**PROJECT**

Juhi Deshpande CWID: A20430435

***Project Topic:*** Big Data Analysis on H1-B Visa Petitions 2011-16 using Hive, SparkSQL and Pig and their performance comparisons

***Introduction:***

The H1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for H1B visa, an US employer must offer a job and petition for H1B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college/ higher education (Masters, Ph.D.) and work in a full-time position. The dataset consists of the H1-B data of the employees between the year 2011-16 with approximately 3 million records. It includes columns like case status, employer name, SOC name, job title, fulltime position status, prevailing wage, year filed, worksite.

***Project Description:***

The project solves the following test cases:

1. Maximum, Minimum, Average based on the wage for the years
2. Which worksite has the popular job title?
3. Which job title has the case status as “Certified”, “Certified -withdrawn”, “Denied”, “Withdrawn”?
4. How many H1-B filers are “Certified” having the fulltime status as “Yes”?
5. What is the general trend of the employer’s H1-B status in the period 2011-16?
6. What is the general trend of the employer name with respect to the year?
7. Is the number of petitions with Data Engineer job title increasing over time?
8. Which part of the US has the most Data Engineer jobs?
9. Which industry has the greatest number of Data Scientist positions?
10. Which employers file the most petitions each year?

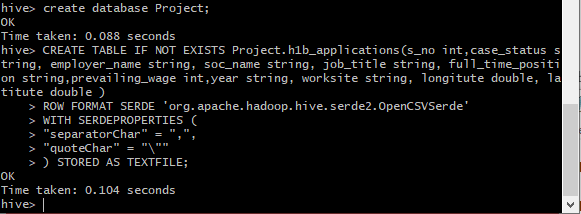
The project is on a dataset that contains 100,000 to 500,000 records and performance analysis is done using Hive with Tez and MapReduce, SparkSQL and Pig. It compares the performance with SparkSQL, Hive and Pig. It would load the data from the csv file and store it into the table where the processing is carried out.

***Dataset Description:***

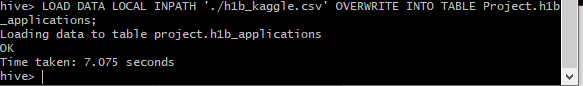
* Case\_status – Status of the application
* Employer\_name – Name of employer registered in H1b visa application
* Soc\_name – Occupation code for employment
* Job\_title – Job title for employment
* Full\_time\_position – Whether application is for full time position or part-time position
* Prevailing\_wage – The most frequent wage
* Year – The application year
* Worksite – The address of the employer worksite

