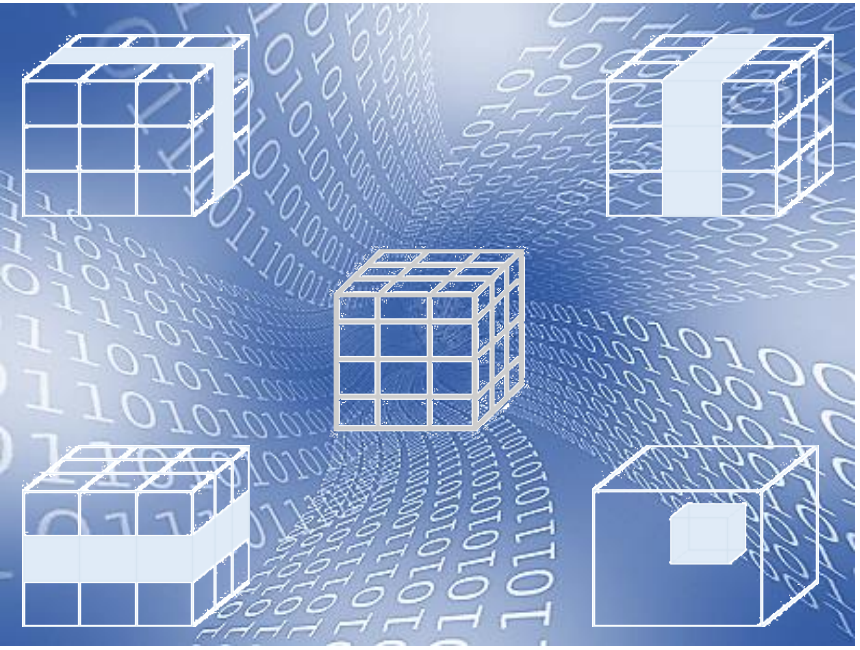
Prof. Dr. Bastian Amberg
School of Business & Economics
Department of Information Systems

Freie Universität Berlin



Business Intelligence

02 Data Warehouse – Overview & OLAP

Prof. Dr. Bastian Amberg
(summer term 2024)
24.4.‘24

Schedule

Freie Universität Berlin

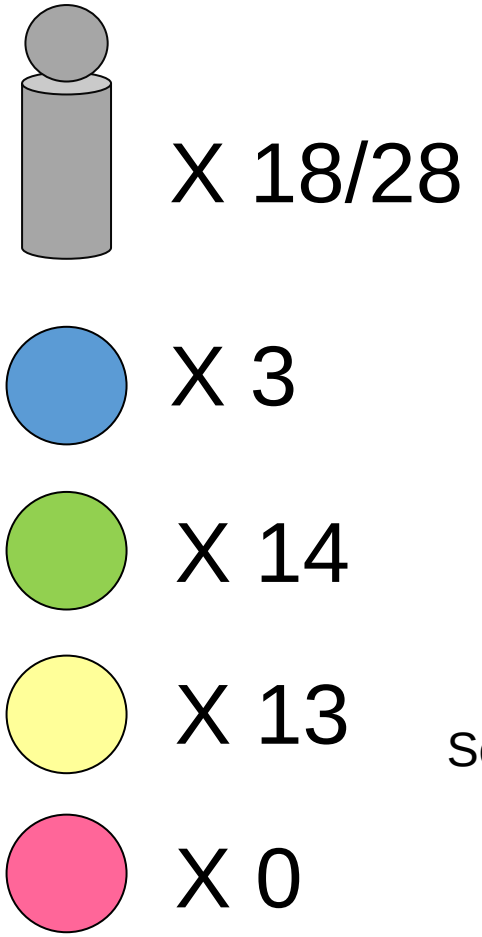
		Wed., 10:00-12:00		Fr., 14:00-16:00 (Start at 14:30)	Self-study	
	W1	17.4.	(Meta-)Introduction	19.4.	Python-Basics	Chap. 1
	W2	24.4.	Data Warehouse – Overview & OLAP	26.4.	*[Blockveranstaltung SE Prof. Gersch]*	Chap. 2
Basics	W3	1.5.		3.5.	Data Warehouse Modeling I	Chap. 3
	W4	8.5.	Data Warehouse Modeling 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* *1.) Co-Create your exam* *2.) Earn bonus points for the exam*	Chap. 2
	W7	29.5.	Data Preparation	31.5.		
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	|
	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
Ref.					*Klausur 1.Termin ~ 22.7. bis 3.8.* *Klausur 2.Termin ~ 23.9. bis 5.10.*	Projektbericht

Case Study

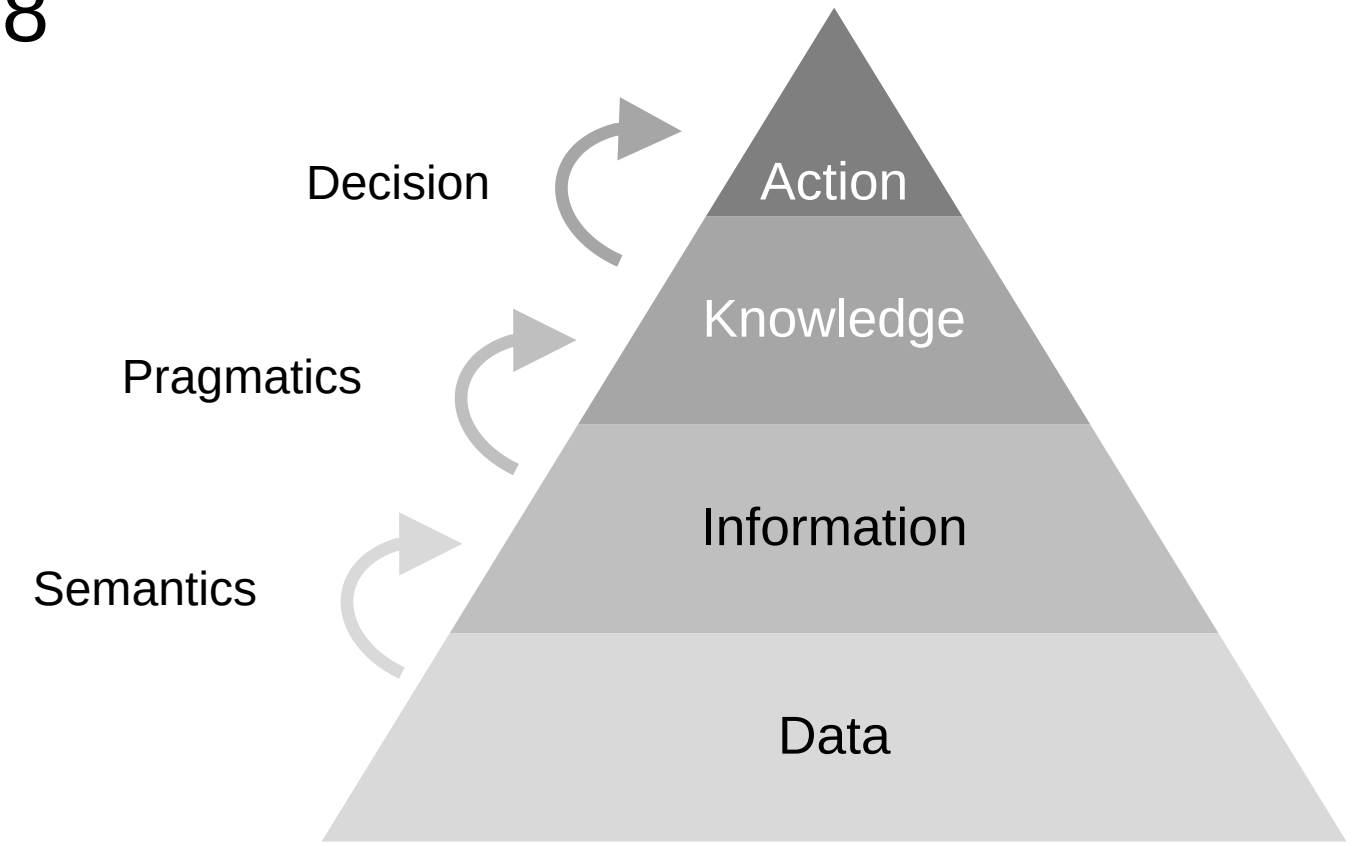
Last lesson

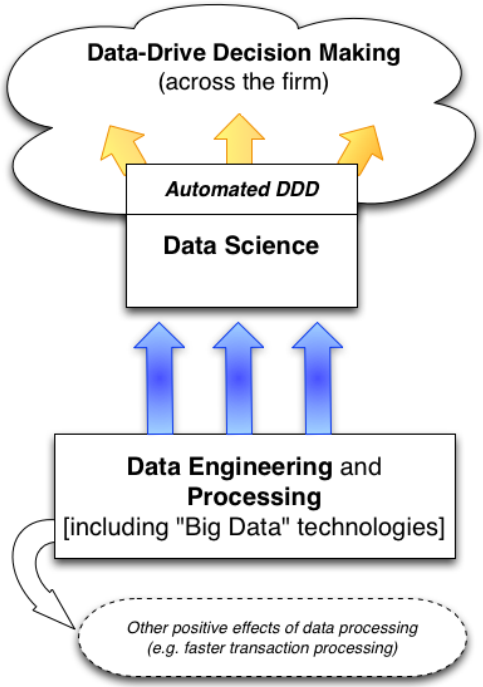
Some data about you…

Data-driven Decision Making (DDD) is a *"practice of basing decisions on the analysis of data rather than purely on intuition."*



