International Journal of AI, BigData, Computational and Management Studies
Noble Scholar Research Group | Volume 6, Issue 4, PP 142-153, 2025
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I4P116

Original Article

AI-Powered ETL Automation for Compliant Data Migration

Anusha Joodala
Independent Researcher, USA.

Received On: 22/09/2025 ***Revised On: 25/10/2025*** ***Accepted On: 05/11/2025*** ***Published On: 16/11/2025***

Abstract: *Data governance poses an increasingly complex challenge for organizations that operate on multiple heterogeneous systems, including various clouds, to ensure compliance with numerous regulations, like GDPR or HIPAA. Whereas traditional approaches to ETL (Extract–Transform–Load) have problems with manual schema mapping, weak transformation logic, and traceability, AutoETL-C provides a solution to these problems with a suggested depreciation of ETL, as it incorporates AI to deliver an automated and policy-compliant ETL. The AI to automate ETL creates a more seamless experience for users, combining (i* IJAIBDCMS-V6I4P116 *neural schema matching via domain-adapted language models, (ii) hybrid transformation through rules and learning, automated constraint discovery, and invasive trace law, (iii) data quality monitoring with continuous anomaly detection, (iv) and full-coverage lineage capture, including all governance and compliance limited controls. AutoETL-C was tested on three migration scenarios—core-to-lakehouse for banking, EHR modernization for healthcare, and ERP-to-warehouse for retail—public and synthetic datasets as well as anonymized enterprise data were used. AutoETL-C outperformed its manually configured ETL, template-based ETL and ML-assisted mapping, setting a new industry standard while proving compliance with Shelby Control Model documentation, HIPAA Security Rule controls, and data maps. Overall, results suggest that migration risk and time-to-value can be substantially improved with augmented, privacy-aware data engineering.*

Keywords: *ETL, Data Migration, GDPR, HIPAA, AI, Schema Matching, Data Lineage, DataOps, Compliance Automation.*

1. Introduction

Data migrations are now part of major enterprises' digital transformation plans. Mergers, system upgrades, adopting cloud infrastructures, and lakehouse consolidations are some of the widespread transformations. However, such data migrations often encounter challenges, including data heterogeneity, legacy semantics, and compliance and regulation issues. Issues such as inconsistent data formats, complex-embedded business logic, and old-new system interconnection issues often lead to data migration problems in heterogeneous systems. Addressing these issues is further complicated by compliance and regulation issues, especially related to data privacy and protection. Additionally, the shift to modern data architectures, including lakehouses and cloud environments, highlights the need for reliable and efficient data migration as organizations move away from legacy on-premise systems. The unbelievable potential of big data drives organizations to undertake such complex transformations. However, the difficulties in managing such complex transformations lead to extended timelines, budget overruns, and compliance and regulation issues. To reduce risks and improve data migration workflows, organizations have increasingly begun to adopt augmented data integration and data fabric patterns to reduce manual effort and improve governance using AI, machine learning, and automation. In the coming period, reliance on people-led integration activities will markedly reduce, as technology will perform those activities, as stated in industry surveys and market reports. According to the underlying principles of the research conducted by Gartner, the augmented data management and DataOps frameworks will help achieve quality indicators of timeliness, precision, and scalability in data migration while maintaining data lineage and compliance. AI technology will certainly help perform integration activities by automating mundane tasks, improving data mapping and ensuring compliance across the migration pipeline. Overall, AI facilitates secure and flexible data migration processes in response to the needs of current businesses.

But, the need for operational efficiency is not the only thing that is driving these changes. The GDPR and HIPAA laws make it necessary to justify how data moves to ensure it meets legal requirements. For example, GDPR Article 20 requires organizations to facilitate data portability in a "structured, commonly used, and machine-readable format." This means organizations must ensure that data is transferable between systems and keep it private. Also, personal data, especially in EU countries, requires the use of Standard Contractual Clauses (SCCs) for cross-border transfers, complicating data migration efforts. The healthcare industry is also subject to the same rules as HIPAA, the U.S. Security Standard policies that describe data integrity, and reasonable assurance policies which include multifactor

authentication (MFA) along with encryption, and oversight of vendors. This type of compliance can also complic the design of the migration solution as well as the legal auditing and documentation of the entire migration path. Thus, organizations have to deal with the legal issues alongside the technical issues to make sure the transformations are legal.

While AI assisted ETL (Extract-Transform-Load) tools are available, research on ETL focuses largely on the efficiencies gained in transforming and moving data, and detecting anomalies. Although these studies are valuable, many of the compliance controls are ignored. Documenting the legal compliance issues of data migrations is complex, especially issues of proving portability, cross-border transfers, privacy risks, and auditable lineage from extraction to post-load validation. This is especially problematic for organizations in the banking, healthcare, and retail industries, where the cost of non-compliance is devastating. Most of the research in this space applies AI and ML to portions of the data transformations and anomaly detection within the ETL process. While these techniques have been proven to work, there are no compliance controls integrated within the workflow, making the technique ineffective in practice.

This paper attempts to reduce the research gap by presenting AutoETL-C, The first ETL automation framework which migrates data and embedding the automated compliance check systems through every stage. The major contributions of AutoETL-C are:

- Schema and Semantic Mapping: AutoETL-C uses domain-adapted language models for learning schema mappings and discovering the semantics relations between source and target systems. It also implements constraint inference so that the mappings comply with business rules and are legally sound.
- Policy-Aware Transformations: The framework performs compliance-embedded transformations that integrate the principles of data minimization and purpose limitation within the design of the system during seamless data migration. This design encompasses the principles to ensure only required data is transferred during migration, which lessens the risk of infringement of privacy, and complies with regulations such as the GDPR.
- Streaming Data Quality & Anomaly Detection: The AutoETL-C framework undergoes seamless migration while implementing constant monitoring for data quality, as well as embedding real-time anomaly detection with the data quality assessment to ensure its operational consistency with the other systems. This allows the system to automatically retrain its mappings and transformations during migration, ensuring data consistency and integrity.
- End-to-End Lineage & Compliance Artifacts: As data flows from a source to a target system, AutoETL-C builds visible, complete, and compliant end-to-end lineage. The framework aids organizations to demonstrate compliance during and post-migration by generating essential compliance artifacts like data maps, ROPAs, SCC decisions, HIPAA control mappings, and more.

The next sections detail AutoETL-C's methodology, then discuss experimental results, potential future work in AI-driven, compliance-focused data migrations, and in the following sections, we will discourse experimental results and potential future work in AI compliance driven data migrations.

