Big Data Analysis using Hadoop on H-1B VISA

Purba Mondal

H1B DataBig Data Analysis using Hadoop

Contents

Big Data Analysis using Hadoop

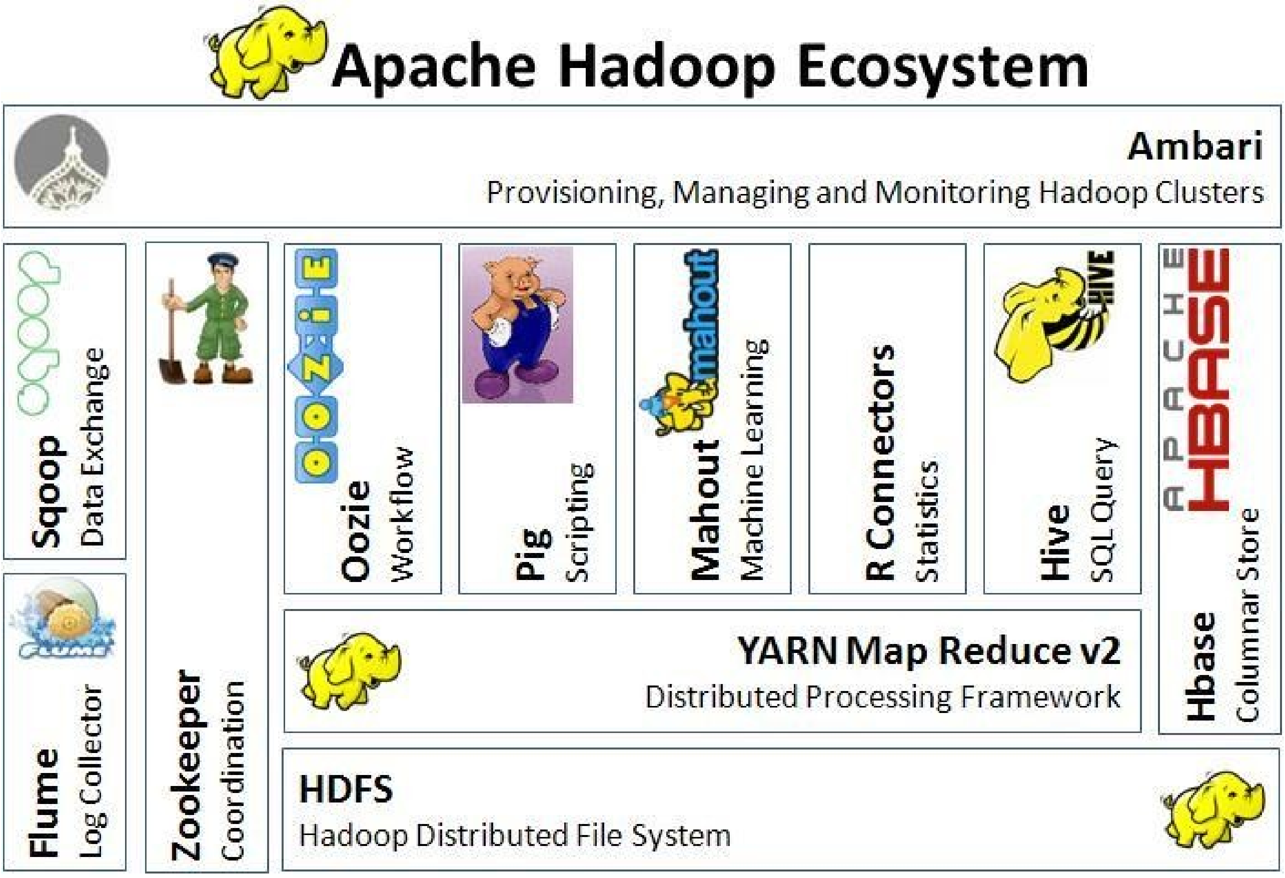
1. Introduction

Big Data: Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture,sharing, storage, transfer, visualization, and updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision-making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk. The characteristics of Big Data are:

Volume: The quantity of generated and stored data.
Variety: The type and nature of the data.
Velocity: The speed at which the data is generated and processed to meet the demands and challenge that lies in the path of growth and development.

