

A FAIR-Compliant Data Lakehouse for Seismological Workflows: Architecture and Experimental Evaluation

Elidiel Dantas da Costa

Federal University of Rio Grande do Norte (UFRN)

Martin A. Musicante

`martin.musicante@ufrn.br`

Federal University of Rio Grande do Norte (UFRN)

software

Keywords: Seismological Data Lakehouse, FAIR data principles, Data lineage, Traceability, Earth science informatics, Seismological workflows.

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9519288/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A FAIR-Compliant Data Lakehouse for Seismological Workflows: Architecture and Experimental Evaluation

Elidiel Dantas da Costa^{1*} and Martin A. Musicante¹

^{1*}Department of Informatics and Applied Mathematics, Federal University of Rio Grande do Norte (UFRN), Natal, RN, Brazil.

*Corresponding author(s). E-mail(s): elidieldantas@ppgsc.ufrn.br;
Contributing authors: martin.musicante@ufrn.br;

Abstract

Processing seismological data poses significant challenges for its integration, governance, and reuse within scientific workflows. Existing infrastructures often lack unified mechanisms to ensure reproducibility, traceability, and operational efficiency. This study proposes a domain-oriented Data Lakehouse architecture that supports seismological workflows by integrating data ingestion, storage, processing, and access within a unified, scalable framework. A transversal governance layer embeds metadata management and end-to-end lineage tracking as intrinsic components, enabling traceability, interoperability, and reproducibility. The architecture is implemented using open-source Big Data technologies, including Apache Hadoop, Apache Spark, Apache Airflow, and Apache Atlas, combined with domain-specific tools for seismic data processing. A proof-of-concept using real-world continuous seismic data streams from Brazilian monitoring stations is used to validate our proposal. The evaluation adopts a multi-method approach, combining quantitative analysis, expert-based assessment, and Fuzzy Comprehensive Evaluation. Experimental results show that the architecture efficiently processes heterogeneous seismic datasets while reducing manual intervention and improving data organization and traceability compared to traditional semi-manual workflows. Experts reported high acceptance levels, with average scores above 4.6 (of 5), and 90% expressed willingness to adopt the proposed solution. These findings demonstrate that the proposed architecture provides a scalable, reproducible, and FAIR-aligned framework for seismological data management.

Keywords: Seismological Data Lakehouse; FAIR data principles; Data lineage; Traceability; Earth science informatics; Seismological workflows.

1 Introduction

The exponential growth in the volume, heterogeneity, and complexity of geoscientific data has introduced significant challenges for data integration, governance, and reuse. Seismology, in particular, generates continuous streams of waveform data, sensor metadata, and event catalogs, typically stored in heterogeneous formats such as MiniSEED and SEED [1, 2]. This diversity often leads to fragmented workflows and isolated data silos, limiting interoperability and hindering scientific reproducibility.

Although Brazil lies on the South American Plate, intraplate seismicity has been consistently observed over recent decades. Historical events, such as the João Câmara (1986) and Pacajus (1980) earthquakes, motivated the establishment of permanent seismic monitoring initiatives [3, 4]. In this context, institutions such as the Brazilian Seismographic Network (RSBR) and the Seismological Laboratory of the Federal University of Rio Grande do Norte (LabSIS/UFRN) play a central role in the acquisition, curation, and dissemination of seismic data [5–7]. However, the continuous increase in data volume, velocity, and heterogeneity requires more advanced infrastructures capable of supporting efficient integration, governance, and long-term reuse.

Recent studies have explored the adoption of Big Data technologies and data lake-based architectures to address these challenges. For example, the INSIDE platform leverages distributed frameworks such as Hadoop, Spark, and Kafka for large-scale data processing [8], but lacks domain-specific governance mechanisms. The Data-verses approach emphasizes FAIR-aligned and researcher-centric data curation [9], yet remains largely domain-agnostic. Other works focus on organizational strategies to mitigate the emergence of data swamps [10], but do not address semantic metadata management or domain-specific scientific workflows. More recent proposals, such as FLAMI [11] and BigFAIR [12], introduce FAIR-compliant lakehouse architectures and improve interoperability; however, they still present limitations in automation, end-to-end lineage tracking, and validation in real operational environments.

Despite these advances, a critical gap remains: existing solutions do not provide unified, domain-oriented infrastructures that natively integrate FAIR principles—findability, accessibility, interoperability, and reusability—into seismological data workflows while simultaneously supporting automation, traceability, and scalability in operational contexts [13, 14]. This limitation is particularly critical in regional monitoring systems, where workflows are often semi-manual, fragmented, and difficult to reproduce, hindering both scientific analysis and decision-making processes.

To address this gap, this study proposes a domain-oriented Seismological Data Lakehouse architecture tailored to the workflows of Northeastern Brazil. The proposed solution integrates automated data ingestion, distributed processing, metadata governance, and end-to-end lineage tracking within a cohesive and scalable framework, enabling the systematic management of seismological data from acquisition to final analytical products.

The main contributions of this work are as follows:

- The design and implementation of a domain-specific Data Lakehouse architecture that integrates ingestion, storage, processing, and access within a unified framework;

- The incorporation of metadata governance and end-to-end data lineage as intrinsic architectural components, enabling traceability and reproducibility;
- A multi-method evaluation combining quantitative analysis, expert-based assessment, and Fuzzy Comprehensive Evaluation to validate the proposed solution;
- An experimental validation using real-world seismic data, demonstrating the operational feasibility and practical applicability of the architecture.

By bridging the gap between data engineering and domain-specific scientific workflows, this work advances the state of the art in Earth science informatics by providing a reproducible, scalable, and FAIR-aligned infrastructure for seismological data management, while also supporting the broader adoption of open science practices.

2 Methods

2.1 Research Design

This study adopts the Design Science Research (DSR) paradigm to guide the development and evaluation of a computational artifact, namely a FAIR-compliant Seismological Data Lakehouse architecture. DSR is particularly suitable for this research as it focuses on the creation and rigorous evaluation of artifacts designed to solve identified problems in real-world contexts.

The research is structured into three main phases: (i) problem identification and requirements definition, based on the analysis of limitations in existing seismological workflows; (ii) artifact design and implementation, involving the development of the proposed architecture; and (iii) evaluation, conducted through a combination of experimental validation and expert-based assessment.

