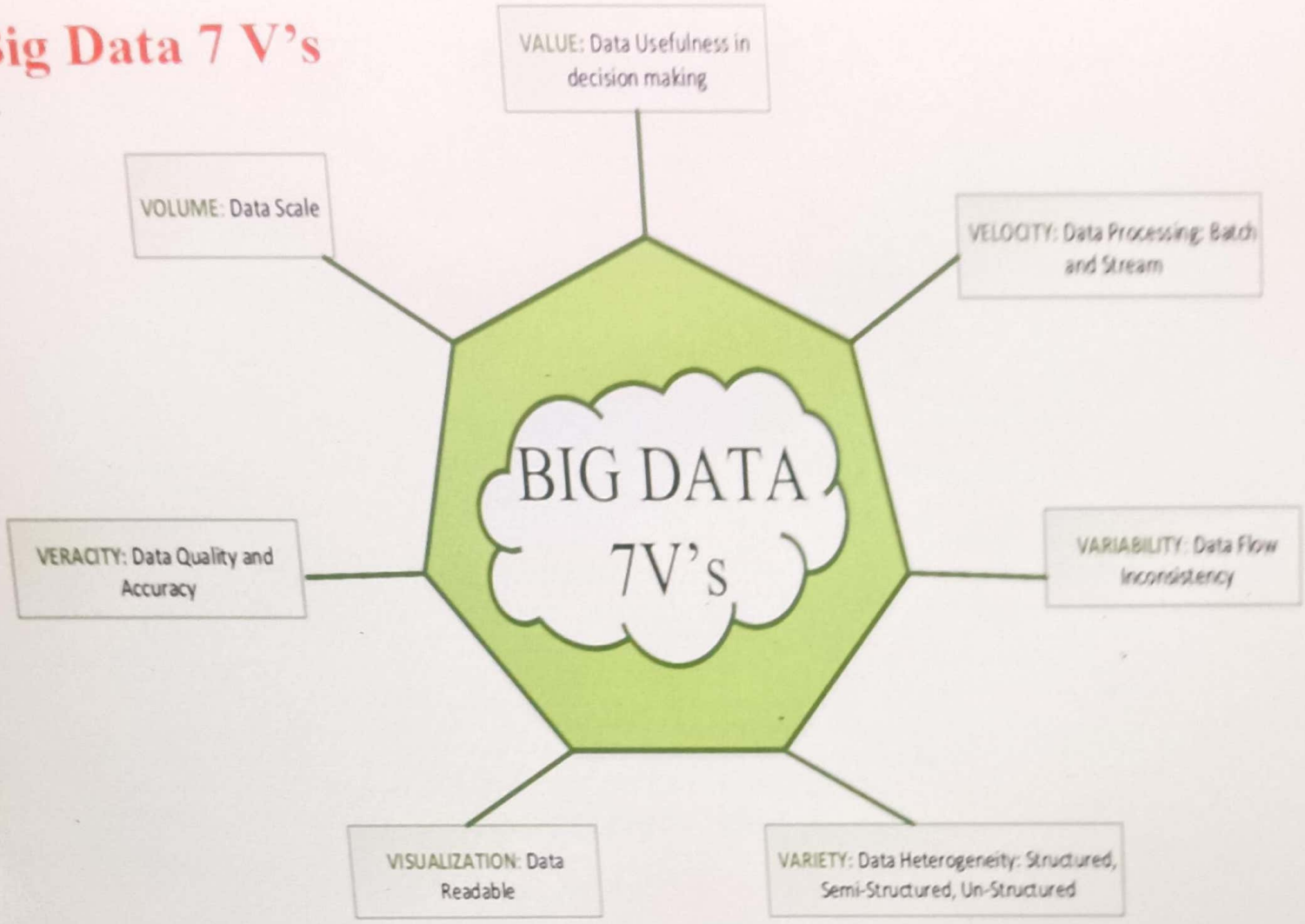
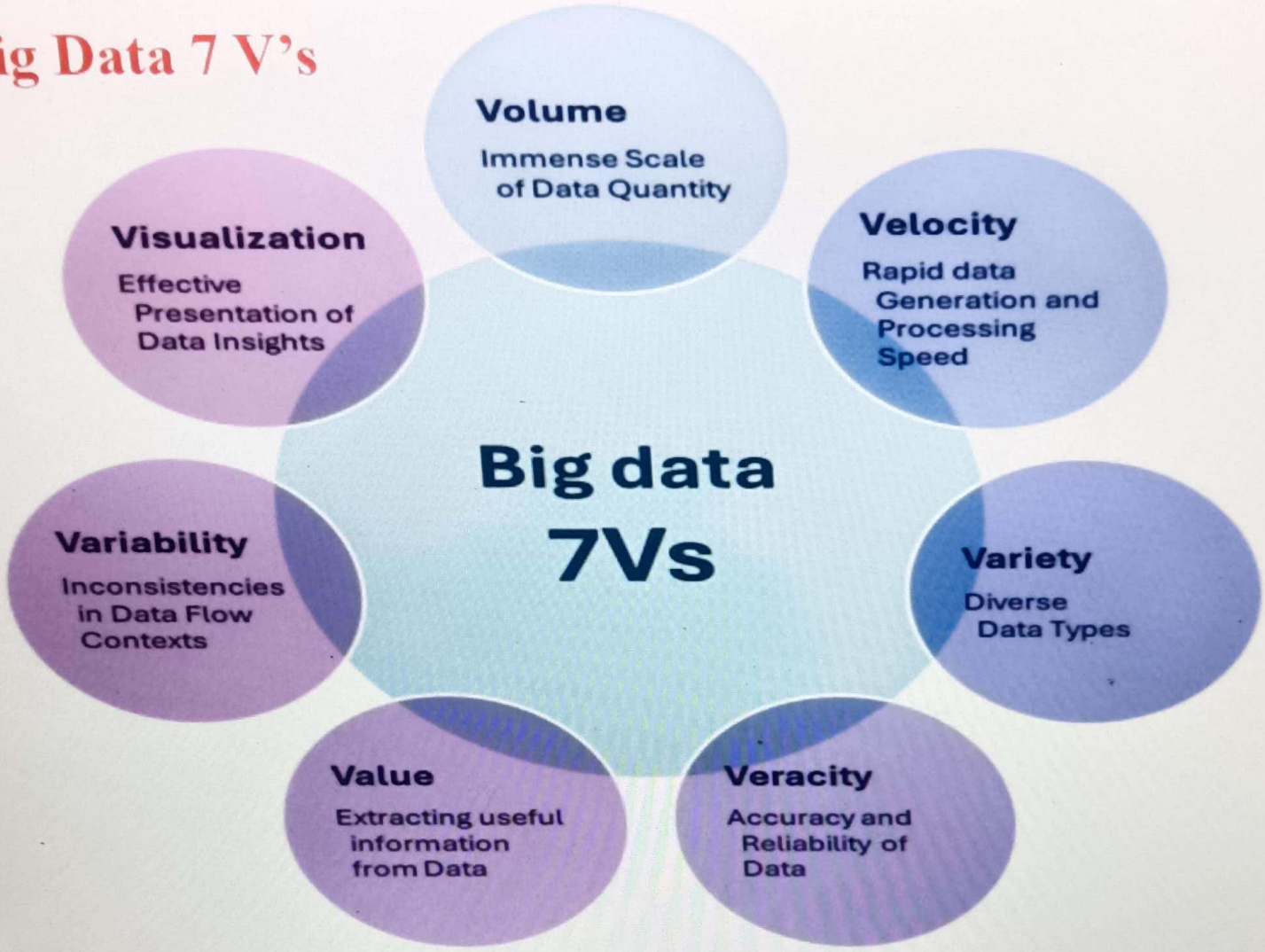
