

# **DATA ANALYTICS**

## **ASSIGNMENT – 1**

### **INTRODUCTION ABOUT HADOOP:**

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.

Hadoop is an open-source software framework for storing and processing big data. It was created by Apache Software Foundation in 2006, based on a white paper written by Google in 2003 that described the Google File System (GFS) and the MapReduce programming model. The Hadoop framework allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. It is used by many organizations, including Yahoo, Facebook, and IBM, for a variety of purposes such as data warehousing, log processing, and research.

### **1.2 VERSIONS OF HADOOP:**

- 1.Hadoop 0.20.x(2009)
- 2.Hadoop 1.x(2011)
- 3.Hadoop 2.x(2013)

- 4.Hadoop 3.x(2017)
- 5.Hadoop 3.1 and 3.2
6. Hadoop 3.3 (2020)
7. Hadoop 3.4 and Beyond (Future Directions)

### **1.3 SYSTEM REQUIREMENTS FOR HADOOP:**

#### **General Requirement:**

##### **JAVA:**

1. Hadoop requires a Java Runtime Environment (JRE) or Java Development Kit (JDK) version 8 or higher. The recommended version is JDK 8.
2. Ensure that the JAVA\_HOME environment variable is set to the path of your Java installation.

##### **MEMORY:**

1. Minimum: 8 GB of RAM (for basic testing and development).
2. Recommended:16-64 GB of RAM.

##### **STORAGE:**

1. Minimum: 10 GB of free disk space.
2. Recommended: Several terabytes of disk space.

#### **Operating System Requirements:**

##### **Windows:**

##### **Java:**

Hadoop requires a Java Runtime Environment (JRE) or Java Development Kit (JDK) version 8 or higher. The recommended version is JDK 8.

##### **Dependencies:**

1. Cygwin or windows Subsystem for Linux (WSL) can be used to emulate a Linux-like environment.
2. Ensure SSH is installed and configured (via Cygwin or WSL).

## Linux:

### 1. Supported Distributions:

CentOS, Red Hat Enterprise Linux (RHEL), Ubuntu, Debian, and SUSE.

### 2. Dependencies:

- SSH must be installed and configured for password-less login for the Hadoop user.
- Native libraries (e.g., zlib, openssl) should be installed for performance improvements.

## macOS:

### 1. Java:

Hadoop requires a Java Runtime Environment (JRE) or Java Development Kit (JDK) version 8 or higher. The recommended version is JDK 8.

### 2. Dependencies:

- Ensure that SSH is enabled and configured.
- Install Hadoop via Homebrew or manually.

### 3. Additional Tools:

Xcode command-line tools.








## 1.4 INSTALLATION STEPS ONE BY ONE WITH COMMANDS WITH ITS EXPLANATION:

### Setting Up Environment Variables:

To edit environment variables, go to Control Panel > System > click on the “Advanced system settings” link

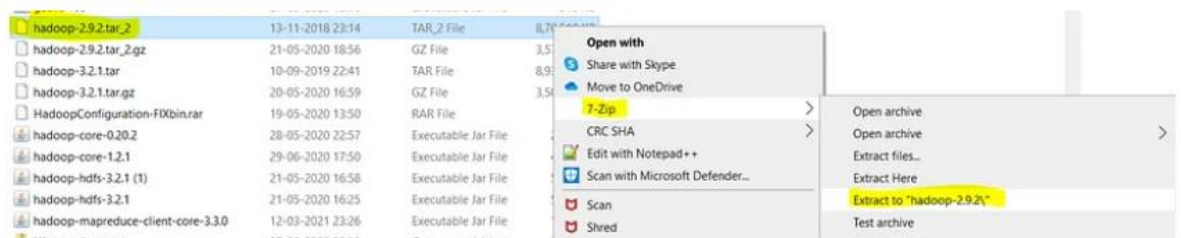
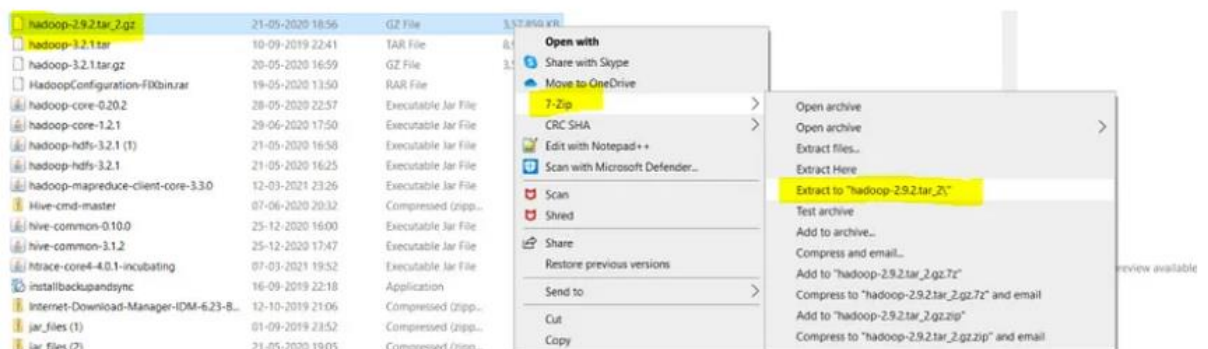
---

