Big Data (HDFS) Assignment

Problem Statement: Using Command line of HDFS, perform following tasks.

**a) Create a directory /hadoop/hdfs/ in HDFS**

--create parent directory as -p

hdfs dfs -mkdir -p /hadoop/hdfs

--verify

hdfs dfs -test -d /hadoop/hdfs

echo $?

--view created directory: hdfs show under hadoop

hdfs dfs -ls /hadoop

**b) Create a temp directory in Hadoop. Run HDFS command to delete “temp” directory.**

--create directory

hdfs dfs -mkdir /hadoop/temp

--The -r option ensures that if the directory contains files or subdirectories, they will be deleted as well.

hdfs dfs -rm -r /temp

**c)List all the files/directories for the given hdfs destination path.**

--list all files and directories

--If you omit the path, it will list the contents of the current working directory in HDFS.

hdfs dfs -ls /hadoop/hdfs/

**d)Command that will list the directories(not files) in /hadoop folder.**

hdfs dfs -ls -d /hadoop/\*/

-ls: Lists the status of the specified files and directories.

-d: Restricts the listing to directories only (the wildcard \*/ ensures only directories are matched).

/hadoop/\*/: The path pattern to match all directories under /hadoop.

This command will list all directories inside the /hadoop directory on HDFS.

**e) Command to list recursively all files in hadoop directory and all subdirectories in hadoop directory**

hdfs dfs -ls -R /hadoop/

-R: Recursively lists all files and directories within the specified directory and its subdirectories.

/hadoop: The directory for which you want to list all files and subdirectories.

This command will output a recursive listing of all files and directories within the /hadoop directory and any subdirectories.

**f) List all the directory inside /hadoop/hdfs/ directory which starts with 'dir'.**

hdfs dfs -ls -d /hadoop/hdfs/dir\*/

note: list all directories within /hadoop/hdfs/ whose names start with "dir".

**g) Create a temp.txt file. Copies this file from local file system to HDFS**

--create a simple text file using the echo

echo "This is a temporary file." > temp.txt

-- just want an empty file:

touch temp.txt

--Copy the temp.txt File(from the local file system) to HDFS

hdfs dfs -put temp.txt /hadoop/hdfs/

--verify the File Copy

hdfs dfs -ls /hadoop/hdfs/

or

hdfs dfs -copyFromLocal temp.txt /hadoop/hdfs

**h) Copies the file from HDFS to local file system.**

hdfs dfs -get /hadoop/hdfs/temp.txt /home/user/

hdfs dfs -get: Copies files from HDFS to the local file system.

/user/hadoop/temp.txt: The path to the file in HDFS that you want to copy.

/home/user/: The local directory where you want to copy the file.

or

hdfs dfs -copyToLocal /hadoop/hdfs /home/local

**i)Command to copy from local directory with the source being restricted to a local file reference.**

--This copies temp.txt from HDFS to the /home/user/ directory on your local system.

hdfs dfs -get /hadoop/hdfs/temp.txt /home/user/

or

hdfs dfs -copyFromLocal /local/path/to/file /hdfs/directory/

**j) Command to copies to local directory with the source being restricted to a local file reference.**

--Assume you have a local file located at /home/user/temp.txt, and you want to copy it to the /user/hadoop/ directory in HDFS. The command would b

hdfs dfs -put /home/user/temp.txt /hadoop/hdfs/

/home/user/temp.txt: This is the path to the local file you want to copy.

/user/hadoop/: This is the destination directory in HDFS where you want the file to be copied.

Ensure that the local file path is correct and that you have the necessary permissions to write to the HDFS destination directory.

or

hdfs dfs -get /hdfs/path/to/file /local/directory/

**k) Command to move from local directory source to Hadoop directory.**

hdfs dfs -moveFromLocal /home/user/temp.txt /hadoop/hdfs/

l)Deletes the directory and any content under it recursively.

hdfs dfs -rm -r /hadoop/hdfs/temp

hdfs dfs -rm: Command to remove files or directories.

-r: Option to remove directories and their contents recursively.

/user/hadoop/temp: The path to the directory you want to delete in HDFS.

m)List the files and show Format file sizes in a human-readable fashion.

hdfs dfs -du -h /user/hadoop/

hdfs dfs -du: Displays the sizes of files and directories.

-h: Human-readable format (e.g., KB, MB, GB).

/user/hadoop/: The path to the directory or file you want to list.

hdfs dfs -du -h -R /user/hadoop/

Listing All Files in a Directory: If you want to list all files and directories recursively with their sizes in a human-readable format,

you can combine -du -h with the -R option:

or

hdfs dfs -ls -h /hadoop/

**n) Take a source file and outputs the file in text format on the terminal.**

hdfs dfs -cat /hadoop/hdfs/example.txt

Note:

The cat command reads the content of a file and outputs it to the standard output (usually the terminal in plain text format.).

**o) Display the content of the HDFS file test on your /user/hadoop2 directory.**

hdfs dfs -cat /user/hadoop2/test

Note:

--This command will display the contents of the test file in the /user/hadoop2 directory on your HDFS.

**p) Append the content of a local file test1 to a hdfs file test2.**

hdfs dfs -appendToFile /home/local/test1 /hadoop/hdfs/test2

Note:

This command allows you to append data from a local file(test1) to an existing file(test2) in HDFS.

**q) Show the capacity, free and used space of the filesystem**

hdfs dfsadmin -report

--This command provides detailed information about the HDFS cluster, including the total, used, and remaining space.

--This command will output a report that includes:

Configured Capacity: The total storage capacity of the HDFS cluster.

Present Capacity: The actual capacity available after considering factors like reserved space.

DFS Used: The amount of space currently used by HDFS files.

DFS Remaining: The free space available in the HDFS cluster.

DFS Used%: The percentage of the filesystem that is used.

This information helps you monitor the storage utilization of your HDFS cluster.

or

filesytem Size Used Available Use%

hdfs://quickstart.cloudera:8020 54.5 G 837.1 M 41.8 G 1%

**r) Shows the capacity, free and used space of the filesystem. Add parameter Formats the sizes of files in a human-readable fashion.**

hdfs dfs -df -h

Note:

This command will display:

Filesystem: The HDFS URI.

Size: The total capacity of the filesystem.

Used: The amount of space currently used.

Available: The free space available.

Use%: The percentage of the filesystem that is used.

The -h option formats the sizes in a human-readable format, showing them in KB, MB, GB, etc., making it easier to interpret the data.

**s) Show the amount of space, in bytes, used by the files that match the specified file pattern.**

hdfs dfs -du -s /user/hadoop2/test\*

Note:

--To show the amount of space, in bytes, used by the files that match a specified file pattern in HDFS

-du: Shows the disk usage.

-s: Displays a summary of disk usage for each directory or file that matches the pattern.

**t) Show the amount of space, in bytes, used by the files that match the specified file pattern. Formats the sizes of files in a human-readable fashion.**

hdfs dfs -du -h /user/hadoop2/test\*

--this command

Size: The amount of space used by each file or directory, formatted in a human-readable way (e.g., KB, MB, GB).

Location: The HDFS path of each file or directory.

**u) Check the health of the Hadoop file system.**

hdfs fsck /

Note:

--This command checks the HDFS for inconsistencies and reports the overall health of the filesystem.

--to check the health of the Hadoop file system

--This command checks the entire Hadoop file system (starting from the root directory /) and reports on issues such as missing blocks,

over-replicated blocks, under-replicated blocks, and corrupted files.

hdfs fsck / -files -blocks -locations

-files: Lists all files in the report.

-blocks: Lists all blocks in the report.

-locations: Lists block locations.

This will give you a comprehensive view of the health of your HDFS.

**v) Command to turn off the safemode of Name Node.**

hdfs dfsadmin -safemode leave

Note:

--This command exits safemode, allowing the NameNode to resume normal operations, such as accepting write requests.

Safemode is a read-only mode for the Hadoop cluster, typically used during startup, maintenance, or recovery processes.

**w) HDFS command to format NameNode.**

hdfs namenode -format

Note:

This command initializes the HDFS metadata by formatting the NameNode's file system.

This process involves creating a new namespace and resetting the file system, which will result in the loss of all data stored in HDFS.

Therefore, it should be used with caution, typically only during the initial setup of the Hadoop cluster or when you want to start from scratch.

Important: Formatting the NameNode will delete all data and metadata in HDFS, so ensure that you have backups or are certain

that you want to perform this action.

**x) Create a file named hdfstest.txt and change it number of replications to 3.**

1.hadoop fs -put hdfstest.txt /hadoop/hdfs/hdfstest.txt

2.hadoop fs -setrep -w 3 /hadoop/hdfs/hdfstest.txt

Note:

1.Create the file in HDFS:

You can use the hdfs dfs -put command to upload a local file or create a file directly using a command like echo or cat.

Here's how to create a file from local content:

echo "Sample content for hdfstest.txt" | hdfs dfs -put - /user/hadoop2/hdfstest.txt

2.hdfs dfs -setrep -w 3 /user/hadoop2/hdfstest.txt

-setrep: Command to set the replication factor.

-w: Waits for the replication to be completed before returning.

3: The desired replication factor.

/user/hadoop2/hdfstest.txt: The HDFS path to the file.

This command will change the replication factor of hdfstest.txt to 3,

ensuring that there are three copies of the file's data distributed across the HDFS cluster.

**y) Write command to display number of replicas for hdfstest.txt file.**

hdfs dfs -stat %r /user/hadoop2/hdfstest.txt

**z) Write command to Display the status of file “hdfstest.txt” like block size, filesize in bytes.**

hdfs dfs -stat '%b %s %r' /user/hadoop2/hdfstest.txt

Note:

%b: Block size in bytes.

%s: File size in bytes.

%r: Replication factor.

-----------

For a detailed overview that includes more file attributes like permissions, modification time, etc., you can use:

hdfs dfs -stat %l,%s,%b,%r,%a,%t /user/hadoop2/hdfstest.txt

%l: Length of the file (size in bytes).

%s: Size of the file in bytes.

%b: Block size in bytes.

%r: Replication factor.

%a: Access time of the file.

%t: Modification time of the file.

aa) Write HDFS command to change file permission from rw – r – r to rwx-rw-x for hdfstest.txt.

hdfs dfs -chmod 765 /user/hadoop2/hdfstest.txt

Note:

Explanation of the permissions:

rwx (read, write, and execute) for the owner.

rw- (read and write) for the group.

r-x (read and execute) for others.

In octal notation, these permissions correspond to 764:

7 (binary 111): Owner has read (4), write (2), and execute (1) permissions.

6 (binary 110): Group has read (4) and write (2) permissions.

4 (binary 100): Others have read (4) and execute (1) permissions.

So, 764 sets the permissions to rwx-rw-r-x for hdfstest.txt.