An Explorative Analysis of H1B Petitions data

INFO H516: Applied Cloud Computing for Data Intensive Sciences

Project Report

Under the guidance of
Sunandan Chakraborty

By
Mani Manjusha Kottala
Computer & Information Science, IUPUI
&
Yashwanth Reddy Kuruganti
Computer & Information Science, IUPUI

Table of Contents .....Page

I. Abstract

The world today is driven by data. IBM says that 90 percent of data in world is generated in the last two years [1]. On an average Facebook generates around 10TB of data daily. Around 7TB of data is generated by Twitter daily. To deal with such huge amount of data there should be a right technology to efficiently store and able to manipulate or provide a means to operate on such data. One such technology useful for Data Science is Cloud computing. This allows use to store data on a platform and seamlessly provides the user with an interface to act upon. We have made use of Spark learnt in this coursework to achieve our goal working on H1B dataset.

II. Introduction

What exactly is H1B visa? H1B visa is an employment based visa that authorizes an employee to work temporarily in any part of the US. First an employer must sponsor the employee to be eligible for H1B petition application. After LCA approves or certifies an employee, his/her petition only then he/she is eligible for H1B visa application. Undergraduate/Graduate students or professional students from streams like Science, Technology, Engineering and Math with a degree are eligible for H1B petition. This H1B allows an employee to legally work a certain amount of time which may be around 6 years. However, the initial period would be a time span of 3 years [4]. Due to its dual intent nature, employee's family will also be eligible to stay in US. In the following sections problem description, dataset description, implementation, results, comparison are discussed.

III. Problem Description

Information Analytics causes us to get some extraordinary bits of knowledge from the enormous lumps of information accessible. Each application in the present time is driven by information, and having the capacity to get the bits of knowledge from this information makes our lives significantly more easier and accomplish awesome outcomes. Following the past Presidential Elections, there has been a great deal of disarray in the United States about the H - 1B visas and talented workers, this influenced us to consider investigating and dissecting the H - 1B petitions so that we as International Students can extract insights which are justified regardless of the time we put in.

We will be able to discover bits of knowledge like, which part of the United States has the biggest number of Data Scientists or which Part has minimal number of Software Engineers, which industry has the most information researcher petitions for H - 1B visas, we will likewise be discovering a few bits of knowledge like which work part has the most astounding achievement rate in H - 1B visas.

IV. Dataset Description

Our interest made us to search Internet extensively to find the right dataset containing information about H1B. Finally, we have found H1B petitions dataset on Kaggle [2] sizing around

600MB containing 3.1 million records for years 2011 to 2016. This dataset is resulted after removing unnecessary attributes from the raw dataset. Also, status of each H1B petition after Labor Condition Application [3] processing is stored in Case status field of the dataset.

a. Case Status: This field contains four values: Certified, Certified-Withdrawn, Withdrawn and Denied which can be explained as follows:
    - Certified: This means that the candidate is now eligible for H1B application approval
    - Certified-Withdrawn: Candidate is eligible for approval, but the employer withdraws the petition due to various reasons (say termination of employee).
    - Withdrawn: Applied for LCA processing, but withdraws before certifying.
    - Denied: Not eligible for H1B application
b. Employer Name: Denotes name of the employer who sponsors for a petition
c. Soc Name: Occupational name or department of the employee
d. Job Title: Title or position of the job offered by the employer
e. Full Time Position: Determines if the job is a full time based or contract based. Contains Y or N values
f. Prevailing Wage (In USD): Minimum salary of the employee offered by the employer to be eligible for petition approval.
g. Year: Year in which employer sponsors H1B petition for its employee.
h. Worksite: City and state of the employer located at.
i. lon: longitude co-ordinate of the employer location
j. lat: latitude co-ordinate of the employer location

V. Implementation

In this section, steps involving data preprocessing, processing dataset in Spark and sample analysis are discussed.

A. Data Preprocessing

First step before analyzing or querying on the dataset, one should make sure that the data is free from misleading values. Data cleaning is the process in which corrupt or inconsistent records are identified and removed without effecting the originality of data. As far as H1B petition dataset is considered, there are few null records with only case status, but no other attributes defined. We have successfully filtered the dataset by removing all these blank records.

Second step is to select the attributes that would be more interesting to analyze the dataset, which can be done by feature selection. Out of all the columns in the raw dataset, we have removed few columns that are not useful for data analysis and extracted the most important columns. This resulted dataset is now used for further analysis.

