**Sciences du Nord Nature Science and Technology**
**Vol. 2, No. 02, December 2025: pp. 87-104**
**E-ISSN: 3046-6504**
**P-ISSN: 3047-4922**
**DOI: 10.71238/snnst.v2i02.197**

# Modernizing Enterprise Data Governance with AI-Powered Lineage Intelligence, End-to-End Data Traceability, and Automated Metadata Management in SAP Datasphere

**Rajendra Asuri** [1]
[1] Senior BTP Consultant Asurirajendra.sap@gmail.com

**ARTICLE HISTORY**
**Received Oct, 2025** **Revised Nov, 2025** **Accepted Dec, 2025**

**ABSTRACT**

Trust becomes difficult to sustain when organizations cannot explain data origins, transformations, or consumption. This study explores the modernization of enterprise data governance through AI-powered lineage intelligence, end-to-end traceability, and automated metadata management in SAP Datasphere. The research problem concerns disconnected metadata, opaque data movement, labor-intensive impact assessment, and inconsistent accountability across systems. A mixed-method design combines measures of lineage coverage, metadata completeness, dependency-detection accuracy, and governance response time with qualitative evidence from data stewards, architects, analysts, and compliance specialists. Results show that intelligent lineage discovery clarifies data relationships, exposes hidden dependencies, shortens impact analysis, and improves readiness for audits and regulatory reviews. A unified governance framework is introduced to connect metadata harvesting, semantic modeling, lineage visualization, policy monitoring, stewardship workflows, and human verification. This framework advances practice by linking technical data flows with business definitions, ownership responsibilities, and governance controls. Practical implications include stronger data quality, reliable analytical assets, responsible AI adoption, and faster decision-making. Academic significance arises from positioning lineage intelligence as a component of governance maturity. The study concludes that SAP Datasphere can support transparent and accountable data ecosystems when automated intelligence operates alongside human stewardship, offering a scalable governance direction for regulated industries and data-intensive enterprises.

***Keywords:*** *AI-Powered Data Lineage; Automated Metadata Management; End-to-End Data Traceability; Enterprise Data Governance; SAP Datasphere*

## INTRODUCTION

Trusted data has become a strategic requirement for organizations that depend on analytics, enterprise planning, regulatory reporting, and artificial intelligence [1]. Business decisions increasingly rely on information assembled from operational applications, cloud platforms, data warehouses, external sources, and analytical environments. Data accuracy alone cannot establish confidence unless organizations can also explain its origin, meaning, ownership, transformation history, and intended use. Reliable governance must therefore connect data quality, business semantics, stewardship, security, and traceability throughout the information lifecycle [2]. Such integration enables organizations to treat data as an accountable enterprise asset rather than an isolated technical resource [3].

Fragmented metadata remains a persistent obstacle to achieving this level of trust [4]. Technical descriptions, business definitions, ownership records, quality indicators, and policy

classifications are frequently distributed across disconnected repositories and organizational units. Inconsistent naming conventions and duplicated semantic models further weaken the shared understanding of critical information [5]. Governance teams often struggle to determine which assets exist, how they relate to one another, and whether established controls have been applied consistently [6]. These conditions create uncertainty across reporting, analytics, compliance assessment, and enterprise decision-making.

Manual lineage documentation offers limited support for complex and continuously changing data landscapes [7]. Spreadsheets, architecture diagrams, catalog entries, and individual knowledge can describe selected data flows, but these records often become incomplete or outdated when sources, transformations, or analytical models change. Detailed dependency analysis becomes especially difficult when data passes through multiple integration, preparation, semantic modeling, and consumption stages. Governance investigations consequently require considerable effort to identify upstream origins and downstream effects. Automated lineage discovery provides a more scalable basis for maintaining accurate and timely representations of data movement [8].

SAP Datasphere provides an enterprise context for combining data integration, semantic modeling, cataloging, governed access, and analytical consumption. Its ability to connect distributed data while preserving business meaning supports governance across heterogeneous landscapes [9]. Metadata collected from connections, objects, transformations, models, and consuming assets can establish a foundation for lineage analysis and impact assessment. Business semantics can further connect technical structures with definitions, responsibilities, and organizational policies. Effective governance within this environment nevertheless requires a coordinated framework that extends beyond technical connectivity and isolated metadata collection.

Existing research explains data governance frameworks, metadata management, data catalogs, data quality, provenance, and decentralized data architectures from several perspectives [10], [11], [12]. Limited attention has been directed toward integrating these elements into a unified governance framework specifically aligned with SAP Datasphere. Current literature also provides insufficient analysis of how intelligent lineage discovery can connect technical dependencies with business meaning, ownership, policy controls, and stewardship activities. This knowledge gap restricts the practical application of automated metadata intelligence in enterprise governance programs. The central research problem concerns how SAP Datasphere can support transparent, scalable, and accountable governance across the complete data lifecycle.

This study aims to develop an integrated framework for modernizing enterprise data governance through AI-powered lineage intelligence, end-to-end data traceability, and automated metadata management. The first objective is to define the relationships among metadata discovery, semantic enrichment, lineage capture, impact analysis, and governance control. The second objective is to examine how automation can identify dependencies, incomplete metadata, undocumented transformations, and governance risks. The third objective is to establish appropriate responsibilities for data owners, stewards, architects, analysts, and compliance professionals. The final objective is to evaluate how the proposed framework can improve lineage coverage, metadata completeness, governance responsiveness, auditability, and confidence in enterprise data.

Three research questions guide the investigation. The first question examines how automated metadata discovery and lineage intelligence can improve source-to-consumption visibility within SAP Datasphere. The second question considers how technical lineage can be

connected with business definitions, ownership responsibilities, quality expectations, and governance policies. The third question evaluates the extent to which the proposed framework can improve dependency detection, change-impact analysis, stewardship coordination, and governance response. The study contributes a structured approach that combines architectural, operational, semantic, and organizational dimensions of data governance.

The paper focuses on enterprise data governed through SAP Datasphere and connected analytical landscapes. The analysis covers governance foundations, research design, metadata intelligence architecture, automated lineage discovery, operational control, stewardship validation, and comparative evaluation. Results are examined through measures associated with lineage coverage, metadata completeness, dependency identification, impact-analysis efficiency, and governance responsiveness. Strategic implications address trusted analytics, regulatory readiness, responsible artificial intelligence, data product accountability, and organizational governance maturity. The concluding discussion consolidates the findings, acknowledges implementation limitations, and identifies opportunities for broader enterprise validation.

## LITERATURE REVIEW

### *2.1 Foundations of Intelligent Data Governance and Lineage Management*

Enterprise data governance has evolved from a compliance-oriented activity into a strategic discipline that coordinates how data is defined, controlled, shared, protected, and used [13]. Earlier governance practices concentrated mainly on standards, access permissions, data quality rules, and organizational responsibilities within individual systems. Expanding analytical landscapes introduced distributed sources, cloud services, semantic models, and cross-functional data consumption, requiring broader governance mechanisms [14]. Modern governance therefore combines organizational accountability with metadata-driven visibility and technology-enabled control [15]. This progression supports a continuous governance model capable of responding to changing data relationships, business requirements, and regulatory obligations.

Metadata provides the descriptive foundation required to understand and govern enterprise information throughout its lifecycle. Business metadata explains terminology, ownership, business rules, calculation logic, and accepted purposes of use. Technical metadata describes schemas, tables, columns, data types, connections, transformations, and structural relationships. Operational metadata records execution status, processing frequency, usage patterns, refresh history, and system activity, while governance metadata captures classifications, policies, quality expectations, access conditions, and stewardship responsibilities. Coordinating these metadata categories creates a multidimensional representation of data that serves technical specialists, business users, governance teams, and compliance professionals [16].

Data lineage describes the movement and transformation of information across sources, integration processes, models, and consuming applications. Provenance provides supporting evidence about data origin, processing history, responsible parties, and contextual conditions affecting its creation or modification [17]. Traceability connects these concepts by enabling authorized users to follow data from its initial source to its final analytical or operational use. Complete traceability supports root-cause investigation, dependency analysis, change assessment, audit preparation, and verification of reported information. Lineage becomes valuable for

governance when technical data flows are interpreted alongside business definitions, ownership, policies, and quality expectations.

Clear ownership and active stewardship establish accountability for the meaning, condition, protection, and appropriate use of enterprise data. Data owners provide decision authority, approve governance expectations, and determine acceptable access and usage conditions for defined data domains. Data stewards translate these expectations into metadata standards, quality controls, classifications, issue-resolution procedures, and operational guidance. Architects, engineers, analysts, and compliance specialists contribute complementary expertise for maintaining technical integrity and regulatory alignment. Coordinated responsibilities prevent governance from becoming an isolated administrative function and embed accountability within regular data management activities [18].

Semantic modeling strengthens governance by connecting physical data structures with recognizable business concepts and analytical meaning. Raw tables and technical attributes have limited value when users cannot relate them to customers, products, transactions, organizational units, or performance measures. Semantic definitions establish consistent relationships, measures, hierarchies, associations, and contextual descriptions across analytical assets. SAP Datasphere supports this alignment by enabling data integration and modeling while preserving relevant business context across connected environments. Business-aware metadata consequently improves discovery, interpretation, reuse, and confidence without requiring every consumer to understand underlying technical implementations.

Conventional metadata catalogs commonly provide searchable inventories, descriptions, classifications, ownership fields, and basic relationship views [19]. Their effectiveness declines when metadata collection depends heavily on manual entry, irregular review, and voluntary maintenance by distributed teams. Static catalog records may fail to reflect recent schema modifications, revised transformation logic, changing dependencies, or emerging governance risks. Disconnected cataloging and processing environments can also separate documented meaning from actual data movement and operational behavior. These limitations reduce catalog reliability and increase the effort required to validate whether metadata accurately represents the governed landscape.

AI-supported metadata discovery can strengthen cataloging by identifying patterns, classifications, semantic similarities, potential relationships, and incomplete descriptions across large collections of assets. Automated harvesting can capture technical and operational metadata directly from connections, models, transformations, and processing activities [20]. Intelligent assistance can propose business terms, detect related objects, highlight unusual dependencies, and prioritize records requiring stewardship attention. Human review remains essential because inferred relationships and classifications may be incomplete, context-dependent, or unsuitable for sensitive governance decisions. Effective implementation therefore combines automated discovery with explicit confidence assessment, approval workflows, and accountable stewardship.

The conceptual foundation of this study integrates metadata intelligence, semantic modeling, lineage analysis, governance control, and human accountability within the SAP Datasphere context. SAP Datasphere functions as the central environment connecting data sources, transformation processes, business models, governance responsibilities, and analytical consumption. Metadata discovery supplies descriptive evidence, while lineage intelligence and impact analysis reveal movement, dependency, and change consequences. Policy control and

stewardship workflows connect technical observations with ownership, quality, compliance, and trusted use. The figure presents this ecosystem as a coordinated relationship between metadata categories, governance capabilities, lifecycle stages, and enterprise outcomes.

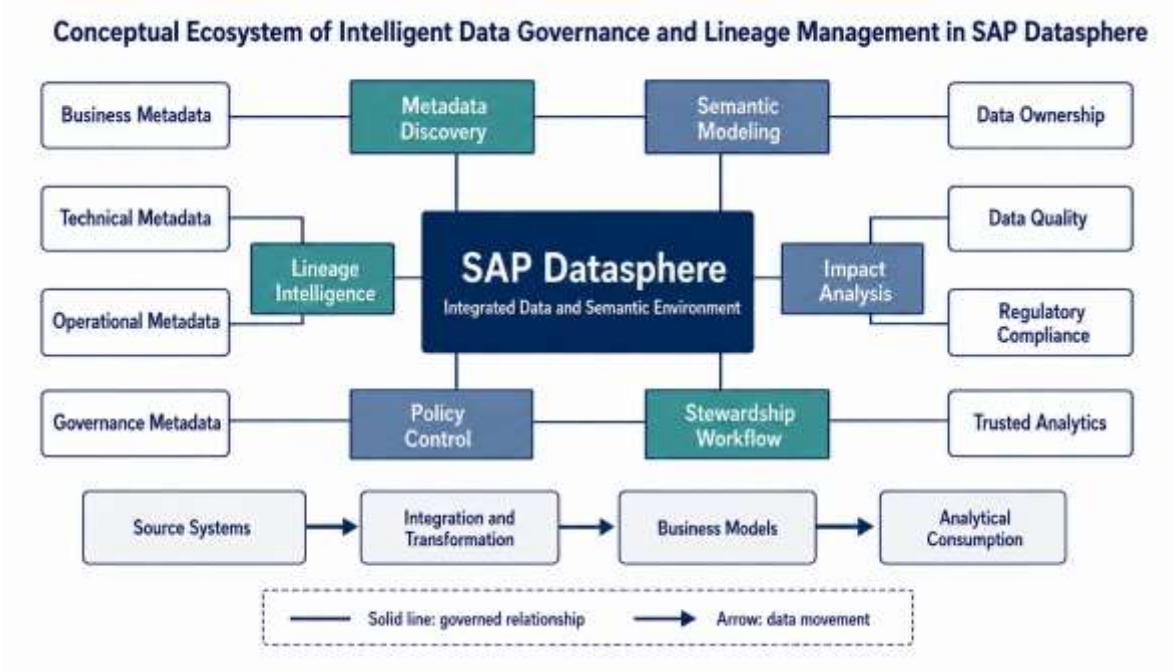



**Figure 1. Enterprise Evolution of SAP SRM Toward Clean Core Architecture**

## METHOD

### *3.1 Research Design and Enterprise Governance Assessment*

A mixed-method research design was selected to examine both measurable governance performance and the organizational conditions influencing its effectiveness. Quantitative analysis evaluates changes in lineage coverage, metadata completeness, dependency detection, and governance response, while qualitative inquiry captures practitioner experience and contextual interpretation. Parallel evidence streams reduce dependence on a single measurement approach and provide a more balanced assessment of the proposed framework. Evidence integration occurs through triangulation, consistency checking, and expert verification before comparative conclusions are established [21]. This design supports a detailed investigation of technical outcomes, governance practices, human responsibilities, and enterprise readiness.

The enterprise assessment scenario represents a heterogeneous data landscape connected through SAP Datasphere for governed integration, semantic modeling, and analytical consumption. Source environments include operational applications, relational databases, cloud services, prepared datasets, transformation processes, business models, and reporting assets. Participating roles comprise data owners, data stewards, enterprise architects, data engineers, analysts, and compliance professionals. Assessment boundaries extend from source registration and metadata acquisition to transformation analysis, semantic interpretation, policy review, and downstream usage. The scenario provides a controlled basis for examining how intelligent lineage and automated metadata management operate across interconnected enterprise assets.

Quantitative evaluation measures were defined to assess governance performance through consistent and interpretable indicators. Lineage coverage measures the proportion of relevant assets

and relationships represented across source-to-consumption pathways. Metadata completeness examines the availability of required business, technical, operational, and governance attributes for selected data objects [22]. Dependency detection accuracy evaluates whether upstream and downstream relationships are identified correctly, while governance response time records the effort required to investigate and address identified issues. Additional measures include impact-analysis duration, unresolved metadata exceptions, ownership assignment coverage, and policy-control completion.

Qualitative practitioner assessment captures insights that numerical indicators cannot adequately explain. Structured interviews with data stewards examine metadata usability, ownership clarity, issue prioritization, and the practical value of automated suggestions. Architect reviews assess structural consistency, integration dependencies, modeling alignment, and the interpretability of lineage relationships. Analyst observations explore data discovery, business-context comprehension, confidence in analytical assets, and the effort required to validate information. Compliance evaluations consider traceability, control evidence, policy visibility, accountability, and readiness for internal or external review.

Governance maturity is evaluated across the interconnected dimensions of people, process, metadata, technology, and control [23]. The people dimension considers role clarity, stewardship capability, executive sponsorship, collaboration, and accountability. Process maturity covers governance workflows, issue resolution, change management, approval procedures, and continuous monitoring. Metadata and technology dimensions examine descriptive completeness, semantic consistency, automation, integration, lineage capture, and analytical accessibility. Control maturity evaluates policy enforcement, access oversight, quality validation, audit evidence, risk identification, and corrective action.

Baseline and proposed-state comparison provides the principal structure for interpreting governance improvement. The baseline represents conventional practices characterized by fragmented metadata, manual lineage documentation, disconnected responsibilities, and reactive issue investigation. The proposed state applies automated metadata discovery, intelligent lineage analysis, semantic enrichment, impact assessment, policy alignment, and stewardship validation. Identical evaluation measures are used across both states to support a consistent comparison of performance and governance maturity. Observed differences are interpreted according to their operational significance rather than relying only on percentage improvement.

Data collection combines system-derived evidence, structured assessment records, practitioner observations, and governance documentation. Metadata inventories, lineage relationships, transformation records, issue logs, ownership assignments, quality indicators, and response durations form the quantitative dataset. Interview guides, review templates, observation records, and compliance checklists provide standardized qualitative evidence. Validation procedures include duplicate removal, completeness checking, relationship verification, metric normalization, and comparison with expert-confirmed reference records. Conflicting findings are reviewed through evidence triangulation and documented resolution decisions before inclusion in the final analysis.

Reliability is supported through consistent measurement definitions, repeatable assessment procedures, standardized instruments, and transparent calculation rules. Reproducibility requires documentation of scenario boundaries, selected assets, metadata requirements, evaluation formulas,

validation decisions, and comparison conditions. Limitations include dependence on scenario-based evidence, variations in organizational maturity, incomplete source metadata, and potential subjectivity within practitioner judgments. Results should therefore be interpreted as evidence of framework feasibility rather than universal performance claims across every enterprise environment. The figure summarizes the complete mixed-method design, evidence paths, validation procedures, maturity dimensions, and comparative assessment sequence.

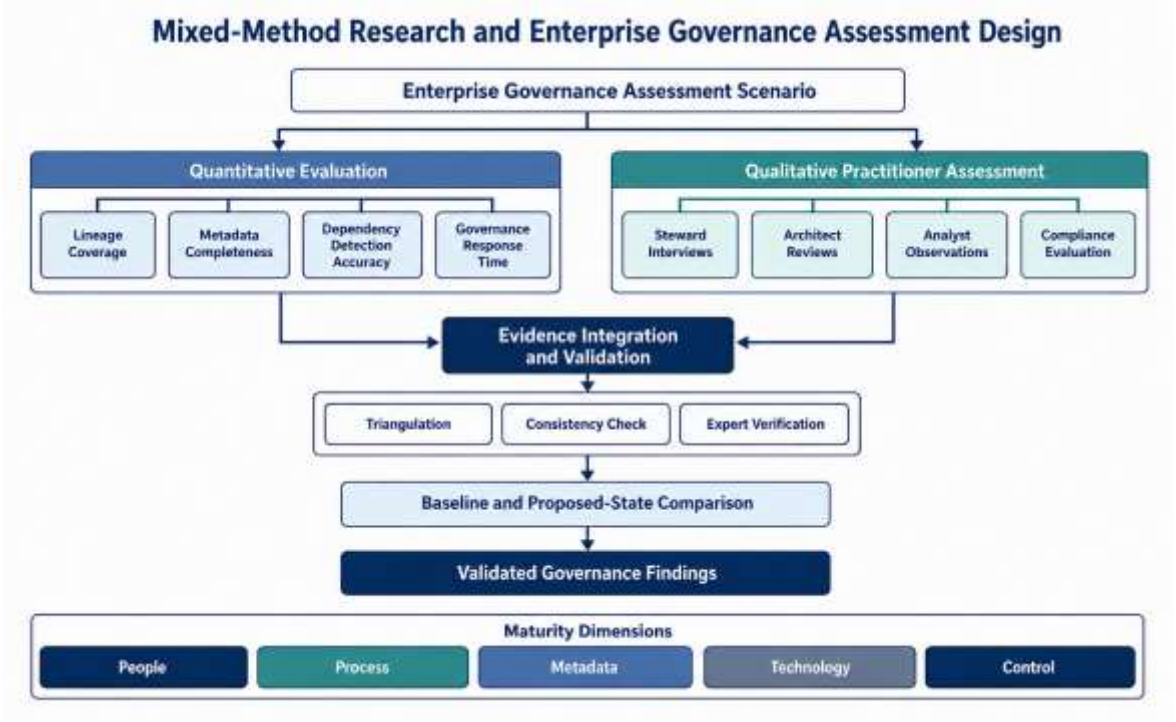



**Figure 2. Mixed-Method Research and Enterprise Governance Assessment Design**

### *3.2 Integrated Governance and Metadata Intelligence Framework*

The proposed framework establishes governance as a continuous capability embedded within enterprise data operations. Its principles combine accountability, transparency, semantic consistency, interoperability, automation, and controlled human oversight. Metadata remains connected to data assets throughout ingestion, transformation, modeling, sharing, and analytical consumption. Risk-based controls apply stronger validation to sensitive, regulated, or operationally important information resources. This approach enables SAP Datasphere to support trusted data use across distributed business and technical environments [24].

