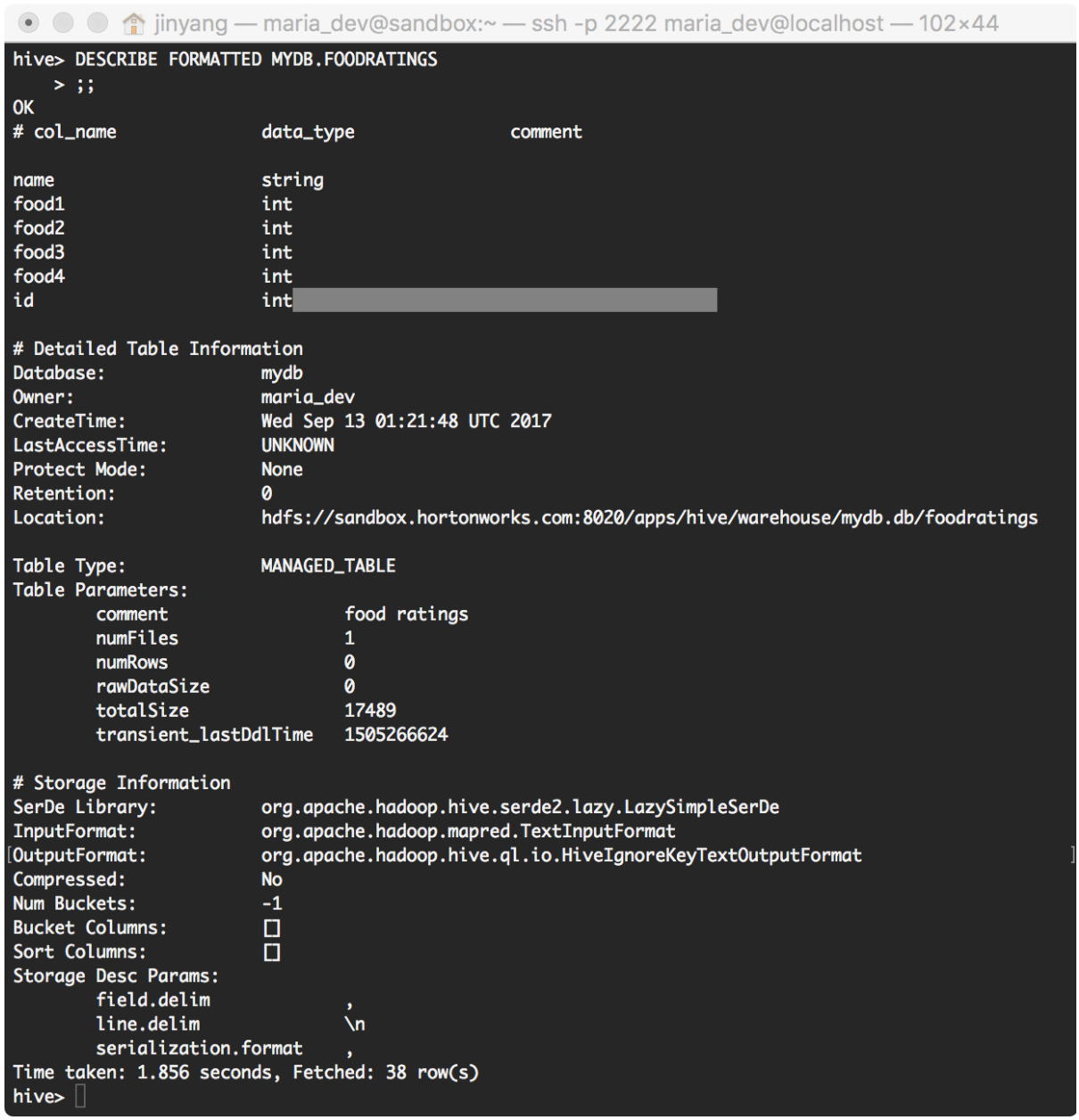
