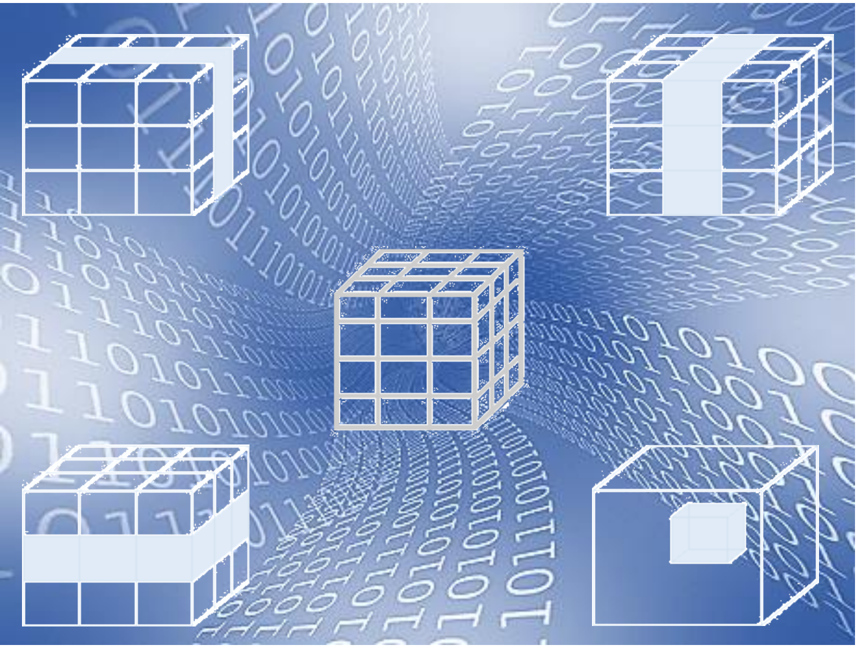
Prof. Dr. Bastian Amberg
School of Business & Economics
Department of Information Systems



Business Intelligence

03 Data Warehouse – OLAP & Modeling I

Prof. Dr. Bastian Amberg
(summer term 2024)
~~3.5.2024~~ 8.5.2024

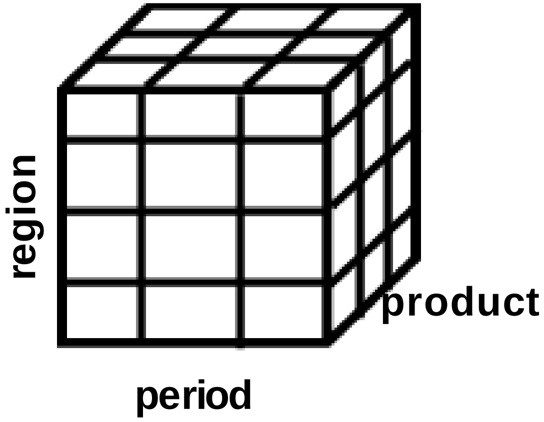
Schedule

		Wed., 10:00-12:00		Fr., 14:00-16:00 (Start at 14:30)		Self-study		
	W1	17.4.	(Meta-)Introduction	19.4.		Python-Basics	Chap. 1	
	W2	24.4.	Data Warehouse – Overview & OLAP	26.4.	*[Blockveranstaltung SE Prof. Gersch]*		Chap. 2	
Basics	W3	1.5.		3.5.	Data Warehouse Modeling I		Chap. 3	
	W4	8.5.	Data Warehouse Modeling I & II	10.5.	Data Mining Introduction			
	W5	15.5.	CRISP-DM, Project understanding	17.5.	Python-Basics-Online Exercise	Python-Analytics	Chap. 1	
	W6	22.5.	Data Understanding, Data Visualization	24.5.	*No lectures, but bonus tasks*		Chap. 2	
	W7	29.5.	Data Preparation	31.5.	*1.) Co-Create your exam* *2.) Earn bonus points for the exam*			
Main Part	W8	5.6.	Predictive Modeling I	7.6.	Predictive Modeling II (10:00 -12:00)	BI-Project	Start	
	W9	12.6.	Fitting a Model I	14.6.	Python-Analytics-Online Exercise		|	
	W10	19.6.	*Guest Lecture*	21.6.	Fitting a Model II		|	Case Study
	W11	26.6.	How to avoid overfitting	28.6.	What is a good Model?		|	
	W12	3.7.	Project status update Evidence and Probabilities	5.7.	Similarity (and Clusters) From Machine to Deep Learning I		|	
Deep-ening	W13	10.7.		12.7.	From Machine to Deep Learning II		|	
	W14	17.7.	Project presentation	19.7.	Project presentation		End	
					Klausur 1.Termin ~ 22.7. bis 3.8. *Klausur 2.Termin ~ 23.9. bis 5.10.*	Projektbericht		

Ref.

Last lesson

- ✓ Operational databases vs. Data warehouses (vs. Data lakes)
- ✓ Basic architecture of a data warehouse system
- ✓ Analytical data are represented by multidimensional data models
  Distinguish facts and dimensions!
- o How to extract information? (→OLAP)
- o How can multidimensional data models be developed and stored?



Kahoot-Fragen zu den Inhalten
www.kahoot.it
(über Smartphone oder Laptop)
PIN folgt

Ref.

Agenda

Freie Universität Berlin

(1) Online Analytical Processing (OLAP)

Different query methods

Properties of OLAP

Common OLAP functionality

(2) Modeling layers

Basic Elements of multidimensional modeling

Conceptual modeling

Logical modeling

Physical modeling

Ref.

Query methods

Three means to query databases

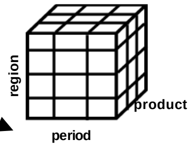
Decision makers need flexible and easy access to data in order to do complex analysis

Programmed reports

- arbitrarily modifiable
- programmer required for changes

Query languages

- standardized and powerful
- difficult to learn
- e.g. SQL, QBE



OLAP

- flexible ad-hoc querying
- possible without expertise

dBase code for “Which are the properties of the products of the department ‚Mobile Computing‘?”:

```
use PRODUCTS
copy to TMP
use TMP
delete for producttype <> ‘MOBILE’
total on PRODUCTS to RESULT
display all
```

SQL query for ”Which are the properties of the products of the department ‚Mobile Computing‘?”:

```
SELECT *
FROM Products
WHERE producttype = ‘MOBILE’
```

Using SQL for multidimensional querying is difficult:

- Several **(inner) queries** (and joins) needed in many cases
- Queries often become quite complex
- Difficult to do time series analysis
- Limited ways for doing statistical calculations

SQL query for “What was the average sales of the department *“Mobile Computing”* to *Government customers* for the *third quarter* of calendar year 2001?”

```
SELECT customer, ROUND(AVG(sales),2)as average,
      ROUND(MIN(sales),2)as minimum,…,
FROM units_cube_cubeview
WHERE time_calendar_year = 'Q3_2001'
  AND product_ldsc = ‚MOBILE‛
  AND customer_market_segme_prnt
      = 'MARKET_SEGMENT_GOV‛
  AND channel_level = 'TOTAL_CHANNEL‛
GROUP BY customer
ORDER BY customer;
```

Ref.

Agenda

Freie Universität Berlin

(1) Online Analytical Processing (OLAP)

Different query methods

Properties of OLAP &

Common OLAP functionality

(2) Modeling layers

Basic Elements of multidimensional modeling

Conceptual modeling

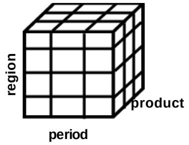
Logical modeling

Physical modeling

Ref.

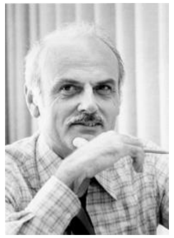
Online Analytical Processing (OLAP)

Let's focus on end users, and their access to data marts by OLAP systems



OLAP systems

- combine querying and interactive analysis
- present a multidimensional view on data



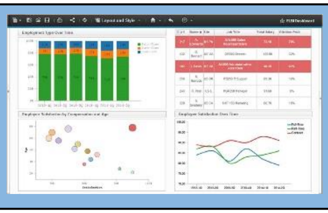
OLAP was introduced by E. F. Codd (one of the founding fathers of relational data bases) in 1993, who established 12 rules to define OLAP

OLAP functionality

- video for illustration
  (exemplary https://www.youtube.com/watch?v=V37vPxlxUwo)

A more concise definition of OLAP is FASMI

Fast	OLAP systems deliver responses to analyze queries within seconds (ideally maximum 5 – 20 seconds)
Analysis of	Cope with any business logic and statistical analysis that is relevant to the user: Mathematic modeling, time series analysis, goal seeking, what-if, drill-down etc., but no programming
Shared	Multiple user access and varying roles with necessary security requirements for confidentiality.
Multidimensional	Truly multidimensional conceptual view of the data
Information	



Bildquelle: IBM Cognos Analytics

Ref. Codd et al. (1993), Chamoni/Gluchowski (2000); Pendse/Creeth (1995)

OLAP functions

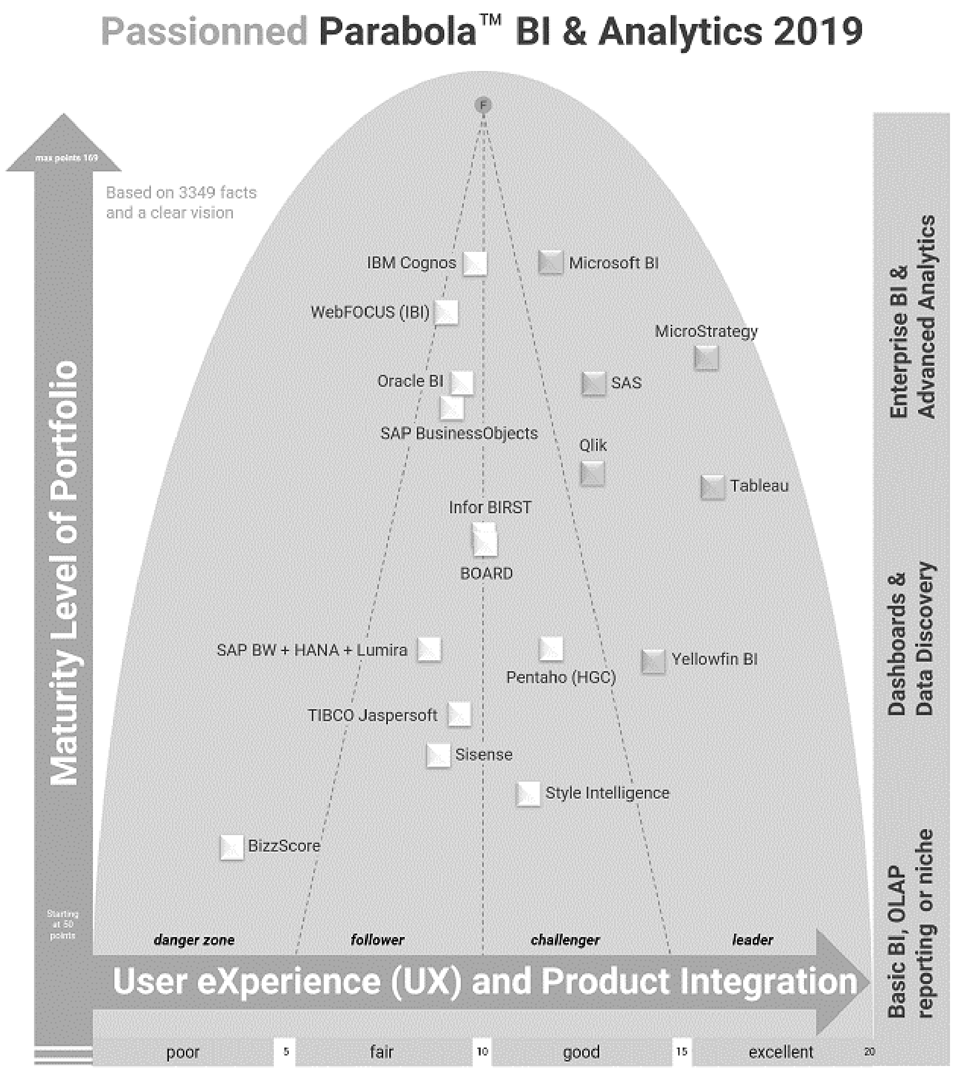
OLAP tools provide a number of standard features

- Different representation modes:
  - absolute as well as relative representation of data
  - 3-dimensional analysis using layers
  - various calculation options (internal or plug-ins)

- Special cube operators provide browsing functions:
  - drilling
    - drill up/down ⇒ detailing/aggregating along a dimension
    - drill through ⇒ access to operational databases
    - ...
  - pivoting (rotating) ⇒ switch rows and columns
  - slicing ⇒ reduce number of dimensions
  - dicing ⇒ cutting parts out of the current cube (filtering)

- Various visualization options

OLAP Tools -> part of BI Tools…



https://www.passionned.com/bi/tools/

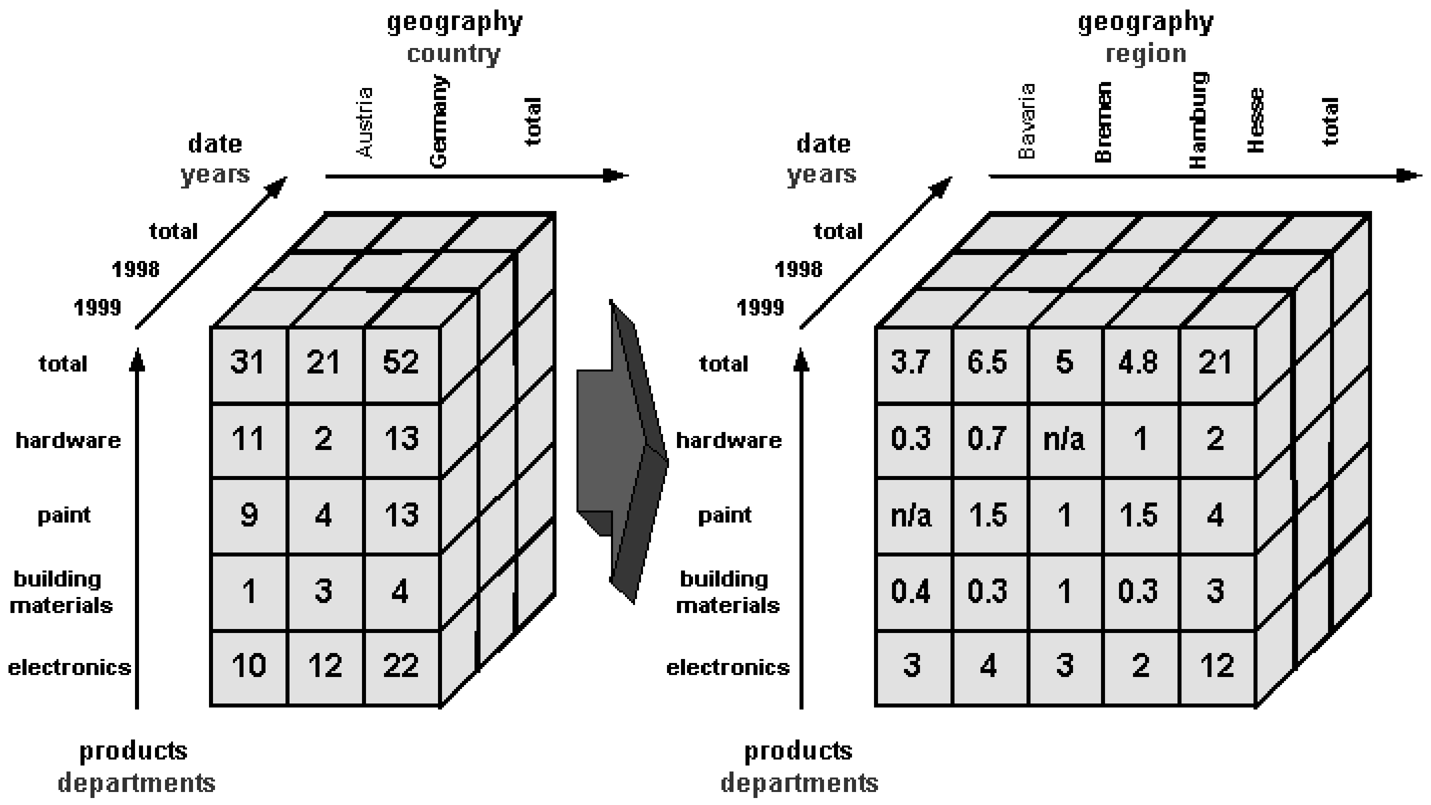
See https://www.passionned.com/bi/#list-business-intelligence-tools for an up-to-date list with detailed information about BI Tools, April 2024

Ref.

Drilling down

More details for specific dimensions

“Show the regions of Germany in detail.”

