intelliTide

An authoritative definition of Master Data and it's characteristics

Ramesh Prabhala
Dr. Satheesh Ramachandran
Kartik Nanda

WHITEPAPER

The Question

Since my early days into Master Data Management, the question of how to define Master Data preceded my curiosity of how you manage it. I frequently heard this definition "data that does not change frequently over time or data that does not change at all". So, it is true that there is a static characteristic to Master data as a function of time but that didn't sound right as a definition. But let's hear what the pros have to say.

Gartner:
"Master data is the consistent and uniform set of identifiers and extended attributes that describes the core entities of the enterprise…"

Techopedia:
"Master data refers to data units that are non-transactional, top level and relational business entities or elements that are joinable in observable ways"

Webopedia:
"Master data is any information that is considered to play a key role in the core operation of a business. Master data may include data about clients and customers, employees, inventory, suppliers, analytics and more"

And finally, **Wikipedia**:
"Master data represents the business objects that contain the most valuable, agreed upon information shared across an organization. It gives context to business activities and transactions…". It goes on to say *"Master data is usually non-transactional in nature, but in some cases gray areas exist where transactional processes and operations may be considered master data by an organization".*

And therein lies the problem. To disambiguate Master Data from Transaction Data, some experts seem to imply Master Data is not just data that enables core business operations but one that is either wholly or somewhat static in state.

So, What is Master Data?

Though none of the definitions are incorrect, they do seem inadequate to offer a precise description of Master data though Gartner comes close. Let's define the characteristics of Master data to help us arrive at its definition.

The Data in "Master Data" Is Attributes

Data is too broad a term so we need to first clarify what "Data" in Master Data implies. The data in question is really a set of fields belonging to an entity in business sense like a product, a customer, a vendor, a location etc. These fields - called Attributes in Master Data Management parlance - contain values that help identify an entity or values that describe a certain property of an entity.

Attributes Consist of Identifier(s)

A person has a name that identifies him or her. He and she may also have a SSN. A product will have a UPC code or some identifying attribute like a Product ID or SKU ID. The purpose of these identifying attribute or attributes is to uniquely identify an entity. One stated goal of Master Data Management is after all to create a single source of truth. To attain this goal, we must select one or more attributes that individually or grouped together provide a uniqueness to the entity. However, the question is why can't

a single attribute suffice? It very well may within a certain data **space** and **density**. But as the space expands and density increases the probability that a scalar identifier will suffice decreases. Let's study these concepts of space and density in depth.

Impact of Entity Space on Identifiers

Let's start with a person entity - me. Consider my name. In most parts of the world including here in America - Ramesh Prabhala - sounds unique. The probability that it is unique within a company is very high. Within an American city is still very high. Within the county of America its probability of singularity relatively diminishes but it is still high. Now expand the scope to the entire world and there is a good chance there is someone else by the same name. Thus, the probability of name (or identifier) collision increases as the **Entity Space** increases. To uniquely identify me in the universe my name alone may not be enough. We may need to combine my name, with my date of birth and place of birth at a minimum to greatly increase the probability to uniquely identify me.

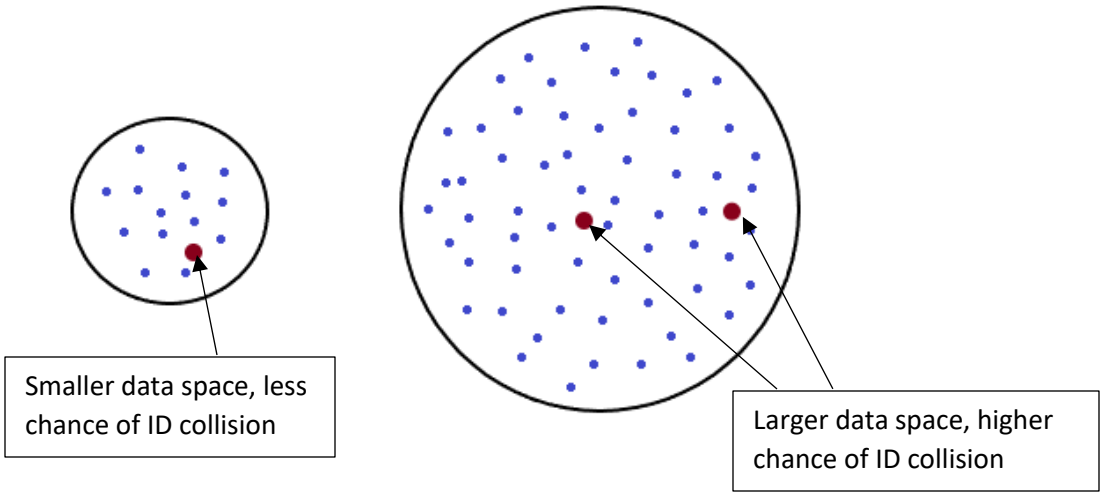
***Note** that we could have considered my SSN as my unique identifier but it will be limited to US. It will lose its relevance as the scope expands beyond the US. Even within the US, many enterprises are hesitant to store SSN due to data protection concerns so most data sets will not have this information.*

When the entity in question is not a person but a product whose scope is within an enterprise, isn't a scalar identifier feasible? In theory it is if an enterprise wide unique identifier assignment strategy was established and enforced from the beginning. That's not always the case. In real life, products are assigned single valued unique identifier that's unique only to the system where they are stored. Another system within the same enterprise may generate the same identifier and assign it to a different product. Thus, to identify a product uniquely we may need to combine multiple identifying attributes of the product to create a composite key until we are assured the combination has attained uniqueness. The circumstances are different but this is analogous to the person entity example I mentioned earlier.

To summarize the probability that a single identifier will suffice as a unique identifier is inversely proportional to the size of the entity space. Here is the equation (which warrants its own article):

$$P(U\_S\_ID) \propto 1/e^{Size(ES)}$$

Where, U_S_ID = Uniqueness of a scalar identifier
And ES = Entity space



Impact of Entity Density on Identifiers

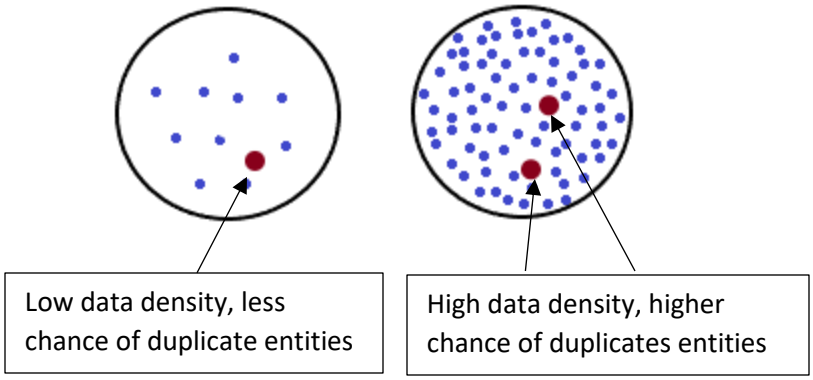
In the context of Master Data, we are rarely talking about the data space being as wide as the universe. After all, how many times would you ever have the need to identify a Ramesh Prabhala or a John Smith in the entire universe? Our scope is confined usually to an enterprise. Why is it hard to identify a product (or any other entity) within an Enterprise? Businesses usually have multiple systems – ERP, SAP, PLM and other systems where same product could be duplicated or different identifiers assigned to the same product residing in different systems. This causes a problem when data is warehoused in a single system. The exercise of creating Master Data involves bringing products (or other entities) into a single store and creating unique single source of truth product record which could result in having to deal with this identity problem.

This is where the problem of data density arises. That is even when the data space is small, as more data is injected into the space the possibility of uniqueness reduces and the probability of identifier collision increases. Once again, we could end up having to consider multiple key attributes as keys to de-duplicate products since different products could be assigned the same primary identifier or same product has multiple copies with different primary identifiers. Different strategies are employed to create a unique record (which is outside the scope of this paper) but even after a unique record is created within this space, a combination of identifiers would still be needed to identify future duplicate records being imported into this space. This space where data is warehoused and single source of truth entities are created in the Enterprise is called Master Data Management.

To summarize the probability that multiple attributes will be required to uniquely identify an entity is directly proportional to the density of entity space. Here is the equation:

$$P(IDs) \alpha \text{ Density } (ES)$$

Where IDs = Multiple identifiers
And ES = Entity space



Before we move to the next segment, we need to recap a couple of lessons we learnt so far. One, that attributes play a key role as Master Data elements since they have the ability to uniquely identity entities. Two, usually in the enterprise more than one attribute is required to uniquely identify entities in an entity space.

A Math Viewpoint of Entity Identification

One more thing before we move on. For those from a math background, the problem of attempting to establish uniqueness in the entity space has been explored before. Conceptualizing in math terms, the attributes are dimensionality of the space that we are locating the entities in. To establish uniqueness, as the number of entities increases, we have to locate these entities in a higher dimensional attribute space. Certain attributes (or dimensions) exhibit low entropy and some very high (like SSN). Rarely, a single higher entropy dimension is enough to establish uniqueness. However often we need to combine dimensions to attain uniqueness especially as the entity cardinality increases. The higher the entropy the lesser the number of dimensions required to uniquely identify the entity in a given cardinality. The cardinality of the entity set is the number of entities in a given set or space. Referring to our discussion above, the cardinality will increase either due to the expansion of the entity space or increasing the density of data in the same space or both.

Attributes Consist of Describers

As previously stated, Master Data consists of Describers in addition to Identifiers. Let's dig deeper into what describers are. Let's start with an example. Here is a person with a fairly common name:

Name: John Smith
SSN: 123-45-6789

Imagine the SSN is real. The SSN uniquely identifies this person but offers little description of the person. You can infer that this person probably lives in the US (he has SSN) and is most likely a male (from the name). However, the fact that we were able to uniquely identify this person via their SSN is of little value from an operational standpoint. Imagine we knew a few more facts about this person:

DOB: 01/12/1970
Address: Austin, TX, USA
Gender: Male
Race: Caucasian

Now a profile emerges. We now know the age and some demographic information. How will it help? If you were a doctor and John Smith is your patient, you can place John in a certain patient risk profile. Mind you that this data on John is still too limited to create a useful patient profile. But if you knew a bit more like his family history and current conditions?

Chronic conditions: Diabetes
Father's cause of death: Prostate Cancer
Occupation: Accountant

Now you know a lot more about John to determine what risk profile he fits into so you can prescribe suitable tests and recommend precautions he needs to take to manage those risks. All the above are descriptive Master Data fields belonging to this person. Not just a physician but credit risk agencies, marketing companies and others will benefit from knowing more about John.

Let's switch from person to product. What does this information tell us?

Name: Samsung LED UHD NU6900 Series TV
SKU: 6268403

It tells us that this is a Samsung LED TV and is an ultra-high definition. But little else. It does uniquely identify the product but not universally. We don't know who is selling this so the SKU # is of no use for identification purposes. Let me add a few more descriptors:

Seller: Best Buy
Screen Size: 50"
Resolution: 2160p (4K)
Smart TV: Yes
Sound: Dolby Digital
Model #: UN50NU6900FXZA

Now we have some data pertaining to this product that makes it useful to achieve a key goal – to sell it. Note that the model # also universally identifies the product. Since we know the seller is Best Buy the aforementioned SKU # helps identify the product within Best Buy's inventory should a customer need to call customer service or sales to purchase the product.

What we have established is that - to sell a product or to diagnose and treat a patient or to market to a customer - we need to be able to both uniquely identify and describe an entity. Master Data combines identifiers and describers to achieve these goals. In its canonical form Master Data is simply reduced to:

Master Data = f(a1n) + f(a2n)
Where a1n = one or more identifier attributes
And a2n = one or more describer attributes

Now that we have a grasp of the definition of Master Data, let's explore a couple of its other characteristics.

Other Characteristics of Master Data

The Data in Master Data is Plural

By now this much should be obvious – the Data in Master Data consists of attributes which in turn identify and describe the entity they are associated with. Multiple attributes or fields are needed to accomplish this objective. A single field or piece of information will not suffice to achieve this goal; therefore, Master Data is always plural.

Duration of State is a Symptom, Not the Cause

We started this article with a definition of Master Data that defined it as a function of time – *data that does not change frequently over time is Master Data*. This is incorrect. It is true that attributes or fields of a Product or Person or another business entity that constitute its Master Data do not change frequently. Remember the Master Data fields identify and describe an entity and something that changes very frequently cannot in theory be used to either identify or describe anything. If my name or SSN changes frequently it cannot identify me. If my ethnicity, date of birth or address change every day then none of these attributes can describe me well. However, just because an attribute is static or relatively static does not automatically qualify it as Master Data if it does not contribute to the identification or description of the entity.

Thus, the relatively static state of Master Data is just the symptom of its inherent characteristics that qualified it as Master Data rather than the sole criteria of its qualification.

Conclusion

Considering the growing significance of Master Data Management in the enterprise world today, gaining an understanding of the precise definition as well important characteristics of Master Data is of great importance in order to manage it. In this article, in the course of explaining the concepts of Master Data we hope to have also illustrated why Master Data plays such a significant role in the operations of an enterprise.

Author

Ramesh Prabhala is the founder of IntelliTide - a Data Science Platforms and Services company which uses the power of Data Science, Machine Learning, Cloud and Big Data to improve efficiency, productivity and financial outcomes. Prior to founding IntelliTide, he worked in various technical management, engineering and consulting roles and has extensive experience in enterprise systems and data management. His detailed profile is on LinkedIn https://www.linkedin.com/in/ramesh-prabhala-00525/

Contributors

Dr. Satheesh Ramachandran is the Chief Data Science advisor for IntelliTide. He is an experienced data scientist with a broad background in applied statistics, data mining, text mining, forecasting and operational research, with over two decades of experience in multiple domains. Dr. Ramachandran is an Engineering graduate from Indian Institute of Technology with a Masters and Ph.D. from Texas A&M. Dr. Ramachandran's LinkedIn profile is https://www.linkedin.com/in/satheeshr/

Kartik Nanda is a Data Science and AI advisor to IntelliTide. His expertise is Artificial Intelligence, Signal Processing and Algorithms, spanning two decades and multiple domains including consumer electronics, e-retail, renewable energy and food supply chain. Kartik is a Engineering graduate from Indian Institute of Technology and a Masters in Computer Science from University of Notre Dame. Kartik's LinkedIn profile is https://www.linkedin.com/in/knanda