



# **BIG DATA & HADOOP**

## **THE FUTURE OF THE INFORMATION ECONOMY!**

H1-B VISA PETITIONS  
(2011-2016)

[H1-B CASE STUDY](#)

Anchal Sadhu | Big Data and Hadoop

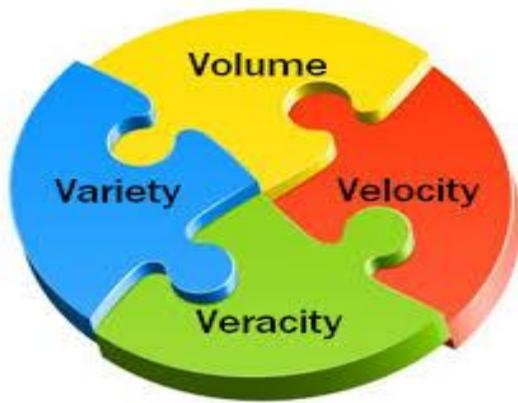
## A Brief Introduction about Various Technologies

### BIG DATA



Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information private

## 4V'S OF Big Data



### Volume

It is estimated that, on an average, 2.3 trillion gigabytes of data are generated every day. Forget analyzing, simply capturing such quantities of data is impractical. Most companies in the US have at least 100,000 gigabytes of data stored; and almost all of them will tell you that they aren't collecting enough data.

The right approach is to fight the urge of making your company's server a data dump. Efforts must be made to employ the right software to filter the relevant data.

### Variety

Along with quantity, the diversity of data is equally important. The variety in data can be in terms of the devices or the sources of data generation.

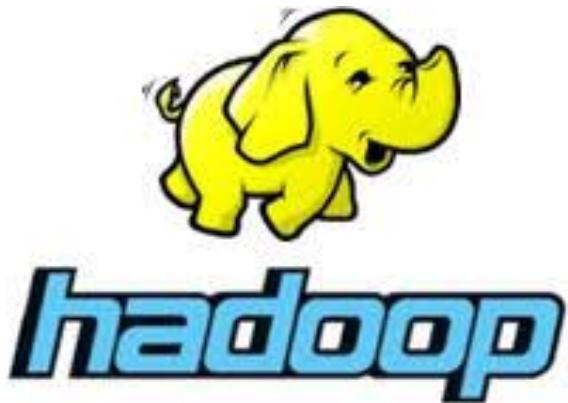
## **Velocity**

Not only is the volume of data ever increasing, but the rate of data generation (from the Internet of Things, social media, etc.) is increasing as well.

## **Veracity**

Not all data is good. In fact, unfiltered data is more likely to be bad than good. Although data quality and usability depend largely on the source, you can never rule out junk. This unreliability of data makes many business heads reluctant to rely on information analysis. That's the wrong approach.

## HADOOP

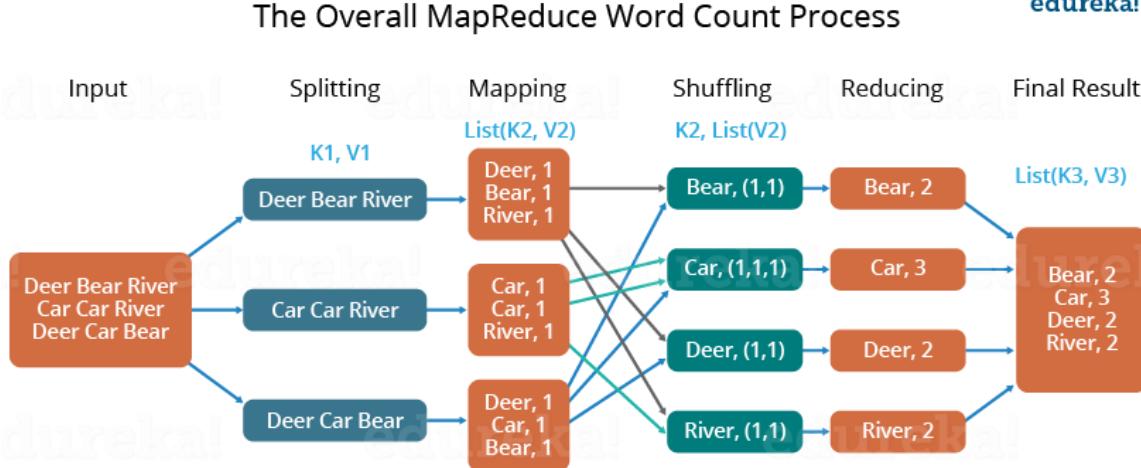


Apache Hadoop is an open-source software framework used for distributed storage and processing of very large data sets.

It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are a common occurrence and should be automatically handled by the frame.

# MapReduce

edureka!

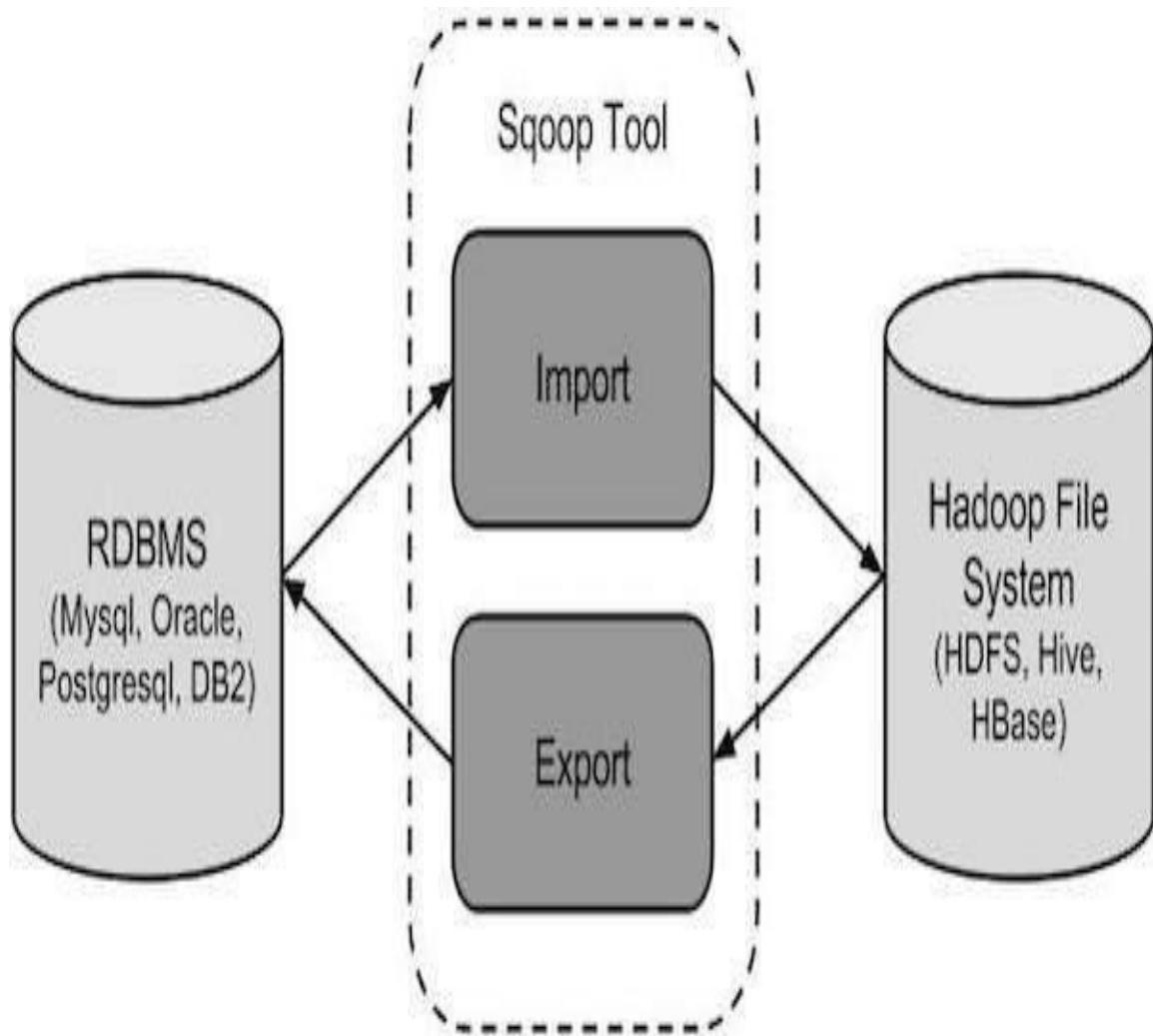


- ❑ Hadoop MapReduce (Hadoop Map/Reduce) is a software framework for distributed processing of large data sets on compute clusters of commodity hardware. It is a sub-project of the Apache Hadoop project. The framework takes care of scheduling tasks, monitoring them and re-executing any failed tasks.
- ❑ A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically, both the input and the output of the job are stored in a file-system.

## SQOOP

SQOOP is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases.





## HIVE



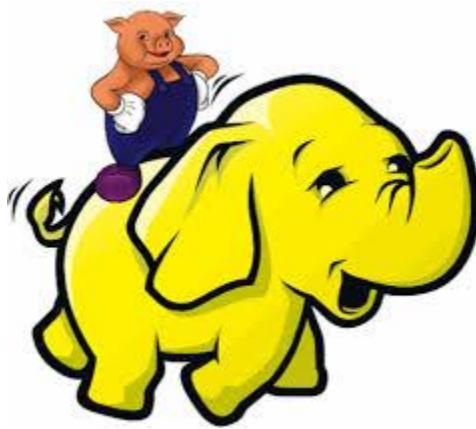
Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis.

Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

It is built on top of Hadoop and developed by Facebook. Hive provides a way to query the data using a SQL-like query language called HiveQL (Hive query Language).

Internally, a compiler translates HiveQL statements into MapReduce jobs, which are then submitted to Hadoop framework for execution.

## Apache PIG



Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin.

Pig can execute its Hadoop jobs in Map Reduce, Apache Tez, or Apache Spark. Pig Latin abstracts the programming from the Java Map Reduce idiom into a notation which makes Map Reduce programming high level, similar to that of SQL for RDBMS.

Pig Latin can be extended using User Defined Functions (UDFs) which the user can write in Java, Python, JavaScript, Ruby or Groovy and then call directly from the language.

## **H1-B CASE STUDY**

**PROJECT OBJECTIVE** --- The H1-B is an employment –based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply H1-B Visa, an US employer must offer a job and petition for H1-B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college/higher education (Masters, Ph.D.) and work in a full-time position.

### **Hardware Requirements-**

8 GB RAM

64Bit OS

### **Technology Requirements-**

- Apache Hadoop
- MapReduce
- Hive
- Pig
- SQOOP

### **Software Used–**

- VMware
- Ubuntu
- Eclipse
- MySQL

### **Assumptions-**

- VMware Workstation – Configurations are set correctly.

- Ubuntu is lying on the Virtual Box and it is powered on
- Hadoop Folder must be extracted and all the services of the Hadoop is running. Configuration to be made in the XML are set.
- Confirmation Box Below that Everything is Set Right.

```
hduser@ubuntu: ~
[1]:~$ jps
27172 Jps
3349 ResourceManager
2822 NameNode
3190 SecondaryNameNode
2970 DataNode
3466 NodeManager
[2]:~$
```

## **Datasets Required-**

- H1-B CASE Applications Data
- The Dataset has nearly 3 million records.

## **The Dataset given below—**

### **The columns in the dataset include:**

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

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

Certified Withdrawn: LCA was approved but later withdrawn by employer

Withdrawn: LCA was withdrawn by employer before approval

Denied: LCA was denied by DOL

**2. EMPLOYER\_NAME:** Name of employer submitting labour condition application.

**3. SOC\_NAME:** The Occupational name associated with the SOC\_CODE.

SOC\_CODE is the occupational code associated with the job being requested for temporary labour condition, as classified by the Standard Occupational Classification (SOC) System.

**4. JOB\_TITLE:** Title of the job

**5. FULL\_TIME\_POSITION:** Y = Full Time Position; N = Part Time Position

**6. PREVAILING\_WAGE:** Prevailing Wage for the job being requested for temporary labour condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer's minimum requirements for the position.

**7. YEAR:** Year in which the H1B visa petition was filed

**8. WORKSITE:** City and State information of the foreign worker's intended area of employment

**9. LONGITUDE:** longitude of the Worksite

**10. LATITUDE:** latitude of the Worksite

```

hive (h1b_project)> describe h1b_final;
OK
s_no          int
case_status   string
employer_name string
soc_name      string
job_title     string
full_time_position string
prevailing_wage bigint
year          string
worksite      string
longitude     double
latitude      double
Time taken: 0.239 seconds, Fetched: 11 row(s)
hive (h1b_project)>

```

## Outcome of this Project:

To generate reports and hence,

We will be performing analysis on the H1B visa applicants between the years 2011-2016. After analysing the data, we can derive the following facts.

- 1 a) Is the number of petitions with Data Engineer job title increasing over time?  
b) Find top 5 job titles who are having highest avg growth in applications. [ALL]
  
- 2 a) Which part of the US has the most Data Engineer jobs for each year?  
b) find top 5 locations in the US who have got certified visa for each year. [certified]
  
- 3) Which industry(SOC\_NAME) has the most number of Data Scientist positions?  
[certified]

- 4) Which top 5 employers file the most petitions each year? - Case Status - ALL
- 5) Find the most popular top 10 job positions for H1B visa applications for each year?  
 a) for all the applications  
 b) for only certified applications.
- 6) Find the percentage and the count of each case status on total applications for each year. Create a line graph depicting the pattern of All the cases over the period of time.
- 7) Create a bar graph to depict the number of applications for each year [All]
- 8) Find the average Prevailing Wage for each Job for each Year (take part time and full time separate). Arrange the output in descending order - [Certified and Certified Withdrawn.]
- 9) Which are the employers along with the number of petitions who have the success rate more than 70% in petitions. (total petitions filed 1000 OR more than 1000)?
- 10) Which are the job positions along with the number of petitions which have the success rate more than 70% in petitions (total petitions filed 1000 OR more than 1000)?
- 11) Export result for question no 10 to MySQL database.

```
hduser@ubuntu:~ at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:736)
at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:681)
at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:621)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:0 cannot recognize input near 'describe' 'h1b_final' '<EOF>'
hive (h1b_project)> describe h1b_final;
OK
s_no          int
case_status    string
employer_name  string
soc_name       string
job_title      string
full_time_position  string
prevailing_wage bigint
year           string
worksite       string
longitude      double
latitude       double
Time taken: 0.239 seconds, Fetched: 11 row(s)
hive (h1b_project)> select * from h1b_final limit 3;
OK
A 1      CERTIFIED-WITHDRAWN   UNIVERSITY OF MICHIGAN  BIOCHEMISTS AND BIOPHYSICISTS  POSTDOCTORAL RESEARCH FELLOW  N      36
067    2016    ANN ARBOR, MICHIGAN   -93.7430378  42.2808256
2      CERTIFIED-WITHDRAWN   GOODMAN NETWORKS, INC.  CHIEF EXECUTIVES      CHIEF OPERATING OFFICER Y      242674  2016  PL
ANO, TEXAS   -96.6988856  33.0198431
3      CERTIFIED-WITHDRAWN   PORTS AMERICA GROUP, INC.  CHIEF EXECUTIVES      CHIEF PROCESS OFFICER Y      193066  20
16     JERSEY CITY, NEW JERSEY -74.0776417  40.7281575
Time taken: 0.236 seconds, Fetched: 3 row(s)
hive (h1b_project)> 
```

**Let's create a table to load the h1b applicant's data as shown below:**

- CREATE TABLE h1b\_applications (s\_no int, case\_status string, employer\_name string, soc\_name string, job\_title string, full\_time\_position string, prevailing\_wage bigint, year string, worksite string, longitude double, latitude double) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITH SERDEPROPERTIES ( "separatorChar" = ",", "quoteChar" = "\"" ) STORED AS TEXTFILE;
- load data local inpath '/home/hduser/Downloads/H1Project/h1b.csv' overwrite into table h1b\_applications;
- CREATE TABLE h1b\_app2(s\_no int, case\_status string, employer\_name string, soc\_name string, job\_title string, full\_time\_position string, prevailing\_wage bigint, year string, worksite string, longitude double, latitude double) row format delimited fields terminated by '\t' STORED AS TEXTFILE;
- INSERT OVERWRITE TABLE h1b\_app2 SELECT regexp\_replace (s\_no, "\t", ""), regexp\_replace (case\_status, "\t", ""), regexp\_replace (employer\_name,

```
"\t", ""), regexp_replace (soc_name, "\t", ""), regexp_replace (job_title, "\t", ""),
regexp_replace (full_time_position, "\t", ""), prevailing_wage,
regexp_replace (year, "\t", ""), regexp_replace (worksit, "\t", ""), regexp_replace
(longitude, "\t", ""), regexp_replace (latitude, "\t", "")) FROM h1b_applications
where case_status != "NA";
```

- CREATE TABLE h1b\_final (s\_no int, case\_status string, employer\_name  
string, soc\_name string, job\_title string, full\_time\_position  
string, prevailing\_wage bigint, year string, worksite string, longitude  
double, latitute double)  
row format delimited  
fields terminated by '\t'  
STORED AS TEXTFILE;
  
- INSERT OVERWRITE TABLE h1b\_final SELECT s\_no,  
case when trim(case\_status) = "PENDING QUALITY AND COMPLIANCE  
REVIEW -  
UNASSIGNED" then "DENIED"  
when trim(case\_status) = "REJECTED" then "DENIED"  
when trim(case\_status) = "INVALIDATED" then "DENIED"  
else case\_status end,  
employer\_name, soc\_name, job\_title, full\_time\_position,  
case when prevailing\_wage is null then 100000  
else prevailing\_wage end, year, worksite, longitude, latitute FROM h1b\_app2;

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

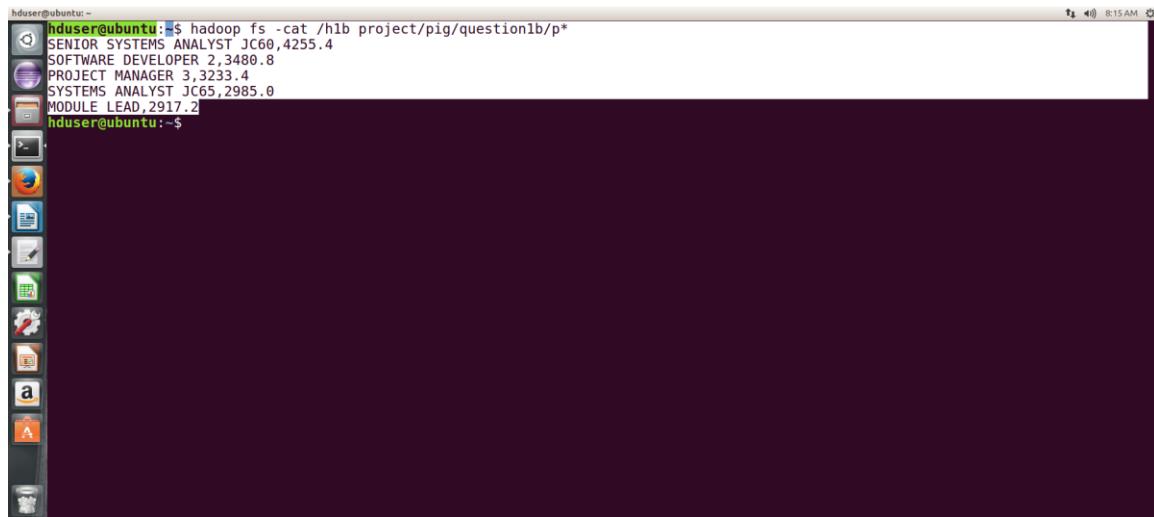
## Solution:

```
hduser@ubuntu:~  
STATISTICIANS,572  
COMPUTER AND INFORMATION RESEARCH SCIENTISTS,419  
OPERATIONS RESEARCH ANALYSTS,380  
Computer and Information Research Scientists,181  
COMPUTER OCCUPATIONS, ALL OTHER,160  
hduser@ubuntu:~$ hadoop fs -put /home/hduser/Downloads/H1Project/Q1(a).jar /MapReduce  
bash: syntax error near unexpected token `'  
hduser@ubuntu:~$ hadoop fs -put /home/hduser/Downloads/H1Project/Q1a.jar /MapReduce  
hduser@ubuntu:~$ hadoop jar /home/hduser/Downloads/H1Project/q1a.jar h1b_final.DataEngineerGrowth /MapReduce/h1b_final /MapReduce/DataEngineerGrowthOutput  
Not a valid JAR: /home/hduser/Downloads/H1Project/q1a.jar  
hduser@ubuntu:~$ hadoop jar /home/hduser/Downloads/H1Project/Q1a.jar h1b_final.DataEngineerGrowth /MapReduce/h1b_final /MapReduce/DataEngineerGrowthOutput  
hduser@ubuntu:~$ hadoop fs -cat /MapReduce/DataEngineerGrowthOutput/p*  
2011 1  
2012 100  
2013 50  
2014 33  
2015 25  
2016 20  
hduser@ubuntu:~$ hadoop fs -put /home/hduser/Downloads/H1Project/Q4.jar /MapReduce  
hduser@ubuntu:~$ hadoop jar /home/hduser/Downloads/H1Project/Q4.jar h1b_final.Top5Employees /MapReduce/h1b_final /MapReduce/Top5EmployeesOutput  
hduser@ubuntu:~$ hadoop fs -cat /MapReduce/Top5EmployeesOutput/p*  
TATA CONSULTANCY SERVICES LIMITED 2011,5416  
MICROSOFT CORPORATION 2011,4253  
DELOITTE CONSULTING LLP 2011,3621  
WIPRO LIMITED 2011,3028  
COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION 2011,2721  
INFOSYS LIMITED 2012,15818  
WIPRO LIMITED 2012,7182
```

