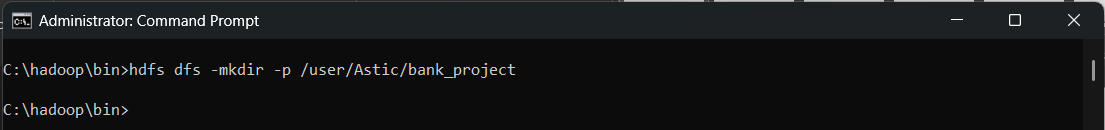
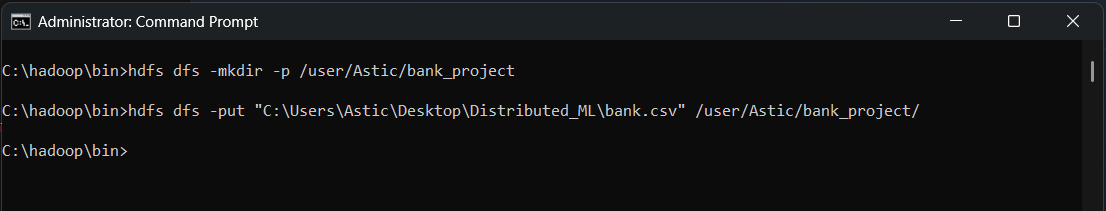
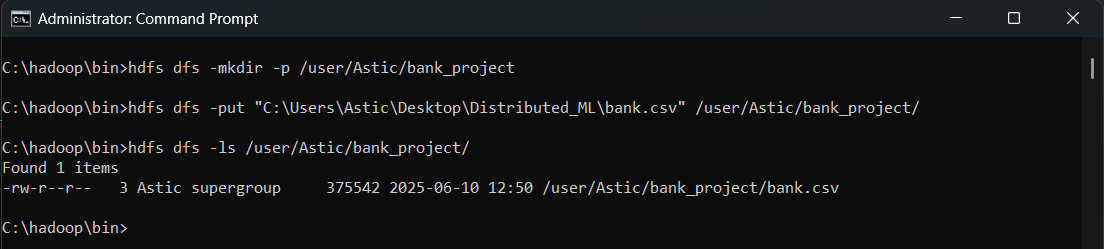
**HADOOP**

