Using Virtual Box 6.1.22

Hortonworks Data Platform (HDP) on Sandbox -2.6.5

Hortonworks Docker Sandbox is a preinstalled Hadoop environment that has a bunch of Associated Technologies. Already Ste up for you. Click on it and Hit Start. Pushing the power button.

Wait for the CentOS Linux to install

Go to grouplens.org to download the MovieLens 100K dataset and Decompress into a folder

u.data – actual rating of the data 4 columns of info, Userid, Movie ID, Rating, Timestamp.

u.item – Metadata for the movie list

To get to the session you type: localhost:<your HDP sandbox number> and you will arrive to the welcome page.

Click on the Launch and login using these Credentials: maria\_dev, pw: maria\_dev

You will arrive into Ambari which lets us visualise what is happening in our Hadoop cluster. You don’t have to worry about running on any system. Not a lot of difference when u are running on one machine or multiple machine.

Import the Movie data and use HIVE to let SQL Queries on top of that data. Use u.data, tab delimiter data. u.item, movie items to movie names which has | delimiter,

Ambari, click on the 3x3 matrix on the top left and click on Use Hive View.

Click on the Upload Table, use the CSV importer. Field Delimiter change it to how the dataset is crafted. And then Input Table name, Change the column names and then upload table. It will generate a table within Hive that basically a view of this data that is sitting on our Hadoop instance.

Write these SQL Queries

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

FROM ratings

GROUP BY movie\_id

ORDER BY ratingCount DESC;

SELECT name

FROM movie\_names

WHERE movie\_id = 50;