www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1054-1058

e-ISSN : 2583-1062
Impact Factor : 7.001

# DATA PROVENANCE AND INTEGRITY IN BIG DATA SYSTEMS: ENSURING TRUSTWORTHY ANALYTICS

**Dr. A. Antony Prakash[1], Shreeparameshwaran.S[2], Praveen A[3], Swikkin Arockiyaraj S[4]**

[1]Assistant Professor, Department of Information Technology, St. Joseph's college, Trichy, Tamilnadu, India.

[2,3,4]PG Student, Department of Information Technology, St. Joseph's college, Trichy, Tamilnadu, India.

DOI: https://www.doi.org/10.58257/IJPREMS43957

## ABSTRACT

As big data systems play a bigger role in decision-making across industries, it is critical to make sure the data they analyze is reliable and trustworthy. Data integrity and transparency in data-driven analytics are greatly aided by data provenance, which is the documentation of the sources, changes, and ownership of data. This study examines the difficulties and approaches involved in maintaining data integrity and provenance in the setting of large data systems. It examines different provenance tracking strategies, such as audit trails, blockchain technology, and metadata management, emphasizing how well they guarantee data accuracy, validity, and ancestry. The study also highlights the significance of reliable data for producing actionable insights by discussing the consequences of data integrity for analytics. Data provenance offers insight into the creation, alteration, and aggregation of data, enabling the detection of irregularities or discrepancies and the verification of data accuracy. It becomes difficult to trace this ancestry in a distributed, complicated big data environment where data comes from several sources and is constantly changing. Effective provenance tracking methods are examined in this study, covering data lineage tools, metadata management, and cutting-edge technologies like blockchain and decentralized ledgers that improve data security and transparency. A wide range of best practices are suggested in this article for businesses to use, such as integrating automated provenance tracking technologies, using secure data validation methods, and creating rules to guarantee accountable and transparent data management.

**Keywords:** Metadata-Based, Blockchain, Data Lineage Tracking, Digital Signatures, Cryptographic.

## 1. INTRODUCTION

Big data's widespread use has transformed industries by allowing governments, corporations, and researchers to extract knowledge from enormous volumes of data. These revelations support data-driven decision-making, direct the creation of policies, and foster the advancement of cutting-edge technologies. However, the problem of guaranteeing the data's validity, integrity, and dependability has grown more crucial than ever before as its volume, variety, and velocity continue to rise. Data that is inaccurate, distorted, or lacking can result in conclusions that are wrong, which can cause expensive errors, injury to one's reputation, and even harm to individuals or society as a whole.

The idea of data provenance is fundamental to guaranteeing the reliability of big data. Tracking and documenting the sources, changes, and movement of data inside a system is known as data provenance. It provides a way to track the lifetime of data, from its inception to its ultimate state, enabling users to confirm its legitimacy, spot possible error sources, and guarantee that the data utilized in analytics is correct and comprehensive. The entire analytics process can be undermined by data manipulation and integrity breach in the absence of adequate data provenance.

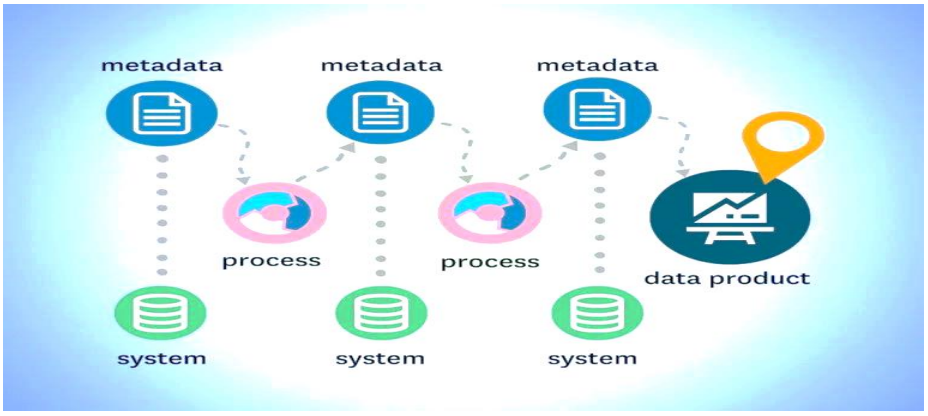


**Fig 1** Data Provenance Vs. Data Lineage

Data integrity, which runs parallel to provenance, is the maintenance of data's accuracy and consistency across the course of its lifecycle. Maintaining data integrity in big data systems is a difficult undertaking. Maintaining data

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1054-1058
e-ISSN : 2583-1062
Impact Factor : 7.001

integrity becomes difficult when data is frequently handled by different entities, analyzed by different algorithms, and dispersed across several systems. Significant hazards, like data corruption or tampering, might arise from a lack of data integrity, producing biased, deceptive, or untrustworthy analytics.

In big data systems, the reliability of analytics depends on the link between data provenance and data integrity. Verification of data's proper sourcing, transformation, and preservation, as well as its accuracy in reflecting the real-world phenomena it depicts, are essential components of trustworthy analytics. Poor decision-making and a decline in stakeholder confidence can result from data provenance errors or data integrity breaches, which seriously impair the validity of the insights obtained from big data.

This study investigates how data integrity and provenance contribute to reliable analytics in big data systems. It looks at the difficulties in determining and preserving the provenance of data as well as the strategies and resources available to guarantee data integrity across time. The study also emphasizes how crucial it is to implement strong data governance frameworks in order to set ethical, transparent, and accountable guidelines for big data management. The study illustrates how businesses may protect the reliability of their analytics by putting in place efficient data provenance and integrity procedures using case studies and actual situations.

## 2. LITERATURE REVIEW

Shuang Wu, Fereidoon Sadri, and Sherali Zeadally [1] The study assesses the efficacy of current methods in various scenarios by classifying them into metadata-based strategies, data lineage, and decentralized technologies like blockchain. Additionally, it emphasizes how data provenance may be used in data-driven systems to guarantee data integrity, ease compliance, and enable repeatability. The survey also covers provenance information retrieval and storage, with an emphasis on distributed storage solutions and database architectures. In order to support the increasing complexity of big data systems, the study concludes by outlining future research areas and highlighting the necessity of scalable, effective, and privacy-preserving provenance solutions.

M. K. O'Neill [2], Data provenance, according to the article, is the recording of the sources, movements, and changes of data as it passes through different systems and procedures. Particularly in systems where data is changed, aggregated, or passed through several components (as in scientific computing, corporate intelligence, and big data analytics), provenance is crucial for data verification, auditability, and reliability. The authors examine various methods for gathering and preserving provenance data, with an emphasis on data lifecycle tracking and related metadata.

A. M. T. Balazinska [3], Examine the difficulties in maintaining data integrity in various big data architectures, such as cloud-based, distributed, and real-time systems. Important provenance methods are divided into groups according to how they are applied in different stages of the data processing pipeline, such as auditing, provenance tracking, and metadata management. The trade-offs between performance, scalability, and provenance data granularity are also covered in the study. The study ends with a discussion of outstanding issues and future research paths in the field of big data data provenance, highlighting the need for effective, safe, and private solutions that meet the growing demands of contemporary data systems.

S. L. Yang, X. Zhang, H. J. Cao [4], It covers a range of provenance tracking strategies, such as data lineage methods, metadata-based approaches, and the application of decentralized technologies like blockchain. The use of data provenance in fields including real-time analytics, compliance, reproducibility in scientific research, and data quality assurance is also examined in the article. The authors also point out a number of urgent issues, including as privacy, scalability, and integrating provenance into heterogeneous systems. The study concludes by outlining possible avenues for future research, such as the creation of effective provenance models, privacy-preserving strategies, and systems for smooth integration across various data platforms and applications.

S. K. Rai, N. N. A. G. Mahmud [5], provides a thorough analysis of the methods and systems intended to guarantee the integrity and provenance of data in distributed cloud storage settings. It examines several provenance models, such as those based on metadata and cryptography, and talks about how well they work with various cloud architectures, including decentralized, hybrid, and multi-cloud systems. The difficulties in preserving provenance and integrity in cloud storage are also highlighted in the article, including issues with scalability, performance overhead, privacy, and the difficulty of managing dispersed data. The study concludes by outlining unresolved research difficulties and potential avenues for future development, such as the requirement for effective and safe provenance tracking systems that can be easily integrated with current cloud storage systems while taking privacy and scalability concerns into account.

X. Zhang, W. Li, Q. Zhang [6], author examine several methods for building trust, including integrity verification, provenance tracking, and safe data storage. Talk about how these methods might be included into big data systems.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1054-1058
e-ISSN : 2583-1062
Impact Factor : 7.001

The study also highlights important obstacles to preserving confidence in big data analytics, such as problems with data security, privacy, and bias detection in analytical models. The study ends by outlining potential avenues for future research to enhance big data analytics' openness, dependability, and ethical considerations. It also stresses the significance of strong frameworks in ensuring the credibility of data-driven judgments.

## 3. TECHNIQUES FOR ENSURING DATA PROVENANCE

Several techniques are employed to capture and track data provenance:

### 3.1 Metadata-Based Approaches

Attaching metadata to data at different points in its lifecycle, such as when it is gathered, modified, or stored, is known as metadata-based provenance. The source, the transformation processes, the storage location, and the users engaged are all included in this metadata.

### 3.2 Data Lineage Tracking

A specialized method for monitoring the movement of data across different phases in a processing pipeline is called data lineage. This is especially helpful in intricate data operations that involve numerous aggregations or transformations. Users can learn how raw data has been processed and whether the final dataset is reliable by following the lineage.

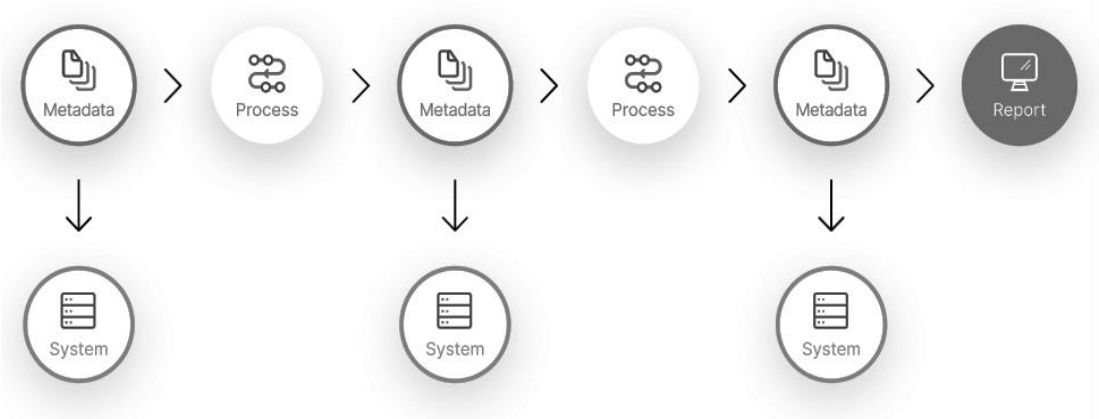


**Fig 2** Data Lineage Tracking

### 3.3 Blockchain and Distributed Ledger Technology

For unchangeable provenance monitoring, blockchain and other decentralized ledger systems present a viable option. By producing a visible and impenetrable record of data, these technologies guarantee that information cannot be changed without being discovered.

### 3.4 Provenance in Databases

In database systems, provenance is often captured using system-level logs or database triggers that record every change made to the data. This is useful for tracking updates, deletions, and additions within a database.

## 4. ENSURING DATA INTEGRITY

To preserve data integrity, especially in distributed and cloud-based systems, several strategies and techniques are employed:

### 4.1 Cryptographic Hashing

By giving each piece of data a distinct fingerprint, cryptographic algorithms like hash functions guarantee data integrity. To identify any tampering, the hash value of the data is computed and compared to the original hash value whenever it is transferred or stored.

### 4.2 Digital Signatures and Certificates

By enabling the owner or creator of the data to sign it using a private key, digital signatures aid in confirming the data's legitimacy. Then, using the matching public key, recipients may confirm the signature and make sure the data hasn't been changed.

### 4.3 Secure Storage Systems

Even in the case of hardware malfunctions or network outages, data is safely saved and recoverable without corruption thanks to distributed storage systems like those based on replication or erasure coding.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1054-1058
e-ISSN : 2583-1062
Impact Factor : 7.001

## 5. CHALLENGES IN ENSURING DATA PROVENANCE AND INTEGRITY IN BIG DATA SYSTEMS

Ensuring **data provenance** and **data integrity** in big data systems is crucial for maintaining trust, transparency, and accuracy in data-driven decision-making. However, due to the large scale, complexity, and distributed nature of modern big data systems, there are several challenges that need to be addressed

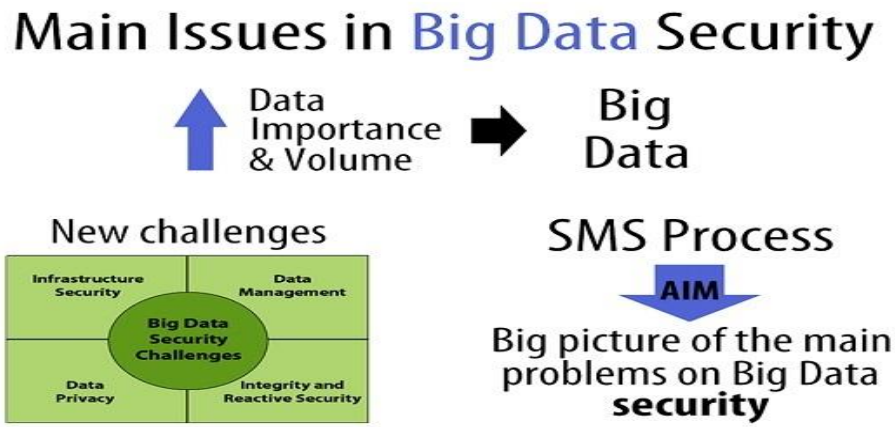


**Fig 3** Data Provenance and Integrity

### 5.1 Complexity and Scale

Effective provenance tracking and management is challenging due to the massive volume of data in big data systems. Furthermore, ensuring data integrity is made more difficult by the complexity of distributed systems, where data is stored across numerous nodes or cloud environments.

### 5.2 Overhead in Performance

Performance overhead is frequently introduced by methods like integrity verification and provenance tracking, particularly when working with big datasets. A major challenge is finding a compromise between insuring integrity, tracing provenance, and sustaining high performance.

### 5.3 Privacy Issues

Privacy issues can occasionally arise from tracking the source of data and guaranteeing its integrity, particularly in sensitive data environments like healthcare or finance. To guarantee adherence to data protection laws (such as GDPR and HIPAA), privacy-preserving techniques must be used.

### 5.4 Integration with Existing Systems

Integrating provenance and integrity mechanisms into existing data systems, especially legacy systems, can be difficult and resource-intensive. Additionally, ensuring that these mechanisms are compatible across different cloud providers, databases, and storage solutions adds complexity.

## 6. FUTURE DIRECTIONS

The future of data provenance and integrity in big data systems lies in developing **scalable**, **efficient**, and **secure** solutions that can seamlessly integrate with diverse data platforms. Research directions include:

**Automated Provenance Capture**: Developing methods to automate the capture of provenance in dynamic environments without excessive overhead.

**Privacy-Preserving Techniques**: Ensuring that provenance and integrity mechanisms do not compromise user privacy, especially in sensitive domains.

**Blockchain-Based Solutions**: Exploring more widespread use of blockchain and distributed ledger technologies for tamper-proof provenance tracking.

**Cross-Domain Provenance**: Addressing the challenge of integrating provenance information across heterogeneous systems (e.g., different databases, cloud platforms).

## 7. CONCLUSION

Big data analytics' credibility is based on the provenance and integrity of its data. While integrity guarantees that data stays secure and true, provenance offers accountability and transparency. Organizations may guarantee the validity and dependability of their data-driven decisions by implementing efficient systems for monitoring data lineage,

confirming data transformations, and maintaining integrity through cryptographic mechanisms. But there are still a lot of obstacles to overcome, especially in the areas of integration, privacy, and scalability. Enabling the next generation of reliable big data analytics solutions will depend on how well these issues are resolved.

## 8. REFERENCE

[1] Shuang Wu, Fereidoon Sadri, and Sherali Zeadally, "A Survey of Data Provenance Techniques in Big Data Systems", ACM Computing Surveys, 2021

[2] Kennesaw State University: M. K. O'Neill, C. M. M. K. R. M. T. J.," Data Provenance: A Survey of Techniques and Applications", Journal of Computer Science and Technology, 2015

[3] A.M. T. Balazinska, K. S. Lakshmanan, L. Zhang, and S. Thirumuruganathan, "Ensuring Data Integrity in Big Data: A Survey of Provenance Models and Approaches", IEEE Transactions, 2016

[4] S. L. Yang, X. Zhang, H. J. Cao, and W. G. Yang, "A Survey on Data Provenance: Challenges and Future Directions", Data & Knowledge Engineering Journal, 2018

[5] S. K. Rai, N. N. A. G. Mahmud, S. N. Kumar, and A. S. Singh, "Provenance and Data Integrity in Distributed Cloud Storage: A Survey of Mechanisms and Challenges", International Journal of Cloud Computing and Services Science, 2019.

[6] X. Zhang, W. Li, Q. Zhang, and J. Liu, "Trustworthy Big Data Analytics: Approaches and Challenges", Future Generation Computer Systems, 2018.

[7] Anil Kumar, Rajesh Kumar, and Manish Kumar, "Big Data Analytics: Tools and Techniques for a Competitive Edge", Journal of Big Data, 2020

[8] Lei Zhang, Fei Deng, Guanglu Zhao, and Ming Yu, "Applications of Big Data Analytics in Healthcare Systems: A Survey", IEEE Access, 2021

[9] M. S. Islam, R. S. Rahman, and M. A. Hossain, "Big Data Analytics for Healthcare: A Survey of Techniques, Applications, and Research Trends", Health Information Science and Systems, 2018

[10] S. S. Bakhshi, S. T. A. Shah, and R. K. Gupta, "Big Data Analytics for Business Applications: Opportunities and Challenges", Journal of Big Data, 2021.

[11] M. A. Hossain, M. M. Rahman, and H. R. U. Karim, "A Survey on Big Data Analytics: Challenges, Techniques, and Applications", International Journal of Computer Science and Information Security, 2018

[12] R. K. Gupta, A. M. Tiwari, and N. K. Gupta, "Scalable Big Data Analytics Using Apache Spark and Hadoop Ecosystem", Journal of Computing and Data Science, 2020.

[13] A.K. Sharma, M. A. K. Parveen, and K. L. Yadav, "Big Data Analytics for Predictive Modeling: A Case Study on Customer Behavior" , Expert Systems with Applications, 2021.

[14] A. M. T. Balazinska, K. S. Lakshmanan, L. Zhang, and S. Thirumuruganathan, "Ensuring Data Integrity in Big Data: A Survey of Provenance Models and Approaches," IEEE Transactions, 2016.

[15] S. Wu, F. Sadri, and S. Zeadally, "A Survey of Data Provenance Techniques in Big Data Systems," ACM Computing Surveys, 2021.