**Setup Hadoop 3.4.1 on a Two-Node Cluster (Without YARN)**

This guide explains how to set up **Hadoop 3.4.1** on a **two-node cluster** using **VMware** with the following configuration:

* **Machine 1 (192.168.1.16 - vmware1)**: **NameNode + DataNode 1**
* **Machine 2 (192.168.1.20 - vmware2)**: **DataNode 2**
* **Machine 3 (192.168.1.21 - vmware3)**: **DataNode 3**

**1. System Configuration**

**1.1. Install Java**

Hadoop 3.4.1 requires **Java 11**. Install Java 11:

Do it on 2 machines

sudo apt update

sudo apt install openjdk-11-jdk -y

**1.2. Configure SSH (On Both Machines)**

**Create user for hadoopuser**

sudo adduser hadoopuser

sudo usermod -aG sudo hadoopuser

**Set Up SSH Key-Based Authentication (On NameNode Only)**

ssh-keygen -t rsa -P ""

cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

**Copy SSH Key to DataNode**

On **vmware1 (NameNode)**, run:

ssh-copy-id hadoopuser @192.168.1.16

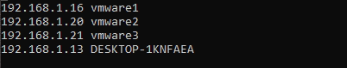
ssh-copy-id hadoopuser @192.168.1.20

ssh-copy-id hadoopuser @192.168.1.21

If you can log in without a password, SSH is configured correctly.

**1.3 Initial DNS Server (both machine)**

**Edit /etc/hosts:**



**2. Install Hadoop 3.4.1**

**2.1. Download and Extract Hadoop on Both Machines**

cd /opt

wget https://dlcdn.apache.org/hadoop/common/hadoop-3.4.1/hadoop-3.4.1.tar.gz

tar -xvzf hadoop-3.4.1.tar.gz

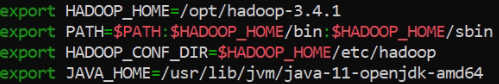
rm -rf hadoop-3.4.1.tar.gz

chown -R hadoopuser:hadoopuser /opt/hadoop-3.4.1/

su hadoopuser

**2.2. Set Environment Variables**

Add the following to ~/.bashrc:



* Save file ~/.bashrc

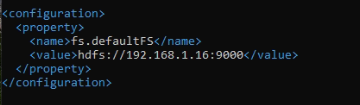
source ~/.bashrc

**3. Configure Hadoop**

All configurations should be made on **vmware1 (NameNode)** first and then copied to **vmware2 (DataNode) and vmware3(DataNode)**.

**3.1. Configure core-site.xml**

Edit $HADOOP\_HOME/etc/hadoop/core-site.xml on **vmware1 and vmware2 and vmware3**:



**When Do You Need core-site.xml?**

* When setting up **a Hadoop cluster** and defining the **NameNode**.
* When changing **security settings**.
* When optimizing **network performance and RPC communication**.

🔥 **core-site.xml is essential because it tells Hadoop where to store data and how system components communicate.**

**3.2. Configure hdfs-site.xml**

**What Does hdfs-site.xml Do?**

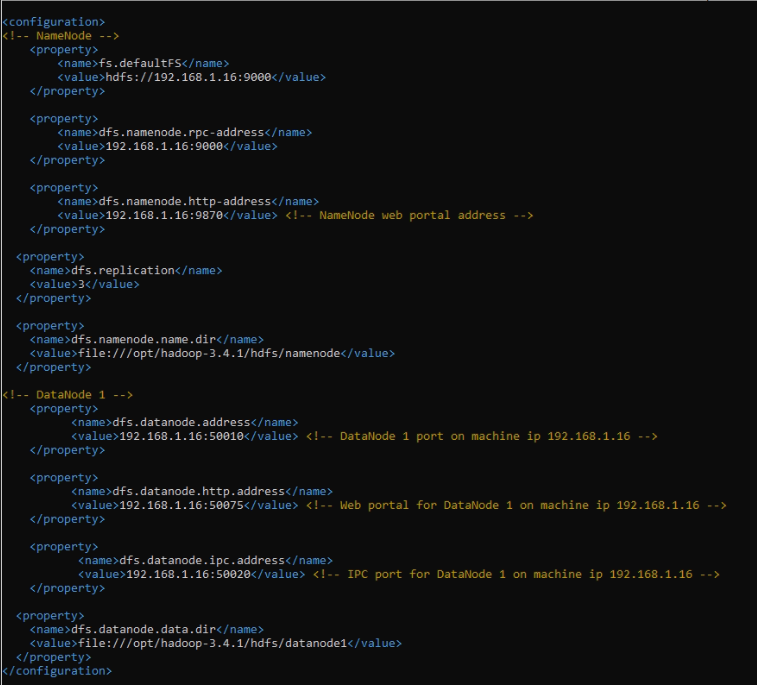
* Defines **HDFS storage locations** (NameNode & DataNode directories).
* Configures **replication factors** (how many copies of data blocks are stored across the cluster).
* Specifies **network addresses and ports** for NameNode and DataNodes.
* Sets up **HDFS communication** between NameNode and DataNodes.

Edit $HADOOP\_HOME/etc/hadoop/hdfs-site.xml on **vmware1**:

cd /opt/hadoop-3.4.1

mkdir -p hdfs/namenode

mkdir -p hdfs/datanode1



**Explanation**:

**NameNode**

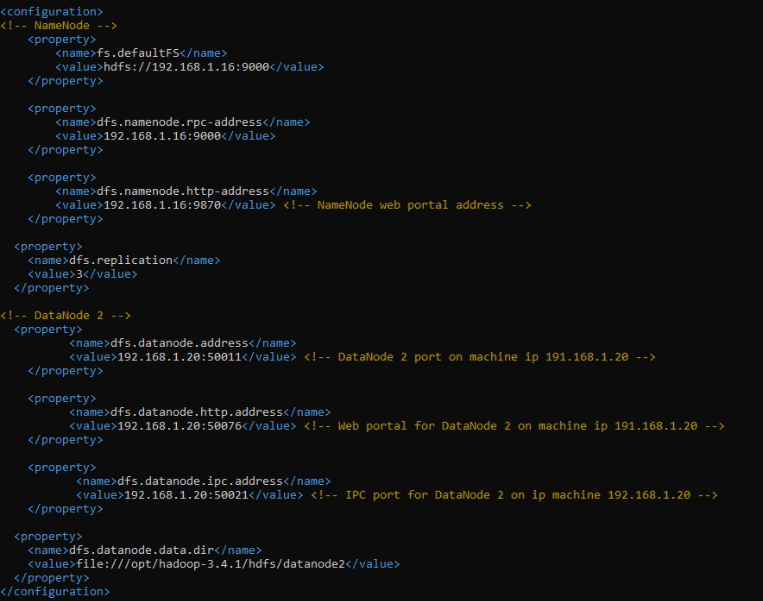
* **fs.defaultFS** → Defines the **HDFS NameNode URL** (hdfs://192.168.1.16:9000).
* **dfs.namenode.rpc-address** → RPC communication port for NameNode.
* **dfs.namenode.http-address** → Web UI for NameNode at **http://192.168.1.16:9870**.
* **dfs.replication** → Each file block will be **replicated 2 times** across DataNodes.
* **dfs.namenode.name.dir** → The local directory on the NameNode machine where HDFS metadata is stored (/opt/hadoop-3.4.1/hdfs/namenode).

**DataNode**

* **dfs.datanode.address** → DataNode **data transfer port (50010)**.
* **dfs.datanode.http-address** → Web UI for DataNode at **http://192.168.1.16:50075**.
* **dfs.datanode.ipc.address** → IPC (Inter-Process Communication) port for DataNode (50020).
* **dfs.datanode.data.dir** → Directory where **HDFS data blocks are stored** (/opt/hadoop-3.4.1/hdfs/datanode1).

Edit $HADOOP\_HOME/etc/hadoop/hdfs-site.xml on **vmware2**:

mkdir -p hdfs/datanode2



**Explanation**:

**NameNode**

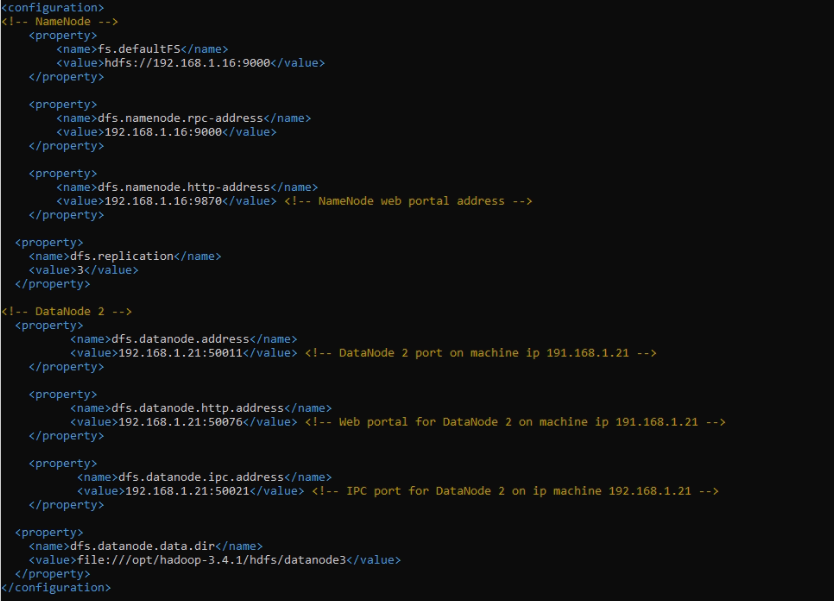
* **fs.defaultFS** → Defines the **HDFS NameNode URL** (hdfs://192.168.1.16:9000).
* **dfs.namenode.rpc-address** → RPC communication port for NameNode.
* **dfs.namenode.http-address** → Web UI for NameNode at **http://192.168.1.16:9870**.
* **dfs.replication** → Each file block will be **replicated 2 times** across DataNodes.
* **dfs.namenode.name.dir** → The local directory on the NameNode machine where HDFS metadata is stored (/opt/hadoop-3.4.1/hdfs/namenode).

**DataNode**

* **dfs.datanode.address** → Specifies the **DataNode IP and port (192.168.1.20:50011)**, used for **block storage communication**.
* **dfs.datanode.http-address** → Defines the **web UI for DataNode 2** at **http://192.168.1.20:50076**, where you can monitor its status.
* **dfs.datanode.ipc.address** → Sets the **Inter-Process Communication (IPC)** port (50021) for DataNode 2.
* **dfs.datanode.data.dir** → Specifies the **storage directory** (/opt/hadoop-3.4.1/hdfs/datanode2), where **HDFS stores actual data blocks** for this DataNode.

Edit $HADOOP\_HOME/etc/hadoop/hdfs-site.xml on **vmware3**:

mkdir -p hdfs/datanode3



**Explanation**:

**NameNode**

* **fs.defaultFS** → Defines the **HDFS NameNode URL** (hdfs://192.168.1.16:9000).
* **dfs.namenode.rpc-address** → RPC communication port for NameNode.
* **dfs.namenode.http-address** → Web UI for NameNode at **http://192.168.1.16:9870**.
* **dfs.replication** → Each file block will be **replicated 2 times** across DataNodes.
* **dfs.namenode.name.dir** → The local directory on the NameNode machine where HDFS metadata is stored (/opt/hadoop-3.4.1/hdfs/namenode).

**DataNode**

* **dfs.datanode.address** → Specifies the **DataNode IP and port (192.168.1.21:50011)**, used for **block storage communication**.
* **dfs.datanode.http-address** → Defines the **web UI for DataNode 3** at **http://192.168.1.21:50076**, where you can monitor its status.
* **dfs.datanode.ipc.address** → Sets the **Inter-Process Communication (IPC)** port (50021) for DataNode 3.
* **dfs.datanode.data.dir** → Specifies the **storage directory** (/opt/hadoop-3.4.1/hdfs/datanode3), where **HDFS stores actual data blocks** for this DataNode.

**3.3. Configure workers**

Edit $HADOOP\_HOME/etc/hadoop/workers on **vmware1 and vmware2**:

vmware1

vmware2

vmware3

**4. Format the NameNode**

Run the following command **only on NameNode (vmware1)**:

hdfs namenode -format

**5. Start HDFS**

Run the following command **on NameNode (vmware1)**:

start-dfs.sh

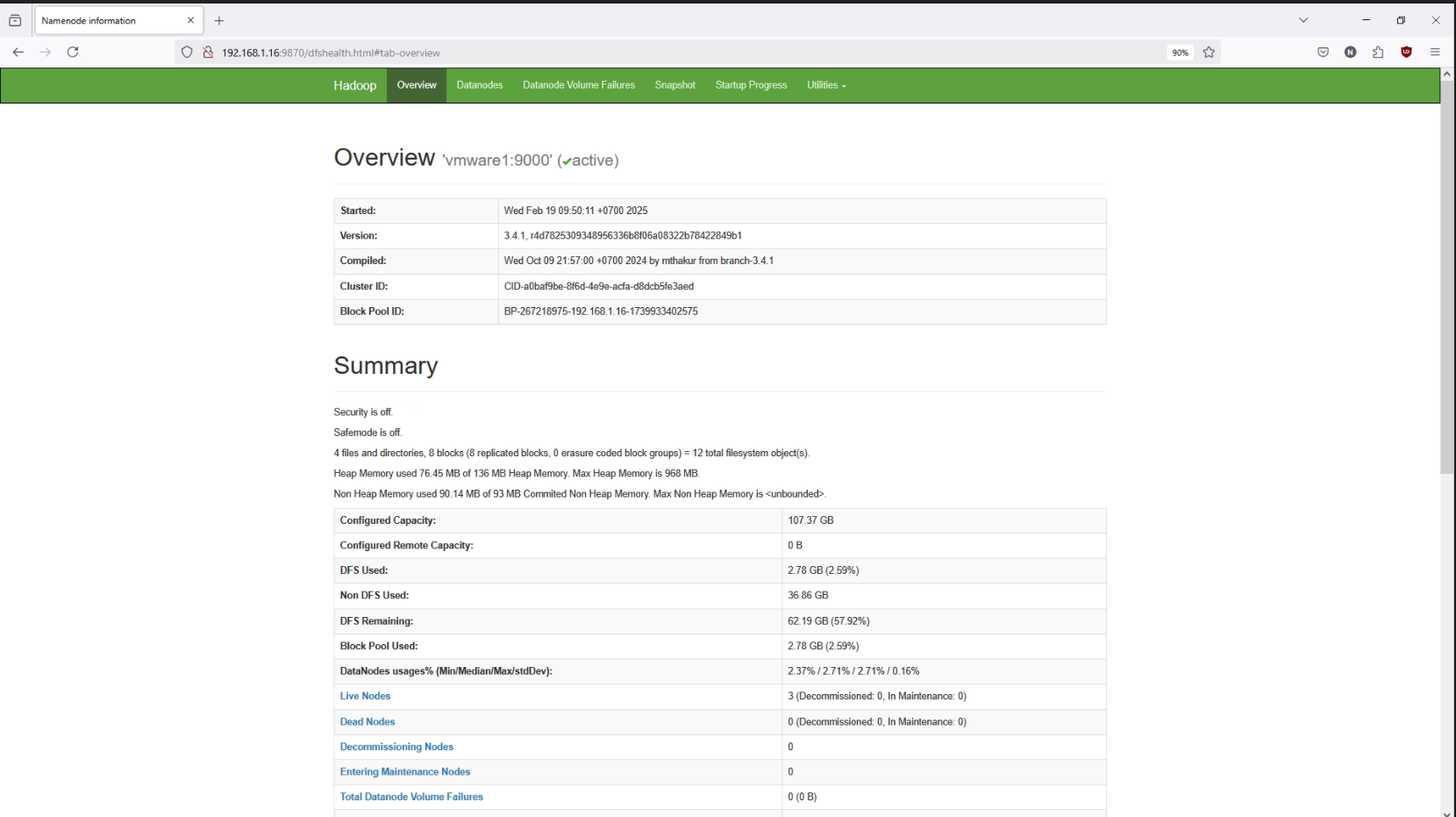
Check if DataNodes are connected:

hdfs dfsadmin -report

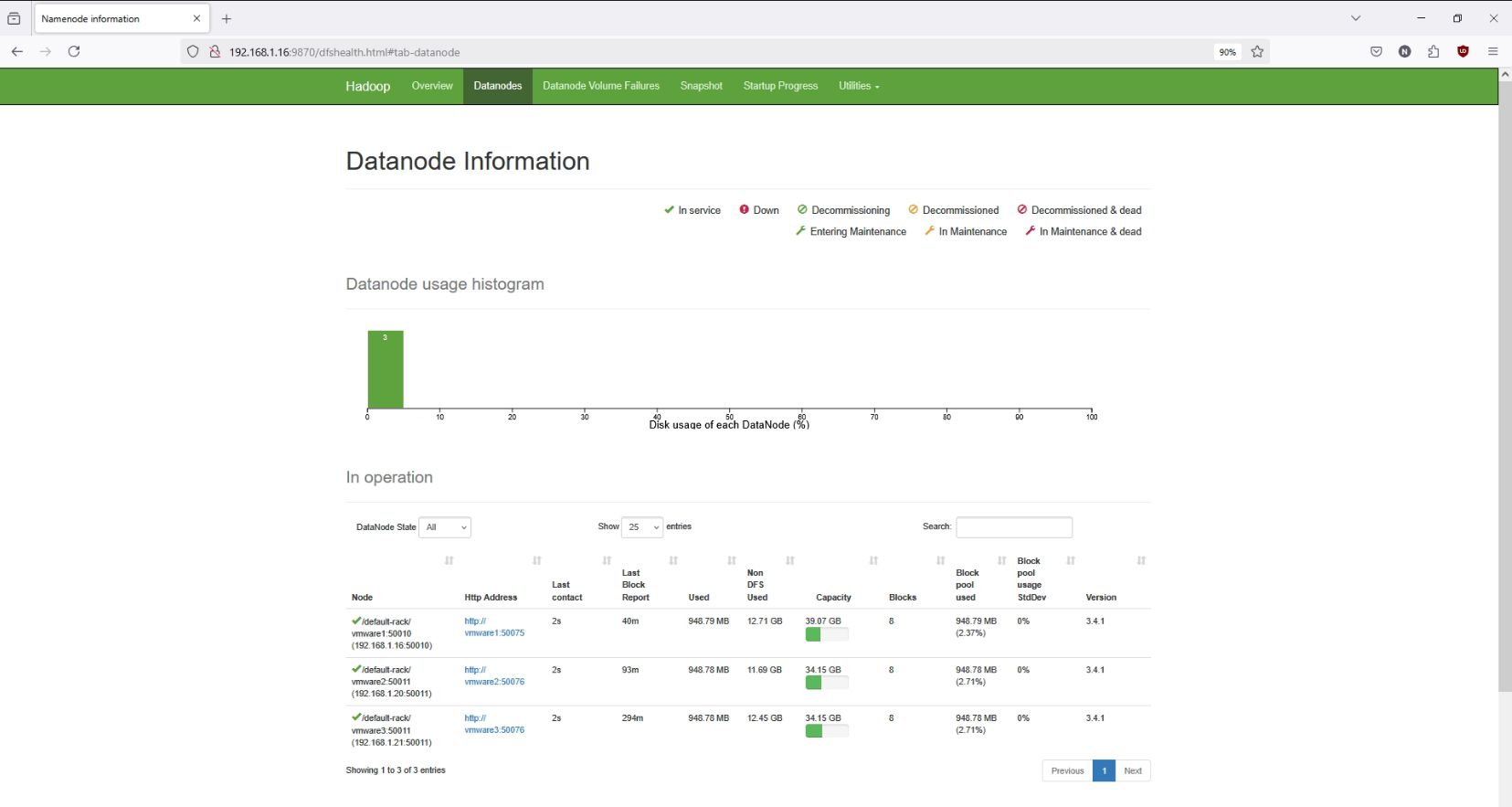
**6. Verify HDFS**

Check Hadoop Web UI:

* **NameNode UI**: [http://192.168.1.16:9870](http://192.168.1.15:9870)



* **DataNode UI** (for each node): <http://vmware1:50075> , <http://vmware2:50076> , <http://vmware3:50076>



**7. Stop HDFS**

To stop HDFS, run:

stop-dfs.sh

**8. Push file to system**

hdfs dfs -mkdir /data

hdfs dfs -put /home/hadoopuser/PS\_20174392719\_1491204439457\_log.csv /data/

hdfs dfs -put /home/hadoopuser/ Synthetic\_Financial\_datasets\_log.csv /data/

