Data Integrity Problems in High-Volume High-Velocity Data Ingestion

Rameshbabu Lakshmanasamy, Senior Data Engineer, Jewelers Mutual Group

Girish Ganachari, Senior Data Engineer, Cervello

Abstract:

In the era of bigdata, and never ending data push from IoT devices, the IT infrastructure are built to be scalable to handle the huge batch loads or continuous streaming live data. However, the big question is how can be establish the data integrity. How can we make sure no data is lost from Origin till the destination passing through numerous touch points enroute ? How can we ensure the quality with continuous inflow ? Should the inflow be suspended to perform the DQ checks? Or should it be a totally independent parallel activity.

Let’s explore.

Key words: Quality Data Management, Data Pipelines, AutoScalling, Large Scale Streaming Data, Performance, IoT, Data Cleansing

Introduction:

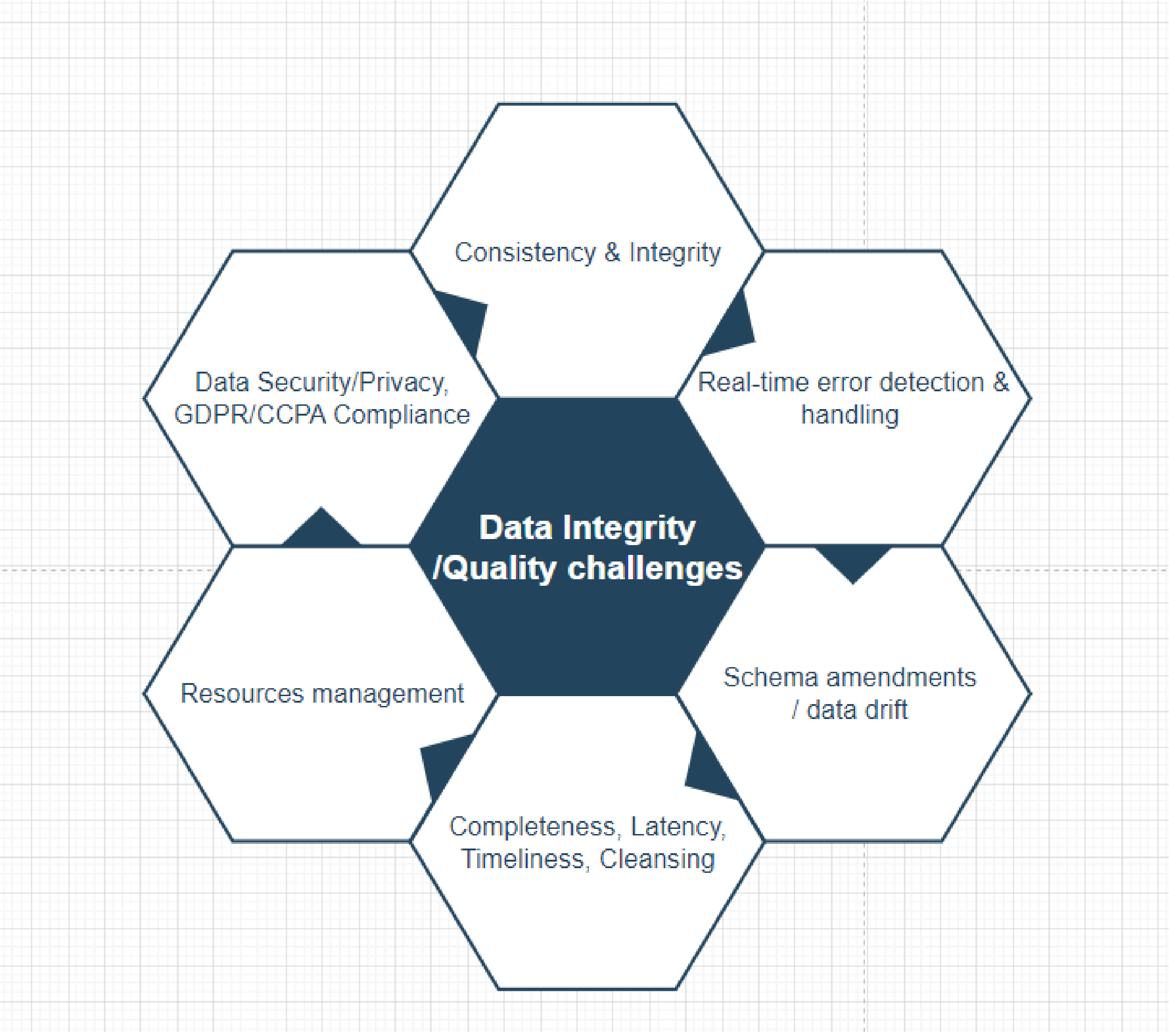
As the business needs of high-volume and faster data keep growing every passing day, it also comes with complications/challenges for the IT Leaders on handling and implementing this successfully, with the goal of high data integrity besides the availability of data on time. Honestly, there is no one-size-fits-all solution. It depends on the architecture and exact volume and sensitive nature of the data.

