Sujan Subedi 11232939 5(C2)

PRACTICAL -03

Implement the following file management tasks in Hadoop:-

> Adding files and directories

> Retrieving files from HDFS to local file system

> Deleting files from HDFS

1)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

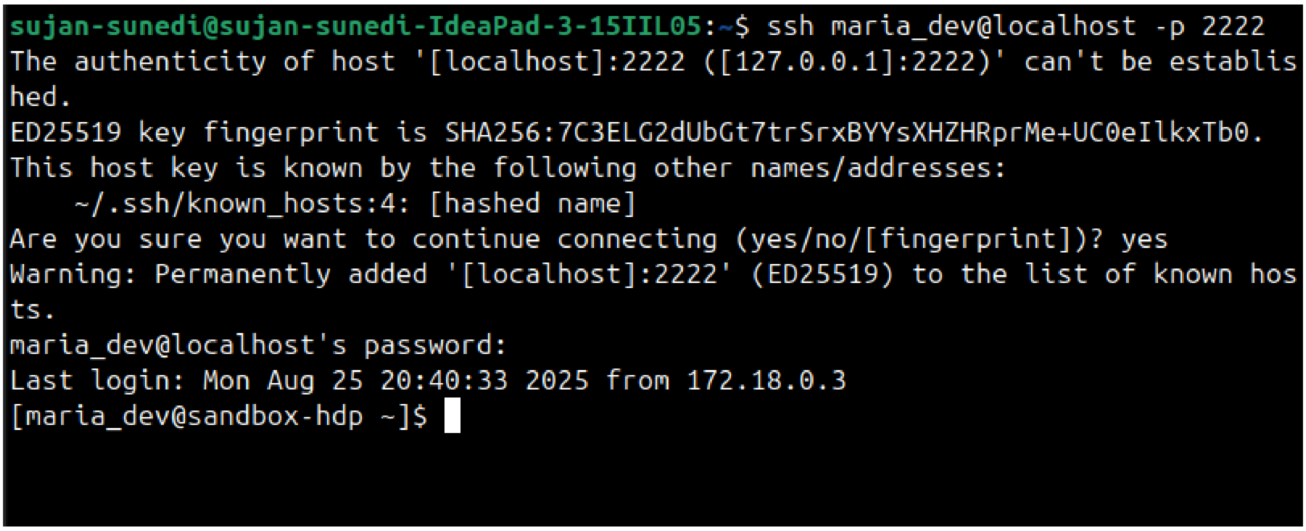
solution:
On **Ubuntu**, I used the built-in SSH tool to connect to the Hadoop Sandbox running inside VirtualBox.

The Sandbox SSH port mapping is as follows (based on the practical):

- **Host**: `localhost`
- **Port**: `2222` (mapped in VirtualBox from guest VM port 22)
- **Username**: `maria_dev`
- **Password**: `maria_dev`

So, instead of using PuTTY, I ran the following command in my Ubuntu terminal:

Command:- ssh maria_dev@localhost -p 2222



```
sujan-sunedi@sujan-sunedi-IdeaPad-3-15IIL05:~$ ssh maria_dev@localhost -p 2222
The authenticity of host '[localhost]:2222 ([127.0.0.1]:2222)' can't be establis
hed.
ED25519 key fingerprint is SHA256:7C3ELG2dUbGt7trSrxBYYsXHZHRprMe+UC0eIlkxTb0.
This host key is known by the following other names/addresses:
    ~/.ssh/known_hosts:4: [hashed name]
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '[localhost]:2222' (ED25519) to the list of known hos
ts.
maria_dev@localhost's password:
Last login: Mon Aug 25 20:40:33 2025 from 172.18.0.3
[maria_dev@sandbox-hdp ~]$ 
```

