http://ybhavesh.blogspot.in/2015/12/poc-generate-analytics-from-product.html

/\* Develop Pig program to extract data for the following KPIs \*/

FILE1 = LOAD '/home/bhavesh/weblog.txt' using PigStorage(',') as (IP:chararray,Category1:chararray,Category2:chararray,page:int,statuscode:int);

GRPD = group FILE1 by IP;

IPCOUNT = foreach GRPD generate group,COUNT(FILE1.IP);

/\* Count of page views by individual user ie IP, count(\*) \*/

STORE IPCOUNT into '/home/bhavesh/Count\_by\_Pageview\_by\_IndividualUser/' using PigStorage(',');

/\*(Exclude status code other than 200) \*/

FILE\_STATUS\_CODE\_200 = FILTER FILE1 by statuscode == 200;

/\*

Top 2 and bottom 2 records

Category, total\_number\_views

page, total\_number\_views

IP, total\_number\_of\_views

\*/

GRPD\_CATEGORY1 = group FILE\_STATUS\_CODE\_200 by Category1;

CNT\_FOR\_CATEGORY1 = FOREACH GRPD\_CATEGORY1 generate group,COUNT(FILE\_STATUS\_CODE\_200.Category1) as COUNTING;

SORTED\_CATEGORY1\_DEC = ORDER CNT\_FOR\_CATEGORY1 by COUNTING DESC;

TOP\_2\_CATEGORY1 = limit SORTED\_CATEGORY1\_DEC 2;

STORE TOP\_2\_CATEGORY1 into '/home/bhavesh/Top2\_By\_Category1/' using PigStorage(',');

SORTED\_CATEGORY1\_ASC = ORDER CNT\_FOR\_CATEGORY1 by COUNTING ASC;

BOTTOM\_2\_CATEGORY1 = limit SORTED\_CATEGORY1\_ASC 2;

STORE BOTTOM\_2\_CATEGORY1 into '/home/bhavesh/Bottom2\_By\_Category1/' using PigStorage(',');

GRPD\_CATEGORY2 = group FILE\_STATUS\_CODE\_200 by Category2;

CNT\_FOR\_CATEGORY2 = FOREACH GRPD\_CATEGORY2 generate group,COUNT(FILE\_STATUS\_CODE\_200.Category2) as COUNTING;

SORTED\_CATEGORY2\_DEC = ORDER CNT\_FOR\_CATEGORY2 by COUNTING DESC;

TOP\_2\_CATEGORY2 = limit SORTED\_CATEGORY2\_DEC 2;

STORE TOP\_2\_CATEGORY2 into '/home/bhavesh/Top2\_By\_Category2/' using PigStorage(',');

SORTED\_CATEGORY2\_ASC = ORDER CNT\_FOR\_CATEGORY2 by COUNTING ASC;

BOTTOM\_2\_CATEGORY2 = limit SORTED\_CATEGORY2\_ASC 2;

STORE BOTTOM\_2\_CATEGORY2 into '/home/bhavesh/Bottom2\_By\_Category2/' using PigStorage(',');

GRPD\_PAGES = group FILE\_STATUS\_CODE\_200 by page;

CNT\_FOR\_PAGE = FOREACH GRPD\_PAGES generate group,COUNT(FILE\_STATUS\_CODE\_200.page) as COUNTING;

SORTED\_PAGE\_DEC = ORDER CNT\_FOR\_PAGE by COUNTING DESC;

TOP\_2\_PAGE = limit SORTED\_PAGE\_DEC 2;

STORE TOP\_2\_PAGE into '/home/bhavesh/Top2\_By\_PAGE/' using PigStorage(',');

SORTED\_PAGE\_ASC = ORDER CNT\_FOR\_PAGE by COUNTING ASC;

BOTTOM\_2\_PAGE = limit SORTED\_PAGE\_ASC 2;

STORE BOTTOM\_2\_PAGE into '/home/bhavesh/Bottom2\_By\_PAGE/' using PigStorage(',');

GRPD\_IP = group FILE\_STATUS\_CODE\_200 by IP;

CNT\_FOR\_IP = FOREACH GRPD\_IP generate group,COUNT(FILE\_STATUS\_CODE\_200.IP) as COUNTING;

SORTED\_IP\_DEC = ORDER CNT\_FOR\_IP by COUNTING DESC;

TOP\_2\_IP = limit SORTED\_IP\_DEC 2;

