HDFS FILE SYSTEM

HDFS is a distributed file system that handles large data sets running on commodity hardware. It is used to scale a single Apache Hadoop cluster to hundreds (and even thousands) of nodes. HDFS is one of the major components of Apache Hadoop, the others being MapReduce and YARN.

The goal of HDFS:

1. Fast recovery from hardware failures
2. Access to streaming data
3. Accommodation of large data sets
4. Portability

Hadoop----- 1) HDFS and 2) Map Reduce

HDFS is derived from Google File System (GFS)

Map Reduce is derived from Google Map Reduce (GMR)

* **File system:** It is layer acting between software and hardware.

Example of File system are NTFS, EXT and HDFS etc.

* **Block:** Big file into small chunks

Example NTFS 16 K, EXT 512 K etc.

* **Client and Servers**: Client sends the request and Server responds.
* **Types of file system**: Standalone-NTFS, EXT etc.

Distributed- HDFS, S3, CFS.

* **Types of distributed system**: Master and slave - Hadoop

Peer to Peer – Cassandra

* **Process:** Program in execution is called process.
* **Daemon process:** Background Process is called Daemon process.
* **Node:** Individual physical or virtual machine
* **Cluster:** Group of nodes
* **Client API:** Application program interface
* **Block:** 1 B- 64 MB (default)

128 MB (default)

* **Replication:** 1 B = 3R(default)

**An example of HDFS:**

* Consider a file that includes the phone numbers for everyone in the United States; the numbers for people with a last name starting with A might be stored on server 1, B on server 2, and so on.
* With Hadoop, pieces of this phonebook would be stored across the cluster, and to reconstruct the entire phonebook, your program would need the blocks from every server in the cluster.
* To ensure availability if and when a server fails, HDFS replicates these smaller pieces onto two additional servers by default. (The redundancy can be increased or decreased on a per-file basis or for a whole environment; for example, a development Hadoop cluster typically doesn’t need any data redundancy.) This redundancy offers multiple benefits, the most obvious being higher availability.
* The redundancy also allows the Hadoop cluster to break up work into smaller chunks and run those jobs on all the servers in the cluster for better scalability. Finally, you gain the benefit of data locality, which is critical when working with large data sets.

Diagram

Description automatically generated

Diagram

Description automatically generated

A picture containing diagram

Description automatically generated

**MAP REDUCE**

MapReduce performs the processing of large data sets in distributed and parallel manner.

MapReduce consists of two distinct tasks Map and Reduce

Two essential daemons of MapReduce: Job Tracker and Task Tracker.

Diagram

Description automatically generated

Diagram

Description automatically generated

MapReduce is a programming paradigm that enables massive scalability across hundreds or thousands of servers in a Hadoop cluster. As the processing component, MapReduce is the heart of Apache Hadoop. The term "MapReduce" refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.

MapReduce programming offers several benefits to help you gain valuable insights from your big data:

* Scalability. Businesses can process petabytes of data stored in the Hadoop Distributed File System (HDFS).
* Flexibility. Hadoop enables easier access to multiple sources of data and multiple types of data.
* Speed. With parallel processing and minimal data movement, Hadoop offers fast processing of massive amounts of data.
* Simple. Developers can write code in a choice of languages, including Java, C++ and Python.