**DATA ANALYTICS**

**ASSIGNMENT-1**

` **Introduction to Hadoop:**

Apache Hadoop is a robust, open-source framework designed to facilitate the distributed storage and processing of massive datasets across clusters of commodity hardware. The Hadoop ecosystem comprises several key components, including the Hadoop Distributed File System (HDFS) for reliable data storage, MapReduce for parallel processing, YARN for resource management, and a suite of tools like Hive, Pig, and Spark for enhanced data analysis and querying.

**1.1 History of Hadoop:**

Apache Hadoop, an open-source framework, was created in the mid-2000s by Doug Cutting and Mike Cafarella. Inspired by Google's MapReduce and Google File System (GFS) papers, it was developed to handle large-scale data processing and storage across distributed systems. Officially launched in 2006 at Yahoo!, Hadoop revolutionized big data management with its scalable and fault-tolerant architecture. Over time, Hadoop's ecosystem expanded to include core components like HDFS, YARN, and MapReduce, along with tools such as Hive and Pig, solidifying its role as a fundamental technology in modern data analytics

**1.2 Versions of Hadoop:**

Hadoop has seen numerous versions since its inception. Here are some key versions:

Hadoop 0.20.x (2009)

Hadoop 1.x (2011)

Hadoop 2.x (2013)

Hadoop 3.x (2017)

Hadoop 3.1 and 3.2

Hadoop 3.3 (2020)

Hadoop 3.4 and beyond (future directions)

**1.3 System Requirements for Hadoop (All OS)**

**Hardware Requirements:**

**Minimum Requirements:**

1. Memory (RAM)

- Minimum: 8 GB per node.

2. Storage

- Minimum: 500 GB per node.

3. CPU

- Minimum: Quad-core processors.

4. Network

- High-speed network (1 Gbps or higher) for inter-node communication.

**Software Requirements:**

1. Operating System

- Linux-based OS (e.g., CentOS, Ubuntu, Red Hat Enterprise Linux) is recommended for stability and compatibility.

2. Java

- Java Development Kit (JDK) 8 or later versions. Hadoop is Java-based, so a compatible JDK is essential.

3. SSH

- Secure Shell (SSH) must be set up for password-less login between nodes in the cluster.

4. Hadoop Distribution

- Download the latest stable release of Hadoop from the Apache Hadoop website or use distributions from vendors like Cloudera, Hortonworks, or MapR, which provide additional tools and support.

**Configuration Considerations:**

1. Cluster Size

- Determine the number of nodes based on data size and processing needs. A small cluster might consist of 4-10 nodes, while large clusters can have hundreds of nodes.

2. Master and Slave Nodes

- Master nodes (NameNode, ResourceManager) require more memory and CPU resources.

- Slave nodes (DataNode, NodeManager) focus on storage and processing, so ensure they have ample disk space and sufficient memory.

3. Network Configuration

- Proper network configuration is crucial for Hadoop's performance. Ensure low-latency and high-bandwidth connections between nodes.

* 1. **Installation Steps One by One with Commands with Explanation**

1.Install Java: Ensure Java Development Kit (JDK) is installed, as Hadoop requires Java to run.

2. \*Create a Hadoop User\*: Create a dedicated user for running Hadoop, and switch to this user for the installation process.

3. \*Install SSH and Configure Password-less Login\*: Install SSH to enable remote management of Hadoop nodes, and set up password-less SSH login for seamless operations.

4. \*Download and Extract Hadoop\*: Obtain Hadoop 3.3.6 from the official Apache website, then extract and move it to a suitable directory.

5. \*Configure Environment Variables\*: Set environment variables to point to your Hadoop installation, ensuring that Hadoop commands are accessible from the command line.

6. \*Configure Hadoop\*: Edit key configuration files to set up core Hadoop properties, including file system settings, replication factors, and resource management.

7. \*Create Hadoop Directories\*: Establish necessary directories for Hadoop's file system metadata and data storage.

Command: .mkdir -p /home/hadoop/hadoop\_data/hdfs/namenode

mkdir -p /home/hadoop/hadoop\_data/hdfs/datanode

8. \*Format HDFS\*: Initialize the Hadoop Distributed File System (HDFS) to prepare it for storing data.

Command: hdfs namenode -format

9. \*Start Hadoop Daemons\*: Launch Hadoop services for HDFS and YARN (Yet Another Resource Negotiator) to start the cluster.

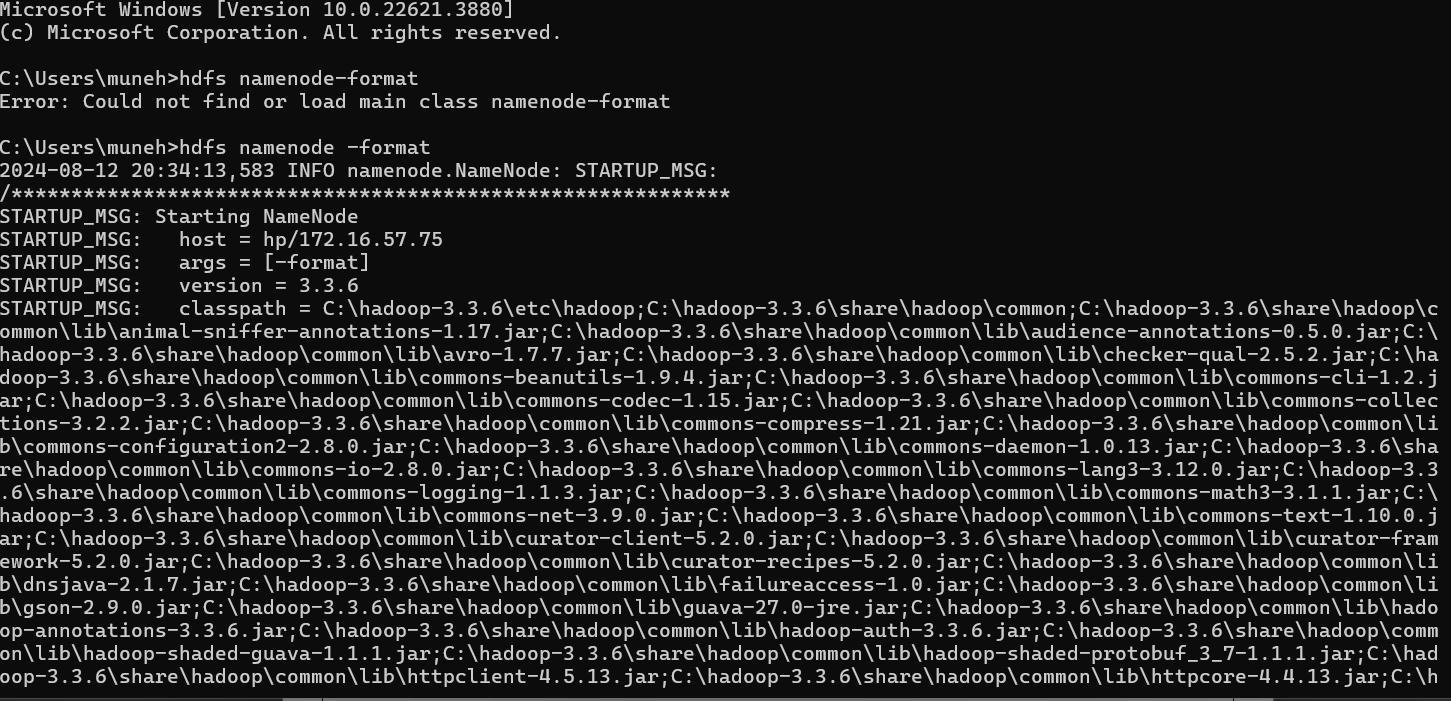
Command: start-dfs.sh

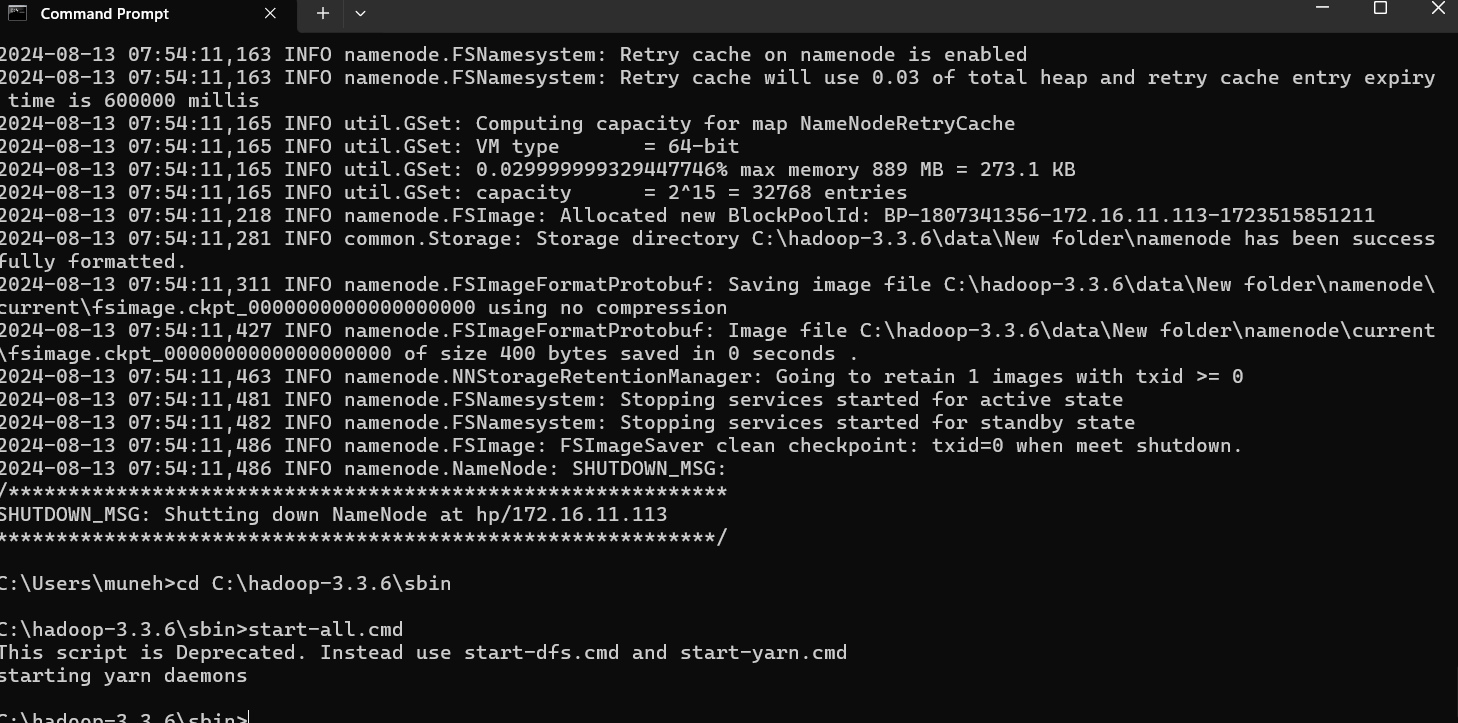
Start-yarn.sh

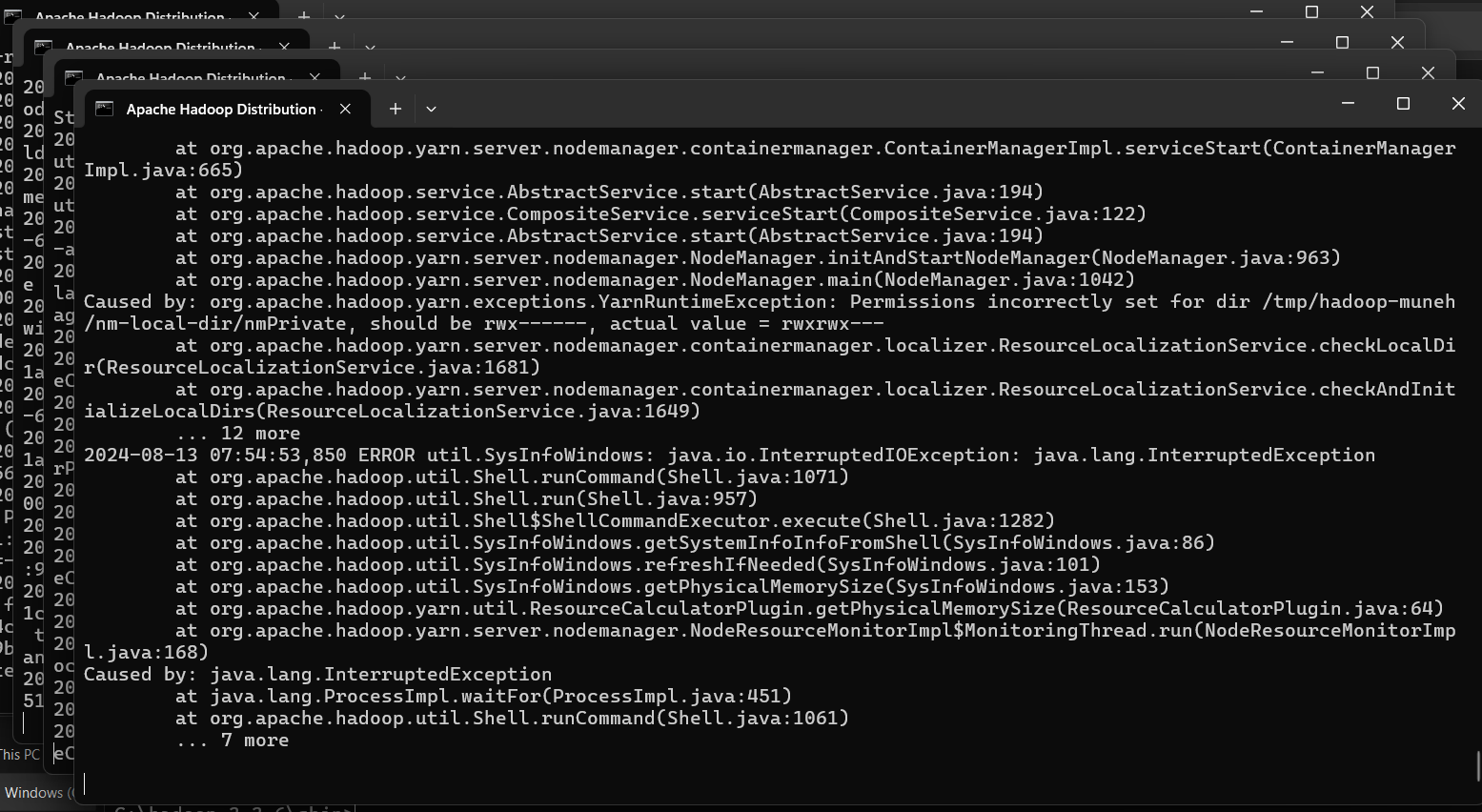
10. \*Verify Installation\*: Check that Hadoop daemons are running correctly and verify the setup through web interfaces provided by Hadoop.

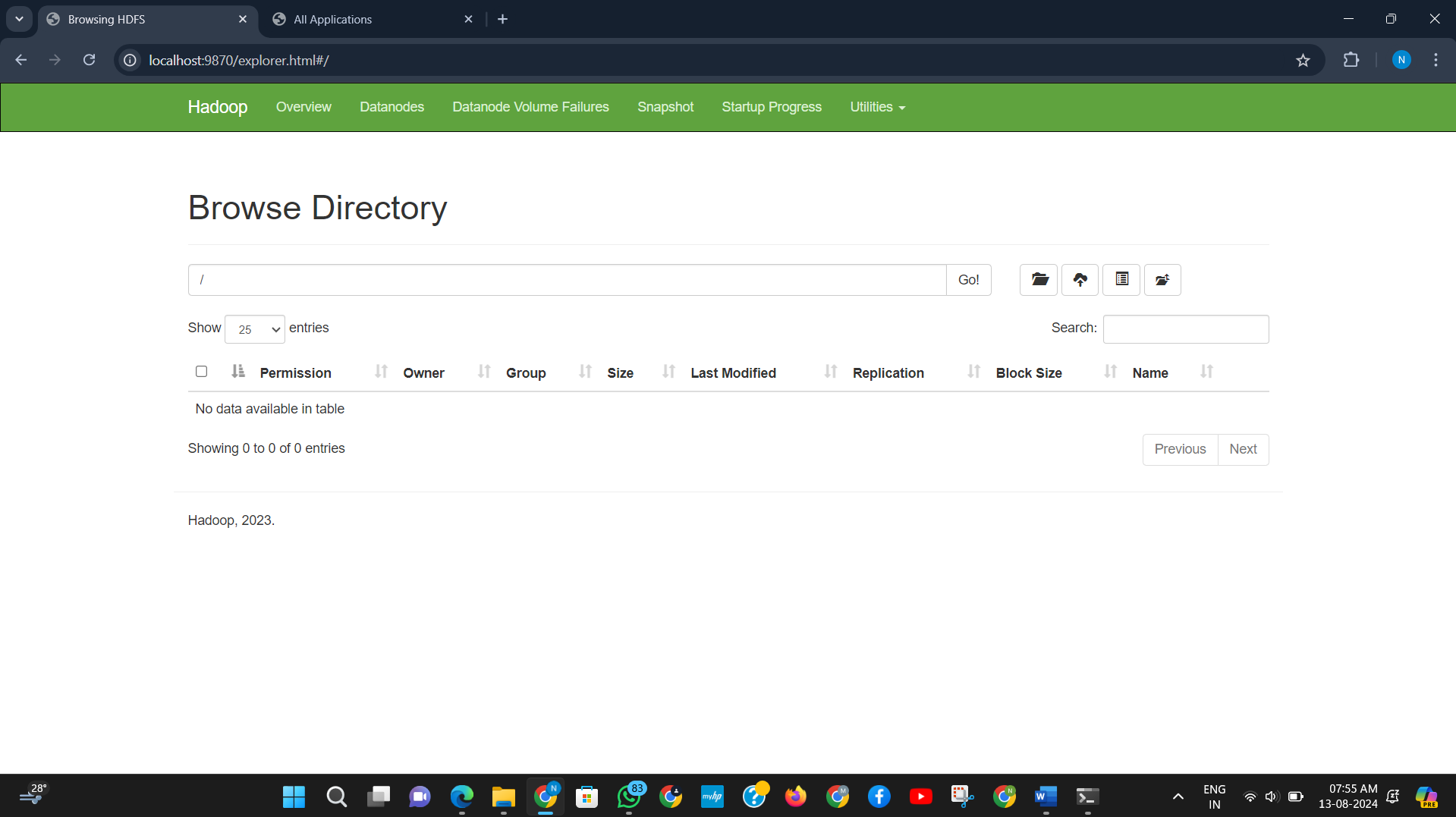
Command: .jps

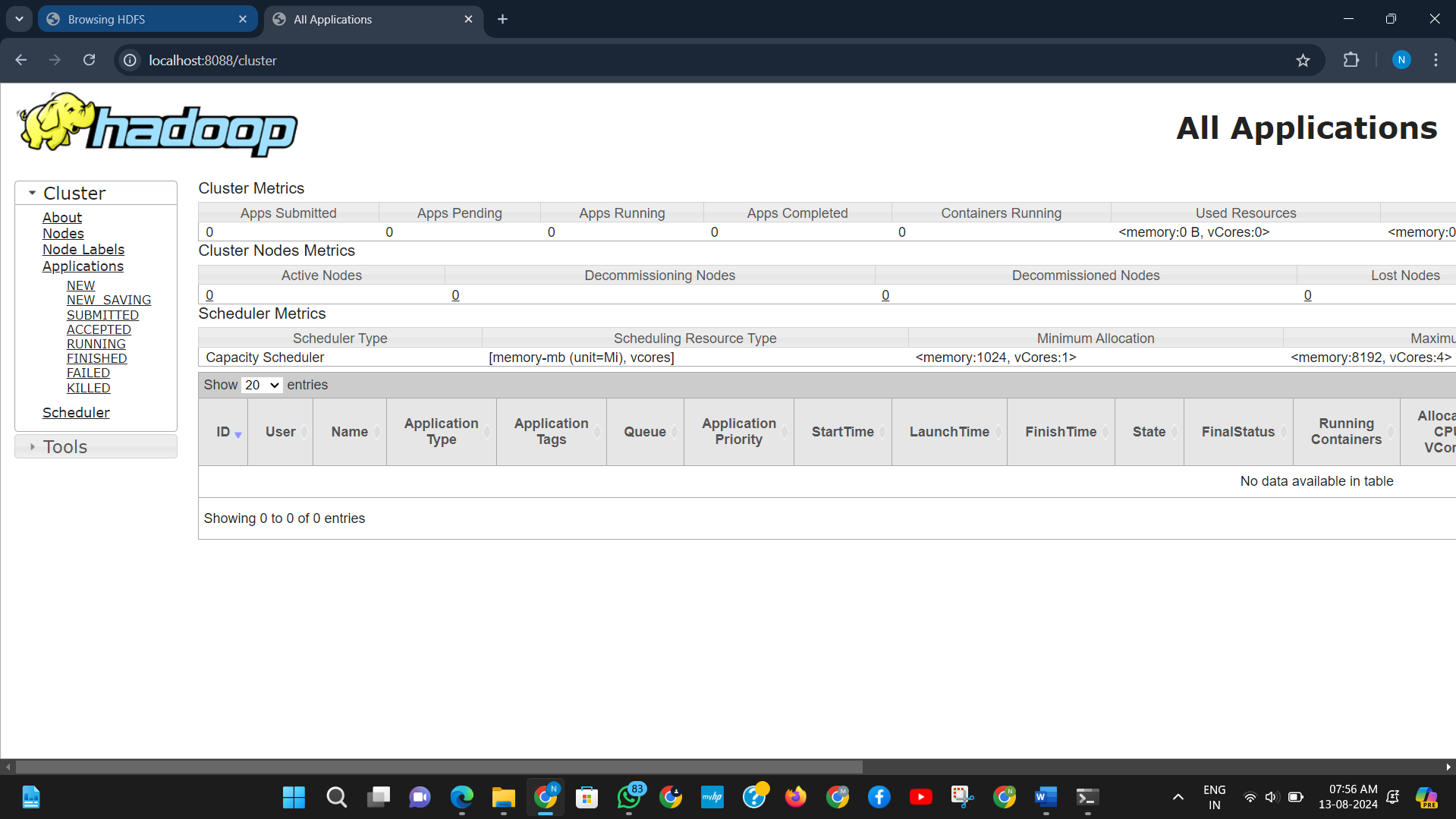
**INSTALLATION SCREENSHOTS**

****

****

****

****

****