Ruchi Sharma A20429225

Magic Number: 125235

```
[ruchisharma26@sandbox-host ~]$ scp -P 2222 ~/Desktop/BigData/HW4/TestDataGen.class maria_dev@localhost:/maria_dev/home
maria_dev@localhost's password:
/home/ruchisharma26/Desktop/BigData/HW4/TestDataGen.class: No such file or directory
[ruchisharma26@sandbox-host ~]$ scp -P 2222 ~/Desktop/BigData/HW4/TestDataGen.class maria_dev@localhost:/maria_dev/home
maria_dev@localhost's password:
Authentication failed.
lost connection
[ruchisharma26@sandbox-host ~]$ scp -P ~/Desktop/BigData/HW4/TestDataGen.class maria_dev@localhost:/home/maria_dev
usage: scp [-12346BCpqrv] [-c cipher] [-F ssh_config] [-i identity_file]
           [-l limit] [-o ssh_option] [-P port] [-S program]
           [[user@]host1:]file1 ... [[user@]host2:]file2
[ruchisharma26@sandbox-host ~]$ scp -P 2222 ~/Desktop/BigData/HW4/TestDataGen.class maria_dev@localhost:/home/maria_dev
maria_dev@localhost's password:
/home/ruchisharma26/Desktop/BigData/HW4/TestDataGen.class: No such file or directory
[ruchisharma26@sandbox-host ~]$ java TestDataGen
-bash: java: command not found
[ruchisharma26@sandbox-host ~]$ ssh -p 2222 maria_dev@localhost
maria_dev@localhost's password:
Last login: Mon Feb 11 21:07:43 2019 from 172.17.0.1
[maria_dev@sandbox-hdp ~]$ java TestDataGen
Magic Number = 125235
[maria_dev@sandbox-hdp ~]$
```

Exercise 1)

a) Foodratings

```
hive> describe FORMATTED MyDb.foodratings;
OK
# col_name              data_type               comment

name                    string                  food critic name
food1                   int                     rating 1
food2                   int                     rating2
food3                   int                     rating3
food4                   int                     rating4
id                      int                     UserID (FK)

# Detailed Table Information
Database:               mydb
Owner:                  maria_dev
CreateTime:             Sun Feb 17 02:00:08 UTC 2019
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://sandbox-hdp.hortonworks.com:8020/apps/hive/warehouse/mydb.db/foodratings
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        comment                 food rating by critic
        numFiles                0
        numRows                 0
        rawDataSize             0
        totalSize               0
        transient_lastDdlTime   1550368808

# 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: 0.564 seconds, Fetched: 38 row(s)
hive>
```

b) Foodpalces

```
hive> describe FORMATTED MyDb.foodplaces;
OK
# col_name              data_type               comment

id                      int
place                   string

# Detailed Table Information
Database:               mydb
Owner:                  maria_dev
CreateTime:             Sun Feb 17 02:18:48 UTC 2019
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://sandbox-hdp.hortonworks.com:8020/apps/hive/warehouse/mydb.db/foodplaces
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        numFiles                0
        numRows                 0
        rawDataSize             0
        totalSize               0
        transient_lastDdlTime   1550369928

# 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: 0.545 seconds, Fetched: 33 row(s)
```

Exercise 2) (Magic Number: 125235)

```
hive> LOAD DATA LOCAL INPATH '/home/maria_dev/foodratings125235.txt' OVERWRITE INTO TABLE foodratings;
Loading data to table mydb.foodratings
Table mydb.foodratings stats: [numFiles=1, numRows=0, totalSize=17499, rawDataSize=0]
OK
Time taken: 1.39 seconds
hive> SELECT min(food3), max(food3), avg(food3) from foodratings;
Query ID = maria_dev_20190217025434_cd7ec6ff-eae4-4f56-b249-b789c3a30735
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_1550356157440_0003)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.14 s
--------------------------------------------------------------------------------
OK
1       50      25.861
Time taken: 11.85 seconds, Fetched: 1 row(s)
hive>
```

Command: LOAD DATA LOCAL INPATH '/home/maria_dev/foodratings125235.txt' OVERWRITE INTO TABLE foodratings;

Hive Command: SELECT min(food3), max(food3), avg(food3) from foodratings;

Exercise 3) (Magic Number: 125235)

```
hive> Select name,min(food1),max(food1),avg(food1) from foodratings GROUP BY foodratings.name;
Query ID = maria_dev_20190217030905_775d9a4d-7efc-4042-ad32-f71d610f6662
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_1550356157440_0004)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.29 s
--------------------------------------------------------------------------------
OK
Jill    1       50      26.383084577114428
Joe     1       50      25.601990049751244
Joy     1       50      25.425
Mel     1       50      24.439393939393938
Sam     1       50      24.66
Time taken: 12.908 seconds, Fetched: 5 row(s)
hive>
```

