**H1B visa dataset**

**BIG DATA - HADOOP**

SATHYA VB|Professional Diploma in Digital Transformation – Big Data with Hadoop |

23.01.2018

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

S180030100337

NIIT

**H1B visa Data set Analysis**

**A PROJECT REPORT**

***Submitted by***

***SATHYA VB***

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

***of***

**PROFESSIONAL DIPLOMA**

***in***

**BIG DATA WITH HADOOP**

**NIIT, CHENNAI**

**JANUARY 2018**

**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 **Mr.Oliver** 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.

**BIGDATA OVERVIEW**

“90% of the world’s data was generated in the last few years.”

Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year. The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. If you pile up the data in the form of disks it may fill an entire football field. The same amount was created in every two days in 2011, and in every ten minutes in 2013. This rate is still growing enormously. Though all this information produced is meaningful and can be useful when processed, it is being neglected.

**BIGDATA**

Big Data is a collection of large datasets that cannot be processed using traditional computing techniques. It is not a single technique or a tool, rather it involves many areas of business and technology.

**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.

Thus Big Data includes huge volume, high velocity, and extensible variety of data. 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.

**BIG DATA CHALLENGES**

The major challenges associated with big data are as follows:

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

**BIGDATA SOLUTIONS**

**TRADITIONAL APPROACH**

In this approach, an enterprise will have a computer to store and process big data. For storage purpose, the programmers will take the help of their choice of database vendors such as Oracle, IBM, etc. In this approach, the user interacts with the application, which in turn handles the part of data storage and analysis.

**Limitation**

This approach works fine with those applications that process less voluminous data that can be accommodated by standard database servers, or up to the limit of the processor that is processing the data. But when it comes to dealing with huge amounts of scalable data, it is a hectic task to process such data through a single database bottleneck.

**GOOGLE’S SOLUTION**

Google solved this problem using an algorithm called MapReduce. This algorithm divides the task into small parts and assigns them to many computers, and collects the results from them which when integrated, form the result dataset. Using the solution provided by Google, Doug Cutting and his team developed an Open Source Project called HADOOP. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others. In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

**HADOOP**

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage. At its core, Hadoop has two major layers namely:

* Processing/Computation layer (MapReduce),
* Storage layer (Hadoop Distributed File System).

**ADVANTAGES OF HADOOP**

* Scalable
* Cost Effective
* Flexible
* Fast
* Resilient to failure
* Compatible on all the platforms since it is Java based.

**HADOOP ECOSYSTEM**

**HDFS**

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules:

* **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules.
* **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.

**MAPREDUCE**

MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework.

**YARN**

**Apache Yarn** – “**Y**et **A**nother **R**esource **N**egotiator” is the resource management layer of **Hadoop**. The Yarn was introduced in Hadoop 2.x. Yarn allows different data processing engines like graph processing, interactive processing, stream processing as well as batch processing to run and process data stored in**HDFS** (Hadoop Distributed File System).

**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 analyzing easy. It provides SQL type language for querying called HiveQL or HQL. It stores schema in a database and processed data into HDFS. Traditional SQL queries must be implemented in the [MapReduce](https://en.wikipedia.org/wiki/MapReduce) Java API to execute SQL applications and queries over distributed data. Hive provides the necessary SQL abstraction to integrate SQL-like queries ([HiveQL](https://en.wikipedia.org/wiki/HiveQL)) into the underlying Java without the need to implement queries in the low-level Java API. Since most data warehousing applications work with SQL-based querying languages, Hive aids portability of SQL-based applications to Hadoop.

**PIG**

Pig is a high level programming language useful for analyzing large data sets. Pig uses PigLatinlanguage. Pig was a result of development effort at Yahoo! Pig enables people to focus more on **analyzing bulk data sets and to spend less time in writing Map-Reduce programs.** Similar to Pigs, who eat anything, the Pig programming language is designed to work upon any kind of data. That's why the name, Pig!

Pig has two execution modes:

* Local mode : In this mode, Pig runs in a single JVM and makes use of local file system. This mode is suitable only for analysis of small data sets using Pig.
* Map Reduce mode: In this mode, queries written in Pig Latin are translated into MapReduce jobs and are run on a Hadoop cluster (cluster may be pseudo or fully distributed). MapReduce mode with fully distributed cluster is useful of running Pig on large data sets.

**HBase**

HBase is a distributed column-oriented database built on top of the Hadoop file system. It is an open-source project and is horizontally scalable. HBase is a data model that is similar to Google’s big table designed to provide quick random access to huge amounts of structured data. It leverages the fault tolerance provided by the Hadoop File System (HDFS). One can store the data in HDFS either directly or through HBase. Data consumer reads/accesses the data in HDFS randomly using HBase. HBase sits on top of the Hadoop File System and provides read and write access.

**Sqoop**

Sqoop imports data from external sources into Hadoop components like HDFS, HBase and Hive. It also exports data from Hadoop to other sources. Sqoop works with relational databases such as Teradata, Netezza, Oracle, MySQL, Postgres etc.

**Flume**

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

**Other Components**

* **Ambari**– A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig, and Sqoop.
* **Avro**– A data serialization system.
* **HCatalog**- Helps data processing tools read and write data on the grid. It supports MapReduce and Pig.
* **Mahout**– A scalable machine learning and data mining library.
* **Oozie** - Workflow scheduler that was developed as part of the Apache Hadoop project. It manages how workflows start and execute, and also controls the execution path.
* **Zookeper**– A high-performance coordination service for distributed applications.

**HADOOP INSTALLATION**

Hadoop is supported by GNU/Linux platform and its flavors. Therefore, we have to install a Linux operating system for setting up Hadoop environment. In case you have an OS other than Linux, you can install a Virtualbox software in it and have Linux inside the Virtualbox.

**Installing Java**

Java is the main prerequisite for Hadoop.0

**user@laptop(local directory):~$** **sudo apt-get update**

**user@laptop:~$** **sudo apt-get install default-jdk**

**user@laptop:~$ java -version**

**java version "1.7.0\_65"**

**Adding a dedicated Hadoop user**

**user@laptop:~$ sudo addgroup hadoop**

**Adding group `hadoop' (GID 1002) ...**

**Done.**

**user@laptop:~$ sudo adduser --ingroup hadoop hduser**

**Adding user `hduser' ...**

**Adding new user `hduser' (1001) with group `hadoop' ...**

**Creating home directory `/home/hduser' ...**

**Installing SSH**

**ssh** has two main components:

1. **ssh** : The command we use to connect to remote machines - the client.
2. **sshd** : The daemon that is running on the server and allows clients to connect to the server.

The **ssh** is pre-enabled on Linux, but in order to start **sshd** daemon, we need to install **ssh** first. Use this command to do that :

**user@laptop:~$ sudo apt-get install ssh**

This will install ssh on our machine. If we get something similar to the following, we can think it is setup properly:

**user@laptop:~$ which ssh**

**/usr/bin/ssh**

**user@laptop:~$ which sshd**

**/usr/sbin/sshd**

**Create and Setup SSH Certificates**

Hadoop requires SSH access to manage its nodes, i.e. remote machines plus our local machine. For our single-node setup of Hadoop, we therefore need to configure SSH access to localhost.

**user@laptop:~$ su hduser**

**Password:**

**hduser@laptop:/home/user$ cd ~**

**hduser@laptop:~$ ssh-keygen -t rsa -P ""**

**Your public key has been saved in /home/hduser/.ssh/id\_rsa.pub.**

**hduser@laptop:~$ cat $HOME/.ssh/id\_rsa.pub >> $HOME/.ssh/authorized\_keys**

The second command adds the newly created key to the list of authorized keys so that Hadoop can use ssh without prompting for a password

**Install Hadoop**

**hduser@laptop:~$ su hduser**

**wget http://archive.apache.org/dist/hadoop/common/hadoop-2.6.0/hadoop-2.6.0.tar.gz**

**hduser@laptop:~$ tar xvzf hadoop-2.6.0.tar.gz**

**a folder would be created by the name of hadoop-2.6.0**

**rename this folder to hadoop**

**mv hadoop-2.6.0 hadoop**

We want to move the Hadoop installation to the **/usr/local/hadoop** directory using the following command:

**hduser@laptop:~$ sudo mv hadoop /usr/local/**

**[sudo] password for hduser:**

**hduser is not in the sudoers file. This incident will be reported.**

This error can be resolved by logging in as a root user, and then add **hduser** to **sudo**:

**hduser@laptop:~/hadoop-2.6.0$ su user (root user)**

**Password:**

**user@laptop:/home/hduser$ sudo adduser hduser sudo**

**[sudo] password for user:**

**Adding user `hduser' to group `sudo' ...**

**Adding user hduser to group sudo**

**Done.**

Now, the **hduser** has root priviledge, we can move the Hadoop installation to the**/usr/local/hadoop** directory without any problem.

**user@laptop:/home/hduser$ sudo su hduser**

**cd ~**

**hduser@laptop:~$ sudo mv hadoop /usr/local/**

**hduser@laptop:~$ sudo chown -R hduser:hadoop /usr/local/hadoop**

**Setup Configuration Files**

**1. ~/.bashrc:**

Before editing the **.bashrc** file in our home directory, we need to find the path where Java has been installed to set the **JAVA\_HOME** environment variable using the following command:

**hduser@laptop update-alternatives --config java**

**There is only one alternative in link group java (providing /usr/bin/java): /usr/lib/jvm/java-7-openjdk-amd64/jre/bin/java**

**Nothing to configure.**

Now we can append the following to the end of **~/.bashrc**:

**hduser@laptop:~$ nano ~/.bashrc OR gedit ~/.bashrc**

**#HADOOP VARIABLES START**

**export JAVA\_HOME=/usr/lib/jvm/java-7-openjdk-amd64**

**export HADOOP\_INSTALL=/usr/local/hadoop**

**export PATH=$PATH:$HADOOP\_INSTALL/bin**

**export PATH=$PATH:$HADOOP\_INSTALL/sbin**

**export HADOOP\_MAPRED\_HOME=$HADOOP\_INSTALL**

**export HADOOP\_COMMON\_HOME=$HADOOP\_INSTALL**

**export HADOOP\_HDFS\_HOME=$HADOOP\_INSTALL**

**export YARN\_HOME=$HADOOP\_INSTALL**

**export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_INSTALL/lib/native**

**export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_INSTALL/lib"**

**#HADOOP VARIABLES END**

**hduser@laptop:~$ source ~/.bashrc**

note that the JAVA\_HOME should be set as the path just before the '.../bin/':

**hduser@laptop:~$ javac -version**

**javac 1.7.0\_75**

**2. /usr/local/hadoop/etc/hadoop/hadoop-env.sh**

We need to set JAVA**\_**HOME by modifying hadoop-env.sh file.

**hduser@laptop:~$ nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh**

**export JAVA\_HOME=/usr/lib/jvm/java-7-openjdk-amd64**

**3. /usr/local/hadoop/etc/hadoop/core-site.xml**:

The /usr/local/hadoop/etc/hadoop/core-site.xml file contains configuration properties that Hadoop uses when starting up.   
This file can be used to override the default settings that Hadoop starts with.

**hduser@laptop:~$ sudo mkdir -p /app/hadoop/tmp**

**hduser@laptop:~$ sudo chown -R hduser:hadoop /app/hadoop/tmp**

**hduser@laptop:~$ nano /usr/local/hadoop/etc/hadoop/core-site.xml**

**<configuration>**

**<property>**

**<name>hadoop.tmp.dir</name>**

**<value>/app/hadoop/tmp</value>**

**<description>A base for other temporary directories.</description>**

**</property>**

**<property>**

**<name>fs.default.name</name>**

**<value>hdfs://localhost:54310</value>**

**<description>The name of the default file system. A URI whose**

**scheme and authority determine the FileSystem implementation. The**

**uri's scheme determines the config property (fs.SCHEME.impl) naming**

**the FileSystem implementation class. The uri's authority is used to**

**determine the host, port, etc. for a filesystem.</description>**

**</property>**

**</configuration>**

**4. /usr/local/hadoop/etc/hadoop/mapred-site.xml**

By default, the /usr/local/hadoop/etc/hadoop/ folder contains   
/usr/local/hadoop/etc/hadoop/mapred-site.xml.template   
file which has to be renamed/copied with the name mapred-site.xml:

**hduser@laptop:~$ cp /usr/local/hadoop/etc/hadoop/mapred-site.xml.template /usr/local/hadoop/etc/hadoop/mapred-site.xml**

**hduser@laptop:~$ nano /usr/local/hadoop/etc/hadoop/mapred-site.xml**

The mapred-site.xml file is used to specify which framework is being used for MapReduce.

**<configuration>**

**<property>**

**<name>mapred.job.tracker</name>**

**<value>localhost:54311</value>**

**<description>The host and port that the MapReduce job tracker runs**

**at. If "local", then jobs are run in-process as a single map**

**and reduce task.**

**</description>**

**</property>**

**</configuration>**

**5. /usr/local/hadoop/etc/hadoop/hdfs-site.xml**

The /usr/local/hadoop/etc/hadoop/hdfs-site.xml file needs to be configured for each host in the cluster that is being used.   
It is used to specify the directories which will be used as the namenode and thedatanode on that host.

Before editing this file, we need to create two directories which will contain the namenode and the datanode for this Hadoop installation.   
This can be done using the following commands:

**hduser@laptop:~$ sudo mkdir -p /usr/local/hadoop\_store/hdfs/namenode**

**hduser@laptop:~$ sudo mkdir -p /usr/local/hadoop\_store/hdfs/datanode**

**hduser@laptop:~$ sudo chown -R hduser:hadoop /usr/local/hadoop\_store**

**hduser@laptop:~$ nano /usr/local/hadoop/etc/hadoop/hdfs-site.xml**

**<configuration>**

**<property>**

**<name>dfs.replication</name>**

**<value>1</value>**

**<description>Default block replication.**

**The actual number of replications can be specified when the file is created.**

**The default is used if replication is not specified in create time.**

**</description>**

**</property>**

**<property>**

**<name>dfs.namenode.name.dir</name>**

**<value>file:/usr/local/hadoop\_store/hdfs/namenode</value>**

**</property>**

**<property>**

**<name>dfs.datanode.data.dir</name>**

**<value>file:/usr/local/hadoop\_store/hdfs/datanode</value>**

**</property>**

**</configuration>**

**Format the New Hadoop Filesystem**

Now, the Hadoop file system needs to be formatted so that we can start to use it. The format command should be issued with write permission since it creates **current** directory   
under **/usr/local/hadoop\_store/hdfs/namenode** folder:

**hduser@laptop:~$ hadoop namenode -format**

Note that hadoop namenode -format command should be executed once before we start using Hadoop. If this command is executed again after Hadoop has been used, it'll destroy all the data on the Hadoop file system.

**Starting Hadoop**

Now it's time to start the newly installed single node cluster.   
We can use **start-all.sh** or (**start-dfs.sh** and **start-yarn.sh**)

**hduser@laptop:~$ start-all.sh**

**This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh**

**15/04/18 16:43:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable**

**Starting namenodes on [localhost]**

**localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-namenode-laptop.out**

**localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-datanode-laptop.out**

**Starting secondary namenodes [0.0.0.0]**

**0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hduser-secondarynamenode-laptop.out**

**15/04/18 16:43:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable**

**starting yarn daemons**

**starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resourcemanager-laptop.out**

**localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduser-nodemanager-laptop.out**

We can check if it's really up and running:

**hduser@laptop:~$ jps**

**9026 NodeManager**

**7348 NameNode**

**9766 Jps**

**8887 ResourceManager**

**7507 DataNode**

**7350 Secondary Namenode**

**Stopping Hadoop**

We run **stop-all.sh** or (**stop-dfs.sh** and **stop-yarn.sh**) to stop all the daemons running on our machine:

**hduser@laptop:/usr/local/hadoop/sbin$ stop-all.sh**

**This script is Deprecated. Instead use stop-dfs.sh and stop-yarn.sh**

**15/04/18 15:46:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable**

**Stopping namenodes on [localhost]**

**localhost: stopping namenode**

**localhost: stopping datanode**

**Stopping secondary namenodes [0.0.0.0]**

**0.0.0.0: no secondarynamenode to stop**

**15/04/18 15:46:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable**

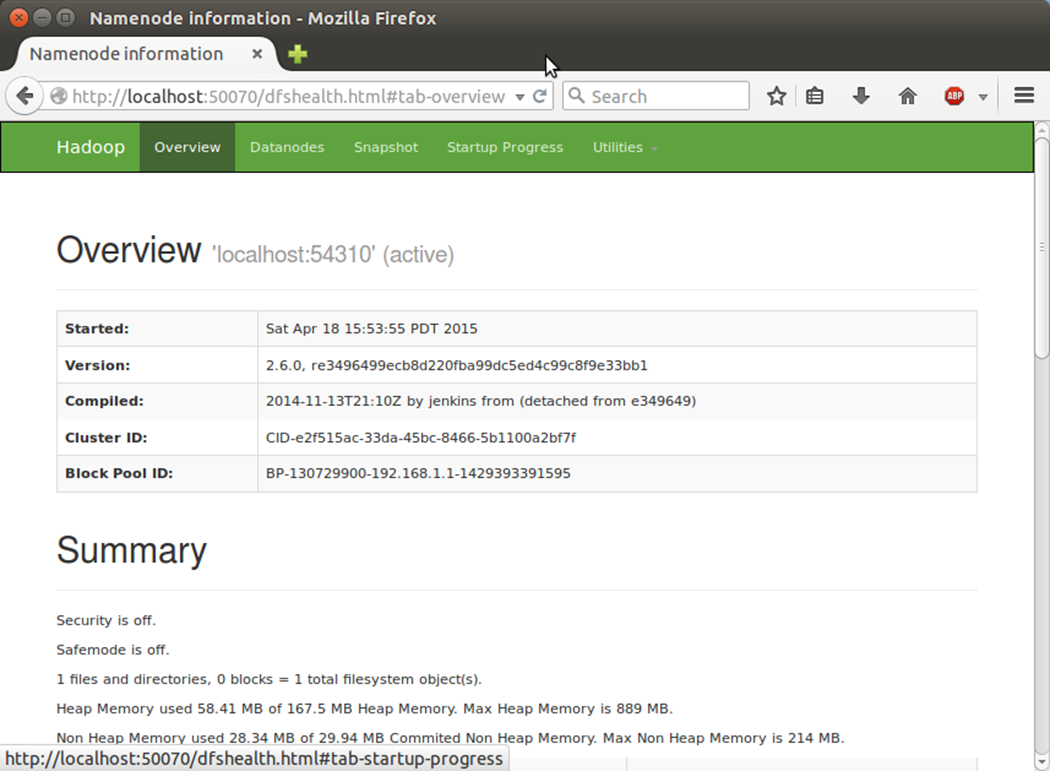
**stopping yarn daemons**

**stopping resourcemanager**

**localhost: stopping nodemanager**

**no proxyserver to stop**

**http://127.0.0.1:50070/ - web UI of the NameNode daemon**

****

**9)** P**ROJECT :** **H1-B Case Study**

**Project overview:** The H1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for H1B visa, an US employer must offer a job and petition for H1B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college/ higher education (Masters, Ph.D.) and work in a full-time position.

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

**Tasks:**

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 average growth in applications. [ALL]

1. 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 certified visa for each year. [certified]

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

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

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

b) for only certified applications.

1. 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.

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

1. 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.]

1. 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) ?

1. 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)?

1. Export result for question no 10 to MySql database.

**Dataset Sample and Explanation:**

1st Column: Sl.no

Data Eg: 1

Description: The serial number column offers unique numbers to each records.

2ndColumn: Case Status

Data Eg: CERTIFIED, CERTIFIED-WITHDRAWN, DENIED, WITHDRAWN.

Description: 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

3rd Column: Employer Name

Data Eg: University of Michigan

Description: Name of employer submitting labour condition application.

4thColumn: Soc\_name

Data Eg: BIOCHEMISTS AND BIOPHYSICISTS

Description: 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.

5th Column: Job Title

Data Eg: Post Doctoral Research Fellow

Description: Title of the job

6th Column: Full\_Time\_Position

Data Eg: Y,N

Description: Y = Full Time Position; N = Part Time Position

7th Column: Prevailing Wage

Data Eg:36067

Description: 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 intended employment. The prevailing wage is based on the employer’s minimum requirements for the position.

8th Column: Year

Data Eg:2011

Description: Year in which the H1B visa petition was filed

9th Column: Worksite

Data Eg: Ann Arbour, Michigan

Description: City and State information of the foreign worker’s intended area of employment

10th Column: Longitude

Data Eg:-83.7430378

Description: longitude of the Worksite

11th Column: Latitude

Data Eg:42.2808256

Description: latitude of the Worksite

**Cleansing the Dataset:**

**Introduction:**

The raw dataset provided had values enclosed within “” and columns separated by “,” and also the records may contain invalid or erroneous records like NA. Hence it is necessary to clean the Dataset and make is processable before carrying out the analysis.

Step 1:

The records from the raw dataset are loaded into a table h1b where the SERDE functions are used to separate the values with “”. Then the data is loaded into the table from local file system.

CREATE TABLE h1b\_applications(s\_no int,case\_status string, employer\_name string, soc\_name string, job\_title string,

full\_time\_position string,prevailing\_wage bigint,year string, worksite string, longitute double, latitute double )

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

"separatorChar" = ",",

"quoteChar" = "\""

) STORED AS TEXTFILE;

load data local inpath '/home/hduser/h1b.csv' overwrite into table h1b\_applications;

Step 2:

Another table h1b\_app2 is created with field delimiter “\t” and data is loaded from the h1b applications table but the field separators “,” is replaced by “\t” while loading the data. Then a condition is specified to avoid records with NA in the case status.

CREATE TABLE h1b\_app2(s\_no int,case\_status string, employer\_name

string, soc\_name string, job\_title string, full\_time\_position string,prevailing\_wage bigint,year string, worksite string, longitute double, latitute double ) row format delimited

fields terminated by '\t'

STORED AS TEXTFILE;

INSERT OVERWRITE TABLE h1b\_app2 SELECT regexp\_replace(s\_no, "\t", ""),

regexp\_replace(case\_status, "\t", ""), regexp\_replace(employer\_name,

"\t", ""), regexp\_replace(soc\_name, "\t", ""), regexp\_replace(job\_title, "\t", ""), regexp\_replace(full\_time\_position, "\t", ""), prevailing\_wage, regexp\_replace(year, "\t", ""), regexp\_replace(worksite, "\t", ""), regexp\_replace(longitute, "\t", ""), regexp\_replace(latitute, "\t",

"") FROM h1b\_applications where case\_status != "NA";

Step 3:

Another table h1b\_final is created and records from the previous table are loaded but the

case status column is filtered only for CERTIFIED, CERTIFIEDWITHDRAWN,DENIED AND WITHDRAWN. Then this table offers the final cleansed dataset with which all the analysis is carried out.

CREATE TABLE h1b\_final(s\_no int,case\_status string, employer\_name

string, soc\_name string, job\_title string, full\_time\_position string,prevailing\_wage bigint,year string, worksite string, longitute double, latitute double ) row format delimited fields terminated by '\t'

STORED AS TEXTFILE;

INSERT OVERWRITE TABLE h1b\_final SELECT s\_no,

case when trim(case\_status) = "PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED" then "DENIED"

when trim(case\_status) = "REJECTED" then "DENIED"

when trim(case\_status) = "INVALIDATED" then "DENIED" else case\_status end,

employer\_name, soc\_name, job\_title, full\_time\_position, case when prevailing\_wage is null then 100000 else prevailing\_wage end,

year, worksite, longitute, latitute

FROM h1b\_app2;

**Solution and Explanation:**

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

**Code:**

import org.apache.hadoop.io.Text;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.conf.\*;

import org.apache.hadoop.fs.\*;

import org.apache.hadoop.mapreduce.lib.input.\*;

import org.apache.hadoop.mapreduce.lib.output.\*;

**Mapper code**

public class proj1

{

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

{

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

{

try{

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

//String pet = str[4];

//String year =(str[7]);

if(str[4].equals("DATA ENGINEER"))

{

context.write(new Text(str[7]),new Text(str[4]));

}

}

catch(Exception e)

{

System.out.println(e.getMessage());

}

}

}

**Reducer Code**

public static class ReduceClass extends Reducer<Text,Text,Text,LongWritable>

{

// private LongWritable result = new LongWritable();

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

{

long count=0;

//String job1 ="";

//String case\_status="";

for (Text val : values)

{

count++;

}

context.write(key, new LongWritable (count));

}

}

**//Driver Code**

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

Configuration conf = new Configuration();

//conf.set("name", "value")

//conf.set("mapreduce.input.fileinputformat.split.minsize", "134217728");

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

job.setJarByClass(proj1.class);

job.setMapperClass(MapClass.class);

//job.setCombinerClass(ReduceClass.class);

job.setReducerClass(ReduceClass.class);

job.setNumReduceTasks(2);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(Text.class);

job.setOutputKeyClass(Text.class);

job.setOutputValueClass(LongWritable.class);

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

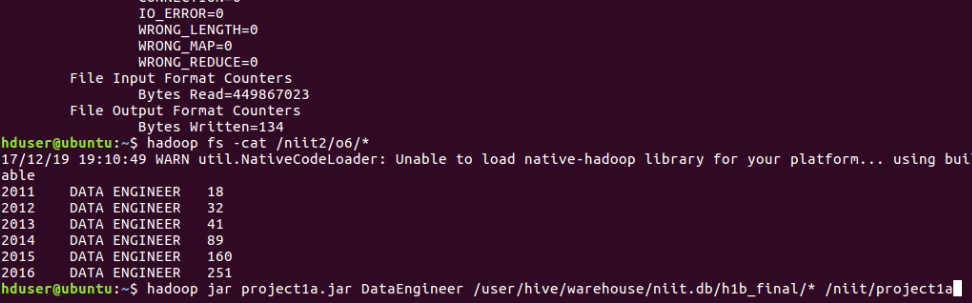
FileOutputFormat.setOutputPath(job, new Path(args[1]));

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

}

}

OUTPUT:

****

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

Solution: This question has been solved using Pig Latin.

Code:

h1b\_final= load '/user/hive/warehouse/project.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:double,year:chararray,worksite:chararray,longitute:double,latitute:double);

filjob = filter h1b\_final by year=='2011';--dump filjob;grpjt = group filjob by $4;grpcs1 = foreach grpjt generate group, COUNT(filjob.$1);--dump grpcs;filjob = filter h1b\_final by year=='2012';--dump filjob;

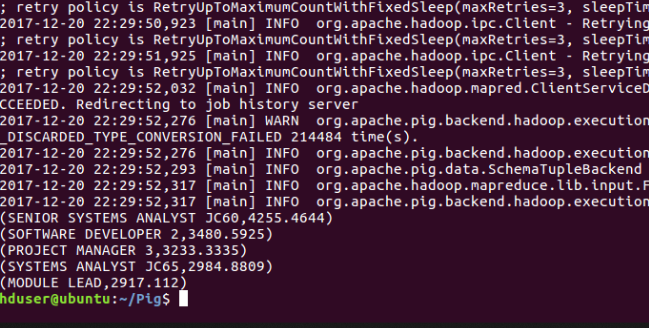
grpjt = group filjob by $4;grpcs2 = foreach grpjt generate group, COUNT(filjob.$1);

filjob = filter h1b\_final by year=='2013';

--dump filjob;grpjt = group filjob by $4;grpcs3 = foreach grpjt generate group, COUNT(filjob.$1);filjob = filter h1b\_final by year=='2014';--dump filjob;grpjt = group filjob by $4;

grpcs4 = foreach grpjt generate group, COUNT(filjob.$1);filjob = filter h1b\_final by year=='2015';--dump filjob;grpjt = group filjob by $4;grpcs5 = foreach grpjt generate group, COUNT(filjob.$1);filjob = filter h1b\_final by year=='2016';--dump filjob;

grpjt = group filjob by $4;grpcs6 = foreach grpjt generate group, COUNT(filjob.$1);jngr = join grpcs1 by $0 ,grpcs2 by $0 ,grpcs3 by $0 ,grpcs4 by $0 ,grpcs5 by $0 ,grpcs6 by $0;--dump jngr;jngr = foreach jngr generate $0 , $1 , $3 ,$5 ,$7 ,$9 ,$11;growth = foreach jngr generate $0 , (float)(($2-$1)/$1)\*100, (float)(($3-$2)/$2)\*100,(float)(($4-$3)/$3)\*100, (float)(($5-$4)/$4)\*100, (float)(($6-$5)/$5)\*100;avggw = foreach growth generate $0 , ($1+$2+$3+$4+$5)/5;lm = limit (order avggw by $1 desc) 5;dump lm;



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

Solution: This problem has been solved using the Map Reduce method using Java programming

Code:

import java.io.\*;

import java.util.TreeMap;

import org.apache.hadoop.io.NullWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.io.LongWritable;

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.conf.\*;

import org.apache.hadoop.fs.\*;

import org.apache.hadoop.mapreduce.lib.input.\*;

import org.apache.hadoop.mapreduce.lib.output.\*;

public class proj2a {

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

{

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

{

try{

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

if((str[4].equals("DATA ENGINEER"))&&(str[1].equals("CERTIFIED")))

{

String c = str[4]+"\t"+str[7];

context.write(new Text(str[8]),new Text(c));

}

}

catch(Exception e)

{

System.out.println(e.getMessage());

}

}

}

public static class yearPartitioner extends Partitioner<Text,Text>

{

public int getPartition(Text key, Text values, int numReduceTasks) {

//String b[]="";

String[] b=values.toString().split("\t");

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

{

return 0;

}

else if(b[1].equals("2012"))

{

return 1;

}

else if(b[1].equals("2013"))

{

return 2;

}

else if(b[1].equals("2014"))

{

return 3;

}

else if(b[1].equals("2015"))

{

return 4;

}

else

{

return 5;

}

}

}

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

{

public TreeMap<Long, Text> tm = new TreeMap<Long, Text>();

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

{

long count=0;

//String year="";

//String job="";

String myVal="";

for(Text val:values)

{

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

count++;

myVal = str[1]+"\t"+key+"\t"+str[0];

}

String myValue = myVal+"\t"+count;

tm.put(new Long(count), new Text(myValue));

if(tm.size()>1)

{

tm.remove(tm.firstKey());

}

}

public void cleanup(Context con) throws IOException, InterruptedException

{

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

{

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

}

}

}

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

Configuration conf = new Configuration();

//conf.set("name", "value")

//conf.set("mapreduce.input.fileinputformat.split.minsize", "134217728");

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

job.setJarByClass(proj2a.class);

job.setMapperClass(MapClass.class);

job.setPartitionerClass(yearPartitioner.class);

//job.setCombinerClass(ReduceClass.class);

job.setReducerClass(ReduceClass.class);

job.setNumReduceTasks(6);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(Text.class);

job.setOutputKeyClass(NullWritable.class);

job.setOutputValueClass(Text.class);

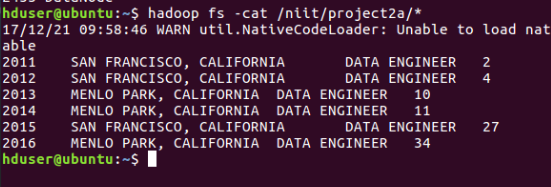
FileInputFormat.addInputPath(job, new Path(args[0]));

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

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

}

}



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

Solution:

Code:

import java.io.\*;

import java.util.TreeMap;

import org.apache.hadoop.io.NullWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.io.LongWritable;

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.conf.\*;

import org.apache.hadoop.fs.\*;

import org.apache.hadoop.mapreduce.lib.input.\*;

import org.apache.hadoop.mapreduce.lib.output.\*;

public class proj2b {

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

{

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

{

try{

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

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

{

String a = str[1]+"\t"+str[7];

context.write(new Text(str[8]),new Text(a));

}

}

catch(Exception e)

{

System.out.println(e.getMessage());

}

}

}

public static class yearPartitioner extends Partitioner<Text,Text>

{

public int getPartition(Text key, Text values, int numReduceTasks) {

//String b[]="";

long c=0;

String[] b=values.toString().split("\t");

c = Long.parseLong(b[1]);

if(c==2011)

{

return 0;

}

else if(c==2012)

{

return 1;

}

else if(c==2013)

{

return 2;

}

else if(c==2014)

{

return 3;

}

else if(c==2015)

{

return 4;

}

else

{

return 5;

}

}

}

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

{

// private LongWritable result = new LongWritable();

public TreeMap<Long, Text> tm = new TreeMap<Long, Text>();

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

{

long count=0;

//String year="";

//String job="";

String myVal="";

for(Text val:values)

{

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

count++;

myVal = str[1]+"\t"+key;

}

String myValue1 = myVal+"\t"+count;

tm.put(new Long(count), new Text(myValue1));

if(tm.size()>5)

{

tm.remove(tm.firstKey());

}

}

public void cleanup(Context con) throws IOException, InterruptedException

{

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

{

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

}

}

}

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

Configuration conf = new Configuration();

//conf.set("name", "value");

//conf.set("mapreduce.input.fileinputformat.split.minsize", "134217728");

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

job.setJarByClass(proj2b.class);

job.setMapperClass(MapClass.class);

job.setPartitionerClass(yearPartitioner.class);

//job.setCombinerClass(ReduceClass.class);

job.setReducerClass(ReduceClass.class);

job.setNumReduceTasks(6);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(Text.class);

job.setOutputKeyClass(NullWritable.class);

job.setOutputValueClass(Text.class);

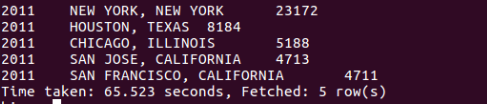
FileInputFormat.addInputPath(job, new Path(args[0]));

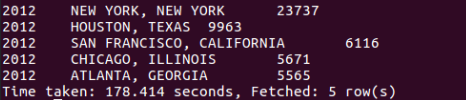
FileOutputFormat.setOutputPath(job, new Path(args[1]));

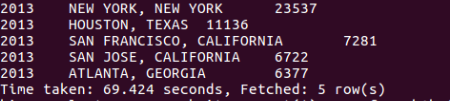
System.exit(job.waitForCompletion(true) ? 0 : 1);

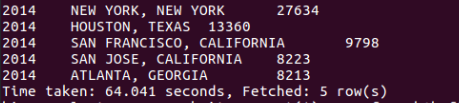
}

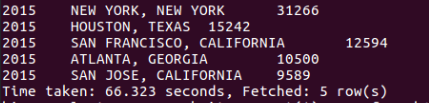
}

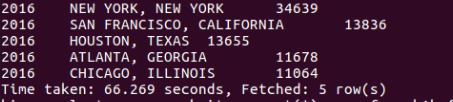










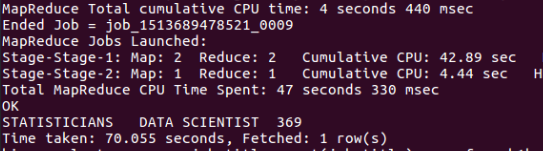


3) Which industry(SOC\_NAME) has the most number of Data Scientist positions? [certified]

Solution: This query is solved using hive with SQL programming

Code:

select SOC\_NAME,count(case\_status) as ind from h1b\_final where case\_status = 'certified' and job\_title ='Data scientist' group by SOC\_NAME order by ind desc limit 1 ;



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

solution: This query is solved using hive with SQL programming.

Code:

select employer\_name,year,count(case\_status) as top from h1b\_final where year = '2011'

group by employer\_name,year order by top desc limit 5;

select employer\_name,year,count(case\_status) as top from h1b\_final where year = '2012' group by employer\_name,year order by top desc limit 5;

select employer\_name,year,count(case\_status) as top from h1b\_final where year = '2013' group by employer\_name,year order by top desc limit 5;

select employer\_name,year,count(case\_status) as top from h1b\_final where year = '2014' group by employer\_name,year order by top desc limit 5;

select employer\_name,year,count(case\_status) as top from h1b\_final where year = '2015' group by employer\_name,year order by top desc limit 5;

select employer\_name,year,count(case\_status) as top from h1b\_final where year = '2016' group by employer\_name,year order by top desc limit 5;



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

Solution: This query is solved using hive with SQL programming.

Code:

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

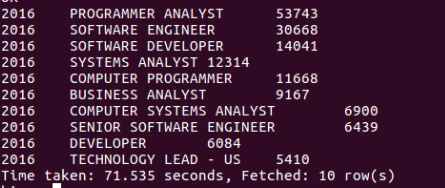
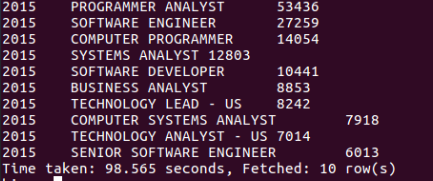
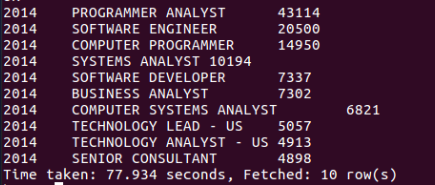
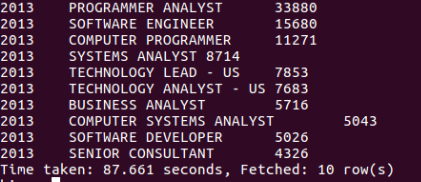
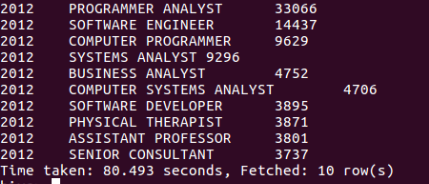
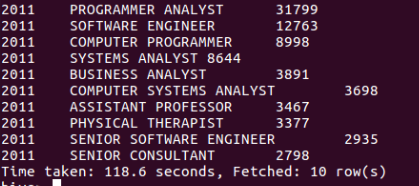
select job\_title,year,count(case\_status ) as pop from h1b\_final where year = '2012' group by job\_title,year order by pop desc limit 10;

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

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

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

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



5) b) Find the most popular top 10 job positions for H1B visa applications for each year? For only certified applications.

