**Readme File**

# Baseball Statistics using Apache Pig and Hive.

**Team Members :**

**Gaurav Dhingra : gxd150230**

**Paritosh Sundriyal : pxs157330**

**Sarthak Mehra : sxm147431**

**Sanket Prabhu : srp140430**

**Note :**

**All the Scripts of Pig and Hive along with the Datasets have been successfully uploaded on the Hadoop Directory /user/gxd150230/**

We have used Apache Pig and Hive for extracting the Data from the large Datasets and thereby running Queries on that dataset to find out the Top Batter, Top Pitcher and the Top Fielder for each of the year in the range.

Top Batter had been identified by keeping in mind different Attributes and corresponding to each attribute one batter is selected

Same has been the case with the Fielders and Pitchers.

Although the Dataset is from 1871 to 2011 but we filtered the data from 2005 onwards so as to have a glimpse of what we want to achieve.

We will be storing output of the Pig in each of the cases on HDFS which will be loaded later on in Hive and hence we can run Queries against the Data we loaded in Hive.

Full Datasets used are Batting.csv, Fielding.csv, Pitching.csv

Partial Datasets just for testing are Batting1.csv, Fielding1.csv,Pitching1.csv

All the datasets have been loaded on Hadoop.

We have created different Pig Scripts for the players :

