

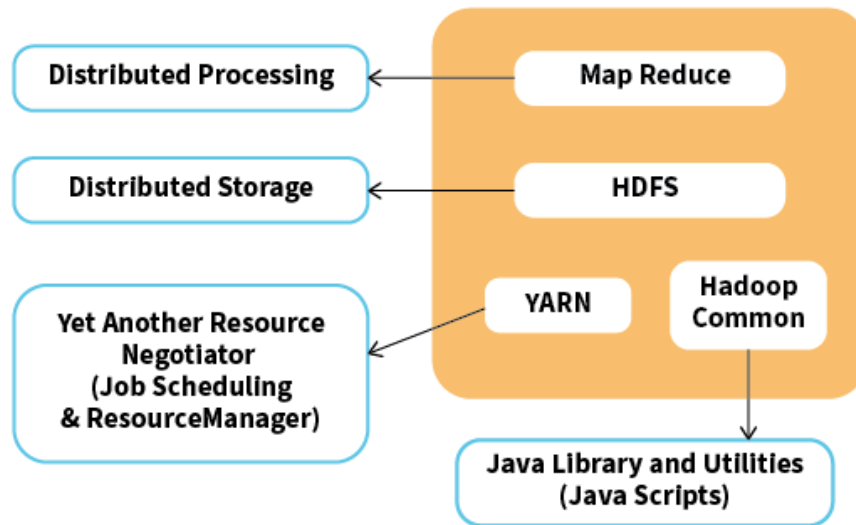


Name:	Prerna Sunil Jadhav
Sap Id:	60004220127
Class:	T. Y. B. Tech (Computer Engineering)
Course:	Big Data Infrastructure Laboratory
Course Code:	DJ19CEEL6011
Experiment No.:	02

**AIM:** Install Hadoop on a Single Node Cluster.

### WHAT IS HADOOP & WHY IS IT IMPORTANT?

- ✚ Hadoop is an open-source software programming framework for storing a large amount of data and performing the computation. Its framework is based on Java programming with some native code in C and shell scripts.
- ✚ 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.
- ✚ Hadoop has two main components:
  - HDFS (Hadoop Distributed File System): This is the storage component of Hadoop, which allows for the storage of large amounts of data across multiple machines. It is designed to work with commodity hardware, which makes it cost-effective.
  - YARN (Yet Another Resource Negotiator): This is the resource management component of Hadoop, which manages the allocation of resources (such as CPU and memory) for processing the data stored in HDFS.
  - Hadoop also includes several additional modules that provide additional functionality, such as Hive (a SQL-like query language), Pig (a high-level platform for creating MapReduce programs), and HBase (a non-relational, distributed database).
  - Hadoop is commonly used in big data scenarios such as data warehousing, business intelligence, and machine learning. It's also used for data processing, data analysis, and data mining. It enables the distributed processing of large data sets across clusters of computers using a simple programming model.
- ✚ Hadoop is important as one of the primary tools to store and process huge amounts of data quickly. It does this by using a distributed computing model which enables the fast processing of data that can be rapidly scaled by adding computing nodes.
- ✚ Hadoop Architecture
  - Hadoop stands as a robust platform for storing and processing vast amounts of data. It serves as a key solution for storing and analysing data from diverse sources, including databases, web servers, and file systems.
  - Built on the MapReduce programming algorithm, Hadoop architecture comprises four key components, each playing a crucial role in managing and processing extensive datasets.
    - HDFS (Hadoop Distributed File System)
    - MapReduce
    - YARN (Yet Another Resource Negotiator)
    - Common Utilities or Hadoop Common



#### INSTALLATION:

Install Hadoop 2.9.1 on Windows 10

First download the Hadoop 2.9.1 from the below link.

