**A PROJECT ON H1B\_ANALYSIS**

**Submitted by**

**Prajwala Shetty**

**Project Overview**

**H1B\_Visa Analysis:**

The H1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for H1B visa, an US employer must offer a job and petition for H1B visa with the US immigration department.. We will be performing analysis on the H1B visa applicants between the years 2011-2016. After analyzing the data, we can derive the following facts.

Various Technologies like Map Reduce, Pig and Hive are used to perform operations on these data and the values are yielded.

**What is big data?**

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. It is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks.

### Characteristics of 'Big Data'

**(1)Volume –** The name 'Big Data' itself is related to a size which is enormous. Size of data plays very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with 'Big Data'.

**(2)Variety –** The next aspect of 'Big Data' is its **variety**.

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Now days, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. is also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analysing data.

**(3)Velocity –** The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

**(4)Variability –** This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

## **What Comes Under Big Data?**

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

* **Black Box Data**: It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
* **Social Media Data**: Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
* **Stock Exchange Data**: The stock exchange data holds information about the ‘buy’ and ‘sell’ decisions made on a share of different companies made by the customers.
* **Power Grid Data**: The power grid data holds information consumed by a particular node with respect to a base station.
* **Transport Data**: Transport data includes model, capacity, distance and availability of a vehicle.
* **Search Engine Data**: Search engines retrieve lots of data from different databases.

The data in it will be of three types.

* **Structured data**: Relational data.
* **Semi Structured data**: XML data.
* **Unstructured data**: Word, PDF, Text, Media Logs.

## **Benefits of Big Data**

Big data is really critical to our life and its emerging as one of the most important technologies in modern world. Follow are just few benefits which are very much known to all of us:

* Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.
* Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.
* Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

## **Big Data Challenges**

The major challenges associated with big data are as follows:

* Capturing data
* Curation
* Storage
* Searching
* Sharing
* Transfer
* Analysis
* Presentation

To fulfill the above challenges, organizations normally take the help of enterprise servers.

**Hadoop**

**Hadoop** is an open-source software framework for storing **data** and running applications on clusters of commodity hardware. It provides massive storage for any kind of **data**, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

## **Hadoop Architecture**

Hadoop framework includes following four modules:

* **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules. These libraries provides file system and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.
* **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.
* **Hadoop Distributed File System (HDFS):** A distributed file system that provides high-throughput access to application data.
* **Hadoop Map Reduce:** This is YARN-based system for parallel processing of large data sets.

**Hadoop Distributed File System (HDFS)**

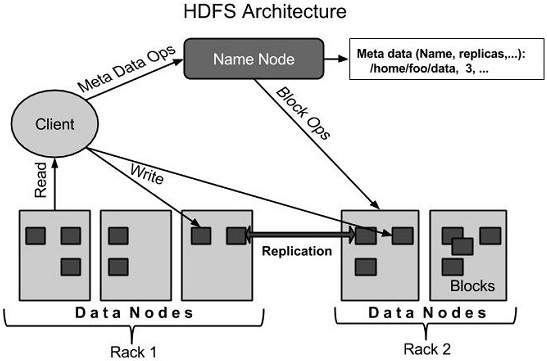
Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware.

## **Features of HDFS**

* It is suitable for the distributed storage and processing.
* Hadoop provides a command interface to interact with HDFS.
* The built-in servers of name node and data node help users to easily check the status of cluster.
* Streaming access to file system data.
* HDFS provides file permissions and authentication.

## **HDFS Architecture**

Given below is the architecture of a Hadoop File System.

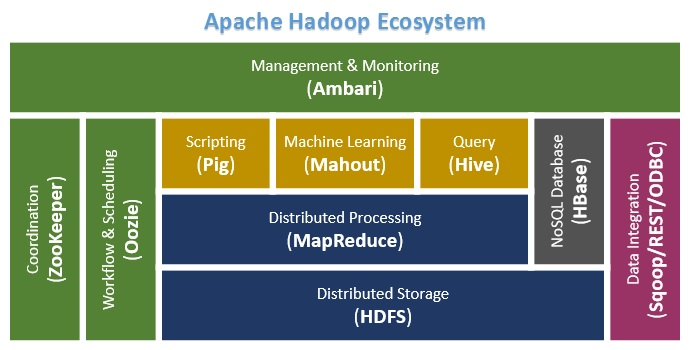


HDFS follows the master-slave architecture and it has the following elements.

## **Goals of HDFS**

* **Fault detection and recovery**: Since HDFS includes a large number of commodity hardware, failure of components is frequent. Therefore HDFS should have mechanisms for quick and automatic fault detection and recovery.
* **Huge datasets**: HDFS should have hundreds of nodes per cluster to manage the applications having huge datasets.
* **Hardware at data**: A requested task can be done efficiently, when the computation takes place near the data. Especially where huge datasets are involved, it reduces the network traffic and increases the throughput.

**Apache Hadoop Ecosystem**



**Map Reduce**

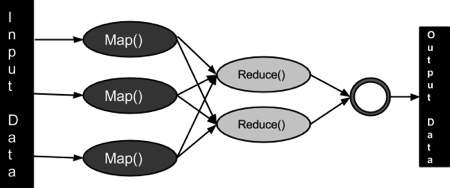
**Map Reduce** is a [programming model](https://en.wikipedia.org/wiki/Programming_model) and an associated implementation for processing and generating [big data](https://en.wikipedia.org/wiki/Big_data) sets with a [parallel](https://en.wikipedia.org/wiki/Parallel_computing), [distributed](https://en.wikipedia.org/wiki/Distributed_computing) algorithm on a [cluster](https://en.wikipedia.org/wiki/Cluster_(computing)).

It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster.

The Map Reduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map Reduce implies, the reduce task is always performed after the map job.

## **The Algorithm**

* Generally Map Reduce paradigm is based on sending the computer to where the data resides!
* Map Reduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.
  + **Map stage**: The map or mapper’s job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
  + **Reduce stage**: This stage is the combination of the **Shuffle**stage and the **Reduce** stage. The Reducer’s job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.
* During a Map Reduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
* The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.
* Most of the computing takes place on nodes with data on local disks that reduces the network traffic.
* After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.



**Hive**

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analysing easy. **Apache Hive**, allows SQL developers to write Hive Query Language (HQL) statements that are similar to standard SQL statements; now you should be aware that HQL is limited in the commands it understands, but it is still pretty useful. HQL statements are broken down by the Hive service into hive jobs and executed across a Hadoop cluster.

**ADVANTAGES OF USING APACHE HIVE**

1. Fits the low level interface requirement of Hadoop perfectly.

2. Supports external tables which make it possible to process data without actually storing in HDFS.

3. It has a rule based optimizer for optimizing logical plans.

4. Supports partitioning of data at the level of tables to improve performance.

5. Meta store or Metadata store is a big plus in the architecture which makes the lookup easy.

**DISADVANTAGES OF USING APACHE HIVE**

1. No support for update and delete.

2. No support for singleton inserts. Data is required to be loaded from a file using LOAD command.

3. No access control implementation.

4. Correlated sub queries are not supported.

**Apache Pig**

**Apache Pig** is a platform for analysing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

**Advantages**

* **Ease of programming.** It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.
* **Optimization opportunities.** The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.
* **Extensibility.** Users can create their own functions to do special-purpose processing.

**Sqoop**

Apache Sqoop efficiently transfers bulk data between Apache Hadoop and structured data stores such as relational databases. Sqoop helps offload certain tasks (such as ETL processing) from the EDW to Hadoop for efficient execution at a much lower cost. Sqoop can also be used to extract data from Hadoop and export it into external structured data stores. 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.

**Flume**

**Apache Flume** is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log data, events (etc...) from various web serves to a centralized data store.

**Oozie**

Apache Oozie is a server-based [workflow](https://en.wikipedia.org/wiki/Workflow) [scheduling](https://en.wikipedia.org/wiki/Scheduling_(computing)) system to manage [Hadoop](https://en.wikipedia.org/wiki/Apache_Hadoop" \o "Apache Hadoop) jobs.

Workflows in Oozie are defined as a collection of control flow and action [nodes](https://en.wikipedia.org/wiki/Vertex_(graph_theory)) in a [directed acyclic graph](https://en.wikipedia.org/wiki/Directed_acyclic_graph). Control flow nodes define the beginning and the end of a workflow (start, end, and failure nodes) as well as a mechanism to control the workflow execution path (decision, fork, and join nodes). Action nodes are the mechanism by which a workflow triggers the execution of a computation/processing task. Oozie provides support for different types of actions including Hadoop [Map Reduce](https://en.wikipedia.org/wiki/MapReduce), Hadoop distributed file system operations, [Pig](https://en.wikipedia.org/wiki/Pig_(programming_tool)), [SSH](https://en.wikipedia.org/wiki/Secure_Shell), and [email](https://en.wikipedia.org/wiki/Email).

**Zookeeper**

Zookeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications. Each time they are implemented there is a lot of work that goes into fixing the bugs and race conditions that are inevitable. Because of the difficulty of implementing these kinds of services, applications initially usually skimp on them, which make them brittle in the presence of change and difficult to manage. Even when done correctly, different implementations of these services lead to management complexity when the applications are deployed.

**PROJECT DESCRIPTION**

**H1B\_Visa Analysis:**

The H1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for H1B visa, an US employer must offer a job and petition for H1B visa with the US immigration department.. We will be performing analysis on the H1B visa applicants between the years 2011-2016. After analyzing the data, we can derive the following facts.

The goal of the project is get insights into **H-1B** applications over the year 2011 to 2016 by performing analysis using different technologies like MapReduce,Pig,Hive

Dataset column description.

1. **CASE\_STATUS:**Status associated with the last significant event or decision. Valid values include.
   * "Certified": Employer filed the LCA, which was approved by DOL
   * "Certified-Withdrawn": LCA was approved but later withdrawn by employer
   * "Denied": LCA was denied by DOL
   * "Withdrawn": LCA was withdrawn by employer before
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.FULL\_TIME\_POSITION:
   * **Y** = Full Time Position.
   * **N** = Part Time Position.
5. **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.
6. **YEAR:** Year in which the H1B visa petition was filed.
7. **WORKSITE:** City and State information of the foreign worker’s intended area of employment.
8. **lon:** longitude of the Worksite.
9. **lat:** latitude of the Worksite.