**Batters :**

**batter_run.pig**

**batter_hit.pig**

**batter_atbat.pig**

**batter_onbase.pig**

**batter_average.pig**

**Fielders:**

**fielder_PO.pig**

**fielder_performance.pig**

**fielder_assist.pig**

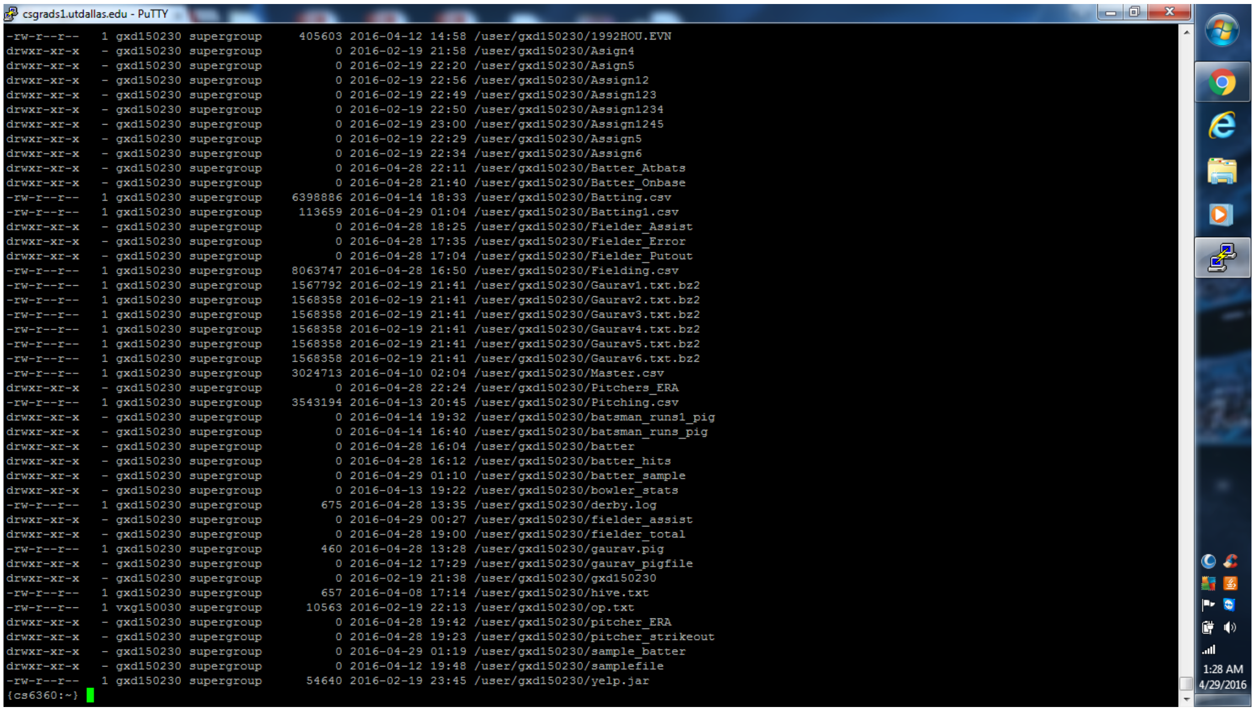
**Pitchers :**

**pitcher_strike.pig**

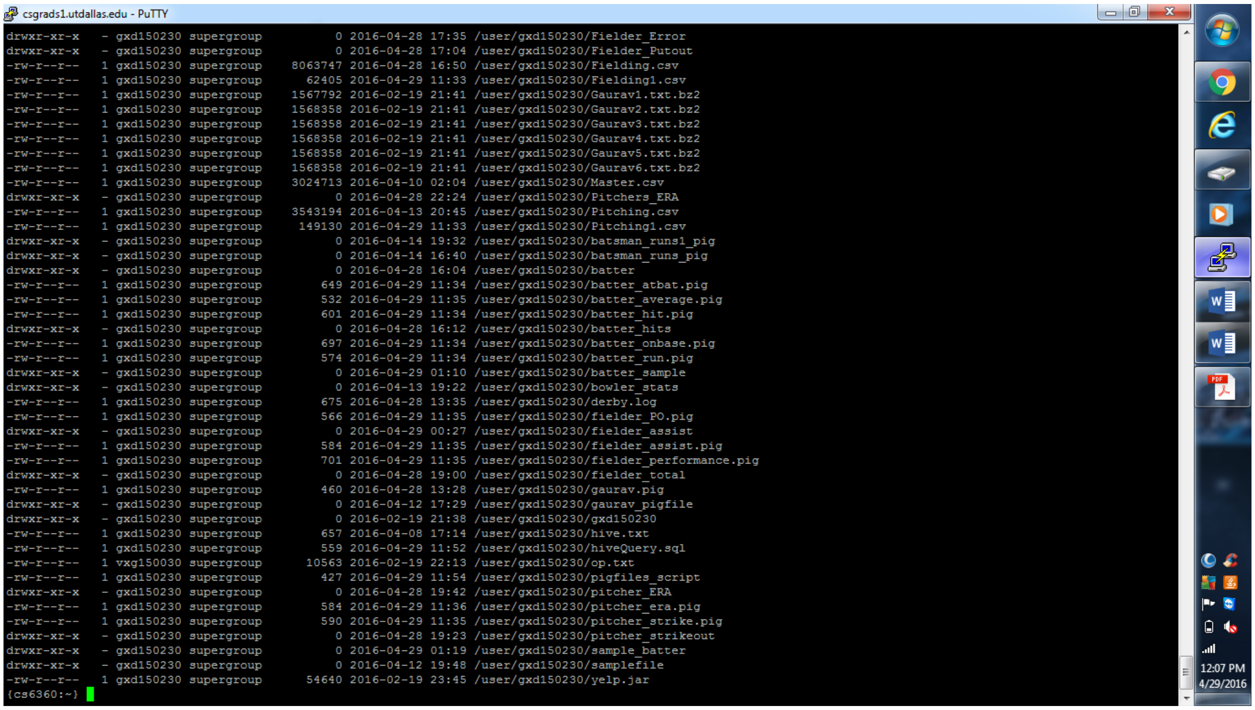
**pitcher_era.pig**

We ran all the scripts successfully and were able to store the output on Hadoop corresponding to each of the Scripts.

