Experiment 2

Aim : Install Hadoop on a Single Node Cluster.

Theory :

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.
Hadoop consists of four main modules:
1. Hadoop Distributed File System (HDFS) - A distributed file system that runs on standard or low-end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support of large datasets.
2. Yet Another Resource Negotiator (YARN) - Manages and monitors cluster nodes and resource usage. It schedules jobs and tasks.
3. MapReduce - A framework that helps programs do the parallel computation on data. The map task takes input data and converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduce tasks to aggregate output and provide the desired result.
4. Hadoop Common - Provides common Java libraries that can be used across all modules.

How Hadoop Works?

● Hadoop makes it easier to use all the storage and processing capacity in cluster servers, and to execute distributed processes against huge amounts of data.
● Hadoop provides the building blocks on which other services and applications can be built.Applications that collect data in various formats can place data into the Hadoop cluster by using an API operation to connect to the NameNode.
● The NameNode tracks the file directory structure and placement of “chunks” for each file, replicated across DataNodes.
● To run a job to query the data, provide a MapReduce job made up of many map and reduce tasks that run against the data in HDFS spread across the DataNodes.
● Map tasks run on each node against the input files supplied, and reducers run to aggregate and organize the final output.
● The Hadoop ecosystem has grown significantly over the years due to its extensibility.
● Today, the Hadoop ecosystem includes many tools and applications to help collect, store, process, analyze, and manage big data.
● Some of the most popular applications are:

➔ Spark - An open source, distributed processing system commonly used for big data workloads. Apache Spark uses in-memory caching and optimized execution for fast performance, and it supports general batch processing, streaming analytics, machine learning, graph databases, and ad hoc queries.

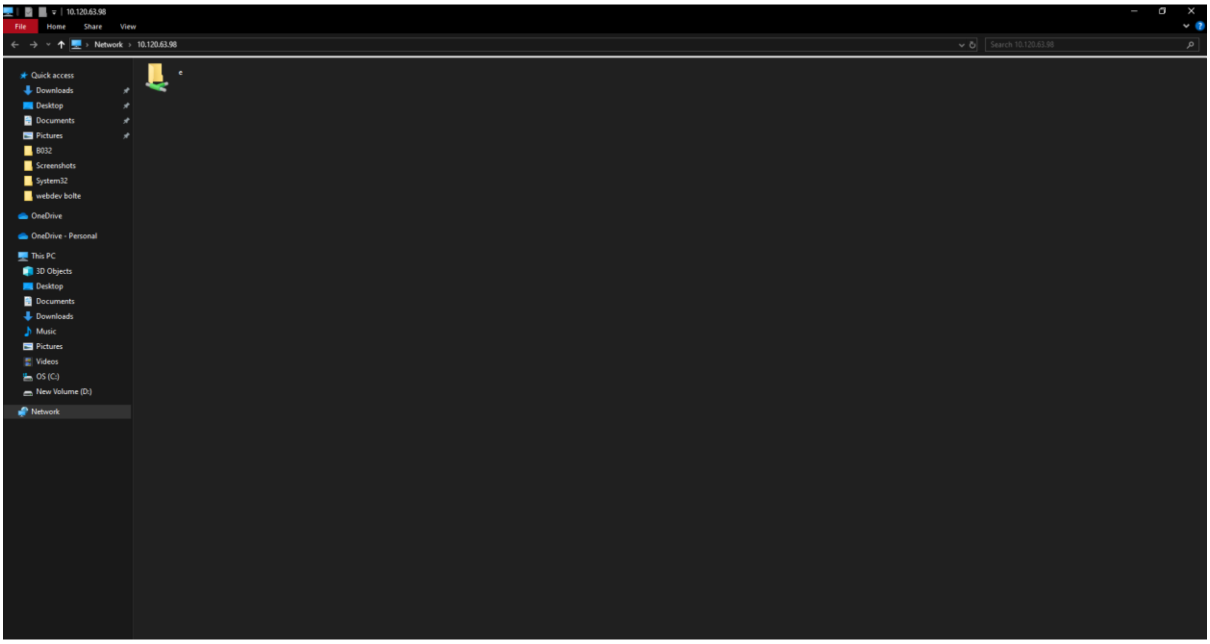
➔ Presto - An open source, distributed SQL query engine optimized for low-latency, ad-hoc analysis of data. It supports the ANSI SQL standard, including complex queries, aggregations, joins, and window functions. Presto can process data from multiple data sources including the Hadoop Distributed File System (HDFS) and Amazon S3.

