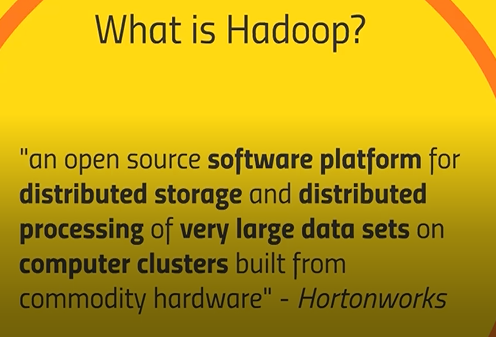
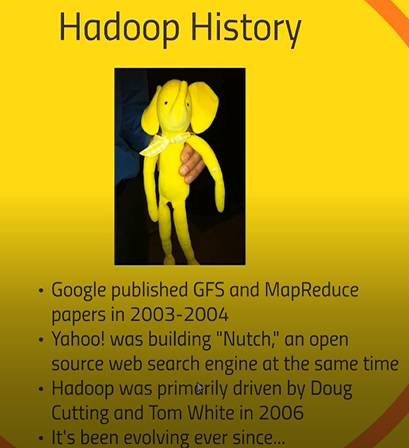
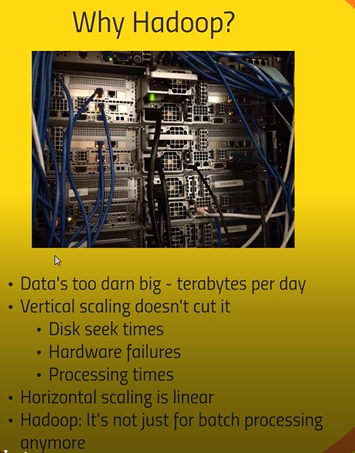
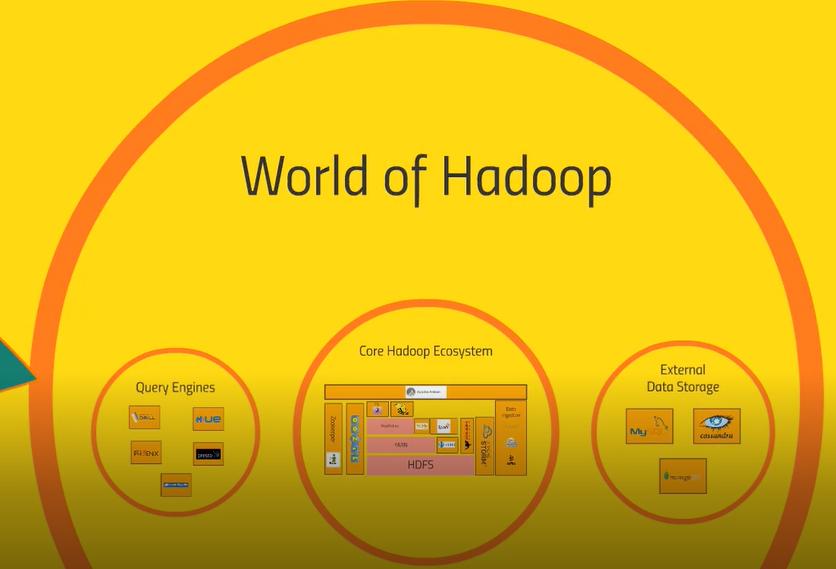
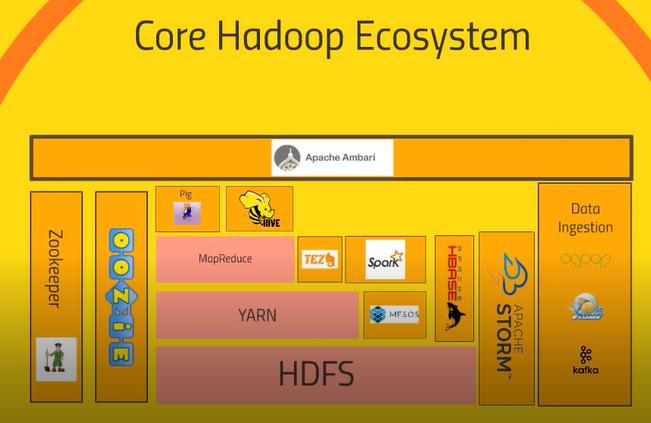
­HADOOP INSTALLATION ON WINDOWS



­







Let’s understand the relationship of all these Hadoop ecosystem-



1. Base of everything is Hadoop distributed file system. It distributed the storage of big data across the cluster of computers.
2. It also very good for fault tolerant system & it is distributed storage of the hadoop



1. Sitting on top of hdfs is YARN (YET ANOTHER RESOURCE NEGOTIATOR)
2. Data storage & Data processing is also part of the Hadoop. data storage take care of hdfs whereas Data process take care by YARN
3. YARN is the system that manages the resources on computing cluster. It decide which task runs when, what nods are available for extra work & which are not available
4. It is the heart for keep the cluster working well. Based on RESOURCE NEGOTIATOR this we can build one interesting application on top of that which is called MAP REDUCE



1. Map reduce is a programming model or stack of layers process your data entire cluster
2. You need to write down map & reduce for both different script
3. It has ability to transform the data in parallel across entire compute cluster in very efficient manner & reducer aggregate the data together
4. Map reduce & yarn kind of same things in the Hadoop. But now they got split up because other application are build on top of yarn



1. If you don’t want to write the code in java, python or map reduce & if you are more familiar with scripting language, sql style syntax so pig is very good tool
2. Pig is very high level programming API that allows to write simple script look like sql , club with query and get the solution for more complex problems
3. Pig will run on top of the map reduce which turns go through yarn and hdfs to actually to process the data



1. **HIVE** also sit on top of map-reduce & solves similar problem what pig does & it looks more directly sql database
2. so **HIVE** is taking sql query & making distributed the data sitting on file system somewhere look like sql databases
3. This is the database connect to odbc & execute sql query from Hadoop cluster
4. If you are familiar is sql then HIVE would be the best tool for this



1. Apache ambari it sits in top of everything. It gives view of cluster visualize what running on cluster & what system using how much sources and it also give view to execute some hive query or import databases into hive & executes pig something like that
2. Apache ambari actually use by hortonwork



1. Mesos is part of Hadoop proper & it is alternative to yarn & meso also act as a associator negotiator



1. You can write your spark script using python, scala, r, sql , java
2. This is very fast and powerful technology
3. It handles sql query, handle machine learning problem
4. It also handles strumming data



1. It also act as similar like spark
2. Tez produce more optimal plan to produce more query
3. Tez is conjunction with hive to accelerate to efficient answer from the cluster



1. Hbase is the way of exposing the data on your cluster to transactional platform
2. HBASE is also called as no sql db & it is colonal data store
3. It is really very fast database which is meant for large transaction & it is appropriate for hitting for web applications, website for transaction
4. It can expose the data that store on your cluster & that data may be transpose some way by spark, MapReduce or something else provide very fast way of exposing result to other system



1. Storm will process the streaming data
2. That data came from sensors, weblogs and also spark streaming also doing the same thing
3. It is not be an batch thing



1. OOZIE this helps to scheduling the jobs
2. You have task that on Hadoop cluster it involves many different steps
3. Oozie is helping to scheduling to all the job & it runs on lively In consistency basis



1. This is one of the technologies for coordinate everything on cluster
2. It will keep track of which nods will up and which nodes will down & very reliable way to keeping track of share states across the cluster for different application even though when nodes goes down.
3. Zookeeper will track who is the current master node is & who is up and who is down



1. How the data getting into your cluster and hdfs from various external sources



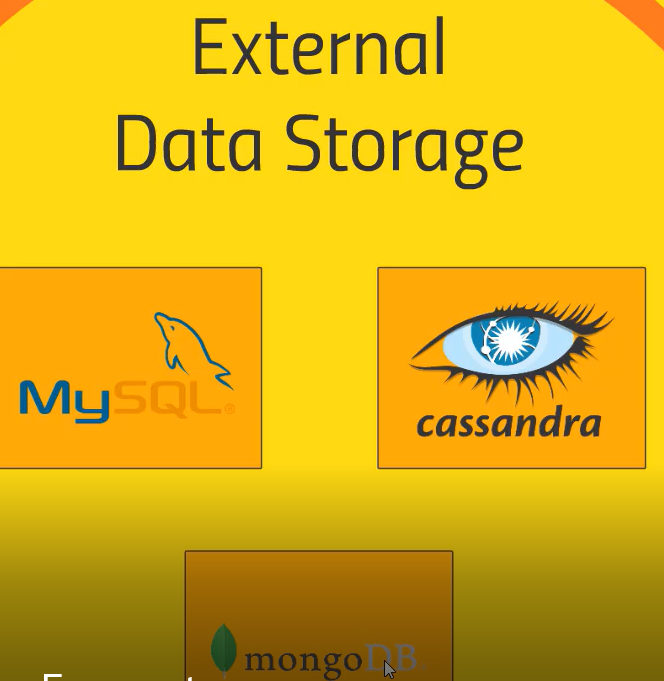
1. Sqoop is the way of actually ting your Hadoop database into relational database anything that talk about odbc or jdbc transform from sqoop into hdfs file system
2. Basically sqoop is the connector between Hadoop and legacy datbases



1. Transforming weblogs at very large scale and very reliability to your cluster.
2. Flume can listen from weblogs coming from websites in realtime & published into cluster processing by storm or spark streaming



1. Kafka will collect the data from any sort from cluster of pcs , cluster of webserver and broadcast that into Hadoop cluster well.
2. This will solve the problem of data ingestion



1. H-base also come under this category.
2. Expose the data for real time usage and these are data bases used in H-base
3. My-sql, Cassandra and mongo db all external databases are might be integrated with your cluster



1. Also there are several engine sitting on top of the cluster



1. Apache drill it actually allow to sql query work across wild range of no sql databases
2. It can actually talk to HBASE database & Cassandra & monogdb as well and combine the result well & write the query to across all the datastores & bring back all together.



1. Hue also query interactively creating querying work with hive and hbase. For cloudera it takes the role of ambari & it sits on top of everything.
2. 



1. Phonix & presto also kind of sql style query across the entire range data storage technology



1. Apache zeppelin is the kind of notebook approach of UI which interactive with cluster

SKIP CLOUDERA SET UP ::: (( pg-6 to pg-13 ))

We have discussed following points: -

- How to install Hadoop in windows 10

- How to install Hadoop on VirtualBox

- How to download cloudera QuickStart vm for VirtualBox

- Big data tutorial

To install the Hadoop, we required to install 3 Applications -

1. 7 ZIP software – extracting the files
2. Virtual Box
3. Cloudera image – Cloudera Quick start VM

**STEPS TO INSTALL THE SOFTWARE: -**

1-Download and install Winzip software from windows which is free

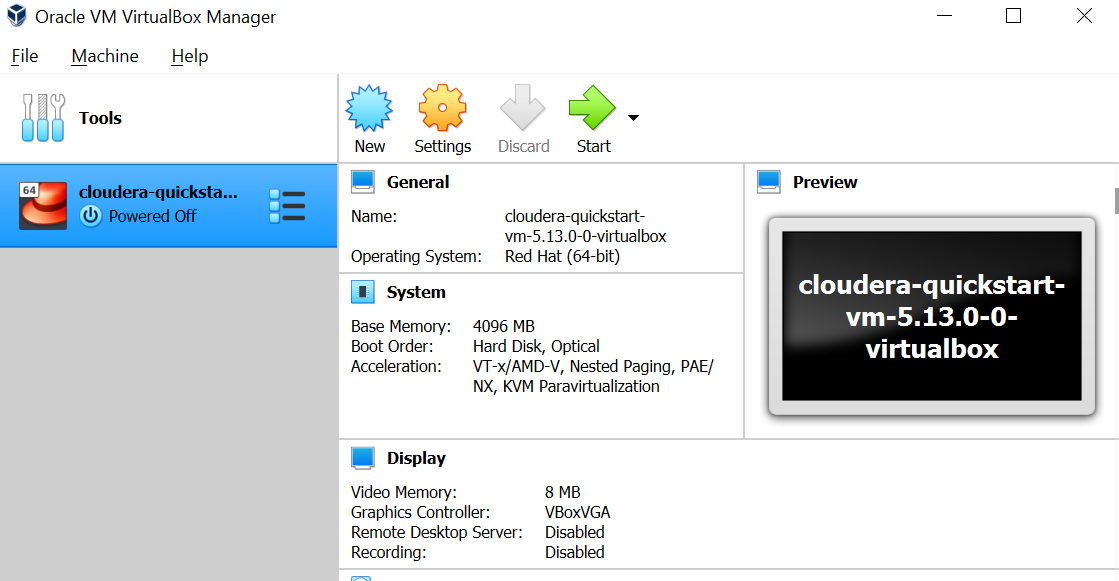
<https://www.winzip.com/en/download/winzip/>

2-Download & install virtual Box for windows –

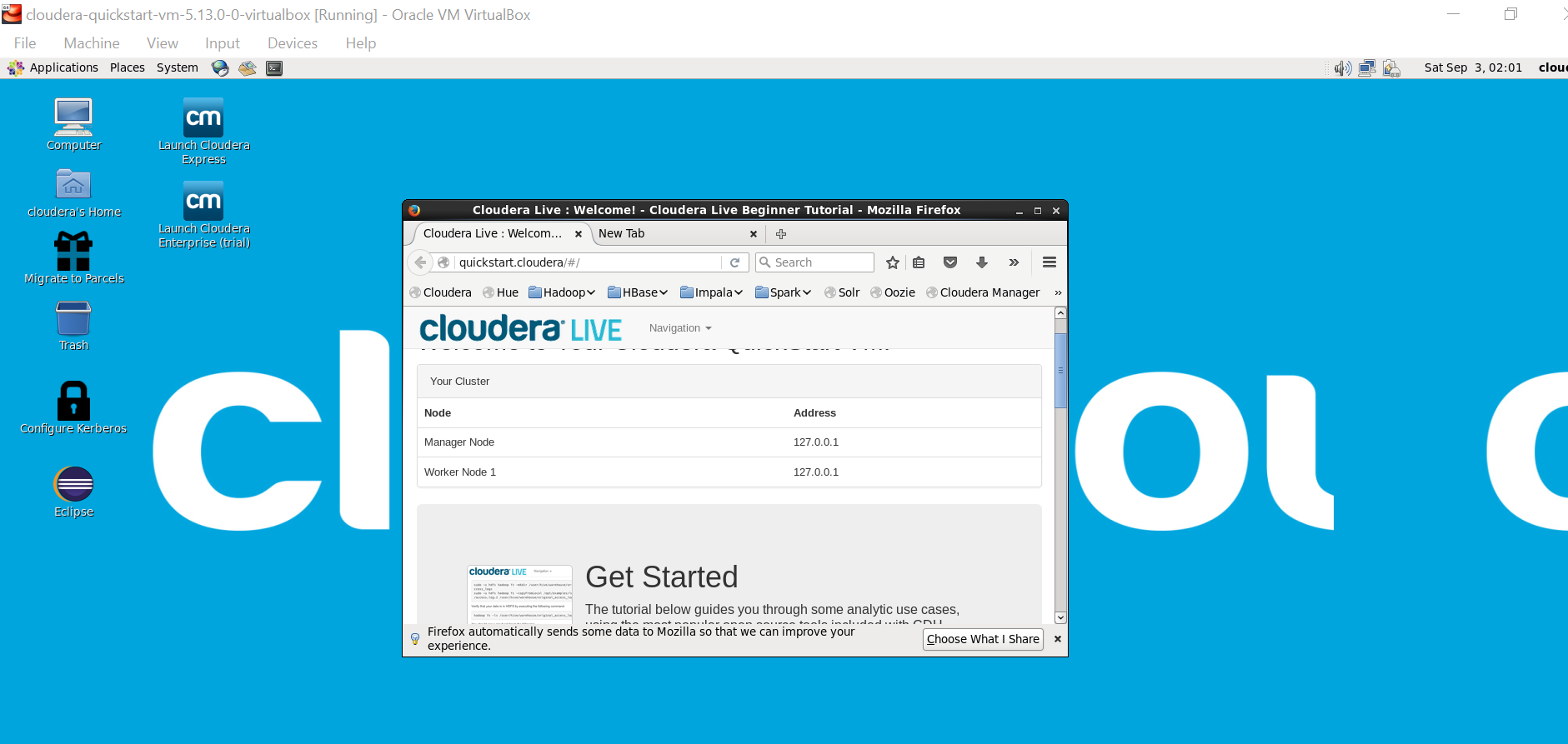
<https://www.virtualbox.org/wiki/Downloads>

3-Download Cloudera for Installation of Hadoop

Open virtual box 🡪 File 🡪 Import Appliance 🡪 Open the Cloudera 🡪 Appliance setting 🡪 Import 🡪 It will take some time for installation 🡪 VM Box 🡪 select on cloudera-quickstart 🡪 Start 🡪 virtual box is start and going to reboot 🡪 this will be installed in red hat operating system

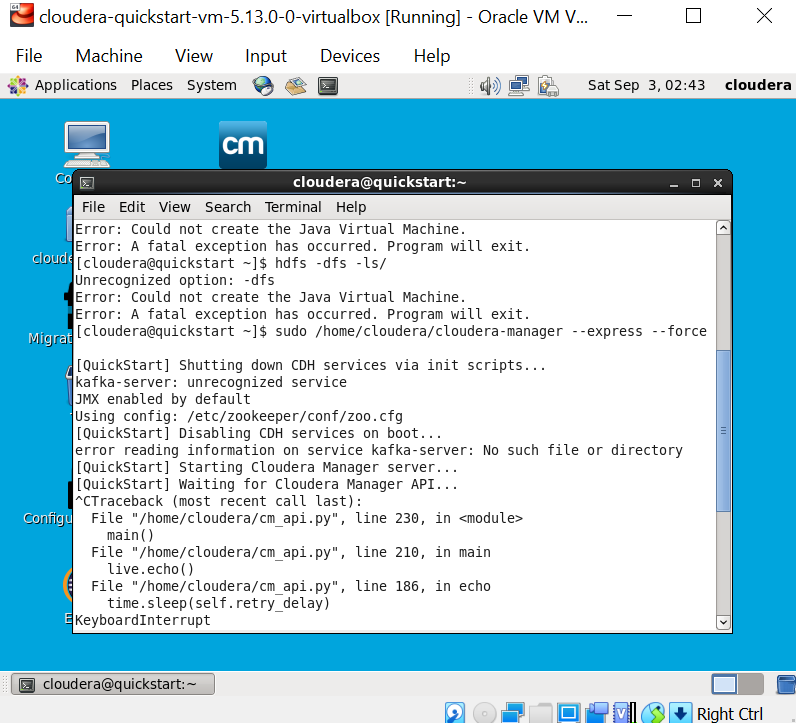


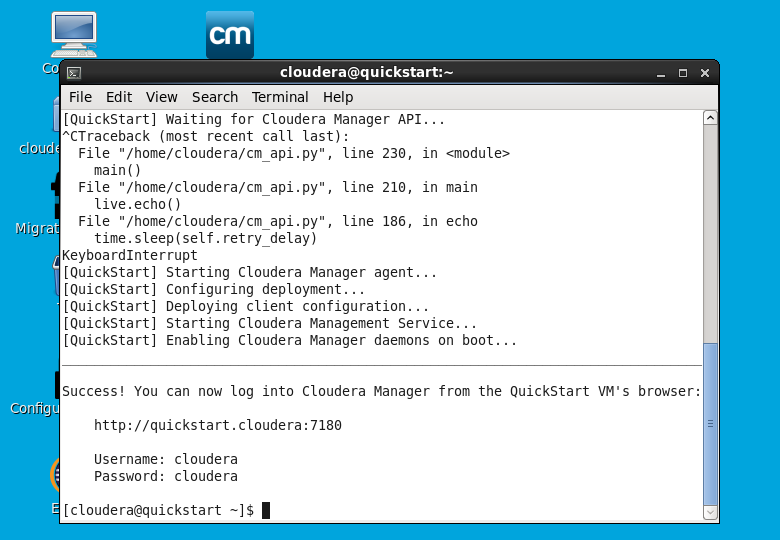
After install the cloud era successfully, you can see the below application screen



Click on cmd prompt 🡪 hostname 🡪 hdfs dfs -ls / 🡪 if you find error – then try this command 🡪

sudo /home/cloudera/cloudera-manager --express –force





Finally, you will generate <http://quickstart.cloudera:7180> || user name & pwd – cloudera

(Cloudera manager is the manage all the Hadoop services)

After successfully installed 🡪 copy the <https://quickstart.cloudera:7180> 🡪 it will ask for username & password 🡪 you are able to see the services are running

**Steps to Install HADOOP 🡪**

1-Download and install Winzip software from windows which is free

<https://www.winzip.com/en/download/winzip/>

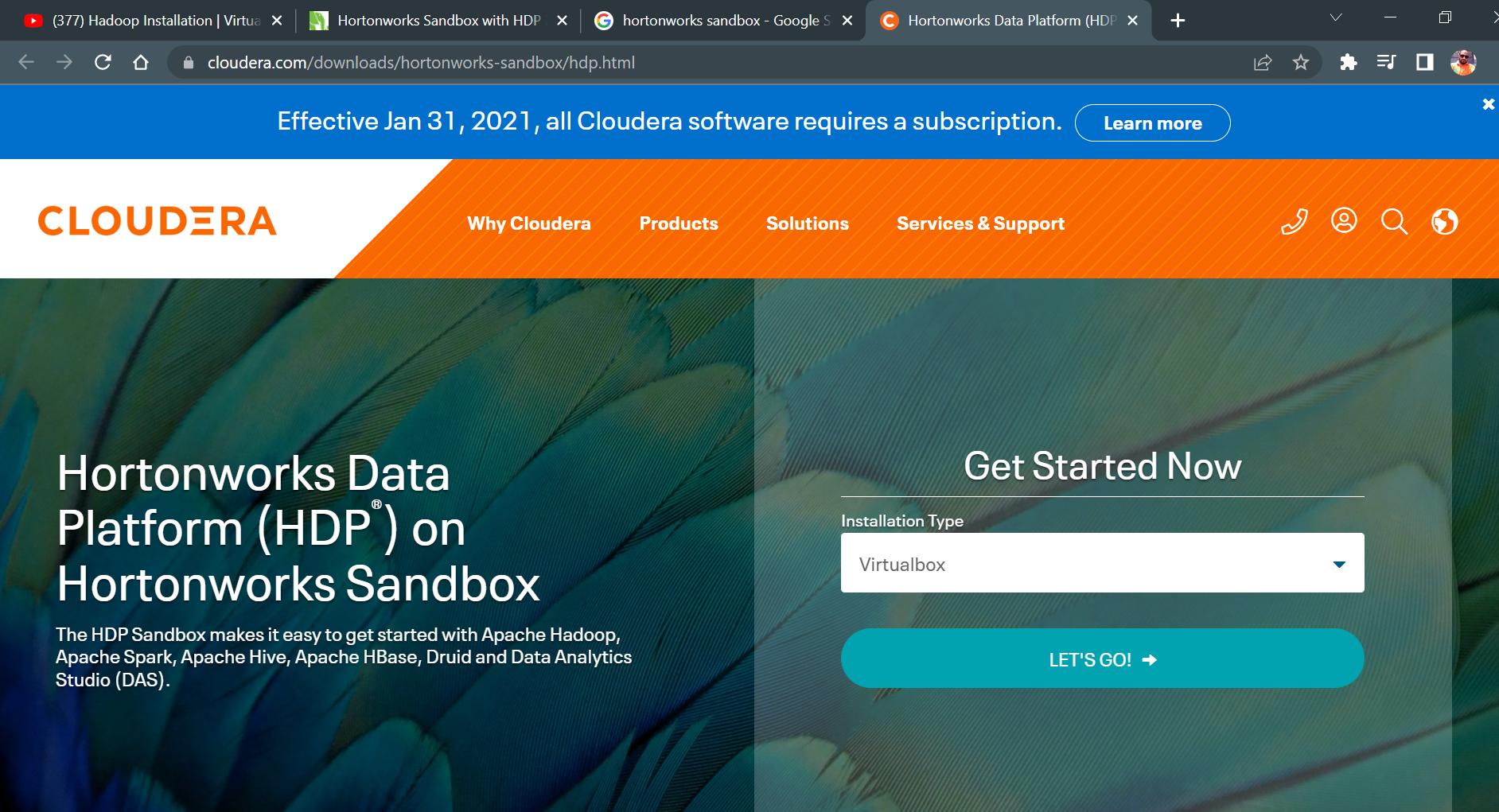
2-Download & install virtual Box for windows –

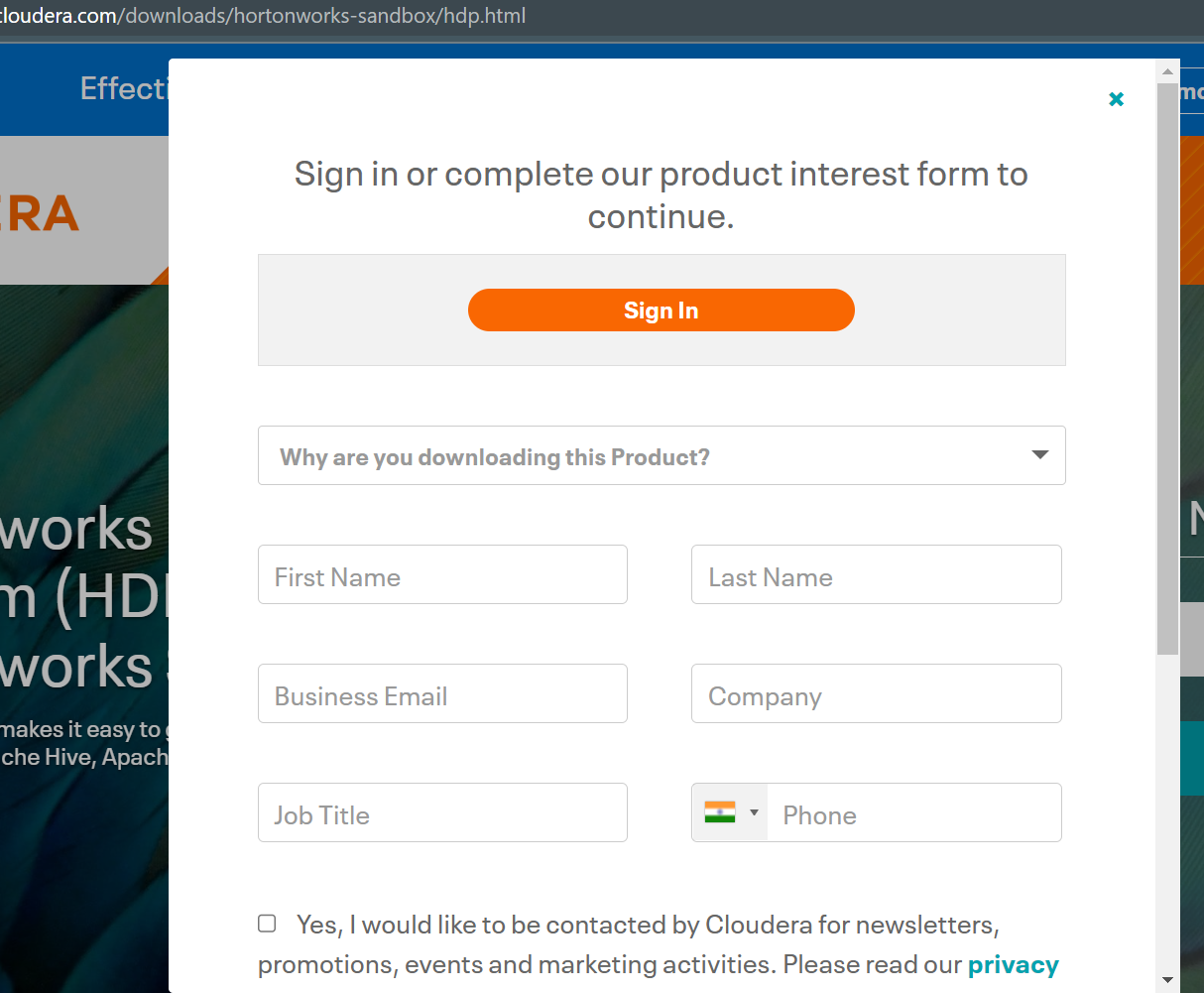
<https://www.virtualbox.org/wiki/Downloads>

3-Downlaod sandbox from below website –

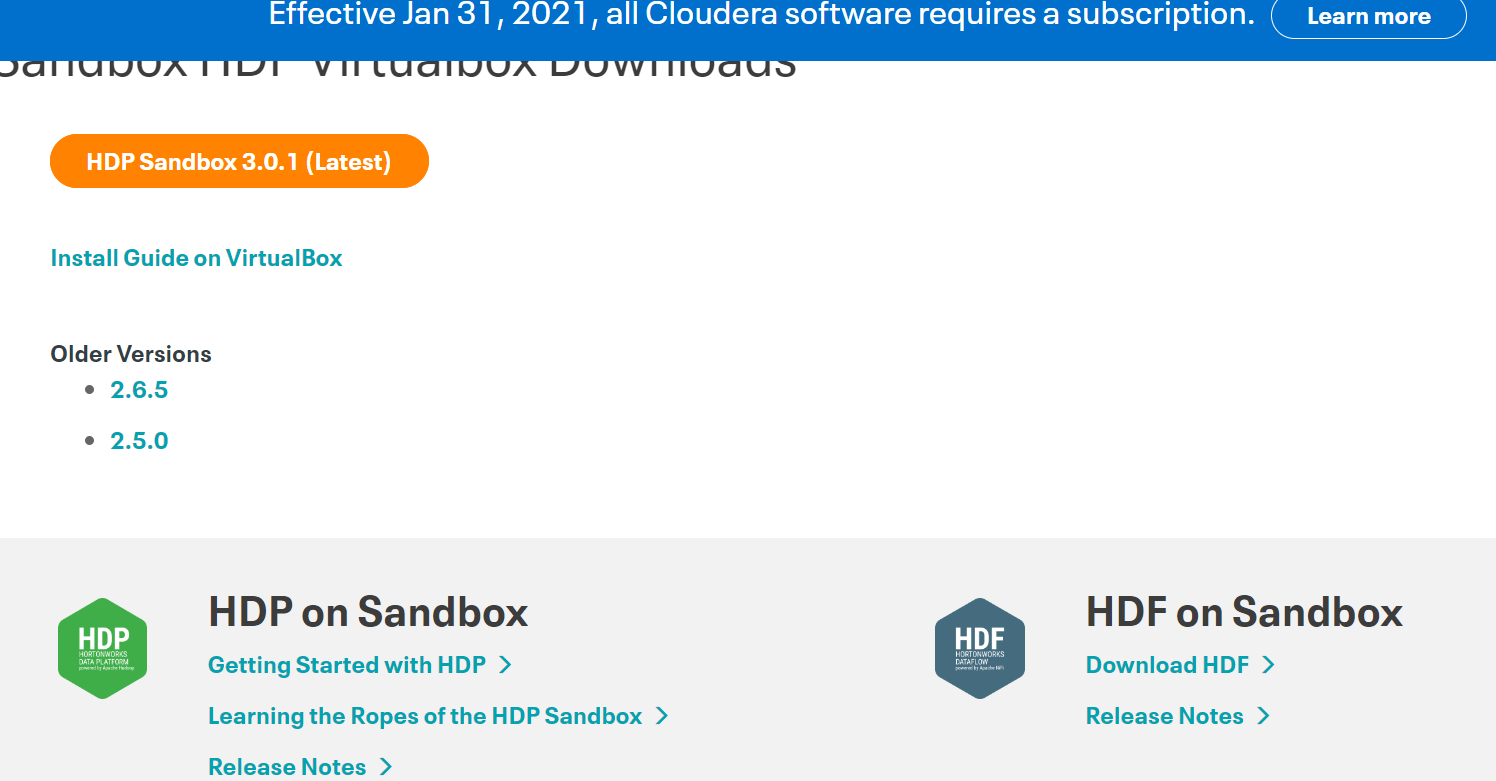
<https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html> -->

