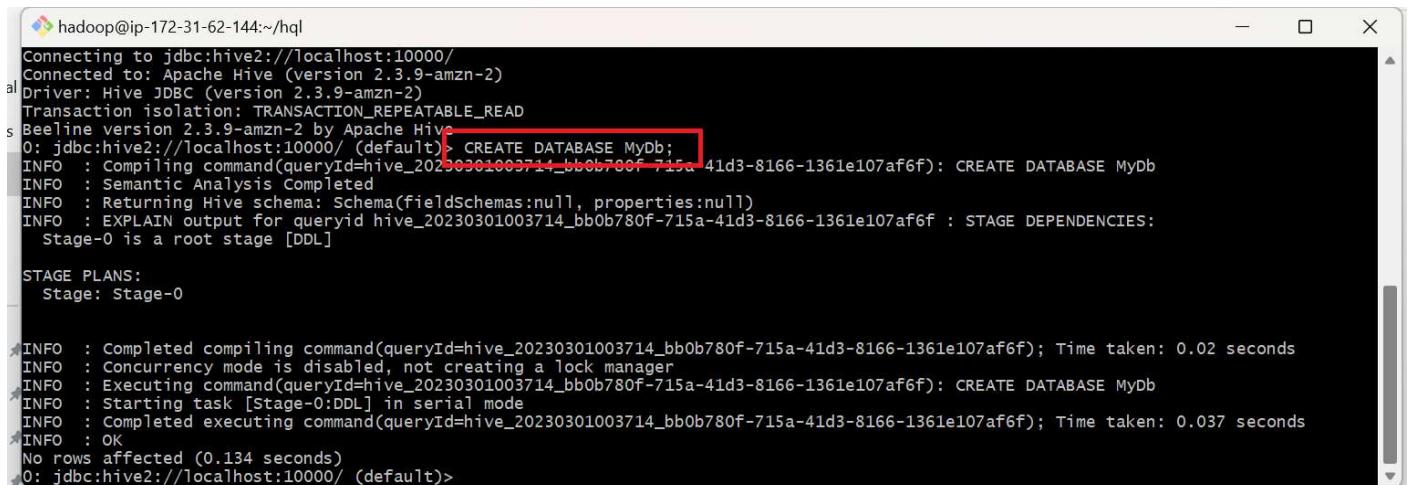
```

hadoop@ip-172-31-62-144:~]
[hadoop@ip-172-31-62-144 ~]$
[hadoop@ip-172-31-62-144 ~]$
[hadoop@ip-172-31-62-144 ~]$
[hadoop@ip-172-31-62-144 ~]$
[hadoop@ip-172-31-62-144 ~] $ ls
TestDataGen.class
[hadoop@ip-172-31-62-144 ~]$
[hadoop@ip-172-31-62-144 ~]$
[hadoop@ip-172-31-62-144 ~] $ java TestDataGen
Magic Number = 173981
[hadoop@ip-172-31-62-144 ~] $ ls
foodplaces173981.txt TestDataGen.class
foodratings173981.txt
[hadoop@ip-172-31-62-144 ~] $ |

```

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

CREATE DATABASE MyDb;



```

hadoop@ip-172-31-62-144:~/hql
Connecting to jdbc:hive2://localhost:10000/
Connected to: Apache Hive (version 2.3.9-amzn-2)
Driver: Hive JDBC (version 2.3.9-amzn-2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
s Beeline version 2.3.9-amzn-2 by Apache Hive
0: jdbc:hive2://localhost:10000/ (default) > CREATE DATABASE MyDb;
INFO : Compiling command(queryId=hive_20230301003714_bb0b780f-715a-41d3-8166-1361e107af6f): CREATE DATABASE MyDb
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301003714_bb0b780f-715a-41d3-8166-1361e107af6f : STAGE_DEPENDENCIES:
Stage-0 is a root stage [DDL]
STAGE PLANS:
 Stage: Stage-0

INFO : Completed compiling command(queryId=hive_20230301003714_bb0b780f-715a-41d3-8166-1361e107af6f); Time taken: 0.02 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301003714_bb0b780f-715a-41d3-8166-1361e107af6f): CREATE DATABASE MyDb
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301003714_bb0b780f-715a-41d3-8166-1361e107af6f); Time taken: 0.037 seconds
INFO : OK
No rows affected (0.134 seconds)
0: jdbc:hive2://localhost:10000/ (default)>

```

USE MyDB;

```

hadoop@ip-172-31-62-144:~/hql
INFO : Executing command(queryId=hive_20230301003714_bb0b780f-715a-41d3-8166-1361e107af6f): CREATE DATABASE MyDb
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301003714_bb0b780f-715a-41d3-8166-1361e107af6f); Time taken: 0.037 seconds
INFO : OK
No rows affected (0.134 seconds)
0: jdbc:hive2://localhost:10000/ (default)> USE MyDb;
INFO : Compiling command(queryId=hive_20230301003819_ac9a5f75-40ba-4670-b2e5-6b3e5cc96d59): USE MyDb
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301003819_ac9a5f75-40ba-4670-b2e5-6b3e5cc96d59 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0

INFO : Completed compiling command(queryId=hive_20230301003819_ac9a5f75-40ba-4670-b2e5-6b3e5cc96d59); Time taken: 0.024 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301003819_ac9a5f75-40ba-4670-b2e5-6b3e5cc96d59): USE MyDb
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301003819_ac9a5f75-40ba-4670-b2e5-6b3e5cc96d59); Time taken: 0.007 seconds
INFO : OK
No rows affected (0.056 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 ',';