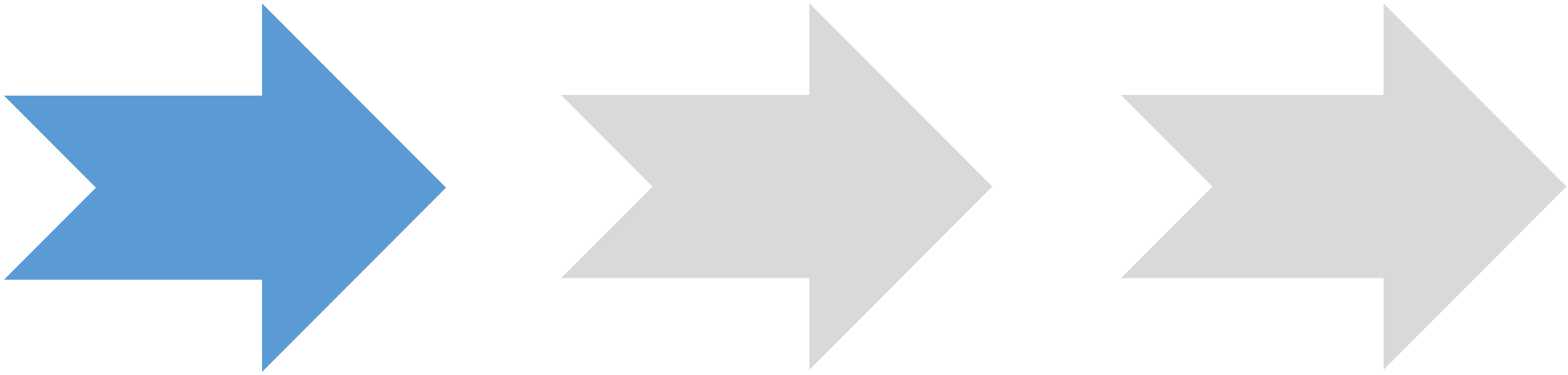
Transforming data to information to knowledge to action





Ref.

Agenda



(1) An illustrative example, operational vs. analytical issues

(2) Basic outline of data warehouses (DWHs)

Distinguishing operational databases from DWHs

Architecture of a DWH system

Data within the DWH

(3) Online Analytical Processing (OLAP)

Different query methods

Properties of OLAP

Common OLAP functionality

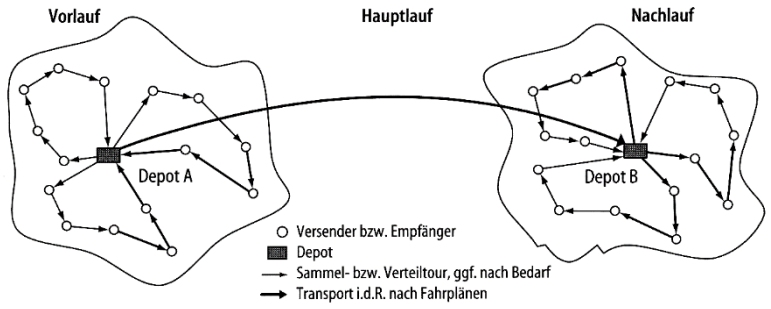
Ref.

An illustrative example

Logistics service providers

Logistics service providers transform freight temporally and spatially (transportation services)

Services offered by logistics service providers differ in
- type and weight of goods,
- time of transportation,
- and price.



The 4 “R” of logistics:
the *right product* at the *right time* and at the *right place* in the *right quality*

Courier service: individually attended transportation of small goods. Transport occurs in the shortest possible time with high reliability

Express service: transportation of goods without weight and size limit

Parcel service: transportation of goods that are limited in volume

Integration of all processes along the supply chain in order to control transportation

Status of entities (i.e., goods for transportation) is very important

Data collection by Enterprise Resource Planning (ERP) systems

Ref.

Img. 天然ガス 2008 | Wikimedia (cc by sa)

An express letter from Germany to the US (1/2)

Freie Universität Berlin

✓	**Luftfrachtbrief: 1421247542 Unterschrieben für von: U RODRIGUEZ**	**Dienstag, November 06, 2012 am 10:37** **Herkunftsgebiet:** HANNOVER - brausnchweig - GERMANY **Zielgebiet:** OMAHA, NE - omaha - USA		**JD013056300600107105**
11	Verlässt DHL-Niederlassung in CINCINNATI HUB - USA	CINCINNATI HUB, OH - USA	05:36	JD013056300600107105
10	Verzollung abgeschlossen in CINCINNATI HUB - USA	CINCINNATI HUB, OH - USA	03:22	JD013056300600107105
9	Sendung sortiert in CINCINNATI HUB - USA	CINCINNATI HUB, OH - USA	03:22	JD013056300600107105
8	Ankunft in der DHL-Niederlassung in CINCINNATI HUB - USA	CINCINNATI HUB, OH - USA	01:23	JD013056300600107105
Donnerstag, November 01, 2012		**Ort**	**Zeit**	**Stücke**
7	Sendung im Transit durch NEW YORK CITY GATEWAY - USA	NEW YORK CITY GATEWAY, NY - USA	10:43	JD013056300600107105
6	Verlässt DHL-Niederlassung in LEIPZIG - GERMANY	LEIPZIG - GERMANY	03:21	JD013056300600107105
5	Sendung sortiert in LEIPZIG - GERMANY	LEIPZIG - GERMANY	00:22	JD013056300600107105
Mittwoch, Oktober 31, 2012		**Ort**	**Zeit**	**Stücke**
4	Ankunft in der DHL-Niederlassung in LEIPZIG - GERMANY	LEIPZIG - GERMANY	23:25	JD013056300600107105
3	Verlässt DHL-Niederlassung in HANNOVER - GERMANY	HANNOVER - GERMANY	20:58	JD013056300600107105
2	Sendung sortiert in HANNOVER - GERMANY	HANNOVER - GERMANY	19:31	JD013056300600107105
1	Sendung abgeholt	HANNOVER - GERMANY	15:07	

Ref.

An express letter from Germany to the US (2/2)

Freie Universität Berlin

✓	**Luftfrachtbrief: 1421247542 Unterschrieben für von: U RODRIGUEZ**	**Dienstag, November 06, 2012 am 10:37** **Herkunftsgebiet:** HANNOVER - brausnchweig - GERMANY **Zielgebiet:** OMAHA, NE - omaha - USA		**JD013056300600107105**
Dienstag, November 06, 2012		**Ort**	**Zeit**	**Stücke**
19	Sendung zugestellt - übernommen von : U RODRIGUEZ	omaha	10:37	JD013056300600107105
18	Sendung in Zustellung	OMAHA, NE - USA	08:17	JD013056300600107105
Montag, November 05, 2012		**Ort**	**Zeit**	**Stücke**
17	Sendung in Zustellung	OMAHA, NE - USA	17:00	JD013056300600107105
16	Erfolgloser Zustellversuch, Empfänger nicht zu Hause	OMAHA, NE - USA	08:42	JD013056300600107105
Freitag, November 02, 2012		**Ort**	**Zeit**	**Stücke**
15	Sendung zur Aufbewahrung in DHL-Niederlassung	OMAHA, NE - USA	20:37	JD013056300600107105
14	Sendung in Zustellung	OMAHA, NE - USA	17:38	JD013056300600107105
13	Erfolgloser Zustellversuch, Empfänger nicht zu Hause	OMAHA, NE - USA	16:21	JD013056300600107105
12	Ankunft in der DHL-Zustellbasis in OMAHA - USA	OMAHA, NE - USA	08:32	JD013056300600107105

DHL: Processing of approx. 70,000 shipments / night per transshipment facility (35 transshipment facilities in Germany)

Ref.

Operational Issues vs. Analytical Issues

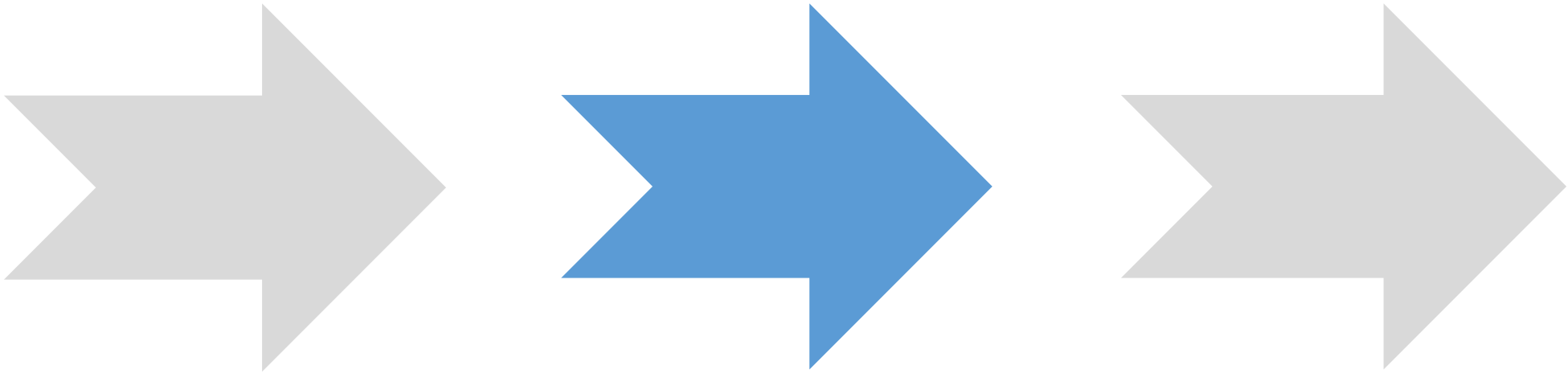
Differences?

- Wo befindet sich die Sendung XYZ zurzeit?
- An welchen Stationen wurde sie umgeschlagen?
- Ist eine Beschädigung der Sendung eingetreten?
- Wann wird sie voraussichtlich ankommen?
- Wer hat den Erhalt der Sendung quittiert?

- Zu wie viel Prozent ist die Sortieranlage in Cincinnati ausgelastet?
- Welcher Verteilung folgt die Verweildauer für den Hub Leipzig?
- Lohnt ein Direktflug Leipzig – Cincinnati bzw. Hannover – New York?
- Wieviel Prozent der Sendungen werden termingerecht ausgeliefert?
- Welche Kosten verursacht die Zustellung gegenüber dem Transport?