This approach ensures both practical relevance and methodological rigor, as the artifact is not only theoretically grounded but also validated under realistic operational conditions.

2.2 LabSIS/UFRN Operational Workflow

The workflow currently adopted by the Seismological Laboratory of the Federal University of Rio Grande do Norte (LabSIS/UFRN) is semi-automated, involving data acquisition, processing, validation, and dissemination.

Seismic data are continuously acquired from RSBR stations and transmitted in MiniSEED format to local servers. These data are processed using SeisCompP3¹ to generate seismograms and perform preliminary event detection.

Despite partial automation, critical steps such as phase picking, event validation, and data consolidation still require manual intervention, typically involving spreadsheets and external communication tools. This results in data fragmentation, limited interoperability, and lack of end-to-end traceability.

¹SeisCompP3 is an open-source software framework for real-time seismological data acquisition, processing, and analysis, developed by GFZ German Research Centre for Geosciences. Available online: <https://www.seiscomp.de>

Figure 1 illustrates the integration of this workflow into the proposed Data Lakehouse architecture, where manual and disconnected processes are replaced by an automated and unified pipeline.

This workflow is used in this study as a baseline for comparison, enabling the assessment of improvements introduced by the proposed architecture in terms of automation, traceability, and data governance.

2.3 Data Sources

The dataset used in this study was obtained from LabSIS/UFRN and includes records from RSBR stations and partner institutions [5]. The data comprise:

- Waveform data in MiniSEED format (triaxial sensors HHE, HHN, HHZ);
- Station metadata in dataless format;
- Seismic event records extracted from SeisComP3;
- Complementary historical datasets for regional seismic analysis [4, 6].

These datasets are characterized by high volume, heterogeneity, and structural variability, underscoring the need for an integrated architecture that supports standardized processing, metadata management, and reuse.

2.4 Proposed Architecture

The proposed architecture (Figure 2) instantiates the DSR artifact to address the limitations identified in traditional seismological workflows.

The proposed Data Lakehouse architecture is organized into four layers: ingestion, storage, processing, and data access. A transversal governance layer supports all stages, providing metadata management, data lineage, and policy enforcement.

The ingestion layer automates data acquisition using Apache Airflow and Python-based pipelines integrated with ObsPy. The storage layer relies on HDFS, enabling scalable and distributed storage of raw and processed data. The processing layer employs Apache Spark and domain-specific tools such as SeisComP3 for data transformation and analysis. The access layer provides interactive interfaces through JupyterLab, APIs, and visualization tools.

The governance layer, implemented using Apache Atlas, ensures metadata standardization, lineage tracking, and compliance with FAIR principles across all layers.

Table 1 compares the proposed architecture with representative approaches, highlighting its integrated and FAIR-oriented design.

2.5 Data Processing and Curation

Data processing is performed using Apache Spark, enabling the transformation of raw seismic data into structured formats such as JSON. MiniSEED files are converted using Python and ObsPy, followed by cleaning, normalization, and metadata enrichment procedures.

The curation process follows the Digital Curation Centre (DCC) lifecycle model [15], ensuring systematic data management from acquisition to preservation (Figure 3).

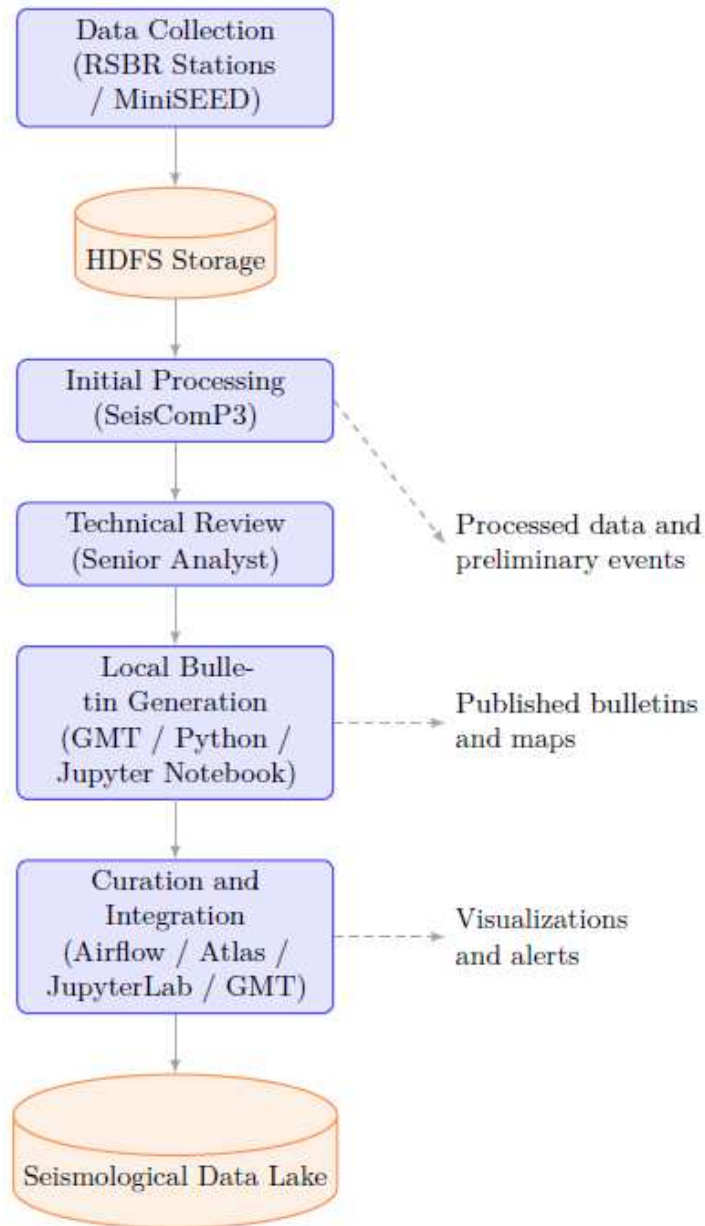


Fig. 1 Operational workflow of LabSIS/UFRN integrated with the Seismological Data Lakehouse.

