Beyond Functional Testing: QA Challenges in Industrial Analytical Pipelines

Estevan Delabari
Atlântico Institute
Fortaleza, Ceará, Brazil
estevan_delabari@atlantico.com.br

Sara Guimaraes
Atlântico Institute
Fortaleza, Ceará, Brazil
sara_guimaraes@atlantico.com.br

Vitor Linhares
Atlântico Institute
State University of Ceará
Fortaleza, Ceará, Brazil
vitor.fontenele@aluno.uece.br

Murilo Coelho
Atlântico Institute
State University of Ceará
Fortaleza, Ceará, Brazil
paulo_coelho@atlantico.com.br

Paulo Maia
Atlântico Institute
State University of Ceará
Fortaleza, Ceará, Brazil
pauloh.maia@uece.br

Abstract

Modern industrial analytical pipelines in Data Science projects introduce QA challenges that extend beyond traditional software validation, including temporal consistency, cross-layer traceability, data semantics, and analytical correctness. This paper presents an industrial experience report on adapting QA practices within a multi-layer Lakehouse architecture in a critical infrastructure project. Initially focused on structural validations, the QA process evolved toward semantic and domain-aware analytical validations through exploratory validations and continuous collaboration between QA, Data Engineering, and analytical teams. During the project execution, the QA team identified inconsistencies related to temporal duplication, ingestion losses, analytical view behavior, and transformation-rule ambiguities. In the context of this project, the experience highlighted the need to adapt traditional QA approaches to deal with domain-dependent analytical pipelines.

Keywords

QA, Analytical Pipelines, Critical Systems, Software Quality

1 Introduction

Modern critical infrastructures increasingly rely on data-driven systems to support monitoring, operational analysis, anomaly detection, and intelligent decision-making processes [4, 7]. In this context, industrial analytical pipelines and Lakehouse-based architectures have become essential for integrating heterogeneous data sources, processing large-scale sensor data, and enabling analytical and Artificial Intelligence (AI) applications [2, 5]. Architectures based on distributed processing and multi-layer data organization (e.g., Bronze, Silver, and Gold layers [3]) are frequently adopted due to their scalability, traceability, and analytical capabilities, although they also significantly increase the complexity of analytical processing and validation workflows [6].

However, validating analytical pipelines in such environments introduces challenges that differ significantly from those found in traditional software testing. In conventional software systems, Quality Assurance (QA) activities are commonly centered around unit, integration, system, and functional testing practices aimed at validating software behavior and expected application flows [8]. In analytical pipelines, however, failures are not limited to incorrect software behavior or broken functionalities. Problems may emerge as silent inconsistencies across data layers, temporal incoherence, semantic ambiguities, ingestion failures, duplicated events, transformation errors, or analytically misleading information that propagates through the pipeline without explicit failures [9]. The Data Quality literature highlights that data reliability is not restricted to structural correctness or integrity, but also involves contextual consistency, interpretability, and fitness for analytical use [10]. In multi-layer analytical architectures, these dimensions become increasingly critical due to the dependency on successive transformations and distributed processing stages [1].

This paper presents an industrial experience report combined with a risk-oriented QA framework proposal derived from a real-world industrial data analytics project in a critical infrastructure environment. During the project, the QA team identified multiple categories of distributed and domain-dependent failures that were not adequately addressed by conventional testing strategies. Based on these observations, the team structured a validation methodology focused on risk analysis, cross-layer traceability, data consistency, and domain-aware analytical validation. The main contribution of this work is the systematization of a risk-oriented QA method for industrial analytical pipelines derived from real-world validation scenarios observed in large-scale analytical pipelines, aiming to support the identification and classification of inconsistencies in these environments.

Furthermore, during the execution of the hydroelectric project analyzed in this work, several inconsistencies could not be identified solely through traditional functional validations. In multiple scenarios, the apparently incorrect behavior of the pipeline depended directly on the interpretation of domain rules and analytical transformations applied across Bronze, Silver, and Gold layers. Additionally, situations initially interpreted as defects were later understood as expected pipeline behavior once aligned with transformation rules, statistical aggregation requirements, and domain-specific operational semantics. Although derived from a hydroelectric context, the reported challenges and strategies are relevant to QA professionals, data engineers, and analytical teams working with distributed ingestion, layered processing, and data-driven industrial environments.

2 Context and Motivation

This study was conducted at the Atlântico Institute, a Brazilian Research and Technology Organization (RTO) located in Northeastern Brazil and accredited by Embrapii, a Brazilian agency that fosters industry-oriented research and innovation initiatives. The industrial partner involved in this study operates in the hydroelectric power generation sector.

