CIS 528-01 Term Paper: H-1B Visa Application Data Analysis

Spriha Batar (305053299)

Premdeep Parasa (305094210)

Sumani Pidikiti (305065103)

Hasitha Reddy Pittu (305084421)

(Group E)

**Abstract**

The H-1B is a non-immigrant visa in the United States under the Immigration and National Act that allows U.S. employers to temporarily employ foreign workers in specialty occupations. H-1B work authorization is strictly limited to employment by the sponsoring employer.

In our analysis, the first step is to take some sample data (around 600MB) from an online repository in .csv format. Next, we create a Cluster using Databricks and upload the dataset into the cluster. Using simple Spark SQL queries we analyze the data for different types of status of the applications, wages and dates of employment etc. The main advantage of using Spark SQL commands is that it has a huge standard library also with python and with many inbuilt functions.

**1. Introduction**

* 1. **Background**

The regulations define H1-B as a "specialty occupation" where highly specialized knowledge in a field of human endeavor including but not limited to areas such as Sciences, technology, Engineering, Mathematics etc. and requiring the attainment of a bachelor's degree or its equivalent as a minimum. If a foreign worker in H-1B status quits or is dismissed from the sponsoring employer, the worker must either apply for or be granted a change of status to another non-immigrant status, find another employer, or leave the U.S.

**1.2 Overview**

Our analysis on the project is primarily focused on the H-1B Visa applications which is primarily used by immigrants from other countries to obtain a work permit in the United States of America. We used the data set consisting of such applications from the year 2011 to 2014. We have analyzed the patterns of all the applications recorded using Spark SQL on Databricks. We queried the database to check the number of applications that have been certified, denied, withdrawn etc. and to check the factors that help in analyzing this dataset.

**2. Technical Specifications**

**System Requirements on Databricks**

* **Cluster Type** – Hadoop 1
* **Memory** – 6 GB
* **Version** – Spark 1.6.1
* **Number of Nodes** – 1
* **Drivers** – 0.88 Cores

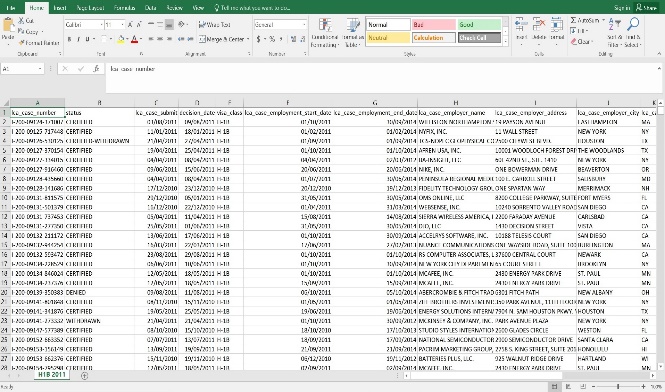
**3. Project Description**

* 1. **Project Scope**

Data analysis is done on the applications submitted for H-1B Visa for the years 2011 to 2014. We have done analysis on the types of statuses of these applications in these years. In accordance with that, we continued analyzing the results of those results. We will visualize the results of these analysis through Excel using pivot table, maps, pie charts and bar charts.

* 1. **Dataset**

We acquired the dataset from an online repository**.** The data covers the inspections from 2011 to 2014. Format of the data is Comma Separated Value (.csv) and size of dataset is around 600 MB with more than 1.7 Billion records. The data primarily consists of case number, status of the applications, submission and decision dates, employer and employee addresses, employment start and end dates etc.