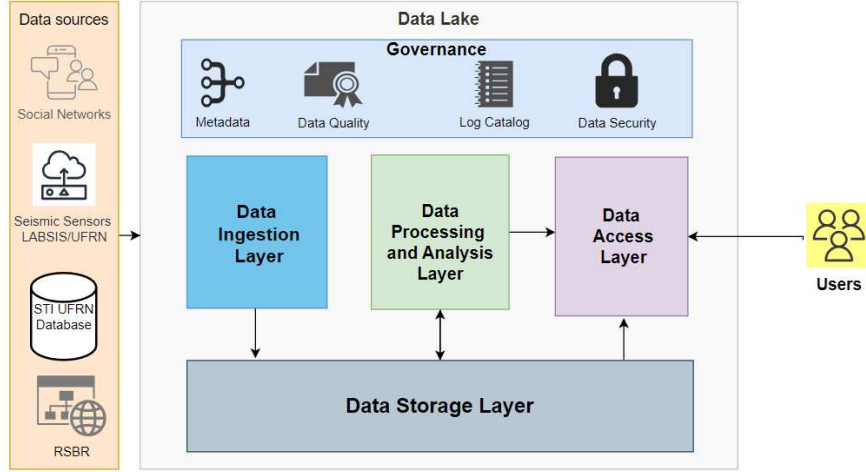


Fig. 2 Proposed Seismological Data Lakehouse architecture for Northeastern Brazil.

Table 1 Comparison between the proposed architecture and related approaches

Layer/Aspect	Proposed Architecture	Related Approaches
Data Collection & Ingestion Storage	Automated orchestration with Airflow, Python, and ObsPy Distributed storage using HDFS	Manual or semi-automated ingestion Centralized or cloud-based storage
Processing	Scalable processing with Spark and SeisComp3	Limited or batch-oriented processing
Data Access	Interactive access via Jupyter-Lab and APIs	Limited access mechanisms
Governance & Lineage	FAIR-aligned metadata and full lineage with Atlas	Limited metadata and lineage support

Metadata are registered and managed in Apache Atlas, establishing explicit relationships between datasets, stations, events, and derived products. Semantic interoperability is supported through the adoption of PROV-O for lineage representation, as well as Dublin Core and DCAT for dataset description. Domain-specific vocabularies based on IRIS/FDSN standards are incorporated to ensure consistency in seismological metadata.

Interactive seismic bulletins with map-based visualization (Figure 4) support geospatial exploration and facilitate data interpretation and reuse.

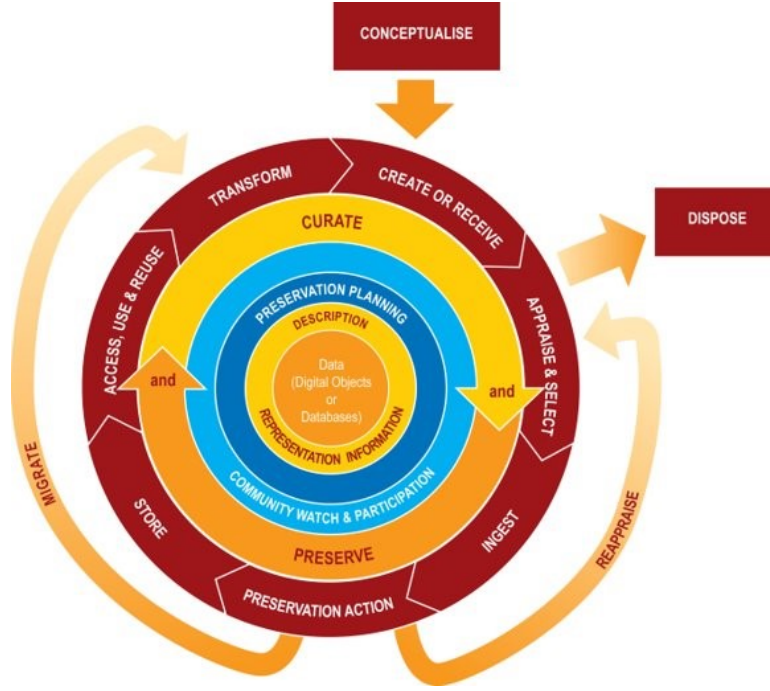


Fig. 3 Digital Curation Centre (DCC) lifecycle model applied to seismological data[15].

2.6 Implementation Environment

The architecture was implemented using open-source technologies, including Apache Hadoop, Apache Spark, Apache Airflow, Apache Atlas, and JupyterLab [8]. These components were integrated into a unified environment supporting scalable storage, distributed processing, workflow orchestration, and metadata governance.

Computational Environment

All experiments were conducted on a local workstation equipped with an Intel Core i5-7200U processor (4 logical cores), 16 GB of RAM, and 1.2 TB of storage. The system runs Fedora Linux 42 (64-bit), with Linux kernel version 6.17.6-200.fc42.x86_64.

Although the infrastructure does not correspond to a distributed cluster, it provides a realistic academic and institutional setting for evaluating the proposed architecture. This configuration demonstrates that the solution is feasible even in resource-constrained environments, while still supporting scalable data processing and metadata governance.

2.7 Evaluation Methodology

The evaluation adopts a multi-method approach to ensure robustness and validity, combining quantitative, qualitative, and fuzzy-based analyses.

A quantitative evaluation was conducted using structured Likert-scale questionnaires administered to 10 domain experts. The questionnaire used for data collection,

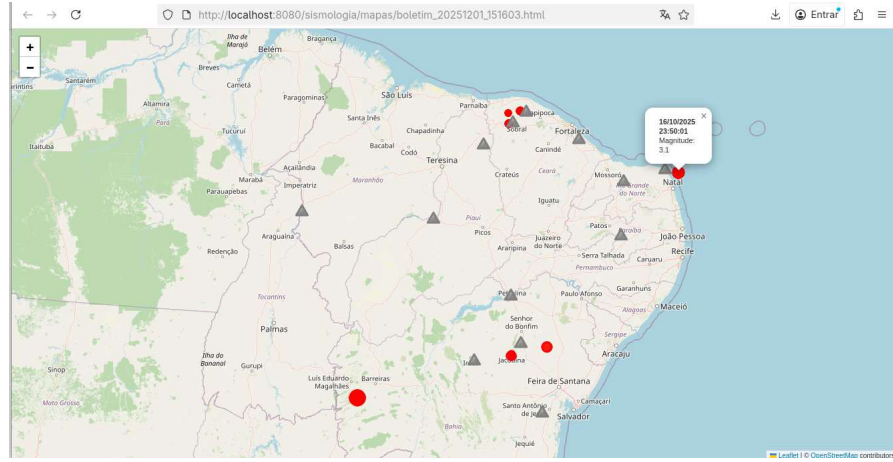


Fig. 4 Interactive seismic bulletin map showing stations (triangles) and seismic events (circles proportional to magnitude).

