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

Managing Data Conversion Complexity in Legacy System Modernization: Challenges Risk and QA Strategies

Sridhar Gajavelli

Abstract: Legacy system modernization increasingly depends on the accurate and reliable conversion of heterogeneous, complex, and historically accumulated data into modern technological environments. This study examined the challenges, risks, and quality assurance strategies associated with data conversion during legacy system modernization, with particular emphasis on data complexity, conversion risk, QA maturity, automation, and conversion performance. A quantitative-dominant mixed-method approach was adopted, integrating project-level characteristics, practitioner assessments, conversion metrics, and statistical analysis. Data Conversion Complexity Index (DCCI), Conversion Risk Index (CRI), Quality Assurance Maturity Index (QAMI), Automation Level Score (ALS), Conversion Error Rate (CER), Reconciliation Success Rate (RSR), and Conversion Success Index (CSI) were used as the principal analytical measures. The findings indicated that data-quality complexity, business-rule complexity, schema incompatibility, and system dependencies were among the most important contributors to conversion difficulty. Incorrect source-to-target mapping, referential-integrity failure, incomplete extraction, and business-rule errors emerged as the most critical risks. Conversion complexity showed a significant negative relationship with conversion success, whereas QA maturity and automation were positively associated with reconciliation success and overall modernization performance and negatively associated with conversion error rates. Regression analysis further identified QA maturity as the strongest positive predictor of conversion success, while conversion complexity exerted a significant negative effect. The multivariate analysis demonstrated a clear separation between complexity- and risk-related factors and QA-, automation-, and performance-related variables. The study concludes that successful legacy modernization requires an integrated, risk-aware, and quality-driven data conversion framework in which profiling, validation, reconciliation, automation, exception management, and continuous monitoring are embedded throughout the conversion lifecycle.

Keywords: *Legacy system modernization; data conversion; data migration; conversion complexity; quality assurance; risk management; data validation; automation; reconciliation; data integrity*

Introduction

Legacy modernization and data conversion

Legacy information systems continue to support critical business functions across Insurance,banking, healthcare, telecommunications, government, manufacturing, retail, and other data-intensive sectors (Kansara, 2021). Although these systems often provide operational stability, many were developed using outdated programming languages, proprietary databases, rigid data structures, and architectures that are increasingly difficult and expensive to maintain. Organizations are consequently modernizing legacy environments by adopting cloud platforms, distributed databases, microservices, enterprise data platforms, and software-as-a-service solutions (Assunção et al., 2025). Within these modernization initiatives, data conversion represents one of the most technically demanding and risk-sensitive activities. Data stored according to legacy formats, schemas, encoding standards, and business rules must be transformed into structures compatible with the target environment without compromising its meaning, accuracy, completeness, or usability. The effectiveness of data conversion therefore has a direct influence on whether modernization programs achieve operational continuity and intended business value (Haque & Rahman, 2021).

Complexity of legacy data environments

Test Architect, Coforge Ltd

Data conversion becomes particularly challenging because legacy systems frequently contain heterogeneous, highly interconnected, and poorly documented datasets accumulated over many years (Ponnusamy & Eswararaj, 2023). Historical databases may contain duplicate records, inconsistent formats, missing values, obsolete codes, incompatible data types, and fields whose original business meanings are no longer clearly understood. Multiple applications can also maintain different representations of the same customers, products, transactions, or operational processes. As organizations attempt to consolidate these sources within modern platforms, simple field-to-field transfer is rarely adequate (Mountford & Cai, 2023). Conversion activities may involve data profiling, cleansing, normalization, deduplication, schema mapping, transformation, enrichment, validation, and reconciliation. Furthermore, undocumented dependencies between databases and applications can introduce unexpected conversion failures. Understanding these relationships is essential because an apparently minor transformation in one dataset may produce downstream consequences for reporting, analytics, interfaces, and business processes (Zhou et al., 2019).

Risks associated with data conversion

The complexity of legacy data creates several technical, operational, and business risks during modernization. Data loss is among the most serious concerns because incomplete extraction or incorrect transformation can permanently remove valuable historical information (Gudivada et al., 2027). Data corruption may occur when source values are converted into incompatible target data types, while incorrect mapping rules can change the semantic meaning of migrated records. Duplicate records, truncated values, inaccurate calculations, character-encoding problems, and broken relationships between tables may remain undetected until the modernized system becomes operational. Conversion failures can consequently disrupt business processes, create unreliable reports, affect regulatory compliance, and reduce user confidence in the new system (Shneiderman, 2020). These risks become more significant in large-scale modernization programs involving millions or billions of records and multiple source platforms (uralidhar Sarabu, 2024). Effective risk management therefore requires systematic identification of conversion-related vulnerabilities before deployment rather than depending primarily on post-migration correction.

Quality assurance in conversion processes

Quality assurance is fundamental to ensuring that converted data accurately represents its legacy source while satisfying the structural and functional requirements of the target platform (Alexakis et al., 2025). QA for data conversion extends beyond conventional application testing because the primary concern is not only whether the system functions correctly but also whether the underlying information has been transformed without unintended modification. Effective QA strategies commonly incorporate source-to-target mapping validation, record-count reconciliation, field-level comparisons, referential integrity verification, duplicate detection, exception analysis, and validation of business rules. Automated testing can further improve coverage by evaluating large datasets that would be impractical to verify manually (Vigo et al., 2013). Statistical sampling and risk-based testing can complement automated validation, particularly when complete record-level comparison requires substantial computational resources. Establishing measurable acceptance criteria is equally important so that conversion quality can be evaluated objectively before production deployment (Dach et al., 2012).

Automation and continuous validation

Growing data volumes and increasingly complex modernization architectures have created a strong requirement for automated and repeatable validation processes. Automated data profiling can identify anomalies before conversion, while transformation testing can determine whether predefined mapping and business rules have been implemented correctly (Chirumamilla, 2024). Reconciliation frameworks can compare source and target datasets using record counts, aggregates, checksums, key fields, and financial or operational control totals. Integrating validation procedures into data conversion pipelines enables organizations to identify errors earlier and reduces dependence on manual inspection. Continuous logging and exception reporting also provide traceability for investigating failed or suspicious records (Hassan et al., 2020). Nevertheless, automation does not completely eliminate the need for domain expertise, because technically valid transformations may still misrepresent business

meaning when legacy rules are insufficiently documented.

