Introduction to Hadoop framework

Presented by Mandisa

Table of content

- Introduction to Hadoop
- Hadoop Components
- Comparison of Hadoop versions
- Hadoop use-cases
- Questions/ Break
- Lab 1 demonstration

Introduction to Hadoop

- Big data as a concept is a vast amount of data (structured and unstructured) collected as means to extract meaningful insights .

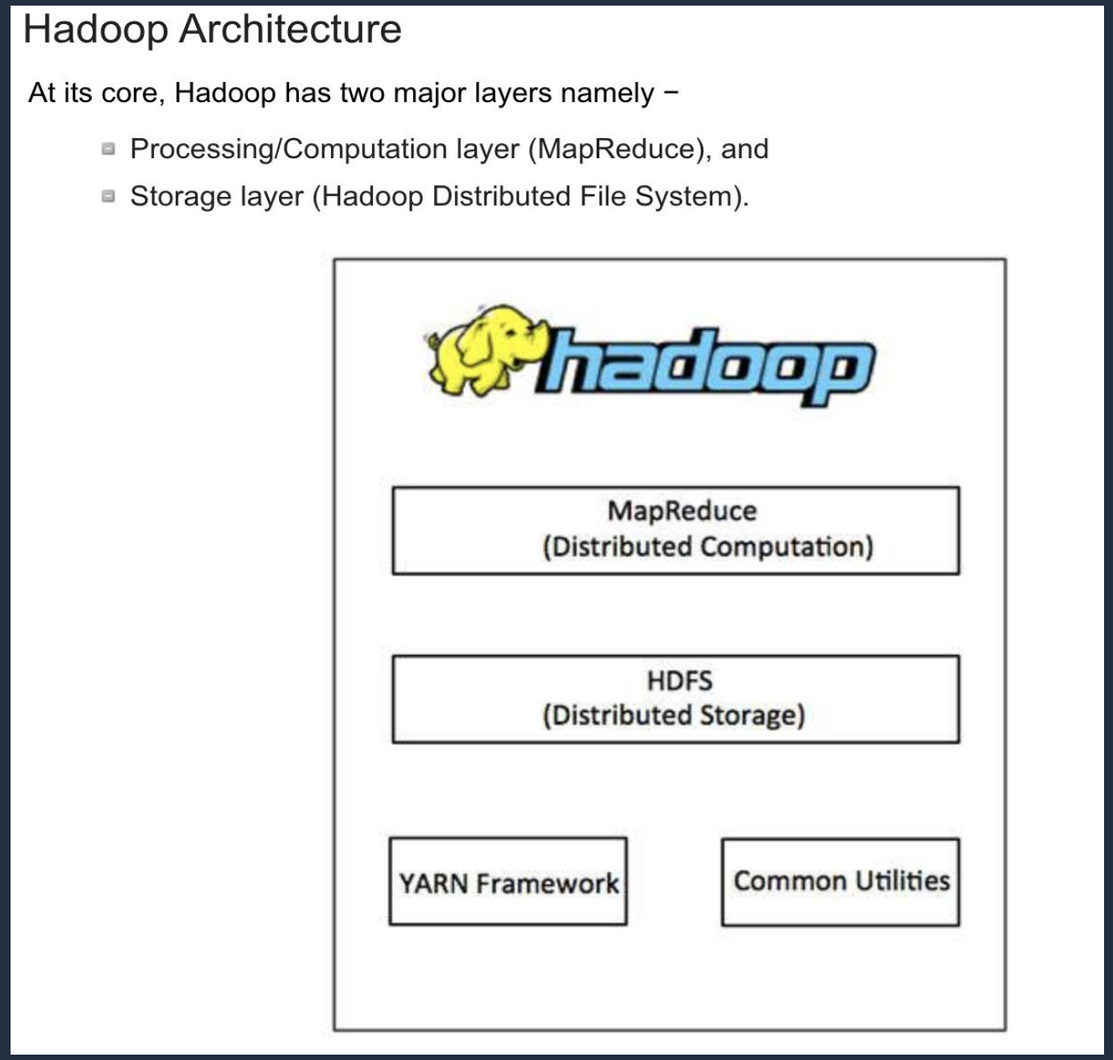
- What are the requirements of big data?
    - High storage space
    - Processing power
    - Scalability

