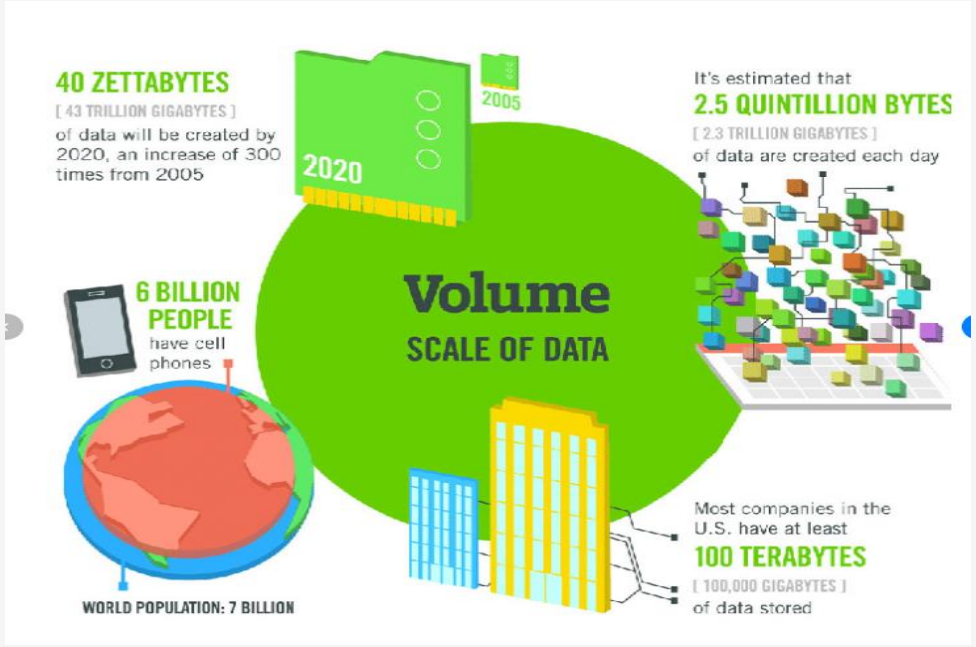
Hadoop Concepts

What is Big Data ?

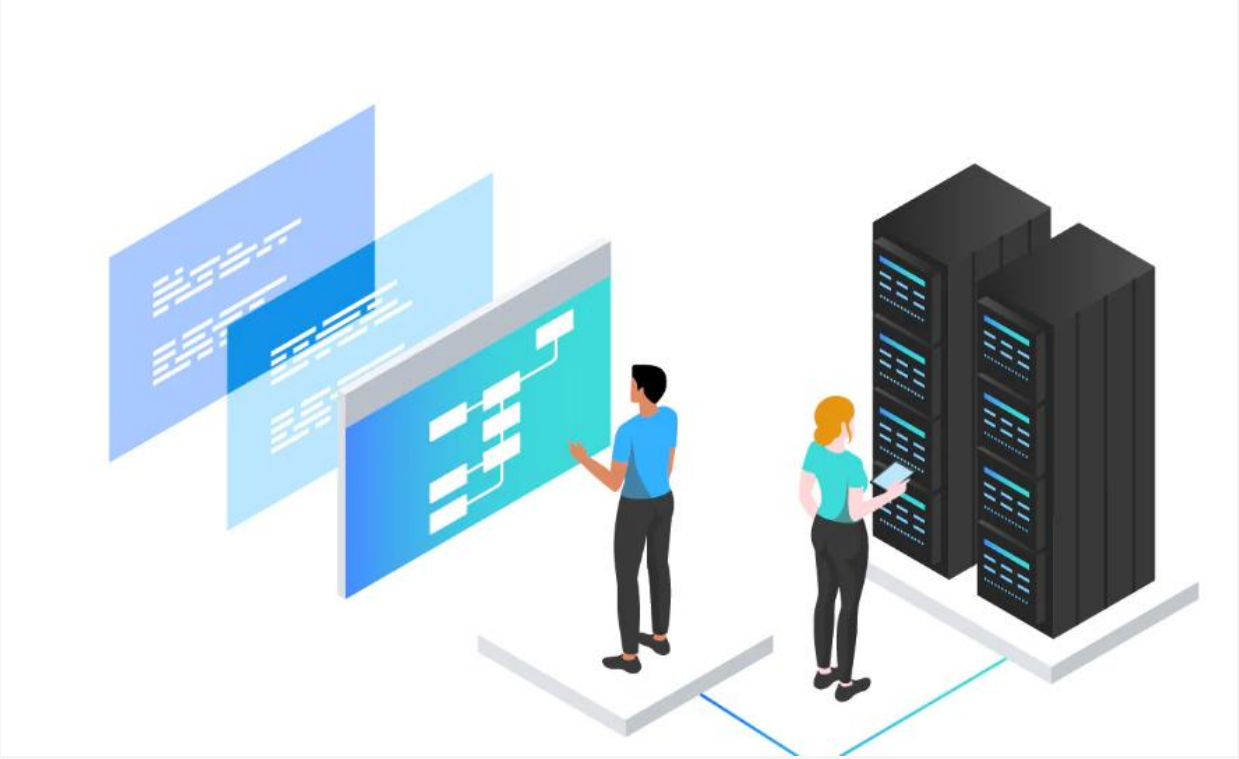
It describes basically Huge volumes of Data .

