Topofminds

Top of Minds Report series

Data Warehouse and definitions for integration

Background

Sometimes organizations have only a few source systems in its IT landscape and others have 100+ systems with overlapping information areas from different departments, divisions and countries etc. Because of different reporting and business analytic requirements, this information has to be integrated for a common usage. To do this we often use a Data Warehouse. Since integrating the information is the main reason to move data into a Data Warehouse. The Data Warehouse main purpose it to see to that the information is integrated, made accessible and usable throughout the organization independently what system, process, department, division and country the data was created.

So the Data Warehouses primary aspect is integration of information. The integration has two major parts. Data model as such, to represent the common view of information. This is both a logical and well as a physical representation. The other big part of integration is Semantic integration. Model and Semantic integration are very tightly intertwined and they are the two sides of the Information Integration coin.

In this report we will look at definitions from a Data Warehouse perspective, and break down what we need and how to think when working with them.

Target audience

This paper turns to new as well as experienced people in the Data Warehouse community.

About the author

Patrik Lager is a senior specialist at Top of Minds. Patrik is specialized in data warehousing architecture, information modeling, data modeling and ETL design. He has a long and vast experience in working within the bank & finance area and also telecom. Mr. Lager holds a BSc in Computer Science from Linköping University. He is a member of the Data Vault Standardization Institute.

Comments on this white paper can be sent to patrik.lager@topofminds.se.

About Top of Minds

Top of Minds is a specialized company that offers services in the data warehousing and business intelligence area. We are premier partner in the Nordic countries on the data warehouse development using an agile approach where we use Data Vault to increase the benefits of the projects we participate in. We are a company with great focus on competence, our expertise and our clients' expertise and we are constantly working to spread our knowledge.
Visit www.topofminds.se for further details.

Focus on the Definitions

Data Warehousing is about one thing and one thing only – Information Integration. Everything else is secondary. Information Integration needs one thing and one thing only - Definitions. Everything else is secondary.

Information Integration and definitions

Information integration is about the ability of sharing and using information throughout an organisation independently what country/division/department/process/system created or updated the information. Information integration can only be attained through the usage of Definitions. Without the Definitions you can't succeed with information integration and by that your Data Warehouse will fail.

In this report we will look closer at Definitions and see why and how they are needed to build a Data Warehouse.

Definitions

Data Warehousing is about Information Integration and to work with Information Integration is about interpreting commonalities in our information universe. To do that we need Definitions that describe which commonality we are aiming for. From a Data Warehouse perspective; that is source system independent definitions of the information that is about to be integrated. I will describe the four levels of Concepts that are needed and how they are used from an information integration perspective.

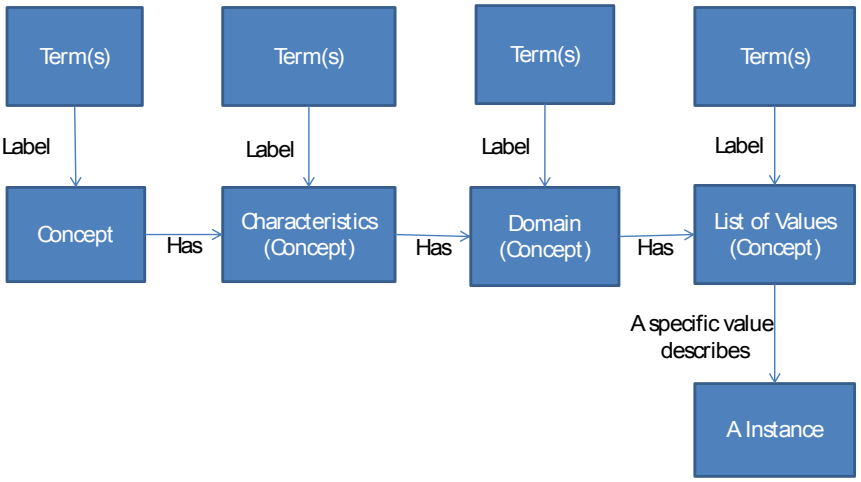


Figure 1: Information Integration Model

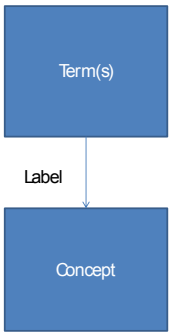
Term

Term can often be replace with the word “Name”. It is nothing more than a label, which consists of a word or a group of words that refers to a definition of a Concept (see below). The term “Car” itself does not give us any insight of what “Car” means. Example, if I tell you that I have a car, you could have a basic understanding what kind of “thing” I have because you “understand” the definition of a “Car”, the understanding comes from the fact that you know and understand the definition of the concept that has the term “car”. Turn it around, any word you don’t understand, not considering the problem of language, homonyms and synonyms, what it means is a Term for a concept you are unfamiliar, don’t have/understand a definition of. When you do not understand a term you can’t really communicate about it either. A Term does not work as a definition; it will only work as a “knowledge bearer” for those who know the definition of the concept that the Term labels.

