International Journal of Social Impact
ISSN: 2455-670X
Volume 9, Issue 1, DIP: 18.02.031/20240901
DOI: 10.25215/2455/0901031
www.ijsi.in |January-March, 2024

Research Paper

Data Observability: Ensuring Trust in Data Pipelines

Krishna Prasanth Brahmaji Kanagarla[1]*

ABSTRACT

Data observability is the key process you need to implement to guarantee the consistency of the data pipeline in terms of credibility. The features of data quality, data freshness, lineage, and schema changes checked in real-time help prevent problems before they accumulate, and impact the data pipeline. As the field of data observability is still in its infancy, this paper aims at identifying which components it could be composed of, what tools it could ideally comprise, and what value it can bring to businesses, with an accentuation of recommendations for its implementation in contemporary data workloads.

Keywords: *Data visibility, data accuracy, datasets delivery, constant monitoring, data origin, schema evolution, outliers, decision making process, risks management, work productivity*

I. INTRODUCTION

Data observability is the practice of ensuring the availability, quality, and efficiency of data pipelines with holistic monitoring of information on their wellbeing. It means that it makes it possible for an organization to maintain discipline on the quality, frequency, origin, and even structural evolution of data in real-time, and therefore, any data discrepancies are detected and addressed on time. The conventional approach to monitoring is the system level, which means data is collected on CPU usage, memory and the servers. Data observability, on the other hand, is focused on challenges related to data processing, such as pipeline, data or data age issues, to have trust in data flowing through the pipeline [1]. In the US and globally firms use data for decision-making, new products or services development, and to sustain competitive advantage. Data pipelines may be unreliable and this can cause issues such as wrong results, loss of money and damage to reputation.

II. AIMS AND OBJECTIVES

Aim*:* The multi-faceted phenomenon of data observability for the purpose of stabilizing and solidifying data ecosystems.

Objectives

- To evaluate the project for assessing the strengths and weaknesses of data observability systems in making pipelines trustworthy there is a need to define specific objectives.
- To assess the complex data workflow and compare the efficacy of several data observability tools that can contribute to that realm.

[1]Sara Software Systems, LLC, USA
*Corresponding Author

Received: March 05, 2024; Revision Received: March 18, 2024; Accepted: March 30, 2024

- To explore the data, observability can support better decision-making and reduce some potential risks.
- To provide recommendations on how best to implement data observability frameworks and align them with organization performance.

III. RESEARCH QUESTIONS

- What is the improvement of data pipeline trustworthiness, how well current data observability frameworks fare?
- What does Monte Carlo, Soda, and Data band offer, and what is out of bounds for these data observability tools?
- What does data versatility affect decision-making and risk management in the context of data-oriented enterprises?
- Which strategies are best practice when it comes to the implementation of data observability frameworks at an organization?

IV. LITERATURE REVIEW

The journal aims at identifying the evolution of data pipeline monitoring practices.

There is little innovation about monitoring a data pipeline, it has been with this for several years now, having started off as basic infrastructure monitoring and progressing to address the higher-level complexities of the data workflow. Initial techniques were based mainly on such generic system parameters as CPU utilization, memory occupancy, and server availability to deduce the status of a pipeline [2]. Nevertheless, these metrics merely revealed little of what happened in terms of the data pumped in the pipelines, thus taking a long time to identify problems such as data staleness, data schema changes, or data loss. As the number of organizations turning to data-driven decision-making continues to grow, stakeholders realized that was a requirement for enhanced insight into the functionality of data pipeline facilities. On The other hand, this led to the development of data monitoring tools, some of which were designed to identify relatively simple forms of data activity variation of volume or frequency [3]. These tools helped in enhancing the pipeline reliability however these were not very smart tools to manage some of the associated complexities of deep learning such as lineage of data or run time quality control checks. Data observability has filled this gap effectively by providing full pipeline visibility and giving organizations the necessary tools to address pipeline concerns before they affect downstream pipeline segments or decision making processes.

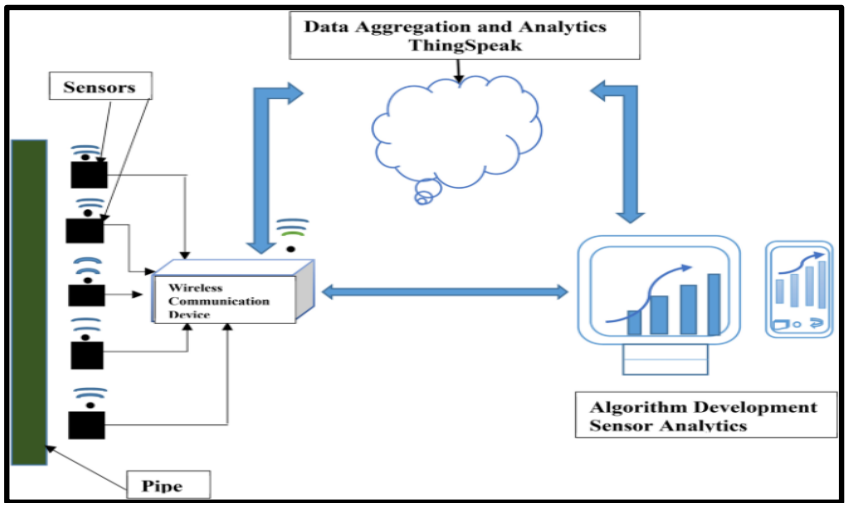

Fig.1. Petroleum pipeline monitoring

Further analysis of some of the tools, such as Monte Carlo, Soda, and Data band

Software tools used today in data observability such as Monte Carlo, Soda, and Data Band have made a significant change when it comes to monitoring data pipelines in organizations. This is particularly useful when data anomalies in terms of volume, schema or distribution have to be detected automatically with the help of machine learning as Monte Carlo does [4]. It has a lineage tracking feature, which helps an organization to isolate the source of the problem related to data. Soda, however, puts simplicity and flexibility on the foreground, focusing on serving as a lightweight tool used for DQ monitoring only. It allows users to create specific checks to key KPIs and works well in the current High performing data environments, thus making it a preferred choice for varied teams. Data band is centered around pipeline monitoring as a means of attaining observability, in addition to offering metadata management. The tracking of resource use in addition to data flow is another special quality that makes the system suitable for organizations, making both performance and cost efficient [5]. With support of the most popular orchestration tools such as Apache Airflow, Data band helps the teams keep control over their data pipelines' operations.

Discussion of Ideas for Observability Frameworks

Observability stems from control theory wherein the exploitation of output data to assess the state of the internal structure is explained. Observability frameworks in the data pipeline space are built to offer insights on the well-being of pipelines via the measurement of data regularity, quality, and provenance aside from this the alteration of the schemas. A robust observability framework relies on three pillars: logs, metrics, and traces [6]. Logs are more detailed records of the activities taking place in pipelines and are useful to try and determine what happened in a certain incident or in case of an error. Metrics can allow concrete data about pipeline performance to be determined including processing rate or throughput volume for example in order to establish a pattern or identify abnormality [7]. Traces are essentially the equivalent of mapping by chronicling the data's movement throughout a flow of interconnected systems. Through integration of these elements the observability frameworks enable organizations to transition from simple monitoring to prevention style of operation in matters that affect business.

Get acquainted with real life monuments made by non-profit organizations from the United States.

There are some organizations that have implemented data observability in organizations based in the United States to improve the stability of data pipelines. A chief e-commerce firm used Monte Carlo to identify and correct data issues in real-time to avoid failure of its recommendation algorithm [8]. Applying automated checks, the company increased its data availability by 80%, and thus received increased customer satisfaction. A financial services firm started using Data band to centrally observe its many data processing pipelines, running in various clouds. The ability of the platform to track lineage helped the firm to understand constraints or issues with pipelines and solve them to offer fast, accurate reporting and decisions. On the other hand, a healthcare analytics provider that incorporated Soda in order to meet tight requirements to data quality [9]. They were able to achieve high accuracy of the data which is paramount for regulatory reporting and the care of patients through the use of checkpoints which the platform gave the provider flexibility to customize. Based on these case studies, it can review and explain how data observability tools strengthened trust and reduced downtime and aided in superior decisions across various industries.

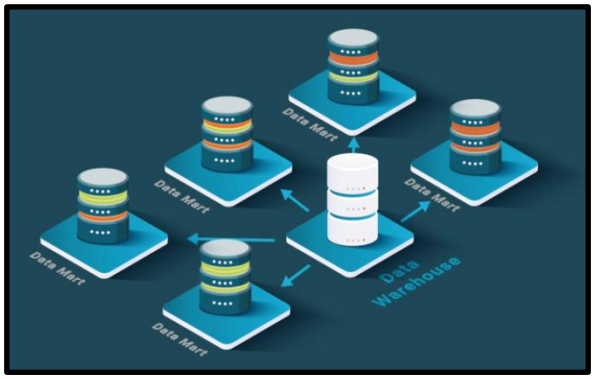

Fig.2. Architectural Patterns for Data Pipelines

Literature Gap

There is rarely a comparative discussion and comparative assessments of tools such as Monte Carlo, Soda, and Data band in existing literature on data observability. These demonstrate their functionalities; few examine intimately the practical applicability of such tools across organizations with reference to integration and cost [10]. Moreover, there is also a low number being written about some of the issues that could occur when applying data observability frameworks to contextually legacy systems or industries that must adhere to specific guidelines such as the financial services industry or the healthcare industry. Even the theoretical proposition of observability lacks consideration of the real-world issues like an appropriate implementation of observability to the organization's objectives. Filling these gaps is important for obtaining recommendations pointing towards the optimization of data observability in practice.

V. METHODOLOGY

The method that is used underpinning this research also called for an interpretivism paradigm, where observability of data is subjected to the author's perception of what the user observes in the specific context in which it finds the user. The interpretivist approach ensures that this study aligns with its goal of uncovering the dynamics and complexity of data observability in organizations [11]. This journal focuses on enriching the existing understanding of organizations' attitudes and practices regarding data observability through the prioritization of ***qualitative information***. The research returns to ***secondary sources*** as the main source of information; gathered from journal articles, industry reports, case studies, and documentation of items such as Monte Carlo, Soda, and Data band. Collection of ***secondary data*** is advantageous since it incorporates the use of time and less cash as compared to the primary data collection method, in addition, secondary data offers a pool of various perceptions as well as better datasets [12]. The data sources are chosen very carefully in order to make sure that the information coming from the sources is relevant, credible and the potential researcher has all the necessary information regarding the chosen topic. The analysis of the collected data a ***qualitative thematic*** data analysis method is used. This method is particularly appropriate for use in interpretivist research since it enables one to make patterns or themes or ***qualitative data***. By applying TAs approach, the complexity of the obtained data sets is summarized in the form of themes that correspond to the objectives of analyzing the components of data observability and its tools, as well as assessing their business effects.

The data analysis involves identifying coding for issues regarding implementation and integration of data observability tools, shortcomings in the organizations, and effects on the decision-making system. ***Thematic analysis*** is especially suitable for secondary data

analysis because of its ability to reorganize identical ideas from various sources into comprehensible themes. Through the use of a **qualitative thematic analysis**, the research obtains a systematic but an open approach to the interpretation of the collected data. This method also renders understandings of how data observability frameworks are conceived and enacted within a variety of settings or practice contexts and the nuances of proceedings that quantitative approaches may well miss [13]. The use of interpretivism means that the views of participants are central, while ***qualitative analysis*** and ***secondary data analysis*** mean that the study can form a robust methodological argument for achieving the objectives and answering the research questions. This way, it avoids subjective results and guarantees that the solutions are applicable and grounded on a variety of organizational experiences together with real-life practice in data observability. It helps to finally generate actionable suggestions for building trust in data pipelines.

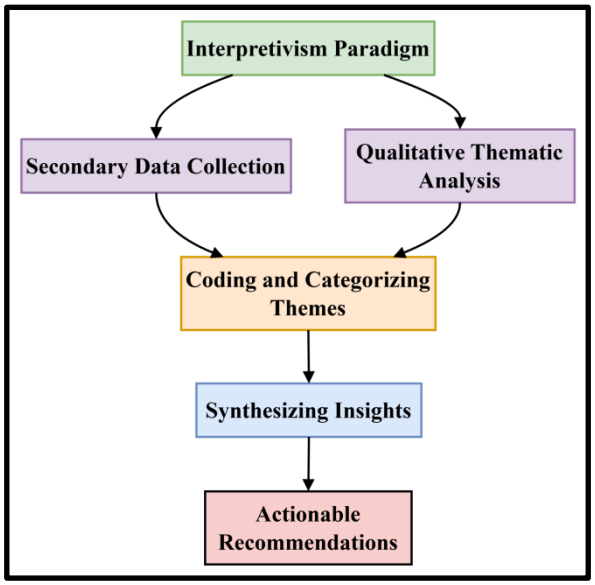

Fig .3. Qualitative research methodology diagram

VI. DATA ANALYSIS

Theme 1: Key aspects of the data observability and their purpose to build trust to the pipeline.

Data quality, data freshness, data lineage, schema changes, and most importantly, the exploration of anomalies are at the center of data observability for building that trust in data pipelines. Data quality is about the quality of the data, in terms of its relevance, validity, and timeliness, in order for organizations to use data to make important decisions. This can be achieved by maintaining the principle of data freshness so as to meet the requirement on data timeliness, which is especially important in real-time applications or analysis [14]. Lineage is a documentation of how data moves from one system to another, and it helps with problem-resolution processes. Schema changes ensure there are no compatibility problems and no pipeline stops, while anomaly detection is used to identify those unusual patterns or trends in the data that could mean the processes are wrong. Collectively, these elements provide a comprehensive perspective of the pipeline health from which companies can be able to confront challenges early enough before negatively affecting the other processes downstream [15]. When combined into a single observability framework, such elements help businesses keep their data accurate and minimize operational risks.

Theme 2: An evaluation of Data Observability Tools and their influence on the Pipeline Reliability.

Tools such as Monte Carlo, Soda, and Data band are among those that have cropped up to offer improved data observability to solve pipeline reliability issues. Monte Carlo automates identification of new outliers in the data, lineage tracking of data assets, and schema validation, providing users with full workflow coverage [16]. Soda offers flexibility, which is supported by the possibility of creating additional checks. This format allows increasing data quality and integrates well with contemporary solutions. Resource monitoring and tracking of data flows, cyber assets, and dependencies in Data band guarantees both effectiveness and quality. These tools enhance reliability by identifying flaws and solving them before they cause any disruptions, reduce pipe time and increase efficiency of pipeline. Still, applicability of the solutions depends on such issues as scalability of the solutions and their degree of integration as well as the price [17]. In critiquing these tools, business entities can be in a position to adapt for the best tool that suits their organizational needs and per merits of pipeline improvement.

Theme 3: The Difference Between Data Observability and Other Monitoring Approaches.

Data observability and traditional monitoring differ in a number of ways in terms of their scope and practices. Conventional monitoring is based on the indicators of absolute system monitoring including servers, CPU, memory, and others mainly directed to identifying possible problems in the IT infrastructure. Additionally, effective for codifying and analyzing organic technical concerns such as downtime or hardware issues, it is insufficient to navigate intricate data processes [18]. Data observability, by contrast, goes a step further and covers aspects related to data, such as quality, age, lineage, and occurrences of deviations. It provides detailed visibility into the movement and actions of data through pipes, so potential data problems can be addressed before negatively affecting other systems. For example, data observability can alert on schema changes or any other deviation from what is expected so that the problems can be solved on time [19]. This differentiation puts data observability as a superior process more tailored to address modern data contexts.

Theme 4: The Impact of Data Observability in Decision Making and Risk Management in Businesses.

Data observability all have a shot at helping improve business decision-making since data applied to analytics and reporting have to be accurate. Thus, organizations adopting data observability frameworks are able to identify and address certain problems before they lead to incorrect conclusions and decisions [20]. This way, observability enhances the control of data quality and pipeline stability and avoids deficiencies which provoke system failures relying on data. For instance, while real-time anomaly detection can be beneficial in time-point security applications like fraud detection or material logistic chain, it could be disadvantageous in other applications of machine tools. In addition, data observability is helpful to regulatory necessities as it ensures the data integrity necessary to avoid punitive measures [21]. These capabilities combined serve to improve the operations effectiveness, minimize risk and offer competitive advantage in core strategic vulnerable areas.

VIII. CONCLUSION

Observability of data is to facilitate trust in data delivery pipelines, data quality, data dependability and data utility. Due to its capability in handling issues more than mere tracking, it enables organizations to come up with the right decision, minimize risks and improve on business flow. Monte Carlo, Soda, and Data band etc. are examples of how it can revolutionize the data management process. Since data remains a key driver of change,

maintaining best data observability practices can remain highly relevant to sustaining business advantage and avoid regulatory pitfalls.

REFERENCES

[1] Shankar, S. and Parameswaran, A., 2021. Towards observability for production machine learning pipelines. arXiv preprint arXiv:2108.13557.

[2] Rupprecht, L., Davis, J.C., Arnold, C., Gur, Y. and Bhagwat, D., 2020. Improving reproducibility of data science pipelines through transparent provenance capture. Proceedings of the VLDB Endowment, 13(12), pp.3354-3368.

[3] Li, B., Peng, X., Xiang, Q., Wang, H., Xie, T., Sun, J., & Liu, X. (2022). Enjoy your observability: an industrial survey of microservice tracing and analysis. Empirical Software Engineering, 27, 1-28.

[4] Therrien, J. D., Nicolaï, N., & Vanrolleghem, P. A. (2020). A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. Water Science and Technology, 82(12), 2613-2634.

[5] Fedushko, S., Ustyianovych, T., & Gregus, M. (2020). Real-time high-load infrastructure transaction status output prediction using operational intelligence and big data technologies. Electronics, 9(4), 668.

[6] Karumuri, S., Solleza, F., Zdonik, S., & Tatbul, N. (2021). Towards observability data management at scale. ACM Sigmod Record, 49(4), 18-23.

[7] Seedat, N., Imrie, F., & van der Schaar, M. (2022). Dc-check: A data-centric ai checklist to guide the development of reliable machine learning systems. arXiv preprint arXiv:2211.05764.

[8] Hurl, B., Cohen, R., Czarnecki, K., & Waslander, S. (2020, October). Trupercept: Trust modelling for autonomous vehicle cooperative perception from synthetic data. In 2020 IEEE Intelligent Vehicles Symposium (IV) (pp. 341-347). IEEE.

[9] Burgess, Mark, and Andras Gerlits. "Continuous integration of data histories into consistent namespaces." arXiv preprint arXiv:2204.00470 (2022).

[10] Schnabel, G., Sjöstrand, H., Hansson, J., Rochman, D., Koning, A. and Capote, R., 2021. Conception and software implementation of a nuclear data evaluation pipeline. Nuclear Data Sheets, 173, pp.239-284.

[11] Sundaram, A., Abdel-Khalik, H.S., Roberson, D. and El Hariri, M., 2022. Data recovery via covert cognizance for unattended operational resilience. Progress in Nuclear Energy, 151, p.104317.

[12] Hartati, S., 2022. Innovative Approaches to Monitoring in Enterprise Production Systems. Sage Science Review of Applied Machine Learning, 5(1), pp.81-96.

[13] Shankar, S., Garcia, R., Hellerstein, J.M. and Parameswaran, A.G., 2022. Operationalizing machine learning: An interview study. arXiv preprint arXiv:2209 .09125.

[14] Bento, A., Correia, J., Filipe, R., Araujo, F. and Cardoso, J., 2021. Automated analysis of distributed tracing: Challenges and research directions. Journal of Grid Computing, 19(1), p.9.

[15] Nestorov, A.M., Berral, J.L., Misale, C., Wang, C., Carrera, D. and Youssef, A., 2022, November. Floki: a proactive data forwarding system for direct inter-function communication for serverless workflows. In Proceedings of the Eighth International Workshop on Container Technologies and Container Clouds (pp. 13-18).

[16] van der Goes, M., 2021, September. Scaling enterprise recommender systems for decentralization. In Proceedings of the 15th ACM Conference on Recommender Systems (pp. 592-594).

[17] Munoz-Arcentales, A., López-Pernas, S., Pozo, A., Alonso, Á., Salvachúa, J. and Huecas, G., 2020. Data usage and access control in industrial data spaces: Implementation using FIWARE. Sustainability, 12(9), p.3885.
[18] Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E. and Sicilia, M.A., 2021. Traceability for trustworthy AI: a review of models and tools. Big Data and Cognitive Computing, 5(2), p.20.
[19] Patel, P. and Uddin, M.N., 2022. AI for algorithmic auditing: mitigating bias and improving fairness in big data systems. International Journal of Social Analytics, 7(12), pp.39-48.
[20] Perrault, A., Fang, F., Sinha, A. and Tambe, M., 2020. Artificial intelligence for social impact: Learning and planning in the data-to-deployment pipeline. AI Magazine, 41(4), pp.3-16.
[21] Haider, L. 2021. Artificial intelligence in ERP (Bachelor's thesis). Metropolia University of Applied Sciences, Finland.

Acknowledgment
The author(s) appreciates all those who participated in the study and helped to facilitate the research process.

Conflict of Interest
The author(s) declared no conflict of interest.

How to cite this article: Kanagarla, K.P.B. (2024). Data Observability: Ensuring Trust in Data Pipelines. *International Journal of Social Impact*, *9*(1), 286-293. DIP: 18.02.031/20240901, DOI: 10.25215/2455/0901031