---

**1 b) Find top 5 job titles who are having highest average growth in applications. [ALL] (PIG)**

## Solution:



A screenshot of a terminal window on an Ubuntu desktop. The terminal shows the command `hadoop fs -cat /hbl project/pig/questionlb/p*` being run, and its output, which lists several job titles and their counts:

```
hduser@ubuntu:~$ hadoop fs -cat /hbl project/pig/questionlb/p*
SENIOR SYSTEMS ANALYST JC60,4255.4
SOFTWARE DEVELOPER 2,3480.8
PROJECT MANAGER 3,3233.4
SYSTEMS ANALYST JC65,2985.0
MODULE LEAD,2917.2
hduser@ubuntu:~$
```

---

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

**Solution:**

```

hduser@ubuntu:~$ hadoop fs -put /home/hduser/Downloads/H1Project/USPart.jar /MapReduce
hduser@ubuntu:~$ hadoop jar /home/hduser/Downloads/H1Project/USPart.jar h1b final.USPartDataEngineer /MapReduce/h1b final /MapReduc
e/USPartDataEngineerOutput
hduser@ubuntu:~$ hadoop fs -cat /MapReduce/USPartDataEngineerOutput/p*
SEATTLE, WASHINGTON 2011,20
SAN FRANCISCO, CALIFORNIA 2011,4
SAN MATEO, CALIFORNIA 2011,3
WALTHAM, MASSACHUSETTS 2011,2
TALLAHASSEE, FLORIDA 2011,1
SEATTLE, WASHINGTON 2012,30
SAN FRANCISCO, CALIFORNIA 2012,10
PONTIAC, MICHIGAN 2012,3
SAN MATEO, CALIFORNIA 2012,2
WOODLAND HILLS, CALIFORNIA 2012,1
SEATTLE, WASHINGTON 2013,46
SAN FRANCISCO, CALIFORNIA 2013,17
MENLO PARK, CALIFORNIA 2013,12
NEW YORK, NEW YORK 2013,6
ATLANTA, GEORGIA 2013,5
SEATTLE, WASHINGTON 2014,45
SAN FRANCISCO, CALIFORNIA 2014,34
MENLO PARK, CALIFORNIA 2014,21
NEW YORK, NEW YORK 2014,18
MOUNTAIN VIEW, CALIFORNIA 2014,13
SEATTLE, WASHINGTON 2015,61
NEW YORK, NEW YORK 2015,41
MENLO PARK, CALIFORNIA 2015,23
MOUNTAIN VIEW, CALIFORNIA 2015,18
SAN MATEO, CALIFORNIA 2015,15
SEATTLE, WASHINGTON 2016,128
SAN FRANCISCO, CALIFORNIA 2016,90
NEW YORK, NEW YORK 2016,70
MENLO PARK, CALIFORNIA 2016,39
IRVINE, CALIFORNIA 2016,18

```

---

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

**Solution:**

Select worksite, count (case\_status) as counted, year from h1b\_final where year ='2011' and case\_status='CERTIFIED' group by worksite, year order by counted desc limit 5;

```

hive (h1b_project)> select worksite,count(case_status) as counted,year from h1b_final where year ='2011' and case_status='CERTIFIED'
' group by worksite,year order by counted desc limit 5;
Query ID = hduser_20171013114354_aade2bca0-5563-4ad1-b906-4538ab7ef10c
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0068, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0068/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0068
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 11:44:09,936 Stage-1 map = 0%, reduce = 0%
2017-10-13 11:44:36,153 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 9.96 sec
2017-10-13 11:44:41,969 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 14.25 sec
2017-10-13 11:44:45,358 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 16.14 sec
2017-10-13 11:44:47,559 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 16.81 sec
2017-10-13 11:45:06,838 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 20.89 sec
MapReduce Total cumulative CPU time: 20 seconds 890 msec
Ended Job = job_1507706430268_0068
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0069, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0069/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0069

```

```

2017-10-13 11:44:41,969 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 14.25 sec
2017-10-13 11:44:45,358 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 16.14 sec
2017-10-13 11:44:47,559 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 16.81 sec
2017-10-13 11:45:06,838 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 20.89 sec
MapReduce Total cumulative CPU time: 20 seconds 890 msec
Ended Job = job_1507706430268_0068
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0069, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0069/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0069
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 11:45:29,339 Stage-2 map = 0%, reduce = 0%
2017-10-13 11:45:39,236 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.82 sec
2017-10-13 11:46:01,948 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.38 sec
MapReduce Total cumulative CPU time: 3 seconds 380 msec
Ended Job = job_1507706430268_0069
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2  Cumulative CPU: 20.89 sec  HDFS Read: 449887245 HDFS Write: 378478 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1  Cumulative CPU: 3.38 sec  HDFS Read: 383504 HDFS Write: 150 SUCCESS
Total MapReduce CPU Time Spent: 24 seconds 270 msec
OK
NEW YORK, NEW YORK      23172    2011
HOUSTON, TEXAS          2011
CHICAGO, ILLINOIS       5188     2011
SAN JOSE, CALIFORNIA   4713     2011
SAN FRANCISCO, CALIFORNIA 4711    2011
Time taken: 128.557 seconds, Fetched: 5 row(s)
hive (h1b project)> 
```

`select worksite, count(case_status) as counted, year from h1b_final where year ='2012' and case_status='CERTIFIED' group by worksite, year order by counted desc limit 5;`

```

hduser@ubuntu:~ 
Time taken: 128.557 seconds, Fetched: 5 row(s)
hive (h1b_project)> select worksite,count(case_status) as counted,year from h1b_final where year ='2012' and case_status='CERTIFIED'
, group by worksite,year order by counted desc limit 5;
Query ID = hduser_20171013114932_203f6221-3a8c-487f-9a94-b13a3ef87d84
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job 1507706430268_0070, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0070/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0070
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 11:49:43,905 Stage-1 map = 0%, reduce = 0%
2017-10-13 11:50:05,344 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 9.37 sec
2017-10-13 11:50:06,616 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 9.85 sec
2017-10-13 11:50:09,137 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 11.42 sec
2017-10-13 11:50:44,133 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 15.2 sec
MapReduce Total cumulative CPU time: 15 seconds 200 msec
Ended Job = job 1507706430268_0070
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job 1507706430268_0071, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0071/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0071

```

```

hduser@ubuntu:~ 
2017-10-13 11:50:05,344 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 9.37 sec
2017-10-13 11:50:06,616 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 9.85 sec
2017-10-13 11:50:09,137 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 11.42 sec
2017-10-13 11:50:44,133 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 15.2 sec
MapReduce Total cumulative CPU time: 15 seconds 200 msec
Ended Job = job 1507706430268_0070
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job 1507706430268_0071, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0071/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0071
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 11:51:02,056 Stage-2 map = 0%, reduce = 0%
2017-10-13 11:51:10,654 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.78 sec
2017-10-13 11:51:21,393 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.22 sec
MapReduce Total cumulative CPU time: 3 seconds 220 msec
Ended Job = job 1507706430268_0071
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2   Cumulative CPU: 15.2 sec   HDFS Read: 449887245 HDFS Write: 384986 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1   Cumulative CPU: 3.22 sec   HDFS Read: 390012 HDFS Write: 146 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 420 msec
OK
NEW YORK, NEW YORK      23737    2012
HOUSTON, TEXAS         9963     2012
SAN FRANCISCO, CALIFORNIA 6116    2012
CHICAGO, ILLINOIS      5671     2012
ATLANTA, GEORGIA       5565     2012
Time taken: 109.819 seconds, Fetched: 5 row(s)
hive (h1b project)>
```

`select worksite, count(case_status) as counted, year from h1b_final where year ='2013' and case_status='CERTIFIED' group by worksite, year order by counted desc limit 5;`

```

hduser@ubuntu:~$ hive (hb project)> select worksite,count(case_status) as counted,year from hb_final where year ='2013' and case_status='CERTIFIED' group by worksite,year order by counted desc limit 5;
Query ID = hduser_20171013115212_0fbdbb52-56c6-438e-b8de-2254cdf689e0
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0072, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0072/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0072
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 11:52:21,618 Stage-1 map = 0%, reduce = 0%
2017-10-13 11:52:43,087 Stage-1 map = 17%, reduce = 0%, Cumulative CPU 8.96 sec
2017-10-13 11:52:47,148 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 10.78 sec
2017-10-13 11:52:48,281 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 11.39 sec
2017-10-13 11:52:49,372 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.34 sec
2017-10-13 11:53:08,828 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 15.89 sec
MapReduce Total cumulative CPU time: 15 seconds 890 msec
Ended Job = job_1507706430268_0072
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0073, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0073/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0073

```

```

hduser@ubuntu:~$ 2017-10-13 11:52:47,148 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 10.78 sec
2017-10-13 11:52:48,281 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 11.39 sec
2017-10-13 11:52:49,372 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.34 sec
2017-10-13 11:53:08,828 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 15.89 sec
MapReduce Total cumulative CPU time: 15 seconds 890 msec
Ended Job = job_1507706430268_0072
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0073, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0073/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0073
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 11:53:27,912 Stage-2 map = 0%, reduce = 0%
2017-10-13 11:53:35,309 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.44 sec
2017-10-13 11:53:45,152 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.23 sec
MapReduce Total cumulative CPU time: 3 seconds 230 msec
Ended Job = job_1507706430268_0073
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 15.89 sec HDFS Read: 449887245 HDFS Write: 368643 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.23 sec HDFS Read: 373669 HDFS Write: 150 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 120 msec
OK
NEW YORK, NEW YORK      23537    2013
HOUSTON, TEXAS        11136    2013
SAN FRANCISCO, CALIFORNIA    7281    2013
SAN JOSE, CALIFORNIA     6722    2013
ATLANTA, GEORGIA       6377    2013
Time taken: 93.917 seconds, Fetched: 5 row(s)
hive (hb project)>

```

select worksite, count(case\_status) as counted, year from hb\_final where year ='2014' and case\_status='CERTIFIED' group by worksite, year order by counted desc limit 5;

```
hduser@ubuntu: ~
Time taken: 93.917 seconds, Fetched: 5 row(s)
hive (h1b_project)> select worksite,count(case_status) as counted,year from h1b_final where year ='2014' and case_status='CERTIFIED'
, group by worksite,year order by counted desc limit 5;
Query ID = hduser_20171013115504_1abef3ec-51ba-4097-aad5-733b86283243
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0074, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0074/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0074
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 11:55:14,591 Stage-1 map = 0%, reduce = 0%
2017-10-13 11:55:36,772 Stage-1 map = 17%, reduce = 0%, Cumulative CPU 10.0 sec
2017-10-13 11:55:37,989 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 10.8 sec
2017-10-13 11:55:39,036 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.03 sec
2017-10-13 11:55:54,641 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 15.21 sec
MapReduce Total cumulative CPU time: 15 seconds 210 msec
Ended Job = job_1507706430268_0074
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0075, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0075/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0075
Ended Job = job_1507706430268_0075
```

```
hduser@ubuntu: ~
2017-10-13 11:55:36,772 Stage-1 map = 17%, reduce = 0%, Cumulative CPU 10.0 sec
2017-10-13 11:55:37,989 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 10.8 sec
2017-10-13 11:55:39,036 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.03 sec
2017-10-13 11:55:54,641 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 15.21 sec
MapReduce Total cumulative CPU time: 15 seconds 210 msec
Ended Job = job_1507706430268_0074
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0075, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0075/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0075
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 11:56:08,253 Stage-2 map = 0%, reduce = 0%
2017-10-13 11:56:16,847 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.58 sec
2017-10-13 11:56:25,547 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.06 sec
MapReduce Total cumulative CPU time: 3 seconds 60 msec
Ended Job = job_1507706430268_0075
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 15.21 sec HDFS Read: 449887245 HDFS Write: 375176 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.06 sec HDFS Read: 380202 HDFS Write: 150 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 270 msec
OK
NEW YORK, NEW YORK      27634    2014
HOUSTON, TEXAS       13360    2014
SAN FRANCISCO, CALIFORNIA   9798    2014
SAN JOSE, CALIFORNIA     8223    2014
ATLANTA, GEORGIA        8213    2014
Time taken: 81.721 seconds, Fetched: 5 row(s)
hive (h1b project)>
```

`select worksite, count(case_status) as counted, year from h1b_final where year ='2015' and case_status='CERTIFIED' group by worksite, year order by counted desc limit 5;`

```

hduser@ubuntu:~$ Time taken: 81.721 seconds, Fetched: 5 row(s)
hive (hb_project)> select worksite,count(case_status) as counted,year from hlb_final where year ='2015' and case_status='CERTIFIED'
' group by worksite,year order by counted desc limit 5;
Query ID = hduser_20171013115709_42a10130-9974-4470-9f20-456df45bd5e0
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0076, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0076/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0076
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 11:57:18,821 Stage-1 map = 0%,  reduce = 0%
2017-10-13 11:57:38,236 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 7.83 sec
2017-10-13 11:57:39,382 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 8.32 sec
2017-10-13 11:57:41,654 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 10.2 sec
2017-10-13 11:57:53,441 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 13.3 sec
MapReduce Total cumulative CPU time: 13 seconds 300 msec
Ended Job = job_1507706430268_0076
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0077, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0077/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0077

```

```

hduser@ubuntu:~$ 2017-10-13 11:57:38,236 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 7.83 sec
2017-10-13 11:57:39,382 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 8.32 sec
2017-10-13 11:57:41,654 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 10.2 sec
2017-10-13 11:57:53,441 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 13.3 sec
MapReduce Total cumulative CPU time: 13 seconds 300 msec
Ended Job = job_1507706430268_0076
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0077, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0077/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0077
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 11:58:07,520 Stage-2 map = 0%,  reduce = 0%
2017-10-13 11:58:15,048 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.35 sec
2017-10-13 11:58:23,961 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.9 sec
MapReduce Total cumulative CPU time: 2 seconds 900 msec
Ended Job = job_1507706430268_0077
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 13.3 sec HDFS Read: 449887245 HDFS Write: 382790 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.9 sec HDFS Read: 387816 HDFS Write: 152 SUCCESS
Total Mapreduce CPU Time Spent: 16 seconds 200 msec
OK
NEW YORK, NEW YORK      31266  2015
HOUSTON, TEXAS      15242  2015
SAN FRANCISCO, CALIFORNIA    12594  2015
ATLANTA, GEORGIA     10500  2015
SAN JOSE, CALIFORNIA   9589   2015
Time taken: 76.598 seconds, Fetched: 5 row(s)
hive (hb project)>
```

S

select worksite, count(case\_status) as counted, year from h1b\_final where year ='2016' and case\_status='CERTIFIED' group by worksite, year order by counted desc limit 5;

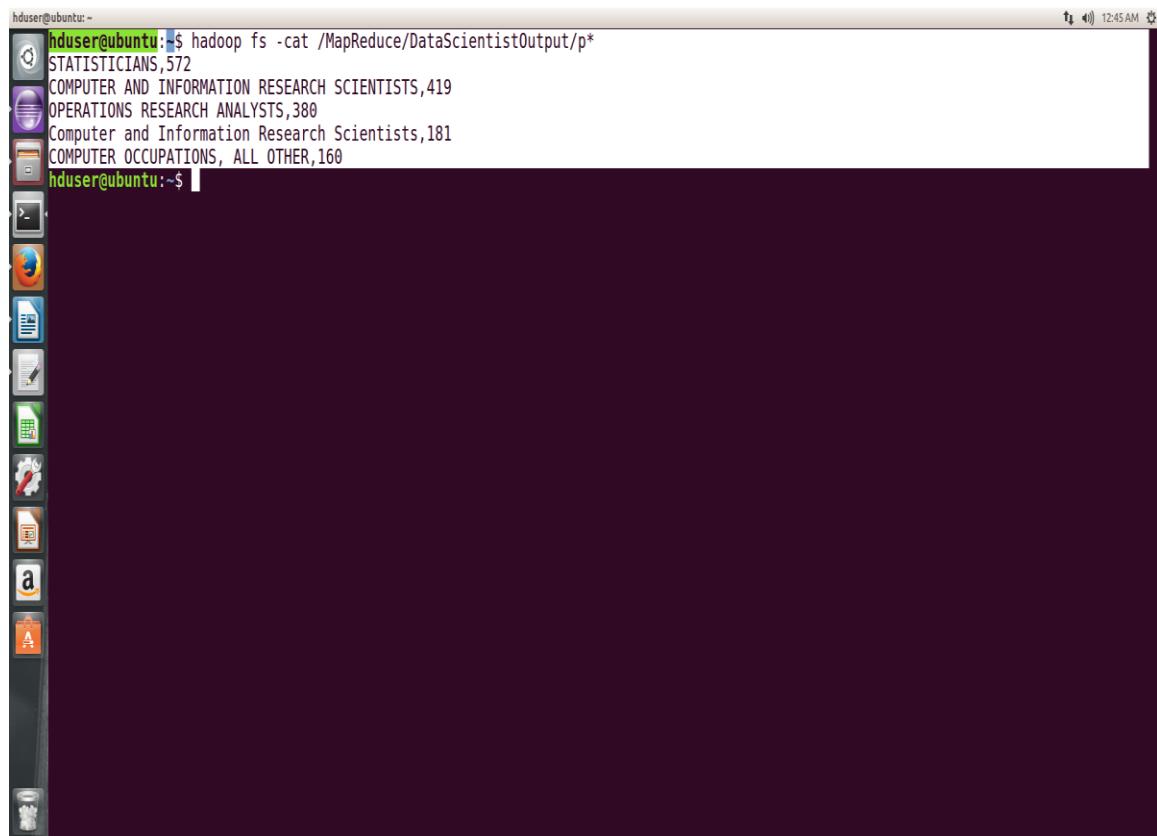
```
hduser@ubuntu:~$ 
hduser@ubuntu:~$ Time taken: 76.598 seconds. Fetched: 5 row(s)
hive (hb_project)> select worksite,count(case_status) as counted,year from h1b_final where year ='2016' and case_status='CERTIFIED'
' group by worksite,year order by counted desc limit 5;
Query ID = hduser_20171013115909_e1681950-46fc-4679-a7b7-dd9fd90dfeef
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0078, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0078/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0078
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 11:59:16,667 Stage-1 map = 0%, reduce = 0%
2017-10-13 11:59:36,171 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 7.9 sec
2017-10-13 11:59:37,210 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 9.23 sec
2017-10-13 11:59:49,211 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 12.52 sec
MapReduce Total cumulative CPU time: 12 seconds 520 msec
Ended Job = job_1507706430268_0078
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0079, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0079/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0079
hduser@ubuntu:~$
```

```
hduser@ubuntu:~$ 
2017-10-13 11:59:16,667 Stage-1 map = 0%, reduce = 0%
2017-10-13 11:59:17,171 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 7.9 sec
2017-10-13 11:59:37,210 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 9.23 sec
2017-10-13 11:59:49,211 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 12.52 sec
MapReduce Total cumulative CPU time: 12 seconds 520 msec
Ended Job = job_1507706430268_0078
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0079, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0079/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0079
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 12:00:03,476 Stage-2 map = 0%, reduce = 0%
2017-10-13 12:00:11,017 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.4 sec
2017-10-13 12:00:20,571 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.66 sec
MapReduce Total cumulative CPU time: 2 seconds 660 msec
Ended Job = job_1507706430268_0079
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 12.52 sec HDFS Read: 449887245 HDFS Write: 378330 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.66 sec HDFS Read: 383356 HDFS Write: 150 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 180 msec
OK
hduser@ubuntu:~$ 
hduser@ubuntu:~$ NEW YORK      34639    2016
hduser@ubuntu:~$ SAN FRANCISCO   13836    2016
hduser@ubuntu:~$ HOUSTON, TEXAS  13655    2016
hduser@ubuntu:~$ ATLANTA, GEORGIA 11678    2016
hduser@ubuntu:~$ CHICAGO, ILLINOIS 11064    2016
hduser@ubuntu:~$ Time taken: 72.697 seconds, Fetched: 5 row(s)
hive (hb_project)> 
```

