H1-B US VISA DATA ANALYSIS

BIG DATA HADOOP

G SOWMYA LAKSHMI | Professional Diploma in Digital Transformation-Big Data with Hadoop | 02.05.2017

Tools used: Apache Hadoop Framework- HDFS, Mapreduce Framework,HiveQL,Pig, Sqoop and MSExcel for generating graphs.

S170030600329
NIIT Limited

NIIT

H1-B US VISA DATA ANALYSIS

A PROJECT REPORT

Submitted by

G SOWMYA LAKSHMI

in the partial fulfillment for the award of the
of

PROFESSIONAL DIPLOMA
in

BIG DATA WITH HADOOP

NIIT, CHENNAI

MAY 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. Kotteswari** 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 means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, technqiues and frameworks. To understand the phenomenon that is big data, it is often described using five Vs: Volume, Velocity, Variety, Veracity and Value.

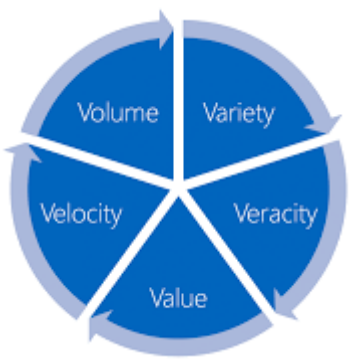
-**Volume**: It refers to the vast amounts of data generated every second. Just think of all the emails, twitter messages, photos, video clips, sensor data etc.

-**Variety** refers to the different types of data we can now use. In the past we focused on structured data .In fact, 80% of the world's data is now unstructured, and therefore can't easily be put into tables (think of photos, video sequences or social media updates).

-**Velocity** refers to the speed at which new data is generated and the speed at which data moves around

-**Veracity** refers to the messiness or trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content)

-**Value**:It is all well and good having access to big data but unless we can turn it into value it is useless.

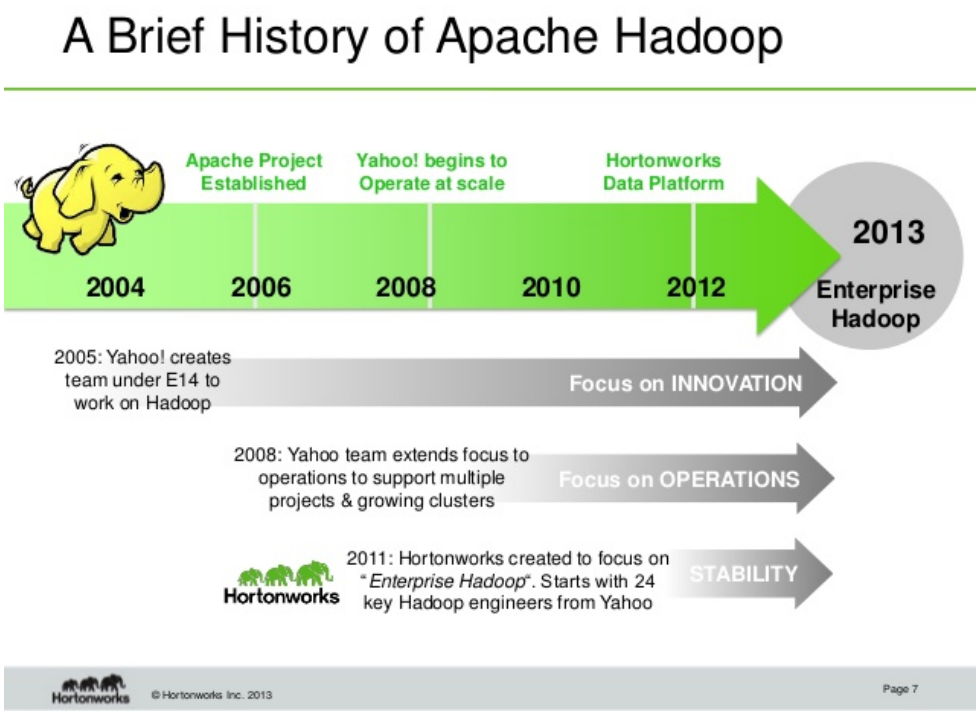


Apache Hadoop

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.



