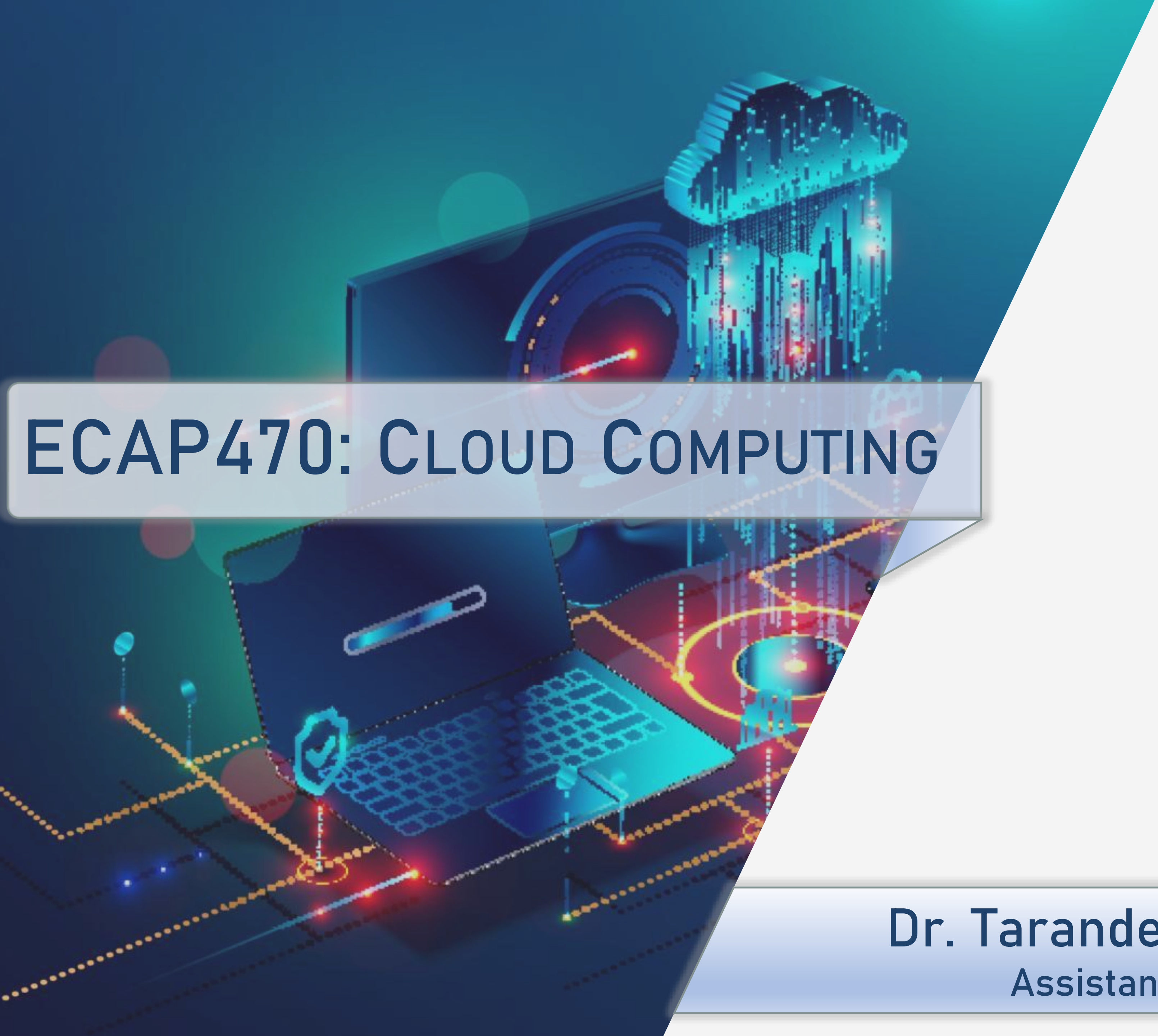


ECAP470: Cloud Computing

Dr. Tarandeep Kaur
Assistant Professor

Learning Outcomes

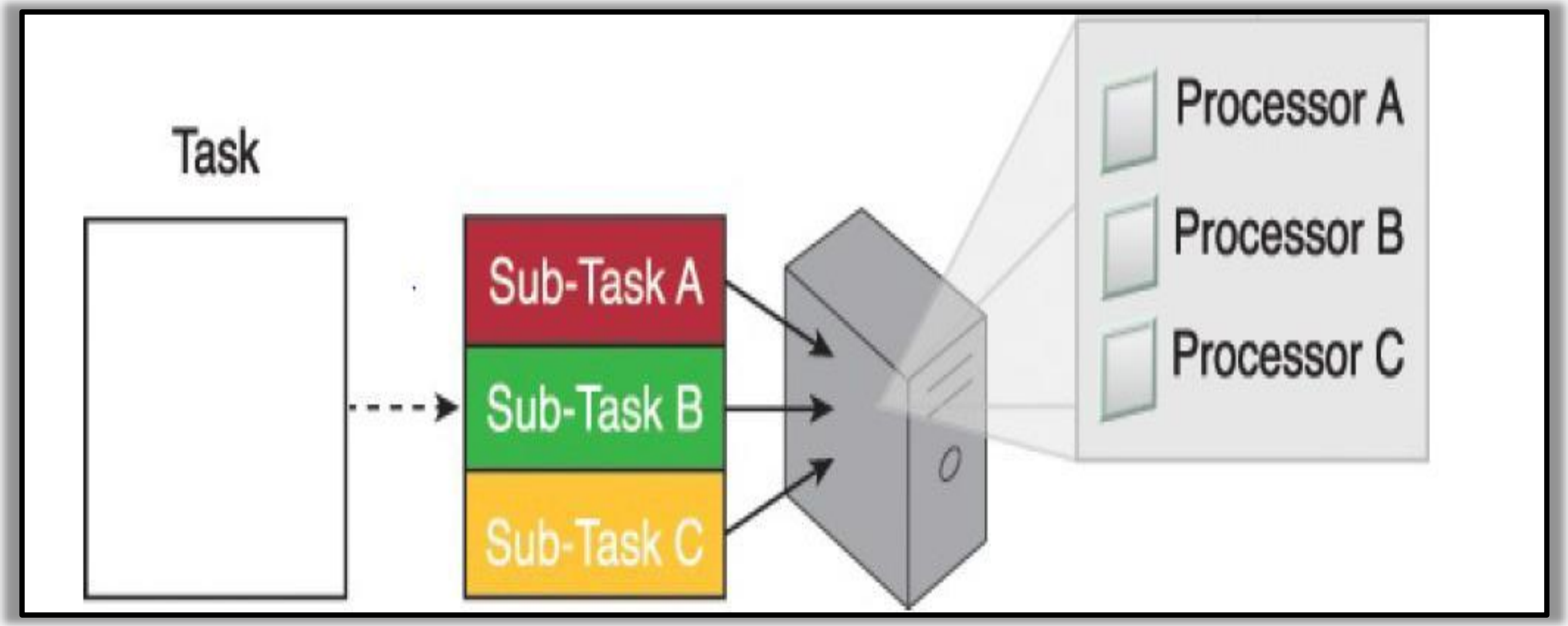
After this lecture, you will be able to,

- ✓ Learn about data processing and Hadoop Processing.
- ✓ Understand the Hadoop Framework and its modules.

