# ***Problem Statement 3****

## PIG: Write a pig script to find list of companies toping in complaint chart (companies with maximum number of complaints).

**Step 1: Start pig in mapreduce mode.**

*$ mr-jobhistory-daemon.sh start historyserver*

*$ pig*

```
[acadgild@localhost project2.1]$ mr-jobhistory-daemon.sh start historyserver
starting historyserver, logging to /usr/local/hadoop-2.6.0/logs/mapred-acadgild-historyserver-localhost.localdomain.out
```

```
[acadgild@localhost project2.1]$ pig
2017-08-25 17:08:53,786 INFO  [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
2017-08-25 17:08:53,789 INFO  [main] pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2017-08-25 17:08:53,789 INFO  [main] pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-08-25 17:08:53,904 [main] INFO  org.apache.pig.Main - Apache Pig version 0.14.0 (r1640057) compiled Nov 16 2014, 18:02:05
2017-08-25 17:08:53,904 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/project2.1/pig_1503661133904.log
2017-08-25 17:08:53,951 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2017-08-25 17:08:54,378 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtr
acker.address
2017-08-25 17:08:54,378 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-08-25 17:08:54,378 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://loca
lhost:9000
2017-08-25 17:08:54,383 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.used.genericoptionsparser is deprecated. Instead, use m
apreduce.client.genericoptionsparser.used
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-08-25 17:08:54,689 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin
-java classes where applicable
2017-08-25 17:08:55,270 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

**Step 2: register piggybank.jar:**

*REGISTER piggybank.jar;*

*DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage;*

```
grunt> REGISTER piggybank.jar;
grunt>
grunt> DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage;
grunt>
```

CSVExcelStorage will be used to load CSV file into pig relation.

**Step 3: Load file into relation:**

*A = LOAD '/flume_sink2/*' USING CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');*

*describe A;*

```
grunt> A = LOAD '/flume_sink2/*' USING CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2017-08-26 16:49:09,750 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max
2017-08-26 16:49:09,750 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-08-26 16:49:09,751 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> describe A;
Schema for A unknown.
```

Relation ‘A’ has been loaded with CSV file along with below details:

Separator Character: “,”

Multi line treatment of the record: *NO_MULTILINE*

Line Break Type: UNIX

Header of CSV: SKIP_INPUT_HEADER = this will ensure that header of the CSV will not be loaded into relation.

**Step 4: Select the required columns from dataset:**

**Dataset Description**

```
Below is the description of the data set

 Column heading              index  Description
 Date received               0      date on which consumer filed the
                                    complaint

 Product                     1      Type of the product
 Sub-product                 2      Sub product type
 Issue                       3      Issue faced by the consumer
 Sub-issue                   4      Any sub issues if exists
 Consumer complaint          5      Detailed description of complaint
 narrative
 Company public response     6      Company's public response to the
                                    complaint
 Company                     7      Name of the company
 State                       8      State from which consumer filed the
                                    complaint
 ZIP code                    9      Zip code
 Submitted via               10     Channel from which complaint was
                                    submitted
 Date sent to company        11     Date on which consumer forum forwarded
                                    the complaint to company
 Company response to         12     Company's   response to the consumer
 consumer
 Timely response?            13
 Consumer disputed?          14
 Complaint ID                15     Unique complaint id

This data is comma delimited.
```

*B = FOREACH A GENERATE (chararray)$7 AS companyName, (int)$15 AS complainID;*

```
grunt> B = FOREACH A GENERATE (chararray)$7 AS companyName, (int)$15 AS complainID;
grunt>
```

**Step 5: Group by companyName and calculate total complains against it:**

*C = GROUP B BY companyName;*

*D = FOREACH C GENERATE group AS companyName, COUNT(B.complainID) AS complaintCount;*

```
grunt> C = GROUP B BY companyName;
grunt> D = FOREACH C GENERATE group AS companyName, COUNT(B.complainID) AS complaintCount;
grunt>
```

**Step 6: order records by most number of complains:**

*E = ORDER D BY complaintCount DESC;*

```
grunt> E = ORDER D BY complaintCount DESC;
grunt>
```

**Step 7: Store the result in '/user/acadgild/project/USAConsumer/ProblemStatement3' HDFS directory:**

*STORE E INTO '/user/acadgild/project/USAConsumer/ProblemStatement3';*

```
2017-08-26 18:57:07,521 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2017-08-26 18:57:07,537 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.2.0   0.14.0  acadgild        2017-08-26 18:55:48     2017-08-26 18:57:07     GROUP_BY,ORDER_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1503739494588_0029  1       1       10      10      10      10      3       3       3       3       A,B,C,D GROUP_BY,COMBINER
job_1503739494588_0030  1       1       3       3       3       3       3       3       3       3       E       SAMPLER
job_1503739494588_0031  1       1       3       3       3       3       4       4       4       4       E       ORDER_BY        /user/acadgild/project/USAConsumer/ProblemStatement3,

Input(s):
Successfully read 0 records from: "/flume_sink2/*"

Output(s):
Successfully stored 0 records in: "/user/acadgild/project/USAConsumer/ProblemStatement3"

Counters:
Total records written : 0
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1503739494588_0029  ->      job_1503739494588_0030,
job_1503739494588_0030  ->      job_1503739494588_0031,
job_1503739494588_0031


2017-08-26 18:57:07,540 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 18:57:07,545 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 18:57:07,597 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 18:57:07,600 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 18:57:07,651 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 18:57:07,655 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 18:57:07,688 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2017-08-26 18:57:07,689 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 18:57:07,696 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 18:57:07,743 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 18:57:07,749 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 18:57:07,786 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 18:57:07,791 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 18:57:07,834 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2017-08-26 18:57:07,835 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 18:57:07,838 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 18:57:07,885 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 18:57:07,890 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 18:57:07,930 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 18:57:07,933 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 18:57:07,981 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2017-08-26 18:57:07,981 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

**Step 8: Check the content in HDFS directory:**

*$ hadoop fs -ls /user/acadgild/project/USAConsumer/ProblemStatement3/*

*$ hadoop fs -cat /user/acadgild/project/USAConsumer/ProblemStatement3/**

```
[acadgild@localhost project2.2]$ hadoop fs -ls /user/acadgild/project/USAConsumer/ProblemStatement3/
17/08/26 18:59:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2017-08-26 18:57 /user/acadgild/project/USAConsumer/ProblemStatement3/_SUCCESS
-rw-r--r--   1 acadgild supergroup      97859 2017-08-26 18:57 /user/acadgild/project/USAConsumer/ProblemStatement3/part-r-00000
[acadgild@localhost project2.2]$ hadoop fs -cat /user/acadgild/project/USAConsumer/ProblemStatement3/*
17/08/26 18:59:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
Bank of America 51127
Wells Fargo     37182
JPMorgan Chase  29583
Experian        24720
Equifax 24706
Citibank        22136
TransUnion      20054
Ocwen   18868
Capital One     13636
Nationstar Mortgage     11401
U.S. Bancorp    8267
GE Capital Retail       7898
Ditech Financial LLC    7455
Navient 6768
PNC Bank        6128
HSBC    5581
Encore Capital Group    5321
Amex    4756
SunTrust Bank   4224
Discover        3934
TD Bank 3856
Select Portfolio Servicing, Inc 3602
RBS Citizens    2719
Portfolio Recovery Associates, Inc.     2654
Fifth Third Bank        2488
Enhanced Recovery Company, LLC  2259
Seterus 2242
Barclays        2162
M&T Bank        2064
BB&T Financial  2006
Regions 1940
Ally Financial Inc.     1847
Santander Bank US       1729
```

PIG output have been stored successfully with the data separated by TAB (\t) which shows companyName and Count of complains against same company in descending order.

**PIG script is save with name ProblemStatement3.PIG which can be run using below command:**

**$ pig <filepath>/ProblemStatement3.PIG**

## SQOOP: Export the results to mysql.

**Step 1: start mysql/services:**

*$ sudo service mysqld status*

*$ sudo service mysqld start*

*$ sudo service mysqld status*

```
[acadgild@localhost project2.1]$ sudo service mysqld status
[sudo] password for acadgild:
mysqld is stopped
[acadgild@localhost project2.1]$ sudo service mysqld start
Starting mysqld:                                           [  OK  ]
[acadgild@localhost project2.1]$ sudo service mysqld status
mysqld (pid  9463) is running...
[acadgild@localhost project2.1]$
```

$ mysql -u root

```
[acadgild@localhost project2.1]$ mysql -u root
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 2
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
```

Above command launches mysql with user root.

**Step2: create table CompaniesWithMaxComplains column companyName & complaintCount:**

*use db1;*

*show tables;*

*create table CompaniesWithMaxComplains*

*(*

*companyName varchar(75),*

*complaintCount int(6)*

*);*

*describe CompaniesWithMaxComplains;*

*select * from CompaniesWithMaxComplains;*

```
mysql> use db1;
Database changed
mysql> show tables;
+---------------------+
| Tables_in_db1       |
+---------------------+
| ForwardSameDay      |
| TimelyResponse      |
| customer            |
| statewiseBPL80      |
| statewiseBPLacheived |
+---------------------+
5 rows in set (0.00 sec)

mysql> create table CompaniesWithMaxComplains
    -> (
    -> companyName varchar(75),
    -> complaintCount int(6)
    -> );
Query OK, 0 rows affected (0.00 sec)

mysql> describe CompaniesWithMaxComplains;
+---------------+------------+------+-----+---------+-------+
| Field         | Type       | Null | Key | Default | Extra |
+---------------+------------+------+-----+---------+-------+
| companyName   | varchar(75) | YES  |     | NULL    |       |
| complaintCount | int(6)     | YES  |     | NULL    |       |
+---------------+------------+------+-----+---------+-------+
2 rows in set (0.00 sec)

mysql> select * from CompaniesWithMaxComplains;
Empty set (0.00 sec)

mysql>
```

