SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Department of Information Technology

19ITE305 – BIG DATA ANALYTICS

III B.Tech. IT/ VI SEMESTER

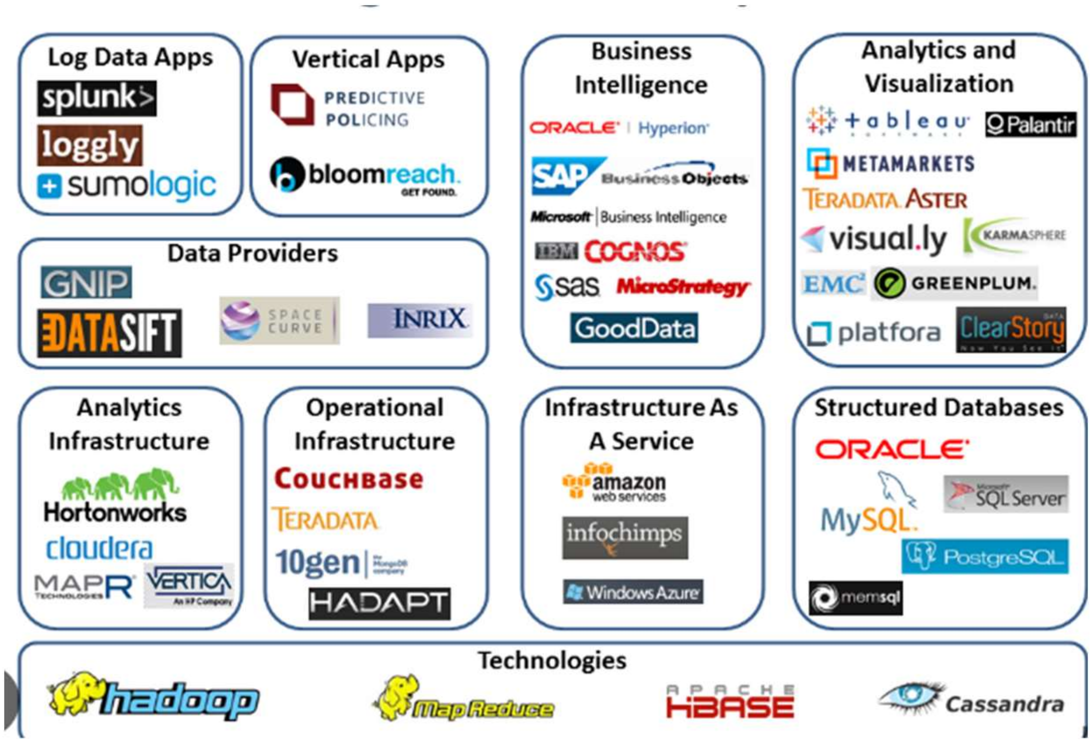
UNIT II : INTRODUCTION TO TECHNOLOGY LANDSCAPE

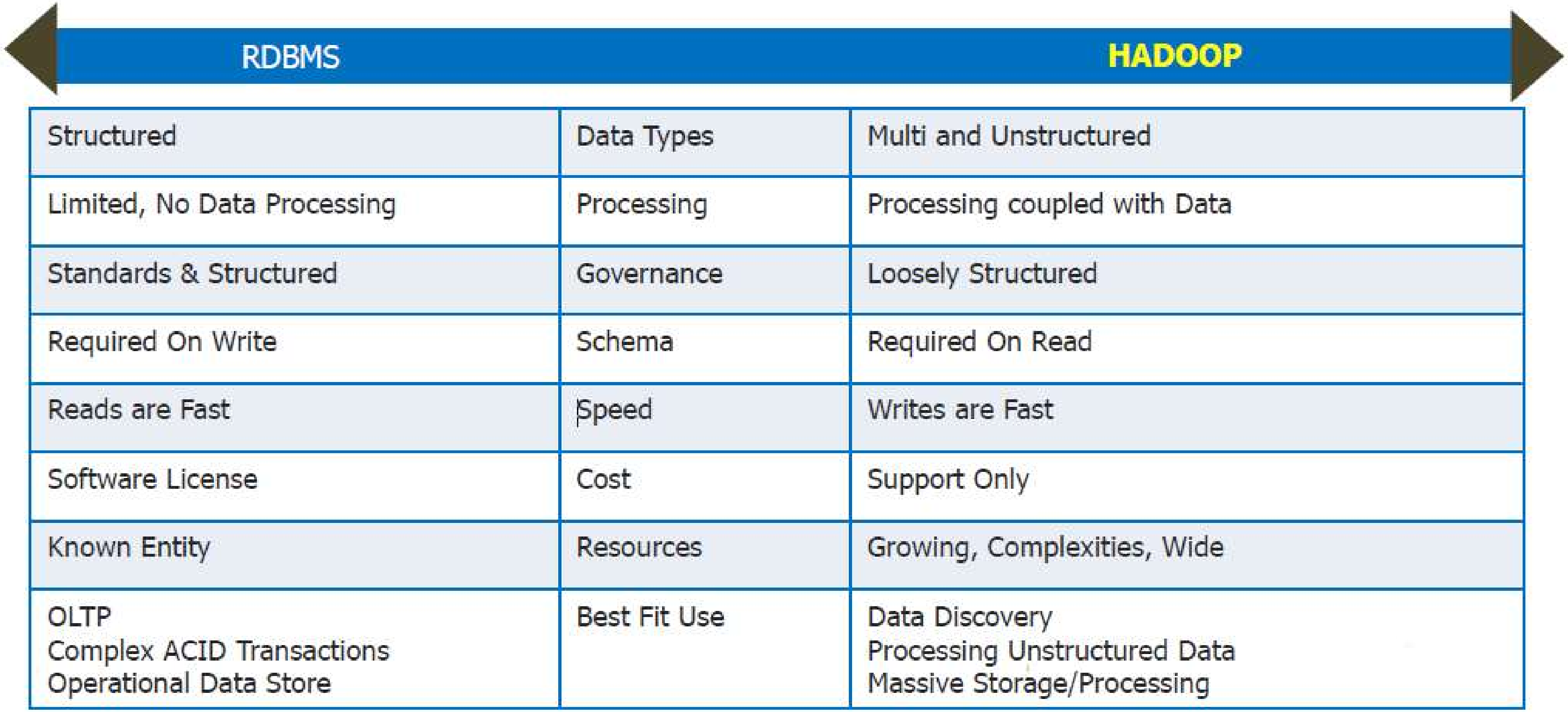
Topic 4 : Hadoop Overview

NoSQL, Comparison of SQL and NoSQL, Hadoop - RDBMS Versus Hadoop - Distributed Computing Challenges – Hadoop Overview - Hadoop Distributed File System - Processing Data with Hadoop - Managing Resources and Applications with Hadoop YARN - Interacting with Hadoop Ecosystem

Big Data Technology Landscape

- NoSQL
- Hadoop





RDBMS		HADOOP
Structured	Data Types	Multi and Unstructured
Limited, No Data Processing	Processing	Processing coupled with Data
Standards & Structured	Governance	Loosely Structured
Required On Write	Schema	Required On Read
Reads are Fast	Speed	Writes are Fast
Software License	Cost	Support Only
Known Entity	Resources	Growing, Complexities, Wide
OLTP Complex ACID Transactions Operational Data Store	Best Fit Use	Data Discovery Processing Unstructured Data Massive Storage/Processing

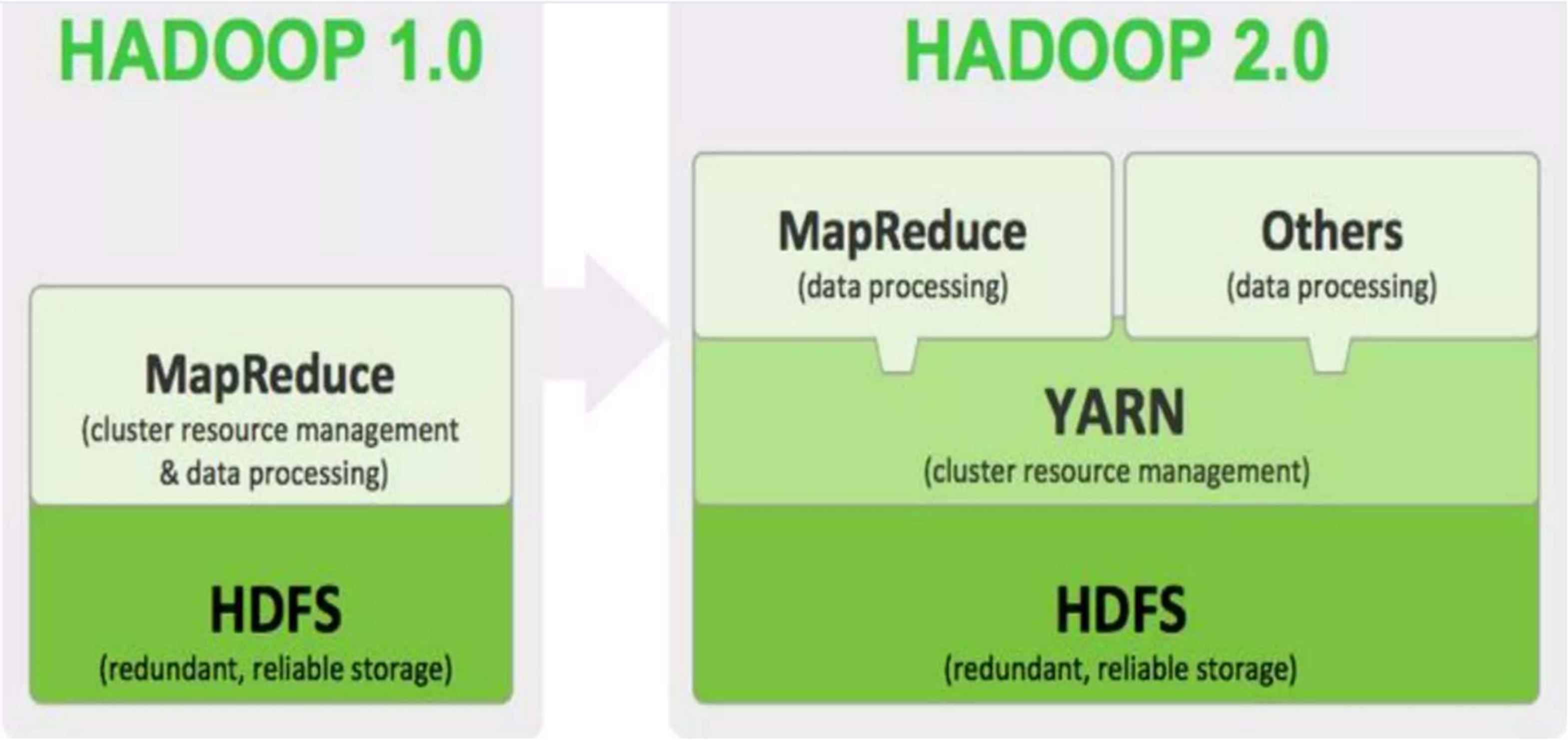
Key advantage of Hadoop

1. Stores data in its native format (HDFS).
2. Scalable : store and distribute very large clusters
3. Cost effective: reduced cost/TB of storage and processing.
4. Resilient to failure : Fault tolerant due to data replication on multiple nodes in the cluster.

Key advantage of Hadoop

5. Flexibility: supports any data (SD, SSD and USD) analysis such as email conversations, social media data analysis, click-stream data analysis, log analysis, data mining, market campaign analysis, etc.
6. Fast: move code to data paradigm. [Process Migration]

Version of Hadoop



HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)

