

CS595 HW4

E1)

```
jinyang — maria_dev@sandbox:~ — ssh -p 2222 maria_dev@localhost — 102x44
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>
```

# 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: foodratings127789.txt

COLUMN_STATS_ACCURATE	{"BASIC_STATS":"true"}
comment	food places
numFiles	0

13 00:56:20 2017 from 104.194.0.9.180 on ssh attempt sin rawDataSize successful 10 in. 31:27 2017 totalSize 194.99.180 0 - packet_transient_lastDdlTime t:1505266143.105.100000000

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> 
```

# col_name	data_type
name	string
food1	int
food2	int
food3	int
food4	int
id	int

Detailed Table Information

Database:	mydb
Owner:	mari
LastAccessTime:	UNKN
Protect Mode:	None
Retention:	0
Location:	hdfs
Table Type:	MANA

Table Parameters:

comment	
numFiles	
numRows	
rawDataSize	
totalSize	

Storage Information

SerDe Library:	org.
InputFormat:	org.
OutputFormat:	org.
Compressed:	No
Num Buckets:	-1
Bucket Columns:	□
Sort Columns:	□
Storage Desc Params:	
field.delim	
line.delim	
serialization.format	

Time taken: 1.856 seconds, Fetched: 35 row(s)

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;;
```

```
hive> select min(food3), max(food3), avg(food3) from Mydb.foodratings;;
Query ID = maria_dev_20170914193057_018e3d1b-6520-42e7-b261-2e95e46e3977
Total jobs = 1
Launching Job 1 out of 1 104.194.99.180 on ssh:notty
Tez session was closed. Reopening...
Session re-established 9.180
packet_write_wait: Connection to 52.183.36.160 port cs595 credits.pages
Status: Running (Executing on YARN cluster with App id application_1505251432032_0019)
-----
Attempts: 0 VERTICES last STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
----- 2017-09-14 10:30:00
Map 1 ..... SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... SUCCEEDED 1 1 0 0 0 0
-----
VERTICES: 02/02 [=====>] 100% ELAPSED TIME: 4.32 s
-----
OK
1      50      25.716
Time taken: 10.375 seconds, Fetched: 1 row(s)
```

E3)

Magic Number 127789

Query:

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

```
[hive> select name, min(food1), max(food1), avg(food1) from Mydb.foodratings group by name;;
Query ID = maria_dev_20170914193346_2dfc0892-dc1e-4bb9-abbb-87654143c8b9
Total jobs = 1
Launching Job 1 out of 1

3:00:56:38 UTC 2017 from 104.194.99.180 on ssh:notty
Status: Running (Executing on YARN cluster with App id application_1505251432032_0019)
:27 2017 from 104.194.99.180

VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----[REDACTED]-----
Map 1 ..... SUCCEEDED 1 1 0 0 0 0
Reducer 2 JTC 2017 SUCCEEDED 1 1 0 0 0 0
-----[REDACTED]-----
VERTICES: 02/02 [==>] 100% ELAPSED TIME: 3.19 s
-----[REDACTED]-----
OK
Jill    1      50      24.68780487804878
Joe     1      50      25.550505050505052
Joy     1      50      24.497487437185928
Mel     1      50      26.114832535885167
Sam     1      48      24.095238095238095
Time taken: 3.771 seconds, Fetched: 5 row(s)
```

E4)

```
[hive] describe formatted mydb.foodratingspart;;  
OK  
# col_name          data_type      comment  
food1              int           date can be used for partitioning (and query  
food2              int           ApacheHive  
food3              int             
food4              int             
id                 int             
  
# Partition Information  
# col_name          data_type      comment  
name               string           
SPACE SHORTCUT  
# 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$ ls  
$ 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:  2017 from 40.86.186.117 on ssh:notty  
$ Sort Columns:    2017 from 40.86.186.117 on ssh:notty  
$ Storage Desc Params:  
  field.delim      ,  
  line.delim       \n  
  serialization.format ,  
Time taken: 0.503 seconds, Fetched: 38 row(s)  
hive>   
Example:  
create table table_name (  
  id                      int,  
  dtDontQuery            string,  
  name                    string  
)  
partitioned by (date string)  
Now your users will still query on "where date = ..."  
original values.  
Here's an example statement to create a part  
$ hive -e "CREATE TABLE page_view(viewTime  
  page_url STRING, referrer_u  
  COMMENT 'IP Address'  
  COMMENT 'This is the page view'  
  dt STRING, count  
  PARTITIONED BY(dt STRING, count  
  STORED AS SEQUENCEFILE;  
$ The statement above creates the page_view  
columns (including comments). The table is a  
format in the files is assumed to be field-delim
```

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  azure@ sandbox:~ — ssh azureSandbox — 80x24
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
80:desktop jinyang$ cd hw4
80:hw4 jinyang$ ls
dataGen.class
80:hw4 jinyang$ java TestdataGen
Number = 127789
80:hw4 jinyang$ ls
dataGen.class foodrates127789.txt foodratings127789.txt
80:hw4 jinyang$ ssh azureSandbox
Status: Running (Executing on YARN cluster with App id application_1505251432032_0020)
Number = 127789
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';;
```

```
[hive] select min(food2), max(food2), avg(food2) from mydb.foodratingspart where name = 'Mel' or name = 'Jill';;
Query ID = maria_dev_20170918192847_2645d85e-8210-497c-8da7-2a0a55eba349
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1505251432032_0021)

-----  

      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: 4.18 s  

-----  

OK
1      50      25.915458937198068
Time taken: 4.799 seconds, Fetched: 1 row(s)
Try to use this alternate query:  

hive> ]
```

E6)

```
[hive> select fp.place, avg(ft.food4) from foodplaces fp join foodratings ft on (fp.id = ft.id and fp.place='Soup Bowl') group by fp.place ;;
Query ID = maria_dev_20170918210739_54effaf4-5382-4c99-8ab6-5c9bc32207a4
Total jobs = 1
Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application_1505251432032_0022)

-----

| VERTICES        | STATUS    | TOTAL | COMPLETED | RUNNING | PENDING | FAILED | KILLED |
|-----------------|-----------|-------|-----------|---------|---------|--------|--------|
| Map 1 .....     | SUCCEEDED | 1     | 1         | 0       | 0       | 0      | 0      |
| Map 2 .....     | SUCCEEDED | 1     | 1         | 0       | 0       | 0      | 0      |
| Reducer 3 ..... | SUCCEEDED | 1     | 1         | 0       | 0       | 0      | 0      |


-----VERTICES: 03/03 [—————>] 100% ELAPSED TIME: 5.84 s
-----OK
Soup Bowl      25.86320754716981
Time taken: 6.609 seconds, Fetched: 1 row(s)
hive> ]
```

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.