**1. Start the Cloudera VM**

1. **Open Oracle VirtualBox**.
2. Select the imported **Cloudera Quickstart VM** (usually named cloudera-quickstart-vm).
3. Click **Start**.
4. Wait for the VM to boot (it takes a few minutes).
5. Login to the desktop (username: cloudera, password: cloudera).

**✅ 2. Open the Terminal in Cloudera**

1. Once the desktop loads, **double-click on the "Terminal"** icon or open it from the Applications menu.
2. Run:

bash

CopyEdit

hive

This opens the **Hive CLI (Command Line Interface)** where you can execute HiveQL queries.

**✅ 3. Verify Hadoop is Running (Optional)**

If you need to verify Hadoop:

bash

CopyEdit

hadoop version

To list running Hadoop daemons:

bash

CopyEdit

jps

Look for:

* NameNode
* DataNode
* SecondaryNameNode
* ResourceManager
* NodeManager

If services aren't running, you can start them using:

bash

CopyEdit

sudo /home/cloudera/cloudera-manager --express

**✅ 4. Create and Use a Hive Database**

sql

CopyEdit

CREATE DATABASE flight\_db;

USE flight\_db;

**✅ 5. Create a Hive Table**

**Option A: Internal Table**

sql

CopyEdit

CREATE TABLE flight\_info (

flight\_id INT,

airline STRING,

origin STRING,

destination STRING,

dep\_time STRING,

arr\_time STRING,

dep\_delay INT,

arr\_delay INT,

flight\_date STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

**Option B: External Table (if CSV is in local path)**

sql

CopyEdit

CREATE EXTERNAL TABLE flight\_info\_external (

flight\_id INT,

airline STRING,

origin STRING,

destination STRING,

dep\_time STRING,

arr\_time STRING,

dep\_delay INT,

arr\_delay INT,

flight\_date STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/cloudera/flights/';

**✅ 6. Load CSV Data into Hive Table**

1. Place your flights\_2008.csv file in the Cloudera VM (e.g., in /home/cloudera/).
2. Use:

sql

CopyEdit

LOAD DATA LOCAL INPATH '/home/cloudera/flights\_2008.csv' INTO TABLE flight\_info;

**✅ 7. Insert New Row**

sql

CopyEdit

INSERT INTO TABLE flight\_info VALUES (1003, 'SpiceJet', 'DEL', 'PNQ', '07:30', '09:30', 5, -3, '2008-01-02');

**✅ 8. Alter Table**

sql

CopyEdit

ALTER TABLE flight\_info ADD COLUMNS (status STRING);

**✅ 9. Join with Another Table (Example)**

First create another table:

sql

CopyEdit

CREATE TABLE airline\_info (

airline STRING,

country STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

Then join:

sql

CopyEdit

SELECT f.airline, a.country

FROM flight\_info f

JOIN airline\_info a ON (f.airline = a.airline);

**✅ 10. Create Index**

sql

CopyEdit

CREATE INDEX flight\_index

ON TABLE flight\_info (flight\_id)

AS 'COMPACT'

WITH DEFERRED REBUILD;

**✅ 11. Find Average Departure Delay Per Day in 2008**

sql

CopyEdit

SELECT flight\_date, AVG(dep\_delay) AS avg\_dep\_delay

FROM flight\_info

WHERE flight\_date LIKE '2008%'

GROUP BY flight\_date;

**🧠 Tips**

| **Task** | **Command/Action** |
| --- | --- |
| Start Hive CLI | hive |
| Run Hadoop Daemons | jps or start via services |
| Create DB | CREATE DATABASE db\_name; |
| Exit Hive | exit; |
| Check tables | SHOW TABLES; |
| Describe table | DESCRIBE table\_name; |

**✅ Step 1: Download Cloudera Quickstart VM**

Cloudera has discontinued the official Quickstart VM, but you can **still get it from third-party sources or community repositories**.

⚠️ **Note**: Be careful while downloading from unofficial sources. Use reputable repositories or ask your instructor if they have provided the .ova file.

You can try this trusted link:

* 🔗 Cloudera Quickstart VM (via Archive.org)
  + Download the **cloudera-quickstart-vm-5.13.0-0-virtualbox.zip**
  + Extract it to get the .ova file.

**✅ Step 2: Install Oracle VirtualBox (if not already installed)**

1. Go to <https://www.virtualbox.org/>
2. Download and install for Windows.

**✅ Step 3: Import the Cloudera VM into VirtualBox**

1. Open **Oracle VirtualBox**.
2. Click on **File > Import Appliance**.
3. Browse to the downloaded .ova file (e.g., cloudera-quickstart-vm-5.13.0-0-virtualbox.ova).
4. Click **Next**, then **Import**.
5. Wait for the VM to import (can take 5–10 minutes).

**✅ Step 4: Start the Cloudera VM**

1. Select the imported VM (cloudera-quickstart-vm) in VirtualBox.
2. Click **Start**.
3. Wait for Linux to boot up (first time takes a bit longer).
4. Login with:
   * **Username:** cloudera
   * **Password:** cloudera

**✅ Step 5: Start Hive, Hadoop, etc.**

1. Open the **Terminal** inside the VM.
2. Start Hive CLI:

bash

CopyEdit

hive

1. If you want to run Hadoop commands:

bash

CopyEdit

hadoop version

jps

**📂 Optional: Share Files from Windows to Cloudera VM**

To load datasets into Hive:

1. In VirtualBox, go to **Settings > Shared Folders** for the Cloudera VM.
2. Add a folder (e.g., your Windows folder containing flights.csv).
3. Enable **Auto-mount**.
4. After booting the VM, the shared folder will be mounted in /media/sf\_<foldername>.

Then you can load files like this:

sql

CopyEdit

LOAD DATA LOCAL INPATH '/media/sf\_flights/flights\_2008.csv' INTO TABLE flight\_info;

**✅ Summary**

| **Task** | **Action** |
| --- | --- |
| Download VM | Archive.org Cloudera Quickstart VM |
| VM Username/Password | cloudera / cloudera |
| Start Hive | hive in terminal |
| Load Data into Hive | LOAD DATA LOCAL INPATH ... INTO TABLE ... |
| Shared Folder (Optional) | Use Settings > Shared Folders in VirtualBox |