2)To go in the Hadoop system give the command-

***hadoop fs -ls**

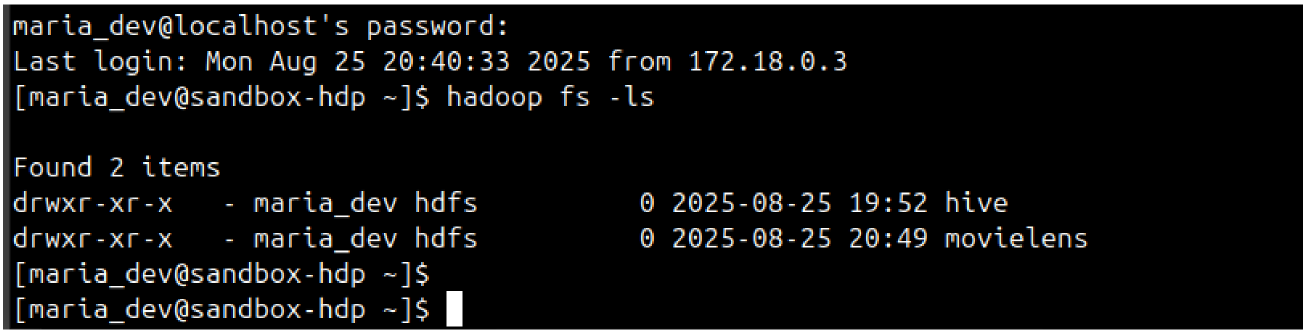
The command **hadoop fs -ls** is used to **list files and directories stored in Hadoop Distributed File System (HDFS)** or other supported file systems (like local FS, S3, etc., depending on configuration).

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

Output:



```
maria_dev@localhost's password:
Last login: Mon Aug 25 20:40:33 2025 from 172.18.0.3
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls

Found 2 items
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 19:52 hive
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 20:49 movielens
[maria_dev@sandbox-hdp ~]$
[maria_dev@sandbox-hdp ~]$
```