Need for an integrated conversion framework

Many modernization projects address data mapping, conversion risk, and quality assurance as separate technical activities, even though they are closely interconnected. A mapping decision determines how legacy information is transformed, transformation complexity influences the probability of errors, and QA determines whether those errors can be detected before implementation (Li et al., 2016). An integrated approach is therefore required to connect data complexity assessment, risk identification, conversion execution, validation, and exception management throughout the modernization lifecycle. Such an approach can support prioritization of high-risk datasets, selection of appropriate testing techniques, and establishment of traceable controls for critical data elements (Raji et al., 2020).

Focus of the study

Against this background, the present study examines the management of data conversion complexity in legacy system modernization with particular emphasis on major conversion challenges, associated risks, and quality assurance strategies. The study explores how heterogeneous legacy structures, data-quality problems, transformation rules, system dependencies, and large data volumes influence conversion outcomes. It further evaluates risk-control mechanisms and QA practices capable of improving data accuracy, completeness, consistency, integrity, and traceability during migration to modern platforms. By integrating conversion complexity, risk management, and quality assurance within a unified perspective, the study seeks to provide a systematic framework for improving the reliability and effectiveness of enterprise legacy-system modernization initiatives.

Methodology

Research design

The study adopted a quantitative-dominant mixed-method research design to investigate the complexity, risks, and quality-assurance practices associated with data conversion during legacy system modernization. The methodology combined structured assessment of modernization projects, survey-based evaluation of practitioners involved in data migration activities, and examination of project-level conversion and validation records. The quantitative component was used to determine relationships among data conversion complexity, conversion risks, QA maturity, automation, and modernization outcomes, while qualitative observations obtained from project documentation and practitioner responses were used to interpret the major sources of conversion failure. The unit of analysis was the legacy-system modernization initiative, with particular attention to the conversion of structured and semi-structured data from legacy databases, flat files, proprietary systems, and enterprise applications into modern relational, cloud-based, or distributed data platforms.

Study population and data collection

The study considered professionals directly involved in enterprise data conversion and modernization projects, including data engineers, database administrators, ETL developers, solution architects, QA engineers, business analysts, project managers, and modernization specialists. Participants were required to have practical experience in at least one legacy data migration or conversion project. Data were collected using a structured questionnaire supplemented by project-level technical information where available. The questionnaire consisted of sections addressing characteristics of legacy systems, data quality, conversion complexity, transformation requirements, migration risks, QA procedures, automation practices, and conversion outcomes. Most perception-based variables were assessed using a five-point Likert scale ranging from 1, representing very low or strongly disagree, to 5, representing very high or strongly agree. Technical project measures such as data volume, number of source systems, conversion error rate, validation coverage, reconciliation success, and conversion duration were recorded using their corresponding numerical units.

Legacy system and project characteristics

Baseline characteristics were recorded to describe the modernization environment and control for differences among projects. The principal parameters included legacy-system age in years, number of source applications, number of source databases, data volume processed, number of tables or data entities, number of source-to-target mappings, target platform type, conversion

duration, number of conversion cycles, and percentage of historical data included in migration. Data volume was measured in gigabytes or terabytes and standardized before multivariate analysis where necessary. System heterogeneity was additionally represented through the number of distinct data-storage formats, database technologies, encoding standards, and application sources involved in each project. These characteristics were treated as contextual or control variables because they could independently influence conversion difficulty and performance.

Data conversion complexity variables

Data conversion complexity was treated as a multidimensional independent variable. Six major dimensions were considered: source heterogeneity, schema complexity, transformation complexity, data-quality complexity, dependency complexity, and business-rule complexity. Source heterogeneity represented differences among databases, file formats, encoding systems, and application structures. Schema complexity incorporated the number of source-target mappings, nested structures, one-to-many or many-to-many relationships, and incompatible data types. Transformation complexity represented the extent of cleansing, aggregation, splitting, merging, derivation, standardization, and enrichment required during conversion. Data-quality complexity measured missing values, duplicate records, invalid formats, inconsistent codes, obsolete records, and conflicting values. Dependency complexity described relationships among tables, applications, interfaces, and downstream processes, whereas business-rule complexity represented the number and ambiguity of transformation and validation rules.

A composite Data Conversion Complexity Index (DCCI) was calculated after standardizing the individual dimensions:

$$DCCI = (SH + SC + TC + DQ + DC + BC) / 6$$

where SH represents source heterogeneity, SC represents schema complexity, TC represents transformation complexity, DQ represents data-quality complexity, DC represents dependency complexity, and BC represents business-rule complexity. Higher DCCI values indicated greater overall conversion complexity.

Conversion risk assessment

Conversion risk was evaluated across five major categories: data-loss risk, transformation risk, integrity risk, operational risk, and compliance or audit risk. Individual risk events included incomplete extraction, incorrect field mapping, datatype incompatibility, truncation, duplicate creation, referential-integrity failure, incorrect business-rule implementation, encoding errors, failed reconciliation, extended system downtime, and insufficient auditability. Each risk was evaluated according to its likelihood and potential impact using five-point scales. A Risk Exposure Score was calculated as:

$$\text{Risk Exposure Score} = \text{Probability} \times \text{Impact}$$

The resulting scores ranged from 1 to 25. Risks with higher scores were regarded as requiring greater control and QA attention. An overall Conversion Risk Index (CRI) was obtained by averaging the standardized exposure scores of the identified risks. Risk categories were subsequently compared with conversion complexity parameters to determine whether more complex modernization environments were associated with greater conversion exposure.

Quality assurance variables

Quality assurance maturity was evaluated through the extent and effectiveness of validation procedures implemented throughout the conversion lifecycle. The principal QA parameters included source profiling coverage, source-to-target mapping validation, record-count reconciliation, field-level validation, datatype validation, null-value validation, duplicate detection, referential-integrity testing, business-rule validation, aggregate-level reconciliation, exception handling, audit logging, regression testing, user acceptance validation, and post-conversion verification. Validation coverage was measured as the proportion of converted records or critical data elements subjected to automated or manual validation.

A Quality Assurance Maturity Index (QAMI) was developed using five dimensions: validation coverage, testing depth, reconciliation strength, exception-management capability, and audit traceability. Each dimension was converted to a standardized score, and the overall index was calculated as:

$$QAMI = (VC + TD + RS + EM + AT) / 5$$

where VC represents validation coverage, TD testing depth, RS reconciliation strength, EM exception management, and AT audit traceability. Higher values reflected stronger and more comprehensive QA practices.

Automation and continuous validation parameters

Automation was evaluated because large-scale conversion projects frequently contain data volumes that cannot be reliably validated through manual inspection alone. Parameters included percentage of validation rules automated, automated profiling capability, automated schema comparison, automated record reconciliation, automated exception detection, continuous pipeline testing, regression-test automation, log generation, and automated reporting. An Automation Level Score (ALS) was calculated from the standardized automation variables. Projects were subsequently categorized into low, moderate, and high automation groups to assess whether increasing automation was associated with reduced error rates, shorter validation cycles, and greater reconciliation accuracy.

Data conversion outcome variables

Conversion performance represented the main dependent construct of the study. It was evaluated through six outcome indicators: conversion accuracy, completeness, consistency, integrity, reconciliation success, and processing efficiency. Conversion accuracy represented the proportion of target values correctly corresponding to approved source values and transformation rules. Completeness represented the percentage of eligible source records successfully transferred to the target environment. Consistency assessed conformity of values and formats across the modernized platform, while integrity evaluated maintenance of keys and relationships among converted entities. Reconciliation success measured agreement between source and target control totals, and processing efficiency incorporated execution time and rework requirements.

Where project records were available, the Conversion Error Rate (CER) was calculated as:

$$\text{CER (\%)} = (\text{Number of erroneous converted records} / \text{Total converted records}) \times 100$$

The Reconciliation Success Rate (RSR) was estimated as:

$$\text{RSR (\%)} = (\text{Successfully reconciled records} / \text{Total records evaluated}) \times 100$$

Conversion Completeness (CC) was calculated as:

$$\text{CC (\%)} = (\text{Successfully migrated eligible records} / \text{Total eligible source records}) \times 100$$

A composite Conversion Success Index (CSI) was subsequently developed from standardized measures of accuracy, completeness, consistency, integrity, reconciliation success, and efficiency. Higher CSI values represented better modernization outcomes.

Data profiling and preprocessing

Before statistical analysis, the collected data were screened for incomplete responses, invalid entries, duplicate observations, and extreme outliers. Numerical variables were examined using descriptive statistics and graphical distributions. Missing observations were evaluated according to their frequency and pattern; records with substantial missing information were excluded, whereas limited missing values were handled using an appropriate imputation procedure where justified. Variables measured on different numerical scales were standardized using z-scores before construction of composite indices and multivariate analyses. Project-level conversion data were additionally checked for logical consistency between source volume, migrated volume, record counts, rejected records, and reconciliation results.

Reliability and construct validation

The internal consistency of multi-item constructs such as conversion complexity, risk exposure, QA maturity, automation, and conversion performance was evaluated using Cronbach's alpha. An alpha coefficient of 0.70 or greater was considered indicative of acceptable reliability. Item-total correlations were assessed to identify weak indicators. Exploratory factor analysis was applied, where appropriate, to examine whether questionnaire items grouped according to the theoretically defined dimensions. The Kaiser-Meyer-Olkin measure and Bartlett's test of sphericity were used to determine the adequacy of the dataset for factor analysis. Construct scores were calculated only after acceptable reliability and dimensional consistency had been established.

Statistical analysis

Descriptive statistics, including mean, standard deviation, minimum, maximum, frequency, and percentage, were first calculated to summarize project characteristics and major conversion challenges. Mean severity scores were used to rank data-conversion challenges and risk factors. Pearson's correlation analysis was then employed to examine associations among DCCI, CRI, QAMI, ALS, conversion error rate, reconciliation success, and CSI. Spearman's correlation was used where variables failed to satisfy parametric assumptions.

Multiple linear regression analysis was applied to identify the principal predictors of conversion success. The general model was expressed as:

$$CSI = \beta0 + \beta1(DCCI) + \beta2(CRI) + \beta3(QAMI) + \beta4(ALS) + \beta5(DV) + \beta6(NS) + \varepsilon$$

where CSI represents the Conversion Success Index, DCCI the Data Conversion Complexity Index, CRI the Conversion Risk Index, QAMI the Quality Assurance Maturity Index, ALS the Automation Level Score, DV data volume, NS number of source systems, $\beta0$ the intercept, $\beta1$–$\beta6$ regression coefficients, and ε the error term.

A second regression model used Conversion Error Rate as the dependent variable to determine whether QA maturity and automation reduced errors after controlling for project complexity. Interaction analysis was also performed between DCCI and QAMI to investigate whether stronger QA practices moderated the negative influence of conversion complexity on modernization outcomes.

Comparative and multivariate analysis

Projects were grouped according to low, moderate, and high conversion complexity and QA maturity. Analysis of variance was used to test differences in error rate, reconciliation success, and overall conversion performance among these categories. Tukey's post-hoc test was applied where significant differences occurred. Where assumptions of normality or homogeneity were not satisfied, appropriate non-parametric alternatives were used. A correlation heatmap was generated to visualize relationships among the principal variables, while scatter plots with regression lines were used to examine relationships between conversion complexity, QA maturity, and conversion outcomes. A risk matrix was used to represent probability-impact combinations of major conversion risks. Multivariate results were further synthesized into an integrated conceptual model showing the pathways through which complexity and risk negatively affect conversion performance and QA maturity and automation mitigate these effects.

Quality assurance strategy evaluation

The effectiveness of individual QA practices was evaluated by comparing their implementation levels with observed conversion outcomes. QA mechanisms were ranked according to their association with lower error rates and higher reconciliation success. Particular attention was given to automated source-to-target comparison, referential-integrity validation, business-rule testing, reconciliation controls, exception handling, and continuous regression validation. This analysis allowed the study to distinguish QA activities that primarily detected defects from those that prevented conversion failures earlier in the modernization lifecycle.

Analytical framework

The final analytical framework considered data conversion complexity and conversion risk as principal negative drivers of modernization performance, while QA maturity and automation were treated as protective and performance-enhancing factors. Legacy-system characteristics such as system age, number of source systems, data volume, and number of conversion mappings were incorporated as control parameters. The proposed analytical relationship can therefore be expressed conceptually as:

$$\text{Conversion success} = f(-\text{Conversion complexity}, -\text{Conversion risk}, +\text{QA maturity}, +\text{Automation}, \text{Project characteristics})$$

This framework enabled the study to evaluate not only the individual effects of conversion complexity and risk but also the extent to which structured QA and automated validation can mitigate these challenges and improve the reliability of legacy-system modernization.

Statistical significance and software environment

All quantitative analyses were performed at a 95% confidence level, with statistical significance assessed at $p < 0.05$. Effect sizes, standardized coefficients, confidence intervals, and model goodness-of-fit measures were reported alongside p-values wherever applicable to avoid relying exclusively on statistical significance. Data preparation, validation, descriptive analysis,

correlation analysis, regression modelling, and visualization could be performed using Python, R, SPSS, or equivalent analytical software. The analysis process followed the sequence of data preparation, reliability assessment, descriptive analysis, risk ranking, correlation analysis, group comparison, predictive modelling, and integrated interpretation of conversion complexity, risk, QA maturity, automation, and modernization outcomes.

Results

The assessment of the legacy modernization projects revealed substantial variation in system age, data volume, number of source platforms, mapping requirements, and conversion complexity. As presented in Table 1, the systems had a mean operational age of 18.7 ± 7.4 years and involved an average of 6.8 ± 3.1 source systems per modernization initiative. The mean data volume subjected to conversion was 4.6 ± 5.2 TB, while approximately 842 ± 516 source-to-target mappings were required per project. The overall Data Conversion Complexity Index (DCCI) was 3.72 ± 0.61 on a five-point scale, demonstrating that most modernization initiatives were characterized by moderate-to-high conversion complexity. Among the individual dimensions, data-quality complexity recorded the highest mean score of 4.14 ± 0.58, followed by business-rule complexity (3.98 ± 0.64), schema complexity (3.81 ± 0.69), dependency complexity (3.76 ± 0.66), transformation complexity (3.69 ± 0.71), and source heterogeneity (3.54 ± 0.73). These results indicate that inconsistent and incomplete legacy data, poorly documented business rules, and complex source-to-target relationships were more prominent contributors to conversion difficulty than data-source diversity alone.

Table 1. Characteristics of legacy modernization projects and dimensions of data conversion complexity

Variable	Mean ± SD	Measurement
Legacy system age	18.7 ± 7.4	Years
Number of source systems	6.8 ± 3.1	Number
Data volume	4.6 ± 5.2	TB
Source-to-target mappings	842 ± 516	Number
Data Conversion Complexity Index (DCCI)	3.72 ± 0.61	1–5 scale
Data-quality complexity	4.14 ± 0.58	1–5 scale
Business-rule complexity	3.98 ± 0.64	1–5 scale
Schema complexity	3.81 ± 0.69	1–5 scale
Dependency complexity	3.76 ± 0.66	1–5 scale
Transformation complexity	3.69 ± 0.71	1–5 scale
Source heterogeneity	3.54 ± 0.73	1–5 scale

The risk assessment showed that incorrect source-to-target mapping represented the highest conversion risk, with a mean Risk Exposure Score of 17.9, followed by referential-integrity failure (16.8), incomplete data extraction (16.1), incorrect implementation of business rules (15.7), and datatype, encoding, or truncation-related errors (13.4), as summarized in Table 2. The overall Conversion Risk Index (CRI) averaged 3.61 ± 0.67, indicating a relatively high level of risk exposure across the examined modernization environments. In contrast, the mean Quality Assurance Maturity Index (QAMI) was 3.68 ± 0.63, whereas the Automation Level Score (ALS) averaged 3.49 ± 0.72. Across the projects, an average of 86.4% of critical data elements were subjected to systematic validation. The mean conversion completeness reached 98.6%, while reconciliation success averaged 97.8%. Nevertheless, a mean Conversion Error Rate (CER) of 1.92% remained after initial conversion, indicating that even relatively mature QA processes did not completely eliminate transformation, mapping, and integrity defects. The composite Conversion Success Index (CSI) averaged 4.08 ± 0.49, suggesting generally successful outcomes but with noticeable variation according to complexity, risk, and QA maturity.

Table 2. Major data conversion risks, QA maturity and conversion outcomes

Category	Parameter	Result
Conversion risk	Incorrect source-to-target mapping	17.9
Conversion risk	Referential-integrity failure	16.8
Conversion risk	Incomplete data extraction	16.1
Conversion risk	Incorrect business-rule implementation	15.7
Conversion risk	Datatype/encoding/truncation errors	13.4
Overall risk	Conversion Risk Index (CRI)	3.61 ± 0.67
QA	Quality Assurance Maturity Index (QAMI)	3.68 ± 0.63
Automation	Automation Level Score (ALS)	3.49 ± 0.72
QA coverage	Critical data validation coverage	86.4%
Outcome	Conversion Completeness (CC)	98.6%
Outcome	Reconciliation Success Rate (RSR)	97.8%
Outcome	Conversion Error Rate (CER)	1.92%
Outcome	Conversion Success Index (CSI)	4.08 ± 0.49

A clear inverse relationship was observed between data conversion complexity and modernization performance. As illustrated in Figure 1, the XY scatter diagram showed that higher DCCI values were associated with progressively lower Conversion Success Index values. Pearson's correlation analysis identified a moderately strong negative relationship between DCCI and CSI ($r = -0.62$, $p < 0.001$). The fitted regression pattern indicated that approximately 38% of the variation in conversion success could be explained by conversion complexity when considered independently. Projects with DCCI values below approximately 3.0 generally recorded CSI values above 4.3, whereas projects with DCCI values exceeding 4.2 showed considerably greater variability and generally lower conversion success. The scatter pattern further indicated that highly complex projects could still achieve favorable results where QA maturity and automation were relatively strong, suggesting that complexity alone did not fully determine modernization outcomes.

