**PySpark Case Study: Online banking**

**Done by: S.YAZHINI**



**Introduction:**

Online banking plays a crucial role in managing financial transactions effectively and efficiently. With the vast amount of data generated daily by banking operations, analyzing it provides valuable insights for decision-making.

This project leverages advanced tools like Apache Spark to process large datasets related to loans, credit cards, and transactions, enabling deep insights into customer behaviors and banking operations.

**Background:**

The banking sector generates diverse datasets encompassing loans, credit cards, and transactions. To unlock the potential of this data, tools like Apache Spark, known for its high-speed processing capabilities. For this project, datasets were sourced from Kaggle, a renowned platform for data science projects. By cleaning and analyzing the data, this project demonstrates how modern tools facilitate quick and insightful analysis of financial data.

**The Problem/Goal:**

The primary goal of this project is to extract actionable insights from large-scale datasets to aid banking institutions in better understanding their operations and customer behaviours. Key problems addressed include:

* Identifying patterns in loan categories and customer income levels.
* Recognizing customers with high-risk behaviours (e.g., frequent returned cheques).
* Evaluating customer eligibility for credit cards and other financial services.
* Analysing transaction behaviours to identify trends in deposits and withdrawals.

**Benefits:**

* **Enhanced Decision-Making**: Provides actionable insights into customer behaviours and risk patterns.
* **Improved Customer Targeting**: Identifies high-value customers eligible for additional financial products.
* **Operational Efficiency**: Streamlines data analysis through powerful tools like Apache Spark.
* **Risk Mitigation**: Highlights high-risk customers based on income, cheque returns, and spending patterns.
* **Customer Retention**: Personalized offers for active and eligible customers improve satisfaction and loyalty.

**Proposed Solution:**

The solution involves analysing the three datasets (loans, credit, and transactions) using Apache Spark (Pyspark). Key tasks include:

1. **Data Cleaning**: Ensuring consistency and handling missing or invalid entries in datasets.
2. **Loan Analysis**:
   * Counting loans by category.
   * Identifying high-value borrowers (e.g., loans > 1 lakh).
   * Highlighting customers with income above ₹60,000 and those with high expenditures.
3. **Credit Card Analysis**:
   * Determining the number of credit card users in specific regions (e.g., Spain).
   * Evaluating bank members active and eligible for credit cards.
4. **Transaction Analysis**:
   * Extracting maximum and minimum deposit/withdrawal amounts.
   * Summarizing account balances and transaction counts by date.
   * Identifying customers with withdrawals exceeding ₹1 lakh.

**Tools and Techniques Used:**

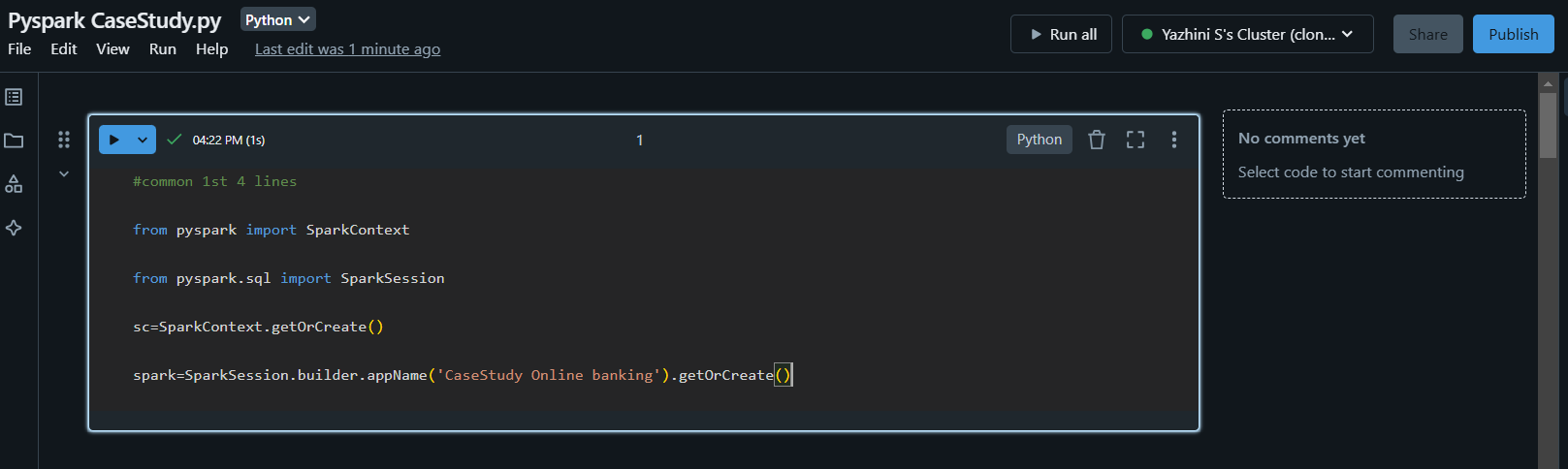
1. Data Bricks
2. PySpark with clusters in Data Bricks

**Datasets used:**

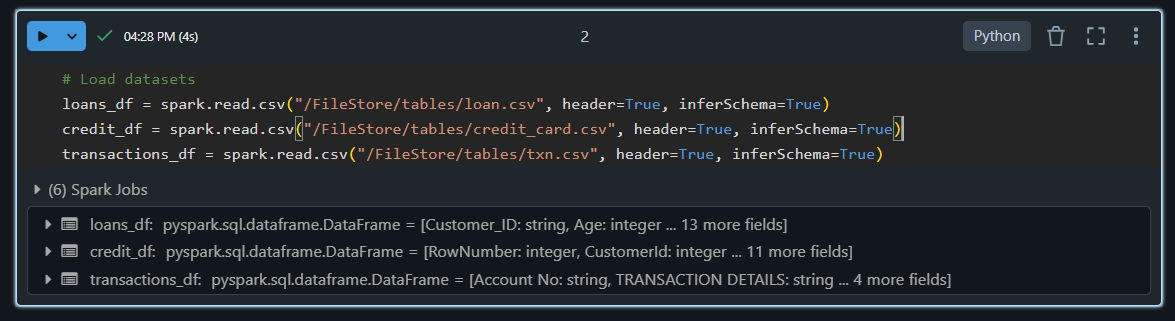
1. [Loan.csv](https://drive.google.com/file/d/1Z1xkaAfOkjLFWqGxducj4L__6i-SKLt5/view?usp=drive_link)
2. [Credit card.csv](https://drive.google.com/file/d/101jTYv7Jh0rPJE7Q971b_yRbmbzC6F39/view?usp=drive_link)
3. [Txn.csv](https://drive.google.com/file/d/1Rkjrbfz-MPm1Rxgaq1p4iFe3hFJH5SCG/view?usp=drive_link)

**Results:**

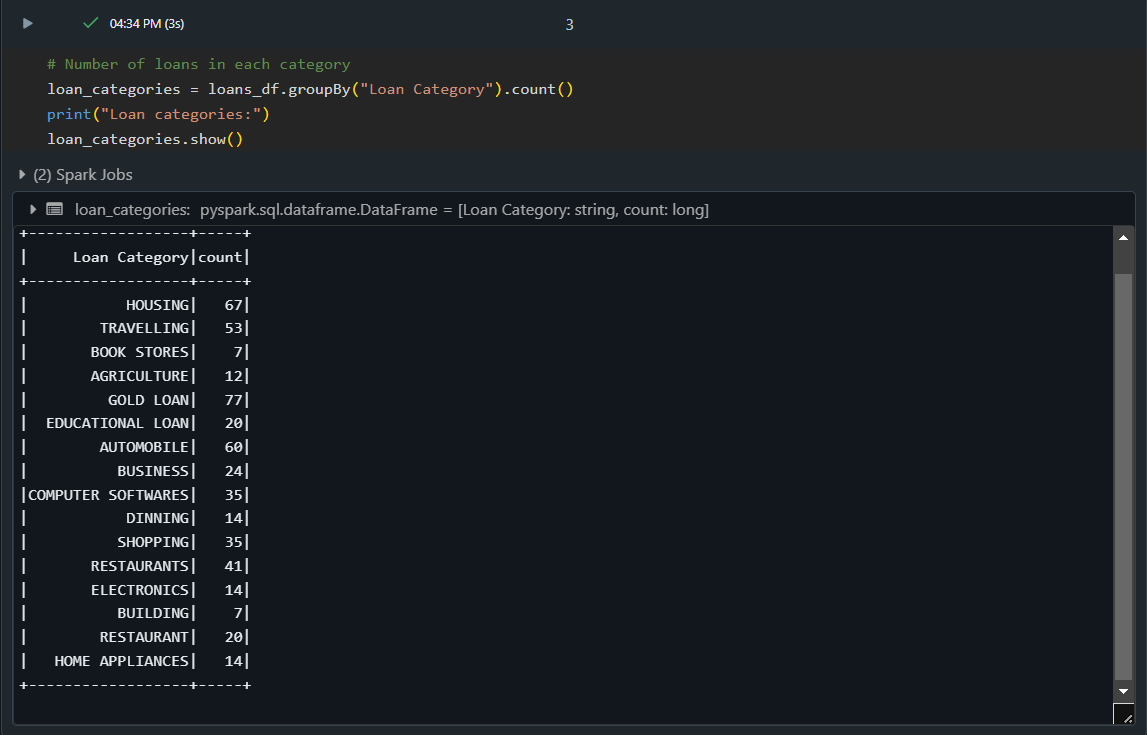
**#Initialising the Spark**

****

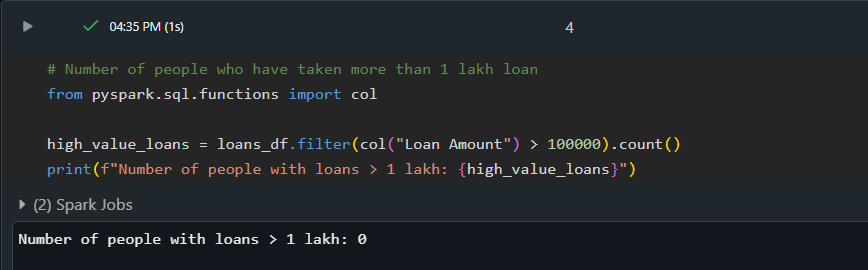
**# Load datasets**



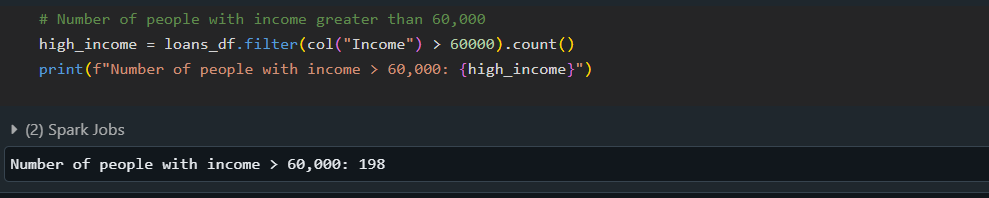
**# Number of loans in each category**

****

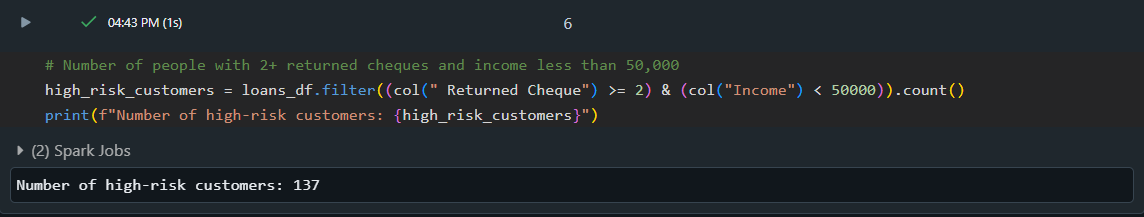
**# Number of people who have taken more than 1 lakh loan**

****

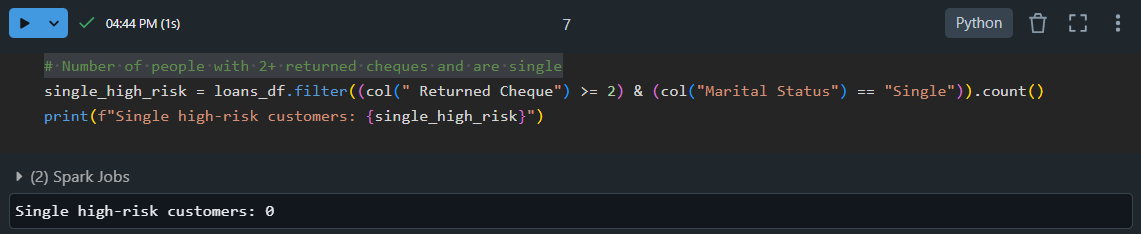
**# Number of people with income greater than 60,000**

****

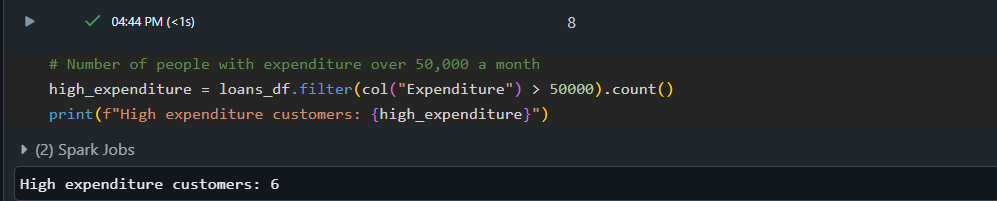
**# Number of people with 2+ returned cheques and income less than 50,000**

****

**# Number of people with 2+ returned cheques and are single**

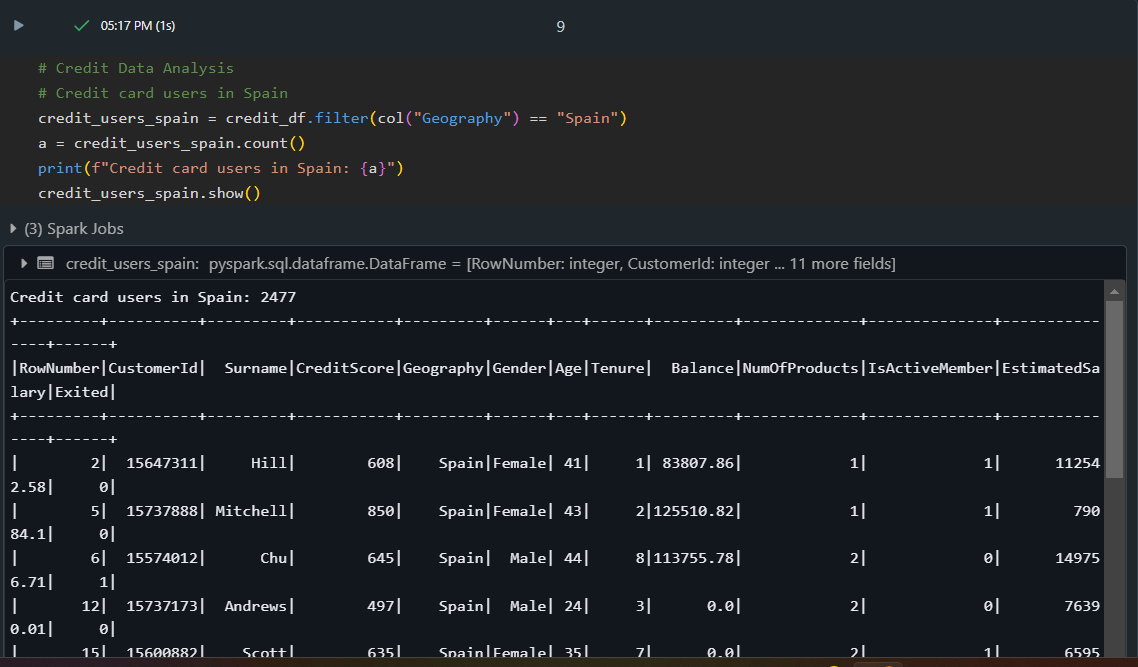
****

**# Number of people with expenditure over 50,000 a month**

****

**# Credit Data Analysis**

**# Credit card users in Spain**

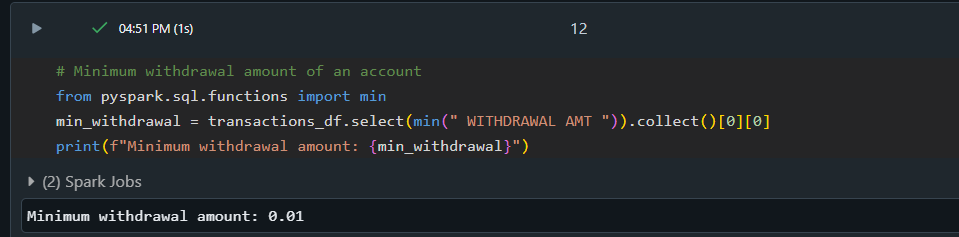
****

**# Transactions Data Analysis**

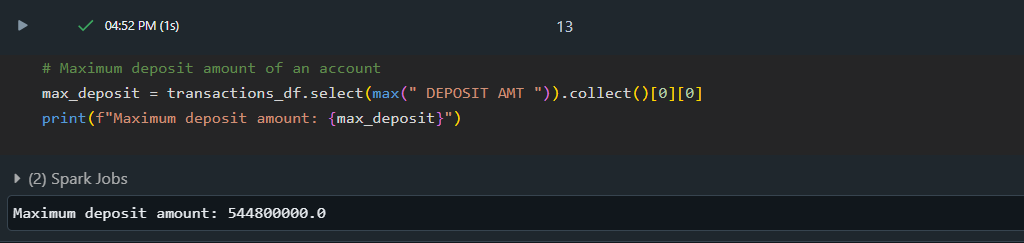
**# Maximum withdrawal amount in transactions**

****

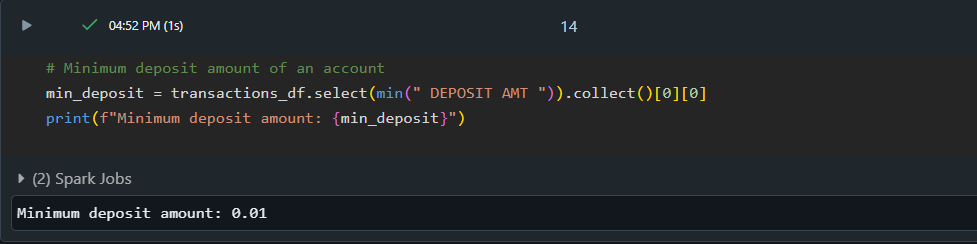
**# Minimum withdrawal amount of an account**

****

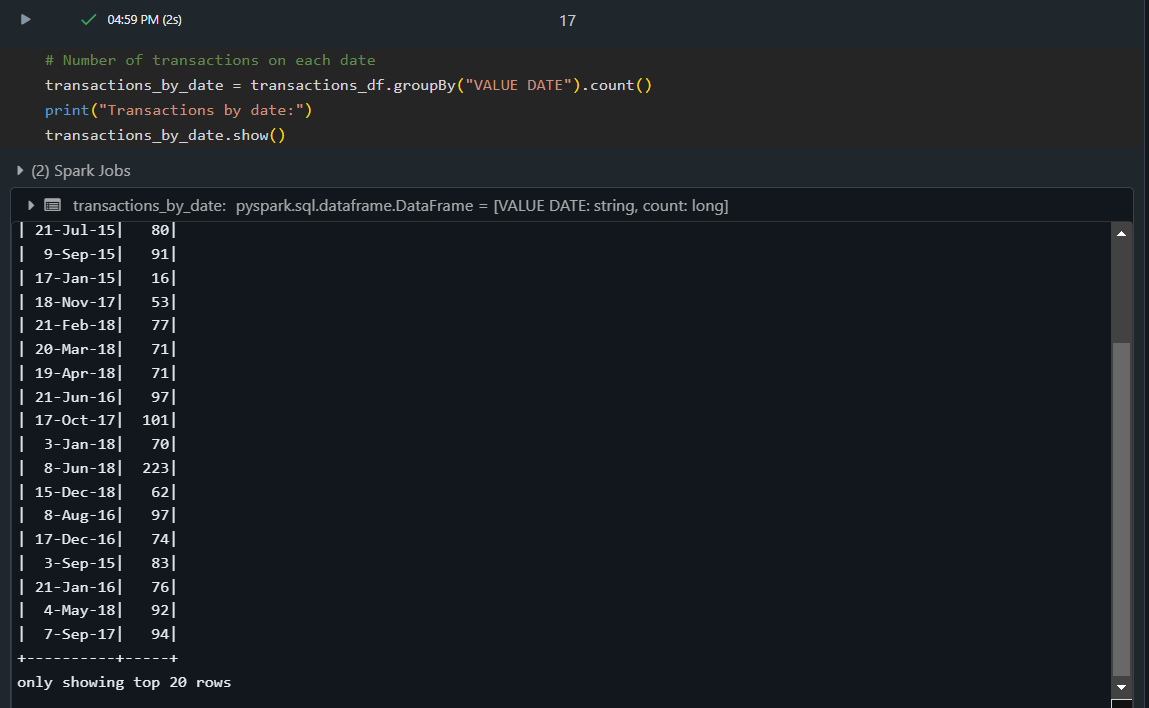
**# Maximum deposit amount of an account**

****

**# Minimum deposit amount of an account**

****

**# Number of transactions on each date**

****

**CONCLUSION:**

The Online Banking Analysis project demonstrates how advanced data processing frameworks like Apache Spark can be leveraged to analyse large datasets in the banking sector. By utilizing Spark's powerful data processing capabilities and Hive for structured querying, the project successfully extracts valuable insights from loan, credit, and transaction data.

Key findings from the analysis include:

* Identification of loan patterns based on loan amounts, customer income, and expenditure, helping to highlight high-value borrowers and those at risk due to frequent cheque returns or high spending.
* Analysis of credit card eligibility and activity, which provides insights into customer engagement and targeting opportunities.
* Transaction data analysis revealed critical metrics such as maximum/minimum withdrawal and deposit amounts, along with patterns in large transactions, enabling banks to monitor high-risk behaviours.

The ability to process and analyse large volumes of data quickly not only enhances operational efficiency but also supports better decision-making, risk mitigation, and targeted marketing efforts. Ultimately, this project illustrates the power of modern data processing tools in extracting actionable insights from financial data, improving customer relationships, and optimizing banking operations.