Command: Select name, min(food1), max(food1), avg(food1) from foodratings GROUP BY foodratings.name;

Exercise 4) **(Magic Number: 125235)**

```
hive> CREATE TABLE IF NOT EXISTS MyDb.foodratingspart(food1 int , food2 int, food3 int , food4 int, id int )
    > PARTITIONED BY (name String)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ","
    > STORED AS TEXTFILE;
OK
Time taken: 0.735 seconds
hive> describe FORMATTED MyDb.foodratingspart;
OK
# col_name              data_type               comment

food1                   int
food2                   int
food3                   int
food4                   int
id                      int

# Partition Information
# col_name              data_type               comment

name                    string

# Detailed Table Information
Database:               mydb
Owner:                  maria_dev
CreateTime:             Sun Feb 17 03:23:19 UTC 2019
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://sandbox-hdp.hortonworks.com:8020/apps/hive/warehouse/mydb.db/foodratingspart
Table Type:             MANAGED_TABLE
Table Parameters:
        transient_lastDdlTime   1550373799

# 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: 0.59 seconds, Fetched: 36 row(s)
hive>
```

CREATE TABLE IF NOT EXISTS MyDb.foodratingspart (food1 int, food2 int, food3 int, food4 int, id int)

PARTITIONED BY (name String)

ROW FORMAT DELIMITED FIELDSTERMINATED BY “,”

STORED AS TEXTFILE;

Exercise 5) Configuration:

```
Time taken: 0.59 seconds, Fetched: 36 row(s)
hive> SET hive.exec.dynamic.partition=true;
hive> SET hive.exec.dynamic.partition.mode=non-strict
    > ;
hive> set hive.exec.max.dynamic.partitions=1000;
hive> set hive.exec.max.dynamic.partitions.pernode=1000;
hive> INSERT OVERWRITE TABLE foodratingspart
    > PARTITION (name)
    > SELECT food1,food2,food3,food4,id,name
    >
    >
    > FROM foodratings
    > DISTRIBUTE BY name;
Query ID = maria_dev_20190217062746_90313de1-0767-462e-b4fa-51eb5e9935f9
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_1550356157440_0005)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.14 s
--------------------------------------------------------------------------------
Loading data to table mydb.foodratingspart partition (name=null)
         Time taken to load dynamic partitions: 1.794 seconds
        Loading partition {name=Joe}
        Loading partition {name=Jill}
        Loading partition {name=Joy}
        Loading partition {name=Sam}
        Loading partition {name=Mel}
         Time taken for adding to write entity : 1
Partition mydb.foodratingspart{name=Jill} stats: [numFiles=1, numRows=201, totalSize=2668, rawDataSize=2467]
Partition mydb.foodratingspart{name=Joe} stats: [numFiles=1, numRows=201, totalSize=2651, rawDataSize=2450]
Partition mydb.foodratingspart{name=Joy} stats: [numFiles=1, numRows=200, totalSize=2677, rawDataSize=2477]
Partition mydb.foodratingspart{name=Mel} stats: [numFiles=1, numRows=198, totalSize=2645, rawDataSize=2447]
Partition mydb.foodratingspart{name=Sam} stats: [numFiles=1, numRows=200, totalSize=2657, rawDataSize=2457]
OK
```

```
ve.BooleanObjectInspector
hive> Select min(food2),max(food2),avg(food2) from foodratingspart where name='Mel'or name='Jill';
Query ID = maria_dev_20190217064447_3a15e4c2-ae13-4a6d-8427-93d0d1216460
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_1550356157440_0006)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.84 s
--------------------------------------------------------------------------------
OK
1       50      25.007518796992482
Time taken: 12.998 seconds, Fetched: 1 row(s)
hive>
```

Setting Dynamic Partitioning:

SET hive.exec.dynamic.partition=true;

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

set hive.exec.max.dynamic.partitions=1000;

set hive.exec.max.dynamic.partitions.pernode=1000;

(As provided configuration were not sufficient at my end, I browsed and found these two commands and kept it as configuration command)

Load Data from Non-Partitioned to Partitioned:

INSERT OVERWRITE TABLE foodratingspart

PARTITION (name)

SELECT food1, food2, food3,food4,id,name

FROM foodratings;

Hive Command:

Select min(food2), max(food2), avg(food2) from foodratingspart where name='Mel' or name='Jill';

Exercise 6)

(Magic Number: 125235)

LOAD DATA LOCAL INPATH '/home/maria_dev/foodplaces125235.txt' OVERWRITE INTO TABLE