Solution: This query is solved using hive with SQL programming.

Code:

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

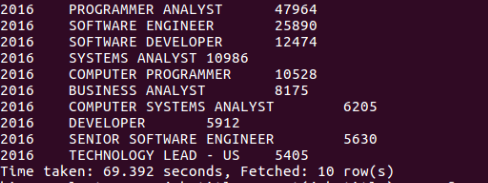
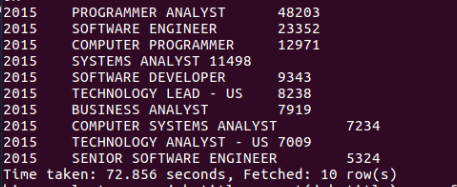
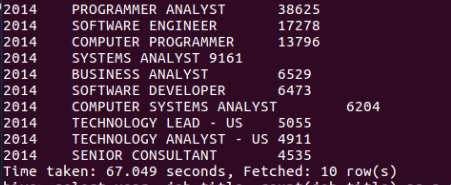
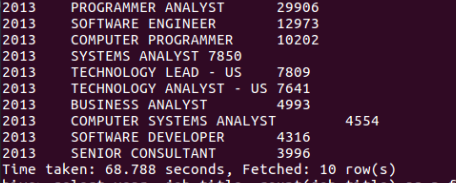
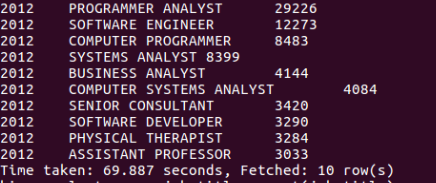
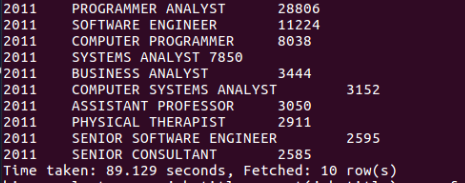
select job\_title,year,count(case\_status ) as pop from h1b\_final where year = '2012' and case\_status='CERTIFIED' group by job\_title,year order by pop desc limit 10;

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

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

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

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



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.

Solution: This question has been solved using Pig Latin.

Code:

h1b\_final = load '/user/hive/warehouse/project.db/h1b\_final' using PigStorage('\t') as (sno:int, case\_status:chararray, emp\_name:chararray, soc\_name:chararray, job\_title:chararray, full\_time\_pos:chararray, wage:long, year:chararray, worksite:chararray, longitude:double, lattitude:double);

year\_grp = group h1b\_final by $7;

total = foreach year\_grp generate group, (float)COUNT(h1b\_final.$1) as tot;

--dump total;

grp = group h1b\_final by ($7,$1);

case\_total = foreach grp generate flatten(group), (float)COUNT(h1b\_final.$1) as per;

--dump case\_total;

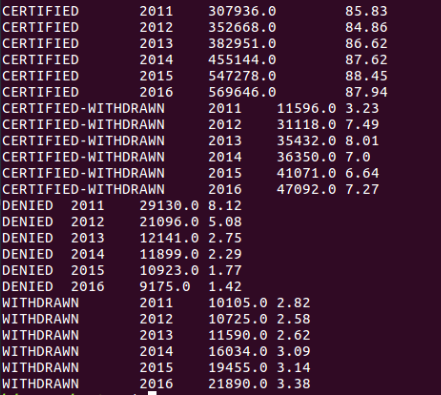
join\_grp = join total by $0,case\_total by $0;

--dump join\_grp;

final = foreach join\_grp generate $0,$3,$4,ROUND\_TO(($4/$1)\*100,2);

--dump final;

store final into '/home/hduser/Documents/project pig querries/project6';



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

Solution: This problem has been solved using the Map Reduce method using Java programming.

Code:

import java.io.\*;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.conf.\*;

import org.apache.hadoop.fs.\*;

import org.apache.hadoop.mapreduce.lib.input.\*;

import org.apache.hadoop.mapreduce.lib.output.\*;

public class proj7{

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

{

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

{

try{

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

String app = str[1];

String year =(str[7]);

context.write(new Text(year),new Text(app));

}

catch(Exception e)

{

System.out.println(e.getMessage());

}

}

}

public static class ReduceClass extends Reducer<Text,Text,Text,LongWritable>

{

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

{

long count=0;

//String job1 ="";

//String case\_status="";

for (Text val : values)

{

count++;

}

context.write(key, new LongWritable (count));

}

}

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

Configuration conf = new Configuration();

//conf.set("name", "value")

//conf.set("mapreduce.input.fileinputformat.split.minsize", "134217728");

Job job = Job.getInstance(conf, "app Count");

job.setJarByClass(proj7.class);

job.setMapperClass(MapClass.class);

//job.setCombinerClass(ReduceClass.class);

job.setReducerClass(ReduceClass.class);

job.setNumReduceTasks(2);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(Text.class);

job.setOutputKeyClass(Text.class);

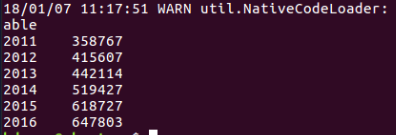
job.setOutputValueClass(LongWritable.class);

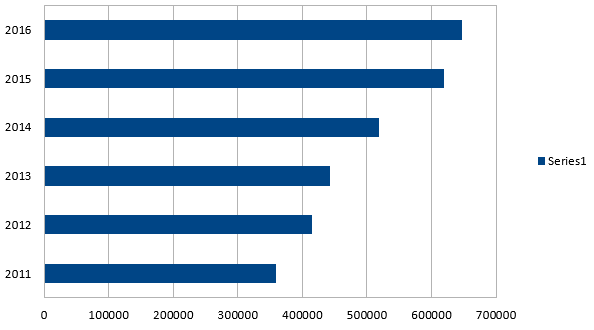
FileInputFormat.addInputPath(job, new Path(args[0]));

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

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

}}





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.]

Code:

select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='Y' and year='2011' group by job\_title,full\_time\_position,year order by average desc;

select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='N' and year='2011' group by job\_title,full\_time\_position,year order by average desc;

select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='Y' and year='2012' group by job\_title,full\_time\_position,year order by average desc;

select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='N' and year='2012' group by job\_title,full\_time\_position,year order by average desc;

select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='Y' and year='2013' group by job\_title,full\_time\_position,year order by average desc;

select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='N' and year='2013' group by job\_title,full\_time\_position,year order by average desc;

select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='Y' and year='2014' group by job\_title,full\_time\_position,year order by average desc;

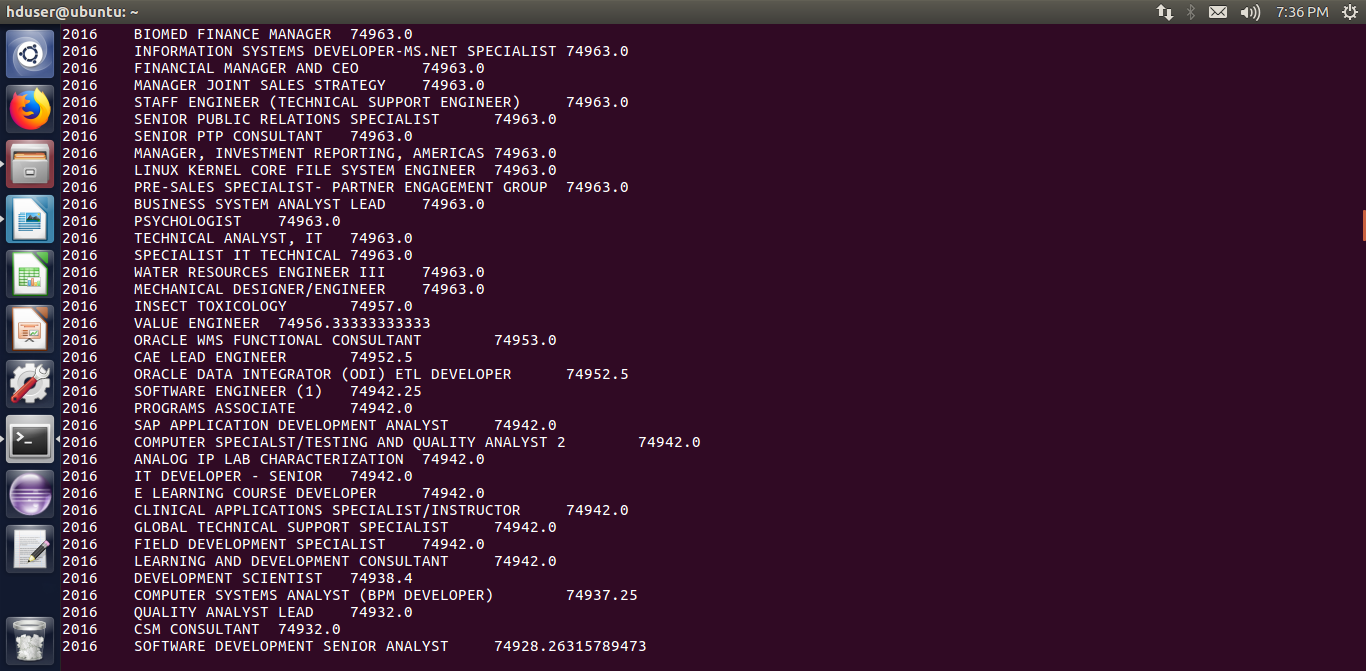
select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='N' and year='2014' group by job\_title,full\_time\_position,year order by average desc;