including the mapping between items ($Q1$ – $Q22$) and evaluation criteria, is provided in Appendix A and is publicly available online². The anonymized responses obtained from the participants are also publicly available³. The responses were originally given in Portuguese because they were collected from native-speaking domain experts and subsequently analyzed by the authors. The questions and answers were automatically translated to English. The assessment covered key dimensions, including performance, scalability, usability, and compliance with FAIR principles. Although the sample size is limited, it is consistent with Design Science Research practices, which commonly employ expert-based validation.

To address subjectivity and uncertainty inherent in human judgment, the results were further analyzed using the Fuzzy Comprehensive Evaluation (FCE) method [16]. This approach enables the transformation of qualitative perceptions into quantitative measures, providing a more nuanced and reliable evaluation.

Qualitative data were collected through open-ended questions, allowing experts to provide detailed feedback on strengths, limitations, and potential improvements. This complements the quantitative analysis by offering contextual insights into the perceived value of the architecture.

In addition, an experimental evaluation using real-world seismic data was conducted to assess performance, scalability, and operational feasibility. This includes analyzing data ingestion efficiency, metadata management, and integration across architectural layers.

FAIR compliance was systematically evaluated across its four core dimensions—findability, accessibility, interoperability, and reusability—based on established guidelines [13, 14].

²Online questionnaire (Google Forms)

³Anonymized evaluation dataset

This multi-dimensional evaluation strategy enhances the reliability and validity of the findings by integrating empirical evidence with expert-based assessment and uncertainty modeling.

3 Results

This section presents a comprehensive evaluation of the proposed Seismological Data Lakehouse architecture. The assessment was conducted with domain experts from Computer Science, Seismology, Data Science, and Geophysics, considering FAIR principles [13, 14], data governance [17], and technical performance.

The results are structured into five complementary dimensions: (i) quantitative evaluation, (ii) experimental validation with real data, (iii) traceability analysis, (iv) fuzzy-based evaluation, and (v) qualitative and FAIR compliance analysis. This multi-method approach provides a robust validation by combining statistical evidence, empirical experimentation, and expert perception. The questionnaire used for data collection, including the mapping between items (Q1–Q22) and evaluation criteria, is provided in Appendix A, ensuring transparency and reproducibility of the assessment.

3.1 Evaluator Profile

The evaluation involved 10 experts with heterogeneous backgrounds and varying levels of professional experience. Table 2 summarizes the distribution of participants across three experience groups.

To increase the assessment’s reliability, a weighting scheme was applied, giving greater weight to more experienced evaluators. This strategy is aligned with Design Science Research practices, where expert knowledge is considered a critical factor in artifact validation [6].

Table 2 Distribution of evaluators by experience group

Group	Experience	N	Weight
Junior	Less than 5 years	3	0.20
Intermediate	6 to 10 years	5	0.35
Senior	More than 10 years	2	0.45

The diversity of expertise ensures that the evaluation captures both technical and domain-specific perspectives, strengthening the external validity of the results.

3.2 Quantitative Evaluation

The quantitative results (Table 3) are derived from Likert-scale responses (Q4–Q18). All evaluated criteria achieved mean scores above 4.5, indicating strong acceptance and high agreement among participants.

The highest-rated criteria (4.8) include findability, accessibility, ingestion, meta-data governance, scalability, and scientific adequacy. These results indicate that the

Table 3 Quantitative evaluation of the Seismological Data Lakehouse architecture

Evaluation Criterion	Average	Std. Dev.
Data findability	4.8	0.4
Metadata clarity	4.7	0.5
Accessibility of data/metadata	4.8	0.4
Interface adequacy	4.6	0.5
Use of recognized standards	4.7	0.5
Integration with external tools	4.7	0.5
Metadata sufficiency for reuse	4.7	0.5
Traceability	4.5	0.7
Data ingestion and organization	4.8	0.4
Metadata governance	4.8	0.4
Pipeline automation	4.7	0.5
Scalability	4.8	0.4
Scientific adequacy	4.8	0.4
Metadata modeling	4.6	0.5
Overall Evaluation	4.6	0.5

architecture effectively addresses core FAIR dimensions, particularly data discoverability and reuse.

The relatively low standard deviations (0.4–0.7) suggest strong consensus among evaluators, reinforcing the robustness of the results. Notably, traceability presented the highest variability (0.7), indicating a more critical perspective among experts. Despite this, its mean score (4.5) remains high, suggesting that while the feature is highly valued, expectations regarding lineage granularity are also elevated.

Overall, the quantitative evaluation demonstrates that the architecture achieves a balanced performance across technical, organizational, and FAIR-related dimensions.

3.3 Proof of Concept

To complement the perception-based evaluation, an experimental validation was conducted using real seismological data from RSBR stations [5, 7]. The objective of this evaluation is to assess the operational performance, scalability, and practical applicability of the proposed architecture under realistic conditions.

The ingestion pipeline processed:

- 825 MiniSEED files (HHE, HHN, HHZ);
- 14 dataless files;
- Approximately 6.6 GB of data covering the period from October 1 to October 22, 2025.

The workflow, orchestrated using Apache Airflow, completed execution in 57.44 minutes, demonstrating the efficiency of the proposed architecture. This corresponds to an average ingestion rate of approximately 115 MB per minute, highlighting the architecture’s ability to efficiently handle continuous, heterogeneous seismic data streams.

All experiments were conducted on a local workstation equipped with an Intel® Core™ i5-7200U processor (4 logical cores), 16 GB of RAM, and 1.2 TB of storage, running Fedora Linux 42 (64-bit). Despite the absence of a distributed cluster environment, the results demonstrate that the proposed architecture delivers robust performance even under resource-constrained conditions.

