Analysis on H1b visa in the United States using Bigdata.

Technologies used: Hive, sqoop, HBase, Mapreduce, Apache Hadoop, Mysql

Result are visualized using Tableau.

Varun Sundaram

26th April 2019

Email @ varunsundaram@live.in

Visa application analysis in United States using Big Data Ecosystem.

Varun Sundaram

Abstract— **This project aims to perform data analysis on H1B type visa, formally known as Labor Condition Application (LCA) of the United States for temporary foreign workers. The dataset was openly available under the disclosure of data category for public use. For this project data from the year, 2012 to 2016 has been used to analyze the h1b visa petition which was filled by the employers. Some of the investigation areas like popular jobs, job title and employer who file more number for the h1b petition has been discussed. The Hadoop bigdata ecosystem was used to analyze this project with the help of the hive and MapReduce framework. Based on this research the result has been visualized using tableau. MapReduce paradigm was extensively used to analyze this project to access the data directly from the source MySQL database and output was stored in HBase. To process this analysis, in a better way it is recommended to use good high configured system.**

Keywords: H1b Visa, Big data, HDFS, Hive, HBase, MySQL, MapReduce, Design Patterns, Visualizations

I. INTRODUCTION

The United States of America generally mentioned as land of opportunities, it attracts millions of immigrants across the globe. The main gateway to the United States in H1B work type visa. Generally, employers file the LCA on behalf of the employees and provide the work authorization. H1b type visa is non-immigrant visa which allows the foreign workers to temporarily work in the United States in speciality occupations. This type of visa will be provided to a person who is highly skilled and specialized in technical and practical knowledge towards the work.

This project aims to do a analysis related to H1B type visa and provide an insight about it. The analysis was performed based on the open dataset which was taken from the Kaggle for the year 2012 to 2016. The attributes like case status, employer name, job title, prevailing wage, worksite etc. are used to perform this analysis.

Some of the main areas are covered in this analysis such as, popular job position based on the visa petition, a quick comparison of the organization which apply more number of h1b visa and provide sponsorship to the employees, Industry which hire more number of data scientists for the work, and understanding the growth percentage of the data engineer position based on the visa petition filing. In recent times more numbers of analysis are being performed using the machine learning techniques and other state-of-heart methods. However, the distributed processing Hadoop environment also provide a better vision for performing these types of analysis. Apache Hadoop is considered as a one-stop solution to store and analyzes the big data in the distributed environment. The connectivity between the other traditional database and the Hadoop was much easier and simpler. Hence, hadoop attracts many users and provide a better solution for their requirement. There are several tools and technology which are adopted easily to the Hadoop environment. For this project, Mysql has been used to store all the data at the initial stage later it was transferred to Hadoop for follow-up the analysis.
Choosing different tool and technology is the key aspect for analyzing data on the Hadoop distributed environment. This project has been carried out employing Apache Hadoop as a distributed environment and analysis performed along with the MapReduce framework, hive for processing, databases like MySQL, HBase have been used to store the data. Sqoop was used to transfer data from the MySQL to Hadoop distributed file system. Several design patterns are used in MapReduce to carry out this analysis. The design patterns like numerical summarization, top 10 design are used in java programming. The outcome of this project aims to identify the work demand pertaining to the skill sets and the employers who are

willing to sponsor the visa for the employees. The result of this project has been visualized in the tableau which provides a better and simple understanding of the job market in the United States.

The contribution of this project aimed to answer the following research questions:

1) Is the number of petitions increasing over the year for Data engineers?
2) Which Industry in the United States hire more number of Data Scientist?
3) Which organization file more number of h1b visa petition?
4) Which are the most popular jobs in the United States for the year 2016 and 2015?

II. Related Works

Currently, more than 200 million international immigrants are living in the United States of America. The largest immigrants are settled over there for work. It is clearly witnessed H1b visa played a central role by bringing more people to the United States [1]. The H1b visa analysis, in this area doesnt seem to have more research related to big data. However, few of the works are carried out based on the machine learning method using the input as location, salary, job type etc [2].

