Get a grip on data modeling

Get a grip on data modeling

Bert Dingemans

Author: Ir. Ing. Bert Dingemans
Cover design: Rudy Barten

Foreword

A book about data modeling in this digital age, isn't that strange? When I look at my daily work in which data modeling takes a central place then I think not. That is why a reference work of various modeling techniques has been created.

This book consists of two parts. Each part looks at data and modeling from a different perspective. The first part describes the different types of modeling. This is depicted on a simple framework to indicate which aspect of data modeling is supported in the modeling form in question.

In the second part a data model is developed of a fictitious case the Alberto ice cream parlors. The data model is an elaboration of all the discussed data modeling methods. In the model a hybrid model is developed that makes all notation methods together into one joint and integrated model.

Without the help of a number of people, this book would not have reached its current form. I would like to mention a number of people by name. My daughter Karin revised this document and made improvements. The members of the Metadata working group of DaMa -NL reviewed the initial whitepapers and made suggestions for improvement.

Culemborg, January 2025.

Modeling techniques

Information and data modeling is an important part of many fields in business and information science. Think for example of information analysis, data or enterprise architecture, system design, data exchange, service orientation, microservices, database design and data flow models.
Due to these different manifestations, a large number of modeling methods have emerged over the years around information and data modeling. Each modeling method has its own specific characteristics, which on the one hand can be used for certain groups of stakeholders and on the other hand can be used to make a certain aspect area of data clear.
When developing information or data models, one can therefore search for which modeling technique best suits the target group or stakeholders. But also most clearly maps out the relevant aspects. Choosing the right modeling method can contribute to the adequate transfer of information about the domain that is communicated to the stakeholders.
Given the many modeling techniques, aspect areas and stakeholders, finding a good modeling technique is not always easy. Reason to write a book within the architecture assistant in which the most important data modeling techniques are discussed based on a simple framework.

Information and Data Modeling

Introduction

This chapter introduces a number of chapters on data modeling techniques and describes the frameworks and principles that we will use when describing the different modeling formats.

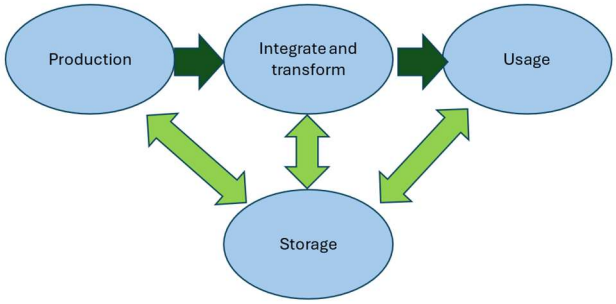
History

Data has been around for centuries, perhaps even since we first started writing things down. Over the years, the amount of data and the complexity of data structures have increased. The advent of automated applications has caused this development to explode.
In recent years, there have been several developments in the application of data. Think of Data Analysis, Business Intelligence, Artificial Intelligence and Data Management. These social trends indicate how extensive data applications are today.
This increase in complexity in data has led to a growing need for information and data models. The first forms of data modeling date back to the 1960s. For example, Codd introduced the relational data model in 1969. This model is still used in the field of databases.
In addition to relational models, many new data modeling techniques have emerged over the years. Think of object-oriented models such as UML or flow diagrams such as DFD. This has led to a multitude of different information and data modeling methods.

Data lifecycle

Information and data can be viewed in many ways. Many forms of modeling assume a certain point of view. However,
In this book we would like to use frameworks that can be used in any situation to plot the different concepts in a modeling approach.
Over the years I have started using a number of frameworks that allow me to represent different aspects of data modeling. In this book we use three of them that we will explain in more detail in this and the following sections. We use this standardized way to describe each type of modeling.

The first framework is the data lifecycle. This lifecycle shows how data appears in a limited number of steps. The image below shows the framework.



