API Data Engineer Skill-Based Assessment

With this skills-based assessment, we are looking to understand how you approach data engineering problems, specifically how you handle JSON data, perform data transformations, and present and communicate findings.

This skill-based assessment is a hypothetical scenario designed to assess a number of skills that we are looking for in candidates for the Data Engineer position, as well as to provide candidates with a sense of what some typical tasks for this position might include.

We anticipate this assessment should take approximately 4-8 hours to complete. Please share your deliverables in the form of a Python code file or notebook and a PowerPoint (or equivalent) presentation within one week of receiving the assessment. You will be required to present on your skills-based assessment during your interview with our office.

Task 1: Download the included CSVs and perform the required processing and analysis tasks.

The City of Syracuse lets residents submit issues related to city services through its non-emergency 311 system, SYRCityLine. Included with this assignment is a CSV of randomized data meant to simulate actual submissions from residents (SCF_SBA.csv) and two CSVs of Service Level Agreements or SLAs (SCF_SLA.csv and Updated_SLA.csv). The City's SLAs denote the timeframe within which the City commits to resolving specific issues and closing the request.

Please complete the following tasks:

1. Load and Inspect the CSV Files:
    a. Read the SCF_SBA.csv file and join it with SCF_SLA.csv file into a pandas DataFrame.
    b. Display the first few rows to understand the structure and content of the data.
2. Rename Columns:
    a. Ensure all column names are clear, descriptive, and follow a consistent naming convention.
3. Datetime Transformation:
    a. Identify columns containing datetime information and convert these columns into datetime format using pandas.
    b. Extract relevant information, such as day, month, year, or time, into separate columns.
4. SLA Analysis:
    a. Determine whether specific time durations fall with the SLAs. The SLA threshold is contained in the Sla in hours column.
    b. Create a new column indicating SLA compliance.
    c. Now take a look at the Updated_SLA.csv file and use it to update the relevant SLAs. See how the City performs at meeting these new commitments.

Please note, these datasets were created for this skills-based assessment and are not real data.

Task 2: Conduct an analysis of Illegal Setouts

An illegal setout occurs when a resident puts out trash that does not conform to the City's rules for trash and sanitation. The Mayor wants to better understand this problem and has asked the API team to look into it. Using the provided data, conduct an analysis of illegal setouts and the SLAs.

Task 3: Create a presentation summarizing your approach, results, and insights.

Please create a presentation that describes the work you did, including data transformations, analysis into illegal setouts, visualizations, and key insights and potential recommendations.

Deliverables:

1. Python code file or notebook showing your data processing steps.
2. Final PowerPoint presentation file.

Tools and Technologies:

- Programming language: Python, PySpark, SQL.
- Libraries: pandas, matplotlib, seaborn (optional for visualizations).
- Other: Microsoft PowerPoint or other presentation software.

Notes:

- You may ask clarifying questions if needed but are expected to work independently.
- Be prepared to explain your approach and decisions during the presentation.