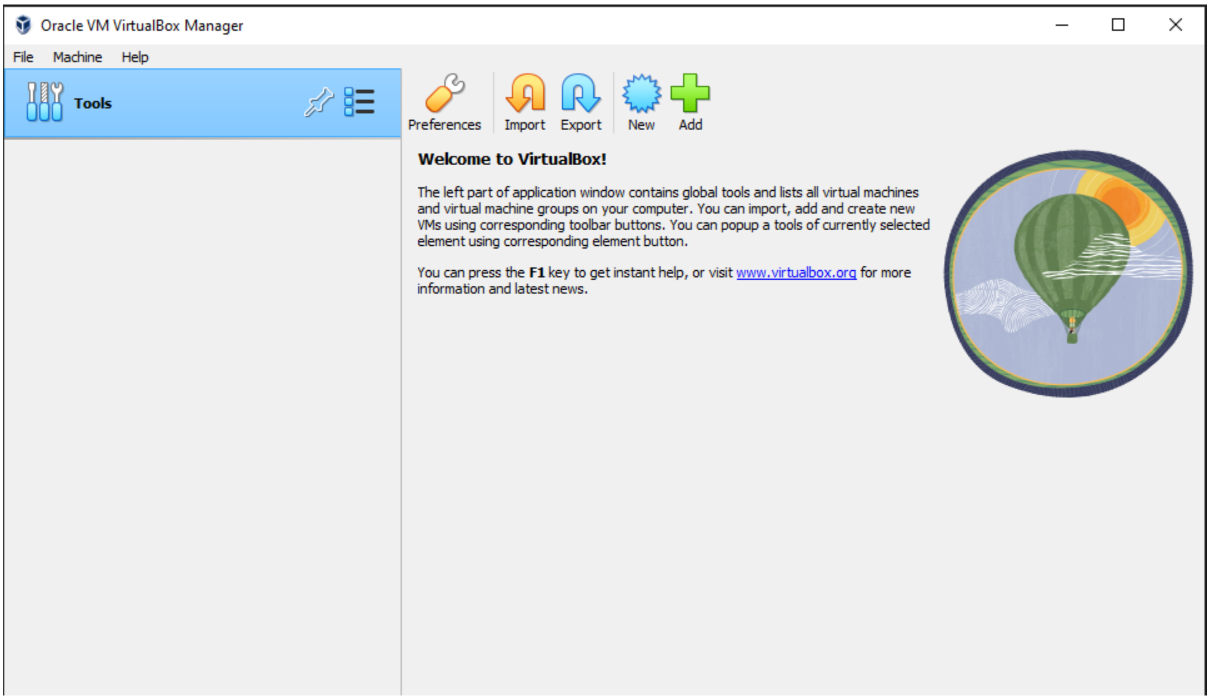
➔ Hive - Allows users to leverage Hadoop MapReduce using a SQL interface, enabling analytics at a massive scale, in addition to distributed and fault-tolerant data warehousing.

➔ HBase - An open source, non-relational, versioned database that runs on top of Amazon S3 (using EMRFS) or the Hadoop Distributed File System (HDFS). HBase is a massively scalable, distributed big data store built for random, strictly consistent, real-time access for tables with billions of rows and millions of columns.

➔ Zeppelin - An interactive notebook that enables interactive data exploration.

Screenshots :

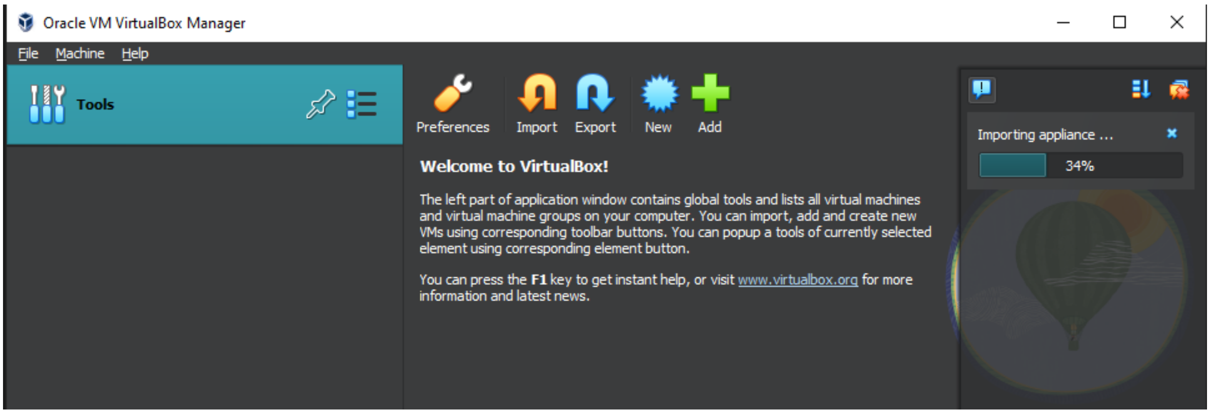




Welcome to VirtualBox!

The left part of application window contains global tools and lists all virtual machines and virtual machine groups on your computer. You can import, add and create new VMs using corresponding toolbar buttons. You can popup a tools of currently selected element using corresponding element button.

You can press the **F1** key to get instant help, or visit www.virtualbox.org for more information and latest news.



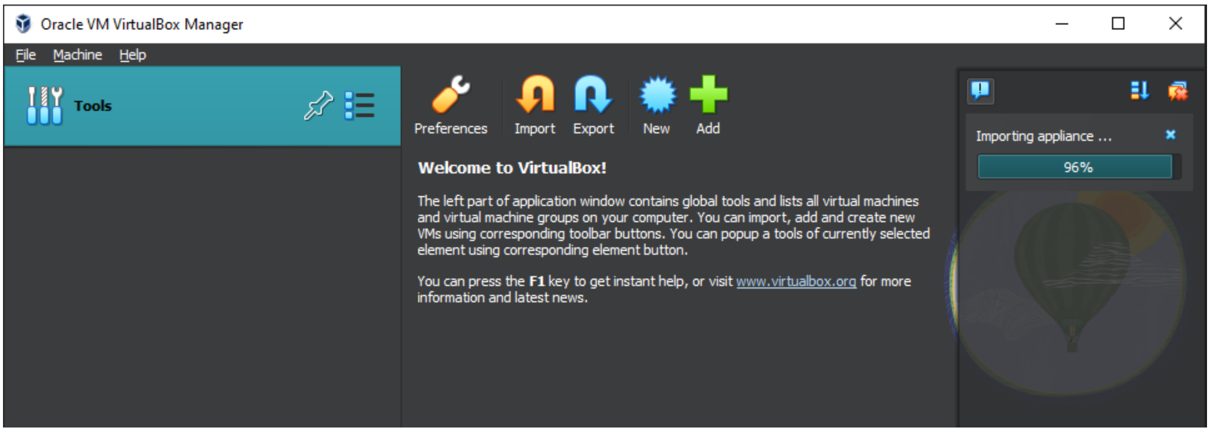
Importing appliance ...

34%

Welcome to VirtualBox!

The left part of application window contains global tools and lists all virtual machines and virtual machine groups on your computer. You can import, add and create new VMs using corresponding toolbar buttons. You can popup a tools of currently selected element using corresponding element button.

You can press the **F1** key to get instant help, or visit www.virtualbox.org for more information and latest news.



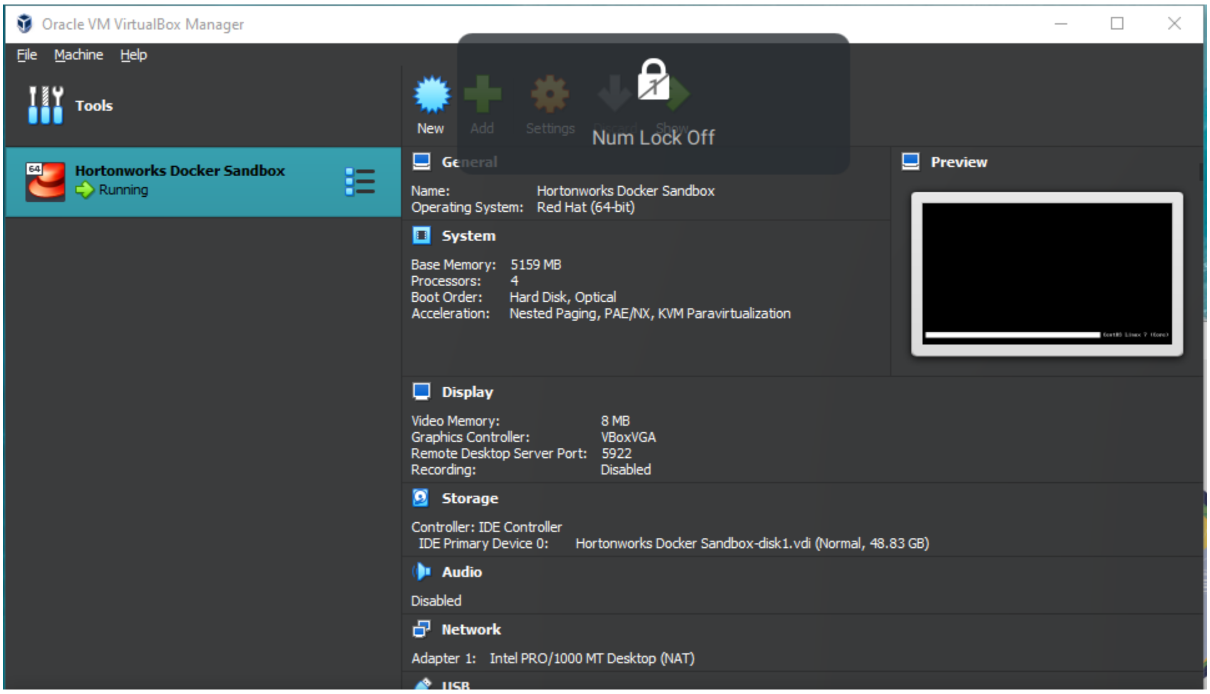
Importing appliance ...

96%

Welcome to VirtualBox!

The left part of application window contains global tools and lists all virtual machines and virtual machine groups on your computer. You can import, add and create new VMs using corresponding toolbar buttons. You can popup a tools of currently selected element using corresponding element button.

You can press the **F1** key to get instant help, or visit www.virtualbox.org for more information and latest news.



Hortonworks Docker Sandbox
Running

General

Name: Hortonworks Docker Sandbox
Operating System: Red Hat (64-bit)

System

Base Memory: 5159 MB
Processors: 4
Boot Order: Hard Disk, Optical
Acceleration: Nested Paging, PAE/NX, KVM Paravirtualization

Display

Video Memory: 8 MB
Graphics Controller: VBoxVGA
Remote Desktop Server Port: 5922
Recording: Disabled

Storage

Controller: IDE Controller
IDE Primary Device 0: Hortonworks Docker Sandbox-disk1.vdi (Normal, 48.83 GB)

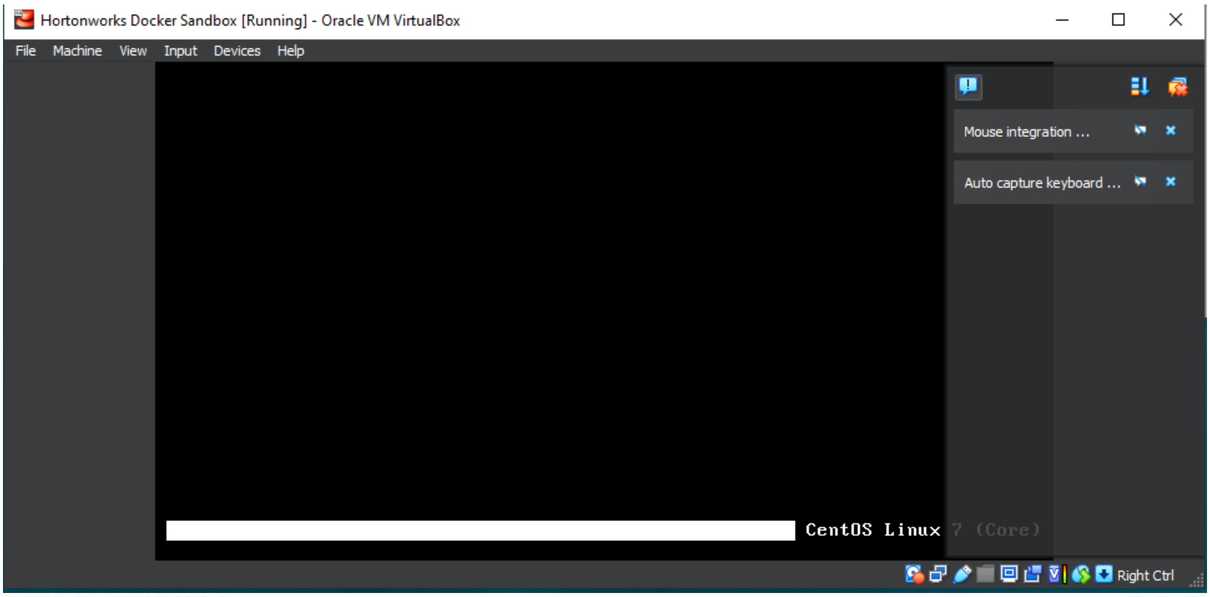
Audio

Disabled

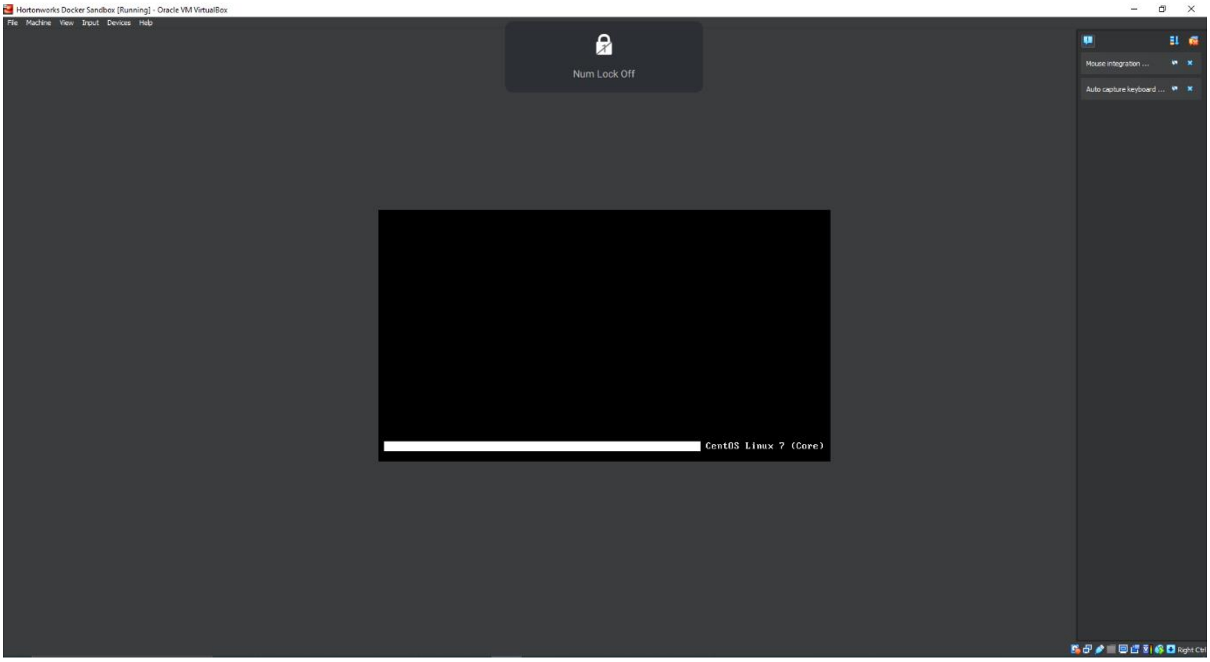
Network

Adapter 1: Intel PRO/1000 MT Desktop (NAT)

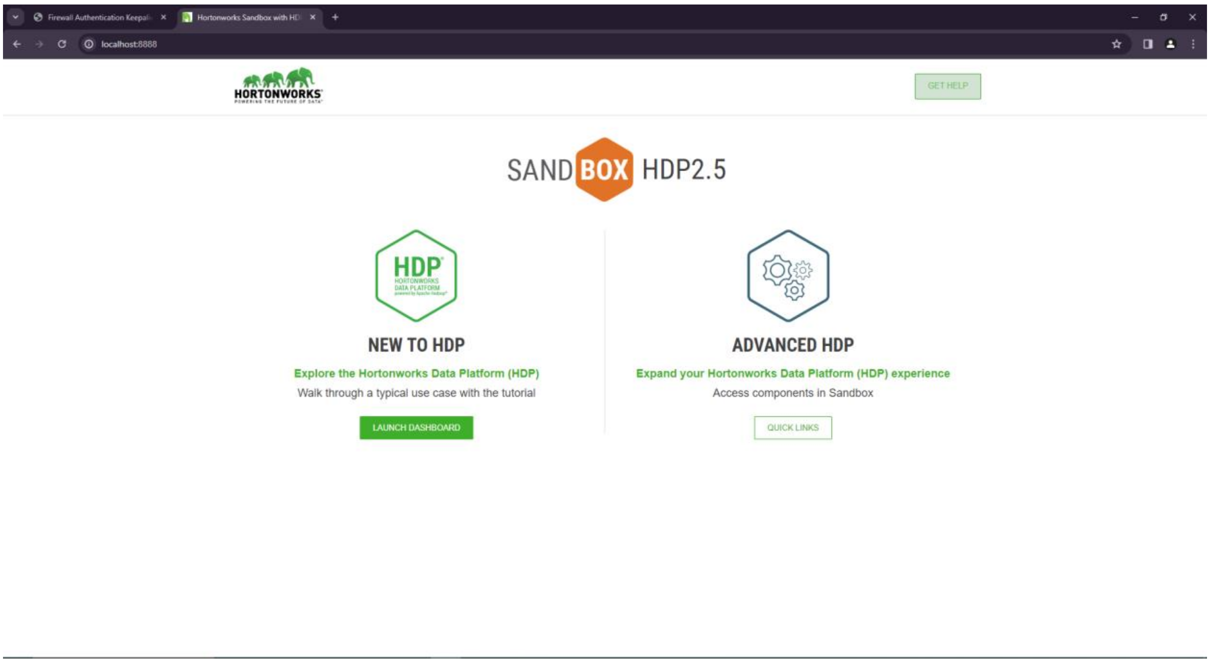
USB


Hortonworks Docker Sandbox [Running] - Oracle VM VirtualBox
File
Machine
View
Input
Devices
Help
Mouse integration ...
Auto capture keyboard ...
Right Ctrl

```
CentOS Linux 7 (Core)
```


Hortonworks Docker Sandbox [Running] - Oracle VM VirtualBox
File
Machine
View
Input
Devices
Help
Num Lock Off
Mouse integration ...
Auto capture keyboard ...
Right Ctrl

```
CentOS Linux 7 (Core)
```



NEW TO HDP

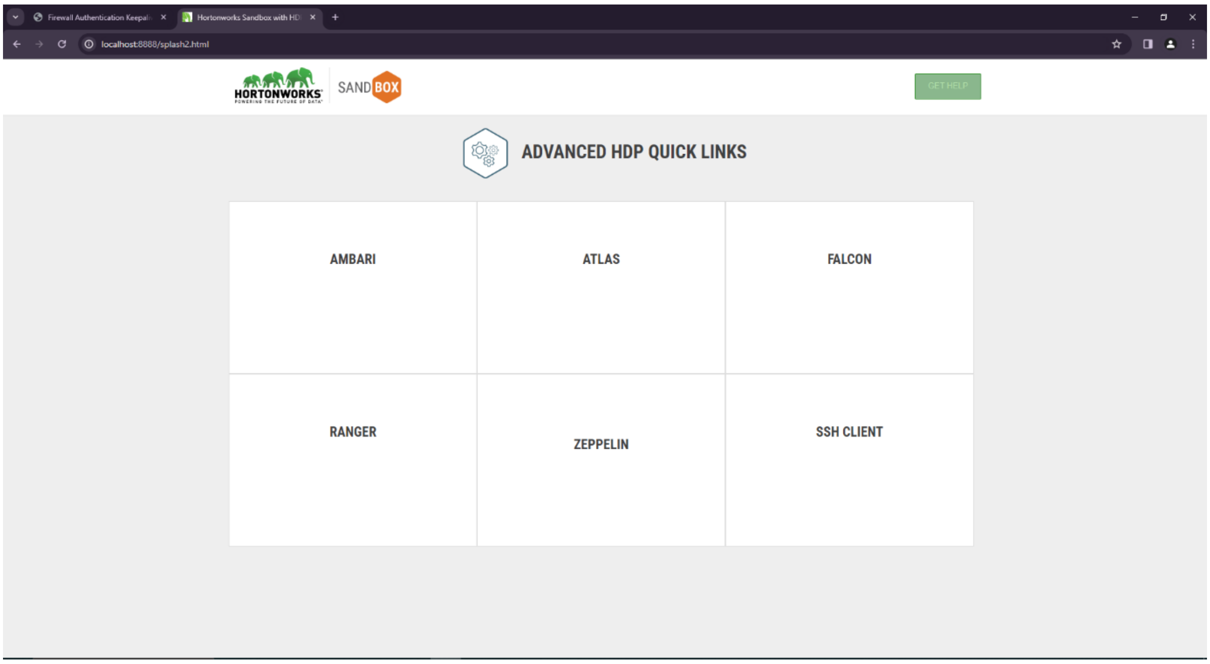
Explore the Hortonworks Data Platform (HDP)

