HDFS Commands

Cluster Architecture:

152
(2FMI)
③

↓ ↓
① ②

actual data
blocks

HDFS installed

Linux O.S [master] (name node)

[worker] Linux O.S

[worker] Linux O.S

files in Linux ≠ files in HDFS
(completely separate)
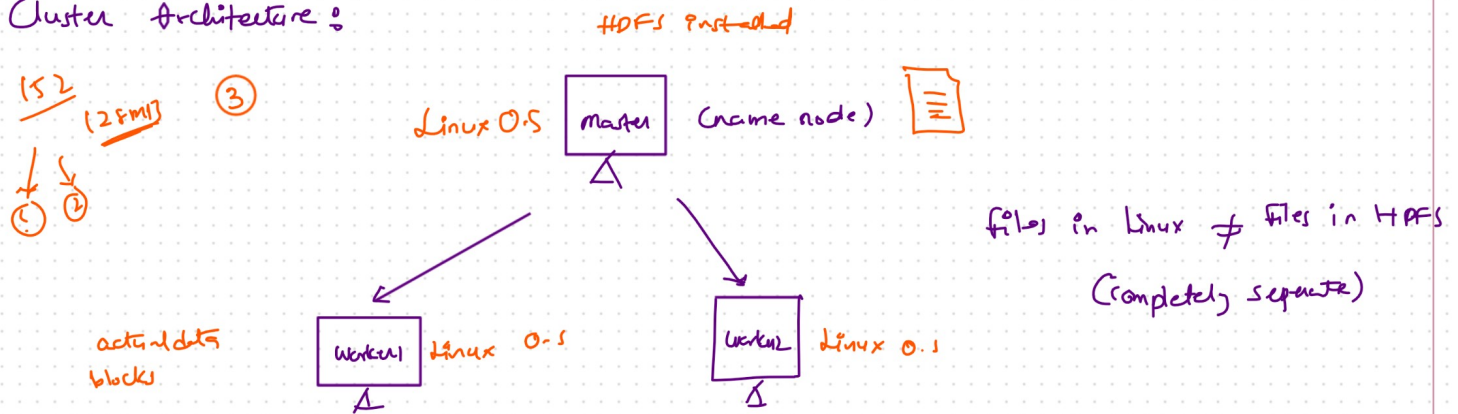
HDFS Command Syntax

2 - ways to execute HDFS commands

✓① hadoop fs [command]

② hdfs dfs [command]

Directory Structure:

Linux : /home/username

HDFS : /user/username

Essential HDFS Commands

< 1, Directory Operations

#List files & director

$ hadoop fs -ls

$ hadoop fs -ls /user/<directory>

Create directory

$ hadoop fs -mkdir /user/<path>/test-data
                              ↓

remove directory

$ hadoop fs -rmdir /usr/<path>/ directory_name

↑

test_data

② File Transfer Operations

copy from local to HDFS

$ hadoop fs -put localfile.txt /usr/destination_directory/

copy from HDFS to local

$ hadoop fs -get /usr/ destination_directory/ hdfs_file.txt /path/local_destination)

Copy within HDFS

$ hadoop fs -cp /source/path /destination/path

move within HDFS

$ hadoop fs -mv /source/path /destination/path

③ File Viewing Operations

View File Contents

$ hadoop fs -cat /usr/ directory-path / file.txt

view first few lines

$ hadoop fs -head /usr/ directory-path / file.txt

View (cat few lines

```
$ hadoop fs -tail /user/directory-path/file.txt
```

Limitation

No Change Directory (cd) Command:

Why: HDFS commands are executed from linux terminal, because we are accessing data nodes remotely from master node

Sol: Use HDFS web interface for browsing

Configuration files

cd (etc/hadoop/conf) [master node]

① hdfs-site.xml   ② core-site.xml

```
<property>
    <name>dfs.replication</name>
    <value>2</value>
    <description>
      Default block replication. The actual number of replications can be
      specified when the file is created. The default is used if replication
      is not specified in create time.
    </description>
</property>
```

```
<property>
    <name>fs.gs.block.size</name>
    <value>134217728</value>
    <final>false</final>
    <source>Dataproc Cluster Properties</source>
</property>
```

q: 241 MB file

Block 1 : 128 MB

Block 2 : 113 MB (remaining)

Replication: each Block 1 & Block 2 are stored in 2 nodes

Viewing Block Details

→ HDFS web interface

⇒ navigate : Utilities → Browse the File System

→ Click on file to see block information

Nodes

* always use absolute paths in HDFS commands

  : hadoop fs -ls /user/hdfs/directory)

* We terminal & commands instead of web UI for professional development

* practice commands manually : rather than using AI assistance