<https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.9.1/hadoop-2.9.1.tar.gz>

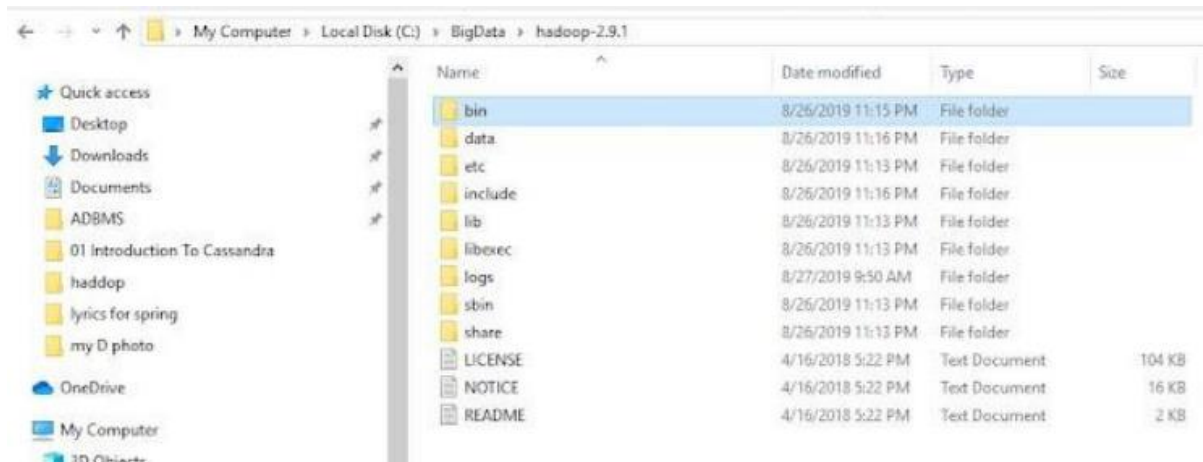


Create a folder path as below and copy the downloaded msi into this folder.

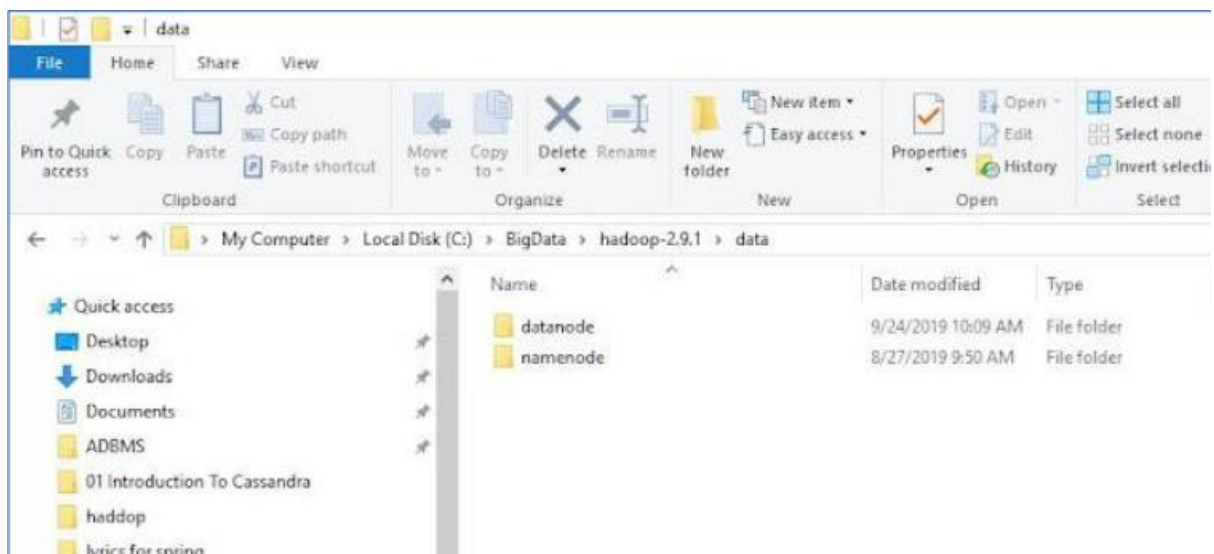
Path:- 'C:/BigData/hadoop-2.9.1'



Academic Year: 2022-2023



Go to C:/BigData/hadoop-2.9.1 and create a folder 'data'. Inside the 'data' folder create two folders 'datanode' and 'namenode'.



