Query 4

Which top 5 employers file the most petitions each year?

Platform used: Mapreduce

2 programs were used here.

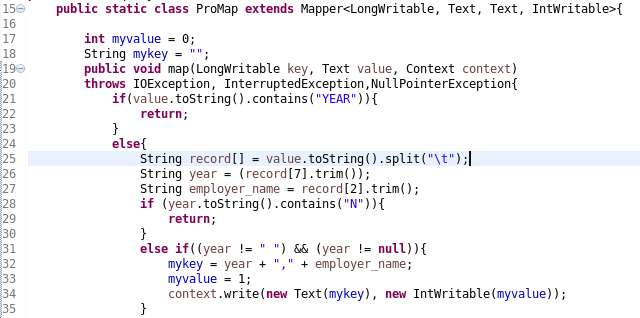
query4.java to calculate the number of petitions for each year by each employer an query4\_top5.java to sort and find the top 5 employers in that respect.

Program- query4.java

Source input file:

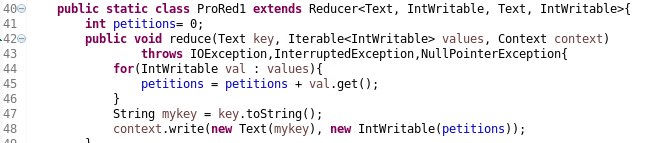
/user/hive/warehouse/project.db/h1b\_app2

Mapper



As usual record[7] and, record[2 ] were put in year and employer\_name variable respectively, from the table h1b\_app2. The key was a combination of year and employer name, because we need to count the number of petitions for each employer in each year. For value we took an integer variable as 1, which we will later use to add and count the petitions in the reducer. The essential part in this piece of logic is taking care of the null values in the year column. As you can see I have worked extensively to making sure that there is no null value in the year variable when it is being written to memory.

Reducer



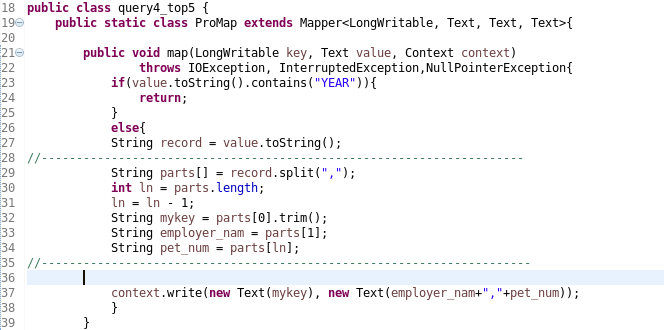
The reducer part is fairly simply as we just took the keys and added all the 1s that we had accumulated from the mapper to get the total petitons of each employer in one year.

Program- query4\_top5

Source input file:

The output file from the execution of query4.java

Mapper



In the mapper we split each record, from the output file of query4.java, into employer\_nam, mykey(which is the year) and pet\_num(number of petitions for each employer). The key was the year value and value was the combination of employer\_nam and the pet\_num, so that we can write into memory the count for each year and sort the same in the reducer.

Reducer



For sorting we used a TreeMap as shown from lines 45 to 57. The sorting was done on the pet\_num which we had obtained from the mapper. We have put the sorted results into an array petits . In lines 62 to 67 we take the size of the array and in the loop we write the contents of the array till the size reduces from 4 to 0(array positions). Thus giving us the top 5 employers for each year.

I have attached the codes below:





The final output is below:

