

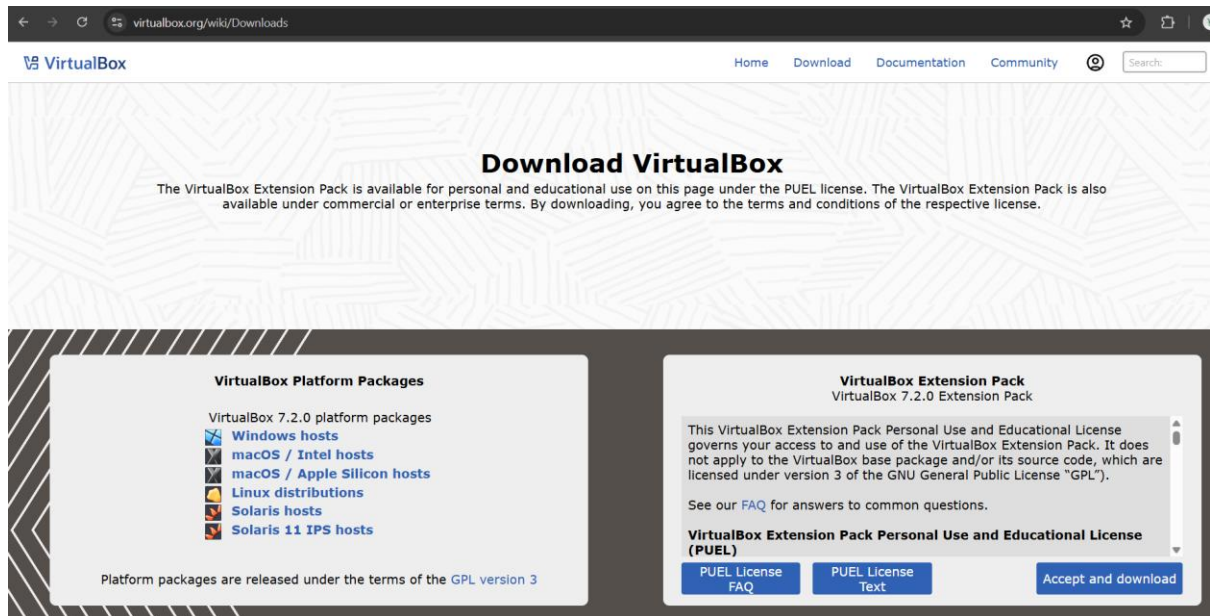
## Practical No.2

**Aim:** Steps to install Hadoop.

Step 1) From the browser download the Virtual box

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

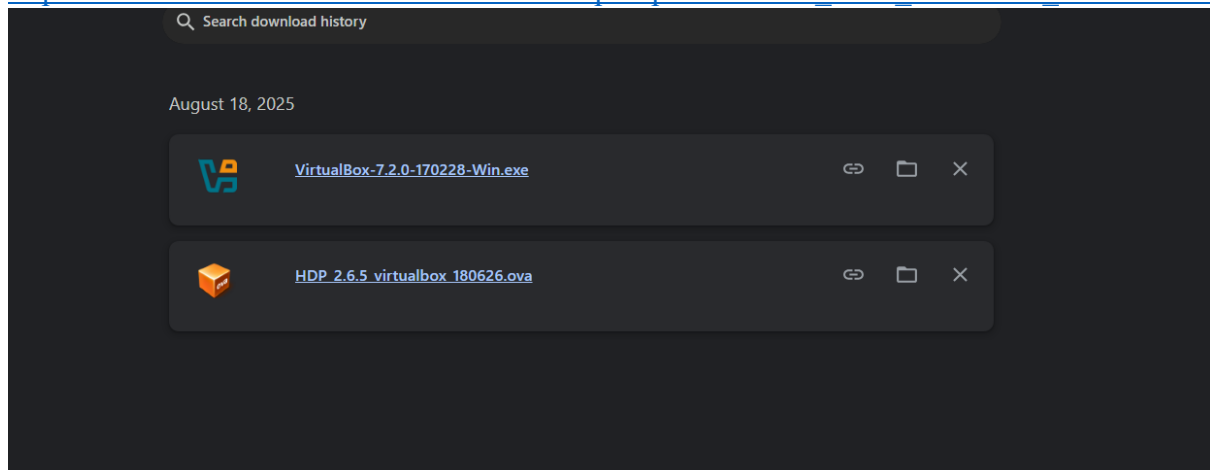
Download the software for the windows hosts



Step 2) Install the HDP Sandbox

Link for hdp sandbox:

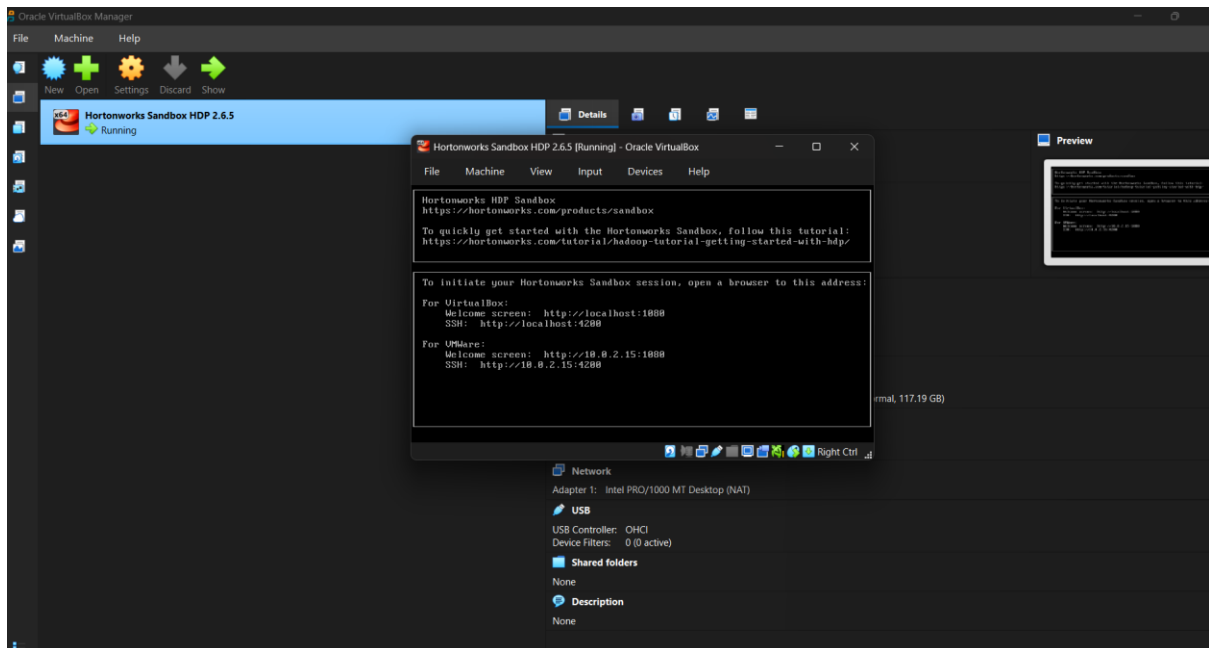
[https://archive.cloudera.com/hwx-sandbox/hdp/hdp-2.6.5/HDP\\_2.6.5\\_virtualbox\\_180626.ova](https://archive.cloudera.com/hwx-sandbox/hdp/hdp-2.6.5/HDP_2.6.5_virtualbox_180626.ova)



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

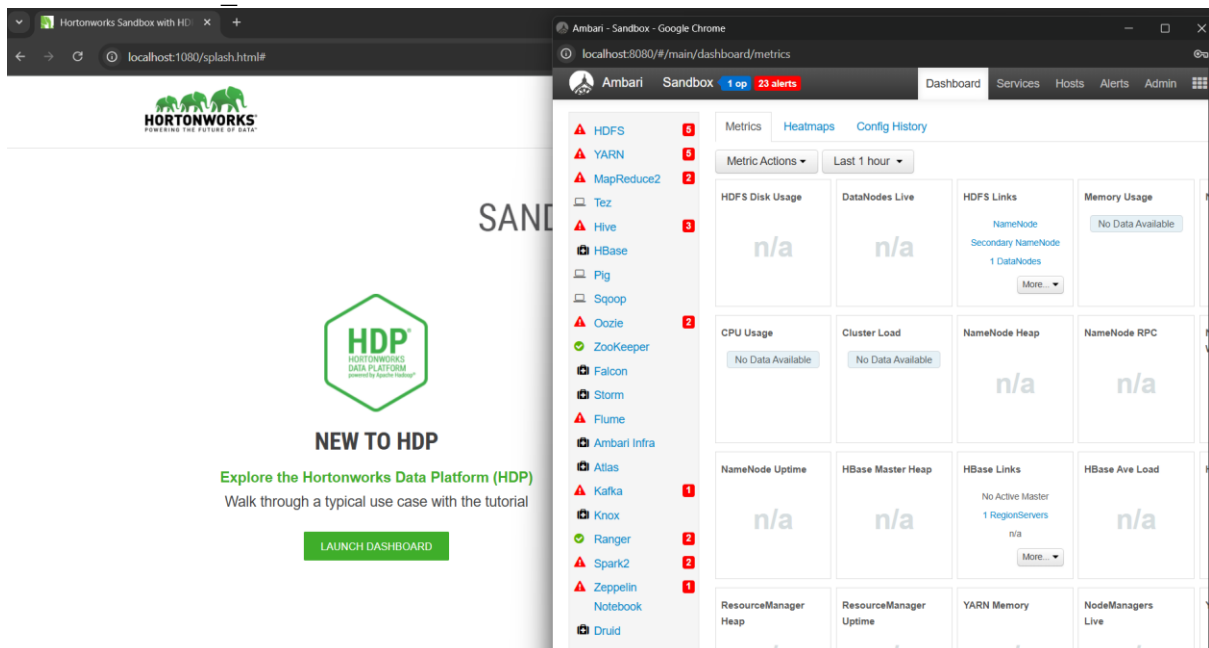
Go & visit on the address provided on the virtual box

