Chapter 2: Hadoop Fundamentals

→ Hadoop Environment Setup

1) System Requirements

- GNU/Linux operating system required
- Can use VirtualBox software for setup
- Java installation prerequisite

→ HDFS Operations and Commands

1) Starting and Stopping HDFS

- Format NameNode: (first time only)

```
$ hadoop namenode -format
```

- Start HDFS (starts NameNode & DataNod

```
$ start-dfs.sh
```

- Stop HDFS

```
$ stop-dfs.sh
```

2) Basic File Operations

↳ Listing / Navigation in HDFS

- List files/directories: $ hadoop fs -ls <path>
- List user home: $ hadoop fs -ls
- List root dir.: $ hadoop fs -ls /
- Disk usage: $ hadoop fs -du <path>

↳ Directory Management

- Create dir: $ hadoop fs -mkdir <path>
- Remove dir: $ hadoop fs -rm -r <path>

↳ Data Transfer

1) Local to HDFS

- Put File: $ hadoop fs -put <localSrc> <dest>

2) HDFS to Local

- Get File: $ hadoop fs -get <src> <localDest>

3) Within HDFS

- Copy files: $ hadoop fs -cp <src> <dest>
- Move files: $ hadoop fs -mv <src> <dest>

↳ File Content Operation

- Display File Content: $ hadoop fs -cat <filename>
- Verify File: $ hadoop fs -ls <file path>

→ Core Commands

- -ls: Lists dir. conlents with permissions owner, size, modification date
- -mkdir: Create directories
- -put: Copy from local to HDFS
- -get: Copy from HDFS to local sys
- -cp: Copy within HDFS
- -mv: move files/dir within HDFS
- -rm: remove files/dir
- -cat: Display file conlent
- -du: Show disk usage for files
- -stat: print file/dir information
- -help: return usage info of a specific command <cmd_name>

→ HDFS Data Flow

A) File Read Process

1. Client Request : HDFS client want to read a file
2. Ask NameNode about blocks location
3. and get them
3. Direct Connection : Client connects directly to DataNodes
4/5 Client reads data block from multiple DataNodes simultaneously
6. Close Connection

B) File Write Process

1. Client ask to create /write a new File
2. NameNode verifies if File exists and user has permission and it creates File entry in namespace
3. Client write data blocks in packets
4/5. DataNode Pipeline : Data Flows through a chain of DataNodes (replication)
6. Complete and close : NameNode updates metadata when writing is Finished