

Hadoop Task1

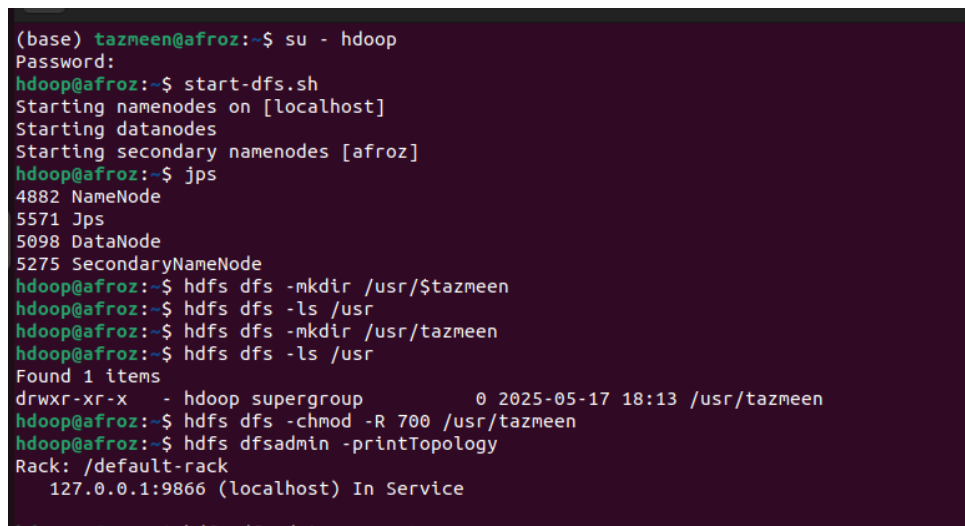
Tazmeen Afroz

May 2025

Task 0: Hadoop Installation (Single Node Setup)

- Installed Hadoop in pseudo-distributed (single-node) mode.
- Configured core-site.xml, hdfs-site.xml, mapred-site.xml, yarn-site.xml
- Successfully started NameNode, DataNode, ResourceManager, and NodeManager.

Verification Screenshots:

A terminal window with a dark purple background showing the steps to start Hadoop in pseudo-distributed mode. The user switches to the 'hdoop' user and runs 'start-dfs.sh', which starts the NameNode and DataNode. Then 'jps' is run, showing the processes: NameNode (4882), Jps (5571), DataNode (5098), and SecondaryNameNode (5275). Finally, 'hdfs dfs -mkdir /usr/\$tazmeen' is run, and 'ls /usr' shows the directory. Other commands include 'hdfs dfs -ls /usr', 'hdfs dfs -mkdir /usr/tazmeen', 'hdfs dfs -ls /usr', 'hdfs dfs -chmod -R 700 /usr/tazmeen', and 'hdfs dfsadmin -printTopology' which shows the rack information: /default-rack, 127.0.0.1:9866 (localhost) In Service.

```
(base) tazmeen@afroz:~$ su - hdoop
Password:
hdoop@afroz:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [afroz]
hdoop@afroz:~$ jps
4882 NameNode
5571 Jps
5098 DataNode
5275 SecondaryNameNode
hdoop@afroz:~$ hdfs dfs -mkdir /usr/$tazmeen
hdoop@afroz:~$ hdfs dfs -ls /usr
hdoop@afroz:~$ hdfs dfs -mkdir /usr/tazmeen
hdoop@afroz:~$ hdfs dfs -ls /usr
Found 1 items
drwxr-xr-x  - hdoop supergroup          0 2025-05-17 18:13 /usr/tazmeen
hdoop@afroz:~$ hdfs dfs -chmod -R 700 /usr/tazmeen
hdoop@afroz:~$ hdfs dfsadmin -printTopology
Rack: /default-rack
127.0.0.1:9866 (localhost) In Service
```

Figure 1: Hadoop Environment Setup

Task 1: Creating Your Directory Space

Command Used:

```
hdfs dfs -mkdir -p /usr/tazmeen
```

Task 2: Understanding the System

Commands Used:

```
hdfs dfsadmin -printTopology
hdfs dfsadmin -report
hdfs fsck /
hadoop fsck / -files -blocks -locations
```

