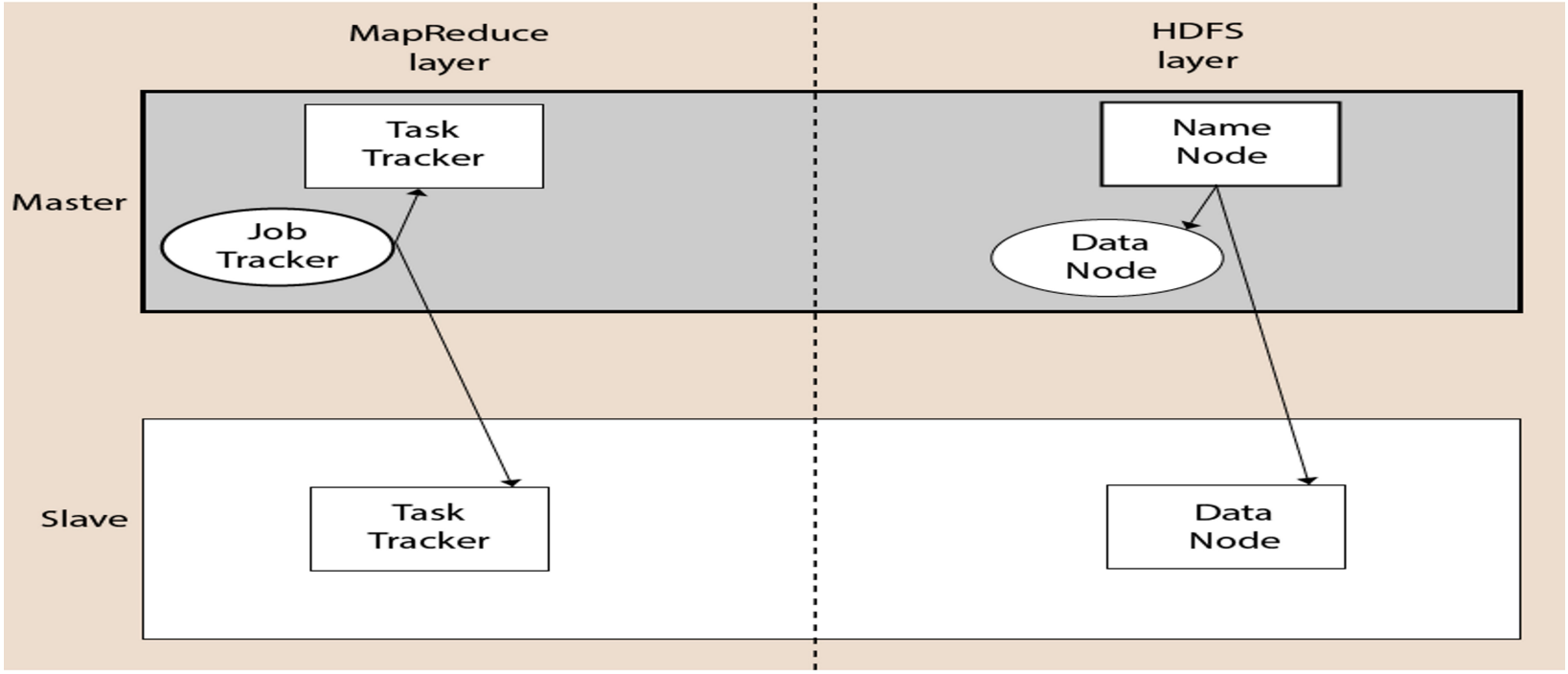
# Day-4 Hadoop Distributed File System

08 May 2024 09:50

## Hadoop Distributed File System




The Java language is used to develop HDFS. So any machine that supports Java language can easily run the NameNode and DataNode software.

### NameNode:

- It is a single master server exist in the HDFS cluster.

- As it is a single node, it may become the reason of single point failure.
- It manages the file system namespace by executing an operation like the opening, renaming and closing the files.
- It simplifies the architecture of the system.

### DataNode

- The HDFS cluster contains multiple DataNodes.
- Each DataNode contains multiple data blocks.
- These data blocks are used to store data.
- It is the responsibility of DataNode to read and write requests from the file system's clients.
- It performs block creation, deletion, and replication upon instruction from the NameNode.

### Job Tracker

- The role of Job Tracker is to accept the MapReduce jobs from client and process the data by using NameNode.
- In response, NameNode provides metadata to Job Tracker.

### Task Tracker

- It works as a slave node for Job Tracker.
- It receives task and code from Job Tracker and applies that code on the file. This process can also be called as a Mapper.

## MapReduce Layer

The MapReduce comes into existence when the client application submits the MapReduce job

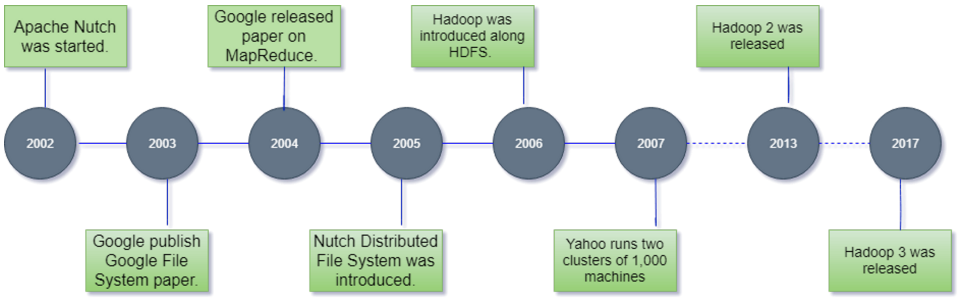
to Job Tracker. In response, the Job Tracker sends the request to the appropriate Task Trackers. Sometimes, the TaskTracker fails or time out. In such a case, that part of the job is rescheduled.

## Advantages of Hadoop

- **Fast:** In HDFS the data distributed over the cluster and are mapped which helps in faster retrieval. Even the tools to process the data are often on the same servers, thus reducing the processing time. It is able to process terabytes of data in minutes and Peta bytes in hours.
- **Scalable:** Hadoop cluster can be extended by just adding nodes in the cluster.
- **Cost Effective:** Hadoop is open source and uses commodity hardware to store data so it really cost effective as compared to traditional relational database management system.
- **Resilient to failure:** HDFS has the property with which it can replicate data over the network, so if one node is down or some other network failure happens, then Hadoop takes the other copy of data and use it. Normally, data are replicated thrice but the replication factor is configurable.

## History of Hadoop

The Hadoop was started by Doug Cutting and Mike Cafarella in 2002. Its origin was the Google File System paper, published by Google.




Let's focus on the history of Hadoop in the following steps: -

- In 2002, Doug Cutting and Mike Cafarella started to work on a project, **Apache Nutch.** It is

an open source web crawler software project.

- While working on Apache Nutch, they were dealing with big data. To store that data they have to spend a lot of costs which becomes the consequence of that project. This problem becomes one of the important reason for the emergence of Hadoop.
- In 2003, Google introduced a file system known as GFS (Google file system). It is a proprietary distributed file system developed to provide efficient access to data.
- In 2004, Google released a white paper on Map Reduce. This technique simplifies the data processing on large clusters.
- In 2005, Doug Cutting and Mike Cafarella introduced a new file system known as NDFS (Nutch Distributed File System). This file system also includes Map reduce.
- In 2006, Doug Cutting quit Google and joined Yahoo. On the basis of the Nutch project, Dough Cutting introduces a new project Hadoop with a file system known as HDFS (Hadoop Distributed File System). Hadoop first version 0.1.0 released in this year.
- Doug Cutting gave named his project Hadoop after his son's toy elephant.
- In 2007, Yahoo runs two clusters of 1000 machines.
- In 2008, Hadoop became the fastest system to sort 1 terabyte of data on a 900 node cluster within 209 seconds.
- In 2013, Hadoop 2.2 was released.
- In 2017, Hadoop 3.0 was released.

| Year | Event |
|---|---|
| 2003 | Google released the paper, Google File System (GFS). |
| 2004 | Google released a white paper on Map Reduce. |
| 2006 | • Hadoop introduced.<br>• Hadoop 0.1.0 released.<br>• Yahoo deploys 300 machines and within this year reaches 600 machines. |

| | |
|---|---|
| 2007 | • Yahoo runs 2 clusters of 1000 machines.<br>• Hadoop includes HBase. |
| 2008 | • YARN JIRA opened<br>• Hadoop becomes the fastest system to sort 1 terabyte of data on a 900 node cluster within 209 seconds.<br>• Yahoo clusters loaded with 10 terabytes per day.<br>• Cloudera was founded as a Hadoop distributor. |
| 2009 | • Yahoo runs 17 clusters of 24,000 machines.<br>• Hadoop becomes capable enough to sort a petabyte.<br>• MapReduce and HDFS become separate subproject. |
| 2010 | • Hadoop added the support for Kerberos.<br>• Hadoop operates 4,000 nodes with 40 petabytes.<br>• Apache Hive and Pig released. |
| 2011 | • Apache Zookeeper released.<br>• Yahoo has 42,000 Hadoop nodes and hundreds of petabytes of storage. |
| 2012 | Apache Hadoop 1.0 version released. |
| 2013 | Apache Hadoop 2.2 version released. |
| 2014 | Apache Hadoop 2.6 version released. |
| 2015 | Apache Hadoop 2.7 version released. |
| 2017 | Apache Hadoop 3.0 version released. |
| 2018 | Apache Hadoop 3.1 version released. |