Big Data analysis on H-1B Visa Application in the United State

Technical Report · June 2019

CITATIONS
0

READS
881

1 author:

Jyoti Chavda
National College of Ireland
6 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Big Data analysis on H-1B Visa Application in the United State

JYOTI CHAVDA
National College of Ireland
x18114831

***Abstract*—H-1B visas are a class of non-immigrant visas for temporary foreign workers in the U.S. It is the most common visa held after completing its higher education for an international student. It is the biggest visa application of the country for inmate workers and is a crucial channel for highly skilled immigrants. So, it is essential for foreign employers to analyze the job opportunity available in the business market of United State also to keep a to record of number of visa approved in each year. Analyzing of this pattern is essential for both academics and business perspective. This project considers H-1B visa application data for last 6 years to analyze the job trend in united states. This data includes case-status, Employer name, occupational name, title of the job, prevailing wage and city and state information. The analysis is carried out by using the big data application in which relational databases like Hbase and MySql are used to store the data and Hive, Pig and MapReduce is used to generate the design pattern. After this visualization is done by making use of R programming language and tableau software. The result of this analysis is used in employment search and recruiting strategies to determining the pattern and best time of year and the top geographical areas.**

***Index Terms*—Hadoop, HBase, MySql, Hive, Pig, MapReduce, Sqoop, H-1B visa application, R, Tableau.**

I. Introduction

As the global mobility of students continues to increase, many students abroad are looking for opportunities to stay in their countries of learning. A United State employer offers a job for this visa and applies for an H-1B visa. This approved request is a permit for this employer to obtain a visa stamp and work in the US. In the United States, the Labor Department, hiring cost is 30% of the potential earnings for the first year of the employee [1]. Not the only foreign professionals who can flourish from the H-1B advantages are foreign professionals, employers also have a lot to profit from. H1B visa, which offers the best professional in the world to enter the USA and work for your company, offers people with educational backgrounds in science, technology, engineering, and mathematics. Comparing the income of H1B visa holder to that of US naturalizes workers, holders of H1B visa do not suffer from lower wages than us born workers [2].

To analyze the benefits of H1B visa in the united state using a large amount of data which gives the insights of the visa application and job trend in the United State. The new platform like big data application was used. Big data technologies analyze large amounts of information and help find the insights of large data [2]. It is a framework that enables the distribution by simple programming models of larger data sets across clusters of cheap commodity hardware. significant increase in the number of users, search, map and supply of the information according to preferences. Big Data is used due to the scalability and flexibility issues as the storage ability and data size increase immensely [3].

In its initial generation, Apache Hadoop provided distributed data storage and resource system with an Open Source big data solution. The process of knowledge is extremely important for data technology. The SQL quires of conventual relational databases are still the most convenient analysis tool to be utilized by people without programming backgrounds [4]. Hadoop is the most commonly distributed system works on big data, which is capable of translating the SQL query into MapReduce tasks performed with HDFS. Apache hive is more reliable, which is the same as SQL and works on big data, in Hadoop. Apache hive and pig which turns the script into a series of MapReduce jobs which helps to retrieve the information from big data.

In this project, a large volume of H1b datasets is used which contain various parameters to analyze. This data has almost 3million records which unable to give the proper insights and appropriate information about the visa application and job trend market in the United State.

- To carried out the important information like which trending job market in United State?
- Finding the top job position in each year?
- Average prevailing wage for job in each year?
- Total number of petitions in each year?

Effective analysis of these complex issues will be very useful for foreign employers who willing to work in the United State. In this regards, Section 2 contains related literature reviews of the various framework of big data with different tools for query solving, such as Hive, Pig, and MapReduce. Section 3 dedicated to following up approach. Section 3 briefly explain the visualization of the query and results lastly section 4 consist conclusion.

II. RELATED WORK

Countries are keen to retain international graduates of their national institutions as they compete for highly skilled talent. A relatively new political phenomenon is the category of visa or programs that have been specially designed for international student of a national institution. A domestic university degree is the prerequisite of these programs that directly involve

higher education in migration issues and put higher education institution on the states frontier [5]. Thus, developed policies that allow higher education to be used as a magnet and filter for qualified migrants. The policies of the United Kingdom led to the improvement of a new immigration channel by which people to a country for study and are then permanently or temporarily hired to work.

