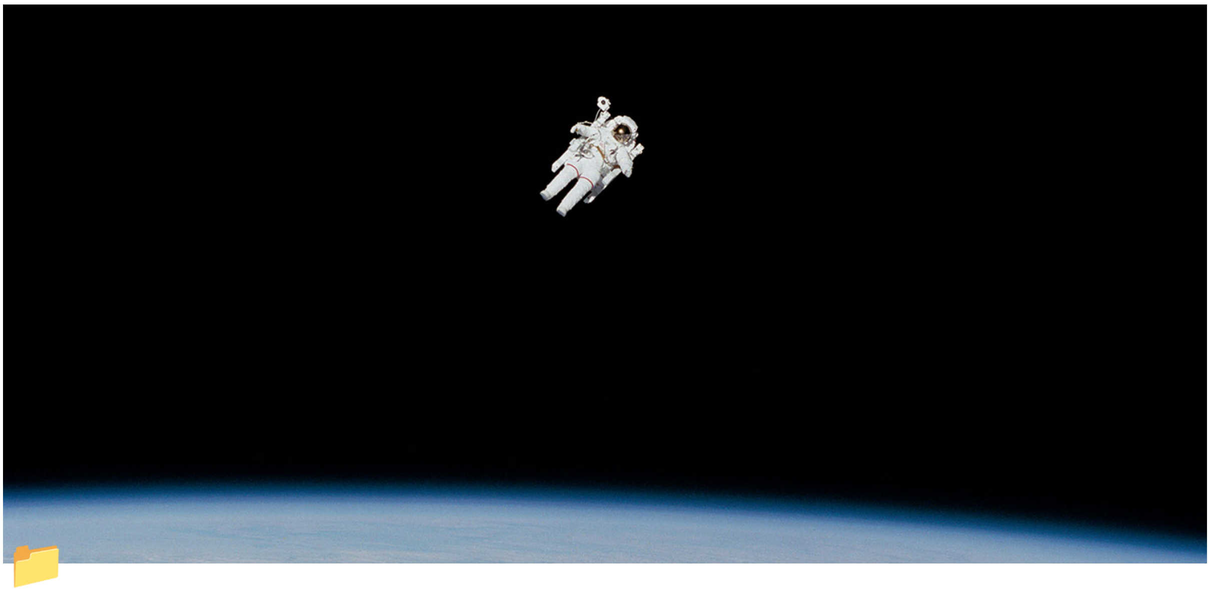


📁

Hadoop

Singh Rounak.

[BIG Data - 3vs]

Volume - Data cannot be stored in local machine, HDD etc, needs more space

Variety - Structured and unstructured data - eg- Social media data

Velocity - Fast processing.

Veracity - refers to Data Quality.

Examples -

Amazon, Netflix, Spotify Recommendation Engines

UBER, Hailo App Sensor and Geodata

Googla Now/Apple Siri

Tesla'a Autopilot

Google Analytics - Log data of websites

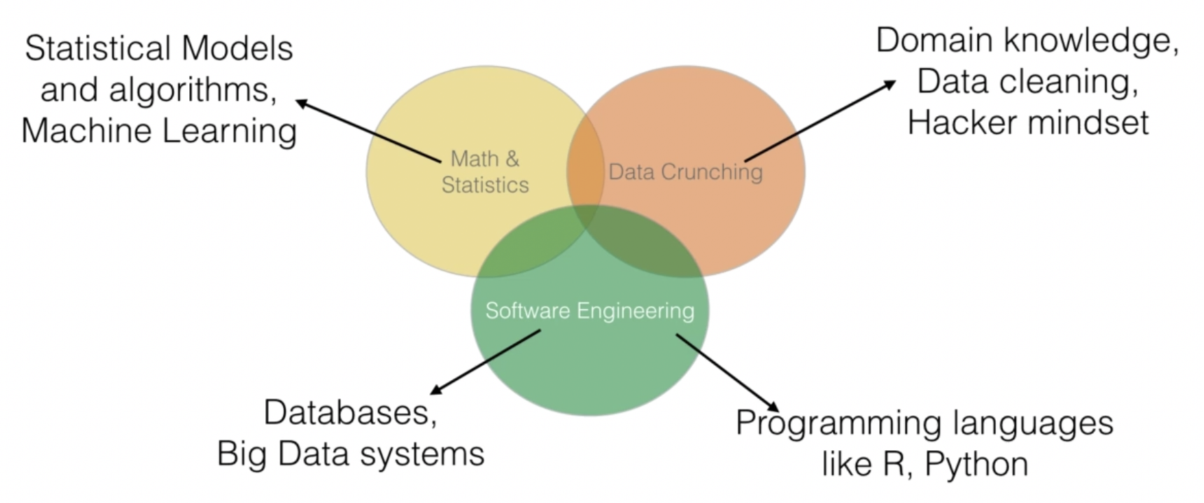
Stock Market Trading Data.

Graph data - Intelligence agencies.

DATA SCIENCE

Data Scientist - Uses statistical and mathematical tools to get more insights from data - Data Crunching and knowledge of Software Engineering.

Data Science is widely used for Big Data Applications



HADOOP

→ Reliable and Scalable Software to store and process Big Data. [created by Yahoo] based on Google's File System, Falls under the Apache umbrella.

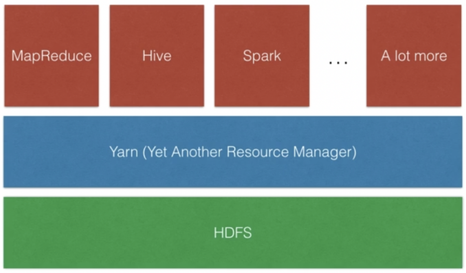
→ Runs on Commodity Hardware

→ Lower cost per GB

→ Petabytes of data in a single GB

***Data Processing Application - Code written to perform data processing over a large distributed system.

A Big Data processing cluster is shown as follows:



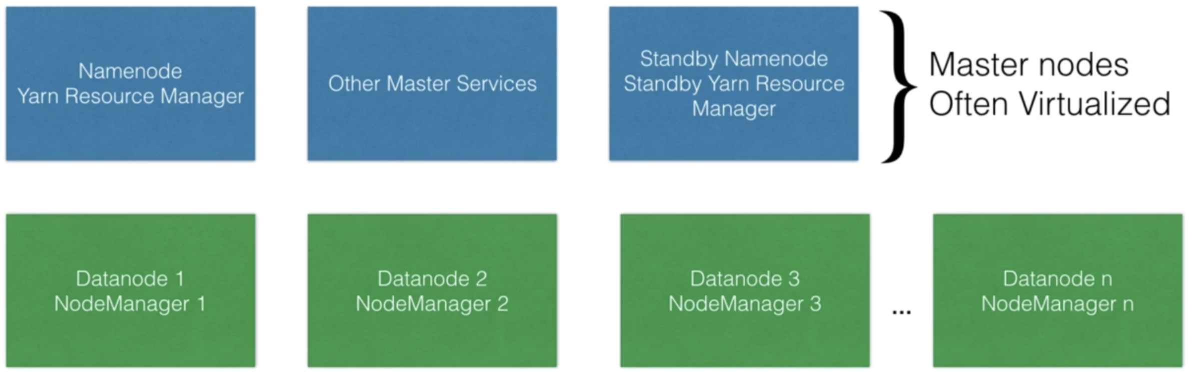
1. YARN - interface to which we submit our applications to - Manages cluster's CPU and Memory
2. Applications can be submitted using any of the processing engines built on top of Hadoop.
3. Eg of Application - WordCounter for a text file stored in HDFS using MapReduce.

Hive uses SQL | Spark implemented using Python, Java, R, Scala.

Hadoop as a Distributed System.

In the following example, the data(worker) nodes can be scaled horizontally whereas the Master nodes cannot.

- The data nodes / worker nodes (in green) can scale horizontally



Namenode
Yarn Resource Manager

Other Master Services

Standby Namenode
Standby Yarn Resource Manager

Master nodes
Often Virtualized

Datanode 1
NodeManager 1

Datanode 2
NodeManager 2

Datanode 3
NodeManager 3

...

Datanode n
NodeManager n

NameNode works on Master Node, Node Manager runs on Worker Node.

Working with HADOOP-

→ Since Hadoop is a distributed system its operation cannot be compared to a single machine operation.

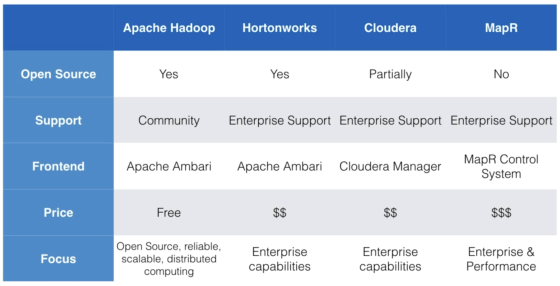
→ A Developer cannot assume that he has access to all the data at the same time, So they have to be able to write data processing applications for Hadoop in a different manner.

→ Implemented as a cluster of nodes in a distributed system.

→It can automatically recover from node failures [Code within the application possibly gets executed twice]

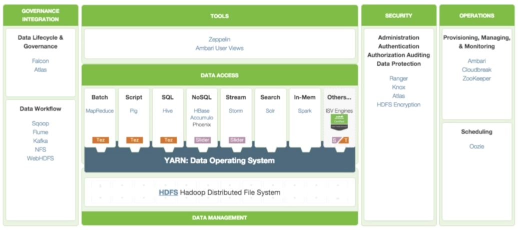
Hadoop replaces traditional Databases - as they cannot scale horizontally. [Cost of Storage is Linear for Hadoop and exponential for Databases.]

Hadoop Distributions



	Apache Hadoop	Hortonworks	Cloudera	MapR
Open Source	Yes	Yes	Partially	No
Support	Community	Enterprise Support	Enterprise Support	Enterprise Support
Frontend	Apache Ambari	Apache Ambari	Cloudera Manager	MapR Control System
Price	Free	$$	$$	$$$
Focus	Open Source, reliable, scalable, distributed computing	Enterprise capabilities	Enterprise capabilities	Enterprise & Performance

Hortonworks Data Platform



Source: hortonworks.com

To run Hadoop on your laptop, you need a virtual machine, A Hadoop cluster requires minimum of 3 nodes.

Hadoop = Vagrant + VirtualMachine

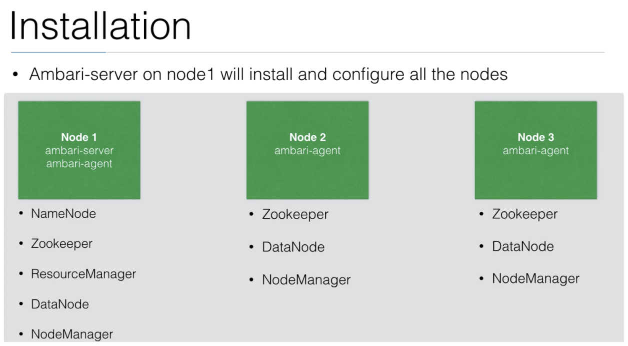
Vagrant - to create and configure virtual development environments.

→ It is lightweight, Reproducible and Portable

→It is basically a wrapper around the VirtualBox/KVM/AWS/Docker/VMWare or HyperV.

→ It helps in creating identical development environments for Operations and Developers.

→Environments created are disposable.



Installation

- Ambari-server on node1 will install and configure all the nodes

Node 1
ambari-server
ambari-agent

- NameNode
- Zookeeper
- ResourceManager
- DataNode
- NodeManager

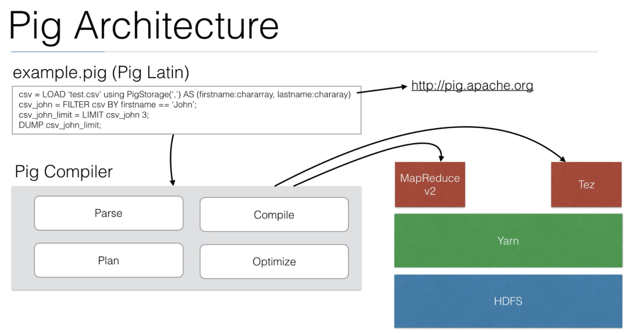
Node 2
ambari-agent

- Zookeeper
- DataNode
- NodeManager

Node 3
ambari-agent

- Zookeeper
- DataNode
- NodeManager

PIG



Pig Architecture

example.pig (Pig Latin)

```
csv = LOAD ‘test.csv’ using PigStorage(‘,’) AS (firstname:chararray, lastname:chararay)
csv_john = FILTER csv BY firstname == ‘John’;
csv_john_limit = LIMIT csv_john 3;
DUMP csv_john_limit;
```

http://pig.apache.org

Pig Compiler

Parse

Compile

Plan

Optimize

MapReduce v2

Tez

Yarn

HDFS