HIVE – CASE – STUDY

ANALYSIS FOR CLICK STREAM DATA

We are now dealing with a data set where web sites of several countries in the world are included. Our Primary task is to store this data on Hadoop and Compute it.

Environment Setup:

1. If you’re using Cloudera or HDP sand box, it easy to get this data set, if you’re using any cloud or third party lab environment push your dataset by using putty or WinSCP.

2. Push your data set to HDFS by using: hdfs dfs –put dataset.

3. As it is a JSON file, download the hive SerDe from: <http://files.cloudera.com/samples/hive-serdes-1.0-SNAPSHOT.jar> and make sure you move the file to hdfs

4. In hive add your SerDe by a command

Add jar hive-serdes-1.0-SNAPSHOT.jar;

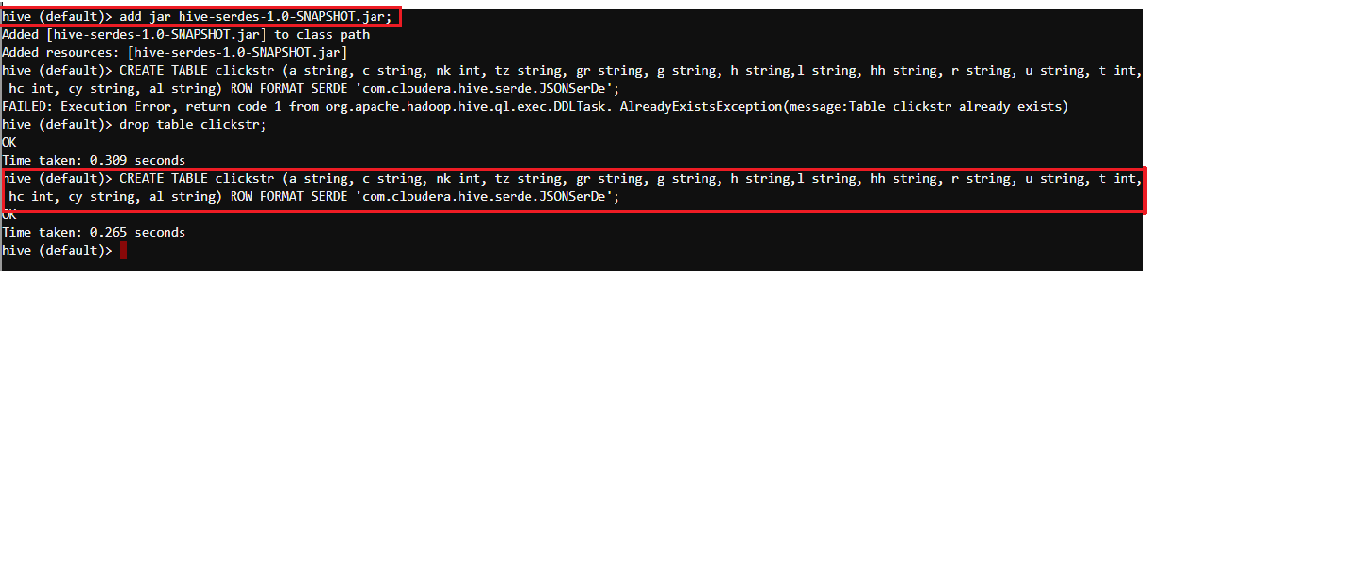
5. Once you’re done with these prior steps you are ready to play with your data set in Hive.

