Coinbase Analytics Engineering

Intern Assessment

11-25-2025

Overview

This assessment evaluates your ability to build a data pipeline, analyze data quality issues, and design solutions for production systems. You'll create a working data pipeline and then analyze a real-world data quality problem.

Time Estimate: 4-6 hours
Submission: Submit as a zip file or GitHub repository link

Part 1: Build a Data Pipeline (3-4 hours)

Task

Create a Python project that:

1. **Fetches cryptocurrency price and volume data** from the public Coinbase API for BTC-USD and ETH-USD over the period from 11/17/25 - 11/24/25. The data is available at https://api.exchange.coinbase.com and documentation is available at https://docs.cdp.coinbase.com/api-reference/exchange-api/rest-api/products/get-product-candles
2. **Stores the data** in a DuckDB or SQLlite database with an appropriate schema
3. **Visualizes** the hourly volume and average price for each trading pair

Notes

- Project must be **runnable end-to-end** following steps in README.md
- **No secrets/credentials** in code or repository
- Code should be **well-organized and documented**
- **Feel free to add any additional insights or features to your project**

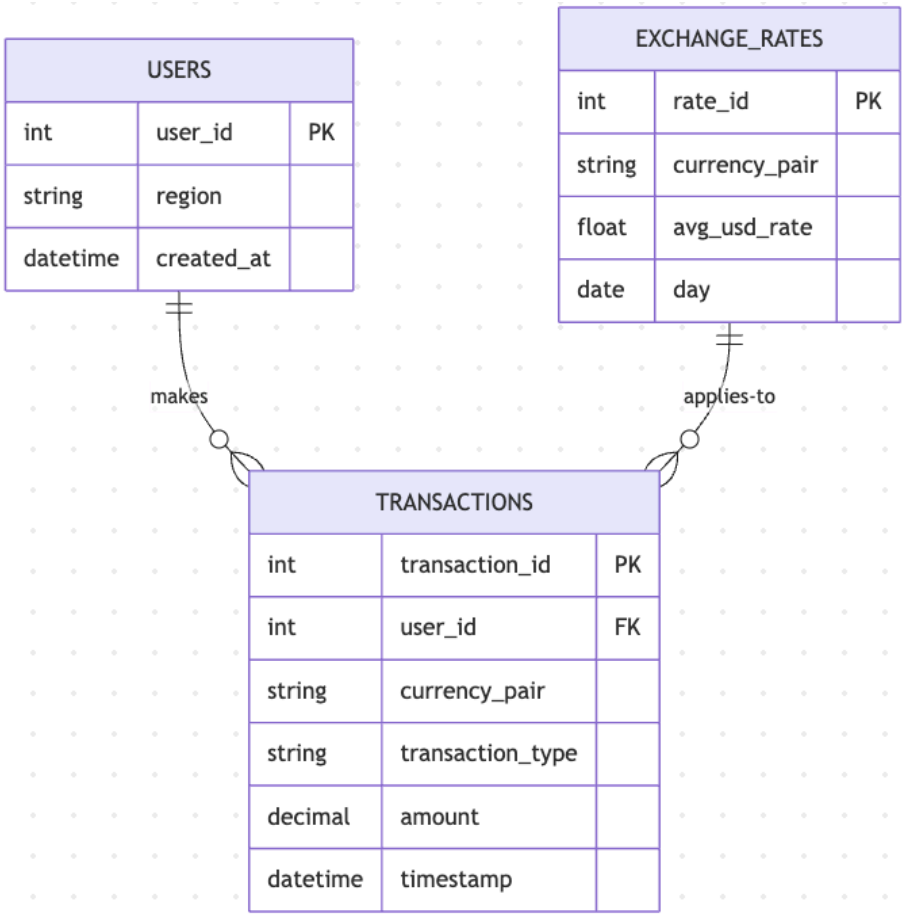
Part 2: Data Quality Analysis (1-2 hours)

Scenario

You're working on Coinbase's internal analytics pipeline. The pipeline aggregates total cryptocurrency trading volume across all supported coins using several internal data sources and presents it to executives in a daily dashboard. It uses complicated logic and sources contrasted with the Coinbase exchange API used in the previous question which is a raw feed from the Coinbase exchange.

Data Model

The internal volume calculation uses the following database schema:



The following SQL query is used to calculate volume internally:

```
None
SELECT day, SUM(transactions.amount * exchange_rates.avg_usd_rate) AS volume
FROM users
JOIN transactions
  ON users.user_id = transactions.user_id
JOIN exchange_rates
  ON transactions.currency_pair = exchange_rates.currency_pair
  AND transactions.timestamp::DATE = exchange_rates.day
WHERE transactions.day BETWEEN CURRENT_DATE()-7 AND CURRENT_DATE()
GROUP BY 1
```

Your Task

Yesterday, Brian Armstrong noticed a discrepancy: the internal dashboard showed $3 billion in trading volume on the Coinbase exchange, but the public Coinbase API showed $2.8 billion for the same time period. Your task is to diagnose the problem and propose solutions.

Answer the following (either in a markdown file or separate document):

1. Root Cause Analysis

What could cause the volume discrepancy?

List hypothetical causes. Consider:

- Data quality issues
- Logical issues
- Data source differences

2. Investigation Plan

How would you systematically diagnose this issue?

Outline your approach:

- What would you check first, second, third? How would you verify or rule out each hypothesis?

3. Prevention Strategy

What data quality checks or monitoring would you add to catch this automatically in the future?

Be specific about:

- What you would validate (e.g., specific validation rules)

- How you would implement it (e.g., automated checks, alerts, dashboards)
- When checks would run (e.g., during ingestion, daily reconciliation)

Submission Format:

- Zip file containing all files, OR
- GitHub repository link (make sure it's accessible)

What We're Looking For:

- Problem-solving approach
- Data engineering fundamentals
- Ability to think through data quality issues
- Clear communication and documentation

Resources

- **Coinbase API Documentation**: https://docs.cloud.coinbase.com/exchange/reference/exchangerestapi_getproductcandles
- **DuckDB Documentation**: https://duckdb.org/docs/
- **UV Documentation**: https://github.com/astral-sh/uv