VENTRA IT

# Программируем ваш успех

## Тестовое задание для Junior Data Engineer

Вы работаете с данными крупной табачной компании. Маркетинговая команда заинтересована в анализе продаж и клиентской активности в различных магазинах компании.

У вас есть три таблицы с данными:

1. stores — информация о магазинах:
   - store_id (INT) — уникальный идентификатор магазина.
   - city (VARCHAR) — город, в котором расположен магазин.
   - state (VARCHAR) — регион, где находится магазин.

2. customers — информация о клиентах:
   - customer_id (INT) — уникальный идентификатор клиента.
   - name (VARCHAR) — имя клиента.
   - signup_date (DATE) — дата регистрации клиента.
   - store_id (INT) — идентификатор магазина, к которому относится клиент (где он чаще всего совершает покупки).

3. transactions — информация о транзакциях:
   - transaction_id (INT) — уникальный идентификатор транзакции.
   - customer_id (INT) — идентификатор клиента, совершившего транзакцию.
   - store_id (INT) — идентификатор магазина, в котором была совершена транзакция.
   - transaction_date (DATE) — дата транзакции.
   - category (VARCHAR) — категория товара, связанная с транзакцией (например, "Табак", "Аксессуары").
   - amount (DECIMAL) — сумма покупки.

## Задание

1. SQL часть
   - Определите общее количество транзакций, совершенных клиентами в каждом магазине в 2023 году. Выведите store_id, city, state и transactions_count.
   - Найдите всех клиентов, которые зарегистрировались в 2023 году и сделали хотя бы одну покупку за первый месяц после регистрации. Выведите customer_id, name и количество транзакций (transactions_count).
   - Выведите три категории товаров с наибольшей общей суммой продаж (total_sales) за 2023 год, указав category, total_sales.

2. Python часть
   - Напишите Python-скрипт, который:
   - Создает базу данных SQLite и создает в ней три таблицы (stores, customers, transactions), используя предоставленные ниже данные.
   - Выполняет описанные SQL-запросы и сохраняет результаты в три CSV-файла: store_transactions_2023.csv, new_customers.csv, top_categories.csv.
   - Реализуйте функцию analyze_sales_growth, которая:
   - Принимает DataFrame с суммой продаж по месяцам за 2023 год (например, из top_categories.csv).

- Возвращает DataFrame с колонками month и growth_rate, где growth_rate — процентное изменение суммы продаж по сравнению с предыдущим месяцем.

**Данные для заполнения таблиц**
**tobacco_company_data.xlsx**

## Вопросы по Airflow

- Что такое DAG?
- Назовите три главных компонента архитектуры Airflow?
- Можно ли менять код DAG'a прямо в веб версии приложения?