```

hadoop@ip-172-31-62-144:~/hql
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301003819_ac9a5f75-40ba-4670-b2e5-6b3e5cc96d59 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [DDL]

STAGE PLANS:
  Stage: Stage-0

INFO : Completed compiling command(queryId=hive_20230301003819_ac9a5f75-40ba-4670-b2e5-6b3e5cc96d59); Time taken: 0.024 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301003819_ac9a5f75-40ba-4670-b2e5-6b3e5cc96d59): USE MyDb
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301003819_ac9a5f75-40ba-4670-b2e5-6b3e5cc96d59); Time taken: 0.007 seconds
INFO : OK
No rows affected (0.056 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 ',';
INFO : Compiling command(queryId=hive_20230301004054_d04670d9-9a2b-4fcf-8f60-cd997fa683b9): CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) row format delimited fields terminated by ',';
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301004054_d04670d9-9a2b-4fcf-8f60-cd997fa683b9 : 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
        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_20230301004054_d04670d9-9a2b-4fcf-8f60-cd997fa683b9); Time taken: 0.064 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301004054_d04670d9-9a2b-4fcf-8f60-cd997fa683b9): CREATE TABLE foodratings(name string,food1 int,food2 int,food3 int,food4 int,id int) row format delimited fields terminated by ',';
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301004054_d04670d9-9a2b-4fcf-8f60-cd997fa683b9); Time taken: 0.13 seconds
INFO : OK
No rows affected (0.216 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> |

```

ALTER TABLE foodratings change food1 food1 int comment 'commenting food1';

```

hadoop@ip-172-31-62-144:~/hql
+-----+
| SerDe Library:          | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL |
| InputFormat:             | org.apache.hadoop.mapred.TextInputFormat        | NULL |
| OutputFormat:            | org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat | NULL |
| Compressed:              | No                                              | NULL |
| Num Buckets:             | -1                                             | NULL |
| Bucket Columns:          | []                                             | NULL |
| Sort Columns:             | []                                             | NULL |
| Storage Desc Params:    | NULL                                           | NULL |
|                         | field.delim                                     | ,      |
|                         | serialization.format                          | ;      |
+-----+
36 rows selected (0.499 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food1 food1 int comment 'commenting food1'
;
INFO : Compiling command(queryId=hive_20230301004835_56675b41-a469-4c7c-bbf8-d6f90d2b113b): ALTER TABLE foodratings change food1 food1 int comment 'commenting food1'
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301004835_56675b41-a469-4c7c-bbf8-d6f90d2b113b : 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_20230301004835_56675b41-a469-4c7c-bbf8-d6f90d2b113b); Time taken: 0.05 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301004835_56675b41-a469-4c7c-bbf8-d6f90d2b113b): ALTER TABLE foodratings change food1 food1 int comment 'commenting food1'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301004835_56675b41-a469-4c7c-bbf8-d6f90d2b113b); Time taken: 0.078 seconds
INFO : OK
No rows affected (0.15 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

40°F Cloudy 6:48 PM 2/28/2023

ALTER TABLE foodratings change food2 food2 int comment 'Now commenting food2';

```

hadoop@ip-172-31-62-144:~/hql
+-----+
| Alter Table
|   type: rename column
|   old name: MyDb.foodratings

INFO : Completed compiling command(queryId=hive_20230301004835_56675b41-a469-4c7c-bbf8-d6f90d2b113b); Time taken: 0.05 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301004835_56675b41-a469-4c7c-bbf8-d6f90d2b113b): ALTER TABLE foodratings change food1 food1 int comment 'commenting food1'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301004835_56675b41-a469-4c7c-bbf8-d6f90d2b113b); Time taken: 0.078 seconds
INFO : OK
No rows affected (0.15 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food2 food2 int comment 'Now commenting food2';
INFO : Compiling command(queryId=hive_20230301004935_7815ca9c-4544-48a1-b5c7-5c1c5f529744): ALTER TABLE foodratings change food2 food2 int comment 'Now commenting food2'
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301004935_7815ca9c-4544-48a1-b5c7-5c1c5f529744 : 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_20230301004935_7815ca9c-4544-48a1-b5c7-5c1c5f529744); Time taken: 0.051 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301004935_7815ca9c-4544-48a1-b5c7-5c1c5f529744): ALTER TABLE foodratings change food2 food2 int comment 'Now commenting food2'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301004935_7815ca9c-4544-48a1-b5c7-5c1c5f529744); Time taken: 0.075 seconds
INFO : OK
No rows affected (0.147 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> |