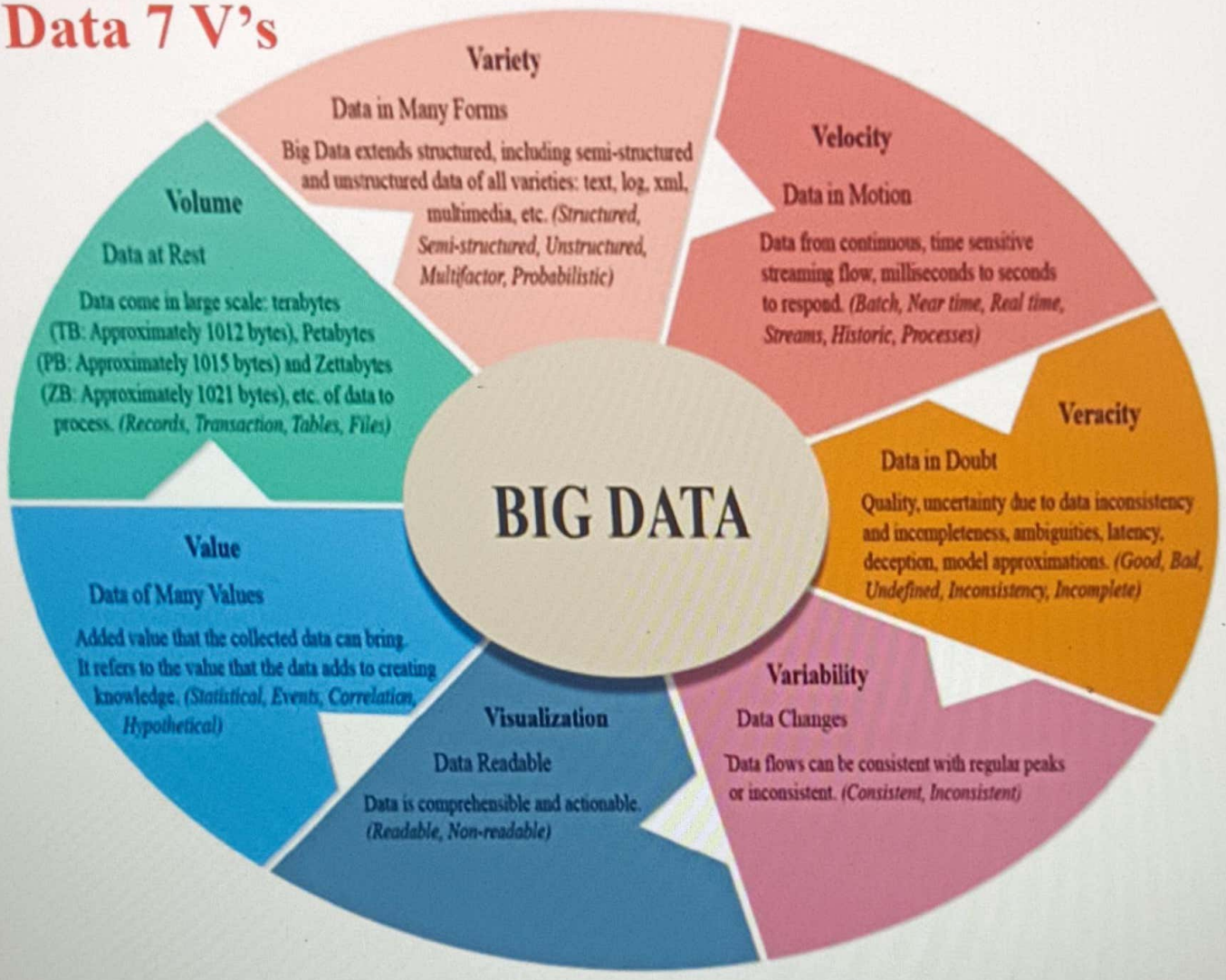
Big Data 7 V's