Parallel Data Processing

- Parallel data processing involves the simultaneous execution of multiple sub-tasks that collectively comprise a larger task.
- Goal: To reduce the execution time by dividing a single larger task into multiple smaller tasks that run concurrently.

Parallel Data Processing

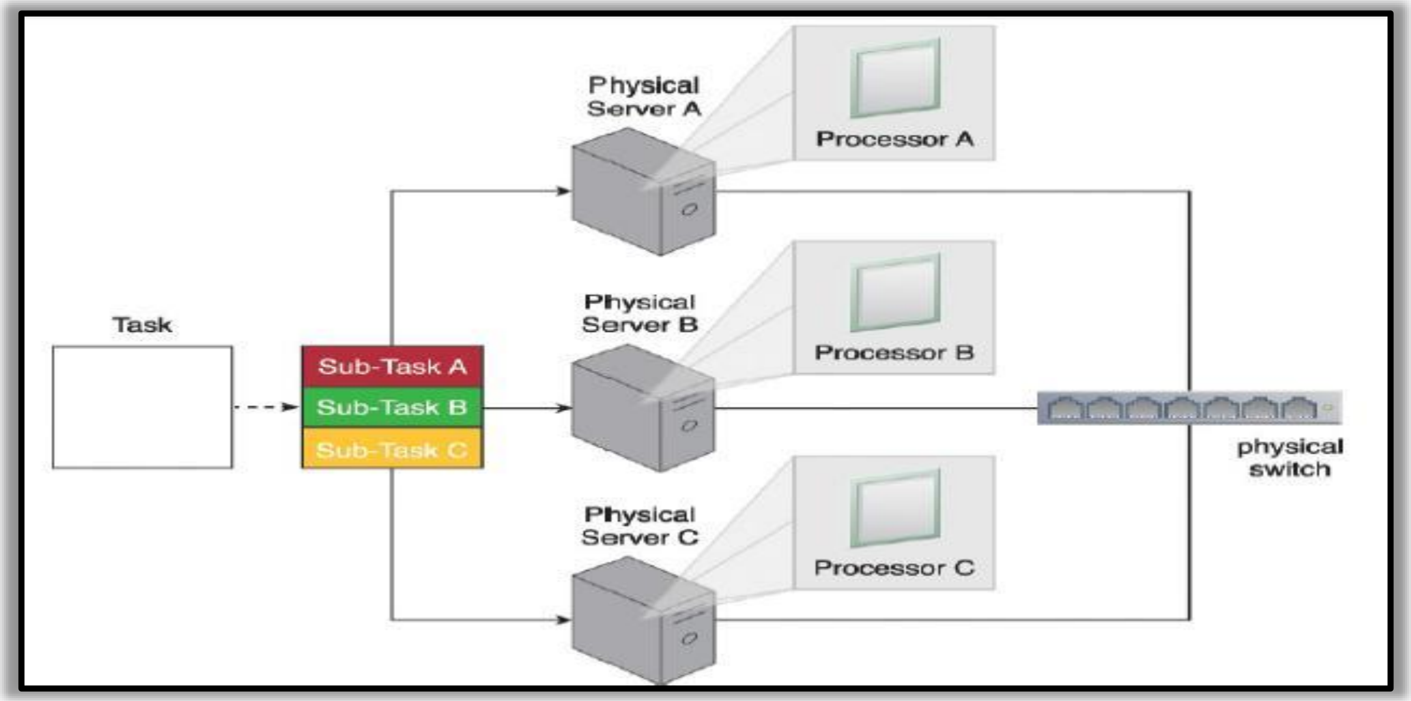


Distributed Data Processing

- Closely related to parallel data processing.
- Follows the principle of “divide-and-conquer”.
- Always achieved through physically separate machines that are networked together as a cluster.

Distributed Data Processing

A task is divided into three sub-tasks that are then executed on three different machines sharing one physical switch.



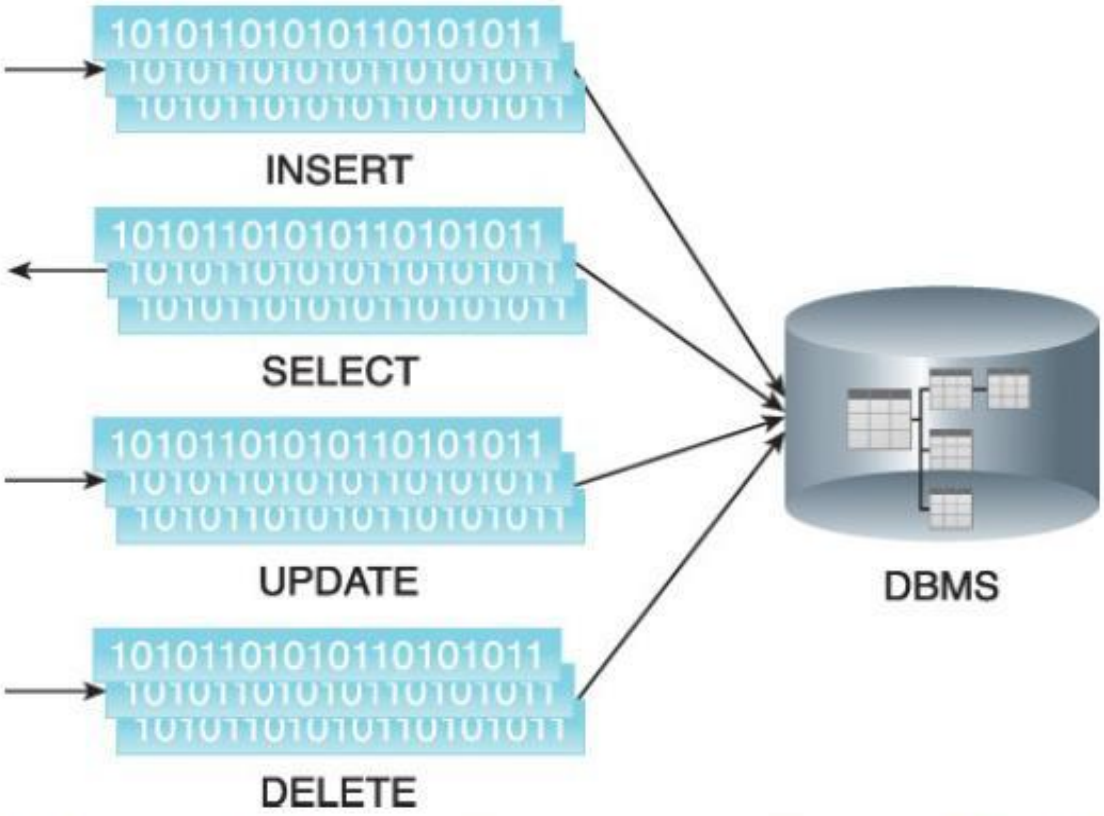
Processing Workloads in Big Data

- Amount and nature of data that is processed within a certain amount of time.
- Workloads are usually divided into two types:
  - Batch
  - Transactional

Batch Processing Workload

- Also known as offline processing.
- In batch mode, data is processed offline in batches and the response time could vary from minutes to hours. As well, data must be persisted to the disk before it can be processed.
- Involves processing data in batches and usually imposes delays, which in turn results in high-latency responses.

