INTELLIGENCE · BUSINESS INTELLIGENCE

Master Data Management and the Single Source of Truth

When different systems hold different versions of the same customer, product or supplier, business intelligence produces contradictory answers. This article explains master data management, the methods used to create a single trusted record, the organisational challenges involved and what the "single source of truth" can and cannot achieve.

ModQu research article · 2,002 words · 21 references (APA 7th edition)

Ask a large company how many customers it has, and the answer may depend on whom you ask. The sales system counts every account opened by a salesperson; the billing system counts every entity that receives an invoice; the online store counts every registered login; the support desk counts every caller. The same business may appear in each system under slightly different names, addresses and identifiers. A product may carry one code in manufacturing, another in the warehouse and a third in the catalogue. When business intelligence draws on these systems, it inherits their inconsistencies, and reports contradict one another.

Master data management (MDM) addresses this problem. It is the set of processes, governance and technology that creates and maintains a single, consistent and trusted version of an organisation's core business entities. This article explains what master data are, why they fragment, how MDM works and why it is as much an organisational challenge as a technical one.

What master data are

Organisations deal with several kinds of data. **Transactional data** record events: an order, a payment, a shipment. **Master data** describe the core entities that those events involve: customers, products, suppliers, employees, locations and accounts. Loshin (2009) describes master data as the business objects that are shared across an organisation's applications and processes and that give transactions their meaning. An order is meaningful only because it links a particular customer to particular products at a particular place.

Master data change relatively slowly compared with transactions, but they are used everywhere. An error in a single customer record, such as a wrong address or a duplicate account, can affect every order, invoice and report associated with that customer. DAMA International (2017) treats master and reference data management as one of the core functions of data management for this reason.

Defining a master entity is less straightforward than it seems. Kent (1978) observed that even basic questions, such as what counts as one customer, one product or one event, involve choices. Is a company with three subsidiaries one customer or four? Is a product sold in two pack sizes one product or two? Different departments answer these questions differently, for good reasons, and those differences are a root cause of fragmented master data.

Why master data fragment

Fragmentation usually results from history rather than poor decisions. Ross et al. (2006) describe how organisations accumulate systems over time, each bought or built for a particular function, with its own data. Mergers and acquisitions add whole new sets of systems. Each system maintains its own copy of customers, products and suppliers, and the copies drift apart as each is updated independently.

The consequences appear throughout the business:

- **Inconsistent reporting.** Sales by customer differ between finance and sales reports because the two systems group accounts differently.
- **Operational errors.** Shipments go to outdated addresses; customers receive duplicate mailings; credit limits are applied to one account but not another belonging to the same customer.
- **Missed insight.** The organisation cannot see the full relationship with a customer who buys through several channels, because the customer appears as several unrelated records.
- **Compliance risk.** Regulations that require organisations to know their customers, or to respond to individuals' requests about their personal data, are hard to meet when records are scattered.

Redman (1998) documented how poor data quality of this kind imposes costs at operational, tactical and strategic levels, from wasted effort to poor decisions and lost trust.

How master data management works

Master data management combines several activities.

Identifying and matching records

The first technical task is to find records in different systems that refer to the same real-world entity. This is the problem of **record linkage** or **entity resolution**. Fellegi and Sunter (1969) provided its statistical foundation: by comparing fields such as name, address and date of birth, and weighting agreements and disagreements by how informative they are, one can estimate the probability that two records match. Pairs above one threshold are treated as matches, pairs below another as non-matches, and those in between are sent for human review.

Elmagarmid et al. (2007) review the many techniques developed to detect duplicate records, from simple string comparison to machine learning. Christen (2012) describes the full process of data matching, including cleaning and standardising data before comparison, reducing the number of comparisons needed in large datasets, and evaluating the accuracy of matches. Matching is never perfect: false matches merge different entities, and missed matches leave duplicates. The balance between them should reflect the costs of each kind of error.

Creating a golden record

Once matching records have been found, MDM creates a single **golden record** combining the best available information from each source. Rules determine which source is trusted for which attribute: perhaps the billing system for addresses, the sales system for contact names and an external registry for company identifiers. These rules, called survivorship rules, must be agreed by the business, not invented by technical staff.