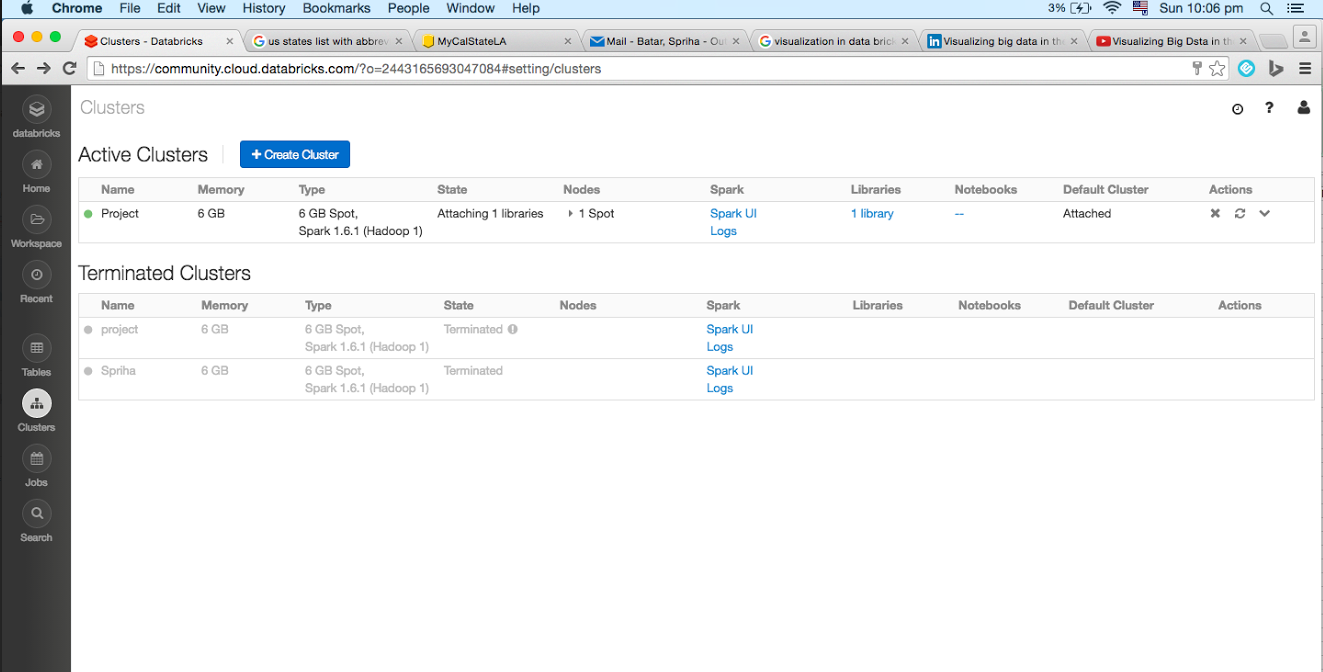
* 1. **Storage Deployment**

The storage is deployed using a Databricks cluster which is built on Spark.

* 1. **Databricks Cluster Deployment**

This is deployed from the Databricks portal, configured to have a Spark version 1.6.1 and the type being Hadoop.

Its resources are comprised of one nodes with 0.88 cores as its drivers and a memory of 6 GB.



* 1. **Data Upload**

The dataset is uploaded on to the cluster by creating a new table on the Databricks platform. As the data is around 600 MB, it might take some time for the data to upload. Data uploaded in the system is in Comma Separated Value (CSV) format.

* 1. **Work Flow**



Data is downloaded in .csv format. Databricks aims to help with cloud-based big data processing. Cluster is created on Databricks and the data is uploaded. Data is put into a table and using a notebook, queries were executed. Visualization is done on the result of the queries.

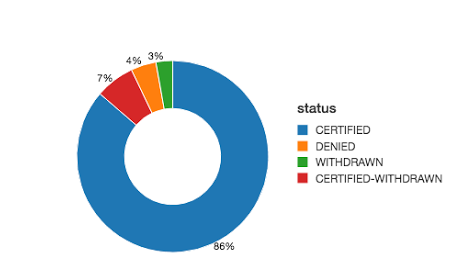
* 1. **Querying the database and Visualization**

The Spark SQL queries are executed on a notebook created in the Databricks platform.

**Table Creation and Data Insertion:** A table is created in the Databricks platform where we can drag and drop our data in .csv format from our source system.

**Select Queries:** *SELECT*queries are used to check the data from the table and validate it to get desired output from the given dataset.

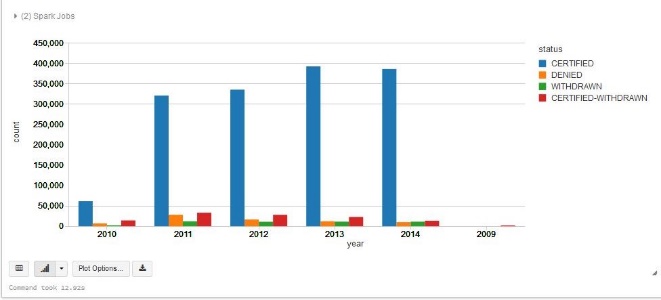
|  |
| --- |
| **select status, count(status) as Status\_count from project1 group by status;** |



The chart above shows the percentage of the different kinds of statuses of the H-1B Applications.The majority was Certified applications (86%) and the least was withdrwan applications(3%)

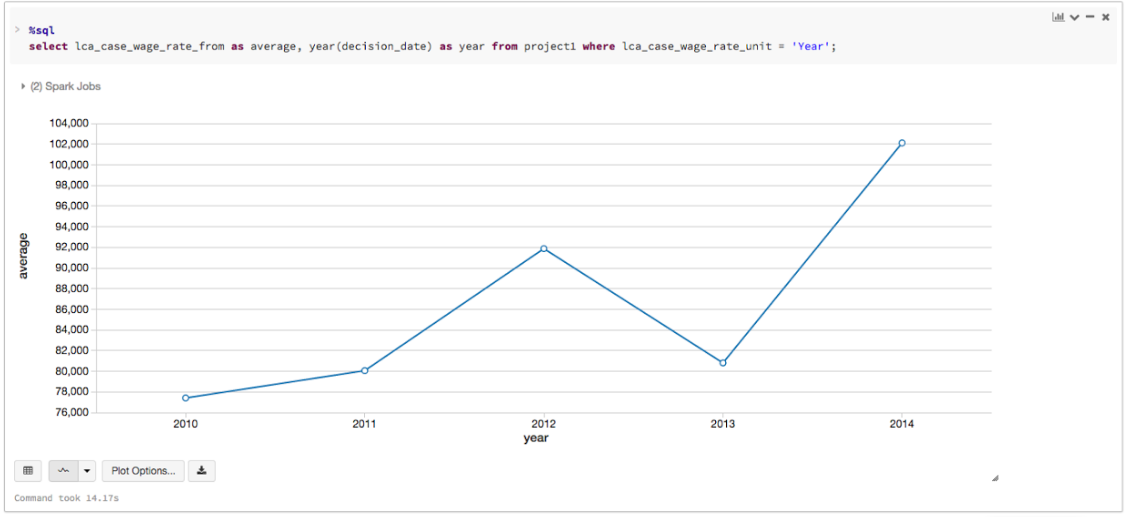
|  |
| --- |
| **select status, year(lca\_case\_submit) as year, count(status) as count from project1 group by status, year(lca\_case\_submit);** |

The above query shows the detailed status of applications over every year. Below is the chart for the same. Year 2013 was recorded with the most number of certified applications.



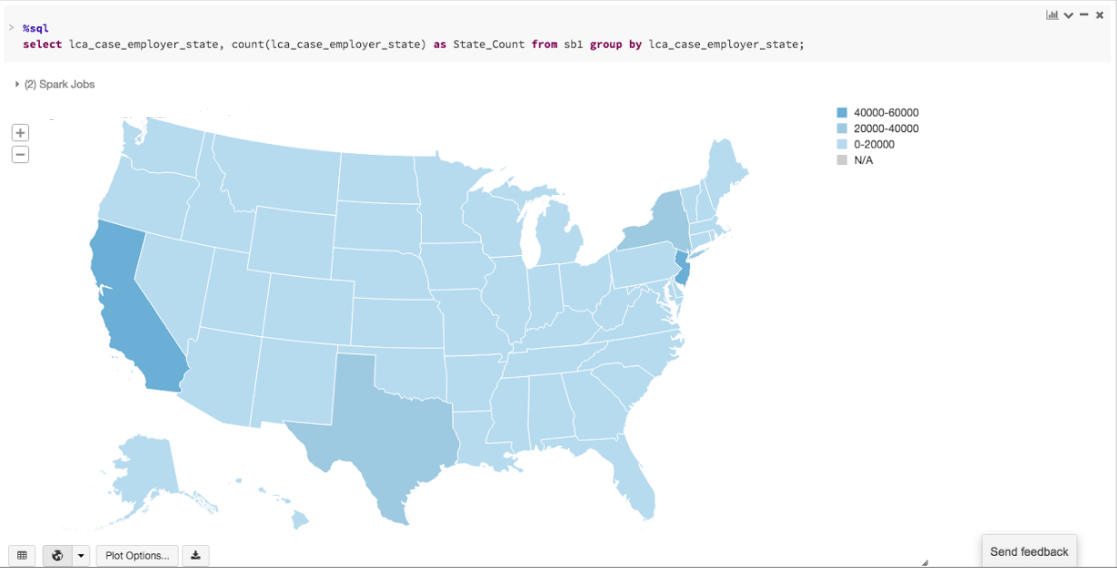
|  |
| --- |
| **select lca\_case\_wage\_rate\_from as average, year(decision\_date) as year from project1 where lca\_case\_wage\_rate\_unit = ‘Year’;** |