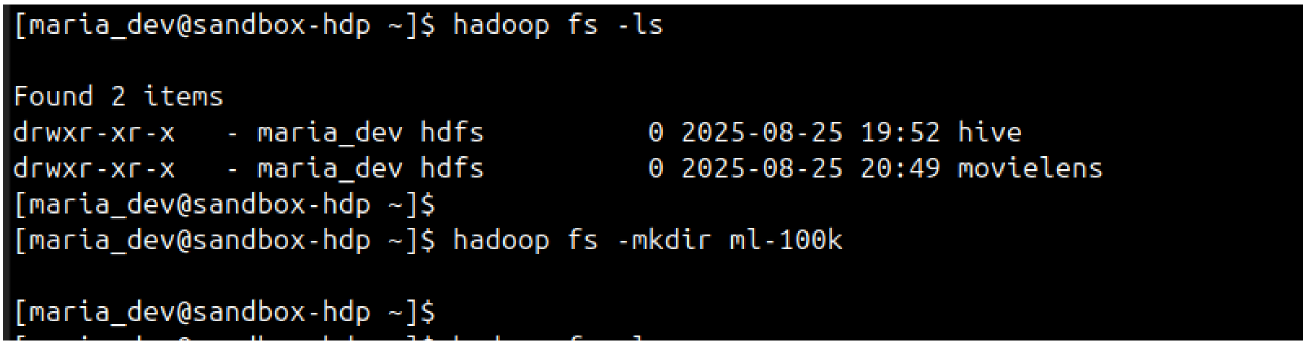
*Hadoop fs -mkdir

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



```
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls

Found 2 items
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 19:52 hive
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 20:49 movielens
[maria_dev@sandbox-hdp ~]$
[maria_dev@sandbox-hdp ~]$ hadoop fs -mkdir ml-100k

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

Command- hadoop fs -mkdir ml-100k.

*hadoop fs -ls

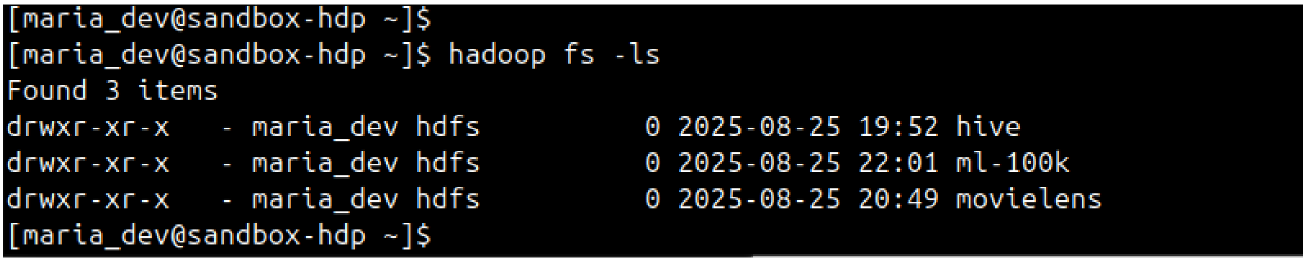
The **hadoop fs -ls** command is used to **list files and directories in Hadoop Distributed File System (HDFS)** or in any other file system supported by Hadoop (like local FS, S3, etc., depending on configuration)

📌 **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)**

Output:



```
[maria_dev@sandbox-hdp ~]$
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls
Found 3 items
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 19:52 hive
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 22:01 ml-100k
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 20:49 movielens
[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:

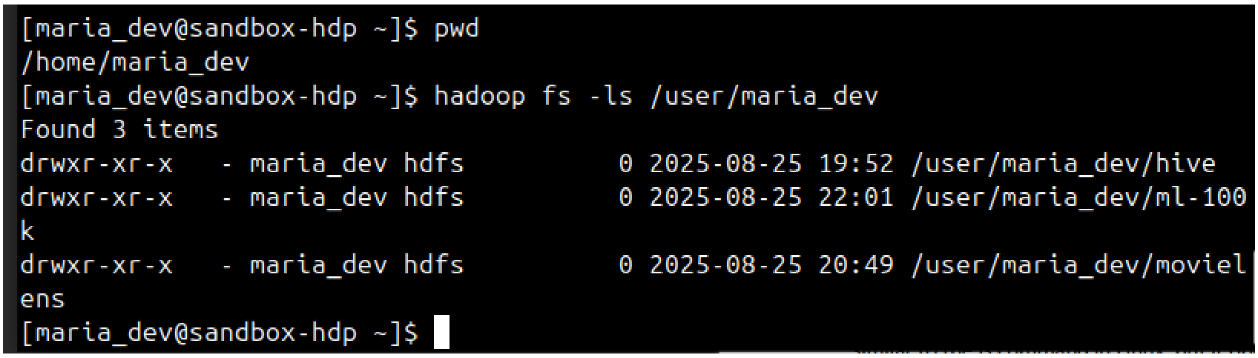
- To display the list of files/directories in a given HDFS directory.
- To view metadata like **permissions, owner, group, file size, replication factor, modification date, and path**.

***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.

Output:



```
[maria_dev@sandbox-hdp ~]$ pwd
/home/maria_dev
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls /user/maria_dev
Found 3 items
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 19:52 /user/maria_dev/hive
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 22:01 /user/maria_dev/ml-100
k
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 20:49 /user/maria_dev/moviel
ens
[maria_dev@sandbox-hdp ~]$ 
```