Hadoop: It is an open source software framework for distributed storage and distributed processing of very large scale datasets on computer clusters built from commodity hardware. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce. Hadoop splits files into large blocks and distributes them across nodes in a cluster. To process data, Hadoop transfers packaged code for nodes to process in parallel based on the data that needs to be processed. This approach takes advantage of data locality – nodes manipulating the data they have access to – to allow the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

2. The Hadoop Ecosystem :



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 co**mputation.**

 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.

 Zookeeper: A high-performance coordination service for distributed applications.

The Hadoop framework is composed of the following modules:

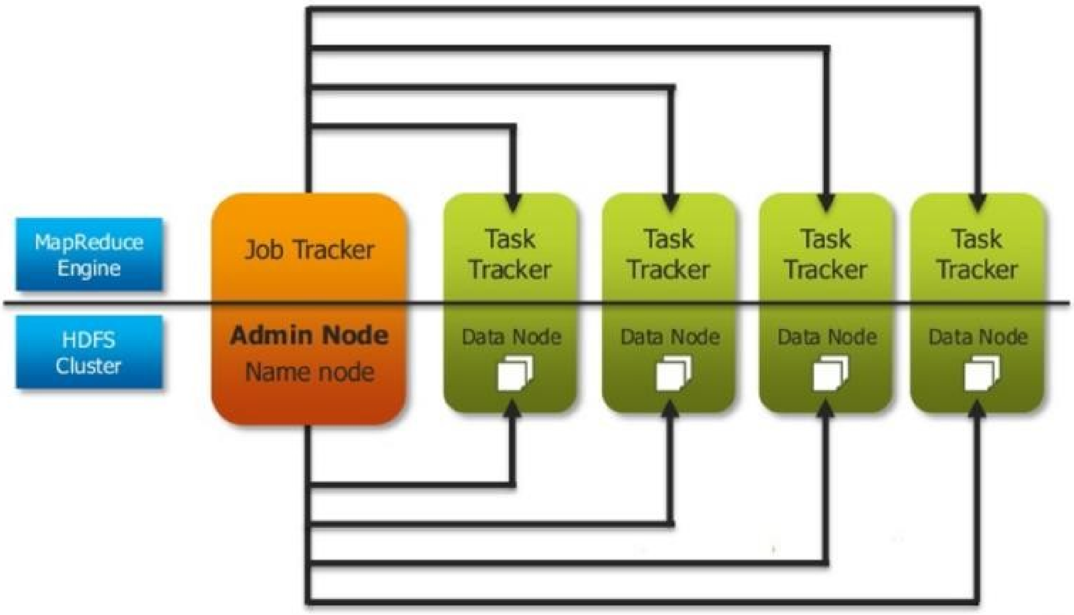
Hadoop Distributed File System: The Hadoop Distributed File System (HDFS) is designed to
store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly
attached storage and execute user application tasks.

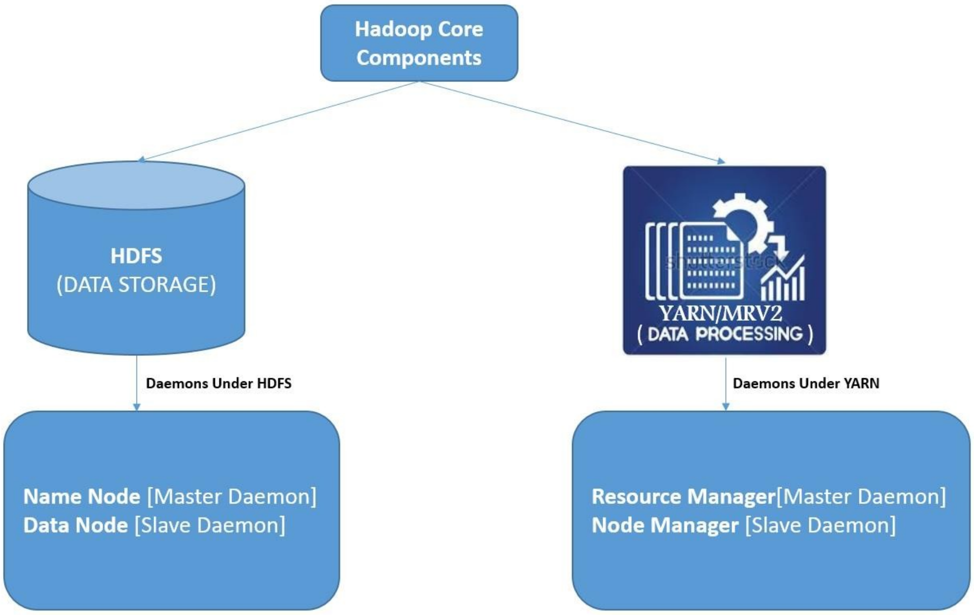
Why is Hadoop important?