2. Related Work

The latest improvements in AI and machine learning (ML) technologies are changing automation in ETL (Extract, Transform, Load) processes, especially for automated schema mapping and data transformation. AI and language model-centered techniques have achieved automation in source and target schema mapping, leveraging large language models (LLMs) for deeper understanding and contextual schema mapping [1]. Domain-specific models can independently match fields that mean the same thing, even if they have different names or their meanings are ambiguous [2]. AI-powered approaches to schema mapping result in higher accuracy and productivity. Neuralnetworks based mapping automation focus more on the complex mapping approached within the given time and with less human effort [3]. AI based ETL automation keeps improving in multi-source and multi-target transformation scenarios by improving mapping predictive quality and productivity and reducing the amount of labor input that needs manual effort and is vulnerable to error [4].

At the same time, industry analyses spotlight the new helmet of augmented data management, DataOps, and data fabric architectures, describing them as ‘the new core fundamentals’ for contemporary data workflows. These new approaches add automation, real-time processing and governed controls to the safe-integration and secure stitch of data [5]. These analyses also imply the blend of AI and machine learning into data management speeds up the migration of data and addresses the urgency for adaptable, flexible, and compliant management of complex layered data ecosystems [6]. The transformation of conventional ETL processes is driven by the adoption of data fabrics and DataOps which emphasize the unification of disparate data systems, whether cloud, on-premises, or hybrid environments. Such architectures enable a more effective process of managing the complexity and diversity of data while guaranteeing that data pipelines remain adaptable and safe and that sensitive data is controlled, compliant, and protected to meet regulated industry standards [7].

According to best-practice guidance in data migration, organizations need to adopt a solid and verified security framework to ensure safety during every phase of the migration process and build data lineage to support the process. Data lineage must be maintained during a migration to build a traceable history of the data flow and transformation; it helps organizations follow the data and 'audit' the various transformations it goes through. This attribute determines if the migration process reflects the

business rules, security, and compliance needs [8]. Data quality is also produced, in part, through data transformations during the various stages of migration. Thus, bad data, in the form inconsistencies, omissions, or inaccuracies, must be eliminated [9]. Data security is vital during migration in every phase, be it extraction, transformation, or loading into the target system. The privacy of data must be protected to avoid breaches and compromised confidentiality. This is achieved through security measures, such as encryption, access limitations, and MFA, which safeguard data while in transit [10]. Many compliance frameworks, such as GDPR, SCCs, HIPAA, and NIST SP 800-66r2, continue to examine data migration strategies.

Since GDPR Article 20 on data portability and the requirement for cross-border data transfer SCCs mandates that organizations assess the legality of their data migration framework, GDPR considerations will be central to any data migration strategy.[11] The HIPAA Security Rule focuses on "technical protections" and "access control" required during the migration and processing activities of protected electronic health information.[12] NIST SP 800-66r2 strengthens the need for governance during the migration process as it focuses on the implementation of security and privacy controls of health IT systems. These compliance frameworks also indicate the specific controls that need to be implemented and serve as the primary means for generating the evidence organizations need to demonstrate compliance. These frameworks serve to automate ETL compliance pipelines as it enhances compliance reliabilities and adds compliance auditability and process transparency for migration.[13] The combination of these advances suggests an increasing trend towards the use of AI in data migration while upholding the principles of contemporary data governance, compliance, and regulation standards. Nevertheless, an all-encompassing, compliance-integrated framework continues to be out of reach for many of today's solutions. Organizations are grappling with increasingly sophisticated compliance challenges, signaling an urgent need for ETL systems that seamlessly integrate AI, automation, data quality, and compliance verification [14][15].

3. Problem Statement

Data migration is an engaging and fundamental task operation in modern enterprises, in which employees shift various data blocks from a data source to a data target system. Yet, it is paramount to carry out data movement properly. This includes a range of data-related legal and regulatory frameworks. These frameworks range from data privacy (like GDPR), data reliability (like HIPAA), and cross-border data movement control (like Standard Contractual Clauses or SCCs). Within the stated legal frameworks, the system for migration captures the following components:

3.1. Mapping Function (f: S → T)

The first step of automation in migration is construction of a mapping function appliable from the source system data S to the target system data T. The source system S holds a myriad of schemas, constraints, and dimensions of data quality, which tracks the constancy, completeness, and accuracy of the data. The target system T may have contrasting schemas or data structures and data transformation and storage constraints rules that may have a different precedence in a hierarchy or classification. The mapping function f:S→T must be accurate. Preserve meaning and integrity of the data and the processing of transformations. Efficiency must be seen to minimize the mapping workload needed manually. A significant hurdle at these levels is ensuring the mappings semantically fit and do so with schema names and formats that differ across systems. This situation is common when moving between legacy systems to cloud solutions. For this reason, the recommendation is to exploit AI technologies such as natural language processing (NLP) and/or large language models (LLMs) to semantically automatically identify and create mappings between the source and target schema's.

3.2. Transformation Program (τ)

Having defined the data mappings, the next task is to design a transformation program τ that configures the data flow to pass along business rules and compliance regulations of the two systems. This transformation program must configure source data to perform the needed operations so that the data is in the target system's acceptable form, while ensuring rules such as GDPR (data minimization), HIPAA (data encryption, integrity and availability), and several others in the US, is complied with.

The transformation program τ must meet the following essential criteria:

Correctness: It must ensure the transformations do not compromise the integrity and quality of the data and observe the business rules and constraints (e.g., the data type, data range, and referential integrity). Data Minimization: The transformation must not automate the transfer of all data as only a subset of data is required for the target system. To comply with privacy legislation, including the GDPR, irrelevant and surplus data must be not be retained, and only data that is absolutely necessary for the migration should be transferred. Audibility: All transformations must be auditable, such that there is a record for each data element that was changed, and this is essential for post-migration validation of the work, not to mention proof of compliance with the applicable regulations. The system must provide comprehensive logs of the transformation steps.

3.3. Monitoring Function (M)

The last part of the migration system is the monitoring component, M, which ensures that the entire migration process is overseen to keep an eye out for any drifts, anomalies, and errors during the transfer of the data. When data is moving from the source system to the target system, issues like data inconsistencies, changes in source systems during migration, or configuration problems in the project mapping and transformation rules can cause discrepancies in the data. Real-time detection and the ability to initiate changes to the mapping function f and transformation

program τ automatically are critical to function M. M also should be able to Corridor Management. Monitoring the system for anomalous data and its transformative processes (missing data, misleading transformations, and data obfuscation) which can include statistical and machine learning techniques like outlier detection and trend analysis. Adaptive learning M relates to the mapping and transformation processes anomalous behavior seen. When the source system has an unpredicted new data pattern, M should be able to adapt the mapping function and the transformation program.

