Graded Assignment: Data Engineering & ETL Case Study – *Retail Banking Transactions*

Overview

A retail bank has provided transactional data covering **customers, accounts, transactions, branches, and credit cards** for the past few years. The goal is to integrate these datasets and answer key business questions using **PySpark DataFrame APIs**. This assignment tests your ability to work with multiple datasets, perform ETL, and apply analytical queries.

You are required to submit the **IPYNB** or **HTML** file with detailed steps, code, and outputs.

Datasets

- **customer_dim** → Customer details
- **account_dim** → Account details (savings/current)
- **transaction_fact** → Debit/Credit transactions
- **branch_dim** → Branch information
- **card_dim** → Credit/Debit card information

Questions (100 Marks Total)

Q1. (10 Marks)
Load all the above files into Spark DataFrames using **SparkSession**. Print schema of each DataFrame.

Q2. (20 Marks)
Join all the DataFrames and create a new DataFrame called **Bank_FullData** such that duplicate columns are removed.

Q3. (10 Marks)
Convert the **Transaction_Date** column into DateType. Print schema and display top 5 records with the converted date column.

Q4. (10 Marks)
Find the **top 5 customers** who have done the **highest total transaction amount**.

Q5. (10 Marks)
Create a new column **Transaction_YearMonth** in the format `YYYY-MM` from the Transaction_Date. Display first 10 rows.

Q6. (10 Marks)
Find the **customer who has made the maximum number of transactions using Credit Card**.

Q7. (10 Marks)
Using a **Window function**, calculate the **running total of transactions per account** (ordered by transaction date).

Q8. (10 Marks)
Count how many **unique customers opened accounts in 2018** and how many of them are **still active in 2021**.

Q9. (10 Marks)
Find the **top 3 branches** with the highest average transaction amount in 2020.

Q10. (10 Marks)
Save the output of Q9 as a file named **branch_avg_txn_2020.json**.

Marking Scheme

- Q1: 10 Marks
- Q2: 20 Marks
- Q3–Q10: 10 Marks each
  Total: 100 Marks
