**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 2011-2015. After analyzing the data, we can derive the following facts.

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

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

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

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

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

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

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.

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.

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

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.

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

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 more than 1000)?

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

SUCCESS RATE % = (Certified + Certified Withdrawn)/Total x 100

Bottom of Form

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 Valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”.

Certified: Employer filed the LCA, which was approved by DOL

Certified Withdrawn: LCA was approved but later withdrawn by employer

Withdrawn: LCA was withdrawn by employer before approval

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

In the data, few columns are enclosed by double quotes and also we have comma’s in a single column and the column is enclosed by double quotes. So we have used hive csv serve to load the data. In the quoteChar, we have given **“(**double quote**).**So this will take the column value in between the double quotes.

Let’s create a table to load the h1b applicant’s data as shown below.

|  |  |
| --- | --- |
|  | CREATE TABLE h1b\_applications(s\_no int,case\_status string, employer\_name string, soc\_name string, job\_title string, full\_time\_position string,prevailing\_wage int,year string, worksite string, longitute double, latitute double )    ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'    WITH SERDEPROPERTIES (    "separatorChar" = ",",    "quoteChar" = "\""    ) STORED AS TEXTFILE; |

Use all the following tools

HDFS

MapReduce - any 4 programs

Hive - any 4 programs

Pig - any 4 programs

Sqoop

Mysql

Excel

Project should Menu based using shell script in Unix

**Case Status**

1.certified

2.certified withdrawn

3. denied

4. withdrawn

5 pending quality and compliance review

6. invalidated

7 rejected

8.NA