

Automated Setup & Installation Guide for Hadoop Single Node Cluster Environment (Pseudo Distributed mode)

**Developed & Tested
by
RAJU CHAL
LKM , Accenture- ATCI**

Context:-

We will be using automated script for installation & configurations of “**Hadoop/Spark Single Node Cluster**” on Ubuntu (18.04 LTS) Linux .

Installation from Ubuntu console

```
$ sudo apt-get install unzip
```

```
$ wget
```

```
https://github.com/rajuchal/hadoop\_light\_cloud/archive/master.zip
```

```
$ unzip master.zip
```

```
Archive:  hadoop_light_cloud.zip
  inflating: hadoop_light_cloud/core-site.xml
  extracting: hadoop_light_cloud/dataset.zip
  inflating: hadoop_light_cloud/hbase-env.sh
  inflating: hadoop_light_cloud/hbase-site.xml
  inflating: hadoop_light_cloud/hdfs-site.xml
  inflating: hadoop_light_cloud/hive-config.sh
  inflating: hadoop_light_cloud/hive-env.sh
  inflating: hadoop_light_cloud/hive-site.xml
  extracting: hadoop_light_cloud/hosts
  inflating: hadoop_light_cloud/install.sh
  inflating: hadoop_light_cloud/mapred-site.xml
  extracting: hadoop_light_cloud/masters
  inflating: hadoop_light_cloud/my.cnf
  extracting: hadoop_light_cloud/regionserver
  extracting: hadoop_light_cloud/slaves
  inflating: hadoop_light_cloud/spark-defaults.conf
  inflating: hadoop_light_cloud/spark-env.sh
  inflating: hadoop_light_cloud/yarn-site.xml
```

```
$ mv hadoop_light_cloud-master hadoop_light_cloud
```

```
$ cd hadoop_light_cloud/
```

```
$ chmod 755 install.sh
```

\$./install.sh

```
mysql: [Warning] Using a password on the command line interface can be insecure.
mysql: [Warning] Using a password on the command line interface can be insecure.
mysql: [Warning] Using a password on the command line interface can be insecure.
mysql: [Warning] Using a password on the command line interface can be insecure.
mysql: [Warning] Using a password on the command line interface can be insecure.
metastore_db @MySQL server created
/home/ubuntu/bigdata
Downloading Hadoop
Downloading Spark
```

```
20/07/22 11:28:05 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
20/07/22 11:28:05 INFO spark.SparkContext: Successfully stopped SparkContext
20/07/22 11:28:05 INFO util.ShutdownHookManager: Shutdown hook called
20/07/22 11:28:05 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-f44cd999-1054-4a89-8a47-d3b886839bbf
20/07/22 11:28:05 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-6c4fbb6c-8975-4654-b9e8-841d1cb6699b
Your environment is ready
```

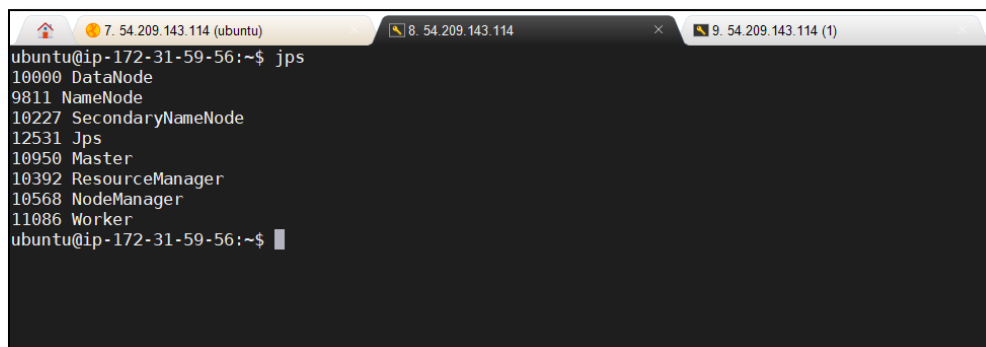
[Note :- it will take 15 minutes to complete the installations]

\$ cd

\$ clear

\$ source .bashrc

\$ jps



```
ubuntu@ip-172-31-59-56:~$ jps
10000 DataNode
9811 NameNode
10227 SecondaryNameNode
12531 Jps
10950 Master
10392 ResourceManager
10568 NodeManager
11086 Worker
ubuntu@ip-172-31-59-56:~$
```

\$ hdfs dfs -ls

Found 1 items

drwxr-xr-x - ubuntu supergroup 0 2020-07-22 11:27 wordcount

```
ubuntu@ip-172-31-68-92:~$ ls bigdata/
cassandra hadoop hbase hive java kafka mongodb mysql-connector pig sbt scala spark sqoop
ubuntu@ip-172-31-68-92:~$
```

```
$ pyspark --master spark://localhost:7077
```

```
Python 3.6.9 (default, Apr 18 2020, 01:56:04)
[GCC 8.4.0] on linux
Type "help", "copyright", "credits" or "license()" for more information.
20/07/22 11:31:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

      _ _ _ _ _
     / _ _ _ _ \
    / _ _ _ _ \
   / _ _ _ _ \
  / _ _ _ _ \
 / _ _ _ _ \
/_ _ _ _ _ \

version 2.4.5

Using Python version 3.6.9 (default, Apr 18 2020 01:56:04)
SparkSession available as 'spark'.
>>>
```

```
>>> quit()
```

```
ubuntu@ip-172-31-68-92:~$ spark-shell --master spark://localhost:7077
```

```
20/07/22 11:33:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://ip-172-31-68-92.ec2.internal:4040
Spark context available as 'sc' (master = spark://localhost:7077, app id = app-20200722113326-0002).
Spark session available as 'spark'.
Welcome to

      _ _ _ _ _
     / _ _ _ _ \
    / _ _ _ _ \
   / _ _ _ _ \
  / _ _ _ _ \
 / _ _ _ _ \
/_ _ _ _ _ \

version 2.4.5

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_131)
Type in expressions to have them evaluated.
Type :help for more information.
```

```
scala> :q
```

```
ubuntu@ip-172-31-68-92:~$
```

Stop All the Services

```
$ stop-dfs.sh
```

```
Stopping namenodes on [localhost]
localhost: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
```

```
$ stop-yarn.sh
```

```
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
localhost: nodemanager did not stop gracefully after 5 seconds: killing with kill -9
no proxyserver to stop
```

```
$ stop-master.sh
```

```
stopping org.apache.spark.deploy.master.Master
```

```
$ stop-slaves.sh
```

```
localhost: stopping org.apache.spark.deploy.worker.Worker
```

```
$ jps
```

```
24410 Jps
```

Start All the Services

```
$ start-dfs.sh
```

```
Starting namenodes on [localhost]
```

```
localhost: starting namenode, logging to /home/ubuntu/bigdata/hadoop/logs/hadoop-ubuntu-namenode-ip-172-31-68-92.out
```

```
localhost: starting datanode, logging to /home/ubuntu/bigdata/hadoop/logs/hadoop-ubuntu-datanode-ip-172-31-68-92.out
```

```
Starting secondary namenodes [0.0.0.0]
```

```
0.0.0.0: starting secondarynamenode, logging to /home/ubuntu/bigdata/hadoop/logs/hadoop-ubuntu-secondarynamenode-ip-172-31-68-92.out
```

```
$ start-yarn.sh
```

```
starting yarn daemons
```

```
starting resourcemanager, logging to /home/ubuntu/bigdata/hadoop/logs/yarn-ubuntu-resourcemanager-ip-172-31-68-92.out
```

```
localhost: starting nodemanager, logging to /home/ubuntu/bigdata/hadoop/logs/yarn-ubuntu-nodemanager-ip-172-31-68-92.out
```

```
$ start-master.sh
```

```
starting org.apache.spark.deploy.master.Master, logging to /home/ubuntu/bigdata/spark/logs/spark-ubuntu-org.apache.spark.deploy.master.Master-1-ip-172-31-68-92.out
```

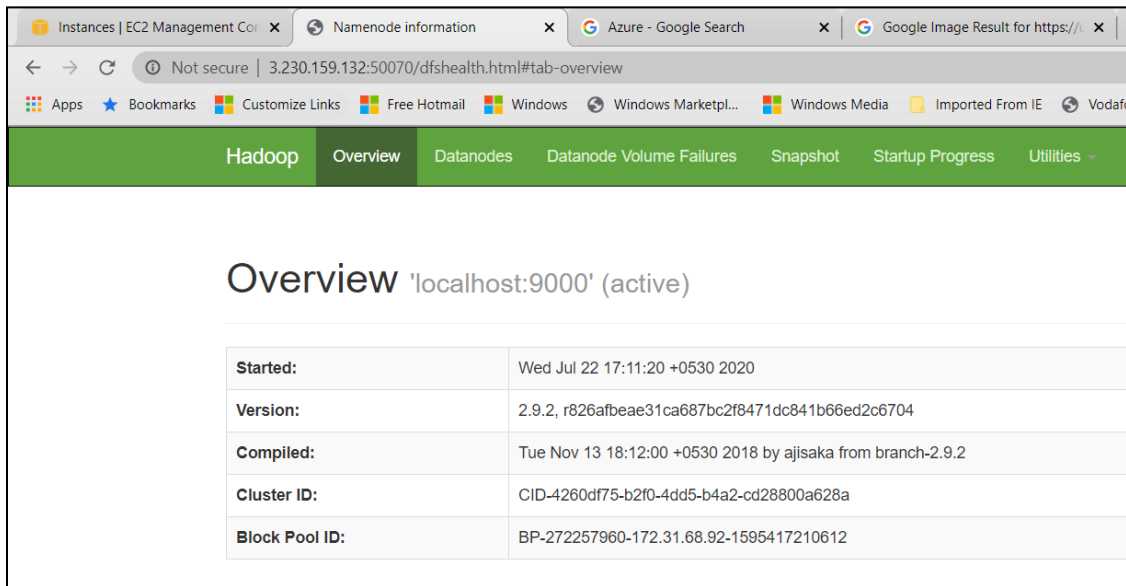
```
$ start-slaves.sh
```

```
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /home/ubuntu/bigdata/spark/logs/spark-ubuntu-org.apache.spark.deploy.worker.Worker-1-ip-172-31-68-92.out
```

```
ubuntu@ip-172-31-68-92:~$ jps
```

Check Namenode web interface

<http://<public ip >:50070>

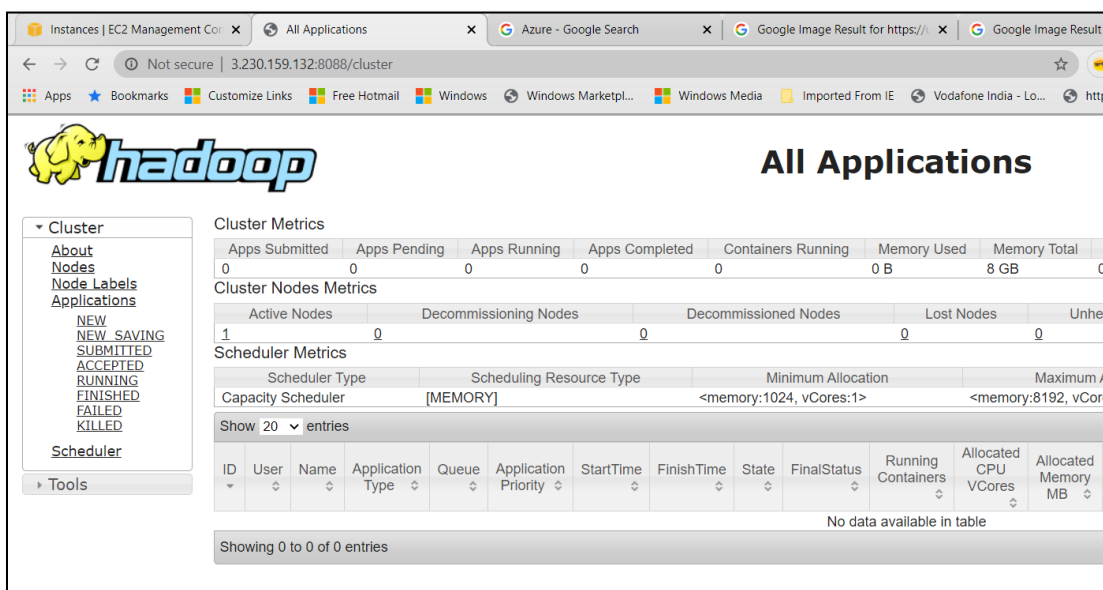


The screenshot shows the Hadoop Namenode web interface. The browser tabs include 'Instances | EC2 Management Console', 'Namenode information', 'Azure - Google Search', and 'Google Image Result for https://'. The address bar shows 'Not secure | 3.230.159.132:50070/dfshealth.html#tab-overview'. The navigation bar has tabs for 'Hadoop', 'Overview', 'Datanodes', 'Datanode Volume Failures', 'Snapshot', 'Startup Progress', and 'Utilities'. The main content area is titled 'Overview 'localhost:9000' (active)' and contains a table with the following information:

Started:	Wed Jul 22 17:11:20 +0530 2020
Version:	2.9.2, r826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled:	Tue Nov 13 18:12:00 +0530 2018 by ajisaka from branch-2.9.2
Cluster ID:	CID-4260df75-b2f0-4dd5-b4a2-cd28800a628a
Block Pool ID:	BP-272257960-172.31.68.92-1595417210612

Check Resource Manager web interface

<http://<public ip >:8088>



The screenshot shows the Hadoop Resource Manager web interface. The browser tabs include 'Instances | EC2 Management Console', 'All Applications', 'Azure - Google Search', and 'Google Image Result for https://'. The address bar shows 'Not secure | 3.230.159.132:8088/cluster'. The navigation bar has tabs for 'Cluster', 'All Applications', 'Jobs', 'Containers', 'Scheduler', and 'Tools'. The main content area is titled 'All Applications' and contains a table with the following information:

Cluster Metrics	
Apps Submitted	Apps Pending
0	0

Cluster Nodes Metrics	
Active Nodes	Decommissioning Nodes
1	0

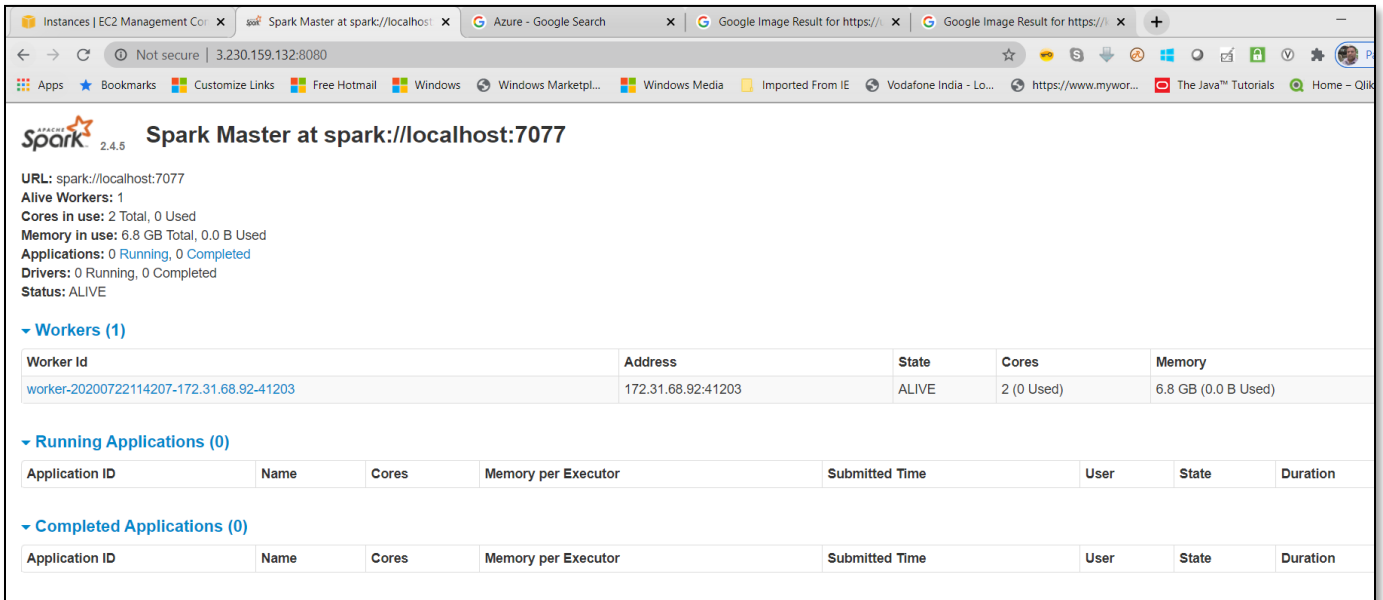
Scheduler Metrics	
Scheduler Type	Scheduling Resource Type
Capacity Scheduler	[MEMORY]

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB
No data available in table												

Showing 0 to 0 of 0 entries

Check Spark Master web interface

<http://<public ip >:8080>



Spark Master at spark://localhost:7077

URL: spark://localhost:7077
 Alive Workers: 1
 Cores in use: 2 Total, 0 Used
 Memory in use: 6.8 GB Total, 0.0 B Used
 Applications: 0 Running, 0 Completed
 Drivers: 0 Running, 0 Completed
 Status: ALIVE

▼ Workers (1)

Worker Id	Address	State	Cores	Memory
worker-20200722114207-172.31.68.92-41203	172.31.68.92:41203	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)

▼ Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	----------------	------	-------	----------

▼ Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	----------------	------	-------	----------

Stop the Ubuntu Instance

\$ sudo init 0