This project aims to do the analysis using the Hadoop bigdata and its related technologies. Bigdata application has become more essential nowadays since the data volume are growing high. The traditional relational database is now being used only for storing the data. However, big data ecosystem provided end to end facilities such as analyzing, storing and visualization. Bigdata technology has different layer data querying, data access and management layer [3]. Based on the requirement the Hadoop environment is fully scalable. Hadoop jobs are generally running on several languages based on user needs. The main processing framework for hadoop was MapReduce. [4] The hadoop cluster provides high performance when it has a good configuration. MapReduce paradigm generally needs some better memory when the input data are more. It split into map and reduce task to process the data. MapReduce can be performed in several languages like java, python etc. The MapReduce by default has fault tolerant character [5].

Data processing in hadoop was more efficient than other distributed processing. HBase is one of the NoSQL distributed database which was developed on the top of the Hadoop distributed file system (HDFS). The hybrid architecture generally enables the faster search and retrieval of the growing data. When the data stored in hdfs and retrieved through HBase provide good performance [6]. Analyzing the data in hadoop provide a complete solution with its other supported packages. Hive is the one of key resource in hadoop, it is generally used a warehouse and query the data. When it comes to data transfer sqoop play a vital role. There is no other better tool available than sqoop to transfer the data from the relational database to the hdfs. Utilizing all these will provide a better way to process and analyze the data [7].

In [8] utilized the apache pig with hadoop and performed the data analysis. The volume of the data seems to be more. However, the expected result has been obtained by utilizing the hadoop bigdata. There are several area hadoop can be implemented. Since it provides high availability and despite any failures, multiple data copy has been stored in fully distributed mode. In this project aimed to utilize the different technologies like hive, MapReduce paradigm for data analysis and process, HBase and MySQL to store the data.

III. SYSTEM SPECIFICATION

Hadoop distribution	Apache Hadoop 2.7.7
No. of Data Nodes	1 node
Memory	4gb
V core CPU	4
Data Disk	50gb
CPU clock speed	2.30 GHz
Operating System	Ubuntu 18.04

A. Justification of tools used

The utilization of the tools in this analysis has been explained in 1

IV. METHODOLOGY

The project methodology has been explained in the below sections.

1) Data description.
2) Work Flow.
3) Data sourcing.
4) Data Preparation.
5) Data Processing.
   - Method 1
   - Method 2
   - method 3
6) Data Visualization.

Technology / Tools	Purpose	Justification
Apache Hadoop	Supports the bigdata ecosystem and to store the processed data in hdfs.	Open source and supports mutiple technology.
MapReduce	Processing framework for the project	Java classes are extensively used here.
Hive	Data Warehouse for hadoop	Inbuilt ware house for Hadoop and easy to process data using Hql.
Sqoop	Data transfer from RDBMS	Used to offload data from MySQL and store in hdfs
HBase	Nosql database used for storing final output	Since it runs at top of hdfs it is used for accessing the data easily.
MySQL	used for initial storage purpose	one of the better performing db with normal resources.
R	Used for cleaning the dataset	Cleaning and data transformation can be done easily
Tableau	To visualize Output data	Provide a better visualization.
Eclipse IDE	Used to connect MapReduce and perform the analysis	Better user interface for Java programming
Winscp	Secure file copy from ubutu machine to local	All the final output file has been taken from machine using winScp
Putty	Secure remote shell access	Ubuntu terminal was observed bit slow hence used putty to connect the machine.