The above query is used to determine the average starting wage for the employees over the years. The highest was in 2014 and the least being in 2010. There is a significant rise in the wages of the applicants



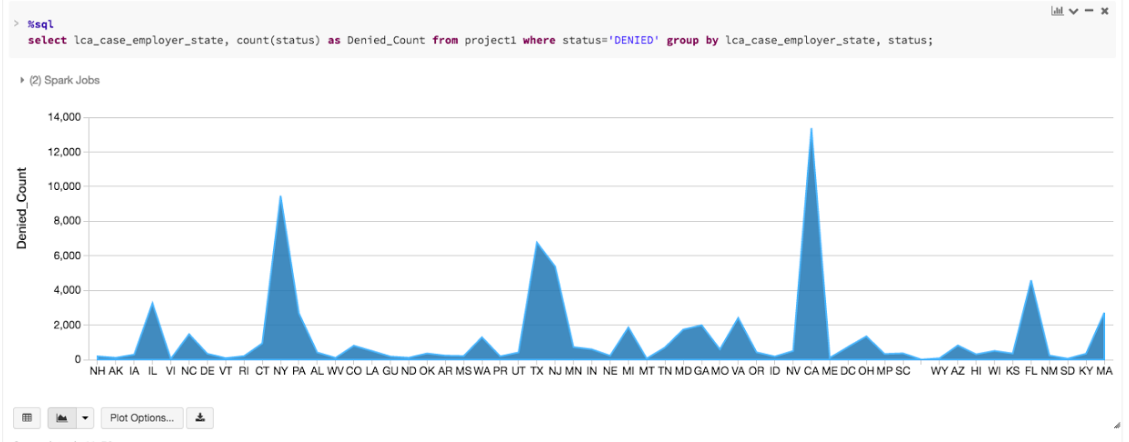
|  |
| --- |
| **select lca\_case\_employer\_state, count (lca\_case\_employer\_state) as State\_Count from project1 group by lca\_case\_employer\_state;** |

This query is used to generate the data which shows the count of employers in different states across the United States. The geographical map below shows the intensity of the applications from each state.



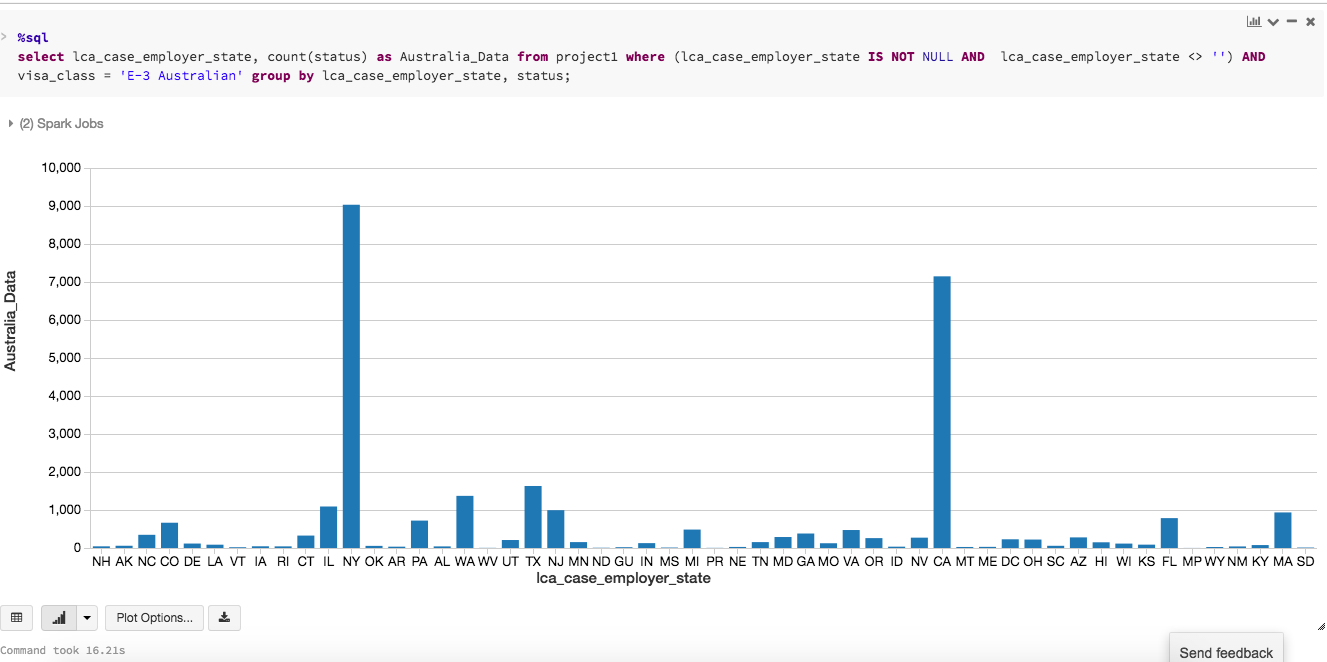
|  |
| --- |
| **select lca\_case\_employer\_state, count(status) as Denied\_Count from project1 where status=‘DENIED’ group by lca\_case\_employer\_state, status;** |

