Paper SAS1822-2015

Master Data and Command Results: Combine Master Data Management with SAS® Analytics for Improved Insights

Ron Agresta, SAS Institute Inc.

ABSTRACT

It's well known that SAS® is the leader in advanced analytics but often overlooked is the intelligent data preparation that combines information from disparate sources to enable confident creation and deployment of compelling models. Improving data-based decision making is among the top reasons why organizations decide to embark on master data management (MDM) projects and why you should consider incorporating MDM functionality into your analytics-based processes. MDM is a discipline that includes the people, processes, and technologies for creating an authoritative view of core data elements in enterprise operational and analytic systems. This paper demonstrates why MDM functionality is a natural fit for many SAS solutions that need to have access to timely, clean, and unique master data. Because MDM shares many of the same technologies that power SAS analytic solutions, it has never been easier to add MDM capabilities to your advanced analytics projects.

INTRODUCTION

As frequent users of SAS software already know, the advanced analytic functionality found throughout SAS applications is unparalleled for its level of sophistication. The power of the results generated by any analytic process can be diminished or amplified by the quality of underlying data on which the results are based. It only stands to reason that poor data quality leads to inaccurate analysis, and that leads to poor decision making.

Master data management (MDM) is a means to combat these problems. MDM is a discipline that incorporates the people, processes, technologies, and methodologies for creating a complete, correct, and authoritative view of the enterprise as reflected in its core data elements. MDM is often called a "golden record" or "best record" creation process. For analytic processes, MDM provides a best record view of enterprise data assets to downstream business intelligence, reporting, and analytic environments.

SAS offers a broad range of MDM features built on best-in-class data quality and data integration functionality. Building MDM processing into your SAS analytics and reporting workflows can dramatically improve the reliability and usability of results.

WHO NEEDS MDM?

Improved data-based decision making, operational efficiency, risk management, and customer intelligence are among the top reasons why organizations decide to embark on MDM projects. Consider for a moment what advanced analytics projects might look like without robust MDM capabilities built in.

TRAPPED INSIGHTS

As you will learn, MDM is not usually a unidirectional process. The power of master data comes in sharing it with interested systems. If, for example, your analytic process generates a "customer value" score that approximates how likely a customer is to continue to use your company's products or services, it will not be much use to your organization unless all customer touchpoint systems have that metric available to them. Customer service representatives armed with this information can better cater to individual customer needs. Without MDM, exciting insights created by analytic processes that have the potential to drive operational interactions cannot be easily pushed to a myriad of operational systems. This type of system-to-system integration is too difficult to maintain as applications evolve. A centralized master data repository solves this problem by making the customer value score available to any system that interacts with it.

INCORRECT AGGREGATIONS

Many analytic calculations rely on categorical data to provide insights to the measures found in the source data. Categories can be used for aggregation. If your enterprise data systems contain data inconsistencies, precise analyses with aggregated data might be more difficult unless categorical differences found in the different systems can be resolved. Consider a simple example of a report you are designing that assumes you will have 50 correctly standardized and unique codes to represent all states in the United States. What happens to your results when invalid state codes are present? If report filters are built on predetermined state abbreviation lists, values associated with unknown state codes might end up being grouped in an “other” category. Worse still, the values might simply be dropped from your report altogether.

OPTIMISTIC RESULTS

Imagine an exercise in which you are tasked with reporting on the total number of customers that your company has, analyzing their behavior, and making predictions about what will happen when your customer base doubles in size. Now consider the reliability of your predictions if your customer count is understated or overstated by 25%. The source of this type of discrepancy can often be attributed to duplicate data in your customer-centric systems. Most organizations have several systems that record customer interactions or that store customer account information, and the contents of each are rarely in agreement. This is not uncommon—most companies design any number of data manipulation processes to reconcile diverse customer information before it heads downstream to analytics or reporting applications. Optimistic use of data elements assumed to be clean, consistent, and duplicate-free can lead to incorrect assumptions in analytic models, which feed skewed results to downstream systems.

UNDISCOVERED RISK

A common analytics-based activity is discovering and quantifying an organization’s exposure to risk and fraud. What if you cannot reliably identify unique companies or individuals in your various systems? Any effort you make to attribute potentially risky or fraudulent activities to known parties can break down when you cannot account for the common aberrations inherent in working with data that is pulled from different data systems. Master data processes can help you identify strong links for entity relationship analysis. This, in turn, helps your risk analysis process understand that it was not five different individual customers who withdrew inconspicuous amounts of money from their account—it was the same individual who withdrew now very conspicuous amounts of money from their account five times in quick succession, thereby flagging their account for investigative action.

NONCOMPLIANCE

Organizations of all sizes are increasingly encumbered by regulatory mandates that provide necessary transparency through reporting to customers, boards of directors, and regulating agencies. These regulatory requirements can turn even relatively simple data integration problems into tangled cross-division negotiations for resources. For example, while it should be straightforward for a bank to identify all clients with a total transaction volume above a certain regulatory threshold, the many systems that contain this data across different functional areas or lines of business all define “client” slightly differently. If a common representation of the client is not implemented, any report generated for compliance purposes must be called into question as no one can be quite sure if the calculated values found in the report represent the true state of affairs. Consequently, erroneous accounting made on this untrustworthy data could have far reaching consequences.

MASTER DATA

MDM DEFINED

To fully understand and take advantage of the benefits of MDM, you first must become familiar with standard MDM functionality and corresponding methodologies. MDM is a discipline that includes the people, processes, technologies, and methodologies for creating an authoritative view of core data elements that empower critical enterprise operational and analytic systems. This core data is typically described as the “nouns” in your business, such as customer, supplier, patient, health care organization,

or product. It's also important to know what master data is not. Generally speaking, actions between or among two or more nouns in your business (a customer "buys" a product) and that are usually referred to as transactions are not master data. You will find that data that is separated in silos are ripe candidates for mastering. Each data silo has its own business rules and data structures, and these might not be compatible from system to system. MDM frees data held in isolation and combines it with like data, thus multiplying its utility.

While there are many different approaches to MDM and just as many use cases, one common and convenient way to delineate MDM initiatives is by the way data flows through your information technology landscape. The two most common categories of use cases are separated into operational MDM and analytic MDM. Operational MDM generally means that as you manage master data, you are integrating master data elements back into the systems that were the source of the data in the first place. The unique customer best records, which contain a combination of data elements from your billing system, call center, and reservation system, are made available to those systems, so they might benefit from the collective view of key customer information. Operational MDM usually incorporates both batch and real-time processing.

Analytic MDM, on the other hand, still generates entity best records in much the same way as an operational MDM framework would, often in batch processes. But rather than providing a round trip for the generated master data back to source systems, analytic MDM usually pushes the master data in one direction toward a data mart or other repository that provides source data to analytic, decision management, or reporting systems.

The means to discover and manage master data are the same in both operational and analytic MDM, but the use of the master data is quite different. Both approaches can be used together, and indeed they often are combined by more technologically mature organizations. For the purposes of this paper though, we focus on analytic MDM because it is in these types of use cases where we commonly see the beneficial relationships between master data and advanced analytic processes.

MDM COMPONENTS

Regardless of the style of MDM in play, profiling, cleansing, data matching, and best record creation (also known as survivorship) are core components of any MDM program. These processes create master data and allow it to be managed over time. Almost any MDM technology that you use, including SAS, includes the following critical functionality:

- Profiling – mark "JON MCALLISTER" as non-null STRING data 14 characters long
- Parsing – transform "JON MCALLISTER" into "JON" | "MCALLISTER"
- Casing – transform "JON MCALLISTER" into "Jon McAllister"
- Identification – categorize "JON MCALLISTER" as "NAME"
- Verification – augment "650 S Griffin St, Dallas, TX" with a postal code of "75202"
- Standardization – transform "Dallas Texas 752021200" into "Dallas, TX 75202-1200"
- Geocoding – augment "650 S Griffin St, Dallas, TX 75202" with "32.775340, -96.802039"
- Matching – determine "JON MCALLISTER, 650 S Griffin St, Dallas, TX" to be a match of "JONATHAN MCALLISTER, 650 Griffin Street Dallas Texas 752021200"
- Survivorship – choose "214-356-8890" over "356-8890" because it contains more information
- Monitoring – generate an alert when an update to "JON MCALLISTER" sets a corresponding gender code to "F"

Once the key data in your systems has been through this data workflow with business rules turned to your data domain, it will have been corrected, matched, and best records will have been created. Good MDM applications wrap all of these core feature areas into an integrated and repeatable process. The general gist of this process is probably not foreign to many long-time SAS users. Most everyone who has worked in SAS over the years has used clever DATA step code to clean up data ahead of analysis

processes and maybe even created one or more macros that can be called on demand from different programs. In some ways, this is similar to master data creation.

The critical difference is a concept called data governance, which provides the repeatability, transparency, and auditability that does not naturally come with embedding data quality and data matching business rules directly in the data processing code. Instead, the logic that provides the smarts to standardize German addresses or parse Spanish names comes from a shared repository, so that any data process uses, and by extension enforces, an approved set of enterprise data quality and matching standards. For example, for industries that are heavily regulated, being able to provide documentation on data transformation behind earnings numbers is not a luxury but a necessity.

The “what” of MDM is relatively standard regardless of whether you plan to develop MDM functionality yourself, work with vendors to find a best of breed solution, or look to existing technology in your portfolio that might provide the core elements needed for mastering data. The “where”, however, can be quite different, depending on your overall information technology landscape.

MDM processes can run most anywhere: on a stand-alone MDM server, in relational database stored procedures, in Hadoop, or in SAS code, just to name a few places. The requirements of your MDM project will dictate how you architect your MDM data flows. But whatever way you choose to implement the core data quality and matching functionality for your project, without data governance you are very likely to spend a lot of time on the initial MDM data load only to have it fall apart as additional data from additional systems with addition business requirements is added. Data governance engenders trust when all stake holders understand how the data they use for important decision support and reporting processes came together.

MDM ADOPTION

If you have not integrated MDM functionality into your SAS analytic solutions yet, don’t feel bad—you are in the majority. Not many organizations claim to have MDM programs fully integrated with their analytics or reporting environments. Those that do have already started down the road toward MDM maturity. Many understand that implementing MDM is an opportunity to improve already productive processes but might not know where to start or might not be disciplined enough in their overall information management strategy to feel confident about achieving the results that MDM promises.

One way to gauge your organization’s readiness for MDM is to perform a maturity model evaluation. There are many variations of information technology maturity models available to help companies understand their ability to successfully adopt new technologies. Below is one such simplified model. Use these phases and recommendations to help guide your adoption of MDM.

Level 1: Unaware

Ignorance is bliss. You do not yet realize the magnitude of your data management problem. You are not concerned about correcting data issues because you have not noticed or felt the effects of poor data quality on your quarterly sales reports or stock price. Enjoy it while it lasts.

Recommendation: You probably have not started to evaluate MDM technologies or even reviewed MDM literature in general. Now would be a good time to start.

Level 2: Reactive

Inventory cannot be reconciled with sales data, and you believe it to be a data problem, not an “it fell off the back of the truck” problem. It’s good that you understand what is happening, but you can only react to the latest data disaster. You can merely fight new fires for so long before you run out of hoses.

Recommendation: Look to turn your ad hoc data management corrections into a formal MDM initiative. Use data profiling techniques to help identify potential master data sources and pinpoint problems. Discuss data issues with your proto-data governance council, made up of data owners, domain experts, and key business stakeholders.

Level 3: Managed

Through careful data profiling and process improvement, you now have a handle on the definition of master data in your organization. Emergency data issues are fewer and farther between, which has left you time to build or acquire a robust framework for enterprise master data. Not all systems are integrated but the roadmap is beginning to take shape.

Recommendation: Seek executive sponsorship and codify your data governance policies. Continue to fine-tune MDM processes and evaluate the ability of your MDM framework to handle additional systems, data types, and means of interaction. Adjust accordingly.

Level 4: Strategic

Your organization has validated and approved an MDM framework and has committed full-time resources to a data governance committee that has full executive support. The MDM system plays an integral part in your organization's decision support and reporting infrastructure. The improvements to these environments are demonstrable and you are ready to declare victory.

Recommendation: Like any good framework, your MDM processes need to be able to adjust to a changing environment and still deliver expected benefits. Now is not the time to polish your trophy. Instead, adopt a stance of continuous improvement. Refine existing MDM processes and plan for new systems, new master data domains, and new technologies.

COMMAND RESULTS

Now that you understand what MDM is and the perils of ignoring it, and you have assessed your institutional readiness for MDM adoption, you are ready to jump into MDM. Full MDM projects can become quite complex and often involve many different groups and technology systems. This can be a deterrent to getting started, so begin with a small project and work toward a quick win. Select one report that is in need of improvement and follow these steps.

1. Measure current processes –Without creating a baseline using current processes, you will not be able to measure the impact of using MDM techniques. Start by looking for reports that rely on the nouns of your business. For example, find or generate a report that shows the number of unique suppliers that your organization has and record the amount spent with each vendor last quarter.
2. Profile source data –Understand where candidate master data is and gauge its overall quality. You will need to bring into alignment the data in different systems that represents suppliers, so look for areas of agreement in your data by reviewing profile reports and begin to tease out potential data quality and matching rules.
3. Define master data –Supplier information has been located through careful analysis of the profile results. Agree on the core data elements of the supplier domain and define an MDM target structure around these data elements.
4. Use core MDM features –This is the heart of your MDM process. The earlier profile results will turn into data quality and matching rules. Standardize data, identify duplicate records, and generate best records for each matched set of data. These best records should be a close approximation for the total number of unique suppliers you do business with. It will not be exact—no match process is perfect—but with careful profile analysis, data quality remediation, and an appropriate match strategy you should get pretty close.
5. Combine master data with transactions – You can answer interesting questions with master data alone (like the count of unique suppliers), but in most cases master data is combined with transactional data before analysis is performed. In this example, the now clean and uniquely identified suppliers will be joined with transactions for all contributor records to each match cluster. This will

allow for calculation of the amount spent with each supplier while accounting for differences in the supplier data that might have otherwise prevented an accurate result.

6. Run analyses –Rerun your spend analysis report with the MDM processing in place.
7. Compare results – Evaluate the differences between the initial report and the one with MDM built in. Where poor data quality, a high level of data duplication, and several data sources are present, the improvements are likely to be dramatic.

CONCLUSION

As you have learned, MDM is a natural fit for many SAS solutions that need access to timely, clean, and unique versions of master data such as parties (individual, organization, company, patient, physician, investigator, and so on), products (part, financial asset, automobile, and so on) or locations (site, store, plant, school, warehouse, and so on). It should come as no surprise that SAS has all of the core MDM features available within its software portfolio when high quality data plays such a critical role in producing high quality analytic results that users have come to rely on from SAS software.

Because these MDM enabling features are built on many of the same technologies that power SAS analytic solutions with metadata shared between them, it is a straightforward exercise to add MDM processing to your SAS advanced analytics projects. As you have seen, there are real benefits to incorporating MDM techniques into the data preparation phase of your data analysis or reporting project, especially when data bound for your report comes from more than one source system. Whether it's improved lift on marketing campaigns, smarter fraud detection, or more precise forecasting results, integrating MDM with SAS analytics makes good business sense.

REFERENCES

Rausch, Nancy, et al. 2015. "What's New in SAS Data Management." *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings15/SAS1390-2015.pdf.

Rineer, Brian. 2015. "Garbage In, Gourmet Out: How to Leverage the Power of the SAS® Quality Knowledge Base." *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings15/SAS1852-2015.pdf.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ron Agresta
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
Ron.Agresta@sas.com
http://www.sas.com