LET’S CREATE A DATABASE AND LOAD OUR DATA IN HIVE TABLES

1. Create a Hive table for our JSON data format.

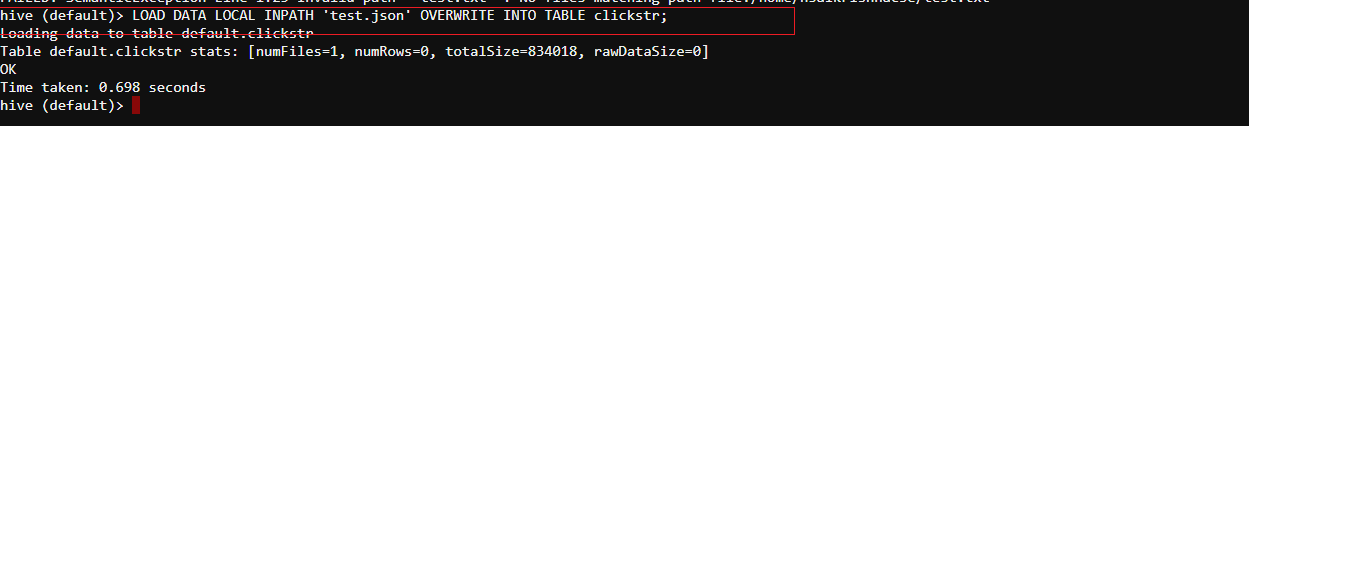
ADD JAR <localpath where hive json serde jar is located>

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';