Explanation of the framework:

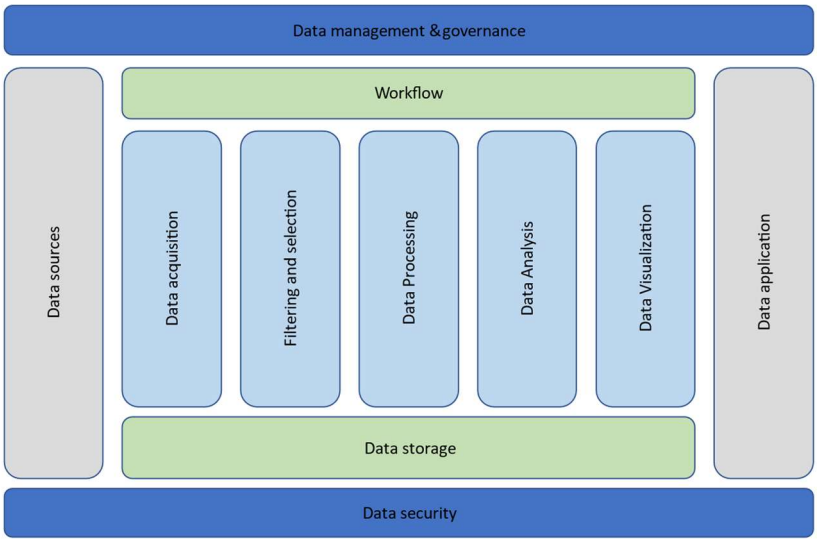
- **Production:** data is produced and therefore comes into being. This can be in many ways. Think of people who enter data via forms, logs and the like in information systems or devices such as smart meters and mobile phones that produce data.
- **Usage:** data that is produced will be used at some point and location, for example when making decisions based on data analysis , alerting for deviations in data production or use in work processes (on the basis of which decisions are made).
- **Integration and transport:** Sometimes there is a physical or model-related distance between the data that is produced and the data that is used. Think for example of the differences between transactional and DWH data models. But also of the place where the data is produced (smart meter in the field) and the place where it is used (control center).
- **Storage:** Storage makes data persistent, for example when there is a time difference between data production and use or if the data can be reused at a later time or in a different context in a different form of use. In addition, stored data can be used to increase the efficiency of production and use (reuse of data). Reason to show an arrow from each step in

the life cycle in which the data can flow in both directions (from and to storage).

A data life cycle is sometimes assumed to represent the origin, mutation and end of data entities. This is not the case. Although this is an interesting type of modeling (see for example the Object Event Table), this life cycle has a more limited scope. If desired, it is possible to place multiple life cycles in a row if this is relevant. The section on stakeholders provides an example of how the life cycle can be applied in explaining data modeling forms.

Data pipe

The data pipe is a detailing of the data lifecycle and is often used in (big) data integration projects. It is actually a framework in which you can map different project activities, deliverables and modeling techniques. This helps to map the complexity in a simple way. The image below shows the data pipe.



The data pipe is read from left to right and shows the steps that need to be taken to be able to use the production of data in a data application (data usage). In addition, a number of extra dimensions

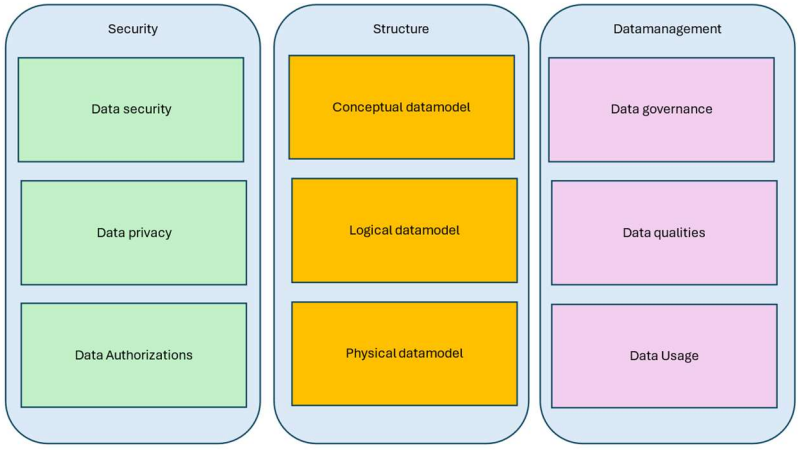
have been added, namely data management and security aspects. In the list below a brief explanation of the components:

- **Data sources:** data collection that is used as raw material for the data application.
- **Data acquisition:** activities that achieve the acquisition of relevant data sources.
- **Filtering and selection:** editing the data from the data sources to those datasets that are relevant for the application.
- **Processing:** transformation, manipulation and enrichment to make the data model suitable for adequate data analysis, visualization and application.
- **Analysis:** Activities in which the transformed data is used to search for connections, patterns or statistical relationships.
- **Visualization:** Making the analysis results visible to support analysts or for presentation to other stakeholders.
- **Application:** use of data in various forms of decision-making.
- **Data storage:** storage of the data and intermediate products for later use in subsequent steps or other data flows.
- **Security:** security and privacy aspects of data.
- **Data management:** data is an asset and therefore management must take place to increase or maintain the value of data.
- **Workflow:** automating and standardizing processing steps on the data.

As mentioned, this is a detailing of the data lifecycle and focuses on the integration aspects. This may mean that this framework is not relevant for every data modeling technique.

Framework

In addition to the more dynamic lifecycle aspect, it is possible to map data modeling techniques onto an architecture framework. This framework is specific to data modeling and consists (as is often the case) of a three by three framework. The image below shows this framework.

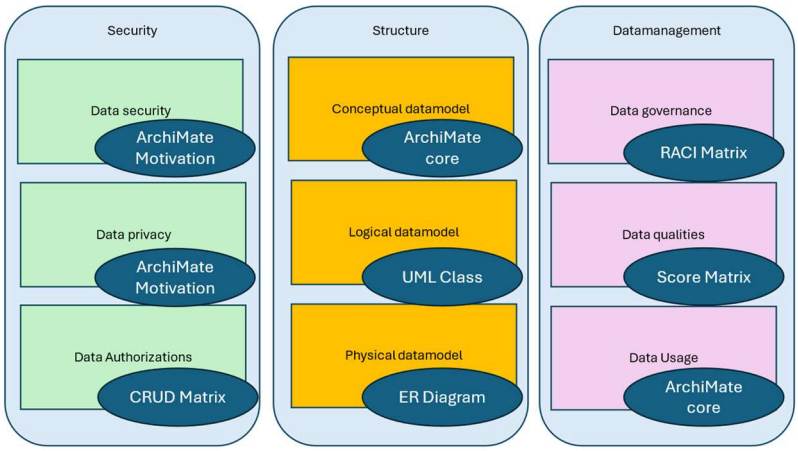


This framework is based on three columns:

- **Structure:** Modeling the structure of data as it is stored or used.
- **Physical data model** : modeling of data based on the technical platform or implementation, for example in data storage and transport. This is often specific and based on a chosen technology.
- **Logical model** : A platform-independent model usually used for an abstract representation of a physical data model.
- **Conceptual model** : the most abstract model for describing data structures. It often leaves out details of structure but addresses how data relates to other concepts such as business processes, information systems, or organizational units
- **Security:** data is the carrier of the information content and security is therefore an important aspect to protect this information.
- **Data authorizations** : models to describe how and for which person or application data entities are accessible.
- **Data privacy** : privacy is a special form of security aimed at protecting personal data.
- **Security** : measures and risks surrounding the use and in particular the misuse of data collections.

