CLARITY

Data QA Engineer Assignment

Objective: Conduct an in-depth data quality analysis for e-commerce data to uncover complex data quality issues. Your findings should be well-documented, including detailed analyses, insights on underlying causes, and forward-thinking recommendations.

Scenario: As a Data QA Engineer, you're assigned to validate and enhance data quality for ecommerce critical analytics data. The business relies on this data for personalized marketing, operational efficiency, and decision-making. The analytics team has recently identified discrepancies in trend analysis and is looking to the Data QA team to uncover the causes.

Tasks:

1. Data Quality Audit with Advanced Insights

   Perform a comprehensive data quality audit to uncover inconsistencies, errors, and anomalies in various quality dimensions, including:

   - Completeness: Missing fields and null values.
   - Accuracy: Unrealistic or outlier values (e.g., prices too low or too high).
   - Consistency: Data format inconsistencies, mismatches between related fields, and unexpected variations within categorical data.
   - Timeliness: Identify if the dataset includes future-dated records or inaccurate timestamps.

Deliverable: A detailed report with identified issues, examples, and percentages of affected records.

2. **Root Cause and Anomaly Detection**

   - Provide in-depth insights into why these issues may be occurring, considering potential sources of errors.
   - Identify any patterns or anomalies in the data (e.g., future dates appearing only in specific categories).

Deliverable: A section in your report with probable root causes and patterns of anomalies, supported by example

3. **Advanced Impact Analysis**

- Explain how data quality issues might impact business decisions and operations, considering aspects such as customer satisfaction and revenue.

- Provide a hypothetical example showing how incorrect data could skew reports.

Deliverable: A short report (2-3 paragraphs) on the anticipated business impact.

4. **Solution Design and Strategy for Data Quality Improvement**

    - Propose solutions and a basic framework for ongoing data quality assurance, including key quality indicators (KQIs).

    - Highlight anticipated challenges in implementing these solutions.

Deliverable: A structured report outlining recommended solutions, proposed framework, KQIs, and anticipated challenges.

5. **Automated Data Validation Script**

Write a script (Python or SQL) to automate basic data quality checks on the dataset:

- Identify missing values and log rows with any `None` or `NaN` values.

- Flag future-dated entries in the `OrderDate` field.

- Detect records with unusual `Quantity` values (e.g., values < 0 or > 100).

- Validate `TotalAmount` by checking if it logically aligns with `Quantity * Price`.

Deliverable: Submit your script with a README explaining how to run it and interpret the output.

Submission Guidelines:

Format: Submit findings and recommendations in a comprehensive PDF report or structured spreadsheet including the scripting code pushed on GitHub.

Deadline: Submit within 2 days.

Evaluation Criteria: Depth of analysis, logical reasoning, innovative solution design, and attention to detail.

Good Luck ☺