Distributing and synchronising

The golden record is useful only if systems use it. Dreibelbis et al. (2008) describe several architectural styles. In a **registry** style, the hub stores only links between records in source systems, leaving the data where they are. In a **consolidation** style, the hub collects and cleans master data for reporting but does not send corrections back. In a **coexistence** or **centralised** style, the hub becomes the authoritative source, and other systems receive updates from it, as in Figure 1. More centralised styles deliver more consistency but require more change to existing systems and processes.

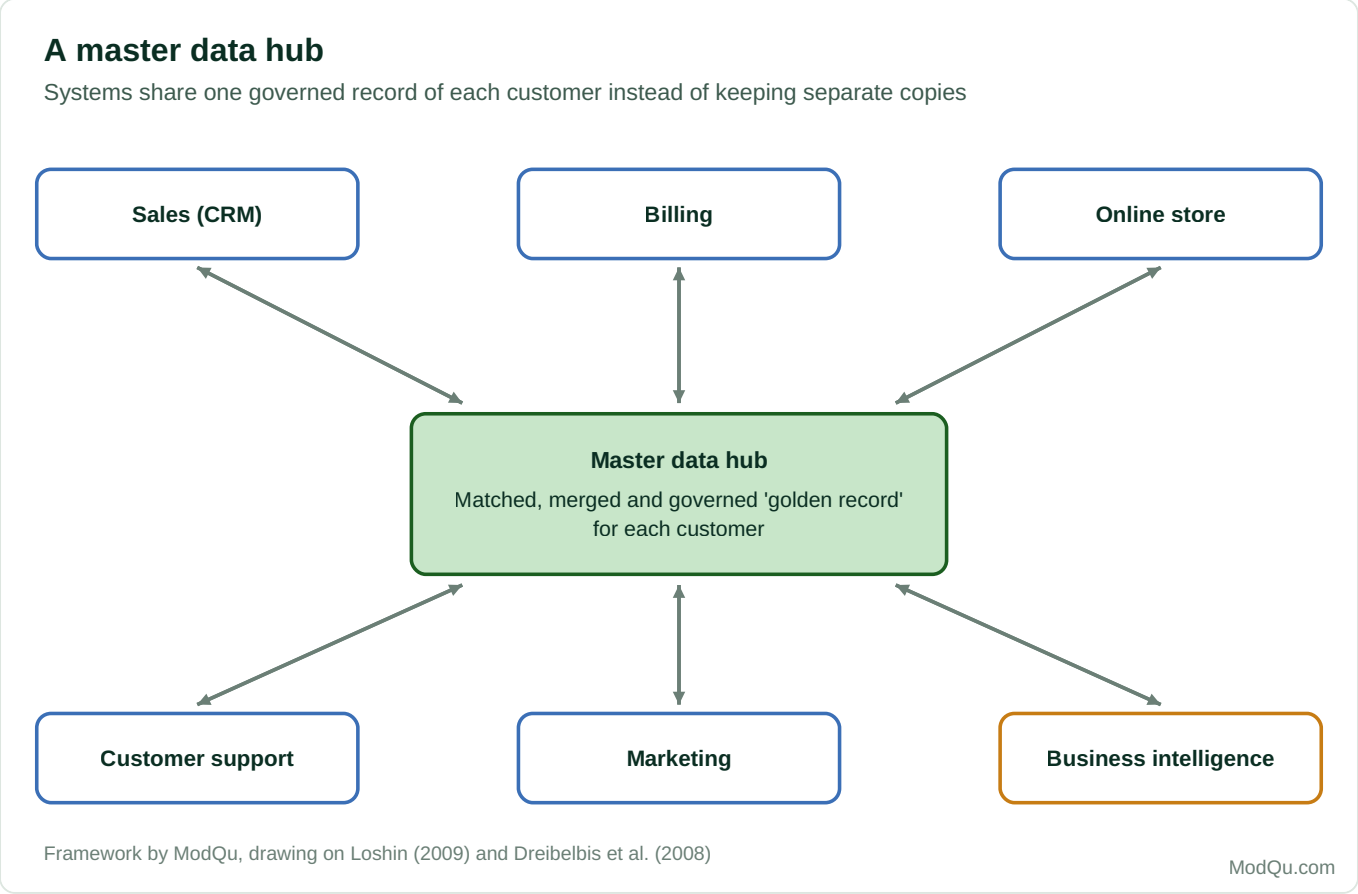


Figure 1. A master data hub. Instead of each system keeping its own copy, systems read and update one governed record. Framework by ModQu, drawing on Loshin (2009) and Dreibelbis et al. (2008).

Connecting to business intelligence

Master data are the backbone of the dimensions in a data warehouse. Kimball and Ross (2013) emphasise that **conformed dimensions**, consistent definitions of customer, product, date and other dimensions shared across all parts of the warehouse, are what allow figures from different business processes to be compared and combined. Inmon (2005) similarly stresses integration as a defining feature of the warehouse. Without consistent master data, a warehouse merely collects inconsistencies in one place.

Reference data

A close relative of master data is **reference data**: the lists of permitted values that classify other data. Country and currency codes, product categories, industry classifications, sales regions and reasons for returning goods are all reference data. They change rarely, but when they differ between systems they cause the same problems as fragmented master data. If one system groups sales into eight regions and another into ten, regional reports cannot be reconciled; if one uses an outdated industry classification, analyses of customers by industry are distorted.

DAMA International (2017) treats reference data management alongside master data management because the disciplines are similar: agree an authoritative list, assign an owner, publish it to every system that needs it and manage changes deliberately. Where external standards exist, such as international codes for countries and currencies, using them avoids unnecessary differences and makes it easier to combine internal data with outside sources. Changes to reference data should be versioned, so that historical reports can still be interpreted using the categories that applied at the time.

The organisational challenge

Experience shows that MDM projects fail more often for organisational than technical reasons. Smith and McKeen (2008), drawing on discussions with practitioners, describe MDM as requiring sustained business commitment, clear ownership and agreement across departments, and warn that it is often sold as a technology solution when it is mainly a matter of governance and process.