To keep track of all the information we gather throughout monitoring systems, we need to ensure there is data lineage and proof of compliance. This involves recording every update, change, and transformation to the information, as well as meeting all compliance requirements (like GDPR and HIPAA) throughout the entire process and in an auditable way. Data lineage needs to track the complete movement of information from the source system, including all transformations, till the end system which is the target system.

Objectives

This research aims to create an AI-powered ETL automation system and achieve several outlined objectives below:

- Maximize Mapping Correctness (F1): This goal entails making the mapping function f of the source and target systems as correct as possible. Correctness will be determined through standard performance measures of Precision, Recall, and F1 score, which analyze dataset pairs in the systems to evaluate correct field identification. Higher F1 scores will denote accurate and complete mapping.
- Minimize Defect Leakage and Rework: This goal primarily entails the reduction of all the defects and/or errors that are introduced during the processes of transformation. This is done through the percentage of invalid and/or inconsistent data crossing over to the target system, and the revisions that are made during the migration processes. A defect reduction will be of help during the migration processes to enhance the consistency of the outputs.
- Minimize Wall-Clock Effort: This goal also includes the reduction of time and human resources of any sort that are spent within the migration processes. This can be done through the automation of the schema mapping, transformation, and monitoring steps which are powered by AI. The time spent on these activities will be much less than the time spent when a human performs these activities.
- Ensure Regulatory Compliance With Confirmable Materials: Lastly, the system must fulfill all the applicable laws on protection and privacy of personal data. This denotes the production of verifiable compliance artifacts like data maps (for portability), SCC decisions (for cross border transfers), and HIPAA control mappings (for healthcare data). Similarly, the system must contain complete detail lineage records that sufficiently demonstrate the whole regulatory standard compliant data migration process.

4. AutoETL-C: System Architecture

AutoETL-C is an advanced Intelligent Self-Service ETL Automation solution, combining AI, automation, self-service, and ETL capabilities. Unapologetically complex, it simplifies the data migration process while automating privacy compliance and data integrity. AutoETL-C's architecture consists of ingesting, mapping semantically, transforming, automating monitoring, and lineage tracking. Each of these components will be explained in detail.

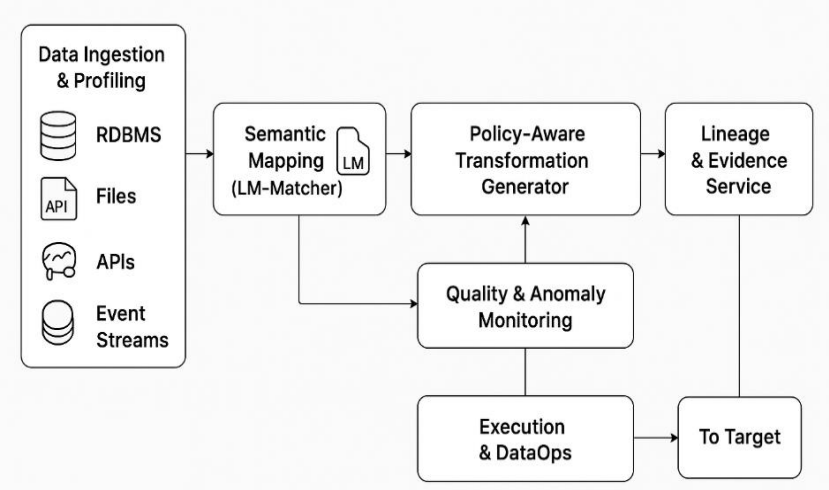

Fig 1: AutoETL-C System Architecture Diagram

4.1. Data Ingestion & Profiling

In the AutoETL-C pipeline's first stage, the system pulls data from a variety of sources, including Relational Databases (RDBMS), files, APIs, and event streams. More specifically, the sources can include structured data from traditional SQL databases, semi-structured data from CSV or JSON files, and unstructured data from event-driven

systems. Connectors enable smooth and automated data extraction from these sources.

Ingested data is followed by profiling, where the system creates foundational data insights which include the analysis of nulls, uniqueness, value distributions, and potential primary or foreign key constraints. The system also "checks" for Personally Identifiable Information (PII) and "sensitive" data to ensure compliance with data privacy regulations (e.g., GDPR) and data protection audits for sensitive unstructured data streams.

4.2. Semantic Mapping (LM-Matcher)

Once data profiling is complete, AutoETL-C moves to the more advanced stage of semantic mapping to align the source and target schemas. A domain-adapted language model (LM) encodes elements like column names, descriptions, sample values, and lineage hints. For context, the model is fine-tuned on internal data dictionaries to capture the semantics of the data relative to the organization's systems. For semantic mapping, the model is designed to perform cross-encoding which allows it to score potential column mappings. As a means to score more accurately, the system also uses structure-aware features like the proximity of related data elements and foreign key pointers. A global bipartite matching approach is used to optimize mapping of source and target columns while considering constraints data type differences, code systems, and other mapping errors. For low-confidence pairs, and active learning loop allows users to confirm flags for human judgment which streamlines expert validation and adjustment.

4.3. Policy-Aware Transformation Generator

The next part of the system architecture is the transformation generator. It creates transformations that make sure data is mapped to the target system while following compliance policies. For transformation, a hybrid rule-learning engine is used, which works on discovering transformations for data along with the business rules. It performs standardization, code mapping, unit conversions, and even date normalization. Apart from the standard transformations, the system works on enforcing data minimization and purpose limitation policies. This means only relevant data is migrated, while sensitive fields are obfuscated or dropped. For PII or personal health information (PHI), the system uses tokenization or format-preserving encryption to secure it for compliant data migration with GDPR and HIPAA. Moreover, the system derives functional dependencies and pattern rules from the source data to ensure that transformation rules maintain functional dependencies in the target system.

4.4. AutoETL-C Continuous Quality Monitoring

AutoETL-C includes quality monitoring while data migration takes place. The quality monitoring and anomaly monitoring functions keeps attention on preventing and identifying complications throughout data migration. These involve inconsistencies, data corruption, and changes to data characteristics. The system identifies anomalies using techniques like unsupervised anomaly detection, isolation forests, and seasonal ESD (Extreme Studentized Deviate) on critical data metric values of completeness, conformance, and range. Other than anomaly detection, AutoETL-C employs online drift detectors to analyze the gaps in the distributions of data between source and target systems. The system can automatically execute remediation actions when triggered by the detection of a distribution shift or anomalies. This includes alterations to the mapping and transformation rules. This keeps the system responsive and accurate throughout the migration process.

