SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V10I10P102

Volume 10 Issue 10, 6-10, October 2023
© 2023 Seventh Sense Research Group®

Original Article

Connecting the Dots: How Data Lineage Helps in Effective Data Governance

Sivakumar Ponnusamy[1], Pankaj Gupta[2]

[1]*Senior Data Engineer, Cognizant Technology Solutions, Richmond, VA, USA.*
[2]*Principal Data Engineer, Discover Financial Services, USA.*

[1]*Corresponding Author : psivakumarmca@gmail.com*

Received: 21 August 2023 Revised: 30 September 2023 Accepted: 13 October 2023 Published: 31 October 2023

Abstract *- The research provides the intertwined realms of Data Lineage and Data Governance, two crucial facets of contemporary data management within organizations. Data Lineage, as the flow and transformation of data through pipelines, is explored in tandem with Data Governance, which provides the principles and frameworks for effective data management. The article elucidates the prerequisites for successful data governance, highlighting the pivotal role played by executive support, clear business objectives, comprehensive data inventories, and robust security measures. The research further discusses how Data Lineage aligns with the fundamental principles of Data Governance, including data transparency, accountability, quality assurance, security, and regulatory compliance. By tracing the origin and evolution of data, Data Lineage ensures that data is trustworthy and can be relied upon for informed decision-making.*

Keywords *- Data lineage, Data governance, Data quality, Regulatory compliances.*

1. Introduction

In the world of computing, there are various components connected to each other and fulfills the users' objectives. Data lineage and data governance are two critical pillars of effective data management within organizations. Data lineage involves tracing the journey of data as it moves through various processes and transformations, whereas, on the other hand, data governance sets the principles and frameworks for managing data as a valuable asset. Collectively, they ensure data quality, security, compliance, and transparency, all of which are essential for informed decision-making. While considering the significance of Data lineage, there is always a gap in research and development in all domains. Certain reasons describe the importance of data lineage, which is further connected to multiple tools. For instance, to establish effective data lineage, organizations can utilize various tools and methodologies, such as open-source tools like Apache Atlas and Apache NiFi, commercial tools like Informatica Axon and Collibra, or cloud-based tools like AWS Glue Data Brew, Google Cloud Data Catalog, and Microsoft Azure Purview. These tools help organizations track and visualize data, perform impact analysis, and manage metadata effectively.

2. Literature Review

2.1. Data Lineage

In the data pipelines, the data travels and gets transformed into various shapes. The flow and the transformation are referred to as the Data lineage. This is a data management domain where a large amount of information is stored in electronic form. It performs certain roles in organizations when used for commercial purposes; for instance, it helps organizations understand how data is sourced, processed, and consumed. Data lineage includes destination tracking, processing and transformation, data quality and validation, impact analysis, troubleshooting, and debugging [1].

2.2. Data Governance

Data governance is the set of principles and frameworks for effective data management. Data governance assists in terms of data security and regulations related to data. Data governance involves defining processes, roles, responsibilities, and rules for managing and using data as a valuable organizational asset [2].

2.3. Effective Data Governance

Effective data governance is required in every organization to run their operations smoothly and sustainably. Three factors that can always be expected out of good data governance include quality, security, and compliance. Data governance paves the path for data quality assurance, data security, regulatory compliance, risk management, data accountability, cost efficiency, and data interoperability[3]. Data governance works on the principles of establishing data standards, definitions, and validation rules. Data governance ensures that data is accurate, consistent, and dependable. With the increase in modern data collecting and handling, good data governance includes security measures to protect sensitive and confidential information. The data governance principles may vary from one organization to others based on their priorities and the nature of the data collected. For example, the healthcare sector in the USA must collect its data according to HIPAA compliance [4].

2.4. Pre-Requisites for Effective Data Governance

Data governance is not a simple task to conduct, as this is a process which includes various steps and requires pre-conditions before its complete implementation. The result-oriented data governance demands the fulfillment of the conditions that could support its operations. Some of the key pre-requisites for data governance include:

- The Executive support and sponsorships provide the resources and communicate their importance throughout the organization.
- The business objectives should be clear and define the outcomes of the data governance.
- Strict and effective data governance policies must be developed.
- The availability of a comprehensive data inventory and catalog that documents all data assets within the organization [5].
- Appointing the data custodians responsible for overseeing specific data domains or data sets.
- Implementing robust data security measures, including access controls, encryption, data masking, and regular security audits.

Regularly review and adapt data governance policies and procedures to compete with changing business needs and technology advancements [6].

