**HIVE JSON DATA FOR USA GOVT**

**Case Study:**

**Query1:**

**The top 10 most popular sites in terms of clicks.**

**Query2:**

**The top-10 most popular sites for each country**

**Query 3:**

**The Top 10 most popular sites for each month.**

**Step 1:**

**hive>add jar /usr/local/hadoop/hive/lib/hive-serdes-1.0-SNAPSHOT.jar;**

**Step 2: Create hive table based on the data format**

**hive> CREATE TABLE clickstr (a string, c string, nk int, tz string, gr string, g string,h string,l string,hh string,r string,u string,t int,hc int,cy string, al string) ROW FORMAT serde 'com.cloudera.hive.serde.JSONSerDe';**

**OK**

**Time taken: 0.206 seconds**

**Step 3: Load data into the table**

**hive> load data local inpath '/home/training/Desktop/pigOperator/UsaGovData.txt' OVERWRITE INTO Table clickstr;**

**Copying data from file:/home/training/Desktop/pigOperator/UsaGovData.txt**

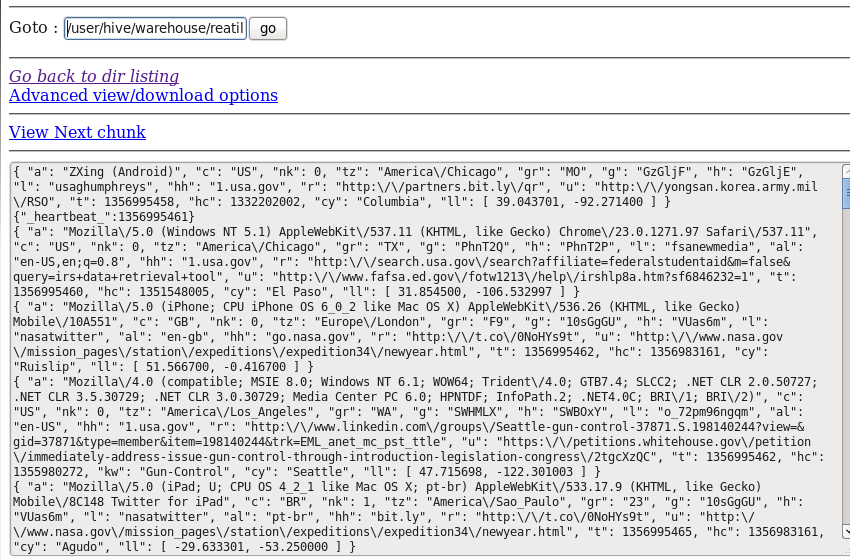
**Copying file: file:/home/training/Desktop/pigOperator/UsaGovData.txt**

**Loading data to table reatil.clickstr**

**Table reatil.clickstr stats: [num\_partitions: 0, num\_files: 1, num\_rows: 0, total\_size: 832182, raw\_data\_size: 0]**

**OK**

**Time taken: 0.245 seconds**

****

**Step 4: For query tuning we are partioning the above table and loading the data into it.**

**hive> CREATE EXTERNAL TABLE IF NOT EXISTS clickstrpart (a string, nk int,g string, h string,l string, hh string,r string,u string,t bigint,gr string,cy string,tz string,hc bigint,al string) PARTITIONED BY(c string) row format delimited fields terminated by '\t';**