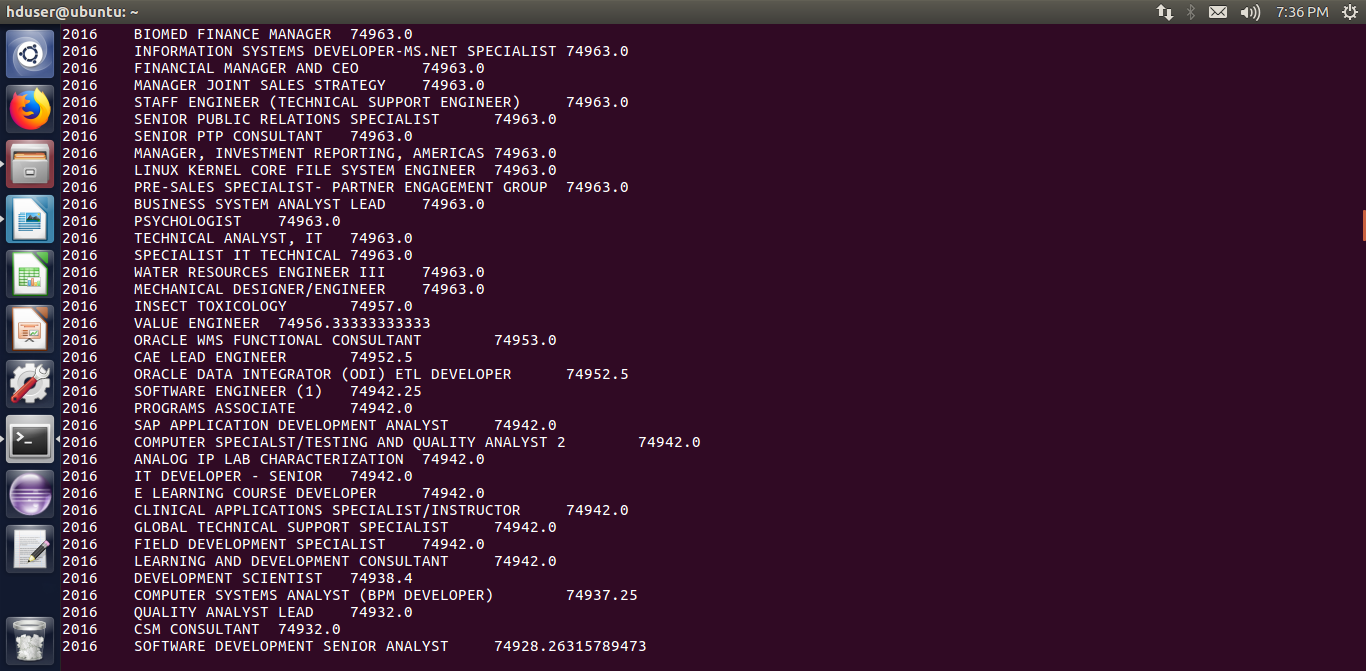
select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='Y' and year='2015' group by job\_title,full\_time\_position,year order by average desc;

select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='N' and year='2015' group by job\_title,full\_time\_position,year order by average desc;

select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='Y' and year='2016' group by job\_title,full\_time\_position,year order by average desc;

select job\_title, full\_time\_position, year, avg(prevailing\_wage) as average from h1b\_final where full\_time\_position ='N' and year='2016' group by job\_title,full\_time\_position,year order by average desc;





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)

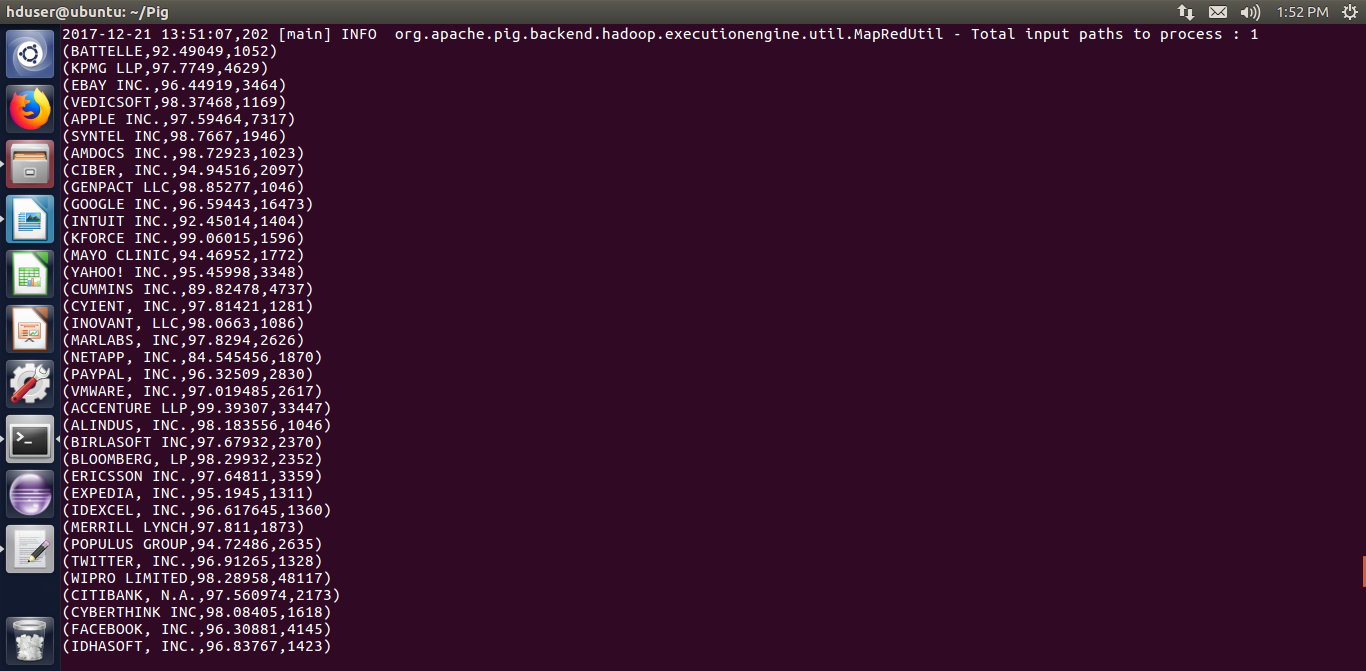
h1b\_final= load '/user/hive/warehouse/project.db/h1b\_final' using PigStorage('\t') AS (s\_no:int,case\_status:chararray,employer\_name:chararray,soc\_name:chararray,job\_title:chararray,full\_time\_position:chararray,prevailing\_wage:double,year:double,worksite:chararray,longitute:double,latitute:double);

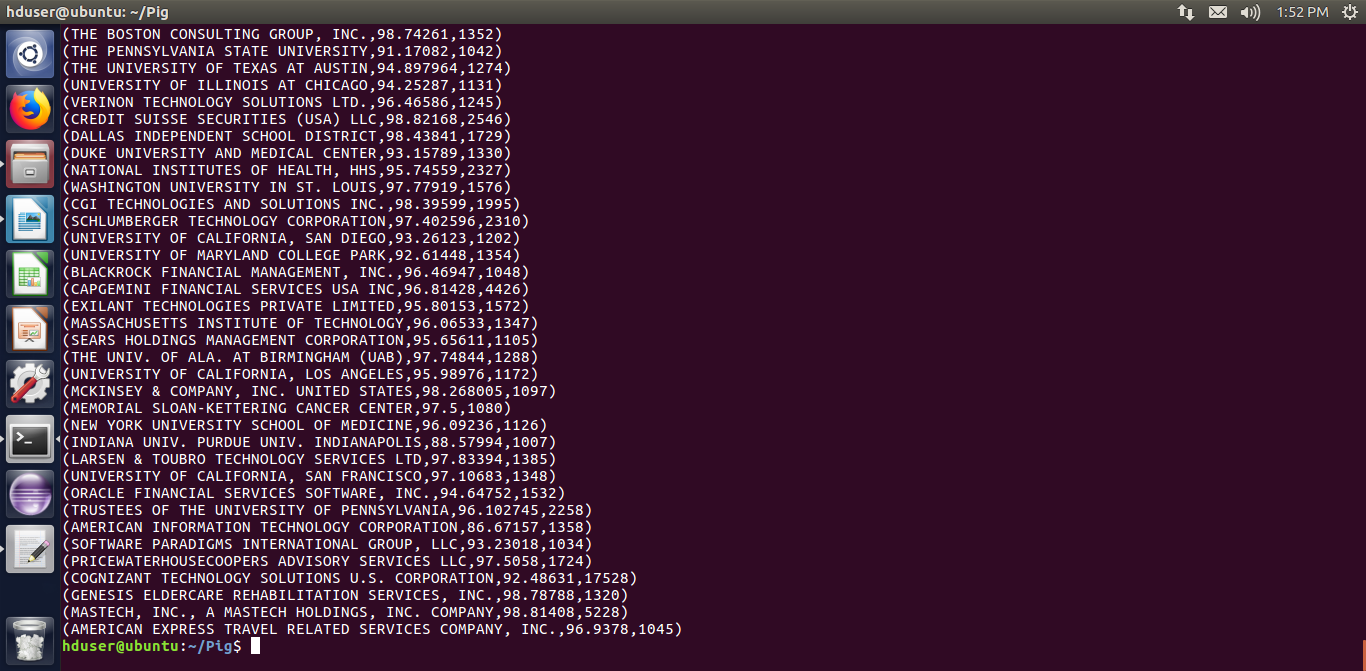
empgrp = GROUP h1b\_final BY $2;--dump empgrp;wordcount = FOREACH empgrp GENERATE group, COUNT(h1b\_final.$1);--dump wordcount;

casefil = filter h1b\_final by case\_status=='CERTIFIED' or case\_status=='CERTIFIED-WITHDRAWN';

case\_status1 = GROUP casefil by $2; wordcount1 = FOREACH case\_status1 GENERATE group, COUNT(casefil.$1);--dump wordcount1;joinb = join wordcount by $0,wordcount1 by $0;--dump joinbag;out = FOREACH joinb GENERATE $0,(float)$3/$1\*100, $1;--dump out;out1 = filter out by $1>70 and $2>=1000;

dump out1;





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)?

