# **Analytics Engineering Data Challenge**

You have 3 different files, each containing different information:

* user information,
* transactional information,
* package delivery information.

Based on these files, do the following:

1. Explore the three files and, based on your findings, write a data dictionary for each file and a UML diagram. **Done**
2. Create a merged file that consolidates the information. **Done**

Make sure that:

* + Date formats are standardized to YYYY-MM-DD,
  + Timestamp fields have the Mexico City time zone,
  + Numeric fields have two decimal places,
  + Text strings do not exceed 256 characters.

1. Design a DAG for the production deployment of your merged table creation. Based on that, answer the following questions:
   * Would you choose ETL or ELT? Why? **I chose ELT because first it was necessary to load the files into a DataFrame; then, the transformations were easy due to the availability of the data.**
   * Which part of the process would you emphasize most? Why? **The cleaning is the most important in this way, data integrity is ensured, along with certainty and confidence that the data is correct.**
2. Write SQL queries to answer the following:
   1. How many users are there in total? **10,000**

**query = "SELECT COUNT (user\_id) AS totalUsers**

**FROM mergedFileSql"**

**result = pd.read\_sql\_query(query, conn)**

**print(result)**

* 1. How many users have transacted? **10,000**

**query = """**

**SELECT COUNT(DISTINCT user\_id) AS transactions**

**FROM mergedFileSql**

**WHERE transaction\_id IS NOT NULL**

**"""**

**result = pd.read\_sql\_query(query, conn)**

**print(result)**

* 1. What is the card delivery rate? **24.82%**

**query = """**

**SELECT**

**(COUNT(CASE WHEN delivery\_status = 'Delivered' THEN 1 END) \* 100.0 / COUNT(\*)) AS delivery\_rate**

**FROM**

**mergedFileSql**

**"""**

**result = pd.read\_sql\_query(query, conn)**

**print(result)**

* 1. Which is the most efficient package carrier? **DHL**

**query = """**

**SELECT**

**courier, COUNT(\*) AS Deliveries**

**FROM**

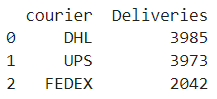
**mergedFileSql**

**GROUP BY courier**

**"""**

**result = pd.read\_sql\_query(query, conn)**

**print(result)**

****

* 1. Which are the top 10 merchants and categories with the most transactions?

**query = """**

**SELECT**

**name, transaction\_type, amount**

**FROM**

**mergedFileSql**

**ORDER BY amount DESC**

**LIMIT 10;**

**"""**

**result = pd.read\_sql\_query(query, conn)**

**print(result)**

