Realizing the Value of MDM Through the Data Lake

Semarchy

A Whitepaper by Dirk Garner

Garner Consulting

©2018 Garner Consulting, Semarchy.
All Rights Reserved.

AUDIENCE

- Technology VPs and Directors
- Data VPs and Directors
- Enterprise and Solution Architects
- Data and Engineers
- Business Analysts
- Data Stewards
- Customer Experience Engineers

WHAT YOU'LL LEARN

Although it is common knowledge that MDM depends on effective data management for success it is less known that a traditional relational database management system (RDBMS) is not the only backend data storage system that can support Master Data Management (MDM). There are more emergent and flexible options regarding storage technologies and data pipelines available. In fact, the use of a Data Lake brings new efficiencies to MDM heretofore not possible when depending on RDBMS and Data Warehouses. This whitepaper will outline the path to delivering business value quickly by using agile MDM powered by your data lake and a performance-optimized semantic layer.

©2018 Garner Consulting, Semarchy.
All Rights Reserved.

EXECUTIVE SUMMARY

Master your Data in the Data Lake: your organization stands to reap significant benefits by mastering data within a data lake of co-located diverse data sets within a common ecosystem sharing processing and storage methods and reducing redundancy in data transport, processing, and storage.

Effective Data Lake Architecture: There is no standard data lake architecture but there are common approaches and methods proven to deliver the best results from data lakes and these principles are prerequisites to extract maximum value from the pairing of MDM with a data lake.

Future-Proof Your Organization: Implementing agile MDM within a flexible data architecture allows for quick remediation of the unknown and unforeseeable future requirements of your organization's evolution.

Cut the Cord: Avoid delays of onboarding data to the data warehouse as a prerequisite to MDM. By de-coupling ETL from an MDM project we can move to agile approaches and accelerate project inception and time to value for MDM.

Bottom Line: Through this approach, you improve data accessibility and data quality while avoiding information blindness by providing of all the data, all the time, all cleaned and ready for analytics, BI, data science, machine learning, and all of you data-driven initiatives.

WHY MASTER DATA?

The mastering of data is essential to earn the trust and confidence of business leadership. Data mastering is the process of combining data elements about similar entities from multiple data sources through processes of validation, enrichment, de-duplication, and record matching, to provide a single accurate perspective of each entity as represented by the data. Future business decisions, both large and small, will be based on analytics and reports based on your data and those reports are no better than the accuracy and completeness of the underlying information. Because of this, it is critical to profile, clean, and master your data prior to aggregating and summarizing it into decision support assets so that the resulting information is accurate and complete and instills confidence throughout the organization and reporting audience.

MASTER DATA SOURCES: THEN & NOW

Historically, MDM works best when functioning synergistically with primary systems of record whether accessing these in transactional systems or data warehouses. Data mastering initiatives have traditionally depended on integration with either or both transactional and analytical data stored solely in relational database management systems (RDBMS), most notably Data Warehouses to which transactional data has been copied and analytical workloads run upon.

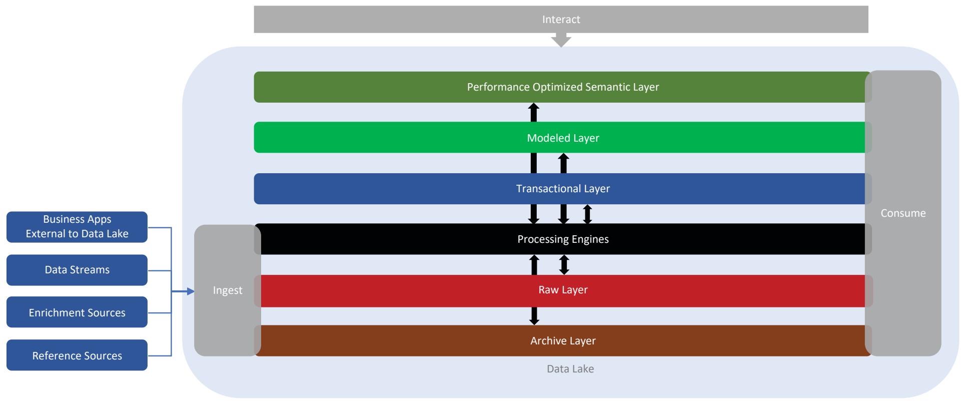
Data Warehouses rarely serve real time uses cases or other analytics on data in motion. Contrast that with the data lake design pattern in which flexibility of ingest, storage and consumption methods, serve a wide variety of use cases such as data science, machine learning, transactional processes, ERP, canned reporting and real-time analytics. When considering your data storage methods that support data mastering initiatives, it doesn't seem to be widely known that a data lake can support MDM just as well or better than traditional RDBMS methods, and data lakes can do so with lower cost and greater flexibility and scaling than a traditional data warehouse.

A well-designed data lake can serve all traditional use cases and many new use cases due to the flexible ingest, multiple processing methods, and multiple storage zones with each component being fit for purpose and integrated to complement each other forming a cohesive multifunctional scalable whole that far exceeds the capabilities of a traditional data warehouse.

ARCHITECTING THE VALUE OF A DATA LAKE

As big data pushed traditional data warehouses beyond their boundaries of effectiveness, efficiency, and capability, it became evident that something new was required to handle these growing needs. Thus, the data lake emerged with all the challenges and shortcomings of any adolescent technology or process and was subject to the associated growing pains, false starts, and failures that would be expected under these circumstances. The first data lakes tended to be rushed into place with little thought to architecture, capabilities, or scaling and these directionless approaches eventually led to the all too common perception that data lakes were nothing more than a dumping ground, a data swamp, or data graveyard or some other disparaging label implying that archiving all of the data without clear purpose or manner to consume was a meaningless pursuit. In most cases of failed data lake attempts the root cause could be traced to misaligned or non-existent business objectives, poor technology choices, or improper layer/zone strategy. It is unfortunate, but commonly true, that these early data lakes frequently were championed by technology rather than the business.

Data lakes can provide flexible ingest methods, flexible storage methods, coexistent transactional and analytical workloads, and performant data accessibility, all of which enable numerous business use cases, agile Multi Domain MDM among them. A data lake must serve the ingest needs of the all of the different types and velocities of data in use by the business, processing needs to stage the data the way the business needs it, and accessibility to meet the expectations of the business systems and analysts. Although specific data lake implementation architectures can vary widely, there are core principles that are key to an effective data lake and these are common across organizations with successful data lake implementations. These core aspects are listed here.



Data lakes can provide flexible ingest methods, flexible storage methods, coexistent transactional and analytical workloads, and performant data accessibility, all of which enable numerous business use cases, agile Multi Domain MDM among them.

©2018 Garner Consulting, Semarchy.
All Rights Reserved.

INGEST

The data Ingest zone includes all necessary methods of onboarding data to the lake. Capabilities might include ETL, ELT, flat files, FTP, streaming data, and other flexible data inflow capabilities. The target storage zones or layers of ingest processes would most likely be the Raw layer with some data possibly going directly to Archive or into Processing prior to storage or consumption. For example, data required for retention only, with little likelihood of retrieval, might go directly to Archive whereas reference data might be modeled through the Processing layer and move directly to the Modeled layer. Transactional data from external sources would likely target the Raw layer and then get remodeled through the Processing layer to be comingled with the internal transactional data in the Modeled layer and ultimately optimized in the Performance layer. Data from enrichment sources might be called on demand or stored in the Raw, Archive or Modeled layers for later use, modeling, or processing.

RAW

The two primary purposes of the Raw zone are 1.) to allow for quick and easy onboarding of data without lengthy ETL development and 2.) to store data in its original format with all original values to be consumed later as needed to serve any possible use case. Data in the Raw layer might be stored as flat files in a relational data store, or any number of popular NoSql methods. The Raw layer also serves to provide lower cost storage as compared to the Transactional, Modeled, and Performance layers. Much of the data within the Raw layer could be consumed via the schema on read method which is a unique and useful aspect of the data lake paradigm.

ARCHIVE

The Archive layer represents the lowest cost online storage intended to preserve data for retention requirements or low frequency retrieval. Backup copies of other zones might be stored here for short or long-term needs and this layer might participate in your high availability and/or disaster recovery strategy.

PROCESSING

The Processing layer is likely to include several complex processing capabilities such as in-memory processing, machine learning and AI methods alongside more basic functions such as data modeling and transport. As the name implies, if data is being moved, changed, or analyzed, it is being done using technologies in this zone. Data mastering rules such as matching, merging, and splits would be performed within the Processing layer with subsequent transport and storage functions carried out afterward. Other synergistic data governance functions such as data profiling, data cleansing, and metadata collection, would be performed in the Processing layer as well.

TRANSACTIONAL

The transactional zone is home to production applications such as ERP, HR, financials, and other applications used in running the business. The coexistence of data brought about by these applications residing in the same ecosystem as data governance, reporting, and analytical process provides the foundation for the advantages of basing your MDM program on a data lake.

©2018 Garner Consulting, Semarchy.
All Rights Reserved.

MODELED

The Modeled layer is home to clean, refined, complete, and accurate data modeled to fit various business requirements. In this layer you might have data warehouses, data marts, sandbox space for data science, OLAP cubes, and other targeted data assets. All data in this layer should be searchable via SQL and BI tools.

CONSUME

The Consume and Interact zones differ only in the way data is consumed. The Consume zone serves consumption through automation, or code. Example use cases would be consumption through ETL jobs, REST services calls, canned reports, flat file export, FTP, and stream or event publishing.

INTERACT

The interact zone is the entry point that serves human access to the data through BI tools, data science queries, SQL editors, and other data analysis methods. The primary difference between Interact and Consume is the access permissions and optimization strategy. The Interact zone mainly accesses the Performance layer and allows for a performant conversation with data, whereas the Consume layer access all layers as needed.

DEFINING THE PERFORMANCE OPTIMIZED SEMANTIC LAYER

Beyond those core aspects of a successful data lake, we'll need to consider the performance requirements of both read and write operations for all business use cases. The performance layer is also the point of business access presented through a semantic interface making it a key component since data from the Performance layer is consumable by a wide audience throughout the business and potentially external to the business.

The semantic layer represents a plain language view of data organized by business domains using business vocabulary. A semantic layer provides a navigable and understandable perspective of information and is an ideal output for MDM and other data governance initiatives through the delivery of clean accurate, whole data delivered in a user-friendly business-centric presentation appropriate for each MDM domain.

There are various manners of optimizing a semantic layer for performance depending on the underlying technologies and data structures. Examples of technologies that can provide performant semantic layers include OLAP databases, column stores, data federation or virtualization, and several of the NoSql data store offerings. Performance must be considered prior to the modeling of data for the semantic layer to take the best advantage of whatever technology capabilities are extant within your chosen platform. Choosing the right technology depends on several factors, among them: your current eco system, source and target data stores and transport methods, your budget, your staff's technical skillset, and vendor relationships.

©2018 Garner Consulting, Semarchy.
All Rights Reserved.

Operating your MDM program within a well-crafted data lake provides your organization with the unique opportunity to intercept upstream data in need of cleaning and mastering during the earliest move or model processes and make changes once, then propagate those changes throughout the lake rather than engineering multiple data flows to multiple locations prior to cleansing and mastering operations.

BETTER TOGETHER: MDM, DATA LAKE, & THE PERFORMANCE LAYER

In today's world, it's not enough for MDM to deliver intelligent methods, 100% accuracy and a user-friendly interface; keeping pace in today's competitive markets requires an advanced approach to multi-domain MDM that provides flexible individualized implementation styles. The data lake architecture approach described herein provides a synergistic relationship among the data lake, MDM and data governance in general. On one hand, data governance is more critical in a data lake than in a data warehouse due to the former having little control over ingest and the latter having rigid control on ingest. On the other hand, we don't want to impede the flexibility native to the data lake by imposing rigid data governance. Balancing these considerations is a key value point of this approach and will allow your organization to fully evolve its data governance program through agile multi-domain MDM implementations with the least effort needed in building and maintaining the required data processing pipelines.

Core to the principles of the data lake is the concept of schema on read. Underpinning this concept is the principle of acting on the most relevant data when needed, leaving other data at rest without effort or processing. In other words, the data processing tasks needed for data cleansing, modeling, and mastering need only be performed on the data required for your immediate business requirements saving countless hours, CPU cycles, and disk space by not processing the data not currently needed and that which may never be needed. In this manner we can pursue action on the highest ROI, highest priority data without the need to process all data just to act on the results of a tiny amount of that processing. The remaining, currently not needed, data can be left in place in the Raw layer until such a time, if ever, it too might be needed by the business.

Transactional and analytical workloads can be designed to take best advantage of the paradigm of 'land first -- process as needed'. The primary advantages of this approach are 1.) to rapidly get all data to a common location with the least amount of effort, and 2.) to act only on the necessary data in the most necessary means to meet the current requirements. If we have designed our data lake appropriately, the 'act on' part of that statement will be done by any one of several fit-for-purpose technology capabilities each chosen for its ability, usability, and performance characteristics.

With this approach, you have transactional data alongside reference data, raw data from numerous supporting systems, and analytical data, all in the same ecosystem using shared and common fit-for-purpose processing engines and storage zones. Here is where the greatest value of the data lake best serves your MDM goals. Operating your MDM program within a well-crafted data lake provides your organization with the unique opportunity to intercept upstream data in need of cleaning and mastering during the earliest move or model processes and make changes once, then propagate those changes throughout the lake rather than engineering multiple data flows to multiple locations prior to cleansing and mastering operations. The value of mastering data through first-touch cleansing and mastering at the source and making transformations, edits, and corrections alongside other transactional batch and real time methods allows us to act once, and only once, for each transform, modeling, or transport function without the need for redundant efforts repeated in multiple systems or on multiple copies of the data. By reducing redundancy and needing to act only once on data, you eliminate many of the problems encountered when mastering data and essentially all the errors involved in duplicative processing.