Fig. 1. Justification of tools used

Field Name	Description
S_NO	Unique key
CASE_STATUS	CASE_STATUS Status associated with the last significant event or decision. Valid values include "Certified," "Certified-Withdrawn," Denied," and "Withdrawn"
EMPLOYER_NAME	Name of employer submitting labor condition application.
SOC_NAME	Occupational name associated with the SOC_CODE
JOB_TITLE	Title of the job.
FULL_TIME_POSITION	Y = Full Time Position; N = Part Time Position.
PREVAILING_WAGE	Prevailing Wage for the job being requested for temporary labor condition
YEAR	Year of the H1b visa processed detail
WORKSITE	City and State information of the foreign worker's intended area of employment.
LON	longitude of the Worksite.
LAT	latitude of the Worksite.

Fig. 2. Dataset Description

A. Data description

H1b is the employment based non-immigrant visa category which allows the temporary foreign workers to work in the United States. For this kind of visa, the employer must need to file a petition to the federal government the United States. This dataset contains in reference to 2 visa petition details for the period of the year 2011 to 2016. The dataset has 10 attributes like employer name, job title, case status etc. Based on these details the dataset has been analyzed using the bigdata eco system. The choice of choosing this dataset since, it gives profound insight about the H1b visa category and the details of the employers, who are more likely to sponsor vise for the skilled foreign workers. The outcome of this project helps the both employee and employers to understand the job demand in the United States based on the job title and the organization requirement.

B. Workflow

Project work-flow design has been explained in reference to 3

C. Data Sourcing

The data set contains ten columns and more than one million rows which were loaded in Hadoop pseudo-distributed model. Dataset has been extracted from [1].

[1]**https://www.kaggle.com/nsharan/h-1b-visa**

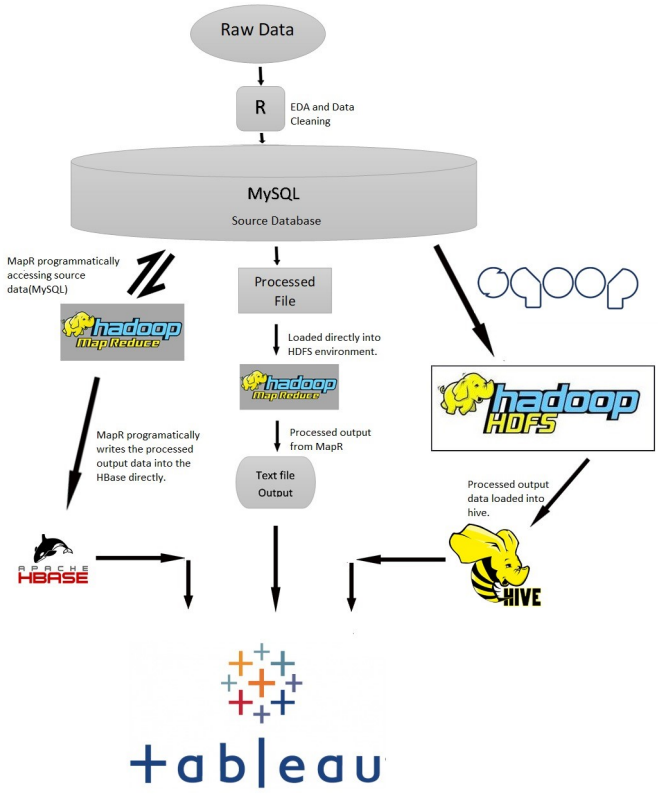

Fig. 3. Workflow

The dataset was split into two based on the processing required in the MapReduce and hive. Since there are 3 different processes has been carried in this project, splitting of the existing data would be considered a

better way to run in Hadoop and process it.

D. Data Preparation

In this process, the data has been cleaned and transformed according to the project need. The dataset which is acquired from kaggale had some of the missing values. All these were being removed and cleaned using R studio. Later the null values, special characters, and some of the numeric function were being taken care before loading the data into the initial source database MySQL. Since the pre-processing of the dirty dataset is considered as one of the important tasks Cleaning the data must be performed prior to processing the data into Hadoop to get the expected results.

E. Data processing

There are 3 different methods has been carried out in this project for processing the data.