At each seconds data is increasing exponentially due to various user activities. In the past for analyzing the H1B visa in the United State, big data techniques were used. The author [3] built the recommender system by applying the complex queries on Hadoop platform by making use of various tools such as hive, pig, Sqoop. Analyzing this tool is based on the data volume, file format and many more. For this research used H1B visa application used which is available in CSV form. After analysis author compares the performance of these tools and the results show that a quick recommender system can be built on a dataset saved in CSV file format using a different query tool.

Big data technologies are now widely used in a number of industries, in this research author [1] investigate the job trend in New York by using big data techniques which include Hadoop framework. Data set used by the researcher is contain few gigabyte data. Handling of such a large volume of data can be difficult with conventional resources. Hadoop proved the ease to handle the large volume of data and various tools to examine the insights of the datasets. The main objective of this study is to understand the principle of data visualization which is applied to the massive data set. The study data analysis is helpful in the hiring industry for business perspective. This study demonstrates the application of big data in the field of job analysis also provide them to analyze the job trend by considering the job position and wage trend in New York.

Hadoop is an open source of implementation where MapReduce provides fault-tolerance and scalability for large data analysis. The author [6] developed and implemented the efficient structure for a MapReduce system is the management of quick data loading, quick query processing, highly effective storage space usage and strong adaptability to dynamic working characteristics. MapReduce cannot control cluster records directly. They must use the distributed file system on a cluster level. This approach is used in our project to carried out the result of MapReduce.

III. Methodology

This section gives the information about data collection, preprocessing of the data, technologies and approach which is used to carried out the result.

A. Data Collection

The dataset used for this project was collected from Kaggle https://www.kaggle.com/nsharan/h-1b-visa. Kaggle is an open data portal and it is a publicly available source. This data contains information about H1B visa petitions from the year 2011 to 2016, with approximately 3million records overall. This dataset includes information like case status, employer name, worksite coordinates, job title, prevailing wage, occupation code, and year filed, location information, longitude, and latitude. This data is generated by the Office of Foreign Labor Certification in the United State which gives the information about H-1B visa including immigration programs. It is very important to know the chance of acceptance of visa based on past records because of the lower acceptance of an H-1B visa and application charges.

The column of dataset includes following information,

1) Case_Status: The status compared with the last past decision. which include Certified, Certified-Withdrawn, Denied, and Withdrawn
2) Employer_Name: Employer name who applying for job application
3) Soc_Name: Name of the department, which consist different r type of jobs
4) Job_Title: contain the information of title of the job.
5) Prevailing Wage: The salary for the job requested for temporary any full time employment.
6) Year: year in which the H-1B visa requested has been submitted.
7) Worksite: contain the information of city and state where expected jobs are available for foreign workers.
8) Lon and Lat: give the details of longitude and latitude of worksite.

B. Data Prepossessing

The available raw data are unclear and cannot be analyzed straight away. The data were made accessible for quick investigation by transform the data in required form. It is done with various RStudio libraries like dplyr, tidyr to remove unwanted characters, missing values and special character. Unused columns are deleted from the datasets. Raw data contain different data type which is not compatible with MySQL database, so this need to convert in appropriate datatype this will done using RStudio. There are many columns who have comma in-between which create problem while loading data into database. This special charecters, unwanted blank spaces and NA values are removed by using the MySQL queries. For further analysis necessary data were extracted.

C. Process Architecture and Approach

The process architecture is comprising following components,

1) Input data: Kaggle data which is available in csv format.
2) Hadoop framework: Use Hadoop platform to perform query using Hive, Pig, Sqoop and MapReduce.
3) Input Database: MySQL is used to store input data for further query.
4) Output Database: HBase is used to store output data for visualization.

D. Process Flow

Process flow diagram shown in the fig.**??**which follows the following steps.

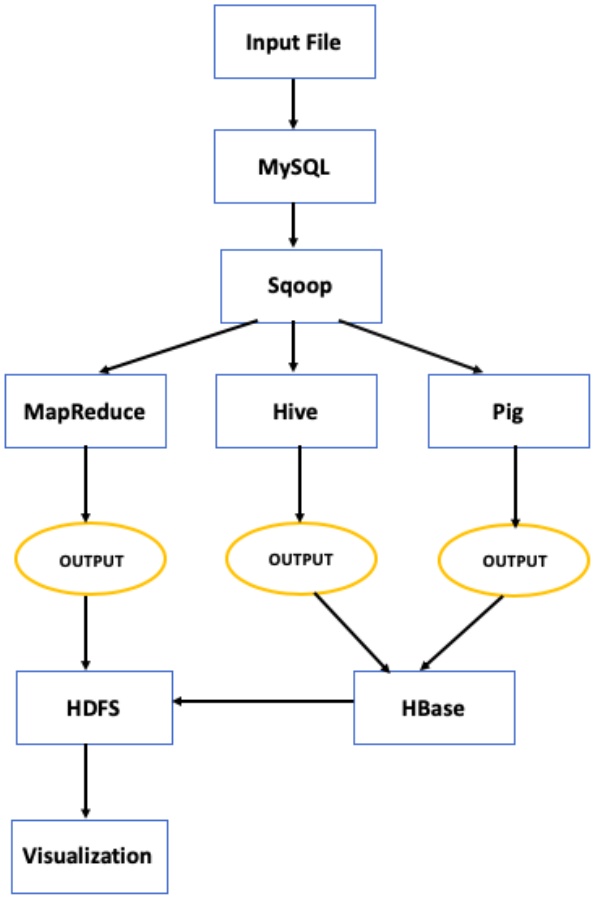

Fig. 1. Architecture Diagram

1) Download data from open data portal.
2) Input data loaded into MySQL database.
3) Import data to Hadoop query tools (hive, pig, MapReduce) using Sqoop import command.
4) final output stored in HBase and hdfs.
5) Hdfs output was used for visualization.

Before implementation of this steps it is required to start all the services of Hadoop and HBase and check these services running properly. Figure 2 show the running services.

```
hduser@jyoti-VirtualBox:~$ jps
3664 HQuorumPeer
3846 HRegionServer
2310 NameNode
2886 ResourceManager
3047 NodeManager
2471 DataNode
3962 Jps
2687 SecondaryNameNode
3727 HMaster
```

Fig. 2. Hadoop and HBase services

1) **MySQL:**
Foremost step is to load input CSV file in mysql database. Originally downloaded row data contain number of missing values, unwanted characters and spatial characters which are remove in MySQL database. After this clean data was stored in MySQL database using the MySQL load command. The main reason behind to store input data into mysql database, it provides the features to clean data sets into database, execution time is less and capable to handle the large volume of data.

```
mysql> use h1b;
Database changed
mysql> show tables;
+---------------+
| Tables_in_h1b |
+---------------+
| h1b_data      |
| h1b_visa_data |
+---------------+
2 rows in set (0.00 sec)

mysql>  DELETE FROM h1b_data WHERE year = 0;
Query OK, 0 rows affected (7.20 sec)

mysql> select count(*) from h1b_data;
+----------+
| count(*) |
+----------+
|  3002372 |
+----------+
1 row in set (6.55 sec)
```

Fig. 3. MySQL commands

2) **Sqoop:**
The data is retrieved from the mysql database by making use of Sqoop import command directly in to hive, pig and MapReduce. Following figure 4, figure 5 and figure 6 show the sqoop command which is uesd for importing data.

```
hduser@jyoti-VirtualBox:/usr/local/sqoop$ sqoop import --connect jdbc:mysql://127.0.0.1/h1b --username root --password root --autoreset-to-one-mapper --table h1b_data --hive-import --hive-overwrite --hive-table h1b_data
```

Fig. 4. Importing Table into Hive

```
hduser@jyoti-VirtualBox:/usr/local/sqoop$ sqoop import --connect jdbc:mysql://localhost/h1b --username root --password root --table h1b_visa_data --m 1
```

Fig. 5. Importing Table into Pig

```
hduser@jyoti-VirtualBox:/usr/local/sqoop$ sqoop import --connect jdbc:mysql://127.0.0.1/h1b --username root --password root --table h1b_data -m 1 --columns case_status--target-dir /user/hduser/h1b_table/Mapreduce/test_column/
```

Fig. 6. Importing column into MapReduce

3) **Hive:**
In hive two task were performed by using following command. External table were created to store the out put of this query in HBase table. figure 8 Various functions like group by, limit, filter are to perform the different task.

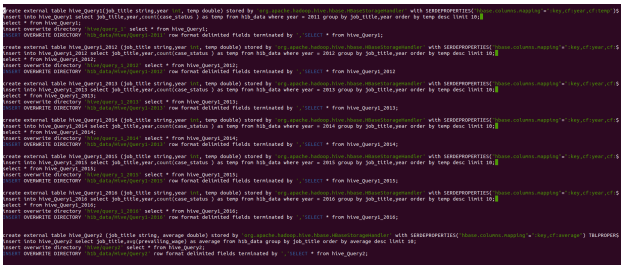

Fig. 7. Hive Query

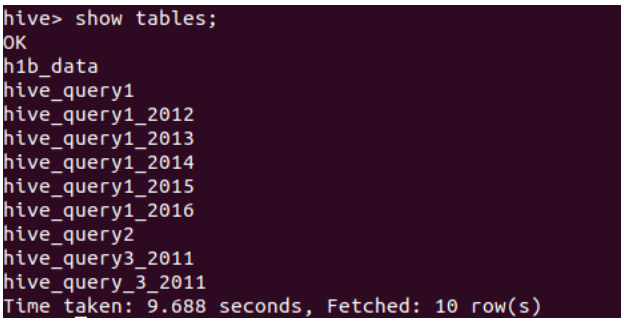

Fig. 8. Hive Table

4) **Pig:**
In pig one task was performed and similarly output of the task was stored in to the HBase as shown in figure 9.

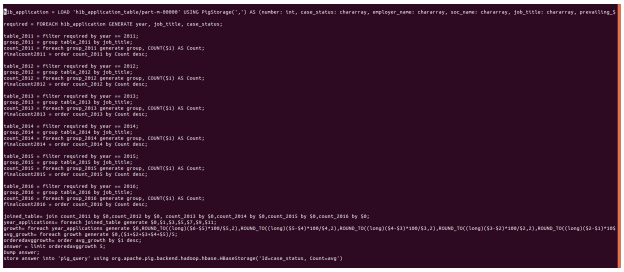

Fig. 9. Pig Task

5) **MapReduce:**
To perform the MapReduce task java program was written in eclipse in which three MapReduce task was performed. After creating the required claases extrenal jar file was added which include Hadoop jar files. Before running the java code jar is developed. By using this jar file code was run in terminal. In this data is read from the hdfs and output is also stored in the hdfs. Similar way other task was execute in the eclipse as shown in figure 10.

Fig. 10. MapReduce Task

6) **HBase:**
All the output of tasks are store in HBase where table are created to store the data. figure 11 shows the ouptput table of Hive and Pig. and all the output is also stored in the hdfs in txt form. this data is further used for visualization. For visualizing the data tableau was used.

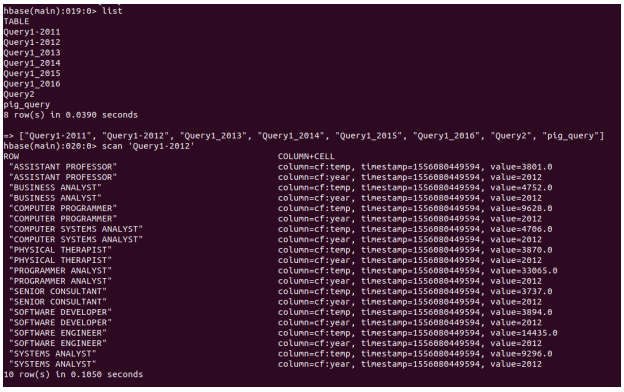

Fig. 11. HBase Tables

IV. Experiment and Result

A. *Query-1*

The total number of records for different case status that is how many applications is certified or denied? Java classes have been implemented for mapper, driver, and reducer. The main objective is to analyze the status of the H-1B visa application from 2011 to 2016. In mapper class, the variable was defined where case status is considered as a token variable. In IntSumReducer, for each status row has been recorded and stored in sum variable and count of unique status was written. Fig12 shows the certified application is 26,15,623 which is

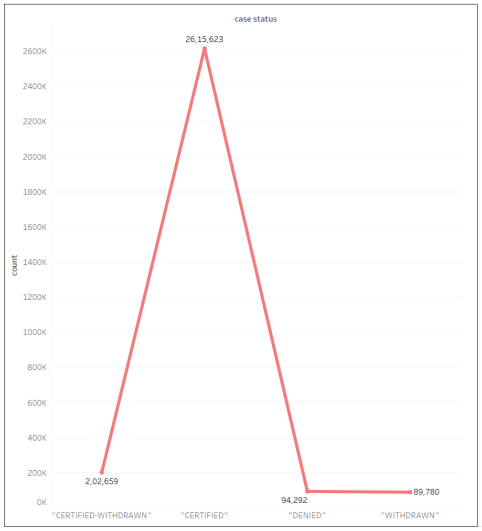

Fig. 12. Count of Case Status

higher that is almost 2milion workers are eligible for an H1B

visa. And only 94 thousand visa application is rejected because of various reasons from 2011 to 2016. There is a number of people who willing to work in the united states. This suggests that there is a number of people who are willing to work in the united states.

B. Query-2

Find the most popular job position for H-1B visa application in each year? By using the Hive, this query is implemented. In this data the MySQL database is loaded directly. Various conditions are used to find out the population job positions in United State. Group by a function are used to group the year wise data further this it is combined for the visualization purpose. Furthermore external table was created to store the hive output in HBase. The fig.2 gives information about the

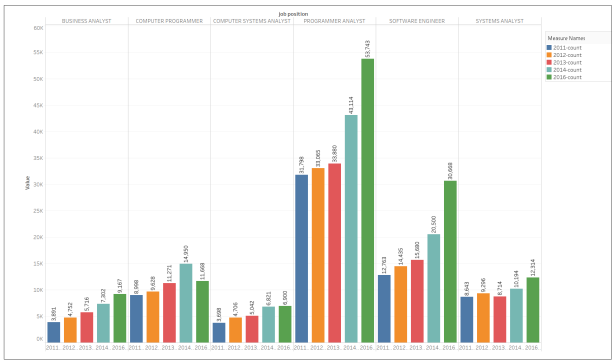

Fig. 13. Job Title Growth in Application

job position trend in the market. People who applied for H1B visa for computer system analyst are less as compared to others and the ratio of programmer analyst are higher. That means opportunity in programmer analyst is more and it increases each year. This visualization gives a clear idea about the market trend in the US.

C. Query-3

Which job title has the highest growth in H1b Visa application? By using the Hive, this query is implemented. In this data the MySQL database is loaded directlye. Firstly, data is filtered by each year separately and stored in different variables, each variable is joint and group by year for the count. By making use of this average growth of job title in the application is determined. As shown in fig14 Senior system analyst job title is in trending among the others. A number of a senior system analyst who applied for the visa application. This indicates that senior system analyst is in demand in the United States job market rapidly increased among others.

D. Query-4

Is the number of petitions with data engineering title increases over the time? In this case provide the average number of data engineering count which increase over the time. For this query java classes are implemented which include diver, mapper and reducer Mapper defines the columns that contain

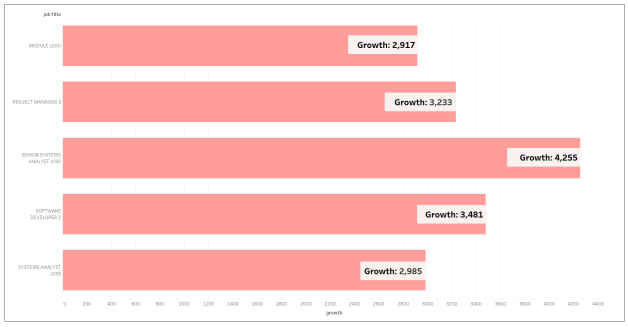

Fig. 14. popular job position in Application

data engineer title as string and year as double. For filtering out the data engineering job and condition is applied to both the column. If condition is true, then only further analysis is processed. In reducer year is stored in array index. At the starting sum is defined as zero. If value is less then array index then only count is written. Fig.4 show trend of the data engineering job over the year. The demand of the data engineering is highest in 2016and less in 2012.

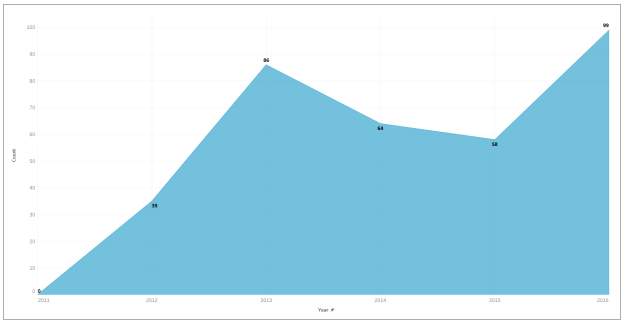

Fig. 15. Average data engineering job

E. Query-5

Which part of the United State has a more number of data engineering job each year? To perform this query MapReduce is used. Which contain four different java classes in which mapper class is used to read the data block and processed as the intermediate output which generate key-value pair. this output is given to reducer as input. Here, data partitioned class is implemented which classifies each year and provide the result of each year. Fig16 shows that location in the United State where data engineering job is available. Seattle-Washington has a greater number of data engineering jobs and Tallahassee-Florida has less number of job available. According to this foreign employer target the Washington and New York state in the United State for their work.

F. Query-6

Find the average prevailing wage for each job? This query is processed through hive in which powerful job analysis and summarization are done using big data. It supports SQL like

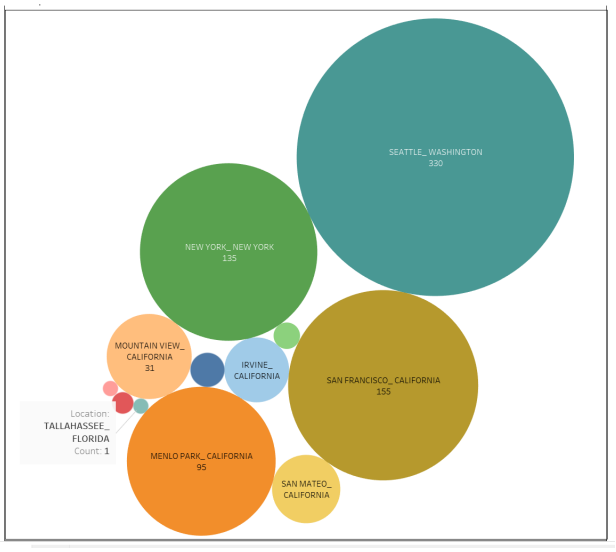

Fig. 16. popular location for data engineering job

queries that automatically translate into MapReduce jobs while running on the Hadoop platform. To fulfill this query average function are used for calculating the average wage and group by job title. This will consider all year and combine them. Fig17 gives information about the average wage of the job in

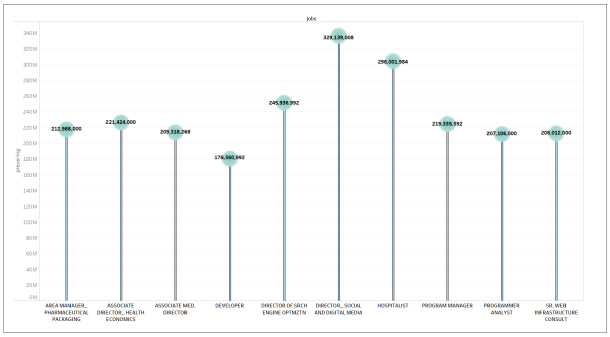

Fig. 17. Growth in prevailing wage

the United State. Wage are differing for every job title. Senior system analyst has high growth in prevailing wage and module lead has a comparatively low prevailing wage.

V. Conclusion

In order to identify the pattern and best times of the year and top geographical region in job search and hiring methods. This project represents the implementation of the MapReduce programming framework with a different design pattern, apache hive, and apache pig are applied on the H1B visa application dataset. After applying the big data techniques, it gives a clear idea to external employers to analyze available employment opportunity on the United States business market and keep track of the number of approved visas each year. Relational database like MySQL, in which input data is store and HBase is used to store the output of the query.

References

[1] P. Kale and S. Balan, "Big data application in job trend analysis," *2016 IEEE International Conference on Big Data (Big Data)*, pp. 4001–4003, 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7841089h/

[2] M. Boyd and S. Tian, "Is stem education portable? country of education and the economic integration of stem immigrants," *Journal of International Migration and Integration*, vol. 19, no. 4, pp. 965–1003, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/imig.12302

[3] A. Gupta, M. Saxena, and R. Gill, "Performance analysis of rdbms and hadoop components with their file formats for the development of recommender systems," *2018 3rd International Conference for Convergence in Technology (I2CT)*, pp. 1–6, 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8529480

[4] M. R. K. M. N. S. Rajeshi Tanwar, Prof. Kailash Patidar, "Bigdata analysis using hadoop ecosystems on cloud plateform," *JASC: Journal of Applied Science and Computations*, vol. 5, pp. 170–178, 2018.

[5] A. Grimm, "Studying to stay: Understanding graduate visa policy content and context in the united states and australia," *International Organization of Migresion*, 2019.

[6] Y. He, R. Lee, Y. Huai, Z. Shao, N. Jain, X. Zhang, and Z. Xu, "Rcfile: A fast and space-efficient data placement structure in mapreduce-based warehouse systems," pp. 1199–1208, April 2011.