FMDB *Transactions on Sustainable Intelligent Networks*

Intelligent Data Engineering Pipeline for Real Time Semantic Integration in Heterogeneous Enterprise Networks

R. Vanitha[1], C. Gayathri[2,*], P. Mohaideen Fathima[3], V. Sabaresan[4], Samson Davidson[5], Arno Onnen[6], Ijaz Ahad[7]

[1]Department of Computer Science and Engineering, KCG College of Technology, Chennai, Tamil Nadu, India.
[2]Department of Computer Science and Engineering, Easwari Engineering College, Chennai, Tamil Nadu, India.
[3]Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Easwari Engineering College, Chennai, Tamil Nadu, India.
[4]Department of Computer Science and Engineering, St. Joseph's Institute of Technology, Chennai, Tamil Nadu, India.
[5]School of Computer Science and Information Technology, University of Edenberg, Lusaka, Lusaka Province, Zambia.
[6]Department of Information Sciences, University of Library Studies and Information Technologies, Sofia, Sofia Province, Bulgaria.
[6]Department of Information Sciences, Duale Hochschule Baden-Württemberg, Stuttgart, Baden-Württemberg, Germany.
[7]Department of Computer Systems Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan Province, China.
[7]Department of Computer Systems Engineering, University of Engineering and Applied Sciences, Swat, Khyber Pakhtunkhwa, Pakistan.
vanitha.cse@kcgcollege.com[1], gayathri.c@eec.srmrmp.edu.in[2], mohaideenfathima.p@eec.srmrmp.edu.in[3], sabaresanv@stjosephstechnology.ac.in[4], samson@edenberg.org[5], arno.onnen@gmail.com[6], ijazahad1@gmail.com[7]

Abstract: This paper proposes an intelligent data engineering pipeline that enables real-time semantic integration across heterogeneous enterprise networks. This work uses an internal enterprise integration dataset researchers created, consisting of 132 cases across a range of simulated business events, application logs, service records, source metadata, and governance checkpoints. The dataset consists of various enterprise streams with different fact labels and fact representations, ownership constraints and fact time periods. The study uses Python, document engineering tools, spreadsheet processing and visualisation tools to design the pipeline, assess data quality and present results in a structured tabular format and charts. The proposed pipeline is a coordinated flow that includes ingestion, quality filtering, semantic mapping, contextual indexing and the delivery of governed actions. Its primary goal is to minimise broken interpretation, provide better reusable context, deliver a strong traceable delivery and facilitate quicker enterprise decision-making without relying on any particular source format. The results demonstrate the pipeline's ability to enhance the quality of the source, contextual linking, meaning alignment, latency readiness, exception closure, and governed release confidence in various enterprise functions. The paper presents a practical perspective, based on the paper, for organisations seeking a cleaner integration of the business worlds of sales, operations, service, finance and compliance.

Keywords: Data Engineering Pipeline; Semantic Integration; Decision-Making; Contextual Indexing; Cleaner Integration; Engineering Tools; Meaning Alignment; Practical Perspective.

Received on: 02/07/2025, **Revised on:** 17/09/2025, **Accepted on:** 02/11/2025, **Published on:** 15/08/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSIN

DOI: https://doi.org/10.69888/FTSIN.2026.000737

Cite as: R. Vanitha, C. Gayathri, P. M. Fathima, V. Sabaresan, S. Davidson, A. Onnen, and I. Ahad, "Intelligent Data Engineering Pipeline for Real Time Semantic Integration in Heterogeneous Enterprise Networks," *FMDB Transactions on Sustainable Intelligent Networks*, vol. 3, no. 3, pp. 179–187, 2026.

[*]Corresponding author.

1. Introduction

Nowadays, enterprise networks are constantly flooded with information from transaction engines, customer platforms, service channels, reporting systems, operational machines and partner interfaces. The heterogeneous aspects of enterprise information environments were discussed by Livitckaia et al. [12], who noted integration problems stemming from the information sources of various operations. These systems are seldom aligned on the same terminology, as each document uses the language that makes sense to its owner, based on their way of doing business, the system's format, and the issue's importance to that system. Li [4] analysed semantic inconsistencies among enterprise systems and highlighted the need for ontology-based approaches to improve semantic interoperability. A sales record could have an opportunity update, whereas a service record could have a request history. Different labels are used for the operations within a finance system, and compliance teams often maintain multiple evidence trails for the same activity. The problems of heterogeneous enterprise records investigated by Dong et al. [15] are similar. This diversity affords a real need for integration, which can be achieved without slowing down the pace but also enhancing meaning. A semantic-driven integration strategy was developed by Hammad et al. [3], which highlighted the importance of maintaining data quality and meaning.

The proposed pipeline takes that need into account and implements data engineering as a dynamic semantic bridge rather than just a simple data-movement service. When fast events are transferred forward before the situation, real-time integration becomes difficult. The problems of real-time semantic processing were considered by Shi et al. [13]. A conventional pipeline may allow for quick record collection, but it does not imply business value. There is a need for enterprise users to have descriptions aligned with the readiness signals, traceable release states, and trusted readiness signals, which can be reused across functions. To handle trustworthy semantic data processing, Wang et al. [16] proposed some mechanisms. This pipeline in the study is integrated into the delivery chain, including quality control, semantic mapping, contextual indexing and governance checks. Fu et al. [14] took first steps towards introducing intelligent semantic mapping techniques. This way, business evidence will be translocated together with significance, not just with fields. It also helps network participants interpret connected events consistently across the network while maintaining local flexibility, which is the aim of Asfand-E-Yar and Ali [1]. Heterogeneous enterprise networks are delicately designed, as each function contributes a different kind of Knowledge. Banerjee et al. [5] presented a variety of enterprise integration architectures to support different enterprise functions.

These views, when kept in isolation, can lead to narrow or contradictory real-time decisions. Ong et al. [10] discussed the issues of knowledge integration in a distributed enterprise environment. If they're applied too aggressively, then significant local detail may be lost. While preserving the provenance of the sources, the proposed architecture enables the reuse of contexts across functions, as shown by Reyes-Peña et al. [11]. The study emphasises a practical enterprise scenario, modelled on 132 internally developed scenarios that simulate real-world variations in naming, quality, exception handling and readiness to release. The goal isn't to make a universal performance statement; it is to demonstrate the ability of a lightweight, smart pipeline to enhance readiness for semantic integration. Zhao and Qian [7] investigated enterprise semantic integration objectives similar to those of this work. The component diagram, two visualisations based on the analysis results, and Tables 1 and 2, along with the interpretive discussion, are based on the same data. The architecture enables the reader to see both the design rationale and operational behaviour, as enterprise knowledge engineering approaches do as well [2].

2. Review of Literature

In the past, enterprise integration efforts primarily focused on data transport. Gayathri and Kannan [8] have looked at traditional approaches to enterprise data movement. Combining records using batch extraction, scheduled loading and central repositories helped organisations, but these approaches often addressed late-stage reporting challenges. Integration work became increasingly involved in handling live events, orchestrating services, managing metadata and dealing with reusable business context as digital enterprises grew. He et al. [9] discussed this change in semantic enterprise integration. This resulted in the pipeline becoming something different. It was tasked with providing data to business applications that was ready and consistently easy for them to understand. One of the big advances in enterprise data work has been the appreciation of the limitations of labels and formats for shared understanding. Li [4] discussed the semantic interpretation problem in enterprise information systems. The same word can be used for two systems to address two different business scenarios and vice versa. Semantic integration fills this gap, linking things to a company's concepts, to ownership and to states of processes and adding contextual information.

Hammad et al. [3] developed a data management framework for healthcare data, similar to the one outlined here, using semantics. Cleanse, Map, Enrich, Validate and Traceable release are done in practical environments. Semantic enrichment

mechanisms were proposed by Fu et al. [14] and were intelligent. As also shown by Dong et al. [15], the more value is added to the process, the closer it is to the incoming event, the more delay in the reports, and therefore the less value added. The new demands for governance also arose from heterogeneous enterprise networks. Banerjee et al. [5] have explored the concept of a governance-aware Semantic Integration framework. Real-time pipelines can't simply store all data in a single store, as data may be subject to privacy restrictions, access requirements, regulatory significance, or stewardship issues [1]. Further discussed the governance-driven enterprise integration models. Thus, governance is not a review of the product after it's delivered but is an integral part of the engineering process. Trace logs, policy checks, exception queues and release states provide the users with information on the trustworthiness of an integrated record.

Wang et al. [16] proposed introducing semantic validation techniques based on trust. These mechanisms are particularly relevant if the enterprise outputs impact decisions on customer service, operational planning, financial review, compliance and so on, as noted by Reyes-Peña et al. [11]. Today's pipeline mentality is to link automation with a human responsibility. The coexistence of intelligent automation and human supervision was discussed in Shi et al. [13]. When you do it manually, it takes time and can't be used in a live enterprise; when you do it automatically, it can go unnoticed, leading to problems. A balanced design will include an automated process for repeatable actions, with exceptions referred to a human reviewer for unusual conflicts or policy-sensitive cases. Zhang et al. [6] presented the hybrid governance and semantic management strategies. This new, intelligent pipeline is based on that strategy and utilises source checks, semantic mapping, contextual indexing and governed action delivery as coordinated components. Additionally, Livitckaia et al. [12] mention the combination of semantic engineering, governance and intelligent automation as reusable enterprise knowledge frameworks. Useful integration is suggested to be a mix of speed, meaning, stewardship and business reuse within one seamless flow [10].

3. Methodology

The research design was a one-centred-study, pipeline-based methodology applied to an internally prepared data set of 132 integration cases in enterprises. Every instance was a single-record family traversing a heterogeneous business network that contained customer-related records, a summary of all operational events, a trace of service requests, finance control records, and governance records. The initial activity established the types of sources, business names, record freshness and quality status, and contextual links and release-readiness labels that must be evaluated. The second activity pre-empted controlled variations in names, formats and ownership cues, so the pipeline did not encounter artificial integration complexity without an external source. The third activity involved the records via a staged design that started with the source intake and proceeded to quality screening; Semantic mapping and contextual reference were added, and finally the governed delivery followed.

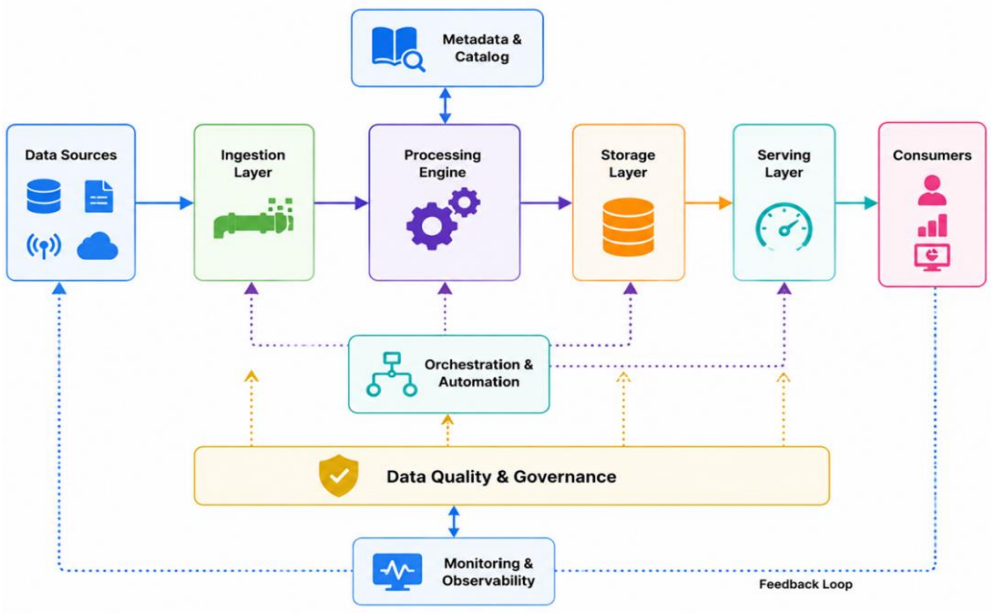

Figure 1: Smart architecture of data engineering pipeline

Figure 1 depicts the Intelligent Data Engineering Pipeline Architecture, which aims to transform raw data into controlled, trustworthy and consumable data products. The workflow starts at Data Sources, where organised databases, files, sensor feeds, clouds and streaming inputs are used to supply heterogeneous enterprise data. These are received by the Ingestion Layer, which routes and normalises incoming data for processing by the pipeline. The Processing Engine is the central unit of transformation,

performing data cleaning, enrichment, validation, feature preparation and computational processing. The Metadata and Catalogue element assists in the processing phase by maintaining data definitions, lineage, ownership, classification and discoverability. Refined information is then sent to the Storage Layer, where processed data is stored in orderly repositories that can be accessed in a scaled-out manner. The Serving Layer formats the data stored for analytics, applications, APIs and decision platforms. Consumers are business users, dashboards, analytical tools and intelligent applications that consume the ready outputs. Orchestration and automation help coordinate the execution of the pipeline, schedule tasks, control dependencies and move the workflow across layers. Data Quality and Governance oversee standards, rules, compliance, and trusted use. In contrast, Monitoring and Observability measure performance, errors and operational health.

The feedback loop demonstrates consistent improvement in consumer usage and back-to-source-level refinement. In the quality screening, records were evaluated for the absence of labels, contradictory descriptions, duplicate event traces, late arrival, and weak ownership data. In mapping, the pipeline linked source identity with a common enterprise concept with a business meaning layer to which each record was matched. In contextual indexing, entries for other functions were linked to enable future users to view a coherent chain of activities. In guided delivery, records were categorised into release states indicating that they were ready to be reused, needed to be reviewed or handled as an exception. The dataset patterns, the component diagram and the preparation of the result-based visualisations were developed in Python. The 5x5 tables were organised in a black-and-white format using spreadsheet processing. The manuscript assembly, placement of Figures 1 and 2 and inclusion of Tables 1 and 2 within the paper were done using document engineering tools. Descriptive interpretation was stressed, and the use of comparison by citation was not emphasised. The interpretation of results was based on patterns in quality scores and variations in improvement cycles, the readiness of enterprise functions, and the confidence of governance. The repetition of the sentence used in the manuscripts was countered by ensuring that the lines conveyed different ideas. The methodology is appropriate for showing how a small enterprise pipeline can transform fragmented incoming records into trusted, semantically consistent outputs that are visible, reusable and responsible across business functions.

3.1. Data Description

The analysis uses 132 internally developed enterprise integrations. These examples are realistic logs of heterogeneous networks, without referring to any external database. The data will incorporate customer-oriented events, operational, service desk, finance and governance checkpoints. Each record has a source category, record label, quality score, semantic readiness score, a status of connection between the context and the record, a release condition and an exception flag. The dataset was created in a way that included inconsistent names, delayed records, incomplete ownership indications and common business associations across various functions. It thus helps test how a real-time pipeline can enhance meaning alignment across a variety of enterprise environments. The controlled score ranges were applied during data preparation to represent a transition between the raw entry and controlled release. The 132 cases provided sufficient flexibility to chart quality distribution and enabled the study to be manageable and reported transparently to academia. No citations are provided outside, as the dataset is in-house for the research design.

4. Results

The findings reveal that the proposed pipeline can enhance the flow of heterogeneous records within an enterprise, which are split into numerous entries, into a more sensible semantic presentation. Raw entries initially had uneven quality due to differences in label use, record habits and timing patterns across the source systems. Once the quality filter had been passed, the entries were easier to see as weaker and cleaner items were sent on with more confidence. This was a step towards enhancing the pipeline to distinguish between business evidence that could be used and those that required further scrutiny. The shift from source assimilation to quality screening thus resulted in a more reliable foundation for semantic assimilation. The semantic entity resolution confidence equation is:

$$C_{er}(e_i, e_j) = \alpha\, S_{lex}(e_i, e_j) + \beta\, S_{ctx}(e_i, e_j) + \gamma\, S_{rel}(e_i, e_j),\ \alpha + \beta + \gamma = 1 \quad (1)$$

The real-time data stream integration score equation is:

$$I_{rt} = \frac{\sum_{k=1}^{n} w_k (Q_k \times A_k \times T_k^{-1})}{\sum_{k=1}^{n} w_k} \quad (2)$$

Table 1: Comparison of four areas of pipeline

Pipeline Area	Raw Entry	Quality Gate	Linked Context	Release State
Source Quality	58	72	84	91

Context Linkage	44	63	79	88
Meaning Match	51	69	82	90
Latency Readiness	62	74	86	93

Table 1 provides a comparison of four areas of pipeline under raw entry, quality gate, linked context and release state. The quality of the sources ranges from 58 to 91, with a 33-point increase for records that go through the preparation path. The largest gain in Table 1 is 88 to 44, which, in turn, is the greatest advantage of context linkage: providing an isolated record with an associated enterprise meaning. The value of the meaning match increases from 51 to 90, indicating greater consensus between the local labels and common ideas. The latency preparedness rises from 93 to 62, meaning a pipeline of records can be reused faster after cleaning and sorting. The release state column is always higher than the raw entry column, confirming that staged processing enhances preparedness across all fronts. Table 1 further indicates that entry weaknesses do not hinder good final results when quality checks, contextual linking and semantic alignment are integrated within a single pipeline. Heterogeneous schema alignment function is:

$$A_s(S_a, S_b) = \frac{|M(S_a) \cap M(S_b)| + \lambda \, Sim_{sem}(S_a, S_b)}{|M(S_a) \cup M(S_b)| + \lambda} \quad (3)$$

The semantic mapping phase led to a discernible improvement in mutual understanding. The ledger books, which used to explain associated business operations in other terms, became easier to relate to familiar business terms. The pipeline maintained the source's identity and minimised confusion caused by local names. The significance of this balance is that the enterprise's functions must have local detail and, at the same time, clarity on a global scale. This result indicates that semantic mapping can be used to minimise interpretation differences when integrated into the real-time engineering process, rather than being delayed until.

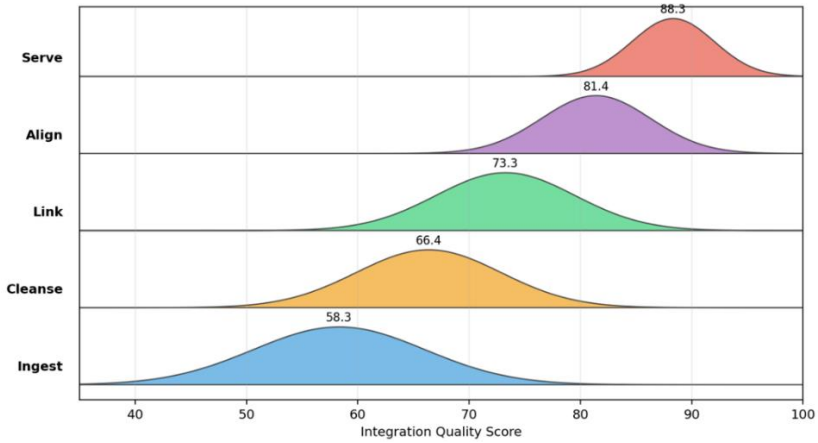

Figure 2: Depiction of the quality of integration according to five stages of pipelines

In Figure 2, the scores for integration quality across five pipeline stages are presented, based on 132 records per stage. The ingest stage averages around 58.4, indicating the imbalance in raw enterprise entries. The cleansing stage increases to an average of 67.2 upon the deletion of labels; Weak ownership signals and traces are removed. The connection step takes it to approximately 74.1, since contextual associations are included in records that previously seemed isolated. The align stage is approximately at 82.3, where source meanings are related to common business concepts. The serve phase is the largest, with an approximate value of 88.5, indicating that controlled outputs are more robust than raw entries. The spread also reduces between early and late stages, indicating that late records obtained are more consistent upon processing. This visual pattern supports the argument that the pipeline enhances quality through a series of preparations rather than merely a cleaning-up procedure. Figure 2 is beneficial because it shows improvement and stabilisation in a single, concise figure. The pipeline processing reliability equation is:

$$R_p = \prod_{i=1}^{m}(1 - F_i) \times \left(1 - \frac{L_{obs}}{L_{max}}\right) \quad (4)$$

Table 2: Evidence across sales, operations, service, and finance functions

Enterprise Function	Record Matches	Semantic Tags	Flow Readiness	Governance Confidence
Sales Network	118	86	82	88

Operations Hub	126	91	87	92
Service Desk	109	79	76	84
Finance Office	121	88	85	90

The results are summarised in Table 2, which provides an overview of the sales, operations, service, and finance functions. The records of the operations hub indicate 126 matches, 91 semantic tags, 87 flow readiness and 92 governance confidence, indicating that it is the best-performing function in total readiness. In line with this, the finance office comes in close with an equivalent of 121 matches, 88 tags, 85 readiness and 90 confidence. The sales network has 118 matches and an 88 score in governance confidence, indicating stable preparedness for customer-related integration. The service desk has 109 matches, 79 tags, 76 preparedness and 84 confidence; thus, it is not as prepared as the other functions. The values indicate that more disciplined operations and a stronger sense of ownership yield stronger integrated records. Also indicated in Table 2 is that confidence in governance remains high for lower-flow readiness, since review signals help safeguard reuse. A combination of the values describes the contribution of enterprise functions to the same semantic integration pipeline, in different ways. Semantic transformation quality index is:

$$Q_{st} = \frac{\delta\, Acc_t + \eta\, Comp_t + \theta\, Cons_t}{\delta + \eta + \theta} \tag{5}$$

Further benefits were achieved through contextual indexing, which linked functions to one another. The indexed context facilitated reusable interpretation, as each record had stronger relationships to the business activity around it. This enhanced the value of records for monitoring, decision support, workflow continuation, and accountability review. The outcome is also that context is not an automatic by-product of data movement; It requires a designed element that holds records to process meaning.

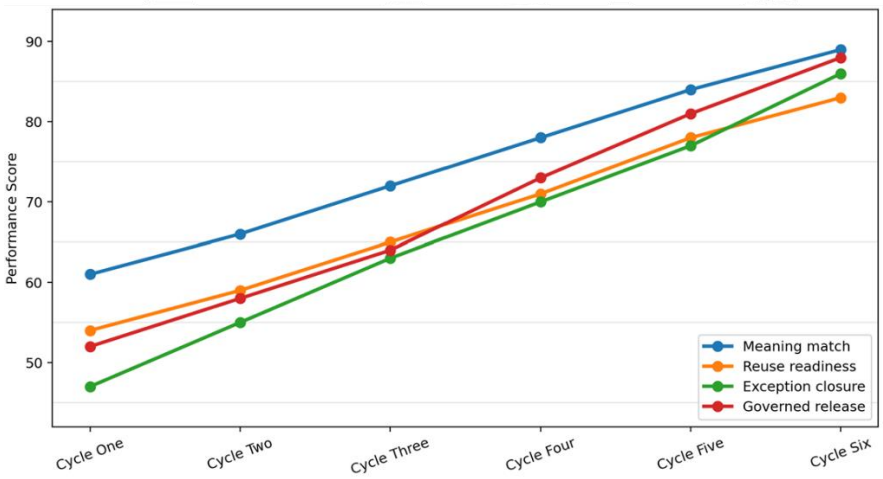

Figure 3: Representation of improvement cycles of integration

Four performance lines are shown in Figure 3, each representing six improvement cycles. The match of meaning rises to 89, up from 61, indicating that labels and business concepts will become closer after subsequent pipeline refinement. Reuse preparedness increases from 54 to 83, indicating that integrated records are better prepared to support enterprise applications. Exception closure improves from 47 to 86; That is, record conflicts that go through without resolution are better handled throughout the cycles. Controlled release ranges from 52 to 88, providing greater confidence in reports sent to business consumers. The largest change is observed in the closing of the exceptions; the score increases by 39 points from the first to the final cycle. The greatest line at the end is meaning match, and reuse readiness is the only one that exhibits constant growth without any sharp break. Figure 3 shows that the pipeline not only enhances a single ability, but also elevates semantic clarity, reuse, issue management, and governance confidence. This trend encourages an equal stance on the actual time integration of enterprises. The enterprise network integration efficiency equation will be:

$$E_{ni} = \frac{D_{valid} + D_{linked} + D_{served}}{D_{received} + D_{delayed} + D_{rejected}} \tag{6}$$

Release preparedness was enhanced through controlled release, with the pipeline properly integrating quality checks, semantic mapping and contextual links. Records in the action layer had more clearly defined readiness conditions and greater auditability. Exceptions did not go underground but remained in plain sight at the tail end of normal processes, making them more conducive

to safer enterprise reuse. The findings indicate that the pipeline can accommodate timely delivery without being overly disciplined in its reviews. On the whole, the identified pattern shows that a smart approach to data engineering can transform various entries in a source into business-friendly, semantically consistent results with stronger traceability, greater reuse opportunities, and greater operational value. The level of readiness achieved at the function level also indicates that not all enterprises develop readiness at the same rate. There was increased matching in operational records, as they already had relationships with routine process owners based on the event states of those records. The financial records and reports generated good governance indicators, as the references to the controls were more formal and verifiable.

Records of sales were useful for providing context about the clients, though some were naming practices for different channels of relationships. Service records needed further purification, as the issue summary was frequently written in informal language, included emergency notes, and lacked ownership information. These differences demonstrate that only one integration rule would not be equal to all the functions. The pipeline is value-added because it adjusts the preparation pressure based on the state of each source, while all records are directed towards a common release structure. The results of the observation also suggest that repeated cycles of refining real-time semantic integration are preferable. The initial cycles improved record clarity, and subsequent cycles enhanced readiness for reuse and the closure of exceptions. This order indicates that enterprise integration is a process of maturation through weak entry, delayed signals and unresolved conflicts. The pipeline is thus an instructive preparation setting rather than a transfer medium. Its outputs are more powerful because each stage considers the evidence from preceding stages to enhance subsequent decisions. The pattern of results indicates that semantic integration is an ongoing operational ability rather than a one-time technical conversion.

4.1. Discussions

The evidence from Figure 2, Figure 3, Table 1 and Table 2 indicates that real-time semantic integration is better achieved when the enterprise pipeline treats quality, meaning, context and governance as interrelated responsibilities. Figure 2 shows that the quality score increases from approximately 58.4 to 88.5 during ingest-to-serve, indicating a significant improvement across the processing stages. It's a significant operation, as raw enterprise records often arrive with their labels incomplete and using local terminology. The increase in cleansing, linking, aligning, and serving indicates that each part offers a specific benefit over handing over the entire responsibility to final delivery. This interpretation is further reinforced in Table 1, which shows that the source quality reaches 91 and the context linkage reaches 88. The context linkage gain is particularly significant because when records are not connected, heterogeneous enterprise networks tend to fail. The meaning match has also shifted from 51 to 90, which is conducive to the main role of semantic integration. Latency preparedness improved from 62 to 93, indicating that the pipeline not only prepares records for reuse in real time but also makes them more interpretable. These values demonstrate that the architecture considers readiness and business understanding. The cycle-based view of improvement is added to Figure 3. Meaning match increases by 61 to 89, reuse preparedness increases by 54 to 83, exception abortion increases by 47 to 86 and controlled release increases by 52 to 88. The greatest increase is seen with exception closure (+39). This implies that through repeated refinement, the pipeline is more successful in managing conflict, missing cues and ambiguity of ownership. The following cycle has a balanced profile as well since all four lines do have ends above 80.

This balance is important, since a pipeline with good alignment of meaning but poor governance would result in risk, whereas good governance would slow business value in the case of a lack of reuse readiness. Table 2 indicates that enterprise functions vary in their roles in semantic integration. The overall profile of the operations hub is the most promising, with 126 matches in records and 92 in governance confidence. The finance office (121) and the sales network (118) have the same number of matches and the same confidence. The service desk has fewer matches (109) and lower flow readiness (76), suggesting that service records may need to be cleansed with greater care and mapped accordingly. The discussion thus suggests that the proposed pipeline enhances common preparedness, although there is a noticeable variation in the level of functions. This visibility is beneficial because it prevents managers from losing sight of the integrated enterprise’s view as they focus on underperforming areas. Table 1 and Figure 3 also show a correlation. Table 1 indicates that context linkage has shifted from 44 to 88, and Figure 3 indicates that the governed release has shifted from 52 to 88 across the cycles. These two readings converge on the same level of finality, implying that the greater the contextual connectedness, the greater the release confidence. Table 1 and Figure 3 are close, with 90 and 89 in the meaning match, respectively, indicating agreement between the staged and cycle views. The closeness is a sign of semantic improvement, apparent in the preparation for operation and in the reworking/making-over.

Latency preparedness is 93 in Table 1, which supports the argument that the pipeline enhances timely utilisation and the quality of interpretation. Thus, there are no isolated improvements in the evidence, but an integrated readiness pattern throughout the whole integration flow. The fact that a record with a source quality of 58 cannot be considered enterprise-ready, whereas a release state of 91 means the same kind of record can be reliable once prepared in a staged fashion. The starting point at 44 indicates a significant gap in raw enterprise networks, and the ending value of 88 demonstrates the capabilities of indexing and mapping to address that gap. Table 2 values for service desk are still below those of the other functions; hence, the work to be

improved would be better labelling of issues, clearer notes of ownership and improved speed of routing exceptions. It can be used as benchmark areas within the same organisation for operations and finance, as their governance confidence scores are 92 and 90, respectively. These results will help follow an effective approach to improvement rather than an overall assertion of equal performance.

5. Conclusion

When the intelligent data engineering pipeline takes in data from the source, applies quality filters, performs semantic mapping, contextual indexing, and delivers it with governance, the study concludes that real-time semantic integration in a heterogeneous enterprise network can be supported. The 132 internally prepared instances demonstrate the usefulness of raw records that undergo staged preparation. As seen in Figure 2, quality scores are higher at the serve stage than at the ingest stage. As seen in Figure 3, the quality scores are consistently higher at the meaning match stage, reuse readiness stage, the exception closure stage, and the governed release stage. Table 1 shows that this is very positive in terms of the improvements in source quality, context linkage, meaning match and latency readiness. As Table 2 indicates, the patterns of readiness for each function are stronger in operations and finance, but patterns across all functions are useful. The principal benefit of the pipeline is that it allows the maintenance of source identity and the generation of shared meaning. This equilibrium enables enterprise choices based on the local detail and integrated context. The design remains compact and traceable, enabling organisations that need real-time interpretation across multiple systems to benefit. The results also indicate that governance is not an add-on step but should be incorporated into the engineering flow. The proposed pipeline provides an obvious basis for semantic integration, and it is a more rapid, consistent and accountable way of handling data than operating with a collection of data islands.

5.1. Limitations

There are several limitations to the study. This data set includes 132 cases prepared in-house and can be used to demonstrate, but not to generalise, at the enterprise level. The source patterns were created to mimic realistic heterogeneous records; however, other, more complex access rules, legacy constraints, fuzzy ownership boundaries and weird record histories can occur in actual organisations. The pipeline was assessed using descriptive Indicators, Tables 1 and 2 and Figures 2 and 3 and not on a production network. Additionally, the architecture employs fewer components to maintain a clean design; specialised functionality, such as advanced security routing, deep lineage inspection, user role negotiation, and cross-region retention, is not represented as separate components. The result values can be compared across stages and functions but are not exhaustive for industry conditions. The other constraint is that it is not from an external source. Therefore, there is no comparison of the proposed pipeline with any designated commercial pipeline platform or published benchmarks. Despite the above restrictions, the paper offers a logical academic design and transparent interpretation of the results for real-time semantic integration.

5.2. Future Scope

Future studies can be carried out to further expand the pipeline in live enterprise trials in larger networks, longer observation times and to support more diverse business functions. Other applications that can be studied include supplier portals, customer experience platforms, audit repositories and distributed operational systems, which are tested to determine whether semantic readiness can be maintained under increased traffic. The pipeline can also be enhanced by implementing more robust privacy control, more comprehensive trace review, adaptive exception handling and role-aware delivery. Future designs might have a feedback element to enable business users to fix poor mappings and enhance subsequent outputs. A second helpful course of action is to get the approach tested in industry-specific applications, including banking, administration in the healthcare and education sectors, logistics and manufacturing services. Larger datasets can provide insight into the sustainability of gains at higher volumes, greater variety and complexity of sources, and more complex policies. Additional efforts can also be used to create dashboards that enable continual monitoring of readiness, unresolved exceptions and variation in manager-reported function levels. These extensions would help make the proposed design enterprise-friendly and governable in the long term.

Acknowledgement: The authors sincerely acknowledge KCG College of Technology, Easwari Engineering College, St. Joseph's Institute of Technology, University of Edenberg, University of Library Studies and Information Technologies, Duale Hochschule Baden-Württemberg, University of Electronic Science and Technology of China, and University of Engineering and Applied Sciences for their valuable academic support, research facilities, and institutional resources.

Data Availability Statement: The data supporting this study are maintained by the authors and may be obtained from the corresponding author upon reasonable request, subject to applicable institutional and ethical requirements.

Funding Statement: The authors confirm that this research and manuscript preparation were completed without receiving any external funding, sponsorship, grants, or financial assistance.

Conflicts of Interest Statement: The authors declare that they have no financial, professional, personal, or competing interests related to this research, and all relevant sources have been properly cited and acknowledged.

Ethics and Consent Statement: The authors confirm that the study was conducted in accordance with recognised ethical standards, with informed consent obtained from all participants and appropriate measures implemented to protect confidentiality and privacy.

References

1. M. Asfand-E-Yar and R. Ali, "Semantic Integration of Heterogeneous Databases of Same Domain Using Ontology," *IEEE Access*, vol. 8, no. 4, pp. 77903–77919, 2020.
2. B. Ben Mahria, I. Chaker, and A. Zahi, "A novel approach for learning ontology from relational database: from the construction to the evaluation," *Journal of Big Data*, vol. 8, no. 1, pp. 1–22, 2021.
3. R. Hammad, M. Barhoush, and B. H. Abed-Alguni, "A Semantic-Based Approach for Managing Healthcare Big Data: A Survey," *Journal of Healthcare Engineering*, vol. 2020, no. 1, pp. 1–13, 2020.
4. G. Li, "Improving Biomedical Ontology Matching Using Domain-Specific Word Embeddings," in *Proceedings of the 4th International Conference on Computer Science and Application Engineering*, Sanya, China, 2020.
5. S. Banerjee, T. Goto, N. C. Debnath, and A. Sarkar, "Ontology driven query language for NoSQL databases," in *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, Emden, Germany, 2017.
6. H. Zhang, Y. Guo, Q. Li, T. J. George, E. A. Shenkman, and J. Bian, "Data integration through ontology-based data access to support integrative data analysis: A case study of cancer survival," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Kansas City, Missouri, United States of America, 2017.
7. S. Zhao and Q. Qian, "Ontology based heterogeneous materials database integration and semantic query," *AIP Advances*, vol. 7, no. 10, pp. 1–19, 2017.
8. M. Gayathri and R. J. Kannan, "Ontology Based Indian Medical System," *Materials Today: Proceedings*, vol. 5, no. 1, pp. 1974–1979, 2018.
9. Y. He, H. Yu, E. Ong, Y. Wang, Y. Liu, A. Huffman, H. H. Huang, J. Beverley, J. Hur, X. Yang, L. Chen, G. S. Omenn, B. Athey, and B. Smith, "CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis," *Scientific Data*, vol. 7, no. 1, pp. 1–5, 2020.
10. E. Ong, L. L. Wang, J. Schaub, J. F. O'Toole, B. Steck, A. Z. Rosenberg, F. Dowd, J. Hansen, L. Barisoni, S. Jain, I. H. de Boer, M. T. Valerius, S. S. Waikar, C. Park, D. C. Crawford, T. Alexandrov, C. R. Anderton, C. Stoeckert, C. Weng, A. D. Diehl, C. J. Mungall, M. Haendel, P. N. Robinson, J. Himmelfarb, R. Iyengar, M. Kretzler, S. Mooney, Y. He, and Kidney Precision Medicine Project, "Modelling kidney disease using ontology: Insights from the Kidney Precision Medicine Project," *Nature Reviews Nephrology*, vol. 16, no. 11, pp. 686–696, 2020.
11. C. Reyes-Peña, M. Tovar, M. Bravo, and R. Motz, "An ontology network for Diabetes Mellitus in Mexico," *Journal of Biomedical Semantics*, vol. 12, no. 10, pp. 1–18, 2021.
12. K. Livitckaia, V. Koutkias, E. Kouidi, M. Van Gils, N. Maglaveras, and I. Chouvarda, "'OPTImAL': An ontology for patient adherence modeling in physical activity domain," *BMC Medical Informatics and Decision Making*, vol. 19, no. 4, pp. 1–15, 2019.
13. C. Shi, B. Hu, W. X. Zhao, and P. S. Yu, "Heterogeneous Information Network Embedding for Recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 357–370, 2019.
14. T. Y. Fu, W. C. Lee, and Z. Lei, "Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore, 2017.
15. Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Nova Scotia, Canada.
16. X. Wang, Y. Lu, C. Shi, R. Wang, P. Cui, and S. Mou, "Dynamic Heterogeneous Information Network Embedding with Meta-Path Based Proximity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1117–1132, 2022.