Big Data Analytics using Hadoop

By – Ramit Kundu.

NIIT Camac Street.

(S171111700093)

Big Data:

Big Data is also a **data** but with a **huge size.** Big Data is a term used to describe collection of data that is huge in size and yet growing exponentially with time. These data’s are so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

Examples: Social Network, Sensors, Stock etc.

Characteristics of Big Data:

* Velocity- The term Big Data itself is related to enormous data and size.
* Variety- The next thing about Big Data is its different variety.
* Velocity- It refers to the speed at which data is being is generated.
* Veracity- Trustworthiness of data.

Hadoop:

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.

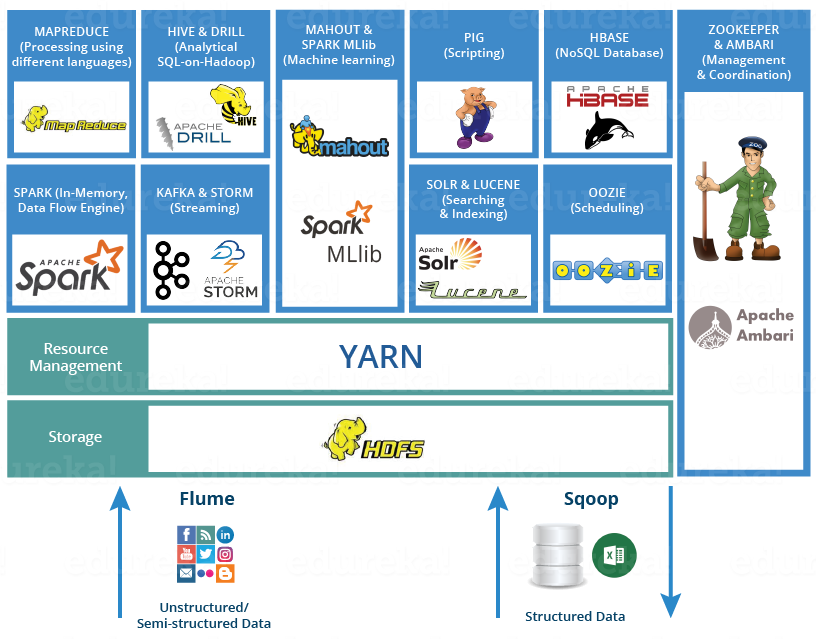
Different modules:-

The modules are:

* 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 Ecosystem:

Hadoop Ecosystem is neither a programming language nor a service, it is a platform or framework which solves big data problems.



* [HDFS](https://www.edureka.co/blog/hadoop-ecosystem#hdfs) - Hadoop Distributed File System
* [YARN](https://www.edureka.co/blog/hadoop-ecosystem#yarn)- Yet Another Resource Negotiator
* [MapReduce](https://www.edureka.co/blog/hadoop-ecosystem#mapreduce)- Data processing using programming
* [Spark](https://www.edureka.co/blog/hadoop-ecosystem#apache_spark)- In-memory Data Processing
* [PIG](https://www.edureka.co/blog/hadoop-ecosystem#apache_pig), [HIVE](https://www.edureka.co/blog/hadoop-ecosystem#apache_hive) - Data Processing Services using Query (SQL-like)
* [HBase](https://www.edureka.co/blog/hadoop-ecosystem#apache_hbase)- NoSQL Database
* [Mahout](https://www.edureka.co/blog/hadoop-ecosystem#apache_mahout), Spark  - Machine Learning
* [Zookeeper](https://www.edureka.co/blog/hadoop-ecosystem#apache_zookeeper)- Managing Cluster
* [Oozie](https://www.edureka.co/blog/hadoop-ecosystem#apache_oozie)- Job Scheduling
* [Flume](https://www.edureka.co/blog/hadoop-ecosystem#apache_flume), [Sqoop](https://www.edureka.co/blog/hadoop-ecosystem#apache_sqoop) - Data Ingesting Services

HDFS:

[Hadoop Distributed File System](https://www.edureka.co/blog/hdfs-tutorial)  is the backbone of Hadoop Ecosystem. HDFS is the one, which makes it possible to store different types of large data sets (i.e. structured, unstructured and semi structured data). It helps us in storing our data across various nodes and maintaining the log file about the stored data (metadata). HDFS has two core components, i.e. Name Node and Data Node.

The Name Node is the main node and it doesn’t store the actual data. It contains metadata, just like a log file.

On the other hand, all the data is stored on the Data Nodes and hence it requires more storage resources. These Data Nodes are commodity hardware in the distributed environment. This is one reason, why Hadoop solutions are very cost effective.

