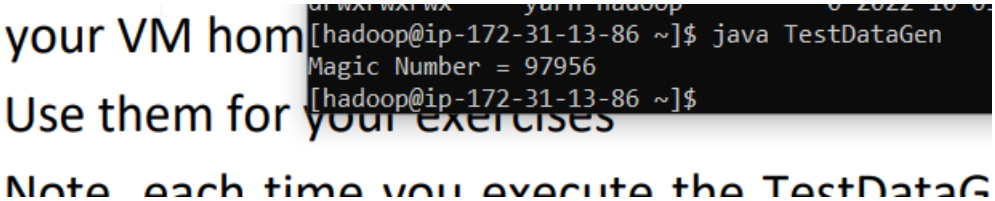


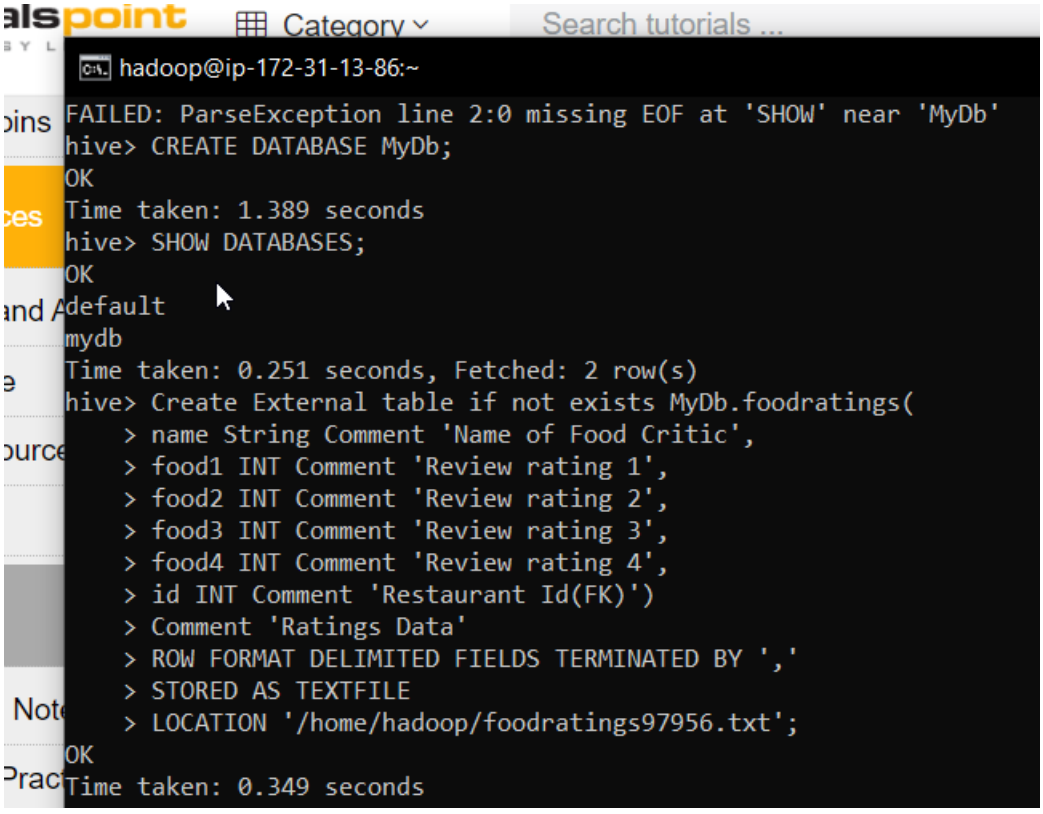
Magic Number: 97956

1.

```
CREATE DATABASE MyDb;

SHOW DATABASES;

CREATE EXTERNAL TABLE IF NOT EXISTS MYDB.foodplaces(
   id INT Comment 'Restaurant ID (PK)',
   place STRING Comment 'Name of Res')
   Comment 'Rest details'
   ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
   STORED AS TEXTFILE
   LOCATION '/home/hadoop/foodplaces97956.txt';
```



DESCRIBE FORMATTED MyDb.foodratings;

```
hive> DESCRIBE FORMATTED MyDB.foodratings;
OK
# col_name              data_type               comment

name                    string                  Name of Food Critic
food1                   int                     Review rating 1
food2                   int                     Review rating 2
food3                   int                     Review rating 3
food4                   int                     Review rating 4
id                      int                     Restaurant Id(FK)

# Detailed Table Information
Database:               mydb
Owner:                  hadoop
CreateTime:             Sat Oct 01 08:21:44 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://ip-172-31-13-86.ec2.internal:8020/home/hadoop/foodratings97956.txt
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        comment                 Ratings Data
        transient_lastDdlTime   1664612504

# 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
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             ,
        serialization.format    ,
Time taken: 0.147 seconds, Fetched: 33 row(s)
```