YARN -> Yet Another Resource Negotiator

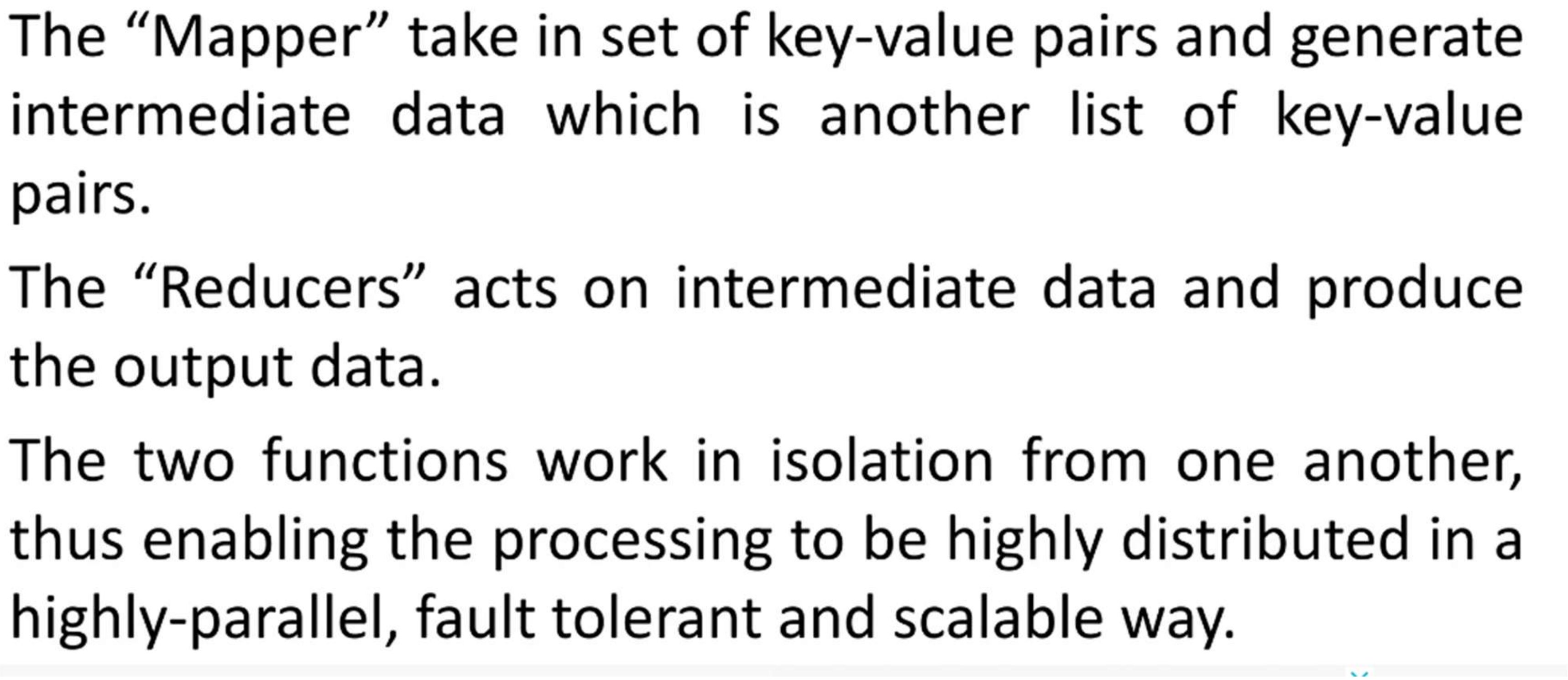
Hadoop 1.0

1. Data storage framework (HDFS):
   - General purpose file system.
   - It is schema-less and stores data files of any format.
   - Stores data files as close to their original format and this provides needed flexibility and agility.

Hadoop 1.0

2. Data processing framework:
   - MapReduce model (Google’s popular model)
   - Uses two functions: Map and Reduce functions to process the data.

Data Processing Framework

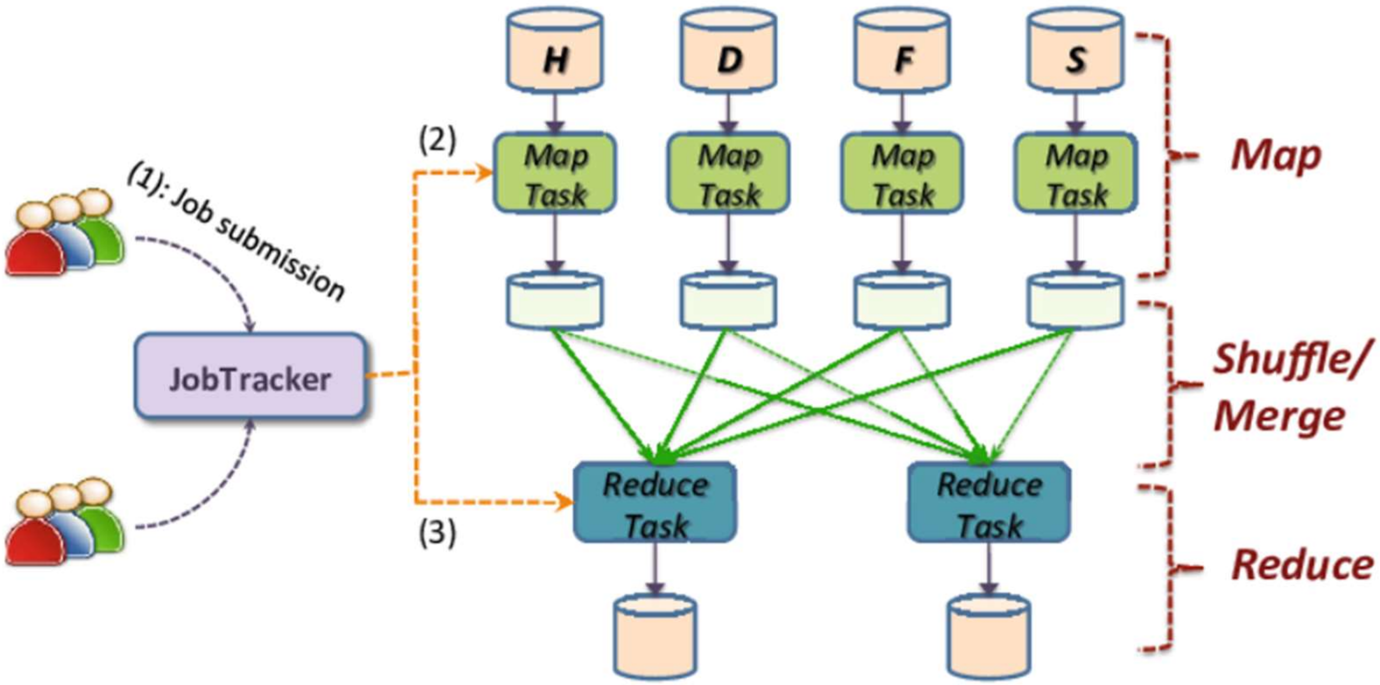


The “Mapper” take in set of key-value pairs and generate intermediate data which is another list of key-value pairs.

The “Reducers” acts on intermediate data and produce the output data.

The two functions work in isolation from one another, thus enabling the processing to be highly distributed in a highly-parallel, fault tolerant and scalable way.

Data Processing Framework



Limitations

Requires expertise in MapReduce programming and Java.

It supports only batch processing.

It is tightly coupled with MapReduce and hence every data for analysis has to be transformed into MapReduce structure.

Hadoop 2.0

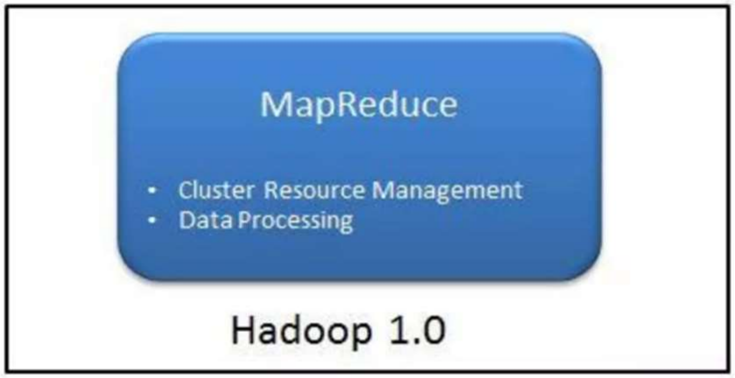
- Apache **Hadoop 2** (**Hadoop 2.0**) is the second iteration of the **Hadoop** framework for distributed data processing.
- **Hadoop 2** adds support for running non-batch applications through the introduction of YARN, a redesigned cluster resource manager that eliminates **Hadoop's** sole reliance on the MapReduce programming model.

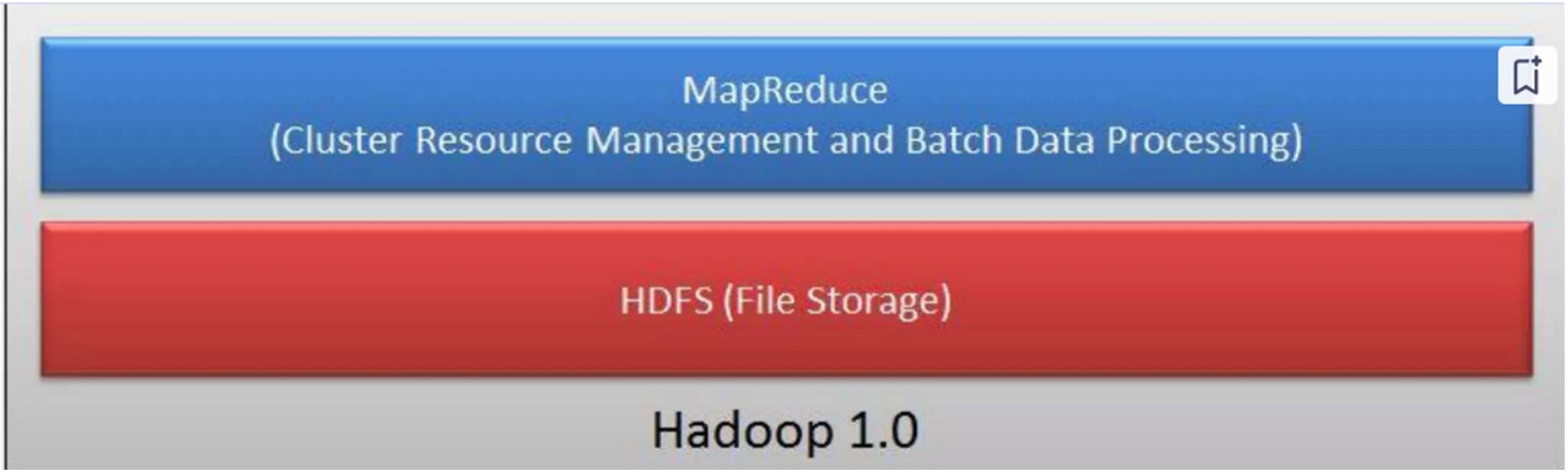
YARN

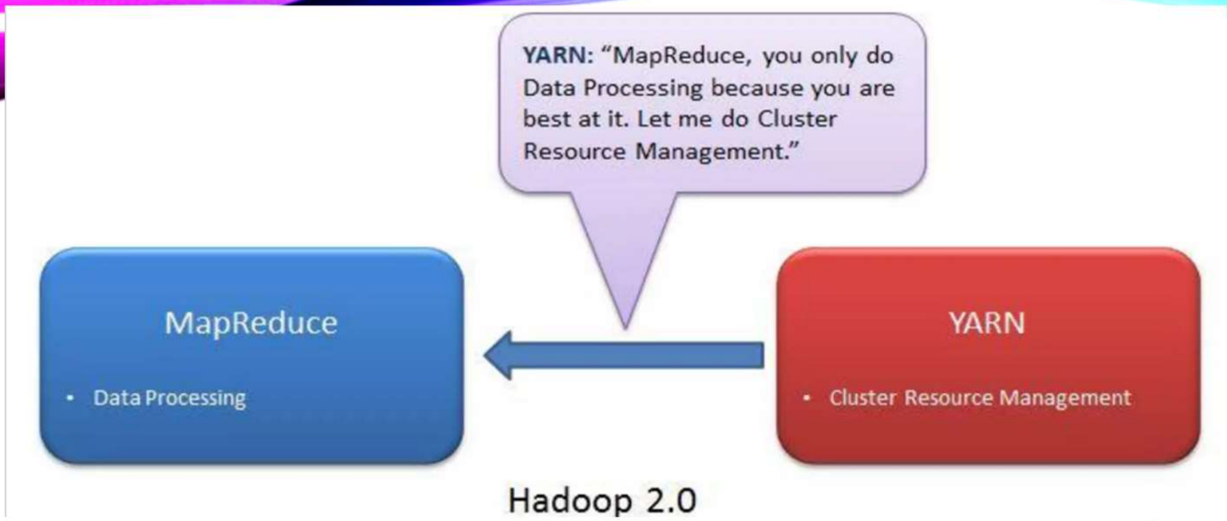
- YARN framework is responsible for Cluster resource management.
- Cluster resource management means managing the resources of the Hadoop Clusters. Resources means Memory, CPU etc.
- YARN took over task of cluster management from MapReduce and MapReduce is streamlined to perform Data Processing only in which it is best.

