CS 185C – Introduction to Big Data Analytics

H-1B Visa Data Analysis

Submitted to : Prof. Peter Zadrozny

Submitted by : Team#2

Sailee Choudhary

Tin Hon Ng

Anumeha Umang Shah

Date: 10/22/2014

Contents

Introduction

Description

The H-1B program allows employers to temporarily employ foreign workers in the U.S. on a nonimmigrant basis in specialty occupations of distinguished merit and ability. A specialty occupation requires the theoretical and practical application of a body of specialized knowledge and a bachelor's degree or the equivalent in the specific specialty (e.g. sciences, medicine, health care, education, biotechnology, and business specialties, etc.). Current laws limit the annual number of qualifying foreign workers who may be issued a visa or otherwise be provided H-1B status to 65,000.

With an in-depth analysis of these data sets, we can get an idea of potential professions and potential work areas within the States. The data provides an insight into job market statistics for past eight years which can be very useful information for people from different professions.

Objective

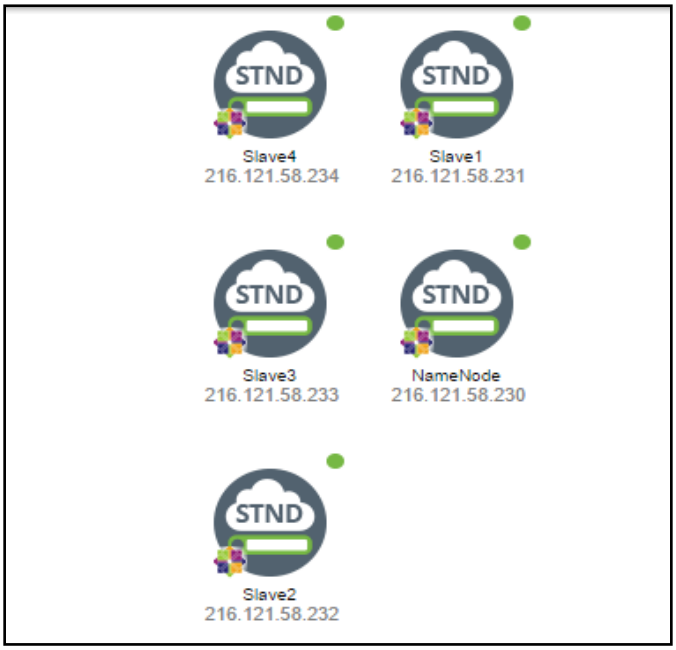
Using H-1B Visa data set(s) we intend to analyze the job market statistics from 2006 to 2013 with the objective of speculating the co-relation between professions, employers, dates, salaries and locations. We use Hadoop and Hive in order to perform this analysis. The main objective is to load data into Hadoop and use Hive to trigger appropriate queries in order to arrive at suitable conclusions that are relevant and useful to others.

Cluster Description

The Hadoop cluster comprises of 5 servers, one master node and four slave nodes. The master node is extra large server with size of 8GB and 4 slave nodes each of size 4 GB.

OS : CentOS

Architecture : 64-bit



Credentials

Password for root : Noeasywayout203

Port: 8000

To login to Cloudera Manager : http://216.121.58.230:7180/cmf/login

Hostname	Public IP	Private IP
master.bigdata.com	216.121.58.230	192.168.1.101
slave1.bigdata.com	216.121.58.231	192.168.1.102
slave2.bigdata.com	216.121.58.232	192.168.1.103
slave2.bigdata.com	216.121.58.233	192.168.1.104
slave3.bigdata.com	216.121.58.234	192.168.1.105

Phase 1 - Loading and Verifying Data

Dataset Description

The issuing of visas for highly skilled workers has been a topic of debate recently in the United States. The H-1B is a non-immigrant visa in the United States under the Immigration and Nationality Act, section 101(a)(15)(H). It allows U.S. employers to temporarily employ foreign workers in specialty occupations.

The dataset consists of data spread over the years from 2006 to 2013. It provides an insight about each applicant who filed for H-1B, his status of visa, employer who sponsored the visa. The dataset is in comprehensible format, and so an in-depth analysis is possible.

Where we obtained the data

The data fields have been extracted from Office of Foreign Labor Certification (OFLC) applications tables for the years 2006 to 2013.

The data was available at:

https://app.enigma.io/search/source/us.gov.dol.oflc.h1b

How we obtained the data

Immigration reform is on many people's minds and especially a hot topic in the technology industry as tech executives lobby hard to bring more international talent and make the process to hire international talent easier. Mark Zuckerberg has even weighed in on the topic saying that "immigration reform and visa reform are necessary to boost the U.S. economy and job market." With that notion in mind, we wanted to see the visa data analyzed so we could come up with our own conclusions. With the help of Enigma.io, we obtained CSV files of H-1B data stats of 2006 to 2013 from the Department of Labor.

Issues with obtaining the data

We used all the publicly available datasets. It required no registration or identification to access, and had no legal restrictions on their use.

Structure and Key Fields of the data set(s)

Structure of Data : comma delimited text format (.csv)

Key Fields of Data sets:

Field Name	Description
submitted_date	Application submission date.
case_no	Unique case no. of each filed application.
employer_name	Employer's Name
employer_city	Employer's City
employer_state	Employer's State
employer_postal_code	Employer's Postal code.
total_workers	Total number of foreign workers being requested for temporary labor certification
employment_start_date	Proposed beginning date of employment.
employment_end_date	Proposed ending date of employment.
job_title	Job title
dol_decision_date	Date on which the last significant event or decision was recorded by the ETA National Processing Center
status	Status associated with the last significant event or decision. Valid values include Certified, Certified-Withdrawn, Denied, and Withdrawn□□
wage_rate	Employer's proposed wage rate
wage_rate_unit	Unit of pay for proposed wage rate
workloc1_city	Address city of the intended are in which the

	foreign worker is expected to be employed
workloc1_state	Address state of the intended are in which the foreign worker is expected to be employed
pw_1	Prevailing wage rate for location
pw_unit_1	Prevailing wage unit for location - hour/year
visa_application_year	The year in which the application was filed

Issues with the data set(s)

1. For year 2006, the year format is “month/date/year” which is different format from the rest of the years which have the format “month-date-year”. In order to have all the dates in same format, we wrote a short python script to copy the original file to a new file with the “/“ replace by “-“.

2. The field names in each data file were different. So, we had to synchronize all the fields and come up with same names for the fields across all the tables.

3. The employer name had comma "," in some records. In order to avoid the incorrect parsing, we had to change the delimiter.

Pre-processing (ETL) of the data

The main aim for preprocessing the data was to synchronize all the columns of every CSV file from 2006 to 2013 and combine them into one big CSV file. Since every file for each year has different number of columns, we had to sync all the columns before appending them into a big CSV file.
The overview of steps that we followed are:

1. import CSV files into phpmyadmin.
2. manipulate the columns to make them the same for each table.
3. export those tables back to CSV files.

Steps:

1. For year 2006, the year format is “month/date/year” which is different format from the rest of the years which have the format “month-date-year”. As a result, we wrote a short python script to copy the original file to a new file with the “/“ replace by “-“.

```
import fileinput
import re

#execute this python script like this:  python process.py inputfile.csv
f = open('h1b_table_2006_3.csv', 'w') #creating a outputfile called h1b_table_2006_3.csv
for line in fileinput.input(): #traverse every line of the file
        line = re.sub(r'([0-9]+)/([0-9]+)/([0-9]+)', r'\1-\2-\3', line.rstrip()) #replace mm/dd/yyyy with mm-dd-yy
        f.write(line + '\n') #write each line ended with newline character into the output file
f.close()
```

2. Splitting the CSV file in three parts for each year :

phpmyadmin restricts the size of the file that can be uploaded. The files in our dataset exceeded the maximum size that phpmyadmin allowed. As a result, we had to split them into smaller files so that the system could handle the upload and we won’t therefore get the timeout error and lose some part of the data.

To perform this, we used the following linux command:

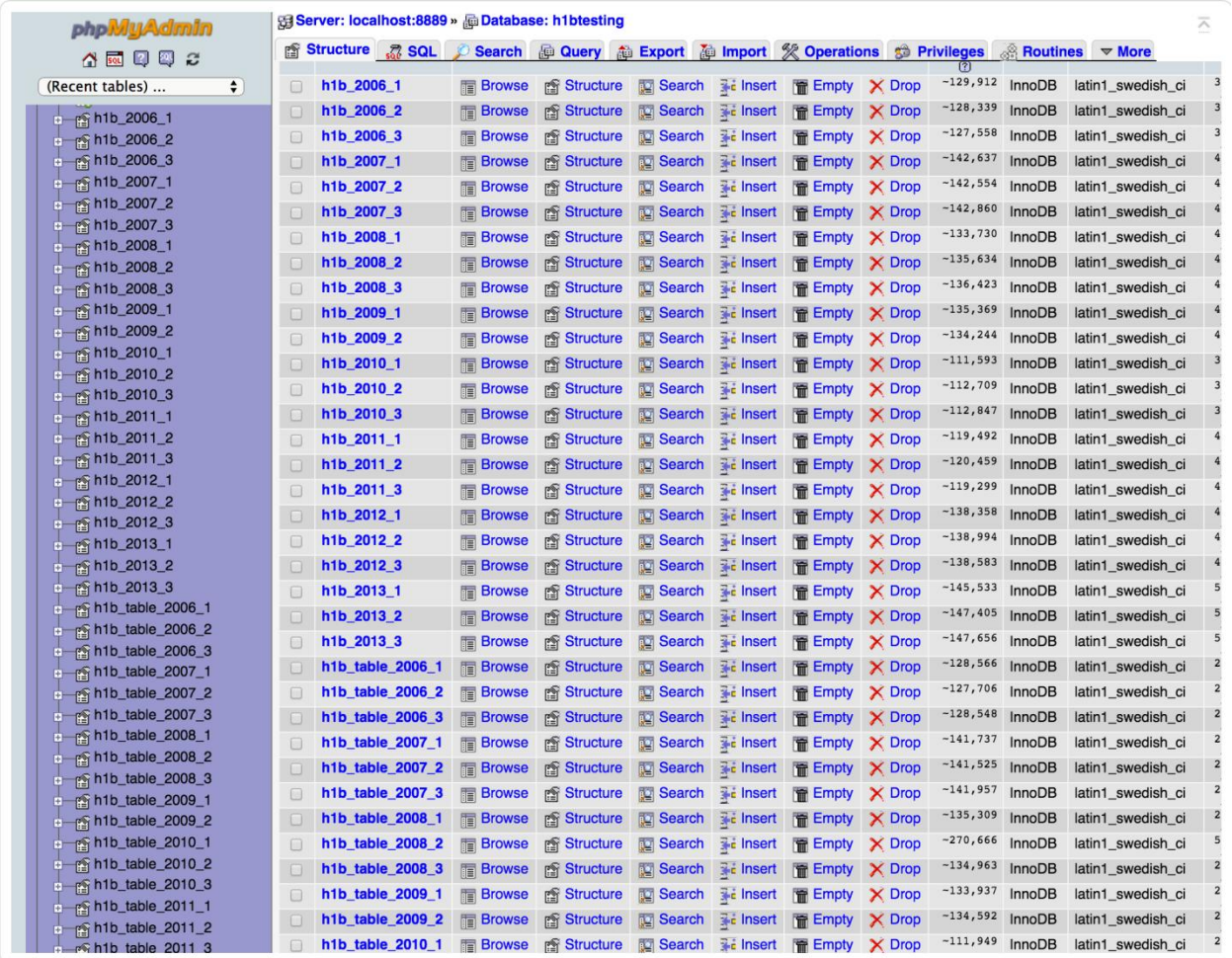
$ split -l number_of_lines filename

For every file, we split them into three parts as we as create three tables. For example for we create h1b_2010_1, h1b_2010_2, h1b_2010_3 for one year and they are the location for those uploaded CSV files.

3. Altering tables

We altered the tables to fit the final columns.

After the above steps, we have a list of tables h1b_20** for holding the original columns, and we also have a list of tables h1b_table_20** for holding the final decided schema. Here is a screen shot showing part of the two lists of tables:



4. Exporting tables to CSV files

After altering all the tables, we exported them as pipe "|" delimited CSV file. The reason for changing the delimiter was that the employer name had comma "," in some records. In order to

avoid the incorrect parsing, we decided to use some other unique separator that was not included in any of the field's records and therefore we decided "|".

5. Combing the CSV files together as a big file

This step involved combining of all the CSV files into one big CSV file so that hive only loads the input file once and creates a table representing all the years of h1b application.

$ cat *.csv > newfilename.csv

For our case : cat *.csv > h1b_applications.csv

Loading the data

1. For loading the data into Hive, first we created a database.

```
$ CREATE DATABASE h1b_visa;
```

2. After creating the database, we created an external table h1b_application. The EXTERNAL keyword lets you create a table and provide a LOCATION so that Hive does not use a default location for this table. This comes in handy if you already have data generated. When dropping an EXTERNAL table, data in the table is NOT deleted from the file system. We created table using following Hive query:

```
CREATE EXTERNAL TABLE h1b_application
(
submitted_date struct<year:INT,month:INT,date:INT>,
case_no STRING,
employer_name STRING,
employer_city STRING,
employer_state STRING,
employer_postal_code STRING,
total_workers INT,
employment_start_date struct<year:INT, month:INT, date:INT>,
employment_end_date struct<year:INT, month:INT, date:INT>,
job_title STRING,
dol_decision_date struct <month:INT, date:INT, year:STRING>,
status STRING,
wage_rate INT,
wage_rate_unit STRING,
workloc1_city STRING,
workloc1_state STRING,
pw_1 INT,
pw_unit_1 STRING,
visa_application_year STRING
```

```
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|'
COLLECTION ITEMS TERMINATED BY '-'
STORED AS TEXTFILE;
```

3. After creating the table, we moved the CSV file to Hadoop file system using the command:

$ hadoop fs -put h1b_applications.csv

4. We then moved the file from HDFS to the created table in hive directory.

$ hadoop fs -mv h1b_applications.csv /user/hive/warehouse/h1b_visa.db/h1b_application/

Data Verification

We validated the loaded data in four phases: validating whether all the columns have been defined properly, validating the number of records loaded, selecting random records and matching them against the CSV files, extracting and matching the first and last records for each year with those in CSV files .

1. Validating the columns

We used DESCRIBE to check that all the columns has been defined and stored properly.

$ describe h1b_application;

```
hive> describe h1b_application;
OK
submitted_date          struct<year:int,month:int,date:int>     None
case_no                 string                  None
employer_name           string                  None
employer_city           string                  None
employer_state          string                  None
employer_postal_code    string                  None
total_workers           int                     None
employment_start_date   struct<year:int,month:int,date:int>     None
employment_end_date     struct<year:int,month:int,date:int>     None
job_title               string                  None
dol_decision_date       struct<month:int,date:int,year:string>  None
status                  string                  None
wage_rate               int                     None
wage_rate_unit          string                  None
workloc1_city           string                  None
workloc1_state          string                  None
pw_1                    int                     None
pw_unit_1               string                  None
visa_application_year   string                  None
Time taken: 0.104 seconds, Fetched: 19 row(s)
hive>
```

2. Validating the number of records

We used COUNT to count the total no of rows and matched with the total no of lines in the CSV files.

$ select count(*) form h1b_application;

```
hive> select count(*) from h1b_application;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_1411083433145_0073, Tracking URL = http://master.bigdata.com:8088/proxy/application_1411083433145_0073/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1411083433145_0073
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 1
2014-10-06 21:56:17,245 Stage-1 map = 0%,  reduce = 0%
2014-10-06 21:56:24,457 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 1.77 sec
2014-10-06 21:56:25,489 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 5.21 sec
2014-10-06 21:56:26,519 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.83 sec
2014-10-06 21:56:27,552 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.83 sec
2014-10-06 21:56:28,581 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.83 sec
2014-10-06 21:56:29,625 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.83 sec
2014-10-06 21:56:30,674 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.83 sec
2014-10-06 21:56:31,704 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.83 sec
2014-10-06 21:56:32,746 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.83 sec
2014-10-06 21:56:33,801 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.83 sec
2014-10-06 21:56:34,845 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.83 sec
2014-10-06 21:56:35,875 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.64 sec
2014-10-06 21:56:36,908 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.64 sec
MapReduce Total cumulative CPU time: 10 seconds 640 msec
Ended Job = job_1411083433145_0073
MapReduce Jobs Launched:
Job 0: Map: 3  Reduce: 1   Cumulative CPU: 10.64 sec   HDFS Read: 565510017 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 640 msec
OK
3172096
Time taken: 31.981 seconds, Fetched: 1 row(s)
hive>
```

3. Validating random data records

We selected random data to confirm that data has been stored as defined in the create table statement and compared it with the csv file

$ select * from h1b_application limit 10;

