POC #: Generate Analytics from a Product based Company Web Log.

by Bhavesh *on* 19:21

POC #: Generate Analytics from a Product based Company Web Log.

The POC is based on Analysing a **Product based Company Web Log.**

Industry: E-Commerce

1. Load weblog data into HDFS using HDFS client
2. Develop Pig program to load log and perform analytics on IP Category-1 Category-2 page, status_code
   - 2.1. Count of page views by individual user ie [IP, count(*)]
   - 2.2. Top / Bottom 2: catagery-1/ catagery-2 / page /users (Exclude status code other than 200)

     Top 2 and bottom 2 records
     - Category, total_number_views
     - page, total_number_views
     - IP, total_number_of_views
   - 2.3. Total page views / Category wise pageviews / Unique pageviews
     - page,total_number_of_views
     - category, total_views
     - page, total_number_of_unique_views
   - 2.4. Count of status code = 200 / 404 / 400 / 500
     - status_code, count
3. Load results into tables in MySql Database using Sqoop.

CODING DETAILS

Pig Script

weblog.pig

```
/* 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 ',';

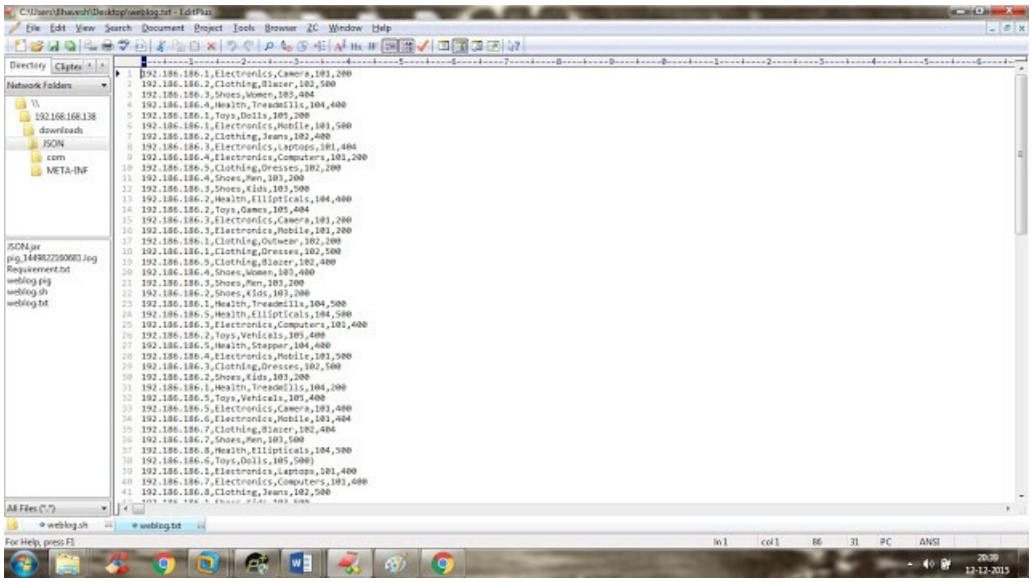
sqoop export --connect jdbc:mysql://localhost/mysql --username root --password root --table TotalCategory2Count --export-dir /home/bhavesh/hive/TotalCategory2Count/part-r-00000 --input-fields-terminated-by ',';

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 ',';

Input Data



Script Execution

```
root@ubuntu:/home/bhavesh/POC/WEBLOG_POC# ./weblog.sh
Pig Script starts here
rmr: cannot remove /home/bhavesh/Total_Category1_Count: No such file or directory.
rmr: cannot remove /home/bhavesh/Bottom2_By_PAGE: No such file or directory.
rmr: cannot remove /home/bhavesh/Top2_By_Category1: No such file or directory.
rmr: cannot remove /home/bhavesh/Top2_By_IP: No such file or directory.
rmr: cannot remove /home/bhavesh/Total_PAGES_Count: No such file or directory.
rmr: cannot remove /home/bhavesh/Bottom2_By_IP: No such file or directory.
rmr: cannot remove /home/bhavesh/Bottom2_By_Category2: No such file or directory.
Deleted hdfs://localhost:8020/home/bhavesh/Status_Code_Count
Deleted hdfs://localhost:8020/home/bhavesh/Count_by_Pageview_by_IndividualUser
Deleted hdfs://localhost:8020/home/bhavesh/Page_Total_Number_Of_Unique_Views
Deleted hdfs://localhost:8020/home/bhavesh/Total_Category2_Count
Deleted hdfs://localhost:8020/home/bhavesh/Bottom2_By_Category1
```

