Transforming Data Governance with Metadata: A Key to Big Data Management

Mr. Ashish N. Patil[1]
Research Scholar, Bharati Vidyapeeth (Deemed to be University) College Of Engineering Pune, India
Assistant Professor, SSPM's JCEP KM Gad, India

Dr. Prakash R. Devale[2]
Professor, Bharati Vidyapeeth (Deemed to be University) College Of Engineering Pune, India

Abstract:

In the era of big data, the volume, variety, and velocity of data generation has significant challenges for effective governance. Traditional approaches often are dependent on the direct governance of original datasets, which can lead to inefficiencies, data privacy concerns, and scalability issues. This paper presents the potential of metadata in addressing these challenges, positioning it as crucial asset for enhancing data governance strategies. By integrating metadata descriptive, structural and administrative data organizations can streamline processes such as data access, security, compliance, and lifecycle management without directly interacting with the raw data. This metadata-driven approach ensures better scalability, preserves data privacy, and enhances overall data management efficiency. The study investigates use case related to credit card fraud detection using various machine and deep learning algorithms, highlighting how metadata governance frameworks can offer a more sustainable solution for managing the complexities of big data.

Keywords: Metadata, Schema Inference, Data Governance, Big Data

1. Introduction:

Due to evolution of big data many industries have unlocked opportunities for insight, innovation, and decision-making. However, with this ever growing of data there is significant challenge associated with data governance and management of vast, diverse, and complex datasets effectively. Traditional data governance approaches are often dependent on the direct management of the original datasets. While these methods offer control, they are increasingly issues regarding the scale and complexity of big data environments, raising concerns around data privacy, and operational efficiency.

Metadata often termed as the data about data emerges as a powerful tool for transforming governance practices. Unlike the direct governance of raw data, metadata provides a layer of abstraction that can describe, structure, and manage data without manipulating the original content. By focusing on metadata, organizations can more efficiently control access, ensure compliance, safeguard privacy, and track data lineage, all while reducing the risks and complexities associated with handling the original data (Marijn, et. al 2020)

Data lakes now days have become an essential component of big data architectures, offering a centralized repository to store vast amounts of structured, semi-structured, and unstructured data. While data lakes provide flexibility and accessibility, their ungoverned nature can lead to disorganization often referred to as the "data swamp" problem. Metadata plays an important role in overcoming these issues, by enabling data lakes to be more effectively managed and organized. With metadata-driven governance, organizations can impose structure on the data within lakes, making it easier to discover, classify, and ensure the quality of data assets (Hassan Mehmood, et.al 2019). This enables efficient retrieval, better security practices, and compliance with regulatory requirements, all while maintaining the scalability and flexibility of the data lake architecture. (Y. Chen, et.al 2018)

The objective of this paper provides critical role of metadata in modernizing data governance frameworks within the big data ecosystem. It argues that metadata-driven governance not only enhances operational efficiency but also provides a scalable, secure, and flexible approach to managing data assets. Through various industry use cases and a review of emerging technologies, such as artificial intelligence, this paper highlights how organizations can utilize potential of metadata to navigate the complexities of big data governance. Moreover, it addresses how this shift to metadata governance can offer a sustainable path forward as data volumes continue to grow exponentially

2. Background and Related Work

Effective data governance is important for ensuring trustworthy artificial intelligence (AI) systems. (Janssen et al. 2020) provide a comprehensive framework for data governance that can be utilized for AI applications, having emphasis on principles such as data quality evaluation, minimizing data access permissions, and the role of data stewards in managing data processes and algorithms. Their work provides significance regarding the importance of stewardship in maintaining the integrity and reliability of data used in AI, aiming to foster trust and accountability in AI systems.

Cuzzocrea et al.2021 focuses on models, frameworks, and techniques to improve the efficiency of data lake processing through data semantics. This paper highlights the challenges and methodologies associated with managing big data lakes, including modern metadata management, machine learning tools, and privacy-preserving techniques. Cuzzocrea's work is pivotal for understanding the difficulties of handling vast amounts of data and the role of metadata in optimizing data lake operations.

The concept of metadata management is further explored by (Alserafi et al. 2016), who focus on content metadata in data lakes. Their approach involves schema annotation and systematic extraction and management of metadata, utilizing schema matching and ontology alignment techniques. This systematic process helps in organizing data content, enhancing the ability to manage metadata effectively within data lakes. In the context of metadata quality, a study addresses various aspects of metadata management, including quality attributes, metrics, and policies. This research provides insights into metadata quality management and suggests directions for future improvements. Waterworth et al. (2021) contribute to the field by examining automated metadata extraction for smart buildings using neural language processing methods. Their work demonstrates the application of transfer learning to tag building sensors with semantic tags, achieving a neural language model accuracy of 71% on a real-world dataset. This research highlights the feasibility and potential of using advanced NLP techniques for metadata extraction and management in smart environments.

Jahnke and Otto (2023) discuss the role of data catalogues in enterprise metadata management. Their study emphasizes the need for standardized practices and federated metadata management to support the decentralization of data management. Their typology provides a framework for focusing on appropriate data catalogue applications within enterprise settings. Overall, these studies collectively offer a comprehensive view of current methodologies and challenges in data governance and metadata management. They provide valuable insights into improving data quality, efficiency, and management practices across various domains and applications.

Table 1: Comparative study of data and metadata and its importance for data governance

Parameter	Data	Metadata	Importance of Metadata for Data Governance
Data Type	Integer, float, string, binary, etc.	Descriptive, structural, and administrative metadata.	Metadata provides detailed descriptions, ensuring accurate data handling and processing.
Granularity	Level of detail in the data (e.g., transaction-level, aggregated data).	Context at various levels (file, record, field level).	Fine-grained metadata ensures precise data control and governance.
Schema Management	Data adheres to a schema that defines its structure.	Describes and manages the schema, including attributes, constraints, and relationships.	Metadata-driven schema management supports dynamic and scalable data governance.
Indexing	Data indexing improves retrieval speed and efficiency.	Metadata itself can serve as an index, enhancing data searchability and access.	Proper metadata indexing optimizes data governance by improving data discoverability.
Version Control	Tracks different versions of data over time.	Provides version history, changes, and timestamps for data.	Metadata-driven version control ensures accurate tracking of data changes, vital for auditing and compliance.
Data Provenance	Involves the history of data from origin to its current state.	Captures the source, transformation steps, and derivation processes.	Metadata enables comprehensive data provenance, ensuring traceability and accountability.
Data Integration	Data from different sources, potentially with different formats.	Data mappings, transformations, and integration processes.	Metadata-driven integration ensures consistency and accuracy during data merging and transformation.
Scalability	Ability to manage increasing data volumes.	Managing the metadata as data volumes grow.	Efficient metadata management ensures scalable and sustainable data governance frameworks.
Performance Monitoring	Data operations, queries, and transactions.	Usage statistics and bottlenecks.	Metadata enables proactive monitoring and optimization of data systems for improved governance.

3. Proposed System Architecture:

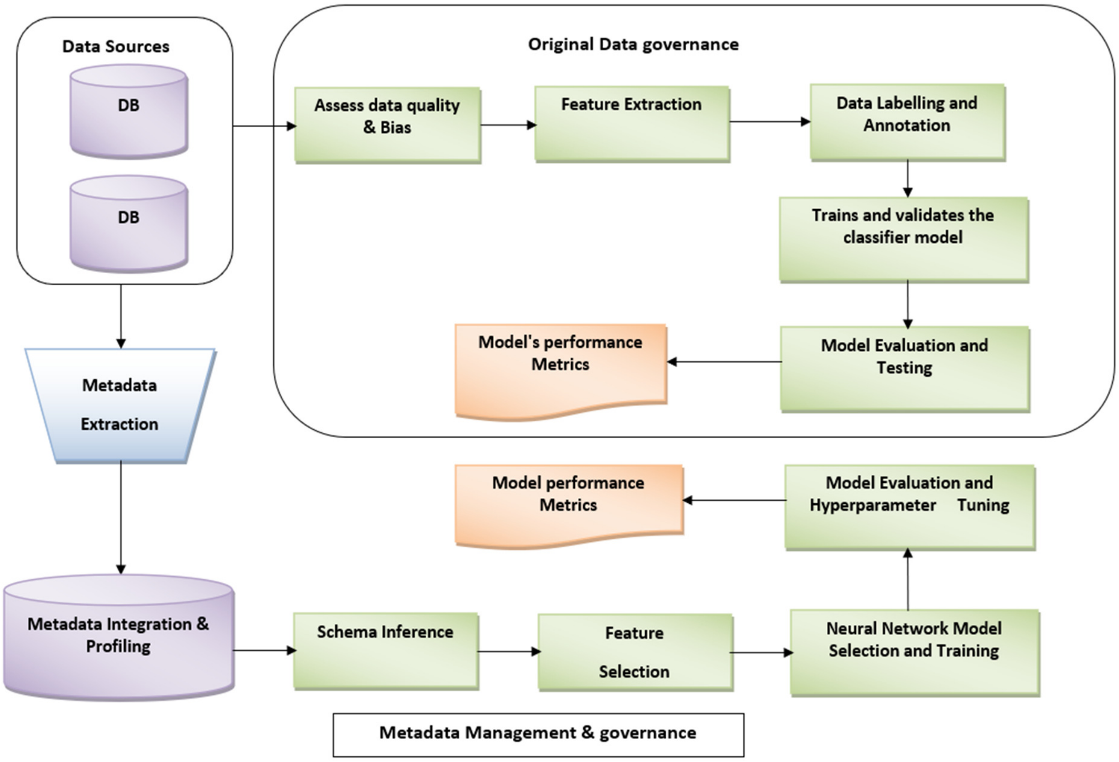

Data Sources
DB
DB
Original Data governance
Assess data quality & Bias
Feature Extraction
Data Labelling and Annotation
Trains and validates the classifier model
Model Evaluation and Testing
Model's performance Metrics
Metadata Extraction
Model performance Metrics
Model Evaluation and Hyperparameter Tuning
Metadata Integration & Profiling
Schema Inference
Feature Selection
Neural Network Model Selection and Training
Metadata Management & governance

Fig 1. Meta data Management and governance

3.1 Original Data Governance:

Original data governance has focus on managing and ensuring the quality and integrity of the data. This process commences with a thorough assessment of data quality and bias, ensuring that the data used for model training is both reliable and fair. Through evaluation of completeness, accuracy, and potential biases within dataset, organizations can minimize the risk of skewed outcomes and enhance the model's performance in case of real-world scenarios.

Further, this process involves extracting relevant features from the dataset, which are important for training machine learning algorithms. Data labelling and annotation are then performed with the aim to structure the dataset for allowing effective supervised learning. This governance of data ensures that when models are trained, on high-quality, well-annotated data, leading to robust and reliable AI systems. The governance process is completed with the evaluation and testing of the model, where various performance metrics are generated to assess the model's effectiveness and identify areas for improvement, ensuring that the AI system meets the desired performance standards before deployment.

3.2 Metadata Governance:

Metadata governance is centred on the extraction, integration, and management of metadata, which provides contextual information about the data. This process starts with metadata extraction from data sources, where key information such as schema, data profiles, and relationships within the dataset are aggregated. Metadata is then integrated and profiled to infer the schema and identify any inconsistencies, ensuring that the metadata aligns with the actual structure and content of the dataset. This alignment is important for maintaining consistency and accuracy in data management, especially in complex environments like big data lake.

The inferred schema and metadata help in selection of features that will be used in the training of neural network models, ensuring that the models are optimized based on a comprehensive understanding of the data. Finally, the process includes the evaluation of these models and the fine-tuning of hyperparameters, with the performance metrics providing valuable feedback for continuous improvement. Through effective metadata governance, organizations can enhance the data privacy, transparency, discoverability, and overall quality of their AI systems, making them more reliable and easier to manage.

4. Implementation Details:

4.1 Dataset Details:

In order to provide significant importance of metadata management over original data, dataset from Kaggle was utilized where training was performed into two aspects on original data and metadata. Various Machine learning and deep learning algorithms are trained, tested and validated

Table 2. Dataset Description used for Implementation

Description	Simulated credit card transaction dataset containing legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2022
Coverage	Credit cards of 1000 customer conducting transactions with a pool of 800 merchant
No of features	22 features
Size, Class Distribution	1.8 million Transactions, Imbalanced dataset with 9.6k fraud transactions.

4.2 Fraud classification model analysis

Five classification models on original dataset are trained while artificial neural network is trained on extracted metadata. Extracted metadata features are transaction date, time, location of merchant and cc_number which are selected as input for training neural network

4.2.1 Dataset splitting:

Under sampling (NearMiss) technique is used to handle imbalanced nature of dataset. NearMiss selectively under samples the majority class instances to ensure proximity to minority class instances.

Dataset split ratio in %: 65:35

Hyperparameters setup for neural network:

- Epochs: 100
- 4 dense layers
- batch size: 128
- Loss Function: Binary Cross Entropy
- Optimizer and learning rate: Adam, 0.001

Table: 3 Details of Models trained on original data and meta data

Models on Original Data	Model on Extracted Metadata
Logistic Regression, Decision Tree, Random Forest Classifier, XGBoost, K-Nearest Neighbor	Artificial neural network Sequential model with 4 dense layers with activation function as Relu for input and hidden layer and sigmoid for output layer.

4.3 Results and Discussion:

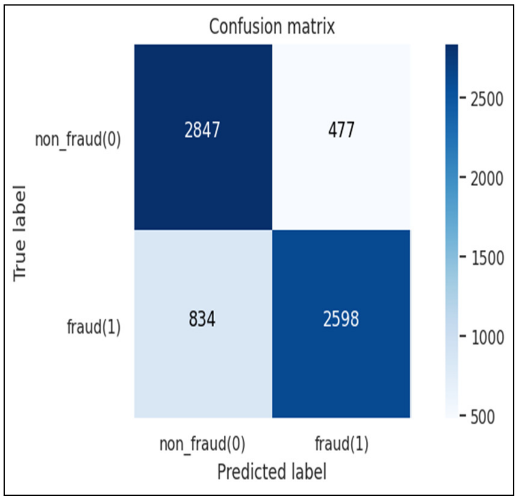


Fig 2. Confusion Matrix of Logistic Regression

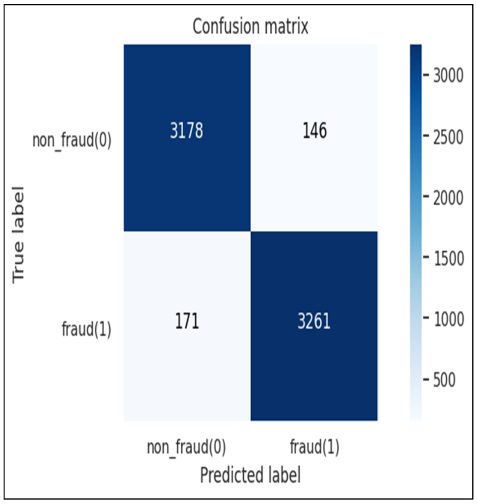


Fig 3. Confusion Matrix of Decision Tree Classifier

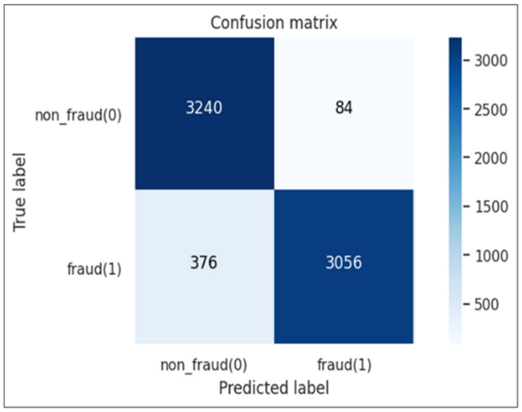


Fig 4. Confusion Matrix of Random Forest Classifier

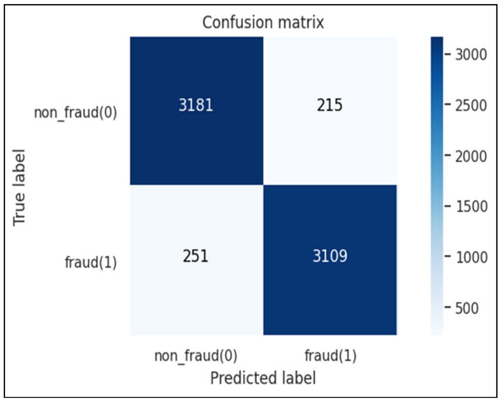


Fig 5. Confusion Matrix of XGBoost Classifier

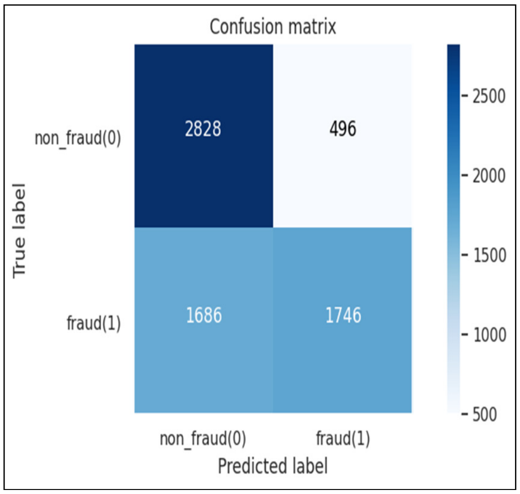


Fig 7. Confusion Matrix of KNN Classifier

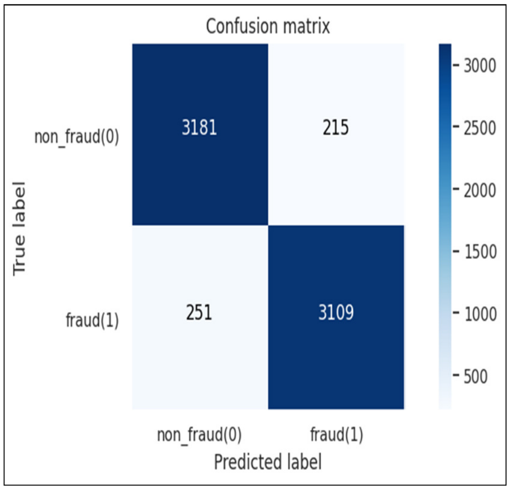
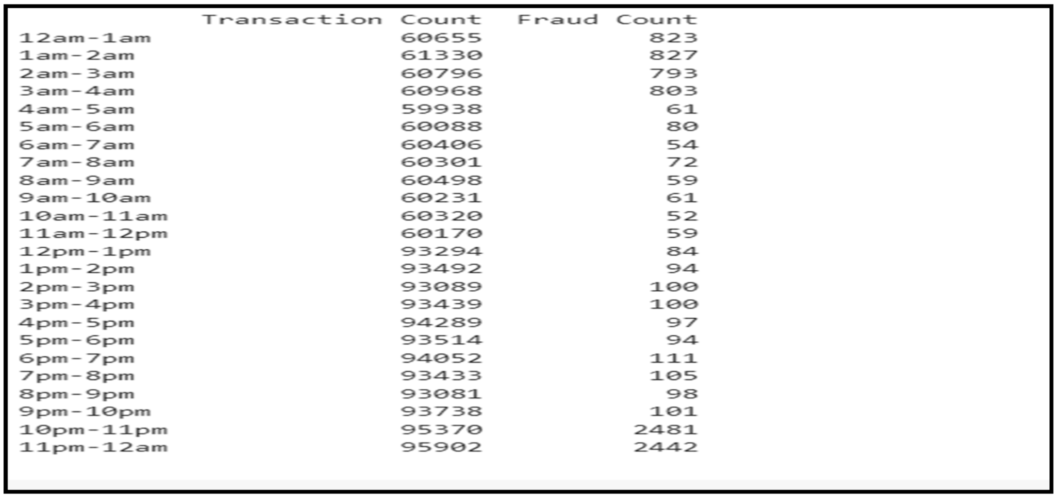


Fig 8. Confusion Matrix ANN on Metadata

4.3.1 Use case of Metadata Extraction:

1. Time distribution of transactions

This use case involves extracting and organizing the temporal metadata to visualize and understand how transactions are distributed across different time intervals (e.g., by hours, days, or months). Such analysis can reveal valuable insights, such as identifying peak transaction times, periods of low activity, and potential anomalies in transaction patterns. For example, a bank might observe an unusual spike in transactions during late hours, which could indicate fraudulent activities or system inefficiencies. By leveraging the time distribution metadata, businesses can optimize their operations, improve system performance during high-traffic periods, and detect suspicious transaction behaviours in real-time, leading to enhanced security and service reliability.



```
             Transaction Count   Fraud Count
12am-1am                 60655           823
1am-2am                  61330           827
2am-3am                  60796           793
3am-4am                  60968           803
4am-5am                  59938            61
5am-6am                  60088            80
6am-7am                  60406            54
7am-8am                  60301            72
8am-9am                  60498            59
9am-10am                 60231            61
10am-11am                60320            52
11am-12pm                60170            59
12pm-1pm                 93294            84
1pm-2pm                  93492            94
2pm-3pm                  93089           100
3pm-4pm                  93439           100
4pm-5pm                  94289            97
5pm-6pm                  93514            94
6pm-7pm                  94052           111
7pm-8pm                  93433           105
8pm-9pm                  93081            98
9pm-10pm                 93738           101
10pm-11pm                95370          2481
11pm-12am                95902          2442
```

Fig 9. Transaction count and Fraud count

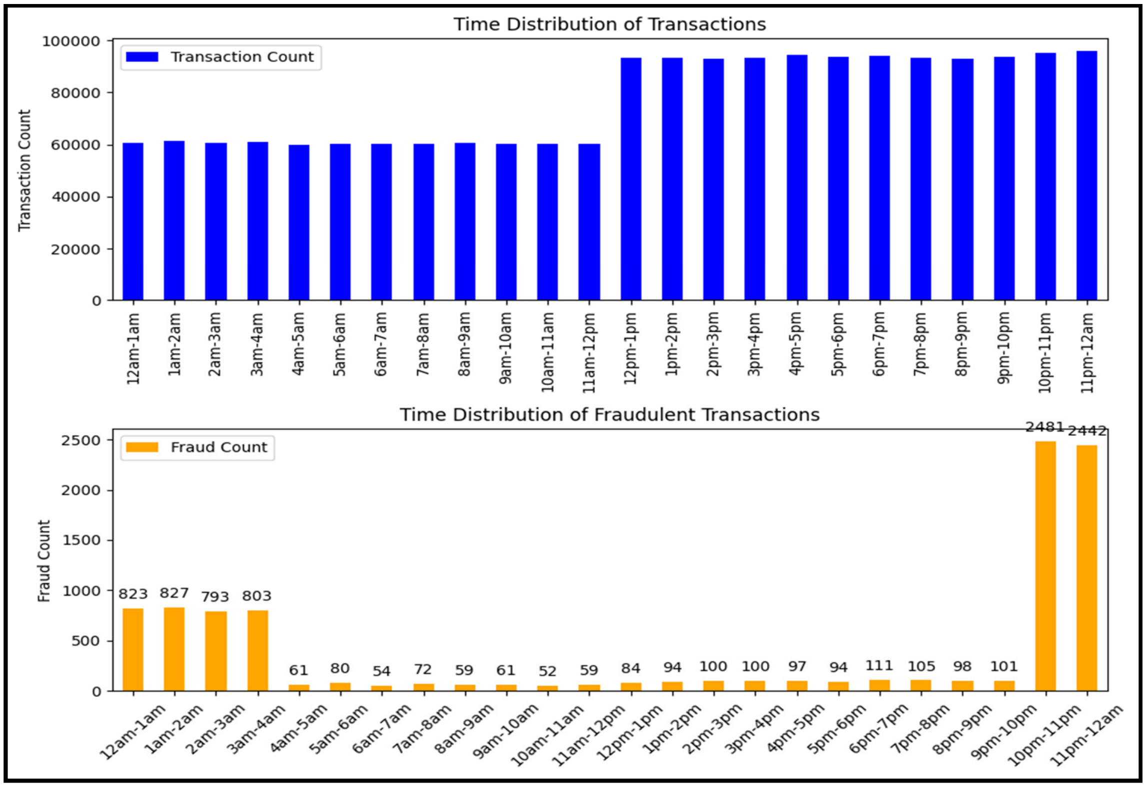
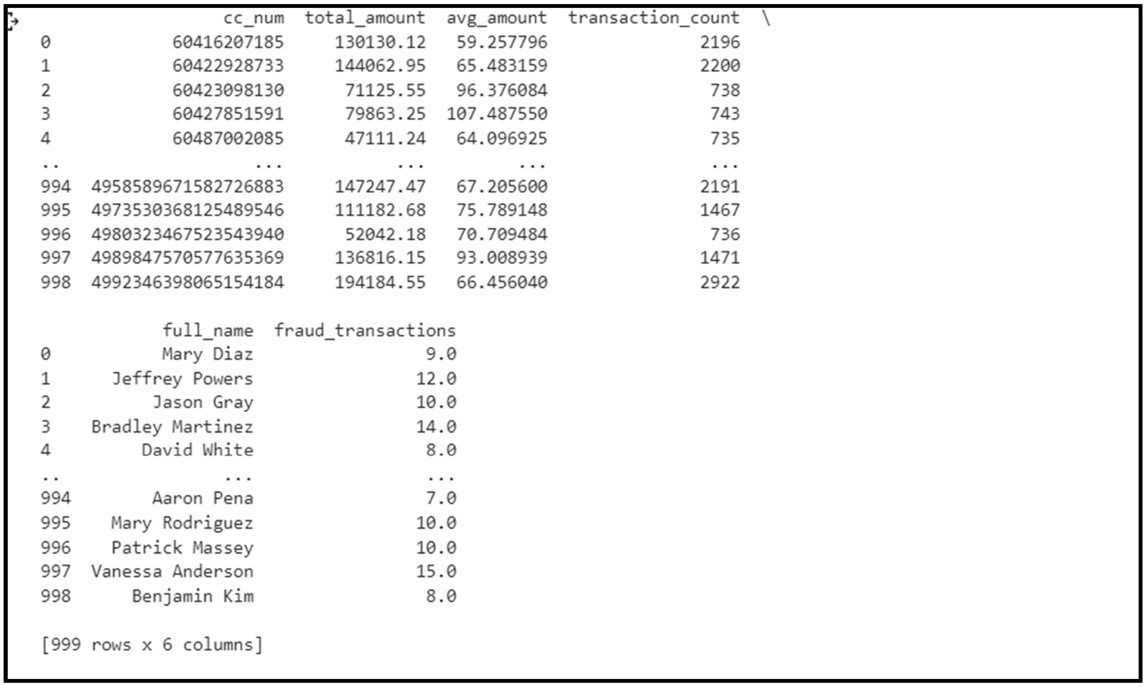


Fig 10. Distribution of transactions and fraudulent transactions with respect to time

2. Analysis of Credit card number

This use case involves extracting metadata related to credit card numbers, such as their structure, issuer details (e.g., Visa, Mastercard), and geographic distribution based on the card's issuer identification number (IIN). Such analysis can help detect inconsistencies in the credit card data, such as invalid card formats or unexpected issuers for specific regions. Moreover, analyzing this metadata can be beneficial in detecting fraud, as patterns in card usage (e.g., unusual transactions across different countries in a short time) may suggest card cloning or misuse. Through metadata-driven credit card analysis, businesses can enhance their fraud detection mechanisms and ensure compliance with payment security standards like PCI-DSS, thereby protecting both consumers and the organization from potential breaches and financial losses.



```
                  cc_num  total_amount  avg_amount  transaction_count  \
0            60416207185     130130.12   59.257796               2196   
1            60422928733     144062.95   65.483159               2200   
2            60423098130      71125.55   96.376084                738   
3            60427851591      79863.25  107.487550                743   
4            60487002085      47111.24   64.096925                735   
..                   ...           ...         ...                ...   
994  4958589671582726883     147247.47   67.205600               2191   
995  4973530368125489546     111182.68   75.789148               1467   
996  4980323467523543940      52042.18   70.709484                736   
997  4989847570577635369     136816.15   93.008939               1471   
998  4992346398065154184     194184.55   66.456040               2922   

             full_name  fraud_transactions  
0            Mary Diaz                 9.0  
1       Jeffrey Powers                12.0  
2           Jason Gray                10.0  
3     Bradley Martinez                14.0  
4          David White                 8.0  
..                 ...                 ...  
994         Aaron Pena                 7.0  
995    Mary Rodriguez                 10.0  
996     Patrick Massey                10.0  
997  Vanessa Anderson                 15.0  
998        Benjamin Kim                8.0  

[999 rows x 6 columns]
```

Fig 11. Analysis of credit card based upon extracted metadata

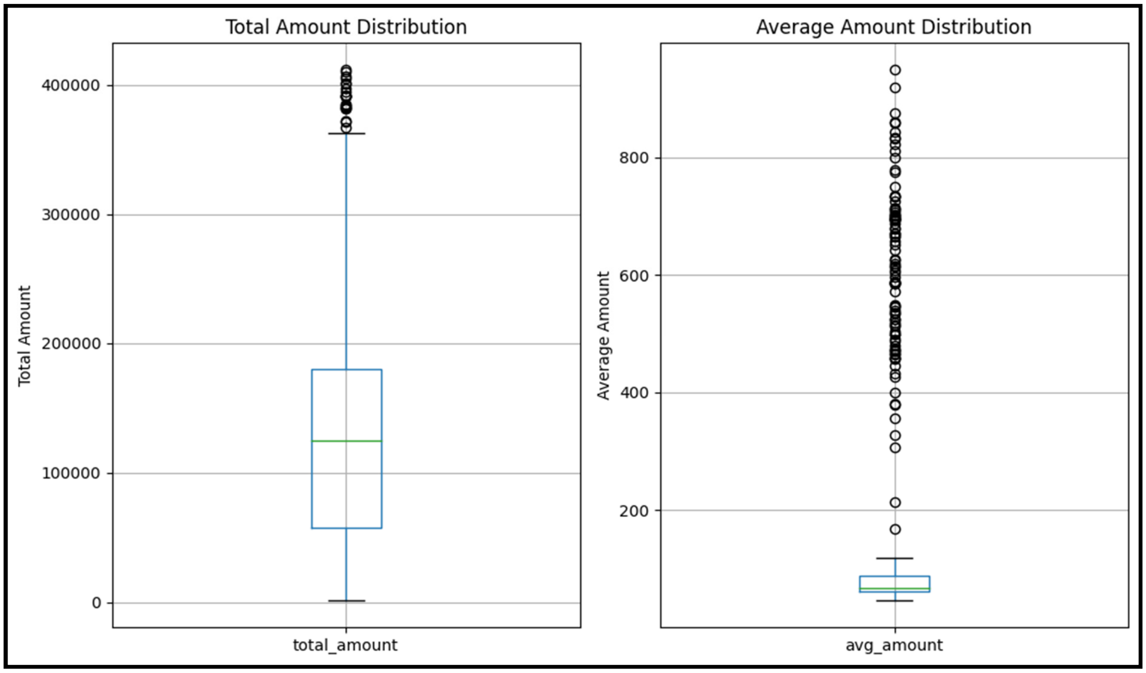


Fig 12. Total amount and average amount distribution

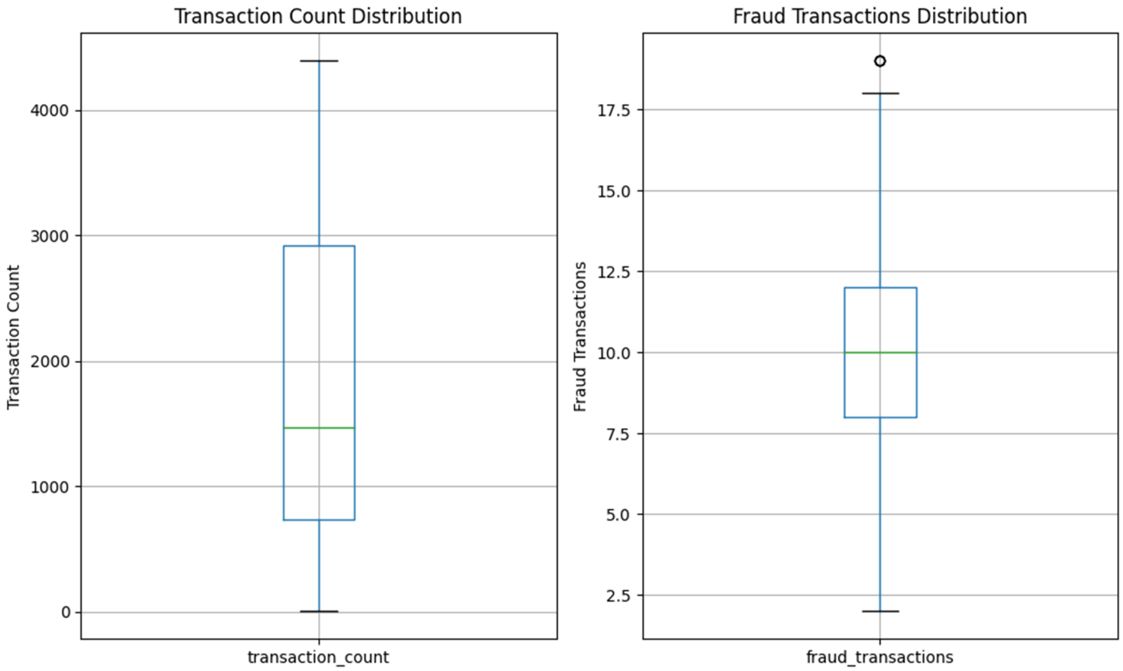


Fig 13. Transaction count distribution and fraud transaction distribution

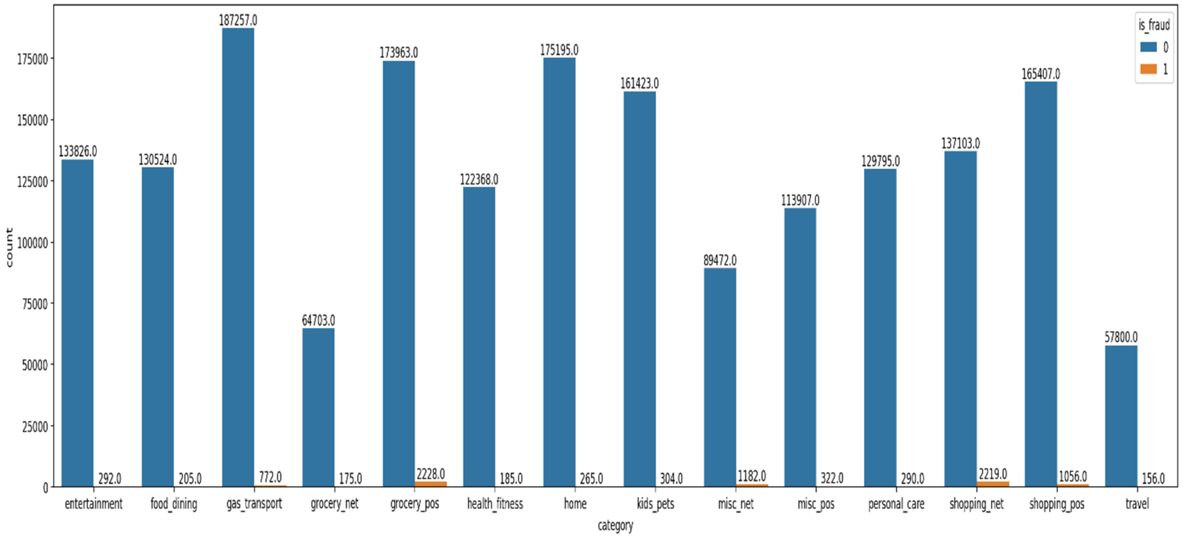


Fig 14. Number of frauds detected across various sectors based upon metadata extracted

Table 4: Training and Testing Accuracy on original data and metadata

Type of Data	Model	Training accuracy	Testing Accuracy
Original Data Governance	Logistic Regression	79.60	80.59
	Decision Tree	100	95.3
	Random Forest Classifier	96.42	93.19
	XgBoost	98.04	**96.56**
	KNN	75.98	67.70
Metadata Governance	ANN	**95.43**	**93.11**

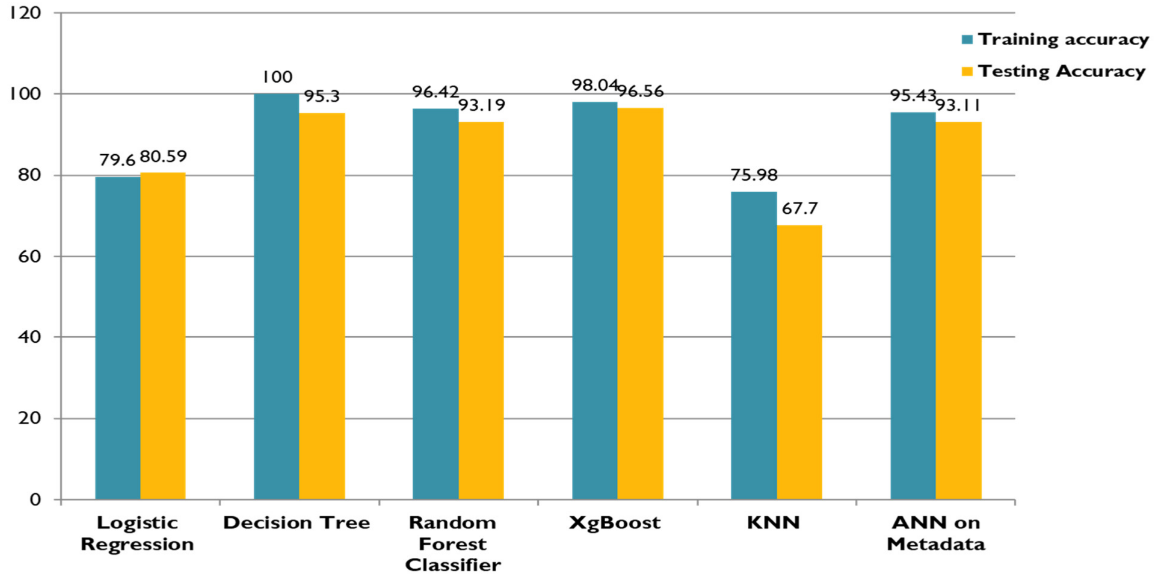
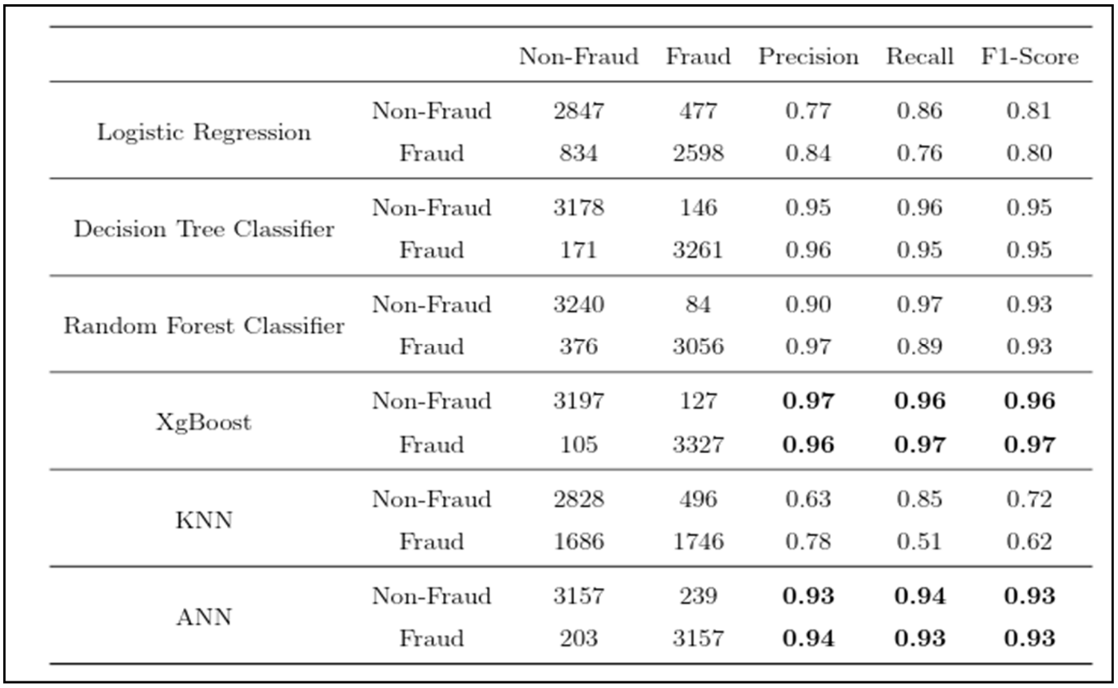


Fig 14. Training and Testing Accuracy of machine and deep learning algorithms

Table 4: Performance of various evaluation metrics for binary classification of classes into fraud and not fraud



		Non-Fraud	Fraud	Precision	Recall	F1-Score
Logistic Regression	Non-Fraud	2847	477	0.77	0.86	0.81
	Fraud	834	2598	0.84	0.76	0.80
Decision Tree Classifier	Non-Fraud	3178	146	0.95	0.96	0.95
	Fraud	171	3261	0.96	0.95	0.95
Random Forest Classifier	Non-Fraud	3240	84	0.90	0.97	0.93
	Fraud	376	3056	0.97	0.89	0.93
XgBoost	Non-Fraud	3197	127	**0.97**	**0.96**	**0.96**
	Fraud	105	3327	**0.96**	**0.97**	**0.97**
KNN	Non-Fraud	2828	496	0.63	0.85	0.72
	Fraud	1686	1746	0.78	0.51	0.62
ANN	Non-Fraud	3157	239	**0.93**	**0.94**	**0.93**
	Fraud	203	3157	**0.94**	**0.93**	**0.93**

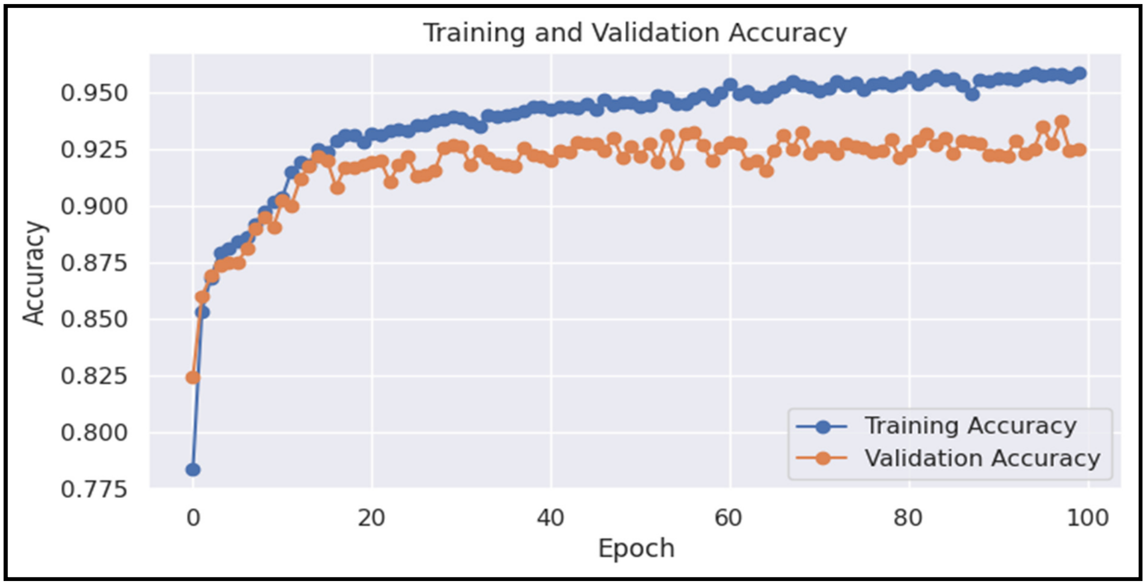


Fig 15. Training and Testing accuracy on Metadata extracted using ANN

Conclusion:

In this study, the analysis of both the original dataset and the extracted metadata has revealed significant patterns within the metadata itself. These patterns have demonstrated their utility in enhancing predictive analysis, leading to more optimized performance across the various models used. This outcome highlights the significant role of effective metadata management and governance in the big data ecosystem.

By integrating metadata rather than focusing solely on the original data, this research explores the potential for metadata to inform decision-making processes, improve data quality, and enhance system efficiency. The findings affirm that:

i. **Predictive Value of Metadata**: The study showcased how patterns embedded in metadata could inform predictions, aiding in more accurate analysis and improving the reliability of the models.

ii. **Optimization of Performance:** The effective governance of metadata led to noticeable optimizations in performance, particularly in processing times and resource management. This indicates that a focus on metadata can lead to tangible benefits beyond traditional data analysis methods.

iii. **Governance as a Key Component:** The governance structures implemented in this research ensured the consistency and quality. It becomes evident that the governance of metadata an integral part of data management.

iv. **Scalability and Adaptability:** As metadata governance frameworks evolve, they offer scalable solutions that can be adapted across different domains, demonstrating versatility and the capacity to enhance various data-driven applications.

This research affirms the growing importance of metadata in predictive analytics and system optimization. By promoting better management and governance of metadata, organizations can harness the power of this often-overlooked resource to enhance their operations and drive innovation. The findings suggest that future research should continue to explore new governance strategies, particularly in the integration of advanced technologies such as blockchain and encryption, which could further strengthen metadata management practices and their applications in diverse sectors.

References:

1. Y. Chen, H. Chen and P. Huang, "Enhancing the data privacy for public data lakes," in 2018 IEEE International Conference on Applied System Invention (ICASI), Chiba, Japan, 2018, pp. 1065-1068, doi: 10.1109/ICASI.2018.8394461.
2. M. Janssen, P. Brous, and E. Estevez, "Data governance: Organizing data for trustworthy Artificial Intelligence," Government Information Quarterly, vol. 37, no. 3, 2020, Art. no. 101493, doi: 10.1016/j.giq.2020.101493.
3. H. Mehmood, E. Gilman, M. Cortes, P. Kostakos, A. Byrne, K. Valta, S. Tekes, and J. Riekki, "Implementing big data lake for heterogeneous data sources," in 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW), 2019, pp. 23-26, doi: 10.1109/ICDEW.2019.00008.
4. M. Kumar, "Engaging in Big Data Transformation in the GCC," IDC White Paper, Sponsored by IBM, Dec. 2015.
5. Alserafi, A. Abelló, O. Romero, and T. Calders, "Towards Information Profiling: Data Lake Content Metadata Management," in 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016, pp. 1193-1198, doi: 10.1109/ICDMW.2016.0177.
6. Kulkarni, "A Study on Metadata Management and Quality Evaluation in Big Data Management," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 4, no. 7, July 2016, pp. 230-235.
7. S. Nandyala and H. K. Kim, "Big and Meta Data Management for U-Agriculture Mobile Services," International Journal of Software Engineering and Its Applications, vol. 10, no. 2, 2016, pp. 11-24.
8. M. Klettke, H. Awolin, U. Störl, D. Muller, and S. Scherzinger, "Uncovering the Evolution History of Data Lakes," in 2017 IEEE International Conference on Big Data, 2017, pp. 246-255, doi: 10.1109/BigData.2017.8257921.
9. J. P. Belaud, J. Le Lann, and F. Nourel, "Big Data for Agri-Food 4.0: Application to Sustainability Management for By-Products Supply Chain," Computers in Industry, vol. 111, 2019, pp. 41-50.
10. M. Janssen, et al., "Data governance: Organizing data for trustworthy Artificial Intelligence," Government Information Quarterly, vol. 37, no. 3, 2020, Art. no. 101493.
11. P. L. Martínez, F. J. García, and D. Navarro, "A Big Data-Centric Architecture Metamodel for Industry 4.0," Future Generation Computer Systems, vol. 125, pp. 263-284, 2021.
12. Cuzzocrea, "Big Data Lakes: Models, Frameworks, and Techniques," in 2021 Conference on Big Data and Smart Computing (BigComp), 2021, doi: 10.1109/BigComp51126.2021.00010.
13. Safder, A. Visvizi, T. Noraset, R. Nawaz, and S. Tuarob, "Deep Learning-based Extraction of Algorithmic Metadata in Full-Text Scholarly Documents," Information Processing and Management, 2020, doi: 10.1016/j.ipm.2020.102269.
14. R. B. Suresh Kumar, "Credit Card Fraud Detection Using Artificial Neural Networks," Global Transitions Proceedings, 2021, doi: 10.1016/j.gltp.2021.01.006.
15. Waterworth, S. Sethuvenkatram, and Q. Z. Sheng, "Advancing Smart Building Readiness: Automated Metadata Extraction Using Neural Language Processing Methods," Advances in Applied Energy, 2021, doi: 10.1016/j.adapen.2021.100041.
16. N. Jahnke and B. Otto, "Data Catalogs in the Enterprise: Applications and Integration," Datenbank-Spektrum, vol. 23, no. 1, pp. 89-96, 2023, doi: 10.1007/s13222-023-00445-2.
17. Atlan, "Gartner Data Governance," [Online]. Available: https://atlan.com/gartner-data-governance.
18. Ayman Alserafi, Alberto Abell´o, Oscar Romero, Toon Calders, Towards Information Profiling: Data Lake Content: Metadata Management, 2016 IEEE 16th International Conference on Data Mining Workshops.
19. Brian Stein, Alan Morrison," The enterprise data lake: Better integration and deeper analytics, Technology Forecast: Rethinking integration", Issue 1, 2014, Retrieved 25, Aug. 2017:

www.pwc.com/us/en/technology-forecast/2014/cloudcomputing/assets/pdf/pwc-technology-forecast-data-lakes.pdf.

20. Hassan Alrehamy Coral Walker, Personal Data Lake With Data Gravity Pull, 2015 IEEE Fifth International Conference on Big Data and Cloud Computing, 26-28 Aug. 2015, Dalian, China.
21. Natalia Miloslavskaya, Alexander Tolstoy, Application of Big Data, Fast Data and Data Lake Concepts to Information Security Issues, 2016 4th International Conference on Future Internet of Things and Cloud Workshops.