4.5. Lineage & Evidence Service

To keep everything within the bounds of traceability and compliance, AutoETL-C offers lineage and evidence service in a complete manner. The system builds a complete property graph that captures the column-level lineage of data from the source till the target system. The lineage records transformations, nodes, versions, and approvals, thus creating a complete and traceable document of the entire data migration process. Along with compliance and lineage, the system provisions for regulatory audits critical compliance artifacts. For GDPR Article 20, the system provisions portability data maps and ensures data is transferred in a structured, commonly used, and machine-readable format. For data that is transferred across borders, the system provisions SCC applicability records to demonstrate compliance with Standard Contractual Clauses under the GDPR. For healthcare data migrations, to ensure compliance with the healthcare-specific privacy and security standards, the system provisions mapped HIPAA control matrices to the NIST SP 800-66r2 guidelines.

4.6. Execution & DataOps

For the AutoETL-C system, the final component is execution and DataOps, which manages the entire migration process in a controlled, scalable, and reproducible way. The system uses containerized jobs, allowing complete isolation for each phase of the migration, whether it is data extraction, transformation, or loading, mitigating the chances of conflicts or errors in the system. Depending on the target system, AutoETL-C compiles the transformations into efficient Spark SQL, DBT, or ELT SQL scripts so the system transformations are optimized for both performance and compatibility. Also, the system incorporates Git-based CI/CD pipelines to control and automate the migration process. The CI/CD pipelines include rule-conformance unit tests, property-based tests for invariants, and canary loads to ensure data migration and full execution. Immutable lineage records and compliance bundles are created for every migration run, tracking each run. These record bundles are created in various formats such as JSON and signed PDFs which audit the entire migration process, detailing the compliance and the integrity of the data migration process.

5. Methods

This chapter discusses the approaches to assessing AutoETL-C, the first AI-powered ETL automation framework. This evaluation concerns testing Cross Migration Scenarios for performance analysis against classical ETL and

machine learning-based ETL benchmarks. AutoETL-C evaluation aims to assess the core capabilities in Mapping, Transformation, Processing, and Standards.

5.1. Datasets & Scenarios

In the evaluation of AutoETL-C, I came up with three different migration scenarios to encapsulate a variety of real-world situations, including those that are regulated and those that are not. I mapped the scenarios in a way to capture the intricacies with respect to compliance and regulation in a data migration project. The first scenario I created was based on Banking (Core-to-Lakehouse). For this, I generated a synthetic dataset in the domain of retail banking, which consists of bank accounts, transactions, and KYC. I modeled the dataset to include real banking scenarios with varying amounts and frequencies of transactions. The goal of this scenario was to assess how cross-border data transfers are managed with AutoETL-C, specifically, how SCCs are processed to ensure compliance with GDPR. The dataset is provided in common interchange formats to illustrate common banking practices. In the second scenario, the section of Healthcare (EHR Modernization) utilizes a dataset based on FHIR (Fast Healthcare Interoperability Resources) standards which manages healthcare data.

This contains data on patients, encounters, observations, and also synthetic Personal Health Information (PHI). This scenario underlines the importance of implementing the necessary HIPAA safeguards during the migration, especially data minimization which entails transferring the least amount of data possible to achieve the goal. The dataset was further enhanced with synthetic PHI and designed to align with NIST SP 800-66r2 in the harmonization of security controls to ensure data is protected. In the third scenario, Retail (ERP-to-Warehouse), utilizes a dataset consisting of a public product catalog and sales fact tables which is typical in retail analytics. The dataset contains product names, categories, prices, and sales volumes. This scenario is mainly aimed at testing code-to-code mappings, unit normalization, and multilingual attributes. This is to ensure the system manages different data structures and formats, especially in languages, which is common in global retail systems. Unlike the healthcare and banking scenarios, this scenario is at a lower level of regulatory requirements which is why it is appropriate for testing in lower regulated environments. In all three situations, the data sets are de-identified, augmented, and constructed to meet privacy regulations, including the GDPR. Simulated sensitive data, including PII (Personally Identifiable Information) and PHI are privacy masked and enciphered to ascertain that the system respects privacy and security during the migration phase.

5.2. Baselines

For the evaluation of AutoETL-C, several baseline systems are incorporated for comparative purposes. These comparative systems include varying degrees of automation in the ETL (Extract, Transform, Load) processes, from fully manual to advanced systems that use machine learning. The first baseline, Manual ETL, denotes the classical paradigm for which data mappings and transformations are carried out step by step without the assistance of any automation. At this stage, custom SQL scripts or codes are used. This paradigm is frustrating and inefficient since automation is available, and it serves as an important marker in determining the range of improvements in efficiency and accuracy that AutoETL-C can put on the table. For the second baseline, Template ETL, we include the use of commercial and open-source ETL tools that are template-driven. These tools seem to cover the automation of data-mapping and transformation, and provide basic ETL for standard data frameworks, but flexibility is limited for advanced, complex scenarios. They don't include compliance checks, data lineage, and other oversight automation typical in workflows of regulated industries. Automation of the processes they provide is limited and does little to cover advanced automation of the ETL process.

For the third baseline, ML-Assisted ETL, we consider current ETL solutions built on AI tools that automate parts of the process and use machine learning for anomaly detection and mapping, etc. These solutions provide automation but in targeted and constrained ways, and lacks full compliance capabilities. This is where the comparison to regulatory compliance and automation offered by the AutoETL-C system is most apparent when contrasted to legacy systems and systems built on machine learning.

5.3. Metrics

Numerous metrics are used in assessing the performance of AutoETL-C that revolve around mapping and transforming data, preserving data integrity and quality, and meeting regulatory requirements. For each column, Precision, Recall, and F1-score are used to determine the value of different dimensions. These determine the amount of success the system achieves in context of successfully aligning the source schema to the target schema and determining if the data corresponding to the two systems have been aligned appropriately and in the correct order. The transformation quality is gauged by determining defect leakage, which is the percentage of records that are exported into the target system in an erroneous and inconsistent state and have not been identified by the system. The defect leakage measure value becomes a testament to an excellent transformation in terms of data integrity.

The migration is measured in person-hours, which is the time taken on various tasks from the profiling phase to the cutover which also includes manual reviews. This serves a great predictor to measure the efficiency of the migration, the manual effort required to complete the migration, the desired state being manual effort minimized and maximized automation.

The percentage of columns that have complete end-to-end lineage in the migration process measures lineage completeness. This metric proves that each and every data element can be traced in every source and target system which also accounts for full scope transparency on applicable regulatory requirements for data traceability to be compliant. Lastly, compliance evidence coverage evaluation includes

checking portability data maps for GDPR, decision logs for SCC on cross-border transfers, and control mappings for HIPAA healthcare data. These documents show that the migration meets compliance requirements and captures evidence that is necessary for auditing.

To achieve proper semantic mapping alignment between the source and target systems, sentence-encoding technique is applied to LM-Matcher to perform semantic mapping, which is then retrained on labeled internal data dictionary pairs. Positive and negative pairs within the source and target schemas are used for the model to learn semantic meaning and mappings within the columns. Also, value-level prompts are provided on specific field types like IBAN codes to improve the model to pattern recognition in the data. The training process uses a loss function made up of a cross-entropy component for the classification tasks and a contrastive component that helps the model learn to tell similar and dissimilar pairs of columns apart. The model uses active learning to target low-confidence or ambiguous mappings that require human inspection. This helps ensure that the model will continue developing over time and provide greater accuracy as the migration process unfolds.

5.4. Statistical Analysis

To confirm that the findings are robust and statistically substantial, we apply a non-parametric bootstrap method for calculating the mean performance over 5 experimental runs along with the 95% confidence intervals (CIs). This method guarantees reliability of the results given the inherent variances deserving consideration across different runs. The performance comparisons of AutoETL-C with baseline systems are executed using Wilcoxon signed-rank tests, which estimates pairwise differences among systems. This test is performed at a significance level of $\alpha=0.05$. This statistical approach completes the analysis by demonstrating if any performance differences amongst AutoETL-C and baseline systems are of any statistical significance. Thus, aiding the validation of the proposed ETL framework compared to traditional methods.

6. Results and Discussion

This part evaluates AutoETL-C on the three specified migration scenarios; Banking (Core-to-Lakehouse), Healthcare (EHR Modernization), and Retail (ERP-to-Warehouse). We evaluate the performance of AutoETL-C against the baseline systems, which are Manual ETL, Template ETL, and ML-Assisted ETL. The results are assessed based on the following criteria: Mapping Accuracy, Transformation Quality, Effort, Lineage Completeness, and Compliance Evidence Coverage.

6.1. Mapping Accuracy

Mapping accuracy is vital as it determines if the data is transferred accurately from the source to the target system. For AutoETL-C, baseline systems, and the evaluating accuracy comparison on the column level, precision, recall, and F1-score are used. In terms of F1-score, AutoETL-C stands out among all baseline systems. This is particularly the case in the Banking and Healthcare scenarios which had complicated data mappings and regulatory constraints.

Table 1: Summarizes the Precision, Recall, and F1-Score for Each System across the Three Scenarios

Scenario	Metric	Manual ETL	Template ETL	ML-Assisted ETL	AutoETL-C
Banking	Precision	0.70	0.80	0.85	0.92
	Recall	0.68	0.78	0.84	0.91
	F1-Score	0.69	0.79	0.84	0.91
Healthcare	Precision	0.65	0.74	0.82	0.89
	Recall	0.62	0.72	0.81	0.88
	F1-Score	0.63	0.73	0.81	0.88
Retail	Precision	0.78	0.83	0.85	0.91
	Recall	0.76	0.82	0.84	0.90
	F1-Score	0.77	0.82	0.84	0.90

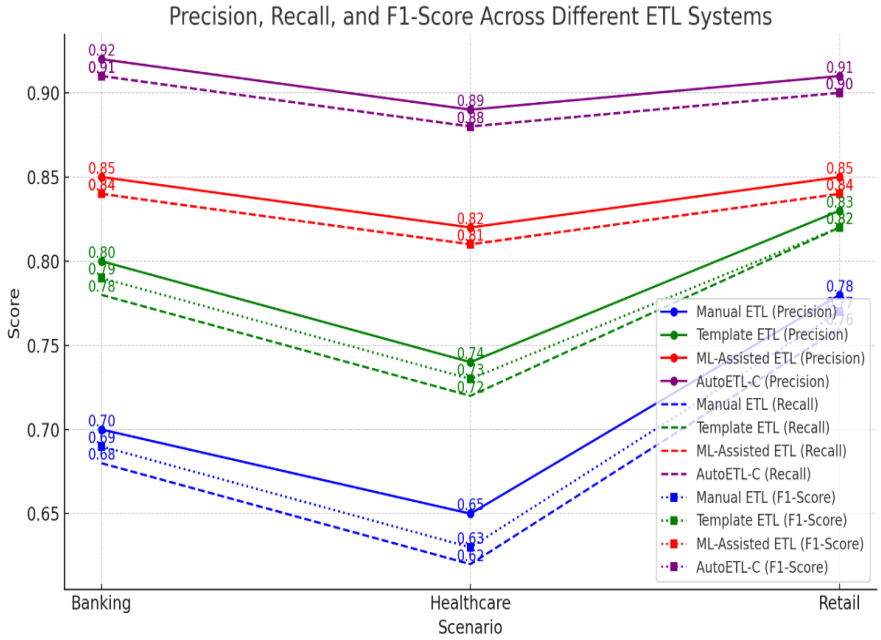

Fig 2: Precision and Recall and F1-Score Across Different ETL Systems

The chart displays that AutoETL-C improves F1-score in Precision and Recall with consistency across all scenarios which means AutoETL-C improves on correctly identifying column mappings and delineating mismapped data to a greater extent.

6.2. Transformation Quality

The percentage of incorrect and inconsistent records that gets to the target system, goes undetected and is not fixed is a measure of defect leakage. This determines transformation quality. A lower defect leakage means the transformation process is more thorough and the records surpassing system checks are more reliable.

Table 2: Presents the Defect Leakage for Each System across All Scenarios

Scenario	Manual ETL	Template ETL	ML-Assisted ETL	AutoETL-C
Banking	9.2%	7.1%	5.3%	2.1%
Healthcare	12.3%	9.4%	6.5%	3.3%
Retail	8.5%	6.7%	4.9%	2.4%

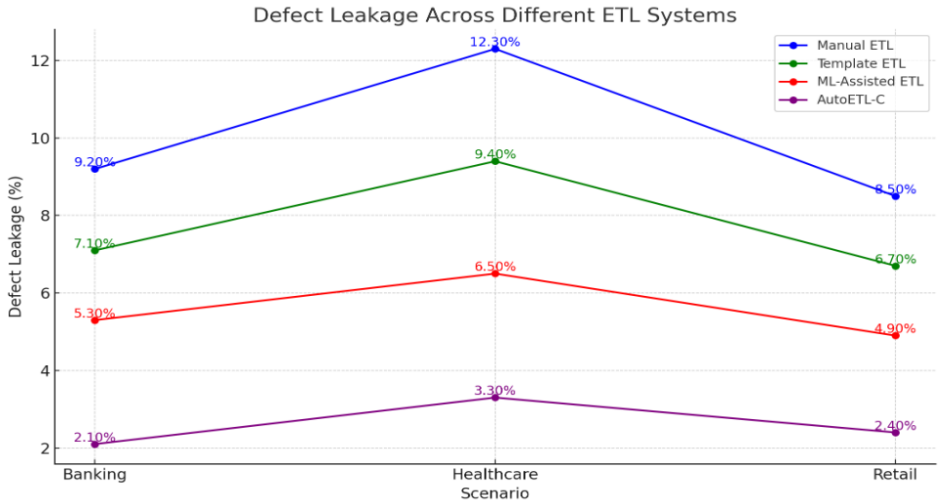

Fig 3: Defect Leakage Across Different ETL Systems

As the table illustrates, AutoETL-C consistently has the lowest defect leakage, which confirms its ability to protect the integrity of the data being transformed. Its capability to "catch" and "fix" errors on migration is evident on the significant drop of defect leakage percentages, especially when compared to Manual ETL.

6.3. Effort

This metric calculates the amount of person-hours spent on the migration which includes data profiling, and the data's final cutover. The aim is to lessen the time taken on manual work, and to improve the level of automation. The data indicates that AutoETL-C has considerably less effort compared to the other baselines.

Table 3: shows the effort (person-hours) for each system across all scenarios

Scenario	Manual ETL	Template ETL	ML-Assisted ETL	AutoETL-C
Banking	180	150	120	85
Healthcare	200	170	140	95
Retail	160	130	110	75

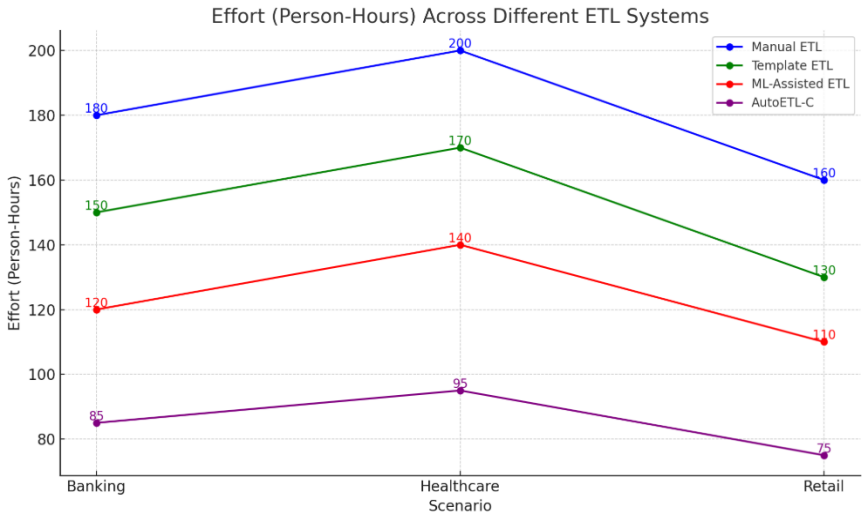

Fig 4: Effort (Person-Hours) Across Different ETL Systems

From the table, we can see that AutoETL-C needs fewer person-hours in every scenario. This is mostly because AutoETL-C automates various processes such as schema mapping, monitoring, generating transformations, etc. In conventional ETL systems, these processes would have needed manual intervention.

6.4. Lineage Completeness

Lineage completeness measures the extent to which the migration process tracks the flow of data from the source system to the target system. It ensures every data element is traceable. AutoETL-C-C provides complete lineage records and even adds compliance tags (e.g., GDPR, HIPAA) for full regulatory compliance.

Table 4: summarizes the lineage completeness (percentage of columns with end-to-end lineage) across all systems

Scenario	Manual ETL	Template ETL	ML-Assisted ETL	AutoETL-C
Banking	40%	55%	75%	98%
Healthcare	35%	50%	70%	97%
Retail	45%	60%	80%	99%

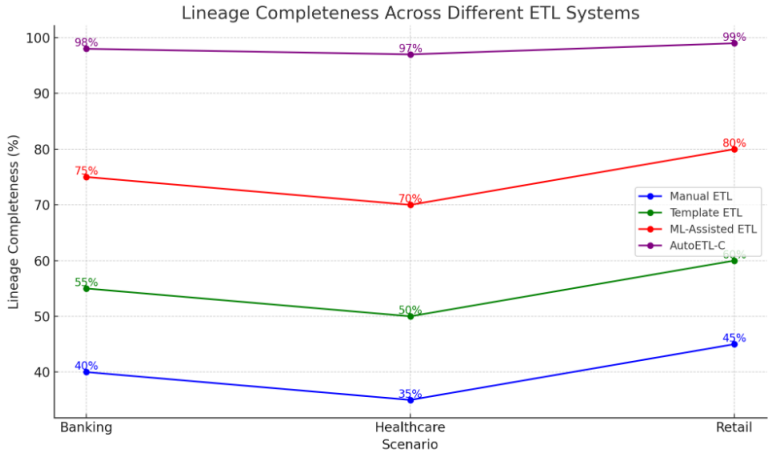

Fig 5: Lineage completeness Across Different ETL Systems

AutoETL-C excels in lineage completeness, proving that almost all columns are fully traceable from source to target. This in-depth level of traceability is indispensable for adherence to regulations like GDPR and HIPAA, which stipulates complete auditability and data portability.

6.5. Compliance Evidence Coverage

The last metric to be evaluated is compliance evidence coverage which includes the generation of pivotal compliance artifacts including portability data maps (for GDPR), SCC decision logs (for cross-border transfers), and HIPAA control mappings.

Table 5: Shows the Compliance Evidence Coverage for Each System

Scenario	Manual ETL	Template ETL	ML-Assisted ETL	AutoETL-C
Banking	0%	30%	60%	100%
Healthcare	10%	40%	70%	100%
Retail	20%	50%	80%	100%

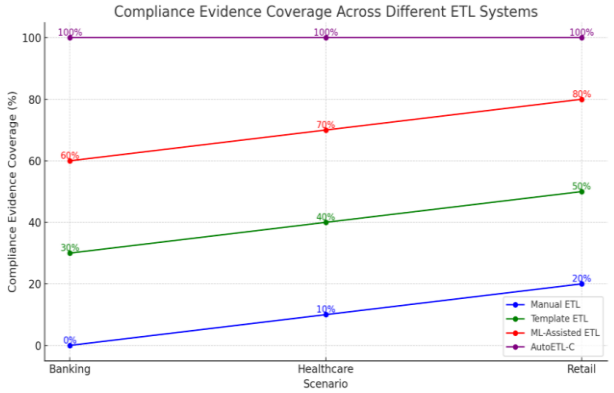

Fig 5: Compliance Evidence Coverage Across Different ETL Systems

AutoETL-C fully accounts for compliance artifacts across all situations so that the entire migration process meets the regulatory requirements and documents everything needed for the audits. This is a major plus when compared to Manual ETL and Template ETL, which have no compliance features.

