GlobalLogic®
A Hitachi Group Company

Data Quality Solutions for Stream and Batch Data Processing

by Mark Norkin, Engineering Consultant.

The Enterprise Data Lake Platform capability is old news. With a plethora of Big Data Technologies at your disposal, aggregating data from Online Transaction Processing (OLTP) systems into a single pile is easy. But as we all know, the data we get from different locations in our Enterprise is dirty, modeled in different ways, and difficult to join.

Generally, this data is not consumable by downstream business-oriented consumers for running their business and monetizing it. Poor data quality has a negative impact on business decisions. To overcome these challenges posed by enterprise data, this white paper gives you insights into solutions in the marketplace and provides designs for implementing and supporting data quality as part of the overall Data Platform taxonomy.

We distinguish streaming and batch approaches for data processing, and review how data quality capability can be incorporated into each of them.

Context and Problem

As businesses strive to extract insights and make informed decisions from their vast data assets, ensuring the reliability, accuracy, and usability of data has become paramount. The emergence of big data has transformed the way organizations collect, process, and analyze data. With the exponential growth of data volumes and the diversity of data sources, big data projects present unique challenges in maintaining data quality. Traditional approaches to data quality management often struggle to keep pace with the scale, complexity, and velocity of big data.

One of the primary challenges posed by big data projects is the sheer volume and variety of data being processed. The influx of data from numerous sources, including structured, unstructured, and semi-structured data, introduces complexities that traditional data quality frameworks may struggle to address. The velocity at which data is generated and the need for real-time or near real-time processing necessitate agile and efficient data quality management processes. These processes enable organizations to effectively manage and utilize data in fast-paced, time-sensitive environments.

The impact of poor data quality can be severe and far-reaching. Inaccurate, incomplete, or inconsistent data can lead to flawed analysis, erroneous insights, and misguided decision-making. Organizations heavily relying on data engineering to drive business strategies, operational efficiency, and customer experience cannot afford to overlook data quality considerations.

Data Quality in the Enterprise Data Platform

An Enterprise Data Platform represents a high-level flow of data-related events and tasks, and Figure 1 shows important components and roles within the data platform. Data quality is located in the Source of Truth & Governance logical component along with Data Profile and Lineage. This component is preceded by the Data Lake components that represent the ingestion, wrangling, and refining of data.

Data sources can be internal or external. *Ingestion* is the process of transporting data from data sources to a target destination where the received data is required to align with a defined structure or format. The ingestion process can be performed in multiple ways. Ingestion can be scheduled and triggered by a request, or it can be done on an ad-hoc basis as needed.

Wrangling is the process of converting and organizing raw data into a format usable for business needs such as analytics, visualizations, and more. Tasks performed during the wrangling process include cleaning data, removing duplicate data, merging data, filtering data, and making the data ready for use by consumers and applications.

