DATA ANALYTICS ASSIGNMENT

HADOOP

INTRODUCTION :

Hadoop is an open-source software framework that is used for storing and processing large amounts of data in a distributed computing environment. It is designed to handle big data and is based on the MapReduce programming model, which allows for the parallel processing of large datasets . Its framework is based on Java programming with some native code in C and shell scripts.

1.1 HISTORY OF HADOOP :

Apache Software Foundation is the developers of Hadoop, and it's co-founders are **Doug Cutting** and **Mike Cafarella**. It's co-founder Doug Cutting named it on his son's toy elephant. In October 2003 the first paper release was Google File System. In January 2006, MapReduce development started on the Apache Nutch which consisted of around 6000 lines coding for it and around 5000 lines coding for HDFS. In April 2006 Hadoop 0.1.0 was released.

1.2 VERSIONS OF HADOOP:

Important versions of Hadoop are :

1 Hadoop 3.3.6 (August 2024)
2 Hadoop 3.3.5 (April 2024)
3 Hadoop 3.3.0 (June 2021)
4 Hadoop 3.2.2 (November 2020)
5 Hadoop 2.10.1 (October 2020)
6 Hadoop 2.9.2 (April 2019)
7 Hadoop 2.8.5 (October 2018)
8 Hadoop 2.7.7 (June 2018)
9 Hadoop 1.2.2 (June 2014)

1.3 SYSTEM REQUIREMENTS :

Hardware Requirements:

Processor:

Minimum: Multi-core processor (e.g., Intel Core i3 or equivalent)

Recommended: Multi-core processor with higher clock speed (e.g., Intel Xeon or equivalent)

Memory (RAM):

Minimum: 4 GB per node

Recommended: 8 GB or more per node

Storage:

Minimum: At least 500 GB of disk space per node

Recommended: 1 TB or more per node, preferably with SSDs for better performance

Network:

Minimum: 1 Gbps Ethernet

Recommended: 10 Gbps Ethernet or higher for large clusters

Software Requirements:

Operating System:

Supported: Linux (various distributions like CentOS, Ubuntu, Debian), Unix

Unsupported: Windows is not officially supported, but Hadoop can run on Windows with some additional configuration

Java:

Minimum: Java 8 (for most Hadoop versions)

Recommended: Java 11 or later, depending on the Hadoop version

Dependencies:

Python: Some Hadoop components may require Python (typically version 2.7 or 3.x)

SSH: SSH should be configured for passwordless login between nodes in the cluster.

1.4 INSTALLATION STEPS:

Step 1: Java installation

Java Development Kit (JDK):

- Hadoop requires Java to run. Download and install the latest JDK from the Oracle website.
- Set the JAVA_HOME environment variable to the path where Java is installed.

Step 2: Extract Hadoop

Extract the downloaded Hadoop to a directory like C:\Hadoop.

Step 3: Set Environment Variables

Add Hadoop bin directory to PATH:

Add C:\Hadoop\bin to your PATH environment variable.

Set Hadoop environment variables:

Create a new system variable HADOOP_HOME and set it to C:\Hadoop.

Add HADOOP_HOME\bin to your PATH.

Step 4: Editing Hadoop files

Create a folder data in the Hadoop directory and two sub folders , namenode and datanode.

These folders are important because files on HDFS reside inside these datanode.