**OK**

**Time taken: 0.306 seconds**

**Step 5: Insert data into Partitioned table, by using select clause:**

**Using Data partition**

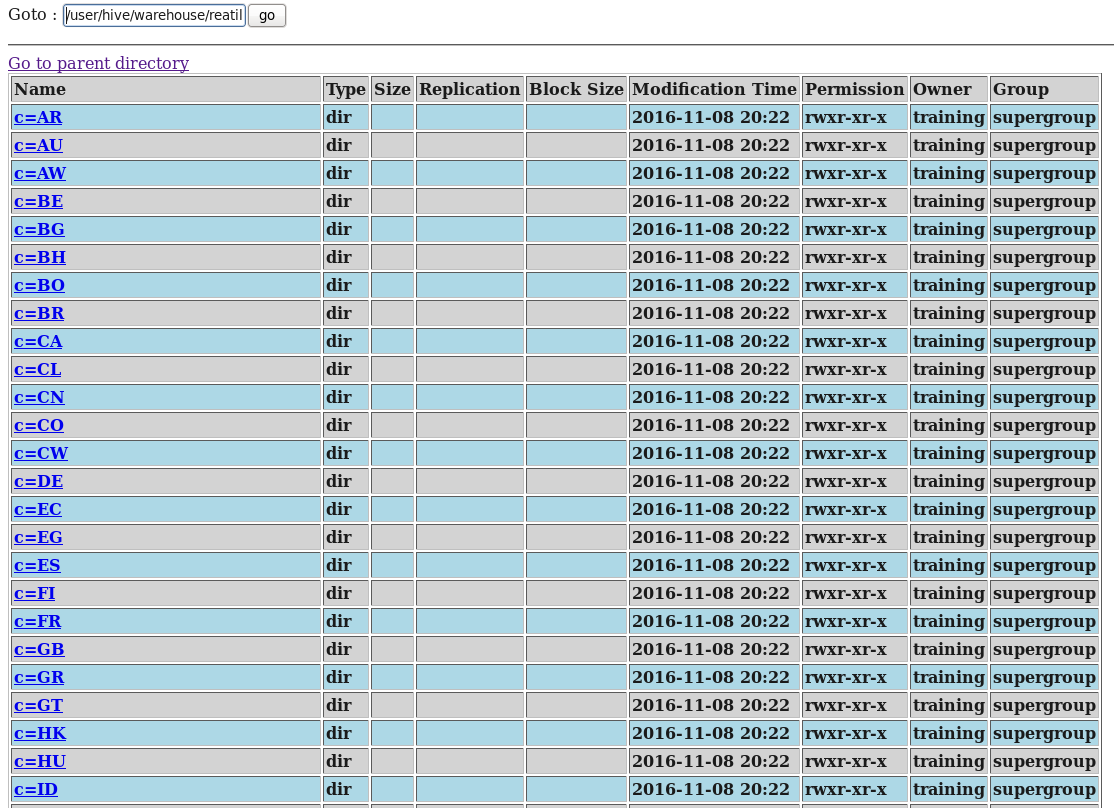