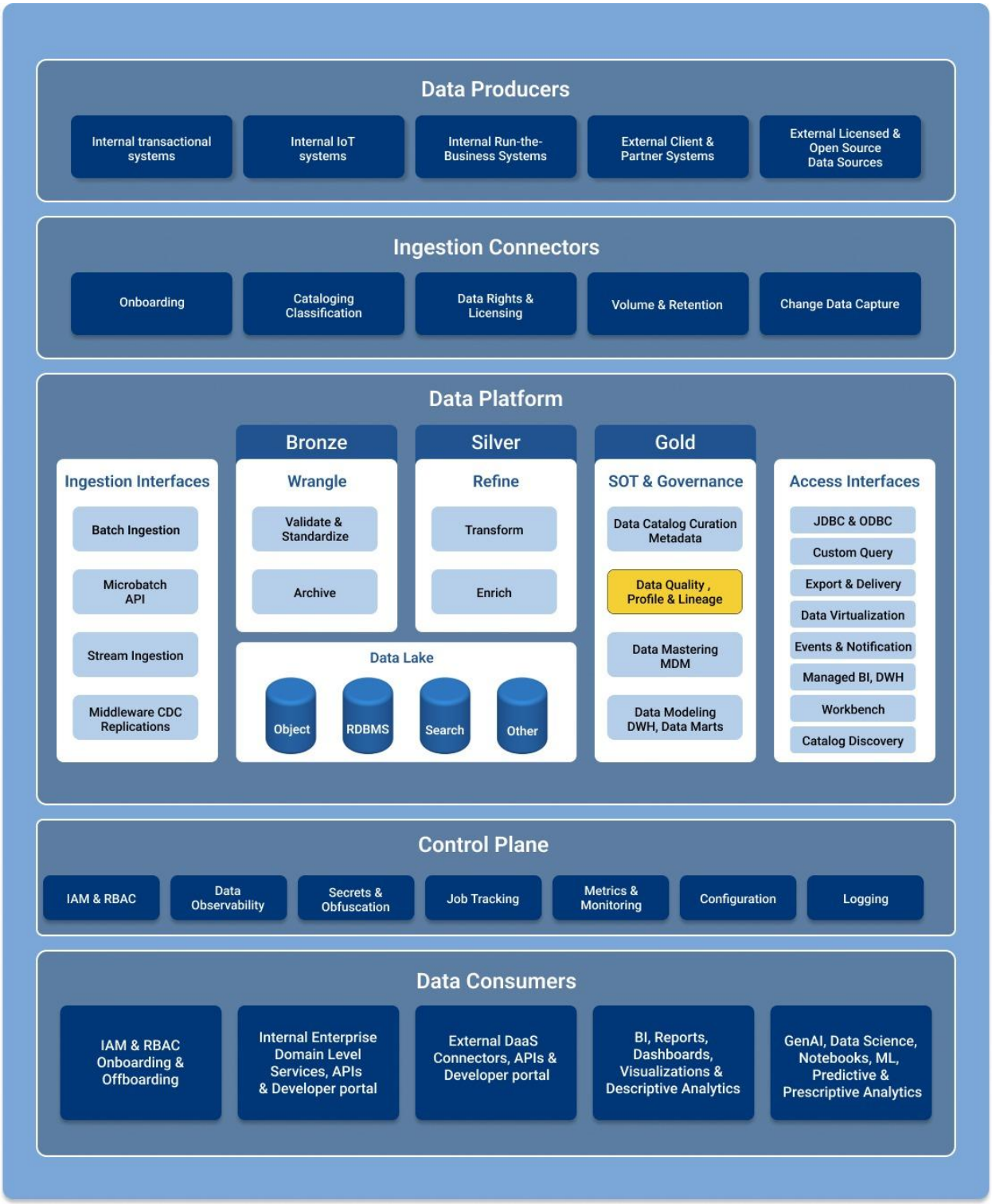

Figure 1 – Enterprise Data Platform

Key Dimensions, Requirements, and Stakeholders

Key Dimensions of Data Quality

Data quality refers to the completeness, accuracy, consistency, timeliness, and validity of data. It is the measure of how well data meets the requirements and expectations of its intended use. High-quality data is reliable, trustworthy, and fit for its intended purpose, enabling organizations to make informed decisions, drive operational efficiency, and achieve business objectives. Figure 2 shows the key dimensions of data quality.

Figure 2 - Key Dimensions of Data Quality

Dimensions of Data Quality	**Description**
Completeness	Data should be comprehensive, with all relevant attributes and fields properly populated and without any missing values.
Accuracy	The degree to which data correctly represents the real-world entities or events it describes.
Consistency	Data is uniform and conforms to predefined rules and standards across different systems or sources.
Timeliness	The freshness and currency of data ensure that it is up-to-date and reflects the current state of the business or environment.
Validity	The data meets predefined rules, constraints, and integrity checks to ensure its correctness and adherence to business requirements.

Data Quality Requirements

Data quality requirements are specific criteria and standards that define the level of quality expected from data within an organization or project. These requirements establish the guidelines for how data should be captured, stored, processed, and used to ensure that it is accurate, reliable, consistent, and relevant. Data quality requirements stem from various sources and considerations:

Business Objectives and Needs - Data quality requirements are often driven by the overarching business objectives and needs. Organizations define the level of data quality necessary to support their strategic goals. For instance, if the business objective is to make informed decisions based on data analytics, the data quality requirements will emphasize accuracy and consistency.

Regulatory and Compliance Standards - Many industries are subject to regulatory requirements that dictate the quality of data they collect and store. Compliance standards often mandate data accuracy, privacy, and security. Organizations must ensure that their data quality practices align with these regulations.

Stakeholder Expectations - Data quality requirements are influenced by the expectations of stakeholders, including customers, partners, and investors. Stakeholders require accurate and trustworthy data to make decisions, and their expectations shape the data quality standards.

Data Usage Scenarios - Different use cases and scenarios demand varying levels of data quality. For example, real-time applications may prioritize low latency and immediate data availability, while analytical applications may focus on accuracy and completeness over time.

Data Criticality - Not all data is equally critical. Organizations classify their data based on its importance to business operations. Critical data, such as financial records or customer information, requires higher data quality standards.

Data Source Characteristics - Data quality requirements are influenced by data sources. Data collected from automated systems, manual data entry, or external sources may have varying levels of inherent quality. Requirements need to be adjusted based on the data source.

Data Lifecycles - Data quality requirements might change at different stages of the data lifecycle. For example, during data collection, accuracy and validity may be emphasized, while in data archiving, long-term consistency may be more crucial.

Industry Best Practices - Industry-specific best practices and benchmarks can guide organizations in setting data quality standards. These practices provide a reference for what constitutes good data quality in a particular field.

Data Governance Policies - Organizations often have data governance policies encompassing data quality practices. These policies define roles, responsibilities, and processes for maintaining data quality.

Historical Data Issues - Past data quality issues and challenges inform future data quality requirements. Organizations learn from mistakes and improve their standards to prevent similar issues in the future.

Technological Capabilities - The tools and technologies used to manage data can impact data quality requirements. Advanced technologies may allow for more automated data validation and cleansing, affecting the defined standards.

Operational and Analytical Needs - Operational systems may prioritize data availability and consistency for day-to-day processes, while analytical systems may require high accuracy for meaningful insights.

In essence, data quality requirements result from a combination of organizational goals, industry standards, stakeholder expectations, and practical considerations. It's important for organizations to carefully analyze these factors and define clear and comprehensive data quality requirements to ensure that their data serves as a reliable foundation for decision-making and business processes.

Key Stakeholders For Maintaining Data Quality

Maintaining data quality involves a team effort from various personnel, each with distinct roles and responsibilities. Here is a list of stakeholders and their responsibilities.

Role	Responsibilities
Data Stewards	Define and enforce data quality standards, policies, procedures, and data governance. Oversee data quality initiatives, identity issues, and improve overall data integrity.
Data Quality Analysts	Assess and monitor data for accuracy, completeness, and consistency. Identify anomalies, validate data against predefined criteria, and collaborate with teams to resolve data quality issues.
Data Engineers	Design and develop data pipelines, transform, and transport data while ensuring data quality. Implement data quality checks during data integration and ETL processes.
Data Scientists	Analyze data to extract insights. Collaborate with data engineers and data quality analysts to ensure that data used for analysis is of high quality and suitable for accurate insights.
Data Governance Managers	Establish data governance policies and guidelines, ensuring that data quality efforts align with business goals. Oversee the overall data quality strategy.
Database Administrators	Manage databases and ensure that they are properly configured for data quality. Optimize database performance, ensure data integrity, and implement necessary controls.
Business Analysts	Collaborate closely with data quality analysts to understand business requirements and translate them into data quality rules. They provide context for data quality expectations.
IT Support Teams	Manage technical infrastructure and tools used for data management. Assist in implementing and maintaining data quality tools and technologies.
Compliance and Legal Teams	Ensure that data quality practices adhere to legal and regulatory requirements. Monitor data handling processes to ensure compliance with data protection and privacy laws.

Quality Assurance (QA) Teams	Test data quality processes and procedures to ensure they are functioning as expected. Identify potential gaps and areas for improvement.
Chief Data Officer (CDO)	Oversees the organization's data strategy and ensures that data quality is a key focus. Collaborates with various teams to establish data quality goals and priorities.
Data Consumers and Users	Validate the data they use and report any discrepancies or issues they encounter to the data quality team.

Data Quality For Stream and Batch Data Processing

Data quality applies to two important approaches for handling and managing big data: batch data processing and stream data processing.

Batch data processing involves processing large volumes of data in batches or groups. Data is collected over a specific period, stored, and then processed as a whole. Batch processing is typically used for non-real-time applications that can tolerate some delay in data analysis or insights.

Stream data processing refers to the real-time or near-real-time processing of continuous data streams. It involves ingesting, processing, and analyzing data as it arrives, typically in small, incremental portions called events or records. Stream data processing is commonly used for time-sensitive applications where immediate insights or actions are required.

Both types of data processing have their respective use cases and advantages depending on the requirements of the application. Organizations often use a combination of these approaches, depending on the nature of their data, processing needs, and business objectives. The following table, which illustrates the key differences between stream and batch data processing, is followed by an explanation of these features.