Then Set Hadoop Environment Variables

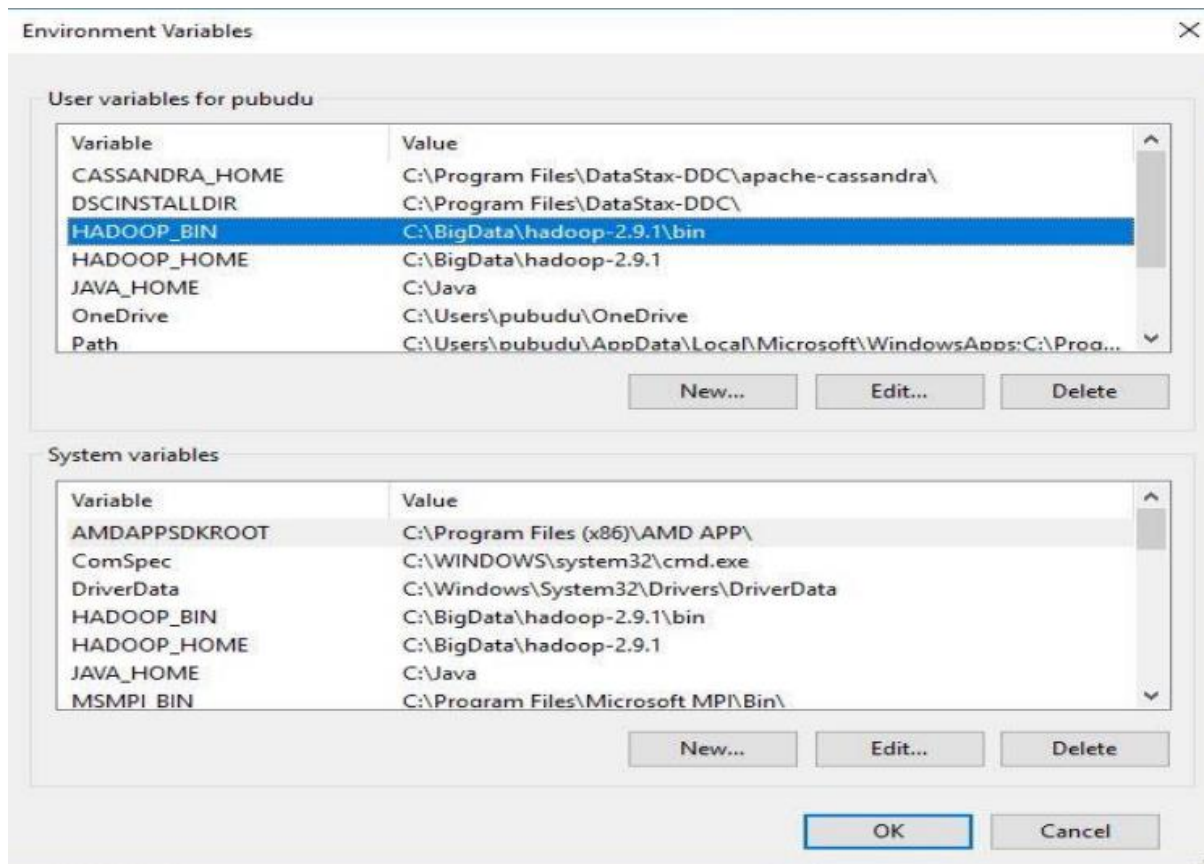
**HADOOP\_HOME="C:\BigData\hadoop-2.9.1"**

**HADOOP\_BIN="C:\BigData\hadoop-2.9.1\bin"**

**JAVA\_HOME=<JDK installation location>"**

To set these variables, go to My Computer or This PC. Right click --> Properties --> Advanced

System settings --> Environment variables. Click New to create a new environment variables



To validate the above setting, open new cmd and check the output.

```
echo %HADOOP_HOME%
```

```
echo %HADOOP_BIN%
```

```
echo %PATH%
```



