# Data Engineer Home Assignment

### Estimated Time: 5–7 Hours

### Deliverable Format: GitHub Repository

---

## Objective

Design and implement a data pipeline using Python, Trino, and PostgreSQL.

This assignment evaluates your skills in:

- Building modular, production-ready Python code
- Fetching and validating API data
- Managing incremental loads
- Deploying and querying a Trino environment with PostgreSQL
- Writing analytical SQL over both raw and aggregated tables

---

## API to Use

**SpaceX Launches API**
https://api.spacexdata.com/v4/launches/latest

You are required to **only fetch the *latest* data** from the API. Simulate this as if it's a real-time incremental source, e.g., via scheduled batch.

---

## What You Need to Build

You are expected to:

### 1. Deploy a Local Data Stack

- Use **Docker Compose** to spin up:
    - **Trino**
    - **PostgreSQL**
- Configure Trino as the query engine for PostgreSQL as a data source

## 2. Build a Python Ingestion Script

Create a script or module that:

- Fetches the **latest launch** data from the API
- Parses and validates the data
- Inserts the launch into a **raw table** in PostgreSQL
-

  This table should be **append-only** and support **incremental ingestion**

## 3. Create and Maintain an Aggregation Table

Your pipeline should also:

- Generate an **aggregated table** in PostgreSQL with metrics like:
  - Total launches
  - Total successful launches
  - Average payload mass
  - Average delay between scheduled and actual launch times

The aggregation logic should be in Python or SQL and kept **up to date** when new data is ingested.

# SQL Exercises

After loading the data and aggregations, write **SQL queries** to answer:

## 1. Launch Performance Over Time

How has the success rate of launches evolved year over year?

## 2. Top Payload Masses

List the top 5 launches with the heaviest total payload mass.

## 3. Launch Delay Breakdown

Show average and max delay (in hours) between scheduled and actual launch times, grouped by year.

### 4. Launch Site Utilization

How many launches have occurred at each launch site, and what's the average payload per site?

## Deliverables

Please submit a link to a GitHub repository with the below structure.

- Source code in a `src/` directory
- Docker Compose files in a `docker/` directory
- SQL scripts in a `sql/` directory
- A `README.md` file including:
  - Setup instructions
  - Design choices and assumptions
  - How to test and run the ingestion and aggregation

If you need to explain any assumptions or discuss limitations, include that in the `README.md`.

If you have any concerns and feel you don’t understand the assigmnet you can contact - ofir@zeronetworks.com