CREATE EXTERNAL TABLE IF NOT EXISTS MYDB.foodplaces(

id INT Comment 'Restaurant ID (PK)',

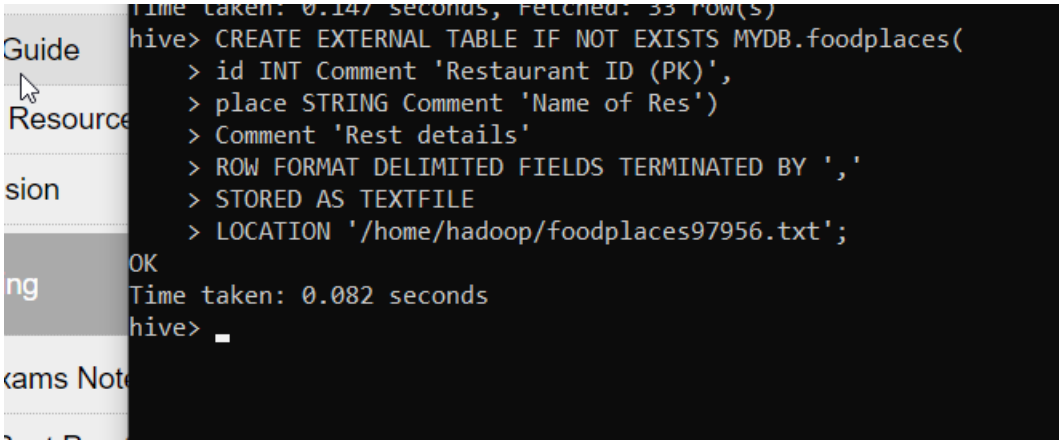
place STRING Comment 'Name of Res')

Comment 'Rest details'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/home/hadoop/foodplaces97956.txt';



DESCRIBE FORMATTED MyDb.foodplaces;

```
hive> DESCRIBE FORMATTED MyDB.foodplaces;
OK
# col_name              data_type               comment

id                      int                     Restaurant ID (PK)
place                   string                  Name of Res

# Detailed Table Information
Database:               mydb
Owner:                  hadoop
CreateTime:             Sat Oct 01 08:27:38 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://ip-172-31-13-86.ec2.internal:8020/home/hadoop/foodplaces97956.txt
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        comment                 Rest details
        transient_lastDdlTime   1664612858

# 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
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             ,
        serialization.format    ,
Time taken: 0.067 seconds, Fetched: 29 row(s)
hive>
```

2.

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

```
hive> select "food3" AS Column_name, min(food3) as MIN, max(food3) AS MAX, avg(food3) AS AVG from MyDb.foodratings;
Query ID = hadoop_20221001083217_3173b14e-dd9d-48bf-b9e7-59c15207970e
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1664609429932_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.51 s
--------------------------------------------------------------------------------
OK
food3   1       50      25.348
Time taken: 14.771 seconds, Fetched: 1 row(s)
hive>
```

3.

SELECT name, min(food1) AS MIN, max(food1) as MAX, AVG(food1) as AVG from MyDb.foodratings Group by name;

```
hive> SELECT name, min(food1) AS MIN, max(food1) as MAX, AVG(food1) as AVG from MyDb.foodratings Group ny name;
FAILED: ParseException line 1:97 missing BY at 'ny' near '<EOF>'
line 1:100 extraneous input 'name' expecting EOF near '<EOF>'
hive> SELECT name, min(food1) AS MIN, max(food1) as MAX, AVG(food1) as AVG from MyDb.foodratings Group by name;
Query ID = hadoop_20221001083424_93670229-6565-4bd4-b9e9-400b38a1ed61
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664609429932_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.61 s
--------------------------------------------------------------------------------
OK
Jill    1       50      24.737373737373737
Joe     1       50      24.555555555555557
Joy     1       50      25.646464646464647
Mel     1       50      24.989583333333332
Sam     1       50      24.653658536585365
Time taken: 5.285 seconds, Fetched: 5 row(s)
```

4.

```
CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodratingspart (
    food1 INT Comment 'Review rating 1',
   food2 INT Comment 'Review rating 2',
    food3 INT Comment 'Review rating 3',
   food4 INT Comment 'Review rating 4',
    id INT Comment 'Restaurant ID (FK)')
    Comment 'Rating data'
    PARTITIONED BY(name STRING Comment 'Name of food Critic')
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    STORED AS TEXTFILE;
```

