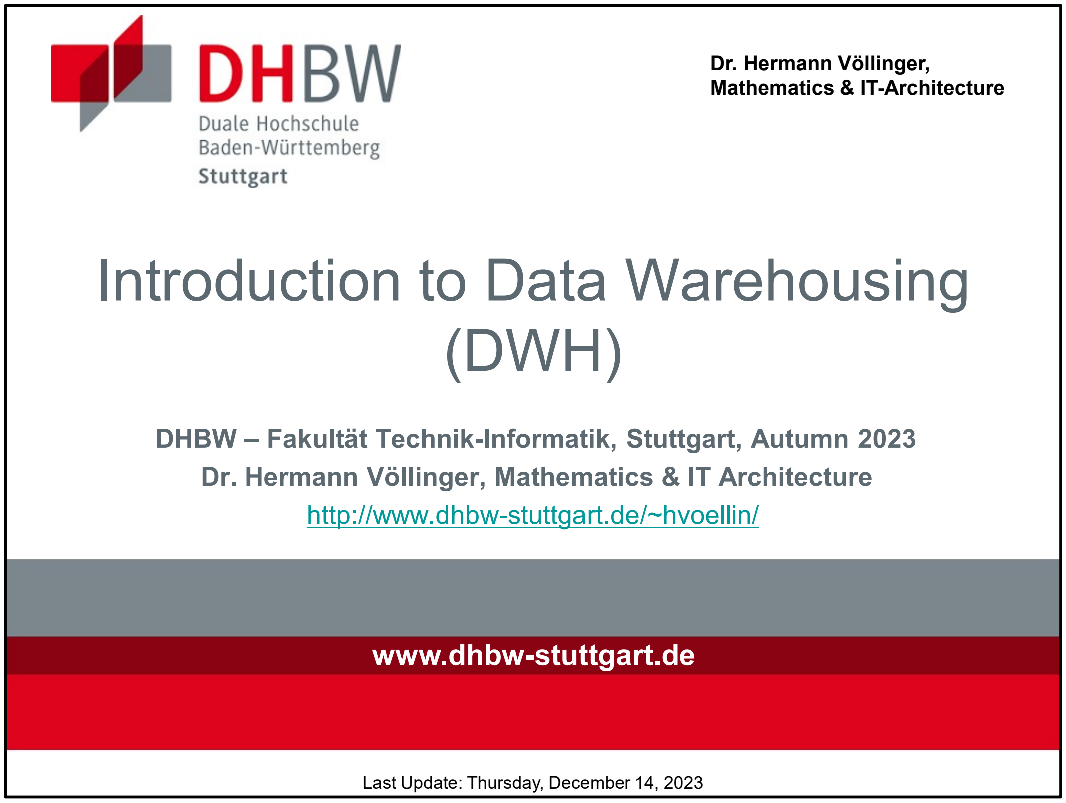


Duale Hochschule Baden-Württemberg Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Introduction to Data Warehousing (DWH)

DHBW – Fakultät Technik-Informatik, Stuttgart, Autumn 2023
Dr. Hermann Völlinger, Mathematics & IT Architecture
http://www.dhbw-stuttgart.de/~hvoellin/

www.dhbw-stuttgart.de

Last Update: Thursday, December 14, 2023

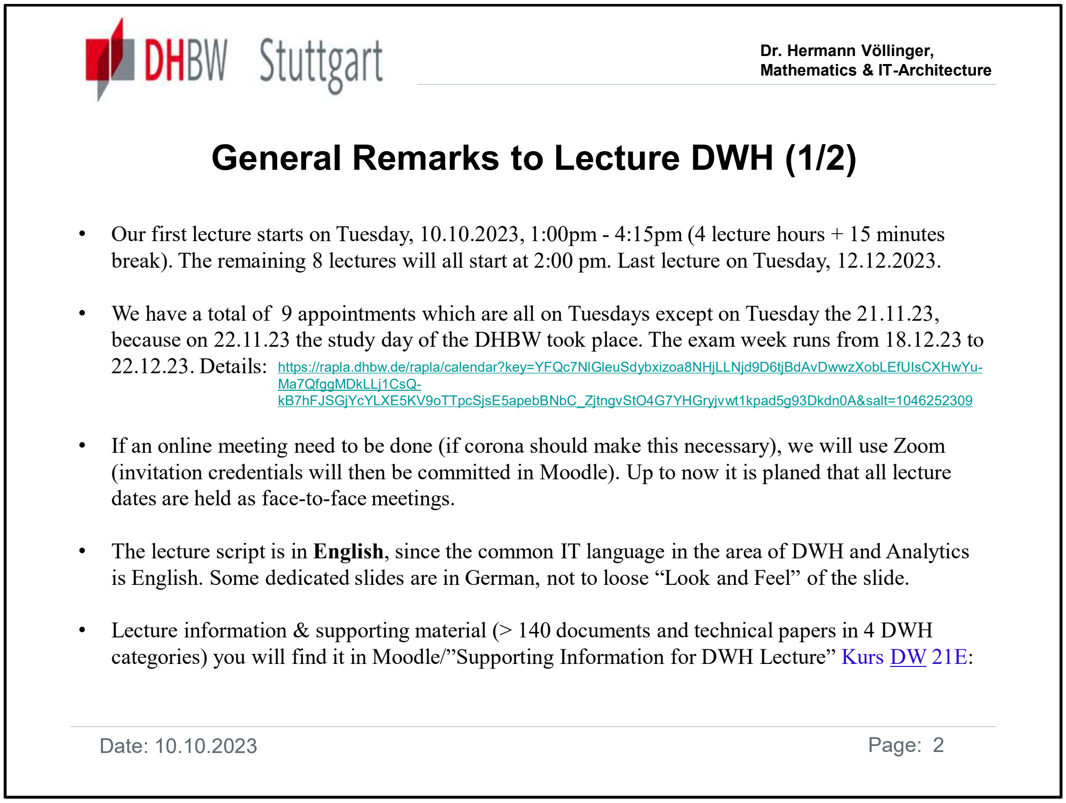
Data-Management and **DWH are the basics for Digitalization and Artificial Intelligence (AI).** As you may know and which may be also **important for your future** - the government of Baden-Württemberg build an **AI Innovation Park:** "….. In order for this seal of quality to gain national and international appeal and become a global success, a world-class artificial intelligence (AI) innovation park is to be realized with the support of the state on the basis of a resolution passed by the state government…" See: https://www.wirtschaft-digital-bw.de/ki-made-in-bw/innovationspark-kuenstliche-intelligenz-baden-wuerttemberg

Examples - Data Usage for Digitalization: „**Digitalisierung im Fussball**":
https://www.unity.de/de/branchen/sport-und-arenamarkt/digitalisierung-im-fussball

See also an actual list of **AI. i.e. Machine Learning (ML) applications**, see a summary paper of the Machine-Learning User-Group Stuttgart (MLUGS)**: "Machine Learning Orientation";** by *Uwe Sterr; 2021-01-19:*
http://machinelearningintro.uwesterr.de/whatML.html#whatML

To prepare the DWH for such future applications see also the TDWI Webinar**:** "**So machen Sie Ihr Data Warehouse zukunftsfähig**": https://youtu.be/DAsn9scb0ts

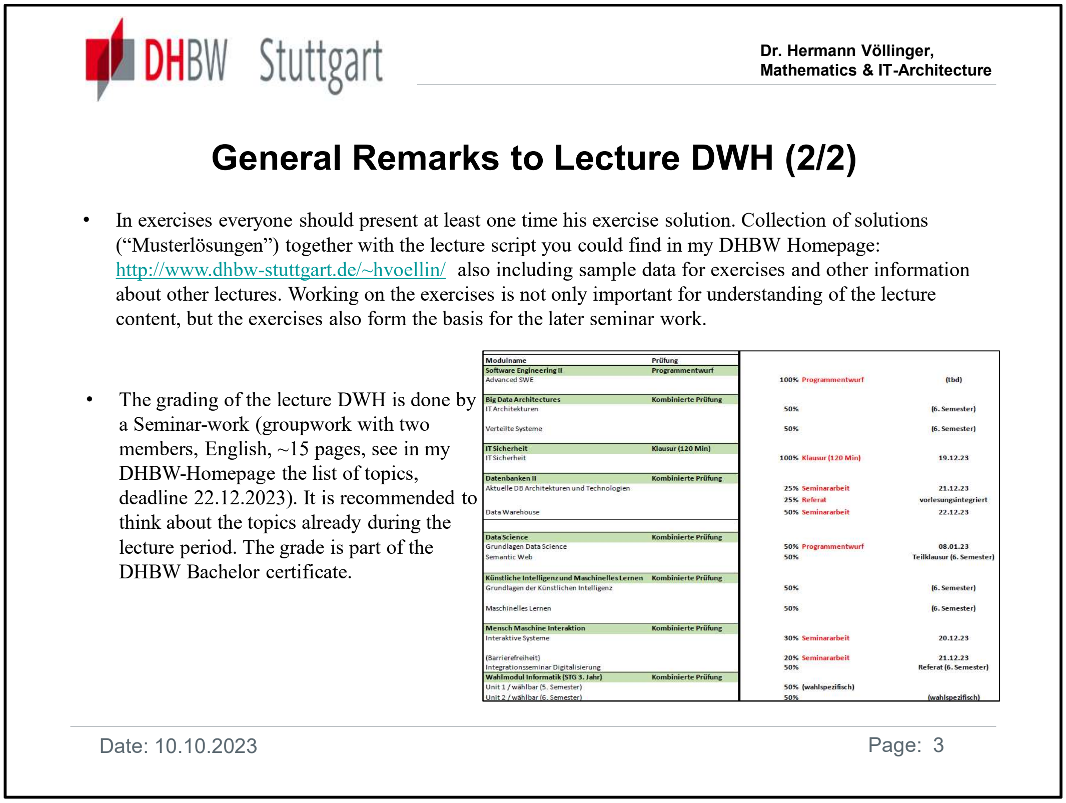
For more themes in detail in the area of Data Warehouse see for example the information of the homepage of the German company It-novum: https://it-novum.com/big-data-analytics/data-warehouse/ "**Data Warehouse - Agiles Data Warehouse mit Data Vault**"



General Remarks to Lecture DWH (1/2)

- Our first lecture starts on Tuesday, 10.10.2023, 1:00pm - 4:15pm (4 lecture hours + 15 minutes break). The remaining 8 lectures will all start at 2:00 pm. Last lecture on Tuesday, 12.12.2023.
- We have a total of 9 appointments which are all on Tuesdays except on Tuesday the 21.11.23, because on 22.11.23 the study day of the DHBW took place. The exam week runs from 18.12.23 to 22.12.23. Details: https://rapla.dhbw.de/rapla/calendar?key=YFQc7NlGleuSdybxizoa8NHjLLNjd9D6tjBdAvDwwzXobLEfUIsCXHwYu-Ma7QfggMDkLLj1CsQ-kB7hFJSGjYcYLXE5KV9oTTpcSjsE5apebBNbC_ZjtngvStO4G7YHGryjvwt1kpad5g93Dkdn0A&salt=1046252309
- If an online meeting need to be done (if corona should make this necessary), we will use Zoom (invitation credentials will then be committed in Moodle). Up to now it is planed that all lecture dates are held as face-to-face meetings.
- The lecture script is in **English**, since the common IT language in the area of DWH and Analytics is English. Some dedicated slides are in German, not to loose "Look and Feel" of the slide.
- Lecture information & supporting material (> 140 documents and technical papers in 4 DWH categories) you will find it in Moodle/"Supporting Information for DWH Lecture" Kurs DW 21E:

Many terms and catchwords of DWH are in language **English** and are also used in German language, like "Data Warehouse", "Core DWH Layer", "Business Intelligence", "Data Science" "Business Performance" and "Machine Learning", etc.
Hint: for translation from English to German and vis versa I recommend the translation tool DeepL, which is also a good example of the usage of Data Management (DWH) for a Machine Learning algorithm.
See Wikipedia: "… **DeepL Translator** is a free neural machine translation service launched on 28 August 2017 and developed by DeepL GmbH (Linguee), based in Cologne, Germany.[1][2][3] It has received positive press asserting that it is more accurate and nuanced than Google Translate.[4][5]] …"



General Remarks to Lecture DWH (2/2)

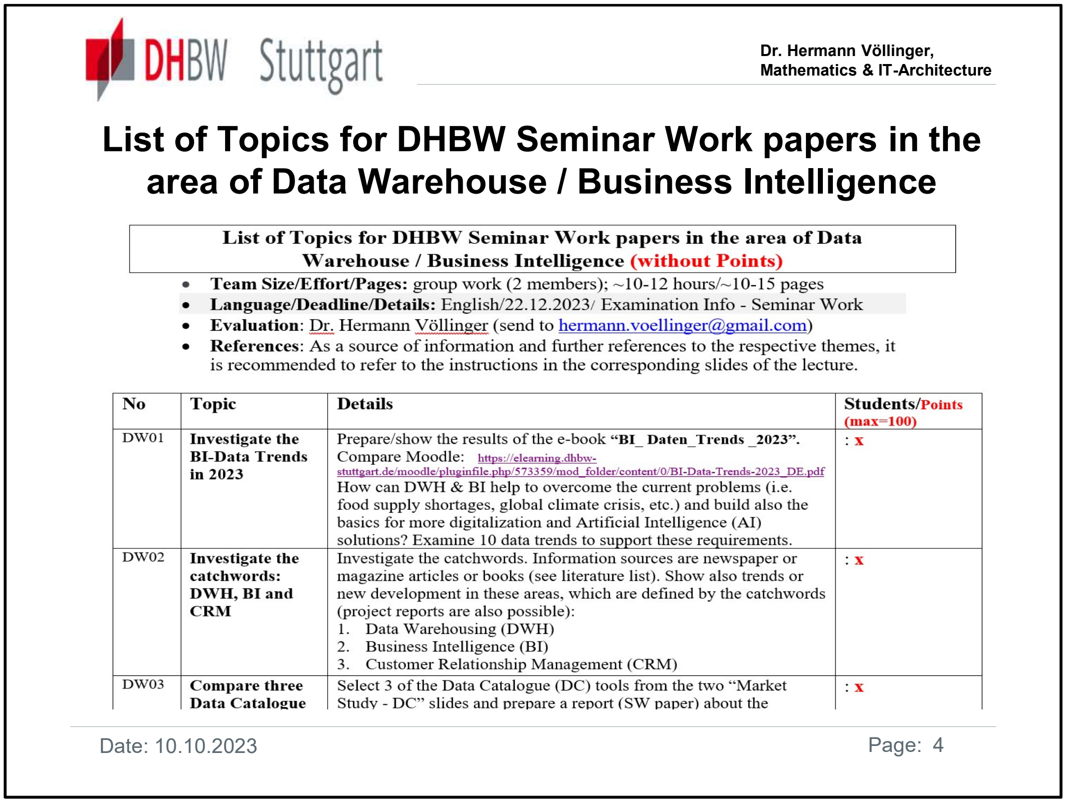
- In exercises everyone should present at least one time his exercise solution. Collection of solutions (“Musterlösungen”) together with the lecture script you could find in my DHBW Homepage: http://www.dhbw-stuttgart.de/~hvoellin/ also including sample data for exercises and other information about other lectures. Working on the exercises is not only important for understanding of the lecture content, but the exercises also form the basis for the later seminar work.

- The grading of the lecture DWH is done by a Seminar-work (groupwork with two members, English, ~15 pages, see in my DHBW-Homepage the list of topics, deadline 22.12.2023). It is recommended to think about the topics already during the lecture period. The grade is part of the DHBW Bachelor certificate.

Modulname	Prüfung			
Software Engineering II	**Programmentwurf**			
Advanced SWE		100%	Programmentwurf	(tbd)
Big Data Architectures	**Kombinierte Prüfung**			
IT Architekturen		50%		(6. Semester)
Verteilte Systeme		50%		(6. Semester)
IT Sicherheit	**Klausur (120 Min)**			
IT Sicherheit		100%	Klausur (120 Min)	19.12.23
Datenbanken II	**Kombinierte Prüfung**			
Aktuelle DB Architekturen und Technologien		25%	Seminararbeit	21.12.23
		25%	Referat	vorlesungsintegriert
Data Warehouse		50%	Seminararbeit	22.12.23
Data Science	**Kombinierte Prüfung**			
Grundlagen Data Science		50%	Programmentwurf	08.01.23
Semantic Web		50%		Teilklausur (6. Semester)
Künstliche Intelligenz und Maschinelles Lernen	**Kombinierte Prüfung**			
Grundlagen der Künstlichen Intelligenz		50%		(6. Semester)
Maschinelles Lernen		50%		(6. Semester)
Mensch Maschine Interaktion	**Kombinierte Prüfung**			
Interaktive Systeme		30%	Seminararbeit	20.12.23
(Barrierefreiheit)		20%	Seminararbeit	21.12.23
Integrationsseminar Digitalisierung		50%		Referat (6. Semester)
Wahlmodul Informatik (STG 3. Jahr)	**Kombinierte Prüfung**			
Unit 1 / wählbar (5. Semester)		50%	(wahlspezifisch)	
Unit 2 / wählbar (6. Semester)		50%		(wahlspezifisch)

In the lecture DWH you can achieve 50% of the points in the study subject "Databases II" (A3.04). The other points you get in the lecture “Advanced Structures of Databases" (WS2023: Mr. Abbas Rashidi). According to Mr. Rashidi, his lecture this semester is mainly about NoSQL databases... The students should work out topics in the environment themselves. "...but I think in the area of Big Data there could be interesting topics from the DWH lecture that the students can take with them for their work...“

Here are some scenarios of **NoSQL databases in DWH context:**
Semi-structured data processing: MongoDB or Couchbase can be useful when you need to process data with variable or changing schemas. This may be the case when handling JSON or XML data loaded into data warehouses.
Large volumes of data: Apache Cassandra or HBase are capable of efficiently processing and storing large volumes of data, which can be beneficial in big data warehousing scenario..
Real-time analytics and streaming: Apache Kafka, which are designed for streaming processing, can be used in combination with data warehouses to enable real-time analytics of streaming data.
However, it is **important to note** that the use of NoSQL databases in a data warehousing system is usually as **a complement to relational databases**. Relational databases remain the main source of structured data and the central hub for business intelligence and data analytics, while NoSQL databases cover specific requirements for unstructured and semi-structured data.



List of Topics for DHBW Seminar Work papers in the area of Data Warehouse / Business Intelligence

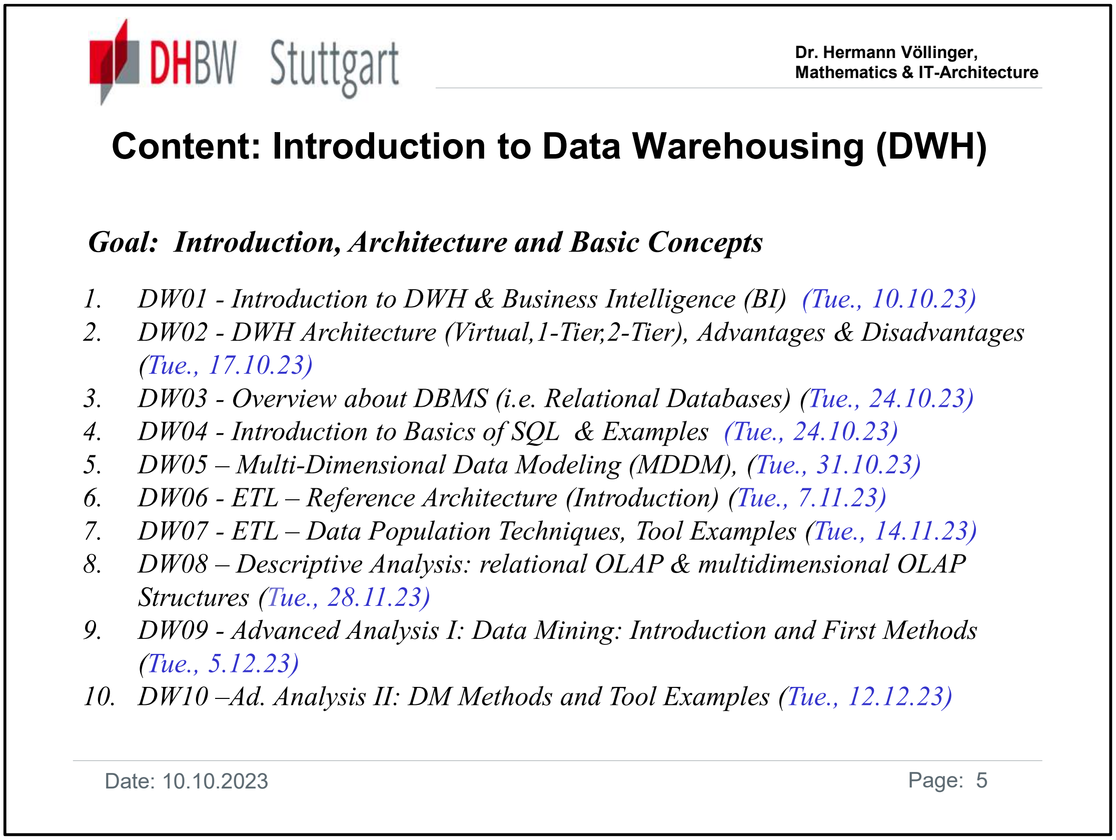
List of Topics for DHBW Seminar Work papers in the area of Data Warehouse / Business Intelligence (without Points)

- **Team Size/Effort/Pages:** group work (2 members); ~10-12 hours/~10-15 pages
- **Language/Deadline/Details:** English/22.12.2023/ Examination Info - Seminar Work
- **Evaluation**: Dr. Hermann Völlinger (send to hermann.voellinger@gmail.com)
- **References**: As a source of information and further references to the respective themes, it is recommended to refer to the instructions in the corresponding slides of the lecture.

No	Topic	Details	Students/Points (max=100)
DW01	**Investigate the BI-Data Trends in 2023**	Prepare/show the results of the e-book **"BI_ Daten_Trends _2023".** Compare Moodle: https://elearning.dhbw-stuttgart.de/moodle/pluginfile.php/573359/mod_folder/content/0/BI-Data-Trends-2023_DE.pdf How can DWH & BI help to overcome the current problems (i.e. food supply shortages, global climate crisis, etc.) and build also the basics for more digitalization and Artificial Intelligence (AI) solutions? Examine 10 data trends to support these requirements.	: x
DW02	**Investigate the catchwords: DWH, BI and CRM**	Investigate the catchwords. Information sources are newspaper or magazine articles or books (see literature list). Show also trends or new development in these areas, which are defined by the catchwords (project reports are also possible): 1. Data Warehousing (DWH) 2. Business Intelligence (BI) 3. Customer Relationship Management (CRM)	: x
DW03	**Compare three Data Catalogue**	Select 3 of the Data Catalogue (DC) tools from the two "Market Study - DC" slides and prepare a report (SW paper) about the	: x

See all further details, like evaluation sheet, structure proposal and much more in the DHBW Moodle for the course TINF21E: Kurs DW 21E

Remark: The grading of the seminar paper is also the grading of the knowledge you have acquired in the lecture. All topics of the seminar paper are deepening's of the knowledge of the lecture. For this reason, **references to the lecture** are absolutely necessary.



Content: Introduction to Data Warehousing (DWH)

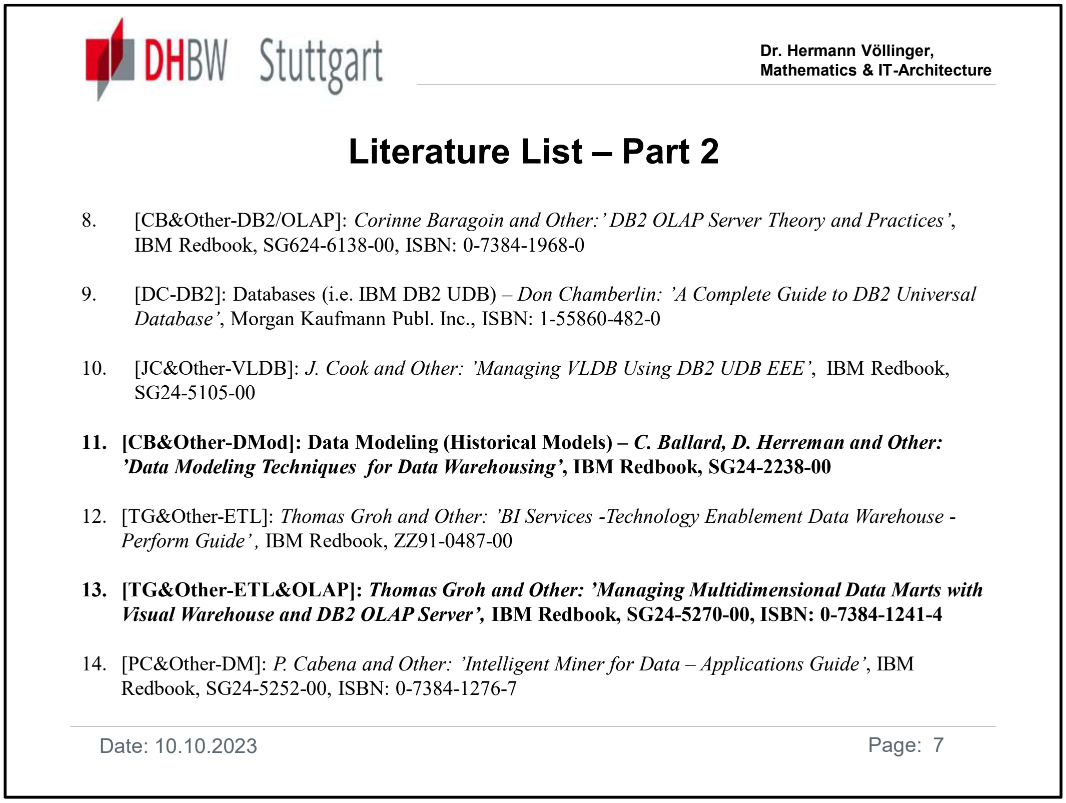
Goal: Introduction, Architecture and Basic Concepts

1. *DW01 - Introduction to DWH & Business Intelligence (BI) (Tue., 10.10.23)*
2. *DW02 - DWH Architecture (Virtual,1-Tier,2-Tier), Advantages & Disadvantages (Tue., 17.10.23)*
3. *DW03 - Overview about DBMS (i.e. Relational Databases) (Tue., 24.10.23)*
4. *DW04 - Introduction to Basics of SQL & Examples (Tue., 24.10.23)*
5. *DW05 – Multi-Dimensional Data Modeling (MDDM), (Tue., 31.10.23)*
6. *DW06 - ETL – Reference Architecture (Introduction) (Tue., 7.11.23)*
7. *DW07 - ETL – Data Population Techniques, Tool Examples (Tue., 14.11.23)*
8. *DW08 – Descriptive Analysis: relational OLAP & multidimensional OLAP Structures (Tue., 28.11.23)*
9. *DW09 - Advanced Analysis I: Data Mining: Introduction and First Methods (Tue., 5.12.23)*
10. *DW10 –Ad. Analysis II: DM Methods and Tool Examples (Tue., 12.12.23)*

Literature List – Part 1

1. [BD-DWH]: *Barry Devlin 'Data Warehouse....'*, Addison-Wesley, ISBN: 0-201-96425-2
2. **[RK-DWH]: *R. Kimball 'The Data Warehouse Toolkit.'*, John Wiley & Sons, NY 1996, ISBN: 0-471-15337-0**
3. **[AB&HG-DWH]: *Andreas Bauer, Holger Günzel (Hrsg.): 'Data Warehouse Systeme - Architektur, Entwicklung, Anwendung'* DPunkt Verlag Heidelberg 2004, 3. Auflage, ISBN: 978-3-89864-540-9**
4. [RK-DWH/TK]: *R. Kimball and Other: 'The Data Warehouse Lifecycle Toolkit'*, John Wiley & Sons, NY 1998, ISBN: 0-471-25547-5
5. [SE-DWH/BI]: *Stefan Eckrich and Other: 'From Multiplatform Operational Data to Data Warehousing and Business Intelligence'*, IBM Redbook, SG24-5174-00, ISBN: 0-7384-0032-7
6. [VAC&Other-BI/390]: *V. Anavi-Chaput and Other: 'Business Intelligence Architecture on S/390 – Presentation Guide'*, IBM Redbook, SG24-5641-00, ISBN: 0-7384-1752-1
7. **[DM-MD]: *David Marco: 'Building &Managing the Meta Data Repository'*, John Wiley & Sons 2000, ISBN: 0-471-35523-2**

IBM Redbooks can be found in the Internet under : http://www.redbooks.ibm.com



Literature List – Part 2

8. [CB&Other-DB2/OLAP]: *Corinne Baragoin and Other:’ DB2 OLAP Server Theory and Practices’*, IBM Redbook, SG624-6138-00, ISBN: 0-7384-1968-0

9. [DC-DB2]: Databases (i.e. IBM DB2 UDB) – *Don Chamberlin: ’A Complete Guide to DB2 Universal Database’*, Morgan Kaufmann Publ. Inc., ISBN: 1-55860-482-0

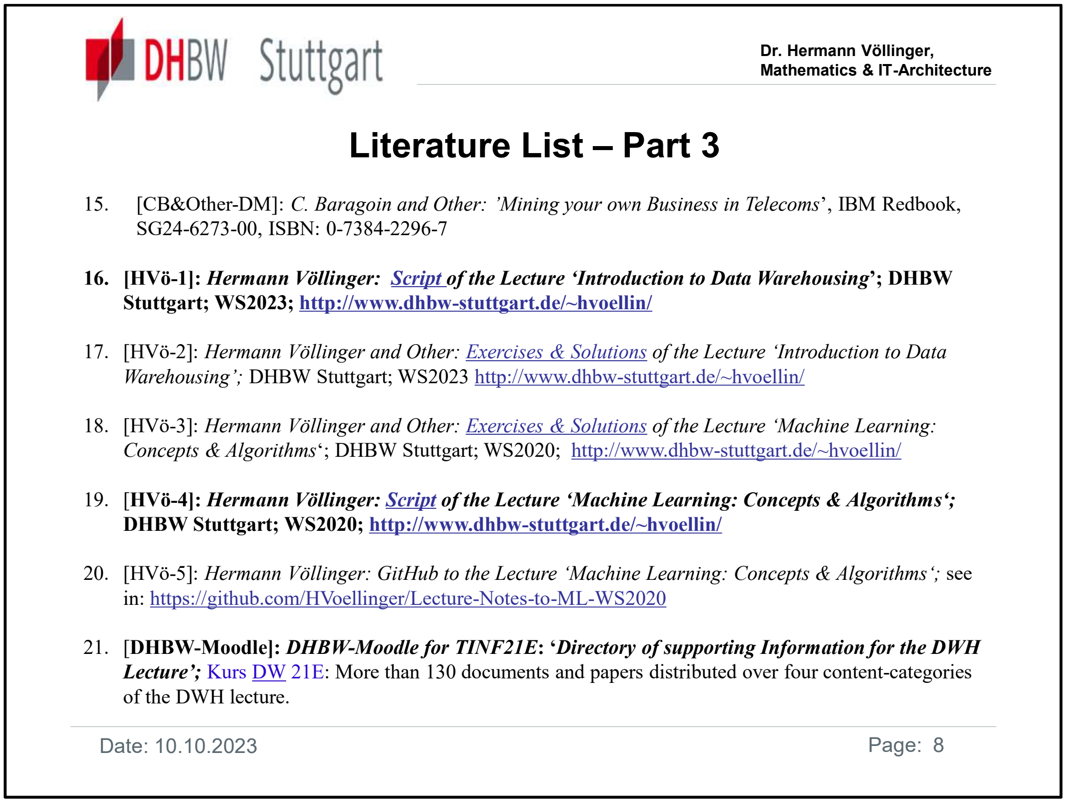
10. [JC&Other-VLDB]: *J. Cook and Other: ’Managing VLDB Using DB2 UDB EEE’*, IBM Redbook, SG24-5105-00

11. **[CB&Other-DMod]: Data Modeling (Historical Models) – *C. Ballard, D. Herreman and Other: ’Data Modeling Techniques for Data Warehousing’*, IBM Redbook, SG24-2238-00**

12. [TG&Other-ETL]: *Thomas Groh and Other: ’BI Services -Technology Enablement Data Warehouse - Perform Guide’ ,* IBM Redbook, ZZ91-0487-00

13. **[TG&Other-ETL&OLAP]: *Thomas Groh and Other: ’Managing Multidimensional Data Marts with Visual Warehouse and DB2 OLAP Server’,* IBM Redbook, SG24-5270-00, ISBN: 0-7384-1241-4**

14. [PC&Other-DM]: *P. Cabena and Other: ’Intelligent Miner for Data – Applications Guide’*, IBM Redbook, SG24-5252-00, ISBN: 0-7384-1276-7



Literature List – Part 3

15. [CB&Other-DM]: *C. Baragoin and Other: 'Mining your own Business in Telecoms'*, IBM Redbook, SG24-6273-00, ISBN: 0-7384-2296-7

16. **[HVö-1]: *Hermann Völlinger: Script of the Lecture 'Introduction to Data Warehousing'*; DHBW Stuttgart; WS2023; http://www.dhbw-stuttgart.de/~hvoellin/**

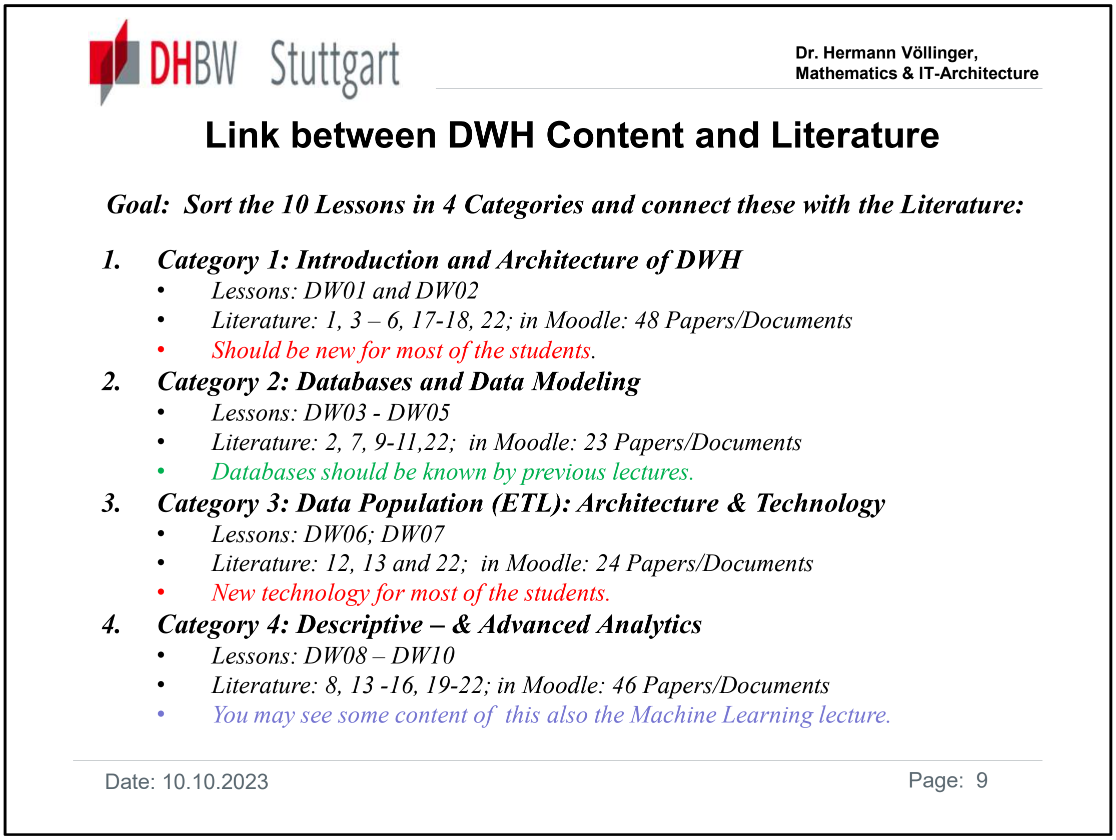
17. [HVö-2]: *Hermann Völlinger and Other: Exercises & Solutions of the Lecture 'Introduction to Data Warehousing';* DHBW Stuttgart; WS2023 http://www.dhbw-stuttgart.de/~hvoellin/

18. [HVö-3]: *Hermann Völlinger and Other: Exercises & Solutions of the Lecture 'Machine Learning: Concepts & Algorithms*'; DHBW Stuttgart; WS2020; http://www.dhbw-stuttgart.de/~hvoellin/

19. **[HVö-4]: *Hermann Völlinger: Script of the Lecture 'Machine Learning: Concepts & Algorithms';* DHBW Stuttgart; WS2020; http://www.dhbw-stuttgart.de/~hvoellin/**

20. [HVö-5]: *Hermann Völlinger: GitHub to the Lecture 'Machine Learning: Concepts & Algorithms';* see in: https://github.com/HVoellinger/Lecture-Notes-to-ML-WS2020

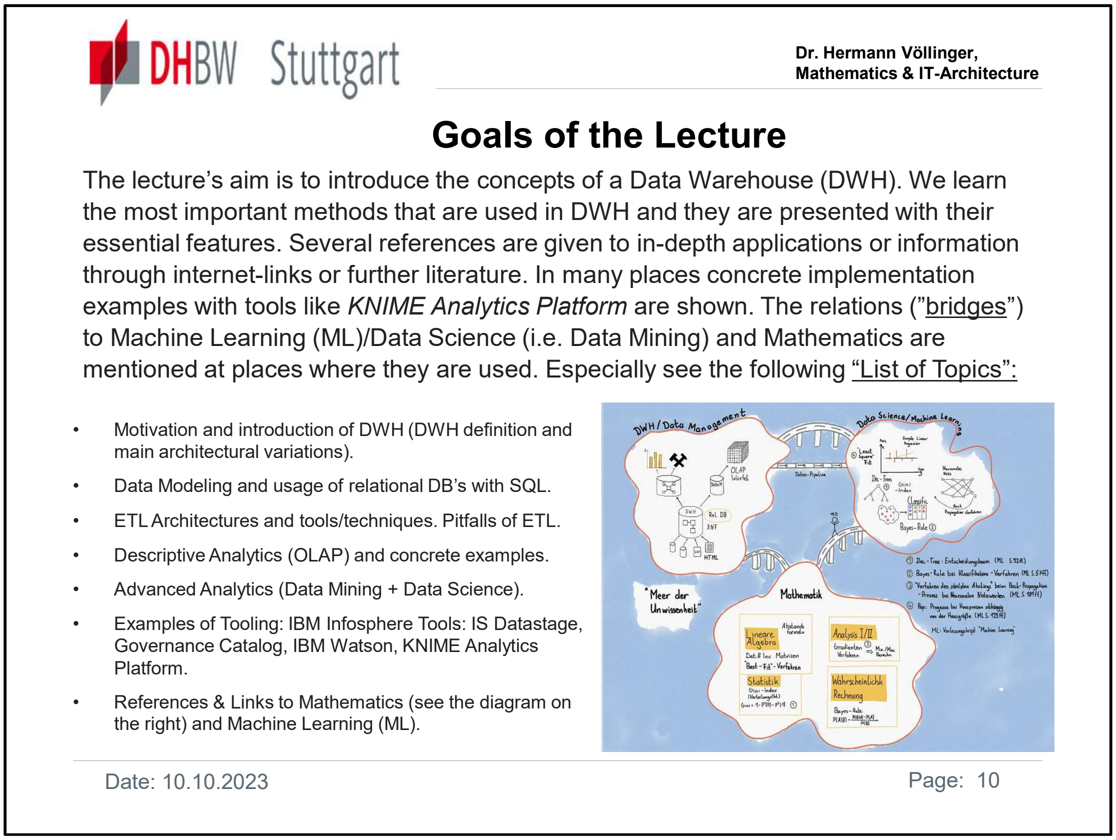
21. **[DHBW-Moodle]: *DHBW-Moodle for TINF21E*: '*Directory of supporting Information for the DWH Lecture';*** Kurs DW 21E: More than 130 documents and papers distributed over four content-categories of the DWH lecture.



Link between DWH Content and Literature

Goal: Sort the 10 Lessons in 4 Categories and connect these with the Literature:

1. ***Category 1: Introduction and Architecture of DWH***
   - *Lessons: DW01 and DW02*
   - *Literature: 1, 3 – 6, 17-18, 22; in Moodle: 48 Papers/Documents*
   - *Should be new for most of the students.*
2. ***Category 2: Databases and Data Modeling***
   - *Lessons: DW03 - DW05*
   - *Literature: 2, 7, 9-11,22; in Moodle: 23 Papers/Documents*
   - *Databases should be known by previous lectures.*
3. ***Category 3: Data Population (ETL): Architecture & Technology***
   - *Lessons: DW06; DW07*
   - *Literature: 12, 13 and 22; in Moodle: 24 Papers/Documents*
   - *New technology for most of the students.*
4. ***Category 4: Descriptive – & Advanced Analytics***
   - *Lessons: DW08 – DW10*
   - *Literature: 8, 13 -16, 19-22; in Moodle: 46 Papers/Documents*
   - *You may see some content of this also the Machine Learning lecture.*



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Goals of the Lecture

The lecture's aim is to introduce the concepts of a Data Warehouse (DWH). We learn the most important methods that are used in DWH and they are presented with their essential features. Several references are given to in-depth applications or information through internet-links or further literature. In many places concrete implementation examples with tools like *KNIME Analytics Platform* are shown. The relations ("bridges") to Machine Learning (ML)/Data Science (i.e. Data Mining) and Mathematics are mentioned at places where they are used. Especially see the following "List of Topics":

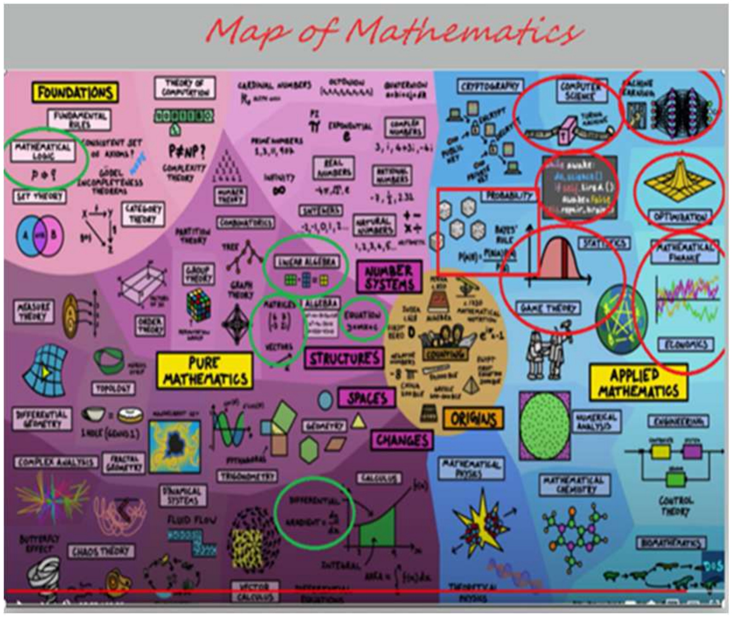
- Motivation and introduction of DWH (DWH definition and main architectural variations).
- Data Modeling and usage of relational DB's with SQL.
- ETL Architectures and tools/techniques. Pitfalls of ETL.
- Descriptive Analytics (OLAP) and concrete examples.
- Advanced Analytics (Data Mining + Data Science).
- Examples of Tooling: IBM Infosphere Tools: IS Datastage, Governance Catalog, IBM Watson, KNIME Analytics Platform.
- References & Links to Mathematics (see the diagram on the right) and Machine Learning (ML).

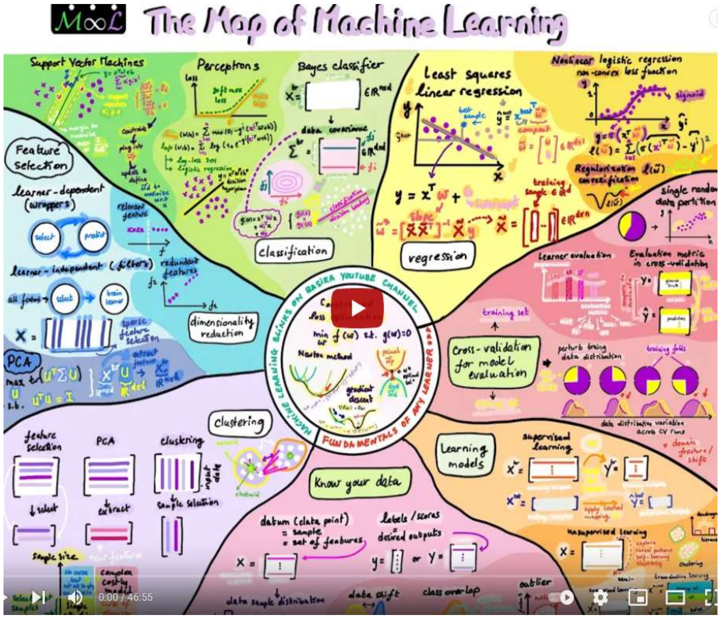
Date: 10.10.2023

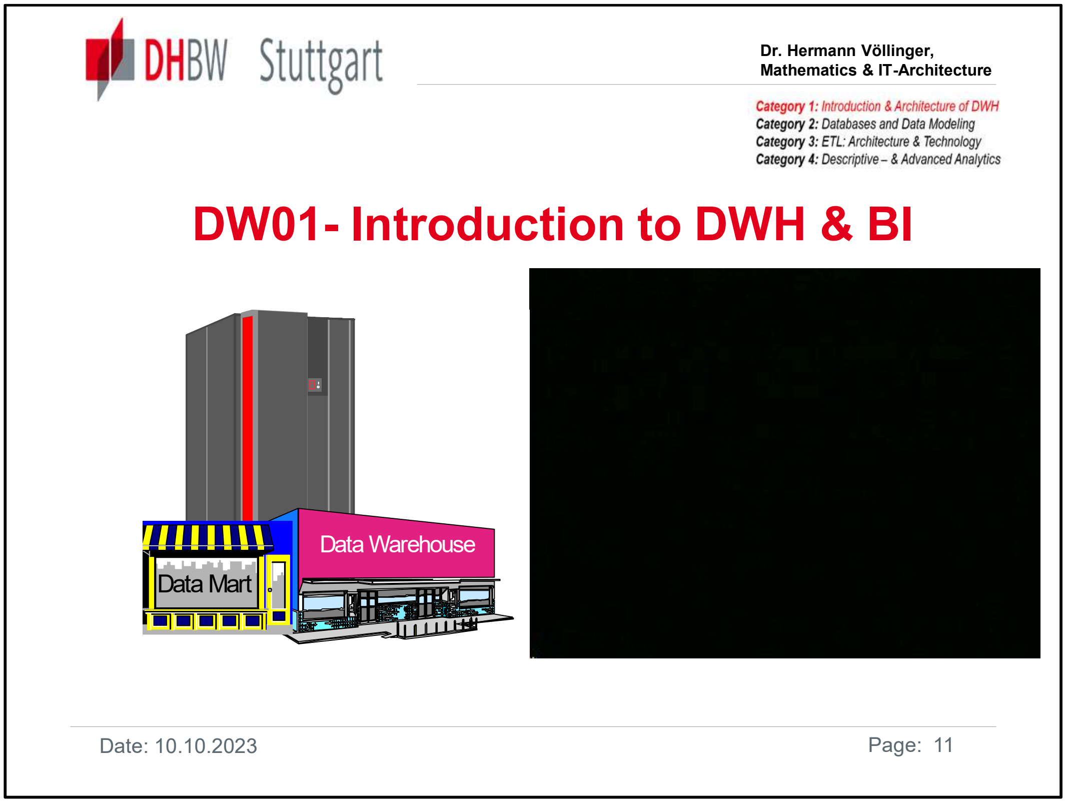
Page: 10

See "Map of Mathematics": https://www.flickr.com/photos/95869671@N08/32264483720

See "The Map of Machine Learning": https://www.youtube.com/watch?v=eDRng1eMKRA







DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Category 1: Introduction & Architecture of DWH
Category 2: Databases and Data Modeling
Category 3: ETL: Architecture & Technology
Category 4: Descriptive – & Advanced Analytics

DW01- Introduction to DWH & BI

Date: 10.10.2023 Page: 11

The embedded video is an old IBM commercial – "The stinky Greece cheese".
See the video source as "cheese.mpeg" at Kurs DW 21E

The story was this: Analysis had shown that a badly smelling Greece cheese was not profitable, so by default it should be replaced.
But by using data mining analysis, it was found out that the majority of the customers who bought this cheese also bought some high margin products, e.g. the fine wines, etc., so to not risk losing valuable customers to the competition, it was decided to keep it on the shelves.

Motivation - What is Business Intelligence (BI) – the Problem

Water, Water Everywhere …

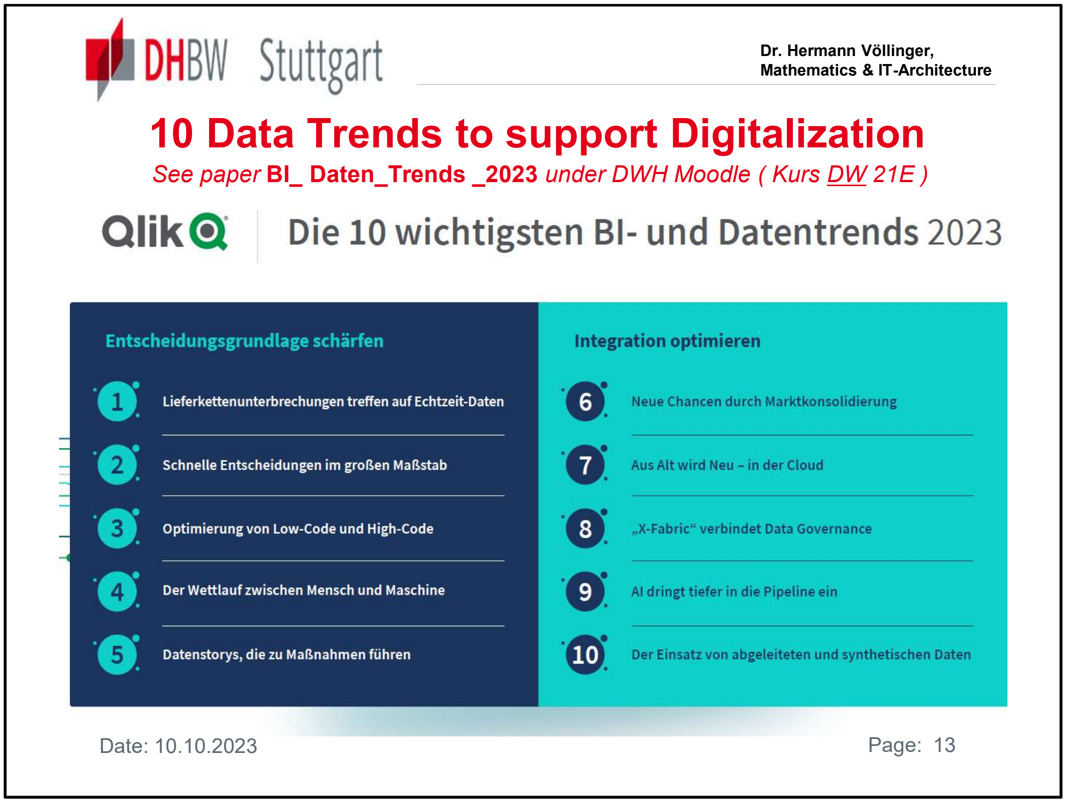
…But not a drop to drink

Data, Data Everywhere…

…But none to help me think

Competition

The slide shows the main problem in today‘s business:
We have a big mass of data existing in transaction system, but we can not use it, since the data is not available or accessible.



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

10 Data Trends to support Digitalization

See paper ***BI_ Daten_Trends _2023*** *under DWH Moodle (Kurs DW 21E)*

Qlik | Die 10 wichtigsten BI- und Datentrends 2023

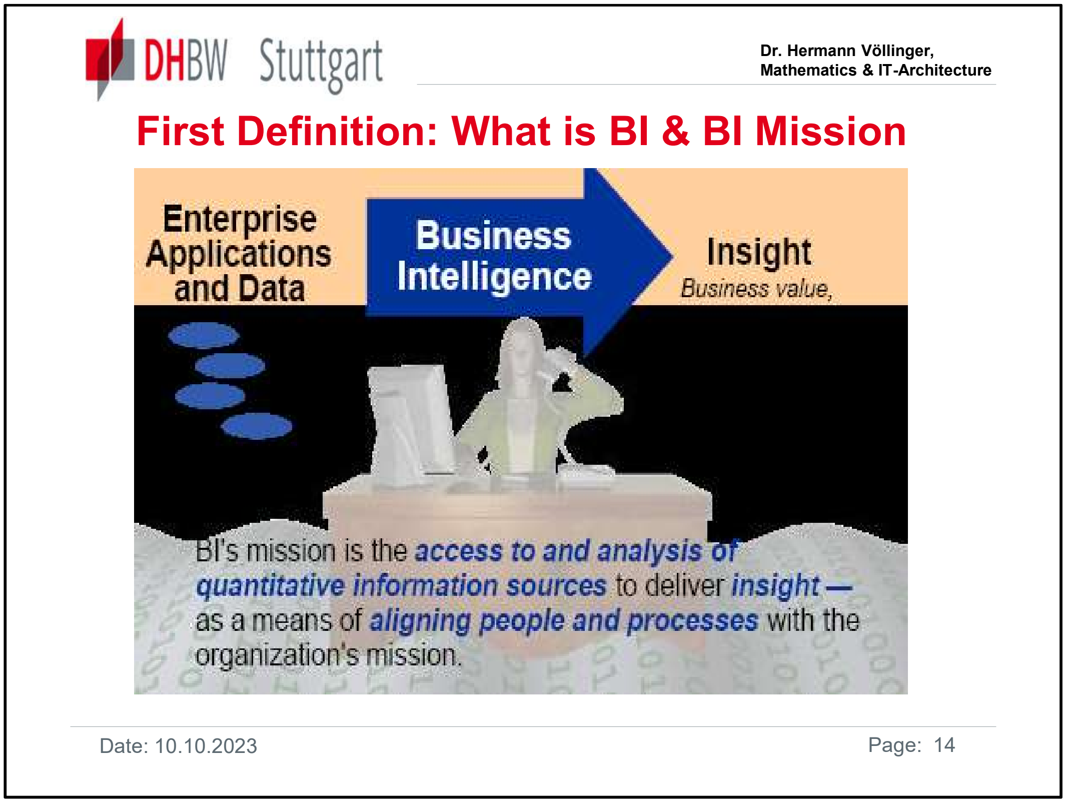
Entscheidungsgrundlage schärfen

1. Lieferkettenunterbrechungen treffen auf Echtzeit-Daten
2. Schnelle Entscheidungen im großen Maßstab
3. Optimierung von Low-Code und High-Code
4. Der Wettlauf zwischen Mensch und Maschine
5. Datenstorys, die zu Maßnahmen führen

Integration optimieren

6. Neue Chancen durch Marktkonsolidierung
7. Aus Alt wird Neu – in der Cloud
8. „X-Fabric“ verbindet Data Governance
9. AI dringt tiefer in die Pipeline ein
10. Der Einsatz von abgeleiteten und synthetischen Daten

See: also Exercise E1.1



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

First Definition: What is BI & BI Mission

Enterprise Applications and Data

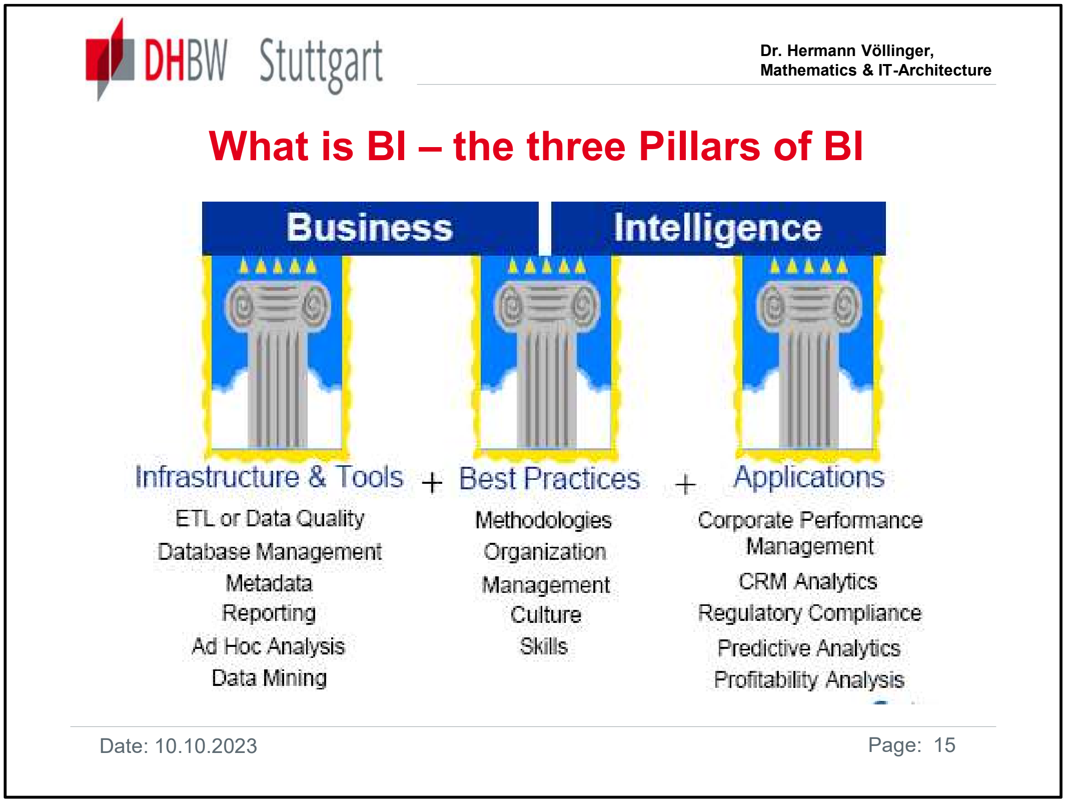
Business Intelligence

Insight
Business value,

BI's mission is the ***access to and analysis of quantitative information sources*** to deliver ***insight*** — as a means of ***aligning people and processes*** with the organization's mission.

Date: 10.10.2023 Page: 14

BI has to mission to get more knowledge about my business.
But before I can get to this knowledge, I must prepare the infrastructure for analysis and BI , i.e. I must build a DWH (this is the rough idea of Data Warehousing).



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

What is BI – the three Pillars of BI

Business Intelligence

Infrastructure & Tools + Best Practices + Applications

Infrastructure & Tools
- ETL or Data Quality
- Database Management
- Metadata
- Reporting
- Ad Hoc Analysis
- Data Mining

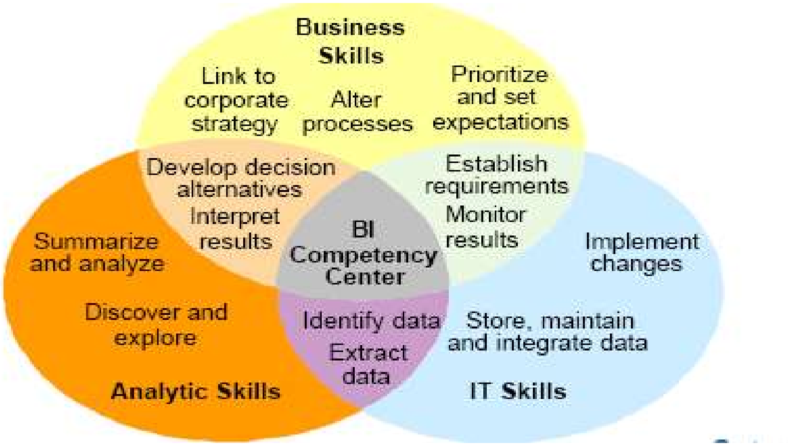
Best Practices
- Methodologies
- Organization
- Management
- Culture
- Skills

Applications
- Corporate Performance Management
- CRM Analytics
- Regulatory Compliance
- Predictive Analytics
- Profitability Analysis

Date: 10.10.2023 Page: 15

BI is based on these three pillars.
Pillar one is necessary for pillar three.
Pillar two is the method to build a successful DWH/BI solution.

What is BI – the BI Competency Center



Business Skills

Link to corporate strategy

Alter processes

Prioritize and set expectations

Develop decision alternatives

Interpret results

Establish requirements

Monitor results

BI Competency Center

Summarize and analyze

Implement changes

Discover and explore

Identify data

Extract data

Store, maintain and integrate data

Analytic Skills

IT Skills

DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

BI - Getting the Answers you Need

What products sold well this month in Germany?

What types of marketing campaigns should we deploy?

This history of sales in USA by city, by product shows we still have tremendous potential

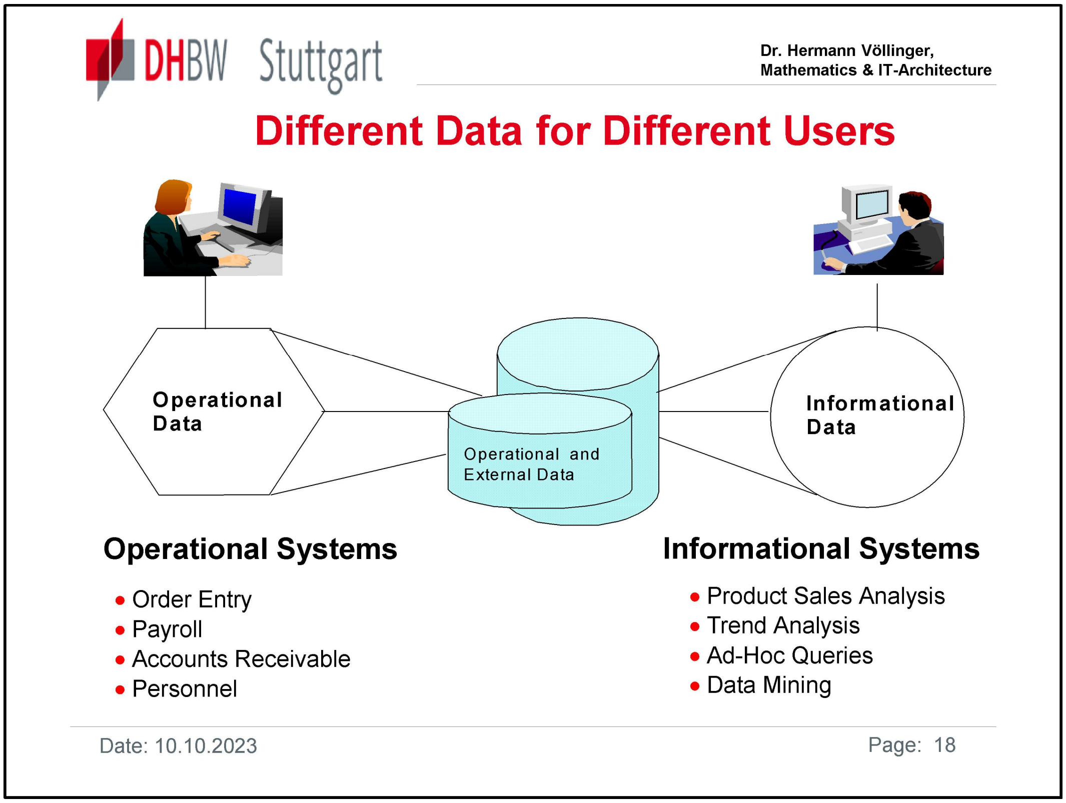
Look at these sales numbers compared to the last 3 years

Let's have a campaign on these 3 products where profit is over 15%

Date: 10.10.2023 Page: 17

The slide shows the questions and goals of a Business Analysis in a company. The answer to these questions is called 'Analytical CRM'.
We will learn more about this in the following chapters

- This slide shows people asking the kind of questions that good data analysis and query tools do an excellent job of answering. Questions such as "Which products are selling best?", and "What campaigns should we execute?" can provide real business advantages for companies if they are answered correctly, efficiently and promptly. The problem is to provide the answers needed for decision making requires information. To get information, you must start with data -- data which is available from many sources internal to the organization and from external sources.
- Organizations today do not suffer from a lack of data, but from an **abundance of redundant and inconsistent** data that is difficult to manage effectively, is increasingly difficult to access and difficult to use for decision making purposes. No matter how good the tool is, if the database being accessed does not have the right data in the right form, the answers will be unsatisfactory.
- What is required is an architected solution that makes the best use of all the available data and transforms it into a more appropriate format for decision making. What is required is Data Warehousing.
- Data Warehousing makes the best use of all available data and turns it into data structures, which could be accessed and analyzed by analytical processes and tools like OLAP and Data Mining. So we get out of data the valuable information.



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Different Data for Different Users

Operational Systems

- Order Entry
- Payroll
- Accounts Receivable
- Personnel

Informational Systems

- Product Sales Analysis
- Trend Analysis
- Ad-Hoc Queries
- Data Mining

Operational and Informational Applications

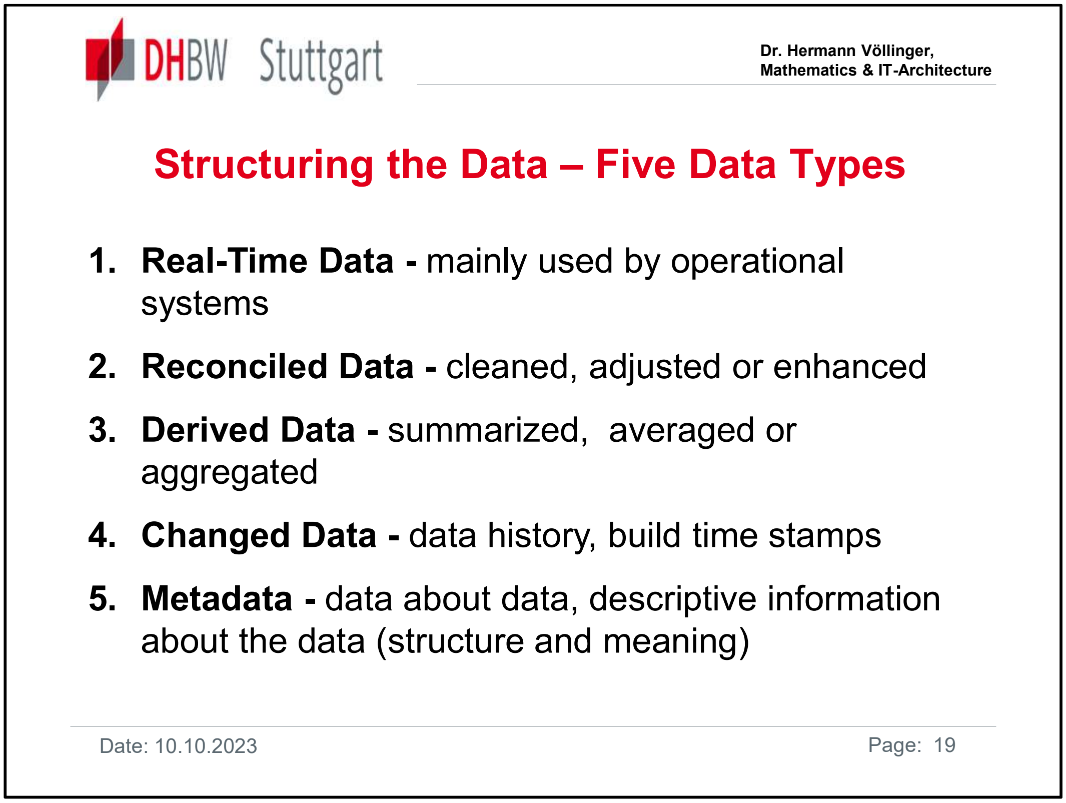
Much of the data inside organizations is captured and used by the day-to-day **operational applications** that help run the business, e.g. order entry, payroll, etc. These applications are optimized for operational purposes and the data typically is real-time and constantly changing. It is typically not in a format or structure that is easily understood or easily accessed by end users for informational purposes. These systems are also often called **Legacy Systems.**

Informational applications require a more stable source of data that is in a format more suitable for analysis, e.g. query and reporting, trend and point-in-time analysis, data mining. The data is usually "transformed", to remove data of use only in the operational environment, change the data format, and eliminate anomalies to improve data quality. Transformation also makes the data more easy to understand and use by business users. The data is stable, consistent as of some point-in-time and will not be changed until the end user desires. It may reflect a history of values, i.e. changing values over time.

Different Data for Different Uses

Most organizations need two data environments, one for operational and one for informational applications. The original data for both types of applications may be the same, but the data used by these two applications is fundamentally different. Using the same data environment to support both these applications can compromise the performance, capability, and benefit they provide.

A data warehouse provides a separate physical data store to better support informational applications (‘**dispositive’** data). The data to support these informational applications, however, does need to be organized in specific ways.



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Structuring the Data – Five Data Types

1. **Real-Time Data -** mainly used by operational systems
2. **Reconciled Data -** cleaned, adjusted or enhanced
3. **Derived Data -** summarized, averaged or aggregated
4. **Changed Data -** data history, build time stamps
5. **Metadata -** data about data, descriptive information about the data (structure and meaning)

Date: 10.10.2023 Page: 19

There are 5 main data types involved in data warehousing:

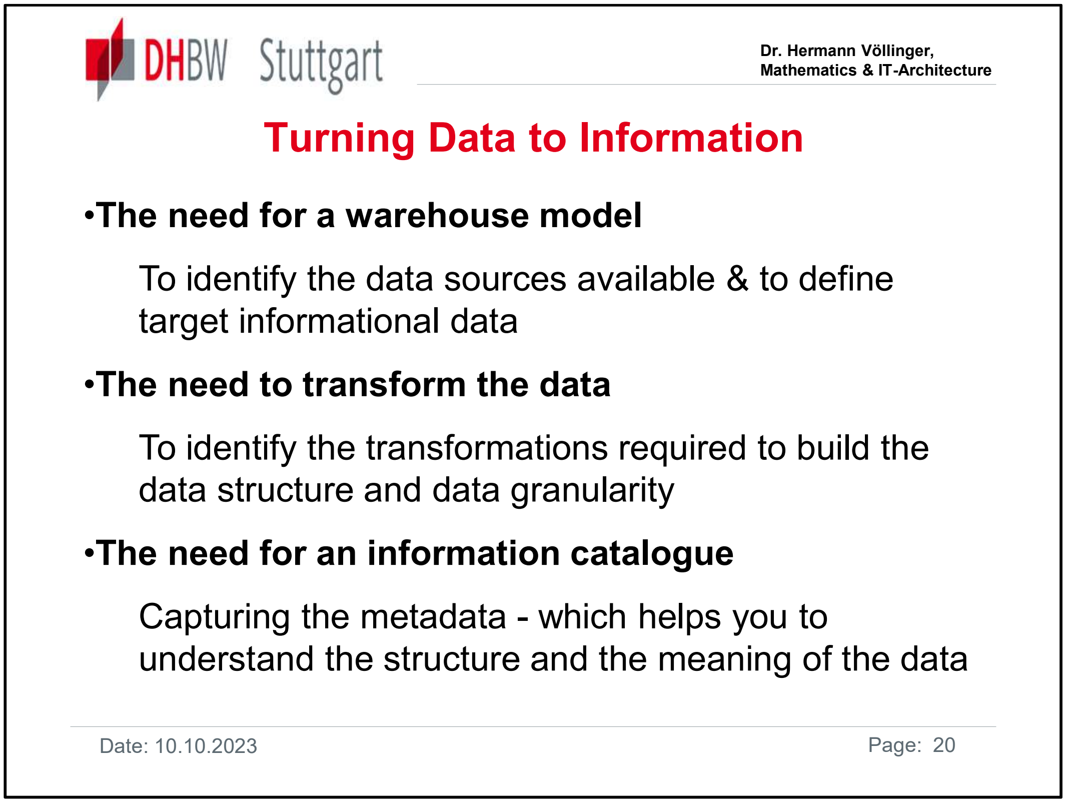
Real-Time Data: Typically used only by operational applications. It contains all the individual, detailed data values, each update overlaying the previous value so there exists no history of changes. It may exist redundantly in multiple locations, which may or may not be synchronized together, e.g. data from several bank branches. It may also be inconsistent in representation, meaning or both. Real-time data normally requires some transformation before being used by informational applications.

Reconciled Data: Real-time data that has been cleaned, adjusted, or enhanced to provide a source of quality data for use by informational applications.

Derived Data: Summarized, averaged or aggregated data from multiple sources of real-time or reconciled data, providing detailed data in a suitable format for decision making, reducing resource requirements and increasing application response time.

Changed Data: Data that can kept as a continuum and reflects the data history. It is a record of all the changes to selected real-time data, time-stamped to document the level of currency. Since all changes are included, trend or point-in-time analysis can be achieved. This does present issues such as how to archive the data to make retrieval easy when desired and could mean storage of a large volume of data.

Metadata: Provides descriptive information on what data is available for accessing, exactly what it represents, how current it is, etc. for the development of informational applications. Building a data warehouse requires capturing both data and metadata. Typically metadata is used for database design and application development, but with data warehousing, metadata definitions are also transformed into business terms for end users and an information catalog is provided to make it easy for end users to search for and use the metadata.



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Turning Data to Information

- **The need for a warehouse model**

  To identify the data sources available & to define target informational data

- **The need to transform the data**

  To identify the transformations required to build the data structure and data granularity

- **The need for an information catalogue**

  Capturing the metadata - which helps you to understand the structure and the meaning of the data

Date: 10.10.2023 Page: 20

Need for a data warehouse model

➢ To enable the movement of data into the informational databases (data warehouses), the "master" data sources must be identified. This sometimes requires an "inventory" of the current operational data to make sure the appropriate data elements are accounted for. This information is used to create the definitions for the target informational databases. The formal term for this activity is "creating a data warehouse data model". This model will define the target informational data. We will see these in ore detail in a following chapter of the lecture.

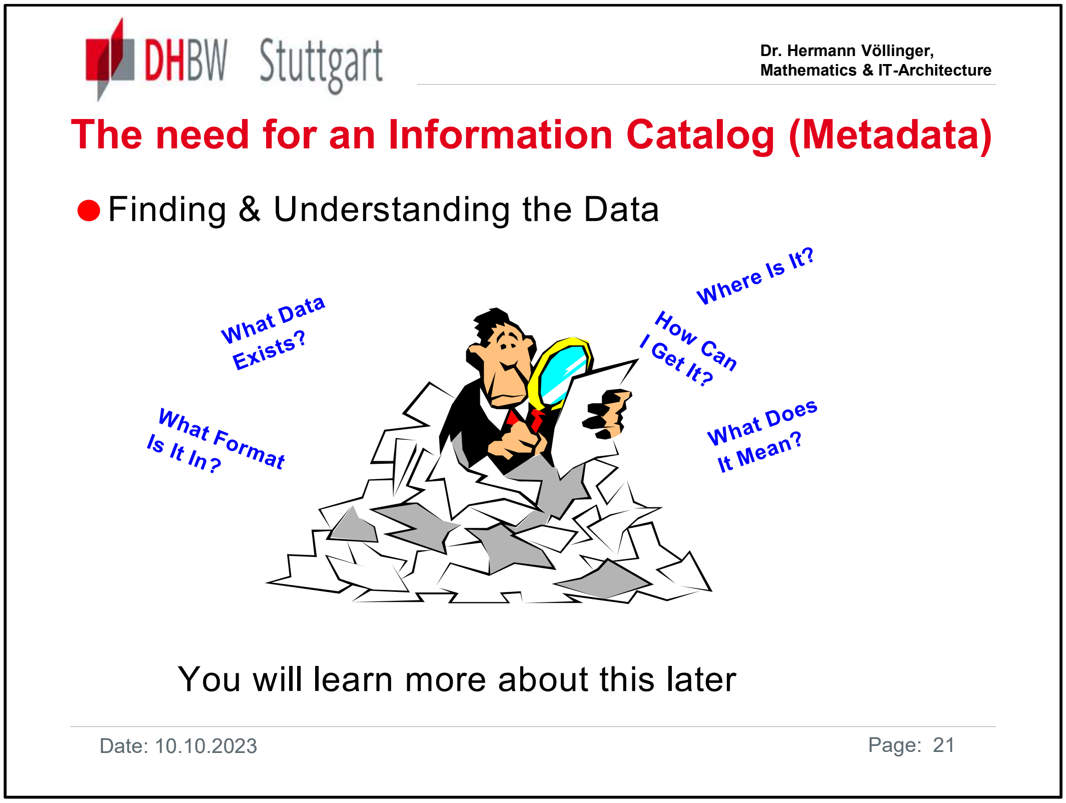
Need to transform the data

➢ Next, the transformations and enhancements that are needed must be defined. For example, rather than sending all the detailed data only summary information will be sent; or only a subset of the detailed data sent based on specific criteria. Data could be enhanced by such actions as adding a time value. Based on these transformation requirements, the data in the warehouse could consist of **reconciled, derived,** and/or **changed** data**.**

Need for an information catalog

➢ Another part of the process involves capturing the metadata that defines the informational data in the data warehouse. This metadata is used to populate an information catalog that enables users to understand what data is in the data warehouse and exactly what it means. Metadata is key in data warehousing as it provides users a "window" to the world of data in the data warehouse.

Before looking more closely at the data-to-information process, lets first look at the data types involved.



The need for an Information Catalog (Metadata)

- Finding & Understanding the Data

What Data Exists?

Where Is It?

How Can I Get It?

What Format Is It In?

What Does It Mean?

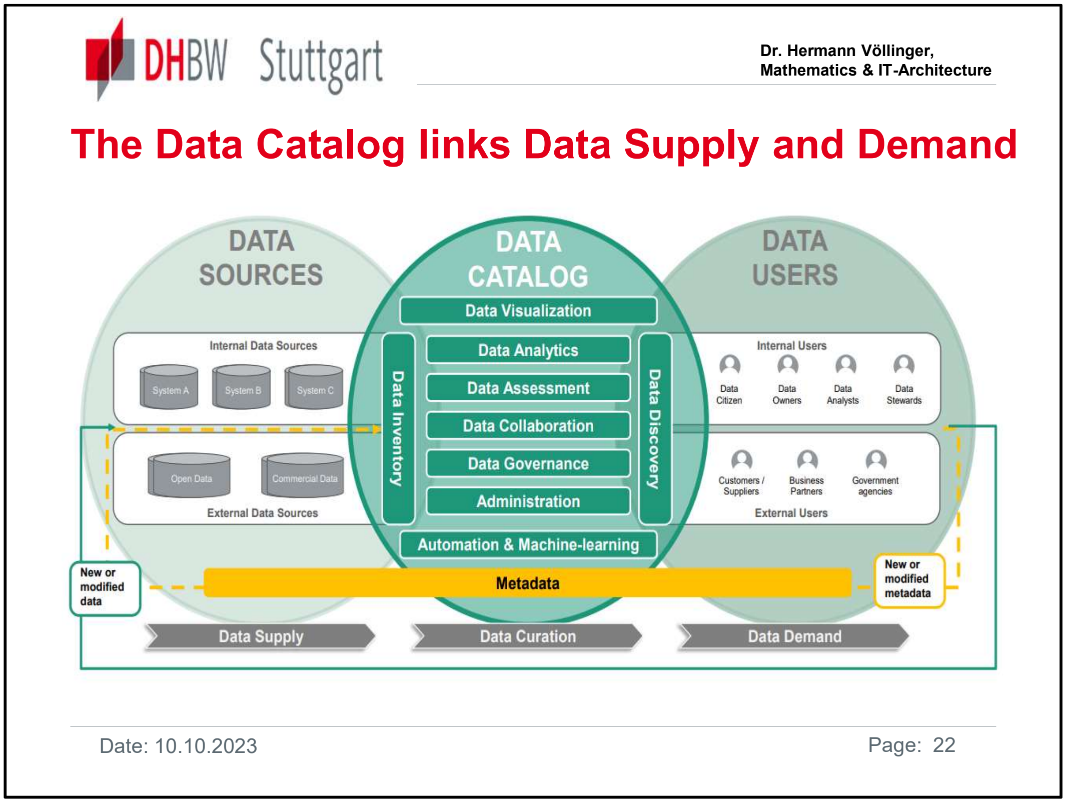
You will learn more about this later

Having set up the data warehouse and provided a data store that is optimized for informational applications, it is important to provide end users with a **way to find and understand the data** in the data warehouse. What is required is an Information Catalog for users, with data descriptions in business terms they can understand.

➢ To access the data, users must first know **what is available** in the data warehouse. This leads to the requirements for an information catalog. To achieve this, the metadata from the source databases is captured and used in the definition of the data warehouse target tables. That source metadata, and metadata that describes any newly defined data elements, provide the base for the data warehouse model. It will have two uses: "technical" metadata used by the administrator for representing the data model, and as input to a "business-oriented" metadata catalog.

➢ **Catalog Information for the Administrators and End Users:** For the administrator, metadata contains the description of the data environment and is the basis for the data warehouse model. It will provide the basis for a "technical" metadata catalog that will support the associations that enable actual access to the data. This technical catalog is used to enable the creation of the business-oriented information catalog.

➢ **The Information Catalog:** For the end user, the information catalog is the key that opens up a world of information. It describes what data exists, where it is located, its currency and so on. It may also be able to catalog and describe information objects, such as queries, charts and reports. All of these descriptions are in business terms that end users can understand. This helps to assure they access and use the correct data and information objects. The data warehouse entry point for the end user is an information catalog: it tells the end user what exists, what it means, and how to use it, all expressed in business terms.



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

The Data Catalog links Data Supply and Demand

DATA SOURCES
DATA CATALOG
DATA USERS

Data Visualization
Data Analytics
Data Assessment
Data Collaboration
Data Governance
Administration
Data Inventory
Data Discovery
Automation & Machine-learning

Internal Data Sources
System A
System B
System C
Open Data
Commercial Data
External Data Sources

Internal Users
Data Citizen
Data Owners
Data Analysts
Data Stewards
Customers / Suppliers
Business Partners
Government agencies
External Users

New or modified data
Metadata
New or modified metadata

Data Supply
Data Curation
Data Demand

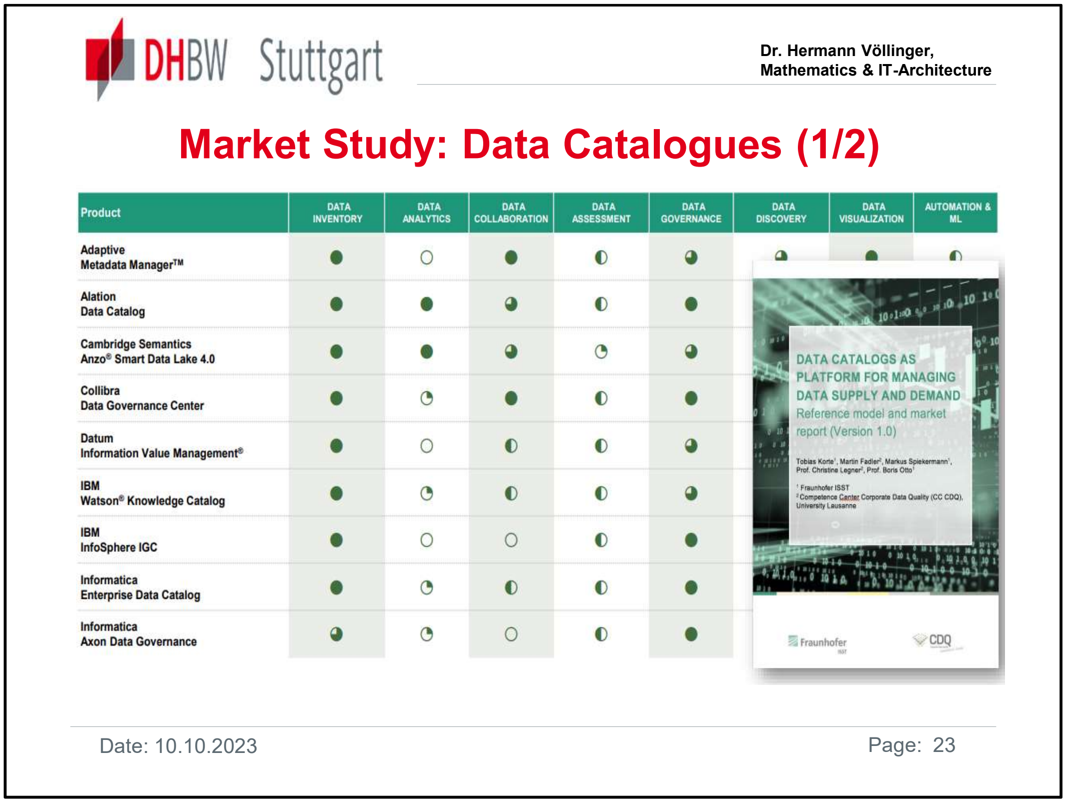
There is also some more detailed information available:

“Data Catalog Reference Model and Market Study” of the CC CDQ: includes important definitions, a reference model for the evaluation of features of a data catalog, also an overview about existing DC tools in the marketplace: https://www.cc-cdq.ch/data-catalogs
See also the slides 8-12 in the presentation «CDQ – Data Catalogs»

CDQ Consulting: supports the conception, evaluation and implementation of Data Catalogs: https://www.cdq.ch/data-catalog

Further research activities: Together with leading companies, CC-CDQ conduct research on topics such as data strategies, data governance or machine learning for data management: https://www.cc-cdq.ch/data-management-research

See also: https://www.alation.com/wp-content/uploads/Data_Catalogs_Creating_a_Single_Source_of_Reference.pdf



Market Study: Data Catalogues (1/2)

Product	DATA INVENTORY	DATA ANALYTICS	DATA COLLABORATION	DATA ASSESSMENT	DATA GOVERNANCE	DATA DISCOVERY	DATA VISUALIZATION	AUTOMATION & ML
Adaptive Metadata Manager™	●	○	●	◐	◕	◕	●	◐
Alation Data Catalog	●	●	◕	◐	●			
Cambridge Semantics Anzo® Smart Data Lake 4.0	●	●	◕	◔	◕			
Collibra Data Governance Center	●	◔	●	◐	●			
Datum Information Value Management®	●	○	◐	◐	◕			
IBM Watson® Knowledge Catalog	●	◔	◐	◐	◕			
IBM InfoSphere IGC	●	○	○	◐	●			
Informatica Enterprise Data Catalog	●	◔	◐	◐	●			
Informatica Axon Data Governance	◕	◔	○	◐	●			

DATA CATALOGS AS PLATFORM FOR MANAGING DATA SUPPLY AND DEMAND

Reference model and market report (Version 1.0)

Tobias Korte[1], Martin Fadler[2], Markus Spiekermann[1], Prof. Christine Legner[2], Prof. Boris Otto[1]

[1] Fraunhofer ISST

[2] Competence Canter Corporate Data Quality (CC CDQ), University Lausanne

Fraunhofer ISST

CDQ

See also [DHBW-Moodle] the document “Value_Data-Catalog_eBook.pdf”

Market Study: Data Catalogues (2/2)

Product	DATA INVENTORY	DATA ANALYTICS	DATA COLLABORATION	DATA ASSESSMENT	DATA GOVERNANCE	DATA DISCOVERY	DATA VISUALIZATION	AUTOMATION & ML
Oracle Enterprise Metadata Mgmt.	●	○	◐	○	◐	◔	●	◐
Podium Data Market Place	●	◐	◕	◐	●			
SAP Information Steward	◕	○	○	◕	●			
SAP Data Hub	◕	◕	○	◐	◕			
Waterline Smart Data Catalog	●	◔	◐	◔	●			
Zaloni Data Management Platform	●	◐	●	◕	●			

DATA CATALOGS AS PLATFORM FOR MANAGING DATA SUPPLY AND DEMAND

Reference model and market report (Version 1.0)

Tobias Korte[1], Martin Fadler[2], Markus Spiekermann[1], Prof. Christine Legner[2], Prof. Boris Otto[1]

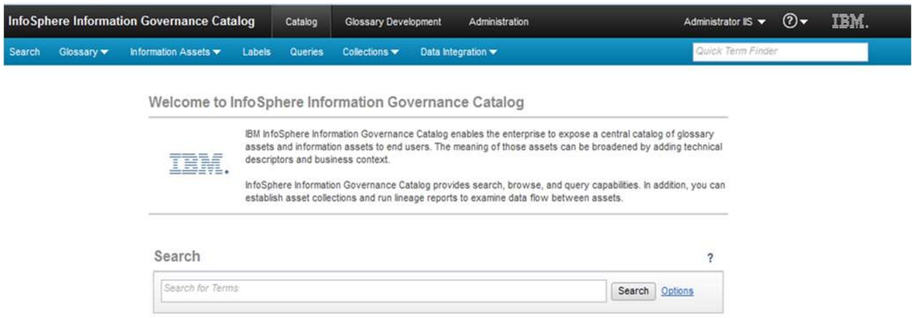
[1] Fraunhofer ISST

[2] Competence Canter Corporate Data Quality (CC CDQ), University Lausanne

Demo: IBM Information Governance Catalog (IGC)

- Allows you to understand where information came from and where it is used
- A key enabler to regulatory compliance and the IBM Data Governance Maturity Model

- Cross-tool reporting on:
  - Data movement and lineage
  - Business meaning
  - Impact of changes
  - Dependencies
  - Data lineage for BI Reports



Welcome to InfoSphere Information Governance Catalog

IBM InfoSphere Information Governance Catalog enables the enterprise to expose a central catalog of glossary assets and information assets to end users. The meaning of those assets can be broadened by adding technical descriptors and business context.

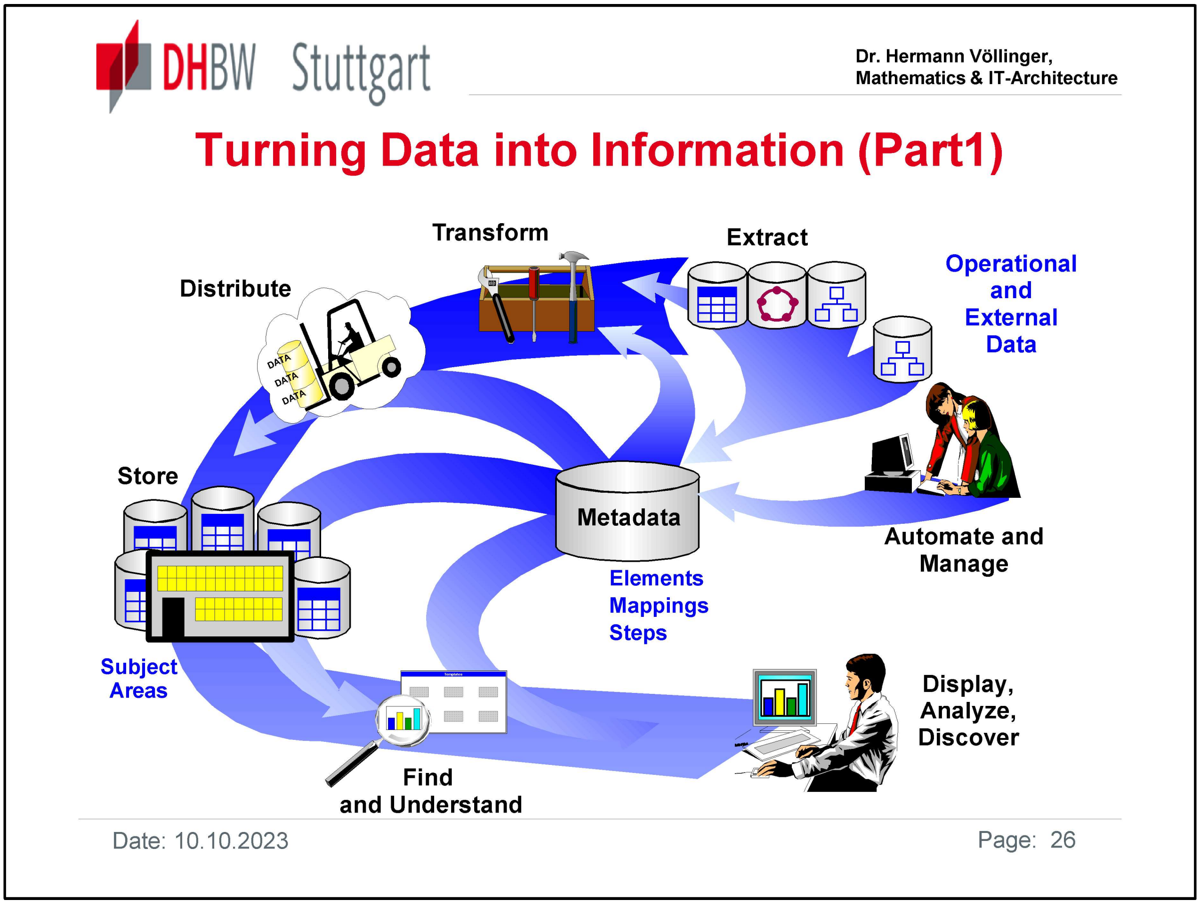
InfoSphere Information Governance Catalog provides search, browse, and query capabilities. In addition, you can establish asset collections and run lineage reports to examine data flow between assets.

IBM InfoSphere® Information Governance Catalog (IGC) is an interactive, web-based tool that enables users to create, manage, and share an enterprise vocabulary and classification system in a central catalog. It helps users to understand the business meaning of their assets and provides search, browse, and query capabilities. In addition, users can establish asset collections and run lineage reports to examine data flow between assets.

See the following YouTube video in See also in [DHBW-Moodle] the following documents and videos:

- “Guided IGC Demo.pdf”, “Guided IGC Demo - Deutsch.pdf”
- “Demo-IGC-Part1.mp4” & “Demo-IGC-Part2.mp4”
- “Demo_Qutput-IGC-Part2-Data_Lineage*.pdf”

See the value of having enterprise data assets cataloged in a central repository that identifies
relationships between data objects and business terms and metadata. Such a repository provides companies with a repeatable, consistent method of tracking data usage, quality, and lineage, in order to meet industry-specific regulatory requirements; i.e. understand the following features:
* Explore business terms, governance policies, and rules in IGC.
* Generate and review lineage analysis to validate and raise confidence in BI reports.
* Examine the rules that have been applied to the data used in the risk data aggregation process.



Turning Data into Information (Part1)

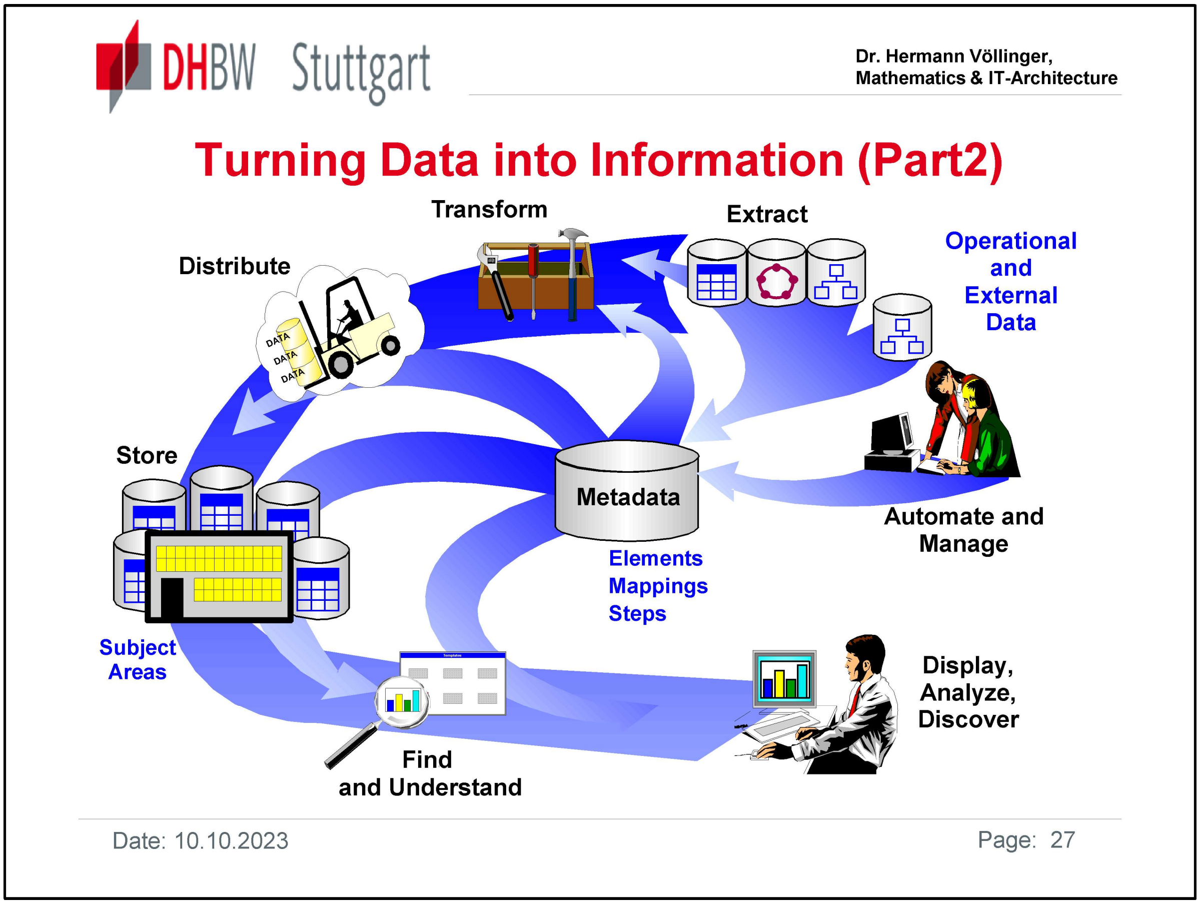
➢**In general, the data-to-information process takes the following steps:**

➢ Data is **extracted** from internal operational data and from data available in sources external to the enterprise. Administrative support is required to define the input data, transformation requirements, method of storing the data in the warehouse, and its presentation to the end user. Since this is an ongoing process, management and automation of the process tasks is important to minimize the burden of operations and have tools that make administration easy.

➢ Based on the types of information applications required, decisions will be made on how the data needs to be **transformed**, e.g. what operational data is not required and can be deleted, how data should be changed to make it understandable by end users. Some data may need to be enhanced (summaries rather than all the detailed data), and some might be aggregated from multiple different data sources into a single data element. Once transformed, data is placed in the warehouse data stores based on a selected business subject area structure.

➢ Data in a data warehouse is typically organized by "**subject areas**". Examples are "customer", "product", "employee", or functional business areas (Finance). With the data organized in the data warehouse in this way, it can be more consistent, at known levels of currency, minimally redundant, of better quality, and more easily available to end users for access and analysis.

➢ **Business views** of the data are developed as part of the structure and made available to users for easy access to the data for business purposes. By mapping the data to a business view and presenting it to the end user in the context of a chart, graph, or report, the data becomes information-information that can be used for better informed business decisions.



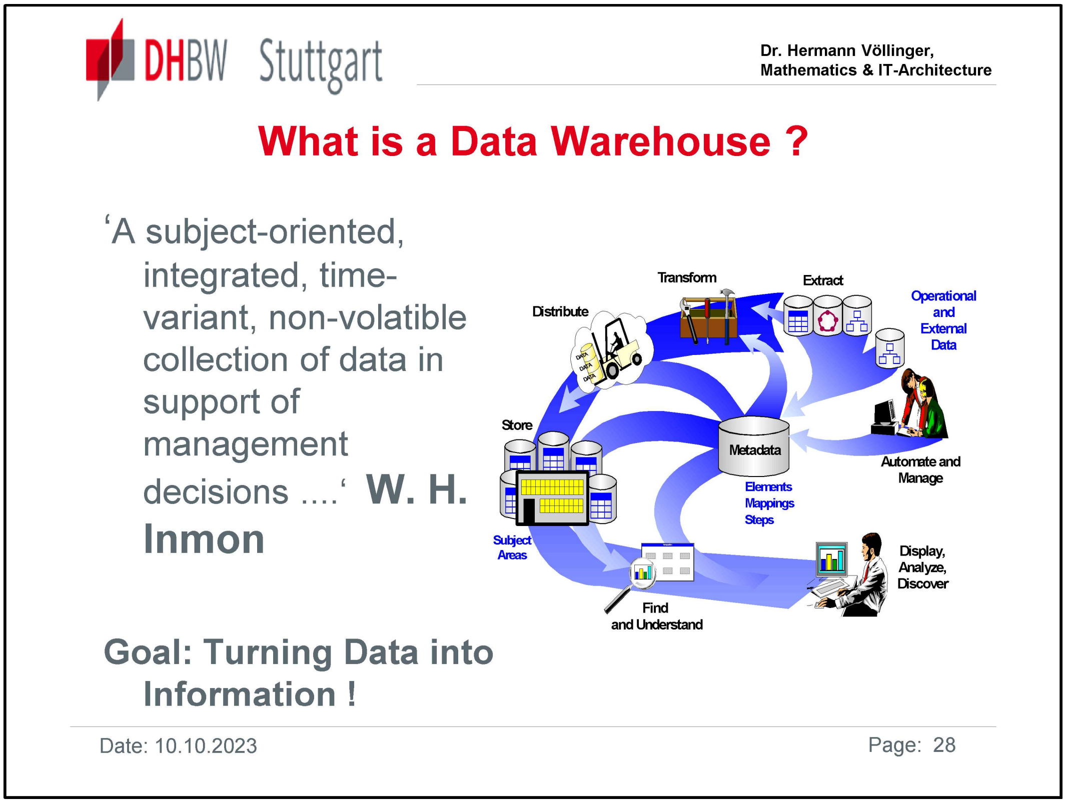
Turning Data into Information (Part2)

Definitions of the data elements, transformation processes, and business views are stored as **metadata:**

➢ **Technical Metadata** (data elements, transformation mappings) is used to build and maintain the data warehouse processes.

➢ **Business Metadata** is used to help end users understand what data is in the warehouse, in business terms they can understand.

➢ **A metadata store**, or information catalog, is built and maintained by the data warehouse administrators. Metadata management is key both for the processes involved in data warehousing and for enabling end users' access to the data.



What is a Data Warehouse ?

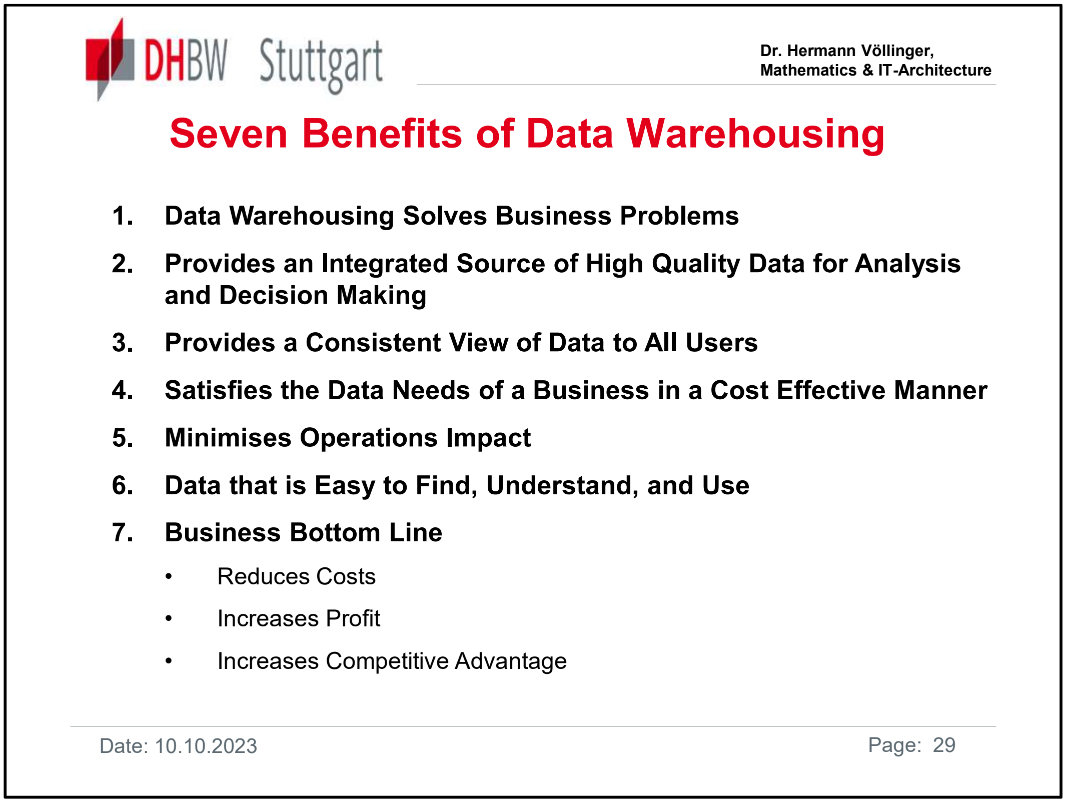
'A subject-oriented, integrated, time-variant, non-volatile collection of data in support of management decisions ....' **W. H. Inmon**

Goal: Turning Data into Information !

The slide shows also the Data Warehouse definition of W. H. Inmon, who is one of the most famous architects of Data Warehouse community.

Data in a data warehouse are build, created and transformed out of operational data such that they full-fill the following four attributes:

1. Subject-oriented
2. Integrated
3. Time-variant
4. Non-volatile



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Seven Benefits of Data Warehousing

1. **Data Warehousing Solves Business Problems**
2. **Provides an Integrated Source of High Quality Data for Analysis and Decision Making**
3. **Provides a Consistent View of Data to All Users**
4. **Satisfies the Data Needs of a Business in a Cost Effective Manner**
5. **Minimises Operations Impact**
6. **Data that is Easy to Find, Understand, and Use**
7. **Business Bottom Line**
   - Reduces Costs
   - Increases Profit
   - Increases Competitive Advantage

Date: 10.10.2023 Page: 29

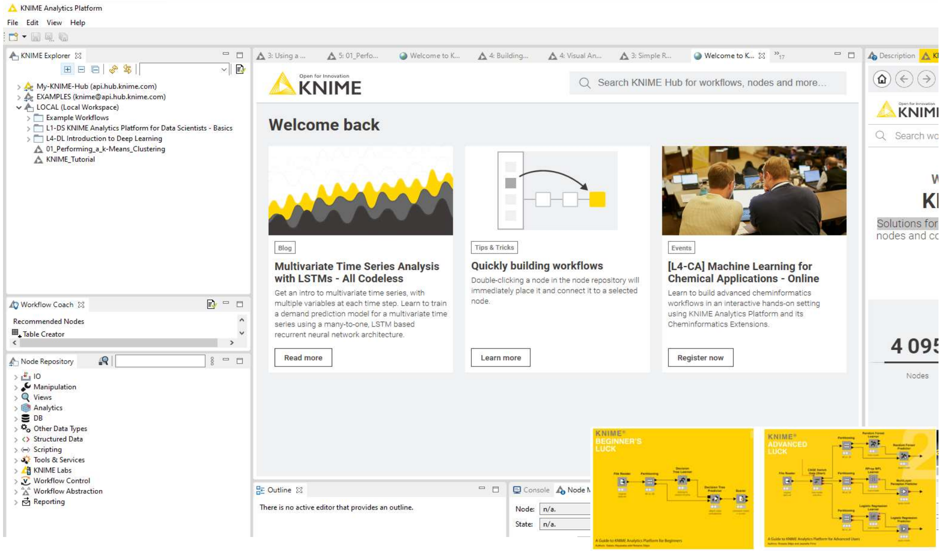
Data warehousing provides an excellent approach for transforming data into useful and reliable information to support the analytical and decision-making techniques that are so important in today's competitive environment, such as data mining, multidimensional data analysis, query and reporting. Making use of these techniques along with data warehousing can result in more informed decision-making capability and lead to significant business advantages:

1. Data Warehousing helps **solve business problems** by catering for the needs of business users
2. Provides **an integrated source** of high-quality data for analysis and decision making - population of the warehouse from a broad range of data sources (both operational and external) provides a single data store, with easy access to source, summarized and historical data, for more informed decision making.
3. Provides a **consistent view** of data to all users improving effectiveness
4. **Satisfies the business data needs** in a cost-effective manner by reducing I/S workload, automating data flow and process tasks to relieve I/S from many of the manual administrative and operations tasks, and significantly reducing the resources required to operate and maintain the data warehouse.
5. **Minimizes operations impact** by providing a separate physical data store optimized for informational applications.
6. **Data that is easy to find, understand, and use -** enables users to directly access the data to perform their own data analysis and create their own queries and reports by providing an information catalog which makes users aware of what data is in the warehouse, what it means in business terms, and how to use it.
7. **Business Bottom Line –** reduces costs, increase profit, increases advantage

DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Solution Platform for DWH/Analytics: KNIME



- My-KNIME-Hub (api.hub.knime.com)
- EXAMPLES (knime@api.hub.knime.com)
- LOCAL (Local Workspace)
  - Example Workflows
  - L1-DS KNIME Analytics Platform for Data Scientists - Basics
  - L4-DL Introduction to Deep Learning
  - 01_Performing_a_k-Means_Clustering
  - KNIME_Tutorial

Recommended Nodes
Table Creator

- IO
- Manipulation
- Views
- Analytics
- DB
- Other Data Types
- Structured Data
- Scripting
- Tools & Services
- KNIME Labs
- Workflow Control
- Workflow Abstraction
- Reporting

There is no active editor that provides an outline.

Open for Innovation KNIME

Search KNIME Hub for workflows, nodes and more...

Welcome back

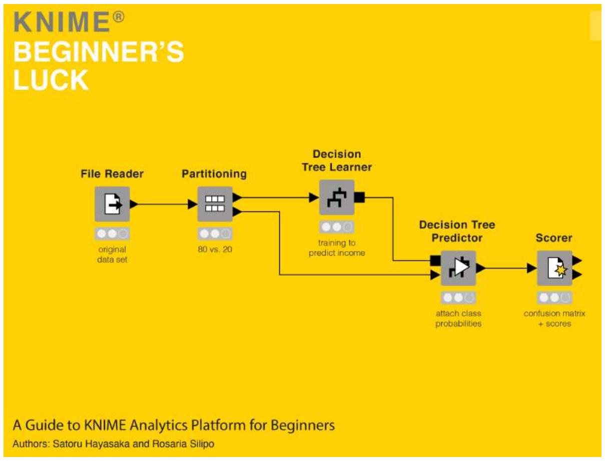
Blog
Multivariate Time Series Analysis with LSTMs - All Codeless
Get an intro to multivariate time series, with multiple variables at each time step. Learn to train a demand prediction model for a multivariate time series using a many-to-one, LSTM based recurrent neural network architecture.
Read more

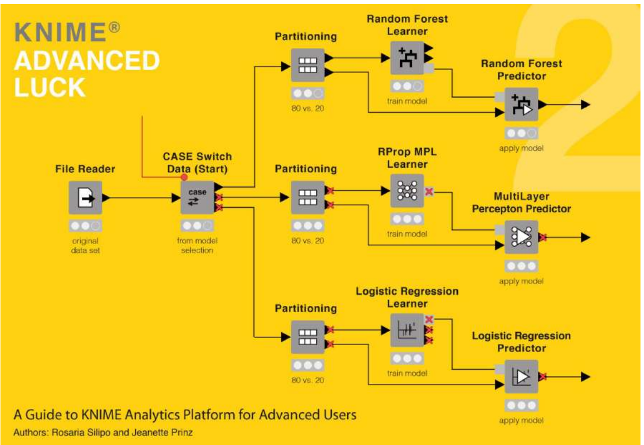
Tips & Tricks
Quickly building workflows
Double-clicking a node in the node repository will immediately place it and connect it to a selected node.
Learn more

Events
[L4-CA] Machine Learning for Chemical Applications - Online
Learn to build advanced cheminformatics workflows in an interactive hands-on setting using KNIME Analytics Platform and its Cheminformatics Extensions.
Register now

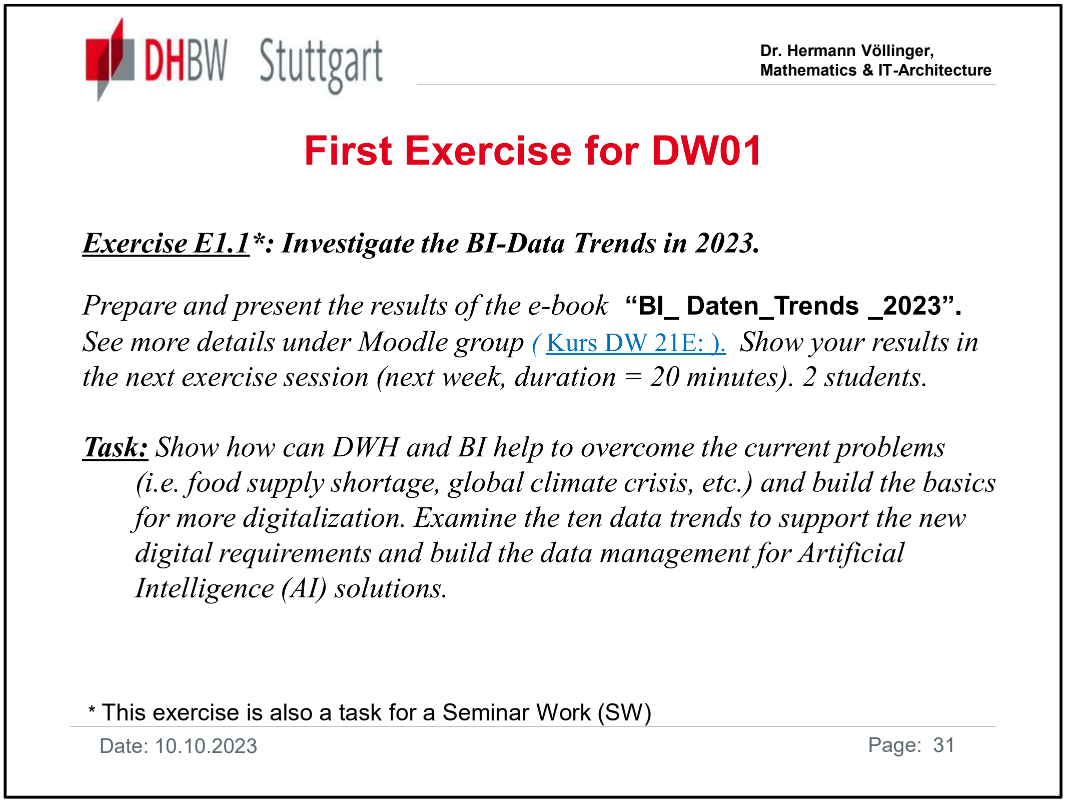
Date: 10.10.2023 Page: 30

KNIME is also used by DHBW Lörrach (working group of Prof. Dr. Klemens Schnattinger). See the following YouTube video about Data Science solutions implemented in KNIME: https://mediaserver.dhbw-loerrach.de/permalink/v1261a6b6c5b5783vmdf/iframe/?login=no
The installation and the basic features of KMIME are described in the “KNIME-BeginnersGuide.pdf”. The advanced features of KNIME are described in the “KNIME-AdvancedUserGuide.pdf” and KNIME_Workbench_Guide.pdf”. All three documents can be found in the first category of the “Supporting Information for DWH Lecture” in the Course-Moodle: Kurs DW 21E


KNIME® BEGINNER'S LUCK
File Reader
original data set
Partitioning
80 vs. 20
Decision Tree Learner
training to predict income
Decision Tree Predictor
attach class probabilities
Scorer
confusion matrix + scores
A Guide to KNIME Analytics Platform for Beginners
Authors: Satoru Hayasaka and Rosaria Silipo



KNIME will be used in 4 exercises: 1. Installation and first insights; 2. Data Manipulation & Aggregation; 3. Performing a KNIME Basics solution and 4. Deep Learning - Image Classification (MNIST Solution).



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

First Exercise for DW01

Exercise E1.1: Investigate the BI-Data Trends in 2023.*

Prepare and present the results of the e-book **"BI_ Daten_Trends _2023".** *See more details under Moodle group (Kurs DW 21E:). Show your results in the next exercise session (next week, duration = 20 minutes). 2 students.*

Task: *Show how can DWH and BI help to overcome the current problems (i.e. food supply shortage, global climate crisis, etc.) and build the basics for more digitalization. Examine the ten data trends to support the new digital requirements and build the data management for Artificial Intelligence (AI) solutions.*

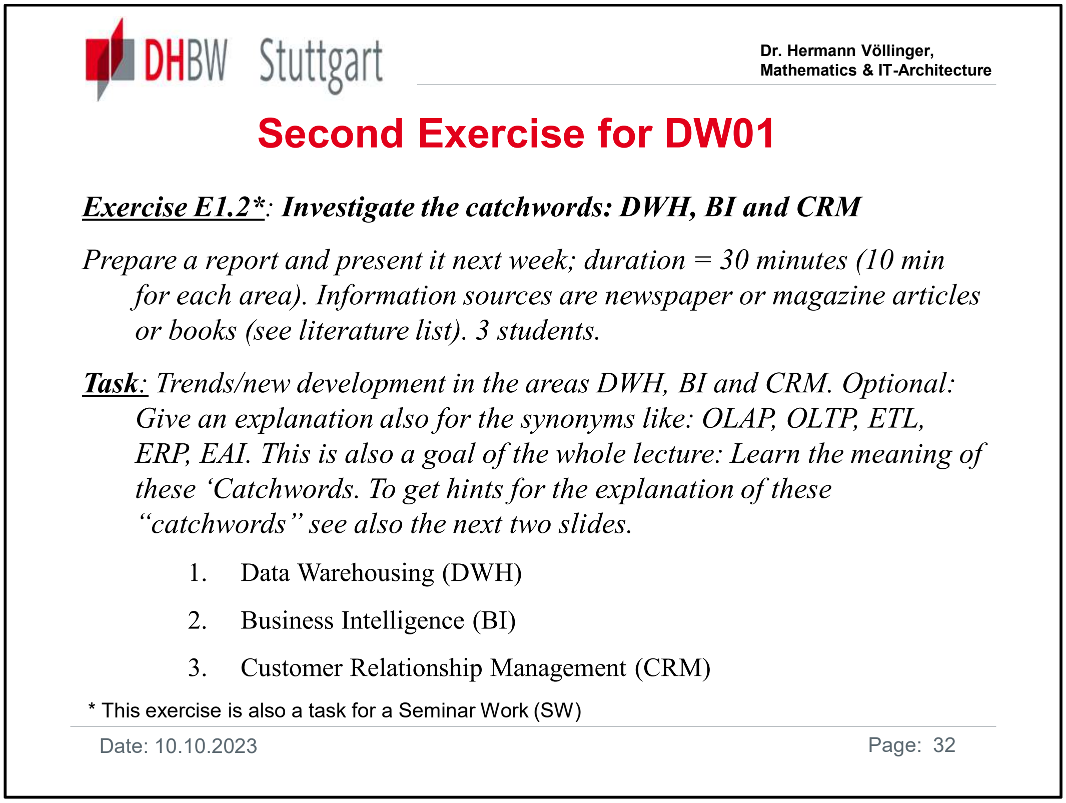
* This exercise is also a task for a Seminar Work (SW)

Date: 10.10.2023 Page: 31

Hint to E1.1: See the comments given in the lesson…

Further Information:

1. So kann Business Intelligence von KI profitieren (bi-scout.com)
2. Mündlein, U.: *Was bringt die Zukunft?* Von CRM Manager: https://www.crmmanager.de/crm-trends-2019-was-bringt-die-zukunft
3. Balentine, C. : *Software Magazine*. Von http://www.softwaremag.com/multi-cloud-strategies-for-bi
4. See also: BI_Daten_Trends_2022.pdf in Moodle/…Category1



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Second Exercise for DW01

Exercise E1.2*: Investigate the catchwords: DWH, BI and CRM*

Prepare a report and present it next week; duration = 30 minutes (10 min for each area). Information sources are newspaper or magazine articles or books (see literature list). 3 students.

***Task**: Trends/new development in the areas DWH, BI and CRM. Optional: Give an explanation also for the synonyms like: OLAP, OLTP, ETL, ERP, EAI. This is also a goal of the whole lecture: Learn the meaning of these 'Catchwords. To get hints for the explanation of these "catchwords" see also the next two slides.*

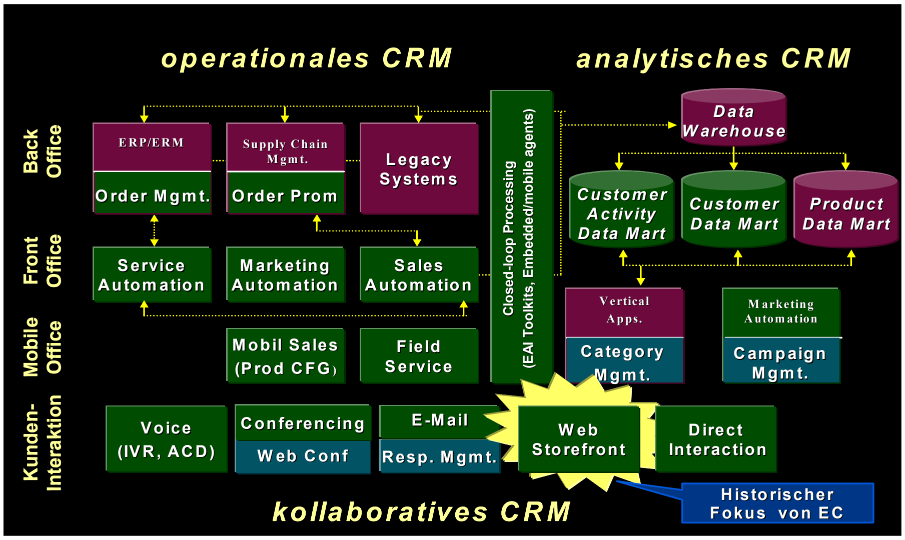
1. Data Warehousing (DWH)
2. Business Intelligence (BI)
3. Customer Relationship Management (CRM)

* This exercise is also a task for a Seminar Work (SW)

Further Information:

1. Litzl, N.: *Was ist Business Intelligence – BI?* Von BigData Insider: https://www.bigdata-insider.de/was-ist-business-intelligence-bi-a-563185/
2. Luber, S., & Litzel, N. : *Was ist ein Data Warehouse?* Von BigData Insider: https://www.bigdata-insider.de/was-ist-ein-data-warehouse-a-606701
3. Böttcher, G.: *CRM im Trend*. Von Springer Professional: https://www.springerprofessional.de/crm/vertriebssteuerung/crm-im-trend/16522806

Hints to E1.2: The BI / CRM Topology



operationales CRM

analytisches CRM

Back Office

Front Office

Mobile Office

Kunden-Interaktion

ERP/ERM

Order Mgmt.

Supply Chain Mgmt.

Order Prom

Legacy Systems

Service Automation

Marketing Automation

Sales Automation

Mobil Sales (Prod CFG)

Field Service

Closed-loop Processing (EAI Toolkits, Embedded/mobile agents)

Data Warehouse

Customer Activity Data Mart

Customer Data Mart

Product Data Mart

Vertical Apps.

Category Mgmt.

Marketing Automation

Campaign Mgmt.

Voice (IVR, ACD)

Conferencing

Web Conf

E-Mail

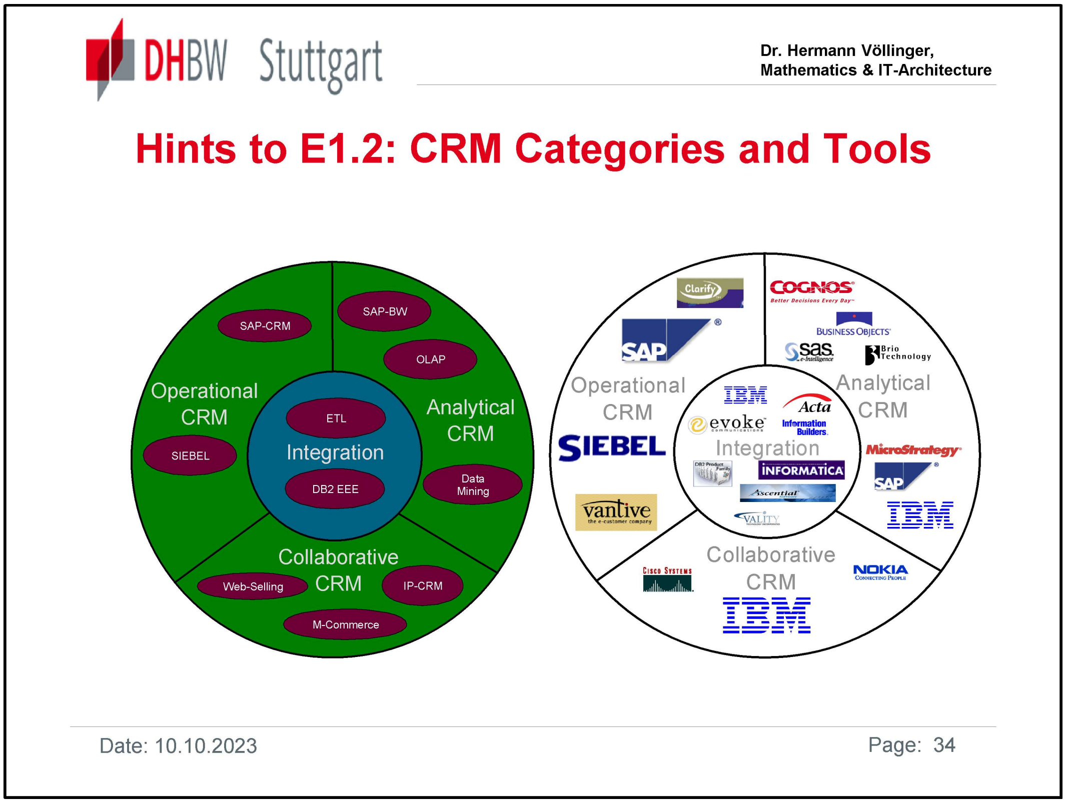
Resp. Mgmt.

Web Storefront

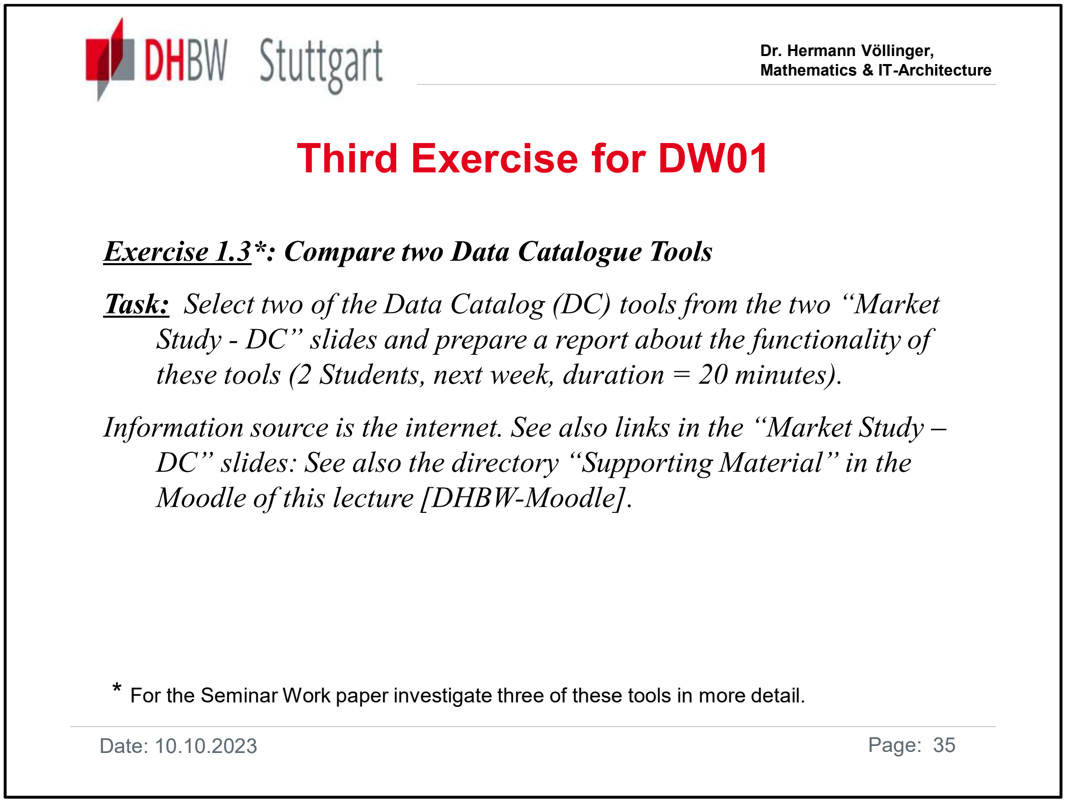
Direct Interaction

Historischer Fokus von EC

kollaboratives CRM



Hints to E1.2: CRM Categories and Tools



Third Exercise for DW01

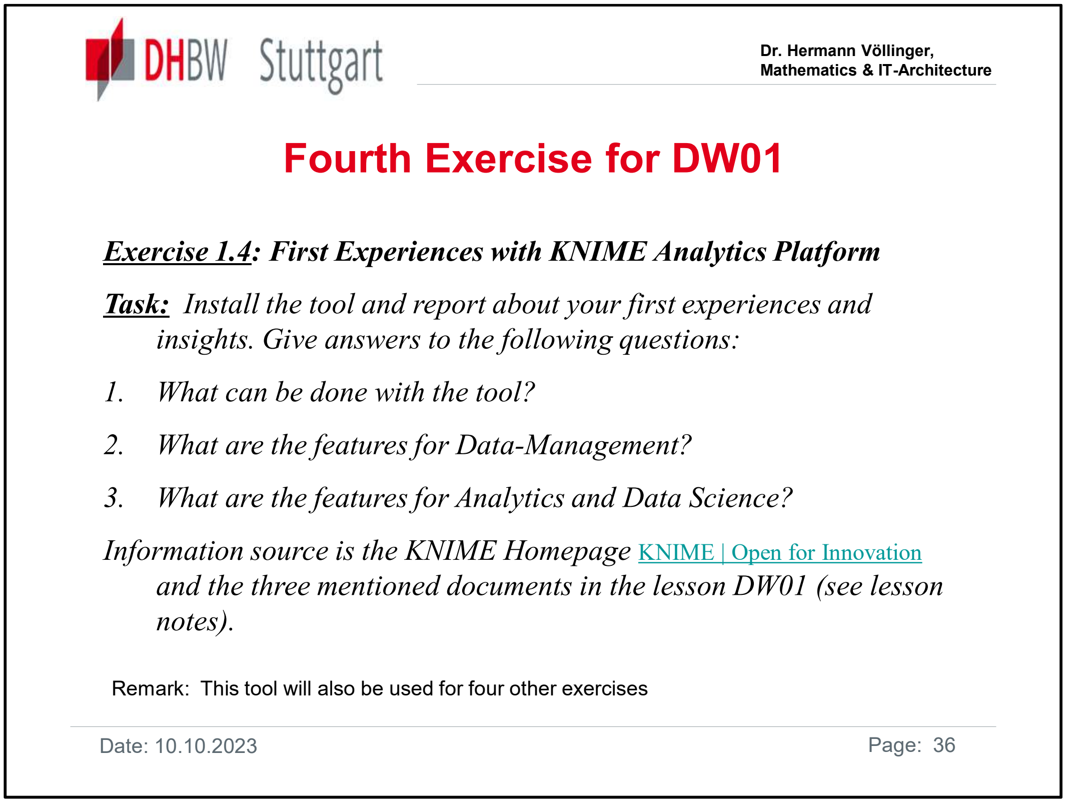
Exercise 1.3**: Compare two Data Catalogue Tools

Task: *Select two of the Data Catalog (DC) tools from the two "Market Study - DC" slides and prepare a report about the functionality of these tools (2 Students, next week, duration = 20 minutes).*

Information source is the internet. See also links in the "Market Study – DC" slides: See also the directory "Supporting Material" in the Moodle of this lecture [DHBW-Moodle].

* For the Seminar Work paper investigate three of these tools in more detail.

See also the information and links in the previous corresponding lecture slides.



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Fourth Exercise for DW01

***Exercise 1.4**: First Experiences with KNIME Analytics Platform*

***Task:** Install the tool and report about your first experiences and insights. Give answers to the following questions:*

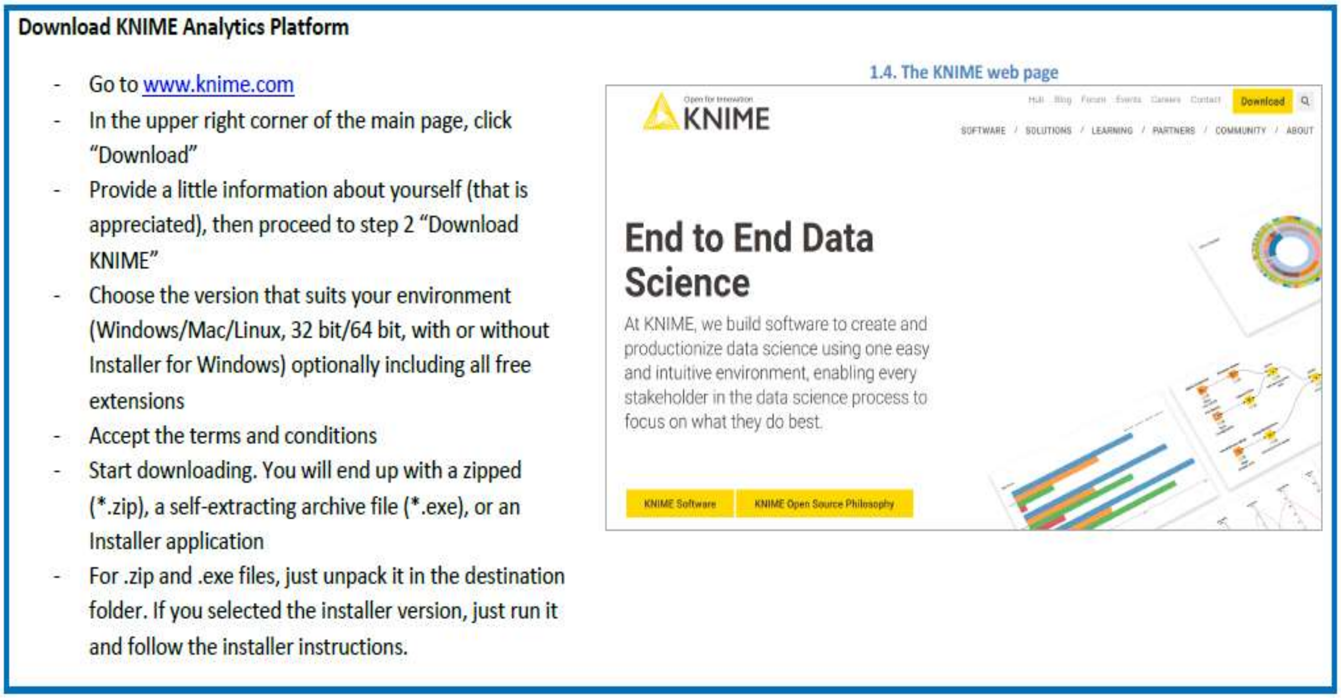
1. *What can be done with the tool?*
2. *What are the features for Data-Management?*
3. *What are the features for Analytics and Data Science?*

Information source is the KNIME Homepage KNIME | Open for Innovation *and the three mentioned documents in the lesson DW01 (see lesson notes).*

Remark: This tool will also be used for four other exercises

Date: 10.10.2023 Page: 36

The installation of KMIME is described in the “KNIME-BeginnersGuide.pdf”.
The document can be found in the first category of the “Supporting Information for DWH Lecture” in the Course-Moodle: Kurs DW 21E



Download KNIME Analytics Platform

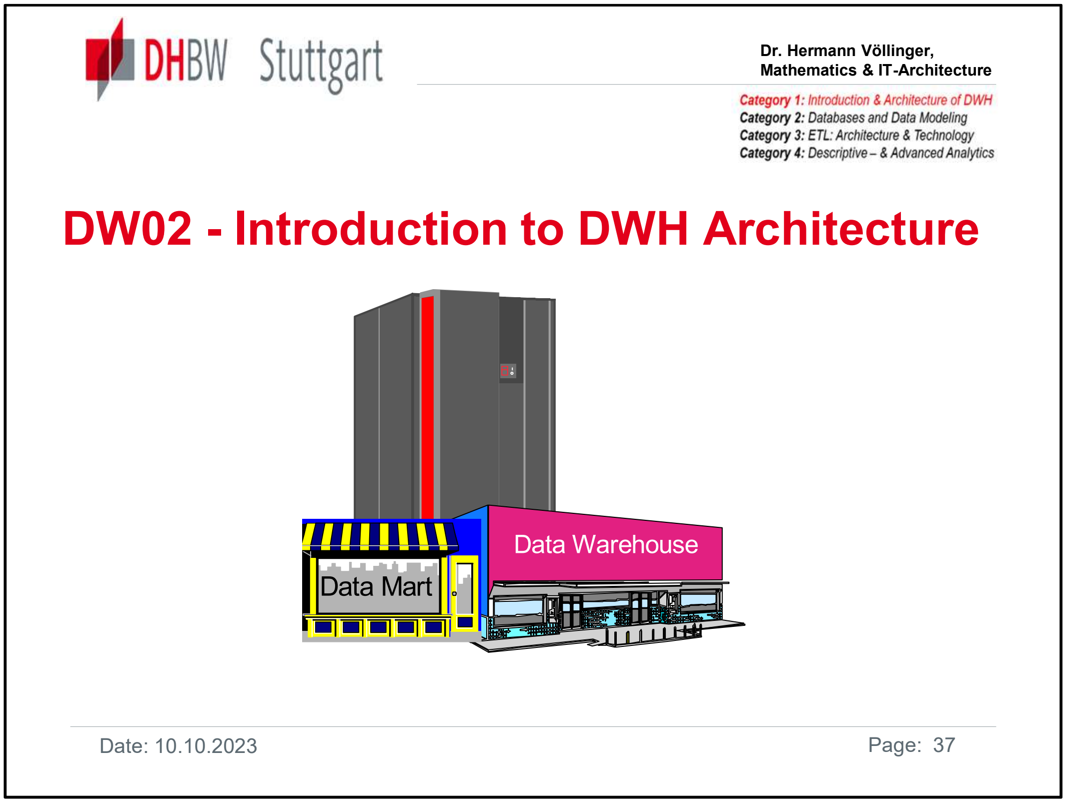
- Go to www.knime.com
- In the upper right corner of the main page, click “Download”
- Provide a little information about yourself (that is appreciated), then proceed to step 2 “Download KNIME”
- Choose the version that suits your environment (Windows/Mac/Linux, 32 bit/64 bit, with or without Installer for Windows) optionally including all free extensions
- Accept the terms and conditions
- Start downloading. You will end up with a zipped (*.zip), a self-extracting archive file (*.exe), or an Installer application
- For .zip and .exe files, just unpack it in the destination folder. If you selected the installer version, just run it and follow the installer instructions.

1.4. The KNIME web page

End to End Data Science

At KNIME, we build software to create and productionize data science using one easy and intuitive environment, enabling every stakeholder in the data science process to focus on what they do best.

Open for Innovation
KNIME
Hub Blog Forum Events Careers Contact
Download
SOFTWARE / SOLUTIONS / LEARNING / PARTNERS / COMMUNITY / ABOUT
KNIME Software
KNIME Open Source Philosophy

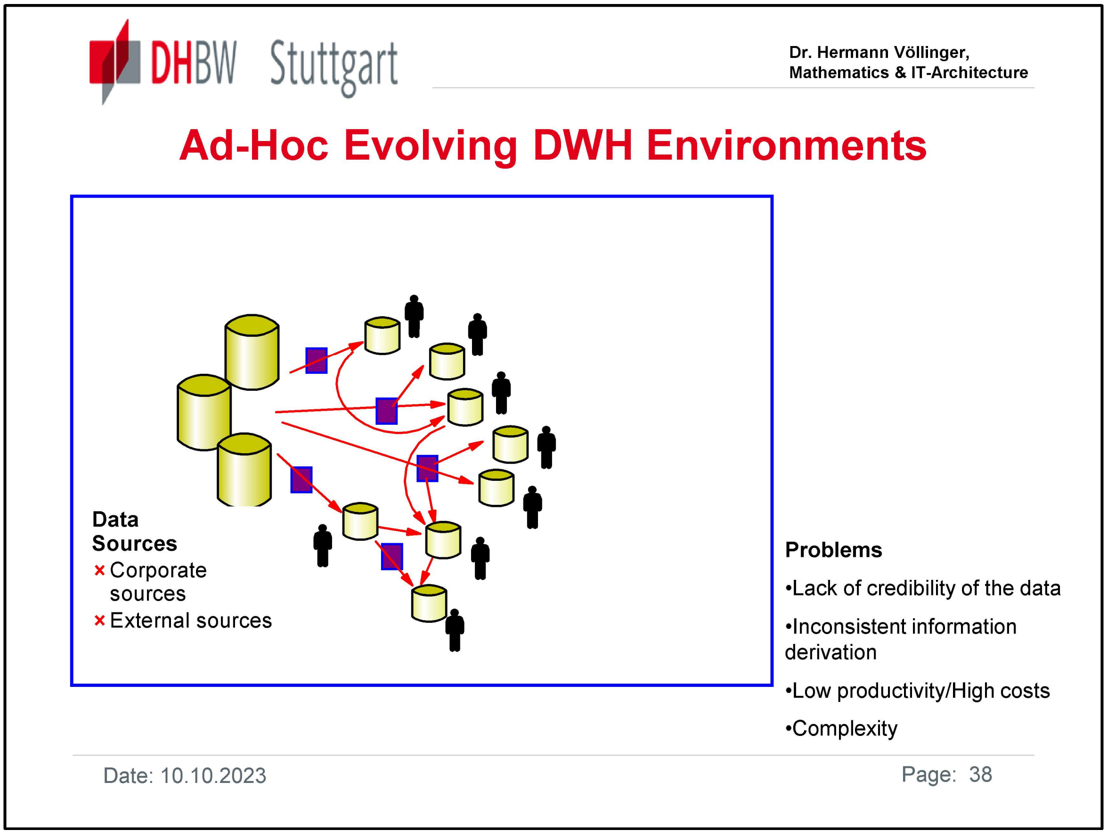


DW02 - Introduction to DWH Architecture

The following chapter gives an introduction into the architecture of a data warehouse. It explains the architectural ideas behind a Data Warehouse solution. It shows aspects of DWH architecture. It explains the different architectural concepts. What can go wrong, etc…..

In especially the following topics are covered:

- “Setting the Scene” with several steps on the way to a Central Data Warehouse (CDW)
- The possible architectural approaches from 0-Tier DWH to a 2-Tier DWH
- The “Big Picture” of a DWH architecture
- Not using architectural concepts: “What can Go Wrong?”
- The Data Warehouse Data Layers
- The IBM DWH Reference Architecture
- Example of a DWH Architecture for a Financial Market customer



Ad-Hoc Evolving DWH Environments

Data Sources
- × Corporate sources
- × External sources

Problems

•Lack of credibility of the data

•Inconsistent information derivation

•Low productivity/High costs

•Complexity

Date: 10.10.2023 Page: 38

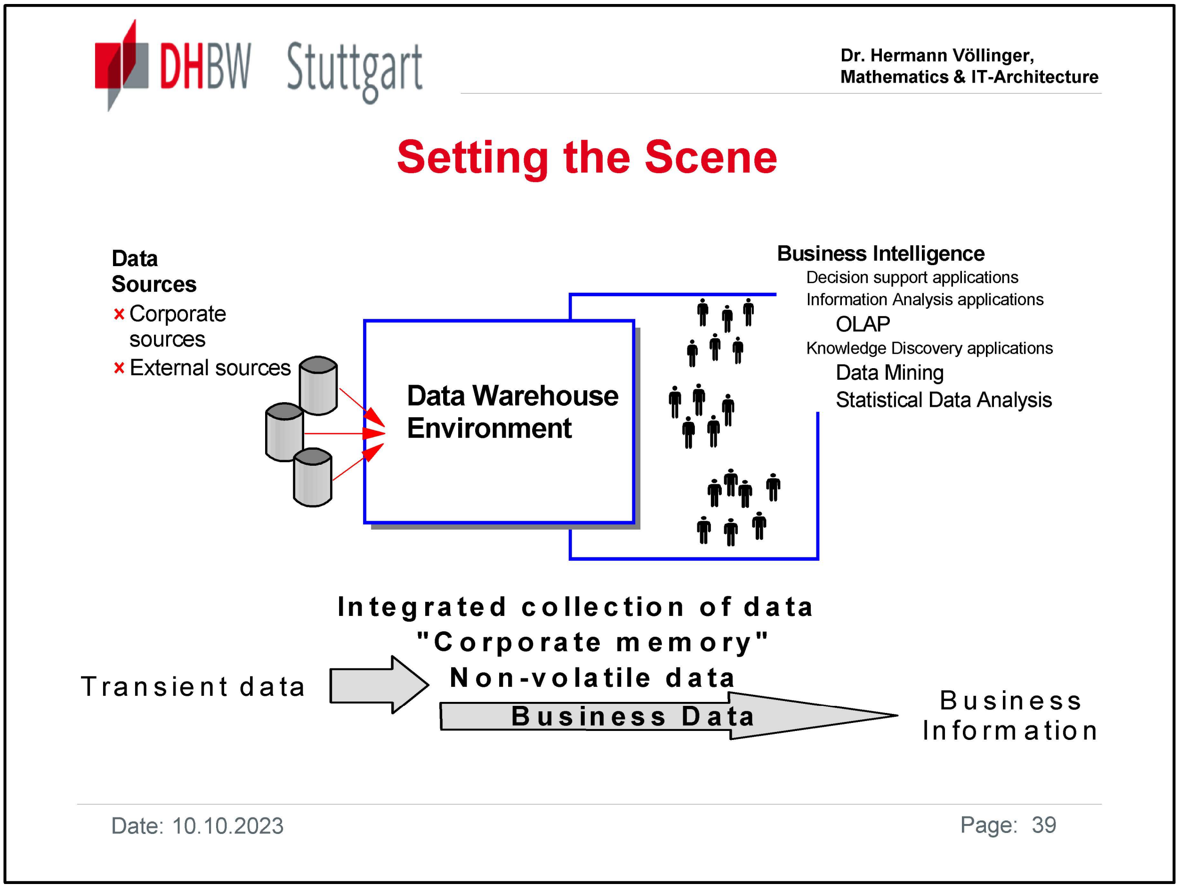
The slide shows the typical situation, which you will see in typical customer environments. This current scenario is build without control over several years in the enterprise..

Special solutions, special data population processes are developed for independent data marts over the years.

You can also see on the slide, that the ETL processes even get data from other data marts ---→ complex process management.

Summary: The 5 main problems are:

1. Lack of credibility in the data
2. **2.** Inconsistent data derivation
3. Complexity of processes
4. Low productivity
5. High costs



Setting the Scene

Data Sources

- × Corporate sources
- × External sources

Data Warehouse Environment

Business Intelligence

Decision support applications
Information Analysis applications
OLAP
Knowledge Discovery applications
Data Mining
Statistical Data Analysis

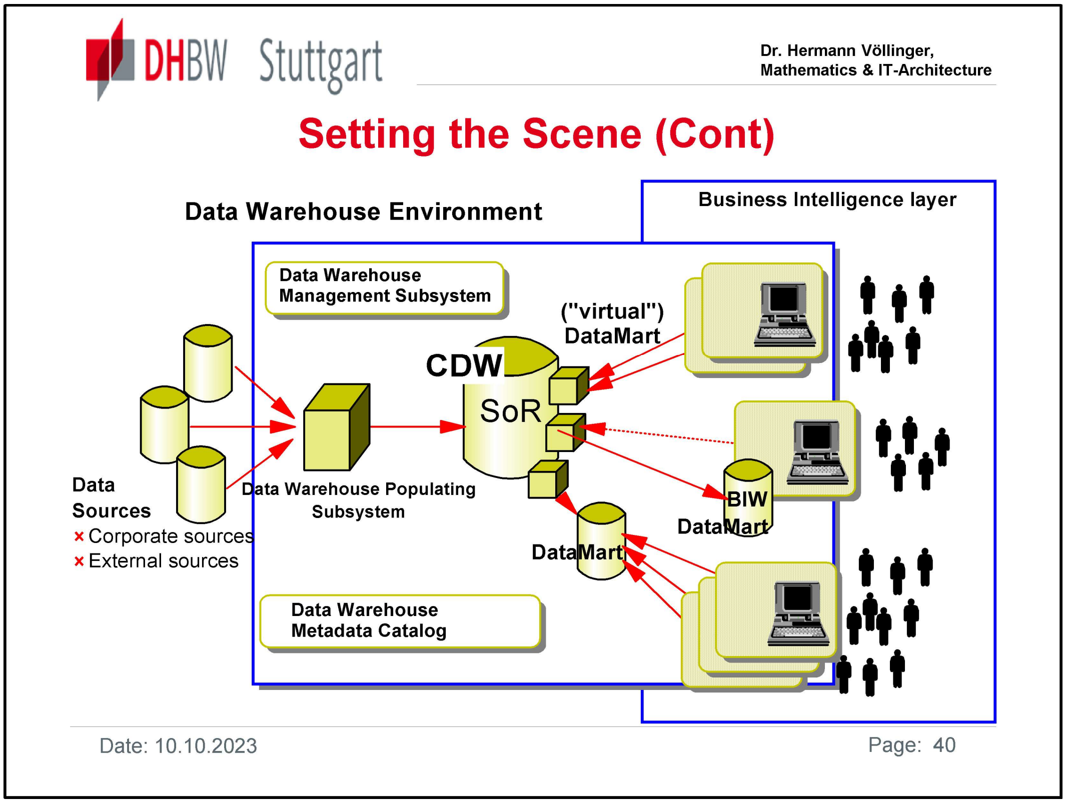
Integrated collection of data
"Corporate memory"
Non-volatile data

Transient data → **Business Data** → Business Information

The picture show the first step – building a integrated enterprise wide data DWH. This results in a data pool with ‘good‘ data..

Main problems and goals which have to be solved for this step are:

- ➢ Storage of Data (non-volatile)
- ➢ Structure the data (data model)
- ➢ Integrate data from different sources
- ➢ Build historical data (historical data model)
- ➢ Consistent data (run check and plausibility processes)
- ➢ Access to the data (create an user- and security-concept)
- ➢ Build a metadata repository with technical and business metadata



Dr. Hermann Völlinger, Mathematics & IT-Architecture

Setting the Scene (Cont)

Data Warehouse Environment

Business Intelligence layer

Data Warehouse Management Subsystem

("virtual") DataMart

CDW

SoR

Data Sources

- × Corporate sources
- × External sources

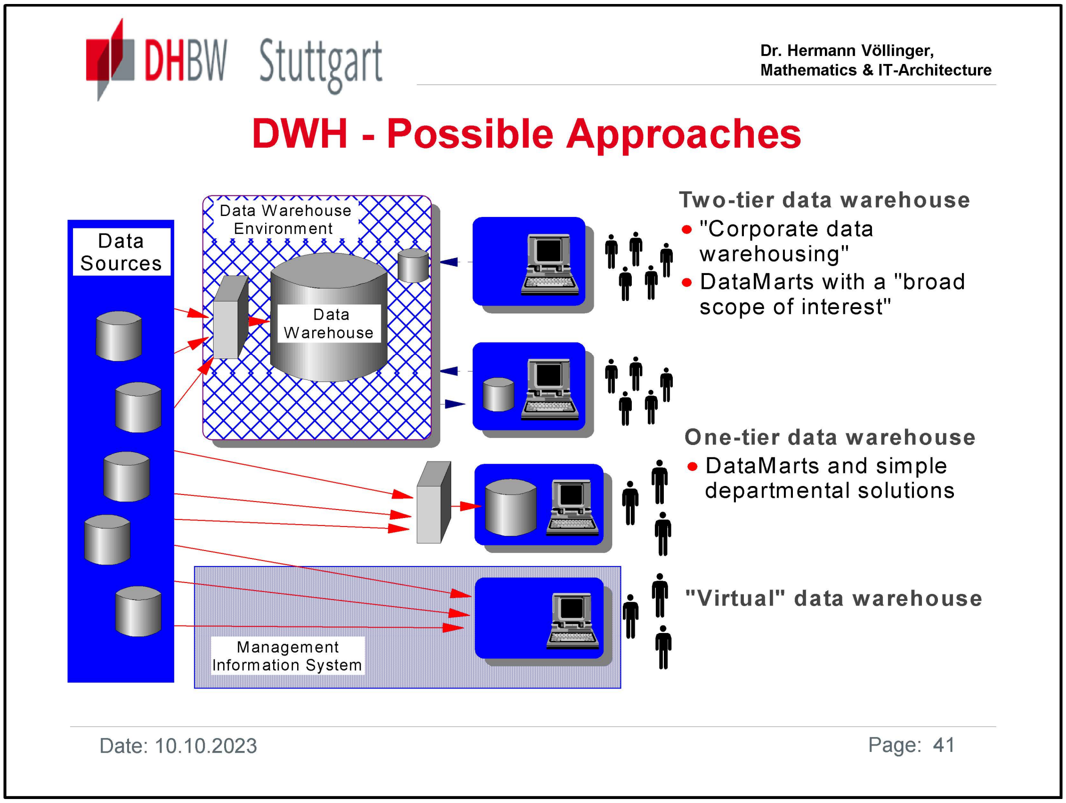
Data Warehouse Populating Subsystem

BIW DataMart

DataMart

Data Warehouse Metadata Catalog

See the next step in this slide:
After we have build the Central Data Warehouse (CDW), which has all detail data in it (therefore also called System of Record –SoR). Data Marts are build for special applications and user groups.



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

DWH - Possible Approaches

Two-tier data warehouse

- "Corporate data warehousing"
- DataMarts with a "broad scope of interest"

One-tier data warehouse

- DataMarts and simple departmental solutions

"Virtual" data warehouse

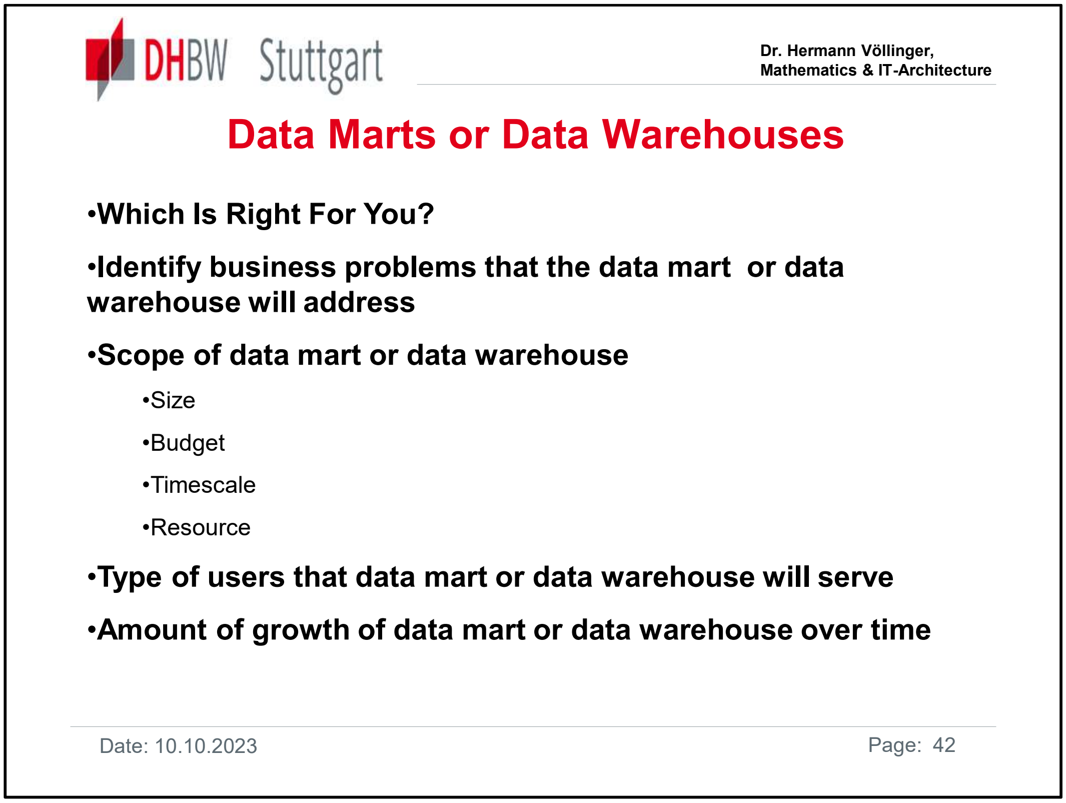
Date: 10.10.2023 Page: 41

What are the possible architectural approaches: at the planning stage in a data warehouse project, it is important to decide what the scope of the warehouse is going to be. Historically data warehouse implementations have been global or enterprise-wide in scope, but this does not necessarily have to be the case. There are 3 main implementation alternatives for a DWH:

➢ **Global Warehouse (2-tier architecture)** -here the primary responsibility for the overall architecture and implementation belongs to the IT department. Requirements and implementation priority would be based on the needs of the enterprise as a whole. This global warehouse could be physically centralized, decentralized, or logically centralized and physically distributed over multiple platforms and locations. The design could include support for any number of data marts, but these would not be the same as the stand-alone and dependent data marts already mentioned. The data marts are designed specifically to be part of the global data warehouse and are populated from it.

➢ **Stand-alone Data Mart (1-tier architecture)** - enables a department/workgroup to implement a data mart covering one particular business subject area, e.g. sales and marketing, or finance, with minimal, or no, impact on the IT department. It may require some technical skills, but these could be managed by the workgroup. This approach could also apply to a smaller organization that might not have the support of the IT department.

➢ **Virtual Data Warehouse (0-tier architecture) -** similar to the stand-alone data mart, except that data is not stored in an own database. The data from the source system are visualized / accessed by the applications. The data connectivity to the data sources managed by the IT dep. is required. These connectivity impacts the operational system. But implementation should still have minimal impact on the IT department. The department decides what data is accessed, the frequency of access, and may even provide tools and skills necessary to extract the data.



Data Marts or Data Warehouses

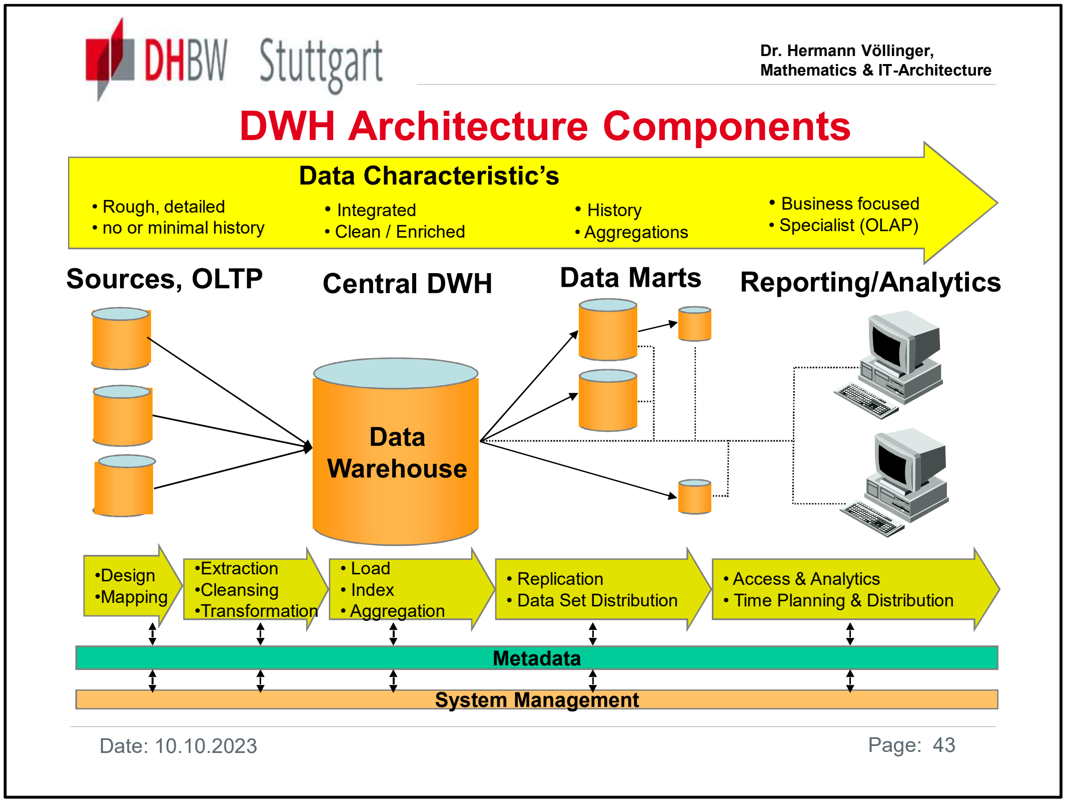
- **Which Is Right For You?**
- **Identify business problems that the data mart or data warehouse will address**
- **Scope of data mart or data warehouse**
  - Size
  - Budget
  - Timescale
  - Resource
- **Type of users that data mart or data warehouse will serve**
- **Amount of growth of data mart or data warehouse over time**

Providing a choice of implementation leaves some companies wondering whether it would be best to start with a data mart and build up to a data warehouse, or go the opposite route and build a data warehouse that can be broken down into data marts.

➢ **Data marts work like data warehouses** - moving data from source to target databases. But data marts are different than data warehouses, because they serve specific business purposes/solve particular business problems. Although they still have to collect, transform and clean data, there's less of it, so therefore you'll spend less money, less time, and fewer resources to put data marts in place and maintain them.

➢ The first step is to **identify the business problems** that need solving by the data warehouse and find the best solution to these business needs. The other main considerations are those shown in this slide. Of course, if your decision has been to start with a data mart, you can still evolve to a data warehouse. Note, however, that building a warehouse from several data mart requires strategic planning. One way may be to start with a small number of data marts, and spend up-front time building the architecture for the data warehouse as you build the data marts.

➢ You can get up and running quickly now with data marts, and can evolve over time to include the tools you need to grow your data mart to a full-scale warehousing solution for your enterprise, as long as you make the right planning decisions at the start.



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

DWH Architecture Components

Data Characteristic's

- Rough, detailed
- no or minimal history

- Integrated
- Clean / Enriched

- History
- Aggregations

- Business focused
- Specialist (OLAP)

Sources, OLTP

Central DWH

Data Marts

Reporting/Analytics

Data Warehouse

- Design
- Mapping

- Extraction
- Cleansing
- Transformation

- Load
- Index
- Aggregation

- Replication
- Data Set Distribution

- Access & Analytics
- Time Planning & Distribution

Metadata

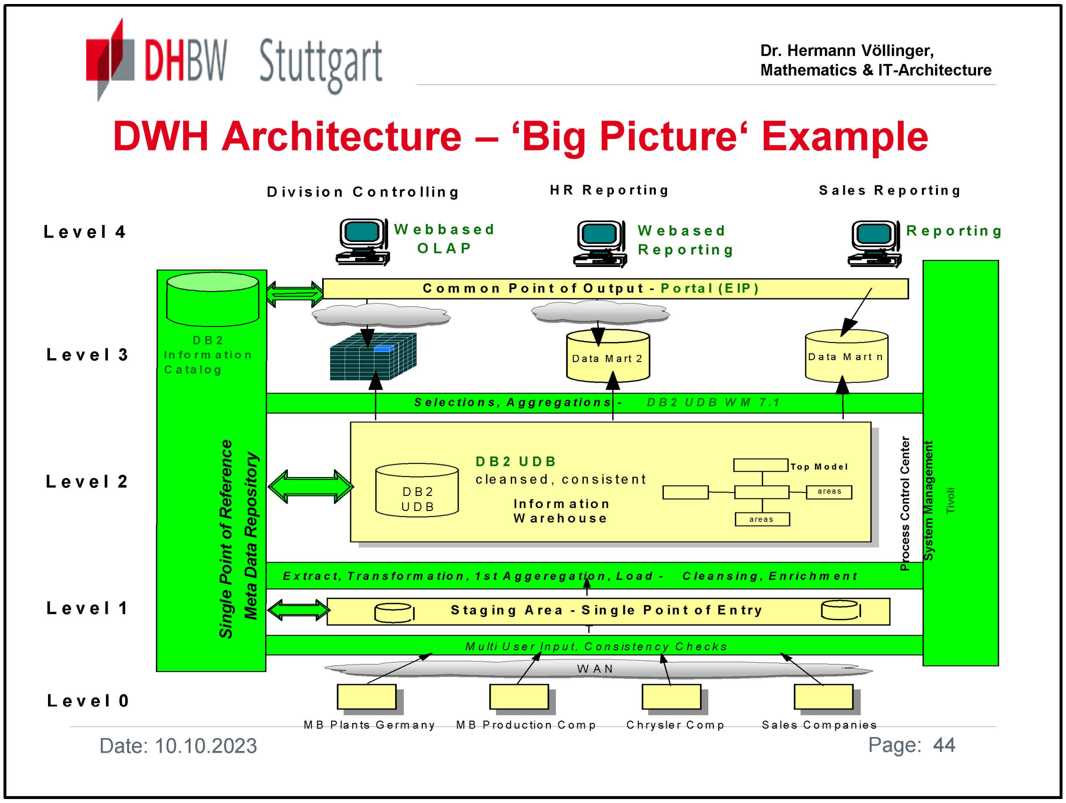
System Management

Date: 10.10.2023 Page: 43

The slide show now the full scope architecture from both sides:

- Processes workflow (see bottom line)
- Data characteristics (see top line)

As general layers over the DWH, we see the Meta Data Layer and the System Management Layer.

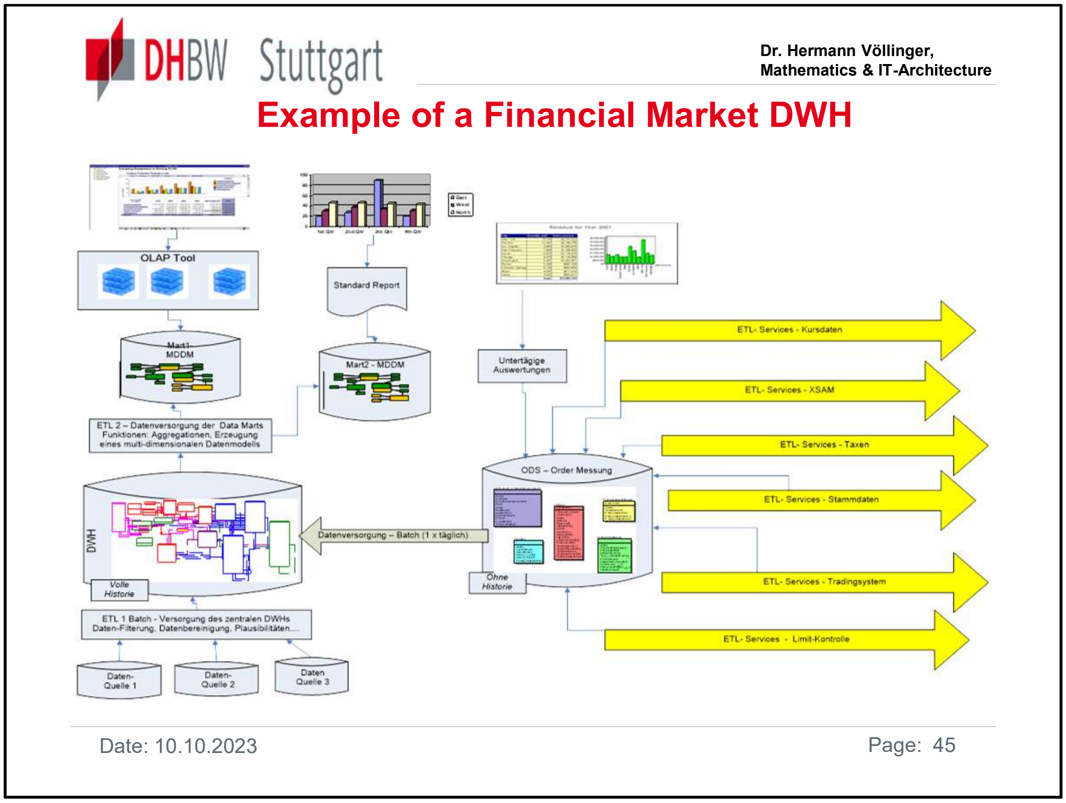


DWH Architecture – ‘Big Picture‘ Example

Shows the same as before, but with ‘bottom-top‘ approach of the layers.

It includes also the names of some concrete tools.

The image can be used as a blueprint for a concrete DWH



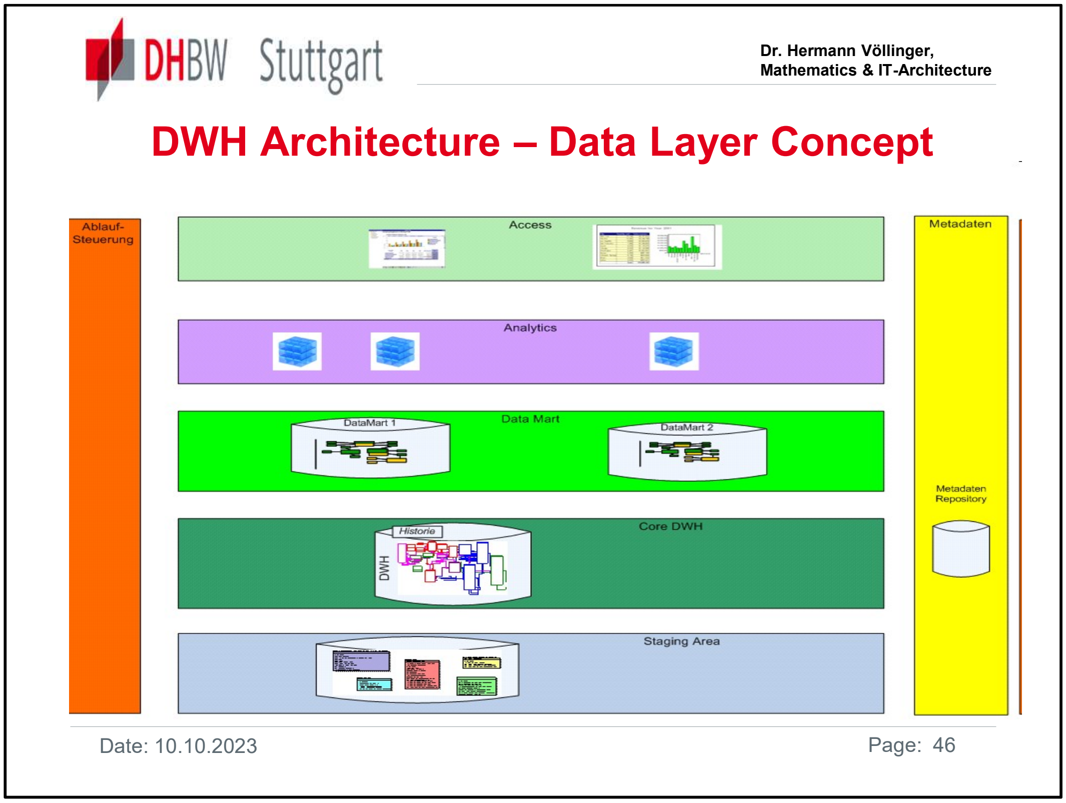
Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Example of a Financial Market DWH

100
80
60
40
20
0
1st Qtr
2nd Qtr
3rd Qtr
4th Qtr
East
West
North
Revenue for Year 2001
OLAP Tool
Standard Report
Mart1 MDDM
Mart2 - MDDM
Untertägige Auswertungen
ETL- Services - Kursdaten
ETL- Services - XSAM
ETL- Services - Taxen
ETL 2 – Datenversorgung der Data Marts Funktionen: Aggregationen, Erzeugung eines multi-dimensionalen Datenmodells
ODS – Order Messung
ETL- Services - Stammdaten
DWH
Datenversorgung – Batch (1 x täglich)
Volle Historie
Ohne Historie
ETL- Services - Tradingsystem
ETL 1 Batch - Versorgung des zentralen DWHs Daten-Filterung, Datenbereinigung, Plausibilitäten....
ETL- Services - Limit-Kontrolle
Daten-Quelle 1
Daten-Quelle 2
Daten Quelle 3

Date: 10.10.2023 Page: 45

Another concrete blueprint of a Financial Data Warehouse (“Stock Market”)



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

DWH Architecture – Data Layer Concept

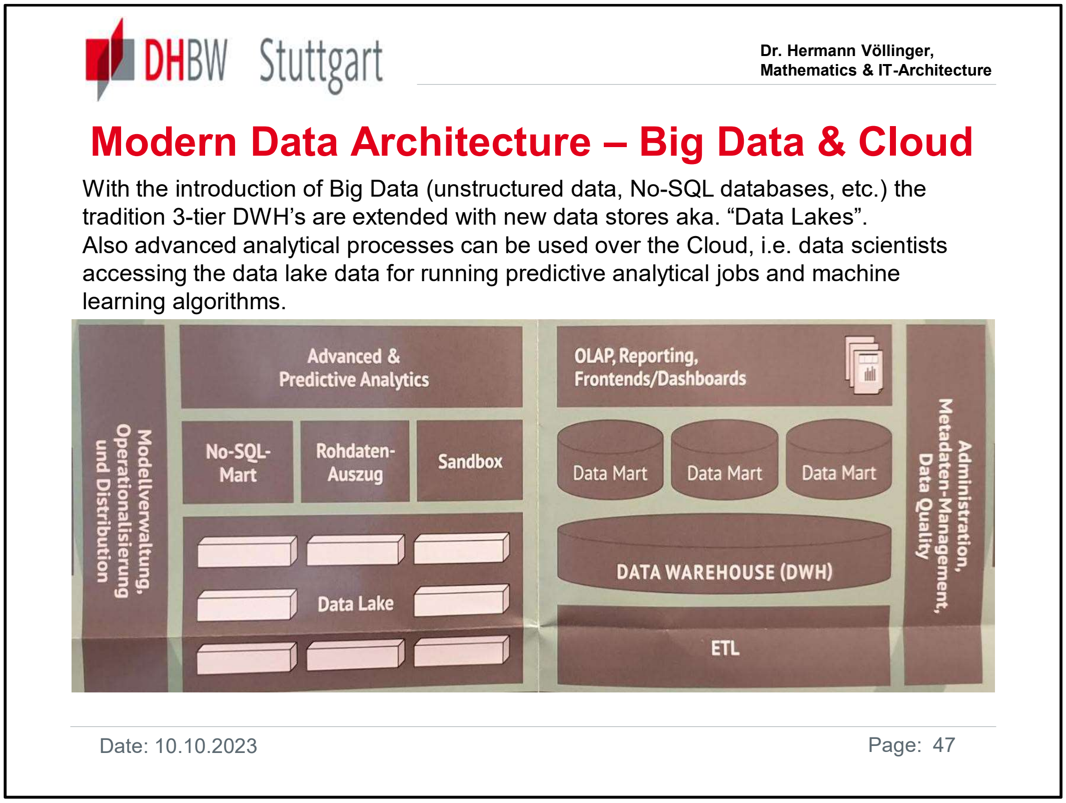
Date: 10.10.2023 Page: 46

The tasks of **the 5 Data Layers** and the properties of data in the layer are**:**

1. „**Staging Area (SA)**“ - collects the data from the source systems. The ETL process accomplishes some technical „Plausi“ rules during the load. Data are here put down as detail data in the database. As soon as the completeness of the data is guaranteed, the data are loaded into the next data layer. Data are stored for technical reasons some days, afterwards the data are deleted.
2. „**Core-/Central-DWH (CDW)**“ – the next layer ist the „Core- or Central- DWH“- Technical computations (transformations of data, i.e ETL) take place. The most important technical transformation are accomplished here. The data are also filtered during the loading. Only the fields defined in the data model (see also chapters for data modeling) are to be filled. The data in the CDW are historized and archived. The data in the CWH are not queried, i.e. evaluations „do not reach through“ to the CDW Layer.
3. „**Data Mart (DM)**“ – the data are in a multi-dimensional data model. Measures and dimensions are defined. This data model is optimized for the special type of analysis. Front-end tools access this data. For instance an OLAP cube (multi-dimensional data base) from these data can be easily built. The data structure is optimal for the demanded technical evaluation / analysis. The metadata layer contains the repository of metadata. This repository can be accessed and administrated for example with the IBM Information Governance Catalog (IGC). This tool processes technical and also business metadata. Which metadata should be exacted is defined in the metadata concept.
4. „**Analytics (AN)**“ – this layer contains all data for analytic evaluations. Exemplarily the OLAP can be cubes for a certain evaluation. The historization and archiving of these data are specified in the concept for the historization and archiving.
5. „**Access (AC)“** - this Layer contains reports in the form of Excel, HTML or pdf files. These files can be accessed by portal. The historization and archiving of these data are specified in the concept for the historization and archiving.

In addition to these pure data layers, we have also two more process-oriented layers:

- the **Metadata Layer** contains the repository of metadata. This repository can be accessed and administered for instance by IGC. This tool processes technical and business metadata. Which types of metadata are stored here, is defined Metadata Management concept.
- The next layer is the **Layer for Flow Control** (in German: **„Ablaufsteuerung“) -** over this layer all processes of the DWH are started, steered and supervised. For this the necessary metadata are also stored in the metadata layer.



Modern Data Architecture – Big Data & Cloud

With the introduction of Big Data (unstructured data, No-SQL databases, etc.) the tradition 3-tier DWH’s are extended with new data stores aka. “Data Lakes”.
Also advanced analytical processes can be used over the Cloud, i.e. data scientists accessing the data lake data for running predictive analytical jobs and machine learning algorithms.

Modellverwaltung, Operationalisierung und Distribution

Advanced & Predictive Analytics

No-SQL-Mart

Rohdaten-Auszug

Sandbox

Data Lake

OLAP, Reporting, Frontends/Dashboards

Data Mart

Data Mart

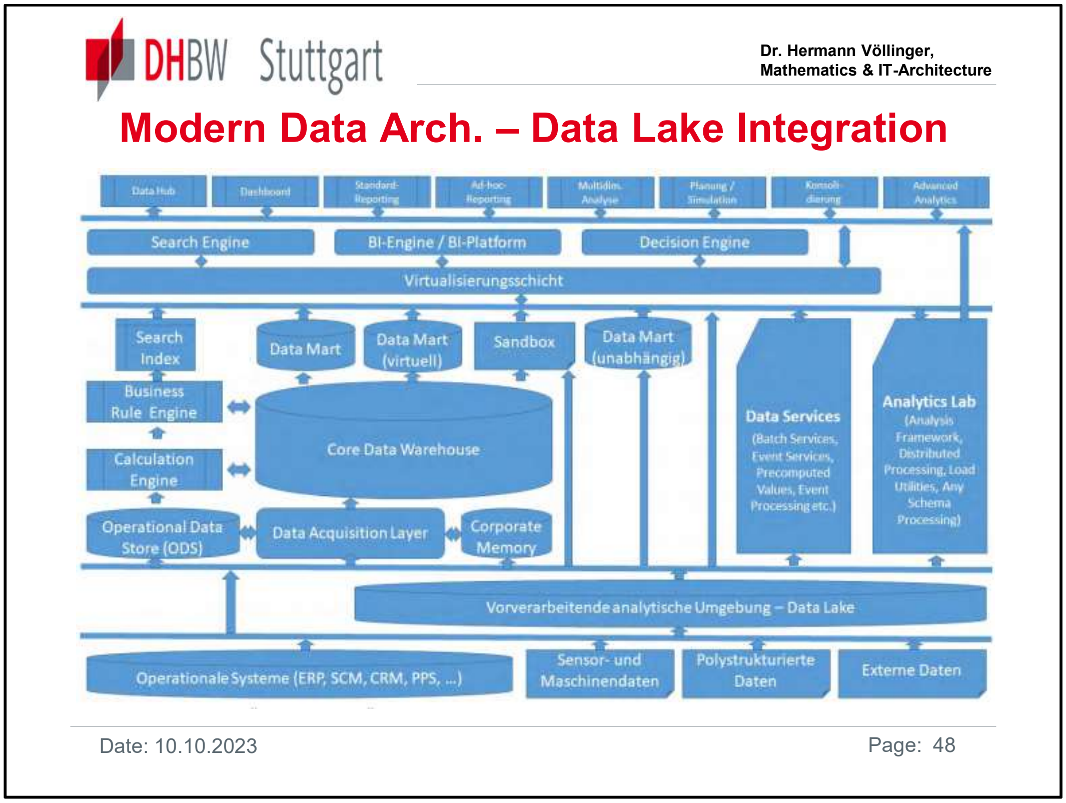
Data Mart

DATA WAREHOUSE (DWH)

ETL

Administration, Metadaten-Management, Data Quality

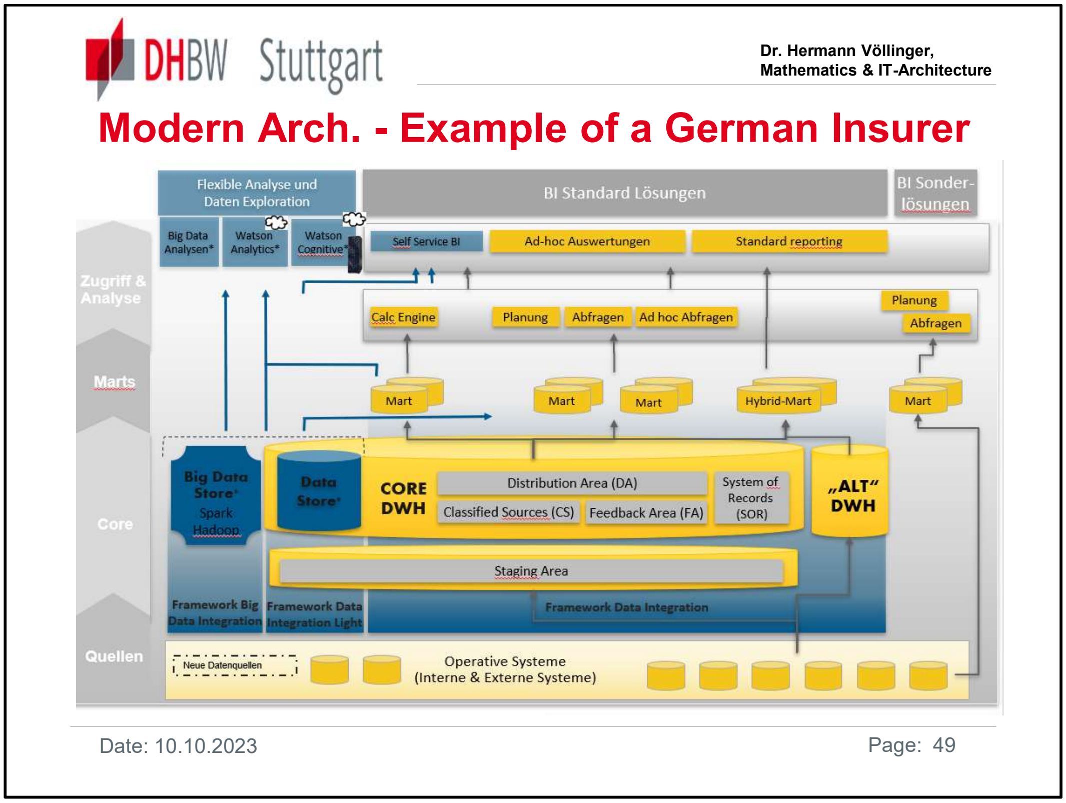
See also in [DHBW-Moodle] the document ”Trends_in_establishing_a_Data_Driven_Enterprise.pdf”



Modern Data Arch. – Data Lake Integration

Compare also the BI-Spektrum article about the co-existence of Data Lakes and DWHs in a modern BI Architecture:

https://www.sigs.de/artikel/datenseen-in-der-bi-landschaft



Modern Arch. - Example of a German Insurer

Flexible Analyse und Daten Exploration

BI Standard Lösungen

BI Sonder-lösungen

Zugriff & Analyse

Big Data Analysen*

Watson Analytics*

Watson Cognitive*

Self Service BI

Ad-hoc Auswertungen

Standard reporting

Calc Engine

Planung

Abfragen

Ad hoc Abfragen

Planung

Abfragen

Marts

Mart

Mart

Mart

Hybrid-Mart

Mart

Core

Big Data Store* Spark Hadoop

Data Store*

CORE DWH

Distribution Area (DA)

Classified Sources (CS)

Feedback Area (FA)

System of Records (SOR)

„ALT" DWH

Staging Area

Framework Big Data Integration

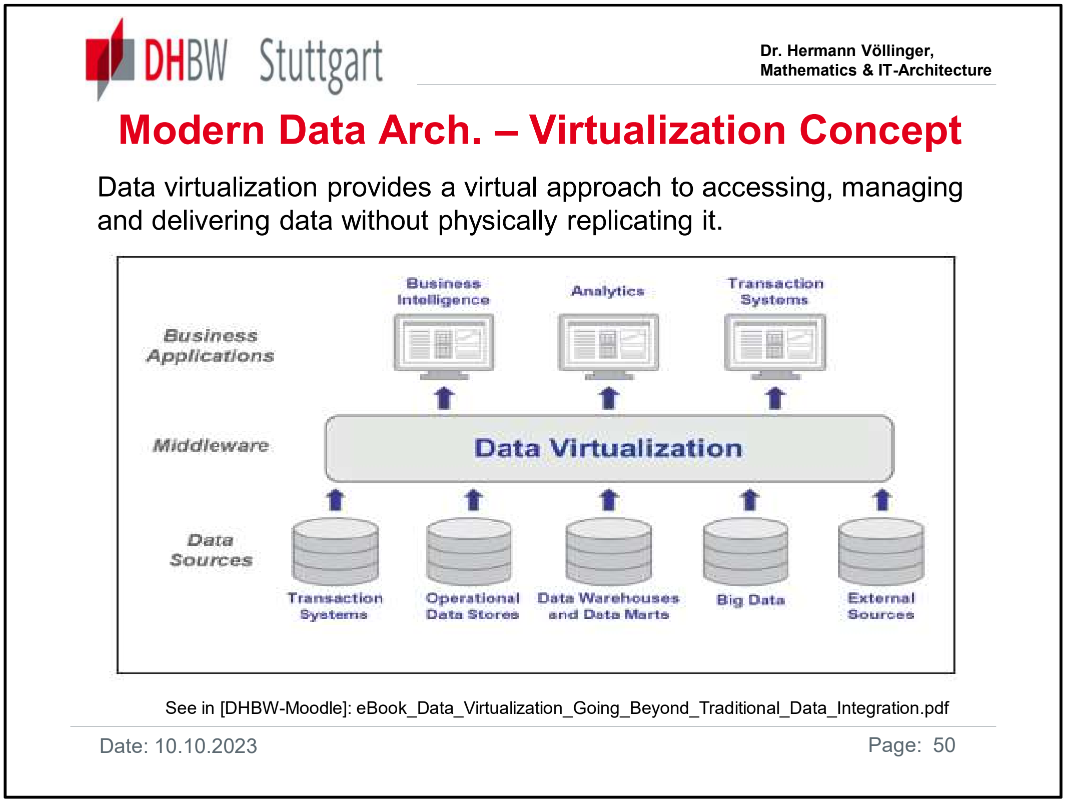
Framework Data Integration Light

Framework Data Integration

Quellen

Neue Datenquellen

Operative Systeme (Interne & Externe Systeme)



Modern Data Arch. – Virtualization Concept

Data virtualization provides a virtual approach to accessing, managing and delivering data without physically replicating it.

See in [DHBW-Moodle]: eBook_Data_Virtualization_Going_Beyond_Traditional_Data_Integration.pdf

Data virtualization is a logical data layer that integrates all enterprise data distributed across disparate systems, manages unified data for centralized security and governance, and makes it available to analysts in real time.

See also in [DHBW-Moodle]:

- “WhitePaper-Customer-Data-Platform_using_Data-Virtualisation.pdf”
- “YouTube-Pandemic_accelerated_data_virtualization.pdf”

Definition: Data virtualization is a technology that allows data from various sources, which may be distributed across different systems and locations, to be accessed and queried in real-time as if it were a single, unified data source. It creates a logical abstraction layer over disparate data sources.

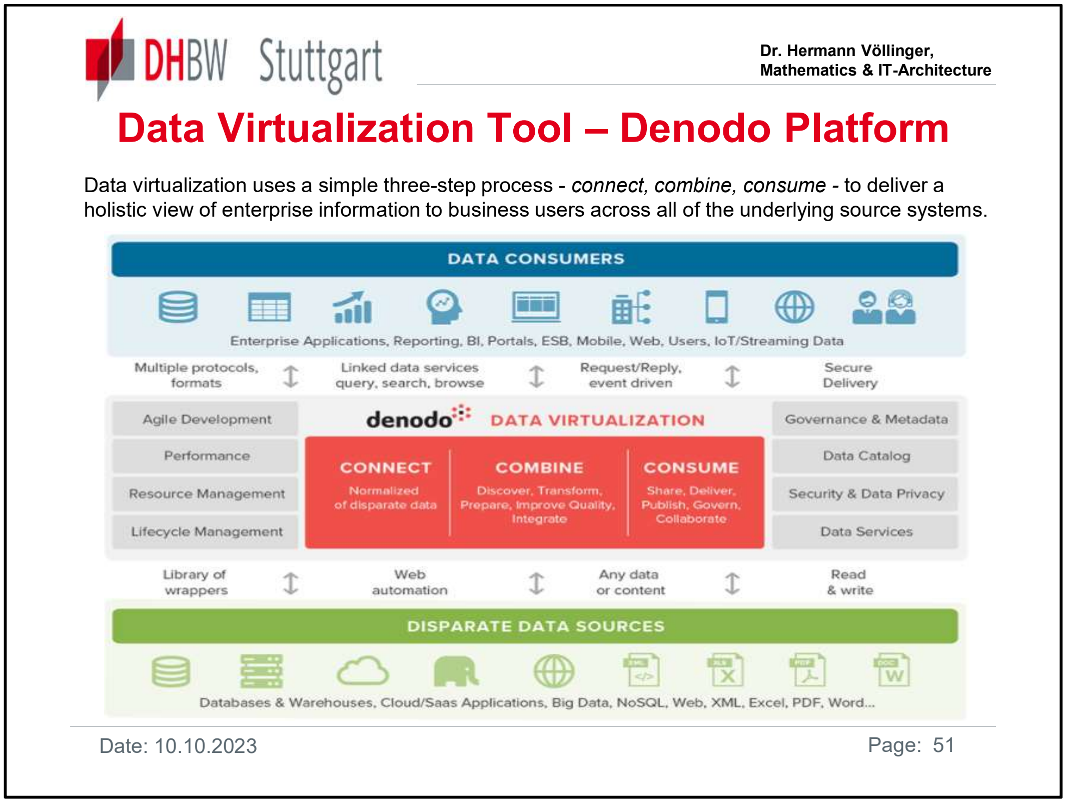
Key Features:

Data Integration: Data virtualization integrates data from multiple sources, including databases, cloud services, and APIs, without physically moving or copying the data.

Real-Time Access: Users can access and query the data in real-time, without the need for ETL (Extract, Transform, Load) processes.

Data Federation: It federates data from different sources and presents it in a unified manner, making it easier for analysts and applications to work with.

Data Security: Data virtualization often includes security features to control access to the underlying data sources.



Data Virtualization Tool – Denodo Platform

Data virtualization uses a simple three-step process - *connect, combine, consume* - to deliver a holistic view of enterprise information to business users across all of the underlying source systems.

DATA CONSUMERS

Enterprise Applications, Reporting, BI, Portals, ESB, Mobile, Web, Users, IoT/Streaming Data

Multiple protocols, formats ↕ Linked data services query, search, browse ↕ Request/Reply, event driven ↕ Secure Delivery

Agile Development | Performance | Resource Management | Lifecycle Management

denodo DATA VIRTUALIZATION

CONNECT: Normalized of disparate data

COMBINE: Discover, Transform, Prepare, Improve Quality, Integrate

CONSUME: Share, Deliver, Publish, Govern, Collaborate

Governance & Metadata | Data Catalog | Security & Data Privacy | Data Services

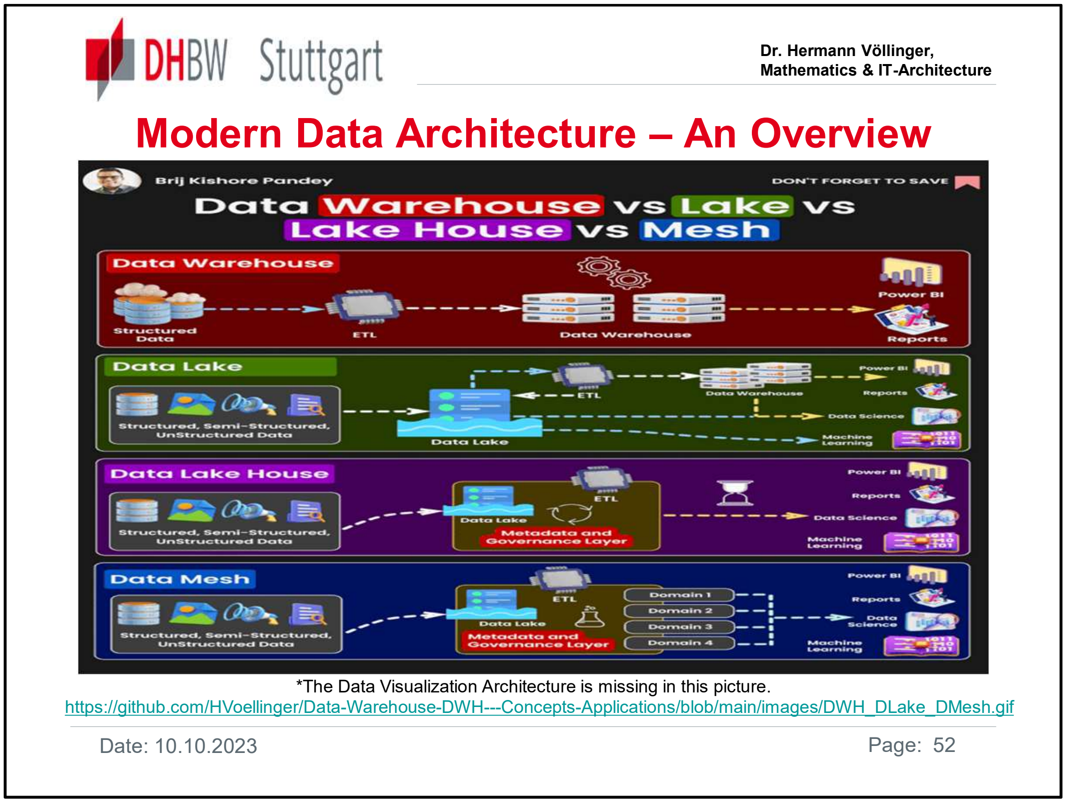
Library of wrappers ↕ Web automation ↕ Any data or content ↕ Read & write

DISPARATE DATA SOURCES

Databases & Warehouses, Cloud/Saas Applications, Big Data, NoSQL, Web, XML, Excel, PDF, Word...

Denodo is a leading provider in the field of data virtualization and offers companies agile and high performance data integration, data abstraction and data services in real-time from a variety of different sources such as enterprise data, cloud data or Big Data and unstructured data.

Denodo's customers from all major industries have benefited from faster and easier access to access to unified business information for their agile BI, Big Data Analytics, web and cloud integration, single-view applications, and enterprise data services. Significantly increase their flexibility and significantly increase profitability. Denodo is well-funded, profitable and privately held. For more information visit [Datenvirtualisierung | Denodo]



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Modern Data Architecture – An Overview

*The Data Visualization Architecture is missing in this picture.
https://github.com/HVoellinger/Data-Warehouse-DWH---Concepts-Applications/blob/main/images/DWH_DLake_DMesh.gif

Date: 10.10.2023 Page: 52

See for Data Mesh (German: “Daten-Gitter“, “Daten-Geflecht“): https://www.datamesh-architecture.com/

Data Mesh-Definition: Data mesh is a relatively new approach to managing and scaling data in large organizations. It emphasizes the decentralization of data ownership and processing, treating data as a product, and using domain-oriented cross-functional teams to manage data at scale.

Key Principles:

Domain-Oriented Ownership: Data is owned and managed by the domain or business unit that generates it, rather than by a central data team.

Data Products: Data is treated as a product, with clear documentation, versioning, and accessibility.

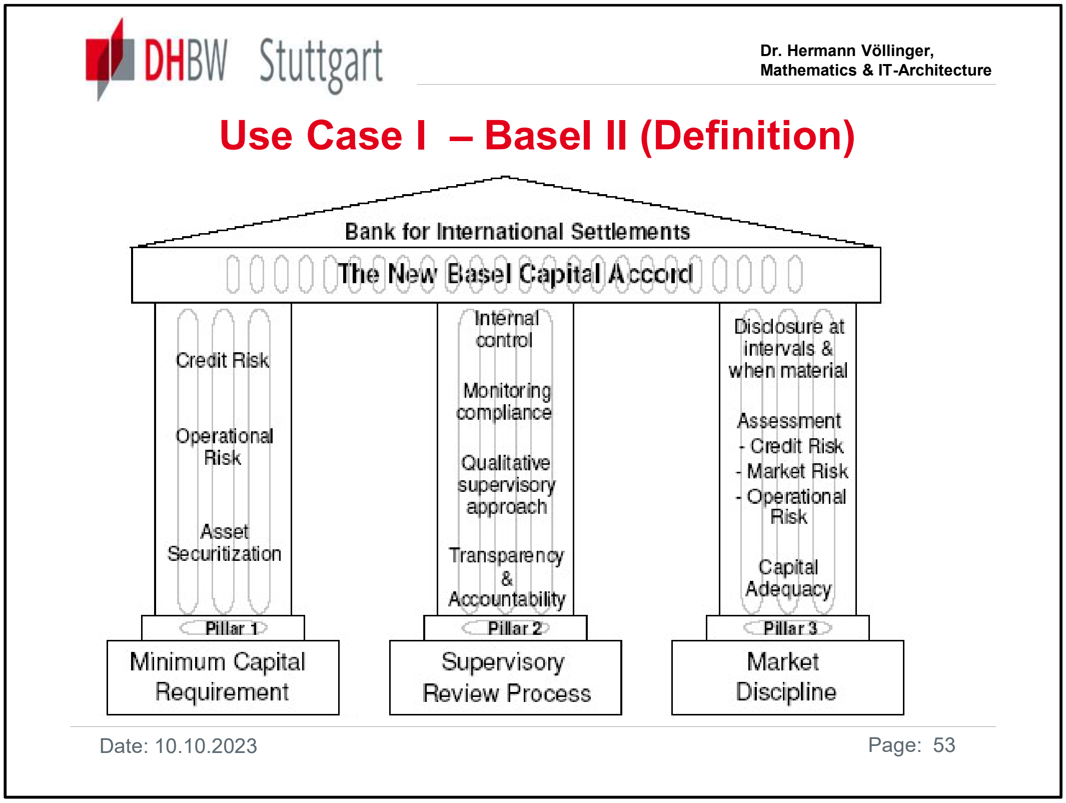
Self-serve Data Infrastructure: Data infrastructure is designed to be self-serve, enabling domain teams to manage their own data pipelines and analytics.

Federated Query: Data can be queried across the organization through a federated approach, allowing teams to access data they need without needing to move or centralize it.

Remark: Data mesh and data virtualization are not the same. They are two distinct concepts in the field of data management and analytics, each with its own purpose and approach.

In summary, data mesh is more of a conceptual and organizational framework for managing data at scale, emphasizing decentralized ownership and treating data as a product. Data virtualization, on the other hand, is a technology-driven approach for integrating and accessing data from various sources in real-time without the need for physical data movement. While they can complement each other in some contexts, they serve different purposes and focus on different aspects of data management and analytics.

See Project example: “Data Mesh Success for Fifth Third Bank”: Watch this webinar



Use Case I – Basel II (Definition)

Bank for International Settlements

The New Basel Capital Accord

Credit Risk

Operational Risk

Asset Securitization

Pillar 1

Minimum Capital Requirement

Internal control

Monitoring compliance

Qualitative supervisory approach

Transparency & Accountability

Pillar 2

Supervisory Review Process

Disclosure at intervals & when material

Assessment
- Credit Risk
- Market Risk
- Operational Risk

Capital Adequacy

Pillar 3

Market Discipline

Context

Introduces key features of Basel II, the Pillar Concept and proposed timeline.

Key Points

-Basel II covers credit, market and operational risk

-There are 3 pillars – P1, Minimum capital charge, P2, Supervisory review and P3, Market Discipline

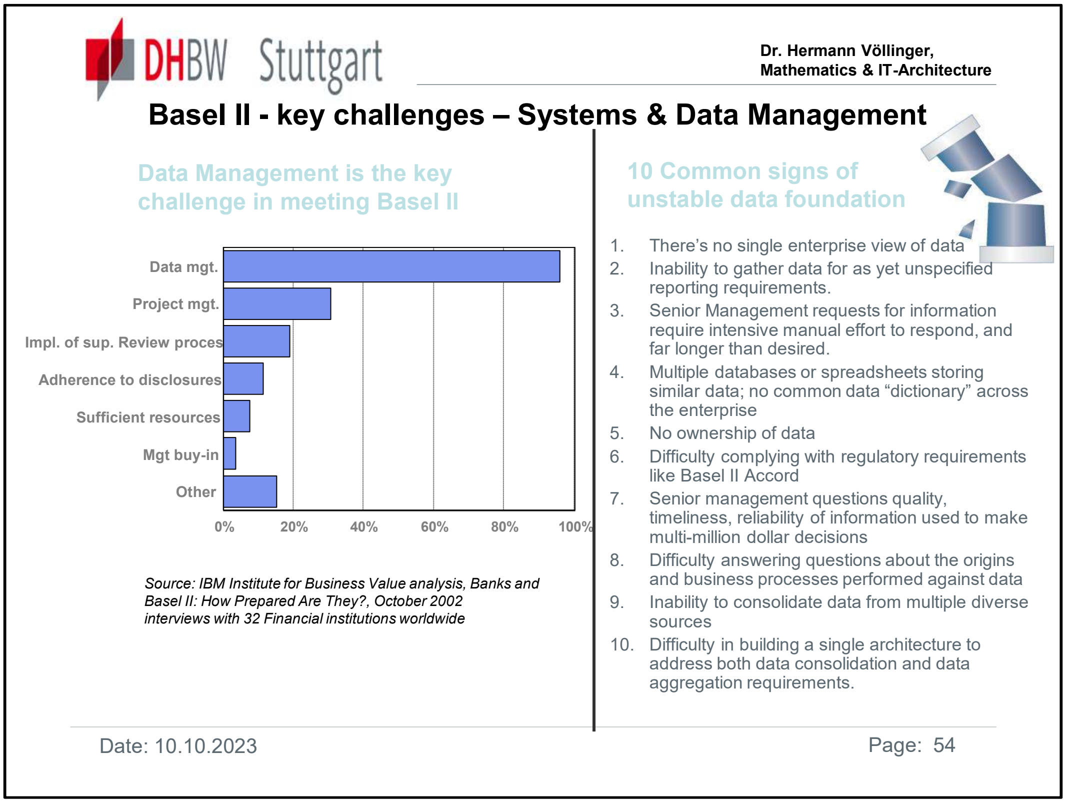
-IBM focus is credit and operational risk as majority of banks will have some solution for their market risk issues (this area is also well covered by risk systems vendors)

-Live by beginning of 2007 but there is a significant amount of preparatory work to be completed well before this date (important to act)

-Option to choose strategy – from basic to advanced

-Basic less advantageous as it requires significant levels of regulator input

-Advanced can deliver significantly higher benefits


DHBW Stuttgart
Data mgt.
Project mgt.
Impl. of sup. Review proces
Adherence to disclosures
Sufficient resources
Mgt buy-in
Other
0%
20%
40%
60%
80%
100%

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Basel II - key challenges – Systems & Data Management

Data Management is the key challenge in meeting Basel II

Source: IBM Institute for Business Value analysis, Banks and Basel II: How Prepared Are They?, October 2002 interviews with 32 Financial institutions worldwide

10 Common signs of unstable data foundation

1. There's no single enterprise view of data
2. Inability to gather data for as yet unspecified reporting requirements.
3. Senior Management requests for information require intensive manual effort to respond, and far longer than desired.
4. Multiple databases or spreadsheets storing similar data; no common data "dictionary" across the enterprise
5. No ownership of data
6. Difficulty complying with regulatory requirements like Basel II Accord
7. Senior management questions quality, timeliness, reliability of information used to make multi-million dollar decisions
8. Difficulty answering questions about the origins and business processes performed against data
9. Inability to consolidate data from multiple diverse sources
10. Difficulty in building a single architecture to address both data consolidation and data aggregation requirements.

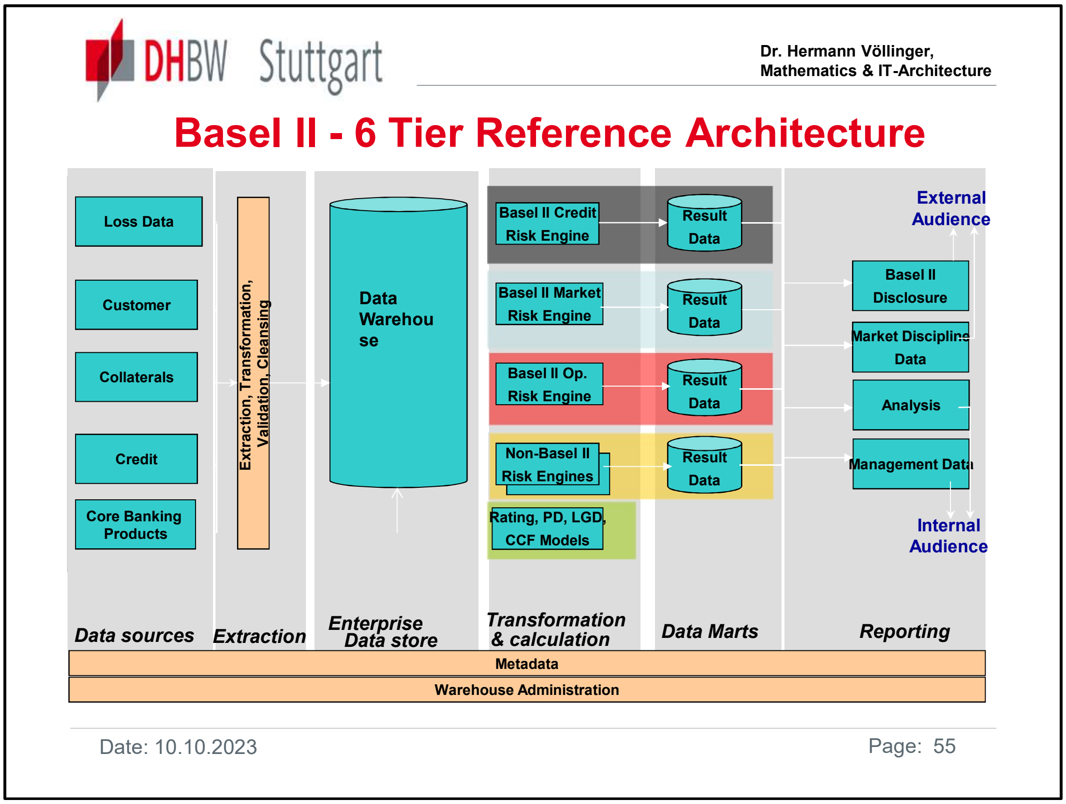
Date: 10.10.2023 Page: 54

Data Management is the key challenge - despite these data problems we see today the need for Basel II to fulfil many **IT Requirements.** These are essential for the success of Basel II:

1. **Standardized data practices,** across all divisions and sources
2. **Extensive time series** data for credit risk drivers
3. **Capacity for massive data volumes**
4. **Sophisticated methodologies** to model the influence of credit risk drivers.
5. **New data modelling processes** (credit rating, LGD, PD, EAD)
6. **Open, flexible architecture** to integrate all kinds of banking models
7. **Internal** and **external reporting**
8. **Certain analytics,** e.g. calculation of migration matrixes
9. **Integration** with calculation engines

What are Banks doing about it (status 2003):

- Most Banks have decided on their strategies during 2003
- Banks are now doing data gap analysis based on CP3 to know what data to collect and ensure it is available from source systems
- RFIs and RFPs are being issued for Basel II data storage solutions (Credit Risk Database or Data Warehouse).
- During 2004 most Banks will implementing their Basel II data collection strategies and will start collecting data for Basel 2 compliance. **3 years data needed by 2007.**
- 2004 was the Data Management Infrastructure Year



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Basel II - 6 Tier Reference Architecture

Loss Data
Customer
Collaterals
Credit
Core Banking Products

Extraction, Transformation, Validation, Cleansing

Data Warehouse

Basel II Credit Risk Engine
Basel II Market Risk Engine
Basel II Op. Risk Engine
Non-Basel II Risk Engines
Rating, PD, LGD, CCF Models

Result Data
Result Data
Result Data
Result Data

External Audience
Basel II Disclosure
Market Discipline Data
Analysis
Management Data
Internal Audience

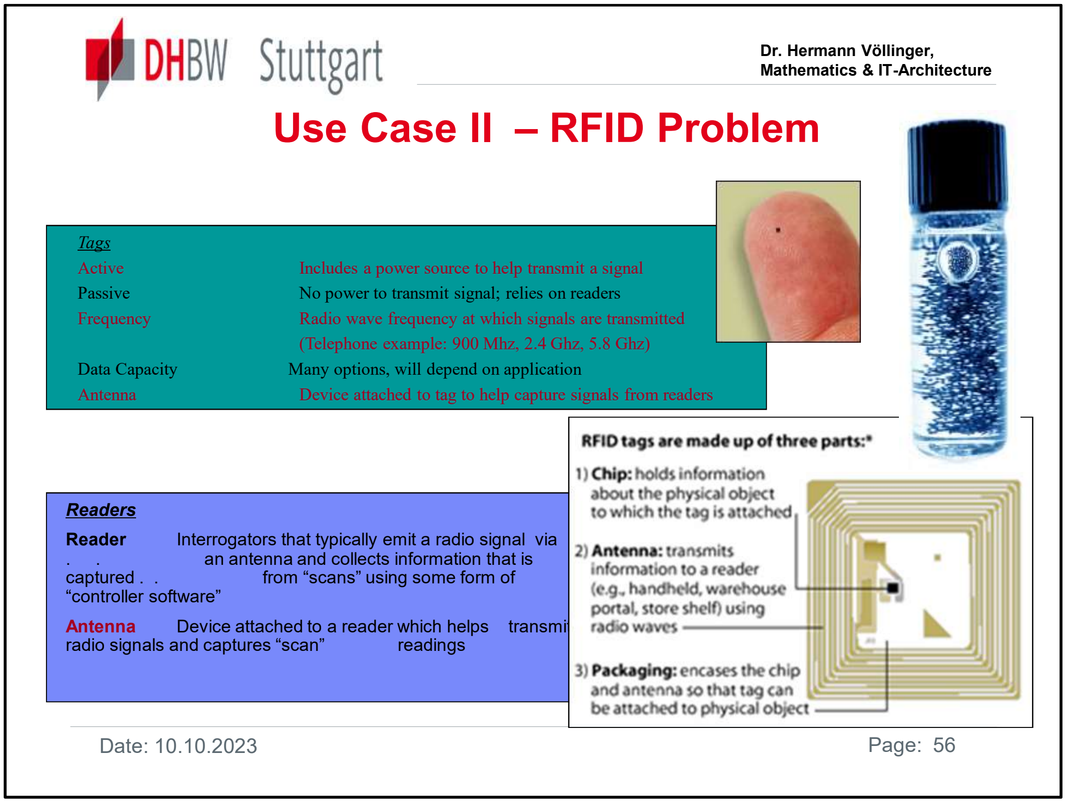
Data sources *Extraction* *Enterprise Data store* *Transformation & calculation* *Data Marts* *Reporting*

Metadata

Warehouse Administration

Date: 10.10.2023 Page: 55

The picture shows the IBM proposal for IT Infrastructure.
It shows a 6-Tier model.
We have a Central DWH. We have Marts as a second, specialised and enriched data store. The reporting layers gives output to internal and also external audience.
Metadata & Administration layers should work over all tiers.
Build up of CDW should start with 2004, since you need three years of data history. The calculation and the reporting facilities should work by end of 2006, beginning with 2007.



Use Case II – RFID Problem

Tags

Active	Includes a power source to help transmit a signal
Passive	No power to transmit signal; relies on readers
Frequency	Radio wave frequency at which signals are transmitted (Telephone example: 900 Mhz, 2.4 Ghz, 5.8 Ghz)
Data Capacity	Many options, will depend on application
Antenna	Device attached to tag to help capture signals from readers

Readers

Reader Interrogators that typically emit a radio signal via . . an antenna and collects information that is captured . . from “scans” using some form of “controller software”

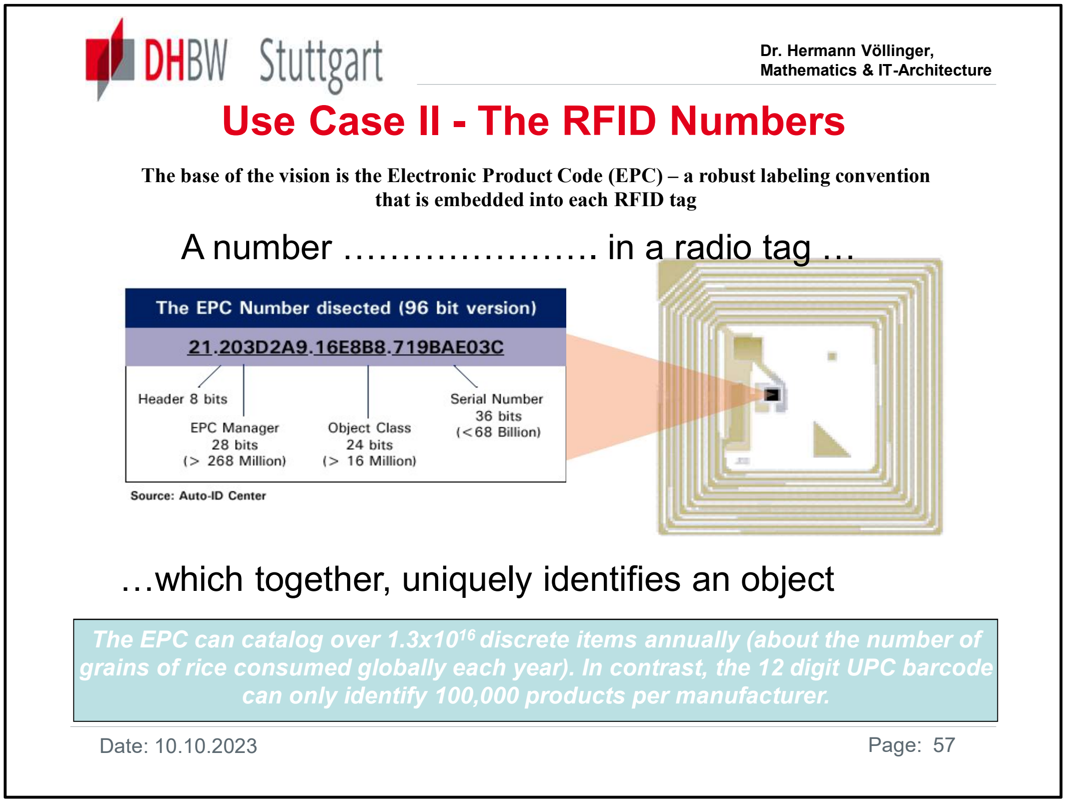
Antenna Device attached to a reader which helps transmi radio signals and captures “scan” readings

RFID tags are made up of three parts:*

1) **Chip:** holds information about the physical object to which the tag is attached

2) **Antenna:** transmits information to a reader (e.g., handheld, warehouse portal, store shelf) using radio waves

3) **Packaging:** encases the chip and antenna so that tag can be attached to physical object



Use Case II - The RFID Numbers

The base of the vision is the Electronic Product Code (EPC) – a robust labeling convention that is embedded into each RFID tag

A number …………………. in a radio tag …

The EPC Number disected (96 bit version)

21.203D2A9.16E8B8.719BAE03C

Header 8 bits

EPC Manager
28 bits
(> 268 Million)

Object Class
24 bits
(> 16 Million)

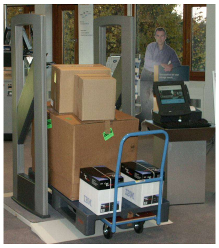
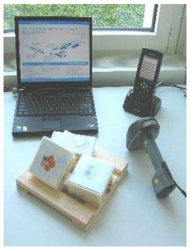
Serial Number
36 bits
(<68 Billion)

Source: Auto-ID Center

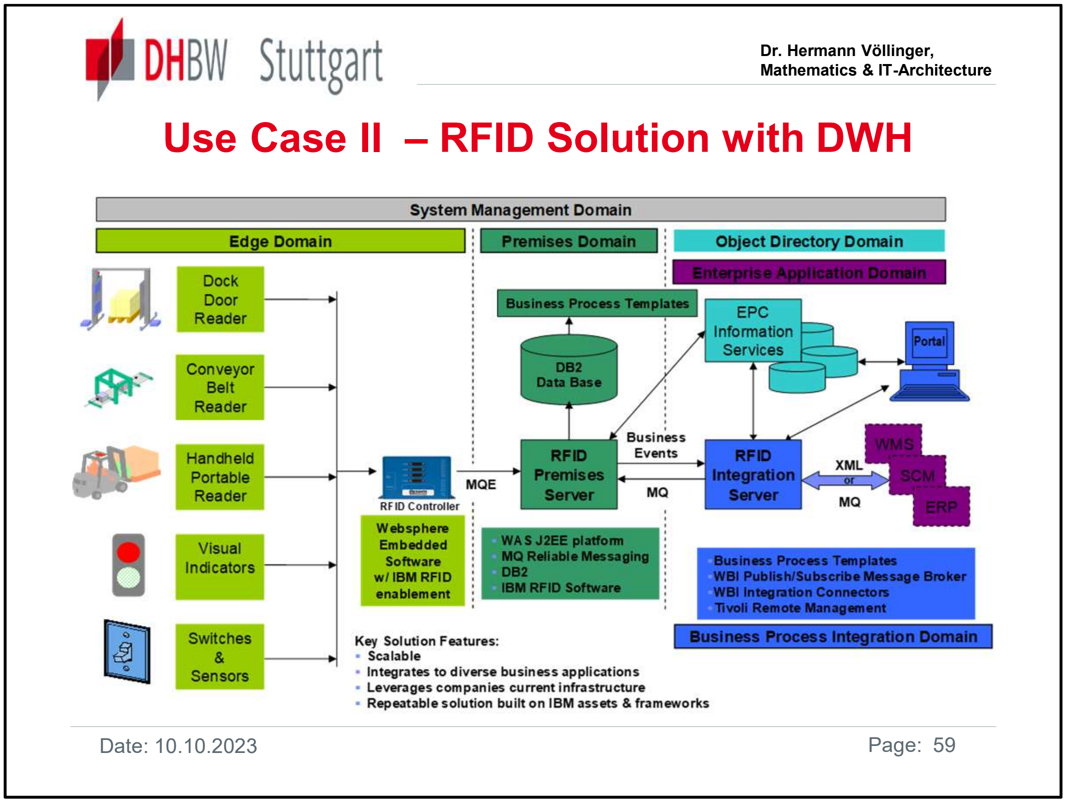
…which together, uniquely identifies an object

The EPC can catalog over 1.3x10^{16} discrete items annually (about the number of grains of rice consumed globally each year). In contrast, the 12 digit UPC barcode can only identify 100,000 products per manufacturer.

Use Case II – The RFID Infrastructure

			
RFID Self-Checkout	**Distribution Center Palette Control (DC Exit)**	**RFID/AutoID Warehouse**	**EPC RFID Demo**
A supermarket scenario similar to the IBM Commercial „Supermarket“	Verify palette packaging before leaving the distribution center	An order pickup scenario	Represent 3 different points in the supply chain via portals (retail store, retail DC, supplier)

Use Case II – RFID Solution with DWH



System Management Domain

Edge Domain

Premises Domain

Object Directory Domain

Enterprise Application Domain

Dock Door Reader

Conveyor Belt Reader

Handheld Portable Reader

Visual Indicators

Switches & Sensors

RFID Controller

Websphere Embedded Software w/ IBM RFID enablement

MQE

Business Process Templates

DB2 Data Base

RFID Premises Server

- WAS J2EE platform
- MQ Reliable Messaging
- DB2
- IBM RFID Software

Business Events

MQ

EPC Information Services

Portal

RFID Integration Server

XML or MQ

WMS

SCM

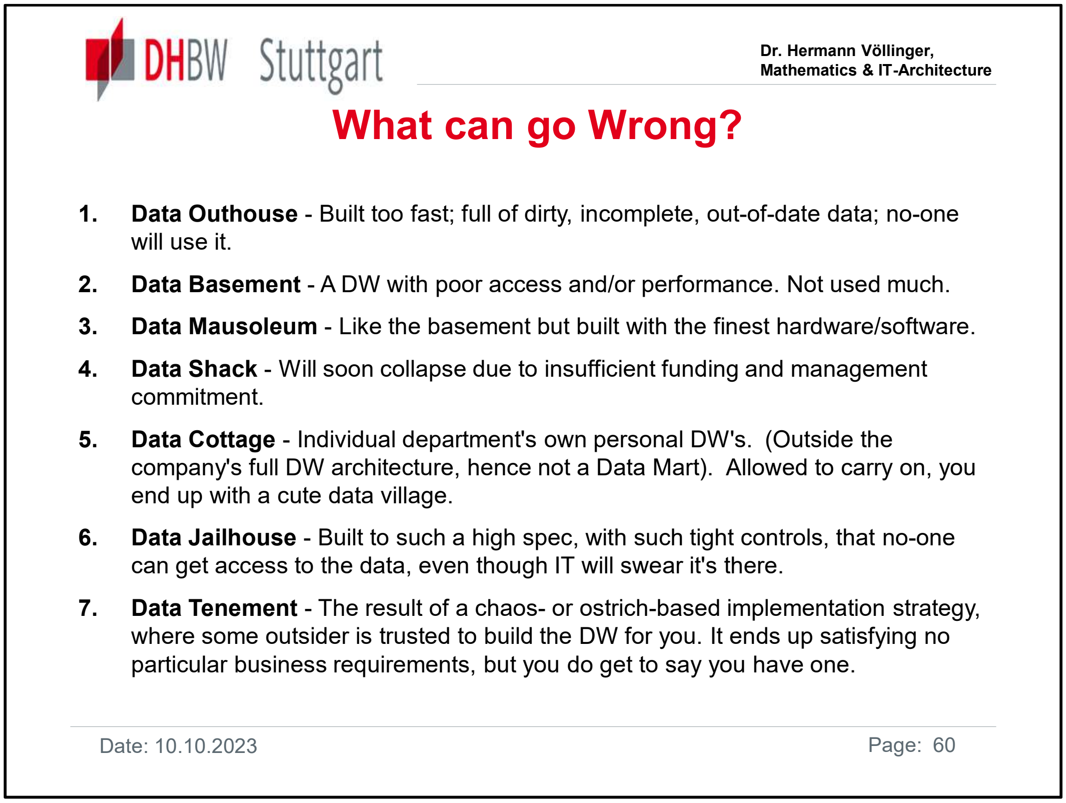
ERP

- Business Process Templates
- WBI Publish/Subscribe Message Broker
- WBI Integration Connectors
- Tivoli Remote Management

Business Process Integration Domain

Key Solution Features:

- Scalable
- Integrates to diverse business applications
- Leverages companies current infrastructure
- Repeatable solution built on IBM assets & frameworks



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

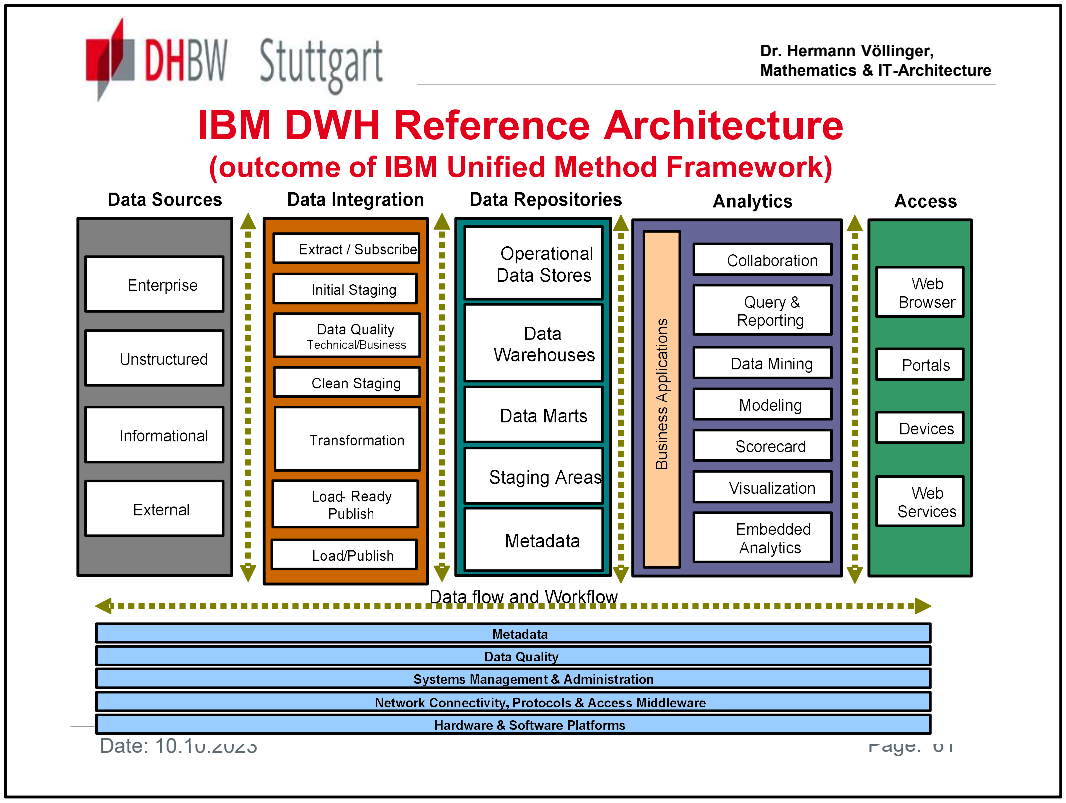
What can go Wrong?

1. **Data Outhouse** - Built too fast; full of dirty, incomplete, out-of-date data; no-one will use it.
2. **Data Basement** - A DW with poor access and/or performance. Not used much.
3. **Data Mausoleum** - Like the basement but built with the finest hardware/software.
4. **Data Shack** - Will soon collapse due to insufficient funding and management commitment.
5. **Data Cottage** - Individual department's own personal DW's. (Outside the company's full DW architecture, hence not a Data Mart). Allowed to carry on, you end up with a cute data village.
6. **Data Jailhouse** - Built to such a high spec, with such tight controls, that no-one can get access to the data, even though IT will swear it's there.
7. **Data Tenement** - The result of a chaos- or ostrich-based implementation strategy, where some outsider is trusted to build the DW for you. It ends up satisfying no particular business requirements, but you do get to say you have one.

Date: 10.10.2023 Page: 60

See slide itself for a description:
The German eequivalence words are:

- Data Outhouse – ‘Schellschuss‘
- Data Basement - ‘Grab‘
- Data Mausoleum - ‘Mausoleum‘
- Data Shack - ‘Schatten‘
- Data Cottage - ‘Hütte‘
- Data Jailhouse - ‘Gefängnis‘
- Data Tenement - ‘Muster-/Mietwohnung‘



IBM DWH Reference Architecture

(outcome of IBM Unified Method Framework)

Data Sources
- Enterprise
- Unstructured
- Informational
- External

Data Integration
- Extract / Subscribe
- Initial Staging
- Data Quality Technical/Business
- Clean Staging
- Transformation
- Load- Ready Publish
- Load/Publish

Data Repositories
- Operational Data Stores
- Data Warehouses
- Data Marts
- Staging Areas
- Metadata

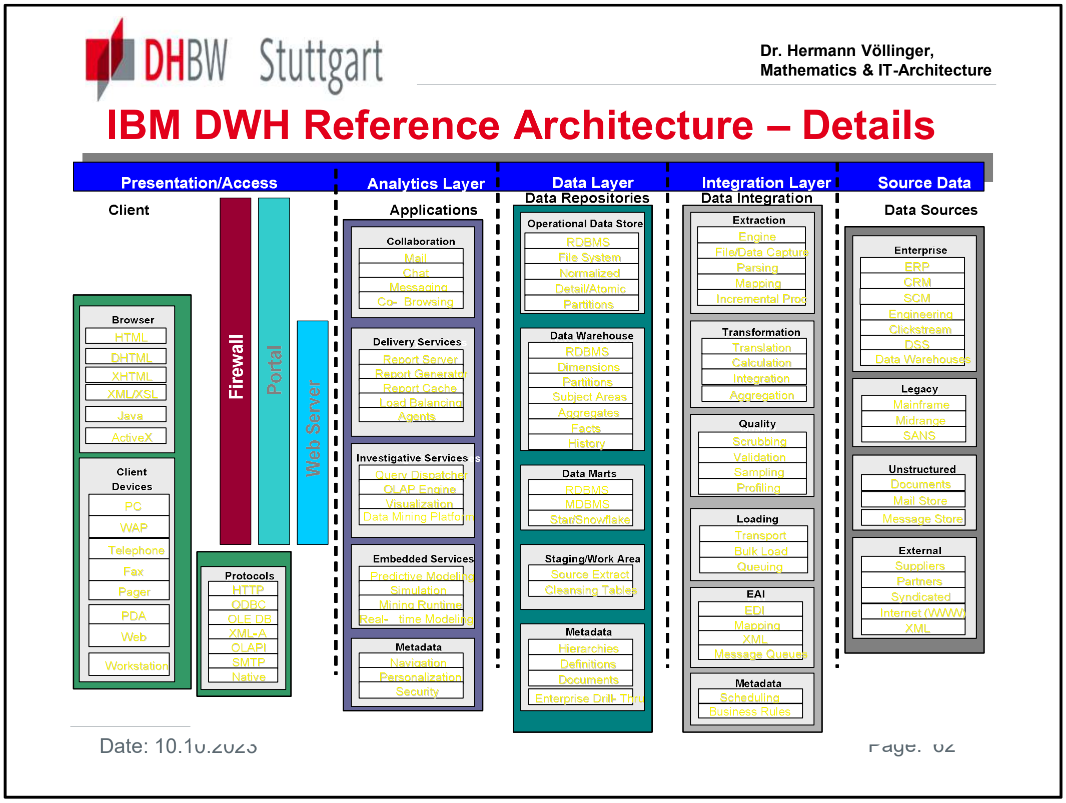
Analytics
- Business Applications
- Collaboration
- Query & Reporting
- Data Mining
- Modeling
- Scorecard
- Visualization
- Embedded Analytics

Access
- Web Browser
- Portals
- Devices
- Web Services

Data flow and Workflow

- Metadata
- Data Quality
- Systems Management & Administration
- Network Connectivity, Protocols & Access Middleware
- Hardware & Software Platforms

Date: 10.10.2023 Page: 61



IBM DWH Reference Architecture – Details

Presentation/Access

Client

Browser
- HTML
- DHTML
- XHTML
- XML/XSL
- Java
- ActiveX

Client Devices
- PC
- WAP
- Telephone
- Fax
- Pager
- PDA
- Web
- Workstation

Firewall

Portal

Web Server

Protocols
- HTTP
- ODBC
- OLE DB
- XML-A
- OLAPI
- SMTP
- Native

Analytics Layer

Applications

Collaboration
- Mail
- Chat
- Messaging
- Co- Browsing

Delivery Services
- Report Server
- Report Generator
- Report Cache
- Load Balancing
- Agents

Investigative Services
- Query Dispatcher
- OLAP Engine
- Visualization
- Data Mining Platform

Embedded Services
- Predictive Modeling
- Simulation
- Mining Runtime
- Real- time Modeling

Metadata
- Navigation
- Personalization
- Security

Data Layer

Data Repositories

Operational Data Store
- RDBMS
- File System
- Normalized
- Detail/Atomic
- Partitions

Data Warehouse
- RDBMS
- Dimensions
- Partitions
- Subject Areas
- Aggregates
- Facts
- History

Data Marts
- RDBMS
- MDBMS
- Star/Snowflake

Staging/Work Area
- Source Extract
- Cleansing Tables

Metadata
- Hierarchies
- Definitions
- Documents
- Enterprise Drill- Thru

Integration Layer

Data Integration

Extraction
- Engine
- File/Data Capture
- Parsing
- Mapping
- Incremental Proc

Transformation
- Translation
- Calculation
- Integration
- Aggregation

Quality
- Scrubbing
- Validation
- Sampling
- Profiling

Loading
- Transport
- Bulk Load
- Queuing

EAI
- EDI
- Mapping
- XML
- Message Queues

Metadata
- Scheduling
- Business Rules

Source Data

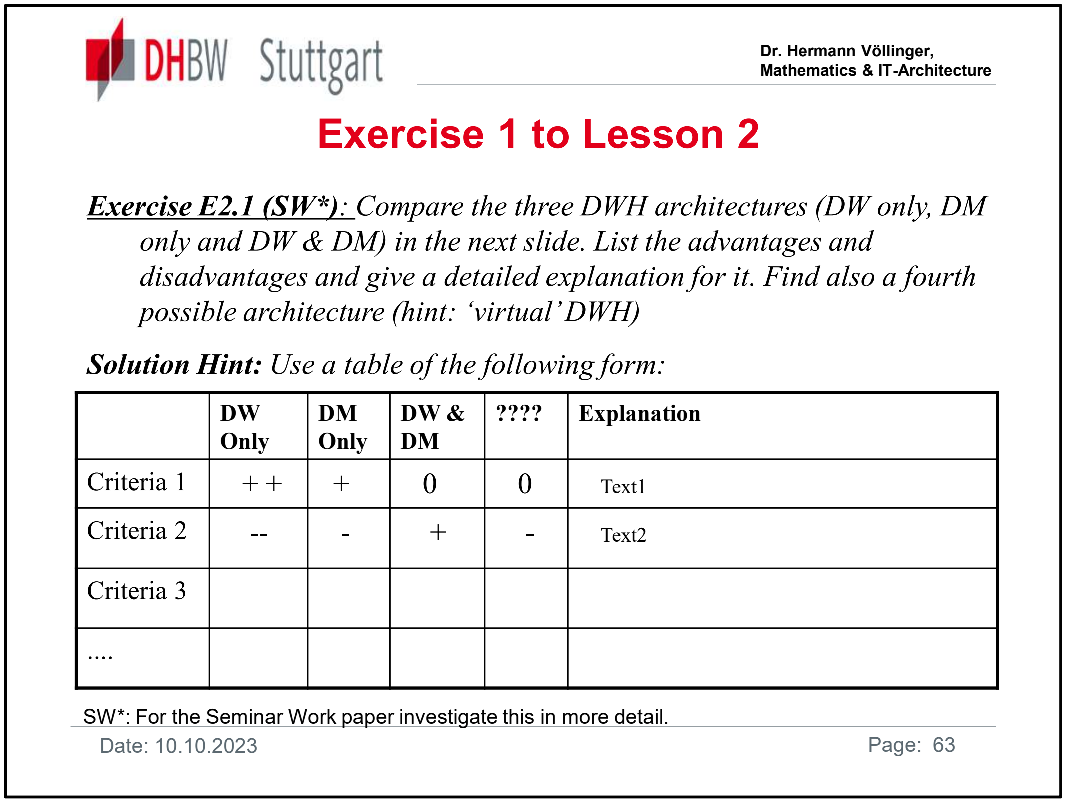
Data Sources

Enterprise
- ERP
- CRM
- SCM
- Engineering
- Clickstream
- DSS
- Data Warehouses

Legacy
- Mainframe
- Midrange
- SANS

Unstructured
- Documents
- Mail Store
- Message Store

External
- Suppliers
- Partners
- Syndicated
- Internet (WWW)
- XML



Exercise 1 to Lesson 2

Exercise E2.1 (SW*): *Compare the three DWH architectures (DW only, DM only and DW & DM) in the next slide. List the advantages and disadvantages and give a detailed explanation for it. Find also a fourth possible architecture (hint: 'virtual' DWH)*

Solution Hint: *Use a table of the following form:*

	DW Only	DM Only	DW & DM	????	Explanation
Criteria 1	+ +	+	0	0	Text1
Criteria 2	--	-	+	-	Text2
Criteria 3					
….					

SW*: For the Seminar Work paper investigate this in more detail.

Further Information:

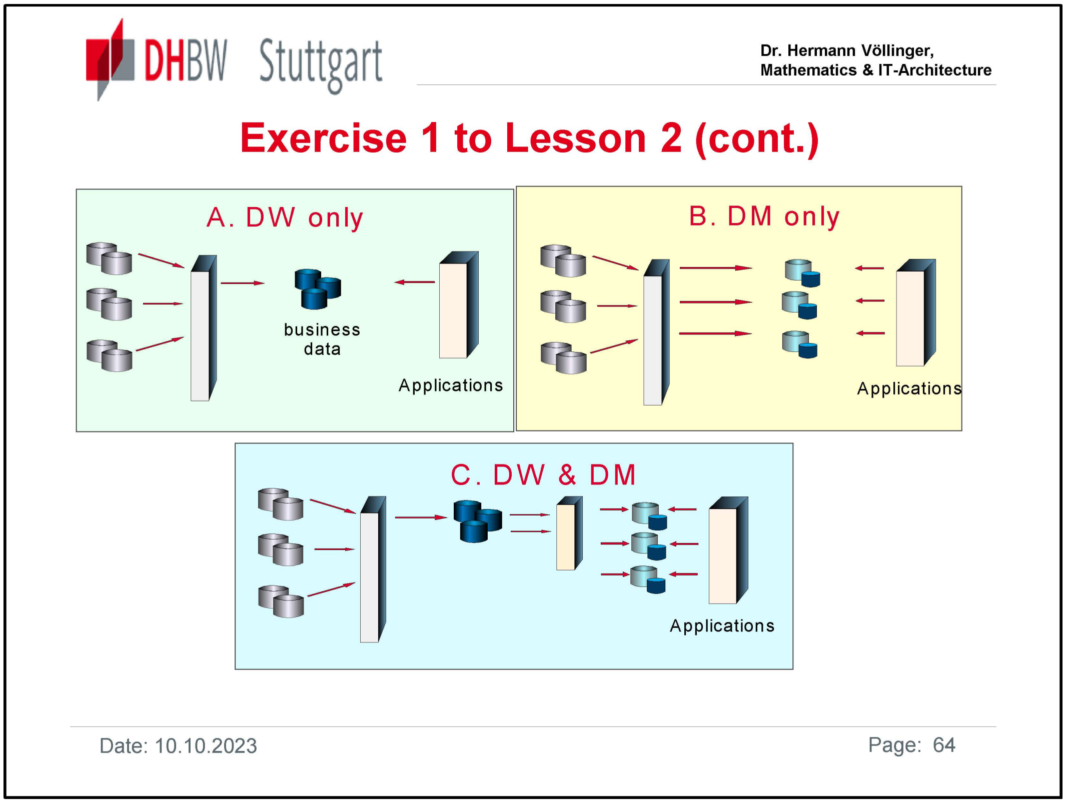
1. Datenbanken Verstehen, *Data Warehouse mit Data Marts.* Available: https://www.datenbanken-verstehen.de/data-warehouse/data-warehouse-grundlagen/data-warehouse-komponenten/data-warehouse-architektur/

2. Datenbanken Verstehen, *Zentrales Data Warehouse.* Available: https://www.datenbanken-verstehen.de/data-warehouse/data-warehouse-grundlagen/data-warehouse-komponenten/data-warehouse-architektur/

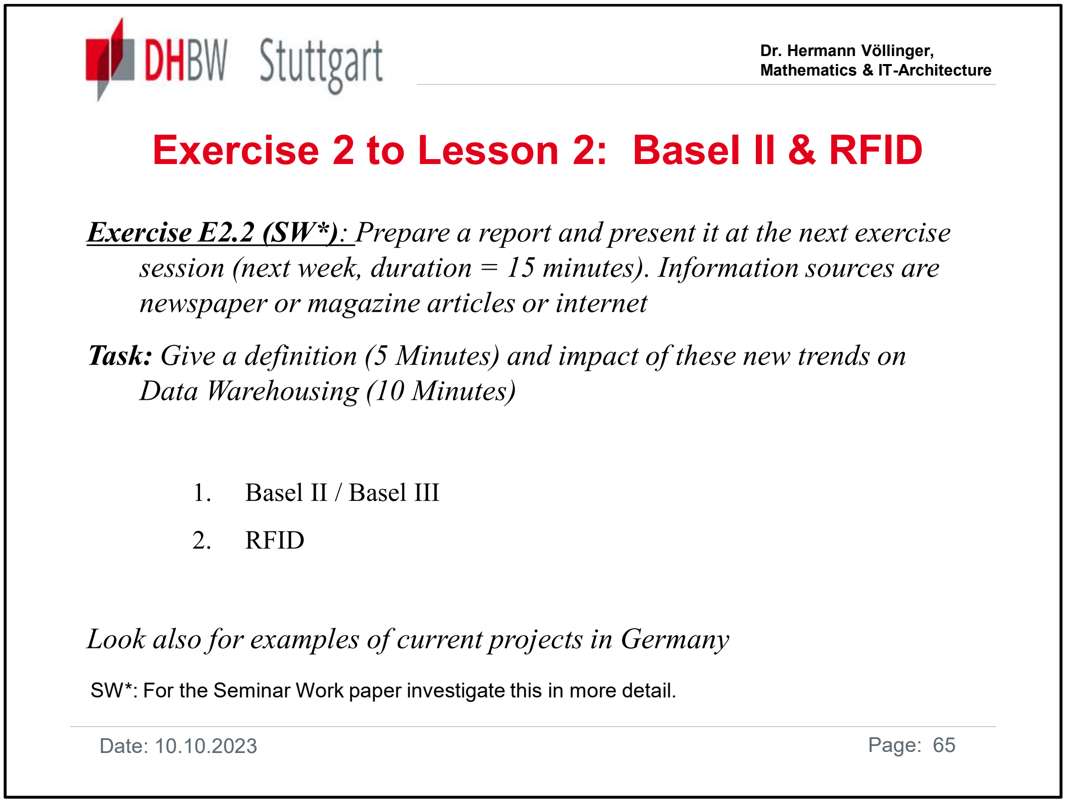
3. Shana Pearlman, *Was ist ein Data Mart?* Available: https://de.talend.com/resources/what-is-data-mart

4. N. L. Michael Matzer, *Das Virtual Data Warehouse verhilft zur schnelleren digitalen Transformation.* Available: https://www.bigdata-insider.de/das-virtual-data-warehouse-verhilft-zur-schnelleren-digitalen-transformation-a-617756

5. Klaus Manhart. *BI-Datenmanagement (Teil 2): Das Data Warehouse: Das Core Data Warehouse*. 2008. Available: Grundlagenserie Business Intelligence: BI-Datenmanagement (Teil 2): Das Data Warehouse - TecChannel Workshop



Exercise 1 to Lesson 2 (cont.)



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise 2 to Lesson 2: Basel II & RFID

Exercise E2.2 (SW*): *Prepare a report and present it at the next exercise session (next week, duration = 15 minutes). Information sources are newspaper or magazine articles or internet*

Task: *Give a definition (5 Minutes) and impact of these new trends on Data Warehousing (10 Minutes)*

1. Basel II / Basel III
2. RFID

Look also for examples of current projects in Germany

SW*: For the Seminar Work paper investigate this in more detail.

Date: 10.10.2023 Page: 65

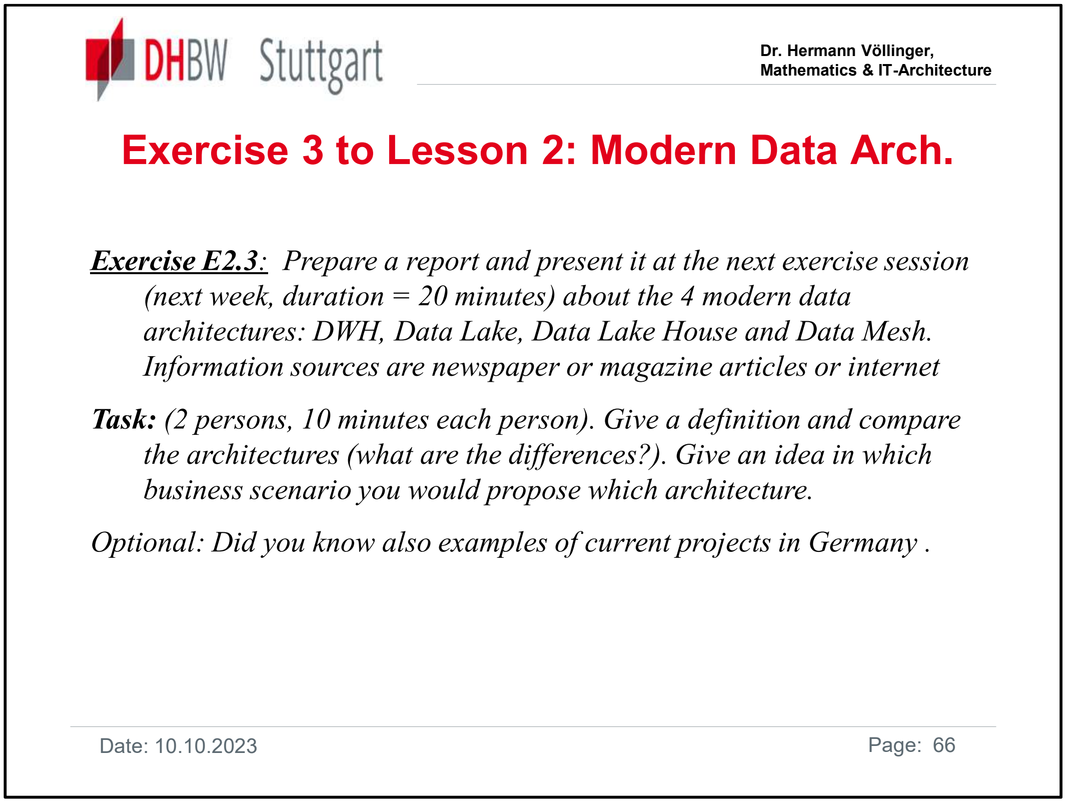
Further Information:

1. Gonzales, H., Han, J., Li, X., & Klabjan, D. (2006). Warehousing and Analyzing Massive RFID Data Sets. In *In Proc. of the 22nd ICDE Conf*. Atlanta: IEEE. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.558.4847&rep=rep1&type=pdf

2. Moody's Analytics. (2011). *Regulation Guide: An Introduction*. Retrieved from https://www.moodysanalytics.com/-/media/whitepaper/2011/11-01-03-regulation-guide-introduction.pdf

3.

4. Deutsche Bundesbank. (2018). Die Fertigstellung von Basel III. In *Monatsbericht - Januar 2018*. Retrieved from: https://www.bundesbank.de/resource/blob/693494/65d2da4c437491434426497fd0ff47fd/mL/2018-01-basel-3-data.pdf



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise 3 to Lesson 2: Modern Data Arch.

***Exercise E2.3**: Prepare a report and present it at the next exercise session (next week, duration = 20 minutes) about the 4 modern data architectures: DWH, Data Lake, Data Lake House and Data Mesh. Information sources are newspaper or magazine articles or internet*

***Task:** (2 persons, 10 minutes each person). Give a definition and compare the architectures (what are the differences?). Give an idea in which business scenario you would propose which architecture.*

Optional: Did you know also examples of current projects in Germany .

Date: 10.10.2023 Page: 66

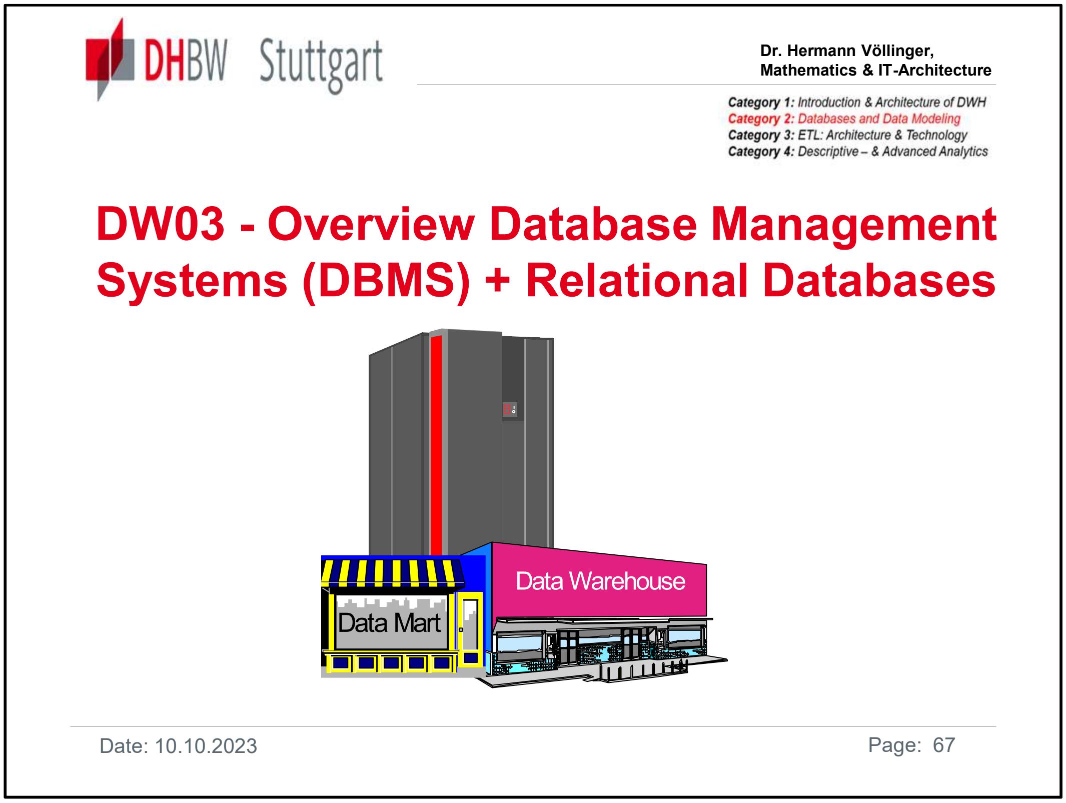
Source to Data Mesh: https://www.datamesh-architecture.com/

Data Mesh-Definition: Data mesh is a relatively new approach to managing and scaling data in large organizations. It emphasizes the decentralization of data ownership and processing, treating data as a product, and using domain-oriented cross-functional teams to manage data at scale.

Key Principles:

- **Domain-Oriented Ownership**: Data is owned and managed by the domain or business unit that generates it, rather than by a central data team.
- **Data Products**: Data is treated as a product, with clear documentation, versioning, and accessibility.
- **Self-serve Data Infrastructure**: Data infrastructure is designed to be self-serve, enabling domain teams to manage their own data pipelines and analytics.
- **Federated Query**: Data can be queried across the organization through a federated approach, allowing teams to access data they need without needing to move or centralize it.

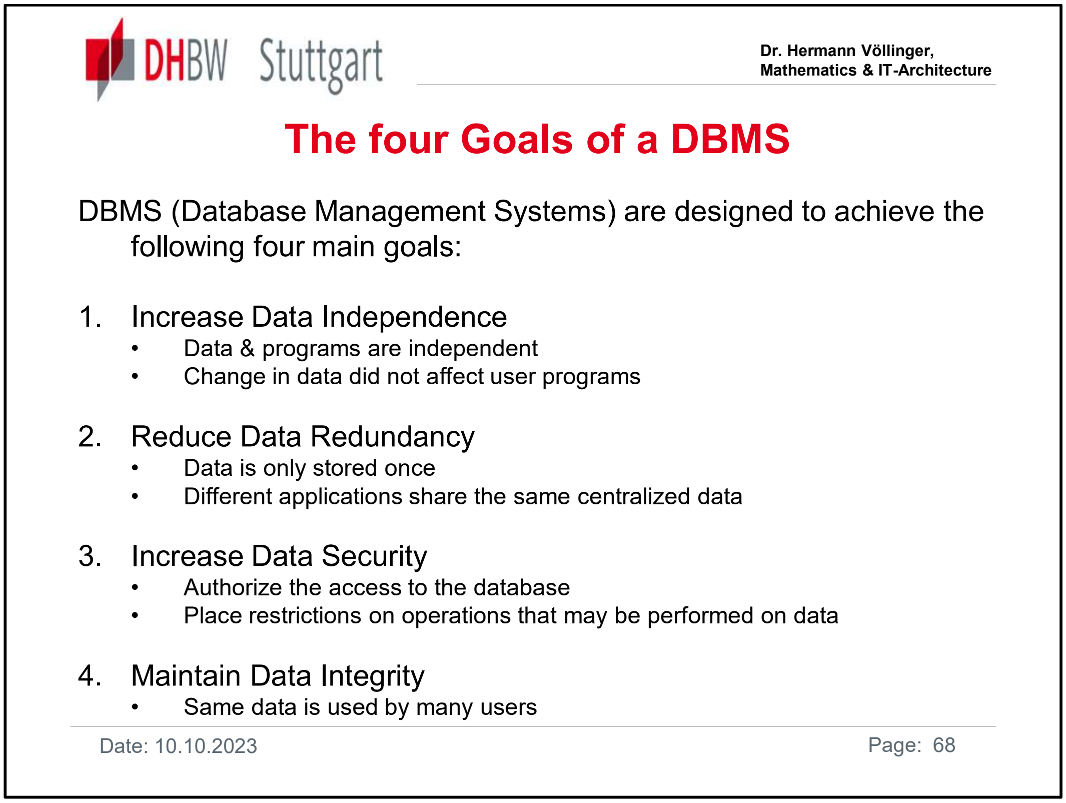
Remark: Data mesh and data virtualization are not the same. They are two distinct concepts in the field of data management and analytics, each with its own purpose and approach: **In summary**, data mesh is more of a conceptual and organizational framework for managing data at scale, emphasizing decentralized ownership and treating data as a product. Data virtualization, on the other hand, is a technology-driven approach for integrating and accessing data from various sources in real-time without the need for physical data movement. While they can complement each other in some contexts, they serve different purposes and focus on different aspects of data management and analytic.



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Category 1: *Introduction & Architecture of DWH*
Category 2: *Databases and Data Modeling*
Category 3: *ETL: Architecture & Technology*
Category 4: *Descriptive – & Advanced Analytics*

DW03 - Overview Database Management Systems (DBMS) + Relational Databases



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

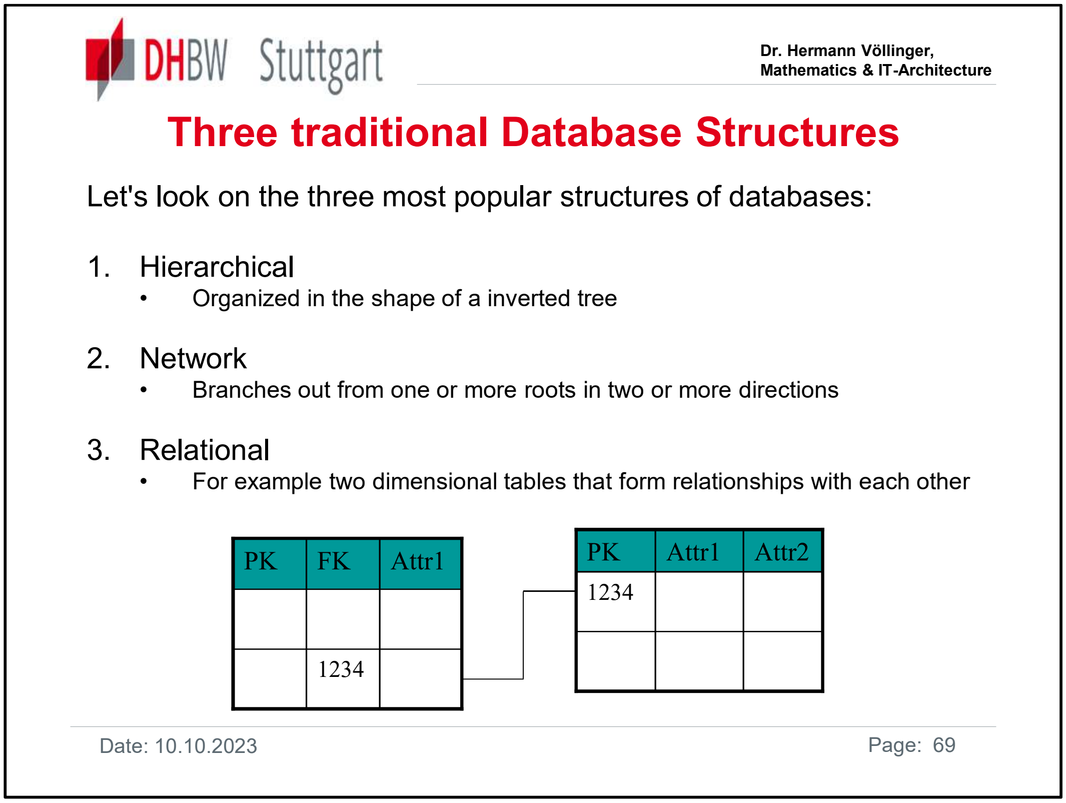
The four Goals of a DBMS

DBMS (Database Management Systems) are designed to achieve the following four main goals:

1. Increase Data Independence
   - Data & programs are independent
   - Change in data did not affect user programs
2. Reduce Data Redundancy
   - Data is only stored once
   - Different applications share the same centralized data
3. Increase Data Security
   - Authorize the access to the database
   - Place restrictions on operations that may be performed on data
4. Maintain Data Integrity
   - Same data is used by many users

Date: 10.10.2023 Page: 68

In traditional systems data and programs are dependent. This means if the data is changed the programs have also to be changed.-→ in database systems the processing of data and programs are **independent.**
In a convenient file systems data (for example a customer address) is stored in many files (for example: customer record, purchase order, accounts receivable) -→ in a database the data is only once stored (**no data redundancy**)
Databases use password protection to get access to DB, also further restrictions how to uses data (i.e. read or write protection)-→ **security**
Data Integrity is important, since the data are shared by many users



Three traditional Database Structures

Let's look on the three most popular structures of databases:

1. Hierarchical
   - Organized in the shape of a inverted tree
2. Network
   - Branches out from one or more roots in two or more directions
3. Relational
   - For example two dimensional tables that form relationships with each other

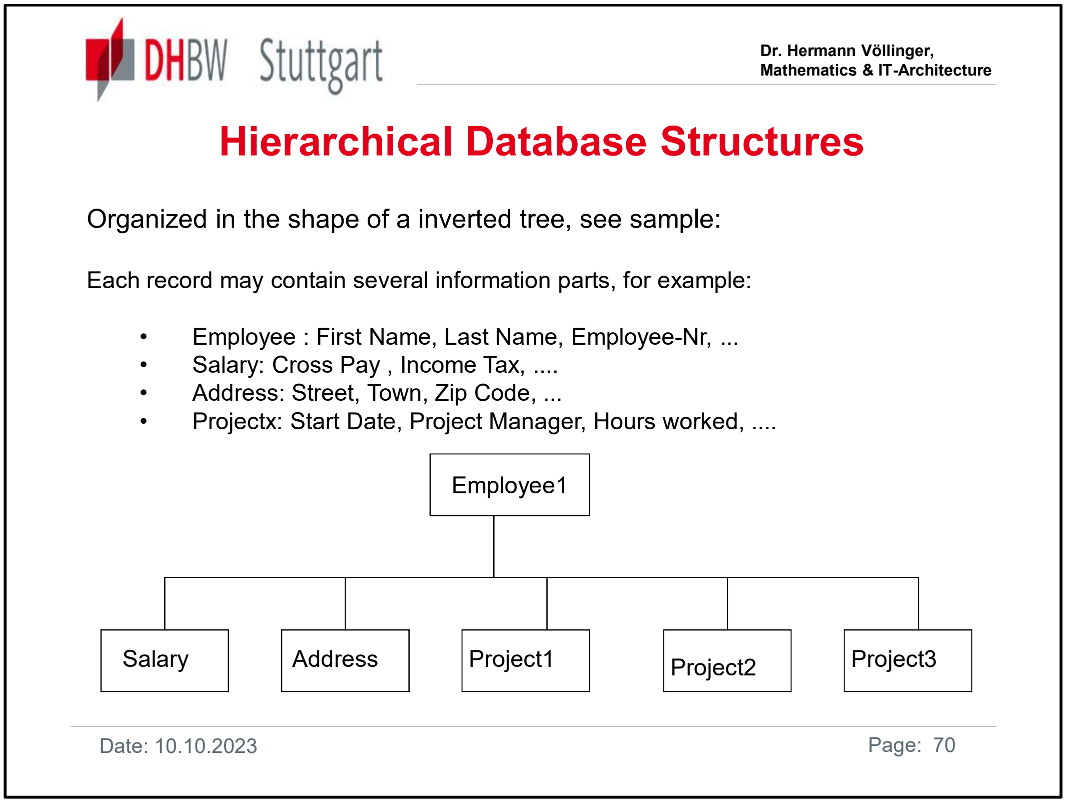
PK	FK	Attr1
	1234	

PK	Attr1	Attr2
1234		

These are the most popular DBMS. Beside there exist also some other DBs like:
VSAM - Virtual **S**torage **A**ccess **M**ethod, i.e. a key-sequenced data sets or files with an index containing extensive data set and volume information.
VSAM Overview: In the early 1970s, IBM introduced a collection of three data set organizations - sequential, indexed, and direct-access, together with the access methods and utilities to be used on the mainframe operating systems. This collection of data set organizations is called the Virtual Storage Access Method (VSAM).
The word *virtual* relates to the fact that VSAM was introduced at approximately the same time as the initial IBM virtual storage operating systems OS/VS1 and OS/VS2. VSAM was developed to replace the Indexed Sequential Access Method (ISAM), which is a much older technology. ISAM has major processing overheads which IBM wanted to improve

What is VSAM?

VSAM is one of several access methods that defines the technique by which data is stored and retrieved. A GET/PUT interface is used to transfer data from a direct access storage device (DASD) to an application program. VSAM does not support data stored on tape. VSAM is used to organize and access data, and maintain information about this data which is stored or referenced in a catalogue.
VSAM data sets must be catalogued in an integrated catalogue facility (ICF) structure. Records are arranged by an index key or by relative byte addressing. VSAM uses direct or sequential processing of fixed and variable length records on DASD.
There are two major parts to VSAM: catalogue management and record management.



Hierarchical Database Structures

Organized in the shape of a inverted tree, see sample:

Each record may contain several information parts, for example:

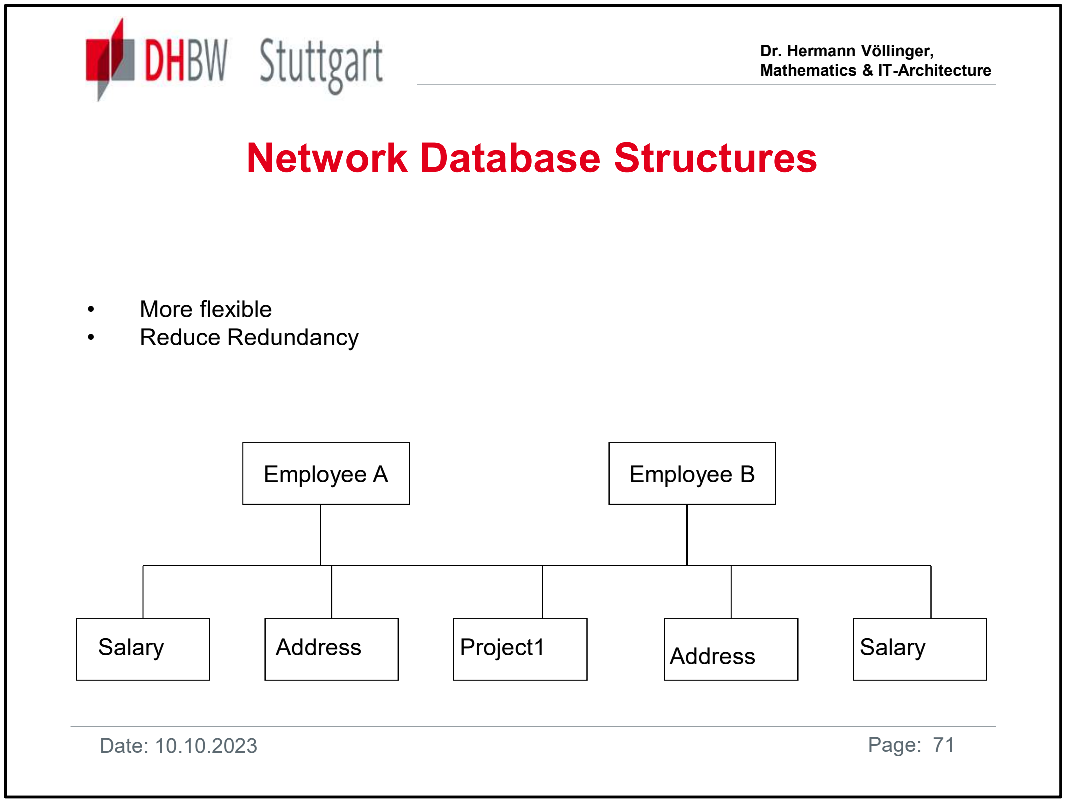
- Employee : First Name, Last Name, Employee-Nr, ...
- Salary: Cross Pay , Income Tax, ...
- Address: Street, Town, Zip Code, ...
- Projectx: Start Date, Project Manager, Hours worked, ...

Employee1

Salary
Address
Project1
Project2
Project3

See an example on the slide:
An employee can work on more than one project only. There exist one parent data set and several children data sets.
The technological reason for this structure are the access possibilities of magnet-tapes.
Like the network model the **hierarchical model** stores its data in a series of **records**, which have a set of field values attached to it. It collects all the instances of a specific record together as a **record type**. These record types are the equivalent of tables in the relational model, and with the individual records being the equivalent of rows.
To create links between these record types, the hierarchical model uses **Parent Child Relationships**. These are a 1:N mapping between record types. This is done by using trees, like set theory used in the relational model, "borrowed" from mathematics.
Unlike the network model, the hierarchical model is only able to cope with a single tree and is not able to cope with linking between branches or over multiple layers. The hierarchical model is more structured that the network model, since it only allows a single tree, with no links between layers in different branches of the tree. This means that it has a much more structured form than the relational model, this was done to improve throughput for transactions (adding, updating and deleting records) and to increase the simplicity of the interface for users.
An example for such a database is IMS: **IMS**- **I**nformation **M**anagement **S**ystem

For **more details** see: http://de.wikipedia.org/wiki/Information_Management_System



Network Database Structures

- More flexible
- Reduce Redundancy

Employee A

Employee B

Salary

Address

Project1

Address

Salary

One can build for example out of two hierarchical employee databases (db) one db, if the project is the same. Project1 occurs once, but both employees are connected to it -→ more flexible and reduce redundancy.

This is more flexible because each db has it‘s own set of rules which define the relationships between the records. A network db is similar to a hierarchical db, except the rules are not so strict.

Key Concepts: The network db model was first introduced in 1971 by CODASYL Data Base Task Group and because of this is sometimes called the DBTG model. It is therefore a contemporary of the Relational Model, both in terms of its age and its basis in research done in the 1960's.

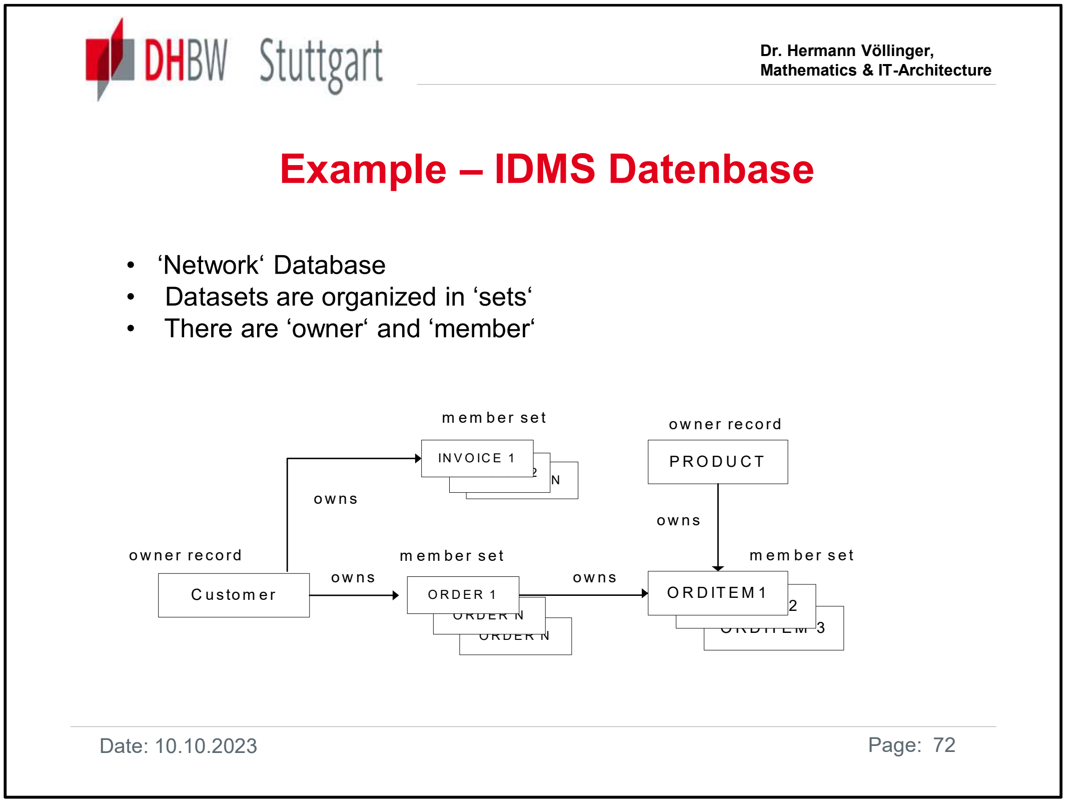
It is called the Network Model because it represents the data it contains in the form of a network of records and sets which are related to each other, forming a network of links.

To do this it uses records, record types and set types which we shall discuss later. Because this model is only used in legacy systems, and is being phased out over time, we shall only cover the basic information about it in this section.

•**Records** are sets of related data values. These are equivalent to rows in the relational model. They store the name of the record type, attributes associated with it and the format for these attributes.

•**Record Types** are set of records of same type. These are the equivalent of tables in a relational model.

•**Set Types**, are named, 1:N relationships between 2 record types. These do not have a direct parallel in the relational model, the closest comparison is to a query statement which has joined two tables together. This makes the network model faster to run certain queries but does not allow the flexibility of the relational model. **More details:** http://en.wikipedia.org/wiki/IDMS



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

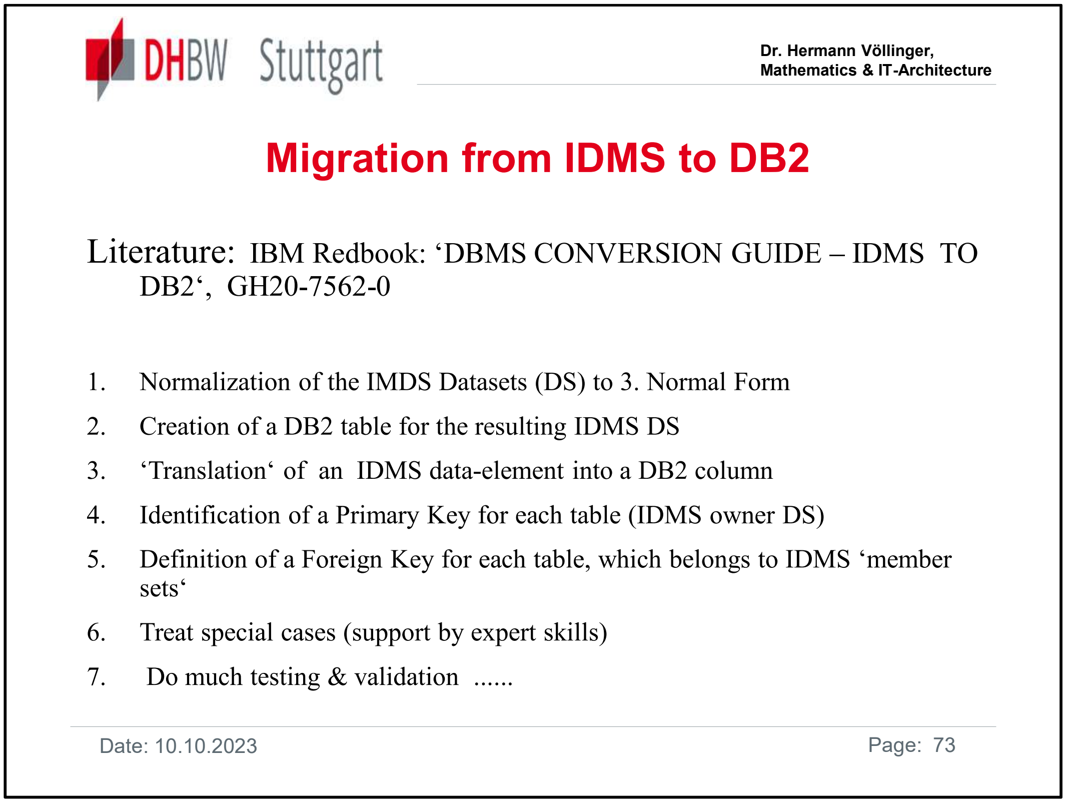
Example – IDMS Datenbase

- ‘Network‘ Database
- Datasets are organized in ‘sets‘
- There are ‘owner‘ and ‘member‘

Date: 10.10.2023

Page: 72

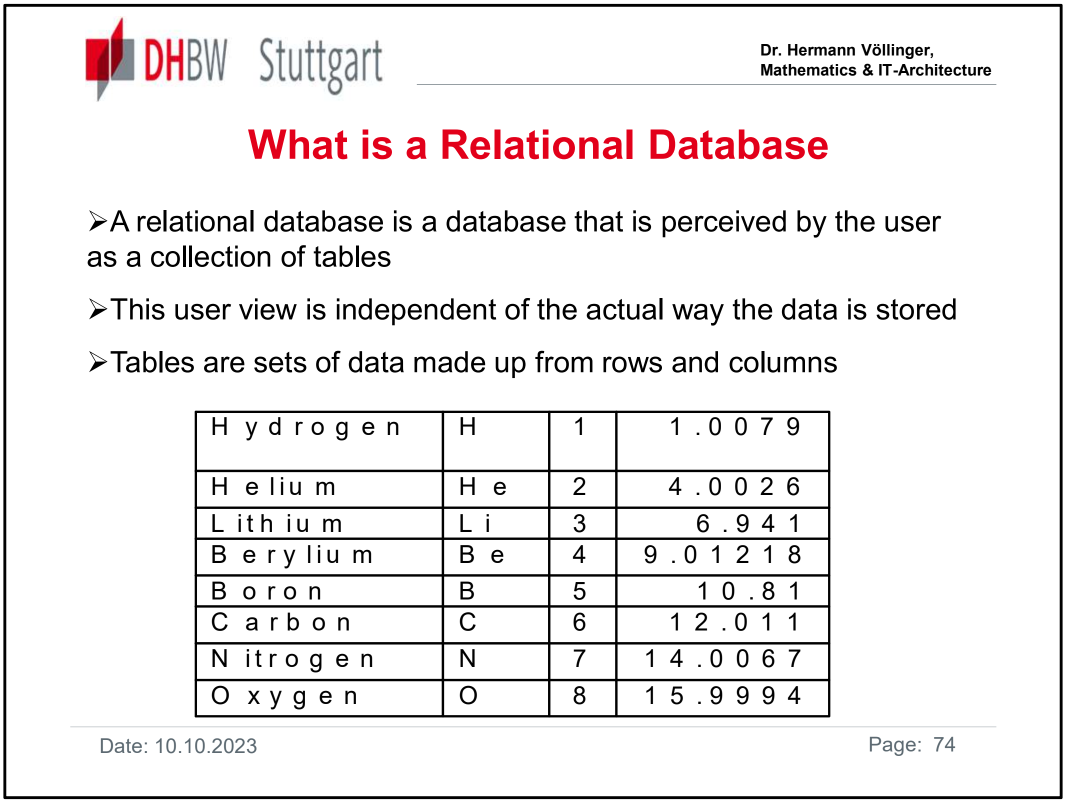
The network model is not commonly used today to design database systems, however, there are a few instances of it being used by companies as part of a legacy system. Because it represents data in a format closer to the way in which it stores it than the other models, it can be faster than relational systems.
Unfortunately, this increase in speed is at the cost of its adaptability.



Migration from IDMS to DB2

Literature: IBM Redbook: ‘DBMS CONVERSION GUIDE – IDMS TO DB2‘, GH20-7562-0

1. Normalization of the IMDS Datasets (DS) to 3. Normal Form
2. Creation of a DB2 table for the resulting IDMS DS
3. ‘Translation‘ of an IDMS data-element into a DB2 column
4. Identification of a Primary Key for each table (IDMS owner DS)
5. Definition of a Foreign Key for each table, which belongs to IDMS ‘member sets‘
6. Treat special cases (support by expert skills)
7. Do much testing & validation ......



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

What is a Relational Database

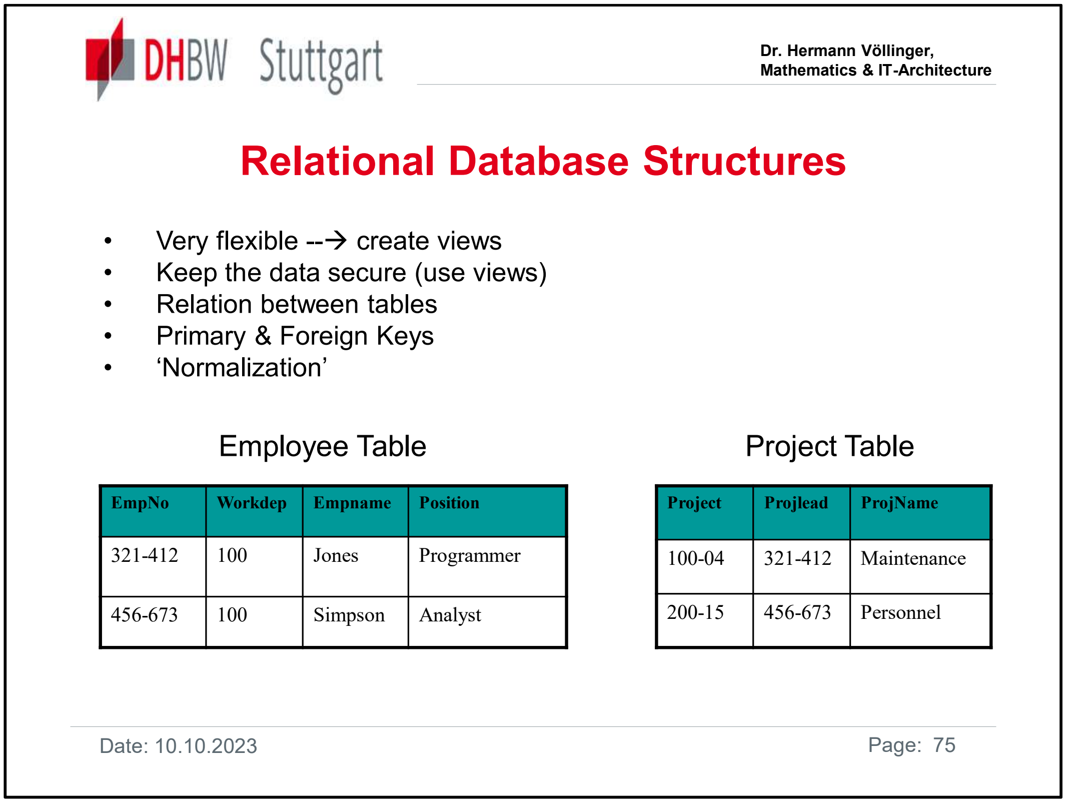
- ➢A relational database is a database that is perceived by the user as a collection of tables
- ➢This user view is independent of the actual way the data is stored
- ➢Tables are sets of data made up from rows and columns

Hydrogen	H	1	1.0079
Helium	He	2	4.0026
Lithium	Li	3	6.941
Berylium	Be	4	9.01218
Boron	B	5	10.81
Carbon	C	6	12.011
Nitrogen	N	7	14.0067
Oxygen	O	8	15.9994

Date: 10.10.2023

Page: 74

You are probably familiar with the term relational database.
In a relational database the user views everything as a set of tables.
These tables are made up of rows and columns. They might be very large.



Relational Database Structures

- Very flexible --→ create views
- Keep the data secure (use views)
- Relation between tables
- Primary & Foreign Keys
- ‘Normalization’

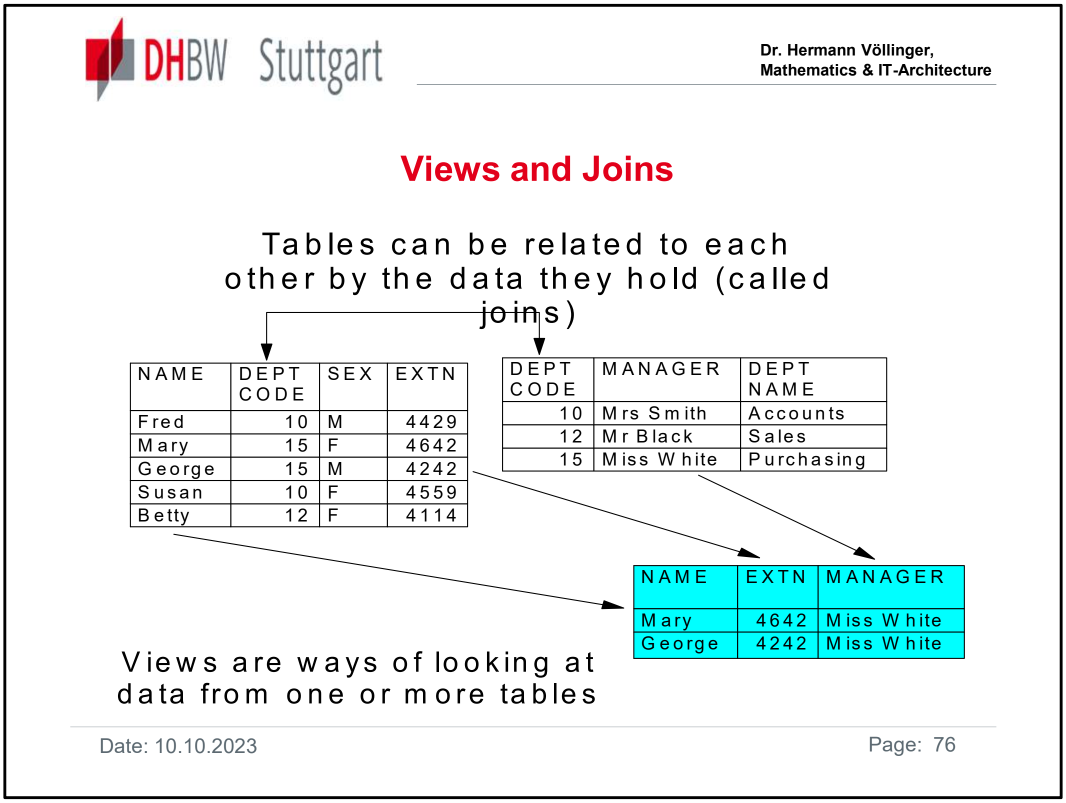
Employee Table

EmpNo	Workdep	Empname	Position
321-412	100	Jones	Programmer
456-673	100	Simpson	Analyst

Project Table

Project	Projlead	ProjName
100-04	321-412	Maintenance
200-15	456-673	Personnel

These are the two tables which represent employee and project information, salary and address information could be stored in other tables.
You can create a relation between the two tables by the EmpNo and the ProjLead attribute.



Views and Joins

Tables can be related to each other by the data they hold (called joins)

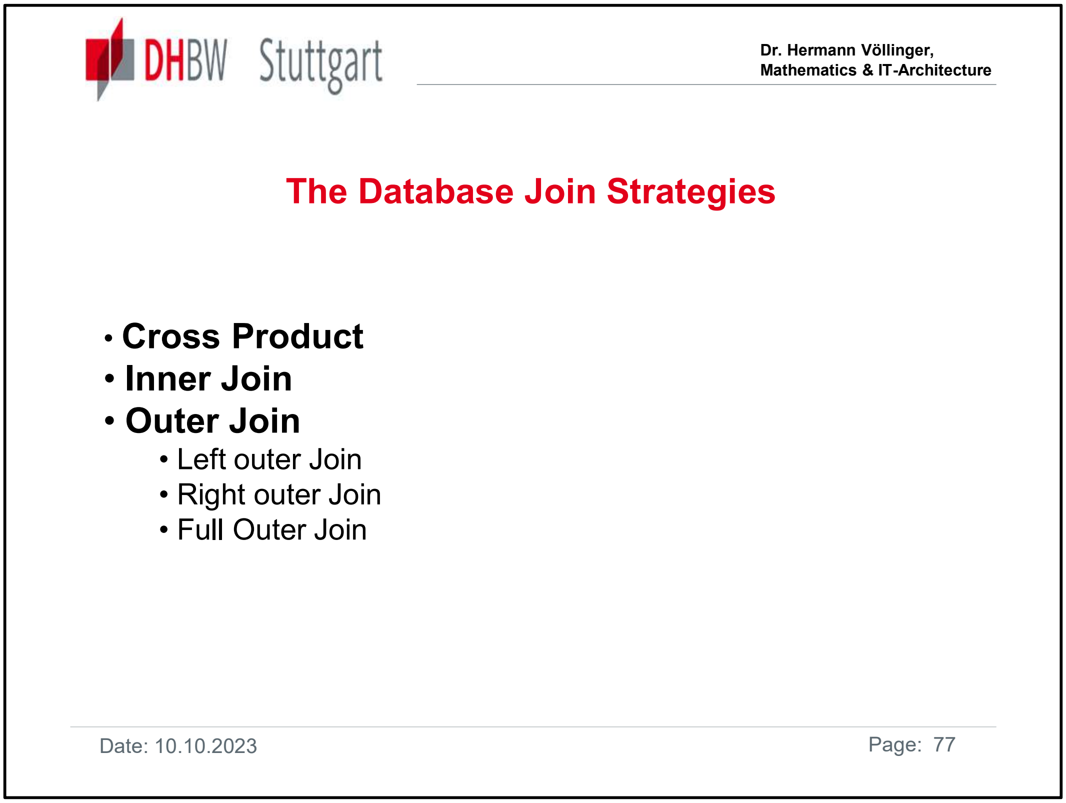
NAME	DEPT CODE	SEX	EXTN
Fred	10	M	4429
Mary	15	F	4642
George	15	M	4242
Susan	10	F	4559
Betty	12	F	4114

DEPT CODE	MANAGER	DEPT NAME
10	Mrs Smith	Accounts
12	Mr Black	Sales
15	Miss White	Purchasing

NAME	EXTN	MANAGER
Mary	4642	Miss White
George	4242	Miss White

Views are ways of looking at data from one or more tables

The information you are looking for is hold in different tables.
To get the relevant Information you are looking for, one has to **join** over different tables using a common key, here: DEPT CODE
These joins can be **performance bottlenecks** for large tables.
The process of combining data from two or more tables is called **joining tables**. The database manager forms all combinations of rows from the specified tables. For each combination, it tests the join condition.
A join condition is a search condition, with some restrictions. For a list of restrictions refer to the SQL Reference.
Note that the data types of the columns involved in the join condition do not have to be identical; however, they must be compatible. The join condition is evaluated the same way as any other search condition, and the same rules for comparisons apply.
If you do not specify a join condition, all combinations of rows from tables listed in the FROM clause are returned, even though the rows may be completely unrelated.
The result is referred to as the **cross product** of the two tables.
The two main types of joins are **inner joins** and **outer joins**. So far, in all of our examples we have used the inner join.



The Database Join Strategies

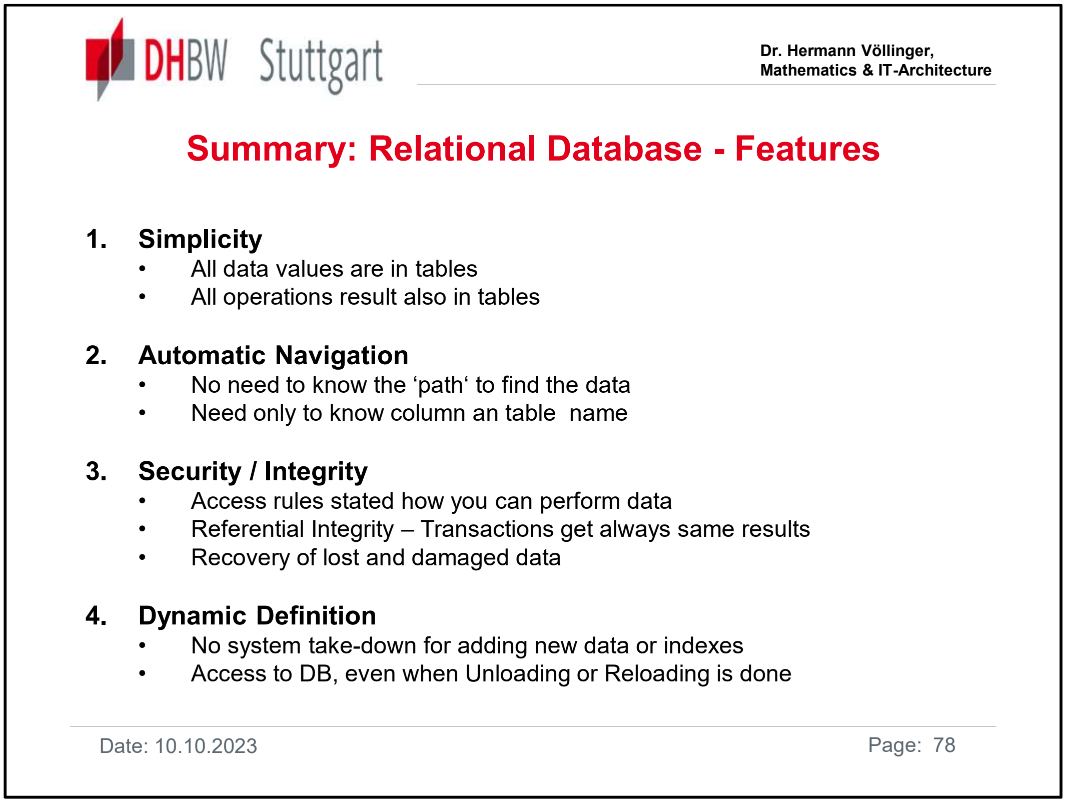
- **Cross Product**
- **Inner Join**
- **Outer Join**
  - Left outer Join
  - Right outer Join
  - Full Outer Join

Inner joins keep only the rows from the cross product that meet the join condition.
If a row exists in one table, but not the other, the information is not included in the result table.
Outer joins are a concatenation of the inner join and rows from the left table, right table, or both tables that are missing from the inner join.
When you perform an outer join on two tables, you arbitrarily assign one table as the left table and the other one as the right table.

There are three types of outer joins:

1. **left outer join** includes the inner join and the rows from the left table that are not included in the inner join.
2. **right outer join** includes the inner join and the rows from the right table that are not included in the inner join.
3. **full outer join** includes the inner join and the rows from both the left and right tables that are not included in the inner join.

Use the SELECT statement to specify the columns to be displayed.
In the FROM clause, list the name of the first table followed by the keywords LEFT OUTER JOIN, RIGHT OUTER JOIN or FULL OUTER JOIN.
Next you need to specify the second table followed by the ON keyword. Following the ON keyword, specify the join condition to express a relationship between the tables to be joined.



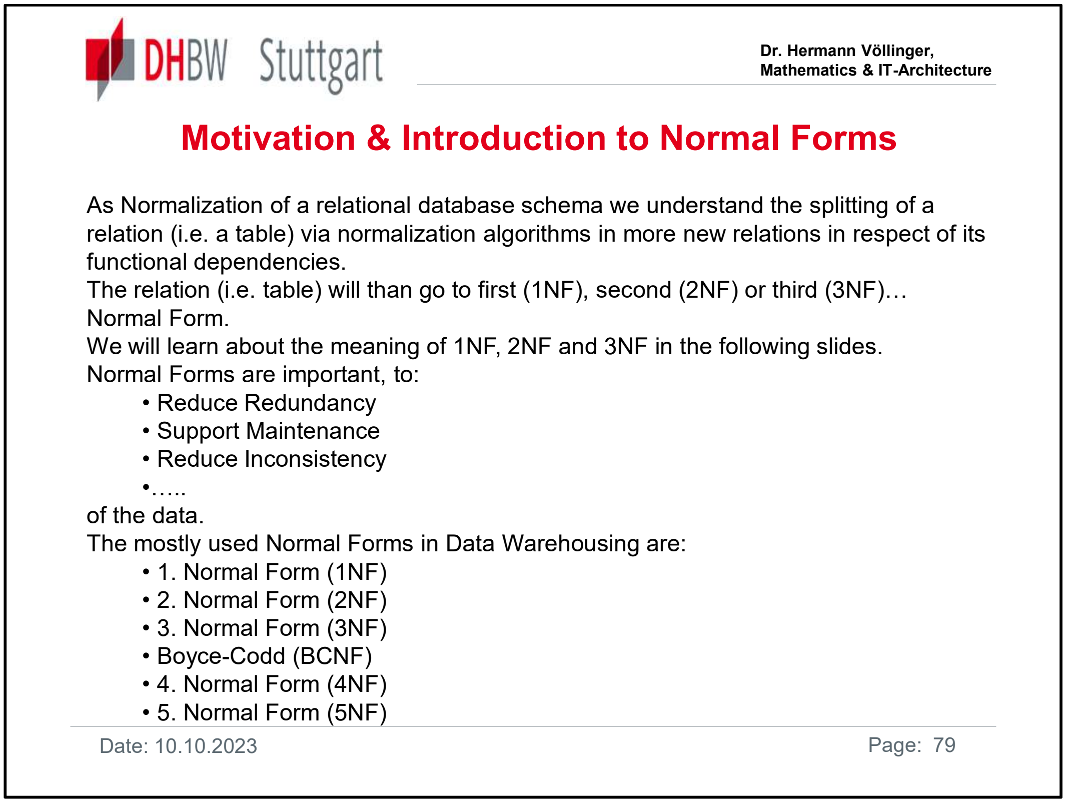
Summary: Relational Database - Features

1. **Simplicity**
   - All data values are in tables
   - All operations result also in tables

2. **Automatic Navigation**
   - No need to know the ‘path‘ to find the data
   - Need only to know column an table name

3. **Security / Integrity**
   - Access rules stated how you can perform data
   - Referential Integrity – Transactions get always same results
   - Recovery of lost and damaged data

4. **Dynamic Definition**
   - No system take-down for adding new data or indexes
   - Access to DB, even when Unloading or Reloading is done



Motivation & Introduction to Normal Forms

As Normalization of a relational database schema we understand the splitting of a relation (i.e. a table) via normalization algorithms in more new relations in respect of its functional dependencies.
The relation (i.e. table) will than go to first (1NF), second (2NF) or third (3NF)… Normal Form.
We will learn about the meaning of 1NF, 2NF and 3NF in the following slides.
Normal Forms are important, to:

- Reduce Redundancy
- Support Maintenance
- Reduce Inconsistency
- …..

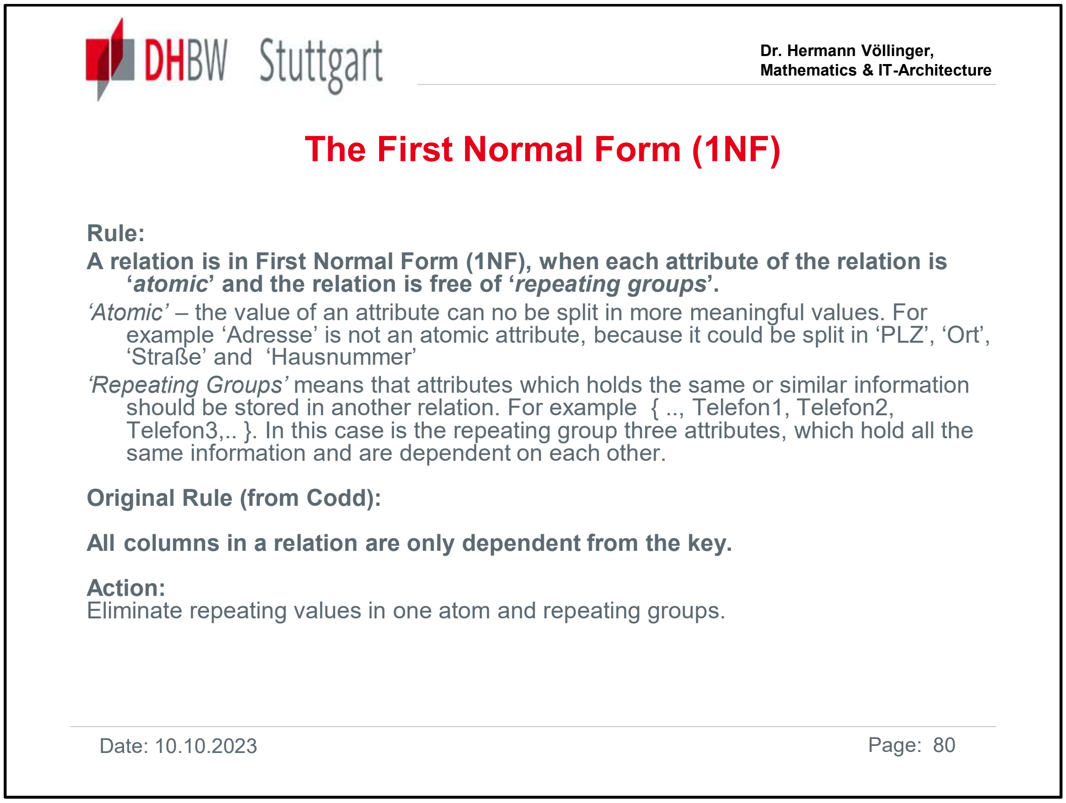
of the data.
The mostly used Normal Forms in Data Warehousing are:

- 1. Normal Form (1NF)
- 2. Normal Form (2NF)
- 3. Normal Form (3NF)
- Boyce-Codd (BCNF)
- 4. Normal Form (4NF)
- 5. Normal Form (5NF)

During the Normalization we split the columns of tables inside a database in new columns, for example we split addresses in post-code, town, street and house number, or we relate these columns with other tables for example a customer is related with a customer table via a concrete key.
The goal of the normalization is to reduce Redundancy (same or double information), to omit Anomalies (i.e. two data content which could not be true for both). Normalization will reduce the effort for maintenance of a database and will also guarantee the consistence of the data..
For example when we change the address of a customer in a database we have to look for all address information of the customer in a database and have to change them all. In a normalized database we have only to change one dataset, because the customer is stored only once in the database.
……………..
In some cases (which we will see later) it makes sense not to normalize the database. Because of :

- Better query performance
- More simpler queries
- or some more reasons

See as a reference:
https://en.wikipedia.org/wiki/Database_normalization



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

The First Normal Form (1NF)

Rule:

A relation is in First Normal Form (1NF), when each attribute of the relation is '*atomic*' and the relation is free of '*repeating groups*'.

'Atomic' – the value of an attribute can no be split in more meaningful values. For example 'Adresse' is not an atomic attribute, because it could be split in 'PLZ', 'Ort', 'Straße' and 'Hausnummer'

'Repeating Groups' means that attributes which holds the same or similar information should be stored in another relation. For example { .., Telefon1, Telefon2, Telefon3,.. }. In this case is the repeating group three attributes, which hold all the same information and are dependent on each other.

Original Rule (from Codd):

All columns in a relation are only dependent from the key.

Action:
Eliminate repeating values in one atom and repeating groups.

Date: 10.10.2023 Page: 80

For further definition of normal forms we need some formal notations:

Functional dependency: Attribute B has a functional dependency on attribute A if, for each value of attribute A, there is exactly one value of attribute B. For example, Employee Address has a functional dependency on Employee ID, because a particular Employee Address value corresponds to every Employee ID value. An attribute may be functionally dependent either on a single attribute or on a combination of attributes.
It is not possible to determine the extent to which a design is normalized without understanding what functional dependencies apply to the attributes within its tables; understanding this, in turn, requires knowledge of the problem domain.

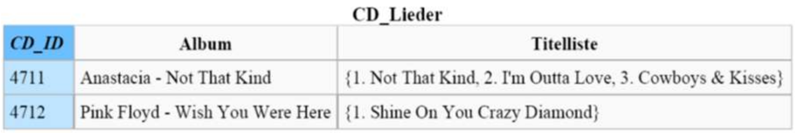
Trivial functional dependency: A trivial functional dependency is a functional dependency of an attribute on a superset of itself.
{Employee ID, Employee Address}-→ {Employee Address} is trivial, as is {Employee Address} → {Employee Address}.

Full functional dependency: An attribute is fully functionally dependent on a set of attributes X if it is 1. functionally dependent on X, and 2. not functionally dependent on any proper subset of X.
{Employee Address} has a functional dependency on {Employee ID, Skill}, but not a full functional dependency, for it is also dependent on {Employee ID}.

Transitive dependency: A transitive dependency is an indirect functional dependency, one in which X→Z only by virtue of X→Y and Y→Z.

Example for First Normal Form *(‘Atomic‘)*

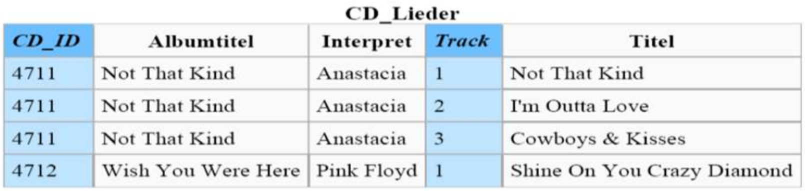
The following table is not in First Normal Form (*examples are from WIKIPEDIA).
The attribute ‘*Album*‘ has information about *Interpret* and *CD Title* ......



CD_Lieder

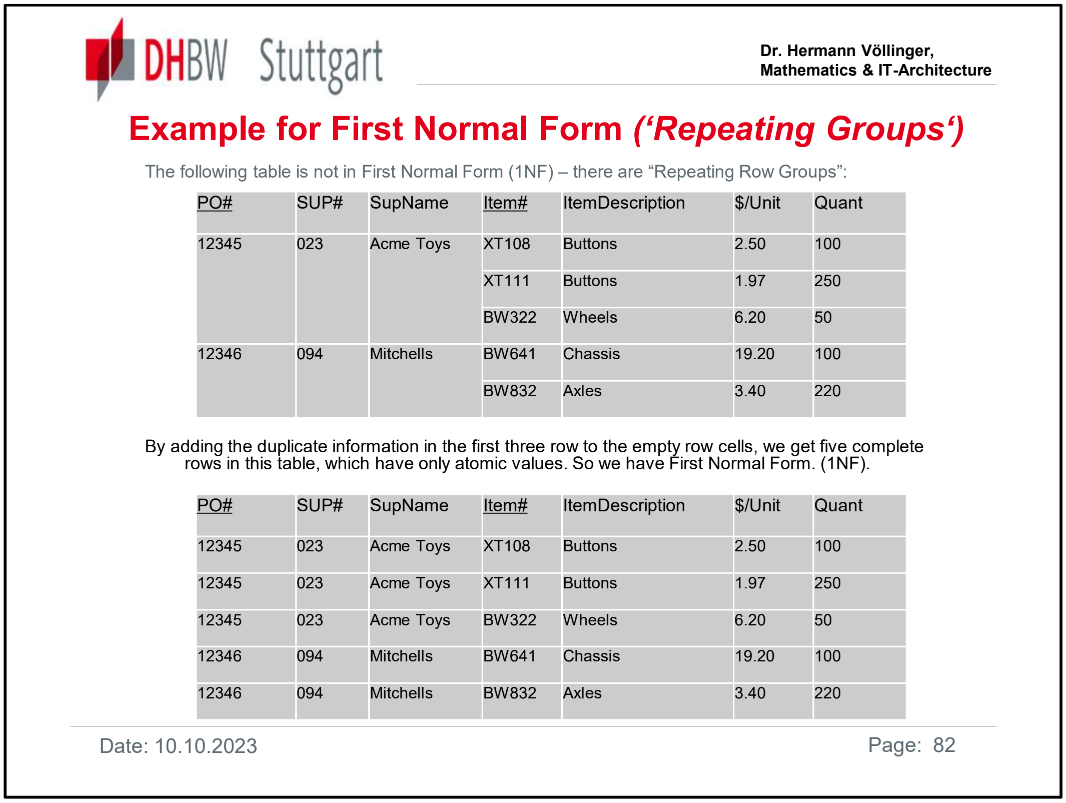
CD_ID	Album	Titelliste
4711	Anastacia - Not That Kind	{1. Not That Kind, 2. I'm Outta Love, 3. Cowboys & Kisses}
4712	Pink Floyd - Wish You Were Here	{1. Shine On You Crazy Diamond}

The attributes *‘Album‘* and ‘*Titelliste*’ are split in atomic attributes. ‘*Titelliste*’ is split in ‘*Track*’ and *‘Titel’.*



CD_Lieder

CD_ID	Albumtitel	Interpret	*Track*	Titel
4711	Not That Kind	Anastacia	1	Not That Kind
4711	Not That Kind	Anastacia	2	I'm Outta Love
4711	Not That Kind	Anastacia	3	Cowboys & Kisses
4712	Wish You Were Here	Pink Floyd	1	Shine On You Crazy Diamond



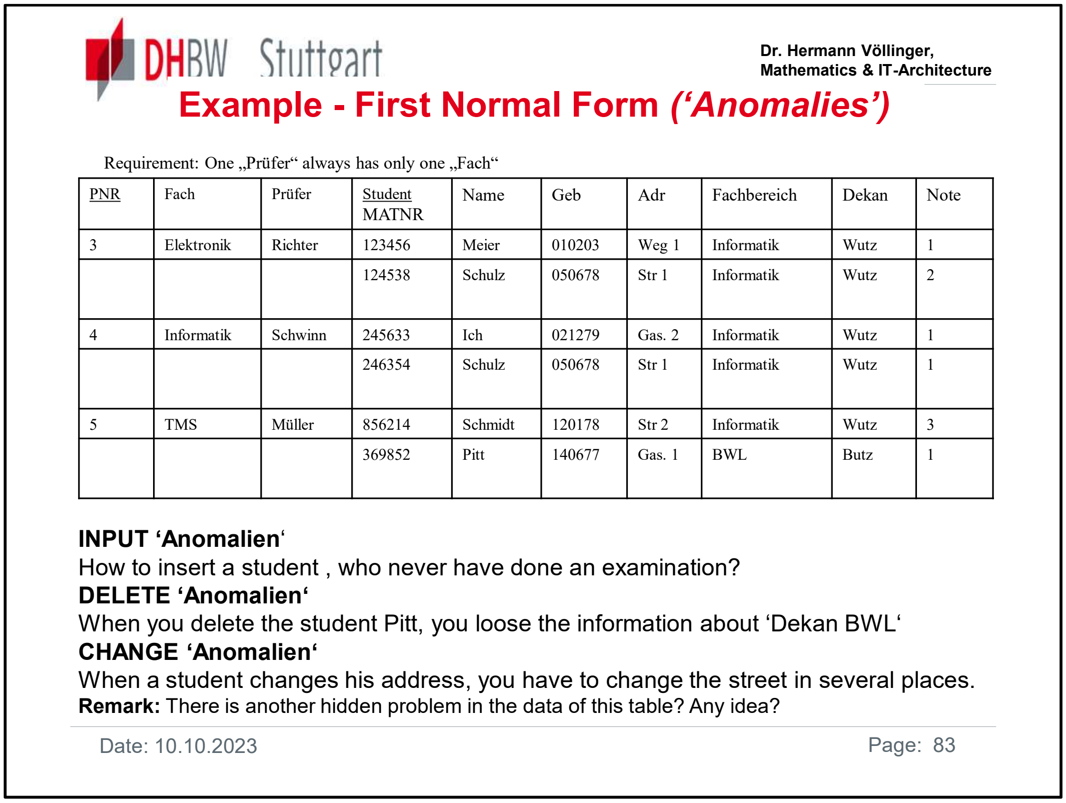
Example for First Normal Form *(‘Repeating Groups‘)*

The following table is not in First Normal Form (1NF) – there are “Repeating Row Groups”:

PO#	SUP#	SupName	Item#	ItemDescription	$/Unit	Quant
12345	023	Acme Toys	XT108	Buttons	2.50	100
			XT111	Buttons	1.97	250
			BW322	Wheels	6.20	50
12346	094	Mitchells	BW641	Chassis	19.20	100
			BW832	Axles	3.40	220

By adding the duplicate information in the first three row to the empty row cells, we get five complete rows in this table, which have only atomic values. So we have First Normal Form. (1NF).

PO#	SUP#	SupName	Item#	ItemDescription	$/Unit	Quant
12345	023	Acme Toys	XT108	Buttons	2.50	100
12345	023	Acme Toys	XT111	Buttons	1.97	250
12345	023	Acme Toys	BW322	Wheels	6.20	50
12346	094	Mitchells	BW641	Chassis	19.20	100
12346	094	Mitchells	BW832	Axles	3.40	220

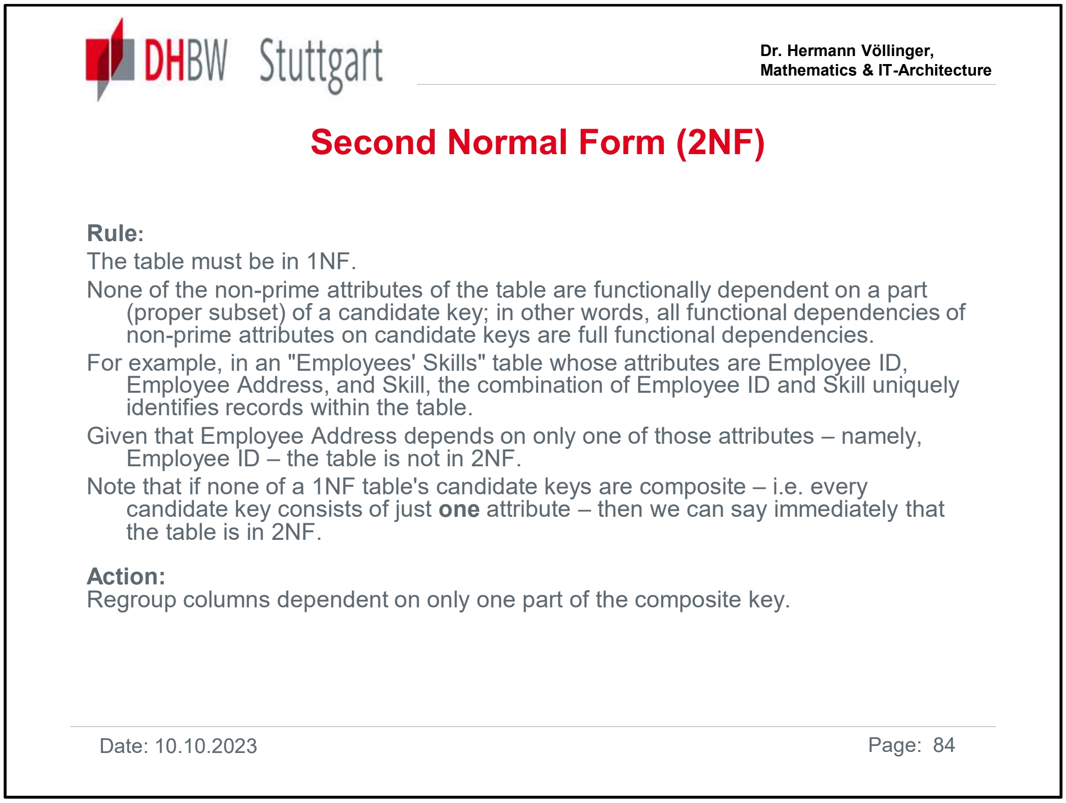


Example - First Normal Form *('Anomalies')*

Requirement: One „Prüfer" always has only one „Fach"

PNR	Fach	Prüfer	Student MATNR	Name	Geb	Adr	Fachbereich	Dekan	Note
3	Elektronik	Richter	123456	Meier	010203	Weg 1	Informatik	Wutz	1
			124538	Schulz	050678	Str 1	Informatik	Wutz	2
4	Informatik	Schwinn	245633	Ich	021279	Gas. 2	Informatik	Wutz	1
			246354	Schulz	050678	Str 1	Informatik	Wutz	1
5	TMS	Müller	856214	Schmidt	120178	Str 2	Informatik	Wutz	3
			369852	Pitt	140677	Gas. 1	BWL	Butz	1

INPUT 'Anomalien'
How to insert a student , who never have done an examination?
DELETE 'Anomalien'
When you delete the student Pitt, you loose the information about 'Dekan BWL'
CHANGE 'Anomalien'
When a student changes his address, you have to change the street in several places.
Remark: There is another hidden problem in the data of this table? Any idea?



Second Normal Form (2NF)

Rule:

The table must be in 1NF.

None of the non-prime attributes of the table are functionally dependent on a part (proper subset) of a candidate key; in other words, all functional dependencies of non-prime attributes on candidate keys are full functional dependencies.

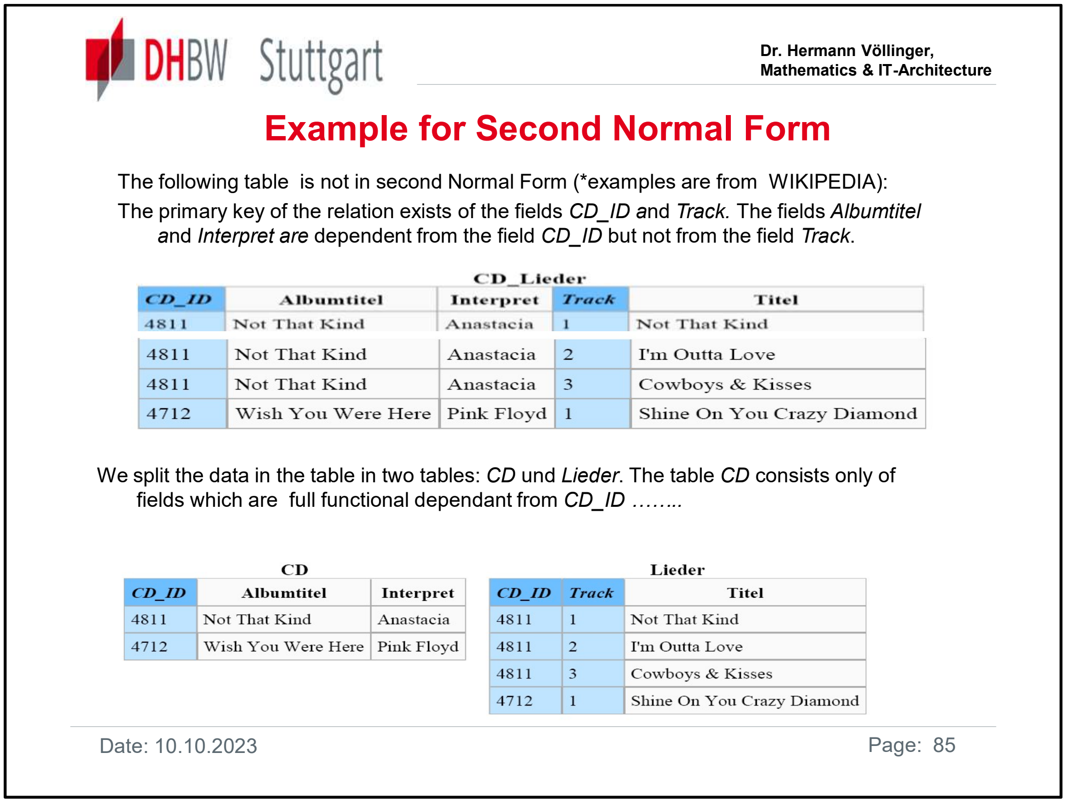
For example, in an "Employees' Skills" table whose attributes are Employee ID, Employee Address, and Skill, the combination of Employee ID and Skill uniquely identifies records within the table.

Given that Employee Address depends on only one of those attributes – namely, Employee ID – the table is not in 2NF.

Note that if none of a 1NF table's candidate keys are composite – i.e. every candidate key consists of just **one** attribute – then we can say immediately that the table is in 2NF.

Action:

Regroup columns dependent on only one part of the composite key.



Example for Second Normal Form

The following table is not in second Normal Form (*examples are from WIKIPEDIA):

The primary key of the relation exists of the fields *CD_ID* and *Track*. The fields *Albumtitel* and *Interpret are* dependent from the field *CD_ID* but not from the field *Track*.

CD_Lieder

CD_ID	Albumtitel	Interpret	*Track*	Titel
4811	Not That Kind	Anastacia	1	Not That Kind
4811	Not That Kind	Anastacia	2	I'm Outta Love
4811	Not That Kind	Anastacia	3	Cowboys & Kisses
4712	Wish You Were Here	Pink Floyd	1	Shine On You Crazy Diamond

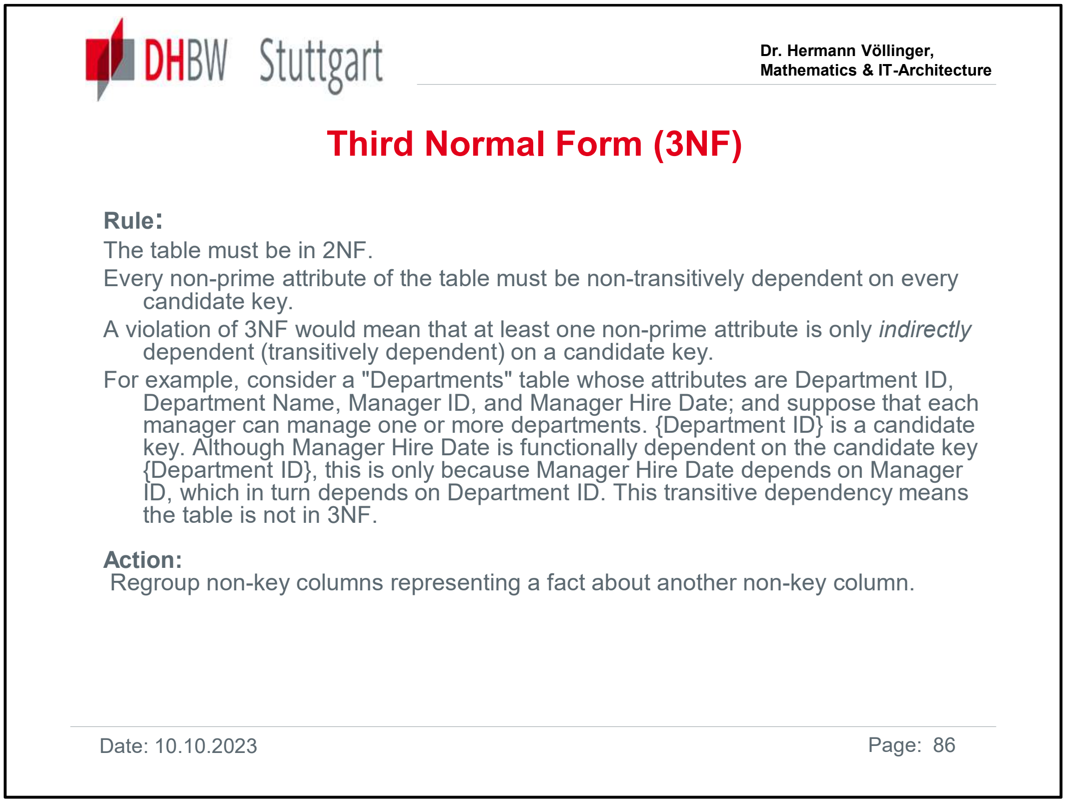
We split the data in the table in two tables: *CD* und *Lieder*. The table *CD* consists only of fields which are full functional dependant from *CD_ID* ……..

CD

CD_ID	Albumtitel	Interpret
4811	Not That Kind	Anastacia
4712	Wish You Were Here	Pink Floyd

Lieder

CD_ID	*Track*	Titel
4811	1	Not That Kind
4811	2	I'm Outta Love
4811	3	Cowboys & Kisses
4712	1	Shine On You Crazy Diamond



Third Normal Form (3NF)

Rule:

The table must be in 2NF.

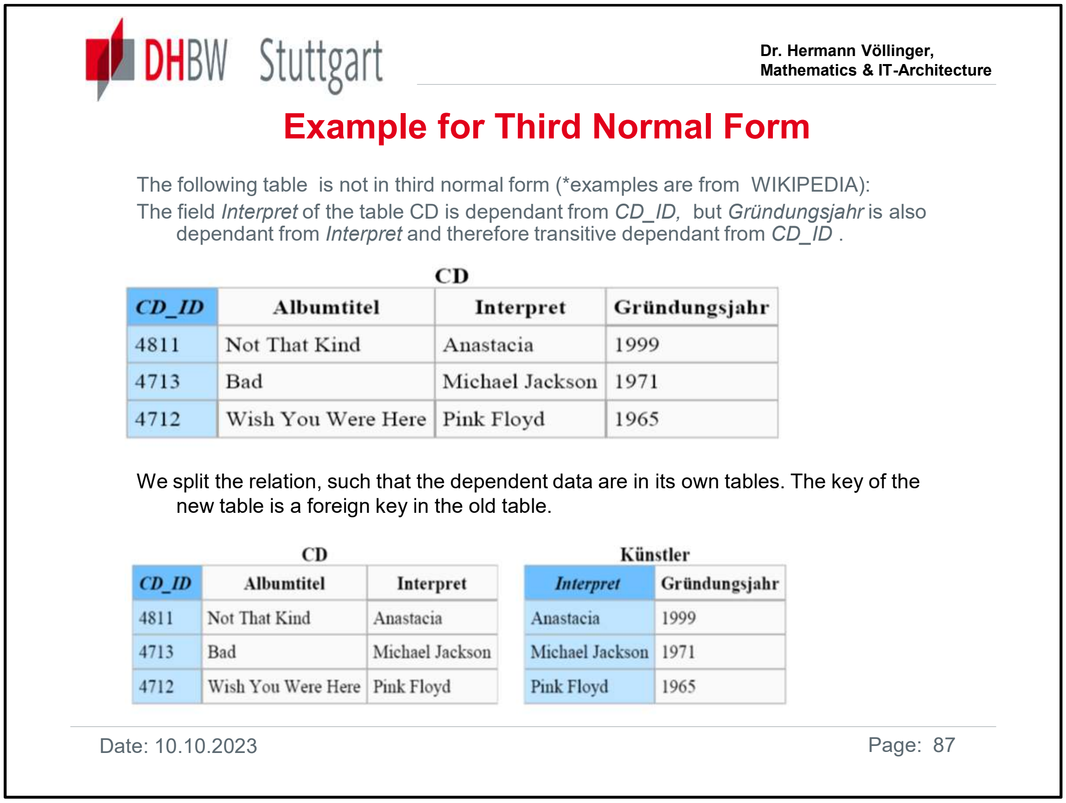
Every non-prime attribute of the table must be non-transitively dependent on every candidate key.

A violation of 3NF would mean that at least one non-prime attribute is only *indirectly* dependent (transitively dependent) on a candidate key.

For example, consider a "Departments" table whose attributes are Department ID, Department Name, Manager ID, and Manager Hire Date; and suppose that each manager can manage one or more departments. {Department ID} is a candidate key. Although Manager Hire Date is functionally dependent on the candidate key {Department ID}, this is only because Manager Hire Date depends on Manager ID, which in turn depends on Department ID. This transitive dependency means the table is not in 3NF.

Action:

Regroup non-key columns representing a fact about another non-key column.



Example for Third Normal Form

The following table is not in third normal form (*examples are from WIKIPEDIA):

The field *Interpret* of the table CD is dependant from *CD_ID,* but *Gründungsjahr* is also dependant from *Interpret* and therefore transitive dependant from *CD_ID* .

CD

CD_ID	Albumtitel	Interpret	Gründungsjahr
4811	Not That Kind	Anastacia	1999
4713	Bad	Michael Jackson	1971
4712	Wish You Were Here	Pink Floyd	1965

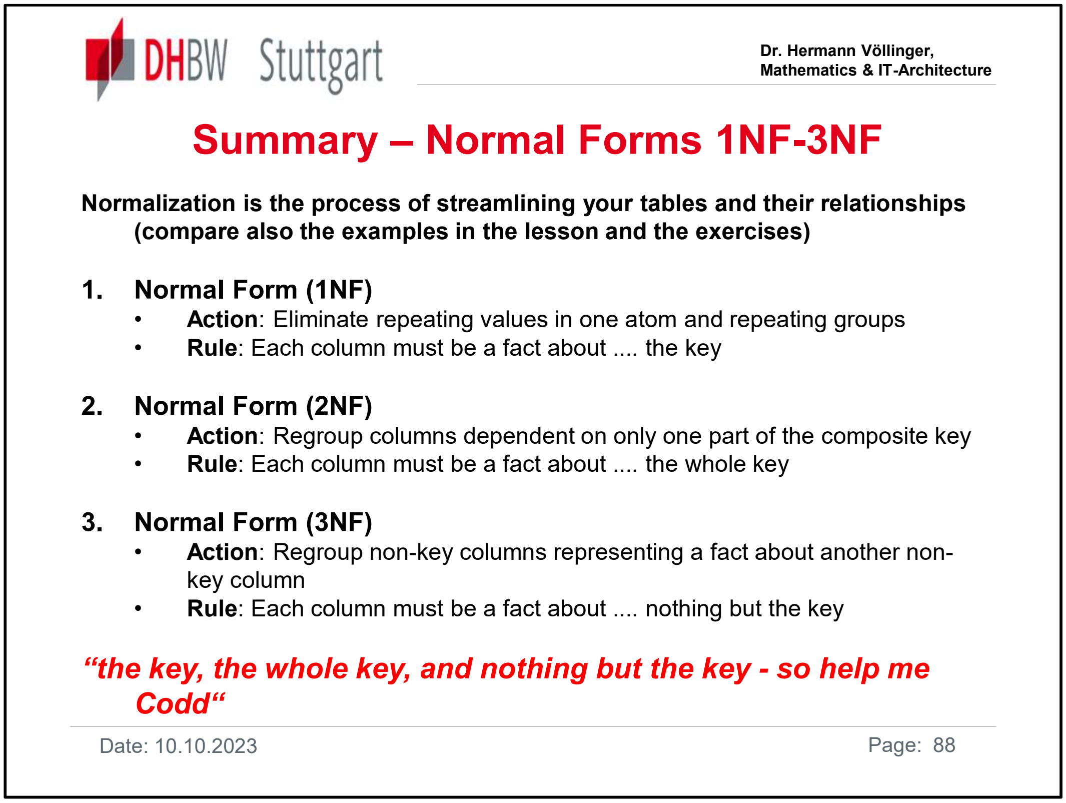
We split the relation, such that the dependent data are in its own tables. The key of the new table is a foreign key in the old table.

CD

CD_ID	Albumtitel	Interpret
4811	Not That Kind	Anastacia
4713	Bad	Michael Jackson
4712	Wish You Were Here	Pink Floyd

Künstler

Interpret	Gründungsjahr
Anastacia	1999
Michael Jackson	1971
Pink Floyd	1965



Summary – Normal Forms 1NF-3NF

Normalization is the process of streamlining your tables and their relationships (compare also the examples in the lesson and the exercises)

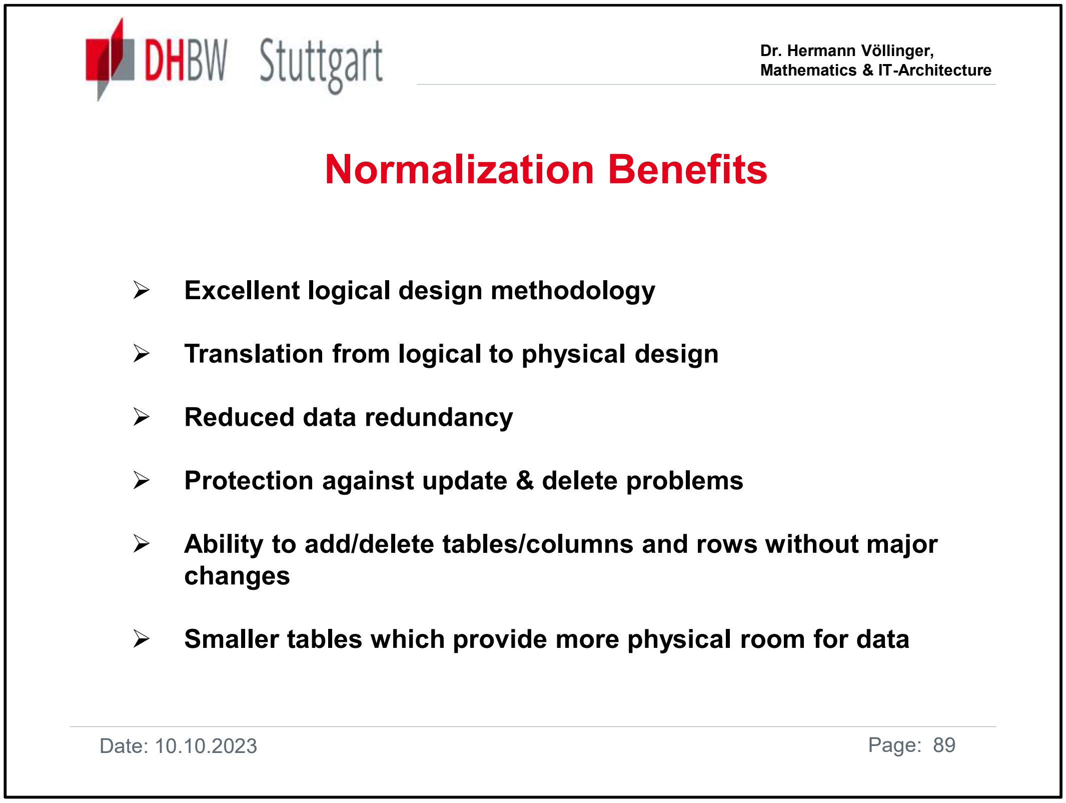
1. **Normal Form (1NF)**
   - **Action**: Eliminate repeating values in one atom and repeating groups
   - **Rule**: Each column must be a fact about .... the key

2. **Normal Form (2NF)**
   - **Action**: Regroup columns dependent on only one part of the composite key
   - **Rule**: Each column must be a fact about .... the whole key

3. **Normal Form (3NF)**
   - **Action**: Regroup non-key columns representing a fact about another non-key column
   - **Rule**: Each column must be a fact about .... nothing but the key

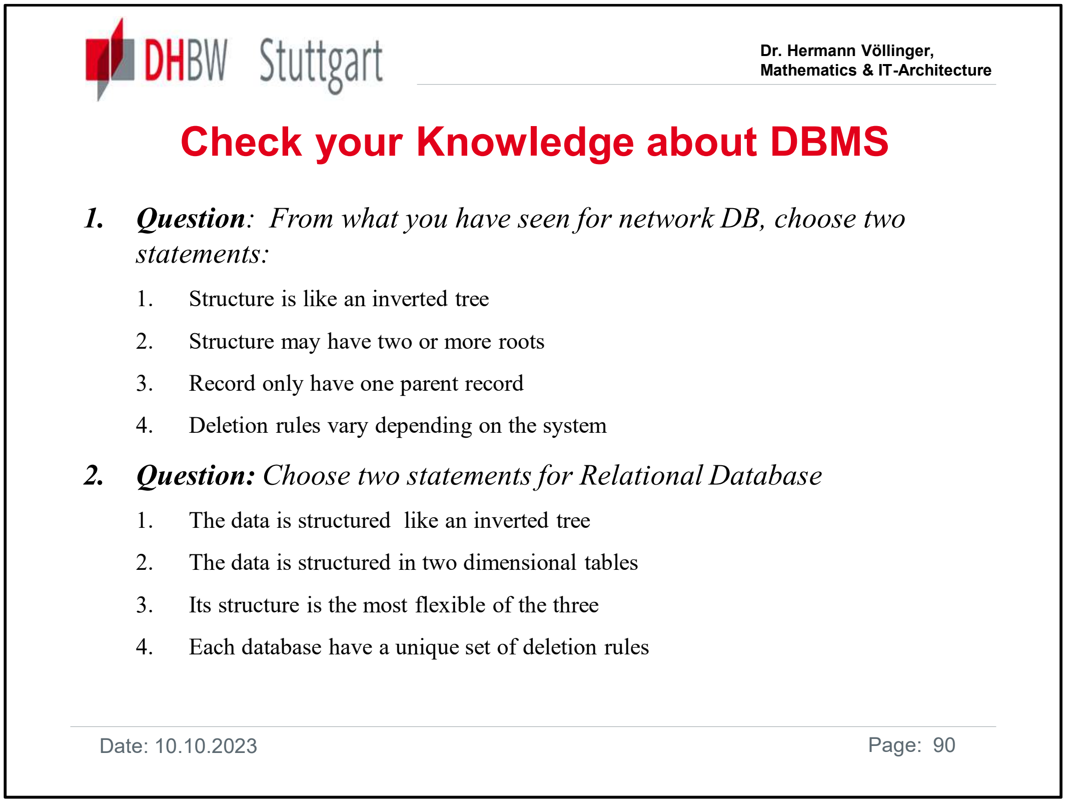
"the key, the whole key, and nothing but the key - so help me Codd"

The database community has developed a series of guidelines for ensuring that databases are normalized. These are referred to as normal forms and are numbered from one (the lowest form of normalization, referred to as first normal form or 1NF) through five (fifth normal form or 5NF). In practical applications, you'll often see 1NF, 2NF, and 3NF along with the occasional 4NF. Fifth normal form is very rarely seen.



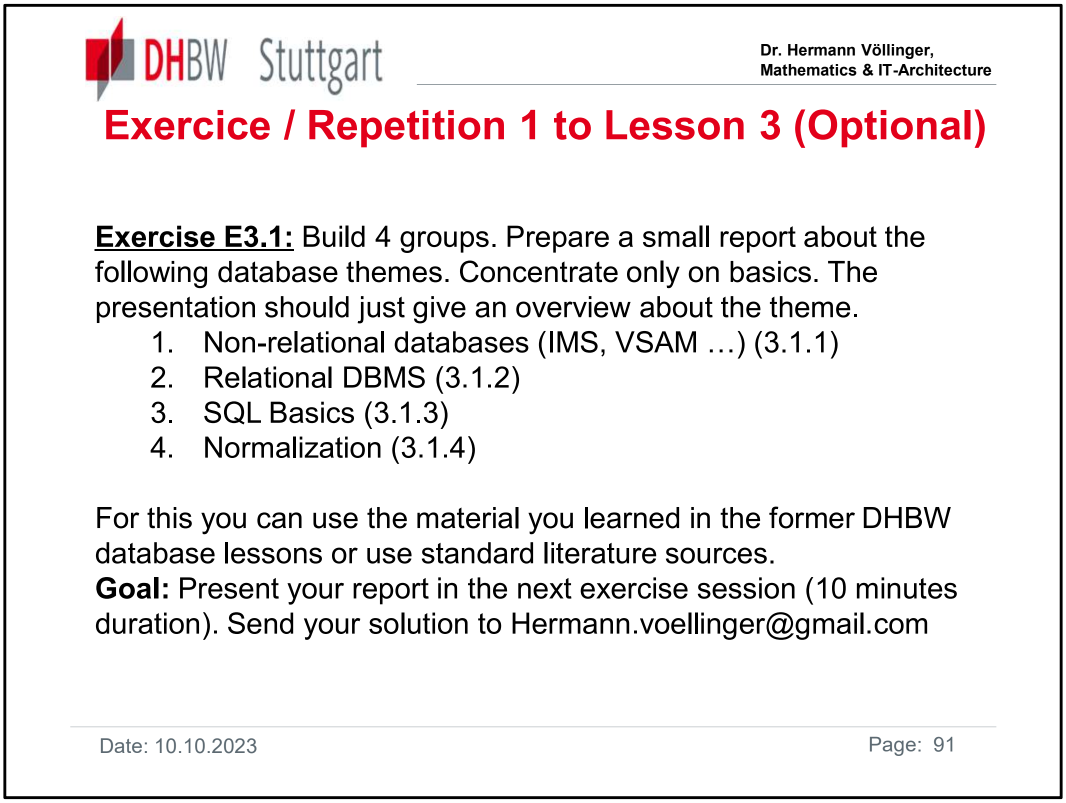
Normalization Benefits

- Excellent logical design methodology
- Translation from logical to physical design
- Reduced data redundancy
- Protection against update & delete problems
- Ability to add/delete tables/columns and rows without major changes
- Smaller tables which provide more physical room for data



Check your Knowledge about DBMS

1. ***Question**: From what you have seen for network DB, choose two statements:*
   1. Structure is like an inverted tree
   2. Structure may have two or more roots
   3. Record only have one parent record
   4. Deletion rules vary depending on the system
2. ***Question:** Choose two statements for Relational Database*
   1. The data is structured like an inverted tree
   2. The data is structured in two dimensional tables
   3. Its structure is the most flexible of the three
   4. Each database have a unique set of deletion rules

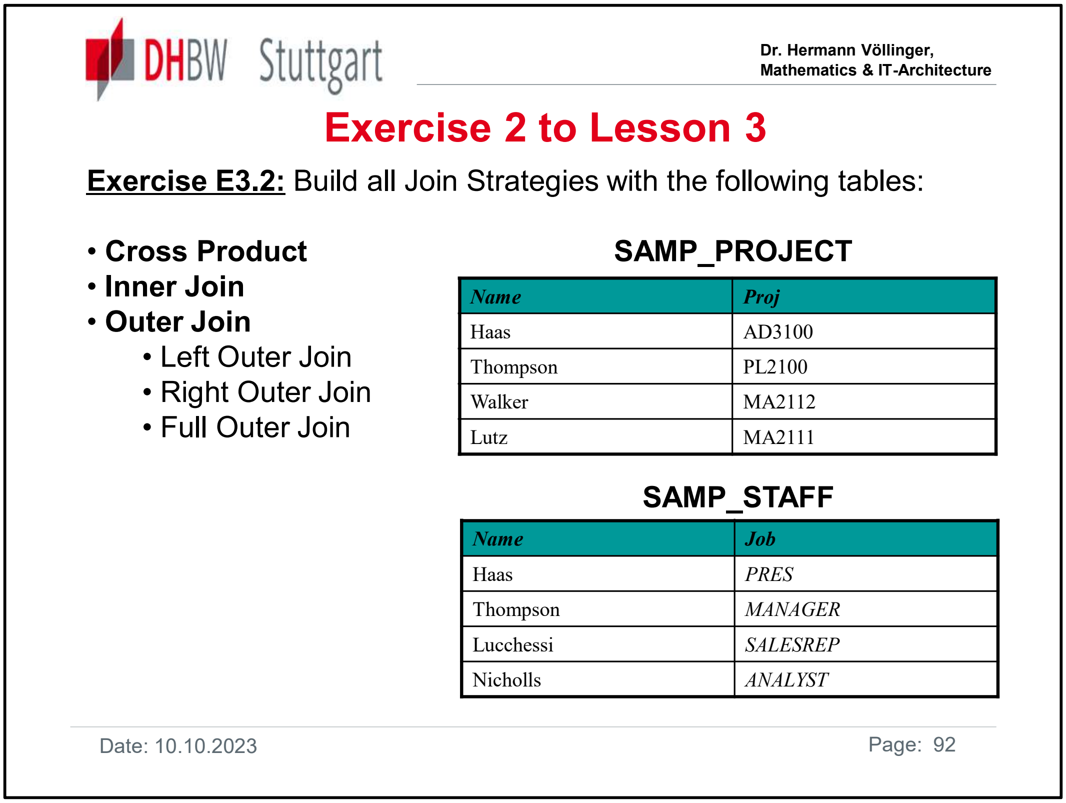


Exercice / Repetition 1 to Lesson 3 (Optional)

<u>Exercise E3.1:</u> Build 4 groups. Prepare a small report about the following database themes. Concentrate only on basics. The presentation should just give an overview about the theme.

1. Non-relational databases (IMS, VSAM …) (3.1.1)
2. Relational DBMS (3.1.2)
3. SQL Basics (3.1.3)
4. Normalization (3.1.4)

For this you can use the material you learned in the former DHBW database lessons or use standard literature sources.
Goal: Present your report in the next exercise session (10 minutes duration). Send your solution to Hermann.voellinger@gmail.com



Exercise 2 to Lesson 3

Exercise E3.2: Build all Join Strategies with the following tables:

- **Cross Product**
- **Inner Join**
- **Outer Join**
  - Left Outer Join
  - Right Outer Join
  - Full Outer Join

SAMP_PROJECT

Name	*Proj*
Haas	AD3100
Thompson	PL2100
Walker	MA2112
Lutz	MA2111

SAMP_STAFF

Name	*Job*
Haas	*PRES*
Thompson	*MANAGER*
Lucchessi	*SALESREP*
Nicholls	*ANALYST*

The following example produces the **cross product** of the two tables. A join condition is not specified, so every combination of rows is present:

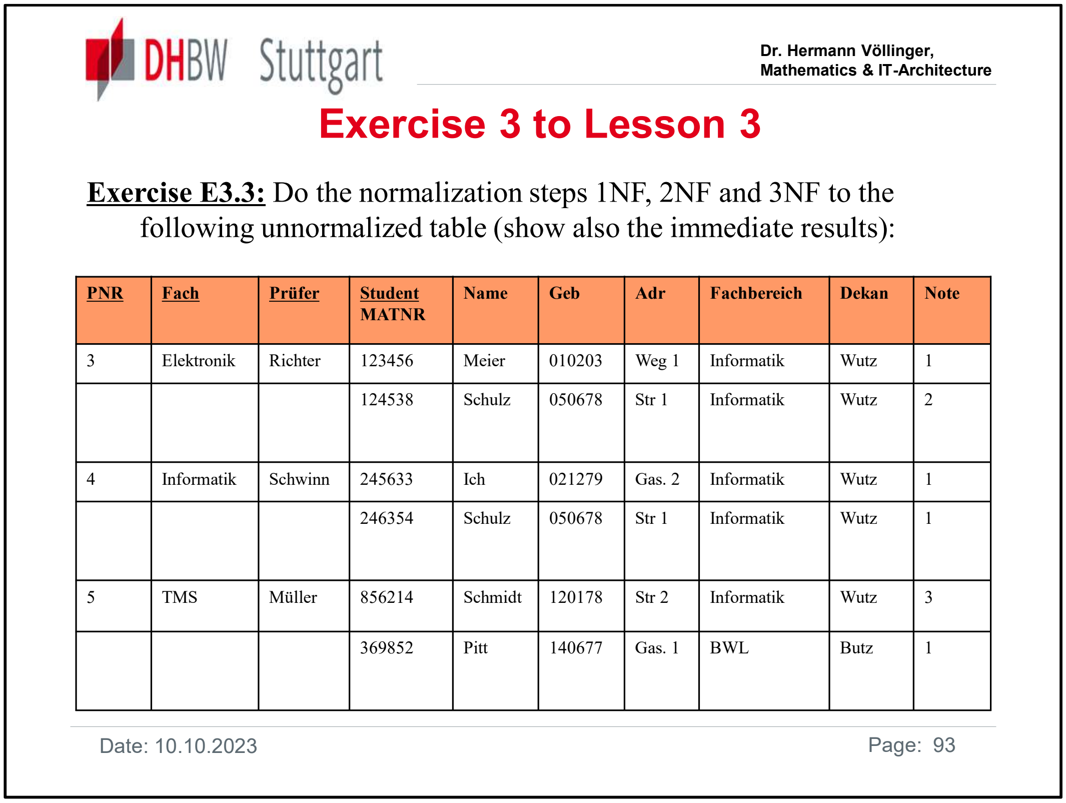
```
  SELECT
SAMP_PROJECT.NAME,  SAMP_PROJECT.PROJ,
SAMP_STAFF.NAME, SAMP_STAFF.JOB
  FROM  SAMP_PROJECT, SAMP_STAFF
```

The following example produces the **inner join** of the two tables. The inner join lists the full-time employees who are assigned to a project:

```
  SELECT
SAMP_PROJECT.NAME,  SAMP_PROJECT.PROJ,
SAMP_STAFF.NAME, SAMP_STAFF.JOB
  FROM  SAMP_PROJECT, SAMP_STAFF
  WHERE SAMP_STAFF.NAME = SAMP_PROJECT.NAME
```

Alternately, you can specify the inner join as follows:

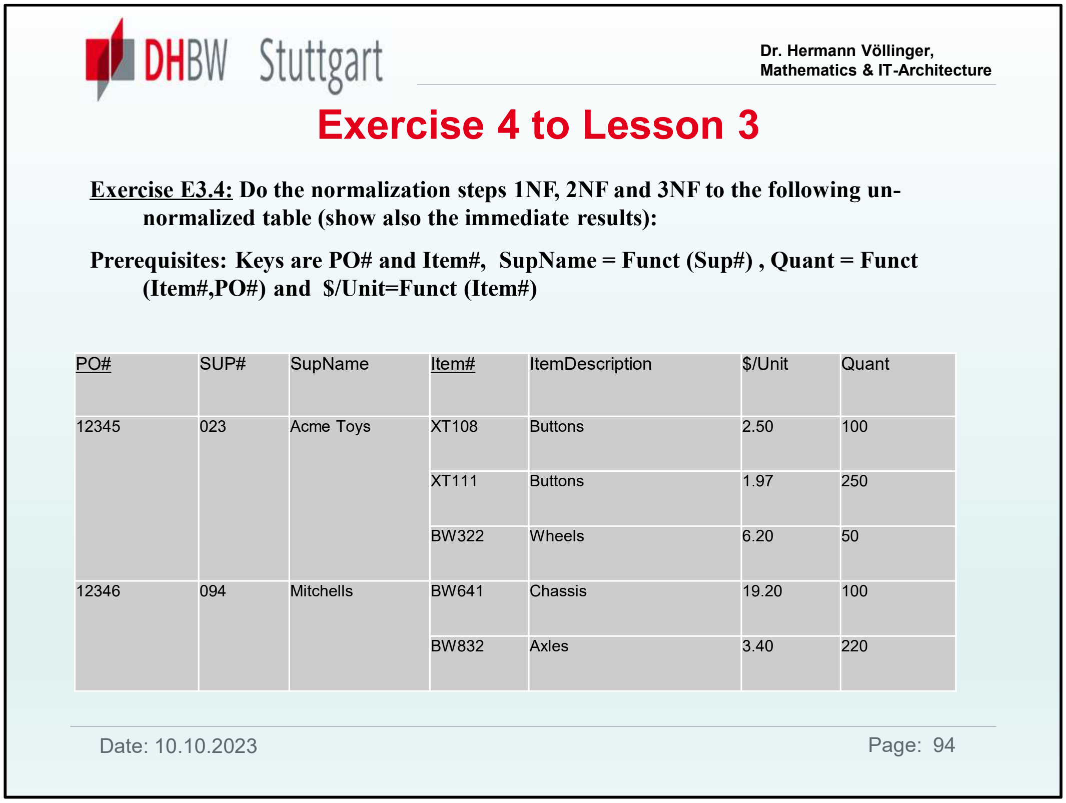
```
  SELECT
SAMP_PROJECT.NAME,  SAMP_PROJECT.PROJ,
 SAMP_STAFF.NAME, SAMP_STAFF.JOB
  FROM
SAMP_PROJECT  INNER JOIN SAMP_STAFF   ON
SAMP_STAFF.NAME = SAMP_PROJECT.NAME
```



Exercise 3 to Lesson 3

Exercise E3.3: Do the normalization steps 1NF, 2NF and 3NF to the following unnormalized table (show also the immediate results):

PNR	Fach	Prüfer	Student MATNR	Name	Geb	Adr	Fachbereich	Dekan	Note
3	Elektronik	Richter	123456	Meier	010203	Weg 1	Informatik	Wutz	1
			124538	Schulz	050678	Str 1	Informatik	Wutz	2
4	Informatik	Schwinn	245633	Ich	021279	Gas. 2	Informatik	Wutz	1
			246354	Schulz	050678	Str 1	Informatik	Wutz	1
5	TMS	Müller	856214	Schmidt	120178	Str 2	Informatik	Wutz	3
			369852	Pitt	140677	Gas. 1	BWL	Butz	1

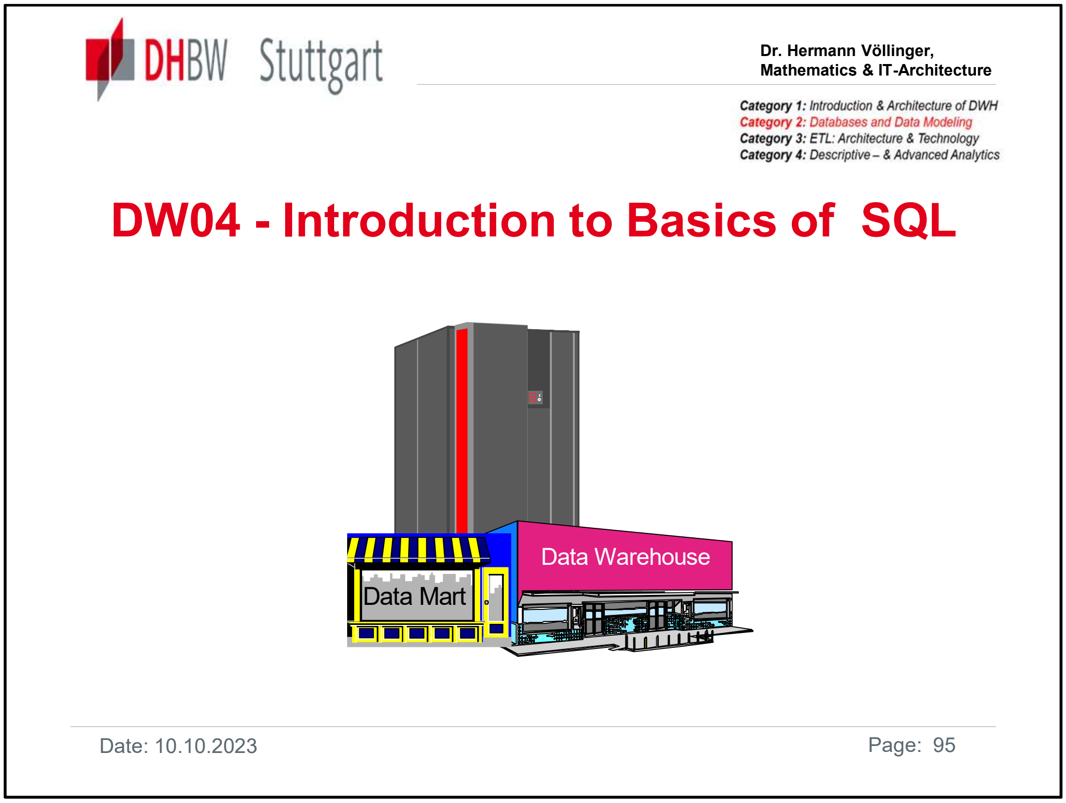


Exercise 4 to Lesson 3

Exercise E3.4: Do the normalization steps 1NF, 2NF and 3NF to the following un-normalized table (show also the immediate results):

Prerequisites: Keys are PO# and Item#, SupName = Funct (Sup#) , Quant = Funct (Item#,PO#) and $/Unit=Funct (Item#)

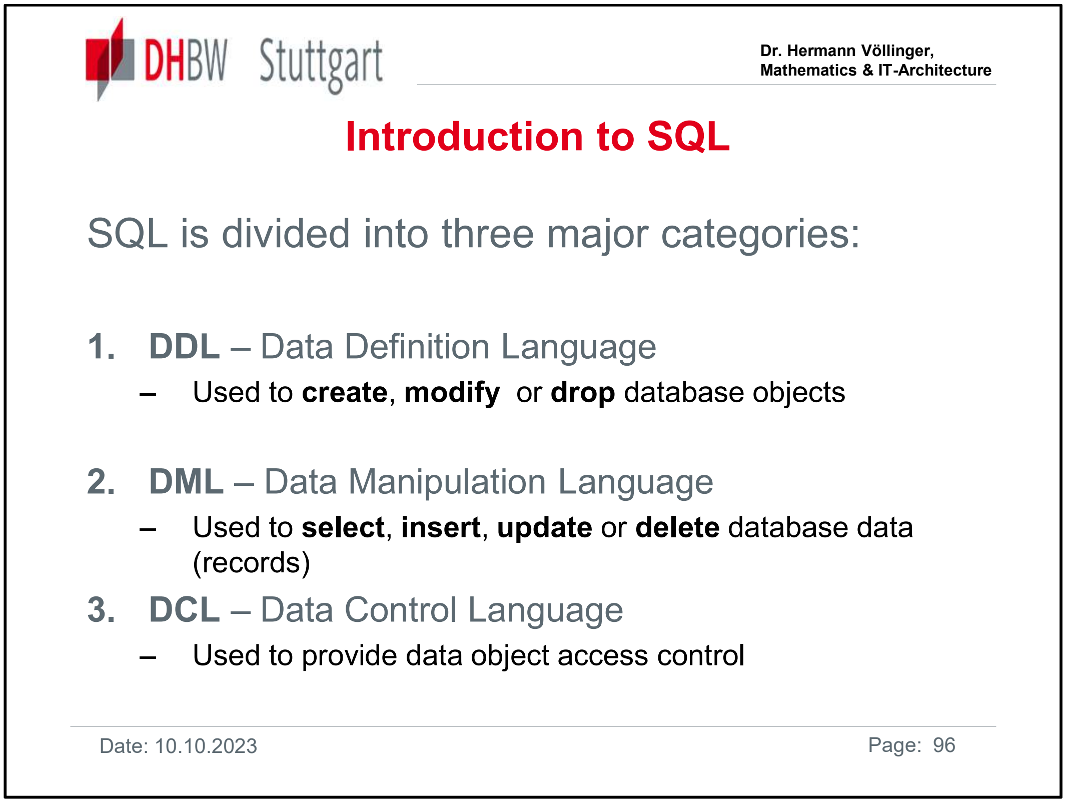
PO#	SUP#	SupName	Item#	ItemDescription	$/Unit	Quant
12345	023	Acme Toys	XT108	Buttons	2.50	100
			XT111	Buttons	1.97	250
			BW322	Wheels	6.20	50
12346	094	Mitchells	BW641	Chassis	19.20	100
			BW832	Axles	3.40	220



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

***Category 1:** Introduction & Architecture of DWH*
***Category 2:** Databases and Data Modeling*
***Category 3:** ETL: Architecture & Technology*
***Category 4:** Descriptive – & Advanced Analytics*

DW04 - Introduction to Basics of SQL



Introduction to SQL

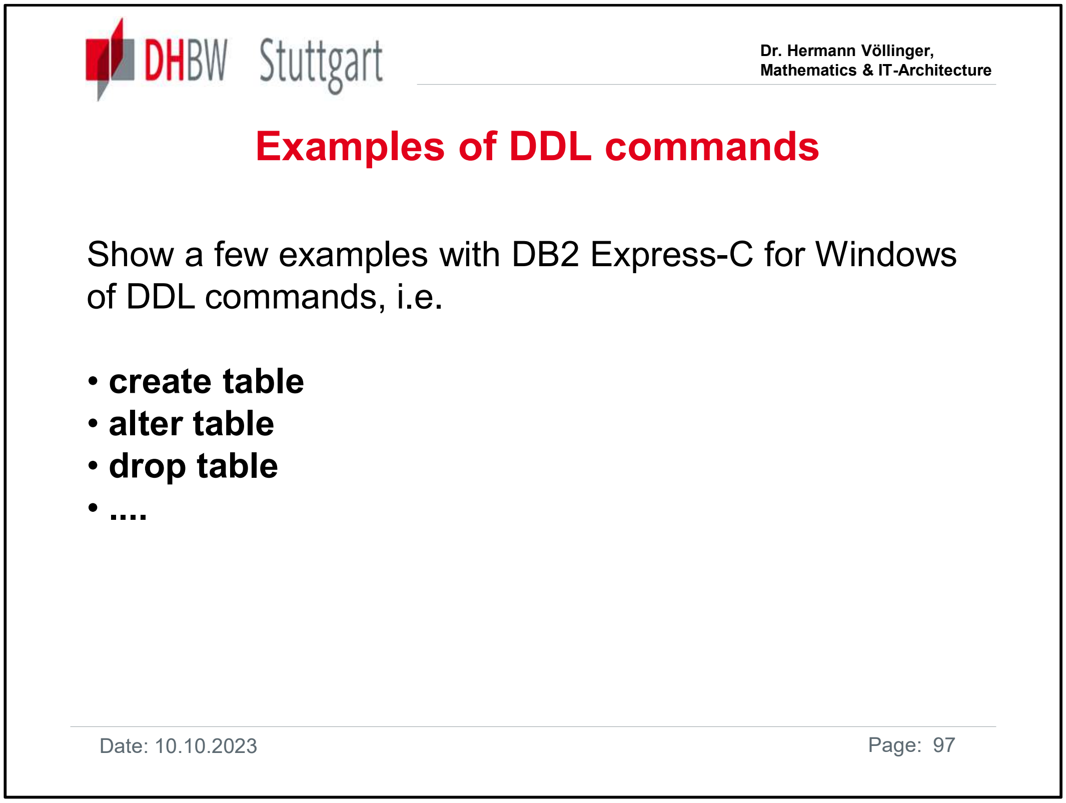
SQL is divided into three major categories:

1. **DDL** – Data Definition Language
   – Used to **create**, **modify** or **drop** database objects

2. **DML** – Data Manipulation Language
   – Used to **select**, **insert**, **update** or **delete** database data (records)

3. **DCL** – Data Control Language
   – Used to provide data object access control

The standard language of relational database access is Structured Query Language (SQL). SQL is not a programming language. It was designed for the single purpose of accessing structured data.



Examples of DDL commands

Show a few examples with DB2 Express-C for Windows of DDL commands, i.e.

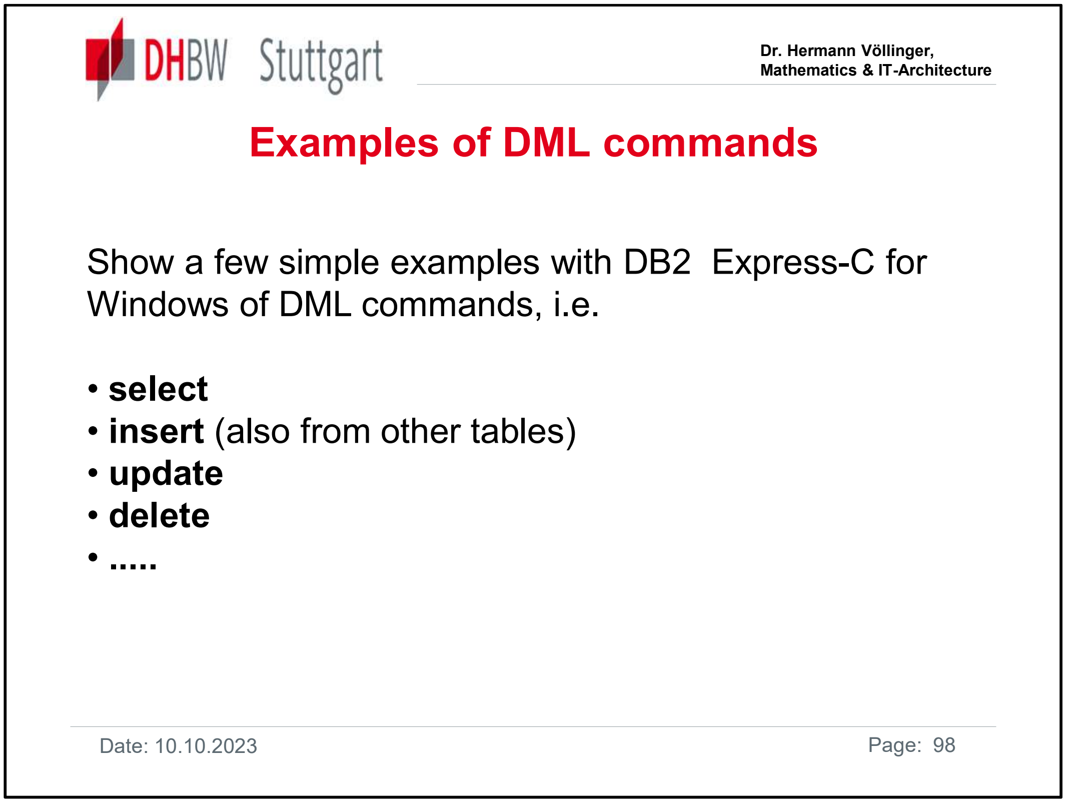
- **create table**
- **alter table**
- **drop table**
- **....**

Creating Tables

Create your own tables using the CREATE TABLE statement, specifying the column names and types, as well as *constraint*s. Constraints are not discussed in in this course.The following statement creates a table named PERS:

```
CREATE TABLE PERS
(ID SMALLINT NOT NULL,NAME VARCHAR(9),
DEPT SMALLINT WITH DEFAULT 10,JOB CHAR(5),
YEARS SMALLINT,SALARY DECIMAL(7,2),
COMM DECIMAL(7,2),BIRTH_DATE DATE)
```

This statement creates a table with no data in it. The next section describes how to insert data into a new table. As shown in the example, you specify both a name and a data type for each column.
NOT NULL is optional and may be specified to indicate that null values are not allowed in a column. Default values are also optional.
There are many other options you can specify in a CREATE TABLE statement, such as unique constraints or referential constraints. For more information about all of the options, see the CREATE TABLE statement in the SQL Reference.© Copyright IBM Corp. 1993, 2000
https://bportaluri.com/wp-content/downloads/docs/db2y0e70.pdf



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Examples of DML commands

Show a few simple examples with DB2 Express-C for Windows of DML commands, i.e.

- **select**
- **insert** (also from other tables)
- **update**
- **delete**
- **.....**

Date: 10.10.2023 Page: 98

Inserting Data: When you create a new table, it does not contain any data. To enter new rows into a table, you use the INSERT statement. This statement has two general forms:
- With one form, you use a VALUES clause to specify values for the columns of one or more rows. The next three examples insert data into tables using this general form.
- With the other form, rather than specifying VALUES, you specify a fullselect to identify columns from rows contained in other tables and/or views.

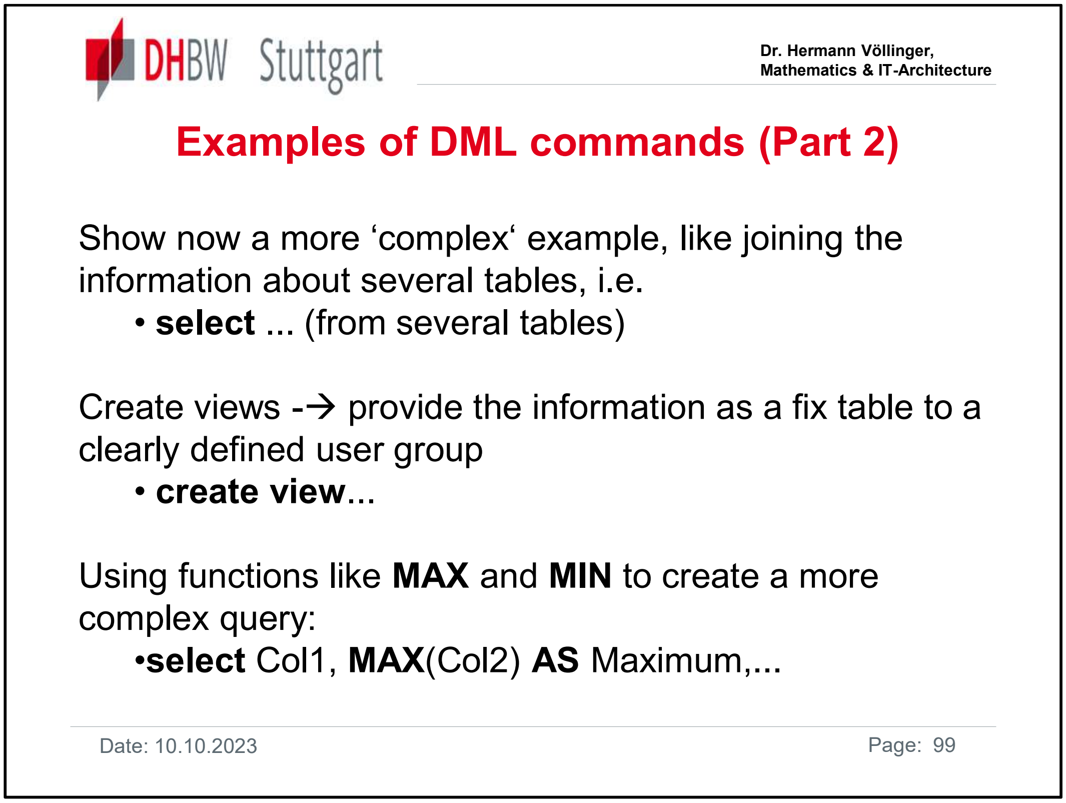
Fullselect is a select statement used in INSERT or CREATE VIEW statements or following a predicate. A fullselect that is enclosed in parentheses is commonly referred as a sub-query.

Depending on the default options that you have chosen when creating your table, for every row you insert, you either supply a value for each column or accept a default value. The default values for the various data types are discussed in the SQL Reference.

The following statement uses a VALUES clause to insert one row of data into the PERS table:
INSERT INTO PERS **VALUES** (12,'Harris',20,'Sales',5,18000,1000,'1950-1-1’)

The following statement uses the VALUES clause to insert three rows into the PERS table where only the IDs, the names, and the jobs are known.
INSERT INTO PERS (NAME,JOB,ID) **VALUES** **(**'Swagerman','Prgmr',500), ('Limoges','Prgmr',510), ('Li','Prgmr',520)

For inserting data from other tables, just execute the following command:
INSERT INTO PERS (ID,NAME,DEPT,JOB,YEARS,SALARY) **SELECT** ID,NAME,DEPT,JOB,YEARS,SALARY **FROM** STAFF WHERE DEPT =38



Examples of DML commands (Part 2)

Show now a more ‘complex‘ example, like joining the information about several tables, i.e.

- **select** ... (from several tables)

Create views -→ provide the information as a fix table to a clearly defined user group

- **create view...**

Using functions like **MAX** and **MIN** to create a more complex query:

- **select** Col1, **MAX**(Col2) **AS** Maximum,...

• To do a **join about two tables,** for example the table **org** and **staff** in the DB2 **SAMPLE** database. You have to take care that the where condition makes sense:

First example:

select deptname, name **from** org, staff
where manager=id

Second example:

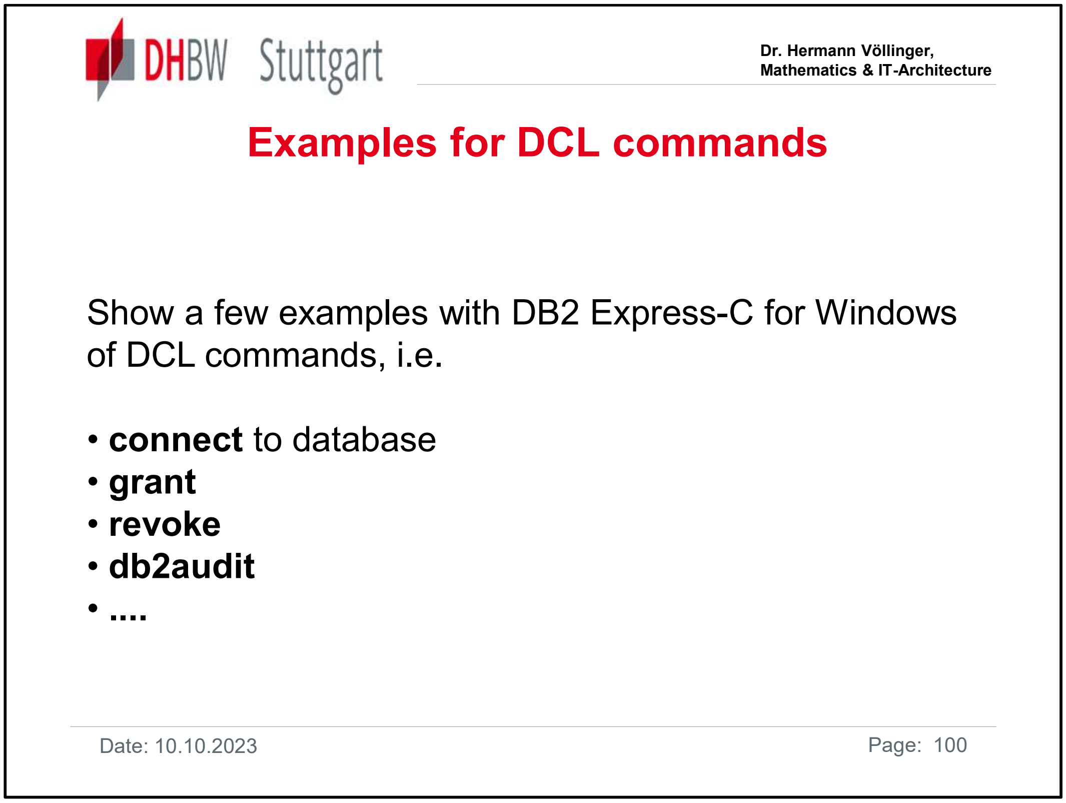
select e.empno, e.firstnme, e.lastname, d.deptno, d.deptname
from employee e, department d
where d.deptno = e.workdept **order by** e.empno

• To **create a view,** compare the following example**:**

create view staff_only
as select id, name, dept, job,years
from staff **where** job<>'Mgr' and dept=20

• To usage of **column functions** like **MAX, MIN** and **COUNT** can be seen in the following example, where we see the high and low salary per department**:**

```
create view High_Low_Department_Salary
as select workdept, MAX(salary) as maximum, MIN(salary) as minimum
from employee
group by workdept
having COUNT(*)>4
```



Examples for DCL commands

Show a few examples with DB2 Express-C for Windows of DCL commands, i.e.

- **connect** to database
- **grant**
- **revoke**
- **db2audit**
- **....**

Connecting to a Database:

You need to connect to a database before you can use SQL statements to query or manipulate it. The CONNECT statement associates a database connection with an user-name.

For example, to connect to the SAMPLE database, type the following command in the DB2 command line processor: **CONNECT TO SAMPLE USER** USERID **USING** PASSWORD

(Be sure to choose a user ID and password that are valid on the server system.)

In this example, USER is USERID and USING is PASSWORD. The following message tells you that you have made a successful connection:

Database Connection Information
Database product =DB2/NT 7.1.0
SQL authorization ID =USERID
Local database alias =SAMPLE

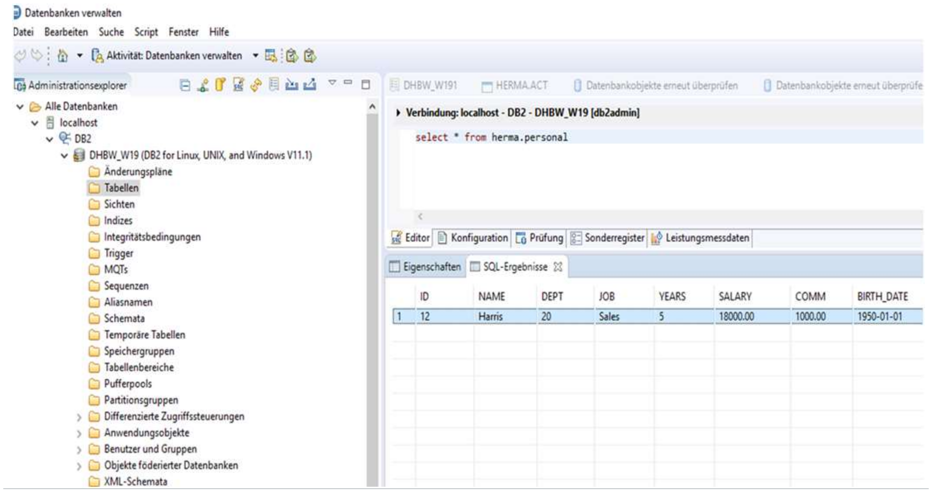
Once you are connected, you can start manipulating the database. For further details on connections, refer to the CONNECT statement in the *SQL Reference*.

Investigating Errors:

Whenever you make a mistake typing in any of the examples, or if an error occurs during execution of an SQL statement, the database manager returns an error message. The error message consists of a message identifier, a brief explanation, and an SQLSTATE. SQLSTATE errors are error codes common to the DB2 family of products. SQLSTATE errors conform to the ISO/ANSI SQL92 standard. For example, if the user ID or password had been incorrect in the CONNECT statement, the database manager would have returned a message identifier of SQL1403N and an SQLSTATE of 08004. The message is as follows:SQL1403N The username and/or password supplied is incorrect.

Demo with IBM Data Studio

Show examples of DDL-, DML- & DCL- commands with IBM Data Studio tools of data in DB2 Express-C Version 11.1. database.



Verbindung: localhost - DB2 - DHBW_W19 [db2admin]

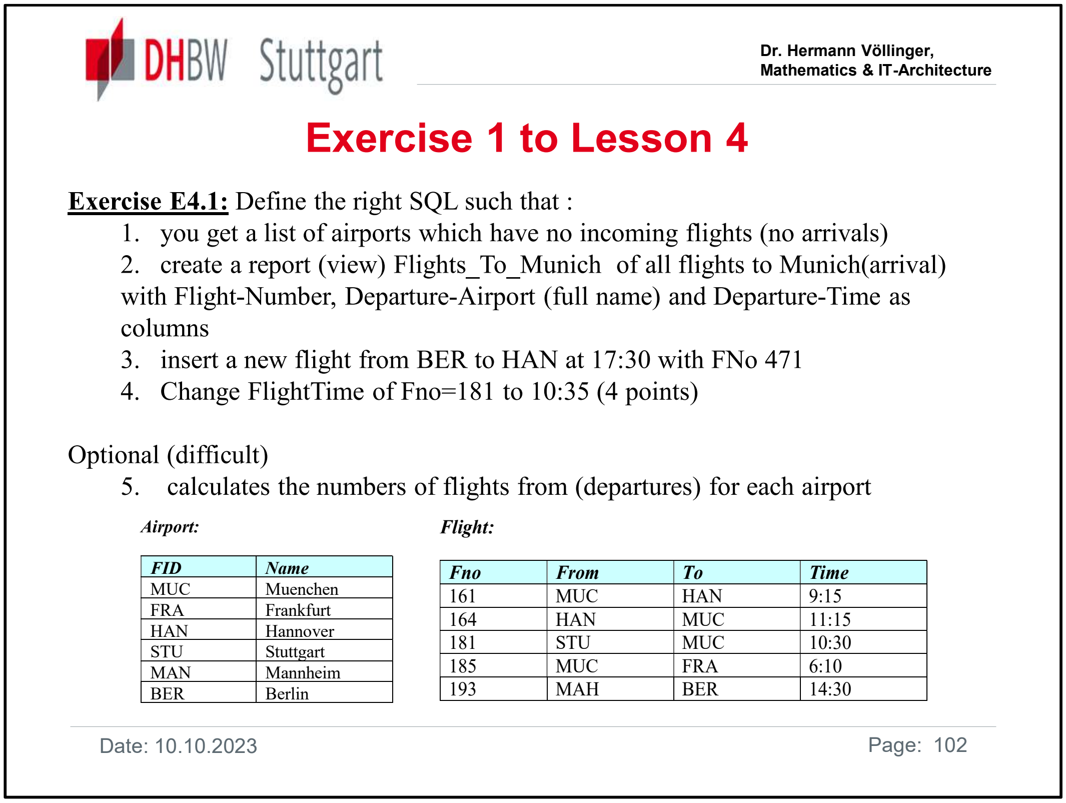
```
select * from herma.personal
```

	ID	NAME	DEPT	JOB	YEARS	SALARY	COMM	BIRTH_DATE
1	12	Harris	20	Sales	5	18000.00	1000.00	1950-01-01

IBM® Data Studio provides application developers with a single integrated development environment that can be used to create, deploy, and debug data-centric applications. Built to extend the Eclipse framework and SQL model components, it combines Eclipse technology and shared repository extensions for database development.

- IBM Data Studio consist of the following components: The **IBM Data Studio client**, which is an Eclipse-based tool that provides an integrated development environment for database and instance administration, routine and Java application development, and query tuning tasks. It can be installed with other IBM software products to share a common environment.
- The **IBM Data Studio web console,** which is a web-based tool with health and availability monitoring, job creation, and database administration tasks.

See also in [DHBW-Moodle] the video: “Demo-SQL_Examples-IBM_Data_Studio.mp4”



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise 1 to Lesson 4

Exercise E4.1: Define the right SQL such that :

1. you get a list of airports which have no incoming flights (no arrivals)
2. create a report (view) Flights_To_Munich of all flights to Munich(arrival) with Flight-Number, Departure-Airport (full name) and Departure-Time as columns
3. insert a new flight from BER to HAN at 17:30 with FNo 471
4. Change FlightTime of Fno=181 to 10:35 (4 points)

Optional (difficult)

5. calculates the numbers of flights from (departures) for each airport

Airport:

FID	*Name*
MUC	Muenchen
FRA	Frankfurt
HAN	Hannover
STU	Stuttgart
MAN	Mannheim
BER	Berlin

Flight:

Fno	*From*	*To*	*Time*
161	MUC	HAN	9:15
164	HAN	MUC	11:15
181	STU	MUC	10:30
185	MUC	FRA	6:10
193	MAH	BER	14:30

Date: 10.10.2023 Page: 102

Ad5 (optional):

```
select name as Departure_Airport, count (*) as Departure_Count
from airport, flight
where fid=from
group by name  union
select name as Departure_Airport, 0 as Departure_Count
from airport
where not exists (select * from flight where from=fid)
order by departure_count
```

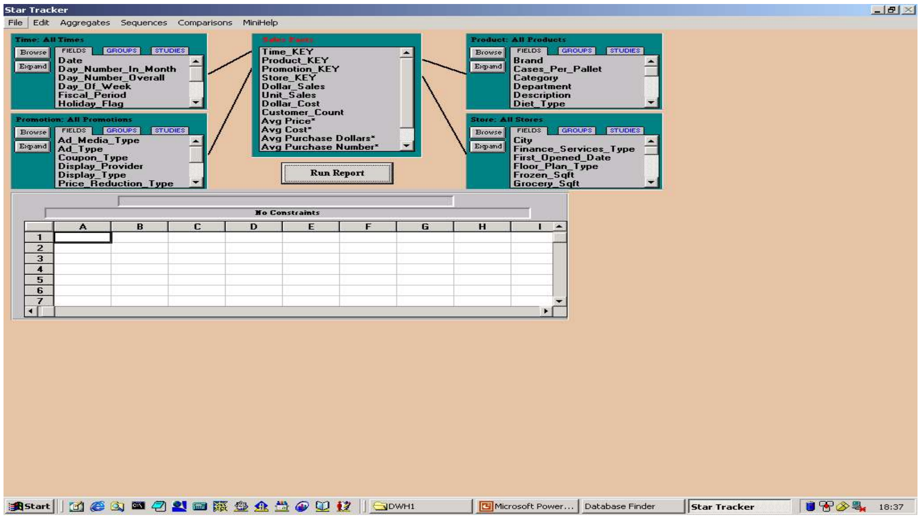
Delivers the following result:

```
DEPATURE_AIRPORT    DEPARTURE_COUNT
-----------------------------   ------------------------------
Berlin                          0
Frankfurt                       0
Hannover                        1
Mannheim                        1
Stuttgart                       1
Muenchen                        2

6 record(s) selected.
```

Exercise 2 to Lesson 4 (First part)

Compare the data model from R. Kimball's Grocery example:



Time: All Times
Date
Day_Number_In_Month
Day_Number_Overall
Day_Of_Week
Fiscal_Period
Holiday_Flag

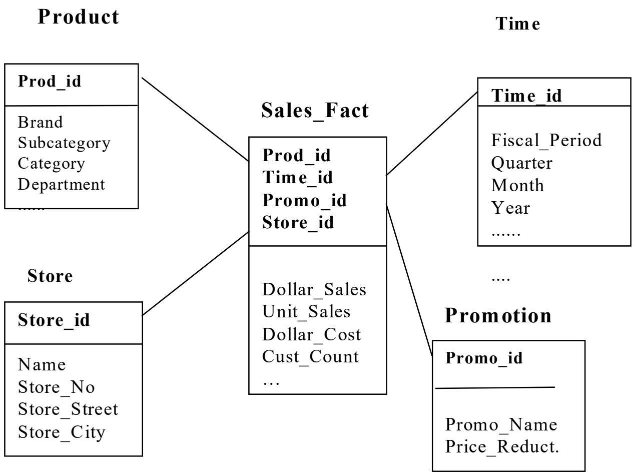
Promotion: All Promotions
Ad_Media_Type
Ad_Type
Coupon_Type
Display_Provider
Display_Type
Price_Reduction_Type

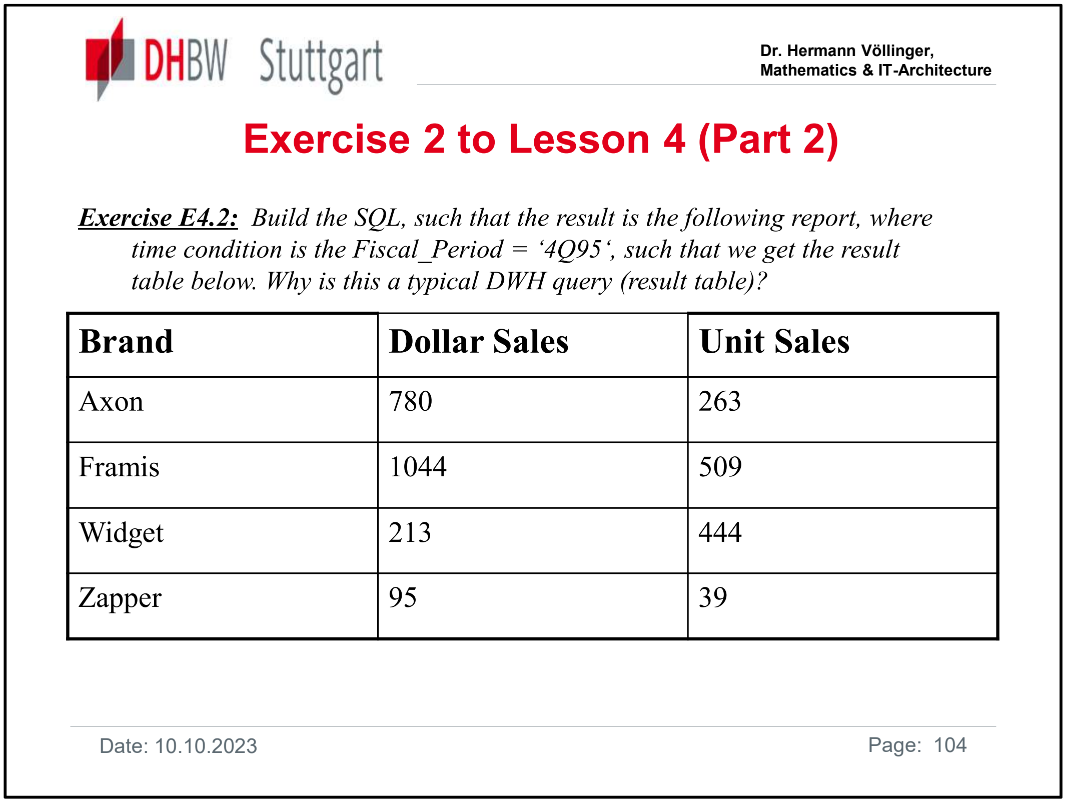
Sales Facts
Time_KEY
Product_KEY
Promotion_KEY
Store_KEY
Dollar_Sales
Unit_Sales
Dollar_Cost
Customer_Count
Avg Price*
Avg Cost*
Avg Purchase Dollars*
Avg Purchase Number*

Product: All Products
Brand
Cases_Per_Pallet
Category
Department
Description
Diet_Type

Store: All Stores
City
Finance_Services_Type
First_Opened_Date
Floor_Plan_Type
Frozen_Sqft
Grocery_Sqft

Consider the following Star Schema:





Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise 2 to Lesson 4 (Part 2)

Exercise E4.2: *Build the SQL, such that the result is the following report, where time condition is the Fiscal_Period = ‘4Q95‘, such that we get the result table below. Why is this a typical DWH query (result table)?*

Brand	Dollar Sales	Unit Sales
Axon	780	263
Framis	1044	509
Widget	213	444
Zapper	95	39

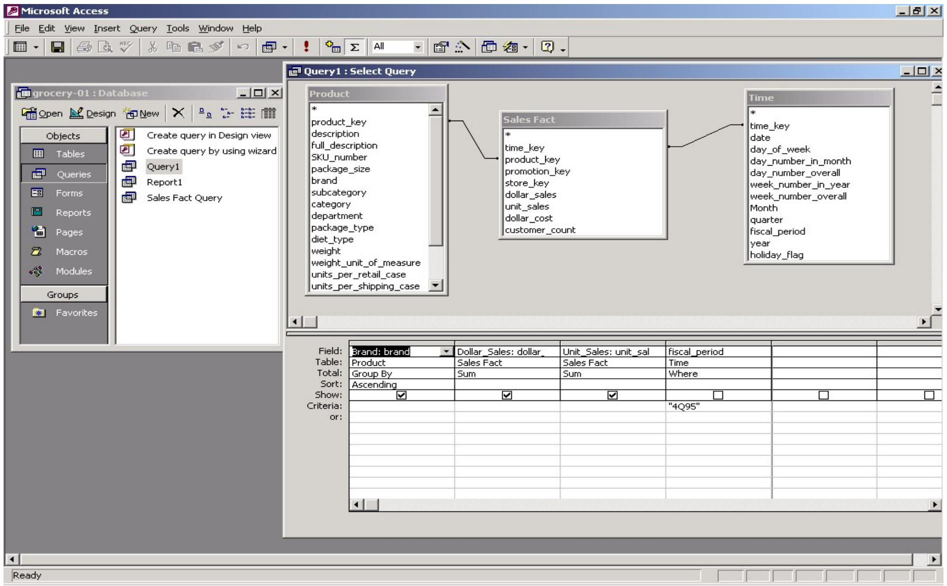
Solution with Standard SQL(for example with DB2):

```
SELECT p.brand AS Brand, Sum(s.dollar_sales) AS Dollar_Sales, Sum(s.unit_sales) AS
Unit_Sales
FROM  sales_fact s, product p, time t
WHERE p.product_key = s.product_key
      AND s.time_key = t.time_key
      AND t.fiscal_period="4Q95"
GROUP BY  p.brand
ORDER BY  p.brand
```

By using the **SQL Wizard** (Design View) in the database **Microsoft Access,** we see the following ‘**Access SQL**‘:

```
SELECT Product.brand AS Brand, Sum([Sales Fact].dollar_sales) AS Dollar_Sales,
Sum([Sales Fact].unit_sales) AS Unit_Sales
FROM ([Sales Fact] INNER JOIN [Time] ON [Sales Fact].time_key = Time.time_key) INNER
JOIN Product ON [Sales Fact].product_key = Product.product_key
WHERE (((Time.fiscal_period)="4Q95"))
GROUP BY Product.brand
ORDER BY Product.brand;
```

Solution with MS Access SQL Wizard

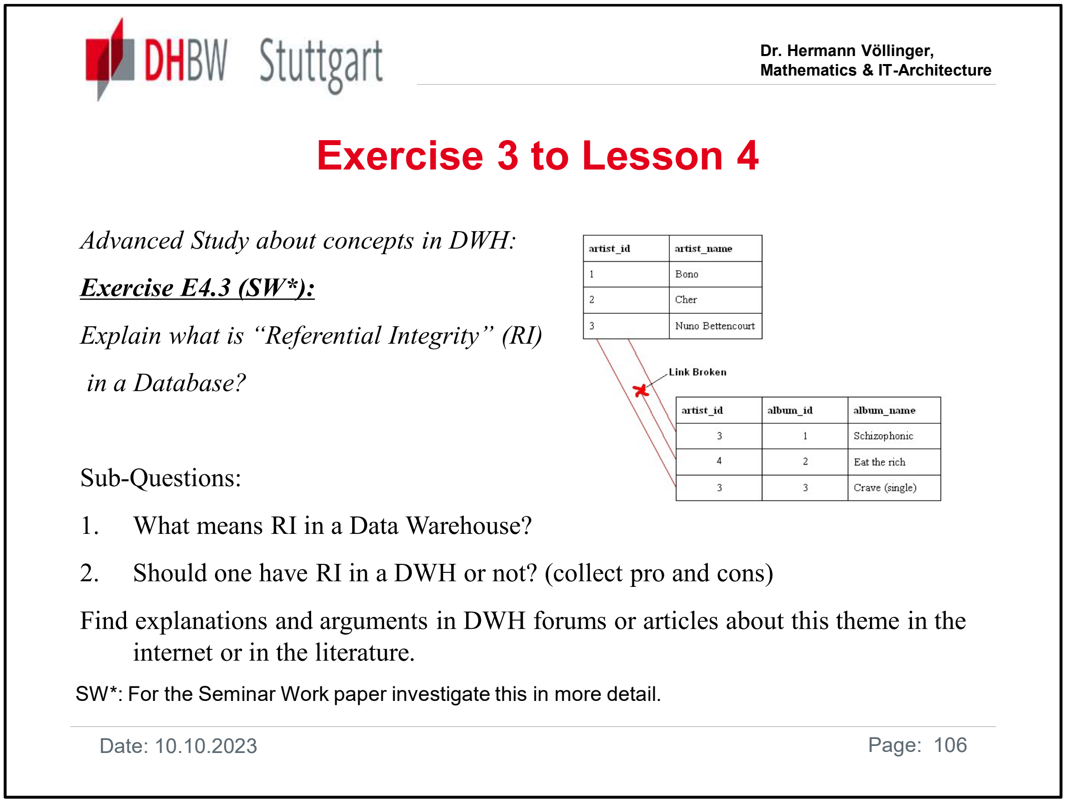


Field:	Brand: brand	Dollar_Sales: dollar_	Unit_Sales: unit_sal	fiscal_period		
Table:	Product	Sales Fact	Sales Fact	Time		
Total:	Group By	Sum	Sum	Where		
Sort:	Ascending					
Show:	☑	☑	☑	□	□	□
Criteria:				"4Q95"		
or:						

Running the above SQL results in the following table:

Brand	Dollar_Sales	Unit_Sales
American Corn	39872,23	41544
Big Can	36375,16	39643
Chewy Industries	33765,57	43612
Cold Gourmet	64938,83	26145
Frozen Bird	70598,67	28611
National Bottle	23791	26099
Squeezable Inc	65020,68	41949
Western Vegetable	50685,69	27998

See also in [DHBW-Moodle] the video: “Demo-Exercise_ E4.2-MS_Access.mp4”



Exercise 3 to Lesson 4

Advanced Study about concepts in DWH:

Exercise E4.3 (SW*):

Explain what is "Referential Integrity" (RI) in a Database?

artist_id	artist_name
1	Bono
2	Cher
3	Nuno Bettencourt

Link Broken

artist_id	album_id	album_name
3	1	Schizophonic
4	2	Eat the rich
3	3	Crave (single)

Sub-Questions:

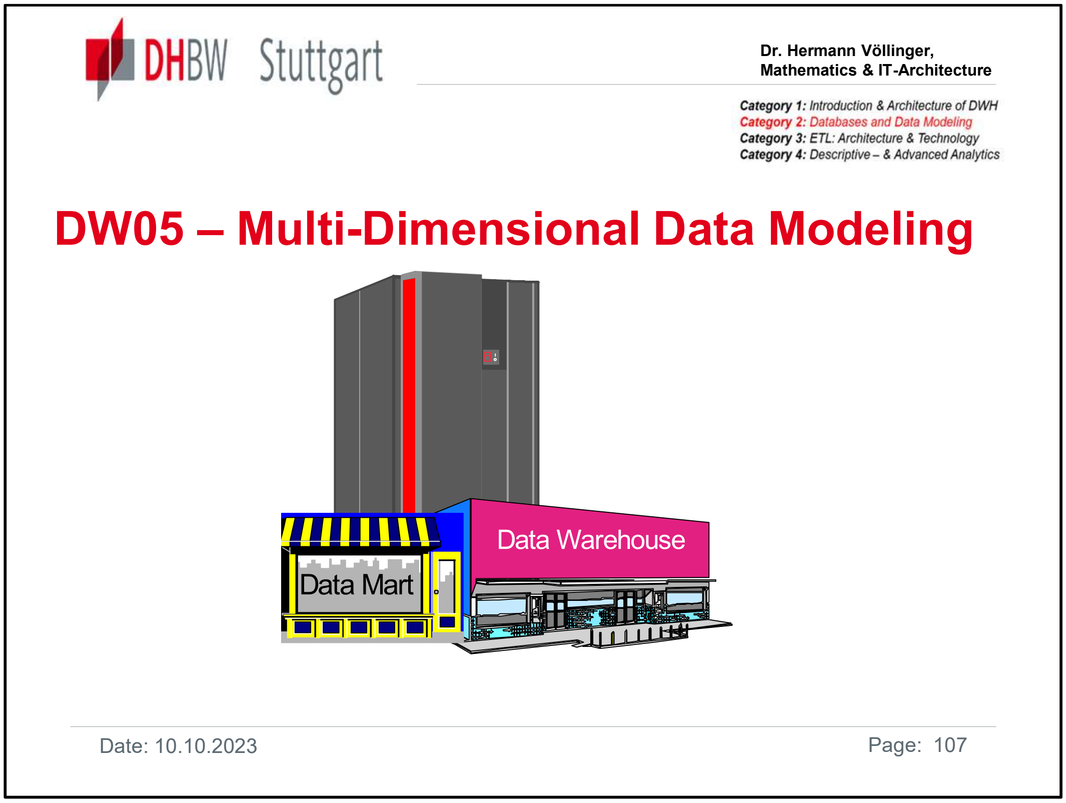
1. What means RI in a Data Warehouse?
2. Should one have RI in a DWH or not? (collect pro and cons)

Find explanations and arguments in DWH forums or articles about this theme in the internet or in the literature.

SW*: For the Seminar Work paper investigate this in more detail.

For more information see the following links:

- https://en.wikipedia.org/wiki/Referential_integrity
- http://stackoverflow.com/questions/2188352/when-is-referential-integrity-not-appropriate
- https://datawarehouseinfo.com/implementing-referential-integrity-in-a-data-warehouse-a-controversial-decision-with-a-lasting-impact/
- https://www.theta.co.nz/news-blogs/tech-blog/referential-integrity-and-its-role-in-data-warehousing-part-one/
- https://www.theta.co.nz/news-blogs/tech-blog/referential-integrity-and-its-role-in-data-warehousing-part-two/



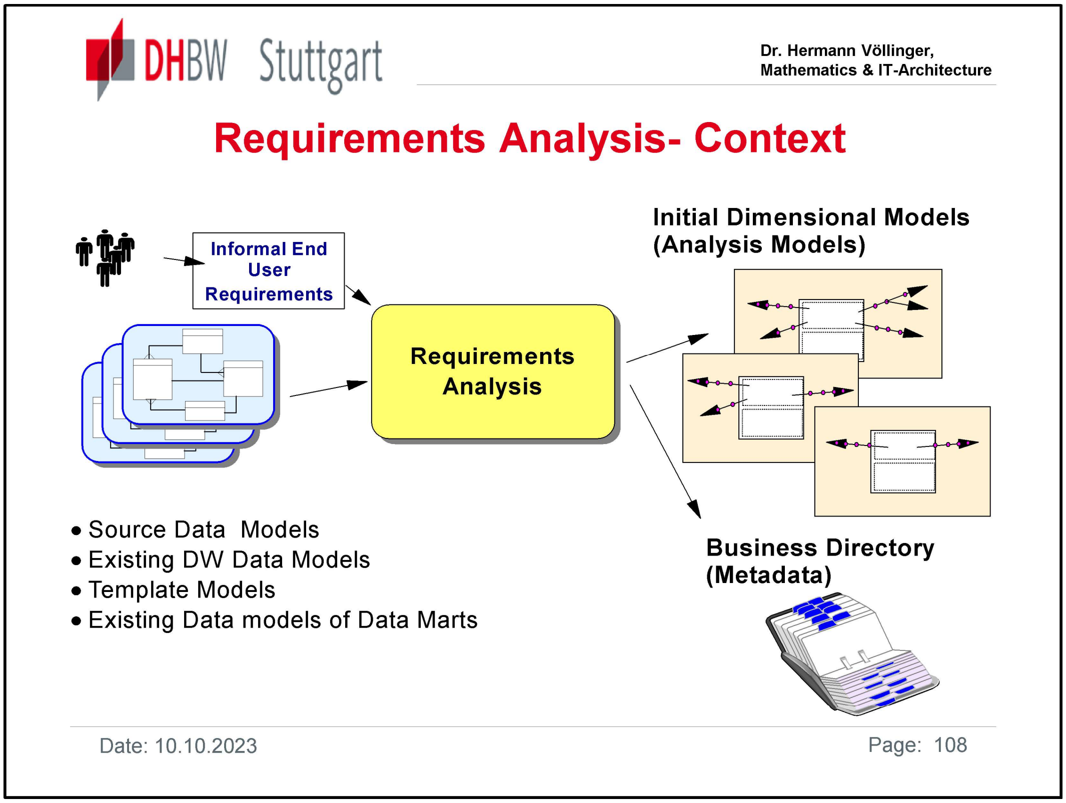
DW05 – Multi-Dimensional Data Modeling

The following chapter gives an introduction into the methods and concepts of Multi Dimensional Data Modeling (MDDM) in the process of building a data warehouse. It explains the architectural ideas behind a Multi Dimensional Model.

It is build to support business needs in arranging data in such a structure that business users can easy ask questions about their business success. This success is measured in "measures and dimensions". These concepts together with 4 other basic concepts are shown and in the "Six Basic Concepts" of MDDM.

So to build an initial multi-dimensional data model, the following six base elements have to be identified:

1. **Measures**
2. **Dimensions**
3. **Grains of dimensions and granularities of measures and facts**
4. **Facts**
5. **Dimension hierarchies**
6. **Aggregation levels**



Requirements Analysis- Context

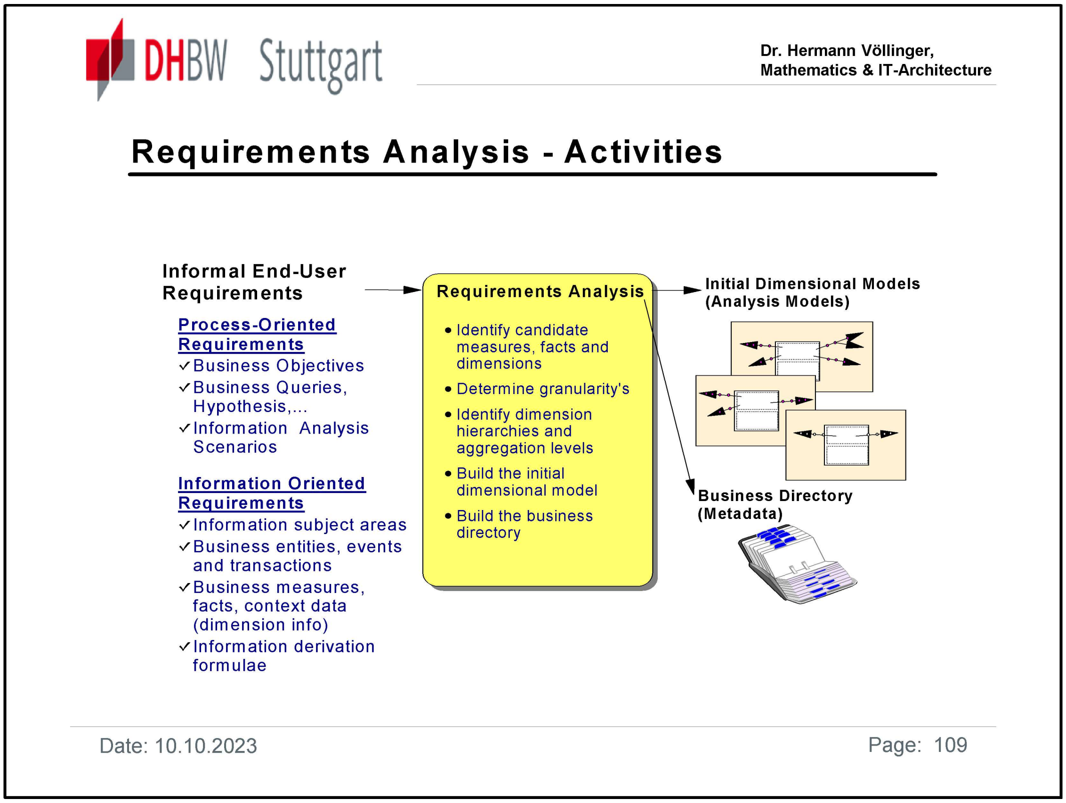
Informal End User Requirements
Requirements Analysis
Initial Dimensional Models (Analysis Models)
Business Directory (Metadata)

- Source Data Models
- Existing DW Data Models
- Template Models
- Existing Data models of Data Marts

Deliverables (‘the output of the process’) consist of a combination of:

• So called **initial dimensional data models**, symbolically represented by the box-and-arrow diagrams on the slide. We will show later how to best construct such initial dimensional models

• The **business directory or metadata definitions** of all element of the MDDM



Requirements Analysis - Activities

Informal End-User Requirements

Process-Oriented Requirements

- ✓Business Objectives
- ✓Business Queries, Hypothesis,...
- ✓Information Analysis Scenarios

Information Oriented Requirements

- ✓Information subject areas
- ✓Business entities, events and transactions
- ✓Business measures, facts, context data (dimension info)
- ✓Information derivation formulae

Requirements Analysis

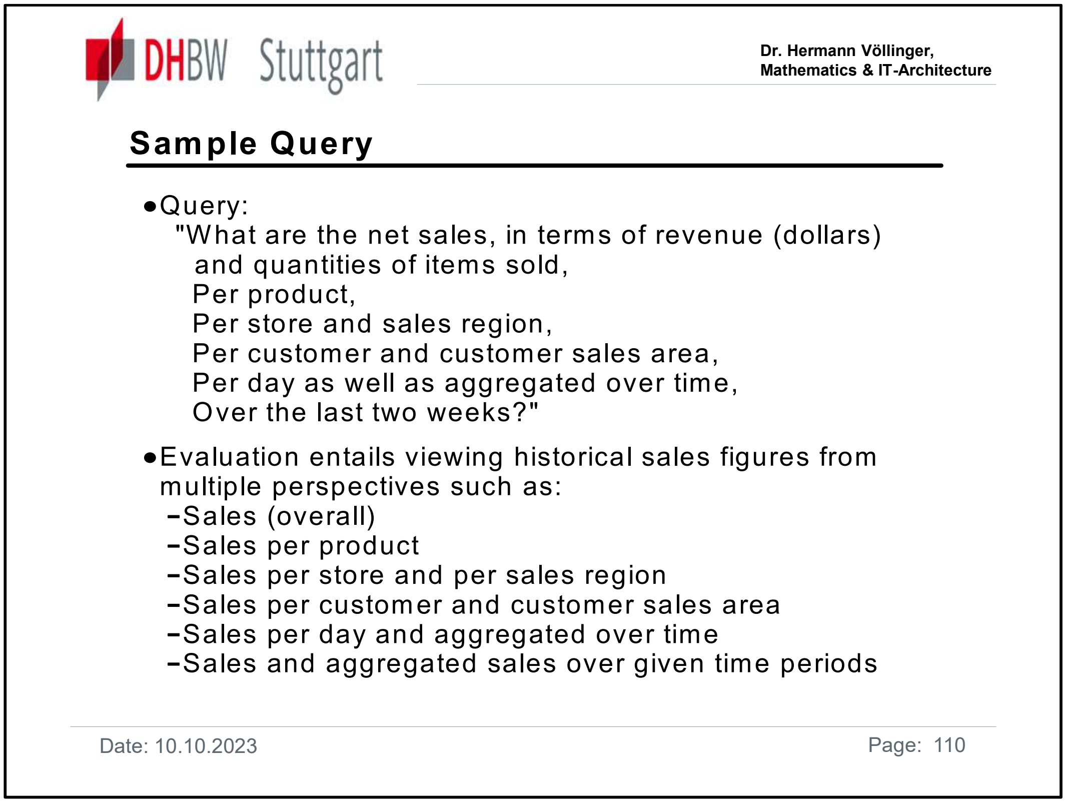
- Identify candidate measures, facts and dimensions
- Determine granularity's
- Identify dimension hierarchies and aggregation levels
- Build the initial dimensional model
- Build the business directory

Initial Dimensional Models (Analysis Models)

Business Directory (Metadata)

End user requirements suitable for a data warehouse modeling project can be classified in two mayor categories:

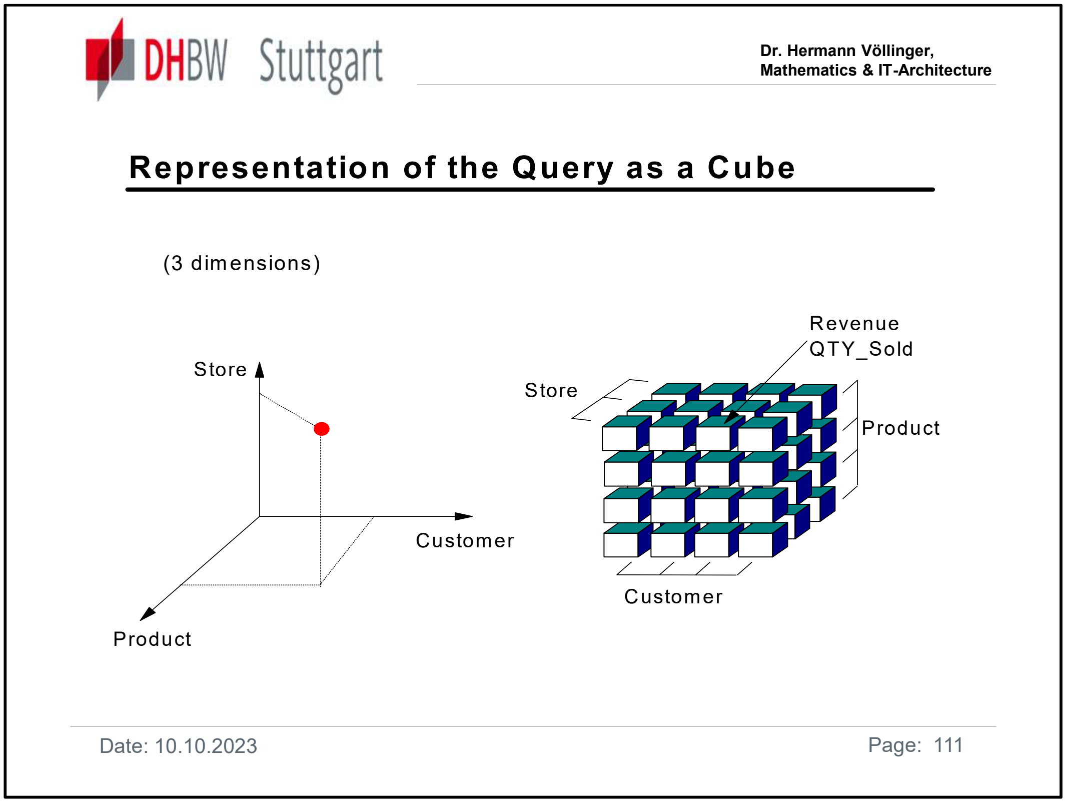
1. **Process oriented requirements**
2. **Information oriented requirements**



Sample Query

- Query:
  "What are the net sales, in terms of revenue (dollars)
  and quantities of items sold,
  Per product,
  Per store and sales region,
  Per customer and customer sales area,
  Per day as well as aggregated over time,
  Over the last two weeks?"
- Evaluation entails viewing historical sales figures from multiple perspectives such as:
  - Sales (overall)
  - Sales per product
  - Sales per store and per sales region
  - Sales per customer and customer sales area
  - Sales per day and aggregated over time
  - Sales and aggregated sales over given time periods

For developing the base concepts of multi-dimensional data modeling, we will use the sample query presented here.
The query is a rather straightforward example of a sales analysis query.

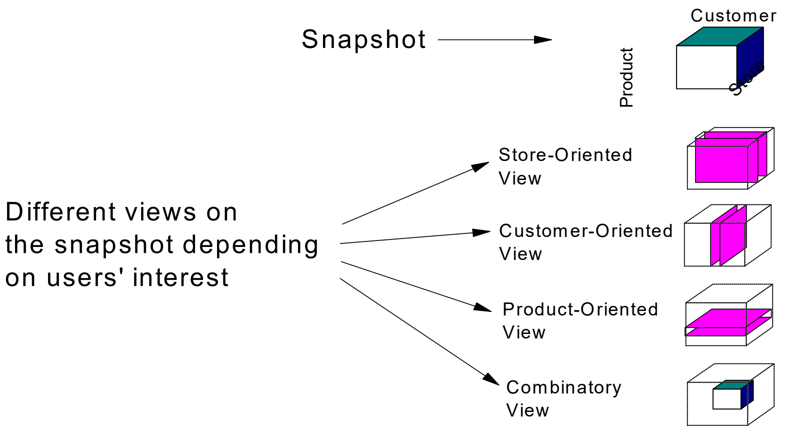


Representation of the Query as a Cube

(3 dimensions)

Multi-dimensional data models could be presented using cubes or using a mathematical notation technique representing points in a multi-dimensional space, for example: **QTY_SOLD = F(S,P,C,t)**

Presentation of the Query as a Cube : Usage



Snapshot

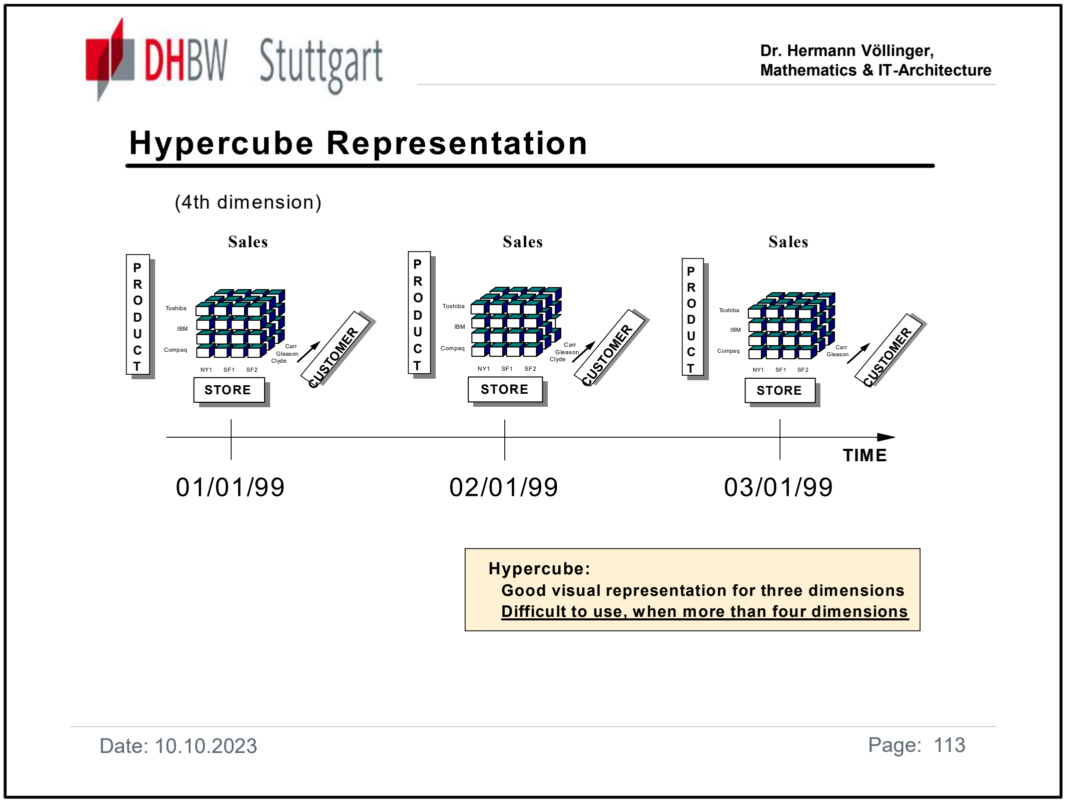
Different views on the snapshot depending on users' interest

Store-Oriented View

Customer-Oriented View

Product-Oriented View

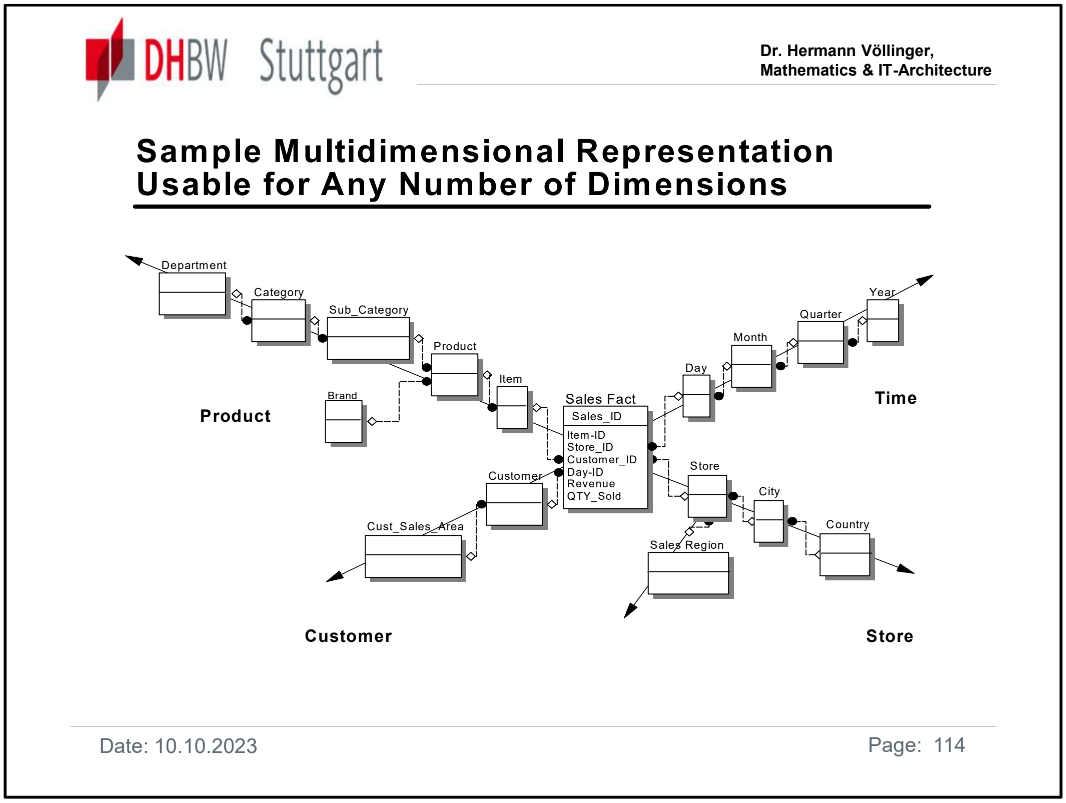
Combinatory View



Hypercube Representation

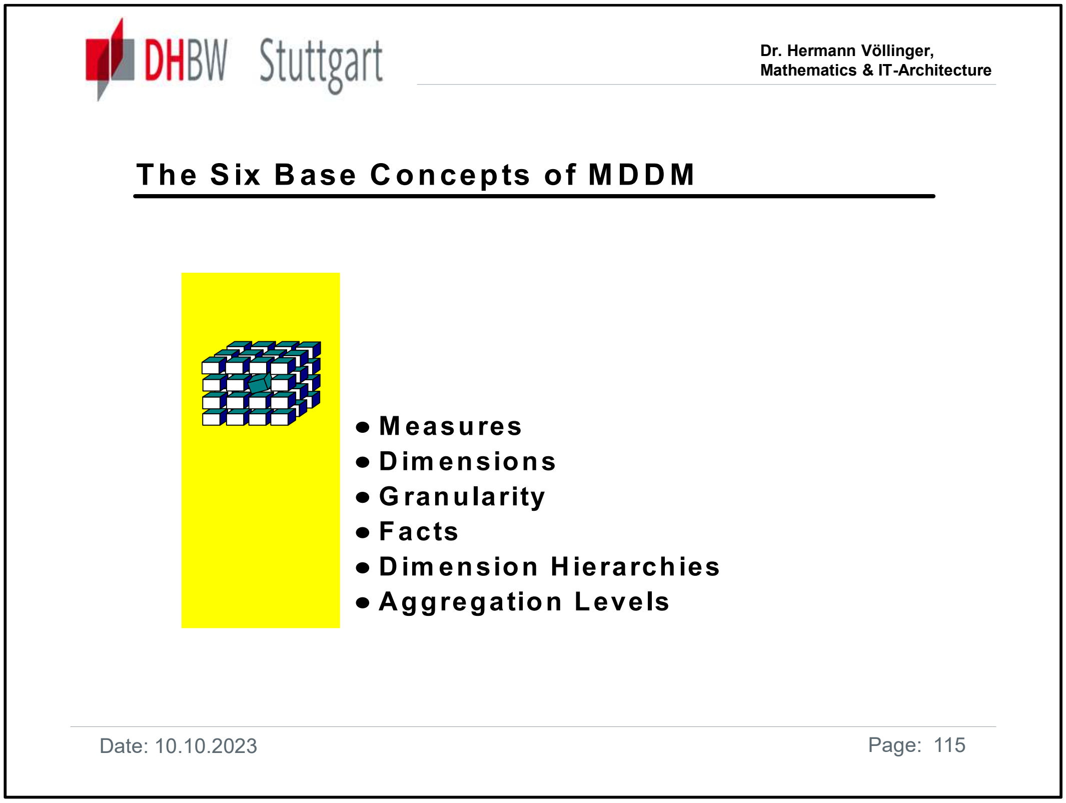
Hypercube:
Good visual representation for three dimensions
Difficult to use, when more than four dimensions

If more than three dimensions are present in the solution, the “cube“ or three-dimensional space representation is no longer usable.
The principle of the cube can be extendedd to “hyper-cube“



Sample Multidimensional Representation Usable for Any Number of Dimensions

The schema presented here show how we will produce the initial dimensional data models we have mentioned before.



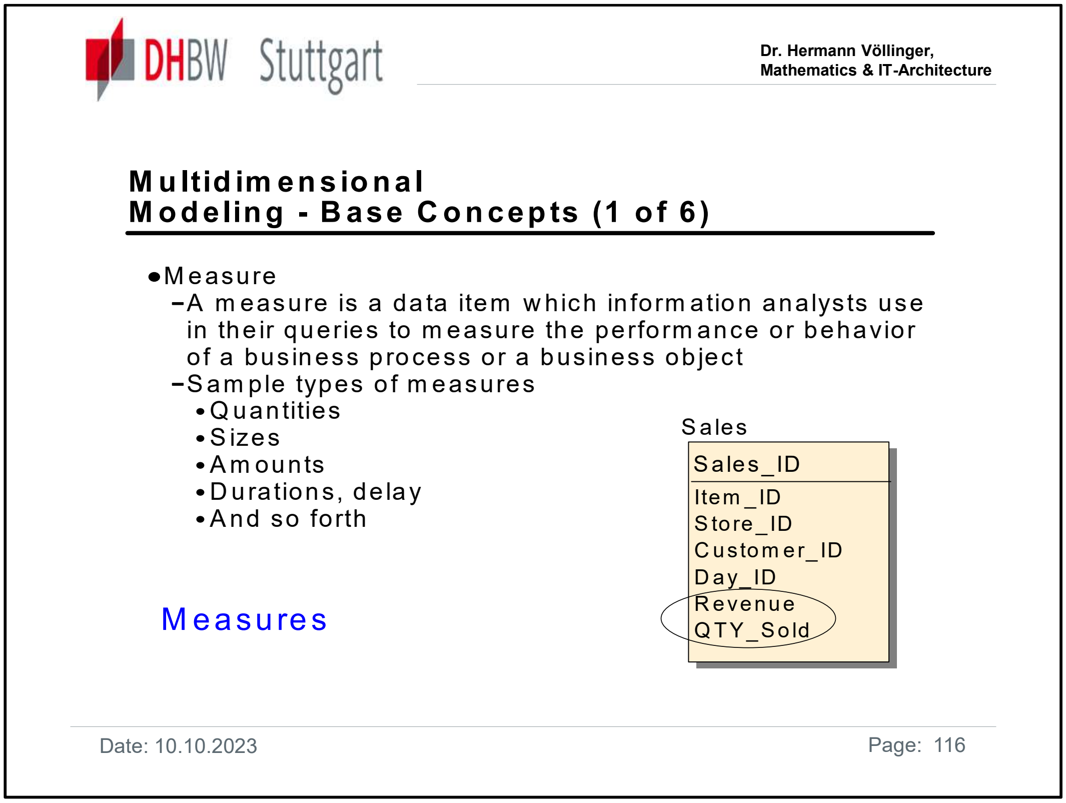
The Six Base Concepts of MDDM

- Measures
- Dimensions
- Granularity
- Facts
- Dimension Hierarchies
- Aggregation Levels

To build an initial multi-dimensional data model, the following six base elements have to be identified:

1. **Measures**
2. **Dimensions**
3. **Grains of dimensions and granularities of measures and facts**
4. **Facts**
5. **Dimension hierarchies**
6. **Aggregation levels**

We will produce our initial multi-dimensional data models base on an analysis of given queries → **query-oriented approach**
Also other approaches are possible.



Multidimensional Modeling - Base Concepts (1 of 6)

- Measure
  - A measure is a data item which information analysts use in their queries to measure the performance or behavior of a business process or a business object
  - Sample types of measures
    - Quantities
    - Sizes
    - Amounts
    - Durations, delay
    - And so forth

Measures

Sales

Sales_ID

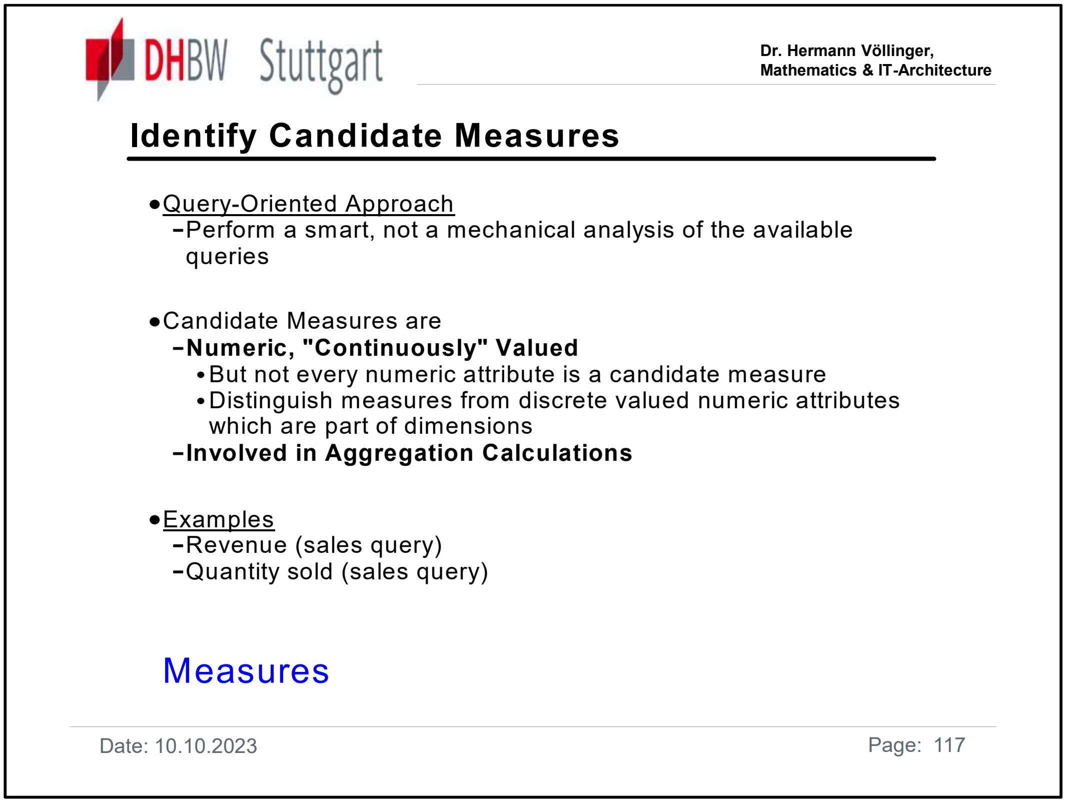
Item_ID
Store_ID
Customer_ID
Day_ID
Revenue
QTY_Sold

Using a query oriented approach, candidates can be identified by analyzing the given end-user query.

The sample query presented before basically include two such measures:

1. **quantity of items sold**
2. **sales revenue**

Remark about KPI: KPI – Key Performance Indicator is a common known synonym for the most important measures of a business.



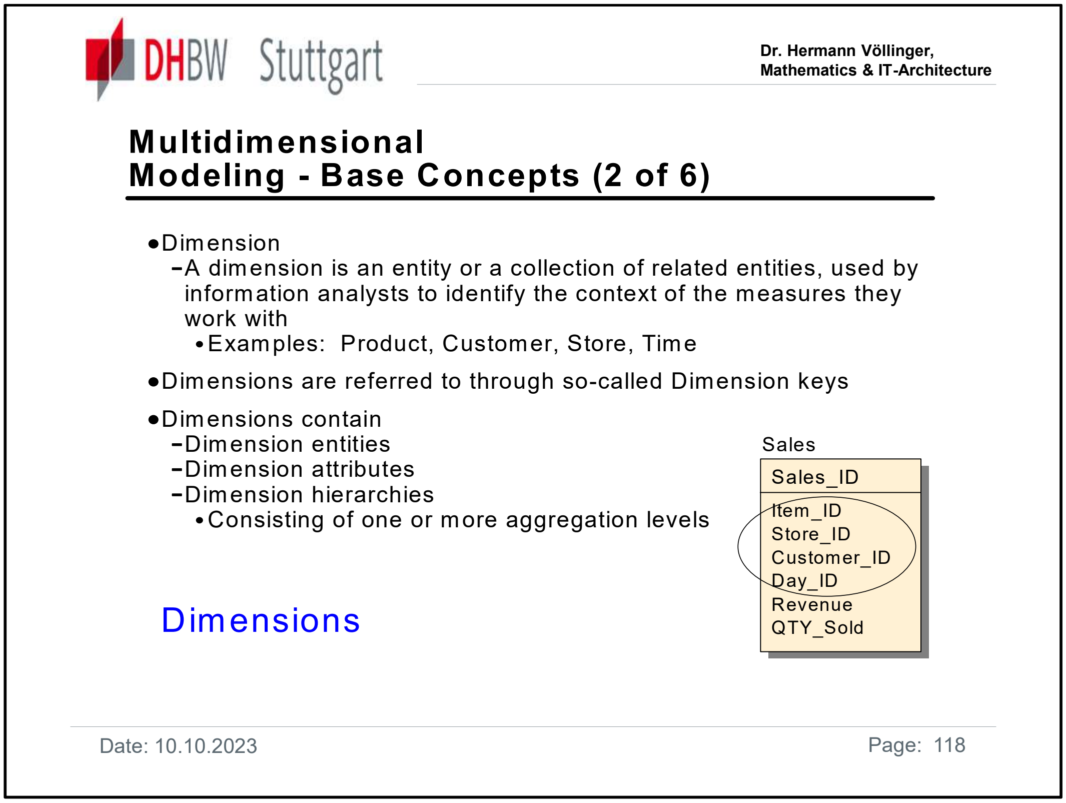
Identify Candidate Measures

- Query-Oriented Approach
  - Perform a smart, not a mechanical analysis of the available queries

- Candidate Measures are
  - **Numeric, "Continuously" Valued**
    - But not every numeric attribute is a candidate measure
    - Distinguish measures from discrete valued numeric attributes which are part of dimensions
  - **Involved in Aggregation Calculations**

- Examples
  - Revenue (sales query)
  - Quantity sold (sales query)

Measures



Multidimensional Modeling - Base Concepts (2 of 6)

- Dimension
  - A dimension is an entity or a collection of related entities, used by information analysts to identify the context of the measures they work with
    - Examples: Product, Customer, Store, Time
- Dimensions are referred to through so-called Dimension keys
- Dimensions contain
  - Dimension entities
  - Dimension attributes
  - Dimension hierarchies
    - Consisting of one or more aggregation levels

Dimensions

Sales

Sales_ID

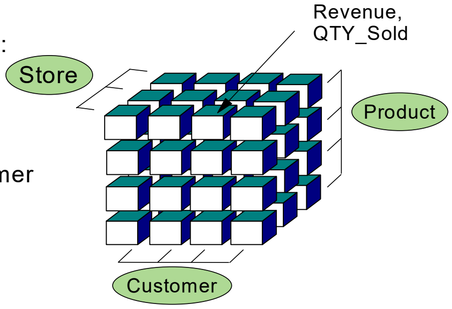
Item_ID
Store_ID
Customer_ID
Day_ID
Revenue
QTY_Sold

Measures require dimensions for their interpretation. As an example, the measure ,'sales revenue' only make sense if we know this value for special item, special customer, at a day and in a certain store -→ we got in or example the four dimensions: **item, store, customer and time.**

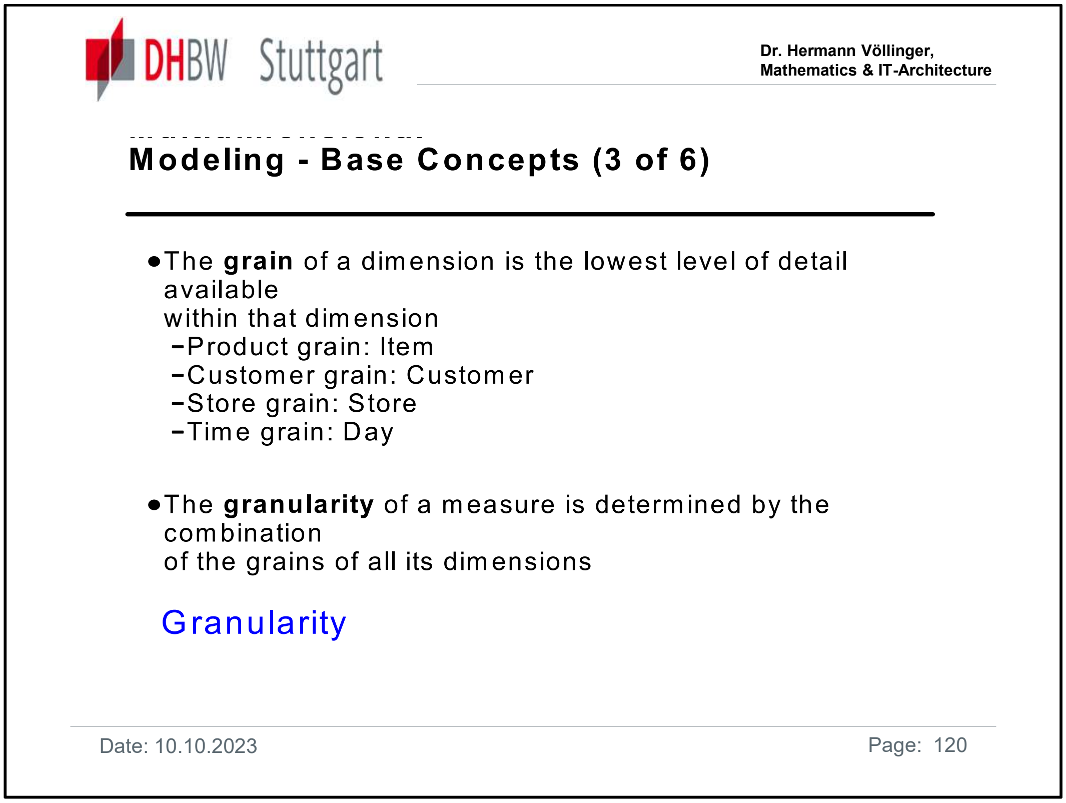
Identify Candidate Dimensions

- Query-Oriented Approach
  - A new dimension shows up each time a query indicates that a measure is aggregated in some way
  - Who, what, where, when, how, ... questions

- Examples
  - Revenue and Quantity sold:
    - Who > Customer
    - What > Product
    - Where > Store
    - When > Time
    - How > Product by Customer



Dimensions



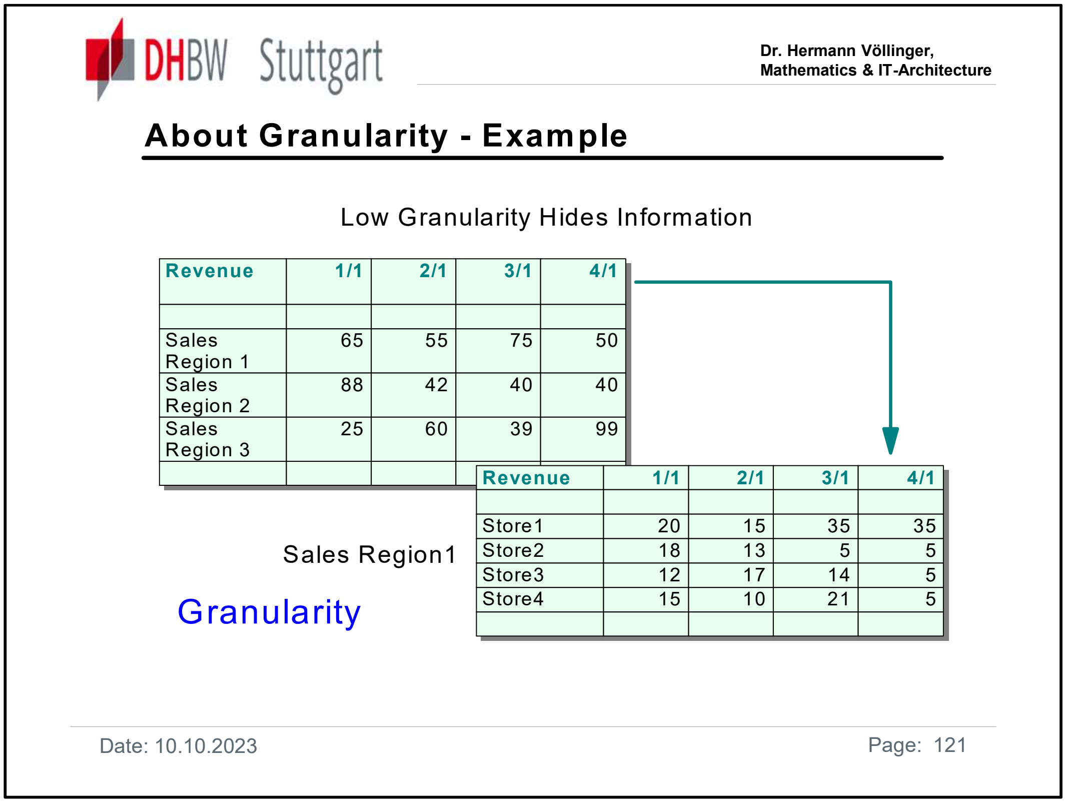
Modeling - Base Concepts (3 of 6)

- The **grain** of a dimension is the lowest level of detail available within that dimension
  - Product grain: Item
  - Customer grain: Customer
  - Store grain: Store
  - Time grain: Day

- The **granularity** of a measure is determined by the combination of the grains of all its dimensions

Granularity

For example the granularity of the measure QTY_SOLD is: **(item, customer, store, day).**
Fine granularity enables fine analysis possibilities, but on the other side it has a big impact on the size of the Data Warehouse.



About Granularity - Example

Low Granularity Hides Information

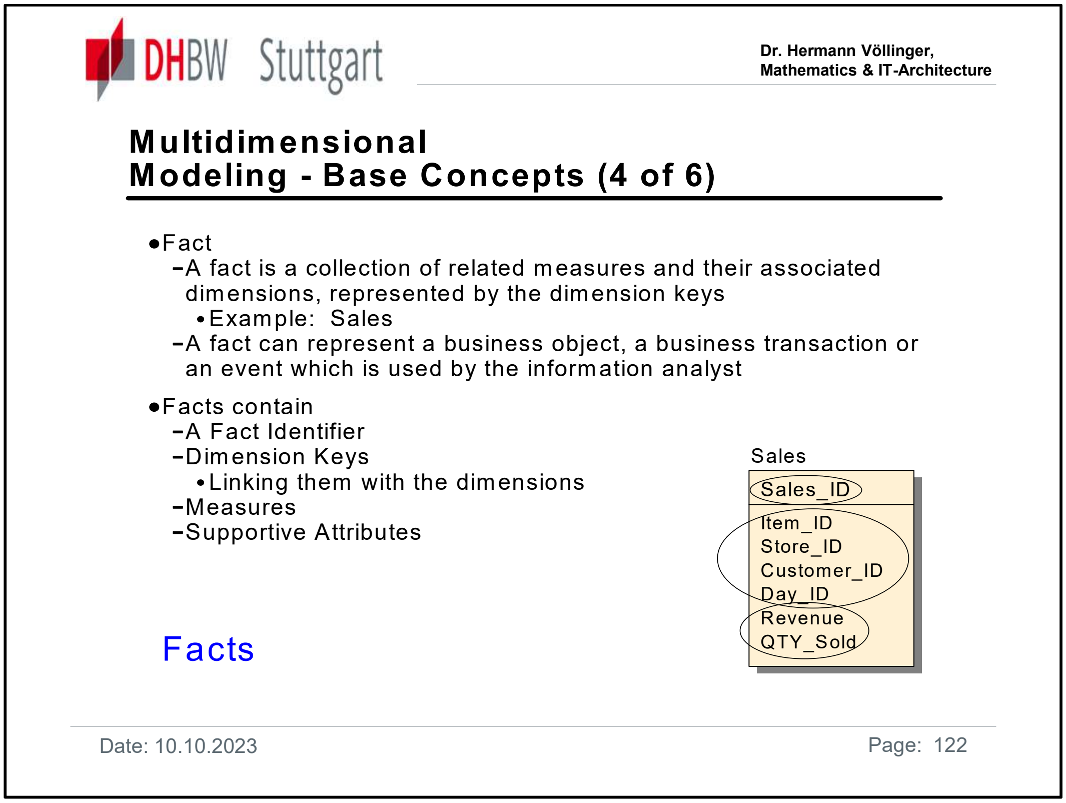
Revenue	1/1	2/1	3/1	4/1
Sales Region 1	65	55	75	50
Sales Region 2	88	42	40	40
Sales Region 3	25	60	39	99

Sales Region1

Revenue	1/1	2/1	3/1	4/1
Store1	20	15	35	35
Store2	18	13	5	5
Store3	12	17	14	5
Store4	15	10	21	5

Granularity

Here we see an example how fine granularity can show ‘hidden‘ information, like that stores in a region are better performers than other.



Multidimensional Modeling - Base Concepts (4 of 6)

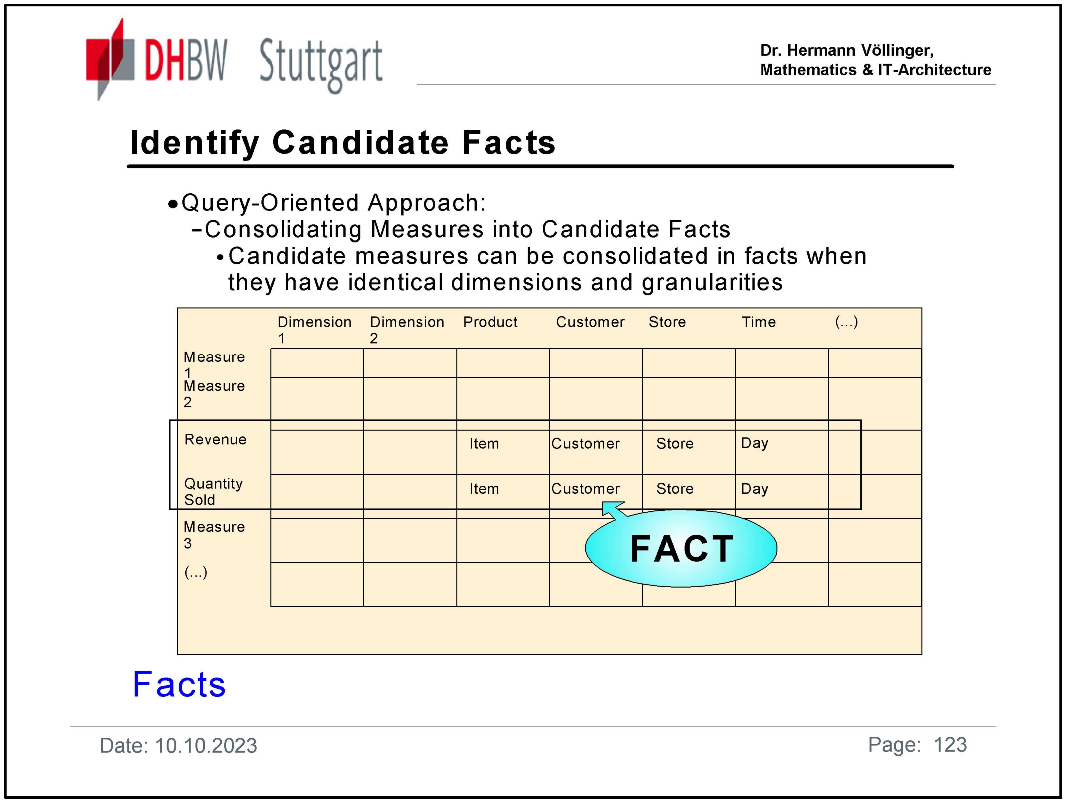
- Fact
  - A fact is a collection of related measures and their associated dimensions, represented by the dimension keys
    - Example: Sales
  - A fact can represent a business object, a business transaction or an event which is used by the information analyst
- Facts contain
  - A Fact Identifier
  - Dimension Keys
    - Linking them with the dimensions
  - Measures
  - Supportive Attributes

Sales

Sales_ID

Item_ID
Store_ID
Customer_ID
Day_ID
Revenue
QTY_Sold

Facts



Identify Candidate Facts

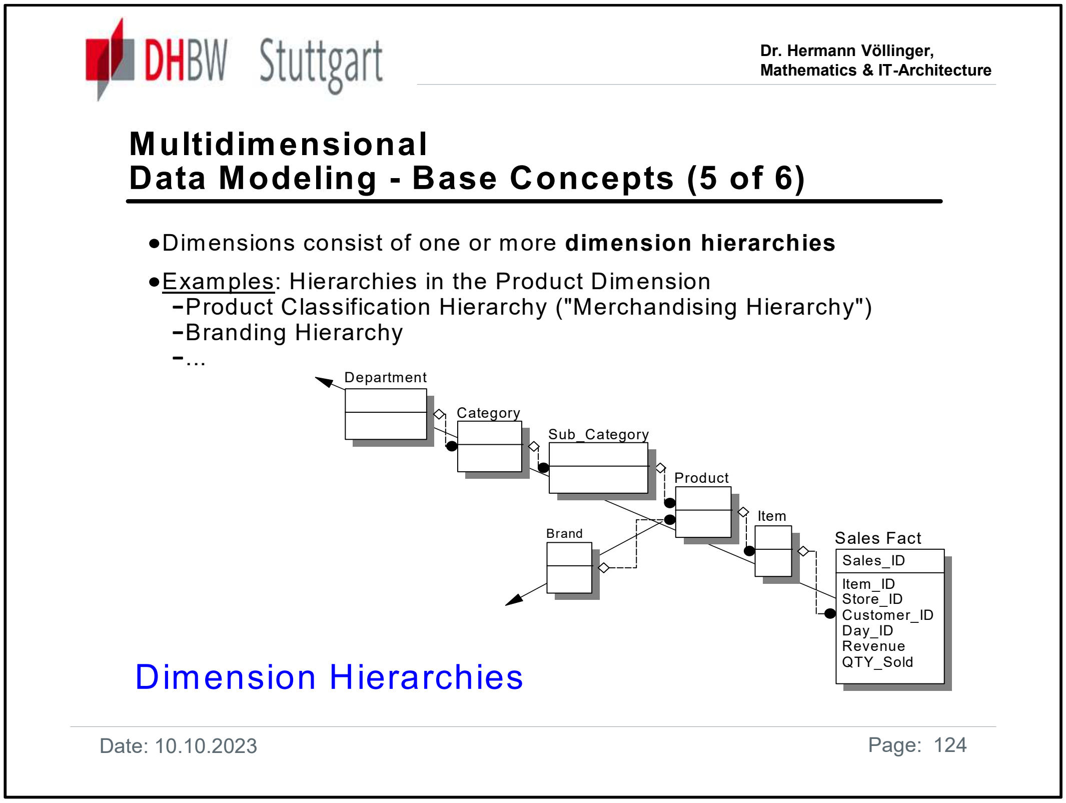
- Query-Oriented Approach:
  - Consolidating Measures into Candidate Facts
    - Candidate measures can be consolidated in facts when they have identical dimensions and granularities

	Dimension 1	Dimension 2	Product	Customer	Store	Time	(...)
Measure 1							
Measure 2							
Revenue			Item	Customer	Store	Day	
Quantity Sold			Item	Customer	Store	Day	
Measure 3							
(...)							

FACT

Facts

Using a query-oriented approach for finding initial modeling artifacts, facts may be identified through consolidating measures which have similar dimensions and identical granularities.
Measures which have same dimensions and same granularity are candidates to be moved in the same fact table.
There exists also more complex methods to find facts.



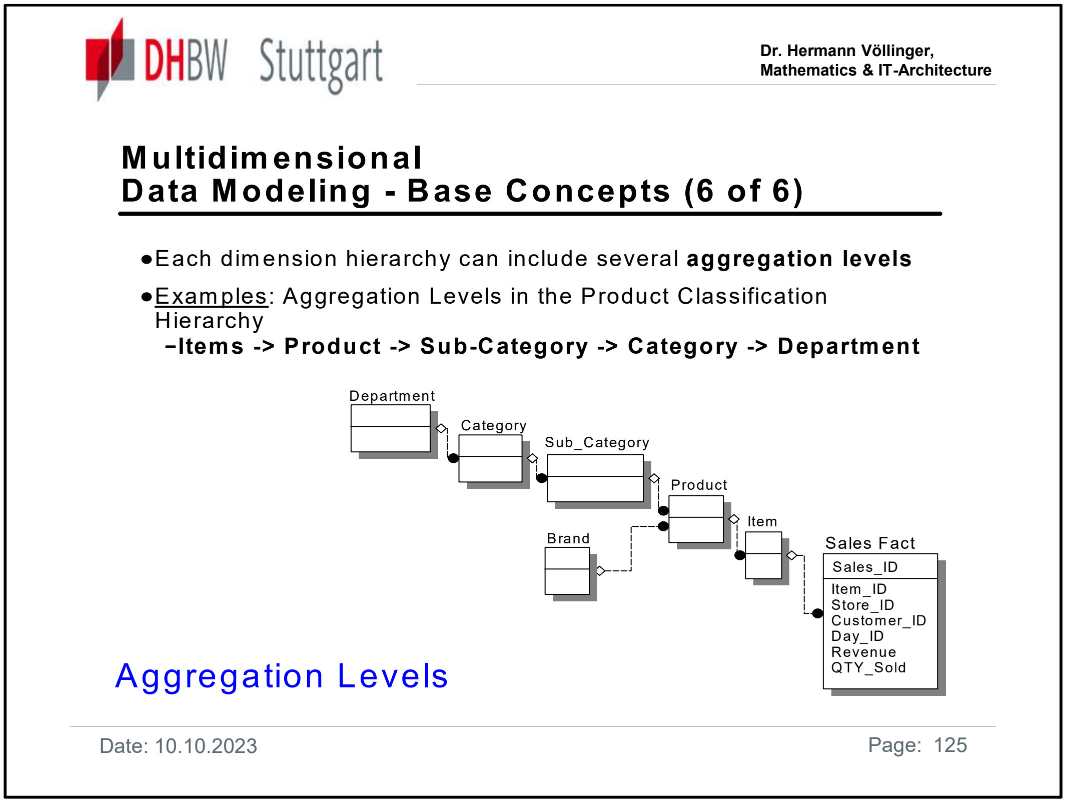
Multidimensional Data Modeling - Base Concepts (5 of 6)

- Dimensions consist of one or more **dimension hierarchies**
- Examples: Hierarchies in the Product Dimension
  - Product Classification Hierarchy ("Merchandising Hierarchy")
  - Branding Hierarchy
  - ...

Dimension Hierarchies

Dimensions consist of one or more dimension-hierarchies.
For example the product dimension in our example has two hierarchies:

1. **‘Merchandising‘ hierarchy**
2. **‘Product Brand‘ hierarchy**



Multidimensional Data Modeling - Base Concepts (6 of 6)

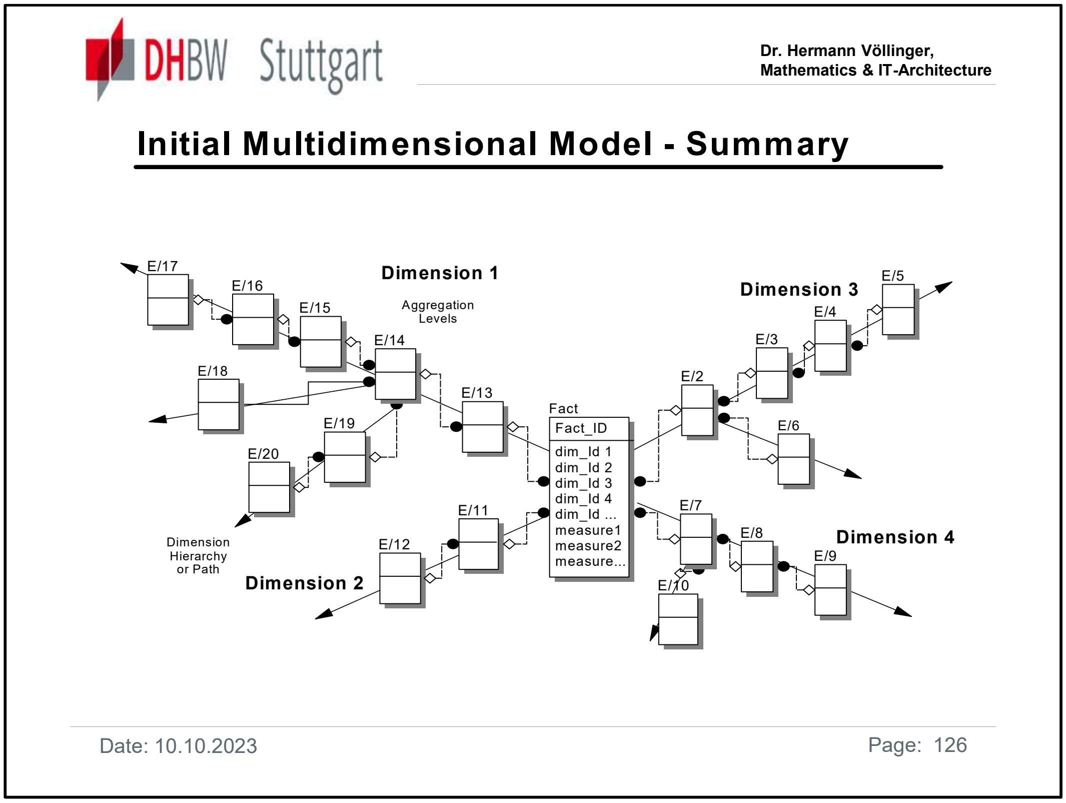
- Each dimension hierarchy can include several **aggregation levels**
- Examples: Aggregation Levels in the Product Classification Hierarchy
  - **Items -> Product -> Sub-Category -> Category -> Department**

Aggregation Levels

Finally, each dimension hierarchy usually include several aggregation levels.
For example:

1. **Item**: 4-pack Duracell AA Alkaline Batteries.
2. **Product**: Duracell AA Alkaline Batteries
3. **Sub-category**: AA Alkaline Batteries
4. **Category**: Batteries
5. **Department**: Supplies

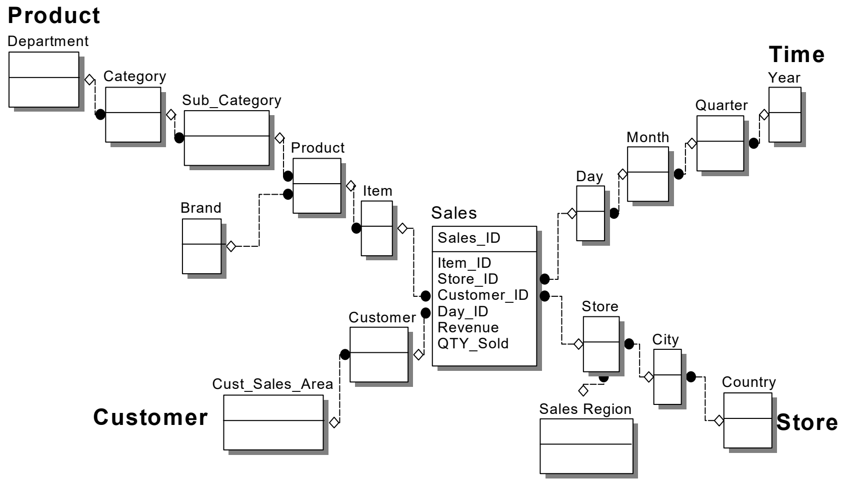
Dimension hierarchies and aggregation levels are used by users when drilling up or down.



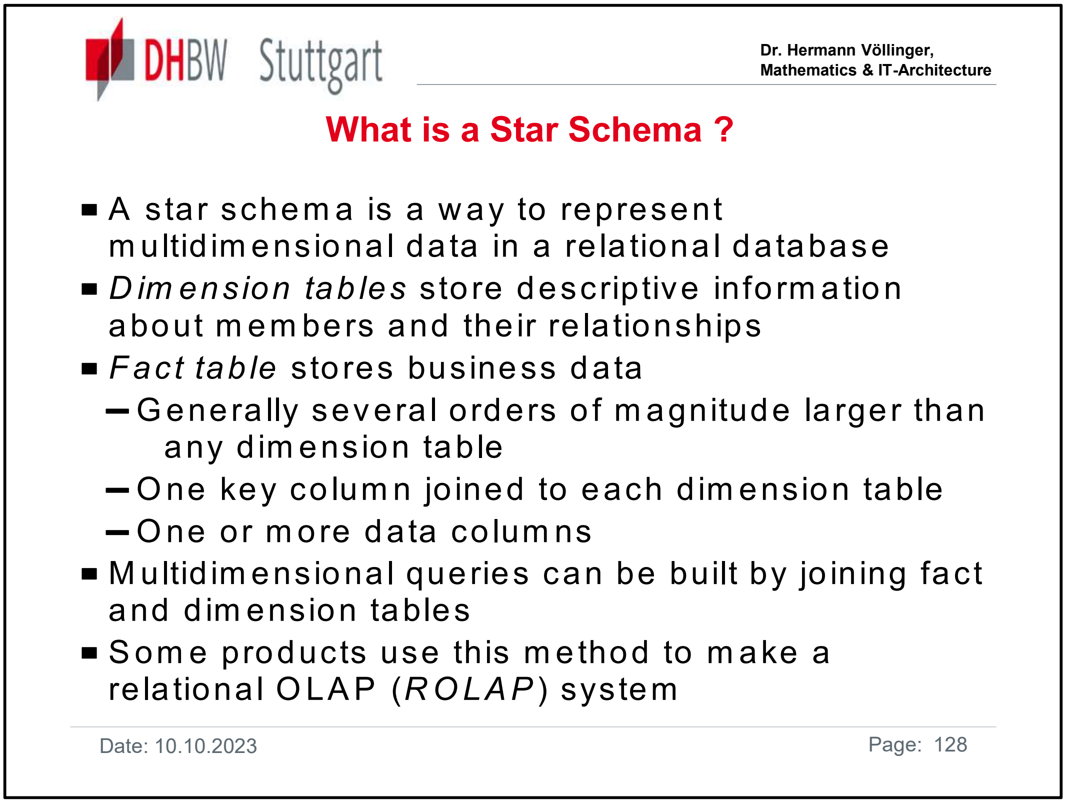
Initial Multidimensional Model - Summary

With these candidate modeling artifacts, initial dimensional models can be built, as illustrated above.

Initial Multidimensional Model - Example



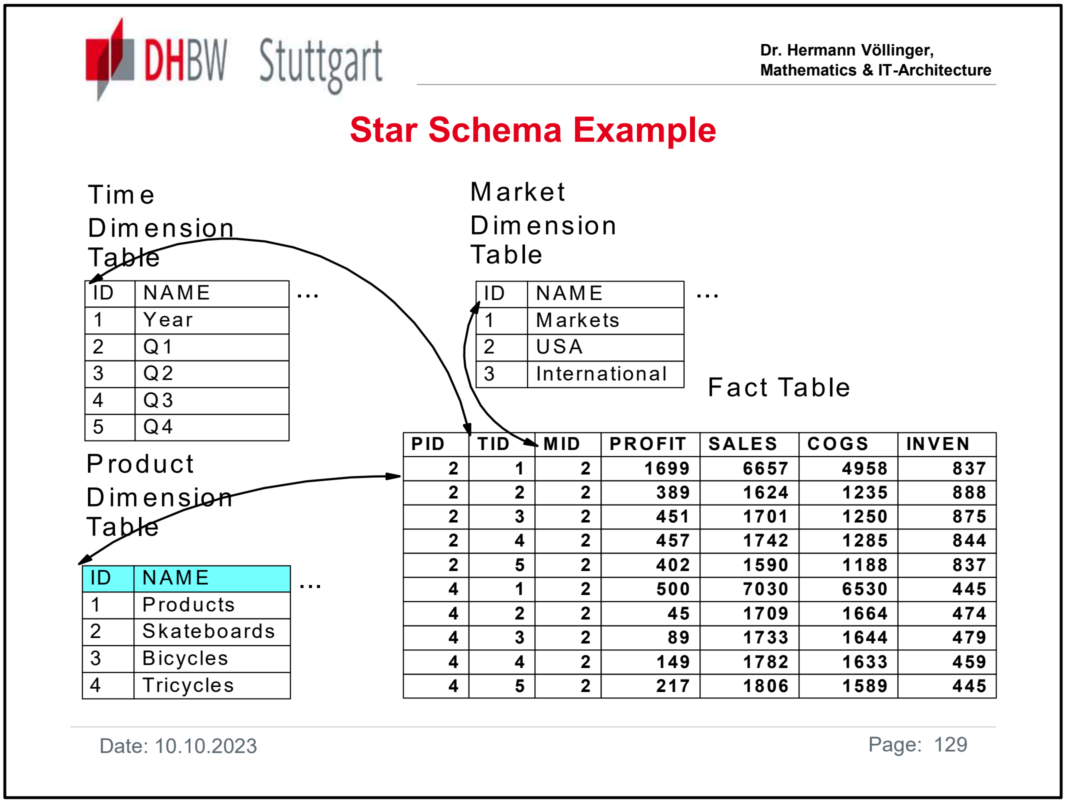
This slide shows **the six base concepts** as they apply to our Sales Query and the initial model that corresponds with that query.



What is a Star Schema ?

- A star schema is a way to represent multidimensional data in a relational database
- *Dimension tables* store descriptive information about members and their relationships
- *Fact table* stores business data
  - Generally several orders of magnitude larger than any dimension table
  - One key column joined to each dimension table
  - One or more data columns
- Multidimensional queries can be built by joining fact and dimension tables
- Some products use this method to make a relational OLAP (*ROLAP*) system

This is a repetition of the chapter about multi-dimensional data modeling (MDDM).



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Star Schema Example

Time Dimension Table

ID	NAME
1	Year
2	Q1
3	Q2
4	Q3
5	Q4

…

Market Dimension Table

ID	NAME
1	Markets
2	USA
3	International

…

Product Dimension Table

ID	NAME
1	Products
2	Skateboards
3	Bicycles
4	Tricycles

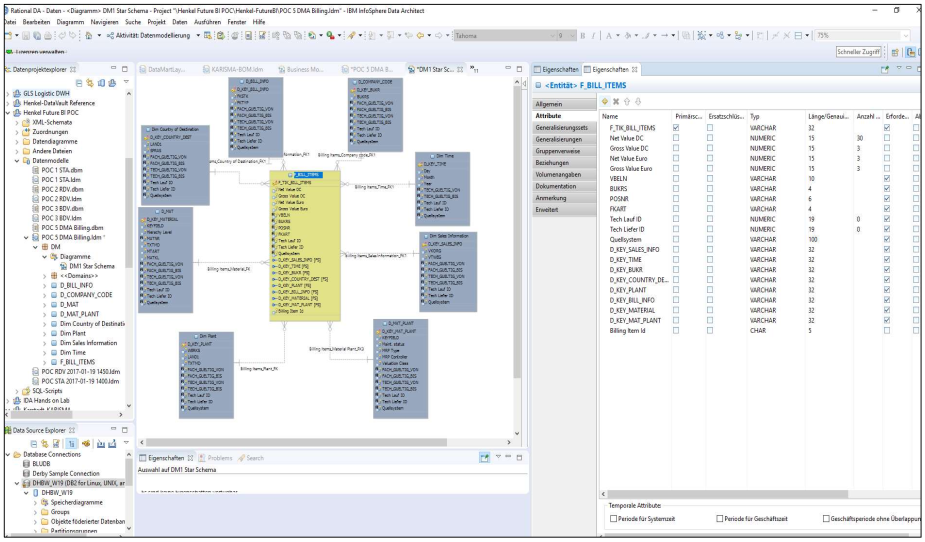
…

Fact Table

PID	TID	MID	PROFIT	SALES	COGS	INVEN
2	1	2	1699	6657	4958	837
2	2	2	389	1624	1235	888
2	3	2	451	1701	1250	875
2	4	2	457	1742	1285	844
2	5	2	402	1590	1188	837
4	1	2	500	7030	6530	445
4	2	2	45	1709	1664	474
4	3	2	89	1733	1644	479
4	4	2	149	1782	1633	459
4	5	2	217	1806	1589	445

This is an example of a simple 3-dim. STAR SCHEMA.
If you use SQL, the downside is that joining and selecting is done each time a query is build. Especially with large DB, this will cost lot of processing and make the system slow ----→ **Killer Query**

Demo1: IBM Infosphere Data Architect (IDA)



Rational DA - Daten - <Diagramm> DM1 Star Schema - Projekt "\Henkel Future BI POC\Henkel-FutureBI\POC 5 DMA Billing.ldm" - IBM InfoSphere Data Architect

<Entität> F_BILL_ITEMS

Name	Primärsc...	Ersatzschlüs...	Typ	Länge/Genaui...	Anzahl ...	Erforde...
F_TIK_BILL_ITEMS	☑	□	VARCHAR	32		☑
Net Value DC	□	□	NUMERIC	15	30	□
Gross Value DC	□	□	NUMERIC	15	3	□
Net Value Euro	□	□	NUMERIC	15	3	□
Gross Value Euro	□	□	NUMERIC	15	3	□
VBELN	□	□	VARCHAR	10		☑
BUKRS	□	□	VARCHAR	4		☑
POSNR	□	□	VARCHAR	6		☑
FKART	□	□	VARCHAR	4		☑
Tech Lauf ID	□	□	NUMERIC	19	0	☑
Tech Liefer ID	□	□	NUMERIC	19	0	☑
Quellsystem	□	□	VARCHAR	100		☑
D_KEY_SALES_INFO	□	□	VARCHAR	32		☑
D_KEY_TIME	□	□	VARCHAR	32		☑
D_KEY_BUKR	□	□	VARCHAR	32		☑
D_KEY_COUNTRY_DE...	□	□	VARCHAR	32		☑
D_KEY_PLANT	□	□	VARCHAR	32		☑
D_KEY_BILL_INFO	□	□	VARCHAR	32		☑
D_KEY_MATERIAL	□	□	VARCHAR	32		☑
D_KEY_MAT_PLANT	□	□	VARCHAR	32		☑
Billing Item Id	□	□	CHAR	5		□

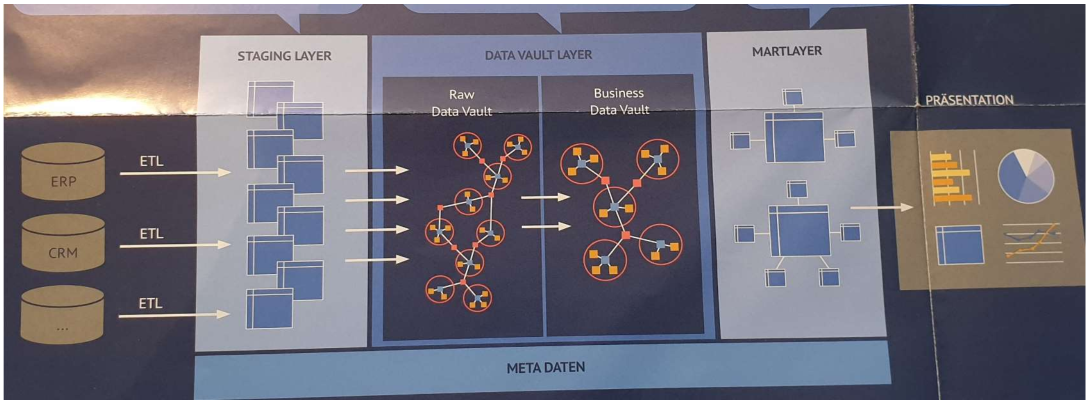
Temporale Attribute:
□ Periode für Systemzeit □ Periode für Geschäftszeit □ Geschäftsperiode ohne Überlappung

IBM® InfoSphere® Data Architect is a collaborative enterprise data modeling and design solution that simplify and accelerate integration design for Business Intelligence, Master Data Management and Service-Oriented Architecture initiatives. IDA enables you to work with users at every step of the data design process, from project management to application design to data design. The tool helps to align processes, services, applications, and data architectures.

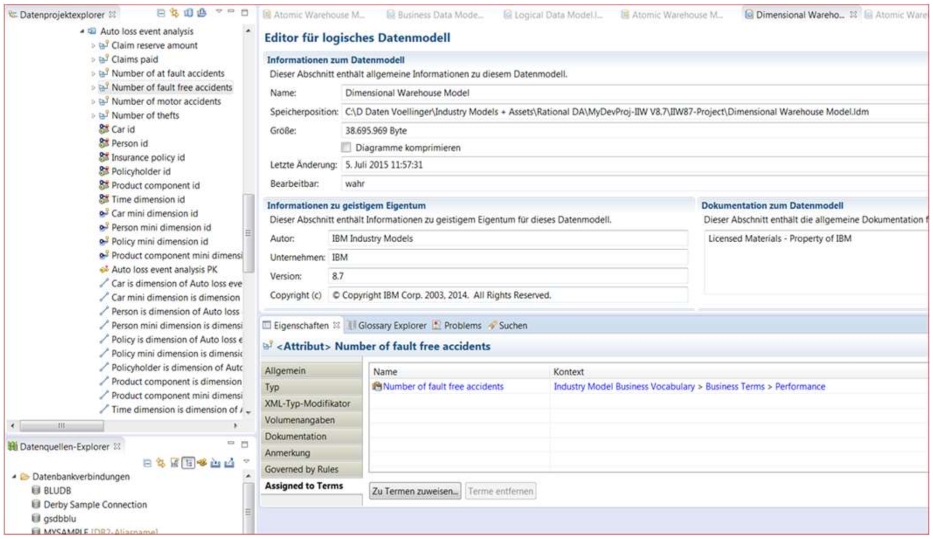
See in [DHBW-Moodle] the video: “Demo-Infosphere_Data_Architect_(IDA).mp4”

We will see in the demo a DWH, which have 4 data layers with different data models. In especially a “Data Vault Layer Model” for the Core DWH:

1. **Staging Area**: Collection of tables which are not linked.
2. **Core DWH:** CDW has layers “Raw Data Vault (RDV)” and “Business Data Vault (BDV)”. RDV and BDV are examples of Data Vault Modeling. https://en.wikipedia.org/wiki/Data_vault_modeling
3. **Data Mart:** Multidimensional data model (STAR schema)



Demo2: Eclipse Plugin “Bridge” of IGC and IDA



Datenprojektexplorer

- Auto loss event analysis
  - Claim reserve amount
  - Claims paid
  - Number of at fault accidents
  - Number of fault free accidents
  - Number of motor accidents
  - Number of thefts
  - Car id
  - Person id
  - Insurance policy id
  - Policyholder id
  - Product component id
  - Time dimension id
  - Car mini dimension id
  - Person mini dimension id
  - Policy mini dimension id
  - Product component mini dimensi
  - Auto loss event analysis PK
  - Car is dimension of Auto loss eve
  - Car mini dimension is dimension
  - Person is dimension of Auto loss
  - Person mini dimension is dimensi
  - Policy is dimension of Auto loss e
  - Policy mini dimension is dimensic
  - Policyholder is dimension of Autc
  - Product component is dimension
  - Product component mini dimensi
  - Time dimension is dimension of /

Datenquellen-Explorer

- Datenbankverbindungen
  - BLUDB
  - Derby Sample Connection
  - gsdbblu
  - MYSAMPLE [DB2-Aliasname]

Atomic Warehouse M... | Business Data Mode... | Logical Data Model.l... | Atomic Warehouse M... | Dimensional Wareho... | Atomic Ware

Editor für logisches Datenmodell

Informationen zum Datenmodell

Dieser Abschnitt enthält allgemeine Informationen zu diesem Datenmodell.

Name: Dimensional Warehouse Model
Speicherposition: C:\D Daten Voellinger\Industry Models + Assets\Rational DA\MyDevProj-IIW V8.7\IIW87-Project\Dimensional Warehouse Model.ldm
Größe: 38.695.969 Byte
☐ Diagramme komprimieren
Letzte Änderung: 5. Juli 2015 11:57:31
Bearbeitbar: wahr

Informationen zu geistigem Eigentum

Dieser Abschnitt enthält Informationen zu geistigem Eigentum für dieses Datenmodell.

Autor: IBM Industry Models
Unternehmen: IBM
Version: 8.7
Copyright (c) © Copyright IBM Corp. 2003, 2014. All Rights Reserved.

Dokumentation zum Datenmodell

Dieser Abschnitt enthält die allgemeine Dokumentation f

Licensed Materials - Property of IBM

Eigenschaften | Glossary Explorer | Problems | Suchen

<Attribut> Number of fault free accidents

Allgemein
Typ
XML-Typ-Modifikator
Volumenangaben
Dokumentation
Anmerkung
Governed by Rules
Assigned to Terms

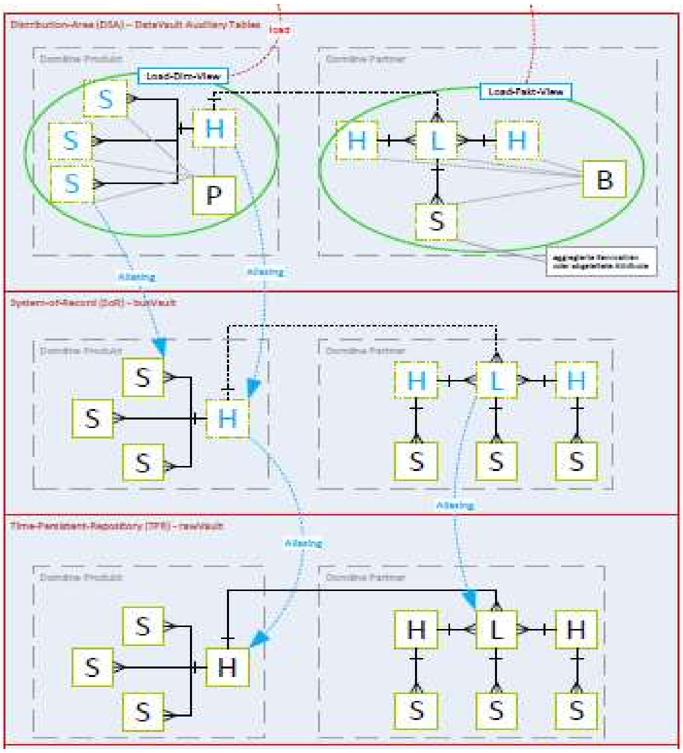
Name	Kontext
Number of fault free accidents	Industry Model Business Vocabulary > Business Terms > Performance

Zu Termen zuweisen... | Terme entfernen

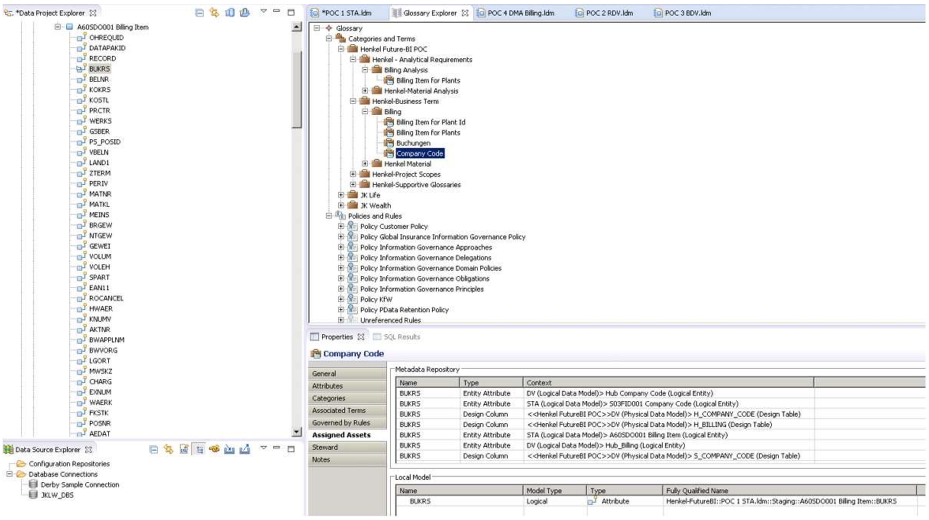
We see the integration of Information Governance Catalog (IGC) as “Eclipse Plugin” in Infosphere Data Architect (IDA). By these we get a very important connection between Data Governance and Data Modelling.

This gives us a bridge between the business-oriented world (for example, business terms…) described in IGC and the associated technical objects (for example, attributes…) defined and developed in the logical data r

<u>The “Data Vault” data layers:</u>


Distribution-Area (DSA) – DataVault Auxiliary Tables
load
Domäne Produkt
Load-Dim-View
S
S
S
H
P
Domäne Partner
Load-Fakt-View
H
L
H
B
S
aggregierte Kennzahlen oder abgeleitete Attribute
Aliasing
Aliasing
System-of-Record (SoR) – busVault
Domäne Produkt
S
S
S
H
Domäne Partner
H
L
H
S
S
S
Aliasing
Time-Persistent-Repository (TPR) – rawVault
Aliasing
Domäne Produkt
S
S
S
H
Domäne Partner
H
L
H
S
S
S

Demo2: Term in IGC with 7 “Assigned Assets” in IDA



Metadata Repository

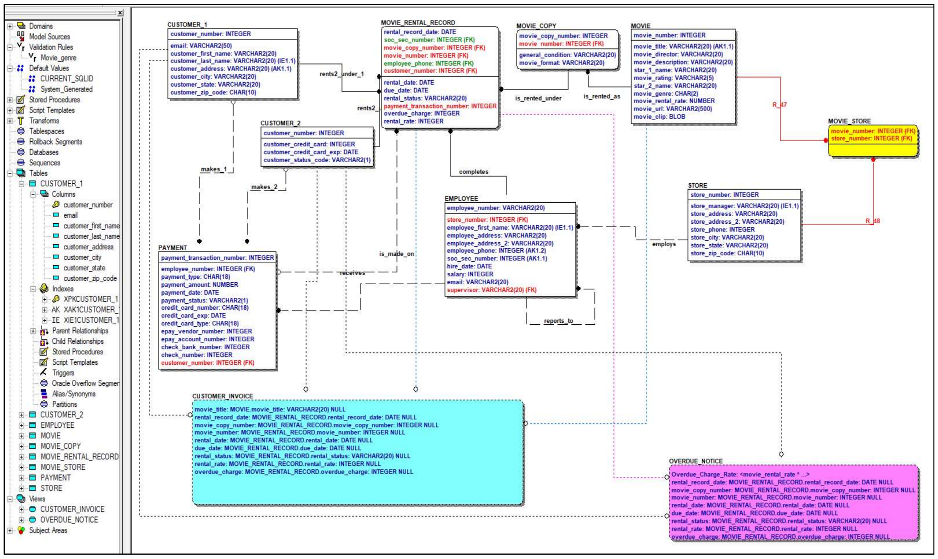
Name	Type	Context
BUKRS	Entity Attribute	DV (Logical Data Model)> Hub Company Code (Logical Entity)
BUKRS	Entity Attribute	STA (Logical Data Model)> S03FID001 Company Code (Logical Entity)
BUKRS	Design Column	<<Henkel FutureBI POC>>DV (Physical Data Model)> H_COMPANY_CODE (Design Table)
BUKRS	Design Column	<<Henkel FutureBI POC>>DV (Physical Data Model)> H_BILLING (Design Table)
BUKRS	Entity Attribute	STA (Logical Data Model)> A60SDO001 Billing Item (Logical Entity)
BUKRS	Entity Attribute	DV (Logical Data Model)> Hub_Billing (Logical Entity)
BUKRS	Design Column	<<Henkel FutureBI POC>>DV (Physical Data Model)> S_COMPANY_CODE (Design Table)

Local Model

Name	Model Type	Type	Fully Qualified Name
BUKRS	Logical	Attribute	Henkel-FutureBI::POC 1 STA.ldm::Staging::A60SDO001 Billing Item::BUKRS

In the above example, we see the business term "Company Code" to which the field with name "BUKRS" belongs in 7 tables in different data layers of the DWH.
This information explains the business meaning of this field and also identifies where this field occurs in the DWH. This is of crucial importance for the analysis/evaluation of the data as well as for the data population process (ETL) when loading data into the DWH.

Demo2: erwin Data Modeler (eDM)



CUSTOMER_1
customer_number: INTEGER
email: VARCHAR2(50)
customer_first_name: VARCHAR2(20)
customer_last_name: VARCHAR2(20) (IE1.1)
customer_address: VARCHAR2(20) (AK1.1)
customer_city: VARCHAR2(20)
customer_state: VARCHAR2(20)
customer_zip_code: CHAR(10)

MOVIE_RENTAL_RECORD
rental_record_date: DATE
soc_sec_number: INTEGER (FK)
movie_copy_number: INTEGER (FK)
movie_number: INTEGER (FK)
employee_phone: INTEGER (FK)
customer_number: INTEGER (FK)
rental_date: DATE
due_date: DATE
rental_status: VARCHAR2(20)
payment_transaction_number: INTEGER
overdue_charge: INTEGER
rental_rate: INTEGER

MOVIE_COPY
movie_copy_number: INTEGER
movie_number: INTEGER (FK)
general_condition: VARCHAR2(20)
movie_format: VARCHAR2(20)

MOVIE
movie_number: INTEGER
movie_title: VARCHAR2(20) (AK1.1)
movie_director: VARCHAR2(20)
movie_description: VARCHAR2(20)
star_1_name: VARCHAR2(20)
movie_rating: VARCHAR2(5)
star_2_name: VARCHAR2(20)
movie_genre: CHAR(2)
movie_rental_rate: NUMBER
movie_url: VARCHAR2(500)
movie_clip: BLOB

MOVIE_STORE
movie_number: INTEGER (FK)
store_number: INTEGER (FK)

CUSTOMER_2
customer_number: INTEGER
customer_credit_card: INTEGER
customer_credit_card_exp: DATE
customer_status_code: VARCHAR2(1)

EMPLOYEE
employee_number: VARCHAR2(20)
store_number: INTEGER (FK)
employee_first_name: VARCHAR2(20) (IE1.1)
employee_address: VARCHAR2(20)
employee_address_2: VARCHAR2(20)
employee_phone: INTEGER (AK1.2)
soc_sec_number: INTEGER (AK1.1)
hire_date: DATE
salary: INTEGER
email: VARCHAR2(20)
supervisor: VARCHAR2(20) (FK)

STORE
store_number: INTEGER
store_manager: VARCHAR2(20) (IE1.1)
store_address: VARCHAR2(20)
store_address_2: VARCHAR2(20)
store_phone: INTEGER
store_city: VARCHAR2(20)
store_state: VARCHAR2(20)
store_zip_code: CHAR(10)

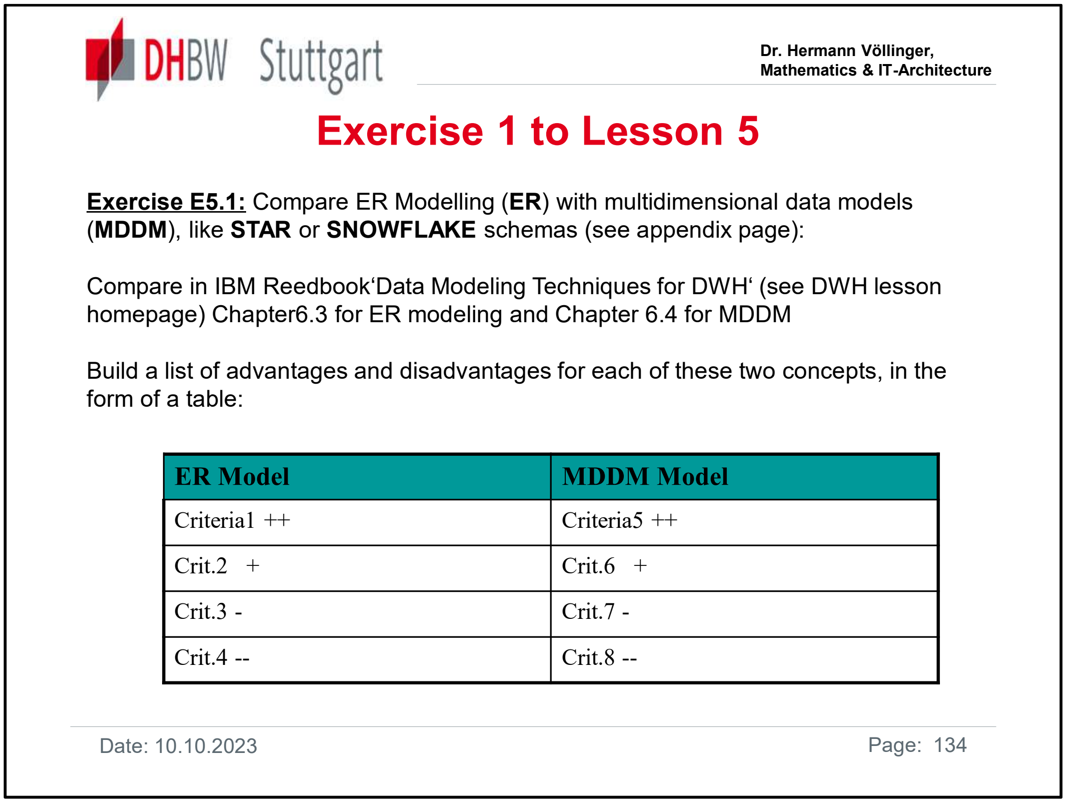
PAYMENT
payment_transaction_number: INTEGER
employee_number: INTEGER (FK)
payment_type: CHAR(18)
payment_amount: NUMBER
payment_date: DATE
payment_status: VARCHAR2(1)
credit_card_number: CHAR(18)
credit_card_exp: DATE
credit_card_type: CHAR(18)
epay_vendor_number: INTEGER
epay_account_number: INTEGER
check_bank_number: INTEGER
check_number: INTEGER
customer_number: INTEGER (FK)

CUSTOMER_INVOICE
movie_title: MOVIE.movie_title: VARCHAR2(20) NULL
rental_record_date: MOVIE_RENTAL_RECORD.rental_record_date: DATE NULL
movie_copy_number: MOVIE_RENTAL_RECORD.movie_copy_number: INTEGER NULL
movie_number: MOVIE_RENTAL_RECORD.movie_number: INTEGER NULL
rental_date: MOVIE_RENTAL_RECORD.rental_date: DATE NULL
due_date: MOVIE_RENTAL_RECORD.due_date: DATE NULL
rental_status: MOVIE_RENTAL_RECORD.rental_status: VARCHAR2(20) NULL
rental_rate: MOVIE_RENTAL_RECORD.rental_rate: INTEGER NULL
overdue_charge: MOVIE_RENTAL_RECORD.overdue_charge: INTEGER NULL

OVERDUE_NOTICE
Overdue_Charge_Rate: <movie_rental_rate * ...>
rental_record_date: MOVIE_RENTAL_RECORD.rental_record_date: DATE NULL
movie_copy_number: MOVIE_RENTAL_RECORD.movie_copy_number: INTEGER NULL
movie_number: MOVIE_RENTAL_RECORD.movie_number: INTEGER NULL
rental_date: MOVIE_RENTAL_RECORD.rental_date: DATE NULL
due_date: MOVIE_RENTAL_RECORD.due_date: DATE NULL
rental_status: MOVIE_RENTAL_RECORD.rental_status: VARCHAR2(20) NULL
rental_rate: MOVIE_RENTAL_RECORD.rental_rate: INTEGER NULL
overdue_charge: MOVIE_RENTAL_RECORD.overdue_charge: INTEGER NULL

https://erwin.com/products/erwin-data-modeler/

https://www.youtube.com/watch?v=YTh43ylo_3o#action=share



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise 1 to Lesson 5

Exercise E5.1: Compare ER Modelling (**ER**) with multidimensional data models (**MDDM**), like **STAR** or **SNOWFLAKE** schemas (see appendix page):

Compare in IBM Reedbook'Data Modeling Techniques for DWH' (see DWH lesson homepage) Chapter6.3 for ER modeling and Chapter 6.4 for MDDM

Build a list of advantages and disadvantages for each of these two concepts, in the form of a table:

ER Model	MDDM Model
Criteria1 ++	Criteria5 ++
Crit.2 +	Crit.6 +
Crit.3 -	Crit.7 -
Crit.4 --	Crit.8 --

Date: 10.10.2023 Page: 134

The BI logical data models are compared here:

Entity-relationship: An entity-relationship logical design is data-centric in nature. In other words, the database design reflects the nature of the data to be stored in the database, as opposed to reflecting the anticipated usage of that data.

Because an entity-relationship design is not usage-specific, it can be used for a variety of application types: OLTP and batch, as well as business intelligence.

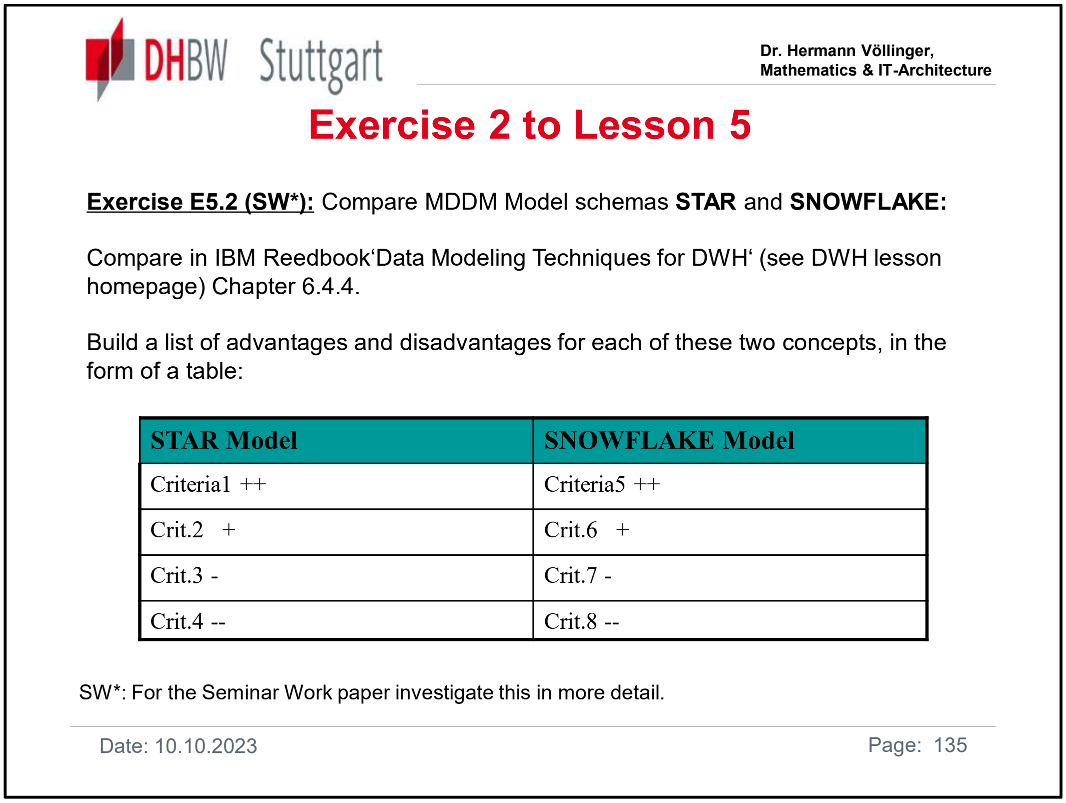
This same usage flexibility makes an entity-relationship design appropriate for a data warehouse that must support a wide range of query types and business objectives.

Star schema: The star schema logical design, unlike the entity-relationship model, is specifically geared towards decision support applications.

The design is intended to provide very efficient access to information in support of a predefined set of business requirements. A star schema is generally not suitable for general-purpose query applications.

A **star schema** consists of a central fact table surrounded by dimension tables and is frequently referred to as a multidimensional model. Although the original concept was to have up to five dimensions as a star has five points, many stars today have more than five dimensions. The information in the star usually meets the following guidelines:

- A fact table contains numerical elements
- A dimension table contains textual elements
- The primary key of each dimension table is a foreign key of the fact table
- A column in one dimension table should not appear in any other dimension table.



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise 2 to Lesson 5

Exercise E5.2 (SW*): Compare MDDM Model schemas **STAR** and **SNOWFLAKE:**

Compare in IBM Reedbook'Data Modeling Techniques for DWH' (see DWH lesson homepage) Chapter 6.4.4.

Build a list of advantages and disadvantages for each of these two concepts, in the form of a table:

STAR Model	SNOWFLAKE Model
Criteria1 ++	Criteria5 ++
Crit.2 +	Crit.6 +
Crit.3 -	Crit.7 -
Crit.4 --	Crit.8 --

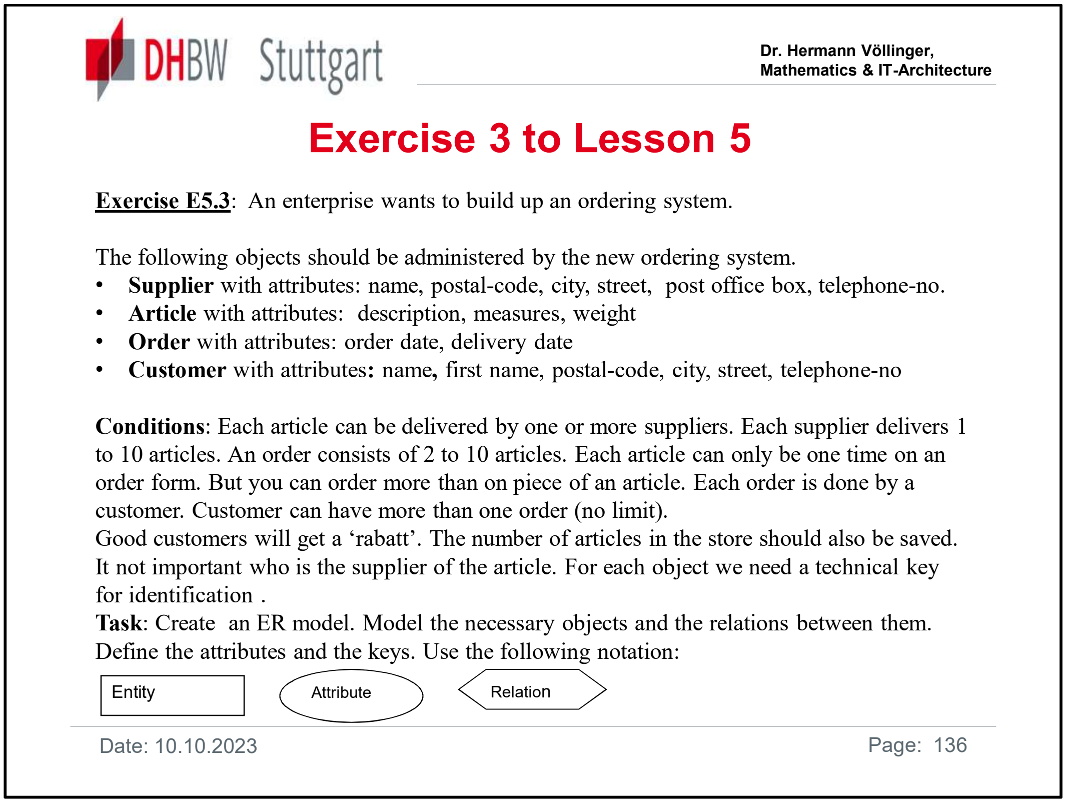
SW*: For the Seminar Work paper investigate this in more detail.

Date: 10.10.2023 Page: 135

Snowflake Schema The snowflake model is a further normalized version of the star schema. When a dimension table contains data that is not always necessary for queries, too much data may be picked up each time a dimension table is accessed. To eliminate access to this data, it is kept in a separate table off the dimension, thereby making the star resemble a snowflake.

The key advantage of a snowflake design is improved query performance. This is achieved because less data is retrieved and joins involve smaller, normalized tables rather than larger, denormalized tables. The snowflake schema also increases flexibility because of normalization and can possibly lower the granularity of the dimensions.

The disadvantage of a snowflake design is that it increases both the number of tables a user must deal with and the complexities of some queries. For this reason, many experts suggest refraining from using the snowflake schema. Having entity attributes in multiple tables, the same amount of information is available whether a single table or multiple tables are used.



Exercise 3 to Lesson 5

Exercise E5.3: An enterprise wants to build up an ordering system.

The following objects should be administered by the new ordering system.

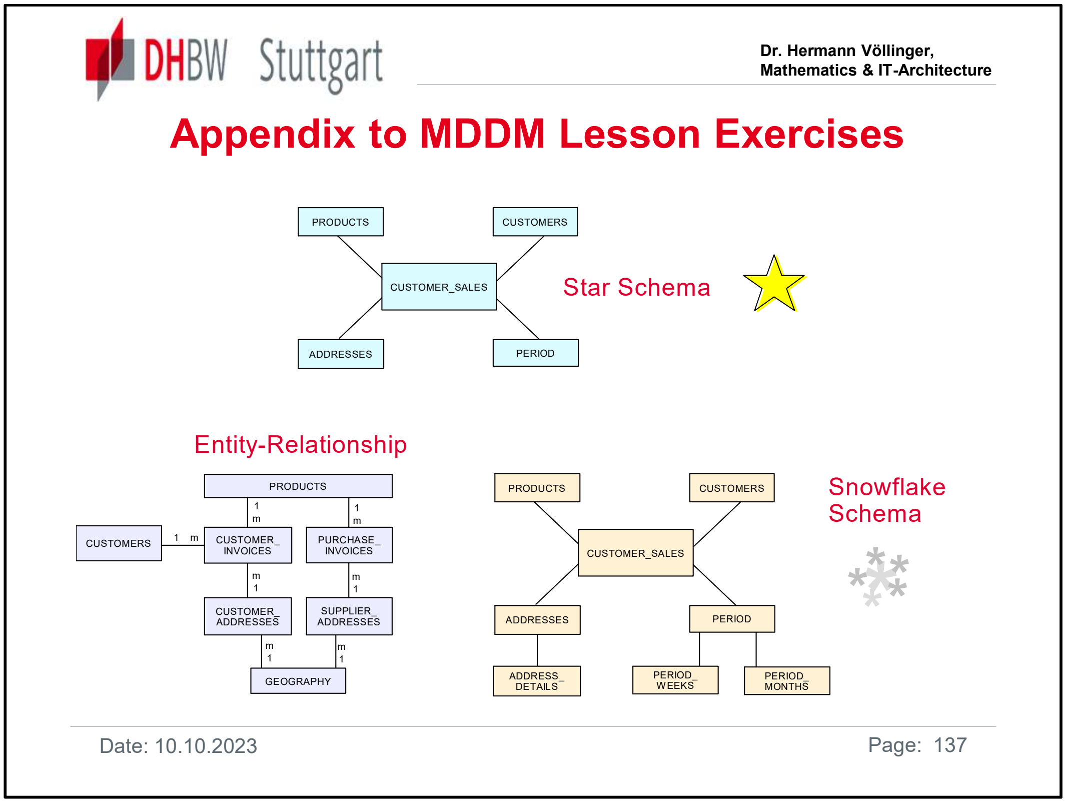
- **Supplier** with attributes: name, postal-code, city, street, post office box, telephone-no.
- **Article** with attributes: description, measures, weight
- **Order** with attributes: order date, delivery date
- **Customer** with attributes**:** name**,** first name, postal-code, city, street, telephone-no

Conditions: Each article can be delivered by one or more suppliers. Each supplier delivers 1 to 10 articles. An order consists of 2 to 10 articles. Each article can only be one time on an order form. But you can order more than on piece of an article. Each order is done by a customer. Customer can have more than one order (no limit).
Good customers will get a 'rabatt'. The number of articles in the store should also be saved. It not important who is the supplier of the article. For each object we need a technical key for identification .
Task: Create an ER model. Model the necessary objects and the relations between them. Define the attributes and the keys. Use the following notation:

Entity

Attribute

Relation



Appendix to MDDM Lesson Exercises

Star Schema

PRODUCTS

CUSTOMERS

CUSTOMER_SALES

ADDRESSES

PERIOD

Entity-Relationship

PRODUCTS

CUSTOMERS

CUSTOMER_ INVOICES

PURCHASE_ INVOICES

CUSTOMER_ ADDRESSES

SUPPLIER_ ADDRESSES

GEOGRAPHY

1 m

Snowflake Schema

PRODUCTS

CUSTOMERS

CUSTOMER_SALES

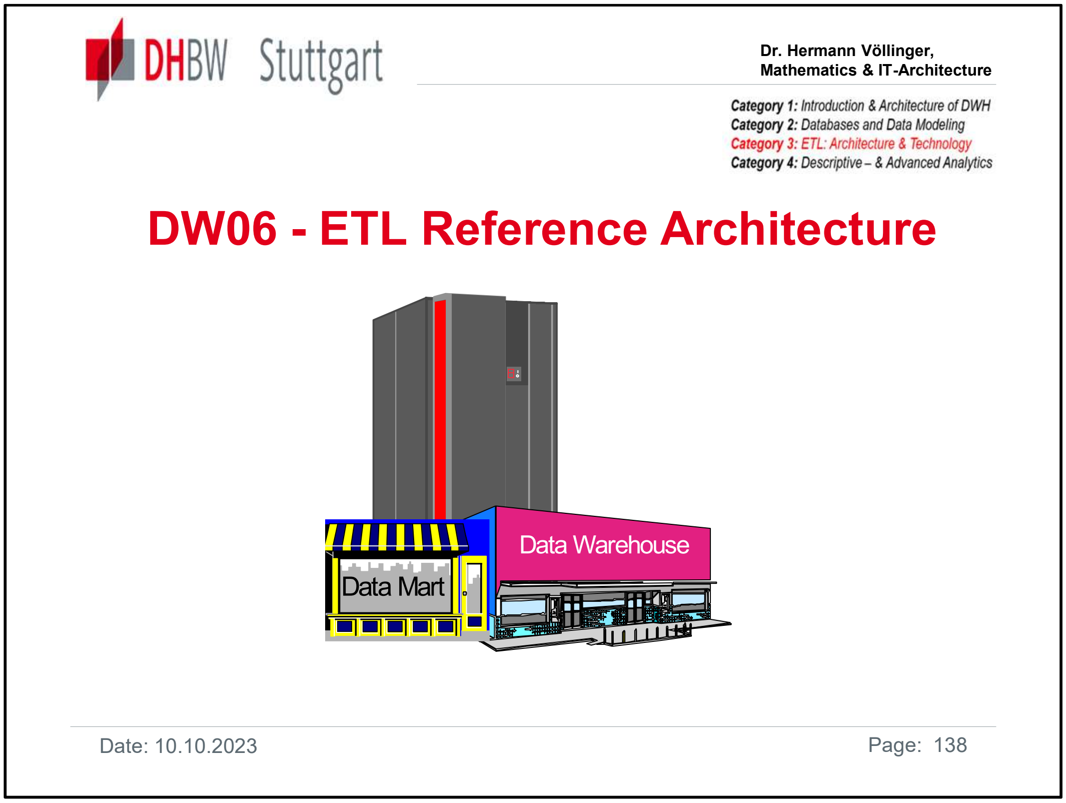
ADDRESSES

PERIOD

ADDRESS_ DETAILS

PERIOD_ WEEKS

PERIOD_ MONTHS



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Category 1: Introduction & Architecture of DWH
Category 2: Databases and Data Modeling
Category 3: ETL: Architecture & Technology
Category 4: Descriptive – & Advanced Analytics

DW06 - ETL Reference Architecture

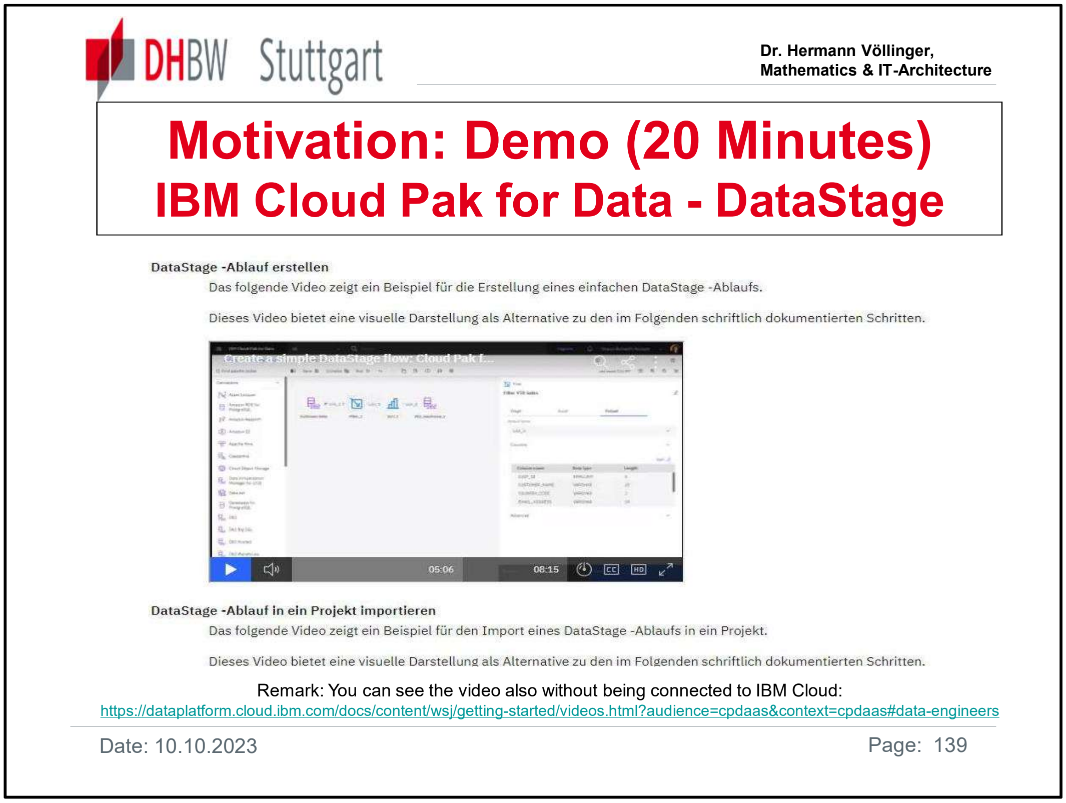
Date: 10.10.2023 Page: 138

Data integration encompasses the processes that serve to manage and centralize the data flows from different sources - with the aim of goal of using this data as an aid to decision-making. For more information see in [DHBW-Moodle] the document **<< Leitfaden_fuer_die_Datenintegration.pdf >>.** This guide discusses the following topics:

- Data integration and its importance
- The traditional approach to data integration: ETL (Extract, Transform, Load)
- The newer approach enabled by the cloud: ELT (Extract, Load, Transform)
- The benefits of automating the data integration process
- The evaluation and implementation of data integration tools

The following **chapter DW06** gives an **introduction into the ETL Reference Architecture** as part of the overall architecture of a data warehouse. It explains the architectural ideas behind an ETL solution. It shows in detail the different components of an ETL Reference Architecture. It also explains the three **steps of a successful ETL Strategy**. See the details:

- Three steps of a successful ETL Strategy
  Discover, Prepare, Transform
- ETL components PSS and Pipe
- Metadata Layer and Warehouse Management
- The different Data Layers
  1. Source Systems of a DWH
  2. Core Data Warehouse (CDW)
  3. Data Mart (DM)
- CDW Archive (Data and Process)



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Motivation: Demo (20 Minutes)
IBM Cloud Pak for Data - DataStage

DataStage -Ablauf erstellen

Das folgende Video zeigt ein Beispiel für die Erstellung eines einfachen DataStage -Ablaufs.

Dieses Video bietet eine visuelle Darstellung als Alternative zu den im Folgenden schriftlich dokumentierten Schritten.

DataStage -Ablauf in ein Projekt importieren

Das folgende Video zeigt ein Beispiel für den Import eines DataStage -Ablaufs in ein Projekt.

Dieses Video bietet eine visuelle Darstellung als Alternative zu den im Folgenden schriftlich dokumentierten Schritten.

Remark: You can see the video also without being connected to IBM Cloud:
https://dataplatform.cloud.ibm.com/docs/content/wsj/getting-started/videos.html?audience=cpdaas&context=cpdaas#data-engineers

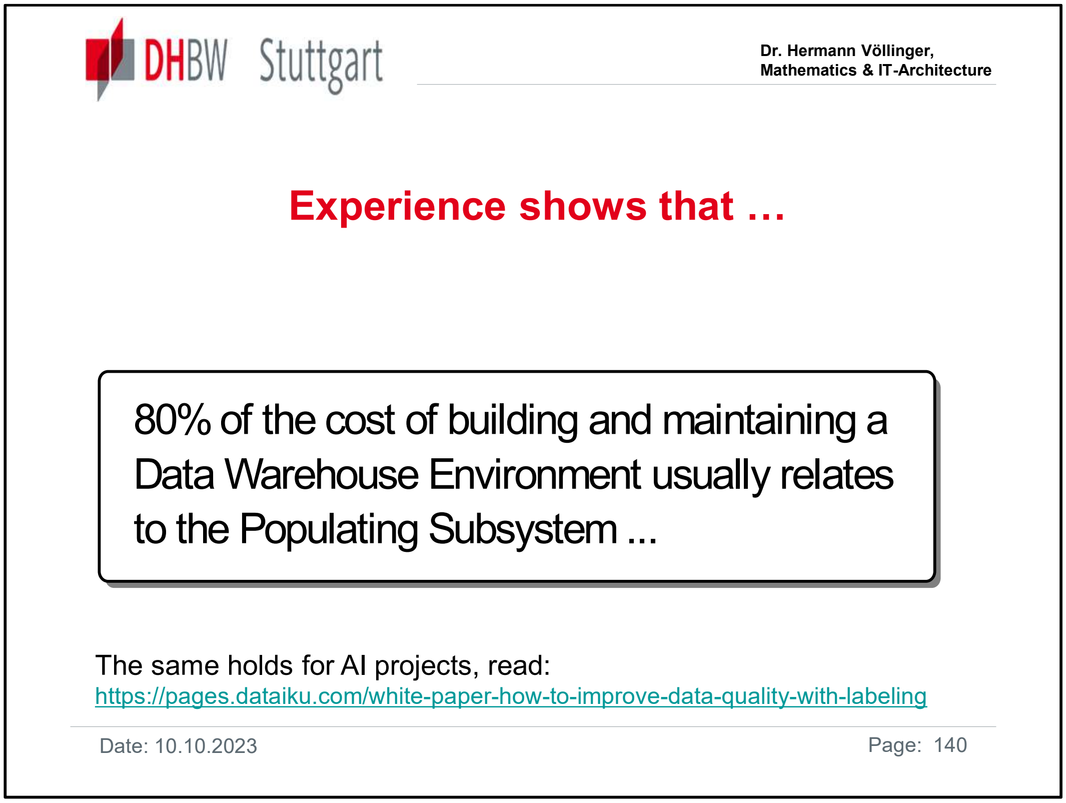
Date: 10.10.2023 Page: 139

.

The DataStage flow contains four nodes: the original data set, a filter node, a sort node and the transformed data set. Perform the following steps to **create the DataStage flow**:

1. Click New Asset > DataStage flow.
2. Enter a name and a description and then click Create.
3. Click on Connectors and then drag the Asset Browser node into the creation area.
4. Select Data asset > customers.csv and click Add.
5. In the node palette, expand the Phases section and drag the Filter node into the creation area.
6. To link the nodes, click on the blue arrow on the Customers.csv node and drag it to the Filter node.
7. In the Phases section, drag the Sort node into the creation area.
8. Connect the Filter node to the Sort node.
9. Expand the Connectors section and drag the Asset Browser node into the creation area.
10. Select Data asset > customers.csv and click Add. You will change the file name later so that you do not overwrite the 'customer.csv' file.
11. Connect the Sort node to this last Customers.csv node.

As a next step you can check the results.

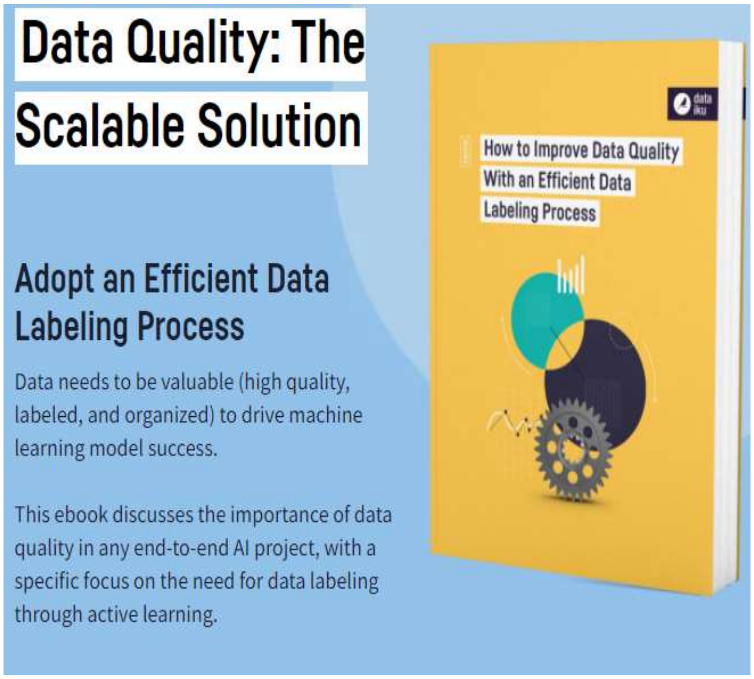


Experience shows that …

80% of the cost of building and maintaining a Data Warehouse Environment usually relates to the Populating Subsystem ...

The same holds for AI projects, read:
https://pages.dataiku.com/white-paper-how-to-improve-data-quality-with-labeling

This is a **well-known fact** coming from the practice. The reason lies in the high efforts for **data cleansing** and **data enrichment** (i.e. data quality tasks). The data quality processes are very complex and therefore we need this effort to finish this work. The same is true for AI projects, see: https://pages.dataiku.com/white-paper-how-to-improve-data-quality-with-labeling
Siehe auch [Moodle-DHBW] (…./Supp. Info…/Category3)



Data Quality: The Scalable Solution

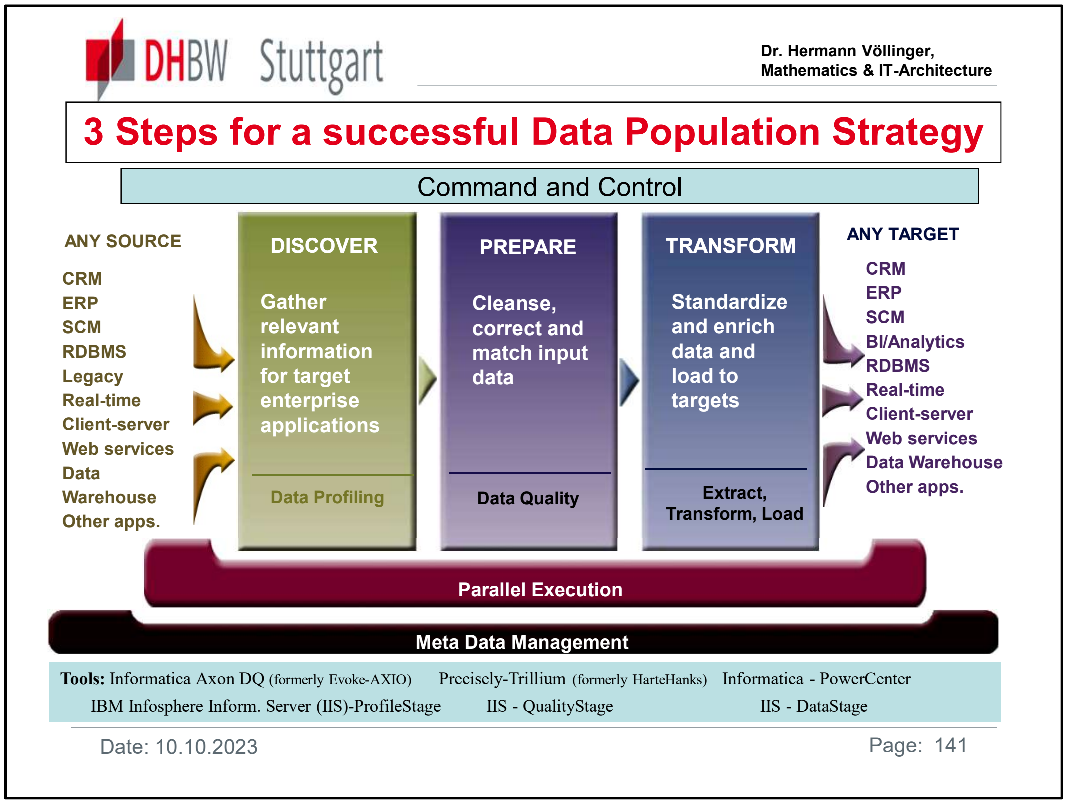
Adopt an Efficient Data Labeling Process

Data needs to be valuable (high quality, labeled, and organized) to drive machine learning model success.

This ebook discusses the importance of data quality in any end-to-end AI project, with a specific focus on the need for data labeling through active learning.

AI Projects Spend 80% of Their Time on Data Preparation Tasks

- The benefits of active learning, such as the ability to lower the number of label-related tasks and cost of data labeling necessary for a model to reach the required accuracy.
- Challenges associated with active learning and how AI tools and processes can help overcome them.
- Use cases of active learning at work and examples that support why labeled data is such a valuable asset.



3 Steps for a successful Data Population Strategy

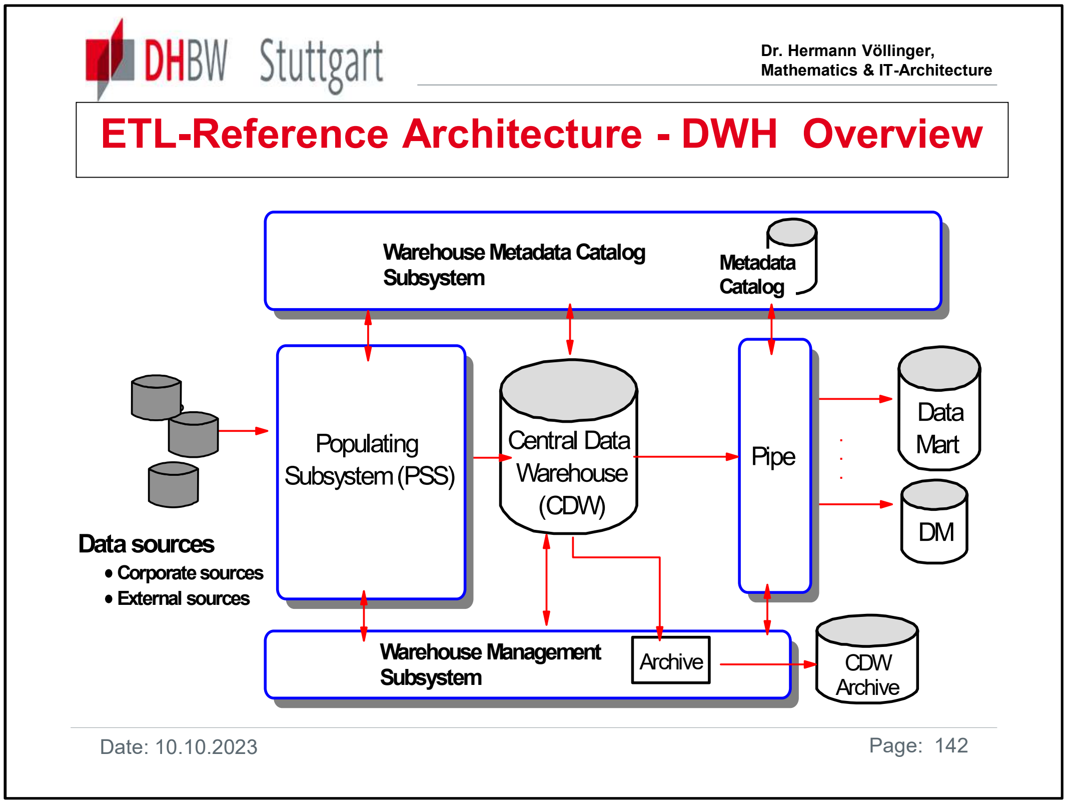
Tools: Informatica Axon DQ (formerly Evoke-AXIO) Precisely-Trillium (formerly HarteHanks) Informatica - PowerCenter
IBM Infosphere Inform. Server (IIS)-ProfileStage IIS - QualityStage IIS - DataStage

The process of data population is not only the TRANSFORMATION part, but also the steps of DISCOVER and PREPARE:
DISCOVER: make an analysis of all data sources, i.e. run a data profiling process, which will find for example inconsistencies and also bad data quality.
PREPARE: cleanse the data and correct the bad or corrupted data structures.

For more Information about profiling and data quality tools see the **Gartner Quadrant for Data Quality tools**: Kurs DW 21E

The three leading Tools in the quadrant are:

1. Informatica: “Informatica (cloud) data quality (IDQ)” & “Informatica Axon DQ (formerly known as Similarity Systems Evoke AXIO)” and other.
2. IBM: Information Server DQ; Infosphere QualityStage & Watson Knowledge Catalogue (former: Information Governance Catalogue)
3. SAP: SAP Information Stewart & SAP Data Intelligence and other.



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

ETL-Reference Architecture - DWH Overview

Warehouse Metadata Catalog Subsystem

Metadata Catalog

Data sources
- Corporate sources
- External sources

Populating Subsystem (PSS)

Central Data Warehouse (CDW)

Pipe

Data Mart

DM

Warehouse Management Subsystem

Archive

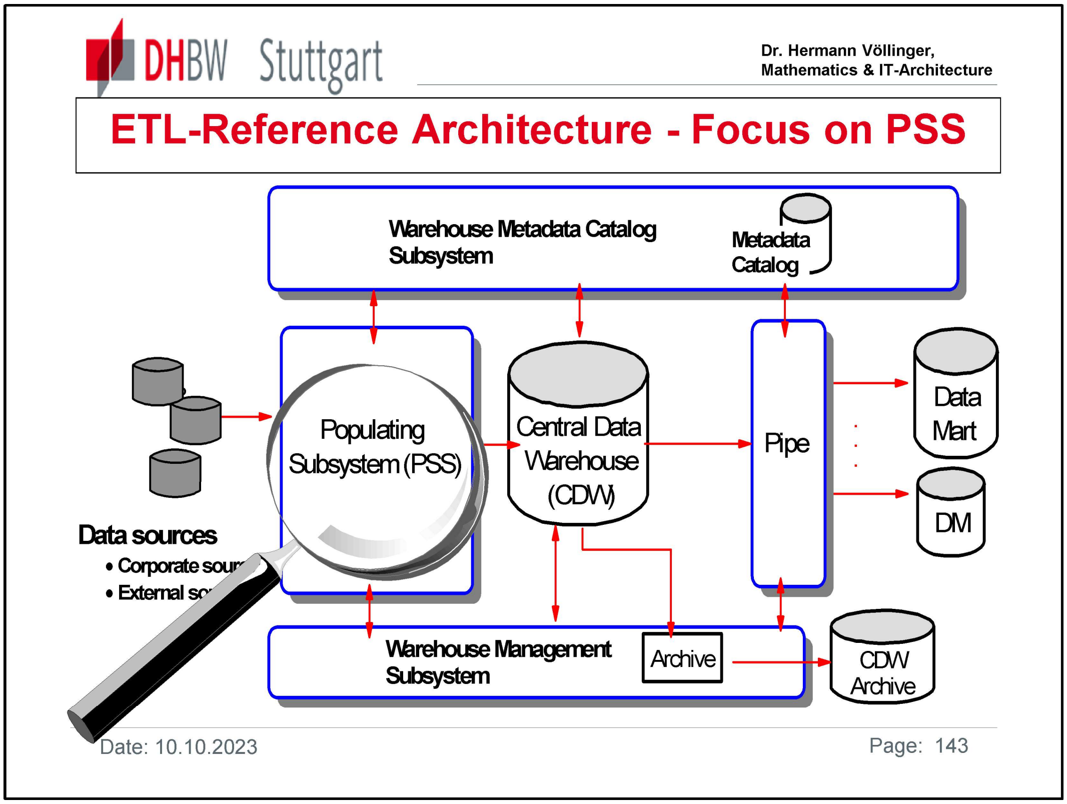
CDW Archive

Date: 10.10.2023 Page: 142

The slide show the major components of a DWH Reference Architecture:

1. Warehouse Metadata Layer
2. Warehouse Management Layer
3. Populating Subsystem -PSS
4. Central Data Warehouse -CDW
5. Pipe: CDW -→ DM (this is the ETL process from CDW to DM)
6. Data Marts - DM
7. CDW Archive

The ETL Reference Architecture describes in detail the PSS component. So we will also show in the next slides in more detail this component.



ETL-Reference Architecture - Focus on PSS

Warehouse Metadata Catalog Subsystem

Metadata Catalog

Populating Subsystem (PSS)

Central Data Warehouse (CDW)

Pipe

Data Mart

DM

Data sources
- Corporate sour
- External so

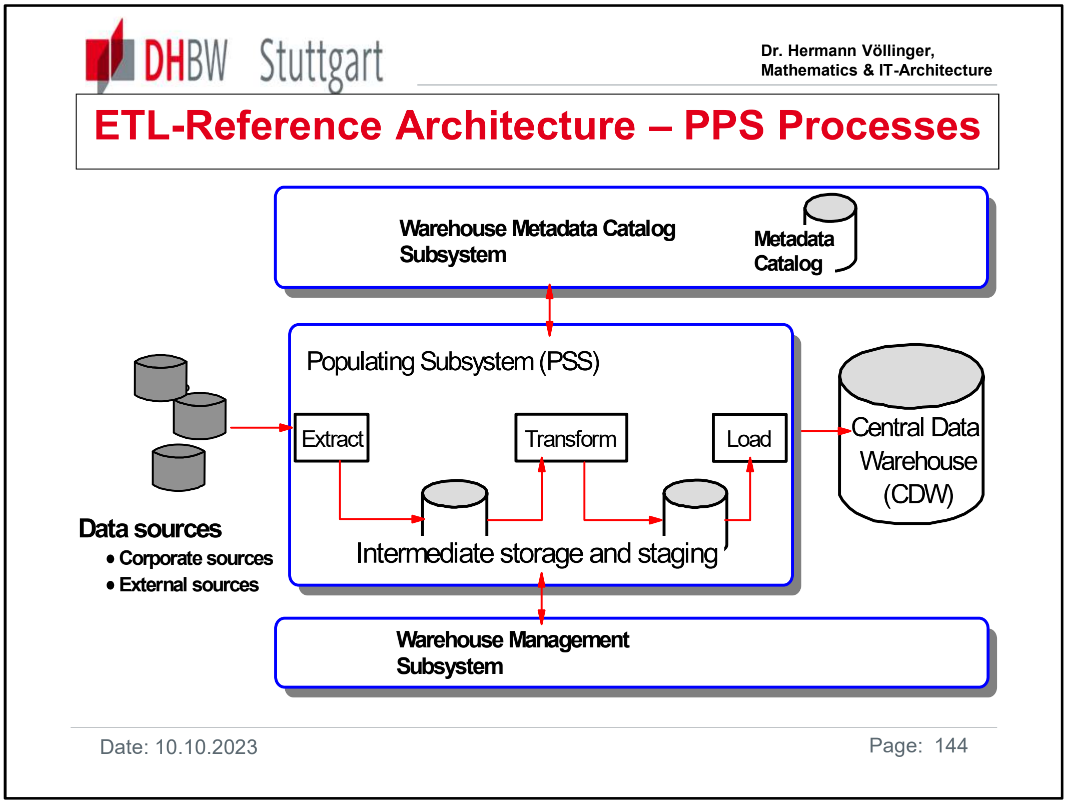
Warehouse Management Subsystem

Archive

CDW Archive

Consider now the PSS in more detail:
This is the ETL process from the source systems to the CDR.
Main Functions are:

- Extract data from sources (IMS, SAP, VSAM, ...)
- Filter the data for valid data records
- Repair data inconsistence
- Cleanse bad data
- Run plausibility Checks
- Transform
- Load the data into DBMS (use also DB functionality)



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

ETL-Reference Architecture – PPS Processes

Warehouse Metadata Catalog Subsystem

Metadata Catalog

Populating Subsystem (PSS)

Extract

Transform

Load

Intermediate storage and staging

Central Data Warehouse (CDW)

Data sources
- Corporate sources
- External sources

Warehouse Management Subsystem

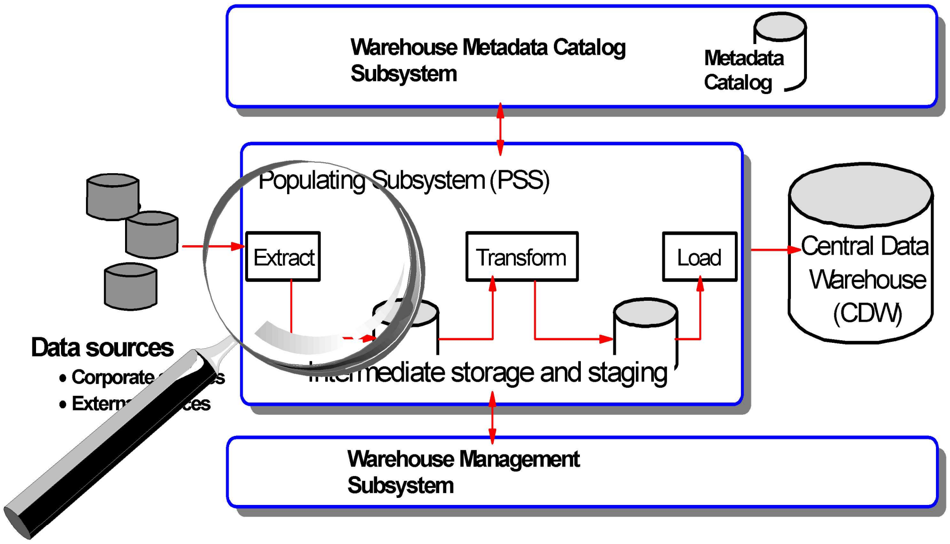
Date: 10.10.2023

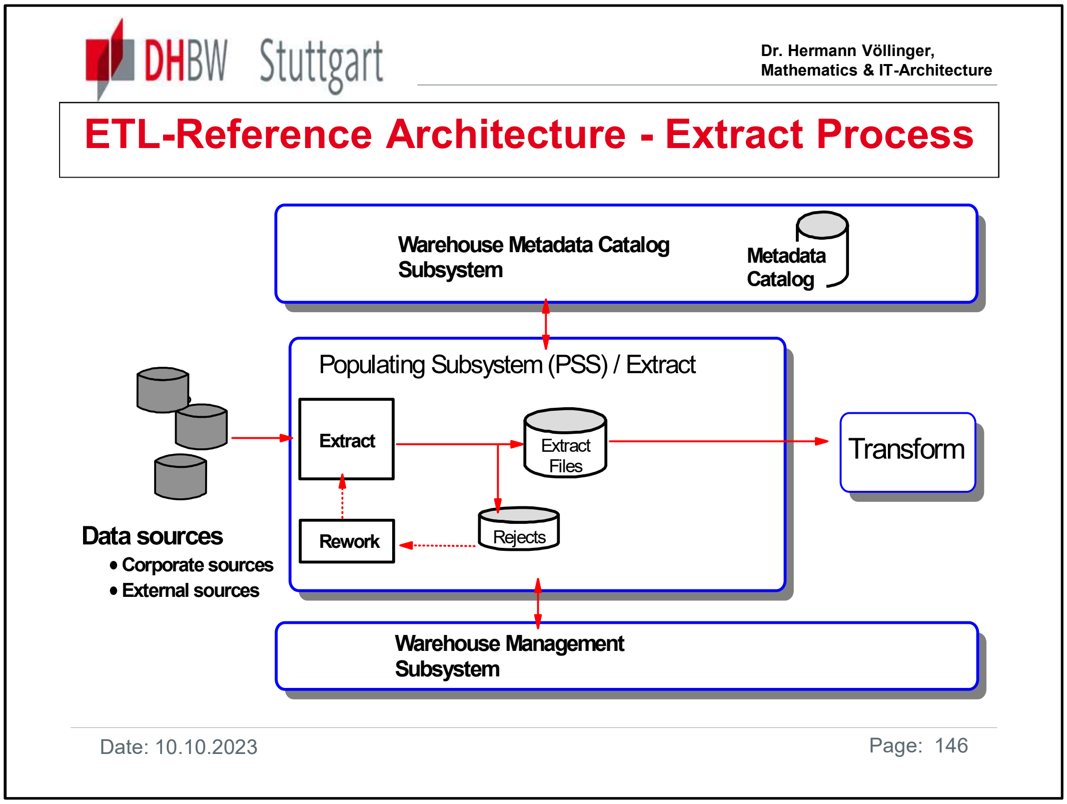
Page: 144

This is the first part of our reference architecture:

- **Extract**
  - Capture & Copy data from Operational Systems
  - Many different data structures, for example: hierarchical DBMS, Flat Files, Legacy Systems, ERP systems or other external sources
- **Transform**
  - Filter and Check (Plausibility)
  - Cleanse bad data
  - Manipulate operational data to conform with CDW standards
  - Build CDW rows
- **Load**
  - Fast load new data into CDW tables
  - Use for Load DBMS Functions if applicable

ETL-Reference Architecture – Extract Process





ETL-Reference Architecture - Extract Process

Warehouse Metadata Catalog Subsystem

Metadata Catalog

Populating Subsystem (PSS) / Extract

Extract

Extract Files

Rejects

Rework

Transform

Data sources
- Corporate sources
- External sources

Warehouse Management Subsystem

Extract services

<u>Making all required source data available for the populating subsystem</u>

- Full or partial snapshot of the data source
- Record of changes that occurred on the data source

<u>Building and managing Extract files (EF)</u>

- Static Capture (Snapshot) EF
- Incremental Capture EF

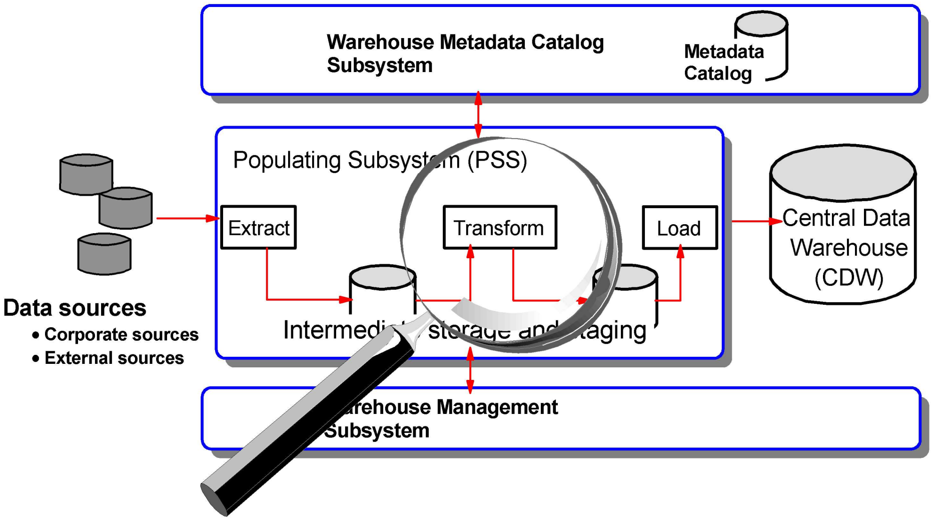
<u>Filtering of inappropriate data and records (rejects)</u>

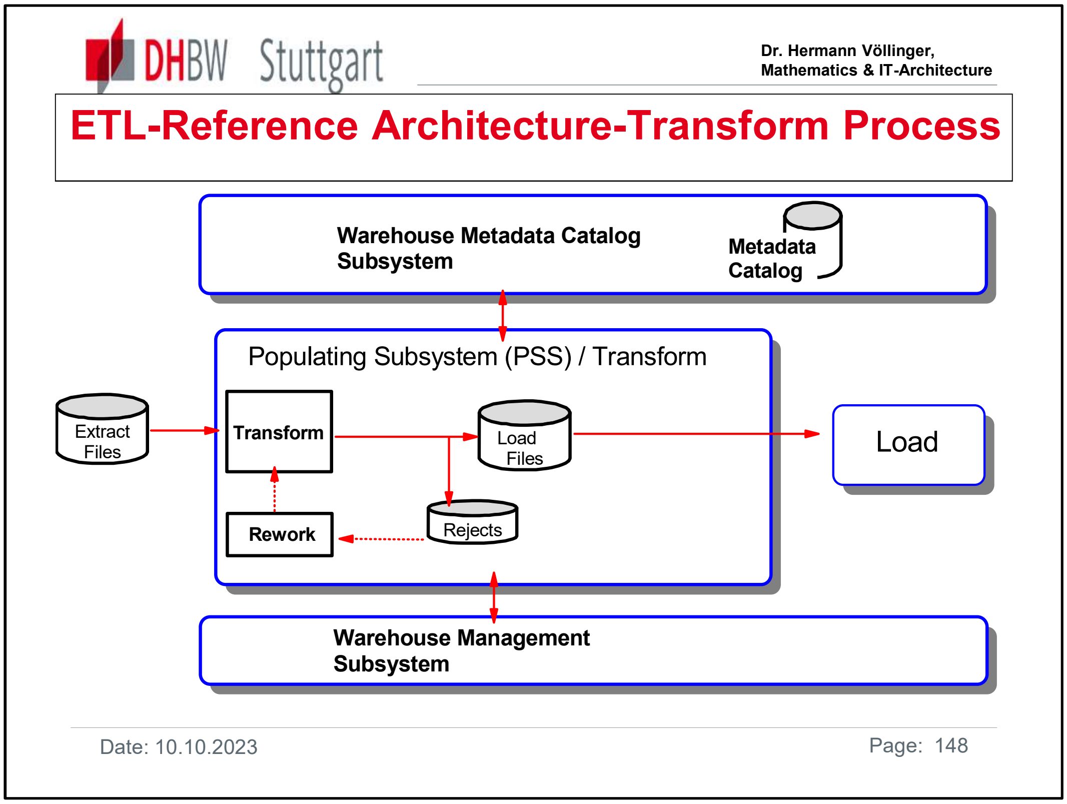
- Minimal
- Provide rework capabilities

<u>Provide Metadata</u>

- Statistics gathering

ETL-Reference Architecture–Transform Process





Dr. Hermann Völlinger,
Mathematics & IT-Architecture

ETL-Reference Architecture-Transform Process

Warehouse Metadata Catalog Subsystem

Metadata Catalog

Populating Subsystem (PSS) / Transform

Extract Files

Transform

Load Files

Rejects

Rework

Load

Warehouse Management Subsystem

Transform Component

<u>Transform Control Manager</u>

- Recognize "common format“ extract files and call appropriate transformation services

<u>Transformation services</u>

- Structural transformations
- Content transformations
- Functional transformations

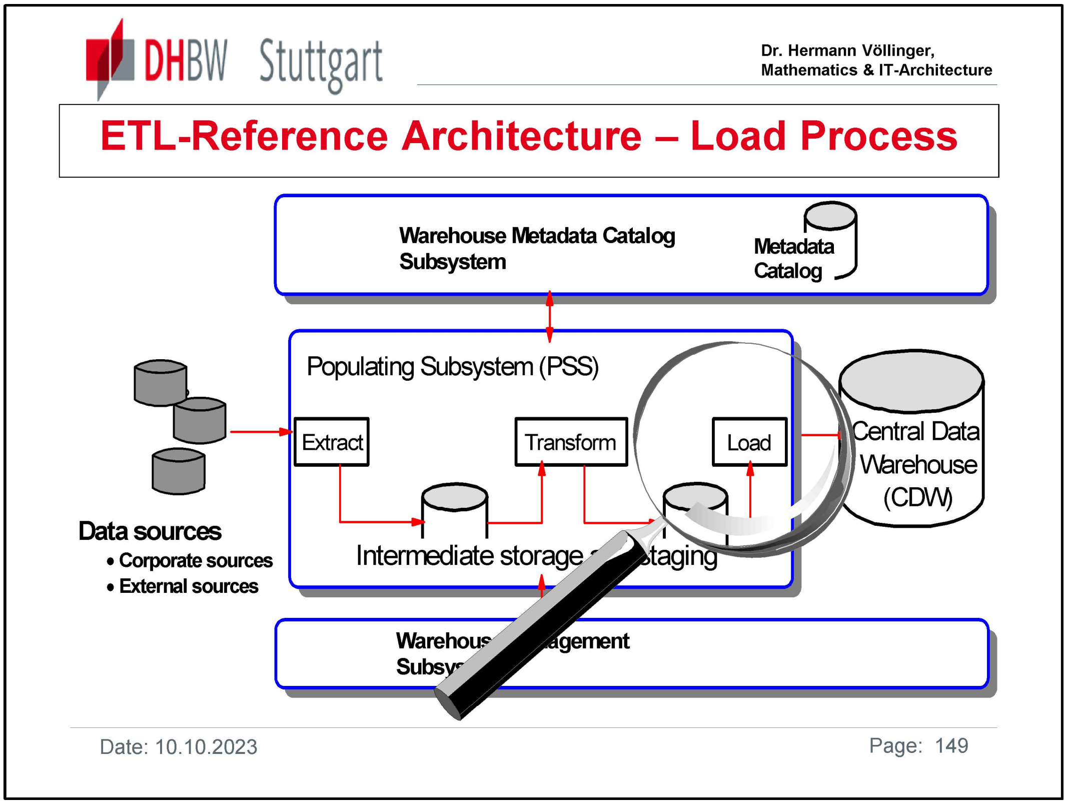
<u>Building and managing Load Files</u>

<u>Filtering of inappropriate records (rejects)</u>

- Provide rework capabilities

<u>Provide Metadata</u>

- Statistics gathering



ETL-Reference Architecture – Load Process

Warehouse Metadata Catalog Subsystem

Metadata Catalog

Populating Subsystem (PSS)

Extract

Transform

Load

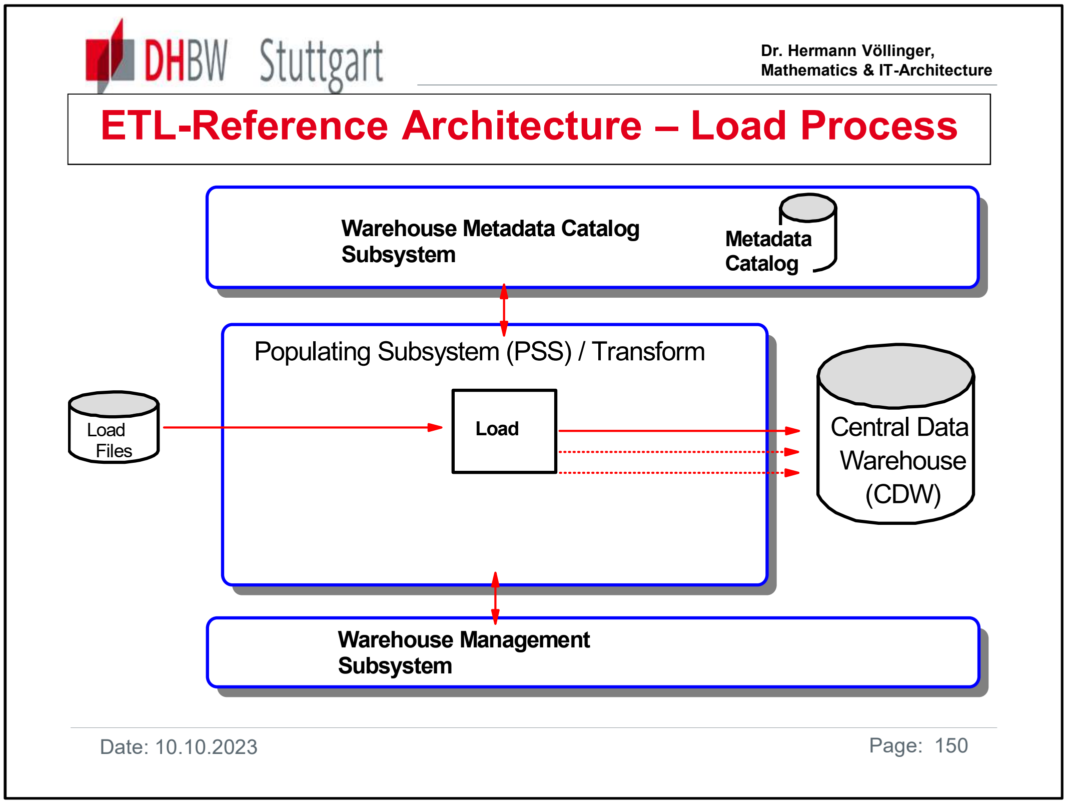
Intermediate storage staging

Central Data Warehouse (CDW)

Data sources

- Corporate sources
- External sources

Warehous agement
Subsys



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

ETL-Reference Architecture – Load Process

Warehouse Metadata Catalog Subsystem

Metadata Catalog

Populating Subsystem (PSS) / Transform

Load Files

Load

Central Data Warehouse (CDW)

Warehouse Management Subsystem

Date: 10.10.2023

Page: 150

Load Component

Load Control Manager

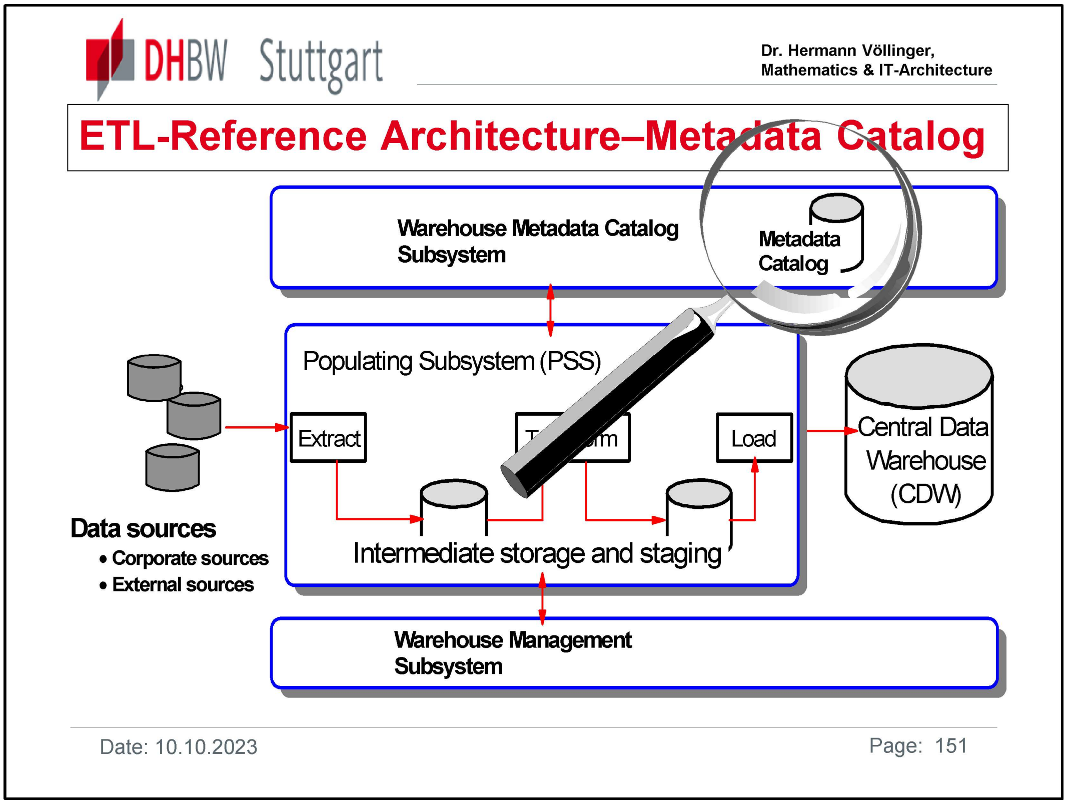
- Recognize load files and call appropriate load services
- Recognize status of the apply files and decide appropriate actions (start, restart prune, ...)
- Manage cascading effects of the load
- Manage locking/unlocking of the CDW
- Split to parallel load processes

Load services

- Read load files and load them into relevant parts of the CDW

Provide Metadata

- Statistics gathering



ETL-Reference Architecture–Metadata Catalog

Warehouse Metadata Catalog Subsystem

Metadata Catalog

Populating Subsystem (PSS)

Extract

Transform

Load

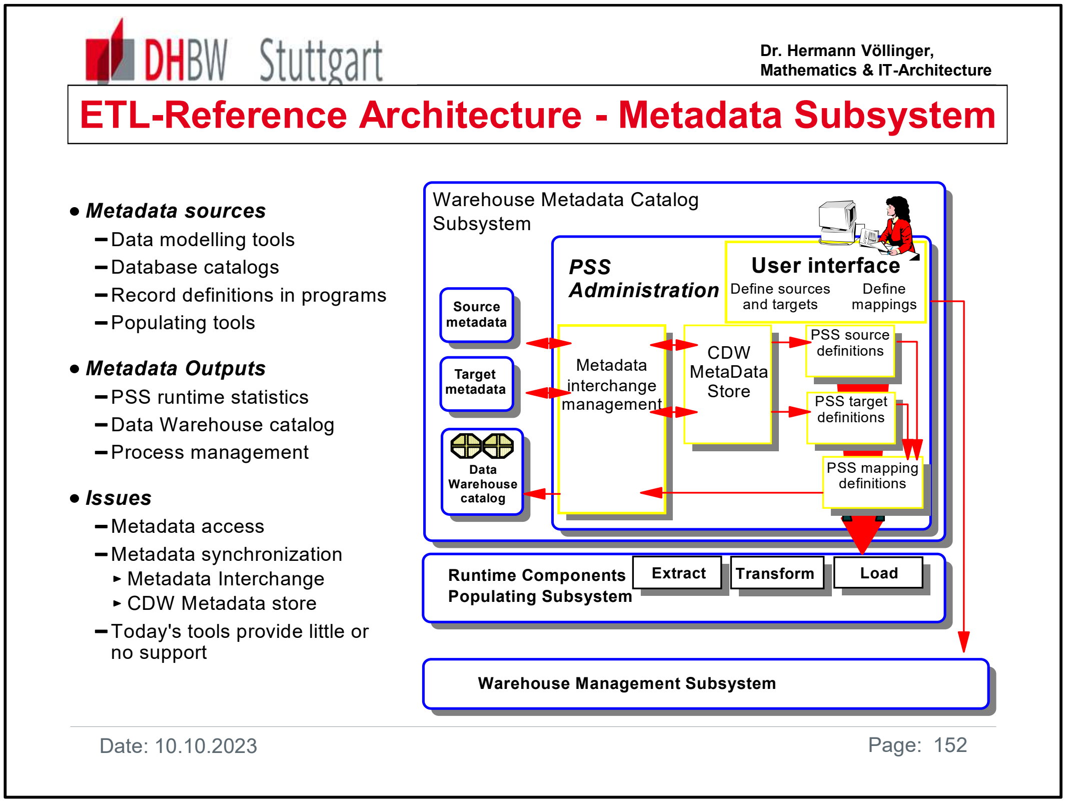
Intermediate storage and staging

Central Data Warehouse (CDW)

Data sources

- Corporate sources
- External sources

Warehouse Management Subsystem



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

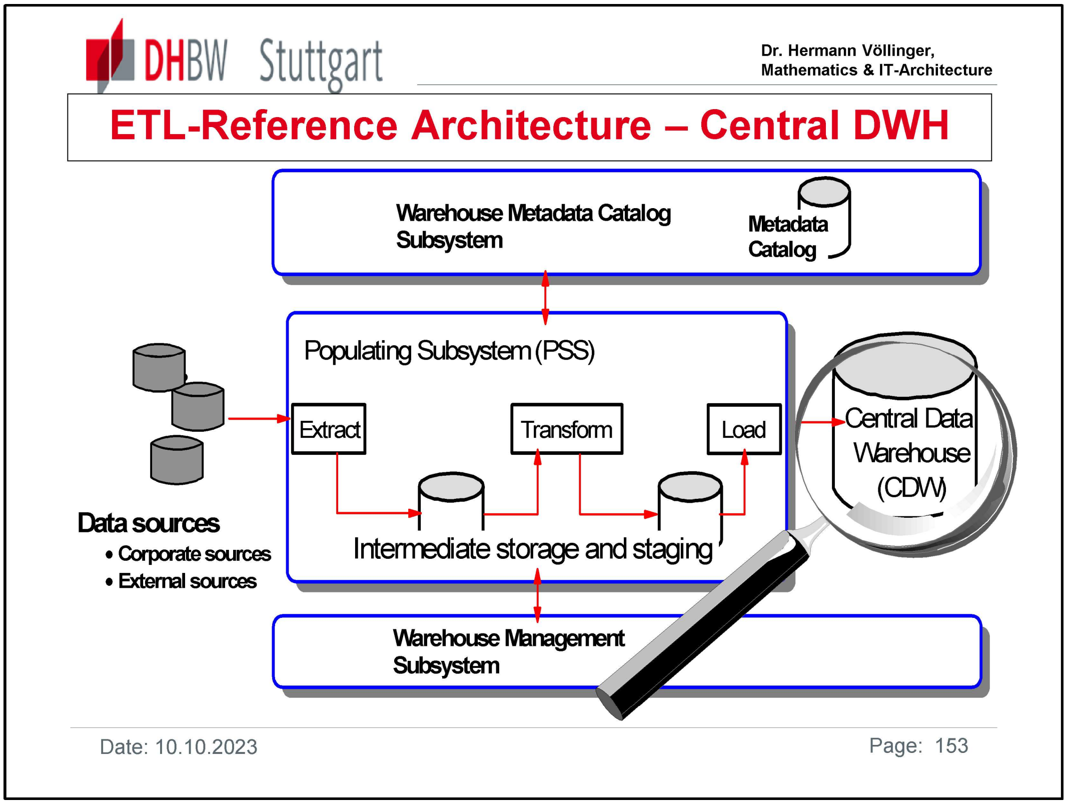
ETL-Reference Architecture - Metadata Subsystem

- ***Metadata sources***
  - Data modelling tools
  - Database catalogs
  - Record definitions in programs
  - Populating tools
- ***Metadata Outputs***
  - PSS runtime statistics
  - Data Warehouse catalog
  - Process management
- ***Issues***
  - Metadata access
  - Metadata synchronization
    - Metadata Interchange
    - CDW Metadata store
  - Today's tools provide little or no support

Date: 10.10.2023 Page: 152

This slide show the metadata layer, which is key to understand a good DWH strategy. When it comes to data warehousing, metadata plays a critical role in managing, organizing, and understanding the data stored in the warehouse.
We have to distinguish between **technical metadata** and *business metadata*. It‘s important that the knowledge in the metadata is published to the end users, such that this knowledge is not lost.
Key for a good metadata strategy is the usage of metadata standards. Several **metadata standards and frameworks** are relevant in the context of data warehousing:

- **Common Warehouse Metamodel (CWM):** CWM is an OMG (Object Management Group) standard, which is designed for metadata interchange between data warehousing and business intelligence tools. It provides a common way to describe and exchange metadata related to data warehousing, including data modeling, ETL (Extract, Transform, Load) processes, and OLAP (Online Analytical Processing) cubes. For more details see the book “CWM Specification.pdf” in Moodle/ Supporting Material../Category2
- **Metadata Modeling and Management (M3):** M3 is an open standard for managing and exchanging metadata in the context of data warehousing and business intelligence.
- **Information Management Metamodel (IMM):** IMM is an open standard from The Data Warehousing Institute (TDWI) that focuses on metadata in data warehousing environments. It defines a common framework for representing metadata about data structures, ETL processes, and reporting.
- etc.



ETL-Reference Architecture – Central DWH

Warehouse Metadata Catalog Subsystem

Metadata Catalog

Populating Subsystem (PSS)

Extract

Transform

Load

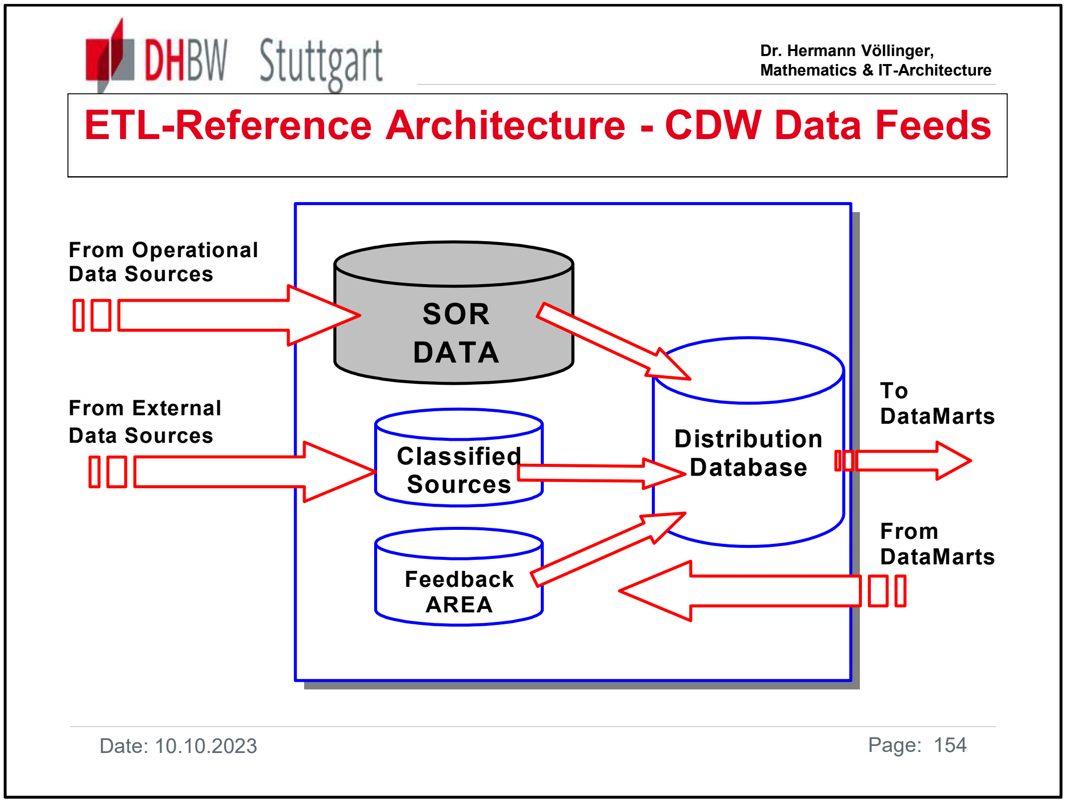
Intermediate storage and staging

Central Data Warehouse (CDW)

Data sources

- Corporate sources
- External sources

Warehouse Management Subsystem



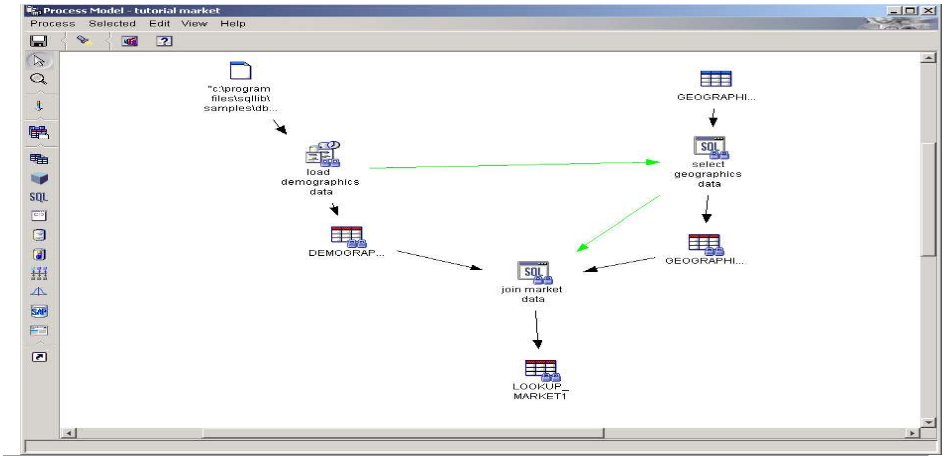
ETL-Reference Architecture - CDW Data Feeds

The slide shows the main data layers of a CDW:

- SoR Data, which get the data out of the operational systems
- External data sources
- Feedback area
- Distribution Database (sometimes also called ‘output layer‘)

Exercise1 to Lesson 6 - DB2 WH-Manager (Part1)

Exercise E6.1: Define the underlying SQL for the loading of Lookup_Market table:



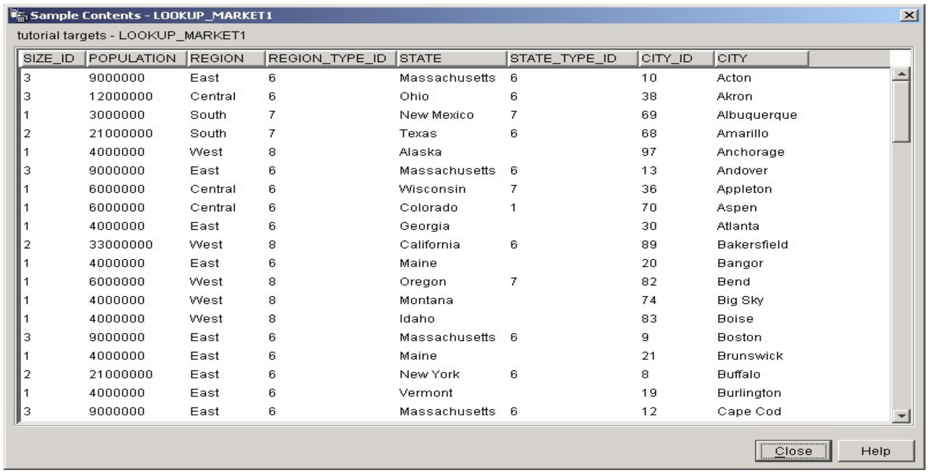
Join the demographics & geographics tables

Remark: DB2 Warehouse Manager uses SQL SELECT statement to extract data from warehouse sources
and generates an INSERT statement to insert the data into the warehouse target table.

For more details see on the homepage:
[HVö-2]: *Hermann Völlinger and Other: Exercises & Solutions of the Lecture 'Introduction to Data Warehousing';* DHBW Stuttgart; WS2023 http://www.dhbw-stuttgart.de/~hvoellin/

Exercise1 to Lesson 6 - DB2 WH-Manager (Part2)

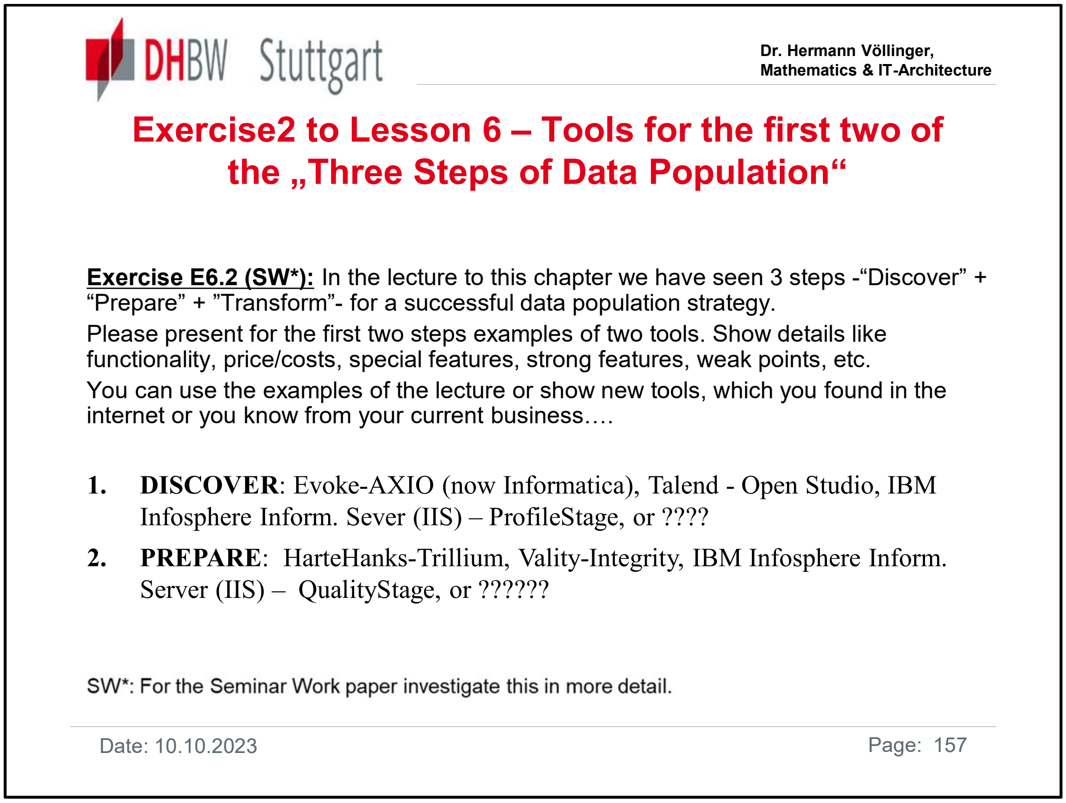
The structure of the target table Lookup_Market1 table can be seen in the following screenshot:



tutorial targets - LOOKUP_MARKET1

SIZE_ID	POPULATION	REGION	REGION_TYPE_ID	STATE	STATE_TYPE_ID	CITY_ID	CITY
3	9000000	East	6	Massachusetts	6	10	Acton
3	12000000	Central	6	Ohio	6	38	Akron
1	3000000	South	7	New Mexico	7	69	Albuquerque
2	21000000	South	7	Texas	6	68	Amarillo
1	4000000	West	8	Alaska		97	Anchorage
3	9000000	East	6	Massachusetts	6	13	Andover
1	6000000	Central	6	Wisconsin	7	36	Appleton
1	6000000	Central	6	Colorado	1	70	Aspen
1	4000000	East	6	Georgia		30	Atlanta
2	33000000	West	8	California	6	89	Bakersfield
1	4000000	East	6	Maine		20	Bangor
1	6000000	West	8	Oregon	7	82	Bend
1	4000000	West	8	Montana		74	Big Sky
1	4000000	West	8	Idaho		83	Boise
3	9000000	East	6	Massachusetts	6	9	Boston
1	4000000	East	6	Maine		21	Brunswick
2	21000000	East	6	New York	6	8	Buffalo
1	4000000	East	6	Vermont		19	Burlington
3	9000000	East	6	Massachusetts	6	12	Cape Cod

For more details see on the homepage:
[HVö-2]: *Hermann Völlinger and Other: Exercises & Solutions of the Lecture 'Introduction to Data Warehousing'*; DHBW Stuttgart; WS2023
http://www.dhbw-stuttgart.de/~hvoellin/



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise2 to Lesson 6 – Tools for the first two of the „Three Steps of Data Population“

Exercise E6.2 (SW*): In the lecture to this chapter we have seen 3 steps -“Discover” + “Prepare” + ”Transform”- for a successful data population strategy.

Please present for the first two steps examples of two tools. Show details like functionality, price/costs, special features, strong features, weak points, etc.

You can use the examples of the lecture or show new tools, which you found in the internet or you know from your current business….

1. **DISCOVER**: Evoke-AXIO (now Informatica), Talend - Open Studio, IBM Infosphere Inform. Sever (IIS) – ProfileStage, or ????
2. **PREPARE**: HarteHanks-Trillium, Vality-Integrity, IBM Infosphere Inform. Server (IIS) – QualityStage, or ??????

SW*: For the Seminar Work paper investigate this in more detail.

Date: 10.10.2023 Page: 157

For more Information about profiling and data quality tools see the **Gartner Quadrant for Data Quality tools**: Kurs DW 21E

For more information see in [DHBW-Moodle] the following documents:

1. Gartner-Market Guide for Data Preparation Tools.pdf
2. Precicsly_Data-Maturity+Data-Integration.pdf

Extract of the second document:

Competitive advantage through data maturity: Data Maturity is about how companies manage their data to answer key questions that have a significant impact on their bottom line, for example, to improve marketing efficiency, reduce operating costs, increase revenue, optimize the supply chain, and improve product offerings and customer service.

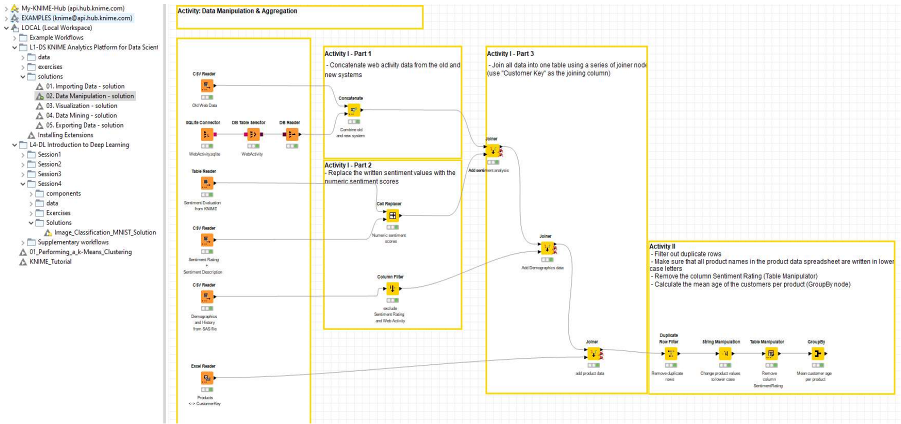
Data maturity phases are a way to measure a company's digital maturity in terms of data management and data analytics. Data Maturity provides insight into where data resides, how it is managed, its quality and the types of questions it answers. **The six data maturity phases** below explain how companies can enable effective decision-making in each phase and gain critical competitive advantage through a centralized business view:

Phase 1: Manual data collection and management
Phase 2: Automated data collection, integration and management
Phase 3: Data quality
Phase 4: Master data consolidation
Phase 5: Centralized overall view of the enterprise
Phase6: Achieving results: Data analysis to improve decision making

Exercise 3 to Lesson 6 - Data Manipulation & Aggregation in the KNIME Platform

Exercise E6.3: **Data Manipulation and Aggregation using KNIME Platform**

Homework for 2 Persons: Rebuild the KNIME Workflow (use given solution) for Data Manipulation & Aggregation and give technical explanations to the solution steps (see image):


My-KNIME-Hub (api.hub.knime.com)
EXAMPLES (knime@api.hub.knime.com)
LOCAL (Local Workspace)
Example Workflows
L1-DS KNIME Analytics Platform for Data Scient
data
exercises
solutions
01. Importing Data - solution
02. Data Manipulation - solution
03. Visualization - solution
04. Data Mining - solution
05. Exporting Data - solution
Installing Extensions
L4-DL Introduction to Deep Learning
Session1
Session2
Session3
Session4
components
data
Exercises
Solutions
Image_Classification_MNIST_Solution
Supplementary workflows
01_Performing_a_k-Means_Clustering
KNIME_Tutorial
CSV Reader
Old Web Data
SQLite Connector
WebActivity.sqlite
DB Table Selector
WebActivity
DB Reader
Table Reader
Sentiment Evaluation from KNIME
CSV Reader
Sentiment Rating + Sentiment Description
CSV Reader
Demographics and History from SAS file
Excel Reader
Products <-> CustomerKey
Concatenate
Combine old and new system
Cell Replacer
Numeric sentiment scores
Column Filter
exclude Sentiment Rating and Web Activity
Joiner
Add sentiment analysis
Joiner
Add Demographics data
Joiner
add product data
Duplicate Row Filter
Remove duplicate rows
String Manipulation
Change product values to lower case
Table Manipulator
Remove column SentimentRating
GroupBy
Mean customer age per product

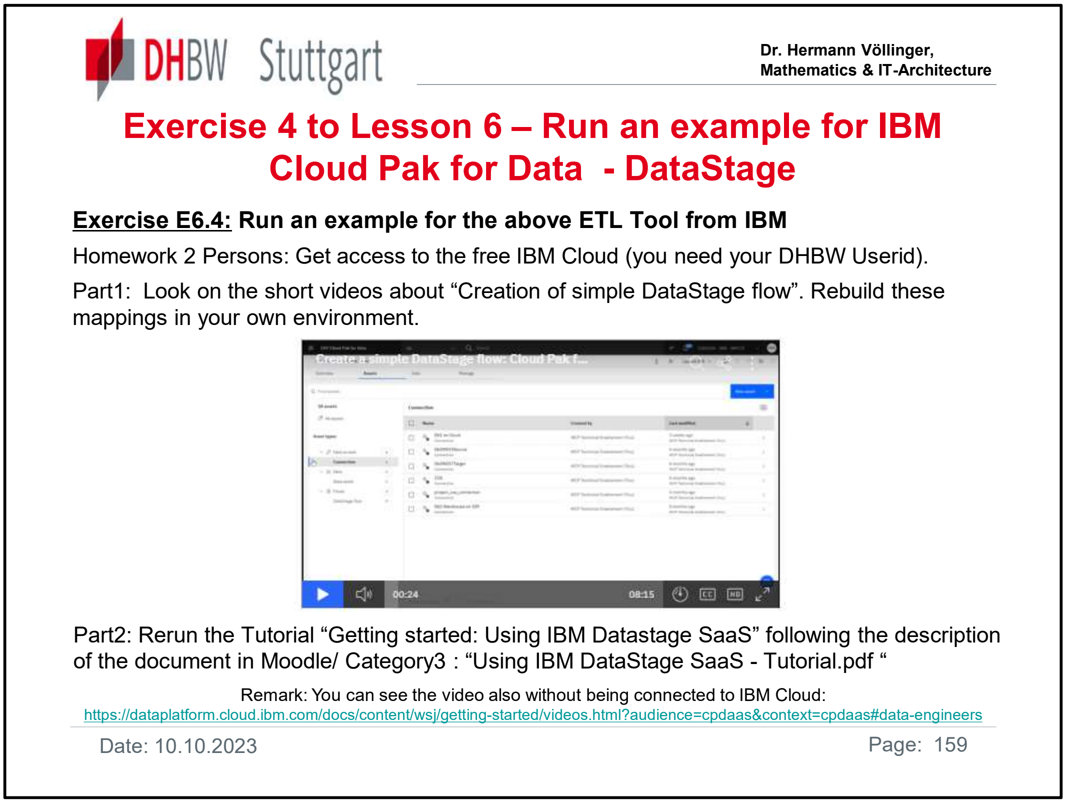
Activity: Data Manipulation & Aggregation

Activity I - Part 1
- Concatenate web activity data from the old and new systems

Activity I - Part 2
- Replace the written sentiment values with the numeric sentiment scores

Activity I - Part 3
- Join all data into one table using a series of joiner node (use "Customer Key" as the joining column)

Activity II
- Filter out duplicate rows
- Make sure that all product names in the product data spreadsheet are written in lower case letters
- Remove the column Sentiment Rating (Table Manipulator)
- Calculate the mean age of the customers per product (GroupBy node)



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise 4 to Lesson 6 – Run an example for IBM Cloud Pak for Data - DataStage

Exercise E6.4: Run an example for the above ETL Tool from IBM

Homework 2 Persons: Get access to the free IBM Cloud (you need your DHBW Userid).

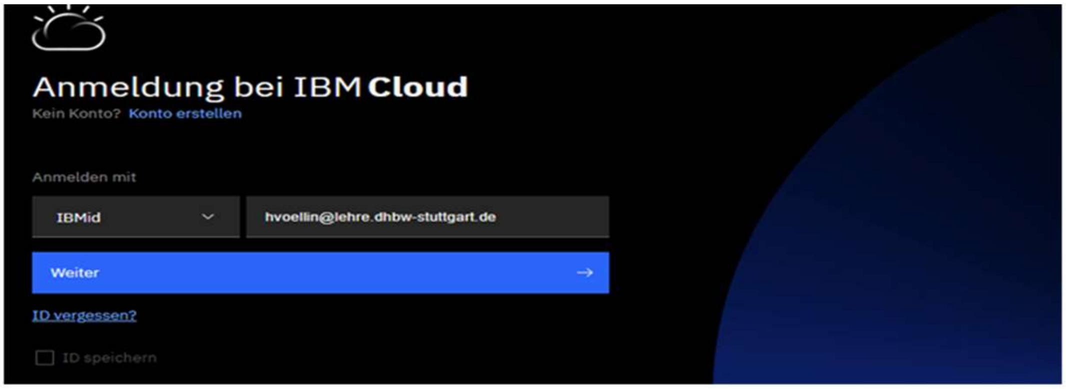
Part1: Look on the short videos about “Creation of simple DataStage flow”. Rebuild these mappings in your own environment.

Part2: Rerun the Tutorial “Getting started: Using IBM Datastage SaaS” following the description of the document in Moodle/ Category3 : “Using IBM DataStage SaaS - Tutorial.pdf “

Remark: You can see the video also without being connected to IBM Cloud:
https://dataplatform.cloud.ibm.com/docs/content/wsj/getting-started/videos.html?audience=cpdaas&context=cpdaas#data-engineers

Date: 10.10.2023 Page: 159

At a first step you need to create an account with your DHBW Userid (“academic program” to get a free “lite access” to the IBM Cloud platform:



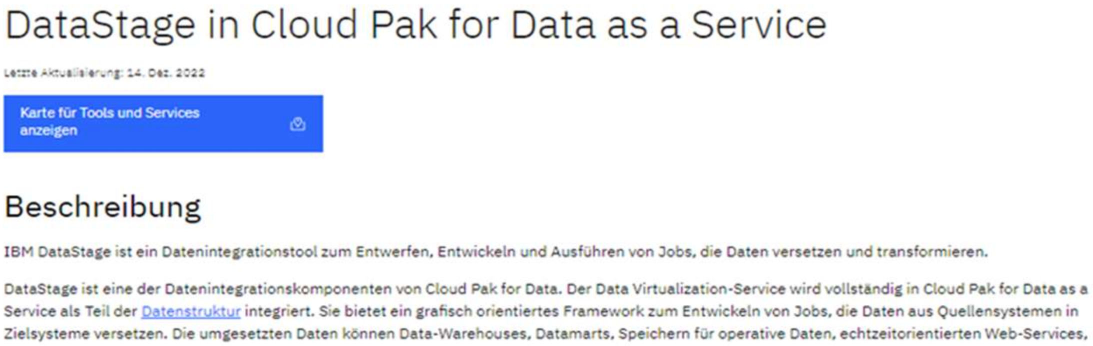
Anmeldung bei IBM **Cloud**

Kein Konto? Konto erstellen

Anmelden mit

IBMid hvoellin@lehre.dhbw-stuttgart.de

In a second step search for “Datastage” and then start the usage of the tool …..



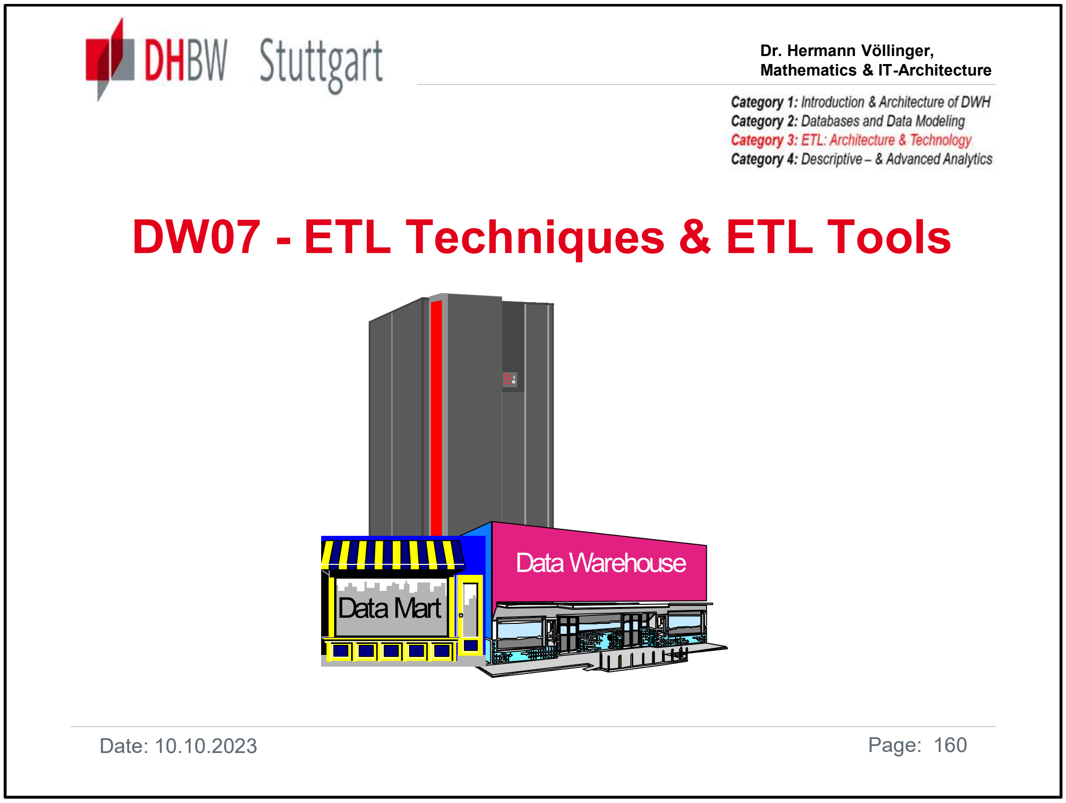
DataStage in Cloud Pak for Data as a Service

Letzte Aktualisierung: 14. Dez. 2022

Beschreibung

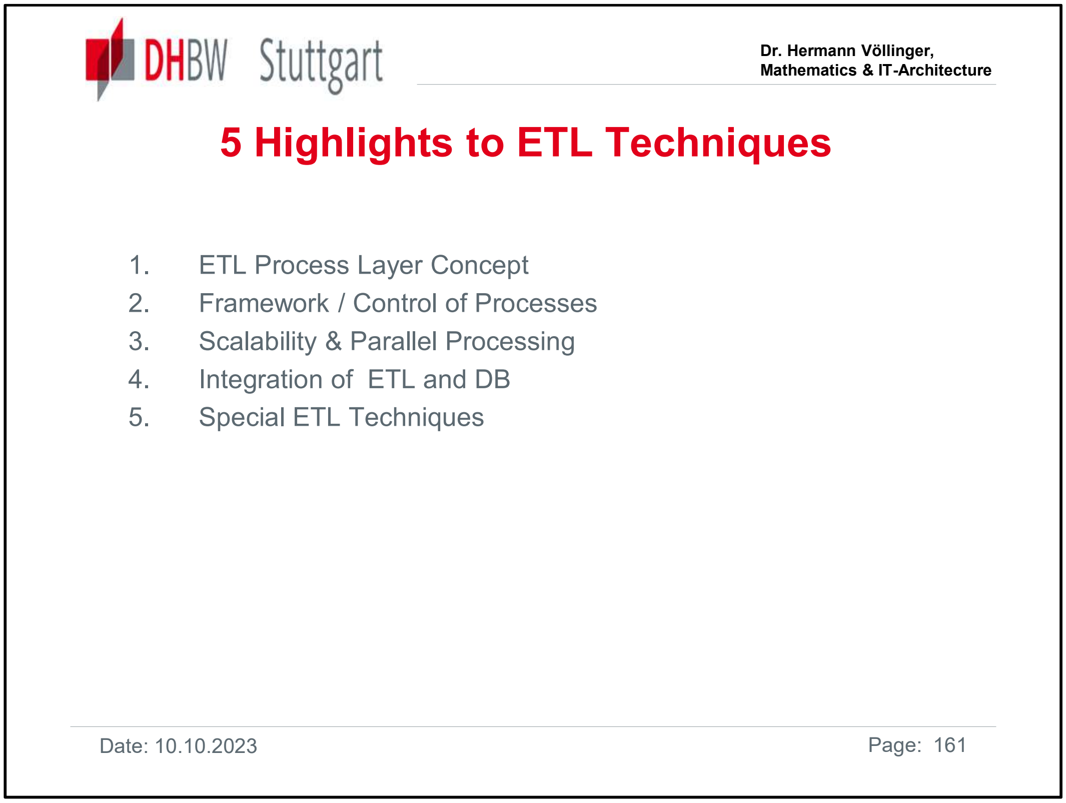
IBM DataStage ist ein Datenintegrationstool zum Entwerfen, Entwickeln und Ausführen von Jobs, die Daten versetzen und transformieren.

DataStage ist eine der Datenintegrationskomponenten von Cloud Pak for Data. Der Data Virtualization-Service wird vollständig in Cloud Pak for Data as a Service als Teil der Datenstruktur integriert. Sie bietet ein grafisch orientiertes Framework zum Entwickeln von Jobs, die Daten aus Quellensystemen in Zielsysteme versetzen. Die umgesetzten Daten können Data-Warehouses, Datamarts, Speichern für operative Daten, echtzeitorientierten Web-Services,



Category 1: *Introduction & Architecture of DWH*
Category 2: *Databases and Data Modeling*
Category 3: *ETL: Architecture & Technology*
Category 4: *Descriptive – & Advanced Analytics*

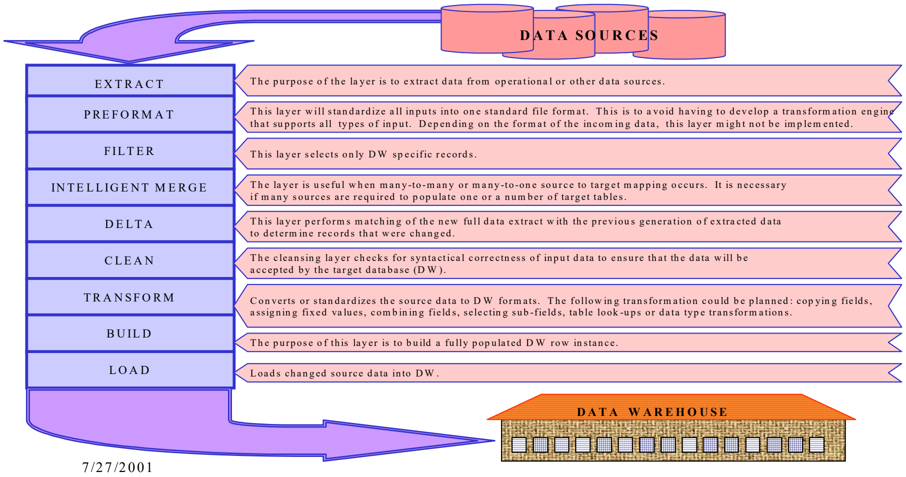
DW07 - ETL Techniques & ETL Tools



5 Highlights to ETL Techniques

1. ETL Process Layer Concept
2. Framework / Control of Processes
3. Scalability & Parallel Processing
4. Integration of ETL and DB
5. Special ETL Techniques

Generic ETL Process Layers



DATA SOURCES

EXTRACT

The purpose of the layer is to extract data from operational or other data sources.

PREFORMAT

This layer will standardize all inputs into one standard file format. This is to avoid having to develop a transformation engine that supports all types of input. Depending on the format of the incoming data, this layer might not be implemented.

FILTER

This layer selects only DW specific records.

INTELLIGENT MERGE

The layer is useful when many-to-many or many-to-one source to target mapping occurs. It is necessary if many sources are required to populate one or a number of target tables.

DELTA

This layer performs matching of the new full data extract with the previous generation of extracted data to determine records that were changed.

CLEAN

The cleansing layer checks for syntactical correctness of input data to ensure that the data will be accepted by the target database (DW).

TRANSFORM

Converts or standardizes the source data to DW formats. The following transformation could be planned: copying fields, assigning fixed values, combining fields, selecting sub-fields, table look-ups or data type transformations.

BUILD

The purpose of this layer is to build a fully populated DW row instance.

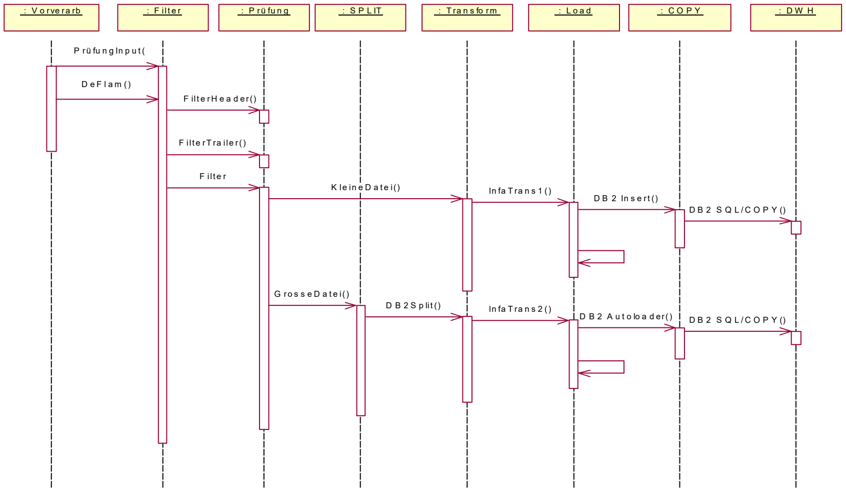
LOAD

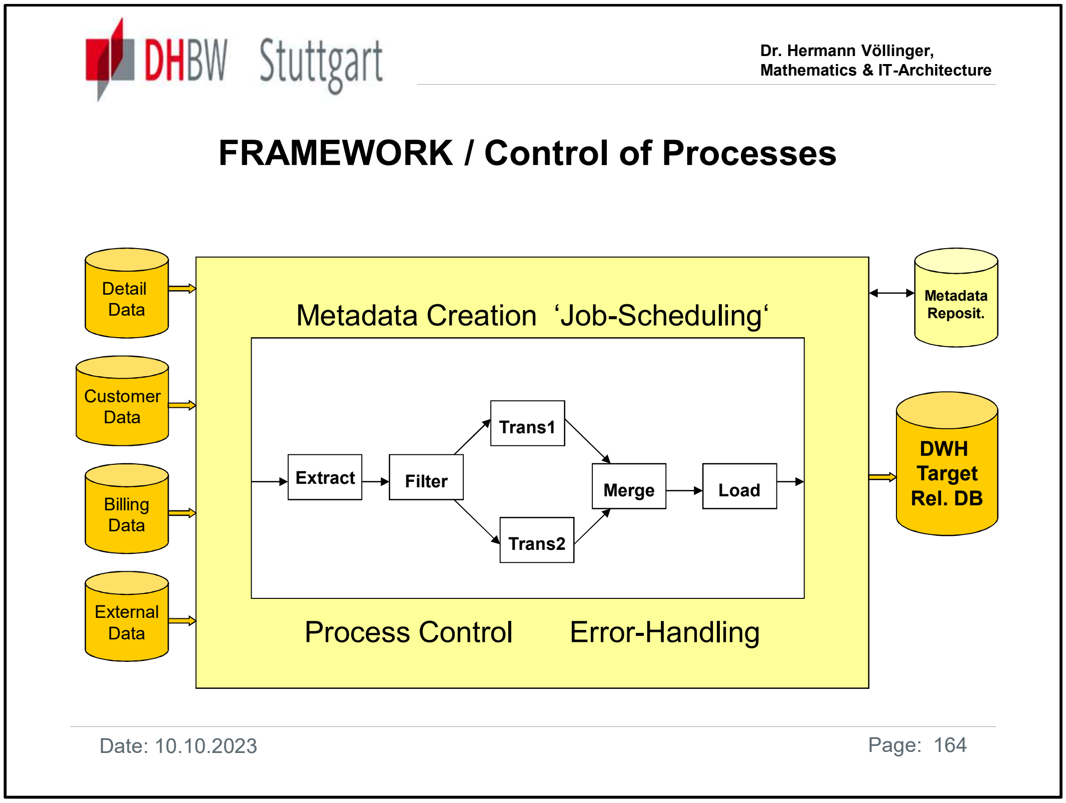
Loads changed source data into DW.

DATA WAREHOUSE

7/27/2001

ETL Layer Concept (Example)

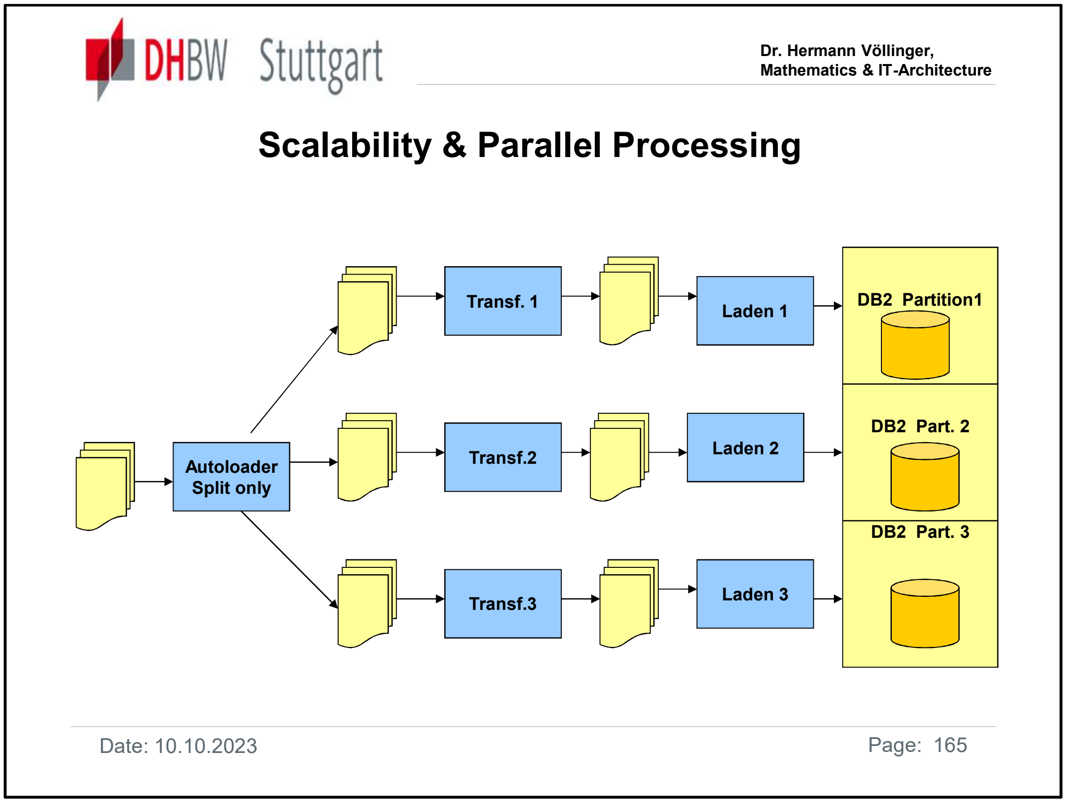




FRAMEWORK / Control of Processes

The whole ETL process will be controlled by the Framework including the following functions:

- Initialization of ETL process (check completeness of data sources)
- Start/Stop of the ETL process
- Job-Scheduling (using triggers)
- Definition and control of parallel ETL processes
- Control all processes (through process parameters out of metadata)
- Error handling and protocol errors as metadata
- Creation of technical metadata (protocol and ETL statistics)
- Control the interface to the database system (DB2)
- Delete temporary files



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

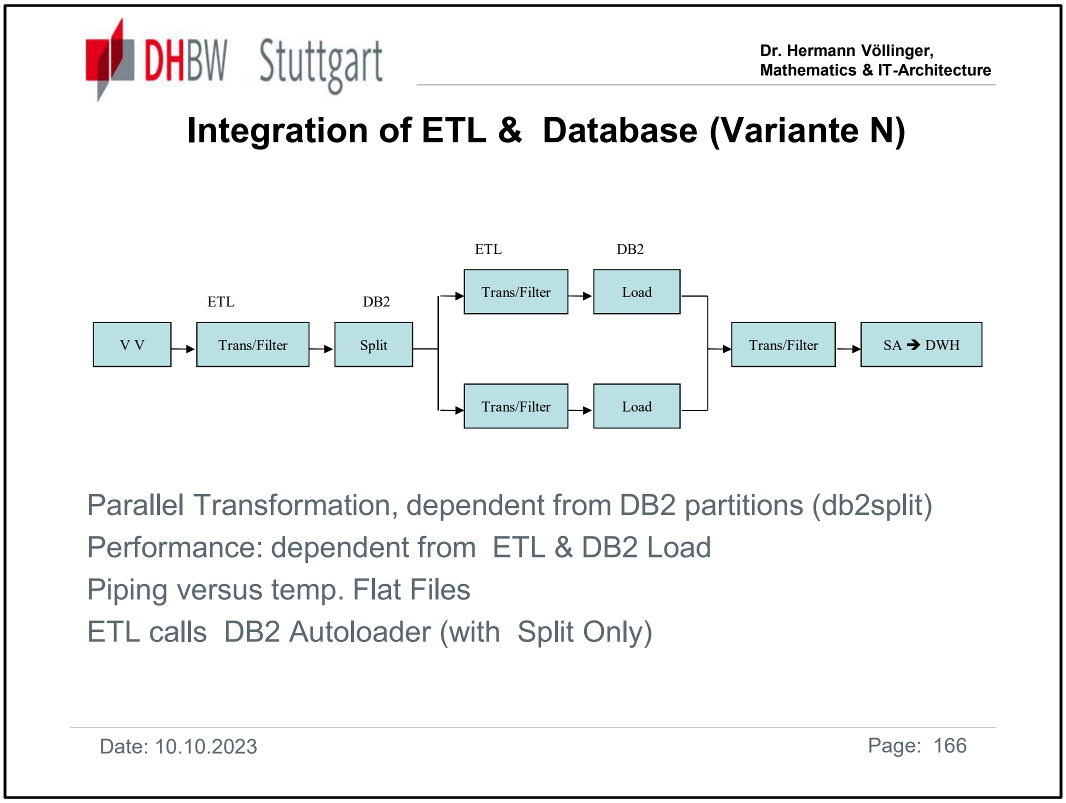
Scalability & Parallel Processing

Date: 10.10.2023

Page: 165

The slide shows an example for parallel processing:
We see the following main steps:

1. Split of the input data in our example with DB2 Autoloader (SPLIT ONLY)
2. Distribution on parallel jobs (dependent on number of CPU‘s)
3. Parallel Transformation
4. Parallel Load into DB2 partitions



Integration of ETL & Database (Variante N)

ETL

DB2

ETL

DB2

V V

Trans/Filter

Split

Trans/Filter

Load

Trans/Filter

Load

Trans/Filter

SA → DWH

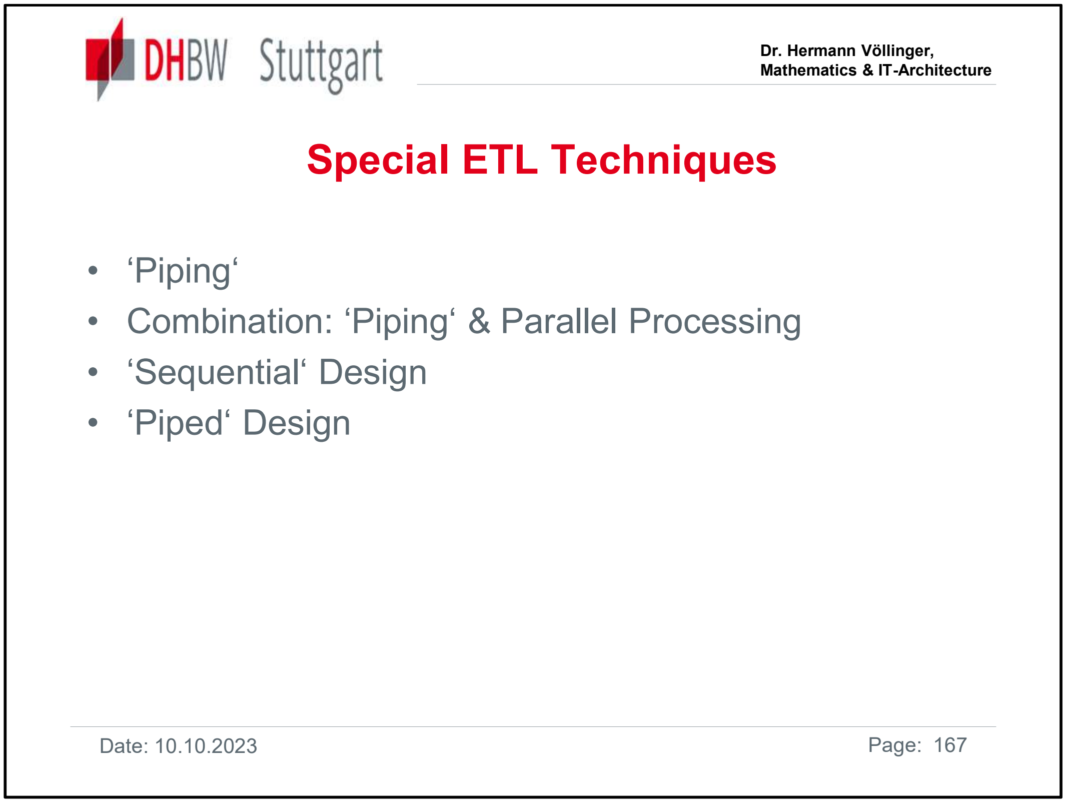
Parallel Transformation, dependent from DB2 partitions (db2split)

Performance: dependent from ETL & DB2 Load

Piping versus temp. Flat Files

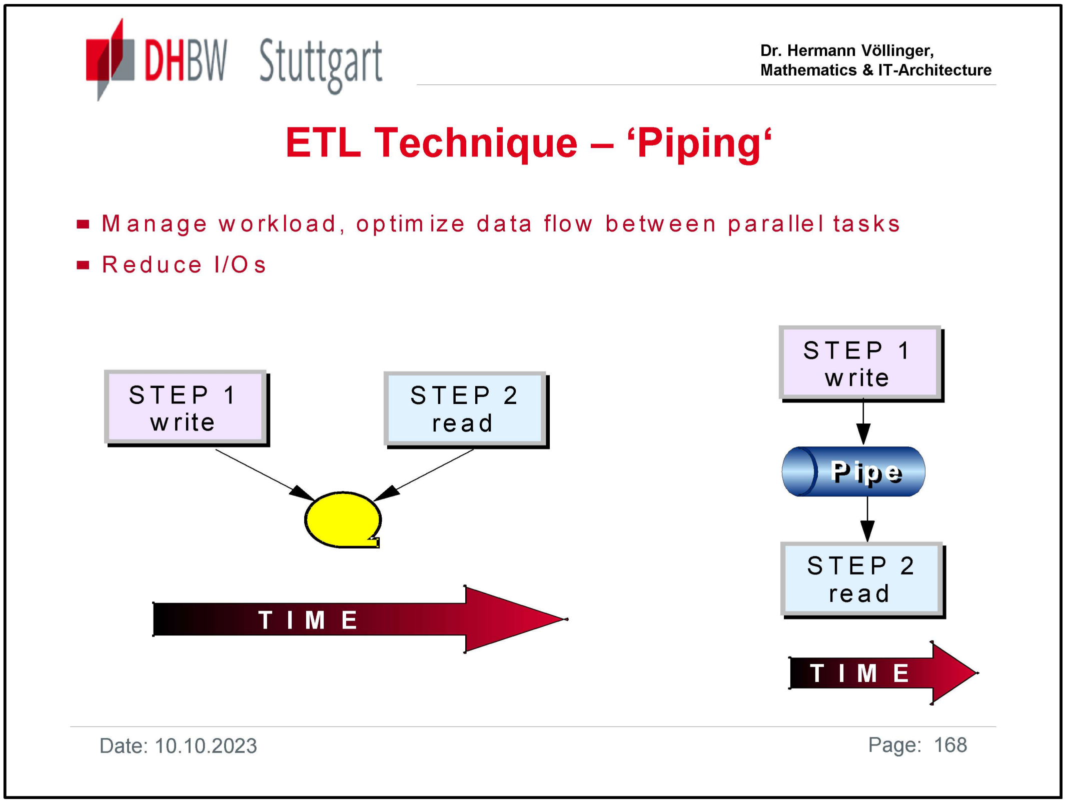
ETL calls DB2 Autoloader (with Split Only)

We see one possible integration scenario.



Special ETL Techniques

- ‘Piping‘
- Combination: ‘Piping‘ & Parallel Processing
- ‘Sequential‘ Design
- ‘Piped‘ Design



ETL Technique – ‘Piping‘

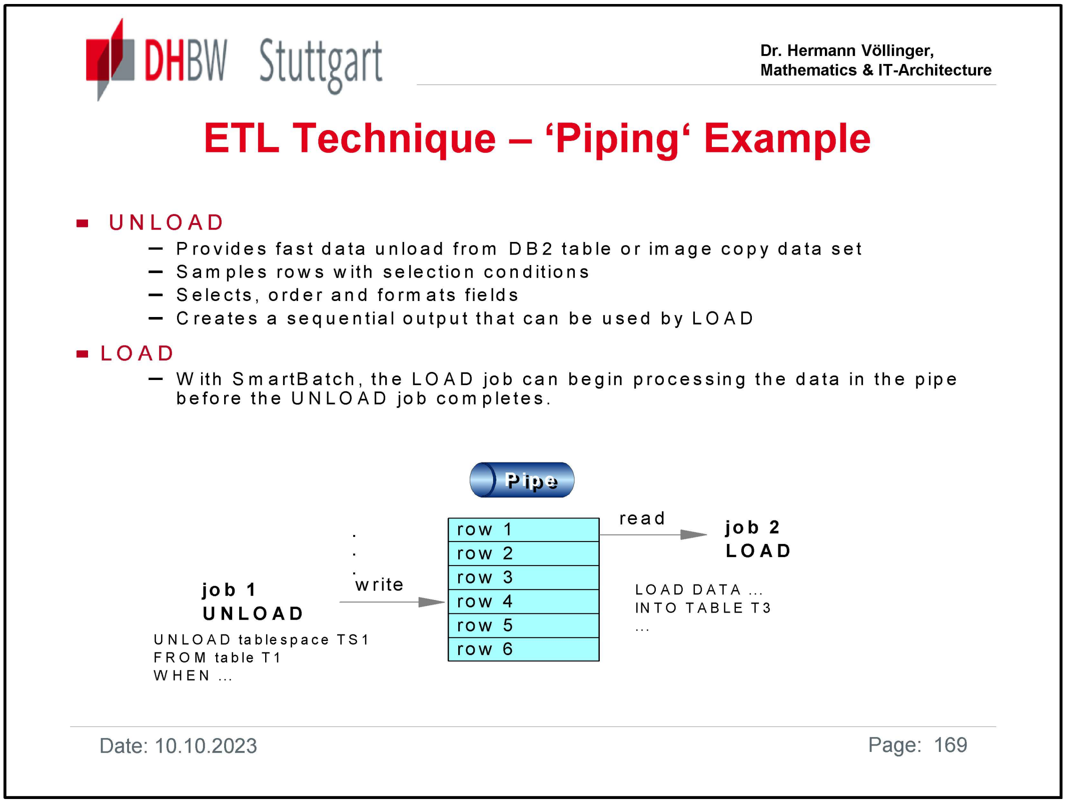
- Manage workload, optimize data flow between parallel tasks
- Reduce I/Os

STEP 1
write
STEP 2
read
TIME
STEP 1
write
Pipe
STEP 2
read
TIME

In this technique, we see a **piped execution** for running batch jobs:
Pipes increase parallelism and provide data in memory. They allow job steps to overlap and eliminate the I/Os incurred in passing data sets between them. Additionally, if a tape data set is replaced by a pipe, tape mounts can be eliminated.

The left part of our figure shows a traditional batch job, where job step 1 writes data to a sequential data set. When step 1 has completed, step 2 starts to read the data set from the beginning.
The right part of our figure shows the use of a batch pipe. Step 1 writes to a batch pipe which is written to memory and not to disk. The pipe processes a block of record at a time. Step 2 starts reading data from the pipe as soon as step 1 has written the first block of data. Step 2 does not need to wait until step 1 has finished. When step 2 has read the data from pipe, the data is removed from pipe.

Using this approach, the total elapsed time of the two jobs can be reduced and the need for temporary disk or tape storage is removed.



ETL Technique – ‘Piping‘ Example

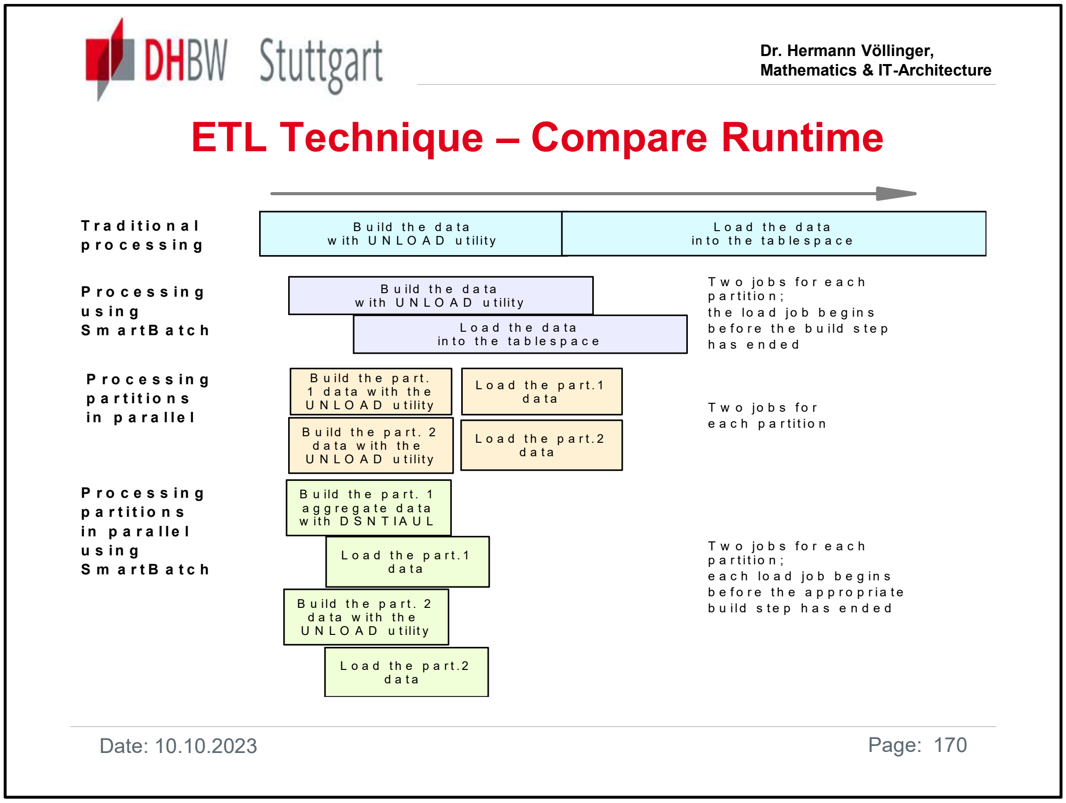
- UNLOAD
  - Provides fast data unload from DB2 table or image copy data set
  - Samples rows with selection conditions
  - Selects, order and formats fields
  - Creates a sequential output that can be used by LOAD
- LOAD
  - With SmartBatch, the LOAD job can begin processing the data in the pipe before the UNLOAD job completes.

row 1
row 2
row 3
row 4
row 5
row 6

```
UNLOAD tablespace TS1
FROM table T1
WHEN ...
```

```
LOAD DATA ...
INTO TABLE T3
...
```

Here we show how to **parallelize the tasks** of data extraction and loading data from a central data warehouse to a data mart.
The data is extracted from the data warehouse using the UNLOAD utility of the database. The UNLOAD writes the unloaded records to a batch pipe. As soon as one block of records has been written to the pipe database LOAD utility can start to load data into the target table in the data mart. At the same time, the unload utility continues to unload the rest of the data from the source tables.



ETL Technique – Compare Runtime

Traditional processing

Build the data with UNLOAD utility

Load the data into the tablespace

Processing using SmartBatch

Build the data with UNLOAD utility

Load the data into the tablespace

Two jobs for each partition; the load job begins before the build step has ended

Processing partitions in parallel

Build the part. 1 data with the UNLOAD utility

Load the part.1 data

Build the part. 2 data with the UNLOAD utility

Load the part.2 data

Two jobs for each partition

Processing partitions in parallel using SmartBatch

Build the part. 1 aggregate data with DSNTIAUL

Load the part.1 data

Build the part. 2 data with the UNLOAD utility

Load the part.2 data

Two jobs for each partition; each load job begins before the appropriate build step has ended

Compare Runtime for combined parallel and piped executions:

Based on the characteristics of the population process, parallel and piped processing can be combined.

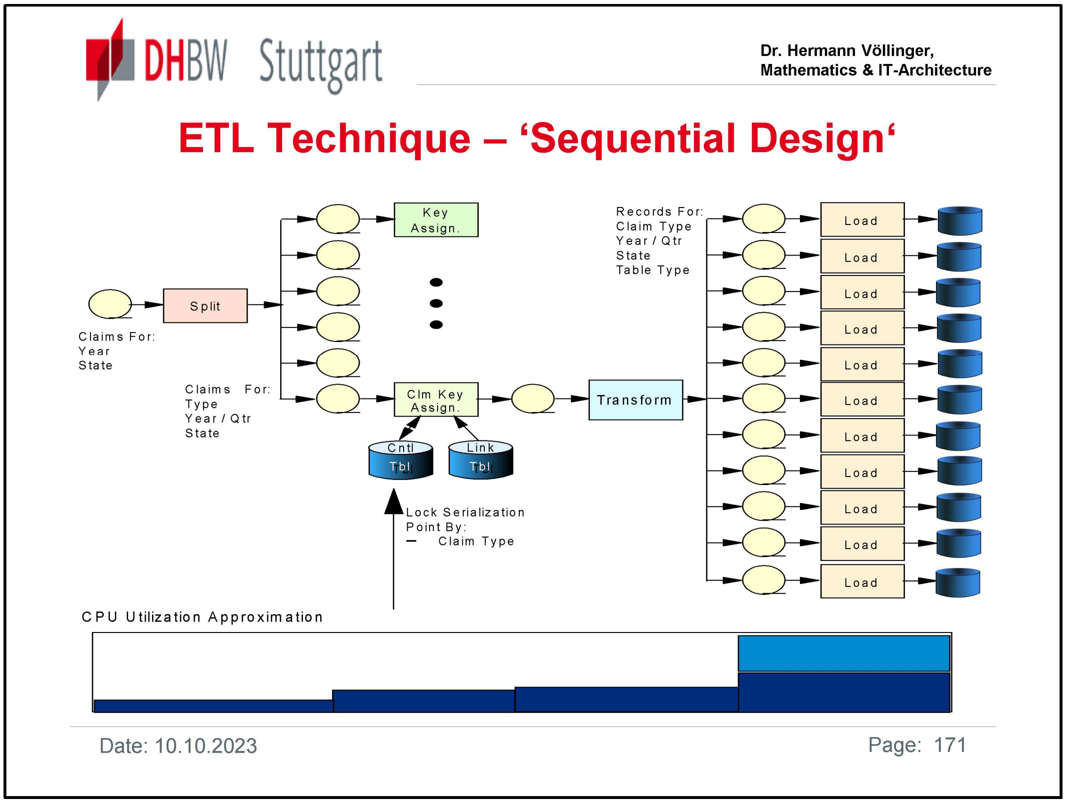
The first scenario shows the traditional solution. The first job step creates the extraction file and the second job step loads the extracted data into the target DB2 table.

In the **second scenario**, we consider those two jobs as two units of work executing in parallel. One unit of work reads from the DB2 source table and writes into the pipe, while the other unit of work reads from the pipe and loads the data into the target DB2 table.

In the **third scenario**, both the source and the target table have been partitioned. Each table has two partitions. Two jobs are created and both jobs contain two job steps. The first job step extracts data from a partition of the source table to a sequential file and the second job step loads the extracted sequential file into a partition of the target table. The first job does this for the first partition and the second job does this for the second partition.