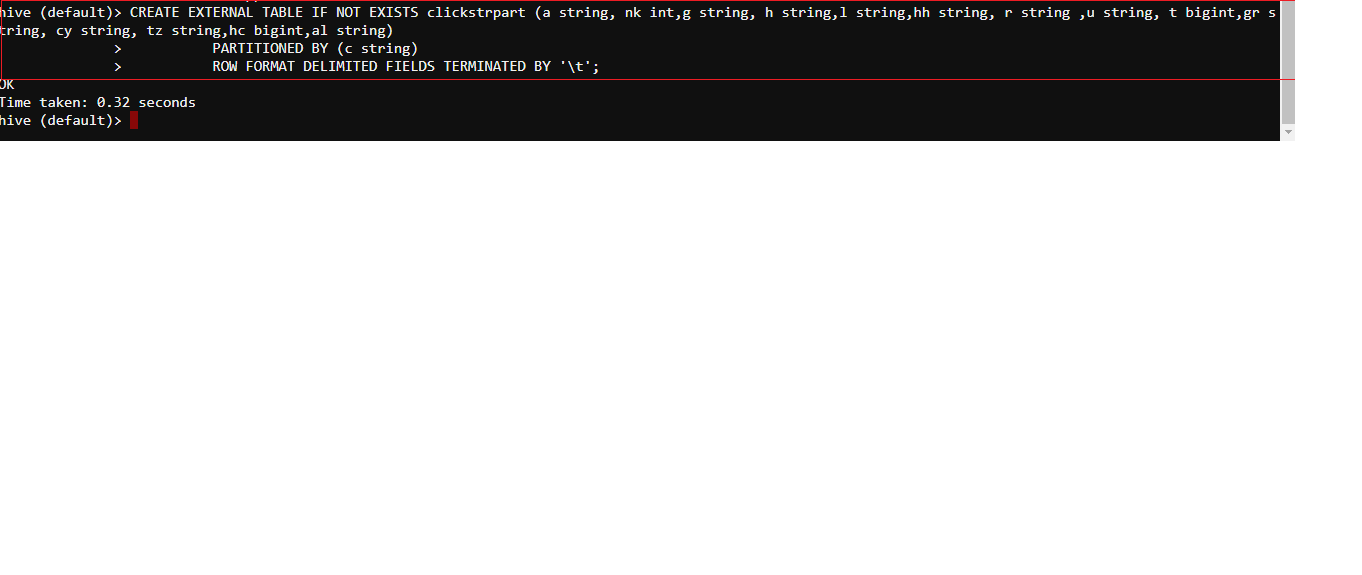


2. Load your Json data into the table you created.

LOAD DATA LOCAL INPATH '/home/user/UsaGovData.txt' OVERWRITE INTO TABLE clickstr;



3. Perform Dynamic partition on the above table for query tuning and load the data into it.

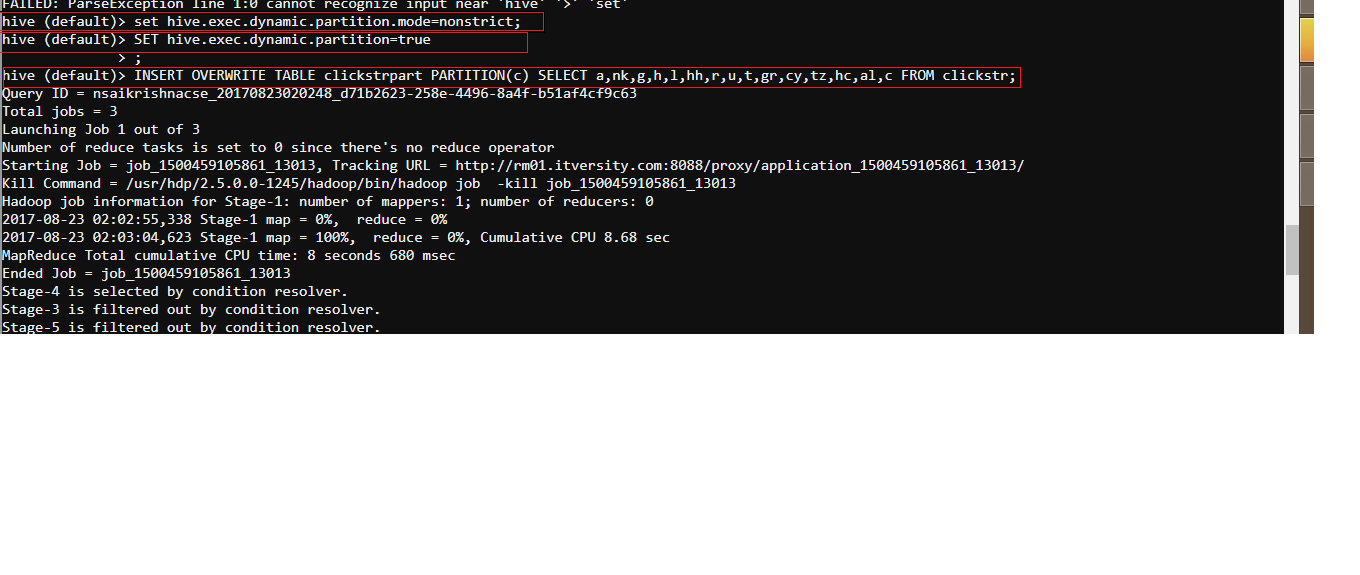


4. Now Enable Dynamic partition and load the data into it.

hive> SET hive.exec.dynamic.partition=true

hive> set hive.exec.dynamic.partition.mode=nonstrict

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;



LET’S PERFORM QUERIES:

PLEASE TAKE A LOOK INTO THE JSON SCHEMA VIZ THE COLUMNS WE HAVE:

{

"a": USER\_AGENT,

"c": COUNTRY\_CODE, # 2-character iso code

"nk": KNOWN\_USER, # 1 or 0. 0=this is the first time we've seen this browser

"g": GLOBAL\_BITLY\_HASH,

"h": ENCODING\_USER\_BITLY\_HASH,

"l": ENCODING\_USER\_LOGIN,

"hh": SHORT\_URL\_CNAME,

"r": REFERRING\_URL,

"u": LONG\_URL,

"t": TIMESTAMP,

"gr": GEO\_REGION,

"ll": [LATITUDE, LONGITUDE],

"cy": GEO\_CITY\_NAME,

"tz": TIMEZONE # in http://en.wikipedia.org/wiki/Zoneinfo format

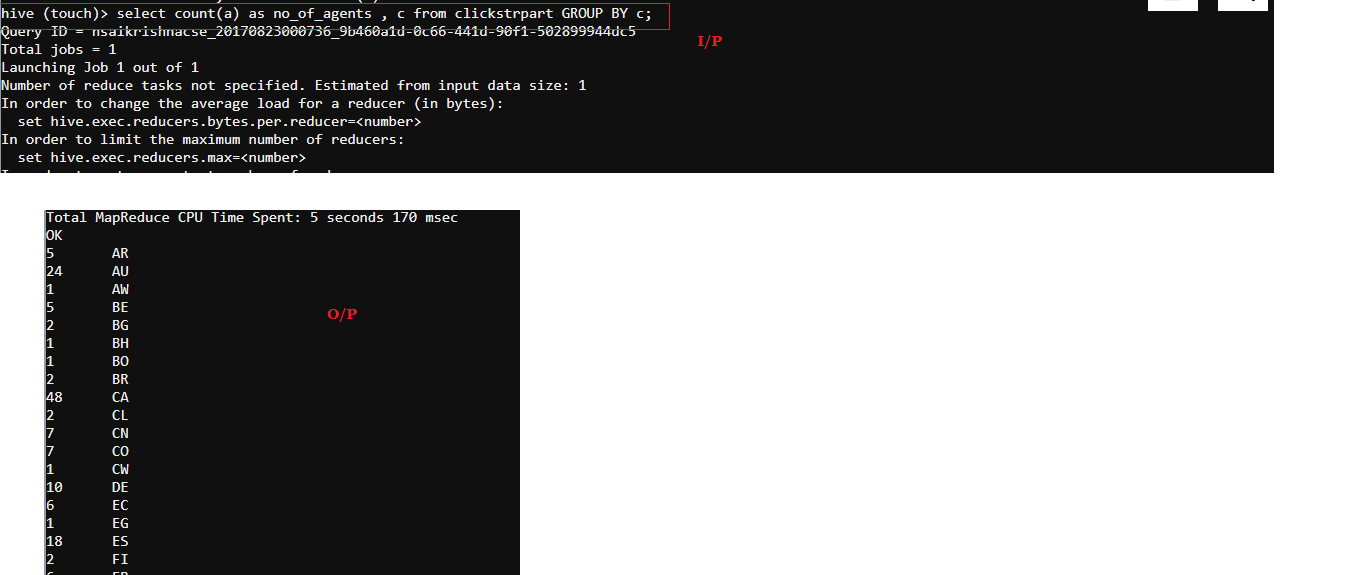
"hc": TIMESTAMP OF TIME HASH WAS CREATED,

"al": ACCEPT\_LANGUAGE http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html#sec14.4