- Volume – TB ,Trillion, Millions of data generated each day
- Variety – Structured(oracle ,Mysql),semi structured(XML,CSV,JSON) unstructured(Audio,Video,image,log files)
- Velocity – 900 Million photos ,600 million tweets etc
- Veracity – Poor data and unclean data



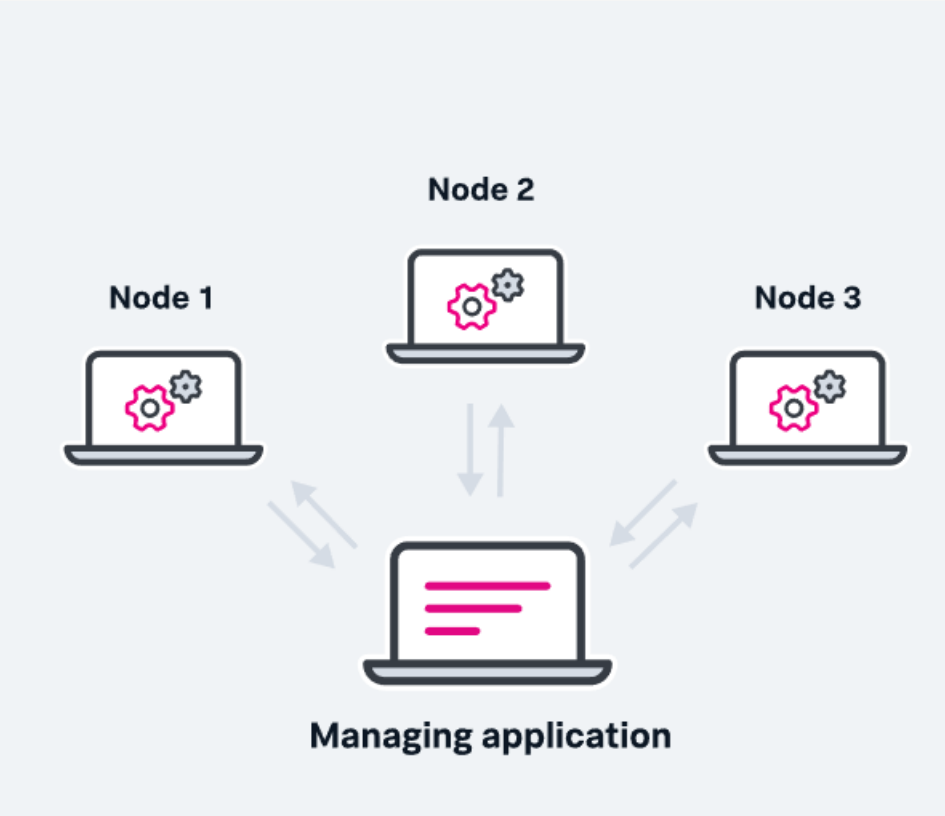
Why Industry is Moving to Big data

- **Data Processing** – Process huge volumes of data in a timely manner
- **Data storage** – Massive data storage where traditional system can't able to hold
- **Scale** – scale easily as data grows but traditional systems have that challenges with serious limitations



Distributed system

- Build a system where many smaller systems comes together in a distributed fashion where you can add resources.
- Distributed systems are linearly scalable
- Increasing the resources increased the speed of processing -→ 2X resources= 2X speed
- Horizontal scaling
- Big Data systems are based on distributed Architecture