We always communicate with the Name Node while writing the data. Then, it internally sends a request to the client to store and replicate data on various Data Nodes.

MapReduce:

It is the core component of processing in a Hadoop Ecosystem as it provides the logic of processing. In other words, MapReduce is a software framework which helps in writing applications that processes large data sets using distributed and parallel algorithms inside Hadoop environment.

In a MapReduce program, Map () and Reduce () are two functions.

TheMap function performs actions like filtering, grouping and sorting.

While Reduce function aggregates and summarizes the result produced by map function.

The result generated by the Map function is a key value pair (K, V) which acts as the input for Reduce function.

Pig:

Apache Pig is a high-level language platform developed to execute queries on huge datasets that are stored in HDFS using Apache Hadoop. It is similar to SQL query language but applied on a larger dataset and with additional features. The language used in Pig is called Pig Latin. It is very similar to SQL. It is used to load the data, apply the required filters and dump the data in the required format. It requires a Java runtime environment to execute the programs. Pig converts all the operations into Map and Reduce tasks which can be efficiently processed on Hadoop.

Hive:

The Apache Hive is a data warehouse software which facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.

Features of Hive:

* It stores schema in a database and processed data into HDFS.
* It is designed for OLAP.
* It provides SQL type language for querying called HiveQL or HQL.
* It is familiar, fast, scalable, and extensible.

HBase:

HBase is an open source, non-relational distributed database. In other words, it is a NoSQL database. It supports all types of data and that is why, it’s capable of handling anything and everything inside a Hadoop ecosystem.

It is modelled after Google’s Big Table, which is a distributed storage system designed to cope up with large data sets. The HBase was designed to run on top of HDFS and provides Big Table like capabilities.

It gives us a fault tolerant way of storing sparse data, which is common in most Big Data use cases. The HBase is written in Java, whereas HBase applications can be written in REST, Avro and Thrift APIs.

Flume:

The Flume is a service which helps in ingesting unstructured and semi-structured data into HDFS .It gives us a solution which is reliable and distributed and helps us in collecting, aggregatingand moving large amount of data sets.

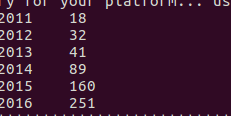
It helps us to ingest online streaming data from various sources like network traffic, social media, email messages, log files etc. in HDFS.

Sqoop:

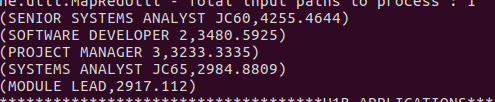
Flume only ingests unstructured data or semi-structured data into HDFS. While Sqoop can import as well as export structured data from RDBMS or Enterprise data warehouses to HDFS or vice versa.

H1B Project

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



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



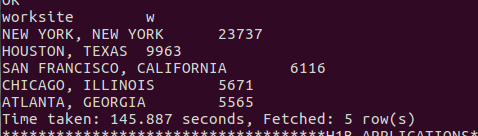
2a) Which part of the US has the most Data Engineer jobs for each year?



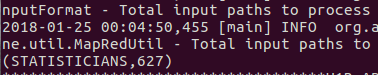


2b) find top 5 locations in the US who have got certified visa for each

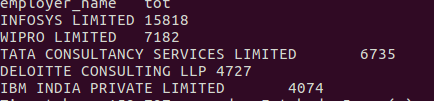
Year [certified]?



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



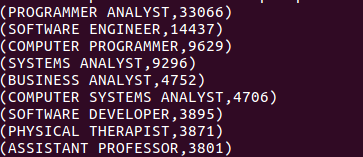
4) Which top 5 employers file the most petitions each year? - Case Status – ALL

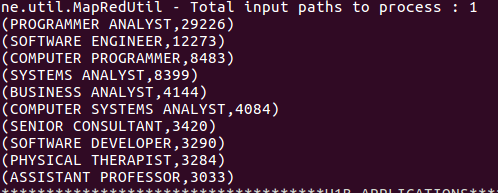


5) Find the most popular top 10 job positions for H1B visa applications for each year?

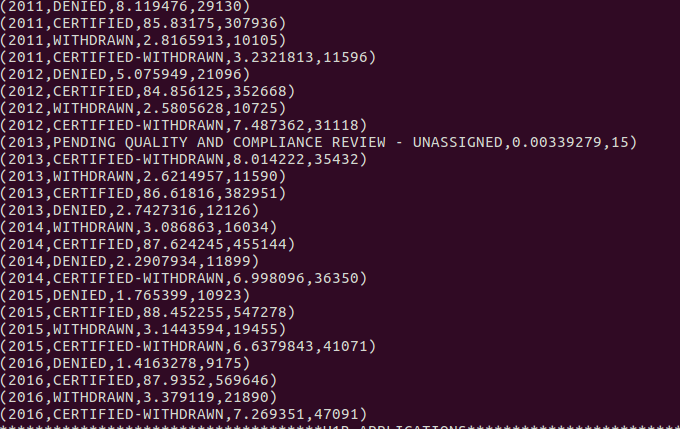
a) For all the applications

b) For only certified applications.