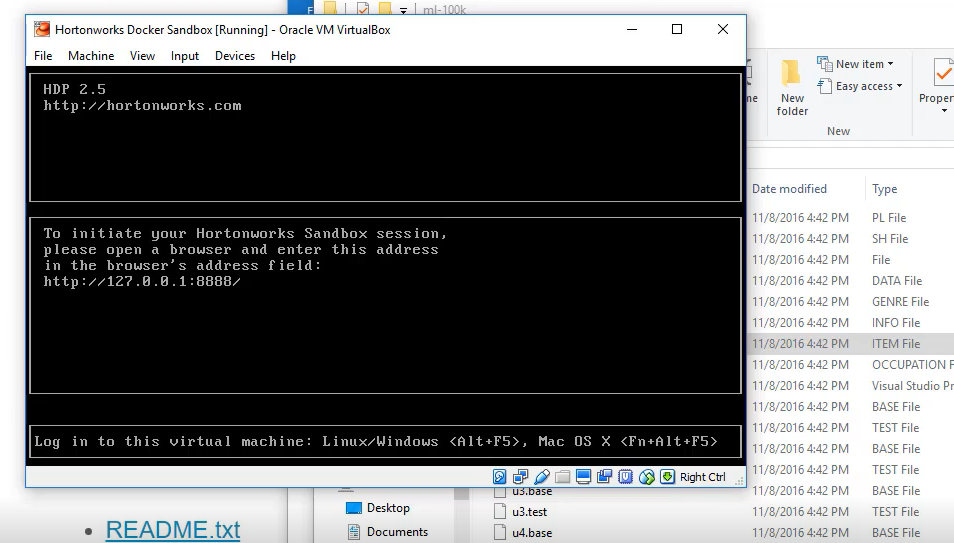
select Virtual box 🡪 Lets Go 🡪 sign in 🡪





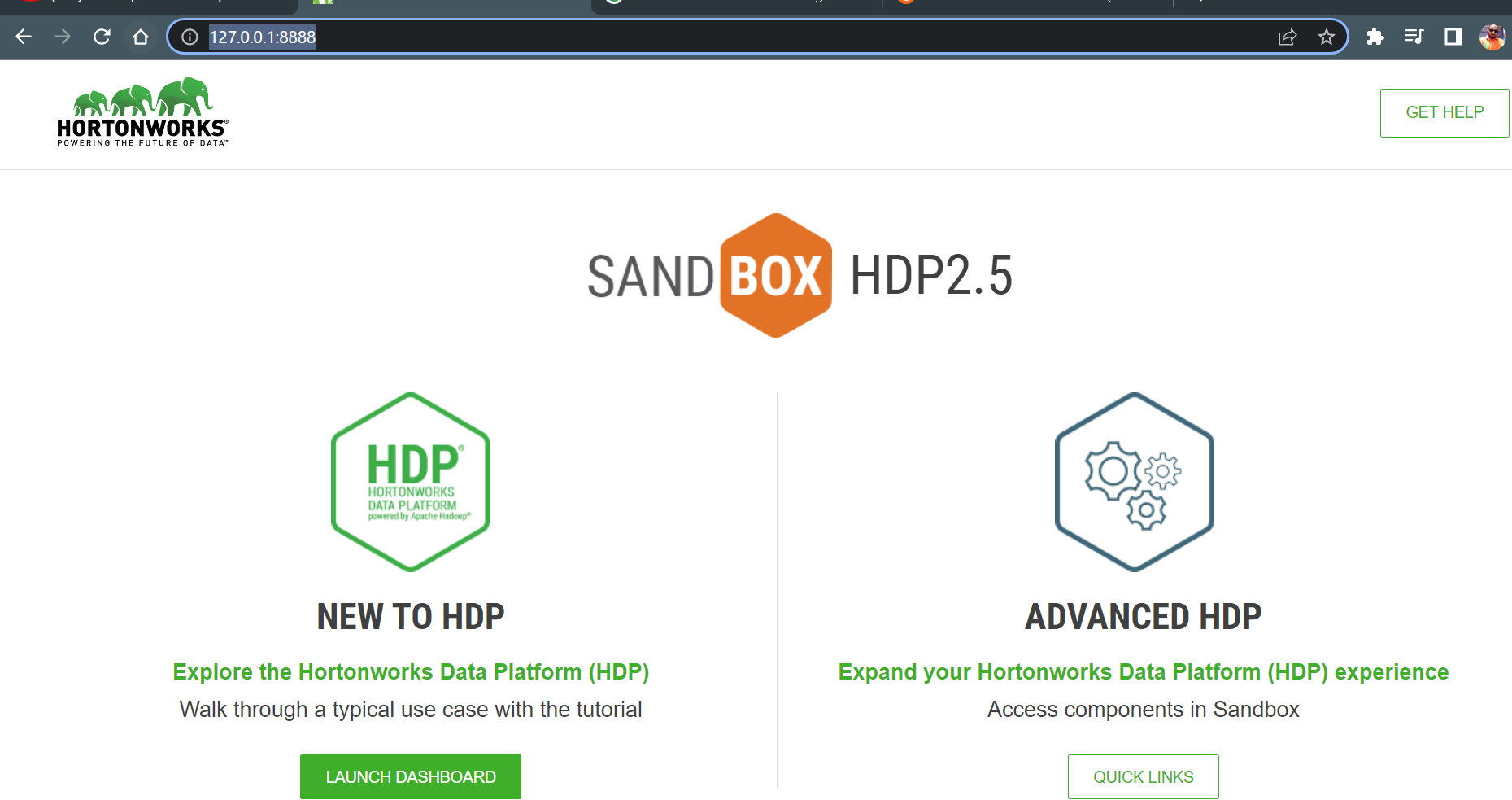
After registration 🡪 need to select for 🡪2.5.0 (11gb) :: 2.6.5. (20gb) :: 3.0.1 (22gb) 🡪 then software will be installed which contains 11gb ( you need have proper internet & enough memory space)



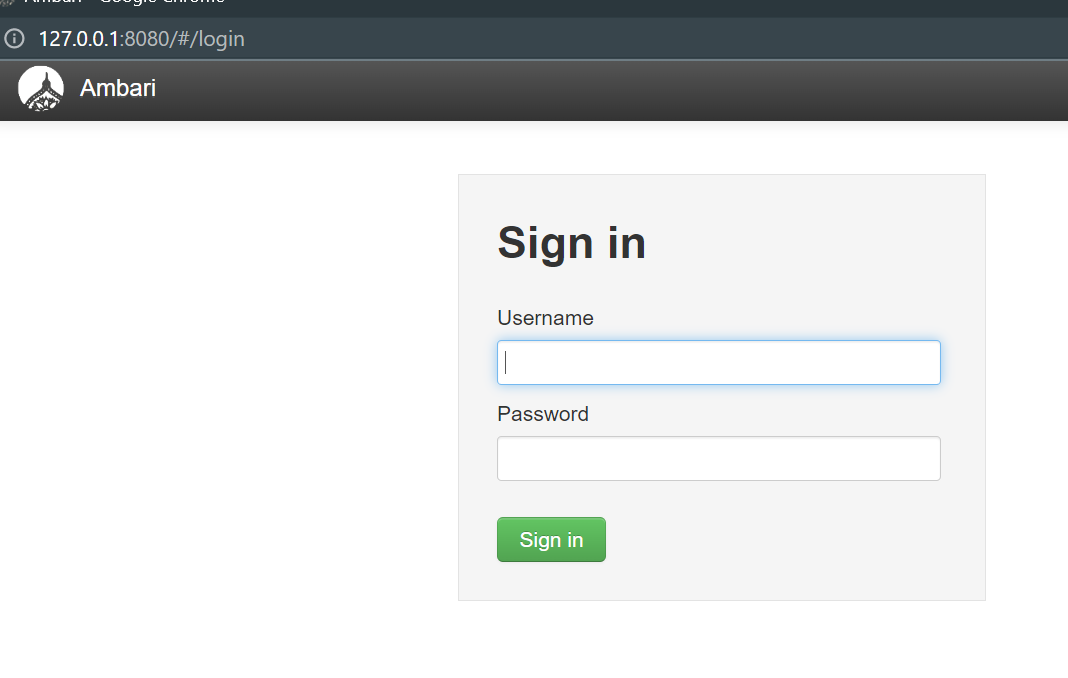
After installation click on the download file (hdp\_2.5\_virtualbo.ove) 🡪change the ram to 4gb 🡪 import 🡪 after successfully installation below screen will appear

Please copy the path above path : <http://127.0.0.1:8888/> 🡪

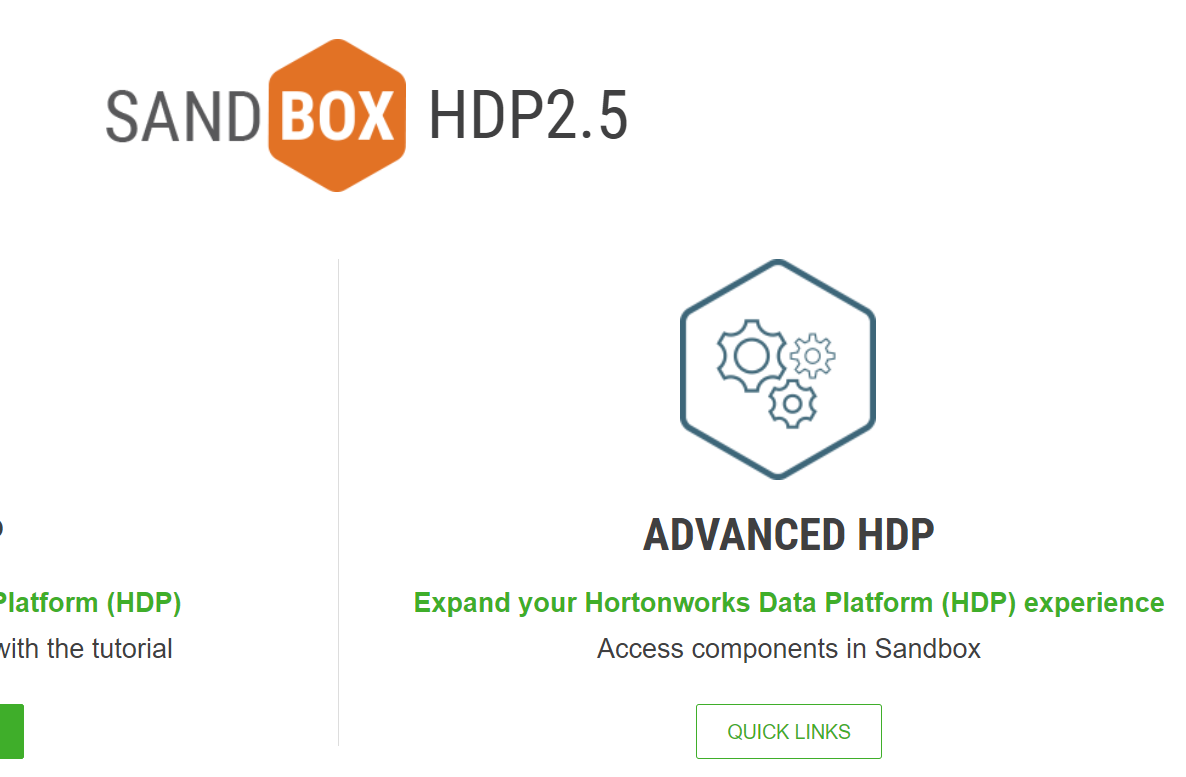
If this path is not works : <http://127.0.0.1:8080/> ||| Login & pwd :: raj\_ops



