SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601– Data Science and Analytics

III Year / VI Semester

DATA ANALYTICAL FRAMEWORKS

Topic 2: HDFS

HDFS

- The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner.

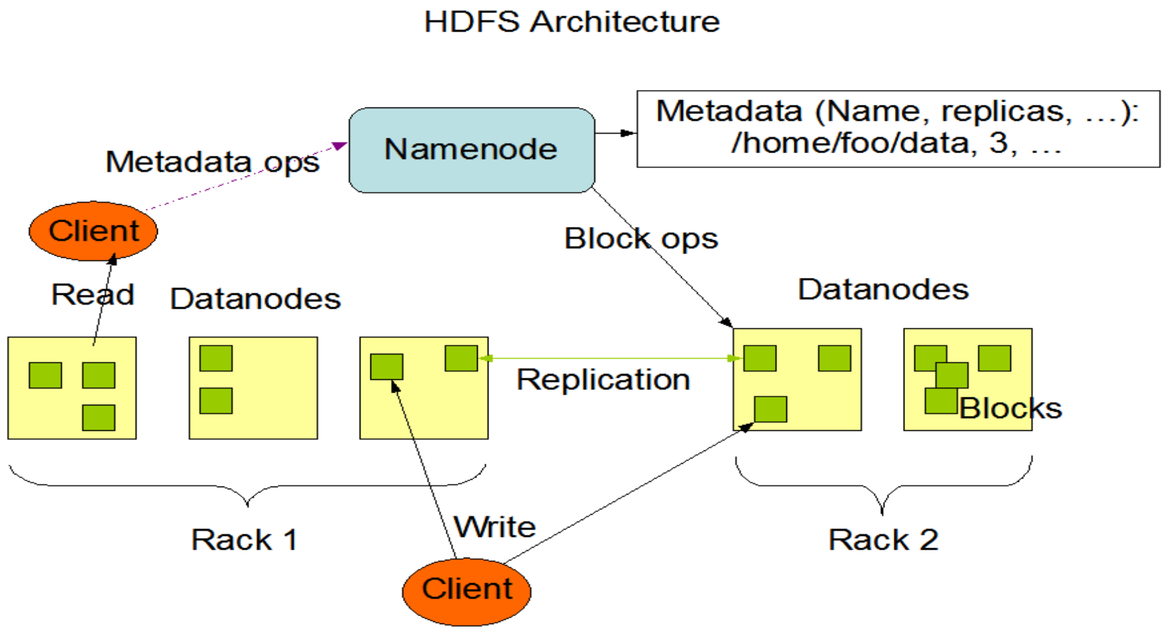
- • Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware.

- • HDFS holds very large amount of data and provides easier access. To store such huge data, the files split into blocks are stored across multiple machines.

- • These files are stored in redundant fashion to rescue the system from possible data losses in case of failure.

HDFS

- HDFS Architecture
- HDFS uses a master/slave architecture where master consists of a single Name Node that manages the file system metadata and one or more slave. Data Nodes that store the actual data.



HDFS

Name Node

- The namenode is the commodity hardware that contains the GNU/Linux operating system and the namenode software. It is a software that can be run on commodity hardware.

- The system having the namenode acts as the master server and it does the following tasks:

- • Manages the file system namespace.
- • Stores metadata for the files, like the directory structure of a typical File System.
- • Regulates client's access to files.
- • It also executes file system operations such as renaming, closing, and opening files and directories. It also determines the mapping of blocks to DataNodes.

HDFS

Data Node:

- The datanode is a commodity hardware having the GNU/Linux operating system and datanode software. These nodes manage the data storage of their system.

- A file in an HDFS namespace is split into several blocks and those blocks are stored in a set of Data Nodes.

- Data nodes store and retrieve blocks when they are requested by client or name node.

- They report back to name node periodically, with list of blocks that they are storing.

- The data node also perform operations such as block creation, deletion and replication as stated by the name node.

THANK YOU