

HADOOP INSTALLATION IN WINDOWS

BRIJESH JOTANIYA - 20BSIT019

ISHA MALLI - 20BSIT079

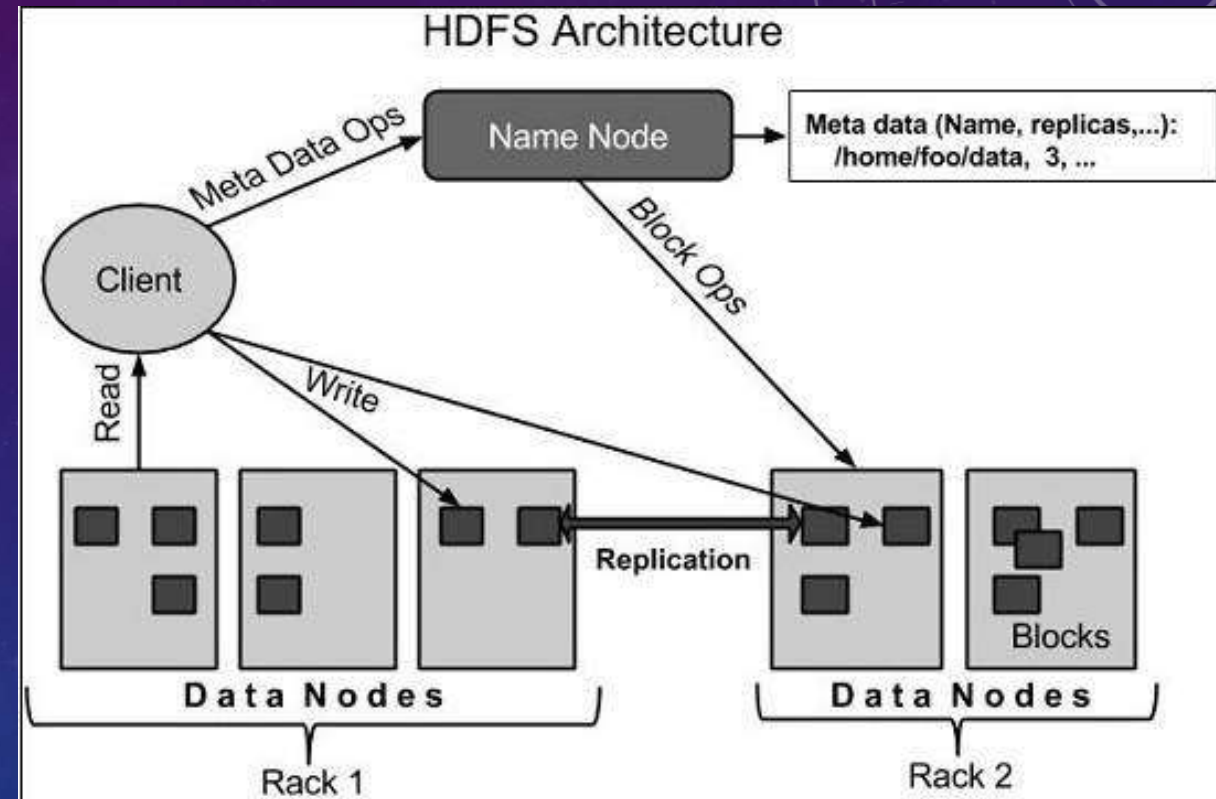
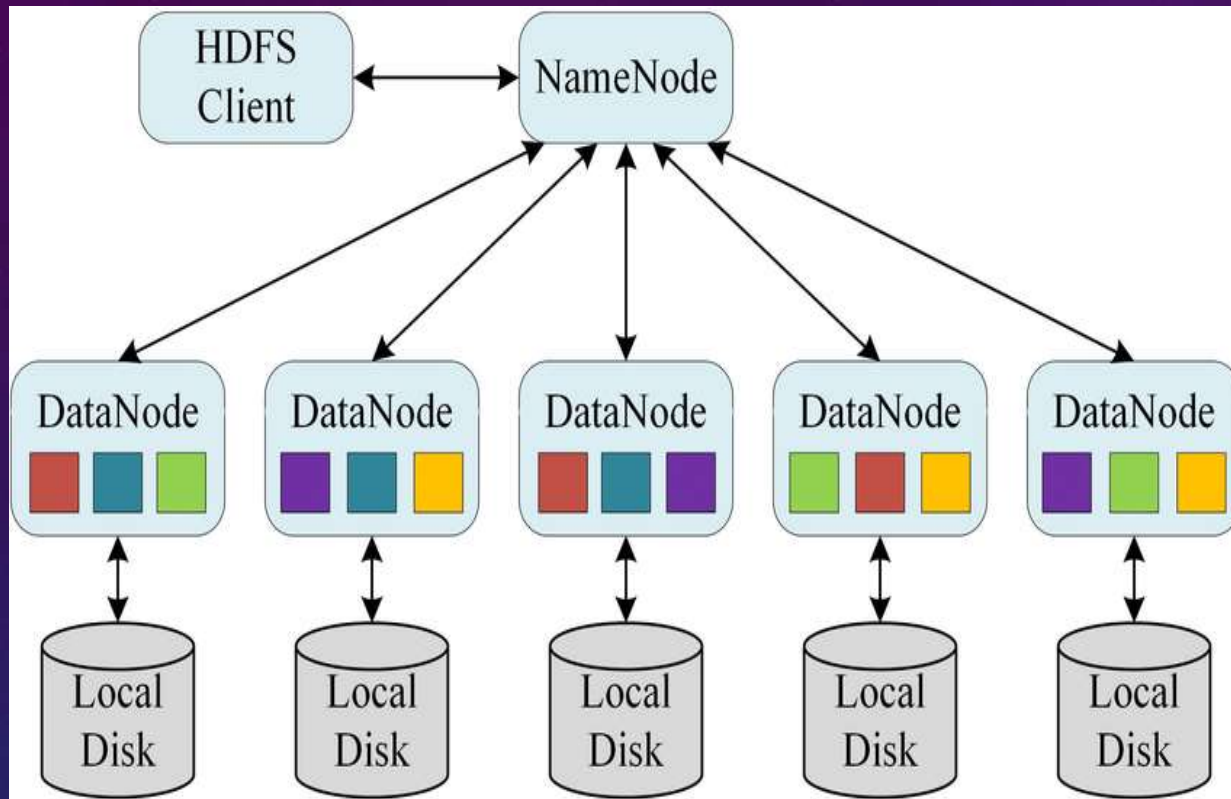
Introduction To Hadoop -:

- Hadoop is an open source framework that allows to store & process large data sets in a parallel & distributed manner.
- Two main components HDFS & MapReduce.
- Hadoop Distributed File System(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 -:

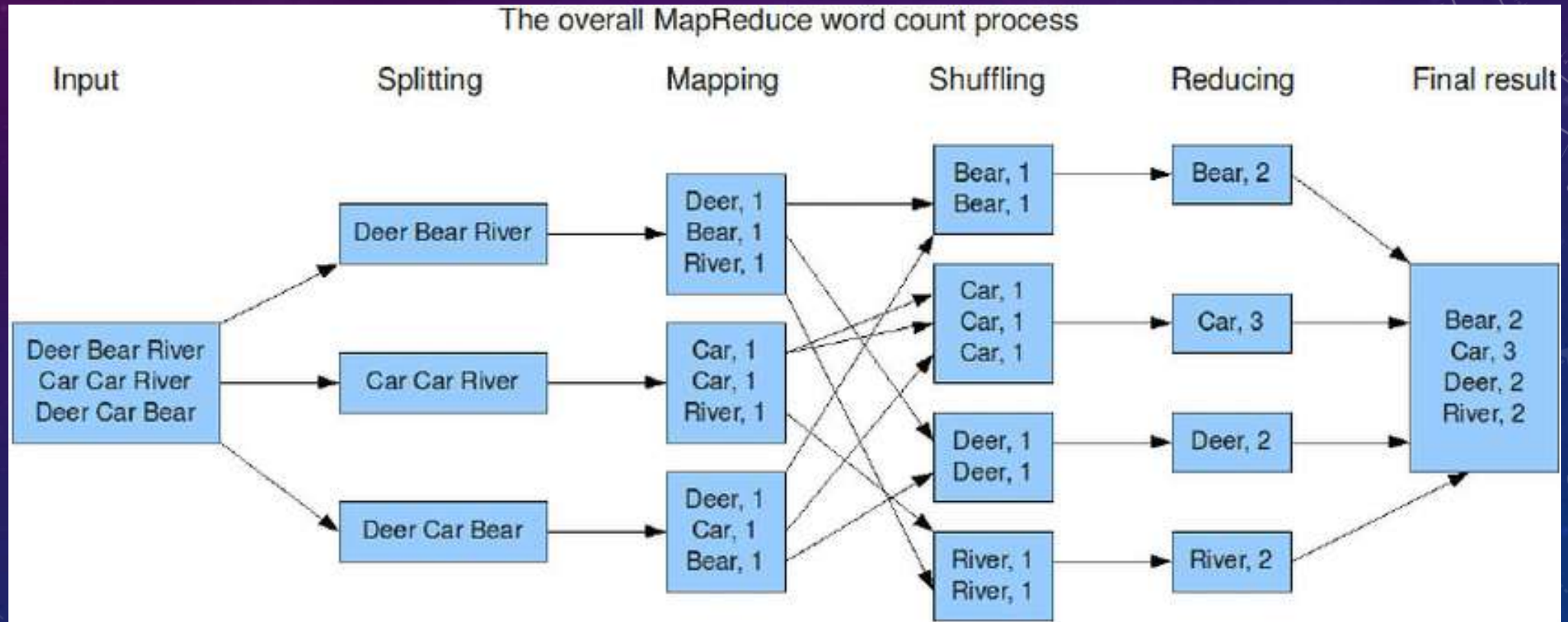


- HDFS client -:
 - With the help of Name node HDFS client can read and write the file.
- Name node -:
 - Name node is also known as Master node.
 - Main task is how to store the data. It stores the data in distributed manner.
 - It has file system namespace which have info related data stored in data 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 the actual data storage.
 - It will divide file into 128 MB partition.
 - The DataNode performs the low level read and write requests from the file system's client.

MapReduce -:

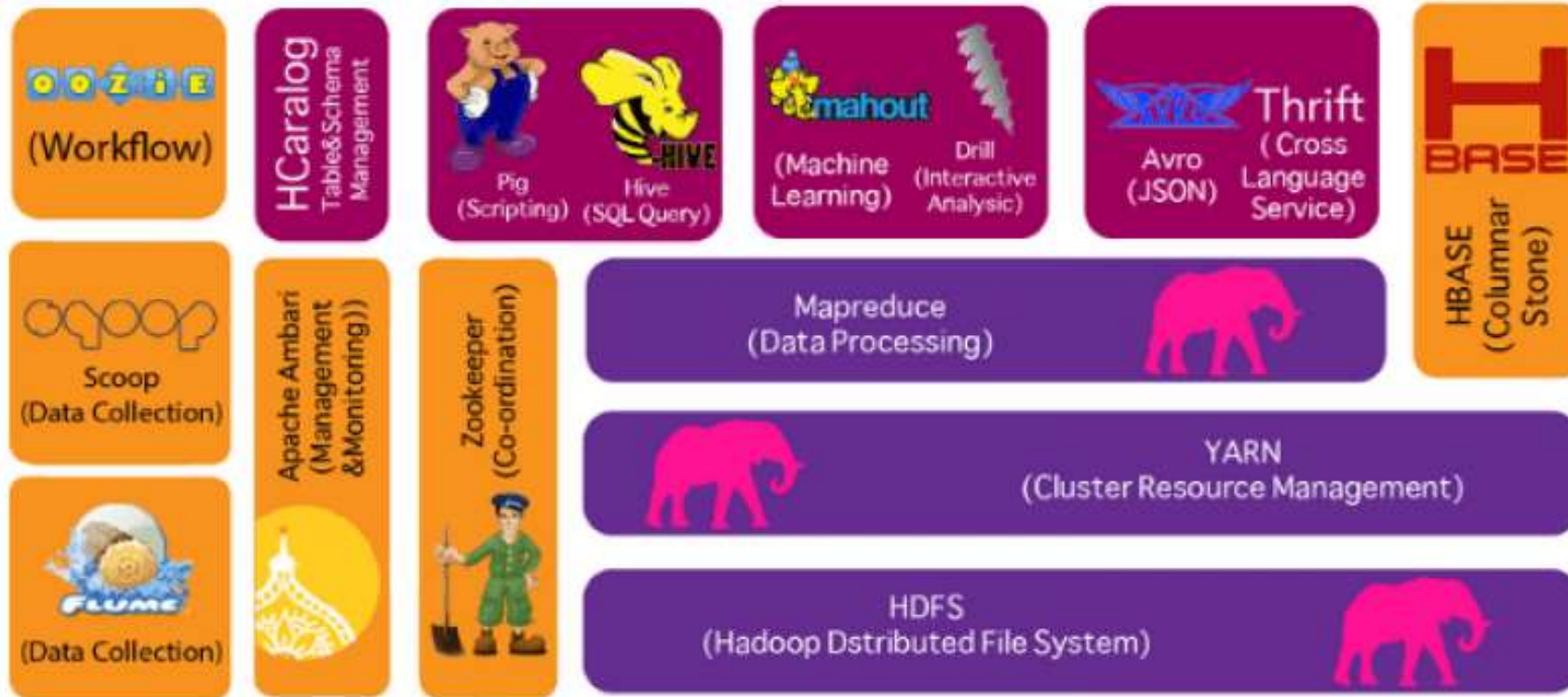
- It is core component of Apache Hadoop.
- MapReduce performs the processing of large data sets in distributed and parallel manner.
- MapReduce consists of two distinct tasks – Map And Reduce.
- Two essential daemons of MapReduce :
 - Job Tracker -: Schedules jobs and tracks the assign jobs to Task tracker.
 - 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.

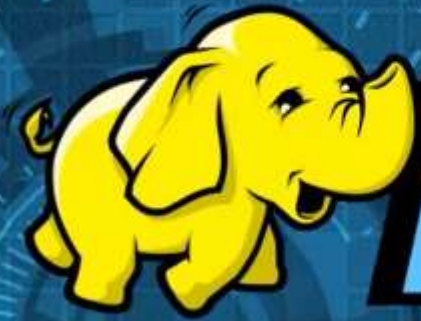
Example -:



Hadoop Ecosystem -:

Top Hadoop Ecosystem Components





hadoop

Big Data

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

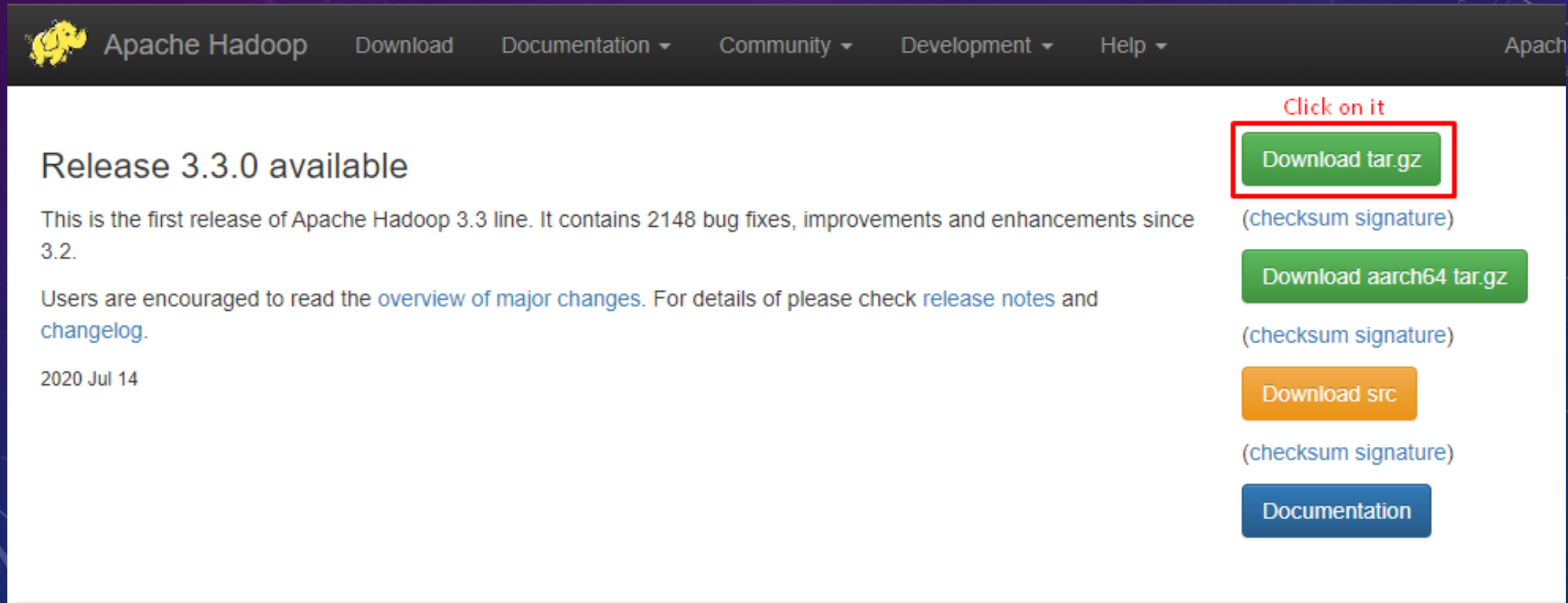
<https://www.oracle.com/in/java/technologies/javase/javase8-archive-downloads.html>

Download and install jdk based on your computer specification.

Today (1)
 jdk-8u202-windows-x64.exe

Download Hadoop Binaries

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



The screenshot shows the Apache Hadoop website's release page for version 3.3.0. The page has a dark navigation bar at the top with the Apache Hadoop logo and links for Download, Documentation, Community, Development, and Help. The main content area is white and features the heading "Release 3.3.0 available". Below this, a paragraph states that this is the first release of the Apache Hadoop 3.3 line, containing 2148 bug fixes, improvements, and enhancements since version 3.2. It encourages users to read the overview of major changes and check the release notes and changelog for details. The date "2020 Jul 14" is displayed. On the right side, there are four buttons: "Download tar.gz" (highlighted with a red box and the text "Click on it"), "Download aarch64 tar.gz", "Download src", and "Documentation". Each button is followed by a link for "(checksum signature)".

Apache Hadoop Download Documentation Community Development Help

Release 3.3.0 available

This is the first release of Apache Hadoop 3.3 line. It contains 2148 bug fixes, improvements and enhancements since 3.2.

Users are encouraged to read the [overview of major changes](#). For details of please check [release notes](#) and [changelog](#).

2020 Jul 14





[Click on it](#)
Download tar.gz
([checksum signature](#))

Download aarch64 tar.gz
([checksum signature](#))














Download src
([checksum signature](#))