Please find the attached Screenshot :



```
-rw-r--r--   1 gxd150230 supergroup     405603 2016-04-12 14:58 /user/gxd150230/1992HOU.EVN
drwxr-xr-x   - gxd150230 supergroup          0 2016-02-19 21:58 /user/gxd150230/Asign4
drwxr-xr-x   - gxd150230 supergroup          0 2016-02-19 22:20 /user/gxd150230/Asign5
drwxr-xr-x   - gxd150230 supergroup          0 2016-02-19 22:56 /user/gxd150230/Assign12
drwxr-xr-x   - gxd150230 supergroup          0 2016-02-19 22:49 /user/gxd150230/Assign123
drwxr-xr-x   - gxd150230 supergroup          0 2016-02-19 22:50 /user/gxd150230/Assign1234
drwxr-xr-x   - gxd150230 supergroup          0 2016-02-19 23:00 /user/gxd150230/Assign1245
drwxr-xr-x   - gxd150230 supergroup          0 2016-02-19 22:29 /user/gxd150230/Assign5
drwxr-xr-x   - gxd150230 supergroup          0 2016-02-19 22:34 /user/gxd150230/Assign6
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 22:11 /user/gxd150230/Batter_Atbats
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 21:40 /user/gxd150230/Batter_Onbase
-rw-r--r--   1 gxd150230 supergroup    6398886 2016-04-14 18:33 /user/gxd150230/Batting.csv
-rw-r--r--   1 gxd150230 supergroup     113659 2016-04-29 01:04 /user/gxd150230/Batting1.csv
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 18:25 /user/gxd150230/Fielder_Assist
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 17:35 /user/gxd150230/Fielder_Error
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 17:04 /user/gxd150230/Fielder_Putout
-rw-r--r--   1 gxd150230 supergroup    8063747 2016-04-28 16:50 /user/gxd150230/Fielding.csv
-rw-r--r--   1 gxd150230 supergroup    1567792 2016-02-19 21:41 /user/gxd150230/Gaurav1.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    1568358 2016-02-19 21:41 /user/gxd150230/Gaurav2.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    1568358 2016-02-19 21:41 /user/gxd150230/Gaurav3.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    1568358 2016-02-19 21:41 /user/gxd150230/Gaurav4.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    1568358 2016-02-19 21:41 /user/gxd150230/Gaurav5.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    1568358 2016-02-19 21:41 /user/gxd150230/Gaurav6.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    3024713 2016-04-10 02:04 /user/gxd150230/Master.csv
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 22:24 /user/gxd150230/Pitchers_ERA
-rw-r--r--   1 gxd150230 supergroup    3543194 2016-04-13 20:45 /user/gxd150230/Pitching.csv
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-14 19:32 /user/gxd150230/batsman_runs1_pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-14 16:40 /user/gxd150230/batsman_runs_pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 16:04 /user/gxd150230/batter
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 16:12 /user/gxd150230/batter_hits
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-29 01:10 /user/gxd150230/batter_sample
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-13 19:22 /user/gxd150230/bowler_stats
-rw-r--r--   1 gxd150230 supergroup        675 2016-04-28 13:35 /user/gxd150230/derby.log
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-29 00:27 /user/gxd150230/fielder_assist
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 19:00 /user/gxd150230/fielder_total
-rw-r--r--   1 gxd150230 supergroup        460 2016-04-28 13:28 /user/gxd150230/gaurav.pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-12 17:29 /user/gxd150230/gaurav_pigfile
drwxr-xr-x   - gxd150230 supergroup          0 2016-02-19 21:38 /user/gxd150230/gxd150230
-rw-r--r--   1 gxd150230 supergroup        657 2016-04-08 17:14 /user/gxd150230/hive.txt
-rw-r--r--   1 vxg150030 supergroup      10563 2016-02-19 22:13 /user/gxd150230/op.txt
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 19:42 /user/gxd150230/pitcher_ERA
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 19:23 /user/gxd150230/pitcher_strikeout
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-29 01:19 /user/gxd150230/sample_batter
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-12 19:48 /user/gxd150230/samplefile
-rw-r--r--   1 gxd150230 supergroup      54640 2016-02-19 23:45 /user/gxd150230/yelp.jar
{cs6360:~}
```



```