From an operational perspective, the ingestion process was fully automated, without manual intervention, ensuring the consistency and repeatability of the data pipeline. This represents a substantial improvement in operational efficiency and reproducibility when compared to traditional workflows, which typically involve manual data handling and fragmented processing stages.

Regarding metadata governance, Apache Atlas registered:

- 14 seismological stations;
- 825 ingested data entities;
- 11 seismic events.

These results demonstrate that the architecture supports not only scalable data ingestion but also structured metadata management and consistent data organization. In addition, the integration between ingestion, processing, and governance layers enables the systematic construction of relationships between datasets, stations, and events, which are essential for ensuring data traceability and reproducibility.

To further contextualize these results, a comparison with the traditional workflow adopted at the Seismological Laboratory of UFRN (LabSIS) is presented below.

The traditional workflow relies heavily on manual intervention across multiple stages of the data lifecycle. Seismic data are initially processed using SeisComP3, where event identification is performed visually by a technician. Once detected, events are manually validated and committed to the system database. Subsequently, relevant metadata are informally shared via mobile communication tools, such as WhatsApp, and later transcribed into spreadsheets for further analysis and reporting. Monthly seismic bulletins and event maps are also generated manually, resulting in a fragmented, labor-intensive workflow with limited traceability and reproducibility.

In contrast, the proposed Data Lakehouse architecture introduces a fully integrated and automated pipeline that significantly reduces manual intervention and enhances data management capabilities. Raw seismic data are automatically ingested into HDFS using Apache Airflow and ObsPy-based pipelines. Metadata extraction and cataloging are systematically performed and registered in Apache Atlas, enabling the structured representation of seismological entities such as stations, waveform data, and seismic events. Event data, previously processed by domain experts, are curated and integrated into the architecture, allowing the generation of higher-level entities such as seismic bulletins.

This transformation enables the automated generation of analytical products, including structured datasets (CSV), reports (PDF), and interactive geospatial visualizations. Most importantly, the architecture ensures end-to-end data lineage, allowing complete traceability from raw waveform acquisition to final data products. Compared to the traditional workflow, the proposed approach eliminates fragmentation,

improves reproducibility, and provides a scalable and auditable infrastructure aligned with FAIR principles.

Table 4 summarizes the main differences between the traditional workflow and the proposed architecture.

Table 4 Comparison between the traditional LabSIS workflow and the proposed Data Lakehouse architecture

Aspect	Traditional Workflow (LabSIS)	Our Proposed Architecture
Data ingestion	Manual/semi-automated	Fully automated (Airflow + ObsPy)
Event detection	Visual/manual	Semi-automated and structured
Data sharing	WhatsApp, spreadsheets	Integrated data pipeline
Metadata management	Unstructured (Excel)	Centralized (Apache Atlas)
Traceability	Not supported	End-to-end lineage
Bulletin generation	Manual	Automated
Reproducibility	Limited	High
Data integration	Fragmented	Unified

Furthermore, tasks that previously required several hours of manual coordination are now executed in approximately 57.44 minutes within an automated pipeline, demonstrating a substantial improvement in operational efficiency.

To further validate the architecture’s governance and lineage capabilities, a traceability analysis was conducted using a real seismic event.

The analyzed event occurred in Itapé-Brazil on October 14, 2025 (magnitude 4.5). The traceability chain established in Apache Atlas includes:

- Raw MiniSEED data stored in HDFS;
- Ingestion records linked to specific stations;
- Station metadata (geographical and operational attributes);
- Event metadata (magnitude, time, location);
- Derived analytical products.

This case study demonstrates that the proposed architecture enables end-to-end lineage, ensuring transparency and reproducibility in scientific workflows. Such capability is essential for FAIR compliance, particularly regarding provenance and reusability.

In addition to the experimental results, the outcomes obtained from real-world data processing were validated by domain experts from LabSIS/UFRN, who are directly involved in the operational seismic monitoring workflow. The experts confirmed that the proposed architecture not only preserves data integrity and consistency but also significantly improves data organization, accessibility, and traceability compared to the existing manual, decentralized processes.

From a practical perspective, the transition from a fragmented, human-dependent workflow to an integrated, automated data pipeline resulted in tangible operational gains, including reduced processing time, the elimination of redundant manual tasks,

and improved reliability of seismic data products. These findings reinforce the practical applicability of the proposed approach and demonstrate its effectiveness in addressing real-world challenges in seismological data management.

3.4 Fuzzy Comprehensive Evaluation (FCE)

To account for uncertainty and subjectivity in expert judgments, a Fuzzy Comprehensive Evaluation (FCE) approach was applied [16]. Likert-scale responses were mapped to Triangular Fuzzy Numbers and weighted by evaluator expertise (Junior: 0.20; Intermediate: 0.35; Senior: 0.45).

The defuzzified results confirmed the robustness of the quantitative findings, with all criteria achieving scores above 4.18. The highest values were observed for findability, ingestion, and metadata modeling (4.62), reinforcing the architecture’s effectiveness in structuring and managing complex seismic datasets.

Traceability received the lowest score (4.18), reflecting stricter expert evaluation criteria rather than a functional limitation. Overall, the FCE results provide a more nuanced interpretation by incorporating uncertainty and reinforcing the evaluation’s consistency.

3.5 FAIR Compliance Analysis

The FAIR compliance analysis compared traditional manual workflows with the proposed architecture. Significant improvements were observed in:

- Persistent identification of data;
- Metadata richness and centralization;
- Explicit linkage between data and metadata;
- Provenance and lineage tracking.

These improvements directly support the FAIR principles, particularly findability, accessibility, and reusability. However, semantic interoperability remains partially addressed, indicating the need for broader adoption of domain ontologies and controlled vocabularies.

3.6 Qualitative Evaluation

Qualitative feedback corroborates the quantitative results, highlighting traceability and metadata governance as the most valuable features of the architecture. The use of Apache Atlas was particularly emphasized as a key enabler of transparency and reproducibility.

Automation and integration capabilities were also positively evaluated, especially in reducing manual intervention and improving workflow efficiency. Identified limitations include the complexity of raw data processing and opportunities for improving usability and final data products.

3.7 Adoption Potential and External Validity

The architecture demonstrated strong practical relevance, with 90% of evaluators expressing willingness to adopt it in future projects.

The interdisciplinary profile of participants enhances the external validity of the results, suggesting applicability beyond the specific case study. Although the sample size ($n=10$) is consistent with expert-based evaluation in Design Science Research, future studies should expand validation to larger and more diverse user groups.

Overall, the results indicate that the proposed architecture achieves high levels of acceptance, robustness, and alignment with FAIR principles, representing a significant advancement over traditional seismological data management workflows.

4 Discussion

The results demonstrate that the proposed architecture effectively achieves its primary objectives, particularly in terms of scalability, metadata governance, and alignment with FAIR principles [13, 14]. More importantly, the combined evidence from quantitative evaluation, experimental validation, and comparative analysis indicates that the architecture not only meets these objectives but also provides a substantial improvement over traditional seismological workflows.

The experimental evaluation with real-world data confirms the operational feasibility of the proposed approach. The ingestion of approximately 6.6 GB of heterogeneous seismic data within 57.44 minutes demonstrates that the architecture can efficiently handle realistic workloads while maintaining consistency and automation. When interpreted alongside the comparative analysis presented in Section 3.4, these results highlight a clear transition from fragmented, manually driven processes to an integrated and automated data pipeline. This transition significantly reduces human intervention, minimizes data handling errors, and improves the repeatability of scientific workflows.

A key contribution of this work lies in the explicit implementation of end-to-end traceability. The results show that the architecture enables complete lineage tracking from raw waveform acquisition to derived analytical products, as demonstrated in the traceability case study. This capability directly addresses one of the most critical limitations in traditional seismological data management: the often-incomplete or entirely absent provenance information. In this context, traceability should be understood not merely as a technical feature but as a fundamental requirement for ensuring reproducibility, auditability, and scientific transparency.

From a broader perspective, the high evaluation scores observed across FAIR-related dimensions—particularly findability, accessibility, and reusability—indicate that the proposed architecture effectively mitigates long-standing challenges in data-intensive geosciences, including data fragmentation, limited interoperability, and insufficient governance mechanisms [18, 19]. These findings suggest that the architecture is not only technically robust but also aligned with the increasing demand for FAIR-compliant infrastructures in Earth science informatics.

Table 5 Comparison between the proposed architecture and related works

Aspect	INSIDE	Dataverses	FLAMI	Our Proposed Architecture
Domain focus	Forensic/IT logs	Generic, researcher-centric	Geosciences (partial)	Seismology (domain-specific)
Ingestion	Distributed (Hadoop, Kafka)	Manual/semi-automated	Batch-oriented	Automated (Airflow, ObsPy)
Processing	Large-scale, generic	Limited analytics	FAIR-compliant workflows	Distributed (Spark + SeisComP3)
Governance	Basic logging	FAIR-aligned metadata	FAIR principles, limited lineage	Apache Atlas, full lineage + policies
FAIR compliance	Not explicit	Strong, but domain-agnostic	Partial (interoperability focus)	Strong alignment (F, A, R; partial I)
Automation	Event-driven pipelines	Manual curation	Limited automation	Continuous ingestion + automated pipelines
Validation	Conceptual/technical	Research repositories	Conceptual, limited case studies	Real data + expert evaluation

As summarized in Table 5, the proposed architecture differentiates itself from existing approaches by integrating domain specialization, automation, and governance within a unified framework. While INSIDE emphasizes scalability, it lacks domain-aware metadata and governance mechanisms. Dataverses advances FAIR data publication but remains largely domain-agnostic, limiting its applicability to specialized scientific workflows. FLAMI introduces FAIR principles within a lakehouse paradigm but presents limitations in automation and validation in real operational contexts. In contrast, the proposed architecture combines continuous ingestion, distributed processing, semantic metadata management, and end-to-end lineage tracking, thereby addressing both technical and domain-specific requirements cohesively.

An important observation concerns the evaluation of traceability. Although it received comparatively lower scores in quantitative and fuzzy analyses, qualitative feedback consistently identified it as one of the most valuable features of the architecture. This apparent discrepancy suggests that experts apply stricter evaluation criteria for traceability because of its critical role in scientific workflows. In practice, even partial improvements in lineage tracking can significantly enhance transparency and reproducibility, reinforcing the importance of this capability.

More broadly, the results indicate a shift from traditional, manual workflows toward integrated, automated data ecosystems. By consolidating ingestion, processing, governance, and access within a unified architecture, the proposed solution reduces operational complexity while improving consistency and reproducibility. This transformation is particularly relevant in the context of natural hazard monitoring, where timely, reliable, and traceable data are essential for decision-making processes [6].

Another key contribution is integrating data governance as a core architectural component rather than an external layer. By embedding metadata management, lineage tracking, and cataloging directly into the Data Lakehouse, the architecture aligns with emerging data engineering paradigms that position governance as a central element of data systems [20]. This design choice strengthens FAIR compliance and

supports the development of sustainable, interoperable, and scalable scientific data infrastructures.

4.1 Research Limitations

Despite the positive results, several limitations must be acknowledged in order to properly contextualize the scope and applicability of the proposed architecture.

First, the evaluation was conducted with a relatively small sample of ten domain experts. Although this is consistent with Design Science Research practices, it limits the statistical generalizability of the findings. The results should therefore be interpreted as indicative of expert perception rather than definitive evidence of broad adoption potential.

Second, the experimental validation was performed within a specific regional context, focusing on seismological data from Northeastern Brazil. While the architectural design is conceptually transferable, its adoption across different institutional, geological, or operational environments may require adaptation, particularly regarding data heterogeneity, infrastructure constraints, and organizational workflows.

Third, semantic interoperability remains only partially addressed. Although the architecture adheres to established standards such as PROV-O and DCAT, it does not yet fully incorporate domain ontologies for seismology. As a result, interoperability is supported at a structural level, but could be further enhanced through richer semantic representations and controlled vocabularies.

Fourth, although end-to-end data lineage is supported, the granularity and level of automation of traceability mechanisms can still be improved. In particular, more fine-grained provenance tracking and automated lineage capture across all processing stages would further strengthen reproducibility and auditability.

Finally, the current implementation does not fully incorporate real-time streaming or advanced analytical capabilities, such as machine learning-based event detection or anomaly identification. These features are increasingly relevant in modern seismological systems, particularly for early warning and rapid response scenarios.

These limitations highlight important directions for future work, including integrating real-time data processing frameworks, adopting domain-specific ontologies, expanding experimental validation to larger, more diverse user groups, and incorporating advanced analytics techniques. Addressing these aspects will further enhance the robustness, scalability, and applicability of the proposed architecture in broader Earth science contexts.

5 Conclusion

This paper presented the design, implementation, and evaluation of a Seismological Data Lakehouse architecture to support the integration, curation, and governance of heterogeneous seismic data in Northeastern Brazil. By evolving from a traditional Data Lake paradigm to a Data Lakehouse model, the proposed approach enables the integration of scalable data processing with FAIR-aligned metadata management within a unified framework.

The results provide consistent evidence that the architecture effectively supports the end-to-end lifecycle of seismological data through an automated and integrated pipeline. The combination of experimental validation, expert-based evaluation, and Fuzzy Comprehensive Evaluation demonstrates not only technical feasibility but also strong acceptance among domain specialists. In particular, the ability to efficiently process heterogeneous datasets while ensuring metadata consistency and end-to-end traceability highlights the operational relevance of the proposed solution.

A central contribution of this work is the explicit integration of data governance and lineage as intrinsic architectural components. Unlike existing approaches that treat governance as an auxiliary layer, the proposed architecture embeds metadata management, provenance tracking, and data organization directly into the data lifecycle. This design choice addresses critical limitations in traditional seismological workflows, particularly regarding fragmentation, limited reproducibility, and lack of transparency.

Furthermore, the architecture provides a domain-oriented and extensible framework that can be adapted to other Earth science applications facing similar data management challenges. By aligning with FAIR principles and supporting automated data pipelines, the proposed solution contributes to the development of reproducible, scalable, and interoperable scientific data infrastructures.

Despite these contributions, some limitations remain. The evaluation was conducted with a relatively small number of experts, and semantic interoperability is only partially addressed through existing standards. In addition, real-time data streaming and advanced analytics capabilities, such as machine learning-based event detection, are not yet fully integrated.

Future work will focus on extending semantic interoperability through domain-specific ontologies, improving the granularity and automation of lineage mechanisms, and incorporating real-time processing and advanced analytical techniques. These enhancements are expected to further strengthen the architecture’s applicability in scenarios such as early warning systems and large-scale seismic analysis.

In conclusion, this study demonstrates that integrating FAIR principles, scalable processing, and data governance within a Data Lakehouse architecture is not only feasible but also highly effective in addressing the challenges of modern seismological data management. The proposed approach represents a significant step toward the development of transparent, reproducible, and operationally viable data infrastructures for Earth science informatics.

Supplementary information. Supplementary material related to this article, including the evaluation questionnaire and the anonymized responses dataset, is available online as described in the Data Availability section.

Acknowledgements. The authors would like to thank the Seismological Laboratory of the Federal University of Rio Grande do Norte (LabSIS/UFRN) for providing access to seismic data and technical support. The authors also acknowledge the domain experts who participated in the evaluation process and contributed valuable feedback.

Declarations

Funding

This study was partially supported by the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) through a graduate scholarship awarded to the first author (September 2025–February 2026).

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The seismological data used in this study were obtained from the Seismological Laboratory of the Federal University of Rio Grande do Norte (LabSIS/UFRN). These data are available only under license and are not publicly available. These data are available from the authors upon reasonable request and with permission of LabSIS/UFRN.

The evaluation materials, including the questionnaire and the anonymized responses dataset, are publicly available at:

- [Online questionnaire \(Google Forms\)](#)
- [Anonymized evaluation dataset](#)

The responses are provided in Portuguese, as they were collected from native-speaking domain experts.

Materials availability

Not applicable.

Code availability

The code and implementation artifacts developed for this study will be made publicly available upon acceptance of the manuscript.

Author contribution

Conceptualization: E.D.C. and M.A.M.; methodology: E.D.C.; software: E.D.C.; validation: E.D.C. and M.A.M.; formal analysis: E.D.C.; investigation: E.D.C.; data curation: E.D.C.; writing—original draft preparation: E.D.C.; writing—review and

editing: M.A.M.; visualization: E.D.C.; supervision: M.A.M.; project administration: M.A.M. All authors have read and agreed to the published version of the manuscript.

Appendix A Evaluation Questionnaire

This appendix presents the questionnaire used to evaluate the proposed Seismological Data Lakehouse architecture. The instrument was originally developed in Portuguese and subsequently translated into English to support broader dissemination and reproducibility.

The questionnaire and the anonymized responses collected from domain experts are publicly available at:

- Questionnaire: [Online questionnaire \(Google Forms\)](#)
- Anonymized responses dataset: [Anonymized evaluation dataset](#)

The responses are provided in Portuguese because they were collected from native-speaking participants and subsequently analyzed by the authors.

The questionnaire is structured into six sections, covering evaluator profile, FAIR principles, technical aspects, general evaluation, and qualitative feedback. Responses were collected using a 5-point Likert scale, where 1 corresponds to *Strongly disagree* and 5 to *Strongly agree*.

The explicit mapping between questionnaire items (*Q1–Q22*) and the evaluation criteria discussed in Section 3 ensures full transparency and traceability of the analysis.

A.1 Evaluator Profile

- Q1.* Area of expertise: Seismology; Geophysics; Computer Science; Data Science; Other
- Q2.* Years of professional experience: Less than 5 years; Between 5 and 10 years; More than 10 years
- Q3.* Level of experience with seismological data: Basic; Intermediate; Advanced

A.2 FAIR Principles Evaluation

Scale: 1 = Strongly disagree 5 = Strongly agree

Findable:

- Q4.* Seismological data are easily findable in the Data Lake by station, event, or time period.
- Q5.* The registered metadata clearly identifies the content and origin of the data.

Accessible:

- Q6.* Access to data and metadata is clear and technically feasible.
- Q7.* Interfaces and APIs provide adequate means to retrieve data and associated information.

