SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601– Data Science and Analytics

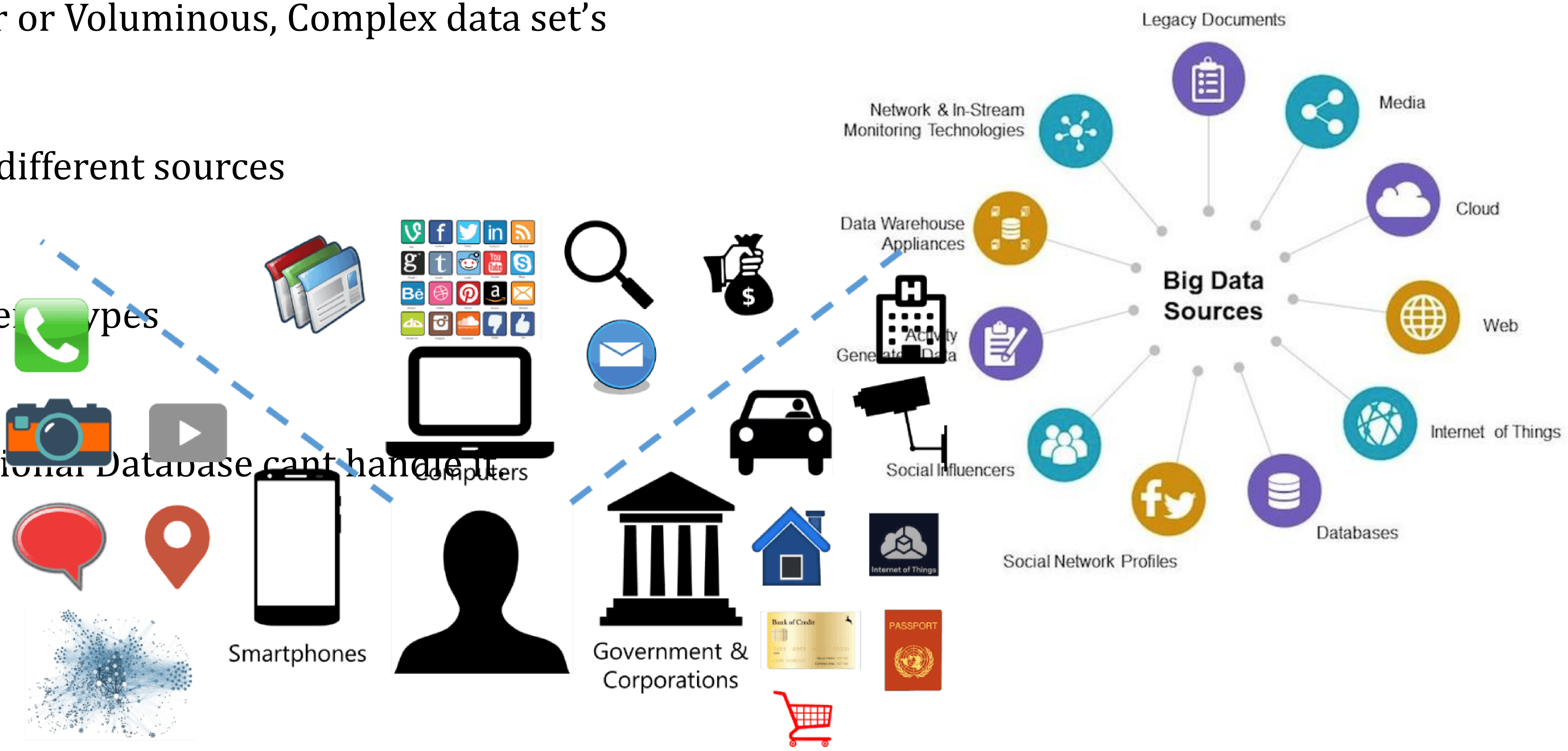
III Year / VI Semester

DATA ANALYTICAL FRAMEWORKS

Topic 1: Hadoop

What is Big Data?

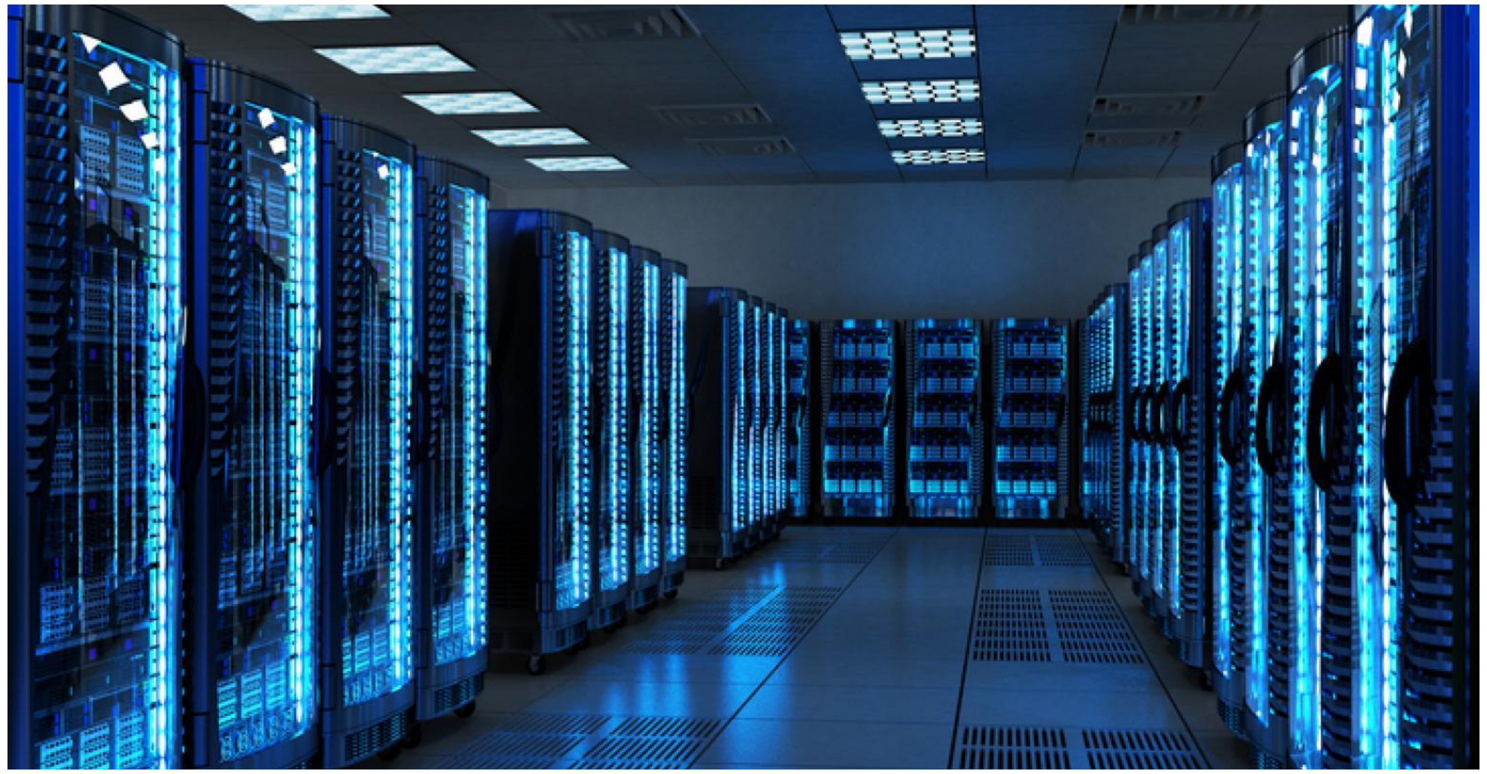
- Larger or Voluminous, Complex data set’s
- From different sources
- Different types
- Traditional Database cant handle it



Hadoop

Why Hadoop?

- Massive Storage



Hadoop

Why Hadoop?

- Faster Processing

Hadoop

- Hadoop is an Apache open source framework
- Written in java
- Allows distributed processing of large datasets across clusters of computers
- Hadoop is an open source software framework
- Used for sorting and processing big data in distributed way

Hadoop

Core Components of Hadoop

Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

Hadoop Distributed File System (HDFS): A distributed file system that provides High-throughput access to application data.

Hadoop MapReduce: A software-programming model for parallel processing of large data sets.

Hadoop Yet Another Resource Negotiator (YARN): This is a framework for job scheduling and cluster resource management. A resource management framework for scheduling and handling resource requests from distributed applications.

Hadoop

Core Components of Hadoop

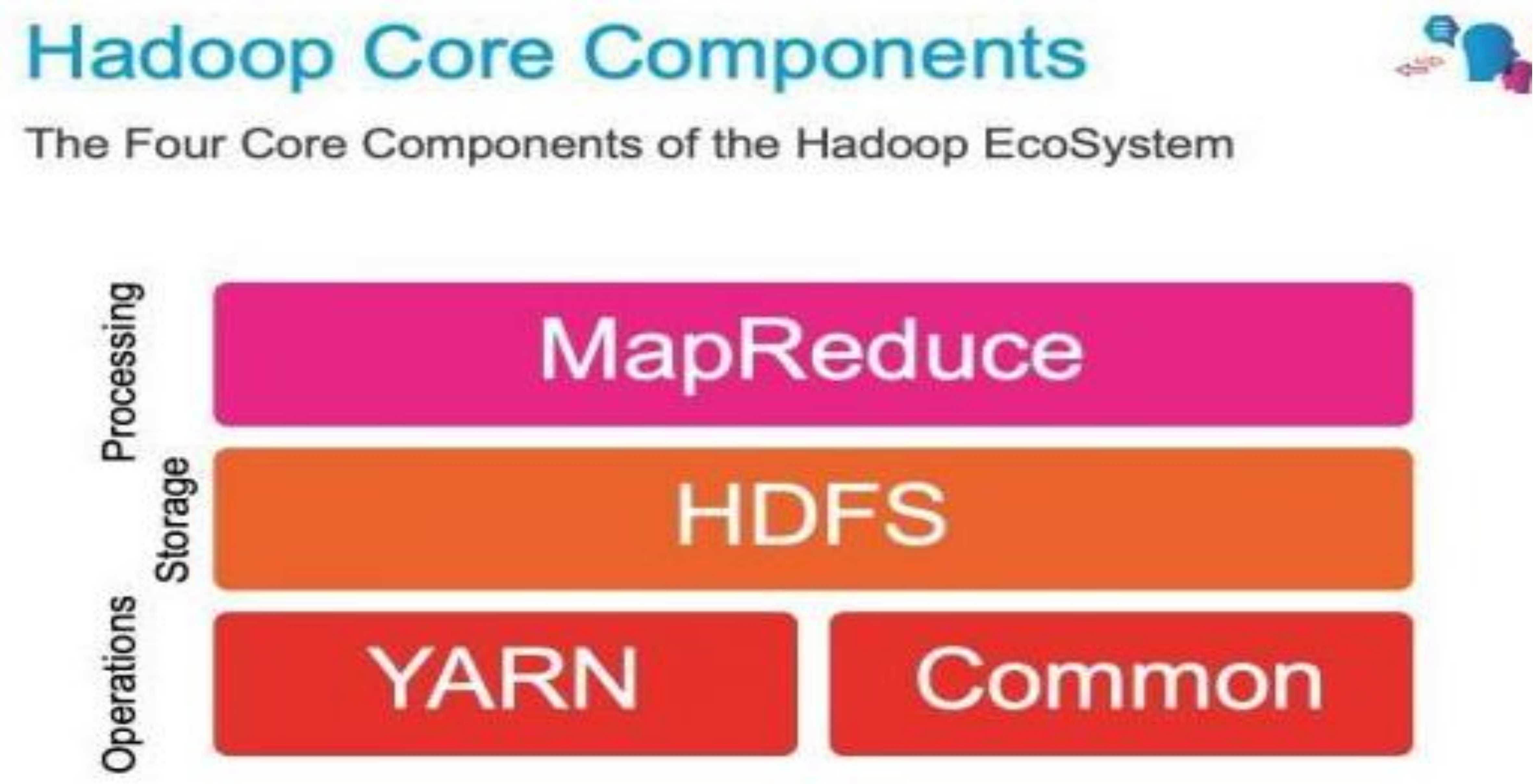
Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

Hadoop Distributed File System (HDFS): A distributed file system that provides High-throughput access to application data.

Hadoop MapReduce: A software-programming model for parallel processing of large data sets.

Hadoop Yet Another Resource Negotiator (YARN): This is a framework for job scheduling and cluster resource management. A resource management framework for scheduling and handling resource requests from distributed applications.

Hadoop



Hadoop Core Components

The Four Core Components of the Hadoop EcoSystem

Processing

MapReduce

Storage

HDFS

Operations

YARN

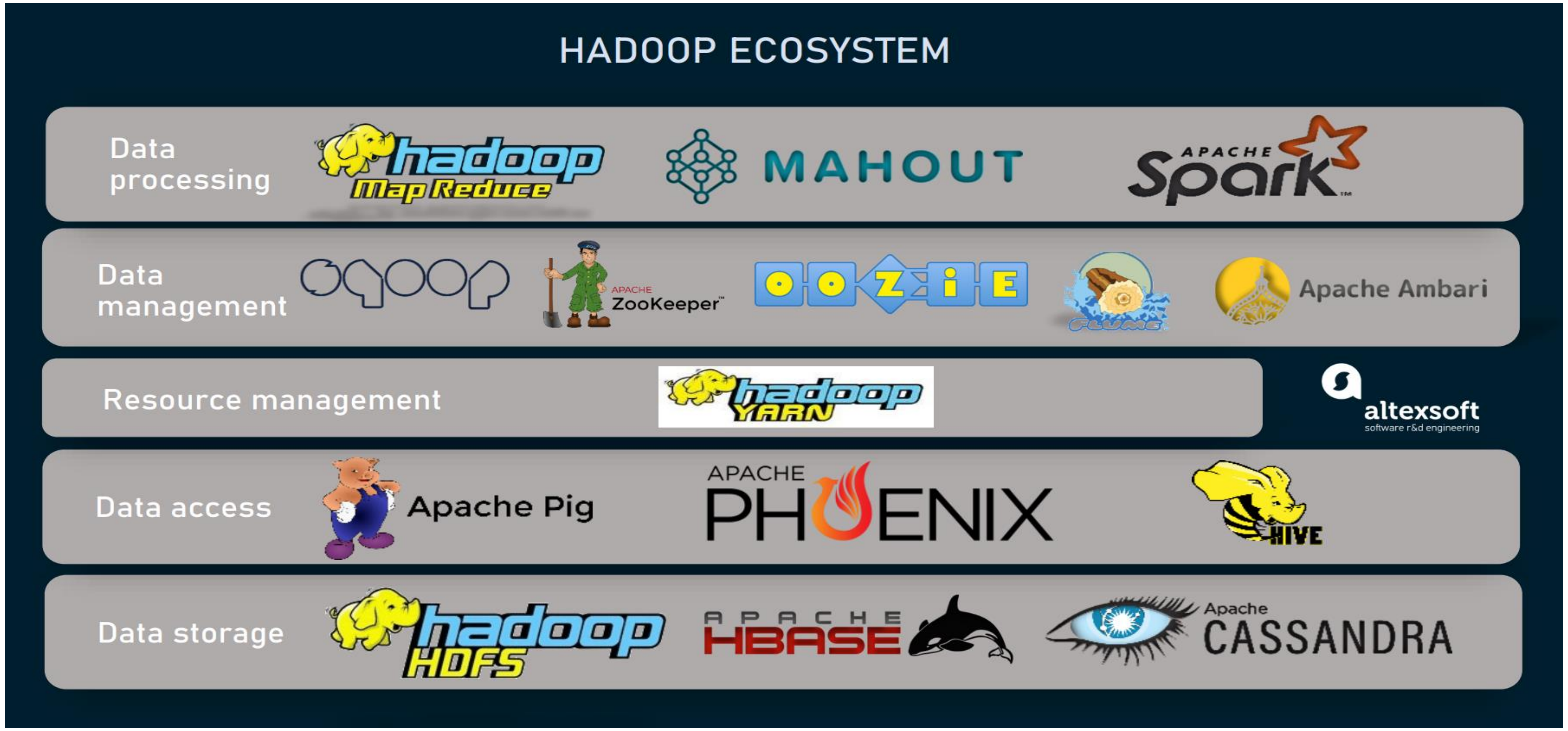
Common

Hadoop Ecosystem

Hadoop Ecosystem

Hadoop ecosystem support projects to enhance the functionality of hadoop core components. The Eco Projects are as follows

1. HIVE
2. PIG
3. SQOOP
4. HBASE
5. FLUME
6. OOZIE
7. MAHOUT



Hadoop Ecosystem

Hadoop Conceptual Layer

It is conceptually divided into Data Storage Layer which stores huge volumes of data Data Processing Layer which processes data in parallel to extract richer and meaningful insights from data.

High-Level Architecture of Hadoop

Hadoop is distributed Master-Slave Architecture. Master Node is known as Name Node and slave nodes are known as DataNodes.

Master HDFS: Its main responsibility is partitioning the data storage across the slave nodes. It also keeps track of locations of data on DataNodes.

Master MapReduce: It decides and schedules computation task on slave nodes.

Hadoop Ecosystem

Why not RDBMS?

RDBMS is not suitable for storing and processing large files, images and videos. RDBMS is not a good choice when it comes to advanced analytics involving machine learning

PARAMETERS	RDBMS	HADOOP
System	Relational Database Management System.	Node based flat structure
Data	Suitable for structures data	Suitable for structured, unstructured data. Supports variety of data formats in real time such as XML, JSON, text based flat file formats, etc.
Processing	OLTP	Analytical, Big Data Processing
Choice	When data needs consistent relationship	Big data processing, which does not require any consistent relationship between data
Processor	Needs expensive hardware or high-end processor to store huge volumes of data.	In a hadoop cluster, a node requires only a processor, a network card, and few hard drives.
Cost	Cost around $10,000 to $14,000 per terabytes of storage	Cost around $4,000 per terabytes of storage.

Hadoop Ecosystem

Why not RDBMS?

RDBMS is not suitable for storing and processing large files, images and videos. RDBMS is not a good choice when it comes to advanced analytics involving machine learning

PARAMETERS	RDBMS	HADOOP
System	Relational Database Management System.	Node based flat structure
Data	Suitable for structures data	Suitable for structured, unstructured data. Supports variety of data formats in real time such as XML, JSON, text based flat file formats, etc.
Processing	OLTP	Analytical, Big Data Processing
Choice	When data needs consistent relationship	Big data processing, which does not require any consistent relationship between data
Processor	Needs expensive hardware or high-end processor to store huge volumes of data.	In a hadoop cluster, a node requires only a processor, a network card, and few hard drives.
Cost	Cost around $10,000 to $14,000 per terabytes of storage	Cost around $4,000 per terabytes of storage.

THANK YOU