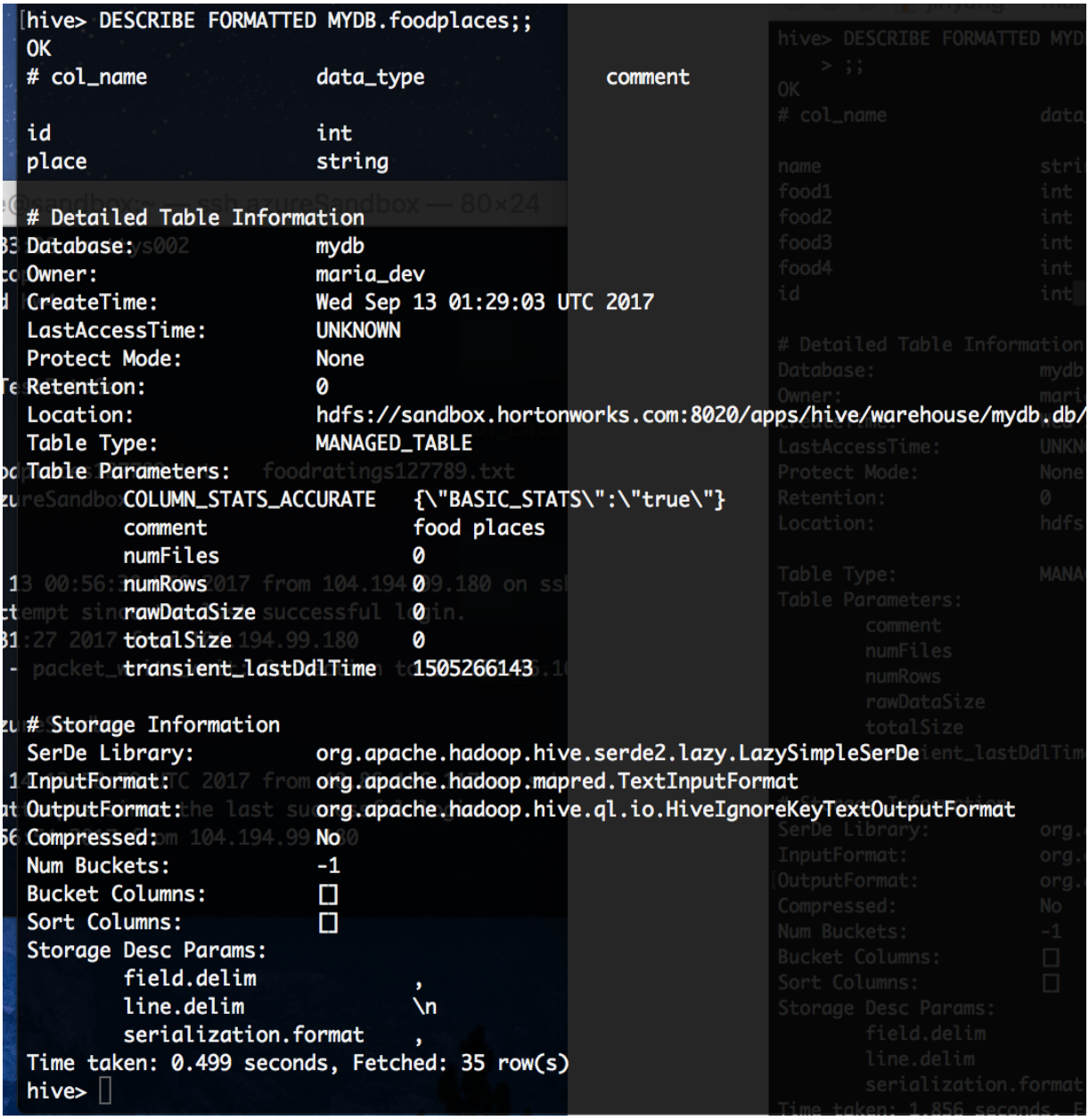
CS595 HW4