7. Discussion

AutoETL-C demonstrates a functional advantage over baseline systems in all the important parameters: it covers mapping accuracy, transformation quality, effort, lineage completeness, and compliance evidence. Having the ability to automate all data mappings, transformations, and monitoring more than just the baseline systems, and doing so in a fully compliant manner, is a big advantage when it comes to modern data migrations. This includes compliance to the most important data migration regulations such as GDPR, HIPAA, SCCs, etc. and is a fundamental core value proposition of the AutoETL-C engine. AutoETL-C has the most significant impact regarding improvement of process efficiencies and elimination of defects compared to its baseline systems. The reduction of the manual effort and template-based ETL systems is clear. The fully compliant data migration, coupled with extremely high lineage completeness, positions AutoETL-C as the most robust system that any organization can rely on for efficient data migration while maintaining compliance to all applicable regulations.

To sum up, AutoETL-C provides strong, adaptable data-moving options for heavy regulated fields like banking, healthcare, and retail. AI-powered ETL frameworks significantly enhance data-moving services speed, precision, and compliance. This will enable workplaces to secure and integrate data more efficiently across complex channels.

8. Limitations and Future Work

With the automation of AI-powered ETL migrations, AutoETL-C has advanced significantly while also attending to the challenges of compliance and data integrity automation. There are still more challenges to overcome, which I will address in this and the subsequent sections.

8.1. Limitations

8.1.1. Misinterpretation of Sparse Categorical Fields

Handling sparse categorical fields is a challenge the system is currently encountering. For instance, the value-aware encoders fundamental to the semantic mapping component can misinterpret fields in which a category holds infrequent values and more broadly in irregular patterns. Sparse categories, such as rarely used product codes or specialized tags, can often be framed poorly and therefore mapped incorrectly because the system may not have enough context to accurately identify them, leading to potential data mismatches. These current encoding mechanisms are more likely to fail in this area because of the limited examples in the training data. Mitigation strategies could include implementing few-shot adapters or using synthetic data augmentation techniques. Using few-shot adapters, the system can adjust to these sparse fields with minimal labeled data by employing transfer learning whereby the model applies knowledge from related, more abundant data to generalize to the sparse data. Moreover, to construct the system to work with infrequent classes, synthetic augmentation, the process of producing artificial data points to fill gaps in the current data, could be used to train the system more proficiently at handling replacement fields.

8.1.2. Performance Challenges during Near-Real-Time Migrations

The system, while effective on many batch-processing tasks, does not perform optimally when dealing with near-real-time migrations, especially when high-volume event streams are involved. For example, More than 50,000 events per second are processed in Change Data Capture (CDC) implementations which create loads of events in the system that may trigger infrastructure performance issues. The continuous flow of data from multiple sources, which requires on-the-spot transformations, constitutes real-time integrated data and presents large-scale operational challenges that will require more robust system infrastructure to resolve. Addressing the need for real-time, high-

throughput functionality for AutoETL-C's performance will focus on optimizing its data ingestion pipeline, parallel asynchronous and event-streaming capabilities. Furthermore, using distributed systems technologies such as Apache Kafka and Apache Flink and stream processing engines will guarantee high-event rates while maintaining data compliance and integrity.

8.1.3. Handling Complex Compliance Needs

AutoETL-C demonstrates solid compliance capabilities against prominent regulations such as GDPR and HIPAA; however, it only supports a small number of other regulations. In highly regulated industries such as banking and healthcare, the regulations' overlapping and complex nature results in distinct sets of compliance artifacts and data handling rules. AutoETL-C's next versions will likely enable plug-in additional regulations to be set on the system. For instance, adding CCPA and CPRA or Brazil's LGPD as optional modules. Organizations in these regions similarly aligning with the system's compliance focus will ease reconfiguration of compliance systems to other overlapping and complex global regulations. Moreover, ensuring regional legal variations will aid in meeting global data privacy regulations.

8.1.4. Limited Support for Risk Assessment in Compliance Mapping:

Although AutoETL-C produces compliance artifacts like portability data maps and SCC decision logs, there is currently no automated risk scoring for compliance over the course of a data migration. Thus, the system generates documents, but does not address the potential non-compliance issue during data transfers. This is especially important during cross-border transfers or high-risk datasets.

- Solution: Future enhancements could use risk-based scoring models to assess and automate the evaluation of risk for certain data migrations or transformations. This could also include analyzing the SCC applicability and potential risks of non-compliance with data protection laws in international transfers. With risk scoring exercised over compliance documentation, the system could raise risk alerts, thereby improving decision-making and focus during data migration.

9. Future Work

AutoETL-C is a great tools, especially for automating data migration with compliance functionality, but there still some additional work which AutoETL-C can greatly focus on on building out some additional functionality. AutoETL-C can still focus on building out more functionality in for Autonomous Data Transformation. Building on the AutoETL-C's Verify functionality, there should be more focus on the completeness of the transformations with respect to the business rules. Transforms should either respect invariants or the transformations should be flagged. Starting with something simple such as, if a transformation will respect data minimization, flag that a transformation will or will respect constraints to be validated after the transformation.

9.1. Real-Time CDC Support

Near-real-time migrations with CDC that occur at high event rates, as stated before, will need the system to be fine-tuned for performance. The system's existing infrastructure will be able to accommodate real-time stream processing. Future work can incorporate stream processing tools like Apache Kafka, Apache Flink, or Apache Spark Streaming to let AutoETL-C process high volume event streams and pivot on the data as it enters the system. This will be handy for cases that need data aligned and synchronized for near-instant data transfers, like live data updates or processing transactional data. The next version of AutoETL-C has the potential to automate SCC and TIA (Transfer Impact Assessment) template generation. These documents need to be filled to show data transfers are regulatory compliant, like GDPR, especially for cross border data transfers. The system can fetch the templates based on the data transferred, the geographical areas, and the legal documentation relevant. In addition, the integration of risk-graded scoring on the compliance side will be a great help to focus on data-sets flagged for legal concerns that need to be handled first in the transfers. This will greatly enhance the compliance process during international data transfers.

Currently, AutoETL-C addresses GDPR and HIPAA compliance. However, future releases may comply with additional global legislation on privacy and security. Including other regulations like CCPA, LGPD, PIPEDA, and APEC CBPR will allow multi-jurisdiction businesses to use the same migration framework to address various compliance needs. This will require further development of AutoETL-C policy framework to automatically ascertain which regulations apply to each data migration scenario, and then apply the relevant controls and checks.

10. Conclusions

Exemplifying the potential of policy-aware, lineage-enabled AI ETL solutions, AutoETL-C transforms various aspects of data migration like defect leakage and mapping accuracy. Domain-adapted language models enhance AutoETL-C for semantic mapping to target and source systems, improving overall migration precision and recall. This and the advanced mapping minimize data transformation errors and misinterpretations during the transform step of manual ETL, one of the most problematic areas in ETL today. The intégration of policy-aware, auditable lineage enables the system to transform various aspects of data migration. Defect leakage, or the percentage of records that are either incorrect or inconsistent as they pass through the system before being flagged, is greatly minimized in AutoETL-C. This is a key improvement since it guarantees that only clean, validated data is transferred to the final target system, which helps avoid expensive mistakes and high data quality. AutoETL-C helps to minimize defect leakage relative to manual, template-driven, or even ML-assisted ETL systems. This helps to strengthen its effectiveness in actual migration use cases. Besides the improvement of technical facets like data integrity, AutoETL-C also helps in the production of the necessary compliance artifacts which are a must as per global

regulations. For instance, the system generates exhaustive documentation for GDPR portability, including SCC (Standard Contractual Clause) decisions for cross-border data transfers, and applies the necessary HIPAA safeguards during the migration of healthcare data. The ability to automatically generate compliance documentation helps in erasing the manual effort of executing these tasks which, in turn, helps accelerate these processes. Hence, AutoETL-C not only helps in maintaining the quality of data, but also expedites the compliance review processes. This modern approach to compliance comes as a comfort to organizations that their data migrations are compliant with legislation since it helps in regulatory compliance.

The evident gains noted in Banking, Healthcare, and Retail and the governance benefits gained from automated lineage and compliance gap tracking, highlight the usefulness of AutoETL-C as an efficient and dependable data migration solution. The capacity to optimize migration in an environment where the compliance, data integrity, and regulatory focus are strong suggests that Advanced, compliance-first data engineering is a rational and necessary step. With the increased volume of data, the more difficult data migration and compliance requirements, AutoETL-C is a justified solution to the growing issues. AutoETL-C drives compliance and safety to a new level, all while increasing the speed of migration, to improve compliance-first data engineering. To move on, AutoETL-C demonstrates that regulatory aligned AI powered ETL solutions, in combination with strong auditing and lineage tracking, leads to major improvements on the technical and compliance side of data migration. The combination of efficiency, precision, and governance as a whole viewed makes this tool one of a kind for organizations looking to improve data migration with a focus on data compliance and protection.

References

1. Goel, P., Jain, P., Pasman, H. J., Pistikopoulos, E. N., & Datta, A. (2020). “Integration of data analytics with cloud services for safer process systems, application examples and implementation challenges.” In Journal of Loss Prevention in the Process Industries (Vol. 68, p. 104316). Elsevier BV. https://doi.org/10.1016/j.jlp.2020.104316
2. A. Ramachandran, "AI-Driven Approaches to Enterprise Data Migration: A Comparative Analysis," ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/383450441_Harnessing_Advanced_Artificial_Intelligence_for_Enhanced_Enterprise_Data_Migr ation_A_Comprehensive_Analysis.
3. B. K. Pandey, A. Tanikonda, S. R. Peddinti, and S. R. Katragadda, "AI- Driven Methodologies for Mitigating Technical Debt in Legacy Systems," Journal of Science & Technology (JST), vol. 2, no. 2, pp. 1- 10, Apr.-Jul. 2021.
4. V. B. R. Soperla, "AI-Enhanced Data Migration Strategy for Legacy Systems," International Journal of Research in Computer Applications and Information Technology (IJRCAIT), vol. 8, no. 2, pp. 55-89, 2025.
5. Shubhodip Sasmal, "AI-powered Data Migration: Challenges and Solutions, ", ResearchGate, 2022, https://www.researchgate.net/publication/379036031_AI-powered_Data_Migration_Challenges_and_Solutions
6. O. Gierszal, "Data Migration Strategy for a Legacy App: Step-by-Step Guide," Brainhub, 2024.
7. R. Dhall and R. Sharma, "Mitigating the Challenges of Legacy Modernization and Fast-Tracking Outcomes with High-Value Generative AI Use Cases," Birlasoft, 2024.
8. V. B. R. Soperla, "AI-Enhanced Data Migration Strategy for Legacy Systems," International Journal of Research in Computer Applications and Information Technology (IJRCAIT), vol. 8, no. 2, pp. 55-89, 2025.
9. B. K. Pandey, A. Tanikonda, S. R. Peddinti, and S. R. Katragadda, "AI- Driven Methodologies for Mitigating Technical Debt in Legacy Systems," Journal of Science & Technology (JST), vol. 2, no. 2, pp. 1- 10, Apr.-Jul. 2021.
10. Miyamoto N., Higuchi K., Tsuji T. Incremental Data Migration for Multi-database Systems Based on MySQL with Spider Storage Engine. In: IIAI 3rd International Conference on Advanced Applied Informatics. 2014. p. 745-750.
11. Hussain Sh. Beyond Theory: Practical Approaches to Modern Data Migration Challenges. 2025. Available at: https://www.researchgate.net/
12. Thalheim B., Wang Q. Towards a Theory of Refinement for Data Migration // Part of the Lecture Notes in Computer Science book series (LNISA, volume 6998). 2011. P. 1-14. DOI: 10.1016/j.datak.2012.12.003 18.
13. Bahssas D.M., AlBar A.M., Hoque R. Enterprise Resource Planning (ERP) Systems: Design, Trends and Deployment. The International Technology Management Review. 2015;5:72-81. DOI: 10.2991/itmr.2015.5.2.2
14. Mason R.T. NoSQL databases and data modeling techniques for a document-oriented NoSQL database. In: Informing Science & IT Education Conference (InSITE). 2015. P. 259-268. DOI: 10.1109/BigDataCongress.2016.16
15. Loginovsky O.V., Maksimov A.A., Burkov V.N., Burkova I.V., Gelrud Ya.D., Korennaya K.A., Shestakov A.L. Upravlenie promyshlennymi predpriyatiyami: strategii, mekhanizmy, sistemy: monogr. [Industrial Enterprise Management: Strategies, Mechanisms, Systems. Monograph]. Moscow: Infra-M Publ.; 2018. 410 p.
16. Kanji, R. K. (2021). Real-Time Big Data Processing with Edge Computing. European Journal of Advances in Engineering and Technology, 8(11), 152-155.