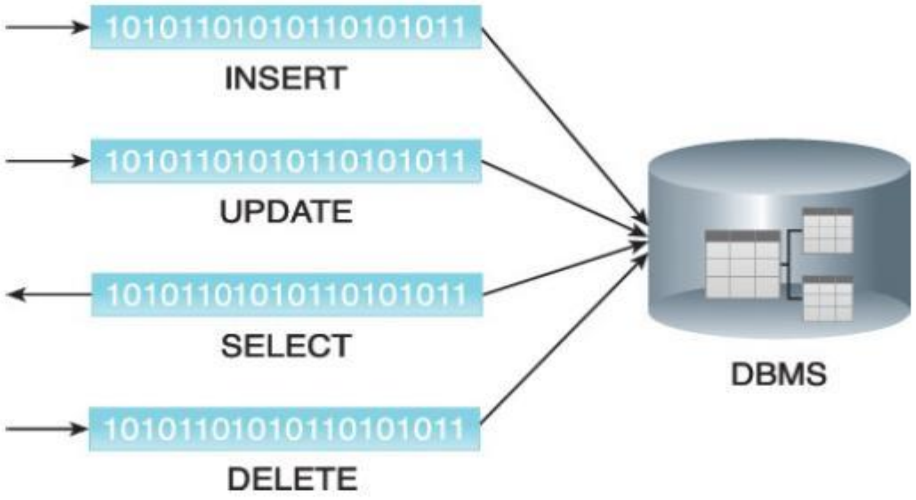
Batch Processing Workload



A batch workload can include grouped read/writes to INSERT, SELECT, UPDATE and DELETE.

Transactional Processing Workload

- Also known as online processing.
- Follows an approach whereby data is processed interactively without delay, resulting in low-latency responses.

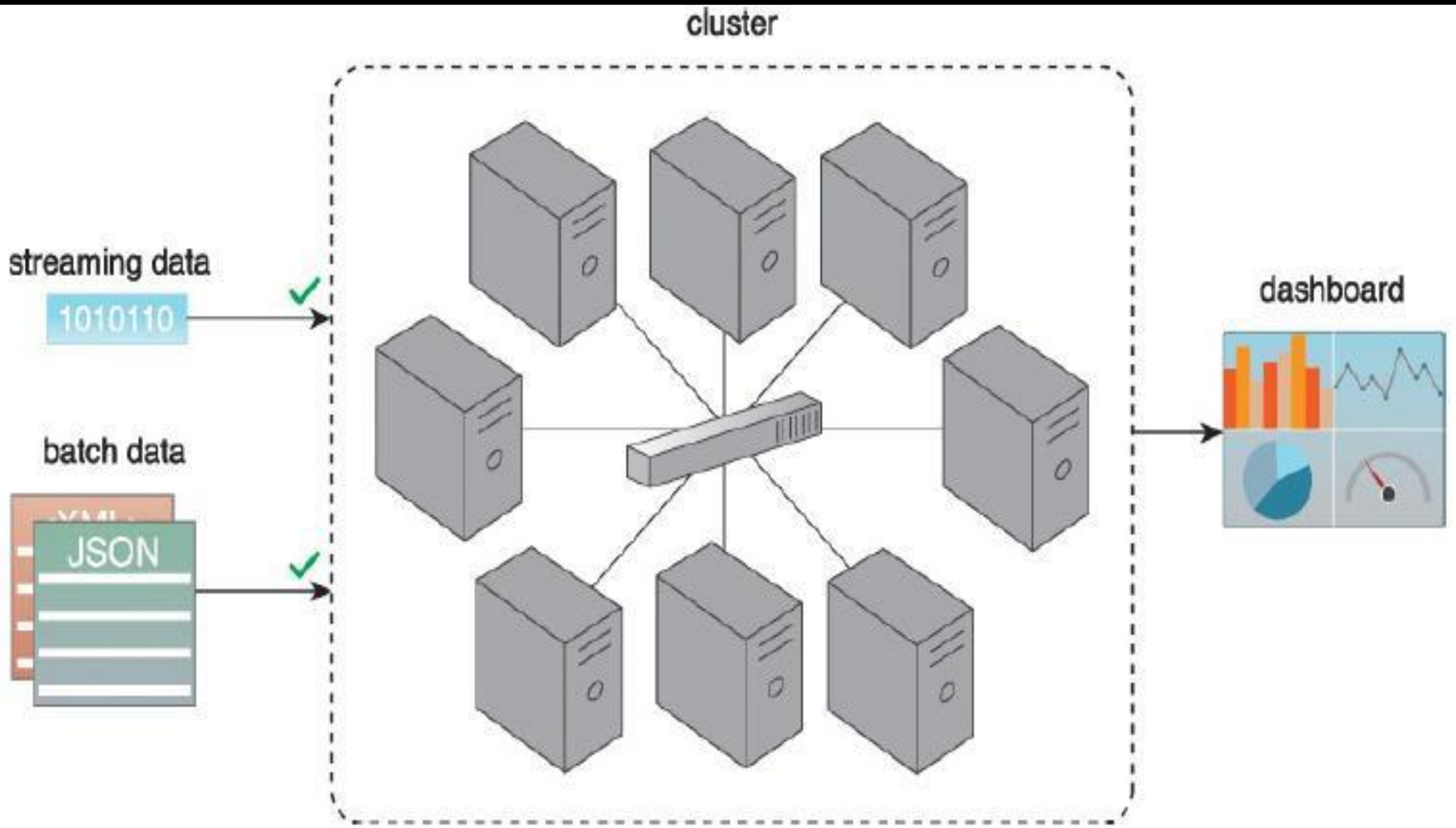


5 Transactional workloads have few joins and lower latency responses than batch workloads.

Cluster Processing

- Provide necessary support to create horizontally scalable storage solutions.
- Also provides the mechanism to enable distributed data processing with linear scalability.

Cluster Processing



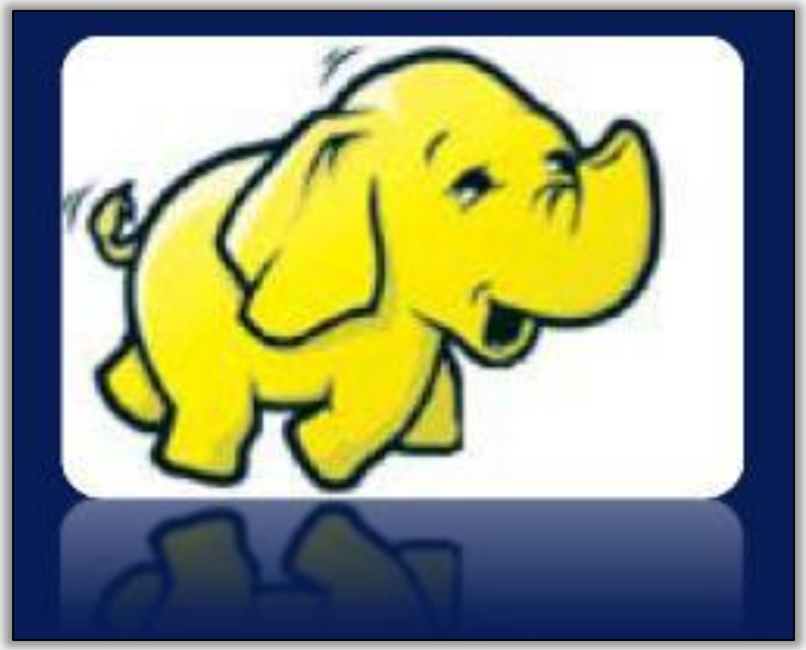
A cluster can be utilized to support batch processing of bulk data and realtime processing of streaming data.