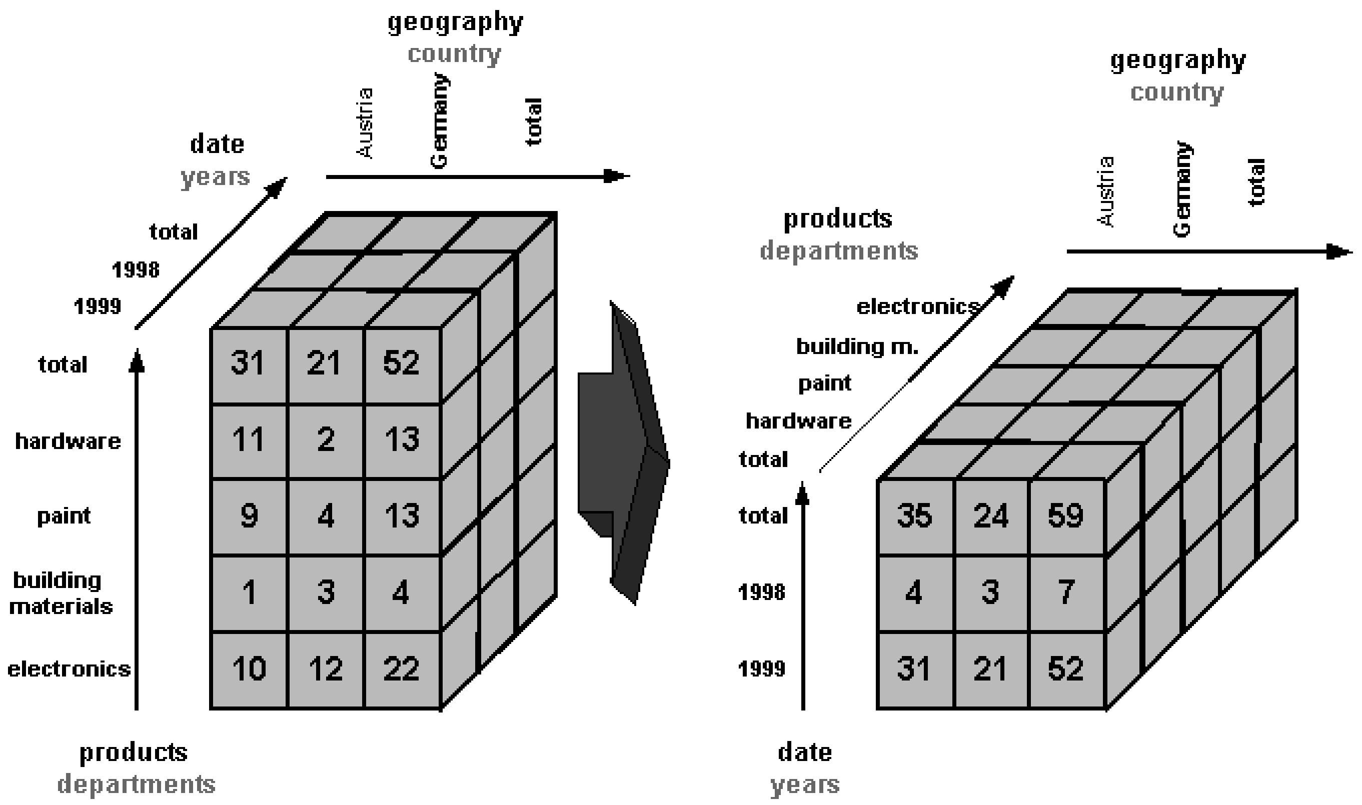


Ref.

Pivoting

Rotate the cube

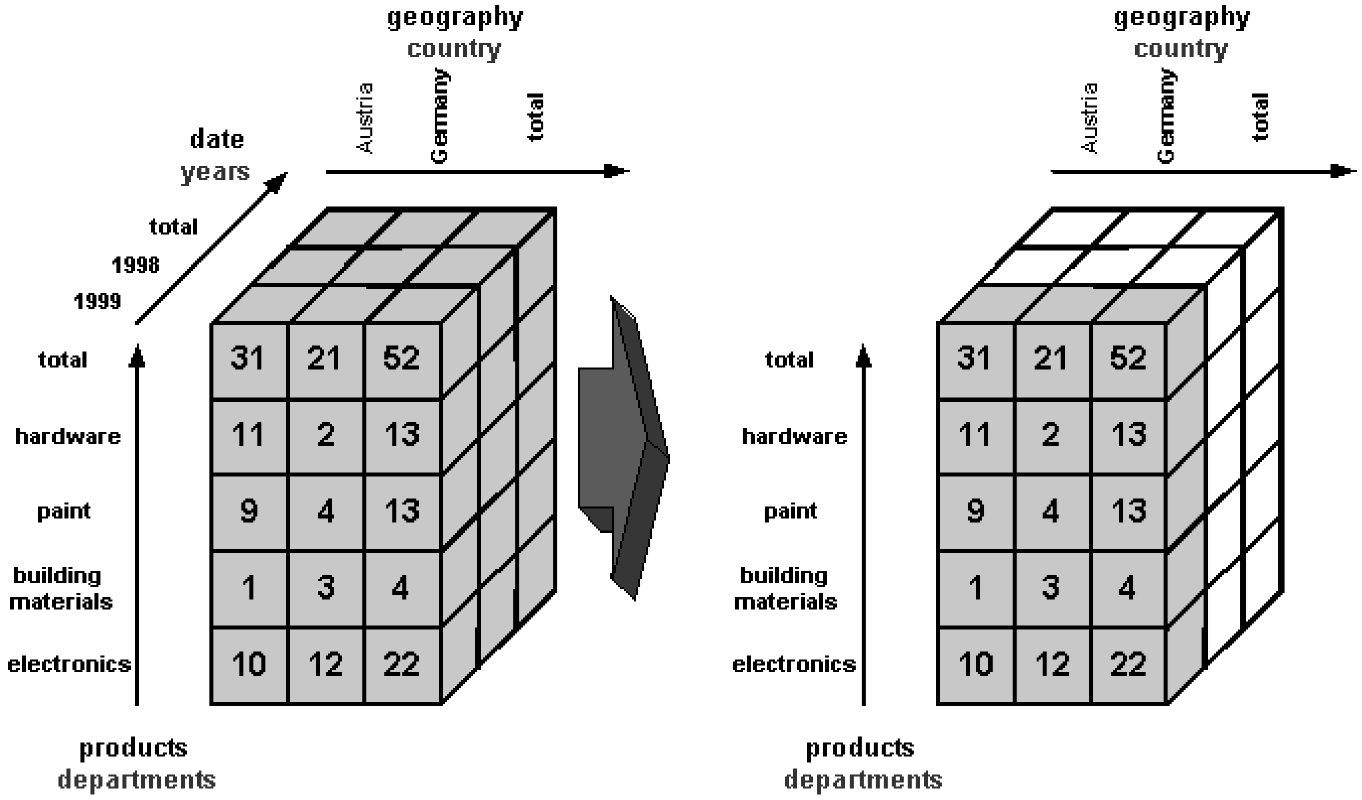
“Show year by country instead of product by country”



Ref.

Slicing

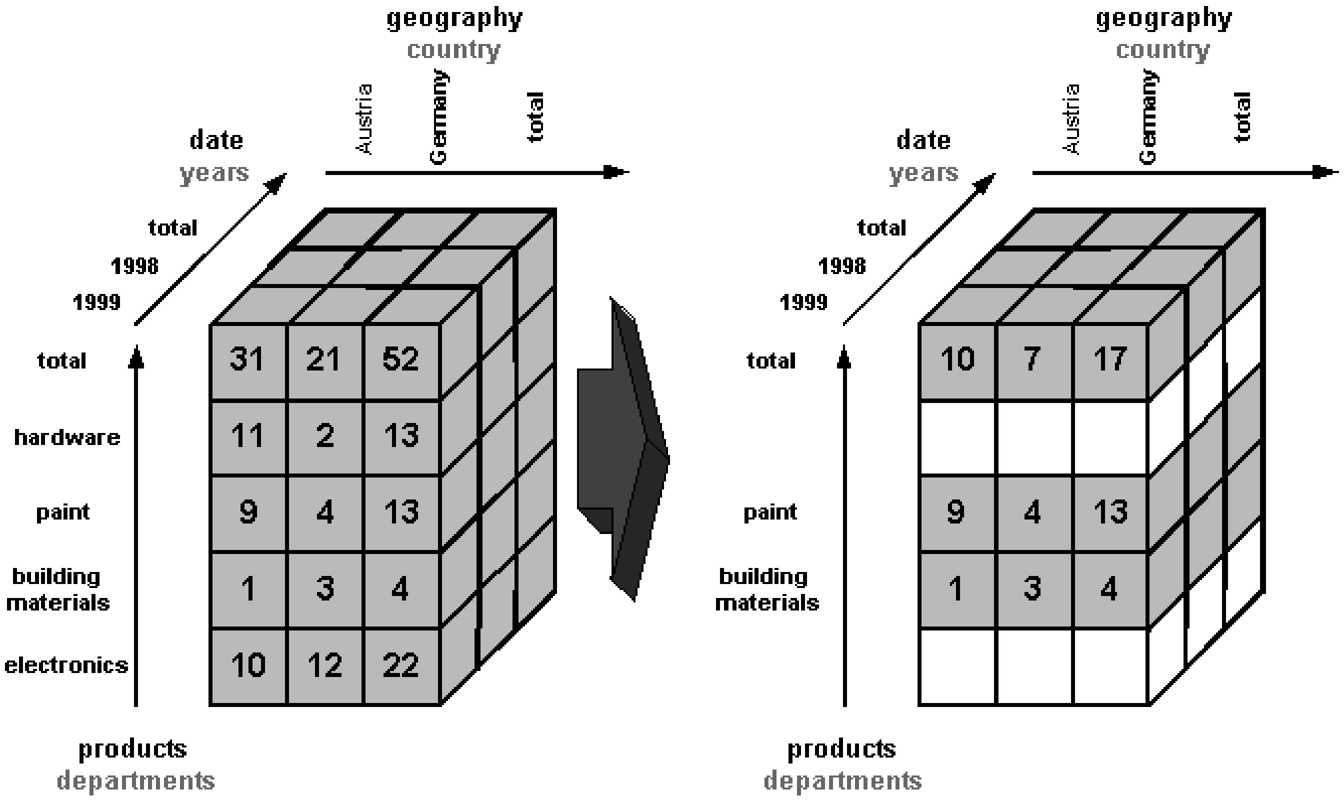
“Show only the values for 1999.”



Ref.

Dicing

“Show only the values for the departments ‘paint’ & ‘building materials’ for all the countries and all the years.”



Ref.

OLAP vs. OLTP

On-line Transactional Processing (OLTP)

- Common way of transactional processing (INSERT, UPDATE, DELETE)
- Primarily used on operational databases (day-by-day business)
- Treats microscopic transactions (e.g., by processing single accounting transactions or order transactions)
- Does not support strategic decisions, but controls and runs subsequent operations

```
66.249.76.[redacted] - - [14/Oct/2012:03:47:21 +0200] "GET /index.php/de/ HTTP/1.1" 200 23963
178.154.211.[redacted] - - [14/Oct/2012:03:49:28 +0200] "GET /robots.txt HTTP/1.1" 200 370
66.249.76.[redacted] - - [14/Oct/2012:04:00:40 +0200] "GET / HTTP/1.1" 303 -
66.249.76[redacted] - - [14/Oct/2012:04:00:41 +0200] "GET /index.php/de/ HTTP/1.1" 200 23961
123.125.71.[redacted] - - [14/Oct/2012:04:19:44 +0200] "GET / HTTP/1.1" 303 -
220.181.108.[redacted]- - [14/Oct/2012:04:19:44 +0200] "GET / HTTP/1.1" 303 -
66.249.76.[redacted] - - [14/Oct/2012:04:30:46 +0200] "GET /index.php/de/konferenzen-uebersicht/konferenzuebersicht HTTP/1.1" 200 20598
180.76.5.[redacted]- - [14/Oct/2012:04:35:20 +0200] "GET / HTTP/1.1" 303 -
```

	OLTP	OLAP
data	operational transactions	management analysis data
user friendliness	low	high
granularity	microscopic	macroscopic
up-to-dateness	current status	historic snapshots
main operations	update (read/write)	query and calculate (read only)
storage efficiency	high	lower
tools	e.g. SQL	proprietary tools

➢ OLAP vs. OLTP in a nutshell

Ref. https://www.youtube.com/watch?v=iw-5kFzIdgY (IBM Technology Video, last access April 2024)

Pros and cons of OLAP

Pro:

- Wide applicability of the method
- OLAP presents quite exact results
- Method is plausible

Con:

- OLAP requires a lot of user interaction
- OLAP regularly requires quite a lot of computing ressources
- Difficult to use automated data mining routines in combination with OLAP

Ref.

Agenda

(1) Online Analytical Processing (OLAP)

Different query methods

Properties of OLAP

Common OLAP functionality

(2) Modeling layers

Basic Elements of multidimensional modeling

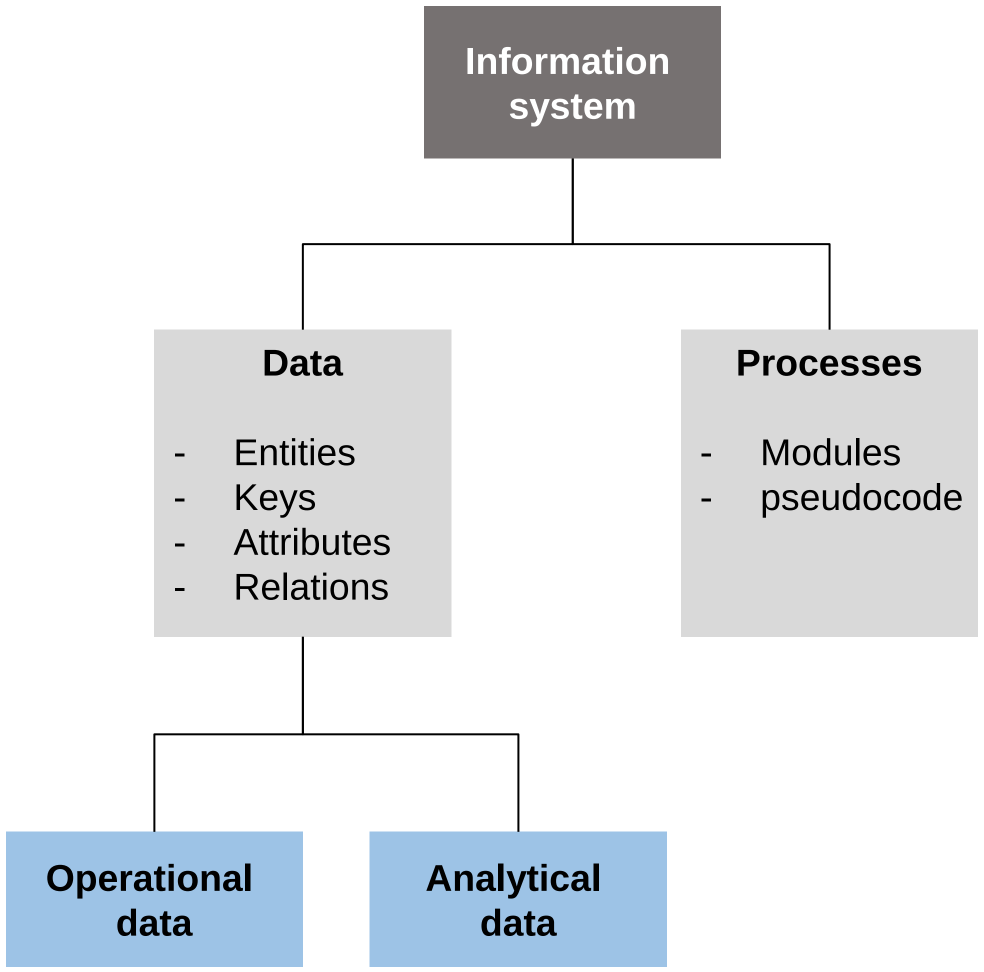
Conceptual modeling

Logical modeling

Physical modeling

Ref.

Modeling of information systems



Ref.

Modeling goals

Operational databases

- Optimize storage efficiency and response time
- Model data which is
  - fine-grained (many details)
  - dynamic (many updates)

- Normalize data
- Minimize redundancy
- Provide data integrity
  - Avoid update anomalies
  - Avoid deletion anomalies
  - Avoid insertion anomalies

Analytical databases

- Support the decision making process
- Maximize user-friendliness and querying efficiency
- Model data which is
  - coarse-grained (less details)
  - static (less updates)

- Data is denormalized
- Redundancy minimization is secondary

→ *Mirror different **views on business measures** within the model*

Ref.

Image: CTSI-Global | Flickr (cc by-sa 2.0)

Multidimensional modeling

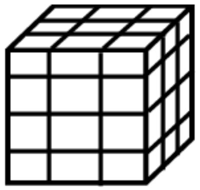
Basic Elements

Common steps compared to operational databases

- Leave out operational data
  (not all attributes necessary)
- Include time dimension
- Integrate pre-calculated attributes
- Reduce join operations

Basic elements of multidimensional models

- Facts
- Dimensions
- Categories
- Aggregation functions



Cube

Measures

Facts

Dimensions

Dimensions

Categories

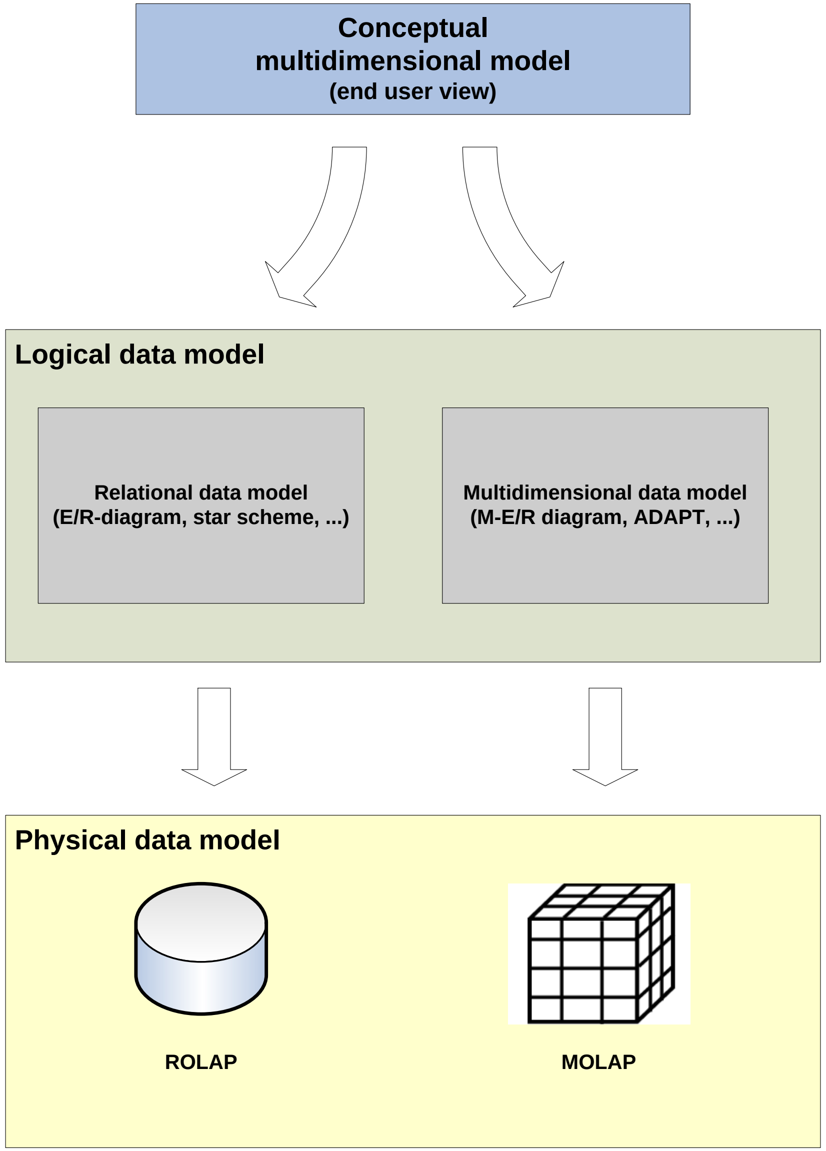
Aggregate functions

Ref.

Multidimensional modeling

Major steps

1. Identify **facts** and **dimensions**
2. Create a **conceptual** data model
3. Derive a **logical** data model from the semantic model
4. Derive a **physical** data model from the logical model



Ref. Hahne (2006)

Facts (= business measures)

Multidimensional models are designed according to the needs of decision makers

➢ Business measures are in the center of interest of decision makers

Definition of business measure:

> *"Business measures are compressed mostly numeric measurements, which refer to important matters of fact within the company and which represent them in a concentrated manner. They provide information about business issues and thereby provide important support for the decision processes within the company."*
> (Langenbeck, 1997, highlights added)

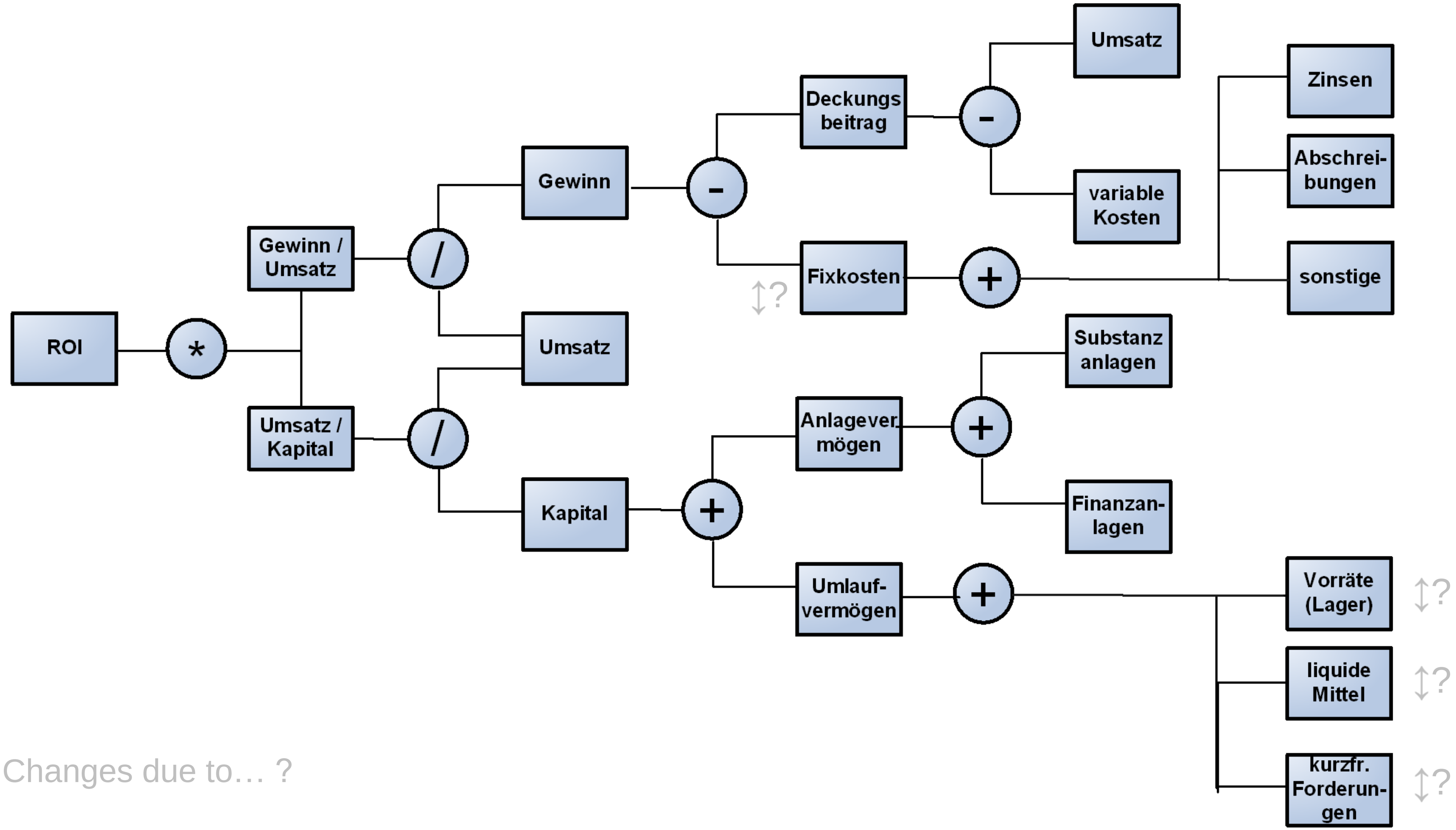
Example business measures: revenues, profits, sales, ROI …

Identification of business measures is one of the basic tasks in multidimensional modeling

Ref.

Example business measure system: ROI

~ „Erfolg im Verhältnis zum eingesetzten Kapital“, „Gewinn in Prozent des investierten Kapitals“, …



Changes due to… ?

Ref.

Dimensions

Decision makers want to **analyze business measures** from **different views** (dimensions)

- ➢ Several dimensions are arranged around one fact

“What amount were the sales revenues for hard disks within the past quarter?”

fact: sales revenues

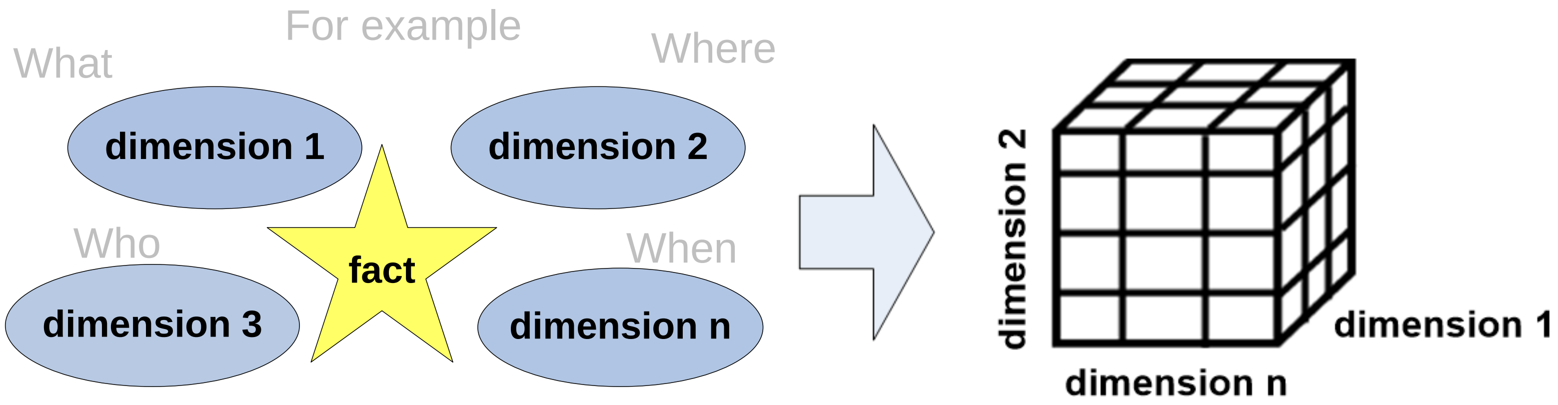
dimensions: range of products, time

“How profitable has our Africa department been on software?”

“What is our growth on A-customers throughout the last quarter?”

During the modeling process, business measures and their set of dimensions are determined

Link to multi-dimensional data structures:



Ref.

Dimensions and categories

Dimension = *finite* **set of categories** which are semantically related to each other with respect to business matters

Categories of one dimension represent a **different levels of aggregation** of the associated business *measures* (facts)

Categories are also known as aggregation objects

An example

dimension: “date”

Four categories: day ⇒ month ⇒ quarter ⇒ year

Resp.: “Sales revenues for hard disks within the past day, month, quarter, year, …?”

Ref. Vaisman/Zimanyi (2014)

Categories

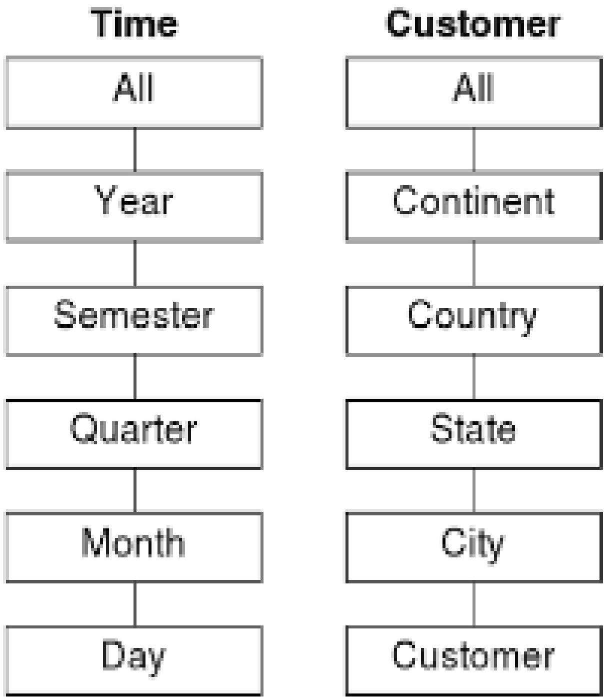
A category is represented by a varying set of elements

- e.g., country = [Germany, Austria, Switzerland], quarter = [q1, q2, q3, q4]

Each dimension consists of **at least one** (real) category

- a category represents the level of granularity
  "**Log-category**" (e.g., "day" in dimension "date")
- a virtual category, encompassing all the others, like
  "**all-category**": encompasses all elements of the Log-category

Number of categories of a dimension is not limited



Dimension schema

Ref.

Facts and aggregation functions

Aggregation function = formula defining the value of facts with respect to the different categories of a dimension

Facts can be classified with respect to aggregation functions:

Additive facts
(distributive aggregation function)

Simple addition possible throughout all the categories of all the associated dimensions

e.g., units sold

Semi-additive facts
(algebraic aggregation function)

Simple addition only possible for a selected number of the categories of the associated dimensions

e.g., not additive over time, but maybe over regions

e.g., current stock/ inventory level, current balance amount

Non-additive facts
(holistic aggregation function)

Simple addition operations not sufficient

e.g., types of average values or ratio values

e.g., temperature

Ref.

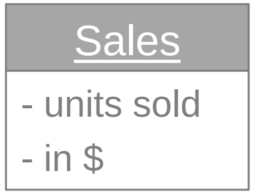
Special types of facts

Fact groups, dimensional and virtual facts

Fact group

set of facts featuring *a common set* of dimensions

e.g., *units sold* and *sales in $* per day

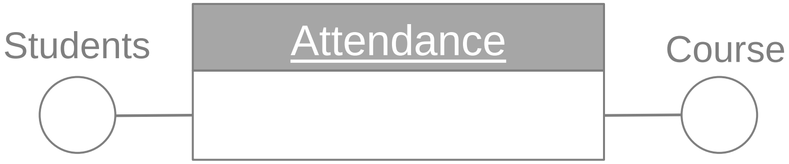


Virtual fact (a.k.a. “factless fact”):

Association between dimensions alone defines the (nominal) fact

e.g., students attendance in class (students, class, time) – virtual fact (0/1) to ask for: *how many students attended class x?*

Relational implementation: only keys in fact table

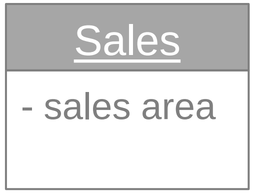


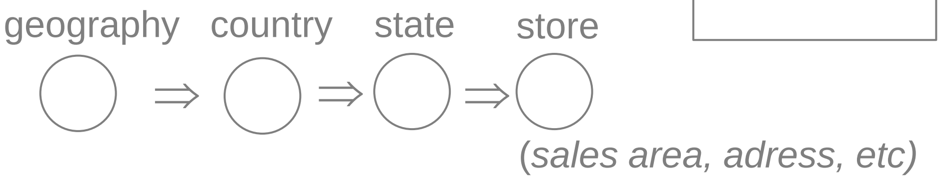
Dimensional fact:

fact is associated with **only one dimension**

dimensional facts are often numerical, non-dimensional attributes of a dimension’s category

e.g., sales area (dimension “geography”)





Ref.

Agenda

(1) Online Analytical Processing (OLAP)

Different query methods

Properties of OLAP

Common OLAP functionality

(2) Modeling layers

Basic Elements of multidimensional modeling

Conceptual modeling

Logical modeling

Physical modeling

Ref.

Conceptual Modeling

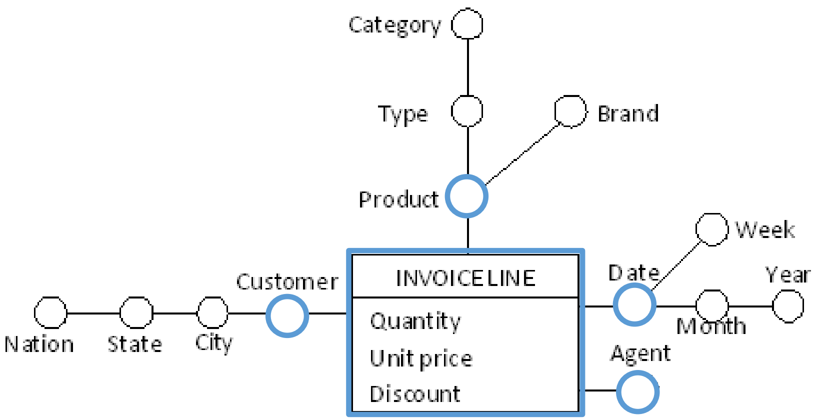
Dimensional fact model (or fact scheme)

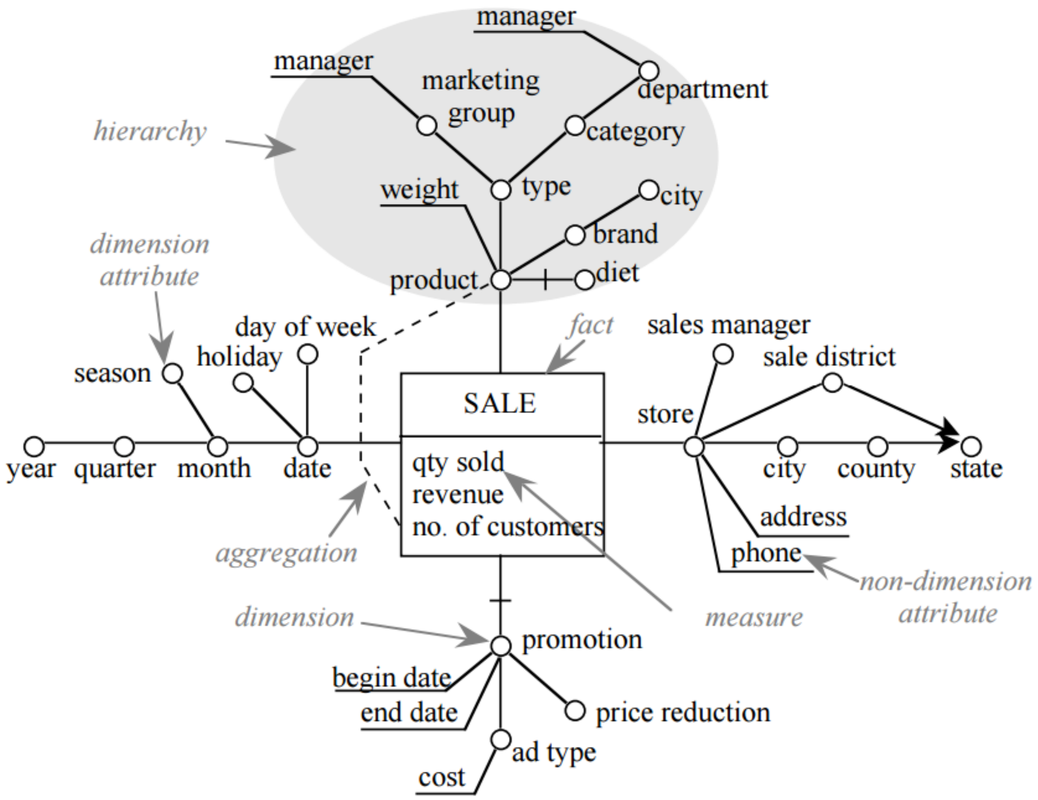
Categories of a dimension arranged in a non-cyclic graph, directed between all-category and log-categories

Categories can have an arbitrary number of (non recursive) relations between each other

Several aggregation paths (e.g., sum/count/mean) may be included in the graph

Hierarchies are discrete attributes and define the granularities of facts (i.e., product -> type -> category)





Ref. Golfarelli et al. (1998), helpful reading (with exercises) by A. Prosser & M.-L. Ossimitz, Hahne (2006)

Image: see ref. (link), Matteo.golfarelli (2010) | Wikimedia

Exercise

Conceptual Modeling (Dimensional fact model)

Design a **conceptual model** for the local Food Company:

Conny's Corner Shop

Your managers need to keep themselves up to date on the number of items in the company's inventory. They especially want to keep an eye on their products with regards to location, and time.

- Conny's Corner Shop sells a range of snacks and beverages. Both categories have different types of products, such as juices and water, as well as prezels and crackers.
- The products are sold under different brands.
- Products have different package types, sizes, and weights.
- They store products in different stores across Europa

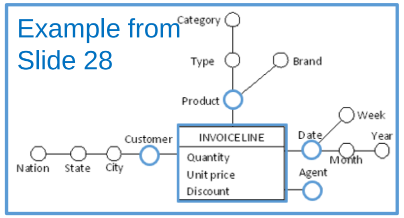
10 Min.

Show your conceptual model as a dimensional fact model. Make reasonable assumptions if necessary.

Ref.

Non-& Cross-dimensional attributes

Dimensional fact model



There may be **various types of relations** between individual categories of one (or more) dimension(s)

Categories having 1:1 relations can be summarized into a single category

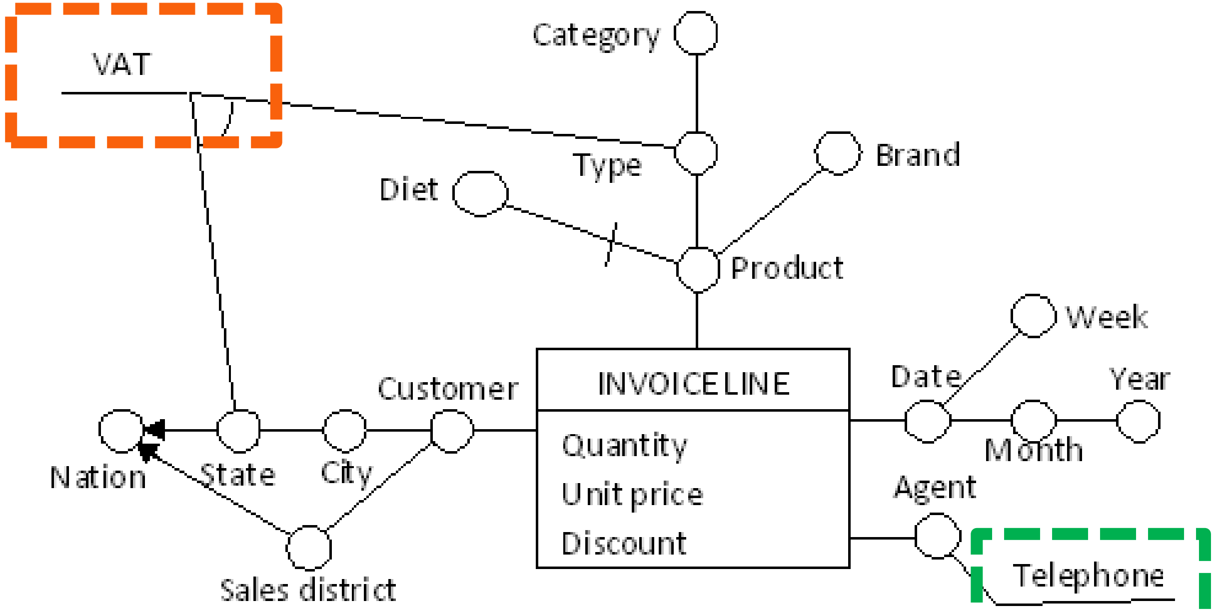
When previous categories become non-dimensional, **descriptive attributes**
Non-dimensional attributes provide additional information and have 1:1 relations with the corresponding categories (phone-No. cannot be aggregated)

OLAP-compliant queries encompassing non dimensional attributes are generally not supported

Categories having **1:N** or N:M relations

When value is defined by multiple categories (e.g., product type and state. For instance, VAT for water or books in Germany vs. USA) they become cross-dimensional attributes

N:M - example



1:N - example

Nation ← State ← City ← Customer

1:N 1:N 1:N

1:1 - example

Ref. Golfarelli et al. (1998), helpful reading (with exercises) by A. Prosser & M.-L. Ossimitz, Hahne (2006)

Image: Matteo.golfarelli (2010) | Wikipedia (cc by-sa 3.0)