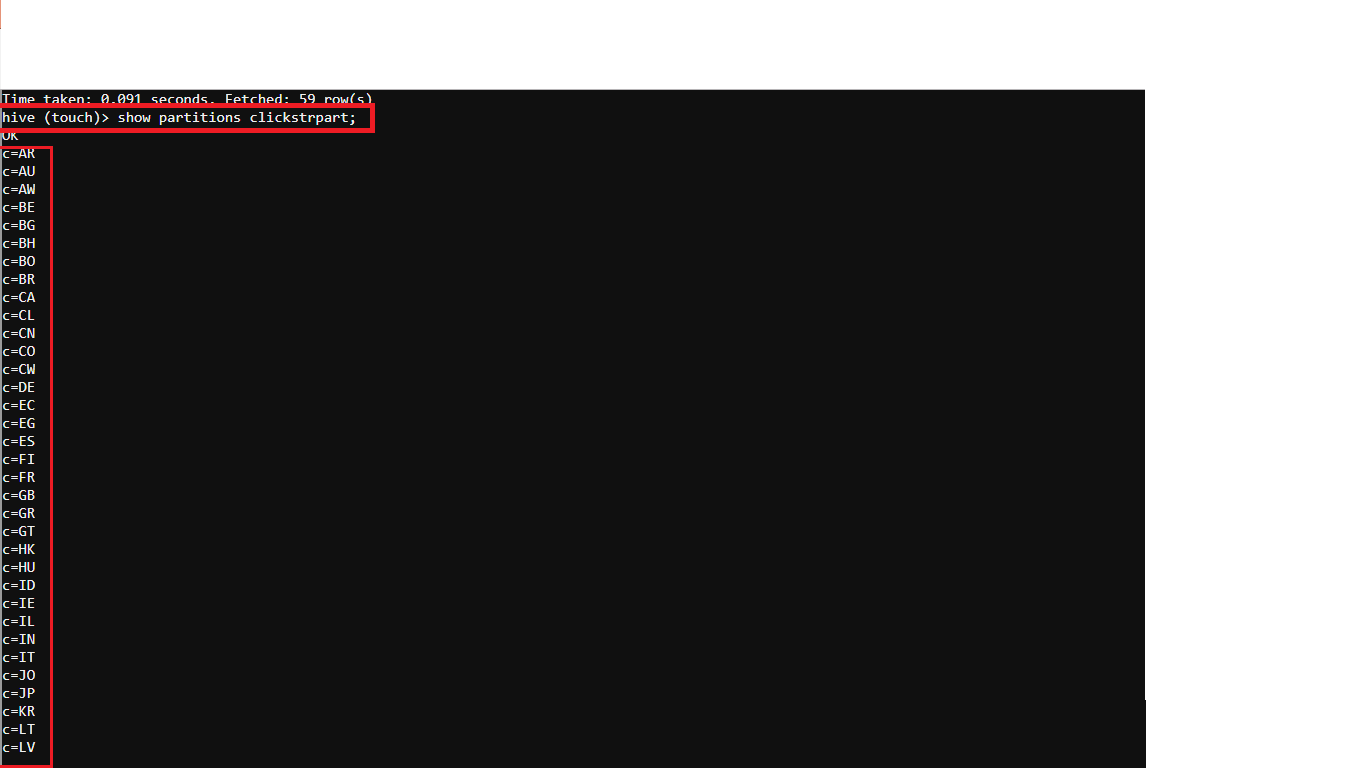
}

1. Write a hive query to find the agents based on the country code.

