Apache Hadoop

Introduction

- Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computing.
- It provides a software framework for distributed storage and processing of Big Data using the MapReduce Model.

Trivia

- The original authors of Hadoop are Doug Cutting and Mike Cafarella.
- The name Hadoop comes from a yellow toy elephant that was owned by Doug Cutting’s son.
- Hadoop was first released to the public in 2006.
- The current stable version is 3.3.2 (as of March 06, 2022).

Implementation

- The main implementation language is Java with some native code written in C and command-line utilities written in shell scripts.
- Though Java is commonly used for MapReduce jobs, through the use of Hadoop Streaming any programming language can be used to write MapReduce tasks.
- Requires JRE 1.6+ to run.

Base Modules

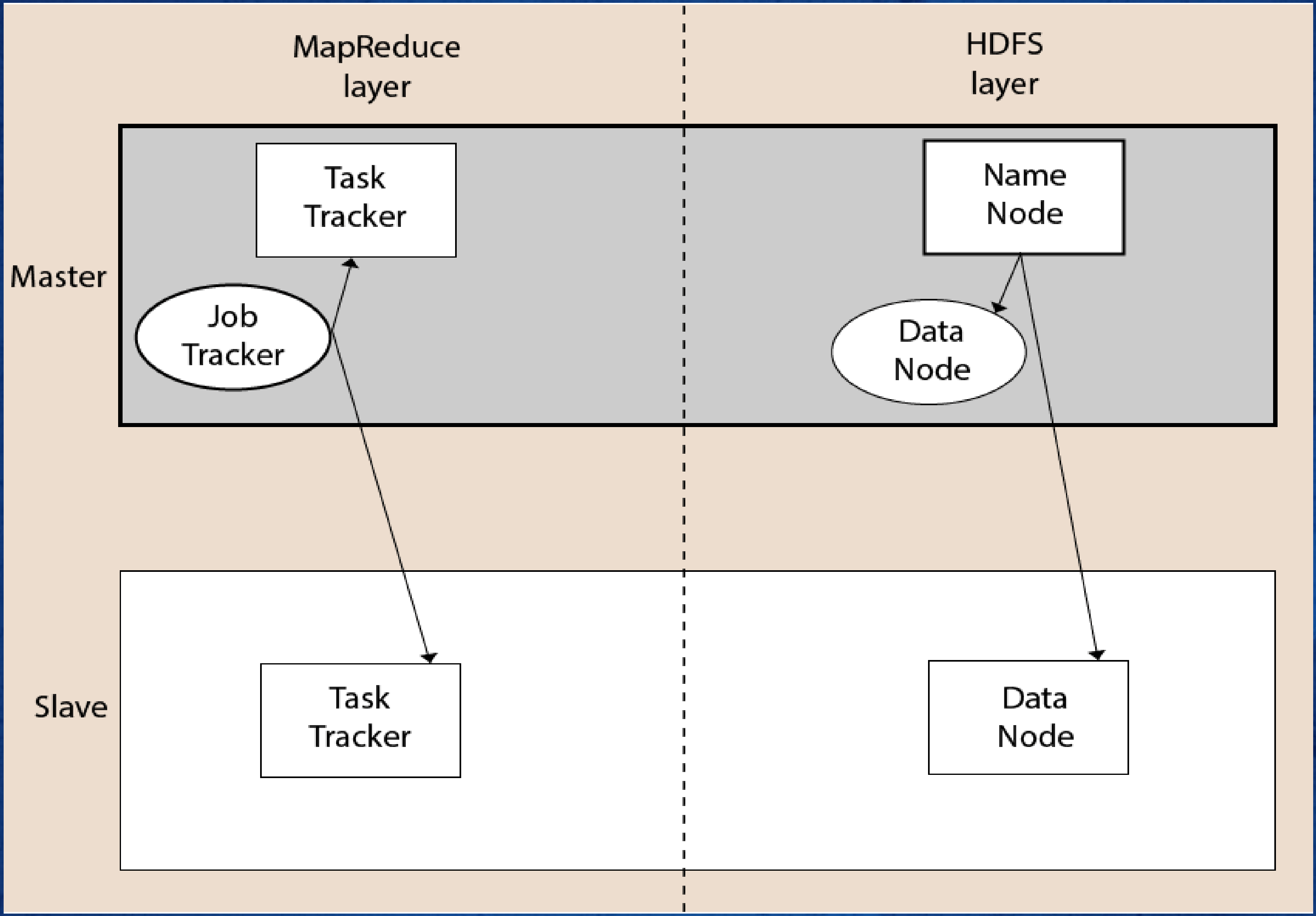
- Hadoop Common: Libraries and utilities needed by other modules.
- Hadoop Distributed File System (HDFS): A distributed file system which stores data on commodity hardware.

- Hadoop YARN (Yet Another Resource Negotiator): A platform responsible for managing computing resources in a cluster and using them to schedule user’s applications. (2012)
- Hadoop MapReduce: An implementation of the MapReduce programming model for large scale data processing.
- Hadoop Ozone: An object store for Hadoop. (2020)

Architecture

- Hadoop runs in a master-slave architecture.
- Hadoop consists of the Hadoop Common package which allows file system and OS level abstractions, a MapReduce engine (MapReduce/MR1 or YARN/MR2) and the Hadoop Distributed File System (HDFS).
- The Hadoop Common package contains JAR files and scripts required to run Hadoop.

- A Hadoop instance is divided into HDFS and MapReduce.
- HDFS is used to store the data and MapReduce is used to process the data.


MapReduce layer
HDFS layer
Master
Task Tracker
Job Tracker
Name Node
Data Node
Slave
Task Tracker
Data Node

HDFS

- The Hadoop Distributed File System (HDFS) is a distributed, scalable and portable file system written in Java for the Hadoop framework.
- HDFS works in a master-slave configuration with one master (Name Node) and multiple slaves (Data Nodes).

- HDFS has the following services:
  - Name Node
  - Secondary Name Node
  - Job Tracker
  - Data Node
  - Task Tracker.
- Top three services are Master services and the rest are Slave services.

HDFS Services

- Name Node:
  - Only one present in a cluster.
  - Tracks files, manages filesystem and has metadata of all stored data within it.
  - Has a direct contact with the client.
- Secondary Name Node:
  - Also called checkpoint node.
  - Takes care of the file system metadata checkpoints found in the Master node.

- Job Tracker:
  - Receives MapReduce execution requests from the client.
  - It then talks to the NameNode for the location of data to be processed upon which the NameNode returns the metadata of the required data.

- Data Node:
  - Stores data as blocks.
  - Actually stores the data into the HDFS.
  - Is a slave node to the Name Node.
- Task Tracker:
  - Slave node to the Job Tracker.
  - Takes the job (with the code) from the Job Tracker and applies the code to the data file.

MapReduce Engine

- The client submits a MapReduce job to a JobTracker which has a single instance over the entire cluster.
- The JobTracker then pushes the job onto an available TaskTracker nodes while making sure that the work is kept as close as possible to the data.

Advantages

- Fast
- Scalable
- Cost Effective
- Resilient to failure

Applications

- Log or clickstream analytics.
- Marketing Analytics.
- Machine Learning and Data Mining.
- Image Processing.
- XML Message Processing.
- Web Crawling.
- Archival work.

Thanks