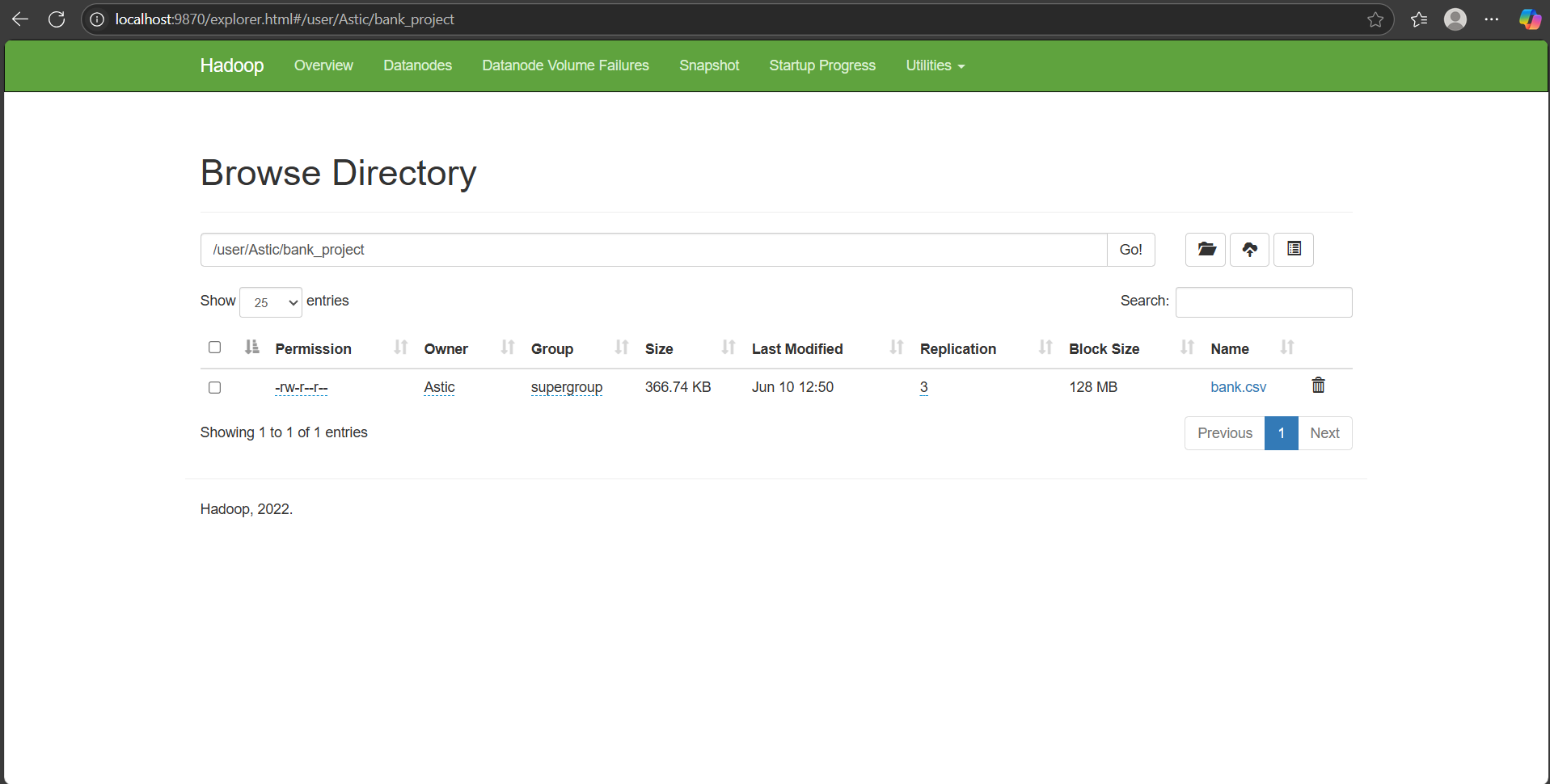
1. **Data Ingestion:**

* Create a directory in HDFS and transfer the banking dataset from the local system to the HDFS directory.



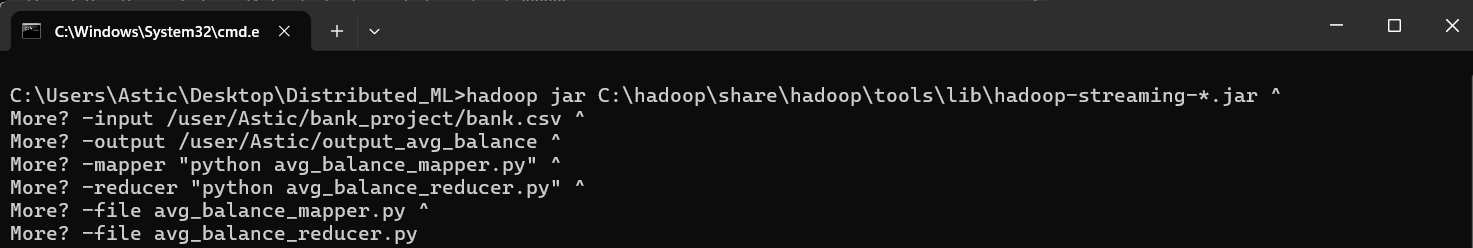


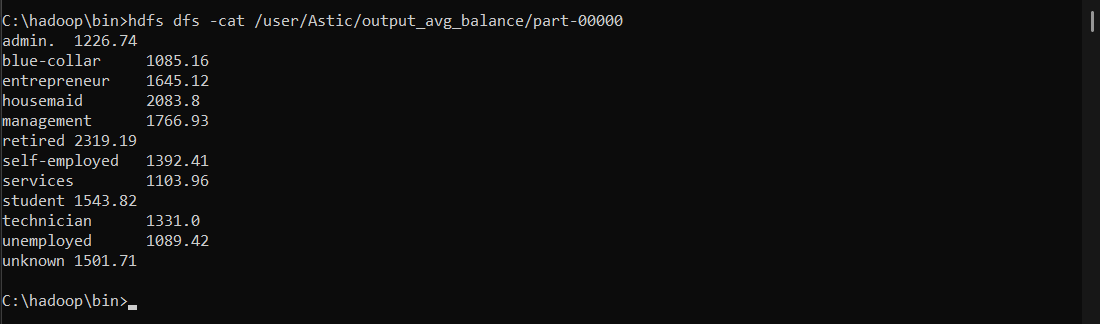




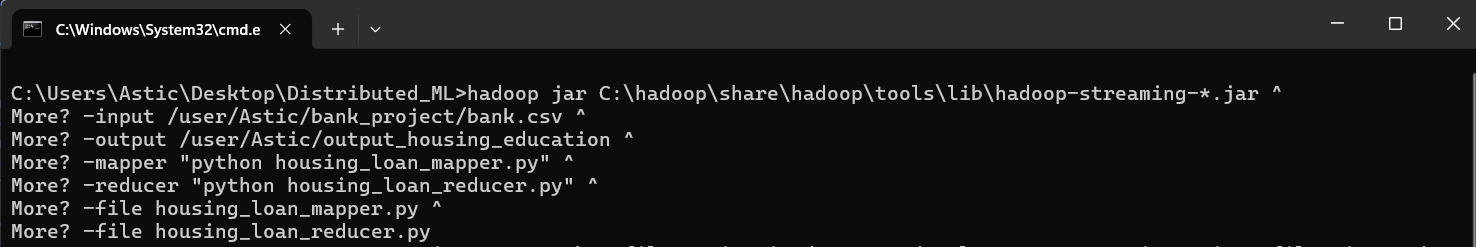
1. **Data Transformation with MapReduce:**

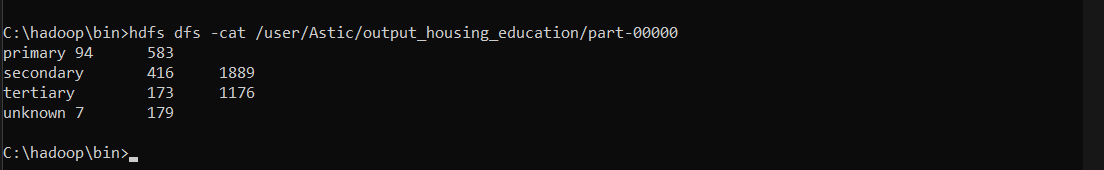
* Write a MapReduce program in Python that calculates the average account balance for each job type.



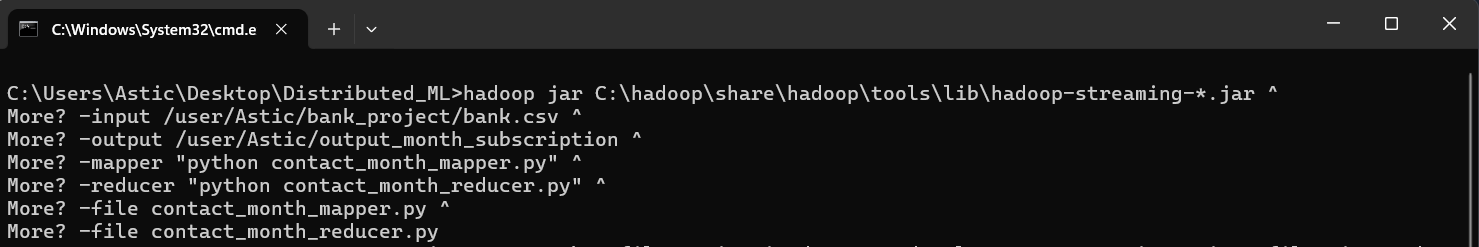


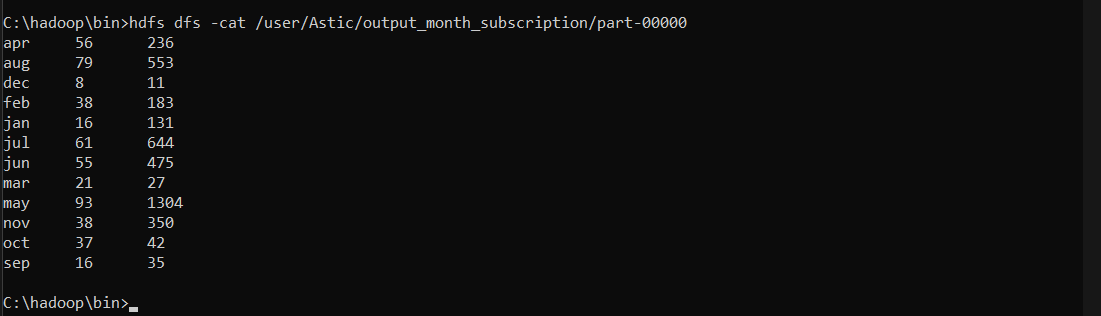
* Write another MapReduce program that counts the number of individuals with and without a housing loan in each education category.





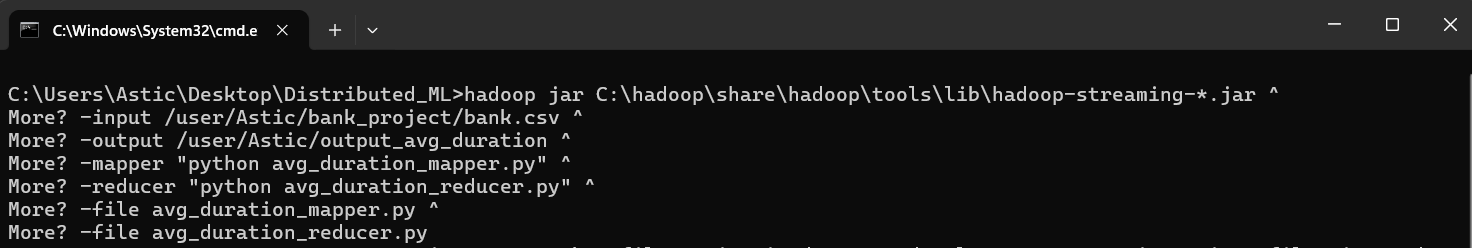
* Perform a MapReduce job to determine the number of clients contacted in each month and their subscription status to term deposits ('y' column).

