Query 1

Part A: Is the number of petitions with Data Engineer job title increasing over time?

Platform used : HIVE

Query for the above question :

hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/h1b\_query1\_a' row format delimited fields terminated by ','

> select year, job\_title, count(\*) from h1b\_app2 where job\_title = 'DATA ENGINEER' group by year;

Explanation:

What I have done in the query is select the count of the records in the h1b\_app2 table by grouping it on year and job title. I have selected only those record where the job\_title column has ‘DATA ENGINEER’ as value, in the where clause. I saved the output in a folder. So yes the number of applications for Data Engineering is increasing every year.Please double click on the output file below:



Part B: Find top 5 job titles who are having highest growth in applications.

Platform used : Mapreduce

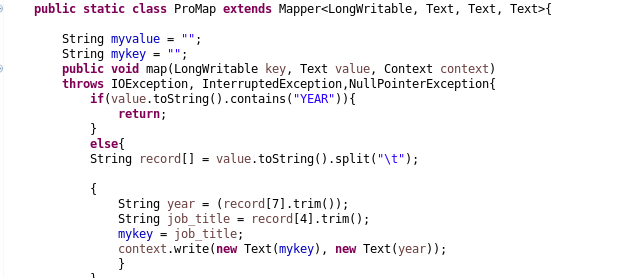
Two programs were used here. They are qyery1b.java to calculate the average growth and query1b\_top5 .java to sort and find the top 5 job titles in terms of highest average growth.

Explanation:

Program- query1b.java

Source input file-

/user/hive/warehouse/project.db/h1b\_app2

Mapper

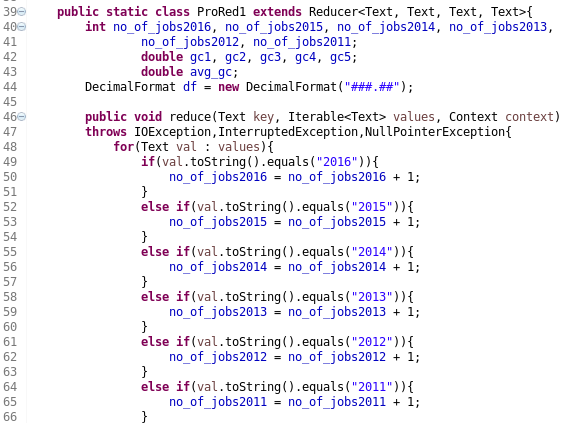
As you can see I have applied an IF condition to avoid the very first record on the table h1b\_app2 as it consists of a header as shown below



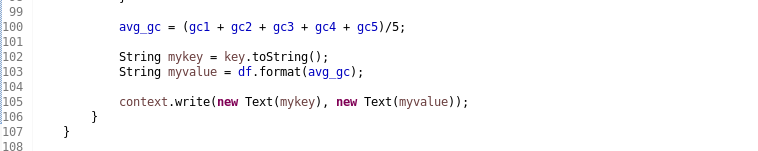
So if the record contains a string as year it will return and start with the next record.

From the records I have taken the job\_title(4th position in the array) as the key and year(7th position in the array) as the value.

Reducer







In the reducer The code basically takes each key (job\_title) and counts the number of entries for that key into separate variable for year 2011, 2012, 2013, 2014, 2015 and 2016. This will represent the number of applications for each job in each year. We take this to find the growth in each of the 5 growth cycles(line 69 to line 96). Then we calculate the average growth for each job and write the job\_title as key and average growth cycle as value, to the memory(from line 99 to line 105).

The whole program is attached below:



Program- query1b\_top5

Source input file:

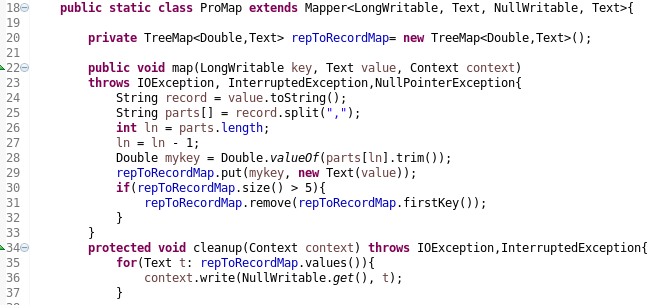
Output file from executing the above program.

Explanation:

This program only consists of a mapper. Here I have used TreeMap to collect the data, from the output file of query1b.java, in a TreeMap. The objective is to set the output data in order of their average growths . If the the TreeMap size exceeds 5 we are removing the first record by its key. This is to get the top 5 records.

The average growth might go beyond 100 because there are certain jobs which may have 0 applications for a particular year.

Mapper



The output file is attached below:



Program below:

