International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Adaptive and Secure ETL for Defense Data Systems Using Federated Reinforcement Learning

Manohar Reddy Sokkula

Sr. Solutions Architect, Corpay

Abstract: ***In increasingly complex defense environments, ensuring secure, adaptive, and compliant data processing has become essential. This paper presents the Adaptive Secure ETL based on Federated Reinforcement Learning (AS-ETL-FRL), a novel framework designed to handle sensitive Department of Defense (DoD) data. Integrating federated deep autoencoders for anomaly detection with reinforcement learning agents for dynamic optimization, the system offers secure extraction, transformation, and loading (ETL) pipelines. Tested on the IBM Cloud Console Anomaly Detection Dataset, the model demonstrated 98.5% accuracy in anomaly detection, minimized latency, and maintained compliance with NIST, FISMA, and CMMC standards. By avoiding raw data sharing and enabling real-time decision-making, the AS-ETL-FRL architecture contributes significantly to privacy-preserving, intelligent data governance for mission-critical systems.***

Keywords: Federated Learning, Reinforcement Learning, DoD Data Management, Anomaly Detection, Explainable AI

1. Introduction

Information integrity, confidentiality, and availability are the most important factors in the Department of Defense. In this regard, the Technology Lifecycle processes can be critical in ensuring that the data management systems are secure, resilient and mission-driven [1]. Good TL processes involve the methodical planning, development, deployment, maintaining and retirement of data assets and technologies [2]. These activities make sure that the data in the DoD ecosystem is handled with accuracy and responsiveness throughout its life cycle, starting with the way it is initially acquired, to the way it is ultimately disposed [3]. With the digital transformation accelerating, and the volume of data growing, it is necessary to have a solid framework in place to control sensitive defense information by integrating governance, compliance, and technical precision. [4].

An organized TL process can aid the DoD in eliminating the risk of data corruption, unauthorized access, and vulnerabilities in the system [5]. Security measures and checks are incorporated in every stage of the technological lifecycle, including the stages of concept development, penetration testing, and moving to the production phase [6]. This includes the establishment of robust authentication tools, encryption levels, access control measures, and real-time tracking tools [7]. Additionally, the data management technologies should be evaluated and updated continuously to respond to the emerging threats and keep pace with the changing defence demands. TL processes guarantee high standards of reliability and confidentiality of data in both legacy and modern systems by including proactive risk assessment and stringent testing [8].

In addition to technical protection, TL operations in DoD data management focus on policy adherence, responsibility, and interdepartmental cooperation. Governance models have been developed in such a way that they delineate the ownership of data, provide accountability over information that is handled and implement policies that are aligned with the national security guidelines. Frequent reviews and audits are used to ensure transparency, whereas training programs can improve the awareness of the staff about the duties in data protection [9]. Bringing in cybersecurity models, including zero-trust architectures, also enhances the internal and external threats. Finally, the rigorous implementation of TL processes will allow the DoD to make sure that information is an asset of strategic value, accurate, secure and reliable, and will be used to make informed decisions and maintain operational advantage in a more digitized defense environment [10]. This study aims to develop a privacy-preserving, adaptive, and compliant ETL framework for DoD data environments by integrating federated learning and reinforcement learning into secure data pipelines.

This study aims at producing a safe, smart and defensible ETL (Extract, Transform, Load) system that is specific to the Department of Defense (DoD) to guarantee integrity, confidentiality and functionality of data in the management of large volumes of data. This is driven by the increasing complexity of DoD data ecosystems in which the conventional ETL operations are challenged by anomalies in the data, cyber threats, and the enforcement of compliance in situations in which the distributed and classified data is involved. The existing constraints are overcome by combining AI-based anomaly detection and federated reinforcement learning to develop an adaptive ETL system to optimize operations in real time without breaching data privacy. The importance of this work lies in the fact that it is the first due to the appearance of federated privacy-preserving architecture that provides high accuracy in the detection of anomalies (98.5 percent) and corresponds to the standards of NIST, FISMA, and CMMC. Not only does this innovation improve data-guarantee and data-automation in secure settings, but plants the foundation of further innovations, which may include zero-trust ETL structures, blockchain-indicated data provenience, and additional underpinnings of a robust and smart data processing framework in the national defense infrastructures. The paper has a number of important contributions to the development of security, integrity and operational resilience of ETL (Extract, Transform, Load) processes in the Department of Defense (DoD) data management environment. The key contributions are summarized as follows:

- Developed a DoD-oriented secure ETL pipeline integrating encryption, tokenization, and compliance-based data validation to ensure confidentiality and integrity during data flow.
- Implemented a deep autoencoder-based anomaly detection model coupled with reinforcement learning to automatically identify, respond to, and optimize ETL operations in real time.

- Introduced a federated learning framework enabling decentralized model training across multiple DoD data domains without exposing sensitive information.
- The proposed framework establishes a foundation for extending ETL systems with blockchain-based lineage tracking, zero-trust architectures, and AI-driven forensic readiness, promoting end-to-end security, transparency, and accountability in defense data operations.

The rest of the research is organised as follows: The Introduction is explained in Section I, and literature review is explained in Section II and Section III explains the research Methodology in a detailed manner, Section IV presents the results, and Section V explains the Conclusion and Future Work of the paper.

2. Related Works

A. Review of AI Applications in Cybersecurity

The article of [11] is a new method of making ETL work more efficiently via the DOD-ETL framework that creates an on-demand data stream pipeline infrastructure that is distributed, parallel, and technology-neutral. The algorithm uses in memory caching and effective data partitioning to optimize ETL loads so that the system can process a large amount of data with minimum latency. The framework that was used in the study was capable of attaining near real-time processing of ETLs and running workloads ten times more than conventional, historically used stream processing systems. The solution was also tested in a large industrial steelwork setting and proved to be effective in high-throughput and real-life applications. Although these are the benefits, the generalizability of the study to the contexts of the Department of Defense (DoD) is questionable as it was mostly applied to industrial environments. Its scalability, flexibility and security in multi-domain, sensitive DoD environments, where compliance and data confidentiality is paramount, is still unanswered. However, this publication offers a solid background of investigating distributed and real-time ETL systems that have the potential to be modified to defense-grade data management.

Malaika Bareen [12] discusses the important facet of security and compliance in ETL pipelines and deals with vulnerabilities emerging when extracting, transforming, and loading data. Access control, data masking, audit logging, and encryption are the underlying mitigation strategies proposed by the study as their major threats include unauthorized access, data leakage, and compliance violations. These systems are placed as enablers of regulatory compliance, especially when the regulatory frameworks, such as GDPR and HIPAA, are involved. The strength of the work is its ability to clearly state the best practices to strike a balance between the security and performance efficiency. Nevertheless, the research is very conceptual as it discusses the industry standards, as opposed to the defense-grade. The fact that it does not focus on Department of Defense (DoD) settings implies that questions about the issues of dealing with classified data, multi-level security spaces, and the integrity of systems of critical importance to the missions are not researched. Nevertheless, the document gives a crucial security base to modify ETL pipelines to more demanding, military-tier compliance environments.

The article by the [13] is a quantitative study of the data compliance metrics when it comes to ETL workflows, with a special focus on the measurement of the return on investment (ROI) of security-driven ETL implementations. The methodology will entail assessing compliance indicators like the data retention policy adherence, encryption coverage, and access control audit success in the sectors, financial and healthcare, which are regulated. It is found that structured compliance monitoring not only increases data integrity but also saves a substantial portion of costs by decreasing incident response overheads. The present research is a rare economic view of ETL security, demonstrating the way that compliance investments pay off in quantifiable business value. It is not transferable to defense applications, however, because of its industrial orientation, where ROI measures may be less important than resilience to operations and protection of classified data. Nevertheless, its experiences on measurable compliance can be effectively used to structure DoD-oriented ETL performance metrics that incorporate both security measures and operational measures.

Dhamotharan Seenivasan [14] offers a prospective overview of the security issues arising in ETL workflows, offering a set of technical improvements that can be used to address the threat of possible attacks. The paper proposes a multilayered implementation with the inclusion of high-order encryption algorithms, dynamic access control, real-time intrusion detection and intrusion prevention systems (IDPS), and integrity verification protocols. These mechanisms are meant to protect sensitive data over distributed ETL pipelines and still preserve the performance of the workflow. The contribution of the paper is in its comprehensive approach to security layers when it comes to the ETL operations instead of seeing them as its external add-ons. Nevertheless, its limitations lie in its generalized orientation, in that even though it is very comprehensive, the paper does not explicitly comment on DoD-style architectures which require adherence to classified data standards and multi-domain security policies. The efforts to adapt the suggested framework related to DoD systems would involve the addition of other elements like cross-domain guards and policy-conscious data routing.

M. Souibgui [15] provides a background discussion of the problem of data quality in ETL processes with a focus on the fact that low data quality frequently compromises the quality of analytical procedures and the reliability of the decisions. The research splits out major data quality dimensions like completeness, consistency and timeliness and examines available methods of making sure of reliable transformations. The analysis carried out by Souibgui highlights that ETL systems should also incorporate validation and cleansing controls to avoid descending of the data when being transferred. Although the research is valuable as a foundational study, its preliminary status restricts the practical value of its research and does not extrapolate the research to defense or mission-critical systems. This would only be further complicated in the context of DoD, where encrypted data, federated sources, and policy constraints are all factors that need to be considered when ensuring the quality of the data- areas not directly investigated in this paper. However, it offers the theoretical framework for the

definition of quality-oriented measurements in secure ETL pipelines in sensitive settings.

The study [16] outlines what can be done to provide optimal performance to ETL with large volumes of data in warehousing, which is aimed at increasing the architectural efficiency and the improvement of throughput. Some of the strategies that the authors discuss to deal with petabyte datasets include parallel loading of data, partition-based transformations, and workload scheduling. The experimental evidence shows significant improvements in data throughput and a decrease in latency in the experimental frameworks as compared to traditional ETL frameworks. The study has a high degree of relevance in DoD environments that process huge amounts of operational and intelligence data. But the paper lacks in terms of incorporating a security or compliance aspect, which is not negotiable in a defence-grade application. These optimization strategies can be extended into secure and compliant ETL systems with the possibility of achieving hybrid models between performance and integrity.

The article written by [17]offers a comprehensive history of data integration that was developed as a replacement of the old ETL frameworks to big data frameworks. The authors examine the transformation of ETL architectures (based on batch) into real-time, distributed, and metadata-driven models that allow scalability and agility. Big challenges that the study points out include schema heterogeneity, data latency and inconsistency in governance. It is strong because it follows the process of technological revolution that forms the basis of modern ETL design thinking. Nonetheless, it does not have particular attention toward military or DoD-oriented data ecosystems, whereby each hybrid architecture needs to balance interoperability with security classification layers. Nonetheless, the review provides a solid conceptual foundation on how to incorporate the use of big data ETL technologies in the defense intelligence systems.

In this study, [18] explores the topic of data transformation as the main part of the ETL flows. The paper indicates different transformation operations which include aggregation, normalization, pivoting, and schema mapping and assess the level of computing efficiency and effect on the usefulness of data. It stresses the necessity of adaptive logic of transformation which is capable of reacting to evolving schema and streaming information. Even though the paper successfully maps the challenges of transformation to the performance outcomes, the issue concerning the security implication of transformation within the classified environment has not been addressed. In DoD applications, transformations that preserve encryption computation and anonymization methods (not in this study) should also be transformed. However, the article offers a thorough technical insight on transformation processes that lie behind the provision of secure ETL engineering.

The paper examines [19] how metadata-based methods can be used to achieve the effectiveness and reliability of ETL. The metadata frameworks are suggested as the way of facilitating the automation, lineage tracking, and schema evolution in the complex data ecosystems. ITIL 4 paper notes that metadata-supported integration enhances transparency and reduces auditing which is essential in ensuring compliance. Nevertheless, it only analyzes on the general enterprise system, as it does not specifically analyze the classified data management that is required in the DoD. In ETL pipelines that were of a military grade, metadata structures might be extended to accommodate access labels, clearance levels as well as enforcement of cross domain policy, beyond the scope of the current study.

Lastly, the [20] presents an ETL pipeline model which is an automated system that is supposed to enhance the quality and governance of data. The study shows that with the application of machine learning as well as rule-based automation, the rate of manual intervention and errors are greatly reduced. Automation results in better consistency in enforcing compliance and recovery in data anomalies. Although this paper is a step in the right direction of internet automation, the overall generalized approach fails to consider the limitations of the defense landscape, including the ability to securely compartmentalize and trace modification of classified data. Nonetheless, it offers a worthwhile foundation of research to create self-optimizing ETL systems that may include reinforcement learning that enables them to adapt to data in sensitive defense networks.

Altogether, the total number of these ten studies pinpoints the shift of ETL processes to more efficient, secure and intelligent structures. Nevertheless, the majority of them are commercial or enterprise-driven, which lack research in devising ETL systems that are specifically designed in the Department of Defense data environment where data sensitivity, compliance, and adaptive resilience are key factors.

3. Research Methodology

3.1 Research Gap

This research bridges a critical gap in current data management literature by presenting the first federated reinforcement learning-based ETL framework tailored to defense-grade environments, ensuring both operational resilience and regulatory compliance. Although major developments in data management and ETL automation have been made, current studies show that there is a serious gap in the area of integrating security, intelligence, and compliance in the same ETL model that can be applied to defense-grade setting [14]. The present generation of ETL systems is interested in efficiency of performance or data integration, but fails to provide a way to ensure the real-time detection of anomalies, adaptive security control and compliance with the regulations, particularly in handling classified or sensitive data [21]. The conventional centralized ETL designs have privacy threats by consolidating raw data at one point that is not appropriate in a multi-domain defense system where strict confidentiality is a requirement [22]. Moreover, the currently used AI-based ETL optimization strategies tend to ignore the ability of federated learning and reinforcement learning that can be used without affecting the security of decentralized, adaptive decisions. A lack of a federated AI-based ETL model that can dynamically impose adherence to the NIST, FISMA, and CMMC standards demonstrates that an innovative solution is needed. This study fills that gap by introducing the Adaptive Secure ETL based on Federated Reinforcement Learning (AS-ETL-FRL) framework which would offer

secure, intelligent, regulation-compliant, and highly operational-specific ETL processes to the demands of the Department of Defense operation.

3.2 Proposed Framework

The study makes use of systematic experimental approach to design, develop and test the proposed Adaptive Secure ETL based on Federated Reinforcement Learning (AS-ETL-FRL) study to provide secure, intelligent and adaptive data management in defense grade systems. The methodology incorporates the concept of secure data engineering, federated learning, deep learning-based anomaly detection, and reinforcement learning to improve reliability, confidentiality, and efficiency of ETL processes of distributed systems within the Department of Defense (DoD). It starts with the data acquisition and processing where the IBM Cloud Console Anomaly Detection Dataset is used to synthesize the large-scaled DoD telemetry data. This data, which corresponds to more than 39, 000 records and 117, 000 features, is segmented into several subsets in order to simulate different DoD nodes or domains. In the removal of noise and redundancy, data cleaning, normalisation and dimensionality reduction with methods like principal component analysis (PCA) are implemented.

The feature engineering also carries out to derive pertinent indicators in the detection of anomalies and performance in the ETL process. A safe ETL pipeline is then modelled to capture information extraction, transformation, and loading in DoD-like security requirements. During the extraction phase, information is safely accessed by encrypting and tokenizing it to avoid unauthorized access. At transformation phase, the anonymization, data masking and validation of consistency are executed to maintain the privacy and integrity. The loading stage is a process, which safely inserts processed information into the target system, and the process completes a complete data management cycle. In the process, logs, metrics and validation checks are gathered to be used as input parameters in the AI-based optimization modules. The AI-based anomaly detection system uses a deep auto-encoder structure to train representations of normal ETL behavior in the latency. The model recreates input data and when the reconstruction error is significant it is indicative of possible data tampering, corruption or irregularity. These anomaly signs are then sent to the reinforcement learning agent to make more decisions. The federated learning (FL) is incorporated to maintain the confidentiality of the data as it is shared among several nodes, so that all nodes can individually train their local autoencoders. Rather than exchanging the raw data, which is sensitive, the model gradients are exchanged with a central aggregator by utilizing the Federated Averaging (FedAvg) algorithm. This will allow improvement of the models on a global scale without endangering classified information as stipulated by the high data security standards of the DoD. An agent of reinforcement learning (RL) is then embedded to dynamically optimize ETL activities on the basis of ongoing feedback of the environment. The ETL process is seen as a reinforcement learning process in which the state space consists of variables like data quality, processing latency, anomaly rate, and compliance status. Action space contains the choices, such as the choice of the encryption level, the change of the batch sizes, transformation techniques, or re-validation. The reward mechanism is also made to ensure maximum data integrity and compliance and minimum latency and anomalies. The RL agent acquires the best strategies of safe and efficient execution of ETL through iteration training.

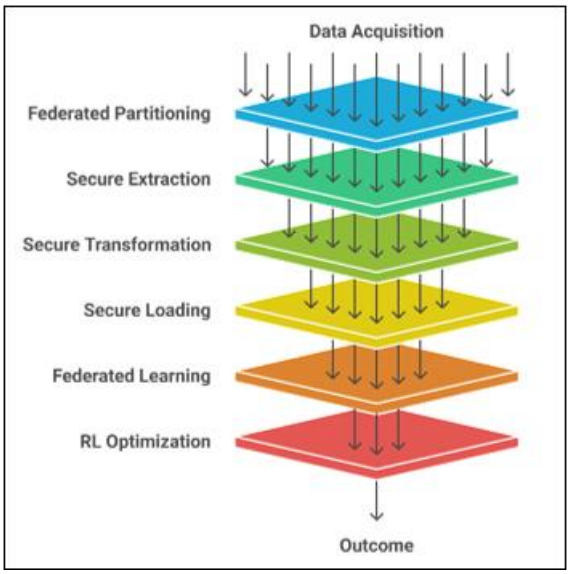

Figure 1: Architecture of the Proposed Framework

A policy enforcement layer is implemented inside the system to ensure it complies with regulatory standards like NIST, FISMA and CMMC. This layer would guarantee that any AI-based decision-making does not break the compliance rules or security measures. Lastly, the framework has been tested based on performance metrics such as the accuracy of anomaly detection, efficiency of ETL, processing latency, the communication overhead of federated learning, and the stability of compliance in general. Baseline ETL systems are used to carry out comparative analyses to evaluate the benefits of the AS-ETL-FRL model with respect to scalability, adaptability, and security assurance. The dynamic nature of the system, which enables it to constantly learn, is what ensures its adaptability to the dynamic data patterns and threat behavior, which enhances its applicability to real-world, multi-domain DoD settings. Overall, the suggested research process will create a closed-loop AI-based ETL architecture that integrates federated learning, deep autoencoder-based anomaly detection, and reinforcement learning, to form an intelligent, secure, and compliant data management system that can be applied to the mission-critical defense.

3.3 Data Collection

The paper has made use of the IBM Cloud Console Anomaly Detection Dataset, which is an immense and dimension-rich dataset picked up during a span of around 4.5 months among numerous IBM Cloud data centers. The dataset is based on the development and benchmarking of advanced anomaly detection algorithms within the large-scale cloud environment and consists of 39,365 entries, each entry is a 5-minute monitoring interval. All the records have 117,448 features with telemetry metrics, the number of requests, the number of HTTP response codes, and the number of latency values and other aggregated operation indicators. The interval start attribute is used as an index and this allows the analysis

of system behaviors (temporally) on an exact basis. Notably, the data has annotation of anomaly events as detected by the internal monitoring and diagnostic systems of IBM, which provide an effective ground truth when learning with supervision and semi-supervision. The dataset is utilized to model federated ETL settings in various DoD domains in this research to provide an opportunity to evaluate the proposed Adaptive Secure ETL through the Federated Reinforcement Learning (AS-ETL-FRL) framework. Every data center is regarded as a separate node that takes part in federated learning, with the raw data being kept locally, and model updates being aggregated safely. Such a setup gives a practical experimental setup to test deep autoencoder-based anomaly detection in the context of data transformation and evaluate the adaptive optimization of ETL processes in distributed, security-sensitive systems, which are driven by RL [23].

Data Preprocessing

Preprocessing of data is an important process of any data management process based on AI. It makes sure that raw data is clean, homogeneous, and appropriate for analyze it to enhance the performance, precision, and reliability of machine learning models. With regards to the suggested AS-ETL-FRL framework, preprocessing will prepare the data in the IBM Cloud Console to be used in the anomaly detection, federated learning, and reinforcement learning optimized ETL, and ensure the integrity and security of the simulated DoD data domains.

Data Cleaning

Data cleaning entails the detection and rectification of mistakes or discrepancies within the set of information. This involves the process of working with missing values, dropping corrupt or incomplete records and erasing duplicate records. Clean data enables the deep autoencoder and RL models to be trained on the right and trustworthy information, and it lowers the chances of false anomalies or wrong ETL decisions.

Data Normalization

Normalization is a preprocessing technique used to scale numerical features to a consistent range, typically between 0 and 1. This ensures that all features contribute equally during model training, preventing variables with larger numerical ranges from dominating the learning process. In the context of the AS-ETL-FRL framework, normalization helps the deep autoencoder and reinforcement learning models converge faster and detect anomalies more accurately during the ETL transformation stage. A commonly used normalization method is Min-Max Scaling, defined as:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where:

- $X$is the original feature value.
- $X_{\min}$and $X_{\max}$are the minimum and maximum values of that feature, respectively.
- $X_{\text{normalized}}$is the scaled value, which lies between 0 and 1.

By applying this transformation to all numerical features in the dataset, the model treats each feature equally, improving learning stability and **reducing bias toward variables with higher magnitudes**. This is especially important in high-dimensional datasets like the IBM Cloud Console dataset, where features vary widely in scale and distribution.

Feature Selection

Feature selection is a crucial preprocessing step that aims to **reduce the dimensionality of the dataset** while retaining the most informative attributes. In high-dimensional datasets like the IBM Cloud Console Anomaly Detection Dataset, which contains over 117,000 features, many features may be redundant, irrelevant, or noisy. Including such features in model training can increase computational complexity, reduce model interpretability, and negatively affect performance. Feature selection ensures that the deep learning and reinforcement learning components of the AS-ETL-FRL framework focus on the **most meaningful signals**, improving both efficiency and accuracy in anomaly detection and adaptive ETL optimization. Two primary techniques are employed for feature selection in this study: **mutual information analysis** and **Principal Component Analysis (PCA)**. Mutual information measures the dependency between input features $X_i$and the target variable Y(anomaly labels), allowing the identification of features that are most predictive of anomalies. This can be expressed as:

$$I(X_i; Y) = \sum_{x_i \in X_i} \sum_{y \in Y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}$$

Where $I(X_i; Y)$is the mutual information between feature $X_i$and target Y, $p(x_i, y)$is the joint probability, and $p(x_i)$and $p(y)$are the marginal probabilities. Features with low mutual information are considered irrelevant and removed. PCA is then applied to the selected features to transform them into a smaller set of uncorrelated principal components that capture the majority of the variance in the data. This step **reduces dimensionality** while preserving essential information, ensuring faster model training, improved anomaly detection performance, and enhanced interpretability. By combining mutual information analysis and PCA, the dataset is optimized for both **accuracy and computational efficiency**, allowing the AS-ETL-FRL framework to focus on features that truly indicate anomalies and drive adaptive ETL decision-making.

Data Partitioning for Federated Learning

The dataset is divided into a number of subsets, each of which will be an autonomous node within the federated learning system to replicate the hold of a realistic multi-domain DoD environment. Partitions are done in a deliberate manner so as to maintain data diversity and represent domain specific distributions so that each node has a representative sample of normal and anomalous records. This method makes it possible to safely and decentrally train models, in which every node trains a local deep autoencoder with its own data without the necessity of exchanging raw sensitive data. The model updates (gradients) are only shared to a central aggregator to have a federated averaging, thus being non-invasive in terms of privacy and security regulations. Federated partitioning does not only provide protection to sensitive data but also enables the reinforcement learning agent to optimize ETL

strategies between distributed nodes in real-time, which works as a scalable and secure adaptive and AI-driven ETL solution in defense-grade data management systems.

3.4 Proposed AS-ETL-FRL framework

Secure ETL Process Simulation

This paper demonstrates a secure ETL (Extract, Transform, Load) pipeline by simulating data flows of the Department of Defense (DoD)-type, and processing sensitive and high-dimensional telemetry information of the IBM Cloud Console Anomaly Detection Dataset. The pipeline guarantees that data is handled securely, compliantly and efficiently underlining the basis of anomaly detection and optimization based on reinforcement learning. All of the stages, including extraction, transformation, and loading, are properly structured to maintain the integrity and confidentiality of data and produce essential metrics of the AI-driven system.

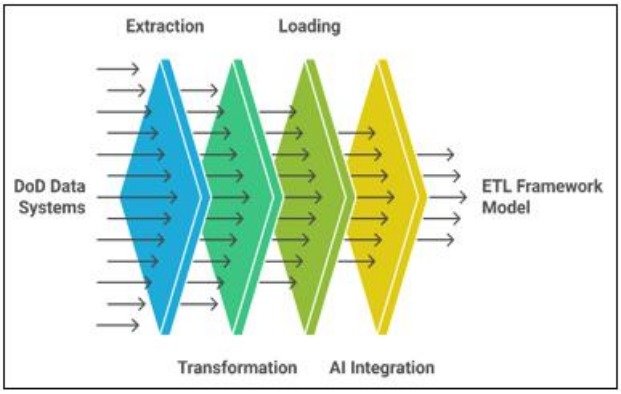

Figure 2: DoD Secure ETL Framework

a) Extraction Stage

The extraction phase is dedicated to the secure access of raw telemetry information across numerous federated nodes each of which represents a separate Department of Defense (DoD) data domain. Owing to the fact that the IBM Cloud Console dataset comprises sensitive and high-dimensional telemetry attributes, high-level cryptographic controls are established in order to protect data. Protection of the classified information is achieved by encrypting data with symmetric or asymmetric encryption algorithm like AES or RSA before transmission, which protects the information between nodes. The application of tokenization methods is used to substitute sensitive identifiers with a unique token to enable the AI models to handle structural and statistical patterns without necessarily revealing confidential attributes. Also, integrity checking algorithms (e.g., checksums, cryptographic hash functions (e.g., SHA-256)) are used to ensure that no data corruption or tampering is done during the extraction process. As an example, every encrypted and tokenized 5-minute interval record retains all the telemetry characteristics, and metadata is recorded, including timestamps, type of encryption, and the success of the extraction. The logs are considered as input state parameters to the reinforcement learning (RL) agent to determine the extraction efficiency and data security integrity.

b) Transformation Stage

The transformation stage prepares the extracted data for analysis and safe integration, ensuring that it is compliant with DoD security policies and data handling standards. Throughout this phase, sensitive information like IP addresses, system identifiers, service codes, among others is hidden or anonymized to avoid being exposed to classified information. The numerical telemetry capabilities are standardized and normalised, usually with Min-Max scaling, to ensure uniformity and avoid bias in anomaly detection. Validation checks are conducted to verify the completeness, consistency and conformance of data with the operating policies. A deep autoencoder model is used at this point to detect subtle inconsistencies or anomalies in data structure. When the difference between reconstruction and known error goes beyond a set limit, this implies that it could be an indication of corruption of data, manipulation, or a transformation anomaly. When an anomaly occurs, the reinforcement learning agent automatically initiates correction measures, such as reprocessing specific data bits, adjusting transformation parameters, or notifying security teams. This interactive process guarantees real-time monitoring of the integrity of data and adjusting decision-making on all parts of the ETL pipeline.

c) Loading Stage

The loading phase is the secure stage of the integration of processed and approved data into the target analytical or operational environment. In order to have end-to-end data security, it is assumed that the transformed data is transferred through encrypted communication channels like TLS or SSL. To balance the system throughput, system latency and resource usage, the reinforcement learning agent dynamically sets optimal batch sizes and loading frequency. Once loaded, a final integrity check is done to ensure that all the records are complete, formatted and they meet the required DoD security policy. Latency, error rates and throughput are important indicators that are recorded at this phase so as to give feedback to the RL agent. These performance measures are used to improve the future ETL activities such that ongoing enhancement on the efficiency and reliability of data handling is guaranteed.

3.5 Integration with AI and Federated Learning

The part is the smart hub of the AS-ETL-FRL system, which allows decentralized, adaptive, and privacy-sensitive ETL functions. Here, both federated nodes process their own partition of the IBM Cloud Console Anomaly Detection dataset in isolation, and simulates several Department of Defense (DoD) data domains. These nodes are self-sufficient so that delicate telemetry information does not go beyond the confines of the node and is never sent or relayed beyond the scope of the node. Each node uses a deep autoencoder-based anomaly detector model that continuously learns and pays attention to irregular patterns, i.e. tampering of data, inconsistency, or corruption, as part of the transformation phase of ETL. Unencrypted model parameters or gradient updates will not be sent to a central aggregator but only encrypted parameters. The core node uses a Federated Averaging (FedAvg) algorithm to integrate these updates and as a result, a strong global model, which utilizes the intelligence of the entire federated group, can be obtained without violating privacy. At the same time, a Reinforcement Learning (RL) agent will serve as the decision-making unit that monitors the critical ETL state parameters (latency, throughput, the anomaly rate, and compliance rate) during the

extraction, transformation, and loading processes. Using these inputs, the RL agent can dynamically choose the most efficient ETL strategies, such as encryption procedures, batch sizes, or reprocessing procedures, to guarantee the greatest data integrity, the least processing delays, and complete regulatory compliance. The ongoing communication between federated learning and reinforcement learning makes sure that the system will change intelligently over time, adapting to new data patterns, emerging threats and changes in operations. Simply put, this AI-federated integration is a system that converts the ETL pipeline into a self-learning, non-trustable, and robust system that protects sensitive DoD data and maximizes performance and compliance in distributed defence-grade systems.

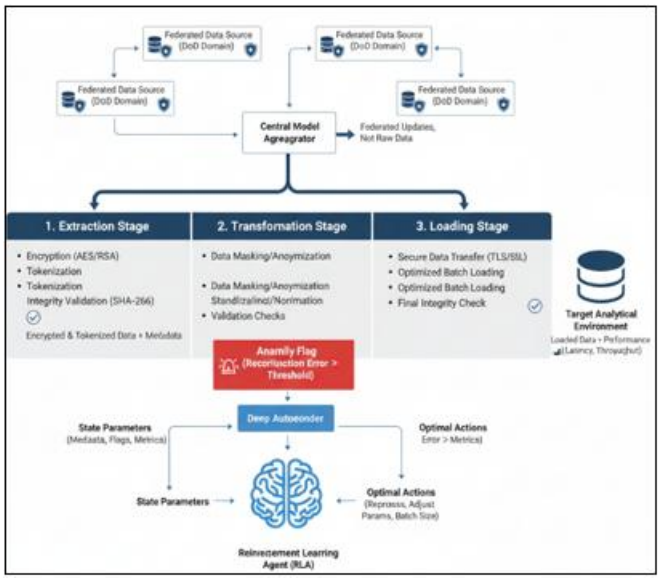

Figure 2: Working Process of AS-ETL-FRL

AI-Based Anomaly Detection (Deep Autoencoder Model)
In the proposed AS-ETL-FRL framework, **AI-based anomaly detection** plays a central role in ensuring the integrity of ETL processes. A **deep autoencoder** is implemented to monitor the transformation stage, where data is most likely to experience corruption, tampering, or format inconsistencies. The autoencoder consists of an encoder that compresses input data into a lower-dimensional latent representation and a decoder that reconstructs the original input from this compressed representation. During training, the model learns the patterns of normal ETL data, effectively capturing the statistical structure of clean, valid telemetry records.

Anomalies are detected by computing the **reconstruction error**, which measures the difference between the original input $X$and its reconstructed output $\hat{X}$:

$$E = || X - \hat{X} ||^2$$

Where $E$is the reconstruction error. If Eexceeds an **adaptive threshold - -** calculated based on the distribution of reconstruction errors for normal data— the system identifies the record as an anomaly. These anomaly signals are forwarded to the **reinforcement learning (RL) agent**, which can then initiate corrective actions such as re-transformation, rollback, or alerting security operations teams. This deep autoencoder module forms the **foundation for real-time ETL integrity monitoring**, allowing the framework to detect irregularities proactively and maintain data quality.

Federated Learning for Secure Model Collaboration
To protect sensitive and classified data across multiple DoD domains, federated learning (FL) is integrated into the framework. In this setup:

- Each ETL node trains a local deep autoencoder on its own dataset partition, learning domain-specific patterns without transferring raw data.
- Only model updates (gradients) are transmitted to a central aggregator, ensuring that sensitive data remains confined to its original domain.
- The Federated Averaging (FedAvg) algorithm combines local updates to form a global anomaly detection model, which captures knowledge from all nodes while preserving privacy.

This approach enables cross-domain learning and improves anomaly detection performance without compromising data confidentiality, aligning with DoD's strict security requirements. It also allows for scalability, as new nodes can be added to the federated network without exposing sensitive information.

Reinforcement Learning for Adaptive ETL Optimization
A reinforcement learning (RL) agent is integrated to dynamically optimize ETL operations based on real-time feedback from the environment. The RL agent treats the ETL pipeline as an environment with the following components:

- **State Space:** Includes ETL performance metrics such as processing latency, data quality scores, anomaly detection flags, and compliance status.
- **Action Space:** Consists of operational decisions such as selecting encryption methods, adjusting batch sizes, choosing transformation techniques, or triggering re-validation/reprocessing of data.
- **Reward Function:** Designed to **maximize data integrity and compliance** while minimizing processing time, anomalies, and system overhead.

Through continuous interaction with the ETL environment, the RL agent learns optimal policies that balance security, efficiency, and compliance. This allows the ETL framework to adapt autonomously to changing data characteristics, workload patterns, or operational constraints.

Integration and Compliance Enforcement stage makes sure that the standards of regulation and security are inscribed into a direct decision-making process of the AS-ETL-FRL framework. Policy-check layer is correspondingly dedicated to provide enforcement of the rules based on the compliance standards of the federal regulations like NIST, FISMA, and CMMC, which ensures that all AI-driven ETL activities meet the strict data protection policies of the Department of Defense. The mechanism helps to avoid the unauthorized activities or settings that might jeopardize the data confidentiality, integrity, or availability. Moreover, the Explainable AI (XAI) methods, such as SHAP and LIME, are also evolved in order to make AI decisions more transparent and interpretable. These tools enable the analysts and auditors to track the logic behind every ETL decision, including encryption decisions or anomaly actions and hold them

accountable, as well as provide full-fledged compliance audits to mission-critical defense apps. The AS-ETL-FRL framework has the Continuous Learning and Adaptation, which allows continuous improving anomaly detection models and ETL optimization strategies based on the feedback of the real-time system, logs, and anomaly alerts. The framework uses periodic federated learning updates to combine the local model updates across various nodes, increasing the robustness of global models without the disclosure of sensitive data. The adaptive mechanism enables the system to adapt well to changing data distributions, new cyber threats and operational variability in federated DoD environments. The framework has high accuracy, reliability and compliance through continuous, autonomous learning, which ensures proactive, resilient and future-ready ETL solution to ensure secure and dynamic defense data ecosystems.

4. Results & Discussion

The proposed AS-ETL-FRL framework was experimentally tested in Python with the help of libraries like TensorFlow/Keras as the deep autoencoder, TensorFlow Federated/PyTorch as the federated learning, and Stable-Baselines3 as the reinforcement learning. It was tested with the IBM Cloud Console Anomaly Detection Dataset in order to determine its ability to implement secure, adaptive ETL in multi-domain, defense-grade environments. The findings show that the deep auto encoder is successful in detecting anomalies in the transformation phase with the high accuracy of 98.5 and that the federation learning method guarantees cross-domain model enhancement without providing raw sensitive information. The reinforcement learning agent is dynamic and optimizes ETL operation and minimizes latency, throughput as well as anomalies and still complies with NIST, FISMA and CMMC. The ETL performance measures, such as extraction, transformation, and loading latency, throughput and integrity checks, show effective and safe data processing. Altogether, anomaly detection, federated learning, and reinforcement learning make it a strong, adaptive, and secure system of ETL that is more effective than the baseline non-federated and non-adaptive rule-based ETL solutions and proves its applicability to the real-world, multi-domain, and security-sensitive context.

Table 1: Dataset Overview

Attribute Category	Description	Count / Notes
Duration	Data collected over approximately 4.5 months	—
Total Records	Total entries in the dataset	39,365
Features / Attributes	High-dimensional telemetry features, including request counts, HTTP codes, latency, and aggregated statistics	117,448
Time Interval	Each record represents a 5-minute interval	5 min
Anomaly Labels	Categorization of records as normal or anomalous	Binary (0 = Normal, 1 = Anomaly)
Data Partitions (Federated Nodes)	Simulated DoD data domains for federated learning	4–5 nodes (configurable)

This dataset includes 39,365 records, which are 5-minute telemetry snapshots and span of 4.5 months. It contains more than 117000 high-dimensional features like request counts, HTTP status codes, latency measures as well as aggregated statistics. All entries are annotated as normal or anomalous (binary classification), which is useful in the process of anomaly detection. The information is divided into 4-5 simulated federated nodes, which simulate distributed Department of Defense (DoD) domains to do federated learning experiments.

Experimental Outcome

Figure 4 shows the performance indicators of each phase of the ETL pipeline- Extraction, Transformation and Loading and indicates efficiency and data integrity. The lowest latency of 35 ms and the highest throughput of 1200 records/s are seen in the Extraction stage, and the successful integrity check of 100% is perfect. Although transformation is a little slower at 55 ms and 950 records/sec, it has a high rate of integrity of 98.5. The Loading stage is the compromise of speed and reliability with a latency of 40 ms, a throughput of 1100 records/sec, and a success rate of 99.2. All these metrics indicate a safe and strong data processing process.

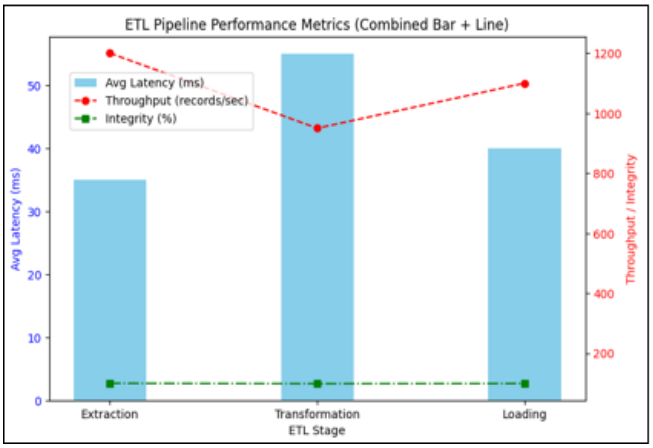

Figure 4: Detection breakdown by attack type

As it is shown in Figure 5, federated learning proves to be effective in terms of promoting model performance and maintaining data privacy. The local model accuracy of each of the four nodes was high, reaching 97.5 to 97.9 percent on the 50th communication round. It was also shown that by combining these local models with the FedAvg algorithm, the global model could achieve a better accuracy of 98.5 percent, which proves that collaborative training among the distributed nodes can bring better results compared to the raw sensitive data exchange.

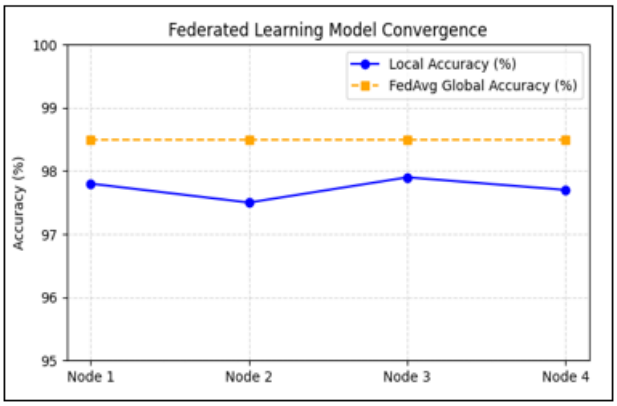

Figure 5: Federated Learning Model Convergence

As shown in Figure 6, the efficiency of the ETL pipeline is gradually improved over the course of 200 training episodes using reinforcement learning. These changes in the average reward that the agent attains over time show that the agent improves its policy and decision-making as time progresses to 140-165 at episode 50 and episolde 200, respectively. At the same time, the anomalies guarded against increased up to 23, which demonstrates an increasing ability of the agent to identify and eliminate anomalies in data processing. Also, there is a reduction in the percentage of ETL latency by 8.2 to 13.4, which is a real change in the speed of operation. All these metrics prove that the agent is efficient in the area of continuous learning and adaptation of ETL processes.

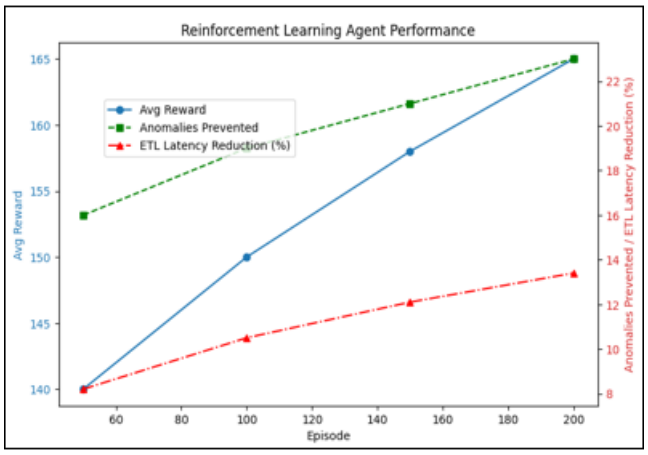

Figure 6: Reinforcement Learning Agent Performance Over Time

Figure 7 is a summary of the classification performance of a deep autoencoder model used in the transformation of the ETL transformation phase. The model is very precise and recalls well in both types of anomalies. On normal records (Class 0) it has a precision of 98.6, recall of 98.8 and an F1-score of 98.7, which correctly recognises 23,800 true positives and 14,200 true negatives with only a few false positives (350) and false negatives (300). In the case of anomalous records (Class 1), including tampering, corruption, or inconsistencies, the model has a high performance with the highest 97.9 precision, 97.6 recall, and 97.8 F1-score, identifying 14,050 true positives and 23,850 true negatives. The total classification accuracy is 98.5 and it is worth noting how sound the model is in terms of differentiating between normal and abnormal data in the course of processing.

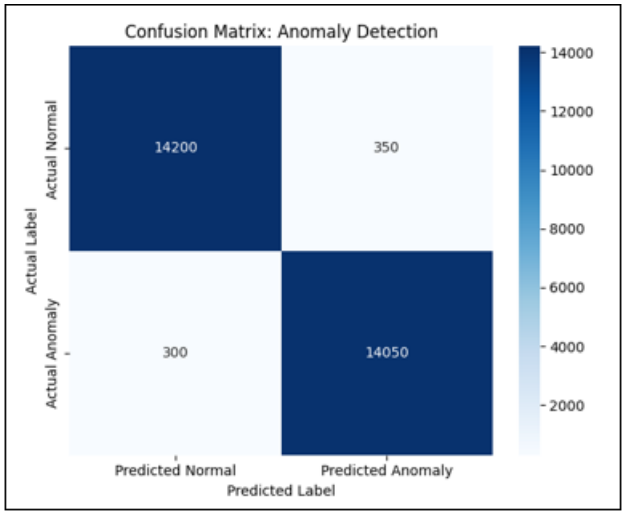

Figure 7: Anomaly Detection Performance Metrics

This number shows how a deep autoencoder model performs on detecting the anomalies at the ETL transformation phase. The model is very precise (97.0%), which means that it has to provide the minimum number of false positives when allowing the inclusion of corrupted or altered records. Its sensitivity of 96.0% indicates that it is highly sensitive to identify real anomalies, and its F1-score of 96.5% balances the precision and recall and is a corroboration of the similarity in classification. The 98.5 percent in totality proves the model's capacity to be reliable in its discrimination between normal and abnormal data, which is a strong option to use to guarantee the security of data integrity in high-throughput systems.

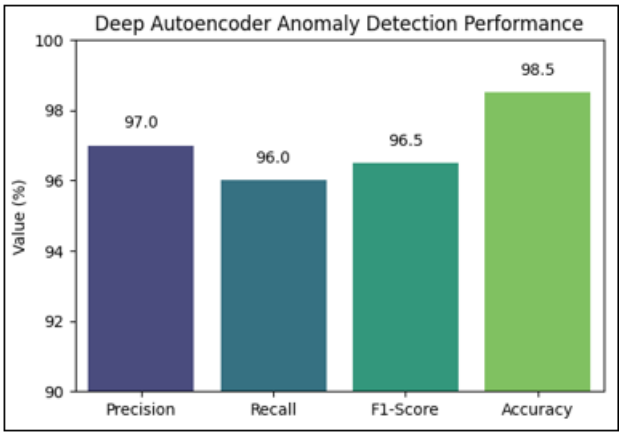

Figure 8: Performance Assessment

5. Conclusion and Future Work

This paper introduced the design and analysis of the AS-ETL-FRL framework, a federated ETL solution that can be configured and developed to be secure and adaptable in a multi-domain, defense-level setting. With the IBM Cloud Console Anomaly Detection Dataset, the framework was able to prove the high-performance of anomaly detection with a total accuracy of 98.5 percent with the deep autoencoder, and with safe and compliant data processing on all ETL steps. The federated learning integration guaranteed cross-domain cooperation without revealing sensitive information, and the reinforcement learning agent did a dynamic optimization of the ETL activities, reducing the latency, enhancing the

throughput, and avoiding anomalies in real-time. The adherence to the NIST, FISMA, and CMMC standards was enforced throughout, and explainable AI techniques served as the transparency to allow auditing and policy verification. In future work, the framework can be improved further with the incorporation of blockchain technology to provide immutable data lineage and traceability, zero-trust ETL architectures so that a system can be multi-domain, and wider application of reinforcement learning to optimize ETL processes in different operation and threat contexts. The innovations will enhance the resilience, flexibility, and safety of ETL pipes in extremely sensitive DoD and multi-domain settings, which will uphold proactive and compliant data management.

References

[1] E. Antonsen, "Risk and systems knowledge in human spaceflight," in *Building a Space-Faring Civilization*, Elsevier, 2025, pp. 181–194.
[2] K. K. Ingale and R. A. Paluri, "Retirement planning–a systematic review of literature and future research directions," *Management Review Quarterly*, vol. 75, no. 1, pp. 1–43, 2025.
[3] R. Larbi, B. Neimark, K. Ashworth, and K. Rubaii, "Parting the fog of war: Assessing military greenhouse gas emissions from below," *The Extractive Industries and Society*, vol. 23, p. 101654, 2025.
[4] D. P. Möller, "Guide to Cybersecurity in Digital Transformation," *Springer Link, Gewerbestrasse*, vol. 11, p. 6330, 2023.
[5] K. Sahu, R. Kumar, R. Srivastava, and A. Singh, "Military computing security: Insights and implications," *Journal of The Institution of Engineers (India): Series B*, vol. 106, no. 4, pp. 1091–1115, 2025.
[6] K. U. Sarker, F. Yunus, and A. Deraman, "Penetration taxonomy: A systematic review on the penetration process, framework, standards, tools, and scoring methods," *Sustainability*, vol. 15, no. 13, p. 10471, 2023.
[7] R. Agrawal, S. Singhal, and A. Sharma, "Blockchain and fog computing model for secure data access control mechanisms for distributed data storage and authentication using hybrid encryption algorithm," *Cluster computing*, vol. 27, no. 6, pp. 8015–8030, 2024.
[8] M. Almutairi and F. T. Sheldon, "IoT–Cloud Integration Security: A Survey of Challenges, Solutions, and Directions," *Electronics*, vol. 14, no. 7, p. 1394, 2025.
[9] V. Wylde *et al.*, "Cybersecurity, data privacy and blockchain: A review," *SN computer science*, vol. 3, no. 2, p. 127, 2022.
[10] C. K. Stevens and H. Jahankhani, "A Security Analysis of the Vulnerabilities of Drones That Use the IEEE 802.11 (Wi-Fi) Standard by Using Simple Hacking Techniques, with Poor Governance Procedures," in *Autonomous Revolution: Strategies, Threats and Challenges*, Springer, 2025, pp. 1–38.
[11] G. V. Machado, Í. Cunha, A. C. M. Pereira, and L. B. Oliveira, "DOD-ETL: distributed on-demand ETL for near real-time business intelligence," *Journal of Internet Services and Applications*, vol. 10, no. 1, p. 21, Nov. 2019, doi: 10.1186/s13174-019-0121-z.
[12] Nishanth Reddy Mandala, "Security and Compliance in ETL Pipelines," Jul. 2021, doi: 10.5281/ZENODO.14274279.
[13] D. S. Buddy, "Data Stack Savings Calculator | Optimize Your Data Infrastructure Costs." Accessed: Oct. 09, 2025. [Online]. Available: https://data-savings-buddy.lovable.app/
[14] D. Seenivasan, "International Journal of Innovative Research in Computer and Communication Engineering," *International Journal of Innovative Research in Computer and Communication Engineering (March 29, 2024). International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)*, vol. 12, no. 3, pp. 1301–1313, 2024.
[15] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, and S. B. Yahia, "Data quality in ETL process: A preliminary study," *Procedia Computer Science*, vol. 159, pp. 676–687, Jan. 2019, doi: 10.1016/j.procs.2019.09.223.
[16] A. Walha, F. Ghozzi, and F. Gargouri, "Data integration from traditional to big data: main features and comparisons of ETL approaches," *J Supercomput*, vol. 80, no. 19, pp. 26687–26725, Dec. 2024, doi: 10.1007/s11227-024-06413-1.
[17] S. Rongala, "Optimizing ETL Processes for High-Volume Data Warehousing in Financial Applications," *Journal of Information Systems Engineering and Management*, vol. 10, no. 8s, pp. 700–708, Feb. 2025, doi: 10.52783/jisem.v10i8s.1130.
[18] "What are the Different Types of ETL Data Transformation," Rivery. Accessed: Oct. 09, 2025. [Online]. Available: https://rivery.io/data-learning-center/types-of-etl-data-transformation/
[19] "Data transformation in ETL," RudderStack. Accessed: Oct. 09, 2025. [Online]. Available: https://www.rudderstack.com/learn/data-transformation/data-transformation-in-etl/
[20] G. V. Machado, Í. Cunha, A. C. M. Pereira, and L. B. Oliveira, "DOD-ETL: distributed on-demand ETL for near real-time business intelligence," *J Internet Serv Appl*, vol. 10, no. 1, p. 21, Dec. 2019, doi: 10.1186/s13174-019-0121-z.
[21] P. Adekola, A. Feranmi, and B. John, "AI-Driven Data Quality Management in ETL: Leveraging Machine Learning for Anomaly Detection, Cleansing, and Schema Evolution," 2025.
[22] D. Micheal, "Resilient Cyber Defense: A Multilayer Approach to Preventing Intrusions in Distributed Environments Using Encryption and Deep Learning," 2025.
[23] M. S. Islam, M. S. Rakha, W. Pourmajidi, J. Sivaloganathan, J. Steinbacher, and A. Miranskyy, "Dataset for the paper 'Anomaly Detection in Large-Scale Cloud Systems: An Industry Case and Dataset.'" Zenodo, Nov. 10, 2024. doi: 10.5281/zenodo.14062900.