```

40°F Cloudy 6:49 PM 2/28/2023

ALTER TABLE foodratings change food3 food3 int comment 'Now commenting food3';

```

hadoop@ip-172-31-62-144:~/hql
Stage: Stage-0
  Alter Table Operator:
    Alter Table
      type: rename column
      old name: MyDb.foodratings

INFO : Completed compiling command(queryId=hive_20230301004935_7815ca9c-4544-48a1-b5c7-5c1c5f529744); Time taken: 0.051 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301004935_7815ca9c-4544-48a1-b5c7-5c1c5f529744): ALTER TABLE foodratings change food2 food2 int comment 'Now commenting food2'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301004935_7815ca9c-4544-48a1-b5c7-5c1c5f529744); Time taken: 0.075 seconds
INFO : OK
No rows affected (0.147 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food3 food3 int comment 'Now commenting food3';
INFO : compiling command(queryId=hive_20230301005032_a6103385-d0dd-423b-a809-28c27c9d906d): ALTER TABLE foodratings change food3 food3 int comment 'Now commenting food3'
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301005032_a6103385-d0dd-423b-a809-28c27c9d906d : 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_20230301005032_a6103385-d0dd-423b-a809-28c27c9d906d); Time taken: 0.031 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301005032_a6103385-d0dd-423b-a809-28c27c9d906d): ALTER TABLE foodratings change food3 food3 int comment 'Now commenting food3'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301005032_a6103385-d0dd-423b-a809-28c27c9d906d); Time taken: 0.046 seconds
INFO : OK
No rows affected (0.086 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

ALTER TABLE foodratings change food4 food4 int comment 'Now onto food4';

```

hadoop@ip-172-31-62-144:~/hql
Stage: Stage-0
  Alter Table Operator:
    Alter Table
      type: rename column
      old name: MyDb.foodratings

INFO : Completed compiling command(queryId=hive_20230301005032_a6103385-d0dd-423b-a809-28c27c9d906d); Time taken: 0.031 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301005032_a6103385-d0dd-423b-a809-28c27c9d906d): ALTER TABLE foodratings change food3 food3 int comment 'Now commenting food3'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301005032_a6103385-d0dd-423b-a809-28c27c9d906d); Time taken: 0.046 seconds
INFO : OK
No rows affected (0.086 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change food4 food4 int comment 'Now onto food4';
INFO : compiling command(queryId=hive_20230301005109_28f36989-035c-4489-83fb-a4399cd4fcf6): ALTER TABLE foodratings change food4 food4 int comment 'Now onto food4'
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301005109_28f36989-035c-4489-83fb-a4399cd4fcf6 : 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_20230301005109_28f36989-035c-4489-83fb-a4399cd4fcf6); Time taken: 0.051 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301005109_28f36989-035c-4489-83fb-a4399cd4fcf6): ALTER TABLE foodratings change food4 food4 int comment 'Now onto food4'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301005109_28f36989-035c-4489-83fb-a4399cd4fcf6); Time taken: 0.051 seconds
INFO : OK
No rows affected (0.13 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> |

```

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

```

hadoop@ip-172-31-62-144:~/hql
Stage: Stage-0
  Alter Table Operator:
    Alter Table
      type: rename column
      old name: MyDb.foodratings

INFO : Completed compiling command(queryId=hive_20230301005109_28f36989-035c-4489-83fb-a4399cd4fcf6); Time taken: 0.051 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301005109_28f36989-035c-4489-83fb-a4399cd4fcf6): ALTER TABLE foodratings change food4 food4 int comment 'Now onto food4'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301005109_28f36989-035c-4489-83fb-a4399cd4fcf6); Time taken: 0.051 seconds
INFO : OK
No rows affected (0.13 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> ALTER TABLE foodratings change id id int comment 'This is ID';
INFO : compiling command(queryId=hive_20230301005146_bc3e9a0e-d7ff-4e5f-86a9-99f3753106ee): ALTER TABLE foodratings change id id int comment 'This is ID'
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301005146_bc3e9a0e-d7ff-4e5f-86a9-99f3753106ee : 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_20230301005146_bc3e9a0e-d7ff-4e5f-86a9-99f3753106ee); Time taken: 0.045 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301005146_bc3e9a0e-d7ff-4e5f-86a9-99f3753106ee): ALTER TABLE foodratings change id id int comment 'This is ID'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301005146_bc3e9a0e-d7ff-4e5f-86a9-99f3753106ee); Time taken: 0.053 seconds
INFO : OK
No rows affected (0.121 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

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

```

hadoop@ip-172-31-62-144:~/hql
+-----+-----+-----+
| col_name          | data_type        | comment          |
+-----+-----+-----+
| # col_name        | data_type        | comment          |
| name              | string           | NULL             |
| food1             | int              | NULL             |
| food2             | int              | commenting food1 |
| food3             | int              | Now commenting food2 |
| food4             | int              | Now commenting food3 |
| id                | int              | Now onto food4 |
|                   | NULL             | This is ID      |
| # Detailed Table Information |
| Database:         | NULL             | NULL             |
| Owner:            | mydb             | NULL             |
| CreateTime:       | hadoop           | NULL             |
| LastAccessTime:  | Wed Mar 01 00:40:54 UTC 2023 | NULL             |
| Retention:        | UNKNOWN          | NULL             |
| Location:         | 0                | NULL             |
|                   | hdfs://ip-172-31-62-144.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratings | NULL
| Table Type:       | MANAGED_TABLE    | NULL             |
| Table Parameters:| NULL             | NULL             |
|                   | last_modified_by | NULL             |
|                   | last_modified_time| hadoop           |
|                   | numFiles          | 1677631907      |
|                   | numRows           | 1                |
|                   | rawDataSize       | 0                |
|                   | totalSize          | 17478            |
|                   | transient_lastDdlTime| 1677632225      |
|                   | NULL              | NULL             |
|                   | NULL              | 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       | ,                |
|                   | serialization.format | ,                |
+-----+-----+-----+
37 rows selected (0.075 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> -

```

CREATE TABLE foodplaces(id int,places string) row format delimited fields terminated by ',';

```

hadoop@ip-172-31-62-144:~/hql
+-----+-----+-----+
| InputFormat: | org.apache.hadoop.mapred.TextInputFormat | NULL |
| OutputFormat: | org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat | NULL |
| Compressed: | No | NULL |
| Num Buckets: | -1 | NULL |
| Bucket Columns: | [] | NULL |
| Sort Columns: | [] | NULL |
| Storage Desc Params: | NULL | NULL |
| | field.delim | , |
| | serialization.format | , |
+-----+-----+-----+
38 rows selected (0.077 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE TABLE foodplaces(id int,places string) row format delimited fields terminated by ',';
INFO : Compiling command(queryId=hive_20230301005336_1be69d14-ba98-4d06-96b2-0e7fef49b585): CREATE TABLE foodplaces(id int,places string) row fo
rmat delimited fields terminated by ','
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301005336_1be69d14-ba98-4d06-96b2-0e7fef49b585 : STAGE DEPENDENCIES:
Stage-0 is a root stage [DDL]

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

INFO : Completed compiling command(queryId=hive_20230301005336_1be69d14-ba98-4d06-96b2-0e7fef49b585); Time taken: 0.02 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301005336_1be69d14-ba98-4d06-96b2-0e7fef49b585): CREATE TABLE foodplaces(id int,places string) row fo
rmat delimited fields terminated by ','
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301005336_1be69d14-ba98-4d06-96b2-0e7fef49b585); Time taken: 0.065 seconds
INFO : OK
No rows affected (0.102 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

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

```

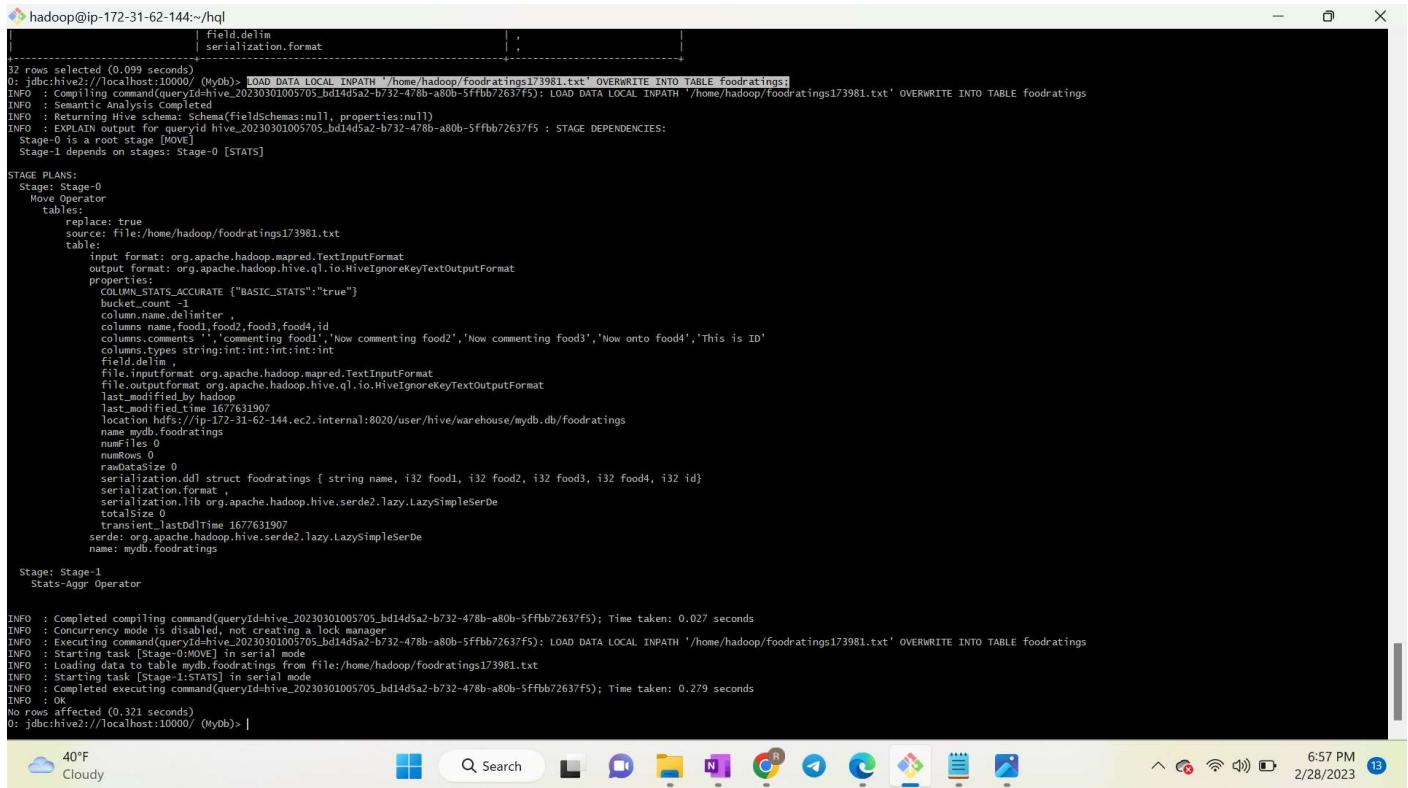
hadoop@ip-172-31-62-144:~/hql
+-----+-----+-----+
| col_name | data_type | comment |
+-----+-----+-----+
| # col_name | data_type | comment |
| id | int | NULL |
| places | string | NULL |
| # Detailed Table Information | | |
| Database: | mydb | NULL |
| Owner: | hadoop | NULL |
| CreateTime: | Wed Mar 01 00:53:36 UTC 2023 | NULL |
| LastAccessTime: | UNKNOWN | NULL |
| Retention: | 0 | NULL |
| Location: | hdfs://ip-172-31-62-144.ec2.internal:8020/user/hive/warehouse/mydb.db/foodplaces | NULL |
| Table Type: | MANAGED_TABLE | NULL |
| Table Parameters: | numFiles | 1 |
| | numRows | 0 |
| | rawDataSize | 0 |
| | totalSize | 59 |
| | transient_lastDdlTime | 1677633826 |
| # 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 | , |
| | serialization.format | , |
+-----+-----+-----+
31 rows selected (0.067 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/foodratings76011.txt' OVERWRITE INTO TABLE foodratings;



```

hadoop@ip-172-31-62-144:~/hql
[...]
| field.delim      | : |
| serialization.format | : |
:
0 rows selected (0.099 seconds)
0: jdbc:hive2://localhost:10000 (MyDB) > LOAD DATA LOCAL INPATH '/home/hadoop/foodratings173981.txt' OVERWRITE INTO TABLE foodratings;
INFO : Compiling command(queryId:hive_20230301005705_bd14d5a2-b732-478b-a80b-5ffbb72637f5): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings173981.txt' OVERWRITE INTO TABLE foodratings
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301005705_bd14d5a2-b732-478b-a80b-5ffbb72637f5 : STAGE DEPENDENCIES:
Stage-0 is a root stage (MOVE)
Stage-1 depends on stages: Stage-0 [STATS]
STAGE PLANS:
Stage: Stage-0
  Move Operator
    table:
      replace: true
      source: file:/home/hadoop/foodratings173981.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.name Food1,Food2,Food3,Food4,
          columns.comments ', commenting Food1','Now commenting Food2','Now commenting Food3','Now onto Food4','This is ID'
          columns.types string:int:int:int:int
          field.delim file.inputformat.org.apache.hadoop.mapred.TextInputFormat
          file.outputformat.org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat
          last_modified_by hadoop
          last_modified_time 167631907
          location hdfs://ip-172-31-62-144.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratings
          name mydb.foodratings
          numFiles 0
          numRows 0
          rawDataSize 0
          serde org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
          totalSize 0
          transient_lastDFOTime 167631907
          serde: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
          name: mydb.foodratings
Stage: Stage-1
  Stats-Agg Operator
:
INFO : Completed compiling command(queryId:hive_20230301005705_bd14d5a2-b732-478b-a80b-5ffbb72637f5); Time taken: 0.027 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId:hive_20230301005705_bd14d5a2-b732-478b-a80b-5ffbb72637f5): LOAD DATA LOCAL INPATH '/home/hadoop/foodratings173981.txt' OVERWRITE INTO TABLE foodratings
INFO : Starting task [Stage-0:MOVE] in serial mode
INFO : Loading data to table mydb.foodratings from file:/home/hadoop/foodratings173981.txt
INFO : Starting task [Stage-1:STATS] in serial mode
INFO : Completed executing command(queryId:hive_20230301005705_bd14d5a2-b732-478b-a80b-5ffbb72637f5); Time taken: 0.279 seconds
INFO : No rows affected (0.321 seconds)
0: jdbc:hive2://localhost:10000 (MyDB) >

```

6:57 PM 2/28/2023 13

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

```

hadoop@ip-172-31-62-144:~/hql
    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_20230301005829_3b68e9cf-d2d8-4219-8067-3c9a7b1cc3bf); Time taken: 0.64 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301005829_3b68e9cf-d2d8-4219-8067-3c9a7b1cc3bf): SELECT MIN(food3) AS MINIMUM, MAX(food3) AS MAXIMUM, AVG(food3) AS AVERAGE FROM foodratings
INFO : Query ID = hive_20230301005829_3b68e9cf-d2d8-4219-8067-3c9a7b1cc3bf
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 MINIMUM, ...foodratings(Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1677629474816_0002)

INFO : Map 1: 0/1      Reducer 2: 0/1
INFO : Map 1: 0/1      Reducer 2: 0/1
INFO : Map 1: 0(+1)/1  Reducer 2: 0/1
INFO : Map 1: 0/1      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_20230301005829_3b68e9cf-d2d8-4219-8067-3c9a7b1cc3bf); Time taken: 17.951 seconds
INFO : OK

+-----+-----+-----+
| minimum | maximum | average |
+-----+-----+-----+
| 1       | 50      | 25.549 |
+-----+-----+-----+
1 row selected (18.668 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> |

