Project Title: Automated Data Quality Checks Using SQL Server

Monday, 26 May 2025

---

1. Project Overview

This project aims to automate routine data quality checks on enterprise datasets using SQL Server. A stored procedure was developed to run multiple checks across different dimensions (customer, transaction, and time), and systematically log the results into a centralized audit table. This process ensures consistency, traceability, and readiness for dashboard integration or alert mechanisms.

---

2. Objective

- Ensure data accuracy and consistency through automated validations.
- Track the number of failed or flagged records from key business tables.
- Enable scheduled or on-demand execution of checks.
- Support monitoring of data quality trends over time.

---

3. Technology Stack

- SQL Server (T-SQL)
- Stored Procedures
- Common Table Expressions (CTEs)
- SQL Server Agent (for scheduling) – *optional*
- Power BI / Tableau (for visualization) – *optional*

---

4. Methodology

A stored procedure named DATA_AUTOMATION3 was created. It performs the following operations:

1. **Checks for Duplicate Customers**

- Identifies customers with the same CUSTOMER_KEY and NAME occurring more than once in the CUSTOMER_DIM table.
- Result is the count of such duplicates.

2. **Customers Spending Above Average**
    - Joins CUSTOMER_DIM and FACT_TABLE to compute total spending per customer.
    - Compares individual spending with the overall average and counts customers above average.
3. **Mid-Range Sales Volume**
    - Filters the FACT_TABLE and TIME_DIM to retrieve total sales from:
        - Last 15 days of the previous month.
        - First 15 days of the current month.
4. **High-Earning Banks**
    - Aggregates total earnings by BANK_NAME from TRANS_DIM and FACT_TABLE.
    - Categorizes banks as "More earning banks" if their total exceeds the average.
    - Counts how many banks fall into this category.

Each result is inserted into the DATA_AUTOMATION log table, along with a check name, description, failed count, and timestamp.

---

5. Audit Table Schema

Table Name: DATA_AUTOMATION
Columns:

- CHECK_NAME (VARCHAR): Name of the quality check
- CHECK_DESCRIPTION (VARCHAR): Description of the check logic
- FAILED_COUNT (INT): Number of records that failed the check or met the condition
- CHECK_DATE (DATETIME): Timestamp of when the check was run *(optional: added for traceability)*

---

6. Sample Output

CHECK_NAME	CHECK_DESCRIPTION	FAILED_COUNT	CHECK_DATE
Duplicate customers	Identifies multiple identical records in CUSTOMER_DIM	25	2025-05-26 10:15:00
More than average	Customers with spending above average	123	2025-05-26 10:15:00
Mid range sales	Sales from split month periods	2,300,000	2025-05-26 10:15:00
More earning banks	Number of banks earning above average	6	2025-05-26 10:15:00

---

7. Key Features

- **Automation-Ready**: Can be scheduled via SQL Server Agent for regular execution.
- **Reusable and Extensible**: Additional checks can be added with minimal effort.
- **Centralized Logging**: All outputs are stored in one table for monitoring and reporting.
- **Traceability**: Each entry includes a timestamp for audit and trend analysis.

---

8. Future Enhancements

- **Email Alerts**: Integrate alerts when thresholds are breached.
- **Run ID Tracking**: Add a unique RUN_ID to group checks by execution.
- **Dashboard Integration**: Build a Power BI or Tableau dashboard to visualize trends.
- **Parameterization**: Allow filters by date, customer segment, or product line for targeted checks.

---

9. How to Use

- **Manual Execution**:
  Run the stored procedure using:
  EXEC DATA_AUTOMATION3;

- **Scheduled Execution** *(Recommended)*:
  Use SQL Server Agent to schedule the procedure on a daily/weekly/monthly basis.

---

10. Summary

This project demonstrates the implementation of a robust, scalable, and automated data quality framework in SQL Server. By centralizing and automating checks, it reduces manual effort, ensures consistent monitoring, and supports proactive data governance practices. It can be further enhanced to include real-time alerting and interactive visualizations, forming a comprehensive data quality monitoring solution.