**Using Hive:**

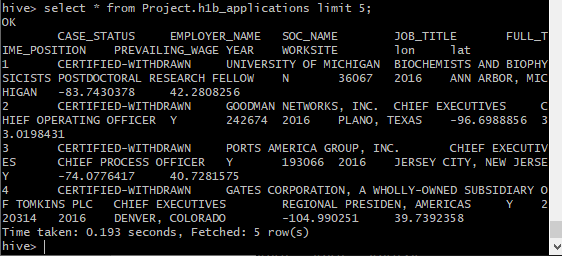
Table created

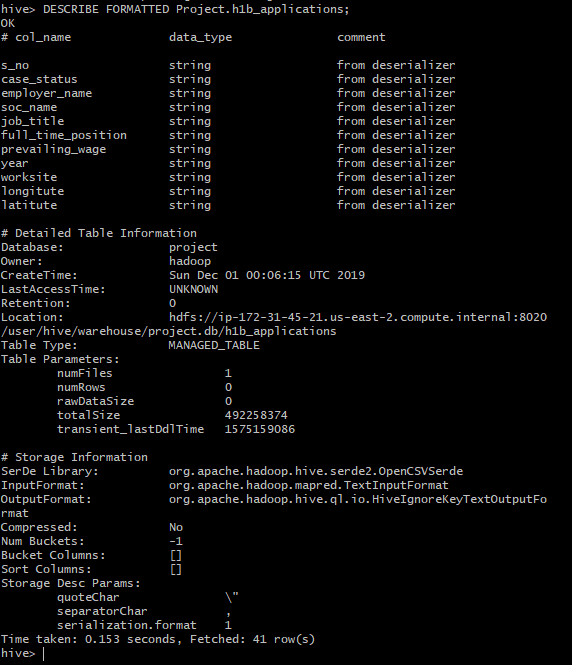


Loading data



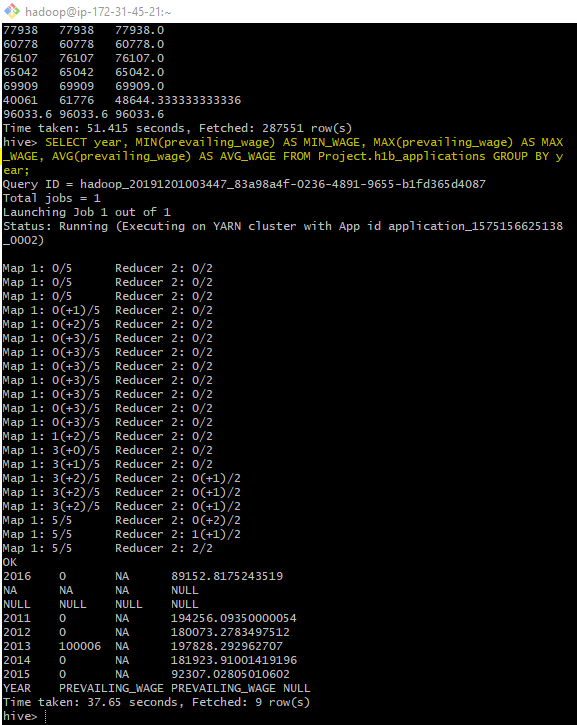
Selecting data



Describing data

**Testcases implemented:**

1. Maximum, Minimum, Average based on the wage for the years

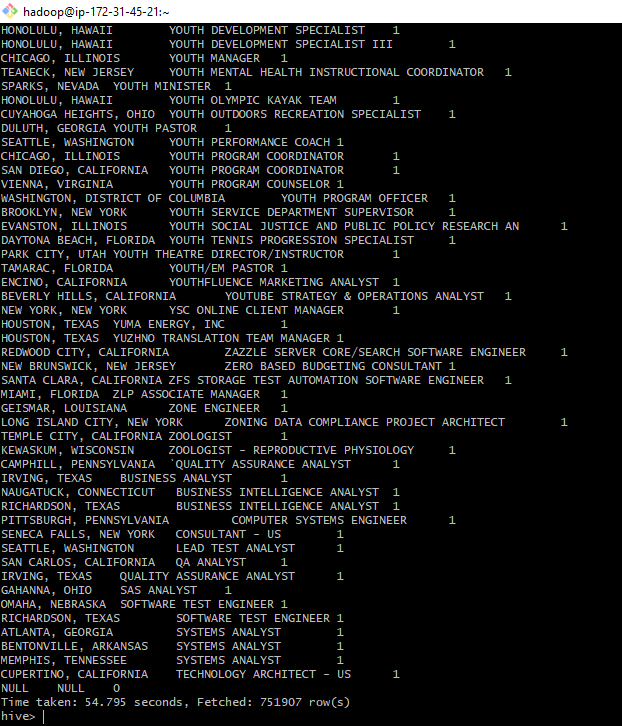


1. Which worksite has the popular job title?

select worksite,job\_title, count(job\_title) as popular\_jobtitle from Project.h1b\_applications

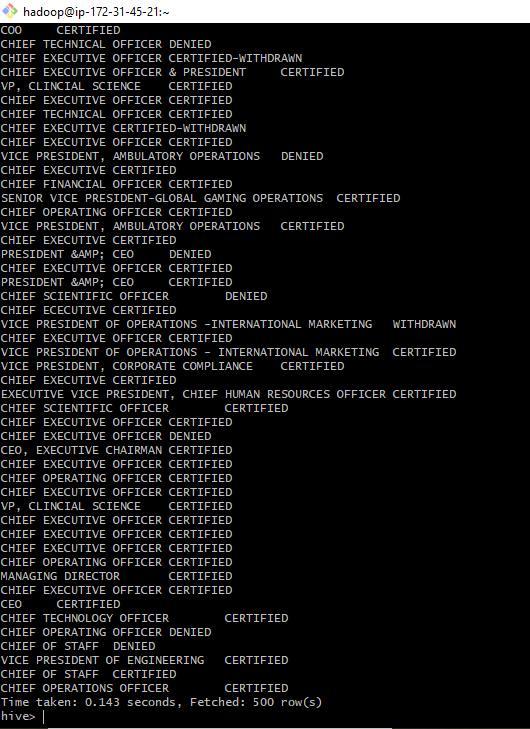
group by worksite, job\_title

order by popular\_jobtitle desc;

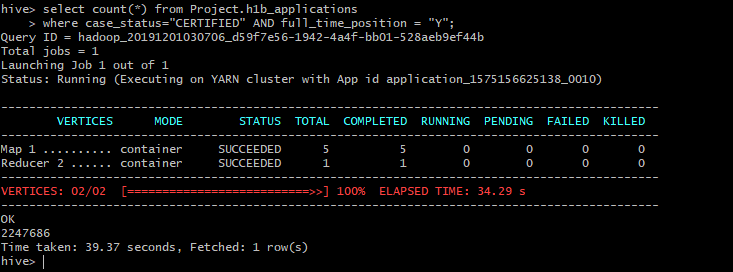


1. Which job title has the case status as “Certified”, “Certified -withdrawn”, “Denied”, “Withdrawn”? (Data Limited to 500 rows)

select job\_title, case\_status FROM Project.h1b\_applications;



4)How many H1-B filers are “Certified” having the fulltime status as “Yes”?

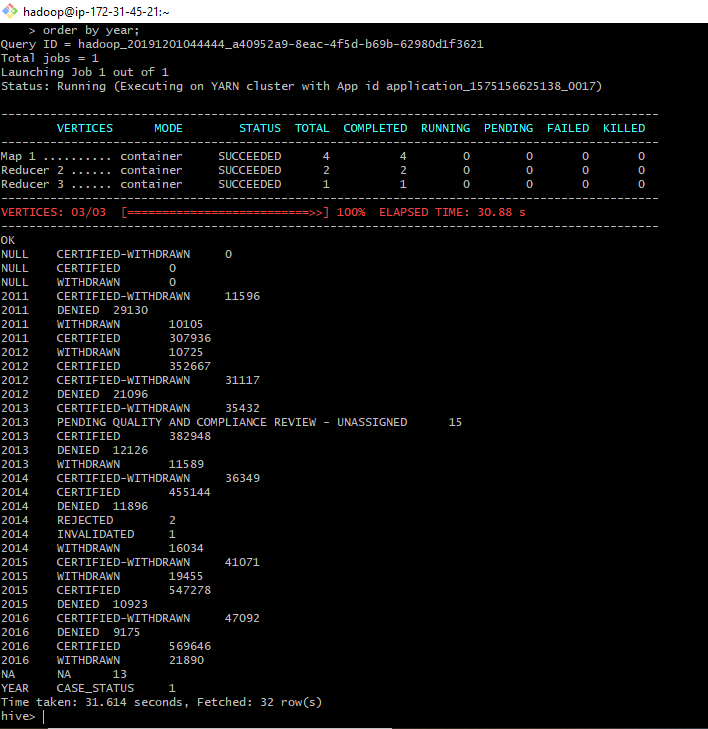


5)What is the general trend of the employer’s H1-B status in the period 2011-16?

select year,case\_status,count(year) from Project.h1b\_applications

group by year,case\_status

order by year;



6.What is the general trend of the employer name with respect to the year?

select year,employer\_name, count(year)

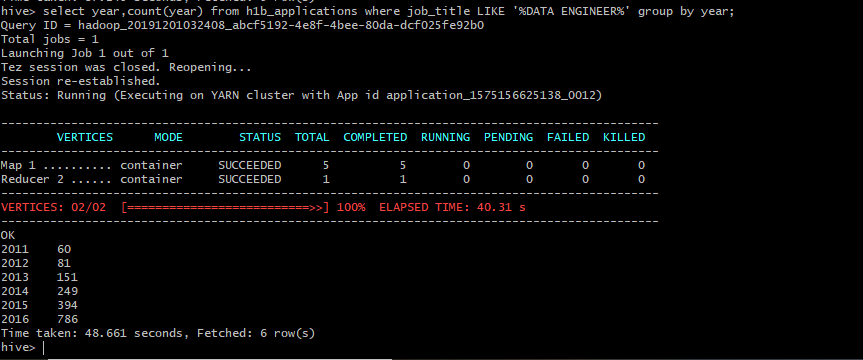
from Project.h1b\_applications

group by year,employer\_name

order by year;



7.Is the number of petitions with Data Engineer job title increasing over time?



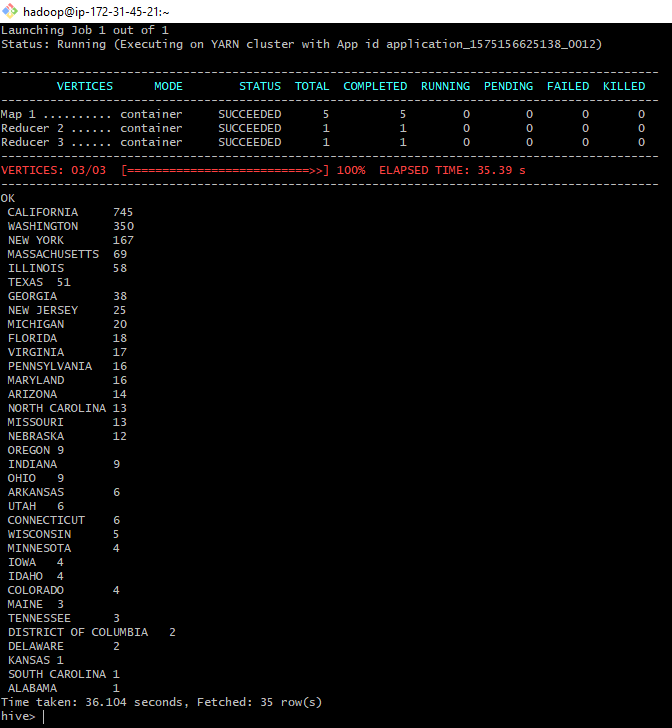
8)Which part of the US has the most Data Engineer jobs?

select split(worksite,'[,]')[1] as state, count(split(worksite,'[,]')[1]) as job\_cnt

from Project.h1b\_applications where job\_title LIKE '%DATA ENGINEER%'

group by split(worksite,'[,]')[1]

order by job\_cnt desc;



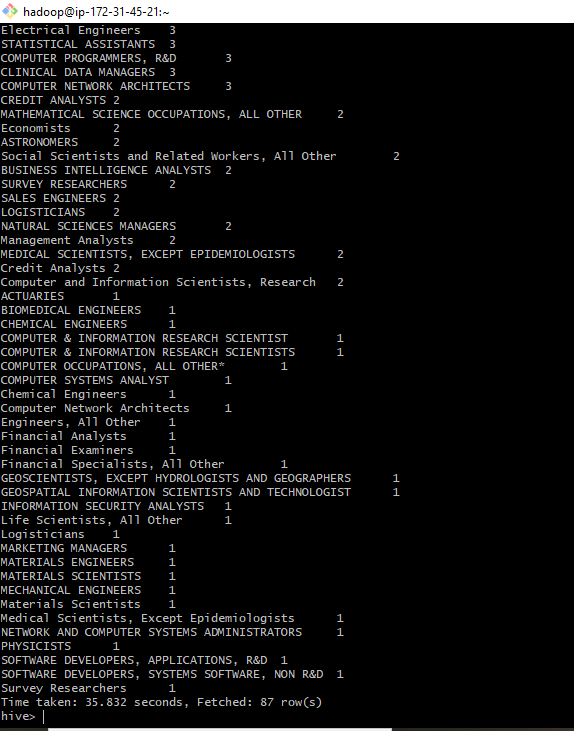
9)Which industry has the greatest number of Data Scientist positions?

select soc\_name, count(soc\_name) as job\_cnt

from Project.h1b\_applications where job\_title LIKE '%DATA SCIENTIST%'

group by soc\_name

order by job\_cnt desc;



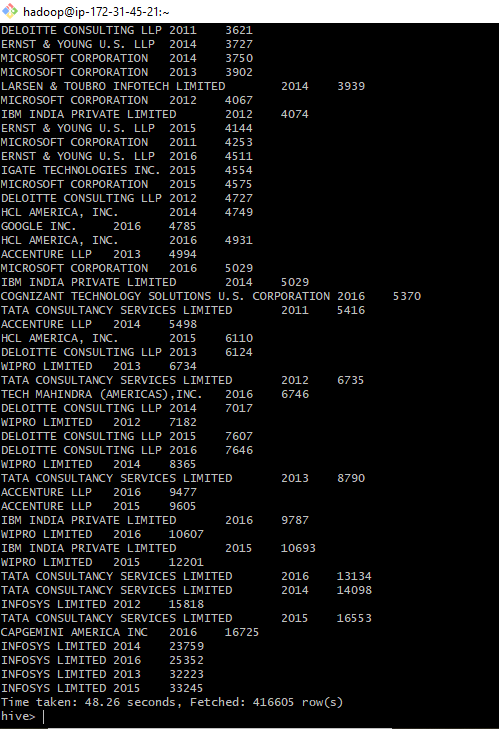
10)Which employers file the most petitions each year?

select employer\_name, year,count(year) as cnt

from Project.h1b\_applications

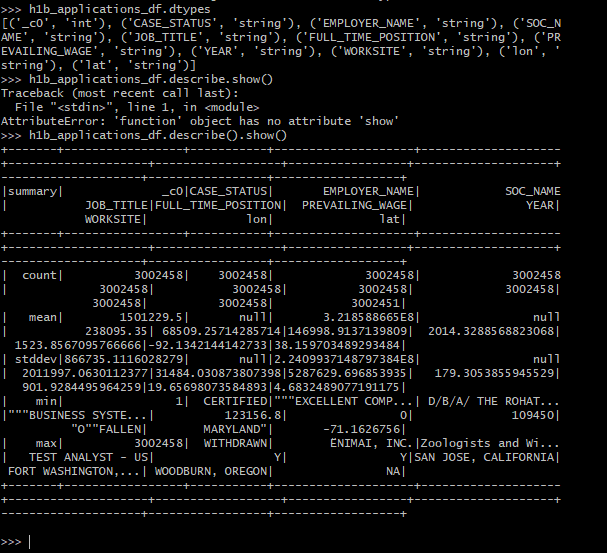
group by year,employer\_name

order by cnt;

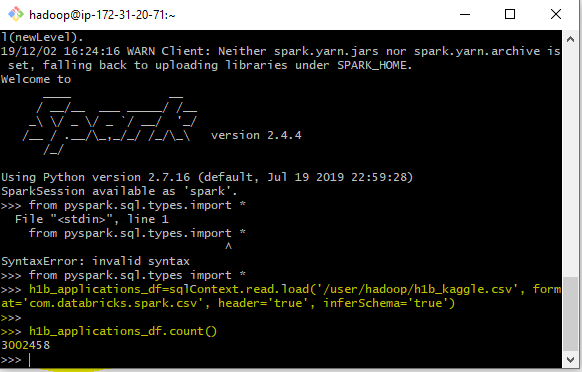


**Using SparkSQL:**

Describing data:



Loading data:

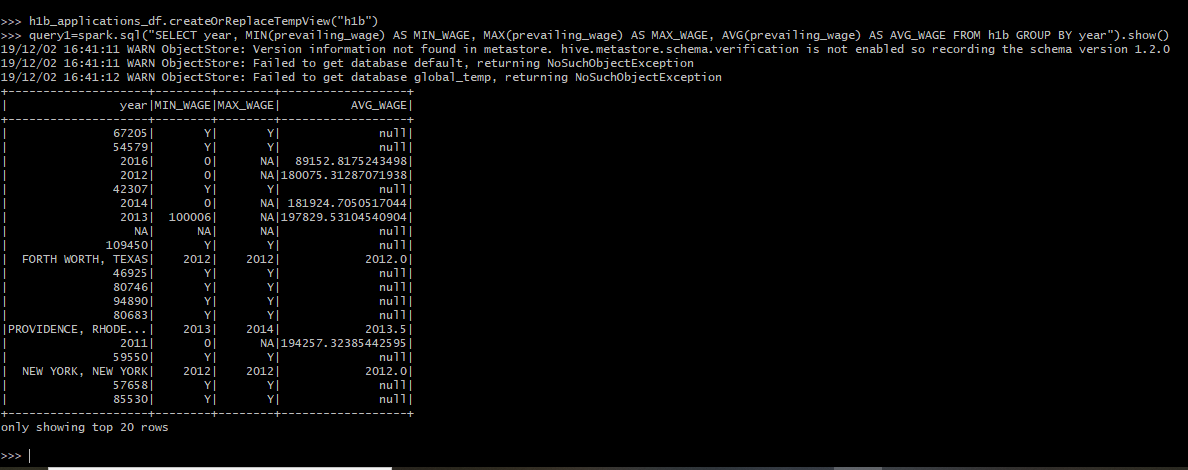


Creating table

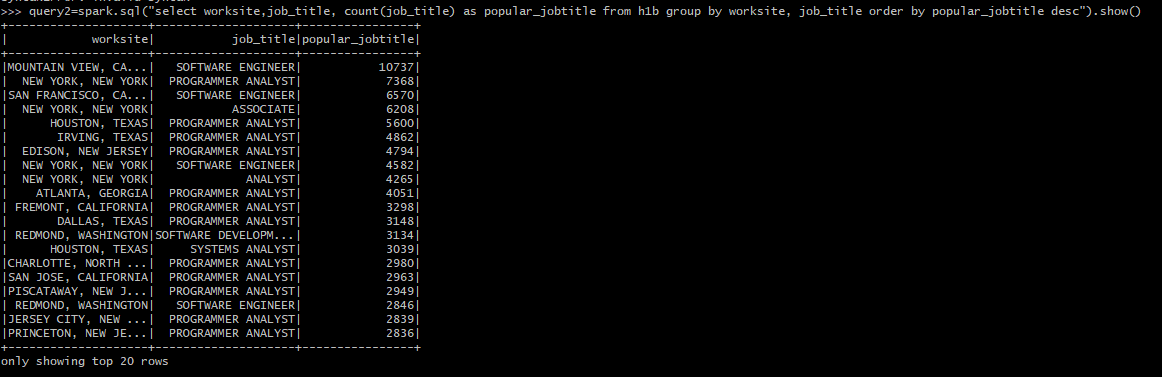


­­

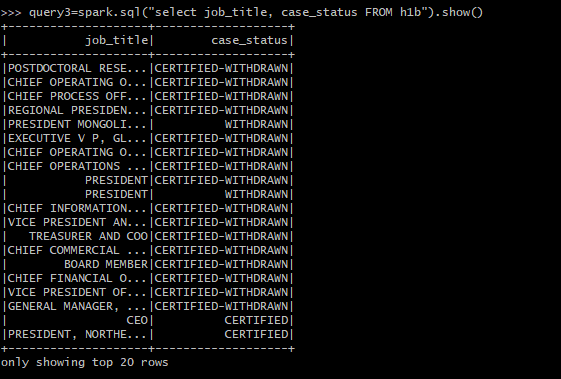
1.Maximum, Minimum, Average based on the wage for the years



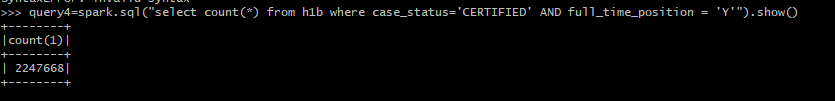
2.Which worksite has the popular job title?



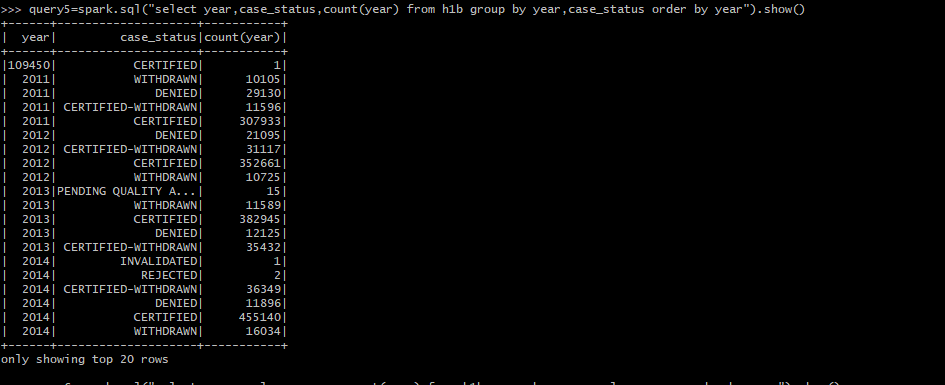
Query 3 Which job title has the case status as “Certified”, “Certified -withdrawn”, “Denied”, “Withdrawn”?



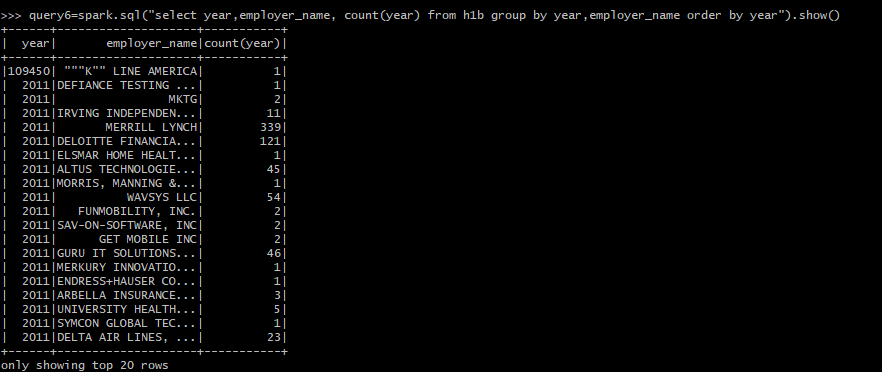
Query 4 How many H1-B filers are “Certified” having the fulltime status as “Yes”?



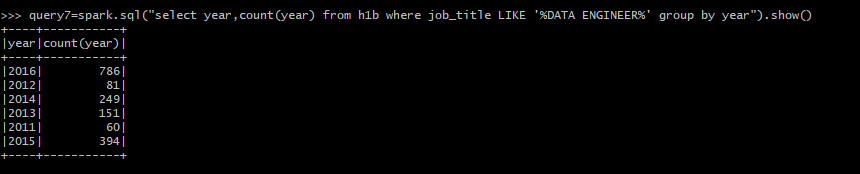
Query 5 What is the general trend of the employer’s H1-B status in the period 2011-16?



6. What is the general trend of the employer name with respect to the year?

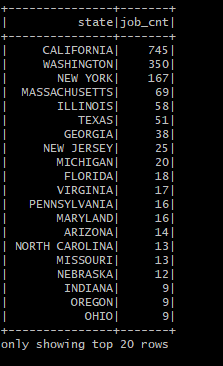


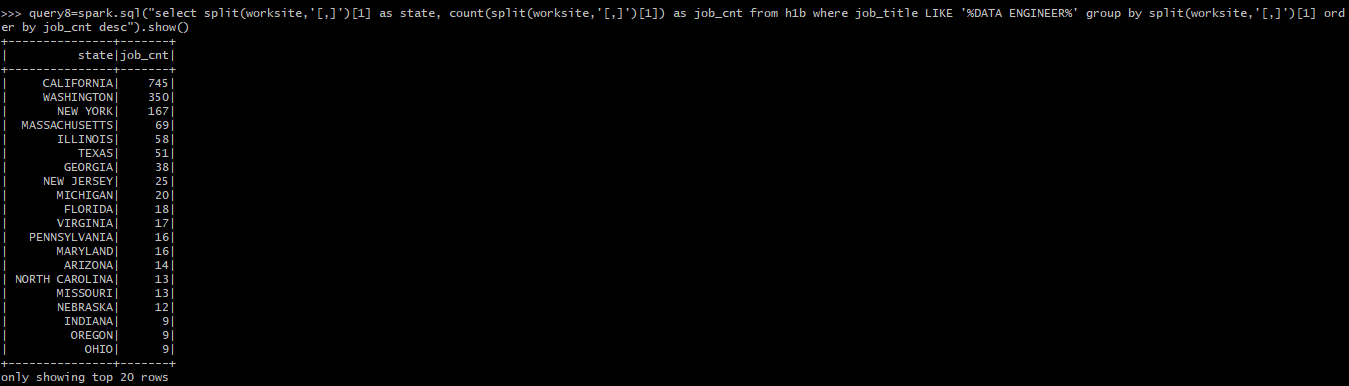
7.Is the number of petitions with Data Engineer job title increasing over time?



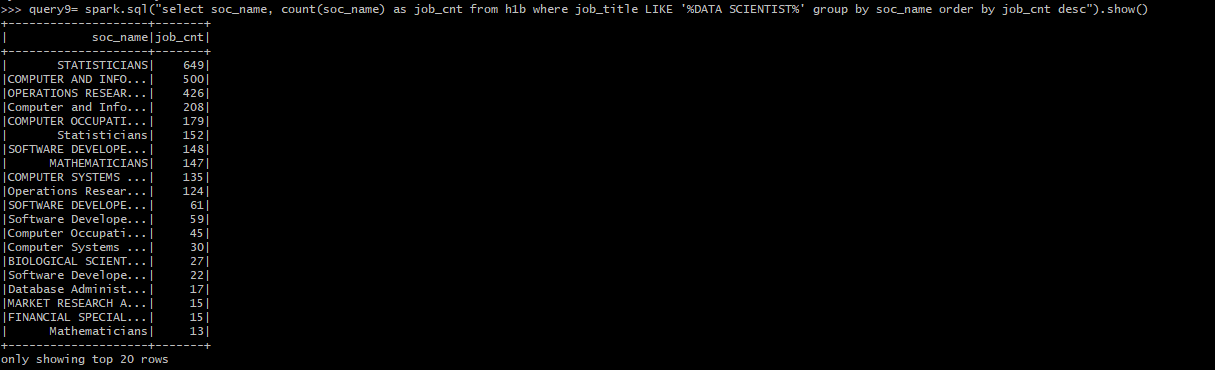
Query 8 Which part of the US has the most Data Engineer jobs?