```

Cloudy



6:59 PM 2/28/2023 13

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-172-31-62-144:~/hql
    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_20230301010656_89a20dfe-9f9d-4a22-bb42-53d2a617428c); Time taken: 0.183 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301010656_89a20dfe-9f9d-4a22-bb42-53d2a617428c): SELECT name, MIN(food1), MAX(food1), AVG(food1) FROM foodratings GROUP BY name
INFO : Query ID = hive_20230301010656_89a20dfe-9f9d-4a22-bb42-53d2a617428c
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 : Tez session was closed. Reopening...
INFO : Session re-established.
INFO : Status: Running (Executing on YARN cluster with App id application_1677629474816_0003)

INFO : Map 1: -/- Reducer 2: 0/2
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(+0)/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_20230301010656_89a20dfe-9f9d-4a22-bb42-53d2a617428c); Time taken: 15.529 seconds
INFO : OK
+-----+-----+-----+-----+
| name | _c1 | _c2 | _c3 |
+-----+-----+-----+-----+
| jill | 1   | 50  | 26.42439024390244 |
| Joe  | 1   | 50  | 23.802816901408452 |
| Joy  | 1   | 50  | 26.04102564102564 |
| Mel  | 1   | 50  | 25.918478260869566 |
| Sam  | 1   | 50  | 25.423645320197043 |
+-----+-----+-----+-----+
5 rows selected (15.754 seconds)
0: jdbc:hive2://localhost:10000/ (Mydb)>
```

Exercise 4)

In MyDb create a partitioned table called ‘foodratingspart’.

```

hadoop@ip-172-31-62-144:~/hql
INFO : Status: Running (Executing on YARN cluster with App id application_1677629474816_0003)
INFO : Map 1: /-- Reducer 2: 0/2
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(+0)/2
INFO : Map 1: 1/1 Reducer 2: 2/2
INFO : Completed executing command(queryId=hive_20230301010656_89a20dfe-9f9d-4a22-bb42-53d2a617428c); Time taken: 15.529 seconds
INFO : OK
+---+---+---+---+
| name | _c1 | c2 | _c3 |
+---+---+---+---+
| Jill | 1 | 50 | 26.42439024390244 |
| Joe | 1 | 50 | 23.80281690108452 |
| Joy | 1 | 50 | 26.04102564102564 |
| Mel | 1 | 50 | 25.918478260869566 |
| Sam | 1 | 50 | 25.423645320197043 |
+---+---+---+---+
5 rows selected (15.754 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> CREATE TABLE foodratingspart (food1 int, food2 int, food3 int, food4 int, id int) partitioned by (name string) row format delimited file
IDs terminated by ',';
INFO : Compiling command(queryId=hive_20230301011159_167b4c6f-fb63-4196-825d-d2f45da805c6): CREATE TABLE foodratingspart (food1 int, food2 int, food3 int, food4 int, id int) p
artitioned by (name string) row format delimited fields terminated by ','
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldsSchemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301011159_167b4c6f-fb63-4196-825d-d2f45da805c6 : STAGE_DEPENDENCIES:
Stage-0 is a root stage [DDL]
STAGE PLANS:
Stage: Stage-0
  Create Table Operator:
    Create Table
      columns: food1 int, food2 int, food3 int, food4 int, id int
      field delimiter: ,
      input format: org.apache.hadoop.mapred.TextInputFormat
      output format: org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat
      partition columns: name string
      serde name: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
      name: MyDb.foodratingspart

INFO : Completed compiling command(queryId=hive_20230301011159_167b4c6f-fb63-4196-825d-d2f45da805c6); Time taken: 0.02 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301011159_167b4c6f-fb63-4196-825d-d2f45da805c6): CREATE TABLE foodratingspart (food1 int, food2 int, food3 int, food4 int, id int) p
artitioned by (name string) row format delimited fields terminated by ','
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301011159_167b4c6f-fb63-4196-825d-d2f45da805c6); Time taken: 0.044 seconds
INFO : OK
No rows affected (0.079 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

40°F Cloudy 7:12 PM 2/28/2023

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

```

hadoop@ip-172-31-62-144:~/hql
INFO : Completed compiling command(queryId=hive_20230301011256_e51dd630-aa53-46c7-97fc-c90ce0c88743); Time taken: 0.038 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301011256_e51dd630-aa53-46c7-97fc-c90ce0c88743): DESCRIBE FORMATTED MyDb.foodratingspart
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230301011256_e51dd630-aa53-46c7-97fc-c90ce0c88743); Time taken: 0.041 seconds
INFO : OK
+---+---+---+
| col_name | data_type | comment |
+---+---+---+
| # col_name | data_type | comment |
| Food1 | int | NULL |
| Food2 | int | NULL |
| Food3 | int | NULL |
| Food4 | int | NULL |
| id | int | NULL |
| # Partition Information | data_type | comment |
| # col_name | data_type | comment |
| name | string | NULL |
| # Detailed Table Information | data_type | comment |
| Database: | NULL | NULL |
| Owner: | hadoop | NULL |
| CreateTime: | Wed Mar 01 01:11:59 UTC 2023 | NULL |
| LastAccessTime: | UNKNOWN | NULL |
| Retention: | 0 | NULL |
| Location: | hdfs://ip-172-31-62-144.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart | NULL |
| Table Type: | MANAGED_TABLE | NULL |
| Table Parameters: | NULL | NULL |
| | COLUMN_STATS_ACCURATE | {"BASIC_STATS":"true"} |
| | numFiles | 0 |
| | numPartitions | 0 |
| | numRows | 0 |
| | rawDataSize | 0 |
| | totalSize | 0 |
| | transient_lastDdlTime | 1677633119 |
| # Storage Information | data_type | comment |
| Serde Library: | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | NULL |
| InputFormat: | org.apache.hadoop.mapred.TextInputFormat | NULL |
| OutputFormat: | org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat | 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 (0.112 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

40°F Cloudy 7:13 PM 2/28/2023

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.

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.

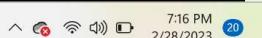
SET hive.exec.dynamic.partition = true;

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