Interoperable:

- Q8.* The architecture adopts recognized standards for data description and interoperability (e.g., PROV-O, DCAT, FDSN/IRIS).
- Q9.* Data can be easily integrated with external tools and systems (e.g., ObsPy, GMT).

Reusable:

- Q10.* The available metadata are sufficient to enable data reuse in new studies.
- Q11.* The recorded traceability allows reproduction of results and derived products.

A.3 Technical Evaluation

- Q12.* The architecture facilitates the ingestion and organization of large volumes of seismological data.
- Q13.* Metadata governance contributes to data curation and reliability.
- Q14.* Pipeline automation improves efficiency and reliability of data processing.
- Q15.* The architecture is scalable for future growth in data volume and new sources.

A.4 General Evaluation

- Q16.* The proposed solution meets the specific needs of scientific and seismological data.
- Q17.* The metadata model is adequate to represent seismic events, stations, and processes.

A.5 Evaluator Perception

Scale: 1 = Very poor ... 5 = Excellent

- Q18.* Overall evaluation of the architecture.
- Q19.* What do you consider the main strengths of the proposed architecture?
- Q20.* What aspects could be improved or further developed?
- Q21.* Would you adopt this architecture in future research or operational projects? (Yes/No/Maybe)

A.6 Final Comments

- Q22.* Additional comments, suggestions, or general observations.

References

- [1] International Federation of Digital Seismograph Networks (FDSN): miniSEED 3 Specification. Accessed: 2026-04-13 (2023). <https://docs.fdsn.org/projects/miniseed3/en/latest/>

- [2] International Federation of Digital Seismograph Networks (FDSN): SEED Reference Manual: Standard for the Exchange of Earthquake Data, Version 2.4. Accessed: 2026-04-13 (2012). <https://www.fdsn.org/pdf>
- [3] Correia, P.B.: Origem dos terremotos no nordeste. *ComCiência* **117** (2010)
- [4] Pereira, L.L.S.: Terremotos de João Câmara/RN dos anos 1980: um contexto epistemológico à luz da teoria antropológica do didático. Master's thesis, Universidade Federal do Rio Grande do Norte, Natal, Brazil (2023)
- [5] Carvalho, J.M., Barros, L.V., Zahradník, J., Von Huelsen, M.G., Ferreira, V.M.: Earthquake relocation, focal mechanism and stress field determination in central Brazil. *Journal of South American Earth Sciences* **97**, 102411 (2020) <https://doi.org/10.1016/j.jsames.2019.102411>
- [6] Fonseca, J.A.S., Lasocki, S., Nascimento, A.F.: Probabilistic estimation of the source component of seismic hazard in north-eastern Brazil. *Heliyon* (2024)
- [7] Borges, R.G., Assumpção, M.S., Almeida, M.C.F., Almeida, M.d.S.S.: Seismicity and seismic hazard in the continental margin of southeastern Brazil. *Journal of Seismology* **24**(6), 1205–1224 (2020) <https://doi.org/10.1007/s10950-020-09941-4>
- [8] Valle, J.M., Souza, G., Fidelis, S., Araújo, A., Cacho, N., Silva, I., Budke, J., Sales, H., Filgueira, M.W., Lopes, F.: Big data platform for analysing crime evidences. In: *Proceedings of the 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 113–119. IEEE, Oxford, UK (2020). <https://doi.org/10.1109/BigDataService49289.2020.00025>
- [9] Vargas-Solar, G., Darmont, J., Adorjan, A., Espinosa-Oviedo, J.A., Hara, C., Loudcher, S., Motz, R., Musicante, M., Zechinelli-Martini, J.-L.: Dataversifying Natural Sciences: Pioneering a Data Lake Architecture for Curated Data-Centric Experiments (2024). <https://arxiv.org/abs/2403.20063>
- [10] Karkalikova, E., Murdjeva, A.: Organization of data in data lake: Real-life practice. In: *Proceedings of the International Conference on Advanced Computer Information Technologies*, p. 407 (2021)
- [11] Castro, J.P.C., Vasconcelos, G.F.X., Vargas-Solar, G., Aguiar, C.D.: Building fair-compliant lakehouses with flami. In: *Proceedings of the International Conference on Advanced Information Systems Engineering*, pp. 113–129. Springer, Cham, Switzerland (2025)
- [12] Castro, J.P.C., Romero, L.M.F., Carniel, A.C., Aguiar, C.D.: The bigfair architecture: Enabling big data analytics in fair-compliant repositories. *ACM Journal of Data and Information Quality* **17**(4), 1–28 (2025) <https://doi.org/10.1145/3774755>

- [13] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., Silva Santos, L.B., Bourne, P.E., *et al.*: The fair guiding principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016) <https://doi.org/10.1038/sdata.2016.18>
- [14] Jacobsen, A., Kaliyaperumal, R., Silva Santos, L.O.B., Mons, B., Schultes, E., Roos, M., Thompson, M.: A generic workflow for the data fairification process. *Data Intelligence* **2**(1-2), 56–65 (2020) https://doi.org/10.1162/dint_a_00028
- [15] Digital Curation Centre: Digital Curation. Accessed: 2025-01-02 (2025). <https://www.dcc.ac.uk/about/digital-curation>
- [16] He, Z., Fang, W.: Research data management in institutional repositories: An architectural approach using data lakehouses. *Digital Library Perspectives* **41**(1), 145–178 (2025) <https://doi.org/10.1108/DLP-02-2024-0022>
- [17] Boukraa, D., Bala, M., Rizzi, S.: Metadata management in data lake environments: A survey. *Journal of Library Metadata* **24**(4), 215–274 (2024) <https://doi.org/10.1080/19386389.2024.2359310>
- [18] Miloslavskaya, N., Tolstoy, A.: Big data, fast data and data lake concepts. In: *Procedia Comput. Sci.*, vol. 88, pp. 300–305 (2016). <https://doi.org/10.1016/j.procs.2016.07.439>
- [19] Nambiar, A., Mundra, D.: An overview of data warehouse and data lake in modern enterprise data management. *Big Data Cogn. Comput.* **6**(4), 132 (2022) <https://doi.org/10.3390/bdcc6040132>
- [20] Gorelik, A.: *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science*. O'Reilly Media, Sebastopol, CA, USA (2019)