Hive shell Output:

```
hive> select * from h1b_application limit 10;
OK
{"year":2006,"month":5,"date":25}       I-06145-2563305 Vensiti Inc.    Irving  TX      75038   1       {"year":2006,"month":10,
"date":1}       {"year":2009,"month":10,"date":1}       Systems Analyst/Consultant      {"month":5,"date":25,"year":"2006 "}    C
ertified        50000   Year    Dallas  TX      45802   NULL    2006
{"year":2006,"month":5,"date":25}       I-06145-2563307 SOFTWARE PUNDITS & ASSOCIATES   BURLINGTON      MA      01803   1       {
"year":2006,"month":5,"date":25}        {"year":2008,"month":9,"date":5}        PROGRAMMER ANALYST      {"month":5,"date":25,"ye
ar":"2006 "}    Certified       70000   Year    BURLINGTON      MA      49483   NULL    2006
{"year":2006,"month":5,"date":25}       I-06145-2563309 DURR SYSTEMS, INC.      Plymouth        MI      48170   1       {"year":
2006,"month":6,"date":20}       {"year":2009,"month":6,"date":19}       SOFTWARE ENGINEER       {"month":5,"date":25,"year":"200
6 "}    Certified       72654   Year    AUBURN HILLS    MI      63211   NULL    2006
{"year":2006,"month":5,"date":25}       I-06145-2563313 Kyocera Wireless Corporation    San Diego       CA      92121   1       {
"year":2006,"month":11,"date":11}       {"year":2009,"month":11,"date":10}      Staff Software Engineer {"month":5,"date":25,"ye
ar":"2006 "}    Certified       96348   Year    San Diego       CA      81286   NULL    2006
{"year":2006,"month":5,"date":25}       I-06145-2563315 LANCESOFT INC   CHANTILLY       VA      20151   1       {"year":2006,"mo
nth":10,"date":1}       {"year":2009,"month":9,"date":30}       PROGRAMMER / ANALYST    {"month":5,"date":25,"year":"2006 "}    C
ertified        48000   Year    CHANTILLY       VA      41200   NULL    2006
{"year":2006,"month":5,"date":25}       I-06145-2563957 US Tech Solutions, Inc  Jersey City      NJ      07302   1       {"year":
2006,"month":10,"date":1}       {"year":2009,"month":9,"date":30}       Programmer Analyst      {"month":5,"date":25,"year":"200
6 "}    Certified       48000   Year    JERSEY CITY     NJ      43964   NULL    2006
{"year":2006,"month":5,"date":25}       I-06145-2563959 Inforeem, Inc.  Edison  NJ      08820   1       {"year":2006,"month":10,
"date":1}       {"year":2009,"month":9,"date":30}       Computer Systems Analyst        {"month":5,"date":25,"year":"2006 "}    C
ertified        44100   Year    EDISON  NJ      44013   NULL    2006
{"year":2006,"month":5,"date":25}       I-06145-2563961 BNP Paribas     New York        NY      10019   1       {"year":2006,"mo
nth":10,"date":1}       {"year":2009,"month":9,"date":30}       Associate       {"month":5,"date":25,"year":"2006 "}    Certifie
d       95000   Year    King of Prussia PA      72134   NULL    2006
{"year":2006,"month":5,"date":25}       I-06145-2563965 SKAYS, INC.     HOUSTON TX      77035   1       {"year":2006,"month":6,"
date":28}       {"year":2008,"month":12,"date":9}       SYSTEMS ACCOUNTANT      {"month":5,"date":25,"year":"2006 "}    Certifie
d       42848   Year    HOUSTON TX      42848   NULL    2006
{"year":2006,"month":5,"date":25}       I-06145-2563967 Perry Mill Supply, Co.  Erie    PA      16501   1       {"year":2006,"mo
nth":10,"date":1}       {"year":2009,"month":10,"date":1}       Programmer Analyst      {"month":5,"date":25,"year":"2006 "}    C
ertified        35000   Year    Erie    PA      32490   NULL    2006
Time taken: 0.087 seconds, Fetched: 10 row(s)
hive>
```

CSV file screenshot:

```
2006-05-25|I-06145-2563305|Vensiti Inc.|Irving|TX|75038|1|2006-10-01|2009-10-01|Systems Analyst/Consultant|5-25-2006 |Certified|
50000|Year|Dallas|TX|45802|NULL|2006
2006-05-25|I-06145-2563307|SOFTWARE PUNDITS & ASSOCIATES |BURLINGTON|MA|01803|1|2006-05-25|2008-09-05|PROGRAMMER ANALYST|5-25-20
06 |Certified|70000|Year|BURLINGTON|MA|49483|NULL|2006
2006-05-25|I-06145-2563309|DURR SYSTEMS, INC.|Plymouth|MI|48170|1|2006-06-20|2009-06-19|SOFTWARE ENGINEER|5-25-2006 |Certified|7
2654|Year|AUBURN HILLS|MI|63211|NULL|2006
2006-05-25|I-06145-2563313|Kyocera Wireless Corporation|San Diego|CA|92121|1|2006-11-11|2009-11-10|Staff Software Engineer|5-25-
2006 |Certified|96348.93|Year|San Diego|CA|81286|NULL|2006
2006-05-25|I-06145-2563315|LANCESOFT INC|CHANTILLY|VA|20151|1|2006-10-01|2009-09-30|PROGRAMMER / ANALYST|5-25-2006 |Certified|48
000|Year|CHANTILLY|VA|41200|NULL|2006
2006-05-25|I-06145-2563957|US Tech Solutions, Inc|Jersey City|NJ|07302|1|2006-10-01|2009-09-30|Programmer Analyst|5-25-2006 |Cer
tified|48000|Year|JERSEY CITY|NJ|43964|NULL|2006
2006-05-25|I-06145-2563959|Inforeem, Inc.|Edison|NJ|08820|1|2006-10-01|2009-09-30|Computer Systems Analyst|5-25-2006 |Certified|
44100|Year|EDISON|NJ|44013|NULL|2006
2006-05-25|I-06145-2563961|BNP Paribas|New York|NY|10019|1|2006-10-01|2009-09-30|Associate|5-25-2006 |Certified|95000|Year|King
of Prussia|PA|72134|NULL|2006
2006-05-25|I-06145-2563965|SKAYS, INC.|HOUSTON|TX|77035|1|2006-06-28|2008-12-09|SYSTEMS ACCOUNTANT|5-25-2006 |Certified|42848|Ye
ar|HOUSTON|TX|42848|NULL|2006
2006-05-25|I-06145-2563967|Perry Mill Supply, Co.|Erie|PA|16501|1|2006-10-01|2009-10-01|Programmer Analyst|5-25-2006 |Certified|
35000|Year|Erie|PA|32490|NULL|2006
2006-05-25|I-06145-2563969|THUMBPLAY, INC|New York|NY|10012|1|2006-10-01|2009-09-30|Director, Marketing & Product |5-25-2006 |Ce
rtified|80000|Year|New York|NY|77438|NULL|2006
2006-05-25|I-06145-2563971|SOLTIUS, INC|EDISON|NJ|08820|1|2006-10-01|2009-09-30|Computer Systems Analyst|5-25-2006 |Certified|44
100|Year|EDISON|NJ|44013|NULL|2006
2006-05-25|I-06145-2563973|INTONE NETWORKS, INC|PARLIN|NJ|08859|1|2006-06-05|2009-06-04|MECHANICAL ENGINEERING|5-25-2006 |Certif
ied|55000|Year|PARLIN|NJ|47611|NULL|2006
2006-05-25|I-06145-2563975|Verizon Data Services Inc.|Coppell|TX|75019|1|2006-06-26|2007-06-25|Systems Engineer|5-25-2006 |Certi
fied|56000|Year|Temple Terrace|FL|51542|NULL|2006
2006-05-25|I-06145-2563977|Satyam Computer Services LTD.|Vienna|VA|22182|1|2006-05-25|2009-05-25|PROGRAMMER ANALYST|5-25-2006 |C
```

4. Validating first and last rows for each year.

We selected the first row and last row for each year from 2006 to 2013. These rows should match to first and last rows of original data. We only provide screenshot results for years 2006 and 2013. However, we performed check for first and last record for all the years.

a. First record for 2006

$ select * from h1b_application where case_no = 'I-06145-2563305';

```
hive> select * form h1b_application where case_no ='I-06145-2563305';
FAILED: ParseException line 1:14 missing FROM at 'h1b_application' near '<EOF>'
hive> select * from h1b_application where case_no ='I-06145-2563305';
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1411083433145_0074, Tracking URL = N/A
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1411083433145_0074
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 0
2014-10-06 22:03:25,583 Stage-1 map = 0%,  reduce = 0%
2014-10-06 22:03:32,802 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.29 sec
2014-10-06 22:03:33,850 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.29 sec
2014-10-06 22:03:34,878 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.29 sec
2014-10-06 22:03:35,914 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.62 sec
2014-10-06 22:03:36,991 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.62 sec
MapReduce Total cumulative CPU time: 13 seconds 620 msec
Ended Job = job_1411083433145_0074
MapReduce Jobs Launched:
Job 0: Map: 3   Cumulative CPU: 13.62 sec   HDFS Read: 565510017 HDFS Write: 162 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 620 msec
OK
{"year":2006,"month":5,"date":25}       I-06145-2563305 Vensiti Inc.    Irving  TX      75038   1       {"year":2006,"month":10,
"date":1}       {"year":2009,"month":10,"date":1}       Systems Analyst/Consultant      {"month":5,"date":25,"year":"2006 "}   C
ertified        50000   Year    Dallas  TX      45802   NULL    2006
Time taken: 23.782 seconds, Fetched: 1 row(s)
hive>
```

b. Last record for 2006

$ select * from h1b_application where case_no = 'I-06145-2563303';

```
hive> select * from h1b_application where case_no ='I-06145-2563303';
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1411083433145_0075, Tracking URL = http://master.bigdata.com:8088/proxy/application_1411083433145_0075/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1411083433145_0075
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 0
2014-10-06 22:04:49,411 Stage-1 map = 0%,  reduce = 0%
2014-10-06 22:04:56,674 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.3 sec
2014-10-06 22:04:57,702 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.3 sec
2014-10-06 22:04:58,739 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.3 sec
2014-10-06 22:04:59,769 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.59 sec
2014-10-06 22:05:00,795 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.59 sec
MapReduce Total cumulative CPU time: 13 seconds 590 msec
Ended Job = job_1411083433145_0075
MapReduce Jobs Launched:
Job 0: Map: 3   Cumulative CPU: 13.59 sec   HDFS Read: 565510017 HDFS Write: 163 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 590 msec
OK
{"year":2006,"month":5,"date":25}       I-06145-2563303 LANCESOFT INC   CHANTILLY       VA      20151   1       {"year":2006,"mo
nth":10,"date":1}       {"year":2009,"month":9,"date":30}       PROGRAMMER / ANALYST    {"month":5,"date":25,"year":"2006 "}   C
ertified        48000   Year    CHANTILLY       VA      41200   NULL    2006
Time taken: 23.829 seconds, Fetched: 1 row(s)
hive>
```

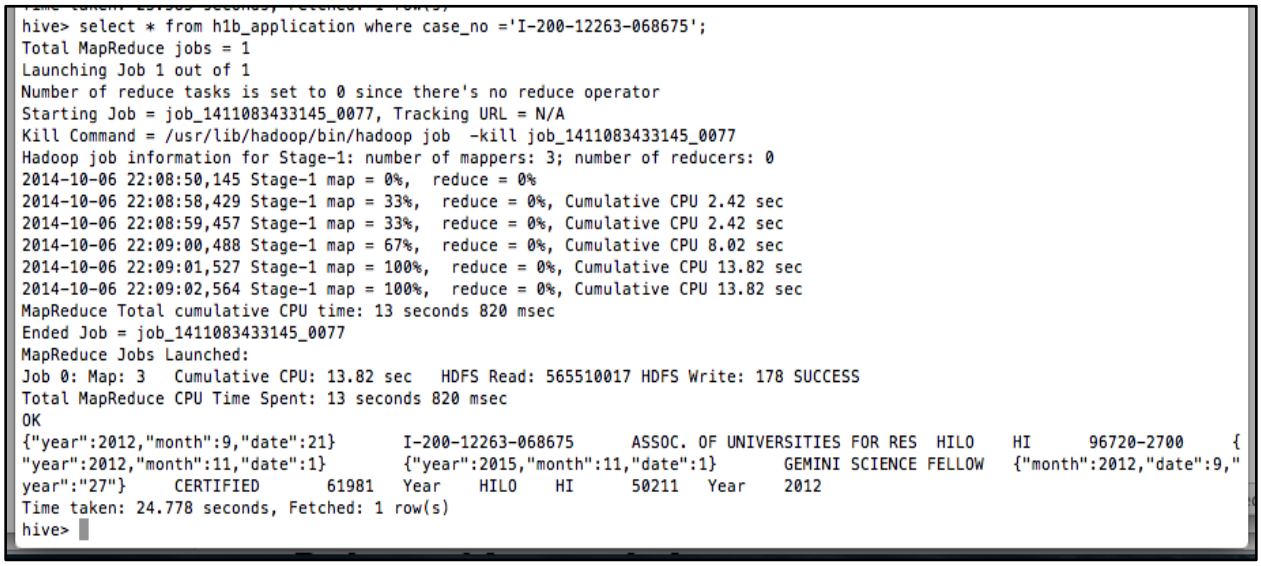
c. First record for 2013

$ select * from h1b_application where case_no = 'I-200-12272-415836';

```
hive> select * from h1b_application where case_no ='I-200-12272-415836';
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1411083433145_0076, Tracking URL = http://master.bigdata.com:8088/proxy/application_1411083433145_0076/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1411083433145_0076
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 0
2014-10-06 22:07:14,370 Stage-1 map = 0%,  reduce = 0%
2014-10-06 22:07:21,596 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.44 sec
2014-10-06 22:07:22,622 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.44 sec
2014-10-06 22:07:23,650 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.44 sec
2014-10-06 22:07:24,710 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 14.02 sec
2014-10-06 22:07:25,742 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 14.02 sec
MapReduce Total cumulative CPU time: 14 seconds 20 msec
Ended Job = job_1411083433145_0076
MapReduce Jobs Launched:
Job 0: Map: 3   Cumulative CPU: 14.02 sec   HDFS Read: 565510017 HDFS Write: 166 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 20 msec
OK
{"year":2012,"month":9,"date":28}       I-200-12272-415836      LAW OFFICE OF XIN MIAO, LLC     FLUSHING        NY      11354   {
"year":2013,"month":2,"date":1} {"year":2016,"month":1,"date":31}       LAW CLERK       {"month":2012,"date":10,"year":"01"}   W
ITHDRAWN        38280   Year    FLUSHING        NY      38272   Year    2013
Time taken: 23.585 seconds, Fetched: 1 row(s)
hive>
```

d. Last record for 2013

$ select * from h1b_application where case_no = 'I-200-12263-068675';



```
hive> select * from h1b_application where case_no ='I-200-12263-068675';
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1411083433145_0077, Tracking URL = N/A
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1411083433145_0077
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 0
2014-10-06 22:08:50,145 Stage-1 map = 0%,  reduce = 0%
2014-10-06 22:08:58,429 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.42 sec
2014-10-06 22:08:59,457 Stage-1 map = 33%,  reduce = 0%, Cumulative CPU 2.42 sec
2014-10-06 22:09:00,488 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 8.02 sec
2014-10-06 22:09:01,527 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.82 sec
2014-10-06 22:09:02,564 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.82 sec
MapReduce Total cumulative CPU time: 13 seconds 820 msec
Ended Job = job_1411083433145_0077
MapReduce Jobs Launched:
Job 0: Map: 3   Cumulative CPU: 13.82 sec   HDFS Read: 565510017 HDFS Write: 178 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 820 msec
OK
{"year":2012,"month":9,"date":21}       I-200-12263-068675      ASSOC. OF UNIVERSITIES FOR RES  HILO    HI      96720-2700      {
"year":2012,"month":11,"date":1}        {"year":2015,"month":11,"date":1}       GEMINI SCIENCE FELLOW   {"month":2012,"date":9,"
year":"27"}     CERTIFIED       61981   Year    HILO    HI      50211   Year    2012
Time taken: 24.778 seconds, Fetched: 1 row(s)
hive>
```

Phase 2 - Building and Verifying Queries

1. Top 5 States offering most H-1B jobs since 2010

Objective of the query: To find the top states with most number of applications since the year 2010

Description of the query :

The US States where you choose to search for an H1B Visa-sponsored job can dramatically affect your career.

Having conducted extensive annual trends analysis and studied visa sponsorship patterns over the last half decade, the top 5 States may slide up and down a few spots in the rankings, but below are the States that most consistently present the highest numbers of Companies that sponsor for H1B visas / the highest levels of opportunity for H1B job seekers. These top 5 States present great target locations for 2013.

Query :

```
select visa_application_year, employer_state, count(*) as count from h1b_application where visa_application_year in('2010','2011','2012','2013') group by visa_application_year, employer_state order by visa_application_year, count desc;
```

Query Output:

1. 2010:

2010	CA	58327
2010	NJ	36641
2010	NY	35636
2010	TX	27539
2010	IL	17909

2. 2011:

2011	CA	62054
2011	NJ	42461
2011	NY	35344
2011	TX	29171
2011	IL	19209

3. 2012:

```
2012    CA      68856
2012    NJ      53791
2012    TX      48365
2012    NY      35989
2012    IL      21526
```

4. 2013 :

```
2013    CA      74633
2013    TX      66760
2013    NJ      53977
2013    NY      34288
2013    IL      24796
```

Execution time of query:

```
2013    FM      2
Time taken: 63.688 seconds, Fetched: 227 row(s)
```

Verification of query: Verified the result for year 2012 and state TX

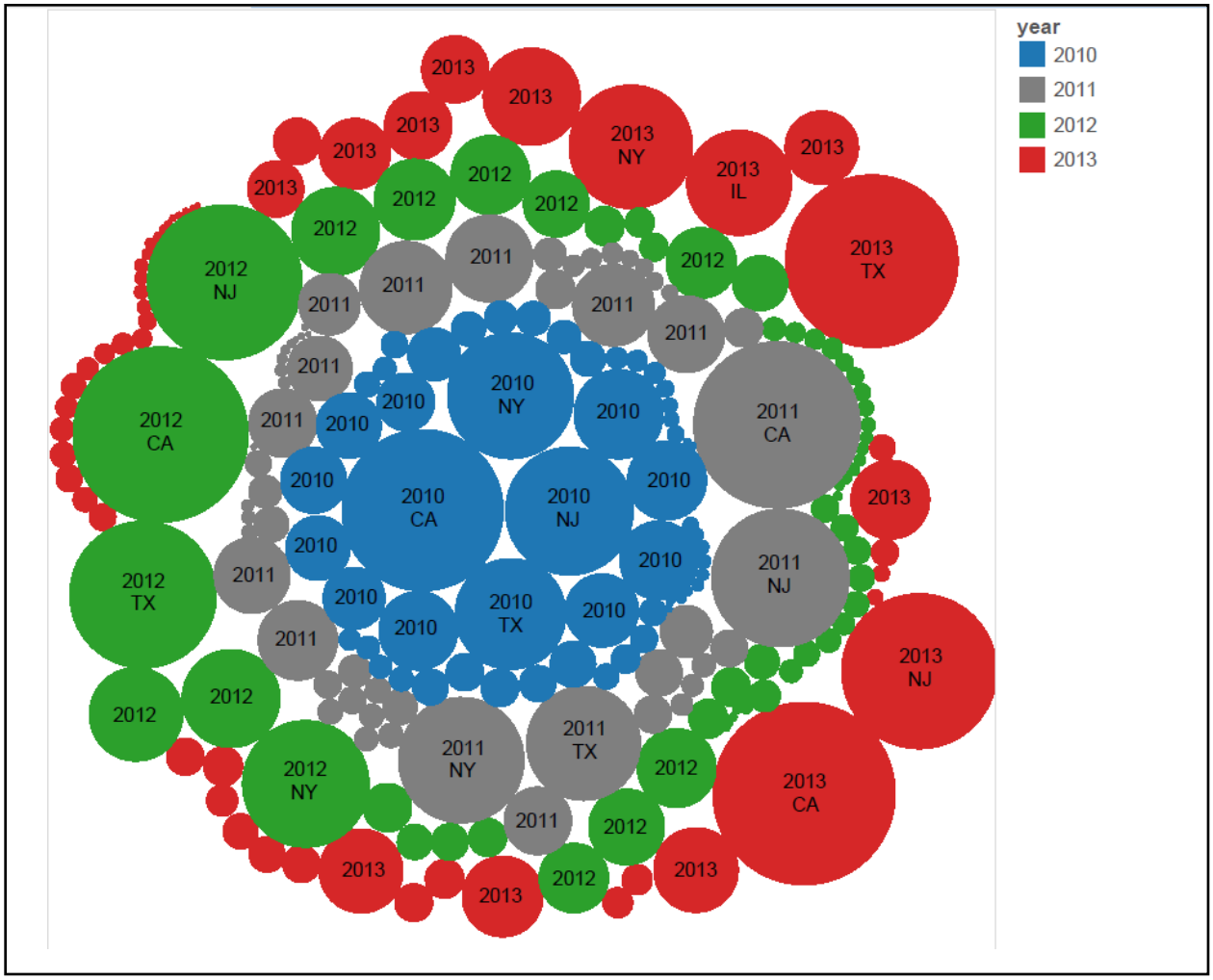
Query:

select visa_application_year, employer_state, count(*) from h1b_application where employer_state = 'TX' and visa_application_year = '2012' group by visa_application_year, employer_state;

Result:

```
Job 0: Map: 3  Reduce: 1   Cumulative CPU: 15.94 sec   HDFS Read: 565510017 HDFS Write: 14 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 940 msec
OK
2012    TX      48365
Time taken: 32.867 seconds, Fetched: 1 row(s)
hive>
```

Visualization



Conclusion:

- Here we find that CA is on top of the list with highest no of applications for all the years.
- TX has picked up in the last 4 years and has come to 2nd position in no of H1B applications with 66760 applications in 2013 from the 4th position in 2010.

