­­­­­­

**DSE-3264 & Big Data Analytics Laboratory Manual**

**What is HDFS??**

Hadoop comes with a distributed file system called HDFS. In HDFS data is distributed over several machines and replicated to ensure their durability to failure and high availability to parallel application. It is cost effective as it uses commodity hardware. It involves the concept of blocks, data nodes and node name.

**HDFS building Blocks:**

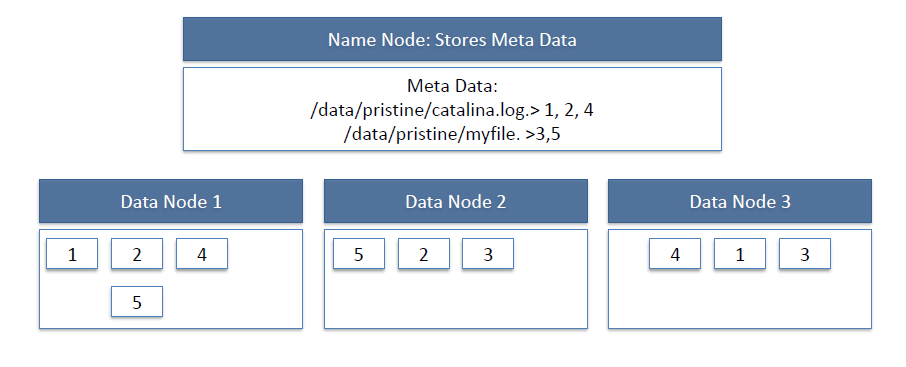
**Blocks:** A Block is the minimum amount of data that it can read or write.HDFS blocks are 128 MB by default, and this is configurable.Files in HDFS are broken into block-sized chunks, which are stored as independent units. Unlike a file system, if the file is in HDFS is smaller than block size, then it does not occupy full blocks size, i.e., 5 MB of file stored in HDFS of block size 128 MB takes 5MB of space only. The HDFS block size is large just to minimize the cost of seek.

**Name Node:** HDFS works in master-worker pattern where the name node acts as master. Name Node is controller and manager of HDFS as it knows the status and the metadata of all the files in HDFS; the metadata information being file permission, names, and location of each block. The metadata is small, so it is stored in the memory of the name node, allowing faster access to data. Moreover, the HDFS cluster is accessed by multiple clients concurrently, so all this information is handled by a single machine. The file system operations like opening, closing, renaming etc. are executed by it.

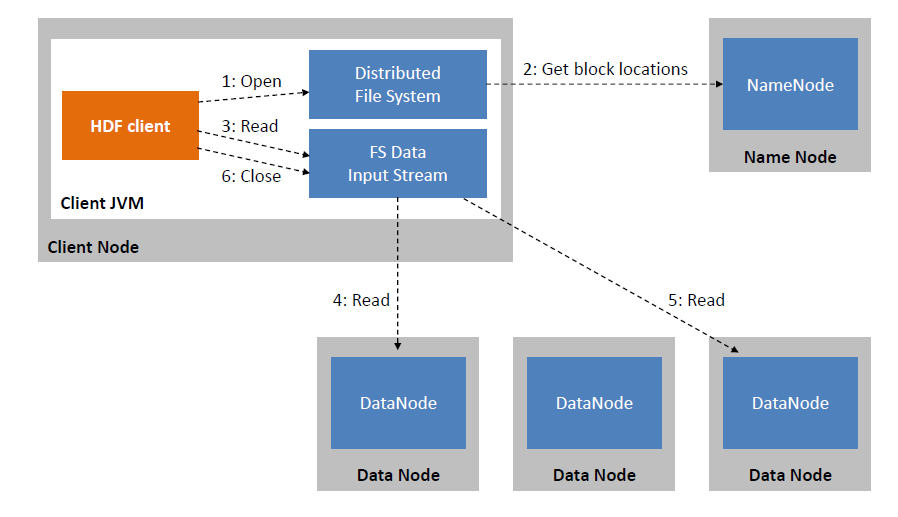
**Data Node:** They store and retrieve blocks when they are told to; by client or name node. They report back to the name node periodically, with a list of blocks that they are storing. The data node being a commodity hardware also does the work of block creation, deletion and replication as stated by the name node.

**Secondary Name Node:** It is a separate physical machine which acts as a helper of name node. It performs periodic check points. It communicates with the name node and takes snapshots of meta data which helps minimize downtime and loss of data.

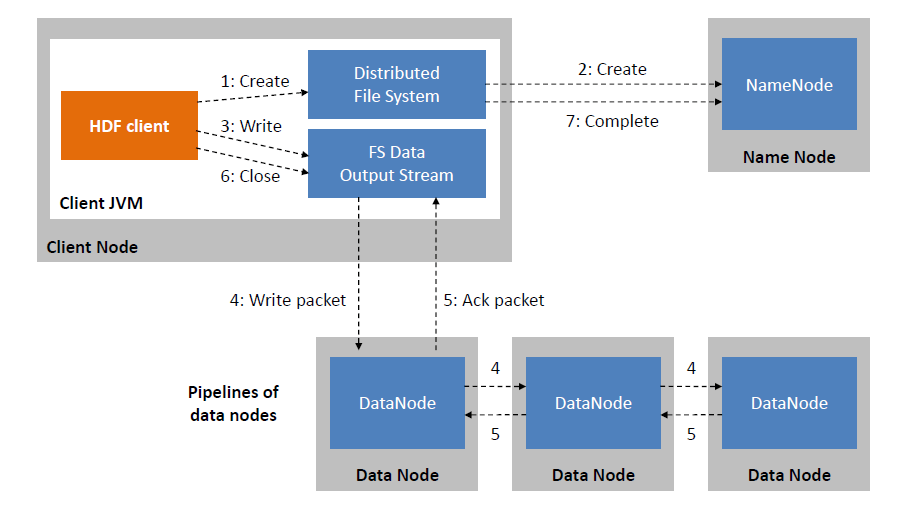
* **HDFS Data Node and Name Node Image:**



* **HDFS Read workflow**:



* **HDFS Write:**



* To use the HDFS commands, first you need to start the Hadoop services using the following command:

*$sbin/start-all.sh*

* To check the Hadoop services are up and running use the following command:

*$jps*

* Check the data node service is up by running ***jps*** commond at client side.



* To perform various file operations, use the prefix ***Hadoop fs/ hdfs dfs*** as a prefix for each command
* To create a directory. In Hadoop *dfs*there is no home directory by default. So, let us first create it using ***mkdir*** command.

**Week 1 Exercise: Hadoop Distributed File System**

1. List all Linux file operations and execute each one of them in Linux CLI.
2. Interact with HDFS using command line interface to understand the basic working structure of Hadoop cluster. Using Hadoop CLI, demonstrate the following commands to:

* Create a directory in HDFS.
* create an empty file
* copy files/folders from local file system to hdfs store.
* print file contents.
* copy files/folders from hdfs store to local file system.
* move file from local to hdfs
* copy files within hdfs
* move files within hdfs
* size of each file in directory
* total size of directory/file
* last modified time of directory or path
* change the replication factor of a file/directory in HDFS.
* List the contents of a directory in HDFS.
* Remove a file from HDFS.
* Change File Permissions
* Changing File Ownership
* Checksum Calculation
* File Concatenation
* File Compression/Decompression
* lsFile Block Location Information
* File Encryption/Decryption

1. Use web interface to monitor Name node manager, resource manager, and Data node status.