Hadoop Framework

- Hadoop is an open-source framework for large-scale data storage and data processing that is compatible with commodity hardware.
- A software framework for storage and large scale processing of data-sets on clusters of commodity hardware.

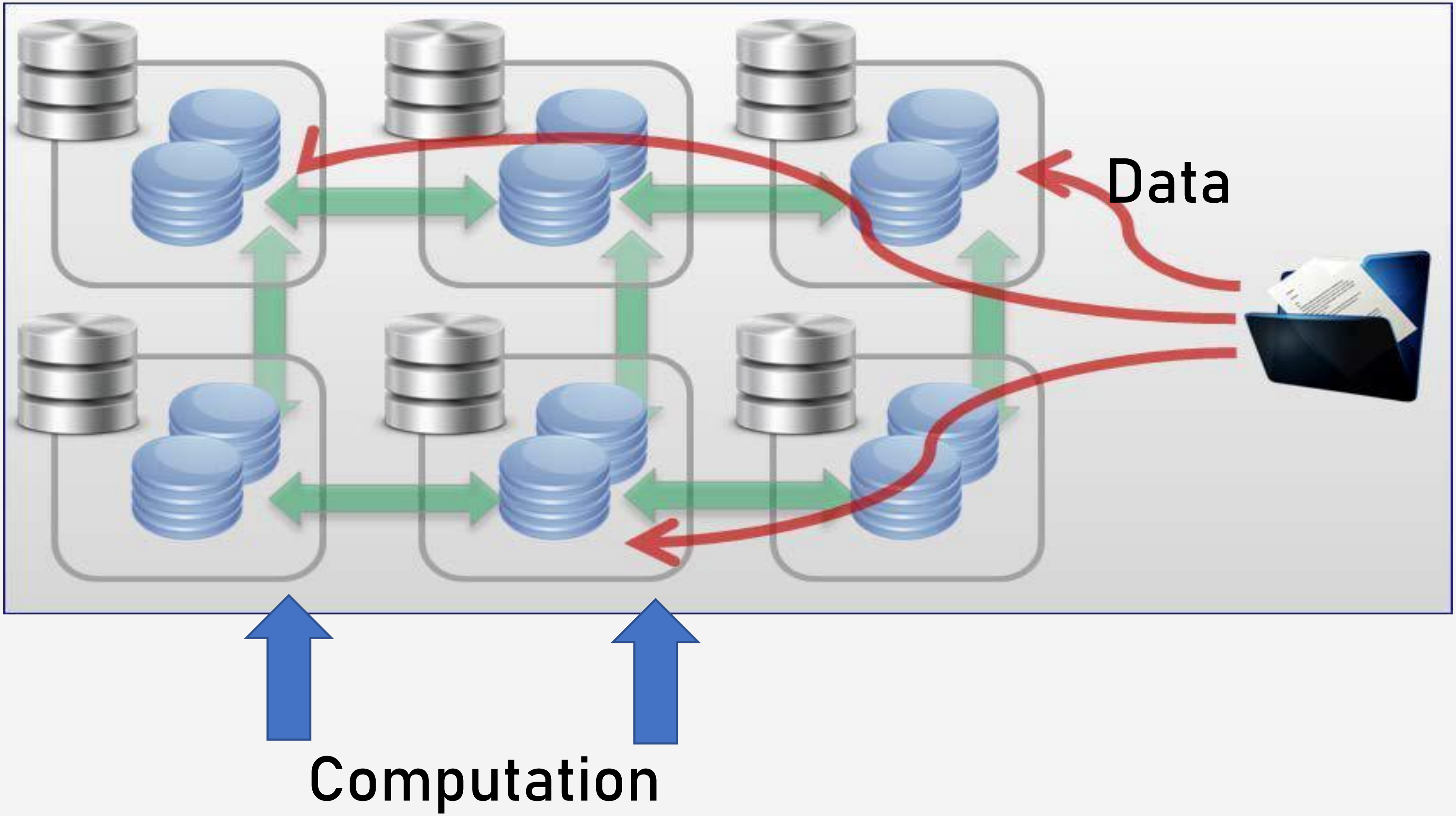
Hadoop Framework

- Created by Doug Cutting and Mike Cafarella in 2005.
- Named the project after son's toy elephant.



Hadoop Framework

Moving Computation to Data- Computation is moved to data.



Hadoop Framework

Scalability Issues Resolved at Hadoop's Core

Reliability Issues

Hadoop Framework

Hadoop Offers:

- New Approach to Data
- New Kinds of Analysis

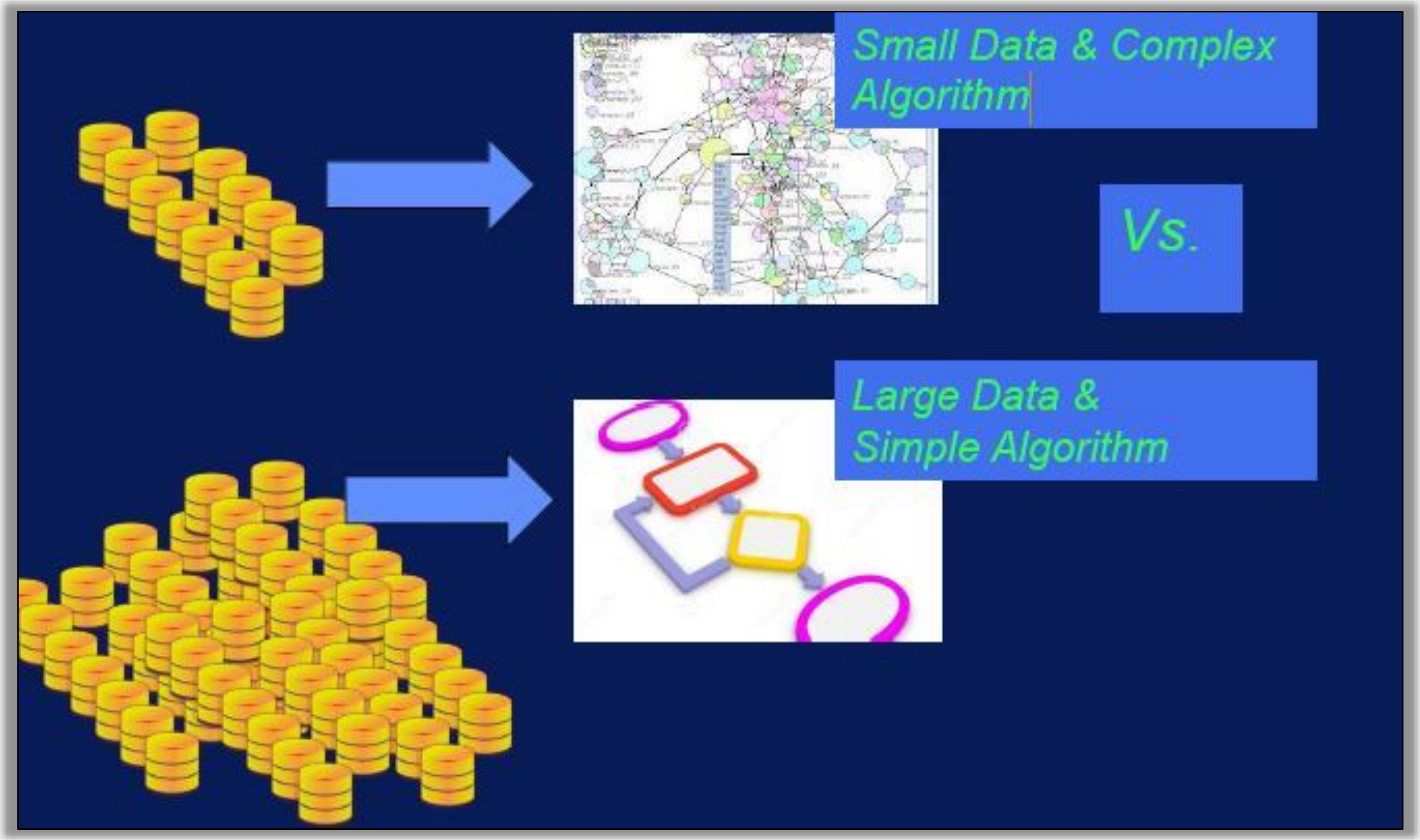
Unstructured Data

Schema-on Read Style

Hadoop Framework

- Can be used as an ETL engine or as an analytics engine for processing large amounts of structured, semi structured and unstructured data.
- From an analysis perspective, Hadoop implements the MapReduce processing framework.

Hadoop Framework

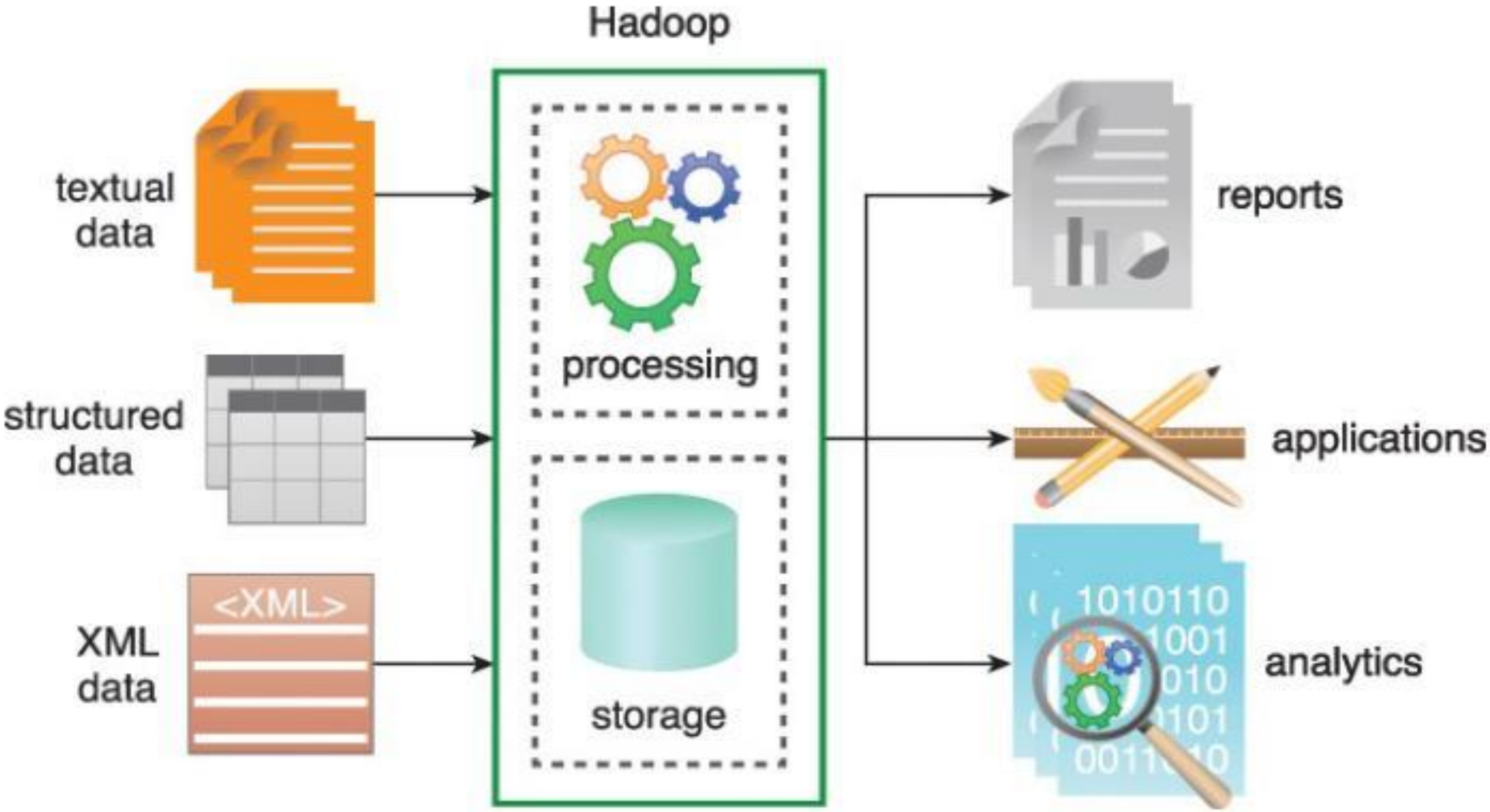


Small Data & Complex Algorithm

Vs.

