PRACTICAL NO.-2

Aim: Installation of Hadoop VM on single node.

DESCRIPTION: -

Hadoop is an open-source framework for distributed storage and processing of large datasets using the MapReduce programming model. In a production setup, Hadoop is deployed on a cluster of multiple nodes. However, for learning and testing purposes, it can be installed on a **single-node VM (Hortonworks HDP Sandbox)**.

This sandbox provides a pre-configured Hadoop environment with supporting tools like **Ambari** for cluster management and **Hive** for data analysis. By running the Hadoop VM in VirtualBox, users can practice data loading, querying, and big data operations without setting up a complex multi-node cluster.

STEPS OF INSTALLATION: -

Step 1: Download VirtualBox

- Open browser and go to: https://www.virtualbox.org/wiki/Downloads
- Download VirtualBox software for Windows hosts.

Step 2: Download HDP Sandbox

- Visit the Cloudera platform: https://www.cloudera.com/
- As per the new updates (since 2023), HDP version changed.
- Use the archived link to download HDP Sandbox for VirtualBox : HDP Sandbox 2.6.5

Step 3: Import Sandbox in VirtualBox

- Open **VirtualBox** → click **File** → **Import Appliance**.
- Select the downloaded .ova file and import it.
- After importing, click **Start** to boot the Hadoop VM.

Step 4: Access Hadoop via Ambari

- Once the VM is running, VirtualBox will display a web access address.
- Copy that address into your browser to open the **Ambari Dashboard**.
- Login with:
  - **Username**: maria_dev
  - **Password**: maria_dev

Step 5: Download Dataset

- Go to: https://grouplens.org/datasets/movielens/
- Download the dataset ml-100k.zip.
- Extract the files locally after download.

Step 6: Import Dataset into Hive (Ambari Tool)

- In **Ambari Dashboard**, go to **Hive View**.
- Click **Upload Table** → Select **CSV file**.

1. Ratings Table

- **File:** u.data
- **Delimiter:** \t (tab = 9)
- **Table Name:** ratings
- **Columns:**
    - user_id
    - movie_id
    - rating
    - rating_time
- Click **Upload**.

2. Movie Names Table

- **File:** (from dataset, movie names file)
- **Delimiter:** | (ASCII 124)
- **Table Name:** movie_name
- **Columns:**
    - movie_id
    - name
- Click **Upload**.

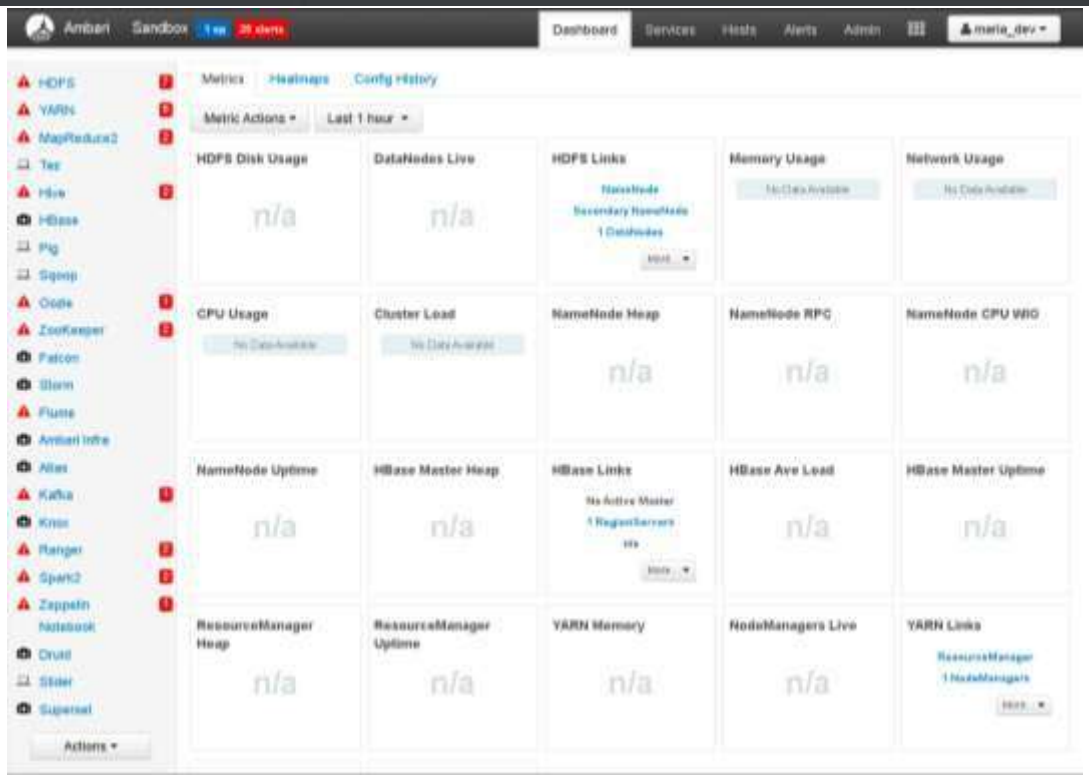
Step 7: Run SQL Queries in Hive

Query 1 – Count ratings per movie:

```
SELECT movie_id, COUNT(movie_id) AS ratingCount
FROM ratings
GROUP BY movie_id
ORDER BY ratingCount DESC;
```

Query 2 – Get movie name for ID = 50:

```
SELECT name
FROM movie_name
WHERE movie_id = 50;
```



CONCLUSION: -

Hadoop VM (HDP Sandbox) was successfully installed on a single node using VirtualBox. The environment was accessed through Ambari, datasets were imported into Hive, and SQL queries were executed to analyze the data. This setup provides a simplified but powerful platform for practicing Big Data operations.