Verihubs

Technical Test
Full Stack Data Scientist

PT Verihubs Inteligensia Nusantara
2025

Submission Instructions

- Submit your work in the form of a GitHub repository and share the link via WhatsApp to Sinta Kartika (+6282154016032).
- You have 7 days to complete the assignment, starting from the date this document is provided.
- The submission can be in the form of a Jupyter Notebook or DuckDB’s built-in notebook, as long as all required files and documentation are included in the GitHub repository.
- For any questions or clarifications regarding the task, please contact Sinta Kartika via WhatsApp before the submission deadline.

Full Stack Data Scientist Test

You are assigned to work on an eCommerce dataset to analyze and optimize sales performance. Your task is to build a data pipeline that extracts, transforms, and loads (ETL) sales data, generates meaningful insights, and visualizes key business metrics. This test aims to evaluate your technical capability in data engineering, analysis, and visualization, reflecting real-world scenarios in Verihubs' data-driven decision-making process.

Dataset

- You will be working with the following dataset:
    - Amazon Sale Rep (Kaggle). URL: https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data (Download **Amazon Sale Report.csv**)

Tools & Environment Requirements

1. Python 3.8 or above
2. Dagster
3. DuckDB
4. Pandas, Matplotlib/Seaborn (or equivalent visualization library)
5. Jupyter Notebook (optional)

Tasks

1. Setup Environment
    a. Create a Dagster environment that:
        i. Loads the dataset from the provided CSV file.
        ii. Inserts the loaded data into a DuckDB database.
2. Data Modeling
    a. Create a DuckDB table that:
        i. Tracks monthly total revenue divided by product category (assume all transactions are successful).
    b. Create another DuckDB table that:
        i. Tracks the number of daily orders, divided by order status.
3. Data Visualization (Python)
    a. Generate a visual chart showing daily orders by status.
    b. Identify and display which month is the most profitable based on total revenue.

Expected Deliverables

- A fully functional Dagster project (with clear folder structure and code comments).
- DuckDB database with the two required tables.
- Python script or notebook for visualization.
- README.md explaining:
    - Setup instructions.
    - How to run the pipeline.
    - Summary of findings (including most profitable month).

Evaluation Notes

The assessment will focus on both functionality and best practices. Candidates are encouraged to write clean, modular, and well-documented code that reflects production-level standards.

Scoring Matrix

- **Correctness & Functionality**
  - Accuracy of pipeline results, correctness of queries, and reliability of data outputs.
- **Dagster Best Practices & Structure**
  - Proper use of ops/jobs, repository setup, and modular project organization.
- **Visualization & Output**
  - Clarity, readability, and insightfulness of charts and final analysis.
- **Code Quality & Readability**
  - Code is clean, well-organized, and properly documented.
- **Bonus**
  - Creative approaches, performance optimization, or additional meaningful insights beyond the main task.