**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?**

Select count(\*) as kill,year from h1b where job\_title="DATA ENGINEER" group by job\_title,year order by kill ;

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

Select count(\*) as kill,job\_title from h1b group by job\_title order by kill desc limit 5;

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

select war4,m.kill,m.year from(select max(thop) as war4,b.year as req from (Select count(\*) as thop,year,worksite as kill from h1b where job\_title="DATA ENGINEER"group by year,worksite) b group by b.year)n join (select count(\*) as thop1,year,worksite as kill from h1b where job\_title="DATA ENGINEER" group by year,worksite)m on n.req=m.year and n.war4=m.thop1 order by year;

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

Select \* from( Select rank() over (partition by year order by kill desc) as thop,worksite,year from (Select count(\*) as kill,worksite,year from h1b where case\_status="CERTIFIED" or case\_status="CERTIFIED-WITHDRAWN" group by worksite,year ) a ) n where thop<6;

**3)Which industry has the most number of Data Scientist positions?**

Select soc\_name,count(\*) as war1 from h1b where job\_title="DATA SCIENTIST" group by soc\_name order by war1 desc limit 1 ;

**4)Which top 5 employers file the most petitions each year?**

select thop,employer\_name,year from(select rank() over (partition by year order by war1 desc) as thop,employer\_name,year from(Select count(\*) as war1,employer\_name,year from h1b group by employer\_name,year) a)b where thop<6 and year is not null;

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

select \* from(Select rank() over (partition by year order by war desc) as thop,job\_title,year from (Select count(\*) as war,year,job\_title from h1b group by job\_title,year ) a)b where b.thop<11 and year is not null;

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

Select case\_status,year,(maharshi\*100)/a.thop from (Select count(\*) as maharshi ,case\_status,year from h1b group by case\_status,year) b,(select count(\*) as thop from h1b) a order by case\_status;

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

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

**8) Find the average Prevailing Wage for each Job for each Year (take part time and full time separate)**

select avg(prevailing\_wage),job\_title,year from h1b group by year,job\_title,FULL\_TIME\_POSITION order by job\_title;

**9) Which are top ten employers who have the highest success rate in petitions?**

Select employer\_name,(kill\*100)/kill1 as top10 from (select count(\*) as kill1,1 as tojoin from h1b)b join (Select count(\*) as kill,1 as tojoin,employer\_name from h1b where case\_status="CERTIFIED" or case\_status="CERTIFIED WITHDRAWN" group by employer\_name) a on a.tojoin=b.tojoin order by top10 limit 10;

**10) Which are the top 10 job positions which have the highest success rate in petitions?**

Select job\_title,(thop\*100)/kill1 as top10 from (Select count(\*) as thop,job\_title from h1b where case\_status=” CERTIFIED “ or case\_status=”CERTIFIED WITHDRAWN” group by job\_title )b,(Select count(\*) as kill1 from h1b) a order by top10 limit 10;

**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