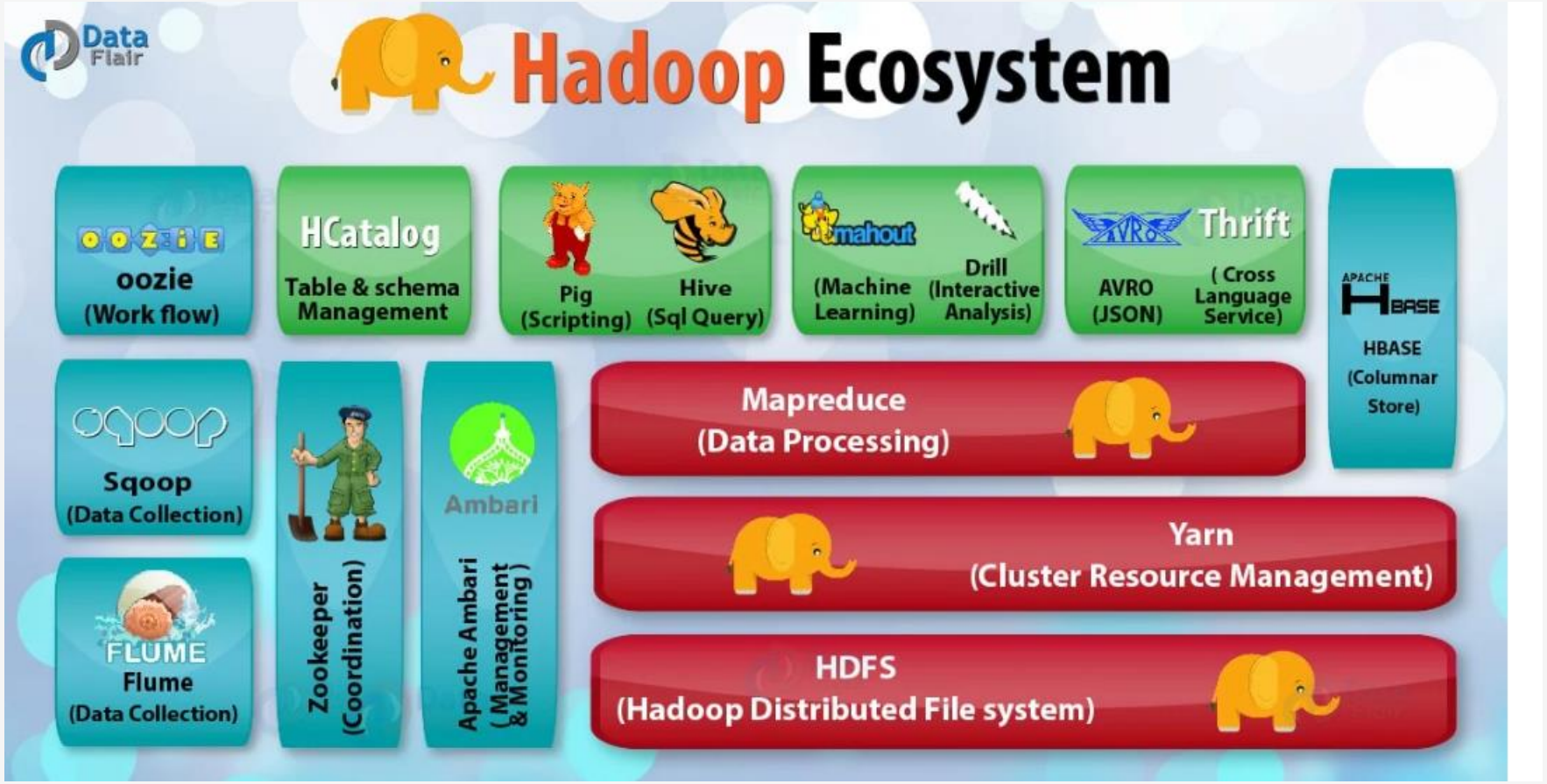


A distributed system is a computing environment in which various components are spread across multiple computers (or other computing devices) on a network. These devices split up the work, coordinating their efforts to complete the job more efficiently than if a single device had been responsible for the task.

Hadoop

- **HDFS refers to Hadoop Distributed File System**
- **Map Reduce is for Distributed Data Processing**
- **Hadoop 1.0 will have only HDFS & Map Reduce**
- **Hadoop 2.0 Released to address major performance improvements [Map Reduce,Yarn,HDFS]**
- **YARN – Yet Another Resource Negotiator [mainly responsible for resource Management]**

Hadoop Ecosystem



Hadoop Ecosystem [Name Node Vs Data Node]

Name Node :

It is also known as *Master* node. NameNode does not store actual data or dataset. NameNode stores Metadata i.e. number of **blocks**, their location, on which Rack, which Datanode the data is stored and other details. It consists of files and directories.

Tasks :

•Manage file system namespace.
•Regulates client's access to files.
•Executes file system execution such as naming, closing, opening files and directories.