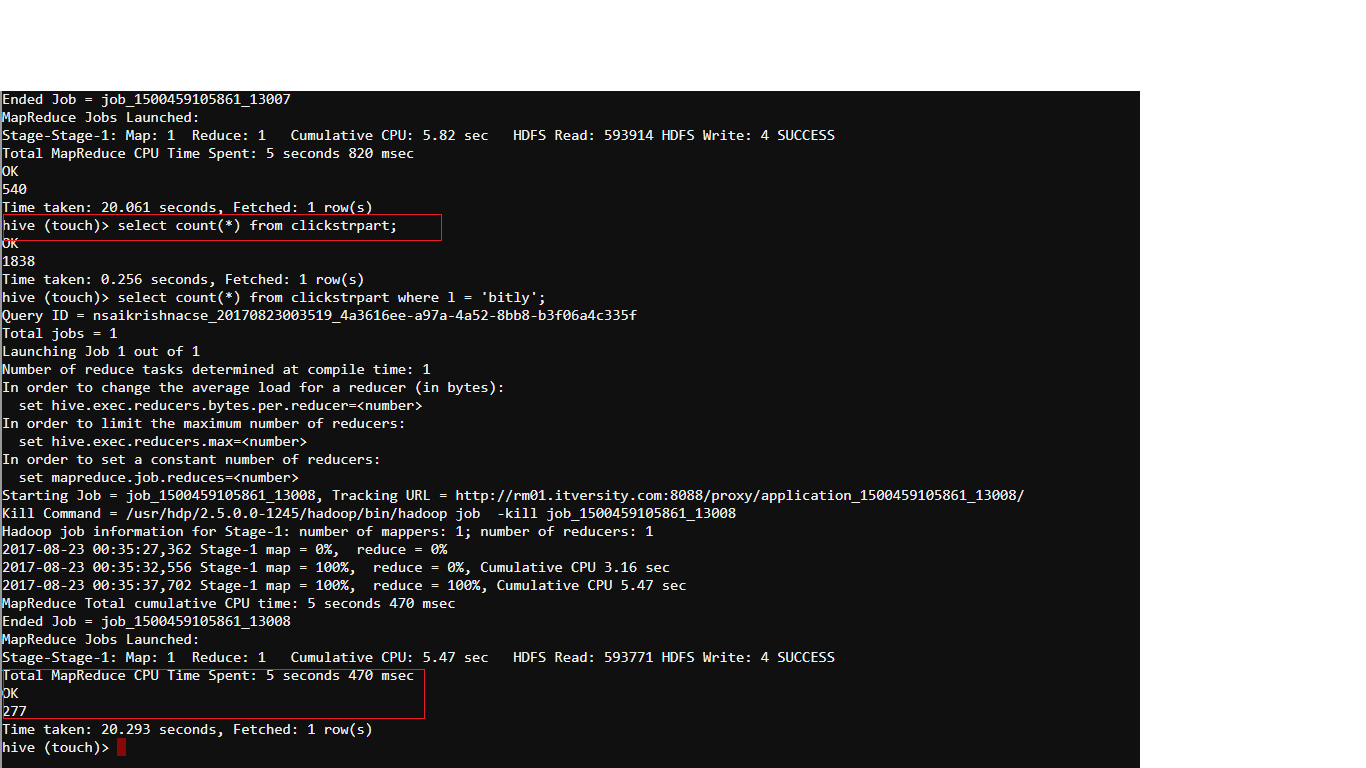
Ans: select count (a) as no\_of\_agents, c from clickstrpart GROUP BY c;