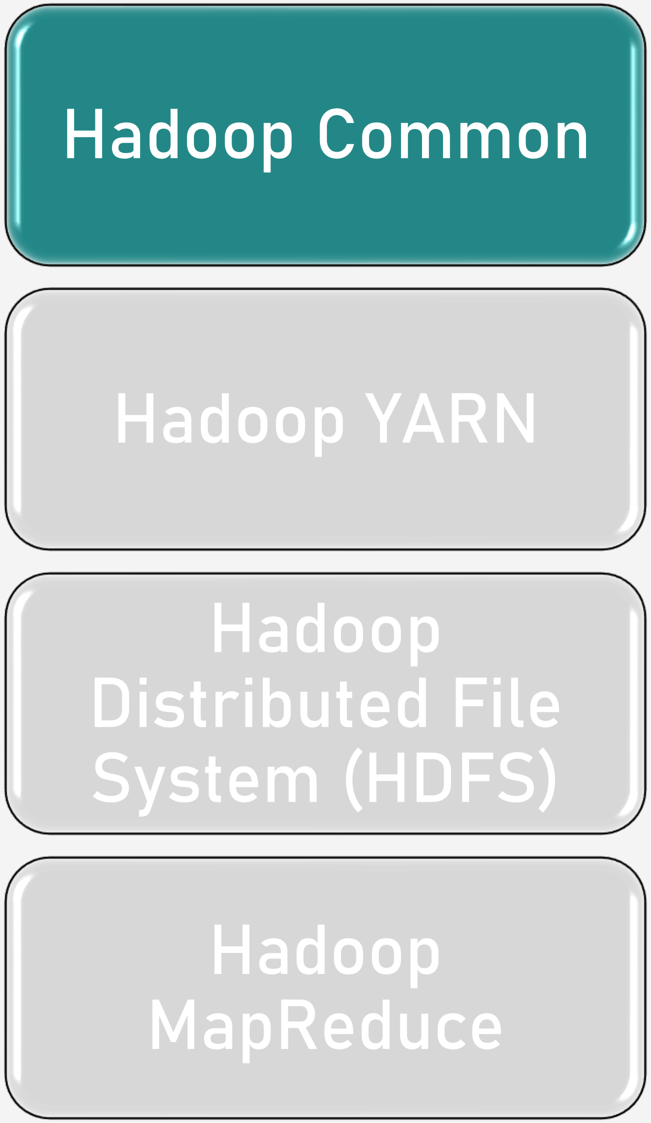
Large Data & Simple Algorithm

Hadoop Framework



Hadoop is a versatile framework that provides both processing and storage capabilities.

Hadoop Modules



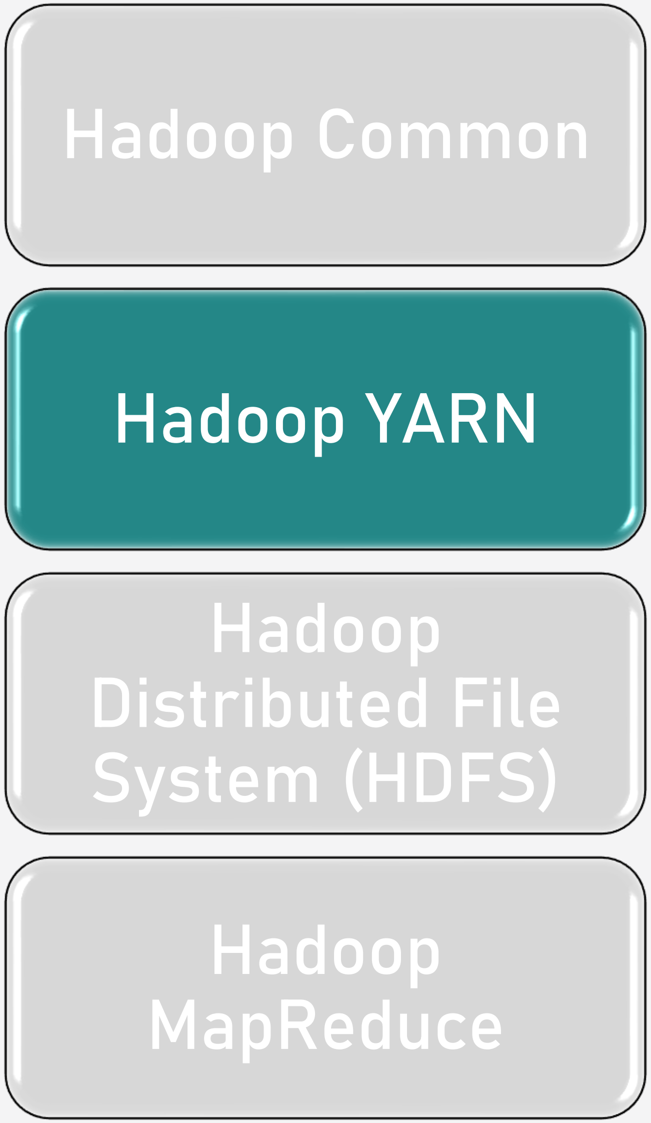
Hadoop Common

Hadoop YARN

Hadoop Distributed File System (HDFS)

Hadoop MapReduce

Hadoop Modules



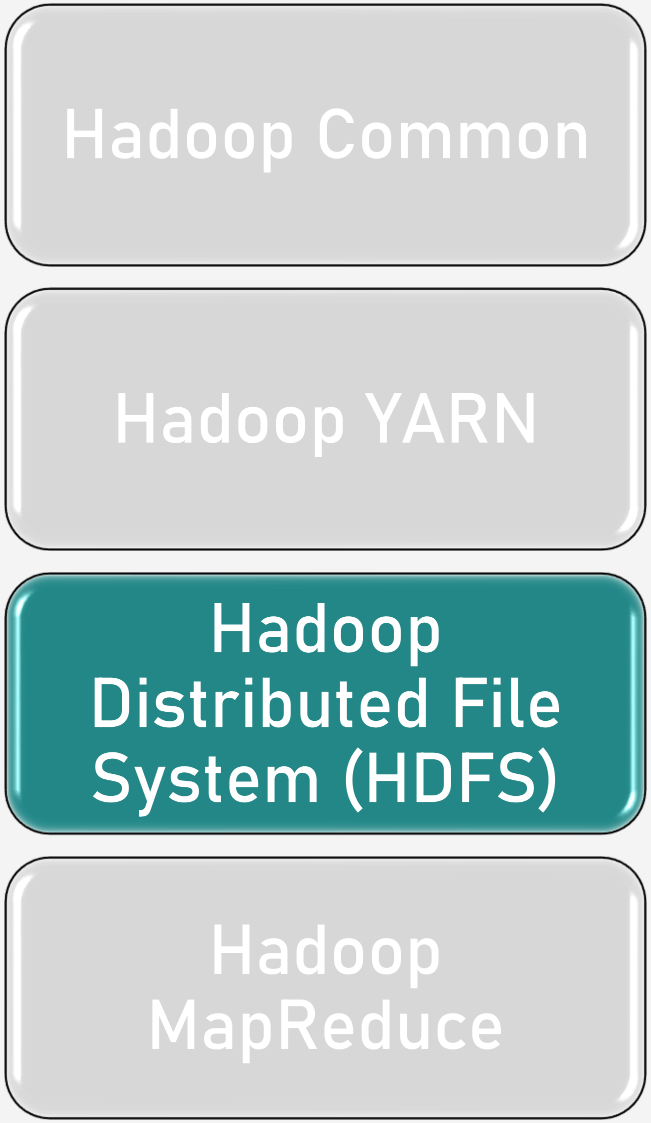
Hadoop Common

Hadoop YARN

Hadoop Distributed File System (HDFS)