1) **Method 1:** The cleaned dataset has been initially stored in MySQL before processing. Mysql is one of the best open source databases which provide high performance. Also, the read/write operation is much faster in MySQL. The size of the acquired dataset is more than 1 million hence, MySQL has been considered as best fit to handle the high volume of data without consuming more resource.

1) Source database: Mysql
2) Processing framework: MapReduce
3) Output data storage: Hbase

a) ***Design Flow****:*

- Initially, raw H1b data has been acquired, cleaned, and stored in the Mysql.
- Appropriate column and table were created to store data in MySQL database prior to processing to MapReduce.
- Programmatically the source data (Mysql) has been accessed by the MapReduce processes.
- Prior, processing the MapReduce appropriate table has been created in HBase to store the processed MapReduce output data.
- Post-processing in MapReduce output has been stored in HBase programmatically.

2) Method 2: The different approach has been used here in this method, with the slight change of storing the output data. The output data was processed using the MapReduce framework and been stored as a text file and later converted into csv for the purpose of visualization.

1) Source database: Mysql
2) Processing framework: MapReduce
3) Output data: Stored as a text file in local

a) ***Design Flow****:*

- The dataset which was stored in MySQL has been exported as a text file in local.
- The exported local file was used as input for MapReduce.
- The generated MapReduce output has been taken and used for visualization.

3) Method 3: In this method, Sqoop was used to offload the data which was stored in MySQL and import the data to hdfs. The required MySQL connection string has been provided to connect and import the data source. Sqoop provides high compression and the data transfer is much faster.

1) Source database: Mysql.
2) Data Iingestion: Sqoop.
3) Data imported in HDFS.
4) Data process/storage: Hive warehouse.

a) ***Design Flow****:*

- The source data which was stored in MySQL has been accessed using the sqoop.
- With the help of sqoop, data was read from MySQL database and exported to hdfs.
- The appropriate table has been created in Hive and loaded the data from hdfs which was imported by sqoop.
- The processed data in the hive has been taken in the local system and used for visualization.

F. Data visualization

The visual representation of the data is more important and final part of this project. In general, visualization delivers and fulfil the overall goal and make the audience to understand the project easily. All the results which are processed in the bigdata ecosystem are taken as the output file, with the help of tableau it was visualized. Totally 4 business queries are being framed and they are being answered visually.

V. RESULTS

A. Research Question 1:

Is the number of petitions increasing over the year for data engineers?

This query has been processed in the MapReduce framework. The objective of this query is to identify the total growth percentage of the Data engineer position for the H1B visa petition from the year 2012 to 2016. The output has been visualized in 4. Based on the result It has been concluded that there was some average increase in data engineer role when compared

yearly. However, the overall comparison 2012 to 2017 some inconsistent trend has been observed.

Design pattern used: For this MapReduce framework, numerical summarization pattern has been used. Along with that external source output pattern using the Immutable Bytes writable class has been developed. Few customizations were done for reading the record and get the data from source MySQL.