query8=spark.sql("select split(worksite,'[,]')[1] as state, count(split(worksite,'[,]')[1]) as job\_cnt from h1b where job\_title LIKE '%DATA ENGINEER%' group by split(worksite,'[,]')[1] order by job\_cnt desc").show()

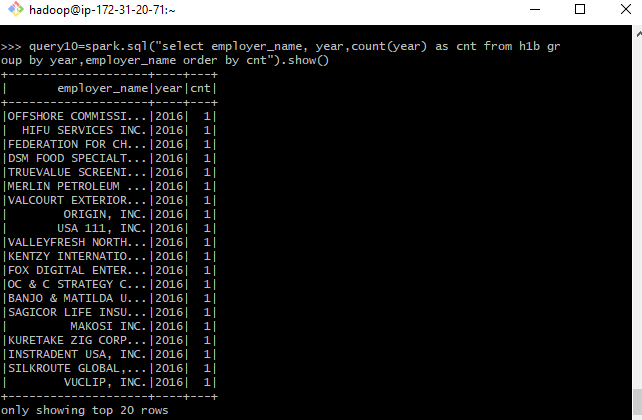




Query 9 Which industry has the greatest number of Data Scientist positions?

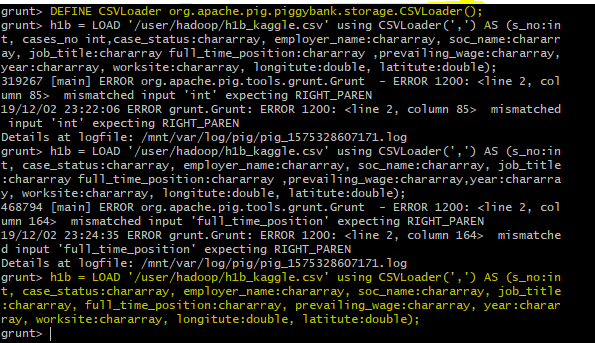


Query 10 Which employers file the most petitions each year?



**Using Pig**

Loading data

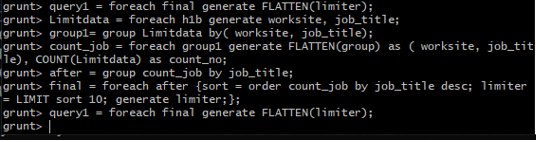


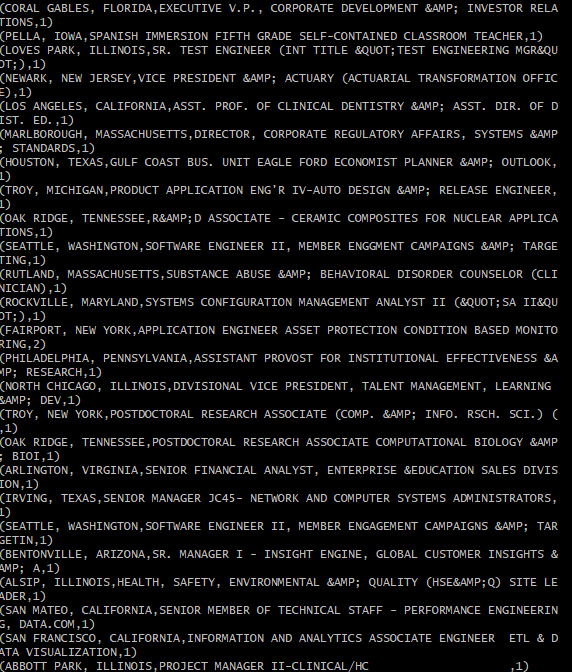
Showing data



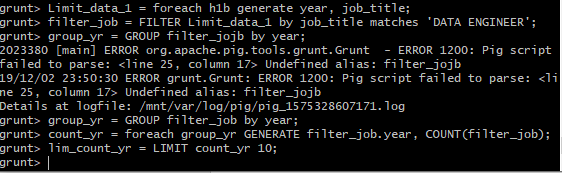


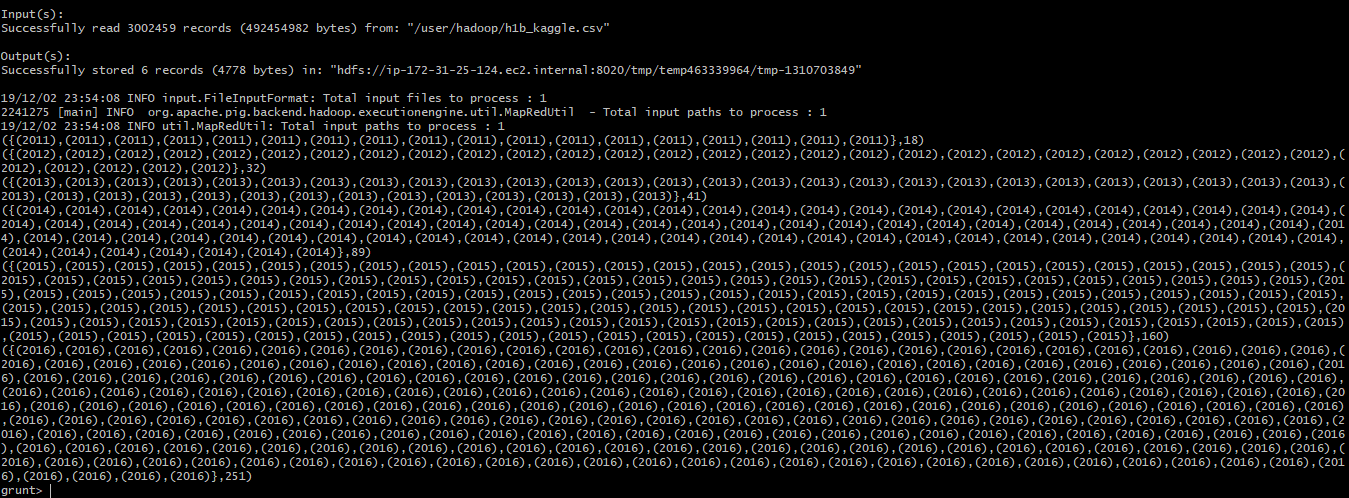
1. Which worksite has popular job title?



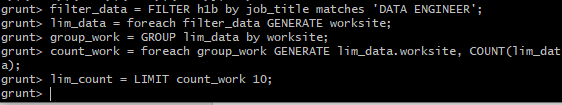


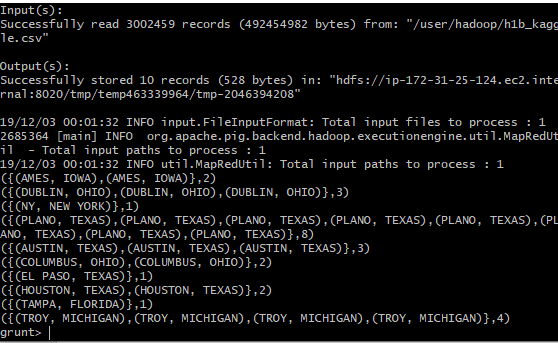
2.Is number of petitions with Data Engineer job title increasing over time?



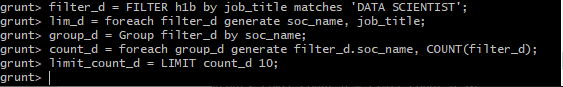


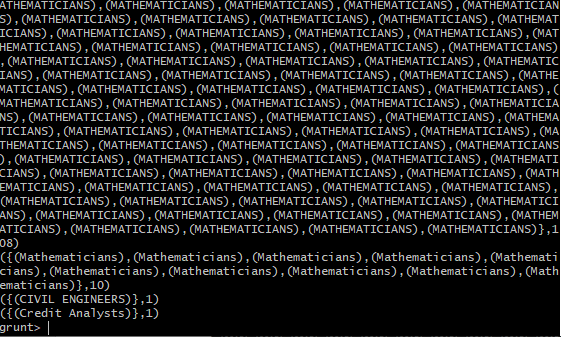
3.Which part of US has most data engineer jobs?

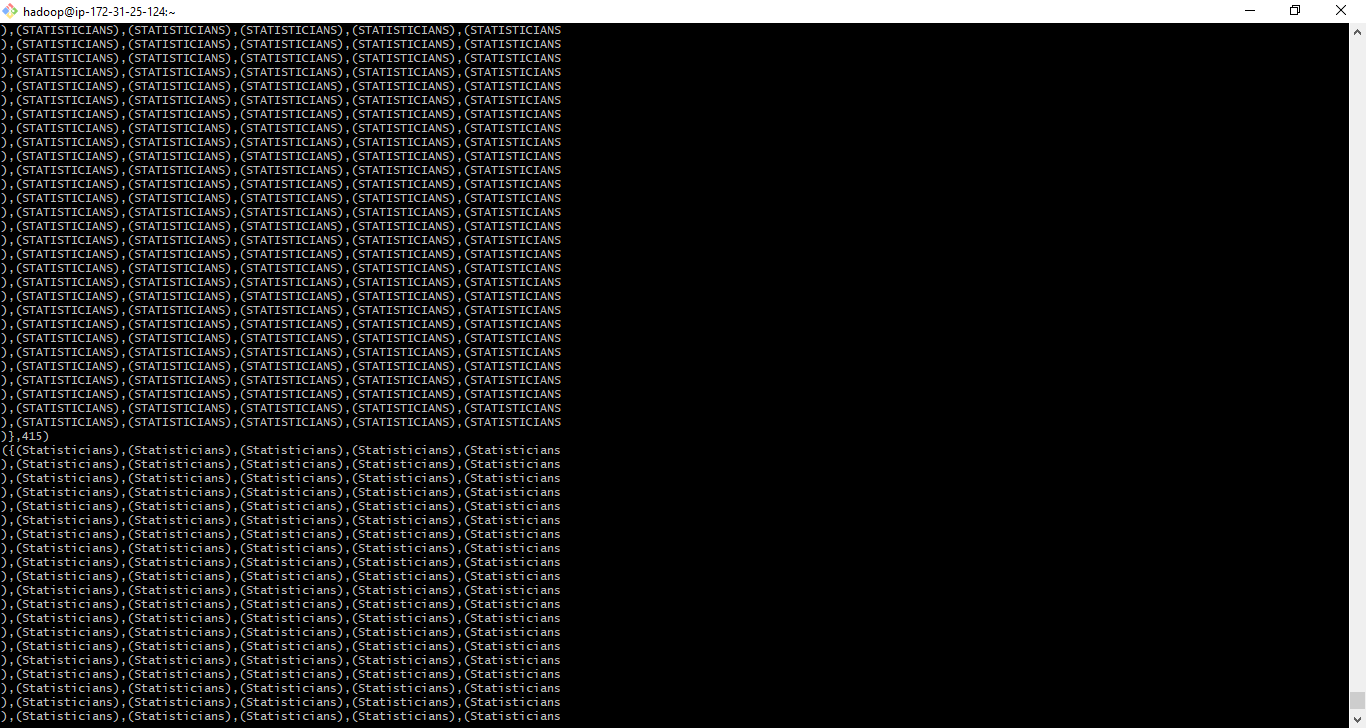




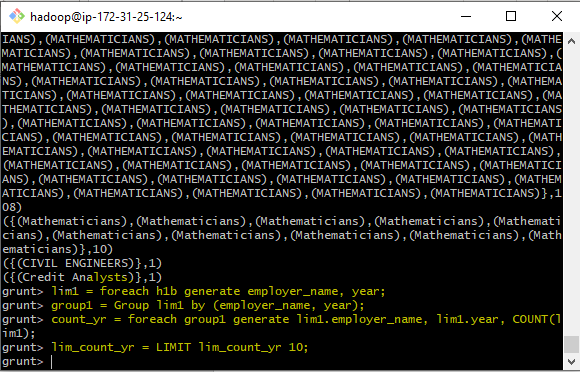
4. Which industry has greatest number of Data scientist position?