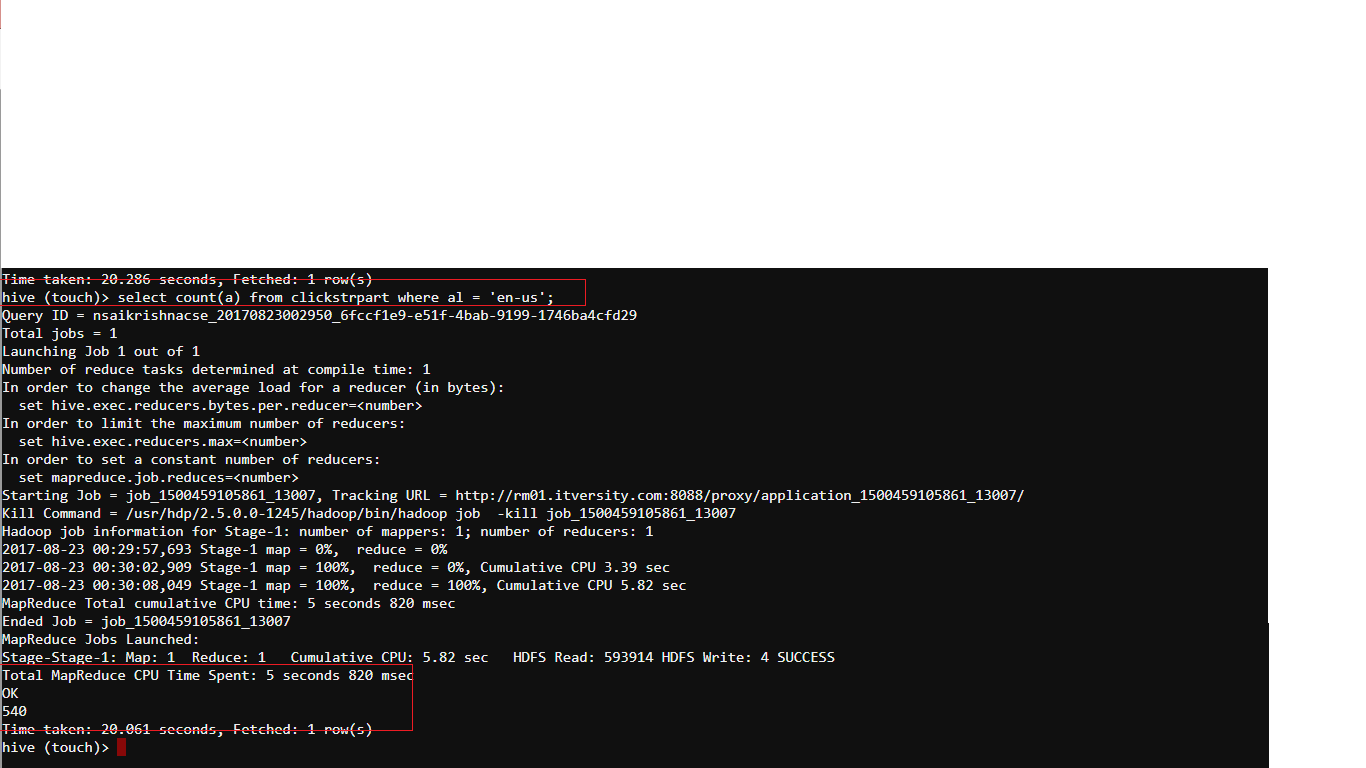
2. Write a query to find out the partitions.



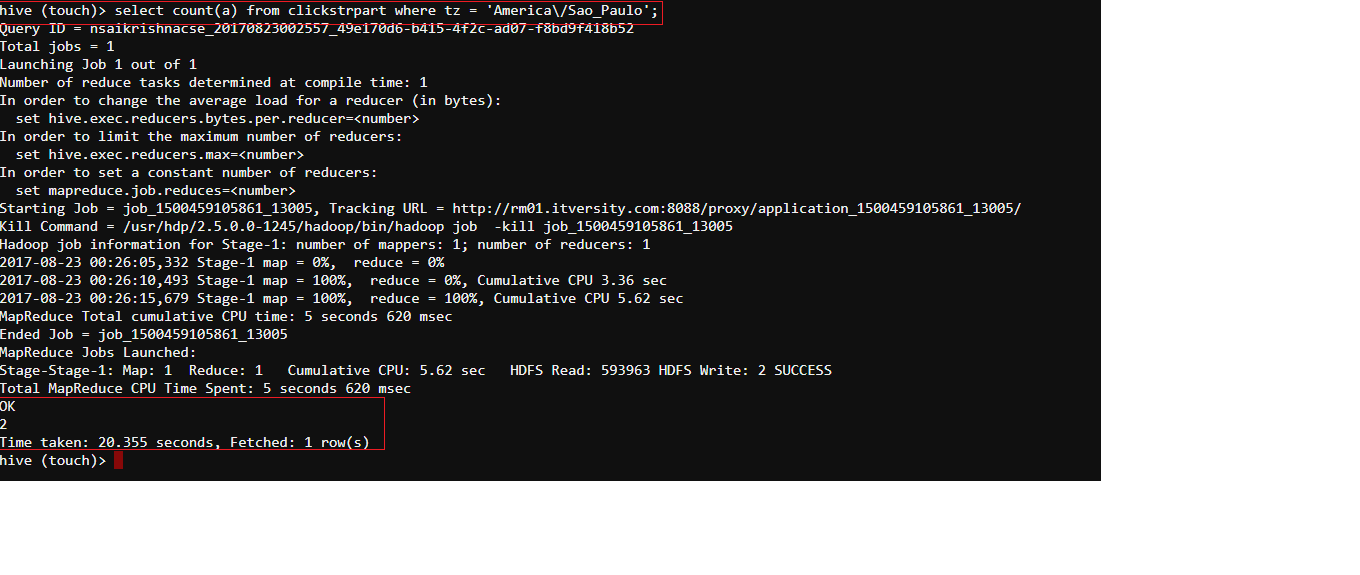
3. Write a query to find the users as per ENCODING\_USER\_LOGIN.

