Mastering the Essentials of Master Data Management

Introduction:

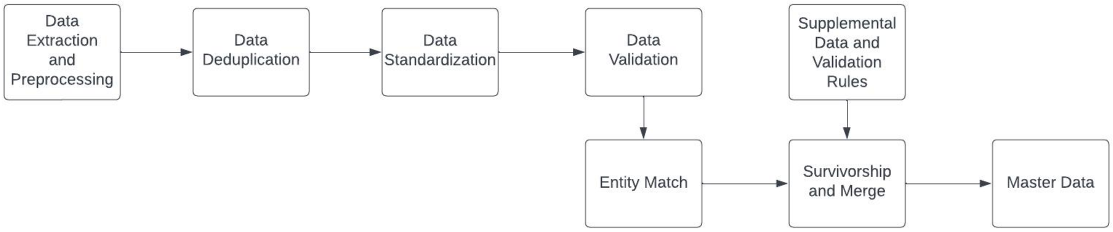
In a rapidly evolving world, data becomes the heartbeat of an organization, influencing strategies and fueling success. Organizations are becoming increasingly dependent on data for every aspect of their operations and recognizing the critical importance of having a centralized version of data.

Master Data Management is a comprehensive method of enabling an organization to link and merge all its critical data into one master record which becomes a common point of reference.

It is a crucial enterprise-wide strategy that treats data as a valuable asset and ensures data consistency, quality, stewardship, and workflow processes. MDM involves the techniques, policies, standards, and tools that consistently define and manage the critical data of an organization to provide a unified single point of reference. It centralizes data management and ensures data accuracy. [2]

In other words, MDM can streamline data sharing between personnel and departments within an organization. Even more than that, MDM can facilitate computing in multiple system architectures, platforms, and applications.

The goals of MDM are to assist in data integration, increase consistency and correctness, improve data quality, and reduce costs associated with data management. In this article, we will integrate provider data from data sources by applying master data management techniques of subjecting the data through extraction, validation, data standardization, Match, merge, and survivorship rules.



Data extraction and preprocessing:

This involves handling missing values and removing duplicates, and errors in the data. Apply data format rules to ensure the right format is applied across datasets.
For healthcare Data, provider Identifiers are always integers, dates are standardized in a single format, and text values are alphanumeric without any non-machine readable characters.

Building technical validation rules on data and capturing error statistics gives the opportunity to ensure improved data quality.

For example. Capture errors where a field that shows effective and termination dates are not equal and the termination date is always greater than the effective Date. Capture errors on Tax ID that are not 9-byte digits. These are some preliminary validation errors that can prevent corrupted data from entering the system.

Data Standardization

Given the diverse and complex nature of organizational data, inconsistencies in formats and structures can act as roadblocks to effective analysis.

For addresses, address verification and standardization tools like Melissa, Smarty, USPS, and GeoAPIfy provide ready-to-use solutions to validate and authenticate the addresses and give output in a standard format. IDQ address validation also provides outputs and checks that indicate if the address was valid or corrected or the likelihood of the address being deliverable.

It is essential for a healthcare organization to store accurate names of providers, and members as it is a crucial aspect of communication and personalization in enhancing customer experience.

Fuzzy matching algorithms utilize similarity functions between different data values and allow name matching for values with slight differences, variations, or errors. There are several popular fuzzy matching techniques:

- Levenshtein Distance (Edit Distance):
  This algorithm calculates the minimum number of single-character edits (insertions, deletions, or substitutions) needed to change one word into another. [3]

- Soundex:
  Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English. It's useful for identifying homophones. [1]

- Jaro-Winkler Distance:
  A measure of similarity between two strings. It’s a variant of the Jaro distance metric but with a boost for strings that have the same prefix.

- Trigram Matching:
  A trigram is a contiguous sequence of n characters from a given sequence. Trigram matching is very useful when searching text fields for terms that sound similar, due to typos or different possible spellings.

- Hamming Distance:
  It's used for error detection or error correction. It measures the minimum number of substitutions required to change one string into the other.

- FuzzyWuzzy Python Library:

This library uses Levenshtein Distance to calculate the differences between sequences in a simple-to-use package. These techniques can be used singularly or can be combined to develop robust fuzzy matching solutions. They all have their respective strengths and weaknesses, and the choice among them will usually depend on the specific application and business requirements.

Data match techniques

This technique refers to identifying and linking multiple records within the same dataset that correspond to the same entities.
Here are two ways of linking data:

a. Deterministic match:

Deterministic matching is a data matching technique where records are linked based on an exact match across one or several identifiers. This approach uses certain specific (and typically unique) attributes to find matches between databases.

For example, In the context of provider data, for practitioners, a match on the provider's profile can be based on NPI, for organizations profile can be matched on NPI, Tax Identifier, and organization specialty. Assign the same key to all the different variations of the same entity that match to each other based on identifiers.

b. Probabilistic match:

Probabilistic matching, also often called fuzzy matching, is a data matching technique that identifies and links records that are likely to correspond to the same entity, even if they do not match exactly. This method uses statistical models to assign likelihood scores to pairs of records based on the values of different matching variables.

Probabilistic matching often involves a manual review step for pairs with scores that fall within an uncertain range. This can increase the accuracy of the process, but it can also be resource-intensive.

Data hierarchy recognition

Data hierarchy recognition is the process of identifying and understanding the hierarchical relationships within a dataset. It involves recognizing and categorizing different levels of data, and how each level is interconnected or grouped.

In many cases, hierarchical relationships can provide valuable insights and can help in structuring the data more efficiently. In the healthcare context for provider data, data hierarchy might involve levels like provider profile, provider address, provider affiliations on address, and provider contracts with health plan on address.

Datasets should be segregated based on the domain they represent and establish a hierarchy within the domain. The datasets can be related to entities in other domains. It's important to set entity relationships trickling down through interconnected domains.

Business rules

Building a business rule framework engine is very crucial during every stage of data processing for several reasons as stated below.

Automation and Efficiency:

It enables the automation of decision-making based on predefined business rules, allowing for faster and more efficient operations.

Consistency:

By encoding and executing business rules automatically, organizations can ensure consistent application of policies and practices across different scenarios and departments.

Adaptability:

Business rules engines are designed to allow easy changes and updates to the rules without having to modify the core application code. This makes it easier for a business to adapt to changes in the environment, such as regulatory updates or new business strategies.

Transparency and Traceability:

The rules encoded in the engine can be inspected and audited, providing greater transparency for stakeholders and easier traceability for compliance needs.

Reduction of Errors:

Automated application of business rules helps reduce human errors that might occur during the manual execution of policies and procedures.

Integration and Interoperability:

These engines often have capabilities to integrate with other systems, allowing a seamless flow of data and helping maintain consistency across multiple systems.

Reduced Dependency on Technical Teams:

Some business rules management solutions allow non-technical users (like business analysts) to define, test, and even modify the rules, reducing dependency on IT or developer teams.

In a healthcare context, business rules will not just capture errors or perform verification of information, they can be extended to augment data.

In examples of provider data, primary care physicians can only have certain specialties. If they are not marked as a PCP, the rule engine can perform intelligent determination of their role. Selfdeclared portals like NPPES, and CAQH can be sources to enhance provider information that allows patients to make informed choices while selecting care.

The challenge in matching rules is the volume of data that needs to be processed. If there are too many variations of the same data, the process becomes resource-intensive. Performing data cleansing via business rules can reduce the volume of match processing.
For example, due to data entry errors, the same provider can be contracted to a health plan on different overlapping dates, making the dates consistent will allow faster processing of data.

Survivorship

Match survivorship, also known as record or data survivorship, is a concept in data management that refers to the process of determining which data values will be retained in the master record after duplicate records are merged.

Typically, this process is guided by certain business rules or data stewardship guidelines. Survivorship rules can be automated by building intelligent systems to read and apply systematically on datasets.

Several key points about match survivorship include:

Quality Importance:

The goal of match survivorship is to ensure that the most accurate, complete, and up-to-date information gets preserved in the consolidated record.

Rule-Based Decision Making:

Survivorship rules often dictate which source to prioritize when determining the surviving data element. For example, the rule could be as simple as "the most recent data wins", or it could be more complex involving the reliability or authority of different data sources.

Critical in Master Data Management (MDM):

Match survivorship is a fundamental part of MDM, where the objective is to create a single, unified view of data entities (like customers, products, or providers) across the organization.

Impact on Customer Experience and Operations:

The decisions made in match survivorship can have a significant impact on customer relations, operations, analytics, and legal compliance.

Bad decisions can lead to data quality problems, miscommunications, customer dissatisfaction, operational inefficiencies, and compliance issues.

So, in essence, match survivorship is part of the data matching and merging process that decides what data "survives" in the master record, ensuring the highest value and most reliable data is retained.

In a healthcare provider database, there might be duplicate entries for a physician with slightly different contact information. One entry might have an older office address, while a newer entry contains the current office address. Match survivorship would determine that the most recent address be retained in the ultimate, single record for that provider.

Another Example would be a provider organization that has a partial name and full name. Match survivorship would choose the full Name and be retained as a single record.

What is the Significance of MDM in creating a unified view of critical Business entities

Consistent and Accurate Data:

MDM ensures that critical business entities, such as customers, products, and employees, are represented consistently across the organization. This consistency eliminates discrepancies and ensures that data is accurate and reliable.

Single Source of Truth:

MDM creates a centralized and authoritative source of master data, often referred to as a "single source of truth." This ensures that everyone within the organization accesses and uses the same accurate information.

Improved Decision-Making:

By providing a unified view of critical business entities, MDM enhances decision-making processes. Decision-makers have access to reliable and consistent information, enabling them to make informed and strategic choices.

Efficient Data Integration:

MDM facilitates the integration of data from various sources. It harmonizes data formats, resolves inconsistencies, and standardizes information, making it easier to integrate and analyze data across different systems and databases.

Enhanced Data Quality:

MDM contributes to improved data quality by enforcing data governance policies and standards. This includes data cleansing, deduplication, and validation processes that ensure the accuracy and completeness of master data. Streamlined Business Processes: With a unified view of critical business entities, organizations can streamline their business processes. This includes more efficient customer relationship management, inventory management, and other operational activities.

Adaptability to Change:

MDM systems are designed to be adaptable to changes in organizational structures, mergers, acquisitions, or changes in business strategies. This adaptability ensures that master data remains accurate and relevant over time.

Facilitates Data Governance:

MDM contributes to effective data governance by establishing and enforcing policies for data management. This includes defining data ownership, access controls, and ensuring compliance with regulatory requirements.

Reduces Data Redundancy:

MDM eliminates data redundancy by consolidating duplicate records and ensuring that master data is represented in a non-redundant manner. This results in a more efficient and manageable data environment. Facilitates

Data Quality Improvement:

MDM processes often include data quality improvement initiatives. Addressing data quality issues at the master data level, MDM contributes to maintaining high data quality throughout the organization.

Supports System Integration:

MDM provides a foundation for integrating various systems within an organization. This is especially critical in environments where different systems need to share and reference the same critical business entities.

In conclusion, MDM is essential for organizations seeking to manage their critical business data effectively. By creating a unified view of critical business entities, MDM ensures data consistency, reliability, and integrity, laying the groundwork for informed decision-making and efficient business operations.

Cited resources:

[1]https://documentation.xojo.com/topics/text_handling/searching_text_using_the_soundex_algorithm.html

[2]A. Cleven and F. Wortmann, "Uncovering Four Strategies to Approach Master Data Management," 2010 43rd Hawaii International Conference on System Sciences, Honolulu, HI, USA, 2010, pp. 1-10, doi: 10.1109/HICSS.2010.488. keywords: {Marketing and sales;Conference management;Quality management;Information management;Management information systems;Taxonomy;Production;Information analysis;Error analysis;Raw materials},

[3] https://www.educative.io/answers/the-levenshtein-distance-algorithm