Big Data Analytics Lab

Assignment 4

Garvin Chanderia

MITU20BTCS0100

2203132

LYISA 1

# AIM:

File Management in Hadoop

# THEORY:

Hadoop, an open-source distributed storage and processing framework, has its own unique file system known as the Hadoop Distributed File System (HDFS). HDFS is designed to handle large-scale data across clusters of commodity hardware. Efficient file management is crucial within Hadoop to ensure data reliability, availability, and scalability.

Key Concepts:

1. Blocks and Replication:

HDFS divides files into fixed-size blocks (default 128MB or 256MB). Each block is replicated across multiple datanodes for fault tolerance. Replication ensures data durability in case of node failures.

2. Namenode and Datanode:

The Namenode is the central metadata management server that tracks the file system hierarchy and block locations. Datanodes store actual data blocks and report their status to the Namenode.

3. Data Integrity:

HDFS ensures data integrity using checksums. Each block's checksum is computed and verified during read operations to detect and correct corruption.

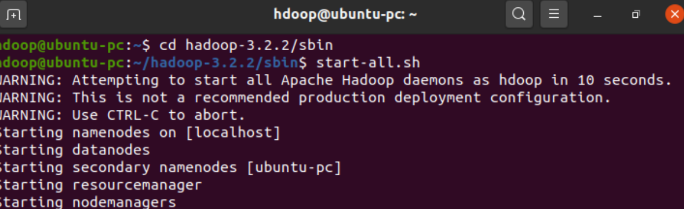
4. Rack Awareness:

HDFS maintains rack awareness to optimize data placement. It ensures that replicas are stored across different racks to enhance fault tolerance and reduce network traffic.

# CODE:

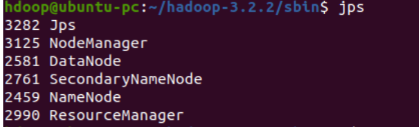
Step1: To use HDFS commands, start the Hadoop services using the following command:

sbin/start-all.sh



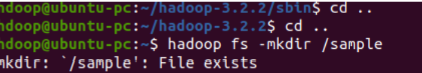
To check if Hadoop is up and running:

Jps



mkdir:

To create a directory, similar to Unix ls command.



cp:

Copy files from one directory to another within HDFS, similar to Unix cp command.



rm:

Remove a file from HDFS, similar to Unix rm command. This command does not delete directories. For

recursive delete, use command -rm -r.



Put:

Transfer and store a data file from local systems to the Hadoop file system using the put command.



Cat:

Initially, view the data from HDFS using cat command.

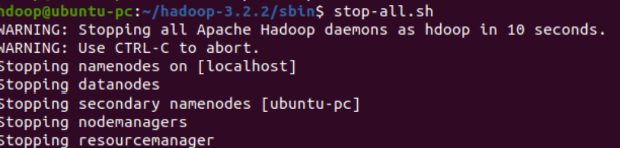


Get:

Gets the file from HDFS to the local file system using get command.



You can shut down the HDFS by using the following command.



# CONCLUSION:

File management in Hadoop, facilitated by HDFS, is a fundamental aspect of the framework's architecture. The division of data into blocks, data replication, and fault tolerance mechanisms ensure data reliability and availability in large-scale distributed environments. HDFS's features and design principles make it a robust choice for storing and managing data in Hadoop clusters