Step 5 : Editing configuration files

Configure core-site.xml

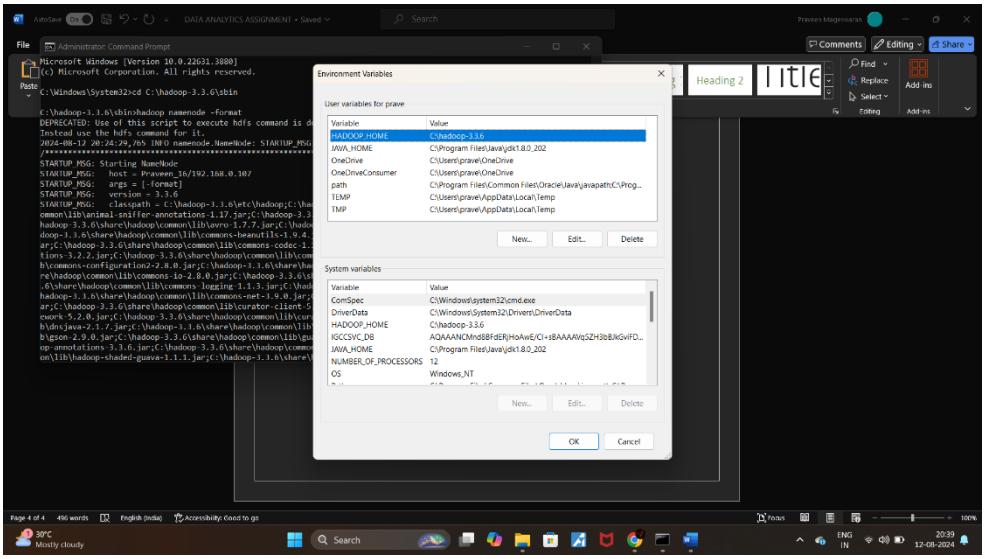
Configure hdfs-site.xml

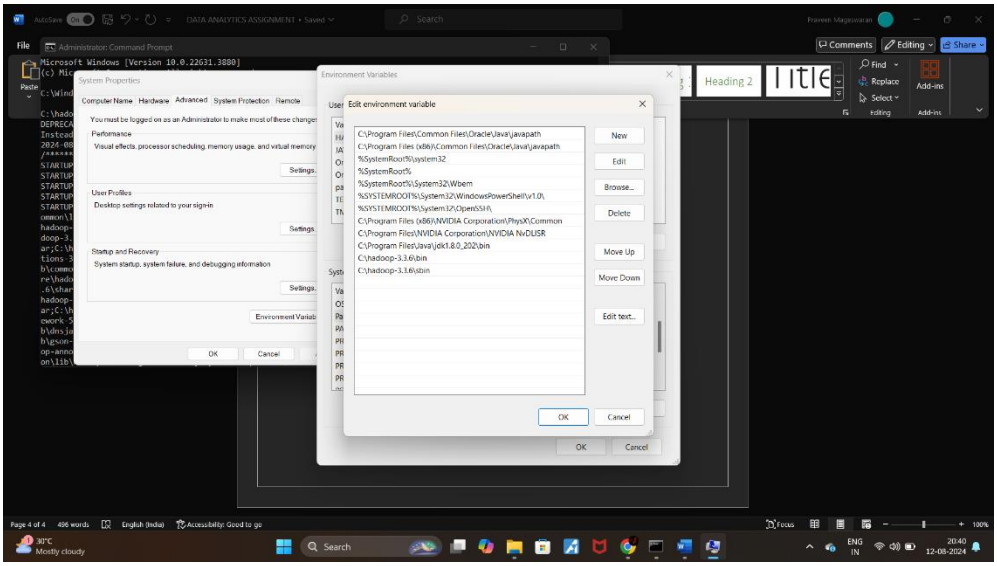
Configure mapred-site.xml