E1)


jinyang — maria_dev@sandbox:~ — ssh -p 2222 maria_dev@localhost — 102×44
hive> DESCRIBE FORMATTED MYDB.FOODRATINGS
> ;;
OK
col_name data_type comment
name string
food1 int
food2 int
food3 int
food4 int
id int
Detailed Table Information
Database: mydb
Owner: maria_dev
CreateTime: Wed Sep 13 01:21:48 UTC 2017
LastAccessTime: UNKNOWN
Protect Mode: None
Retention: 0
Location: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/mydb.db/foodratings
Table Type: MANAGED_TABLE
Table Parameters:
comment food ratings
numFiles 1
numRows 0
rawDataSize 0
totalSize 17489
transient_lastDdlTime 1505266624
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 ,
line.delim \n
serialization.format ,
Time taken: 1.856 seconds, Fetched: 38 row(s)
hive>

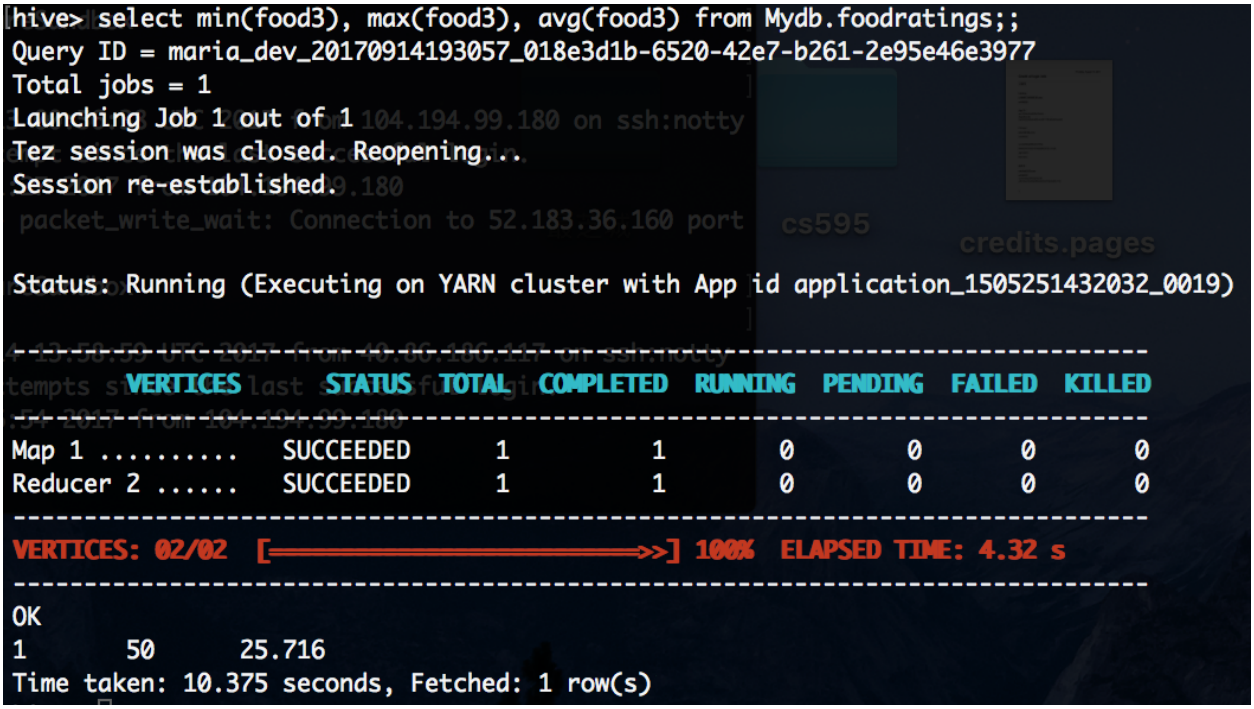

hive> DESCRIBE FORMATTED MYDB.foodplaces;;
OK
col_name data_type comment