The architectural design consists of enterprise source, metadata harvesting, semantic enrichment, lineage intelligence, governance control, and trusted data product layers. Each layer performs a defined responsibility while exchanging contextual metadata through governed relationships and interfaces. Enterprise sources provide structural, operational, and processing information for classification and dependency analysis. Cross-layer controls maintain access conditions, ownership assignments, quality expectations, retention rules, and audit evidence. This layered arrangement separates technical responsibilities while preserving an integrated enterprise governance process.

Automated metadata harvesting captures schemas, tables, columns, transformations, schedules, interfaces, and processing changes from connected systems. Scheduled extraction reduces

dependence on manually maintained inventories that frequently become incomplete or outdated. Change capture identifies added assets, altered structures, removed fields, and modified processing dependencies. Metadata quality checks evaluate completeness, consistency, uniqueness, timeliness, and conformity before catalog registration. The resulting inventory provides a current technical foundation for governance, discovery, and lineage intelligence.

Semantic enrichment converts harvested technical descriptions into contextual information that business practitioners can interpret. Business terms, classifications, domains, relationships, and definitions are associated with relevant physical and logical assets. Pattern analysis and similarity assessment recommend connections between technical fields and approved business vocabulary. Data stewards review proposed mappings to prevent ambiguous definitions or unsuitable classifications from entering the governed environment. Semantic context connects enterprise meaning with the structures used to produce reports, models, and analytical measures.

Integrated lineage intelligence combines technical processing paths with the business purposes surrounding enterprise information. Technical lineage records movement across sources, transformations, views, models, interfaces, and downstream analytical products. Business lineage explains how governed terms, performance indicators, policies, and decisions depend on those processing relationships. Dependency mapping supports impact analysis by identifying assets and consumers affected by structural or operational changes. Combined traceability allows practitioners to move between implementation details and business interpretations.

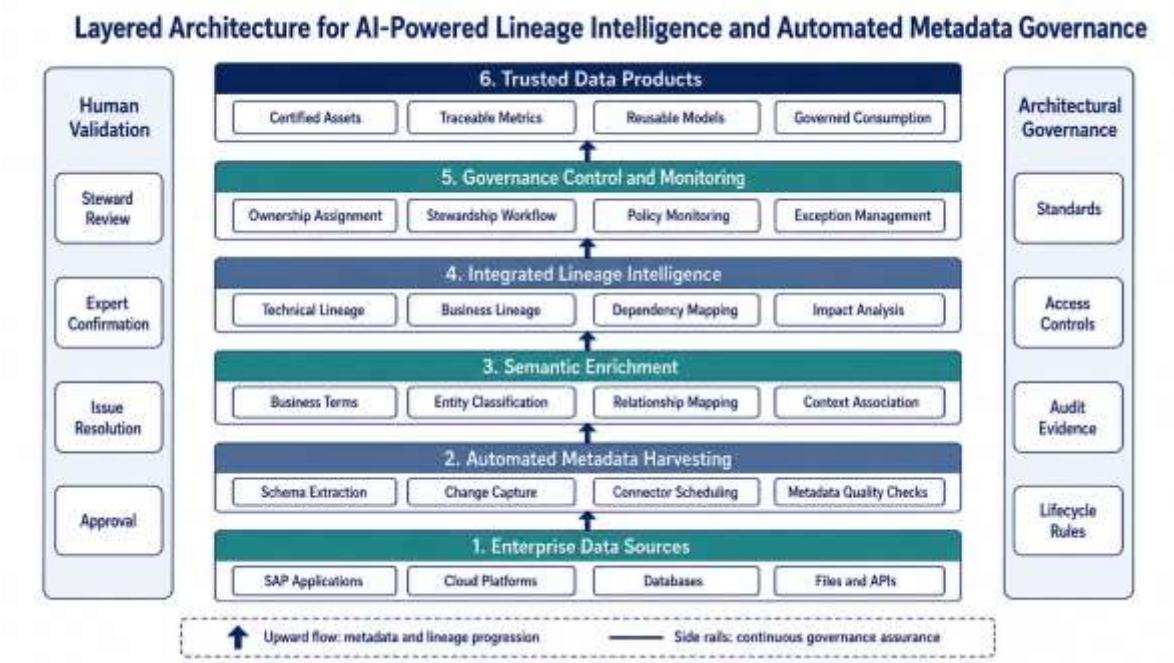



**Figure 3. Layered Architecture for AI-Powered Lineage Intelligence and Automated Metadata Governance**

Ownership and stewardship controls assign accountable participants to data domains, definitions, policies, products, and identified exceptions. Owners authorize governance requirements, while stewards maintain classifications, quality rules, definitions, and supporting evidence. Workflow controls route certification requests, proposed changes, unresolved issues, and policy deviations to appropriate reviewers. Role-based responsibilities clarify approval authority

and corrective action across technical and organizational boundaries. Human accountability remains essential when automated recommendations require business, operational, or regulatory interpretation.

Policy monitoring evaluates governed assets against ownership, access, classification, quality, retention, and certification requirements. Detected violations generate traceable exceptions containing affected assets, policy references, severity levels, and remediation responsibilities. Exception management supports prioritization, escalation, investigation, resolution tracking, evidence collection, and approved risk acceptance. Lineage-aware monitoring reveals downstream products that may inherit risk from an upstream failure or unauthorized modification. Continuous control evaluation improves governance responsiveness while preserving documented professional judgment.

Trusted data products emerge when automated evidence is combined with steward review, expert confirmation, and formal approval. Human validation examines semantic accuracy, lineage completeness, policy interpretation, business relevance, and unresolved exceptions. Certified products provide traceable measures, reusable models, documented ownership, quality expectations, and governed consumption conditions [25]. Validation outcomes return to metadata processes so classifications, relationships, controls, and recommendations can improve. The framework connects metadata intelligence with accountable governance and reliable enterprise data consumption.

**Table 1. Governance Framework Components, Responsibilities, Controls, and Expected Enterprise Outcomes**

| Component | Responsibility | Main control | Expected outcome |
|---|---|---|---|
| Data sources | Supply enterprise data | Source registration | Visible and governed sources |
| Metadata harvesting | Collect technical details | Scheduled capture | Updated metadata inventory |
| Semantic enrichment | Add business meaning | Steward review | Clear data understanding |
| Lineage intelligence | Track data movement | Dependency validation | End-to-end traceability |
| Ownership and stewardship | Assign responsibility | Approval workflow | Clear accountability |
| Policy monitoring | Check governance rules | Compliance evaluation | Earlier issue detection |
| Exception management | Manage policy failures | Resolution tracking | Controlled correction |
| Human validation | Confirm automated results | Expert verification | Accurate governance decisions |
| Trusted data products | Provide approved information | Certification | Reliable analytical use |

### *3.3 Automated Lineage Discovery and End-to-End Traceability*

Automated lineage discovery begins by extracting metadata from the systems connected to SAP Datasphere. The extraction process collects schemas, tables, views, columns, data types, interfaces, and transformation details. Scheduled scans keep the metadata inventory updated when source structures or processing rules change. Standard connectors reduce the effort required to document assets from separate enterprise platforms. This collected metadata provides the foundation for reliable lineage construction.

Upstream discovery identifies the systems, objects, and transformations that contribute data to a selected asset. Downstream discovery determines which models, reports, applications, and users depend on that asset. Both directions are required because data issues can originate from a source or affect later consumption. Automated discovery follows metadata relationships across connected processing stages. This approach provides a clearer view of enterprise data dependencies.

Object-level lineage tracks relationships among databases, tables, views, models, and analytical products. It presents a broad view of how major data objects move through the enterprise environment. Column-level lineage provides more detailed information by tracing individual fields across transformations. It shows how a source column contributes to a calculated measure, report field, or business indicator. Combining both levels supports technical analysis and business understanding.

Transformation mapping records the operations applied while data moves between sources and consumption points. These operations may include joins, filters, aggregations, calculations, mappings, and data type conversions. Each transformation is connected to its input and output objects within the lineage graph. Clear mapping allows practitioners to understand how final values were produced. It also helps identify incorrect rules or unexpected changes in analytical results.

Dependency identification reveals relationships among source objects, transformation processes, semantic models, and downstream products. Direct dependencies show immediate connections between two assets, while indirect dependencies expose longer processing chains. SAP Datasphere can use these relationships to support navigation across connected data objects. Dependency information also helps owners determine which teams use a particular asset. This visibility improves coordination during maintenance and issue resolution.

Change-impact analysis evaluates the possible consequences of modifying a source, column, transformation, or model. The analysis identifies affected downstream objects, reports, measures, and business users before the change is implemented. Impact information helps teams estimate risk and plan suitable testing activities. High-risk changes can be routed to data owners or stewards for approval. This process reduces unexpected failures across enterprise reporting and analytics.

Source-to-consumption traceability connects original data with its final use in reports, analytical models, and business decisions. It provides evidence of where data originated, how it changed, and which governed products consumed it. Complete traceability supports audit investigation, quality analysis, regulatory review, and operational problem resolution. Business users gain confidence when important measures can be traced to verified sources. Technical teams also receive a practical path for locating processing failures.

Automated lineage may contain gaps when systems expose limited metadata or transformations occur outside governed environments. Missing connections can also result from custom code, manual uploads, external tools, or undocumented processing activities. Gap detection

highlights incomplete paths and uncertain relationships that require further examination. Data stewards and technical experts review supporting evidence before confirming or correcting the proposed lineage. Human verification therefore ensures that automated traceability remains accurate, complete, and suitable for governance.

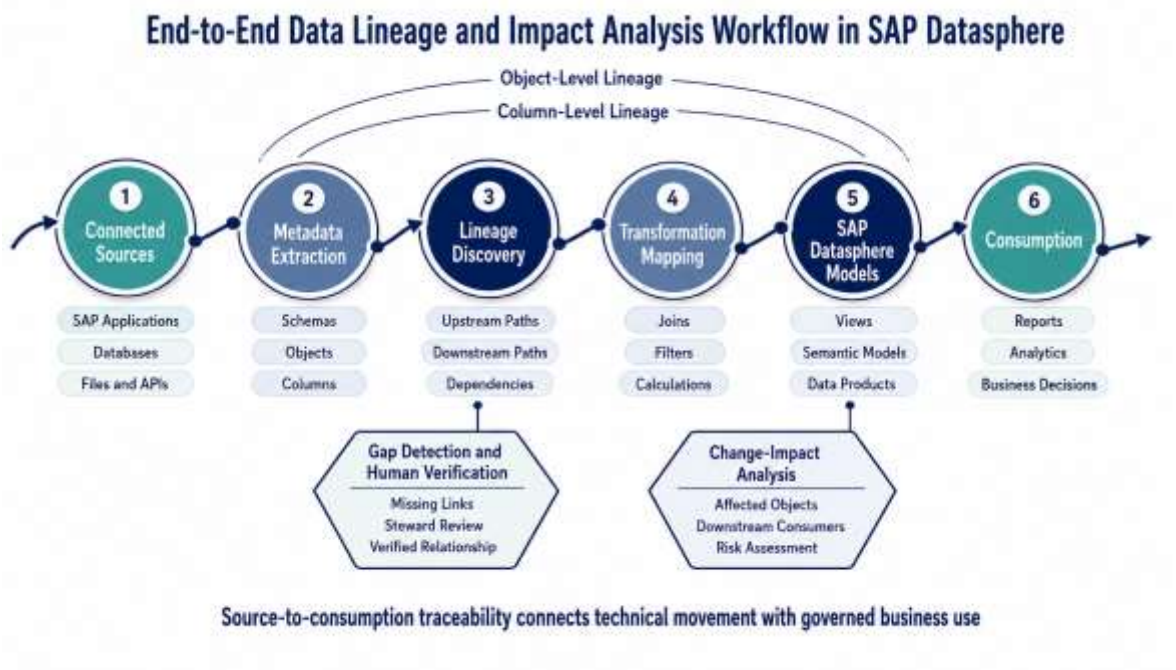



**Figure 4. End-to-End Data Lineage and Impact Analysis Workflow in SAP Datasphere**

### *3.4 Operational Governance, Risk Control, and Stewardship Validation*

Continuous metadata monitoring observes changes in data assets, schemas, relationships, and processing activities. It helps governance teams identify new objects, altered structures, and missing descriptions. Monitoring also shows whether important metadata remains complete and current. Regular checks reduce the risk of using outdated information in reports and analytical models. SAP Datasphere can support this process through connected metadata and governed data relationships.

Governance issue detection identifies missing ownership, broken lineage, incomplete classifications, and policy failures. Each issue is recorded with its affected asset, source, severity, and responsible participant. High-risk issues receive priority because they may affect important reports or regulated information. Clear issue records support faster investigation and prevent problems from remaining unnoticed. This process converts monitoring results into practical governance actions.

Data quality controls evaluate completeness, accuracy, consistency, validity, and timeliness [26]. Compliance controls check whether governed assets follow access, retention, classification, and usage requirements. Failed checks can prevent an asset from receiving trusted or certified status. Lineage information helps identify where a quality problem began and which outputs may be affected. Combined controls support dependable data use across enterprise operations.

Sensitive-data risk management begins with the classification of personal, financial, confidential, and regulated information [27]. Classified assets receive suitable access restrictions, handling rules, and monitoring requirements [28]. Lineage analysis shows where sensitive fields

move and which downstream products consume them. Unexpected movement or unauthorized use can then be identified more quickly. These controls reduce exposure while supporting approved business access.

Alerts notify responsible participants when metadata, quality, lineage, or compliance problems are detected. Each alert includes the affected asset, issue type, severity, and expected response. Remediation workflows assign corrective actions to owners, stewards, or technical teams. Resolution evidence records what was changed, who approved it, and when the issue was closed. This structured process improves response time and strengthens accountability.

Data stewards maintain business definitions, classifications, ownership details, quality expectations, and lineage relationships. They review governance issues and coordinate corrections with technical and business teams. Owners approve important rules, access decisions, and risk treatments for their data domains. Clear responsibilities prevent delays caused by uncertain authority or incomplete communication. Stewardship therefore connects automated monitoring with accountable enterprise action.

Automated discovery can infer relationships among sources, columns, transformations, models, and reports. Human reviewers must verify inferred relationships before they become trusted governance evidence. Stewards compare recommendations with business definitions, technical logic, and supporting records. Incorrect links are rejected, while confirmed relationships are approved and documented. Human validation improves lineage accuracy without removing the benefits of automation.

Operational resilience depends on reliable monitoring, controlled remediation, documented ownership, and recoverable governance records. Audit trails preserve issue histories, approvals, policy results, lineage changes, and resolution evidence. This information helps enterprises explain how data was governed and how risks were handled. Repeated issues can be studied to improve controls and prevent future failures. Continuous review creates a practical cycle of monitoring, correction, validation, and improvement.

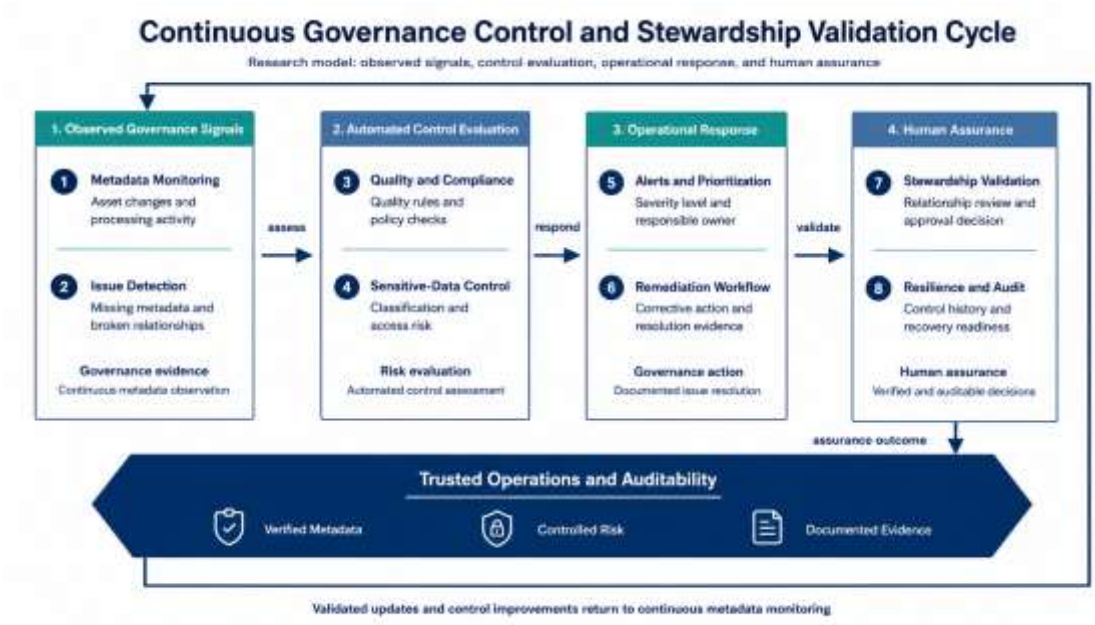


**Figure 5. Continuous Governance Control and Stewardship Validation Cycle**

## RESULTS AND DISCUSSION

### *4.1 Results, Comparative Evaluation, and Strategic Implications*

The baseline assessment revealed fragmented metadata, incomplete lineage records, and inconsistent ownership assignments across the enterprise data environment. Conventional governance processes depended heavily on spreadsheets, interviews, and manually maintained catalog entries. Lineage information was often limited to selected objects and rarely extended from source systems to analytical consumption. Governance teams therefore required considerable time to identify dependencies and evaluate proposed changes. These conditions reduced confidence in data products and increased the effort required for audit preparation.

The comparative evaluation used a controlled enterprise scenario representing connected source systems, transformation processes, semantic models, and analytical outputs in SAP Datasphere. Conventional governance practices were compared with the proposed lineage intelligence framework under the same operational conditions. Five measures were examined: lineage coverage, metadata completeness, dependency-detection accuracy, governance response time, and impact-analysis time. Quantitative results were reviewed together with practitioner observations and stewardship verification. The values represent scenario-based outcomes and should not be interpreted as universal production benchmarks.

The proposed framework improved governance performance by combining automated metadata harvesting, semantic enrichment, and integrated technical and business lineage. Metadata collected from connected systems was associated with transformation logic, ownership information, business definitions, and policy classifications. Automated dependency discovery reduced the need to reconstruct relationships through manual investigation. Stewardship validation remained part of the process for confirming inferred associations and resolving uncertain mappings. This combination supported greater automation while preserving accountable human oversight.

Lineage coverage increased from 42 percent under conventional governance to 91 percent after implementation of the proposed framework. This represents an improvement of 49 percentage points across the assessed source-to-consumption pathways. Metadata completeness increased from 55 percent to 93 percent because technical attributes were combined with semantic definitions, ownership assignments, and operational context. The results indicate that automated harvesting alone was insufficient without semantic enrichment and stewardship review. The integrated approach produced a more complete representation of enterprise data assets and their relationships.

Dependency-detection accuracy increased from 61 percent to 94 percent when object-level and column-level relationships were evaluated together. More accurate dependency information strengthened change-impact analysis by identifying affected transformations, models, reports, and downstream consumers. Governance response time decreased from 48 hours to 12 hours, representing a 75 percent reduction. Impact-analysis time decreased from 32 hours to 6 hours, representing an 81.3 percent reduction. These improvements demonstrate how accessible lineage evidence can shorten investigation and decision cycles.

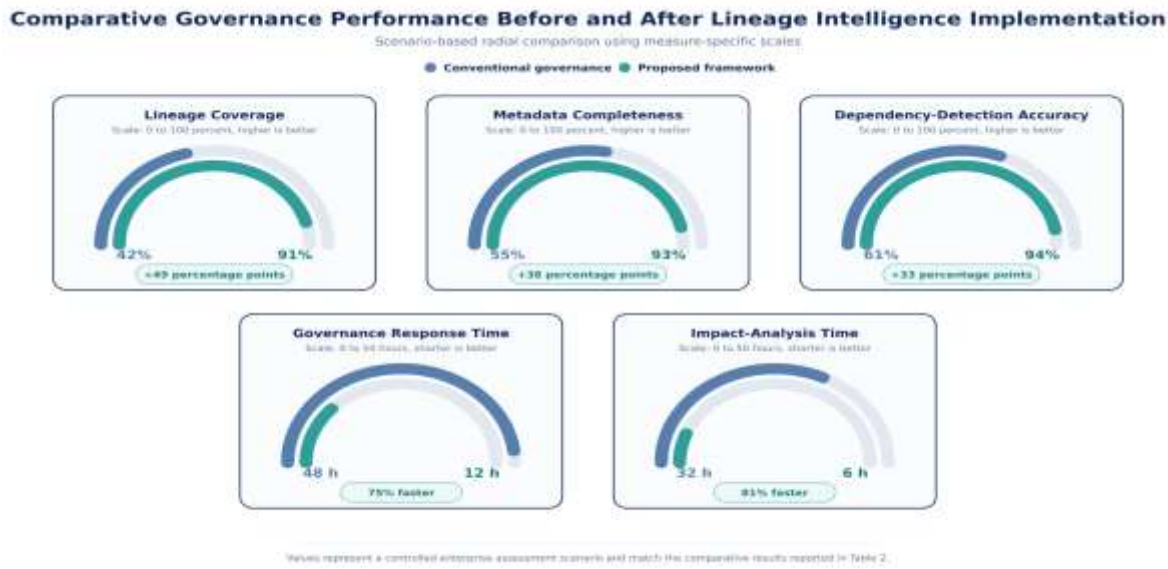



**Figure 6. Comparative Governance Performance Before and After Lineage Intelligence Implementation**

The comparative results show that lineage intelligence improved both governance quality and operational efficiency within the assessment scenario. Higher coverage and completeness provided stronger evidence for ownership, policy enforcement, compliance review, and data quality investigation. Faster response and impact analysis allowed governance teams to address issues before they affected a wider group of data consumers. Practitioner verification remained important when source metadata was incomplete or transformation logic could not be interpreted reliably.

From a strategic perspective, the proposed framework can support trusted data products, regulatory readiness, and more reliable enterprise analytics [29]. Shared lineage and metadata evidence can improve cooperation among data owners, stewards, architects, analysts, and compliance teams. Operationally, automated discovery can reduce repetitive documentation work and direct practitioner attention toward exceptions requiring judgment. Organizations can also use the resulting evidence to prioritize remediation according to downstream impact and business criticality. Successful implementation nevertheless requires clear ownership, controlled access, metadata standards, and sustained stewardship participation.

The study contributes a unified view of automated metadata management, lineage intelligence, semantic context, and human governance controls in SAP Datasphere. It demonstrates how governance performance can be evaluated through measures covering both information quality and response efficiency. The scenario-based design, however, does not represent every source technology, organizational structure, or regulatory environment. Performance may vary according to connector availability, metadata quality, transformation complexity, and practitioner expertise. Future evaluation should therefore examine larger production environments, longer operating periods, and additional measures of auditability, resilience, and user trust.

**Table 2. Comparative Results for Conventional Governance and the Proposed Lineage Intelligence Framework**

| Evaluation measure | Conventional governance | Proposed framework | Comparative result |
|---|---|---|---|
| Lineage coverage | 42% | 91% | Increase of 49 percentage points |
| Metadata completeness | 55% | 93% | Increase of 38 percentage points |
| Dependency-detection accuracy | 61% | 94% | Increase of 33 percentage points |
| Governance response time | 48 hours | 12 hours | 36 hours faster, 75% reduction |
| Impact-analysis time | 32 hours | 6 hours | 26 hours faster, 81.3% reduction |

## CONCLUSION

Enterprise data governance increasingly depends on the ability to establish trust across distributed data sources, transformation processes, semantic models, and analytical products. Fragmented metadata and manually maintained lineage records limit visibility into how enterprise data is created, modified, and consumed. This study examined these limitations within the SAP Datasphere environment and proposed an integrated governance framework. The framework combines automated metadata management, lineage intelligence, semantic context, policy controls, and accountable stewardship. This combination provides a practical foundation for improving enterprise data traceability and governance consistency.

The proposed framework treats metadata as an active governance resource rather than a passive catalog record. Automated harvesting collects technical, operational, business, and governance metadata from connected enterprise systems. Semantic enrichment associates these records with business definitions, ownership responsibilities, sensitivity classifications, and quality expectations. Integrated metadata enables governance teams to understand both the technical movement and business meaning of data. This approach strengthens the reliability of information used for operational decisions, regulatory reporting, and enterprise analytics.

End-to-end lineage intelligence represents the central capability of the proposed governance approach. Object-level and column-level relationships connect source systems, integration processes, transformations, semantic models, and analytical outputs. Automated dependency discovery identifies upstream and downstream relationships that conventional documentation practices may overlook. Change-impact analysis uses these relationships to determine how modifications could affect dependent data products and consumers. Human verification remains necessary when transformation logic is incomplete, ambiguous, or unavailable through connected metadata.

SAP Datasphere provides an appropriate environment for integrating data, metadata, semantic modeling, and governance responsibilities. The proposed architecture extends this environment through coordinated harvesting, enrichment, lineage discovery, policy monitoring, and stewardship validation. Data owners remain accountable for business meaning and acceptable

use, while stewards manage metadata quality and governance exceptions. Architects and technical teams maintain structural accuracy, transformation mappings, and system dependencies. Shared responsibilities reduce governance isolation and support consistent control across enterprise data domains.

The scenario-based evaluation demonstrated measurable improvement across the selected governance indicators. Lineage coverage increased from 42 percent to 91 percent, while metadata completeness increased from 55 percent to 93 percent. Dependency-detection accuracy increased from 61 percent to 94 percent through integrated technical and business lineage. Governance response time decreased from 48 hours to 12 hours, and impact-analysis time decreased from 32 hours to 6 hours. These results indicate that accessible metadata and lineage evidence can improve both governance quality and operational efficiency.

Strategic benefits extend beyond faster documentation and dependency analysis. Trusted lineage supports regulatory readiness, data quality management, sensitive-data protection, risk assessment, and accountable analytical consumption. Reliable metadata also helps enterprises define and maintain trusted data products with clear ownership and measurable governance conditions. Operational teams can direct their effort toward exceptions and high-impact issues instead of repetitive relationship discovery. Governance leaders can use the resulting evidence to prioritize investment according to business criticality, compliance exposure, and downstream dependency.

Implementation success depends on organizational readiness as well as technical capability. Incomplete source metadata, unsupported connectors, complex transformation logic, and inconsistent business terminology may reduce automated discovery accuracy. Stewardship capacity, ownership clarity, access controls, and metadata standards also influence the sustainability of the framework. Scenario-based results cannot represent every enterprise architecture, industry requirement, or regulatory environment. Organizations should therefore introduce the framework incrementally and validate its performance against their own systems, policies, and operating conditions.

The study contributes an integrated perspective on intelligent governance by connecting automated metadata management, semantic enrichment, technical lineage, business traceability, and human accountability. Its conceptual and evaluation structures can support future research into governance maturity, metadata quality, lineage confidence, and trusted data-product management. Further studies should evaluate the framework across larger production environments and longer operating periods. Additional research may examine cross-platform lineage, federated governance, automated policy interpretation, and explainable relationship inference [30]. Enterprise adoption of these capabilities can establish a more transparent, resilient, and trusted foundation for data-driven operations.

## REFERENCES

[1] B. van der Sloot and E. Keymolen, "Can we trust trust-based data governance models?," *Data Policy*, vol. 4, Art. no. e45, 2022, doi: 10.1017/dap.2022.36.

[2] B. M. V. Bernardo, H. São Mamede, J. M. P. Barroso, and V. M. P. D. dos Santos, "Data governance and quality management: Innovation and breakthroughs across different fields," *J. Innov. Knowl.*, vol. 9, no. 4, Art. no. 100598, 2024, doi: 10.1016/j.jik.2024.100598.