Benefits

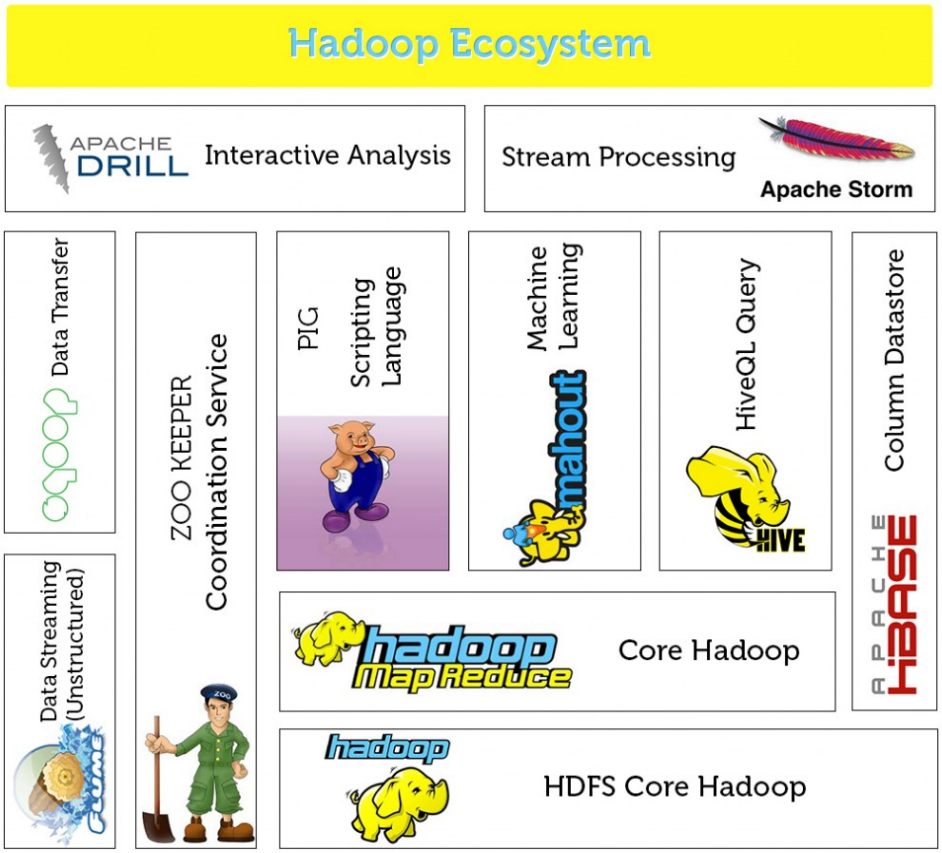
Some of the reasons organizations use Hadoop is its' ability to store, manage and analyze vast amounts of structured and unstructured data quickly, reliably, flexibly and at low-cost.

- **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 tools

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

- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.
- **HBase**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout**: A Scalable machine learning and data mining library.
- **Pig**: A high-level data-flow language and execution framework for parallel computation.
- **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.
- **Tez**: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to processdata for both batch and interactive use-cases. Tez is being adopted by HiveTM, PigTM and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace HadoopTM MapReduce as the underlying execution engine.
- **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.

Why H1- B VISA?

The US H-1B visa is a non-immigrant visa that allows US companies to employ graduate level workers in specialty occupations that require theoretical or technical expertise in specialized fields such as in IT, finance, accounting, architecture, engineering, mathematics, science, medicine, etc. Any professional level job that usually requires anyone to have a bachelors degree or higher can come under the H-1B visa for specialty occupations. The US employer petitions for the H-1B Visa in the US which has a duration of up to 6 years.Applying for a non-immigrant visa is generally quicker than applying for a US green Card, therefore the H-1B visa is popular for companies wishing to bring in staff for long-term assignment in the US. However, because of the lack of available visas employers frequently have to look at applying for other visa categories.

Eligibility for applying to Visa

The US H1-B visa is designed to be used for staff in specialty occupations. The job must meet one of the following criteria to qualify as a specialty occupation:

- Have a minimum entry requirement of a Bachelor's or higher degree or its equivalent.

- The degree requirement for the job is common to the industry or the job is so complex or unique that it can be performed only by an individual with a degree.

- The employer normally requires a degree or its equivalent for the position.

- The nature of the specific duties is so specialized and complex that the knowledge required to perform the duties is usually associated with the attainment of a bachelor's or higher degree.