B. Spark Processing

Above resulted dataset which contains 3.1 million records in csv format is now uploaded to Aspen cluster. Spark supports structured data processing along with optimization through interfaces like internal structure of data and its computation. Spark Session is created through which the file is read and processed. After this step a data frame is created and can be further used for querying through Spark SQL. Before executing SQL queries on the dataset, a temporary view or table is created. Each result returned from this query is treated as a data frame on which several actions can be performed. All these temporary views are session scoped.

Out of all the analysis we have done, below is the list of few important queries we have scripted:

- ✓ Certified petitions per year/ Range of petitions over years
- ✓ Retrieve Average salary per year to get a petition certified
- ✓ Retrieve average salary per year for all the years and for each state for all the years
- ✓ Top employers sponsoring highest no of petitions
- ✓ For each employer name get all soc_names(departments/portfolios/streams) in descending order for all the years
- ✓ Which state has highest salary/average salary
- ✓ Which state has highest no of jobs
- ✓ Top companies based on proportion of success to total petitions
- ✓ Minimum salary to be eligible for certified petition
- ✓ Getting top 10 employers per year and their respective sum of each case status types
- ✓ For a specific job role identify the top 5 companies that offer H1B based on full time position or part time/contract

How complex or useful information does the query return, every detail can only be better understood when visualized. So further section demonstrates few visualizations based on the raw output resulted from each Spark query.

VI. Results Explained

Each spark SQL statement returns a data frame which can be displayed on the terminal or stored in an external file. If the resulted output is of less details, it can be understood easily by a user. But when large amount of information is displayed in raw format or spread over excel sheets, user can extract very less details which makes the entire analysis useless. So, data

visualization helps a human eye to better process and understand the variation and recognize similarities. In the following sub-sections several visualizations have been discussed which were plotted using Tableau. For better visualization and live interaction, please refer this link: https://cs.iupui.edu/~mkottala/H1B%20analysis.html

a. Petition per Year

This visualization is plotted considering each Year from 2011 to 2016 against total no of petitions sponsored for respective year. No of petitions sponsored keeps on increasing year by year.

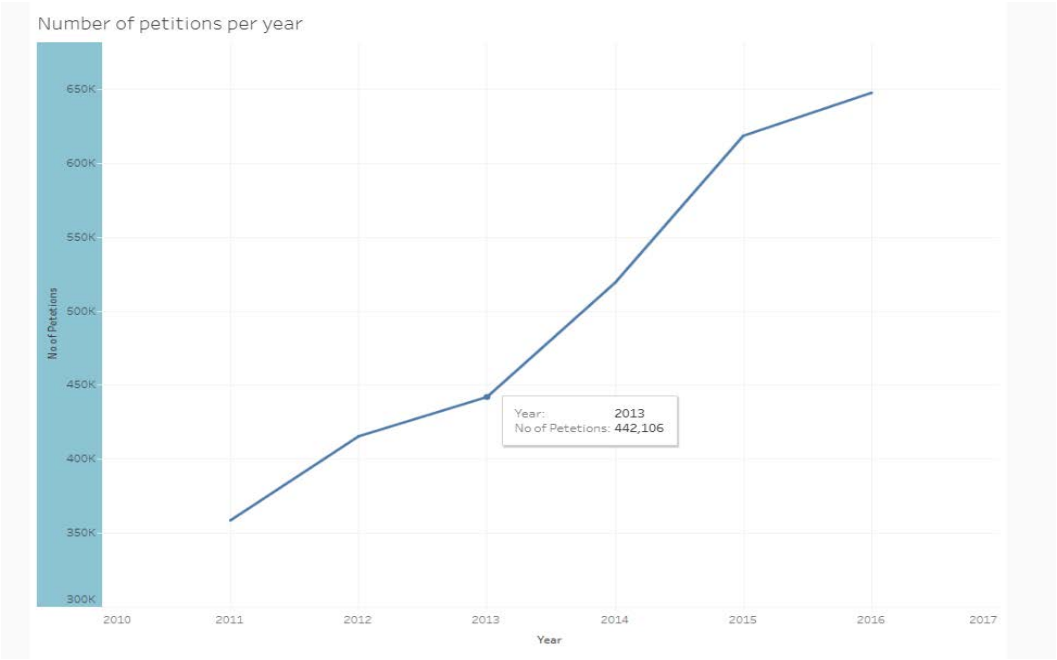
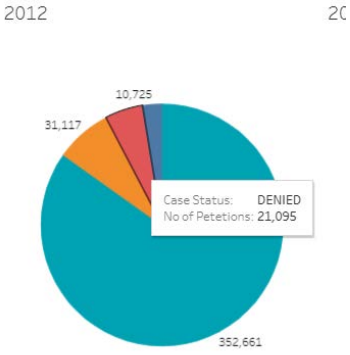


Fig 1. Total no of petitions per year

But what about case status distribution per year over all the years. Next visualization deals with this problem.

b. Case status distribution

Each year, all the petitions sponsored by the employers have a case status id and case status assigned by LCA. After looking at the trend in no of petitions per year, we thought of analysing no of petitions certified or denied over the years through multiple pie charts. Below is a sample screenshot showing share of each case status in the year 2012. On hover, each pie show the no of petitions fell under respective case status each year. 6 pie charts are plotted each denoting case status for the years from 2011 to 2016.



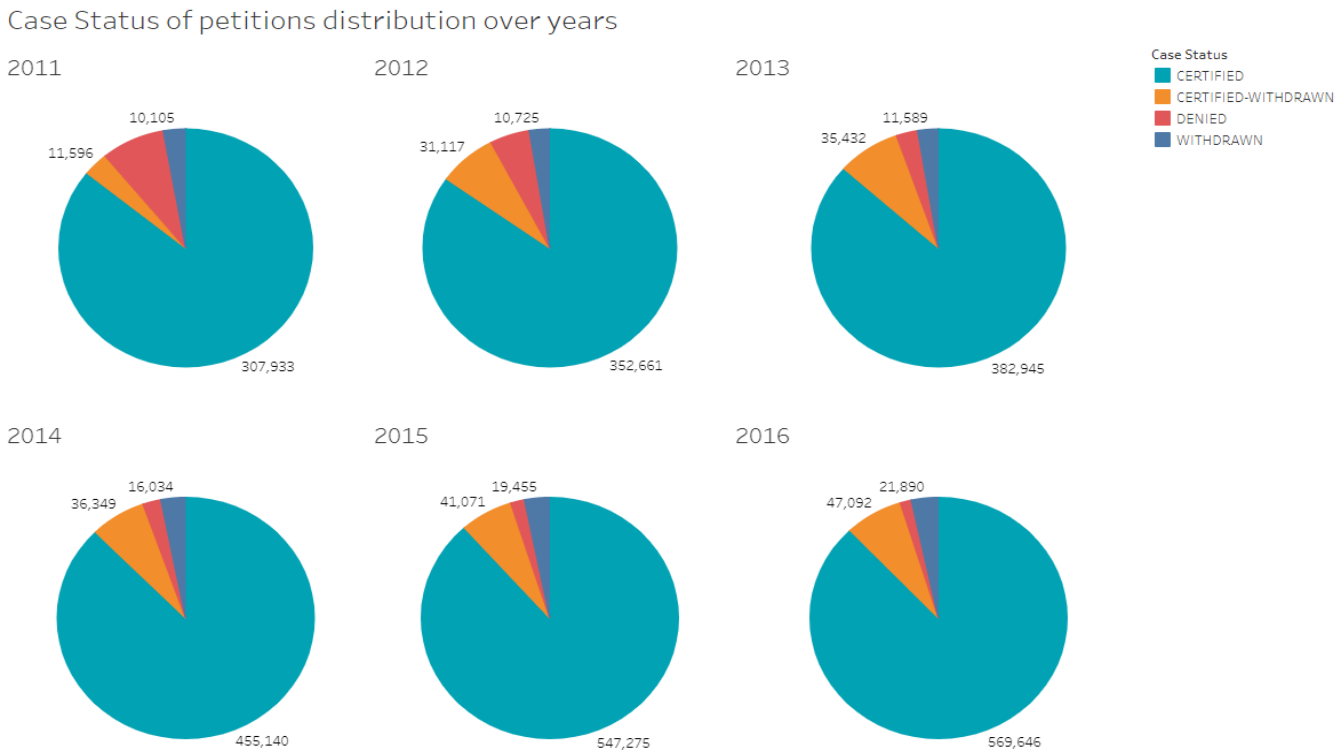


Fig 2. Case status distribution over years

As the no of petitions keeps on increasing year by year, no of certified petitions also kept on increasing. One interesting interpretation from the above chart is that, no of petitions getting denied keeps on reducing over the years.

c. Average salary per state & No of jobs per state

For this visualization, aggregation is performed over states and calculated average salary for all the states in US. On the other side data is also aggregated year wise. So an user can interact with the graph by selecting an option from the list of years provided on the right side. On selection Fig 3 showing average salary and Fig 4 showing no of jobs are retrieved on screen by utilizing the results from the Spark query.

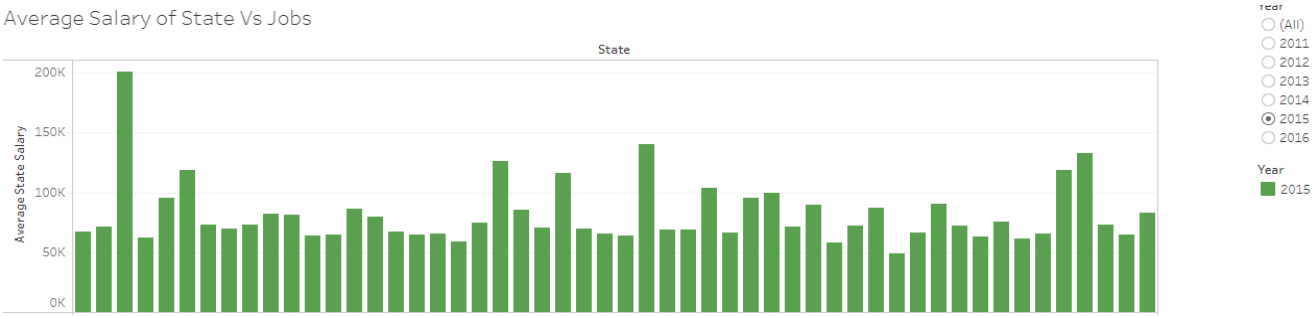


Fig 3. Average salary of all the states in 2015

From the above graph one can observe that, Arizona has the highest average salary followed by Nebraska, Washington etc., in the year 2015 compared to all other states.

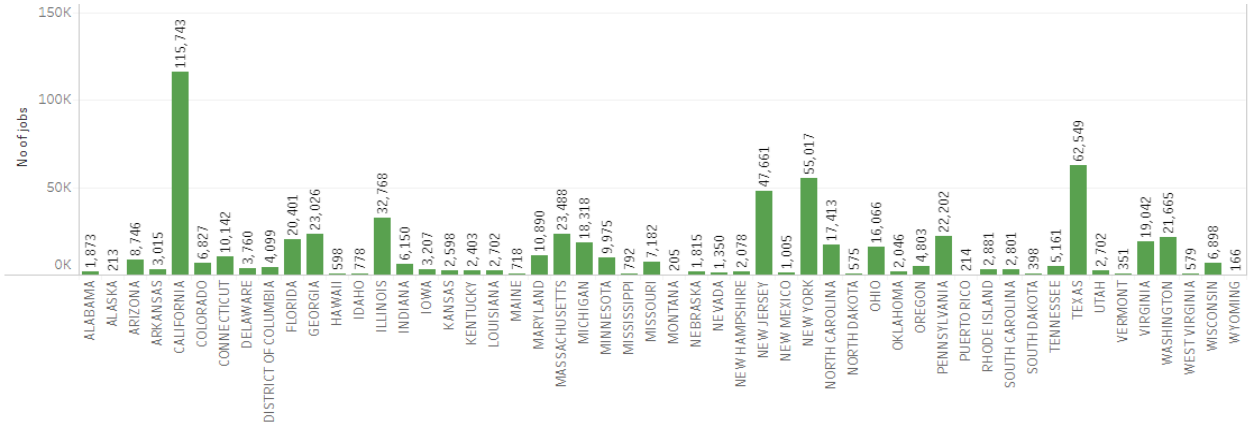


Fig 4. No of jobs for all states in 2015

On hovering the bar graph, information like Year, State name and its respective average salary, no of jobs are displayed in the tool tip. Fig 4 shows that, highest no of jobs is offered by California followed by Texas, New York etc., which contrasts with the highest average salary.

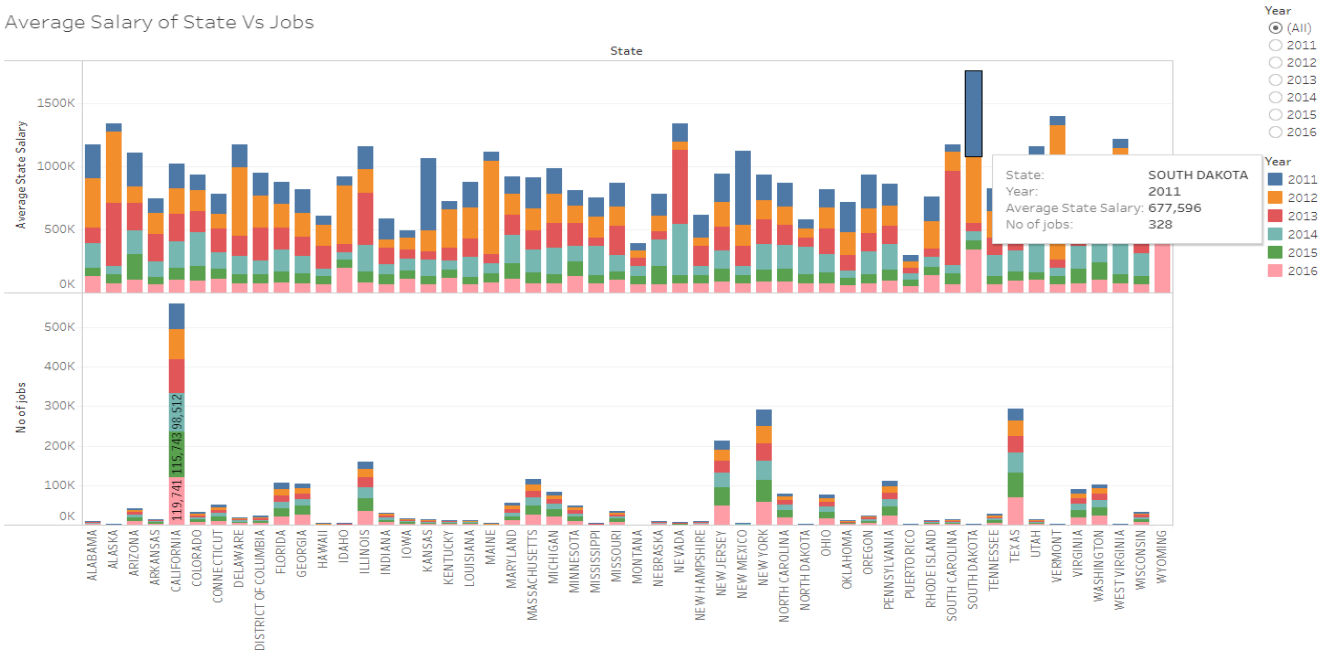


Fig 5. Average salary and No of jobs for all states and Years

Fig 5 denotes average salary distribution for all the states when selected for all years. Each color on the bar plot denotes each year from 2011 to 2016. Among all Wyoming has the least no of jobs offered, but with a good average salary. Overall, South Dakota has the highest average salary over years and California has the highest no of jobs sponsored.

d. Top five Employers for specific job role

Below grouped bar plot visualizes multiple attirbutes like No of H1B petitions sponsored by an employer and its corresponding job title or employer position along with full time status.

