H1B data Analysis using Mongodb and Node js

Haritha Damarla
Indiana University Bloomington

Aravind Bharatha
Indiana University Bloomington

Abstract

With the recent developments, there are many people who wanted to work for an US employer and also at the same time employers at USA are looking for foreign talent to hire for their company. One way to make it happen is, the company should provide the work permit by raising H1B visa. So in this project we wanted to analyze the H1B dataset and provide various analyzations that helps in taking an important decisions. We have analyzed the data in the terms of which company can sponsor more H1B visas, what are trends in applicants when compared over different years, which field has more demand and what are the statuses of applicants who applied for H1B.

1 Introduction

Now a days we are required to deal with the unstructured data, also though initially data is structured, as the time forwards there will be unpredictable changes to our data that keeps changing with the requirements at different times. The solution to deal with such data is a NoSQL database which stores all the data related to an entity in a single document.

There are many such NoSQL databases such as MongoDB, Cassandra, Reddis, Neo4J etc.

For our project we chose MongoDB as it provides best of both SQL and NoSQL world, and all the data inside is stored in the format of documents and we need not restrict the data to predefined columns. Also, MongoDB provides High-Performance and High-Availability and has a powerful query system.

And for data visualization we have used chart.js and google map api visualization as it is easy to use, open source, responsive and provides representing the data in various other ways.

We have displayed all the results of the queries with appropriate visualizations in the website by using the technologies Node JS, Express JS and EJS

2 Problem Statement

In United States H1-B visa is required to be sponsored by the employers to the recruit the foreign employees into their company. In recent days there are many changes happening to it and we can observe higher demands in either ways i.e. people are craving to work for US employers also at the same time employers wanted to hire the foreign talent.

Hence, we have taken a dataset from Kaggle with about 10,00,000 records providing the data on the same. The data has the below fields on which have drew various patterns.

1) Case_Status: It describes about the application status. It has been given 4 values as Certified, Denied, Withdrawn, Certified Withdrawn

2) Employer_Name: Name of employer at US who is sponsoring the Visa

3) SOC_Name: Code that is related to the job sponsored

4) Job_Title: Title of the job

5) Full_Time_Position: Has been given 2 values Y/N

6) Year: Year for which Visa has been filed

7) WorkSite: location of the job

8) lon: longitude co-ordinate of the location

9) lat: latitude co-ordinate of the location

3 Tools and Technologies:

- Node js
- Express js
- Ejs
- Chart js
- Mongodb
- MongoClient
- Html
- Css
- Javascript
- JQuery
- Google Maps Api and Visualization library
- Sublime

4. IMPLEMENTATION:

We have developed a simple webpage to represent all the charts using node js and express js from the server side. Node js will host the code of our application and express js is used to create a server.

```
npm init
```

The above command creates server.js file, that hosts the code of H1B application.

The below command helps in installing the express js, which helps in getting our server up and running.

```
npm install --save express
```

We have used MongoClient library in node js to connect with mongo db. The below command helps in establishing the connection to database from the code, and thus helps in running the queries against the database and retrieves the results from it

```
npm i mongodb -- save
```

The retrieved results and brought up and presented on the UI using HTML, CSS, JavaScript, JQuery and used CSS for styling.

5 Approach and Analysis

We have collected the data from Kaggle in CSV format, and created a database in MongoDB and loaded the data into MongoDB by running the command from the system command prompt

```
mongoimport -d h1b -c h1bdata --type csv --file h1b_kaggle.csv –headerline
```

The above command helps in loading the csv data into h1b database created inside our mongodb.

Following which as discussed in the Implementation section, we have installed tools related to technologies Express JS,

Node JS, EJS to present the various result we are analyzing into a website.

We have analyzed the data in the below perspectives

a) Which company is sponsoring more H1-B VISAS?

b) Which area of USA has more approved/rejected/withdrawn Visa Petitions?

c) What are the trends in the applications in the years 2016 vs 2015?

d)Which role is in more demand in USA right now?

e) What are the status of different applications?

a) Which company is sponsoring more H1-B VISAS?

The query we used to run on the db are:

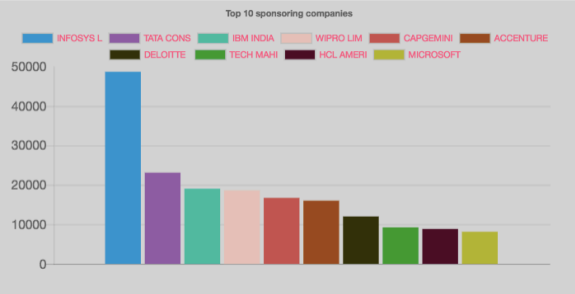
```
dbo.collection ("h1b").aggregate ([

{$group :{ _id:"$EMPLOYER_NAME", total :{ $sum: 1}}},

{$sort :{"total": -1}}

])
```

This query produces the results of top 10 sponsoring companies, and we tried representing the data in a bar chart also with this visualization we can understand that Infosys pvt Limited company offers more H1-B Visas, and hence employees who are looking for a sponsor in USA can apply for Infosys to improve their chance of working in USA.



b) Which role is in more demand in USA right now?