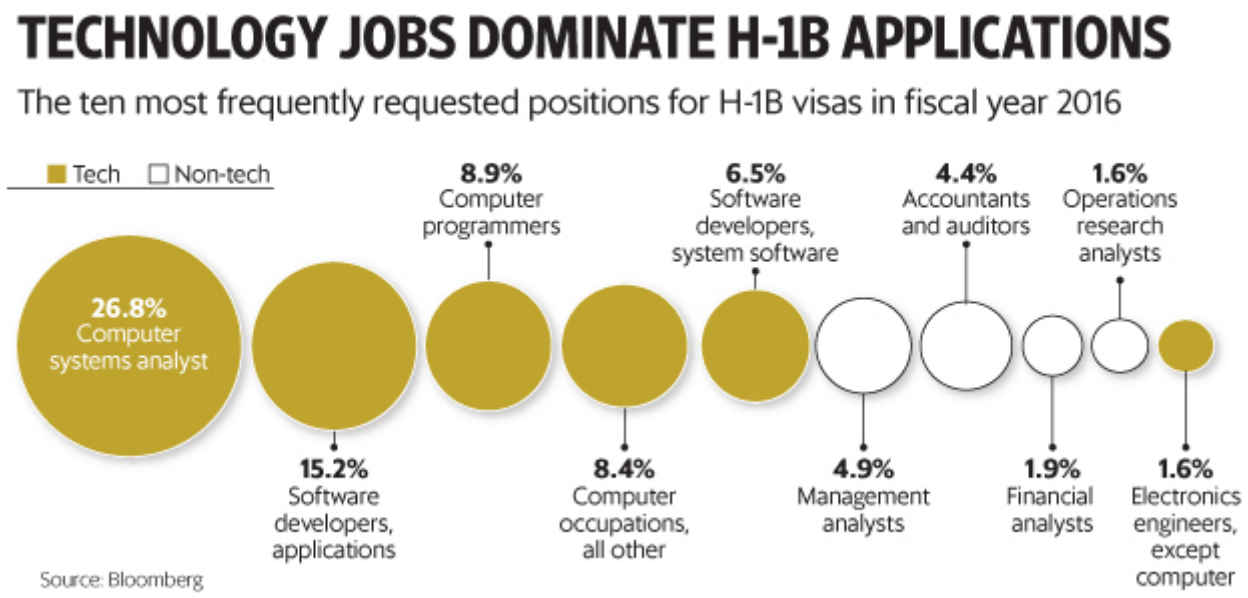
Length of stay

The H-1B visa is initially granted for up to three years, but may then be extended to a maximum of six years.

Even though the H-1B visa is a non-immigrant visa, it is one of the few US visa categories recognized as dual intent, meaning the H-1B visa holder can apply for and obtain a US Green Card while in the US on an H-1B visa. If you are still in the US on an H-1B visa and wish to remain in the US for more than six years, you can apply for permanent residency in the US to receive a Green Card. If you do not gain permanent residency prior to the expiration of your H-1B visa, then you must live outside the US for at least one year before reapplying for another H or L visa.

New Policies:

The US government has signalled a major policy shift in the H1B work visa programme, making computer programmers ineligible for the non-immigrant visas by default, firing a shot across the bow of India's largest outsourcing firms. On the face of it, the order looks to be levelling the playing field for Silicon Valley's biggest firms, such as Google Inc., Apple Inc. and Facebook Inc., which have for long complained that they lose out in the race to acquire highly skilled immigrant talent to large outsourcing firms such as Tata Consultancy Services Ltd (TCS) and Infosys Ltd, which are often the biggest recipients of the H1B visas.As per the regulations of the act, if a foreign worker quits or is dismissed as an employee, the worker must either apply for or be granted non-immigrant status or leave the US. The duration of stay, under this visa, is three years which can be extended to a maximum six years. An extension of ten years is applicable, exclusively for those who are engaged in defence-related project work in the United States.



PROJECT OUTLINE

TITLE : H1-B US VISA ANALYSIS

INPUT : Sample of US-Visa Data

DATA FIELDS : S_no,Case_status, employer_name,Soc_name, full time position, Job title, Prevailing wage, Year, worksite, longitude and latitude.

PURPOSE : Finding the non-Eligible Immigrants to work in USA.

DURATION : The non-immigrants applied from 2011-2015

PROJECT IMPLEMENTATION

Prerequisite : Hadoop Distributed File System access, Hive, Pig, Sqoop are installed in the node where Hadoop is installed.

Mode : Hadoop standalone mode.

PRE-PROCESS

- Input Data is in the form of CSV format. Each field is surrounded by double quotes(“ “) and delimited by comma(,)
- Input Data is loaded in Hive table of an h1b database. The create table command is as follows:

  hive(h1b_app)> CREATE TABLE h1b_pro(s_no int,case_status string, employer_name string, soc_name string, job_title string, full_time_position string,prevailing_wage int,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 COMMAND:

hive(h1b_app)> Load data local inpath ‘/home/hduser/h1bapp.csv overwrite into table h1b_app;

USE the following query to remove double quotes and to use delimiter as Tab.

hive(h1b_app)> INSERT OVERWRITE LOCAL DIRECTORY '/home/hduser/Desktop'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
select * from h1b_pro;

Merge the Hive mapper outputs into a single file using Pig:

h1b= LOAD '/home/sowmya/Desktop/h1b-test' 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,year, worksite1:chararray, worksite2:chararray, longitute, latitute);

aa = foreach h1b generate $0,$1,$2,$3,$4,$5,$6,$7,$8,$9,$10,$11;

Store aa into '/home/hduser/join' using PigStorage('\t');

fs -getmerge /home/hduser/join /home/hduser/project;

HIVE:

Tool: Hive

The input data in loaded in HiveQL.

1. Top 10 job positions which have high success rate in petitions.

hive> Select ROUND(((count(s_no)/t.tot)*100),2) as ctn,job_title from h1b_pro, (select count(s_no) as tot from h1b_pro)t where case_status ='CERTIFIED' or case_status= 'CERTIFIED-WITHDRAWN' group by job_title,t.tot order by ctn desc limit 10;

```
hduser@sowmya-ubuntu: ~
c
2017-04-29 19:42:14,280 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 7.35
sec
MapReduce Total cumulative CPU time: 7 seconds 350 msec
Ended Job = job_1493474794542_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 2   Cumulative CPU: 64.17 sec   HDFS Read: 492290
706 HDFS Write: 16521774 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.35 sec   HDFS Read: 1652654
4 HDFS Write: 254 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 11 seconds 520 msec
OK
7.97    PROGRAMMER ANALYST
3.84    SOFTWARE ENGINEER
2.25    COMPUTER PROGRAMMER
1.98    SYSTEMS ANALYST
1.35    SOFTWARE DEVELOPER
1.26    BUSINESS ANALYST
1.11    COMPUTER SYSTEMS ANALYST
0.94    TECHNOLOGY LEAD - US
0.87    TECHNOLOGY ANALYST - US
0.85    SENIOR SOFTWARE ENGINEER
Time taken: 62.234 seconds, Fetched: 10 row(s)
hive>
```

2) Find the average Prevailing Wage for each Job for each Year (take part time and full time separate)

hive> select ROUND((AVG(prevailing_wage)),2), job_title,year,full_time_position from h1b_pro where full_time_position!='NULL' group by full_time_position,job_title,year order by full_time_position desc;

```
68661.0 EPC PLANNER     2016    N
52228.8 ENVIRONMENTALLY PREFERRED PURCHASING COORDINATOR         2016    N
69389.0 ENVIRONMENTAL/PERMITS ENGINEER II       2016    N
50315.0 ENVIRONMENTAL, HEALTH & SAFETY ADVISOR  2016    N
55286.4 ENVIRONMENTAL TEST ENGINEER     2015    N
54433.6 ENVIRONMENTAL TEST ENGINEER     2013    N
50294.4 ENVIRONMENTAL SYSTEMS ANALYST   2012    N
49951.2 ENVIRONMENTAL SPECIALISTS       2016    N
43784.0 ENVIRONMENTAL SPECIALIST, AIR AND WATER (ASSOCIATE 6)   2016    N
35297.6 ENVIRONMENTAL SPECIALIST / RESEARCH ASSOCIATE   2015    N
53071.2 ENVIRONMENTAL SPECIALIST        2015    N
56049.07        ENVIRONMENTAL SPECIALIST        2013    N
42697.2 ENVIRONMENTAL SPECIALIST        2011    N
38709.0 ENVIRONMENTAL SERVICE ENGINEER TECHNICIAN       2016    N
53435.0 ENVIRONMENTAL SCIENTIST/GEOLOGIST       2016    N
60715.0 ENVIRONMENTAL SCIENTIST (ENVIRONMENTAL ENGINEER)        2016    N
50112.59        ENVIRONMENTAL SCIENTIST 2016    N
51168.0 ENVIRONMENTAL SCIENTIST 2014    N
52012.13        ENVIRONMENTAL SCIENTIST 2012    N
48333.0 ENVIRONMENTAL SCIENCE TEACHER   2016    N
61193.6 ENVIRONMENTAL SCIENCE INSTRUCTOR        2013    N
33000.0 ENVIRONMENTAL SCIENCE AND PROTECTION TECHNICIAN 2016    N
39353.6 ENVIRONMENTAL PROJECT MANAGER   2016    N
44325.0 ENVIRONMENTAL PROJECT ENGINEER  2016    N
35464.0 ENVIRONMENTAL PROJECT DESIGNER  2016    N
29494.4 ENVIRONMENTAL PROGRAM ASSISTANT 2016    N
37294.0 ENVIRONMENTAL POLICY ANALYST    2016    N
55827.0 ENVIRONMENTAL PLANNER   2016    N
38438.4 ENVIRONMENTAL MICROBIOLOGIST    2014    N
58177.6 ENVIRONMENTAL MANAGER   2013    N
58011.0 ENVIRONMENTAL INFORMATICS SOFTWARE DEVELOPER    2016    N
52124.8 ENVIRONMENTAL HYGIENIST 2016    N
57907.0 ENVIRONMENTAL HEALTH AND SAFETY SPECIALIST      2016    N
69805.0 ENVIRONMENTAL HEALTH & SAFETY SPECIALIST III    2016    N
99860.8 ENVIRONMENTAL HEALTH & SAFETY SPECIALIST        2011    N
43118.0 ENVIRONMENTAL HEALTH & SAFETY PRACTITIONER      2016    N
40580.8 ENVIRONMENTAL FIELD SCIENTIST   2016    N
65811.0 ENVIRONMENTAL FATE SCIENTIST    2016    N
38563.2 ENVIRONMENTAL EVENT SPECILIST   2014    N
53269.0 ENVIRONMENTAL ENGINEERS 2016    N
34444.8 ENVIRONMENTAL ENGINEERING TECHINCIANS   2016    N
54870.0 ENVIRONMENTAL ENGINEERING SPECIALIST 2  2016    N
44637.0 ENVIRONMENTAL ENGINEER IV       2016    N
```