A Comparison of Features For Batch Processing vs Stream Processing	Batch Processing	Stream Processing
Data collection and storage over a period of time	✅	❌
Higher latency	✅	❌
Bulk processing	✅	❌
Scheduled execution	✅	❌
Continuous data ingestion	❌	✅
Low latency	❌	✅
Event-driven	❌	✅
Dynamic and evolving data	❌	✅

Features of Batch Data Processing

Data Collection and Storage - Before processing, data is collected and stored over a certain period, often in distributed file systems or databases.

Higher Latency - Batch processing operates on data collected within a specific time window or batch interval, which can range from minutes to hours or even longer. As a result, the processing latency is higher when compared to stream processing.

Bulk Processing - Data is analyzed and processed in larger chunks or batches, allowing for efficient computations on large datasets.

Scheduled Execution - Batch jobs are typically scheduled to run at specific intervals or predefined times, processing the accumulated data in each batch.

Features of Stream Data Processing

Continuous Data Ingestion - Data is continuously ingested and processed as it becomes available, without waiting for the entire dataset to be collected.

Low Latency - Processing delays are minimized and real-time insights on data streams with low latency are often available in milliseconds or seconds.

Event-Driven - Data processing is driven by individual events or records where each event triggers specific actions or computations.

Dynamic and Evolving Data - Data is constantly changing and evolving with a requirement for real-time analysis and adaptation to changing data patterns.

Implementing Data Quality Solutions

While the underlying principles of data quality apply to both streaming and batch data pipelines, the implementation approach differs to accommodate the specific characteristics of each pipeline type. Some data quality features are common to both stream and batch data processing.

Data exploration is a key feature that applies to both scenarios. Within the context of big data and analytics, it refers to the process of analyzing and understanding the characteristics, structure, and content during onboarding of a new data source. It involves examining the data to gain insights, discover patterns, and identify potential issues or opportunities for data integration and utilization.

Although data quality techniques can be implemented in many locations across the spectrum of sources, aggregation, curation, and service, let's examine data quality techniques that specifically apply within the context of batch and stream processing.

When onboarding a new data source, data exploration helps in several ways.

Understanding Data Structure - Data exploration allows you to understand the structure of the data source, including the format, organization, and schema of the data. This helps in planning the data integration process and mapping the data to the target systems or data models.

Identifying Data Quality Issues - By exploring data, you can identify potential data quality issues such as missing values, inconsistencies, outliers, or data format problems. This information is crucial for data cleansing and transformation activities that ensure the quality and reliability of data.

Discovering Data Relationships - Data exploration helps in discovering relationships and dependencies within the data. It enables you to identify connections between different data elements, tables, or entities. This understanding is valuable for establishing data relationships during data integration and enhancing data consistency and accuracy.

Uncovering Data Patterns and Insights - Through data exploration, you can uncover hidden patterns, trends, or anomalies in the data. This can provide valuable insights for data analysis, reporting, and decision-making. Exploratory data analysis techniques, such as visualization or statistical analysis, can be applied to gain a deeper understanding of the data.

Planning Data Integration and Transformation - By exploring the data source, you can determine the necessary data integration and transformation steps required to align the data with the target systems or data models. This includes activities such as data mapping, data standardization, and data enrichment.

The analysis done during the data exploration phase helps to define key data quality metrics and data quality rules. These metrics and rules serve as a requirement for implementing data quality checks within both streaming and batch data pipelines for newly onboarded data.

Data Quality for Batch Data Processing

Workflow orchestration tools such as Apache Airflow are frequently used to implement batch data processing. These tools enable the creation of data pipelines consisting of interrelated tasks, defining their execution order and dependencies.Once the data that batch data pipeline processes has clear data quality metrics and data quality rules defined, these rules can be implemented by integrating data quality checks into the data pipeline tasks.

Example

Let's consider a batch processing example from a GlobalLogic project. The client was a US-based company in the advertising industry that evaluated the quality of online ad placements needed to migrate the data platform from an on-premise environment to AWS in preparation for the company's IPO. Several daily batch data pipelines were implemented by using Apache Airflow as a workflow orchestration tool. To ensure that the migrated data pipeline in the cloud environment produces has the same data quality as as the on-premise version of that same data pipeline, a regression testing framework was developed for data pipelines that compares two datasets i.e. on-prem version of the dataset vs cloud version of the same dataset. The following figure shows the sequence of events for data pipelines.