id int
place string
Detailed Table Information
Database: mydb
Owner: maria_dev
CreateTime: Wed Sep 13 01:29:03 UTC 2017
LastAccessTime: UNKNOWN
Protect Mode: None
Retention: 0
Location: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/mydb.db/
Table Type: MANAGED_TABLE
Table Parameters:
COLUMN_STATS_ACCURATE {\"BASIC_STATS\":\"true\"}
comment food places
numFiles 0
numRows 0
rawDataSize 0
totalSize 0
transient_lastDdlTime 1505266143
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 ,
line.delim \n
serialization.format ,
Time taken: 0.499 seconds, Fetched: 35 row(s)
hive>

E2)
Magic Number 127789

Query:

hive> load data local inpath '/home/maria_dev/foodratings127789.txt' into table mydb.foodratings;;

hive> select min(food3), max(food3), avg(food3) from Mydb.foodratings;;

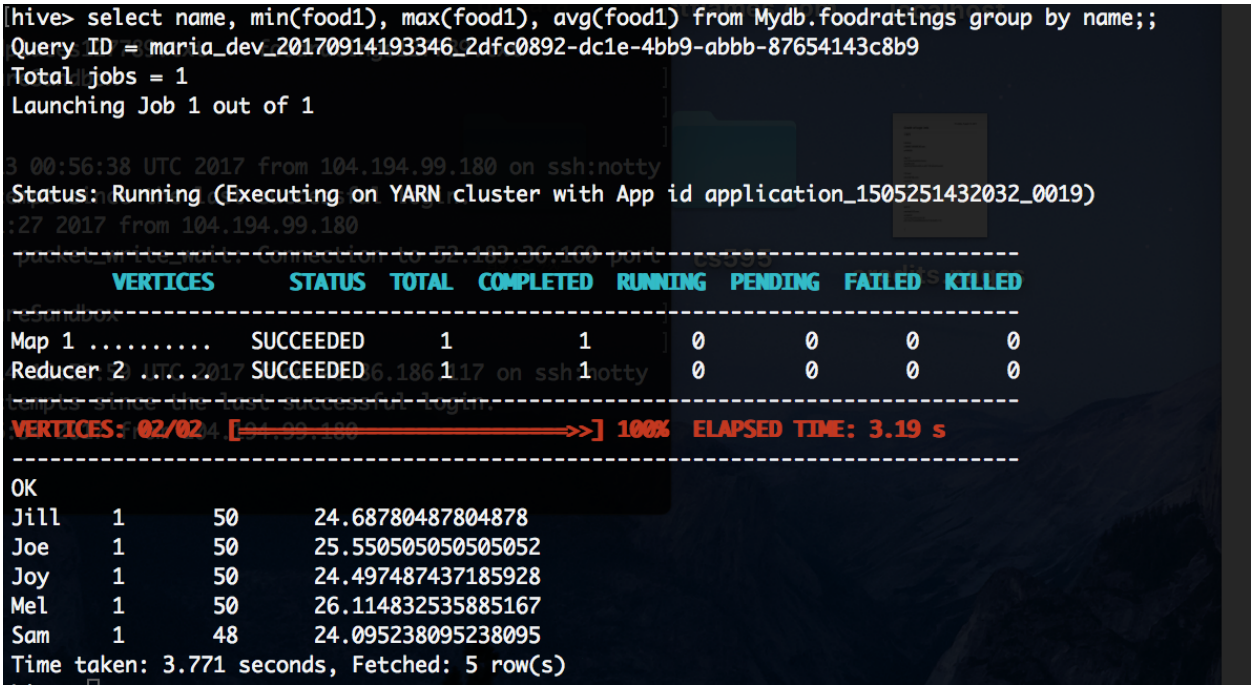


E3)