3) **Which are top ten employers who have the highest success rate in petitions?**

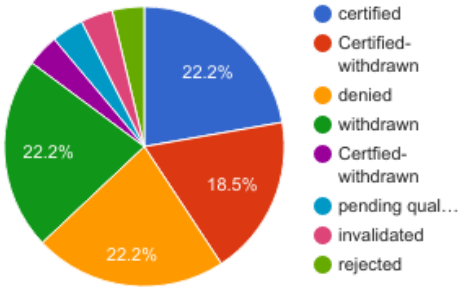
hive>select ROUND(((count(s_no)/t.tot)*100),2) as ctn,employer_name from h1b_pro,(select count(s_no) as tot from h1b_pro)t where case_status ='CERTIFIED' or case_status= 'CERTIFIED-WITHDRAWN' group by employer_name,t.tot order by ctn desc limit 10;

```
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1493474794542_0009
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-04-29 20:00:31,413 Stage-1 map = 0%,  reduce = 0%
2017-04-29 20:00:49,152 Stage-1 map = 37%,  reduce = 0%, Cumulative CPU 36.87 sec
2017-04-29 20:00:52,268 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 44.84 sec
2017-04-29 20:00:53,320 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 46.62 sec
2017-04-29 20:01:01,693 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 54.79 sec
2017-04-29 20:01:03,762 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 64.13 sec
MapReduce Total cumulative CPU time: 1 minutes 4 seconds 130 msec
Ended Job = job_1493474794542_0009
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1493474794542_0010, Tracking URL = http://sowmya-ubuntu:8088/proxy/application_1493474794542_0010/
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1493474794542_0010
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-04-29 20:01:14,566 Stage-2 map = 0%,  reduce = 0%
2017-04-29 20:01:19,926 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 4.41 sec
2017-04-29 20:01:26,154 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 7.31 sec
MapReduce Total cumulative CPU time: 7 seconds 310 msec
Ended Job = job_1493474794542_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 2   Cumulative CPU: 64.13 sec   HDFS Read: 492290928 HDFS Write: 11502572 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.31 sec   HDFS Read: 11507350 HDFS Write: 271 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 11 seconds 440 msec
OK
4.33    INFOSYS LIMITED
2.14    TATA CONSULTANCY SERVICES LIMITED
1.58    WIPRO LIMITED
1.2     DELOITTE CONSULTING LLP
1.11    ACCENTURE LLP
1.0     IBM INDIA PRIVATE LIMITED
0.84    MICROSOFT CORPORATION
0.75    HCL AMERICA, INC.
0.6     ERNST & YOUNG U.S. LLP
0.56    LARSEN & TOUBRO INFOTECH LIMITED
Time taken: 61.774 seconds, Fetched: 10 row(s)
hive>
```

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

hive>select count(s_no) as count,round(((count(s_no)/t.tot)*100),2) as ctn,year,case_status from h1b_pro,(select count(s_no) as tot from h1b_pro)t where year!='NULL' or year!='NA' group by case_status,year,t.tot order by ctn desc;

GRAPHICAL REPRESENTATION



PIG

- The tool used for further analysis is Pig. Data set is loaded is a bag in Pig grunt shell or by using the .pig script file. Script files are used for code reusability.

