Designing Data and information Systems

Cloud Foundations-Data Fundamentals

[Pick the date]

[Type the company name]

Chhaya Nikam

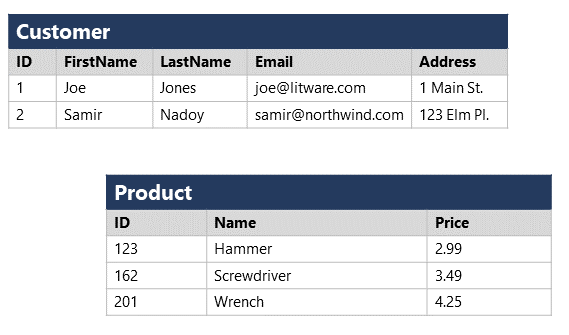
# What is Data

Data is a collection of facts such as numbers, descriptions, and observations used to record information. Data structures in which this data is organized often represents entities that are important to an organization (such as customers, products, sales orders, and so on). Each entity typically has one or more attributes, or characteristics (for example, a customer might have a name, an address, a phone number, and so on).

You can classify data as structured, semi-structured, or unstructured.

## Structured data

Structured data is data that adheres to a fixed schema, so all of the data has the same fields or properties. Most commonly, the schema for structured data entities is tabular - in other words, the data is represented in one or more tables that consist of rows to represent each instance of a data entity, and columns to represent attributes of the entity. For example, the following image shows tabular data representations for Customer and Product entities.



Structured data is often stored in a database in which multiple tables can reference one another by using key values in a relational model; which we'll explore in more depth later.

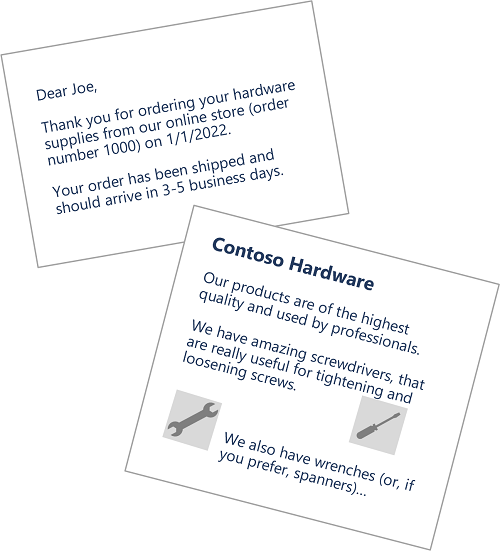
## Semi-structured data

Semi-structured data is information that has some structure, but which allows for some variation between entity instances. For example, while most customers may have an email address, some might have multiple email addresses, and some might have none at all.

One common format for semi-structured data is JavaScript Object Notation (JSON). The example below shows a pair of JSON documents that represent customer information. Each customer document includes address and contact information, but the specific fields vary between customers.

## Unstructured data

Not all data is structured or even semi-structured. For example, documents, images, audio and video data, and binary files might not have a specific structure. This kind of data is referred to as unstructured data.



## What are Data Stores

Organizations typically store data in structured, semi-structured, or unstructured format to record details of entities (for example, customers and products), specific events (such as sales transactions), or other information in documents, images, and other formats. The stored data can then be retrieved for analysis and reporting later.

There are two broad categories of data store in common use:

* File stores
* Databases

<https://www.sap.com/insights/what-is-data-modeling.html>

### Data modeling is the process of diagramming data flows. When creating a new or alternate database structure, the designer starts with a diagram of how data will flow into and out of the database. This flow diagram is used to define the characteristics of the data formats, structures, and database handling functions to efficiently support the data flow requirements. After the database has been built and deployed, the data model lives on to become the documentation and justification for why the database exists and how the data flows were designed.

The data model that results from this process provides a framework of relationships between data elements within a database as well as a guide for use of the data. Data models are a foundational element of software development and analytics. They provide a standardized method for defining and formatting database contents consistently across systems, enabling different applications to share the same data.