VALUE: Data Usefulness in decision making
VOLUME: Data Scale
VELOCITY: Data Processing: Batch and Stream
BIG DATA 7V's
VERACITY: Data Quality and Accuracy
VARIABILITY: Data Flow Inconsistency
VISUALIZATION: Data Readable
VARIETY: Data Heterogeneity: Structured, Semi-Structured, Un-Structured

Big Data 7 V's


Big data
7Vs
Volume
Immense Scale of Data Quantity
Velocity
Rapid data Generation and Processing Speed
Variety
Diverse Data Types
Veracity
Accuracy and Reliability of Data
Value
Extracting useful information from Data
Variability
Inconsistencies in Data Flow Contexts
Visualization
Effective Presentation of Data Insights

Big Data 7 V's


Variety
Data in Many Forms
Big Data extends structured, including semi-structured and unstructured data of all varieties: text, log, xml, multimedia, etc. (Structured, Semi-structured, Unstructured, Multifactor, Probabilistic)
Velocity
Data in Motion
Data from continuous, time sensitive streaming flow, milliseconds to seconds to respond. (Batch, Near time, Real time, Streams, Historic, Processes)
Volume
Data at Rest
Data come in large scale: terabytes (TB: Approximately 1012 bytes), Petabytes (PB: Approximately 1015 bytes) and Zettabytes (ZB: Approximately 1021 bytes), etc. of data to process. (Records, Transaction, Tables, Files)
BIG DATA
Veracity
Data in Doubt
Quality, uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception, model approximations. (Good, Bad, Undefined, Inconsistency, Incomplete)
Value
Data of Many Values
Added value that the collected data can bring. It refers to the value that the data adds to creating knowledge. (Statistical, Events, Correlation, Hypothetical)
Variability
Data Changes
Data flows can be consistent with regular peaks or inconsistent. (Consistent, Inconsistent)
Visualization
Data Readable
Data is comprehensible and actionable. (Readable, Non-readable)

Hadoop

- The Apache Hadoop software library is a framework for distributed processing of large data sets across clusters of computers using simple programming models.
- Hadoop is an open-source software programming framework. The framework of Hadoop is based on Java Programming Language with some native code in shell script and C language.

