Frisco Analytics

Implementing Master Data Management at Scale on Databricks with AI

Whitepaper

Authored by: Roz King, Chief Architect at Frisco Analytics

Table of Contents

Executive Summary 2
Introduction 3
Understanding Master Data Management 4
Challenges of Master Data Management at Scale 5
Databricks as a Platform for Data and AI 6
Databricks Capabilities for Data Quality and Governance 7
Leveraging AI and Generative AI in Master Data Management 8
Implementing MDM Workflows on Databricks with AI 10
LakeFusion: A Databricks-Native MDM Solution 11
LakeFusion's Role in Streamlining MDM on Databricks 12
Conclusion 13
Cited Sources 14

Executive Summary

Master Data Management (MDM) is paramount for establishing a unified, trusted "**golden record**" of core business entities [2, 5]. It is essential for driving **data quality**, enhancing **operational efficiency**, enabling **better decision-making**, and ensuring **compliance** [4, 5]. However, implementing MDM at scale faces considerable challenges, including tackling **poor data quality**, integrating **fragmented data** across silos, managing **complex data transformations**, meeting stringent **governance requirements**, and achieving operational **scalability** [5, 6].

The **Databricks platform**, founded on the **Lakehouse architecture** [9, 10, 19], offers a unified environment uniquely suited for large-scale data processing and management [9]. It provides essential capabilities like **Delta Lake** for reliable storage with **ACID transactions** and **schema enforcement** [13, 20], **Delta Live Tables (DLT)** for building robust data pipelines with built-in quality checks using expectations [13, 14], and **Unity Catalog** for centralized data and AI governance covering access control, auditing, and lineage [11, 12, 15]. Databricks also provides **Lakehouse Monitoring** for integrated quality tracking [16].

Crucially, **Artificial Intelligence (AI)** and **Generative AI (GenAI)** are revolutionizing MDM by automating many labor-intensive tasks [7, 8]. AI/ML automates data discovery, quality processes (profiling, cleansing, standardization), and, most importantly, sophisticated **matching and merging** for entity resolution [7, 8, 25]. GenAI further streamlines workflows and enhances user interaction [24, 27]. While AI significantly boosts automation, a **hybrid approach** combining AI insights with human oversight is vital for handling complex cases and building trust in the mastered data [28, 29, 30].

Implementing MDM workflows on Databricks is effectively achieved using the **Medallion architecture** (Bronze, Silver, Gold) [17, 18] to progressively refine data quality [31]. The **Silver layer** is typically where core MDM processes like matching, merging, and golden record creation are executed [18, 43], leveraging Databricks' scalable compute.

Addressing the need for a dedicated, powerful solution, **LakeFusion, developed by Frisco Analytics, is a native MDM platform built specifically for Databricks** [1, 2, 3]. LakeFusion directly addresses large-scale MDM challenges by leveraging **AI**, including GenAI models, for advanced **entity resolution**, matching, merging, and standardization [1, 3]. Operating seamlessly within the Databricks ecosystem, LakeFusion utilizes the Medallion architecture for data flow and Unity Catalog for governance [3]. Its workflow automates the creation of unified golden records by consolidating data from raw sources through structured matching and review processes [3].

Frisco Analytics acts as a key thought partner for Databricks customers [1], combining expertise in AI-driven insights and advanced analytics with the LakeFusion platform [1]. This partnership accelerates **time to value**, improves **data quality**, streamlines data consolidation, enhances efficiency, and supports compliance [1], enabling organizations to effectively manage master data at scale and accelerate their data and AI initiatives on Databricks [1].

Introduction

Master Data Management (MDM) stands as a cornerstone practice for modern data-driven enterprises. It is the discipline of creating a single, authoritative, and consistent view – often called a "**golden record**" – for crucial business entities such as customers, products, suppliers, and locations [4, 5]. By integrating, de-duplicating, reconciling, and enriching data from numerous internal and external sources, MDM establishes a reliable **source of truth**, essential for improving data quality, enhancing operational efficiency, supporting compliance, and enabling confident decision-making [4, 5, 37].

Implementing MDM at scale, especially in complex organizations, presents substantial challenges related to data fragmentation, inconsistent quality, complex integration requirements, and the need for robust governance [5, 6]. Successfully tackling these requires a modern, scalable, and intelligent platform.

The **Databricks platform** is widely recognized as a unified data intelligence platform that converges data, analytics, and AI capabilities [9, 10]. Built on the Lakehouse architecture, Databricks integrates features for ETL, BI, AI, and governance [9], making it well-equipped to handle the demanding data processing and analytics requirements of MDM at scale [31]. Its ability to handle large batches and near real-time processing is particularly relevant for MDM workflows.

Furthermore, **Artificial Intelligence (AI)** and **Machine Learning (ML)** are increasingly vital in modern MDM, automating complex tasks like data cleansing and entity resolution [7, 8, 24]. AI capabilities enhance data quality, integration, and analysis [7, 8].

This report explores the implementation of Master Data Management at scale on the Databricks platform, demonstrating how its capabilities provide a strong foundation. Critically, it highlights how **LakeFusion, developed by Frisco Analytics, is a leading, native MDM platform** built specifically to leverage the Databricks ecosystem and address these challenges effectively [1, 2, 3]. LakeFusion utilizes **AI, including Databricks GenAI models, Large Language Models (LLMs), and Vector Search**, to automate and enhance core MDM processes like intelligent **entity resolution** [1, 3]. By operating natively within Databricks, LakeFusion aims to provide a streamlined solution for achieving data consistency and accelerating data and AI initiatives [1,

3]. **Frisco Analytics serves as a key thought partner** in this space, bringing expertise in AI-driven data management to enable customers' success on the Databricks platform [1].

Understanding Master Data Management

Master Data Management (MDM) is fundamentally about ensuring that critical business data is accurate, consistent, and reliable across an enterprise [4, 36]. It involves implementing technology, tools, and processes to manage and govern this crucial data [4]. The core output of MDM is the creation of a single, authoritative "**golden record**" or "**best version of the truth**" for key entities like customers, products, suppliers, and locations [5]. This is achieved by consolidating, de-duplicating, reconciling, and enriching data sourced from various internal and external systems [5]. Master data represents the foundational data of an organization, essential for operations and relatively static [4, 37]. It serves to enhance analytical and transactional data by providing consistent entity definitions. MDM directly tackles problems caused by fragmented, inconsistent, and inaccurate data spread across disparate applications [5].

The benefits of a mature MDM practice are substantial, particularly for large, global organizations or those with complex data landscapes [5]. It is considered a necessity for data-driven strategies [37]. Key advantages include:

- **Improved Data Quality:** Establishing a single source reduces errors, inconsistency, and duplication [37, 36].
- **Enhanced Efficiency:** Automating processes and reducing manual data cleaning tasks boosts productivity [37, 4].
- **Reduced Integration Costs:** Centralizing master data simplifies integrating various sources [37].
- **Better Decision-Making:** Reliable, consistent data provides a trusted foundation for analytics and strategic choices [5, 37, 4].
- **Stronger Compliance & Risk Management:** Standardized, secured data management supports regulatory requirements and risk mitigation [5, 37, 36].
- **Elimination of Redundancy:** Conflicting data sources and duplicates are resolved [5].
- **Improved Customer Experience:** A consistent view of customers across touchpoints prevents disjointed interactions [5].

Implementing MDM requires integrating several disciplines:

- **Data Quality:** Cleansing, transforming, and repairing data inconsistencies [5, 4].
- **Data Governance:** Defining policies, standards, and roles for data stewardship and ownership [5, 4, 38].
- **Data Integration:** Connecting disparate data sources effectively [5, 4, 38].

- **Technical Capabilities:** Data modeling, hierarchy management, **matching and merging** algorithms, workflow automation, versioning, and ongoing maintenance [4, 5].

Reference data, such as codes or categories, is a type of master data that requires management within an MDM framework [5, 4, 36]. Common master data domains include customer, product, supplier, employee, and location data [5, 4, 36, 37].

Ultimately, successful MDM implementation is a strategic, business-driven initiative that integrates people, processes, and technology, requiring active business participation and change management [5, 4].

Challenges of Master Data Management at Scale

Implementing and maintaining Master Data Management (MDM) at enterprise scale, particularly within large organizations characterized by numerous disparate systems and vast data volumes, poses significant challenges [6, 35]. These complexities can make achieving the benefits of MDM seem like a daunting task [6].

A foremost challenge is addressing persistent **poor data quality**, which manifests as inaccuracies, inconsistencies, incompleteness, and particularly, a high prevalence of **duplicate records** across various systems [6, 7, 41]. Resolving these issues often requires substantial manual effort [6], and poor data quality directly undermines data analysis, operational efficiency, and decision-making [6, 41]. Creating a **single source of truth** or "golden record" is complicated by the existence of multiple conflicting records for the same entity, hindering a unified 360-degree view [6, 7]. Maintaining data accuracy as sources evolve is an ongoing battle [7].

Integrating data across numerous, often siloed, source systems presents another significant hurdle [6, 7]. Complex data modeling, mapping, and transformation logic are required to reconcile diverse data structures and formats [6]. Large-scale **data migration**, especially from legacy systems, adds further complexity, involving meticulous planning, cleansing, and validation to overcome issues like incomplete or corrupt data.

Navigating intricate **data governance** requirements, such as defining policies, roles, and procedures, is fundamental yet challenging [35, 6, 7, 39]. Adhering to evolving **regulatory compliance** laws like GDPR, CCPA, and HIPAA, which demand strict data protection, access control, and transparency, adds significant complexity [40]. **Data security**, preventing unauthorized access and breaches, is paramount [41, 39], requiring measures like encryption, access controls, and data masking [39].

Operational hurdles include managing the explosion in **data volume** [39] and ensuring the **scalability** of MDM processes, which can be limited by traditional systems [6]. Effective **change management** and user training are often underestimated but vital for successful adoption [6, 39]. Fostering **collaboration** across disparate departments, each with its own data practices, is essential for unified MDM and governance [6, 35, 39].

Overcoming these large-scale challenges requires a modern, scalable, and intelligent MDM platform, ideally one that is tightly integrated with the underlying data infrastructure.

Databricks as a Platform for Data and AI

Databricks positions itself as the data intelligence platform that unifies data, analytics, and AI [9, 10]. At its core is the **Lakehouse architecture** [9, 10, 19], designed to bring together the best features of data lakes (cost-effective storage, flexibility) and data warehouses (structure, governance, performance) [10, 19]. This architecture aims to simplify data estates by breaking down traditional silos [9] and provide a highly scalable foundation for building, deploying, sharing, and maintaining data, analytics, and AI solutions effectively [9, 10].

The platform integrates a comprehensive set of capabilities essential for modern data management:

- **ETL and Orchestration:** Supports robust data pipelines for both batch and streaming data [9, 42].
- **Serverless SQL Analytics:** Delivers low query latency and high reliability for demanding BI and analytics workloads, leveraging **Databricks SQL warehouses** [9, 21].
- **Real-time Analytics, AI, and Applications:** Provides the infrastructure to support real-time data processing and build sophisticated AI applications [9, 10].
- **Collaborative Data Science:** Offers a platform for teams to work together on data science and machine learning projects at scale [9].
- **AI Capabilities:** Enables applying advanced analytics and ML techniques directly to data within the platform [10]. This includes features like **AI Functions**, built-in functions that apply AI models (e.g., text translation, sentiment analysis) directly on data tables, scalable for production use cases [22, 34].

Databricks is built on **open source and open standards**, embracing technologies like **Delta Lake**, Apache Spark, and MLflow [9]. **Delta Lake** provides the optimized storage layer with crucial features like **ACID transactions** and **schema enforcement**, ensuring data reliability and performance [13, 20]. Apache Spark underpins the platform's ability to handle massively scalable workloads through distributed computing [10, 42].

A key component for enterprise data management is **Unity Catalog**. This unified **governance** solution covers data and AI assets across Databricks workspaces [10, 11, 12], offering

centralized **access control**, auditing, and lineage [11, 12, 15]. **Delta Sharing**, an open protocol, allows secure sharing of live data without costly replication [9, 10].

By supporting diverse data formats and processing massive volumes through its distributed computing engine, Databricks provides a powerful and scalable environment [9, 10, 42]. This robust foundation is essential for implementing demanding data processes like Master Data Management, especially when seeking to leverage advanced capabilities like AI.

Databricks Capabilities for Data Quality and Governance

Establishing and maintaining high **data quality** and enforcing stringent **governance** are non-negotiable requirements for effective Master Data Management. The Databricks Lakehouse Platform provides a suite of built-in features specifically designed to address these crucial pillars, underpinned by **Delta Lake** for fundamental data reliability [13, 20].

Delta Lake Features for Reliability and Management

Delta Lake brings **ACID (Atomicity, Consistency, Isolation, Durability) transactions** to data lakes, ensuring data integrity and consistency, particularly vital for updates and upserts in MDM workflows [13, 20]. ACID transactions guarantee that concurrent read and write operations are reliably handled, and changes are atomic, complete, and isolated [13]. The built-in **Time Travel** feature allows accessing previous versions of a table based on timestamp or version, simplifying rollbacks to recover from errors or analyze historical states [13]. The `VACUUM` operation supports data retention policies by removing older, unneeded data files [13]. Delta Lake's **high concurrency** capabilities enable multiple pipelines to load data quickly and efficiently into a single table [13].

Ensuring Data Validity and Structure

Schema Enforcement (Schema Validation) is a core Delta Lake feature that automatically blocks data writes if the schema of the incoming data does not match the target table's schema [13, 20]. This prevents unintended data type mismatches or introduction of unexpected columns [13]. For handling evolving data sources, **Schema Evolution** allows Delta tables to automatically adapt by adding new columns [13]. More controlled changes are possible with Schema Overwrite or manual updates [13]. Auto Loader, used for data ingestion, includes features like schema hints and controlled schema evolution to assist with data validity [13].

Delta Live Tables (DLT) for Pipeline Quality

Delta Live Tables (DLT) simplifies the construction of reliable data pipelines and significantly enhances **data quality** management through its declarative approach [13, 18]. DLT allows defining **Expectations**, which are built-in validation and integrity checks applied to data flowing

through pipelines [13, 14]. These expectations use SQL boolean statements per record [14]. Users define how violations are handled using the `ON VIOLATION` clause:

- **`WARN`**: Records are kept, but violations are reported in logs and metrics [14].
- **`DROP ROW`**: Invalid records are automatically discarded [14].
- **`FAIL UPDATE:`** The entire pipeline update fails and rolls back the transaction [14]. DLT pipelines automatically log data quality metrics and violation counts, which can be monitored via the UI or by querying the automatically logged event logs [13]. Grouping expectations and implementing **quarantining logic** for bad data are also supported [13].

Unified Governance with Unity Catalog

Unity Catalog acts as a unified **governance solution** for data and AI assets across Databricks workspaces [11, 12]. It delivers centralized, ANSI SQL-based **access control** that applies granular permissions on catalogs, schemas, tables, and views [11, 12, 32]. Access is managed centrally at the account level and enforced across all attached workspaces [11, 12]. This is fundamental for securing sensitive master data. Unity Catalog automatically captures user-level **audit logs** for all actions against the metastore, available via system tables, providing transparency and traceability essential for compliance [11, 12, 15]. It also offers built-in, runtime **data lineage tracking**, capturing how data assets are created and used across queries down to the **column level**, including links to related notebooks, jobs, and dashboards [11, 12, 15]. Lineage data is consolidated across workspaces and accessible programmatically or in the Catalog Explorer [15]. For **data discovery**, Unity Catalog enables tagging and documenting data assets and provides a search interface [11, 12]. Its capabilities are foundational for modern data governance on Databricks [27, 34], providing the strict controls necessary for master data.

Monitoring Data and Model Quality

Lakehouse Monitoring is an integrated service for monitoring the statistical properties and quality of data in Delta tables and ML models registered in Unity Catalog [13, 16]. It provides automated profiling, data drift detection, and visualization via an auto-generated dashboard [16]. Customizable metrics and alerts can be set up and linked to Unity Catalog lineage for root-cause analysis [13].

These comprehensive capabilities provide a powerful environment for managing the reliability, quality, and governance of data, forming the essential bedrock upon which sophisticated MDM processes can be built and managed at scale.

Leveraging AI and Generative AI in Master Data Management

The complexity and scale of modern data environments necessitate innovative approaches to Master Data Management (MDM). **Artificial intelligence (AI)** and **machine learning (ML)** offer

significant potential to transform MDM processes, automating labor-intensive tasks and improving accuracy, particularly for large and diverse datasets [7, 8, 28, 29]. The increasing volume and complexity of data highlight the need for AI-driven automation [7, 8].

AI and ML techniques automate many traditionally manual MDM activities:

- **Data Discovery and Onboarding:** Automating the identification and classification of potential master data across sources, including mapping fields to data models using techniques like semantic tagging and NLP [7, 8, 30].
- **Data Quality Management:** Automating profiling, cleansing, standardization, and validation using ML and NLP to improve accuracy, scalability, and productivity [7, 8, 24, 29]. AI can suggest and enforce quality rules [7].
- **Automated Matching and Merging:** This is a critical application for **entity resolution** and creating the "**golden record**" [7, 8, 24, 25, 29]. AI algorithms identify duplicate records and recommend consolidation using advanced techniques that can combine rule-based and machine learning approaches [7, 28]. AI-native data mastering significantly automates curation [28, 29].
- **Relationship Discovery:** Automatically identifying and mapping relationships between different master data domains, contributing to knowledge graphs [7, 24, 30].
- **Data Governance Support:** Automating the mapping of business glossary terms, policies, and data ownership to master data assets [7, 24].
- **Data Privacy and Security:** Identifying and classifying sensitive data and associating relevant policies and masking rules automatically [7].

Generative AI (GenAI) is particularly impacting data management by simplifying data discovery, understanding, and quality processes, boosting efficiency [24, 26]. Integrating Large Language Models (LLMs) can streamline profiling, modeling, and integration [24, 28]. LLMs also improve the user experience by enabling natural language interactions [27]. This aligns with the trend towards more user-friendly, potentially AI-driven, self-service MDM tools [24].

A prime example of a modern platform leveraging these capabilities is **LakeFusion, which incorporates AI, including Databricks GenAI models, LLMs, and Vector Search, for advanced entity resolution, matching, and merging processes** [3]. This showcases how AI is directly applied within a native Databricks MDM solution.

While AI and GenAI offer substantial advantages, it's important to note that complex or ambiguous edge cases in data matching often require human judgment and domain expertise [28, 29, 30]. A **hybrid approach**, combining automated AI analysis with human review and validation, is crucial for ensuring accuracy, building trust in the mastered data, and addressing the limitations of purely automated systems [28, 29, 30].

Implementing MDM Workflows on Databricks with AI

Implementing Master Data Management (MDM) workflows on the Databricks platform leverages its robust capabilities and architectural patterns to process and curate data at scale. The **Lakehouse architecture** on Databricks provides a structured approach through its **Medallion architecture** (Bronze, Silver, Gold layers) [17, 18], which naturally aligns with the staged refinement of data required for MDM [31].

- **Bronze Layer:** Serves as the landing zone for **raw data** from various sources, ingested incrementally with minimal changes [17, 31].
- **Silver Layer:** Data is cleansed, validated, and conformed here [17]. This is the layer where core MDM processes, such as **matching, merging, and deduplication**, take place to create the "**Enterprise view**" and foundational "**golden records**" for business entities [18, 43]. Data quality checks are actively enforced [17].
- **Gold Layer:** This final layer provides curated, refined data, often aggregated and optimized for specific consumption patterns (BI, analytics, AI) [17, 18]. It combines the golden records from the Silver layer with other relevant data to answer business questions [43].

Building these MDM workflows on Databricks requires implementing complex logic for data quality, standardization, survivorship (determining the best value for an attribute from conflicting sources), and entity resolution (matching records). While organizations can develop entirely custom code for these processes or use generic accelerators [35, 45], this can be complex and time-consuming at scale.

This is where a dedicated, native platform offers significant advantages. **LakeFusion is purpose-built as a native MDM solution on Databricks, providing pre-built, AI-powered capabilities** that accelerate the implementation of these critical MDM workflows within the Medallion framework [1, 3]. LakeFusion's platform is designed to handle the complexities of matching, merging, and standardization, leveraging the power of Databricks' underlying infrastructure.

LakeFusion fully utilizes Databricks' processing power, including technologies like Spark and Photon, which are essential for applying sophisticated matching and merging algorithms efficiently across large datasets [3, 42]. It relies on Delta Lake's **ACID transactions** and data consistency features to maintain the integrity of master data [13, 20]. LakeFusion can also integrate with data quality enforcement mechanisms like Delta Live Tables (DLT) **Expectations**, helping to ensure only quality data enters the MDM process [13, 14].

Leveraging AI further enhances these workflows. LakeFusion's **Match & Merge functionality utilizes AI, including Databricks GenAI models and LLMs, along with Vector Search**, to perform intelligent entity resolution [3]. This goes beyond traditional deterministic or probabilistic matching rules to identify potential duplicates based on semantic similarity and other advanced techniques, improving accuracy, especially with noisy or inconsistent data. AI can assist in

standardizing values (e.g., recognizing variations of names or addresses) and inferring relationships.

The decision to implement MDM on Databricks with AI can involve choosing between a custom build, leveraging accelerators, or deploying a native platform like LakeFusion. While building allows maximum customization, it requires significant development and maintenance effort. Accelerators provide starting points but still need extensive customization. **LakeFusion offers a faster time to value and a more comprehensive feature set for core MDM processes, leveraging native Databricks capabilities for scalability, performance, and governance.** It provides a structured workflow for bringing data from Bronze sources into curated Golden Records, automating key steps while allowing for necessary human review in complex cases [3].

LakeFusion: A Databricks-Native MDM Solution

Addressing the specific need for effective, scalable, and governed Master Data Management within the Databricks ecosystem, **LakeFusion, developed by Frisco Analytics, stands out as a truly native MDM platform** [1, 2, 3]. It is highlighted as the only MDM solution purpose-built to run entirely within the Databricks environment [44]. This native integration is a key differentiator, ensuring seamless operation, maximum performance, and full leverage of the underlying Databricks infrastructure.

LakeFusion's primary objective is to provide a **single source of truth** and create accurate, unified **golden records** by consolidating fragmented data from diverse sources [1, 2, 44]. It is designed to manage critical business entities including **customer, product, supplier, and patient data**, among others [2, 3].

Key capabilities that LakeFusion brings to MDM on Databricks include:

- **Advanced Entity Resolution and Deduplication:** Utilizing sophisticated algorithms to accurately identify and resolve duplicate records across datasets [1, 2, 3].
- **AI-Powered Match & Merge:** This core functionality leverages the power of **AI, including Databricks GenAI models, LLMs, and Vector Search**, to perform intelligent matching of records and streamline the merging process, handling complex variations and fuzzy matches more effectively than traditional methods [1, 3].
- **Standardization:** Implementing and enforcing predefined business rules to ensure data consistency and format standardization across attributes [2, 3].
- **Unified Workflow:** Providing a structured workflow to move data from raw sources (typically in the Bronze layer) through matching and standardization processes to create validated golden records (often stored in the Silver layer) [3]. This workflow often includes stages for human review of potential matches or conflicts [3].

LakeFusion is deeply integrated with the core Databricks architecture and governance framework. It aligns with the **Medallion Architecture**, facilitating the flow of data from Bronze to Silver and ultimately to Gold layers [3, 17, 18]. Crucially, it fully utilizes **Unity Catalog** for centralized governance, inheriting and applying **role-based access controls**, capturing automated **data lineage**, and generating **audit logs** for actions performed within the MDM process [1, 3]. This ensures that master data management is conducted in a secure, compliant, and transparent manner directly within the unified data ecosystem [1].

By operating natively within Databricks, all processing is performed in place, eliminating the need for data movement to external MDM systems and optimizing performance [1]. The platform is designed for **scalability**, leveraging Databricks' capabilities, including serverless compute options, to handle large volumes of data efficiently [1, 3]. LakeFusion is readily available for deployment and trial via the **Databricks Marketplace** [2, 3].

LakeFusion's Role in Streamlining MDM on Databricks

LakeFusion, developed by Frisco Analytics, is specifically engineered to streamline Master Data Management (MDM) processes and resolve pervasive data inconsistency challenges for organizations utilizing the Databricks Data Intelligence Platform [1, 3]. Its role is to provide a powerful, native solution that simplifies the implementation of complex MDM workflows at scale.

LakeFusion achieves this by providing advanced MDM capabilities directly on Databricks, leveraging the platform's strengths. Key contributions to streamlining MDM include:

1. **Accelerated Time to Value:** As a native platform available on the Databricks Marketplace [2, 3], LakeFusion can be deployed and integrated quickly [1]. It provides pre-built functionalities for core MDM tasks like matching, merging, and standardization [3], significantly reducing the time and effort compared to building custom MDM solutions from scratch on Databricks [3].
2. **Enhanced Data Quality and Consistency:** The platform's sophisticated **AI-powered entity resolution and deduplication algorithms** [1, 2, 3], including its use of **Databricks GenAI models** for the Match & Merge process [3], significantly improve the accuracy of identifying and consolidating duplicate records. The **Standardization** feature helps enforce consistent data formats, further improving quality and reducing errors [3]. This directly contributes to creating reliable **golden records** [1, 2].
3. **Streamlined Data Consolidation:** LakeFusion facilitates the process of unifying fragmented data scattered across various source systems into a single, cohesive view [1, 2]. Its workflow, which typically starts with raw data in the Bronze layer, guides users through the process of defining entities, mapping attributes, and applying matching logic [3].
4. **Improved Operational Efficiency:** By automating key MDM tasks, particularly the labor-intensive processes of matching, merging, and standardization, LakeFusion

reduces the need for manual intervention, freeing up data teams to focus on higher-value activities [1]. Leveraging Databricks' scalable compute ensures these automated processes can handle large volumes efficiently [1, 3].

5. **Robust Governance and Compliance:** LakeFusion's deep integration with **Unity Catalog** provides essential governance capabilities out-of-the-box [1, 3]. This includes centralized access control, automated data lineage, and audit logs, ensuring that the entire MDM process is secure, transparent, and supports regulatory compliance requirements [1]. By managing master data within the governed environment of Unity Catalog, organizations can confidently use this data for sensitive analytics and reporting [1, 3].
6. **Leveraging Databricks Ecosystem Power:** LakeFusion is designed to run all processing within Databricks [3], utilizing the platform's scalability, performance (e.g., Photon engine), and reliability features (Delta Lake ACID transactions) [1]. This eliminates the complexity and overhead associated with integrating separate, external MDM systems and moving data in and out of the lakehouse [1].

Frisco Analytics acts as a key thought partner for Databricks customers navigating the complexities of MDM. As a company specializing in **AI-driven insights and advanced analytics** [1], Frisco Analytics brings deep expertise in data management challenges and modern solutions. They collaborate with Databricks to offer a unified approach [1], guiding customers on how to best structure their data using the Medallion architecture, leverage Unity Catalog for governance, and apply AI techniques for MDM, all within the Databricks platform. Through LakeFusion, they provide the engineered solution that embodies these best practices, enabling businesses to move beyond discussions of abstract concepts to the practical reality of achieving scalable, high-quality master data management and unlocking actionable insights [1].

Conclusion

Master Data Management (MDM) at scale remains a significant challenge for enterprises, driven by the proliferation and increasing complexity of data from myriad sources [6, 7, 8]. Overcoming issues like **data fragmentation**, **duplicate records**, inconsistency, and regulatory compliance requires a robust strategy and powerful tooling [6, 7, 40]. Manual processes alone are simply insufficient for managing data quality and consistency effectively at enterprise scale [7, 8].

The **Databricks platform** provides the essential foundation for tackling these challenges head-on. Its **scalable cloud lakehouse architecture** unifies data storage and processing, eliminating silos and efficiently handling diverse data types and massive volumes [9]. Features like **Delta Lake** ensure data reliability and integrity [13, 20], while **Unity Catalog** provides centralized governance, access control, auditing, and lineage – all critical for securing and managing master data [11, 12, 15].

Furthermore, **Artificial Intelligence (AI)** has become indispensable for modern MDM. AI automates and streamlines complex processes like data discovery, quality checks, and sophisticated **matching and merging** for **entity resolution** [7, 8, 24]. Leveraging AI accelerates processing, minimizes errors, identifies hidden patterns, and enhances data quality and enrichment [8].

Bringing these capabilities together is the key to successful large-scale MDM on Databricks. **LakeFusion, developed by Frisco Analytics, is a dedicated, native, AI-powered MDM solution** built specifically to leverage the power of the Databricks Data Intelligence Platform and address the complexities of master data management directly within the lakehouse [1, 3]. LakeFusion provides pre-built, AI-enhanced functionalities for core MDM processes, including advanced entity resolution, matching, and merging, utilizing **Databricks GenAI models** and other AI techniques [3].

By deploying LakeFusion, organizations can move beyond building custom solutions or struggling with fragmented tools. They gain a scalable, governed, and AI-accelerated platform for creating and managing their **golden records**. **Frisco Analytics acts as a key thought partner** for Databricks customers, providing not just the LakeFusion technology but also expertise in AI-driven data management to ensure customers effectively leverage the Databricks platform for their MDM needs [1].

In conclusion, implementing effective Master Data Management at scale requires a robust platform and intelligent tooling. The Databricks Lakehouse platform provides the necessary foundation, and **LakeFusion by Frisco Analytics provides the specialized, native, AI-powered solution** to streamline MDM processes, accelerate time to value, and enable organizations to confidently manage their master data for better analytics and AI initiatives [1, 3].

Cited Sources

[1] https://www.globenewswire.com/news-release/2025/01/28/3016053/0/en/Frisco-Analytics-Partners-with-Databricks-to-Drive-MDM-and-Business-Value-for-Lakehouse-Architecture.html

[2] https://marketplace.databricks.com/details/502f2f9e-5cfc-482a-b2e8-e22c4024a83e/Frisco-Analytics_LakeFusion-Databricks-Native-MDM

[3] https://www.friscoanalytics.com/lakefusionmdm

[4] https://profisee.com/master-data-management-what-why-how-who/

[5] https://www.informatica.com/resources/articles/what-is-master-data-management.html

[6] https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/master-data-management-the-key-to-getting-more-from-your-data

[7] https://www.informatica.com/blogs/10-ways-ai-improves-master-data-management.html

[8] https://www.leewayhertz.com/ai-in-master-data-management/

[9] https://www.databricks.com/discover/pages/data-lakehouse

[10] https://docs.databricks.com/aws/en/introduction/

[11] https://docs.databricks.com/aws/en/data-governance/unity-catalog

[12] https://learn.microsoft.com/en-us/azure/databricks/data-governance/unity-catalog/

[13] https://www.databricks.com/discover/pages/data-quality-management

[14] https://docs.databricks.com/aws/en/dlt/expectations

[15] https://docs.databricks.com/aws/en/data-governance/unity-catalog/data-lineage

[16] https://docs.databricks.com/en/lakehouse-monitoring/index.html

[17] https://learn.microsoft.com/en-us/azure/databricks/lakehouse/medallion

[18] https://www.databricks.com/glossary/medallion-architecture

[19] https://www.databricks.com/glossary/data-lakehouse

[20] https://docs.delta.io/latest/delta-intro.html#delta-lakehouse

[21] https://www.databricks.com/product/databricks-sql

[22] https://docs.databricks.com/aws/en/large-language-models/ai-functions.html

[23] https://www.informatica.com/about-us/news/news-releases/2025/01/20250114-informatica-strengthens-databricks-partnership-with-native-genai-capabilities-for-databricks-data-intelligence-platform.html

[24] https://barc.com/data-management-trends-developments-2024/

[25]

https://www.reltio.com/resources/press-releases/reltio-expands-ai-ml-innovations-unveils-industry-first-pre-trained-ml-model-for-automated-entity-resolution/

[26] https://www.informatica.com/about-us/claire.html

[27] https://open.spotify.com/episode/3GVsDJ4TGd5AFiiUPVPXPm

[28] https://www.tamr.com/blog/transforming-data-management-insights-from-anthrty-deighton-ceo-of-tamr

[29] https://www.tamr.com/blog/master-data-management-challenges

[30] https://www.francescatabor.com/articles/2023/12/30/master-data-management

[31] https://www.databricks.com/blog/master-data-management-lakehouse-modern-approach

[32] https://learn.microsoft.com/en-us/azure/databricks/data-governance/unity-catalog/best-practices

[33] https://www.databricks.com/blog/machine-learning-unity-catalog-databricks-best-practices

[34] https://learn.microsoft.com/en-us/azure/databricks/data-governance/unity-catalog/get-started

[35] https://profisee.com/blog/master-data-management-implementation-styles

[36] https://www.stibosystems.com/blog/what-is-a-data-domain

[37] https://www.dataversity.net/what-is-master-data-management-and-why-is-it-important/

[38] https://www.revgenpartners.com/insight-posts/3-key-components-for-a-successful-master-data-management-program/

[39] https://www.actian.com/enterprise-data-governance/

[40] https://semarchy.com/blog/data-governance-regulations/

[41] https://semarchy.com/blog/important-mdm-concepts-for-healthcare-data/

[42] https://www.databricks.com/solutions/data-engineering

[43] https://profisee.com/blog/what-makes-data-consumable/

[44] https://azuremarketplace.microsoft.com/en-us/marketplace/apps/1741184698943.lakefusion_installable?tab=overview

[45] https://marketplace.databricks.com/details/110c3b94-207b-4563-9377-857c10c8df73/Frisco-Analytics_DataLake-Sync-Accelerator