Lab Exercise 1

Aim: Perform setting up and Installing Hadoop in Hadoop Distributing File System (HDFS)

Theory:

Hadoop is an open-source framework based on Java that manages the storage and processing of large amounts of data for applications. Hadoop uses distributed storage and parallel processing to handle big data and analytics jobs, breaking workloads down into smaller workloads that can be run at the same time.

Four modules comprise the primary Hadoop framework and work collectively to form the Hadoop ecosystem:

1. **Hadoop Distributed File System (HDFS)**: As the primary component of the Hadoop ecosystem, HDFS is a distributed file system in which individual Hadoop nodes operate on data that resides in their local storage. This removes network latency, providing high-throughput access to application data. In addition, administrators don't need to define schemas up front.
2. **Yet Another Resource Negotiator (YARN)**: YARN is a resource-management platform responsible for managing compute resources in clusters and using them to schedule users' applications. It performs scheduling and resource allocation across the Hadoop system.
3. **MapReduce**: MapReduce is a programming model for large-scale data processing. In the MapReduce model, subsets of larger datasets and instructions for processing the subsets are dispatched to multiple different nodes, where each subset is processed by a node in parallel with other processing jobs. After processing the results, individual subsets are combined into a smaller, more manageable dataset.
4. **Hadoop Common**: Hadoop Common includes the libraries and utilities used and shared by other Hadoop modules.

Procedure:

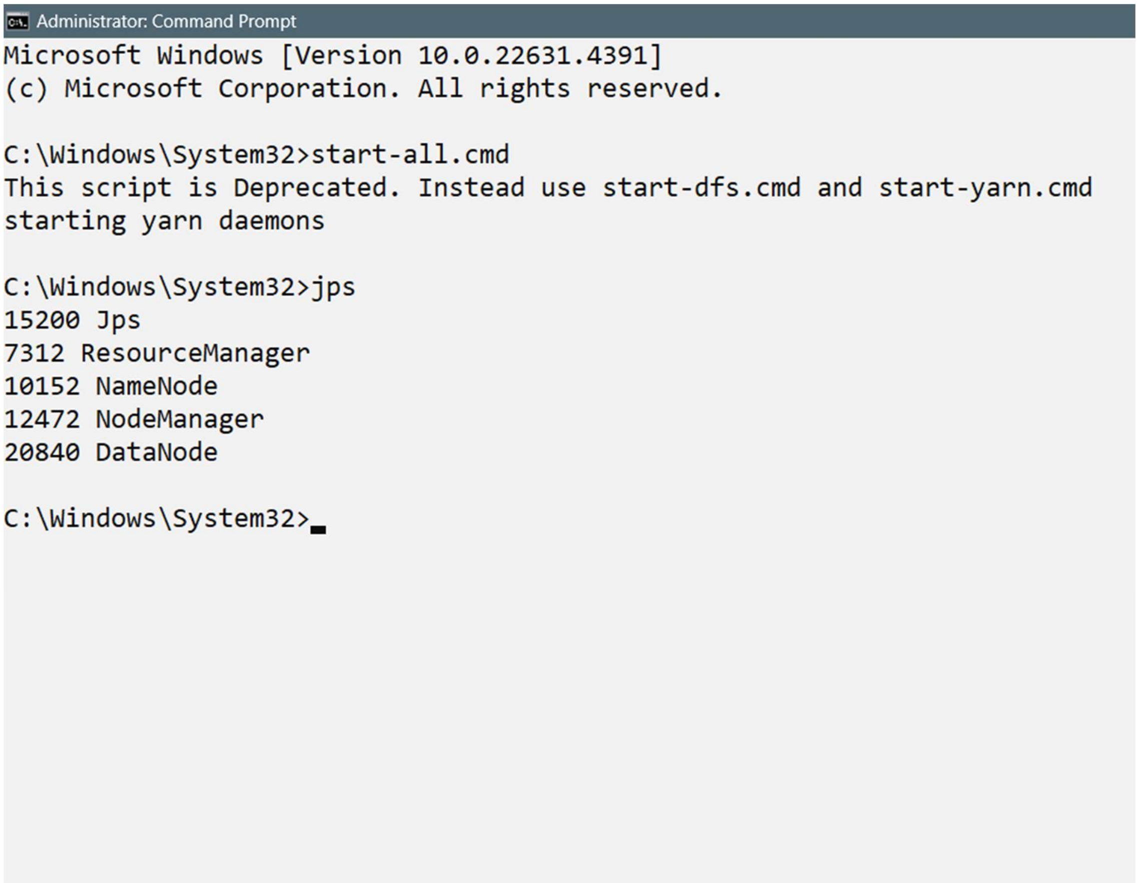
1. Download and install Java in folder created in C drive named "Java" then setup environment variable like below:
   Environment Variable→ New→ JAVA_HOME →C:/java/jdk1.8/bin→ Ok