## Why is data modeling important?

A comprehensive and optimized data model helps create a simplified, logical database that eliminates redundancy, reduces storage requirements, and enables efficient retrieval. It also equips all systems with a ‘single source of truth’ – which is essential for effective operations and provable compliance with regulations and regulatory requirements. Data modeling is a key step in two vital functions of a digital enterprise.

### Software development projects (new or customizations) performed by IT professionals

Before designing and building any software project, there must be a documented vision of what the final product will look like and how it will behave. A big part of that vision is the set of business rules that govern the desired functionality. The other part is the data description – the data flows (or data model) and the database design to support it.

Data modeling keeps a record of the vision and provides a roadmap for the software designers. With the database and data flows fully defined and documented, and systems developed per those specifications, the systems should deliver the expected functionality required to keep the data accurate (assuming that procedures have been properly followed.)

### Analytics and visualization – or business intelligence – a primary decision-making tool for users

With increasing data volumes and growing numbers of users, organizations need a way to turn raw data into actionable information for decision-making. Not surprisingly, the demand for data analytics has grown dramatically. Data visualization makes data even more accessible to users by presenting the data graphically.

Today’s data models transform raw data into useful information that can be turned into dynamic visualizations. Data modeling prepares the data for analysis: cleansing the data, defining the measures and dimensions, and enhancing data by establishing hierarchies, setting units and currencies, and adding formulas.

## What are the types of data modeling?

The three primary data model types are relational, dimensional, and entity-relationship (E-R). There are also several others that are not in general use, including hierarchical, network, object-oriented, and multi-value. The model type defines the logical structure – how the data is stored, logically – and therefore how it is stored, organized, and retrieved.

1. **Relational:** Although “older” in approach, the most common database model still in use today is relational, which stores the data in fixed-format records and arranges data in tables with rows and columns. The most basic type of data model has two elements: measures and dimensions. Measures are numeric values, such as quantities and revenue, used in mathematical calculations like sum or average. Dimensions can be text or numeric. They are not used in calculations and include descriptions or locations. The raw data is defined as a measure or a dimension. Other terminology used in relational database design includes “relations” (the table with rows and columns), “attributes” (columns), “tuples” (rows), and “domain” (set of values allowed in a column). While there are additional terms and structural requirements that define a relational database, the important factor is the relationships defined within that structure. Common data elements (or keys) link tables and data sets together. Tables can also be related explicitly, like parent and child relationships including one-to-one, one-to-many, or many-to-many.
2. **Dimensional:** Less rigid and structured, the dimensional approach favors a contextual data structure that is more related to the business use or context. This database structure is optimized for online queries and [data warehousing](https://www.sap.com/insights/what-is-a-data-warehouse.html) tools. Critical data elements, like a transaction quantity for example, are called “facts” and are accompanied by reference information called “dimensions,” be that product ID, unit price, or transaction date. A fact table is a primary table in a dimensional model. Retrieval can be quick and efficient – with data for a specific type of activity stored together – but the lack of relationship links can complicate analytical retrieval and use of the data. Since the data structure is tied to the business function that produces and uses the data, combining data produced by dissimilar systems (in a data warehouse, for instance) can be problematic.
3. **Entity-Rich (E-R):** An E-R model represents a business data structure in graphical form containing boxes of various shapes to represent activities, functions, or “entities” and lines to represent associations, dependencies, or “relationships.” The E-R model is then used to create a relational database with each row representing an entity and the fields in that row contain attributes. As in all relational databases, “key” data elements are used to link tables together.

## What are the three levels of data abstraction?

There are many types of data models, with different types of possible layouts. The data processing community identifies three kinds of modeling to represent levels of thought as the models are developed.

### Conceptual data model

This is the “big picture” model that represents the overall structure and content but not the detail of the data plan. It is the typical starting point for data modeling, identifying the various data sets and data flow through the organization. The conceptual model is the high-level blueprint for development of the logical and physical models and is an important part of the data architecture documentation.

### Logical data model

The second level of detail is the logical data model. It most closely relates to the general definition of “data model” in that it describes the data flow and database content. The logical model adds detail to the overall structure in the conceptual model but does not include specifications for the database itself as the model can be applied to various database technologies and products. (Note that there may not be a conceptual model if the project relates to a single application or other limited system.)