#### Questions that were answered after performing analysis.

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

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

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

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

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

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

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

a) For all the applications

b) For only certified applications.

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

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

8) find the average Prevailing Wage for each Job for each Year (take part time and full time separate). Arrange the output in descending order.

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

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

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

Solutions:

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

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class Question1a{

public static class MyMapClass extends Mapper<LongWritable,Text,Text,IntWritable>{

public void map(LongWritable key,Text value,Context con) throws IOException, InterruptedException{

String[] str = value.toString().split("\t");

String year = str[7];

String job = str[4];

int count = 1;

if(job.equals("DATA ENGINEER")){

con.write(new Text(year),new IntWritable(count));

}

}

}

public static class MyReduceClass extends Reducer<Text,IntWritable,Text,Text>{

int year = 0;

String growth = "N.A";

String output = " ";

public void reduce(Text key,Iterable<IntWritable> values,Context context) throws IOException, InterruptedException{

int count = 0;

double growthpercentage= 0;

for (IntWritable val:values){

count += val.get();

}

if(year!=0){

growthpercentage = ((double)(count-year)/year)\*100;

String.format("%.2f%%", growthpercentage);

}

output = String.format("%d",count)+","+growthpercentage;

year = count;

context.write(key,new Text(output));

}

}

public static void main(String[] args) throws IOException, ClassNotFoundException, InterruptedException {

Configuration conf = new Configuration();

Job job = Job.getInstance(conf, "Question1a");

job.setJarByClass(Question1a.class);

job.setMapperClass(MyMapClass.class);

job.setReducerClass(MyReduceClass.class);

job.setOutputKeyClass(Text.class);

//job.setNumReduceTasks(0);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(IntWritable.class);

job.setOutputValueClass(Text.class);

FileInputFormat.addInputPath(job, new Path(args[0]));

FileOutputFormat.setOutputPath(job, new Path(args[1]));

System.exit(job.waitForCompletion(true) ? 0 : 1);

}

}

**Output:**

2011 18, 0.0

2012 32, 77.77777777777779

2013 41, 28.125

2014 89, 117.07317073170731

2015 160, 79.7752808988764

2016 251, 56.875

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

|  |
| --- |
| h1b = load '/user/hive/warehouse/h1b\_final' using PigStorage() as (s\_n\_no:int,case\_status:chararray,emp\_name:chararray,soc\_name:chararray,job\_title:chararray,full\_time:chararray,prevailing\_wage:int,year:chararray,worksite:chararray,longitude:int,latitude:int);  a1 = filter h1b by $7=='2011';  b1 = group a1 by $4;  c1 = foreach b1 generate group,COUNT($1);  a2 = filter h1b by $7=='2012';  b2 = group a2 by $4;  c2 = foreach b2 generate group,COUNT($1);  a3 = filter h1b by $7=='2013';  b3 = group a3 by $4;  c3 = foreach b3 generate group,COUNT($1);  a4 = filter h1b by $7=='2014';  b4 = group a4 by $4;  c4 = foreach b4 generate group,COUNT($1);  a5 = filter h1b by $7=='2015';  b5 = group a5 by $4;  c5 = foreach b5 generate group,COUNT($1);    a6 = filter h1b by $7=='2016';  b6 = group a6 by $4;  c6 = foreach b6 generate group,COUNT($1);    combine = join c1 by $0, c2 by $0, c3 by $0, c4 by $0, c5 by $0, c6 by $0;  job\_year\_count = foreach combine generate $0,$1,$3,$5,$7,$9,$11;  growth\_percent = foreach job\_year\_count generate $0, (float)(($6-$5)\*100)/$5,(float)(($5-$4)\*100)/$4, (float)(($4-$3)\*100)/$3, (float)(($3-$2)\*100)/$2, (float)(($2-$1)\*100)/$1;  avg\_growth\_percent = foreach growth\_percent generate $0,(($1+$2+$3+$4+$5)/5);  order\_agp = order avg\_growth\_percent by $1 desc;  top5 = limit order\_agp 5;  dump top5; |

(SENIOR SYSTEMS ANALYST JC60,4255.4644)

(SOFTWARE DEVELOPER 2,3480.5925)

(PROJECT MANAGER 3,3233.3335)

(SYSTEMS ANALYST JC65,2984.8809)

(MODULE LEAD,2917.112)

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

select worksite, case\_status,count(job\_title) as jobs from h1b\_final where job\_title like '%DATA ENGINEER%' and case\_status='CERTIFIED' and year='2011' group by worksite,year,case\_status order by jobs desc limit 1;

SEATTLE, WASHINGTON CERTIFIED 19

select worksite, case\_status,count(job\_title) as jobs from h1b\_final where job\_title like '%DATA ENGINEER%' and case\_status='CERTIFIED' and year='2012' group by worksite,year,case\_status order by jobs desc limit 1;

SEATTLE, WASHINGTON CERTIFIED 26

select worksite, case\_status,count(job\_title) as jobs from h1b\_final where job\_title like '%DATA ENGINEER%' and case\_status='CERTIFIED' and year='2013' group by worksite,year,case\_status order by jobs desc limit 1;

SEATTLE, WASHINGTON CERTIFIED 43

select worksite, case\_status,count(job\_title) as jobs from h1b\_final where job\_title like '%DATA ENGINEER%' and case\_status='CERTIFIED' and year='2014' group by worksite,year,case\_status order by jobs desc limit 1;

SEATTLE, WASHINGTON CERTIFIED 42

select worksite, case\_status,count(job\_title) as jobs from h1b\_final where job\_title like '%DATA ENGINEER%' and case\_status='CERTIFIED' and year='2015' group by worksite,year,case\_status order by jobs desc limit 1;

SEATTLE, WASHINGTON CERTIFIED 60

select worksite, case\_status,count(job\_title) as jobs from h1b\_final where job\_title like '%DATA ENGINEER%' and case\_status='CERTIFIED' and year='2016' group by worksite,year,case\_status order by jobs desc limit 1;

SEATTLE, WASHINGTON CERTIFIED 121

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

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

output:

year worksite status

2011 NEW YORK, NEW YORK 23172

2011 HOUSTON, TEXAS 8184

2011 CHICAGO, ILLINOIS 5188

2011 SAN JOSE, CALIFORNIA 4713

2011 SAN FRANCISCO, CALIFORNIA 4711

select year, worksite,count(case\_status) as status from h1b\_final where case\_status='CERTIFIED' and year='2012' group by worksite,year order by status desc limit 5;

year worksite status

2012 NEW YORK, NEW YORK 23737

2012 HOUSTON, TEXAS 9963

2012 SAN FRANCISCO, CALIFORNIA 6116

2012 CHICAGO, ILLINOIS 5671

2012 ATLANTA, GEORGIA 5565

select year, worksite,count(case\_status) as status from h1b\_final where case\_status='CERTIFIED' and year='2013' group by worksite,year order by status desc limit 5;

year worksite status

2013 NEW YORK, NEW YORK 23537

2013 HOUSTON, TEXAS 11136

2013 SAN FRANCISCO, CALIFORNIA 7281

2013 SAN JOSE, CALIFORNIA 6722

2013 ATLANTA, GEORGIA 6377

select year, worksite,count(case\_status) as status from h1b\_final where case\_status='CERTIFIED' and year='2014' group by worksite,year order by status desc limit 5;

year worksite status

2014 NEW YORK, NEW YORK 27634

2014 HOUSTON, TEXAS 13360

2014 SAN FRANCISCO, CALIFORNIA 9798

2014 SAN JOSE, CALIFORNIA 8223

2014 ATLANTA, GEORGIA 8213

select year, worksite,count(case\_status) as status from h1b\_final where case\_status='CERTIFIED' and year='2015' group by worksite,year order by status desc limit 5;

year worksite status

2015 NEW YORK, NEW YORK 31266

2015 HOUSTON, TEXAS 15242

2015 SAN FRANCISCO, CALIFORNIA 12594

2015 ATLANTA, GEORGIA 10500

2015 SAN JOSE, CALIFORNIA 9589

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

year worksite status

2016 NEW YORK, NEW YORK 34639

2016 SAN FRANCISCO, CALIFORNIA 13836

2016 HOUSTON, TEXAS 13655

2016 ATLANTA, GEORGIA 11678

2016 CHICAGO, ILLINOIS 11064

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

Aimport java.io.IOException;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class Question3 {

public static class MyMapper extends Mapper<LongWritable,Text,Text,IntWritable>

{

public void map(LongWritable key, Text value, Context con) throws IOException, InterruptedException

{

String str[] = value.toString().split("\t");

String soc\_name = str[3];

String job = str[4];

int count = 0;

if(job.contains("DATA SCIENTIST"));

{

count++;

}

con.write(new Text(soc\_name), new IntWritable(count));

}

}

public static class MyReducer extends Reducer<Text,IntWritable,Text,IntWritable>

{

String output = " ";

int most\_pos = 0;

public void reduce(Text key, Iterable <IntWritable>values,Context con)

{

int count = 0;

for(IntWritable val:values)

{

count+= val.get();

if(most\_pos<count);

{

most\_pos = count;

output = key.toString();

}

}

}

public void cleanup(Context con) throws IOException, InterruptedException

{

con.write(new Text(output), new IntWritable(most\_pos));

}

}

public static void main(String[] args) throws IOException, ClassNotFoundException, InterruptedException

{

Configuration conf = new Configuration();

Job job = Job.getInstance(conf,"max data scientist position");

job.setJarByClass(Question3.class);

job.setMapperClass(MyMapper.class);

job.setReducerClass(MyReducer.class);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(IntWritable.class);

job.setOutputKeyClass(Text.class);

job.setOutputKeyClass(IntWritable.class);

FileInputFormat.addInputPath(job,new Path(args[0]));

FileOutputFormat.setOutputPath(job,new Path(args[1]));

System.exit(job.waitForCompletion(true)? 0 : 1);

}

}

Output:

STATISTICIANS 649

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

import java.io.IOException;

import java.util.TreeMap;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.NullWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Mapper.Context;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Partitioner;

import org.apache.hadoop.mapreduce.Reducer;

public class Question4 {

public static class MyMapper extends Mapper<LongWritable,Text,Text,LongWritable>

{

LongWritable one = new LongWritable(1);

public void map(LongWritable key, Text value, Context context) throws IOException,