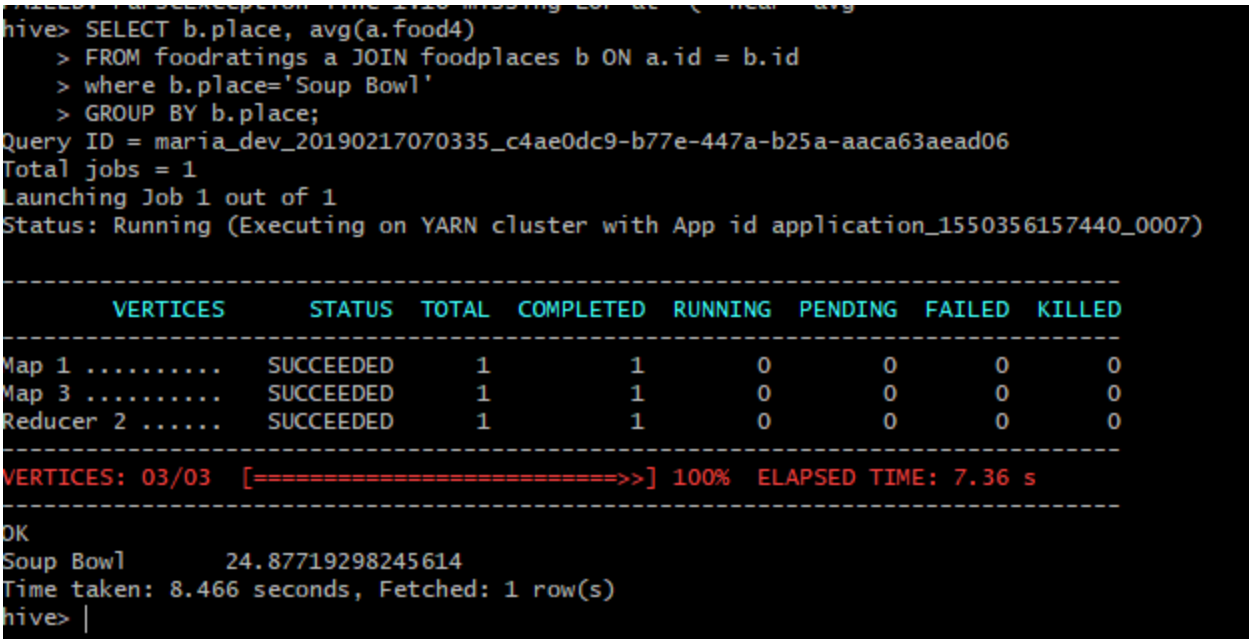
foodplaces;

SELECT b.place avg(a.food4)

FROM foodratings a JOIN foodplaces b ON a.id = b.id

where b.place='Soup Bowl'

GROUP BY b.place;



Exercise 7)

The article “Pig latin: a not-so-foreign language for data processing” describes a new language called Pig Latin. It aimed to provide a middle way between the declaration SQL style language (which many developers find unnatural) and the procedural mapping model (very low level and hard). It also offers anew, interactive debug environment called Pig Pen that can lead to even higher productivity gains.

Pig Latin allows optimization, by reordering the code whereas this is not possible in the opaque Map or Reduce function. Provides Quick start as Pig Latin can work directly on data with a proper function to parse over the content into tuples which avoids time consuming data imports. Allows fully ‘Nested Data Model’ and allows complex, non-atomic data types such as set, map, and tuple to occur as fields of a table. Supports custom processing through UDF (User Defined Functions).

The article also describes a novel debugging environment for Pig, called Pig Pen. In conjunction with the step-by-step nature of our Pig Latin language, Pig Pen makes it easy and fast for users to construct and debug their programs in an incremental fashion.

They also offer a way where we could freeze the progression of program after testing and move further without worrying about the previous freeze code. The Pig system compiles Pig Latin expressions into a sequence of map-reduce jobs, and orchestrates the execution of these jobs on Hadoop, an open-source scalable map-reduce implementation.

This section which is based on conciseness, completeness and realism. In the 6th section, the author, on a high level, talks about the use of Pig Latin from group-by-aggregate and rollup queries to more complex tasks like temporal and session analysis. Section 7 compares the Pig Latin with other technologies at Google, Amazon, etc

Pros: Pig Latin is more natural looking, and reusable compared to SQL. It can operate over plain input files without schema information and costly data import operations. Its nested data model allows complex, non-atomic data types. It comes with an interactive debugging environment. It can work on multiple data sets and can efficiently generate aggregated results. Pig is not tied to Hadoop only and architected to work with other execution platforms. Pig is open source and available for general use.

Cons: Due to its nested data model Pig may generate huge amount of intermediate data. As Pig currently relies on map-reduce jobs, intermediate data need to be generated, transferred and stored in the distributed file system multiple times affecting system throughput and latency.