

## Practical-2

**Aim:** Steps to installation of Hadoop.

Step 1) From the browser download the Virtual box

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

Download the software for the windows hosts

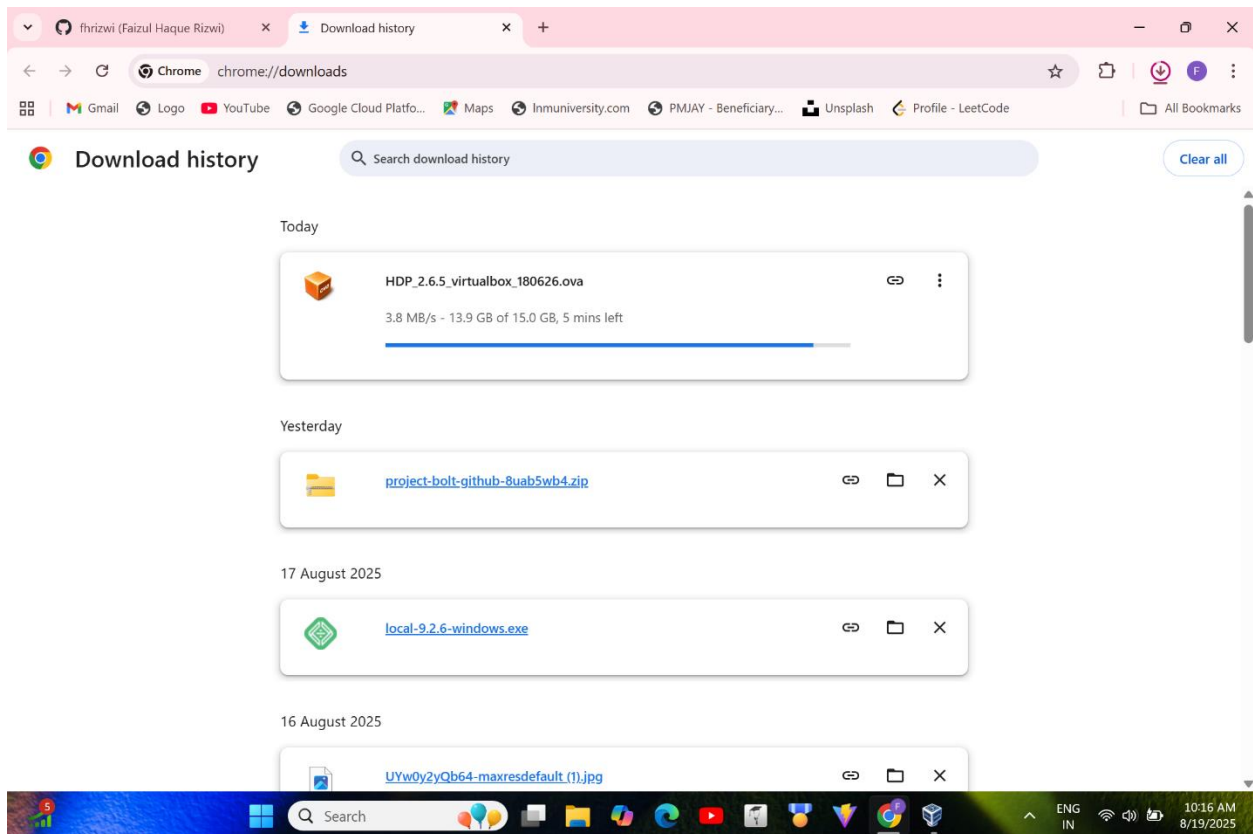
Step 2) Install the **HDP Sandbox**

visit the cloudera platform & search for the hdp sandbox

Link-<https://www.cloudera.com/>

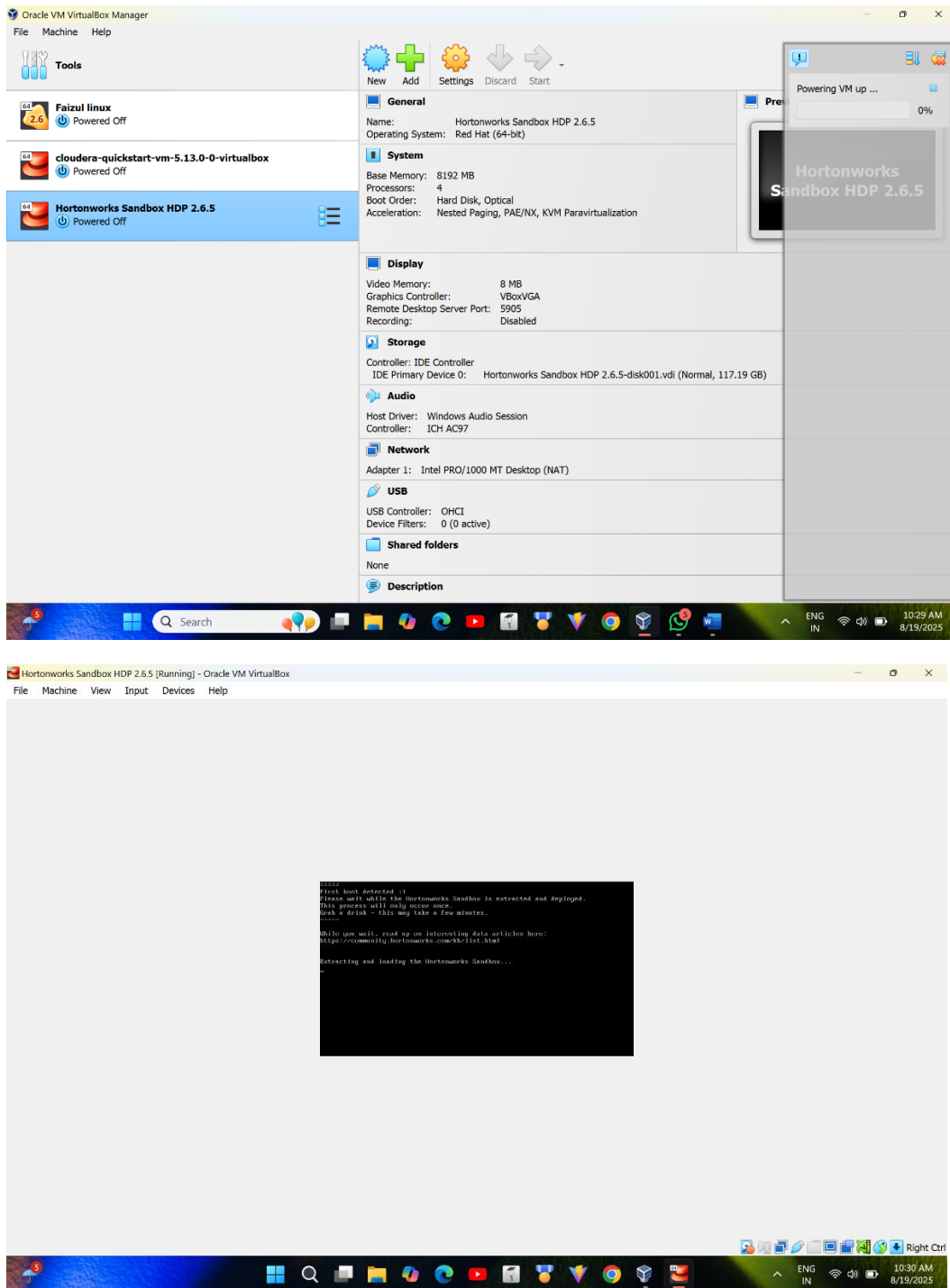
As per new updation since from 2023 the version have been changed so you may get your sandbox through the link provided link below

link for hdp sandbox-[https://archive.cloudera.com/hwx-sandbox/hdp/hdp-2.6.5/HDP\\_2.6.5\\_virtualbox\\_180626.ova](https://archive.cloudera.com/hwx-sandbox/hdp/hdp-2.6.5/HDP_2.6.5_virtualbox_180626.ova)



### Step 3) Import the sandbox in virtual box

After downloading click on import button on the virtual box and import the file  
Hit the start button



Step 4) To visualize what's going on in the Hadoop we may visualize through Ambari  
Go & visit on the address provided on the virtual box  
Click on the address & launch the dashboard  
It requires the username & password

**Username – maria\_dev**

**Password- maria\_dev**

Step 5) Small Activity

Download the dataset from grouplens

**link-<https://grouplens.org/datasets/movielens/>**

older dataset is provided on the website

Hit the download button & download **ml.100k.zip** file

Once you have downloaded extract the data

Step 6) Working on the downloaded dataset

Go into the **ambari** tool and from the menu go into the **hive** view

We will import the data from the local file there to import data open the hive view

After hive view click on the upload table option

Select the csv file type & set the file delimiter type to the 9 (i.e horizontal tab)

Choose the file from your local system (i.e u.data file)

Rename the table name-ratings

**column name-**

user\_id

movie\_id

rating

rating\_time

Hit the upload button

Same for the movie name table

Select the file type as above and set the file delimiter to 124

Rename the table name to movie\_name

**column name-**

Movie\_id

name

After this hit the upload button

Step 7) Write the SQL Query to perform operations

**SQL Query1-**

```
SELECT movie_id, count(movie_id), as ratingcount
FROM ratings
GROUP BY movie_id
ORDER BY ratingCount
DESC
```

After writing the query execute it and see the results

**SQL Query2-**

```
SELECT name
FROM movie_names
WHERE movie_id=50
```

After writing the query execute it and see the results