[3] Q. Zhang, X. Sun, and M. Zhang, "Data matters: A strategic action framework for data governance," *Inf. Manag.*, vol. 59, no. 4, Art. no. 103642, 2022, doi: 10.1016/j.im.2022.103642.

[4] D. Boukraâ, M. Bala, and S. Rizzi, "Metadata management in data lake environments: A survey," *J. Libr. Metadata*,

vol. 24, no. 4, pp. 215–274, 2024, doi: 10.1080/19386389.2024.2359310.

[5] S. Hikmawati, P. I. Santosa, and I. Hidayah, "Improving data quality and data governance using master data management: A review," *IJITEE Int. J. Inf. Technol. Electr. Eng.*, vol. 5, no. 3, pp. 90–101, 2021, doi: 10.22146/ijitee.66307.

[6] D. Petrik, A. Untermann, and H. Baars, "Functional requirements for enterprise data catalogs: A systematic literature review," in *Software Business, Lecture Notes in Business Information Processing*, vol. 500, 2023, pp. 3–18. doi: 10.1007/978-3-031-53227-6_1.

[7] P. Parízek and L. Hermann, "Data lineage analysis for enterprise applications by Manta: The story of Java and C# scanners," in *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice*, 2024, pp. 25–35. doi: 10.1145/3639477.3639739.

[8] Y. Chen, Y. Zhao, and X. Li, "An open dataset of data lineage graphs for data governance research," *Vis. Informatics*, vol. 8, no. 1, pp. 1–5, 2024, doi: 10.1016/j.visinf.2024.01.001.

[9] I. Blohm, F. Wortmann, and C. Legner, "Data products, data mesh, and data fabric: New paradigm(s) for data and analytics?," *Bus. Inf. Syst. Eng.*, vol. 66, no. 5, pp. 643–652, 2024, doi: 10.1007/s12599-024-00876-5.

[10] J. Bode, N. Kühl, and D. Kreuzberger, "Toward avoiding the data mess: Industry insights from data mesh implementations," *IEEE Access*, vol. 12, pp. 95402–95416, 2024, doi: 10.1109/ACCESS.2024.3417291.

[11] K. Bližnák, M. Munk, and A. Pilková, "A systematic review of recent literature on data governance (2017–2023)," *IEEE Access*, vol. 12, pp. 149875–149888, 2024, doi: 10.1109/ACCESS.2024.3476373.

[12] I. A. Machado, C. Costa, and M. Y. Santos, "Data mesh: Concepts and principles of a paradigm shift in data architectures," *Procedia Comput. Sci.*, vol. 196, pp. 263–271, 2022, doi: 10.1016/j.procs.2021.12.013.

[13] R. Abraham, J. Schneider, and J. vom Brocke, "A taxonomy of data governance decision domains in data marketplaces," *Electron. Mark.*, vol. 33, no. 1, Art. no. 22, 2023, doi: 10.1007/s12525-023-00631-w.

[14] P. Wieder and H. Nolte, "Toward data lakes as central building blocks for data management and analysis," *Front. Big Data*, vol. 5, Art. no. 945720, 2022, doi: 10.3389/fdata.2022.945720.

[15] S. Nadal, P. Jovanovic, and B. Bilalli, "Operationalizing and automating data governance," *J. Big Data*, vol. 9, no. 1, Art. no. 117, 2022, doi: 10.1186/s40537-022-00673-5.

[16] R. Eichler, C. Giebler, and C. Gröger, "Modeling metadata in data lakes: A generic model," *Data Knowl. Eng.*, vol. 136, Art. no. 101931, 2021, doi: 10.1016/j.datak.2021.101931.

[17] K. Gierend, F. Krüger, and S. Genehr, "Provenance information for biomedical data and workflows: Scoping review," *J. Med. Internet Res.*, vol. 26, Art. no. e51297, 2024, doi: 10.2196/51297.

[18] D. Paparova, M. Aanestad, and P. Vassilakopoulou, "Data governance spaces: The case of a national digital service for personal health data," *Inf. Organ.*, vol. 33, no. 1, Art. no. 100451, 2023, doi: 10.1016/j.infoandorg.2023.100451.

[19] N. Jahnke and B. Otto, "Data catalogs in the enterprise: Applications and integration," *Datenbank-Spektrum*, vol. 23, no. 2, pp. 89–96, 2023, doi: 10.1007/s13222-023-00445-2.

[20] M. Cherradi, F. Bouhafer, and A. El Haddadi, "Data lake governance using IBM-Watson Knowledge Catalog," *Sci. African*, vol. 21, Art. no. e01854, 2023, doi: 10.1016/j.sciaf.2023.e01854.

[21] K. Goel, N. Martin, and A. H. M. ter Hofstede, "Demystifying data governance for process mining: Insights from a Delphi study," *Inf. Manag.*, vol. 61, no. 5, Art. no. 103973, 2024, doi: 10.1016/j.im.2024.103973.

[22] F. Ridzuan and W. M. N. W. Zainon, "A review on data quality dimensions for big data," *Procedia Comput. Sci.*, vol. 234, pp. 341–348, 2024, doi: 10.1016/j.procs.2024.03.008.

[23] K. A. Jang and W. J. Kim, "Development of data governance components using DEMATEL and content analysis," *J. Supercomput.*, vol. 77, no. 4, pp. 3695–3709, 2021, doi: 10.1007/s11227-020-03405-9.

[24] M. Bacco, A. Kocian, and S. Chessa, "What are data spaces? Systematic survey and future outlook," *Data Br.*, vol. 57, Art. no. 110969, 2024, doi: 10.1016/j.dib.2024.110969.

[25] C. Ducuing and R. H. Reich, "Data governance: Digital product passports as a case study," *Compet. Regul. Netw. Ind.*, vol. 24, no. 1, pp. 3–23, 2023, doi: 10.1177/17835917231152799.

[26] J. Shen, S. Zhou, and F. Xiao, "Research on data quality governance for federated cooperation scenarios," *Electronics*, vol. 13, no. 18, Art. no. 3606, 2024, doi: 10.3390/electronics13183606.

[27] D. Oreščanin, T. Hlupić, and B. Vrdoljak, "Managing personal identifiable information in data lakes," *IEEE Access*, vol. 12, pp. 32164–32180, 2024, doi: 10.1109/ACCESS.2024.3365042.

[28] N. Sheehan, F. Botta, and S. Leonelli, "Unrestricted versus regulated open data governance: A bibliometric comparison of SARS-CoV-2 nucleotide sequence databases," *Data Sci. J.*, vol. 23, no. 1, Art. no. 29, 2024, doi: 10.5334/dsj-2024-029.

[29] J. J. Zygmuntowski, L. Zoboli, and P. F. Nemitz, "Embedding European values in data governance: A case for public data commons," *Internet Policy Rev.*, vol. 10, no. 3, pp. 1–29, 2021, doi: 10.14763/2021.3.1572.

[30] A. Dolhopolov, A. Castelltort, and A. Laurent, "Implementing federated governance in data mesh architecture," *Futur. Internet*, vol. 16, no. 4, Art. no. 115, 2024, doi: 10.3390/fi16040115.