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 17:35 /user/gxd150230/Fielder_Error
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 17:04 /user/gxd150230/Fielder_Putout
-rw-r--r--   1 gxd150230 supergroup    8063747 2016-04-28 16:50 /user/gxd150230/Fielding.csv
-rw-r--r--   1 gxd150230 supergroup      62405 2016-04-29 11:33 /user/gxd150230/Fielding1.csv
-rw-r--r--   1 gxd150230 supergroup    1567792 2016-02-19 21:41 /user/gxd150230/Gaurav1.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    1568358 2016-02-19 21:41 /user/gxd150230/Gaurav2.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    1568358 2016-02-19 21:41 /user/gxd150230/Gaurav3.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    1568358 2016-02-19 21:41 /user/gxd150230/Gaurav4.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    1568358 2016-02-19 21:41 /user/gxd150230/Gaurav5.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    1568358 2016-02-19 21:41 /user/gxd150230/Gaurav6.txt.bz2
-rw-r--r--   1 gxd150230 supergroup    3024713 2016-04-10 02:04 /user/gxd150230/Master.csv
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 22:24 /user/gxd150230/Pitchers_ERA
-rw-r--r--   1 gxd150230 supergroup    3543194 2016-04-13 20:45 /user/gxd150230/Pitching.csv
-rw-r--r--   1 gxd150230 supergroup     149130 2016-04-29 11:33 /user/gxd150230/Pitching1.csv
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-14 19:32 /user/gxd150230/batsman_runs1_pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-14 16:40 /user/gxd150230/batsman_runs_pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 16:04 /user/gxd150230/batter
-rw-r--r--   1 gxd150230 supergroup        649 2016-04-29 11:34 /user/gxd150230/batter_atbat.pig
-rw-r--r--   1 gxd150230 supergroup        532 2016-04-29 11:35 /user/gxd150230/batter_average.pig
-rw-r--r--   1 gxd150230 supergroup        601 2016-04-29 11:34 /user/gxd150230/batter_hit.pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 16:12 /user/gxd150230/batter_hits
-rw-r--r--   1 gxd150230 supergroup        697 2016-04-29 11:34 /user/gxd150230/batter_onbase.pig
-rw-r--r--   1 gxd150230 supergroup        574 2016-04-29 11:34 /user/gxd150230/batter_run.pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-29 01:10 /user/gxd150230/batter_sample
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-13 19:22 /user/gxd150230/bowler_stats
-rw-r--r--   1 gxd150230 supergroup        675 2016-04-28 13:35 /user/gxd150230/derby.log
-rw-r--r--   1 gxd150230 supergroup        566 2016-04-29 11:35 /user/gxd150230/fielder_PO.pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-29 00:27 /user/gxd150230/fielder_assist
-rw-r--r--   1 gxd150230 supergroup        584 2016-04-29 11:35 /user/gxd150230/fielder_assist.pig
-rw-r--r--   1 gxd150230 supergroup        701 2016-04-29 11:35 /user/gxd150230/fielder_performance.pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 19:00 /user/gxd150230/fielder_total
-rw-r--r--   1 gxd150230 supergroup        460 2016-04-28 13:28 /user/gxd150230/gaurav.pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-12 17:29 /user/gxd150230/gaurav_pigfile
drwxr-xr-x   - gxd150230 supergroup          0 2016-02-19 21:38 /user/gxd150230/gxd150230
-rw-r--r--   1 gxd150230 supergroup        657 2016-04-08 17:14 /user/gxd150230/hive.txt
-rw-r--r--   1 gxd150230 supergroup        559 2016-04-29 11:52 /user/gxd150230/hiveQuery.sql
-rw-r--r--   1 vxg150030 supergroup      10563 2016-02-19 22:13 /user/gxd150230/op.txt
-rw-r--r--   1 gxd150230 supergroup        427 2016-04-29 11:54 /user/gxd150230/pigfiles_script
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 19:42 /user/gxd150230/pitcher_ERA
-rw-r--r--   1 gxd150230 supergroup        584 2016-04-29 11:36 /user/gxd150230/pitcher_era.pig
-rw-r--r--   1 gxd150230 supergroup        590 2016-04-29 11:35 /user/gxd150230/pitcher_strike.pig
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-28 19:23 /user/gxd150230/pitcher_strikeout
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-29 01:19 /user/gxd150230/sample_batter
drwxr-xr-x   - gxd150230 supergroup          0 2016-04-12 19:48 /user/gxd150230/samplefile
-rw-r--r--   1 gxd150230 supergroup      54640 2016-02-19 23:45 /user/gxd150230/yelp.jar
{cs6360:~}
```