Answers

1. **How many datanodes are part of the Hadoop topology?**
1 DataNode on my local , 2 on uni servers
2. **What are the IP addresses of these datanodes?**
127.0.0.1 (localhost) on my local ,permission issues on the uni servers
3. **What is the configured and present capacity of the HDFS?**
(Configured Capacity: 9.95 GB, Present Capacity: 9.5 GB (on my local) Configured Capacity: 3303399329792 (3.00 TB) Present Capacity: 2821930455040 (2.57 TB) DFS Remaining: 2821113008128 (2.57 TB) (uni)
4. **What is the default file replication count?**
2

```
Status: HEALTHY
Number of data-nodes: 2
Number of racks:      1
Total dirs:           0
Total symlinks:        0

Replicated Blocks:
Total size:    38 B
Total files:   1
Total blocks (validated): 1 (avg. block size 38 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
Missing blocks: 0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Blocks queued for replication: 0

Erasure Coded Block Groups:
Total size:    0 B
Total files:   0
Total block groups (validated): 0
Minimally erasure-coded block groups: 0
Over-erasure-coded block groups: 0
Under-erasure-coded block groups: 0
Unsatisfactory placement block groups: 0
Average block group size: 0.0
Missing block groups: 0
Corrupt block groups: 0
Missing internal blocks: 0
Blocks queued for replication: 0
FSCK ended at Mon Feb 17 16:05:22 PKT 2025 in 1 milli
```

Figure 2: System Report Output

Task 3: Block Size and Replica Analysis

Initial File: bible.txt

Command Used:

```
hdfs dfs -put bible.txt /usr/tazmeen/bible.txt
hdfs fsck /usr/tazmeen/bible.txt -files -blocks -locations
```

Answers:

1. Default block size: **64kb**
2. Missing replicas before block size change: **0**
3. Command to change block size:

```
hdfs dfs -Ddfs.blocksize=1048576 -put bible.txt /usr/tazmeen/test.txt
```

4. Blocks used after block size change: **5**
5. Missing replicas after change: **0**
6. Why missing replicas? **No missing replicas found.**

```

Status: HEALTHY
Number of data-nodes: 2
Number of racks: 1
Total dirs: 0
Total symlinks: 0

Replicated Blocks:
Total size: 4332554 B
Total files: 1
Total blocks (validated): 5 (avg. block size 866510 B)
Minimally replicated blocks: 5 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 2
Average block replication: 2.0
Missing blocks: 0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Blocks queued for replication: 0

Erasure Coded Block Groups:
Total size: 0 B
Total files: 0
Total block groups (validated): 0
Minimally erasure-coded block groups: 0
Over-erasure-coded block groups: 0
Under-erasure-coded block groups: 0
Unsatisfactory placement block groups: 0
Average block group size: 0.0
Missing block groups: 0
Corrupt block groups: 0
Missing internal blocks: 0
Blocks queued for replication: 0
FSCK ended at Sat May 17 19:40:32 PKT 2025 in 1 milliseconds

```

Figure 3: FSCK Output After Changing Block Size

Task 4: Word Count Application

Command Used:

```

hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-
mapreduce-examples-*.jar wordcount /input /output

```

Answers:

1. {key, value} pair: <word : 1>
2. Mapper threads used: **67**
3. Reducer threads used: **1**
4. Time spent by mappers: **222,182 milliseconds (or 222.182 seconds)** (from log)
5. Time spent by reducers: **43,799 milliseconds (or 43.799 seconds)** (from log)
6. Output file name: **/output/part-r-00000**

```

Data-local Map tasks=0/
Total time spent by all maps in occupied slots (ms)=222182
Total time spent by all reduces in occupied slots (ms)=43799
Total time spent by all map tasks (ms)=222182
Total time spent by all reduce tasks (ms)=43799
Total vcore-milliseconds taken by all map tasks=222182
Total vcore-milliseconds taken by all reduce tasks=43799
Total megabyte-milliseconds taken by all map tasks=227514368
Total megabyte-milliseconds taken by all reduce tasks=44850176
Map-Reduce Framework
  Map input records=99805
  Map output records=99805
  Map output bytes=2916719
  Map output materialized bytes=3116731
  Input split bytes=7102
  Combine input records=0
  Combine output records=0
  Reduce input groups=66839
  Reduce shuffle bytes=3116731
  Reduce input records=99805
  Reduce output records=66838
  Spilled Records=199610
  Shuffled Maps =67
  Failed Shuffles=0
  Merged Map outputs=67
  GC time elapsed (ms)=9389
  CPU time spent (ms)=51210
  Physical memory (bytes) snapshot=20483964928
  Virtual memory (bytes) snapshot=174002524160
  Total committed heap usage (bytes)=31631343616
  Peak Map Physical memory (bytes)=601755648
  Peak Map Virtual memory (bytes)=2563670016
  Peak Reduce Physical memory (bytes)=386494464
  Peak Reduce Virtual memory (bytes)=2567462912
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  MAP_FAILED=0
  NOT_IMPLEMENTED=0
  OTHER=0
  REQUESTED=0
  TIMEOUT=0
  UNRECOVERED_FAILURE=0
  UNKNOWN=0
  WRITER_TIMEOUT=0

```

Figure 4: MapReduce Job Completed