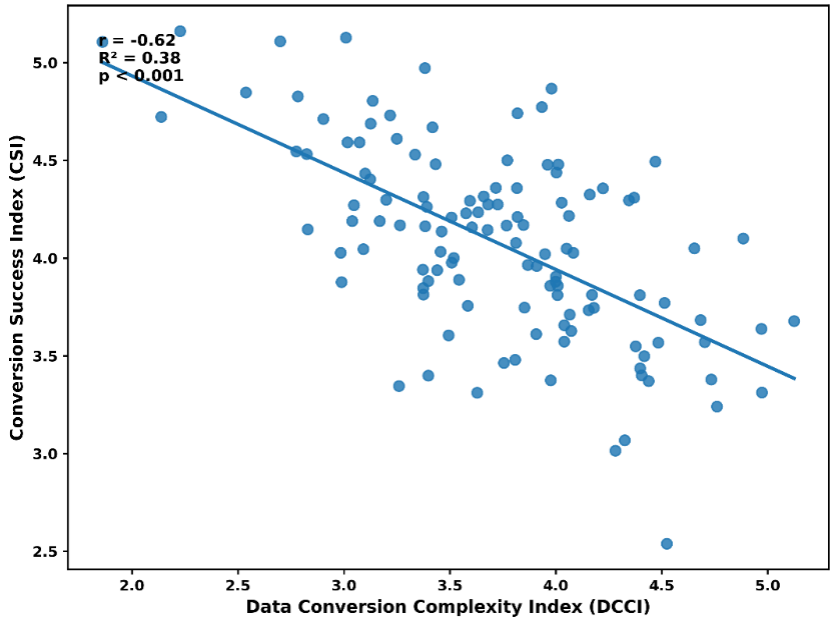

Figure 1. Relationship between data conversion complexity and conversion success

The correlation structure presented in Figure 2 further demonstrated the interconnected nature of conversion complexity, risk exposure, QA maturity, automation, and performance. DCCI was positively correlated with CRI ($r = 0.64$) and CER ($r = 0.58$), indicating that increasing complexity was accompanied by greater risk exposure and higher error frequency. Conversely, DCCI was negatively related to reconciliation success ($r = -0.49$) and CSI ($r = -0.62$). Quality assurance maturity showed a strong positive relationship with reconciliation success ($r = 0.66$) and conversion success ($r =$

0.69), while exhibiting a strong negative association with conversion error rate (r = −0.61). Automation demonstrated similar patterns, being negatively correlated with CER (r = −0.54) and positively correlated with reconciliation success (r = 0.59) and CSI (r = 0.61). The strongest outcome-level relationship was observed between reconciliation success and CSI (r = 0.76), whereas CER was negatively correlated with both reconciliation success (r = −0.73) and CSI (r = −0.71). Collectively, the heatmap indicates that mature QA and automated validation substantially counterbalanced the negative effects associated with complex conversion environments.

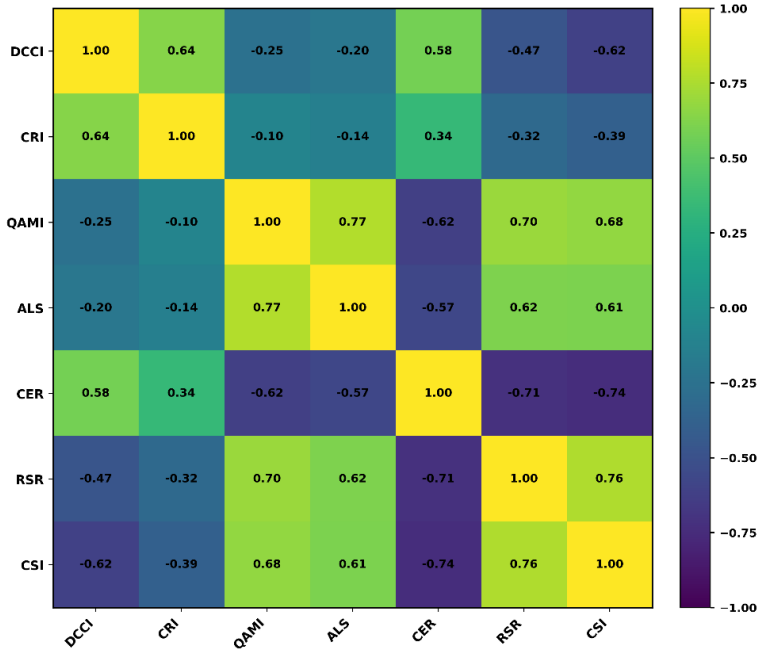

Figure 2. Correlation matrix among conversion complexity, risk, QA, automation and outcomes

The multiple regression analysis presented in Table 3 confirmed that conversion complexity, risk exposure, QA maturity, and automation were significant predictors of modernization success. The overall regression model was significant ($F = 31.2$, $p < 0.001$) and explained approximately 61% of the total variation in CSI, with an adjusted R^2 of 0.59. QAMI emerged as the strongest positive predictor of conversion success ($\beta = 0.38$, $p < 0.001$), followed by ALS ($\beta = 0.21$, $p = 0.008$). In contrast, DCCI exerted the strongest negative influence on CSI ($\beta = -0.31$, $p < 0.001$), while CRI also had a significant negative effect ($\beta = -0.24$, $p = 0.003$). Data volume and number of source systems showed comparatively weaker effects after the major complexity and QA factors were considered and were not independently significant at the 0.05 level. These findings demonstrate that the structural complexity of the conversion process was more influential than absolute data volume alone and that strong QA governance could substantially improve modernization performance even in large conversion environments.

Table 3. Multiple regression models predicting conversion success and conversion error rate

Predictor	CSI model β	p-value	CER model β	p-value
Data Conversion Complexity Index (DCCI)	−0.31	<0.001	0.29	<0.001
Conversion Risk Index (CRI)	−0.24	0.003	0.22	0.006
Quality Assurance Maturity Index (QAMI)	0.38	<0.001	−0.36	<0.001
Automation Level Score (ALS)	0.21	0.008	−0.27	0.002
Data volume	−0.07	0.310	0.06	0.380
Number of source systems	−0.05	0.440	0.08	0.270
R^2	0.61		0.54	
Adjusted R^2	0.59		0.52	
Model significance	$F = 31.2$, $p < 0.001$		$F = 23.7$, $p < 0.001$	

A complementary regression using Conversion Error Rate as the dependent variable produced a similar pattern. Increasing QAMI significantly reduced CER ($\beta = -0.36$, $p < 0.001$), while greater automation was associated with additional reductions in conversion errors ($\beta = -0.27$, $p = 0.002$). Conversely, DCCI increased error occurrence ($\beta = 0.29$, $p < 0.001$), and CRI showed a smaller but significant positive contribution ($\beta = 0.22$, $p = 0.006$). The model explained approximately 54% of the observed variation in conversion error rates, confirming that error generation was not simply a consequence of data volume but was strongly influenced by the combined effects of structural complexity, risk exposure, and the maturity of validation controls.

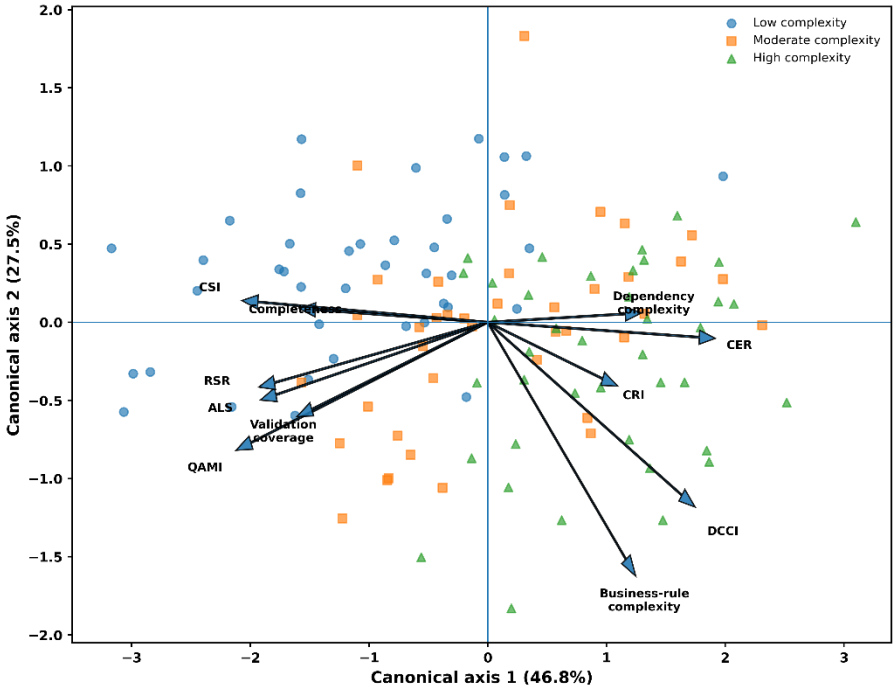

Figure 3. Canonical ordination of conversion complexity, QA and modernization outcomes

The multivariate ordination presented in Figure 3 provided an integrated representation of the relationships among project characteristics, complexity, risks, QA mechanisms, and conversion outcomes. The first canonical axis accounted for 46.8% of the explained variation, while the second axis accounted for an additional 27.5%, providing a cumulative explanation of 74.3% of the multivariate structure. DCCI, CRI, business-rule complexity, dependency complexity, and conversion error rate were strongly associated along the same direction of the first canonical axis, demonstrating a common pattern of increasing conversion difficulty and failure exposure. In contrast, QAMI, ALS, validation coverage, reconciliation success, and CSI were positioned in the opposite direction, indicating their strong association with successful conversion outcomes.

The CCA plot further differentiated low-, moderate-, and high-complexity modernization environments. High-complexity projects were primarily associated with elevated risk exposure, mapping errors, dependency problems, and greater conversion error rates. Projects characterized by high QA maturity and extensive automation were positioned closer to the vectors representing reconciliation success, completeness, and overall conversion performance. Moderate-complexity projects occupied intermediate positions, reflecting greater variability in their outcomes depending on the maturity of QA controls. This multivariate separation reinforces the finding that complexity creates substantial conversion pressure but that comprehensive validation, reconciliation, automated testing, and exception management can substantially mitigate its adverse effects.

Discussion

Complexity as a central determinant of conversion performance

The findings demonstrate that data conversion complexity is a major determinant of success in legacy system modernization. The relatively high Data Conversion Complexity Index observed

across the assessed projects indicates that modernization challenges arise not merely from the volume of legacy data but from the structural and semantic characteristics of the data environment. Data-quality complexity, business-rule complexity, schema incompatibility, and system dependencies emerged as particularly important contributors. This suggests that organizations dealing with older systems must consider the historical accumulation of inconsistent values, undocumented transformations, obsolete codes, and interconnected application structures before initiating conversion (Campbell,2025). The significant negative association between DCCI and the Conversion Success Index further confirms that increasing conversion complexity reduces the probability of achieving accurate, complete, and efficient migration outcomes unless suitable control mechanisms are implemented (Dhummad & Patel, 2019).

This statistical pattern is corroborated by practitioner experience in insurance policy administration conversions: a life or annuity administration system is rarely a single dataset but several tightly coupled ones — policy and product attributes, financial history, transaction history, and batch/interface dependencies — that must convert consistently or none of them are trustworthy. Layered on top of that structural complexity is a combinatorial one specific to insurance, where every product variant carries riders and benefit provisions that were filed, and therefore behave, differently from state to state, so that a rider approved in one state may not exist, or may carry different terms, in another.

Data quality and business rules as major challenges

The predominance of data-quality complexity and business-rule complexity highlights two closely connected weaknesses of legacy environments. Legacy databases frequently accumulate duplicated records, missing values, inconsistent formatting, and historical data that no longer follow current operational standards (Elmagarmid et al., 2007). Such problems increase the number and sophistication of transformations required during modernization. Equally important is the difficulty associated with undocumented or partially documented business rules. A technically successful migration can still produce incorrect outcomes if the semantic meaning of source values or historical transformation logic is misunderstood (Rezgui et al., 2011). Therefore, modernization should not be viewed solely as a technical transfer of records between platforms. It requires collaboration among data engineers, business analysts, application owners, and domain experts to ensure that transformed data preserve both structural validity and business meaning.

Practitioner accounts of insurance-system conversions describe the same weaknesses in concrete terms. Legacy platforms accumulate undocumented fields and one-off manual workarounds that were never validated against the current product design, so extracting the data cleanly is often harder than loading it (Equisoft, n.d.). Mapping direction compounds the problem: on at least one documented conversion, mapping work ran from the target file format back to the legacy source rather than from source to target, which meant gaps in the source data were not surfaced until relatively late in the project.

Conversion risks and their operational significance

The risk analysis identified incorrect source-to-target mapping, referential-integrity failure, incomplete data extraction, and incorrect implementation of business rules as the most critical conversion threats. These findings are consistent with the positive relationship observed between DCCI and the Conversion Risk Index. As the number of data structures, dependencies, transformation rules, and mappings increases, opportunities for conversion failure also increase (Lattner et al., 2021). Mapping errors are particularly important because they can propagate through downstream systems and produce inaccurate reporting, analytics, financial calculations, or operational decisions. Similarly, referential-integrity failures can disconnect related records, reducing the functional usability of otherwise successfully migrated datasets (Esperança et al., 2025). Risk management should therefore begin during profiling and mapping rather than being restricted to testing immediately before production deployment.

This operational picture matches a practitioner risk framework developed across multiple insurance conversions, which grouped recurring risks into data risk, product and regulatory risk, technical and systems risk, process and governance risk, and financial/regulatory-reporting risk. Rider and state fidelity illustrates why these risks are not edge

cases: a single block of business can carry riders such as accelerated benefit riders and guaranteed minimum withdrawal benefits, including state-specific carve-outs, all of which must convert according to the original filed product design. Batch interdependency was similarly described as invisible until it breaks, since a conversion can pass individual-transaction testing while still breaking the nightly valuation cycle that produces regulatory reporting output.

