Task 1

Description:

To configure a basic Distributed File System using HDFS on AWS EC2.

Tools Required

- AWS EC2
- Java
- Hadoop (Single-node setup)

Procedure:

1. Launch an EC2 instance
2. Install Java and Hadoop
3. Configure NameNode and DataNode
4. Start HDFS services
5. Verify HDFS using web UI

Expected Outcome

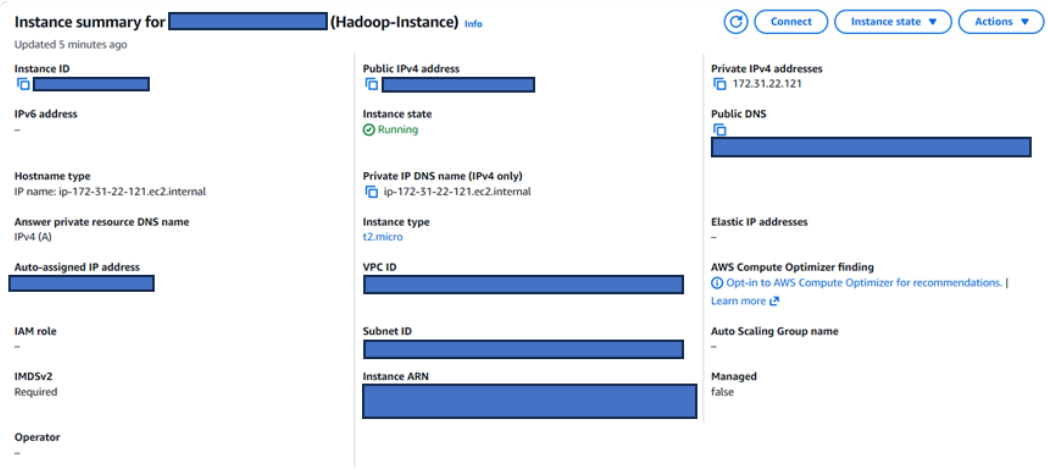
HDFS is successfully configured on AWS EC2.

Prerequisites:

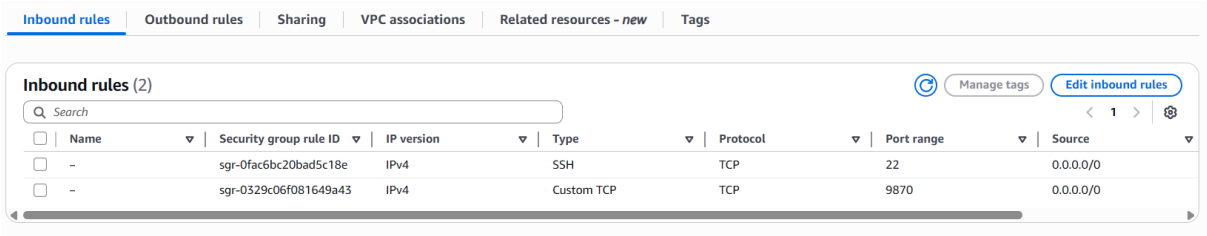
- Active AWS Academy account
  - Access to AWS Management Console
  - EC2 instance (Amazon Linux 2, t2.micro)
- Security Group configured with:
  - SSH (Port 22)
  - Custom TCP (Port 9870)
  - Key pair (.pem file) for SSH access
- Software Requirements:
  - Java (OpenJDK 8 or 11)
  - Hadoop (Version 3.x)

Process:

1. **Launch EC2 Instance**
   - Created an EC2 instance using Amazon Linux 2 AMI.
   - Selected t2.micro instance type.



- Configured Security Group to allow ports 22 and 9870.



- Connected to the instance using SSH.

2. **Install Java**
   - Installed OpenJDK on the EC2 instance.
   - Verified installation using:

     java -version

```
[hadoop@ip-172-31-22-121 ~]$ java -version
openjdk version "1.8.0_482"
OpenJDK Runtime Environment Corretto-8.482.08.1 (build 1.8.0_482-b08)
OpenJDK 64-Bit Server VM Corretto-8.482.08.1 (build 25.482-b08, mixed mode)
[hadoop@ip-172-31-22-121 ~]$
```

3. **Install Hadoop**
   - Downloaded Hadoop 3.x package.
   - Extracted and configured Hadoop environment variables.
   - Verified installation using:

     hadoop version

```
[hadoop@ip-172-31-22-121 ~]$ hadoop version
Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da9266a4f88195058f08fd012bf9c
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bbd
This command was run using /home/hadoop/hadoop/share/hadoop/common/hadoop-common-3.3.6.jar
[hadoop@ip-172-31-22-121 ~]$
```

4. **Configure Hadoop**

- Edited configuration files:
  - core-site.xml (set fs.defaultFS)

```
[hadoop@ip-172-31-22-121 ~]$ cat $HADOOP_HOME/etc/hadoop/core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>fs.defaultFS</name>
  <value>hdfs://localhost:9000</value>
</property>
</configuration>
[hadoop@ip-172-31-22-121 ~]$
```

  - hdfs-site.xml (set replication factor and storage paths)
- Created NameNode and DataNode directories.

5. **Configure SSH**
   - Generated SSH keys for passwordless login.
   - Enabled SSH access to localhost for Hadoop services.
6. **Format NameNode**
   - Initialized HDFS using:

     hdfs namenode -format
7. **Start HDFS Services**
   - Started NameNode and DataNode using:

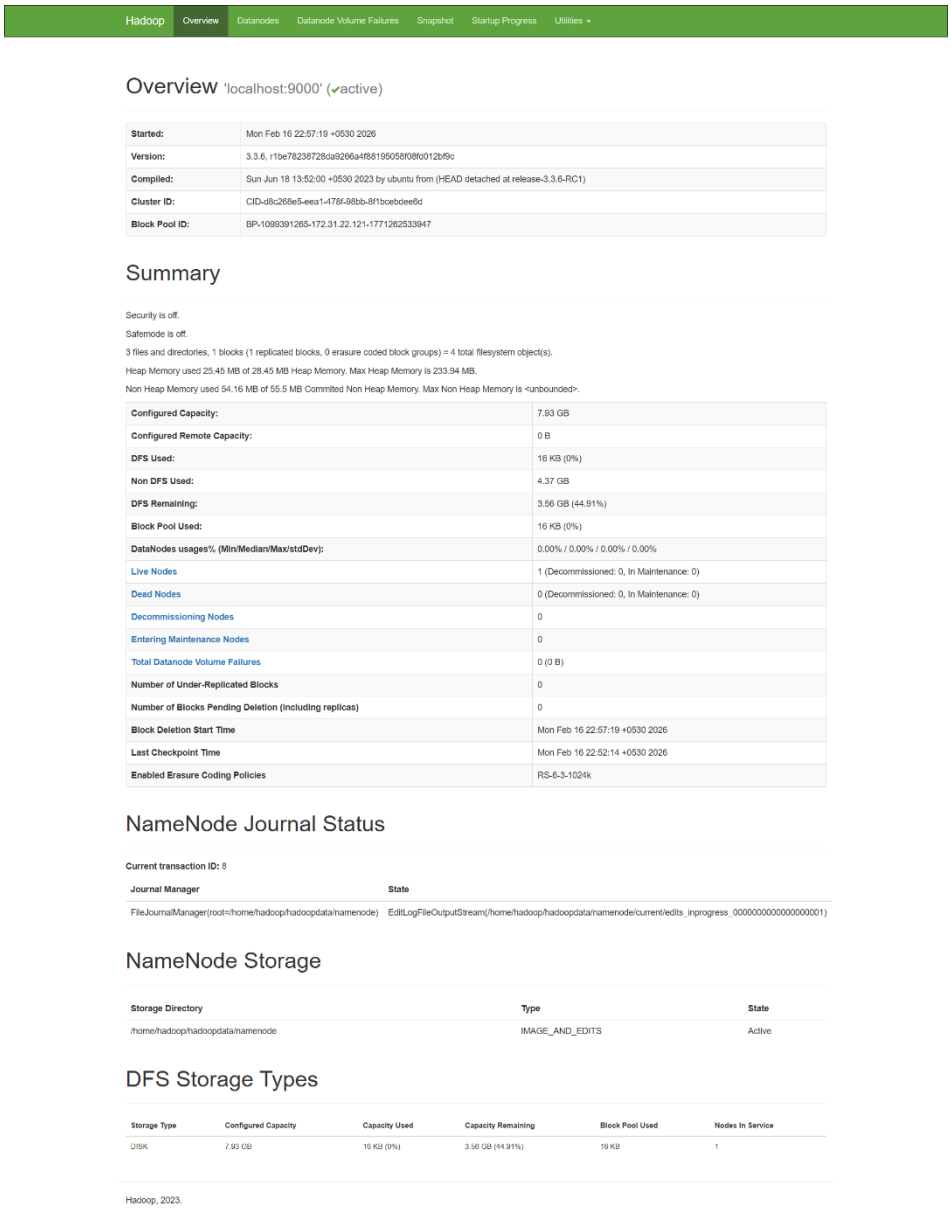
     start-dfs.sh
   - Verified services using:

     Jps

```
[hadoop@ip-172-31-22-121 ~]$ jps
29445 Jps
28615 NameNode
28715 DataNode
28910 SecondaryNameNode
[hadoop@ip-172-31-22-121 ~]$
```

8. **Verify HDFS via Web UI**

- Accessed Hadoop dashboard through:
  http://<Public-IP>:9870
- Confirmed cluster status and active DataNode.



Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities

Overview 'localhost:9000' (active)

Started:	Mon Feb 16 22:57:19 +0530 2026
Version:	3.3.6, r1be78238728da9266a4f8819505 8f08fd012bf9c
Compiled:	Sun Jun 18 13:52:00 +0530 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-d8c268e5-eea1-478f-98bb-8f1bcebdee6d
Block Pool ID:	BP-1098391265-172.31.22.121-1771262533947

Summary

Security is off.
Safemode is off.
3 files and directories, 1 blocks (1 replicated blocks, 0 erasure coded block groups) = 4 total filesystem object(s).
Heap Memory used 25.45 MB of 28.45 MB Heap Memory. Max Heap Memory is 233.94 MB.
Non Heap Memory used 54.16 MB of 55.5 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	7.93 GB
Configured Remote Capacity:	0 B
DFS Used:	16 KB (0%)
Non DFS Used:	4.37 GB
DFS Remaining:	3.56 GB (44.91%)
Block Pool Used:	16 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Mon Feb 16 22:57:19 +0530 2026
Last Checkpoint Time	Mon Feb 16 22:52:14 +0530 2026
Enabled Erasure Coding Policies	RS-6-3-1024k

NameNode Journal Status

Current transaction ID: 8

Journal Manager	State
FileJournalManager(root=/home/hadoop/hadoopdata/namenode)	EditLogFileOutputStream(/home/hadoop/hadoopdata/namenode/current/edits_inprogress_0000000000000000001)

NameNode Storage

Storage Directory	Type	State
/home/hadoop/hadoopdata/namenode	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	7.93 GB	16 KB (0%)	3.56 GB (44.91%)	16 KB	1

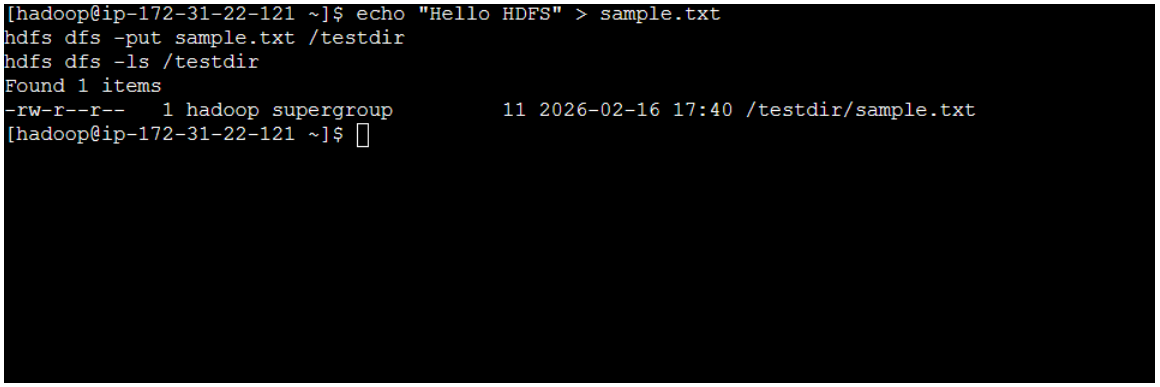
Hadoop, 2023.

9. **Test HDFS Functionality**
   - Created directory in HDFS:
     hdfs dfs -mkdir /testdir

```
[hadoop@ip-172-31-22-121 ~]$ hdfs dfs -mkdir /testdir
hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - hadoop supergroup          0 2026-02-16 17:38 /testdir
[hadoop@ip-172-31-22-121 ~]$
```

- Uploaded a sample file:
  hdfs dfs -put sample.txt /testdir

- Verified file upload using:
  hdfs dfs -ls /testdir



Output:

- EC2 instance successfully launched and running.
- Java installed and verified.
- Hadoop installed and configured properly.
- NameNode and DataNode services started successfully.
- HDFS Web UI accessed through port 9870.
- Directory successfully created in HDFS.
- File uploaded and verified inside HDFS.