This is relatively easier in case of batch loads (even of very huge volume). Systems can afford to wait and introduce a delay in case of DQ controls/validation failures. The source application that generates/pushes the data can also send the aggregate metadata like count-of-records or count-of-hits or count-of-policies or sum-of-amounts or sum-of-claims etc. This metadata can be utilized upon ingestion in various touchpoints to ensure no loss of data.

However, in case of live streaming high velocity/volume transaction data that keeps pouring endlessly (like POS retail data), this needs to be a parallel activity. A robust working effective QDM process is a must here. More importantly if it is very important for a highly critical financial data from any banking organization. Low latency is key here to aid in timely decision making for quality insights, or preventing a loss of any important transaction in case of credit card usage data.

Key Challenges:

Data quality processes become more challenging when integrating data from two different streams. This will involve additional complications of data slicing, dicing, cleansing, transforming, and aggregating to arrive at the trustable data with high integrity. Data will be of no use if you cannot have a 100% trust.


Consistency & Integrity
Data Security/Privacy, GDPR/CCPA Compliance
Real-time error detection & handling
Data Integrity /Quality challenges
Resources management
Schema amendments / data drift
Completeness, Latency, Timeliness, Cleansing

Below are some of the list of main challenges the IT Leaders have :

1. Data Consistency & Integrity
2. Live/Realtime error handling and detection
3. Schema Evolution & data drift
4. Data deduplication/cleansing
5. Completeness
6. Latency & timeliness
7. Data Lineage
8. Scalability of Quality Check processes
9. Data Security/Privacy
10. Resources management
11. Monitoring & Alerting

Challenge & Possible Solutions:

Data Consistency & Integrity: It's challenging to ensure data integrity and consistency across a high-volume & high-velocity streams. Possible solutions are -

a. Hash functions, Checksums implementations as appropriate
b. Setting up quality control checks at various touch points in pipeline

Live/Realtime error handing & detection:

With high-volume and high-velocity streaming data, managing the data quality issues without impacting the ingestion speed is a challenge. Possible solutions are –

a. A robust error handling and logging setup might be appropriate
b. ML models can be employed for detecting anamolies
c. Think through configuring real-time data quality control/validation rules check

Schema evolution & data drift:

There are always schema amendments & data structure enhancements/modifications depending on the changing business need and new change requests. This needs to be accounted/addressed without stopping the ingestion. Halting it might lead to catching up later adding up performance/resource constraints as well. Possible solutions could be –

a. Develop a framework for automated schema changes handling
b. Consider feasible data formats like Parquet or Avro types

Data deduplication:

Finding and performing data deduplication during high-velocity nonstop streams without slowing the ingestion. Possible solutions –

a. Implement the de-dup processes downstream after the initial ingestion layer.
b. Distributed caching for record lookups/references.
c. A bloom filter setup, which is a probabilistic data structure that is based on Hashing.

Data Completeness:

Has all the data that are supposed to have ingested have arrived successfully ? Ensuring this specifically in Internet of Things scenario is very important.

a. Setting up a sequence numbering or checkpoints in data streams to measure.
b. Robust acknowledgement mechanisms
c. Create a data reconciliation processes to identify and document the variances and there upon the missing data record.

