**Project Title:**

**“ Dynamic Data Ingestion and Storage in HDFS with Automated Hive Integration “**

**Objective:**

To design and implement a system that dynamically ingests structured data from an external source, stores it in **Hadoop Distributed File System (HDFS)**, and integrates it into **Hive** for structured querying and visualization. The process also aims to support automation using scripting for efficient data refreshes.

**Problem Statement:**

The task involves fetching data from a specified URL (<https://www2.census.gov/programs-surveys/popest/datasets/>), storing it in HDFS, and creating a corresponding Hive table for querying and analysis. This project showcases how to:

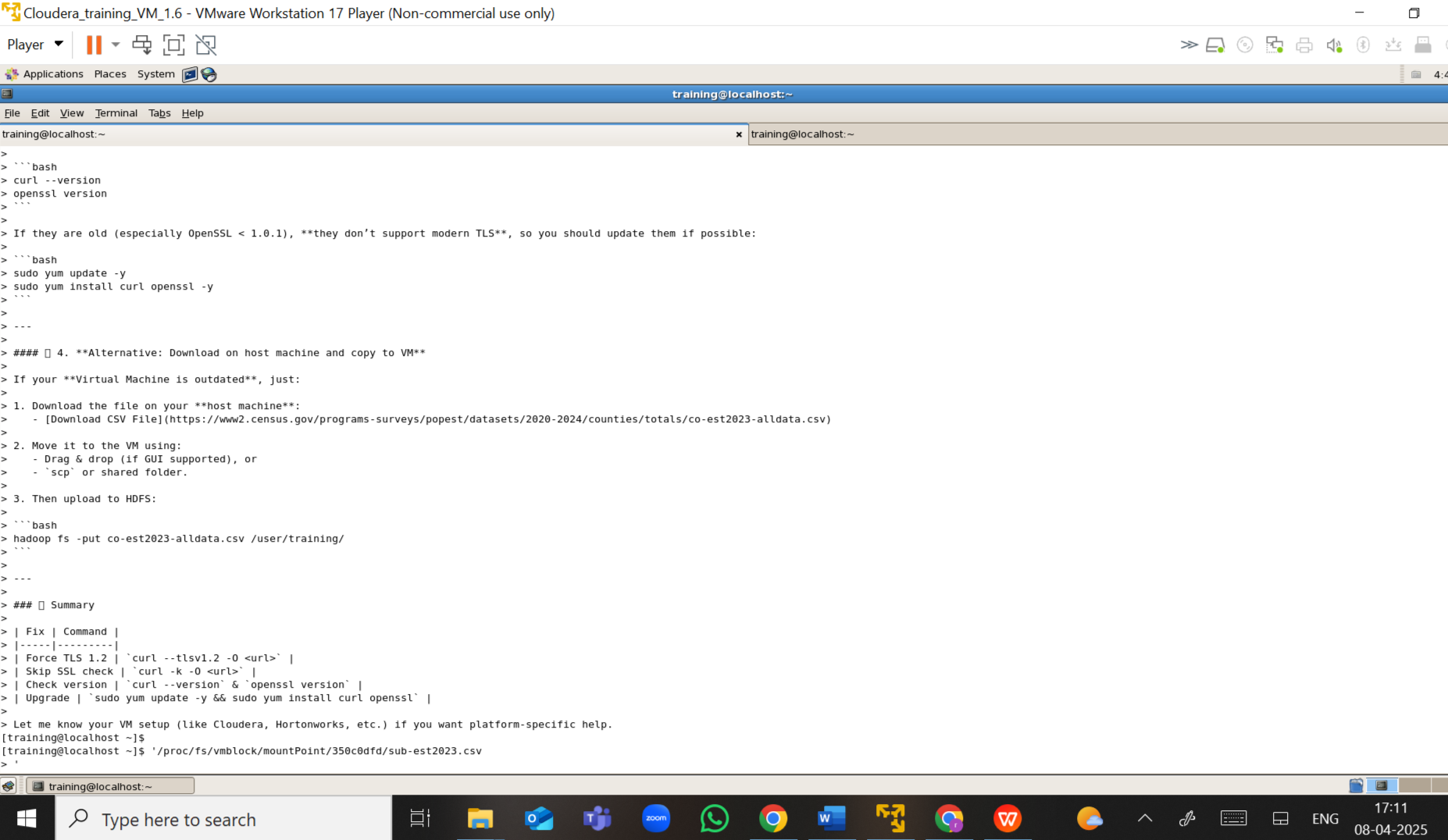
* Ensure data accessibility,
* Determine and align data schema,
* Store and manage data in HDFS,
* Integrate with Hive,
* And automate the entire pipeline using scripting.

**Platform:**

* **Virtual Hadoop Machine** running HDFS and Hive

**Image-1:  
 If your Virtual Machine is outdated, just:**

1. **Download the file on your host machine:**
   * [**Download CSV File**](https://www2.census.gov/programs-surveys/popest/datasets/2020-2024/counties/totals/co-est2023-alldata.csv)
2. **Move it to the VM using:**
   * **Drag & drop (if GUI supported)**
   * **drag the downloaded .csv file from your host desktop into the VM window.**

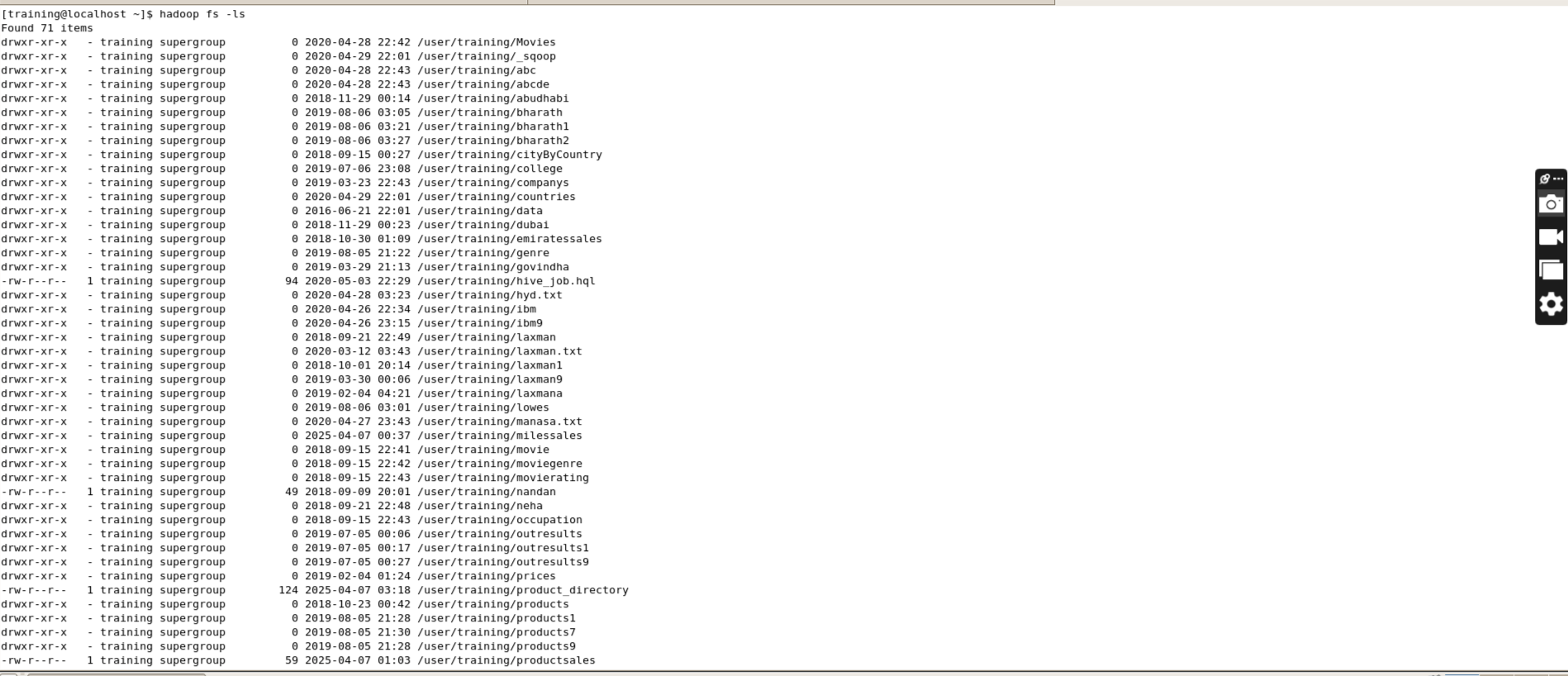


**Image-1: Download on host machine and copy to Virtual machine  
because VM still doesn’t support to downloads.**

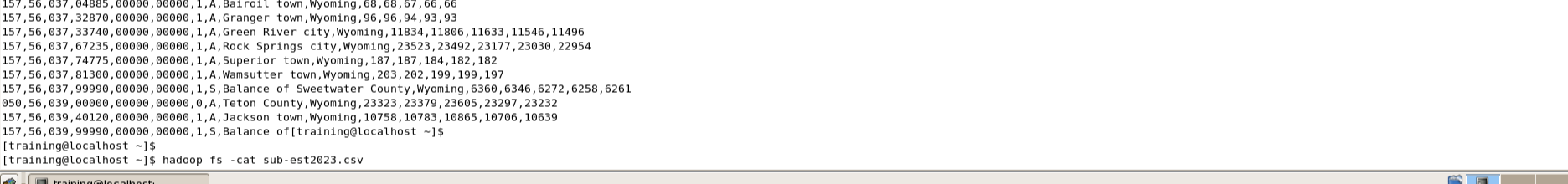
**So, I used drag the downloaded .csv file from your host desktop into the VM window.**

