

DATA ANALYTICS

ASSIGNMENT – 1

INTRODUCTION TO HADOOP:

Apache Hadoop software is an open source framework that allows for the distributed storage and processing of large datasets across clusters of computers using simple programming models. Hadoop is designed to scale up from a single computer to thousands of clustered computers, with each machine offering local computation and storage. In this way, Hadoop can efficiently store and process large datasets ranging in size from gigabytes to petabytes of data.

HISTORY OF HADOOP:

Hadoop has its origins in the early era of the World Wide Web. As the Web grew to millions and then billions of pages, the task of searching and returning search results became one of the most prominent challenges. Startups like Google, Yahoo, and AltaVista began building frameworks to automate search results. One project called Nutch was built by computer scientists Doug Cutting and Mike Cafarella based on Google's early work on MapReduce (more on that later) and Google File System. Nutch was eventually moved to the Apache open source software foundation and was split between Nutch and Hadoop. Yahoo, where Cutting began working in 2006, open sourced Hadoop in 2008.

VERSIONS OF HADOOP:

Hadoop has evolved over time, with several versions released to improve performance, scalability, and functionality.

1. Hadoop 0.x
2. Hadoop 1.x
3. Hadoop 2.x

4. Hadoop 3.x

5. Hadoop 4.x (Upcoming)

Each version of Hadoop brought significant improvements, especially in terms of scalability, resource management, and support for diverse workloads. The transition from Hadoop 1.x to 2.x with YARN was particularly transformative, allowing Hadoop to become the backbone of a broader big data ecosystem.

SYSTEM REQUIREMENTS FOR HADOOP:

Hardware Requirements:

1. Memory (RAM): 8 GB per node.
2. CPU: Dual-core processor.
3. Disk Storage: 1 TB of storage space.
4. Network: 1 Gbps Ethernet.

Software Requirements

1. Operating System:

- **Supported OS:**

- Linux distributions such as CentOS, Ubuntu, or Red Hat Enterprise Linux (RHEL) are most commonly used.
- Windows is supported, but Linux is preferred due to better performance and compatibility.

- **Java:** Java 8 (Hadoop 2.x and 3.x require at least Java 8).

2. Hadoop Dependencies:

- **SSH:** SSH must be installed and configured to allow password-less login for the Hadoop user across the cluster nodes.
- **Python:** Some Hadoop components (like Apache Pig) require Python, so having Python installed is recommended.

INSTALLATION STEPS ONE BY ONE WITH COMMANDS WITH ITS EXPLANATION:

Step 1: Download and install Java

```
C:\Users\OVIYA>java -version
java version "1.8.0_40"
Java(TM) SE Runtime Environment (build 1.8.0_40-b26)
Java HotSpot(TM) 64-Bit Server VM (build 25.40-b25, mixed mode)

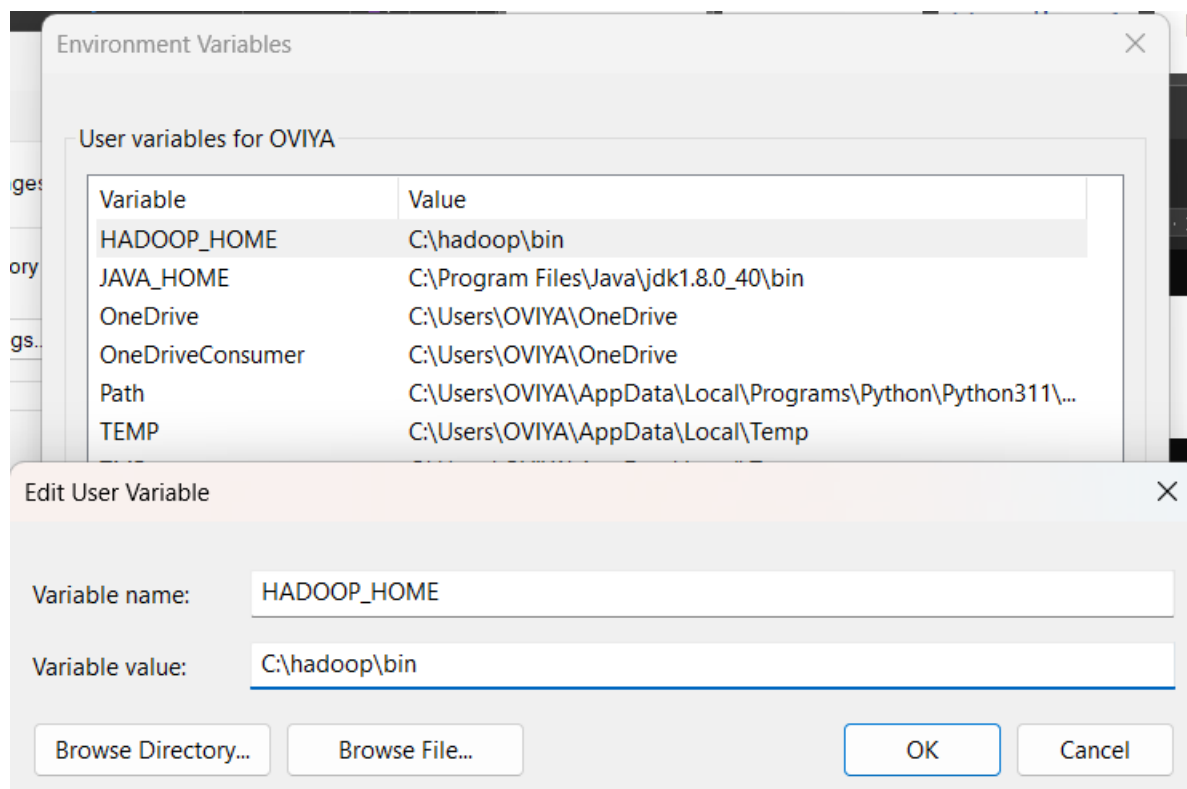
C:\Users\OVIYA>
```

