IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138
INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Review of Data Lineage: Challenges, Tools, Techniques and Approaches

K. Venkateswara Rao
Department of Computer Science and Engineering,
CVR College of Engineering, Hyderabad, Telangana State, India

Abstract: Motivation: Data management has undergone a massive transformation in the past two decades, from the early days of business intelligence using historical data to the big data era for having business insights. The data has been collected and stored in systems without thinking about its lineage. The data lineage covers the lifecycle of data, from its origins, through what happens to the data, when it is processed by different systems, and where it moves from and to over time. Context: If organizations do not know where their data comes from or goes, then they have uncontrolled environments that have risk at various levels. Having uncontrolled data environments means it is very difficult to extract value from data, and the organizations have good chances of being outcompeted. Objective: To identify benefits, challenges of data lineage and evaluation of tools and techniques to implement the same. Methods: The research methodology used in this systematic literature review consists of planning, implementing and results investigation. The first stage involves formulation of the review, recognizing its important requirements, and outlining rules that includes a) Research questions, b) Paper extraction, c) and identification of relevant papers for review. The second stage contains extraction of relevant information from the identified papers and analysis. Last stage constitutes presentation of discussions, solutions, conclusion, and future work. Results: Data lineage challenges, features and comparison of tools, and techniques are presented and discussed. Novelty: Ideal approach for data lineage is proposed and elaborated.

Index Terms - Data Lineage, Data Lineage Benefits, Data Lineage Challenges, Data Lineage Tools, Data Lineage Techniques, Data Lineage Approaches.

I. INTRODUCTION

As organizations have been struggling to manage large amounts of data than ever before, and to answer questions such as where from a given dataset is originated?, How it is transformed?, Who transformed the data?, Why it is transformed?, Where is the transformed dataset?. The organizations have to maintain data flows as well as reasons to answer these questions. If organizations do not maintain information about where their data comes from or goes, they have uncontrolled environments. The uncontrolled data environments pose problems in extracting value from data. The data is the new oil for the organizations. The organizations that fail to extract value from data have good chances of being outcompeted and replaced by organizations that has data lineage in place. More number of systems, data transformations, huge data and legacy systems pose challenges for organizations to provide required information on their data transformation, integration and aggregation processes.

Data lineage has to provide insights into business and technical transformation logic that has been applied on the data. It should be able to trace source of data elements, find their journey through enterprise systems as they are transformed, aggregated, processed and analyzed by various applications and reporting tools. It is important to provide that this data journey can be viewed at least at following two levels of granularity.

i. Summary business lineage [1] to show a summary view of how data has been used by business users, how it is transformed and aggregated as it flows from source to destination.

ii. Detailed technical lineage [2] to track data flows and transformations at lowest levels of granularity through table, columns, and queries.

Data lineage needs to be addressed at both these levels of granularity. Data lineage is aimed to uncover the life cycle of data to show the complete data flow[3] from start to finish with all the details such as how the data got transformed, what is changed, and why.

Data lineage is first documented at data set level manually as overall flows between systems known as horizontal data lineage. It has the benefit of offering a big picture about how customer data flows through enterprise systems. It does not go deep enough when users pick up a point in the horizontal data lineage and dig deeper. This problem is addressed by vertical data lineage. The vertical data lineage go through successive layers of detail to get to the ultimate level that is column-to-report or column-to-column. It also provides details of transformations that happen through these layers of data movement. Vertical data lineage is capable of answering questions such as where a particular data value in a report came from, or how a data value is transformed. It is useful to business intelligence analysts when trying to solve a report discrepancy, or and to data analysts when trying to trying to migrate the data to a new platform.

In data warehousing environment [4, 5], the data lineage helps in tracing warehouse data items back to the original source items, and transformations and aggregations applied through extract-transform-load processes.

1.1 Data Lineage Formal Definition

Let T_I be a set of input tables, T_O be a set of output tables, d be a data item (column or row) from the output set, and P be a transformation or a procedure that takes a set of tables as input and produces another set of tables as output. Then, given $P(T_I) = T_O$ and an output item $d \in T_O$, the set $T_I' \subseteq T_I$ of input data items that contributed to the derivation of data item d is the lineage of d, denoted as $T_I' = P'(d, T_I)$.

The Data Lineage can be modeled and visualized as a directed acyclic graph [6, 7, 8] where nodes are dataset with time, and edges are the transformation with details such as why it is transformed, how it is transformed, who transformed. Though the data lineage offers many benefits, it has involved many challenges, tools and techniques. This paper is organized as follows. Section two outlines research methodology used to review state of the art data lineage. Section three elaborates benefits, Section four discusses challenges, Review and comparison of the tools is provided in section five, Techniques for data lineage are described in section six, Approaches for data lineage are presented in section seven and the last section concludes the paper.

II. RESEARCH METHODOLOGY

The purpose of this systematic review paper is to identify state of the art research in the field of data lineage. The process used in this literature review involves planning, implementing, and results investigation. The first stage involves formulation of the review, identifying its requirements, and rules including a) research questions, b) paper extraction, c) and selection of relevant papers for review. Well known digital libraries and web sources are identified and searched with the keywords related to data lineage to extract the relevant works till date. Older papers having less relevance are rejected and also filtered if papers are duplicated from different sources. The results of this first stage are furnished in Table 1.

Table 1. Digital Libraries and Web sources and Number of Papers collected

Name	Digital Library	Web Source	Number of papers considered
IEEE Digital Library	√		12
ACM Library	√		2
Science Direct (Elsevier)	√		1
Springer	√		3
Google Scholar		√	4
DOAJ Open Access		√	1
Research Gate		√	7

The second stage consists of extracting the relevant information from the selected papers mentioned as references in this article and the last stage presents benefits, challenges, tools and techniques for data lineage. Following sections are the result of this systematics review process.

III. DATA LINEAGE BENIFITS

Data lineage prepares organizations to be data-driven in which data analysis, decision making and addressing regulatory obligations relating to data privacy are based on trusted data [9]. This capability of an organization offers clear benefits to data engineers, data architects and technology teams, and business decision makers as discussed below.

3.1. Improved Data Quality [10] and Integrity

Data lineage helps in tracking the data and eliminating errors, duplicate or redundancy at appropriate source. This enhances data accuracy, quality and integrity. It can also identify problems where source data may be accurate, but an error was introduced in the way the data got processed.

3.2. Fast insight into Data and Improved Business Decisions

Data Lineage enable to better understanding of data and fast insight for accurate analysis. It in turn leads to improved operational efficiency, decision making and identification of new business opportunities through application of techniques from data science, artificial intelligence, data mining, machine learning and deep learning.

3.3. Complying with Relevant Regulations

Data Lineage provides backward lineage of data for tracing of results back to data owners and sources, and for quality assurance [11] and access control. It provides evidence, to regulatory bodies, on where from the data came, who is using it, and how it has been changed. It can support the role of data governance in managing the availability, quality and security [12, 13] of data across an organization. It is key to supporting a variety of regulatory obligations. Privacy regulations need to grant consumers the right to have their personal data deleted, which requires the organizations to know exactly where that data resides.

3.4. System Rationalization and Data Migration

Large scale organizations experience data proliferation and its flow to new systems. It needs to rationalize and simplify data architectures [14] to improve operational efficiency and reduce costs. Maintaining an accurate data lineage is critical to support rationalization projects. Data lineage provides mapping out logical data flows, the precise fields, table joins and transformations supporting individual processes which are key to any migration project.

3.5. Impact Analysis and Change Management

Data lineage allows to understand how data element change will impact downstream systems and reports. It can be used to study how data flows in systems and impact of changes to infrastructure of an enterprise systems, business processes and/or data that could affect specific products and reports.

IV. DATA LINEAGE CHALLENGES

Though data lineage offers significant benefits to the enterprises, it involves Operational, Technical and Data Management challenges as discussed below.

4.1 Operational challenges

Offering data lineage begins with winning management for funding to provide a solution that is expensive, requires lot of human resources, and offers only moderate benefits in early stages of implementation. It involves challenges to answer queries like why an organization requires data lineage?, where is it now with respect to the data lineage?, How much does it cost to acquire required skills into the Organization?, How to address internal cultural issues if any?, How much does it cost now and in future?.

4.2 Technology challenges

The technology challenges are related to automation, Regulations requirements and auditors and regulators queries which are to be answered on demand. Advances in technologies such as Big data [15], Cloud [16], Machine Learning [ML][17, 19], Cyber security [18], Artificial Intelligence[AI] [20] and Data Mining (DM) demands a complex data infrastructure. It involves challenges in finding answers to queries like what extent of manual lineage required?, How data lineage be documented?, How it can be scaled well?, What is the long term strategy for automation, Which areas of the business be covered by the lineage?, What could be the granularity level for the lineage?, What are the skills required and how much does it cost?.

4.3 Data management challenges

It is a complex data management task to implement the data lineage because it involve creation of metadata, availability of huge volumes of data in mixed data formats, mountains of spreadsheets, existence of multiple legacy systems, siloed data and data flows. Data lakes, Big data [15], and data repositories cause issues such as how data is tagged, stored, and linked[21] to other data and systems. When an organization grows through mergers and acquisitions, it poses challenges such as how to map all data sources and data flows?, How to integrate technology and data architectures of new sources? How to incorporate regulatory requirements into data analysis? How to manage and track the obligations across multiple jurisdictions? How metadata [1] changes will impact the existing data lineage? All these challenges make the task of maintaining data lineage a moving target. Other data management challenges include what data is valuable? How to manage redundant data? Is data licensed if it is external? What tools are available to handle these challenges? How to use them.

V. TOOLS FOR DATA LINEAGE

While data lineage offers help to track data and different processes involved in the data flow and their dependencies, metadata management is a key for capturing data flow in an enterprise and presenting data lineage. The tools for data lineage are described below and compared in Table 2.

5.1 Collibra tool automates lineage through enterprise-wide integrated platform with embedded governance and privacy. It captures data lineage in technical and business context.

5.2 Colt [22] tool has good features to allow a data flow network using business-specific metadata such as the business, concept, and product. It represents systems and data stores as nodes in a directed graph. The edges of the graph represents flows of data with standardized taxonomy characteristics.

5.3 Microsoft Purview or Azure Purview is a data governance service that is unified to govern and manage software-as-a-service (SaaS) data in multi-cloud and on-premises.

5.4 Alation is a data lineage tool to offer a lot of data intelligence solutions like data governance, data search and discovery, data analytics, data stewardship, and data transformation. It is an Artificial Intelligence driven tool with behavioral analysis capability to generate actionable insights.

5.5 Atlan tool offers solutions in areas such as data governance, data lineage, data quality, data cataloging, data profiling and discovery, data integration and exploration.

5.6 OvalEdge is based on data catalog for data governance, privacy compliance, quick and credible analytics.

5.7 Octopai gathers metadata automatically from databases, and reports and business intelligence tools. It provides multilayered data catalog to data lineage and analytics teams to discover, trace and interpret the data flows.

5.8 Datameer provides a data lineage and analytics platform that enables the data teams to transform and model the data in the cloud or warehouses using user interface or SQL code.

5.9 CloverDX manages lifecycle of a data pipeline from design, development, deployment and testing. It supports design, run, debug, troubleshoot the data transformations and job work flows using visual designer.

5.10 SQLFlow is an automated data lineage tool for analysis across databases, data warehouses, Hadoop and cloud environments by parsing scripts and stored procedure. It offers visualization of overall data flow.

5.11 Tracer takes tabular database as input and discovers data lineage by analyzing the tables using statistical methods, heuristics, and machine learning techniques.

5.12 Smoke [23] is a lineage enabled system. It optimizes data intensive applications using interactive data profiling and visualizations. It captures lineage using queries at interactive speeds with low overhead.

Table 2. Comparison of Data Lineage Tools

Data Lineage Tool	Automated Data Mapping /Discovery	Data Compliance with Regulation/ Governance /Privacy	System Rationalization and Data Migration	Use of Emerging Technologies	Object Lineage Tracing	Impact Analysis and Change Management	Visualization as Graph/Text
Collibra	Yes	Yes		BI	Yes	Yes	Yes
Colt	Yes	Yes		DM	Yes	Yes	Yes
Microsoft Purview	Yes	Yes			Yes		Yes
Alation	Yes	Yes	Yes	AI	Yes		Yes
Atlan	Yes	Yes	Yes		Yes		Yes
OvalEdge	Yes	Yes	Yes	ML	Yes		Yes
Octopai	Yes	Yes	Yes		Yes		Yes
Datameer	Yes		Yes	ML	Yes		Yes
CloverDX	Yes		Yes				Yes
SQLFlow	Yes	Yes	Yes	BI	Yes	Yes	Yes
Tracer	Yes			ML	Yes		Yes
Smoke	Yes	Yes	Yes		Yes		Yes

VI. TECHNIQUES FOR DATA LINEAGE

The techniques to perform data lineage are discussed in this section.

6.1 Pattern-Based Lineage

This technique uses metadata of tables, columns, and business reports to identify patterns which are similar. As an example, if there is a column with similar values in two data sets, then this technique considers that it is the same data in two stages of its data lineage life cycle. Then this technique connect these two columns in the data lineage chart. As pattern matching techniques do data lineage without looking at the code that transformed or generated the data, these techniques can be used for any database technology. But it may miss some connections between data if data processing logic which is not in metadata, generates data.

6.2 Lineage by Data Tagging

This technique assumes that the data transformation programs tag data in some way. It tracks the tag from origin to finish to discover data lineage. This technique is effective if the transformation tool that tags controls every data movement. If data is generated through transformations without using tool, such data do not possess any tag. In this case, lineage by data tagging does not work.

6.3 Self-Contained Lineage

This technique assume that organizations maintain a data environment in such a way that it offers storage, transformation logic, and master data management (MDM) over metadata. This data environments may be a data lake capable of storing the data in every stages of data lineage lifecycle. This provides lineage without any need to use any tool.

6.4 Lineage by Parsing

This technique relies on processing logic used to generate the data. It does reverse engineering of data processing logic to perform tracing of data to its origin. This technique has to understand all programming languages and logic used to process the data. This is the most advance technique for data lineage generation.

VII. APPROACHES FOR DATA LINEAGE

The scope of data lineage implementation is often determined by enterprise data management strategy, regulatory requirements, and critical data elements of an organization. Building a data linkage system requires to keep track of all the processes in the system that transforms the data. Data has to be mapped at every stage of its transformation. It needs to keep track of columns, tables, views, and reports across databases and extract-transform-load (ETL) jobs of data warehouse. There are two possible approaches for data lineage.

7.1 Use of Automation based Tools Approach

Use of automation techniques such as AI, ML, BI, and Data mining while building data lineage tools. Automated data lineage tools are fast, and deal well with complexity and scale well. This is a work around solution for existing databases till all database management systems are enhanced to have built-in data lineage.

7.2 Lineage as Built-in feature of Database Management Systems Approach

Engineering all existing database management systems to have built-in data lineage feature. This involves design of metadata database for lineage, modifying Data Definition Language, Data Manipulation Language, Application Programming interface, data loading tools, and data transformation applications to create and update metadata tables appropriately. Techniques used in temporal databases, spatiotemporal databases and software configuration management systems can be borrowed to add temporal aspects to the metadata of data lineage. This approach is ideal and provides true and accurate data lineage.