## Index of /dist/hadoop/core/hadoop-2.9.2

Name	Last modified	Size	Description
 <a href="#">Parent Directory</a>		-	
 <a href="#">hadoop-2.9.2-src.tar.gz</a>	2020-07-03 04:37	37M	
 <a href="#">hadoop-2.9.2-src.tar.gz.asc</a>	2020-07-03 04:36	801	
 <a href="#">hadoop-2.9.2-src.tar.gz.mds</a>	2020-07-03 04:36	1.0K	
 <a href="#">hadoop-2.9.2.tar.gz</a>	2020-07-03 04:38	349M	
 <a href="#">hadoop-2.9.2.tar.gz.asc</a>	2020-07-03 04:37	801	
 <a href="#">hadoop-2.9.2.tar.gz.mds</a>	2020-07-03 04:36	1.0K	

## Setting JAVA\_HOME

- Open environment Variable and click on “New” in “User Variable”
- Now , add JAVA\_HOME in variable name and path of Java(jdk) in Variable Value.
- Click OK and we are half done with setting JAVA\_HOME.



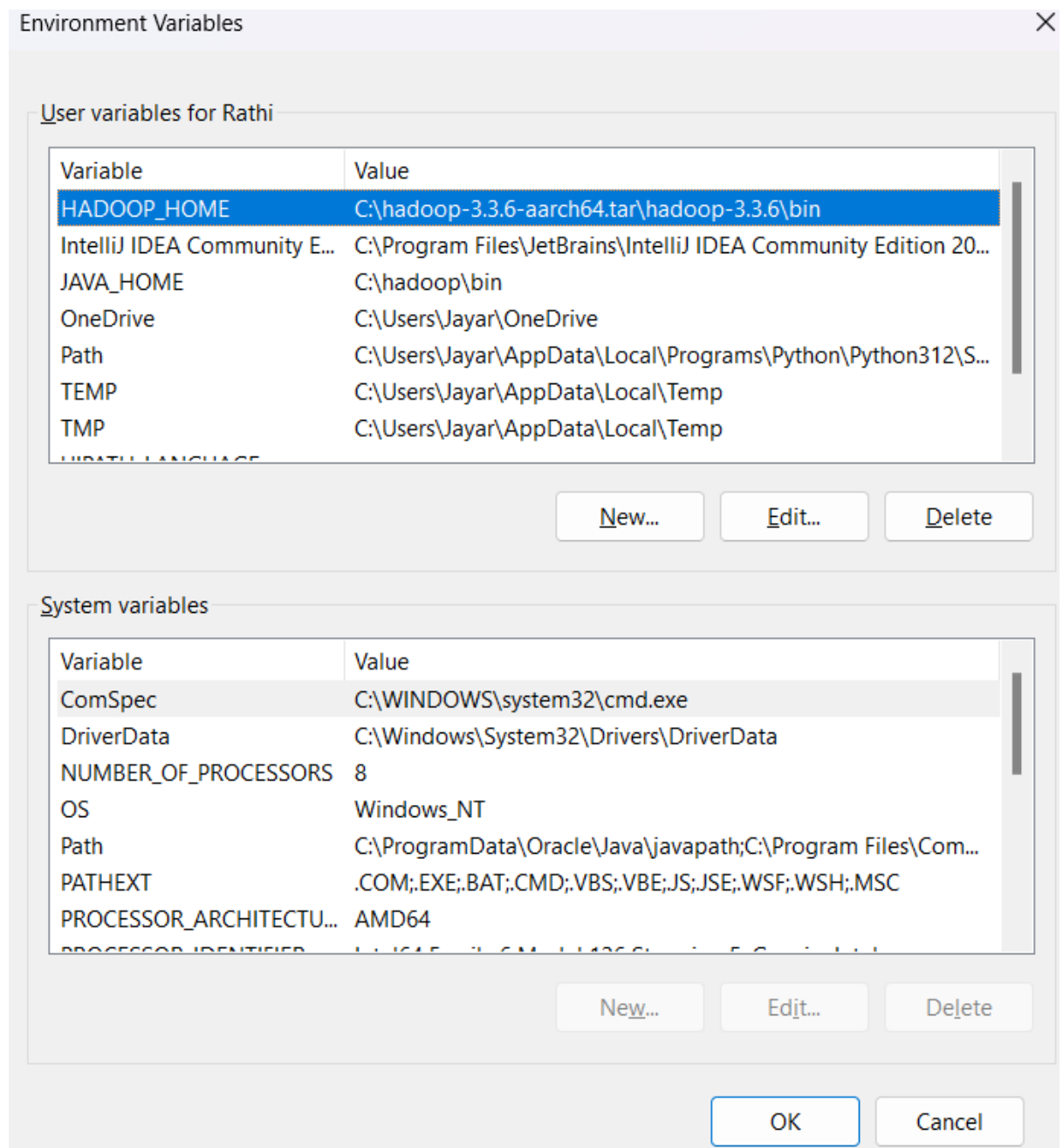
This PC > Shashank (D:) > Shashank > Study > hadoop-2.9.2