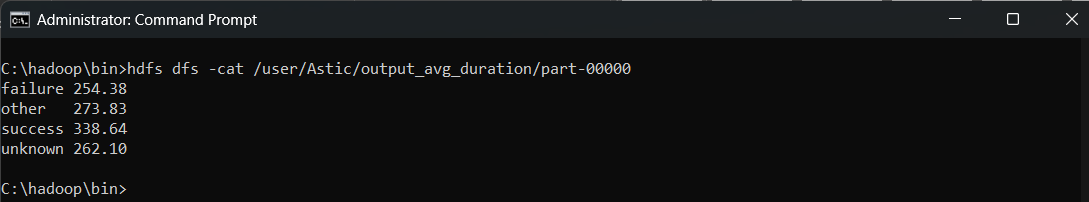




1. **Data Analysis with MapReduce:**

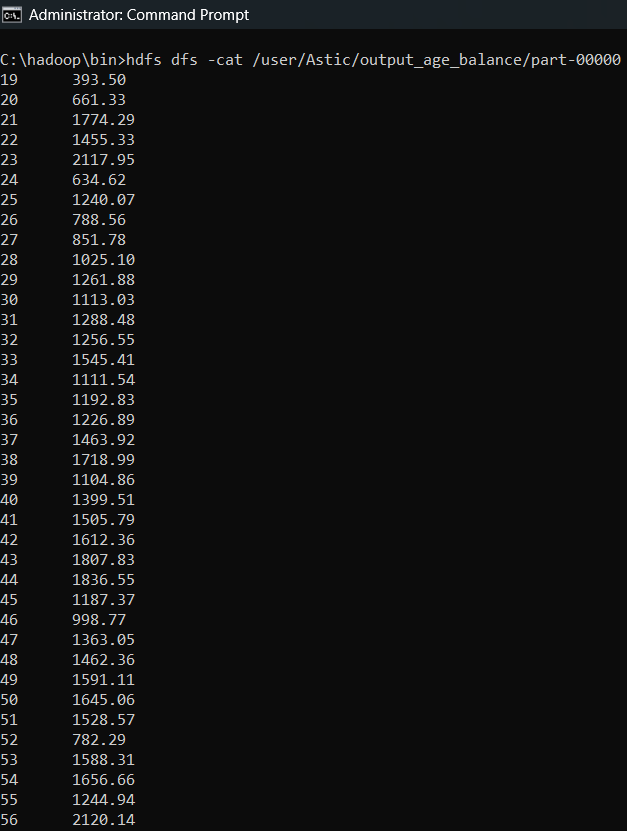
* Analyze the average duration of contact (in seconds) per campaign outcome ('poutcome').

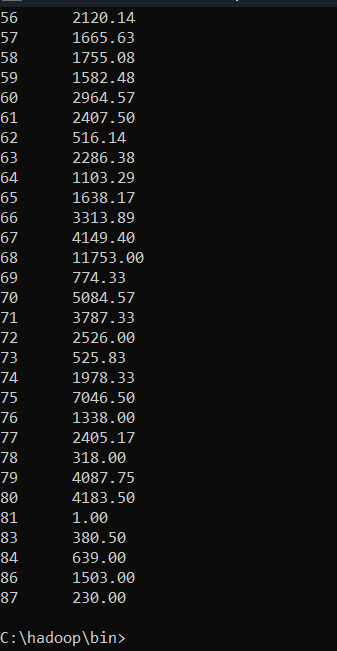




**Summary:** The MapReduce job analyzed the dataset to calculate the average contact duration (in seconds) for each campaign outcome. The results are as follows: ● Failure: Average contact duration is 254.38 seconds. ● Other: Average contact duration is 273.83 seconds. ● Success: Average contact duration is 338.64 seconds. ● Unknown: Average contact duration is 262.10 seconds.

* Examine the relationship between the age of clients and their balance, and present findings in a summarized form.





**Summary**: After running the MapReduce job, the output provides the average account balance for each specific age. Here’s a summary of the key points:

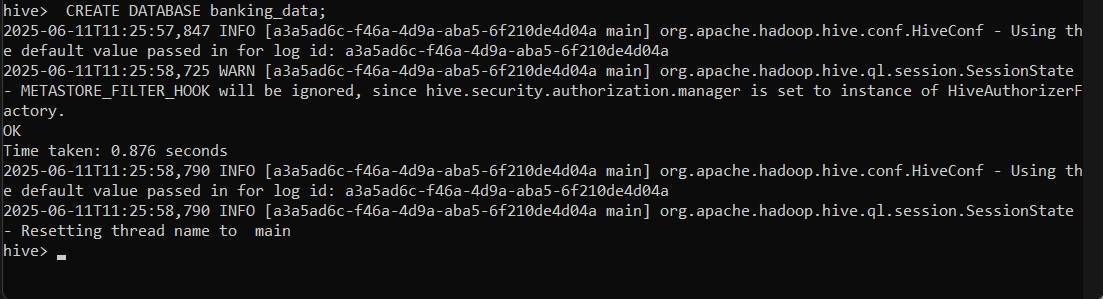
● Age-Specific Averages: The output shows the average balance corresponding to each client’s age. For example, a 23-year-old might have an average balance of 2117.95, while a 25-year-old might have an average balance of 1240.05.

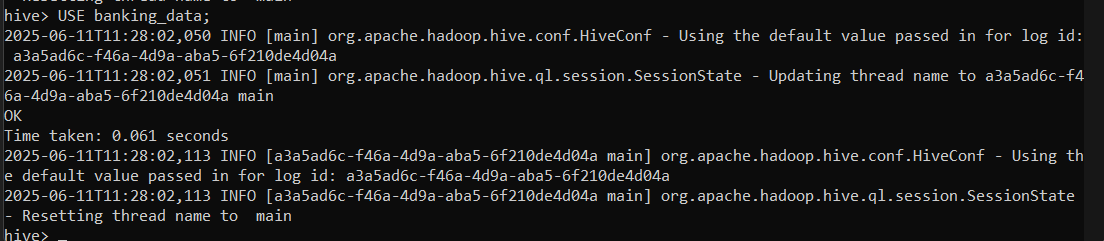
● Trends: If observed across ages, you might notice trends such as: ○ Increase with Age: In some cases, there might be a gradual increase in average balance as age increases. ○ Fluctuations: Certain ages might show higher or lower average balances due to specific financial behaviours or life events.

● Variability: The average balances might fluctuate significantly across different ages, reflecting the diverse financial situations of clients at various life stages.

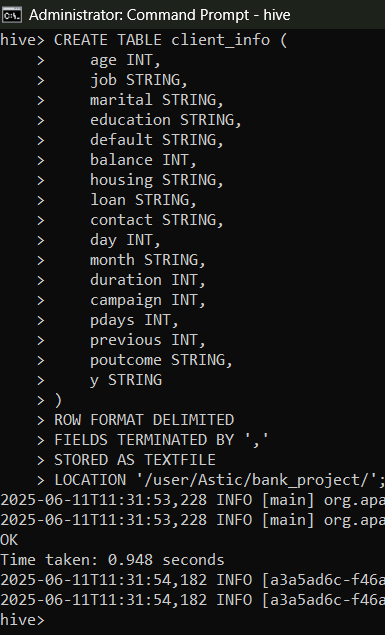
**HIVE**

1. **Data Ingestion and Table Creation**:
   * Create a Hive database named **banking\_data**.

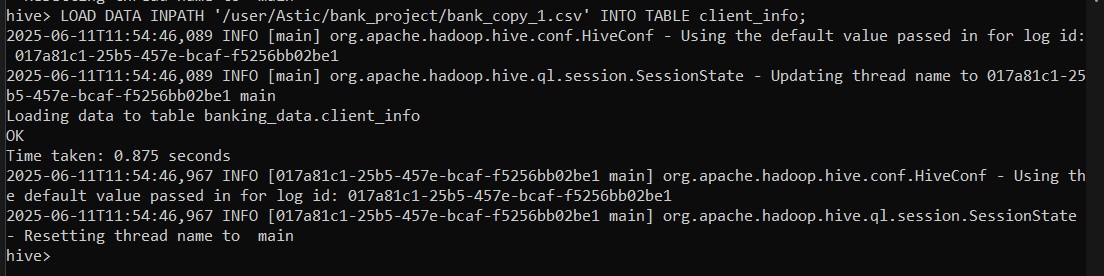


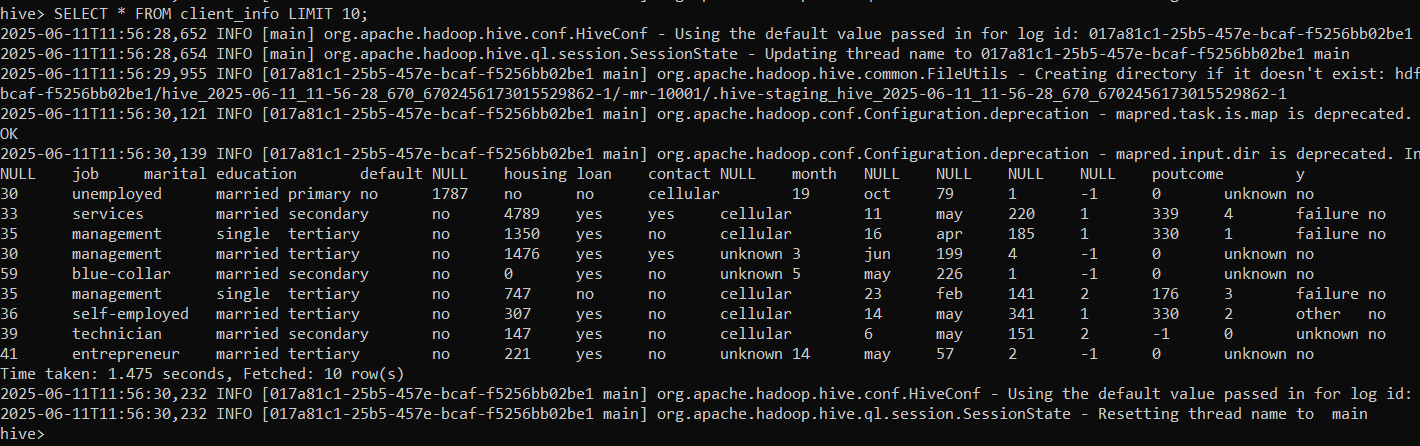


* + Define and create a Hive table **client\_info** with appropriate data types for the **bank.csv** dataset.



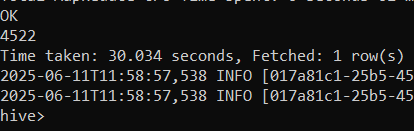
* + Load the data from the **bank.csv** file into the **client\_info** table.