Ref.

Agenda



(1) An illustrative example

(2) Basic outline of data warehouses (DWHs)

Distinguishing operational databases from DWHs

Architecture of a DWH system

Data within the DWH

(3) Online Analytical Processing (OLAP)

Different query methods

Properties of OLAP

Common OLAP functionality

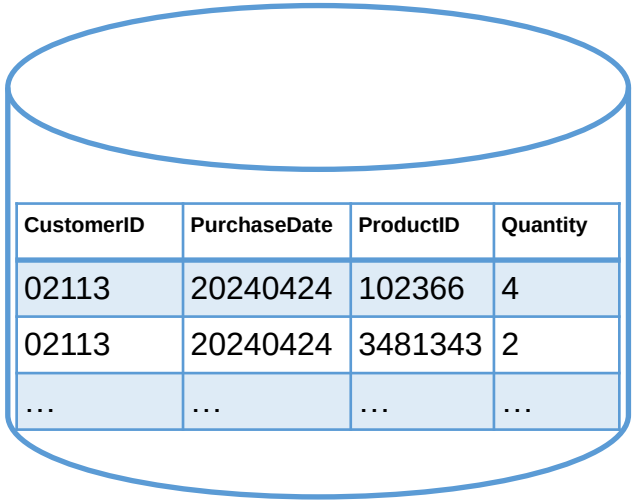
Ref.

Operational databases

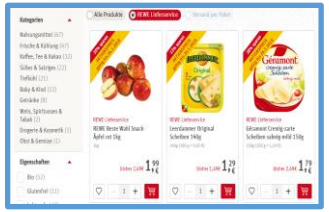
A database for the day-to-day business

Operational databases…

- primarily *support the day-to-day* business (“real-time databases“)
- record operative business transactions
- aim at storing transactional details with *full integrity* and *without redundancy*



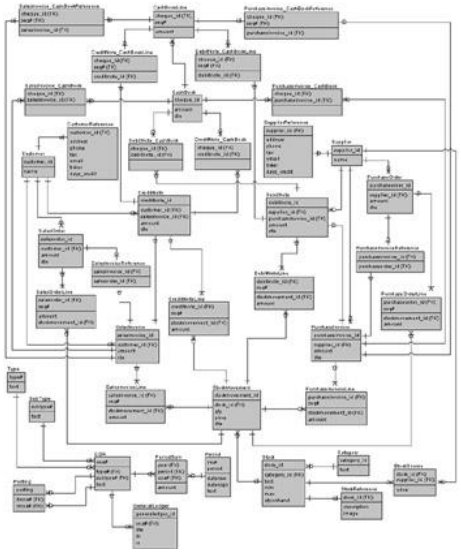
CustomerID	PurchaseDate	ProductID	Quantity
02113	20240424	102366	4
02113	20240424	3481343	2
…	…	…	…



For this reason, data is regularly stored in a **complex** way:

- many details
- many updates
- Highly *normalized (1NF, 2NF, 3NF)*

-> Operational databases are seldomly very user-friendly



Excursus: Normalization

1. Normalform (1NF)
Herstellung der Grundform einer Relation (atomare Attribute)

2. Normalform (2NF)
Auflösung von Teil-abhängigkeiten (abh. Von Primärschlüssel)

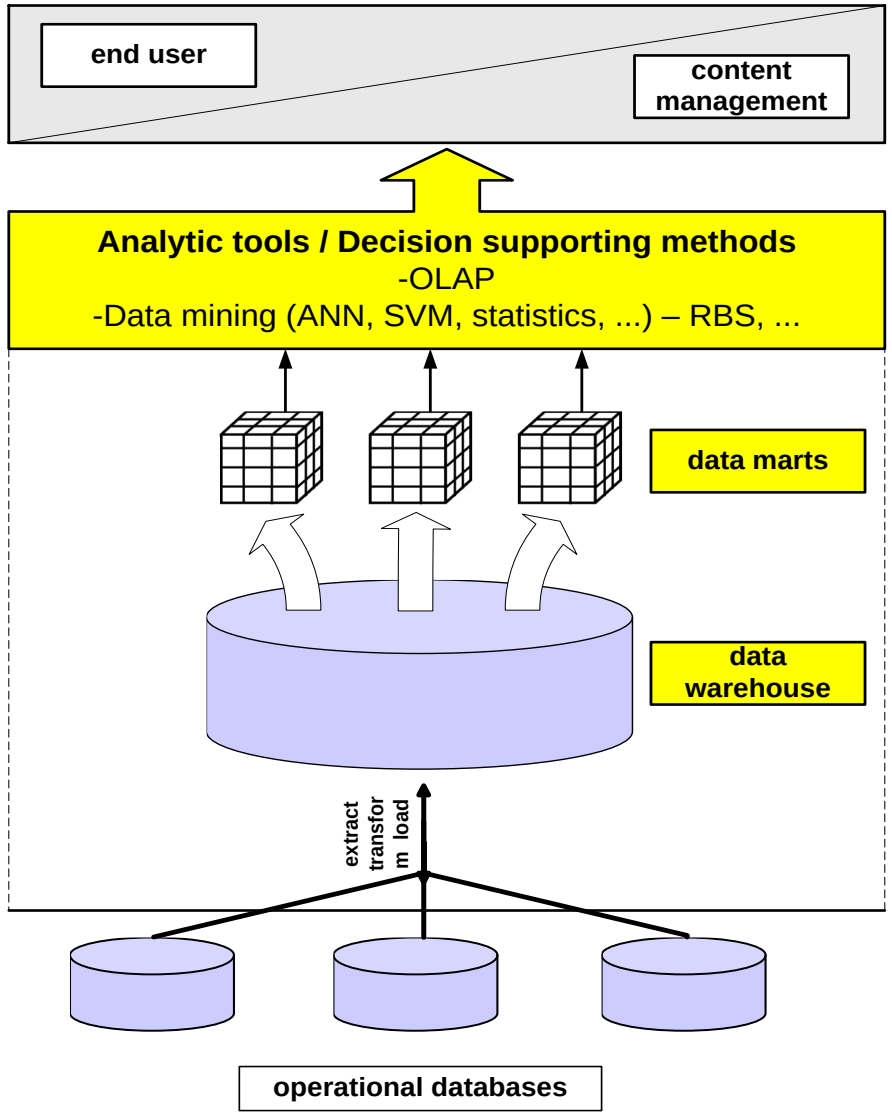
3. Normalform (3NF)
Auflösung transitiver Abhängigkeiten (nicht-Schlüsselattribute sind unabhängig)

Ref.

Img. Powell (2003)

Operational databases and Data Warehouses

Reasons for data warehousing



Ref.

Data Warehouses

- collect data from operational databases
- accumulate **historical data**
- provide the basis for Business Intelligence applications

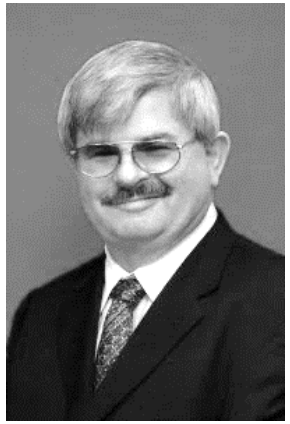
Data warehousing has become a **strategic goal** of many companies

1998: 90% of the 2000 biggest companies worldwide were already developing data warehouses

Reasons for the development of data warehouses:

- Integration of many different data sources into one database
- Creating a better basis for data mining tools
- Controlling the “information flooding” by structuring and aggregating of operative data
- Analyzing tools can be applied to complex questions For example?

Definition of Data Warehouses



W.H. Inmon was the first to provide a definition:

> *"A Data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process."*
> (Inmon, 1992)

Detailed look at the keywords of the definition:

Subject-oriented
data gets organized corresponding to the business context of the particular company

Integrated
data from many different internal and external sources is loaded into the DWH

Time-variant
time series analysis is possible by the means of DWHs

Non-volatile
data is stored persistently and read-only access is provided

Ref.

Operational databases vs. DWHs

A comparison

Another (more concrete) definition of data warehouses:

"A data warehouse is a decision supporting database (analytic database), which is separated from the operational databases, and which is primarily used for decision support in a company.
A data warehouse is always modeled in a multidimensional way and is used for the long-term storage of historic, cleaned, validated, synthesized, operative data from internal and external sources."
(A. Kurz, 1998)

Exercise

Properties	Operational Database	Data Warehouse
Operational data	•	
Detailed data		
Complex data model		
Strategic data		
Processing requires many queries		
Interface end-user oriented		•

Full amount of Data covered

Derived data
(e.g. aggregates -> redundant data)

Many Updates

Ad hoc queries

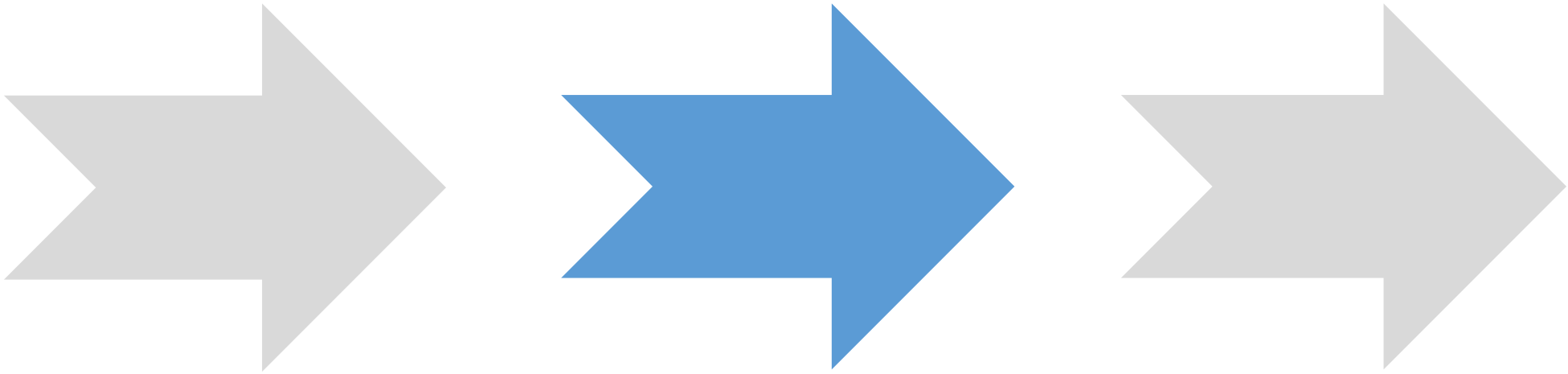
Historical data

Non-redundant data

5 Min.

Ref. Kimball/Ross (2011)

Agenda



(1) An illustrative example

(2) Basic outline of data warehouses (DWHs)

Distinguishing operational databases from DWHs

Architecture of a DWH system

Data within the DWH

(3) Online Analytical Processing (OLAP)

Different query methods

Properties of OLAP

Common OLAP functionality

Ref.