Academic Year: 2022-2023

```
Command Prompt
Microsoft Windows [Version 10.0.17134.1006]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\pubudu>echo %HADOOP_HOME%
C:\BigData\hadoop-2.9.1

C:\Users\pubudu>echo %HADOOP_BIN%
C:\BigData\hadoop-2.9.1\bin

C:\Users\pubudu>echo %PATH%
C:\Program Files (x86)\Common Files\Oracle\Java\javapath;F:\Oracle\product\12.2.0\dbhome_1\bin;C:\Program Files\Microsoft MPI\Bin\;C:\Program Files (x86)\AND APP\bin\x86_64;C:\Program Files (x86)\AND APP\bin\x86;C:\Program Files (x86)\Intel\iCLS Client\;C:\Program Files\Intel\iCLS Client\;C:\WINDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wbem;C:\WINDOWS\System32\WindowsPowerShell\v1.0\;C:\Program Files\Intel\Intel(R) Management Engine Components\DAL;C:\Program Files\Intel\Intel(R) Management Engine Components\IPT;C:\Program Files (x86)\Intel\Intel(R) Management Engine Components\DAL;C:\Program Files (x86)\Intel\Intel(R) Management Engine Components\IPT;C:\Program Files\WIDCOMM\Bluetooth Software\;C:\Program Files\WIDCOMM\Bluetooth Software\syswow64;C:\Program Files (x86)\ATI Technologies\ATI.ACE\Core-Static;C:\Program Files (x86)\Skype\Phone\;C:\WINDOWS\System32\OpenSSH\;C:\Program Files (x86)\Microsoft SQL Server\140\Tools\Binn\;C:\Program Files\Microsoft SQL Server\140\Tools\Binn\;C:\Program Files (x86)\Microsoft SQL Server\140\DTSDTS\Binn\;C:\Program Files\Microsoft SQL Server\140\DTSDTS\Binn\;C:\Program Files\Microsoft SQL Server\140\ODBC\130\Tools\Binn\;C:\Program Files (x86)\Microsoft SQL Server\150\DTSDTS\Binn\;C:\Program Files\dotnet\;C:\Program Files\Microsoft SQL Server\130\Tools\Binn\;C:\Program Files\Microsoft SQL Server\Client SDK\ODBC\170\Tools\Binn\;C:\Program Files (x86)\Microsoft SQL Server\110\DTSDTS\Binn\;C:\Program Files (x86)\Microsoft SQL Server\120\DTSDTS\Binn\;C:\Program Files (x86)\Microsoft SQL Server\130\DTSDTS\Binn\;C:\sqlite3;C:\Program Files\Java\jdk1.8.0_221\bin;C:\Java;C:\BigData\hadoop-2.9.1;C:\BigData\hadoop-2.9.1\bin;C:\BigData\hadoop-2.9.1\sbin;C:\Users\pubudu\AppData\Local\Microsoft\WindowsApps;C:\Program Files\Java\jdk1.8.0_221\bin;C:\Program Files\Java\jdk1.7.0_25\bin;

C:\Users\pubudu>
```

To configure the Hadoop on windows we have to edit below mention files in the extracted

location.

1. hadoop-env.cmd
2. core-site.xml
3. hdfs-site.xml
4. mapred-site.xml
5. yarn-site.xml

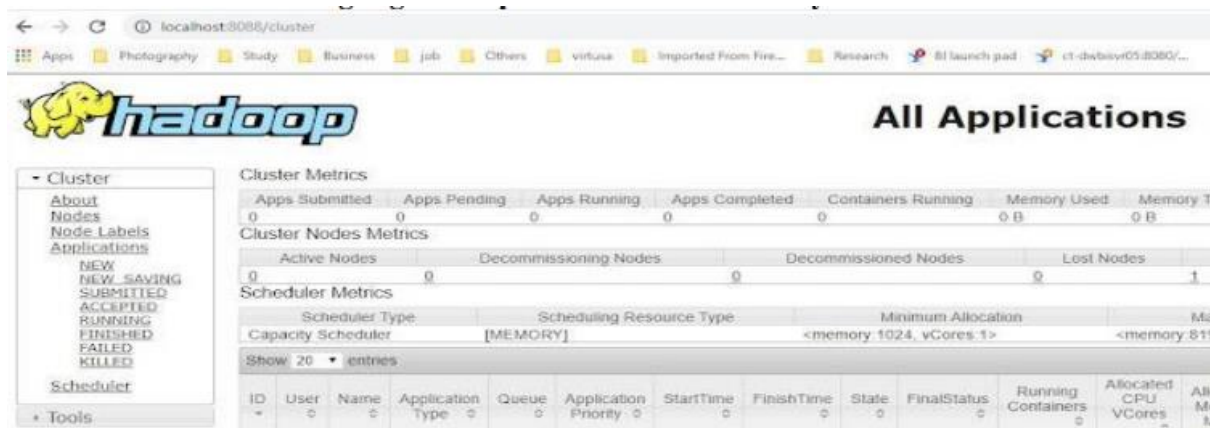
Now you can access all the Hadoop components via web urls.