**Image – 2:**

**“ Use Hadoop commands to list all files and verify that the required file is present “**

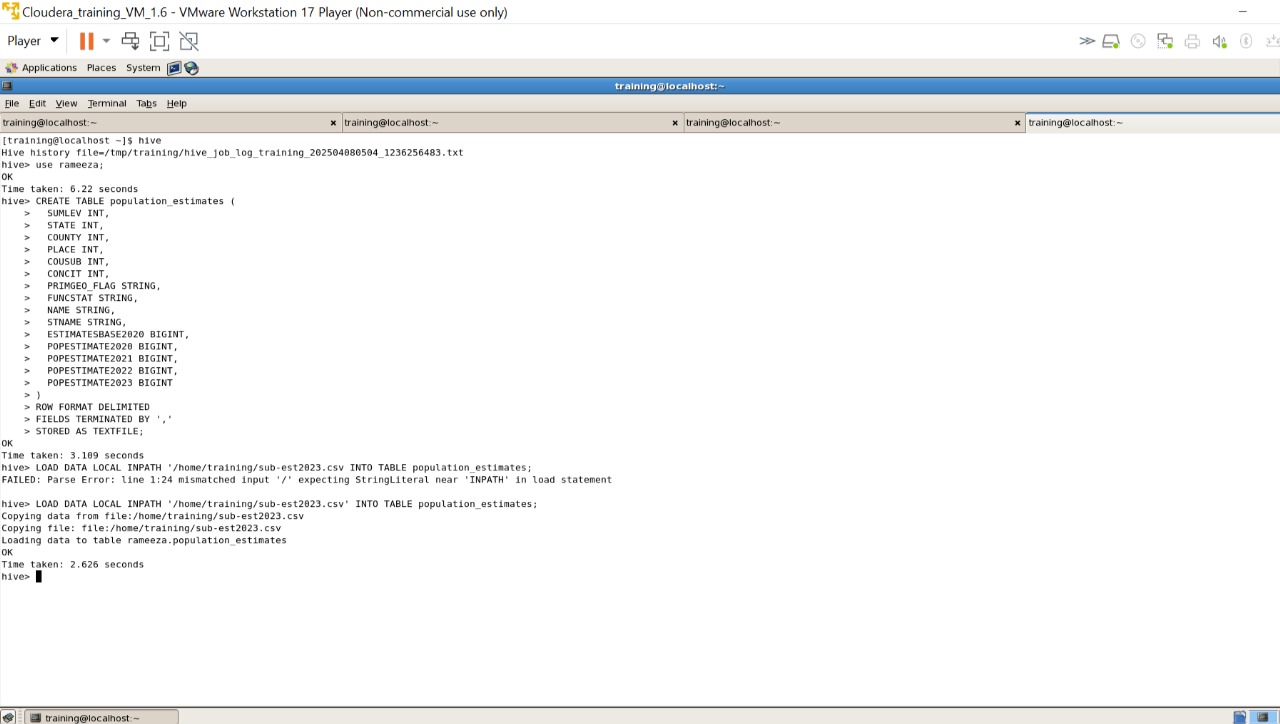


**Image -3 Using Cat command :**



**"The -cat command is used to read the required file from the directory in Hadoop”**

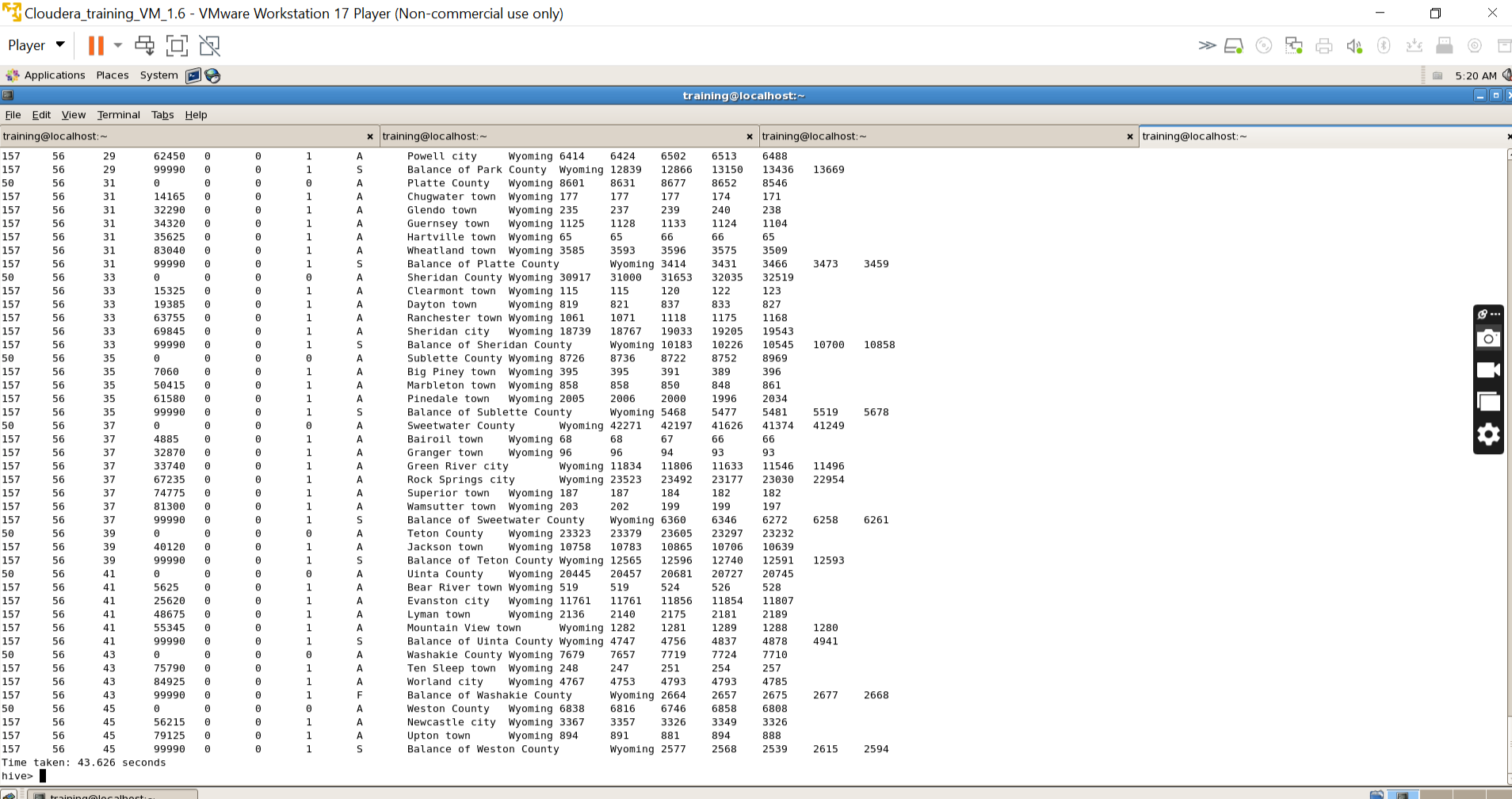
**Use Case:  
This is especially useful to verify data after uploading or downloading files, or before creating Hive tables, ensuring the file has the expected format and data.**

**Image -4 : Hive Environment Setup**

* Start Hive
* Create and switch to a database
* Create table with appropriate schema
* Load CSV data into Hive
* Run queries to analyse the data