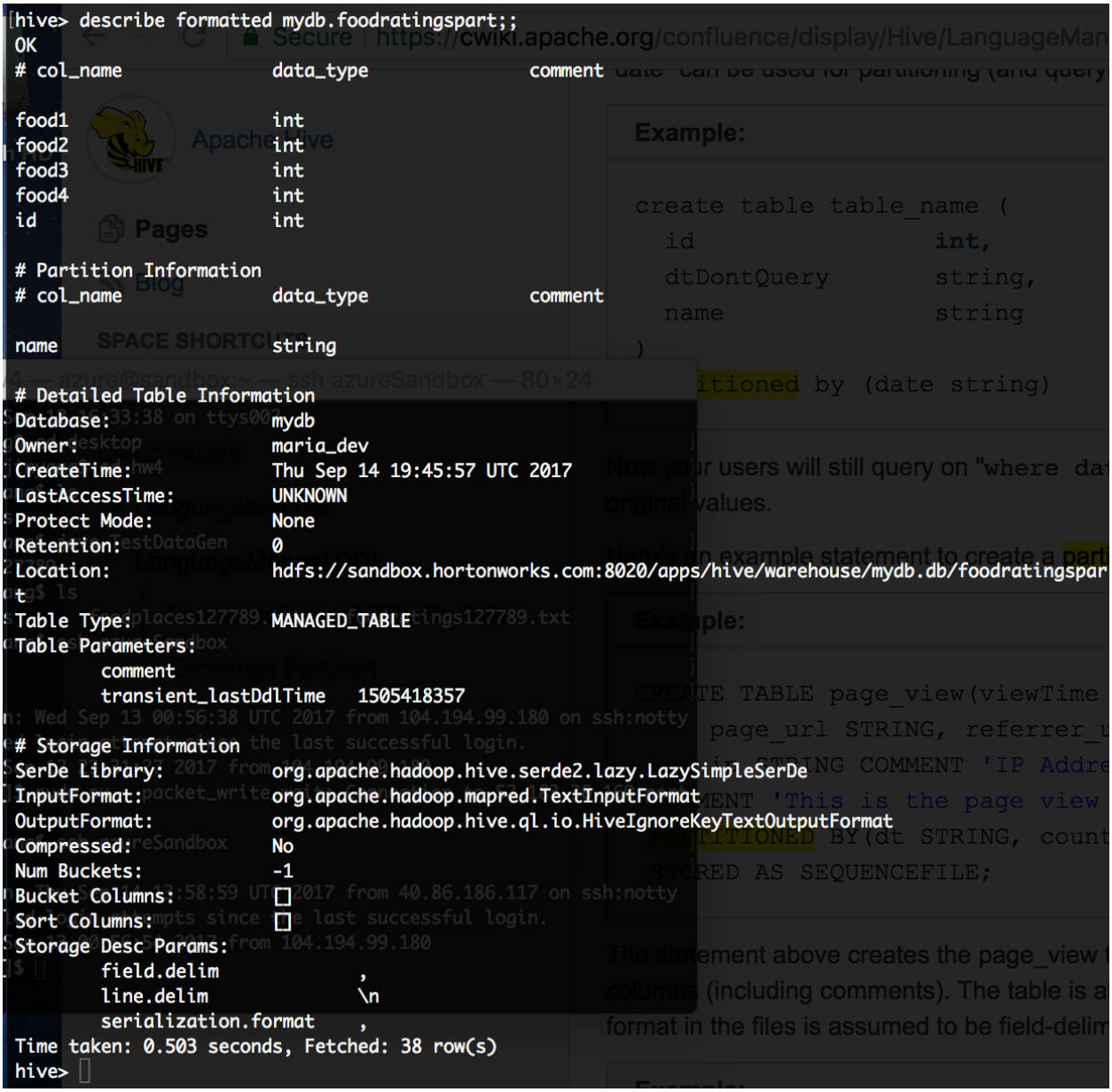
Magic Number 127789

Query:

hive> select name, min(food1), max(food1), avg(food1) from Mydb.foodratings group by name;;