Basic steps concerning DWHs

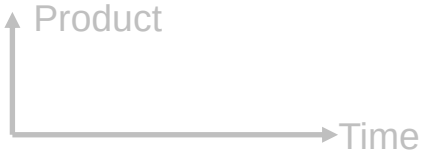
1. Select appropriate attributes from operational databases
2. Add selected data from external sources
3. Transform and load data
4. Store loaded data subject to dimensions
5. (Administrational operations similar to those known from operational databases)
6. Query and analyze based on DWH (reports, OLAP)

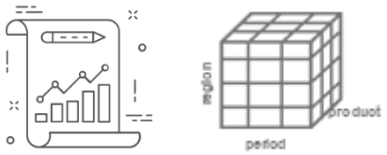
KundenNr	ArtNr	Anzahl	Datum
02113	3481343	2	24.04.2024
…	…	…	…

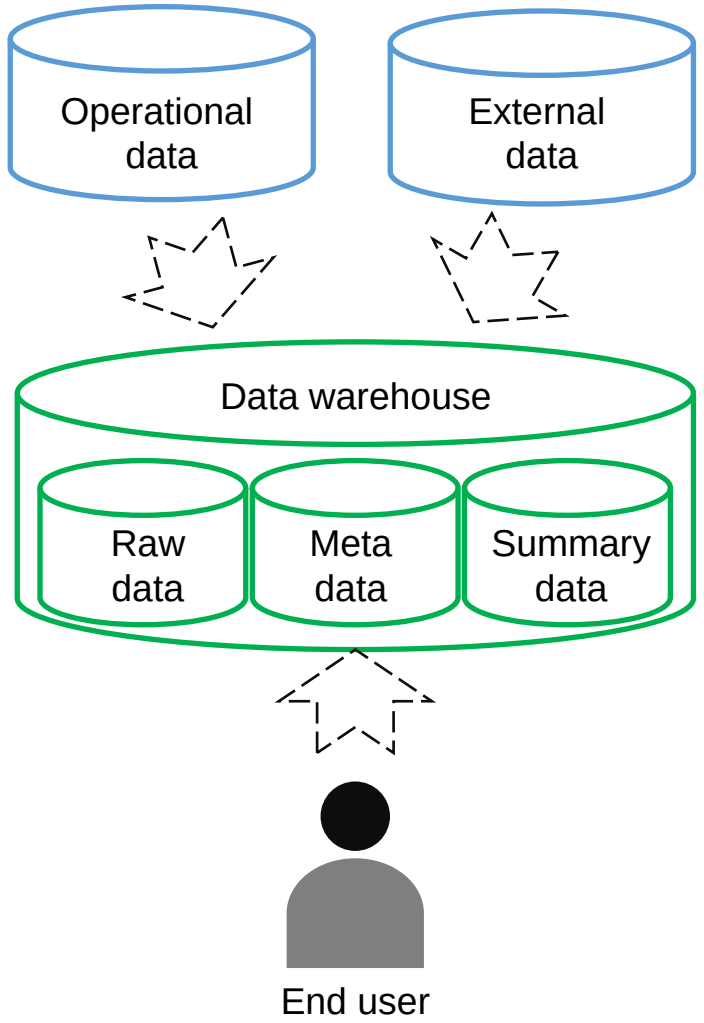
For example?

C_ID	P_ID	Date	..
…	…	…	…

CustomerID	PurchaseDate	ProductID	Quantity
02113	04-24-2024	102366	4
…	…	…	…







Ref.

Key elements of data warehouse systems (1/2)

Data Marts

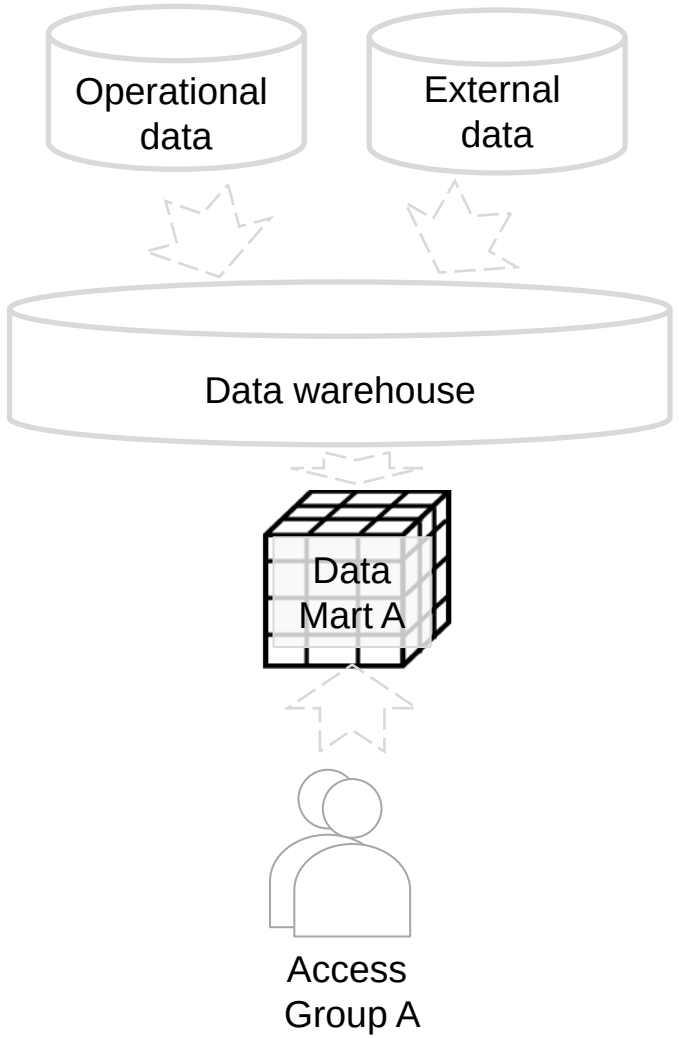
The core of a data warehouse system can be built out of several components:

Data marts

(small) analytic databases **for a special group of people** (e.g. department, or workgroup)

- administration by **local** departments instead of Central IT
- coordination with other analytic databases
- development less complex than creation of larger DWHs
- based on **specialized data models**, which are quite easy to understand and which provide efficient access to data
- end users can be easily involved in the development of the single data marts

- load data from other DWHs or from operational databases
- Distribution of analytic data among data marts is a difficult task
  - Build homogenous user groups
  - assimilate data model to a functional area



Ref. e.g., Sen & Sinha (2005), March & Hevner (2007)

Key elements of data warehouse systems (2/2)

Data Marts, Central Data Warehouse, Enterprise Data Warehouse

Central data warehouse

analytical database provides data transformed and coordinated to local data marts

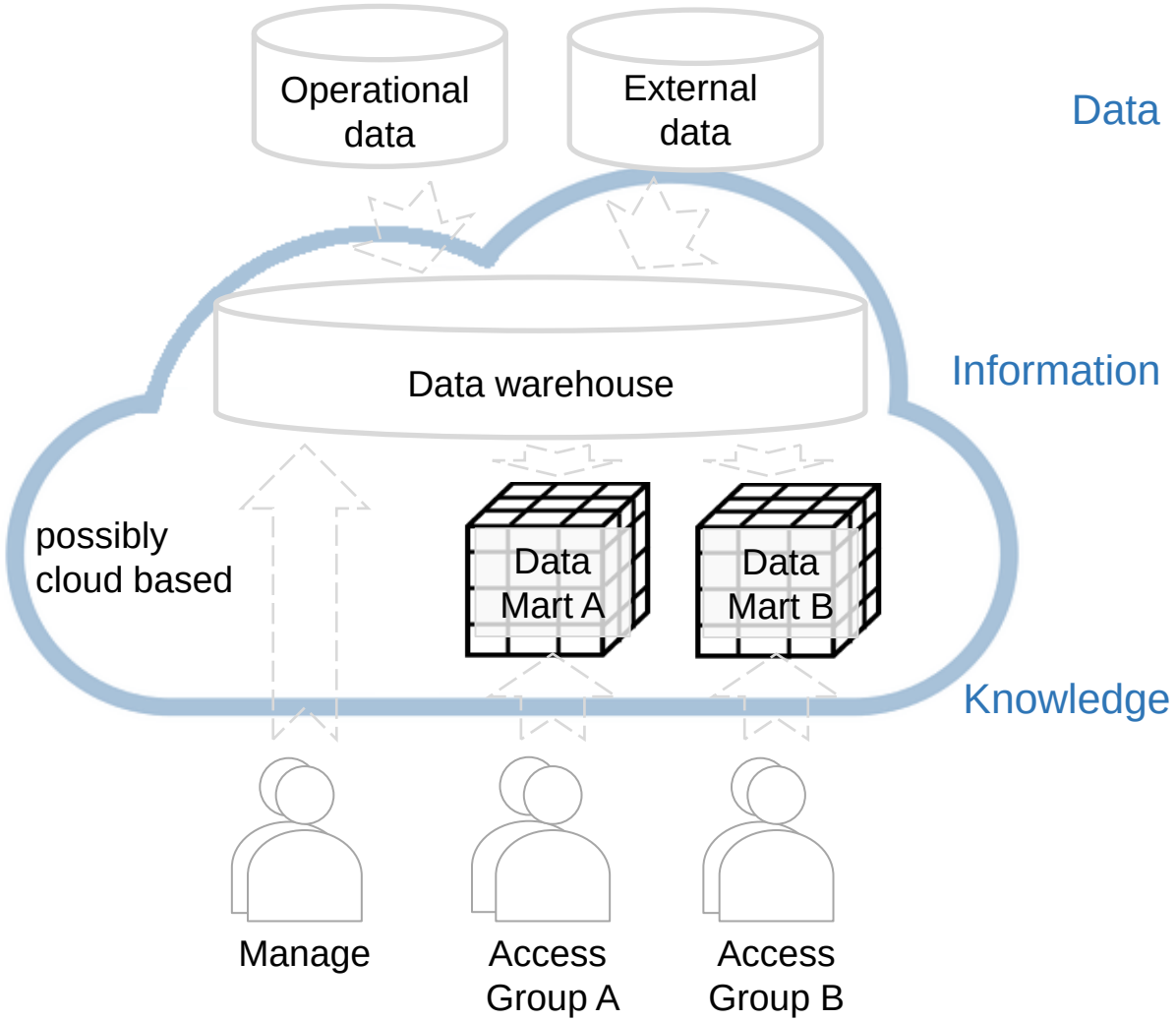
not necessarily providing information for the whole company

Enterprise data warehouse (EDWH)

central data warehouse providing data and information for the whole company

https://aws.amazon.com/de/redshift/
https://www.cloudera.com/products/enterprise-data-hub.html
https://cloud.google.com/bigquery/
https://www.vertica.com/overview/
https://www.ibm.com/cloud/db2-warehouse-on-cloud
https://azure.microsoft.com/de-de/services/sql-data-warehouse/
https://cloud.oracle.com/de_DE/database
https://www.sap.com/germany/products/bw4hana-data-warehousing.html
https://www.snowflake.com/product/
https://www.teradata.de/Products/Cloud
List of April 2024

Typical Issues?



Ref. e.g., Sen & Sinha (2005), March & Hevner (2007)

Data Warehouse vs. Data Lake

Search by yourself

e.g. https://blogs.oracle.com/bigdata/data-lake-database-data-warehouse-difference

Todo for next Friday

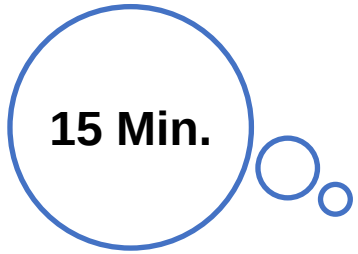
Type of data?

Task?

Users?

Data processing?

Data granularity?



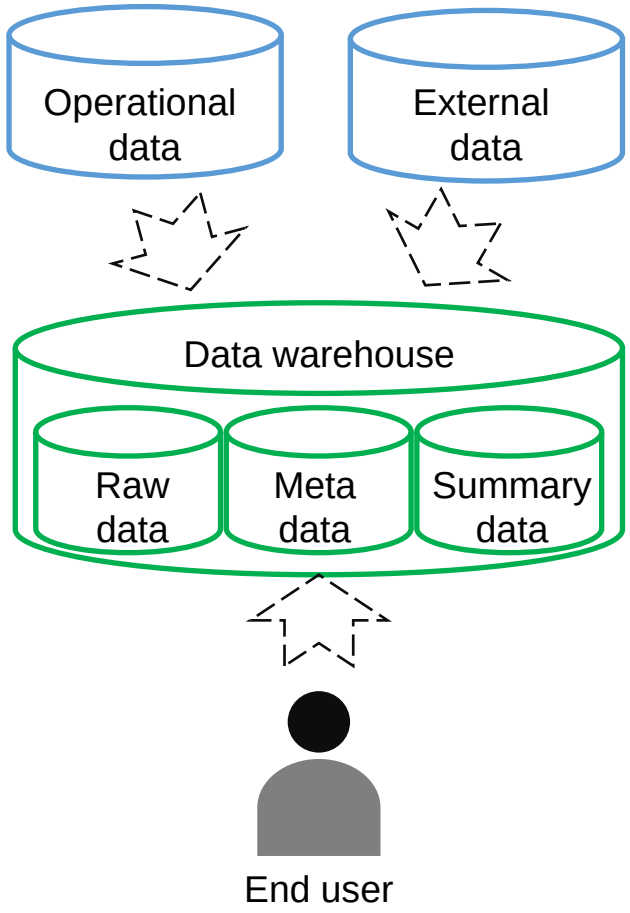
Ref.

Architectures (1/2)

Freie Universität Berlin

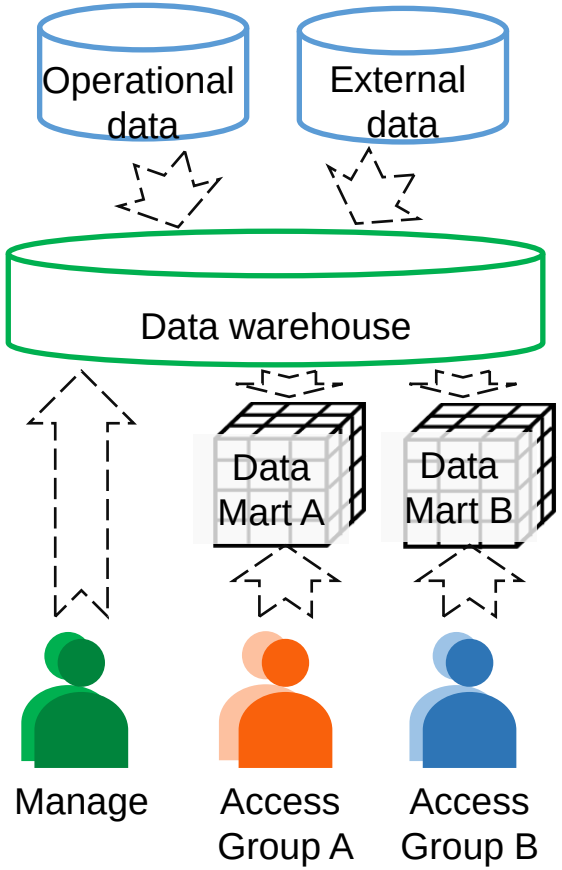
Centralized architecture

All the analytic data is bundled on **one platform**



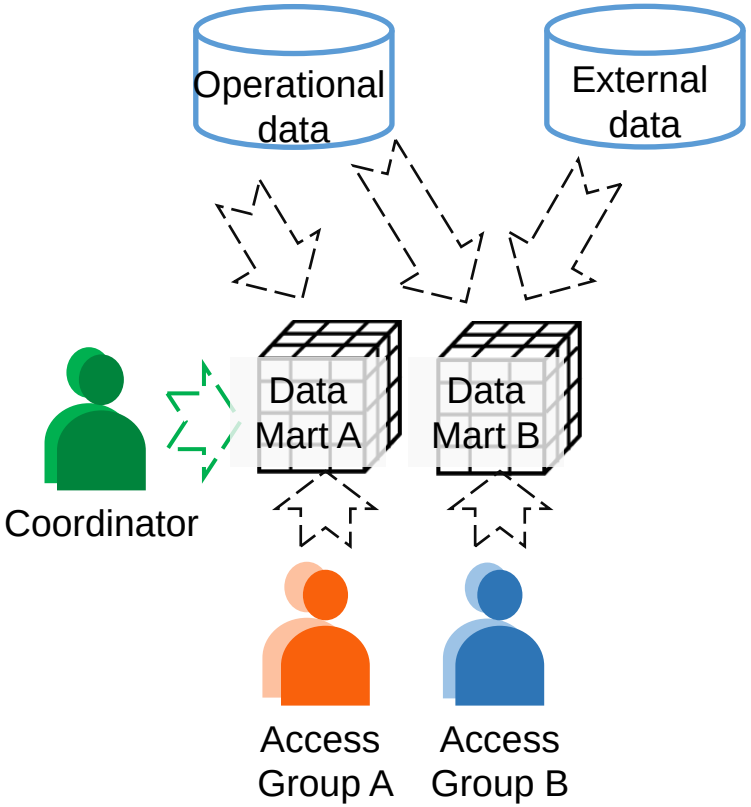
Hierarchical architecture

Local data marts become **coordinated** by an enterprise data warehouse (EDWH)



Enterprise data mart architecture

Central DWH functionally replaced by **coordinated data marts**



Ref.

Architectures (2/2)

Centralized architecture

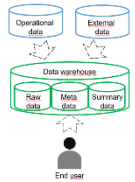
All the analytic data is bundled on **one platform**

Advantages:

- Less redundancy
- Savings concerning hardware (For On-Premise-DWH)

Disadvantages:

- Limited possibilities for modularization
- Development too complex for many companies
- Degree of user friendliness and efficiency is only for small companies sufficient



Ref.

Hierarchical architecture

Local data marts become **coordinated** by an enterprise data warehouse (EDWH)

The EDWH extracts, integrates and distributes data

The data marts are…

- … used for querying and for analyzing data
- … specialized on a special functional field within the company

Coordination of attributes necessary
(bijective relationship between attribute and description:
no homonyms, no synonyms, no aliases)

Example?
Individual = Customer = Person = Employee?
Customer = Client = Consumer ?

Enterprise data mart architecture

Central DWH functionally replaced by **coordinated data marts**

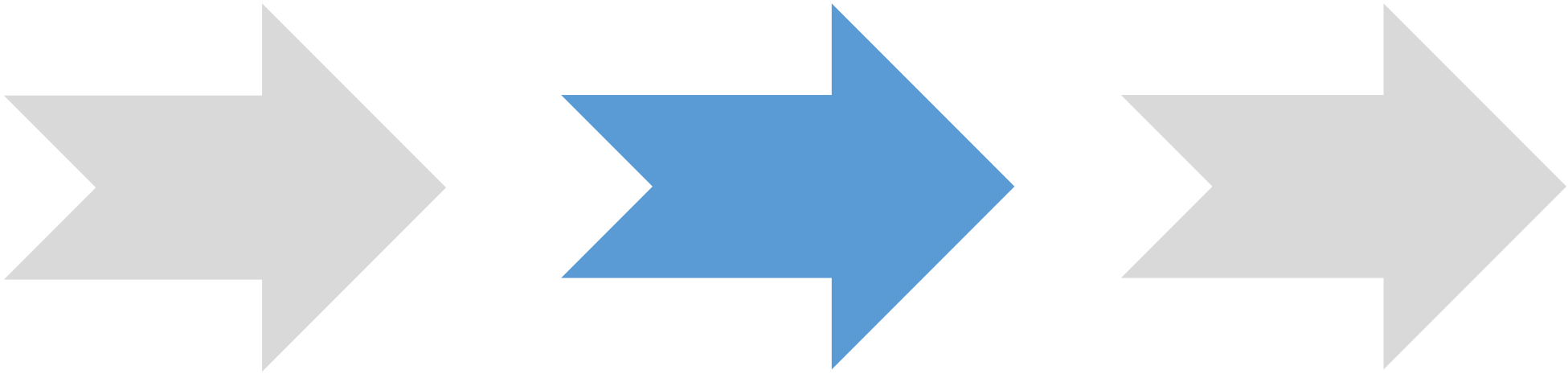
Often based on a distributed database system

Extraordinary focus on the maximization of *intra*modularity & minimization of *inter*modularity of Data Marts

High efforts for coordination required

- Load and access coordination
- Coordination of data model (metadata)

Agenda



(1) An illustrative example

(2) Basic outline of data warehouses (DWHs)

Distinguishing operational databases from DWHs

Architecture of a DWH system

Data within the DWH

(3) Online Analytical Processing (OLAP)

Different query methods

Properties of OLAP

Common OLAP functionality

Ref.

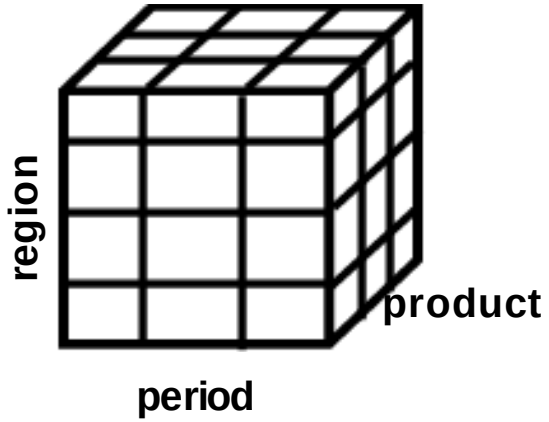
Multidimensional data

Analytical data are represented by multidimensional data models

Modeling user-friendly and close to business

The hypercube data structure is based on **business measures (“facts”)** and **dimensions**

	business measure (or fact)	dimension
purpose	analysis of success subject to several dimensions	selection, aggregation and navigation of facts
examples	quantity turnover, monetary turnover	product, region, period
synonyms	fact, performance measure, key business measure	constraint
datatype	numeric and continuous	symbolic and discrete
data volume	Large (about 70% of the DWH)	small
key	primary key consists of foreign keys of the dimensions	primary keys



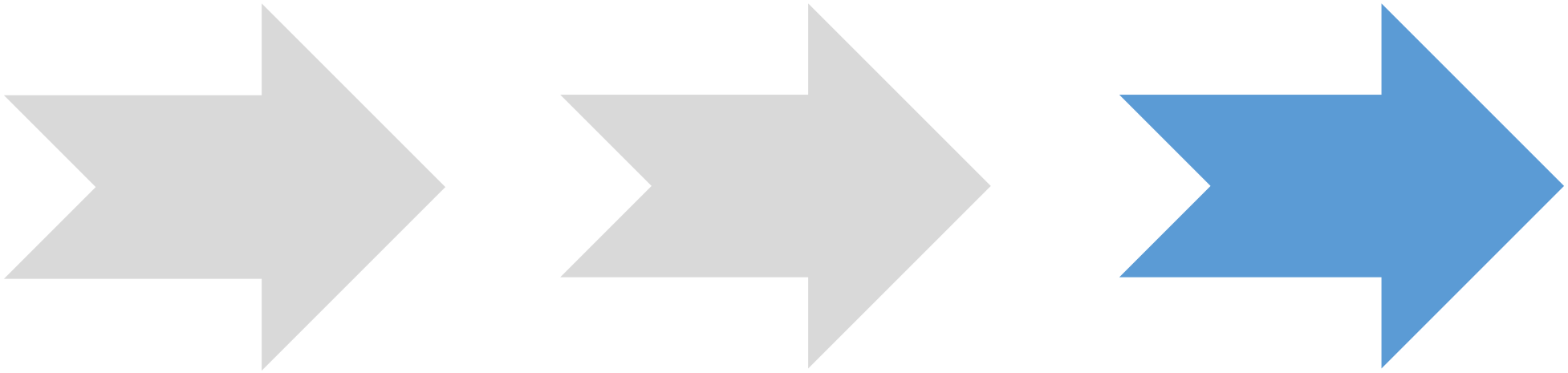
Kahoot-Fragen
www.kahoot.it
(über Smartphone oder Laptop)
PIN folgt

Distinguish facts and dimensions!

(Diese Folie ist nach der Vorlesung mit Lösungen verfügbar)

Ref.

Agenda



(1) An illustrative example

(2) Basic outline of data warehouses (DWHs)

Distinguishing operational databases from DWHs

Architecture of a DWH system

Data within the DWH

(3) Online Analytical Processing (OLAP)

Different query methods

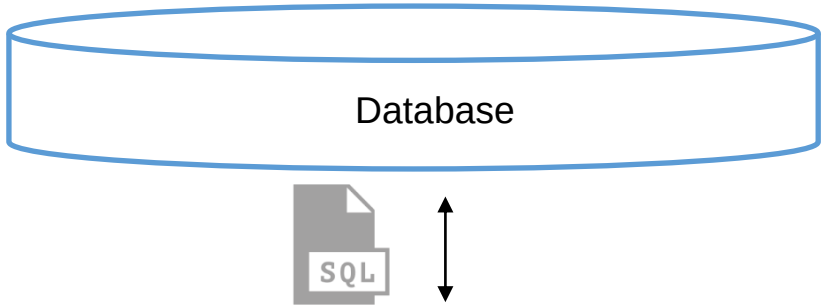
Properties of OLAP

Common OLAP functionality

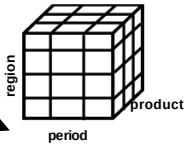
Ref.

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

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
```

Ref.

Query languages

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

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

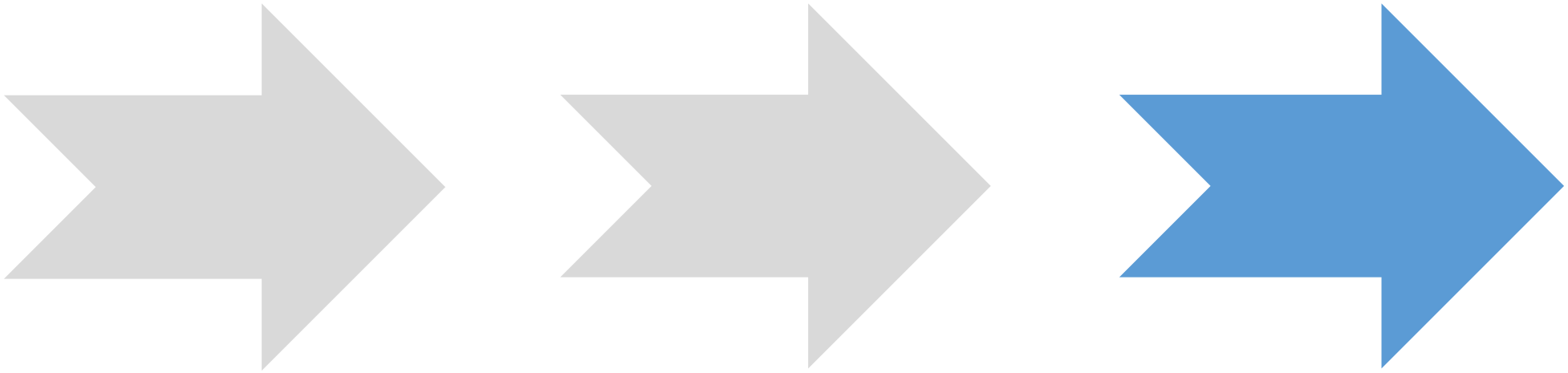
OLAP

- flexible ad-hoc querying
- possible without expertise

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;
```

Agenda



(1) An illustrative example

(2) Basic outline of data warehouses (DWHs)

Distinguishing operational databases from DWHs

Architecture of a DWH system

Data within the DWH

(3) Online Analytical Processing (OLAP)

Different query methods

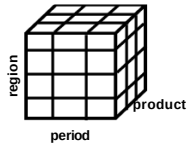
Properties of OLAP

Common OLAP functionality

Ref.

