PRACTICAL – 03

AIM : Implement the following file management tasks in Hadoop:-
> Adding files and directories
> Retrieving files from HDFS to local file system
> Deleting files from HDFS

Description:

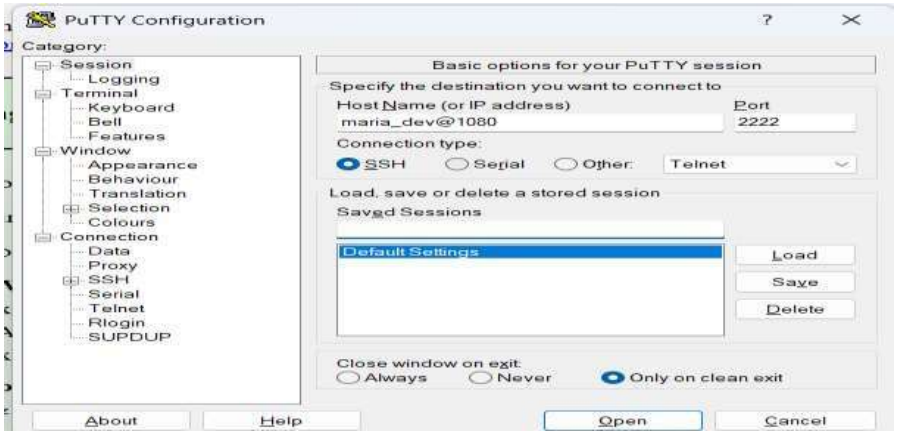
This practical shows how to manage files in Hadoop Distributed File System (HDFS).
Tasks performed include :

1. Creating directories and adding files using hdfs dfs -mkdir and hdfs dfs -put.
2. Retrieving files from HDFS to local system with hdfs dfs -get.
3. Deleting files or directories using hdfs dfs -rm and -rm -r.

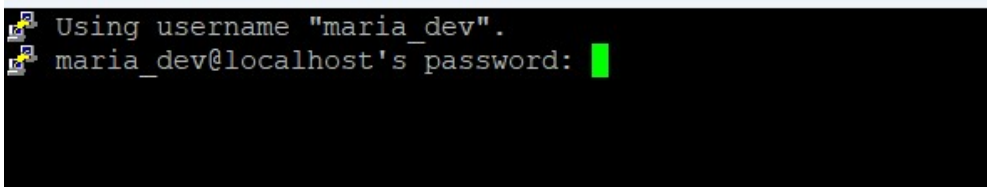
These commands demonstrate the basic file handling operations in a distributed storage system.

PROCEDURE :

- To give commands in HDFS download the platform putty it gets directly connected with the HDFS dashboard and from where you can give commands to add & delete the files
  Download Links - https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html
  After downloading open the file and give following details :
  Host name- maria_dev@1080
  Port- 2222
  Connection type- SSH
  Load server- HDP & Save



After saving you will get to see the command prompt where you have to enter the password which you have been set for your browser dashboard
Password- maria_dev



- To go in the Hadoop system give the command-

`hadoop fs -ls`

The command **hadoop fs -ls** is used to **list files and directories stored in Hadoop Distributed File System (HDFS)** or other supported file systems (like local FS, S3, etc., depending on configuration).
Shows the **files and directories** at the given path.

Displays **metadata**:

- File permissions
- Replication factor
- Owner & group
- File size (in bytes)
- Last modification date & time
- Path

```
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls
Found 1 items
drwxr-xr-x   - maria_dev hdfs          0 2025-08-18 16:29 hive
```

hadoop fs -mkdir

The **hadoop fs -mkdir** command is used to **create new directories in Hadoop Distributed File System (HDFS)** (or any other file system supported by Hadoop, like S3, local FS, etc., depending on your configuration)

⚡ Purpose

- To create a **new directory** in HDFS.

Suppose we will give the command for creating a directory for a movielens dataset
Command –

`hadoop fs -mkdir ml-100k`

```
[maria_dev@sandbox-hdp ~]$ hadoop fs -mkdir ml-100k
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - maria_dev hdfs          0 2025-08-18 16:29 hive
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 06:21 ml-100k
[maria_dev@sandbox-hdp ~]$
```

hadoop fs -ls

The **hadoop fs -ls** command is used to **list files and directories in Hadoop Distributed File System (HDFS)** or in any other file system supported by Hadoop (like local FS, S3, etc., depending on configuration)

⚡ Purpose

- To **view the contents** of a directory in HDFS.
- To **see metadata** of files/directories such as:
  - **Permissions** (read, write, execute)
  - **Replication factor** (for files in HDFS)
  - **Owner** and **Group**
  - **File size** (in bytes)
  - **Modification date & time**
  - **File/Directory name (path)**

```
[maria_dev@sandbox-hdp ~]$ hadoop fs -mkdir ml-100k
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - maria_dev hdfs          0 2025-08-18 16:29 hive
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 06:21 ml-100k
[maria_dev@sandbox-hdp ~]$
```

ls

In **Hadoop**, the ls command is used to **list files and directories** in the Hadoop Distributed File System (**HDFS**)—similar to the ls command in Linux, but it operates on HDFS paths instead of local file system paths.

Purpose:

- To display the list of files/directories in a given HDFS directory.
- To view metadata like **permissions, owner, group, file size, replication factor, modification date, and path**.

pwd

⚡ Purpose of pwd in Hadoop

- pwd stands for **Print Working Directory**.
- It shows the **current working directory in HDFS** where you are operating.
- Useful to confirm your present location before running file operations like ls, put, or get.

```
[maria_dev@sandbox-hdp ~]$ pwd
/home/maria_dev
```

- **wget** http://media.sundog-soft.com/hadoop/ml-100k/u.data
  The above command is used to copy the data from web server to the Hadoop file system.

```
[maria_dev@sandbox-hdp ~]$ wget http://media.sundog-soft.com/hadoop/ml-100k/u.da
ta
--2025-08-25 06:27:27--  http://media.sundog-soft.com/hadoop/ml-100k/u.data
Resolving media.sundog-soft.com (media.sundog-soft.com)... 52.216.219.105, 52.21
7.170.177, 16.15.177.80, ...
Connecting to media.sundog-soft.com (media.sundog-soft.com)|52.216.219.105|:80..
. connected.
HTTP request sent, awaiting response... 200 OK
Length: 2079229 (2.0M) [application/octet-stream]
Saving to: 'u.data'

100%[=======================================================================

2025-08-25 06:27:40 (26.2 MB/s) - 'u.data' saved [2079229/2079229]

[maria_dev@sandbox-hdp ~]$
```

ls

Give the command ls to see whether the data is imported in hdfs
Once it is imported you will see the name as u.data

```
[maria_dev@sandbox-hdp ~]$ ls
u.data
[maria_dev@sandbox-hdp ~]$
```

ls -la

- **Purpose of ls -la (Linux vs Hadoop)**

```
[maria_dev@sandbox-hdp ~]$ ls -la
total 2060
drwx------ 1 maria_dev maria_dev    4096 Aug 25 06:27 .
drwxr-xr-x 1 root      root         4096 Jun 18  2018 ..
-rw------- 1 maria_dev maria_dev      14 Aug 25 05:59 .bash_history
-rw-r--r-- 1 maria_dev maria_dev      18 Sep  6  2017 .bash_logout
-rw-r--r-- 1 maria_dev maria_dev     193 Sep  6  2017 .bash_profile
-rw-r--r-- 1 maria_dev maria_dev     619 Jun 18  2018 .bashrc
-rw-rw-r-- 1 maria_dev maria_dev 2079229 Nov 11  2016 u.data
[maria_dev@sandbox-hdp ~]$
```

- In **Linux**, ls -la lists **all files including hidden ones** (those starting with .), with detailed information (long format).

`hadoop fs  -copyFromLocal u.data ml-100k/u.data`
The file will get copied from local file system to the Hadoop named as u.data

`hadoop fs -ls`
The **hadoop fs -ls** command is used to **list files and directories in Hadoop Distributed File System (HDFS)** or in any other file system supported by Hadoop (like local FS, S3, etc., depending on configuration)

`hadoop fs -rm ml-100k/u.data`

⚡ Purpose

- To **remove (delete) files** from HDFS.
- Works similar to Linux rm, but operates on HDFS.

```
[maria_dev@sandbox-hdp ~]$ hadoop fs -copyFromLocal u.data ml-100k/u.data
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - maria_dev hdfs          0 2025-08-18 16:29 hive
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 06:30 ml-100k
```

`hadoop fs -rmdir ml-100k`
The **hadoop fs -rmdir** command is used to **remove (delete) empty directories from HDFS**.