### Physical data model

The physical database model describes the specifics of how the logical model will be realized. It must contain enough detail to enable technologists to create the actual database structure in hardware and software to support the applications that will use it. Needless to say, the physical data model is specific to a designated database software system. There can be multiple physical models derived from a single logical model if different database systems will be used.

## Data modeling process and techniques

Data modeling is inherently a top-down process, starting with the conceptual model to establish the overall vision, then proceeding to the logical model, and finally the detailed design contained in the physical model.

Building the conceptual model is mostly a process of converting ideas into a graphical form that resembles a programmer developer’s flow chart.

Modern data modeling tools can help you define and build your logical and physical data models and databases. Here are a few typical data modeling techniques and steps:

* **Determine entities and create an entity relationship diagram (ERD).** Entities can be better described as “data elements of interest to your business.” For example, “customer” would be an entity. “Sale” would be another. On an ERD, you document how these different entities relate to each other in your business and which high-level connections exist between them.
* **Define your facts, measures, and dimensions.**A fact is the part of your data that indicates a specific occurrence or transaction, like the sale of a product. Your measures are quantitative, like quantity, revenue, cost, and so on. Your dimensions are qualitative measures, such as descriptions, locations, and dates.
* **Create a data view link using a graphical tool or via SQL queries.** If you are unfamiliar with SQL, the graphical tool is the most intuitive option, allowing you to drag and drop elements into your model and visually build your connections. While creating a view, you have the option to combine tables and even other views into a single output. When you select a source in the graphical view and drag it on top of a source already associated with the output, you will have the option to either join or create a union of these tables.

Modern analytics solutions can also help you select, filter, and connect data sources using a graphical drag-and-drop display. Advanced tools are available for data experts typically working in IT – but users can also create their own stories by visually creating a data model and organizing tables, charts, maps, and other objects to tell a story based on data insights.

[Learn more](https://www.sap.com/products/technology-platform/cloud-analytics.html" \o "" \t "_self)

## Data modeling examples

For any application – whether business, entertainment, personal, or other – data modeling is a necessary early step in designing the system and defining the infrastructure needed to enable the system. This includes any type of transactional system, data processing application set or suite, or any other system that collects, creates, or uses data.

Data modeling is imperative for data warehousing because a data warehouse is a repository for data brought in from multiple sources, which likely have similar or related data in different formats. It is necessary to first map out the warehouse formats and structure in order to determine how to manipulate each incoming data set to conform to the needs of the warehouse design – so that the data will be useful for analysis and data mining. The data model is then an important enabler for analytical tools, executive information systems (dashboards), data mining, and integration with any and all data systems and applications.

In the early stages of design for any system, data modeling is a key prerequisite that all the other steps and stages depend on to establish the foundation upon which all of the programs, functions, and tools rely. The data model is like a common language that allows systems to communicate through their understanding and acceptance of the data as described in the model. This is more important than ever in today’s world of [Big Data](https://www.sap.com/insights/what-is-big-data.html), [machine learning](https://www.sap.com/insights/what-is-machine-learning.html), [artificial intelligence](https://www.sap.com/products/artificial-intelligence.html), [cloud connectivity](https://www.sap.com/insights/what-is-cloud-computing.html), [IoT](https://www.sap.com/insights/what-is-iot-internet-of-things.html" \t "_self), and distributed systems including edge computing.

## Evolution of data modeling

In a very real sense, data modeling has been around for as long as data processing, data storage, and computer programming, although the term itself probably only came into common use around the time that database management systems began to evolve in the 1960s. There’s nothing new or innovative about the concept of planning and architecting a new structure. Data modeling itself has become more structured and formalized as more data, more databases, and more varieties of data have emerged.

Today, data modeling is more essential than ever as technologists struggle with new sources of data (IoT sensors, location-aware devices, clickstreams, social media) along with an onrush of unstructured data (text, audio, video, raw sensor output) – at volumes and velocity that exceed the capabilities of traditional systems. There is now a constant demand for new systems, innovative database structures and techniques, and new data models to tie this new development effort together.

What is Data lake