Hadoop provides a reliable shared storage (HDFS) and analysis system (Map Reduce).
Hadoop is very cost effective as it can work with commodity hardware and does not require expensive high-end hardware.
Hadoop is highly flexible and can process both structured as well as unstructured data.
Hadoop has built-in fault tolerance.
Hadoop works on the principle of write once and read multiple times.

3. Core Components of Hadoop Architecture

HDFS: Hadoop Distributed File System is a distributed file system
YARN – (Yet Another Resource Negotiator) provides resource management for the processesrunning on Hadoop.
MapReduce – It is a parallel processing software framework. It is comprised of two steps. Map step is a master node that takes inputs and partitions them into smaller sub problems and then distributes them to worker nodes. After the map step has taken place, the master node takes the answers to all of the sub problems and combines them to produce output.
Name Node: keep tark of the data node.
Resource Manager:
Job Tracker:
Data Node:
Node Manager:


MapReduce Engine
HDFS Cluster
Job Tracker
Admin Node
Name node
Task Tracker
Data Node
Task Tracker
Data Node
Task Tracker
Data Node
Task Tracker
Data Node


Hadoop Core Components
HDFS
(DATA STORAGE)
YARN/MRV2
(DATA PROCESSING)
Daemons Under HDFS
Daemons Under YARN
Name Node [Master Daemon]
Data Node [Slave Daemon]
Resource Manager[Master Daemon]
Node Manager [Slave Daemon]

4. Map Reduce Framework:

Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data in-parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner .Map takes a data in the form of pairs and returns a list of <key, value> pairs.Using the output of Map, sort and shuffle are applied.Output of sort and shuffle will be sent to reducer phase.In Reducer calculation part is done on list of values for unique keysFinal output will<key, value> will be stored/displayed.

5. Pig:

Apache Pig is a high level data flow platform for execution Map Reduce programs of Hadoop.Pig executes in two modes: Local Mode and MapReduce Mode.

Local Mode:

```
$ pig-x local
```

It executes in a single JVM and is used for development experimenting and prototyping.Local mode works on local file system.

MapReduce Mode:

```
$ pig  or $ pig-xmapreduce
```

The MapReduce mode is also known as Hadoop Mode. In this Pig Latin into MapReduce jobs and executes them on the cluster.

6. Hive:

Hive is a data ware house system for Hadoop. It runs SQL like queries called HQL (Hive query language). Hive Query Language is similar to SQL and gets reduced to map reduce jobs in backend. Hive's default database is derby. Hive's metastore is used to persist schema i.e. table definition(table name, columns, types), location of table files, row format of table files, storage format of files.

7. Configuration:

Software Required:

Virtual Machine: In computing, a virtual machine (VM) is an emulation of a computer system.Virtual machines are based on architectures and provide functionality of a physical computer.Their implementations may involve specialized hardware, software, or a combination.There are different kinds of virtual machines, each with different functions. System virtual machines (also termed virtualizations) provide a substitute for a real machine. They provide functionality needed to execute entire operating systems. A hypervisor uses native execution to share and manage hardware, allowing multiple different environments. Each one being isolated from one another, yet existing on the same physical machine. Modern hypervisors use hardware- assisted virtualization, virtualization-specific hardware, primarily from the host CPUs. Process virtual machines are designed to execute computer programs in a platform-independentenvironment.

