MDM and Data Quality for the Data Warehouse

*Enabling Timely, Confident Decisions and Accurate Reports with Reliable Reference Data*
Table of Contents

Executive Summary ................................................................. 2

Your Business (Intelligence) Is Only as Good as Your Data. ............... 3

Why You Still Need to Master Your Data Before You Master
Your Business (Intelligence) ...................................................... 4

The Technical Challenge: Managing Data in a Data Warehouse ........ 5

MDM Reduces Data Integration Efforts ....................................... 6

Role of Data Quality in BI .......................................................... 7

Summary ...................................................................................... 8
Executive Summary

Many companies have invested in sophisticated business intelligence (BI) systems for optimizing business processes or complying with regulatory reporting requirements yet fail to achieve the results they expected. Why? These systems don’t contain reliable business-critical data about customers, products, channel partners, suppliers, and employees—otherwise known as master or reference data. As a result, business leaders make poor decisions. Compliance officers struggle to comply with regulatory reporting requirements. Business analysts, who prepare reports for their managers, spend too much time manually searching for and reconciling relevant information. And precious IT resources strain to fix data quality issues. All these problems can be attributed to feeding unreliable data into the data warehouse that powers BI. This executive brief explains how the combination of master data management (MDM) and data quality (DQ) significantly enhances the accuracy and reliability of data, enabling timely, confident decisions and accurate reporting.
Your Business (Intelligence) Is Only as Good as Your Data

When accurate data is not made available through BI and reporting systems, the business as a whole suffers. Here are some of the consequences of using unreliable data to feed BI systems that create reports:

- **Business Leaders**: Inaccurate management reports lead to poor decisions. Misunderstandings about which customers are the top 100 based on revenue and which products are the most profitable can result in revenue and profit loss.

- **Compliance Officers**: Compliance initiatives such as Sarbanes-Oxley and Basel II require companies to provide transparency and auditability of their financial and compliance reporting. Inaccurate reports expose the company to serious risks such as steep financial penalties and negative market perceptions, not to mention potential jail time for corporate officers.

- **Business Analysts**: Productivity of business analysts is hampered when they spend excessive time manually searching for and reconciling information across multiple BI systems to update and fix reports. This inefficient effort directly impacts costs and profitability.

These business problems result from not having a single version of the truth for business-critical data about customers, products, channel partners, and suppliers. Also called reference or master data, this data is collected, housed, and managed across disparate systems handling every business process. As a result, overlapping and conflicting reference data needs to be correctly resolved to a single version of the truth to provide valuable and actionable insight. Many organizations have tens, and sometimes hundreds, of different applications maintaining the same core reference objects in tens or hundreds of databases, with overlapping attributes. This complexity makes it very difficult for companies to answer questions such as “Who are our 100 top customers by revenue?” or “How many Part Xs have we sold in the last two quarters?”

The purpose of BI is to report on existing data from multiple systems without prejudice. It may do some aggregation for dimensional analysis, but it is not designed or equipped to create a single version of the truth. Inconsistency in customer or product data from application silos can negatively impact the reliability of analytics running on data warehouses.
Why You Still Need to Master Your Data Before You Master Your Business (Intelligence)

Master data management is necessary to recognize disparate data, resolve it into a single version of the truth, and relate it to other data to derive some meaning. In fact, the initial phases of many MDM projects have focused on creating more reliable data for more accurate BI reporting.

MDM is the controlled process by which the master data is created and maintained as the system of record for the enterprise. MDM is implemented to ensure that the master data is validated as correct, consistent, and complete and can be centrally and dynamically managed in hierarchies. Then this reliable and related master data is fed from the MDM system to the data warehouse, where analysts and business leaders can use it for BI and reporting.

The following example, as Figure 1 illustrates, will highlight the value of MDM for BI. The business leaders at a global imaging system manufacturing company need to know its top 100 customers to resolve a customer service issue that arose after a product had been discontinued. To respond to this request, the business analysts prepared a report of its top 100 customers by revenue using the BI system. Unfortunately, they did not have an MDM system implemented, and the data warehouse that powers the BI system had two entries for one of the top customers. General Electric bought multiple imaging systems in two separate transactions: one transaction valued at $400 million came from the U.S. accounting system under the company name The General Electric Company; the second transaction valued at $300 million came from its European accounting system and under the company name GE. The data warehouse did not know that these two are the same company and reported them as two separate transactions and company entities. As a result, The General Electric Company does not show up as one of the top 100 customers. The information in this report is unreliable and any decisions made on the basis of this report would be incorrect.

This manufacturer was able to resolve this potentially costly issue by implementing an MDM system that correctly identified that The General Electric Company and GE are the same company. It fed this information to the data warehouse, which then correctly aggregated the total from the two transactions as $700 million. Now with MDM in place, The General Electric Company correctly showed up as one of the top 100 customers by revenue in the report. The information in this report is reliable and any business decisions made using this report will be accurate.
Business Reasons Why MDM Is Required for Data Reliability

The Technical Challenge: Managing Data in a Data Warehouse

To make timely, confident decisions, business leaders need access to reliable data, which can be achieved by integrating MDM with a data warehouse. The bulk of the effort involved in building a data warehouse involves reference data management, which causes a lot of work for data warehouse developers; they must code and maintain complex logic for integrating frequently redundant, and often incorrect, reference data from all possible data sources.

Maintaining this custom-coded logic is time consuming and expensive. It adds significant overhead to the data integration process. Duplicate data needs to be consolidated. Conflicts must be evaluated to determine the most reliable representation of data. This consolidation not only ensures reliability of reference data but it also ensures reliability of the keys and dimensions used to aggregate transactional data accurately. This interdependency, which is the cornerstone of data accuracy in the data warehouse, eliminates misinformation and ensures that valuable insights can be gleaned by BI systems.
MDM Reduces Data Integration Efforts

MDM offers a code-free, low-maintenance solution for integrating and managing reference data. When the MDM system becomes the definitive source of dimensional data for the data warehouse, it enables the data warehouse to focus on the data management goals of volume management and data delivery.

Briefly, master data management provides the following capabilities:

- Match-and-merge logic for identifying and consolidating duplicate records from one or more source systems
- Extensive cell-level lineage and history, which supply a detailed audit trail for data content
- A central repository for all relationship data across all sources and applications

These capabilities significantly reduce the overall development and maintenance effort associated with data warehouses as described below:

- **MDM Conformed Dimensions**

  The first step in designing a data warehouse—whether it’s an enterprise data warehouse (EDW), a distributed data warehouse (DDW), or a data mart bus architecture—is to conform the data warehouse’s dimensions. Ralph Kimball, founder of the Kimball Group, is an innovator and educator in data warehousing. Documented in several articles and Wikipedia, he defines a dimension as conformed “when two copies of the dimensions are either exactly the same (including the values of the keys and all the attributes), or else one dimension is a perfect subset of the other.”

  The MDM system becomes the single conformed, integrated, cleansed, and standardized data source for the data warehouse’s conformed dimensions. This system removes much of the transformation effort involved in populating conformed data structures, enabling data warehouse resources to focus on answering business questions rather than integrating data.

- **MDM Slowly Changing Dimensions**

  Slowly changing dimensions are the most effective and most frequently used method for maintaining a history of changes to dimensions. Data warehouse developers issue a new dimension record for each dimension record that undergoes a change in one of its data segmentation attributes. The drawback of this approach is that the dimension tables grow continuously, which stalls load and query performance. The growth problem is compounded because the slowly changing dimension approach is used for all sorts of changes within a dimension record, including changes in attributes such as customer name, marital status, or branch manager name.

  The MDM system resolves this performance problem by providing a history-tracking option. It can record all changes on a reference data entity, liberating the data warehouse to concentrate on tracking the data segmentation changes as slowly changing dimensions. When data warehouse dimensions grow more slowly, they minimize the impact on query and load performance. When users want to query the history on a non-slowly changing dimension for a record, they can drill through to the history tables in the MDM system.

- **MDM and Data Lineage**

  Data warehouse metadata and lineage solutions are generally limited to supplying structural and process lineage. For example, a standard data warehouse implementation should be able to answer such questions as “Which fields in which systems are sources for the Customer Name field in the data warehouse?”, “Which load script updates the Customer Name field in the data warehouse?”, and even “What business rules or transformations affect the Customer Name field in the data warehouse?”
Where data warehouses fall short is in providing adequate “content” lineage. For example, most data warehouses could not answer the question “Why is the Customer Name value for this record ‘Fredd Jones’?” Answering this sort of question usually entails querying across at least the staging tables for all of the possible source systems to find out which source record furnished that value.

MDM removes the need for such time-consuming, manual data tracking because it keeps detailed data lineage for every field on every record (i.e., data lineage is tracked at the cell level). For every cell of data in a base object, the MDM system can identify which source system provided that cell value, and specifically which record in the source system provided the value, along with when the value was last updated. The MDM system also tracks the history of all record merges.

**Role of Data Quality in BI**

Integrating data quality rules and activities (profiling, cleansing, and monitoring) with MDM processes is critical to enhancing the accuracy and value of data assets. Before you start any MDM project, you need to understand the content, quality, and structure of your source data. Data profiling at the source enables data stewards and data warehouse administrators to quickly discover and analyze all data anomalies across all data sources before data goes into the MDM system. This process greatly increases the speed to value from the MDM implementation.

Because data cleansing enhances the accuracy of data, provides completeness, and promotes trustworthiness of the data at the source, it improves the consistency of the data within the MDM system. Once source data enters the MDM system, it undergoes data quality processing, including validation, correction, and standardization. The MDM system stores the complete history of the data before and after it is cleansed, eliminating the need for developers to track the data lineage within the data warehouse.

Finally, data quality metrics empower data warehouse administrators to better monitor the quality of reference data and ensure the continuous use of high-quality data over time.

So, from a technical standpoint, implementing MDM, along with DQ, as the definitive source of master data for a data warehouse simplifies data integration from the extract, transform, and load (ETL) process. This approach significantly reduces the overall development and maintenance effort associated with data warehouses. By establishing data quality metrics and defining data quality targets, data warehouse administrators and data stewards are empowered to better monitor the quality of reference data and ensure the continuous use of high-quality data across the enterprise over time. MDM simplifies the handling of updates to the data warehouse dimensions because all logic for determining what has changed is encapsulated within the MDM system.

Furthermore, the MDM system can remove much of the history-tracking burden from the data warehouse, leaving the data warehouse to manage only those changes that it needs to track for aggregation purposes. This system results in smaller data warehouse dimensions and significant improvements in load and query performance. Ultimately, leveraging MDM and DQ reduces the data integration effort and increases the quality of insights derived from BI and reports, ensuring that the expected value and return on investment (ROI) from the data warehouse initiative that powers BI are achieved.
Summary

When BI systems aren’t using reliable data, business leaders lack confidence and make poor decisions, compliance officers risk filing inaccurate regulatory reports, and business analysts spend excessive time searching for and reconciling data to update and fix reports. MDM and data quality significantly enhance the accuracy and reliability of data within the data warehouse, enabling timely, confident decisions and accurate reporting.

Technically, implementing MDM along with data quality as the definitive source of dimensional data for a data warehouse simplifies data integration complexity from the ETL process. This implementation significantly reduces the overall development and maintenance effort associated with data warehouses.

With good data quality, business insights gained through BI become actionable faster. With increased confidence in data, executives, managers, and employees can recognize and act on new patterns and trends, such as opportunities for growth and potential issues that need attention. And end-to-end data quality management increases auditability and visibility of BI reporting, which is especially valuable for compliance and risk management purposes.
About Informatica MDM

Informatica® MDM™ empowers companies to improve operations with business-user access to consolidated and reliable business-critical data about customers, products, channel partners, suppliers, and employees, as well as the relationships between data. Flexible and proven, Informatica’s multidomain master data management (MDM) system provides comprehensive support for all MDM project requirements—data integration, data profiling, data quality, and master data management—on the same platform. Best-in-class companies choose Informatica MDM because they can start small by addressing any pressing MDM-related business problem with rapid implementation and then expand to address additional business problems across the enterprise, realizing fast time to value, lower TCO, and superior ROI. Visit www.informatica.com/mdm.

Learn More

Learn more about the Informatica Platform. Visit us at www.informatica.com or call +1 650-385-5000 (1-800-653-3871 in the United States).

About Informatica

Informatica Corporation (NASDAQ: INFA) is the world’s number one independent provider of data integration software. Organizations around the world gain a competitive advantage in today’s global information economy with timely, relevant and trustworthy data for their top business imperatives. More than 4,100 enterprises worldwide rely on Informatica to access, integrate and trust their information assets held in the traditional enterprise, off premise and in the cloud.