Walk through a typical use case with the tutorial

ADVANCED HDP

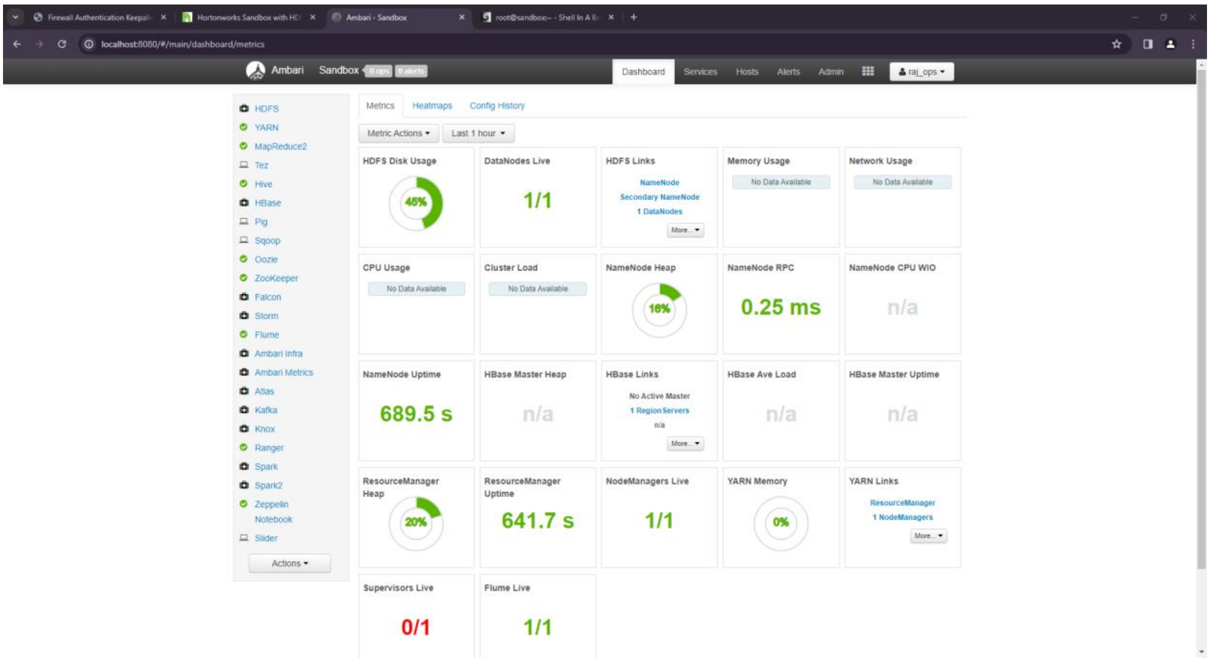
Expand your Hortonworks Data Platform (HDP) experience

Access components in Sandbox

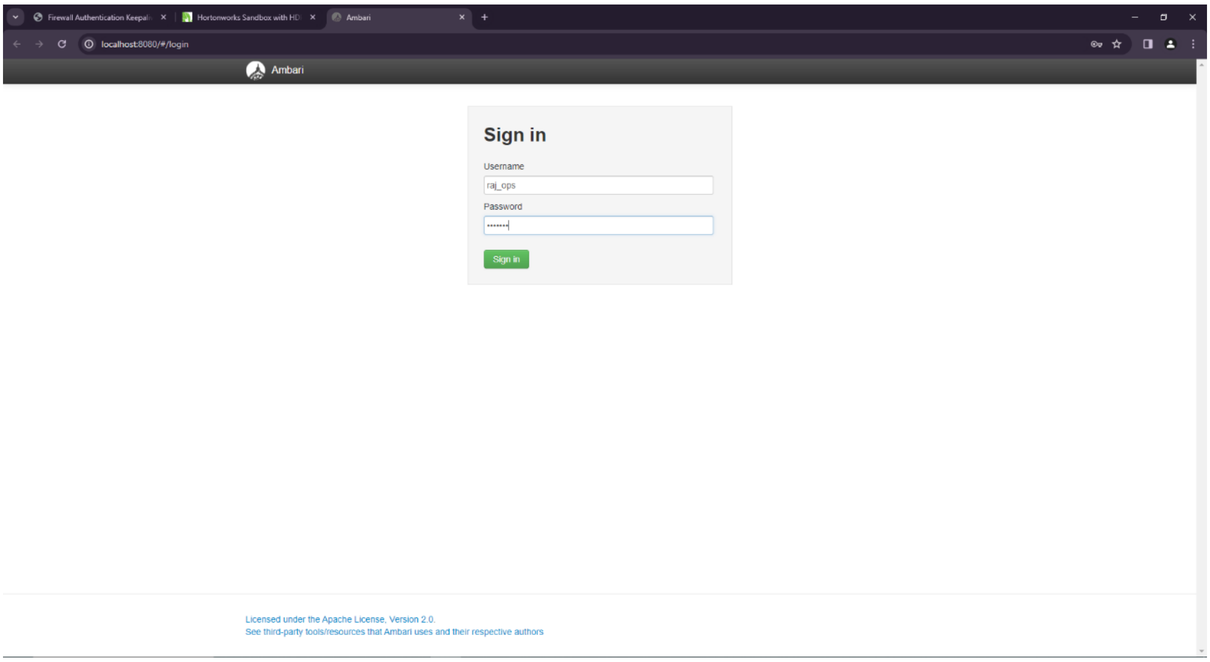

Firewall Authentication Keepali
Hortonworks Sandbox with HD
localhost:8888/splash2.html
HORTONWORKS
SANDBOX
GET HELP