```
Input(s):
Successfully read 86 records (3693 bytes) from: "/home/bhavesh/weblog.txt"

Output(s):
Successfully stored 5 records (35 bytes) in: "/home/bhavesh/Total_PAGES_Count"
Successfully stored 2 records (32 bytes) in: "/home/bhavesh/Bottom2_By_IP"
Successfully stored 2 records (12 bytes) in: "/home/bhavesh/Top2_By_PAGE"
Successfully stored 2 records (20 bytes) in: "/home/bhavesh/Bottom2_By_Category1"
Successfully stored 2 records (21 bytes) in: "/home/bhavesh/Top2_By_Category1"
Successfully stored 17 records (162 bytes) in: "/home/bhavesh/Total_Category2_Count"
Successfully stored 5 records (30 bytes) in: "/home/bhavesh/Status_Code_Count"
Successfully stored 5 records (29 bytes) in: "/home/bhavesh/Page_Total_Number_Of_Unique_Views"
Successfully stored 8 records (132 bytes) in: "/home/bhavesh/Count_by_Pageview_by_IndividualUser"
Successfully stored 2 records (32 bytes) in: "/home/bhavesh/Top2_By_IP"
Successfully stored 5 records (54 bytes) in: "/home/bhavesh/Total_Category1_Count"
Successfully stored 2 records (12 bytes) in: "/home/bhavesh/Bottom2_By_PAGE"
Successfully stored 2 records (22 bytes) in: "/home/bhavesh/Bottom2_By_Category2"
Successfully stored 2 records (15 bytes) in: "/home/bhavesh/Top2_By_Category2"
```

```
2015-12-12 00:47:39,449 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
HIVE Script starts here
Hive history file=/tmp/root/hive_job_log_root_201512120047_177261252.txt
OK
Time taken: 14.124 seconds
Hive history file=/tmp/root/hive_job_log_root_201512120048_694489129.txt
OK
Time taken: 8.908 seconds
Hive history file=/tmp/root/hive_job_log_root_201512120048_1401957871.txt
OK
Time taken: 9.651 seconds
Hive history file=/tmp/root/hive_job_log_root_201512120048_949202696.txt
OK
Time taken: 9.0 seconds
Hive history file=/tmp/root/hive_job_log_root_201512120048_1836695866.txt
OK
```