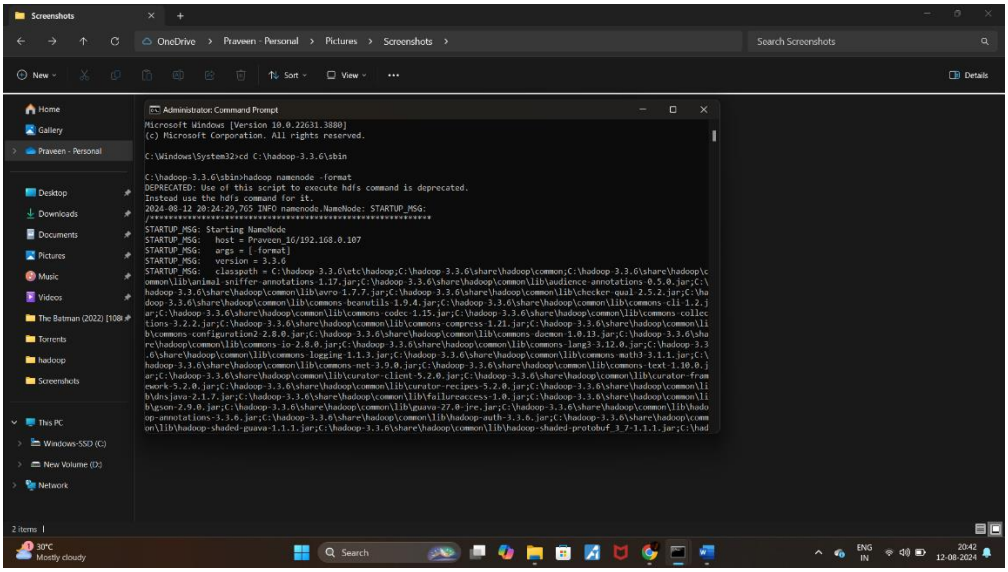
Configure yarn-site.xml

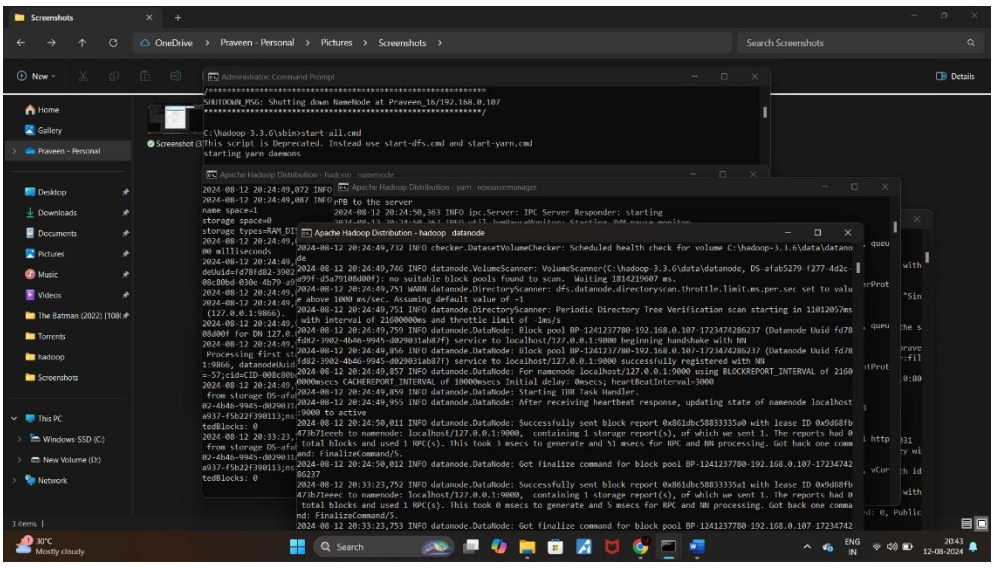
Configure Hadoop-env.cmd

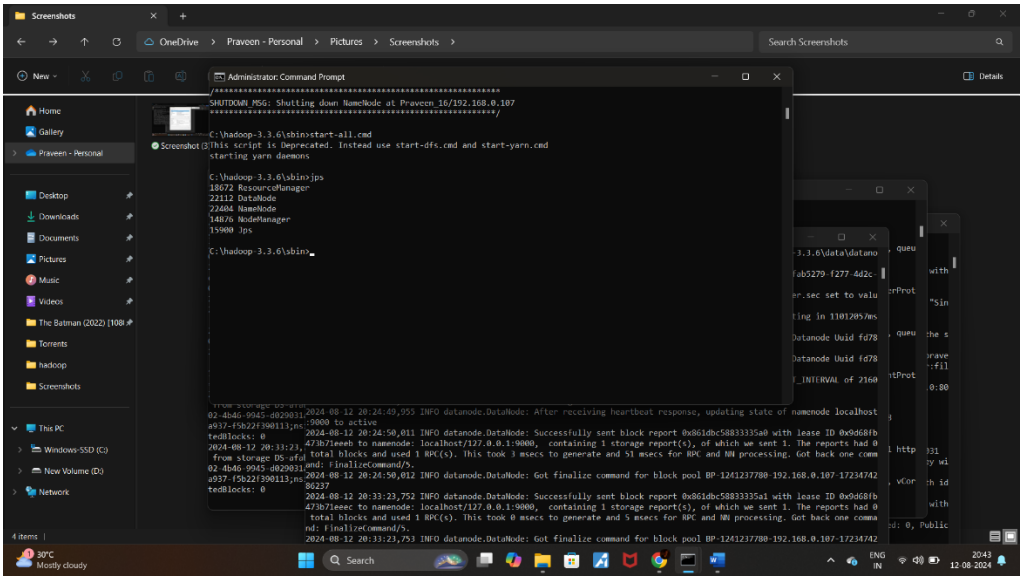
1.5 INSTALLATION SCREENSHOTS:

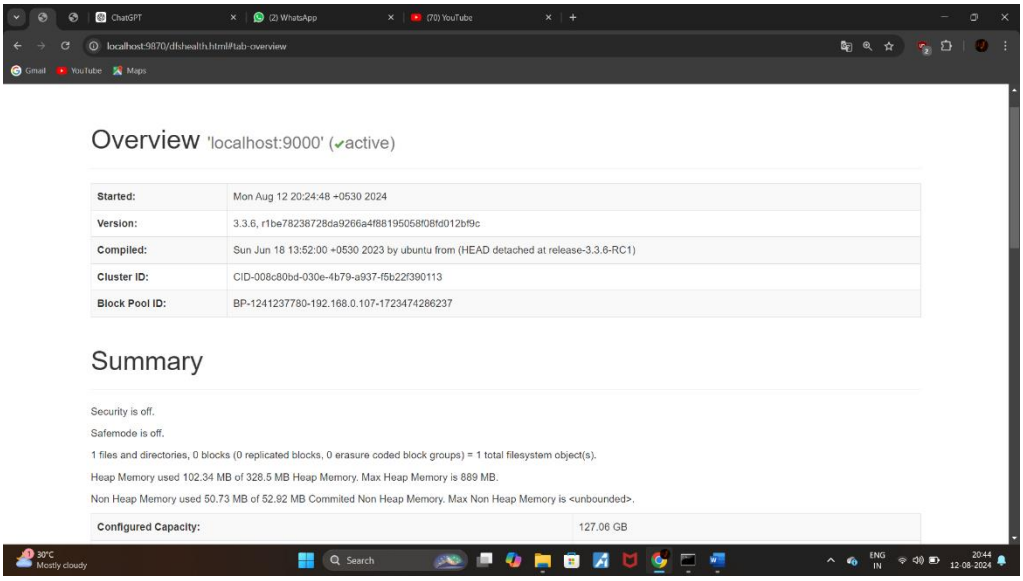


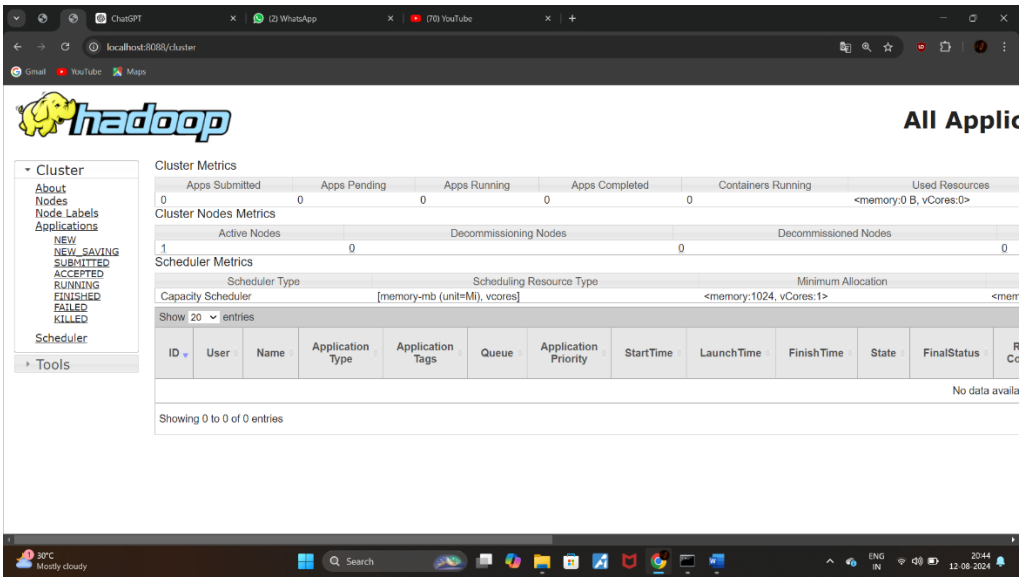

Edit environment variable
C:\Program Files\Common Files\Oracle\Java\javapath
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\Program Files (x86)\NVIDIA Corporation\PhysX\Common
C:\Program Files\NVIDIA Corporation\NVIDIA NvDLISR
C:\Program Files\Java\jdk1.8.0_202\bin
C:\hadoop-3.3.6\bin
C:\hadoop-3.3.6\sbin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel


Administrator: Command Prompt
Microsoft Windows [Version 10.0.22631.3880]
(c) Microsoft Corporation. All rights reserved.
C:\Windows\System32>cd C:\hadoop-3.3.6\sbin
C:\hadoop-3.3.6\sbin>hadoop namenode -format
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
2024-08-12 20:24:29,765 INFO namenode.NameNode: STARTUP_MSG:
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = Praveen_16/192.168.0.107
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.3.6


Administrator: Command Prompt
SHUTDOWN_MSG: Shutting down NameNode at Praveen_16/192.168.0.107
C:\hadoop-3.3.6\sbin>start all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
Apache Hadoop Distribution - hadoop namenode
Apache Hadoop Distribution - yarn resourcemanager
Apache Hadoop Distribution - hadoop datanode


Administrator: Command Prompt
C:\hadoop-3.3.6\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
C:\hadoop-3.3.6\sbin>jps
18672 ResourceManager