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS MyDb.foodratingspart (
    > food1 INT Comment 'Review rating 1',
    > food2 INT Comment 'Review rating 2',
    > food3 INT Comment 'Review rating 3',
    > food4 INT Comment 'Review rating 4',
    > id INT Comment 'Restaurant ID (FK)')
    > Comment 'Rating data'
    > PARTITIONED BY(name STRING Comment 'Name of food Critic')
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.086 seconds
```

DESCRIBE FORMATTED MyDB.foodratingspart;

```
hive> DESCRIBE FORMATTED MyDB.foodratingspart;
OK
# col_name              data_type               comment

food1                   int                     Review rating 1
food2                   int                     Review rating 2
food3                   int                     Review rating 3
food4                   int                     Review rating 4
id                      int                     Restaurant ID (FK)

# Partition Information
# col_name              data_type               comment

name                    string                  Name of food Critic

# Detailed Table Information
Database:               mydb
Owner:                  hadoop
CreateTime:             Sat Oct 01 08:38:28 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://ip-172-31-13-86.ec2.internal:8020/user/hive/warehouse/mydb.db/foodratingspart
Table Type:             EXTERNAL_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        EXTERNAL                TRUE
        comment                 Rating data
        numFiles                0
        numPartitions           0
        numRows                 0
        rawDataSize             0
        totalSize               0
        transient_lastDdlTime   1664613508

# 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
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             ,
        serialization.format    ,
Time taken: 0.087 seconds, Fetched: 43 row(s)
```

5.

Answer:

As number of critics are very low, partition on critic names would help to easy query part of data than partitioning data on number of places which are large.

6.

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

```
hive> SET hive.exec.dynamic.partittion;
hive.exec.dynamic.partittion is undefined
hive> SET hive.exec.dynamic.partition;
hive.exec.dynamic.partition=true
hive> SET hive.exec.dynamic.partition.mode;
hive.exec.dynamic.partition.mode=strict
hive> SET hive.exec.dynamic.partition.mode=non-strict;
hive> SET hive.exec.dynamic.partition.mode;
hive.exec.dynamic.partition.mode=non-strict
hive>
```

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

```
hive> INSERT OVERWRITE TABLE MyDb.foodratingspart
    > PARTITION (name)
    > SELECT food1,food2,food3,food4,id,name FROM MyDB.foodratings;
Query ID = hadoop_20221001084809_d4703165-ec86-4ad6-a3ed-672ba6be7c55
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1664609429932_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.26 s
--------------------------------------------------------------------------------
Loading data to table mydb.foodratingspart partition (name=null)

Loaded : 5/5 partitions.
         Time taken to load dynamic partitions: 0.313 seconds
         Time taken for adding to write entity : 0.002 seconds
OK
Time taken: 12.416 seconds
```

SELECT min(food1) as MIN, MAX(Food1) AS MAX, AVG(food1) as AVG from MyDb.foodratingspart where name ="Jill" or name ="Mel";

```
hive> SELECT min(food1) as MIN, MAX(Food1) AS MAX, AVG(food1) as AVG from MyDb.foodratingspart where name ="Jill" or nam
e ="Mel";
Query ID = hadoop_20221001085131_153301c9-5ac1-4ada-aaeb-c01c20261dc1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664609429932_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.09 s
--------------------------------------------------------------------------------
OK
1       50      24.861538461538462
Time taken: 6.039 seconds, Fetched: 1 row(s)
```

7.

select foodp.place, avg(foodr.food4) as AVG from mydb.foodplaces foodp, Mydb.foodratings foodr where foodp.place ="Soup Bowl" and foodp.id =foodr.id group by foodp.place;

```
Time taken: 0.725 seconds
hive> select foodp.place, avg(foodr.food4) as AVG from mydb.foodplaces foodp, Mydb.foodratings foodr where foodp.place =
"Soup Bowl" and foodp.id =foodr.id group by foodp.place;
Query ID = hadoop_20221001085612_6756f362-678f-493b-8035-99bf080a8256
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664609429932_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 7.29 s
--------------------------------------------------------------------------------
OK
Soup Bowl       25.21578947368421
Time taken: 8.056 seconds, Fetched: 1 row(s)
hive>
```

8.

a. Row format is most useful when user has to access data with respect to row values and need to access many rows at a time. This format is used to read and write data optimally.

Column format is useful when the computation is focused on specific columns without the need to search row values. This format is used to read and compute optimally.

b.

The ability to breakdown a file into smaller parts which are not dependent on each other is called Splitability. Splitability on column file format can be done when the query computation is focused on one column. Which splits the data based on columns thus, making the computation optimized.

c.

Storing data of same type side by side allows us to compress better than stored in row by row.

d.

Parquet is used in Hadoop analytical database like (Implaca). It is specially used in analysing huge dataset with multiple columns for computations.