   Then setup the path in Environment variable:
   Path→Edit→Add→add file path of java/jdk1.8/bin
2. Download and install Apache Hadoop in a C drive and rename as "hadoop" and setup the Environment variable like below:
   Environment Variable→ New→ HADOOP_HOME →C:/hadoop/bin→Ok

Then setup the environment path as follow,

Path→Edit →Add →add file path of hadoop/bin and hadoop/sbin

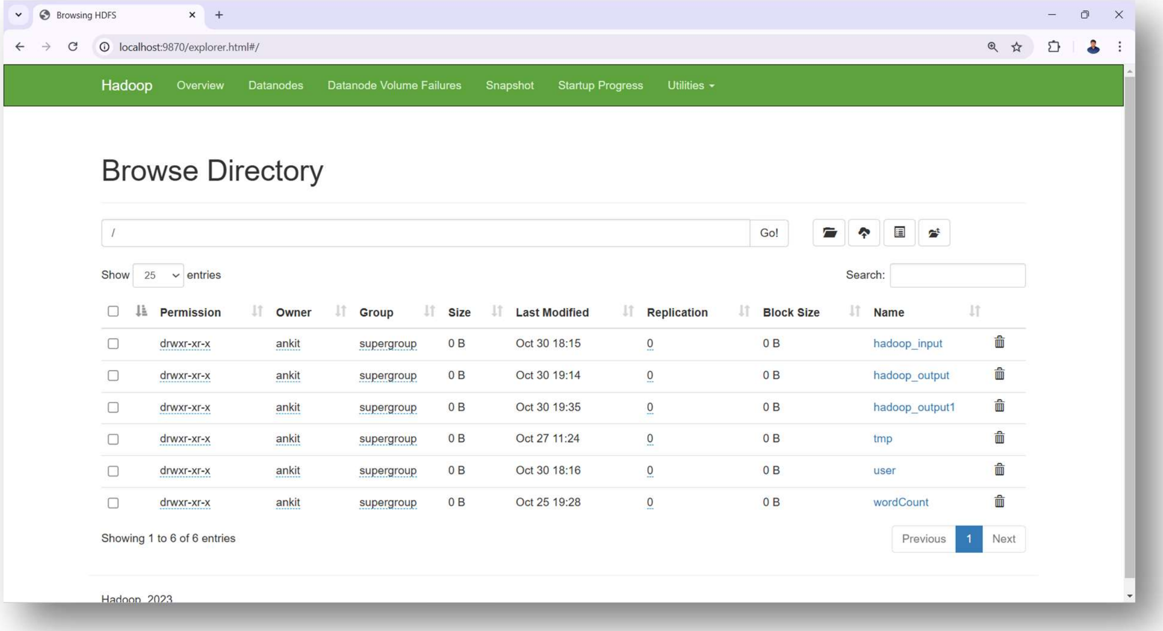
3. Change the hadoop-env folder in hadoop/etc/hadoop path. Add java/jdk1.8 path at the place of %JAVA_HOME%.
4. Add a new folder name “data” in hadoop. And also, two subfolders in data named first “namenode” and second “datanode”.
5. Setup Hadoop core-site.xml, hdfs-site.xml, mapred-site.xml and yarn-site.xml.
6. Open Command Prompt and “run as administrator” and start write the following,
    a. >>> hadoop namenode -format   //it will format your namenode
    b. >>> cd /
    c. >>> cd hadoop/sbin
    d. >>> start-dfs.cmd   //it will start your namenode and datanode
    e. >>>start-yarn.cmd   //starts the development server
    f. >>>jps   //it will show all namenode and datanode running



```
Microsoft Windows [Version 10.0.22631.4391]
(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>jps
15200 Jps
7312 ResourceManager
10152 NameNode
12472 NodeManager
20840 DataNode

C:\Windows\System32>
```

7. To ensure that Hadoop is properly installed, open a web browser and go to **https://localhost:9870** and **https://localhost:8088**. This will launch the web interface for the Hadoop NameNode. You should see a page with Hadoop cluster information.

Output:



Browse Directory

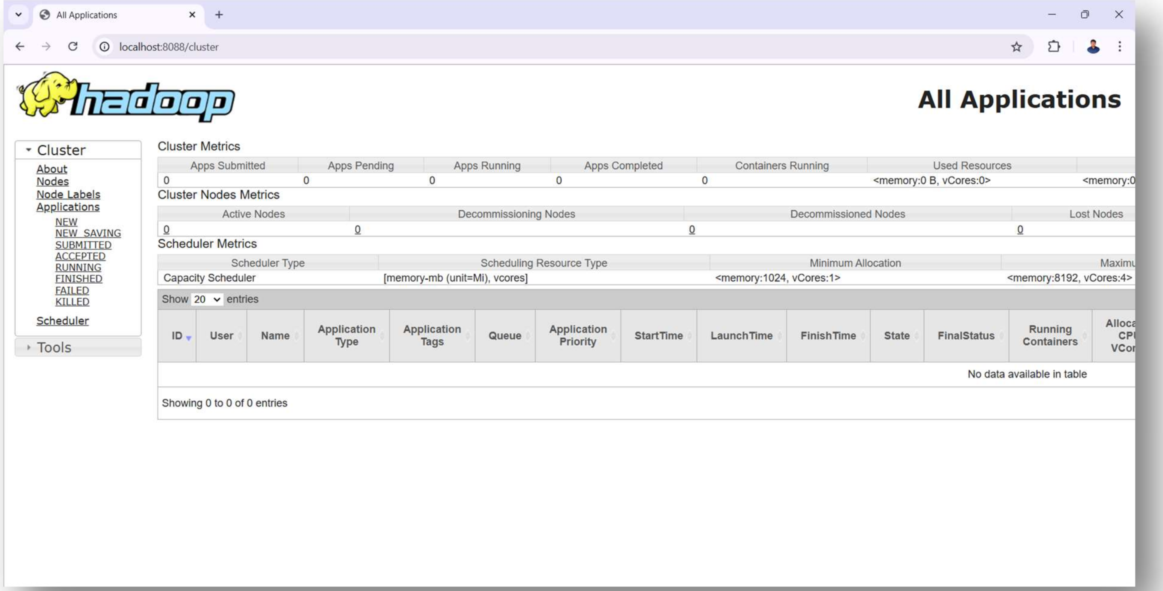
/

Show 25 entries Search:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	drwxr-xr-x	ankit	supergroup	0 B	Oct 30 18:15	0	0 B	hadoop_input	🗑
□	drwxr-xr-x	ankit	supergroup	0 B	Oct 30 19:14	0	0 B	hadoop_output	🗑
□	drwxr-xr-x	ankit	supergroup	0 B	Oct 30 19:35	0	0 B	hadoop_output1	🗑
□	drwxr-xr-x	ankit	supergroup	0 B	Oct 27 11:24	0	0 B	tmp	🗑
□	drwxr-xr-x	ankit	supergroup	0 B	Oct 30 18:16	0	0 B	user	🗑
□	drwxr-xr-x	ankit	supergroup	0 B	Oct 25 19:28	0	0 B	wordCount	🗑

Showing 1 to 6 of 6 entries

Hadoop, 2023



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	
0	0	0	0	0	<memory:0 B, vCores:0>	<memory:0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes
0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximu
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Alloca CPI VCor
												No data available in table	

Showing 0 to 0 of 0 entries

Conclusion:

In this experiment we have learnt about the hadoop and hdfs system and successfully we install hadoop in our local PC.