There is a need to handle the big data Vs (volume, velocity, variety, veracity)!

aws

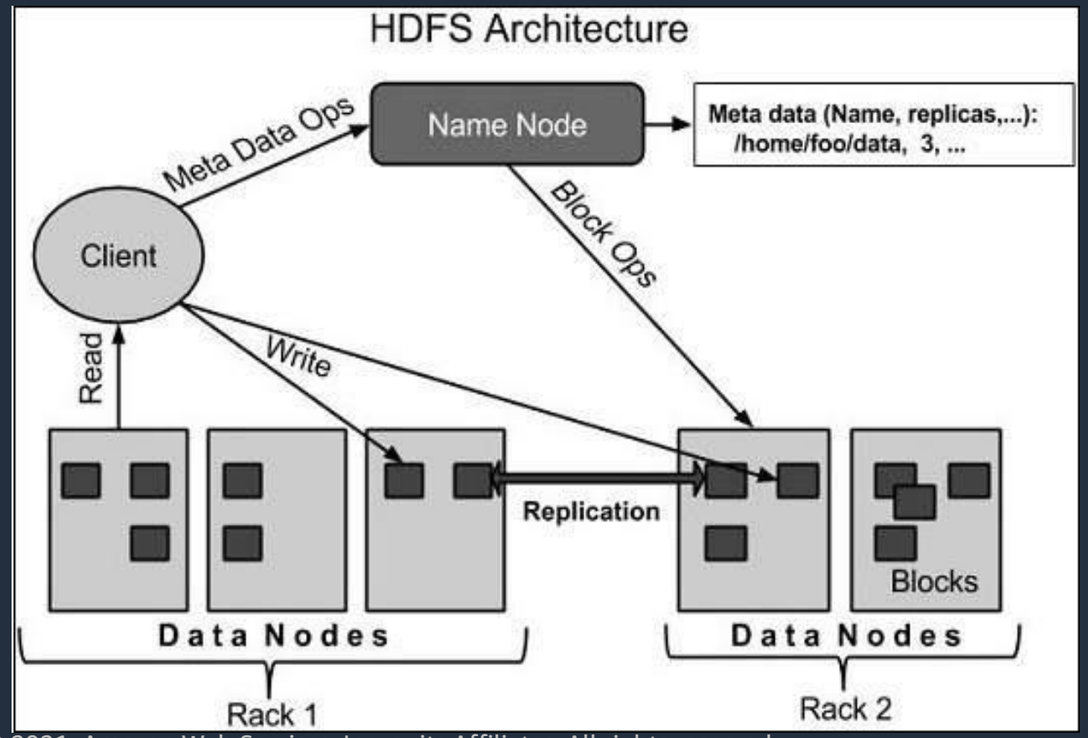
Introduction to Hadoop cont.