```
15/12/12 06:26:12 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `TotalCategory2Count` AS t LIMIT 1
15/12/12 06:26:12 INFO orm.CompilationManager: HADOOP_HOME is /usr/lib/hadoop
15/12/12 06:26:12 INFO orm.CompilationManager: Found hadoop core jar at: /usr/lib/hadoop/hadoop-0.20.2-cdh3u5-core.jar
15/12/12 06:26:15 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-root/compile/756e2f76c551046ecbaeaf7d8cfc0e29/TotalCategory2Coun
15/12/12 06:26:15 INFO mapreduce.ExportJobBase: Beginning export of TotalCategory2Count
15/12/12 06:26:17 INFO input.FileInputFormat: Total input paths to process : 1
15/12/12 06:26:17 INFO input.FileInputFormat: Total input paths to process : 1
15/12/12 06:26:17 INFO mapred.JobClient: Running job: job_201512120558_0034
15/12/12 06:26:18 INFO mapred.JobClient:  map 0% reduce 0%
15/12/12 06:26:26 INFO mapred.JobClient:  map 100% reduce 0%
15/12/12 06:26:28 INFO mapred.JobClient: Job complete: job_201512120558_0034
15/12/12 06:26:29 INFO mapred.JobClient: Counters: 16
15/12/12 06:26:29 INFO mapred.JobClient:   Job Counters
15/12/12 06:26:29 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=9375
15/12/12 06:26:29 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
15/12/12 06:26:29 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
15/12/12 06:26:29 INFO mapred.JobClient:     Launched map tasks=1
15/12/12 06:26:29 INFO mapred.JobClient:     Data-local map tasks=1
15/12/12 06:26:29 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=0
15/12/12 06:26:29 INFO mapred.JobClient:   FileSystemCounters
15/12/12 06:26:29 INFO mapred.JobClient:     HDFS_BYTES_READ=311
15/12/12 06:26:29 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=70360
15/12/12 06:26:29 INFO mapred.JobClient:   Map-Reduce Framework
15/12/12 06:26:29 INFO mapred.JobClient:     Map input records=17
15/12/12 06:26:29 INFO mapred.JobClient:     Physical memory (bytes) snapshot=134377472
15/12/12 06:26:29 INFO mapred.JobClient:     Spilled Records=0
15/12/12 06:26:29 INFO mapred.JobClient:     CPU time spent (ms)=1940
15/12/12 06:26:29 INFO mapred.JobClient:     Total committed heap usage (bytes)=39911424
15/12/12 06:26:29 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=456220672
15/12/12 06:26:29 INFO mapred.JobClient:     Map output records=17
15/12/12 06:26:29 INFO mapred.JobClient:     SPLIT_RAW_BYTES=143
15/12/12 06:26:29 INFO mapreduce.ExportJobBase: Transferred 311 bytes in 13.0613 seconds (23.8107 bytes/sec)
```

Hive Output

```
hive> select * from TotalCategory1Count;
OK
Electronics     21
Shoes   18
Health  17
Clothing        17
Toys    13
Time taken: 7.382 seconds
hive> select * from Bottom2ByPAGE;
OK
104     3
102     4
Time taken: 0.121 seconds
hive> select * from Top2ByCategory1;
OK
Electronics     9
Toys    6
Time taken: 0.203 seconds
hive> select * from Top2ByIP;
OK
192.186.186.1   7
192.186.186.3   6
Time taken: 0.111 seconds
```

```
hive> select * from TotalPAGESCount;
OK
101     21
103     18
102     17
104     17
105     13
Time taken: 0.146 seconds
hive> select * from Bottom2ByIP;
OK
192.186.186.8   1
192.186.186.4   2
Time taken: 0.177 seconds
hive> select * from Bottom2ByCategory2;
OK
Ellipticals     1
Jeans   1
Time taken: 0.208 seconds
hive> select * from StatusCodeCount;
OK
200     27
400     21
404     11
500     24
NULL    0
Time taken: 0.334 seconds
```

```
hive> select * from CountbyPageviewbyIndividualUser;
OK
192.186.186.1   15
192.186.186.2   13
192.186.186.3   13
192.186.186.4   12
192.186.186.5   12
192.186.186.6   8
192.186.186.7   9
192.186.186.8   4
Time taken: 0.191 seconds
hive> select * from PageTotalNumberOfUniqueViews;
OK
101     8
102     8
103     8
104     8
105     5
Time taken: 0.319 seconds
```

```
hive> select * from TotalCategory2Count;
OK
Camera  8
Ellipticals     8
Men     7
Mobile  6
Kids    6
Games   5
Women   5
Blazer  5
Dolls   5
Dresses 5
Treadmills      5
Stepper 4
Laptops 4
Jeans   4
Outwear 3
Vehicals        3
Computers       3
Time taken: 0.26 seconds
hive> select * from Bottom2ByCategory1;
OK
Health  3
Clothing        4
Time taken: 0.122 seconds
```

```
hive> select * from Bottom2ByCategory1;
OK
Health  3
Clothing        4
Time taken: 0.122 seconds
hive> select * from Top2ByCategory2;
OK
Camera  5
Men     3
Time taken: 0.133 seconds
hive> select * from Top2ByPAGE;
OK
101     9
105     6
Time taken: 0.121 seconds
hive>
```

MySQL Output