Hadoop MapReduce

Hadoop Modules



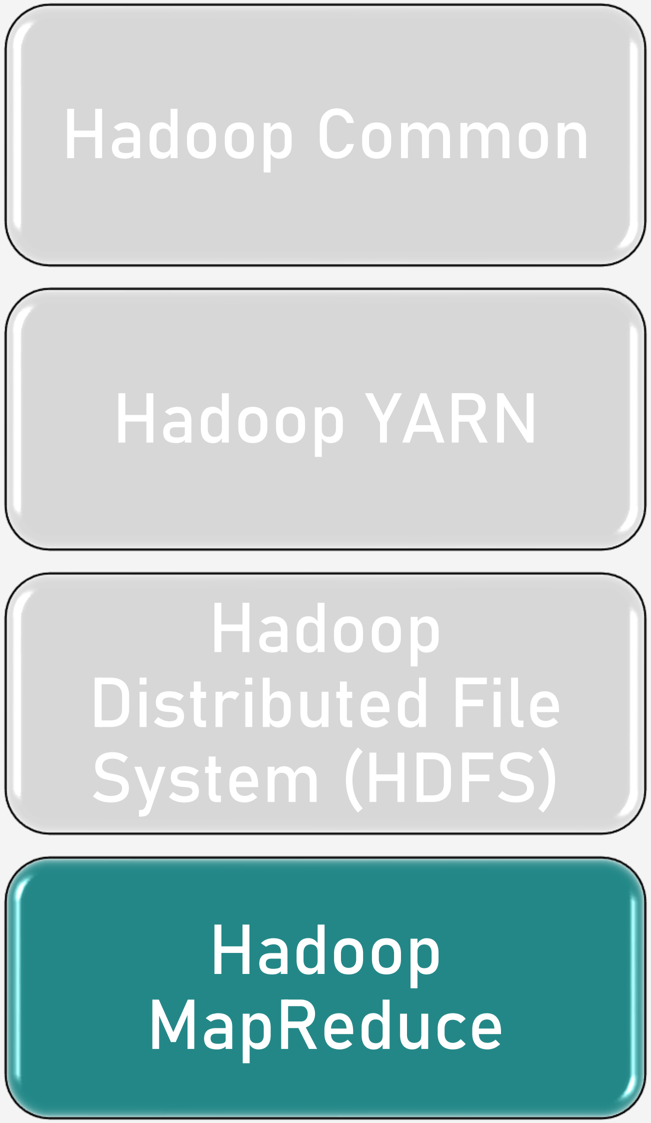
Hadoop Common

Hadoop YARN

Hadoop Distributed File System (HDFS)

Hadoop MapReduce

Hadoop Modules



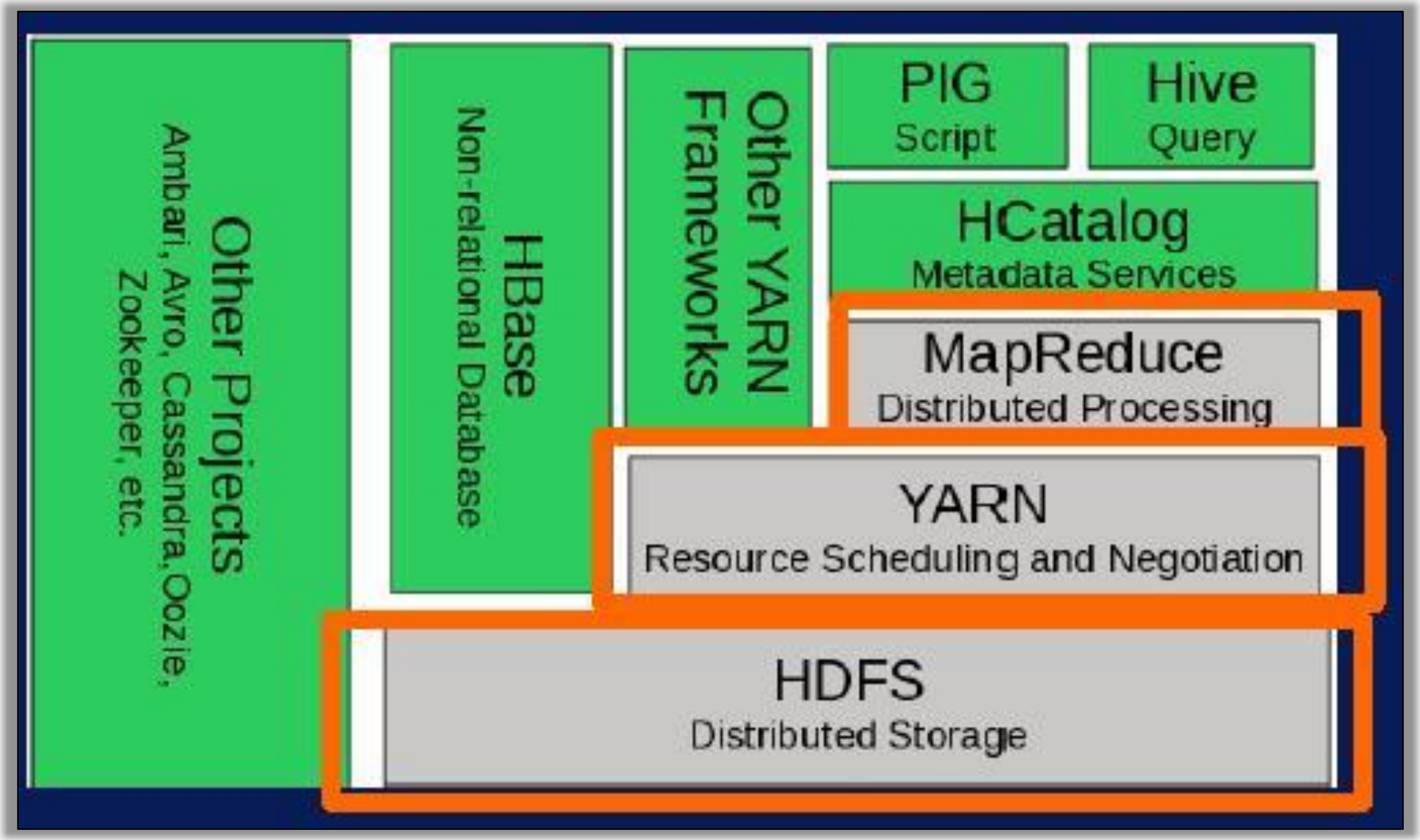
Hadoop Common

Hadoop YARN

Hadoop Distributed File System (HDFS)

Hadoop MapReduce

Hadoop Modules



Hadoop Modules- Hadoop Common

- Hadoop Common refers to the collection of common utilities and libraries that support other Hadoop modules.
- An essential part or module of the Apache Hadoop Framework, along with the Hadoop Distributed File System (HDFS), Hadoop YARN and Hadoop MapReduce.

Hadoop Modules- Hadoop Common

- Assumes that hardware failures are common and that these should be automatically handled in software by the Hadoop Framework.
- Also, known as Hadoop Core.

Hadoop Modules- Hadoop Common

Hadoop Common Package

- Considered as the base/core of the framework as it provides essential services and basic processes such as abstraction of the underlying operating system and its file system.
- Contains the necessary Java Archive (JAR) files and scripts required to start Hadoop.