**Step 3: run sqoop export command to get data from output directory of the pig job to mysql table.**

*sqoop export --connect jdbc:mysql://localhost/db1 *

*--username 'root' -P --table 'CompaniesWithMaxComplains' *

*--export-dir '/user/acadgild/project/USAConsumer/ProblemStatement3/' *

*--input-fields-terminated-by '\t' *

*-m 1*

```
[acadgild@localhost project2.2]$ sqoop export --connect jdbc:mysql://localhost/db1 \
> --username 'root' -P --table 'CompaniesWithMaxComplains' \
> --export-dir '/user/acadgild/project/USAConsumer/ProblemStatement3/' \
> --input-fields-terminated-by '\t' \
> -m 1
Warning: /usr/local/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2017-08-26 19:01:20,674 INFO  [main] sqoop.Sqoop: Running Sqoop version: 1.4.5
Enter password:
2017-08-26 19:01:21,609 INFO  [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2017-08-26 19:01:21,609 INFO  [main] tool.CodeGenTool: Beginning code generation
2017-08-26 19:01:21,871 INFO  [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `CompaniesWithMaxComplains` AS t LIMIT 1
2017-08-26 19:01:21,906 INFO  [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `CompaniesWithMaxComplains` AS t LIMIT 1
2017-08-26 19:01:21,914 INFO  [main] orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop-2.6.0
```

```
2017-08-26 19:01:25,657 INFO  [main] Configuration.deprecation: mapred.cache.files.filesizes is deprecated. Instead, use mapreduce.job.cache.files
2017-08-26 19:01:25,768 INFO  [main] mapreduce.JobSubmitter: Submitting tokens for job: job_1503739494588_0032
2017-08-26 19:01:26,017 INFO  [main] impl.YarnClientImpl: Submitted application application_1503739494588_0032 to ResourceManager at /0.0.0.0:8032
2017-08-26 19:01:26,049 INFO  [main] mapreduce.Job: The url to track the job: http://http://localhost:8088/proxy/application_1503739494588_0032/
2017-08-26 19:01:26,049 INFO  [main] mapreduce.Job: Running job: job_1503739494588_0032
2017-08-26 19:01:33,136 INFO  [main] mapreduce.Job: Job job_1503739494588_0032 running in uber mode : false
2017-08-26 19:01:33,137 INFO  [main] mapreduce.Job:  map 0% reduce 0%
2017-08-26 19:01:38,200 INFO  [main] mapreduce.Job:  map 100% reduce 0%
2017-08-26 19:01:38,238 INFO  [main] mapreduce.Job: Job job_1503739494588_0032 completed successfully
```

Sqoop command completed successfully.

**Step 4: check table in mysql:**

*select * from CompaniesWithMaxComplains;*

```
mysql> select * from CompaniesWithMaxComplains;
+-------------------------------------------------------------+---------------+
| companyName                                                 | complaintCount |
+-------------------------------------------------------------+---------------+
| Bank of America                                             |         51127 |
| Wells Fargo                                                 |         37182 |
| JPMorgan Chase                                              |         29583 |
| Experian                                                    |         24720 |
| Equifax                                                     |         24706 |
| Citibank                                                    |         22136 |
| TransUnion                                                  |         20054 |
| Ocwen                                                       |         18868 |
| Capital One                                                 |         13636 |
| Nationstar Mortgage                                         |         11401 |
| U.S. Bancorp                                                |          8267 |
| GE Capital Retail                                           |          7898 |
| Ditech Financial LLC                                        |          7455 |
| Navient                                                     |          6768 |
| PNC Bank                                                    |          6128 |
| HSBC                                                        |          5581 |
| Encore Capital Group                                        |          5321 |
| Amex                                                        |          4756 |
| SunTrust Bank                                               |          4224 |
| Discover                                                    |          3934 |
| TD Bank                                                     |          3856 |
| Select Portfolio Servicing, Inc                             |          3602 |
| RBS Citizens                                                |          2719 |
| Portfolio Recovery Associates, Inc.                         |          2654 |
| Fifth Third Bank                                            |          2488 |
| Enhanced Recovery Company, LLC                              |          2259 |
| Seterus                                                     |          2242 |
| Barclays                                                    |          2162 |
| M&T Bank                                                    |          2064 |
| BB&T Financial                                              |          2006 |
| Regions                                                     |          1940 |
| Ally Financial Inc.                                         |          1847 |
| Santander Bank US                                           |          1729 |
```

**Step 5: Verify if all data have been exported from HDFS to MySQL:**

Check number of lines in the HDFS file directory:

*$ hadoop fs -cat /user/acadgild/project/USAConsumer/ProblemStatement3/* | wc -l*

```
[acadgild@localhost project2.2]$ hadoop fs -cat /user/acadgild/project/USAConsumer/ProblemStatement3/* | wc -l
17/08/26 19:04:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin
3337
[acadgild@localhost project2.2]$
```

Check the count of the Table *ForwardSameDay* in mysql:

*select count(*) from CompaniesWithMaxComplains;*

```
mysql> select count(*) from CompaniesWithMaxComplains;
+----------+
| count(*) |
+----------+
|     3337 |
+----------+
1 row in set (0.00 sec)

mysql>
```

As compared above all the data has been exported from HDFS to mysql using Sqoop.