VIII. CONCLUSION

Data lineage has been offering improved data quality, integrity, security, availability and data governance. It helps organizations to have regulatory compliance and fast insight into data. It is helping data and system engineers in impact analysis and change management as well as system rationalization and data migration. It has been making the company to have confidence in their data and to work more efficiently and effectively to make more informed business decisions. Operational, Technology and Data management challenges associated with data lineage are identified and discussed well in the paper. Two possible approaches for data lineage are elaborated. As the first approach involves tools, features of various data lineage tools are presented and compared. Need for automation to benefit this approach is emphasized. Various techniques for data lineage are discussed. The second approach is an ideal approach that calls for design of metadata for data lineage and changes in implementation of data definition language, data manipulation language, application programming interface and other data access and data loading tools. As automation becomes more common in all aspects of software engineering, it is only natural that data engineering follows suit from DevOps to DataOps in future.

REFERENCES

[1]. S. Karkošková and O. Novotný. 2021. Design and Application on Business Data Lineage as a part of Metadata Management. International Conference on Computers and Automation (CompAuto). P. 34-39. DOI: https://doi.org/10.1109/CompAuto54408.2021.00014

[2]. Dennis Dosso, Susan B. Davidson and Gianmaria Silvello. 2020. Data Provenance for Attributes: Attribute Lineage. TAPP'20: Proceedings of the 12th USENIX Conference on Theory and Practice of Provenance. DOI: https://dl.acm.org/doi/10.5555/3488890.3488894

[3]. Pushkin E. 2020. Theoretical Model and Practical Considerations for Data Lineage Reconstruction. arXiv preprint arXiv:2001.11506. DOI: https://doi.org/10.48550/arXiv.2001.11506

[4]. M. Jamedžija and Z. Đurić, 2021. Moonlight: A Push-based API for Tracking Data Lineage in Modern ETL processes. 20th International Symposium INFOTEH-JAHORINA (INFOTEH). P. 1-5. DOI: https://doi.org/10.1109/INFOTEH51037.2021.9400667

[5]. Tomingas, K., Järv, P. and Tammet, T. 2016. Discovering Data Lineage from Data Warehouse Procedures. Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. P. 101–110. DOI: https://doi.org/10.5220/0006054301010110

[6]. K. Reddy. 2019. Interactive Graph Data Integration System With Spatial-Oriented Visualization and Feedback-Driven Provenance. IEEE Access. Vol. 7. P. 101336-101344. DOI: https://doi.org/10.1109/ACCESS.2019.2928847

[7]. Nobre, C., Gehlenborg, N., Coon, H., and Lex, A. 2019. Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs. IEEE transactions on visualization and computer graphics. 25(3).P 1543–1558. DOI: https://doi.org/10.1109/TVCG.2018.2811488

[8]. Pokorný, J., Sýkora, J., & Valenta, M. 2019. Data Lineage Temporally Using a Graph Database. Proceedings of the 11th International Conference on Management of Digital EcoSystems. ACM Digital Library. P. 285–291. DOI: https://doi.org/10.1145/3297662.3365794

[9]. A. Bates and W. U. Hassan. 2019. Can Data Provenance Put an End to the Data Breach?. IEEE Security & Privacy. Vol. 17(4). P. 88-93. DOI: https://doi.org/10.1109/MSEC.2019.2913693

[10]. C. Bors, T. Gschwandtner and S. Miksch. 2019. Capturing and Visualizing Provenance From Data Wrangling. IEEE Computer Graphics and Applications. Vol. 39(6). P. 61-75. DOI: https://doi.org/10.1109/MCG.2019.2941856

[11]. Soňa Karkošková. 2022. Data Governance Model To Enhance Data Quality In Financial Institutions. Information Systems Management. DOI: https://doi.org/10.1080/10580530.2022.2042628

[12]. C. Wang et al. 2018. Data Provenance With Retention of Reference Relations. IEEE Access. Vol. 6. P. 77033-77042. DOI: https://doi.org/10.1109/ACCESS.2018.2876879

[13]. Pingcheng Ruan, Gang Chen, Tien Tuan Anh Dinh, Qian Linn, Beng Chin Ooi and Meihui Zhang. 2019. Fine-grained, secure and efficient data provenance on blockchain systems. Proceedings of the VLDB Endowment. Vol 12(9). P. 975–988. DOI: https://doi.org/10.14778/3329772.3329775

[14]. Inês Araújo Machado, Carlos Costa and Maribel Yasmina Santos. 2022. Data Mesh: Concepts and Principles of a Paradigm Shift in Data Architectures. Elsevier Procedia Computer Science. Vol. 196. P. 263-271. DOI: https://doi.org/10.1016/j.procs.2021.12.013

[15]. M. Tang et al. 2019. SAC: A System for Big Data Lineage Tracking. IEEE 35th International Conference on Data Engineering (ICDE). P. 1964-1967. DOI: https://doi.org/10.1109/ICDE.2019.00215

[16]. Sachin Patel, Mrugendrasinh Rahevar and Martin Parmar. 2020. Data Provenance and Data Lineage in the Cloud: A Survey. International Journal of Advanced Science and Technology. 29(05). P. 4883-4900. DOI: http://sersc.org/journals/index.php/IJAST/article/view/13882

[17]. R.M. Thiago, R. Souza, L. Azevedo, E. Figueiredo De Souza Soares, R. Santos, W. Dos Santos, M. De Bayser, M.C. Cardoso, M.F. Moreno and R.F.D.G. Cerqueira. 2020. Managing Data Lineage of O&G Machine Learning Models: The Sweet Spot for Shale Use Case. Proceedings of First EAGE Digitalization Conference and Exhibition. P. 1 – 5. DOI: https://doi.org/10.3997/2214-4609.202032075

[18]. Michael Zipperle, Florian Gottwalt, Elizabeth Chang and Tharam Dillon. 2022. Provenance-based Intrusion Detection Systems: A Survey. ACM Computing Surveys. DOI: https://doi.org/10.1145/3539605

[19]. Hofmann and Felipe Alex. 2020. Tracer: A Machine Learning Approach to Data Lineage. DSpace MIT Libraries. DOI: https://hdl.handle.net/1721.1/127410

[20]. Uchida N, Kaji T, Blake N, Mase M, Ohashi H, Ghosh D, Gupta C, Naono K and Takata M. 2022. Research and Development of AI Trust and Governance. Hitachi Review Special issue. Vol. 71. P. 22-29. DOI: https://www.hitachi.com/rev/archive/2022/r2022-sp/pdf/02.pdf

[21]. M. Wylot, P. Cudré-Mauroux, M. Hauswirth and P. Groth. 2017. Storing, Tracking, and Querying Provenance in Linked Data. IEEE Transactions on Knowledge and Data Engineering. Vol. 29(8). P. 1751-1764. DOI: https://doi.org/10.1109/TKDE.2017.2690299

[22]. Aggour, K.S., Williams, J.W., McHugh, J. and Kumar, V.S. 2017. Colt: concept lineage tool for data flow metadata capture and analysis. Proceedings of the VLDB Endowment. 10(12). P. 1790-1801. DOI: https://doi.org/10.14778/3137765.3137783

[23]. Fotis Psallidas, Eugene Wu. 2018. Demonstration of Smoke: A Deep Breath of Data-Intensive Lineage Applications. SIGMOD '18: Proceedings of the 2018 International Conference on Management of Data. P 1781–1784. DOI: https://doi.org/10.1145/3183713.3193537