Concept

We humans understand that the universe has different “kinds” of things in it and we have the knowledge that every individual thing that exists can be seen as an “instance” of a kind of thing.

A “kind of thing” is what’s called a Concept. While the concept itself never exists in the real world, it is the bearer of a definition we use in our daily lives and helps us to sort out the information we are receiving. Since we do not have the time to communicate with definitions we often use a “Term” to communicate about the concept. If someone tells you that they have a car you use your definition of the concept “Car” to understand what that person is speaking about. You do not need to know the exact car. You just need to understand the definition that the Term “Car” labels to have the ability to communicate about that concept with a person.

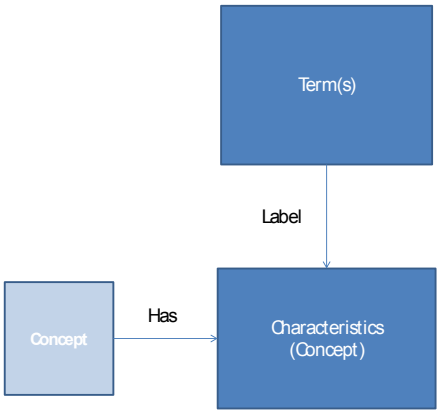


The Concept is the main theme of Information Integration, since it is the idea that Information Integration is about finding out commonality among information, independent of where it was created or used in the organization, to do that we need the definitions of concepts. How would we otherwise know if information from different areas in the organization really belongs to the same Concept?

The definition of a Concept should answer the question: What is that? The better it brings understanding of the question, the better the definition. A Concept with bad a definition can create bad information integration.

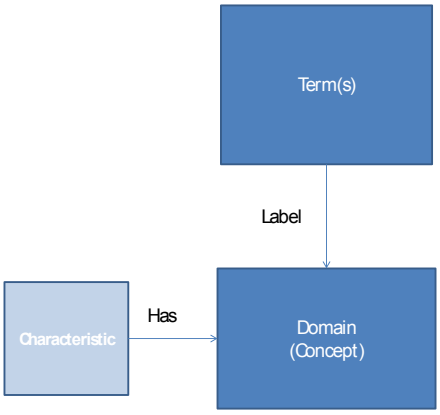
Characteristic

The characteristic is used to hold information to describe the instance of the concept, but it is not the definition of the concept it populates. If we have a concept called "Customer" we might have a Characteristic that is called "Date of Birth", the value in that Characteristic helps us describe the instance (a certain customer) that is part of that Concept, but "Date of Birth" might not be part of the definition of the Concept "Customer". It is Important to understand that each Characteristic is its own Concept, since we need to define the Characteristic. We need to find out the commonality of the Characteristic and therefore it is a Concept itself. So the Characteristic "Name" is the Term which itself needs a definition of its Concept.



If we have the Concept "Bank Account" with its definition, one of the Characteristics might be "Account Balance", which belongs to the Concept "Bank Account". The Characteristic "Account Balance" would need its own Concept definition, because otherwise we would not know what data records that can populate that Characteristic. A simplified example of a definition for the "Account Balance"; *Account Balance is the balance of the bank account at a specific point in time and the value it holds at that specific point in time represent the monetary value of the settled transactions and does not contain any accrued interest.* If we didn't have that definition of the Account Balance we would not know what records should be loaded into that characteristic and maybe even load values that does not have the same meaning, what normally is called "mixing apples with pears. That could be dangerous since we might then use the records in that characteristic and make wrong assumptions.

Domain

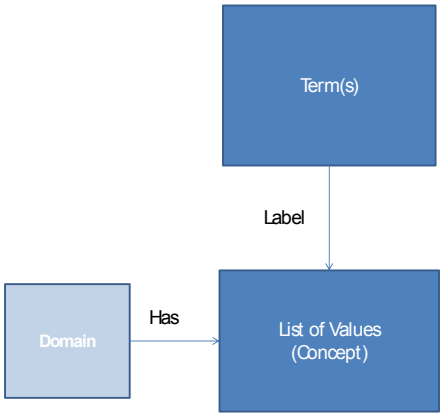


Each characteristic has a Domain; it is within the Domain the characteristic get its values. There are mainly two types of Domains from an information integration perspective; open or closed. Open Domains are those that normally have very little need of integration of values, such as Date. The Characteristic – Date of Birth – takes its values form the Date Domain, from an integration point of view; the "format" of the date is an integration point but not its value. Those Characteristics that has a closed Domain are often much more in need of integration. A Domain is also a Concept and therefore also needs a definition to support information integration. An example could be that we have a Bank Account that has the Characteristic – Account Balance Currency – which sets the currency in which the Account Balance is represented. The Characteristic - Account

Balance Currency uses the Domain - ISO 4217 Currency Code standard. Which itself has a definition. Sometimes the Domain acts as a Characteristic for a Concept and when that happens there is no difference between them.