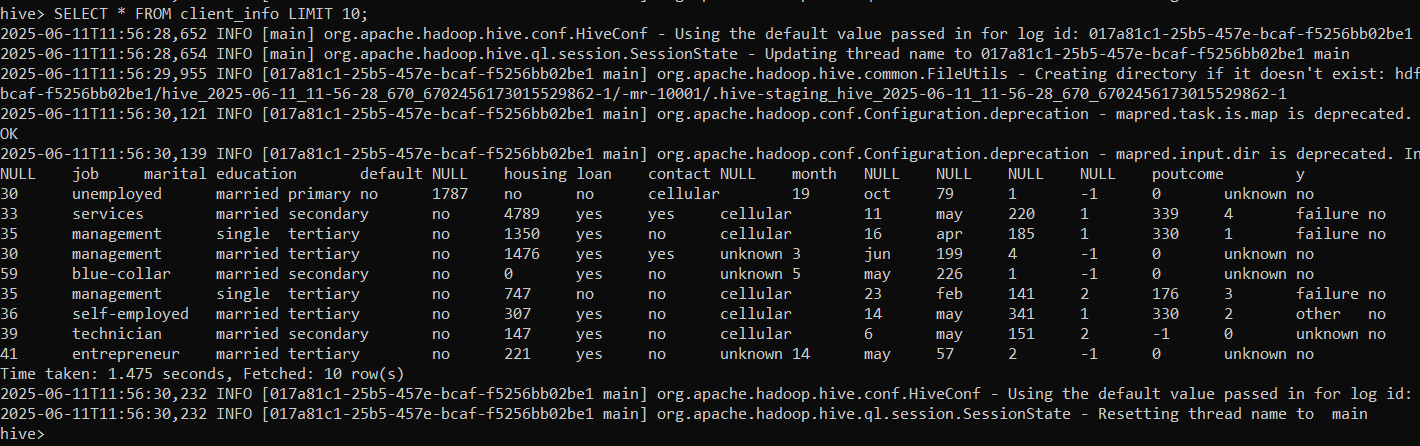


1. **Basic Data Exploration**:
   * Write a HiveQL query to count the total number of clients in the dataset.



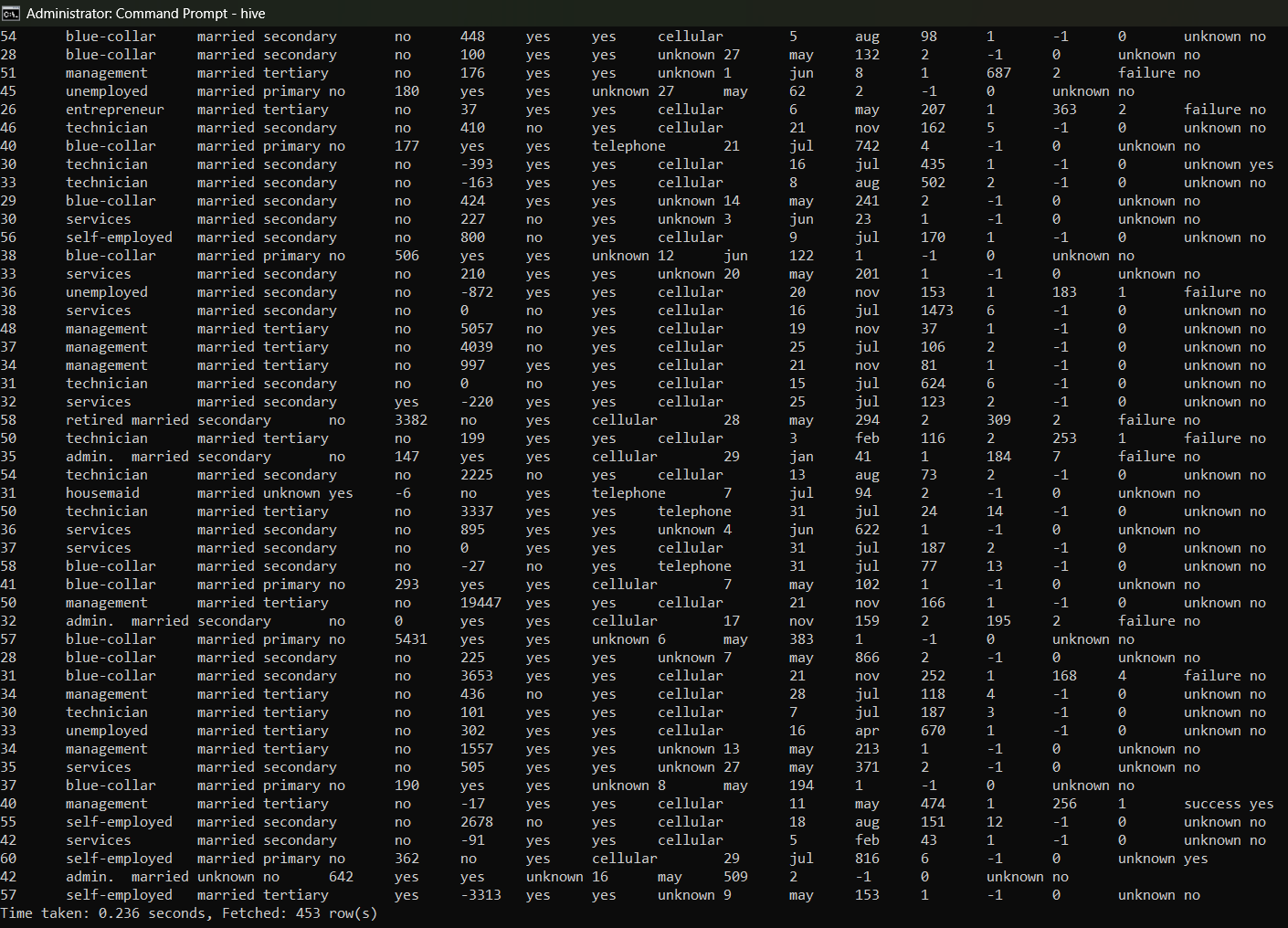


* + Display the first 10 rows of the dataset.

