

APACHE HADOOP WITH CLOUDERA TUTORIAL 2

Venkatesh Vinayakarao

venkateshv@cmi.ac.in

<http://vvtesh.co.in>

Chennai Mathematical Institute

HDFS Tutorial

- Run the Hadoop Container
 - `docker run --hostname=quickstart.cloudera --privileged=true -t -i -p 8888:8888 -p 8080:8080 -p 8088:8088 -p 7180:7180 -p 50070:50070 cloudera/quickstart /usr/bin/docker-quickstart`
- Let us create few folders
 - `hdfs dfs -mkdir <DIR NAME>`
- List those folders
 - `hdfs dfs -ls`

```
[root@quickstart /]# hdfs dfs -mkdir DATA
[root@quickstart /]# hdfs dfs -mkdir DATA2
[root@quickstart /]# hdfs dfs -ls
Found 2 items
drwxr-xr-x  - root supergroup          0 2023-09-09 16:49 DATA
drwxr-xr-x  - root supergroup          0 2023-09-09 16:49 DATA2
[root@quickstart /]#
```

HDFS Tutorial

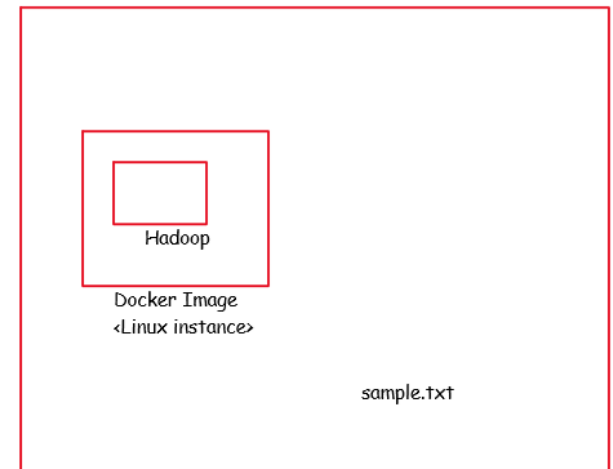
- Check the namenode information
 - <http://localhost:50070/>
- Go to utilities -> browse -> change to /user/root

Browse Directory

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	root	supergroup	0 B	Sat Sep 09 22:19:33 +0530 2023	0	0 B	DATA
drwxr-xr-x	root	supergroup	0 B	Sat Sep 09 22:19:46 +0530 2023	0	0 B	DATA2

Copying Files

- Local machine to docker
 - `docker cp c:/tmp/sample.txt <container id>:/tmp/sample.txt`
- Docker to HDFS
 - `hdfs dfs -put /tmp/sample.txt DATA`
 - `hdfs dfs -copyFromLocal /tmp/sample.txt DATA2`
- Print block count
 - `hdfs fsck /user/root/DATA -files`



Run this in your command prompt (not in Hadoop root)

My Machine
<Windows>

```
C:\Users\vvtes>docker ps
CONTAINER ID   IMAGE                COMMAND              CREATED        STATUS        PORTS
c511541a6c95   cloudera/quickstart "/usr/bin/docker-qui... 18 minutes ago Up 18 minutes 0.0.0.0:7180->7180/tcp, 0.0.0.0:8080->8080/tcp, 0.0.0.0:8088->8088/tcp, 0.0.0.0:8888->8888/tcp, 0.0.0.0:50070->50070/tcp
exciting_kilby

C:\Users\vvtes>docker cp c:\tmp\sample.txt c511541a6c95:/tmp/sample.txt
Successfully copied 2.05kB to c511541a6c95:/tmp/sample.txt
```

hadoop fs vs hdfs dfs

- Some commands are deprecated.

```
[root@quickstart /]# hadoop fsck /user/root/DATA -files -blocks
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
```

```
[root@quickstart /]# hdfs fsck /user/root/DATA -files
Connecting to namenode via http://quickstart.cloudera:50070
FSCK started by root (auth:SIMPLE) from /172.17.0.2 for path /user/root/DATA at Sat Sep 09 17:11:08 UTC 2023
/user/root/DATA <dir>
/user/root/DATA/sample.txt 11 bytes, 1 block(s): OK
Status: HEALTHY
Total size:      11 B
Total dirs:      1
Total files:     1
Total symlinks:  0
Total blocks (validated): 1 (avg. block size 11 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Sat Sep 09 17:11:08 UTC 2023 in 1 milliseconds
```

Some useful commands

```
hdfs dfs -mv DATA/sample.txt DATA2/sample2.txt
hdfs dfs -touchz DATA2/sample3.txt
hdfs dfs -rm DATA2/sample3.txt
hdfs dfs -tail DATA2/sample2.txt
hdfs dfs -cat DATA/sample1.txt
hdfs dfs -cp DATA/output.txt DATA2/output.txt
hdfs dfs -du DATA2
hdfs dfs -stat DATA
hdfs dfs -chmod -R 777 DATA
hdfs dfs -cat DATA/sample1.txt DATA2/sample2.txt
```

Thank You