Name	Date modified	Type	Size
bin	20-09-2020 16:35	File folder	
data	20-09-2020 16:31	File folder	
etc	13-11-2018 20:45	File folder	
include	13-11-2018 20:45	File folder	
lib	13-11-2018 20:45	File folder	
libexec	13-11-2018 20:45	File folder	
logs	20-09-2020 16:41	File folder	
sbin	13-11-2018 20:45	File folder	
share	13-11-2018 20:45	File folder	
LICENSE	13-11-2018 20:45	TXT File	104 KB
NOTICE	13-11-2018 20:45	TXT File	16 KB
README	13-11-2018 20:45	TXT File	2 KB

## Setting HADOOP\_HOME

- Open environment Variable and click on “New” in “User Variable”

- Now, add HADOOP\_HOME in variable name and path of Hadoop folder in Variable Value.
- Click OK and we are half done with setting HADOOP\_HOME.



## Setting Path Variable

- ## Verify the Paths

- ```
echo %JAVA_HOME%

echo %HADOOP_HOME%

echo %PATH%
```

## Editing Hadoop files

## Creating Folders

- Create **DATA folder** in the Hadoop directory
- These folders are important because files on HDFS resides inside the datanode.

[illegible]

Fig. 15:- Formatting Namenode

```

C:\Users\shash>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
starting yarn daemons
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.

C:\Users\shash>

```

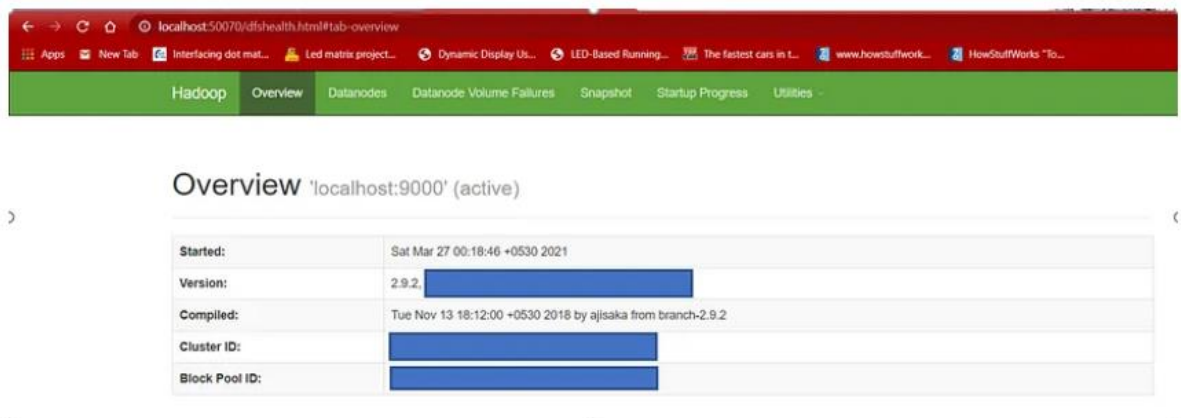


Fig. 18:- Namenode Web UI

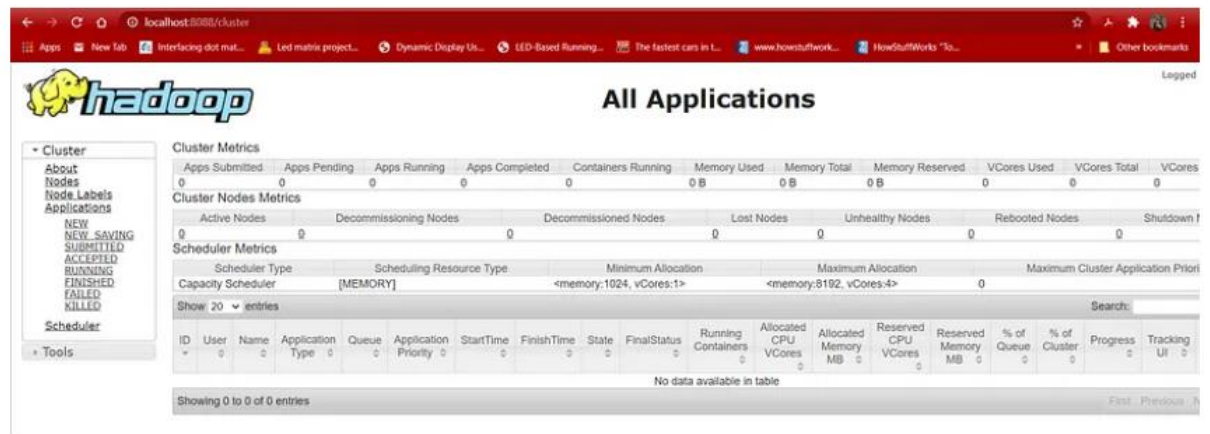


Fig. 19:- Resourcemanager Web UI