⚡Purpose

- To delete **empty directories** in Hadoop Distributed File System (HDFS).
- It is similar to the Linux rmdir command.
- ı Unlike -rm -r, it **cannot delete directories that contain files or subdirectories**.

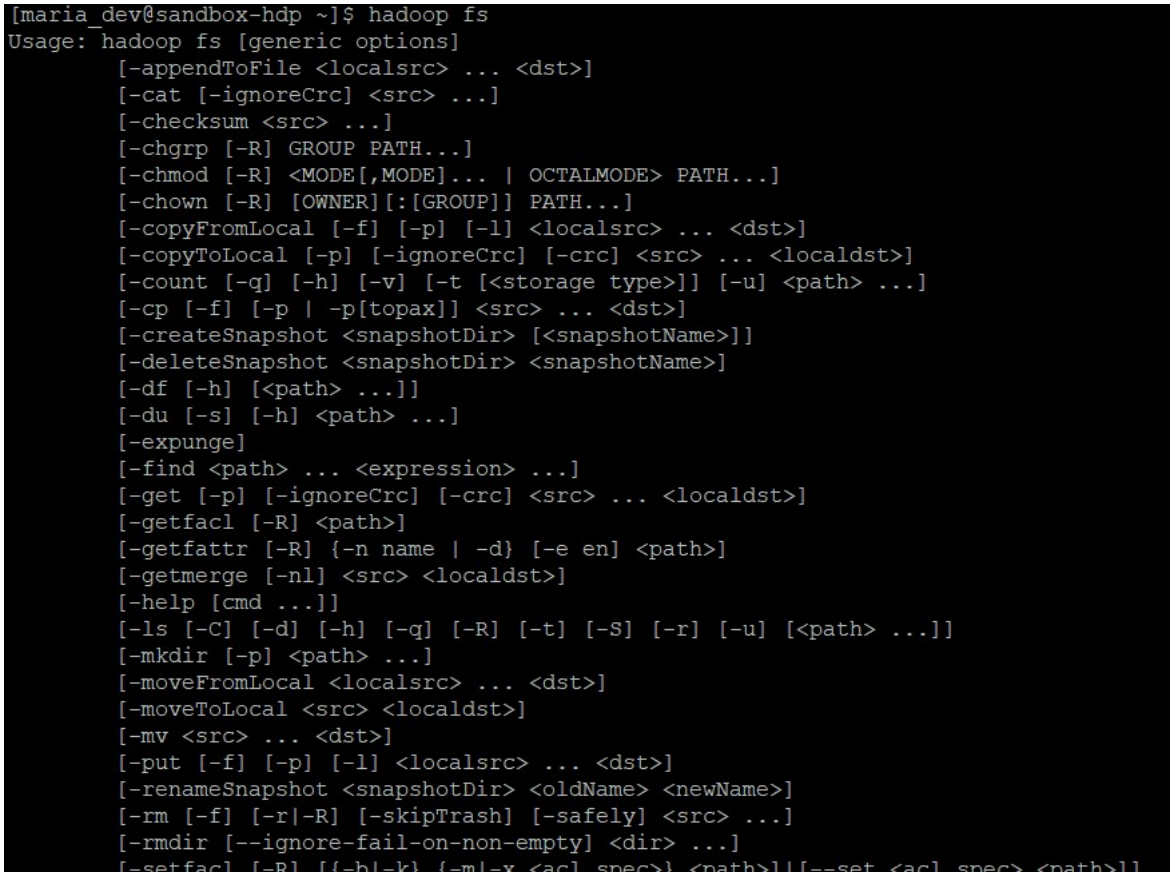
```
[maria_dev@sandbox-hdp ~]$ hadoop fs -rm ml-100k/u.data
25/08/25 06:31:31 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox-hdp.hortonworks.com:8020/user/maria_dev/ml-100k/u.da
dev/.Trash/Current/user/maria_dev/ml-100k/u.data
[maria_dev@sandbox-hdp ~]$
```

`hadoop fs -ls`
The commands checks where the directory is removed from the hadoop

```
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls
Found 2 items
drwx------   - maria_dev hdfs          0 2025-08-25 06:31 .Trash
drwxr-xr-x   - maria_dev hdfs          0 2025-08-18 16:29 hive
[maria_dev@sandbox-hdp ~]$
```

`hadoop fs`

By using this command we may see the activities that we have performed in our Hadoop file system



Conclusion:
Basic file management operations in HDFS were successfully performed, showing how to add, access, and delete data in Hadoop.