**H1-B Case Study**

The H1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for H1B visa, an US employer must offer a job and petition for H1B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college/ higher education (Masters, Ph.D.) and work in a full-time position.

We will be performing analysis on the H1B visa applicants between the years 2011-2016

The dataset has nearly 3 million records.

The dataset description is as follows:

The columns in the dataset include:

* CASE\_STATUS: Status associated with the last significant event or decision. Valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”.

1. Certified: Employer filed the LCA, which was approved by DOL
2. Certified Withdrawn: LCA was approved but later withdrawn by employer
3. Withdrawn: LCA was withdrawn by employer before approval
4. Denied: LCA was denied by DOL

* EMPLOYER\_NAME: Name of employer submitting labour condition application.
* SOC\_NAME: the Occupational name associated with the SOC\_CODE. SOC\_CODE is the occupational code associated with the job being requested for temporary labour condition, as classified by the Standard Occupational Classification (SOC) System.
* 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 labour condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer’s minimum requirements for the position.
* YEAR: Year in which the H1B visa petition was filed
* WORKSITE: City and State information of the foreign worker’s intended area of employment
* lon: longitude of the Worksite
* lat: latitude of the Worksite

PROJECT OUTLINE

|  |  |
| --- | --- |
| **Title** | **Big Data Analysis in Hadoop on H1-B Visa** |
| **Input** | H1B Applications |
| **Data Elements** | S\_No,Case\_Status,Employer\_Name,Soc\_Name,Job\_Title,Full\_Time\_Position,Prevailing\_Wage,Year,Work\_Site,Latitude,Longitude |
| **Analysis Relevance** | Applications between 2011-2016,Maximum petitions filed in year, Top job positions for H1B,Industry(Soc\_Name) having highest job position in H1B application, Part of US having highest job title, Average Prevailing\_Wage for each job in per year, Success\_rate for particular employer in H1B,Success\_rate for particular job position |
| **Purpose** | To provide analysis on top job positions for H1B applications, Work location having maximum applications, Top Employers having highest applicants, Success\_rate of applications applied |
| **Methodology** | Agile |

PROJECT IMPLEMENTATION

**Assumptions:**

1. Hadoop Cluster is running
2. Ecosystem Products (Pig, Hive, Sqoop) are installed
3. MySQL Database is installed
4. H1B Visa Data available on HDFS in ZIP format

**Steps to convert into CSV format**

Step 1: Using tar command extract Zip file

Step 2: Create a table h1b\_applications to read data from extracted file h1b.csv

Step 3: Create a table h1b\_app2 to load data from h1b\_applications after filtering unnecessary data like NA, ’,’ ,’:’ ,’\”’ ,’~’ ,’\t’ ,’.’ ,’!’

Step 4: Create a table h1b\_final to load data from h1b\_app2 after converting prevailing wage and case status to correct data

1 a) Is the number of petitions with Data Engineer job title increasing over time?

Technology used: MapReduce job in Java

Implementation Steps

1. H1B data read line by line in map() method and number of petitions and year are retrieved for DataEngineer.
2. Year and number of petitions for each year values are passed to Reducer.
3. Reducer calculates average growth of DataEngineer for each year and generates output.

1 b) Find top 5 job titles who are having highest avg growth in applications

Technology used: Pig

Implementation Steps

1. Register hive jar files using register command
2. Load data from /user/hive/warehouse/niit\_h1b/h1b\_final into bag data1
3. Filter bag data1 on year 2011
4. Group filter data on job\_title
5. Count applications for that job\_title and year
6. Repeat the steps for years 2012, 2013, 2014, 2015, 2016
7. Join the data for all years 2011, 2012, 2013, 2014, 2015, 2016 using common joining key
8. Calculate average growth from joined data
9. Arrange the result in descending order to get top 5 job\_title
10. Store the result in hdfs file format

2 a) Which part of the US has the most Data Engineer jobs for each year?

Technology used: MapReduce job in Java

Implementation Steps

1. H1B data read line by line in map () method and worksite and year are retrieved for DataEngineer.
2. Year and worksite values are passed to Partitioner.
3. Partitioner partitions data year wise and sends partitioned year and worksite values to Reducer.
4. Reducer calculated number of DataEngineer jobs in each worksite for each year.
5. The result data is sorted in descending order to get top records for each year.

2 b) find top 5 locations in the US who have got certified visa for each year. [Certified]

Technology used: Hive

Supporting Table:

question2b

|  |  |  |
| --- | --- | --- |
| Column Name | Column DataType | Remarks |
| Worksite | string | Work location from US |
| Count | bigint | Number of applicants |
| Year | string | Year applied |

Implementation Steps

1. Retrieve worksite, number of applicants and year from h1b\_final table for each year and certified applications only.
2. Sort the data in descending order
3. Load the data into question2b table

3) Which industry (SOC\_NAME) has the most number of Data Scientist positions?[certified]

Technology used: MapReduce job in Java

Implementation Steps

1. H1B data read line by line in map() method and industry (Soc\_Name) and year are retrieved for Data Scientist and certified applicant.
2. Year and Soc\_Name values are passed to Reducer.
3. Reducer calculates number of DataScientist jobs in each Soc\_Name for each year.
4. The result data is sorted in descending order to get top records for each year.