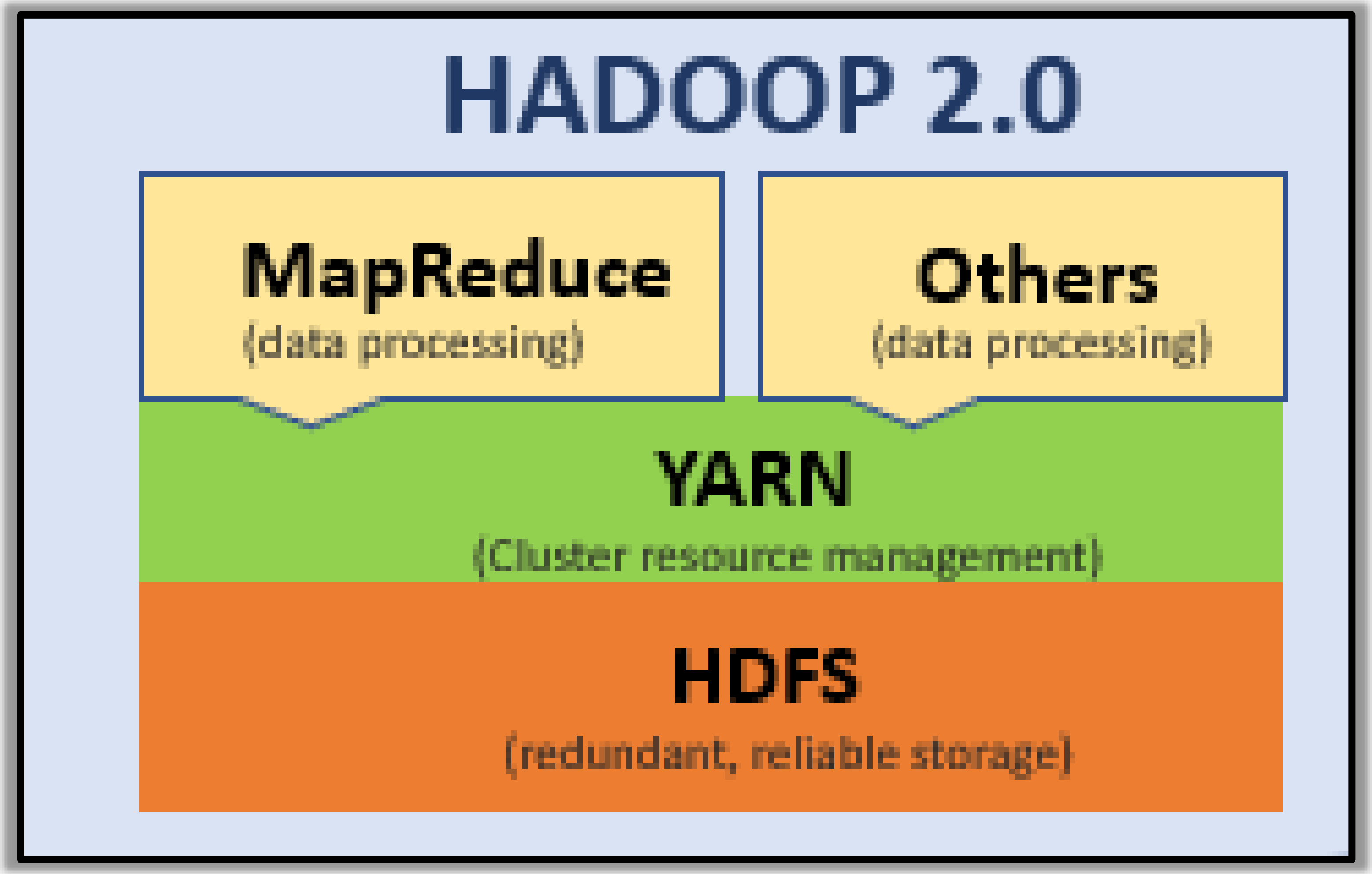
Hadoop Modules- Hadoop Yarn

- One of the main components in Hadoop.
- Stands for Yet Another Resource Negotiator though it is called as Yarn by the developers.
- Resource management platform responsible for managing compute resources in the cluster and using them in order to schedule users and applications.
- Used for job scheduling and resource management.

Hadoop Modules- Hadoop Yarn

- Yarn is the parallel processing framework for implementing distributed computing clusters that processes huge amounts of data over multiple compute nodes.
- Hadoop Yarn allows for a compute job to be segmented into hundreds and thousands of tasks.

Hadoop Modules- Hadoop Yarn



Hadoop Modules- Hadoop Yarn

Major components responsible for all the YARN operations are as follows:

- Resource Manager
- Node Manager
- Application Master

Hadoop Modules- Hadoop Yarn

Resource Manager- Resource Manager allocates the cluster resources.

- Resource Management
- Scheduling Management
- Application Management
- Containers in Hadoop
- Resource Containers

Hadoop Modules- Hadoop Yarn

Yarn Operation- Yarn uses master servers and data servers. There is only one master server per cluster. It runs the resource manager daemon. There are many data servers in the cluster, each one runs on its own Node Manager daemon and the application master manager as required.

Hadoop Modules- Hadoop Yarn

Key Features of Yarn

- Multi-Tenancy
- Sharing Resources
- Cluster Utilization
- Fault Tolerance
- Scalability
- Compatibility

Hadoop Modules- Hadoop Distributed File System (HDFS)

Hadoop Distributed File System- Major component of Apache Hadoop

- Handles large data sets running on commodity hardware.
- Allows multiple files to be stored and retrieved at the same time at an unprecedented speed.
- Used to scale a single Apache Hadoop cluster to hundreds (and even thousands) of nodes.

Hadoop Modules- Hadoop Distributed File System (HDFS)

Goals of HDFS:

- Fast Recovery from Hardware Failures.
- Access to Streaming Data.
- Accommodation of Large Data Sets.
- Portability.

Hadoop Modules- Hadoop Distributed File System (HDFS)

HDFS Key Features	Description
Storing bulks of data	HDFS is capable of storing terabytes and petabytes of data.
Minimum intervention	It manages thousands of nodes without operators’ intervention.
Computing	HDFS provides the benefits of distributed and parallel computing at once.
Scaling out	It works on scaling out, rather than on scaling up, without a single downtime.
Rollback	HDFS allows returning to its previous version post an upgrade.
Data integrity	It deals with corrupted data by replicating it several times.

An Example of HDFS

Example of HDFS

Hadoop Modules- Hadoop Distributed File System (HDFS)

IBM and Cloudera have partnered to offer an industry-leading, enterprise-grade Hadoop distribution, including an integrated ecosystem of products and services to support faster analytics at scale.

Why Should You Use HDFS?

- HDFS stores the same data in multiple sets.
- More suited for batch processing applications rather than for interactive use.
- HDFS works exclusively well for large datasets, offering high-aggregate data bandwidth.
- Highly Profitable

Hadoop Modules- MapReduce

- A programming model that scales data across a lot of different processes.
- MapReduce is defined as the framework of Hadoop, which is used to process a huge amount of data parallelly on large clusters of commodity hardware in a reliable manner.

Hadoop Modules- MapReduce

How does MapReduce in Hadoop make working so easy?

That’s all for now…