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

**[certified] (MapReduce)**

**Solution:**



```
hduser@ubuntu:~$ hadoop fs -cat /MapReduce/DataScientistOutput/p*
STATISTICIANS,572
COMPUTER AND INFORMATION RESEARCH SCIENTISTS,419
OPERATIONS RESEARCH ANALYSTS,380
Computer and Information Research Scientists,181
COMPUTER OCCUPATIONS, ALL OTHER,160
hduser@ubuntu:~$
```

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

### Solution:

```
hduser@ubuntu:~$ hadoop fs -put /home/hduser/Downloads/H1Project/Q4.jar /MapReduce  
hduser@ubuntu:~$ hadoop jar /home/hduser/Downloads/H1Project/Q4.jar h1b final.Top5Employees /MapReduce/h1b final /MapReduce/Top5Emp  
loyeesOutput  
hduser@ubuntu:~$ hadoop fs -cat /MapReduce/Top5EmployeesOutput/p*  
TATA CONSULTANCY SERVICES LIMITED 2011,5416  
WIPRO SOFTCORPORATION 2011,4253  
DELOITTE CONSULTING LLP 2011,3621  
WIPRO LIMITED 2011,3028  
COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION 2011,2721  
INFOSYS LIMITED 2012,15818  
WIPRO LIMITED 2012,7182  
TATA CONSULTANCY SERVICES LIMITED 2012,6735  
DELOITTE CONSULTING LLP 2012,4727  
IBM INDIA PRIVATE LIMITED 2012,4074  
INFOSYS LIMITED 2013,3223  
TATA CONSULTANCY SERVICES LIMITED 2013,8790  
WIPRO LIMITED 2013,6734  
DELOITTE CONSULTING LLP 2013,6124  
ACCENTURE LLP 2013,4994  
INFOSYS LIMITED 2014,23759  
TATA CONSULTANCY SERVICES LIMITED 2014,14098  
WIPRO LIMITED 2014,8365  
DELOITTE CONSULTING LLP 2014,7017  
ACCENTURE LLP 2014,5994  
INFOSYS LIMITED 2015,33245  
TATA CONSULTANCY SERVICES LIMITED 2015,16553  
WIPRO LIMITED 2015,12201  
IBM INDIA PRIVATE LIMITED 2015,10693  
ACCENTURE LLP 2015,9605  
INFOSYS LIMITED 2016,25352  
CAPGEMINI AMERICA INC 2016,16725
```

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

### (a) for all the applications (HIVE)

**Solution:**

```
select job_title,year,count(case_status ) as count_case_status from h1b_final where year = 2011 group by job_title,year order by count_case_status desc limit 10;
```

```
hduser@ubuntu:~$ 
2013 442114
2014 218427
2015 218727
2016 647803
Time taken: 111.086 seconds, Fetched: 6 row(s)
hive (h1b project)> select job_title,year,count(case_status ) as count_case_status from h1b_final where year = 2011 group by job_title,year;
FAILED: SemanticException [Error 10004]: Line 1:129 Invalid table alias or column reference 'temp': (possible column names are: job_title, year, count_case_status)
hive (h1b project)> select job_title,year,count(case_status ) as count_case_status from h1b_final where year = 2011 group by job_title,year;
Query ID = hduser_20171013071433_2ae64441-1ce2-42f7-9633-c5747576fb26
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0052, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0052/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0052
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 07:15:08,278 Stage-1 map = 17%, reduce = 0%, Cumulative CPU 6.45 sec
2017-10-13 07:15:09,477 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 7.98 sec
2017-10-13 07:15:10,522 Stage-1 map = 70%, reduce = 0%, Cumulative CPU 9.06 sec
2017-10-13 07:15:11,562 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 10.81 sec
2017-10-13 07:15:12,602 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 11.97 sec
2017-10-13 07:15:26,417 Stage-1 map = 100%, reduce = 95%, Cumulative CPU 14.89 sec
2017-10-13 07:15:27,452 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 15.48 sec
MapReduce Total cumulative CPU time: 15 seconds 480 msec
Ended Job = job_1507706430268_0052
```

```
hduser@ubuntu:~$ 
Ended Job = job_1507706430268_0052
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0053, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0053/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0053
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 07:15:40,472 Stage-2 map = 0%, reduce = 0%
2017-10-13 07:15:47,801 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.8 sec
2017-10-13 07:15:56,345 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.44 sec
MapReduce Total cumulative CPU time: 3 seconds 440 msec
Ended Job = job_1507706430268_0053
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 15.48 sec HDFS Read: 449886405 HDFS Write: 3490343 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.44 sec HDFS Read: 3495626 HDFS Write: 299 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 920 msec
OK
PROGRAMMER ANALYST 2011 31799
SOFTWARE ENGINEER 2011 12763
COMPUTER PROGRAMMER 2011 8998
SYSTEMS ANALYST 2011 8644
BUSINESS ANALYST 2011 3891
COMPUTER SYSTEMS ANALYST 2011 3698
ASSISTANT PROFESSOR 2011 3467
PHYSICAL THERAPIST 2011 3377
SENIOR SOFTWARE ENGINEER 2011 2935
SENIOR CONSULTANT 2011 2798
Time taken: 84.275 seconds, Fetched: 10 row(s)
hive (h1b project)>
```

```
select job_title, year, count (case_status) as count_case_status from h1b_final where year = 2012 group by job_title, year order by count_case_status desc limit 10;
```

```

hduser@ubuntu:~$ Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.44 sec HDFS Read: 3495626 HDFS Write: 299 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 920 msec
OK
PROGRAMMER ANALYST    2011      31799
SOFTWARE ENGINEER      2011      12763
COMPUTER PROGRAMMER    2011      8998
SYSTEMS ANALYST 2011   8644
BUSINESS ANALYST       2011      3891
COMPUTER SYSTEMS ANALYST 2011   3698
ASSISTANT PROFESSOR    2011      3467
PHYSICAL THERAPIST     2011      3377
SENIOR SOFTWARE ENGINEER 2011   2935
SENIOR CONSULTANT      2011      2798
Time taken: 84.275 seconds, Fetched: 10 row(s)
hive (hb project)> select job_title,year,count(case_status ) as count_case_status from h1b_final where year = 2012 group by job_title,year order by count_case_status desc limit 10;
Query ID = hduser_20171013072217_1df7ddce-7a36-46d4-bfb1-fef770f59448
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0054, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0054/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0054
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 07:22:24,962 Stage-1 map = 0%,  reduce = 0%
2017-10-13 07:22:46,293 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 9.08 sec
2017-10-13 07:22:48,497 Stage-1 map = 37%,  reduce = 0%, Cumulative CPU 10.94 sec
2017-10-13 07:22:50,627 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 12.55 sec

```

```

hduser@ubuntu:~$ Ended Job = job_1507706430268_0054
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0055, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0055/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0055
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 07:23:28,862 Stage-2 map = 0%,  reduce = 0%
2017-10-13 07:23:36,653 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.41 sec
2017-10-13 07:23:44,020 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.84 sec
MapReduce Total cumulative CPU time: 3 seconds 840 msec
Ended Job = job_1507706430268_0055
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 17.56 sec HDFS Read: 449886403 HDFS Write: 3727623 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.84 sec HDFS Read: 3732899 HDFS Write: 293 SUCCESS
Total MapReduce CPU Time Spent: 21 seconds 400 msec
OK
PROGRAMMER ANALYST    2012      33066
SOFTWARE ENGINEER      2012      14437
COMPUTER PROGRAMMER    2012      9629
SYSTEMS ANALYST 2012   9296
BUSINESS ANALYST       2012      4752
COMPUTER SYSTEMS ANALYST 2012   4706
SOFTWARE DEVELOPER     2012      3895
PHYSICAL THERAPIST     2012      3871
ASSISTANT PROFESSOR    2012      3801
SENIOR CONSULTANT      2012      3737
Time taken: 88.56 seconds, Fetched: 10 row(s)
hive (hb project)> 
```

`select job_title,year,count(case_status ) as count_case_status from h1b_final where year = 2013 group by job_title,year order by count_case_status desc limit 10;`

```

hduser@ubuntu:~$ 
hduser@ubuntu:~$ SOFTWARE ENGINEER    2012    14437
hduser@ubuntu:~$ COMPUTER PROGRAMMER   2012    9629
hduser@ubuntu:~$ SYSTEMS ANALYST 2012    9296
hduser@ubuntu:~$ BUSINESS ANALYST   2012    4752
hduser@ubuntu:~$ COMPUTER SYSTEMS ANALYST 2012    4706
hduser@ubuntu:~$ SOFTWARE DEVELOPER   2012    3895
hduser@ubuntu:~$ PHYSICAL THERAPIST   2012    3871
hduser@ubuntu:~$ ASSISTANT PROFESSOR   2012    3801
hduser@ubuntu:~$ SENIOR CONSULTANT   2012    3737
Time taken: 88.56 seconds. Fetched: 10 row(s)
hive (hb_project)> select job_title,year,count(case_status ) as count_case_status from hlb_final where year = 2013 group by job_title,year order by count_case_status desc limit 10;
Query ID = hduser_20171013081118_e49533d0-e13b-45aa-9b91-799e64328df4
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0056, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0056/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0056
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 08:11:29,304 Stage-1 map = 0%, reduce = 0%
2017-10-13 08:11:50,625 Stage-1 map = 17%, reduce = 0%, Cumulative CPU 8.95 sec
2017-10-13 08:11:57,615 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 11.48 sec
2017-10-13 08:12:06,368 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 16.36 sec
2017-10-13 08:12:45,992 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 18.65 sec
2017-10-13 08:12:47,090 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 20.98 sec
MapReduce Total cumulative CPU time: 20 seconds 980 msec
Ended Job = job_1507706430268_0056
Launching Job 2 out of 2

```

```

hduser@ubuntu:~$ 
hduser@ubuntu:~$ Ended Job = job_1507706430268_0056
hduser@ubuntu:~$ Launching Job 2 out of 2
hduser@ubuntu:~$ Number of reduce tasks determined at compile time: 1
hduser@ubuntu:~$ In order to change the average load for a reducer (in bytes):
hduser@ubuntu:~$   set hive.exec.reducers.bytes.per.reducer=<number>
hduser@ubuntu:~$ In order to limit the maximum number of reducers:
hduser@ubuntu:~$   set hive.exec.reducers.max=<number>
hduser@ubuntu:~$ In order to set a constant number of reducers:
hduser@ubuntu:~$   set mapreduce.job.reduces=<number>
hduser@ubuntu:~$ Starting Job = job_1507706430268_0057, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0057/
hduser@ubuntu:~$ Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0057
hduser@ubuntu:~$ Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
hduser@ubuntu:~$ 2017-10-13 08:13:05,597 Stage-2 map = 0%, reduce = 0%
hduser@ubuntu:~$ 2017-10-13 08:13:16,538 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.28 sec
hduser@ubuntu:~$ 2017-10-13 08:13:24,433 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.72 sec
hduser@ubuntu:~$ MapReduce Total cumulative CPU time: 3 seconds 720 msec
hduser@ubuntu:~$ Ended Job = job_1507706430268_0057
hduser@ubuntu:~$ MapReduce Jobs Launched:
hduser@ubuntu:~$ Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 20.98 sec HDFS Read: 449886405 HDFS Write: 3748326 SUCCESS
hduser@ubuntu:~$ Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.72 sec HDFS Read: 3753609 HDFS Write: 300 SUCCESS
hduser@ubuntu:~$ Total MapReduce CPU Time Spent: 24 seconds 700 msec
hduser@ubuntu:~$ OK
hduser@ubuntu:~$ PROGRAMMER ANALYST    2013    33880
hduser@ubuntu:~$ SOFTWARE ENGINEER    2013    15680
hduser@ubuntu:~$ COMPUTER PROGRAMMER   2013    11271
hduser@ubuntu:~$ SYSTEMS ANALYST 2013    8714
hduser@ubuntu:~$ TECHNOLOGY LEAD - US   2013    7853
hduser@ubuntu:~$ TECHNOLOGY ANALYST - US 2013    7683
hduser@ubuntu:~$ BUSINESS ANALYST   2013    5716
hduser@ubuntu:~$ COMPUTER SYSTEMS ANALYST 2013    5043
hduser@ubuntu:~$ SOFTWARE DEVELOPER   2013    5026
hduser@ubuntu:~$ SENIOR CONSULTANT   2013    4326
Time taken: 126.597 seconds. Fetched: 10 row(s)
hive (hb_project)> 
```

`select job_title,year,count(case_status ) as count_case_status from hlb_final where year = 2014 group by job_title,year order by count_case_status desc limit 10;`

```

hduser@ubuntu:~$ SENIOR CONSULTANT      2013      4326
Time taken: 126.597 seconds, Fetched: 10 row(s)
hive (hb project)> select job_title,year,count(case_status ) as count_case_status from h1b_final where year = 2014 group by job title,year order by count_case_status desc limit 10;
Query ID = hduser 20171013083036 d0d6bd5e-edeb-4c07-9e18-119ef029d5aa
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0058, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0058/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0058
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 08:30:43,844 Stage-1 map = 0%, reduce = 0%
2017-10-13 08:31:01,965 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 7.5 sec
2017-10-13 08:31:03,029 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 7.77 sec
2017-10-13 08:31:05,131 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 9.63 sec
2017-10-13 08:31:08,385 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 11.18 sec
2017-10-13 08:31:12,010 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.1 sec
2017-10-13 08:31:20,486 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 16.59 sec
MapReduce Total cumulative CPU time: 16 seconds 590 msec
Ended Job = job_1507706430268_0058
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>

```

```

hduser@ubuntu:~$ Ended Job = job_1507706430268_0058
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0059, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0059/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0059
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 08:31:34,224 Stage-2 map = 0%, reduce = 0%
2017-10-13 08:31:41,595 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.81 sec
2017-10-13 08:31:48,986 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.26 sec
MapReduce Total cumulative CPU time: 3 seconds 260 msec
Ended Job = job_1507706430268_0059
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 16.59 sec HDFS Read: 449886405 HDFS Write: 4190592 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.26 sec HDFS Read: 4195875 HDFS Write: 301 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 850 msec
OK
PROGRAMMER ANALYST      2014      43114
SOFTWARE ENGINEER        2014      20500
COMPUTER PROGRAMMER      2014      14950
SYSTEMS ANALYST 2014    10194
SOFTWARE DEVELOPER       2014      7337
BUSINESS ANALYST         2014      7382
COMPUTER SYSTEMS ANALYST 2014      6821
TECHNOLOGY LEAD - US     2014      5057
TECHNOLOGY ANALYST - US  2014      4913
SENIOR CONSULTANT        2014      4898
Time taken: 74.03 seconds, Fetched: 10 row(s)
hive (hb project)>
```

`select job_title,year,count(case_status ) as count_case_status from h1b_final where year = 2015 group by job_title,year order by count_case_status desc limit 10;`

```

hduser@ubuntu:~$ 
Time taken: 74.03 seconds, Fetched: 10 row(s)
hive (h1b project)> select job_title,year,count(case_status ) as count_case_status from h1b_final where year = 2015 group by job_title,year order by count_case_status desc limit 10;
Query ID = hduser 20171013084657 8fd86025-e036-43cb-bb9f-4319334f8832
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0060, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0060
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0060
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 08:47:13,355 Stage-1 map = 0%, reduce = 0%
2017-10-13 08:47:38,218 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 8.06 sec
2017-10-13 08:47:41,586 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 9.3 sec
2017-10-13 08:47:46,975 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 11.46 sec
2017-10-13 08:47:58,695 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 13.59 sec
2017-10-13 08:48:04,256 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 14.52 sec
2017-10-13 08:48:07,999 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 15.93 sec
2017-10-13 08:48:11,166 Stage-1 map = 100%, reduce = 76%, Cumulative CPU 18.78 sec
2017-10-13 08:48:13,560 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 20.37 sec
MapReduce Total cumulative CPU time: 20 seconds 370 msec
Ended Job = job_1507706430268_0060
Launching Job 2 out of 2
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:

```

```

hduser@ubuntu:~$ 
Ended Job = job_1507706430268_0060
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0061, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0061
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0061
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 08:48:28,619 Stage-2 map = 0%, reduce = 0%
2017-10-13 08:48:37,310 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.66 sec
2017-10-13 08:48:44,786 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 4.34 sec
MapReduce Total cumulative CPU time: 4 seconds 340 msec
Ended Job = job_1507706430268_0061
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 20.48 sec HDFS Read: 449886405 HDFS Write: 4592226 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 4.34 sec HDFS Read: 4597509 HDFS Write: 309 SUCCESS
Total MapReduce CPU Time Spent: 24 seconds 820 msec
OK
PROGRAMMER ANALYST      2015      53436
SOFTWARE ENGINEER        2015      27259
COMPUTER PROGRAMMER      2015      14054
SYSTEMS ANALYST 2015    12803
SOFTWARE DEVELOPER        2015      10441
BUSINESS ANALYST        2015      8853
TECHNOLOGY LEAD - US    2015      8242
COMPUTER SYSTEMS ANALYST 2015      7918
TECHNOLOGY ANALYST - US 2015    7014
SENIOR SOFTWARE ENGINEER 2015      6013
Time taken: 108.726 seconds, Fetched: 10 row(s)
hive (h1b project)> 

```

`select job_title,year,count(case_status ) as count_case_status from h1b_final where year = 2016 group by job_title,year order by count_case_status desc limit 10;`

```

hduser@ubuntu: ~
BUSINESS ANALYST    2015    8853
TECHNOLOGY LEAD - US 2015    8242
COMPUTER SYSTEMS ANALYST 2015    7918
TECHNOLOGY ANALYST - US 2015  7014
SENIOR SOFTWARE ENGINEER 2015    6013
Time taken: 108.726 seconds, Fetched: 10 row(s)
hive (hib_project)> select job_title,year,count(case_status ) as count_case_status from hib_final where year = 2016 group by job_title,year order by count case status desc limit 10;
Query ID = hduser_20171013085130_8ef9ce49-45b8-4c5a-a89a-361dce38c127
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0062, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0062/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0062
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 08:51:37,749 Stage-1 map = 0%, reduce = 0%
2017-10-13 08:51:55,760 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 8.45 sec
2017-10-13 08:51:56,806 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 8.69 sec
2017-10-13 08:51:58,898 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 10.95 sec
2017-10-13 08:52:02,081 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.07 sec
2017-10-13 08:52:11,603 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 16.53 sec
MapReduce Total cumulative CPU time: 16 seconds 530 msec
Ended Job = job_1507706430268_0062
Launching Job 2 out of 2
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:

```

```

hduser@ubuntu: ~
Ended Job = job_1507706430268_0062
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0063, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0063/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0063
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 08:52:24,648 Stage-2 map = 0%, reduce = 0%
2017-10-13 08:52:33,050 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.47 sec
2017-10-13 08:52:40,444 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 4.0 sec
MapReduce Total cumulative CPU time: 4 seconds 0 msec
Ended Job = job_1507706430268_0063
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 16.53 sec HDFS Read: 449886405 HDFS Write: 4754418 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 4.0 sec HDFS Read: 4759701 HDFS Write: 295 SUCCESS
Total MapReduce CPU Time Spent: 20 seconds 530 msec
OK
PROGRAMMER ANALYST    2016    53743
SOFTWARE ENGINEER      2016    30668
SOFTWARE DEVELOPER     2016    14041
SYSTEMS ANALYST        2016    12314
COMPUTER PROGRAMMER    2016    11668
BUSINESS ANALYST       2016    9167
COMPUTER SYSTEMS ANALYST 2016    6900
SENIOR SOFTWARE ENGINEER 2016    6439
DEVELOPER              2016    6084
TECHNOLOGY LEAD - US   2016    5410
Time taken: 71.54 seconds, Fetched: 10 row(s)
hive (hib_project)> .