Solution: This question has been solved using Pig Latin.

Code:

h1b\_final= load '/user/hive/warehouse/project.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:double,year:double,worksite:chararray,longitute:double,latitute:double);

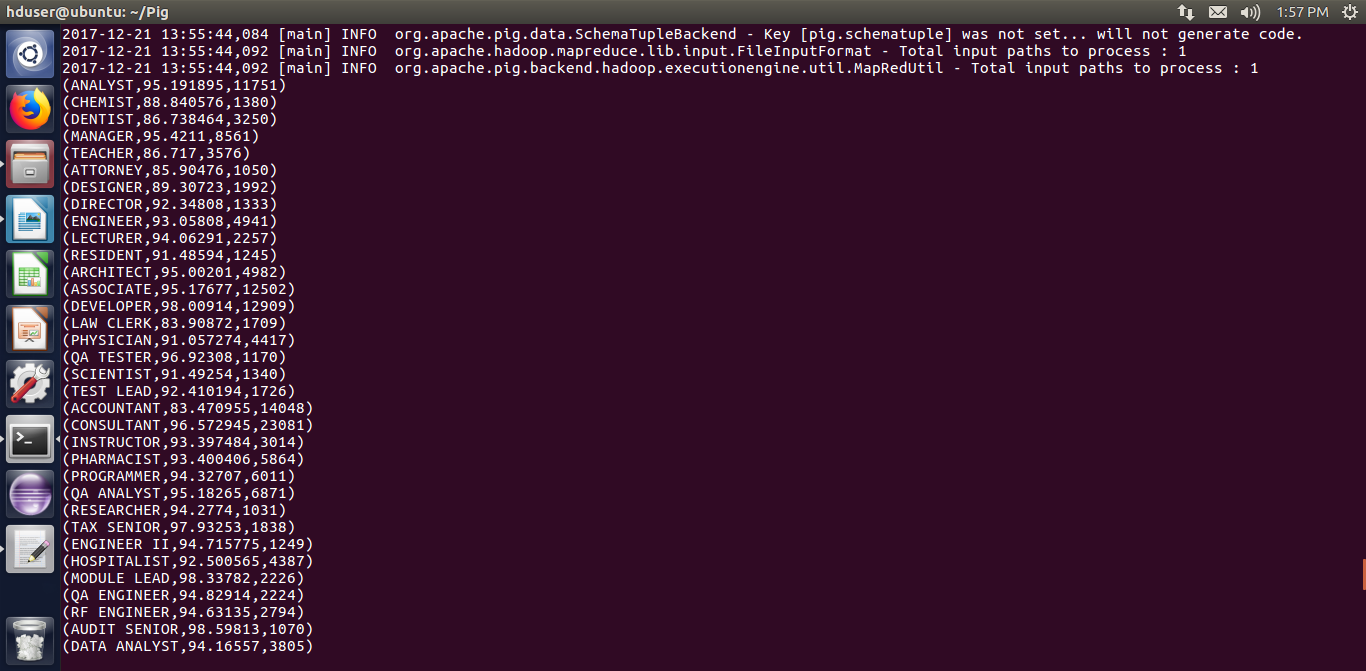
empgrp = GROUP h1b\_final BY $4;--dump empgrp;wordcount = FOREACH empgrp GENERATE group, COUNT(h1b\_final.$1);--dump wordcount;

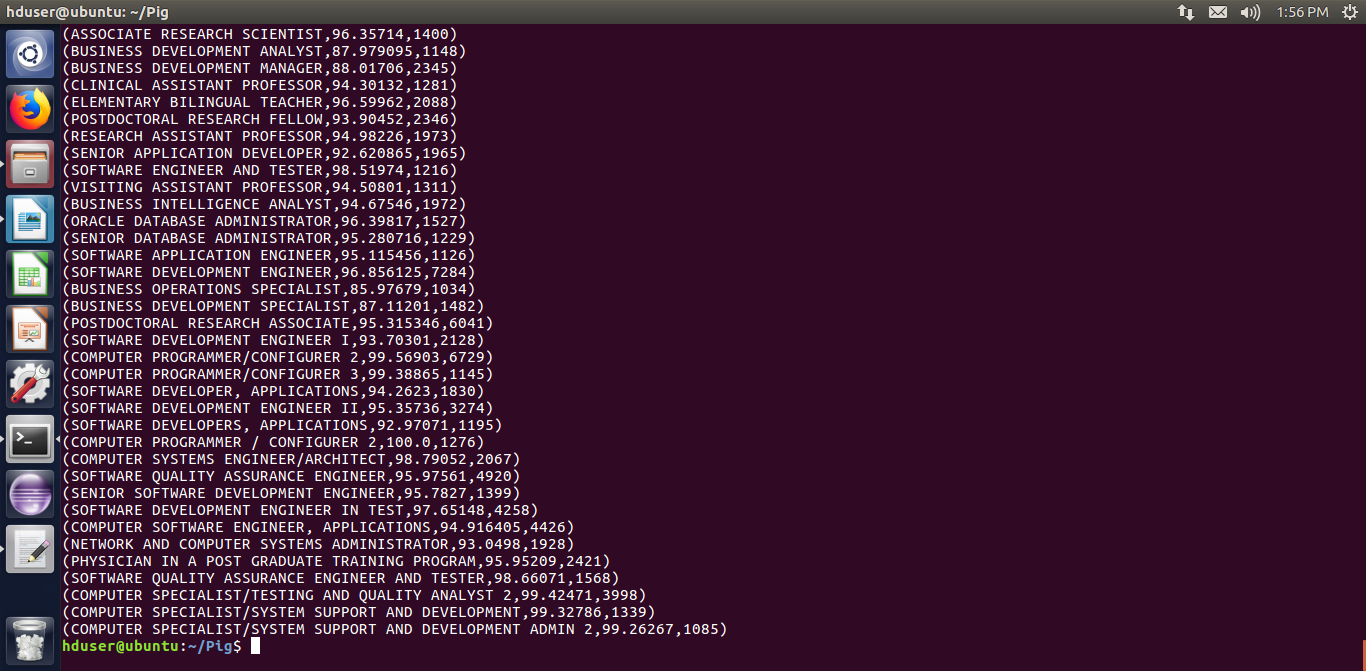
casefil = filter h1b\_final by case\_status=='CERTIFIED' or case\_status=='CERTIFIED-WITHDRAWN';case\_status1 = GROUP casefil by $4; wordcount1 = FOREACH case\_status1 GENERATE group, COUNT(casefil.$1);--dump wordcount1;

joinb = join wordcount by $0,wordcount1 by $0;--dump joinbag;

out = FOREACH joinb GENERATE $0,(float)$3/$1\*100, $1;--dump out;out1 = filter out by $1>70 and $2>=1000;

--dump out1; store out1 into '/niit/project10';





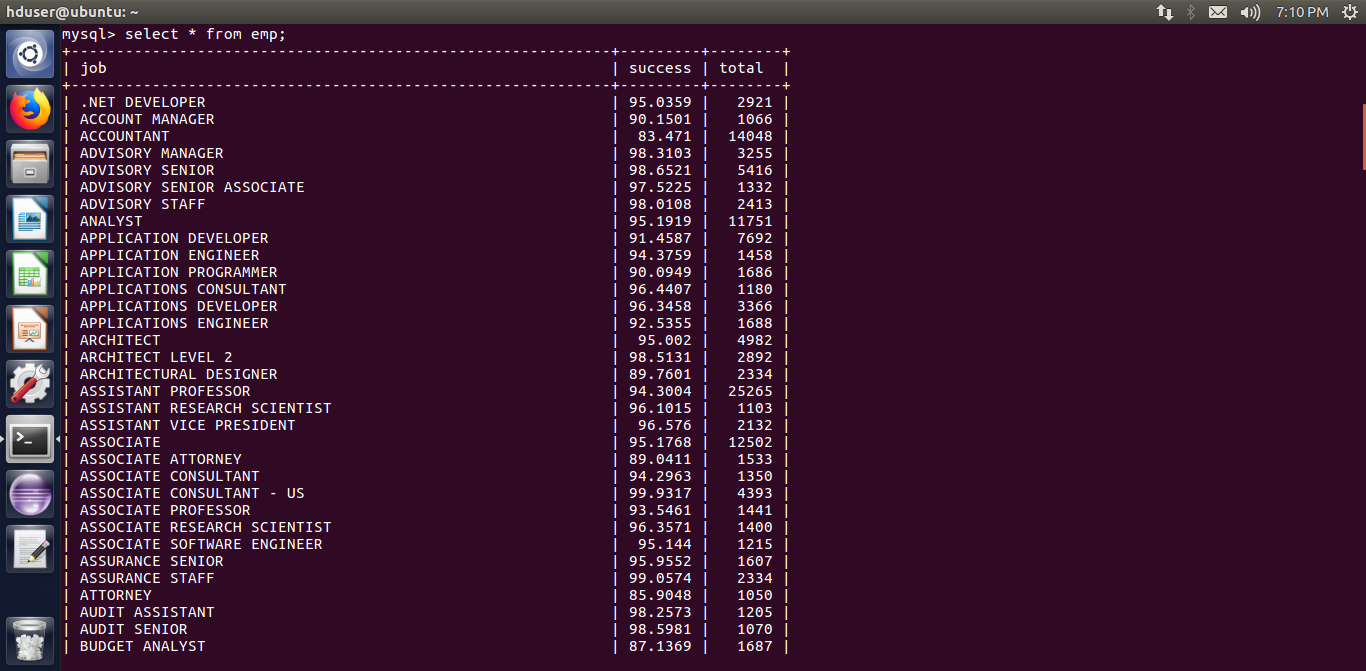
11) Export result for question no 10 to MySql database.

