

# HDFS – Hadoop Distributed File System

Sundharakumar KB

Department of Computer Science and Engineering  
School of Engineering

Shiv Nadar University Chennai

# Big Data - Introduction

- Next gen data warehousing and business analytics
- Rapid pace of innovation
- Datasets whose size is larger than the traditional database to capture, store and analyse data.

# Characteristics of Big data

## Volume

- Data quantity

## Velocity

- Speed

## Variety

- Data Types

# Types of data

- Structured
- Semi – structure
- Unstructured

# Introduction to Hadoop and HDFS

- A **file system** can be thought of as an index or database containing the physical location of every piece of data on the hard drive or any other storage device.
- **Distributed file system** (DFS) is a method of storing and accessing **files** based on a client/server architecture. In a **distributed file system**, one or more central servers store **files** that can be accessed, with proper authorization rights, by any number of remote clients in the network.

# Introduction to Hadoop and HDFS

- Hadoop comes with a distributed file system called as HDFS.
- HDFS is designed to store
  - Very Large files
  - Streaming data access
- HDFS doesn't work for the following applications
  - Low-latency data access
  - Lot of small files

# HDFS blocks

- Every disk has a block size which is the minimum amount of data that it can read and write. File systems blocks are usually few kilobytes in size (512 bytes usually).
- Similar to file systems, files in HDFS are also broken into block-sized chunks, which are stored as independent units.
- HDFS too has blocks but they are much larger in size – 64 MB (Hadoop version 1.x) & 128 MB in version 2.x
- Each block is replicated 'n' times (default replication factor is 3; can be modified) for backup purposes.

# HDFS components – version 1.x

- NameNode
- Secondary NameNode
- DataNode



# Namenode

- HDFS has two types of nodes operating in master-slave pattern.
- Namenode is the master node that manages the file namespace. It also maintains the metadata for all the files and directories.
- NN also knows the datanodes on which all the blocks for a given file are located.
- Data nodes are the work horses and they report to the namenode with the list of blocks that they are storing.

Heartbeat Signal → NN

# Secondary Namenode

- HDFS also runs a secondary Namenode which doesn't act as a namenode by itself.
- NN is a SPOF. To avoid that secondary NN contains a copy of merged namespace image.
- Job Trackers and Task trackers are daemon process that help in job execution.
- Job tracker (v1.x) coordinates the job run.
- Tasktrackers (v1.x) are those that run the tasks which are essentially the splits of a job and assigned by the job tracker.

Namenode

Datanode1



Datanode2



Datanode3



Datanode4



Namenode

Datanode1



Datanode2



Datanode3