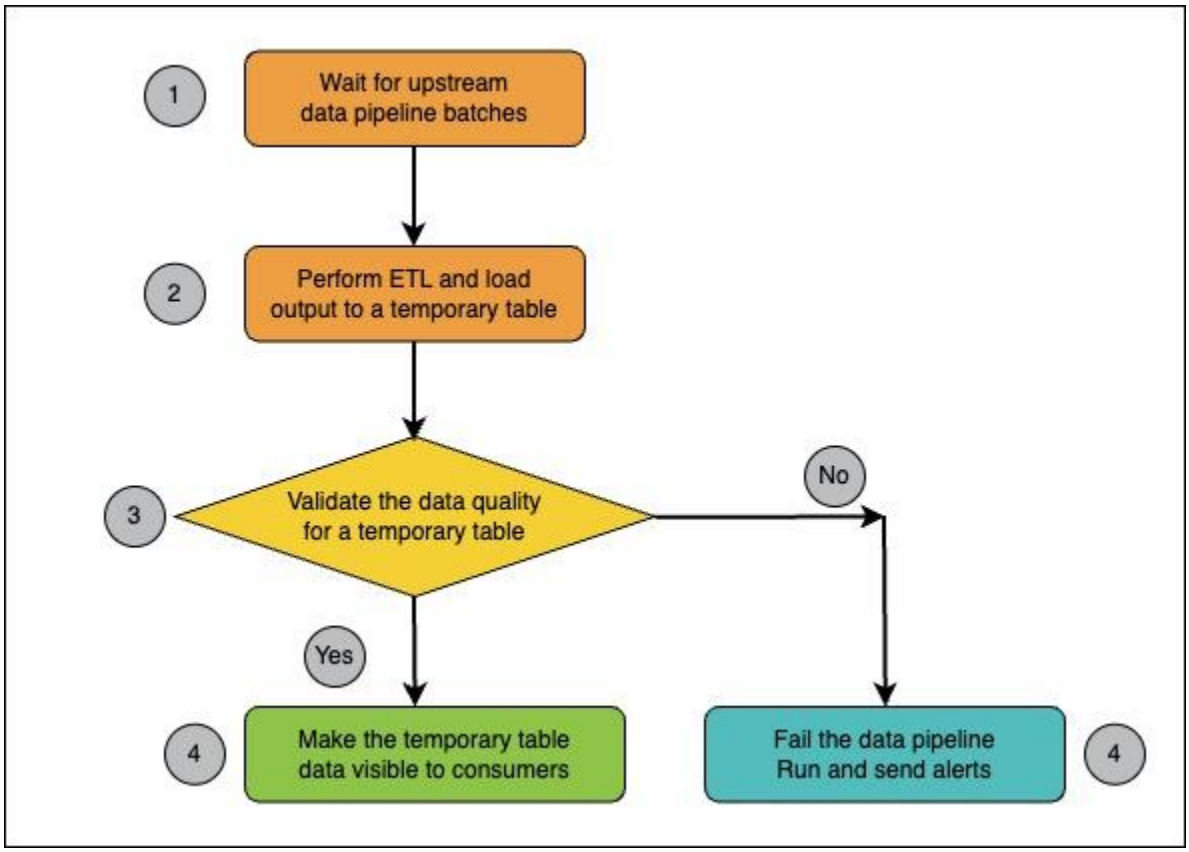

Figure 3 - Batch Processing Example for Data Quality

We note the following:

1. The data pipeline is waiting for its upstream data pipelines to finish so that they produce the dataset(s) upon which this data pipeline depends.
2. When the inputs to the data pipeline are ready, the data pipeline does the data processing according to its own business logic and produces the output to the internal temporary table.
3. The predefined data quality checks are executed against the fresh dataset to see if the data quality matches the business requirements.
4. If the fresh dataset matches the data quality requirements, then the previously internal table changes from being internal to public - meaning it is now visible to the downstream data pipelines for consumption. If the data quality checks fail, the data pipeline run fails and appropriate alerts are sent to data pipeline support engineers. The results are then examined by data engineers and data analysts.

Data Quality for Stream Data Processing

Implementing data quality for a streaming big data pipeline requires a comprehensive approach that addresses data validation, cleansing, monitoring, and remediation in real-time. Figure 4 shows approaches for building data quality into stream data processing.

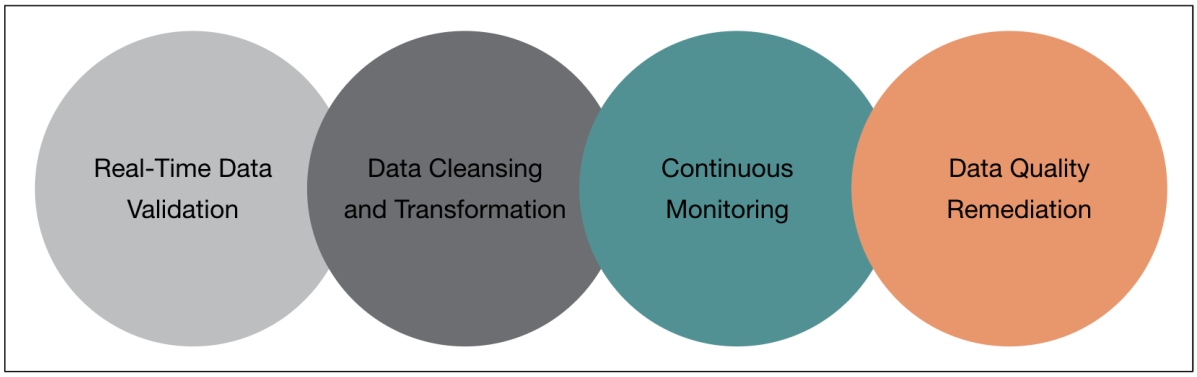

Figure 4 - A 4-Step Approach for Building Data Quality Into Stream Data Processing

Real-Time Data Validation - Data engineers implement code with data validation mechanisms in real time to ensure that incoming data meets the defined quality standards. Validation techniques such as schema validation, data type checks, range checks, and business rule validations are applied. Invalid or non-compliant data should be flagged or rejected, and appropriate notifications or alerts should be generated for further action.

Data Cleansing and Transformation - Data inconsistencies, errors, and outliers are handled by applying data cleansing and transformation techniques in the streaming data. These techniques include data deduplication, data standardization, data normalization, and missing value imputation.

Continuous Monitoring - Establishes real-time monitoring capabilities to track the quality of data flowing through the pipeline. Data quality metrics are monitored and alerts or notifications are issued when deviations or anomalies are detected. Monitoring tools and visualizations provide insights into the health and performance of the streaming pipeline for identifying potential issues.

Data Quality Remediation - Based upon developing processes and mechanisms to address data quality issues in real time. This may involve automated or manual remediation steps such as reprocessing or correcting the data, depending on the severity and impact of the data quality problem. Automated workflows or triggers are implemented to initiate remediation actions promptly.

A stream-processing data pipeline needs to account for possible quality issues with the processed data and requires continued operation even if the data quality is poor for the processed data.

Example

Figure 5 shows an example from a project completed at GlobalLogic. The security analytics system that was implemented had to receive raw logs data from multiple data sources such as Windows logs, Linux logs, and various Cisco devices, etc.