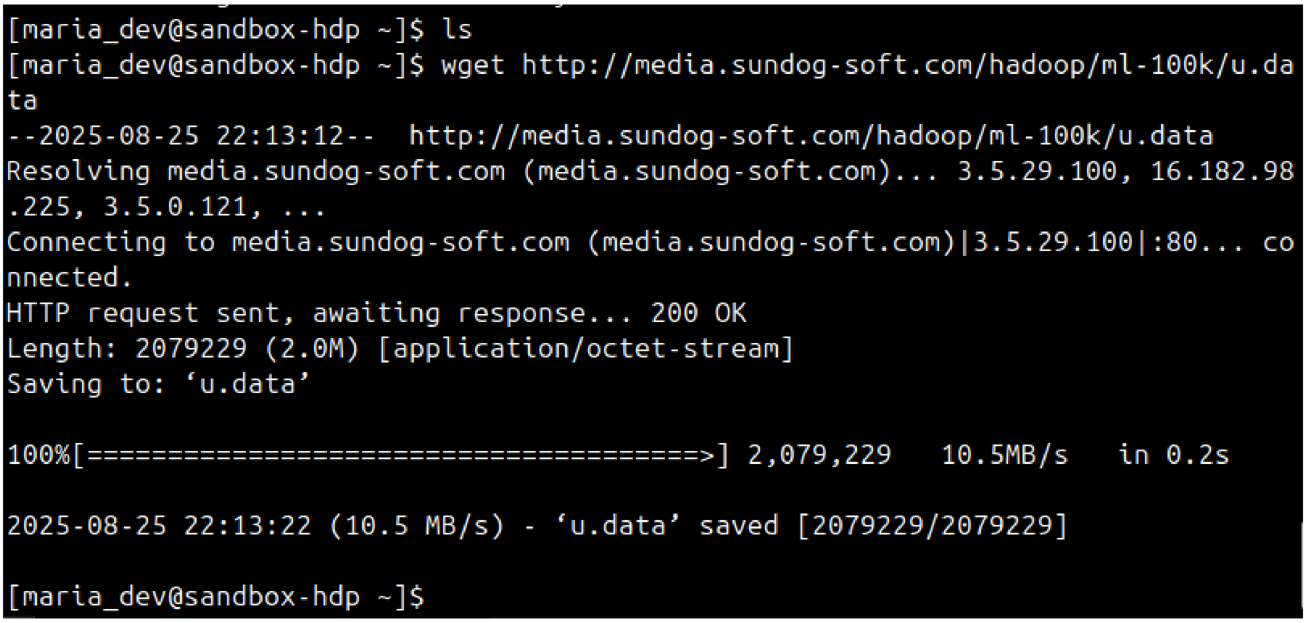
***ls**

Command to display the directory

***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

Output:



```
[maria_dev@sandbox-hdp ~]$ ls
[maria_dev@sandbox-hdp ~]$ wget http://media.sundog-soft.com/hadoop/ml-100k/u.da
ta
--2025-08-25 22:13:12--  http://media.sundog-soft.com/hadoop/ml-100k/u.data
Resolving media.sundog-soft.com (media.sundog-soft.com)... 3.5.29.100, 16.182.98
.225, 3.5.0.121, ...
Connecting to media.sundog-soft.com (media.sundog-soft.com)|3.5.29.100|:80... co
nnected.
HTTP request sent, awaiting response... 200 OK
Length: 2079229 (2.0M) [application/octet-stream]
Saving to: ‘u.data’

100%[======================================>] 2,079,229   10.5MB/s   in 0.2s

2025-08-25 22:13:22 (10.5 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

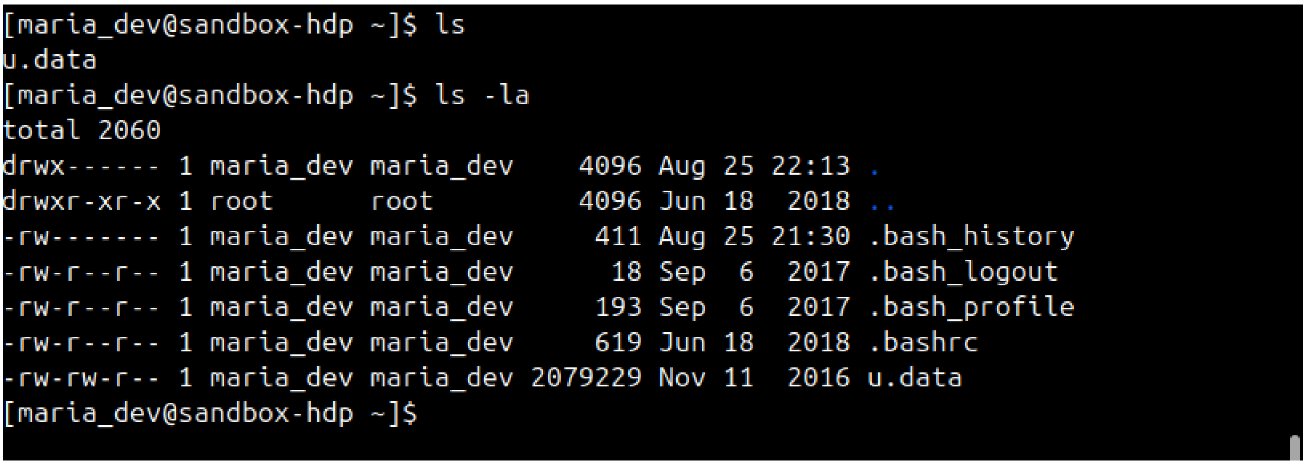
Once it is imported you will see the name as u.data

***ls -la**

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

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

Output:-

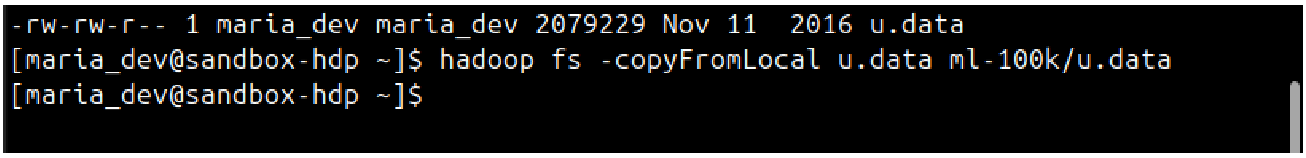


