Problem Statement 1*

PIG: Write a pig script to find no of complaints which got timely response.

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: Take selected column into relation which are required for further analysis:

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 narrative	5	Detailed description of complaint
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 consumer	12	Company's response to the consumer
Timely response?	13	
Consumer disputed?	14	
Complaint ID	15	Unique complaint id

This data is comma delimited.

From above we need only index 15 and 13 , which will be type casted as int and chararray respectively.

B = FOREACH A GENERATE (int)$15 AS complainID, (chararray)$13 AS timelyResponse;

describe B;

```
grunt> B = FOREACH A GENERATE (int)$15 AS complainID, (chararray)$13 AS timelyResponse;
grunt> describe B;
B: {complainID: int,timelyResponse: chararray}
grunt>
```

Step 5: Filter & count for timely response to complaints:

C = FILTER B BY timelyResponse=='Yes';

D = GROUP C all;

E = FOREACH D GENERATE B.timelyResponse, COUNT(C.complainID);

describe E;

```
grunt> C = FILTER B BY timelyResponse=='Yes';
grunt> D = GROUP C all;
grunt> E = FOREACH D GENERATE B.timelyResponse, COUNT(C.complainID);
grunt> describe E;
E: {timelyResponse: chararray,long}
grunt>
```

Step 5: Filter & count the complaints where response wasn't timely:

C = FILTER B BY timelyResponse=='No';

D = GROUP C all;

F = FOREACH D GENERATE B.timelyResponse, COUNT(C.complainID);

describe F;

```
grunt> C = FILTER B BY timelyResponse=='No';
grunt> D = GROUP C all;
grunt> F = FOREACH D GENERATE B.timelyResponse, COUNT(C.complainID);
grunt> describe F;
F: {timelyResponse: chararray,long}
grunt>
```

Step 6: UNION E and F into G:

G = UNION E, F;

describe G;

```
grunt> G = UNION E, F;
grunt> describe G;
G: {timelyResponse: chararray,long}
```

Step 7: Store relation 'G' to location /user/acadgild/project/USAConsumer/ProblemStatement1 in HDFS:

Create directory with below hadoop commands:

$ hadoop fs -mkdir /user/acadgild/project/USAConsumer

$ hadoop fs -ls /user/acadgild/project/

```
[acadgild@localhost project2.2]$ hadoop fs -mkdir /user/acadgild/project/USAConsumer
17/08/26 17:14:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[acadgild@localhost project2.2]$ hadoop fs -ls /user/acadgild/project/
17/08/26 17:14:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 8 items
drwxr-xr-x   - acadgild supergroup          0 2017-08-25 20:24 /user/acadgild/project/StateWiseDevelopment
-rw-r--r--   1 acadgild supergroup     717414 2017-07-15 16:46 /user/acadgild/project/StatewiseDistrictwisePhysicalProgress.xml
-rw-r--r--   1 acadgild supergroup      61111 2017-07-09 14:01 /user/acadgild/project/TitanicData.txt
drwxr-xr-x   - acadgild supergroup          0 2017-08-26 17:14 /user/acadgild/project/USAConsumer
-rw-r--r--   1 acadgild supergroup   69234933 2017-07-08 19:36 /user/acadgild/project/crimes.csv
drwxr-xr-x   - acadgild supergroup          0 2017-07-09 21:16 /user/acadgild/project/female
drwxr-xr-x   - acadgild supergroup          0 2017-07-09 21:16 /user/acadgild/project/male
drwxr-xr-x   - acadgild supergroup          0 2017-07-12 13:40 /user/acadgild/project/twitter
[acadgild@localhost project2.2]$
```

Run below PIG command:

STORE G INTO '/user/acadgild/project/USAConsumer/ProblemStatement1';

```
2017-08-26 17:49:37,675 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2017-08-26 17:49:37,684 [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 17:48:45     2017-08-26 17:49:37     GROUP_BY,FILTER,UNION

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1503739494588_0017  1       1       10      10      10      10      3       3       3       3       A,B,C,D,E,F     MULTI_QUERY,COMBINER
job_1503739494588_0018  2       0       6       6       6       6       0       0       0       0       G       MAP_ONLY        /user/acadgild/project/USAConsumer/ProblemStatement1,

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

Output(s):
Successfully stored 0 records in: "/user/acadgild/project/USAConsumer/ProblemStatement1"

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_0017  ->      job_1503739494588_0018,
job_1503739494588_0018


2017-08-26 17:49:37,685 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 17:49:37,686 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 17:49:37,737 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 17:49:37,742 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 17:49:37,782 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 17:49:37,787 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 17:49:37,822 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2017-08-26 17:49:37,824 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 17:49:37,826 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 17:49:37,865 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 17:49:37,866 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 17:49:37,909 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-08-26 17:49:37,916 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-08-26 17:49:37,950 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning aggregation.
2017-08-26 17:49:37,950 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```


Step 8: Check store files in HDFS:

*$ hadoop fs -ls /user/acadgild/project/USAConsumer/ProblemStatement1/**

*$ hadoop fs -cat /user/acadgild/project/USAConsumer/ProblemStatement1/**