**hive> set hive.exec.dynamic.partition=true;**

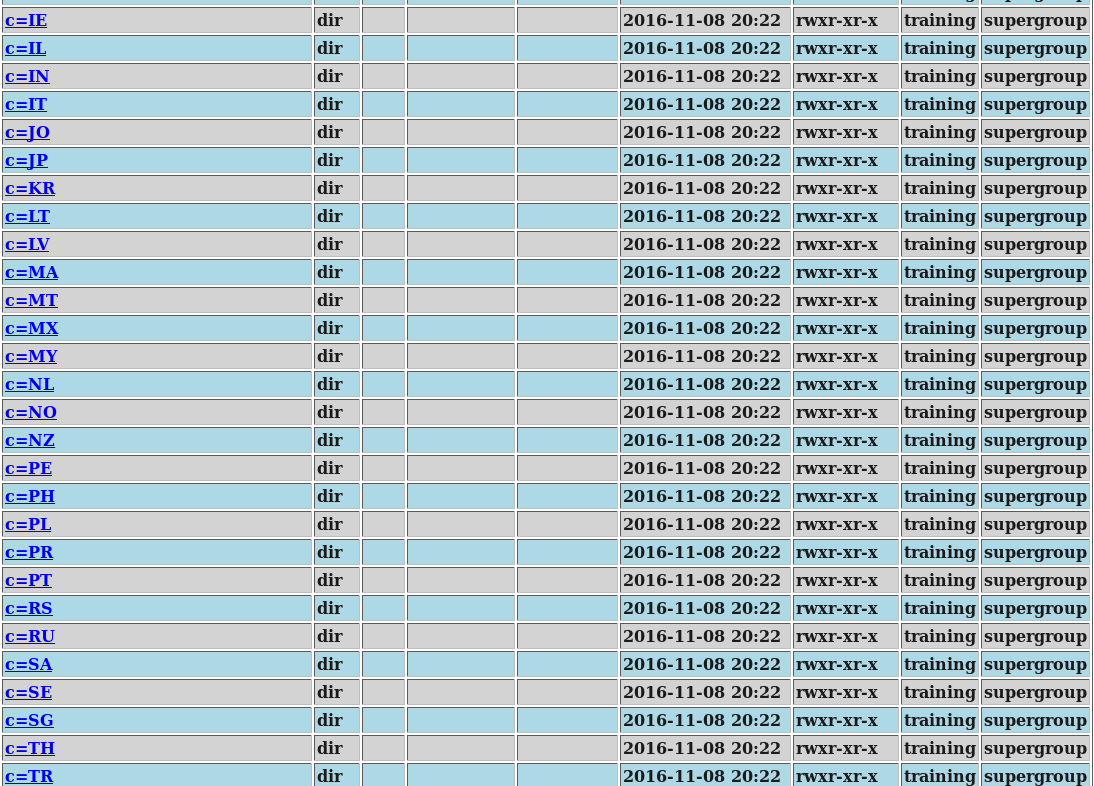
**hive> set hive.exec.dynamic.partition.mode=nonstrick;**

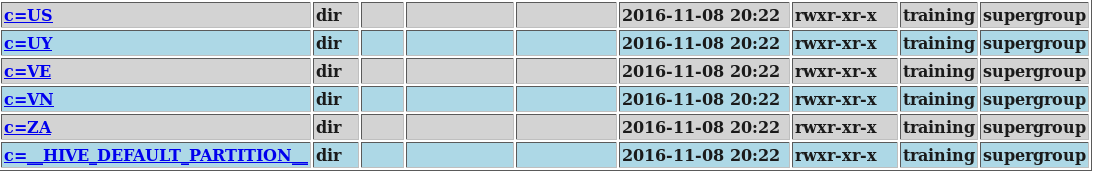
**hive> INSERT OVERWRITE TABLE clickstrpart PARTITION(c) SELECT a,nk,g,h,l,hh,r,u,t,gr,cy,tz,hc,al,c from clickstr;**

**Total MapReduce jobs = 3**

**Launching Job 1 out of 3.0**

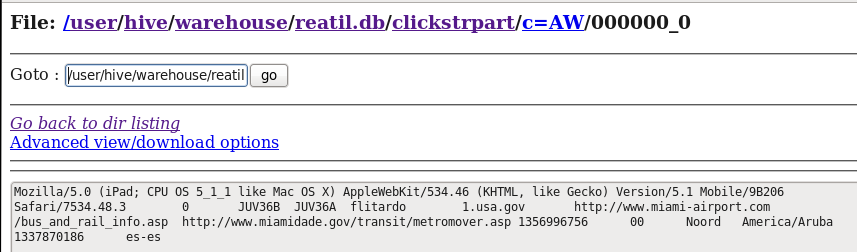
****

****

****

****

****

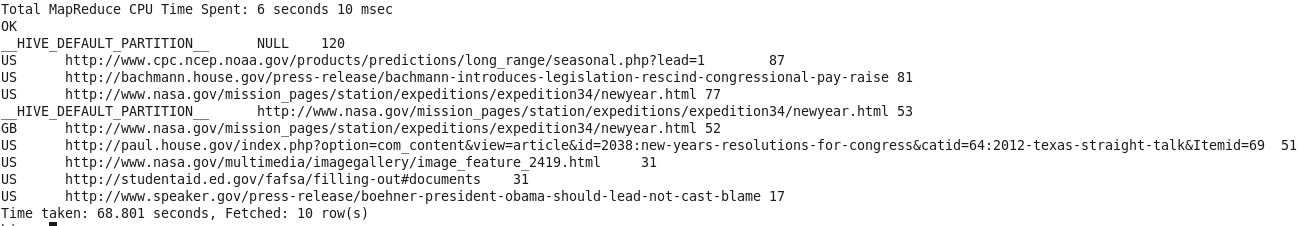
****

**Step 6: Now top 10 sites in term of click is:**

**hive> select c,u,count(\*) as cnt from clickstrpart GROUP BY c,u ORDER BY cnt DESC LIMIT 10;**

**Total MapReduce jobs = 2**

**Launching Job 1 out of 2**

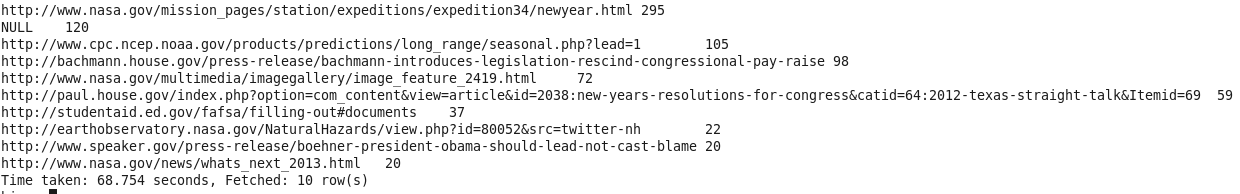
****

**Step 7: top 10 most popular sites for each country will be:**

**hive> select u,count(\*) as cnt from clickstrpart GROUP BY u ORDER BY cnt DESC LIMIT 10;**

**Total MapReduce jobs = 2**

**Launching Job 1 out of 2**

****

**Step 8:**

**The Top 10 most popular sites for each month.**

**Next as we month for the next query i have created the below the MonthUdf using java which will take timestamp as input and return**

**package** month;

**import** java.util.Calendar;

**import** org.apache.hadoop.hive.ql.exec.UDF;

**import** org.apache.hadoop.hive.ql.udf.UDFType;

**import** org.apache.hadoop.io.Text;

@UDFType(stateful = **true**)

**public** **class** MonthUDF **extends** UDF {

**public** **int** evaluate(Text value) {

Calendar cal = Calendar.*getInstance*();

**if** (value == **null**) {

**return** 1;

} **else** {

cal.setTimeInMillis(Long.*valueOf*(value.toString()) \* 1000);

**return** cal.get(Calendar.*MONTH*) + 1;

}

}

}

**Step 8:**

**hive> add jar /home/training/Desktop/pigOperator/monthudf.jar;**

**Step 9:**

**hive> create temporary function month as 'month.MonthUDF';**

**OK**

**Time taken: 0.569 seconds**

**Step 10: Next i have created new table click partition with extra field month and also partitione the table by country so that the quering will be the fast.**

**hive> CREATE EXTERNAL TABLE IF NOT EXISTS clickpartition (a string,nk int,g string,h string,l string,hh string,r string,u string,t bigint,gr string,cy string,tz string,hc bigint,al string, month int) PARTITIONED BY(c string) row format delimited fields terminated by '\t';**

**OK**

**Time taken: 0.665 seconds**

**Step 11:**

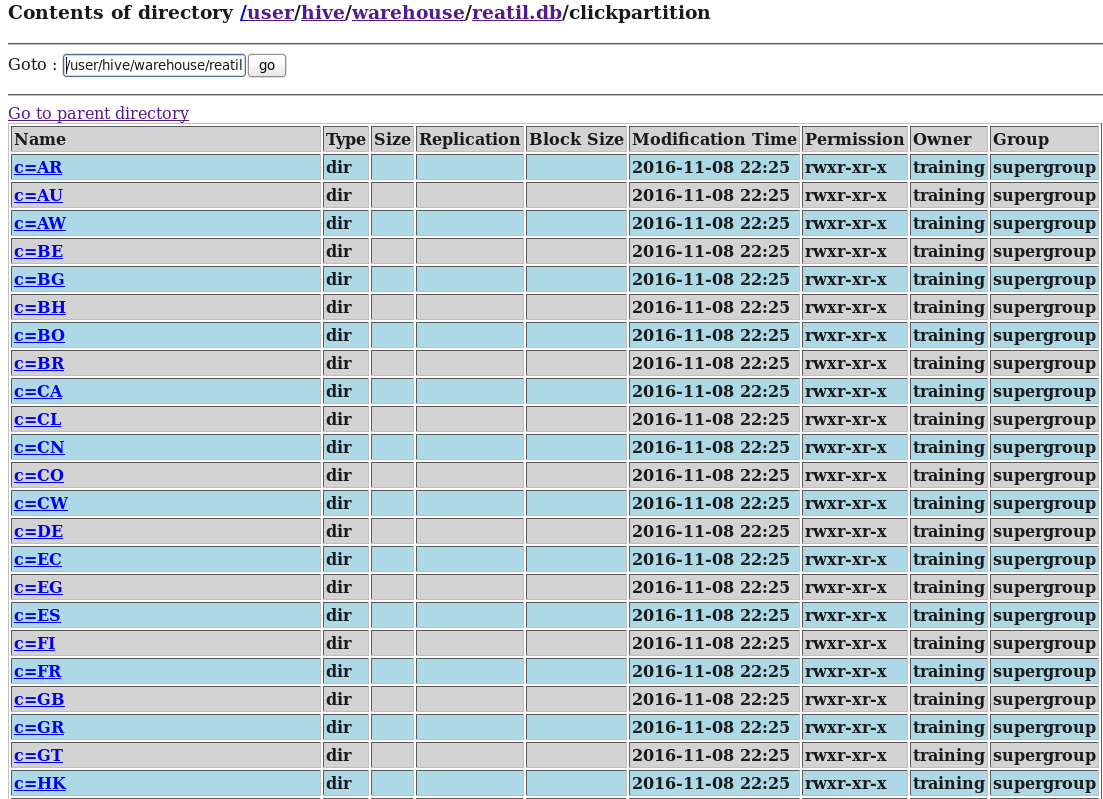
**hive> set hive.exec.dynamic.partition=true;**

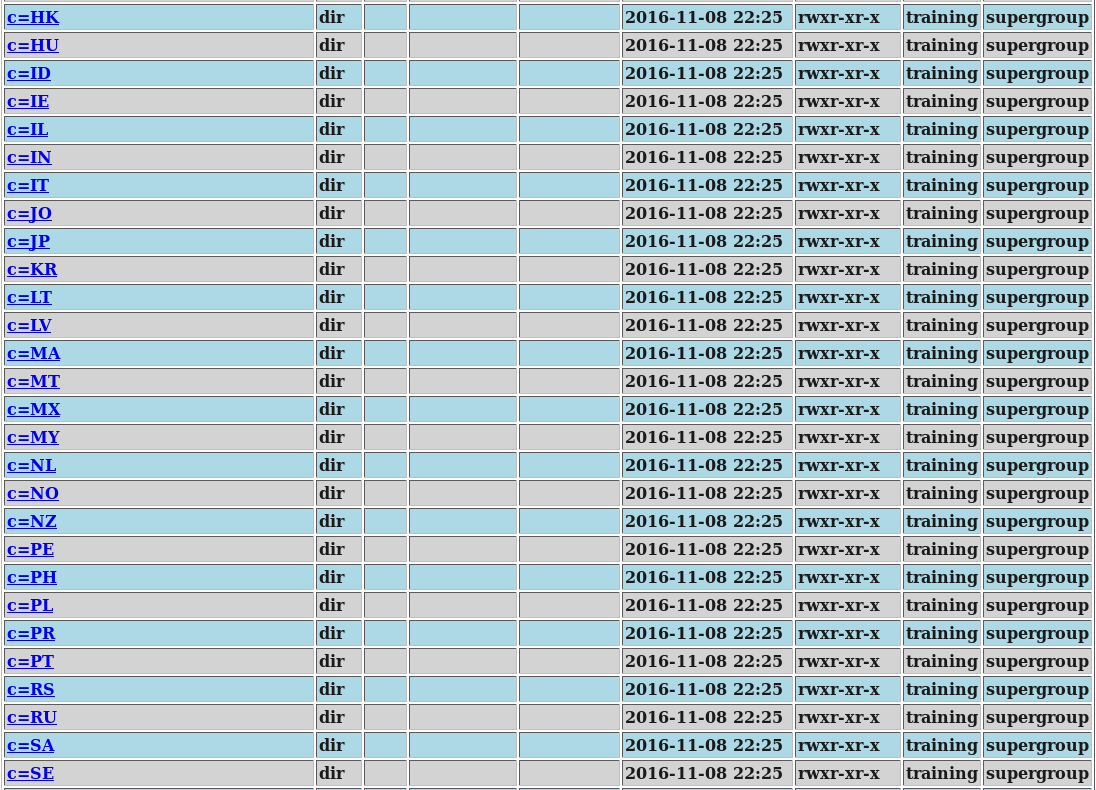
**hive> set hive.exec.dynamic.partition.mode=nonstrick;**