```

## 5 (b) for only certified applications. (HIVE)

**Solution:**

```
select job_title, year, count(case_status) as count_case_status from h1b_final where
year = 2011 and case_status='CERTIFIED' group by job_title, year order by
count_case_status limit 10;
```

```
hduser@ubuntu:~$ Time taken: 72.607 seconds, Fetched: 5 row(s)
hive (h1b_project)> select job_title,year,count(case_status ) as count_case_status from h1b_final where year = 2011 and case_status='CERTIFIED' group by job_title,year order by count_case_status limit 10;
Query ID = hduser_20171013121950_dc9ec35f-e17b-423d-9158-abffedac9545
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0080, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0080/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0080
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 12:19:57.612 Stage-1 map = 0%, reduce = 0%
2017-10-13 12:20:19.214 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 10.21 sec
2017-10-13 12:20:21.913 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 12.23 sec
2017-10-13 12:20:22.948 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.75 sec
2017-10-13 12:20:38.131 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 16.43 sec
MapReduce Total cumulative CPU time: 16 seconds 430 msec
Ended Job = job_1507706430268_0080
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0081, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0081/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0081
```

```
hduser@ubuntu:~$ Ended Job = job_1507706430268_0080
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0081, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0081/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0081
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 12:20:53.412 Stage-2 map = 0%, reduce = 0%
2017-10-13 12:21:01.889 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.61 sec
2017-10-13 12:21:11.649 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 4.48 sec
MapReduce Total cumulative CPU time: 4 seconds 480 msec
Ended Job = job_1507706430268_0081
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 16.43 sec HDFS Read: 449886897 HDFS Write: 3310378 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 4.48 sec HDFS Read: 3315661 HDFS Write: 336 SUCCESS
Total MapReduce CPU Time Spent: 20 seconds 910 msec
OK
[HIQX] COMMERCIAL SPECIALIST (SALES ENGINEER) 2011 1
YOUTH PROGRAM CLINICIAN 2011 1
YOUTH GROUP MINISTRY DIRECTOR 2011 1
YOUTH DIRECTOR 2011 1
YOUTH DEVELOPMENT SPECIALIST 2011 1
YOUTH DEVELOPMENT MANAGER 2011 1
YOUTH COORDINATOR 2011 1
YOUTH ASSOCIATE 2011 1
YOA CHAMBER PROGRAM COORDINATOR 2011 1
YIELD SENIOR PRODUCT ENGINEER 2011 1
Time taken: 82.796 seconds, Fetched: 10 row(s)
hive (h1b_project)> 
```

```
select job_title, year, count(case_status) as count_case_status from h1b_final where
year = 2012 and case_status='CERTIFIED' group by job_title, year order by count_case_status
limit 10;
```

```
hduser@ubuntu:~  
Time taken: 82.796 seconds, Fetched: 10 row(s)  
hive (hbl_project)> select job_title,year,count(case status ) as count_case status from hbl final where year = 2012 and case status  
='CERTIFIED' group by job title,year order by count_case status limit 10;  
Query ID = hduser_20171013122403_67515268-le7d-486b-b066-b3b989e4807d  
Total jobs = 2  
Launching Job 1 out of 2  
Number of reduce tasks not specified. Estimated from input data size: 2  
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 mapreduce.job.reduces=<number>  
Starting Job = job 1507706430268_0082, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0082/  
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0082  
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2  
2017-10-13 12:24:15,101 Stage-1 map = 0%, reduce = 0%  
2017-10-13 12:24:35,363 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 10.16 sec  
2017-10-13 12:24:40,436 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 12.17 sec  
2017-10-13 12:24:41,511 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 13.03 sec  
2017-10-13 12:25:04,947 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 17.12 sec  
MapReduce Total cumulative CPU time: 17 seconds 120 msec  
Ended Job = job 1507706430268_0082  
Launching Job 2 out of 2  
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 mapreduce.job.reduces=<number>  
Starting Job = job 1507706430268_0083, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0083/  
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0083  
[
```

```
hduser@ubuntu:~  
Ended Job = job 1507706430268_0082  
Launching Job 2 out of 2  
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 mapreduce.job.reduces=<number>  
Starting Job = job 1507706430268_0083, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0083/  
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0083  
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1  
2017-10-13 12:25:21,284 Stage-2 map = 0%, reduce = 0%  
2017-10-13 12:25:28,612 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.92 sec  
2017-10-13 12:25:37,281 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.76 sec  
MapReduce Total cumulative CPU time: 3 seconds 768 msec  
Ended Job = job 1507706430268_0083  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 17.12 sec HDFS Read: 449886897 HDFS Write: 3383342 SUCCESS  
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.76 sec HDFS Read: 3388625 HDFS Write: 315 SUCCESS  
Total MapReduce CPU Time Spent: 20 seconds 880 msec  
OK  
LEAD TEST ANALYST      2012      1  
ZOOLOGISTS AND WILDLIFE BIOLOGISTS I      2012      1  
ZOOLOGIST - REPRODUCTIVE PHYSIOLOGY      2012      1  
ZOOKEEPER      2012      1  
ZONING MANAGER      2012      1  
YOUTH PASTOR      2012      1  
YOUTH MUSIC TEACHER      2012      1  
YOUTH LEADERSHIP DIRECTOR      2012      1  
YOUTH COMPLEX SOCCER COACH      2012      1  
YOGURT MANUFACTURER OPERATIONS MANAGER      2012      1  
Time taken: 94.869 seconds, Fetched: 10 row(s)  
hive (hbl_project)> [
```

select job\_title, year, count(case\_status) as count\_case\_status from h1b\_final where year = 2013 and case\_status='CERTIFIED' group by job\_title, year order by count\_case\_status limit 10;

```
hduser@ubuntu:~$ YOGURT MANUFACTURER OPERATIONS MANAGER 2012 1
Time taken: 94.869 seconds, Fetched: 10 row(s)
hive (hib_project)> select job_title,year,count(case status ) as count_case status from h1b final where year = 2013 and case status
= 'CERTIFIED' group by job title,year order by count case status limit 10;
Query ID = hduser_20171013122623_5d94fc77-dd6d-4464-802b-73d0810cfde
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0084, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0084/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0084
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 12:26:31.083 Stage-1 map = 0%, reduce = 0%
2017-10-13 12:26:51.777 Stage-1 map = 17%, reduce = 0%, Cumulative CPU 10.87 sec
2017-10-13 12:26:53.866 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 11.73 sec
2017-10-13 12:26:54.944 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 13.56 sec
2017-10-13 12:27:05.530 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 15.6 sec
2017-10-13 12:27:06.577 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 17.42 sec
MapReduce Total cumulative CPU time: 17 seconds 420 msec
Ended Job = job_1507706430268_0084
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
```

```
hduser@ubuntu:~$ Ended Job = job_1507706430268_0084
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0085, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0085/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0085
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 12:27:19.973 Stage-2 map = 0%, reduce = 0%
2017-10-13 12:27:27.378 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.89 sec
2017-10-13 12:27:34.961 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.45 sec
MapReduce Total cumulative CPU time: 3 seconds 450 msec
Ended Job = job_1507706430268_0085
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 17.42 sec HDFS Read: 449886897 HDFS Write: 3389382 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.45 sec HDFS Read: 3394665 HDFS Write: 352 SUCCESS
Total MapReduce CPU Time Spent: 20 seconds 870 msec
OK
TEST ANALYST - US 2013 1
ZIMBABWE PARTNERSHIP COORDINATOR 2013 1
YOUTH SOCIAL JUSTICE AND PUBLIC POLICY RESEARCH AN 2013 1
YOUTH PROGRAM DIRECTOR 2013 1
YOUTH PROGRAM COORDINATOR 2013 1
YOUTH MENTAL HEALTH INSTRUCTIONAL COORDINATOR 2013 1
YOUTH COUNSELOR II 2013 1
YOGESHWAR SALES INC 2013 1
YIELD ENGINEER / RESEARCHER 2013 1
YIELD ANALYST 2013 1
Time taken: 74.092 seconds, Fetched: 10 row(s)
hive (hib_project)>
```

```
select job_title, year, count(case_status) as count_case_status from h1b_final where year = 2014 and case_status='CERTIFIED' group by job_title, year order by count_case_status limit 10;
```

```
hduser@ubuntu:~$ 
Time taken: 74.092 seconds, Fetched: 10 row(s)
hive (hb project)> select job_title,year,count(case status ) as count case_status from h1b_final where year = 2014 and case_status='CERTIFIED' group by job_title, year order by count_case_status limit 10;
Query ID = hduser_20171013122824_d64d6bb7-3895-45b5-8e26-74dc08928182
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0086, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0086/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0086
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 12:28:31,948 Stage-1 map = 0%,  reduce = 0%
2017-10-13 12:28:50,048 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 8.39 sec
2017-10-13 12:28:53,209 Stage-1 map = 83%,  reduce = 0%, Cumulative CPU 10.94 sec
2017-10-13 12:28:54,255 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 10.94 sec
2017-10-13 12:29:05,069 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 15.46 sec
MapReduce Total cumulative CPU time: 15 seconds 460 msec
Ended Job = job_1507706430268_0086
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0087, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0087/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0087
[...]
```

```
hduser@ubuntu:~$ 
Ended Job = job_1507706430268_0086
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0087, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0087/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0087
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 12:29:19,779 Stage-2 map = 0%,  reduce = 0%
2017-10-13 12:29:28,507 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.65 sec
2017-10-13 12:29:36,287 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.27 sec
MapReduce Total cumulative CPU time: 4 seconds 278 msec
Ended Job = job_1507706430268_0087
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 15.46 sec HDFS Read: 449886897 HDFS Write: 3818622 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 4.27 sec HDFS Read: 3823905 HDFS Write: 301 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 730 msec
OK
(ASSOCIATE) BRAND MANAGER      2014      1
SOFTWARE TEST ENGINEER        2014      1
SENIOR PROJECT LEADER         2014      1
QUALITY ASSURANCE ANALYST    2014      1
QUALITY ASSURANCE ANALYST    2014      1
ZOOLOGIST                      2014      1
ZOOKEEPER                      2014      1
YUZHNO TRANSLATION TEAM MANAGER 2014      1
YUMA ENERGY, INC               2014      1
YSC ONLINE CLIENT MANAGER     2014      1
Time taken: 72.559 seconds, Fetched: 10 row(s)
hive (hb project)> [...]
```

```
select job_title, year, count(case_status) as count_case_status from h1b_final where
year = 2015 and case_status='CERTIFIED' group by job_title, year order by
count_case_status limit 10;
```

```
hduser@ubuntu:~$ YSC ONLINE CLIENT MANAGER 2014 1
Time taken: 72.559 seconds, Fetched: 10 row(s)
hive (h1b_project)> select job_title,year,count(case_status) as count_case_status from h1b_final where year = 2015 and case_status = 'CERTIFIED' group by job_title,year order by count_case_status limit 10;
Query ID = hduser_20171013123013_2ed9adfe-4488-451b-bc08-e6a696e097ef
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0088, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0088/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0088
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 12:30:21,943 Stage-1 map = 0%, reduce = 0%
2017-10-13 12:30:38,881 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 8.35 sec
2017-10-13 12:30:39,936 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 8.61 sec
2017-10-13 12:30:42,027 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 10.75 sec
2017-10-13 12:30:45,273 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 11.61 sec
2017-10-13 12:30:54,815 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 15.34 sec
MapReduce Total cumulative CPU time: 15 seconds 340 msec
a Ended Job = job_1507706430268_0088
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
a Ended Job = job_1507706430268_0088
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0089, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0089/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0089
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 12:31:08,193 Stage-2 map = 0%, reduce = 0%
2017-10-13 12:31:16,769 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.57 sec
2017-10-13 12:31:24,329 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 4.2 sec
MapReduce Total cumulative CPU time: 4 seconds 200 msec
a Ended Job = job_1507706430268_0089
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 15.34 sec HDFS Read: 449886897 HDFS Write: 4176528 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 4.2 sec HDFS Read: 4181811 HDFS Write: 356 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 540 msec
OK
(QUANTITATIVE) SOFTWARE ENGINEER 2015 1
ORACLE APPS DBA 2015 1
ZONE ENGINEER 2015 1
ZLINUX ENGINEER 2015 1
A ZFS STORAGE TEST AUTOMATION SOFTWARE ENGINEER 2015 1
YOUTH PROGRAM DEVELOPER 2015 1
YOUTH PASTOR/ENGLISH MINISTRY PASTOR 2015 1
YOUTH OUTDOORS RECREATION SPECIALIST 2015 1
YOUTH OPERA MUSIC DIRECTOR 2015 1
YOUTH LACROSSE PROGRAM DIRECTOR 2015 1
Time taken: 71.969 seconds, Fetched: 10 row(s)
hive (h1b project)> |
```

Job Title	Year	Count
(QUANTITATIVE) SOFTWARE ENGINEER	2015	1
ORACLE APPS DBA	2015	1
ZONE ENGINEER	2015	1
ZLINUX ENGINEER	2015	1
A ZFS STORAGE TEST AUTOMATION SOFTWARE ENGINEER	2015	1
YOUTH PROGRAM DEVELOPER	2015	1
YOUTH PASTOR/ENGLISH MINISTRY PASTOR	2015	1
YOUTH OUTDOORS RECREATION SPECIALIST	2015	1
YOUTH OPERA MUSIC DIRECTOR	2015	1
YOUTH LACROSSE PROGRAM DIRECTOR	2015	1

```
select job_title, year, count(case_status) as count_case_status from h1b_final where
year = 2016 and case_status='CERTIFIED' group by job_title, year order by
count case status limit 10;
```

```
hduser@ubuntu: ~
YOUTH LACROSSE PROGRAM DIRECTOR 2015 1
Time taken: 71.969 seconds, Fetched: 10 row(s)
hive (h1b.project)> select job_title,year,count(case_status) as count_case_status from h1b_final where year = 2016 and case_status = 'CERTIFIED' group by job_title,year order by count case status limit 10;
Query ID = hduser_20171013123201_ad17aaab-bfb8-4da9-bc01-adf3a17f2146
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0090, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0090
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0090
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 12:32:10,245 Stage-1 map = 0%, reduce = 0%
2017-10-13 12:32:29,851 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 9.3 sec
2017-10-13 12:32:34,060 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 12.51 sec
2017-10-13 12:32:37,292 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 13.57 sec
2017-10-13 12:32:45,784 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 15.43 sec
2017-10-13 12:32:46,851 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 17.69 sec
MapReduce Total cumulative CPU time: 17 seconds 690 msec
Ended Job = job_1507706430268_0090
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
```

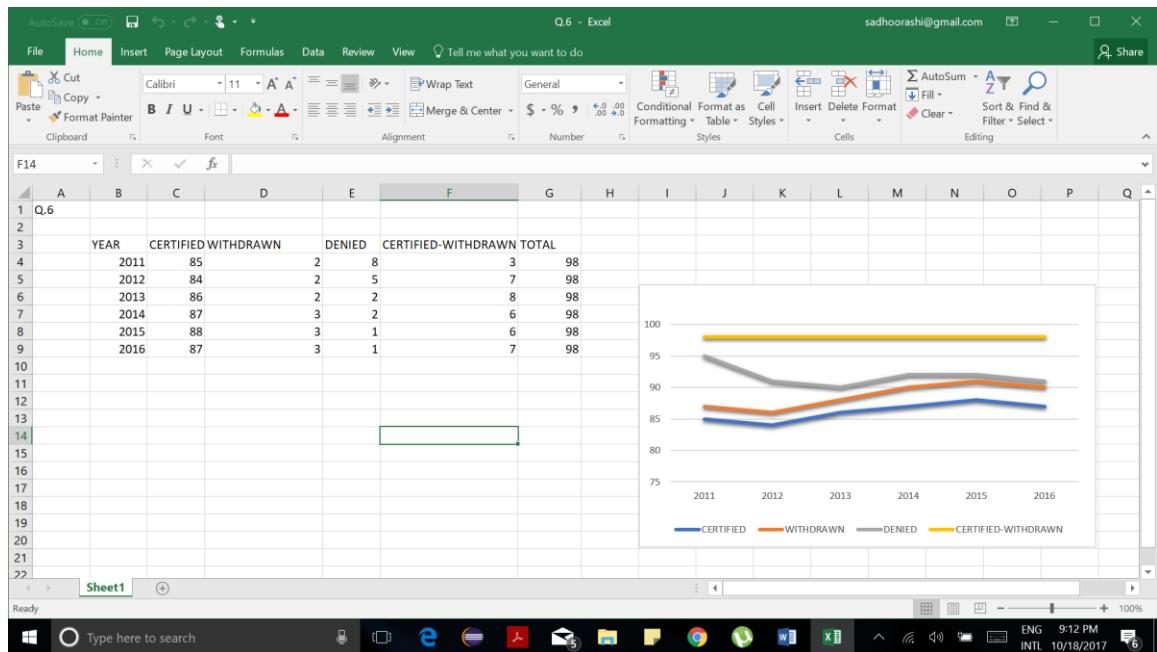
```
hduser@ubuntu: ~
Ended Job = job 1507706430268_0090
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job 1507706430268_0091, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0091
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0091
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 12:33:01,959 Stage-2 map = 0%, reduce = 0%
2017-10-13 12:33:10,700 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.0 sec
2017-10-13 12:33:19,737 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.59 sec
MapReduce Total cumulative CPU time: 3 seconds 590 msec
Ended Job = job 1507706430268_0091
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 17.69 sec HDFS Read: 449886897 HDFS Write: 4332573 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.59 sec HDFS Read: 4337856 HDFS Write: 395 SUCCESS
Total MapReduce CPU Time Spent: 21 seconds 280 msec
00
SR. BUSINESS INTELLIGENCE DEVELOPER 2016 1
SHAREPOINT/SQL DEVELOPER 2016 1
QA ANALYST 2016 1
BUSINESS ANALYST 2016 1
ZONING DATA COMPLIANCE PROJECT ARCHITECT 2016 1
YOUTH THEATRICAL TALENT COORDINATOR 2016 1
YOUTH TENNIS PROGRESSION SPECIALIST 2016 1
YOUTH SPECIALSIT PROGRAM COORDINATOR 2016 1
YOUTH MINISTRY ASSOCIATE 2016 1
YOUTH LITERARY PROGRAMS COORDINATOR 2016 1
Time taken: 80.453 seconds, Fetched: 10 row(s)
hive (h1b.project)>
```

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

**Solution:**

```

hduser@ubuntu:~$ hadoop fs -cat /hadoop/project/pig/question6/p*
2011,DENIED,8,29138
2011,CERTIFIED,85,0,307936
2011,WITHDRAWN,2,0,10105
2011,PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED,3,0,11596
2012,DENIED,5,0,21096
2012,CERTIFIED,84,0,352668
2012,WITHDRAWN,2,0,11596
2012,CERTIFIED-WITHDRAWN,7,0,31118
2013,PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED,0,0,15
2013,WITHDRAWN,2,0,11596
2013,CERTIFIED,86,0,382951
2013,DENIED,3,0,12122
2013,WITHDRAWN,3,0,16034
2014,CERTIFIED,87,0,455144
2014,DENIED,2,0,11596
2014,WITHDRAWN,6,0,36350
2015,DENIED,1,0,10923
2015,CERTIFIED,88,0,47278
2015,WITHDRAWN,4,0,19451
2015,CERTIFIED-WITHDRAWN,6,0,41071
2016,DENIED,3,0,9175
2016,CERTIFIED,87,0,569646
2016,WITHDRAWN,3,0,21890
2016,CERTIFIED-WITHDRAWN,7,0,47092
hduser@ubuntu:~$
```