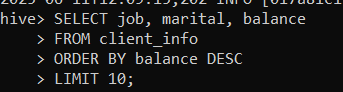


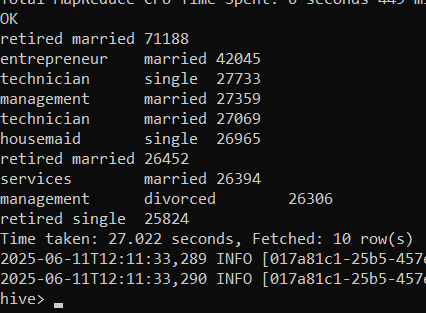
1. **Data Filtering and Sorting**:
   * Retrieve all records of clients who are married and have a personal loan.



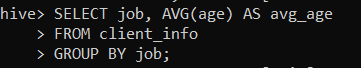


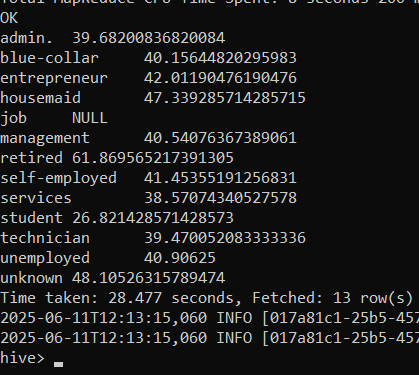
* + List the top 10 clients with the highest balance, displaying their job, marital status, and balance.





1. **Data Aggregation and Grouping**:
   * Calculate the average age of clients for each job category.

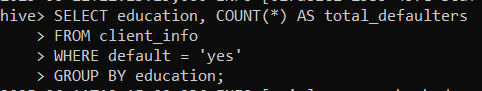


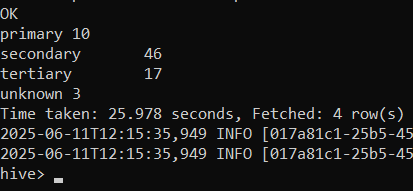


Summary:

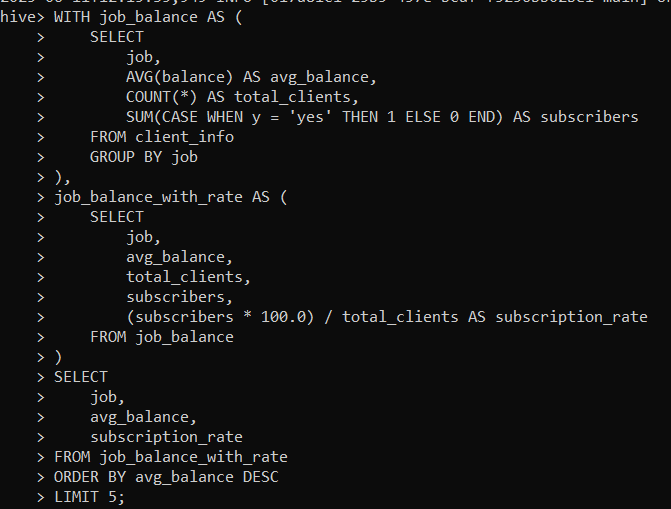
● Average Age by Job Category: The query outputs the average age of clients for each job category. This provides insight into the typical age of clients in different professions. Here, we can see that the average age of the majority of clients for different job categories is between 35 to 45.

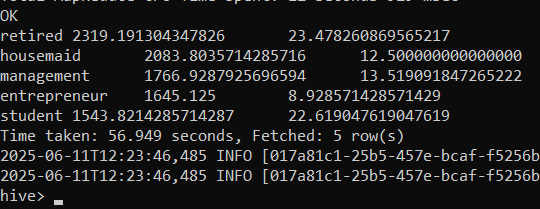
* + Find the total number of clients for each education level who have defaulted on credit.



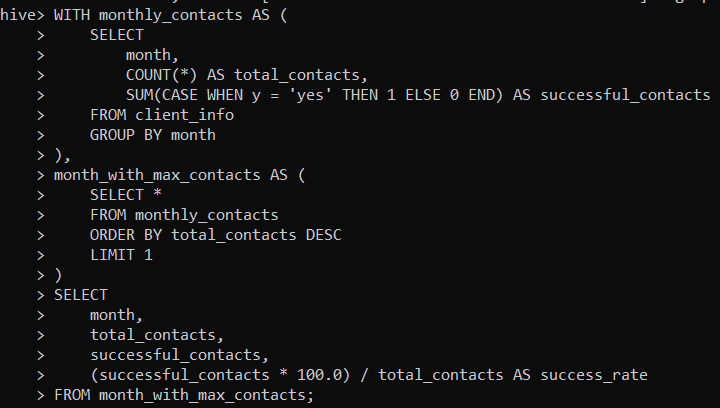


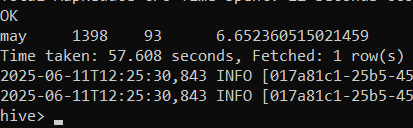
1. **Complex Queries for Insights**:
   * Identify the top 5 job categories with the highest average balance and the percentage of clients in each of these job categories who have subscribed to a term deposit.





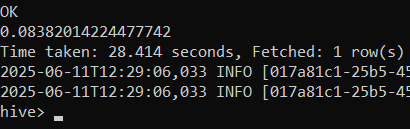
* + Determine the month with the highest number of contacts and the success rate of the campaign in that month (percentage of clients who subscribed to a term deposit).



