Vacation Assignment

Installing Apache Hadoop on Ubuntu (Single-Node Setup Using Virtual Machine)

Course: 23CSE312 - Distributed Systems

Assignment Type: Hands-on Installation

Environment: Personal Laptop

Submission: Screenshots + Short Report + Demo of Working Hadoop

---

1. Objective of the Assignment

- Understand the basic environment required for Big Data processing
- Install and configure **Apache Hadoop** in **single-node mode**
- Gain hands-on experience with **Linux (Ubuntu)** and **HDFS**
- Verify Hadoop installation using basic commands and web UI

---

2. Software & System Requirements

Host System (Student Laptop)

- Operating System: **Windows 10 / 11**
- RAM: **Minimum 8 GB recommended**
- Free Disk Space: **At least 60 GB**

Software to be Installed

Software	Purpose
VirtualBox / VMware	Run Ubuntu as a virtual machine
Ubuntu 22.04 LTS	Linux OS
Java (OpenJDK 8)	Required for Hadoop
Apache Hadoop 3.4.2	Big Data framework

---

3. Virtual Machine Setup (Ubuntu Installation)

Step 1: Install VirtualBox (on Windows)

1. Download from: https://www.virtualbox.org
2. Install VirtualBox and Extension Pack
3. Restart system if required

Step 2: Download Ubuntu ISO

- Download **Ubuntu 22.04 LTS Desktop ISO**
- Link: https://ubuntu.com/download/desktop

Step 3: Create Ubuntu Virtual Machine

1. Open VirtualBox → Click **New**
2. Name: `Ubuntu-Hadoop`
3. Type: `Linux`
4. Version: `Ubuntu (64-bit)`
5. RAM: **4096 MB minimum (8 GB recommended)**
6. CPU: **2 cores minimum**
7. Disk: **50 GB (dynamically allocated)**
8. Attach Ubuntu ISO and start VM

Step 4: Install Ubuntu (Inside VM)

- Click **Install Ubuntu**
- Normal installation
- Enable updates and third-party software
- Create user:
- `Username: hadoop`
- `Password: (remember this)`

---

4. Open Terminal in Ubuntu

After installation:

- Press **Ctrl + Alt + T**
- All commands below must be typed in the **Ubuntu terminal**

---

5. System Update

```
sudo apt update && sudo apt upgrade -y
```

---

6. Install Java (Required for Hadoop)

```
sudo apt install openjdk-8-jdk -y
```

Verify:

```
java -version
```

---

7. Download and Install Hadoop (Latest Stable)

Hadoop Version Used

Apache Hadoop 3.4.2

```
wget https://downloads.apache.org/hadoop/core/hadoop-3.4.2/hadoop-
3.4.2.tar.gz
tar -xvzf hadoop-3.4.2.tar.gz
mv hadoop-3.4.2 hadoop
```

---

8. Set Environment Variables

```
nano ~/.bashrc
```

Add at the bottom:

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

Save → Exit → Apply:

```
source ~/.bashrc
```

Verify:

```
hadoop version
```

---

9. Hadoop Configuration Files

All files are located in:

```
~/hadoop/etc/hadoop/
```

9.1 hadoop-env.sh

```
nano $HADOOP_CONF_DIR/hadoop-env.sh
```

Set:

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

9.2 core-site.xml

```
nano $HADOOP_CONF_DIR/core-site.xml
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

9.3 hdfs-site.xml

```
nano $HADOOP_CONF_DIR/hdfs-site.xml
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>

  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/home/hadoop/hadoopdata/namenode</value>
  </property>

  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:/home/hadoop/hadoopdata/datanode</value>
  </property>
```

```
</configuration>
```

Create directories:

```
mkdir -p ~/hadoopdata/namenode
mkdir -p ~/hadoopdata/datanode
```

9.4 mapred-site.xml

```
cd $HADOOP_CONF_DIR
cp mapred-site.xml.template mapred-site.xml
nano mapred-site.xml
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```

9.5 yarn-site.xml

```
nano $HADOOP_CONF_DIR/yarn-site.xml
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
</configuration>
```

---

10. Configure SSH

```
sudo apt install openssh-server -y
ssh-keygen -t rsa -P ""
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
ssh localhost
exit
```

---

11. Format NameNode (One Time Only)

```
hdfs namenode -format
```

---

12. Start Hadoop Services

```
start-dfs.sh
```

```
start-yarn.sh
```

Verify:

```
jps
```

---

13. Hadoop Web Interfaces

Open browser **inside VM or host system**:

Service	URL
NameNode	http://localhost:9870
ResourceManager	http://localhost:8088

---

14. Test Hadoop (HDFS)

```
hdfs dfs -mkdir /input
hdfs dfs -put ~/.bashrc /input
hdfs dfs -ls /input
```

---

15. Stop Hadoop

```
stop-yarn.sh
stop-dfs.sh
```

---

16. Assignment Submission Guidelines

Each student must submit:

1. **Screenshots** of:
    - Hadoop version command
    - `jps` output
    - NameNode Web UI
    - HDFS file listing
2. **Short Report (2–3 pages)** including:
    - Objective
    - System configuration
    - Steps followed
    - Issues faced (if any)
    - Conclusion

17. Learning Outcomes

- Install Linux on a virtual machine
- Configure Hadoop in single-node mode
- Understand HDFS and YARN basics
- Prepare environment for Big Data processing