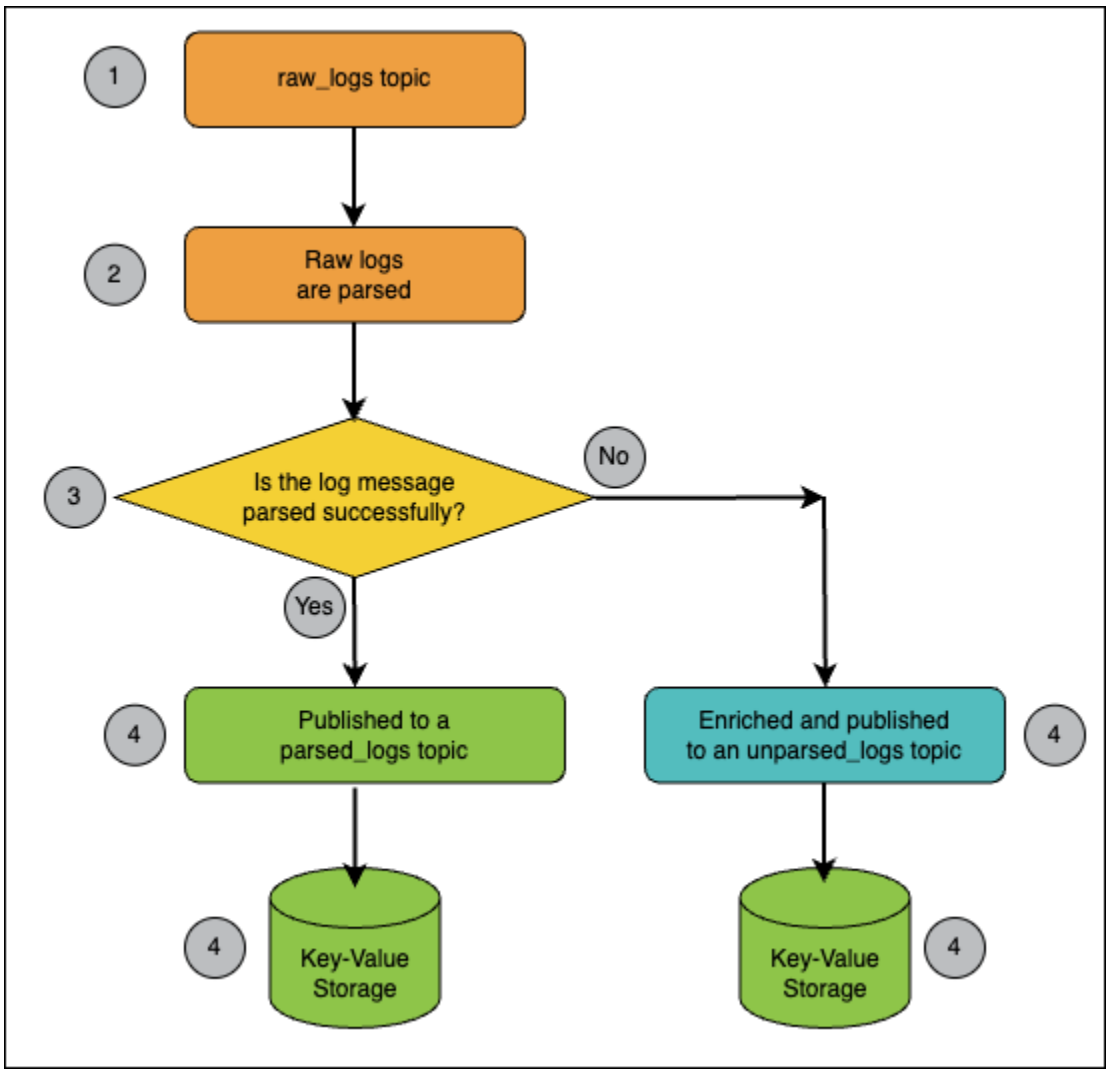

Figure 5 - Raw Logs Data Flow in a Stream Processing Primer

1. The logs are received via the TCP/IP protocol and published without any transformations to the *raw_logs* Apache Kafka topic.
2. A dedicated Apache Spark Streaming job receives the raw log messages and parses them according to business and device type rules.
3. If the log message was parsed successfully, it is published into a dedicated *parsed_logs* Apache Kafka topic. If the message was unsuccessfully parsed due to data quality issues, the raw log message is enriched with additional data quality error information and then published into another dedicated *unparsed_logs* Apache Kafka topic.

4. Successfully parsed log messages are stored in a dedicated Key-Value database according to a predefined schema. Unparsed log messages that had data quality issues were stored in a dedicated blob storage location and were later analyzed by data analysts.
5. After conducting data analysis on the questioned data, there was a separate remediation batch process that allowed us to replay backlog messages with data quality issues. These unparsed log messages are now either fixed manually by data analysts, or the code of the raw logs parsing job (step 2) was changed in order to handle additional data quality use cases.

Proactive vs Reactive Approaches to Data Quality

There are two distinctive approaches to performing data quality checks. The first approach is proactive and consists of data quality checks on the data produced by the data pipeline. The benefits of this approach include:

1. Publishing of output data for consumption after data quality is ensured.
2. A faster feedback loop is possible when the data quality checks are failing. Data professionals such as data analysts and data engineers could react immediately to incidents of failure and take the appropriate measures quickly.

The second approach consists of data quality checks that are performed outside of the data pipeline's main logic and include the following use cases:

1. Unlike the proactive approach, there are no time-sensitive limitations. Therefore, the data quality checks can be computed and time-heavy.
2. There is greater freedom to select data quality technologies and tooling such as data profilers tools. You can avoid integration issues when these technologies and tooling are incorporated into your current streaming or batch data pipeline.