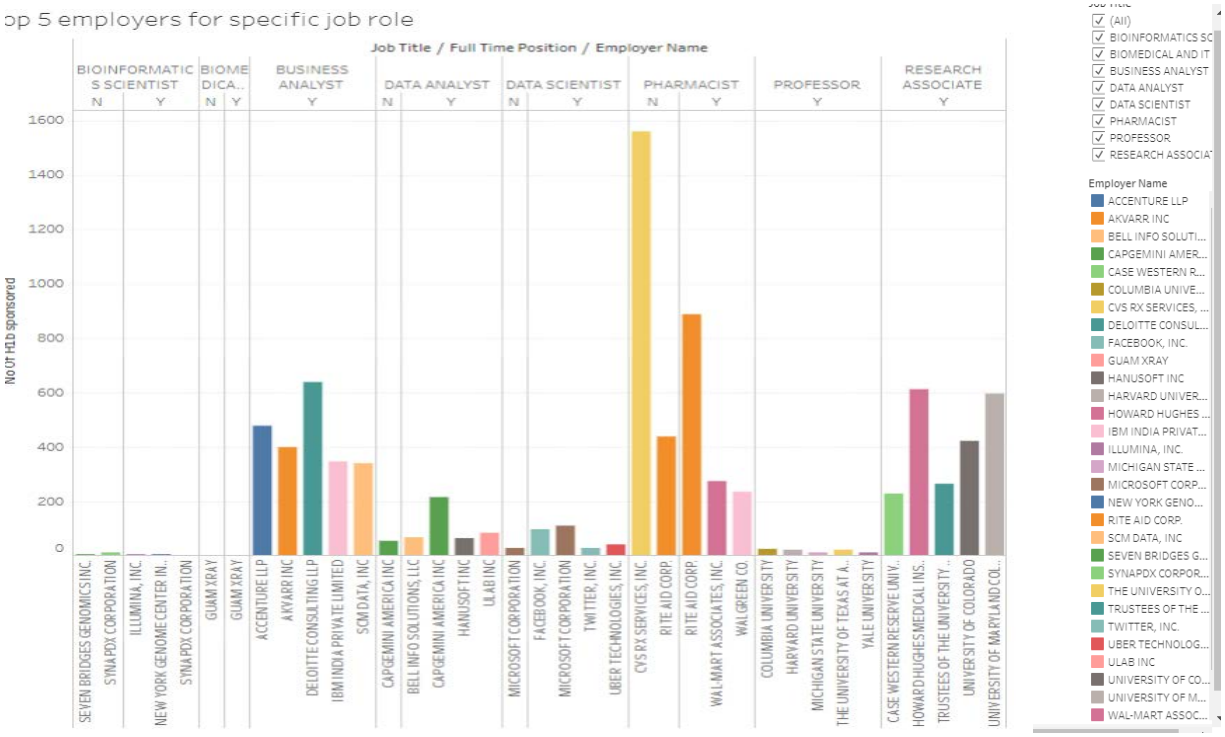


Fig 6. Top 5 employers for specific job roles

To make it more specific for our class, we have taken few job roles like Bioinformatics scientist, Business analyst, Data scientist, Data Analyst, Pharmacist, Research associate and Web Developer. A user can interact with the graph either by selecting a specific job role from the checkboxes available on the right panel. Based on his/her selection, top five employers along with full time position status is displayed as above. In addition, a user can click only on the Employer name on the right, which displays only the respective job title of selected employer.

e. Top Employers based on Success Ratio

For few employee, the no of petitions sponsored would be relative to the no of petitions getting certified. But for few employers this may not be the case. Very less percent of petitions out of applied petitions get sanctioned. This variation can be observed from the below visualization which represents No of petitions on the top Percent of certified petitions on the other half. To interact with the graph, user can select a company or employer name from the drop-down list available on the right side. For example, if Wipro is selected in the drop-down, two charts are displayed on the screen. Over the years, one can observe that No of petitions kept on increasing but the success ratio started decreasing.