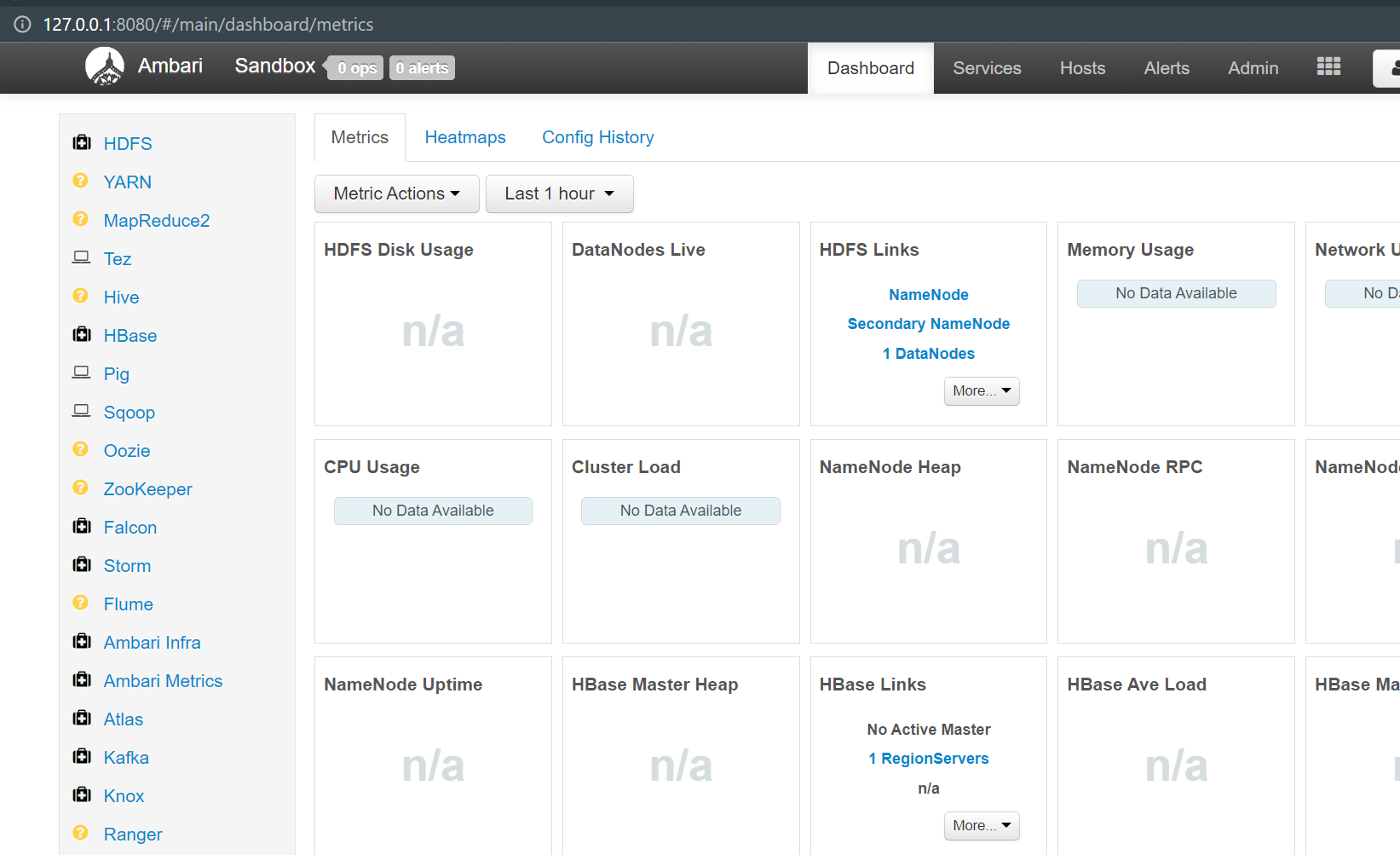
Once you lunch the dashboard 🡪 it will ask for user name & pwd



To get the password 🡪 select on quick links 🡪 click on AMBARI 🡪 username & password: raj\_ops

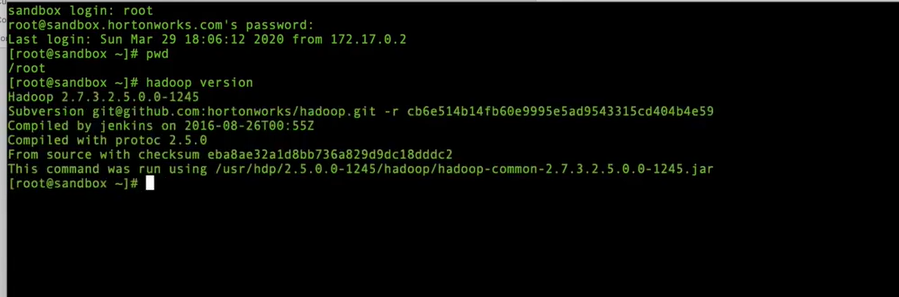


After successfully login we need to get the below screen for successfully installation 🡪 Finally successfully Hadoop installation is completed

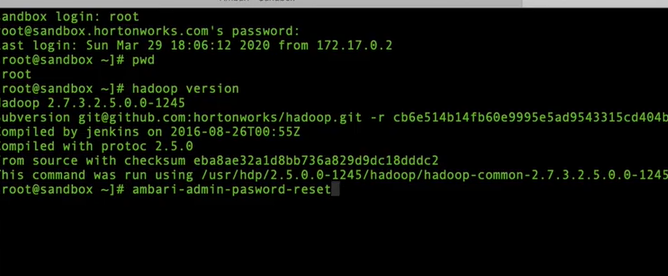


To get the command mode to connect :: localhost:4200 🡪 sandbox login : root & you need to enter your pwd 🡪 Hadoop version





If you want to change to make you own admin & password so you need to To reset the pwd 🡪 use command as : ambari-admin-pasword-reset 🡪 set you pwd



After you change the pwd🡪 signout 🡪 then login with new admin pwd ( better practise is admin (login & pwd) would both same 🡪 Tableau software installation is completed using sandbox

Let’s work on Hadoop:

Log in to Grouplens.org 🡪 <https://grouplens.org/datasets/movielens/> 🡪 click on the link 🡪 download file 🡪 ml-100k.zip 🡪 after extract it contain many files 🡪 u.data 🡪 u.item is the metadata for the movies 🡪

Go to the view 🡪 select on hive view 🡪 upload the table 🡪 setting ( change to 9 (tab(horizontal tab) 🡪 choose the file u.data 🡪 change the table name to ratings 🡪 replace the column name to user\_id, movie\_id, rating , ratingtime 🡪 upload the table

After done go to setting 🡪 in setting change to 124 | 🡪 change table name to movies 🡪 change to column based on your interest 🡪

Let’s go head and run some query 🡪 query 🡪 to check the table 🡪 check on the databases 🡪 refresh the db 🡪 click on default 🡪 you can view the uploaded tables

1. To know which movie id has heighest rating 🡪

SELECT movie\_id, count(movie\_id) as ratingCount

From ratings

GROUP BY movie\_id

ORDER BY ratingCount

DESC;

1. To know top movie which has highest rating

SELECT movie\_id, count(movie\_id) as ratingCount

From ratings

GROUP BY movieid

ORDER BY ratingCount

DESC;