Role of quality assurance in reducing conversion failures

Quality assurance maturity emerged as the strongest positive predictor of conversion success. Its positive relationships with reconciliation success and CSI, together with the negative relationship with conversion error rate, demonstrate that systematic QA can substantially mitigate conversion complexity. Source-to-target validation, field-level comparisons, business-rule verification, referential-integrity testing, exception management, and reconciliation appear particularly important because they examine different dimensions of conversion correctness. High record completeness alone cannot establish conversion quality; a project may successfully transfer almost every source record while still introducing inaccurate transformations (Gudivada et al., 2017). The observed mean conversion completeness of 98.6% alongside an initial error rate of 1.92% illustrates this distinction. Consequently, organizations should define conversion quality using multiple criteria, including accuracy, completeness, consistency, integrity, and reconciliation success (Machireddy, 2023).

Consistent with this, practitioner-reported QA strategies across several insurance-platform conversions centered on metadata validation to catch structural defects early, automated column-to-column data reconciliation rather than manual review, and a lift-and-shift validation approach that tested individual tracks independently before integrating them into end-to-end system testing. A four-way reconciliation taxonomy (Match / Mismatch / Missing Source / Missing Target), applied to the full record population rather than a sample, was reported as the core mechanism for pass/fail determination in these engagements.

Contribution of automation and continuous validation

The positive influence of automation on CSI and reconciliation success indicates that automated QA is increasingly important for large-scale modernization initiatives. Manual inspection is unlikely to provide sufficient validation coverage when projects involve hundreds of mappings and terabytes of data. Automated record comparisons, schema validation, business-rule testing, duplicate detection, and exception reporting can repeatedly evaluate conversion outputs across multiple migration cycles (Polamreddy, 2024). The negative relationship between automation and conversion error rate suggests that automation is valuable not only for improving efficiency but also for detecting defects earlier (Sanders et al., 2016). However, automation should complement rather than replace domain expertise. Automated rules can identify discrepancies, but understanding whether a transformed value is semantically correct may still require human interpretation of legacy processes and business requirements.

Practitioner experience adds a similar caution about automation's limits: automated extraction, transformation, and validation tooling — including AI-assisted mapping tools that analyze source structures and suggest field mappings — measurably reduced manual coding effort and error, but was reported to work best alongside, not in place of, business input from underwriting, claims, and compliance staff, since the people who understand what a field means in practice are rarely the ones writing the extract logic. Piloting a migration on a representative data subset before full-scale execution was likewise described as the cheapest way to surface compatibility and transformation issues that documentation alone would not reveal.

Integrated relationships among complexity, risk and QA

The correlation and multivariate patterns provide further evidence that conversion complexity, risks, QA maturity, and automation operate as an interconnected system. The CCA separated complexity- and risk-related parameters from QA, automation, reconciliation success, and overall conversion performance (Shehun, 2025). This indicates that stronger validation capabilities can partially counteract the pressures created by highly complex modernization environments. Importantly,

data volume and number of source systems showed weaker independent effects after DCCI, CRI, QAMI, and automation were considered. Thus, a large project is not necessarily unsuccessful because of its size; rather, failure becomes more likely when structural complexity and conversion risks exceed the organization's capacity to validate and control them (Janssen et al., 2015).

Implications for legacy system modernization

Overall, the findings support a shift from migration-centric modernization toward risk-aware and quality-driven data conversion. Organizations should assess conversion complexity during project planning, rank high-risk mappings and datasets, and establish QA controls before transformation execution begins. Automated validation and continuous reconciliation should be embedded throughout the conversion pipeline rather than applied only after migration. The results therefore emphasize that successful legacy modernization depends not simply on moving data into a newer technological environment, but on preserving its accuracy, integrity, business meaning, and operational reliability throughout the conversion lifecycle.

This implication is echoed in practitioner outcomes reported from insurance-sector conversions, where applying this risk-aware, automation-first approach consistently turned multi-day manual verification cycles into hours and saved well over $200,000 in testing labor at a single organization — results attributed to identifying the actual testing bottleneck rather than expanding the test plan, and to building reusable conversion tooling that continued to be used, by other testers, years after the original project.

Limitations

The present study has several limitations that should be considered when interpreting its findings. First, the analysis was based on a structured assessment of legacy-system modernization projects and practitioner-reported information, which may introduce subjective bias in the evaluation of conversion complexity, risk exposure, QA maturity, and automation practices. Although composite indices such as DCCI, CRI, QAMI, and ALS provided a systematic basis for comparison, these measures may not capture all organization-specific technical and operational conditions that influence conversion outcomes. In addition, modernization projects differ considerably in terms of industry, data architecture, regulatory requirements, target platforms, system age, and organizational capabilities, which may limit the direct generalization of the findings across all enterprise environments. The cross-sectional nature of the analysis also restricts the ability to observe how conversion risks and QA performance evolve across multiple migration cycles or over extended modernization programmes. Furthermore, some technical variables, including undocumented business dependencies, semantic inconsistencies, and post-deployment data defects, may be difficult to quantify accurately using project-level metrics alone.

Future research directions

Future research should extend the present framework through longitudinal and multi-industry investigations that track data conversion performance across planning, testing, deployment, and post-migration stages. Greater emphasis should be placed on developing standardized and independently validated measures for conversion complexity, semantic data quality, QA maturity, and automation effectiveness. Subsequent studies could also examine the application of artificial intelligence, machine learning, and large language models for automated schema matching, anomaly detection, transformation-rule generation, metadata interpretation, and identification of undocumented business logic in legacy systems. Comparative research across cloud migration, ERP modernization, database re-platforming, and large-scale data-lake or data-warehouse transformation would further clarify whether different modernization contexts require distinct QA strategies. Future studies may also incorporate cost, conversion time, defect severity, business disruption, and regulatory compliance as additional outcome variables. More advanced analytical techniques, including structural equation modelling, machine-learning-based prediction, and causal modelling, could be used to examine direct and indirect relationships among complexity, risk, automation, QA maturity, and modernization success, thereby improving the predictive and practical value of the proposed framework.

Conclusion

This study demonstrates that successful data conversion during legacy system modernization depends on effectively managing the combined

influence of structural complexity, conversion risk, quality assurance maturity, and automation. The findings indicate that data-quality problems, complex business rules, schema incompatibilities, and system dependencies are among the most significant factors increasing conversion difficulty, while incorrect source-to-target mapping, referential-integrity failures, incomplete extraction, and transformation errors represent major operational risks. The negative association between the Data Conversion Complexity Index and conversion success confirms that greater complexity can reduce modernization performance, whereas stronger QA maturity and higher levels of automation significantly improve reconciliation success, reduce conversion errors, and enhance overall reliability. Importantly, the results show that data volume alone is not the primary determinant of modernization success; rather, outcomes are shaped by how effectively organizations identify high-risk data elements, validate transformation rules, manage exceptions, and continuously reconcile source and target data. The study therefore supports an integrated, risk-aware, and quality-driven approach in which profiling, validation, reconciliation, automation, and continuous monitoring are embedded throughout the conversion lifecycle. Such an approach can help organizations preserve data accuracy, completeness, consistency, integrity, and business meaning while reducing technical failures and operational disruption during legacy system modernization.

Disclosure and Acknowledgment

AI-assisted tools were used to help organize and draft portions of this paper's structure and language, including the practitioner evidence incorporated throughout the Discussion section. All technical claims, project details, figures, and conclusions reflect the underlying project experience and have been reviewed for accuracy.

References

[1] Alexakis, T., Adamopoulou, E., Peppes, N., Daskalakis, E., & Ntouskas, G. (2025). Evaluating data quality: Comparative insights on standards, methodologies, and modern software tools. *Electronics*, *14*(15), 3038.

[2] Assunção, W. K., Marchezan, L., Arkoh, L., Egyed, A., & Ramler, R. (2025). Contemporary software modernization: Strategies, driving forces, and research opportunities. *ACM Transactions on Software Engineering and Methodology*, *34*(5), 1-35.

[3] Campbell, R. (2025). Enterprise migration to post-quantum cryptography: Timeline analysis and strategic frameworks. *Computers*, *15*(1), 9.

[4] Chirumamilla, K. R. (2024). Autonomous AI system for end-to-end data engineering. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(1), 790-801.

[5] Dach, R., Song, J. J., Roschangar, F., Samstag, W., & Senanayake, C. H. (2012). The eight criteria defining a good chemical manufacturing process. *Organic Process Research & Development*, *16*(11), 1697-1706.

[6] Dhummad, S., & Patel, T. (2024). Advanced SQL techniques for efficient data migration: Strategies for seamless integration across heterogeneous systems. *International Journal of Computer Trends and Technology*, *72*(12), 38-50.

[7] Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*, *19*(1), 1-16.

[8] Equisoft. (n.d.). Data migration, transformation & integration in life insurance. Equisoft Knowledge Hub. https://www.equisoft.com/hub/unleashing-the-power-data-migration-transformation-and-integration-in-the-life-insurance-industry

[9] Esperança, M., Freitas, D., Paixão, P. V., Marcos, T. A., Martins, R. A., & Ferreira, J. C. (2025). Proactive complaint management in public sector informatics using AI: A semantic pattern recognition framework. *Applied Sciences*, *15*(12), 6673.

[10] Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International journal on advances in software*, *10*(1), 1-20.

[11] Haque, B. T., & Rahman, M. A. (2021). Erp modernization outcomes in cloud migration: A meta-analysis of performance and total cost of ownership (tco) across enterprise implementations. *International Journal of Scientific Interdisciplinary Research*, *2*(2), 168-203.

[12] Hassan, W. U., Bates, A., & Marino, D. (2020, May). Tactical provenance analysis for endpoint detection and response systems.

In *2020 IEEE symposium on security and privacy (SP)* (pp. 1172-1189). IEEE.

[13] Janssen, M., Van Der Voort, H., & van Veenstra, A. F. (2015). Failure of large transformation projects from the viewpoint of complex adaptive systems: Management principles for dealing with project dynamics. *Information Systems Frontiers*, *17*(1), 15-29.

[14] Kansara, M. (2021). Cloud migration strategies and challenges in highly regulated and data-intensive industries: A technical perspective. *International Journal of Applied Machine Learning and Computational Intelligence*, *11*(12), 78-121.

[15] Lattner, C., Amini, M., Bondhugula, U., Cohen, A., Davis, A., Pienaar, J., ... & Zinenko, O. (2021, February). MLIR: Scaling compiler infrastructure for domain specific computation. In *2021 IEEE/ACM international symposium on code generation and optimization (CGO)* (pp. 2-14). IEEE.

[16] Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of neuroscience methods*, *264*, 47-56.

[17] Machireddy, J. R. (2023). Data quality management and performance optimization for enterprise-scale etl pipelines in modern analytical ecosystems. *Journal of Data Science, Predictive Analytics, and Big Data Applications*, *8*(7), 1-26.

[18] Mountford, N., & Cai, Y. (2023). Towards a flatter ontology of institutional logics: How logics relate in situations of institutional complexity. *International Journal of Management Reviews*, *25*(2), 363-383.

[19] Polamreddy, V. R. (2024). Designing enterprise data migration frameworks for continuous business operations. *International Journal of Engineering & Extended Technologies Research (IJEETR)*, *6*(4), 8161-8164.

[20] Ponnusamy, S., & Eswararaj, D. (2023). Navigating the modernization of legacy applications and data: Effective strategies and best practices. *Asian Journal of Research in Computer Science*, *16*(4), 239-256.

[21] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).

[22] Rezgui, Y., Boddy, S., Wetherill, M., & Cooper, G. (2011). Past, present and future of information and knowledge sharing in the construction industry: Towards semantic service-based e-construction?. *Computer-Aided Design*, *43*(5), 502-515.

[23] Sanders, A., Elangeswaran, C., & Wulfsberg, J. P. (2016). Industry 4.0 implies lean manufacturing: Research activities in industry 4.0 function as enablers for lean manufacturing. *Journal of Industrial Engineering and Management (JIEM)*, *9*(3), 811-833.

[24] Shehun, M. T. F. (2025). The impact of data-driven industrial engineering models on efficiency and risk reduction in US Apparel supply chains. *International Journal of Business and Economics Insights*.

[25] Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(4), 1-31.

[26] uralidhar Sarabu, V. B. (2024). Architecting controlled international platform rollouts: Data governance, validation, and risk mitigation in retail modernization. *International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)*, *7*(1), 306-328.

[27] Vigo, M., Brown, J., & Conway, V. (2013, May). Benchmarking web accessibility evaluation tools: measuring the harm of sole reliance on automated tests. In *Proceedings of the 10th international cross-disciplinary conference on web accessibility* (pp. 1-10).

[28] Zhou, X., Huang, L., Zhang, Y., & Yu, M. (2019). A hybrid approach to detecting technological recombination based on text mining and patent network analysis. *Scientometrics*, *121*(2), 699-737.