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

Leveraging Agentic AI for Cost-Effective Master Data Management: A Technical Framework

Viswakanth Ankireddi

Abstract: Master Data Management platforms have traditionally relied on expensive third-party providers to enrich organizational data with hierarchical, geographic, and contextual metadata. The emergence of agentic artificial intelligence technologies, particularly Retrieval-Augmented Generation systems and advanced prompt engineering techniques, presents transformative opportunities to reimagine data enrichment economics and architecture. This technical framework demonstrates how organizations can leverage publicly accessible business information through GenAI-powered extraction, synthesis, and validation processes to dramatically reduce dependency on costly commercial subscriptions. The proposed architecture combines vector databases containing embeddings of millions of public documents, sophisticated natural language processing pipelines specialized for corporate entity recognition, and a Master Control Point server infrastructure that orchestrates data flows while enforcing governance policies. In geographies that pose a huge challenge, like China, Russia, and emerging markets, where public data is inadequate, a hybrid model is strategically planned to combine low costs in the region and automated extraction features. The implementation process should be in phases, starting with proof-of-concept validation in data-rich jurisdictions, scaling to production, which focuses on quality assurance by multi-source cross-referencing, confidence scoring algorithms, and human-in-the-loop validation of low-confidence extractions. Organizations adopting this framework achieve substantial cost reductions while simultaneously improving data freshness, expanding coverage to underserved entity types and geographies, and building proprietary data assets that reduce vendor lock-in and create competitive advantages through superior business intelligence capabilities.

Keywords: *Master Data Management, Agentic Artificial Intelligence, Retrieval-Augmented Generation, Data Enrichment, Hybrid Data Architecture*

1. Introduction

The Master Data Management systems are the foundation of enterprise data ecosystems that assure consistency, accuracy, and control of essential business information within organisations. The issues related to the preservation of high-quality master data have grown even more complicated since organizations struggle to cope with the spread of data over numerous systems and locations. A study conducted by Gartner about data quality indicates that companies have serious inhibitors in their data management programs, and data quality is a major problem that has cost businesses a lot of money and time wastage in operations [1]. The data quality management is determined not just by simple accuracy checks, but also by completeness, consistency, timeliness, and the suitability of use in all sorts of business environments. The conventional treatments of these issues have been the third-party data vendors like Bureau van Dijk and Dun and Bradstreet, who provide costly hierarchical, geographic, and contextual metadata to customer and supplier data.

The economic burden of traditional MDM enrichment has been substantial for enterprises of all sizes. Forrester's Total Economic Impact study of Reltio's Master Data Management solution revealed that organizations implementing modern MDM approaches achieved significant financial benefits, with the study documenting a net present value of thirteen million dollars over three years [2]. This research underscores both the high costs associated with data quality challenges and the substantial returns possible through optimized data management strategies. However, much of the information traditionally sourced from premium providers exists in publicly accessible repositories

Independent Researcher, USA

such as business registries, corporate filings, government databases, and open web sources, creating an opportunity for alternative approaches.

The advent of advanced Agentic AI technologies, particularly Retrieval-Augmented Generation and sophisticated prompt engineering techniques, presents an opportunity to fundamentally reimagine the economics and architecture of MDM data enrichment. This article explores a technical framework for leveraging GenAI to extract, synthesize, and validate public business data at scale, dramatically reducing dependency on costly third-party subscriptions while maintaining or exceeding data quality standards required for enterprise operations.

2. Current State Analysis: The Cost Burden of Traditional MDM Enrichment

The current landscape of Master Data Management enrichment reflects a vendor-dependent ecosystem where organizations face mounting pressures to maintain data quality while controlling escalating costs. A study of statistics of data quality improvement as a result of ETL processes also shows that data quality has a huge influence on organizational performance, with research showing that data quality problems can influence the accuracy of the decisions made, operational efficiency, and customer satisfaction in various areas of businesses [5]. The financial impact goes well beyond what is directly needed in terms of subscription fees to include indirect costs in terms of data clean-up, integration complexity, and opportunity costs of making late or incorrect business decisions. Companies usually have subscriptions to several premium data providers, corporate hierarchies, and credit ratings, and other regional-specific providers, which becomes a tangle of vendor relationships that consumes significant coordination and administration overhead.

The overall cost of ownership of traditional MDM enrichment is not only base platform charges and per-record enrichment charges, but also considerable operational costs associated with integrating the system, repairing data quality, and managing its vendors. Data quality assessment methodologies reveal that comprehensive evaluation of data fitness requires systematic approaches examining multiple dimensions of quality, including accuracy, completeness, consistency, timeliness, and validity across diverse business contexts [6]. Organizations must dedicate substantial resources to validating vendor-supplied data, reconciling conflicts between different sources, and maintaining integration points with downstream systems. The hidden costs of data quality issues manifest through various channels, including customer service inefficiencies from incorrect contact information, supply chain disruptions from outdated supplier records, compliance risks from incomplete regulatory data, and strategic miscalculations based on inaccurate business intelligence. Despite these substantial investments, traditional providers exhibit notable limitations in coverage and quality, particularly for private companies in emerging markets, small and medium enterprises, and rapidly changing organizational structures such as mergers, acquisitions, and restructurings that occur between scheduled data updates.

Cost Category	Description	Impact Level
Base Platform Subscriptions	Annual licensing fees for premium data providers, including global and regional vendors	High
Per-Record Enrichment Charges	Transaction-based pricing for individual entity lookups and data retrievals	High
API Volumetric Pricing	Usage-based costs for programmatic data access through application interfaces	Medium
Integration Development	Technical resources required for connecting multiple vendor platforms to enterprise systems	High
Data Quality Remediation	Personnel dedicated to validating, cleansing, and reconciling vendor-supplied	Medium-High

	information	
Vendor Coordination Overhead	Administrative burden of managing multiple provider relationships and license compliance	Medium
System Maintenance	Ongoing support for integration points, version upgrades, and performance optimization	Medium

Table 1: Traditional MDM Enrichment Cost Components [3, 4]

3. Technical Architecture: GenAI-Powered Data Enrichment Framework

The technical architecture for GenAI-powered MDM enrichment centers on Retrieval-Augmented Generation systems that combine the reasoning capabilities of large language models with the factual grounding of external knowledge bases. Intel's comprehensive guide on implementing RAG systems emphasizes that successful RAG architectures require careful attention to several critical components, including knowledge base construction, retrieval mechanisms, and generation quality [7]. The foundation begins with building robust vector databases containing embeddings of public business documents from thousands of regulatory sources, government registries, and corporate filings worldwide. These vector representations enable semantic search capabilities that go beyond simple keyword matching to understand the contextual meaning and relationships within business information. The retrieval mechanism employs dense vector search algorithms to identify the most relevant document chunks for each enrichment query, followed by re-ranking processes that prioritize authoritative sources and recent information. The generation component utilizes advanced language models fine-tuned on business entity recognition tasks to extract structured information from unstructured text, maintaining strict adherence to predefined schemas while providing confidence scores for each extracted data element.

A comparative study of Retrieval-Augmented Generation chatbots demonstrates that RAG systems can effectively ground language model outputs in factual information from external sources, significantly reducing hallucination rates while improving response accuracy for information extraction tasks [3]. The implementation of RAG for business data enrichment requires several architectural layers working in concert to deliver production-grade results. The data discovery and acquisition layer continuously monitors thousands of public sources for new and updated business information, employing web crawlers that respect robots.txt protocols and rate-limiting requirements while maximizing coverage of available public data. The processing pipeline applies advanced natural language processing algorithms such as named entity recognition using specialized knowledge in corporate structures, extraction of relationships in subsidiary-parent hierarchies, and support of dozens of languages using multi-lingual processing of documents. Quality assurance mechanisms authenticate extracted information by cross-referencing to a number of independent sources, use anomaly detection algorithms to identify suspicious patterns that must be reviewed by a human, and full audit trails to show the provenance of each piece of data. The Master Control Point server uses centralized orchestration of the data flows, including governance policies, API access to the downstream systems, and caching layers that not only maximize performance but also minimize the computation costs caused by redundant computing.

Architecture Layer	Primary Functions	Key Technologies
Data Discovery and Acquisition	Continuous monitoring of public business registries, corporate filings, and regulatory disclosures	Web crawlers, API integrations, document processors
GenAI Processing Pipeline	Extraction of structured information from unstructured public sources using language models	RAG systems, fine-tuned LLMs, NER models, prompt frameworks
Data Validation and Quality Assurance	Cross-referencing multiple sources, confidence scoring, and anomaly detection	Multi-source validation, ML-based scoring, human review workflows
Master Control Point Server	Centralized orchestration, governance enforcement, and API management	Microservices, message queues, caching, metadata storage
Integration and Distribution	Bidirectional connectivity with MDM platforms and downstream business systems	REST APIs, event streaming, and data transformation services

Table 2: GenAI-Powered Enrichment Architecture Components [5, 6]

4. Hybrid Model: Addressing Complex Geographies and Data Gaps

The reality of global business information accessibility necessitates a hybrid approach that strategically combines GenAI-powered extraction of public data with selective use of paid commercial sources for geographic regions and entity types where public information proves insufficient. Best practices in prompt engineering emphasize that effective prompts require clear instructions, appropriate context, and well-defined output formats, with iterative refinement based on actual performance metrics being essential for optimizing accuracy and reliability [9]. For complex geographies such as China, where language barriers, fragmented regulatory databases, and data access restrictions limit the effectiveness of public source extraction, the hybrid model incorporates targeted subscriptions to regional data providers offering specialized coverage at substantially lower costs than global premium vendors. Similarly, for Russia, Middle Eastern countries, and various emerging markets where public data digitization remains incomplete and disclosure requirements vary widely, selective augmentation with commercial data sources fills critical gaps while maintaining overall cost efficiency. The key to economic optimization lies in intelligent routing logic that determines the most cost-effective data sourcing strategy for each entity based on geography, entity size, industry sector, and data availability indicators.

Data fusion and conflict resolution become paramount when combining information from diverse sources with varying levels of authority and reliability. Research on data fusion techniques highlights that effective integration of multi-source information requires systematic approaches to assessing source credibility, resolving conflicts through consensus mechanisms or hierarchical rules, and quantifying uncertainty in final outputs [8]. The hybrid architecture implements sophisticated scoring algorithms that assign reliability weights to different source types, with government registries receiving the highest confidence scores, followed by regulatory filings, commercial data providers, and finally unverified public sources such as company websites and news articles. In case of source conflicts on the same data element, e.g., corporate addresses or relationships with a parent company, the system uses the concept of multi-stage resolution logic where the temporal analysis is used to determine the latest information, the source hierarchy analysis is used to determine the source hierarchy, and finally, the consensus voting, which is used in case the sources concur. High-value entities that represent strategic customers, core suppliers, or major business partners would require the hybrid model to escalate to manual expert research, which would use human judgment and investigative abilities to resolve complex ownership structures, check regulatory compliance status, and provide intelligence not possible in an automated fashion. This tiered approach optimizes the allocation of resources by directing expensive manual research only where it delivers commensurate value while relying on cost-effective automated processing for the majority of entity enrichments.

Tier Classification	Entity Characteristics	Primary Data Source	Fallback Strategy
Tier One - GenAI Primary	Public companies in mature markets, government contractors, and entities with extensive disclosure requirements	Automated extraction from public registries, corporate filings, and open databases	Selective paid augmentation for missing fields
Tier Two - Selective Paid Augmentation	Entities in complex geographies, private companies in restricted markets, and credit and financial data requirements	Regional commercial providers, specialized data vendors	Manual research for critical gaps
Tier Three - Manual Research	High-value strategic relationships, complex ownership structures, sanctioned entities, and newly formed companies	Expert human investigation, specialized compliance screening	Premium provider consultation
Low-Cost External Datasets	Supplementary information across all tiers	Government open data, academic databases, collaborative platforms, and ethical web scraping	Crowd-sourced verification

Table 3: Hybrid Model Entity Sourcing Strategy [7, 8]

5. Implementation Strategy and Operational Considerations

Effective deployment of GenAI-enhanced MDM should have a well-planned staged process that balances technical risks and proves incremental value to the stakeholders, and develops organizational capabilities over time. Organizations implementing AI-driven data enrichment in contemporary environments benefit from automated data pipeline architectures that reduce manual intervention, accelerate processing cycles, and enable continuous improvement through machine learning feedback loops [4]. The journey typically begins with a proof-of-concept phase focused on a single geography with mature public data infrastructure, such as the United States or the United Kingdom, where success probability is highest and validation against existing vendor data enables clear performance benchmarking. During this initial phase, teams establish technical foundations including vector database infrastructure, GenAI model selection and fine-tuning protocols, integration patterns with existing MDM platforms, and quality assurance frameworks that will scale to full production deployment. The proof-of-concept should have success criteria that are clearly spelled out with minimum requirements on field completion rates, measures of accuracy on ground truth samples, processing throughput requirements, and cost per enrichment requirements that are indicative of economic viability.

Making a shift towards the production implementation rather than the proof-of-concept requires strict focus on the quality of data, governance structure, and reliability of work. The methodologies of data quality analysis have placed significance on an ongoing monitoring process covering various aspects such as completeness of the mandated fields, accuracy verified against authoritative sources, consistency within individual records and also across the entire data set, timeliness as a lag between the updates of the source and the availability of the MDM, and fitness to the purpose assessed through the feedback of the downstream business processes [6]. To implement the full-scale quality frameworks, there should be automated validation rules to raise potential issues to human attention, statistical sampling programs to give constant assurance to the correctness of the data, incident response procedures to counter any quality degradation, and continuous improvement mechanisms to leverage quality metrics to influence immediate optimization, model retraining, and prioritization of the source refinements. The data privacy concerns the governance architecture should cover include the GDPR compliance in the case of European organizations and the CCPA compliance in the case of companies based in California, the security

considerations such as encryption, access controls, and audit logging, and compliance with the web scraping ethics and terms of service compliance in the case of the publicly available data sources. Companies that manage to pass through such operating factors by retaining stakeholder trust with effective reporting on performance and responsive fixation of the issue, set themselves in a position to attain significant cost savings and capability augmentation that makes master data management no longer a cost center but a competitive ability that will make business decisions more swift by obtaining more exhaustive and timely business information.

Implementation Phase	Duration	Primary Objectives	Key Deliverables
Proof of Concept	Quarter to the tertile of the year	Technical feasibility validation, quality benchmarking, cost modeling	RAG system prototype, quality metrics baseline, economic viability assessment
MVP Production	Biannual to triannual period	Geographic expansion, core integration, automated workflows	Multi-country coverage, MDM platform connectivity, quality assurance framework
Scale and Optimization	Annual to sesquiannual timeline	Full production deployment, hybrid model implementation, and advanced features	Comprehensive geographic coverage, selective vendor integration, and real-time monitoring
Advanced Capabilities	Sesquiannual to biennial horizon	Innovation deployment, predictive analytics, and continuous improvement	ML-based quality prediction, anomaly detection, and custom industry models

Table 4: Implementation Phase Deliverables and Success Criteria [9, 10]

Conclusion

The transformation of Master Data Management enrichment through Agentic artificial intelligence represents a fundamental shift in how organizations acquire, validate, and leverage business intelligence. Traditional reliance on expensive third-party data providers, while delivering comprehensive coverage, has created unsustainable cost structures and vendor dependencies that limit organizational agility and innovation. The technical framework presented demonstrates that publicly accessible business information, when combined with advanced RAG architectures, sophisticated prompt engineering, and robust quality assurance mechanisms, can deliver comparable or superior data quality at substantially reduced costs. The hybrid model recognizes realistic geographic complexity and data availability limits and logically includes selective commercial sources where public information is not adequate, but overall remains cost-effective economically. Companies that use this framework have realized a variety of strategic benefits beyond direct cost savings, such as faster data refresh cycles, which allow better, more timely making of business decisions, wider coverage of a broader range of previously underserved entities and geographies, less vendor lock-in which increases negotiating leverage, and the development of proprietary data assets that can be seen as the basis of substantial competitive advantages. Master Control Point architecture offers the much-needed orchestration and governance, with seamless integration with the existing enterprise systems and audit trails, as well as keeping up with the privacy standards. Success requires commitment to phased implementation beginning with proof-of-concept validation in favorable geographies, investment in quality assurance infrastructure, including multi-source validation and human review workflows, and organizational development of capabilities in artificial intelligence, data engineering, and governance frameworks. As Agentic AI technologies continue advancing with improving accuracy and decreasing costs, early adopters position themselves to capture compounding advantages through proprietary knowledge bases, optimized processing pipelines, and institutional expertise that competitors will struggle to replicate. The economic argument is convincing as the time

to break even is calculated in terms of quarters as opposed to years, and pay-offs on investment are multiples of the initial spending in a normal planning duration. In addition to the financial indicators, there is the strategic value, such as changing master data management into a dynamic capability that speeds up business intelligence and improves the quality of decision-making, and provides a sustainable competitive advantage in the ever-more data-driven markets.

References

[1] Gartner, "Data Quality: Best Practices for Accurate Insights". [Online]. Available: https://www.gartner.com/en/data-analytics/topics/data-quality

[2] Reltio, "Total Economic Impact Study Finds Reltio's Modern MDM Delivered 366% ROI". [Online]. Available: https://www.reltio.com/resources/press-releases/forrester-total-economic-impact-tei/

[3] Kalindi Vijesh Parekh et al., "A Comparative Study of Retrieval-Augmented Generation (RAG) Chatbots," 2025 International Conference on Automatics, Robotics and Artificial Intelligence (ICARAI), 2025. [Online]. Available: https://ieeexplore.ieee.org/document/11137956

[4] SuperAGI, "Mastering AI-Driven Data Enrichment in 2025: A Beginner's Guide to Automating Your Data Pipeline", 2025. [Online]. Available: https://superagi.com/mastering-ai-driven-data-enrichment-in-2025-a-beginners-guide-to-automating-your-data-pipeline/

[5] Donal Tobin, "Data Quality Improvement Stats from ETL – 50+ Key Facts Every Data Leader Should Know in 2025," Integrate, 2025. [Online]. Available: https://www.integrate.io/blog/data-quality-improvement-stats-from-etl/

[6] R.A. Jonker, "Data quality assessment," Compact, 2012. [Online]. Available: https://www.compact.nl/articles/data-quality-assessment/

[7] Intel, "Implement Retrieval-Augmented Generation (RAG) to Accelerate LLM Application Development," 2024. [Online]. Available: https://www.intel.com/content/www/us/en/goal/how-to-implement-rag.html

[8] Xin (Luna) Dong and Felix Naumann, "Data Fusion – Resolving Data Conflicts for Integration," VLDB, 2009. [Online]. Available: https://lunadong.com/publication/fusion_vldbTutorial.pdf

[9] LaunchDarkly, "Prompt Engineering Best Practices". [Online]. Available: https://launchdarkly.com/blog/prompt-engineering-best-practices/

[10] Philip Beaucamp et al. "Information-Theoretic Cost–Benefit Analysis of Hybrid Decision Workflows in Finance," Entropy (Basel), 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12385591/