```
mysql> select * from TotalCategory1Count;
+-------------+-----------+
| Category1   | countings |
+-------------+-----------+
| Electronics |        21 |
| Shoes       |        18 |
| Health      |        17 |
| Clothing    |        17 |
| Toys        |        13 |
+-------------+-----------+
5 rows in set (0.01 sec)

mysql> select * from Bottom2ByPAGE;
+-------+-----------+
| Pages | countings |
+-------+-----------+
|   104 |         3 |
|   102 |         4 |
+-------+-----------+
2 rows in set (0.00 sec)

mysql> select * from Top2ByCategory1;
+-------------+-----------+
| Category1   | countings |
+-------------+-----------+
| Electronics |         9 |
| Toys        |         6 |
+-------------+-----------+
```

```
mysql> select * from Top2ByIP;
+---------------+-----------+
| IP            | countings |
+---------------+-----------+
| 192.186.186.1 |         7 |
| 192.186.186.3 |         6 |
+---------------+-----------+
2 rows in set (0.00 sec)

mysql> select * from TotalPAGESCount;
+-------+-----------+
| PAGES | countings |
+-------+-----------+
|   101 |        21 |
|   103 |        18 |
|   102 |        17 |
|   104 |        17 |
|   105 |        13 |
+-------+-----------+
5 rows in set (0.00 sec)

mysql> select * from Bottom2ByIP;
+---------------+-----------+
| IP            | countings |
+---------------+-----------+
| 192.186.186.8 |         1 |
| 192.186.186.4 |         2 |
+---------------+-----------+
```

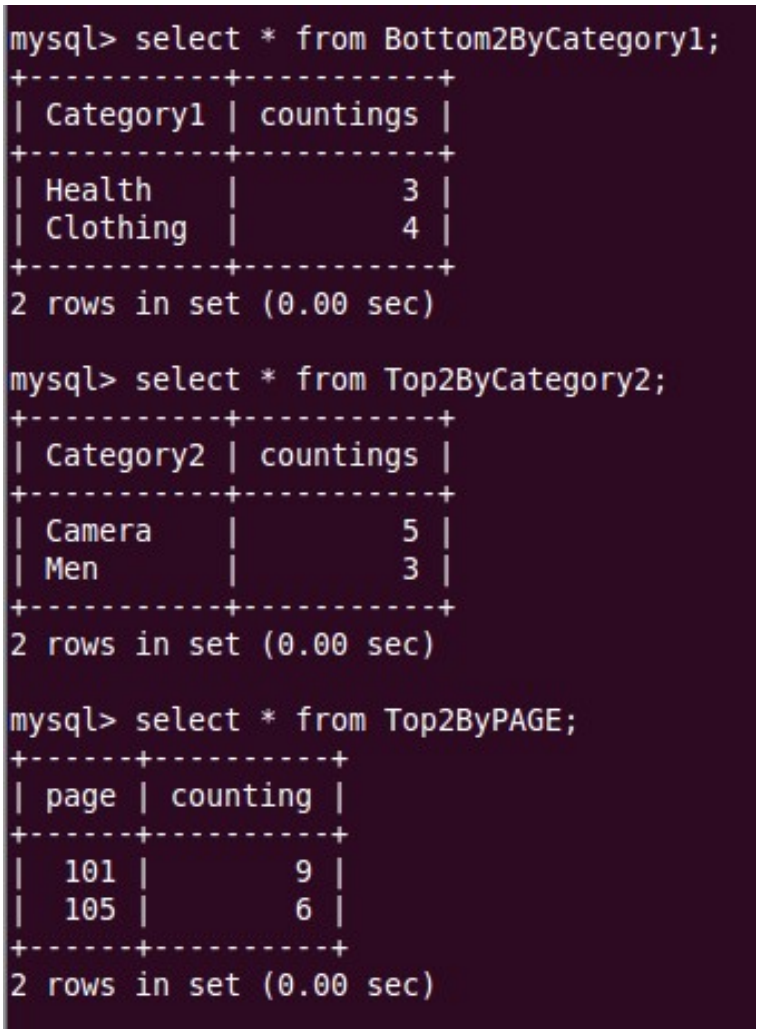
```
mysql> select * from Bottom2ByCategory2;
+-------------+-----------+
| Category2   | countings |
+-------------+-----------+
| Ellipticals |         1 |
| Jeans       |         1 |
+-------------+-----------+
2 rows in set (0.00 sec)

mysql> select * from StatusCodeCount;
+------------+-----------+
| StatusCode | countings |
+------------+-----------+
|        200 |        27 |
|        400 |        21 |
|        404 |        11 |
|        500 |        24 |
|       NULL |         0 |
+------------+-----------+
5 rows in set (0.00 sec)
```

