

INDEX

Prerequisites	2
1. Port Opening for internal cluster communication	2
2. Before Ansible Deployment	2
Ansible Deployment Steps	6
Prerequisites to run hadoop via ansible	6
Setting up ansible configuration	7
Verify the hadoop services are running by accessing the WebUI	13
Connecting Hadoop Service Remotely	17
Submitting spark job remotely	19
Connecting Hive services remotely	20
Run Hbase query remotely	21

Hadoop Ecosystem

Prerequisites

1. Port Opening for internal cluster communication

?????

2. Before Ansible Deployment

- Download the hadoop ecosystem binaries by using the below links:-
 - Hadoop binary:-
<https://archive.apache.org/dist/hadoop/common/hadoop-2.10.2/hadoop-2.10.2.tar.gz>
 - Zookeeper binary:-
<https://archive.apache.org/dist/zookeeper/zookeeper-3.4.6/zookeeper-3.4.6.tar.gz>
 - Hive binary:-
<https://downloads.apache.org/hive/hive-2.3.9/apache-hive-2.3.9-bin.tar.gz>
 - Hbase_binary:-
<https://archive.apache.org/dist/hbase/2.4.15/hbase-2.4.15-bin.tar.gz>
 - Spark_binary:-
<https://archive.apache.org/dist/spark/spark-3.2.2/spark-3.2.2-bin-hadoop2.7.tgz>
 - Flink_binary:-
https://archive.apache.org/dist/flink/flink-1.14.4/flink-1.14.4-bin-scala_2.11.tgz
 - Presto_binary:-
<https://repo1.maven.org/maven2/com/facebook/presto/presto-server/0.278.1/presto-server-0.278.1.tar.gz>

- Upload the above binaries onto ICS Bucket so that ansible will pull binaries from ICS bucket.
- Download s3cmd on server by using the below commands

```
sudo wget
https://github.com/s3tools/s3cmd/releases/download/v2.3.0/s3cmd-2.3.0.tar.gz
sudo wget https://bootstrap.pypa.io/pip/2.7/get-pip.py
```

```
python2 get-pip.py
```

- Check the pip version and the expected output should be the as below

```
[nhadmin@deltextesthm1 ~]$ pip -V
pip 20.3.4 from /usr/lib/python2.7/site-packages/pip (python 2.7)
[nhadmin@deltextesthm1 ~]$
```

- Install python-magic with this command

```
pip install --user python-magic
```

```
details about python 2 support in pip can be found at https://p
Collecting python-magic
  Downloading python_magic-0.4.27-py2.py3-none-any.whl (13 kB)
Installing collected packages: python-magic
Successfully installed python-magic-0.4.27
```

- Install s3cmd

```
tar -xvf s3cmd-2.3.0.tar.gz
cd s3cmd-2.3.0
sudo python ./setup.py install
s3cmd --version
```

```
[nhadmin@deltextesthm1 ~]$ s3cmd --version
s3cmd version 2.3.0
[nhadmin@deltextesthm1 ~]$
```

- Add configurations inside .bashrc

```
export AWS_ACCESS_KEY_ID=access_key
export AWS_SECRET_ACCESS_KEY=secret_key
export AWS_HOST=host_url
export AWS_ENDPOINT=endpoint_url
```

- After this run the following command on the terminal

```
source .bashrc
```

- In order to create bucket run the below command on the node you have installed s3cmd

```
s3cmd mb --ssl --host=${AWS_HOST} --host-bucket= s3://bucket-name
```

```
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$
[nhadmin@deltetestm1 ~]$ s3cmd mb --ssl --host=${AWS_HOST} --host-bucket= s3://demo
Bucket 's3://demo/' created
[nhadmin@deltetestm1 ~]$
```

- In order to insert in the bucket run the below command

```
s3cmd put filename --host=${AWS_HOST} --host-bucket= ls s3://<bucket-name>
```

- In order to list the contents of the bucket run the below command

Ansible Deployment Steps

Prerequisites to run hadoop via ansible

- We have deployed the hadoop ecosystem on RHEL version 7.9(maipo).

```
cat /etc/redhat-release
```

```
[nhadmin@deltextrsthm1 ~]$ cat /etc/redhat-release  
Red Hat Enterprise Linux Server release 7.9 (Maipo)
```

- Ensure that you have python3 and ansible installed on your local system

```
python3 --version
```

```
dhru@dhru-Inspiron-3576:~$ python3 --version
Python 3.10.6
dhru@dhru-Inspiron-3576:~$
```

- To check the ansible version run the following command:

```
ansible --version
```

```
dhru@dhru-Inspiron-3576:~$ ansible --version
ansible [core 2.13.7]
  config file = /etc/ansible/ansible.cfg
  configured module search path = ['/home/dhru/.ansible/plugins/modules', '/usr/share/ansible/plugins/modules']
  ansible python module location = /usr/lib/python3/dist-packages/ansible
  ansible collection location = /home/dhru/.ansible/collections:/usr/share/ansible/collections
  executable location = /usr/bin/ansible
  python version = 3.10.6 (main, Nov 14 2022, 16:10:14) [GCC 11.3.0]
  jinja version = 3.0.3
  libyaml = True
dhru@dhru-Inspiron-3576:~$
```

- If you have not downloaded ansible and python, follow the below link to install
 - For python3: <https://docs.python-guide.org/starting/install3/linux>
 - For ansible: https://docs.ansible.com/ansible/latest/installation_guide/installation_distros.html

Setting up ansible configuration

- Inventory file in ansible will look like this :

```
≡ inventory X
≡ inventory
1
2 [masters]
3 primary zk_myid=1 ansible_host=100.96.20.168 ansible_user=nhadmin
4 secondary zk_myid=2 ansible_host=100.96.20.169 ansible_user=nhadmin
5 [datanode]
6 datanode1 zk_myid=3 ansible_host=100.96.20.170 ansible_user=nhadmin
7 [tasknode]
8 tasknode1 zk_myid=4 ansible_host=100.96.20.171 ansible_user=nhadmin
9 tasknode2 zk_myid=5 ansible_host=100.96.20.172 ansible_user=nhadmin
10
```

As you can see from the above example, you can change **ansible_host** according to the ip address you have available. Similarly if the user is different change the **ansible_user**

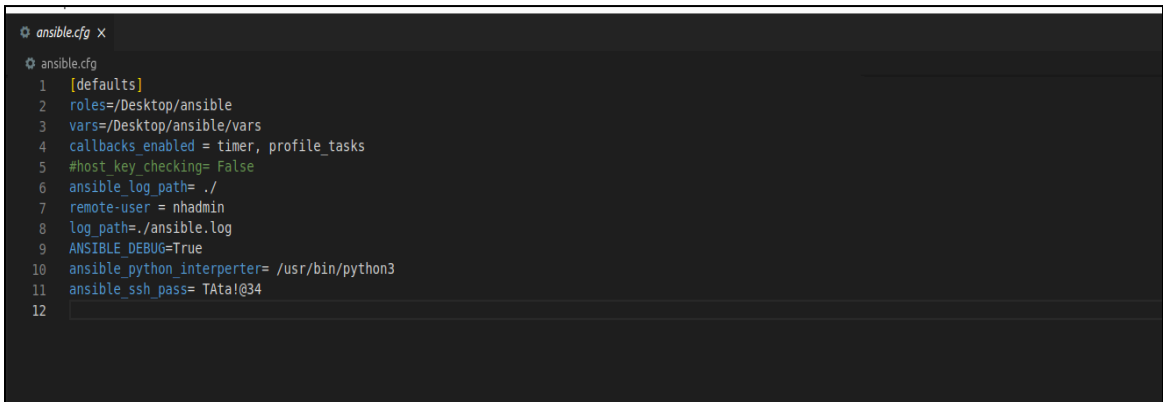
- Setting up Ansible.cfg file:

Give the roles path under "roles" section of the code line

Give path for the global variable under vars section folder

Give ansible_python_interpreter path present in your system

Give the path for the vault file under global vars section in playbook.yml

A screenshot of a code editor showing the contents of an Ansible configuration file named 'ansible.cfg'. The file is opened in a tab labeled 'ansible.cfg x'. The code is as follows:

```
1 [defaults]
2 roles=/Desktop/ansible
3 vars=/Desktop/ansible/vars
4 callbacks_enabled = timer, profile_tasks
5 #host key checking= False
6 ansible_log_path= ./
7 remote-user = nhadmin
8 log_path=./ansible.log
9 ANSIBLE_DEBUG=True
10 ansible_python_interpreter= /usr/bin/python3
11 ansible_ssh_pass= TAta!@34
12
```

- For global variables file stored under vars folder

Copy the ip of all the instances and paste it in the "vars/external_vars.yml" present in the
For eg:-

```
primary: hadoop_primary_ip
secondary: hadoop_secondary_ip
datanode1: hadoop_datanode1_ip
tasknode1: hadoop_tasknode1_ip
tasknode2: hadoop_tasknode2_ip
```



```

vars > ! external_vars.yml
1  # Master and worker node IP's
2
3  #private ip for all nodes
4  primary: 100.96.20.168
5  secondary: 100.96.20.169
6  datanode1: 100.96.20.170
7  tasknode1: 100.96.20.171
8  tasknode2: 100.96.20.172
9
10 ha_zookeeper_quorum: primary:2181,secondary:2181
11
12 #For presto node.properties
13 node_id_master: ip-172-31-68-112-primary
14 node_id_slave1: ip-172-31-66-56-tasknode1
15 node_id_slave2: ip-172-31-77-110-tasknode2
16 node_environment: production
17
18 enable_debug: false
19 jps_debug: true
20 status_debug: true
21 yarn_nodelist_debug: false
22
23 s3_bucket_name: hadoop
24
25 #Linux
26 home_path: /home/nhadmin/
27 java_path: /usr
28 sleep_time: 10
29 user: nhadmin

```

- Change the “**home path**” according to the environment you are working on
- Change the “**user**” according to the environment you are working on
- For rhel linux:-
 - Change the “**s3 bucket name**” by the bucket name you created under "s3_bucket_name" section
 - Change the node_id properties for all the three with the private ip except user "-" instead of "."

(Note: Ensure that you know which is the master node and which are the slave nodes)

- Playbook.yml

```

1  ! playbook.yml
2
3  ---
4  - hosts: all
5    vars_files:
6      - ./vars/external_vars.yml
7      - ./aws_creds.yml
8    roles:
9      - hadoop
10     - zookeeper
11     - hive
12     - mysql
13     - hbase
14     - spark
15     - flink
16     - presto
17     - activate-service
18
19
20

```

- We have given "hosts: all" as there are 3 groups present in the inventory
- Provided global variable path under "vars_files:" section
- We provide all the roles we have used in the script under "roles" section
- Now according to your requirement you can choose what components you want in your hadoop ecosystem.
 - For example:-If you want to hadoop, hbase in your system run the following roles in the playbook.yml

```

1  ! playbook.yml
2
3  ---
4  - hosts: all
5    vars_files:
6      - ./vars/external_vars.yml
7      - ./aws_creds.yml
8    roles:
9      - hadoop
10     - zookeeper
11     # - hive
12     # - mysql
13     - hbase
14     # - spark
15     # - flink
16     # - presto
17     - activate-service
18
19
20

```

- Execute ansible ping command to ensure all the servers are reachable through ansible

```
ansible all -i inventory -m ping -v --ask-pass
```

(In password prompt you need to enter ssh password for nodes)

```

● dhru@dhru-Inspiron-3576:~/Desktop/ansible$ ansible all -i inventory -m ping -v --ask-pass
Using /home/dhru/Desktop/ansible/ansible.cfg as config file
SSH password:
datanode1 | SUCCESS => {
  "ansible_facts": {
    "discovered_interpreter_python": "/usr/bin/python"
  },
  "changed": false,
  "ping": "pong"
}
secondary | SUCCESS => {
  "ansible_facts": {
    "discovered_interpreter_python": "/usr/bin/python"
  },
  "changed": false,
  "ping": "pong"
}
tasknode1 | SUCCESS => {
  "ansible_facts": {
    "discovered_interpreter_python": "/usr/bin/python"
  },
  "changed": false,
  "ping": "pong"
}
tasknode2 | SUCCESS => {
  "ansible_facts": {
    "discovered_interpreter_python": "/usr/bin/python"
  },
  "changed": false,
  "ping": "pong"
}
primary | SUCCESS => {
  "ansible_facts": {
    "discovered_interpreter_python": "/usr/bin/python"
  },
  "changed": false,
  "ping": "pong"
}
● dhru@dhru-Inspiron-3576:~/Desktop/ansible$

```

- Command to start playbook:

```
ansible-playbook -i inventory playbook.yml --ask-vault-pass
```

```

● dhru@dhru-Inspiron-3576:~/Desktop/ansible$ ansible-playbook -i inventory playbook.yml --ask-vault-pass
Vault password:

PLAY [all] *****

TASK [Gathering Facts] *****
Tuesday 24 January 2023  18:02:18 +0530 (0:00:00.131)    0:00:00.131 *****
ok: [datanode1]
ok: [tasknode1]
ok: [primary]
ok: [secondary]
ok: [tasknode2]

TASK [hadoop : Install OpenJDK Java] *****
Tuesday 24 January 2023  18:02:21 +0530 (0:00:03.376)    0:00:03.508 *****

```

- Ensuring all the services are up and running in respective nodes by executing jps command.

```

[nhadmin@delttextesthm1 ~]$ jps
9393 DFSZKFailoverController
67107 QuorumPeerMain
123843 ResourceManager
99891 NameNode
8870 JournalNode
68699 HMaster
61069 StandaloneSessionClusterEntrypoint
30749 RunJar
10990 PrestoServer
11023 Jps
[nhadmin@delttextesthm1 ~]$

```

- After playbook run is successful check the below ips to get the respective service UI:

Service	Hostname	Ports
Namenode	primary	50070
Namenode	secondary	50070
Yarn	primary	8088
Yarn	secondary	8088
Flink	primary	8081
Presto	primary	8080

Verify the hadoop services are running by accessing the WebUI

- Check the UI of namenode primary by typing below URL in browser

hadoop_primary_ip:50070

← → ↻ 🔍 Not secure | 100.96.20.168:50070/dfshealth.html#tab-overview

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Overview 'primary:8020' (active)

Namespace:	ha-cluster
Namenode ID:	nn1
Started:	Mon Jan 23 11:54:25 +0530 2023
Version:	2.10.2, r965d380006fa78b2315668fbc7eb432e1d8200f
Compiled:	Wed May 25 04:05:00 +0530 2022 by ubuntu from branch-2.10.2
Cluster ID:	CID-2e6b060a-d5eb-409f-9b67-66ba01eb803d
Block Pool ID:	BP-535755290-100.96.20.168-1674150489958

Summary

Security is off.

Safe mode is ON. The reported blocks 361 has reached the threshold 0.9990 of total blocks 361. The minimum number of live datanodes is not required. In safe mode extension. Safe mode will be turned off automatically in 11 seconds.

434 files and directories, 364 blocks = 798 total filesystem object(s).

Heap Memory used 93.5 MB of 498 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 50.69 MB of 52 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	36.29 GB
DFS Used:	3.49 MB (0.01%)
Non DFS Used:	6.15 GB
DFS Remaining:	27.79 GB (76.58%)
Block Pool Used:	3.49 MB (0.01%)
DataNodes usages% (Min/Median/Max/stdDev):	0.01% / 0.01% / 0.01% / 0.00%

← → ↻ 🔍 Not secure | 100.96.20.168:50070/dfshealth.html#tab-datanode

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Datanode Information

✓ In service ● Down ⚠ Decommissioned ⚙ Decommissioned & dead ⚡ In Maintenance & dead

Datanode usage histogram

In operation

Show 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓ datanode150010 (100.96.20.170:50010)	http://datanode1:50075	1s	6m	36.29 GB	555	5.36 MB (0.01%)	2.10.2

Showing 1 to 1 of 1 entries

Previous 1 Next

Entering Maintenance

- Check the UI of namenode secondary by typing below URL in browser

hadoop_secondary_ip:50070

← → ↻ ⚙ ⚠ Not secure | 100.96.20.169:50070/dfshealth.html#tab-overview

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Overview 'secondary:8020' (standby)

Namespace:	ha-cluster
Namenode ID:	nn2
Started:	Mon Jan 23 11:54:38 +0530 2023
Version:	2.10.2, r965fd380006fa78b2315668fbc7eb432e1d8200f
Compiled:	Wed May 25 04:05:00 +0530 2022 by ubuntu from branch-2.10.2
Cluster ID:	CID-2e6b060a-dde5-409f-9b67-66ba01eb802d
Block Pool ID:	BP-535755290-100.96.20.168-1674150489958

Summary

Security is off.

Safe mode is ON. The reported blocks 361 has reached the threshold 0.9990 of total blocks 361. The minimum number of live datanodes is not required. In safe mode extension, Safe mode will be turned off automatically in 22 seconds.

434 files and directories, 364 blocks = 798 total filesystem object(s).

Heap Memory used 51.41 MB of 508.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 48.84 MB of 49.5 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	36.29 GB
DFS Used:	3.49 MB (0.01%)
Non DFS Used:	6.15 GB
DFS Remaining:	27.79 GB (76.58%)
Block Pool Used:	3.49 MB (0.01%)
DataNodes usages% (Min/Median/Max/stdDev):	0.01% / 0.01% / 0.01% / 0.00%

← → ↻ ⚙ ⚠ Not secure | 100.96.20.169:50070/dfshealth.html#tab-datanode

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Datanode Information

✓ In service ● Down ⚠ Decommissioned ⚙ Decommissioned & dead ⚡ In Maintenance & dead

Datanode usage histogram

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓ datanode1.50010 (100.96.20.178:50010)	http://datanode1.50075	0s	164m	36.29 GB	555	5.18 MB (0.02%)	2.10.2

Showing 1 to 1 of 1 entries


Previous 1 Next

Entering Maintenance

- Check the UI of yarn resource manager either on primary or secondary by typing below URL in browser

hadoop_primary_ip:8088 or hadoop_secondary_ip:8088

← → ↻ ⚙ Not secure | 100.96.20.169:8088/cluster



All Applications

- Cluster
- About
- Nodes
- Node Labels
- Applications
- NEW
- NEW SAVING
- SUBMITTED
- ACCEPTED
- RUNNING
- FINISHED
- FAILED
- KILLED
- Scheduler
- Tools

Cluster Metrics

Apps Submitted	0	Apps Pending	0	Apps Running	4	Apps Completed	0	Containers Running		Used Resources	<memory:0 B, vCores:0>	Total Resources	<memory:24 GB, vCores:24>	Reserved Resources	<memory:0 B, vCores:0>
----------------	---	--------------	---	--------------	---	----------------	---	--------------------	--	----------------	------------------------	-----------------	---------------------------	--------------------	------------------------

Cluster Nodes Metrics

Active Nodes	0	Decommissioning Nodes	0	Decommissioned Nodes	0	Lost Nodes	0	Unhealthy Nodes	0
--------------	---	-----------------------	---	----------------------	---	------------	---	-----------------	---

Scheduler Metrics

Scheduler Type	Capacity Scheduler	Scheduling Resource Type	[<name=memory-mb default-unit=M type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>]	Minimum Allocation	<memory:1024, vCores:1>	Maximum Allocation	<memory:8192, vCores:1>
----------------	--------------------	--------------------------	--	--------------------	-------------------------	--------------------	-------------------------

Show 20 ▼ entries

ID	User	Name	Application Type	Queue	Application Priority	Start Time	Launch Time	Finish Time	State	Final Status	Running Containers	Allocated CPU V-Cores	Allocated Memory MB	Allocated GPUs	Reserved CPU V-Cores	Reserved Memory MB
application_1674152303270_0004	rhadmin	org.apache.spark.examples.SparkPi	SPARK	default	0	Fri Jan 20 00:17:46 +0550 2023	Fri Jan 20 00:17:56 +0550 2023	Fri Jan 20 00:17:56 +0550 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A
application_1674152303270_0003	rhadmin	org.apache.spark.examples.SparkPi	SPARK	default	0	Fri Jan 20 00:17:44 +0550 2023	Fri Jan 20 00:17:55 +0550 2023	Fri Jan 20 00:17:55 +0550 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A
application_1674152303270_0002	rhadmin	org.apache.spark.examples.SparkPi	SPARK	default	0	Fri Jan 20 00:17:42 +0550 2023	Fri Jan 20 00:17:55 +0550 2023	Fri Jan 20 00:17:55 +0550 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A
application_1674152303270_0001	rhadmin	org.apache.spark.examples.SparkPi	SPARK	default	0	Fri Jan 20 00:03:37 +0550 2023	Fri Jan 20 00:03:48 +0550 2023	Fri Jan 20 00:03:48 +0550 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A

Showing 1 to 4 of 4 entries

- Check the UI of flink on primary by typing below URL in browser

hadoop_primary_ip:8081

The screenshot shows the Apache Flink Dashboard Overview page. The left sidebar contains navigation links: Overview (selected), Jobs, Running Jobs, Completed Jobs, Task Managers, Job Manager, and Submit New Job. The main content area displays the following information:

- Available Task Slots:** 0
- Total Task Slots:** 0
- Task Managers:** 0
- Running Jobs:** 1
 - Finished: 0
 - Canceled: 0
 - Failed: 0
- Running Job List:**

Job Name	Start Time	Duration	End Time	Tasks	Status
State machine job	2023-01-23 11:57:17	27s	-	2	RUNNING
- Completed Job List:** No Data

The top right corner of the dashboard shows the version (1.14.4), commit (895c609 @ 2022-02-25T11:57:14+01:00), and a message icon.

- Check the UI of Presto on primary by typing below URL in browser

hadoop_primary_ip:8080

The screenshot shows the Presto Cluster Overview page. The top right corner displays the version (0.278.1-EC67BA1), environment (PRODUCTION), and uptime (4.71m). The main content area is divided into several sections:

- CLUSTER OVERVIEW:**
 - RUNNING QUERIES:** 0
 - QUEUED QUERIES:** 0
 - BLOCKED QUERIES:** 0
 - ACTIVE WORKERS:** 2
 - RUNNABLE DRIVERS:** 0.00
 - RESERVED MEMORY (B):** 0
 - ROWS/SEC:** 0.00
 - BYTES/SEC:** 0
 - WORKER PARALLELISM:** 0.00
- QUERY DETAILS:**
 - State, source, query ID, resource group, or query text:
 - State: ☒ Running ☒ Queued ☐ Finished ☐ Failed
 - Sort:
 - Refresh Interval:
 - Show:

The bottom of the query details section displays "No queries".

Connecting Hadoop Service Remotely

Prerequisites:

- Ensure that you have java installed in your local system

```
dhru@dhru-Inspiron-3576:~$ java -version
openjdk version "11.0.17" 2022-10-18
OpenJDK Runtime Environment (build 11.0.17+8-post-Ubuntu-1ubuntu222.04)
OpenJDK 64-Bit Server VM (build 11.0.17+8-post-Ubuntu-1ubuntu222.04, mixed mode, sharing)
```

- Download the configured binary files for all components and unzip the binary files
- Copy the configured binary files of all hadoop components on to your local system
- Give the environment path of all the components binaries in .bashrc

```
export JAVA_HOME="/usr"
export HADOOP_HOME="/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/hadoop-jan19/hadoop/hadoop-2.10.2"
export SPARK_HOME="/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/spark-jan19/spark/spark-3.2.2"
export HIVE_HOME="/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/hive-jan19/hive"
export PRESTO_HOME="/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/presto-server-0.278.1-jan19/presto-server-0.278.1"
export HBASE_HOME="/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/hbase-jan19/hbase/hbase-2.4.15"
export FLINK_HOME="/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/flink-jan19/flink/flink-1.14.4"
export ZOOKEEPER_HOME="/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/zookeeper-jan19/zookeeper/zookeeper-3.4.6"
export HADOOP_INSTALL=SHADOOP_HOME
export HADOOP_MAPRED_HOME=SHADOOP_HOME
export HADOOP_HDFS_HOME=SHADOOP_HOME
export YARN_HOME=SHADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=SHADOOP_HOME/lib/native
export HADOOP_CONF_DIR=SHADOOP_HOME/etc/hadoop
export YARN_CONF_DIR=SHADOOP_HOME/etc/hadoop
export HADOOP_OPTS="-Djava.library.path=SHADOOP_HOME/lib/native"
export PATH=$PATH:SHADOOP_HOME/sbin:SHADOOP_HOME/bin:$SPARK_HOME/bin:$HIVE_HOME/bin:$HBASE_HOME/bin:$FLINK_HOME/bin:$PRESTO_HOME/bin:$JAVA_HOME/bin:$ZOOKEEPER_HOME/bin
```

- Run the following command after updating .bashrc

```
source .bashrc
```

- Make a host entry for Hadoop servers inside /etc/hosts

```
sudo vi /etc/hosts
```

```
dhru@dhru-Inspiron-3576:~$ cat /etc/hosts
127.0.0.1    localhost
127.0.1.1    dhru-Inspiron-3576
# The following lines are desirable for IPv6 capable hosts
::1        ip6-localhost ip6-loopback
fe00::0    ip6-localnet
ff00::0    ip6-mcastprefix
ff02::1    ip6-allnodes
ff02::2    ip6-allrouters

100.96.20.168 primary
100.96.20.169 secondary
100.96.20.170 datanode1
100.96.20.171 tasknode1
100.96.20.172 tasknode2
```

- Change the java path in hbase-env.sh file inside /conf folder in Hbase directory

```
# Override text processing tools for use by these launch scripts.
# export GREP="${GREP-grep}"
# export SED="${SED-sed}"
export JAVA_HOME=/usr
export HBASE_PID_DIR=/home/nhadmin/HA/hbase/pids
export HBASE_MANAGES_ZK=false
```

Submitting spark job remotely

- Navigate to spark binary folder present on local machine and execute below command

```
./bin/spark-submit --class org.apache.spark.examples.SparkPi --master yarn
--deploy-mode cluster --conf spark.yarn.am.nodeLabelExpression=TASK --conf
spark.yarn.executor.nodeLabelExpression=TASK
./examples/jars/spark-examples_2.12-3.2.2.jar
```

```
dhru@dhru-Inspiron-3576:~/Desktop/hadoop_configs/OneDrive_1_25-01-2023/spark-jan19/spark/spark-3.2.2_$ ./bin/spark-submit --class org.apache.spark.examples.SparkPi --master yarn --deploy-mode cluster --c
onf spark.yarn.am.nodeLabelExpression=TASK --conf spark.yarn.executor.nodeLabelExpression=TASK ./examples/jars/spark-examples_2.12-3.2.2.jar
23/01/25 13:00:15 WARN Utils: Your hostname, dhru-Inspiron-3576 resolves to a loopback address: 127.0.1.1; using 192.168.29.109 instead (on interface wlp3s0)
23/01/25 13:00:15 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/spark-jan19/spark/spark-3.2.2/jars/spark-unsafe_2.12-3.2.2.jar) to con
structor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
23/01/25 13:00:15 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/01/25 13:00:16 INFO ConfiguredRMFAutoProxyProvider: Falling over to rm2
23/01/25 13:00:16 INFO Client: Requesting a new application from cluster with 3 NodeManagers
23/01/25 13:00:16 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (8192 MB per container)
23/01/25 13:00:16 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
23/01/25 13:00:16 INFO Client: Setting up container launch context for our AM
23/01/25 13:00:16 INFO Client: Setting up the launch environment for our AM container
23/01/25 13:00:16 INFO Client: Preparing resources for our AM container
23/01/25 13:00:16 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
23/01/25 13:00:19 INFO Client: Uploading resource file:/tmp/spark-23749c4c-7a24-4f35-b77d-e10dc0eb6ea1/_spark_libs_14644048794144907593.zip -> hdfs://ha-cluster/user/dhru/.sparkStaging/application_1674
152303270_0006/_spark_libs_14644048794144907593.zip
23/01/25 13:01:48 INFO Client: Uploading resource file:/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/spark-jan19/spark/spark-3.2.2/examples/jars/spark-examples_2.12-3.2.2.jar -> hdfs://ha-clust
er/user/dhru/.sparkStaging/application_1674152303270_0006/spark-examples_2.12-3.2.2.jar
23/01/25 13:01:49 INFO Client: Uploading resource file:/tmp/spark-23749c4c-7a24-4f35-b77d-e10dc0eb6ea1/_spark_conf_2049723406510325432.zip -> hdfs://ha-cluster/user/dhru/.sparkStaging/application_16741
52303270_0006/_spark_conf_2049723406510325432.zip
23/01/25 13:02:02 INFO Client: Uploading resource file:/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/spark-jan19/spark/spark-3.2.2/examples/jars/spark-examples_2.12-3.2.2.jar -> hdfs://ha-clust
er/user/dhru/.sparkStaging/application_1674152303270_0006/spark-examples_2.12-3.2.2.jar
```

```
23/01/25 13:02:10 INFO client: Application report for application_1674152303270_0006 (state: RUNNING)
23/01/25 13:02:11 INFO client: Application report for application_1674152303270_0006 (state: RUNNING)
23/01/25 13:02:12 INFO client: Application report for application_1674152303270_0006 (state: RUNNING)
23/01/25 13:02:13 INFO client: Application report for application_1674152303270_0006 (state: RUNNING)
23/01/25 13:02:14 INFO client: Application report for application_1674152303270_0006 (state: FINISHED)
23/01/25 13:02:14 INFO client:
client token: N/A
diagnostics: N/A
ApplicationMaster host: tasknode2
ApplicationMaster RPC port: 39429
queue: default
start time: 1674631923131
final status: SUCCEEDED
tracking URL: http://secondary:8088/proxy/application_1674152303270_0006/
user: dhru
23/01/25 13:02:14 INFO Client: Deleted staging directory hdfs://ha-cluster/user/dhru/.sparkStaging/application_1674152303270_0006
23/01/25 13:02:14 INFO ShutdownHookManager: Shutdown hook called
23/01/25 13:02:14 INFO ShutdownHookManager: Deleting directory /tmp/spark-8dce3829-18d1-4446-9474-519773753878
23/01/25 13:02:14 INFO ShutdownHookManager: Deleting directory /tmp/spark-23749c4c-7a24-4f35-b77d-e10dc0eb6ea1
```

Connecting Hive services remotely

- Run the following command either on home terminal or spark binary folder:-

```
beeline
```

```
dhru@dhru-Inspiron-3576:~$ beeline
Beeline version 2.3.9 by Apache Hive
beeline>
```

- Run the following command on the beeline terminal:-

```
!connect jdbc:hive2://hadoop_primary_ip:10000 root
```

- Here “root” is the username that you have given during mysql setup
- It will prompt for a password,type the password that you have given during mysql setup

```
dhru@dhru-Inspiron-3576:~/Desktop/hadoop_configs/OneDrive_1_25-01-2023/spark-jen19/spark/spark-3.2.2/bin$ beeline
Beeline version 2.3.9 by Apache Hive
beeline> !connect jdbc:hive2://100.96.20.168:10000 root
Connecting to jdbc:hive2://100.96.20.168:10000
Enter password for jdbc:hive2://100.96.20.168:10000: *****
23/01/25 13:17:15 INFO Utils: Supplied authorities: 100.96.20.168:10000
23/01/25 13:17:15 INFO Utils: Resolved authority: 100.96.20.168:10000
Connected to: Apache Hive (version 2.3.9)
Driver: Hive JDBC (version 2.3.9)
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://100.96.20.168:10000> show databases;
+-----+
| database_name |
+-----+
| default       |
+-----+
1 row selected (1.233 seconds)
0: jdbc:hive2://100.96.20.168:10000>
```

Run Hbase query remotely

Run the following command in home terminal to start hbase shell:-

```
hbase shell
```

```
dhru@dhru-Inspiron-3576:~$ hbase shell
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.hbase.unsafe.HBasePlatformDependent (file:/home/dhru/Desktop/hadoop_configs/OneDrive_1_25-01-2023/hbase-jan19/hbase/hbase-2.4.15/lib/hbase-unsafe-4.1.2.jar) to method java.nio.Bits.unaligned()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.hbase.unsafe.HBasePlatformDependent
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
2023-01-25 15:11:57.495 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.4.15, r35310fcd6b11a1d04d75eb7db2e592dd34e4d5b6, Thu Oct 13 11:42:20 PDT 2022
Took 0.0018 seconds
hbase:001:0>
```

Run the following command to create a table :

```
create 'dhруп', 'personal data', 'professional data'
```

```
Took 0.0018 seconds
hbase:001:0> create 'dhруп', 'personal data', 'professional data'
Created table dhруп
Took 1.1376 seconds
=> Hbase::Table - dhруп
hbase:002:0>
```

Run the following command to put the contents in the table:

```
put 'dhруп','1','personal data:name','raju'
```

```
hbase:002:0> put 'dhруп','1','personal data:name','raju'
Took 0.4712 seconds
hbase:003:0>
```

Run the following command to list the contents of the table:

```
scan 'dhруп'
```

```
hbase:003:0> scan 'dhrup'
ROW                                COLUMN+CELL
1                                  column=personal data:name, timestamp=2023-01-25T15:16:37.688, value=raju
1 row(s)
Took 0.0993 seconds
hbase:004:0> █
```