LAB ASSIGNEMNT 01

IMAAD IMRAN HAJWANE

202101132 / 23

TOPIC: INSTALLATION OF HADOOP

Question:

What is Hadoop? Describe its Architecture.

Answer:

Hadoop is an open-source software framework used for storing and processing large datasets in a distributed computing environment. It is designed to scale up from a single server to thousands of machines, each offering local computation and storage.

Key Features of Hadoop

1. **Scalability:** Hadoop can store and process petabytes of data efficiently.
2. **Fault Tolerance:** Automatically handles hardware failures.
3. **Cost-Effective:** Utilizes commodity hardware to store large datasets.
4. **Flexibility:** Can handle structured and unstructured data.
5. **Speed:** Distributes data across clusters, allowing for faster data processing.

Hadoop Architecture Overview

Hadoop architecture consists of the following main components:

1. **Hadoop Common**
2. **Hadoop Distributed File System (HDFS)**
3. **Yet Another Resource Negotiator (YARN)**
4. **MapReduce**

1. Hadoop Common

Hadoop Common is the set of shared utilities and libraries that support the other Hadoop modules. It includes essential Java libraries, scripts, and commands used throughout the Hadoop ecosystem. These components provide file system abstraction, IO utilities, and a platform for Hadoop modules to interact with the underlying operating system.

2. Hadoop Distributed File System (HDFS)

HDFS is the storage system of Hadoop. It is designed to store large datasets across a distributed cluster of machines. It provides high-throughput access to application data and is highly fault-tolerant.

Key Components of HDFS:

- **NameNode:**
    - **Role:** The NameNode is the master server that manages the file system namespace and controls access to files. It maintains the metadata of all the files and directories in the HDFS, such as the location of data blocks, permissions, and hierarchy.
    - **Functionality:** It keeps track of where data is stored across the cluster and manages operations such as opening, closing, and renaming files and directories.
    - **Fault Tolerance:** The NameNode is a single point of failure, so to improve fault tolerance, Hadoop provides a Secondary NameNode, which takes periodic snapshots of the NameNode's metadata. In newer versions, high availability is achieved using a standby NameNode.

Installation of Hadoop:

x

sudo apt update && sudo apt install openjdk-8-jdk

java -version

sudo apt install ssh

sudo adduser hadoop

su — hadoop

ssh-keygen -t rsa

cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

chmod 640 ~/.ssh/authorized_keys

ssh localhost

su — hadoop

wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz

tar -xvzf hadoop-3.3.6.tar.gz

mv hadoop-3.3.6 hadoop

nano ~/.bashrc

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

export HADOOP_HOME=/home/hadoop/hadoop

export HADOOP_INSTALL=$HADOOP_HOME

export HADOOP_MAPRED_HOME=$HADOOP_HOME

export HADOOP_COMMON_HOME=$HADOOP_HOME

export HADOOP_HDFS_HOME=$HADOOP_HOME

export HADOOP_YARN_HOME=$HADOOP_HOME

export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native

export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin

export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

```
source ~/.bashrc
```

Additionally, configure the JAVA_HOME variable in the hadoop-env.sh file. Open the Hadoop environment configuration file with a text editor:

```
nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
```

```
cd hadoop/

mkdir -p ~/hadoopdata/hdfs/{namenode,datanode}
```

Edit core-site.xml:

```
nano $HADOOP_HOME/etc/hadoop/core-site.xml
```

```
nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
```

Format the Namenode:

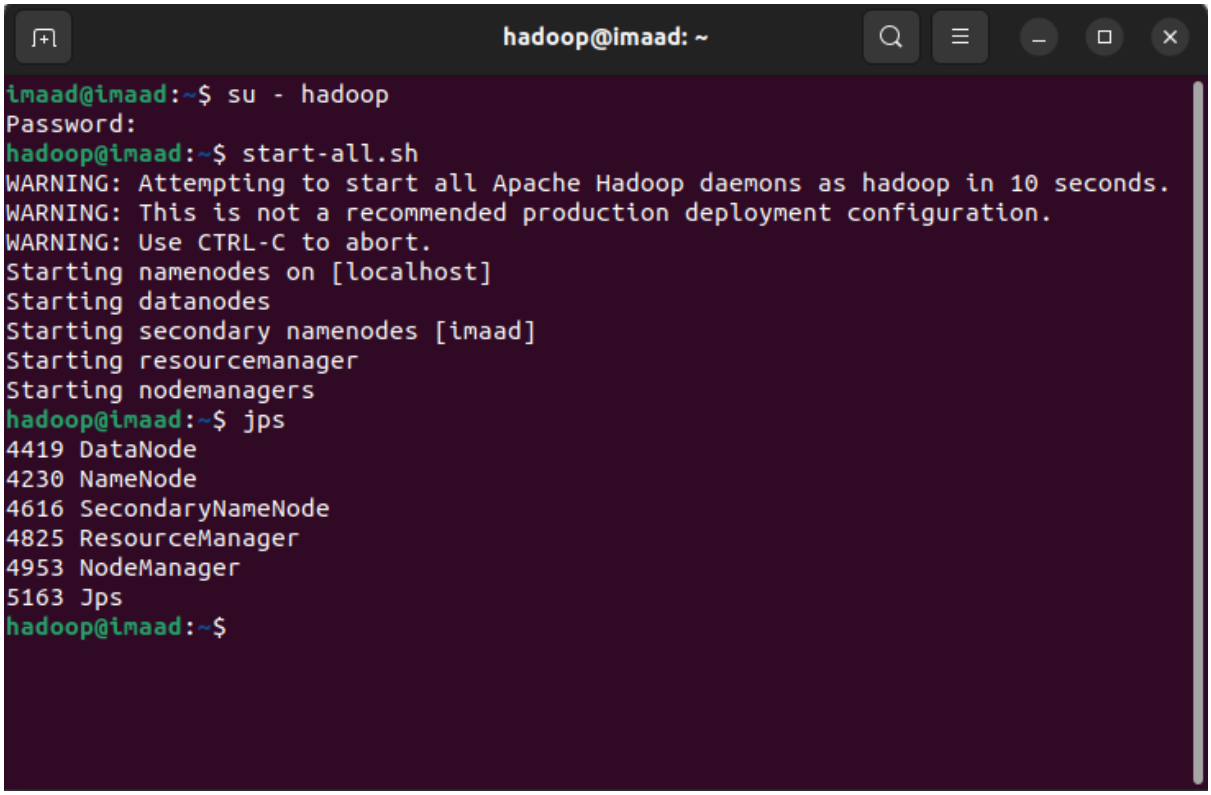
Format the Hadoop Namenode by running the following command:

```
hdfs namenode -format
```

