

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 set up the path in Environment variable:

Path → Edit → Add → add file path of java/jdk1.8/bin

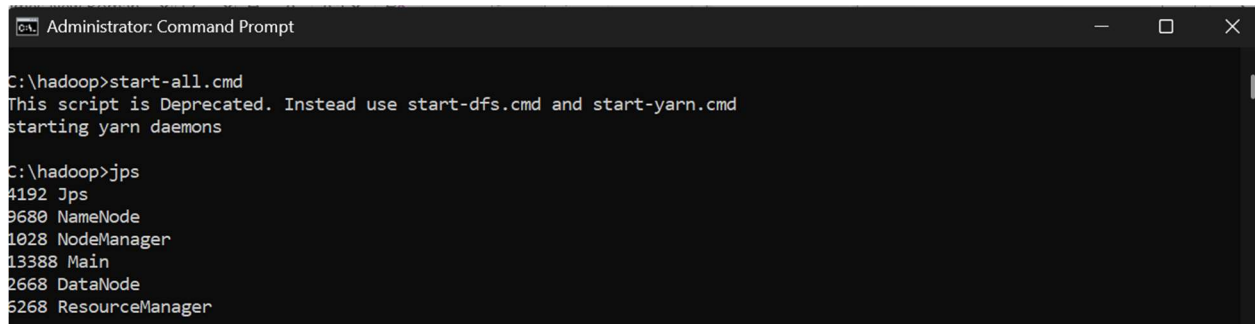
2. Download and install Apache Hadoop in a C drive and rename as "hadoop" and setup the Environment variable like below:

Environment Variable New → HADOOP_HOME → C:/hadoop/binOk

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



```
Administrator: Command Prompt
C:\hadoop>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\hadoop>jps
4192 Jps
9680 NameNode
1028 NodeManager
13388 Main
2668 DataNode
6268 ResourceManager
```

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:

The screenshot shows the 'Browse HDFS' web interface in a browser. The address bar shows 'localhost:9870/explorer.html/'. The interface has a green header with navigation tabs: Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is titled 'Browse Directory'. Below the title is a search bar with the text '/' and a 'Go!' button. There are also icons for file operations. Below the search bar, it says 'Show 25 entries'. A table lists the directory contents with columns: Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. The table contains 8 entries, all with size 0 B and replication 0.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	baghe	supergroup	0 B	Sep 04 10:18	3	128 MB	1.docx
-rw-r--r--	baghe	supergroup	0 B	Sep 04 10:17	3	128 MB	1.txt
drwxr-xr-x	baghe	supergroup	0 B	Sep 24 21:03	0	0 B	StopWordOutput
drwxr-xr-x	baghe	supergroup	0 B	Sep 11 10:54	0	0 B	output
drwxr-xr-x	baghe	supergroup	0 B	Sep 11 11:15	0	0 B	pawan
drwxr-xr-x	baghe	supergroup	0 B	Sep 04 10:20	0	0 B	priyanshu
drwxr-xr-x	baghe	supergroup	0 B	Sep 18 10:53	0	0 B	stopwordElimination
drwxr-xr-x	baghe	supergroup	0 B	Sep 04 11:57	0	0 B	tmp

Snapshot No. 1 (Browse HDFS at LocalHost:9870)

The screenshot shows the 'About the Cluster' web interface in a browser. The address bar shows 'localhost:8088/cluster/cluster'. The interface has a green header with the Hadoop logo and the title 'About the Cluster'. Below the header, there is a sidebar with a 'Cluster' menu and a 'Tools' menu. The main content area displays various cluster metrics and node information. The 'Cluster Metrics' section shows a table with columns: Apps Submitted, Apps Pending, Apps Running, Apps Completed, Containers Running, Used Resources, Total Resources, Reserved Resources, Physical Mem Used %, and Physical VCores Used %. The 'Cluster Nodes Metrics' section shows a table with columns: Active Nodes, Decommissioning Nodes, Decommissioned Nodes, Lost Nodes, Unhealthy Nodes, Rebooted Nodes, and Shutdown Nodes. The 'Scheduler Metrics' section shows a table with columns: Scheduler Type, Scheduling Resource Type, Minimum Allocation, Maximum Allocation, Maximum Cluster Application Priority, and Scheduler Busy %.

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %	Physical VCores Used %
1	0	0	1	0	<memory:0 B, vCores:0>	<memory:8 GB, vCores:8>	<memory:0 B, vCores:0>	73	0

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

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority	Scheduler Busy %
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0	0

Cluster overview

Cluster ID: 1758727781470

ResourceManager state: STARTED

ResourceManager HA state: active

ResourceManager HA zookeeper connection state: Could not find leader elector. Verify both HA and automatic failover are enabled.

ResourceManager RMStateStore: org.apache.hadoop.yarn.server.resourcemanager.recovery.NullRMStateStore

ResourceManager started on: Wed Sep 24 20:59:41 +0530 2025

ResourceManager version: 3.3.6 from 1be78238728da9266a4f88195058f08fd012bf9c by ubuntu source checksum d42eb795a5eadb0feb5e44a7f87a9 on 2023-06-18T08:31Z

Hadoop version: 3.3.6 from 1be78238728da9266a4f88195058f08fd012bf9c by ubuntu source checksum 5652179ad55f76cb287d9c633bb53bbd on 2023-06-18T08:22Z

Snapshot No. 2 (Cluster at LocalHost:8088)

Conclusion:

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