A reactive approach to data quality is based on performing scheduled data quality checks regularly to automatically validate data quality stored within the big data platform. Defining the frequency of the checks is determined by the criticality of the data and business requirements. Configuring the checks to run in the background ensures minimal disruption to ongoing data processing activities.

Regardless of whether you follow a proactive or reactive approach to data quality, when data quality issues are identified, develop processes and workflows to remediate and clean up data. Establish guidelines and procedures for correcting or removing data that fails to meet defined quality standards. Define the responsibilities and ownership for data remediation activities.

Available Tools and Platforms

Here is a list of notable industry tools and resources (in alphabetical order) for building a data quality capability in your data platform:

Apache Airflow's SQL Data Quality Operators - The SQL check operators in the Common SQL provider package provide a simple and effective way to implement data quality checks in Airflow DAGs. Using this set of operators, data engineers can quickly develop a pipeline specifically for checking data quality, or add data quality checks to existing pipelines.

Apache Griffin - An open source Data Quality solution for Big Data, which supports both batch and streaming modes. It offers a set of well-defined data quality domain models, which cover most of data quality problems in general. It also defines a set of data-quality DSLs to help users define their quality criteria. By extending the DSL, users can even implement their own specific features/functions in Apache Griffin.

Deequ - Deequ is a library built on top of Apache Spark for defining "unit tests for data", which measure data quality in large datasets. It is primarily built by the Amazon Web Services team and integrated into AWS as a AWS Glue Data Quality service.

Great Expectations - A library for validating, documenting, and profiling your data to maintain quality and improve communication between teams. It has several notable integrations with major cloud providers and other tools.

IBM InfoSphere - Provides functionality that includes data cleansing and data quality monitoring. With the end-to-end data quality tools, you can understand your data and its relationships, analyze and monitor data quality continuously, cleanse, standardize, and manage data, and maintain data lineage.

Informatica - Enables you to identify, fix, and monitor data quality problems in cloud and on-premises business applications.

SAP - Offers a cloud-based, REST API for address cleansing, geocoding, and reverse geocoding.

SAS Data Quality - Includes data standardization, deduplication, and data correction. Data integration and data quality jobs can be executed concurrently.

Soda.io - Enables data engineers to test data for quality by taking the data quality checks that data engineers prepare and using them to run a scan of datasets in a data source. This tool offers several usage options. Soda Core - is a free, open-source Python library and CLI tool. Soda Library, an extension of Soda Core, enables users to connect to Soda Cloud and offers features and functionality not available with the open-source tool.

Talend - Profiles, cleans, and masks data in real time. The data profiling feature enables you to identify data quality issues, discover hidden patterns, and identify anomalies through summarized statistics and graphical representations.

Conclusion

Data quality is a critical aspect of big data projects. As organizations continue to embrace the power of big data analytics, ensuring the reliability, accuracy, and consistency of data becomes paramount. The insights derived from big data drive key business decisions and strategies, making data quality an essential factor in achieving successful outcomes.

Throughout this whitepaper, we have explored the context, challenges, and aspects of data quality in the realm of big data. The volume, velocity, and variety of data generated pose significant challenges, making it necessary to adopt robust data quality practices. From data exploration and data cleansing to data validation, monitoring, and remediation, each step is crucial in ensuring high-quality data.

We examined the differences between implementing data quality in streaming data pipelines and batch data pipelines and acknowledged the unique characteristics of each approach. Whether you follow real-time or batch processing, data quality checks and measures must be implemented to maintain data integrity and reliability.

In conclusion, data quality in big data is a multifaceted challenge that requires a combination of technical expertise, organizational commitment, and the right set of tools and technologies. Investing in data quality practices not only ensures reliable and accurate insights but also enhances trust in data-driven decision-making processes.

About GlobalLogic

GlobalLogic, a Hitachi Group Company, is a leader in digital product engineering. We help our clients design and build innovative products, platforms, and digital experiences for the modern world. By integrating our strategic design, complex engineering, and vertical industry expertise with Hitachi's Operating Technology and Information Technology capabilities, we help our clients imagine what's possible and accelerate their transition into tomorrow's digital businesses. Headquartered in Silicon Valley, GlobalLogic operates design studios and engineering centers around the world, extending our deep expertise to customers in the automotive, communications, financial services, healthcare & life sciences, media and entertainment, manufacturing, semiconductor, and technology industries.