Latency & timeliness:

Ensuring Low-latency to perform the necessary data quality control checks on high-volume streams is another challenge. Possible solutions are –

a. Wherever applicable, implement parallel processing for data pipelines.
b. Using in-memory processing (e.g. Spark framework)
c. Setup adaptive rate limiting mechanisms.

Data Lineage:

Ensuring the complete data lineage captured and well documented for the high-volume stream situations. Possible solutions –

a. A detailed metadata management – Crucial for Impact assessments and data traceability
b. Employ data cataloging tools for tracking origins and transformations.
c. Implement audit trails in various touchpoints.

Scalability of QC Processes:

Design the QC controls in such a way they also scale up with increasing volumes. Possible solutions in this case –

a. Serverless architectures from cloud-native applications comes handy.
b. Tools capable of handling multiple data formats
c. Create standardized QC rules that is applicable across formats.

Data Security/Privacy:

Preserving data security and ensuring data privacy is not compromised without impacting ingestions or QC validations. Possible solutions are -

a. Configuring RBAC (role based access controls)
b. Secure transmission protocols
c. Data encryption and masking at all places for PII data elements.

Resources management:

Balancing and managing available resources between the ingestion and QC processes that run in parallel. Possible solutions -

a. Auto-Scaling feature provided by cloud platforms
b. Efficient QC algorithms
c. Dynamic resource allocation mechanism based on volume

Monitoring & Alerting:

Design a comprehensive alerting and monitoring mechanisms to keep up with the every increasing data flows. Options are –

a. Real-time dashboards using the data visualization applications/tools.
b. Setting up distributed monitoring systems.

Conclusion:

Addressing these challenges pertaining to high-volume and high-velocity streaming data often involves a combination of below :

1. State-of-the-art advanced technologies (e.g., stream processing, machine learning)
2. Robust architectural designs (e.g., distributed systems, cloud-native solutions)
3. Clearly laid out processes & protocols

IT Leadership must take account of one or more options that suits their applications/pipelines/systems, and build from there.

References:

□ Abiteboul, S., Andersson, M., Batini, C., Delcambre, L., Hess, C., Kersten, M., ... & Weikum, G. (2018). Research directions for principles of data management (dagstuhl perspectives workshop 16151). *Dagstuhl Manifestos*, 7(1), 1-29.

□ Batini, C., & Scannapieco, M. (2016). *Data and information quality: Dimensions, principles and techniques*. Springer.

□ Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14.

□ Ehrlinger, L., & Wöß, W. (2018). Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48.

□ Gao, J., Xie, C., & Tao, C. (2016). Big data validation and quality assurance -- Issuses, challenges, and needs. *2016 IEEE Symposium on Service-Oriented System Engineering (SOSE)*, 433-441.

□ Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, 154, 72-80.

□ Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, 53(1), 148-152.

□ Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015). A survey on data quality: Classifying poor data. *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*, 179-188.

□ Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A data quality in use model for big data. *Future Generation Computer Systems*, 63, 123-130.

□ Taleb, I., Dssouli, R., & Serhani, M. A. (2015). Big data pre-processing: A quality framework. *2015 IEEE International Congress on Big Data*, 191-198.

□ Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5-33.

□ Wende, K. (2007). A model for data governance–Organising accountabilities for data quality management. *18th Australasian Conference on Information Systems*.

□ Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering data quality problems. *Business & Information Systems Engineering*, 61(5), 575-593.

□ Zhu, H., Madnick, S. E., Lee, Y. W., & Wang, R. Y. (2014). Data and information quality research: Its evolution and future. In *Computing Handbook, Third Edition: Information Systems and Information Technology* (pp. 16-1). CRC Press.

□ Ardagna, D., Cappiello, C., Samá, W., & Vitali, M. (2018). Context-aware data quality assessment for big data. *Future Generation Computer Systems*, 89, 548-562.

□ Karkouch, A., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, 73, 57-81.