Documentation

Extract this file in C drive

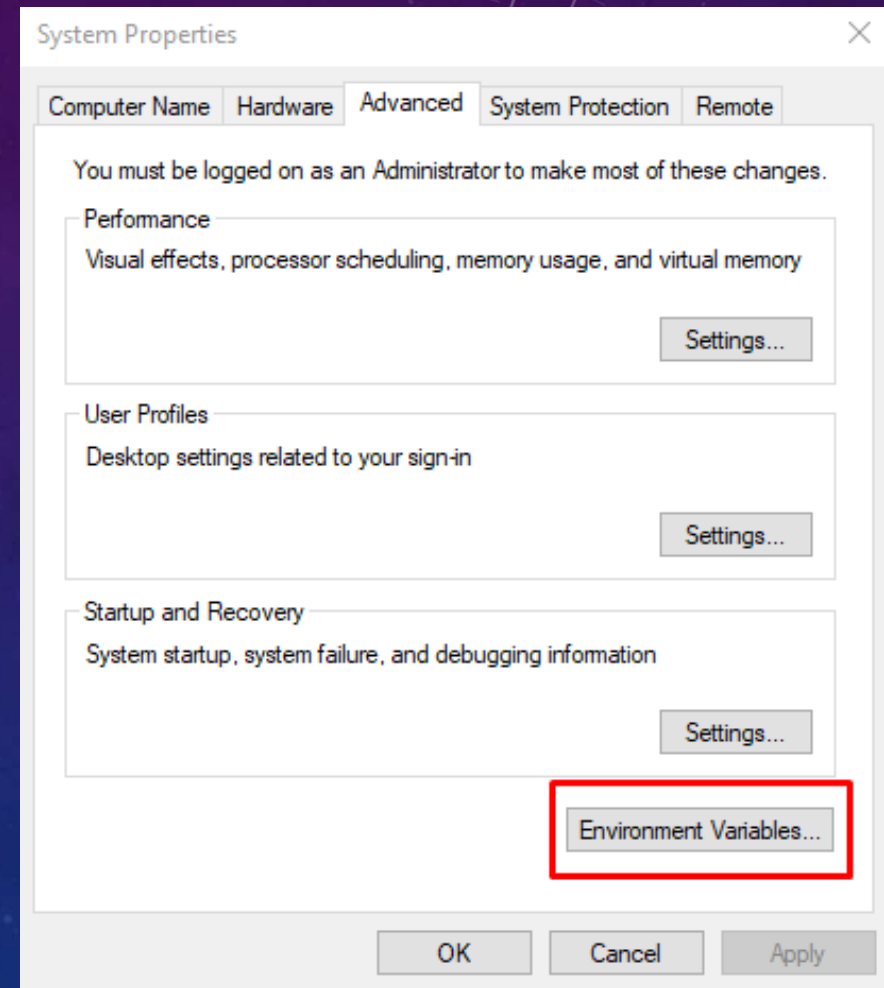
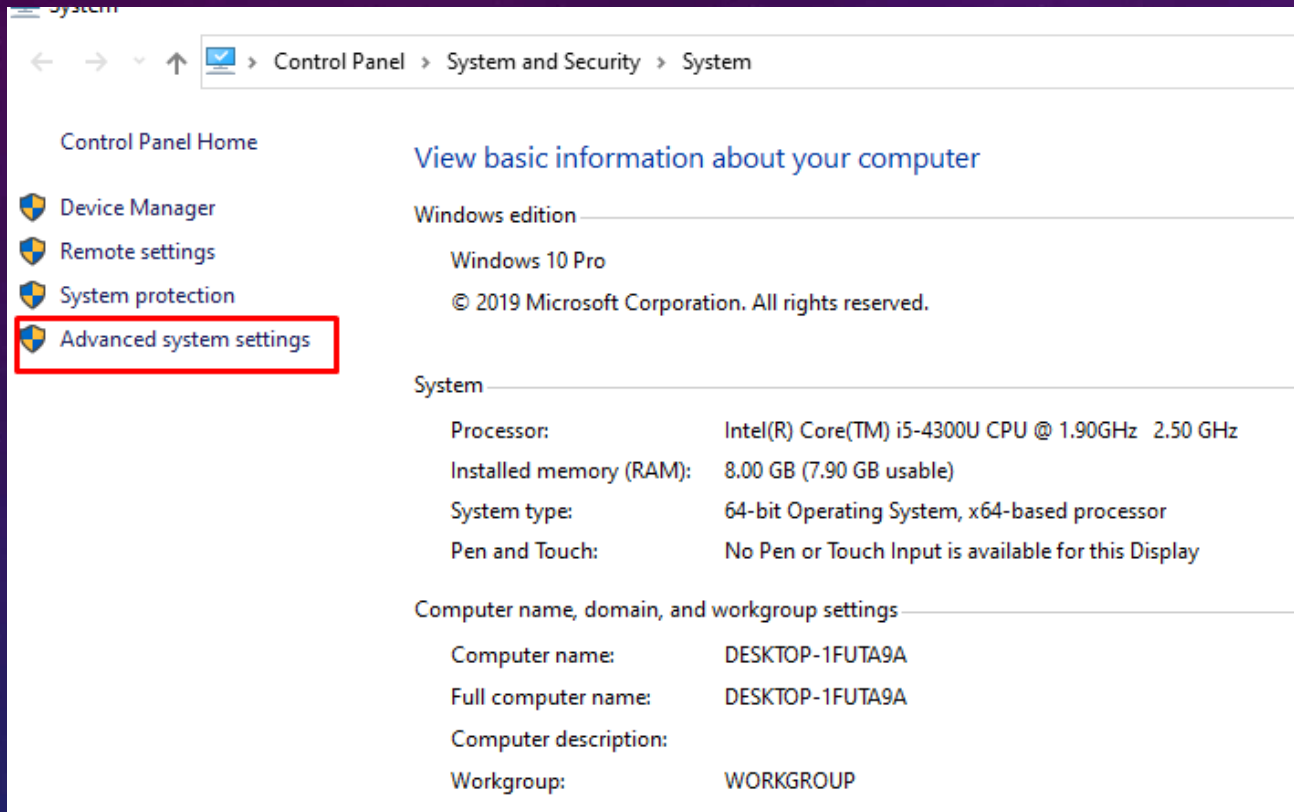
Name	Date modified	Type	Size
 hadoop-3.3.0.tar.gz	11-10-2022 19:21	WinRAR archive	4,89,013 KB
<div><div>Open</div><div> Share with Skype</div><div> Open with WinRAR</div><div> Extract files...</div></div>			