```

hadoop@ip-172-31-62-144:~/hql
+-----+
| CreateTime: | Wed Mar 01 01:11:59 UTC 2023 | NULL |
| LastAccessTime: | UNKNOWN | NULL |
| Retention: | 0 | NULL |
| Location: | hdfs://ip-172-31-62-144.ec2.internal:8020/user/hive/warehouse/mydb.db/foodrat | NULL |
| Table Type: | MANAGED_TABLE | NULL |
| Table Parameters: | NULL | {"BASIC_STATS": "true"} |
| COLUMN_STATS_ACCURATE: | COLUMN_STATS_ACCURATE | 0 |
| numFiles: | numFiles | 0 |
| numPartitions: | numPartitions | 0 |
| numRows: | numRows | 0 |
| rawDataSize: | rawDataSize | 0 |
| totalSize: | totalSize | 0 |
| transient_lastDdlTime: | transient_lastDdlTime | 1677633119 |
| NULL: | NULL | NULL |
| NULL: | 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: | Compressed: No | NULL |
| Num Buckets: | Num Buckets: -1 | NULL |
| Bucket Columns: | Bucket Columns: [] | NULL |
| Sort Columns: | Sort Columns: [] | NULL |
| Storage Desc Params: | Storage Desc Params: NULL | NULL |
| field.delim: | field.delim | , |
| serialization.format: | serialization.format | , |
+-----+
41 rows selected (0.112 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> SET hive.exec.dynamic.partition = true;
No rows affected (0.005 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> SET hive.exec.dynamic.partition.mode = non-strict;
No rows affected (0.004 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)> |

```

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

```

hadoop@ip-172-31-62-144:~/hql
0: jdbc:hive2://localhost:10000/ (MyDb)> SET hive.exec.dynamic.partition.mode = non-strict;
No rows affected (0.004 seconds)
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_20230301012059_049e0e50-d18d-4956-a9d5-c649d44d2a9a): 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_20230301012059_049e0e50-d18d-4956-a9d5-c649d44d2a9a : 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_20230301012059_049e0e50-d18d-4956-a9d5-c649d44d2a9a:4
  DagName:
  Vertices:
    Map 1
      Map Operator Tree:
        TableScan
          alias: foodratings
          Statistics: Num rows: 145 Data size: 17478 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: 17478 Basic stats: COMPLETE Column stats: NONE
            File Output Operator

```

Cloudy 38°F 7:21 PM 2/28/2023

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

```

hadoop@ip-172-31-62-144:~/hql
  GatherStats: false
  MultiFileSpray: false

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

INFO : Completed compiling command(queryId=hive_20230301012204_3ccd2036-9bb2-422c-94cb-f0af9d813812); Time taken: 0.881 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301012204_3ccd2036-9bb2-422c-94cb-f0af9d813812): 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_20230301012204_3ccd2036-9bb2-422c-94cb-f0af9d813812
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_1677629474816_0004)