- Pig operates in two modes- Local and mapreduce mode
- Go to grunt shell by: pig -x local

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

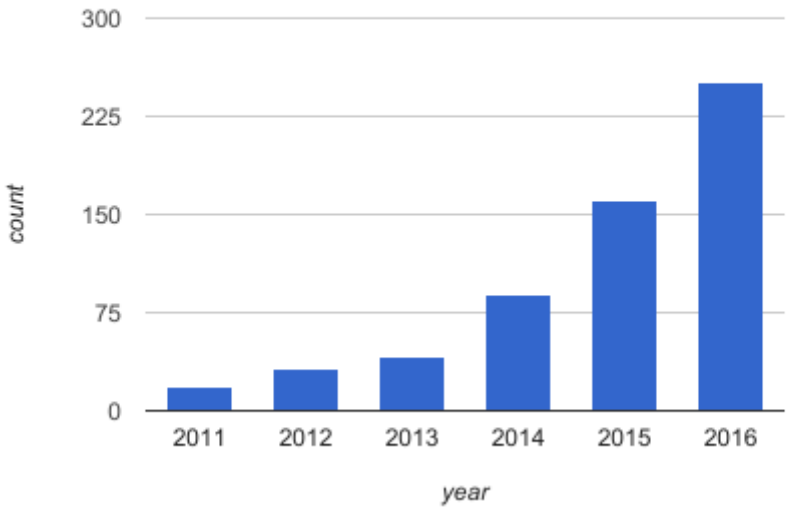
grunt> h1b= LOAD '/home/sowmya/Desktop/mapreduce' using PigStorage('\t') AS (s_no:int,case_status:chararray,employer_name:chararray,soc_name:chararray,job_title:chararray,full_time_position:chararray,prevailing_wage,year, worksite1:chararray,longitute, latitute);

filter_bag= filter h1b by job_title== 'DATA ENGINEER';

group_all= group filter_bag by year;

count_all= foreach group_all generate $0, COUNT(filter_bag) as headcount;

DATA VISUALIZATION



The above graph depicts that the Data Engineer increase over the time rapidly. Thus they are getting more visas certified.

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

h1b= LOAD '/home/sowmya/Desktop/mapreduce' 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,year,worksite1:chararray,longitute, latitute);

groupbyyear= GROUP h1b by (year, job_title);

countcust= foreach groupbyyear generate group as year, COUNT(h1b) as headcount;

orderbycount = order countcust by $1 desc;

limit2= limit orderbycount 5;

dump limit2;

Dump operator is view the results. And that can be stored in Local File System or HDFS.

```
ER
job_local1397506383_0003        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     orderbycount    SAMPLER
job_local1823924649_0005        1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     orderbycount            file:/t
mp/temp-116192712/tmp1580809765,
job_local877122179_0002 14      1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     countcust,groupbyyear,h1b       GROUP_B
Y,COMBINER

Input(s):
Successfully read 3002458 records from: "/home/sowmya/Desktop/mapreduce"

Output(s):
Successfully stored 5 records in: "file:/tmp/temp-116192712/tmp1580809765"

Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local877122179_0002 ->      job_local1397506383_0003,
job_local1397506383_0003        ->      job_local1365126249_0004,
job_local1365126249_0004        ->      job_local1823924649_0005,
job_local1823924649_0005


2017-04-30 20:22:22,307 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-04-30 20:22:22,308 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.b
ytes-per-checksum
2017-04-30 20:22:22,309 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultF
S
2017-04-30 20:22:22,309 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduc
e.jobtracker.address
2017-04-30 20:22:22,309 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-04-30 20:22:22,317 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-04-30 20:22:22,317 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((2016,PROGRAMMER ANALYST),53743)
((2015,PROGRAMMER ANALYST),53436)
((2014,PROGRAMMER ANALYST),43114)
((2013,PROGRAMMER ANALYST),33880)
((2012,PROGRAMMER ANALYST),33066)
grunt>
```

3)Which industry has the most number of Data Scientist positions?

h1b= LOAD '/home/sowmya/Desktop/mapreduce' 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,year, worksite1:chararray,longitute, latitute);

filteration= filter h1b by job_title == 'DATA SCIENTIST';

groupbyemployer= group filteration by employer_name;

countbypos= foreach groupbyemployer generate $0, COUNT(filteration) as headcount;