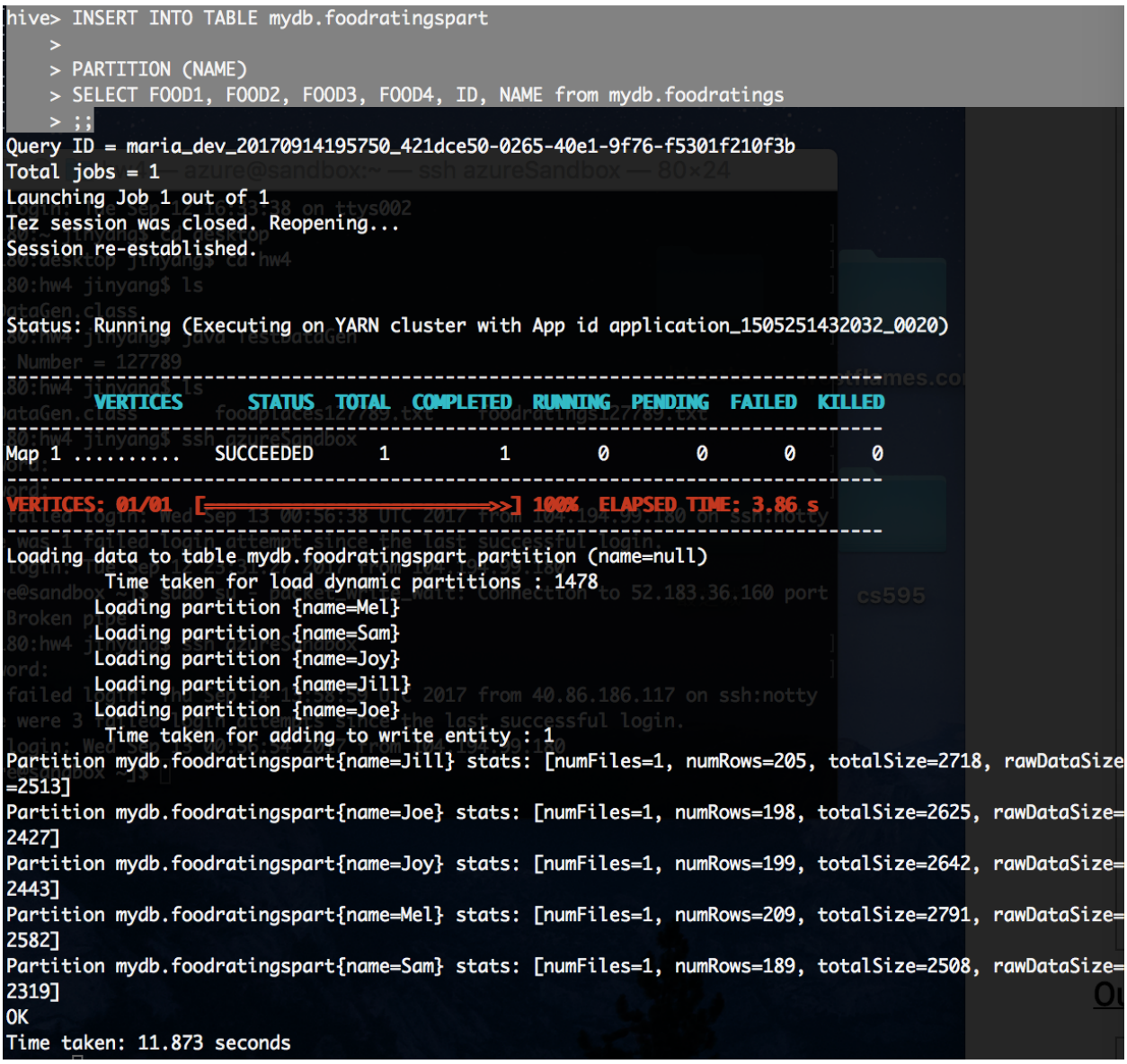
E4)


