

Hands-on lab on Hadoop Cluster (20 mins)



What is a Hadoop Cluster?

A Hadoop cluster is a collection of computers, known as nodes, that are networked together to perform parallel computations on big data sets. The Name node is the master node of the Hadoop Distributed File System (HDFS). It maintains the meta data of the files in the RAM for quick access. An actual Hadoop Cluster setup involves extensive resources which are not within the scope of this lab. In this lab, you will use dockerized hadoop to create a Hadoop Cluster which will have:

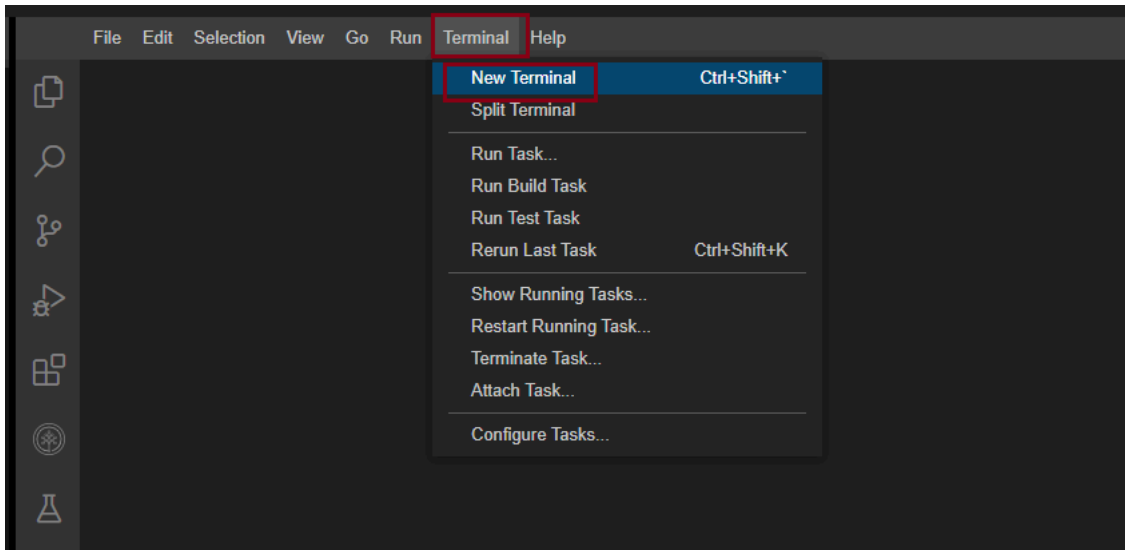
1. Namenode
2. Datanode
3. Node Manager
4. Resource manager
5. Hadoop history server

Objectives

- Run a dockerized Cluster Hadoop instance
- Create a file in the HDFS and view it on the GUI

Set up Cluster Nodes Dockerized Hadoop

1. Start a new terminal



2. Clone the repository to your theia environment.

```
git clone https://github.com/ibm-developer-skills-network/ooxwv-docker_hadoop.git
```

3. Navigate to the docker-hadoop directory to build it.

```
cd ooxwv-docker_hadoop
```

4. Compose the docker application.

```
docker-compose up -d
```

Compose is a tool for defining and running multi-container Docker applications. It uses the YAML file to configure the services and enables us to create and start all the services from just one configuration file.

You will see that all the five containers are created and started.

```
# 3192219afd04 Pull complete
# aa53513fe997 Pull complete
# b0d764123f3e Pull complete
# b04394ddb35d Pull complete
[+] Running 9/9
# Network ooxwv-docker_hadoop_default Created
# Volume "ooxwv-docker_hadoop_hadoop_historyserver" Created
# Volume "ooxwv-docker_hadoop_hadoop_namenode" Created
# Volume "ooxwv-docker_hadoop_hadoop_datanode" Created
# Container nodemanager Started
# Container datanode Started
# Container historyserver Started
# Container namenode Started
# Container resourcemanager Started
```

5. Run the namenode as a mounted drive on bash.

```
docker exec -it namenode /bin/bash
```

6. You will observe that the prompt changes as shown below.

```
theia@theiadocker-lavanyas:/home/project/docker-hadoop$ docker exec
root@d72225e7724e:/#
```

Explore the hadoop environment

As you have learnt in the videos and reading thus far in the course, a Hadoop environment is configured by editing a set of configuration files:

- **hadoop-env.sh** Serves as a master file to configure YARN, HDFS, MapReduce, and Hadoop-related project settings.
- **core-site.xml** Defines HDFS and Hadoop core properties
- **hdfs-site.xml** Governs the location for storing node metadata, fsimage file and log file.
- **mapred-site.xml** Lists the parameters for MapReduce configuration.
- **yarn-site.xml** Defines settings relevant to YARN. It contains configurations for the Node Manager, Resource Manager, Containers, and Application Master.

For the docker image, these xml files have been configured already. You can see these in the directory **/opt/hadoop-3.2.1/etc/hadoop/** by running

```
ls /opt/hadoop-3.2.1/etc/hadoop/*.xml
```

Create a file in the HDFS

1. In the HDFS, create a directory structure named **user/root/input**.

```
hdfs dfs -mkdir -p /user/root/input
```

2. Copy all the hadoop configuration xml files into the input directory.

```
hdfs dfs -put $HADOOP_HOME/etc/hadoop/*.xml /user/root/input
```

3. Create a **data.txt** file in the current directory.

```
curl https://raw.githubusercontent.com/ibm-developer-skills-network/ooxwv-docker_hadoop/master/SampleMapReduce.txt --output data.txt
```

4. Copy the **data.txt** file into **/user/root**.

```
hdfs dfs -put data.txt /user/root/
```

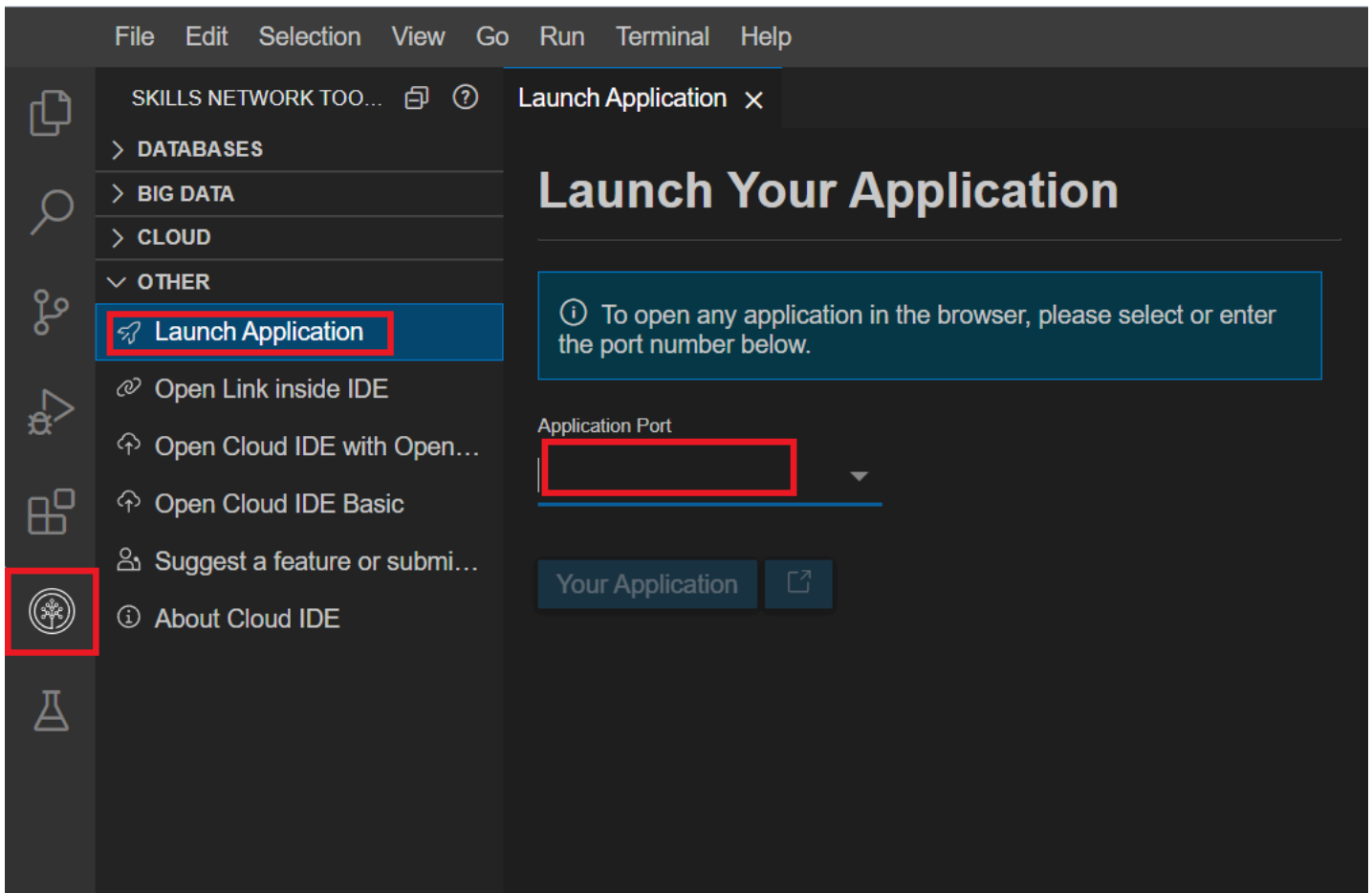
5. Check if the file has been copied into the HDFS by viewing its content.

```
hdfs dfs -cat /user/root/data.txt
```

View the HDFS

1. Click the button below or click on the Skills Network button on the left, it will open the "Skills Network Toolbox". Then click the Other then Launch Application. From there you should be able to enter the port number as **9870** and launch.

View HDFS



2. This will open up the Graphical User Interface (GUI) of the Hadoop node. Click on utilities -> Browse the file system to browse the files.

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Overview 'namenode:9000' (active)

Started:	Mon Jul 12 15:11:20 +0530 2021
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0
Compiled:	Tue Sep 10 21:26:00 +0530 2019 by rohithsharm
Cluster ID:	CID-0dba2137-1551-44b7-8ab3-49a6661cdaf7
Block Pool ID:	BP-936334794-172.18.0.2-1626082572639

3. View the files in the directories that you have just created by clicking on user then root.

Browse Directory

/

Show

25

entries

<input type="checkbox"/>	<div><div></div>Permission</div>	<div><div></div>Owner</div>	<div><div></div>Group</div>	<div><div></div>Size</div>	<div><div></div>Last Modified</div>	<div><div></div>Replicatio</div>
<input type="checkbox"/>	<div><div>drwxr-xr-x</div></div>	<div><div>root</div></div>	<div><div>supergroup</div></div>	<div><div>0 B</div></div>	<div><div>Nov 10 16:00</div></div>	<div><div>0</div></div>
<input type="checkbox"/>	<div><div>drwxr-xr-x</div></div>	<div><div>root</div></div>	<div><div>supergroup</div></div>	<div><div>0 B</div></div>	<div><div>Nov 10 16:01</div></div>	<div><div>0</div></div>

Showing 1 to 2 of 2 entries

Hadoop, 2019.

Browse Directory

Show entries

<input type="checkbox"/>	↓↑ Permission	↓↑ Owner	↓↑ Group	↓↑ Size	↓↑ Last Modified	↓↑ Replication
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Nov 10 16:01	0

Showing 1 to 1 of 1 entries

[Hadoop](#)[Overview](#)[Datanodes](#)[Datanode Volume Failures](#)[Snapshot](#)

Browse Directory

Show entries

<input type="checkbox"/>	↓↑ Permission	↓↑ Owner	↓↑ Group	↓↑ Size	↓↑ L
<input type="checkbox"/>	-rw-r--r--	root	supergroup	6.7 KB	J
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	J

Showing 1 to 2 of 2 entries

4. Notice that the block size is 128 MB though the file size is actually much smaller. This is because the default block size used by HDFS is 128 MB.

5. You can click on the file to check the file into. It gives you information about the file in terms of number of bytes, block id etc.,

File information - data.txt

Download

Head the file (first 32K)

Tail the file (last 32K)

Block information --

Block 0

Block ID: 1073741839

Block Pool ID: BP-1800570971-172.18.0.5-1642502538329

Generation Stamp: 1015

Size: 6858


Availability:

- 1bb0a610767b

Close

Congratulations! You have:

- Deployed Hadoop using Docker
- Created data in HDFS and viewed it on the GUI

 Tweet and share your achievement!

Author(s)

Lavanya T S

© IBM Corporation. All rights reserved.