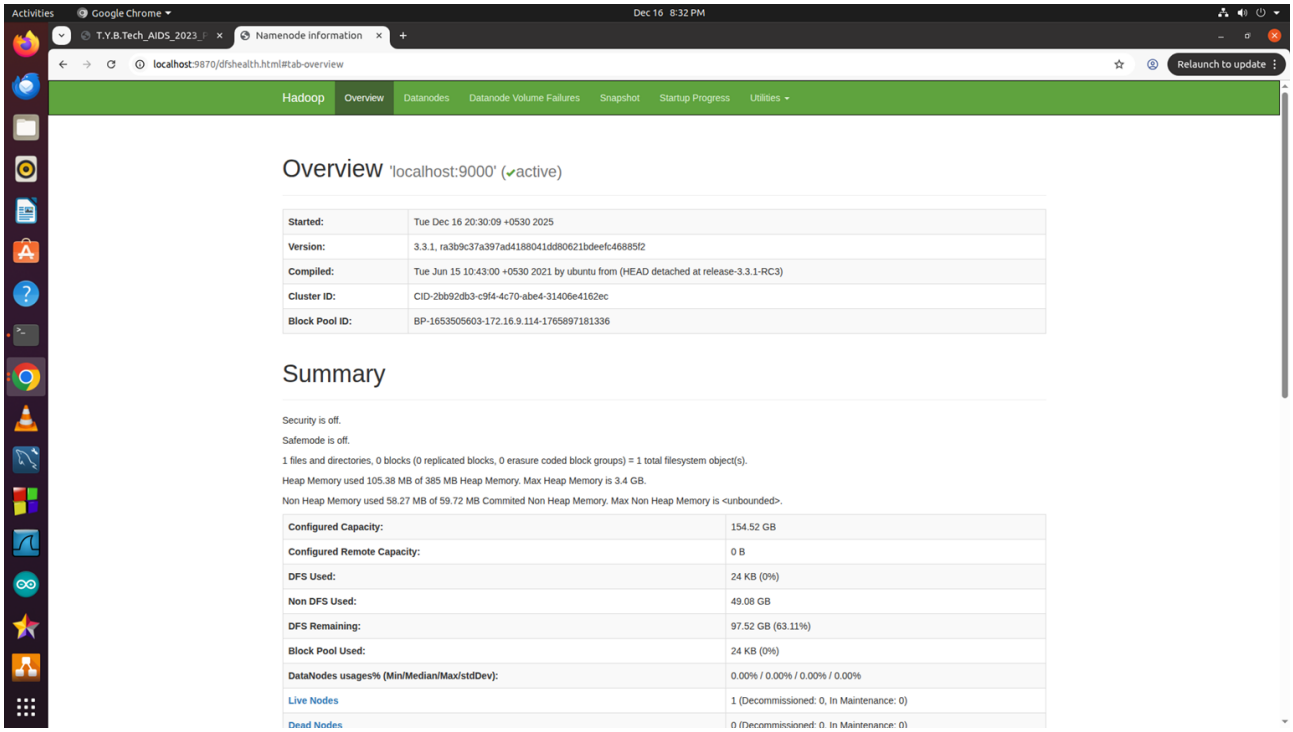
ASSIGNMENT – I

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



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