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

---

## older datasets

### MovieLens 100K Dataset

MovieLens 100K movie ratings. Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies. Released 4/1998.

- [README.txt](#)
- [ml-100k.zip](#) (size: 5 MB, [checksum](#))
- [Index of unzipped files](#)

Permalink: <https://grouplens.org/datasets/movielens/100k/>

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

Upload from Local

File type

CSV

Database

default

Stored as

ORC

Upload from HDFS

Select from local

Choose File | u5.test

Table name

name-rating

Contains endlines?

Upload Table

| user_id | movie_id | rating | rating_time |
|---------|----------|--------|-------------|
| INT     | INT      | INT    | INT         |
| 1       | 3        | 4      | 878542960   |
| 1       | 13       | 5      | 875071805   |
| 1       | 15       | 5      | 875071608   |
| 1       | 18       | 4      | 887432020   |
| 1       | 19       | 5      | 875071515   |
| 1       | 28       | 4      | 875072173   |
| 1       | 29       | 1      | 878542869   |
| 1       | 52       | 4      | 875072205   |
| 1       | 59       | 5      | 876892817   |
| 1       | 83       | 3      | 875072370   |

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

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

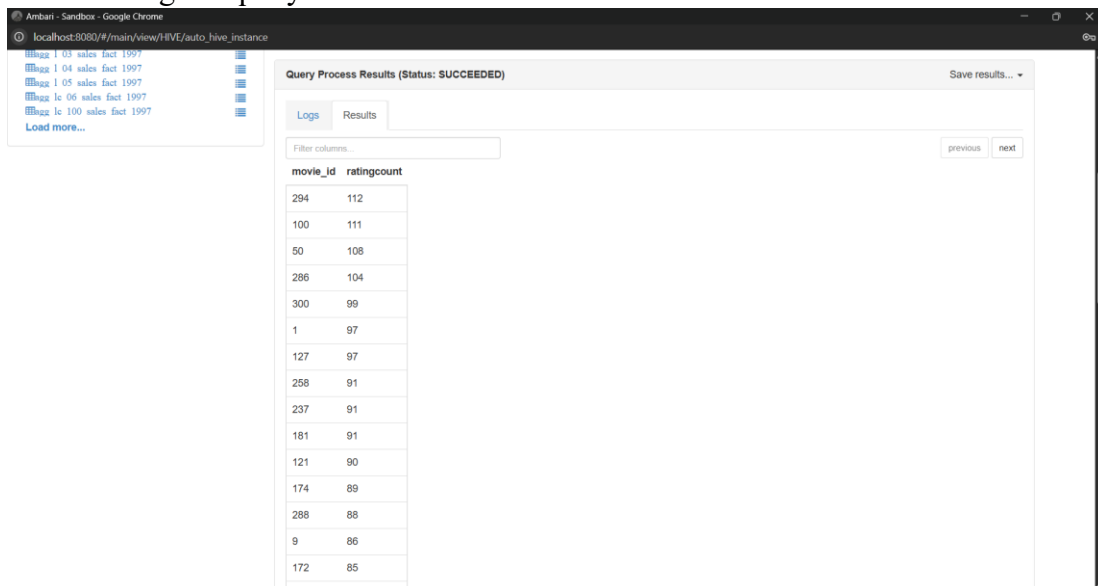
FROM movie\_ratings

GROUP BY movie\_id

ORDER BY ratingCount

DESC;

After writing the query execute it and see the results



Query Process Results (Status: SUCCEEDED)

| movie_id | ratingcount |
|----------|-------------|
| 294      | 112         |
| 100      | 111         |
| 50       | 108         |
| 286      | 104         |
| 300      | 99          |
| 1        | 97          |
| 127      | 97          |
| 258      | 91          |
| 237      | 91          |
| 181      | 91          |
| 121      | 90          |
| 174      | 89          |
| 288      | 88          |
| 9        | 86          |
| 172      | 85          |