The project involved integrating and analytically processing data from heterogeneous industrial and legacy systems to support monitoring, operational analysis, and data-driven decision-making. The environment adopted distributed ingestion pipelines within a multi-layer Lakehouse architecture, supporting historical and incremental ingestion from operational, geotechnical, physicochemical, hydrological, and industrial maintenance systems. Data was processed across Bronze, Silver, and Gold layers through progressive refinement stages commonly adopted in modern analytical architectures [3].

In this architecture, the Bronze layer is responsible for preserving raw data, ingestion history, and source traceability. The Silver layer applies data standardization, transformation, deduplication, typing enforcement, and consistency rules over the ingested datasets. Finally, the Gold layer consolidates curated analytical structures and aggregated datasets intended for dashboards, monitoring activities, operational analysis, and downstream analytical applications.

Initially, QA activities were conducted similarly to traditional ETL validation approaches, focusing mainly on schema verification, ingestion validation, table availability, and persistence consistency across processing stages. However, as analytical transformations evolved throughout the pipeline, more complex issues began to emerge, including temporal inconsistencies, cross-layer divergences, ambiguities involving deduplication logic, unexpected analytical behaviors, and volumetric discrepancies that could not be correctly interpreted without contextual understanding of the applied transformation rules. These scenarios gradually expanded the role of the QA team beyond structural validations, requiring closer interaction with Data Engineering, analytical teams, and domain specialists to properly interpret the semantics and operational behavior of the analytical pipeline.

These observations motivated the development of a structured and risk-oriented QA strategy for industrial analytical pipelines. Rather than focusing exclusively on functional software behavior, the proposed approach aimed to systematically identify inconsistencies, silent failures, data quality deviations, and domain-dependent anomalies throughout the complete analytical lifecycle.

3 Methodology

The proposed QA strategy evolved incrementally as new behaviors and ambiguities emerged throughout the analytical pipelines. Instead of relying on rigid validation procedures defined from the beginning of the project, the QA process progressively incorporated exploratory analyses, reproducible SQL-based validations, and continuous alignment with data engineering and analytical teams.

One representative example involved detecting duplicated timestamps before validating derived displacement calculations. The exploratory validation was initially performed through SQL queries grouping records by sensor identifier and timestamp to identify duplicated temporal events. Once confirmed, this validation became part of the reusable QA workflow because duplicated timestamps directly affected time interval calculations and could propagate division-by-zero errors into downstream analytical views.

Rather than assigning equal effort to every validation activity, the proposed methodology prioritized validation scenarios according to their potential risk to downstream data consumption. Risk was assessed by considering the expected impact on business-critical dashboards and datasets, the probability of silent error propagation across pipeline layers, the recurrence of similar inconsistencies during exploratory analyses, and the complexity of the transformation being validated. This prioritization allowed QA efforts to focus first on scenarios with the highest potential to compromise data reliability, while lower-risk issues were investigated as the understanding of the pipeline matured. Over time, the methodology became organized into six complementary validation dimensions: structural, volumetric, temporal consistency, semantic, analytical, and ingestion traceability validations. In practice, several validation scenarios initially categorized as isolated defects later revealed broader patterns associated with temporal consistency, semantic interpretation, or analytical transformation behavior. This incremental refinement process allowed the QA strategy to evolve from predominantly ingestion-oriented validations toward a more risk-oriented analytical quality approach capable of addressing cross-layer inconsistencies and silent analytical failures.

Structural validations focused on schema consistency, table availability, architectural adherence across layers, and appropriate data typing. These validations revealed scenarios in which analytically relevant numerical fields remained persisted as STRING values, negatively impacting downstream analytical calculations and aggregation reliability. Although these validations initially appeared straightforward, they became increasingly relevant as downstream analytical transformations started relying on numerical aggregations, statistical calculations, and derived indicators generated from intermediate layers. As a result, inconsistencies that would traditionally be considered low-impact at ingestion time became capable of silently propagating analytical distortions across dashboards and consolidated analytical structures.

Volumetric validations focused on understanding incremental ingestion behavior, deduplication effects, and quantitative consistency between analytical layers. In several situations, significant reductions between Bronze and Silver layers were initially interpreted as potential data loss, but were later validated as expected outcomes of incremental persistence strategies. These analyses demonstrated that volumetric discrepancies cannot always be interpreted as evidence of processing failures. In analytical pipelines based on incremental persistence strategies, substantial reductions between layers may represent expected consolidation behavior rather than unintended data loss. Consequently, volumetric validations required continuous alignment with transformation logic and persistence semantics implemented throughout the pipeline lifecycle.

