

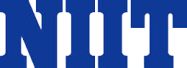
**H1b Application – Related Statistical Analysis**

**BIG DATA – HADOOP**

SUDARSANAN.M.R.|Professional Diploma in Digital Transformation – Big Data with Hadoop | 02.05.2016

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

**312612106083** **NIIT Limited**



**H1B APPLICATION –RELATED STATISTICAL ANALYSIS**

**A PROJECT REPORT**

***Submitted by***

**SUDARSANAN.M.R.**

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

***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. Santhanalakshmi,** 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 extend my thanks to all staff members of NIIT for their kind co-operation for the completion of the project successfully. I am grateful to thank my family and all my friends for their valuable feedback, encouragement and suggestions.

**Introduction**

**Big Data**

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

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

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

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

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

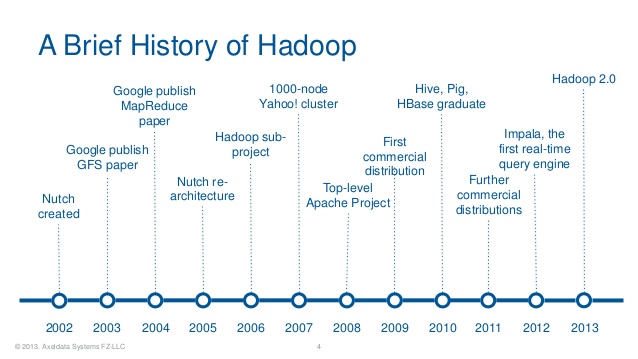
**Apache Hadoop**

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 projects:

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

**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.

**Hadoop-related projects at Apache include:**

**Ambar**: 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. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.

**Avro**: A data serialization system.

**Cassandra**: A scalable multi-master database with no single points of failure.

**Chukwa**: A data collection system for managing large distributed systems.

**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 process data for both batch and interactive use-cases. Tez is being adopted by Hive™, Pig™ and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop™ 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.

**Big Data – Government**

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

**H1b Application –Statistical Analysis**

* Analysis on H1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States.
* We will be performing analysis on the H1B visa applicants between the years 2011-2015.
* It is the purpose of gathering information about the below mentioned fields using Apache Hadoop Tools.
* The dataset has nearly 3 million records.
* Purpose of this Analysis is to produce the Analytic overview of Applicants applying for visa process.

**Tools in Hadoop Ecosystem**

* Apache Hive
* Apache Sqoop
* Apache Pig
* Apache HBASE
* Apache Flume

**Input Data set Details**

* Case status
* SOC name
* SOC number
* Job Title
* Employer name
* Prevailing Wages
* Full time position
* Worksite
* Year
* Latitude
* Longitude

**HIVE Tool**

* Data set is loaded into hive table pro under database named Project.
* 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 int,year string, worksite string,worksite2 string,longitute double, latitute double )

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

WITH SERDEPROPERTIES (

"separatorChar" = ",",

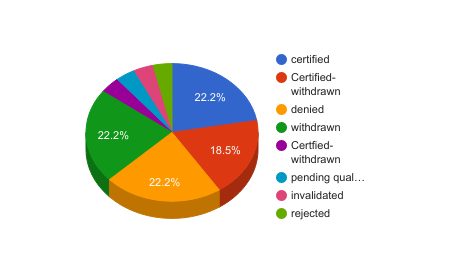
"quote Char" = "\""

) STORED AS TEXTFILE;

**Sample Query in HIVE**

* 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 ,(count(s\_no)/t.tot)\*100 as ctn,year,case\_status from pro,(select count(s\_no) as tot from pro)t group by case\_status,year order by ctn desc;



**Pig Tool**

* Data set is load in a Bag in Pig grunt shell using the following command.
* a= load '/home/sudarshan/Downloads/project1/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,lattitute);

**Pig Latin-Script**

Sample Question

* Which top employers file the most petitions each year?

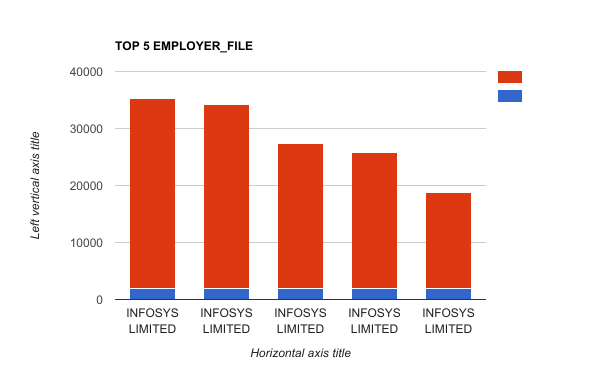
gp= group a by (year,employer\_name);

ct= foreach gp generate group as year, COUNT(a) as headcount;

obc= order ct by $1 desc;

limit2= limit obc 5;

OUTPUT-



**Data Export to Relational Database**

* Job positions which have the highest success rate in petitions has been done in hive tool it is been transfered to mysql database
* This can be achieved using Apache Sqoop
* Sqoop is used to import and export data from or to Relational Databases



**Procedures to Run Sqoop Command**

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

Enter the mysql server: mysql -u root -p

**Create a table in Mysql to export data from HDFS:**

create table h1b\_app(

success\_rate FLOAT NOT NULL,

position VARCHAR(40) NOT NULL,

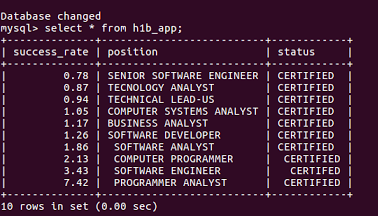
status VARCHAR(40) NOT NULL,

PRIMARY KEY(success\_rate));

**SQOOP QUERY:**

sqoop export --connect jdbc:mysql://localhost/app --username root -P --table h1b\_app --update-mode allowinsert --update-key success\_rate --export-dir /niit/visa.txt --input-fields-terminated-by ',';

**OUTPUT –**



ANALYSIS FROM THE REPORT:

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

* The number of applications submitted for H1-B Visas are increasing repidly every year. So the government needs to select the top candidates
* The top job positions who get certified are Programmer Analyst, Software Enginner 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.