dump countbypos;

```
(UNIVISION INTERACTIVE MEDIA, INC.,1)
(VOICEBOX TECHNOLOGIES CORPORATION,1)
(AMERICAN EXPRESS TRS, COMPANY INC.,1)
(BLUESTAR TECHNOLOGY SOLUTIONS, LLC,1)
(MOTION PICTURES LABORATORIES, INC.,1)
(OMNICOM MEDIA GROUP HOLDINGS, INC.,1)
(TRENDALYTICS INNOVATION LABS, INC.,2)
(CHAR SOFTWARE INC. (DBA LOCALYTICS),1)
(ENDURANCE INTERNATIONAL GROUP, INC.,1)
(PIONEER HI-BRED INTERNATIONAL, INC.,3)
(RUTHS ANALYTICS AND INNOVATION, LLC,1)
(SCHLUMBERGER TECHNOLOGY CORPORATION,11)
(CHILDREN'S HOSPITAL OF ORANGE COUNTY,1)
(E. I. DU PONT DE NEMOURS AND COMPANY,1)
(EXPERIAN INFORMATION SOLUTIONS, INC.,2)
(INSTITUTE FOR MEDICAL RESEARCH, INC.,1)
(T3C, INC. DBA RETAIL SOLUTIONS, INC.,1)
(COMPREHENSIVE HEALTH MANAGEMENT, INC.,1)
(MASTERCARD INTERNATIONAL INCORPORATED,2)
(SEARS HOLDINGS MANAGEMENT CORPORATION,1)
(DAVIDSON KEMPNER CAPITAL MANAGEMENT LP,1)
(GEMOLOGICAL INSTITUTE OF AMERICA (GIA),1)
(MCKINSEY & COMPANY, INC. UNITED STATES,2)
(DAVIDSON KEMPNER CAPITAL MANAGEMENT, LP,1)
(FANATICS RETAIL GROUP FULFILLMENT, INC.,5)
(GLOBAL TOUCHPOINTS, INC. DUNS# 13-8058305,1)
(ACCIDENT FUND INSURANCE COMPANY OF AMERICA,2)
(HEALTHCARE ANALYTICS SERVICES HOLDING INC.,1)
(UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL,1)
(AXIAL NETWORKS, INC. (FORMERLY AXIAL MARKET),1)
(HOUGHTON MIFFLIN HARCOURT PUBLISHING COMPANY,1)
(SONY NETWORK ENTERTAINMENT INTERNATIONAL LLC,2)
(CAESARS ENTERTAINMENT OPERATING COMPANY, INC.,1)
(PHILIPS ELECTRONICS NORTH AMERICA CORPORATION,1)
(PLAYDOM, INC., PART OF THE WALT DISNEY COMPANY,1)
(HEALTHCARE BUSINESS INTELLIGENCE SOLUTIONS INC.,1)
(NOVARTIS INSTITUTE FOR FUNCTIONAL GENOMICS, INC.,1)
(AMERICAN EXPRESS TRAVEL RELATED SERVICES COMPANY,,2)
(ADVANCED INFORMATION MANAGEMENT TECHNOLOGY PARTNER,2)
(AMERICAN EXPRESS TRAVEL RELATED SERVICES CO., INC.,1)
(CAPITAL IQ, INC. (SUBSIDIARY OF THE MCGRAW-HILL CO,1)
(MORNINGSTAR, INC. (HELLOWALLET, A MORNINGSTAR COMPANY),1)
grunt>
```

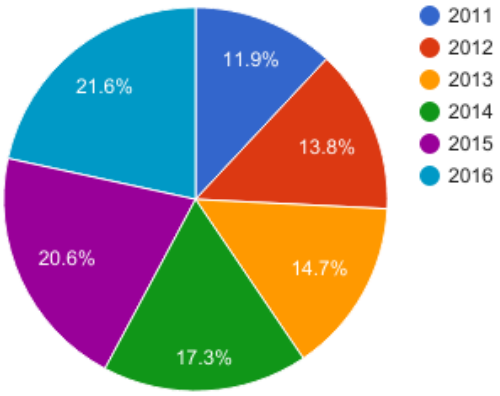
4) Create a bar graph to depict the number of applications for each year

h1b= LOAD '/home/sowmya/Desktop/mapreduce' 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,year, worksite1:chararray,longitute, latitute);

groupbyyear= GROUP h1b by year;

countapp= foreach groupbyyear generate group as year, COUNT(h1b) as headcount;

GRAPHICAL VIEW:



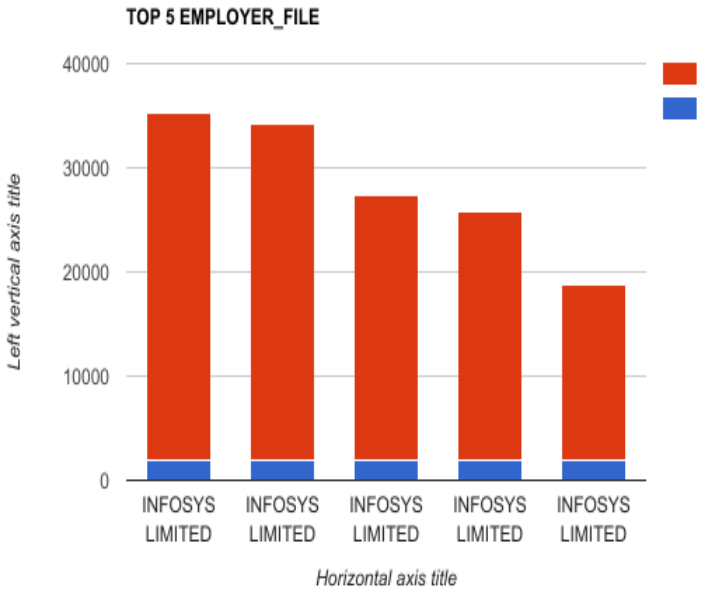
From the above graph, it is evident that number of applications grows every year which makes the government hard to find the eligible candidates.

