



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Big Data Analytics Lab

PMDS507P

Name: **Tufan Kundu**

Registration number: **24MDT0184**

Slot: L29+L30

Digital Assignment 2

Step 1: Go to Command Prompt and type ([start-dfs.cmd](#))

```
Administrator: Command Prompt
C:\Windows\System32>start-dfs.cmd
```

```
Apache Hadoop Distribution - hadoop  datanode
2025-09-17 19:21:43,256 INFO checker.DatasetVolumeChecker: Scheduled health check for volume C:\hadoop\data\dfs\datanode
2025-09-17 19:21:43,283 INFO datanode.VolumeScanner: VolumeScanner(C:\hadoop\data\dfs\datanode, DS-ce1718e4-ed93-4044-adaa-c1113e2a2903): no suitable block pools found to scan. Waiting 1814302017 ms.
2025-09-17 19:21:43,285 WARN datanode.DirectoryScanner: dfs.datanode.directoryscan.throttle.limit.ms.per.sec set to value above 1000 ms/sec. Assuming default value of -1
2025-09-17 19:21:43,286 INFO datanode.DirectoryScanner: Periodic Directory Tree Verification scan starting in 9146178ms with interval of 21600000ms and throttle limit of -1ms/s
2025-09-17 19:21:43,290 INFO datanode.DataNode: Block pool BP-1730253224-172.16.167.193-1758116997780 (Datanode Uuid 055a41d5-ae4-4adc-897f-e1e1a445ac2c) service to localhost/127.0.0.1:9000 beginning handshake with NN: localhost/127.0.0.1:9000.
2025-09-17 19:21:43,346 INFO datanode.DataNode: Block pool BP-1730253224-172.16.167.193-1758116997780 (Datanode Uuid 055a41d5-ae4-4adc-897f-e1e1a445ac2c) service to localhost/127.0.0.1:9000 successfully registered with NN: localhost/127.0.0.1:9000.
2025-09-17 19:21:43,346 INFO datanode.DataNode: For namenode localhost/127.0.0.1:9000 using BLOCKREPORT_INTERVAL of 21600000msecs CACHEREPORT_INTERVAL of 10000msecs Initial delay: 0msecs; heartbeatInterval=3000
2025-09-17 19:21:43,426 INFO datanode.DataNode: After receiving heartbeat response, updating state of namenode localhost:9000 to active
2025-09-17 19:21:43,481 INFO datanode.DataNode: Successfully sent block report 0xe14e3d5af616a65e with lease ID 0xfcbe0d243293633 to namenode: localhost/127.0.0.1:9000, containing 1 storage report(s), of which we sent 1. The reports had 0 total blocks and used 1 RPC(s). This took 4 mssecs to generate and 48 mssecs for RPC and NN processing. Got back one command: FinalizeCommand/5.
2025-09-17 19:21:43,481 INFO datanode.DataNode: Got finalize command for block pool BP-1730253224-172.16.167.193-1758116997780

Apache Hadoop Distribution - hadoop  namenode
7 milliseconds
name space=1
storage space=0
storage types=RAM_DISK=0, SSD=0, DISK=0, ARCHIVE=0, PROVIDED=0, NVDIMM=0
2025-09-17 19:21:42,632 INFO blockmanagement.CacheReplicationMonitor: Starting CacheReplicationMonitor with interval 30000 milliseconds
2025-09-17 19:21:43,334 INFO hdfs.StateChange: BLOCK* registerDatanode: from DatanodeRegistration(127.0.0.1:9866, datanodeUuid=055a41d5-ae4-4adc-897f-e1e1a445ac2c, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-a93d0ce8-6f25-469a-b630-8a7de6c18b46;nsid=160449242;c=1758116997780) storage 055a41d5-ae4-4adc-897f-e1e1a445ac2c
2025-09-17 19:21:43,337 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2025-09-17 19:21:43,337 INFO blockmanagement.BlockReportLeaseManager: Registered DN 055a41d5-ae4-4adc-897f-e1e1a445ac2c (127.0.0.1:9866).
2025-09-17 19:21:43,407 INFO blockmanagement.DatanodeDescriptor: Adding new storage ID DS-ce1718e4-ed93-4044-adaa-c1113e2a2903 for DN 127.0.0.1:9866
2025-09-17 19:21:43,451 INFO BlockStateChange: BLOCK* processReport 0xe14e3d5af616a65e with lease ID 0xfcbe0d243293633: Processing first storage report for DS-ce1718e4-ed93-4044-adaa-c1113e2a2903 from datanode DatanodeRegistration(127.0.0.1:9866, datanodeUuid=055a41d5-ae4-4adc-897f-e1e1a445ac2c, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-a93d0ce8-6f25-469a-b630-8a7de6c18b46;nsid=160449242;c=1758116997780)
2025-09-17 19:21:43,453 INFO BlockStateChange: BLOCK* processReport 0xe14e3d5af616a65e with lease ID 0xfcbe0d243293633: from storage DS-ce1718e4-ed93-4044-adaa-c1113e2a2903 node DatanodeRegistration(127.0.0.1:9866, datanodeUuid=055a41d5-ae4-4adc-897f-e1e1a445ac2c, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-a93d0ce8-6f25-469a-b630-8a7de6c18b46;nsid=160449242;c=1758116997780), blocks: 0, hasStaleStorage: false, processing time: 2 mssecs, invalidatedBlocks: 0
```