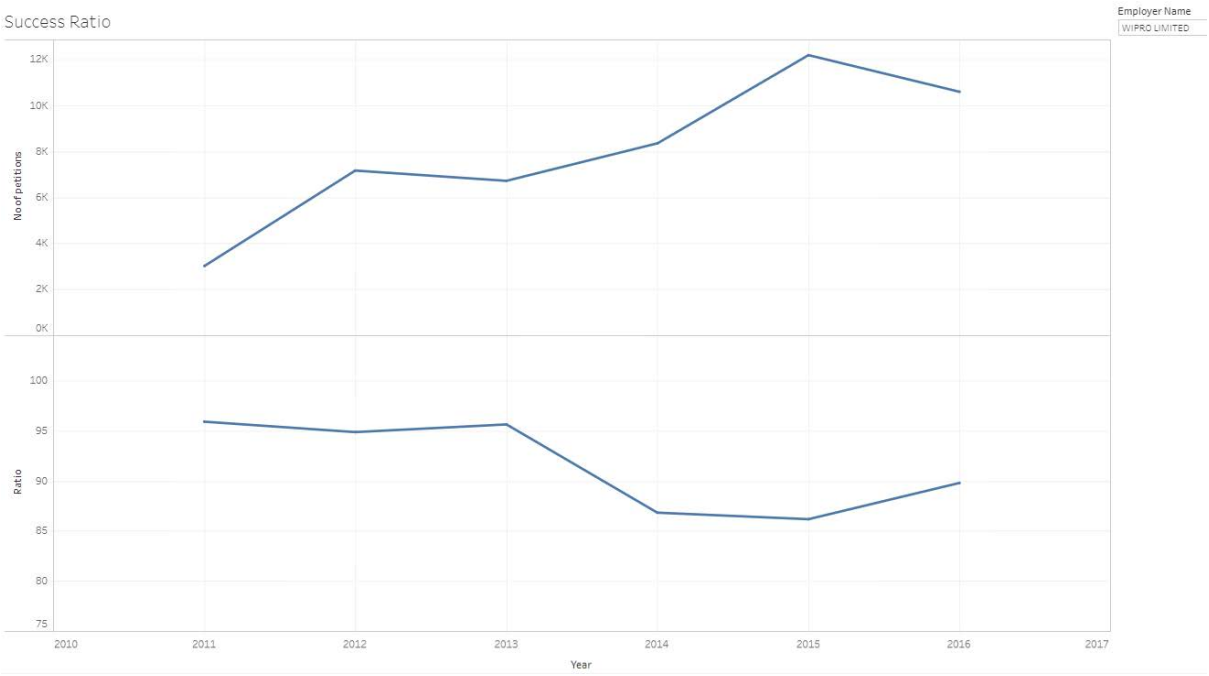


Fig 7. Top Employers along with Success Rate

f. Top 10 Companies per Year

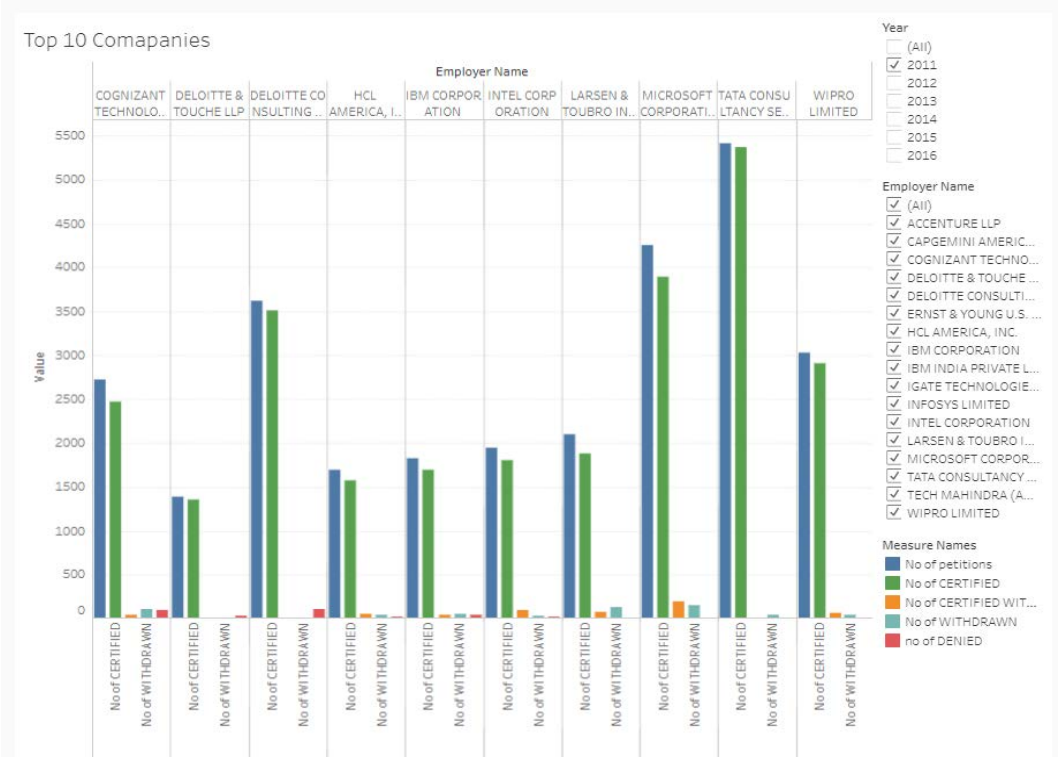


Fig 8. Top 10 Employer sponsoring high petitions per each case status

Below visualization lists top ten employers per each year along with No of petitions based on all the case status types. User can interact with this graph either by selecting a year from the checkbox list or by selecting an employer name from the listed employers. And based on case status, user can understand the variation between no of petitions certified or withdrawn or denied. When an year is selected all the employer names are displayed on the top and count of all the petitions based on case status types.