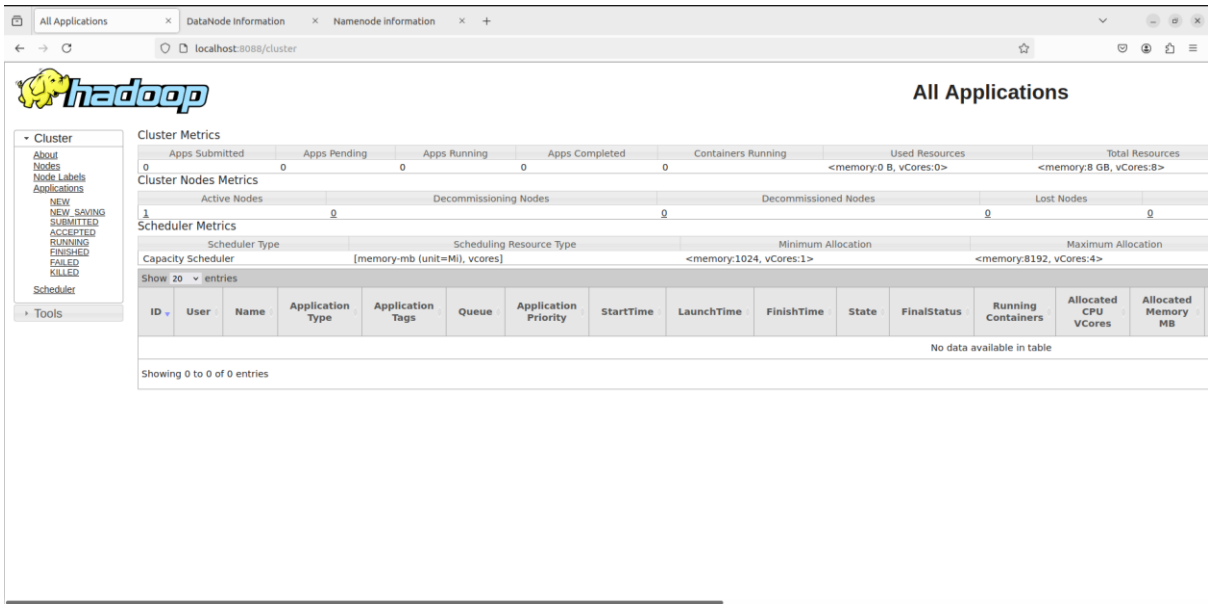
```
start-all.sh
```

> *jps*

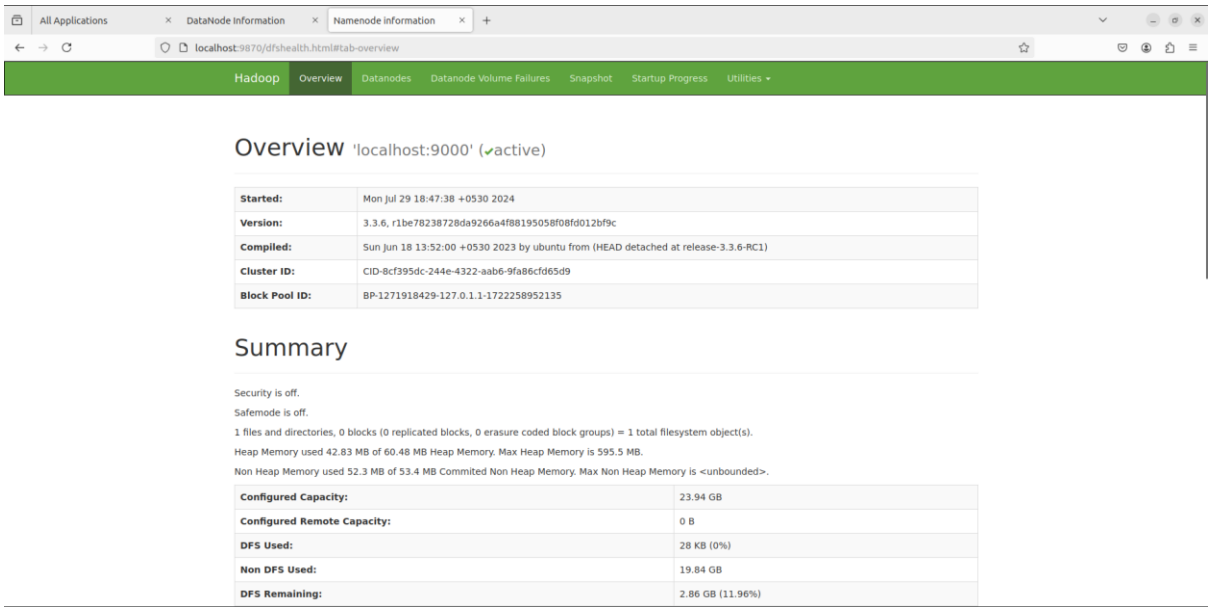
Screenshots:



Resource Manager:



Name Node:



Data Node: