# SESSION 5

## Introduction to Hadoop and its Architecture

# Objectives

- Explain the basics of Big Data and Hadoop
- Describe the architecture of Hadoop

# Basics of Hadoop

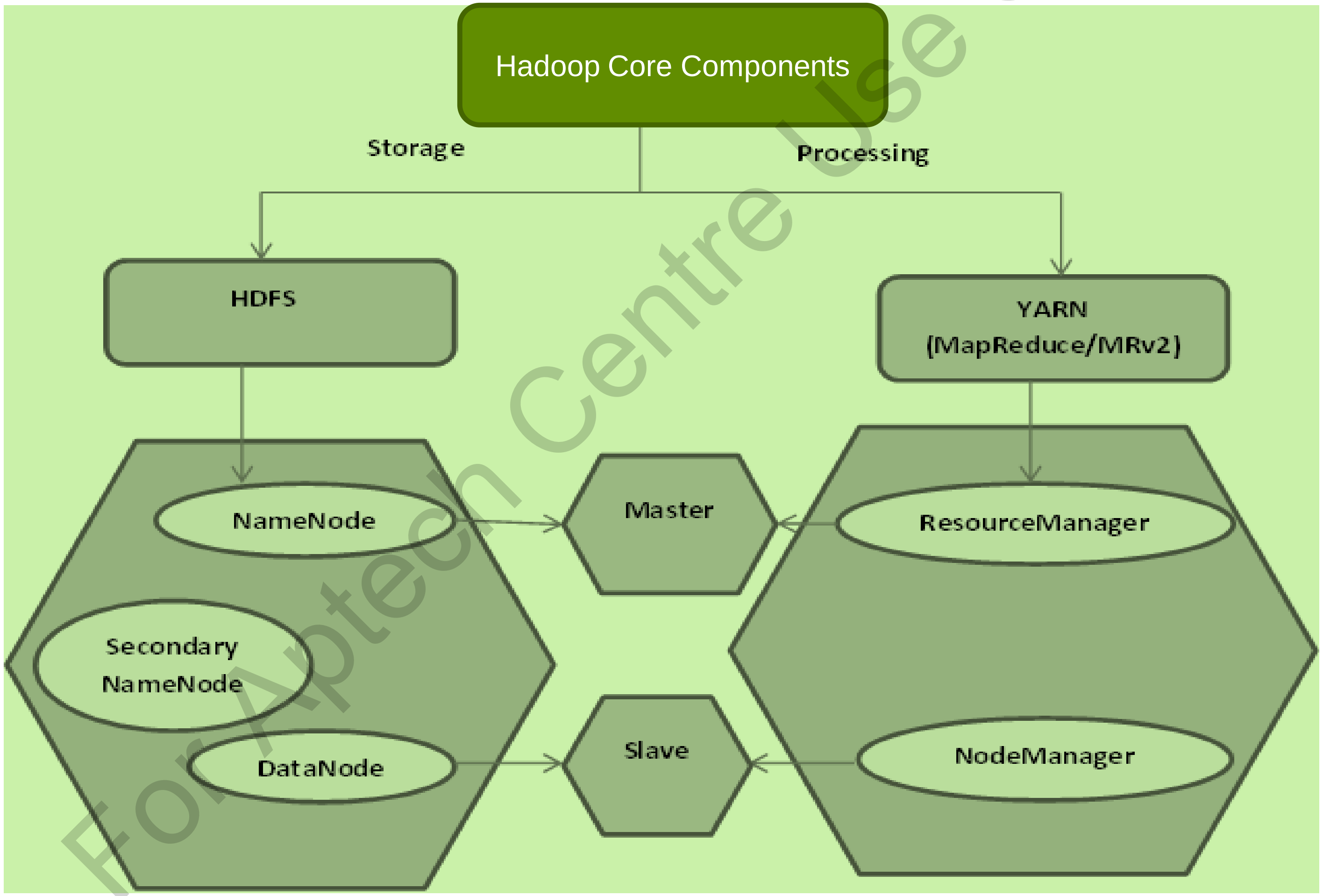
- Referred to as Apache Hadoop
- Is an open-source Java-based software framework
- Developed by Doug Cutting and Mike J
- Designed using MapReduce programming model with an objective to save and process Big Data

Two components of Hadoop are:

**Hadoop Distributed File System (HDFS)**

**Yet Another Resource Negotiator (YARN)**

# Hadoop Components



# Hadoop Architecture

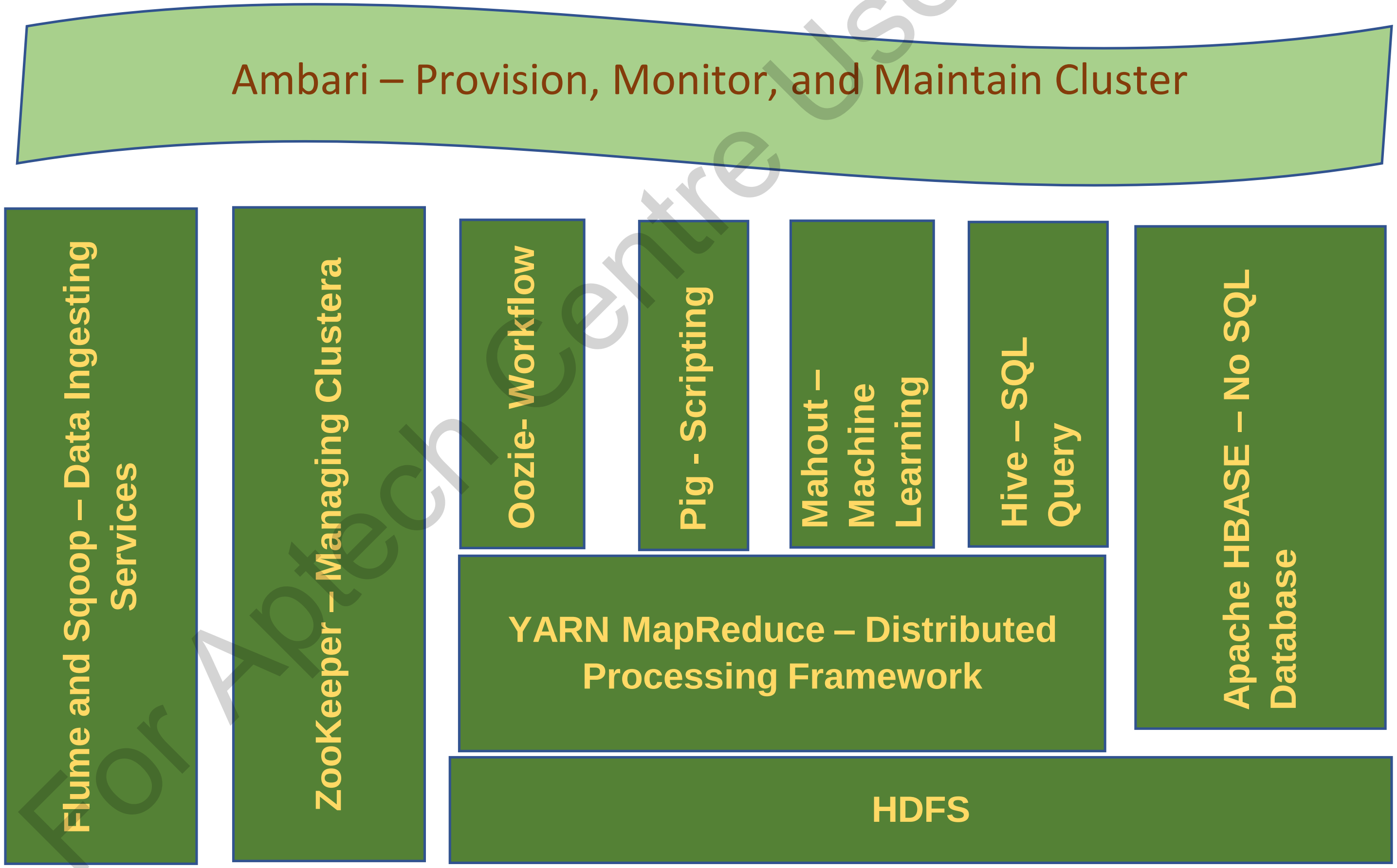
- Hadoop has many smaller components that are responsible for handling different processes.
- These components work together under the Hadoop ecosystem.