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: Thu Sep 14 19:45:57 UTC 2017
LastAccessTime: UNKNOWN
Protect Mode: None
Retention: 0
Location: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/mydb.db/foodratingspar
t
Table Type: MANAGED_TABLE
Table Parameters:
comment
transient_lastDdlTime 1505418357
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 ,
line.delim \n
serialization.format ,
Time taken: 0.503 seconds, Fetched: 38 row(s)
hive>

E5)

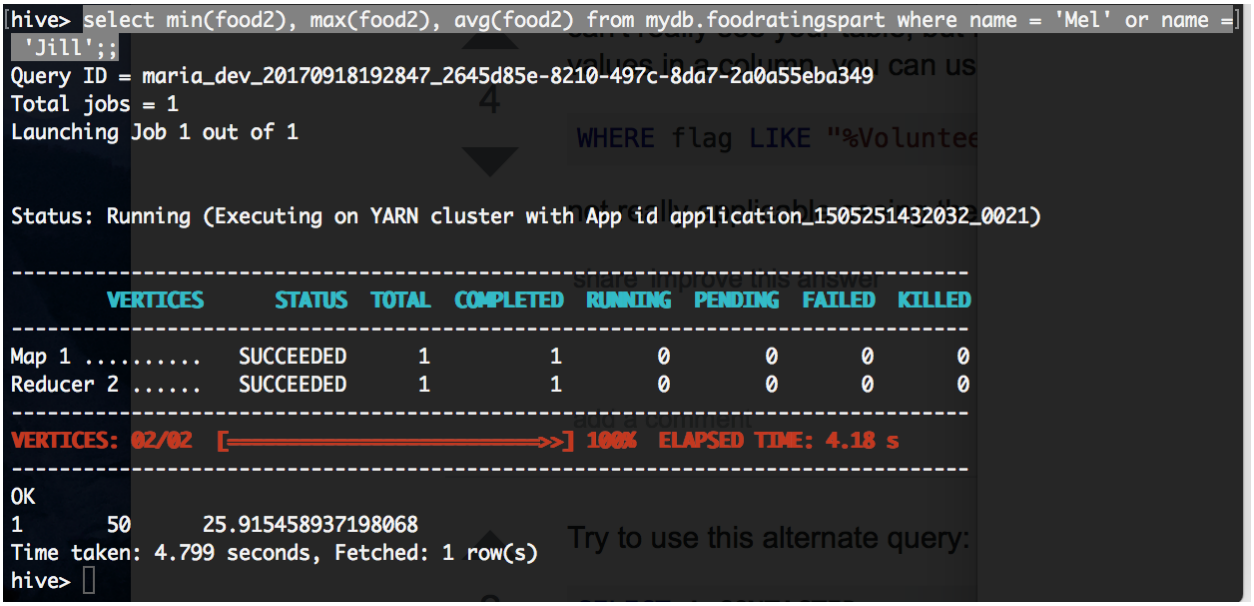
Query:

```
hive> INSERT INTO TABLE mydb.foodratingspart
    >
    > PARTITION (NAME)
    > SELECT FOOD1, FOOD2, FOOD3, FOOD4, ID, NAME from mydb.foodratings
    > ;;
```