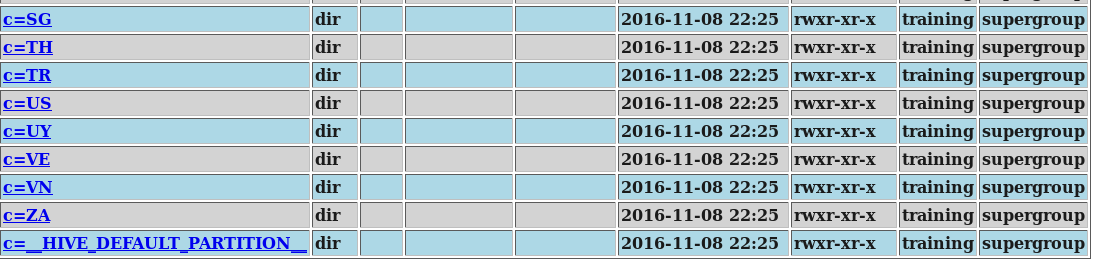
**hive> INSERT OVERWRITE TABLE clickpartition PARTITION(c) SELECT a,nk,g,h,l,hh,r,u,t,gr,cy,tz,hc,al,month(t),c from clickstr;**

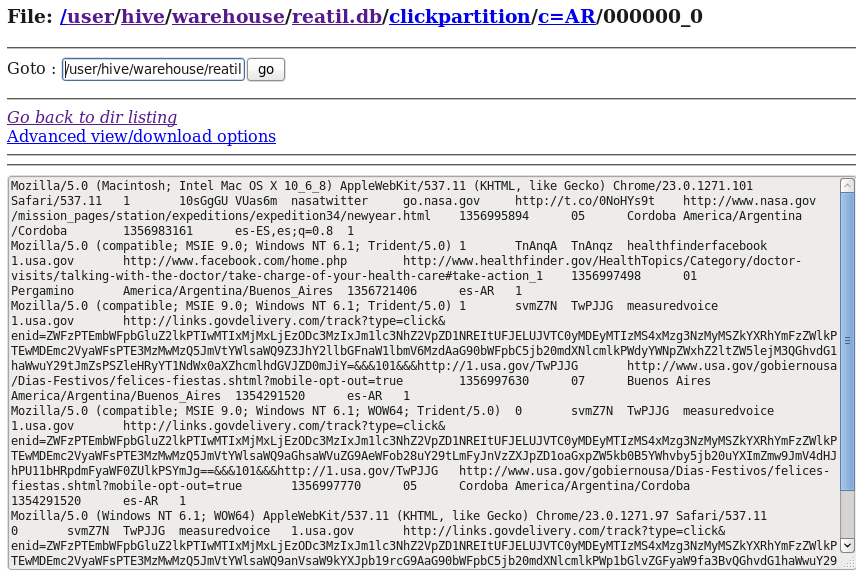
**Total MapReduce jobs = 3**

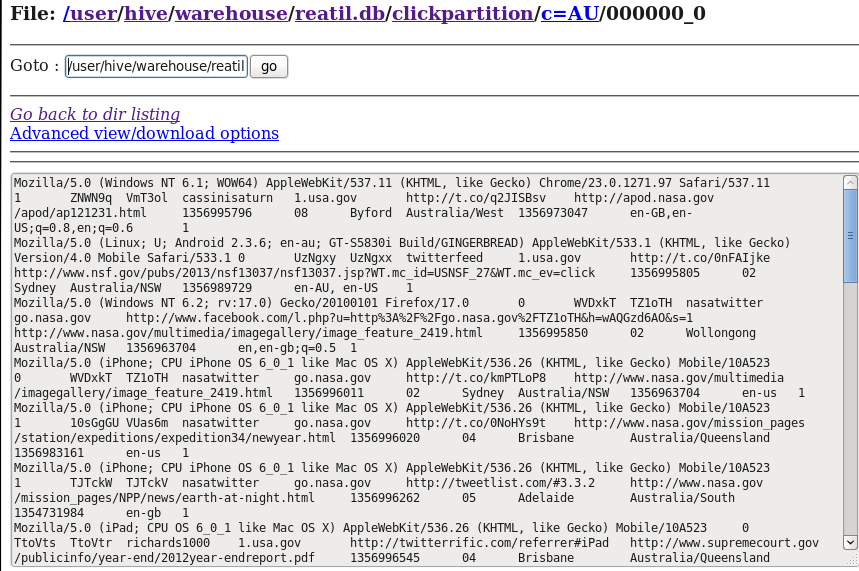
**Launching Job 1 out of 3**

****

****

****

****

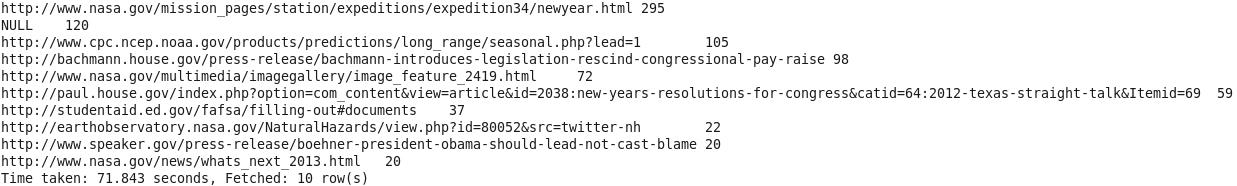
****

**Step 12: The top 10 most popular sites in terms of clicks**

**hive> select u,count(\*) as cnt from clickpartition group by u order by cnt DESC limit 10;**

**Total MapReduce jobs = 2**

**Launching Job 1 out of 2**

****

**Step 13: The top 10 most popular sites for each country.**

**For the above query we need to list top 10 sites in each country group which require some extra code. so i have written RankUDF as below:**

**import** org.apache.hadoop.hive.ql.exec.UDF;