```
[maria_dev@sandbox-hdp ~]$ ls
u.data
[maria_dev@sandbox-hdp ~]$ ls -la
total 2060
drwx------ 1 maria_dev maria_dev    4096 Aug 25 22:13 .
drwxr-xr-x 1 root      root         4096 Jun 18  2018 ..
-rw------- 1 maria_dev maria_dev     411 Aug 25 21:30 .bash_history
-rw-r--r-- 1 maria_dev maria_dev      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 ~]$
```

*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

Output:-

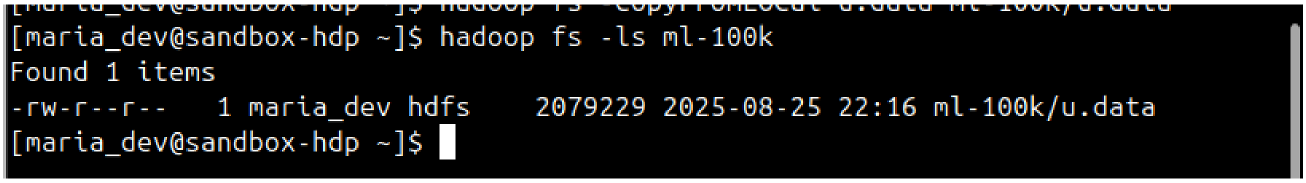


```
-rw-rw-r-- 1 maria_dev maria_dev 2079229 Nov 11  2016 u.data
[maria_dev@sandbox-hdp ~]$ hadoop fs -copyFromLocal u.data ml-100k/u.data
[maria_dev@sandbox-hdp ~]$
```

*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).

Output:-



```
[maria_dev@sandbox-hdp ~]$ hadoop fs -copyFromLocal u.data ml-100k/u.data
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls ml-100k
Found 1 items
-rw-r--r--   1 maria_dev hdfs    2079229 2025-08-25 22:16 ml-100k/u.data
[maria_dev@sandbox-hdp ~]$
```

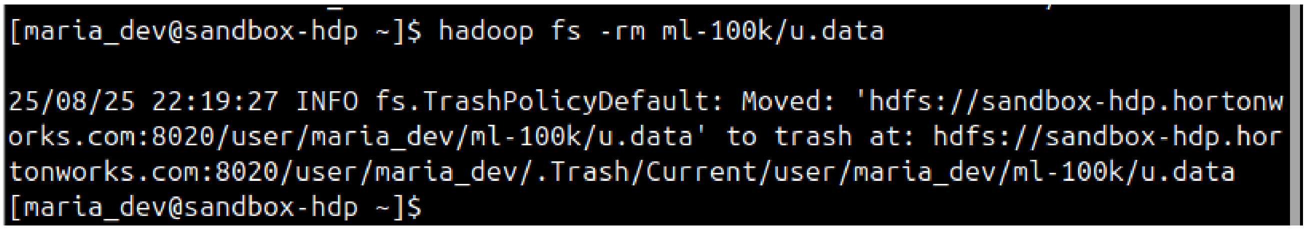
*hadoop fs -rm ml-100k/u.data

📌 **Purpose**

- To **remove (delete) files** from HDFS.

- Works similar to Linux rm, but operates on HDFS.

Output:-