Online Analytical Processing (OLAP)

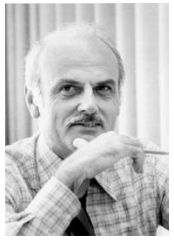
Let's focus on end users, and their access to data marts by OLAP systems



Freie Universität Berlin

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	

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

OLAP functions

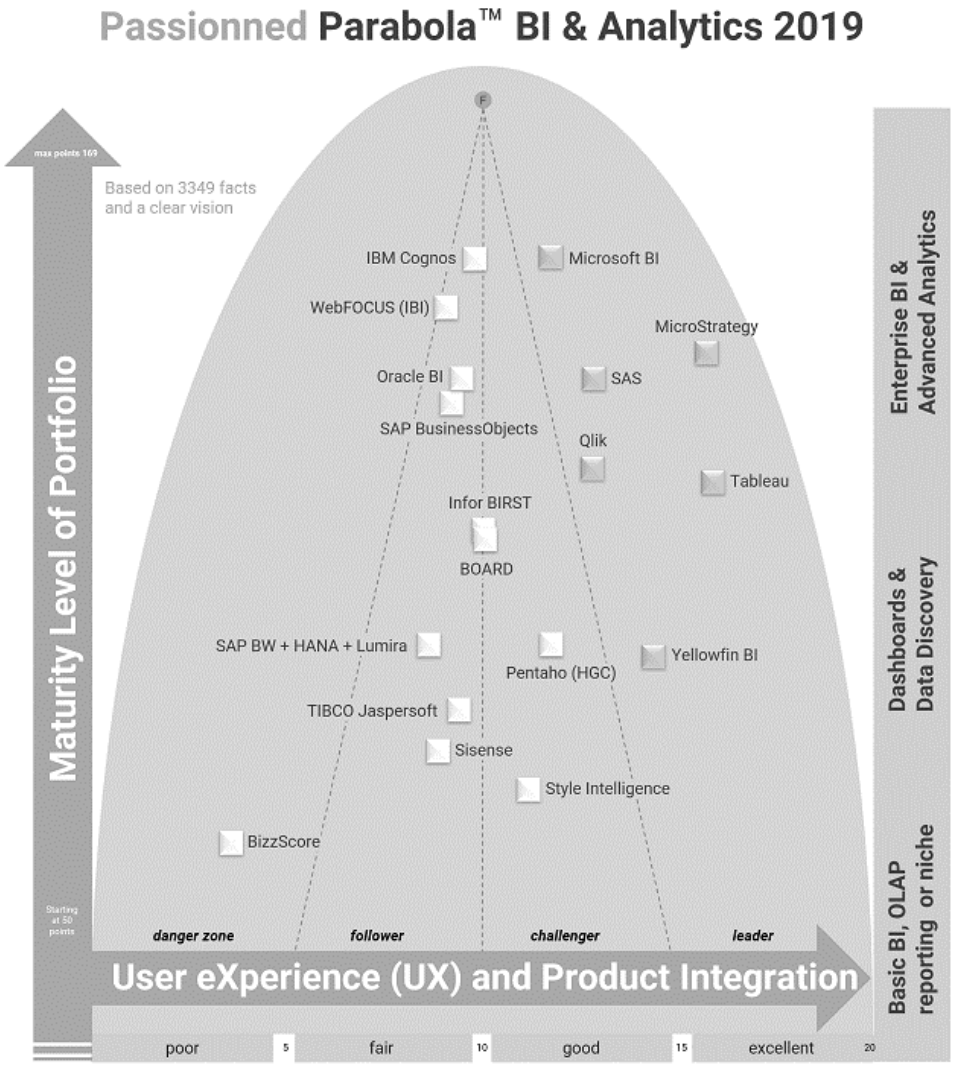
OLAP tools provide a number of standard features

- 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/

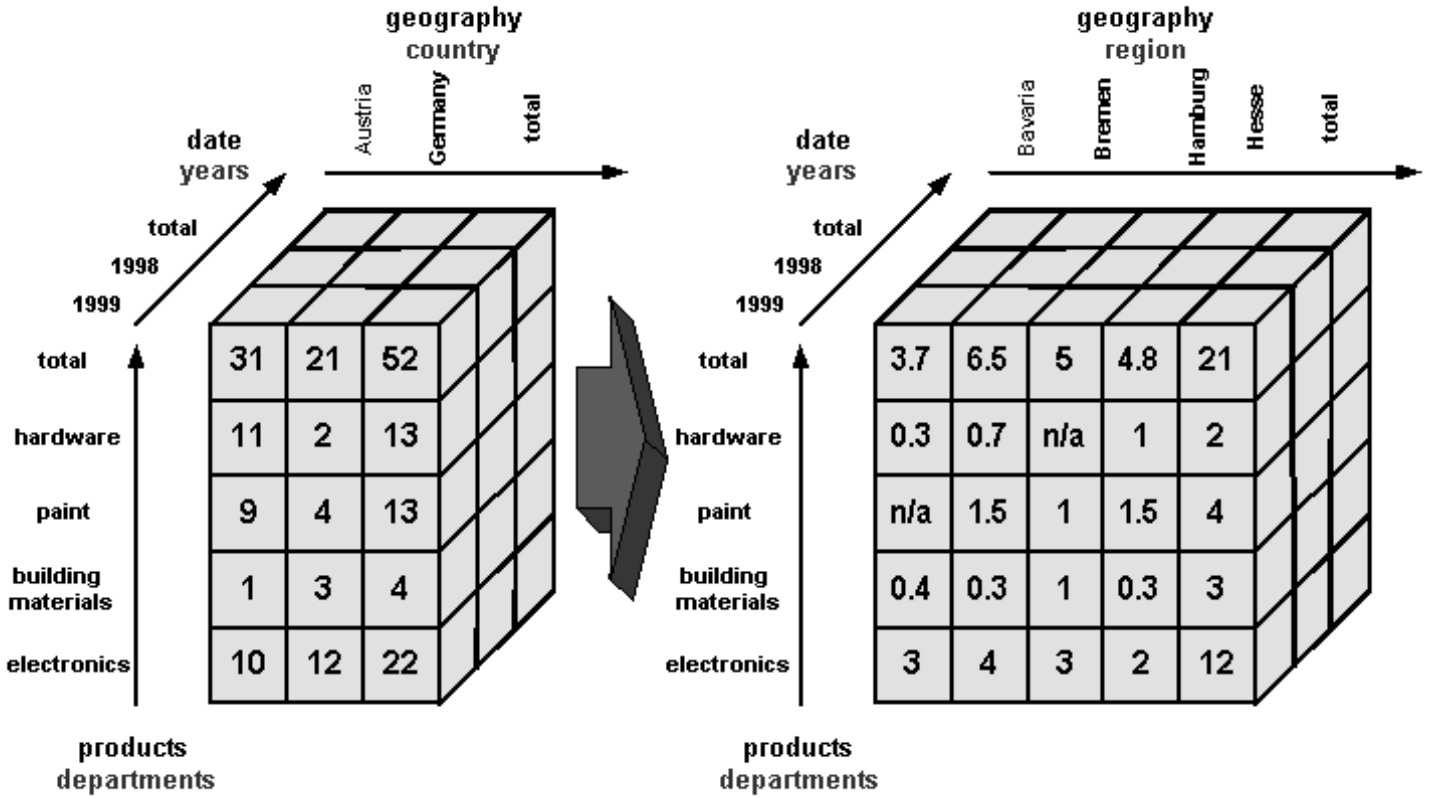
See https://www.passionned.com/bi/#list-business-intelligence-tools for an up-to-date list with detailed information about BI Tools, April 2024

Ref.

Drilling down

More details for specific dimensions

"Show the regions of Germany in detail."

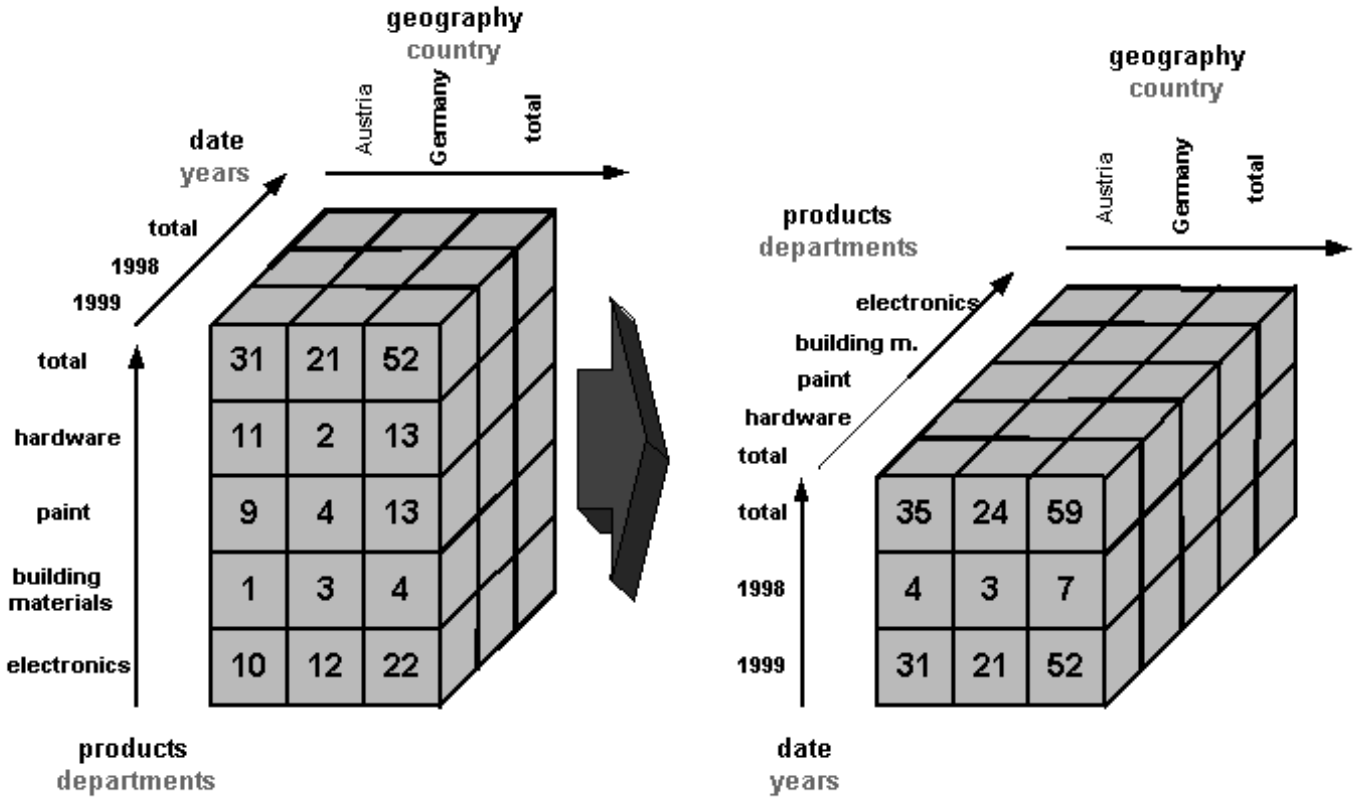


Ref.

