

Candidate Take-Home Task

Overview

Welcome! This challenge is designed to assess your engineering skills while giving you the flexibility to showcase your strengths.


You'll be provided with a dataset containing 10,000 JSON files. Some are clean and structured, while others are messy and inconsistent. Your task is to analyze, process, and derive insights from this data—how you do it is up to you!

This is a time-boxed exercise—expect to spend a few focused hours over the next three days. The goal is *not* to build a perfect solution but to demonstrate your problem-solving approach, engineering skills, and decision-making process.

The Data

Each JSON file represents a user in an advocacy platform, containing information such as:

- User metadata (ID, name, emails, social handles, etc.)
- Advocacy programs they participated in
- Social media analytics (likes, comments, shares, reach, etc.)
- Sales attribution data tracking purchases influenced by posts

 However, not all data is perfect:

 Some data is clean and well-structured

 Some data contains inconsistencies, missing values, incorrect types, and formatting errors

The data is available from here:

<https://drive.google.com/file/d/18Kyzbjw5FAAvxqd254d544xBltjhrK67/view?usp=sharing>

Your Task

This challenge is open-ended—you decide how to approach it. Here are some possible focus areas:

Data Exploration & Analysis

- What does the dataset look like?

- What inconsistencies, anomalies, or patterns do you observe?

2 Data Processing & Engineering

- Handle missing fields, incorrect typing, and formatting issues.
- Normalize and structure the data efficiently.
- Standardize identifiers (social media handles, email formats, timestamps).

3 Software Engineering & System Design

- Design and implement a robust ETL pipeline to process and store clean data.
- Build a scalable backend API to access processed data easily.
- Implement unit tests and validation for data integrity.

4 Insights & Visualizations

- Compute metrics like total engagement per user, platform, or program.
- Identify top advocates based on activity, influence, or conversions.
- Discover patterns, trends, or outliers within sales attribution data.

5 Bonus (If You Have Time) 🚀

- Performance optimization for handling large JSON datasets.
- Data visualization (charts, dashboards, reports).
- Automated pipeline/workflow with CI/CD support.

We don't expect you to cover everything—instead, focus on what best showcases your skills. 🚀

Tech Stack & Deliverables

We encourage you to use technologies you are comfortable with. Possible tools include:

- ♦ Programming languages: Python, JavaScript (Node.js), Java, Go, etc.
- ♦ Frameworks: Flask, FastAPI, Django, Spring Boot, Express.js, etc.
- ♦ Data tools: Pandas, SQL, Spark, Airflow, dbt, etc.
- ♦ Databases: PostgreSQL, MongoDB, Redis, BigQuery, etc.

Our Tech Stack

As a reminder, our tech stack is:

- ♦ Programming languages: TypeScript w/Node.js
- ♦ Frameworks: Angular 19
- ♦ Data tools: SQL, Airflow, Snowflake
- ♦ Databases: MongoDB, Redis, ElasticSearch
- ♦ Cloud: AWS, Heroku

What You Should Submit

Your GitHub Repository must contain:

1. Your code: Scripts, services, or notebooks
2. A `README.md` file explaining:
 - Your approach and thought process
 - Challenges & assumptions made
 - Steps to set up and run your solution
 - Any sample output, visualizations, or insights (if applicable)

We will run and review your solution, so make sure it's easy to set up!



Timeline



You have 3 days to complete this challenge.



Submission is required 24 hours before the review session.



In the review meeting, you will walk us through:

- Your approach to handling structured vs. messy data
 - How your process works & key technical decisions
 - Any challenges & trade-offs you encountered
 - What you'd improve with more time
-



Evaluation Criteria

We're not looking for a perfect solution—we want to see how you think and work through problems.

We'll evaluate:

- ✓ Handling of messy & structured data
 - ✓ Problem-solving & decision-making
 - ✓ Code quality, structure, and readability
 - ✓ Scalability and performance considerations
 - ✓ Ability to extract insights & communicate findings
-



How To Submit



Share your GitHub repository link before the deadline.



Ensure we can easily run your solution with provided instructions.



Questions?

If anything is unclear, feel free to reach out before submitting!

Good luck, and we look forward to seeing your approach! 🎉