STORE TOP\_2\_IP into '/home/bhavesh/Top2\_By\_IP/' using PigStorage(',');

SORTED\_IP\_ASC = ORDER CNT\_FOR\_IP by COUNTING ASC;

BOTTOM\_2\_IP = limit SORTED\_IP\_ASC 2;

STORE BOTTOM\_2\_IP into '/home/bhavesh/Bottom2\_By\_IP/' using PigStorage(',');

/\* Total page views / Category wise pageviews / Unique pageviews

page,total\_number\_of\_views

category, total\_views

page, total\_number\_of\_unique\_views \*/

GRPD\_TOTALPAGES = group FILE1 by page;

CNT\_FOR\_TOTALPAGE = FOREACH GRPD\_TOTALPAGES generate group,COUNT(FILE1.page) as COUNTING;

SORTED\_PAGES\_DEC = ORDER CNT\_FOR\_TOTALPAGE by COUNTING DESC;

STORE SORTED\_PAGES\_DEC into '/home/bhavesh/Total\_PAGES\_Count/' using PigStorage(',');

GRPD\_TOTALCATEGORY1 = group FILE1 by Category1;

CNT\_FOR\_TOTALCATEGORY1 = FOREACH GRPD\_TOTALCATEGORY1 generate group,COUNT(FILE1.page) as COUNTING;

SORTED\_TOTALCATEGORY1\_DEC = ORDER CNT\_FOR\_TOTALCATEGORY1 by COUNTING DESC;

STORE SORTED\_TOTALCATEGORY1\_DEC into '/home/bhavesh/Total\_Category1\_Count/' using PigStorage(',');

GRPD\_TOTALCATEGORY2 = group FILE1 by Category2;

CNT\_FOR\_TOTALCATEGORY2 = FOREACH GRPD\_TOTALCATEGORY2 generate group,COUNT(FILE1.page) as COUNTING;

SORTED\_TOTALCATEGORY2\_DEC = ORDER CNT\_FOR\_TOTALCATEGORY2 by COUNTING DESC;

STORE SORTED\_TOTALCATEGORY2\_DEC into '/home/bhavesh/Total\_Category2\_Count/' using PigStorage(',');

GRPD\_TOTALPAGES\_UNIQUEVIEW = group FILE1 by page;

CNT\_FOR\_TOTALPAGE\_UNIQUEVIEW = FOREACH GRPD\_TOTALPAGES\_UNIQUEVIEW {

internet\_protocol = FILE1.IP;

unique\_internet\_protocol = DISTINCT internet\_protocol;

GENERATE group, COUNT(unique\_internet\_protocol);

};

STORE CNT\_FOR\_TOTALPAGE\_UNIQUEVIEW into '/home/bhavesh/Page\_Total\_Number\_Of\_Unique\_Views/' using PigStorage(',');

/\*

Count of status code = 200 / 404 / 400 / 500

status\_code, count

\*/

GRPD = group FILE1 by statuscode;

STATUS\_CODE\_COUNT = foreach GRPD generate group,COUNT(FILE1.statuscode);

STORE STATUS\_CODE\_COUNT into '/home/bhavesh/Status\_Code\_Count/' using PigStorage(',');

Shell Script

#####################################################################

############################# COMPLETE SCRIPT #######################

### HEADER - PROGRAM NAME - <weblog.sh> ###

### AUTHOR - BHAVESH BHADRICHA ###

### DATE - 11/DEC/2015 ###

### VERSION - 1.0 ###

### DESCRIPTION - Data: It comprises of the information gathered from websites ###

### which contains IP, Two Categories of Product, Pages and Status Code ###

### ###

### Problem Statement: Analyse the data in Hadoop Eco-system to: ###

### 1.Load data into HDFS using HDFS client ###

### ###

### 2. Develop PIG program to parse WEB logs and meaning full result from it ###

### INUPT file Format ###

### IP,Category-1,Category-2,page status\_code ###

### ###

### PIG program to extract data for the following ###

### ###

### 3. Count of page views by individual user ###

### IP, count(\*) ###

### ###

### 4. Top / Bottom 5: catagery-1/ catagery-2 / page /users ###

### (Exclude status code other than 200) ###

### ###

### Top 5 and bottom 5 records ###

### Category, total\_number\_views ###

### page, total\_number\_views ###

### IP, total\_number\_of\_views ###

### ###

### 5. Total page views / Category wise pageviews / Unique pageviews ###

### ###

### page,total\_number\_of\_views ###

### category, total\_views ###

### page, total\_number\_of\_unique\_views ###

### ###

### 6. Count of status code = 200 / 404 / 400 / 500 ###

### status\_code, count ###

### ###

### 7. Load results into tables in MySql Database using Sqoop.? ###

#####################################################################

#####################################################################

#####################################################################

###DEFINING THE LOCAL VARIABLES###

######################################################################

DATE=$(date +"%Y%m%d\_%H%M%S")