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_20230301012204_3ccd2036-9bb2-422c-94cb-f0af9d813812); Time taken: 7.21 seconds
INFO : OK
+-----+
| min | max | average |
+-----+
| 1 | 50 | 24.912596401028278 |
+-----+
1 row selected (8.121 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>
```

Cloudy 38°F 7:22 PM 2/28/2023

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-172-31-62-144:~$ hql
0: jdbc:hive2://localhost:10000/ (MyDb)>
0: jdbc:hive2://localhost:10000/ (MyDb)> LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces173981.txt' OVERWRITE INTO TABLE foodplaces;
INFO : Compiling command(queryId=hive_20230301012346_90f8f14b-5843-4300-9f0c-951640787ed9): LOAD DATA LOCAL INPATH '/home/hadoop/foodplaces173981.txt' OVERWRITE INTO TABLE foodplaces
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: schema(fieldschemas:null, properties:null)
INFO : EXPLAIN output for queryid hive_20230301012346_90f8f14b-5843-4300-9f0c-951640787ed9 : STAGE DEPENDENCIES:
  Stage-0 is a root stage [MOVE]
  Stage-1 depends on stages: Stage-0 [STATS]

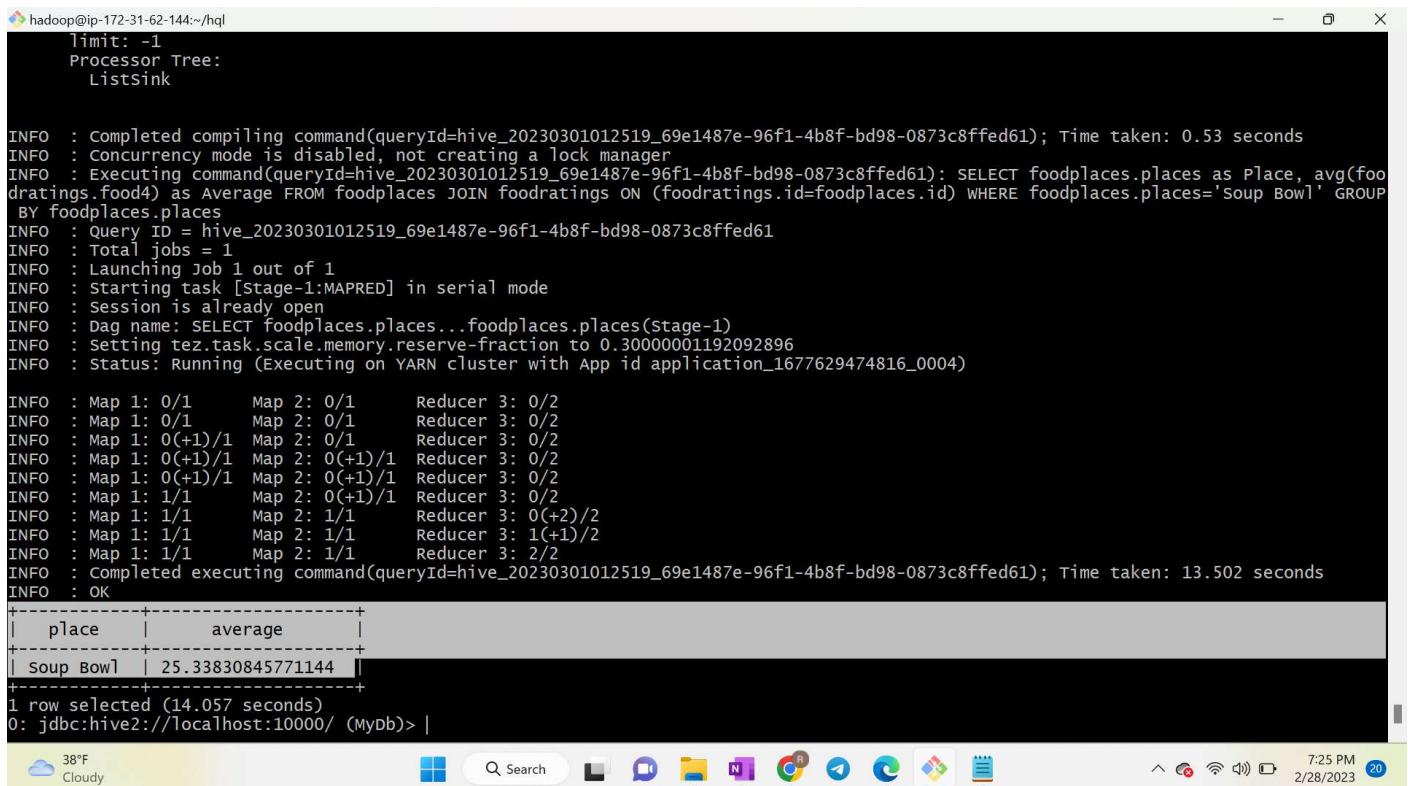
STAGE PLANS:
Stage: Stage-0
  Move Operator
    tables:
      replace: true
      source: file:/home/hadoop/foodplaces173981.txt
      table:
        input format: org.apache.hadoop.mapred.TextInputFormat
        output format: org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat
        properties:
          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
        location hdfs://ip-172-31-62-144.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

```



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-172-31-62-144:~/hql
Limit: -1
Processor Tree:
ListSink

INFO : Completed compiling command(queryId=hive_20230301012519_69e1487e-96f1-4b8f-bd98-0873c8ffed61); Time taken: 0.53 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230301012519_69e1487e-96f1-4b8f-bd98-0873c8ffed61): 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_20230301012519_69e1487e-96f1-4b8f-bd98-0873c8ffed61
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_1677629474816_0004)

INFO : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/2
INFO : Map 1: 0/1      Map 2: 0/1      Reducer 3: 0/2
INFO : Map 1: 0(+1)/1  Map 2: 0/1      Reducer 3: 0/2
INFO : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/2
INFO : Map 1: 0(+1)/1  Map 2: 0(+1)/1  Reducer 3: 0/2
INFO : Map 1: 1/1      Map 2: 0(+1)/1  Reducer 3: 0/2
INFO : Map 1: 1/1      Map 2: 1/1      Reducer 3: 0(+2)/2
INFO : Map 1: 1/1      Map 2: 1/1      Reducer 3: 1(+1)/2
INFO : Map 1: 1/1      Map 2: 1/1      Reducer 3: 2/2
INFO : Completed executing command(queryId=hive_20230301012519_69e1487e-96f1-4b8f-bd98-0873c8ffed61); Time taken: 13.502 seconds
INFO : OK

+-----+
| place |      average      |
+-----+
| Soup Bowl | 25.33830845771144 |
+-----+
1 row selected (14.057 seconds)
0: jdbc:hive2://localhost:10000/ (MyDb)>

```

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:

- When is the most important consideration when choosing a row format and when a column format for your big data file?
- What is “splittability” for a column file format and why is it important when processing large volumes of data?
- What can files stored in column format achieve better compression than those stored in row format?
- Under what circumstances would it be the best choice to use the “Parquet” column file format

