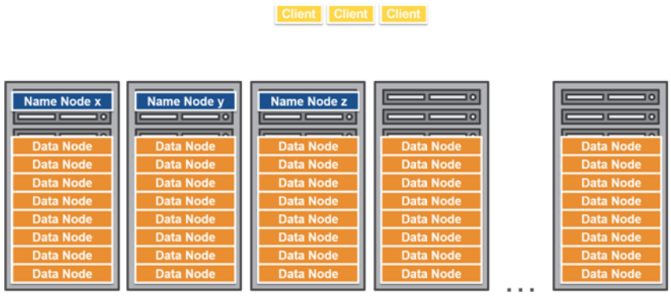
Module 1: Big Data Overview

1. Apache Hadoop

- It's the big data standard which is a scalable data storage and processing framework. And it also could ingest, process, and aggregate external data.
- Results can be exported.
- Function:
  - I. Hadoop Common: Hadoop modules.
  - II. HDFS: A system for store data.
  - III. Hadoop YARN: It's a framework for scheduling and execution of data processing.
  - IV. Hadoop MapReduce: it's YARN-base system for processing large dataset on the cluster.

2. HDFS

- HDFS is from Google DFS (Data File System).
- Data Node stores data in the following figure:



3. YARN

- YARN = **Y**et **A**nother **R**esource **N**egotiator
- **Application:**
  - I. Batch: e.g. original Map/Reduce (v1)
  - II. Interactive: Tez (ASF Incubator) is combined with application of Pig & Hive.
  - III. Online: e.g. HBase
  - IV. Streaming: e.g. Strom