YARN

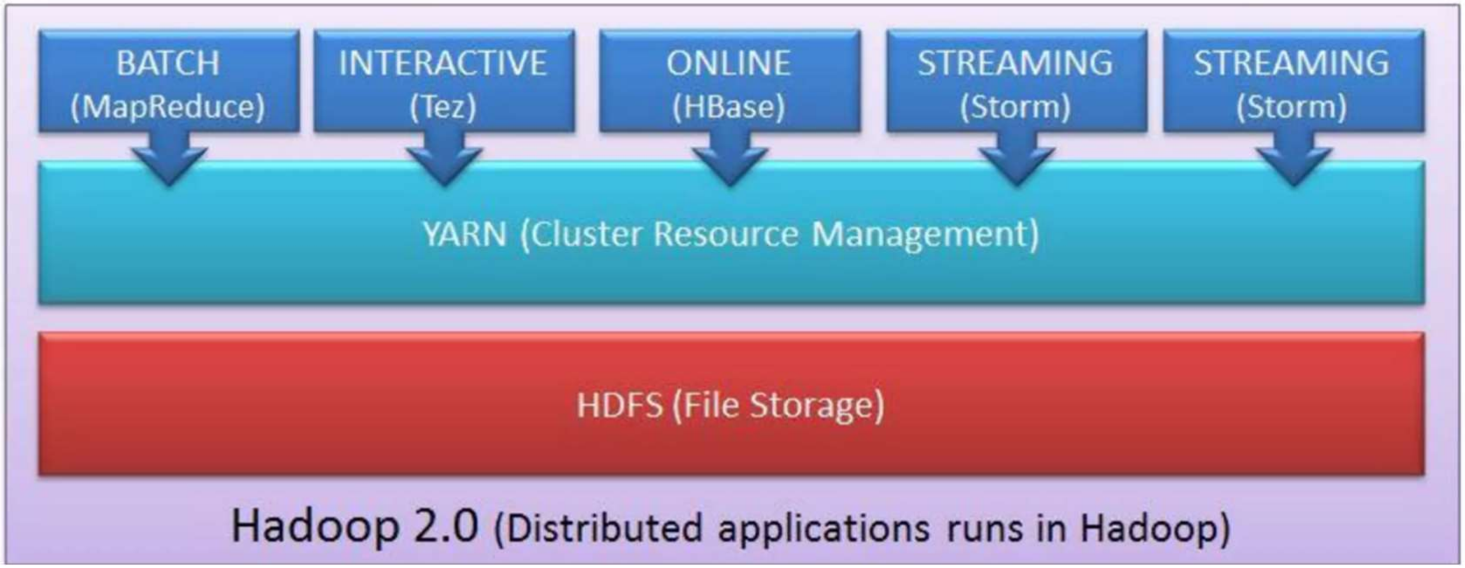
- YARN is the brain of Hadoop Ecosystem. It performs all the processing activities by allocating resources and scheduling tasks.
- YARN co-ordinates the allocation of subtasks of the submitted applications, thus enhances flexibility, scalability and efficiency of the applications.


MapReduce
• Cluster Resource Management
• Data Processing
Hadoop 1.0

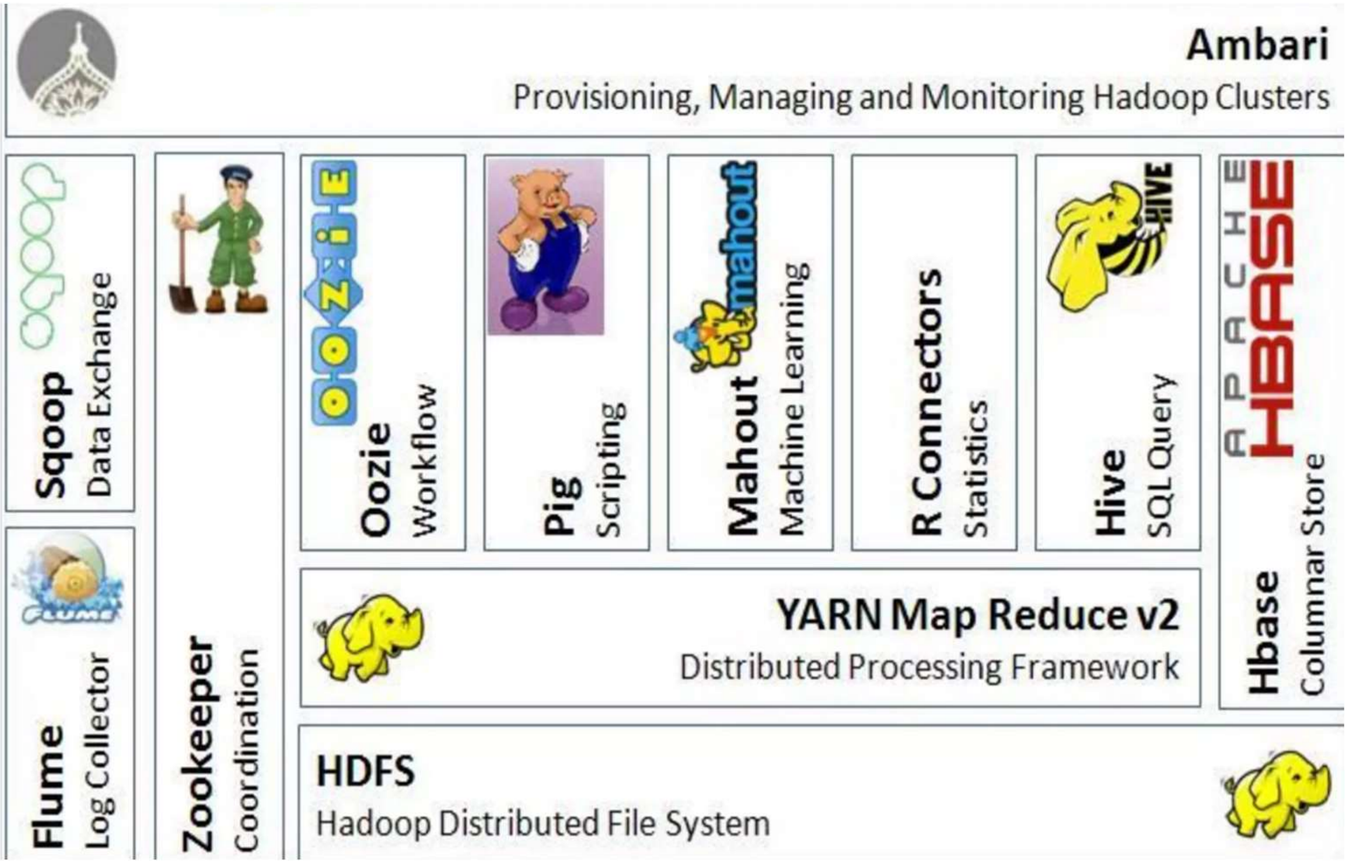

MapReduce
(Cluster Resource Management and Batch Data Processing)
HDFS (File Storage)
Hadoop 1.0



YARN: “MapReduce, you only do Data Processing because you are best at it. Let me do Cluster Resource Management.”



Hadoop Overview



Ambari

Provisioning, Managing and Monitoring Hadoop Clusters

Sqoop

Data Exchange

Flume

Log Collector

Zookeeper

Coordination

Oozie

Workflow

Pig

Scripting

Mahout

Machine Learning

R Connectors

Statistics

Hive

SQL Query

Hbase

Columnar Store

YARN Map Reduce v2

Distributed Processing Framework

HDFS

Hadoop Distributed File System

Hadoop Overview

1. HDFS: It stores different types of large data sets (i.e. structured, unstructured and semi structured data) as close to original form.
2. Hbase (Hadoop's database): HBase is an open source, non-relational distributed database. In other words, it is a NoSQL database.

Hadoop Overview

3. Hive(Facebook): HIVE is a data warehousing component which performs reading, writing and managing large data sets in a distributed environment (Hadoop Cluster) using SQL-like interface. (***HIVE + SQL = HQL)***

Hadoop Overview

4. Pig: It gives a platform for building data flow for ETL processing and analyzing huge data sets.
   - It is also known as Data Flow language.
   - PIG has two parts: **Pig Latin**, the language and **the pig runtime,** for the execution environment. It is similar to Java and JVM.
   - ***10 line of pig latin = approx. 200 lines of Map-Reduce Java code***

Hadoop Overview

5. ZooKeeper: is the coordinator of any Hadoop job which includes a combination of various services in a Hadoop Ecosystem for distributed applications.

Hadoop Overview

6. Oozie: clock and alarm (scheduler) service inside Hadoop Ecosystem.
   – It schedules Hadoop jobs and binds them together as one logical work.
7. Mahout: It provides an environment for creating machine learning applications which are scalable.

Hadoop Overview

8. Flume/Chukwa : which helps in storing unstructured and semi-structured data into HDFS.
   - It is data collection system

9. Sqoop: Import and export structured data from RDBMS or Enterprise data warehouses to HDFS or vice versa.

Hadoop Overview

10. Ambari: It aims at making Hadoop ecosystem more manageable.
    - It is web based tool for **provisioning, managing and monitoring** Apache Hadoop clusters.

TEXT BOOKS

Seema Acharya, Subhashini Chellappan, “Big Data and Analytics”, Wiley Publications, First Edition,2015

REFERENCES

1. Judith Huruwitz, Alan Nugent, Fern Halper, Marcia Kaufman, “Big data for dummies”, John Wiley & Sons, Inc. (2013)
2. Tom White, “Hadoop The Definitive Guide”, O’Reilly Publications, Fourth Edition, 2015
3. Dirk Deroos, Paul C.Zikopoulos, Roman B.Melnky, Bruce Brown, Rafael Coss, “Hadoop For Dummies”, Wiley Publications, 2014
4. Robert D.Schneider, “Hadoop For Dummies”, John Wiley & Sons, Inc. (2012)
5. Paul Zikopoulos, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill, 2012