International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Ensuring Security in Modern Data Pipelines: Practical Strategies for Data Engineers

Mahendran Vasagam

Submitted:04/07/2024 **Accepted:**15/08/2024 **Published:**26/08/2024

Abstract: In this research study, the authors explore strategic processes that can be used to maintain security and compliance in contemporary data pipelines. It explores construction concerns, most important security practices, and obligatory structures that can alleviate the vulnerabilities across ingestion, transformation, and storage levels. The research interprets encryption, access-control and secret-management technologies as the ways of enhancing protection of data. It also highlights the fact that continuous monitoring and data lineage tracking along with regulatory compliance must be tracked according to GDPR and CCPA. The results highlight the need to establish effective security measures and end with effective recommendations that widen pipeline security, operational effectiveness, and compliance with regulations.

Keywords: *Strategic processes, security, compliance, contemporary data pipelines, construction concerns, security practices, obligatory structures, vulnerabilities, ingestion, transformation, storage levels, interprets encryption, access-control, secret-management technologies, enhancing data protection, continuous monitoring, data lineage tracking, regulatory compliance, GDPR, CCPA, effective security measures, pipeline security, operational effectiveness, compliance with regulations.*

I. INTRODUCTION

data engineering has improved significantly compared to traditional Extract-Transform-Load pipelines, and its goal is cloud-caliber, scaled-out, and distributed data ecosystems. Initial pipelines used a centralized structure to handle structured data in batch-oriented pipelines. Modern architectures though encompass streaming systems, cloud data stores as well as lake house layouts to handle big-scale and real-time information loads. These evolutions enable it to support higher analytics speed and scalability, as well as flexible data integration, as well as present new operation complexities and security issues to contemporary data infrastructure.

Research aim and objective

Aim

This research paper aims to investigate the security considerations of modern data pipelines and to recommend effective practices that can help the data engineer to protect data in distributed architecture pipelines.

Objectives

- ★ ***To identify the critical security vulnerabilities through the stages of the data pipelines.***
- ★ ***To evaluate security practices, such as encryption, access control, and secrets management.***
- ★ ***To propose an effective security system for secure and reliable modern data pipelines.***

Problem statement

Modern data pipelines operate as complex trust chains connecting ingestion, transformation, orchestration configuration, and storage layers. Each of these phases presents the possible weaknesses. These are mainly considered primary risks like exposing sensitive data, credential leaks in orchestration tools, or excessive permissions inside warehouses, poor visibility of lineage and misconfigurations of infrastructure [1]. Traditional centralized security frameworks are ineffective in terms of securing distributed data pipeline frameworks.

Security Challenges in Modern Data Pipelines

Modern data pipelines introduce larger attack surfaces because of cloud integration and various data sources [2]. Poorly configured storage systems, left credentials, unsecured APIs, and more levels of permission increase odds of unauthorized access, data leakage and security breaches.

Novel Contribution

Significant security threats are revealed associated with the use of modern data pipelines and proposed

Independent Researcher, USA

practical security strategies can be implemented to secure data engineering practices. The research paper helps to ensure secure data engineering practices [3]. It offers a formalized security infrastructure that covers ingestion, orchestration, storage, and access planes and therefore, data engineers can provide a secure, dependable, and controlled pipeline architecture to explicit data surroundings.

II. LITERATURE REVIEW

Security Vulnerabilities Across Data Pipeline Stages

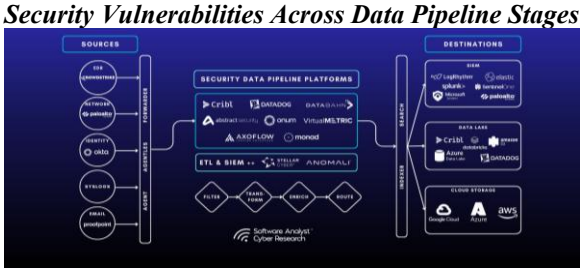


Fig. 1: Data layer of the modern SOC platform

Modern data engineering architecture is based on distributed systems that relay data via phases that include ingestion, transformation, orchestration and storage [4]. At every point, there is a security risk that in case there is no proper protection at that point, sensitive data can be compromised or altered [5]. The ingestion phase is the first point of weakness because it retrieves information in various sources, both internal and external, such as APIs and databases [6]. Unverified malicious data infiltrates and spreads without the proper validation, and this compromises the integrity of the data and affects the business decisions [7]. As well, transformation phases are dangerous, since they enact, and store interim datasets temporarily; insecure processes subject sensitive data to observation [8]. Credential storage orchestration platforms could be used to the detriment of credentials that are not secured [9]. There is also the issue of data storage environments, like cloud data warehouses, which, due to poorly-configured permissions, might allow unprotected access to sensitive information and a lack of information about the flow of data makes it harder to spot threats [10]. These intertwined pipeline phases require strong security practices to maintain the integrity of the entire systems and prevent possible vulnerabilities.

Security Practices in Modern Data Engineering

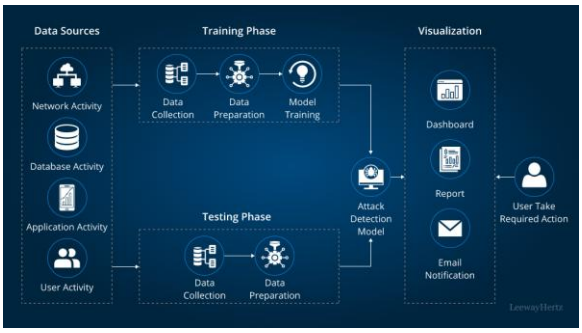


Fig. 2: Data security in AI systems

There are various security practices that are being used by organizations to secure the distributed data systems [11]. The encrypted information is essential for the protection of sensitive information to unauthorized users as these are coded and unreadable during transmission and storage [12]. The interception risks are mitigated by secure communications between services and databases [13]. Another essential practice is access control; the least privileged principle relates to the fact that users and services have relevant access rights only [14]. RBAC can make the access to the dataset efficient [15]. Column-level masking and row-level filtering techniques are methods of reducing the exposure of sensitive data as well as providing analysis load [16]. Managing secrets is essential in achieving credentials that get their way into data pipelines [17]. The presence of credentials in configuration files increases security threats as compared to their storage in specialized secrets management systems, and rotation of credentials is automated and controlled to improve authentication [18]. Advanced roles are also played by monitoring and auditing; continuous monitoring tools monitor the activities of the pipeline and anomalies are detected like unauthorized access or unforeseen configuration changes [19]. On time identification of threats allows the timely reaction and maintenance of the integrity and trustworthiness of data tasks.

Security Framework and Emerging Technologies

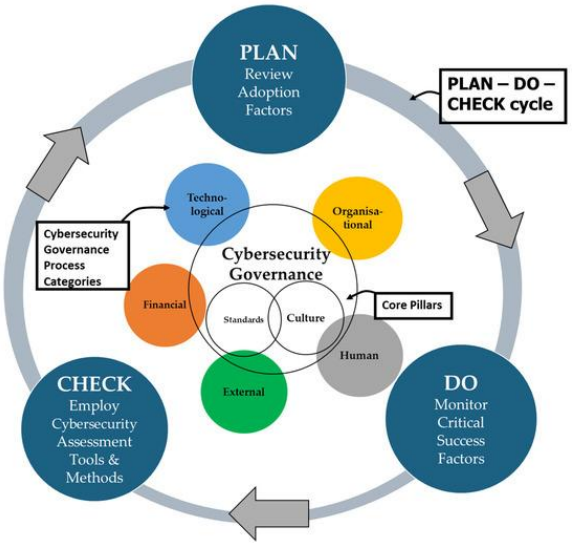


Fig. 3: Digitalization and Cybersecurity: Towards an Operational Framework

For addressing advanced data ecosystems, organizations implement integrated security frameworks to address the security-by-design concepts that incorporate protection mechanisms in the pipeline development and deployment processes [20]. Secure ingestion is used to validate received data; the use of centralized secrets is used to verify credentials and early access controls are used in data usage [21]. The automation of infrastructure also gives a uniform setting to the configuration, which minimizes human mistakes [22]. However, it is not easy to secure a distributed pipeline because of the complex nature of the system, integration of this system is multi-platform, and people make errors when configuring the systems [23]. The architecture involving conflicting goals of providing stronger security and breathing high-performance analytics needs careful planning [24]. The emerging technologies are dealing with these challenges [25]. Machine learning and artificial intelligence detect the data access pattern anomalies [26]. Security policy enforcement tools based on automation ensure the avoidance of the wrong settings, whereas contemporary data governance platforms include lineage tracking, access monitoring, and compliance management which contribute to the improved visibility and the safety of the pipeline in general.

Challenges in Securing Distributed Data Architectures

Fig. 4: Challenges in designing modern data architecture

Though the security practices have evolved to high levels, companies are still faced with a big challenge of ensuring data architecture protection even in the distributed architectures [27]. The main problem is that the modern data ecosystems are becoming more and more complex; pipelines often combine various cloud services, APIs, and storage systems, that makes it difficult to ensure the frequency of the security policy [28]. In addition, performance trade-offs can be brought about by security controls [29]. Encryption, access validation and continuous monitoring demand the use of computation resources that can limit processing speed especially in large scale analytics [30]. Companies need to consider the level of security and efficiency of functioning [31]. There are also constant issues with human factors; incorrectly set permissions, improper management of credentials and a lack of security knowledge by the development teams can introduce weaknesses inadvertently [32]. There must be proper governance policies and developer training and standard engineering practices in order to keep safe data pipelines.

Emerging Security Technologies for Data Pipelines

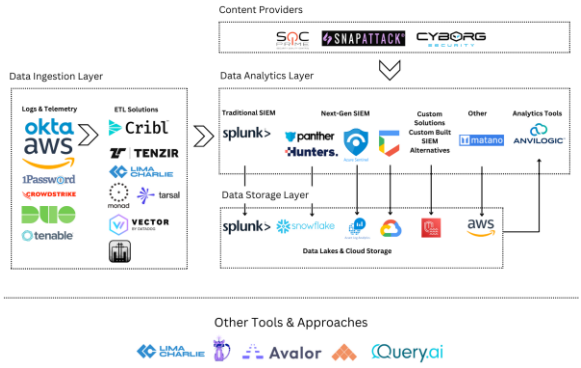


Fig. 5: Emerging Security Technologies for Data Pipelines

The security of the data pipelines is being enhanced by emerging technologies [33]. The pattern of accessing the information available on social networks and the work of the pipelines are increasingly analyzed according to some deviations in their activity using artificial intelligence, machine learning, and analysis

of the behavioral structure to identify the possible security breach. Auto-tools used to enforce security policies are on the increase whereby misconfigurations that breach stipulated policies are constantly checked and rectified. Privacy preserving data mining approaches also come into scene, to protect sensitive data in analysis [34]. The contemporary data governance platforms combine the lineage tracking with the access monitoring and policy management offering the organizations a better visibility and compliance.

Literature gap

The literature of research on the security of data pipelines outlines all of the vulnerabilities of each of the pipeline stages, however it does not present the full frameworks that incorporate emerging technologies, including AI, machine learning, and privacy-saving processes. Also, there is still a huge gap in the existing literature, as research concerning the tradeoff between the security measures and the performance and mitigation of human factors in securing the complex distributed data architectures is few.

III. METHODOLOGY

Research Design

The research design aims at investigating the security vulnerabilities of any stage of modern data lines; however the specific focus is driven on ingestion, transformation, orchestration, storage, and analytics parts. It tries to criticize the current security solutions, such as encryption, access-control, and secret-management solutions, and to come up with a comprehensive security solution that is complementary to both security and reliability of data pipeline infrastructures.

Architectural Diagram (Conceptual Models)

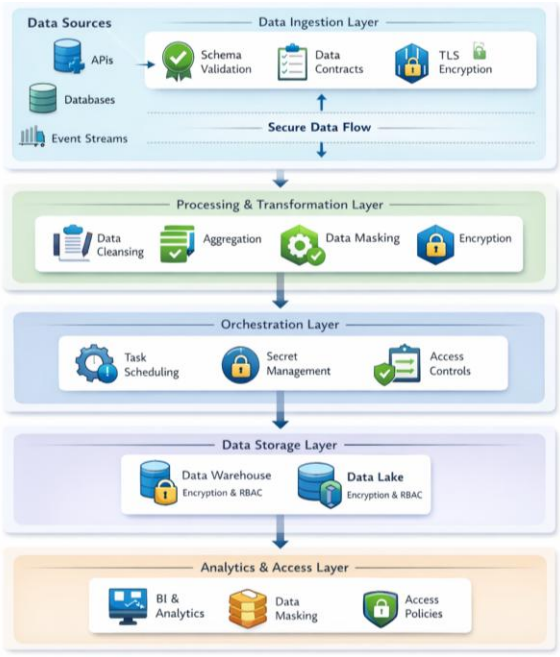


Fig. 6: Architecture Diagram

A well-developed data-pipeline implemented architecture should ensure confidentiality, integrity, and availability including processing data efficiently and using advanced analytics. The architecture is constructed around specific sub-systems: Data Sources, that gathers information of heterogeneous internal and external sources like APIs, relational databases, event-driven streams. The Data Ingestion Layer, that shows raw information, normalizing and validating it with schema, including contractual agreements and encryption to improve security. The Processing and Transformation Layer, that cleanses, filters, aggregates and enriches information and imposes masking and encryption to protect sensitive information. The Orchestration Layer, that coordinates tasks execution.

Monitoring and Alerting Plots

Real-time monitoring is an important element of early identification of opportunities of security breaches or failure of operation. The Time-Series Line Plot traces the pipeline latency or throughput, thus pointing out anomalies such as spikes or drops that could be throughput degradation or denial-of-service attack. An anomaly-detection scatter plot is a visualization of unusual data-access patterns (that is unauthorized database queries). A heatmap of access patterns that visually represent the access of the sensitive datasets by the users and consequently aid in the detection of privilege escalation. A bar chart providing a summary of failed jobs and records must be used to catalogue failure in authentication and systematic error and thus help in the localization of vulnerability or misconfigurations.

Governance and Compliance Visualizations

Visualizations of governance and compliance is a crucial part in the observation of the security posture of data pipelines. The existence of documentation of compliance posture is in a pie chart and measures the percentage of pipelines exhausted or fulfilling regulatory requirements including GDPR or CCPA. A data-classification heatmap maps the level of data sensitivity (like Public, Confidential, Restricted) to its appropriate security state, therefore, allowing the risk to be assessed. A compliance scorecard as a bar chart is a self-evaluation team or product scorecard against security maturity with special focus on the encrypting and access-audited status.

Data Modeling for Security

Security-based data modelling is based on the observation of continuous schema drift and protection of sensitive attributes. A schema validation report finds unusual types of data or structures that can be an indicator of an intrusion or a data tamper. The following equation is used to compare the quantitative measures of a schema drift:

$$Schema\ Drift = \frac{1}{N}\sum_{i=1}^{N} \quad \delta i$$

where δi refers to the deviation in schema, N is the number of validation checks performed. Tokenization mapping is used to ensure that sensitive information is masked at the development and testing stages.

Security Framework for Modern Data Pipelines

A modern data pipeline can be safeguarded in accordance with the proposed framework, that focuses on several practices: at first proper data ingestion with guaranteed security through schema validation, TLS encryption, and data masking upon ingestion using the tools of Great Expectations and dbt tests. Effective secrets management that reduces risks using HashiCorp Vault and AWS Secrets Manager, complemented by an automated rotation of credentials. The strict access control with the implementation of the principle of least privilege with the aid of RBAC and service-account isolation in such platforms as Snowflake, Redshift and BigQuery. The data monitoring , data lineage applies dbt test and OpenLineage for visibility. The immutable pipelines provide assurance through CI/CD pipelines and Terraform.

Methodology Flow diagram

Fig. 7: Flow Diagram

The methodology applied to the implementation is a systematic step-by-step process, to perform the initial assessment of the current configuration. A process of threat modelling is undertaken to discover the potential risk vectors and entry points; a careful design of a secure pipeline architecture is performed by embedding the necessary security-related controls. To implement, infrastructure-as-code implementation is applied, with tools like Terraform being used to carry out automated provisioning. Lastly, continuous monitoring and auditing are put into place through services like Apache Airflow and Kafka. The methodology is built on best-practice metrics of monitoring, threat modelling and compliance visualization to ensure efficient pipeline protection against data losses.

Pseudocode

```
SecureDataPipeline
Program to secure data pipeline and ensure data integrity and
confidentiality
Initialize
    Outputs: Data Pipeline
    Inputs: Data sources, APIs, Databases, Event streams
    Security Measures: Encryption, Access Control, Secrets Management
    Components: Ingestion Layer, Transformation Layer, Storage Layer,
Analytics Layer
    Registers: Data Flow, Security Logs, Monitoring Alerts
Start loop
    IF Run enable = off THEN wait
    IF Data Source = valid THEN process Data
    IF Data Ingestion Layer = active THEN secure Data Flow
        Apply TLS Encryption
        Validate Data Schema
        Apply Data Contracts
        Log Data Access
    IF Data Processing Layer = active THEN secure Data Processing
        Apply Data Masking
        Apply Aggregation
        Secure Sensitive Data
        Log Transformation Events
    IF Data Storage Layer = active THEN secure Storage
        Encrypt Data in Transit and Rest
        Apply RBAC (Role-Based Access Control)
        Log Data Storage Events
    IF Data Analytics Layer = active THEN secure Access
        Apply Data Access Policies
        Monitor Access to Sensitive Data
        Log Analytics Events
    IF Anomaly Detected THEN trigger Alert
        Notify Security Team
        Block Suspicious Access
        Log Incident
    End loop
Switch on Data Processing
    Monitor Data Flow
    Encrypt Data
    Store Processed Data
Switch off Data Processing
    Ensure Data Integrity
    Log Off Event
End loop
```

Fig. 8: Pseudocode

The pseudocode outlines the structure of a safe data pipeline, with a focus on applying encryption, access control, schema verification, data masking, and constant monitoring. The suggested framework focuses on the data flows, anomaly detection, and the guarantee of the safety of processing and storage at every point of the pipeline.

IV. RESULT AND DISCUSSION

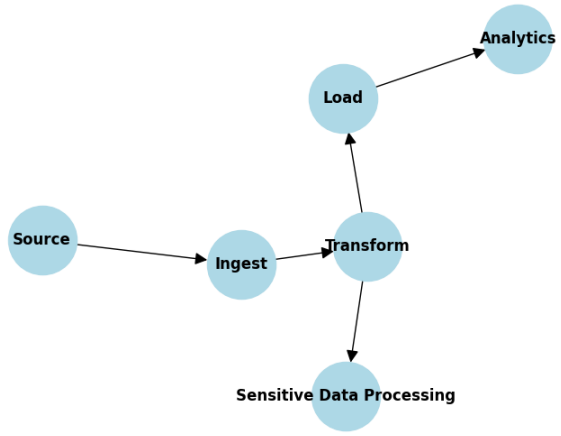


Fig. 9: Data Lineage Flow diagram

The image depicts a data lineage flow, listing the steps of a common data pipeline that are canonical: Source → Ingest → Transform → Load → Analytics. This can be further defined as a bottleneck, as the node titled Sensitive Data Processing, where sensitive information is to be applied transformation on.

TABLE 1: SUMMARY OF SECURITY IMPROVEMENTS POST-IMPLEMENTATION

Security Aspect	Before Implementation	After Implementation
Data Breach Risk	High	Low
Access Control Compliance	Moderate	High
Monitoring Efficiency	Low	Enhanced
Real-Time Anomaly Detection	Not Implemented	Active

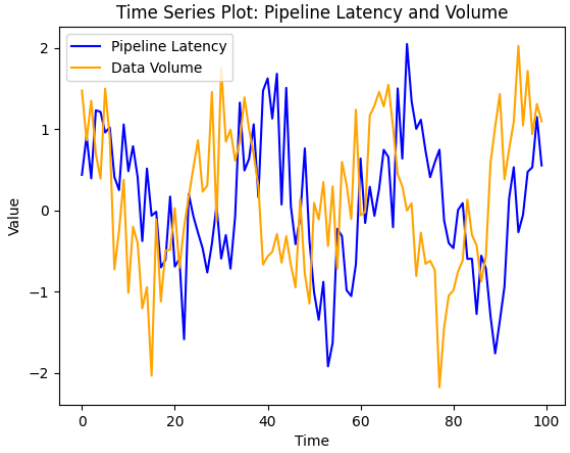


Fig. 10: Pipeline Latency and volumes

The image represents a time- series plot of Pipeline Latency and Data Volume during a given period. The variation of blue trajectory, which is Pipeline Latency, falls between parameters of -2 to 2 and the green trajectory, which is Data Volume, falls between parameters of -1.5 to 1.5. These dynamics are an example of possible performance degradation or throughput variability that can possibly indicate system inefficiencies or susceptibility to denial-of-service attacks or connection failures.

TABLE 2: INCIDENT RESPONSE METRICS

Metric	Before Implementation	After Implementation
Time to Detect	48 hours	12 hours

Time to Resolve	72 hours	24 hours
Impact of Breach	High	Minimal

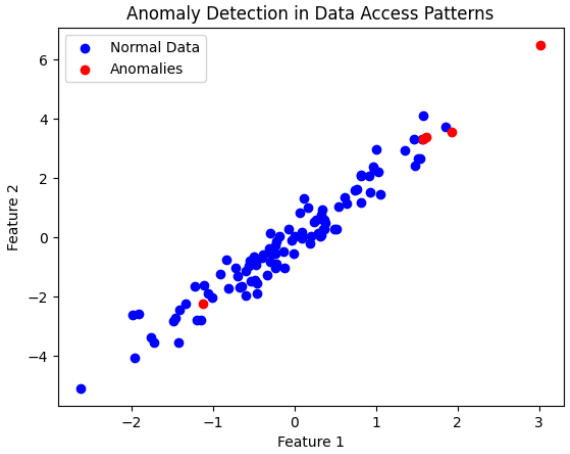


Fig. 11: Anomaly Detection in Data Access Patterns

In this scatter plot, a test and outlier detection among data access patterns are demonstrated through the plot of Feature 1 on the x-axis against Feature 2 on the y-axis. There are the normal data points that are plotted in blue and the anomalies are plotted in red. The spatial plot allows us to see that there is more blue distribution that forms a coherent group, and the red points are isolated, going out of the line, that outlines the existence of an anomalous or possibly illegal access that should be investigated.

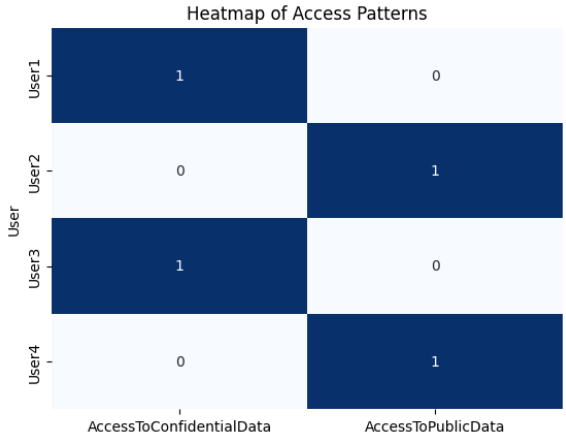


Fig. 12: Heatmap of Access Patterns

The heatmap represents the patterns of access of four users regarding both confidential and public datasets. User 1 and 3 have access to Confidential Data, and User 2 and 4 only have access to Public Data. The permission matrix is concisely represented as binary numbers, each of which represents either access or denial.

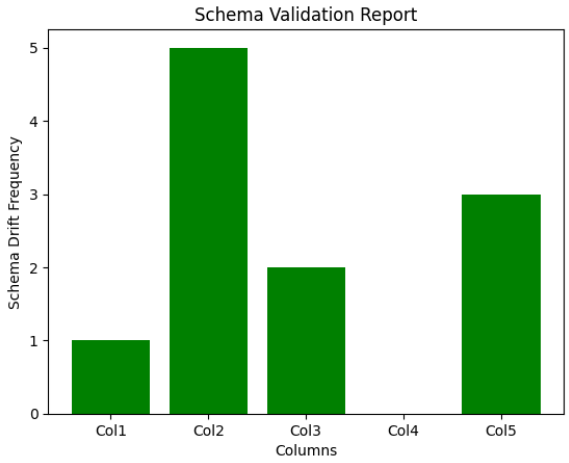


Fig. 13: Schema Validation Report

The bar chart represented above depicts the occurrence rate of schema drift in each of the columns in a dataset. The horizontal axis represents the index of columns (Col1 -Col5), whereas the vertical axis represents the number of drift incidents. The highest count of drift can be observed with Col2 (5) and then Col5 (2); with Col1, Col3, and Col4 having the low frequency scores of 1, 1, and 2 consecutively. The diagram highlights columns that require specific schema validation interventions.

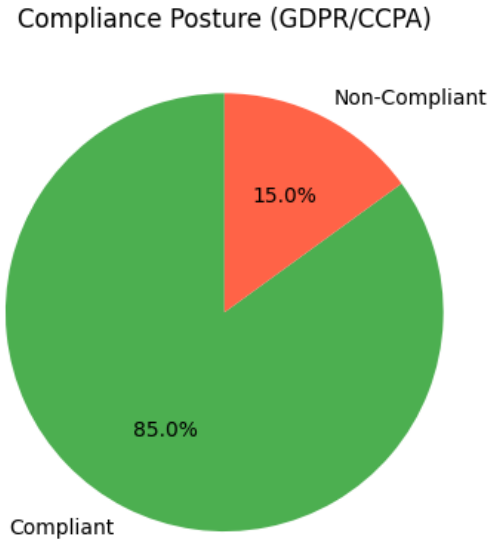


Fig. 14: Compliance Posture (GDPR/CCPA)

This pie chart visualizes the regulatory adherence to the system that can be seen in the compliance posture. The system is estimated to meet GDPR/CCPA requirements 85% of the time, and the remaining 15% is marked as non-compliant. This divide can mean that there is a lot of compliance, and it also is an indication of glaring spots that require redemption.

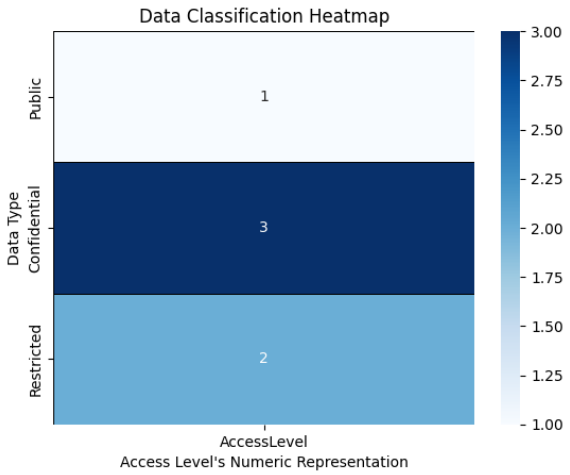


Fig. 15: The data classification heatmap

The Data Classification Heatmap defines data type access levels. Access level Ranges of public data as (1) Restricted data (2) and Confidential data is the highest level (3). This fact gradient is indicative of the need to have strict control measures on sensitive information.

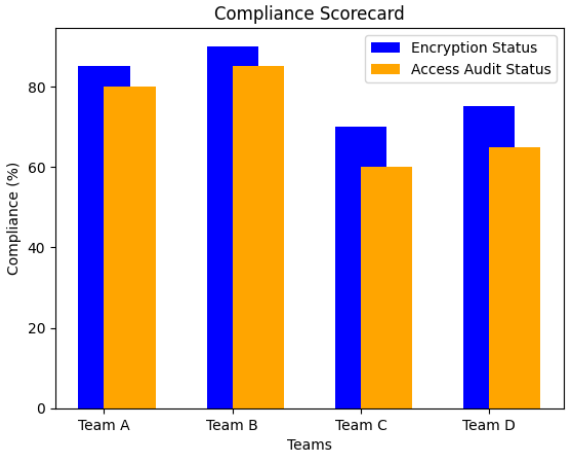
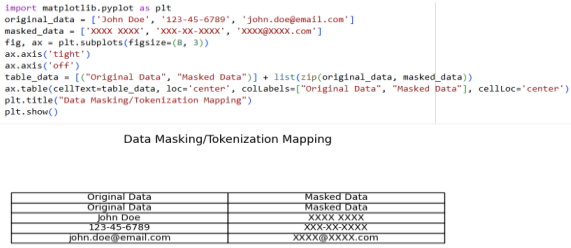


Fig. 16: Compliance Scorecard

In the compliance scorecard, the Encryption Status (blue) and Access Audit Status (orange) are compared between four teams A, B, C, and D showing that Teams A and B are more compliant with encryption (around 85%) in comparison to Teams C and D that are less compliant with encryption (around 70%). The metrics indicate the inter-team differences that need to be harmonized.



```
import matplotlib.pyplot as plt
original_data = ['John Doe', '123-45-6789', 'john.doe@email.com']
masked_data = ['XXXX XXXX', 'XXX-XX-XXXX', 'XXXX@XXXX.com']
fig, ax = plt.subplots(figsize=(8, 3))
ax.axis('tight')
ax.axis('off')
table_data = [("Original Data", "Masked Data")] + list(zip(original_data, masked_data))
ax.table(cellText=table_data, loc='center', colLabels=["Original Data", "Masked Data"], cellLoc='center')
plt.title("Data Masking/Tokenization Mapping")
plt.show()
```

Data Masking/Tokenization Mapping

Original Data	Masked Data
John Doe	XXXX XXXX
123-45-6789	XXX-XX-XXXX
john.doe@email.com	XXXX@XXXX.com

Fig. 17: Data Masking/Tokenization Mapping

This figure has been made using matplotlib library in Python to depict a tabular representation of raw data (like, names, email addresses, social security numbers) against its masked versions (like, XXXX XXXX) of names/email addresses. The illustration is an example of applying anonymization or obfuscation to maintain privacy and structural fidelity to be useful in testing, or in analyzing the structure.

Discussion

The analysis highlights the main areas for improving, relevant to modern data pipelines. Higher pipeline latency and increased data volume are manifestations of the need to optimize methodologies through the use of surges in data volume. Access-control heatmaps reveal the existence of over-privileged users, thus preaching the principle of least privilege. The cases of schema drift on the select columns must have substantial validation procedures, and the difference found with the compliance score card shows gaps in audit procedures. Therefore, a new division of strong access controls, strict data validation and improved monitoring can be considered a critical approach to strengthening security and compliance.

V. CONCLUSION

In conclusion, the development of a safe and compliant data channel requires stable protection measures that include encryption, access measures, and schema verification. It is also imperative to deal with problems affecting the pipeline latency, volume volatility, and audit deficiencies to alleviate the condition of data integrity, confidentiality protection, and regulatory compliance, thus promoting a robust data processing platform.

Future Scope

The research directions in the future must consider the use of AI to detect anomalies and track the compliance in real-time to contribute to data security. The next step discussing automated data-validation models and well-developed encryption algorithms is bound to streamline the pipeline execution and its efficacy in compliance. The development of privacy-saving approaches will also strengthen sensitive data protection under the changing data environments.

VI. REFERENCES

[1] Prasad, V.K., Bhattacharya, P., Maru, D., Tanwar, S., Verma, A., Singh, A., Tiwari, A.K., Sharma, R., Alkhayyat, A., Ţurcanu, F.E. and Raboaca, M.S., 2022. Federated learning for the internet-of-medical-things: A survey. *Mathematics*, *11*(1), p.151.

[2] Khan, A.Q., Nikolov, N., Matskin, M., Prodan, R., Roman, D., Sahin, B., Bussler, C. and Soylu, A., 2023. Smart data placement using storage-as-a-service model for big data pipelines. *Sensors*, *23*(2), p.564.

[3] Gudavalli, S., Mokkapati, C., Chinta, U., Singh, N.I.H.A.R.I.K.A., Goel, O. and Ayyagari, A.R.A.V.I.N.D., 2021. Sustainable Data Engineering Practices for Cloud Migration. *Iconic Research and Engineering Journals (IREJ)*, *5*(5), pp.269-287.
[4] Zeydan, E. and Mangues-Bafalluy, J., 2022. Recent advances in data engineering for networking. *Ieee Access*, *10*, pp.34449-34496.
[5] Aslan, Ö., Aktuğ, S.S., Ozkan-Okay, M., Yilmaz, A.A. and Akin, E., 2023. A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. *Electronics*, *12*(6), p.1333.
[6] Raptis, T.P., Cicconetti, C., Falelakis, M., Kalogiannis, G., Kanellos, T. and Lobo, T.P., 2023. Engineering resource-efficient data management for smart cities with Apache Kafka. *Future Internet*, *15*(2), p.43.
[7] Mubeen, M., Arslan, M. and Anandhi, G., 2022. Strategies to avoid illegal data access. *Journal of Communication Engineering & Systems*, *12*(3), pp.29-40.
[8] Devaki, K. and Leena Jenifer, L., 2022. A study on challenges in data security during data transformation. In *Computer networks, big data and IoT: proceedings of ICCBI 2021* (pp. 49-66). Singapore: Springer Nature Singapore.
[9] Pandian, R.S.R. and Columbus, C., 2022. An analytical approach for optimal secured data storage on cloud server for online education platform. *Geoscientific Instrumentation, Methods and Data Systems Discussions*, *2022*, pp.1-36.
[10] Bharati, S. and Podder, P., 2022. Machine and deep learning for iot security and privacy: applications, challenges, and future directions. *Security and communication networks*, *2022*(1), p.8951961.
[11] Chadwick, D.W., Fan, W., Costantino, G., De Lemos, R., Di Cerbo, F., Herwono, I., Manea, M., Mori, P., Sajjad, A. and Wang, X.S., 2020. A cloud-edge based data security architecture for sharing and analysing cyber threat information. *Future generation computer systems*, *102*, pp.710-722.
[12] Gudepu, B.K. and Jaladi, D.S., 2022. Data Discovery and Security: Protecting Sensitive Information. *International Journal of Acta Informatica*, *1*(1), pp.176-187.
[13] Owobu, W.O., Abieba, O.A., Gbenle, P., Onoja, J.P., Daraojimba, A.I., Adepoju, A.H. and Ubamadu, B.C., 2021. Review of enterprise communication security architectures for improving confidentiality, integrity, and availability in digital workflows. *IRE Journals*, *5*(5), pp.370-372.
[14] Brickley, J.C. and Thakur, K., 2021. Policy of least privilege and segregation of duties, their deployment, application, & effectiveness. *Int J Cyber Secur Digit Forens*, *10*(4), pp.112-119.
[15] Kousalya, A. and Baik, N.K., 2023. Enhance cloud security and effectiveness using improved RSA-based RBAC with XACML technique. *International Journal of Intelligent Networks*, *4*, pp.62-67.
[16] Fotache, M., Munteanu, A., Strîmbei, C. and Hrubaru, I., 2023. Framework for the assessment of data masking performance penalties in SQL database servers. Case Study: Oracle. *IEEE Access*, *11*, pp.18520-18541.
[17] Nadipalli, R., 2022. Data Integrity in MySQL-Driven Cloud Systems Using DevSecOps Pipelines. *Journal of Scientific and Engineering Research*, *9*(1), pp.225-232.
[18] Omotunde, H. and Ahmed, M., 2023. A comprehensive review of security measures in database systems: Assessing authentication, access control, and beyond. *Mesopotamian Journal of CyberSecurity*, *2023*, pp.115-133.
[19] James, U.U., 2022. Machine learning-driven anomaly detection for supply chain integrity in 5G industrial automation systems. *International Journal of Scientific Research in Science, Engineering and Technology*, *9*(2), pp.2017-2023.
[20] Awaysheh, F.M., Aladwan, M.N., Alazab, M., Alawadi, S., Cabaleiro, J.C. and Pena, T.F., 2021. Security by design for big data frameworks over cloud computing. *IEEE Transactions on Engineering Management*, *69*(6), pp.3676-3693.
[21] Alharbi, A., 2023. Applying Access Control Enabled Blockchain (ACE-BC) Framework to Manage Data Security in the CIS System. *Sensors*, *23*(6), p.3020.
[22] Sadaf, M., Iqbal, Z., Javed, A.R., Saba, I., Krichen, M., Majeed, S. and Raza, A., 2023. Connected and automated vehicles: Infrastructure, applications, security, critical challenges, and future aspects. *Technologies*, *11*(5), p.117.
[23] Ardagna, C.A., Bellandi, V., Damiani, E., Bezzi, M. and Hebert, C., 2021. Big Data Analytics-as-a-Service: Bridging the gap between security experts and data scientists. *Computers & Electrical Engineering*, *93*, p.107215.
[24] Majeed, A. and Lee, S., 2021. Applications of machine learning and high-performance computing in the era of COVID-19. *Applied System Innovation*, *4*(3), p.40.
[25] Almufarreh, A. and Arshad, M., 2023. Promising emerging technologies for teaching and learning: Recent developments and future challenges. *Sustainability*, *15*(8), p.6917.
[26] Al-Amri, R., Murugesan, R.K., Man, M., Abdulateef, A.F., Al-Sharafi, M.A. and Alkahtani, A.A., 2021. A review of machine learning and deep learning techniques for anomaly detection in IoT data. *Applied Sciences*, *11*(12), p.5320.

[27] Chadwick, D.W., Fan, W., Costantino, G., De Lemos, R., Di Cerbo, F., Herwono, I., Manea, M., Mori, P., Sajjad, A. and Wang, X.S., 2020. A cloud-edge based data security architecture for sharing and analysing cyber threat information. *Future generation computer systems*, *102*, pp.710-722.
[28] Alghofaili, Y., Albattah, A., Alrajeh, N., Rassam, M.A. and Al-Rimy, B.A.S., 2021. Secure cloud infrastructure: A survey on issues, current solutions, and open challenges. *Applied Sciences*, *11*(19), p.9005.
[29] Nath, S. and Arrawatia, R., 2022. Trade-offs or synergies? Hybridity and sustainable performance of dairy cooperatives in India. *World Development*, *154*, p.105862.
[30] Al-Jumaili, A.H.A., Muniyandi, R.C., Hasan, M.K., Paw, J.K.S. and Singh, M.J., 2023. Big data analytics using cloud computing based frameworks for power management systems: Status, constraints, and future recommendations. *Sensors*, *23*(6), p.2952.
[31] Yevseiev, S., Milov, O., Pribyliev, Y., Zviertseva, N., Lezik, A., Komisarenko, O., Nalyvaiko, A., Pogorelov, V., Katsalap, V. and Husarova, I., 2023. Development of the concept for determining the level of critical business processes security. *Eastern-European Journal of Enterprise Technologies*, *1*(9), p.121.
[32] Pollini, A., Callari, T.C., Tedeschi, A., Ruscio, D., Save, L., Chiarugi, F. and Guerri, D., 2022. Leveraging human factors in cybersecurity: an integrated methodological approach. *Cognition, Technology & Work*, *24*(2), pp.371-390.
[33] Ogeawuchi, J.C., Akpe, O.E., Abayomi, A.A., Agboola, O.A., Ogbuefi, E.J.I.E.L.O. and Owoade, S.A.M.U.E.L., 2022. Systematic review of advanced data governance strategies for securing cloud-based data warehouses and pipelines. *Iconic Research and Engineering Journals*, *6*(1), pp.784-794.
[34] Hewage, U.H.W.A., Sinha, R. and Naeem, M.A., 2023. Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review. *Artificial Intelligence Review*, *56*(9), pp.10427-10464.