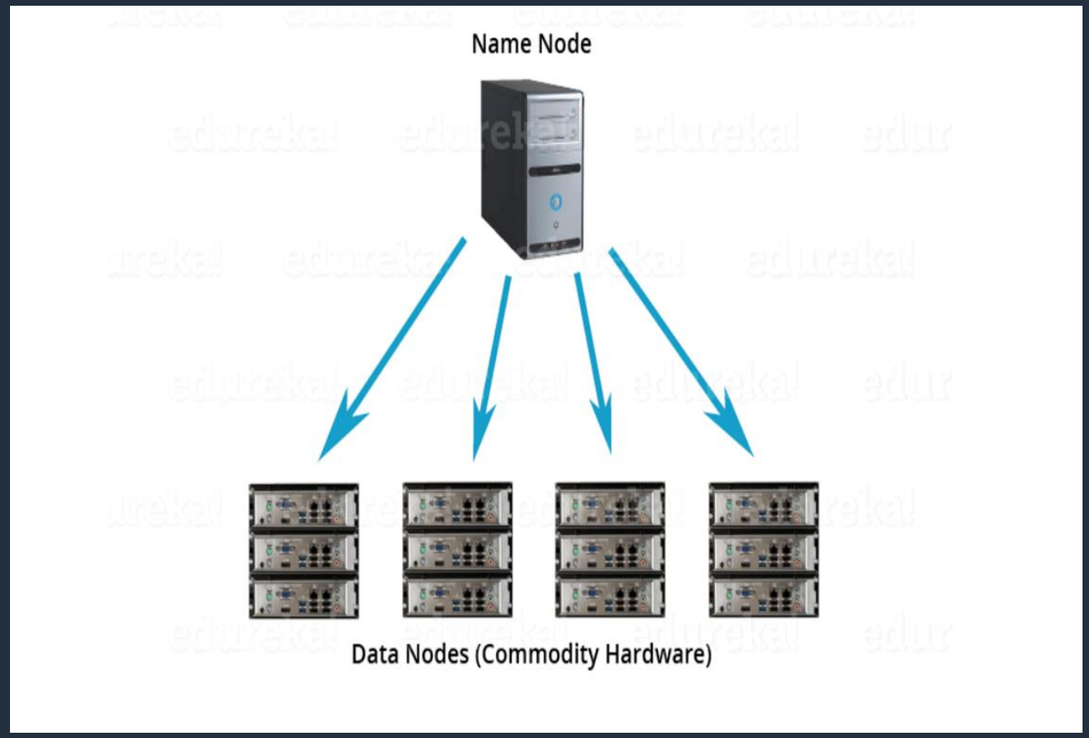
- Hadoop is defined as an open source framework that is used to store, process, and analyse chunks of data.
    - HDFS (Hadoop Distributed File System) for data storage
    - MapReduce/ Yarn (Yet another resource negotiator) for data processing



Hadoop core components: HDFS

- A distributed file system that runs on commodity hardware used for managing and storing data in a form of blocks across the cluster. The configuration is maintained on the cluster using hdfs-site.xml and core-site.xml files.
- Consists of two components namely, name node, and data node.





Hadoop core components: HDFS cont.

Name node

- Runs on a master daemon
- Manages and maintain data nodes
- Records the data metadata in memory such as size, permission, storage location, etc.
- Receives heartbeat and block report from all the data nodes.
    - ‘hdfs dfsadmin –report’, ‘hdfs fsck /’

Data node

- Runs on a slave daemon
- Stores the actual data in disk space.
- Responsible for serving read and write requests from the clients.
- When a data node is down, it does not affect the availability of data or the cluster.

Poll

Which component in Hadoop is aware of the block size and its location?