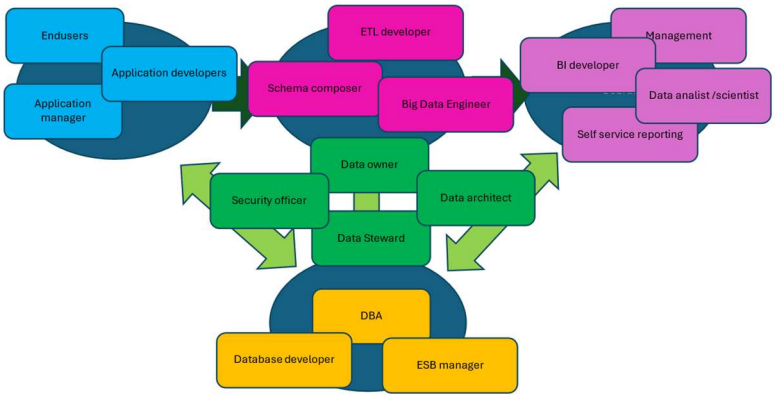
- **Data Management** , data is an asset for many organizations and therefore has value. Data Management are the activities that ensure that this value of data is maintained or increased.
- **Data usage** : how and by which persons and applications is the data used.
- **Data qualities** : data has certain qualities that can be used to determine the value of data. Based on this, data enhancing measures can be taken.
- **Data governance** : organization design to monitor the characteristics that make data an asset for the organization.

The model is shown as a three by three matrix. In fact, three layers can also be defined in this model. The layout of the concepts in the columns already takes this into account. However, the dimensions of the layers in the different columns are different. That is why no layers such as business and application layer are not applicable in this situation. We therefore use the layers but do not display them in the framework.
This framework can be used well to bring the different modeling techniques into proportion to each other. The image below shows how this can be done. This example only shows a part of the modeling techniques but gives a good impression of the application of the framework.



Stakeholders

You hopefully do not draw up (data) models without reason. It is often aimed at informing, convincing or having stakeholders make decisions about a certain aspect of the application of data. This application can of course relate to many aspects. Think for example of data transport and integration, storage, and also in the area of data application in work processes or information systems.
Stakeholders have characteristics that influence the choice of the right modeling form. For example, developers or administrators need more details on the physical or implementation level. However, users or representatives of the business are more interested in how the concepts are applied in an organization. Reason to map and name stakeholders for the different modeling methods.
Stakeholders come in many forms and it is therefore not possible to create a classification or framework. The previously mentioned classification forms are tools that can help with this. The image below shows an example of how stakeholders are depicted on the data lifecycle.



Modeling representations

This section describes a large number of modeling representations. Modeling forms are based on modeling methods. The modeling methods are in fact modeling patterns on the basis of which a modeling form is developed.

Within data modeling, a limited number of modeling representations are used. Some are common, others are only used sporadically. The most important modeling representations are:

- **Lists:** representation of data entities in lists and list views. A frequently used modeling method within modeling tools to quickly get an overview of the available data entities. Within modeling techniques, lists are rarely used because little can be added to the form from a modeling perspective.
- **Trees:** representation of data entities in a hierarchical tree structure with usually one “root” or starting entity to which zero, one or more “ child ” or sibling entities are linked. A characteristic is that a child entity has only one parent entity. An example of a tree is the concept tree or thesaurus.
- **Matrices:** representation of the relationships between two types of entities in a table or matrix representation. Matrix representation is very useful in situations where only one link can exist between two unique types of entities. Often the link (read matrix cell) is enriched with extra selections to add a fourth dimension that characterizes the link. Think of the CRUD matrix.
- **Graphs:** based on mathematical graph theory in which entities (nodes) are linked to each other via associations or links (edges). Graph notation makes it possible to model multiple associations between two entities. In addition, the entities and the associations are often enriched with extra notations such as association types and the cardinalities of a link. An example of a graph modeling method is the UML class diagram.
- **Predicates:** a language-based modeling method where models are often built on the basis of three related entities. For example, the predicate “Student follows Course”. An example of predicate notations is NIAM or FCO-IM.

Modeling techniques

In this chapter we have introduced data modeling and outlined a number of frameworks that can be used to relate the different data modeling approaches to each other.

Table of contents