**EX :5 hdfs commands to handle semi unstructured data sets**

**AIM (Advanced Infrastructure Management) Program**

AIM programs automate HDFS operations, improving efficiency in handling large semi-structured/unstructured datasets.

**Description**

HDFS efficiently manages semi-structured (e.g., JSON, XML, CSV) and unstructured (e.g., images, videos, logs) data across distributed systems. By using tools like Hive, Pig, and Spark, organizations extract valuable insights from these datasets.

**1. Upload a File to HDFS**

**Command:**

sh

CopyEdit

hdfs dfs -put localfile.txt /user/hadoop/

**Description:** Uploads localfile.txt from the local filesystem to HDFS under /user/hadoop/.

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

**2. List Files in HDFS**

**Command:**

sh

CopyEdit

hdfs dfs -ls /user/hadoop/

**Description:** Lists files and directories in the given HDFS path.

**Output:**

bash

CopyEdit

-rw-r--r-- 3 hadoop supergroup 12345 2025-02-21 12:30 /user/hadoop/localfile.txt

**3. Read a File in HDFS**

**Command:**

sh

CopyEdit

hdfs dfs -cat /user/hadoop/localfile.txt

**Description:** Displays the contents of a file stored in HDFS.

**Output:**

css

CopyEdit

Sample semi-structured data content

**4. Remove a File from HDFS**

**Command:**

sh

CopyEdit

hdfs dfs -rm /user/hadoop/localfile.txt

**Description:** Deletes localfile.txt from HDFS.

**Output:**

bash

CopyEdit

Deleted /user/hadoop/localfile.txt

**5. Create a Directory in HDFS**

**Command:**

sh

CopyEdit

hdfs dfs -mkdir /user/hadoop/semi\_structured

**Description:** Creates a new directory in HDFS.

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

**6. Move a File in HDFS**

**Command:**

sh

CopyEdit

hdfs dfs -mv /user/hadoop/localfile.txt /user/hadoop/semi\_structured/

**Description:** Moves a file from one HDFS location to another.

**7. Count Number of Files, Directories, and Bytes**

**Command:**

sh

CopyEdit

hdfs dfs -count /user/hadoop/

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

**Output:**

bash

CopyEdit

3 10 12345678 /user/hadoop/

**8. Display File Summary**

**Command:**

sh

CopyEdit

hdfs dfs -du -h /user/hadoop/

**Description:** Displays disk usage summary for files in a directory.

**Output:**

swift

CopyEdit

12.0 K /user/hadoop/localfile.txt

8.0 M /user/hadoop/semi\_structured/

**9. Merge Multiple Files into One**

**Command:**

sh

CopyEdit

hdfs dfs -getmerge /user/hadoop/semi\_structured/ merged\_output.txt

**Description:** Merges multiple small files into a single file and copies it to the local filesystem.

**10. Run a MapReduce Job on Semi-Structured Data**

**Command:**

sh

CopyEdit

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

-input /user/hadoop/semi\_structured/ \

-output /user/hadoop/output/ \

-mapper /path/to/mapper.py \

-reducer /path/to/reducer.py

**Description:** Runs a MapReduce job using a Python script as a mapper and reducer to process semi-structured data.