Temporal consistency validations addressed timestamp integrity, chronological consistency, duplicated temporal events, and ordering coherence across historical analytical structures. Temporal consistency validations became particularly critical in scenarios involving derived analytical calculations dependent on chronological ordering between measurements. Even small temporal inconsistencies, such as duplicated timestamps or invalid chronological sequences, proved capable of generating cascading analytical failures affecting derived indicators, aggregated views, and operational dashboards consumed by downstream analytical processes.

Semantic validations focused on contextual interpretation of inconsistencies and analytical behaviors. Several situations initially interpreted as defects were later reclassified as semantically valid behaviors after alignment with domain specialists and data engineering teams. This category represented one of the most significant shifts observed during the project because it demonstrated that analytical quality cannot be evaluated exclusively through structural correctness. In several situations, the interpretation of inconsistencies depended directly on contextual operational knowledge, temporal granularity, and transformation rules distributed throughout different stages of the analytical pipeline.

Analytical validations focused on the behavior of Gold-layer structures consumed by dashboards and analytical models. These analyses evaluated the plausibility of derived calculations, aggregation stability, and consistency of analytical transformations under real operational scenarios. These validations expanded the role of QA beyond ingestion correctness by incorporating plausibility analysis, derived metric stability, and analytical behavior verification under real operational conditions. This approach became especially important in Gold-layer structures consumed directly by dashboards and analytical decision-support workflows.

Finally, ingestion traceability validations incorporated operational observability concerns related to incremental ingestion processes, including integrity verification and completeness validation of transferred artifacts before persistence into analytical layers. These operational integrity concerns highlighted that analytical reliability also depends on the observability of ingestion workflows and artifact transfer processes. In scenarios involving incremental ingestion and distributed storage, failures associated with partially transferred or unstable artifacts may silently compromise downstream analytical consistency without generating explicit pipeline interruptions.

4 Discussion

Table 1 summarizes the main validation scenarios identified during the industrial project, including their observed impacts and the corresponding validation outcomes. The observed validation scenarios demonstrated that QA activities in analytical pipelines evolve significantly beyond traditional functional verification strategies. Initially, validations were predominantly focused on schema consistency, ingestion verification, and persistence correctness across processing layers. However, as analytical transformations became more complex and domain-dependent, the validation process progressively expanded toward temporal consistency analysis, volumetric behavior interpretation, semantic contextualization, and analytical plausibility assessment. This evolution revealed that analytical QA requires continuous adaptation as the understanding of the pipeline and its operational semantics matures over time.

The observed validation scenarios demonstrated that analytical inconsistencies frequently propagate across multiple layers without generating explicit software failures or ingestion interruptions. In several situations, the analytical impact of the inconsistencies only became evident after downstream aggregations, derived calculations, or dashboard-level consumption exposed unexpected behaviors. This characteristic reinforced the importance of combining structural validations with contextual, temporal, and analytical consistency analyses throughout the entire pipeline lifecycle.

Several inconsistencies identified during the project could not be classified as conventional software defects. In multiple scenarios, behaviors initially interpreted as failures were later validated as expected outcomes after alignment with domain specialists and data engineering teams. Examples included volumetric reductions caused by incremental persistence strategies, duplicated records associated with different temporal granularities, and null values intentionally preserved to maintain historical consistency. These observations reinforced the importance of contextual and semantic validation approaches in industrial analytical environments, where data interpretation frequently depends on operational rules and domain-specific analytical requirements.

These situations also demonstrated that industrial analytical pipelines frequently contain behaviors that cannot be interpreted exclusively through generic QA criteria. In practice, the distinction between analytical defects and semantically valid behaviors often depended on operational context, persistence strategies, and temporal aggregation rules negotiated between QA, Data Engineering, and analytical stakeholders.

Another important observation was the progressive transformation of exploratory analyses into reproducible and systematic validation procedures. Throughout the project, the QA process increasingly relied on SQL-based validation scenarios, cross-layer comparisons, and operational traceability mechanisms capable of supporting more reliable investigation workflows. This process also expanded the role of QA teams, requiring closer collaboration with data engineering, analytical teams, and domain specialists in order to correctly interpret analytical behaviors and transformation rules distributed throughout the pipeline lifecycle. Within this methodology, ingestion integrity is treated as a specific traceability concern because it focuses on ensuring that transferred artifacts remain complete, stable, and verifiable before entering downstream processing stages.