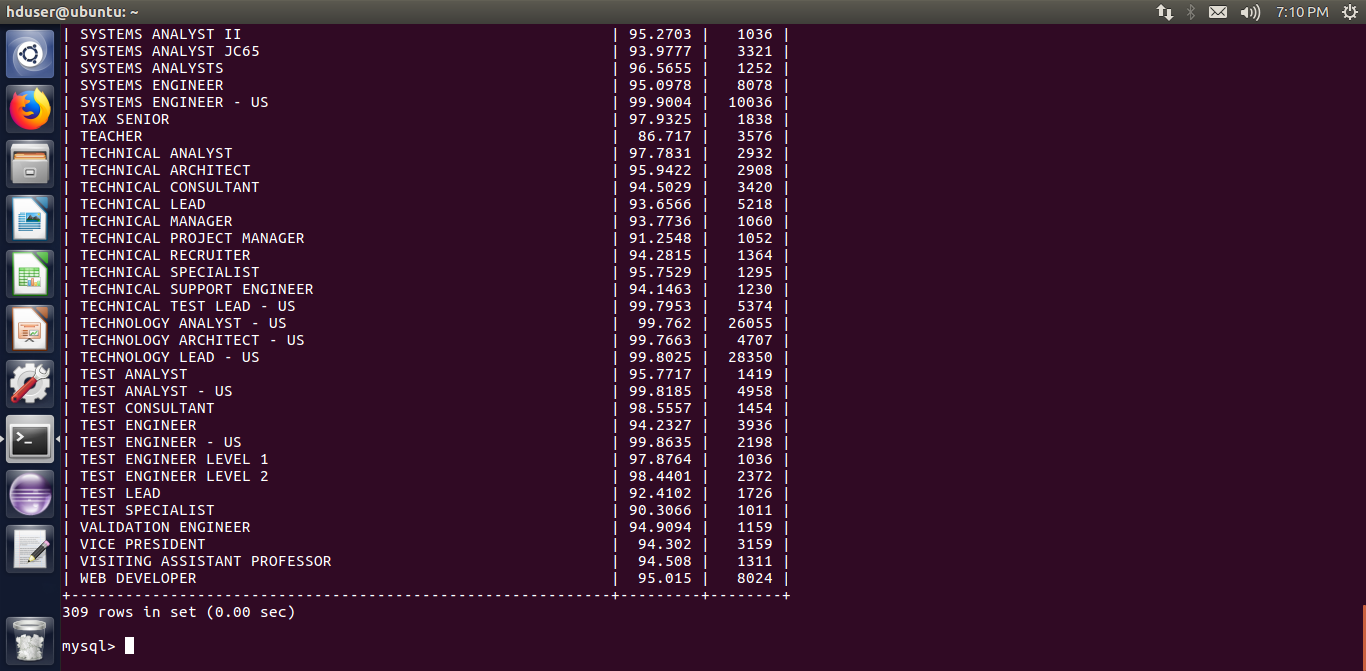
Solution: This query is solved using Sqoop Export command

Code:

create table sq (job VARCHAR(100) NOT NULL, percentage FLOAT NOT NULL,total INT NOT NULL, PRIMARY KEY (job));

sqoop export --connect jdbc:mysql://localhost/project --username root --password '123' --table sq --update-mode allowinsert --update-key job --export-dir /project%2010 --input-fields-terminated-by '\t' ;





Shell Scripting

Overview:

The developed analysis is offered to the user by means of a menu based interface which provides solution to the selected query. The code below offers the menu and acts as an interface between the user and computer. The menu is run on the linux terminal and the options are selected to display the analysis results.

#!/bin/bash

show\_menu()

{

NORMAL=`echo "\033[m"`

MENU=`echo "\033[36m"` #Blue

NUMBER=`echo "\033[33m"` #yellow

FGRED=`echo "\033[41m"`

RED\_TEXT=`echo "\033[31m"`

ENTER\_LINE=`echo "\033[33m"`

echo -e "${MENU}\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*APP MENU\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 1)${MENU} Question 1a ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 2)${MENU} Question 1b ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 3)${MENU} Question 2a ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 4)${MENU} Question 2b ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 5)${MENU} Question 3 ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 6)${MENU} Question 4 ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 7)${MENU} Question 5a ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 8)${MENU} Question 5b ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 9)${MENU} Question 6 ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 10)${MENU} Question 7 ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 11)${MENU} Question 8 ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 12)${MENU} Question 9 ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 13)${MENU} Question 10 ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 14)${MENU} Question 11 ${NORMAL}"

echo -e "${MENU}\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*${NORMAL}"

echo -e "${ENTER\_LINE}Please enter a menu option and enter or ${RED\_TEXT}enter to exit. ${NORMAL}"

read opt

}

function option\_picked()

{

COLOR='\033[01;31m' # bold red

RESET='\033[00;00m' # normal white

MESSAGE="$1" #modified to post the correct option selected

echo -e "${COLOR}${MESSAGE}${RESET}"

}

function getpinCodeBank(){

echo "in getPinCodebank"

echo $1

echo $2

#hive -e "Select \* from AppData where PinCode = $1 AND Bank = '$2'"

}

clear

show\_menu

while [ opt != '' ]

do

if [[ $opt = "" ]]; then

exit;

else

case $opt in

1) clear;

option\_picked "1a) Is the number of petitions with Data Engineer job title increasing over time?";

bash /home/hduser/ProjectH1BFinal/1a.sh

show\_menu;

;;

2) clear;

option\_picked "1b) Find top 5 job titles who are having highest avg growth in applications.";

pig /home/hduser/1b.pig

show\_menu;

;;

3) clear;

option\_picked "2a) Which part of the US has the most Data Engineer jobs for each year?";

bash /home/hduser/ProjectH1BFinal/2a.sh

show\_menu;

;;

4) clear;

option\_picked "2b) find top 5 locations in the US who have got certified visa for each year";

bash /home/hduser/ProjectH1BFinal/2b.sh

show\_menu;

;;

5) clear;

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

hive -e "select SOC\_NAME,count(case\_status) as ind from project.h1b\_final where case\_status = 'certified' and job\_title ='Data scientist' group by SOC\_NAME order by ind desc limit 1 ";

show\_menu;

;;

6) clear;

option\_picked "4)Which top 5 employers file the most petitions each year?";

echo "Enter the year"

read year

echo "You've selected ${year}"

hive -e "select employer\_name,year,count(case\_status) as top from project.h1b\_final where year=$year group by employer\_name,year order by top desc limit 5;"

show\_menu;

;;

7) clear;

option\_picked "5) Find the most popular top 10 job positions for H1B visa applications for each year?a) for all the applications";

echo "Enter the year"

read year

echo "You've selected ${year}"

hive -e "select job\_title,year,count(case\_status ) as pop from project.h1b\_final where year=$year group by job\_title,year order by pop desc limit 10;"

show\_menu;

;;

8) clear;

option\_picked "5) Find the most popular top 10 job positions for H1B visa applications for each year?b) for only certified applications.";

echo "Enter the year"

read year

echo "You've selected ${year}"

hive -e "select job\_title,year,count(case\_status ) as pop from project.h1b\_final where year=$year and case\_status='CERTIFIED' group by job\_title,year order by pop desc limit 10;"

show\_menu;

;;

9) clear;

option\_picked "6) Find the percentage and the count of each case status on total applications for each year. Create a line graph depicting the pattern of All the cases over the period of time.";

pig /home/hduser/6.pig

show\_menu;

;;

10) clear;

option\_picked "7) Create a bar graph to depict the number of applications for each year";

bash /home/hduser/ProjectH1BFinal/7.sh

show\_menu;

;;

11) clear;

option\_picked "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";

echo -e "${MENU}Select Full Time Job or Part Time Job ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 1)${MENU} Full Time Job ${NORMAL}"

echo -e "${MENU}\*\*${NUMBER} 2)${MENU} Part Time Job ${NORMAL}"

read job

case $job in

1) echo "FULL TIME JOB SELECTED"

pig /home/hduser/project/project8y.pig

;;

2) echo "PART TIME JOB SELECTED"

pig /home/hduser/Proj/Proj8n.pig

;;

\*) echo "Please Select one among the option[1-2]";;

esac

show\_menu;

;;

12) clear;

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

pig /home/hduser/9.pig

show\_menu;

;;

13) clear;

option\_picked "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)?";

pig /home/hduser/10.pig

show\_menu;

;;

14) clear;

option\_picked "11) Export result for question no 10 to MySql database.";

sqoop export --connect jdbc:mysql://localhost/project10 --username root --P --table emp --update-mode allowinsert --update-key job --export-dir /niit/project10/\* --input-fields-terminated-by '\t' ;

show\_menu;

;;

\n) exit;

;;

\*) clear;

option\_picked "Pick an option from the menu";

show\_menu;

;;

esac

fi

done