Components that form the Hadoop ecosystem are:

- Hadoop Distributed File System (HDFS)
- Yet Another Resource Negotiator (YARN)
- MapReduce
- Spark
- Pig and Hive
- HBase
- Mahout and Spark Mllib
- Apache Drill
- ZooKeeper
- Oozie
- Flume and Sqoop
- Solr and Lucene
- Ambari

# Hadoop Ecosystem

- Figure shows the ecosystem of Hadoop:


Ambari – Provision, Monitor, and Maintain Cluster
Flume and Sqoop – Data Ingesting Services
ZooKeeper – Managing Clustera
Oozie- Workflow
Pig - Scripting
Mahout – Machine Learning
Hive – SQL Query
YARN MapReduce – Distributed Processing Framework
Apache HBASE – No SQL Database
HDFS

# Hadoop Features

- Features of Hadoop are:

### Flexibility

- It analyzes all types of data - structured, semi-structured, or unstructured

### Reliability

- HDFS of Hadoop is highly fault tolerant

### Cost Effective

- It uses commodity servers, such as PCs and laptops and is economical

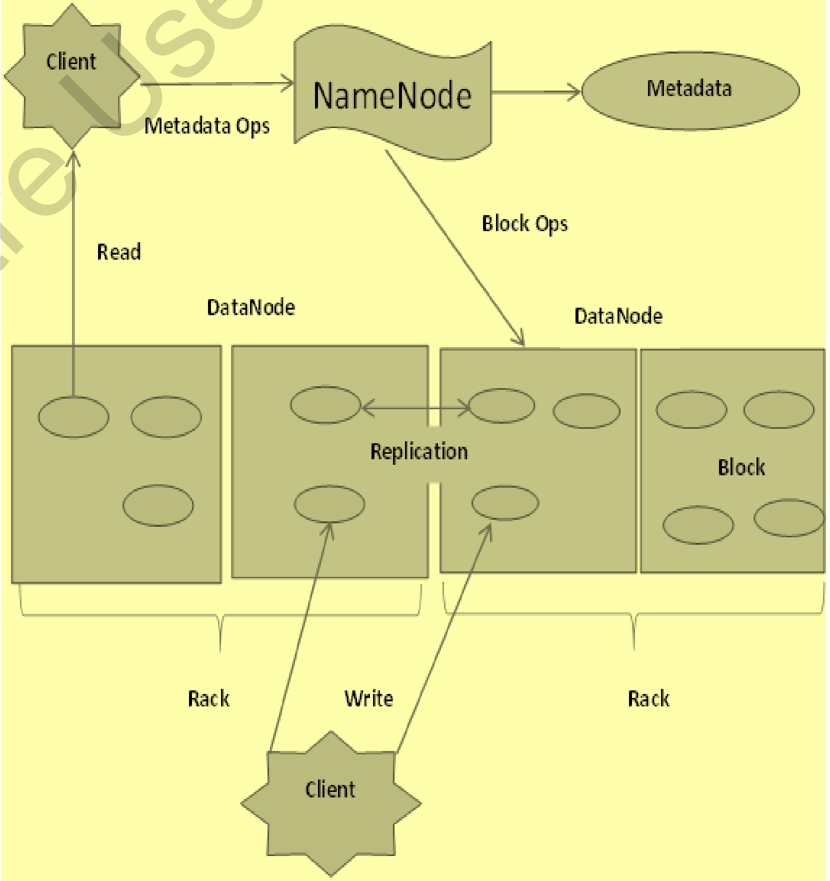
### Scalability

- It is highly scalable and enables to add more nodes at any time

# HDFS Architecture

- It is a main component of Hadoop and is capable of storing large amount of data and providing fast access to data.

- It is highly fault tolerant as its data gets copied to multiple machines.