Domain format

The List of Values in a domain also has to have a defined format to achieve full integration. Example would be that a date has to be represented accordingly to its defined format independently of the source format.

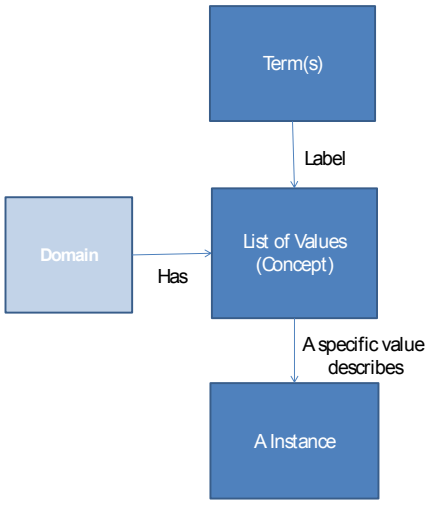


Value

Even the Values in a Characteristic or Domain can represent its own Concept. If we have a Concept called “Organization” we might have a Characteristic that classifies the Organization according to Global Industrial Classification System (GICS), the GICS Characteristic has its own definition and one of the Values the Characteristic uses is the code “10” which stands for the Term “Energy Sector” in the GICS Concept and is defined as follows:

“Energy Sector – the GICS Energy Sector comprises companies whose businesses are dominated by either of the following activities: The construction or provision of oil rigs, drilling equipment and in energy related service and equipment, including seismic data collection. Companies engaged in the exploration, production, marketing, refining and/or transportation of oil and gas products, coal and other consumable fuels” So even Values can have their own definitions and by that are their own Concepts.

Another example is when a Domain Value is a Concept; take the ISO 4217 Currency Code standard, where each currency code has a definition and is its own Concept. From information Integration point it is important to have the ability to integrate Values to support the common information view in the Data Warehouse.



Instance

The idea that a Concept defines a commonality that exists among individual things in the universe is the main theme of Information Integration. The individual things that exist in the universe are the “Instances” of Concepts. An Instance is therefore an individual thing that fulfills the definition of a Concept. The only way to know if an individual thing implements a Concept is to check if it matches the definition of the

Concept. So without definitions of Concepts, there can't be any Information Integration. Through the help of the Values of an Instance that is held in the different Characteristics of a Concept we can analyze and report on Instances. If we have integrated the Concept, Characteristics, Domains and Values according to their definitions, we can use the information throughout the organization, independently of where the information was created or updated and that is the Primary Aspect of a Data Warehouse.

Acceptance of Definition

The usability scope of a definition depends on how large the acceptance is of the definition in your organization. If a definition is created for a division in a company it can't be an Enterprise definition, it's a definition accepted within a division of the company. This might sound unimportant but it is one of the fundamental problems of with the usage of definitions in Information Integration today. It is easy to create an expectation level of the usability of data that does not match the definition scope. If someone tells you they have an Enterprise Data Warehouse, then the definitions used for integration in the Data Warehouse have to be accepted throughout the Enterprise and frankly, they seldom are. This affects the expectation level of how large the organizational coverage is in the Information Integration solution where the definition is used. It is important to be very clear about the width of acceptance and the possible usage of a definition in an organization.

Conclusion

Information Integration in a Data Warehouse is the discipline of understanding what Information Concept each individual data record belong to and the only way to do this is to have definitions of Information Concepts, so without any Definitions there can't be any Information Integration and without Information Integration there can't be any Data Warehouse.

From an Information Integration point of view we have four levels definitions with one sub level of Domain definition

- Concept definition
- Characteristic definition
- Domain definition
    - Value format
- Value definition

What I written here about Definitions are focused on Information Integration in a Data Warehouse environment. If you are interested to know more about the creation and usage of Definitions and the management of those, I urge you to read the book "Definitions in Information Management" by Malcolm D. Chisholm, Ph.D., which will bring you a broader and deeper understanding in the subject of Definitions.

Further reading

The usage of Definitions for information integration are described in ToM Report Series on Data Warehouse -The six levels of integration, that can be downloaded at http://topofminds.se/wp/aktuellt/publicerat/