©2018 Garner Consulting, Semarchy.
All Rights Reserved.

ENHANCED MDM IMPLEMENTATION STYLES THROUGH DATA LAKES

Traditional MDM implementation styles include the Registry hub, Co-existence style, Consolidated approach and the Centralized Transactional hub. Although there are pros and cons associated with each approach, the data lake delivers new enhancements to these styles that can extend and improve technology capabilities and business agility.

REGISTRY

The output of a Registry MDM approach would be accessible via the semantic layer through the Interact or Consume zones and would potentially source input data from the Modeled, Raw and Archive layers by leveraging data processing methods from the Processing zone.

CONSOLIDATED

This MDM and Data lake pairing is essentially an organic Consolidated MDM style in which all data is available, whether copied or native, to a central location. The main value of this pairing is that additional copying is eliminated due to the preexisting coexistence of all the data within the same ecosystem with minimal processing required due to the 'touch early and touch once' approach described above.

COEXISTENCE

The coexistence style, in which the source systems are updated with mastered data, is also greatly simplified through the advantages of having all data organically coexistent without additional effort, but more importantly, the output of the mastering methods need only to store data in the modeled, raw and/or semantic layers within the data lake itself rather than undergoing additional costly and slow ETL processes to transfer data externally to the data lake.

CENTRALIZED

The Centralized, or Transactional, MDM style routes all references or calls to data that is mastered under the authority of the MDM governance program, are made to a single central location potentially through REST calls, SQL queries, etc. And like the other styles, we realize the advantages of all processes sharing and re-using the same data stores and processing engines.

With some or all the transactional and analytical data workloads co-existing in the data lake, any combination of these approaches in a multi MDM domain environment is organic, natural, and to some degree automatic.

SEMARCHY -- A PERFECT FIT FOR THE DATA LAKE

Architectures with flexibility can be tricky to achieve and you must align the capabilities of a data management vendor with the needs of the business; the skill set of the technology team; the expertise and capabilities of the business users; and maintain the flexibility to tailor to the nuances of your business without which, MDM would be unsuccessful. What you don't want is a lengthy customization process and one-off result that is difficult

MDM should be simple, clear, and easy to maintain.

©2018 Garner Consulting, Semarchy.
All Rights Reserved.

to understand and maintain and next to impossible to train new hires to use or maintain. MDM should be simple, clear, and easy to maintain.

Take the time to choose a vendor who will truly partner with you and takes responsibility for the implementation and success of their product. There are products that heap the brunt of the analysis and prep on the client leaving an already overburdened organization to put forth a large amount of effort into an area in which they may have very little expertise.

In this regard *x*DM from Semarchy stands out from the crowd through a collaborative approach to data governance and hybrid Multi-Domain MDM, and native compatibility with Data Lake.

Key highlights of the Semarchy difference:

- Hybrid, multi domain capabilities
- Flexible implementation styles
- Quick time to value
- Agile collaborative MDM
- API interface for consumption and updates

CONCLUSION

As more organizations adopt the data lake architectural design pattern and permeate it throughout their transactional and analytical workflows, new opportunities materialize regarding how MDM and other data governance initiatives add value through this co-existence of source and mastered data without the need to further refine data or build additional pipelines to move the data where it is needed.

Although you cannot eliminate the need for careful planning and preparation to complete a successful master data management initiative, you can significantly pare down the data preparation work required prior to making forward progress through the application of agility in MDM, effective data lake design patterns, and performance optimized semantic access layers. These three powerful forces converge to power your business goals in BI, Analytics, Data Science, Data Governance, self-service and beyond.

Report was sponsored by Semarchy
Semarchy is the Intelligent Data Management™ company. Its *x*DM platform is an innovation in hybrid, multi-vector data management that encompasses the capabilities of Master Data Management (MDM) and Collaborative Data Governance, as well as Data Quality, Enrichment and Workflows. The software leverages smart algorithms and material design to simplify data stewardship, governance and integration. Its platform is implemented via an agile and iterative approach that delivers business value almost immediately, and scales to meet enterprise complexity. The technology is in use at some of the most well-known brands in the US and Europe, supporting all data domains, integration styles, industries and use cases in one environment that adapts to evolving business requirements.

GARNER CONSULTING

Garner Consulting provides research advisory and consulting services supporting data strategy, technology architecture, and business evolution.

Learn more at
garnersoftware.com

+1 844-542-7637

info@garnersoftware.com

Semarchy

Learn more at semarchy.com

+1.650.240.2000

info@semarchy.com

©2018 Garner Consulting, Semarchy.
All Rights Reserved.