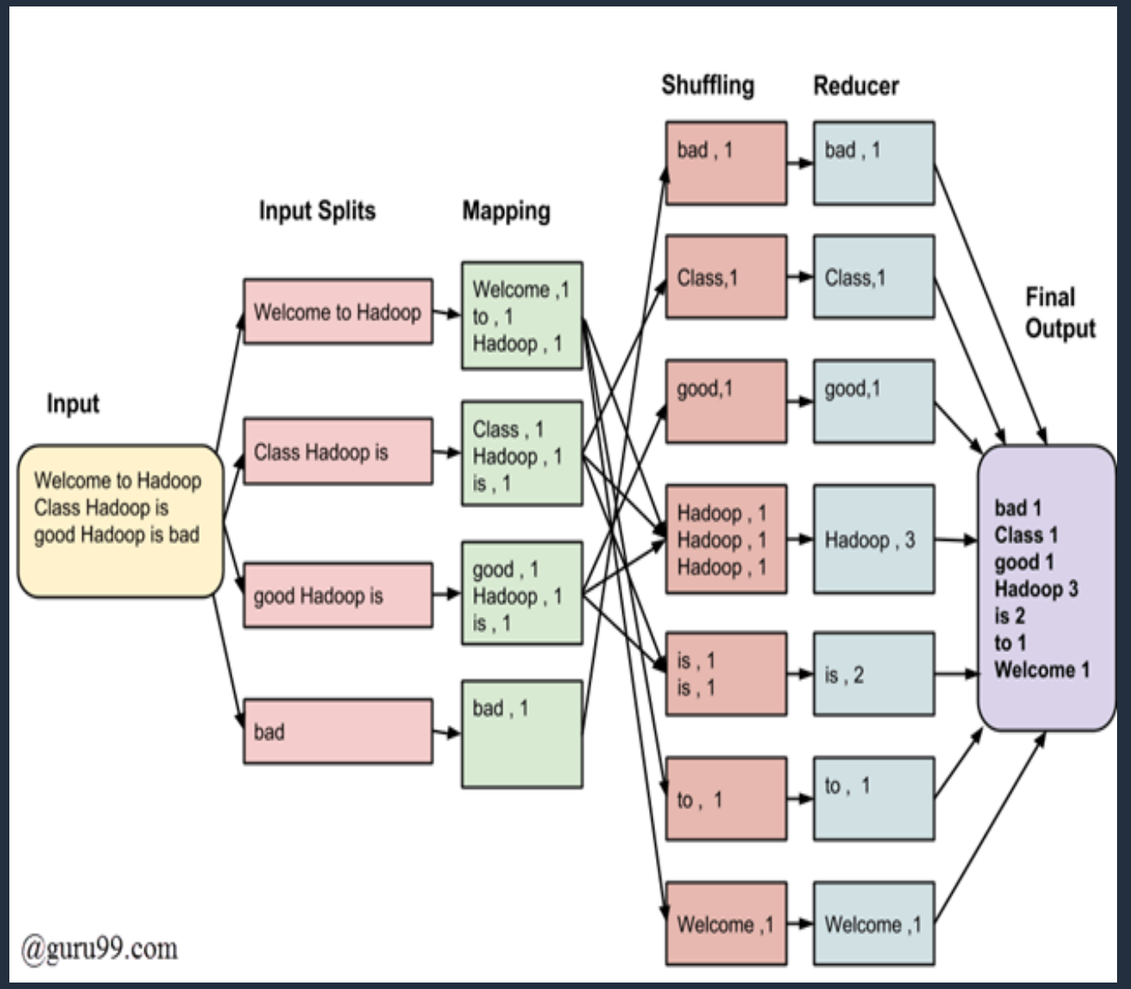
Hadoop core components: Map Reduce

- MapReduce is a software framework and programming model that allows easy writing of applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner.
- There are two phases involved in a MapReduce job:
    - Map – splits and maps data
    - Reduce – shuffles and reduce the data
- The input to each phase is **key-value** pairs
- Limitation:
    - Only supports batch processing not real-time data processing
    - Not suitable for small data files < 128 MB (Block size)

Hadoop core components: MapReduce cont.

For an example MapReduce word count job processes:

- Input splits – the input file is divided into a fixed size of records.
- Mapping - Count a number of occurrences of each word from input splits.
- Shuffling - Same words are grouped together along with their respective frequency.
- Reducing – Summarizes the complete data set using results from the shuffling phase.



Poll

True/ False: The input split process encounters performance issues when the split is configured with the same values as the HDFS block size.

Hadoop core components: YARN

- Apache YARN (Yet Another Resource Negotiator) is a resource management system introduced in Hadoop v2 to improve the MapReduce implementation, but it is generally enough to support other distributed computing paradigms such as Spark, hive, Hbase and others.

- YARN makes use of the components namely, resource manager, node manager, application master and containers for job scheduling, processing of activities by allocating resources and scheduling tasks.

