



Practical - 3

2CS702 – Big Data Analytics

Harshit Gajipara

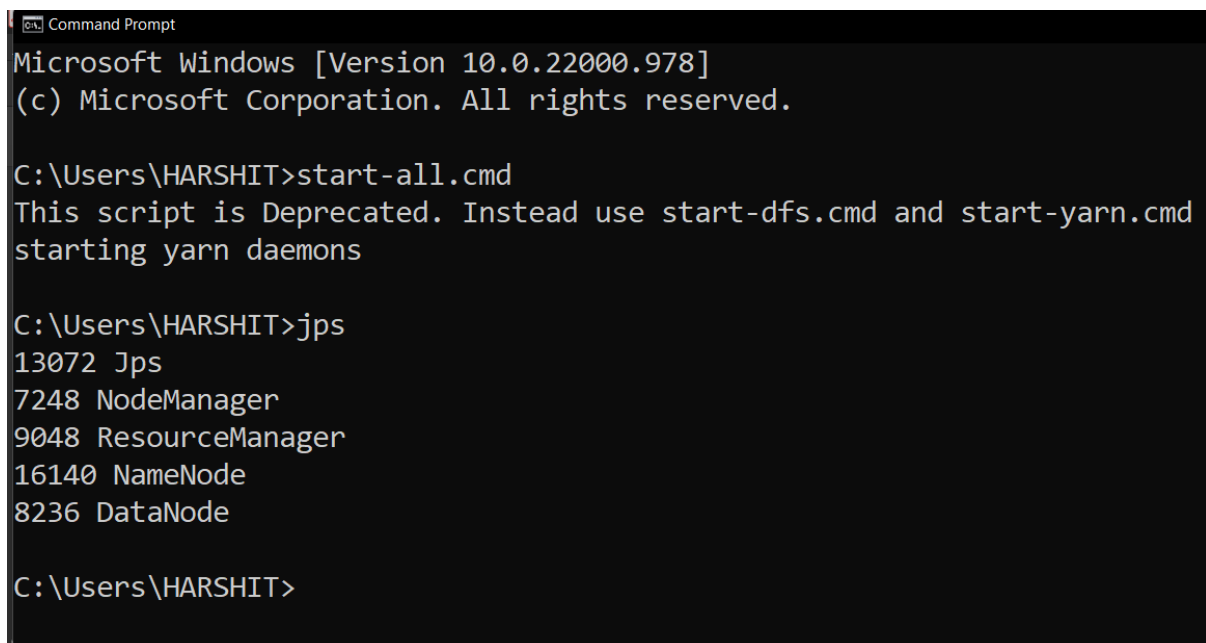
19BCE059

Aim: Setup single node Hadoop cluster and apply HDFS commands on single node Hadoop Cluster.

Commands:

- **jps**

To check hadoop services are up and running.

A screenshot of a Windows Command Prompt window. The title bar says 'Command Prompt'. The text inside shows the Windows version '10.0.22000.978' and copyright information. The user is at the prompt 'C:\Users\HARSHIT>' and has entered 'start-all.cmd'. A message says 'This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd starting yarn daemons'. Then the user enters 'jps' and the output shows five processes: '13072 Jps', '7248 NodeManager', '9048 ResourceManager', '16140 NameNode', and '8236 DataNode'. The prompt is now 'C:\Users\HARSHIT>'.

- **ls**

To list all files and directories in specified folder path

- **mkdir**

To create a directory in hadoop distributed file system

```
C:\Users\HARSHIT>hdfs dfs -mkdir /Directory1

C:\Users\HARSHIT>hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - HARSHIT supergroup          0 2022-09-30 11:59 /Directory1

C:\Users\HARSHIT>
```

- **rmdir**

To remove an empty directory from hdfs

```
C:\Users\HARSHIT>hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - HARSHIT supergroup          0 2022-09-30 12:00 /Directory1
-rw-r--r--   3 HARSHIT supergroup          0 2022-09-30 12:00 /random.txt

C:\Users\HARSHIT>hdfs dfs -rmdir /Directory1

C:\Users\HARSHIT>hdfs dfs -ls /
Found 1 items
-rw-r--r--   3 HARSHIT supergroup          0 2022-09-30 12:00 /random.txt

C:\Users\HARSHIT>
```

- **touchz**

To create empty files in hdfs

```
C:\Users\HARSHIT>hdfs dfs -touchz /random.txt

C:\Users\HARSHIT>hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - HARSHIT supergroup          0 2022-09-30 12:00 /Directory1
-rw-r--r--   3 HARSHIT supergroup          0 2022-09-30 12:00 /random.txt
```

- **chmod**

Used to change permissions of a file. Default is 644, which means user has read and write, group and other have only read.

4 is for read, 2 is for write and 1 is for execute.

```

C:\Users\HARSHIT>hdfs dfs -ls /
Found 1 items
-rw-r--r--  3 HARSHIT supergroup          0 2022-09-30 12:00 /random.txt

C:\Users\HARSHIT>hdfs dfs -chmod 666 /random.txt

C:\Users\HARSHIT>hdfs dfs -ls /
Found 1 items
-rw-rw-rw-  3 HARSHIT supergroup          0 2022-09-30 12:00 /random.txt

C:\Users\HARSHIT>

```

- **copyToLocal or get**

To copy files from hdfs to local system

```

C:\Users\HARSHIT>hdfs dfs -copyToLocal /random.txt

C:\Users\HARSHIT>dir random.txt
Volume in drive C is OS
Volume Serial Number is D4B3-16A5

Directory of C:\Users\HARSHIT

30-09-2022  12:03                0 random.txt
               1 File(s)                0 bytes
               0 Dir(s) 108,672,192,512 bytes free

C:\Users\HARSHIT>