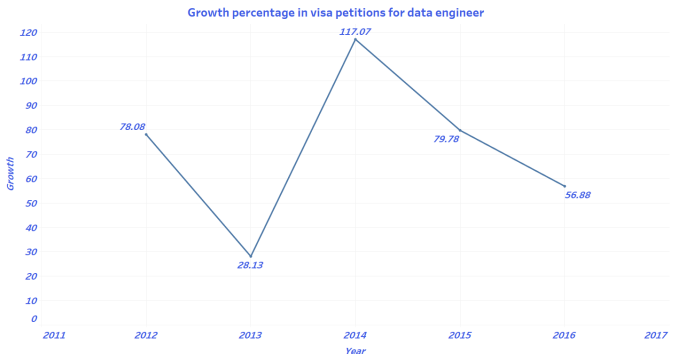

Fig. 4. Number of H1B petitions

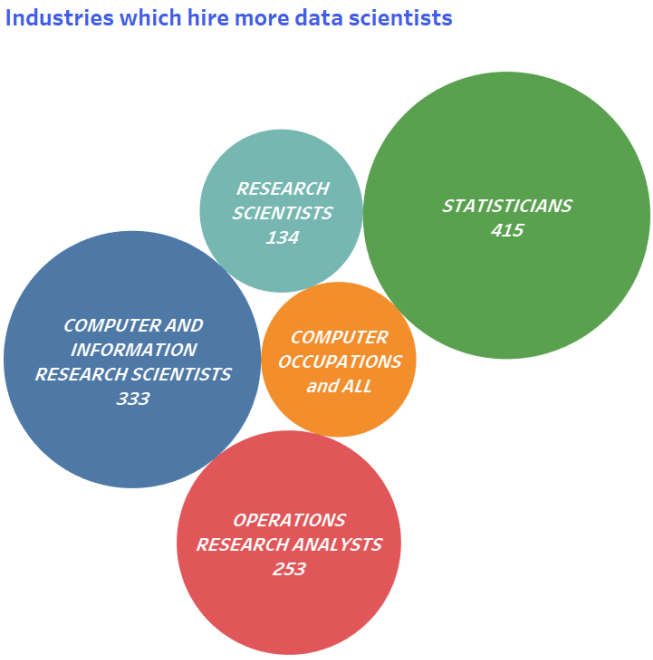

Fig. 5. Data Scientist requirement based on industry

B. Research Question 2:

Which Industry in the United States hire more number of Data Scientist?

This query has been processed in the MapReduce framework. In this, it says about the industry which has more number of data scientists based on the h1b visa petition. There are lots of industry which have data scientist however, the result has been filtered and taken out the top industry which has a number of the data scientist. From the 5 visualization result it depicts the Statistician industry hire number of data scientist role in the United States based on the visa application.

Design pattern used: For this top 10 design pattern has been developed to process this in the MapReduce framework.

C. Research Question 3:

Top organization in the United States who filed most number of h1b visa petition for the year 2016 2015.

To process this query, the hive has been used based on group by function in year and employer. Two separate hql was used for the year 2015 and 2016. The result has been combined and visualized as a comparison 6. The total number of visas filed the petition by the employer along with the year has been compared as shown below. Upon comparing the result based on top organization, Infosys limited has filed a more h1b petition for the year 2016 and 2015 than other organizations.

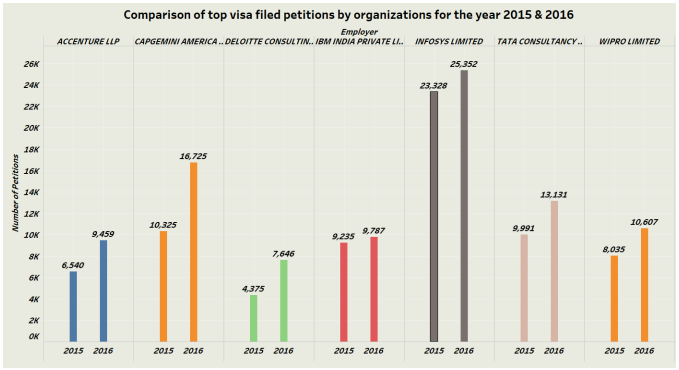

Fig. 6. Visa petition filed based on organization

D. Research Question 4:

Which are the most popular jobs in the United States for the year 2016 and 2015?

Hive has been used to process this query using the hql the job title and year has been done group by an order by the total number of applications. The results say that the programmer analyst role has an

increasing trend with more number application. Hence, it will be considered a more popular job in the United States during the year 2016 and 2015. 7

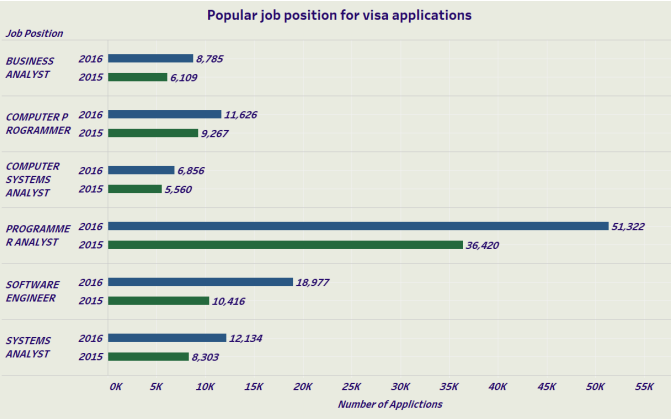

Fig. 7. Popular job in U.S

VI. CHALLENGES FACED

1) JVM and container memory out error has been encountered during running the MapReduce, since the data had more than 1 million record. Few listed parameters are tweaked post that the data was processed in MapReduce.
   - mapreduce.map.memory.mb
   - mapreduce.reduce.memory.mb
   - mapreduce.map.java.opts
   - mapreduce.reduce.java.opts
2) Programmatically accessing the source database MySQL was failed in MapReduce even after adding the appropriate driver connection. Post including the external source output pattern the MapReduce able to fetch the record from MySQL.
3) Since, the dataset had 1 million rows with large number of record frequent memory out error has been observed while analyzing it in hive using hql. Later, the table has been dropped and the dataset has been split and loaded back in hive based on the Year.
4) The actual project has been planned to analyze with the apache spark. However, due to lack of memory the data couldnt be processed. Hence, switched to the MapReduce.

VII. CONCLUSION & FUTURE WORKS

To conclude, in this project H1b visa analysis has been performed for the year 2012 to 2016. Based on the overall result the data engineer and data scientist role are more popular and they are expected to grow more in upcoming days. Many people around the world have their dream to work in the United States. This result and analysis based on h1b visa petition help them and give deep insight into the overall job market. Based on this they can improve their skill set to achieve their goal. To extend this work in future it can be done by including more attributes such as wages per state, visa status, and the state where they process more number of the h1b visa petition. So, it helps to provide more perception about the work permit in the United States. Currently, this analysis has been done using MapReduce and Hive. For storage purpose, MySQL and HBase have been used. The performance of the MapReduce was not up to the level while processing the high volume of data. Also, to run the Hadoop and perform the detailed analysis, the computer hardware and memory plays an important role. But in future, it can be analyzed using Apache Spark and apache kudu for the better and faster analytics purpose by utilizing them in the supercomputers.

REFERENCES

[1] D. Ravindranath, "Visa regulations and labour market restrictions: implications for indian immigrant in the united states," *The Indian Journal of Labour Economics*, vol. 60, no. 2, pp. 217–232, 2017.

[2] E. Salamanca Pacheco, "Implications of the u.s. visa reform for high-skilled mexican migration implicaciones de la reforma de visas estadunidenses," *Norteamrica*, vol. 14, no. 1, 2018.

[3] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431 – 448, 2018.

[4] H. Alshammari, J. Lee, and H. Bajwa, "H2hadoop: Improving hadoop performance using the metadata of related jobs," *IEEE Transactions on Cloud Computing*, vol. 6, pp. 1031–1040, Oct 2018.

[5] S. M. Nabavinejad, M. Goudarzi, and S. Mozaffari, "The memory challenge in reduce phase of mapreduce applications," *IEEE Transactions on Big Data*, vol. 2, pp. 380–386, Dec 2016.

[6] C. Dobre and F. Xhafa, "Parallel programming paradigms and frameworks in big data era," *Int. J. Parallel Program.*, vol. 42, pp. 710–738, Oct. 2014.

[7] G. Mackey, S. Sehrish, J. Bent, J. Lopez, S. Habib, and J. Wang, "Introducing map-reduce patterns to high end computing," in *2008 3rd Petascale Data Storage Workshop*, pp. 1–6, IEEE, 2008.

[8] A. Jain and V. Bhatnagar, "Crime data analysis using pig with hadoop," *Procedia Computer Science*, vol. 78, pp. 571 – 578, 2016.