Step 2: Download Hadoop

```
C:\Users\OVIYA>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 /C:/hadoop/share/hadoop/common/hadoop-common-3.3.6.jar
```

Step 3: Set Environment Variables

Click “New” under System Variables to add a new variable. Enter the variable name “HADOOP_HOME” and the path to the Hadoop folder as the variable value.



Step 4: Setup Hadoop

You must configure Hadoop in this phase by modifying several configuration files. Navigate to the “etc/hadoop” folder in the Hadoop folder. You must make changes to three files:

Open each file in a text editor and edit the properties.

- core-site.xml

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

- hdfs-site.xml

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>C:\hadoop\data\namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>C:\hadoop\data\datanode</value>
  </property>
</configuration>
```

- yarn-site.xml

```
<configuration>
  <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>
```

Step 5: Format Hadoop Name Node

```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.22621.3880]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>cd "C:/hadoop/sbin"

C:\hadoop\sbin>hdfs namenode -format_
```

Step 6: Start Hadoop

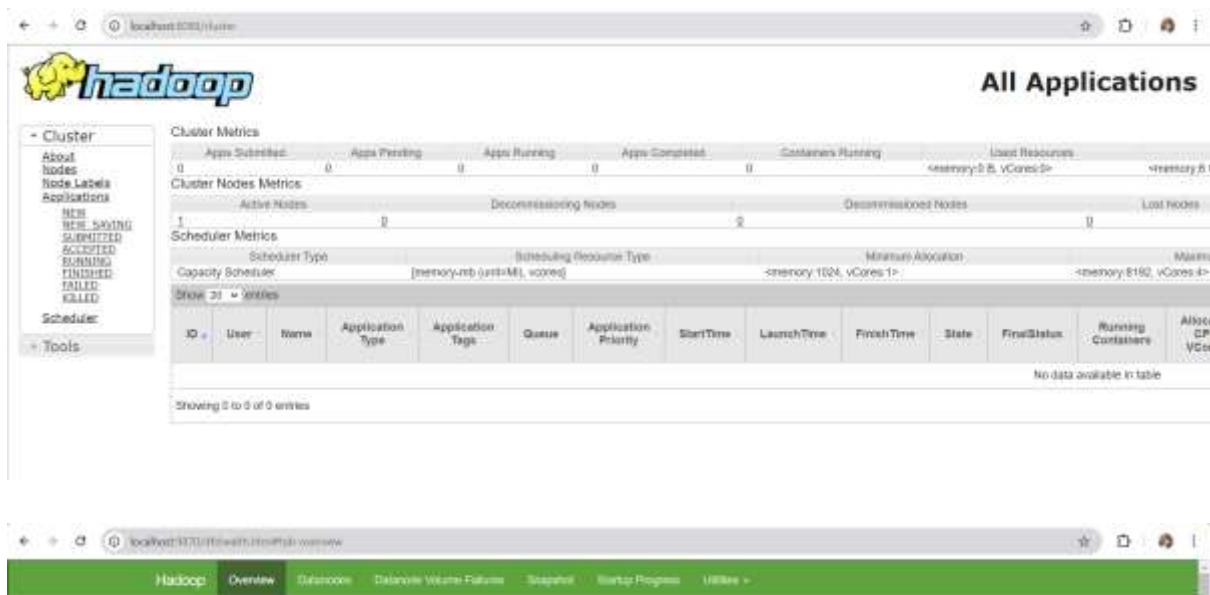
```
C:\hadoop\sbin>start-dfs.cmd

