Data Engineering - Data Quality Exercise

Objective:

Learn to perform data profiling and quality checks on a dataset using a data platform. This exercise will guide you through the process of profiling a table and creating data quality checks.

Prerequisites:

- Access to the data platform interface
- Familiarity with data profiling and quality concepts
- Dataset: `transactions`

Exercise Steps:

Part 1: Data Profiling **Goal:** Profile the `transactions` table to understand the data distribution.

1. Navigate to the data platform's main page.
2. In the search bar, type "transactions" and select the `transactions` table from the results.
3. Observe that the "Data Profile" and "Data Quality" tabs are initially empty.

Action: Profile the `transactions` table.

4. Go to the "Profile" tab.
5. Click on "Create Data Profile Scan".
6. Name the scan (e.g., `dp-2023-transactions-profile`).
7. Select the `transactions` table.
8. Click "Continue".
9. Choose the "sales" dataset.
10. For the output, enter `profile-transactions` as the table name.
11. Initiate the scan and wait for it to complete (this may take a few minutes).
12. Review the data profile results.

Part 2: Data Quality Check **Goal:** Establish data quality checks for the `transactions` table.

1. Go to the "Data Quality" tab.
2. Click on "Create Data Quality Scan".
3. Name the scan (e.g., `dp-2023-transactions-quality`).

4. Again, select the `transactions` table.
5. Click “Continue”.
6. Add rules by choosing “Built-in rule types”.
    - For NULL checks, apply the rule to critical columns such as `transaction_id`, `customer_id`, and `amount`.
    - For format checks, use a Regex check on the `amount` column to ensure it contains only numeric values (use the regex pattern `^\d+(\.\d+)?$`).
7. Continue to the next step and select the “sales” dataset.
8. For the output, enter `check-transactions` as the table name.
9. Start the scan and wait for it to complete.
10. Once finished, review the data quality results.

Part 3: Verification

1. Return to the search bar and type “transactions” to open the table again.
2. Examine the “Data Profile” and “Data Quality” tabs to see the updated information.
3. In the data platform, locate and review the `profile-transactions` and `check-transactions` tables to understand the profiling and quality check results.

Conclusion:

By completing this exercise, you will gain practical experience in performing essential data quality and profiling tasks, which are critical for ensuring the reliability of data in any data-driven decision-making process.