This query is used to determine the count of employers from different states whose applications have been denied. Here also, California is the highest followed by New York and Texas.



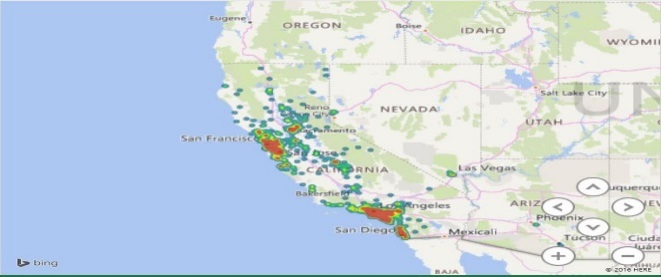
|  |
| --- |
| **select lca\_case\_employee\_state, count(status) as Australia\_Data from project1 where visa\_class = ‘E-3 Australian’ group by lca\_case\_employer\_state, status;** |

The E-3 visa is a United States work permit visa for which only citizens of Australia are eligible. This query is used to determine the count of Australian employess across USA.



|  |
| --- |
| **select lca\_case\_employer\_postal\_code from project1 where lca\_case\_employer\_state = ‘CA’;** |

This query is used to go deeper into the data and determine the number of applicants who are from California. The heat map below shows the places in California where the maximum number of applications have come from. Te areas are mainly, Los Angeles, San Francisco and San Diego.



* 1. **Analytics Tools setup and Integration**

We used Microsoft Excel 2016 to transfer the output from Databricks to visualize some data using geographical maps and heat maps.

* 1. **Testing**

The testing of the end system should conform to various conditions and criteria set during the business requirement analysis. The testing process is in accordance to the acceptance criteria wherein the data can be properly queried and displayed while maintaining accuracy. All queries should produce expected output based on which resultant graphs and reports are generated.

**4. Conclusion**

Majority of the visa applications that are filed are **certified** (86%). Very few of the applications were **withdrawn**(3%) and **denied**(4%). Significant growth in application certified and decline in denials each year. The **highest wage rate** for employees working on H1B visa was noted in the year 2014. (**Around 100K**). There was a dip in the year 2013 where it fell to around 80K. The maximum applications were filed from the state of California and New Jersey. Texas and New York were next in line.The rejections were also seen high in California, New York, New Jersey and Texas. The Australian employees majorly worked for employers in **New York** and **California**. Most number of applications from California are from **Los Angeles** followed by **San Francisco** and **San Diego**. On basis of our analysis, we found that the employees coming from different parts of the world, prefer working in the states of California and New York.

**5. References**

[1] **DATASET URL** **:** [https://app.enigma.io/table/us.gov.dol.oflc.h1b.2011?search[]=(%22international%22%20%22student%22)](https://app.enigma.io/table/us.gov.dol.oflc.h1b.2011?search%5b%5d=(%22international%22%20%22student%22))

[2] **GITHUB CODE** **:**

<https://github.com/pparasa/H1B-Applications-Data-Analysis>

[3] **REFERENCES :**

[https://](https://www.safaribooksonline.com/library/view/mastering-apache-spark/9781783987146/ch08s10.html)[www.safaribooksonline.com/library/view/mastering-apache-spark/9781783987146/ch08s10.html](http://www.safaribooksonline.com/library/view/mastering-apache-spark/9781783987146/ch08s10.html) [4] <https://databricks.com/blog/2015/06/22/understanding-your-spark-application-through-visualization.html>

[5] <https://www.youtube.com/watch?v=xAQQClV3k1E>