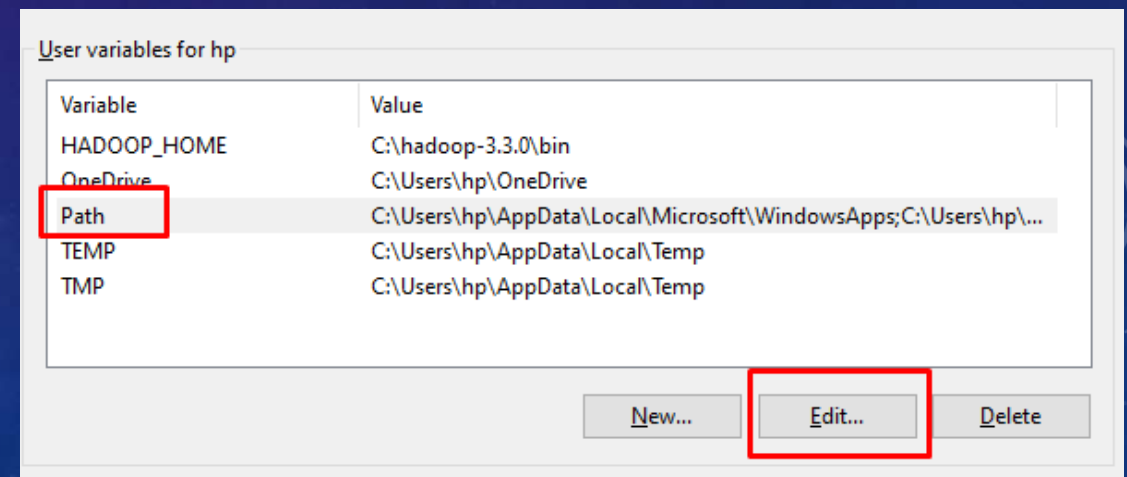
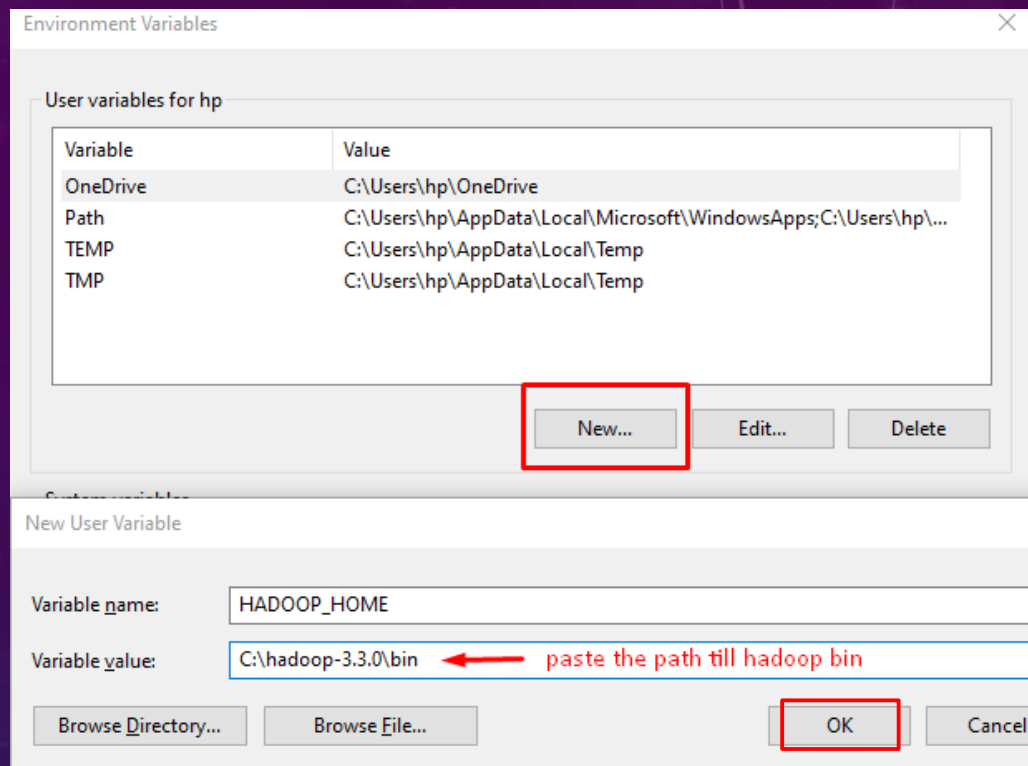
is PC > Local Disk (C:) > hadoop-3.3.0 >

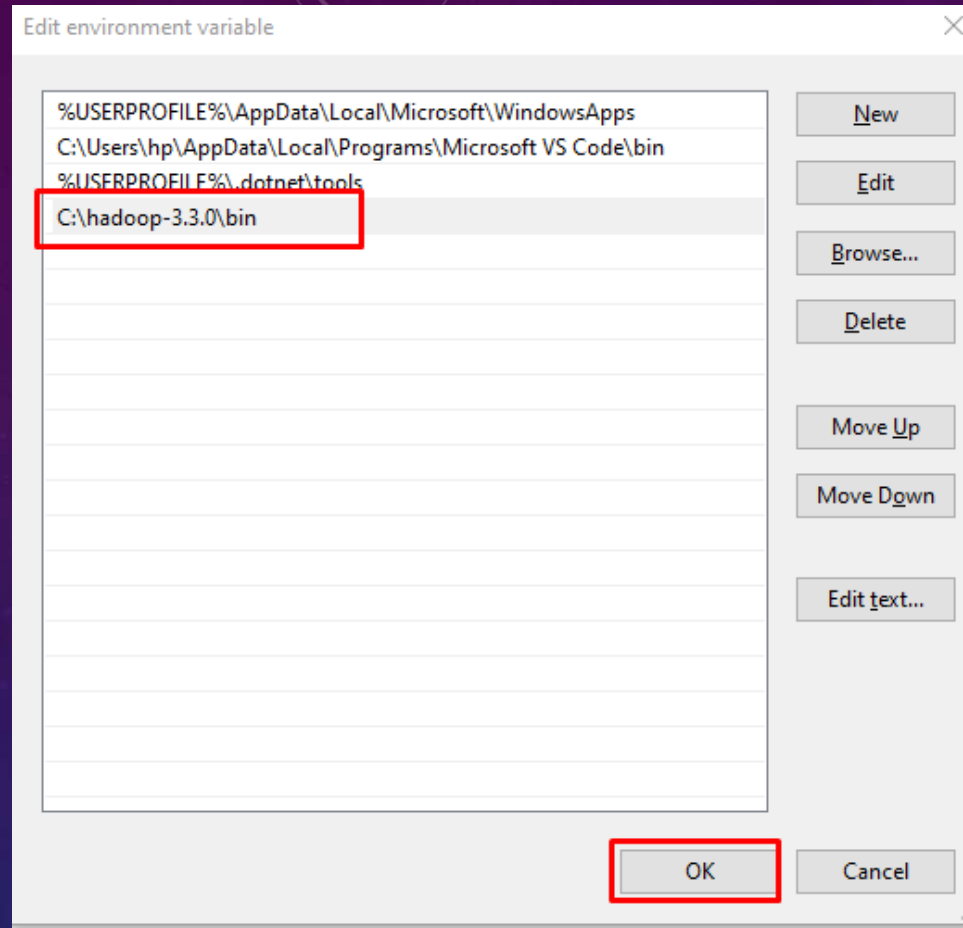
Name	Date modified	Type	Size
 bin	11-10-2022 19:44	File folder	
 etc	11-10-2022 19:47	File folder	
 include	11-10-2022 19:44	File folder	
 lib	11-10-2022 19:47	File folder	
 libexec	11-10-2022 19:44	File folder	
 licenses-binary	11-10-2022 19:44	File folder	
 sbin	11-10-2022 19:44	File folder	
 share	11-10-2022 19:45	File folder	
 LICENSE.txt	24-03-2020 22:53	Text Document	16 KB
 LICENSE-binary	04-07-2020 22:59	File	23 KB
 NOTICE.txt	24-03-2020 22:53	Text Document	2 KB
 NOTICE-binary	24-03-2020 22:53	File	27 KB
 README.txt	24-03-2020 22:53	Text Document	1 KB

← This is unzipped folder

Set path in environment variable -:

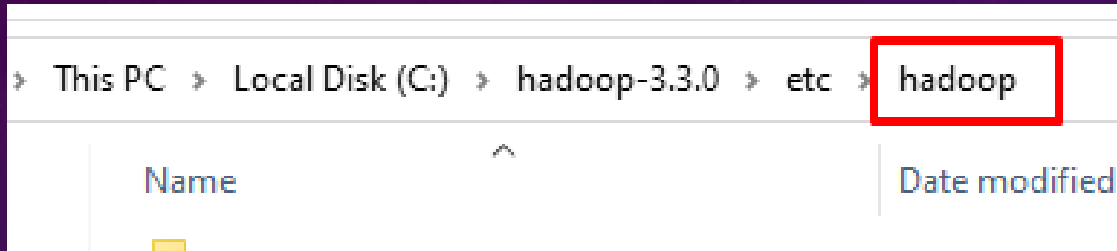






Similarly set the path for the Java.
Set the path till bin..

Edit Few Files(xml Files)-:



Go in this folder

core-site.xml	open this file	07-07-2020 00:16	XML Document	1 KB
---------------	----------------	------------------	--------------	------

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> paste this code between configuration

<property>

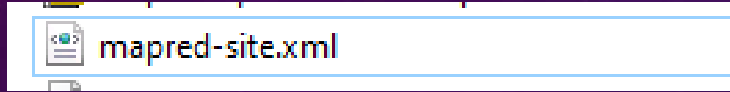
<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

2. mapred-site file



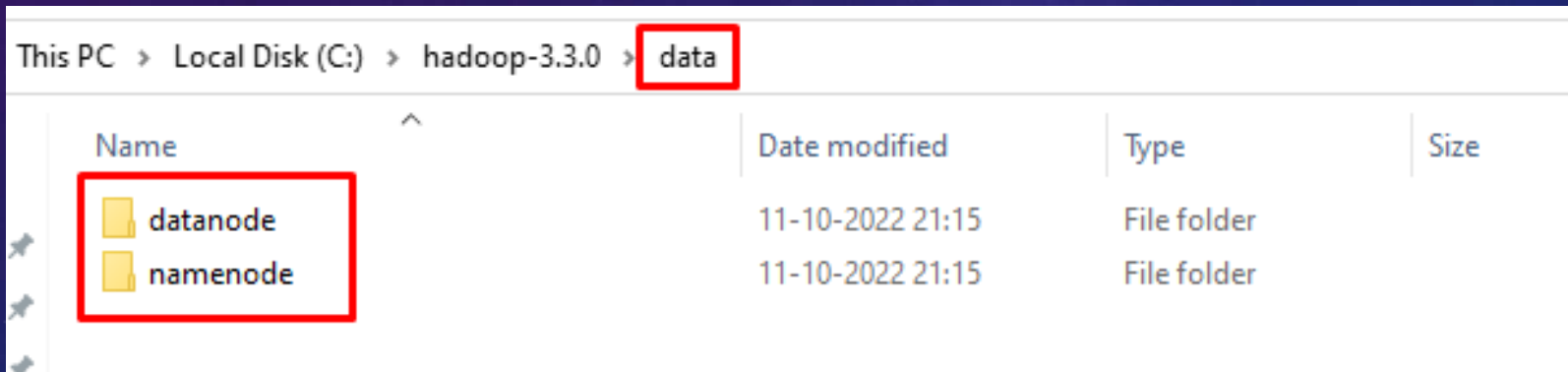
Open thi file and paste the xml code between configuration in file and save.

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

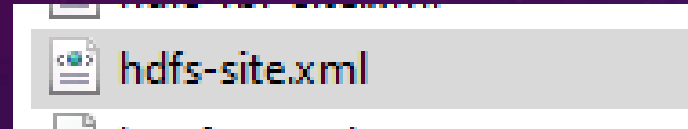
Create folder “data” under “C:\Hadoop-3.3.0”

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

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



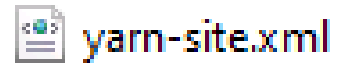
3. hdfs-site file



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



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.

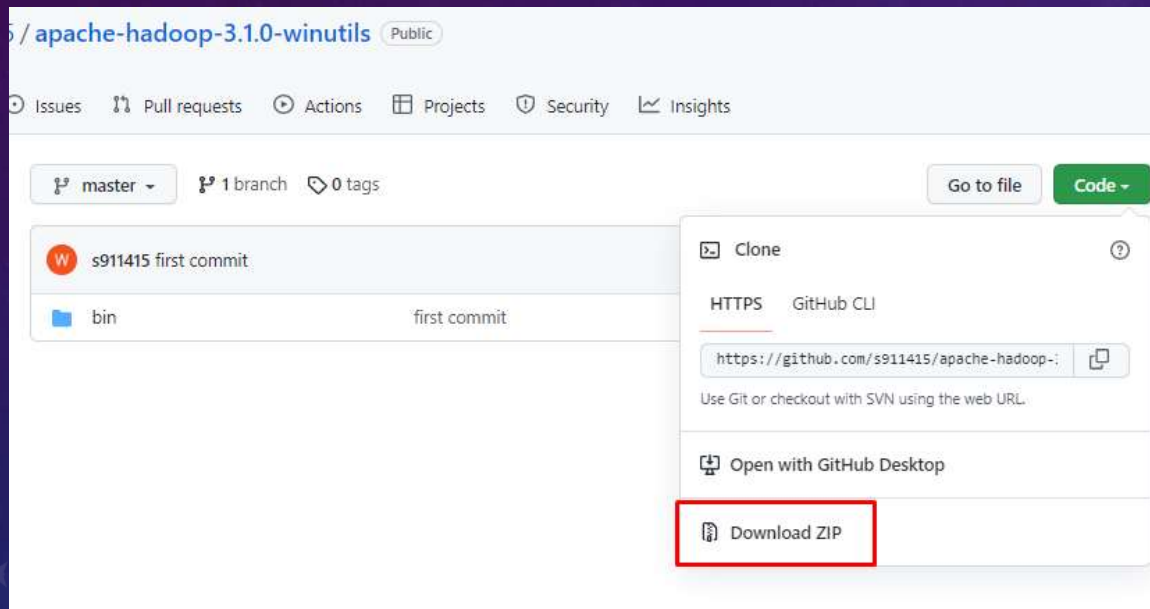
```
@rem remote nodes.
```

```
@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:8088/cluster ☆ ⋮

 **hadoop**

Logged in as: dr.who

All Applications

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	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	8 GB	0 B	0	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																	

Showing 0 to 0 of 0 entries

First Previous Next Last

→ ↺ ⓘ 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