C:\hadoop\sbin>start-yarn.cmd
starting yarn daemons

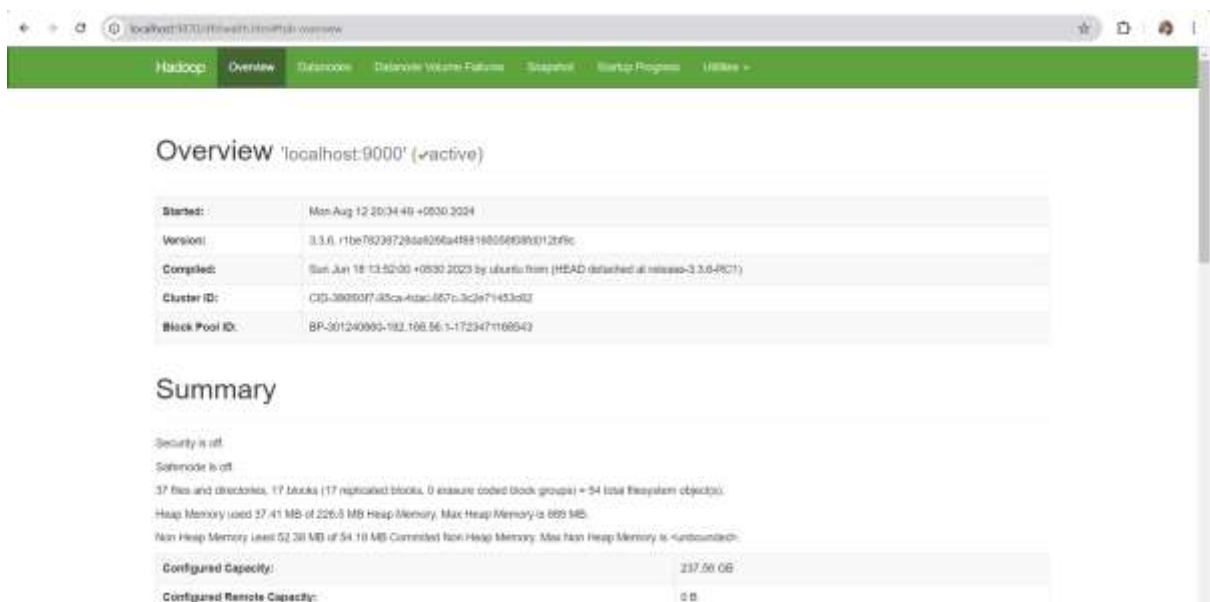
C:\hadoop\sbin>jps
19012 NameNode
22372 DataNode
23848 Jps
27736 ResourceManager
5388 NodeManager

C:\hadoop\sbin>_
```

Step 7: Verify Hadoop Installation



The screenshot shows the Hadoop web interface at localhost:8080/jsp. The 'All Applications' page is displayed, showing cluster metrics and a table of applications. The cluster metrics section includes a table with columns: Apps Submitted, Apps Pending, Apps Running, Apps Completed, Containers Running, and Used Resources. The scheduler metrics section includes a table with columns: Scheduler Type, Scheduling Resource Type, Minimum Allocation, and Maximum Allocation. The applications table is currently empty, showing 'No data available in table'.



The screenshot shows the Hadoop web interface at localhost:8080/dfshealth.html#dfs-overview. The 'Overview' page is displayed, showing cluster information and a summary of the cluster status. The cluster information section includes a table with columns: Started, Version, Compiled, Cluster ID, and Block Pool ID. The summary section includes a table with columns: Security, Softnode, Heap Memory, and Configured Capacity.