Hortonwork sandbox installation

Maimoona Khilji

Institute of Management Science

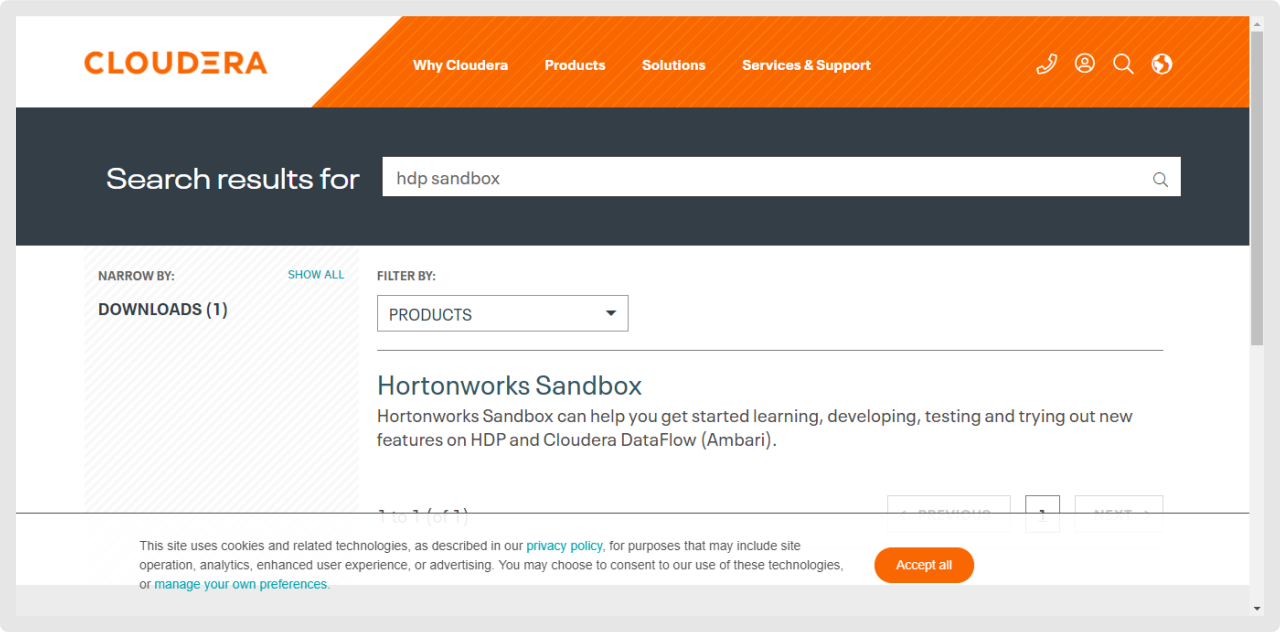
Course Code: Big Data Programming

Imran Ahmad Mughal

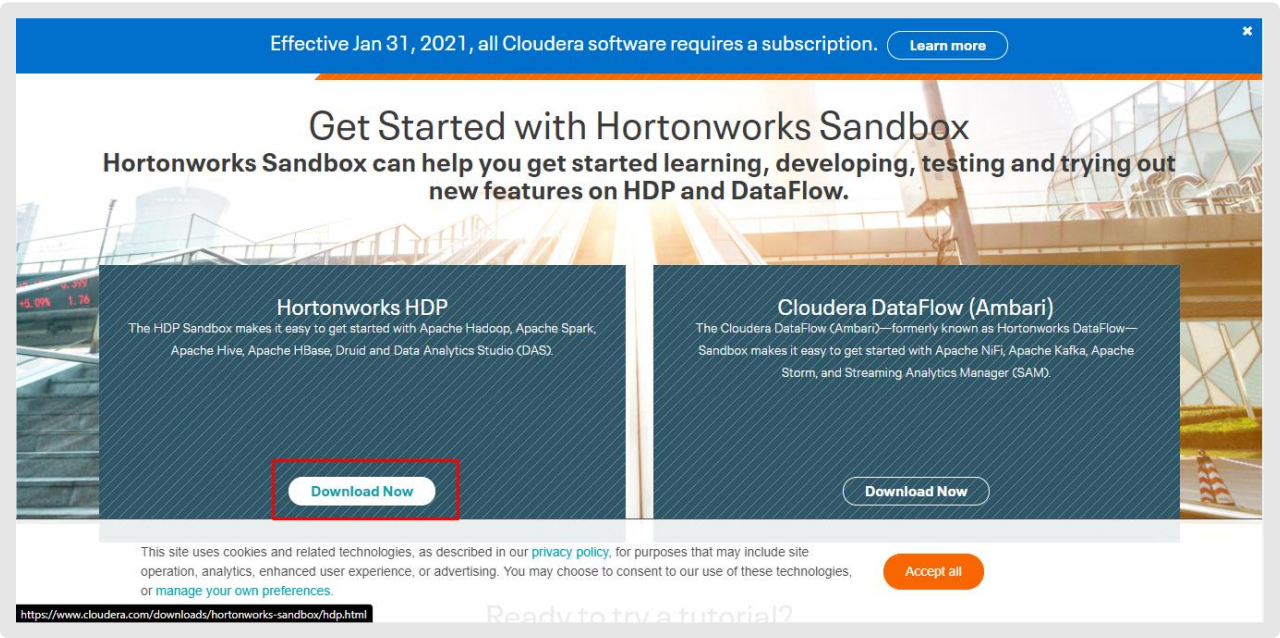
20th October, 2021

Installation of Hortonworks Sandbox

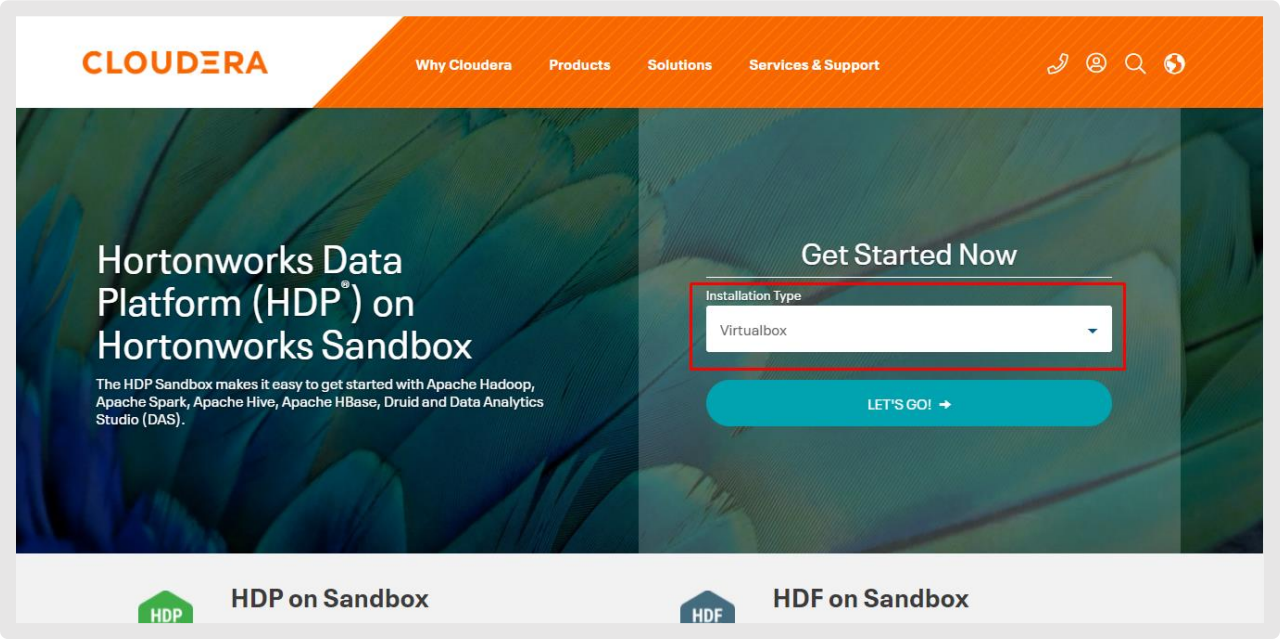
Step-1:



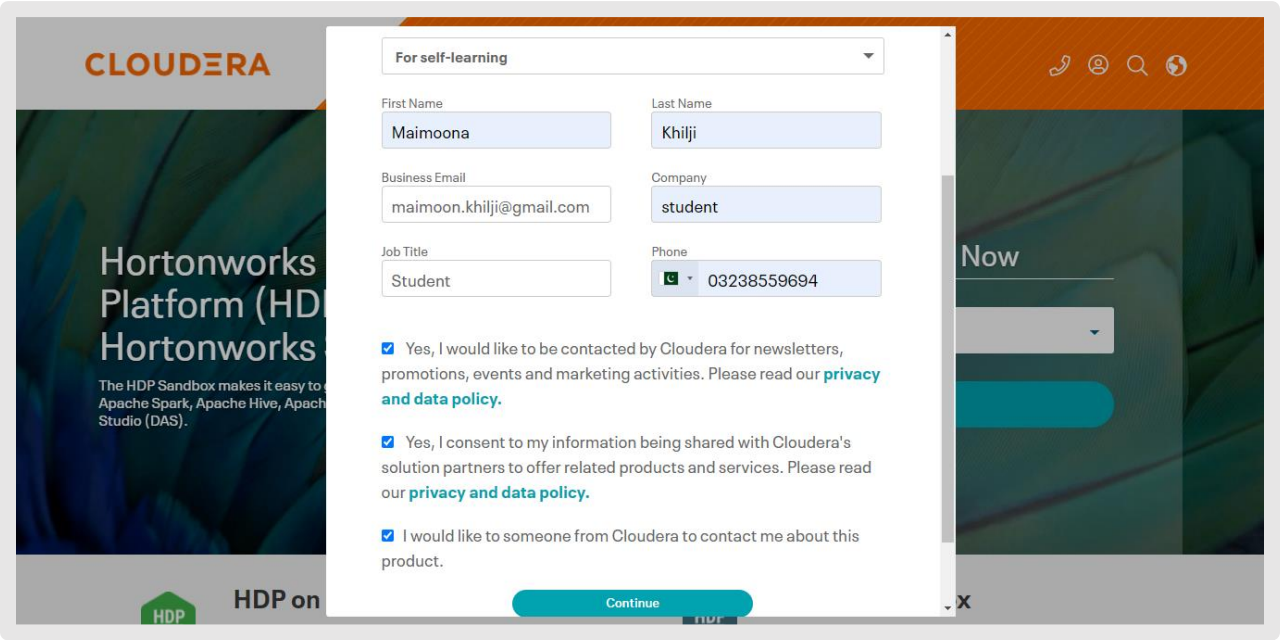
Step-2:



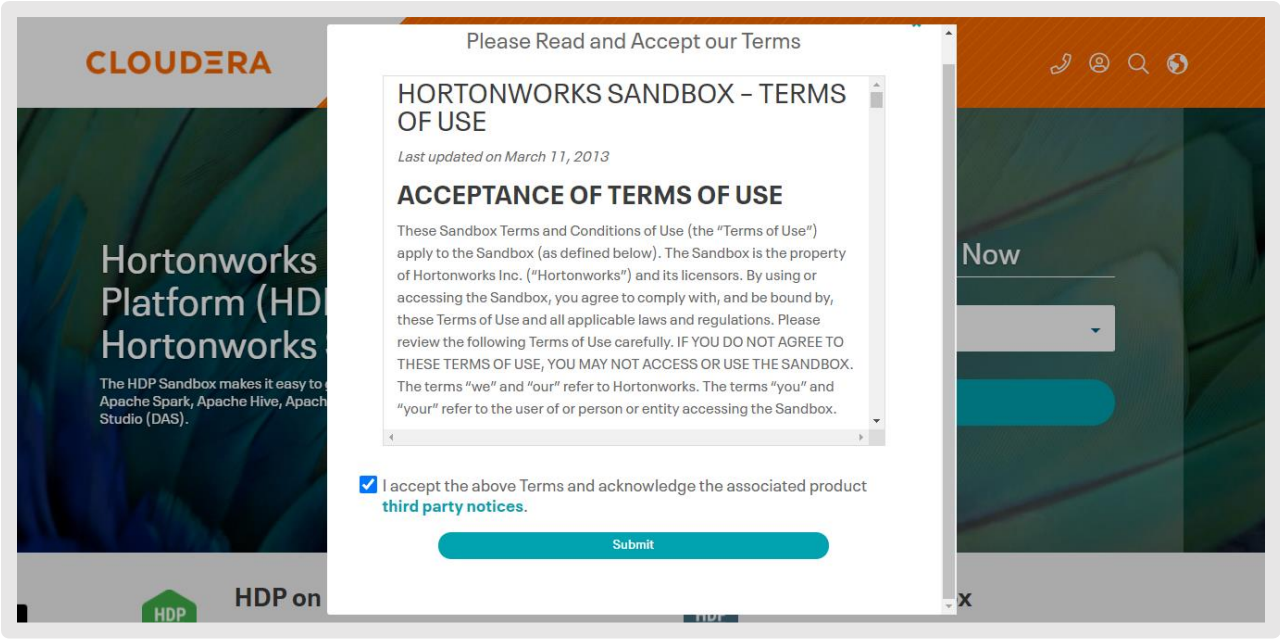
Step-3:


CLOUDERA
Why Cloudera
Products
Solutions
Services & Support
Hortonworks Data Platform (HDP®) on Hortonworks Sandbox
The HDP Sandbox makes it easy to get started with Apache Hadoop, Apache Spark, Apache Hive, Apache HBase, Druid and Data Analytics Studio (DAS).
Get Started Now
Installation Type
Virtualbox
LET'S GO!
HDP on Sandbox
HDF on Sandbox

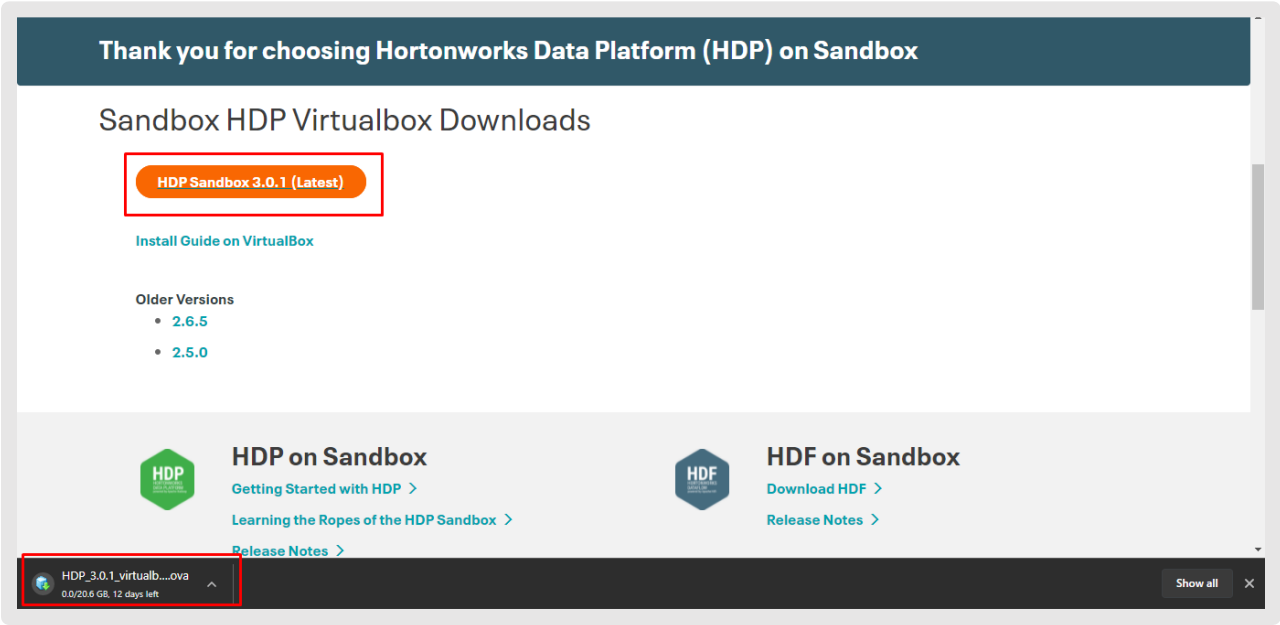
Step-4:


CLOUDERA
For self-learning
First Name
Maimoona
Last Name
Khilji
Business Email
maimoon.khilji@gmail.com
Company
student
Job Title
Student
Phone
03238559694
Yes, I would like to be contacted by Cloudera for newsletters, promotions, events and marketing activities. Please read our privacy and data policy.
Yes, I consent to my information being shared with Cloudera's solution partners to offer related products and services. Please read our privacy and data policy.
I would like to someone from Cloudera to contact me about this product.
Continue

Step-5:

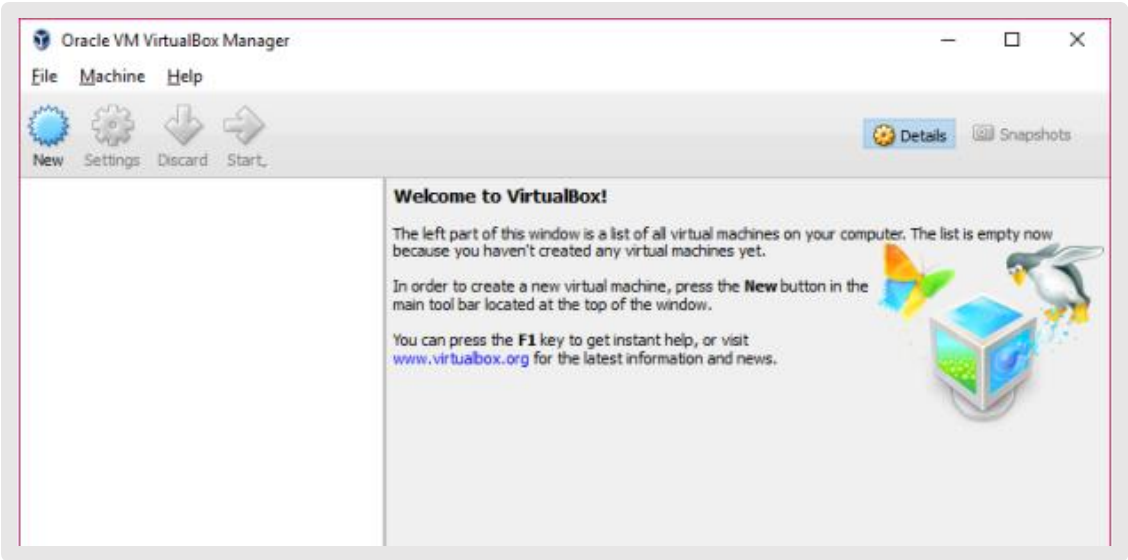

CLOUDERA
Please Read and Accept our Terms
HORTONWORKS SANDBOX – TERMS OF USE
Last updated on March 11, 2013
ACCEPTANCE OF TERMS OF USE
These Sandbox Terms and Conditions of Use (the "Terms of Use") apply to the Sandbox (as defined below). The Sandbox is the property of Hortonworks Inc. ("Hortonworks") and its licensors. By using or accessing the Sandbox, you agree to comply with, and be bound by, these Terms of Use and all applicable laws and regulations. Please review the following Terms of Use carefully. IF YOU DO NOT AGREE TO THESE TERMS OF USE, YOU MAY NOT ACCESS OR USE THE SANDBOX. The terms "we" and "our" refer to Hortonworks. The terms "you" and "your" refer to the user of or person or entity accessing the Sandbox.
I accept the above Terms and acknowledge the associated product third party notices.
Submit
Now
HDP on

Step-6:

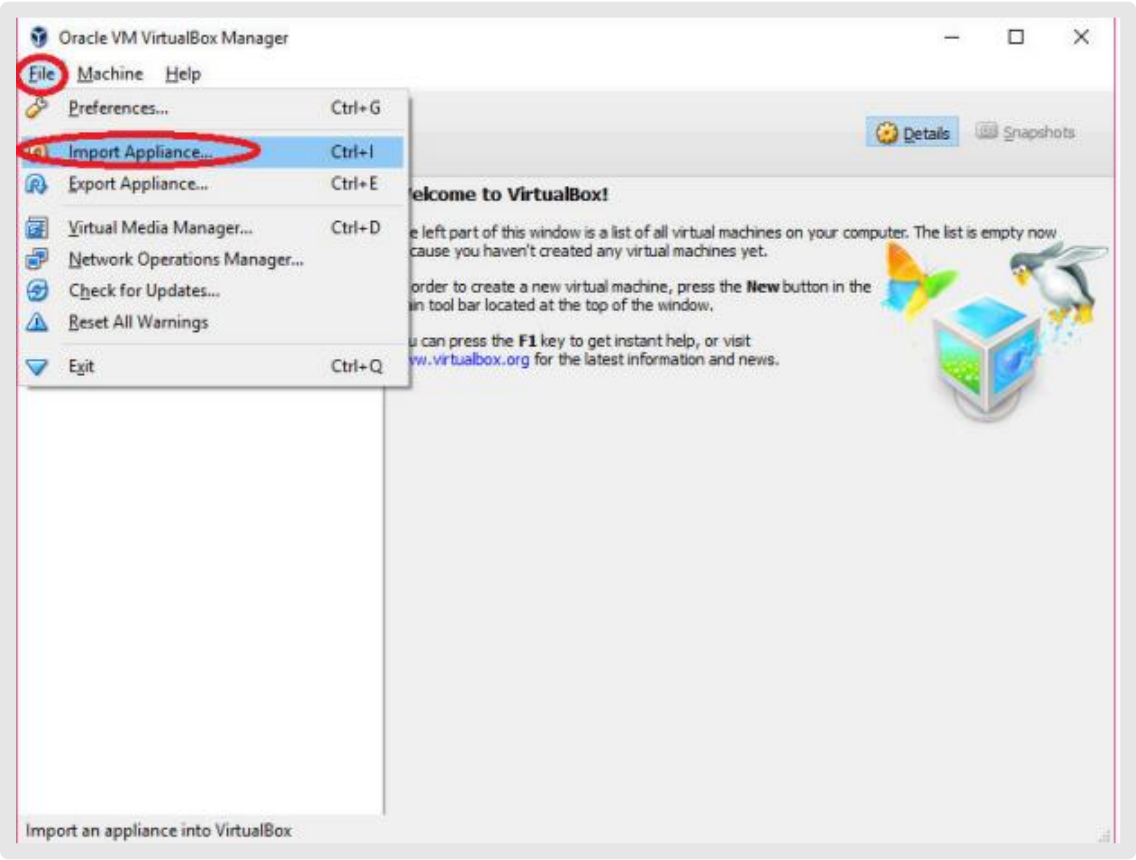

Thank you for choosing Hortonworks Data Platform (HDP) on Sandbox
Sandbox HDP Virtualbox Downloads
HDP Sandbox 3.0.1 (Latest)
Install Guide on VirtualBox
Older Versions
2.6.5
2.5.0
HDP on Sandbox
Getting Started with HDP
Learning the Ropes of the HDP Sandbox
Release Notes
HDF on Sandbox
Download HDF
Release Notes
HDP_3.0.1_virtualb....ova
0.0/20.6 GB, 12 days left
Show all

Step-7:

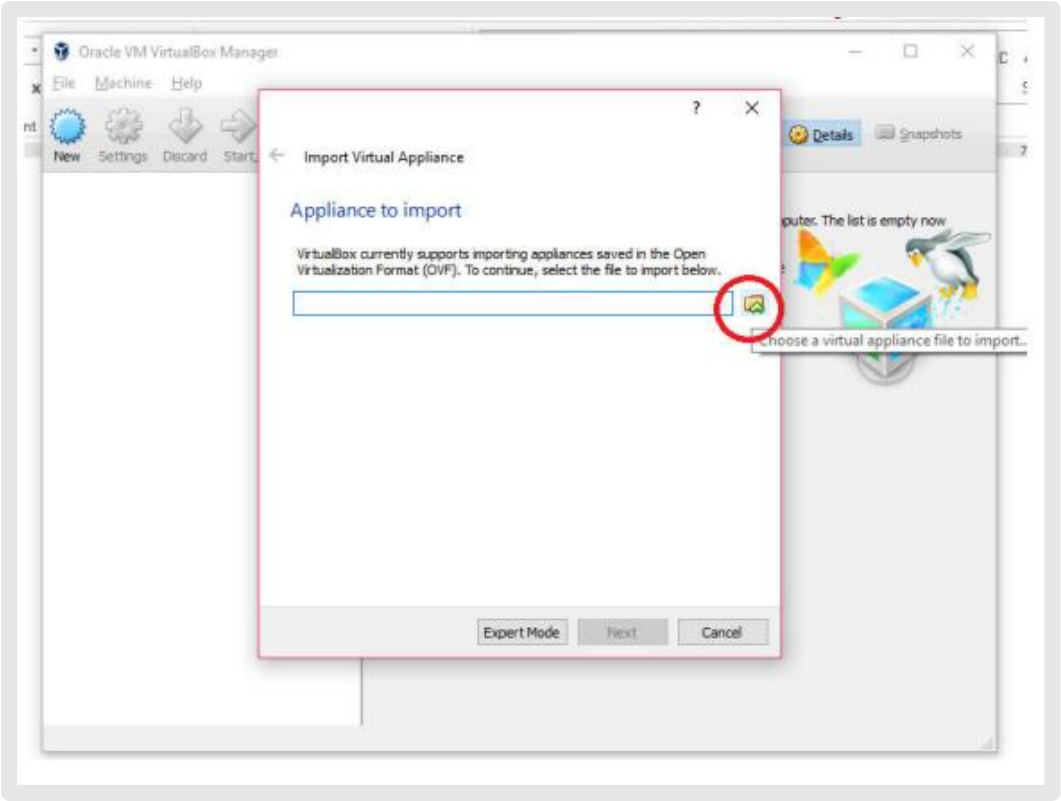
After downloading the sandbox, install it on Oracle Virtual Box.



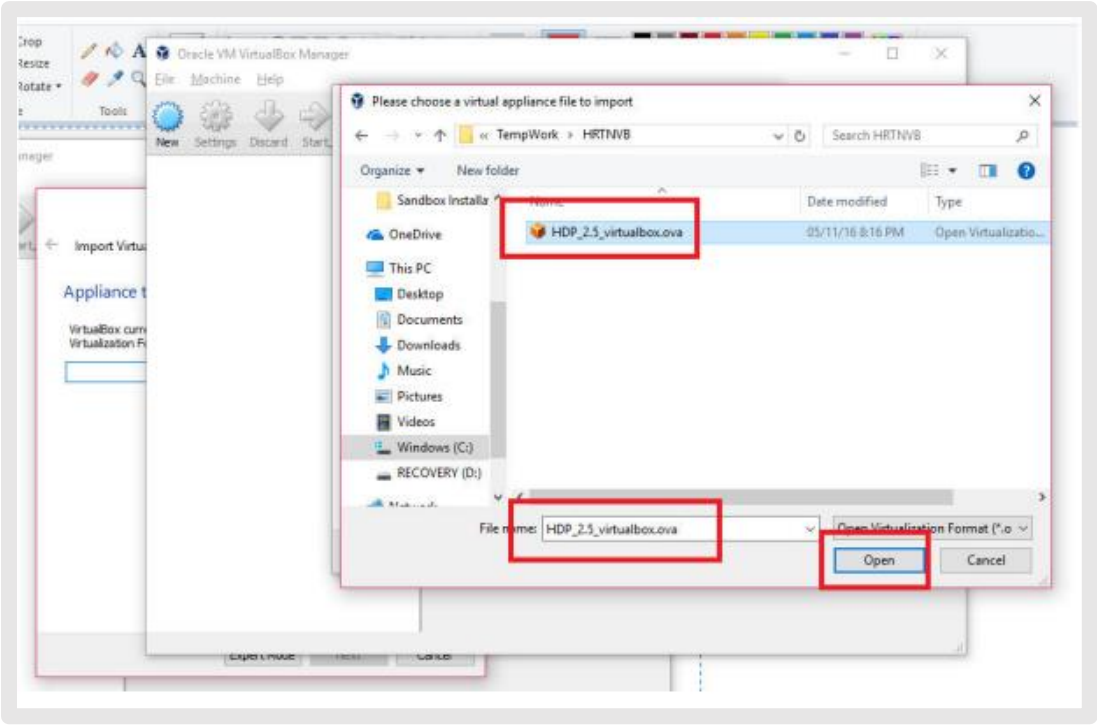
Step-8:



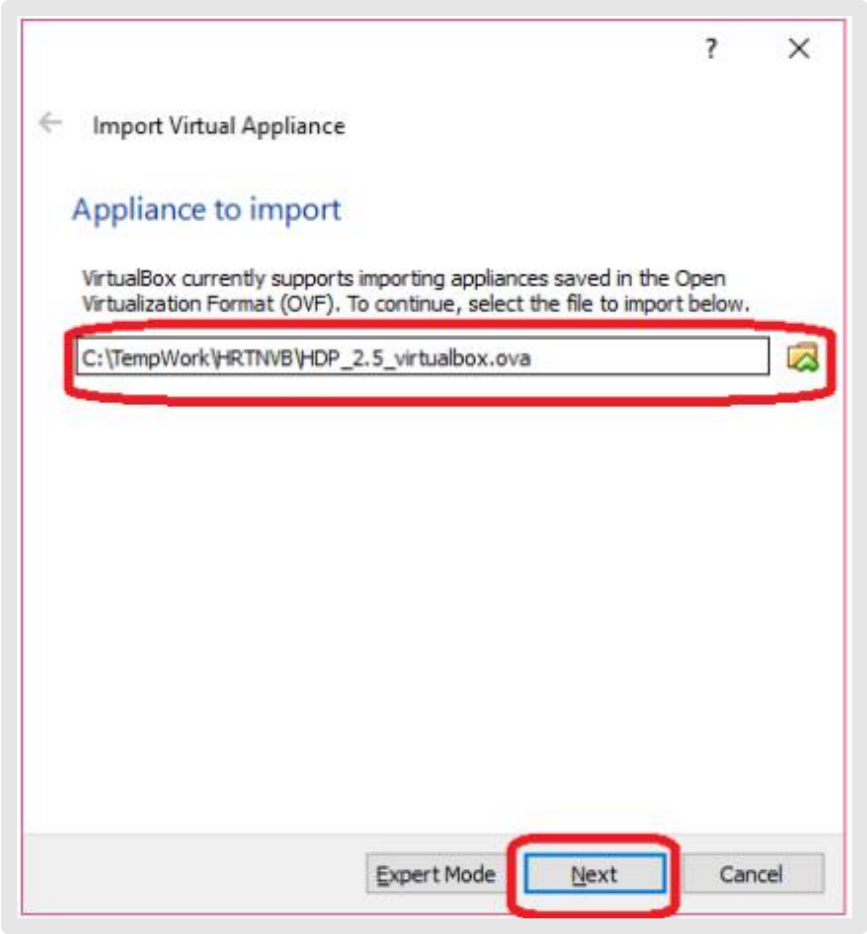
Step-9:



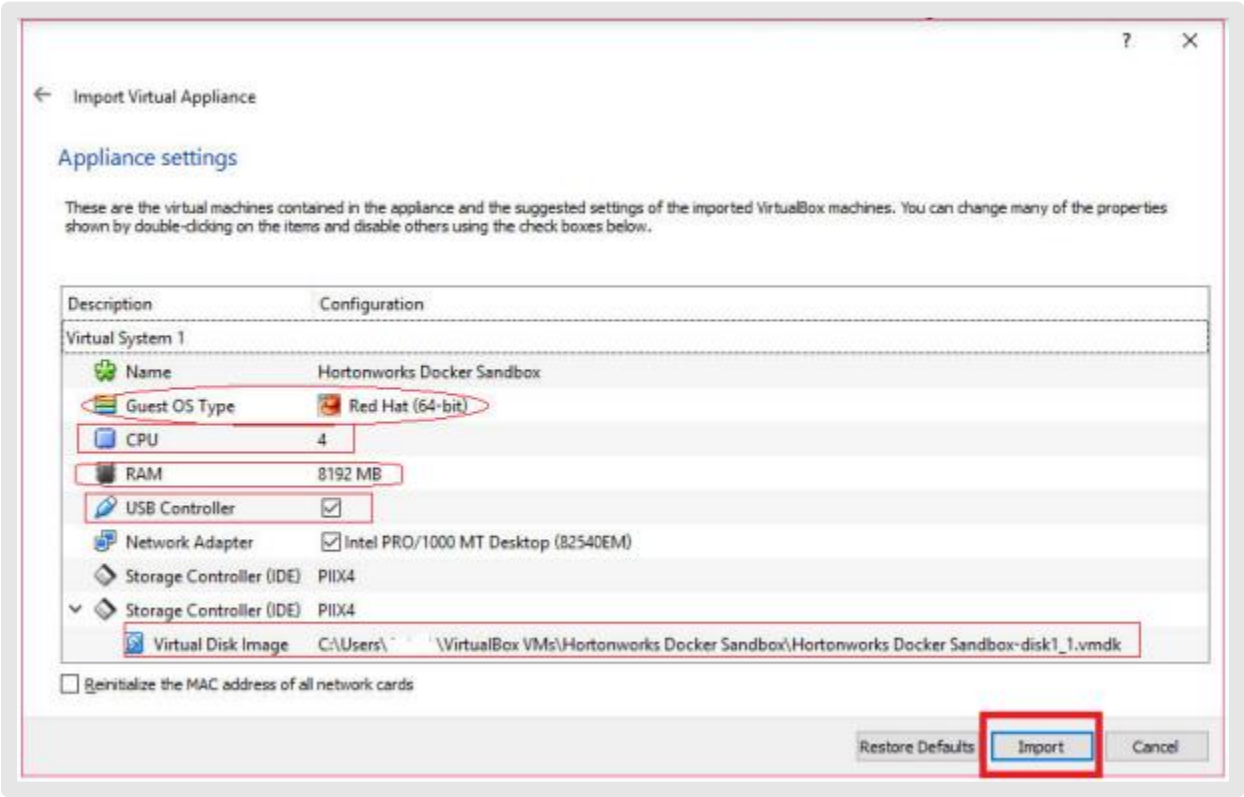
Step-10:



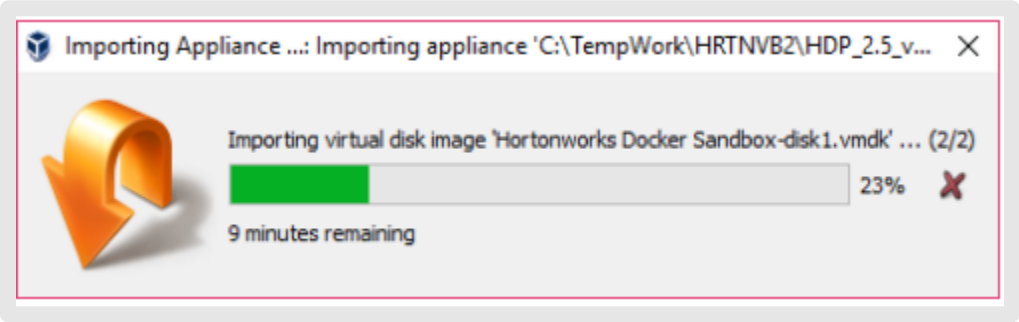
Step-11:



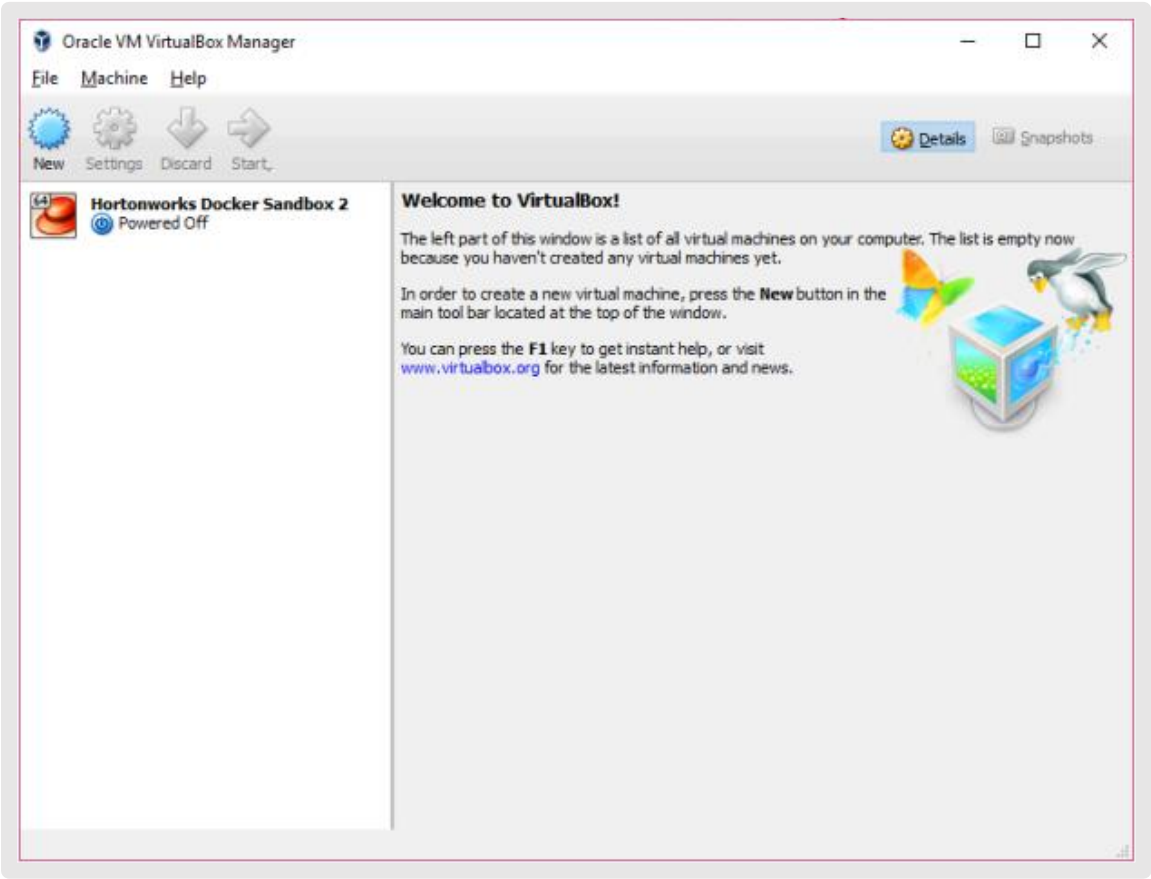
Step-12:



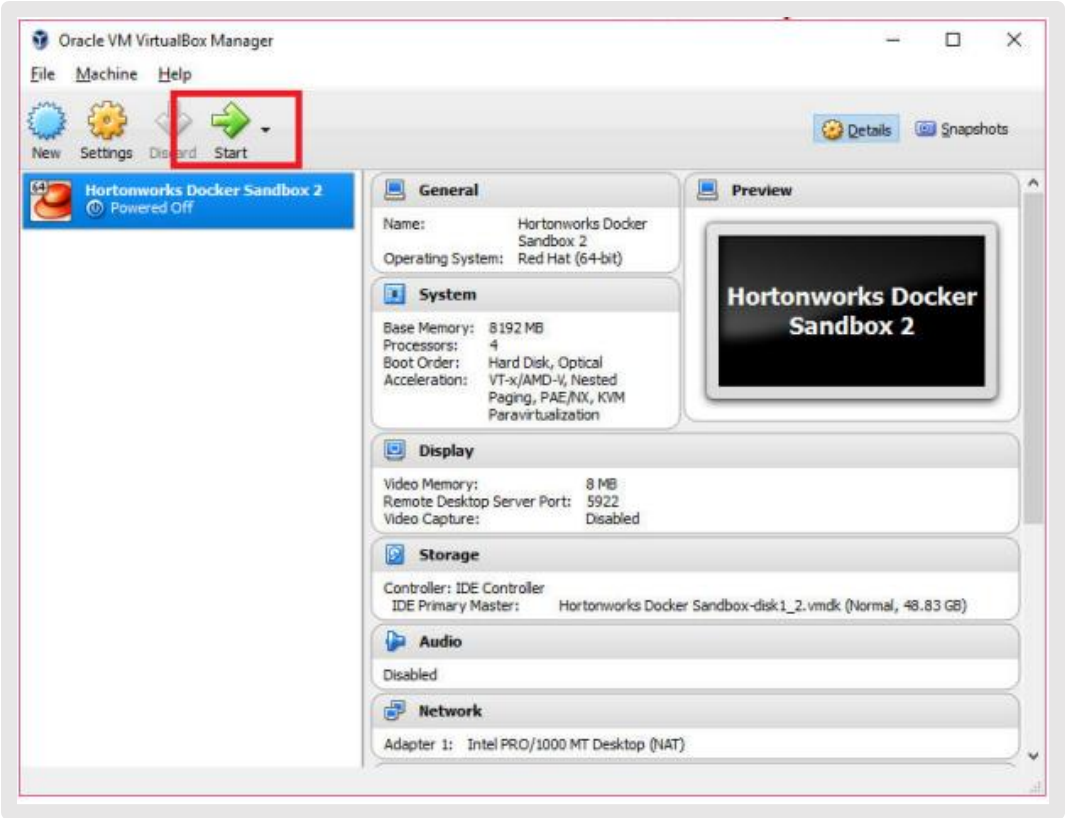
Step-13:



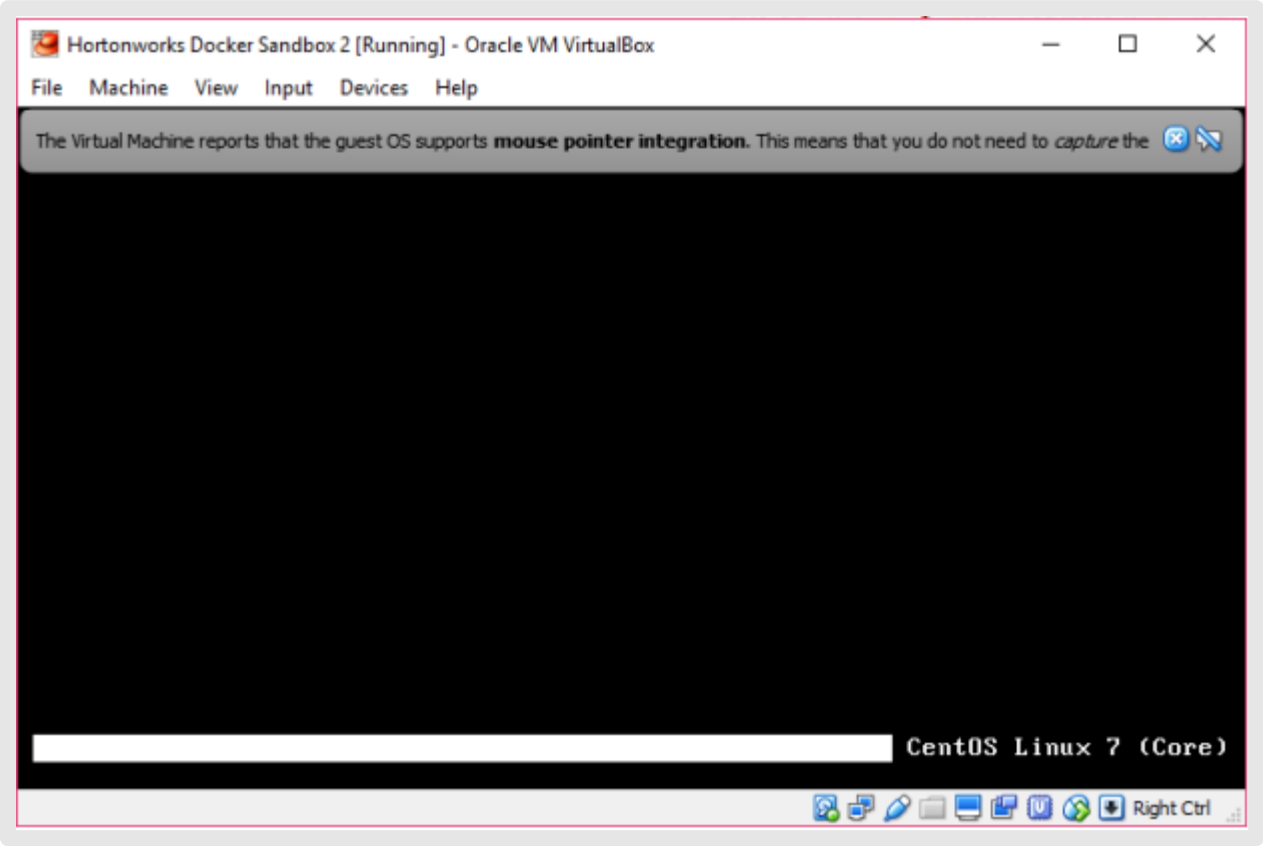
Step-14:



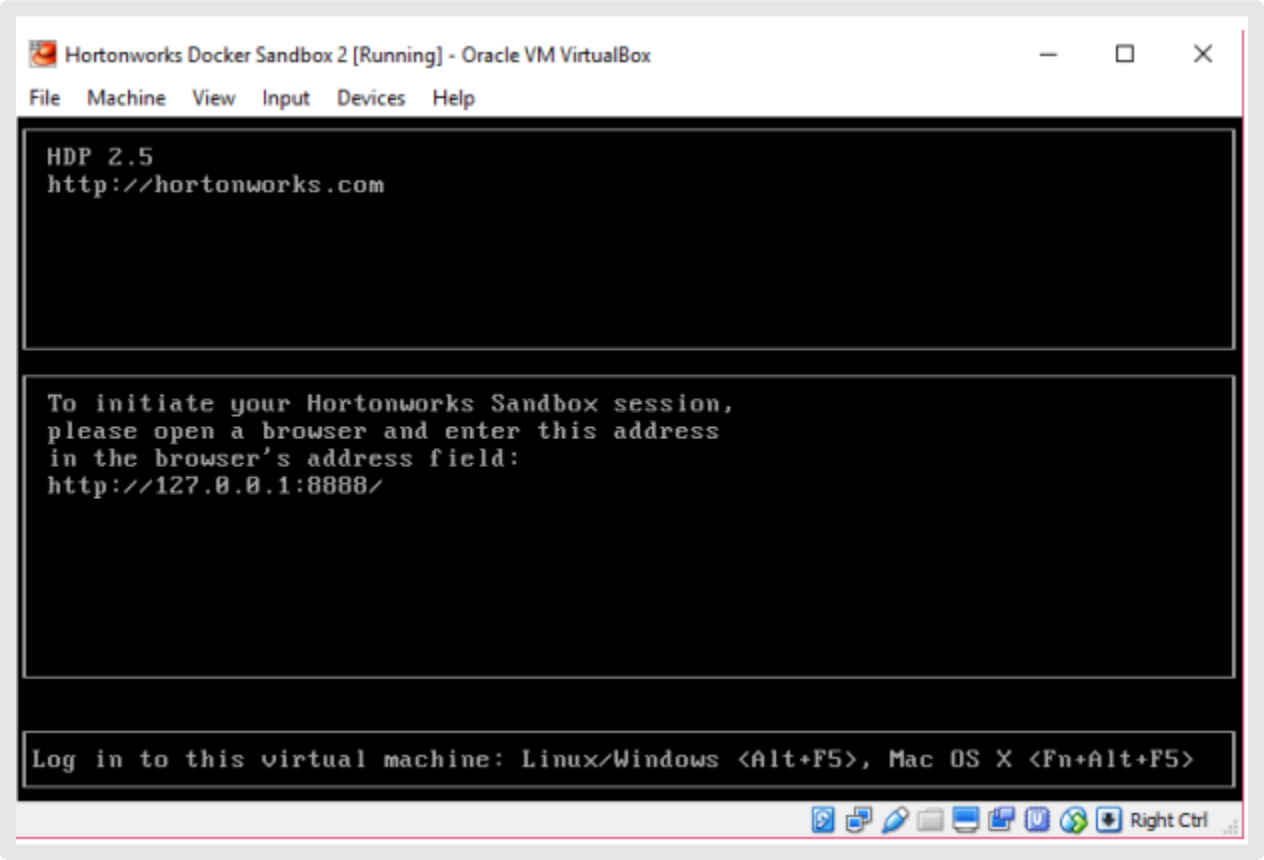
Step-15:



Step-16:

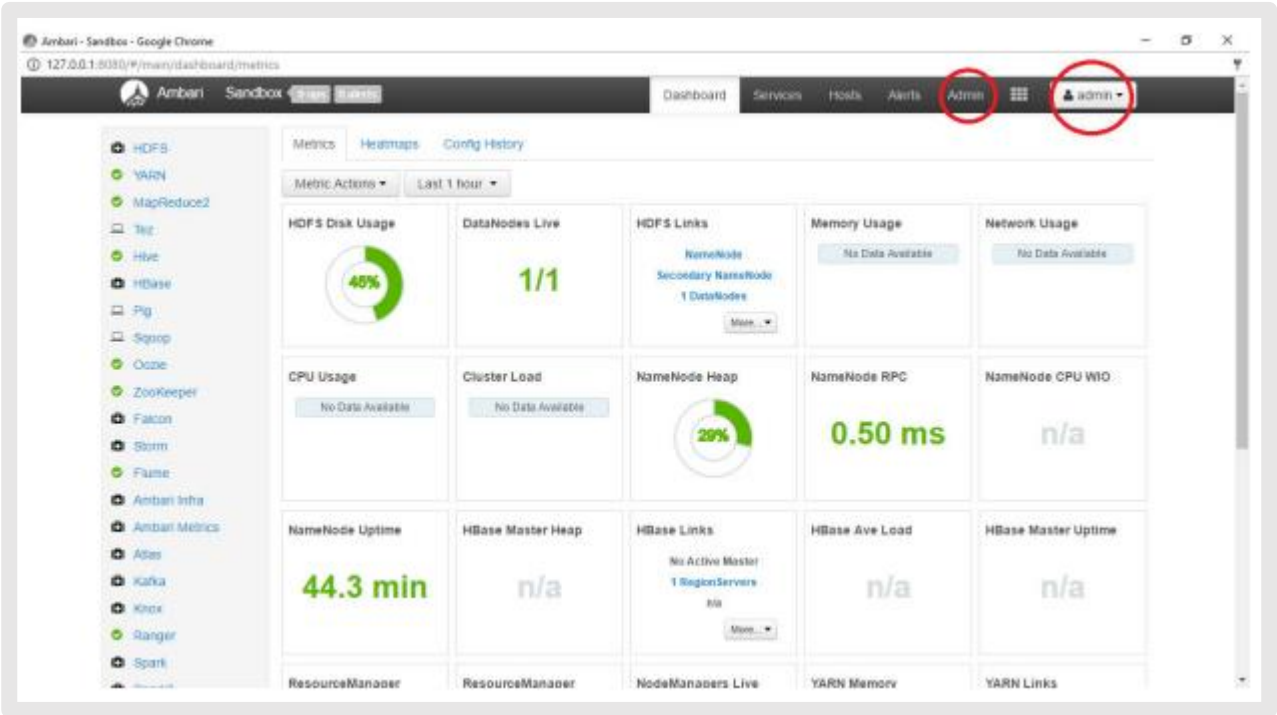


Step-17:



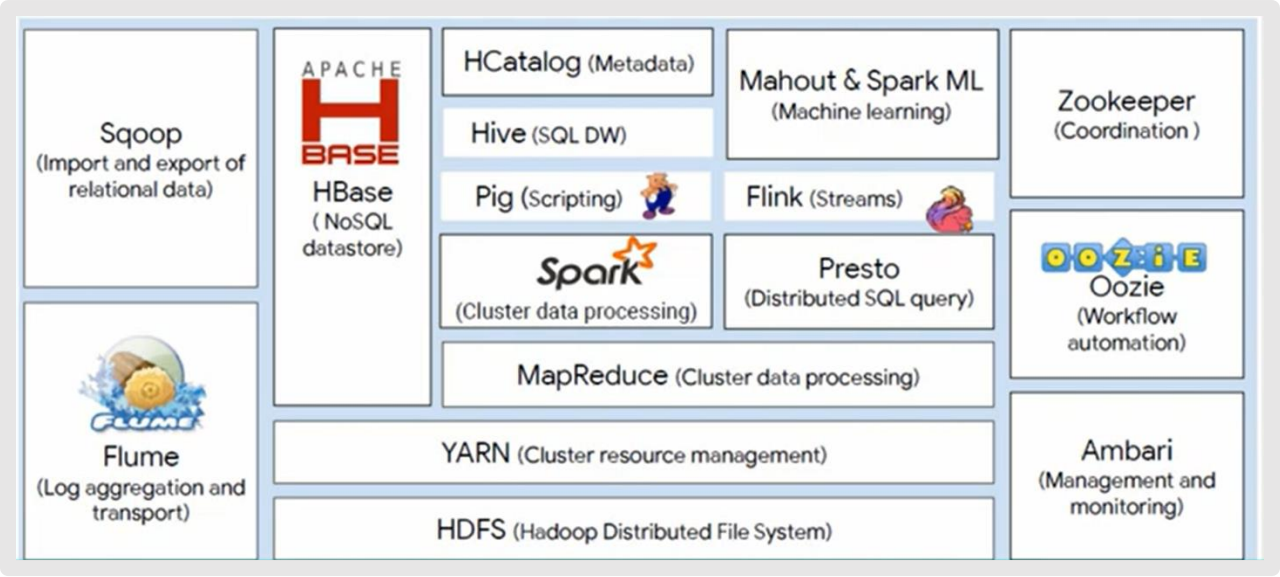
Step-18:

Now open the browser and start working on ambary after signing in.



Hadoop Ecosystem

Hadoop ecosystem is a platform that provides different services to deal big data. It consists of storage, processing, management, access /injection of big data and many more services.



The utilities of Hadoop ecosystem can be explained in basic three layers:

1. Data injection

Modules used for data injection are:

- **Flume:**
  Flume is used for data collection from different resources and then transfer it to one centralized repository. It is used to capture stream of moving data.

- **Sqoop:**
  Sqoop is used for transferring data between relational database and Hadoop.

2. Data storage and processing

Modules used for data storage, management and processing are:

- **HDFS:**
  HDFS is a Hadoop distributed file system. It is a primary data storage system. It is used to scale single cluster to hundreds or thousands of nodes.

- **YARN:**
  YARN stands for Yet Another Resource Negotiator. It is used to manage the resources and job scheduling in Hadoop.

- **Apache HBase:**
  Apache HBase is NoSQL, distributed big data store. It gives the random real-time access to petabytes of data.

- **Apache spark:**

  Apache spark provides an interface for programming the entire clusters.

  (Cluster Data Processing)

- **HCatalog:**
  HCatalog is table storage management tool. It enables users with different data processing tools to easily write data onto a grid.

- **Hive:**
  Hive facilitates users to alter, read, and manage the petabytes of data using SQL.

- **Pig:**
  Pig is a high-level scripting language that enables users to write complex data transformations without knowing high level languages (Like Java).

- **Mahout & Spark ML:**
  Mahout & Spark ML is used for creating scalable machine learning algorithms.

- **Flink:**
  Flink is a stream processing engine which is faster then Spark and Hadoop on the basis of Speed.

- **Presto:**
  Presto is a distributed SQL query engine designed for fast, interactive queries on data on HDFS.

- **MapReduce:**
  MapReduce is used for writing applications that can process huge amount of data on large clusers.

3. Interface

Modules used for interface are:

- **Zookeeper:**
  Zookeeper provides a centralized service for providing configuration information, naming, synchronization and group services over large clusters in distributed systems.

- **Oozie:**
  Oozie is a java web application used to schedule Hadoop jobs. It combines multiple jobs sequentially into one logical unit of work.

- **Ambari:**
  Ambari is an administration tool that is used to manage, monitor the health of Hadoop clusters.

_____