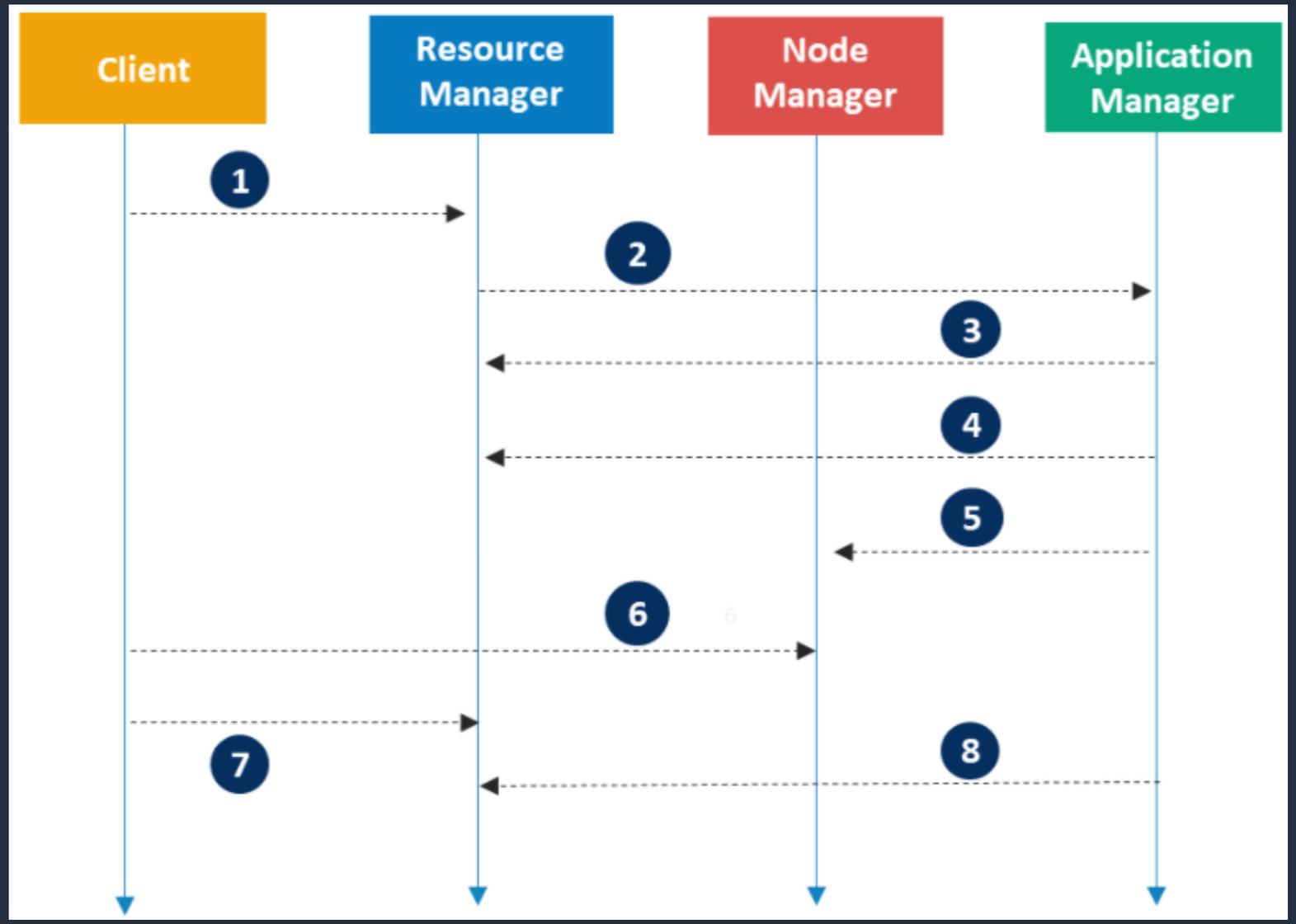
aws

Hadoop core components: YARN cont.

Components:

- Resource manager runs on a master daemon and used to manage resource allocation in the cluster.
- Application manager is responsible for managing user job lifecycle and resource requirements for individual applications.
    - Node Manager and monitors the execution of tasks.
- Node manager runs on a slave daemon and is used for managing the execution of tasks from the data node.
- Containers are packages of resources (RAM, CPU, Network, HDD and others) on a single node used to run jobs.

Hadoop core components: YARN cont.



How application workflow is carried in YARN

aws

Hadoop core components: YARN cont.

Useful commands and tools to monitor YARN applications:

- Yarn rmadmin (https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YarnCommands.html)
    - Reload queue properties
    - Refresh host information at the resource manager

- View/ monitor running yarn applications from the command line
    - Yarn top
    - Yarn application list \
    - Yarn application –kill <application id>

aws

Poll

Which component in YARN is responsible for spinning up containers?