2. Top 5 Employers offering most H-1B jobs since 2010

Objective of the query: To find the top 5 employers since 2010 in terms of filing number of H-1B applications. The goal here is to verify if the top hiring employers are from CA or from other states.

Description of the query :

If you want an H1B visa, the key is planning, focusing and targeting the Companies that are currently sponsoring, and, the jobs in those companies that are available and qualify for H1B visa sponsorship. So, we focused on finding the top employers and state in which those employers file the most number of visa applications.

Query :

select visa_application_year, employer_name, employer_state, count(*) as count from h1b_application where visa_application_year ='2010' group by employer_name, visa_application_year , employer_state order by visa_application_year, count desc limit 5;

Query Output:

1.2010 :

```
Total MapReduce CPU Time Spent: 27 seconds 270 msec
OK
2010    MICROSOFT CORPORATION   WA      4449
2010    WIPRO LIMITED   NJ      3025
2010    DELOITTE CONSULTING LLP PA      2342
2010    INFOSYS TECHNOLOGIES LIMITED    TX      2182
2010    FUJITSU AMERICA, INC.   CA      1714
Time taken: 65.246 seconds, Fetched: 5 row(s)
hive>
```

2.2011:

```
Total MapReduce CPU Time Spent: 26 seconds 720 msec
OK
2011    TATA CONSULTANCY SERVICES LIMI  MD      5416
2011    MICROSOFT CORPORATION   WA      4252
2011    DELOITTE CONSULTING LLP PA      3621
2011    WIPRO LIMITED   NJ      3027
2011    COGNIZANT TECHNOLOGY SOLUTIONS  NJ      2742
Time taken: 65.966 seconds, Fetched: 5 row(s)
hive>
```

3.2012:

```
Total MapReduce CPU Time Spent: 27 seconds 280 msec
OK
2012    INFOSYS LIMITED TX      16318
2012    WIPRO LIMITED   NJ      7182
2012    TATA CONSULTANCY SERVICES LIMI  MD      6736
2012    DELOITTE CONSULTING LLP PA      4727
2012    IBM INDIA PRIVATE LIMITED       NC      4075
Time taken: 67.375 seconds, Fetched: 5 row(s)
hive>
```

4.2013:

```
Total MapReduce CPU Time Spent: 27 seconds 570 msec
OK
2013    INFOSYS LIMITED TX      32256
2013    TATA CONSULTANCY SERVICES LIMI  MD      8790
2013    WIPRO LIMITED   NJ      6734
2013    DELOITTE CONSULTING LLP PA      6124
2013    ACCENTURE LLP   IL      4994
Time taken: 64.346 seconds, Fetched: 5 row(s)
hive>
```

Execution time of query:

1. 2010:

```
Time taken: 65.246 seconds, Fetched: 5 row(s)
hive>
```

2.2011:

```
Time taken: 65.966 seconds, Fetched: 5 row(s)
hive>
```

3. 2012:

```
Time taken: 67.375 seconds, Fetched: 5 row(s)
hive>
```

4. 2013:

```
Time taken: 64.346 seconds, Fetched: 5 row(s)
hive>
```

Verification of query: Verified the result for year 2010 and employer name MICROSOFT CORPORATION

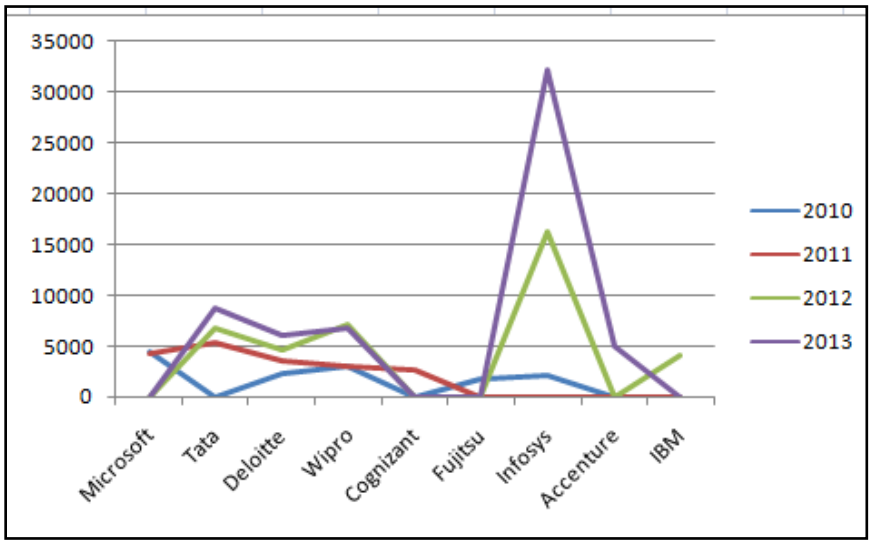
Query :

select visa_application_year, employer_name, employer_state, count(*) as count from h1b_application where visa_application_year ='2010' and employer_name = 'MICROSOFT CORPORATION' group by employer_name, visa_application_year, employer_state ;

Result :

```
Job 0: Map: 3  Reduce: 1   Cumulative CPU: 15.21 sec   HD
Total MapReduce CPU Time Spent: 15 seconds 210 msec
OK
2010    MICROSOFT CORPORATION   WA      4449
Time taken: 32.62 seconds, Fetched: 1 row(s)
```

Visualization



Conclusion:

- INFOSYS LIMITED in TX(Texas) state has come up as number 1 H-1B filing company with no of applications 32256.
- TATA CONSULTANCY SERVICES is on second position even though MD(Maryland) does not come in top 5 states in h1b VISA filing.
- We see that there is a sudden increase in no of H-1B applications for the employer INFOSYS LIMITED in 2013. We searched for INFOSYS and we found that INFOSYS has been recently fined $34 million in U.S. Visa case. The reason was that company was bringing employers on B1 visa instead of H-1B Visa. And now after this case we see that there is drastic increase in number of H-1B visa filled by INFOSYS LIMITED following the case.

http://www.reuters.com/article/2013/11/08/usinfosyssettlementidUSBRE9A70IC20131108
http://timesofindia.indiatimes.com/tech/technews/Infosystopay34mfinewithin10dayssettlementnottoaffectfuturevisas/articleshow/24952527.cms

- This concludes that offshore companies are filling the most H-1B applications.
- US is in dire need of technical professionals.

3. Analysis for State of Maryland

Objective of the query: After realizing Maryland had almost four times applications in 2013 comparing with 2010, we wanted to see the companies there and the number of H-1B they filed each year from 2010 to 2013.

Description of the query :

We see that TATA CONSULTANCY SERVICES are filing more number of applications even if the state of Maryland does not show up in top 5 state in hiring H-1B employees. So, we were curious to know if there were other employers in this state with such large number of application filing.

Query :

```
select employer_name, count(*) as num from h1b_test where employer_state = 'MD' and
visa_application_year = '2010' group by employer_name order by num DESC limit 5;
```

Query Output:

1.2010:

```
Total MapReduce CPU Time Spent: 18 seconds 550 msec
OK
TATA CONSULTANCY SERVICES LIMI  1007
NATIONAL INSTITUTES OF HEALTH,  406
THE JOHNS HOPKINS UNIVERSITY    354
PRINCE GEORGE'S COUNTY PUBLIC   239
UNIVERSITY OF MARYLAND COLLEGE  213
Time taken: 62.122 seconds, Fetched: 5 row(s)
```

2.2011:

```
Total MapReduce CPU Time Spent: 18 seconds 280 msec
OK
TATA CONSULTANCY SERVICES LIMI  5416
NATIONAL INSTITUTES OF HEALTH,  408
JOHNS HOPKINS UNIVERSITY        313
BALTIMORE CITY PUBLIC SCHOOL S  308
UNIVERSITY OF MARYLAND COLLEGE  251
Time taken: 60.869 seconds, Fetched: 5 row(s)
```

3. 2012:

```
OK
TATA CONSULTANCY SERVICES LIMI  6736
NATIONAL INSTITUTES OF HEALTH,  449
JOHNS HOPKINS UNIVERSITY        368
UNIVERSITY OF MARYLAND COLLEGE  219
XPEDITE TECHNOLOGIES INC        205
Time taken: 62.092 seconds, Fetched: 5 row(s)
```

4. 2013:

```
Total MapReduce CPU Time Spent: 18 seconds 670 msec
OK
TATA CONSULTANCY SERVICES LIMI  8790
NATIONAL INSTITUTES OF HEALTH,  420
JOHNS HOPKINS UNIVERSITY        382
UNIVERSITY OF MARYLAND COLLEGE  256
TEKSYSTEMS GLOBAL SERVICES LLC  179
Time taken: 63.39 seconds, Fetched: 5 row(s)
hive>
```

Execution time of query:

1. 2010:

```
Time taken: 62.122 seconds, Fetched: 5 row(s)
```

2. 2011:

```
Time taken: 60.869 seconds, Fetched: 5 row(s)
```

3. 2012:

```
Time taken: 62.092 seconds, Fetched: 5 row(s)
```

4. 2013:

```
Time taken: 63.39 seconds, Fetched: 5 row(s)
```

Verification of query: Verify the result for state MD and employer_name JOHNS HOPKINS UNIVERSITY for visa application year 2013

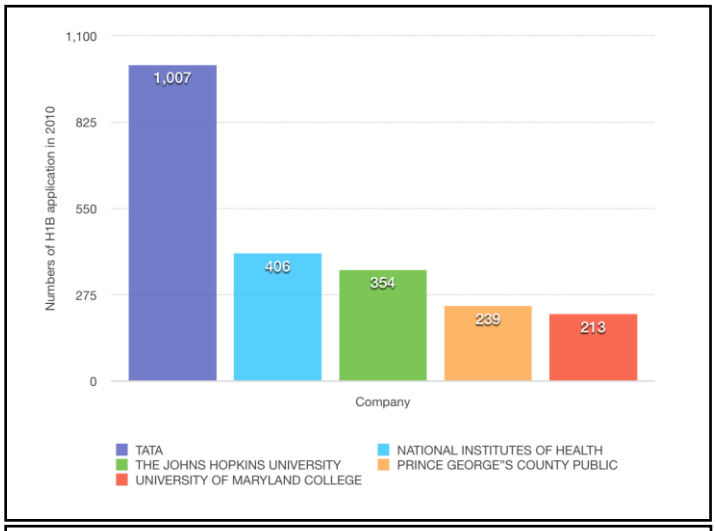
Query :

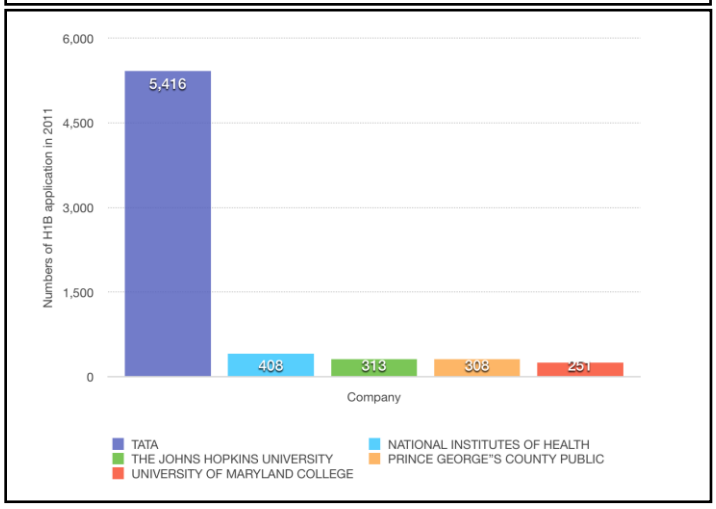
select employer_name, count(*) as num from h1b_test where employer_state = 'MD' and employer_name = 'JOHNS HOPKINS UNIVERSITY' and visa_application_year = '2013' group by employer_name;

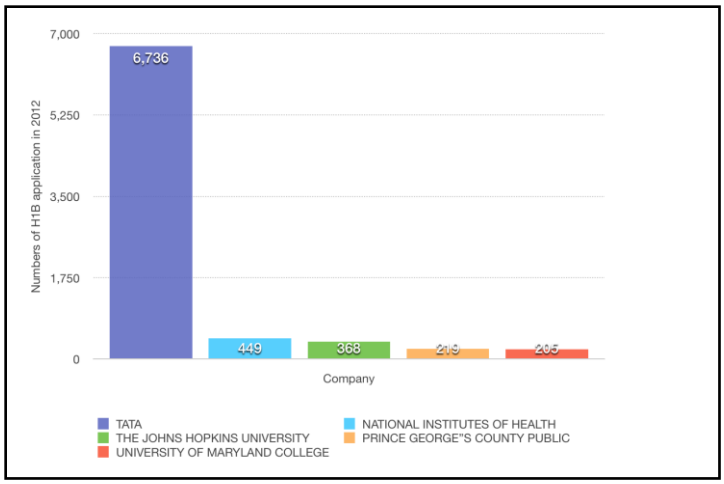
Result :

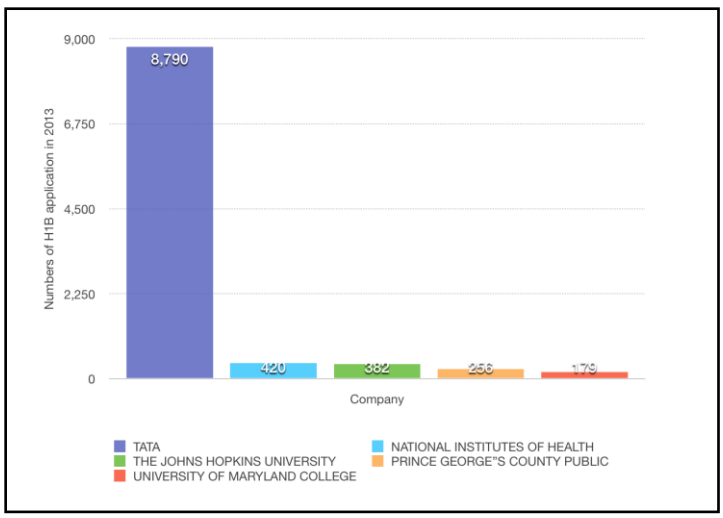
```
Job 0: Map: 3  Reduce: 1   Cumulative CPU: 15.21 sec   HDF
Total MapReduce CPU Time Spent: 15 seconds 210 msec
OK
JOHNS HOPKINS UNIVERSITY        382
Time taken: 35.506 seconds, Fetched: 1 row(s)
hive>
```

Visualization









Conclusion:

- By looking at the Tata Consultancy company in 2010 to 2011, they hired 5 times more employees. We were curious to know regarding this drastic increase in application filing. What we found?
  http://www.computing.co.uk/ctg/news/2026508/india-s-tcs-targets-smes-usd1bn-revenue
  http://en.wikipedia.org/wiki/Tata_Consultancy_Services

- After reading these articles, the reason for sudden increase in the applications was that the company was entering into the small and medium enterprises market. According to their Chief Executive Officer, the company aimed to increase their small and medium enterprise clients.

4. Average Wage Rate and Maximum Wage Rate analysis for the top 5 hiring employers and comparison with the top 5 hiring employers in CA

a. Job titles having maximum wage rate offered by top 5 employers in 2013

Objective of the query: To find the job titles for whom employers are offering the highest wage rate.

Description of the query :

After having the top five employers on the list, we decided to find the job titles for whom they have the maximum wage rate.