Identifying fact groups

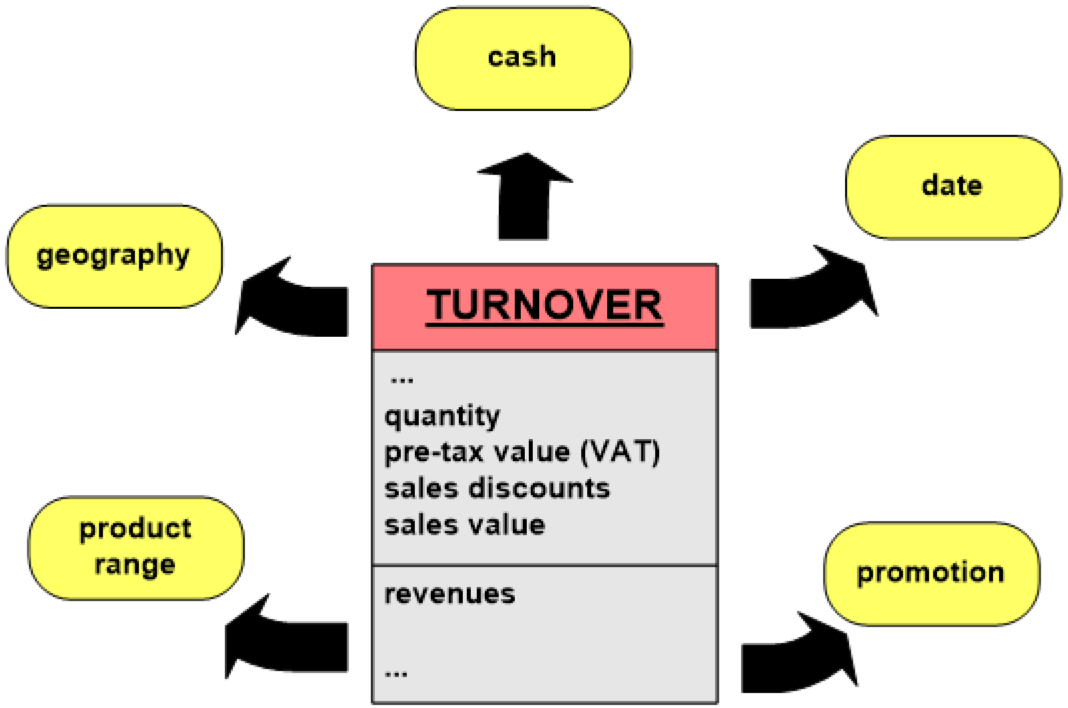
Single Star Scheme

A graphical overview should be created for each fact group

single star scheme

Distinction between materialized facts and derived facts should be drawn

High level of abstraction:
aggregation formulas are commonly not modeled



Remember: **Fact group = set of facts** featuring *a common set* of dimensions

Ref.

Combining fact groups

Multiple Star Scheme

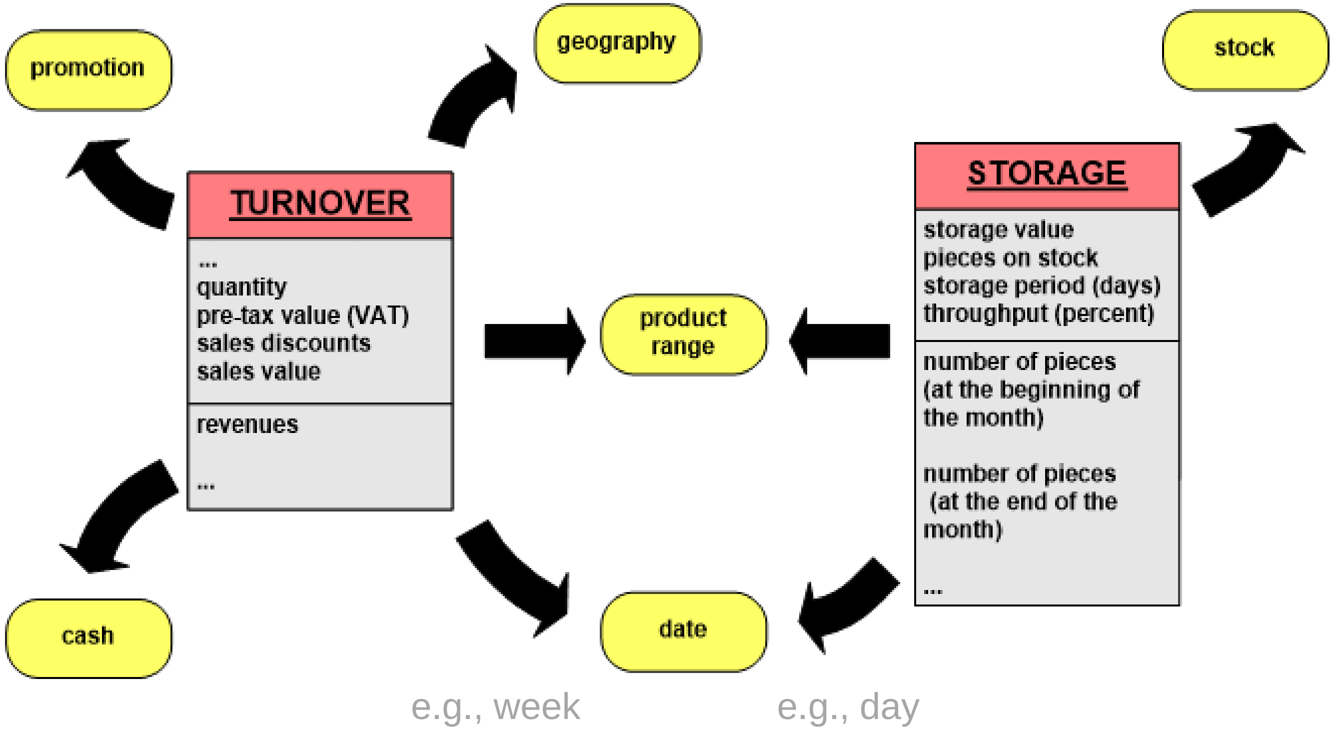
A data warehouse data model encompasses a number of fact groups (**multiple star scheme**)

The sets of associated dimensions of different fact groups may overlap.

Different fact groups may use different log-categories with respect to one common dimension

- e.g., actual / debit values:
  - Actual facts: log-category of dimension date: day
  - Target facts: log-category of dimension date: month

Category used as log-category should be specified in the model (at least if standard log-category is not used)



Ref.

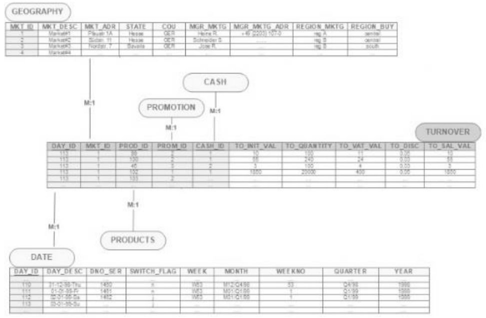
Next Lesson

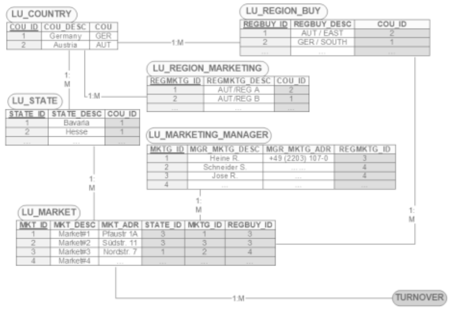
Identify **facts** and **dimensions**

Create a **conceptual** data model

Continuation Data Warehouse Modeling

Derive a logical data model from the semantic model





Derive a physical data model from the logical model

Ref.

Fragen?

- ✓ Online Analytical Processing (OLAP)
  - ✓ Different query methods
  - ✓ Properties of OLAP
  - ✓ Common OLAP functionality

- ✓ Modeling layers
  - ✓ Basic Elements of multidimensional modeling
  - ✓ Conceptual modeling
  - o Logical modeling
  - o Physical modeling

Todos for this Week

1. Support Conny‘s Corner Shop by finishing the conceptual model.
   See exercise on slide 29

2. Python-Basics – Chapter 3
   Kursmaterial > Readings/Übungen > Python Übungen – Jupyter

Ref.

Bibliography

- Böhnlein, M. (2013). *Konstruktion semantischer Data-Warehouse-Schemata*. Springer-Verlag.
- Bulos, D., & Forsman, S. (2000). *Olap Database Design: Delivering on the Promise of the Data Warehouse*. Morgan Kaufmann Publishers Inc..
- Golfarelli, M., Maio, D., & Rizzi, S. (1998). The dimensional fact model: A conceptual model for data warehouses. *International Journal of Cooperative Information Systems*, *7*(02n03), 215-247.
- Hahne M. (2006) Mehrdimensionale Datenmodellierung für analyseorientierte Informationssysteme. In: Chamoni P., Gluchowski P. (eds) Analytische Informationssysteme. Springer, Berlin, Heidelberg
- Jukic, N., Jukic, B., & Malliaris, M. (2008). Online analytical processing (OLAP) for decision support. In Handbook on Decision Support Systems 1 (pp. 259-276). Springer, Berlin, Heidelberg.
- Vaisman, A., & Zimányi, E. (2014). *Data Warehouse Systems*. Springer, Heidelber

Ref.