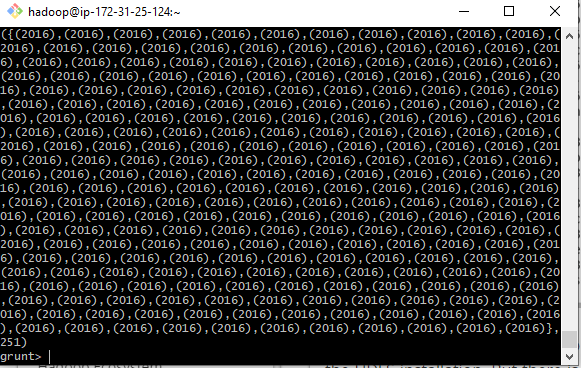


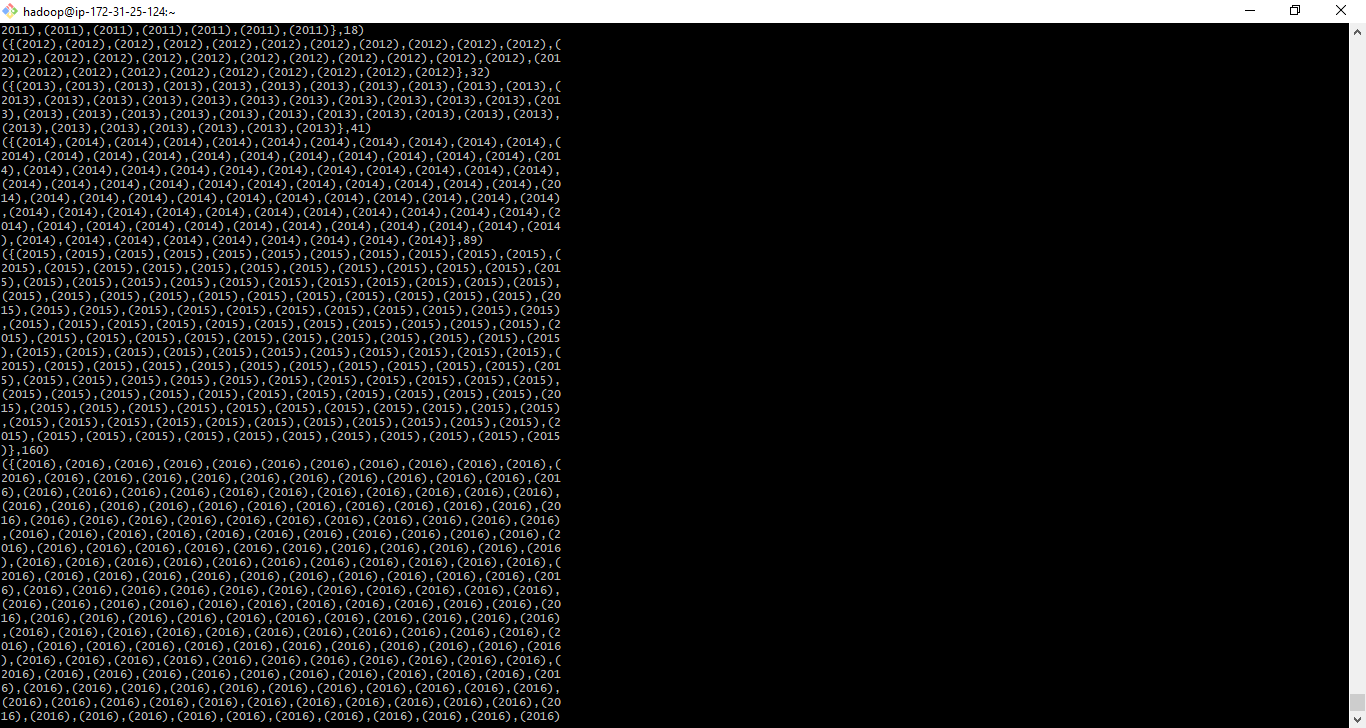




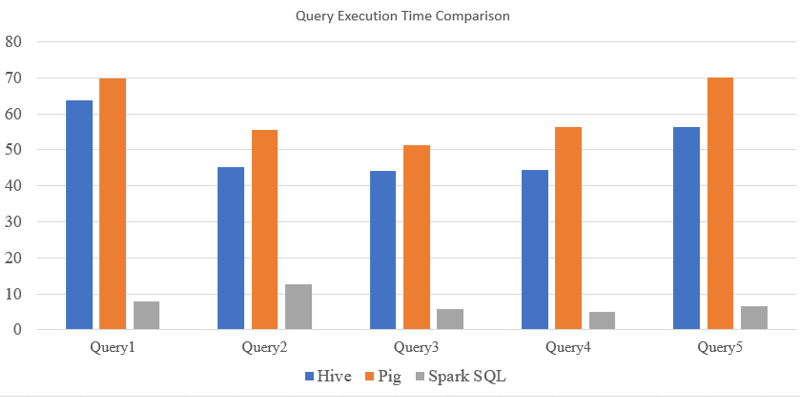
5.Which employer files the most petitions each year?







**Results:**



* The graph shows that SparkSQL has taken the least amount of time for query exeution
* Pig on the other hand takes more time to execute a query
* Hence, we can say that Apache Spark is very fast and can be used for large-scale data processing
* The latency provided with Hive is more than SparkSQL
* Hence for live streaming of data SparkSQL is used

**Conclusion:**

Thus, the project helps in understanding and analysing the job trends for foreigners in US. It shows that how many cases were certified with H1-B and which is the trending field for H1-B. The exploratory data analysis shows a graphical representation of the data in Tableau for providing an insight to the analysis. SparkSQL had the least time in query execution when compared to Hive and Pig. The query performance with Hive was better than Pig. Hence SparkSQL is preferred when using real time stream processing.

**References:**

1. <https://github.com/yugokato/Spark-and-Kafka_IoT-Data-Processing-and-Analytics#2-format-of-sensor-data>
2. <https://nycdatascience.com/blog/student-works/h-1b-visa-petitions-exploratory-data-analysis/>
3. <https://github.com/sharan-naribole/H1B_visa_eda>
4. <https://acadgild.com/blog/h1b-visa-applicants-data-analysis>