Mapreduce Framework:

Tools: Mapreduce, Eclipse

- Input data should be put in HDFS(Hadoop Distrbuted File System)
- Tree map is used in the program which sorts the values in ascending order
- Cleanup( ) function is called at the end of the task.

Graphical View:




```
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1493474794542_0009
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-04-29 20:00:31,413 Stage-1 map = 0%,  reduce = 0%
2017-04-29 20:00:49,152 Stage-1 map = 37%,  reduce = 0%, Cumulative CPU 36.87 sec
2017-04-29 20:00:52,268 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 44.84 sec
2017-04-29 20:00:53,320 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 46.62 sec
2017-04-29 20:01:01,693 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 54.79 sec
2017-04-29 20:01:03,762 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 64.13 sec
MapReduce Total cumulative CPU time: 1 minutes 4 seconds 130 msec
Ended Job = job_1493474794542_0009
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1493474794542_0010, Tracking URL = http://sowmya-ubuntu:8088/proxy/application_1493474794542_0010/
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1493474794542_0010
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-04-29 20:01:14,566 Stage-2 map = 0%,  reduce = 0%
2017-04-29 20:01:19,926 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 4.41 sec
2017-04-29 20:01:26,154 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 7.31 sec
MapReduce Total cumulative CPU time: 7 seconds 310 msec
Ended Job = job_1493474794542_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 2   Cumulative CPU: 64.13 sec   HDFS Read: 492290928 HDFS Write: 11502572 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.31 sec   HDFS Read: 11507350 HDFS Write: 271 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 11 seconds 440 msec
OK
4.33    INFOSYS LIMITED
2.14    TATA CONSULTANCY SERVICES LIMITED
1.58    WIPRO LIMITED
1.2     DELOITTE CONSULTING LLP
1.11    ACCENTURE LLP
1.0     IBM INDIA PRIVATE LIMITED
0.84    MICROSOFT CORPORATION
0.75    HCL AMERICA, INC.
0.6     ERNST & YOUNG U.S. LLP
0.56    LARSEN & TOUBRO INFOTECH LIMITED
Time taken: 61.774 seconds, Fetched: 10 row(s)
hive>
```

DATA EXPORT TO RELATIONAL DATABASE

The OUTPUT data can be used whenever needed, apart from HDFS, it is better to export data to the local database where analysts do simple query in MySQL and do analysis in future.

This can be achieved using Apache Sqoop,

- Sqoop is used to import and export data from or to Relational Databases
- HDFS file can be directly exported to RDBMS

Make an input file for the top 10 job positions who have highest success rate in petitions and put it in HDFS. The command for it:

hduser@sowmya-ubuntu:~$ hadoop fs -put /home/sowmya/visa.txt /niit

STEPS FOR EXPORT:

Enter the mysql server: mysql -u root -p

- Create a database in Mysql.
- **Create a table in Mysql to export data from HDFS.** The table should be as exact schema as of HDFS file. Command is as follows:

```
create table h1b_app(
    success_rate FLOAT NOT NULL,
     position VARCHAR(40) NOT NULL,
     status VARCHAR(40) NOT NULL,
PRIMARY KEY(success_rate));
```

ANALYSIS FROM THE REPORT:

Analysis of H1-B US-Visa is as follows:

- The number of applications submitted for H1-B Visas are increasing rapidly every year. So the government needs to select the top candidates
- The top job positions who get certified are Programmer Analyst, Software Engineer and Computer Programmer.
- The top companies who get certified are Infosys Ltd, TCS Ltd and Wipro Ltd.

SUGGESTIONS FROM THE REPORT:

- With Trump's decision on H-1B, Indian IT companies may suffer a major setback as top companies offering H-1B visa were majorly outsourcing employment to India. It will hit hard the entry level techies. So Young People should really work hard for H1-B Visa
- There is a setback for Computer Programmers according to the new Policies of H1-B VISA. The competition has increased in Engineering Field.