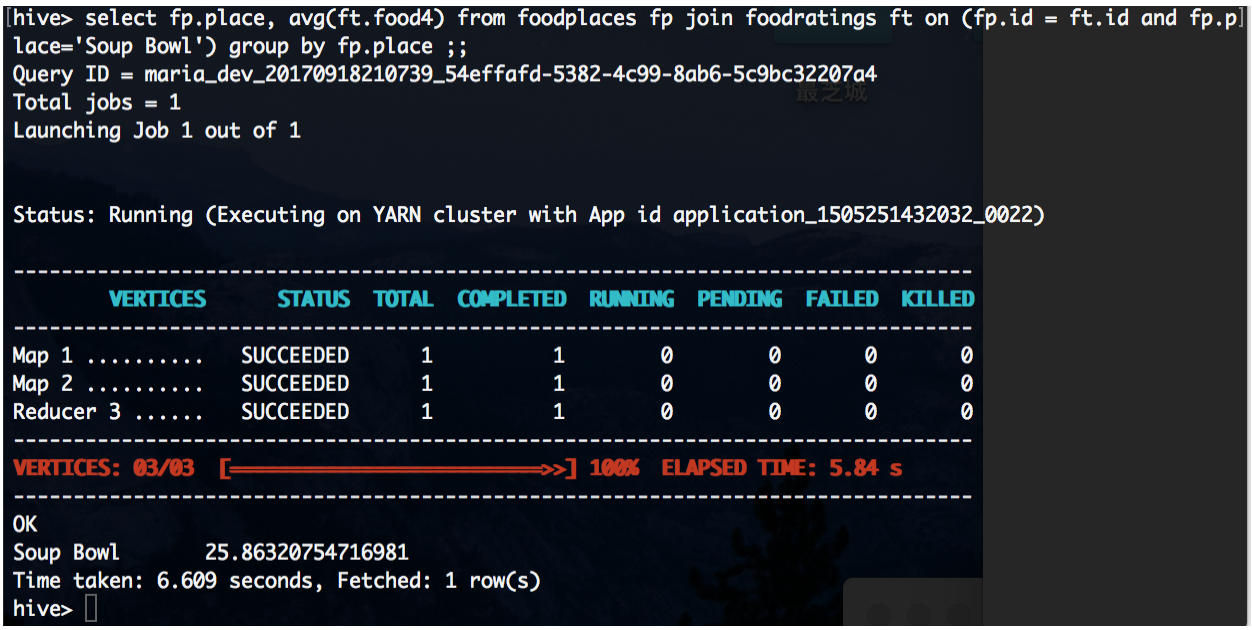


```
hive> INSERT INTO TABLE mydb.foodratingspart
    >
    > PARTITION (NAME)
    > SELECT FOOD1, FOOD2, FOOD3, FOOD4, ID, NAME from mydb.foodratings
    > ;;
Query ID = maria_dev_20170914195750_421dce50-0265-40e1-9f76-f5301f210f3b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.

Status: Running (Executing on YARN cluster with App id application_1505251432032_0020)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 3.86 s
--------------------------------------------------------------------------------
Loading data to table mydb.foodratingspart partition (name=null)
         Time taken for load dynamic partitions : 1478
        Loading partition {name=Mel}
        Loading partition {name=Sam}
        Loading partition {name=Joy}
        Loading partition {name=Jill}
        Loading partition {name=Joe}
         Time taken for adding to write entity : 1
Partition mydb.foodratingspart{name=Jill} stats: [numFiles=1, numRows=205, totalSize=2718, rawDataSize
=2513]
Partition mydb.foodratingspart{name=Joe} stats: [numFiles=1, numRows=198, totalSize=2625, rawDataSize=
2427]
Partition mydb.foodratingspart{name=Joy} stats: [numFiles=1, numRows=199, totalSize=2642, rawDataSize=
2443]
Partition mydb.foodratingspart{name=Mel} stats: [numFiles=1, numRows=209, totalSize=2791, rawDataSize=
2582]
Partition mydb.foodratingspart{name=Sam} stats: [numFiles=1, numRows=189, totalSize=2508, rawDataSize=
2319]
OK
Time taken: 11.873 seconds
```

Hive query:
select min(food2), max(food2), avg(food2) from mydb.foodratingspart where name = 'Mel' or name = ‘Jill';;



E6)



E7)

Pig is a data processing environment in Hadoop that is specifically targeted towards procedural programmers who perform large-scale data analysis. Pig-Latin offers high-level data manipulation in a procedural style.

The advantage is:

Quick - Exploit parallel processing power of a distributed system.

Easy - Be able to write a program or query without a huge learning curve, as well as have some common analysis tasks predefined.

Flexible - transform a data set into workable structure without much overhead. As well as ability to perform customized processing

Transparent - have a say in how the data processing is exited on the system.