InterruptedException {

if (key.get() > 0)

{

String[] token = value.toString().split("\t");

if (!token[1].equals("NA") && token[1] != null && !token[2].equals("NA") && token[2] != null && !token[7].equals("NA") && token[7] != null) {

Text answer = new Text(token[2].replaceAll("\"", "") + "\t" + token[7]);

context.write(answer, one);

}

}

}

}

public static class MyPartitioner extends Partitioner < Text, LongWritable > {

@Override

public int getPartition(Text key, LongWritable value, int numReduceTasks) {

String[] str = key.toString().split("\t");

if (str[1].equals("2011"))

return 0;

if (str[1].equals("2012"))

return 1;

if (str[1].equals("2013"))

return 2;

if (str[1].equals("2014"))

return 3;

if (str[1].equals("2015"))

return 4;

if (str[1].equals("2016"))

return 5;

else

return 6;

}

}

public static class MyReducer extends Reducer< Text, LongWritable, NullWritable, Text > {

private TreeMap < LongWritable,Text > Top5Employers = new TreeMap < LongWritable,Text > ();

long sum = 0;

public void reduce(Text key, Iterable < LongWritable > values, Context context) throws IOException,

InterruptedException {

sum = 0;

for (LongWritable val: values) {

sum += val.get();

}

Top5Employers.put(new LongWritable(sum), new Text(key + "," + sum));

if (Top5Employers.size() > 5)

Top5Employers.remove(Top5Employers.firstKey());

}

protected void cleanup(Context context) throws IOException,

InterruptedException {

for (Text t: Top5Employers.descendingMap().values())

context.write(NullWritable.get(), t);

}

}

public static void main(String args[]) throws IOException, InterruptedException, ClassNotFoundException

{

Configuration conf = new Configuration();

Job job = Job.getInstance(conf, "Top 5 Employers");

job.setJarByClass(Question4.class);

job.setMapperClass(MyMapper.class);

job.setPartitionerClass(MyPartitioner.class);

job.setReducerClass(MyReducer.class);

job.setNumReduceTasks(7);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(LongWritable.class);

job.setOutputKeyClass(NullWritable.class);

job.setOutputValueClass(Text.class);

FileInputFormat.addInputPath(job, new Path(args[0]));

FileOutputFormat.setOutputPath(job, new Path(args[1]));

System.exit(job.waitForCompletion(true) ? 0 : 1);

}

}

Output

(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)

(WIPRO LIMITED,2012,7182)

(TATA CONSULTANCY SERVICES LIMITED,2012,6735)

(DELOITTE CONSULTING LLP,2012,4727)

(IBM INDIA PRIVATE LIMITED,2012,4074)

(INFOSYS LIMITED,2012,15818)

(ACCENTURE LLP,2013,4994)

(DELOITTE CONSULTING LLP,2013,6124)

(WIPRO LIMITED,2013,6734)

(TATA CONSULTANCY SERVICES LIMITED,2013,8790)

(INFOSYS LIMITED,2013,32223)

(ACCENTURE LLP,2014,5498)

(DELOITTE CONSULTING LLP,2014,7017)

(WIPRO LIMITED,2014,8365)

(TATA CONSULTANCY SERVICES LIMITED,2014,14098)

(INFOSYS LIMITED,2014,23759)

(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)

(TATA CONSULTANCY SERVICES LIMITED,2016,13134)

(WIPRO LIMITED,2016,10607)

(IBM INDIA PRIVATE LIMITED,2016,9787)

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.

Ans5a) select year,job\_title,count(case\_status) as popular from h1b\_final where year ='2011' group by year,job\_title order by popular desc limit 10;

2011 PROGRAMMER ANALYST 31799

2011 SOFTWARE ENGINEER 12763

2011 COMPUTER PROGRAMMER 8998

2011 SYSTEMS ANALYST 8644

2011 BUSINESS ANALYST 3891

2011 COMPUTER SYSTEMS ANALYST 3698

2011 ASSISTANT PROFESSOR 3467

2011 PHYSICAL THERAPIST 3377

2011 SENIOR SOFTWARE ENGINEER 2935

2011 SENIOR CONSULTANT 2798

select year,job\_title,count(case\_status) as popular from h1b\_final where year ='2012' group by year,job\_title order by popular desc limit 10;

2012 PROGRAMMER ANALYST 33066

2012 SOFTWARE ENGINEER 14437

2012 COMPUTER PROGRAMMER 9629

2012 SYSTEMS ANALYST 9296

2012 BUSINESS ANALYST 4752

2012 COMPUTER SYSTEMS ANALYST 4706

2012 SOFTWARE DEVELOPER 3895

2012 PHYSICAL THERAPIST 3871

2012 ASSISTANT PROFESSOR 3801

2012 SENIOR CONSULTANT 3737

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

2013 PROGRAMMER ANALYST 33880

2013 SOFTWARE ENGINEER 15680

2013 COMPUTER PROGRAMMER 11271

2013 SYSTEMS ANALYST 8714

2013 TECHNOLOGY LEAD - US 7853

2013 TECHNOLOGY ANALYST - US 7683

2013 BUSINESS ANALYST 5716

2013 COMPUTER SYSTEMS ANALYST 5043

2013 SOFTWARE DEVELOPER 5026

2013 SENIOR CONSULTANT 4326

select year,job\_title,count(case\_status) as popular from h1b\_final where year ='2014' group by year,job\_title order by popular desc limit 10;

2014 PROGRAMMER ANALYST 43114

2014 SOFTWARE ENGINEER 20500

2014 COMPUTER PROGRAMMER 14950

2014 SYSTEMS ANALYST 10194

2014 SOFTWARE DEVELOPER 7337

2014 BUSINESS ANALYST 7302

2014 COMPUTER SYSTEMS ANALYST 6821

2014 TECHNOLOGY LEAD - US 5057

2014 TECHNOLOGY ANALYST - US 4913

2014 SENIOR CONSULTANT 4898

select year,job\_title,count(case\_status) as popular from h1b\_final where year ='2015' group by year,job\_title order by popular desc limit 10;

2015 PROGRAMMER ANALYST 53436

2015 SOFTWARE ENGINEER 27259

2015 COMPUTER PROGRAMMER 14054

2015 SYSTEMS ANALYST 12803

2015 SOFTWARE DEVELOPER 10441

2015 BUSINESS ANALYST 8853

2015 TECHNOLOGY LEAD - US 8242

2015 COMPUTER SYSTEMS ANALYST 7918

2015 TECHNOLOGY ANALYST - US 7014

2015 SENIOR SOFTWARE ENGINEER 6013

select year,job\_title,count(case\_status) as popular from h1b\_final where year ='2016' group by year,job\_title order by popular desc limit 10;

2016 PROGRAMMER ANALYST 53743

2016 SOFTWARE ENGINEER 30668

2016 SOFTWARE DEVELOPER 14041

2016 SYSTEMS ANALYST 12314

2016 COMPUTER PROGRAMMER 11668

2016 BUSINESS ANALYST 9167

2016 COMPUTER SYSTEMS ANALYST 6900

2016 SENIOR SOFTWARE ENGINEER 6439

2016 DEVELOPER 6084

2016 TECHNOLOGY LEAD - US 5410

5b)

select year,job\_title,case\_status,count(case\_status) as popular from h1b\_final where year ='2011' and case\_status= 'CERTIFIED'group by year,job\_title,case\_status order by popular desc limit 10;

2011 PROGRAMMER ANALYST CERTIFIED 28806

2011 SOFTWARE ENGINEER CERTIFIED 11224

2011 COMPUTER PROGRAMMER CERTIFIED 8038

2011 SYSTEMS ANALYST CERTIFIED 7850

2011 BUSINESS ANALYST CERTIFIED 3444

2011 COMPUTER SYSTEMS ANALYST CERTIFIED 3152

2011 ASSISTANT PROFESSOR CERTIFIED 3050

2011 PHYSICAL THERAPIST CERTIFIED 2911

2011 SENIOR SOFTWARE ENGINEER CERTIFIED 2595

2011 SENIOR CONSULTANT CERTIFIED 2585

select year,job\_title,case\_status,count(case\_status) as popular from h1b\_final where year ='2012' and case\_status= 'CERTIFIED'group by year,job\_title,case\_status order by popular desc limit 10;

2012 PROGRAMMER ANALYST CERTIFIED 29226

2012 SOFTWARE ENGINEER CERTIFIED 12273

2012 COMPUTER PROGRAMMER CERTIFIED 8483

2012 SYSTEMS ANALYST CERTIFIED 8399

2012 BUSINESS ANALYST CERTIFIED 4144

2012 COMPUTER SYSTEMS ANALYST CERTIFIED 4084

2012 SENIOR CONSULTANT CERTIFIED 3420

2012 SOFTWARE DEVELOPER CERTIFIED 3290

2012 PHYSICAL THERAPIST CERTIFIED 3284

2012 ASSISTANT PROFESSOR CERTIFIED 3033

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

2013 PROGRAMMER ANALYST CERTIFIED 29906

2013 SOFTWARE ENGINEER CERTIFIED 12973

2013 COMPUTER PROGRAMMER CERTIFIED 10202

2013 SYSTEMS ANALYST CERTIFIED 7850

2013 TECHNOLOGY LEAD - US CERTIFIED 7809

2013 TECHNOLOGY ANALYST - US CERTIFIED 7641

2013 BUSINESS ANALYST CERTIFIED 4993

2013 COMPUTER SYSTEMS ANALYST CERTIFIED 4554

2013 SOFTWARE DEVELOPER CERTIFIED 4316

2013 SENIOR CONSULTANT CERTIFIED 3996

select year,job\_title,case\_status,count(case\_status) as popular from h1b\_final where year ='2014' and case\_status= 'CERTIFIED'group by year,job\_title,case\_status order by popular desc limit 10;

2014 PROGRAMMER ANALYST CERTIFIED 38625

2014 SOFTWARE ENGINEER CERTIFIED 17278

2014 COMPUTER PROGRAMMER CERTIFIED 13796

2014 SYSTEMS ANALYST CERTIFIED 9161

2014 BUSINESS ANALYST CERTIFIED 6529

2014 SOFTWARE DEVELOPER CERTIFIED 6473

2014 COMPUTER SYSTEMS ANALYST CERTIFIED 6204

2014 TECHNOLOGY LEAD - US CERTIFIED 5055

2014 TECHNOLOGY ANALYST - US CERTIFIED 4911

2014 SENIOR CONSULTANT CERTIFIED 4535

select year,job\_title,case\_status,count(case\_status) as popular from h1b\_final where year ='2015' and case\_status= 'CERTIFIED'group by year,job\_title,case\_status order by popular desc limit 10;

2015 PROGRAMMER ANALYST CERTIFIED 48203

2015 SOFTWARE ENGINEER CERTIFIED 23352

2015 COMPUTER PROGRAMMER CERTIFIED 12971

2015 SYSTEMS ANALYST CERTIFIED 11498

2015 SOFTWARE DEVELOPER CERTIFIED 9343

2015 TECHNOLOGY LEAD - US CERTIFIED 8238

2015 BUSINESS ANALYST CERTIFIED 7919

2015 COMPUTER SYSTEMS ANALYST CERTIFIED 7234

2015 TECHNOLOGY ANALYST - US CERTIFIED 7009

2015 SENIOR SOFTWARE ENGINEER CERTIFIED 5324

select year,job\_title,case\_status,count(case\_status) as popular from h1b\_final where year ='2016' and case\_status= 'CERTIFIED'group by year,job\_title,case\_status order by popular desc limit 10;

2016 PROGRAMMER ANALYST CERTIFIED 47964

2016 SOFTWARE ENGINEER CERTIFIED 25890

2016 SOFTWARE DEVELOPER CERTIFIED 12474

2016 SYSTEMS ANALYST CERTIFIED 10986

2016 COMPUTER PROGRAMMER CERTIFIED 10528

2016 BUSINESS ANALYST CERTIFIED 8175

2016 COMPUTER SYSTEMS ANALYST CERTIFIED 6205

2016 DEVELOPER CERTIFIED 5912

2016 SENIOR SOFTWARE ENGINEER CERTIFIED 5630

2016 TECHNOLOGY LEAD - US CERTIFIED 5405

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.

create table if not exists total(total int,year string)

row format delimited

fields terminated by ',';

insert overwrite table total select count(\*),year from h1b\_final where h1b\_final.case\_status is not NULL group by year;

select a.case\_status,count(\*) as case\_total,a.year,ROUND((count(\*)/b.total)\*100,2) as perofcase\_status from h1b\_final a left outer join total b on (a.year=b.year) where a.year is not NULL group by a.case\_status,b.total,a.year order by a.year;

WITHDRAWN 10105 2011 2.82

DENIED 29130 2011 8.12

CERTIFIED 307936 2011 85.83

CERTIFIED-WITHDRAWN 11596 2011 3.23

WITHDRAWN 10725 2012 2.58

DENIED 21096 2012 5.08

CERTIFIED-WITHDRAWN 31118 2012 7.49

CERTIFIED 352668 2012 84.86

WITHDRAWN 11590 2013 2.62

CERTIFIED 382951 2013 86.62

CERTIFIED-WITHDRAWN 35432 2013 8.01

DENIED 12141 2013 2.75

CERTIFIED 455144 2014 87.62

CERTIFIED-WITHDRAWN 36350 2014 7.0

WITHDRAWN 16034 2014 3.09

DENIED 11899 2014 2.29

WITHDRAWN 19455 2015 3.14

CERTIFIED-WITHDRAWN 41071 2015 6.64

DENIED 10923 2015 1.77

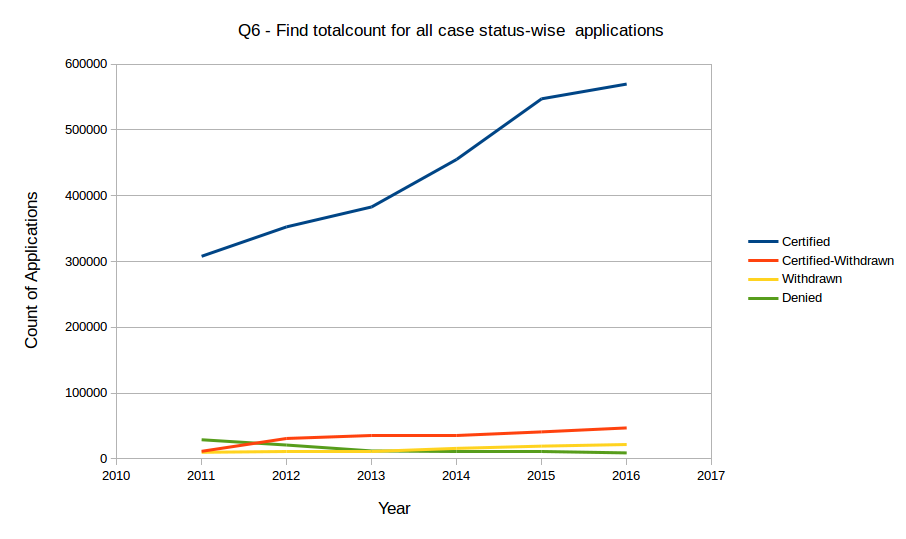
CERTIFIED 547278 2015 88.45

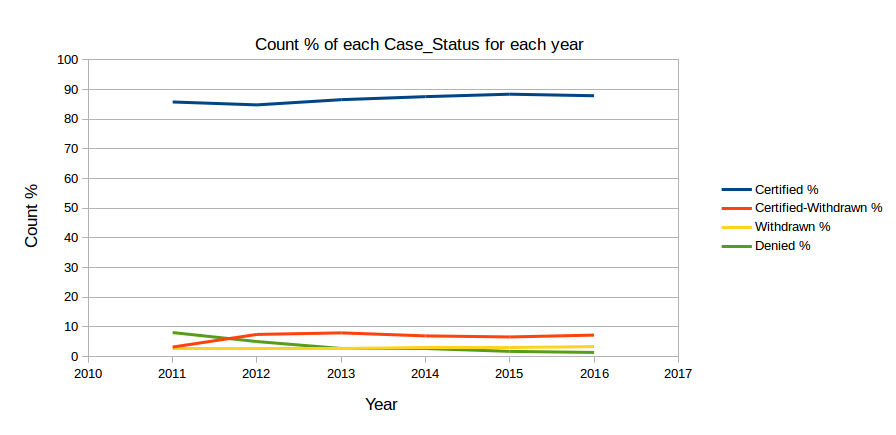
CERTIFIED 569646 2016 87.94

DENIED 9175 2016 1.42

WITHDRAWN 21890 2016 3.38

CERTIFIED-WITHDRAWN 47092 2016 7.27





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

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class Question7 {

public static class MyMapper extends Mapper<LongWritable,Text,Text,Text>

{

public void map(LongWritable key,Text value,Context con) throws IOException, InterruptedException

{

String str[] = value.toString().split("\t");

String year = str[7];

String case\_status = str[6];

con.write(new Text(year), new Text(value));

}

}

public static class MyReducer extends Reducer<Text,Text,Text,IntWritable>

{

public void reduce(Text key,Iterable<Text> values,Context con) throws IOException, InterruptedException

{

int count = 0;

for(Text val: values)

{

count++;

}

con.write(key, new IntWritable(count));

}

}

public static void main(String[] args) throws IOException, ClassNotFoundException, InterruptedException

{

Configuration conf = new Configuration();

Job job = Job.getInstance(conf,"Total applications per year");

job.setJarByClass(Question7.class);

job.setMapperClass(MyMapper.class);

job.setReducerClass(MyReducer.class);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(Text.class);

job.setOutputKeyClass(Text.class);

job.setOutputKeyClass(IntWritable.class);

FileInputFormat.addInputPath(job, new Path(args[0]));

FileOutputFormat.setOutputPath(job, new Path(args[1]));

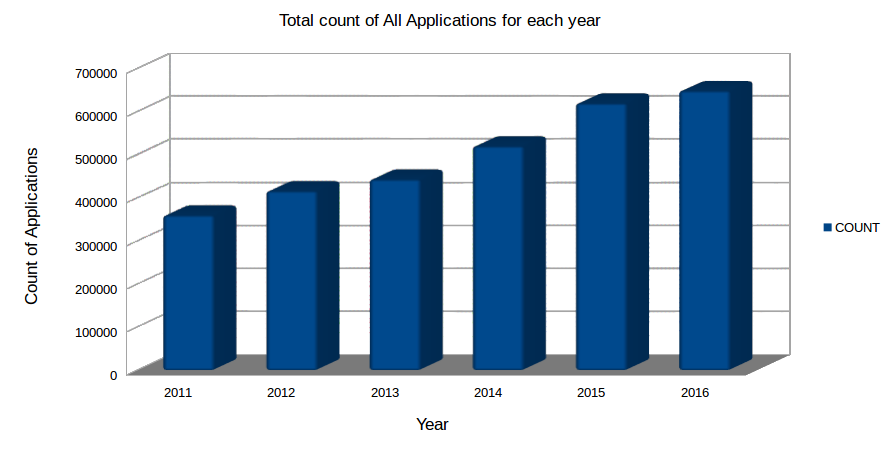
System.exit(job.waitForCompletion(true) ? 0:1);

}

}

Output:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |



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.

FullTime:

h1b = load '/user/hive/warehouse/h1b\_final' using PigStorage('\t') as (s\_no:int, case\_status:chararray,employer\_name:chararray,soc\_name:chararray,job\_tittle:chararray,full\_time\_position:chararray,prevailing\_wage:int,year:chararray,worksite:chararray,longitude:double,latitude:double);

--h1b1 = filter h1b by $1 != 'CASE\_STATUS';

h1b2 = filter h1b by $1=='CERTIFIED';

-- for fulltime

h1b3 = filter h1b2 by $5 == 'Y';

h1b\_a = foreach h1b3 generate $4,$7,$6,$5;

h1b\_2011 = filter h1b\_a by $1=='2011';

h1b\_2012 = filter h1b\_a by $1=='2012';

h1b\_2013 = filter h1b\_a by $1=='2013';

h1b\_2014 = filter h1b\_a by $1=='2014';

h1b\_2015 = filter h1b\_a by $1=='2015';

h1b\_2016 = filter h1b\_a by $1=='2016';

h1b\_groupa = group h1b\_2011 by ($0,$1,$3);

h1b\_groupb = group h1b\_2012 by ($0,$1,$3);

h1b\_groupc = group h1b\_2013 by ($0,$1,$3);

h1b\_groupd = group h1b\_2014 by ($0,$1,$3);

h1b\_groupe = group h1b\_2015 by ($0,$1,$3);

h1b\_groupf = group h1b\_2016 by ($0,$1,$3);

h1b\_counta = foreach h1b\_groupa generate group, ROUND\_TO(AVG(h1b\_2011.$2),2);

h1b\_countb = foreach h1b\_groupb generate group, ROUND\_TO(AVG(h1b\_2012.$2),2);

h1b\_countc = foreach h1b\_groupc generate group, ROUND\_TO(AVG(h1b\_2013.$2),2);

h1b\_countd = foreach h1b\_groupd generate group, ROUND\_TO(AVG(h1b\_2014.$2),2);

h1b\_counte = foreach h1b\_groupe generate group, ROUND\_TO(AVG(h1b\_2015.$2),2);

h1b\_countf = foreach h1b\_groupf generate group, ROUND\_TO(AVG(h1b\_2016.$2),2);

h1b\_all = UNION h1b\_counta,h1b\_countb,h1b\_countc,h1b\_countd,h1b\_counte,h1b\_countf;

h1b\_order = order h1b\_all by $0 desc;

dump h1b\_order;

Part time:

h1b = load '/user/hive/warehouse/h1b\_final' using PigStorage('\t') as (s\_no:int, case\_status:chararray,employer\_name:chararray,soc\_name:chararray,job\_tittle:chararray,full\_time\_position:chararray,prevailing\_wage:int,year:chararray,worksite:chararray,longitude:double,latitude:double);

h1b1 = filter h1b by $1 != 'CASE\_STATUS';

h1b2 = filter h1b1 by $1=='CERTIFIED';

h1b3 = filter h1b2 by $5 == 'N';

h1b\_a = foreach h1b2 generate $4,$7,$6,$5;

h1b\_2011 = filter h1b\_a by $1=='2011';

h1b\_2012 = filter h1b\_a by $1=='2012';

h1b\_2013 = filter h1b\_a by $1=='2013';

h1b\_2014 = filter h1b\_a by $1=='2014';

h1b\_2015 = filter h1b\_a by $1=='2015';

h1b\_2016 = filter h1b\_a by $1=='2016';

h1b\_groupa = group h1b\_2011 by ($0,$1,$3);

h1b\_groupb = group h1b\_2012 by ($0,$1,$3);

h1b\_groupc = group h1b\_2013 by ($0,$1,$3);

h1b\_groupd = group h1b\_2014 by ($0,$1,$3);

h1b\_groupe = group h1b\_2015 by ($0,$1,$3);

h1b\_groupf = group h1b\_2016 by ($0,$1,$3);

h1b\_counta = foreach h1b\_groupa generate group, ROUND\_TO(AVG(h1b\_2011.$2),2);

h1b\_countb = foreach h1b\_groupb generate group, ROUND\_TO(AVG(h1b\_2012.$2),2);

h1b\_countc = foreach h1b\_groupc generate group, ROUND\_TO(AVG(h1b\_2013.$2),2);

h1b\_countd = foreach h1b\_groupd generate group, ROUND\_TO(AVG(h1b\_2014.$2),2);

h1b\_counte = foreach h1b\_groupe generate group, ROUND\_TO(AVG(h1b\_2015.$2),2);

h1b\_countf = foreach h1b\_groupf generate group, ROUND\_TO(AVG(h1b\_2016.$2),2);

h1b\_all = UNION h1b\_counta,h1b\_countb,h1b\_countc,h1b\_countd,h1b\_counte,h1b\_countf;

h1b\_order = order h1b\_all by $0 desc;

dump h1b\_order;

---union of those

--h1b\_overall = union h1b\_all,h1b\_overall;

--dump h1b\_overall;

Output(ft)

((ANALYST, CUSTOMER CONTRACTS OPERATIONS ADMINISTRATOR,2015,Y),47632.0)

((ANALYST, CUSTOMER ANALYTICS (SENIOR),2015,Y),79061.0)

((ANALYST, CUSTOMER ANALYTICS,2015,Y),56708.0)

((ANALYST, CUSTOMER ANALYTICS,2014,Y),56380.2)

((ANALYST, CUSTOMER ANALYTICS,2013,Y),65530.25)

((ANALYST, CUSTOMER ANALYTICS,2011,Y),64189.0)

((ANALYST, CUSTOM RESEARCH,2013,Y),59738.0)

((ANALYST, CRM, TECHNOLOGY OPERATIONS,2015,Y),80995.0)

((ANALYST, CRM, APPLICATIONS,2015,Y),80995.0)

((ANALYST, CRM BUSINESS SYSTEMS,2015,Y),61797.0)

((ANALYST, CRM,2012,Y),37960.0)

((ANALYST, CREDIT/TREASURY,2015,Y),64438.0)

((ANALYST, CREDIT/TREASURY,2012,Y),59987.0)

((ANALYST, CREDIT TRADING,2014,Y),67205.0)

((ANALYST, CREDIT RISK PORTFOLIO,2012,Y),62566.0)

((ANALYST, CREDIT RISK,2013,Y),42598.0)

((ANALYST, CREDIT RESEARCH,2015,Y),65187.0)

((ANALYST, CREDIT POLICY,2014,Y),122845.0)

((ANALYST, CPO REFERENCE DATA,2011,Y),59613.0)

((ANALYST, CPM,2014,Y),50794.0)

((ANALYST, CPG NORTH AMERICA,2015,Y),52333.0)

((ANALYST, CORPORATE STRATEGY & GOVERNMENT RELATIONS,2015,Y),40976.0)

((ANALYST, CORPORATE STRATEGY & DEVELOPMENT,2012,Y),64002.0)

((ANALYST, CORPORATE STRATEGY,2015,Y),39790.0)

((ANALYST, CORPORATE HEOR, CENTER FOR EXCELLENCE,2015,Y),82638.0)

((ANALYST, CORPORATE HEOR, CENTER FOR EXCELLENCE,2014,Y),82659.0)

((ANALYST, CORPORATE FINANCE,2012,Y),86278.0)

((ANALYST, CORPORATE FINANCE,2011,Y),61942.0)

((ANALYST, CORPORATE DEVELOPMENT,2014,Y),67205.0)

((ANALYST, CORPORATE DEVELOPMENT,2012,Y),65936.0)

((ANALYST, CORPORATE CREDIT,2011,Y),59363.0)

((ANALYST, CORPORATE CLIENT SOLUTIONS,2015,Y),66019.0)

((ANALYST, CORPORATE CLIENT SOLUTIONS,2014,Y),66152.5)

((ANALYST, CORPORATE BANKING,2014,Y),66019.0)

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

(INFOSYS LIMITED,130592,99.54055)

(ACCENTURE LLP,33447,99.39307)

(TATA CONSULTANCY SERVICES LIMITED,64726,99.337204)

(HCL AMERICA, INC.,22678,99.26801)

(RELIABLE SOFTWARE RESOURCES, INC.,1992,99.14658)

(NTT DATA, INC.,4611,99.13251)

(ERP ANALYSTS, INC.,1785,99.10364)

(PATNI AMERICAS INC.,3149,99.07907)

(KFORCE INC.,1596,99.06015)

(GENPACT LLC,1046,98.852776)

(SMARTPLAY, INC.,1377,98.83805)

(SYNTEL CONSULTING INC.,3167,98.8317)

(CREDIT SUISSE SECURITIES (USA) LLC,2546,98.82168)

(MASTECH, INC., A MASTECH HOLDINGS, INC. COMPANY,5228,98.81408)

(GENESIS ELDERCARE REHABILITATION SERVICES, INC.,1320,98.78788)

(HORIZON TECHNOLOGIES INC,1731,98.78683)

(SYNTEL INC,1946,98.7667)

(THE BOSTON CONSULTING GROUP, INC.,1352,98.74261)

(AMDOCS INC.,1023,98.729225)

(SAP AMERICA, INC.,1456,98.69505)

(DELOITTE TAX LLP,2501,98.64054)

(MPHASIS CORPORATION,5199,98.63435)

(3I INFOTECH, INC.,2041,98.579124)

(COMPUNNEL SOFTWARE GROUP, INC.,3378,98.57904)

(THE MATHWORKS, INC.,2020,98.46535)

(PERFICIENT, INC.,1366,98.46266)

(DALLAS INDEPENDENT SCHOOL DISTRICT,1729,98.4384)

(CGI TECHNOLOGIES AND SOLUTIONS INC.,1995,98.39599)

(VEDICSOFT,1169,98.37468)

(UNIVERSITY OF PITTSBURGH,1632,98.34559)

(DELOITTE CONSULTING LLP,36742,98.32889)

(BLOOMBERG, LP,2352,98.29932)

(WIPRO LIMITED,48117,98.28959)

(MCKINSEY & COMPANY, INC. UNITED STATES,1097,98.268005)

(MICHIGAN STATE UNIVERSITY,1191,98.23678)

(NVIDIA CORPORATION,1182,98.22335)

(ALINDUS, INC.,1046,98.183556)

(PRICEWATERHOUSECOOPERS, LLP,2529,98.141556)

(CVS RX SERVICES, INC.,2735,98.135284)

(PYRAMID TECHNOLOGY SOLUTIONS, INC,1056,98.106064)

(MICROSOFT CORPORATION,25576,98.091965)

(CYBERTHINK INC,1618,98.08405)

(BARCLAYS SERVICES CORP.,1605,98.068535)

(HARVARD UNIVERSITY,1966,98.06714)

(INOVANT, LLC,1086,98.0663)

(ERNST & YOUNG U.S. LLP,18232,98.05287)

(DELOITTE & TOUCHE LLP,9642,97.99834)

(CAPGEMINI AMERICA INC,16725,97.955154)

(TECH MAHINDRA (AMERICAS) INC.,2102,97.95433)

(SUNERA TECHNOLOGIES, INC,1440,97.916664)

(SRS CONSULTING INC.,1150,97.91304)

(LARSEN & TOUBRO TECHNOLOGY SERVICES LTD,1385,97.83394)

(MARLABS, INC,2626,97.8294)

(CYIENT, INC.,1281,97.81421)

(MERRILL LYNCH,1873,97.811)

(NATSOFT CORPORATION,1137,97.80123)

(WASHINGTON UNIVERSITY IN ST. LOUIS,1576,97.77919)

(KPMG LLP,4629,97.774895)

(THE UNIV. OF ALA. AT BIRMINGHAM (UAB),1288,97.74844)

(APEX TECHNOLOGY SYSTEMS, INC,1060,97.73585)

(LINKEDIN CORPORATION,2194,97.721054)

(HEADSTRONG SERVICES LLC,2587,97.71937)

(SATYAM COMPUTER SERVICES LTD,1622,97.718864)

(BIRLASOFT INC,2370,97.67932)

(ERICSSON INC.,3359,97.64811)

(UBER TECHNOLOGIES, INC.,1006,97.61431)

(DOTCOM TEAM, LLC,1125,97.6)

(APPLE INC.,7317,97.59464)

(PHOTON INFOTECH, INC.,1235,97.57085)

(UNIVERSITY OF UTAH,1069,97.56782)

(CITIBANK, N.A.,2173,97.560974)

(UNIVERSITY OF MINNESOTA,1353,97.560974)

(TEXAS INSTRUMENTS INCORPORATED,1780,97.52809)

(COMPUTER SCIENCES CORPORATION,1089,97.52066)

(PRICEWATERHOUSECOOPERS ADVISORY SERVICES LLC,1724,97.5058)

(MEMORIAL SLOAN-KETTERING CANCER CENTER,1080,97.5)

(CAPITAL ONE SERVICES, LLC,2796,97.49642)

(ORACLE AMERICA, INC.,7684,97.48829)

(CSC COVANSYS CORPORATION,2251,97.46779)

(CITIGROUP GLOBAL MARKETS INC.,1435,97.4216)

(MICROEXCEL, INC,1159,97.41156)

(SCHLUMBERGER TECHNOLOGY CORPORATION,2310,97.402596)

(RITE AID CORP.,1577,97.40012)

(ASTIR IT SOLUTIONS INC.,1955,97.34016)

(SATYAM COMPUTER SERVICES LIMITED,2403,97.33666)

(INTONE NETWORKS INC.,1575,97.20635)

(COMCAST CABLE COMMUNICATIONS, LLC,1214,97.19934)

(BANK OF AMERICA N.A.,4282,97.19757)

(RJT COMPUQUEST, INC.,1662,97.17208)

(CHILDREN'S HOSPITAL CORPORATION,1017,97.148476)

(UNIVERSITY OF CALIFORNIA, SAN FRANCISCO,1348,97.10683)

(VMWARE, INC.,2617,97.019485)

(TESLA MOTORS, INC.,1441,97.01596)

(PRICEWATERHOUSECOOPERS LLP,2719,96.984184)

(HP ENTERPRISE SERVICES, LLC,1149,96.95387)

(AMERICAN EXPRESS TRAVEL RELATED SERVICES COMPANY, INC.,1045,96.9378)

(UNIVERSITY OF MICHIGAN,2893,96.92361)

(TWITTER, INC.,1328,96.91265)

(JPMORGAN CHASE & CO.,7035,96.87278)

(HCL GLOBAL SYSTEMS INC,3677,96.87245)

(IDHASOFT, INC.,1423,96.83767)

(CAPGEMINI FINANCIAL SERVICES USA INC,4426,96.81428)

(CHARTER GLOBAL, INC.,1188,96.801346)

(RANDSTAD TECHNOLOGIES, LP,3419,96.782684)

(SMARTSOFT INTERNATIONAL, INC.,1212,96.78218)

(VEDICSOFT SOLUTIONS LLC,1274,96.78179)

(V-SOFT CONSULTING GROUP, INC,4283,96.75461)

(UNIVERSITY OF ILLINOIS,1196,96.73913)

(THE UNIVERSITY OF CHICAGO,1277,96.711044)

(TECHDEMOCRACY LLC,1027,96.689384)

(VERIZON DATA SERVICES LLC,1635,96.636086)

(EMC CORPORATION,4467,96.61965)

(IDEXCEL, INC.,1360,96.617645)

(GOOGLE INC.,16473,96.59443)

(BLACKROCK FINANCIAL MANAGEMENT, INC.,1048,96.46947)

(VERINON TECHNOLOGY SOLUTIONS LTD.,1245,96.465866)

(HEWLETT-PACKARD COMPANY,1639,96.46126)

(EBAY INC.,3464,96.44919)

(IGATE TECHNOLOGIES INC.,12564,96.442215)

(LARSEN & TOUBRO INFOTECH LIMITED,17457,96.43123)

(T-MOBILE USA, INC.,1457,96.43102)

(UNIVERSITY OF WISCONSIN-MADISON,1115,96.41256)

(GENERAL HOSPITAL CORPORATION,1429,96.36109)

(SYMANTEC CORPORATION,2290,96.33188)

(PAYPAL, INC.,2830,96.32509)

(FACEBOOK, INC.,4145,96.30881)

(COLUMBIA UNIVERSITY,1841,96.30636)

(SAGARSOFT, INC,1082,96.30314)

(PURDUE UNIVERSITY,1076,96.282524)

(SALESFORCE.COM, INC.,2245,96.258354)

(NIIT TECHNOLOGIES LIMITED,1339,96.191185)

(PROKARMA, INC.,1333,96.17404)

(TRUSTEES OF THE UNIVERSITY OF PENNSYLVANIA,2258,96.102745)

(EXPERIS US, INC.,1641,96.09994)

(NEW YORK UNIVERSITY SCHOOL OF MEDICINE,1126,96.09236)

(CERNER CORPORATION,2268,96.075836)

(MASSACHUSETTS INSTITUTE OF TECHNOLOGY,1347,96.06533)

(JOHNS HOPKINS UNIVERSITY,1823,96.05047)

(LEAD IT CORPORATION,1720,96.04651)

(WAL-MART ASSOCIATES, INC.,3670,96.0218)

(UNIVERSITY OF CALIFORNIA, LOS ANGELES,1172,95.98976)

(UST GLOBAL INC.,6363,95.929596)

(THE UNIVERSITY OF IOWA,1569,95.92097)

(DEUTSCHE BANK SECURITIES INC.,1170,95.89744)

(POLARIS SOFTWARE LAB (INDIA) LTD.,1326,95.85219)

(A2Z DEVELOPMENT CENTER, INC.,1025,95.80488)

(EXILANT TECHNOLOGIES PRIVATE LIMITED,1572,95.80153)

(NATIONAL INSTITUTES OF HEALTH, HHS,2327,95.7456)

(THE OHIO STATE UNIVERSITY,1587,95.71519)

(TECHNOSOFT CORPORATION,1625,95.69231)

(SYNECHRON, INC.,3802,95.68648)

(SEARS HOLDINGS MANAGEMENT CORPORATION,1105,95.656105)

(INTRAEDGE, INC.,1254,95.61404)

(ADVENT GLOBAL SOLUTIONS INC.,1048,95.51527)

(HEXAWARE TECHNOLOGIES, INC.,5466,95.49945)

(YAHOO! INC.,3348,95.45998)

(SATYAM COMPUTER SERVICES LTD.,1694,95.33648)

(CMC AMERICAS, INC.,1157,95.332756)

(AMAZON CORPORATE LLC,9026,95.29138)

(DELL MARKETING L.P.,1532,95.234985)

(EXPEDIA, INC.,1311,95.19451)

(CYMA SYSTEMS INC,1269,95.19306)

(MULTIVISION INC.,1502,95.00666)

(UNIVERSITY OF CALIFORNIA, DAVIS,1334,94.97751)

(MOUNT SINAI MEDICAL CENTER,1114,94.97307)

(SYSTEM SOFT TECHNOLOGIES LLC,3102,94.970985)

(CIBER, INC.,2097,94.94516)

(CISCO SYSTEMS, INC.,3140,94.90446)

(THE UNIVERSITY OF TEXAS AT AUSTIN,1274,94.89796)

(HOWARD HUGHES MEDICAL INSTITUTE,1135,94.88987)

(QUALCOMM TECHNOLOGIES, INC.,6113,94.87976)

(BRIGHAM AND WOMEN'S HOSPITAL,1117,94.80752)

(ADOBE SYSTEMS INCORPORATED,1167,94.77292)

(AVANT HEALTHCARE PROFESSIONALS,1006,94.73161)

(UST GLOBAL INC,6355,94.72856)

(POPULUS GROUP,2635,94.72486)

(ORACLE FINANCIAL SERVICES SOFTWARE, INC.,1532,94.64752)

(BROOKHAVEN NATIONAL LABORATORY,1023,94.62366)

(SAPIENT CORPORATION,2237,94.59097)

(AT&T SERVICES, INC.,1201,94.587845)

(LARSEN & TOUBRO LIMITED,3066,94.58578)

(UNIVERSITY OF FLORIDA,1429,94.54164)

(HITACHI CONSULTING CORPORATION,2854,94.49895)

(MAYO CLINIC,1772,94.46953)

(KPIT INFOSYSTEMS, INC.,3114,94.44444)

(QUALCOMM INCORPORATED,3965,94.42623)

(FUJITSU AMERICA, INC.,5309,94.34922)

(ITECH US, INC.,2476,94.34572)

(YALE UNIVERSITY,1852,94.33045)

(UNIVERSITY OF ILLINOIS AT CHICAGO,1131,94.25288)

(UNIVERSITY OF COLORADO,1599,94.12132)

(MANAGEMENT HEALTH SYSTEMS, INC.,2000,94.1)

(UNIVERSITY OF WASHINGTON,1187,94.01853)

(ECLINICALWORKS, LLC,1547,93.92372)

(GLOBALFOUNDRIES U.S. INC.,1391,93.8174)

(TECH MAHINDRA (AMERICAS), INC.,7019,93.78829)

(NORTHWESTERN UNIVERSITY,1439,93.46769)

(CAPGEMINI U.S. LLC,3712,93.3459)

(ORION SYSTEMS INTEGRATORS, INC,1160,93.27586)

(UNIVERSITY OF CALIFORNIA, SAN DIEGO,1202,93.26123)

(SOFTWARE PARADIGMS INTERNATIONAL GROUP, LLC,1034,93.23017)

(DUKE UNIVERSITY AND MEDICAL CENTER,1330,93.1579)

(GOLDMAN, SACHS & CO.,3713,93.10531)

(EMORY UNIVERSITY,1680,93.09524)

(BAYLOR COLLEGE OF MEDICINE,1666,93.03722)

(MICRON TECHNOLOGY, INC.,1934,93.019646)

(QUALCOMM ATHEROS, INC.,1274,93.01413)

(AKAMAI TECHNOLOGIES, INC.,1092,92.948715)

(MORGAN STANLEY & CO. LLC,1669,92.9299)

(UNIVERSITY OF MARYLAND COLLEGE PARK,1354,92.61448)

(INTEL CORPORATION,11415,92.562416)

(BATTELLE,1052,92.49049)

(COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION,17528,92.486305)

(INTUIT INC.,1404,92.45014)

(ITC INFOTECH (USA), INC.,1859,91.931145)

(MEDTRONIC, INC.,1050,91.90476)

(BALTIMORE CITY PUBLIC SCHOOLS,1014,91.51874)

(SIRI INFOSOLUTIONS INC.,1039,91.33782)

(THE PENNSYLVANIA STATE UNIVERSITY,1042,91.17082)

(ADVANCED MICRO DEVICES, INC.,1512,91.137566)

(HSBC BANK USA, N.A.,1110,90.99099)

(L&T TECHNOLOGY SERVICES LIMITED,3722,90.56959)

(MARVELL SEMICONDUCTOR, INC.,1631,89.94482)

(CUMMINS INC.,4737,89.82478)

(IBM INDIA PVT. LTD.,1284,89.563866)

(VIRTUSA CORPORATION,2217,88.76861)

(INDIANA UNIV. PURDUE UNIV. INDIANAPOLIS,1007,88.57994)

(THE UNIVERSITY OF ARIZONA,1037,88.331726)

(IBM CORPORATION,13276,88.19675)

(BROADCOM CORPORATION,2862,88.12019)

(IBM INDIA PRIVATE LIMITED,34219,87.6852)

(PERSISTENT SYSTEMS, INC.,3225,87.06977)

(AMERICAN INFORMATION TECHNOLOGY CORPORATION,1358,86.67158)

(CITRIX SYSTEMS, INC.,1044,85.344826)

(NETAPP, INC.,1870,84.545456)

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

COMPUTER PROGRAMMER / CONFIGURER 2,1276,100.0

ASSOCIATE CONSULTANT - US,4393,99.93171

SYSTEMS ENGINEER - US,10036,99.90036

TEST ANALYST - US,4958,99.818474

CONSULTANT - US,7426,99.81147

TECHNOLOGY LEAD - US,28350,99.80247

TECHNICAL TEST LEAD - US,5374,99.79531

TECHNOLOGY ARCHITECT - US,4707,99.766304

TECHNOLOGY ANALYST - US,26055,99.76204

SENIOR PROJECT MANAGER - US,2774,99.74766

DEVELOPER USER INTERFACE,5247,99.71412

COMPUTER SYSTEMS ANALYST 2,4031,99.70231

SYSTEMS ANALYST - II,1339,99.70127

PROJECT MANAGER - III,1651,99.69715

PROJECT MANAGER - US,7046,99.68777

PROGRAMMER ANALYST - II,3588,99.66555

COMPUTER SYSTEMS ANALYST 3,2170,99.58525

COMPUTER PROGRAMMER/CONFIGURER 2,6729,99.56903

PROGRAMMER ANALYST - I,1432,99.51117

SYSTEMS ANALYST - III,1006,99.50298

COMPUTER SPECIALIST/TESTING AND QUALITY ANALYST 2,3998,99.42471

COMPUTER PROGRAMMER/CONFIGURER 3,1145,99.38865

COMPUTER SPECIALIST/SYSTEM SUPPORT AND DEVELOPMENT,1339,99.32786

COMPUTER SPECIALIST/SYSTEM SUPPORT AND DEVELOPMENT ADMIN 2,1085,99.26267

DATA WAREHOUSE SPECIALIST,1631,99.20294

ASSURANCE STAFF,2334,99.05741

COMPUTER SYSTEMS ENGINEER/ARCHITECT,2067,98.79052

SOFTWARE QUALITY ASSURANCE ENGINEER AND TESTER,1568,98.66071

AUDIT SENIOR,1070,98.59813

TEST CONSULTANT,1454,98.55571

SOFTWARE ENGINEER AND TESTER,1216,98.51974

ARCHITECT LEVEL 2,2892,98.51314

PROGRAMMER/DEVELOPER,1560,98.46154

TEST ENGINEER LEVEL 2,2372,98.44013

MODULE LEAD,2226,98.33782

ADVISORY MANAGER,3255,98.310295

AUDIT ASSISTANT,1205,98.25726

LEAD ENGINEER,11157,98.23429

COMPUTER SPECIALIST,2175,98.206894

CONSULTANT LEVEL 3,1171,98.12126

DEVELOPER,12909,98.00914

ERS SENIOR CONSULTANT,2249,97.95464

TAX SENIOR,1838,97.93253

TEST ENGINEER LEVEL 1,1036,97.87645

PROGRAMMER ANALYST LEVEL 1,2395,97.87057

TECHNICAL ANALYST,2932,97.78308

SOFTWARE DEVELOPMENT ENGINEER IN TEST,4258,97.65148

ADVISORY SENIOR ASSOCIATE,1332,97.52252

SOFTWARE ENGINEER 2,4166,97.16755

ERS CONSULTANT,2170,97.14286

FUNCTIONAL CONSULTANT,1115,97.04036

QA TESTER,1170,96.92308

SOFTWARE DEVELOPMENT ENGINEER,7284,96.856125

PROGRAMMER ANALYSTS,1133,96.82259

COMPUTER SYSTEMS ENGINEER,11090,96.69973

BUSINESS SYSTEM ANALYST,4435,96.685455

LEAD CONSULTANT,2169,96.63439

ELEMENTARY BILINGUAL TEACHER,2088,96.59962

ASSISTANT VICE PRESIDENT,2132,96.57599

CONSULTANT,23081,96.57294

PRINCIPAL CONSULTANT,1836,96.56863

SYSTEMS ANALYSTS,1252,96.5655

SYSTEM ADMINISTRATOR,5048,96.55309

SOLUTIONS ARCHITECT,1915,96.501305

APPLICATIONS CONSULTANT,1180,96.44068

ORACLE DATABASE ADMINISTRATOR,1527,96.39816

SOFTWARE DESIGN ENGINEER,1080,96.388885

PROGRAMMER/ANALYST,9375,96.37334

ASSOCIATE RESEARCH SCIENTIST,1400,96.35714

RESEARCH FELLOW,5981,96.355125

APPLICATIONS DEVELOPER,3366,96.34581

SENIOR SOFTWARE DEVELOPER,10208,96.31661

PROGRAMMER ANALYST,249038,96.1279

SENIOR PROGRAMMER ANALYST,5810,96.127365

ASSISTANT RESEARCH SCIENTIST,1103,96.10154

SENIOR ASSOCIATE,3540,96.01695

SOFTWARE QUALITY ASSURANCE ENGINEER,4920,95.97561

SENIOR MANAGER,1439,95.96942

PHYSICIAN IN A POST GRADUATE TRAINING PROGRAM,2421,95.95209

SYSTEMS ANALYST,61965,95.947716

QUALITY ASSURANCE ANALYST,7326,95.945946

TECHNICAL ARCHITECT,2908,95.94223

PROJECT LEAD,2363,95.93737

SOFTWARE ENGINEER III,1328,95.93374

SOFTWARE ANALYST,1072,95.89552

SR. SYSTEMS ANALYST,1151,95.82971

SOFTWARE ENGINEER 3,1891,95.82232

LEAD DEVELOPER,1049,95.80553

QUALITY ANALYST,2616,95.795105

SENIOR SOFTWARE DEVELOPMENT ENGINEER,1399,95.7827

SOFTWARE QA ANALYST,1112,95.773384

TEST ANALYST,1419,95.77167

SENIOR CONSULTANT,24904,95.7557

TECHNICAL SPECIALIST,1295,95.7529

SENIOR TECHNICAL CONSULTANT,1882,95.74921

SR. PROGRAMMER ANALYST,3716,95.748116

DATA SCIENTIST,1932,95.70393

SAP CONSULTANT,3023,95.69964

COMPUTER PROGRAMMER,70570,95.628456

PROGRAM MANAGER,3920,95.58673

BUSINESS SYSTEMS ANALYST,10110,95.56874

PROGRAMMER / ANALYST,1173,95.566925

MEMBER OF TECHNICAL STAFF,1774,95.54678

ETL DEVELOPER,1841,95.5459

SOFTWARE PROGRAMMER,3577,95.52698

PRINCIPAL SOFTWARE ENGINEER,2257,95.52503

POSTDOCTORAL SCHOLAR,3186,95.480225

SOFTWARE ENGINEER & TESTER,1538,95.44863

MANAGER,8561,95.4211

JAVA DEVELOPER,7596,95.39231

SOFTWARE DEVELOPMENT ENGINEER II,3274,95.35736

POSTDOCTORAL RESEARCH ASSOCIATE,6041,95.315346

SCIENCE TEACHER,1127,95.29725

POSTDOCTORAL ASSOCIATE,5145,95.2964

SENIOR DATABASE ADMINISTRATOR,1229,95.280716

SYSTEMS ANALYST II,1036,95.27027

BUSINESS ANALYST,39681,95.19669

ANALYST,11751,95.1919

QA ANALYST,6871,95.182655

DATABASE ADMINISTRATOR,16665,95.18152

ASSOCIATE,12502,95.17677

SENIOR BUSINESS ANALYST,3402,95.14991

ASSOCIATE SOFTWARE ENGINEER,1215,95.144035

SENIOR SYSTEMS ANALYST,5353,95.142914

SOFTWARE ENGINEER,121307,95.1157

SOFTWARE APPLICATION ENGINEER,1126,95.115456

COMPUTER PROGRAMMER/ANALYST,1122,95.09804

SYSTEMS ENGINEER,8078,95.09779

COMPUTER PROGRAMMER ANALYST,13634,95.071144

DATABASE ANALYST,1050,95.04762

SOFTWARE QA ENGINEER,1169,95.0385

.NET DEVELOPER,2921,95.03595

COMPUTER SYSTEMS ANALYSTS,4728,95.02961

WEB DEVELOPER,8024,95.01495

ARCHITECT,4982,95.00201

HARDWARE ENGINEER,2556,94.99217

RESEARCH ASSISTANT PROFESSOR,1973,94.98226

QUALITY ASSURANCE ENGINEER,3647,94.98218

COMPUTER SOFTWARE ENGINEER, APPLICATIONS,4426,94.916405

VALIDATION ENGINEER,1159,94.9094

DATABASE DEVELOPER,1155,94.89178

COMPUTER SYSTEMS ANALYST,35086,94.86405

POSTDOCTORAL FELLOW,7857,94.858086

SENIOR HARDWARE ENGINEER,1653,94.857834

QA ENGINEER,2224,94.82914

ENGINEER II,1249,94.715775

SR. SOFTWARE ENGINEER,4863,94.715195

STAFF SCIENTIST,1242,94.68599

BUSINESS INTELLIGENCE ANALYST,1972,94.67545

PRINCIPAL ENGINEER,1066,94.65291

RF ENGINEER,2794,94.631355

SYSTEMS ADMINISTRATOR,6659,94.59378

SENIOR SOFTWARE ENGINEER,27133,94.56381

SENIOR JAVA DEVELOPER,1395,94.55197

SOFTWARE TEST ENGINEER,3591,94.54191

VISITING ASSISTANT PROFESSOR,1311,94.50801

TECHNICAL CONSULTANT,3420,94.50292

SOFTWARE DEVELOPER,42907,94.49973

STAFF ENGINEER,1869,94.48903

NETWORK ENGINEER,5422,94.46699

PROJECT MANAGER,20172,94.437836

IT PROJECT MANAGER,2473,94.41973

APPLICATION ENGINEER,1458,94.375854

SOFTWARE ARCHITECT,1878,94.3557

PROGRAMMER,6011,94.327065

VICE PRESIDENT,3159,94.301994

CLINICAL ASSISTANT PROFESSOR,1281,94.30133

ASSISTANT PROFESSOR,25265,94.300415

ASSOCIATE CONSULTANT,1350,94.296295

SENIOR SYSTEMS ENGINEER,2030,94.28571

TECHNICAL RECRUITER,1364,94.281525

RESEARCHER,1031,94.2774

SOFTWARE DEVELOPER, APPLICATIONS,1830,94.2623

STAFF SOFTWARE ENGINEER,2976,94.254036

TEST ENGINEER,3936,94.23273

RESEARCH ASSOCIATE,13623,94.20098

SENIOR PRODUCT MANAGER,1085,94.19355

DATA ANALYST,3805,94.16557

TECHNICAL SUPPORT ENGINEER,1230,94.14634

PROGRAMMER ANALYST II,1059,94.145424

SYSTEM ANALYST,4684,94.12895

POSTDOCTORAL RESEARCHER,2130,94.08451

LECTURER,2257,94.06291

SR. SOFTWARE DEVELOPER,1161,94.05685

SOFTWARE ENGINEER II,2051,94.05168

PRODUCT MANAGER,3367,94.0303

IT CONSULTANT,3497,93.99485

QUALITY ENGINEER,2381,93.99412

SENIOR RESEARCH ASSOCIATE,1015,93.99015

SPEECH LANGUAGE PATHOLOGIST,1381,93.98986

SENIOR ANALYST,1646,93.98542

SYSTEMS ANALYST JC65,3321,93.977715

RESEARCH ENGINEER,1338,93.94619

DESIGN ENGINEER,6325,93.928856

POSTDOCTORAL RESEARCH FELLOW,2346,93.90452

MANUFACTURING ENGINEER,1906,93.80902

SENIOR DESIGN ENGINEER,1209,93.796524

COMPUTER PROGRAMMERS,4963,93.773926

TECHNICAL MANAGER,1060,93.77358

COMPONENT DESIGN ENGINEER,2851,93.75658

NETWORK ADMINISTRATOR,2624,93.71189

SOFTWARE DEVELOPMENT ENGINEER I,2128,93.70301

RESEARCH ASSISTANT,1777,93.69724

QUANTITATIVE ANALYST,1293,93.65816

TECHNICAL LEAD,5218,93.65657

POST DOCTORAL FELLOW,1507,93.56337

ASSOCIATE PROFESSOR,1441,93.54615

RESEARCH SCIENTIST,5154,93.46139

SENIOR DEVELOPER,2994,93.453575

PHARMACIST,5864,93.400406

INSTRUCTOR,3014,93.39748

SENIOR SYSTEMS ANALYST JC60,3069,93.38547

SENIOR ENGINEER,3773,93.37397

PRODUCT ENGINEER,2634,93.24222

ENGINEER,4941,93.05808

NETWORK AND COMPUTER SYSTEMS ADMINISTRATOR,1928,93.04979

LEAD SOFTWARE ENGINEER,1572,93.00255

SOFTWARE DEVELOPERS, APPLICATIONS,1195,92.97071

COMPUTER SYSTEM ANALYST,3753,92.752464

OCCUPATIONAL THERAPIST,4437,92.72031

SENIOR APPLICATION DEVELOPER,1965,92.620865

SENIOR PROJECT MANAGER,1015,92.61084

CLINICAL FELLOW,1146,92.58289

INDUSTRIAL DESIGNER,3619,92.56701

PROJECT ENGINEER,6439,92.56096

PHYSICAL THERAPIST,20207,92.53724

APPLICATIONS ENGINEER,1688,92.535545

SOLUTION ARCHITECT,1994,92.52758

HOSPITALIST,4387,92.50057

TEST LEAD,1726,92.410194

DIRECTOR,1333,92.34808

HOSPITALIST PHYSICIAN,4067,92.3285

SENIOR FINANCIAL ANALYST,1196,92.22408

SYSTEM ENGINEER,2145,92.214455

MANAGER JC50,1874,91.88901

MEDICAL RESIDENT,2336,91.78082

PSYCHIATRIST,1289,91.54383

COMPUTER SOFTWARE ENGINEER,2684,91.50522

SCIENTIST,1340,91.49254

RESIDENT,1245,91.48595

APPLICATION DEVELOPER,7692,91.45866

TECHNICAL PROJECT MANAGER,1052,91.25475

PROCESS ENGINEER,4377,91.22687

PHYSICIAN,4417,91.05728

ENGINEERING MANAGER,1199,90.90909

MECHANICAL ENGINEER,7301,90.83687

RESEARCH ANALYST,1869,90.74371

SPECIAL EDUCATION TEACHER,1721,90.47066

MANAGING CONSULTANT,3873,90.34341

TEST SPECIALIST,1011,90.306625

STRUCTURAL ENGINEER,1094,90.219376

ACCOUNT MANAGER,1066,90.15009

SENIOR SYSTEM ENGINEER,1408,90.12784

APPLICATION PROGRAMMER,1686,90.0949

NEPHROLOGIST,1027,89.97079

ARCHITECTURAL DESIGNER,2334,89.76007

MEDICAL TECHNOLOGIST,3125,89.568

RESIDENT PHYSICIAN,2119,89.476166

SALES ENGINEER,2167,89.386246

PEDIATRICIAN,1214,89.37397

DESIGNER,1992,89.30723

INDUSTRIAL ENGINEER,2093,89.2021

OPERATIONS RESEARCH ANALYST,1946,89.15724

IT SPECIALIST,2585,89.12959

ASSOCIATE ATTORNEY,1533,89.0411

ELECTRONICS ENGINEER,1060,88.86793

CHEMIST,1380,88.840576

MARKETING ANALYST,1573,88.684044

ELECTRICAL ENGINEER,4174,88.42837

MANAGEMENT ANALYST,5386,88.043076

BUSINESS DEVELOPMENT MANAGER,2345,88.01706

BUSINESS DEVELOPMENT ANALYST,1148,87.979095

ELEMENTARY SCHOOL TEACHER,1304,87.34663

STAFF ACCOUNTANT,4491,87.26341

FINANCIAL ANALYST,8515,87.152084

BUDGET ANALYST,1687,87.13693

BUSINESS DEVELOPMENT SPECIALIST,1482,87.11201

SYSTEM ANALYST JC65,1419,86.82171

DENTIST,3250,86.738464

TEACHER,3576,86.717

MARKETING SPECIALIST,2150,86.558136

BUSINESS OPERATIONS SPECIALIST,1034,85.97679

ATTORNEY,1050,85.90476

INTERIOR DESIGNER,1361,85.45187

CIVIL ENGINEER,2257,84.980064

MARKET RESEARCH ANALYST,8934,84.45265

LAW CLERK,1709,83.90872

SALES MANAGER,1232,83.68507

OPERATIONS MANAGER,1785,83.47339

ACCOUNTANT,14048,83.470955

GRAPHIC DESIGNER,5020,83.14741

PUBLIC RELATIONS SPECIALIST,1931,82.70326

FINANCIAL MANAGER,1080,82.40741

MARKETING MANAGER,2230,80.807175

CHIEF EXECUTIVE OFFICER,1095,80.63927

GENERAL MANAGER,1348,78.48665

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

mysql> create database H1B;

Query OK, 1 row affected (0.15 sec)

mysql> use H1B;

Database changed

sqoop export --connect jdbc:mysql://localhost/H1B --username hduser --password 'hduser' --table question10 --update-mode allowinsert --export-dir /niit/p10result --input-fields-terminated-by ',';

mysql> select \* from question10;

**CONCLUSION:**

Hadoop has been very effective solution for companies dealing with the data in perabytes.It has solved many problems in industry related to huge data management and distributed system.As it is open source, so it is adopted by companies widely. And the data is analysed on the basis of various algorithms and the efficient results are yielded.