Hadoop

- The Apache Hadoop software library is a framework for distributed processing of large data sets across clusters of computers using simple programming models.
- Hadoop is an open-source software programming framework. The framework of Hadoop is based on Java Programming Language with some native code in shell script and C language.

Hadoop

- Hadoop is designed to scale up from single servers to thousands of machines, each offering local computation and storage.
- Hadoop can efficiently process all types of data, namely structured data, unstructured, or semi-structured data.
- Hadoop uses Transmission Control Protocol and User Datagram Protocol for communication.

Hadoop

- Rather than relying 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.
- Hadoop provides a flexible and powerful solution for Extract, Transform and Load processing.

Hadoop

- Hadoop is used for storage and processing of big data sets on clusters of commodity hardware. The Hadoop framework includes the following:
  - ✓ Hadoop Distributed File System (HDFS)
  - ✓ Hadoop Yet Another Resource Negotiator (YARN)
  - ✓ Hadoop MapReduce

Hadoop

- A Single Node Hadoop Cluster has all Hadoop Daemons namely Name Node, Data Node, Secondary Name Node, Resource Manager, and Node Manager run on a single machine.

Hadoop

- Hadoop Distributed File System (HDFS) is a high performance distributed file system.
- Hadoop YARN is a framework for job scheduling and cluster resource management.
- Hadoop MapReduce is a system for parallel processing of large data sets that implements the MapReduce model of distributed programming.

Hadoop

- Hadoop is designed to process large volumes of data by dividing the data into smaller chunks, distributing these chunks across a cluster of computers, and processing them in parallel (distributed processing).
- This ability to divide and conquer makes Hadoop extremely powerful for handling big data.
- By distributing the data, Hadoop can process it in parallel on the nodes where the data is located.

Hadoop

- Moving computation is more efficient than moving large data.
- Redundant and reliable
  - ✓ Hadoop replicates data automatically, so when machine goes down there is no data loss.
- Easy to develop distributed applications
  - ✓ Possible to develop a program to run on one machine and then scale it to thousands of machines without changing the program.

Hadoop

- Runs on commodity hardware
  - ✓ Don't have to buy special hardware, expensive RAIDs, or redundant hardware; reliability is built into software.
  - ✓ No need for super computers with high-end storage.

Hadoop

- In a large Hadoop cluster, there are multiple racks. Each rack consists of Data Nodes.
- Communication between the Data Nodes on the same rack is more efficient as compared to the communication between Data Nodes residing on different racks.

Hadoop

- Rack is the collection of around 40 to 50 Data Nodes connected using the same network switch. If the network goes down, the whole rack will be unavailable. A large Hadoop cluster is deployed in multiple racks.

