

ASSIGNMENT – I

TITLE :

Implement a distributed data processing pipeline using Apache Hadoop and Spark.

NAME : Shinde Shubham Dnyandev. **ROLL NO :** 23107121. **BATCH :** B

1)HADOOP :

```
sudo apt-get update  
ssh localhost  
hadoop-3.3.1/bin/hdfs namenode -format  
start-all.sh  
localhost:9870
```

The screenshot shows a Google Chrome window with the URL `localhost:9870/dfshealth.html#tab-overview`. The title bar says "Namenode information". The main content area has a green header tab labeled "Hadoop". Below it are tabs for "Overview", "Datanodes", "Datanode Volume Failures", "Snapshot", "Startup Progress", and "Utilities". The "Overview" tab is active. It displays the following information:

Started:	Tue Dec 16 20:30:09 +0530 2025
Version:	3.3.1, ra3b9c37a397ad4188041dd80621bdeefc46885f2
Compiled:	Tue Jun 15 10:43:00 +0530 2021 by ubuntu from (HEAD detached at release-3.3.1-RC3)
Cluster ID:	C10-2bb92d03-c9f4-4c70-abef-31406e4162ec
Block Pool ID:	BP-1653505603-172.16.9.114-1765897181336

Below this is a "Summary" section with the following details:

Configured Capacity:	154.52 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)
Non DFS Used:	49.08 GB
DFS Remaining:	97.52 GB (63.11%)
Block Pool Used:	24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)

2)SPARK :

Step 1: Check Python Installation
`python3 --version`

Step 2: Set PYTHON Environment Variables
Check Python path:
`which python3`

Edit environment file:
`nano ~/.bashrc`

Add the following lines:
`export PYSPARK_PYTHON=/usr/bin/python3`
`export PYSPARK_DRIVER_PYTHON=/usr/bin/python3`

Apply changes:
source `~/.bashrc`

Step 3: Download Apache Spark

```
cd /opt  
sudo wget https://archive.apache.org/dist/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz
```

Step 4: Extract and Configure Spark

```
sudo tar -xvzf spark-3.5.0-bin-hadoop3.tgz  
sudo mv spark-3.5.0-bin-hadoop3 spark
```

Step 5: Set Spark Environment Variables

```
nano ~/.bashrc
```

Add:

```
export SPARK_HOME=/opt/spark  
export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin
```

Apply changes:

```
source ~/.bashrc
```

Step 6: Start PySpark Shell

```
pyspar
```

Successful installation output:

Spark context Web UI available at <http://localhost:4040>

The screenshot shows a web browser window with the following details:

- Title Bar:** Activities, Google Chrome, VPKBIET | Baramati, T.Y.B.Tech_AIDS_2023, PySparkShell - Environment.
- URL:** localhost:4040/environment
- Content Area:**
 - Environment Tab:** Shows the PySparkShell application UI.
 - Runtime Information:**

Name	Value
Java Home	/usr/lib/jvm/java-8-openjdk-amd64/jre
Java Version	1.8.0_452 (Private Build)
Scala Version	version 2.12.17
 - Spark Properties:**

Name	Value
spark.app.id	local-1767088461588
spark.app.name	PySparkShell
spark.app.startTime	1767088461067
spark.app.submitTime	1767088460761
spark.driver.extraJavaOptions	-Djava.net.preferIPv6Addresses=false -XX:+IgnoreUnrecognizedVMOptions --add-opens=java.base/java.lang=ALL-UNNAMED --add-opens=java.base/java.lang.invoke=ALL-UNNAMED --add-opens=java.base/java.lang.reflect=ALL-UNNAMED --add-opens=java.base/java.io=ALL-UNNAMED --add-opens=java.base/java.net=ALL-UNNAMED --add-opens=java.base/java.nio=ALL-UNNAMED --add-opens=java.base/java.util=ALL-UNNAMED --add-opens=java.base/java.util.concurrent=ALL-UNNAMED --add-opens=java.base/java.util.concurrent.atomic=ALL-UNNAMED --add-opens=java.base/jdk.internal.ref=ALL-UNNAMED --add-opens=java.base/sun.nio.ch=ALL-UNNAMED --add-opens=java.base/sun.nio.cs=ALL-UNNAMED --add-opens=java.base/sun.security.action=ALL-UNNAMED --add-opens=java.base/sun.util.calendar=ALL-UNNAMED --add-opens=java.security.jgss/sun.security.krb5=ALL-UNNAMED -Djdk.reflect.useDirectMethodHandle=false
spark.driver.host	plcomp16
spark.driver.port	46131
spark.executor.extraJavaOptions	-Djava.net.preferIPv6Addresses=false -XX:+IgnoreUnrecognizedVMOptions --add-opens=java.base/java.lang=ALL-UNNAMED --add-opens=java.base/java.lang.invoke=ALL-UNNAMED --add-opens=java.base/java.lang.reflect=ALL-UNNAMED --add-opens=java.base/java.io=ALL-UNNAMED --add-opens=java.base/java.net=ALL-UNNAMED --add-opens=java.base/java.nio=ALL-UNNAMED --add-opens=java.base/java.util=ALL-UNNAMED --add-opens=java.base/java.util.concurrent=ALL-UNNAMED --add-opens=java.base/java.util.concurrent.atomic=ALL-UNNAMED --add-opens=java.base/jdk.internal.ref=ALL-UNNAMED --add-opens=java.base/sun.nio.ch=ALL-UNNAMED --add-opens=java.base/sun.nio.cs=ALL-UNNAMED --add-opens=java.base/sun.security.action=ALL-UNNAMED --add-opens=java.base/sun.util.calendar=ALL-UNNAMED --add-opens=java.security.jgss/sun.security.krb5=ALL-UNNAMED -Djdk.reflect.useDirectMethodHandle=false
spark.executor.id	driven
spark.master	local[*]