```

- **copyFromLocal or put**

To copy files from local system to hdfs

```
C:\Users\HARSHIT>hdfs dfs -copyFromLocal random.txt /

C:\Users\HARSHIT>hdfs dfs -ls /
Found 1 items
-rw-r--r--   3 HARSHIT supergroup          0 2022-09-30 12:12 /random.txt

C:\Users\HARSHIT>
```

- **cp**

To copy files in distributed file system

```
C:\Users\HARSHIT>hdfs dfs -cp /random.txt /random2.txt

C:\Users\HARSHIT>hdfs dfs -ls /
Found 2 items
-rw-r--r--   3 HARSHIT supergroup          0 2022-09-30 12:12 /random.txt
-rw-r--r--   3 HARSHIT supergroup          0 2022-09-30 12:15 /random2.txt
```

- **mv**

To move files from one path location to other

```
C:\Users\HARSHIT>hdfs dfs -mv /random2.txt /dir/random2.txt

C:\Users\HARSHIT>hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - HARSHIT supergroup          0 2022-09-30 12:16 /dir
-rw-r--r--   3 HARSHIT supergroup          0 2022-09-30 12:12 /random.txt

C:\Users\HARSHIT>hdfs dfs -ls /dir
Found 1 items
-rw-r--r--   3 HARSHIT supergroup          0 2022-09-30 12:15 /dir/random2.txt

C:\Users\HARSHIT>_
```

- **rm**

To remove an existing file from dfs

```
C:\Users\HARSHIT>hdfs dfs -rm /random.txt
Deleted /random.txt
```

- **cat**

To view, edit and change content of a file

```
C:\Users\HARSHIT>hdfs dfs -cat /random.txt
2022-09-30 12:22:53,813 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
Hello world!
C:\Users\HARSHIT>
```

- **expunge**

To empty trash available in hdfs

```
HELLO WORLD!
C:\Users\HARSHIT>hdfs dfs -expunge

C:\Users\HARSHIT>hdfs dfs -ls /
Found 2 items
drwxr-xr-x  - HARSHIT supergroup          0 2022-09-30 12:16 /dir
-rw-rw-rw-  3 HARSHIT supergroup        12 2022-09-30 12:21 /random.txt
```

- **find**

To find size of a file or folder in hdfs

```
C:\Users\HARSHIT>hdfs dfs -ls /
Found 2 items
drwxr-xr-x  - HARSHIT supergroup          0 2022-09-30 12:16 /dir
-rw-rw-rw-  3 HARSHIT supergroup        12 2022-09-30 12:21 /random.txt

C:\Users\HARSHIT>hdfs dfs -find / -name "r*"
/dir/random2.txt
/random.txt
```

- **help**

To get help about any commands in hadoop

```
C:\Users\HARSHIT>hdfs dfs -help cp
-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst> :
Copy files that match the file pattern <src> to a destination. When copying
multiple files, the destination must be a directory. Passing -p preserves status
[topax] (timestamps, ownership, permission, ACLs, XAttr). If -p is specified
with no <arg>, then preserves timestamps, ownership, permission. If -pa is
specified, then preserves permission also because ACL is a super-set of
permission. Passing -f overwrites the destination if it already exists. raw
namespace extended attributes are preserved if (1) they are supported (HDFS
only) and, (2) all of the source and target pathnames are in the /.reserved/raw
hierarchy. raw namespace xattr preservation is determined solely by the presence
(or absence) of the /.reserved/raw prefix and not by the -p option. Passing -d
will skip creation of temporary file(<dst>._COPYING_).

C:\Users\HARSHIT>
```

- **tail**

To display content of file from end. By default it displays last 5 lines of file

```
C:\Users\HARSHIT>hdfs dfs -tail /random.txt
2022-09-30 12:44:28,317 INFO sasl.SaslDataTransferClient: SASL encry
teHostTrusted = false
Hello world!
Line 2
Line 3
Line 4
Line 5
Line 6
Line 7
C:\Users\HARSHIT>hdfs dfs -tail -n 2 /random.txt
```

- **test**

It is used for file test operations. It gives 1 if path exists. It gives 0 if it has 0 length or directory doesn't exist at given path. Basically it checks traits of file. Sets the UNIX return code.

- **count**

To count number of files, directories, characters in a file etc.

```
C:\Users\HARSHIT>hdfs dfs -count -v /
      DIR_COUNT      FILE_COUNT      CONTENT_SIZE  PATHNAME
              2              2              12  /
C:\Users\HARSHIT>_
```

- **stat**

To check stats of a particular file in hdfs

```
C:\Users\HARSHIT>hdfs dfs -stat /random.txt
2022-09-30 06:51:52
C:\Users\HARSHIT>_
```

- **usage**

Displays quick syntax help for a command

```
C:\Users\HARSHIT>hdfs dfs -usage cp
Usage: hadoop fs [generic options] -cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>
C:\Users\HARSHIT>_
```

- **du**

To display the size of each file in a directory

```
C:\Users\HARSHIT>hdfs dfs -du /
0    0    /dir
12   36   /random.txt
```


- **du**

Used to display total size of a file in hdfs

```
C:\Users\HARSHIT>hdfs dfs -du /  
du: DEPRECATED: Please use 'du -s' instead.  
12  36  /
```

Output:

We learnt different types of commands that can be implemented using hadoop for more parallel processing of data and efficient data processing.