Pivoting

Rotate the cube

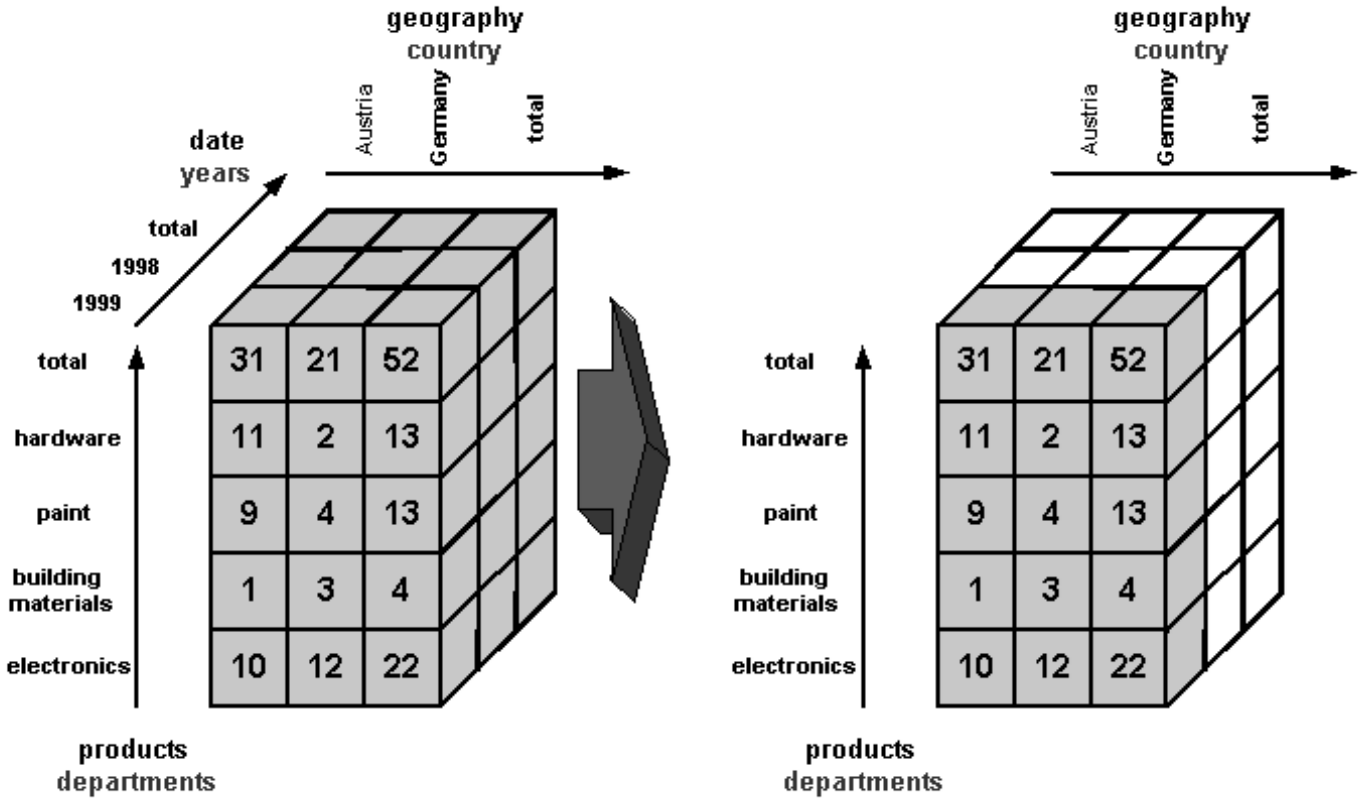
“Show year by country instead of product by country”



Ref.

Slicing

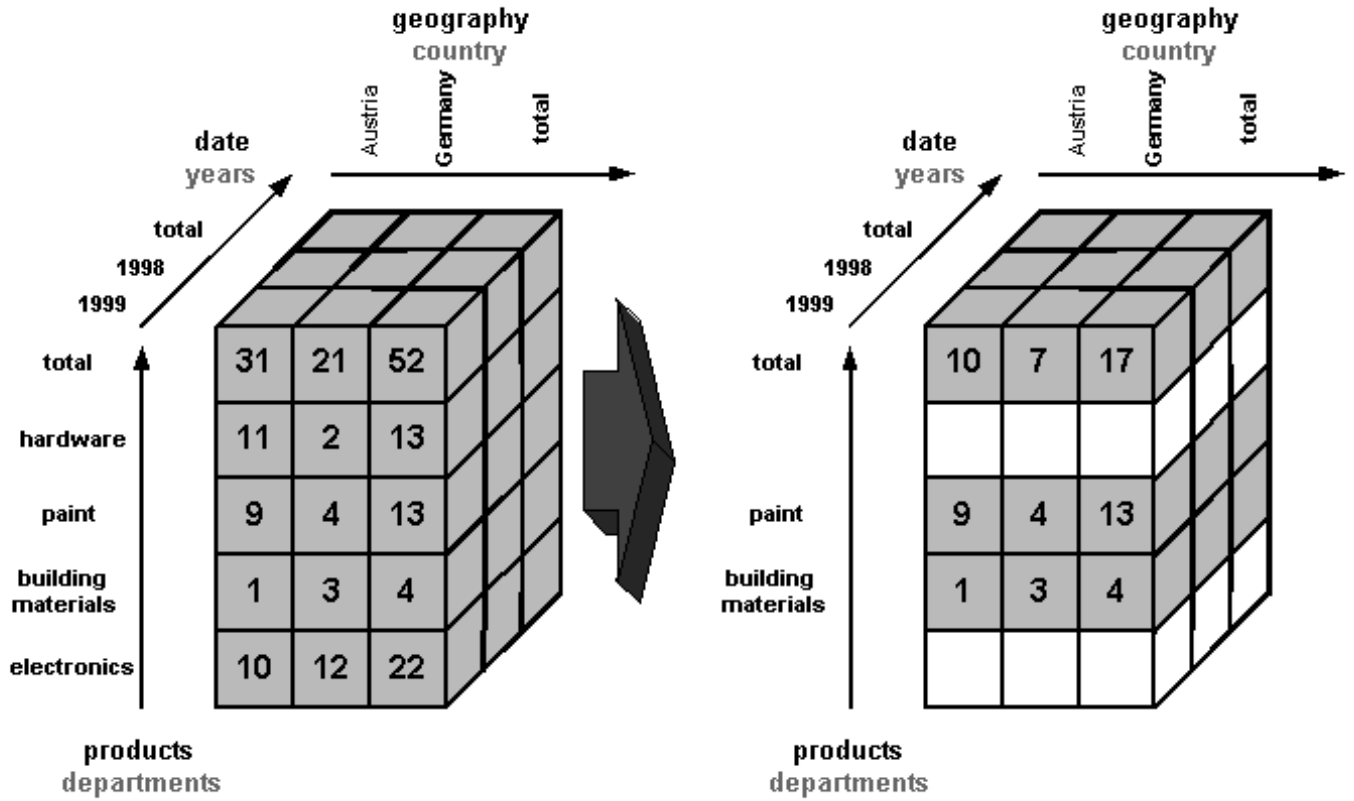
“Show only the values for 1999.”



Ref.

Dicing

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



Ref.

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.

Fragen?

- ✓ An illustrative example
- ✓ Basic outline of data warehouses (DWHs)
  - ✓ Distinguishing operational databases from DWHs
  - ✓ Architecture of a DWH system
  - ✓ Data within the DWH
- ✓ Online Analytical Processing (OLAP)
  - ✓ Different query methods
  - ✓ Properties of OLAP
  - ✓ Common OLAP functionality

Todos for next Friday

1. Data Warehouse vs. Data Lake: What is the difference?
   (Slide 19)

2. Read the short article about recent developments in data warehousing
   “Paradigmenwechsel: Data Warehouses für die Cloud“
   (from iX 5/2020 - Magazin für professionelle Informationstechnik)
   Kursmaterial > Readings/Übungen

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

Ref.

Recommended reading

Lusti, M. (2002): Data Warehousing und Data Mining (esp. Chapter 5)

Kurz, A. (1999): Data Warehousing (esp. Chapters 1 and 4)

Inmon, W.H. (1996): Building the Data Warehouse
(esp. Chapters 1 and 2)

http://www.tdwi.org

Ref.

Bibliography

- Chamoni, P., & Gluchowski, P. (2000). On-Line Analytical Processing (OLAP). *Das Data-Warehouse-Konzept. Architektur–Datenmodelle–Anwendungen. Wiesbaden*.
- Codd, E. F., Codd, S. B., & Salley, C. T. (1993). *Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate*. *Codd and Date*, *32*.
- Kimball, Ralph, and Margy Ross. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.
- March, S. T., & Hevner, A. R. (2007). Integrated decision support systems: A data warehousing perspective. *Decision Support Systems*, *43*(3), 1031-1043.
- Pendse, N., & Creeth, R. (1995). Succeeding with On-Line Analytical Processing. *The OLAP-Report*, *1*.
- Powell, Gavin JT. *Oracle high performance tuning for 9i and 10g*. Digital Press, 2003.
- Sen, A., & Sinha, A. P. (2005). A comparison of data warehousing methodologies. *Communications of the ACM*, *48*(3), 79-84.
- Stucke, Maurice E., and Allen P. Grunes. "Big Data and Competition Policy." (2016).

Ref.