Step 2: Now type (start-yarn.cmd)

```
C:\Windows\System32>start-yarn.cmd
starting yarn daemons
```

```
Apache Hadoop Distribution - yarn  resourcemanager
2025-09-17 19:23:30,632 INFO ipc.Server: IPC Server listener on 8030: starting
2025-09-17 19:23:30,727 INFO ipc.CallQueueManager: Using callQueue: class java.util.concurrent.LinkedBlockingQueue, queueCapacity: 5000, scheduler: class org.apache.hadoop.ipc.DefaultRpcScheduler, ipcBackoff: false, ipcFailOver: false.
2025-09-17 19:23:30,728 INFO ipc.Server: Listener at 0.0.0.0:8032
2025-09-17 19:23:30,731 INFO ipc.Server: Starting Socket Reader #1 for port 8032
2025-09-17 19:23:30,736 INFO pb.RpcServerFactoryPBImpl: Adding protocol org.apache.hadoop.yarn.api.ApplicationClientProtocolPB to the server
2025-09-17 19:23:30,737 INFO ipc.Server: IPC Server listener on 8032: starting
2025-09-17 19:23:30,738 INFO ipc.Server: IPC Server Responder: starting
2025-09-17 19:23:31,135 INFO webproxy.ProxyCA: Created Certificate for OU=YARN-a2852858-8169-46e2-8803-17b3e2efa5ba
2025-09-17 19:23:31,232 INFO recovery.RMStateStore: Storing CA Certificate and Private Key
2025-09-17 19:23:31,234 INFO resourcemanager.ResourceManager: Transitioned to active state
2025-09-17 19:23:31,801 INFO resourcemanager.ResourceTrackerService: NodeManager from node TK(cmPort: 6549 httpPort: 8042) registered with capability: <memory:8192, vCores:8>, assigned nodeId TK:6549
2025-09-17 19:23:31,804 INFO rmnode.RMNodeImpl: TK:6549 Node Transitioned from NEW to RUNNING
2025-09-17 19:23:31,836 INFO capacity.AbstractLeafQueue: LeafQueue: root.default update max app related, maxApplications=10000, maxApplicationsPerUser=10000, Abs Cap:1.0, Cap: 1.0, MaxCap : 1.0
2025-09-17 19:23:31,837 INFO capacity.CapacityScheduler: Added node TK:6549 clusterResource: <memory:8192, vCores:8>
2025-09-17 19:23:31,838 INFO capacity.AbstractLeafQueue: LeafQueue: root.default update max app related, maxApplications=10000, maxApplicationsPerUser=10000, Abs Cap:1.0, Cap: 1.0, MaxCap : 1.0

Apache Hadoop Distribution - yarn  nodemanager
Sep 17, 2025 7:23:30 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.webapp.GenericExceptionHandler to GuiceManagedComponentProvider with the scope "Singleton"
Sep 17, 2025 7:23:31 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.NMWebServices to GuiceManagedComponentProvider with the scope "Singleton"
2025-09-17 19:23:31,124 INFO handler.ContextHandler: Started o.e.j.w.WebAppContext@5fb5ad40{/node/,file:///C:/Users/TUFAN/AppData/Local/Temp/jetty-0_0_0-8042-hadoop-yarn-common-3.4.1_jar--any-5425193700034583051/webapp/,AVAILABLE}{jar:file:/C:/hadoop/share/hadoop/yarn/hadoop-yarn-common-3.4.1.jar!/webapps/node}
2025-09-17 19:23:31,139 INFO server.AbstractConnector: Started ServerConnector@295bfa2a{HTTP/1.1, (http/1.1)}{0.0.0.0:8042}
2025-09-17 19:23:31,140 INFO server.Server: Started @6025ms
2025-09-17 19:23:31,144 INFO webapp.WebApps: Web app node started at 8042
2025-09-17 19:23:31,145 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : TK:6549.
2025-09-17 19:23:31,148 INFO util.JvmPauseMonitor: Starting JVM pause monitor
2025-09-17 19:23:31,170 INFO client.DefaultNoHARMFalloverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8031
2025-09-17 19:23:31,274 INFO nodemanager.NodeStatusUpdaterImpl: Running Applications Size : 0.
2025-09-17 19:23:31,839 INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key with id -2036005023
2025-09-17 19:23:31,841 INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key with id 659971207
2025-09-17 19:23:31,841 INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as TK:6549 with total resource of <memory:8192, vCores:8>
```

Step 3: Now type jps, which will display a list of all currently running Java processes on the local system.

```
Administrator: Command Prompt

C:\Windows\System32>jps
7712 ResourceManager
18356 Jps
12360 DataNode
6172 NodeManager
5996 NameNode
```

Step 4: Stop all the processes by typing (stop-dfs.cmd) and ([stop-yarn.cmd](#))

```
Administrator: Command Prompt

C:\Windows\System32>stop-dfs.cmd
SUCCESS: Sent termination signal to the process with PID 20368.
SUCCESS: Sent termination signal to the process with PID 15064.

C:\Windows\System32>stop-yarn.cmd
stopping yarn daemons
SUCCESS: Sent termination signal to the process with PID 19288.
SUCCESS: Sent termination signal to the process with PID 19596.

INFO: No tasks running with the specified criteria.
```

Step 5: Now try to open all the processes by a single command (start-all.cmd) and check if all the processes are running by the command jps

```
Administrator: Command Prompt

C:\Windows\System32>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
```

```
Apache Hadoop Distribution - hadoop namenode
2025-09-17 19:27:52,603 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2025-09-17 19:27:52,603 INFO blockmanagement.BlockReportLeaseManager: Registered DN 055a41d5-ae4-4adc-897f-e1e1a445ac2c (127.0.0.1:9866).
2025-09-17 19:27:52,734 INFO blockmanagement.DatanodeDescriptor: Adding new storage ID DS-ce1718e4-ed93-4044-adaa-c1113e2a2903 for DN 127.0.0.1:9866
2025-09-17 19:27:52,812 INFO BlockStateChange: BLOCK* processReport 0xa8831ecc2ecdcc40 with lease ID 0x33b7b98f31e34131: Processing first storage report for DS-ce1718e4-ed93-4044-adaa-c1113e2a2903 from datanode DatanodeRegistration(127.0.0.1:9866, datanodeUuid=055a41d5-ae4-4adc-897f-e1e1a445ac2c, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-a93d0ce8-6f25-469a-b630-8a7de6c18b46;nsid=160449242;c=1758116997780)
2025-09-17 19:27:52,815 INFO BlockStateChange: BLOCK* processReport 0xa8831ecc2ecdcc40 with lease ID 0x33b7b98f31e34131: From storage DS-ce1718e4-ed93-4044-adaa-c1113e2a2903 node DatanodeRegistration(127.0.0.1:9866, datanodeUuid=055a41d5-ae4-4adc-897f-e1e1a445ac2c, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-a93d0ce8-6f25-469a-b630-8a7de6c18b46;nsid=160449242;c=1758116997780), blocks: 0, hasStaleStorage: false, processing time: 3 msec, invalidatedBlocks: 0

Apache Hadoop Distribution - hadoop datanode
2025-09-17 19:27:52,523 INFO datanode.DataNode: Block pool BP-1730253224-172.16.167.193-1758116997780 (Datanode Uuid 055a41d5-ae4-4adc-897f-e1e1a445ac2c) service to localhost/127.0.0.1:9000 beginning handshake with NN: localhost/127.0.0.1:9000.
2025-09-17 19:27:52,619 INFO datanode.DataNode: Block pool BP-1730253224-172.16.167.193-1758116997780 (Datanode Uuid 055a41d5-ae4-4adc-897f-e1e1a445ac2c) service to localhost/127.0.0.1:9000 successfully registered with NN: localhost/127.0.0.1:9000.
2025-09-17 19:27:52,620 INFO datanode.DataNode: For namenode localhost/127.0.0.1:9000 using BLOCKREPORT_INTERVAL of 21600000msecs CACHEREPORT_INTERVAL of 10000msecs Initial delay: 0msecs; heartbeatInterval=3000
2025-09-17 19:27:52,762 INFO datanode.DataNode: After receiving heartbeat response, updating state of namenode localhost:9000 to active
2025-09-17 19:27:52,864 INFO datanode.DataNode: Successfully sent block report 0xa8831ecc2ecdcc40 with lease ID 0x33b7b98f31e34131 to namenode: localhost/127.0.0.1:9000, containing 1 storage report(s), of which we sent 1. The reports had 0 total blocks and used 1 RPC(s). This took 8 msec to generate and 92 msec for RPC and NN processing. Got back one command: FinalizeCommand/5.
2025-09-17 19:27:52,865 INFO datanode.DataNode: Got finalize command for block pool BP-1730253224-172.16.167.193-1758116997780

Apache Hadoop Distribution - yarn nodemanager
p/share/hadoop/yarn/hadoop-yarn-common-3.4.1.jar!/webapps/node}
2025-09-17 19:27:53,941 INFO server.AbstractSecretManager: Started ServerConnector@295bf2a{HTTP/1.1, (http/1.1)}{0.0.0.0:8042}
2025-09-17 19:27:53,941 INFO server.Server: Started @7694ms
2025-09-17 19:27:53,944 INFO webapp.WebApps: Web app node started at 8042
2025-09-17 19:27:53,946 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : TK:6739.
2025-09-17 19:27:53,957 INFO util.JvmPauseMonitor: Starting JVM pause monitor
2025-09-17 19:27:53,970 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8031
2025-09-17 19:27:54,023 INFO nodemanager.NodeStatusUpdaterImpl: Running Applications Size : 0
2025-09-17 19:27:54,362 INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key with id 1949176106
2025-09-17 19:27:54,363 INFO security.NMTokenSecretManagerINM: Rolling master-key for container-tokens, got key with id -233391305
2025-09-17 19:27:54,365 INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as TK:6739 with total resource of <memory:8192, vCores:8>

Apache Hadoop Distribution - yarn resourcemanager
608-6369-4a00-9e17-e2cd8d5416f7
2025-09-17 19:27:53,805 INFO recovery.RMStateStore: Storing CA Certificate and Private Key
2025-09-17 19:27:53,813 INFO resourcemanager.ResourceManager: Transitioned to active state
2025-09-17 19:27:54,337 INFO resourcemanager.ResourceTrackerService: NodeManager from node TK(cmpPort: 6739 httpPort: 8042) registered with capability: <memory:8192, vCores:8>, assigned nodeId TK:6739
2025-09-17 19:27:54,339 INFO rmnode.RMNodeImpl: TK:6739 Node Transitioned from NEW to RUNNING
2025-09-17 19:27:54,372 INFO capacity.AbstractLeafQueue: LeafQueue: root.default update max app related, maxApplications=10000, maxApplicationsPerUser=10000, Abs Cap:1.0, Cap:1.0, MaxCap:1.0
2025-09-17 19:27:54,372 INFO capacity.CapacityScheduler: Added node TK:6739 clusterResource: <memory:8192, vCores:8>
2025-09-17 19:27:54,374 INFO capacity.AbstractLeafQueue: LeafQueue: root.default update max app related, maxApplications=10000, maxApplicationsPerUser=10000, Abs Cap:1.0, Cap:1.0, MaxCap:1.0
```

```
Administrator: Command Prompt
C:\Windows\System32>jps
19392 ResourceManager
17988 NodeManager
11560 NameNode
12888 Jps
5576 DataNode
```

Step 6: List out all the root directories by the command (hdfs dfs -ls /)

```
Administrator: Command Prompt
C:\Windows\System32>hdfs dfs -ls /
Found 1 items
drwxr-xr-x  - TK supergroup          0 2025-09-17 19:30 /filename
```

Step 7: Explore localhost:9870/ and localhost:8088/

localhost:9870/dfshealth.html#tab-overview

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview

'localhost:9000' (active)

Started:	Wed Sep 17 19:27:49 +0530 2025
Version:	3.4.1, r4d7825309348956336b8f06a08322b78422849b1
Compiled:	Wed Oct 09 20:27:00 +0530 2024 by mthakur from branch-3.4.1
Cluster ID:	CID-a93d0ce8-6f25-469a-b630-8a7de6c18b46
Block Pool ID:	BP-1730253224-172.16.167.193-1758116997780

Summary

Security is off.

Safemode is off.

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Us
0	0	0	0	0	<memory:0 B, ...

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State
Showing 0 to 0 of 0 entries										

Create Directory:

Step 8: Now create a new directory by command (hdfs dfs -mkdir /filename) and view the list of directories again

```
Administrator: Command Prompt
C:\Windows\System32>hdfs dfs -mkdir /tufan

C:\Windows\System32>hdfs dfs -ls /
Found 2 items
drwxr-xr-x - TK supergroup          0 2025-09-17 19:30 /filename
drwxr-xr-x - TK supergroup          0 2025-09-17 19:34 /tufan
```

Create multiple directories at once:

Step 9: We can create multiple directories at once by the command (hdfs dfs -mkdir -p/path1/path2)

```
Administrator: Command Prompt
C:\Windows\System32>hdfs dfs -mkdir -p /path1/path2

C:\Windows\System32>hdfs dfs -ls /
Found 3 items
drwxr-xr-x - TK supergroup          0 2025-09-17 19:30 /filename
drwxr-xr-x - TK supergroup          0 2025-09-17 19:35 /path1
drwxr-xr-x - TK supergroup          0 2025-09-17 19:34 /tufan
```

Put (upload) a file to HDFS:

Step 10: Now create a text file and upload it into the new directory by command (hdfs dfs -put data.txt /myfolder)

```
Administrator: Command Prompt
C:\Windows\System32>hdfs dfs -put data.txt /myfolder
C:\Windows\System32>hdfs dfs -get /myfolder/data.txt
get: `data.txt': File exists
```

Get(download) a file for HDFS:

Step 11: To download a file from hdfs we can use the command (hdfs dfs -get /myfolder/data.txt)

```
C:\Windows\System32>hdfs dfs -get /myfolder/data.txt
get: `data.txt': File exists
```

Remove a file:

Step 12: To remove a file from any directory, we use the command (hdfs dfs -rm /myfolder/data.txt)

```
C:\Windows\System32>hdfs dfs -rm /myfolder/data.txt
Deleted /myfolder/data.txt
```

Remove a file recursively:

Step 13: To delete a directory, we can use (hdfs dfs -rm -r /myfolder) this will delete all the files/folders permanently

```
C:\Windows\System32>hdfs dfs -rm -r /myfolder
Deleted /myfolder
```

Step 14: For viewing file contents (hdfs dfs -cat /myfolder/data.txt)

```
C:\Windows\System32>hdfs dfs -cat /myfolder/data.txt
My name is Tufan Kundu.
I am pursuing MSc in Data Science at VIT.
I have experience in C, C++, Python, and Java.
My interests include Deep Learning, Machine Learning, and Big Data Analytics.
I enjoy working on projects that involve data-driven problem solving.
```

Step 15: To display the first few lines, we use (hdfs dfs -head /myfolder/data.txt)

```
C:\Windows\System32>hdfs dfs -head /myfolder/data.txt
My name is Tufan Kundu.
I am pursuing MSc in Data Science at VIT.
I have experience in C, C++, Python, and Java.
My interests include Deep Learning, Machine Learning, and Big Data Analytics.
I enjoy working on projects that involve data-driven problem solving.
```

Step 16: To display the last few lines, we use (hdfs dfs -tail /myfolder/data.txt)

```
Administrator: Command Prompt
C:\Windows\System32>hdfs dfs -tail /myfolder/data.txt
My name is Tufan Kundu.
I am pursuing MSc in Data Science at VIT.
I have experience in C, C++, Python, and Java.
My interests include Deep Learning, Machine Learning, and Big Data Analytics.
I enjoy working on projects that involve data-driven problem solving.
```

Step 17: For copying files within hdfs (hdfs dfs -cp /myfolder/data.txt /backup)

```
C:\Windows\System32>hdfs dfs -cp /myfolder/data.txt /backup
```

Step 18: For checking the disk usage (hdfs dfs -du -h /)

```
C:\Windows\System32>hdfs dfs -du -h /
266  266  /backup
0    0    /filename
266  266  /myfolder
0    0    /path1
0    0    /tufan
```

Step 19: To show the file checksum (hdfs dfs -checksum /myfolder/data.txt)

```
Administrator: Command Prompt
C:\Windows\System32>hdfs dfs -checksum /myfolder/data.txt
/myfolder/data.txt      MD5-of-0MD5-of-512CRC32C      000002000000000000000000df69ea702a.
```