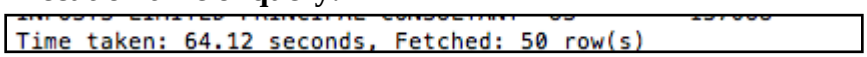
Query:

```
select employer_name, job_title, MAX( wage_rate ) as wage_rate  from h1b_application where visa_application_year=2013 and employer_name in('INFOSYS LIMITED', 'TATA CONSULTANCY SERVICES LIMI', 'WIPRO LIMITED', 'DELOITTE CONSULTING LLP', 'ACCENTURE LLP')  group by employer_name, job_title order by wage_rate desc limit 50;
```

Query Output:

```
Total MapReduce CPU Time Spent: 22 seconds 910 msec
OK
WIPRO LIMITED   BUSINESS SYSTEM ANALYST 20132715
TATA CONSULTANCY SERVICES LIMI  COMPUTER PROGRAMMER     671000
INFOSYS LIMITED TECHNOLOGY ANALYST-US   670083
TATA CONSULTANCY SERVICES LIMI  DATABASE ADMINISTRATOR  607000
DELOITTE CONSULTING LLP PRINCIPAL       380000
DELOITTE CONSULTING LLP SENIOR MANAGER  278300
INFOSYS LIMITED PARTNER - BUSINESS CONSULTING   240750
INFOSYS LIMITED PRINCIPAL LEGAL COUNSEL 227106
ACCENTURE LLP   SALES DIRECTOR - FINANCIAL SER  222352
INFOSYS LIMITED PRINCIPAL - BUSINESS CONSULTIN  202020
DELOITTE CONSULTING LLP MANAGER 200000
DELOITTE CONSULTING LLP SPECIALIST LEADER       190000
INFOSYS LIMITED SENIOR PRINCIPAL - BUSINESS CO  186180
ACCENTURE LLP   CONSULTING- SENIOR MANAGER      186100
ACCENTURE LLP   SENIOR MANAGER, ENERGY  182500
INFOSYS LIMITED INDUSTRY PRINCIPAL - US 180600
WIPRO LIMITED   BUSINESS DEVELOPMENT MANAGER    179213
ACCENTURE LLP   HEALTHCARE SALES DIRECTOR       179005
ACCENTURE LLP   COMPUTER AND INFORMATION SYSTE  175100
DELOITTE CONSULTING LLP SPECIALIST MASTER       175000
INFOSYS LIMITED PRINCIPAL CONSULTANT - US       174240
INFOSYS LIMITED PRINCIPAL TECHNOLOGY ARCHITECT  173340
ACCENTURE LLP   SENIOR MANAGER  170206
ACCENTURE LLP   MANAGER 166300
INFOSYS LIMITED ENGAGEMENT MANAGER      165000
ACCENTURE LLP   STRATEGY MANAGER        164400
TATA CONSULTANCY SERVICES LIMI  DATA ARCHITECT  163714
ACCENTURE LLP   SENIOR APPLICATION ARCHITECT    157352
DELOITTE CONSULTING LLP CONSULTANT      155000
INFOSYS LIMITED PRINCIPAL - PROCESS AND DOMAIN  151440
INFOSYS LIMITED PRINCIPAL CONSULTANT-US 150696
INFOSYS LIMITED SENIOR TECHNOLOGY ARCHITECT -   150480
DELOITTE CONSULTING LLP SENIOR CONSULTANT       150000
ACCENTURE LLP   HNAV DEVELOPER  150000
INFOSYS LIMITED PRINCIPAL - PROCESS & DOMAIN    150000
ACCENTURE LLP   SAP SENIOR MANAGER      147600
ACCENTURE LLP   COMPUTER INFORMATION SYSTEM MA  147400
DELOITTE CONSULTING LLP ASSOCIATE 7     145000
DELOITTE CONSULTING LLP SPECIALIST SENIOR       145000
WIPRO LIMITED   ACCOUNT DELIVERY HEAD   144872
DELOITTE CONSULTING LLP ASSOCIATE 8     144600
INFOSYS LIMITED LEAD CONSULTANT - US    143933
DELOITTE CONSULTING LLP ASSOCIATE 9     142769
ACCENTURE LLP   IMPLEMENTATION MANAGER  142000
ACCENTURE LLP   CONSULTING MANAGER      141200
ACCENTURE LLP   PHARMACEUTICAL R&D MANAGEMENT   140500
INFOSYS LIMITED DELIVERY MANAGER - US   138309
INFOSYS LIMITED TECHNOLOGY ARCHITECT - US       138240
INFOSYS LIMITED SENIOR PROJECT MANAGER - US     137520
INFOSYS LIMITED PRINCIPAL CONSULTANT- US        137068
Time taken: 64.12 seconds, Fetched: 50 row(s)
```

Execution time of query:



Verification of query:

Query :

```
select employer_name, job_title, MAX( wage_rate ) as wage_rate  from h1b_application where visa_application_year=2013 and  employer_name = 'WIPRO LIMITED' and job_title ='BUSINESS SYSTEM ANALYST'  group by employer_name, job_title ;
```

Result :

```
Total MapReduce CPU Time Spent: 18 seconds 980 msec
OK
WIPRO LIMITED   BUSINESS SYSTEM ANALYST 20132715
Time taken: 34.025 seconds, Fetched: 1 row(s)
```

Conclusion:

- Above query show the job titles which are receiving maximum pay from the top five employers.

- We extract all the keywords related to highest pay job titles:
  BUSINESS, COMPUTER, TECHNOLOGY, DATABASE, SALES, CONSULTING, CONSULTANT, MANAGER, ENGINEER, PROGRAMMER, SOFTWARE

b. Average wage rate for the job titles for top 5 employers.

Objective of the query: To find the average wage for each job title by top five employers.

Description of the query :

For each employer, we wanted to find the average wage rate for different job positions.

Query:

```
select employer_name, job_title,
AVG(IF(wage_rate_unit='Hour',(wage_rate*2080),IF(wage_rate_unit='Week',(wage_rate*52),
IF(wage_rate_unit='Year', wage_rate,0)))) as wage_rate from h1b_application where visa_application_year=2013 and  employer_name in('INFOSYS LIMITED', 'TATA CONSULTANCY SERVICES LIMI', 'WIPRO LIMITED', 'DELOITTE CONSULTING LLP', 'ACCENTURE LLP')  and (job_title LIKE '% BUSINESS %'  OR job_title LIKE '% COMPUTER %'  OR job_title LIKE '% TECHNOLOGY %'  OR job_title LIKE '% DATABASE %'  OR job_title LIKE '% SALES %'  OR job_title LIKE '% CONSULTING %'  OR job_title LIKE '% MANAGER %' OR job_title LIKE '% SOFTWARE %' OR job_title LIKE '% ENGINEERING %'  OR job_title LIKE '% PROGRAMMER %')  group by employer_name, job_title order by wage_rate desc;
```

Query Output:

```
Total MapReduce CPU Time Spent: 22 seconds 190 msec
OK
INFOSYS LIMITED PARTNER - BUSINESS CONSULTING   185849.75
ACCENTURE LLP   HEALTHCARE SALES DIRECTOR       179005.0
INFOSYS LIMITED SENIOR PRINCIPAL - BUSINESS CO  167703.625
INFOSYS LIMITED DELIVERY MANAGER - US   136754.5
INFOSYS LIMITED PRINCIPAL - BUSINESS CONSULTIN  136419.72
INFOSYS LIMITED SENIOR TECHNOLOGY ARCHITECT--U  134748.0
TATA CONSULTANCY SERVICES LIMI  INFORMATION TECHNOLOGY PROJECT  125000.0
INFOSYS LIMITED PRINCIPAL TECHNOLOGY ARCHITECT  124885.69230769231
ACCENTURE LLP   STRATEGY & BUSINESS DEVELOPMEN  123000.0
INFOSYS LIMITED SENIOR TECHNOLOGY ARCHITECT-US  120123.7
ACCENTURE LLP   INFORMATION TECHNOLOGY PROJECT  119000.0
INFOSYS LIMITED SENIOR PROJECT MANAGER - ENGIN  118143.8
TATA CONSULTANCY SERVICES LIMI  STRATEGIC BUSINESS MANAGER      115000.0
INFOSYS LIMITED SENIOR TECHNICAL MANAGER - US   112504.0
INFOSYS LIMITED SENIOR TECHNOLOGY ARCHITECT -   110187.08888888889
INFOSYS LIMITED GROUP PROJECT MANAGER - US      108780.56818181818
INFOSYS LIMITED PROGRAM MANAGER - US    105403.79166666667
INFOSYS LIMITED SENIOR ASSOCIATE - BUSINESS CO  105391.08771929824
INFOSYS LIMITED SENIOR PROJECT MANAGER - US     104868.18781725888
INFOSYS LIMITED TECHNICAL MANAGER - US  96558.75
INFOSYS LIMITED SENIOR TECHNOLOGY ARCHITECT     94032.0
INFOSYS LIMITED PROJECT MANAGER - ENGINEERING   92028.62068965517
INFOSYS LIMITED PROJECT MANAGER - US    91025.99804305284
INFOSYS LIMITED PROJECT MANAGER -US     90813.0
ACCENTURE LLP   DATASTAGE BUSINESS INTELLIGENC  90000.0
INFOSYS LIMITED SENIOR ENGINEERING MANAGER - U  85311.5
TATA CONSULTANCY SERVICES LIMI  SENIOR PROGRAMMER ANALYST       85000.0
ACCENTURE LLP   SENIOR SOFTWARE ENGINEER        83000.0
INFOSYS LIMITED ASSOCIATE - BUSINESS CONSULTIN  79245.22727272728
INFOSYS LIMITED TECHNOLOGY LEAD - ENGINEERING   76749.65034965034
INFOSYS LIMITED PRODUCT MANAGER - US    76308.0
INFOSYS LIMITED ENGINEERING MANAGER - US        73689.42857142857
DELOITTE CONSULTING LLP BUSINESS TECHNOLOGY ANLAYST     72666.66666666667
DELOITTE CONSULTING LLP BUSINESS TECHNOLOGY ANALYST     72663.08571428571
DELOITTE CONSULTING LLP BUSINESS TECHNOLOGY ANALTST     72368.18181818182
TATA CONSULTANCY SERVICES LIMI  NETWORK & COMPUTER SYSTEMS ADM  69200.0
TATA CONSULTANCY SERVICES LIMI  APPLICATION PROGRAMMER ANALYST  68050.0
TATA CONSULTANCY SERVICES LIMI  APPLICATION PROGRAMMER LEAD     68000.0
```

Execution time of query:

```
Time taken: 63.083 seconds, Fetched: 65 row(s)
hive>
```

Verification of query:

Query :

select employer_name, job_title,
AVG(IF(wage_rate_unit='Hour',(wage_rate*2080),IF(wage_rate_unit='Week',(wage_rate*52),
IF(wage_rate_unit='Year', wage_rate,0)))) as wage_rate from h1b_application where

visa_application_year=2013 and employer_name ='ACCENTURE LLP' and (job_title LIKE '% BUSINESS %') group by employer_name, job_title order by wage_rate desc;

Result :

```
Total MapReduce CPU Time Spent: 21 seconds 180 msec
OK
ACCENTURE LLP   STRATEGY & BUSINESS DEVELOPMEN  123000.0
ACCENTURE LLP   DATASTAGE BUSINESS INTELLIGENC  90000.0
Time taken: 63.321 seconds, Fetched: 2 row(s)
hive>
```

Conclusion:

- We found average wage for each job position that different employers pay.

c. Maximum wage rate for the above extracted job titles for top 5 employers.

Objective of the query: To find the maximum wage rate for each job title by top five employers.

Description of the query :

For each employer, we wanted to find the maximum wage rate for different job positions.

Query:

select employer_name, job_title, MAX(wage_rate) as wage_rate from h1b_application where visa_application_year=2013 and employer_name in('INFOSYS LIMITED', 'TATA CONSULTANCY SERVICES LIMI', 'WIPRO LIMITED', 'DELOITTE CONSULTING LLP', 'ACCENTURE LLP') and (job_title LIKE '% BUSINESS %' OR job_title LIKE '% COMPUTER %' OR job_title LIKE '% TECHNOLOGY %' OR job_title LIKE '% DATABASE %' OR job_title LIKE '% SALES %' OR job_title LIKE '% CONSULTING %' OR job_title LIKE '% MANAGER %' OR job_title LIKE '% SOFTWARE %' OR job_title LIKE '% ENGINEERING %') group by employer_name, job_title order by wage_rate desc;

Query Output:

```
Total MapReduce CPU Time Spent: 22 seconds 170 msec
OK
INFOSYS LIMITED PARTNER - BUSINESS CONSULTING   240750
INFOSYS LIMITED PRINCIPAL - BUSINESS CONSULTIN  202020
INFOSYS LIMITED SENIOR PRINCIPAL - BUSINESS CO  186180
ACCENTURE LLP   HEALTHCARE SALES DIRECTOR       179005
INFOSYS LIMITED PRINCIPAL TECHNOLOGY ARCHITECT  173340
INFOSYS LIMITED SENIOR TECHNOLOGY ARCHITECT -   150480
INFOSYS LIMITED DELIVERY MANAGER - US   138309
INFOSYS LIMITED SENIOR PROJECT MANAGER - US     137520
INFOSYS LIMITED SENIOR TECHNOLOGY ARCHITECT--U  134748
INFOSYS LIMITED PROGRAM MANAGER - US    133093
INFOSYS LIMITED GROUP PROJECT MANAGER - US      131058
INFOSYS LIMITED SENIOR ASSOCIATE - BUSINESS CO  130811
INFOSYS LIMITED SENIOR PROJECT MANAGER - ENGIN  125561
TATA CONSULTANCY SERVICES LIMI  INFORMATION TECHNOLOGY PROJECT  125000
INFOSYS LIMITED SENIOR TECHNOLOGY ARCHITECT-US  124005
ACCENTURE LLP   STRATEGY & BUSINESS DEVELOPMEN  123000
INFOSYS LIMITED PROJECT MANAGER - US    121056
ACCENTURE LLP   INFORMATION TECHNOLOGY PROJECT  119000
INFOSYS LIMITED SENIOR TECHNICAL MANAGER - US   117241
TATA CONSULTANCY SERVICES LIMI  STRATEGIC BUSINESS MANAGER      115000
INFOSYS LIMITED PROJECT MANAGER - ENGINEERING   104664
INFOSYS LIMITED ASSOCIATE - BUSINESS CONSULTIN  100200
INFOSYS LIMITED TECHNICAL MANAGER - US  99242
ACCENTURE LLP   COMPUTER PROGRAMMER / CONFIGUR  97700
ACCENTURE LLP   SENIOR SOFTWARE ENGINEER        97000
INFOSYS LIMITED SENIOR ENGINEERING MANAGER - U  96138
INFOSYS LIMITED TECHNOLOGY LEAD - ENGINEERING   95198
INFOSYS LIMITED SENIOR TECHNOLOGY ARCHITECT     94032
INFOSYS LIMITED PROJECT MANAGER -US     90813
ACCENTURE LLP   DATASTAGE BUSINESS INTELLIGENC  90000
ACCENTURE LLP   COMPUTER PROGRAMMER CONFIGURER  86500
INFOSYS LIMITED ENGINEERING MANAGER - US        85140
TATA CONSULTANCY SERVICES LIMI  SENIOR PROGRAMMER ANALYST       85000
INFOSYS LIMITED ASSOCIATE BUSINESS ANALYST - U  83866
DELOITTE CONSULTING LLP BUSINESS TECHNOLOGY ANALYST     83400
TATA CONSULTANCY SERVICES LIMI  NETWORK AND COMPUTER SYSTEMS A  81500
```

Execution time of query:

```
Time taken: 63.044 seconds, Fetched: 56 row(s)
hive>
```

Verification of query: Verifying the query for employer_name ACCENTURE LLP;

Query :

select employer_name, job_title, MAX(wage_rate) as wage_rate from h1b_application where visa_application_year=2013 and employer_name ='ACCENTURE LLP' and (job_title LIKE '% BUSINESS %') group by employer_name, job_title;

Result :

```
Job 0: Map: 3  Reduce: 1   Cumulative CPU: 20.16 sec   HDFS Read:
Total MapReduce CPU Time Spent: 20 seconds 160 msec
OK
ACCENTURE LLP   DATASTAGE BUSINESS INTELLIGENC  90000
ACCENTURE LLP   STRATEGY & BUSINESS DEVELOPMEN  123000
Time taken: 34.554 seconds, Fetched: 2 row(s)
hive>
```

Conclusion:

- We found maximum wage for each job position that different employers pay.

d. Finding Top employers in CA in terms of no of visa applications

Objective of the query: To find the employers who file the most number of H-1B visa applications in California.

Description of the query :

After having the top five employers from the States on the list, we wanted to find the employers in California who filed most applications as it is the leading state for filing job applications.

Query:

select employer_name, count(*) as count from h1b_application where visa_application_year ='2013' and employer_state = 'CA' Group by employer_name order by count desc limit 20;

Query Output:

```
Total MapReduce CPU Time Spent: 25 seconds 700 msec
OK
HCL AMERICA, INC.       3013
GOOGLE INC.     2166
IGATE TECHNOLOGIES INC. 2059
INTEL CORPORATION       1947
QUALCOMM TECHNOLOGIES, INC.     1901
ORACLE AMERICA, INC.    1757
APPLE INC.      971
UST GLOBAL INC  897
QUALCOMM INCORPORATED   686
EBAY INC.       592
FACEBOOK, INC.  582
UST GLOBAL INC. 575
CISCO SYSTEMS, INC.     530
MINDTREE LIMITED        524
YAHOO! INC.     504
BROADCOM CORPORATION    480
PERSISTENT SYSTEMS, INC.        477
HEWLETT-PACKARD COMPANY 408
UNIVERSITY OF CALIFORNIA, SAN   382
SALESFORCE.COM, INC.    362
Time taken: 66.116 seconds, Fetched: 20 row(s)
```

Execution time of query:

```
Time taken: 66.116 seconds, Fetched: 20 row(s)
```

Verification of query: Verified the above query for application_year 2013 and employer name HCL AMERICA INC.

Query :

```
select employer_name, count(*) as count  from h1b_application where
visa_application_year ='2013' and employer_state = 'CA' and employer_name ='HCL
AMERICA, INC.' Group by employer_name;
```

Result :

```
Total MapReduce CPU Time Spent: 15 seconds 300 msec
OK
HCL AMERICA, INC.       3013
Time taken: 32.697 seconds, Fetched: 1 row(s)
```

Conclusion:

- HCL America, Inc. filed the most visa applications in the year 2013 followed by Google.

e. Finding average wage rate for top employers in CA

Objective of the query: To find the average wage rate for employers who file the most number of H-1B visa applications in California.

Description of the query :

To find the average wage rate offered by different employers in California in 2013.

