Data Engineer Quiz

The Data Engineer recruitment quiz comprises 3 distinct parts:

Part 1: Algorithmic Thinking

Please select 1 of 2 problems

Problem 1:

In the Thailand the currency is made up of Baht (฿) and Stang (s). There are six coins in general circulation:

$$25s, 50s, ฿1, ฿2, ฿5, ฿10$$

It is possible to make ฿10 in the following way:

$$1 \times ฿5 \ + \ 1 \times ฿2 \ + \ 2 \times ฿1 \ + \ 1 \times 50s \ + \ 2 \times 25s$$

How many different ways can ฿10 be made using any number of coins?

Problem 2:

The number, 197, is called a circular prime because all rotations of the digits: $197, 971, and\ 719$, are themselves prime.

There are thirteen such primes below 100: $2, 3, 5, 7, 11, 13, 17, 31, 37, 71, 73, 79,$ and 97

How many circular primes are there below one million?

Deliverable 1. Code up brute force solution for 1 of 2 problems (it might not finish in a reasonable time). If you pass the exam stage, prepare to explain possible solutions to improve your current approach in relation to runtime during the panel interview.

Part 2: Technical Skills

Candidates showcase their proficiency in relevant technologies and tools. Topics covered may include database management, ETL (Extract, Transform, Load) processes, and data modeling.

Part 2.1 SQL:

Use data from quiz-de.db as a main source

Instructions:

1. find number of new customers in 1997
2. find number return customers in 1997
   (Return customers is defined as customers who make a purchase again after being inactive for more than 90 days)
3. find the most popular top 5 suppliers in 1997
4. find sales_amount group by product category in 1997 (2 decimal places)
   $$sales\_amount = price * quantity;$$
5. find top 3 employee those have best growth sales performance in 1997

Deliverable 2.1. SQL scripts and all results export in csv files name:

- new_customers_1997
- return_customers_1997
- most_popular_suppliers_1997
- sales_amount_group_by_product_category_1997
- top_employee_growth_sales_performance_1997

Part 2.2 Data Pipeline:

Use data from quiz-de.db as a main source

Instructions:

1. **Data Ingestion:**
   - Extract data from the quiz-de.db store database or export it based on your preferred method.
2. **Data Pipeline Creation:**
   - Develop a data pipeline to load the data into a new database. Ensure compatibility between the source and target databases (e.g., MySQL to Postgres).

3. **Data Cleansing:**
    - For the `suppliers` table:
        - Create a new column called `supplier_contact` containing only numeric values extracted from the `Phone` column. If `Phone` is missing, use 0.
        - Clean the `PostalCode` column to contain only numeric values.
    - For the `shippers` table:
        - Create a new column called `shipper_contact` containing only numeric values extracted from the `Phone` column. If `Phone` is missing, use 0.
4. **Aggregate Table Creation:**
    - Create an aggregate table named `product_sales_amount_by_month`
    - Calculate the `sales_amount` as `price * quantity`.
    - Compute the `percentage_change` using the formula

      $$percentage\_change = ((salesAmount[t] / salesAmount[t-1]) - 1) * 100;$$
      $$where\ t = period$$

    - Format the result to display two decimal places.
    - Table Name: `product_sales_amount_by_month`
    - Fields: `Year-Month, product_id, product_name, amount, and percentage_change`. Format `Year-Month` as `%Y-%m`.
5. **Use Case Design:**
    - The company intends to return money to customers over 50 years of age who have made a purchase in 1996 – 1997. Return rates is 50% of purchase amount.
    - To facilitate this program, a data pipeline can be designed to generate a targeted table containing the relevant customer information.
    - Plan, design and develop data pipeline and target data by yourself.

In addition, please address the following questions:

- If data needs to be ingested periodically, how would you modify your current approach?
- Create a data architecture diagram illustrating the various components of your ETL (Extract, Transform, Load) process.
- Explain your methodology for ensuring the correctness of ingested data.

For additional credit, consider the following:

- **Technology Stack:** Utilize tools such as Talend, Alteryx, Python, or an AWS environment for enhanced functionality.
- **Deployment Options:** Opt for a publicly accessible deployment of your service, or code whether it's on the Cloud, On-premises, or within a Git Repository.
- **Comprehensive Documentation:** Provide thorough documentation to facilitate understanding and maintenance.
- **Testing Framework:** Implement robust tests to validate functionality and reliability.
- **Visual Aids:** Create relevant diagrams to enhance clarity and communication.

Remember, these elements contribute to a well-rounded and effective solution.

Deliverable 2.2. You have been provided with straightforward data. Your objective is as follows:

- **Data Cleansing:** Cleanse the data by addressing any inconsistencies, inaccuracies, or missing values.
- **ETL (Extract, Transform, Load):** Perform the necessary ETL processes to prepare the data for analysis.
- **Data Warehouse Loading:** Load the processed data into your preferred database, serving as the Data Warehouse.

Please provide access to your solutions, code, and documents.

Part 3: Technical and Consulting Skills

Successful technology solutions stem from cohesive engineering teams working in tandem. This collaborative spirit extends to presenting the final product to clients. Deliverable 2 focused on internal solution development.

Deliverable 3: Technical Walkthrough Presentation

This stage challenges you to create a clear and concise technical walkthrough presentation. This exercise serves a dual purpose:

- **Solution Communication:** Effectively convey the key elements of your solution to a technical audience.
- **Structured Thinking:** Organize your thought processes into a well-defined, communicative format.

By crafting a compelling presentation, you not only educate stakeholders but also solidify your own understanding of the solution's intricacies.

The goal is to clearly communicate the following points:

- **Tech Walkthrough**
  - Tech Stack & Architecture: Briefly explain the languages, frameworks, and a high-level diagram showcasing the solution's structure.
  - Key Features: Outline the core functionalities delivered.
- **Decision Points & Challenges**
  - Key Decisions: Highlight any critical choices made during development and the rationale behind them.
  - Challenges Faced: Briefly discuss the technical hurdles overcome.
- **Production Readiness**
  - Gap Analysis: Identify any missing features or functionalities crucial for full operation.
  - Implementation Effort: Estimate the major efforts required to transition to a production environment.

Please also elaborate on your answers to the questions posed in part 2 through the technical walkthrough.

We value:

- Communication
- Reproducibility
- Pragmatism
- Code hygiene