22112 DataNode
22404 NameNode
14876 NodeManager
15900 Jps
C:\hadoop-3.3.6\sbin>


localhost:9870/dfshealth.html#tab-overview
Overview 'localhost:9000' (active)
Started: Mon Aug 12 20:24:48 +0530 2024
Version: 3.3.6, r1be78238728da9266a4f88195058f08fd012bf9c
Compiled: Sun Jun 18 13:52:00 +0530 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID: CID-008c80bd-030e-4b79-a937-f5b22f390113
Block Pool ID: BP-1241237780-192.168.0.107-1723474286237
Summary
Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 102.34 MB of 328.5 MB Heap Memory. Max Heap Memory is 889 MB.
Non Heap Memory used 50.73 MB of 52.92 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.
Configured Capacity: 127.06 GB


localhost:8088/cluster
hadoop
All Applic
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Cluster Metrics
Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Used Resources
0 | 0 | 0 | 0 | 0 | <memory:0 B, vCores:0>
Cluster Nodes Metrics
Active Nodes | Decommissioning Nodes | Decommissioned Nodes
1 | 0 | 0 | 0
Scheduler Metrics
Scheduler Type | Scheduling Resource Type | Minimum Allocation
Capacity Scheduler | [memory-mb (unit=Mi), vcores] | <memory:1024, vCores:1>
Show 20 entries
ID | User | Name | Application Type | Application Tags | Queue | Application Priority | StartTime | LaunchTime | FinishTime | State | FinalStatus
No data availa
Showing 0 to 0 of 0 entries