Query:

```
select employer_name,
job_title, AVG(IF(wage_rate_unit='Hour',(wage_rate*2080),IF(wage_rate_unit='Week',(wage_rate*52), IF(wage_rate_unit='Year', wage_rate,0)))) as wage_rate from h1b_application where visa_application_year=2013 and employer_name in('HCL AMERICA, INC.', 'GOOGLE INC.', 'IGATE TECHNOLOGIES INC.', 'INTEL CORPORATION', 'ORACLE AMERICA, INC.','APPLE INC.', 'EBAY INC.', 'FACEBOOK, INC.', 'CISCO SYSTEMS, INC.') and (job_title LIKE '% BUSINESS %' OR job_title LIKE '% COMPUTER %' OR job_title LIKE '% TECHNOLOGY %' OR job_title LIKE '% DATABASE %' OR job_title LIKE '% SALES %' OR job_title LIKE '% CONSULTING %' OR job_title LIKE '% MANAGER %' ) group by employer_name, job_title order by wage_rate desc LIMIT 50;
```

Query Output:

```
MapReduce Jobs Launched:
Job 0: Map: 3  Reduce: 1   Cumulative CPU: 19.47 sec   HDFS Read: 56
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 2.5 sec   HDFS Read: 1436
Total MapReduce CPU Time Spent: 21 seconds 970 msec
OK
ORACLE AMERICA, INC.    SERVICES SALES SENIOR VICE PRE  400000.0
GOOGLE INC.     SOFTWARE ENGINEERING MANAGER    223750.0
GOOGLE INC.     LEAD SOFTWARE ENGINEER  219288.0
FACEBOOK, INC.  DIRECTOR, BUSINESS OPERATIONS   218162.0
APPLE INC.      TEST ENGINEERING MANAGER 3      210000.0
GOOGLE INC.     STAFF SOFTWARE ENGINEER 206900.0
GOOGLE INC.     MANAGER, SOFTWARE ENGINEERING   200000.0
EBAY INC.       PRINCIPAL MTS, SOFTWARE ENGINE  200000.0
FACEBOOK, INC.  MANAGER, SOFTWARE ENGINEERING   194180.0
HCL AMERICA, INC.       OPERATIONS MANAGER - IV 190543.5
APPLE INC.      SOFTWARE ENGINEERING MANAGER    185000.0
GOOGLE INC.     MANAGER - DATABASE ADMINISTRAT  180000.0
HCL AMERICA, INC.       SALES MANAGER - IV      179343.0
APPLE INC.      SOFTWARE DEVELOP MANAGER 2      178500.0
ORACLE AMERICA, INC.    SENIOR SOFTWARE MANAGER 175000.0
APPLE INC.      PRINCIPAL SOFTWARE ENGR/DATA S  170000.0
EBAY INC.       MTS 3, SOFTWARE ENGINEER        170000.0
APPLE INC.      FIRMWARE ENGINEER MANAGER 1     170000.0
APPLE INC.      SOFTWARE DEVELOP MANAGER 3      170000.0
EBAY INC.       SR. PRODUCT MANAGER 2 - TECHNI  165000.0
GOOGLE INC.     SR. SOFTWARE ENGINEER   164600.0
APPLE INC.      HARDWARE DEVELOP MANAGER 2      162000.0
FACEBOOK, INC.  SOFTWARE ENGINEERING MANAGER    161773.0
APPLE INC.      SENIOR IOS LOCATION SOFTWARE E  160000.0
EBAY INC.       SR. MTS, SOFTWARE ENGINEER      160000.0
GOOGLE INC.     SENIOR SOFTWARE ENGINEER        159200.0
ORACLE AMERICA, INC.    SALES CONSULTING SENIOR MANAGE  154596.0
EBAY INC.       MTS 2, SOFTWARE ENGINEER        153894.9090909091
EBAY INC.       MANAGER, SOFTWARE DEVELOPMENT   153142.5
EBAY INC.       SR. PRODUCT MANAGER 1 - TECHNI  152000.0
GOOGLE INC.     SENIOR SOFTWARE ENGINEER, PART  151812.0
HCL AMERICA, INC.       MARKETING MANAGER - III 151486.0
APPLE INC.      SILICON VALIDATION SOFTWARE EN  150000.0
FACEBOOK, INC.  RECRUITING PROGRAM MANAGER - M  150000.0
GOOGLE INC.     QUANTITATIVE BUSINESS ANALYST   148000.0
APPLE INC.      ENGINEERING SERVICES MANAGER 1  147000.0
ORACLE AMERICA, INC.    TELESALES/INTERNET SALES DIREC  145959.0
APPLE INC.      SR. IOS WIFI SOFTWARE ENGINEER  145000.0
```

Execution time of query:

```
Time taken: 64.099 seconds, Fetched: 50 row(s)
hive>
```

Verification of query: Verified the query for employer_name GOOGLE INC and job_title keywords SOFTWARE and BUSINESS.

Query :

```
select employer_name,
job_title,  AVG(IF(wage_rate_unit='Hour',(wage_rate*2080),IF(wage_rate_unit='Week',(wag
e_rate*52), IF(wage_rate_unit='Year', wage_rate,0)))) as wage_rate   from h1b_application
where visa_application_year=2013 and  employer_name= 'GOOGLE INC.'  and ( job_title
LIKE '% SOFTWARE %'   OR job_title LIKE '% BUSINESS %' )   group by employer_name,
job_title order by wage_rate desc LIMIT 10;
```

Result :

```
Total MapReduce CPU Time Spent: 21 seconds 440 msec
OK
GOOGLE INC.     LEAD SOFTWARE ENGINEER  219288.0
GOOGLE INC.     STAFF SOFTWARE ENGINEER 206900.0
GOOGLE INC.     MANAGER, SOFTWARE ENGINEERING   200000.0
GOOGLE INC.     SR. SOFTWARE ENGINEER   164600.0
GOOGLE INC.     SENIOR SOFTWARE ENGINEER        159200.0
GOOGLE INC.     SENIOR SOFTWARE ENGINEER, PART  151812.0
GOOGLE INC.     QUANTITATIVE BUSINESS ANALYST   148000.0
GOOGLE INC.     NEW BUSINESS DEVELOPMENT MANAG  141000.0
GOOGLE INC.     HR BUSINESS PARTNER     128000.0
GOOGLE INC.     EMERGING BUSINESS LEAD,GEOCOMM  125000.0
Time taken: 62.405 seconds, Fetched: 10 row(s)
```

Conclusion:

- Oracle America, Inc. had the highest average wage followed by Google.

f. Maximum wage rate for top CA employers for the above job categories.

Objective of the query: To find the maximum wage rate for employers who file the most number of H-1B visa applications in California.

Description of the query :

To find the average wage rate offered by different employers in California in 2013.

Query:

select employer_name, job_title, MAX(wage_rate) as wage_rate from h1b_application where visa_application_year=2013 and employer_name in('INFOSYS LIMITED', 'TATA CONSULTANCY SERVICES LIMI', 'WIPRO LIMITED', 'DELOITTE CONSULTING LLP', 'ACCENTURE LLP') and (job_title LIKE '% BUSINESS %' OR job_title LIKE '% COMPUTER %' OR job_title LIKE '% TECHNOLOGY %' OR job_title LIKE '% DATABASE %' OR job_title LIKE '% SALES %' OR job_title LIKE '% CONSULTING %' OR job_title LIKE '% MANAGER %' OR job_title LIKE '% SOFTWARE %' OR job_title LIKE '% ENGINEERING %') group by employer_name, job_title order by wage_rate desc limit 50;

Query Output:

```
Total MapReduce CPU Time Spent: 22 seconds 20 msec
OK
ORACLE AMERICA, INC.    SERVICES SALES SENIOR VICE PRE  400000
GOOGLE INC.     SOFTWARE ENGINEERING MANAGER    250000
FACEBOOK, INC.  MANAGER, SOFTWARE ENGINEERING   220000
GOOGLE INC.     LEAD SOFTWARE ENGINEER  219288
FACEBOOK, INC.  DIRECTOR, BUSINESS OPERATIONS   218162
APPLE INC.      TEST ENGINEERING MANAGER 3      210000
GOOGLE INC.     STAFF SOFTWARE ENGINEER 206900
HCL AMERICA, INC.       OPERATIONS MANAGER - IV 206440
GOOGLE INC.     MANAGER, SOFTWARE ENGINEERING   200000
EBAY INC.       PRINCIPAL MTS, SOFTWARE ENGINE  200000
HCL AMERICA, INC.       SALES MANAGER - IV      197850
APPLE INC.      SOFTWARE ENGINEERING MANAGER    185000
EBAY INC.       MANAGER, SOFTWARE DEVELOPMENT   185000
GOOGLE INC.     MANAGER - DATABASE ADMINISTRAT  180000
HCL AMERICA, INC.       SALES MANAGER - III     179213
APPLE INC.      SOFTWARE DEVELOP MANAGER 2      178500
EBAY INC.       MTS 2, SOFTWARE ENGINEER        176748
ORACLE AMERICA, INC.    SENIOR SOFTWARE MANAGER 175000
GOOGLE INC.     SENIOR SOFTWARE ENGINEER        175000
APPLE INC.      PRINCIPAL SOFTWARE ENGR/DATA S  170000
APPLE INC.      SOFTWARE DEVELOP MANAGER 3      170000
APPLE INC.      FIRMWARE ENGINEER MANAGER 1     170000
EBAY INC.       MTS 3, SOFTWARE ENGINEER        170000
EBAY INC.       SR. PRODUCT MANAGER 2 - TECHNI  165000
GOOGLE INC.     SR. SOFTWARE ENGINEER   164600
APPLE INC.      HARDWARE DEVELOP MANAGER 2      162000
FACEBOOK, INC.  SOFTWARE ENGINEERING MANAGER    161773
EBAY INC.       SR. MTS, SOFTWARE ENGINEER      160000
```

Execution time of query:

```
Time taken: 65.346 seconds, Fetched: 50 row(s)
hive>
```

Verification of query: Verified the query for employer FACEBOOK INC for job category keywords BUSINESS and SOFTWARE.

Query :

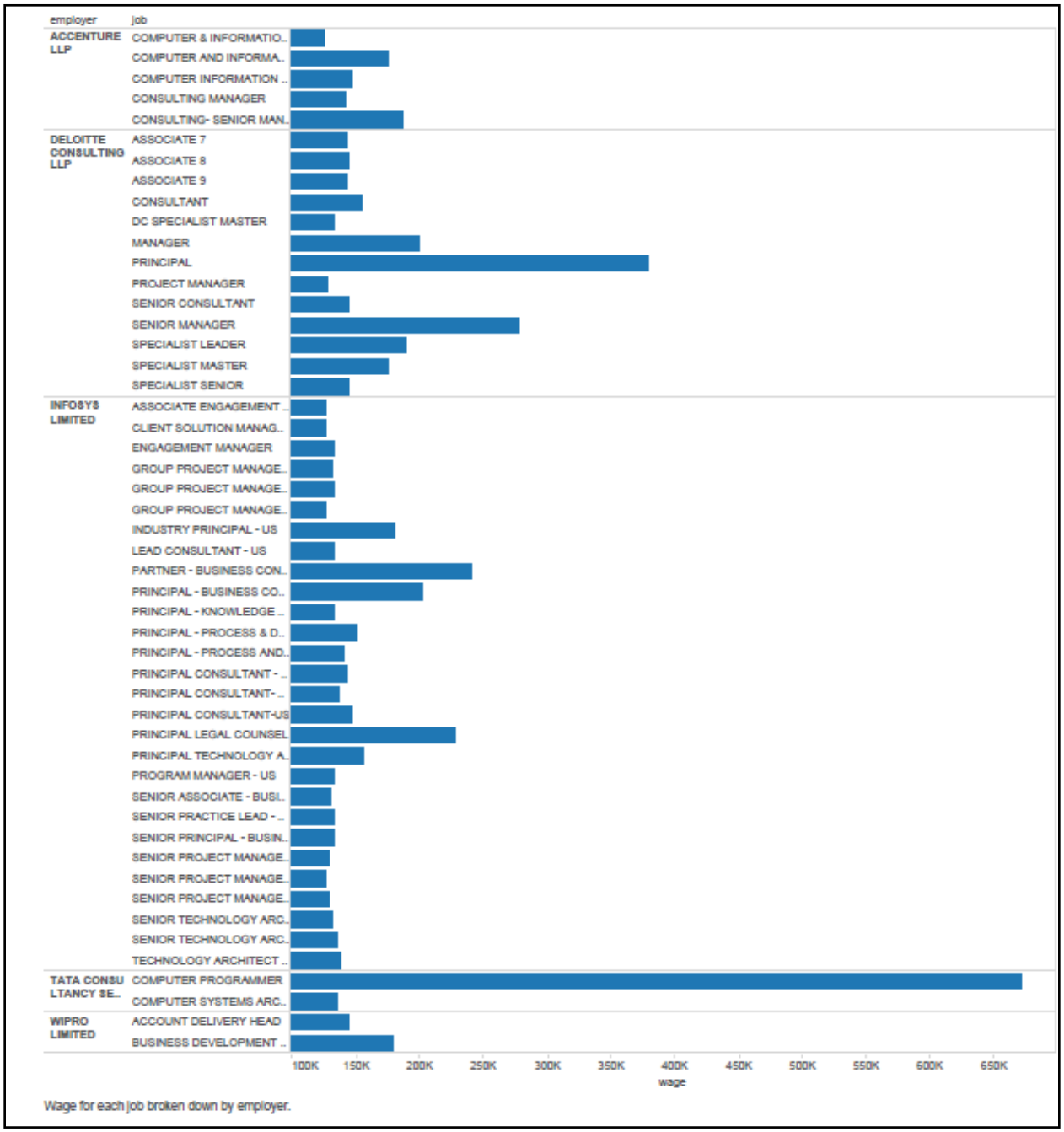
select employer_name, job_title, MAX(wage_rate) as wage_rate from h1b_application where visa_application_year=2013 and employer_name = 'FACEBOOK, INC.' and (job_title LIKE '% BUSINESS %' OR job_title LIKE '% SOFTWARE %') group by employer_name, job_title order by wage_rate desc LIMIT 5;

Result :

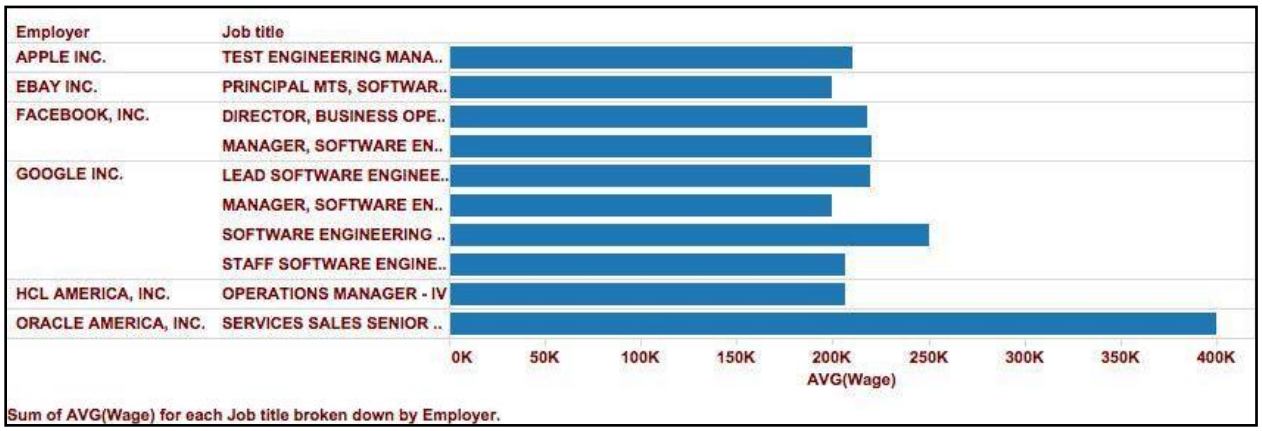
```
Total MapReduce CPU Time Spent: 21 seconds 470 msec
OK
FACEBOOK, INC.  MANAGER, SOFTWARE ENGINEERING   220000
FACEBOOK, INC.  DIRECTOR, BUSINESS OPERATIONS   218162
FACEBOOK, INC.  GLOBAL BUSINESS MANAGER 119142
Time taken: 67.716 seconds, Fetched: 3 row(s)
```

Visualization:

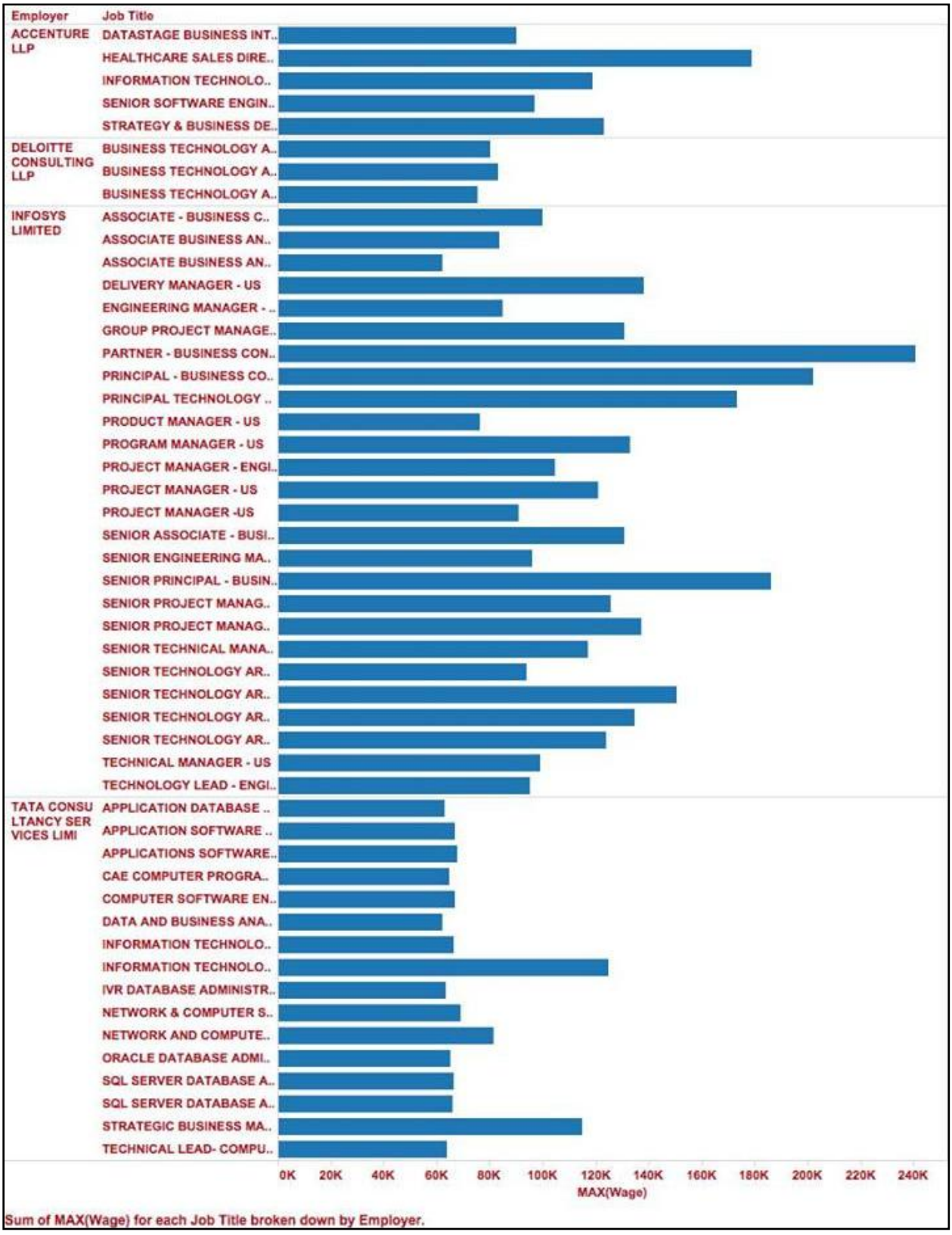
Average wage rate for employer with highest number of applications for all states :



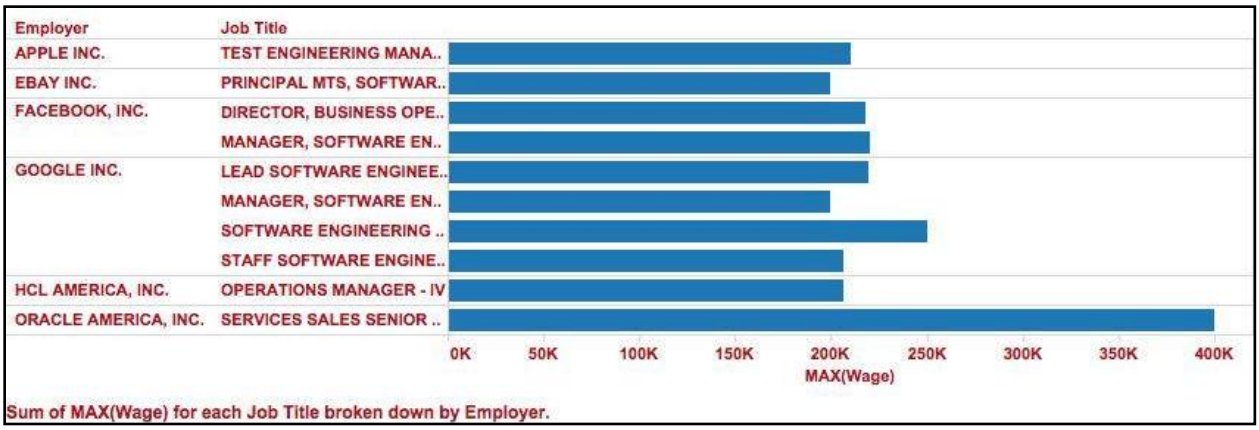
Average wage rate for employer with highest number of applications for CA:



Max wage rate for employer with highest number of applications for all states :



Max wage rate for employer with highest number of applications for CA:



Conclusion:

- Oracle America, Inc. had the maximum wage rate in California.

Comparison Results for CA and other states

- Here we see that though top employers are from other states than CA but employers in CA are offering highest paid job in categories related to (SOFTWARE, COMPUTER).
- The wage rate related to category BUSINESS does not differ much in between other states and CA.
- We also see that GOOGLE INC, FACEBOOK INC, APPLE INC are offering higher wages than all the other companies.

5. California job statistics analysis

Objective of the query: To find if the statistics for California has changed over the years(2010-2013).

Description of the query :

We wanted to find the positions within the employers that had the maximum average wage rate. We analyzed this data over the years 2010 to 2013 to see if there were any significant changes.

Query:

```
select visa_application_year , employer_name,
job_title,  AVG(IF(wage_rate_unit='Hour',(wage_rate*2080),IF(wage_rate_unit='Week',(wage_rate*52), IF(wage_rate_unit='Year', wage_rate,0)))) as wage_rate  from h1b_application where visa_application_year=2010 and  employer_name in('HCL AMERICA, INC.', 'GOOGLE INC.', 'IGATE TECHNOLOGIES INC.', 'INTEL CORPORATION', 'ORACLE AMERICA, INC.','APPLE INC.', 'EBAY INC.', 'FACEBOOK, INC.', 'CISCO SYSTEMS, INC.') and (job_title LIKE '% BUSINESS %'  OR job_title LIKE '% COMPUTER %'  OR job_title LIKE '% TECHNOLOGY %'  OR job_title LIKE '% DATABASE %'  OR job_title LIKE '% SALES %'  OR job_title LIKE '% CONSULTING %'  OR job_title LIKE '% MANAGER %' OR job_title LIKE '% SOFTWARE %' OR job_title LIKE '% ENGINEERING %'  OR job_title LIKE '% PROGRAMMER %')  group by visa_application_year , employer_name, job_title order by wage_rate desc LIMIT 20;
```

Query Output:

1. 2010:

```
Total MapReduce CPU Time Spent: 21 seconds 640 msec
OK
2010    APPLE INC.      IPHONE ENTERPRISE SALES SR. DI  235000.0
2010    GOOGLE INC.     SALES MANAGER (BUSINESS RELATI  198000.0
2010    CISCO SYSTEMS, INC.     DIRECTOR, BUSINESS DEVELOPMENT  171539.0
2010    APPLE INC.      R&D MANAGER 1   166500.0
2010    EBAY INC.       DIRECTOR OF PRODUCT, BUSINESS   160000.0
2010    GOOGLE INC.     MANAGER, SOFTWARE ENGINEERING   160000.0
2010    APPLE INC.      GRAPHICS ENGINEERING MANAGER    150000.0
2010    GOOGLE INC.     PROGRAM MANAGER (STRATEGIC NEG  150000.0
2010    APPLE INC.      SYSTEMS ENGINEERING MANAGER     148720.0
2010    EBAY INC.       PRINCIPLE SOFTWARE ENGINEER     141896.0
2010    EBAY INC.       MANAGER, SOFTWARE DEVELOPMENT   141440.0
2010    GOOGLE INC.     NEW BUSINESS DEVELOPMENT MANAG  141000.0
2010    EBAY INC.       STAFF DATABASE ENGINEER 135000.0
2010    APPLE INC.      GRAPHICS SOFTWARE ENGINEER      135000.0
2010    APPLE INC.      SPG WIRELESS ENGINEERING AND T  135000.0
2010    CISCO SYSTEMS, INC.     MANAGER, SOFTWARE DEVELOPMENT   132355.5
2010    EBAY INC.       STAFF SOFTWARE ENGINEER, SOA    130000.0
2010    APPLE INC.      PORTABLE MAC SOFTWARE ENGINEER  130000.0
2010    APPLE INC.      SENIOR SOFTWARE ENGINEER        127988.2
2010    EBAY INC.       STAFF SOFTWARE ENGINEER 126214.71428571429
Time taken: 63.052 seconds, Fetched: 20 row(s)
```

2. 2011:

```
Total MapReduce CPU Time Spent: 22 seconds 40 msec
OK
2011    INTEL CORPORATION       GRAPHICS SOFTWARE VALIDATION E  7.5178335E7
2011    GOOGLE INC.     DIRECTOR, SOFTWARE ENGINEERING  350000.0
2011    GOOGLE INC.     PRODUCT/PROJECT MANAGER (PEERI  245816.0
2011    GOOGLE INC.     SOFTWARE ENGINEERING MANAGER    202532.0
2011    APPLE INC.      DIRECTOR, WW CHANNEL SALES PRO  200000.0
2011    APPLE INC.      SENIOR WIRELESS TECHNOLOGY ARC  200000.0
2011    APPLE INC.      SENIOR MANAGER - MODULE DEVELO  200000.0
2011    GOOGLE INC.     HUMAN RESOURCES BUSINESS PARTN  196400.0
2011    APPLE INC.      SENIOR BUSINESS PLANNING MANAG  185000.0
2011    CISCO SYSTEMS, INC.     DIRECTOR, INTERNET BUSINESS SO  182270.0
2011    INTEL CORPORATION       SR. VP GENERAL MANAGER DIGITAL  180981.0
2011    INTEL CORPORATION       STRATEGIC BUSINESS DEV/HEAD OF  178256.0
2011    APPLE INC.      IOS ENGINEERING LEAD    175000.0
2011    APPLE INC.      RF/OTA SYSTEMS ENGINEERING LEA  170000.0
2011    GOOGLE INC.     PROJECT MANAGER (PARTNER STRAT  168800.0
2011    GOOGLE INC.     MARKETING MANAGER (AUDIENCE DE  165000.0
2011    APPLE INC.      SENIOR IOS SOFTWARE ENGINEER    165000.0
2011    APPLE INC.      CORE OS RELEASE ENGINEERING LE  160000.0
2011    APPLE INC.      SW ENG SYSTEMS MANAGER 2        150255.0
2011    APPLE INC.      GLOBAL SUPPLY MANAGER (GSM): L  150000.0
Time taken: 65.273 seconds, Fetched: 20 row(s)
```

3. 2012:

```
Total MapReduce CPU Time Spent: 22 seconds 40 msec
OK
2012    APPLE INC.      ASIC DESIGN ENGINEERING SR. DI  226000.0
2012    GOOGLE INC.     STAFF SOFTWARE ENGINEER 215033.33333333334
2012    GOOGLE INC.     SOFTWARE ENGINEERING MANAGER    210000.0
2012    INTEL CORPORATION       ENGINEERING MANAGER (PROGRAM D  191381.0
2012    APPLE INC.      SOFTWARE ENGINEERING APPS MANA  190000.0
2012    FACEBOOK, INC.  SENIOR SOFTWARE ENGINEER        185000.0
2012    GOOGLE INC.     MARKETING MANAGER (PRODUCT MAN  185000.0
2012    APPLE INC.      WW SALES CONTRACTS - SENIOR MA  185000.0
2012    APPLE INC.      SENIOR CAMERA TECHNOLOGY SPECI  180000.0
2012    GOOGLE INC.     SENIOR SOFTWARE ENGINEER        177750.0
2012    APPLE INC.      WW CHANNEL SALES PROG MGR - CA  175000.0
2012    APPLE INC.      SOFTWARE DEVELOP MANAGER 2      170602.0
2012    INTEL CORPORATION       ENGINEERING MANAGER (SOFTWARE   170165.0
2012    EBAY INC.       SR MANAGER MANAGED MARKETPLACE  165000.0
2012    EBAY INC.       MANAGER, SOFTWARE DEVELOPMENT   160000.0
2012    GOOGLE INC.     SR. SOFTWARE ENGINEER   159296.5
2012    FACEBOOK, INC.  MANAGER, BUSINESS OPERATIONS    156000.0
2012    APPLE INC.      SENIOR IOS SOFTWARE ENGINEER    155000.0
2012    APPLE INC.      MARKETING MANAGER 2     152568.0
2012    INTEL CORPORATION       SOFTWARE ENGINEERING MANAGER    151614.75
Time taken: 64.096 seconds, Fetched: 20 row(s)
```

4. 2013:

```
Total MapReduce CPU Time Spent: 22 seconds 10 msec
OK
2013    ORACLE AMERICA, INC.    SERVICES SALES SENIOR VICE PRE  400000.0
2013    GOOGLE INC.     SOFTWARE ENGINEERING MANAGER    223750.0
2013    GOOGLE INC.     LEAD SOFTWARE ENGINEER  219288.0
2013    FACEBOOK, INC.  DIRECTOR, BUSINESS OPERATIONS   218162.0
2013    APPLE INC.      TEST ENGINEERING MANAGER 3      210000.0
2013    GOOGLE INC.     STAFF SOFTWARE ENGINEER 206900.0
2013    GOOGLE INC.     MANAGER, SOFTWARE ENGINEERING   200000.0
2013    EBAY INC.       PRINCIPAL MTS, SOFTWARE ENGINE  200000.0
2013    FACEBOOK, INC.  MANAGER, SOFTWARE ENGINEERING   194180.0
2013    HCL AMERICA, INC.       OPERATIONS MANAGER - IV 190543.5
2013    APPLE INC.      SOFTWARE ENGINEERING MANAGER    185000.0
2013    GOOGLE INC.     MANAGER - DATABASE ADMINISTRAT  180000.0
2013    HCL AMERICA, INC.       SALES MANAGER - IV      179343.0
2013    APPLE INC.      SOFTWARE DEVELOP MANAGER 2      178500.0
2013    ORACLE AMERICA, INC.    SENIOR SOFTWARE MANAGER 175000.0
2013    APPLE INC.      PRINCIPAL SOFTWARE ENGR/DATA S  170000.0
2013    EBAY INC.       MTS 3, SOFTWARE ENGINEER        170000.0
2013    APPLE INC.      FIRMWARE ENGINEER MANAGER 1     170000.0
2013    APPLE INC.      SOFTWARE DEVELOP MANAGER 3      170000.0
2013    EBAY INC.       SR. PRODUCT MANAGER 2 - TECHNI  165000.0
Time taken: 64.135 seconds, Fetched: 20 row(s)
hive>
```

Execution time of query:

1. 2010 :

```
Time taken: 63.052 seconds, Fetched: 20 row(s)
```

2. 2011:

```
Time taken: 65.273 seconds, Fetched: 20 row(s)
```

3. 2012:

```
Time taken: 64.096 seconds, Fetched: 20 row(s)
```

4. 2013:

```
Time taken: 64.135 seconds, Fetched: 20 row(s)
hive>
```

Verification of query: Verified it against the employer HCL AMERICA, INC and GOOGLE INC for job keyword BUSINESS and COMPUTER

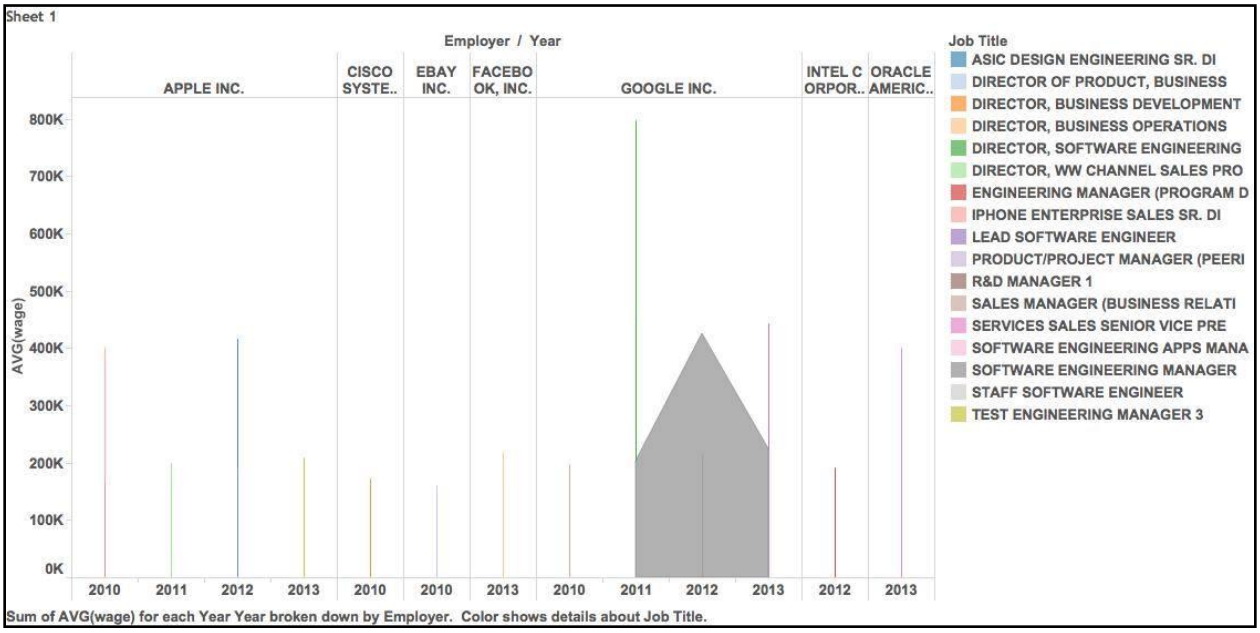
Query :

select visa_application_year , employer_name, job_title, AVG(IF(wage_rate_unit='Hour',(wage_rate*2080),IF(wage_rate_unit='Week',(wage_rate*52), IF(wage_rate_unit='Year', wage_rate,0)))) as wage_rate from h1b_application where visa_application_year=2010 and employer_name in('HCL AMERICA, INC.', 'GOOGLE INC.') and (job_title LIKE '% BUSINESS %' OR job_title LIKE '% COMPUTER %' OR job_title LIKE '% TECHNOLOGY %') group by visa_application_year , employer_name, job_title order by wage_rate desc LIMIT 5;

Result :

```
Total MapReduce CPU Time Spent: 21 seconds 260 msec
OK
2010    GOOGLE INC.     NEW BUSINESS DEVELOPMENT MANAG  141000.0
2010    GOOGLE INC.     PARTNER TECHNOLOGY MANAGER      114000.0
2010    GOOGLE INC.     ADVERTISING TECHNOLOGY TEAM LE  76900.0
2010    GOOGLE INC.     PEOPLE TECHNOLOGY OPERATIONS T  75000.0
2010    GOOGLE INC.     ADVERTISING TECHNOLOGY ASSOCIA  70800.0
Time taken: 62.155 seconds, Fetched: 5 row(s)
```

Visualization



Conclusion:

- GOOGLE INC and APPLE INC are the highest paying employer since 2010 in CA.
- FACEBOOK INC and ORACLE AMERICA INC are moving up and are among the top paying employers in 2013.

6. Analysis for top hiring companies in computer science and software engineering

a. For all states in US

Objective of the query: To find the top employers hiring in the field of computer science, software engineering and analyzing their changing trends over the years.

Description of the query :

In order to cover all the fields related to computer science and software engineering, we extracted the keywords related to this field. These include :
COMPUTER, DATABASE, CONSULTING, TECHNOLOGY, SOFTWARE, ENGINEERING, PROGRAMMER, DEVELOPER, ENGINEER, SYSTEM, CONSULTANT
The query was written in order to find the employers having the maximum number of applicants for particular job title who filed for visa. The query also accounts for the state in which this job title is filed and average wage for it by the employer.

Query:

```
select visa_application_year, employer_name, employer_state, job_title, count(*) as count , visa_application_year,
AVG(IF(wage_rate_unit='Hour',(wage_rate*2080),IF(wage_rate_unit='Week',(wage_rate*52),
IF(wage_rate_unit='Year', wage_rate,0)))) as wage_rate from h1b_application where visa_application_year = '2010' and (job_title LIKE '% COMPUTER %' OR job_title LIKE '% DATABASE %' OR job_title LIKE '% CONSULTING %' OR job_title LIKE '% TECHNOLOGY %' OR job_title LIKE '% SOFTWARE %' OR job_title LIKE '% ENGINEERING %' OR job_title LIKE '% PROGRAMMER %' OR job_title LIKE '% DEVELOPER %' OR job_title LIKE '% ENGINEER %' OR job_title LIKE '% SYSTEM %' OR job_title LIKE '% CONSULTANT %') group by visa_application_year, job_title, employer_name, employer_state having count > 125 order by count desc, wage_rate desc limit 20;
```

Query Output:

1. 2010:

```
Total MapReduce CPU Time Spent: 22 seconds 980 msec
OK
2010    MICROSOFT CORPORATION   WA      SOFTWARE DEVELOPMENT ENGINEER   1736    2010    88945.70161290323
2010    FUJITSU AMERICA, INC.   CA      COMPUTER SOFTWARE ENGINEER, AP  1336    2010    220927.4393712575
2010    WIPRO LIMITED   NJ      BUSINESS SYSTEM ANALYST 438     2010    67827.31278538813
2010    YAHOO! INC.     CA      SOFTWARE ENGINEER (TECHNICAL Y  304     2010    84012.28618421052
2010    AMAZON CORPORATE LLC    WA      SOFTWARE DEVELOPMENT ENGINEER   290     2010    96230.50689655173
2010    ORACLE AMERICA, INC.    CA      SOFTWARE ENGINEER (SOFTWARE DE  231     2010    808222.4805194805
2010    DELOITTE CONSULTING LLP PA      BUSINESS TECHNOLOGY ANALYST     203     2010    67044.81773399014
2010    LARSEN & TOUBRO INFOTECH LIMIT  NJ      COMPUTER PROGRAMMER ANALYST     174     2010    61677.32183908046
2010    QUALCOMM INCORPORATED   CA      SENIOR SOFTWARE ENGINEER        168     2010    89581.30952380953
2010    MICROSOFT CORPORATION   WA      SENIOR SOFTWARE DEVELOPMENT EN  153     2010    117329.46405228759
Time taken: 64.116 seconds, Fetched: 10 row(s)
```

2. 2011:

```
Total MapReduce CPU Time Spent: 22 seconds 850 msec
OK
2011    MICROSOFT CORPORATION   WA      SOFTWARE DEVELOPMENT ENGINEER   1742    2011    94787.46096440873
2011    FUJITSU AMERICA, INC.   CA      COMPUTER SOFTWARE ENGINEER, AP  892     2011    82024.84753363229
2011    WIPRO LIMITED   NJ      BUSINESS SYSTEM ANALYST 523     2011    377302.6673040153
2011    ORACLE AMERICA, INC.    CA      SOFTWARE ENGINEER (SOFTWARE DE  341     2011    176557.71260997068
2011    AMAZON CORPORATE LLC    WA      SOFTWARE DEVELOPMENT ENGINEER   325     2011    768775.4830769231
2011    COGNIZANT TECHNOLOGY SOLUTIONS  NJ      SENIOR SYSTEM ANALYST   281     2011    66042.98932384342
2011    DELOITTE CONSULTING LLP PA      BUSINESS TECHNOLOGY ANALYST     257     2011    69895.10505836576
2011    HEXAWARE TECHNOLOGIES, INC.     NJ      COMPUTER PROGRAMMER ANALYST     212     2011    57097.85849056604
2011    LARSEN & TOUBRO INFOTECH LIMIT  NJ      SOFTWARE ENGINEER AND TESTER    203     2011    57347.32019704433
2011    LARSEN & TOUBRO INFOTECH LIMIT  NJ      COMPUTER PROGRAMMER ANALYST     201     2011    65812.78606965175
2011    MICROSOFT CORPORATION   WA      SENIOR SOFTWARE DEVELOPMENT EN  192     2011    118401.59375
2011    WIPRO LIMITED   NJ      BUSINESS SYSTEM ANALYST LEVEL   183     2011    84214.73224043715
2011    ORACLE AMERICA, INC.    CA      SOFTWARE ENGINEER (APPLICATION  154     2011    88239.81168831169
2011    QUALCOMM INCORPORATED   CA      SENIOR SOFTWARE ENGINEER        151     2011    85482.18543046358
2011    LARSEN & TOUBRO INFOTECH LIMIT  NJ      SOFTWARE ENGINEER & TESTER      131     2011    56800.36641221374
Time taken: 64.121 seconds, Fetched: 15 row(s)
```

3. 2012:

```
Total MapReduce CPU Time Spent: 23 seconds 10 msec
OK
2012    MICROSOFT CORPORATION   WA      SOFTWARE DEVELOPMENT ENGINEER   1416    2012    104239.42443502825
2012    WIPRO LIMITED   NJ      BUSINESS SYSTEM ANALYST 632     2012    70085.5506329114
2012    FUJITSU AMERICA, INC.   VA      COMPUTER SOFTWARE ENGINEER, AP  532     2012    83737.34962406015
2012    ORACLE AMERICA, INC.    CA      SOFTWARE ENGINEER (SOFTWARE DE  515     2012    431564.87378640776
2012    AMAZON CORPORATE LLC    WA      SOFTWARE DEVELOPMENT ENGINEER   474     2012    103011.92194092827
2012    MASTECH, INC., A MASTECH HOLDI  PA      SENIOR SOFTWARE DEVELOPER       463     2012    87539.44060475162
2012    HEXAWARE TECHNOLOGIES, INC.     NJ      COMPUTER PROGRAMMER ANALYST     414     2012    57348.78502415459
2012    WIPRO LIMITED   NJ      TEST ENGINEER LEVEL 2   311     2012    64951.12861736334
2012    DELOITTE CONSULTING LLP PA      BUSINESS TECHNOLOGY ANALYST     301     2012    71028.05315614618
2012    IBM INDIA PRIVATE LIMITED       NC      SENIOR SYSTEM ENGINEER  294     2012    478648.87074829935
2012    QUALCOMM INCORPORATED   CA      SENIOR SOFTWARE ENGINEER        222     2012    85243.93693693694
2012    MICROSOFT CORPORATION   WA      SENIOR SOFTWARE DEVELOPMENT EN  215     2012    131300.83720930232
2012    ORACLE AMERICA, INC.    CA      SOFTWARE ENGINEER (APPLICATION  208     2012    100724.10576923077
2012    LARSEN & TOUBRO INFOTECH LIMIT  NJ      SOFTWARE ENGINEER & TESTER      181     2012    56593.127071823204
2012    SYNECHRON, INC. NJ      SOFTWARE ENGINEER APPLICATIONS  164     2012    67609.48780487805
2012    INFOSYS LIMITED TX      LEAD CONSULTANT - US    154     2012    87186.06493506493
2012    BLOOMBERG, LP   NY      SENIOR SOFTWARE DEVELOPER       152     2012    128376.50657894737
2012    INFOSYS LIMITED TX      SYSTEMS ENGINEER - US   145     2012    61926.606896551726
2012    CMC AMERICAS, INC.      LA      SYSTEMS ENGINEER (15-1099.02)   137     2012    65089.24087591241
Time taken: 64.113 seconds, Fetched: 19 row(s)
```

4. 2013:

```
Total MapReduce CPU Time Spent: 24 seconds 930 msec
OK
2013    INFOSYS LIMITED TX      SYSTEMS ENGINEER - US   2606    2013    62771.381427475055
2013    MICROSOFT CORPORATION   WA      SOFTWARE DEVELOPMENT ENGINEER   1106    2013    108024.14466546112
2013    INFOSYS LIMITED TX      ASSOCIATE CONSULTANT - US       902     2013    65893.93902439025
2013    INFOSYS LIMITED TX      LEAD CONSULTANT - US    673     2013    96306.92570579494
2013    MASTECH, INC., A MASTECH HOLDI  PA      SENIOR SOFTWARE DEVELOPER       586     2013    88494.14505119454
2013    AMAZON CORPORATE LLC    WA      SOFTWARE DEVELOPMENT ENGINEER   559     2013    107612.65474060822
2013    INFOSYS LIMITED TX      TEST ENGINEER - US      559     2013    62272.58139534884
2013    WIPRO LIMITED   NJ      BUSINESS SYSTEM ANALYST 525     2013    110875.12571428572
2013    FUJITSU AMERICA, INC.   PA      COMPUTER SOFTWARE ENGINEER, AP  409     2013    82941.77017114914
2013    ACCENTURE LLP   IL      COMPUTER SPECIALIST / SYSTEM S  379     2013    66260.15831134564
2013    WIPRO LIMITED   NJ      TEST ENGINEER LEVEL 2   374     2013    66037.3101604278
2013    DELOITTE CONSULTING LLP PA      BUSINESS TECHNOLOGY ANALYST     350     2013    72663.08571428571
2013    INFOSYS LIMITED TX      SENIOR TECHNOLOGY ARCHITECT -   315     2013    110187.08888888889
2013    INFOSYS LIMITED TX      PRINCIPAL CONSULTANT - US       301     2013    121976.15282392026
2013    LARSEN & TOUBRO INFOTECH LIMIT  NJ      SOFTWARE ENGINEER & TESTER      290     2013    59447.43448275862
2013    HEXAWARE TECHNOLOGIES, INC.     NJ      COMPUTER PROGRAMMER ANALYST     279     2013    59698.6129032258
2013    MICROSOFT CORPORATION   WA      SENIOR SOFTWARE DEVELOPMENT EN  213     2013    133931.60563380283
2013    QUALCOMM TECHNOLOGIES, INC.     CA      SENIOR SOFTWARE ENGINEER        187     2013    89865.94652406417
2013    IBM INDIA PRIVATE LIMITED       NC      SENIOR SYSTEM ENGINEER  186     2013    67814.15053763441
2013    TATA CONSULTANCY SERVICES LIMI  MD      NETWORK AND COMPUTER SYSTEMS A  176     2013    64919.318181818184
Time taken: 64.057 seconds, Fetched: 20 row(s)
```

Execution time of query:

1. 2010:

```
Time taken: 64.116 seconds, Fetched: 10 row(s)
```

2. 2011:

```
Time taken: 64.121 seconds, Fetched: 15 row(s)
```

3. 2012:

```
Time taken: 64.113 seconds, Fetched: 19 row(s)
```

4. 2013:

```
Time taken: 64.057 seconds, Fetched: 20 row(s)
```

Verification of query: Verified for the year 2013 and job key word COMPUTER and DATABSE.

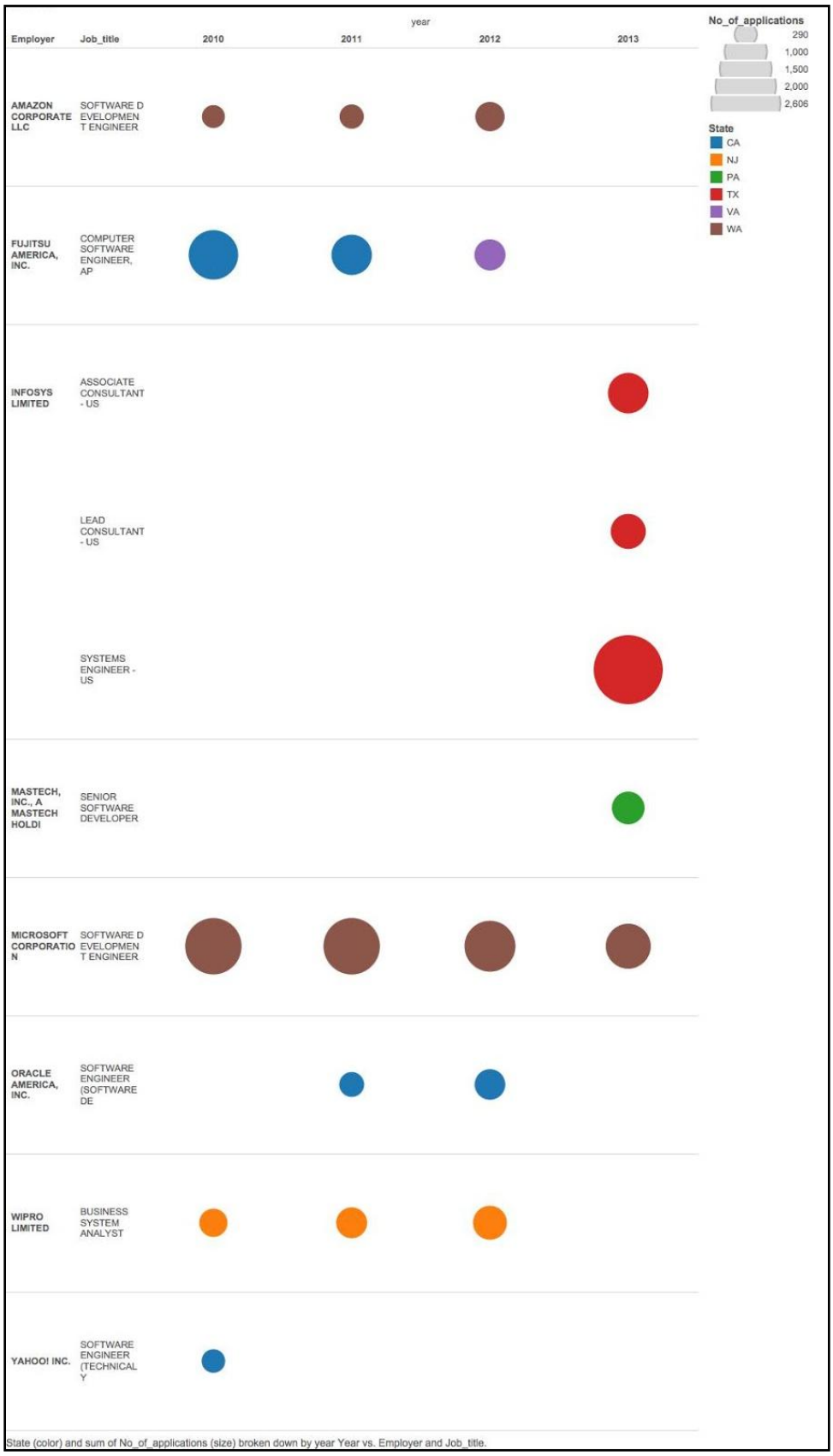
Query :

```
select visa_application_year, employer_name, employer_state, job_title, count(*) as count , AVG(IF(wage_rate_unit='Hour',(wage_rate*2080),IF(wage_rate_unit='Week',(wage_rate*52), IF(wage_rate_unit='Year', wage_rate,0)))) as wage_rate from h1b_application where visa_application_year = '2013' and (job_title LIKE '% COMPUTER %' OR job_title LIKE '% DATABASE %' ) group by visa_application_year, job_title, employer_name, employer_state order by count desc, wage_rate desc limit 5;
```

Result :

```
Total MapReduce CPU Time Spent: 21 seconds 80 msec
OK
2013    TATA CONSULTANCY SERVICES LIMI  MD      NETWORK AND COMPUTER SYSTEMS A  176     64919.318181818184
2013    PERSISTENT SYSTEMS, INC.        CA      SR. COMPUTER PROGRAMMER ANALYS  100     78165.12
2013    PERSISTENT SYSTEMS, INC.        CA      SENIOR COMPUTER PROGRAMMER ANA  69      77105.37681159421
2013    IGATE TECHNOLOGIES INC. CA      NETWORK AND COMPUTER SYSTEMS A  55      66532.72727272728
2013    HCL AMERICA, INC.       CA      NETWORK AND COMPUTER SYSTEMS A  53      65848.09433962264
Time taken: 63.159 seconds, Fetched: 5 row(s)
```

Visualization



b. Top 10 states for year 2013 offering highest jobs in software.

Objective of the query: To find the top employers hiring in the field of computer science, software engineering for 2013.

Description of the query :

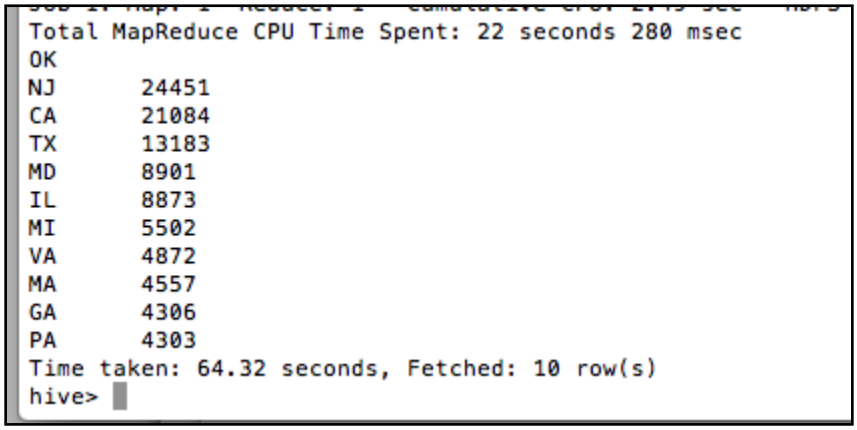
In order to cover all the fields related to computer science and software engineering, we extracted the keywords related to this field. These include :
COMPUTER, DATABASE, CONSULTING, TECHNOLOGY, SOFTWARE, ENGINEERING, PROGRAMMER, DEVELOPER, ENGINEER, SYSTEM, CONSULTANT
The query was written in order to find the employers having the maximum number of applicants for particular job title who filed for visa in year 2013.

