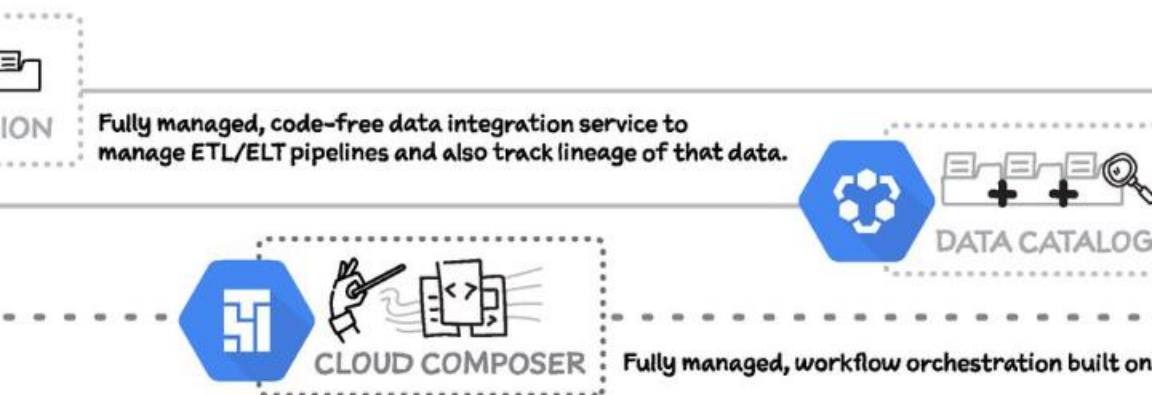
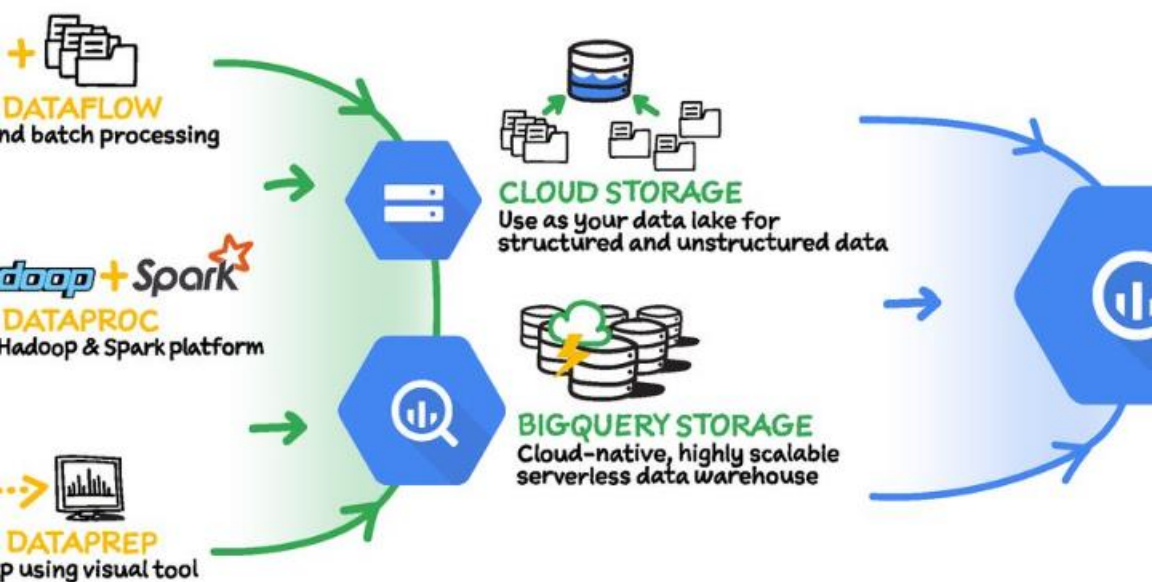


How to build a scalable ANALYTICS PIPE



Smart Data Quality Engine for Data Pipelines

A robust engine that performs automated ETL and data quality checks, ensuring high-quality data in pipelines. Learn how it can revolutionize your data processing.

by Atanu Das



Introduction

1 Overview

Automated Smart Data Quality Engine revolutionizes data quality in dynamic landscapes.

2 Importance of Data Quality

Accurate data is pivotal for strategic decisions; the engine ensures precision in today's data-driven era:

- **Decision Impact:** Accurate data influences strategic decision-making.
- **Business Outcomes:** Precision enhances overall business outcomes.
- **Data-Driven Era:** Crucial for success in today's data-driven ecosystem.
- **Strategic Asset:** The engine is not just a solution; it's a key strategic asset.

Challenges in Data Quality Checks

1 Complexity of ELT and DQ

Data Integration Challenges:

- Diverse source alignment.
- Ensuring consistency during extraction and loading.

Transformations and Cleansing:

- Precision in mapping complexities.
- Thorough identification and cleansing.

Dependency Management:

- Consideration of interconnected processes.
- Careful planning of the order of operations.

Data Quality Checks:

- Definition of meaningful thresholds.
- Continuous, real-time monitoring.

2 Continuous Monitoring

Continuous monitoring in data pipelines is crucial to swiftly detect anomalies and ensure real-time data accuracy.

Error handling is vital for preventing the propagation of faulty data, safeguarding the integrity of downstream processes, and maintaining the reliability of the entire data infrastructure.

Smart Data Quality Engine

1 Definition and Purpose

The Smart Data Quality Engine is a proactive solution that automates the detection of major data issues in real-time. Its scope includes a wide range of DQ checks, such as validating data types, identifying outliers, and ensuring referential integrity.

2 Functionality

The engine should perform major ETL and data quality checks seamlessly. The must-have data quality checks in ETL:

- NULL values test
- Volume tests
- Numeric distribution tests
- Uniqueness tests
- Referential integrity test
- String patterns
- Freshness checks
- Data Observability: The Key to Scaling Data Quality

Threshold-Based Error Handling

1

Explanation of Error Threshold

One key feature of the Smart Data Quality Engine is its threshold-based error handling. For each DQ check, users can set predefined thresholds.

2

Stopping the Pipeline

If the engine detects that the threshold has been breached during a data pipeline run, it will automatically halt the pipeline, preventing the propagation of potentially erroneous data.

3

Warning and Continuation

The engine should detect and warns about errors that are below the threshold, allowing for pipeline continuation.



Benefits of a Smart Data Quality Engine

1 Automation

Reduces the manual effort required for DQ checks, improving efficiency.

2 Real-time Detection

Identifies data issues as soon as they arise, minimizing the impact on downstream processes.

3 Threshold Control

Empowers users to define and enforce acceptable data quality thresholds, ensuring data integrity.

4 Enhanced Reliability

Increases the overall reliability of data pipelines by preventing the propagation of faulty data.

Conclusion

In conclusion, the Smart Data Quality Engine is a proactive and automated solution that addresses the challenges associated with data quality in modern data pipelines. By incorporating this engine into your data infrastructure, you can ensure the consistency, accuracy, and reliability of your data, ultimately leading to more informed decision-making and improved business outcomes.