# HDFS Advantages

## Distributed Storage

- A single huge file is distributed over different nodes on the Hadoop cluster

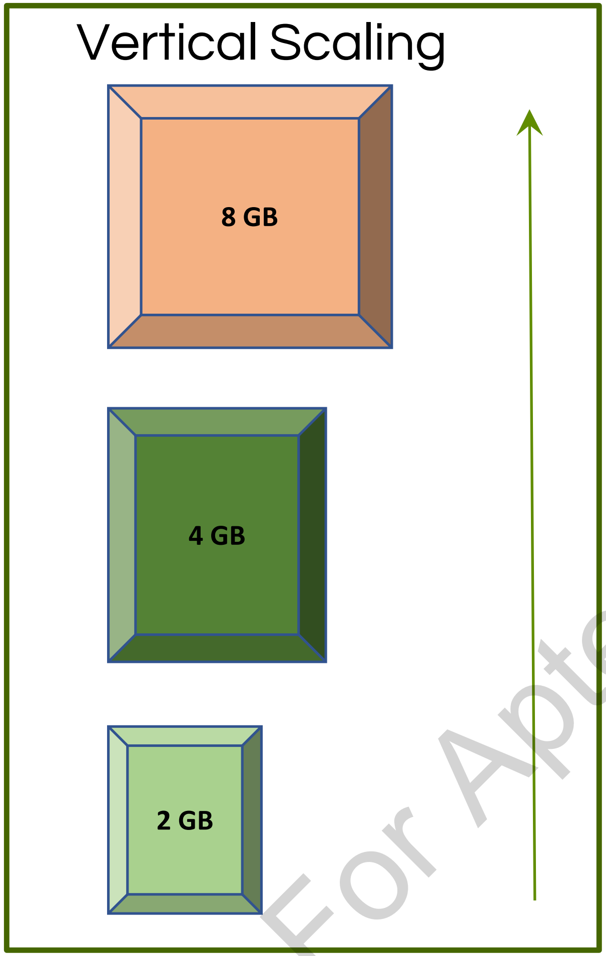
## Distributed and Parallel Computation

- Distributed data over multiple machines working parallel

## Horizontal Scalability

- **Vertical or Scaling Up**: The powerful hardware increases RAM or CPU
- **Horizontal or Scaling Down**: Adding more nodes or machines to existing cluster

# Vertical and Horizontal Scaling

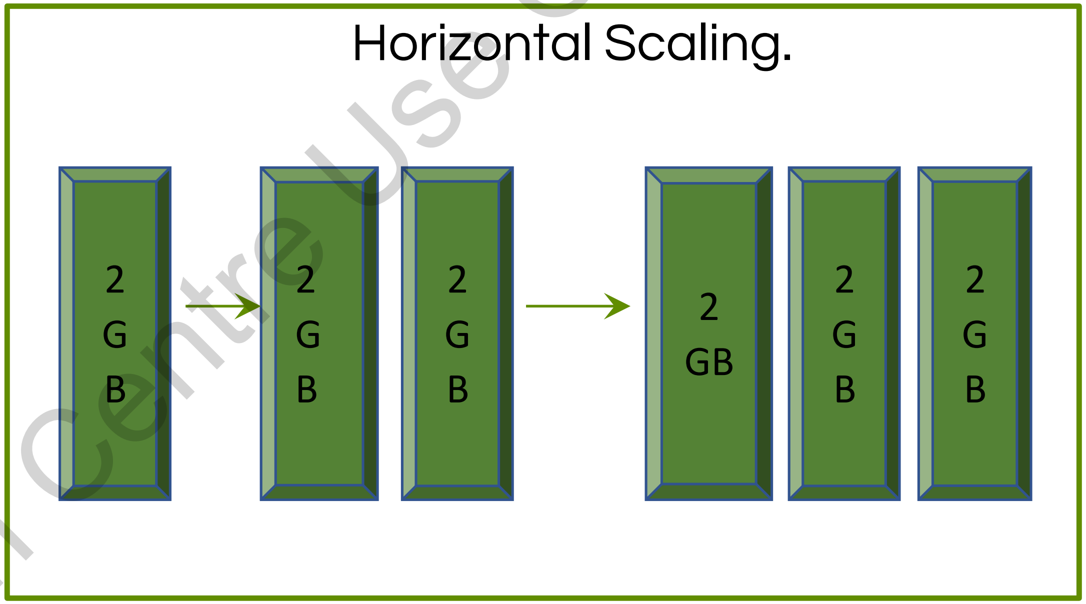


Vertical Scaling

8 GB

4 GB

2 GB



Horizontal Scaling.