Silvola et al. (2011), studying companies managing product and customer master data, found that the main challenges were unclear ownership, inconsistent processes for creating and maintaining records, and a lack of shared understanding of the data's importance. Technology mattered, but only after these conditions were addressed. Otto (2012), in a case study of master data architecture at Bosch, describes how the company had to decide which entities to manage centrally, which to leave to business units, and how to align data architecture with business structure.

Governance and ownership

Master data require clear **decision rights**. Khatri and Brown (2010) frame data governance as deciding who makes which decisions about data. For master data, key decisions include the official definition of each entity, the rules for creating new records, the trusted source for each attribute and the quality standards that must be met. Otto (2011) describes how organisations assign these responsibilities to data owners in the business, data stewards who maintain quality day to day and central teams who provide tools and standards.

Weber et al. (2009) argue that governance must fit the organisation: a highly centralised company may manage master data centrally, while a diversified group may allow business units more autonomy, coordinating only the entities they share. Weill and Ross (2004), writing about information technology governance more broadly, found that organisations that design decision rights deliberately get more value from their technology investments.

Fixing processes, not just data

Cleaning master data once is not enough. If the processes that create records remain unchanged, duplicates and errors return. English (1999) argued that quality must be built into the processes that create data: checking for existing records before creating new ones, validating addresses at entry, and requiring key attributes. Batini et al. (2009) describe improvement methods that combine data cleansing with process redesign.

The limits of a single source of truth

The phrase "single source of truth" is often used to justify MDM. It is a useful aspiration, but it can mislead.

First, some differences between systems are legitimate. Finance may need to treat a corporate group as one customer for credit risk while sales treats each subsidiary separately. A good MDM design records the relationships between entities so that each view can be produced consistently, rather than forcing a single definition on everyone.

Second, the golden record is a representation of reality, not reality itself. Kent (1978) reminds us that data embody choices about how to describe the world. Those choices should be documented and revisited as the business changes.

Third, quality is relative to use. Wang and Strong (1996) showed that users judge data by relevance, timeliness and interpretability as well as accuracy. A master record that is accurate but out of date, or defined in a way that does not fit a department's needs, will not be trusted, and people will start keeping their own copies again.

A more accurate aim is a **single, governed set of definitions and records**, with clear rules about how different views are derived from it.

An example

A distributor formed by the merger of two companies had customer records in two billing systems, a sales system and an online ordering platform. An initial analysis found that about a fifth of its largest customers appeared in more than one system under different names and numbers. Reports of revenue by customer were unreliable, and some customers received two sets of statements.

The company established a governance group with owners from sales, finance and customer service. The group agreed a definition of a customer, a hierarchy linking subsidiaries to parent companies and rules for which system was trusted for each attribute. A matching process linked records across the four systems, with uncertain matches reviewed by customer service staff. New customer records could be created only after a search of the master hub. Within a year, duplicate statements stopped, revenue by customer reconciled between sales and finance, and the company could see, for the first time, how much each corporate group bought across all its channels.

Conclusion

Master data management creates consistent, trusted records of the core entities a business depends on. It relies on statistical matching, agreed rules for combining data and architecture that distributes the result, but above all on governance: clear ownership, shared definitions and processes that keep data right at the source. Pursued as a technology project alone, it rarely succeeds. Pursued as a business discipline, it provides the foundation on which reliable business intelligence is built.

References

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), Article 16. https://doi.org/10.1145/1541880.1541883

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer. https://doi.org/10.1007/978-3-642-31164-2

DAMA International. (2017). *DAMA-DMBOK: Data management body of knowledge* (2nd ed.). Technics Publications.

Dreibelbis, A., Hechler, E., Milman, I., Oberhofer, M., van Run, P., & Wolfson, D. (2008). *Enterprise master data management: An SOA approach to managing core information*. IBM Press.

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 1–16. https://doi.org/10.1109/TKDE.2007.250581

English, L. P. (1999). *Improving data warehouse and business information quality: Methods for reducing costs and increasing profits*. Wiley.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Inmon, W. H. (2005). *Building the data warehouse* (4th ed.). Wiley.

Kent, W. (1978). *Data and reality: Basic assumptions in data processing reconsidered*. North-Holland.

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, *53*(1), 148–152. https://doi.org/10.1145/1629175.1629210

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). Wiley.

Loshin, D. (2009). *Master data management*. Morgan Kaufmann.

Otto, B. (2011). Organizing data governance: Findings from the telecommunications industry and consequences for large service providers. *Communications of the Association for Information Systems, 29*, 45–66. https://doi.org/10.17705/1CAIS.02903

Otto, B. (2012). How to design the master data architecture: Findings from a case study at Bosch. *International Journal of Information Management, 32*(4), 337–346. https://doi.org/10.1016/j.ijinfomgt.2011.11.018

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM, 41*(2), 79–82. https://doi.org/10.1145/269012.269025

Ross, J. W., Weill, P., & Robertson, D. C. (2006). *Enterprise architecture as strategy: Creating a foundation for business execution*. Harvard Business School Press.

Silvola, R., Jaaskelainen, O., Kropsu-Vehkapera, H., & Haapasalo, H. (2011). Managing one master data – challenges and preconditions. *Industrial Management & Data Systems, 111*(1), 146–162. https://doi.org/10.1108/02635571111099776

Smith, H. A., & McKeen, J. D. (2008). Developments in practice XXX: Master data management: Salvation or snake oil? *Communications of the Association for Information Systems, 23*, 63–72. https://doi.org/10.17705/1CAIS.02304

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Weber, K., Otto, B., & Österle, H. (2009). One size does not fit all—A contingency approach to data governance. *Journal of Data and Information Quality, 1*(1), Article 4. https://doi.org/10.1145/1515693.1515696

Weill, P., & Ross, J. W. (2004). *IT governance: How top performers manage IT decision rights for superior results*. Harvard Business School Press.