BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

Autonomous Institute under VTU, Belagavi, Karnataka - 590 018

Yelahanka, Bengaluru, Karnataka - 560 119

Cloud Computing (BCS601) CCA Report

On

"Hadoop Installation and Execution of Simple Application on it"

BACHELOR OF ENGINEERING

in

INFORMATION SCIENCE AND ENGINEERING

by

NAME : Manjunath B L USN : 1BY23IS407

Under the Guidance of

Dr. Srinivas B V

Assistant Professor,

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

Avalahalli, Yelahanka, Bengaluru, Karnataka -560119

May 2025

Evaluation sheet

Name	USN	Environment Setup	Application Implementation	Documentation	Understanding and Analysis	Innovation/ Extra Work	Report	Total
		3	2	5	2	3	5	20
Manjunath B L	1BY23IS407							

	Name	Signature with date
Course Coordinator	Dr. Srinivas B V	
Course Co-Coordinator	Prof. Amulya P(Batch-2)/Prof. Sonne Gowda(Batch-1)	

Content Sheet

1. Introduction

Big data technologies have revolutionized the way we store, process, and analyze large volumes of information. Apache Hadoop is one such powerful framework that enables distributed storage and processing of data using commodity hardware. This report demonstrates the step-by-step procedure to install and configure a single-node Hadoop cluster on a cloud-based Ubuntu system, culminating with the execution of a sample MapReduce application — Word Count.

2. System Requirements

To implement a Hadoop single-node cluster, the following requirements were considered:

- **Operating System**: Pop!_OS 22.04 LTS
- **Memory (RAM)**: Minimum 4 GB
- **Disk Space**: Minimum 20 GB
- **User Privileges**: sudo/root access
- **Software Dependencies**:
    - Java (OpenJDK 8)
    - SSH server
    - Hadoop 3.4.1

3. Hadoop Overview

Hadoop is an open-source framework by Apache that allows for the distributed processing of large data sets across clusters using a simple programming model. Its core components include:

- **HDFS (Hadoop Distributed File System)**: For reliable, distributed storage.
- **YARN (Yet Another Resource Negotiator)**: For cluster resource management.
- **MapReduce**: For parallel data processing.
- **Common Utilities**: Shared tools required by Hadoop modules.

4. Prerequisites Installation

- **Add New User:**

  sudo adduser hadoop_user

  sudo usermod -aG sudo hadoop_user

- **Install Java:**

  sudo apt update

  sudo apt install openjdk-8-jdk -y

- **Verify Java Installation:**

  java -version

- **Install SSH and Configure Passwordless SSH:**

  ssh-keygen

  cat ~/.ssh/id_ed25519.pub >> ~/.ssh/authorized_keys

  ssh localhost

5. Downloading and Installing Hadoop

- **Download Hadoop**:

  zip-file: https://downloads.apache.org/hadoop/common/hadoop-3.4.1/hadoop- 3.4.1.tar.gz

- **Extract and Move**:

  tar -zxvf hadoop-3.4.1.tar.gz

  sudo mv hadoop-3.4.1 /usr/local/hadoop

- **Set Environment Variables in .bashrc**:

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

  export HADOOP_HOME=/usr/local/hadoop

  exportPATH=$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

  export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop

6. Configuring Hadoop for Single Node Cluster

- **core-site.xml**: Set default filesystem to HDFS.

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

- **hdfs-site.xml**: Specified namenode and datanode directories.

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

- **mapred-site.xml**: Set framework name to yarn.

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>yarn.app.mapreduce.am.env</name>
    <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
  </property>
  <property>
    <name>mapreduce.map.env</name>
    <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
  </property>
  <property>
    <name>mapreduce.reduce.env</name>
    <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
  </property>
```

```
</configuration>
```

- **yarn-site.xml**: Configured resource manager and node manager.

```
<configuration>
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>localhost</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
</configuration>
```

7. Starting Hadoop Daemons

- **Started the required services:**

  start-dfs.sh
  start-yarn.sh

- Verified with:

  jps

- Expected output included:

  NameNode
  DataNode
  ResourceManager
  NodeManager
  SecondaryNameNode.

8. Running Sample Application: Word Count

- **Create Input File**:

  echo "The quick brown fox jumps over the lazy dog. The dog barked back at the fox. Quick movements startled the birds nearby." > input.txt

- **Upload File to HDFS**:

  hadoop fs -mkdir /input

  hadoop fs -put input.txt /input

- **Run Word Count Job**:

  hadoop jar wordcount.jar WordCount /input /output

9. Output and Verification

- **To verify the results of the Word Count job:**

  hadoop fs -cat /output/part-r-00000

- **Output:**

Figure 1: Output of Word Count application

This confirms that Hadoop processed the input file and outputted the word frequency count.

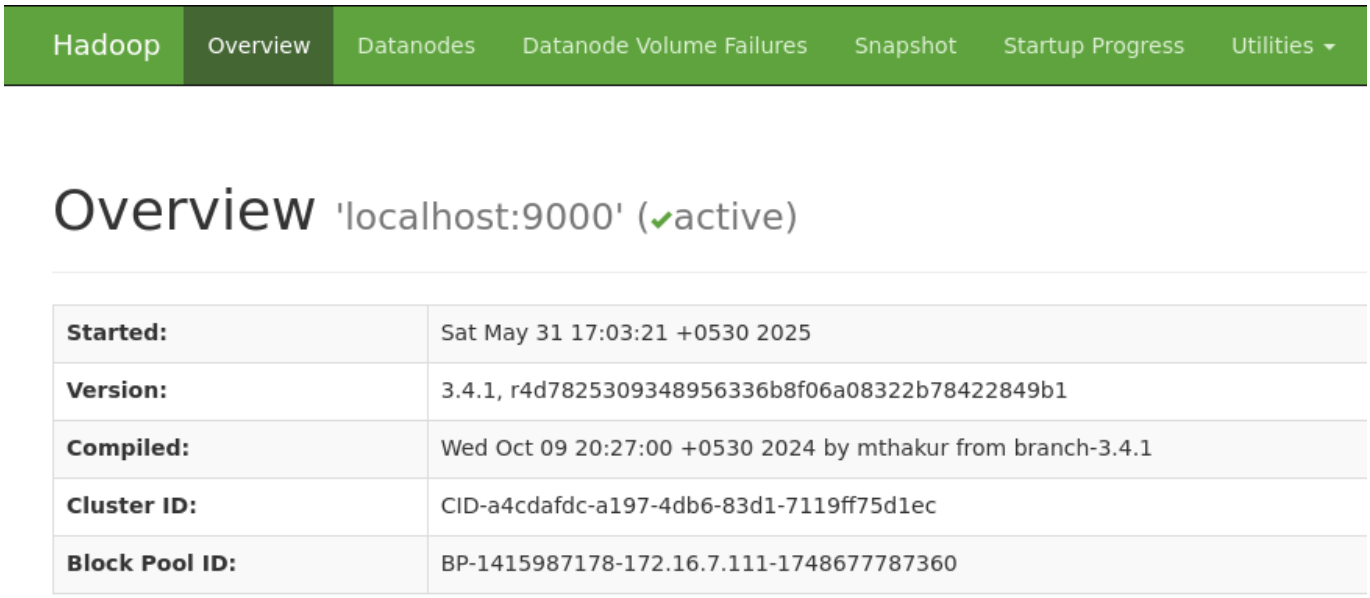

Figure 2: Hadoop Web UI - NameNode Overview

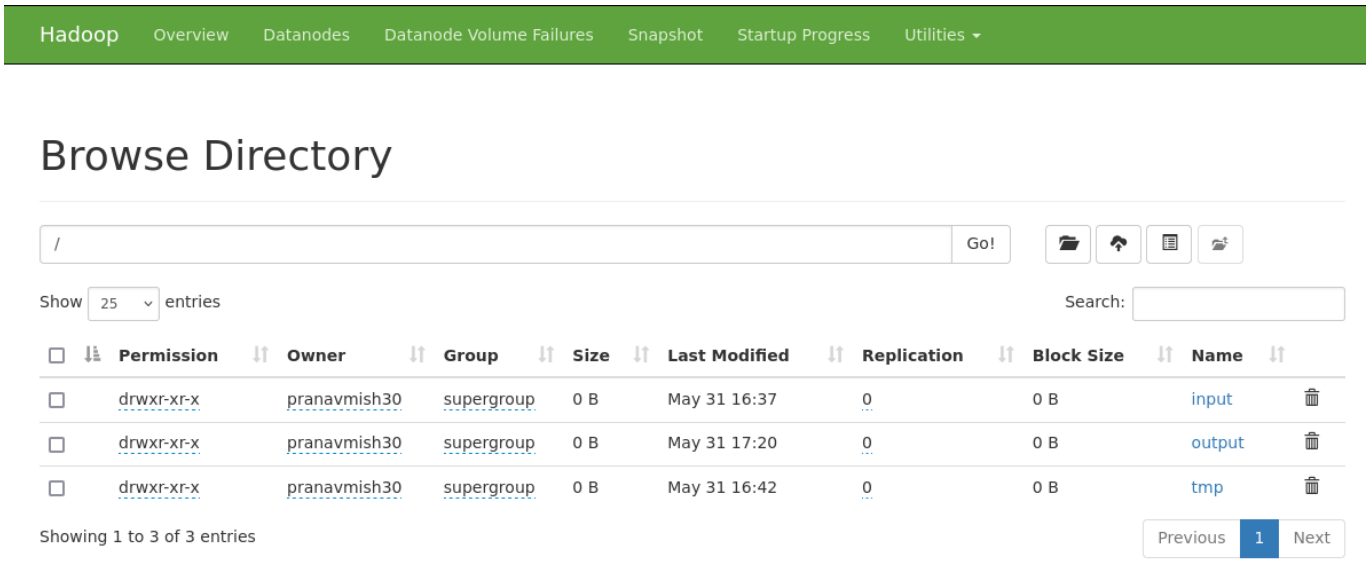

Figure 3: *HDFS Directory Structure via Hadoop Web UI*

10. Conclusion

The successful deployment of a Hadoop single-node cluster on a cloud-based Ubuntu environment demonstrates a practical understanding of distributed data processing frameworks. Through the systematic installation of Java, SSH configuration, environment setup, and Hadoop installation, this report illustrates how foundational components such as HDFS and YARN are integrated and orchestrated to process large-scale data. The execution of a sample MapReduce job, Word Count, further validates the operational readiness and correctness of the cluster.

This implementation not only enhances technical proficiency in configuring and managing big data infrastructure but also lays a strong foundation for scaling towards multi-node clusters and more complex data processing tasks. Mastery of such a setup is a critical step for professionals aiming to work in big data engineering, cloud computing, or data-driven application development.

11. References

- Apache Hadoop Official Documentation
- OpenJDK Documentation
- Ubuntu SSH Configuration Guide
- Google Cloud Compute Engine Documentation
- https://www.youtube.com/watch?v=Sk2ImFdH2uc
- https://codewitharjun.medium.com/install-hadoop-on-ubuntu-operating-system-6e0ca4ef9689
- https://www.youtube.com/watch?v=qgBu8Go1SyM
- https://www.youtube.com/watch?v=Slbi-uzPtnw
- https://www.youtube.com/watch?v=inDC9jgwpWY