ADVANCED HDP QUICK LINKS

AMBARI	ATLAS	FALCON
RANGER	ZEPPELIN	SSH CLIENT



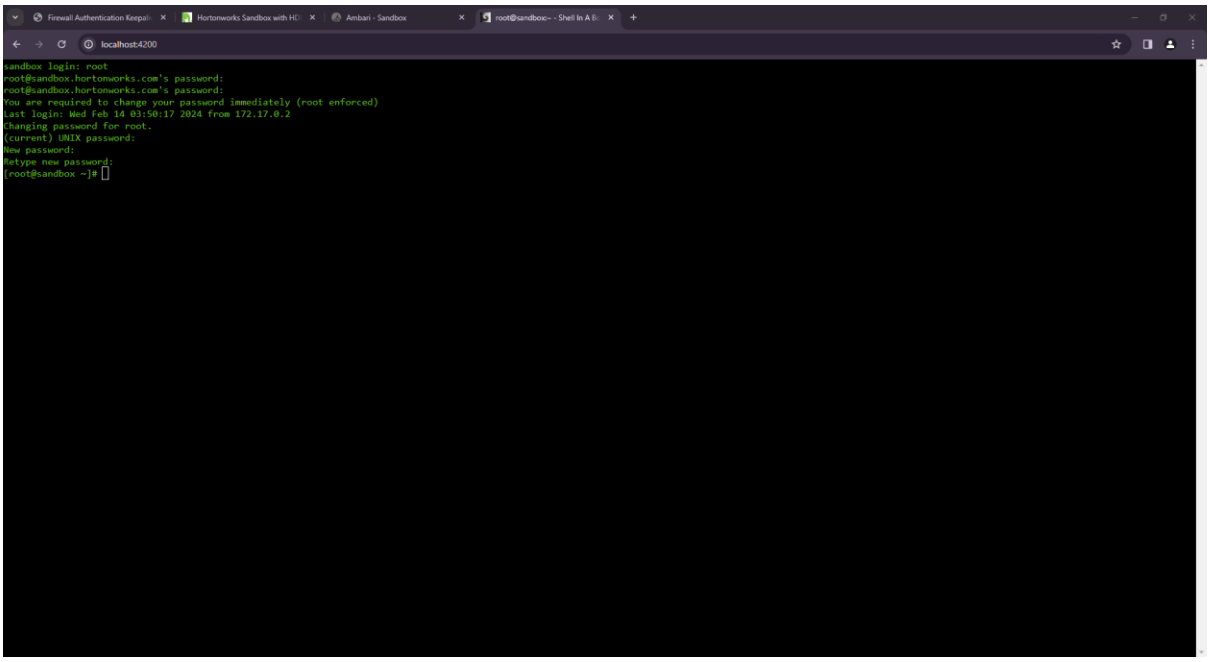
HDFS Disk Usage: 45%
DataNodes Live: 1/1
HDFS Links: NameNode, Secondary NameNode, 1 DataNodes
Memory Usage: No Data Available
Network Usage: No Data Available
CPU Usage: No Data Available
Cluster Load: No Data Available
NameNode Heap: 16%
NameNode RPC: 0.25 ms
NameNode CPU WIO: n/a
NameNode Uptime: 689.5 s
HBase Master Heap: n/a
HBase Links: No Active Master, 1 RegionServers, n/a
HBase Ave Load: n/a
HBase Master Uptime: n/a
ResourceManager Heap: 20%
ResourceManager Uptime: 641.7 s
NodeManagers Live: 1/1
YARN Memory: 0%
YARN Links: ResourceManager, 1 NodeManagers
Supervisors Live: 0/1
Flume Live: 1/1



Sign in

Username: raj_ops

Password: •••••••



Firewall Authentication Keepali
Hortonworks Sandbox with HD
Ambari - Sandbox
root@sandbox:~ - Shell In A Bo
localhost:4200

```
sandbox login: root
root@sandbox.hortonworks.com's password:
root@sandbox.hortonworks.com's password:
You are required to change your password immediately (root enforced)
Last login: Wed Feb 14 03:50:17 2024 from 172.17.0.2
Changing password for root.
(current) UNIX password:
New password:
Retype new password:
[root@sandbox ~]# 
```

Conclusion : We have successfully installed Hadoop.