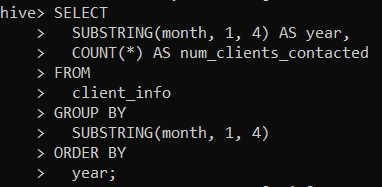


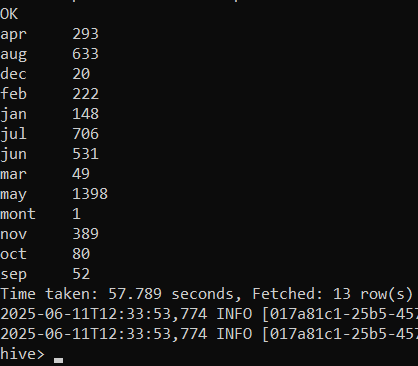
1. **Correlation Analysis**:
   * Calculate the correlation between age and balance for the clients.



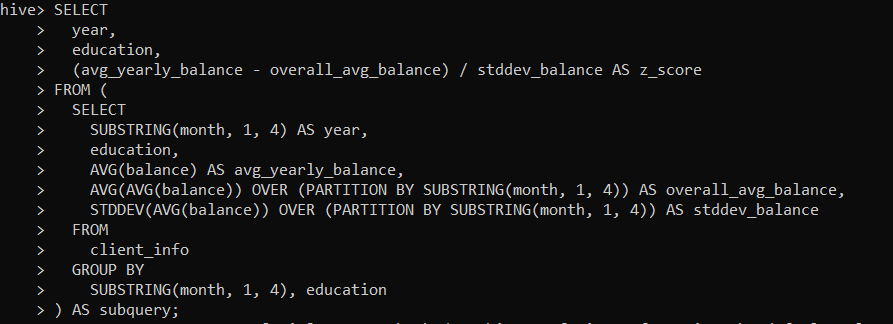


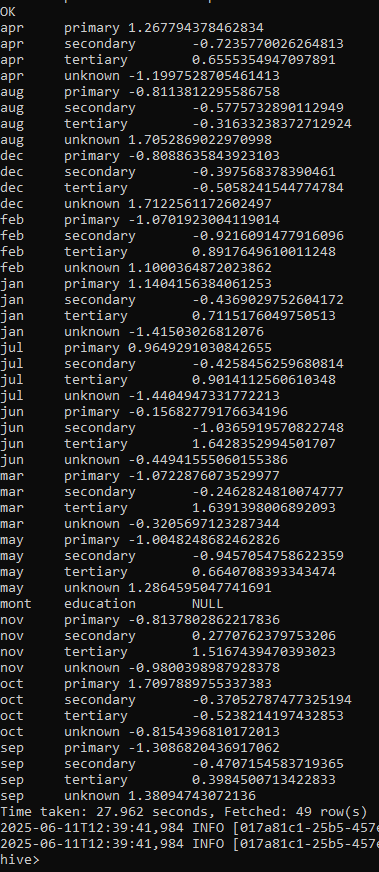
1. **Trend Analysis**:
   * Analyze the year-over-year trend in the number of clients contacted.



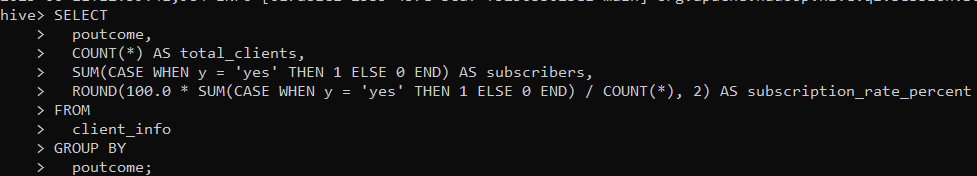


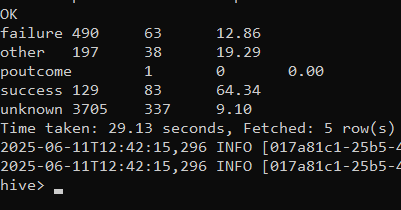
1. **Anomaly Detection**:
   * Identify any unusual patterns in the average yearly balance across different education levels.



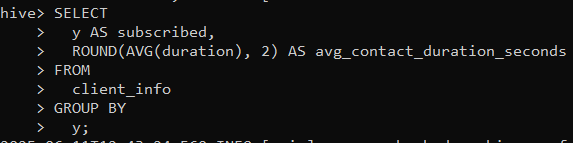


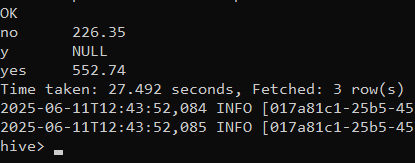
1. **Advanced Analysis**:
   * Analyze the impact of previous campaign outcomes (**poutcome**) on the current campaign's success. Calculate the subscription rate (to term deposits) for each **poutcome** category.





* + Compare the average contact duration for clients who subscribed and who did not subscribe to a term deposit.





**Submission Guidelines:**

* Make a copy of this doc file.
* Perform the analysis in your local system using Hadoop and Hive and provide screenshots of both the **code** and the **output** under each question.
* Upload the doc file with other files and submit it in the submission dashboard.