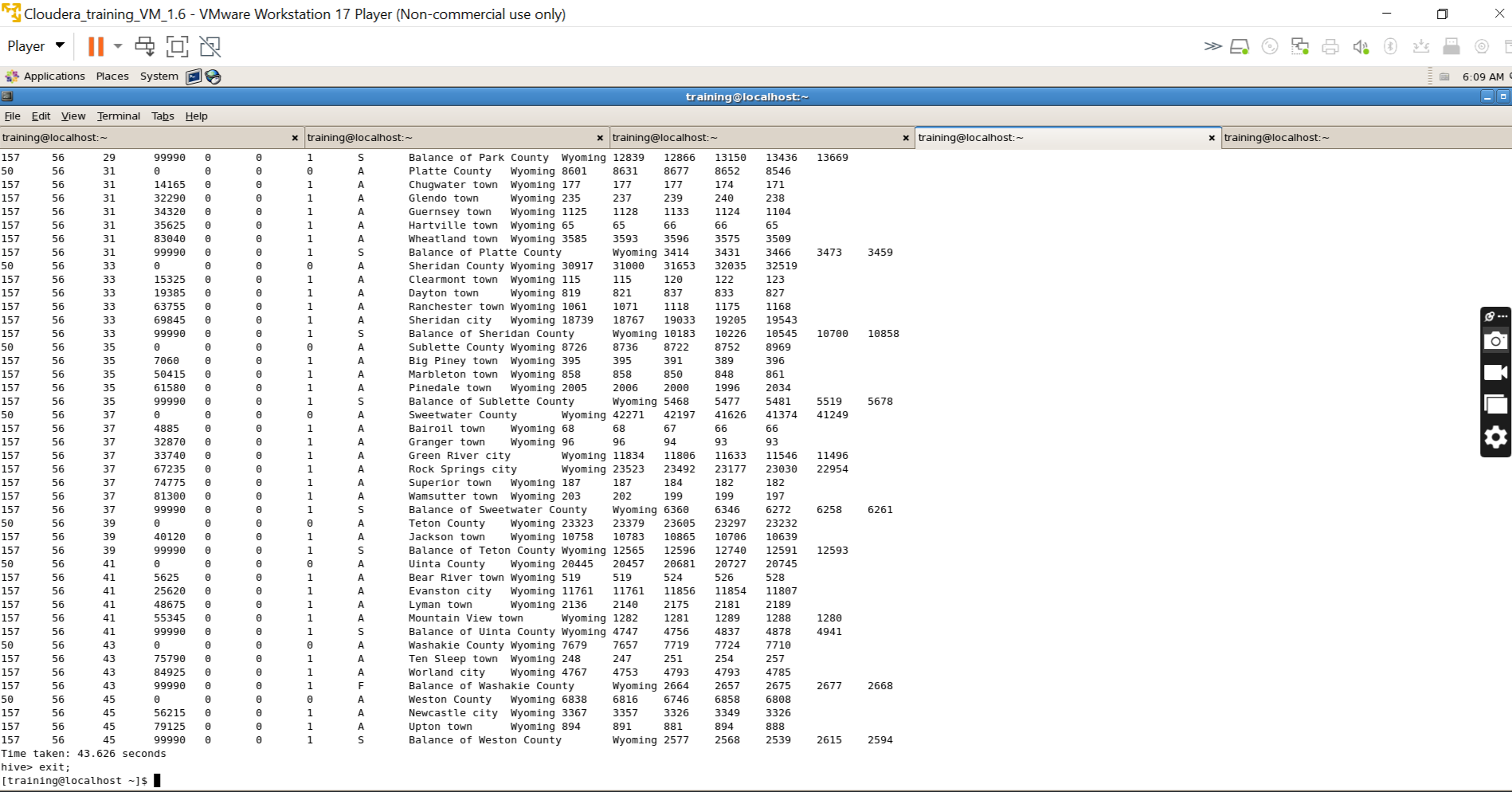
**Image – 5: Verify the Data**

**Run a query to confirm data is loaded correctly**

****

**Command : Select \* From Table\_Name;**

**Image-6:Hive shell Exit;**



**Technologies Used:**

* **Python** – for automation and scripting (optional)
* **Shell Scripting** – for CLI-based automation
* **HDFS (Hadoop Distributed File System)** – for storing big data
* **Hive** – for SQL-like querying and data visualization