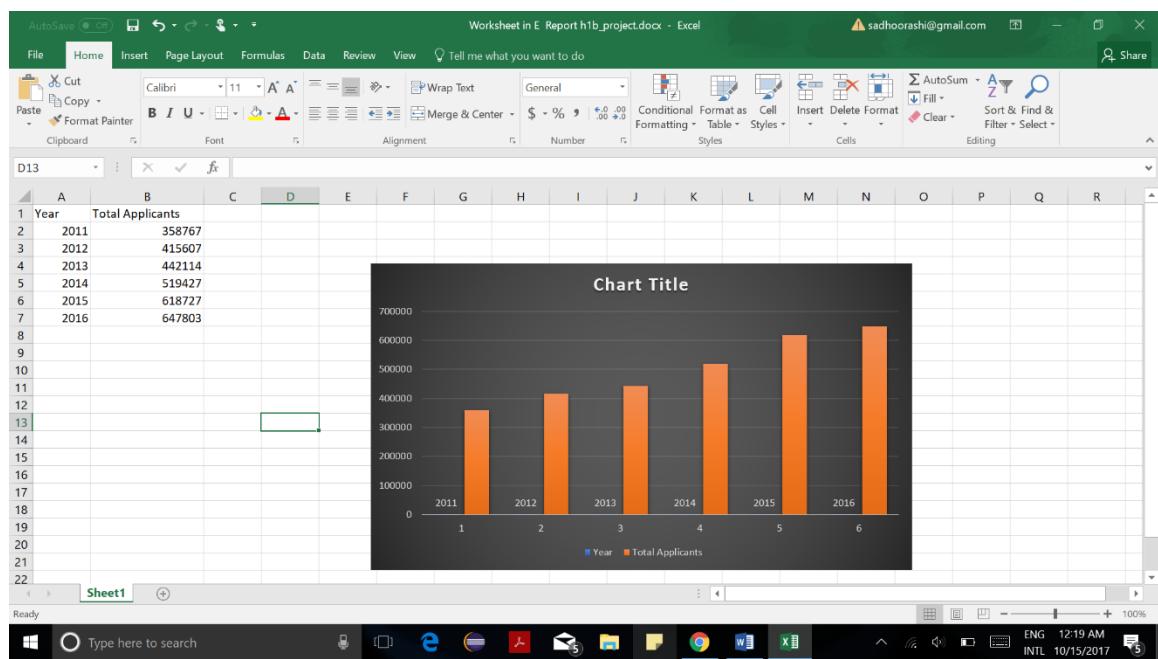


## 7) Create a bar graph to depict the number of applications for each year [All] (HIVE)

### Solution:

```
hduser@ubuntu:~$ YOUTH LITERARY PROGRAMS COORDINATOR 2016 1
Time taken: 80.453 seconds, Fetched: 10 row(s)
hive (hb project)> select year.count(*) as count all from h1b final group by year order by year;
Query ID = hduser_20171014083630_10f455ef-52df-4d94-8246-d1b31d1ac2a3
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0093, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0093/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0093
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-14 08:36:57,149 Stage-1 map = 0%, reduce = 0%
2017-10-14 08:37:45,866 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 21.19 sec
2017-10-14 08:37:47,044 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 22.08 sec
2017-10-14 08:37:51,295 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 23.99 sec
2017-10-14 08:37:52,443 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 24.84 sec
2017-10-14 08:38:10,622 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 29.42 sec
MapReduce Total cumulative CPU time: 29 seconds 420 msec
Ended Job = job_1507706430268_0093
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0094, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0094/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0094
```

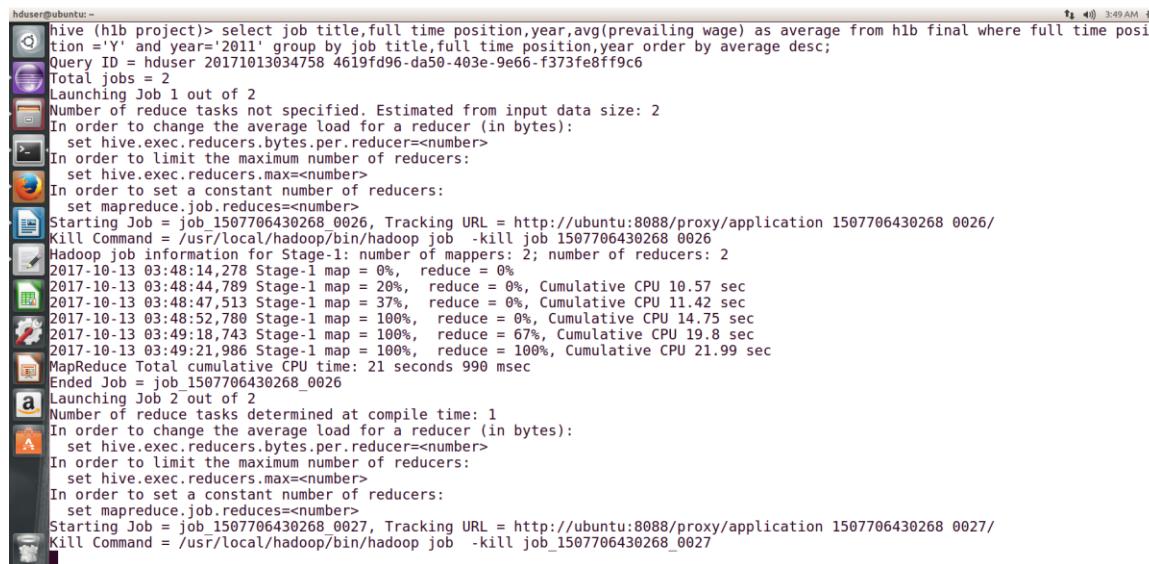
```
hduser@ubuntu:~$ 2017-10-14 08:37:51,205 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 23.99 sec
2017-10-14 08:37:52,443 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 24.84 sec
2017-10-14 08:38:10,622 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 29.42 sec
MapReduce Total cumulative CPU time: 29 seconds 420 msec
Ended Job = job_1507706430268_0093
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0094, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0094/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0094
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-14 08:38:33,558 Stage-2 map = 0%, reduce = 0%
2017-10-14 08:38:44,576 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.71 sec
2017-10-14 08:38:58,861 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 4.48 sec
MapReduce Total cumulative CPU time: 4 seconds 480 msec
Ended Job = job_1507706430268_0094
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 29.42 sec HDFS Read: 449883747 HDFS Write: 348 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 4.48 sec HDFS Read: 5113 HDFS Write: 72 SUCCESS
Total MapReduce CPU Time Spent: 33 seconds 900 msec
OK
2011  358767
2012  415607
2013  442114
2014  519427
2015  618727
2016  647803
Time taken: 150.975 seconds, Fetched: 6 row(s)
hive (hb project)>
```



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

**Solution:**

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2011' group by job_title, full_time_position, year order by average desc;
```



The screenshot shows a terminal window with the following content:

```
hive (h1b project)> select job title,full time position,year,avg(prevailing wage) as average from h1b final where full time position ='Y' and year='2011' group by job title,full time position,year order by average desc;
Query ID = hduser 20171013034758 4619fd96-da50-403e-9e66-f373fe8ff9c6
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0026, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0026/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0026
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 03:48:14,278 Stage-1 map = 0%, reduce = 0%
2017-10-13 03:48:44,789 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 10.57 sec
2017-10-13 03:48:47,513 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 11.42 sec
2017-10-13 03:48:52,780 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 14.75 sec
2017-10-13 03:49:18,743 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 19.8 sec
2017-10-13 03:49:21,980 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 21.99 sec
MapReduce Total cumulative CPU time: 21 seconds 990 msec
Ended Job = job_1507706430268_0026
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0027, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0027/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0027
```

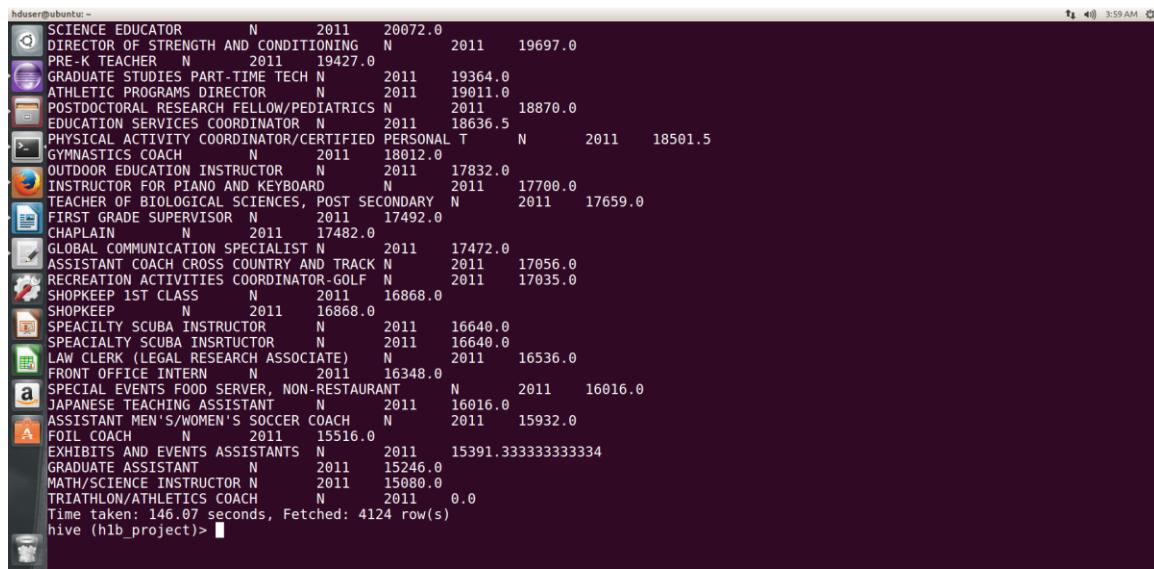
	AUDIO-VISUAL MEDIA SPECIALIST	Y	2011	43035.0
	CLIENT TRAINING SPECIALIST	Y	2011	43035.0
	COMPUTER SPECIALIST, ALL OTHER	Y	2011	43035.0
	JUNIOR ARTIST	Y	2011	43035.0
	GRANT COORDINATOR	Y	2011	43035.0
	COMPUTER SPECIALIST IN PROGRAM DESIGN	Y	2011	43035.0
	BUSINESS INTELLIGENCE AND OPERATIONS COORDINATOR	Y	2011	43035.0
	JR. NETWORK SECURITY ENGINEER	Y	2011	43035.0
	EMAIL DELIVERABILITY ANALYST	Y	2011	43035.0
	JR IMPLEMENTATION ENGINEER	Y	2011	43035.0
	PUBLIC RELATIONS & MARKETING ASSOCIATE	Y	2011	43035.0
	SOFTWARE TESTER SQM CONSULTANT	Y	2011	43035.0
	ASSOCIATE, QUALITY ASSURANCE	Y	2011	43035.0
	WEBSITE MANAGER/DESIGNER	Y	2011	43035.0
	ELECTRICAL ENGINEERING (DESIGN & DRAFTER)	Y	2011	43035.0
	COMPUTER INTEGRATED MANUFACTURING ENGINEER	Y	2011	43035.0
	CLIENT SERVICE CENTER SUPERVISOR	Y	2011	43035.0
	COMMUNICATION SPECIALIST/PR	Y	2011	43035.0
	PUBLIC RELATIONS AND COMMUNICATIONS MANAGER	Y	2011	43035.0
	DATA COLLECTION AND FORENSIC ANALYST	Y	2011	43035.0
	PUBLIC RELATIONS AND COMMUNICATIONS SPECIALIST	Y	2011	43035.0
	WEBSHPERE COMMERCE ADMINISTRATOR	Y	2011	43035.0
	PUBLIC RELATIONS AND PROJECT MANAGER	Y	2011	43035.0
	SR.DOCUMENTUM ARCHITECT	Y	2011	43035.0
	JUNIOR SOFTWARE DEVELOPER E-COMMERCE (JAVA)	Y	2011	43035.0
	PUBLIC RELATIONS AND PROJETC MANAGER	Y	2011	43035.0
	PUBLICITY COORDINATOR	Y	2011	43035.0
	SENIOR ACCOUNT EXECUTIVE, FINANCIAL COMMUNICATIONS	Y	2011	43035.0
	E-COMMERCE SPECIALIST	Y	2011	43035.0
	MANAGER, GLOBAL NETWORK DEVELOPMENT	Y	2011	43035.0
	WEB DESIGN DEVELOPER	Y	2011	43035.0
	QUALITY ASSURANCE (QA) ASSOCIATE	Y	2011	43035.0
	WEB DEVELOPER/IT MANAGER	Y	2011	43035.0
	IT SPECIALIST	Y	2011	43035.0

	CONTENT ENGINEER/WEB DEVELOPER	Y	2011	43035.0
	TAC ENGINEER (U.S.)	Y	2011	43035.0
	TAC ENGINEER (US)	Y	2011	43035.0
	CONTENT SOLUTIONS ANALYST	Y	2011	43035.0
	DATA QUALITY ASSURANCE ANALYST	Y	2011	43035.0
	ASSISTANT HOCKEY COACH	Y	2011	43033.5
	BILINGUAL THIRD GRADE ELEMENTARY SCHOOL TEACHER	Y	2011	43030.0
	BIOPHYSICIST RESEARCH SCIENTIST / ENGINEER	Y	2011	43025.0
	RESTAURANT MANAGER	Y	2011	43016.91836734694
	IT MARKETING COORDINATOR	Y	2011	43014.0
	RESTAURANT MANAGER (OPERATIONS)	Y	2011	43014.0
	IT/GIS DEVELOPER	Y	2011	43014.0
	RESEARCH DATABASE PROGRAMMER	Y	2011	43014.0
	SPECIALTY RESTAURANT MANAGER	Y	2011	43014.0
	LATIN AMERICA MARCOMM SPECIALIST	Y	2011	43014.0
	INTERNATIONAL ACCOUNT ADMINISTRATOR - ASIA	Y	2011	43014.0
	ASSOCIATE DIRECTOR (WINDOWS SERVER ADMINISTRATOR)	Y	2011	43014.0
	PROGRAMMER/ANALYST	Y	2011	43014.0
	ASSOCIATE PROFESSOR OF CHEMISTRY	Y	2011	43010.0
	ASSISTANT PROFESSOR OF ETHNOMUSICOLOGY	Y	2011	43010.0
	AGRICULTURAL MANAGER	Y	2011	43001.8
	JUNIOR DESIGNER, ARCHITECTURE	Y	2011	43000.0
	ELEMENTARY SCHOOL TEACHER	Y	2011	42997.2
	ASSOCIATE APPLICATION DEVELOPER	Y	2011	42994.0
	CRM SIEBEL ACTUATE REPORT WRITER	Y	2011	42994.0
	BUSINESS OPERATIONS SPECIALIST	Y	2011	42994.0
	INVESTMENT ADVISOR FINANCIAL ANALYST	Y	2011	42994.0
	HUMAN RESOURCES COUNSELOR	Y	2011	42994.0
	EMPLOYEE RELATIONS MANAGER	Y	2011	42993.0
	PHARMA DATA CONSULTANT	Y	2011	42991.545454545456
	APPLICATION SYSTEMS ANALYST PROGRAMMER - ASSOCIATE	Y	2011	42979.666666666666
	GENERAL ACCOUNTANT	Y	2011	42976.2
	GREEN COFFEE QUALITY CONTROL COORDINATOR	Y	2011	42973.0
	SENIOR POST-DOCCTORAL RESEARCH FELLOW	Y	2011	42973.0

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2011' group by job_title, full_time_position, year order by average desc;
```

```
hduser@ubuntu:~$ SENIOR COACH AND DEPUTY ACADEMY ADMINISTRATOR Y 2011 15520.0
PROFESSIONAL SKILLS COACH Y 2011 15520.0
ASSISTANT DIRECTOR OF COACHES Y 2011 15420.0
COSMETOLOGISTS Y 2011 15308.0
TECH ASSISTANT ELL (LEAD) Y 2011 15240.0
FARM SOLAR ENERGY INTL INC Y 2011 15204.0
FARM WORKER Y 2011 15204.0
CERTIFIED STRENGTH CONDITIONING SPECIALIST/ATHLETIC SPECIALIST Y 2011 15170.0
LEGISLATIVE SECRETARY Y 2011 15110.0
LECTURER IN PHYSICS Y 2011 15080.0
PURE TALENT TRAINING STYLIST Y 2011 15080.0
SPECIAL CLASS ASSISTANT Y 2011 15070.0
VOLUNTEER IN QUALITY RESEARCH FELLOWSHIP Y 2011 12000.0
Time taken: 0:42.431 seconds Fetched: 61377 row(s)
hive (hdb-project) > select job_title,full_time_position,year,avg(prevailing_wage) as average from hlb_final where full_time_position
> 'N' and year=2011 group by job_title,full_time_position,year order by average desc;
Query ID : hduser_20171013035424_d1552669-fda9-4def-8ef4-6825bb06a832
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0028, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0028/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0028
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 03:54:42,056 Stage-1 map 0%, reduce = 0%
2017-10-13 03:55:24,468 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 13.85 sec
2017-10-13 03:55:26,838 Stage-1 map = 56%, reduce = 0%, Cumulative CPU 14.65 sec
2017-10-13 03:55:27,949 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 15.55 sec
2017-10-13 03:55:32,785 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 16.71 sec
```

