IMPROVING DATA INTEGRATION QUALITY FOR MULTI-SOURCE ANALYTICS

Abstract :

This project focuses on enhancing the quality of data integration processes for multi-source analytics. By addressing data inconsistencies, improving data accuracy, and streamlining data pipelines, we aim to deliver high-quality data for informed decision- making.

Problem Definition :

Key Questions :

1. How to effectively identify and address data inconsistencies across multiple sources?

2. How to ensure data accuracy and completeness during the integration process?

3. How to monitor and maintain data quality over time?

4. How to effectively communicate data quality issues to stakeholders?

Target User :

Data analysts, data scientists, business intelligence professionals, and decision- makers who rely on integrated data for insights and reports.

Goal:-

To establish a robust and efficient data integration process that delivers high-quality, reliable, and timely data for all downstream analytics
applications.

Requirements :

Functional :

1. Data profiling and cleansing capabilities (e.g., identifying and handling missing values, duplicates, outliers).
2. Data validation and verification rules (e.g., data type checks, range checks, uniqueness checks) **.**
3. Data transformation and enrichment capabilities (e.g., data type conversions, data standardization).
4. Data lineage tracking and documentation.
5. Integration with various data sources (e.g., databases, files, APIs).

Non-functional :

1. Scalability and performance to handle large datasets.
2. Maintainability and ease of use for the data quality analyst.
3. Data security and privacy compliance.
4. Robust error handling and logging.

Tools and Platform:

Language

Python : Core programming language for data manipulation, analysis, and quality assessment.

Data Quality Libraries :

1. Pandas : Data manipulation and analysis.
2. Great Expectations : Data profiling, validation, and documentation.
3. PySpark : Scalable data processing for large datasets.

4. DBT (Data Build Tool) : Data transformation and orchestration.
5. Cloud Platforms : IBM Cloud Services.

Implementation Plan :

1. Data Source Assessment : Analyze data sources, identify data quality issues, and define data integration requirements.

2. ETL Pipeline Design : Design and develop ETL pipelines using chosen tools and technologies.

3. Data Quality Rules Implementation : Implement data quality checks and validation rules within the ETL process.

4. Testing and Validation : Conduct thorough testing of the data integration process, including unit tests, integration tests, and user acceptance testing.

5. Deployment and Monitoring : Deploy the solution to the production environment and establish ongoing monitoring and maintenance processes.

Expected Outcome :

1. Improved data accuracy, consistency, and reliability across all analytical datasets.
2. Increased trust in data-driven decisions.
3. Enhanced efficiency and productivity of data analysts and business users.
4. Reduced time spent on data cleaning and validation.
5. Better understanding of data quality issues and their impact on business outcomes.

My contribution in this phase of the project :

Implementation plan and outcome.

Name : Nagayashas A B

Candidate Id : CAN_25553698