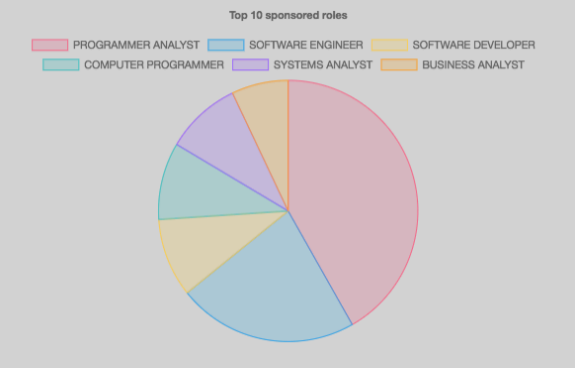
The query we used to run on the db are:

```
dbo.collection("h1b").aggregate([

{$group: {_id: "$JOB_TITLE", total:{ $sum: 1 }}} ,{$sort:{"total": -1}}

])
```

This query produces the results of top 10 sponsored roles, and with this visualization we can understand that Programmer Analyst role is in more demand now and thus infers that employees who are working as Programmer Analysts has more scope and also most parts of USA requires Analysts from foreign countries.



c) What are the trends in the applications in the years 2016 vs 2015?

The query we used to run on the db are:

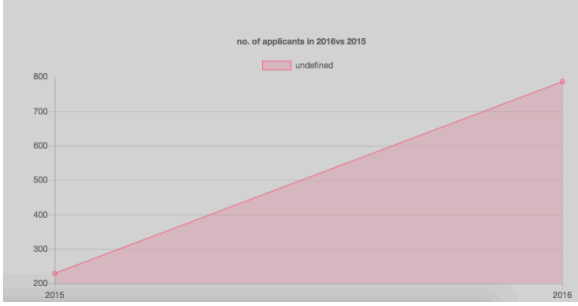
```
dbo.collection ("h1b").aggregate ([
```

```
{$match:{JOB_TITLE:/.*DATA ENGINEER.*/}},

{$group :{ _id:"$YEAR", count :{ $sum: 1},

{$sort:  {"count":1} }])
```

This query produces the results of no. of applicants in different years. With this line graph we can analyze that no. of applicants had increased from 2015 to 2016.



d) What are the status of different applications?

The query we used to run on db is:

```
dbo.collection("h1b").aggregate([

{$match:{JOB_TITLE:/.*DATA ENGINEER.*/ }},

{$group:{_id:{"CASE_STATUS":"$CASE_STATUS"}, count:{ $sum:1 } } },

{ $sort:  {"count":-1} }]
```

This query produces the results of no. of applicants got certified, withdrawn, denied etc.

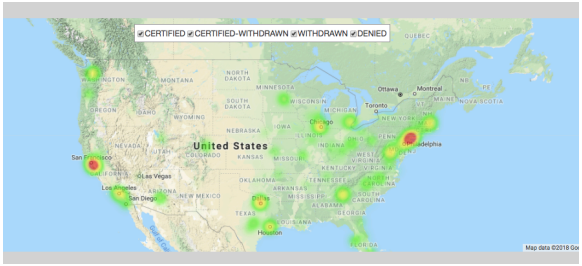
Thus this visualization helps in understanding the status of the applications.

e) Which Location on World map has most number of approvals of H1-B STATUS?

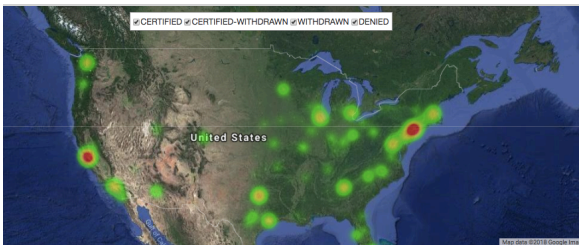
This is the most interesting part of the project. Here we have used heat map visualization and thus helps in understanding which part of the US has more approved applications. Here we have integrated google maps api with our applications, to represent different areas on map. Thus the are with highest number of applications with status as CERTIFIED is represented in red color.

We have also provided the toggle bar option, where user can check and uncheck various statuses, to understand which area has better opportunities.

Terrain view :



Satellite view:



6. EVALUATION AND FUTURE SCOPE

We have developed this project as part of our course, and it helped in introducing and exploring various new technologies.

The website is also built on very minimalistic data that is collected from Kaggle.

In future the same representations and data analyzations can be performed by running queries on bigger data for more realistic representations.

This also helped in understanding how NoSQL data is powerful and important to understand for the current world data change requirements.

The whole project has been developed in a span of 3 days, and we have tried all basic kind of representations, use cases and analysis, but this can be extended to a bigger project by providing the choice to end user to filter out the representations for a specific role.

7. CONCLUSION

With the outcome of the project we could understand that, Infosys pvt Ltd and Tata Consultancy are the top 2 companies that sponsor more H1B visas.

Programmer Analyst and Software Engineer are the top 2 roles in highest demand right now.

Most of the H1-B applications are in CERTIFIED status which says eligible. And there is significant increase in the number of applications from 2015 to 2016.

This project that we developed has given us a great satisfaction of learning new technologies, though it is not ready to be rolled out in the business perspective, but it definitely adds a great value to the learning and doing.

We would like to express our sincere thanks to our professor Ying Ding and AI YiBu for their consistent support throughout the course.

8. REFERENCES:

[1]https://www.kaggle.com/msingh32/h1b-visa

[2]https://developers.google.com/maps/documentation/javascript/examples/layer-heatmap

[3]http://www.chartjs.org/docs/latest/getting-started/

[4]https://codeburst.io/build-a-weather-website-in-30-minutes-with-node-js-express-openweather-a317f904897b