**H1B visa dataset**

**BIG DATA - HADOOP**

Krishnakumar G |Professional Diploma in Digital Transformation – Big Data with Hadoop |

24.10.2017

Tools : Apache Hadoop Framework – HDFS, MapReduce, Hive, Pig, Sqoop, MySql and MS Excel for Data Visualization

S180030100194

NIIT

**H1B visa Data set Analysis**

**A PROJECT REPORT**

***Submitted by***

***KRISHNAKUMAR.G***

***in the partial fulfillment for the award of the course***

***of***

**PROFESSIONAL DIPLOMA**

***in***

**BIG DATA WITH HADOOP**

**NIIT, CHENNAI**

**JUNE 2017**

**ACKNOWLEDGEMENT**

I find immense pleasure to convey my sincere and grateful thanks to **NIIT** and the management for providing necessary facilities in carrying out this project.

I greatly indebted to my Tech Mentor **Ms. Amirtha,** the batch instructor **Mr. Annu** and the SLT faculty **Mr. Sandeep** for constant support throughout the course and also for useful suggestions, constant encouragement and kind advice in bringing out this project as a success.

I express my regards and sincere thanks to the Academic Leader **Ms. Aishwarya** for providing all necessary resource to complete the project. I extend my thanks to all staff members of NIIT for their kind co-operation for the completion of the project successfully. I am grateful to thank my family and all my friends for their valuable feedback, encouragement and suggestions.

**Introduction**

**Big Data**

Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal parallel processing power, analytics capabilities and skills.

While the term “big data” is relatively new, the act of gathering and storing large amounts of information for eventual analysis is ages old. The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the four Vs:

\* **Volume-**Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would’ve been a problem – but new technologies (such as Hadoop) have eased the burden.

\* **Velocity-** Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time

\* **Variety-**Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

\* **veracity-**accuracy of data

4V’s of Big Data

**Apache Hadoop**

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

**History**

The genesis of Hadoop came from the Google File System paper that was published in October 2003. This paper spawned another research paper from Google – MapReduce: Simplified Data Processing on Large Clusters. Development started in the Apache Nutch project, but was moved to the new Hadoop subproject in January 2006. The first committer added to the Hadoop project was Owen O’Malley in March 2006. Hadoop 0.1.0 was released in April 2006 and continues to be evolved by the many contributors to the Apache Hadoop project. Hadoop was named after one of the founder’s toy elephant.

Hadoop Timeline

**Benefits**

Some of the reasons organizations use Hadoop is its’ ability to store, manage and analyze vast amounts of structured and unstructured data quickly, reliably, flexibly and at low-cost.

- **Scalability and Performance** – distributed processing of data local to each node in a cluster enables Hadoop to store, manage, process and analyze data at petabyte scale.

- **Reliability** – large computing clusters are prone to failure of individual nodes in the cluster. Hadoop is fundamentally resilient – when a node fails processing is re-directed to the remaining nodes in the cluster and data is automatically re-replicated in preparation for future node failures.

- **Flexibility** – unlike traditional relational database management systems, you don’t have to created structured schemas before storing data. You can store data in any format, including semi-structured or unstructured formats, and then parse and apply schema to the data when read.

- **Low Cost** – unlike proprietary software, Hadoop is open source and runs on low-cost commodity hardware.

Hadoop framework and Apache projects:

1. **Hadoop Common**: The common utilities that support the other Hadoop modules.

2. **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.

3. **Hadoop YARN**: A framework for job scheduling and cluster resource management.

4. **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

Hadoop-related projects at Apache include:

5. **HBase**: A scalable, distributed database that supports structured data storage for large tables.

6. **Hive**: A data warehouse infrastructure that provides data summarization and ad hoc querying.

7. **Pig**: A high-level data-flow language and execution framework for parallel computation.

8. **Spark**: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.

9. **ZooKeeper**: A high-performance coordination service for distributed applications.

**Why Big Data Analytics?**

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today’s technology, it’s possible to analyze data and get answers from it almost immediately – an effort that’s slower and less efficient with more traditional business intelligence solutions.

**Big Data – Government**

By implementing a big data platform, governments can access vast amounts of relevant information important to their daily functions. The positive effect it can have is nearly endless. It’s so important because it not only allows the government to pinpoint areas that need attention, but it also gives them that information in real time. In a society that moves so quickly from one thing to the next, real-time analysis is vital. It allows governments to make faster decisions, and it allows them to monitor those decisions and quickly enact changes if necessary.

**Uses of h1b visa dataset**

Every country needs basic information on its residents for purposes of planning, development and improvement of the residents' quality of life. Good planning is based on reliable,up-to-date, accurate and detailed information on the state of the society in the country. This information makes it possible to plan better services, improve the quality of life and solve existing problems. Statistical information, which serves as the basis for constructing planning forecasts, is essential for the democratic process since it enables the citizens to examine the decisions made by the government and local authorities, and decide whether they serve the public they are meant to help. For these reasons official statistics are collected and published in all countries, world-wide. Thus, for example,while planning a road system, the planners use information regarding the quantity of people and number of vehicles who are likely to use the road; for if not, the advantages of constructing the road may not justify its cost. Similarly, when planning a school system, there is a need for a forecast of the number of pupils who are likely to need schools, in order to ensure that they will be built in locations with an appropriate number of pupils.

**The importance of the employment details**

The h1b visa data set, wide-range activity, which takes place once a decade in the entire country. Its purpose is to gather information about the general population, in order to present a full and reliable picture of the population in the country - its housing conditions and demographic, social and economic characteristics.The information collected includes data on,

**S\_no**

**Case\_status**

**Employer\_name**

**Job\_title**

**Soc\_name**

**Soc\_code**

**Job\_title**

**Full\_time\_position**

**Previling\_wages**

**Year**

**Work\_site**

**Lon**

**Lat**

**Immigration to the United States**

Immigration to the United States is the international movement of individuals who are not natives or do not possess citizenship in order to settle, reside, study or to take-up employment in the United States. It has been a major source of population growth and cultural change throughout much of the history of the United States. The economic, social, and political aspects of immigration have caused controversy regarding ethnicity, economic benefits, jobs for non-immigrants, settlement patterns, impact on upward social mobility, crime, and voting behavior.

As for economic effects, research suggests that immigration to the United States is beneficial to the US economy. Research, with few exceptions, finds that immigration on average has positive economic effects on the native population, but is mixed as to whether low-skilled immigration adversely affects low-skilled natives. Research finds that immigration either has no impact on the crime rate or that it reduces the crime rate in the United States. Research shows that the United States excels at assimilating first- and second-generation immigrants relative to many other Western countries.

The census is extremely important for documenting the growth of immigrant communities, allocating resources for needed services, and identifying areas where civil rights enforcement may be needed. Department of Homeland Security (DHS), United States of America keeps track of all data from the Bureau of the Census, and do statistical analysis to plan its policy and funds.

**Total Data and Statistics**

total data and statistics of the h1b visa data set . that are total for 20111,2012,2013,2014,2015,2016.that consist of serial no, case status ,job title, employer name, latitude ,longitude, worksite ,prevailing wages, years.

**Benefit of h1b visa**

Everyone, including immigrants, benefits from investments in education, health care, and jobs that are distributed based on census information. And census data are also used in ways that are of special importance to immigrants, including:

1.finding the talented persons in each and every country by giving high pay.

2country economy gets improved because of those immigrants

. And the technology gets peroved

3. standary of the people living in the country gets improved

4. evaluating the effectiveness of equal opportunity employment programs and policies under the Civil Rights Act;

5. allocating funds to school districts for children with limited English language proficiency.

**Term in this Project**

The columns in the dataset include:

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

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

Certified Withdrawn: LCA was approved but later withdrawn by employer

Withdrawn: LCA was withdrawn by employer before approval

Denied: LCA was denied by DOL

* EMPLOYER\_NAME: Name of employer submitting labour condition application.
* SOC\_NAME: the Occupational name associated with the SOC\_CODE. SOC\_CODE is the occupational code associated with the job being requested for temporary labour condition, as classified by the Standard Occupational Classification (SOC) System.
* JOB\_TITLE: Title of the job
* FULL\_TIME\_POSITION: Y = Full Time Position; N = Part Time Position
* PREVAILING\_WAGE: Prevailing Wage for the job being requested for temporary labour condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer’s minimum requirements for the position.
* YEAR: Year in which the H1B visa petition was filed
* WORKSITE: City and State information of the foreign worker’s intended area of employment
* lon: longitude of the Worksite
* lat: latitude of the Worksite

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

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

select year,job\_title,count(s\_no) from h1b\_final where job\_title='DATA ENGINEER' group by year,job\_title order by year;

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

data1 = LOAD '/user/hive/warehouse/h1b.db/h1b\_final' USING PigStorage('\t') as (s\_no:double,case\_status:chararray,employer\_name:chararray,soc\_name:chararray,job\_title:chararray,full\_time\_position:chararray,prevailing\_wage:double,year:chararray,worksite:chararray,longitude,latitude);

--B = LOAD '/user/hive/warehouse/h1b.db/h1b\_final' USING PigStorage('\t') as (s\_no:double,case\_status:chararray,employer\_name:chararray,soc\_name:chararray,job\_title:chararray,full\_time\_position:chararray,prevailing\_wage:double,year:chararray,worksite:chararray,longitude,latitude);

--data1 = UNION A,B;

--dump data1;

data2 = FOREACH data1 GENERATE $1,$4,$7;

--dump data2;

year1 = FILTER data2 BY $2 == '2011';

year2 = FILTER data2 BY $2 == '2012';

year3 = FILTER data2 BY $2 == '2013';

year4 = FILTER data2 BY $2 == '2014';

year5 = FILTER data2 BY $2 == '2015';

year6 = FILTER data2 BY $2 == '2016';

group1 = group year1 by ($1);

group2 = group year2 by ($1);

group3 = group year3 by ($1);

group4 = group year4 by ($1);

group5 = group year5 by ($1);

group6 = group year6 by ($1);

Count1 = foreach group1 generate group, COUNT(year1.$1);

Count2 = foreach group2 generate group, COUNT(year2.$1);

Count3 = foreach group3 generate group, COUNT(year3.$1);

Count4 = foreach group4 generate group, COUNT(year4.$1);

Count5 = foreach group5 generate group, COUNT(year5.$1);

Count6 = foreach group6 generate group, COUNT(year6.$1);

JOIN1 = join Count1 BY $0, Count2 BY $0, Count3 BY $0, Count4 BY $0, Count5 BY $0,Count6 BY $0;

data3 = foreach JOIN1 GENERATE $0,$1,$3,$5,$7,$9,$11;

--dump data3;

average1 = FOREACH data3 GENERATE $0, (DOUBLE)((($2-$1)\*100)/$1+(($3-$2)\*100)/$2+(($4-$3)\*100)/$3+(($5-$4)\*100)/$4+(($6-$5)\*100)/$5);

RESULT1 = FOREACH average1 GENERATE $0, (DOUBLE)($1/5);

--DUMP RESULT1

AVERAGEDESC1= LIMIT (ORDER RESULT1 BY $1 DESC) 5;

DUMP AVERAGEDESC1;

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

load\_data = Load '/user/hive/warehouse/h1b.db/h1b\_final' using PigStorage('\t') as (s\_no,case\_status,employer\_name,soc\_name,job\_title:chararray,full\_time\_position,prevailing\_wage,year:chararray,worksite:chararray,longitute,latitute);

filter\_data\_2011 = filter load\_data by job\_title == 'DATA ENGINEER' and year == '2011';

groupby\_worksite = group filter\_data\_2011 by worksite ;

bag2011 = foreach groupby\_worksite generate group as worksite,COUNT(filter\_data\_2011.job\_title) as total;

final =limit( order bag2011 by total desc)1;

filter\_data\_2012 = filter load\_data by job\_title == 'DATA ENGINEER' and year == '2012';

groupby\_worksite2 = group filter\_data\_2012 by worksite ;

bag2012 = foreach groupby\_worksite2 generate group as worksite,COUNT(filter\_data\_2012.job\_title) as total;

final2 =limit( order bag2012 by total desc)1;

filter\_data\_2013 = filter load\_data by job\_title == 'DATA ENGINEER' and year == '2013';

groupby\_worksite3 = group filter\_data\_2013 by worksite ;

bag2013 = foreach groupby\_worksite3 generate group as worksite,COUNT(filter\_data\_2013.job\_title) as total,filter\_data\_2013.year;

final3 =limit( order bag2013 by total desc)1;

filter\_data\_2014 = filter load\_data by job\_title == 'DATA ENGINEER' and year == '2014';

groupby\_worksite4 = group filter\_data\_2014 by worksite ;

bag2014 = foreach groupby\_worksite4 generate group as worksite,COUNT(filter\_data\_2014.job\_title) as total;

final4 =limit( order bag2014 by total desc)1;

filter\_data\_2015 = filter load\_data by job\_title == 'DATA ENGINEER' and year == '2015';

groupby\_worksite5 = group filter\_data\_2015 by worksite ;

bag2015 = foreach groupby\_worksite5 generate group as worksite,COUNT(filter\_data\_2015.job\_title) as total;

final5 =limit( order bag2015 by total desc)1;

filter\_data\_2016 = filter load\_data by job\_title == 'DATA ENGINEER' and year == '2016';

groupby\_worksite6 = group filter\_data\_2016 by worksite ;

bag2016 = foreach groupby\_worksite6 generate group as worksite,COUNT(filter\_data\_2016.job\_title) as total;

final6 =limit( order bag2016 by total desc)1;

total\_final = union final,final2,final3,final4,final5,final6;

dump total\_final;

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

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.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.NullWritable;

//import org.apache.hadoop.io.NullWritable;

//import org.apache.hadoop.io.NullWritable;

//import org.apache.hadoop.io.NullWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Partitioner;

import org.apache.hadoop.mapreduce.Reducer;

//import org.apache.hadoop.mapreduce.Reducer.Context;

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

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

public class Project\_2b {

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

{

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

{

try{

String[] record=val.toString().toUpperCase().split("\t");

//String job\_title=record[4];

String year=record[7];

String worksite=record[8];

String case\_status=record[1];

String value=1+","+year;

if(case\_status.equals("CERTIFIED"))

{

context.write(new Text(worksite),new Text(value));

}

}

catch(Exception e)

{

System.out.println(e);

}

}

}

public static class part extends Partitioner<Text,Text>

{

@Override

public int getPartition(Text key,Text val,int numReduceTasks)

{

String[] str= val.toString().split(",");

String year=str[1];

if(year.equals("2011"))

{

return 0;

}

else

if(year.equals("2012"))

{

return 1;

}

else

if(year.equals("2013"))

{

return 2;

}

else

if(year.equals("2014"))

{

return 3;

}

else

if(year.equals("2015"))

{

return 4;

}

else

{

return 5;

}

}

}

public static class reduceclass extends Reducer<Text,Text,NullWritable,Text>

{

TreeMap<IntWritable,Text> map=new TreeMap<>();

String year="";

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

{

int count=0;

for(Text str:val)

{

String[] arr = str.toString().split(",");

count+=Integer.parseInt(arr[0]);

year=arr[1];

}

String v1=year+" --- "+count+" -- "+key;

// String val1=count+","+year;

//context.write(key,new Text(count+","+year));

map.put(new IntWritable(count),new Text(v1));

if(map.size()>5)

{

map.remove(map.firstKey());

}

}

// String result = str + "," + max;

//context.write(key,new IntWritable(count));

protected void cleanup(Context context) throws IOException,InterruptedException

{

for(Text s:map.descendingMap().values())

{

context.write(NullWritable.get(),new Text(s));

}

}

}

public static void main(String[] args) throws Exception

{

Configuration conf=new Configuration();

Job job = Job.getInstance(conf, "H1b\_Visa 2b question");

job.setJarByClass(Project\_2b.class);

job.setMapperClass(mapclass.class);

job.setReducerClass(reduceclass.class);

job.setNumReduceTasks(6);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(Text.class);

//job.setCombinerClass(reduceclass.class);

job.setPartitionerClass(part.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);

}

}

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

[certified]

select soc\_name,count(job\_title) as tot from h1b\_final where job\_title='DATA SCIENTIST' and case\_status='CERTIFIED' group by soc\_name order by tot desc limit 1;

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.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.NullWritable;

//import org.apache.hadoop.io.NullWritable;

//import org.apache.hadoop.io.NullWritable;

//import org.apache.hadoop.io.NullWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Partitioner;

import org.apache.hadoop.mapreduce.Reducer;

//import org.apache.hadoop.mapreduce.Reducer.Context;

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

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

public class Project\_4 {

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

{

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

{

try{

String[] record=val.toString().toUpperCase().split("\t");

String emp\_name=record[2];

String year=record[7];

String value=1+","+year;

context.write(new Text(emp\_name),new Text(value));

}

catch(Exception e)

{

System.out.println(e);

}

}

}

public static class part extends Partitioner<Text,Text>

{

@Override

public int getPartition(Text key,Text val,int numReduceTasks)

{

String[] str= val.toString().split(",");

String year=str[1];

if(year.equals("2011"))

{

return 0;

}

else

if(year.equals("2012"))

{

return 1;

}

else

if(year.equals("2013"))

{

return 2;

}

else

if(year.equals("2014"))

{

return 3;

}

else

if(year.equals("2015"))

{

return 4;

}

else

{

return 5;

}

}

}

public static class reduceclass extends Reducer<Text,Text,NullWritable,Text>

{

TreeMap<IntWritable,Text> map=new TreeMap<>();

String year="";

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

{

int count=0;

for(Text str:val)

{

String[] arr = str.toString().split(",");

count+=Integer.parseInt(arr[0]);

year=arr[1];

}

String v1=year+" --- "+count+" -- "+key;

// String val1=count+","+year;

//context.write(key,new Text(count+","+year));

map.put(new IntWritable(count),new Text(v1));

if(map.size()>5)

{

map.remove(map.firstKey());

}

}

// String result = str + "," + max;

//context.write(key,new IntWritable(count));

protected void cleanup(Context context) throws IOException,InterruptedException

{

for(Text s:map.descendingMap().values())

{

context.write(NullWritable.get(),new Text(s));

}

}

}

public static void main(String[] args) throws Exception

{

Configuration conf=new Configuration();

Job job = Job.getInstance(conf, "H1b\_Visa 4th question");

job.setJarByClass(Project\_4.class);

job.setMapperClass(mapclass.class);

job.setReducerClass(reduceclass.class);

job.setNumReduceTasks(6);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(Text.class);

//job.setCombinerClass(reduceclass.class);

job.setPartitionerClass(part.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);

}

}

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

a) for all the applications

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.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Partitioner;

import org.apache.hadoop.mapreduce.Reducer;

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

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

public class Project\_5a {

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

{

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

{

try{

String[] record=val.toString().toUpperCase().split("\t");

String job\_title=record[4];

String year=record[7];

String value=1+","+year;

context.write(new Text(job\_title),new Text(value));

}

catch(Exception e)

{

System.out.println(e);

}

}

}

public static class part extends Partitioner<Text,Text>

{

@Override

public int getPartition(Text key,Text val,int numReduceTasks)

{

String[] str= val.toString().split(",");

String year=str[1];

if(year.equals("2011"))

{

return 0;

}

else

if(year.equals("2012"))

{

return 1;

}

else

if(year.equals("2013"))

{

return 2;

}

else

if(year.equals("2014"))

{

return 3;

}

else

if(year.equals("2015"))

{

return 4;

}

else

{

return 5;

}

}

}

public static class reduceclass extends Reducer<Text,Text,NullWritable,Text>

{

TreeMap<IntWritable,Text> map=new TreeMap<>();

String year="";

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

{

int count=0;

for(Text str:val)

{

String[] arr = str.toString().split(",");

count+=Integer.parseInt(arr[0]);

year=arr[1];

}

String v1=year+" --- "+count+" -- "+key;

map.put(new IntWritable(count),new Text(v1));

if(map.size()>5)

{

map.remove(map.firstKey());

}

}

protected void cleanup(Context context) throws IOException,InterruptedException

{

for(Text s:map.descendingMap().values())

{

context.write(NullWritable.get(),new Text(s));

}

}

}

public static void main(String[] args) throws Exception

{

Configuration conf=new Configuration();

Job job = Job.getInstance(conf, "H1b\_Visa 5a question");

job.setJarByClass(Project\_5a.class);

job.setMapperClass(mapclass.class);

job.setReducerClass(reduceclass.class);

job.setNumReduceTasks(6);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(Text.class);

//job.setCombinerClass(reduceclass.class);

job.setPartitionerClass(part.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);

}

}

b) for only certified applications.

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.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Partitioner;

import org.apache.hadoop.mapreduce.Reducer;

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

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

public class Project\_5b {

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

{

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

{

try{

String[] record=val.toString().toUpperCase().split("\t");

String job\_title=record[4];

String year=record[7];

String case\_status=record[1];

String value=1+","+year;

if(case\_status.equals("CERTIFIED"))

{

context.write(new Text(job\_title),new Text(value));

}

}

catch(Exception e)

{

System.out.println(e);

}

}

}

public static class part extends Partitioner<Text,Text>

{

@Override

public int getPartition(Text key,Text val,int numReduceTasks)

{

String[] str= val.toString().split(",");

String year=str[1];

if(year.equals("2011"))

{

return 0;

}

else

if(year.equals("2012"))

{

return 1;

}

else

if(year.equals("2013"))

{

return 2;

}

else

if(year.equals("2014"))

{

return 3;

}

else

if(year.equals("2015"))

{

return 4;

}

else

{

return 5;

}

}

}

public static class reduceclass extends Reducer<Text,Text,NullWritable,Text>

{

TreeMap<IntWritable,Text> map=new TreeMap<>();

String year="";

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

{

int count=0;

for(Text str:val)

{

String[] arr = str.toString().split(",");

count+=Integer.parseInt(arr[0]);

year=arr[1];

}

String v1=year+" --- "+count+" -- "+key;

map.put(new IntWritable(count),new Text(v1));

if(map.size()>5)

{

map.remove(map.firstKey());

}

}

protected void cleanup(Context context) throws IOException,InterruptedException

{

for(Text s:map.descendingMap().values())

{

context.write(NullWritable.get(),new Text(s));

}

}

}

public static void main(String[] args) throws Exception

{

Configuration conf=new Configuration();

Job job = Job.getInstance(conf, "H1b\_Visa 5b question");

job.setJarByClass(Project\_5b.class);

job.setMapperClass(mapclass.class);

job.setReducerClass(reduceclass.class);

job.setNumReduceTasks(6);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(Text.class);

//job.setCombinerClass(reduceclass.class);

job.setPartitionerClass(part.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);

}

}

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.

CERTIFIEDCOUNT/TOTALCOUNT\*100

hive -e "select a.case\_status,count(a.case\_status) as case\_count,round((count(a.case\_status)/total\*100),2) as case\_percent from h1b.h1b\_final a,h1b.totalcount b where a.year='$year' group by a.case\_status,b.total;"

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

select year,count(\*) from h1b\_final group by 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 - [Certified and Certified Withdrawn.]

select year,job\_title,avg(prevailing\_wage) as totavg from h1b\_final where full\_time\_position='Y' and (case\_status='CERTIFIED' OR case\_status='CERTIFIED-WITHDRAWN') group by year,job\_title order by totavg desc;

select year,job\_title,avg(prevailing\_wage) as totavg from h1b\_final where full\_time\_position='N' and (case\_status='CERTIFIED' OR case\_status='CERTIFIED-WITHDRAWN') group by year,job\_title order by totavg desc;

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) ?

load\_data = load '/user/hive/warehouse/h1b.db/h1b\_final' using PigStorage('\t') as (s\_no:int,case\_status:chararray,employer\_name:chararray,soc\_name:chararray,job\_title:chararray,full\_time\_position:chararray,prevailing\_wage:long,year:chararray,worksite:chararray,longitute:double,latitute:double);

groupby\_emp = group load\_data by employer\_name;

count\_total = foreach groupby\_emp generate group,COUNT(load\_data.case\_status) as totalcount;

filter\_certified = filter load\_data by case\_status == 'CERTIFIED';

group\_certified\_filter\_data = group filter\_certified by employer\_name;

count\_certified = foreach group\_certified\_filter\_data generate group,COUNT(filter\_certified.case\_status) as certifiedcount;

filter\_certified\_withdrawn = filter load\_data by case\_status == 'CERTIFIED-WITHDRAWN';

group\_certified\_withdrawn\_filter\_data = group filter\_certified\_withdrawn by employer\_name;

count\_certified\_withdrawn = foreach group\_certified\_withdrawn\_filter\_data generate group,COUNT(filter\_certified\_withdrawn.case\_status) as

certified\_withdrawncount;

join\_data = join count\_total by $0,count\_certified by $0,count\_certified\_withdrawn by $0;

--dump join\_data;

join\_data1 = foreach join\_data generate $0,$1,$3,$5 ;

successive\_rate = foreach join\_data1 generate $0,(float)($2+$3)/($1)\*100,$1;

final = filter successive\_rate by $1 > 70 and $2 > 1000;

final\_output = order final by $1 desc;

dump final\_output;

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)?

register /usr/local/hive/lib/hive-exec-1.2.1.jar

register /usr/local/hive/lib/hive-common-1.2.1.jar

load\_data = load '/user/hive/warehouse/h1b.db/h1b\_final' using PigStorage('\t') as (s\_no:int,case\_status:chararray,employer\_name:chararray,soc\_name:chararray,job\_title:chararray,full\_time\_position:chararray,prevailing\_wage:long,year:chararray,worksite:chararray,longitute:double,latitute:double);

groupby\_emp = group load\_data by job\_title;

count\_total = foreach groupby\_emp generate group,COUNT(load\_data.case\_status) as totalcount;

filter\_certified = filter load\_data by case\_status == 'CERTIFIED';

group\_certified\_filter\_data = group filter\_certified by job\_title;

count\_certified = foreach group\_certified\_filter\_data generate group,COUNT(filter\_certified.case\_status) as certifiedcount;

filter\_certified\_withdrawn = filter load\_data by case\_status == 'CERTIFIED-WITHDRAWN';

group\_certified\_withdrawn\_filter\_data = group filter\_certified\_withdrawn by job\_title;

count\_certified\_withdrawn = foreach group\_certified\_withdrawn\_filter\_data generate group,COUNT(filter\_certified\_withdrawn.case\_status) as

certified\_withdrawncount;

join\_data = join count\_total by $0,count\_certified by $0,count\_certified\_withdrawn by $0;

--dump join\_data;

join\_data1 = foreach join\_data generate $0,$1,$3,$5 ;

successive\_rate = foreach join\_data1 generate $0,(float)($2+$3)/($1)\*100,$1;

final = filter successive\_rate by $1 > 70 and $2 > 1000;

final\_output = order final by $1 desc;

dump final\_output;

store final\_output INTO '/output/pigout/pigproject10' USING PigStorage('\t');

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

mysql -u root -p'root' -e 'drop database question10;create database if not exists question10;use question10;create table q10(job\_title varchar(100),success\_rate float,petitions int);';

sqoop export --connect jdbc:mysql://localhost/question10 --username root --password 'root' --table q10 --update-mode allowinsert --export-dir /output/pigout/pigproject10/p\* --input-fields-terminated-by '\t'

mysql -u root -p'root' -e 'select \* from question10.q10';

**suggestions:**

1) Native people standard should be improved and more jobs to be created for natives.

2) Number of denied case status is very less . Must examine the candidate before giving the visa clearance .

3) Number of part-time job is full compare to the full time job . can increase the part-time job candidate.

4) Number of job vacancies in the Florida is must be improved.

**Conclusion:**

1) H1b data set project shows that the amount of case status that are with\_drawn and denied in year wise.

2) From the H1b dataset the data engineer job title increases over time.

3) And shows that in year 2011 new jersey

Is the state that has most number of data scientists each year

4) Top job position which the H1b visa is applied and in year wise.

5) The graph shows that number of applications increases over time.

6) Most number of data scientists are available in statistics job location.