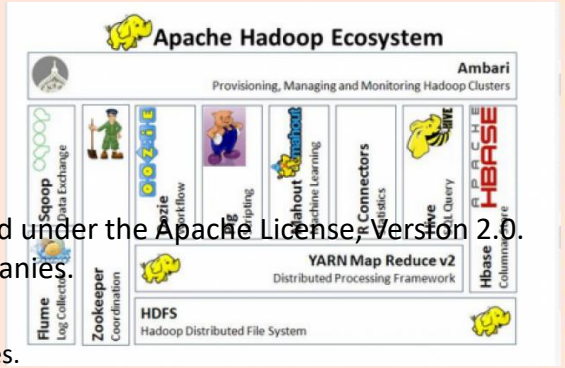
CBTU presents a course on **Big data and Hadoop**

Module 2: hadoop

Section 2.1: Hadoop introduction



BIG DATA

All the logos, trademarks are copyright of the respective companies.

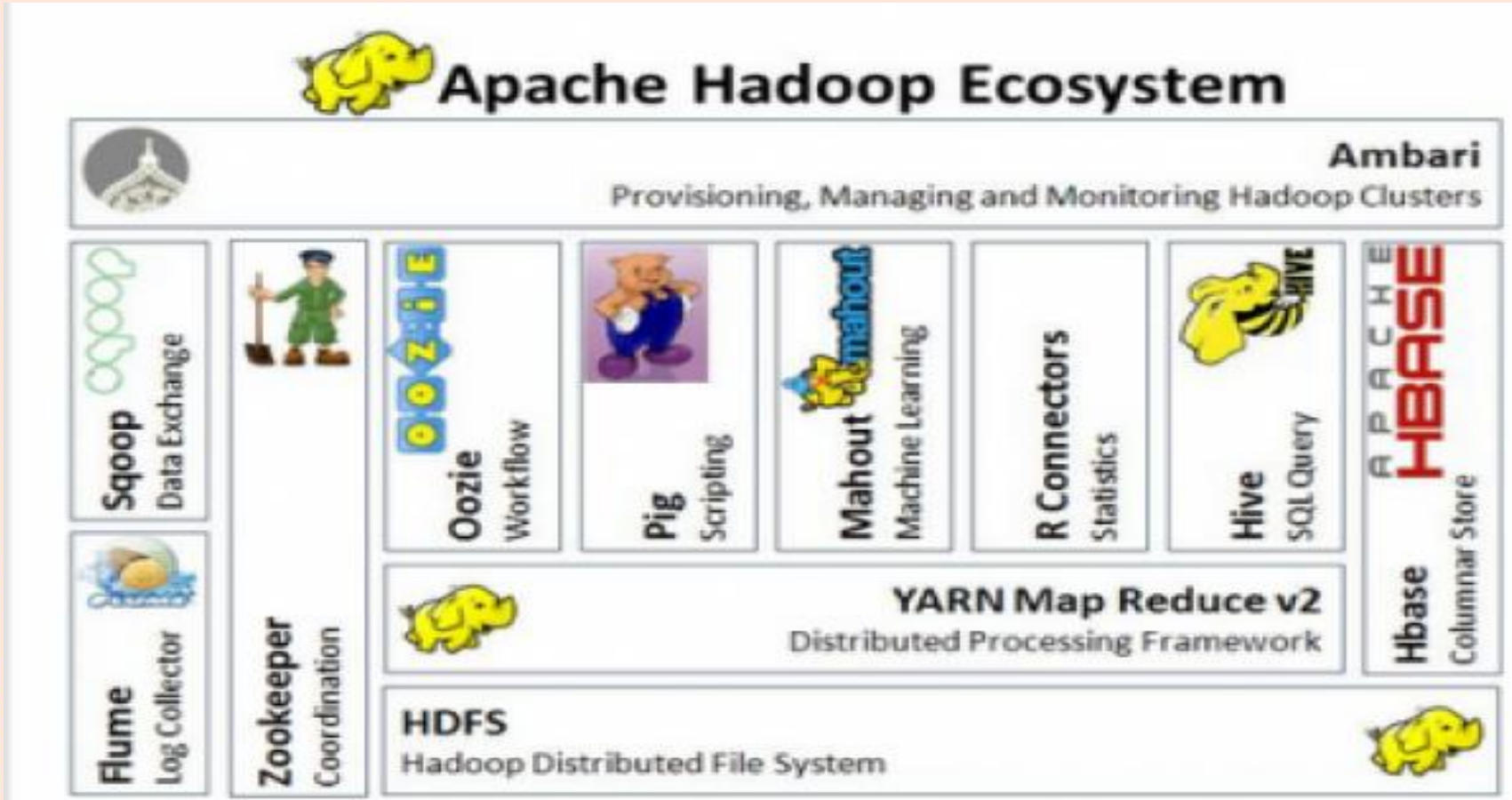
What is Hadoop

- Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models.
  - It is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop and Java

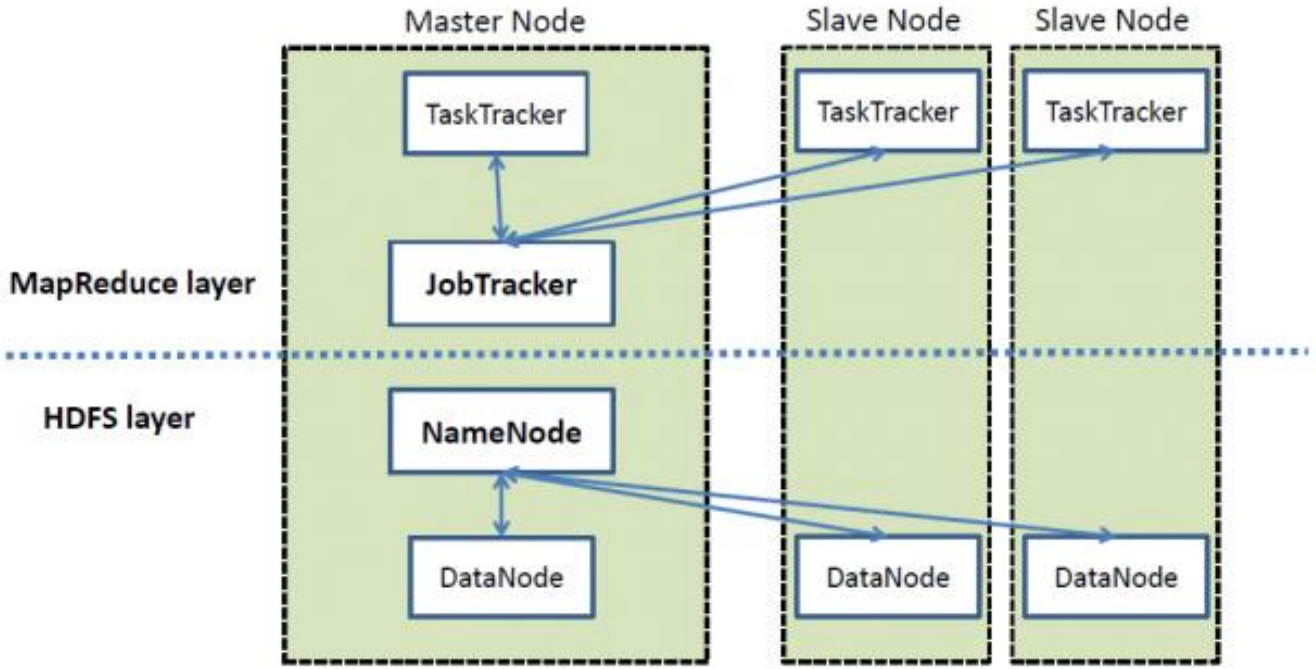
The Hadoop framework is mostly written in the Java language, with some native code in C and command line utilities written as shell scripts.

- Any programming language can be used with "Hadoop Streaming" to implement the "map" and "reduce" parts of the user's program.



Hadoop 2.0 Ecosystem - Apache License 2.0

High Level Architecture of Hadoop



Source: opensource.com

Apache Hadoop framework

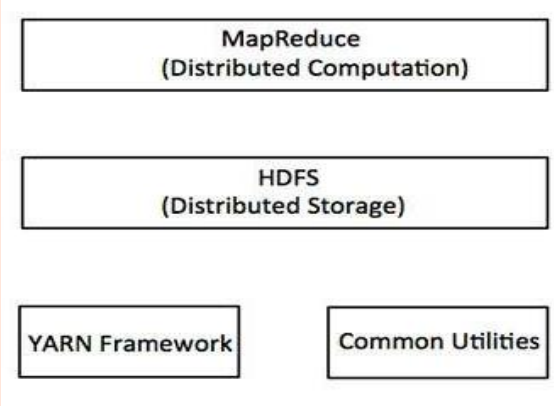
Apache Hadoop framework is composed of the following modules:

- Hadoop Common – contains libraries and utilities needed by other Hadoop modules.
- Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines.
- Hadoop YARN – manages computing resources in clusters .
- Hadoop MapReduce – an implementation of the MapReduce programming model for large scale data processing.

Hadoop Core

Hadoop core has two major components:

- Processing/Computation part (MapReduce)
- Storage part (Hadoop Distributed File System)



Hadoop nodes

Hadoop implementation creates four unique node types for cataloging, tracking, and managing data throughout the infrastructure as below:

- **Name Node (Master):** maintains the index and location of every data node.
- **Data Node (worker/slave):** these are the repositories for the data, and consist of multiple smaller database infrastructures.
- **Client Node:** this represents the user interface to the big data implementation and query engine. The client could be a server or PC with a traditional user interface.
- **Job tracker:** represents the software job tracking mechanism to distribute and aggregate search queries across multiple nodes for ultimate client analysis.

Jobs and talent pool

- With new technologies new data scientists talent pool is in demand to manage and analyze huge data/datasets.
- Huge gap in the demand and availability of the Data science and Big data talent.

Hadoop related projects

- Hbase: Bigtable-like structured storage system for HDFS
- Apache Pig: High-level data-flow language
- Hive: Data warehouse infrastructure
- ZooKeeper: Coordination service for distributed applications.
- Hama: Computing techniques
- Mahout: Scalable Machine Learning algorithms using Hadoop
- Apache Gora: Provides an in-memory data model
- Etc.

Advantages of Hadoop

- Allows to quickly write and test distributed systems.
- Efficient, and it automatically distributes the data and work across the machines.
- Designed to detect and handle failures at the application layer.
- Scalability: Servers/Nodes can be added or removed from the cluster dynamically.
- Open source, Java based and compatible on all the platforms.

Commercial applications

- Machine learning
- Data mining
- Marketing analytics
- Healthcare
- +many others

Prominent users

- Facebook has largest Hadoop cluster in the world with 100s of PBs of storage and the data was growing by roughly half a PB per day.
- Hadoop adoption is widespread: more than half of the Fortune 50 use Hadoop.

Hadoop flavors

- Apache Hadoop
- Hortonworks
- Cloudera
- IBM

Thanks for watching

CBTUniversity.com
learnq@CBTUniversity.com
+91 963 246 5599
/CBTUniversity