We stored all the Pig Commands in one script file named under pigfiles_script

Permissions have been successfully changed with the command:

```
Chmod 755 pigfiles_script
```

**Once we have all the Data on Hadoop we will be loading the data in the Hive and then use hiveQL to run Queries against it.**

**One sample Query is :**

**File is located in the directory /user/gxd150230/Batter_Atbats/**

```
CREATE EXTERNAL TABLE batterAtbat(year string,id string,atbats int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' LOCATION '/user/gxd150230/Batter_Atbats/';
```

Sample Query be like :

```
select * from batterAtbat where id LIKE '%suzukic01%';
```

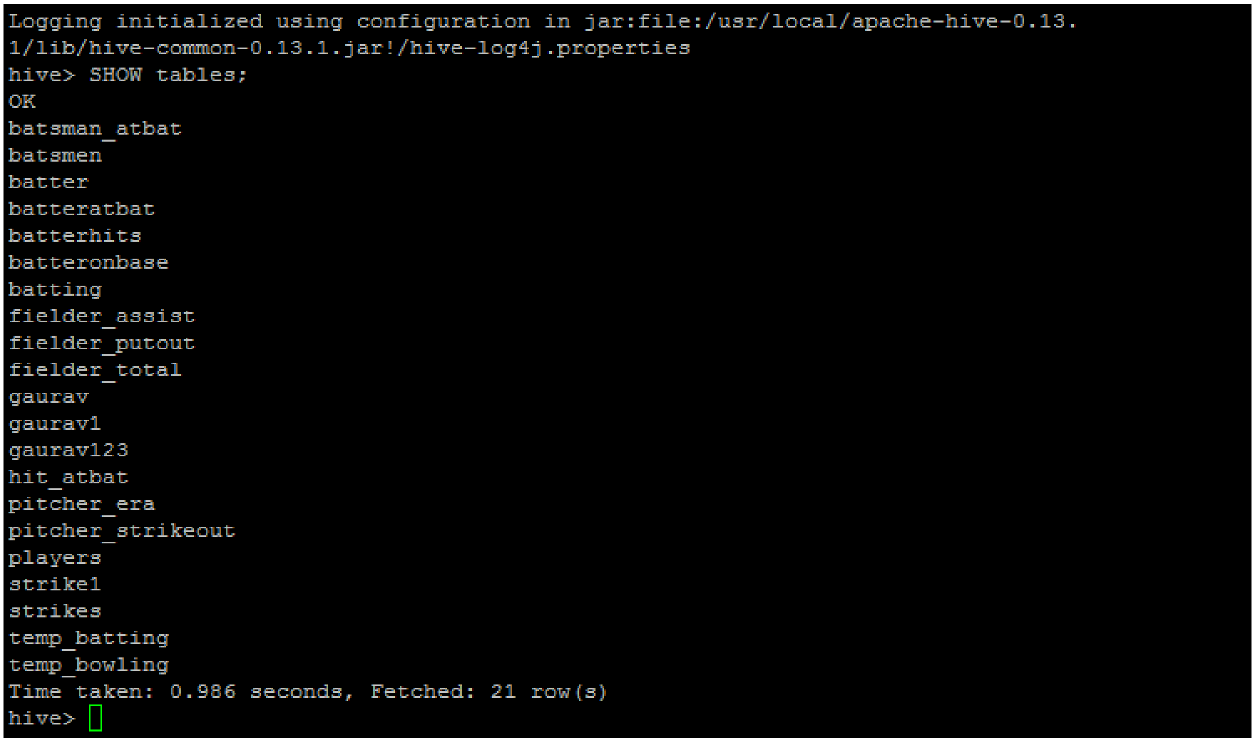
**Output :**

```
2010    suzukic01    680
2011    suzukic01    677
```

All the rest of the Queries are there in the Script.txt which I have attached as part of the source code.

Corresponding to each batter,pitcher and Fielder we have created one Table in Hive.

Please find the below screenshot :



```
Logging initialized using configuration in jar:file:/usr/local/apache-hive-0.13.
1/lib/hive-common-0.13.1.jar!/hive-log4j.properties
hive> SHOW tables;
OK
batsman_atbat
batsmen
batter
batteratbat
batterhits
batteronbase
batting
fielder_assist
fielder_putout
fielder_total
gaurav
gaurav1
gaurav123
hit_atbat
pitcher_era
pitcher_strikeout
players
strike1
strikes
temp_batting
temp_bowling
Time taken: 0.986 seconds, Fetched: 21 row(s)
hive>
```

We created a Hive Script using the following steps :

**vi hiveQuery.sql**

**Content of the Script File :**

**show tables;**

**select * from batterAtbat where id LIKE '%suzukic01%';**

**SELECT * FROM batterAtbat ORDER BY atbats DESC  LIMIT 1;**

**select * from batterOnbase where year LIKE '%2006%';**

**select * from batterhits where year LIKE '%2007%';**

**select * from batter where id LIKE '%pujola101%';**

**select * from pitcher_strikeout where year LIKE '%2008%';**

**SELECT * FROM pither_ERA ORDER BY ERA DESC  LIMIT 1;**

**SELECT * FROM fielder_assist ORDER BY assist DESC  LIMIT 1;**

**select * from fielder_putout where year LIKE '%2008%';**

**select * from fielder_total where year LIKE '%2008%';**

**Steps to run the Hive Script :**

```
hive -f /home/013/g/gx/gxd150230/hiveQuery.sql
```

**hiveQuery.sql has also been uploaded to HDFS**

It will generate the output corresponding to the above Queries for each of the Batters, Pitchers and the Fielders.