A DECENTRALIZED MASTER DATA SOLUTION USING DATA VIRTUALIZATION

Author: **Rick F. van der Lans**

INTENDA

The importance of master data cannot be denied. There are a variety of articles, books, and events that emphasize its importance. Yet for many organizations', it is a struggle to organize their master data in such a way that it can be used by many data consumers. Those who have been able to organize it well, use master data primarily within a limited and controlled set of applications.

Note: In this article, reference data is regarded as a form of master data.

It is important that organizations' focus on making master data available to all potential data consumers. Master data is not only useful for integrating data when it is copied to a data warehouse. Its use benefits data scientists, business users developing ad-hoc reports, apps running on mobile devices enabling customers to manage their own bank account, and so on.

Even if an organization does not have a centralized master data system (MDM), it is still data that can be considered as master data. It is likely stored in a variety of systems, such as the following:

- Master data may be stored in a dedicated MDM to manage all the correct product and customer data.
- Some data stored in production systems can be treated as master data; especially if that data is correct, properly managed, and considered correct by all the users.
- Some personal master data may be stored in spreadsheets or other files.
- Some ERP applications may contain master data.
- Master data may also be defined externally.

For example, the International Classification of Primary Care (ICPC) is accepted in several countries as the standard for coding and classifying medical complaints, symptoms and disorders in general practice. This can be considered as master data.

To bring all that master data together, the first solution that always comes to mind is to copy and store it in a centralized MDM. However, this seemingly simple solution does have its drawbacks, such as:

- The original source of the master data may have a high frequency of change, which makes it difficult to keep the copied and centralized master data up to date.
- The sheer size of the master data may be impractical to copy. Imagine an organization has tens of millions of customers; copying it periodically and keeping it up to date with the centralized MDM may be complicated.
- If the master data in the original source is correct, copying it one-to-one to a centralized MDM is somewhat superfluous, and conflicts with design principles such as data minimization.
- Large organizations may operate two or more MDMs; for example, due to a merger or acquisition. The question that arises is: how is that master data integrated? Is it copied from one system to another?

With all this master data distributed across many systems, it can be a challenge to make everything available in an integrated way to all the data consumers. A decentralized solution may be preferred. Such a decentralized solution can be implemented using data virtualization.

In this case, a data virtualization layer is defined on all the internal and external systems that contain master data. Without copying it, the master data is presented to all the data consumers as integrated master data.

This data virtualization-based solution offers the following advantages:

- The master data can be transformed, filtered, and integrated into a form that simplifies consumption within the data virtualization solution.
- The built-in security mechanism can be used to allow master data to be presented only to those data consumers who are allowed access.
- The same master data can be presented in different forms to different data consumers. This is done without the need to create multiple, physical copies of the master data for different consumers.
- If master data is related to personal data such as customer or patient data, it can be anonymized and/or pseudonymized by the data virtualization solution ensuring that certain data consumers cannot view it. This assists to comply with the GDPR and other data privacy regulations.
- When the data virtualization solution is directly linked to master data stored in a production system, the data consumers will always see up-to-date master data and not copied data with a high data latency.
- When master data needs to be migrated from one MDM to another; for example, because one MDM needs to be phased out, the lift and shift process can be hidden for the data consumers.
- If a system contains master data that is not scalable and fast enough to support specific data consumers; data virtualization features such as caching can be used to easily copy that data to a fast and scalable platform that can support that data consumer. Note that cached master data is managed by the data virtualization solution.
- Data Virtualization allows organizations to start small if they still need to develop a master data system. They can start by first selecting existing systems that contain master data. Next, they define a data virtualization layer on top and make the master data available. If necessary, a more sophisticated MDM system can be deployed later that replace the current solution.

However, it will still operate behind the data virtualization solution to hide from data consumers that the implementation has changed. In general, what data virtualization can do for master data is what it has always been able to do for data itself.

In my book, Data Virtualization for Business Intelligence Architectures, I have included the following definition of data virtualization:

“Data virtualization is the technology that offers data consumers a unified, abstracted, and encapsulated view for querying and manipulating data stored in a heterogeneous set of data stores.”

Perhaps this definition needs to be extended a bit:

“Data virtualization is the technology that offers data consumers a unified, abstracted, and encapsulated view for querying and manipulating data and master data stored in a heterogeneous set of data stores.”

If organizations recognize that their master data is stored in a heterogeneous set of internal and external systems; it is well worth considering data virtualization to virtually integrate all that master data. This includes/Not to mention without physically centralizing the copied master data and to make it available to a wide range of data consumers quickly.

It is important that organizations’ focus on making master data available to all potential data consumers.