```
mysql> select * from CountbyPageviewbyIndividualUser;
+---------------+-----------+
| IP            | countings |
+---------------+-----------+
| 192.186.186.1 |        15 |
| 192.186.186.2 |        13 |
| 192.186.186.3 |        13 |
| 192.186.186.4 |        12 |
| 192.186.186.5 |        12 |
| 192.186.186.6 |         8 |
| 192.186.186.7 |         9 |
| 192.186.186.8 |         4 |
+---------------+-----------+
8 rows in set (0.00 sec)

mysql> select * from PageTotalNumberOfUniqueViews;
+------+-----------+
| page | countings |
+------+-----------+
|  101 |         8 |
|  102 |         8 |
|  103 |         8 |
|  104 |         8 |
|  105 |         5 |
+------+-----------+
5 rows in set (0.00 sec)
```

```
mysql> select * from TotalCategory2Count;
+-------------+-----------+
| Category2   | countings |
+-------------+-----------+
| Camera      |         8 |
| Ellipticals |         8 |
| Men         |         7 |
| Mobile      |         6 |
| Kids        |         6 |
| Games       |         5 |
| Women       |         5 |
| Blazer      |         5 |
| Dolls       |         5 |
| Dresses     |         5 |
| Treadmills  |         5 |
| Stepper     |         4 |
| Laptops     |         4 |
| Jeans       |         4 |
| Outwear     |         3 |
| Vehicals    |         3 |
| Computers   |         3 |
+-------------+-----------+
17 rows in set (0.00 sec)
```



```
mysql> select * from Bottom2ByCategory1;
+-----------+-----------+
| Category1 | countings |
+-----------+-----------+
| Health    |         3 |
| Clothing  |         4 |
+-----------+-----------+
2 rows in set (0.00 sec)

mysql> select * from Top2ByCategory2;
+-----------+-----------+
| Category2 | countings |
+-----------+-----------+
| Camera    |         5 |
| Men       |         3 |
+-----------+-----------+
2 rows in set (0.00 sec)

mysql> select * from Top2ByPAGE;
+------+----------+
| page | counting |
+------+----------+
|  101 |        9 |
|  105 |        6 |
+------+----------+
2 rows in set (0.00 sec)
```

Tags

Author Image

About Bhavesh

Sora Blogging Tips is a blogger resources site is a provider of high quality blogger template with premium looking layout and robust design. The main mission of sora blogging tips is to provide the best quality blogger templates.

Newer Article

POC #: Sensex Log Data Processing (PDF File Processing In Map Reduce)

Older Article

POC #: Analyse Social Bookmarking Sites To Find Insights

RELATED POSTS:

6 comments:

subhadeep110

12 January 2017 at 18:27

hi,

can you please tell me how you made the input data from a web server log. in hive or pig.

Senthil Prabu

Reply

3 February 2017 at 05:43

Hi Bhavesh,

Nice Explanations with codes, please share input files, it wud be great!!!!

Reply

Unknown

13 February 2017 at 09:02

Can u please share input files.
Email: prashant2892@hotmail.com

Thank You

Reply

sushma

22 April 2017 at 15:42

Hi
Can you please share input file.

Reply

harshit

25 October 2017 at 11:08

Hi ,
nice explanation can you please mail me input file
harshitkacker@gmail.com

Reply

gopinath678

17 May 2018 at 08:47

Hi i want to try could you please refer me the input data set

Reply

RECENT POST

CONTACT FORM

Name

Email *

Message *

Send

RECENT IN INTERNET

POPULAR

POC #:
Proces

How to