```
[acadgild@localhost project2.2]$ hadoop fs -ls /user/acadgild/project/USAConsumer/ProblemStatement1/*
17/08/26 17:49:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
-rw-r--r--   1 acadgild supergroup          0 2017-08-26 17:49 /user/acadgild/project/USAConsumer/ProblemStatement1/_SUCCESS
-rw-r--r--   1 acadgild supergroup         11 2017-08-26 17:49 /user/acadgild/project/USAConsumer/ProblemStatement1/part-m-00000
-rw-r--r--   1 acadgild supergroup          9 2017-08-26 17:49 /user/acadgild/project/USAConsumer/ProblemStatement1/part-m-00001
[acadgild@localhost project2.2]$ hadoop fs -cat /user/acadgild/project/USAConsumer/ProblemStatement1/*
17/08/26 17:49:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Yes     455383
No      11490
[acadgild@localhost project2.2]$
```

PIG output have been stored successfully with the data separated by TAB (\t) which shows timelyResponse and Count of complaints.

PIG script is save with name ProblemStatement1.PIG which can be run using below command:

$ pig <filepath>/ProblemStatement1.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 TimelyResponse with column isTimelyResponse & complaintCount:

use db1;

show tables;

create table TimelyResponse

(

isTimelyResponse varchar(3),

complaintCount int(8)

);

describe TimelyResponse;

*select * from TimelyResponse;*

```
mysql> use db1;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+----------------------+
| Tables_in_db1        |
+----------------------+
| customer             |
| statewiseBPL80       |
| statewiseBPLacheived |
+----------------------+
3 rows in set (0.00 sec)

mysql> create table TimelyResponse
    -> (
    -> isTimelyResponse varchar(3),
    -> complaintCount int(8)
    -> );
Query OK, 0 rows affected (0.01 sec)

mysql> describe TimelyResponse;
+-----------------+------------+------+-----+---------+-------+
| Field           | Type       | Null | Key | Default | Extra |
+-----------------+------------+------+-----+---------+-------+
| isTimelyResponse | varchar(3) | YES  |     | NULL    |       |
| complaintCount   | int(8)     | YES  |     | NULL    |       |
+-----------------+------------+------+-----+---------+-------+
2 rows in set (0.00 sec)

mysql> select * from TimelyResponse;
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 'TimelyResponse' \

--export-dir '/user/acadgild/project/USAConsumer/ProblemStatement1/' \

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

-m 1

```
[acadgild@localhost project2.2]$ sqoop export --connect jdbc:mysql://localhost/db1 \
> --username 'root' -P --table 'TimelyResponse' \
> --export-dir '/user/acadgild/project/USAConsumer/ProblemStatement1/' \
> --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 18:01:35,110 INFO  [main] sqoop.Sqoop: Running Sqoop version: 1.4.5
Enter password:
2017-08-26 18:01:37,824 INFO  [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2017-08-26 18:01:37,825 INFO  [main] tool.CodeGenTool: Beginning code generation
2017-08-26 18:01:38,144 INFO  [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `TimelyResponse` AS t LIMIT 1
2017-08-26 18:01:38,174 INFO  [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `TimelyResponse` AS t LIMIT 1
```

```
2017-08-26 18:01:46,771 INFO  [main] mapreduce.JobSubmitter: Submitting tokens for job: job_1503739494588_0019
2017-08-26 18:01:47,007 INFO  [main] impl.YarnClientImpl: Submitted application application_1503739494588_0019 to ResourceManager at /0.0.0.0:8032
2017-08-26 18:01:47,029 INFO  [main] mapreduce.Job: The url to track the job: http://http://localhost:8088/proxy/application_1503739494588_0019/
2017-08-26 18:01:47,029 INFO  [main] mapreduce.Job: Running job: job_1503739494588_0019
2017-08-26 18:01:54,150 INFO  [main] mapreduce.Job: Job job_1503739494588_0019 running in uber mode : false
2017-08-26 18:01:54,151 INFO  [main] mapreduce.Job:  map 0% reduce 0%
2017-08-26 18:02:00,217 INFO  [main] mapreduce.Job:  map 100% reduce 0%
2017-08-26 18:02:00,227 INFO  [main] mapreduce.Job: Job job_1503739494588_0019 completed successfully
```

Sqoop command completed successfully.

Step 4: check table in mysql:

*select * from TimelyResponse;*

```
mysql> select * from TimelyResponse;
+-----------------+---------------+
| isTimelyResponse | complaintCount |
+-----------------+---------------+
| Yes              |         455383 |
| No               |          11490 |
+-----------------+---------------+
2 rows in set (0.00 sec)

mysql>
```

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/ProblemStatement1/ | wc -l*

```
[acadgild@localhost project2.2]$ hadoop fs -cat /user/acadgild/project/USAConsumer/ProblemStatement1/* | wc -l
17/08/26 18:06:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using
2
[acadgild@localhost project2.2]$
```

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

select count(*) from TimelyResponse;

```
mysql> select count(*) from TimelyResponse;
+----------+
| count(*) |
+----------+
|        2 |
+----------+
1 row in set (0.00 sec)

mysql>
```

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