Whitepaper

Composable MDM Architecture

Authors -

Michael Ashwell
VP and GM Data Management,
Mastech InfoTrellis

Jason Kodish
Chief Strategy Officer,
Mastech InfoTrellis

The Evolution of Customer Master Data Management

Philosophers first asked the question thousands of years ago, then punted the question to artists, authors, and playwrights. Then, movie makers took the baton from them. Now, it has all come down to us – the data engineers and data architects - to answer the question, "What is a person?"

It has always been easy to talk conceptually about what a person is, what a business is, and what a product is. But, with the advent of high-scale, structured (and even semi-structured) data environments and architectures, the conceptual definitions need to become black-and-white rules that could be implemented with zero room for interpretation.

This concept of the persistent, immediate identification of a person, a business, an entity, or a relationship is the heart of all master data management, and all master data management, by extension, is the core of any data architecture.

Businesses traditionally had a limited view of the people and/or other business entities they encountered. Collection and management of critical domain master data typically revolved around a small set of well-understood B2C and B2B entities – people, businesses, locations, product, etc. generated from/through finite entry points into the master data environment. As enterprises increased their desire to manage critical data holistically, this led to creating processes to manage this data at an enterprise level - **Master Data Management.** Leveraging master data became increasingly complex due to companies collecting, processing, and storing this data across multiple processes and systems. While companies may have a good base level of master data (e.g., customer name, address, contact information), the exact attributes collected, associated data quality, and data duplication vary from system to system.

Most early efforts implementing master data management processes revolved around extracting known data from disparate systems, standardizing, and improving the data quality, merging and curating the resulting entities, and distributing the managed data back to the enterprise for more efficient and accurate execution of customer business processes.

Software vendors took note, which led to the creation of Master Data Management software platforms (i.e., MDM Platforms) and more specialized data domain solutions (e.g., Customer data Platforms). These platforms have become popular since the early 2,000s and typically provide an integrated set of MDM functions with a heavy focus on data ingestion, standardization, cleansing, matching, curating, and storing set data entities.

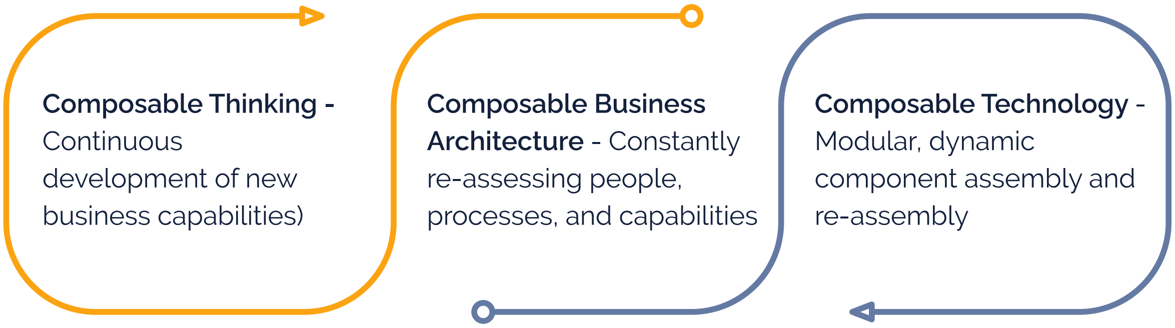
Depending on the specific platform they may also provide integration and distribution functions. However, these typically only work within a range of operational requirements. Larger enterprises usually must pair them with more sophisticated data engineering and integration solutions. Additionally, these platforms offer varying levels of flexibility on what entities (and attributes) are supported and what associated entity relationships can be ingested, discovered, and managed. While they can offer quick Master Data Management "solutions-in-a-box," they fall short of end-to-end enterprise solutions.

As companies began implementing their initial MDM Platform solutions and Master Data Management processes, the demand for complex business requirements grew. Simple management of known defined master data became the entry point for master data management. Businesses started pushing IT departments to manage increasingly complex entities and relationships. It was no longer enough to just have a clean, unified view of known customer data. Providing contextual collapse services (i.e., business unit specific and/or situational entity collapse and composition logic) became imperative. In addition to increasing complex entity types, requirements surfaced around managing complex relationships (e.g., households and social connections). Additionally, operational requirements increased dramatically. Batch data flow requirements evolved to near real-time event pub/sub demands, and eventually, large-scale real-time interface demands became the norm.

Our thesis: MDM software platforms are a tool in the broader arsenal – they are not the answer. Master Data Management is a process that includes multiple tools, platforms, rulesets, and conventions to bring order to chaos.

Composable Architecture Meets MDM

One of the largest evolutions in today's data architecture and strategy is the lean toward Composable Architecture. This fluid ecosystem contains independent systems and components that communicate with each other with the help of APIs. Gartner breaks down Composable Architecture into three parts:



Composable Thinking - Continuous development of new business capabilities)

Composable Business Architecture - Constantly re-assessing people, processes, and capabilities

Composable Technology - Modular, dynamic component assembly and re-assembly

The key is flexibility. Under the new focus on Composability, there are no "set-in-stone" decisions. Processes, tools, and points of connectivity are modular and flexible, which, by definition, makes them perishable. As a result, the supporting MDM capabilities and software that provide the necessary linkage within and around the superstructure must also be flexible, modular, and sentient to learn from new points of connectivity.

Today's Next-generation Master Data Management solutions are rapidly evolving to meet these new requirements. Vendor MDM Platforms are incorporating more sophisticated storage strategies typically encompassing multiple data store technologies (e.g., relational, graph, NoSQL, index Search) to accommodate entity/relationship flexibility and evolving to more sophisticated container-based platforms to support increased operational requirements.

Additionally, enterprises are incorporating vendor MDM Platforms into larger master data and integrated solutions involving multiple components dedicated to various aspects of the master data value chain. **In today's world, enterprises must think of Master Data Management as a dynamic, flexible process involving multiple components and not an MDM hub based on a single software product implementation.** Enterprise master data processes represent multiple distinct functions such as data ingestion, data standardization, data quality remediation, identity resolution, data storage, data stewardship, data publication, and data distribution. In addition, solving for multiple MDM domains that spread across multiple geographies often leads to multiple aggregated hubs and/or sophisticated data distribution solutions that isolate and bring data closer to the edge where decisioning occurs.

Early Wins in the Composable Architecture/MDM Marriage

It's worth noting that MDM platforms are often used in conjunction with multiple downstream custom master data hubs. Combining these components within enterprise data governance and integration architectures can provide tremendous benefits. Data can be fed into these downstream hubs via existing MDM solutions and secondary direct data sources (e.g., marketing prospect data, web interaction history, or social media data). Of particular importance to modern master data solutions are the Data Integration and Data Governance service layers. One advantage of creating a custom master data hub is that it allows an enterprise to closely tie master data functionality into broader enterprise IT strategies and architecture standards. Controlling the data integration interfaces allows a company to abstract vendor MDM Platform interfaces and rapidly deliver specific services for the downstream consumption of master data. It also allows for rapidly creating interfaces for newer data entity types and/or complex party profile objects.

Ultimately, the goal is to deliver an Enterprise Data Fabric that participates in an event-driven sense and response backbone, providing efficient, accurate, and timely data that enables critical business processes and decisions. This type of solution allows an enterprise to use vendor platforms to manage more well know entity types (e.g., customer) in combination with built-in data quality features (e.g., address standardization) and complex data steward functionality. The use of secondary custom master data hubs allows an enterprise to respond more rapidly to changing business requirements by combining traditional known customer data with other types of party data, creating more complex flexible entity types.

Entity Resolution and Complex Party Profiles

A crucial feature of modern Master Data Management solutions is their ability to map an incoming query with limited information to existing stored master data. This feature is used for both data ingestion and distribution. Traditional MDM platforms typically use a structured matching approach, which is highly effective at matching known data records such as customers and products. However, it may not work as well when trying to link secondary unknown entities with limited attributes or provide real-time non-obvious relationship information, such as whether a person has a social connection to an existing known customer.

There are two important requirements to providing the ability to support this type of complex query mapping functionality. The first is Entity Resolution, which in its simplest form, is the process of identifying an entity. Operationally, Entity Resolution matches multiple profiles for the same entity (i.e., Customer) within and across systems and eliminates duplication.

Conventional techniques for entity resolution involve applying deterministic and/or probabilistic matching rules on multiple records of an entity within a business line to create a master record for that entity. However, these methods have their shortcomings. For instance, they may be restricted to matching based on exact ID numbers, such as Social Security Numbers. Their accuracy may be lower than desired, even when fuzzy matches are used for fields such as names and/or addresses. Moreover, these techniques require considerable configuration efforts and frequently encounter performance and scaling issues.

Advanced entity resolution is a strategic asset that surpasses traditional approaches by using a golden profile for entity resolution rather than just a golden record. It generates a 360-degree profile in real time that provides an entity-centric view of each object and its relationships.

The entity resolution process is automated and involves several steps, such as standardizing, normalizing, validating, enhancing, and enriching data. Input records from various sources are processed to form feature sets that define the entity and generate additional or inferred features about it. These features are used to identify direct and indirect relationships of the entity and establish a contextual understanding.

The framework of blocking-matching-clustering is then used to identify candidate profiles for matching from a pool of resolved entities and match and link the entity to the closest cluster. The matching process uses pre-built advanced algorithms and open-source reference libraries for high-precision feature matching. Clustering generates groups of similar and linked entities, thus identifying duplicates/matches across sparse and heterogeneous user profiles.

Machine learning algorithms leverage probabilistic methods to progressively create the entity profile and resolve identities. The process successfully handles typographic, informational, and temporal variations and improves accuracy and completeness with time by self-correcting and reducing false positives and negatives.

This type of incremental framework also allows the addition of third-party data sources to enrich the profiles further with information about beneficial owners, hierarchies, and risk profiles. The enriched contextual awareness and complex relationship information added to basic demographic data can improve the accuracy and completeness of identity matching.

The incremental resolution framework is flexible and continuously procures the latest updates of the entity's profile from its data sources, which are added to existing resolved profiles. The process assigns an enterprise-wide persistent ID to each entity profile and links all identifiers across different LOBs through a "digital keyring model."

The second requirement involves creating, managing, and storing multi-dimensional profile entities that provide the base information for progressively building and managing an entity's data and all associated known relationships. This type of profile offers the flexibility to respond to business requirements and provide complex contextual entity-type service interfaces. For example, enterprises not only want to know the person involved in any encounter but the person within the encounter context. i.e., being able to respond with customer profiles that differentiate between John Smith on the company's website and John Smith using his phone app. The more master data solutions can customize entity interfaces to provide core, extended and contextual data about a party in real-time, the more effective enterprise responses and actions will be.

Integration and Operational Requirements

Changing master data business requirements has also led to changing operational or non-functional technical requirements. Integration with enterprise data engineering, integration, and governance infrastructure is a growing requirement, along with support for batch, near real-time, event publishing, and real-time interfaces. In large enterprises, these requirements can come with very stringent SLAs, response, and throughput requirements. While MDM software products can typically satisfy some of these requirements, MDM solutions must often be decomposed and implemented within larger enterprise architectures. To satisfy these requirements, enterprises must pay attention to multiple levels. Architecturally, asking system components to satisfy multiple differentiated workload types can lead to bottlenecks and unsatisfactory results. In larger master data solutions, separating component functions (e.g., ingestion of complex legacy data, data quality remediation, data stewardship user interface functionality, complex entity resolution/definition, real-time OLTP, and Query processing) into aligned component groupings can lead to better results. This was one of the initial motivations for solutions that separate data distribution functions into downstream data hubs, leaving curation ingestion/cleansing/curation functions to MDM Platforms.

Care must also be given to the associated technical implementation of individual components within the broader master data solution. Appropriate service style architectures should be utilized for components that handle integration, OLTP, and query processing duties, with preference given to container-based microservice implementations. In addition, the data storage implementation of any MDM Platform and/or data hubs will always be one of the most important components within the larger solution. The challenge with modern master data solutions is that the type of data stores that best support complex/flexible data entities (i.e., graph-based databases) can be a challenge to implement in a manner that supports a large enterprise's rigorous SLAs and non-functional requirements. Current best practice calls for deploying multi-storage type subsystems that support both complex flexible data models (i.e., typically a graph data store) and higher operational requirements (some combination of NoSQL, Index Search and/or Relational Datastore). Care should be given to all layers (e.g., physical, network, software) to achieve high-demand requirements.

Data store subsystem implementations should be abstracted through an integration layer with published service interfaces to provide flexibility in the underlying implementation. This means that consumers of MDM Platform and Data Hub data and services should only use the published integration interfaces and not directly access the underlying data stores.

In Summary

Enterprise MDM process requirements have evolved over time to require sophisticated multi-component solutions that work within broader enterprise architecture standards, especially in the world of Composable Architecture. Key MDM functions such as data ingestion, standardization, quality, stewardship, distribution, and entity resolution can be provided by separate or selectively integrated components. Composing MDM solutions in a flexible, efficient architecture can be challenging. Still, the payoff is the base for a well-engineered Data Fabric that enables business decisioning. Understanding the broader picture and how MDM solutions can be built up over time will avoid duplication and enable better business results.

Authors

Michael Ashwell

Michael is a seasoned professional with over 35 years of experience in enterprise architecture, solution development, cloud offerings, global sales, and consulting. He spent 30+ years at IBM where he held various roles, including leading the Data and Analytics Lab Services Cloud COE, and developed several key offerings. Michael has worked internationally across multiple industries and holds a BSBA degree in Information Systems from Xavier University.
In his free time, he enjoys home theater, making wood-fired pizza, and driving his restored 1962 TR3.

Jason Kodish

Jason Kodish, also known as Kodi, has more than 25 years of experience in strategy, transformation, data/analytics, and marketing. He has extensive experience creating data-driven business growth and leading diverse teams - through influence and innovation - from strategy to successful execution. Kodi has pioneered writing and speaking about the "value chain of data" while staying true to his mantra: "Do not be interesting... be instrumental."

In his long career, Kodi has worked globally in all markets and across all business verticals. Kodi's success in data and analytics services is based on his focus on the team and allowing them to grow while doing great work.

He constantly challenges his team to (1) be great, (2) make mistakes, and (3) have fun.

Mastech InfoTrellis partners with enterprises to unlock the value of their data by delivering data tothe people and machines where decisions are made. Our offerings - Data-in-motion, Data-as-an-asset, and Data Activation - activate data at scale and unleash the full potential of decision-making. As a minority-run organization with a presence in nine global locations, including the US, UK, India, and Canada, we have transformed over 250 organizations with our proven expertise in deliveringDigital Transformation.

Pittsburgh I New Jersey I Toronto I Chicago I Boston I Dallas I Orlando I London I Chennai I Noida

Let's Get in Touch

☎ +1 412.787.2100
📍 Pittsburgh, PA
✉ experience@mastechinfotrellis.com