LOGFILE="/home/bhavesh/POC/WEBLOG\_POC/LOG/"$DATE".log"

#####################################################################

### Load data into HDFS using HDFS client ####################################

#####################################################################

hadoop fs -put weblog.txt /home/bhavesh/weblog.txt

########################## PIG Processing ###############################

#### PIG, which splits the data into two parts: Category data and Ratings data ###

#####################################################################

echo "Pig Script starts here"

echo "PIG Script,Weblog Processing" >> $LOGFILE

hadoop fs -rmr /home/bhavesh/Total\_Category1\_Count

hadoop fs -rmr /home/bhavesh/Bottom2\_By\_PAGE

hadoop fs -rmr /home/bhavesh/Top2\_By\_Category1

hadoop fs -rmr /home/bhavesh/Top2\_By\_IP

hadoop fs -rmr /home/bhavesh/Total\_PAGES\_Count

hadoop fs -rmr /home/bhavesh/Bottom2\_By\_IP

hadoop fs -rmr /home/bhavesh/Bottom2\_By\_Category2

hadoop fs -rmr /home/bhavesh/Status\_Code\_Count

hadoop fs -rmr /home/bhavesh/Count\_by\_Pageview\_by\_IndividualUser

hadoop fs -rmr /home/bhavesh/Page\_Total\_Number\_Of\_Unique\_Views

hadoop fs -rmr /home/bhavesh/Total\_Category2\_Count

hadoop fs -rmr /home/bhavesh/Bottom2\_By\_Category1

hadoop fs -rmr /home/bhavesh/Top2\_By\_Category2

hadoop fs -rmr /home/bhavesh/Top2\_By\_PAGE

pig /home/bhavesh/POC/WEBLOG\_POC/weblog.pig

if [ $? -eq 0 ]; then

echo "Succesfully finished PIG Processing " >> $LOGFILE

else

echo "PIG Processing Failed Please check the Log " >> $LOGFILE

fi

############################ HIVE Processing ###############################

###### HIVE will load the Category data and Rating Data into Hive Tables ##########

########################################################################

echo "HIVE Script starts here"

echo "HIVE LOAD data into Table " >> $LOGFILE

hive -e 'drop table if exists TotalCategory1Count';

hive -e 'drop table if exists Bottom2ByPAGE';

hive -e 'drop table if exists Top2ByCategory1';

hive -e 'drop table if exists Top2ByIP';

hive -e 'drop table if exists TotalPAGESCount';

hive -e 'drop table if exists Bottom2ByIP';

hive -e 'drop table if exists Bottom2ByCategory2';

hive -e 'drop table if exists StatusCodeCount';

hive -e 'drop table if exists CountbyPageviewbyIndividualUser';

hive -e 'drop table if exists PageTotalNumberOfUniqueViews';

hive -e 'drop table if exists TotalCategory2Count';

hive -e 'drop table if exists Bottom2ByCategory1';

hive -e 'drop table if exists Top2ByCategory2';

hive -e 'drop table if exists Top2ByPAGE';

hive -e "create external table TotalCategory1Count

(Category1 string,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/TotalCategory1Count'";

hive -e "create external table Bottom2ByPAGE

(Pages int,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/Bottom2ByPAGE'";

hive -e "create external table Top2ByCategory1

(Category1 string,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/Top2ByCategory1'";

hive -e "create external table Top2ByIP

(IP string,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/Top2ByIP'";

hive -e "create external table TotalPAGESCount

(PAGES int,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/TotalPAGESCount'";

hive -e "create external table Bottom2ByIP

(IP string,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/Bottom2ByIP'";

hive -e "create external table Bottom2ByCategory2

(Category2 string,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/Bottom2ByCategory2'";

hive -e "create external table StatusCodeCount

(StatusCode int,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/StatusCodeCount'";

hive -e "create external table CountbyPageviewbyIndividualUser

(IP string,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/CountbyPageviewbyIndividualUser'";