Data Node :

It is also known as *Slave*. HDFS Datanode is responsible for storing actual data in HDFS. Datanode performs **read and write operation** as per the request of the clients. Replica block of Datanode consists of 2 files on the file system. The first file is for data and second file is for recording the block's metadata. HDFS Metadata includes checksums for data. At startup, each Datanode connects to its corresponding Namenode and does handshaking. Verification of namespace ID and software version of DataNode take place by handshaking. At the time of mismatch found, DataNode goes down automatically.

Tasks :

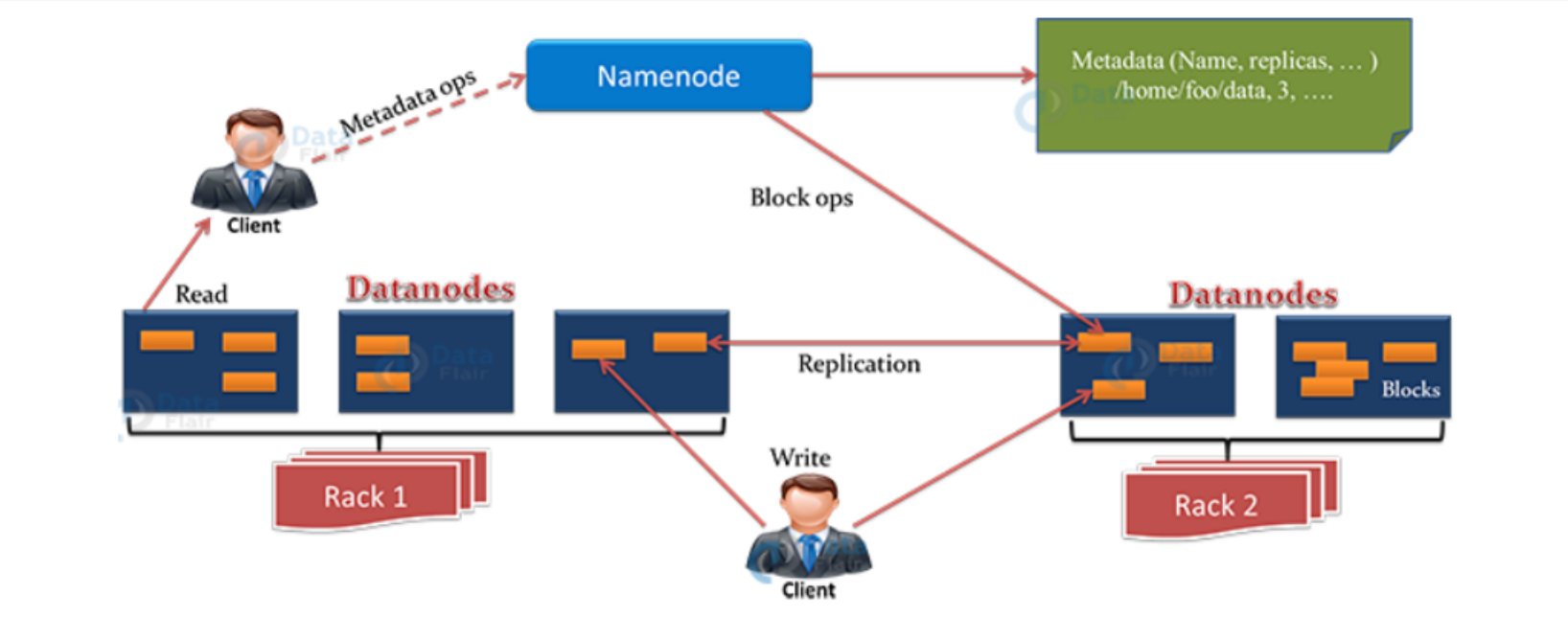
•DataNode performs operations like block replica creation, deletion, and replication according to the instruction of NameNode.
•DataNode manages data storage of the system.

HDFS Architecture

Blocks : HDFS splits huge files into small chunks known as blocks. Block is the smallest unit of data in a filesystem. We (client and admin) do not have any control on the block like block location. NameNode decides all such things.
HDFS default block size is 128 MB. We can increase or decrease the block size as per our need. This is unlike the OS filesystem, where the block size is 4 KB.
If the data size is less than the block size of HDFS, then block size will be equal to the data size.

Hadoop Default Replication factor - 3

Heart Beat :Each data node sends heart beats name node in every 3 secs



Fault Tolerance

Secondary Name node – performs FS image and edit logs merging to get new updated FS image

Checkpointing

Rack Awareness – group of systems placed in different geographical locations. Namenode stores data block in a data node of a rack

Thank You !!!!!!!!!!