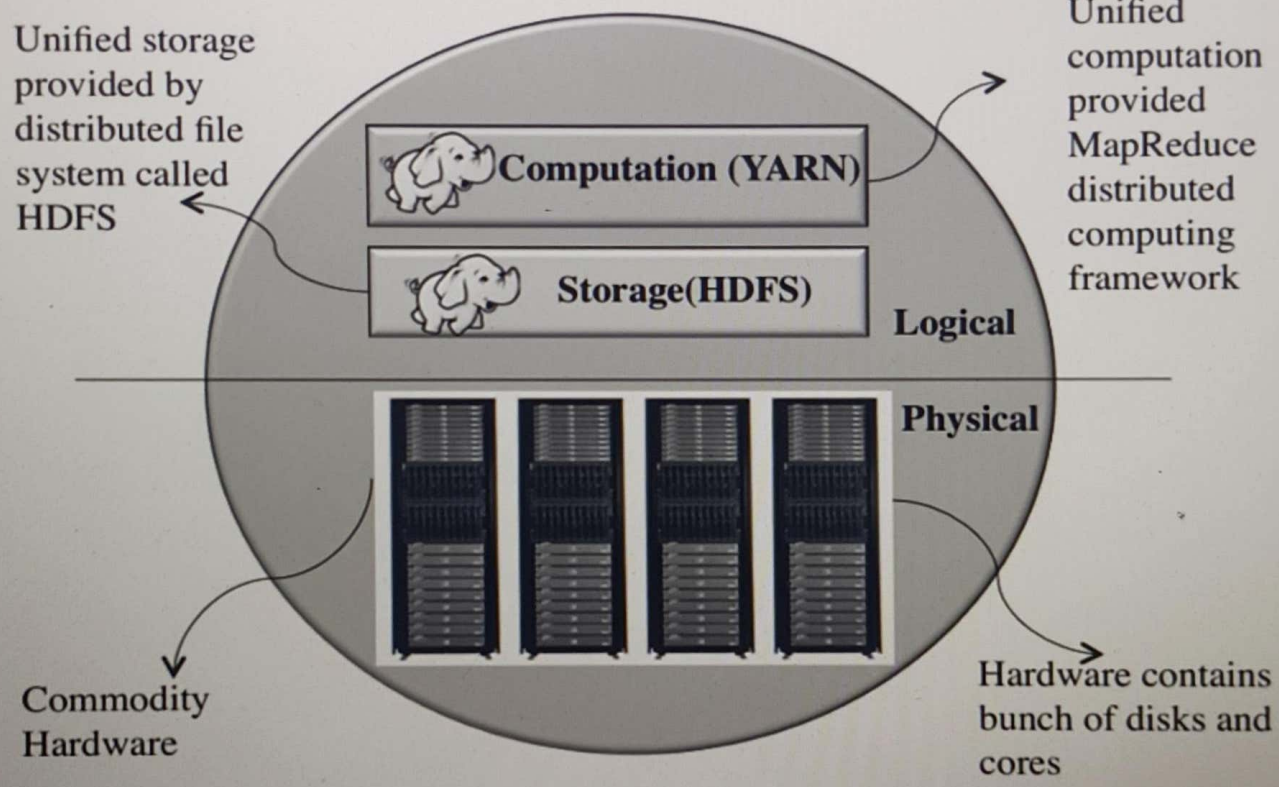
Hadoop

- To achieve the maximum performance from Hadoop and to reduce the network traffic during file read / write, Name Node chooses the Data Nodes on the same rack or nearby racks for data read / write.
- Rack awareness is the concept of choosing the closer DataNode based on rack information.

Hadoop

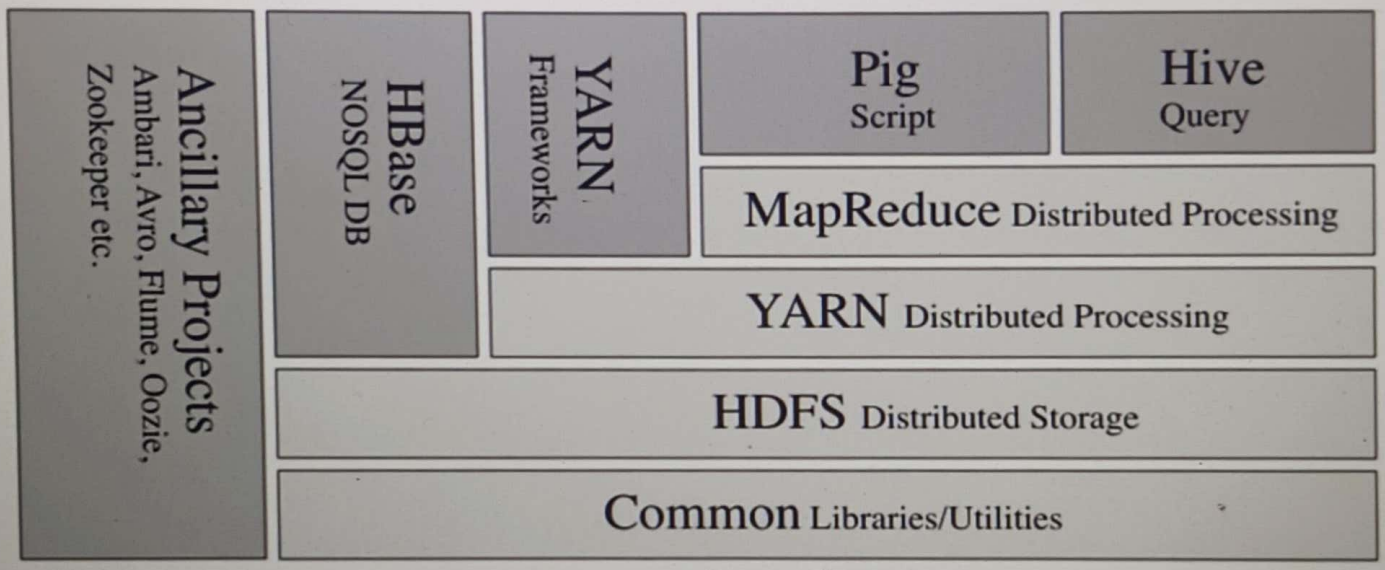
- Data is organized into files and directories. Files are divided into uniform sized blocks and distributed across cluster nodes.
- Blocks are replicated to handle failure.
- Checksums of data are used for corruption detection and recovery.