FINANCIAL ADMINISTRATOR N	2011	43451.0
PRINCIPAL (ART DIRECTOR) N	2011	43430.0
CONSULTANT ON INTERNATIONAL FREIGHT AND LOGISTICS N	2011	43430.0
RECRUITER N	2011	43430.0
NETWORK/COMPUTER SYSTEMS ADMINISTRATOR N	2011	43430.0
TAX CREDIT ANALYST N	2011	43409.0
STAFF ACCOUNTANT-ANALYST N	2011	43409.0
MANAGERIAL ACCOUNTANT (COST ACCOUNTANCY) N	2011	43409.0
FINANCIAL CONTROL STAFF N	2011	43409.0
3D MODELER N	2011	43409.0
MJC AMERICA, LTD. N	2011	43409.0
STAFF ACCOUNTANT N	2011	43409.0
ACCOUNTING N	2011	43409.0
ACCOUNTANT, COST N	2011	43409.0
MANAGEMENT ACCOUNTANT (COST AND SYSTEMS ACCOUNTING) N	2011	43409.0
SR. ACCOUNTANT N	2011	43409.0
PROPERTY ACCOUNTANT N	2011	43409.0
ACCOUNTANT/AUDITOR N	2011	43409.0
CORPORATE ACCOUNTANT / LIAISON N	2011	43409.0
TEXTILE & OPERATIONS COORDINATOR N	2011	43388.0
SALES INVENTORY LOGISTICIAN N	2011	43388.0
BUSINESS OPERATIONS COORDINATOR N	2011	43388.0
PROGRAM ADMINISTRATION SPECIALIST N	2011	43388.0
SCIENTIST I N	2011	43368.0
PHYSICAL EDUCATION TEACHER N	2011	43358.916666666666
WEB DEVELOPER & ADMINISTRATOR N	2011	43347.0
OPERATIONS SPECIALIST N	2011	43346.8
COMMERCIAL DESIGNER N	2011	43346.666666666664
QUANTITATIVE ENGINEER N	2011	43326.0
GENERAL ACCOUNTANT N	2011	43318.125
SALES/MARKETING ANALYST N	2011	43305.0
CONDUCTOR N	2011	43305.0
BUSINESS DEVELOPMENT ANALYST/SPECIALIST N	2011	43305.0
MARKET RESEACH ANALYST N	2011	43305.0

```
buser@ubuntu:~  
hive (h1b_project)> 

| Job Title                                          | Year   | Salary             |
|----------------------------------------------------|--------|--------------------|
| SCIENCE EDUCATOR                                   | N 2011 | 20072.0            |
| DIRECTOR OF STRENGTH AND CONDITIONING              | N 2011 | 19697.0            |
| PRE-K TEACHER                                      | N 2011 | 19427.0            |
| GRADUATE STUDIES PART-TIME TECH                    | N 2011 | 19364.0            |
| ATHLETIC PROGRAMS DIRECTOR                         | N 2011 | 19011.0            |
| POSTDOCTORAL RESEARCH FELLOW/PEDIATRICS            | N 2011 | 18870.0            |
| EDUCATION SERVICES COORDINATOR                     | N 2011 | 18636.5            |
| PHYSICAL ACTIVITY COORDINATOR/CERTIFIED PERSONAL T | N 2011 | 18501.5            |
| GYMNASTICS COACH                                   | N 2011 | 18012.0            |
| OUTDOOR EDUCATION INSTRUCTOR                       | N 2011 | 17832.0            |
| INSTRUCTOR FOR PIANO AND KEYBOARD                  | N 2011 | 17700.0            |
| TEACHER OF BIOLOGICAL SCIENCES, POST SECONDARY     | N 2011 | 17659.0            |
| FIRST GRADE SUPERVISOR                             | N 2011 | 17492.0            |
| CHAPLAIN                                           | N 2011 | 17482.0            |
| GLOBAL COMMUNICATION SPECIALIST                    | N 2011 | 17472.0            |
| ASSISTANT COACH CROSS COUNTRY AND TRACK            | N 2011 | 17056.0            |
| RECREATION ACTIVITIES COORDINATOR-GOLF             | N 2011 | 17035.0            |
| SHOPKEEP 1ST CLASS                                 | N 2011 | 16868.0            |
| SHOPKEEP                                           | N 2011 | 16868.0            |
| SPECIALTY SCUBA INSTRUCTOR                         | N 2011 | 16640.0            |
| SPECIALTY SCUBA INSRTUTOR                          | N 2011 | 16640.0            |
| LAW CLERK (LEGAL RESEARCH ASSOCIATE)               | N 2011 | 16536.0            |
| FRONT OFFICE INTERN                                | N 2011 | 16348.0            |
| SPECIAL EVENTS FOOD SERVER, NON-RESTAURANT         | N 2011 | 16016.0            |
| JAPANESE TEACHING ASSISTANT                        | N 2011 | 16016.0            |
| ASSISTANT MEN'S/WOMEN'S SOCCER COACH               | N 2011 | 15932.0            |
| FOIL COACH                                         | N 2011 | 15516.0            |
| EXHIBITS AND EVENTS ASSISTANTS                     | N 2011 | 15391.333333333334 |
| GRADUATE ASSISTANT                                 | N 2011 | 15246.0            |
| MATH/SCIENCE INSTRUCTOR                            | N 2011 | 15080.0            |
| TRIATHLON/ATHLETICS COACH                          | N 2011 | 0.0                |



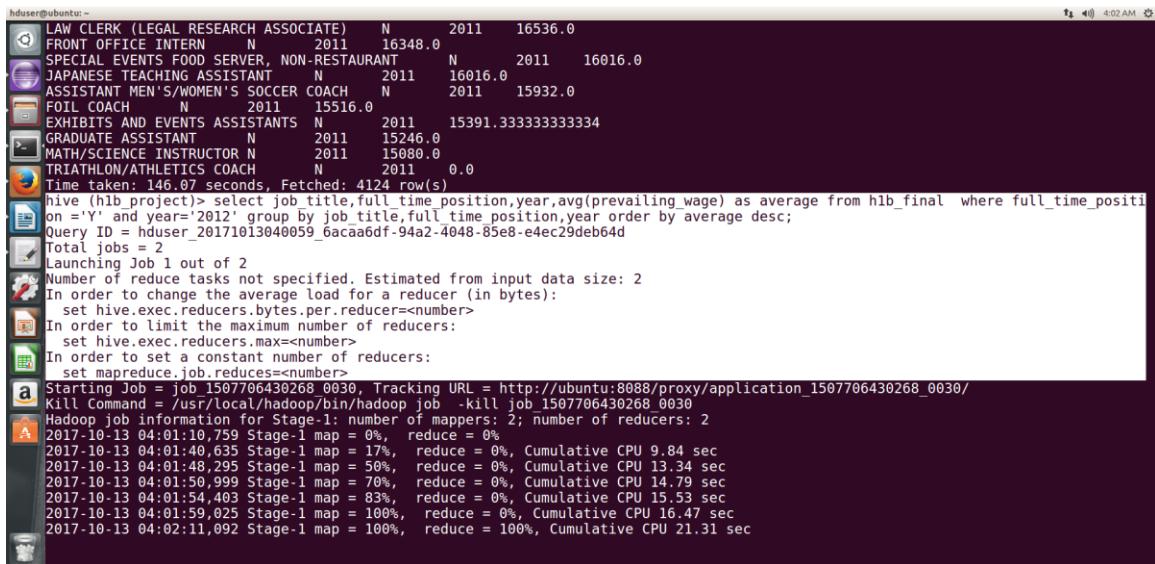
Time taken: 146.07 seconds, Fetched: 4124 row(s)



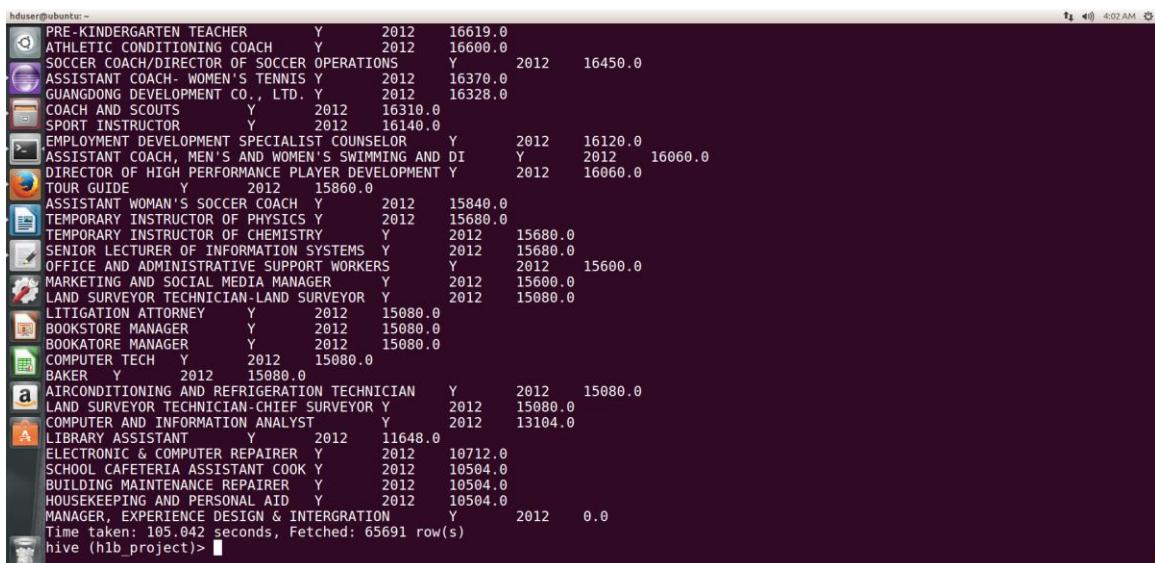
hive (h1b_project)>


```

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2012' group by job_title, full_time_position, year order by average desc;
```



```
hduser@ubuntu:~  
hive (h1b_final) > select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2012' group by job_title, full_time_position, year order by average desc;  
Time taken: 146.07 seconds, Fetched: 4124 row(s)  
hive (h1b_final) > select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2012' group by job_title, full_time_position, year order by average desc;  
Query ID = hduser_20171013040059_6acaa6df-94a2-4048-85e8-e4ec29deb64d  
Total jobs = 2  
Launching Job 1 out of 2  
Number of reduce tasks not specified. Estimated from input data size: 2  
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 mapreduce.job.reduces=<number>  
Starting Job = job_1507706430268_0030, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0030  
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0030  
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2  
2017-10-13 04:01:10,759 Stage-1 map = 0%, reduce = 0%  
2017-10-13 04:01:40,635 Stage-1 map = 17%, reduce = 0%, Cumulative CPU 9.84 sec  
2017-10-13 04:01:48,295 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 13.34 sec  
2017-10-13 04:01:50,999 Stage-1 map = 70%, reduce = 0%, Cumulative CPU 14.79 sec  
2017-10-13 04:01:54,403 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 15.53 sec  
2017-10-13 04:01:59,025 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 16.47 sec  
2017-10-13 04:02:11,092 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 21.31 sec
```



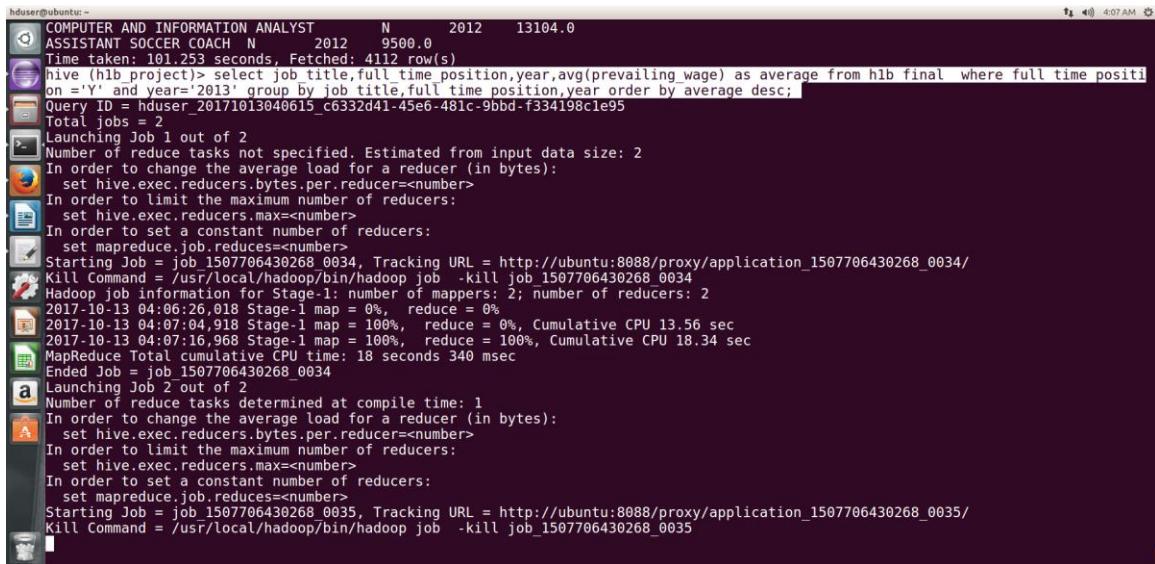
```
hduser@ubuntu:~  
hive (h1b_final) > select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2012' group by job_title, full_time_position, year order by average desc;  
Time taken: 105.042 seconds, Fetched: 65691 row(s)  
hive (h1b_final) >
```

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2012' group by job_title, full_time_position, year order by average desc;
```

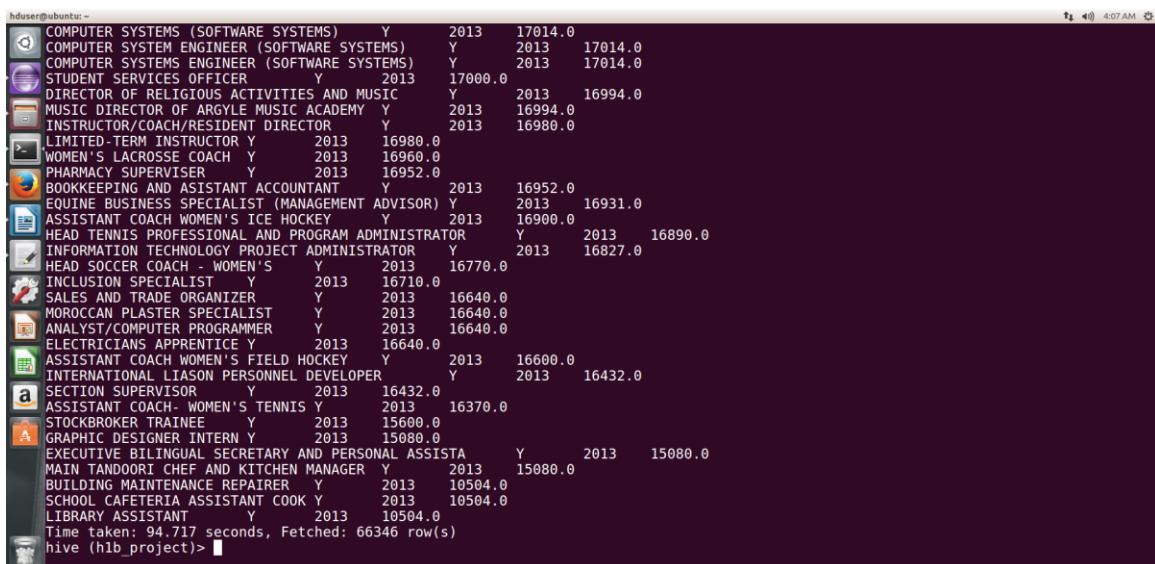
```
hduser@ubuntu:~$ Time taken: 105.042 seconds, Fetched: 65691 row(s)
hive (hib_project)> select job_title,full_time_position,year,avg(prevailing_wage) as average from h1b_final where full_time_position = 'N' and year='2012' group by job_title,full_time_position,year order by average desc;
Query ID = hduser_20171013040340_44defdc9-9604-4145-a7f9-86f7129faec4
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0032, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0032/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0032
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 04:03:51.083 Stage-1 map = 0%, reduce = 0%
2017-10-13 04:04:16.367 Stage-1 map = 17%, reduce = 0%, Cumulative CPU 10.37 sec
2017-10-13 04:04:23.531 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 11.98 sec
2017-10-13 04:04:29.483 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 13.25 sec
2017-10-13 04:04:44.015 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 16.67 sec
MapReduce Total cumulative CPU time: 16 seconds 670 msec
Ended Job = job_1507706430268_0032
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0033, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0033/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0033
```

MUSLIM CHAPLAIN N	2012	26046.25				
INSTRUCTIONAL COORDINATOR /EVENT COORDINATOR N	2012	26000.0				
EVENT COORDINATOR N	2012	26000.0				
INTERN N	2012	26000.0				
LICENSED PROFESSIONAL COUNSELING INTERN N	2012	25937.0				
DIRECTOR, RELIGIOUS ACTIVITIES AND EDUCATION N	2012	25896.0				
INVOICING, ACCOUNTING SPECIALIST N	2012	25896.0				
CHAT BILINGUAL THERAPIST N	2012	25875.0				
RELIGIOUS EDUCATION DIRECTOR N	2012	25875.0				
REHABILITATION COUNSELOR N	2012	25802.0				
VETERINARY TECHNOLOGIST N	2012	25729.0				
RENDERING SPECIALIST N	2012	25708.0				
TEACHER ASSISTANT N	2012	25688.0				
CHILDREN'S PROGRAM COUNSELOR N	2012	25563.0				
CREATIVE DESIGNER N	2012	25521.0				
3D GRAPHIC DESIGNER N	2012	25521.0				
VETERINARY TECHNICIAN N	2012	25230.0				
VETERINARY TECHNICIAN N	2012	25230.0				
CAREER INSTRUCTOR OF PORTUGUESE N	2012	25209.0				
BEHAVIOR SPECIALIST N	2012	25084.0				
RECREATION COORDINATOR N	2012	25064.0				
PRESCHOOL TEACHER N	2012	25015.0				
NEWS REPORTER N	2012	25008.1111111111				
FLORAL DESIGNER N	2012	24960.0				
ESL TEACHER & PROGRAM COORDINATOR N	2012	24960.0				
MUSIC TECHNICIAN N	2012	24960.0				
ASSISTANT TEACHER N	2012	24641.0				
ASSISTANT ACADEMY DIRECTOR N	2012	24564.0				
DIRECTOR OF CHILDREN'S MINISTRY N	2012	24419.0				
DIRECTOR OF RELIGIOUS, CULTURAL AND EDUCATIONAL AC N	2012	24419.0				
PHILOSOPHY INSTRUCTOR N	2012	24419.0				
CHRISTIAN EDUCATION DIRECTOR N	2012	24419.0				
SPANISH PRE-SCHOOL/KINDERGARTEN TEACHER N	2012	24356.0				
POST DOCTORAL RESEARCH FELLOW N	2012	24336.0				

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2013' group by job_title, full_time_position, year order by average desc;
```



```
hduser@ubuntu:~$ COMPUTER AND INFORMATION ANALYST      N    2012    13104.0
hduser@ubuntu:~$ ASSISTANT SOCCER COACH N    2012    9500.0
Time taken: 101.253 seconds. Fetched: 4112 row(s)
hive (h1b_project)> select job_title,full_time_position,year,avg(prevailing_wage) as average from h1b_final where full_time_position = 'Y' and year='2013' group by job_title,full_time_position,year order by average desc;
Query ID = hduser_20171013040615_c6332d41-45e6-481c-9bbd-f334198c1e95
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0034, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0034/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0034
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 04:06:26.018 Stage-1 map = 0%,  reduce = 0%, Cumulative CPU 13.56 sec
2017-10-13 04:07:04.918 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 18.34 sec
MapReduce Total cumulative CPU time: 18 seconds 340 msec
Ended Job = job_1507706430268_0034
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0035, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0035/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0035
Killed Job = job_1507706430268_0035
```



```
hduser@ubuntu:~$ COMPUTER SYSTEMS (SOFTWARE SYSTEMS) Y    2013    17014.0
hduser@ubuntu:~$ COMPUTER SYSTEM ENGINEER (SOFTWARE SYSTEMS) Y    2013    17014.0
hduser@ubuntu:~$ COMPUTER SYSTEMS ENGINEER (SOFTWARE SYSTEMS) Y    2013    17014.0
hduser@ubuntu:~$ STUDENT SERVICES OFFICER Y    2013    17000.0
hduser@ubuntu:~$ DIRECTOR OF RELIGIOUS ACTIVITIES AND MUSIC Y    2013    16994.0
hduser@ubuntu:~$ MUSIC DIRECTOR OF ARGYLE MUSIC ACADEMY Y    2013    16994.0
hduser@ubuntu:~$ INSTRUCTOR/COACH/RESIDENT DIRECTOR Y    2013    16980.0
hduser@ubuntu:~$ LIMITED-TERM INSTRUCTOR Y    2013    16980.0
hduser@ubuntu:~$ WOMEN'S LACROSSE COACH Y    2013    16960.0
hduser@ubuntu:~$ PHARMACY SUPERVISER Y    2013    16952.0
hduser@ubuntu:~$ BOOKKEEPING AND ASSISTANT ACCOUNTANT Y    2013    16952.0
hduser@ubuntu:~$ EQUINE BUSINESS SPECIALIST (MANAGEMENT ADVISOR) Y    2013    16931.0
hduser@ubuntu:~$ ASSISTANT COACH WOMEN'S ICE HOCKEY Y    2013    16900.0
hduser@ubuntu:~$ HEAD TENNIS PROFESSIONAL AND PROGRAM ADMINISTRATOR Y    2013    16890.0
hduser@ubuntu:~$ INFORMATION TECHNOLOGY PROJECT ADMINISTRATOR Y    2013    16827.0
hduser@ubuntu:~$ HEAD SOCCER - WOMEN'S Y    2013    16770.0
hduser@ubuntu:~$ INCLUSION SPECIALIST Y    2013    16710.0
hduser@ubuntu:~$ SALES AND TRADE ORGANIZER Y    2013    16640.0
hduser@ubuntu:~$ MOROCCAN PLASTER SPECIALIST Y    2013    16640.0
hduser@ubuntu:~$ ANALYST/COMPUTER PROGRAMMER Y    2013    16640.0
hduser@ubuntu:~$ ELECTRICIANS APPRENTICE Y    2013    16640.0
hduser@ubuntu:~$ ASSISTANT COACH WOMEN'S FIELD HOCKEY Y    2013    16600.0
hduser@ubuntu:~$ INTERNATIONAL LIASON PERSONNEL DEVELOPER Y    2013    16432.0
hduser@ubuntu:~$ SECTION SUPERVISOR Y    2013    16432.0
hduser@ubuntu:~$ ASSISTANT COACH- WOMEN'S TENNIS Y    2013    16370.0
hduser@ubuntu:~$ STOCKBROKER TRAINEE Y    2013    15600.0
hduser@ubuntu:~$ GRAPHIC DESIGNER INTERN Y    2013    15080.0
hduser@ubuntu:~$ EXECUTIVE BILINGUAL SECRETARY AND PERSONAL ASSISTA Y    2013    15080.0
hduser@ubuntu:~$ MAIN TANDOORI CHEF AND KITCHEN MANAGER Y    2013    15080.0
hduser@ubuntu:~$ BUILDING MAINTENANCE REPAIRER Y    2013    10504.0
hduser@ubuntu:~$ SCHOOL CAFETERIA ASSISTANT COOK Y    2013    10504.0
hduser@ubuntu:~$ LIBRARY ASSISTANT Y    2013    10504.0
Time taken: 94.717 seconds, Fetched: 66346 row(s)
hive (h1b_project)> 
```

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2013' group by job_title, full_time_position, year order by average desc;
```

```
hduser@ubuntu:~$ LIBRARY ASSISTANT Y 2013 10504.0
Time taken: 94.717 seconds, Fetched: 66346 row(s)
hive (hib_project)> select job_title,full_time_position,year,avg(prevailing_wage) as average from h1b_final where full_time_position = 'N' and year='2013' group by job_title,full_time_position,year order by average desc;
Query ID = hduser_20171013040832_8a14b549-bc82-45d0-988d-9ae8a5d9ed4c
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0036, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0036/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0036
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 04:08:42,503 Stage-1 map = 0%, reduce = 0%
2017-10-13 04:09:08,050 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 8.82 sec
2017-10-13 04:09:14,291 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 10.81 sec
2017-10-13 04:09:26,964 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 13.74 sec
MapReduce Total cumulative CPU time: 13 seconds 740 msec
Ended Job = job_1507706430268_0036
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0037, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0037/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0037
hduser@ubuntu:~$
```

```
hduser@ubuntu:~$ PASTORAL ADMINISTRATOR N 2013 22568.0
ASSISTANT HEAD GYMNASTICS COACH N 2013 21840.0
GRADUATE ASSISTANT N 2013 21174.0
DIRECTOR OF CHARLOTTESVILLE INSTITUTE OF CHRISTIAN N 2013 21070.0
SUZUKI PIANO INSTRUCTOR N 2013 20966.0
TUTOR N 2013 20914.0
DFGDFG N 2013 20800.0
ADJUNCT INSTRUCTOR OF MATH N 2013 20654.0
PROGRAM ASSISTANT N 2013 20633.0
INVESTMENT RESEARCH ASSOCIATE N 2013 20155.0
PERFORMANCE COACH N 2013 20092.0
PIANO & FLUTE TEACHER N 2013 20009.0
OUTDOOR EDUCATION INSTRUCTOR N 2013 19926.0
PRIVATE INSTRUCTOR N 2013 19864.0
DIRECTOR OF EDUCATION AND CURRICULUM DEVELOPMENT N 2013 19739.0
SAXOPHONE TEACHER N 2013 19489.0
SCUBA INSTRUCTOR N 2013 19260.0
INSTRUCTOR OF TAIKO DRUMMING N 2013 19240.0
STEM TEACHER N 2013 19136.0
LEAD TEACHER N 2013 17908.0
TEMPORARY ADJUNCT INSTRUCTOR N 2013 17596.0
COSTUMER N 2013 17420.0
FRENCH TEACHING ASSISTANT N 2013 17097.0
JAPANESE TEACHING ASSISTANT N 2013 17076.0
INSTRUCTOR, PSYCHOLOGY N 2013 16972.0
ASSISTANT VOLLEYBALL COACH N 2013 16889.0
DIRECTOR OF RELIGIOUS ACTIVITIES AND EDUCATION N 2013 16764.0
ASSISTANT WOMEN'S SOCCER COACH N 2013 16120.0
COOPERATIVE ENTERPRISE DEVELOPMENT COORDINATOR N 2013 15600.0
CHEMISTRY ADJUNCT N 2013 15392.0
EXECUTIVE BILINGUAL SECRETARY AND PERSONAL ASSISTANT N 2013 15080.0
BILINGUAL SPECIAL EDUCATION TEACHER ASSISTANT N 2013 15080.0
Time taken: 85.799 seconds, Fetched: 3661 row(s)
hive (hib_project)> hduser@ubuntu:~$
```

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2014' group by job_title, full_time_position, year order by average desc;
```

```
hduser@ubuntu:~$ 
ASSISTANT WOMEN'S SOCCER COACH N 2013 16120.0
COOPERATIVE ENTERPRISE DEVELOPMENT COORDINATOR N 2013 15600.0
CHEMISTRY ADJUNCT N 2013 15392.0
EXECUTIVE BILINGUAL SECRETARY AND PERSONAL ASSISTA N 2013 15080.0
BILINGUAL SPECIAL EDUCATION TEACHER ASSISTANT N 2013 15080.0
Time taken: 85.799 seconds, Fetched: 3661 row(s)
hive (hb1_project)> select job_title,full time position,year,avg(prevailing wage) as average from h1b final where full time position = 'Y' and year='2014' group by job title,full time position,year order by average desc;
Query ID = hduser 2017013041042 e2b490cf-a39c-45ca-b2bf-e187eadc50a0
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0038, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0038
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0038
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 04:10:50,258 Stage-1 map = 0%, reduce = 0%
2017-10-13 04:11:18,186 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 8.78 sec
2017-10-13 04:11:20,330 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 9.8 sec
2017-10-13 04:11:27,074 Stage-1 map = 56%, reduce = 0%, Cumulative CPU 10.43 sec
2017-10-13 04:11:28,111 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 13.59 sec
2017-10-13 04:11:31,307 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 15.82 sec
2017-10-13 04:11:40,774 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 17.85 sec
2017-10-13 04:11:41,820 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 20.04 sec
MapReduce Total cumulative CPU time: 20 seconds 40 msec
Ended Job = job_1507706430268_0038
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
```

```
hduser@ubuntu:~$ 
HEAD COACH BRONCO MEN'S TENNIS Y 2014 17220.0
ASSISTANT MEN'S HOCKEY COACH NCAA DIV III Y 2014 17120.0
ASSISTANT PROFESSOR IN CHEMISTRY Y 2014 17060.0
ASSISTANT PROFESSOR IN CHEMISTRY Y 2014 17060.0
DIRECTOR OF TENNIS MARKETING AND OPERATIONS/ INTER Y 2014 17060.0
HEAD SOCCER COACH - MEN Y 2014 17040.0
ASSOCIATE PROFESSOR IN FOREIGN LANGUAGE - PORTUGUE Y 2014 17030.0
HEAD COACH: WOMEN'S SOCCER Y 2014 17030.0
MEN'S SOCCER COACH Y 2014 17030.0
ASSISTANT PROFESSOR IN ITALIAN Y 2014 17030.0
INTERPETER/TRANSLATOR Y 2014 16972.0
INFANT TEACHER Y 2014 16910.0
ASSISTANT COACH MEN'S TENNIS Y 2014 16880.0
RESIDENCE LIFE COORDINATOR Y 2014 16869.0
BABY SITTER / NANNY Y 2014 16868.0
MINISTER FOR CHURCH DEVELOPMENT & DISCIPLESHIP PRO Y 2014 16806.0
CLASS 1/2 TEACHER FOR REGISTERED HOMESCHOOL Y 2014 16800.0
GOLF INSTRUCTOR & ASSISTANT GOLF PROFESSIONAL Y 2014 16780.0
HEAD COACH, WOMEN'S LACROSSE Y 2014 16710.0
INSTRUCTOR IN DRAMA Y 2014 16710.0
DIRECTOR-SPECIAL EVENTS AND PROGRAMS Y 2014 16682.0
CHILD CARE CENTER DIRECTOR Y 2014 16640.0
ASSISTANT COACH, MEN'S GOLF Y 2014 16570.0
ASSISTANT MEN'S AND WOMEN'S SOCCER COACH Y 2014 16560.0
LICENSED PROFESSIONAL COUNSELOR Y 2014 16536.0
SECTION SUPERVISOR Y 2014 16432.0
PROGRAM CONSULTANT Y 2014 15600.0
ENGINEERING TRAINEE Y 2014 15680.0
ORGANIC SEARCH MARKETER Y 2014 15680.0
POSTSECONDARY TEACHER Y 2014 15680.0
ARCHITECTS Y 2014 14580.0
POST-DOCTORAL RESEARCH Y 2014 0.0
Time taken: 91.777 seconds, Fetched: 74369 row(s)
hive (hb1_project)> ■
```

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2014' group by job_title, full_time_position, year order by average desc;
```

```
hduser@ubuntu:~$ hive (h1b_project)> select job_title,full_time_position,year,avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2014' group by job_title,full_time_position,year order by average desc;
Query ID = hduser_20171013041328_71f1b119-ecaa-4229-a8eb-858950a81c0a
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0040, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0040/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0040
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 04:13:36,972 Stage-1 map = 0%,  reduce = 0%
2017-10-13 04:13:54,741 Stage-1 map = 37%,  reduce = 0%, Cumulative CPU 7.27 sec
2017-10-13 04:13:55,800 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 7.87 sec
2017-10-13 04:13:56,843 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.28 sec
2017-10-13 04:14:07,403 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 10.81 sec
2017-10-13 04:14:08,455 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12.33 sec
MapReduce Total cumulative CPU time: 12 seconds 330 msec
Ended Job = job_1507706430268_0040
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0041, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0041/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0041
```

SUBSTITUTE TEACHER	N	2014	20238.0
LECTURER - CHINESE LANGUAGES	N	2014	20196.0
EARLY CHILDHOOD ASSISTANT	N	2014	19801.0
PRESCHOOL TEACHER	N	2014	19244.75
INSTRUCTOR OF TAIKO DRUMMING	N	2014	19198.0
TAEKWONDO INSTRUCTOR	N	2014	19156.0
TAEKWONDO	N	2014	19156.0
SITE DIRECTOR	N	2014	19073.0
MARKETING AND FLORAL COORDINATOR/BARISTA	N	2014	18928.0
TEACHER ASSISTANTS	N	2014	18449.0
ASSISTANT COACH, WOMEN'S SOCCER TEAM	N	2014	18366.0
GALLERY CURATORIAL ASSISTANT	N	2014	18096.0
ASSISTANT MEN'S SOCCER COACH	N	2014	18033.0
DIRECTOR OF COLORADO ACADEMY FIELD HOCKEY	N	2014	18012.0
FIRST LINE SUPERVISOR OF FOOD PREPARATION AND SERV	N	2014	17888.0
FIRST-LINE SUPERVISOR OF FOOD PREPARATION AND SERV	N	2014	17888.0
EXECUTIVE BILINGUAL SECRETARY AND PERSONAL ASSISTA	N	2014	17867.0
MUSICIAN INSTRUMENTAL	N	2014	17846.0
EXECUTIVE DRESSAGE TRAINER	N	2014	17784.0
WOMEN'S BASKETBALL ASSISTANT COACH	N	2014	17555.0
STEAM TEACHER	N	2014	17097.0
MARKETING & ADVERTISING SUPPORT	N	2014	17076.0
INFANT TEACHER	N	2014	16910.0
PERSONAL ASSISTANT/ HAIRSTYLIST	N	2014	16827.0
BOX OFFICE STAFF/PERFORMER	N	2014	16785.0
PRE-K MONTESSORI SCHOOL TEACHER	N	2014	16764.0
MONTESSORI SCHOOL TEACHER	N	2014	16764.0
RESIDENT ARTIST	N	2014	16723.0
ASSISTANT COACH	N	2014	16702.0
BILINGUAL SPECIAL EDUCATIN TEACHER ASSISTANT	N	2014	15080.0
LOBSTER WEIGHING TECH/WELDER	N	2014	15080.0
BILINGUAL SPECIAL EDUCAITON TEACHER ASSISTANT	N	2014	15080.0

Time taken: 70.635 seconds, Fetched: 3795 row(s)

```
hive (h1b_project)> █
```

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2015' group by job_title, full_time_position, year order by average desc;
```

Time taken: 70.635 seconds, Fetched: 3795 row(s)

```
hive (h1b_project)> select job_title,full_time_position,year,avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2015' group by job_title,full_time_position,year order by average desc;
Query ID = hduser_20171013041603_calldaa2-8le0-4865-a37c-5c1d03242bdc
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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=<n>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<n>
Starting Job = job_1507706430268_0042, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0042/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0042
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 04:16:11.016 Stage-1 map = 0%, reduce = 0%
2017-10-13 04:16:28.819 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 7.06 sec
2017-10-13 04:16:29.935 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 7.76 sec
2017-10-13 04:16:32.048 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 9.59 sec
2017-10-13 04:16:35.306 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 11.03 sec
2017-10-13 04:16:39.228 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 11.9 sec
2017-10-13 04:16:47.697 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 15.82 sec
MapReduce Total cumulative CPU time: 15 seconds 820 msec
Ended Job = job_1507706430268_0042
Launching Job 2 out of 2
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=<n>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<n>
```

```
hduser@ubuntu:-
FALCONRY LEADER Y 2015 17014.0
SERVER / PREP COOK Y 2015 17014.0
COMMUNITY AND SOCIAL SERVICE ASSOCIATE Y 2015 17014.0
MONTESSORI ASSISTANT TEACHER Y 2015 17000.0
ASSISTANT COACH TENNIS Y 2015 17000.0
ADJUNCT FACULTY Y 2015 16952.0
FRONT DESK LEAD Y 2015 16931.0
ASSISTANT PROFESSOR OF FOREIGN LANGUAGE - GERMAN Y 2015 16920.0
REGIONAL SOCCER DIRECTOR Y 2015 16890.0
BOYS ACADEMY DIRECTOR Y 2015 16858.5
MARKETING/PHOTOGRAPHER Y 2015 16848.0
CHRISTIAN EDUCATION DIRECTOR Y 2015 16806.0
CAREGIVER Y 2015 16785.0
PERSONAL CARE WORKER Y 2015 16764.0
ASSISTANT PROFESSOR - EC ENGINEERING Y 2015 16750.0
CUSTOMER SERVICE REP Y 2015 16744.0
DIRECTOR OF JUNIOR TENNIS Y 2015 16723.0
HEAD FITNESS AND GOLF INSTRUCTOR Y 2015 16720.0
COLLEGE PROGRAMS COORDINATOR Y 2015 16702.0
BARN WORKER, BREEDING ASSISTAND ANIMAL GROOMER & CARER Y 2015 16660.0
BARN WORKER , BREEDING ASSISTAND ANIMAL GROOMER & CARER Y 2015 16660.0
RELIGIOUS LANGUAGE INSTRUCTOR Y 2015 16640.0
JEWISH OUTREACH AND ENGAGEMENT EDUCATOR Y 2015 16578.0
ASSISTANT PROFESSOR EDUCATION Y 2015 16570.0
RESIDENTIAL ADVISOR Y 2015 16340.0
BAKER, CASHIER, DELI PERSONEL Y 2015 15912.0
AUTHENTIC CHINESE CUISINE CHEF/COOK Y 2015 15600.0
SALES MANAGER, AUSTIN BRANCH Y 2015 15080.0
HUMAN RESOURCE ASSISTANT Y 2015 15080.0
INTERNATIONAL SALES RAP Y 2015 2000.0
ADSF Y 2015 52.0
VOLUNTEER CULTURAL EXCHANGE EDUCATOR Y 2015 0.0
Time taken: 77.016 seconds, Fetched: 81224 row(s)
hive (h1b_project)>
```

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2015' group by job_title, full_time_position, year order by average desc;
```

```
hduser@ubuntu:~$ hive (hb_project)> select job_title,full_time position,year,avg(prevailing wage) as average from h1b_final where full_time position
on ='N' and year='2015' group by job_title,full time position,year order by average desc;
Query ID = hduser_20171013041801_68632cda-6a81-4fe8-a2a9-2cf4678802e1
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0044, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0044/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0044
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 04:18:08.880 Stage-1 map = 0%, reduce = 0%
2017-10-13 04:18:26.148 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 6.99 sec
2017-10-13 04:18:27.219 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 7.43 sec
2017-10-13 04:18:28.248 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 8.82 sec
2017-10-13 04:18:37.758 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 10.21 sec
2017-10-13 04:18:38.793 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 11.61 sec
MapReduce Total cumulative CPU time: 11 seconds 610 msec
Ended Job = job_1507706430268_0044
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0045, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0045/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0045
```

	PROJECT AND TECHNICAL OPERATIONS MANAGER	N	2015	21673.0
Adjuncy Faculty N	2015	21652.0		
PRESCHOOL TEACHERS N	2015	21632.0		
GALLERY AND PRINT ASSOCIATE N	2015	21320.0		
FINANCIAL CLERK N	2015	21257.0		
PIANO AND MUSIC HISTORY TEACHER N	2015	21049.0		
FITNESS CENTER ASSISTANT DIRECTOR N	2015	21008.0		
MASTER TEACHER N	2015	20966.0		
FOREIGN LANGUAGE TEACHER N	2015	20966.0		
OPERATIONS TECHNICIAN N	2015	20800.0		
ASSISTANT COACH FOR STATISTICS - WOMEN'S SPORTS N	2015	20800.0		
OPERATIONS TECHNICIAN N	2015	20800.0		
TUTOR N	2015	20633.0		
MUSIC TEACHER (CLASSICAL GUITAR) N	2015	20300.0		
SEO CLIENT STRATEGIST N	2015	20009.0		
PRESCHOOL TEACHER N	2015	19166.75		
PROGRAMS DIRECTOR N	2015	19094.0		
TEES RESEARCH ENGINEERING ASSOCIATE III N	2015	18803.0		
ENGINEERING EDUCATOR N	2015	18720.0		
DIRECTOR OF COLORADO ACADEMY FIELD HOCKEY N	2015	18678.0		
PHOTOJOURNALIST N	2015	18532.0		
INSTRUCTIONAL MENTOR/SPORT MANAGEMENT N	2015	18387.0		
PROFESSIONAL BOXER N	2015	18200.0		
CFO N	2015	17763.0		
CAREGIVER N	2015	17232.5		
ACCOUNT MARKETING RESEARCH N	2015	17056.0		
MEN'S SOCCER ASSISTANT COACH & RECRUITING COORDINATOR N	2015	16952.0		
SCOUT N	2015	16952.0		
URBAN PLANNING RESEARCH ASSISTANT N	2015	16681.0		
ICE SKATING COACH/INSTRUCTOR N	2015	15600.0		
GOLF PROGRAMS MANAGER N	2015	15680.0		
TEACHER AND INSTRUCTOR N	2015	12000.0		
		Time taken: 66.237 seconds, Fetched: 3839 row(s)		

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2016' group by job_title, full_time_position, year order by average desc;
```

```
hive (h1b_project)> select job_title,full_time_position,year,avg(prevailing_wage) as average from h1b_final where full_time_position = 'Y' and year='2016' group by job_title,full_time_position,year order by average desc;
Query ID = hduser_20171013041945_dd920a4c-dd7c-4f06-91dc-badc9d73bfd4
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0046, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0046/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0046
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 04:19:54,182 Stage-1 map = 0%, reduce = 0%
2017-10-13 04:20:13,675 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 7.81 sec
2017-10-13 04:20:14,726 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 9.37 sec
2017-10-13 04:20:15,749 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 10.67 sec
2017-10-13 04:20:27,329 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 12.45 sec
2017-10-13 04:20:28,389 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 14.56 sec
MapReduce Total cumulative CPU time: 14 seconds 560 msec
Ended Job = job_1507706430268_0046
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0047, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0047/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0047
```

Job Title	Full Time Position	Year	Average Prevailing Wage
PROGRAMMER ANALYST - SALESFORCE (JAVA) DEVELOPER	Y	2016	70034.0
ENGINEER, ANALYTICS	Y	2016	70034.0
PROJECT OFFICER	Y	2016	70034.0
SENIOR PROGRAMMER ANALYST DEVELOPER	Y	2016	70034.0
DEVELOPER USER INTERACE	Y	2016	70034.0
PROGRAMMER WEB-D	Y	2016	70034.0
DATABASE ADMINISTRATOR - I	Y	2016	70034.0
DESIGNER, WEB	Y	2016	70034.0
ASSOCIATE STATISTICAL PROGRAMMER	Y	2016	70034.0
RESEARCH CHEMICAL ENGINEER	Y	2016	70034.0
PERSONALIZATION DEVELOPER	Y	2016	70034.0
MARKETING COORDINATOR	Y	2016	70034.0
TECHNICAL LEAD (WEB DEVELOPMENT)	Y	2016	70034.0
SENIOR SYSTEM MANAGER	Y	2016	70034.0
SENIOR ACCOUNTANT, FINANCIAL REPORTING	Y	2016	70034.0
FIELD SERVICE & COMMISSIONING ENGINEER	Y	2016	70034.0
SENIOR PREDICTIVE MODELER (ANALYST IV, MODELING)	Y	2016	70013.0
BIOSTATISTICIAN III	Y	2016	70013.0
PROGRAMMER/COMPUTER SYSTEMS ANALYST	Y	2016	70013.0
ASSISTANT PROGRAM MANAGER	Y	2016	70013.0
PURCHASING SPECIALIST, RETAIL DEVELOPMENT	Y	2016	70013.0
PHARMACOVIGILANCE OPERATIONS ASSOCIATE	Y	2016	70013.0
DAM SOFTWARE DEVELOPER	Y	2016	70013.0
DEMAND PLANNER, CUSTOMER SUPPLY CHAIN	Y	2016	70013.0
AUDITOR	Y	2016	70013.0
AUDITOR - RISK ADVISORY SERVICES	Y	2016	70013.0
AUDITOR/SENIOR CONSULTANT - RISK ADVISORY SERVICES	Y	2016	70013.0
SOFTWARE ENGINEER I ("SE I")	Y	2016	70013.0
BUSINESS OBJECTS ADMINISTRATOR/DEVELOPER	Y	2016	70013.0
ENGINEER- EMBEDDED	Y	2016	70013.0
GLOBAL SUPPLY MANAGER, LOGISTICS MANAGEMENT	Y	2016	70013.0
SENIOR ASSOCIATE JC60 - OPERATIONS RESEARCH ANALYST	Y	2016	70013.0
EBU SENIOR FINANCIAL ANALYST	Y	2016	70013.0

Time taken: 72.096 seconds, Fetched: 53265 row(s)

hive (h1b\_project)>

```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2016' group by job_title, full_time_position, year order by average desc;
```

```
hduser@ubuntu:~$ Time taken: 72.096 seconds, Fetched: 53265 row(s)
hive (hib_project)> select job_title,full_time_position,year,avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2016' group by job_title,full_time_position,year order by average desc;
Query ID = hduser_20171013042136_f5190f4a-e23a-472d-9da5-c7014a2e9c55
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0048, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0048/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0048
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 04:21:45.266 Stage-1 map = 0%, reduce = 0%
2017-10-13 04:22:03.218 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 8.46 sec
2017-10-13 04:22:04.263 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 9.81 sec
2017-10-13 04:22:14.787 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 11.61 sec
2017-10-13 04:22:15.846 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 13.53 sec
MapReduce Total cumulative CPU time: 13 seconds 530 msec
Ended Job = job_1507706430268_0048
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0049, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0049/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0049
```

```
hduser@ubuntu:~$ ASSISTANT PROFESSOR - MECH/BIOMED ENGR N 2016 17060.0
CLINICAL ASSISTANT PROFESSOR - ENGINEERING N 2016 17060.0
ASSISTANT PROFESSOR - MATERIAL SCIENCE ENGINEERING N 2016 17060.0
ASSISTANT PROFESSOR - CIVIL ENGINEERING N 2016 17060.0
HEAD WOMENS SOCCER COACH N 2016 17050.0
DIRECTOR OF 10 AND UNDER TENNIS/ TENNIS PROFESSIONAL N 2016 17040.0
RESEARCH ASSOCIATE (POST-DOCCTORAL PSYCHOLOGY FELLOW) N 2016 16994.0
POT WASHER/KITCHEN SUPPORT N 2016 16993.0
HEAD OF YOUTH DEVELOPMENT N 2016 16950.0
CLASSROOM TEACHER - CHINESE IMMERSION N 2016 16930.0
NANNY N 2016 16925.0
CAREGIVER TO CHILDREN N 2016 16904.333333333332
DIRECTOR OF MEMBER CLUB SERVICES N 2016 16890.0
COOKS AND SERVICE N 2016 16889.0
ASSISTANT COACH FOR WOMEN'S BASKETBALL N 2016 16850.0
DIRECTOR OF COMMUNITY ACTIVITIES N 2016 16806.0
ARTS ADMINISTRATOR/ARTISTIC CONSULTANT N 2016 16765.0
CLUB HOCKEY COACH N 2016 16702.0
HUMAN RESOURCES MANAGERS N 2016 15756.0
TATTOO ARTIST APPRENTICE N 2016 15600.0
OFFICE ADMINISTRATOR AND SAFETY COORDINATOR N 2016 15080.0
NANNY/DOMESTIC HELP N 2016 15080.0
AUTOMOTIVE BODY AND RELATED REPAIRERS N 2016 15080.0
CHEF - PROFESSIONAL FOOD PRODUCTION N 2016 13575.333333333334
CO-EXECUTIVE DIRECTOR N 2016 12000.0
PLAYER N 2016 3900.0
SEAMSTRESS N 2016 2783.0
IOS AND ANDROID DEVELOPER N 2016 1278.0
CIS MANAGERS N 2016 0.0
E - COMMERCE MANAGER N 2016 0.0
FOREIGN RE INVESTMENT ADVISOR N 2016 0.0
LAW LIBRARY RESEARCHER N 2016 0.0
Time taken: 70.566 seconds, Fetched: 40837 row(s)
hive (hib_project)>
```

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

**Solution:**

```
hduser@ubuntu:~$ hadoop fs -cat /h1b/project/pig/question9/p*
INFOSYS LIMITED 99.54055 130592
ACCENTURE LLP 99.39307 33447
TATA CONSULTANCY SERVICES LIMITED 99.337204 64726
HCL AMERICA, INC. 99.26801 22678
RELIABLE SOFTWARE RESOURCES, INC. 99.14658 1992
NTT DATA, INC. 99.13251 4611
ERP ANALYSTS, INC. 99.10364 1785
PATNI AMERICAS INC. 99.07907 3149
KFORCE INC. 99.06015 1596
GENPACT LLC 98.852776 1046
SMARTPLAY, INC. 98.83805 1377
SYNTEL CONSULTING INC. 98.8317 3167
CREDIT SUISSE SECURITIES (USA) LLC 98.82168 2546
MASTECH, INC., A MASTECH HOLDINGS, INC. COMPANY 98.81408 5228
GENESIS ELDERCARE REHABILITATION SERVICES, INC. 98.78788 1320
HORIZON TECHNOLOGIES INC 98.78683 1731
SYNTEL INC 98.7667 1946
THE BOSTON CONSULTING GROUP, INC. 98.74261 1352
AMDOCS INC. 98.729225 1023
SAP AMERICA, INC. 98.69505 1456
DELOITTE TAX LLP 98.64054 2501
MPHASIS CORPORATION 98.63435 5199
3I INFOTECH, INC. 98.579124 2041
COMPUNNEL SOFTWARE GROUP, INC. 98.57904 3378
THE MATHWORKS, INC. 98.46535 2020
PERFICIENT, INC. 98.46266 1366
DALLAS INDEPENDENT SCHOOL DISTRICT 98.4384 1729
CGI TECHNOLOGIES AND SOLUTIONS INC. 98.39599 1995
VEDICSOFT 98.37468 1169
UNIVERSITY OF PITTSBURGH 98.34559 1632
DELOITTE CONSULTING LLP 98.32889 36742
BLOOMBERG, LP 98.29932 2352
WIPRO LIMITED 98.28959 48117
```

List of Petitioners			
HEADSTRONG SERVICES LLC	97.71937	2587	
SATYAM COMPUTER SERVICES LTD	97.718864	1622	
BIRLASOFT INC	97.67932	2370	
ERICSSON INC.	97.64811	3359	
UBER TECHNOLOGIES, INC.	97.61431	1006	
DOTCOM TEAM, LLC	97.6	1125	
APPLE INC.	97.59464	7317	
PHOTON INFOTECH, INC.	97.57085	1235	
UNIVERSITY OF UTAH	97.56782	1069	
CITIBANK, N.A.	97.560974	2173	
UNIVERSITY OF MINNESOTA	97.560974	1353	
TEXAS INSTRUMENTS INCORPORATED	97.52809	1780	
COMPUTER SCIENCES CORPORATION	97.52066	1089	
PRICEWATERHOUSECOOPERS ADVISORY SERVICES LLC	97.5058	1724	
MEMORIAL SLOAN-KETTERING CANCER CENTER	97.5	1080	
CAPITAL ONE SERVICES, LLC	97.49642	2796	
ORACLE AMERICA, INC.	97.48829	7684	
CSC COVANSYS CORPORATION	97.46779	2251	
CITIGROUP GLOBAL MARKETS INC.	97.4216	1435	
MICROEXCEL, INC	97.41156	1159	
SCHLUMBERGER TECHNOLOGY CORPORATION	97.402596	2310	
RITE AID CORP.	97.40012	1577	
ASTIR IT SOLUTIONS INC.	97.34016	1955	
SATYAM COMPUTER SERVICES LIMITED	97.33666	2403	
INTONE NETWORKS INC.	97.20635	1575	
COMCAST CABLE COMMUNICATIONS, LLC	97.19934	1214	
BANK OF AMERICA N.A.	97.19757	4282	
RJT COMPQUEST, INC.	97.17208	1662	
CHILDREN'S HOSPITAL CORPORATION	97.148476	1017	
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO	97.10683	1348	
VMWARE, INC.	97.019485	2617	
TESLA MOTORS, INC.	97.01596	1441	
PRICEWATERHOUSECOOPERS LLP	96.984184	2719	
HP ENTERPRISE SERVICES, LLC	96.95387	1149	

10) Which are the job positions along with the number of petitions which have the success rate more than 70% in petitions (total petitions filed 1000 OR more than1000)? (PIG)

**Solution:**

```

huser@ubuntu:~$ 2017-10-06 05:31:17,454 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to re
trieve job to compute warning aggregation.
2017-10-06 05:31:17,456 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
huser@ubuntu:~$ hadoop fs -cat /lib/project/pig/question10/p*
COMPUTER PROGRAMMER / CONFIGURER 2 100.0 1276
ASSOCIATE CONSULTANT - US 99.93171 4393
SYSTEMS ENGINEER - US 99.90036 10036
TEST ANALYST - US 99.818474 4958
CONSULTANT - US 99.81147 7426
TECHNOLOGY LEAD - US 99.80247 28350
TECHNICAL TEST LEAD - US 99.79531 5374
TECHNOLOGY ARCHITECT - US 99.766304 4707
TECHNOLOGY ANALYST - US 99.76204 26055
SENIOR PROJECT MANAGER - US 99.74766 2774
DEVELOPER USER INTERFACE 99.71412 5247
COMPUTER SYSTEMS ANALYST 2 99.70231 4031
SYSTEMS ANALYST - II 99.70127 1339
PROJECT MANAGER - III 99.69715 1651
PROJECT MANAGER - US 99.68777 7046
PROGRAMMER ANALYST - II 99.66555 3588
COMPUTER SYSTEMS ANALYST 3 99.58525 2170
COMPUTER PROGRAMMER/CONFIGURER 2 99.56903 6729
PROGRAMMER ANALYST - I 99.51117 1432
SYSTEMS ANALYST - III 99.50298 1006
COMPUTER SPECIALIST/TESTING AND QUALITY ANALYST 2 99.42471 3998
COMPUTER PROGRAMMER/CONFIGURER 3 99.38865 1145
COMPUTER SPECIALIST/SYSTEM SUPPORT AND DEVELOPMENT 99.32786 1339
COMPUTER SPECIALIST/SYSTEM SUPPORT AND DEVELOPMENT ADMIN 2 99.26267 1085
DATA WAREHOUSE SPECIALIST 99.20294 1631
ASSURANCE STAFF 99.05741 2334
COMPUTER SYSTEMS ENGINEER/ARCHITECT 98.79052 2067
SOFTWARE QUALITY ASSURANCE ENGINEER AND TESTER 98.66071 1568
AUDIT SENIOR 98.59813 1070
TEST CONSULTANT 98.55571 1454
COMPUTER SPECIALIST/TESTING AND QUALITY ANALYST 2 99.42471 3998
COMPUTER PROGRAMMER/CONFIGURER 3 99.38865 1145
COMPUTER SPECIALIST/SYSTEM SUPPORT AND DEVELOPMENT 99.32786 1339
COMPUTER SPECIALIST/SYSTEM SUPPORT AND DEVELOPMENT ADMIN 2 99.26267 1085
DATA WAREHOUSE SPECIALIST 99.20294 1631
ASSURANCE STAFF 99.05741 2334
COMPUTER SYSTEMS ENGINEER/ARCHITECT 98.79052 2067
SOFTWARE QUALITY ASSURANCE ENGINEER AND TESTER 98.66071 1568
AUDIT SENIOR 98.59813 1070
TEST CONSULTANT 98.55571 1454
SOFTWARE ENGINEER AND TESTER 98.51974 1216
ARCHITECT LEVEL 2 98.51314 2892
PROGRAMMER/DEVELOPER 98.46154 1560
TEST ENGINEER LEVEL 2 98.44013 2372
MODULE LEAD 98.33782 2226
ADVISORY MANAGER 98.310295 3255
AUDIT ASSISTANT 98.25726 1205
LEAD ENGINEER 98.23429 11157
COMPUTER SPECIALIST 98.206894 2175
CONSULTANT LEVEL 3 98.12126 1171
DEVELOPER 98.00914 12909
ERS SENIOR CONSULTANT 97.95464 2249
TAX SENIOR 97.93253 1838
TEST ENGINEER LEVEL 1 97.87645 1036
PROGRAMMER ANALYST LEVEL 1 97.87057 2395
TECHNICAL ANALYST 97.78308 2932
SOFTWARE DEVELOPMENT ENGINEER IN TEST 97.65148 4258
ADVISORY SENIOR ASSOCIATE 97.52252 1332
SOFTWARE ENGINEER 2 97.16755 4166
ERS CONSULTANT 97.14286 2170
FUNCTIONAL CONSULTANT 97.04036 1115
QA TESTER 96.92308 1170
SOFTWARE DEVELOPMENT ENGINEER 96.856125 7284
PROGRAMMER ANALYSTS 96.82259 1133

```

---

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

## Solution:

```
hduser@ubuntu:~$ nano ~/.bashrc
hduser@ubuntu:~$ sqoop export --connect jdbc:mysql://localhost/h1b project --username root --password root --table question11 --update-mode allowinsert --export-dir /h1b project/pig/question10/p* --input-fields-terminated-by '\t';
Warning: /usr/local/sqoop/.:/accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/.:/zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Sat Oct 14 07:41:41 PDT 2017 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
Note: /tmp/sqoop-hduser/compile/5d39154bb8d47ca2a699a979e5e1f14/question11.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j.impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
hduser@ubuntu:~$ mysql -u root -p -e 'select * from h1b_project.question11';
Enter password:
+-----+-----+-----+
| job_title | success_rate | petitions |
+-----+-----+-----+
| POSTDOCTORAL FELLOW | 94.8581 | 7857 |
| SOFTWARE DEVELOPERS, APPLICATIONS | 92.9707 | 1195 |
| COMPUTER PROGRAMMER / CONFIGURER 2 | 100 | 1276 |
| RESEARCH FELLOW | 96.3551 | 5981 |
| SENIOR HARDWARE ENGINEER | 94.8578 | 1653 |
| QA ENGINEER | 94.8291 | 2224 |
| ASSOCIATE CONSULTANT - US | 99.9317 | 4393 |
| ENGINEER II | 94.7158 | 1249 |
| APPLICATIONS DEVELOPER | 96.3458 | 3366 |
| SR. SOFTWARE ENGINEER | 94.7152 | 4863 |
+-----+-----+-----+
```

```
hduser@ubuntu:~$ mysql -u root -p -e 'select * from h1b project.question11';
Enter password:
+-----+-----+-----+
| job_title | success_rate | petitions |
+-----+-----+-----+
| POSTDOCTORAL FELLOW | 94.8581 | 7857 |
| SOFTWARE DEVELOPERS, APPLICATIONS | 92.9707 | 1195 |
| COMPUTER PROGRAMMER / CONFIGURER 2 | 100 | 1276 |
| RESEARCH FELLOW | 96.3551 | 5981 |
| SENIOR HARDWARE ENGINEER | 94.8578 | 1653 |
| QA ENGINEER | 94.8291 | 2224 |
| ASSOCIATE CONSULTANT - US | 99.9317 | 4393 |
| ENGINEER II | 94.7158 | 1249 |
| APPLICATIONS DEVELOPER | 96.3458 | 3366 |
| SR. SOFTWARE ENGINEER | 94.7152 | 4863 |
| COMPUTER SYSTEM ANALYST | 92.7525 | 3753 |
| SENIOR SOFTWARE DEVELOPER | 96.3166 | 10208 |
| PROGRAMMER ANALYST | 96.1279 | 249038 |
| SENIOR PROGRAMMER ANALYST | 96.1274 | 5810 |
| ASSISTANT RESEARCH SCIENTIST | 96.1015 | 1103 |
| SENIOR ASSOCIATE | 96.0117 | 3540 |
| SOFTWARE QUALITY ASSURANCE ENGINEER | 95.9756 | 4920 |
| SENIOR MANAGER | 95.9694 | 1439 |
| PHYSICIAN IN A POST GRADUATE TRAINING PROGRAM | 95.9521 | 2421 |
| SYSTEMS ANALYST | 95.9477 | 61965 |
| QUALITY ASSURANCE ANALYST | 95.9459 | 7326 |
| TECHNICAL ARCHITECT | 95.9422 | 2998 |
| PROJECT LEAD | 95.9374 | 2363 |
| SOFTWARE ENGINEER III | 95.9337 | 1328 |
| SOFTWARE ANALYST | 95.8955 | 1072 |
| SR. SYSTEMS ANALYST | 95.8297 | 1151 |
| SOFTWARE ENGINEER 3 | 95.8223 | 1891 |
| LEAD DEVELOPER | 95.8055 | 1049 |
| QUALITY ANALYST | 95.7951 | 2616 |
+-----+-----+-----+
```

SENIOR DEVELOPER	93.4536	2994		
PHARMACIST	93.4004	5864		
INSTRUCTOR	93.3975	3014		
SENIOR SYSTEMS ANALYST JC60	93.3855	3069		
SENIOR ENGINEER	93.374	3773		
PRODUCT ENGINEER	93.2422	2634		
ENGINEER	93.0581	4941		
NETWORK AND COMPUTER SYSTEMS ADMINISTRATOR	93.0498	1928		
LEAD SOFTWARE ENGINEER	93.0025	1572		
OCCUPATIONAL THERAPIST	92.7203	4437		
SENIOR APPLICATION DEVELOPER	92.6209	1965		
SENIOR PROJECT MANAGER	92.6188	1015		
CLINICAL FELLOW	92.5829	1146		
INDUSTRIAL DESIGNER	92.567	3619		
PROJECT ENGINEER	92.561	6439		
PHYSICAL THERAPIST	92.5372	20207		
APPLICATIONS ENGINEER	92.5355	1688		
SOLUTION ARCHITECT	92.5276	1994		
HOSPITALIST	92.5066	4387		
TEST LEAD	92.4102	1726		
DIRECTOR	92.3481	1333		
HOSPITALIST PHYSICIAN	92.3285	4067		
SENIOR FINANCIAL ANALYST	92.2241	1196		
SYSTEM ENGINEER	92.2145	2145		
MANAGER JC50	91.889	1874		
MEDICAL RESIDENT	91.7808	2336		
PSYCHIATRIST	91.5438	1289		
COMPUTER SOFTWARE ENGINEER	91.5052	2684		
SCIENTIST	91.4925	1340		
RESIDENT	91.4859	1245		
APPLICATION DEVELOPER	91.4587	7692		
TECHNICAL PROJECT MANAGER	91.2548	1052		
PROCESS ENGINEER	91.2269	4377		
PHYSICIAN	91.0573	4417		

## CONCLUSION

Therefore, from the given dataset of H1-B Applicants within the years 2011-2016, which contained around 3 billion/30 lakh records, we have done a complete analysis on various factors and criteria which has some predictive nature. This predictive analysis on the H1-B visa dataset has revolved around finding top Occupation, States, Employers and Industries that contribute to highest number of H1-B visa applications. This clearly indicates the trends of H1-B Visa filings which has a high correlation with the employer's acceptance rate.