Ans : 

4. Write a Query to find the users as per their language.



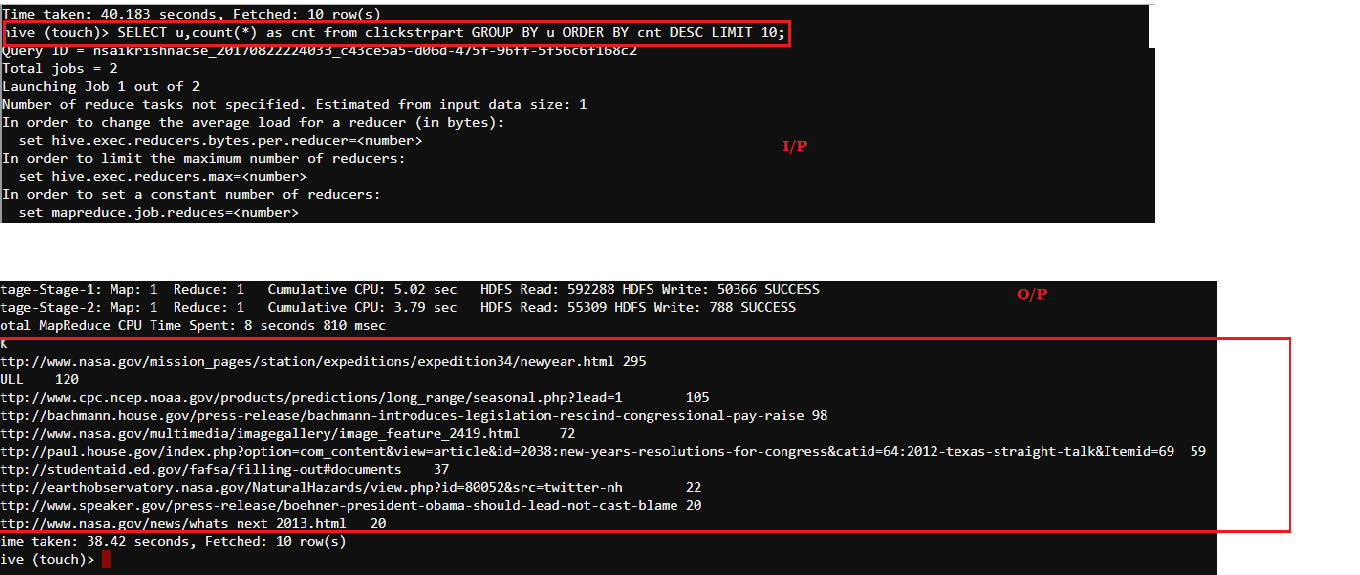
5. Write a query to find the users as per their time zone



6. Write a query to group by the first view and latest view of web sites done by the user.



7. Write a query to find out 10 popular sites in each country by considering LONG\_URL ‘u’.



8. Write a query to find out 10 popular sites in each country by considering both LONG\_URL ‘u’ and Country code ‘c’.

