

# Hadoop Distributed File System (HDFS)



0:05 / 17:43



## Hadoop Distributed File System (HDFS)

- based on the GFS. and provides a distributed file system.
- designed to run on large clusters (thousands of computers).
- Master/slave architecture.

# Hadoop Distributed File System (HDFS)

- based on the GFS. and provides a distributed file system.
- designed to run on large clusters (thousands of computers).
- Master / slave architecture.

Namesnode  Datanodes

Files

# Hadoop Distributed File System (HDFS)

- based on the GFS. and provides a distributed file system.
- designed to run on large clusters (thousands of computers).
- Master / slave architecture.

! Name node

Data nodes

- Files split into several blocks and stored in a set of data nodes.

## Building Blocks of Hadoop

1) Name node (Master) - stores metadata - files



# Hadoop Distributed File System (HDFS)

- based on the GFS. and provides a distributed file system.
- designed to run on large clusters (thousands of computers).
- Master / slave architecture.

1  
namenode

datanodes

- Files split into several blocks and stored in a set of data nodes.

## Building Blocks of Hadoop.

1) namenode (Master) - stores metadata - files

10 seconds

# Hadoop Distributed File System (HDFS)

- based on the GFS. and provides a distributed file system.
- designed to run on large clusters (thousands of computers).
- Master / slave architecture.

Name node

Data nodes

- Files split into several blocks and stored in a set of data nodes.

## Building Blocks of Hadoop

- 1) Name node (Master) - stores metadata - files which are broken down into file blocks, which nodes store those blocks. and the overall health of Dfs.
- server hosting these name node doesn't store user data or perform any computations.

Negative impact of Name node is that if it fails then the entire system is affected.



# Hadoop Distributed File System (HDFS)

- based on the GFS. and provides a distributed file system.
- designed to run on large clusters (thousands of computers).
- Master / slave architecture.

Name node

Data nodes

- Files split into several blocks and stored in a set of data nodes.

## Building Blocks of Hadoop

### 1) Name node (Master)

- server hosting the name node doesn't store user

data or perform any computations.

- stores metadata - files which are broken down into file blocks, which nodes store those blocks. and the overall health of DFS.

Negative impact of Name node is that if it fails then the entire system is affected.

- Name node      Data nodes
- Files split into several blocks and stored in a set of data nodes.

## Building Blocks of Hadoop.

### 1) Name node (Master)

- Server hosting the name node doesn't store user data or perform any computations.
  - Negative impact of Name node is that if it fails then the entire Hadoop cluster will fail.
- stores metadata - files which are broken down into file blocks, which nodes store those blocks. and the overall health of DFS.

### 2) Data Node (Slave). stores the actual data.

- Client can communicate directly with the Data Node.
- It performs reading and writing HDFS blocks to actual files on the local file system.



- Name node      Data nodes
- Files split into several blocks and stored in a set of data nodes.

## Building Blocks of Hadoop.

### 1) Name node (Master)

- Server hosting the name node doesn't store user data or perform any computations.
  - Negative impact of Name node is that if it fails then the entire Hadoop cluster will fail.
- stores metadata - files which are broken down into file blocks, which nodes store those blocks. and the overall health of DFS.

### 2) Data Node (Slave). stores the actual data.

- Client can communicate directly with the Data Node.
- It performs reading and writing HDFS blocks to actual files on the local file system.

### 3. Secondary Name node

- assistant that monitors the state of the cluster.
- Each cluster has one SNN.
- communicates with NN



### 3. Secondary Name node

- assistant that monitors the state of the cluster.
- Each cluster has one SNN.
- communicates with NN



10 seconds



### 3. Secondary Name node

- assistant that monitors the state of the cluster.
- Each cluster has one SNN.
- communicates with NameNode to take snapshots of HDFS metadata at intervals ~~to~~ defined by the cluster configuration.
- Snapshots helps in minimizing the downtime and loss of data.

### 4. Job Tracker

- link between your app. and Hadoop.
-



### 3. Secondary Name node

- assistant that monitors the state of the cluster.
- Each cluster has one SNN.
- communicates with NameNode to take snapshots of HDFS metadata at intervals defined by the cluster configuration.
- Snapshots helps in minimizing the downtime and loss of data.

### 4. Job Tracker

- link between your app. and Hadoop.
- 



- communicates with NameNode to take snapshots of HDFS metadata at intervals defined by the cluster configuration.
- Snapshots helps in minimizing the downtime and loss of data.

#### 4. Job Tracker

- link between your app. and Hadoop.
- determines the execution plan by determining —
  - files to process.
  - assign nodes to different task
  - monitors all tasks as they are running.
- If a task fails, it will automatically re-launch the task.
- Every cluster has only one Job Tracker.

#### 5. Task I



- communicates with NameNode to take snapshots of HDFS metadata at intervals defined by the cluster configuration.
- Snapshots helps in minimizing the downtime and loss of data.

#### 4. Job Tracker

- link between your app. and Hadoop.
- determines the execution plan by determining —
  - files to process,
  - assign nodes to different task
  - monitors all tasks as they are running.
- If a task fails, it will automatically re-launch the task.
- Every cluster has only one Job Tracker.

#### 5. Task I

Snapshots helps in minimizing the downtime and loss of data. Snapshots of HDFS metadata.

#### 4. Job Tracker

- link between your app. and Hadoop.
- determines the execution plan by determining —
  - files to process.
  - assign nodes to different task
  - monitors all tasks as they are running.
- If a task fail, it will automatically re-launch the task.
- Every cluster has only one Job Tracker.

#### 5. Task Tracker

(Heartbeat).

- to constantly communicate with the Job Tracker.
- If Job Tracker fails to receive a heartbeat from a Task Tracker.



Snapshots helps in minimizing the downtime and loss of data. Snapshots of HDFS metadata.

#### 4. Job Tracker

- link between your app. and Hadoop.
- determines the execution plan by determining —
  - files to process.
  - assign nodes to different task
  - monitors all tasks as they are running.
- If a task fail, it will automatically re-launch the task.
- Every cluster has only one Job Tracker.

#### 5. Task Tracker

(Heartbeat).

- to constantly communicate with the Job Tracker.
- If Job Tracker fails to receive a heartbeat from a Task Tracker.



Why you should learn  
DATA SCIENCE IN  
2021 ???  
Complete Roadmap for Data Science |  
Why you should learn Data Science in...

\*\*My personal experience

Complete  
Roadmap 5  
STEPS

18:56

Data Analyst Job | My first job |  
Experience | Data Talks 3

AS

Data Analyst

My Experience

8:09

What is Data Analysis? | Data Analysis  
using Python | Part 4 | Tech Classes

Python



What is Data  
Analysis

PART 4

6:33



17:36 / 17:43