Hadoop


Unified storage provided by distributed file system called HDFS
Computation (YARN)
Storage(HDFS)
Logical
Physical
Unified computation provided MapReduce distributed computing framework
Commodity Hardware
Hardware contains bunch of disks and cores

Hadoop Technology Stack

Ramanujan Computing Centre

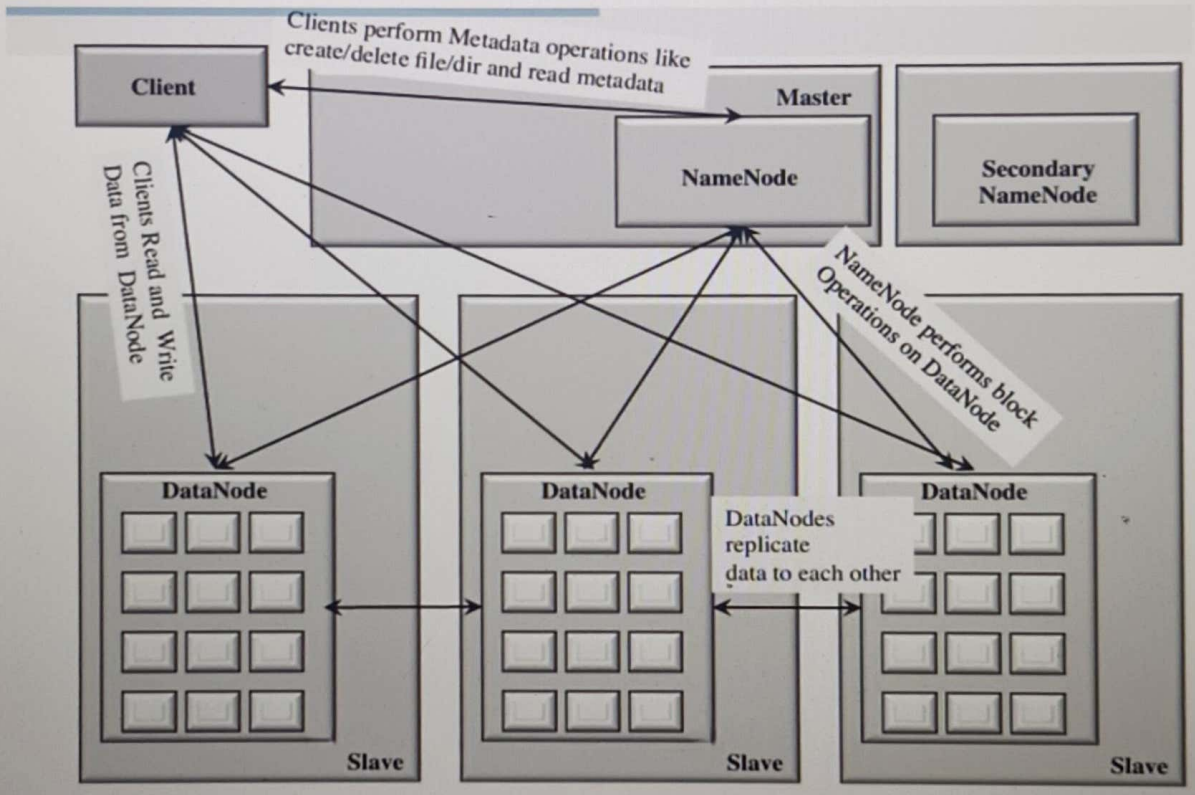


Core Hadoop Modules

Ancillary Projects

HDFS Architecture

Ramanujan Computing Centre



HDFS Architecture

- HDFS uses a Master-Slave architecture with Name Node as Master and Data Node as Slave.
- Each cluster comprises a single Master Node and multiple Slave nodes.
- HDFS breaks Data / Files into small blocks (128 MB each block) and stores on Data Node and each block replicates on other nodes to accomplish fault tolerance.
- Name Node keeps track of blocks written to the Data Node.