Key differences between Hadoop versions

- You can confirm Hadoop version used by the cluster – Hadoop version command.

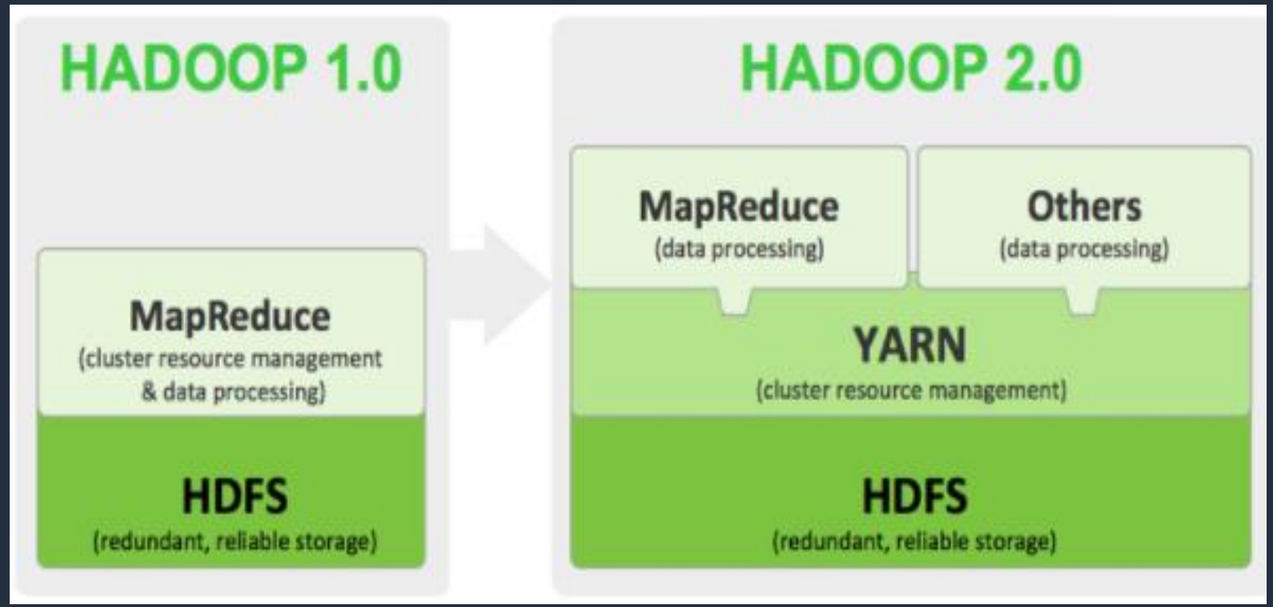
- There are currently 3 Hadoop versions
    - Hadoop 1,
    - Hadoop 2,
    - and Hadoop 3.

aws

Key differences between Hadoop versions: V1 and V2 cont.

Hadoop v2 was introduced to mitigate v1 limitations such as:

- No suitable for real-time and data streaming.
- Runs only MapReduce jobs
- Job tracker is a single point of failure
- No multi-tenancy support
- Scales up to 4000 nodes



Key differences between Hadoop versions: V2 and V3 cont.

Feature	Hadoop v2	Hadoop v3
Min java version required	Java 7	Java 8
Fault-tolerance	Replication	Erasure coding
Storage scheme	3x replication factor for data reliability with 200% overhead	Erasure coding for data reliability with 50% overhead
Yarn Timeline service	Scalability issues	High scalable and reliable
Standby name node	Support only 1	Supports 2 or more
Heap management	Manual tuning of hadoop heap size	Auto tunes heap

Hadoop use-cases

Finance

Government

Sentiment Analysis

Healthcare

Security and Law

Retail

Understanding Customers

Advertisement

aws

Questions

Will resume in 10 minutes

Lab 1: MapReduce job execution

Link for the lab:

https://github.com/aws-support-bigdata-cpt-vls/2021/tree/main/Day%201/Hadoop%20Lab

Will resume in 15 minutes for HDFS Content presentation