HADOOP INSTALLATION IN WINDOWS

BRIJESH JOTANIYA - 20BSIT019

ISHA MALLI - 20BSIT079

Introduction To Hadoop -:

- Hadoop is an open source framework that allows to store & process large data sets in a parallel & distributed manner.
- Two main componenets HDFS & MapReduce.
- Hadoop Distributed File Systm(HDFS) is the primary data storage system used by Hadoop applications.
- MapReduce is the processing unit of Hadoop.

HDFS(Hadoop Distributed File System)

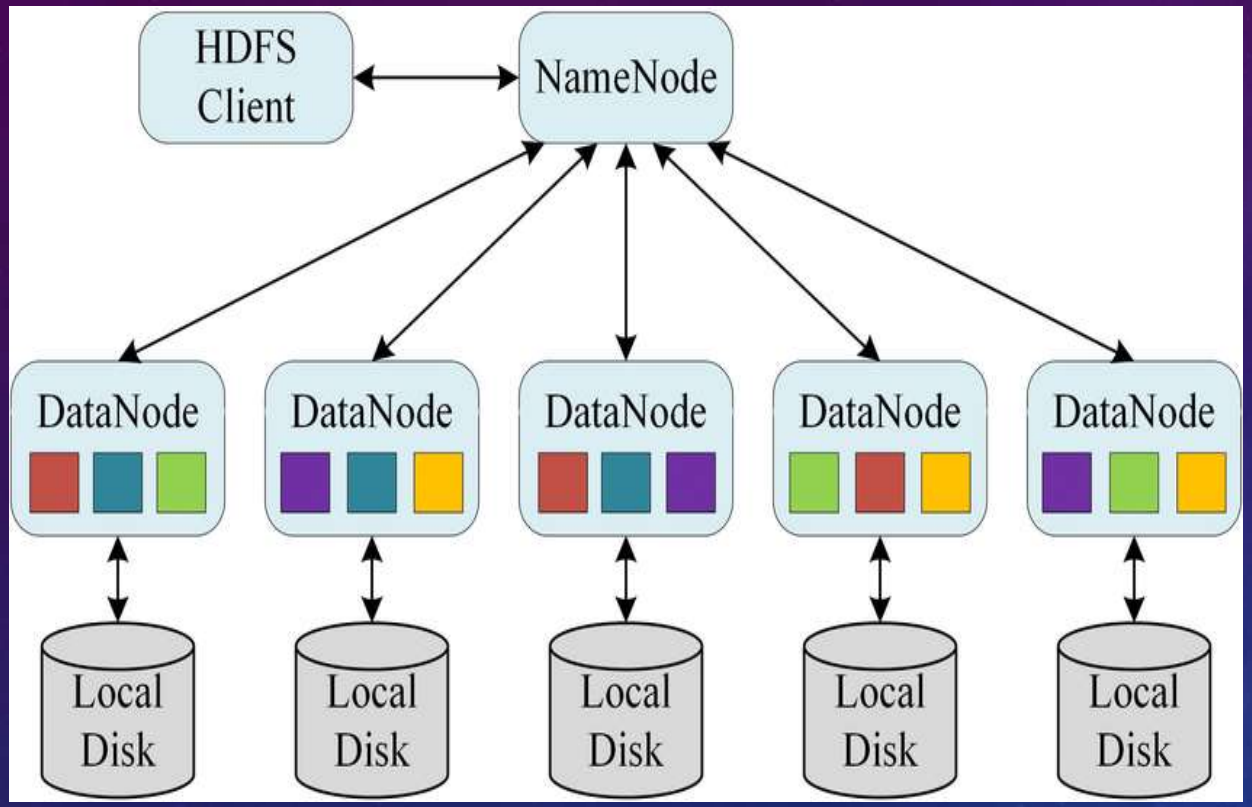
- Hadoop provides one of the most reliable filesystems.
- HDFS (Hadoop Distributed File System) is a unique design that provides storage for *extremely large files* with streaming data access pattern and it runs on *commodity hardware*.
- Component of HDFS -:

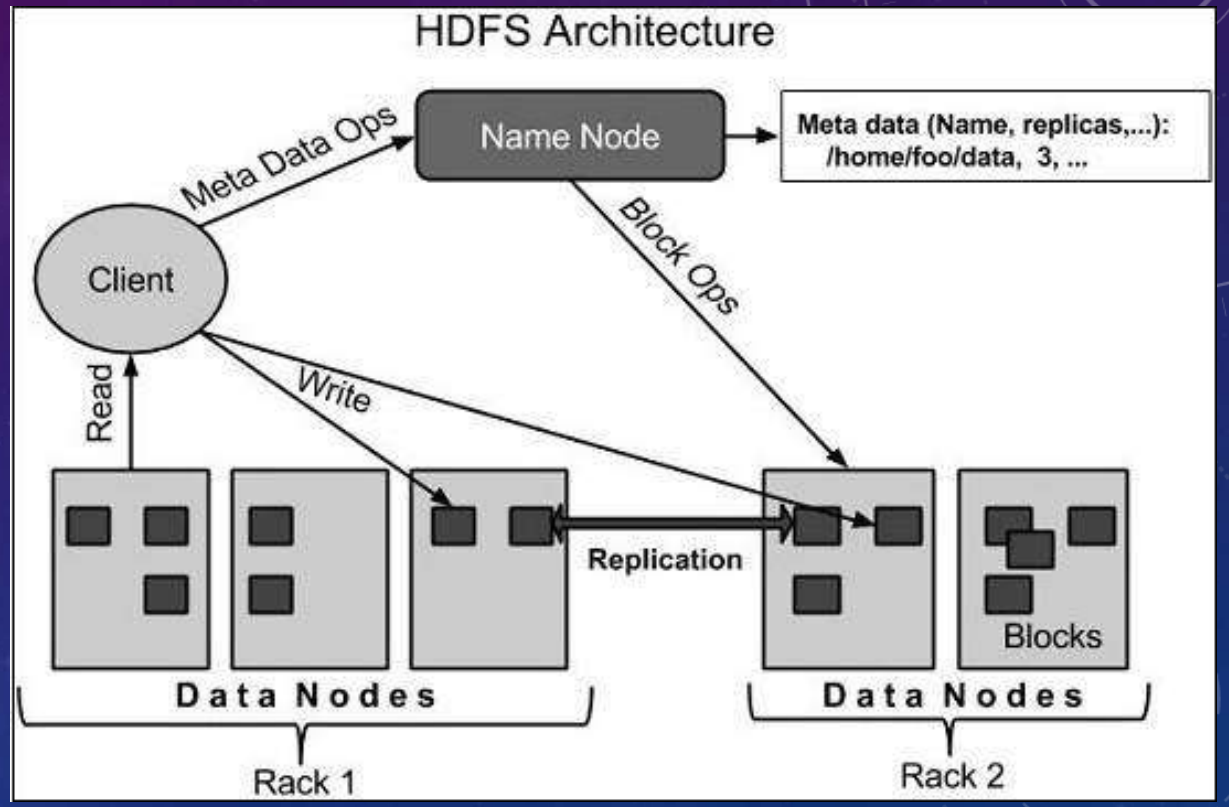
  -> HDFS Client

  -> Name Node

  -> Data Node

HDFS Architecture -:



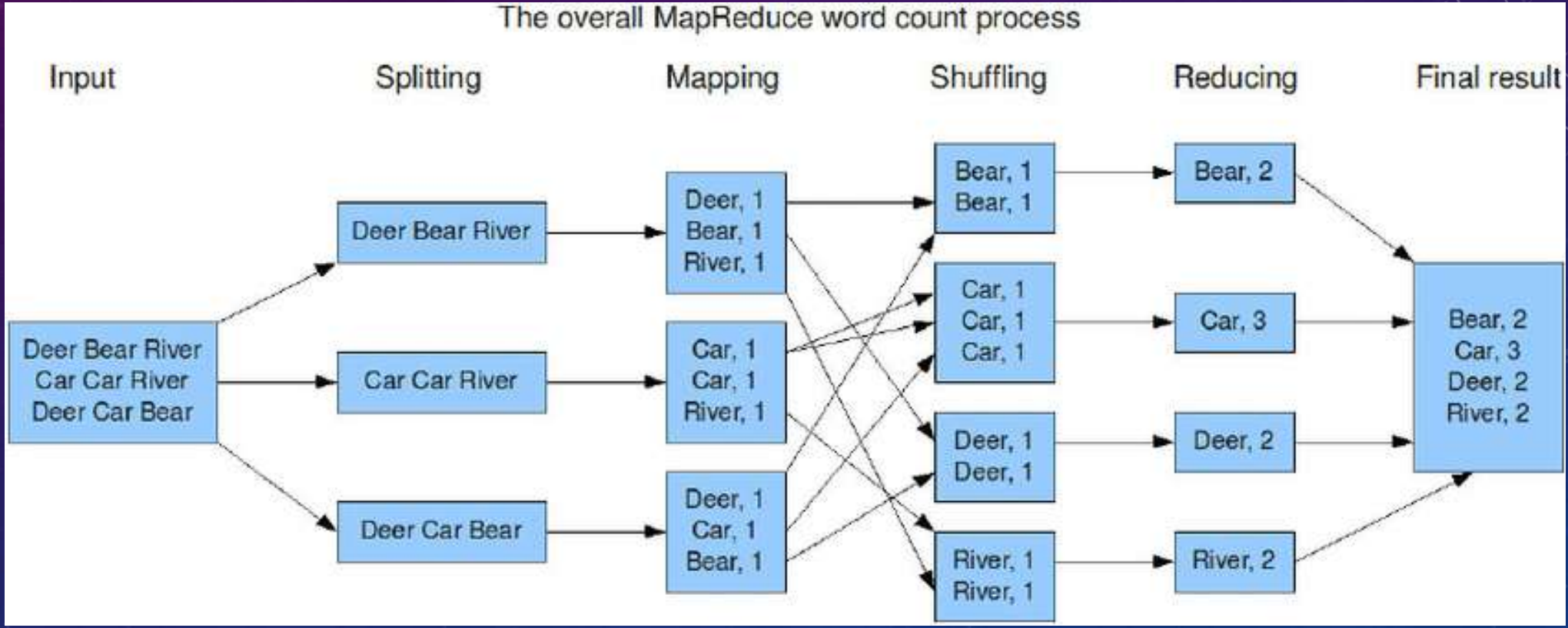


- HDFS client -:
  - With the help of Name node HDFS client can read and write the file.
- Name node -:
  - Name node is also known as Master node.
  - Main task is how to store the data. It storess the data in distributed manner.
  - It has file system namespace which have info related data stored in data node.
  - It works on replication factor.
  - It regularly receives a HeartBeat and a block report from all the DataNode in the cluster that the DataNodes are live.
- Data node -:
  - This is te actual data storage.
  - It will devide file into 128 MB partition.
  - The DataNode perform the low level read and write requests from the file system's client.

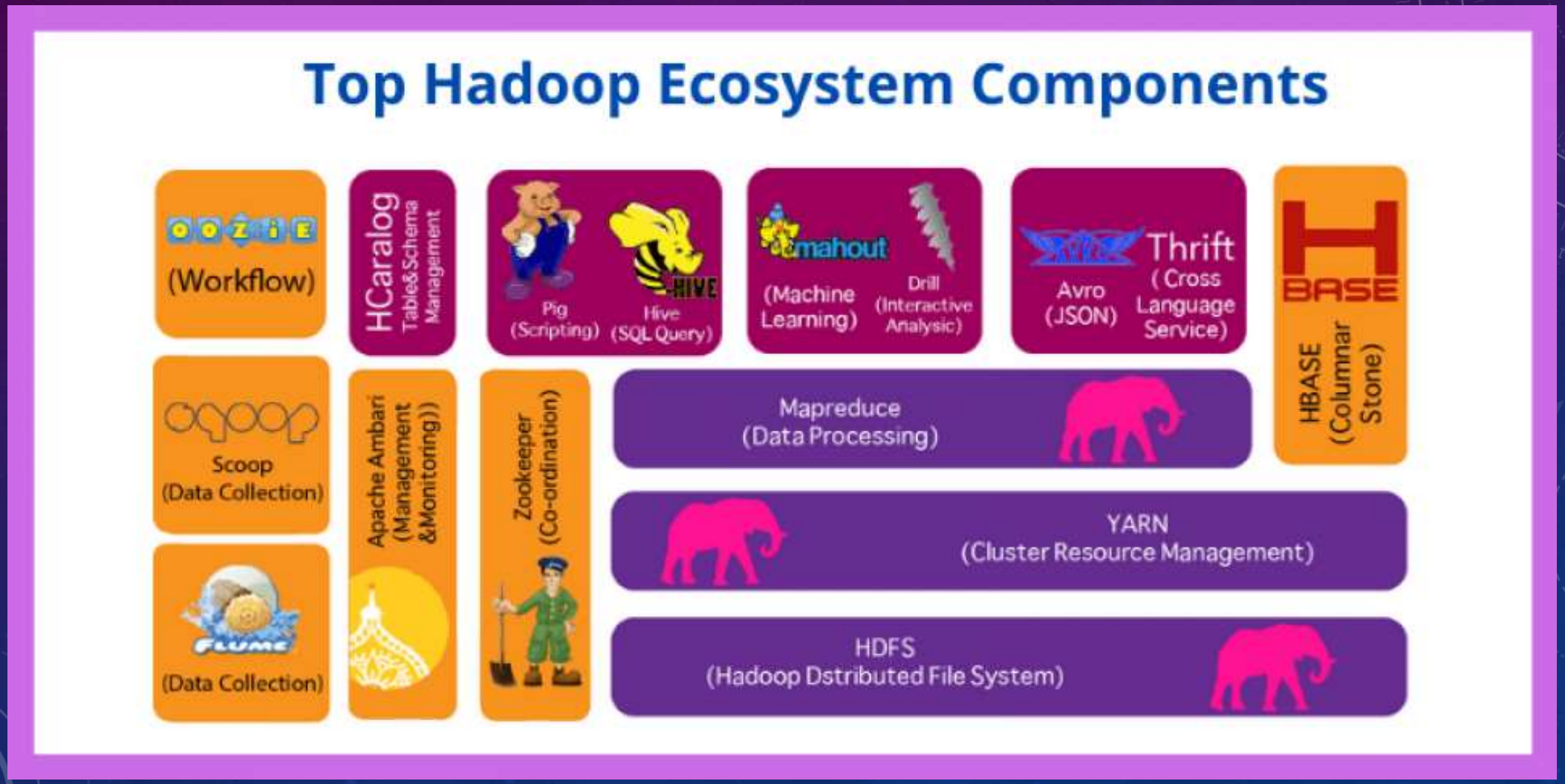
MapReduce -:

- It is core component of Apache Hadoop.
- MapReduce performs the processing of large data sets in distributed and parallel manner.
- MapReduce consists of two distinct tasks – Map And Reduce.
- Two essential daemons of MapReduce :
    - Job Tracker -: Schedules jobs and tracks the assign jobs to Task tracker.
    - Task Traker -: Tracks the task and reports status to JobTracker.
- Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

Example -:


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

Hadoop Ecosystem -:




Big Data
hadoop
Hadoop 3.3.0
Installation on
Windows 10
Programming Epitome – Umarah Qaseem

Hadoop Installation Steps-:

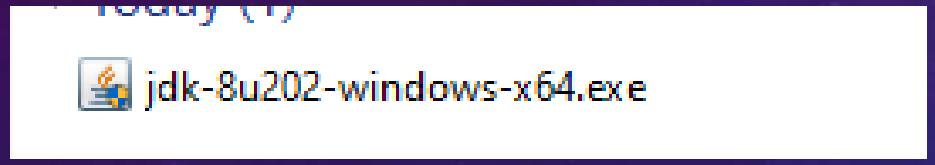
Steps to Install Hadoop

- Install Java JDK 1.8
- Download Hadoop and extract and place under C drive
- Set Path in Environment Variables
- Config files under Hadoop directory
- Create folder datanode and namenode under data directory
- Edit HDFS and YARN files
- Set Java Home environment in Hadoop environment
- Setup Complete. Test by executing start-all.cmd

Download Java jdk 8

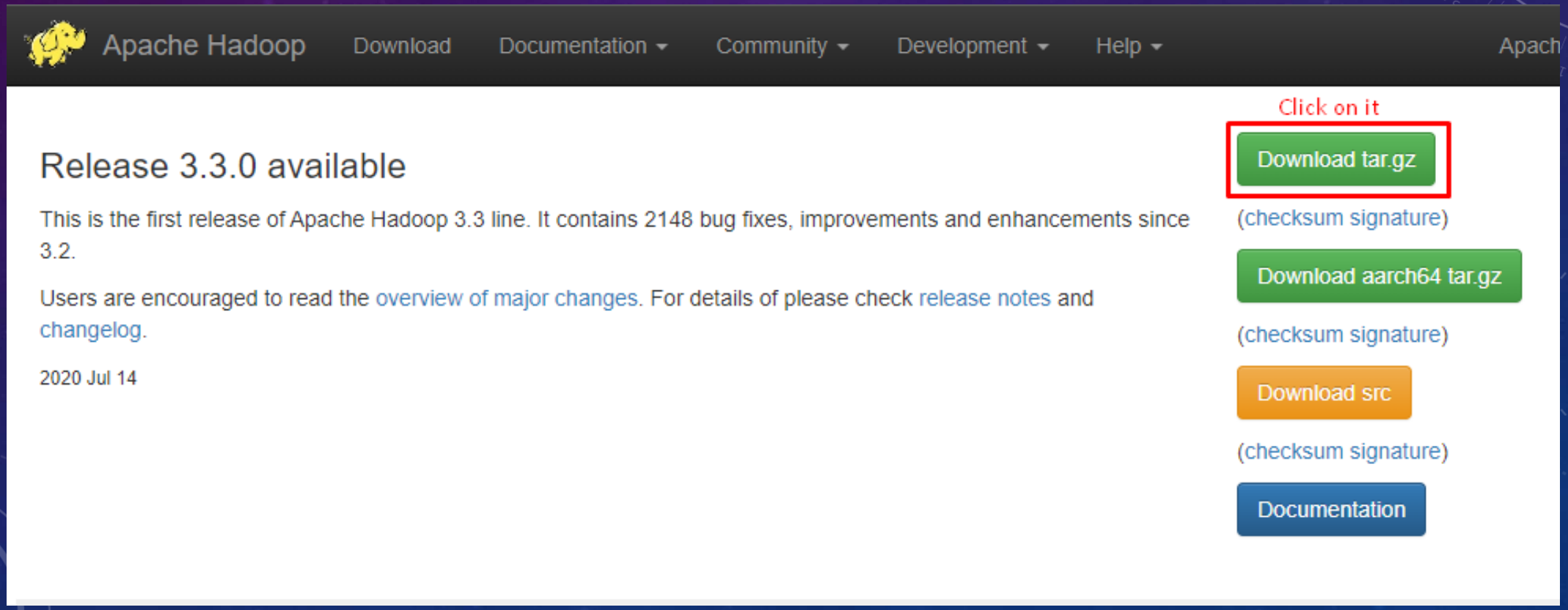
https://www.oracle.com/in/java/technologies/javase/javase8-archive-downloads.html

Download and install jdk based on your computer specification.

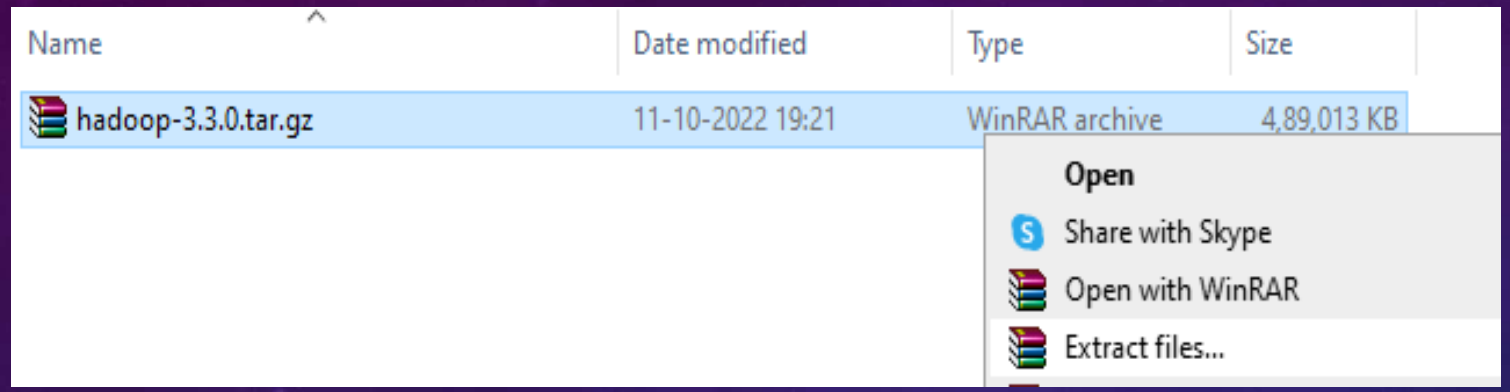


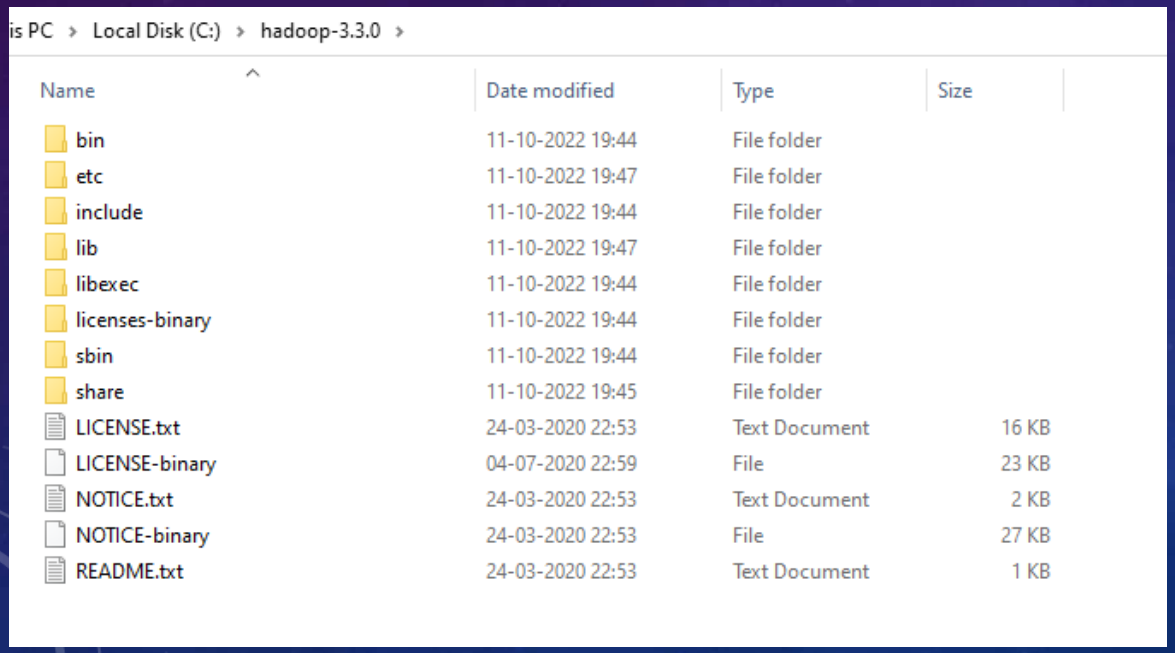
Download Hadoop Binaries

- https://hadoop.apache.org/release/3.3.0.html



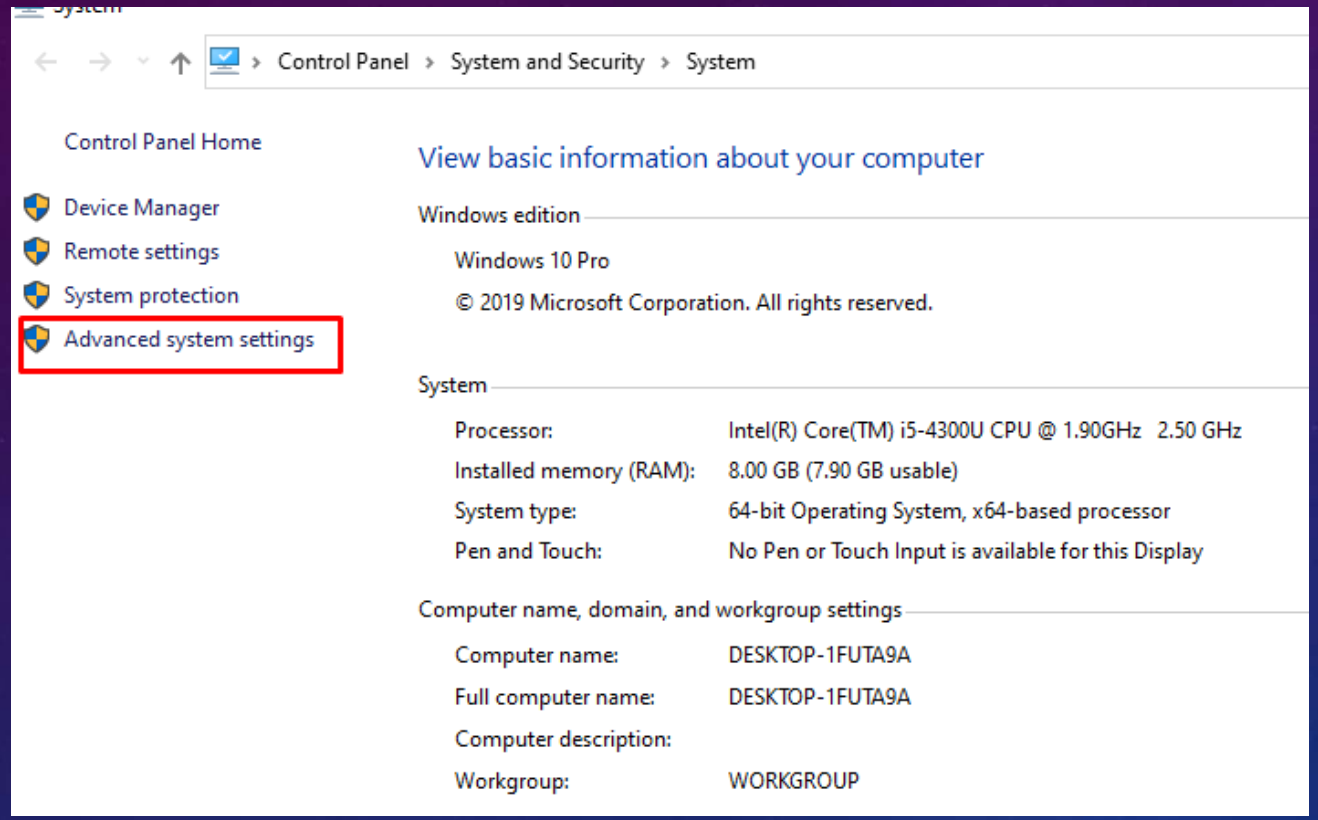
Extract this file in C drive

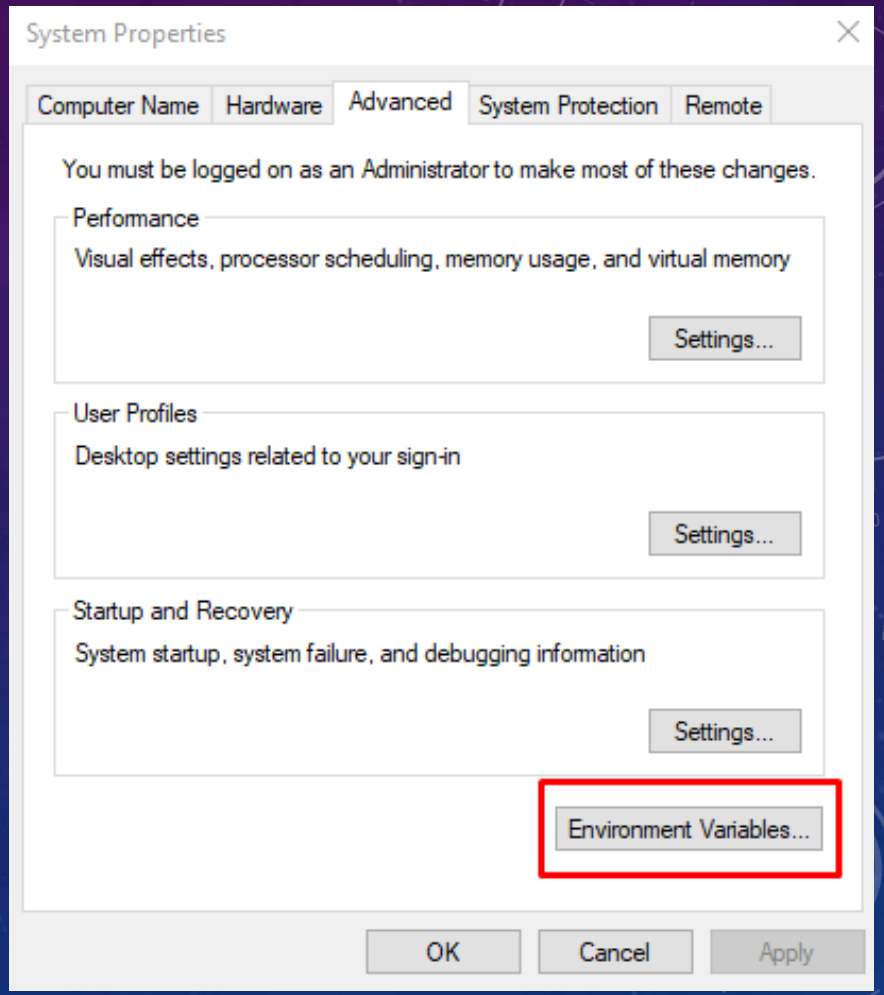


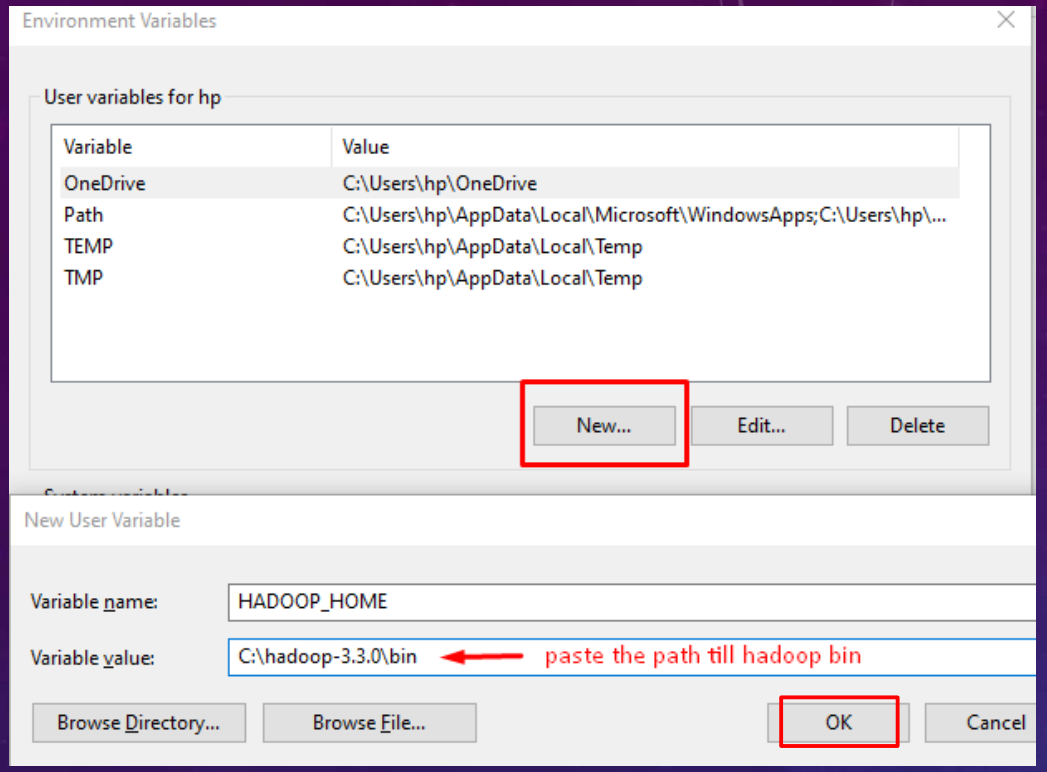


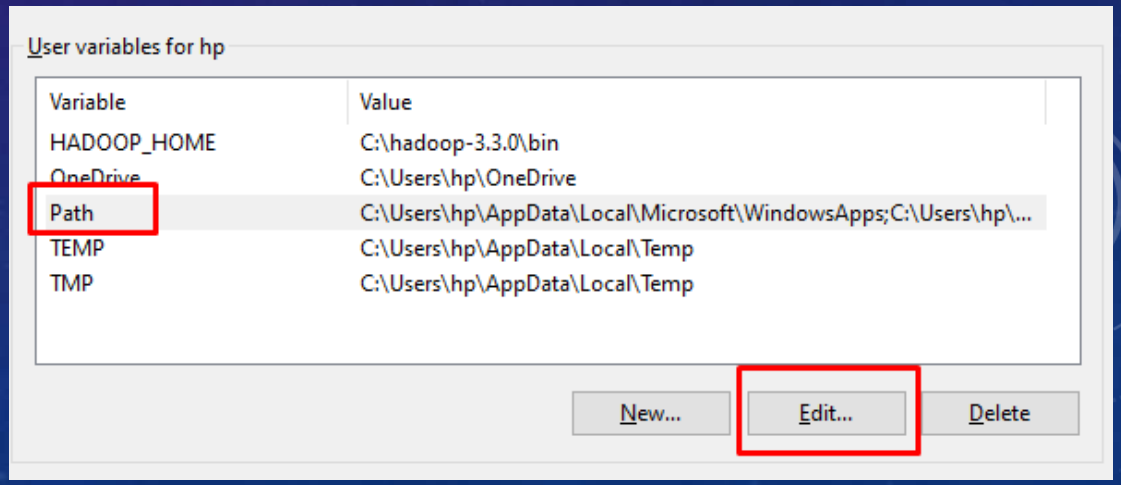
This is unzipped folder

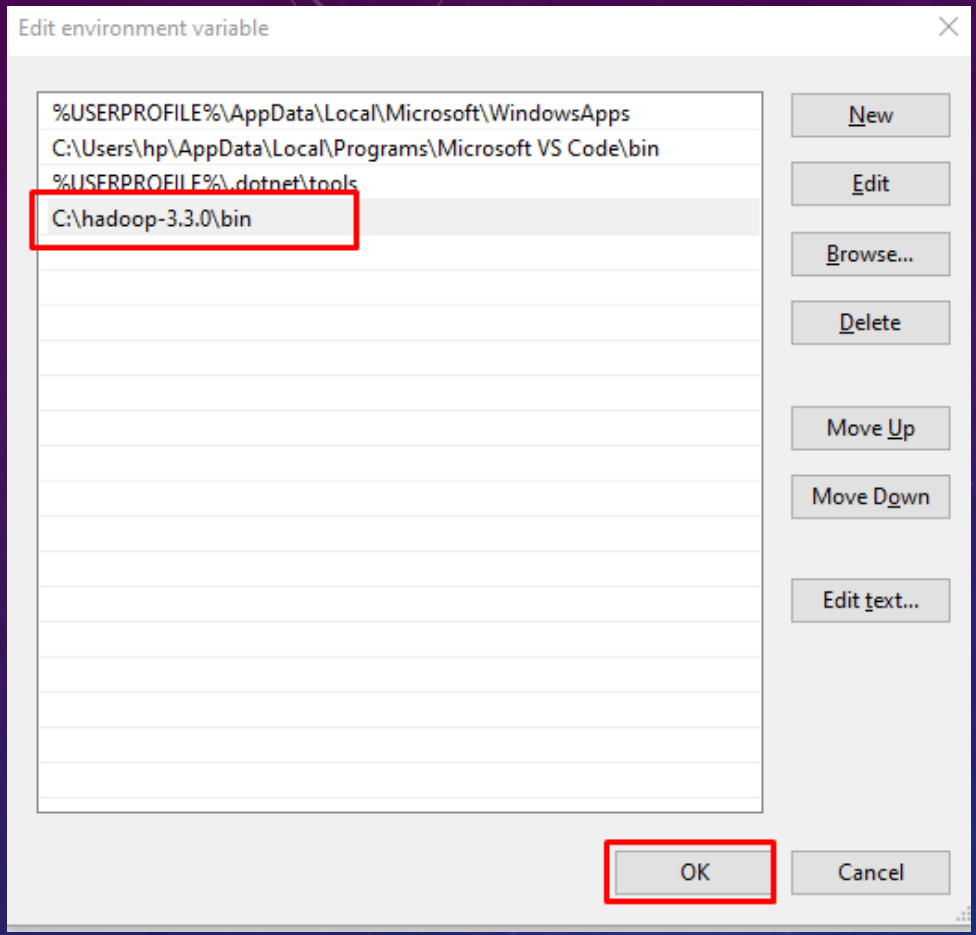
Set path in environment variable -:






Environment Variables
User variables for hp
Variable
Value
OneDrive
C:\Users\hp\OneDrive
Path
C:\Users\hp\AppData\Local\Microsoft\WindowsApps;C:\Users\hp\...
TEMP
C:\Users\hp\AppData\Local\Temp
TMP
C:\Users\hp\AppData\Local\Temp
New...
Edit...
Delete
New User Variable
Variable name:
HADOOP_HOME
Variable value:
C:\hadoop-3.3.0\bin
paste the path till hadoop bin
Browse Directory...
Browse File...
OK
Cancel

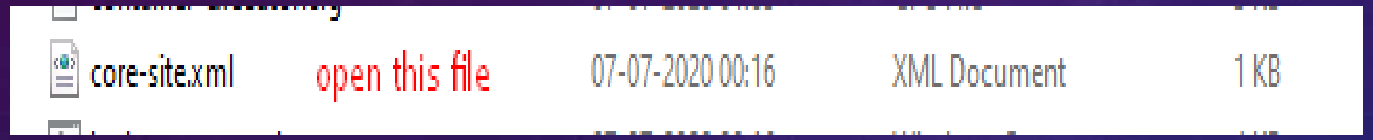

User variables for hp
Variable
Value
HADOOP_HOME
C:\hadoop-3.3.0\bin
OneDrive
C:\Users\hp\OneDrive
Path
C:\Users\hp\AppData\Local\Microsoft\WindowsApps;C:\Users\hp\...
TEMP
C:\Users\hp\AppData\Local\Temp
TMP
C:\Users\hp\AppData\Local\Temp
New...
Edit...
Delete



Similarly set the path for the Java.
Set the path till bin..

Edit Few Files(xml Files)-:

Go in this folder



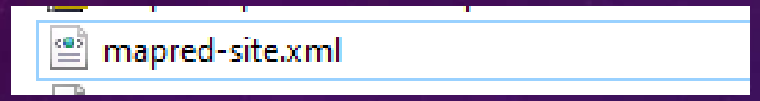
1. core-site file

Edit file C:/Hadoop-3.3.0/etc/hadoop/core-site.xml,
Open the file and paste the xml code in file and save.

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

paste this code between configuration

2. mapred-site file



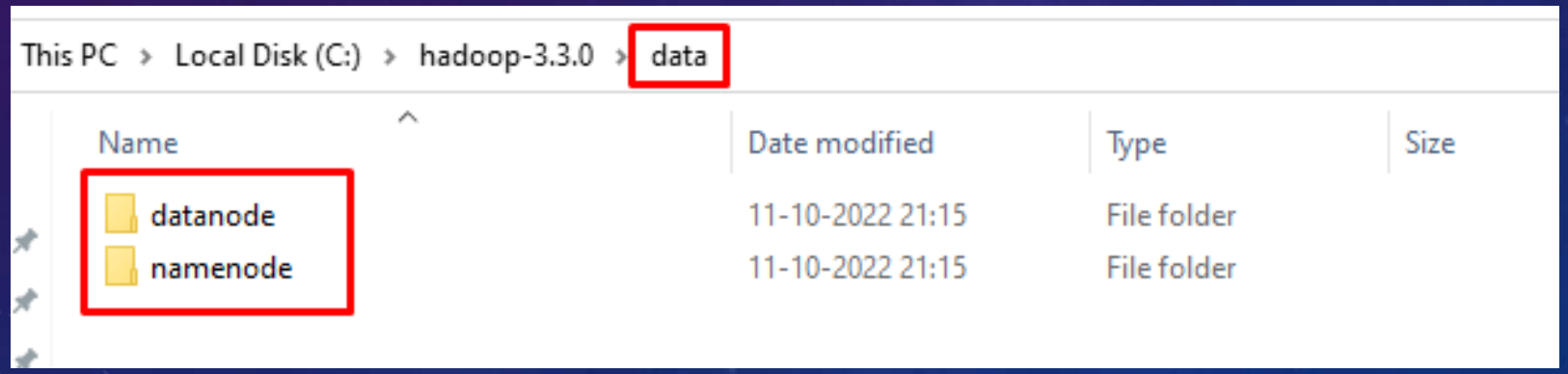
Open thi file and paste the xml code between configuration in file and save.

```
<configuration>
  <property>
      <name>mapreduce.framework.name</name>
      <value>yarn</value>
  </property>
</configuration>
```

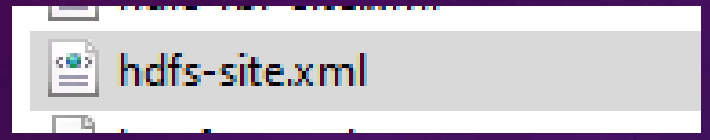
Create folder “data” under “C:\Hadoop-3.3.0”

Create folder “datanode” under “C:\Hadoop-3.3.0\data”

Create folder “namenode” under “C:\Hadoop-3.3.0\data”



3. hdfs-site file



Open thi file and paste the xml code between configuration in file and save.

```
<configuration>
  <property>
      <name>dfs.replication</name>
      <value>1</value>
  </property>
  <property>
      <name>dfs.namenode.name.dir</name>
      <value>/hadoop-3.3.0/data/namenode</value>
  </property>
  <property>
      <name>dfs.datanode.data.dir</name>
      <value>/hadoop-3.3.0/data/datanode</value>
  </property>
</configuration>
```

4. yarn-site file

yarn-site.xml

Open thi file and paste the xml code between configuration in file and save.

```
<configuration>
<property>
            <name>yarn.nodemanager.aux-services</name>
            <value>mapreduce_shuffle</value>
  </property>
  <property>
            <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
            <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
</configuration>
```

Set the JAVA_HOME inside the Hadoop environment

5. Edit file C:/Hadoop-3.3.0/etc/hadoop/hadoop-env.cmd

hadoop-env.cmd

```
@rem set JAVA_HOME in this file, so that it is correctly d
@rem remote nodes.

@rem The java implementation to use.  Required.
set JAVA_HOME=%JAVA_HOME%
```

Open thi file and paste the your java folder path till jdk here.

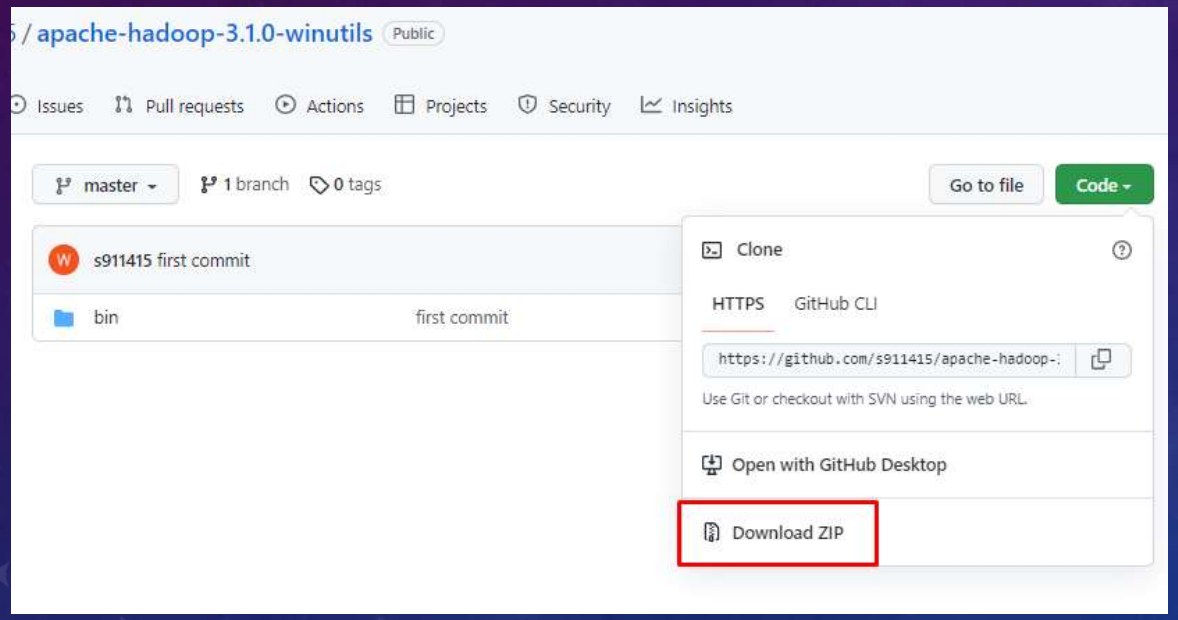
```
@rem remote nodes.

@rem The java implementation to use.  Required.
set JAVA_HOME=C:\Java\jdk1.8.0_202
```

Hadoop Configurations -:

Download

https://github.com/s911415/apache-hadoop-3.1.0-winutils



– After downloading unzip this folder and copy bin folder and replace existing bin folder in C:\Hadoop-3.3.0\bin

Testing -:

Open cmd and type command “hdfs namenode –format”

C:\Windows\System32\cmd.exe

```
Microsoft Windows [Version 10.0.19041.572]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\hadoop-3.3.0\bin>hdfs namenode -format
```

– Open cmd and change directory to C:\Hadoop-3.3.0\sbin
– type start-all.cmd

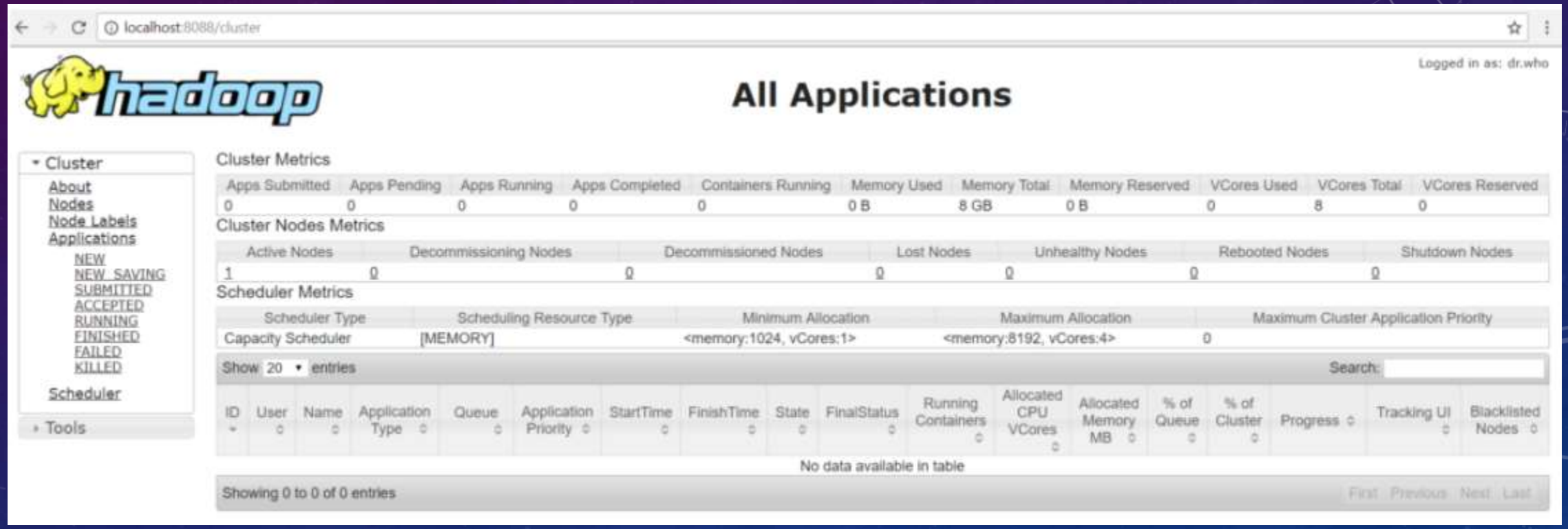
C:\Windows\System32\cmd.exe

```
Microsoft Windows [Version 10.0.19041.572]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\hadoop-3.3.0\sbin>start-all.cmd
```

Make sure these apps are running
– Hadoop Namenode
– Hadoop datanode
– YARN Resource Manager
– YARN Node Manager

Open browser and type localhost:8088



localhost:9870/dfshealth.html#tab-overview

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

Overview 'localhost:9000' (active)

Started:	Tue Oct 11 23:19:34 +0530 2022
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af
Compiled:	Tue Jul 07 00:14:00 +0530 2020 by brahma from branch-3.3.0
Cluster ID:	CID-536979ac-a2bd-49c0-a15e-6b3d58293535
Block Pool ID:	BP-2095255360-192.168.43.27-1665510490303

Summary

Thank You