**HDFS Commands for Handling Unstructured Data Sets**

**Aim**

**To efficiently manage and process unstructured datasets (such as images, videos, logs, and raw text) using Hadoop Distributed File System (HDFS) commands.**

**1. Upload Unstructured Data to HDFS**

**Command:**

sh

CopyEdit

hdfs dfs -put sample\_video.mp4 /user/hadoop/videos/

**Description:** Uploads sample\_video.mp4 from the local filesystem to HDFS under /user/hadoop/videos/.

**Output:** (No output if successful)

**2. List Files in HDFS**

**Command:**

sh

CopyEdit

hdfs dfs -ls /user/hadoop/videos/

**Description:** Lists all files in the /user/hadoop/videos/ directory.

**Output:**

swift

CopyEdit

-rw-r--r-- 3 hadoop supergroup 10485760 2025-02-21 12:30 /user/hadoop/videos/sample\_video.mp4

**3. Read a Text-Based File (Logs, Raw Text)**

**Command:**

sh

CopyEdit

hdfs dfs -cat /user/hadoop/logs/logfile.txt

**Description:** Displays the content of a text-based unstructured file (like logs).

**Output:**

scss

CopyEdit

[INFO] Application started...

[ERROR] Null pointer exception at line 42...

**4. Get File Size**

**Command:**

sh

CopyEdit

hdfs dfs -du -h /user/hadoop/videos/sample\_video.mp4

**Description:** Shows the disk usage of the file in a human-readable format.

**Output:**

swift

CopyEdit

10.0 M /user/hadoop/videos/sample\_video.mp4

**5. Count Files and Total Size in Directory**

**Command:**

sh

CopyEdit

hdfs dfs -count -q /user/hadoop/videos/

**Description:** Displays the number of directories, files, and bytes in the directory.

**Output:**

swift

CopyEdit

1 5 524288000 /user/hadoop/videos/

**6. Remove an Unstructured File from HDFS**

**Command:**

sh

CopyEdit

hdfs dfs -rm /user/hadoop/videos/sample\_video.mp4

**Description:** Deletes the unstructured file from HDFS.

**Output:**

swift

CopyEdit

Deleted /user/hadoop/videos/sample\_video.mp4

**7. Move a File to Another HDFS Location**

**Command:**

sh

CopyEdit

hdfs dfs -mv /user/hadoop/videos/sample\_video.mp4 /user/hadoop/archive/

**Description:** Moves sample\_video.mp4 from /user/hadoop/videos/ to /user/hadoop/archive/.

**8. Copy a File Between HDFS Directories**

**Command:**

sh

CopyEdit

hdfs dfs -cp /user/hadoop/images/image1.png /user/hadoop/backup/

**Description:** Copies an unstructured file (image) to another directory.

**9. Merge Multiple Text-Based Unstructured Files**

**Command:**

sh

CopyEdit

hdfs dfs -getmerge /user/hadoop/logs/merged\_logs.txt

**Description:** Merges multiple log/text files into a single file and downloads it to the local filesystem.

**10. Running a MapReduce Job on Unstructured Data**

For **processing log files, images, or videos**, MapReduce can be used.

**Example: Extracting Keywords from Logs**

**Command:**

sh

CopyEdit

hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-\*.jar \

-input /user/hadoop/logs/ \

-output /user/hadoop/output/ \

-mapper /path/to/log\_mapper.py \

-reducer /path/to/log\_reducer.py

**Description:** Runs a MapReduce job to process log files stored in HDFS.