```
[maria_dev@sandbox-hdp ~]$ hadoop fs -rm ml-100k/u.data

25/08/25 22:19:27 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox-hdp.hortonw
orks.com:8020/user/maria_dev/ml-100k/u.data' to trash at: hdfs://sandbox-hdp.hor
tonworks.com:8020/user/maria_dev/.Trash/Current/user/maria_dev/ml-100k/u.data
[maria_dev@sandbox-hdp ~]$
```

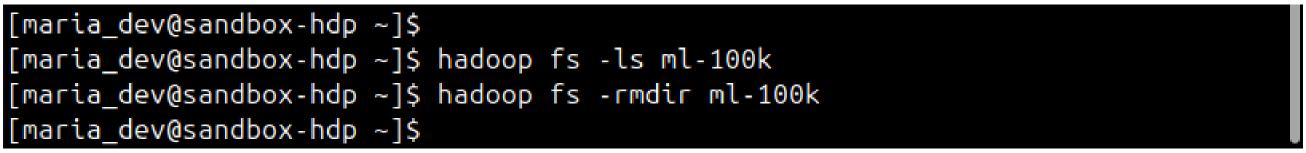
*hadoop fs -rmdir ml-100k

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.
- ⚠ Unlike -rm -r, it **cannot delete directories that contain files or subdirectories**.

Output:-

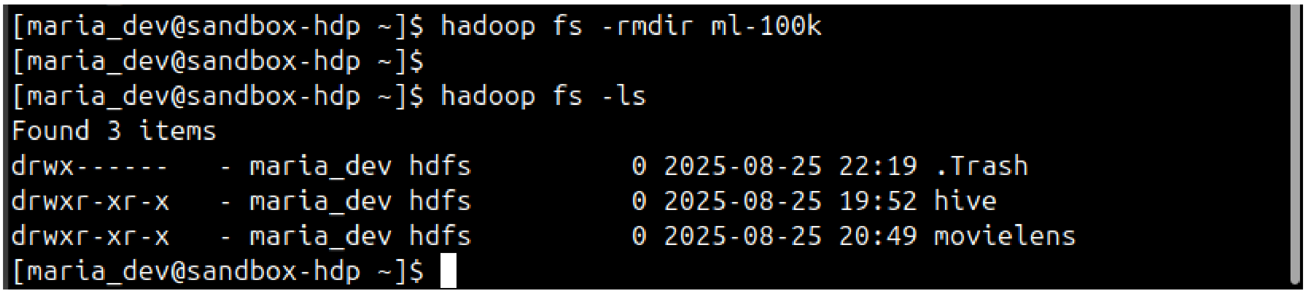


```
[maria_dev@sandbox-hdp ~]$
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls ml-100k
[maria_dev@sandbox-hdp ~]$ hadoop fs -rmdir ml-100k
[maria_dev@sandbox-hdp ~]$
```

*hadoop fs -ls

The commands checks where the directory is removed from the hadoop

Output:-



```
[maria_dev@sandbox-hdp ~]$ hadoop fs -rmdir ml-100k
[maria_dev@sandbox-hdp ~]$
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls
Found 3 items
drwx------   - maria_dev hdfs          0 2025-08-25 22:19 .Trash
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 19:52 hive
drwxr-xr-x   - maria_dev hdfs          0 2025-08-25 20:49 movielens
[maria_dev@sandbox-hdp ~]$
```

*Hadoop fs

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