Ubuntu:Ubuntu is a complete Linux operating system, freely available with both community and professional support. The Ubuntu community is built on the ideas enshrined in the Ubuntu Manifesto: that software should be available free of charge, that software tools should be usable by people in their local language and despite any disabilities, and that people should have the freedom to customize and alter their software in whatever way they see fit.

Database: A database is an organized collection of data. It is the collection of schemas, tables, queries, reports, views, and other objects. The data are typically organized to model aspects of reality in a way that supports processes requiring information, such as modelling the availability of rooms in hotels in a way that supports finding a hotel withvacancies.

A database management system (DBMS) is a computer software application that interacts withthe user, other applications, and the database itself to capture and analyse data. A general-purpose DBMS is designed to allow the definition, creation, querying, update, and administration of databases. Well-known DBMSs include MySQL, PostgreSQL,MongoDB, Microsoft SQL Server, Oracle, Sybase, SAP HANA, and IBM DB2. A database is not generally portable across different DBMSs, but different DBMS can interoperate by using standards such as SQL and ODBC or JDBC to allow a single application to work with more than one DBMS.Database management systems are often classified according to the database model that they support; the most popular database systems since the 1980s have all supported the relational model as represented by the SQLlanguage.

MySQL:

- It is Open source relational database management system.
- Enables you to implement a database with tables, columns and indexes.
- Guarantees the Referential Integrity between rows of various tables.
- Updates the indexes automatically.
- Interprets an SQL query and combines information from various tables

8 .Project Objective

Title BigData Analysis in Hadoop on H1B Data.:
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.

Inputs H1b Data
The dataset has nearly 3 million records.

The dataset description is as follows:
The columns in the dataset include:

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

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

  Certified Withdrawn: LCA was approved but later withdrawn by employer

  Withdrawn: LCA was withdrawn by employer before approval

  Denied: LCA was denied by DOL

- EMPLOYER_NAME: Name of employer submitting labour condition application.
- SOC_NAME: the Occupational name associated with the SOC_CODE. SOC_CODE is the occupational code associated with the job being requested for temporary labour condition, as classified by the Standard Occupational Classification (SOC) System.

- JOB_TITLE: Title of the job
- FULL_TIME_POSITION: Y = Full Time Position; N = Part Time Position
- PREVAILING_WAGE: Prevailing Wage for the job being requested for temporary labour condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer’s minimum requirements for the position.
- YEAR: Year in which the H1B visa petition was filed
- WORKSITE: City and State information of the foreign worker’s intended area of employment
- lon: longitude of the Worksite
- lat: latitude of the Worksite

Analysis Relevance Education, Social, Finance and Planning
Purpose This analysis can be used by government for the
Methodology Agile

9. Project Implementation

Assumptions:

1. Hadoop Cluster is Running
2. Ecosystem Products (Pig, Hive, Sqoop) are installed
3. H1b data is available on HDFS in CSV Format

Prerequisites for All Jobs:

The H1B data is in CSV format and hence needs to be converted in Text format in Hadoop file
system.

Steps for Conversion

Step 1: Create Table in hive to read entire record as text format
Step 2: Create Table to remove different terminator and put a common terminator
Step 3: Create a table to change and recreate only 4 case status.

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

Objective:Count each year Data Engineer job poisition,and check is it increaing or not from previous year

Method use : Map-Reduce

a. Map part: Pass 1 value where ever job position ==Data Engineer (2011,1)
b. Reducer part: Count per year value
c. Check is it increaing from previous year or not

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

Objective:Count each year job poisition group by job poisition and calculate avg growth of that job position, and take top 5

Method use : pig

a. Filter the data per year and group by job position for each group
b. Join each group and theire count, and filter
c. Calculate the percentage for each group, then make avg value
d. take the top 5 job poisition.

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

Objective: Count job poisition for each year and each work site, then sort the result as per value then take top position

Method used: Map-reduce

a. select worksite and year as pass 1 for each value

b. calculate the total value per year paer worksite

c. select top most position of each group

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

Objective: group the data by worksite and year where status = CERTIFIED.

a.give which year want to get ans

b.group by worksire

c.take top 5.