Query 2

Part A: Which part of the US has the most Data Engineer jobs for each year?

Platform used : HIVE

Query for the above question :

hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/h1b\_query2\_a/op2a/2016' row format delimited fields terminated by ','

> select a.year,a.worksite,a.job\_title,count(\*) as cnt from h1b\_app2 a where (a.year = 2016 AND a.job\_title = ‘DATA ENGINEER’) group by a.year,a.worksite,a.job\_title order by cnt desc limit 1;

hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/h1b\_query2\_a/op2a/2015' row format delimited fields terminated by ','

> select a.year,a.worksite,a.job\_title,count(\*) as cnt from h1b\_app2 a where (a.year = 2015 AND a.job\_title = ‘DATA ENGINEER’) group by a.year,a.worksite,a.job\_title order by cnt desc limit 1;

hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/h1b\_query2\_a/op2a/2014' row format delimited fields terminated by ','

> select a.year,a.worksite,a.job\_title,count(\*) as cnt from h1b\_app2 a where (a.year = 2014 AND a.job\_title = ‘DATA ENGINEER’) group by a.year,a.worksite,a.job\_title order by cnt desc limit 1;

hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/h1b\_query2\_a/op2a/2013' row format delimited fields terminated by ','

> select a.year,a.worksite,a.job\_title,count(\*) as cnt from h1b\_app2 a where (a.year = 2013 AND a.job\_title = ‘DATA ENGINEER’) group by a.year,a.worksite,a.job\_title order by cnt desc limit 1;

hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/h1b\_query2\_a/op2a/2012' row format delimited fields terminated by ','

> select a.year,a.worksite,a.job\_title,count(\*) as cnt from h1b\_app2 a where (a.year = 2012 AND a.job\_title = ‘DATA ENGINEER’) group by a.year,a.worksite,a.job\_title order by cnt desc limit 1;

hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/h1b\_query2\_a/op2a/2011' row format delimited fields terminated by ','

> select a.year,a.worksite,a.job\_title,count(\*) as cnt from h1b\_app2 a where (a.year = 2011 AND a.job\_title = ‘DATA ENGINEER’) group by a.year,a.worksite,a.job\_title order by cnt desc limit 1;

Explanation:

In the above queries I am making a count on the worksite and grouping it on year and worksite. Then I am sorting it in descending order based on the count and limiting it to just one record so that I get the job with highest petition each year.

I had save the output files in different folder in accordance to the year, however for the convenience of viewing it I have saved all the outputs in one file. Please click on the file below for the same:



Part B: find top 5 locations in the US who have got certified visa for each year.

Platform used : PIG

Source input file:

/home/cloudera/h1b\_data

Explanations:

Please refer to the pig file below for the queries.



Step1) In the file I filtered all the data to case\_status = = ‘CERTIFIED’ in one bag.

Step2) Then I filtered the above bag to specific year and put them in bags for year 2016, 2015 and so on.

Step3) Grouped each bag from step2 in terms of worksite

Step4) Counted each of the bags from step3 in terms of worksite

Step5)Sorted each bag in step4 in descending order.

Step6)Limited the bags in step5 to 5

Output files are below:

