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 Artificial Intelligence and Data Science

Course Name – Big Data Analytics
III Year / V Semester

Unit 3 – DATA ANALYTICAL FRAMEWORKS

Topic - Introducing Hadoop

Hadoop

- Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models.

- A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers.

- Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

- In short Hadoop is an open source software framework for sorting and processing big data in distributed way on large clusters of commodity hardware

Hadoop

Why Hadoop?

- Its capability to handle massive amounts of data, different categories of data fairly quickly.

- Low cost: It is an open source framework and uses commodity hardware to store enormous quantities of data.

- Computing Power: Hadoop is based on distributed computing model, therefore more number of computing nodes, the more processing power at hand.

- Scalability: When adding more nodes as the system grows and requires less administration.

- Storage Flexibility: Hadoop provides convenience of storing as much as data as one needs and also added flexibility of deciding later as to how to use the stored data.

- Inherent Data Protection: Hadoop protects the data and executing applications against hardware failure. If a node fails it automatically redirects the jobs that had been assigned to this node to the other functional and available nodes.

Hadoop

Hadoop is open source software framework to store and process massive amounts of data in a distributed fashion on large clusters of commodity hardware.

Basically, Hadoop accomplishes two tasks

- Massive data storage
- Faster data processing

Key Aspects of Hadoop

- Open source: It is free to download,
- Frameworks: Means everything that you will need to develop and execute and application is provided programs, tool etc.
- Distributed: Divides and stores data across multiple computers. Computation/processing is done in parallel across multiple connected nodes.
- Massive Storage: Stores colossal, amount of data across nodes of low cost commodity hardware.
- Faster Processing: Large amounts of data is processed in parallel, yielding quick response.

Hadoop

Hadoop Core Components

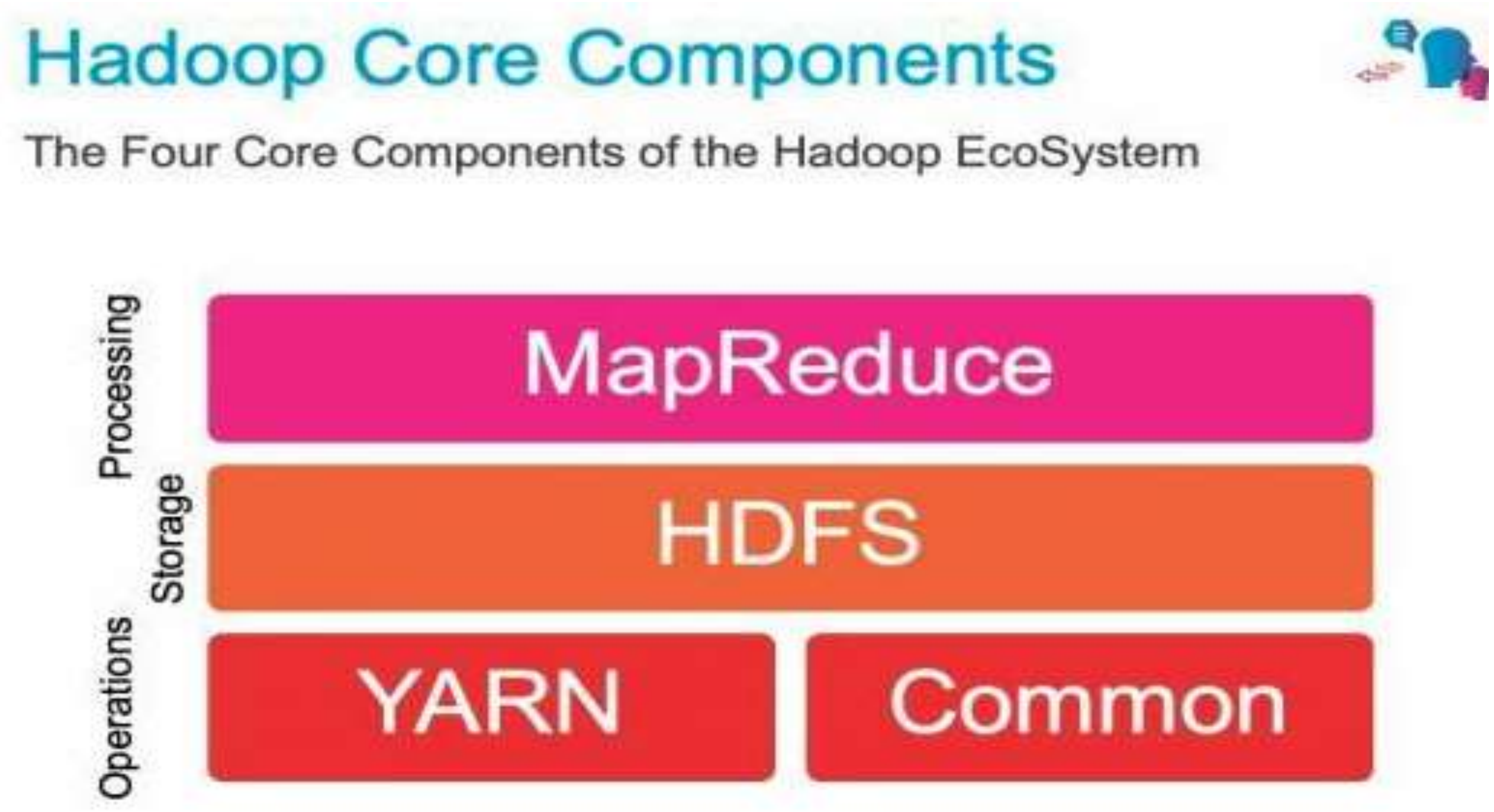
- Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and containsthe necessary Java files and scripts required to start Hadoop.

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

- Hadoop MapReduce: This is YARN-based system for parallel processing of large data sets.

- Hadoop Yet Another Resource Negotiator (YARN): This is a framework for job scheduling and cluster resource management.

Hadoop



Hadoop Core Components

The Four Core Components of the Hadoop EcoSystem

Processing

MapReduce

Storage

HDFS

Operations

YARN

Common

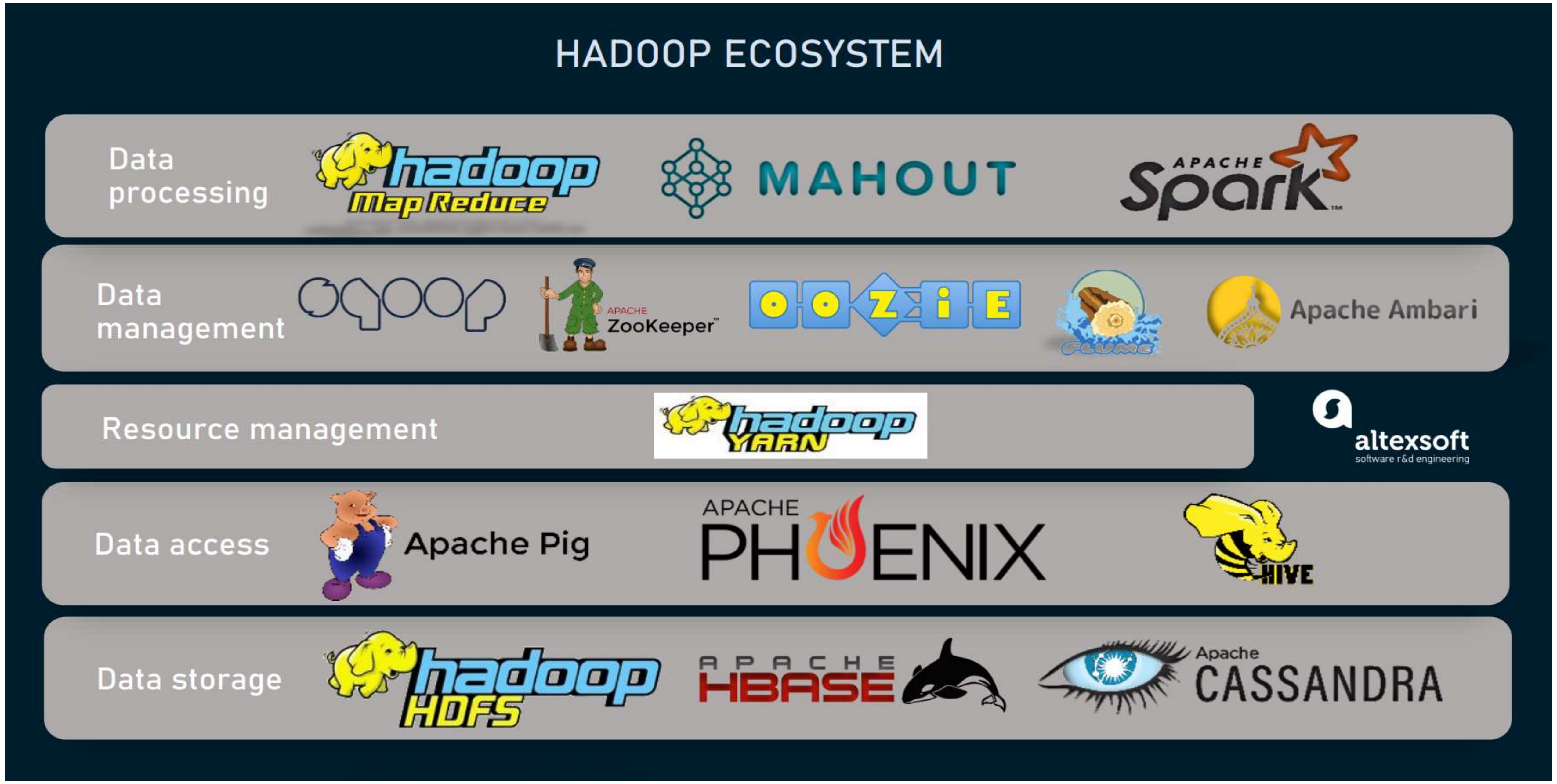
Hadoop

Hadoop Ecosystem

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

- HIVE
- PIG
- SQOOP
- HBASE
- FLUME
- OOZIE
- AMBARI
- MAHOUT
- SPARK
- ZOOKEEPER

Hadoop


HADOOP ECOSYSTEM
Data processing
hadoop Map Reduce
MAHOUT
APACHE Spark
Data management
sqoop
APACHE ZooKeeper
OOZIE
FLUME
Apache Ambari
Resource management
hadoop YARN
altexsoft
software r&d engineering
Data access
Apache Pig
APACHE PHOENIX
HIVE
Data storage
hadoop HDFS
APACHE HBASE
Apache CASSANDRA

THANK YOU