Hadoop Commands

Command 1: hdfs dfs -ls /

Description: The `hdfs dfs ls /` command is used to list the contents of the root directory (`/`) in Hadoop Distributed File System (HDFS). Here's a detailed explanation along with use cases:

Command Breakdown

- `hdfs`: This is the Hadoop commandline utility for interacting with HDFS.
- `dfs`: This specifies that you're using the HDFS command set.
- `ls`: This option lists the contents of a directory, similar to the `ls` command in Unix/Linux.
- `/`: This is the path to the directory you want to list. In this case, `/` refers to the root directory of the HDFS file system.

Functionality

Lists Directory Contents: The command shows all files and subdirectories within the specified path (in this case, the root directory of HDFS). It provides details such as file permissions, owner, group, size, and modification timestamp.

Use Cases

1. Viewing Root Directory Contents:

- **Scenario:** After setting up HDFS or performing maintenance, you want to verify the structure and contents at the root level.
- **Command:** `hdfs dfs -ls /`
- **Output:** Lists directories and files located directly under the root (`/`), such as `/user`, `/tmp`, etc.

2. Checking for Specific Directories:

- **Scenario:** You want to confirm whether a specific directory (e.g., `/user/hadoop`) exists before performing operations on it.
- **Command:** `hdfs dfs -ls /user`
- **Output:** Shows if `/user` contains the `hadoop` directory and other files or directories.

3. Monitoring Directory Contents:

- **Scenario:** You need to check the current state of directories to monitor space usage or check for recent changes.
- **Command:** `hdfs dfs -ls /`
- **Output:** Provides a snapshot of the directory structure and metadata at that moment.

4. Troubleshooting:

- **Scenario:** You encounter issues with HDFS operations and need to verify if the directory structure is as expected.
- **Command:** `hdfs dfs -ls /`

- **Output:** Helps identify missing directories or discrepancies in the file system.

Example Output:

The **Output** from the command `hdfs dfs -ls /` provides a snapshot of the contents within the root directory of Hadoop Distributed File System (HDFS). Here's a detailed explanation of the **Output**:

```plaintext
Found 1 items
drwxrxrx   rawat supergroup       0 20240901 18:54 /test_directory
```

1. `Found 1 items`:

This indicates that there is one item (directory or file) present in the root directory (`/`) of HDFS.

2. `drwxrxrx`:

- `d`: This signifies that the item is a directory. In HDFS, directories and files are listed with a leading `d` for directories and a `` for files.
- `rwxrxrx`: These are the permissions for the directory.
- `rwx`: The owner (`rawat`) has read, write, and execute permissions.
- `rx`: The group (`supergroup`) has read and execute permissions.
- `rx`: Others (everyone else) also have read and execute permissions.

3. ``:

This represents the replication factor of the directory's contents. In this context, it's not applicable as it is a directory.

4. `rawat`:

This is the owner of the directory. The user `rawat` has ownership and full permissions over the `/test_directory`.

5. `supergroup`:

This is the group associated with the directory. The group `supergroup` has permissions to access and execute files within the directory but does not have write permissions.

6. `0`:

This represents the size of the directory. For directories, the size is typically shown as `0` because directories themselves do not have a size; they contain files and other directories.

7. `20240901 18:54`:

This is the timestamp indicating when the directory was last modified or created. In this case, the directory was created or last modified on September 1, 2024, at 18:54 (6:54 PM).

8. `/test_directory`:

This is the path of the item within HDFS. It shows that `/test_directory` is a directory located directly under the root directory (`/`).

Summary

The command `hdfs dfs -ls /` has revealed that the root directory of HDFS contains a single directory named `/test_directory`. This directory is owned by `rawat`, is part of the `supergroup` group, and has standard directory permissions set. It was created or last modified on September 1, 2024.

Command 2: `hdfs dfs -put <localsourcepath> <hdfsdestinationpath>`

Description: The `hdfs dfs put` command is used to copy files or directories from the local filesystem to Hadoop Distributed File System (HDFS). Below is a detailed explanation along with use cases:

Command Breakdown:

- `hdfs`: This is the Hadoop command line utility for interacting with HDFS.
- `dfs`: This specifies that you're using the HDFS command set.
- `put`: This option is used to upload files or directories from the local filesystem to HDFS.
- `<localsourcepath>`: This is the path to the file or directory on your local filesystem that you want to copy.
- `<hdfsdestinationpath>`: This is the path in HDFS where you want to copy the file or directory.

Functionality:

Uploads Files or Directories: The command transfers files or directories from the local system to the specified destination in HDFS. It ensures that local files are available in the distributed file system, enabling their use in Hadoop jobs and applications.

Use Cases:

1. Uploading Individual Files:

- **Scenario**: You have a file named `example.txt` on your local system that you want to make available in HDFS.
- Command: `hdfs dfs -put C:\localfolder\example.txt /test_directory/`
- **Output**: Copies `example.txt` from `C:\localfolder\` on your local system to the `/test_directory/` in HDFS.

2. Uploading Directories:

- **Scenario**: You need to upload an entire directory from your local filesystem to HDFS, including all files and subdirectories.
- **Command:** `hdfs dfs -put C:\localfolder\ /test_directory/`
- **Output**: Copies all contents of `C:\localfolder\` to `/test_directory/` in HDFS.

3. Overwriting Existing Files:

- **Scenario**: You want to ensure that a file in HDFS is updated with a new version from your local system.
- **Command:** `hdfs dfs -put f C:\localfolder\example.txt /test_directory/`
- **Output**: Copies `example.txt` to `/test_directory/` in HDFS, overwriting any existing file with the same name.

4. Ensuring File Availability:

- **Scenario**: You need to ensure that a file is available in HDFS for processing by a Hadoop job or application.
- **Command:** `hdfs dfs -put C:\localfolder\example.txt /test_directory/`
- **Output**: Makes `example.txt` available in HDFS at `/test_directory/`, allowing it to be accessed by Hadoop processes.

Example Command Usage:

To copy a file named `example.txt` from your local directory `C:\localfolder\` to the HDFS directory `/test_directory/`, you would use:

```bash
hdfs dfs -put C:\localfolder\example.txt /test_directory/
```

Explanation:

 `C:\localfolder\example.txt`: Path to the source file on your local machine.

 `/test_directory/`: Destination path in HDFS where the file will be uploaded.

Summary:

The `hdfs dfs -put` command facilitates the transfer of files and directories from the local filesystem to HDFS. It is essential for making local data accessible in a distributed environment, enabling data processing and analysis within the Hadoop ecosystem.

Note: In place of put we can also use CopyFromLocal.

Command 3: `hdfs dfs -copyToLocal /path/to/hdfs/directory /local/path/directory`

Description: It copies file from Hadoop directory to local directory.

Note: In place of CopyToLocal we can also use get.

Command 4: hdfs dfs -cat /path/to/hdfs/directory /local/path/directory | wc -l

Description: This command is used to count the number of lines in a file located in HDFS and output the result.