**public** **final** **class** RankUDF **extends** UDF{

**private** **int** counter;

**private** String last\_key;

**public** **int** evaluate(**final** String key){

**if** ( !key.equalsIgnoreCase(**this**.last\_key) ) {

**this**.counter = 1;

**this**.last\_key = key;

}

**return** **this**.counter++;

}

}

**Step 14:**

**hive> add jar /home/training/Desktop/pigOperator/rankudf.jar;**

**Step 15:**

**hive> create temporary function rank as 'RankUDF';**

**OK**

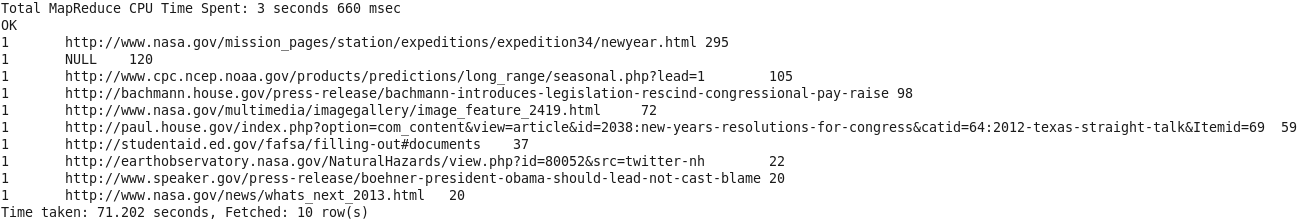
**Time taken: 0.003 seconds**

**Step 16:**

**hive> select month,u,cnt from( select \*, rank(month) as row\_number from ( select month,u,count(\*) as cnt from clickpartition group by month,u order by month, cnt DESC) A ) B where row\_number<=10;**

**Total MapReduce jobs = 2**

**Launching Job 1 out of 2**

****