

EXP NO: 1

INSTALLATION OF HADOOP

INTRODUCTION OF HADOOP:

Hadoop is an open-source framework designed for storing and processing large datasets in a distributed computing environment. Managed by the Apache Software Foundation, Hadoop allows organizations to handle vast amounts of data across clusters of commodity hardware, providing both scalability and fault tolerance. It is built on two core components: the Hadoop Distributed File System (HDFS) and the Yet Another Resource Negotiator (YARN).

1.1 HISTORY OF HADOOP:

Hadoop, an open-source framework overseen by the Apache Software Foundation and written in Java, is designed for storing and processing large datasets across clusters of commodity hardware. It addresses two main big data challenges: data storage and processing. Traditional RDBMS systems fall short due to data heterogeneity. Hadoop's core components are the Hadoop Distributed File System (HDFS) and Yet Another Resource Negotiator (YARN).

Hadoop originated from the Apache Nutch project, started by Doug Cutting and Mike Cafarella in 2002, aimed at creating a search engine to index 1 billion pages. Faced with high costs, they were inspired by Google's GFS and MapReduce papers. In 2006, Cutting, supported by Yahoo, separated the distributed computing parts from Nutch, forming Hadoop. It became an Apache open-source project in 2008. By 2011, Hadoop 1.0 was released, with Hadoop 3.0 following in 2017, significantly enhancing its scalability and performance.

1.2 VERSIONS OF HADOOP:

1. Hadoop 0.x Series (2006-2009):

Initial releases, primarily focused on proving the concept. Introduced basic HDFS and MapReduce functionalities. Limited scalability and stability.

2. Hadoop 1.x Series (2011):

Formal release with significant stability improvements. Featured HDFS for distributed storage and MapReduce for data processing. Used JobTracker and TaskTracker for resource management, with scalability limited to thousands of nodes.

3. Hadoop 2.x Series (2013):

- Major architectural change with the introduction of YARN (Yet Another Resource Negotiator).
- Separated resource management and job scheduling, allowing multiple applications beyond MapReduce.
- Enhanced scalability and support for different processing models.

4.Hadoop 3.x Series (2017):

- Added support for erasure coding, reducing storage overhead.
- Introduced HDFS Federation, allowing multiple namespaces and scaling HDFS clusters.
- Improved resource management and scheduling in YARN, including GPU support.
- Enhanced security and monitoring features.

1.3 INSTALLATION STEPS:

1.INSTALL JAVA

2.INSTALL AND UNZIP HADOOP

3.SETTING UP ENVIRONMENT VARIABLES

User variables for shash

Variable	Value
FLUME_CONF	D:\Shashank\Study\Flume\conf
FLUME_HOME	D:\Shashank\Study\Flume
HADOOP_HOME	D:\Shashank\Study\hadoop-2.9.2
HIVE_HOME	D:\Shashank\Study\Hive-3.1.2
JAVA_HOME	C:\Program Files\Java\jdk1.8.0_221
OneDrive	C:\Users\shash\OneDrive

System variables

Variable	Value
Path	C:\Windows\System32;D:\Softwares\Python\Scripts\;D:\Softw...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC;.PY;.PYW
platformcode	KV
PROCESSOR_ARCHITECTU...	AMD64
PROCESSOR_IDENTIFIER	Intel64 Family 6 Model 142 Stepping 12, GenuineIntel
PROCESSOR_LEVEL	6
PROCESSOR_REVISION	9c0c

4.EDITING CONFIGURATION FILES

Core-site.xml

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

Hdfs-site.xml

```
<configuration>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
<property>
  <name>dfs.namenode.name.dir</name>
  <value>C:\hadoop-3.3.6\data\namenode</value>
</property>
<property>
  <name>dfs.datanode.data.dir</name>
  <value>C:\hadoop-3.3.6\data\datanode</value>
</property>
</configuration>
```

Mapred-site.xml

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

Yarn-site.xml

```
<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>
```

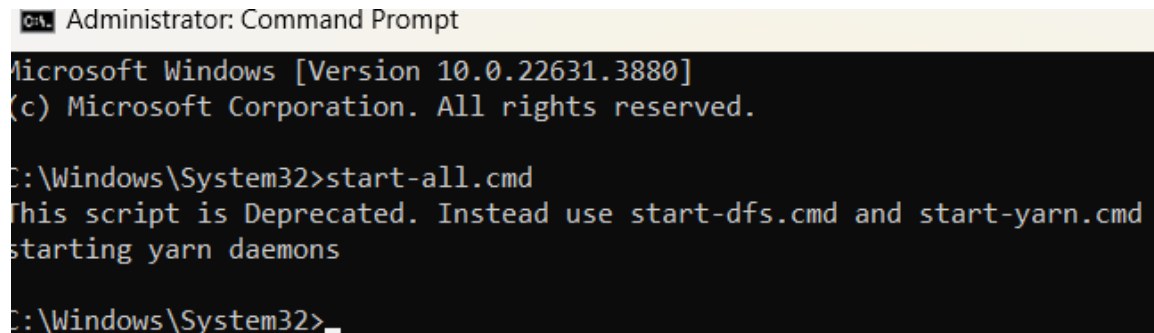
5.REPLACE BIN FOLDER

6.FORMAT NAMENODE AND DATANODE

hadoop namenode -format

hadoop datanode -format

7.START HADOOP



Administrator: Command Prompt

```
Microsoft Windows [Version 10.0.22631.3880]
(c) Microsoft Corporation. All rights reserved.


C:\Windows\System32>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\Windows\System32>
```

```
C:\Windows\System32>jps
27744 Jps
18904 NodeManager
10044 NameNode
17196 ResourceManager
17692 DataNode

C:\Windows\System32>
```

8.RUNNING HADOOP



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	
0	0	0	0	0	<m

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned N
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Alloc
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
Showing 0 to 0 of 0 entries									

Namenode information

localhost:9870/dfshealth.html#tab-overview

ASUS Software Port...MyASUS Software ~...McAfee LiveSafeGmailYouTubeMaps

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview

'localhost:9000' (active)

Started:	Tue Aug 13 07:54:43 +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-9ff10afd-9f2a-4744-9bc6-f45b2553ad17
Block Pool ID:	BP-468074218-172.16.11.91-1722913037250