4) Which top 5 employers file the most petitions each year? - Case Status - ALL

Technology used: MapReduce job in Java

Implementation Steps

1. H1B data read line by line in map () method and Employer\_Name, number of petitions and year are retrieved.
2. Year and Employer\_Name values are passed to Partitioner.
3. Partitioner partitions year wise and send partitioned values to Reducer.
4. Reducer calculates number of Petitions filed for each Employer for each year.
5. The result data is sorted in descending order to get top records for each year

5) Find the most popular top 10 job positions for H1B visa applications for each year?

a) for all the applications

b) for only certified applications.

Technology used: Hive

Supporting Table:

question5a

|  |  |  |
| --- | --- | --- |
| Column Name | Column DataType | Remarks |
| Job\_title | string | Job position |
| Year | string | Applicant year |
| Count | bigint | Number of applications for that job per year |

Implementation Steps

1. Retrieve job\_title, number of applicants and year from h1b\_final table for each year and job\_title
2. Sort the data in descending order
3. Load the data into question5a table

question5b

|  |  |  |
| --- | --- | --- |
| Column Name | Column DataType | Remarks |
| Job\_title | string | Job position |
| Year | string | Applicant year |
| Case\_Status | string | Certified Case\_Status |
| Count | bigint | Number of applications for that job per year |

Implementation Steps

1. Retrieve job\_title, number of applicants, Case\_Status and year from h1b\_final table for each year and job\_title and certified applicants only
2. Sort the data in descending order
3. Load the data into question5b table

6) Find the percentage and the count of each case status on total applications for each year. Create a line graph depicting the pattern of All the cases over the period of time.

Technology used: Pig

Implementation Steps

1. Register hive jar files using register command
2. Load data from /user/hive/warehouse/niit\_h1b/h1b\_final into bag data1
3. Filter bag data1 on not null values for Case\_Status
4. Group filter data on Year
5. Count total applications on Year
6. Filter bag data1 on not null values for Year
7. Group filter data on Year and Case\_Status
8. Count total applications on Year and Case\_Status
9. Join the data using common joining key
10. Calculate percentage from joined data
11. Store the result in hdfs file format
12. Create line graph on result

7) Create a bar graph to depict the number of applications for each year [All]

Technology used: Hive

Implementation Steps

1. Execute the query to calculate number of applications for each year

Select year, count(\*) from h1b\_final group by year order by year;

1. Create bar graph on result

8) Find the average Prevailing Wage for each Job for each Year (take part time and full time separate). Arrange the output in descending order - [Certified and Certified Withdrawn.]

Technology used: Hive

Supporting Table:

question8

|  |  |  |
| --- | --- | --- |
| Column Name | Column DataType | Remarks |
| Job\_title | string | Job name |
| Full\_time\_position | string | Full-time or Part-time job |
| Year | string | Year |
| Avg\_Wage | bigint | Calculated Prevailing\_Wage average |

Implementation Steps

1. Retrieve job\_title, full\_time\_position, Case\_Status and year from h1b\_final table for each year and job\_title and full-time and certified and certified-withdrawn applicants only
2. Calculate average on Prevailing\_Wage for job\_title, year and full\_time\_position
3. Sort average in descending order
4. Repeat the steps a-c for part-time position
5. Repeat the steps a-d for years 2012,2013,2014,2015,2016
6. Load the data into question8 table

9) Which are the employers along with the number of petitions who have the success rate more than 70% in petitions? (total petitions filed 1000 OR more than 1000) ?

Technology used: Pig

Implementation Steps

1. Register hive jar files using register command
2. Load data from /user/hive/warehouse/niit\_h1b/h1b\_final into bag data1
3. Filter bag data1 on not null values for Case\_Status
4. Group filter data on Employer\_Name
5. Count total applications on S\_No
6. Filter data for certified applicants
7. Group filter data on Employer\_Name
8. Count total certified applications on group data
9. Filter data for certified-withdrawn applicants
10. Group filter data on Employer\_Name
11. Count total certified-withdrawn applications on group data
12. Join the data using common joining key
13. Retrieve necessary columns from joined data
14. Calculate success\_rate from joined data
15. Filter data where success\_rate > 70 and number of petitions > 1000
16. Sort the result in descending order
17. Store the result in hdfs file format

10) Which are the job positions along with the number of petitions which have the success rate more than 70% in petitions (total petitions filed 1000 OR more than 1000)?

Technology used: Pig

Implementation Steps

1. Register hive jar files using register command
2. Load data from /user/hive/warehouse/niit\_h1b/h1b\_final into bag data1
3. Filter bag data1 on not null values for Case\_Status
4. Group filter data on Job\_title
5. Count total applications on S\_No
6. Filter data for certified applicants
7. Group filter data on Job\_title
8. Count total certified applications on group data
9. Filter data for certified-withdrawn applicants
10. Group filter data on Job\_title
11. Count total certified-withdrawn applications on group data
12. Join the data using common joining key
13. Retrieve necessary columns from joined data
14. Calculate success\_rate from joined data
15. Filter data where success\_rate > 70 and number of petitions > 1000
16. Sort the result in descending order
17. Store the result in hdfs file format

11) Export result for question no 10 to MySQL database.

Technology used: MySQL, Sqoop

Implementation Steps

1. Create database h1b in MySQL
2. Drop table question11 if it already exists in database h1b
3. Create table question11 in h1b database
4. Using Sqoop export command load the result from hdfs file format into question11 table in h1b database in MySQL
5. Display the result from question11 table