2 GB

2 GB

2 GB

2 GB

2 GB

2 GB

# MapReduce

- It is a Java-based software framework that uses distributed and parallel computing to process large data sets
- It has two important tasks:

### Map

- Individual elements of each data set are broken into value pairs and converted into another data set

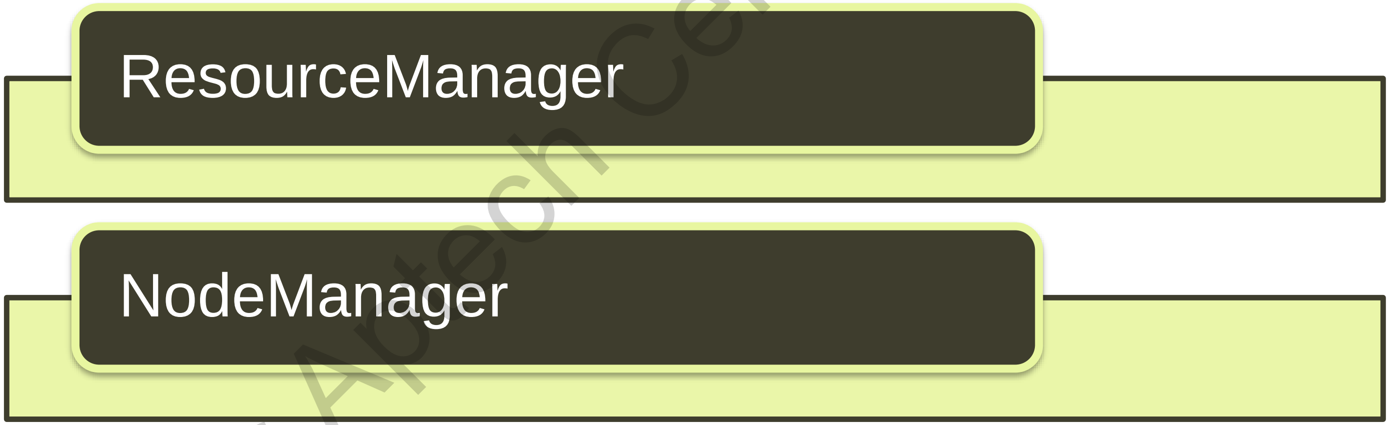
### Reduce

- Output from the map is taken as input and the value pairs are combined and converted into smaller value pairs

- MapReduce framework and HDFS run on the same set of nodes
- Each node contains one master JobTracker and slave TaskTracker.

# YARN

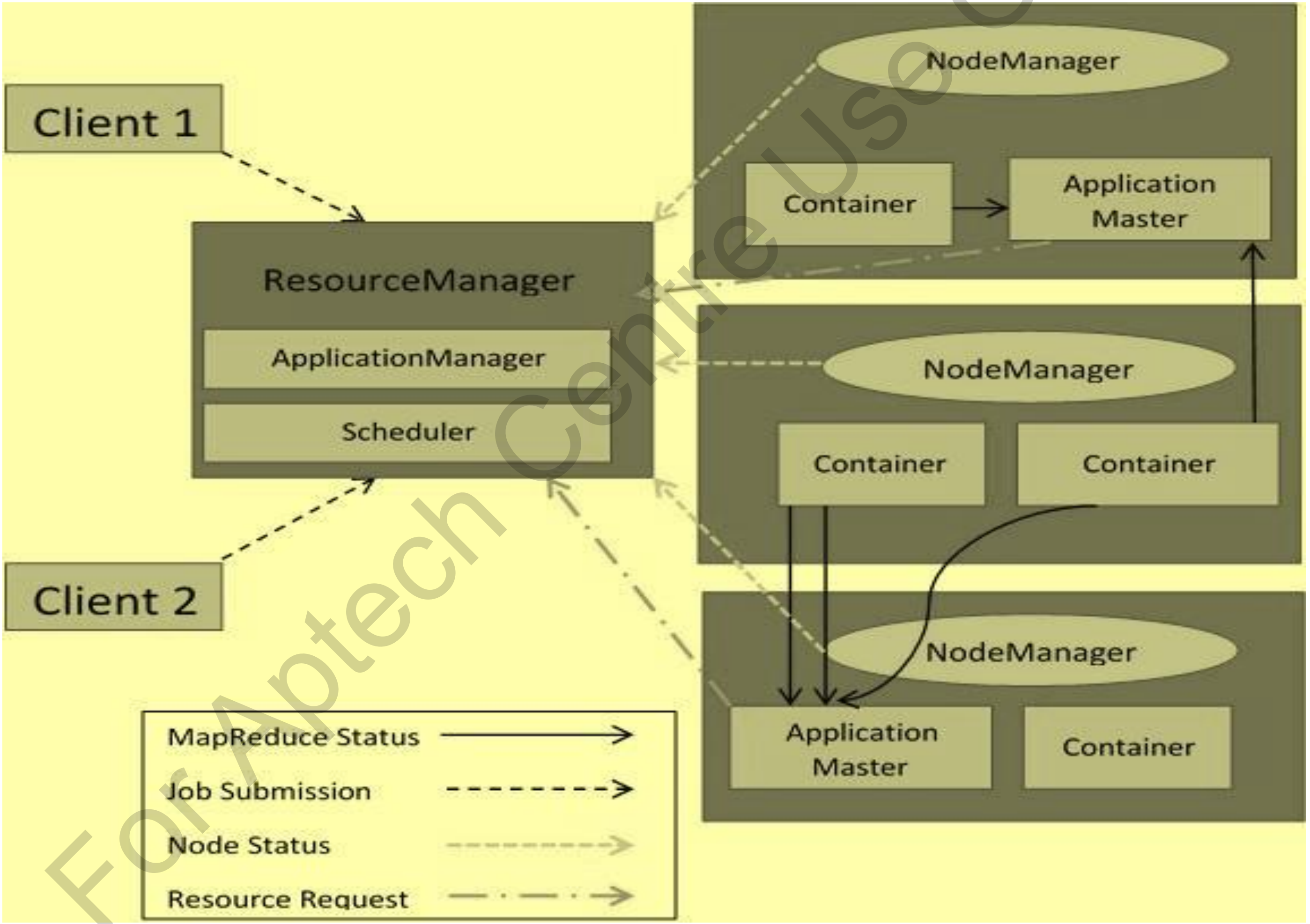
- A cluster management technology which allows multiple data processing engines to handle the stored data and provide to users.
- It performs resource allocation and task scheduling.
- Two major components are:



ResourceManager

NodeManager

# Components of YARN



# ZooKeeper (1-3)

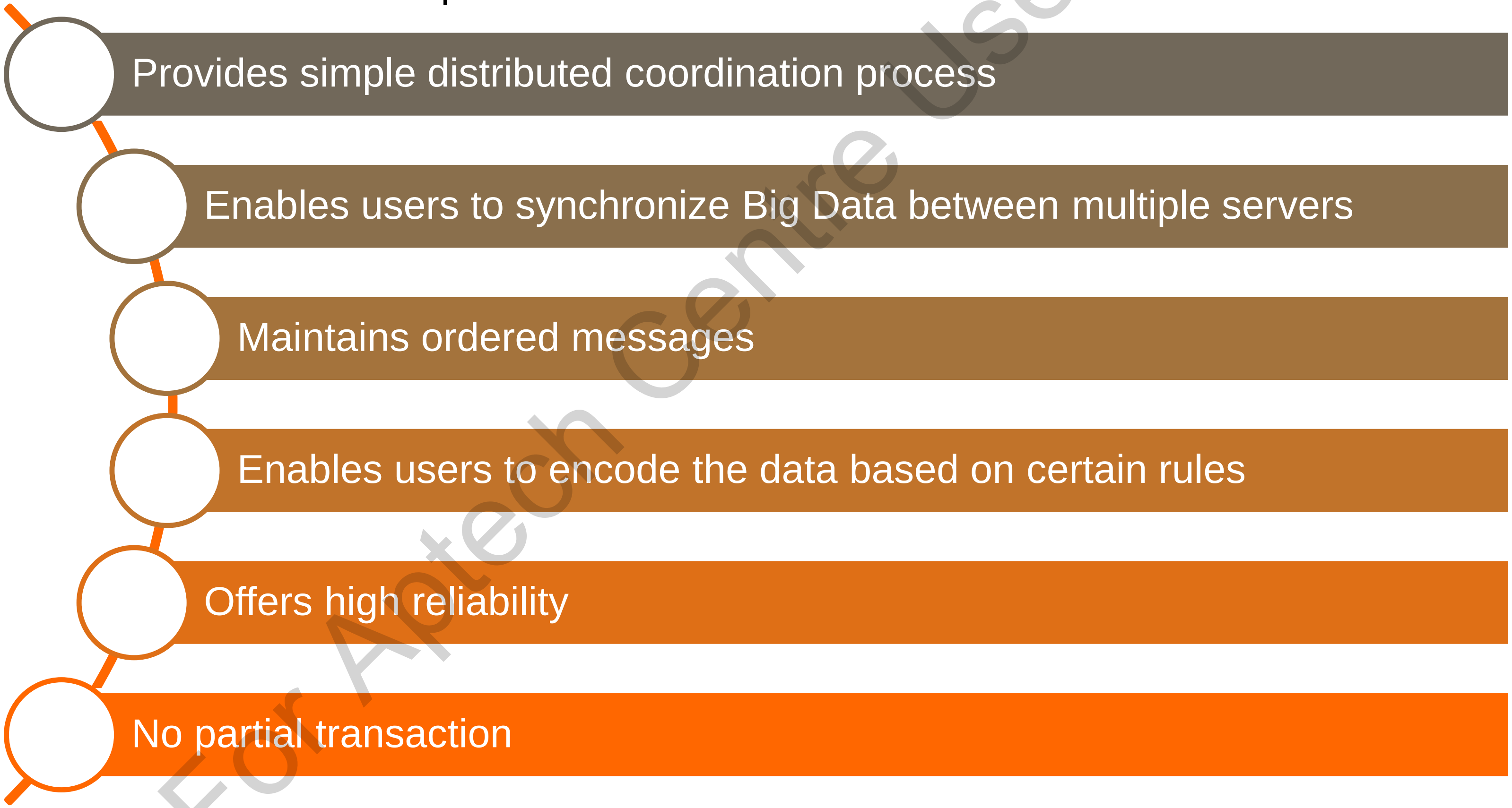
- It is a framework designed to provide distributed coordination service in a Hadoop cluster
- It acts as a coordinator in a Hadoop task
- It consists of combination of numerous services in a Hadoop ecosystem
- It coordinates with different services in a distributed environment
- The nodes use the ZooKeeper service to coordinate and synchronize shared data in a Hadoop cluster
- Hadoop clusters are very large and require centralized management
- Multiple ZooKeeper servers can be implemented where master server synchronizes the top-level servers and communicates with the client machine

# ZooKeeper (2-3)

- Services provided by ZooKeeper are:
  - Naming Service
  - Configuration Management
  - Cluster Management
  - Leader Election
  - Locking and Synchronization Service
  - Highly reliable data registry

# ZooKeeper (3-3)

- Benefits of ZooKeeper are:



- Provides simple distributed coordination process
- Enables users to synchronize Big Data between multiple servers
- Maintains ordered messages
- Enables users to encode the data based on certain rules
- Offers high reliability
- No partial transaction

# Summary (1-2)

- Hadoop is an open-source Java-based software framework.
- Hadoop is designed using MapReduce programming model with an objective to store and process Big Data in a distributed manner.
- Hadoop consists of many functional modules; however, while setting up Hadoop, the two main components required are Hadoop Distributed File System (HDFS) and Yet Another Resource Negotiator (YARN).
- HDFS is highly scalable and stores Big Data across thousands of commodity servers. YARN is responsible for all the resource management and scheduling of jobs.

# Summary (2-2)

- Hadoop can handle very large volumes of organizational data; be it structured or unstructured.
- MapReduce is a Java-based software framework that uses distributed and parallel computing to process large data sets inside the Hadoop environment in a reliable and fault-tolerant manner.
- Apache ZooKeeper is a software framework designed to provide distributed co-ordination service in a Hadoop cluster.