Federation University

ITECH2302 Big Data Management Laboratory - Hadoop

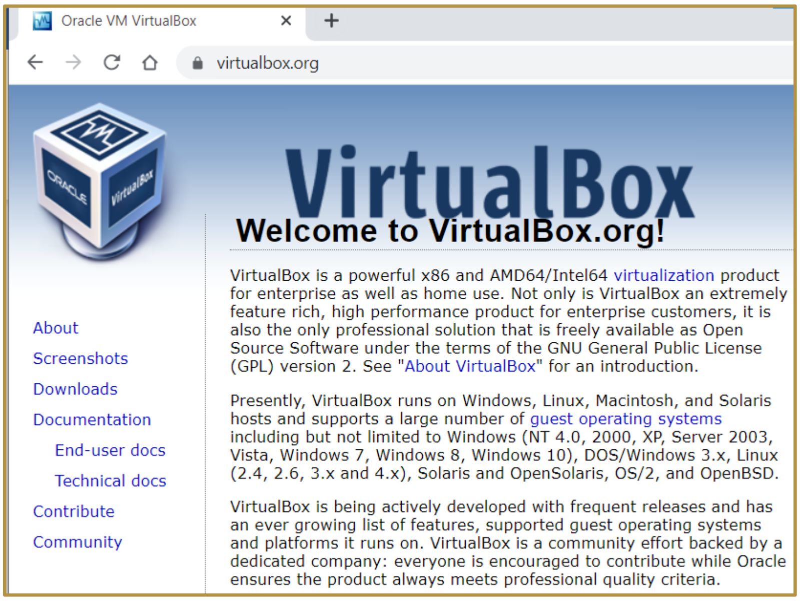
Objectives:

- Installation of Hadoop/Spark environment
- Introduction to Hadoop
- Review questions and activities

Activity 1

Installation of Hadoop/Spark environment

1. Download and install the latest version of VirtualBox for your Operating System:



Welcome to VirtualBox.org!

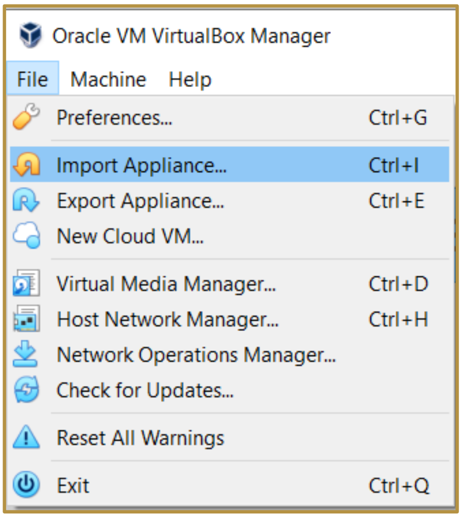
VirtualBox is a powerful x86 and AMD64/Intel64 virtualization product for enterprise as well as home use. Not only is VirtualBox an extremely feature rich, high performance product for enterprise customers, it is also the only professional solution that is freely available as Open Source Software under the terms of the GNU General Public License (GPL) version 2. See "About VirtualBox" for an introduction.

Presently, VirtualBox runs on Windows, Linux, Macintosh, and Solaris hosts and supports a large number of guest operating systems including but not limited to Windows (NT 4.0, 2000, XP, Server 2003, Vista, Windows 7, Windows 8, Windows 10), DOS/Windows 3.x, Linux (2.4, 2.6, 3.x and 4.x), Solaris and OpenSolaris, OS/2, and OpenBSD.

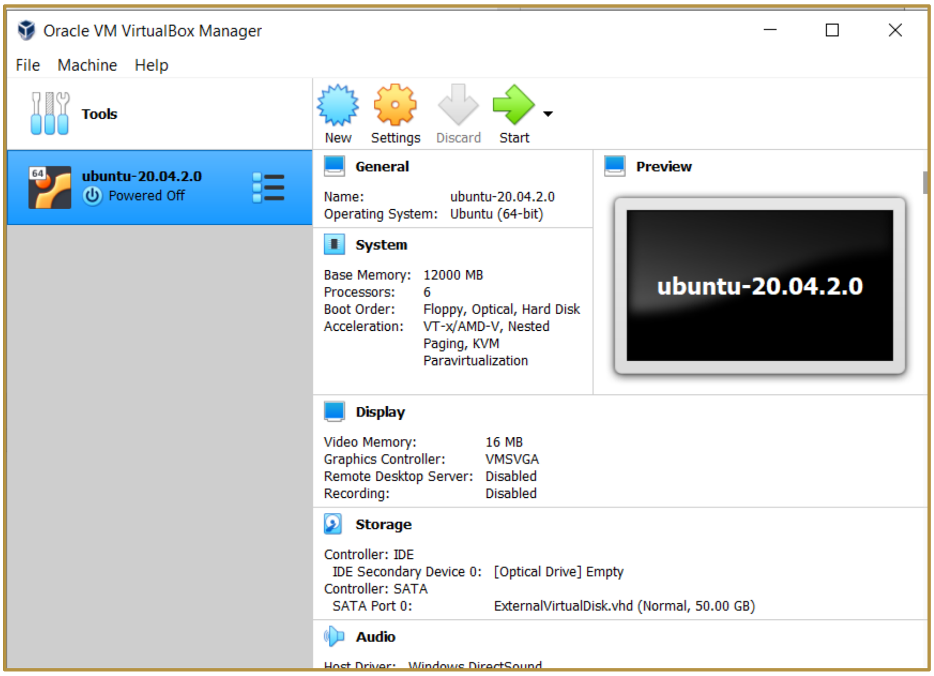
VirtualBox is being actively developed with frequent releases and has an ever growing list of features, supported guest operating systems and platforms it runs on. VirtualBox is a community effort backed by a dedicated company: everyone is encouraged to contribute while Oracle ensures the product always meets professional quality criteria.

2. Download the OVA file for Hadoop/Spark from the FedUni server (refer to Announcements for details).

3. Load the OVA file into Virtualbox as an Appliance:


Oracle VM VirtualBox Manager
File
Machine
Help
Preferences... Ctrl+G
Import Appliance... Ctrl+I
Export Appliance... Ctrl+E
New Cloud VM...
Virtual Media Manager... Ctrl+D
Host Network Manager... Ctrl+H
Network Operations Manager...
Check for Updates...
Reset All Warnings
Exit Ctrl+Q

4. Open the ubuntu operating system by selecting it and clicking the Start icon



ubuntu-20.04.2.0
Powered Off

General

Name: ubuntu-20.04.2.0
Operating System: Ubuntu (64-bit)

System

Base Memory: 12000 MB
Processors: 6
Boot Order: Floppy, Optical, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, KVM Paravirtualization

Display

Video Memory: 16 MB
Graphics Controller: VMSVGA
Remote Desktop Server: Disabled
Recording: Disabled

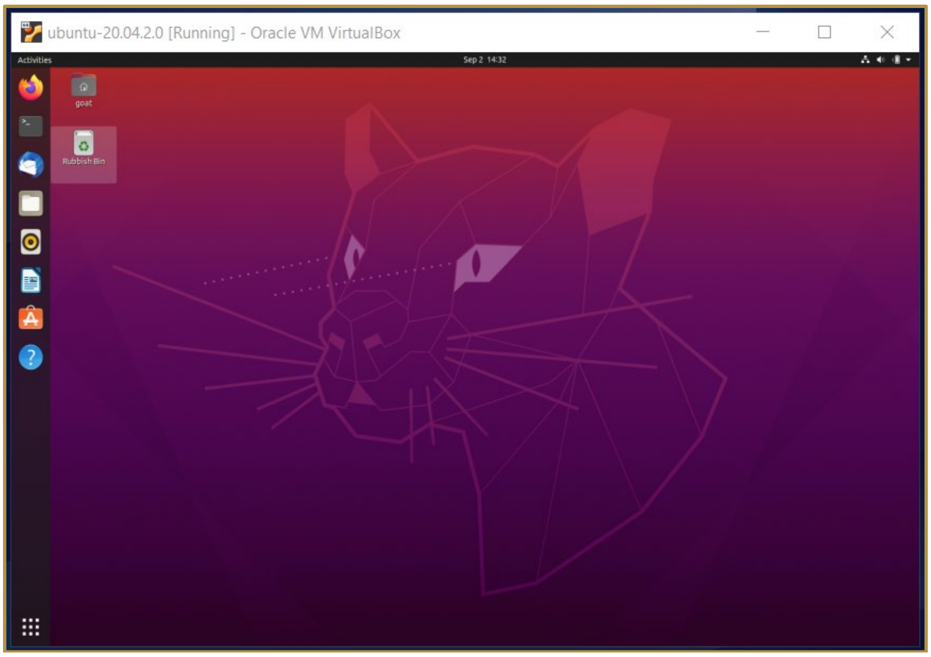
Storage

Controller: IDE
IDE Secondary Device 0: [Optical Drive] Empty
Controller: SATA
SATA Port 0: ExternalVirtualDisk.vhd (Normal, 50.00 GB)

Audio

Host Driver: Windows DirectSound

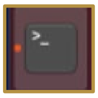
5. Use the provided login details if required (*username*: goat, *password*: goat).

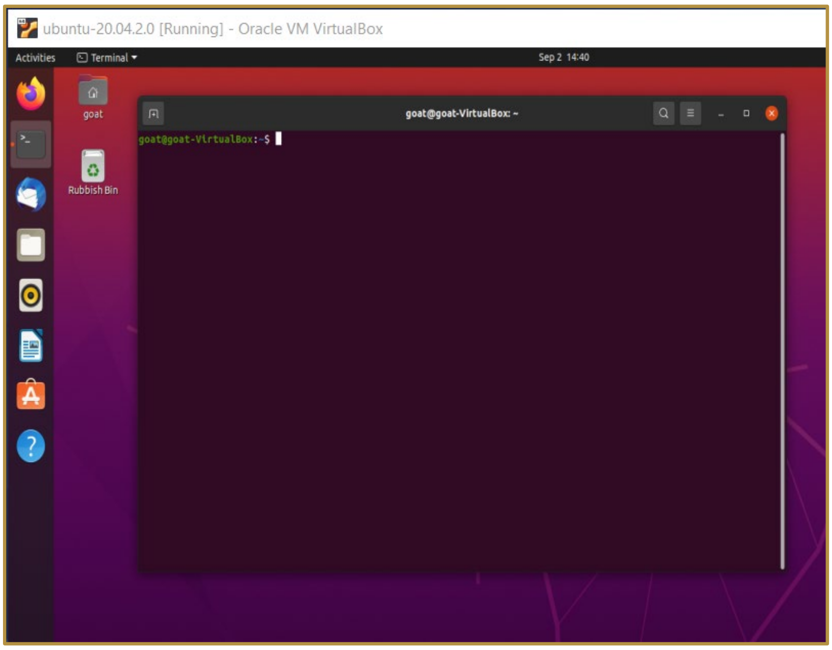


Activity 2

Introduction to Hadoop

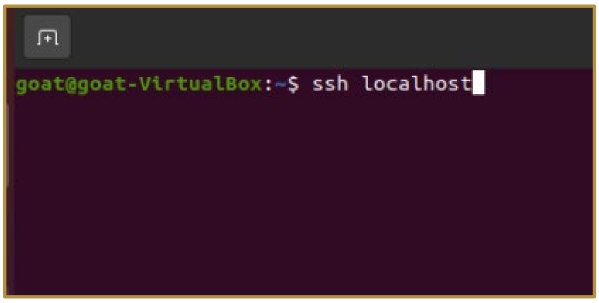
1. Open a terminal with the ubuntu operating system



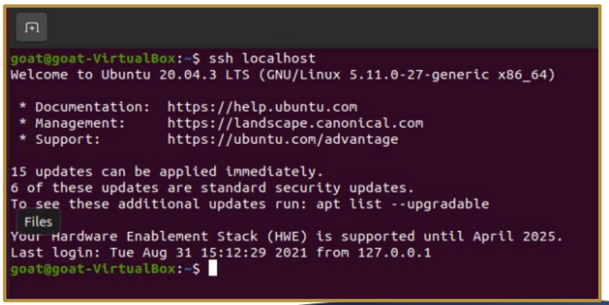


```
goat@goat-VirtualBox:~$
```

2. Familiarise yourself with writing command within the terminal:



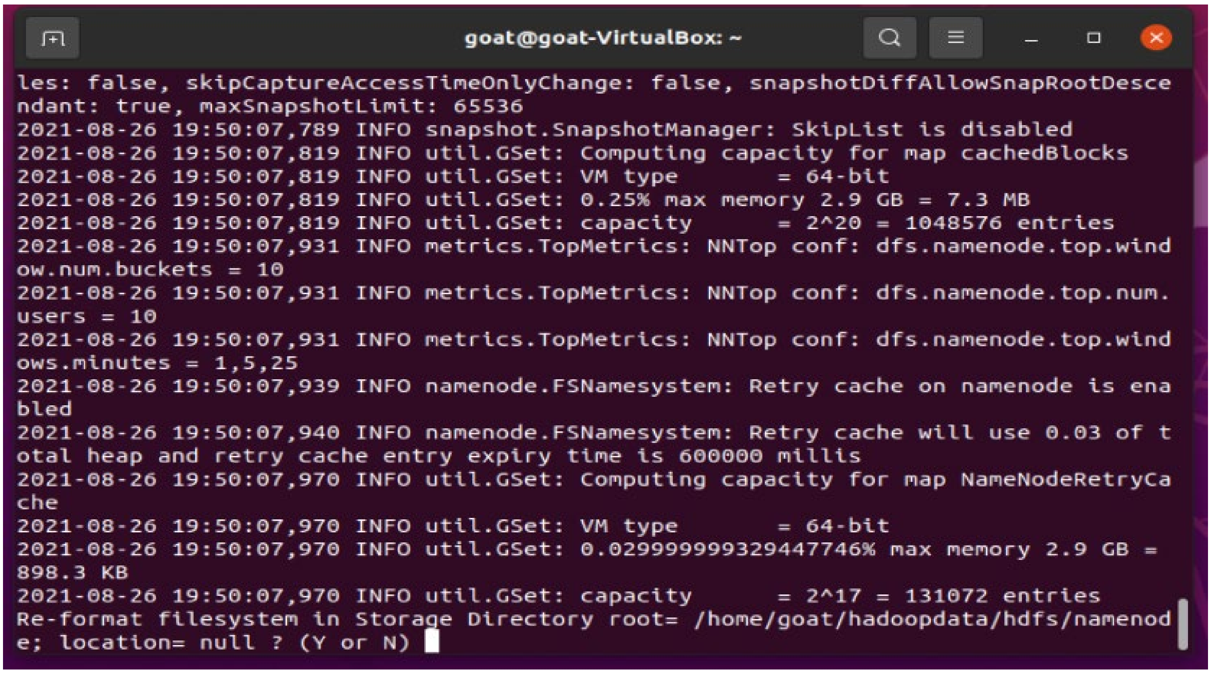
```
goat@goat-VirtualBox:~$ ssh localhost
```



```
goat@goat-VirtualBox:~$ ssh localhost
Welcome to Ubuntu 20.04.3 LTS (GNU/Linux 5.11.0-27-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

15 updates can be applied immediately.
6 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Your Hardware Enablement Stack (HWE) is supported until April 2025.
Last login: Tue Aug 31 15:12:29 2021 from 127.0.0.1
goat@goat-VirtualBox:~$
```

3. The following commands should be written within the terminal:

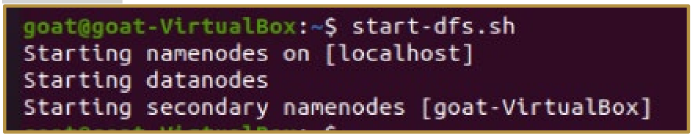
```
ssh localhost
hdfs namenode -format
```



```
les: false, skipCaptureAccessTimeOnlyChange: false, snapshotDiffAllowSnapRootDesce
ndant: true, maxSnapshotLimit: 65536
2021-08-26 19:50:07,789 INFO snapshot.SnapshotManager: SkipList is disabled
2021-08-26 19:50:07,819 INFO util.GSet: Computing capacity for map cachedBlocks
2021-08-26 19:50:07,819 INFO util.GSet: VM type       = 64-bit
2021-08-26 19:50:07,819 INFO util.GSet: 0.25% max memory 2.9 GB = 7.3 MB
2021-08-26 19:50:07,819 INFO util.GSet: capacity      = 2^20 = 1048576 entries
2021-08-26 19:50:07,931 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.wind
ow.num.buckets = 10
2021-08-26 19:50:07,931 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.
users = 10
2021-08-26 19:50:07,931 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.wind
ows.minutes = 1,5,25
2021-08-26 19:50:07,939 INFO namenode.FSNamesystem: Retry cache on namenode is ena
bled
2021-08-26 19:50:07,940 INFO namenode.FSNamesystem: Retry cache will use 0.03 of t
otal heap and retry cache entry expiry time is 600000 millis
2021-08-26 19:50:07,970 INFO util.GSet: Computing capacity for map NameNodeRetryCa
che
2021-08-26 19:50:07,970 INFO util.GSet: VM type       = 64-bit
2021-08-26 19:50:07,970 INFO util.GSet: 0.029999999329447746% max memory 2.9 GB =
898.3 KB
2021-08-26 19:50:07,970 INFO util.GSet: capacity      = 2^17 = 131072 entries
Re-format filesystem in Storage Directory root= /home/goat/hadoopdata/hdfs/namenod
e; location= null ? (Y or N)
```

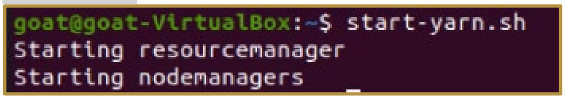
Click Y

```
start-dfs.sh
```



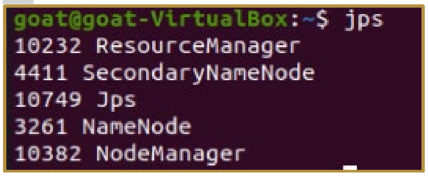
```
goat@goat-VirtualBox:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [goat-VirtualBox]
```

```
start-yarn.sh
```



```
goat@goat-VirtualBox:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```

```
jps
```

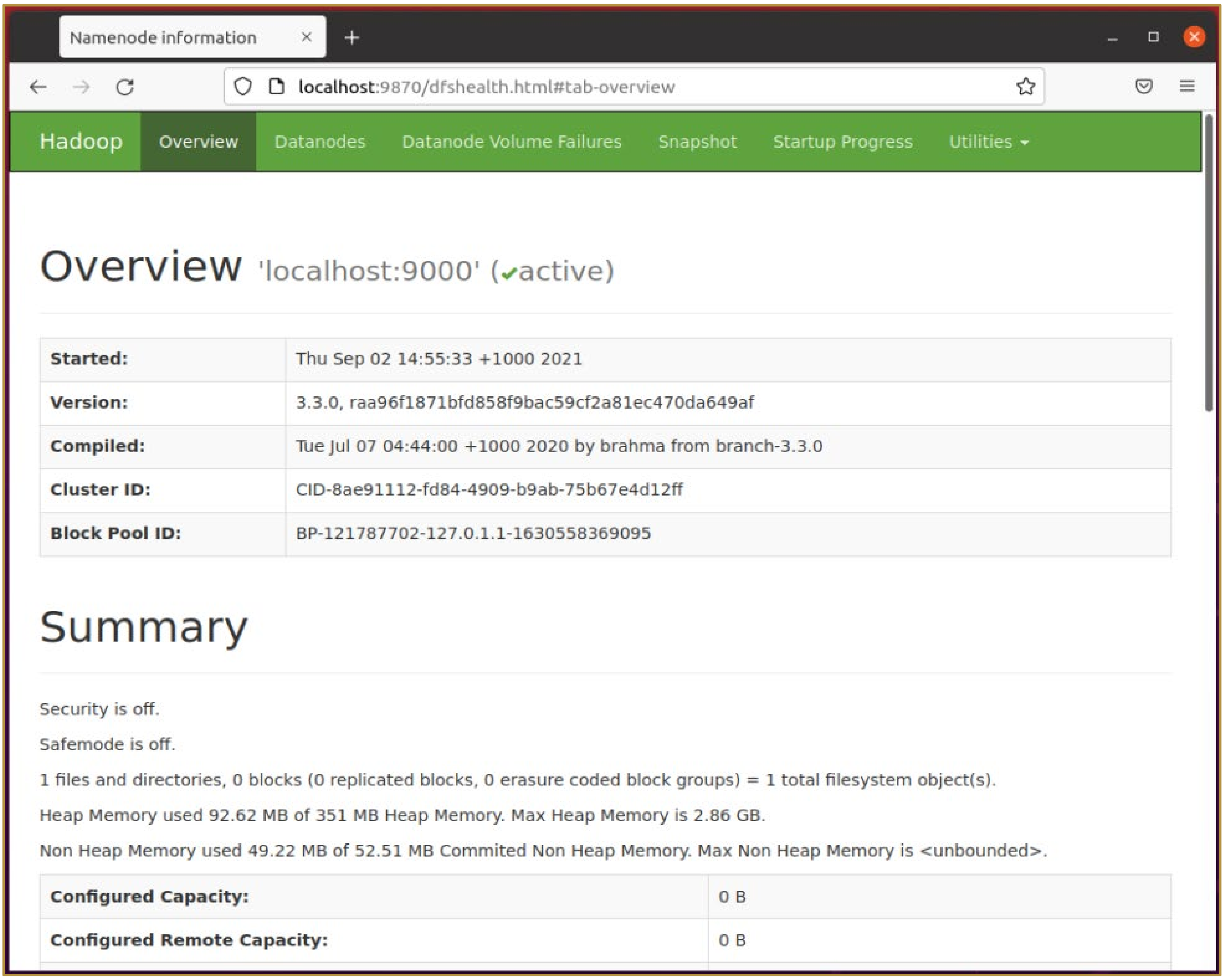


```
goat@goat-VirtualBox:~$ jps
10232 ResourceManager
4411 SecondaryNameNode
10749 Jps
3261 NameNode
10382 NodeManager
```

Open the following URL’s in Firefox:

http://localhost:9870
http://localhost:8088

The pages should look like the following:



Overview 'localhost:9000' (✓active)

Started:	Thu Sep 02 14:55:33 +1000 2021
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af
Compiled:	Tue Jul 07 04:44:00 +1000 2020 by brahma from branch-3.3.0
Cluster ID:	CID-8ae91112-fd84-4909-b9ab-75b67e4d12ff
Block Pool ID:	BP-121787702-127.0.1.1-1630558369095

Summary

Security is off.

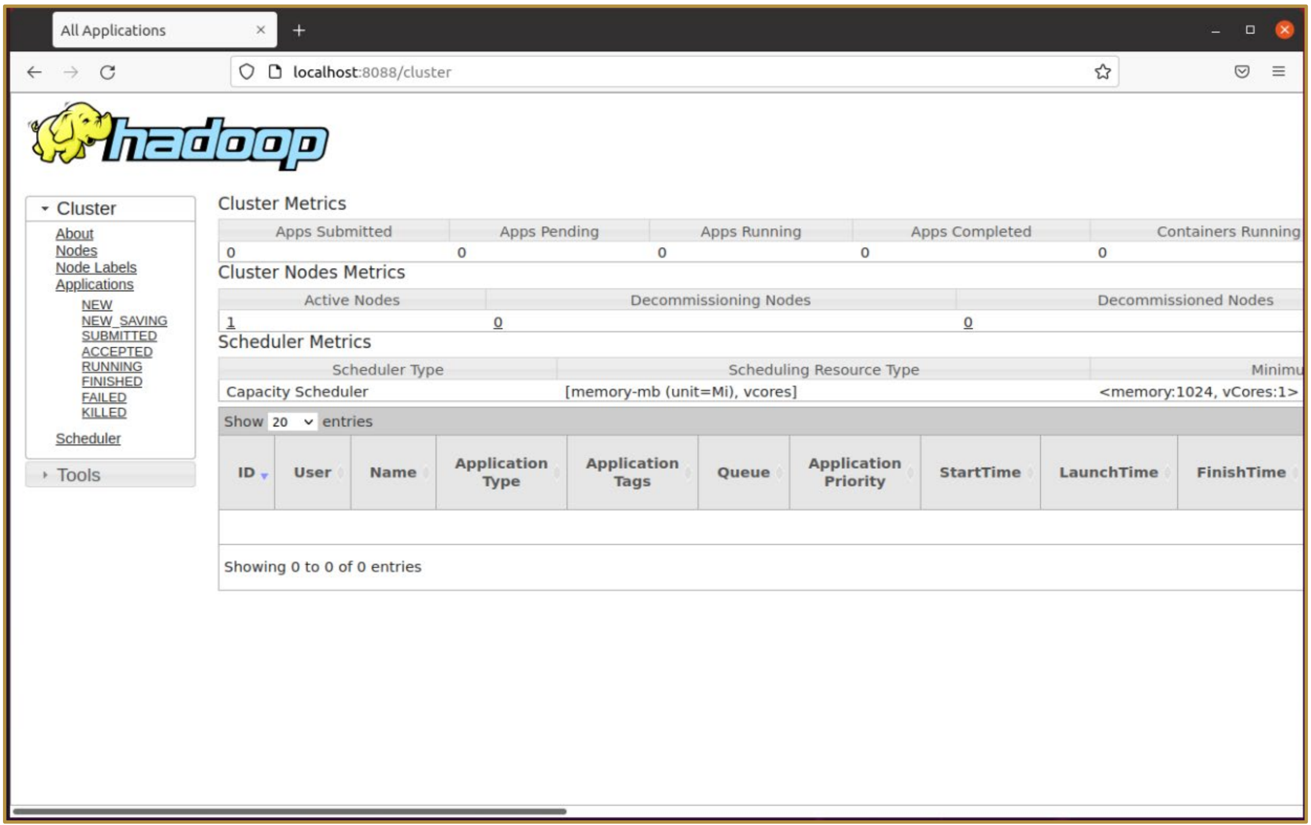
Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 92.62 MB of 351 MB Heap Memory. Max Heap Memory is 2.86 GB.

Non Heap Memory used 49.22 MB of 52.51 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	0 B
Configured Remote Capacity:	0 B



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

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
0	0	0	0	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

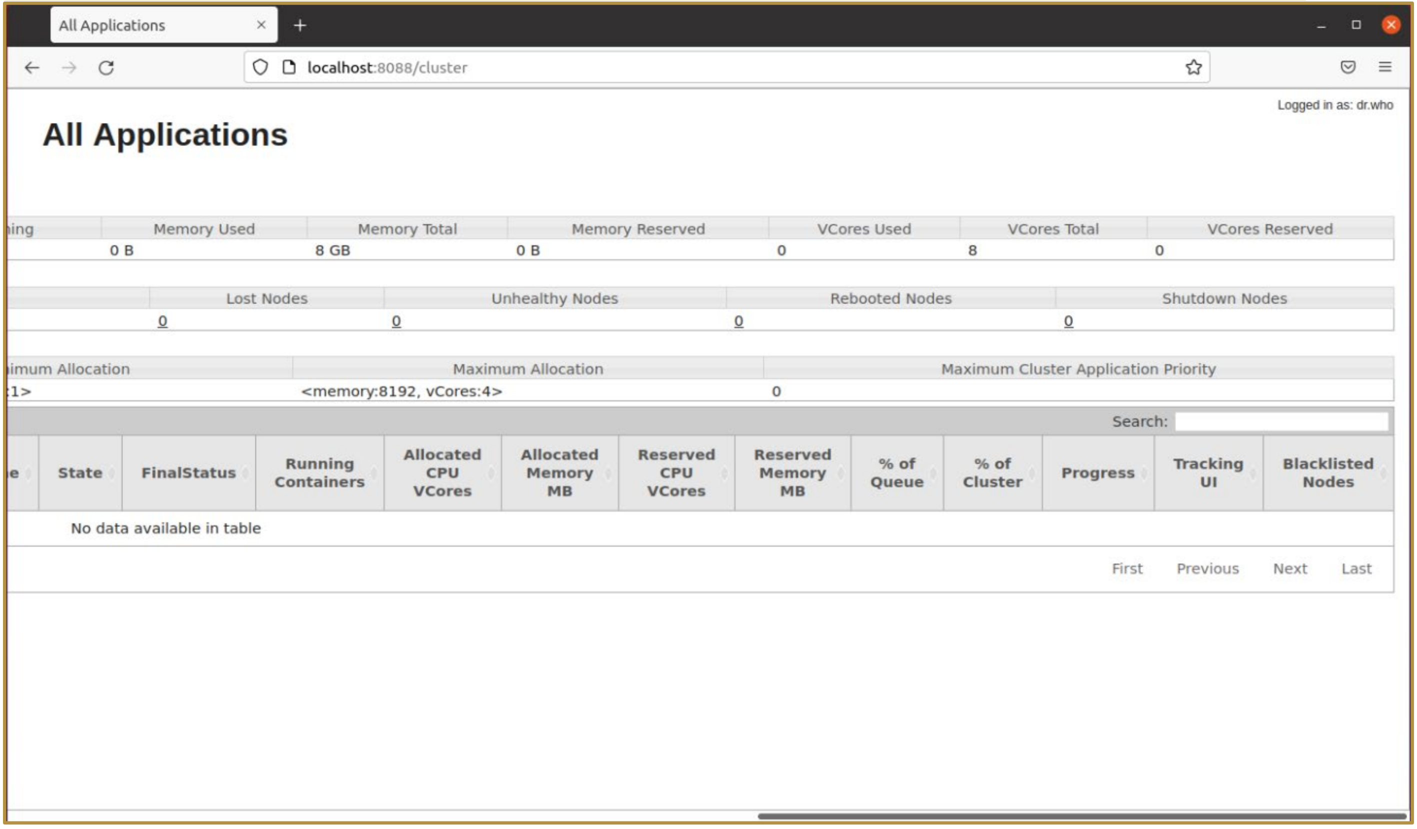
Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimu
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime

Showing 0 to 0 of 0 entries



Logged in as: dr.who

All Applications

ing	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
	0 B	8 GB	0 B	0	8	0

	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
	0	0	0	0

imum Allocation	Maximum Allocation	Maximum Cluster Application Priority
:1>	<memory:8192, vCores:4>	0

Search:

e	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes

No data available in table

First Previous Next Last

4. At the end of your session you can issue the following commands:

```
stop-dfs.sh
stop-yarn.sh
```

Or:

```
stop-all.sh
```

Activity 3

There are many great resources for the whole ecosystem, covering a broad set of topics:

- https://www.edureka.co/blog/hadoop-ecosystem