Query:

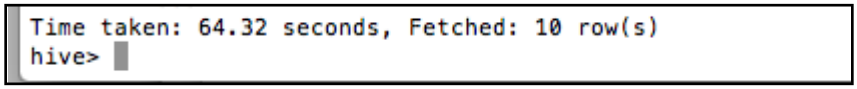
```
select employer_state, count(*) as num from h1b_test where job_title in ('PROGRAMMER ANALYST', 'SOFTWARE ENGINEER', 'COMPUTER PROGRAMMER', 'Programmer Analyst', 'SYSTEMS ANALYST', 'COMPUTER SYSTEMS ANALYST', 'Software Engineer', 'PROGRAMMER/ANALYST', 'SENIOR SOFTWARE ENGINEER', 'SOFTWARE DEVELOPER', 'Systems Analyst', 'COMPUTER PROGRAMMER ANALYST', 'SOFTWARE DEVELOPMENT ENGINEER', 'DATABASE ADMINISTRATOR', 'Computer Programmer', 'TECHNOLOGY LEAD', 'TECHNOLOGY ANALYST', 'SOFTWARE DEVELOPMENT ENGINEER', 'Programmer/Analyst', 'COMPUTER SOFTWARE ENGINEER, AP', 'PROGRAMMER', 'COMPUTER SOFTWARE ENGINEER', 'SYSTEMS ENGINEER', 'COMPUTER SYSTEMS ENGINEER', 'SYSTEMS ADMINISTRATOR', 'PROGRAMMER ANALYST', 'SYSTEM ANALYST', 'Computer Systems Analyst', 'COMPUTER PROGRAMMERS', 'Senior Software Engineer', 'SENIOR PROGRAMMER ANALYST', 'SOFTWARE QUALITY ASSURANCE ENG', 'COMPUTER SYSTEMS ANALYSTS', 'SENIOR SOFTWARE DEVELOPER', 'NETWORK ENGINEER', 'SENIOR SYSTEMS ANALYST', 'COMPUTER SYSTEM ANALYST', 'SOFTWARE PROGRAMMER', 'PROGRAMMER / ANALYST', 'WEB DEVELOPER', 'Software Developer', 'Database Administrator', 'COMPUTER PROGRAMMER/CONFIGURER', 'APPLICATION DEVELOPER', 'NETWORK AND COMPUTER SYSTEMS A', 'SR. SOFTWARE ENGINEER', 'SYSTEM ADMINISTRATOR', 'SOFTWARE DEVELOPMENT ENGINEER', 'ASSOCIATE CONSULTANT', 'LEAD CONSULTANT', 'TEST ENGINEER - US', 'COMPUTER SPECIALIST / SYSTEM S', 'SENIOR SYSTEM ENGINEER', 'NETWORK AND COMPUTER SYSTEMS A', 'COMPUTER SYSTEMS ENGINEER', 'SYSTEMS ENGINEER - US', 'COMPUTER SYSTEMS ANALYST 2', 'LEAD CONSULTANT - US', 'COMPUTER SYSTEMS ENGINEER/ARCH', 'TEST ENGINEER LEVEL 2', 'SENIOR SYSTEMS ANALYST JC60', 'COMPUTER SYSTEMS ANALYST 3', 'SENIOR TECHNOLOGY ARCHITECT -', 'PRINCIPAL CONSULTANT', 'SOFTWARE ENGINEER &
```

TESTER', 'SENIOR SOFTWARE DEVELOPMENT EN', 'TEST ENGINEER LEVEL 1', 'SOFTWARE DEVELOPER 2', 'COMPUTER PROGRAMMER / CONFIGUR', 'TECHNOLOGY LEAD - ENGINEERING', 'SOFTWARE ENGINEER AND TESTER', 'COMPUTER SYSTEMS ARCHITECT', 'COMPUTER SYSTEMS ENGINEERS/ARC', 'SOFTWARE DEV ENGIN', 'SOFTWARE DEVELOPER 3' , 'EMBEDDED SYSTEMS SPECIALIST', 'SOFTWARE ENGINEER 3', 'SR. COMPUTER PROGRAMMER ANALYS', 'SR. PROGRAMMER ANALYST', 'STAFF SOFTWARE ENGINEER', 'SOFTWARE ENGINEER 2', 'SOFTWARE DEVELOPER - II','SOFTWARE DEVELOPER - I', 'SOFTWARE ENGINEER APPLICATIONS', 'SOFTWARE ENGINEER IN TEST', 'SENIOR COMPUTER PROGRAMMER ANA', 'GRAPHICS SOFTWARE ENGINEER', 'SOFTWARE ENGINEER (SOFTWARE DE', 'LEAD CONSULTANT - INFRASTRUCTU', 'COMPUTER SYSTEMS ENGINEER (PRI', 'SOFTWARE DEVELOPER 4', 'HARDWARE DEVELOPER 2') and visa_application_year = '2013' group by employer_state order by num DESC limit 10;

Query Output:



Execution time of query:



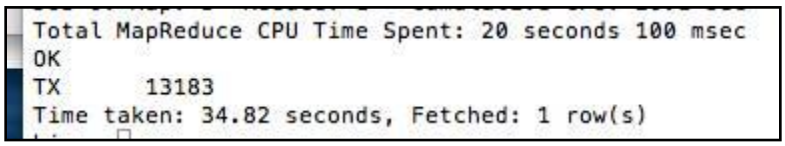
Verification of query: Verified for state TX

Query :

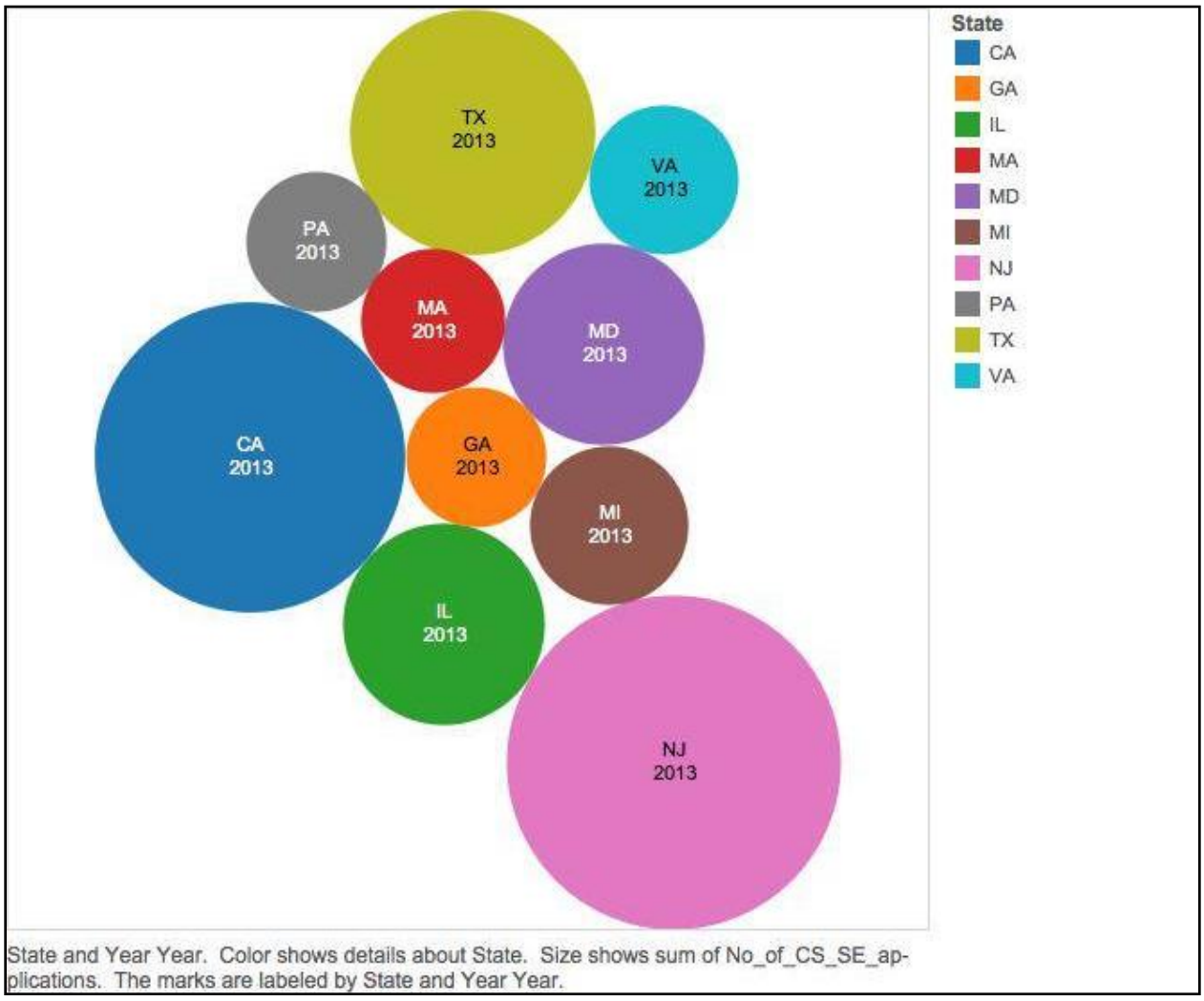
select employer_state, count(*) as num from h1b_test where job_title in ('PROGRAMMER ANALYST', 'SOFTWARE ENGINEER', 'COMPUTER PROGRAMMER', 'Programmer Analyst', 'SYSTEMS ANALYST', 'COMPUTER SYSTEMS ANALYST', 'Software Engineer', 'PROGRAMMER/ANALYST', 'SENIOR SOFTWARE ENGINEER', 'SOFTWARE DEVELOPER', 'Systems Analyst', 'COMPUTER PROGRAMMER ANALYST', 'SOFTWARE

DEVELOPMENT ENGINEER', 'DATABASE ADMINISTRATOR', 'Computer Programmer', 'TECHNOLOGY LEAD', 'TECHNOLOGY ANALYST', 'SOFTWARE DEVELOPMENT ENGINEER', 'Programmer/Analyst', 'COMPUTER SOFTWARE ENGINEER, AP', 'PROGRAMMER', 'COMPUTER SOFTWARE ENGINEER', 'SYSTEMS ENGINEER', 'COMPUTER SYSTEMS ENGINEER', 'SYSTEMS ADMINISTRATOR', 'PROGRAMMER ANALYST', 'SYSTEM ANALYST', 'Computer Systems Analyst', 'COMPUTER PROGRAMMERS', 'Senior Software Engineer', 'SENIOR PROGRAMMER ANALYST', 'SOFTWARE QUALITY ASSURANCE ENG', 'COMPUTER SYSTEMS ANALYSTS', 'SENIOR SOFTWARE DEVELOPER', 'NETWORK ENGINEER', 'SENIOR SYSTEMS ANALYST', 'COMPUTER SYSTEM ANALYST', 'SOFTWARE PROGRAMMER', 'PROGRAMMER / ANALYST', 'WEB DEVELOPER', 'Software Developer', 'Database Administrator', 'COMPUTER PROGRAMMER/CONFIGURER', 'APPLICATION DEVELOPER', 'NETWORK AND COMPUTER SYSTEMS A', 'SR. SOFTWARE ENGINEER', 'SYSTEM ADMINISTRATOR', 'SOFTWARE DEVELOPMENT ENGINEER', 'ASSOCIATE CONSULTANT', 'LEAD CONSULTANT', 'TEST ENGINEER - US', 'COMPUTER SPECIALIST / SYSTEM S', 'SENIOR SYSTEM ENGINEER', 'NETWORK AND COMPUTER SYSTEMS A', 'COMPUTER SYSTEMS ENGINEER', 'SYSTEMS ENGINEER - US', 'COMPUTER SYSTEMS ANALYST 2', 'LEAD CONSULTANT - US', 'COMPUTER SYSTEMS ENGINEER/ARCH', 'TEST ENGINEER LEVEL 2', 'SENIOR SYSTEMS ANALYST JC60', 'COMPUTER SYSTEMS ANALYST 3', 'SENIOR TECHNOLOGY ARCHITECT -', 'PRINCIPAL CONSULTANT', 'SOFTWARE ENGINEER & TESTER', 'SENIOR SOFTWARE DEVELOPMENT EN', 'TEST ENGINEER LEVEL 1', 'SOFTWARE DEVELOPER 2', 'COMPUTER PROGRAMMER / CONFIGUR', 'TECHNOLOGY LEAD - ENGINEERING', 'SOFTWARE ENGINEER AND TESTER', 'COMPUTER SYSTEMS ARCHITECT', 'COMPUTER SYSTEMS ENGINEERS/ARC', 'SOFTWARE DEV ENGIN', 'SOFTWARE DEVELOPER 3' , 'EMBEDDED SYSTEMS SPECIALIST', 'SOFTWARE ENGINEER 3', 'SR. COMPUTER PROGRAMMER ANALYS', 'SR. PROGRAMMER ANALYST', 'STAFF SOFTWARE ENGINEER', 'SOFTWARE ENGINEER 2', 'SOFTWARE DEVELOPER - II','SOFTWARE DEVELOPER - I', 'SOFTWARE ENGINEER APPLICATIONS', 'SOFTWARE ENGINEER IN TEST', 'SENIOR COMPUTER PROGRAMMER ANA', 'GRAPHICS SOFTWARE ENGINEER', 'SOFTWARE ENGINEER (SOFTWARE DE', 'LEAD CONSULTANT - INFRASTRUCTU', 'COMPUTER SYSTEMS ENGINEER (PRI', 'SOFTWARE DEVELOPER 4', 'HARDWARE DEVELOPER 2') and visa_application_year = '2013' and employer_state ='TX' group by employer_state ;

Result :



Visualization



State and Year Year. Color shows details about State. Size shows sum of No_of_CS_SE_applications. The marks are labeled by State and Year Year.

Conclusion:

- New Jersey and California lead in the number of jobs for software in 2013.

7. Analysis of job market from year 2010 to 2013 in terms of h1b applications

Objective of the query: To find how the number of visa applications have changed from 2008 to 2013.

Description of the query:

We wanted to find how the number of applications filed have been changing over these years. We wanted to see the ups and downs in market.

Query:

select employer_state, count(*) as no_of_application from h1b_application where visa_application_year ='2008' GROUP BY employer_state ORDER BY no_of_application desc LIMIT 10 ;

Result:

1. 2008:

```
Total MapReduce CPU Time Spent: 17 seconds 420 msec
OK
CA      83182
NJ      82429
NY      48877
TX      45338
IL      33356
VA      26969
PA      21567
MA      21164
FL      18734
MI      18379
Time taken: 59.929 seconds, Fetched: 10 row(s)
```

2. 2009:

```
Total MapReduce CPU Time Spent: 17 seconds 610 msec
OK
CA      43580
NJ      33001
NY      26650
TX      22676
IL      14331
VA      12714
MA      11882
PA      10912
FL      9877
GA      8227
Time taken: 60.134 seconds, Fetched: 10 row(s)
```

3. 2010:

```
Total MapReduce CPU Time Spent: 17 seconds 410 msec
OK
CA      58327
NJ      36641
NY      35636
TX      27539
IL      17909
PA      14133
FL      13975
MA      13880
VA      12047
GA      9805
Time taken: 61.193 seconds, Fetched: 10 row(s)
```

4. 2011:

```
Total MapReduce CPU Time Spent: 17 seconds 640 msec
OK
CA      62054
NJ      42461
NY      35344
TX      29171
IL      19209
PA      16964
MA      14960
FL      14171
VA      13214
MD      13001
Time taken: 62.165 seconds, Fetched: 10 row(s)
```

5. 2012:

```
Total MapReduce CPU Time Spent: 17 seconds 490 msec
OK
CA      68856
NJ      53791
TX      48365
NY      35989
IL      21526
PA      19685
MA      17345
VA      14711
FL      13986
MD      13971
Time taken: 62.09 seconds, Fetched: 10 row(s)
```

6. 2013:

```
Total MapReduce CPU Time Spent: 17 seconds 620 msec
OK
CA      74633
TX      66760
NJ      53977
NY      34288
IL      24796
PA      21255
MA      16069
MD      15806
VA      14680
MI      14033
Time taken: 60.294 seconds, Fetched: 10 row(s)
```

Execution time of query:

1. 2008:

```
Time taken: 59.929 seconds, Fetched: 10 row(s)
```

2. 2009:

```
Time taken: 60.134 seconds, Fetched: 10 row(s)
```

3. 2010:

```
Time taken: 61.193 seconds, Fetched: 10 row(s)
```

4. 2011:

```
Time taken: 62.165 seconds, Fetched: 10 row(s)
```

5. 2012:

```
Time taken: 62.09 seconds, Fetched: 10 row(s)
```

6. 2013:

```
Time taken: 60.294 seconds, Fetched: 10 row(s)
```

Verification of query: Verified for year 2012

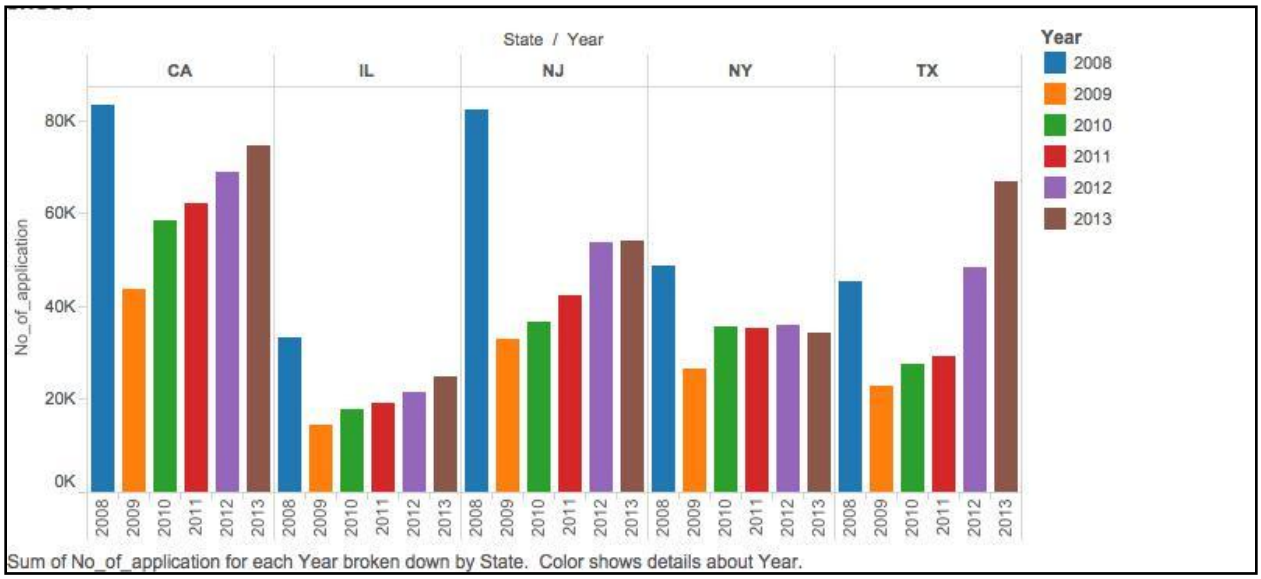
Query :

select employer_state, count(*) as no_of_application from h1b_application where visa_application_year ='2012' and employer_state ='CA' ORDER BY no_of_application desc LIMIT 10 ;

Result :

```
Total MapReduce CPU Time Spent: 17 seconds 710 msec
OK
CA      68856
Time taken: 61.096 seconds, Fetched: 1 row(s)
```

Visualization



Conclusion:

- In 2008, the job market was on peak.
- In 2009, the number of applications are almost half as compared to 2008 due to recession.
- Number of applications have started increasing again from 2010.
- In year 2013, the number of applications are almost close to number of applications in 2008.
- We can conclude that job market has reached the peak again.
- So, Are we close to recession again?

Conclusion

This concludes our journey through the H-1B Data Analysis. We found some very interesting facts about the data and also some not so interesting facts. We encountered many hurdles in this journey but we were able to complete it successfully. We were able to get hands on training on Hadoop and Hive. Using samples of data to test queries, verify queries with other queries after writing them, and also visualizing and making sense of the data was all that we focused.

Thank you for the wonderful experience!