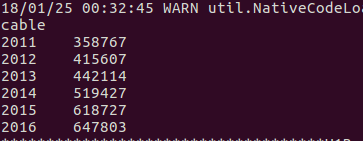




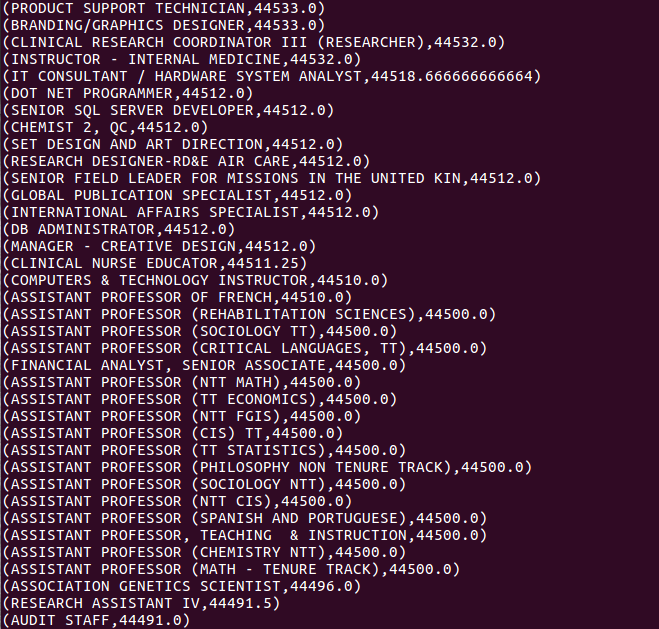
6) Find the percentage and the count of each case status on total applications for each year. Create a line graph depicting the pattern of all the cases over the period of time.



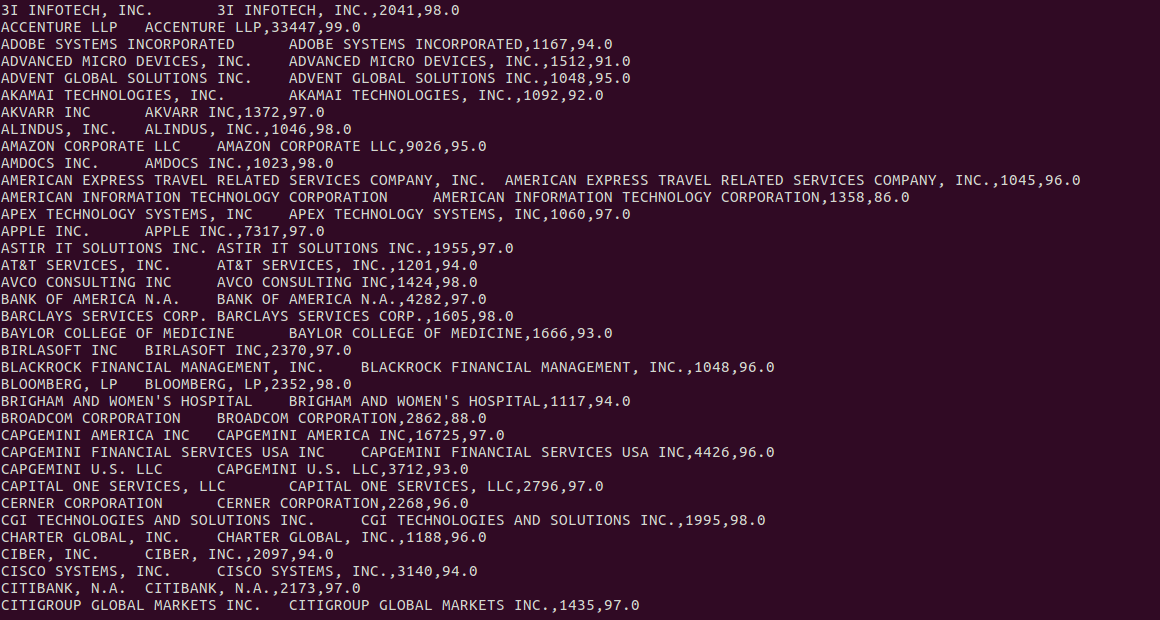
7) Create a bar graph to depict the number of applications for each year [All]



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.]



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)?



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)?