2.5. Principles of Data lineage to support effective Data Governance

Data lineage is the component of data governance, as this is also about the data and its flow through the communication channels. Due to this reason, data lineage supports several key principles and objectives of data governance. One of the basic principles that Data lineage supports is effective Data governance. Based on the research presented by Stuiver & Crompvoets, below are the ten principles,

- Data transparency
- Data accountability
- Data quality assurance
- Data security
- Regulatory compliance
- Data impact analysis
- Data Governance Council Decision-Making
- Data life-cycle management
- Data Versioning
- Data Trust and Confidence [7]

Data transparency is the key requirement of data governance, which is also ensured by the data lineage, as this allows the flow of data within an organization. It enables stakeholders to understand where data comes from, how it is transformed, and where it goes. The tracking and auditing of data usage are necessary for data governance and lineage.

Data lineage assures data accountability by showing who is responsible for various data sets and how data is managed across different processes and systems. This determines the roles and responsibilities, a key aspect of data governance [8].

Data quality is one of the crucial factors in any of the business domains or even in non-business operations. Because good quality data provides the guarantee of good decisions. Data lineage allows the data quality for the organizations to trace data back to its source and identify any transformations or modifications along the way. This capability is essential for identifying and addressing data quality issues [9].

Data security is a major concern for every organization in this world. The abundance of data is available on cloud servers continuously subjected to online threats. Data breaches and alternations are serious concerns that could even invite government actions against the organization. The data lineage is crucial for data security. It assists organizations in identifying points in the data flow where security measures, such as encryption or access controls, should be applied. It also assists in monitoring data access and ensuring compliance with security policies [10].

Data lineage supports the demonstration of regulatory compliance. It enables organizations to show how data is collected, processed, and used, which is necessary for meeting data protection and privacy regulations. It helps in tracking data consent, retention, and usage. The data policies and procedures are developed based on these core principles.

Data impact analysis works when changes are proposed within an organization, whether a system upgrade or a data transformation; data lineage allows for impact analysis. This defines the changes affecting downstream processes and data consumers, facilitating risk assessment and mitigation planning [11].

Data Governance Council Decision-Making has become possible with the help of valuable information for data governance councils or committees when making decisions about data policies, data ownership, and data usage. It ensures that governance decisions are based on a clear understanding of data flow and dependencies.

The data transforms throughout the process of transmission. A complete understanding of the data lineage is required to maintain the data life-cycle management. With the help of that, organizations can develop data retention and archiving policies, ensuring that data is managed throughout its lifecycle in a compliant and efficient manner [12].

Change management is a need for organizations; the transformation of the technological resources available complements this, also due to the change in the processes. The same is true for the data lineage, which satisfies the conditions of data governance. Lineage track changes to data over time. This is important for versioning data and managing changes effectively. It allows organizations to roll back to previous versions if data issues are detected.

Confidence in the data is necessary for good decision-making, and data lineage builds trust and confidence in data. When stakeholders can easily trace the source and transformation of data, they are more likely to trust the information for decision-making, which is a fundamental aspect of Data governance.

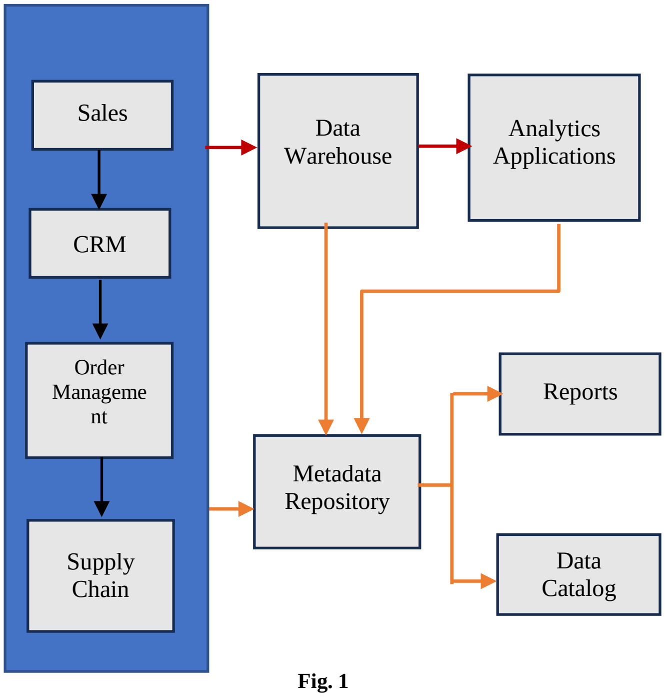

Fig. 1

2.6. Tools and Technologies for Data Lineage

The tools used for the data lineage are based on the principles defined and the nature of the data handling. There are three types of tools for the data lineage, which include open-source tools, commercial tools, and cloud-based tools. The open-source tools commonly used are Apache Atlas and Apache NiFi. The commercial tools used for the data lineage include the Informatica Axon, Collibra, and Erwin Data Intelligence. The commercial tools possess the ability to track and visualize data. The tools used for this purpose can also help provide data showing the impact analysis and metadata management features. There are also cloud-based tools which support the users for the data lineage for the vast usage. Some popular tools are AWS Glue DataBrew, Google Cloud Data Catalog, and Microsoft Azure Purview. These tools help the data governance by paving a path for the Data lineage.

As described in the fig 1. a data-driven supply chain system. This comprises various components such as Sales, CRM, data warehouse, analytical application and data log. The diagram also shows the data flow between the system's different components. For example, sales data flows from the sales component to the CRM component and the data warehouse. The CRM component uses this data to segment customers and develop targeted marketing campaigns. The data warehouse also provides data to the analytics applications, which use it to generate insights into the supply chain.

The tools for the data lineage work on the procedures and processes which comply with the data governance principles. The methods for the data lineage involve examining the technical metadata generated by systems and applications. It includes examining database schemas, ETL (Extract, Transform, Load) processes, data integration workflows, and log files to understand how data moves through the technical infrastructure. Business process mapping is also conducted for this purpose and engages in documenting how data is used and transformed as it moves through different business units and systems. Business process modeling tools can aid in this approach. Data profiling complements the data governance mechanism, and the data lineage tools act based on the profiling. This is done by examining the data's characteristics, such as column names, data types, and transformations; data lineage can be inferred.

The Data lineage utilized the various methods, which have been pointed out above, and use a combination of the above methods to establish comprehensive Data lineage. This approach may involve using tools to automate some aspects of lineage tracking while manually documenting more complex or undocumented data flows. The tools for data lineage are dependent upon the organizational objectives. Data lineage is a critical component of effective data governance, ensuring that you have a clear understanding of how data is used, transformed, and moved throughout your organization, which, in turn, supports data quality, compliance, and decision-making processes.

2.7. Data Governance for Compliance

Data governance is very critical in the banking and finance domain in order to be compliant with regulatory requirements which specify rules for handling personal and sensitive information. Data Lineage is the main component in Data governance, which helps track data from source to ultimate target along with various hops between the flow. A robust data framework with good governance is necessary for meeting compliances and improving operational efficiency and data quality. In order to comply with Anti-Money Laundering (AML) regulations and other data privacy and protection acts, data lineage is crucial. It provides an audit trail of how personal information is collected, processed at different stages, and shared within organizations.

2.8. Data Lineage Best Practices

There is always a difference between professional and non-expert practices in any world process. The same applies to the data governance-related mechanism, and this includes the data lineage as well. The professional organizations establishing best practices for data lineage is crucial to ensure that data is well-governed, transparent, and effectively managed. Organizations have various objectives for storing and utilizing data. In data-intensive organizations, establishing recommended professional practices for data lineage is essential to ensure efficient data management, compliance, and decision-making.

- The data governance framework must be developed, considering the data lineage as the core component. There is a need for clearly defined roles, responsibilities, and processes for data governance, including data lineage practices. This will also outline the principles and processes required for data management in the firm. This is quite critical for certain organizations, such as the healthcare sector or the production sector, as the data drives the executives' decisions.

- The data comprises various elements and terms, which are to be known by the data users. The recommended step for the same purpose is to develop a centralized data dictionary or glossary that defines data elements, terms, and their meanings. This helps ensure consistent terminology and understanding across the organization.

- Metadata implementation is also required for modern data lineage techniques because it allows the capturing and maintaining of metadata for all data assets, including data lineage, data source descriptions, transformations, and business rules. Ensure metadata is accurate and up to date.

- Automation in any process ensures the accuracy and timely delivery of the information.

- The documentation is a critical step in the data lineage, for such causes the investment in tools and automation to automatically capture and document data lineage. These tools can streamline the process and reduce the risk of human errors.

- Develop standardized templates and visual representations for data lineage diagrams. Consistency in visualization makes it easier for stakeholders to interpret and compare different data lineages. The understanding of facts provides effectiveness in the decision-making; when the steps taken are not backed by accurate information, the results can be inappropriate. The new methods for the documentation should be defined, which ensures document data quality metrics and KPIs. This allows you to track data quality issues throughout the lineage and prioritize necessary improvements.

The integration for the data lineage should be made with the controls and data security measures. Ensure that sensitive data is appropriately protected throughout the data flow and enforce security policies consistently. For implementing the new policies related to data governance, the need arises for change management and visioning. Keep track of changes and versions to facilitate rollback in case of errors or discrepancies.

2.9. Implementation Challenges and Mitigation Plan

There are certain challenges in the implementation of the data lineages, some of these include:

- The executive buy-in and sponsorship contribute towards the hurdles in implementing the data lineage in the organizations. For such reason, there is a need to secure executive support by demonstrating the value of Data lineage for informed decision-making, compliance, and Data quality improvement.
- In most cases, the data inventory is outdated or incomplete, which needs quick addressing from the data teams in the organization to smoothen the data lineage. The data profiling and inventory exercises catalog data sources, attributes, and metadata.
- Data silos can be the hurdle in the data lineage; that issue can be resolved with the development of strategies for integrating with legacy systems, including API connections or data extraction mechanisms.
- Data quality in the firms is untrustworthy and does not meet the basic requirement for the implementation of data lineage. The data experts should do data quality checks and validations at various points along the data lineage. Establish data quality metrics and continuously monitor and improve data quality.

Scalability is a greater challenge, as this is driven by the excessive amount of data that is generated daily.

In this environment, scalability becomes a concern for data lineage tools and processes—selecting scalable data lineage tools and architectures to manage growing data volumes. Regularly assess and enhance performance.

3. Conclusion

The research has provided various insights into data trends and the role of Data Lineage in effective data governance. The research suggests that data lineage complements the applicability of data governance. The objectives required from data lineage must align with the data governance principles developed by individual organizations. Data lineage supports various key principles and objectives of data governance, ensuring data quality, security, compliance, and trust. When working in compliance with data governance, data lineage can unlock the full potential of data, enable informed decision-making, and maintain a competitive edge in today's data-driven landscape. The research has demonstrated the effectiveness of data lineage for data governance and its multiple associated benefits. For instance, it ensures data quality, provides confirmation of compliance and audit, and facilitates data documentation. Data lineage is also connected to resource optimization, as organizations leverage their data infrastructure to their advantage.

References

[1] Sughosh V. Kaushik, “Lineage Resource Manager,” *Proceedings of the 2022 International Conference on Management of Data*, pp. 2530–2532, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Healthcare Data Governance, AHIMA, New York, 2022. [Online]. Available: https://www.ahima.org/media/pmcb0fr5/healthcare-data-governance-practice-brief-final.pdf

[3] Ana I. Torre-Bastida et al., “Technological Perspective of Data Governance in Data Space Ecosystems,” *Data Spaces*, Springer, pp. 65-87, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Sergi Nadal et al., “Operationalizing and Automating Data Governance,” *Journal of Big Data*, vol. 9, no. 117, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Elena Nabieva et al., "A Highly Divergent SARS-CoV-2 Lineage B.1.1 Sample in a Patient with Long-Term COVID-19," *MedRxiv*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] COVID-19 Variants Identified in the UK – Latest Updates, gov.uk, 2021. [Online]. Available: https://www.gov.uk/government/news/covid-19-variants-identified-in-the-uk-latest-update

[7] H.J. Stuiver, and J.W.H.C. Crompvoets, “Data Lineage,” *GIM International*, vol. 23, no. 9, pp. 1-5, 2009. [Publisher Link]

[8] Yuan Li et al., “Expansion of Invasive Group A Streptococcus M1UK Lineage in Active Bacterial Core Surveillance, United States, 2019-2021,” *Emerging Infectious Diseases*, vol. 29, no. 10, pp. 2116–2120, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Jasmijn A. Baaijens et al., “Lineage Abundance Estimation for SARS-CoV-2 in Wastewater using Transcriptome Quantification Techniques,” *Genome Biology*, vol. 23, no. 236, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Calum Gabbutt et al., "Lineage Tracing in Human Tissues," *The Journal of Pathology*, vol. 257, no. 4, pp. 501-512, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Haoteng Yan et al., “Lineage Landscape: A Comprehensive Database that Records Lineage Commitment across Species,” *Nucleic Acids Research*, vol. 51, no. 1, pp. 1061–1066, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Louisa Hadj Abed et al., "CellDestiny: A RShiny Application for the Visualization and Analysis of Single-Cell Lineage Tracing Data," *Frontiers in Medicine*, vol. 9, 2022. [CrossRef] [Google Scholar] [Publisher Link]