VII. Comparison

In this section, a brief comparison of cloud computing technology like Spark against traditional non-cloud computing technology is done. Access SQL is good at handling small to medium sized datasets. But when the size of data keeps on increasing, each query takes more than reasonable amount of time to output the result. This disadvantage can be superseded by making use of Spark SQL which outputs the result faster than Access. Below is the table which show the comparison of time taken by four queries across Access and Spark.

Table 1: Time comparison for Access SQL and Spark SQL

Technology	Q1	Q2	Q3	Q4
AccessSQL	2.65	39.53	92.57	110.23
PySpark	2.77	3.27	16.32	17.52

Q1: Count total no of records
Q2: Avg salary for all years per employer
Q3: Top employers per specific job roles
Q4: Case status wise analysis

Query Q1 retrieves total count of records using both Access SQL and Spark SQL in equal amount time. Considering Query Q2 which calculates Average salary per each employer for all the years, Access SQL ran for 40seconds which was 10 times more than Spark SQL. As the complexity of query (multiple joins, case statements etc.,) increases, Access SQL takes lot more time to result the expected output. This advantage of Spark helped us to analyze efficiently and extract useful data from the dataset with minimum time.

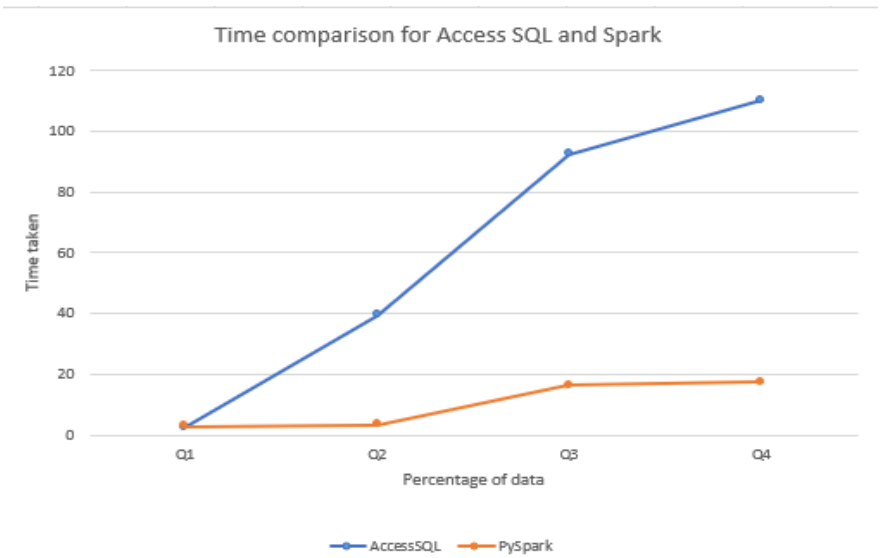


Fig 9. Line chart denoting time taken by Access and Spark

VIII. Future scope

Along with all the above analysis, as an additional step we have tried implementing a logistic regression based prediction model. This model takes an employer's previous years success ratio of h1b petition approval and predicts the outcome as how high the probability for a petition to be approved. So, this can be considered as a future development or an extension to the current analysis.

IX. Conclusion

To conclude, we have learnt how a cloud technology can efficiently deal with a huge dataset and how to process a dataset using Spark. Also, this helped us to analyze H1B petitions data in a better way and extract lot of useful insights, certified petitions pattern, top employers, minimum wage, etc.,

X. Team Contribution

This project involved numerous steps right from finding an appropriate dataset till documenting the report. But we always follow this line *"Teamwork divides the task and multiplies the success"* [5].

All the tasks like Dataset Search, Dataset Preprocessing, Spark SQL, Visualization, Presentation and Documentation were equally shared by both of us.

References

[1] https://www.sciencedaily.com/releases/2013/05/130522085217.htm

[2] https://en.wikipedia.org/wiki/Labor_Condition_Application

[3] https://www.kaggle.com

[4] http://www.h1base.com/visa/work/H1B%20Visa%20Overview/ref/1164/

[5] http://www.inspirationalspark.com/teamwork-quotes.html