To further contextualize the impact of the proposed QA approach, the adoption of systematic validation dimensions contributed to measurable improvements in pipeline reliability and anomaly detection efficiency. The migration toward reusable SQL-based validation workflows reduced the estimated mean investigation effort associated with data anomalies by approximately 40%. Furthermore, the proactive identification of semantic, temporal, and ingestion-related inconsistencies prevented multiple analytical discrepancies from reaching downstream Gold-layer datasets and dashboarding artifacts. Among the observed issues, semantic inconsistencies accounted for approximately 25% of all validation findings, highlighting the relevance of contextual validation strategies within analytical data architectures. These results suggest that risk-oriented

Table 1: Examples From Observed Validation Issues in the Analytical Pipeline

Category	Observed Problem	Impact	Outcome
Structural	Numerical fields persisted as STRING values	Loss of analytical precision and downstream calculation reliability	Need for stricter typing strategies
Volumetric	Aggressive reduction between Bronze and Silver layers	Initial suspicion of unintended data loss	Incremental persistence behavior validated
Temporal Consistency	Temporal duplication causing division-by-zero in analytical views	Failure in derived analytical calculations	Adjustment of analytical transformation logic
Semantic	Duplicated records initially interpreted as inconsistencies	Ambiguity in interpretation and validation criteria	Refinement of contextual validation rules
Analytical	Inconsistencies in derived analytical calculations	Impact on dashboards and analytical consumption	Correction of transformation logic
Traceability	Need to preserve record origin and transformation lineage	Loss of traceability across analytical layers	Inclusion of technical metadata and lineage controls

validation mechanisms can substantially improve analytical trustworthiness, operational traceability, and the overall robustness of modern data pipelines.

5 Threats to Validity

This experience report is based on a single industrial project conducted within the hydroelectric power generation domain. As such, although the reported validation challenges are consistent with characteristics commonly found in modern Lakehouse architectures, the proposed methodology has not yet been evaluated in other industrial domains or analytical platforms. Therefore, the generalizability of the reported observations should be interpreted with caution. Moreover, the proposed QA strategy evolved incrementally throughout the project as the team gained experience with the analytical pipeline and its domain-specific transformation rules. Consequently, part of the observed improvements may be associated with the team's learning curve rather than exclusively with the validation methodology itself. In addition, the approach requires continuous collaboration between QA engineers, Data Engineering teams, and domain specialists, increasing the validation effort and operational cost. Finally, several situations initially classified as defects were later recognized as expected analytical behavior, reinforcing that human validation remains essential to reduce false positives and correctly interpret domain-dependent analytical results.

6 Conclusion

The industrial experience reported in this work demonstrated that modern analytical pipelines require QA strategies capable of simultaneously addressing structural, temporal, volumetric, semantic, analytical, and operational quality concerns. The proposed risk-oriented validation methodology contributed to improving anomaly detection, reducing silent analytical inconsistencies, and increasing confidence in analytical consumption layers. These lessons suggest that QA practices in data-intensive industrial environments must progressively incorporate observability, contextual interpretation, and cross-layer traceability as central elements of analytical quality assurance. Although the reported results are encouraging, they should be interpreted in light of the limitations discussed in Section 5, particularly the single-project evaluation and the dependence on continuous human validation. Future work will focus on automating SQL-based semantic validations and integrating observability tools for real-time data monitoring across the Lakehouse layers.

Artifact Availability

Due to confidentiality agreements and data privacy policies regarding the industrial partner's critical infrastructure, the datasets and full raw prompts used in this study are not publicly available.

Acknowledgements

This work was supported by the Atlântico Institute. We used Large Language Models for language refinement, but all analysis and interpretation remain our sole responsibility.

References

[1] Syed Muhammad Fawad Ali and Robert Wrembel. 2017. From conceptual design to performance optimization of ETL workflows: current state of research and open problems. *The VLDB Journal* 26, 6 (2017), 777–801.
[2] Michael Armbrust, Ali Ghodsi, Reynold Xin, Matei Zaharia, et al. 2021. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR*, Vol. 8. 28.
[3] Databricks. 2023. *What is the Medallion Architecture?* https://www.databricks.com/blog/what-is-medallion-architecture
[4] Amir Gandomi and Murtaza Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management* 35, 2 (2015), 137–144.
[5] Ahmed A Harby and Farhana Zulkernine. 2025. Data lakehouse: a survey and experimental study. *Information Systems* 127 (2025), 102460.
[6] Martin Kleppmann. 2019. *Designing Data-Intensive Applications.* O'Reilly Media.
[7] Favour Olaoye and Kaledio Potter. 2024. AI-Driven Anomaly Detection in Critical Infrastructure. EasyChair Preprint 14553.
[8] Roger S Pressman. 2005. *Software engineering: a practitioner's approach.* Palgrave Macmillan.
[9] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015).
[10] Richard Y Wang and Diane M Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12, 4 (1996), 5–33.