To access Resource Manager go to <http://localhost:8088> from your web browser.





Academic Year: 2022-2023



The screenshot shows the Hadoop All Applications page in a web browser. The address bar shows 'localhost:8088/cluster'. The page has a sidebar with a 'Cluster' menu and a 'Tools' section. The main content area displays 'Cluster Metrics' and 'Scheduler Metrics'.

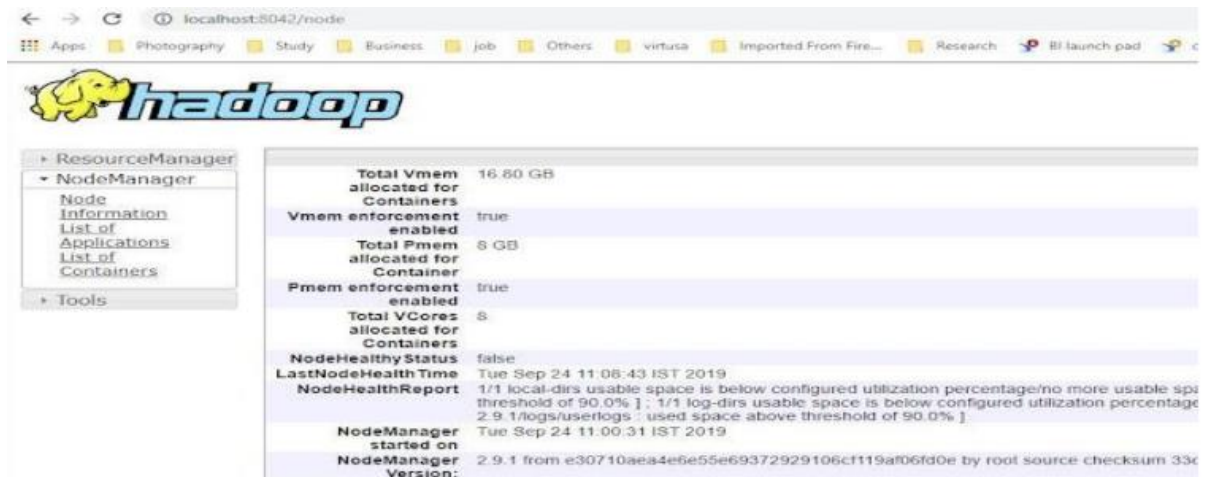
Cluster Metrics							
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory T	
0	0	0	0	0	0 B	0 B	

Cluster Nodes Metrics				
Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	
0	0	0	1	

Scheduler Metrics			
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Ms
Capacity Scheduler	[MEMORY]	<memory 1024, vCores 1>	<memory 81

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Al

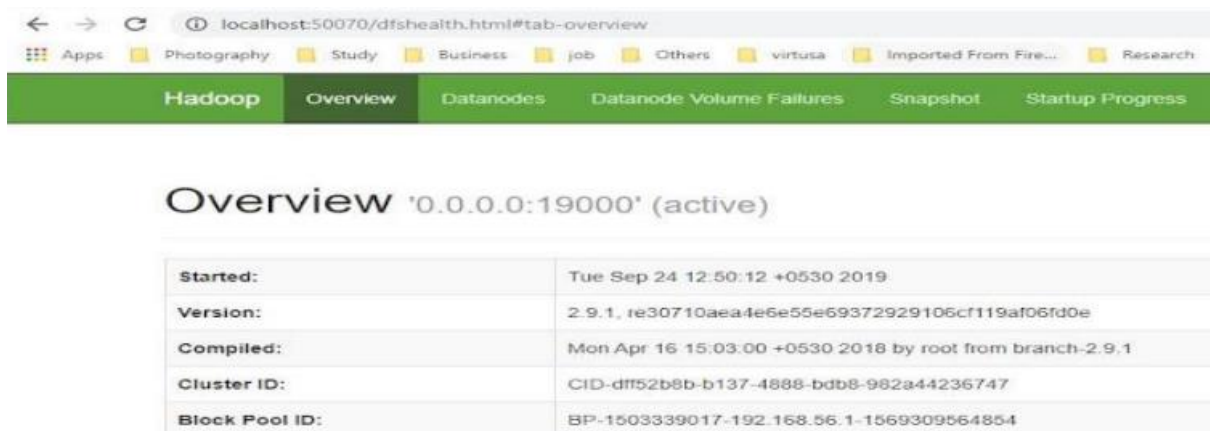
To access Node Manager go to <http://localhost:8042> from your web browser.



The screenshot shows the Hadoop NodeManager page in a web browser. The address bar shows 'localhost:8042/node'. The page has a sidebar with a 'ResourceManager' menu and a 'Tools' section. The main content area displays various metrics for the NodeManager.

Total Vmem allocated for Containers	16.80 GB
Vmem enforcement enabled	true
Total Pmem allocated for Container	8 GB
Pmem enforcement enabled	true
Total VCoers allocated for Containers	8
NodeHealthyStatus	false
LastNodeHealthTime	Tue Sep 24 11:08:43 IST 2019
NodeHealthReport	1/1 local-dirs usable space is below configured utilization percentage/no more usable spi threshold of 90.0% ] : 1/1 log-dirs usable space is below configured utilization percentage 2.9.1/logs/userlogs : used space above threshold of 90.0% ]
NodeManager started on	Tue Sep 24 11:00:31 IST 2019
NodeManager Version:	2.9.1 from e30710aea4e6e55e69372929106cf119af06fd0e by root source checksum 33c

To access Name Node go to <http://localhost:50070> from your web browser



The screenshot shows the Hadoop Overview page in a web browser. The address bar shows 'localhost:50070/dfshealth.html#tab-overview'. The page has a sidebar with a 'Hadoop' menu and a 'Tools' section. The main content area displays the 'Overview' of the NameNode.

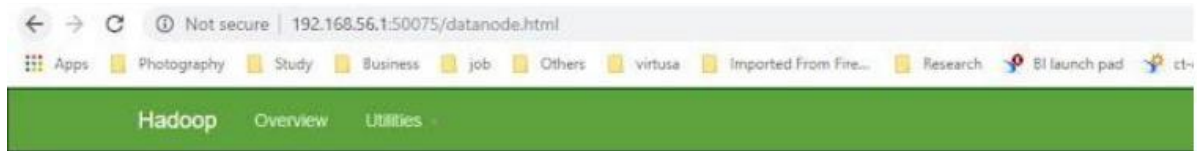
Overview '0.0.0.0:19000' (active)	
Started:	Tue Sep 24 12:50:12 +0530 2019
Version:	2.9.1, re30710aea4e6e55e69372929106cf119af06fd0e
Compiled:	Mon Apr 16 15:03:00 +0530 2018 by root from branch-2.9.1
Cluster ID:	CID-dff52b8b-b137-4888-bdb8-982a44236747
Block Pool ID:	BP-1503339017-192.168.56.1-1569309564854

## Summary



Academic Year: 2022-2023

To access Data Node go to <http://localhost:50075> from your web browser.



## DataNode on 192.168.56.1:50010

Cluster ID:	CID-dff52b8b-b137-4888-bdb8-982a44236747
Version:	2.9.1

## Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Repo
0 0 0 0:19000	BP-1503339017-192.168.56.1-1569309564854	RUNNING	2s	a few seconds

**CONCLUSION:** Hence, we successfully installed Hadoop