hive -e "create external table PageTotalNumberOfUniqueViews

(page int,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/PageTotalNumberOfUniqueViews'";

hive -e "create external table TotalCategory2Count

(Category2 string,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/TotalCategory2Count'";

hive -e "create external table Bottom2ByCategory1

(Category1 string,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/Bottom2ByCategory1'";

hive -e "create external table Top2ByCategory2

(Category2 string,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/Top2ByCategory2'";

hive -e "create external table Top2ByPAGE

(page int,

countings int)

row format delimited

fields terminated by','

lines terminated by '\n'

stored as textfile location '/home/bhavesh/hive/Top2ByPAGE'";

hive -e "load data inpath '/home/bhavesh/Total\_Category1\_Count/part-r-00000' overwrite into table TotalCategory1Count";

hive -e "load data inpath '/home/bhavesh/Bottom2\_By\_PAGE/part-r-00000' overwrite into table Bottom2ByPAGE";

hive -e "load data inpath '/home/bhavesh/Top2\_By\_Category1/part-r-00000' overwrite into table Top2ByCategory1";

hive -e "load data inpath '/home/bhavesh/Top2\_By\_IP/part-r-00000' overwrite into table Top2ByIP";

hive -e "load data inpath '/home/bhavesh/Total\_PAGES\_Count/part-r-00000' overwrite into table TotalPAGESCount";

hive -e "load data inpath '/home/bhavesh/Bottom2\_By\_IP/part-r-00000' overwrite into table Bottom2ByIP";

hive -e "load data inpath '/home/bhavesh/Bottom2\_By\_Category2/part-r-00000' overwrite into table Bottom2ByCategory2";

hive -e "load data inpath '/home/bhavesh/Status\_Code\_Count/part-r-00000' overwrite into table StatusCodeCount";

hive -e "load data inpath '/home/bhavesh/Count\_by\_Pageview\_by\_IndividualUser/part-r-00000' overwrite into table CountbyPageviewbyIndividualUser";

hive -e "load data inpath '/home/bhavesh/Page\_Total\_Number\_Of\_Unique\_Views/part-r-00000' overwrite into table PageTotalNumberOfUniqueViews";

hive -e "load data inpath '/home/bhavesh/Total\_Category2\_Count/part-r-00000' overwrite into table TotalCategory2Count";

hive -e "load data inpath '/home/bhavesh/Bottom2\_By\_Category1/part-r-00000' overwrite into table Bottom2ByCategory1";

hive -e "load data inpath '/home/bhavesh/Top2\_By\_Category2/part-r-00000' overwrite into table Top2ByCategory2";

hive -e "load data inpath '/home/bhavesh/Top2\_By\_PAGE/part-r-00000' overwrite into table Top2ByPAGE";

############################ SQOOP Processing ##############################

###### Pushing the HIVE Tale data into RDBMS Tables via SQOOP #######################

#########################################################################

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table TotalCategory1Count --export-dir /home/bhavesh/hive/TotalCategory1Count/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table Bottom2ByPAGE --export-dir /home/bhavesh/hive/Bottom2ByPAGE/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table Top2ByCategory1 --export-dir /home/bhavesh/hive/Top2ByCategory1/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table Top2ByIP --export-dir /home/bhavesh/hive/Top2ByIP/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table TotalPAGESCount --export-dir /home/bhavesh/hive/TotalPAGESCount/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table Bottom2ByIP --export-dir /home/bhavesh/hive/Bottom2ByIP/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table Bottom2ByCategory2 --export-dir /home/bhavesh/hive/Bottom2ByCategory2/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table StatusCodeCount --export-dir /home/bhavesh/hive/StatusCodeCount/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table CountbyPageviewbyIndividualUser --export-dir /home/bhavesh/hive/CountbyPageviewbyIndividualUser/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table PageTotalNumberOfUniqueViews --export-dir /home/bhavesh/hive/PageTotalNumberOfUniqueViews/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table TotalCategory2Count --export-dir /home/bhavesh/hive/TotalCategory2Count/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table Bottom2ByCategory1 --export-dir /home/bhavesh/hive/Bottom2ByCategory1/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table Top2ByCategory2 --export-dir /home/bhavesh/hive/Top2ByCategory2/part-r-00000 --input-fields-terminated-by ',';

sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table Top2ByPAGE --export-dir /home/bhavesh/hive/Top2ByPAGE/part-r-00000 --input-fields-terminated-by ',';