In the **fourth scenario**, we manage the jobs as four units of work:

1. Unload data from partition 1 to pipe 1
2. Read data from pipe 1 and load into 1. partition of target DB
3. Unload data from part. 2 to pipe 2
4. Read data from pipe 2 and load into 2. partition of target DB



ETL Technique – ‘Sequential Design‘

We show here an example of an **initial load with a sequential design:** In a sequential design the different phases of the process must complete before the next phase can start. The split must, for example, complete before the key assignment can start.

In this example of an insurance company, the input file is split by claim type, which is then updated to reflect the generated identifier, such as claim ID or individual ID.

This data is sent to the transformation process, which splits the data for the load utilities. The load utilities are run in parallel, one for each partition. This process is repeated for every state for every year, a total of several hundred times.

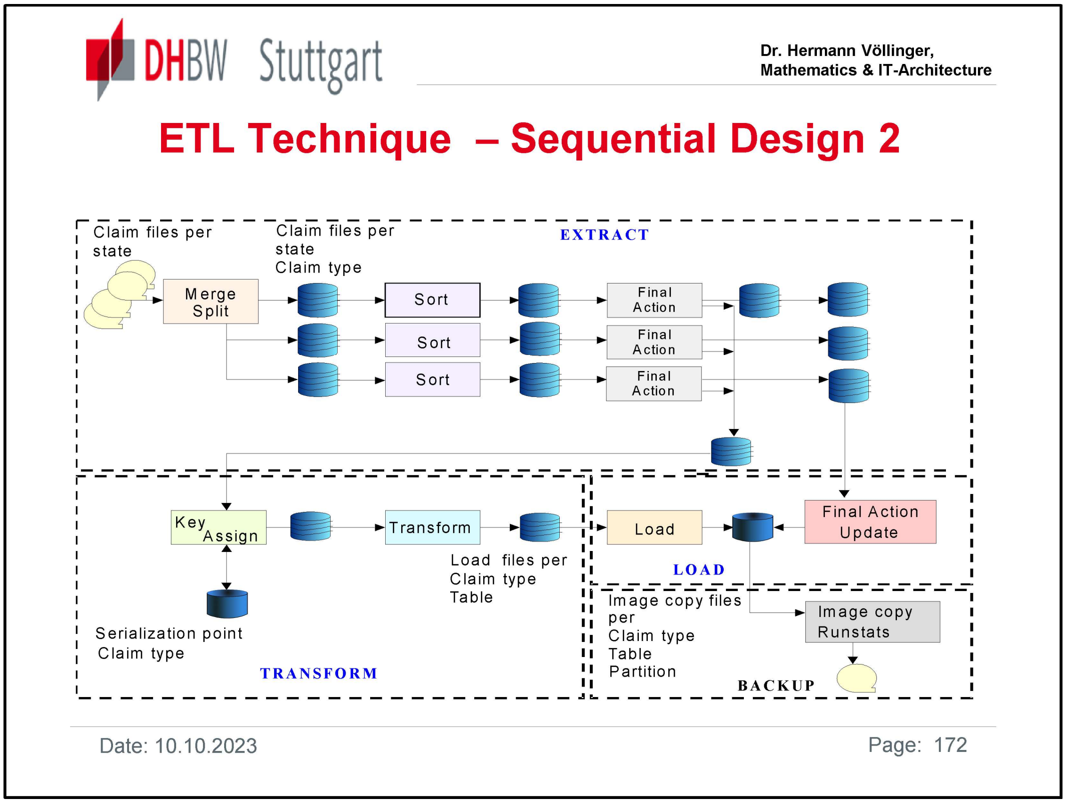
As can be seen at the bottom of the chart, the CPU utilization is very low until the load utilities start running in parallel.

The advantage of the sequential design is that it is easy to implement and uses a traditional approach, which most developers are familiar with.

This approach can be used for initial load of small to medium-sized data warehouses and refresh of data where a batch window is not an issue.

In a VLDB environment with large amounts of data a large number of scratch tapes are needed: one set for the split, one set for the key assignment, and one set for the transform. A large number of reads/writes for each record makes the elapsed time very long. If the data is in the multi-terabyte range, this process requires several days to complete through a single controller. Because of the number of input cartridges that are needed for this solution when loading a VLDB, media failure may be an issue. If we assume that 0.1 percent of the input cartridges fail on average, every 1000th cartridge fails.

This solution has a low degree of parallelism (only the loads are done in parallel), hence the inefficient usage of CPU and DASD. To cope with multi-terabyte data volumes, the solution must be able to run in parallel and the number of disk and tape I/Os must be kept to a minimum.



ETL Technique – Sequential Design 2

Here we show a design similar to the one in the previous figure. This design is based on a customer proof of concept conducted by the IBM Teraplex center in Poughkeepsie. The objectives for the test are to be able to run monthly refreshes within a weekend window. As indicated here, DASD is used for storing the data, rather than tape. The reason for using this approach is to allow I/O parallelism using SMS data striping. This design is divided into four groups:

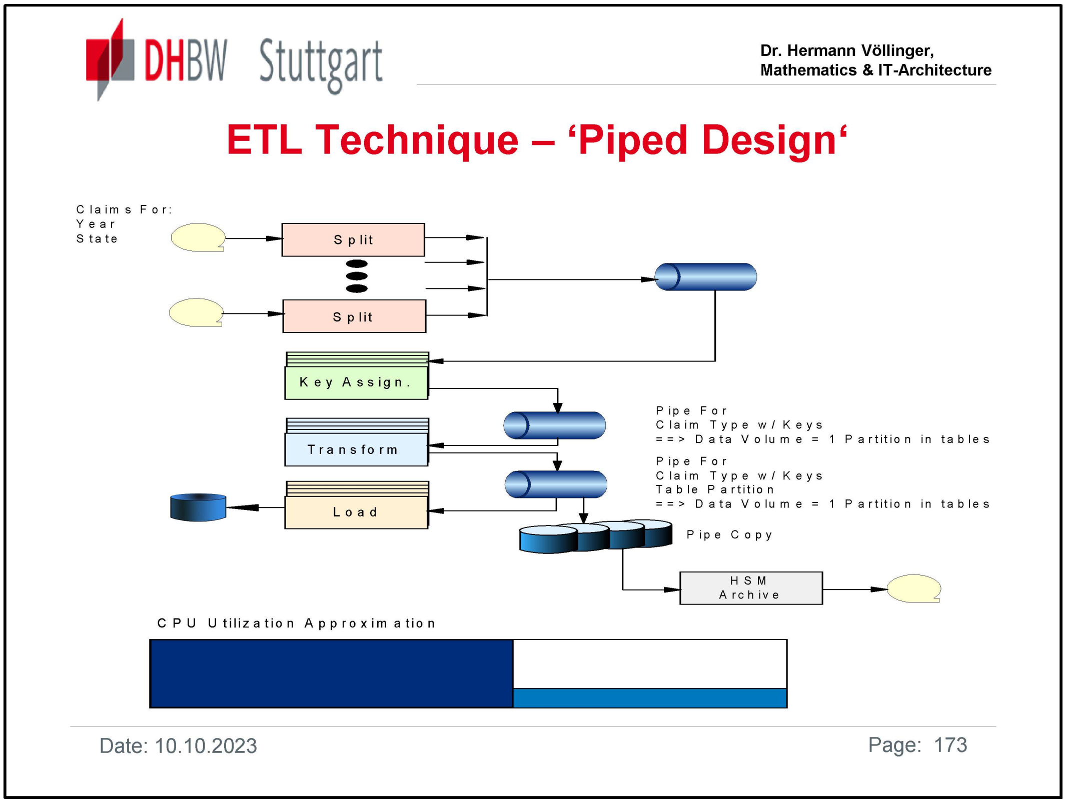
- Extract
- Transform
- Load
- Backup

The extract, transform and backup can be executed while the database is online and available for end user queries, removing these processes from the critical path.

It is only during the load that the database is not available for end user queries. Also, Runstats and Copy can be executed outside the critical path. This leaves the copy pending flag on the table space. As long as the data is accessed in a read-only mode, this does not cause any problem.

Executing Runstats outside the maintenance window may have severe performance impacts when accessing new partitions. Whether Runstats should be included in the maintenance must be decided on a case-by-case basis. The decision should be based on the query activity and loading strategy. If this data is load replaced into a partition with similar data volumes, Runstats can probably be left outside the maintenance part (or even ignored).

Separating the maintenance part from the rest of the processing not only improves end-user availability, it also makes it possible to plan when to execute the non-critical paths based on available resources.



ETL Technique – 'Piped Design'

Claims For:
Year
State

Split

Split

Key Assign.

Transform

Load

Pipe For
Claim Type w/ Keys
==> Data Volume = 1 Partition in tables

Pipe For
Claim Type w/ Keys
Table Partition
==> Data Volume = 1 Partition in tables

Pipe Copy

HSM
Archive

CPU Utilization Approximation

The **piped design** shown here uses pipes to avoid externalizing temporary data sets throughout the process, thus avoiding I/O system-related delays (tape controller, tape mounts, and so forth) and potential I/O failures.

Data from the transform pipes are read by both the DB2 Load utility and the archive process.

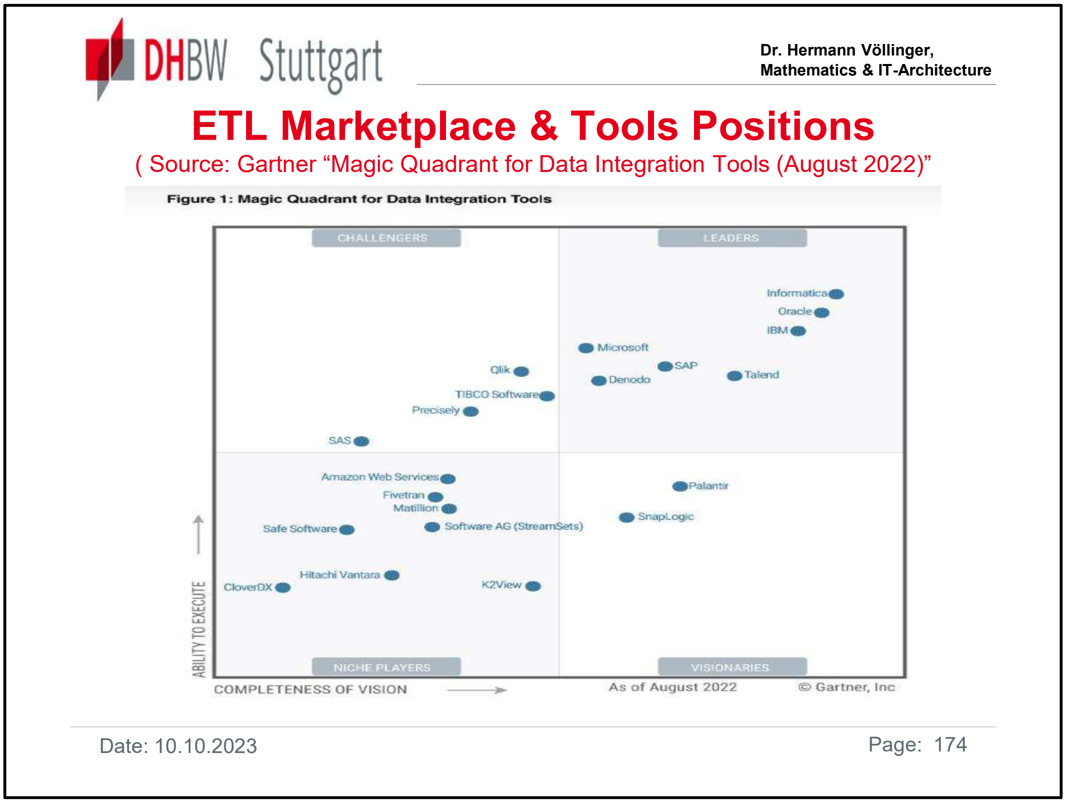
Since reading from a pipe is by nature asynchronous, these two processes do not need to wait for each other. That is, the Load utility can finish before the archive process has finished making DB2 tables available for end users.

This design enables parallelism—multiple input data sets are processed simultaneously.

Using batch pipes, data can also be written and read from the pipe at the same time, reducing elapsed time.

The throughput in this design is much higher than in the sequential design. The number of disk and tape I/O's has been drastically reduced and replaced by memory accesses. The data is processed in parallel throughout all phases.

Since batch pipes concurrently process multiple processes, restart is more complicated than in sequential processing. If no actions are taken, a piped solution must be restarted from the beginning. If restartability is an important issue, the piped design must include points where data is externalized to disk and where a restart can be performed.



ETL Marketplace & Tools Positions

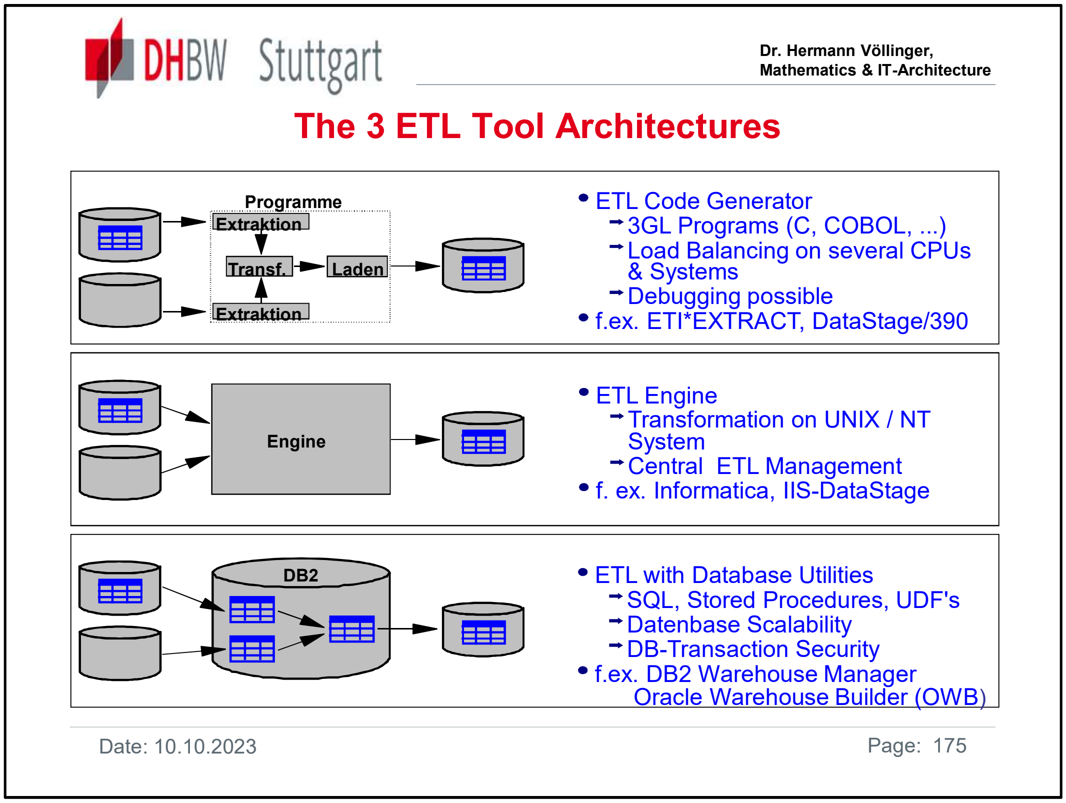
(Source: Gartner “Magic Quadrant for Data Integration Tools (August 2022)”

Figure 1: Magic Quadrant for Data Integration Tools

For more details see the documents “Gartner-Data_Integration_Tools-2022” and “Gartner-Data_Integration_Tools-2021” in our course Moodle:
Kurs DW 21E Category 3: „Data Population Tools…“

Gartner has defined multiple classes of functional capabilities that vendors of data integration tools must possess to deliver optimal value to organizations in support of a full range of data integration scenarios:

- Connectivity/adapter capabilities (data source and target support).
- Data delivery capabilities.
- Data transformation capabilities.
- Metadata and data modeling capabilities.
- Design and development environment capabilities.
- Data governance capabilities (data quality, profiling and mining).
- Deployment options and runtime platform capabilities.
- Operations and administration capabilities.
- Architecture and integration.
- Service-enablement capabilities.



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

The 3 ETL Tool Architectures

- ETL Code Generator
  - 3GL Programs (C, COBOL, ...)
  - Load Balancing on several CPUs & Systems
  - Debugging possible
- f.ex. ETI*EXTRACT, DataStage/390

- ETL Engine
  - Transformation on UNIX / NT System
  - Central ETL Management
- f. ex. Informatica, IIS-DataStage

- ETL with Database Utilities
  - SQL, Stored Procedures, UDF's
  - Datenbase Scalability
  - DB-Transaction Security
- f.ex. DB2 Warehouse Manager
  Oracle Warehouse Builder (OWB)

Date: 10.10.2023 Page: 175

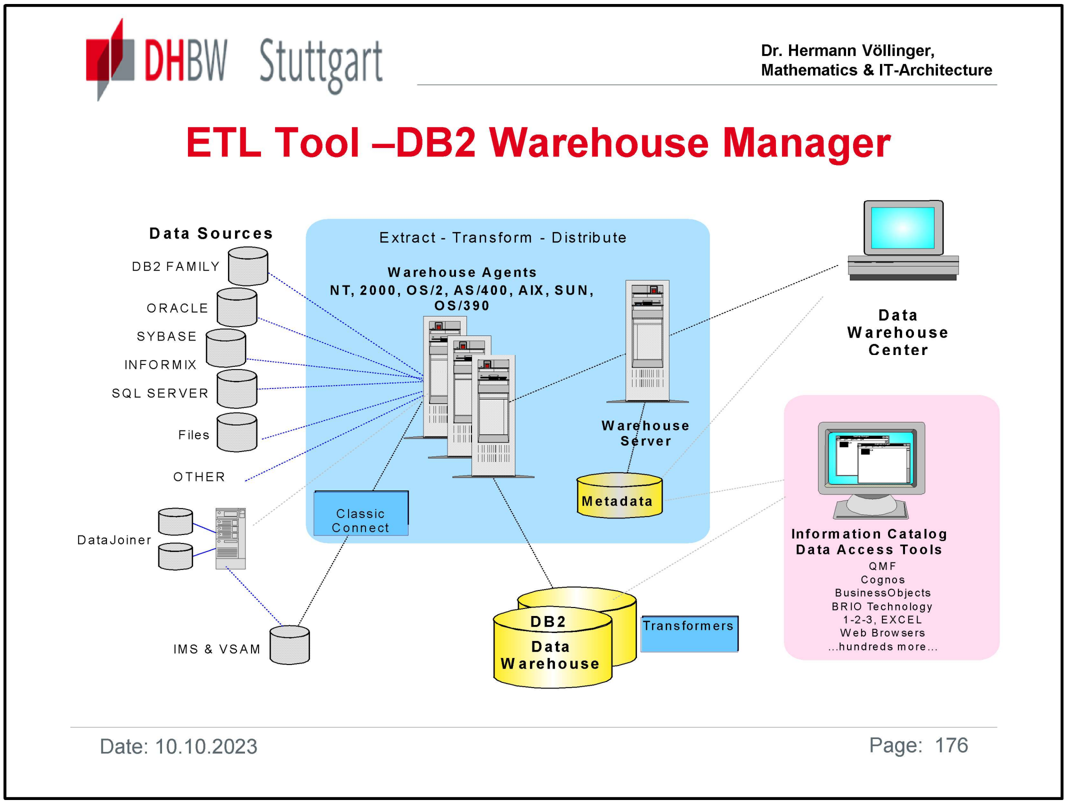
SQL Engines, like DB2 Warehouse Manager first extracts data from the source systems and then loads this data into DB2 tables. It then uses DB2 database technology (SQL, stored procedures, user defined functions) to transform the data inside DB2. With this approach all DB2 features like parallelization, scalability and transaction security are automatically available for the ETL process with no additional effort.

ETL Engines provide a centralized ETL management where all needed functionality is

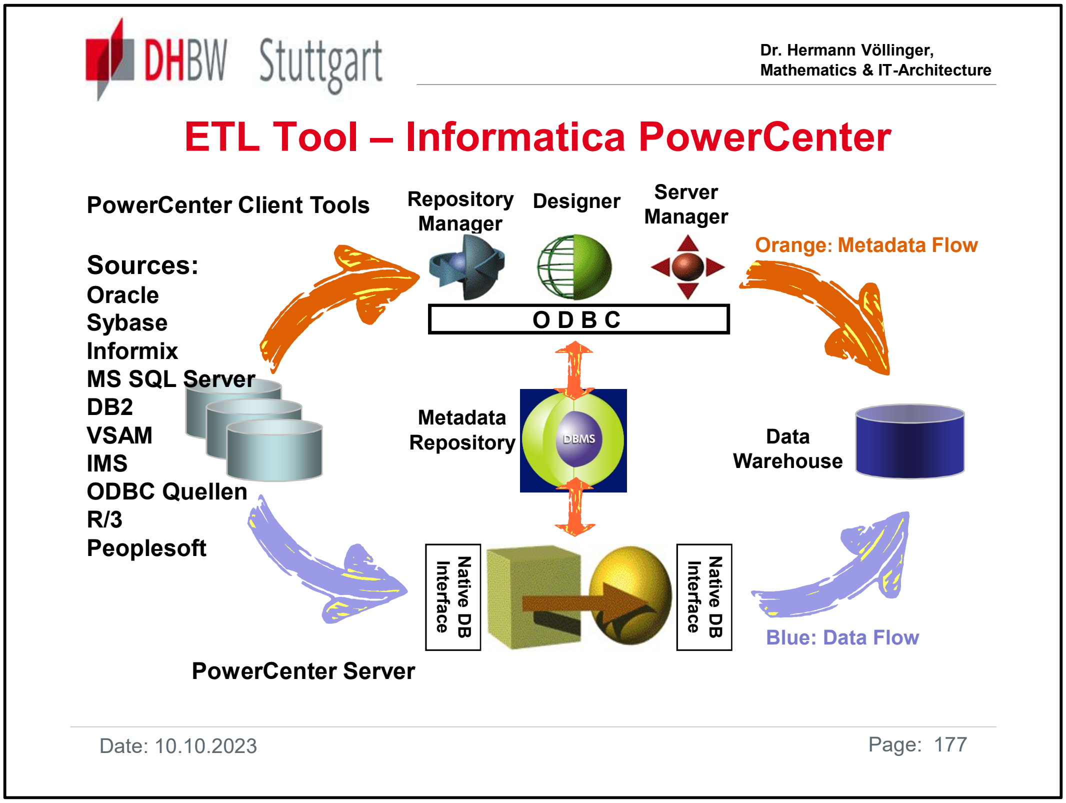
integrated transparently. Data is extracted from the source systems, transformed on a UNIX or Windows NT server and then loaded into the target system. An example for this approach is Ascential DataStage or Informatica.

ETL Code Generators generate 3GL programs (COBOL, C, ABAP, ...) for the extraction, transformation and loading of data. These programs can be run on different servers to achieve load balancing. Debugging and performance tuning is easy because the program source code is available. DataStage/390 is an example for this architecture.

Each of these architectures has its advantages in certain environments.



ETL Tool –DB2 Warehouse Manager



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

ETL Tool – Informatica PowerCenter

PowerCenter Client Tools

Repository Manager

Designer

Server Manager

ODBC

Orange: Metadata Flow

Sources:
Oracle
Sybase
Informix
MS SQL Server
DB2
VSAM
IMS
ODBC Quellen
R/3
Peoplesoft

Metadata Repository

DBMS

Data Warehouse

Native DB Interface

Native DB Interface

Blue: Data Flow

PowerCenter Server

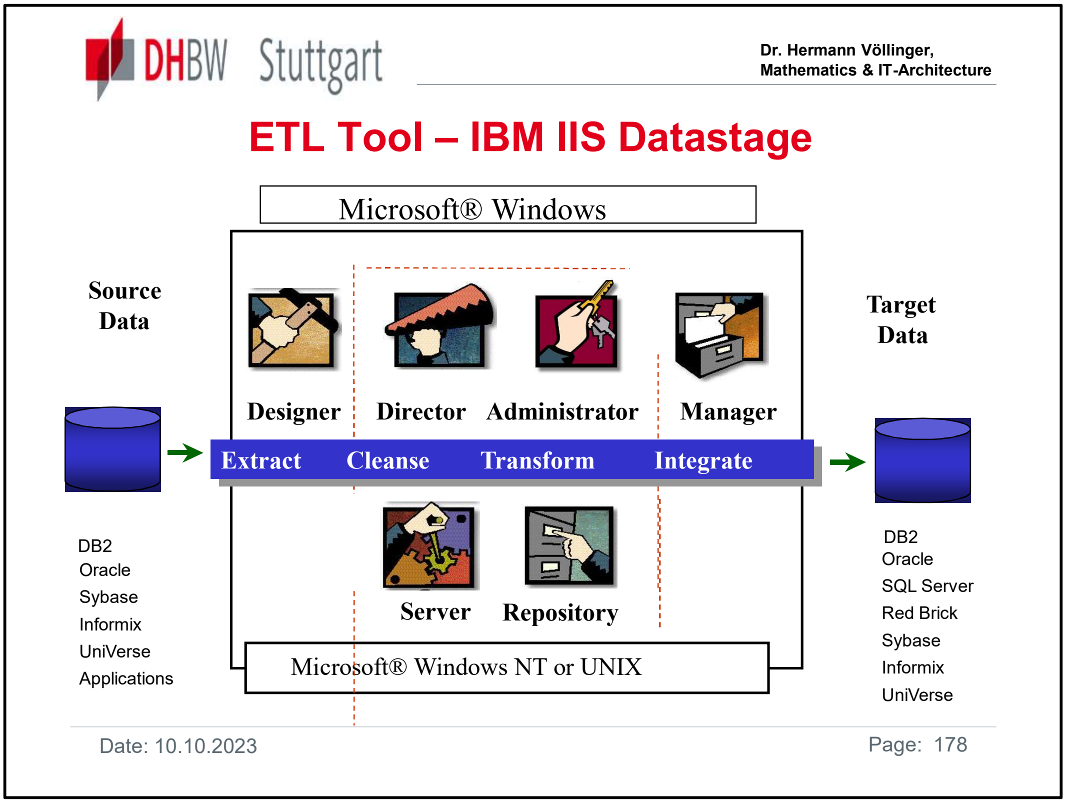
Date: 10.10.2023

Page: 177

To get more information about the functionality of Informatia PowerCenter, you can use the following link:
https://www.informatica.com/products/data-integration/powercenter.html#resources_pc

Informatica PowerCenter Standard Edition is a scalable, high-performing, easy-to-use enterprise data integration software that will greatly improve staff productivity and promote IT and business collaboration. The highly visual environment lets you connect and integrate data from any data source in any data format and automatically generate data integration mappings and workflows, without hand-coding. This edition includes basic XML processing, connectors to foundational data sources, an analyst tool to understand source data, partitioning, administration capabilities, basic data profiling, rapid prototyping, and batch data integration.
PowerCenter Editions with more features and functions are:

- Informatica PowerCenter **Advanced Edition**
- Informatica PowerCenter **Premium Edition**



ETL Tool – IBM IIS Datastage

Microsoft® Windows

Source Data

Designer
Director
Administrator
Manager

Extract
Cleanse
Transform
Integrate

Target Data

DB2
Oracle
Sybase
Informix
UniVerse
Applications

Server
Repository

Microsoft® Windows NT or UNIX

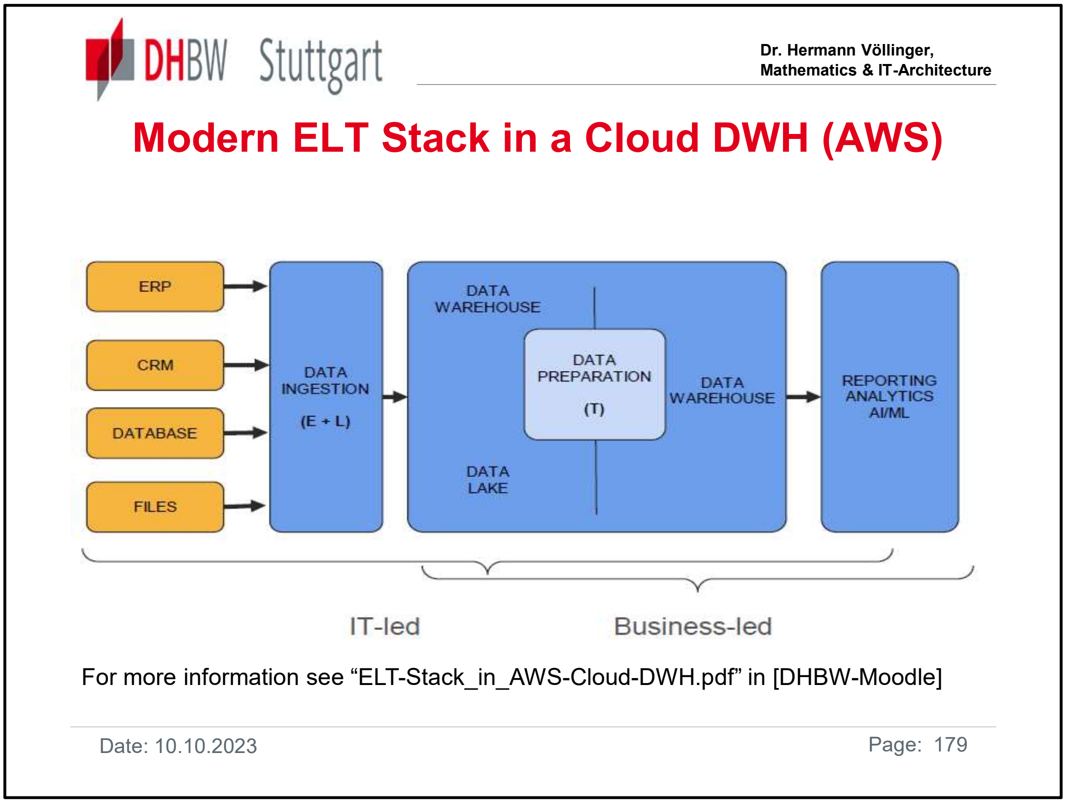
DB2
Oracle
SQL Server
Red Brick
Sybase
Informix
UniVerse

See also a guided tour „Offload Data Warehousing to Hadoop by using DataStage" in the document "Guided Datastage Demo-Deutsch.pdf" in Moodle: https://elearning.dhbw-stuttgart.de/moodle/mod/folder/view.php?id=323626

Use IBM® InfoSphere® DataStage® to load Hadoop and use YARN to manage DataStage workloads in a Hadoop cluster(a registered IBM Cloud Id is needed!): In this demo, you use DataStage to complete extract, transform, and load (ETL) data processing in a traditional enterprise data warehouse. You then offload the data and ETL processing into scalable, high-value Hadoop clusters and data lakes. You will walk through the following tasks and you get experience with the following features:

- Learn how to run DataStage traditional ETL jobs.
- Configure DataStage to run inside Hadoop Clusters.
- Examine execution logs to ensure configuration worked correctly.

To get a concrete examples of data stage jobs in a concrete DWH project, you can also look on the following presentation in [DHBW-Moodle] (…/Supp. Info../Category 3): "DataStage_Sample-Transformations.pdf"

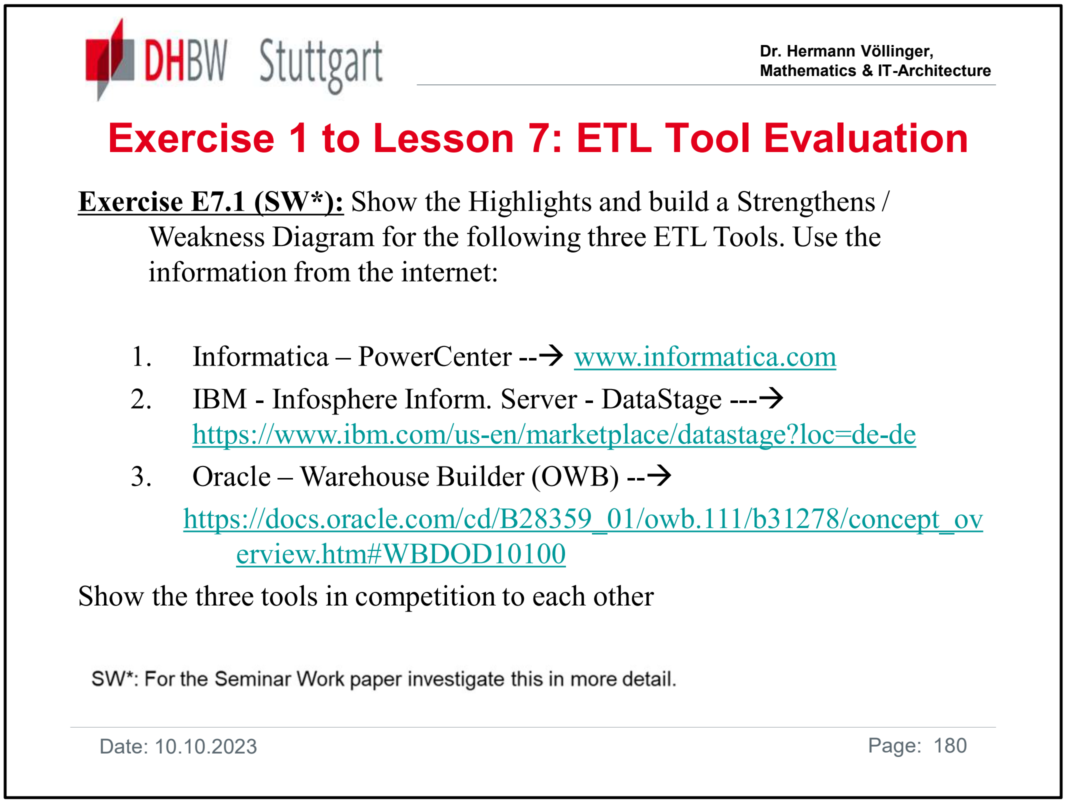


Modern ELT Stack in a Cloud DWH (AWS)

For more information see “ELT-Stack_in_AWS-Cloud-DWH.pdf” in [DHBW-Moodle]

See in [DHBW-Moodle] the YouTube video: “The Modern ELT Stack to Win with Cloud Data Warehousing”
https://www.youtube.com/watch?v=9x_g1P7Fjx4&feature=youtu.be

In this video you get an explanation of the differences between the traditional ETL approach (mostly used in the chapters above) versus the ELT approach which is used in more cloud-oriented DWH architectures. An example is the Amazon Redshift DWH (see the picture above).



Exercise 1 to Lesson 7: ETL Tool Evaluation

Exercise E7.1 (SW*): Show the Highlights and build a Strengthens / Weakness Diagram for the following three ETL Tools. Use the information from the internet:

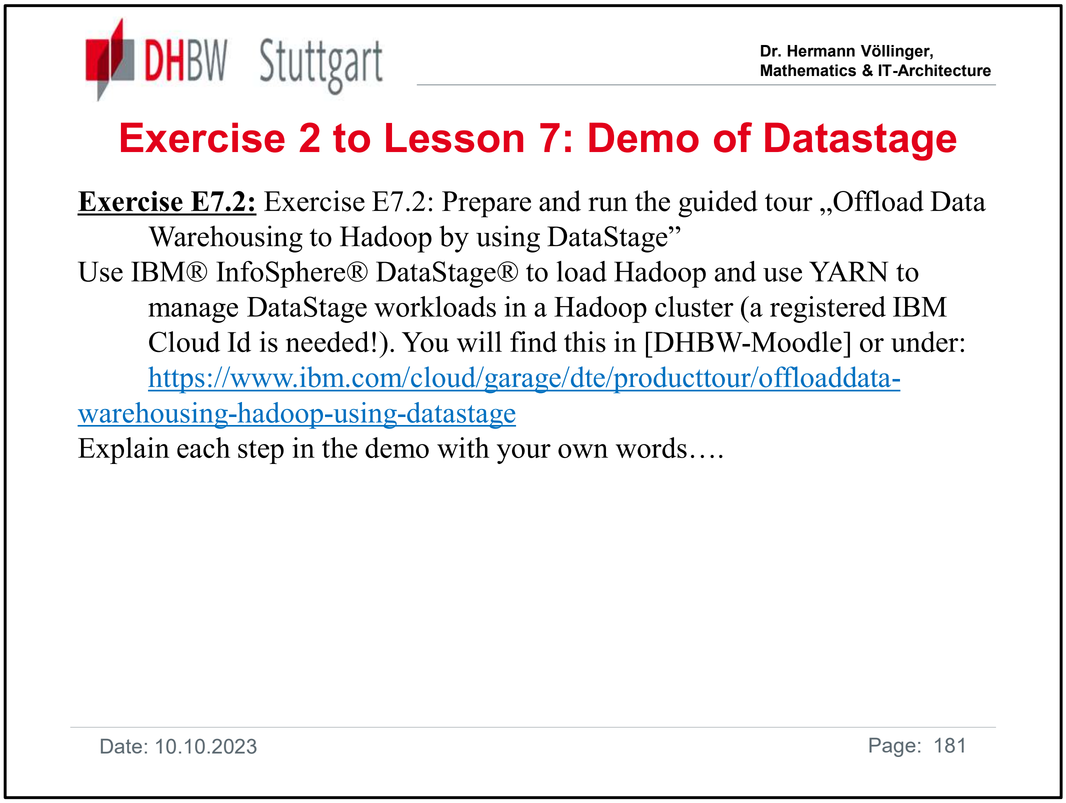
1. Informatica – PowerCenter --→ www.informatica.com
2. IBM - Infosphere Inform. Server - DataStage ---→ https://www.ibm.com/us-en/marketplace/datastage?loc=de-de
3. Oracle – Warehouse Builder (OWB) --→ https://docs.oracle.com/cd/B28359_01/owb.111/b31278/concept_overview.htm#WBDOD10100

Show the three tools in competition to each other

SW*: For the Seminar Work paper investigate this in more detail.

For more details see in [DHBW-Moodle] the documents:

- “Gartner-Data_Integration_Tools-2022.pdf”
- “Gartner-Data_Integration_Tools-2023.pdf”



Exercise 2 to Lesson 7: Demo of Datastage

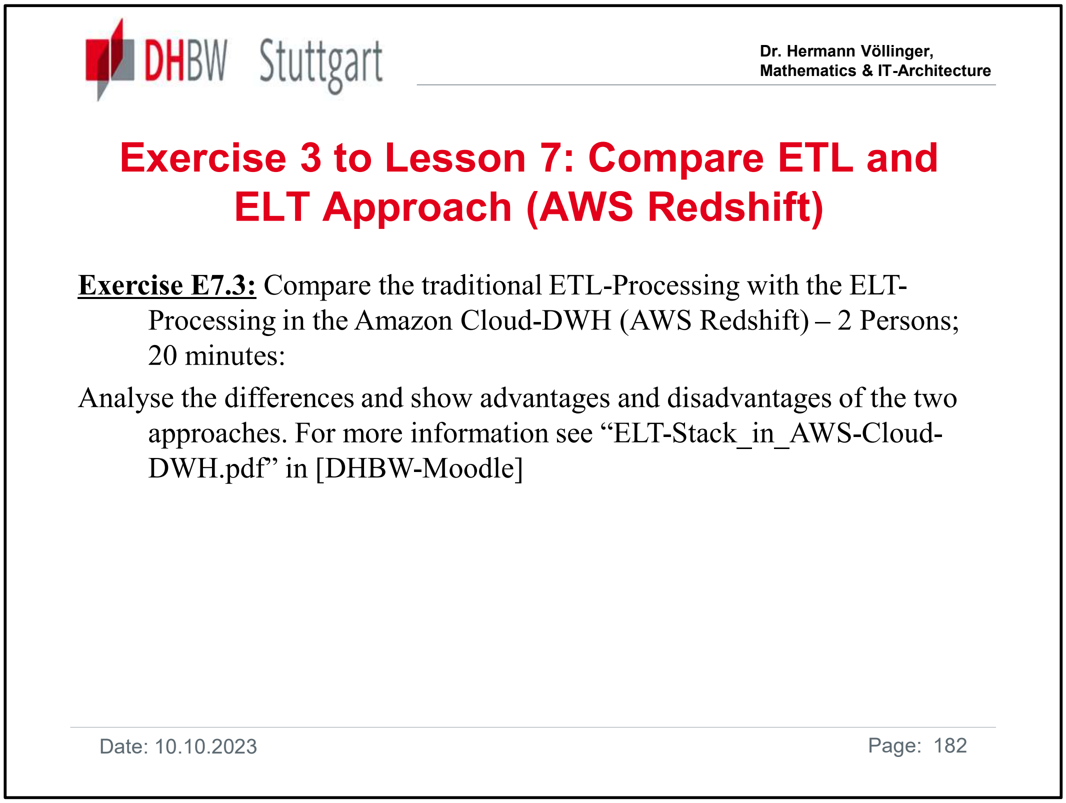
Exercise E7.2: Exercise E7.2: Prepare and run the guided tour „Offload Data Warehousing to Hadoop by using DataStage”

Use IBM® InfoSphere® DataStage® to load Hadoop and use YARN to manage DataStage workloads in a Hadoop cluster (a registered IBM Cloud Id is needed!). You will find this in [DHBW-Moodle] or under: https://www.ibm.com/cloud/garage/dte/producttour/offloaddata-warehousing-hadoop-using-datastage

Explain each step in the demo with your own words….

For more details see in [DHBW-Moodle] the documents:

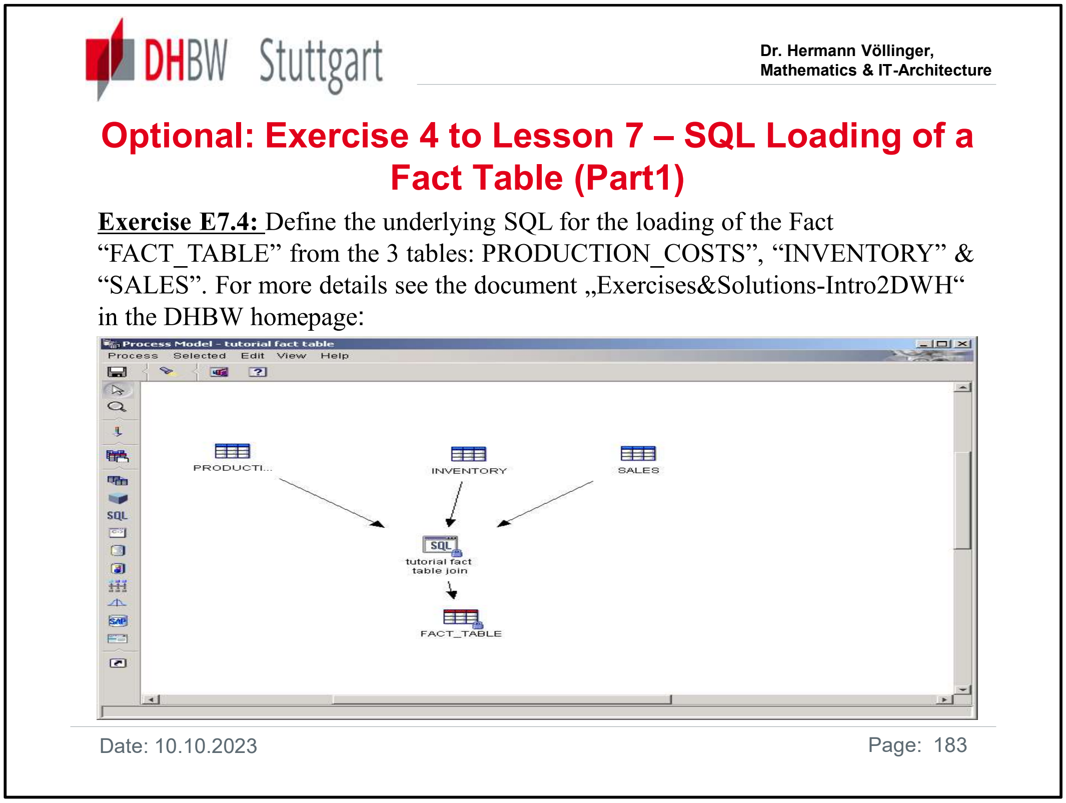
- “Guided Datastage Demo.pdf”
- “Guided Datastage Demo-Deutsch.pdf”



Exercise 3 to Lesson 7: Compare ETL and ELT Approach (AWS Redshift)

Exercise E7.3: Compare the traditional ETL-Processing with the ELT-Processing in the Amazon Cloud-DWH (AWS Redshift) – 2 Persons; 20 minutes:

Analyse the differences and show advantages and disadvantages of the two approaches. For more information see “ELT-Stack_in_AWS-Cloud-DWH.pdf” in [DHBW-Moodle]



Optional: Exercise 4 to Lesson 7 – SQL Loading of a Fact Table (Part1)

Exercise E7.4: Define the underlying SQL for the loading of the Fact “FACT_TABLE” from the 3 tables: PRODUCTION_COSTS”, “INVENTORY” & “SALES”. For more details see the document „Exercises&Solutions-Intro2DWH“ in the DHBW homepage:

Join information from **SALES, PRODUCTION_COSTS,** and **INVENTORY** tables and writes the result in a new table**: FACT_TABLE.**

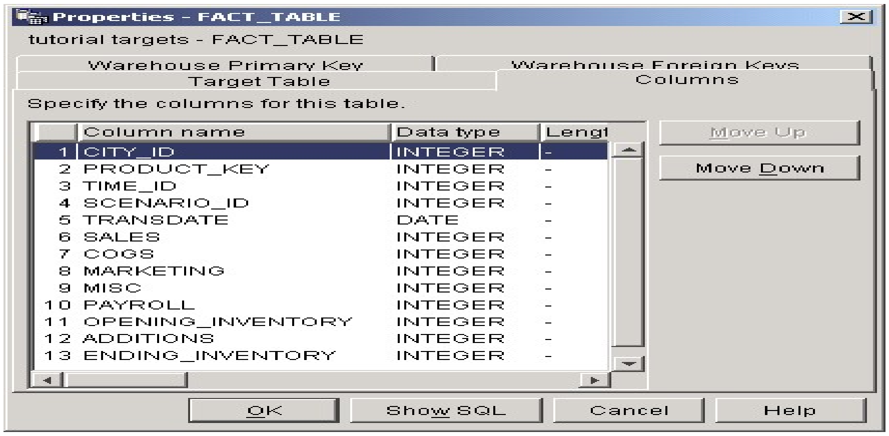
Compute calculated column to derive ending Inventory value.
Generates CITY_ID, TIME_ID, and SCENARIO_ID

Remark: DB2 Warehouse Manager uses SQL Select statement to extract data from warehouse sources and generates an INSERT statement to insert the data into the warehouse target table.

For more details see on the homepage:
[HVö-2]: *Hermann Völlinger and Other: Exercises & Solutions of the Lecture ‘Introduction to Data Warehousing’;* DHBW Stuttgart; WS2021
http://www.dhbw-stuttgart.de/~hvoellin/

Optional: Exercise 4 to Lesson 7 – SQL Loading of a Fact Table (Part2)

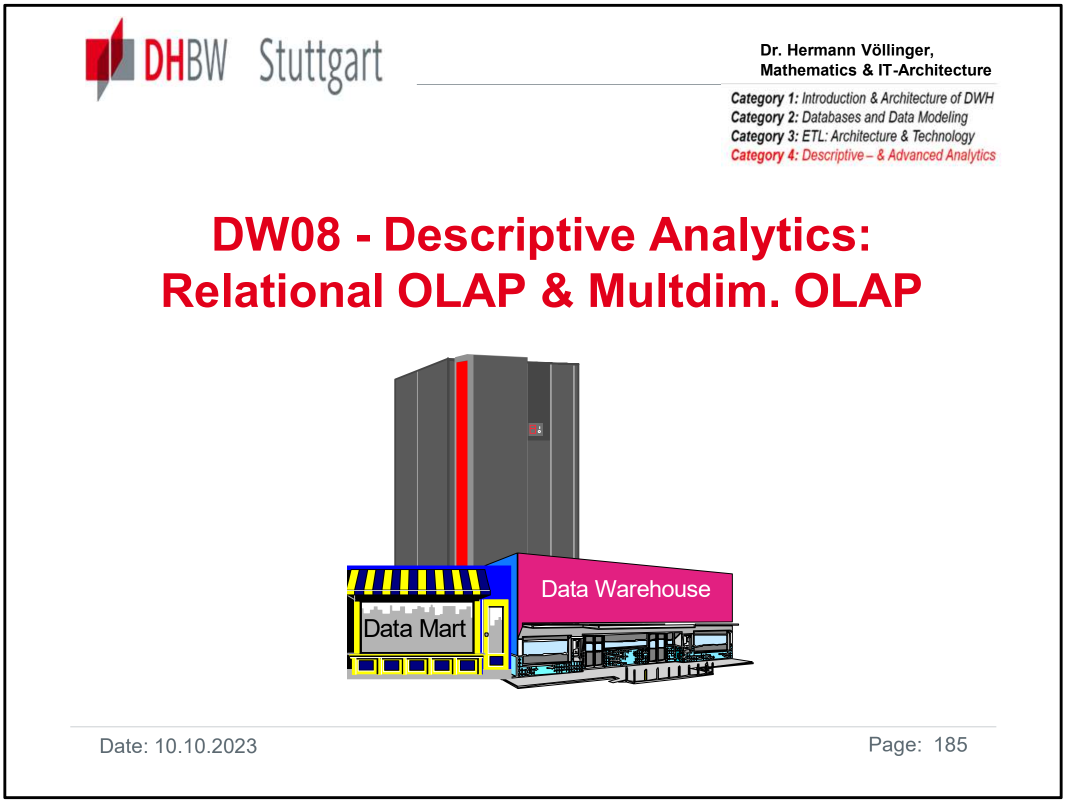
The structure of the target fact table can be seen in the following screenshot:



Specify the columns for this table.

	Column name	Data type	Lengt
1	CITY_ID	INTEGER	-
2	PRODUCT_KEY	INTEGER	-
3	TIME_ID	INTEGER	-
4	SCENARIO_ID	INTEGER	-
5	TRANSDATE	DATE	-
6	SALES	INTEGER	-
7	COGS	INTEGER	-
8	MARKETING	INTEGER	-
9	MISC	INTEGER	-
10	PAYROLL	INTEGER	-
11	OPENING_INVENTORY	INTEGER	-
12	ADDITIONS	INTEGER	-
13	ENDING_INVENTORY	INTEGER	-

Conditions: **For more details see on the homepage:**
[HVö-2]: *Hermann Völlinger and Other: Exercises & Solutions of the Lecture 'Introduction to Data Warehousing';* DHBW Stuttgart; SS2021
http://www.dhbw-stuttgart.de/~hvoellin/



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

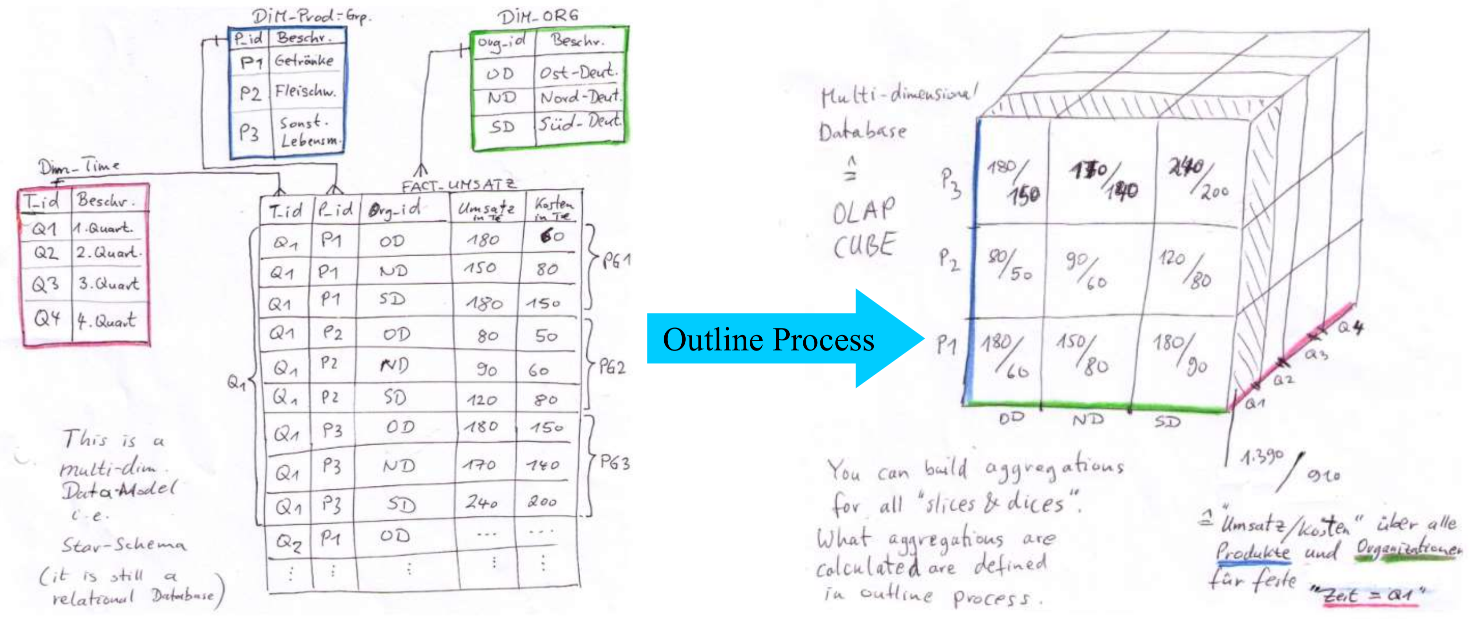
Category 1: Introduction & Architecture of DWH
Category 2: Databases and Data Modeling
Category 3: ETL: Architecture & Technology
Category 4: Descriptive – & Advanced Analytics

DW08 - Descriptive Analytics: Relational OLAP & Multdim. OLAP

Data Mart
Data Warehouse

Date: 10.10.2023 Page: 185

Motivation: In chapter 5 we learned much about Multi Dim. Data Models (MDDM). Out of such a data model a QLAP Cube is build. This process is called “OUTLINE” process. See the following pictures:



DIM-Prod-Grp.

P_id	Beschr.
P1	Getränke
P2	Fleischw.
P3	Sonst. Lebensm.

DIM-ORG

Org-id	Beschr.
OD	Ost-Deut.
ND	Nord-Deut.
SD	Süd-Deut.

Dim-Time

T_id	Beschr.
Q1	1. Quart.
Q2	2. Quart.
Q3	3. Quart
Q4	4. Quart

FACT-UMSATZ

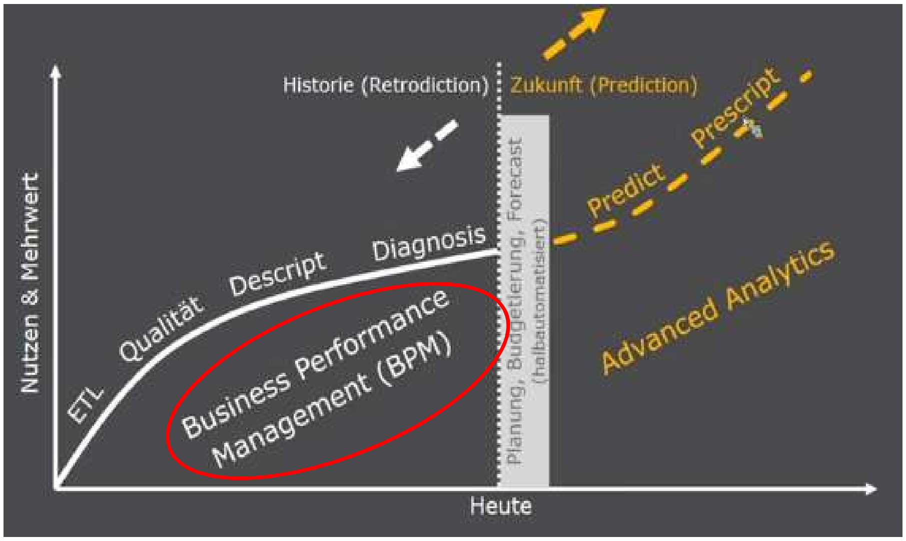
T-id	P_id	Org-id	Umsatz in T€	Kosten in T€
Q1	P1	OD	180	60
Q1	P1	ND	150	80
Q1	P1	SD	180	150
Q1	P2	OD	80	50
Q1	P2	ND	90	60
Q1	P2	SD	120	80
Q1	P3	OD	180	150
Q1	P3	ND	170	140
Q1	P3	SD	240	200
Q2	P1	OD	...	...
⋮	⋮	⋮	⋮	⋮

This is a multi-dim. Data-Model i.e. Star-Schema (it is still a relational Database)

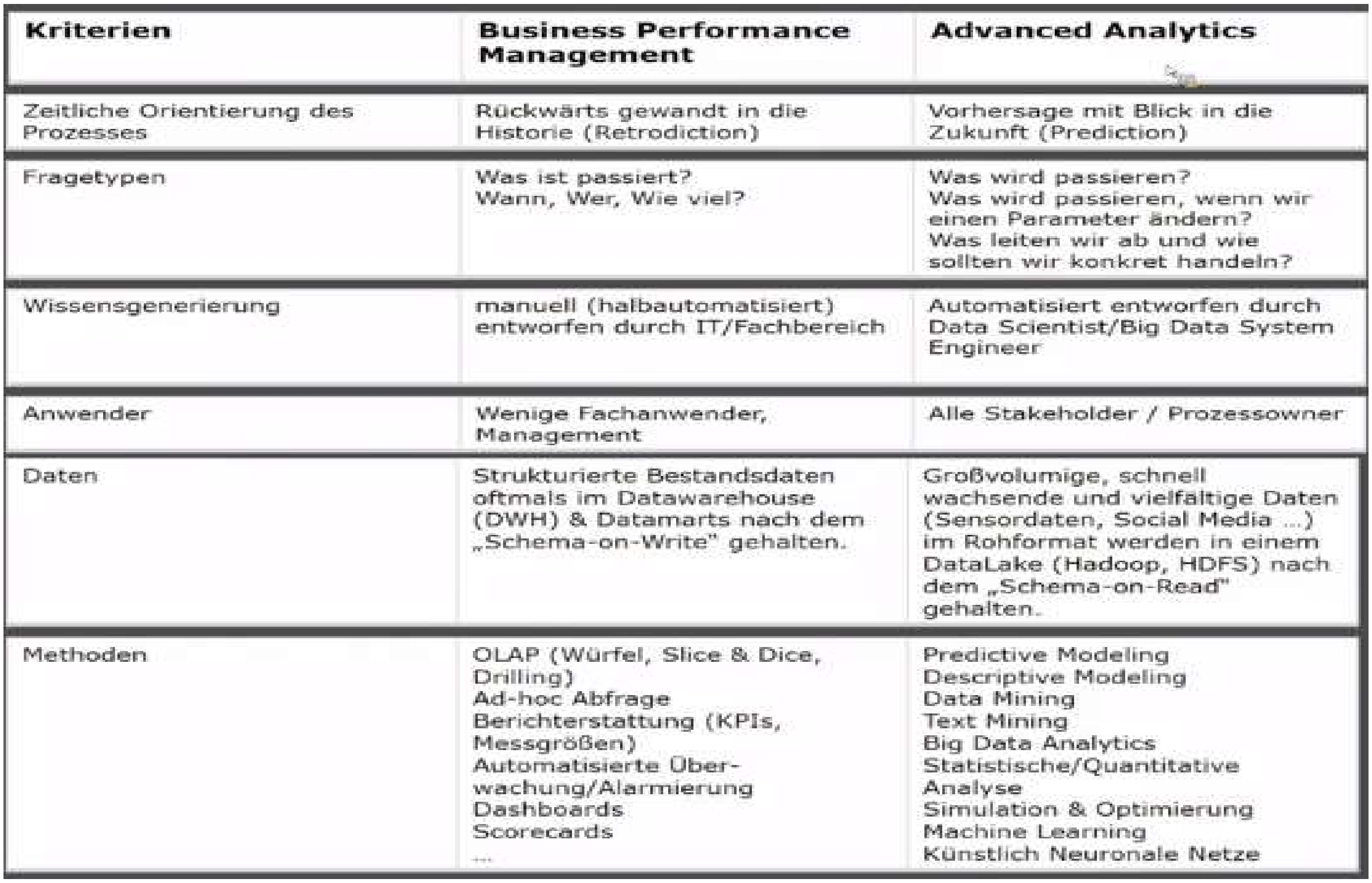
You can build aggregations for all "slices & dices". What aggregations are calculated are defined in outline process.

≙ "Umsatz/Kosten" über alle Produkte und Organizationen für feste "Zeit = Q1"

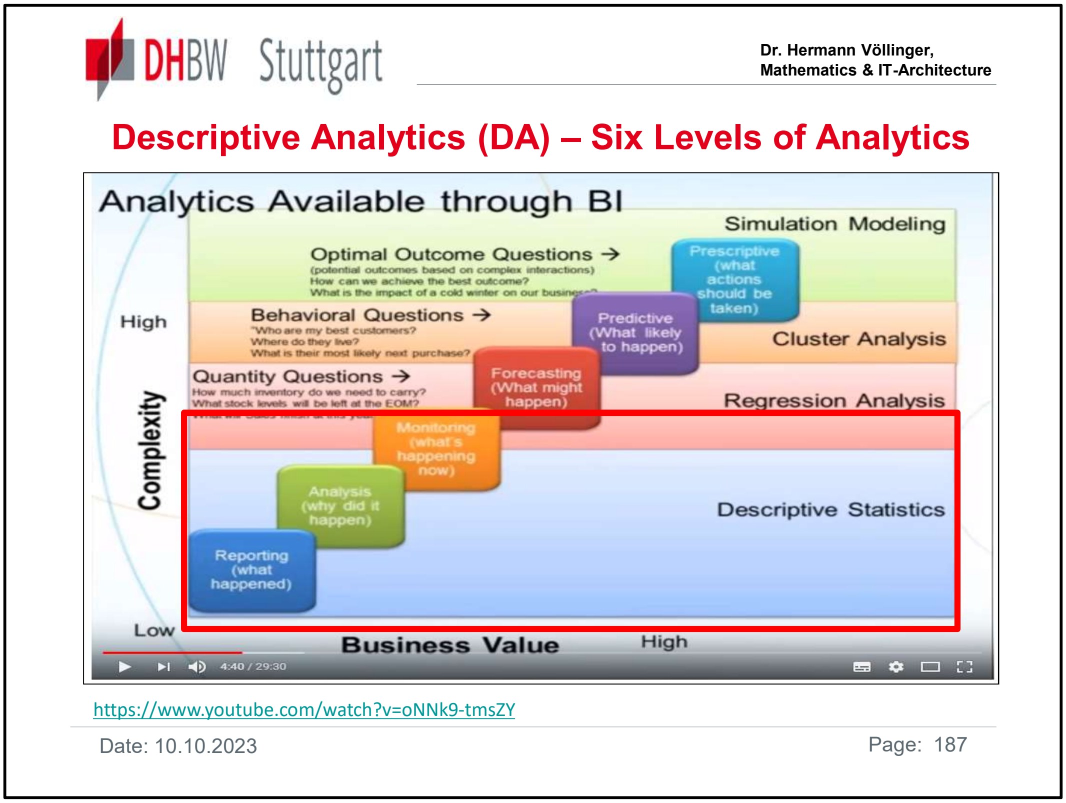
Motivation - From Descriptive to Predictive Analytics



Business Performance Management (BPM) versus Advanced Analytics (AA):



Kriterien	Business Performance Management	Advanced Analytics
Zeitliche Orientierung des Prozesses	Rückwärts gewandt in die Historie (Retrodiction)	Vorhersage mit Blick in die Zukunft (Prediction)
Fragetypen	Was ist passiert? Wann, Wer, Wie viel?	Was wird passieren? Was wird passieren, wenn wir einen Parameter ändern? Was leiten wir ab und wie sollten wir konkret handeln?
Wissensgenerierung	manuell (halbautomatisiert) entworfen durch IT/Fachbereich	Automatisiert entworfen durch Data Scientist/Big Data System Engineer
Anwender	Wenige Fachanwender, Management	Alle Stakeholder / Prozessowner
Daten	Strukturierte Bestandsdaten oftmals im Datawarehouse (DWH) & Datamarts nach dem „Schema-on-Write" gehalten.	Großvolumige, schnell wachsende und vielfältige Daten (Sensordaten, Social Media ...) im Rohformat werden in einem DataLake (Hadoop, HDFS) nach dem „Schema-on-Read" gehalten.
Methoden	OLAP (Würfel, Slice & Dice, Drilling) Ad-hoc Abfrage Berichterstattung (KPIs, Messgrößen) Automatisierte Überwachung/Alarmierung Dashboards Scorecards ...	Predictive Modeling Descriptive Modeling Data Mining Text Mining Big Data Analytics Statistische/Quantitative Analyse Simulation & Optimierung Machine Learning Künstlich Neuronale Netze

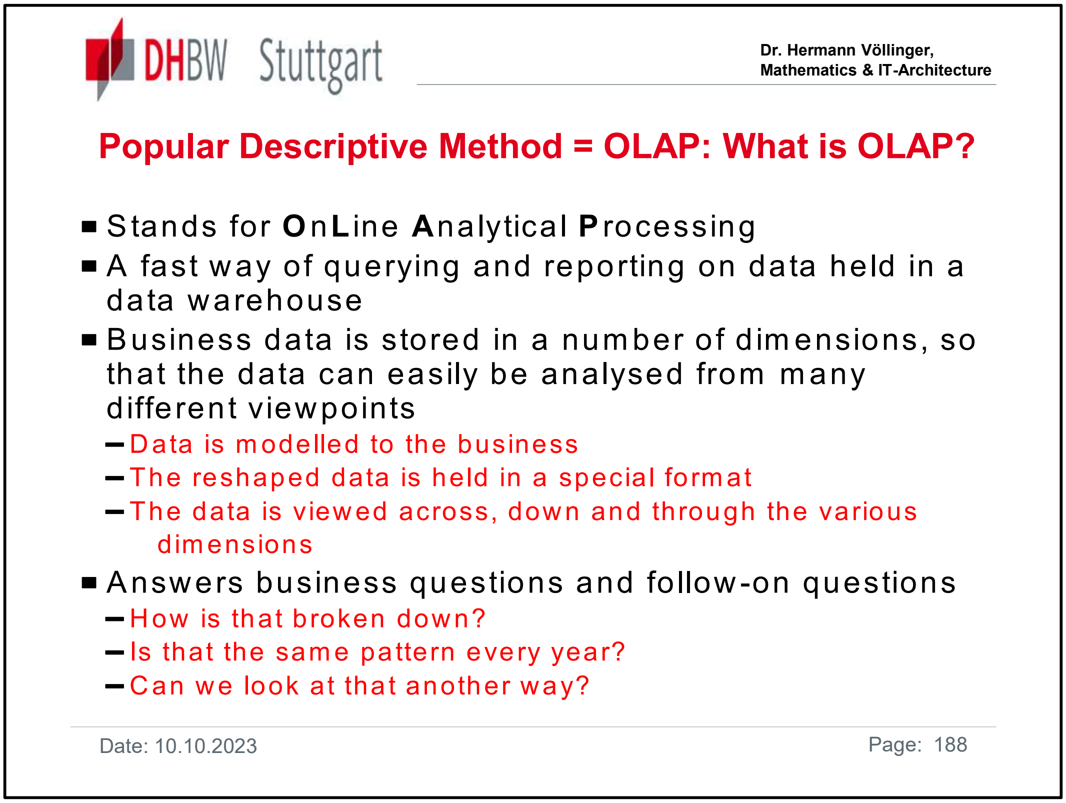


Descriptive Analytics (DA) – Six Levels of Analytics

Analytics Available through BI
Simulation Modeling
Optimal Outcome Questions →
(potential outcomes based on complex interactions)
How can we achieve the best outcome?
What is the impact of a cold winter on our business?
Prescriptive (what actions should be taken)
High
Behavioral Questions →
"Who are my best customers?
Where do they live?
What is their most likely next purchase?
Predictive (What likely to happen)
Cluster Analysis
Quantity Questions →
How much inventory do we need to carry?
What stock levels will be left at the EOM?
Forecasting (What might happen)
Regression Analysis
Complexity
Monitoring (what's happening now)
Analysis (why did it happen)
Descriptive Statistics
Reporting (what happened)
Low
Business Value
High
4:40 / 29:30

https://www.youtube.com/watch?v=oNNk9-tmsZY

1. **Reporting:** All companies regularly use reporting tools to report on their current sales and sales figures. These procedures are standard and do not add much new value compared to the competition. Nevertheless, this procedure is proven and sometimes mandatory for the reporting system.
2. **Analysis:** analysis of the departments and the questions why certain numbers have emerged are common today. Here comes the knowledge of the local parameters and good knowledge of the sales field or sales experience to fruition.
3**. Monitoring:** The monitoring also includes a rating and thus offers:• information about what an intervention does, how well it works, and whether it achieves its goals;• Guidelines for future intervention activities;• an important part of accountability to funding agencies and stakeholders.
Monitoring and evaluation plans should be prepared at the beginning of an intervention development process.
4. **Forecasting:** Forecasts are predictions for the future that are based on past and present data and are mostly generated by analyzing trends. A banking example could be the estimate of the probability of a loan application closing at a future point in time. Prediction is a similar but more general term. Both can relate to formal statistical methods using time series, cross-sectional or longitudinal data, or alternatively to less formal assessment methods. Risk and uncertainty are central to forecasts and forecasts. It is generally considered good practice to indicate the degree of uncertainty of the forecasts.
5**. Predictive Analysis:** A prediction is a statement about a future event. A prediction is often, but not always, based on experience or knowledge. Although future events are necessarily uncertain, so that guaranteed accurate information about the future is often not possible, the forecast can be helpful in making plans for possible developments.
6. **Prescriptive Analysis**: The prescriptive analytic termed the "ultimate limit of analytical ability" involves the application of mathematical and computer science and proposes decision-making options to exploit the results of descriptive and predictive analytics.

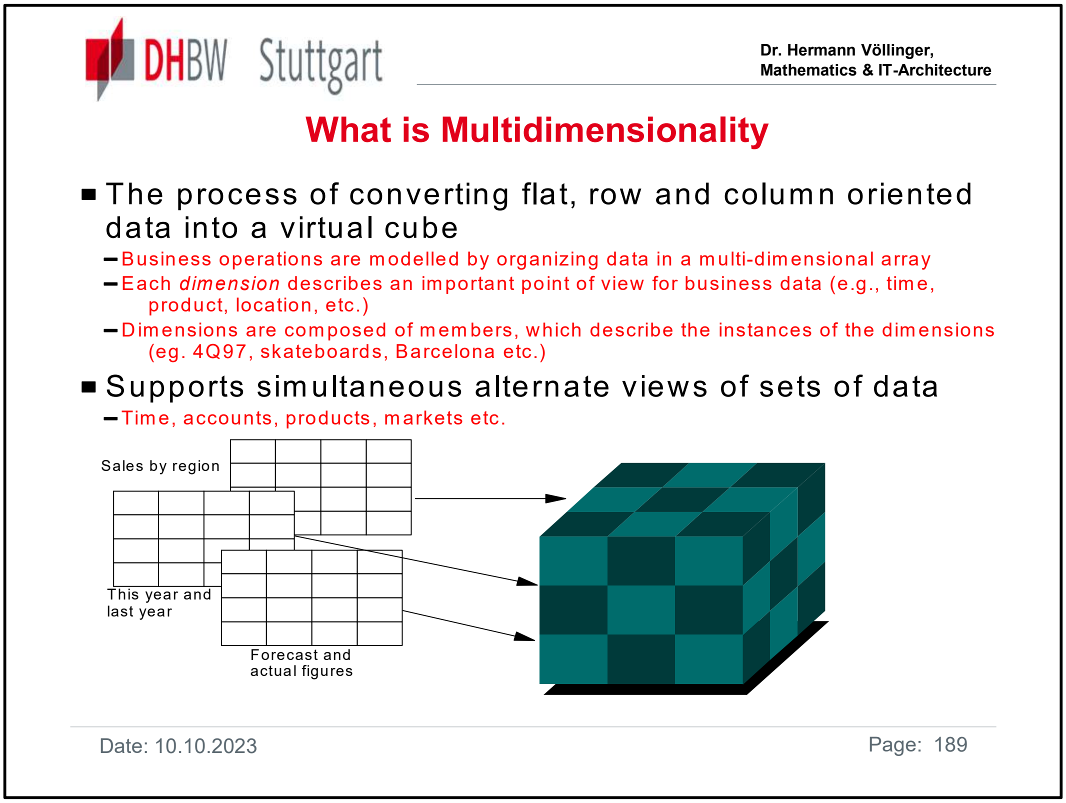


Popular Descriptive Method = OLAP: What is OLAP?

- Stands for **O**n**L**ine **A**nalytical **P**rocessing
- A fast way of querying and reporting on data held in a data warehouse
- Business data is stored in a number of dimensions, so that the data can easily be analysed from many different viewpoints
  - Data is modelled to the business
  - The reshaped data is held in a special format
  - The data is viewed across, down and through the various dimensions
- Answers business questions and follow-on questions
  - How is that broken down?
  - Is that the same pattern every year?
  - Can we look at that another way?

Compare also what we learned before above Multi Dimensional Data Modeling (MDDM).
How is broken down ? ---→ detail of hierarchy (drill down, drill up)
...every year ---→ time dimension
...in another way ? ---→ slice and dice

OLAP tools take you a step beyond query and reporting tools. Via OLAP tools, data is represented using a multidimensional model rather than the more traditional tabular data model.
The traditional model defines a database schema that focuses on modelling a process of function, and the information is viewed as a set of transactions, each which occurred at some single point in time.
The multidimensional model usually defines a star schema, viewing data not as a single event but rather as the cumulative effect of events over some period of time, such as weeks, then months, then years.
With OLAP tools, the user generally views the data in grids or crosstabs that can be pivoted to offer different perspectives on the data. OLAP also enables interactive querying of the data. For example, a user can look at information at one aggregation (such as a sales region) and then drill down to more detail information, such as sales by state, then city, then store.



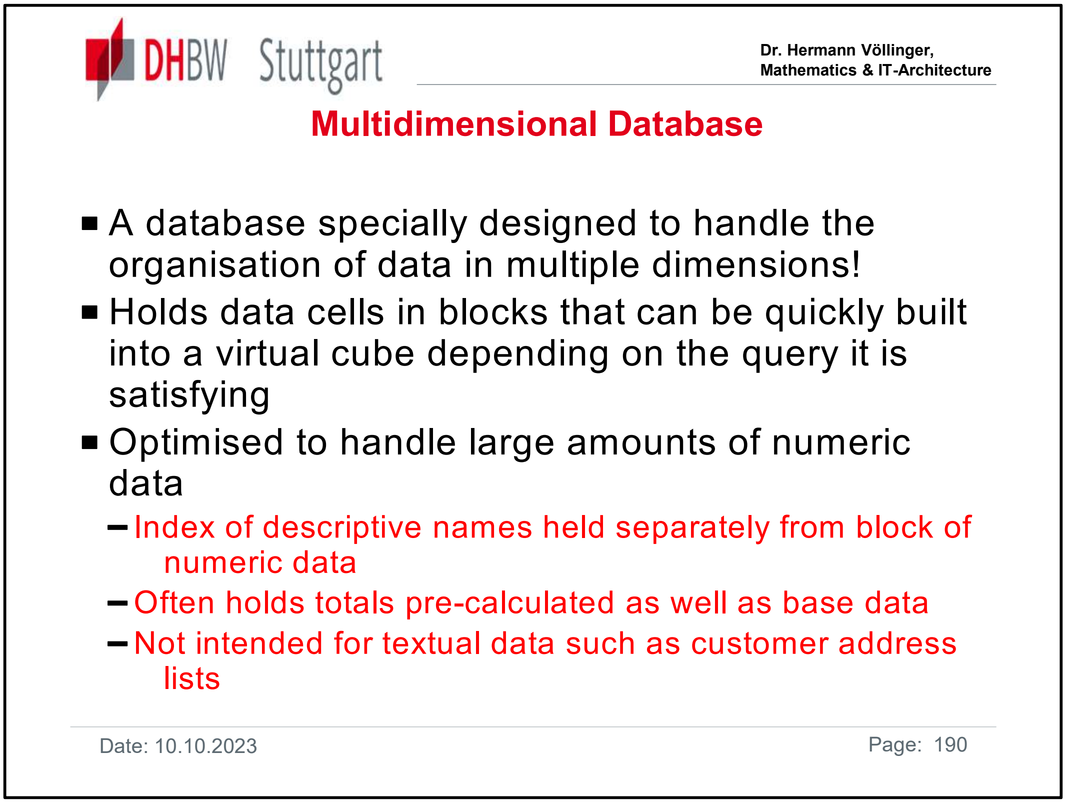
What is Multidimensionality

- The process of converting flat, row and column oriented data into a virtual cube
  - Business operations are modelled by organizing data in a multi-dimensional array
  - Each *dimension* describes an important point of view for business data (e.g., time, product, location, etc.)
  - Dimensions are composed of members, which describe the instances of the dimensions (eg. 4Q97, skateboards, Barcelona etc.)
- Supports simultaneous alternate views of sets of data
  - Time, accounts, products, markets etc.

Multidimensionality is turning data from various sources, relational tables or flat files, into a structure were data is grouped into separate, heterogeneous dimensions.

This is often referred to as a cube. In reality cubes are three dimensional of course, but the term is used for a database holding data from more than three dimensions.

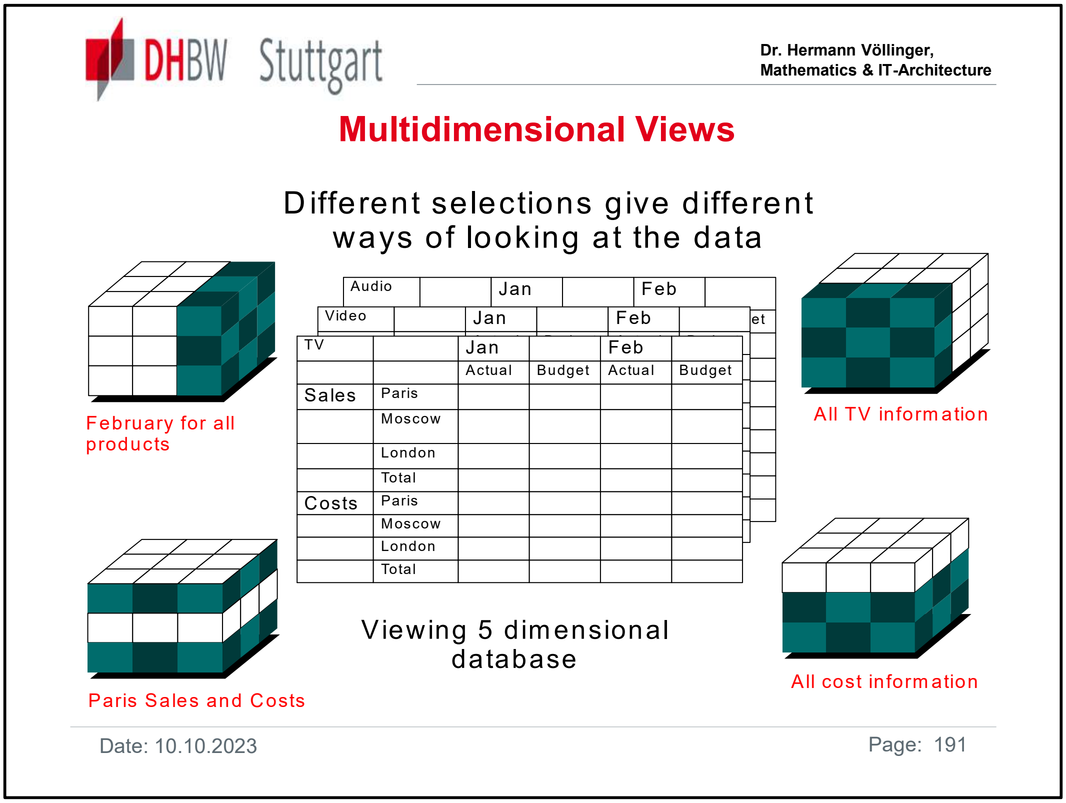
Dimensions are made up of **members**. In the example here , a dimension might be products and skateboards may be a member of that dimension.



Multidimensional Database

- A database specially designed to handle the organisation of data in multiple dimensions!
- Holds data cells in blocks that can be quickly built into a virtual cube depending on the query it is satisfying
- Optimised to handle large amounts of numeric data
  - Index of descriptive names held separately from block of numeric data
  - Often holds totals pre-calculated as well as base data
  - Not intended for textual data such as customer address lists

Some products store the data in a purpose-built multidimensional database, others hold it in a different form, such as relational tables, and construct the dimensions when the data is requested.
A multidimensional database stores all the "cube" data in a way that it can be regarded as multidimensional, not as separate files or tables. The data itself is numeric, the names of the dimensions and members are stored separately.
Another feature of an multidimensional database is that totals are often pre-calculated and stored with the data, not calculated when the data is fetched.



Multidimensional Views

Different selections give different ways of looking at the data

February for all products

All TV information

Audio | Jan | Feb

Video | Jan | Feb

TV		Jan		Feb	
		Actual	Budget	Actual	Budget
Sales	Paris				
	Moscow				
	London				
	Total				
Costs	Paris				
	Moscow				
	London				
	Total				

Viewing 5 dimensional database

Paris Sales and Costs

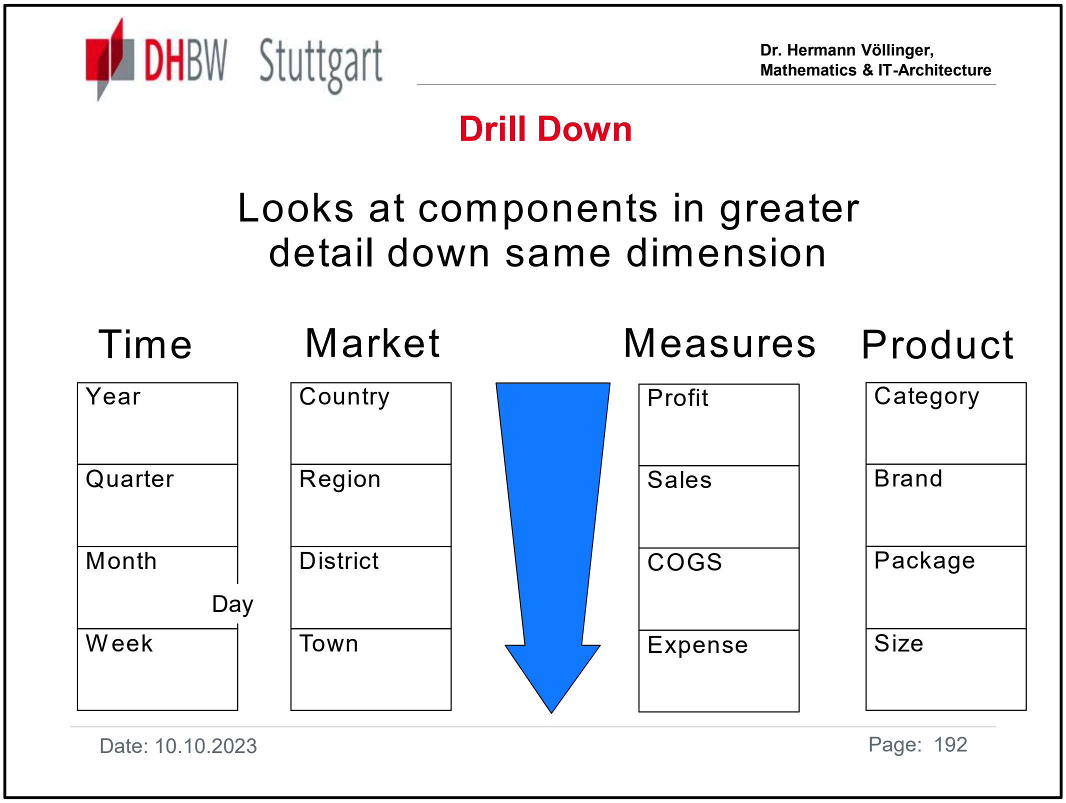
All cost information

Once data is in a cube, it is easy to get all possible combinations of dimensions.
These are multidimensional views of a database.

Here we see an example of a 5-dimensional database. Sales/Costs and Locations are **row** dimensions, budget/actual and time are **column** dimensions, while product (TV, Video, Audio) is a **page** dimension in this example.

They can of course be fetched in another combination of row, column and page.

Round the outside we can see the different selections can be made to show Paris only, all information for TV sales, cost information only, February budget information and so on.



Drill Down

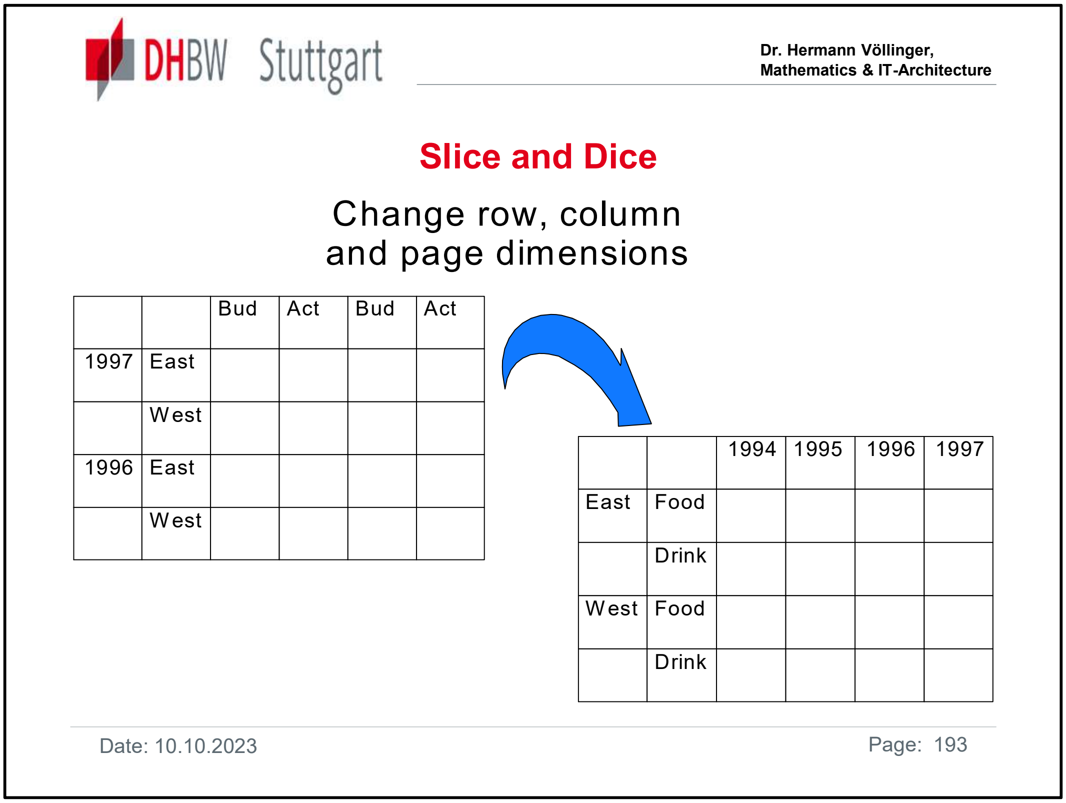
Looks at components in greater detail down same dimension

Time	Market	Measures	Product
Year	Country	Profit	Category
Quarter	Region	Sales	Brand
Month Day	District	COGS	Package
Week	Town	Expense	Size

A term commonly used with OLAP is **drill down.** It simply means going down to the next level of detail.

For example a drill down on year shows information by quarter, a drill down on quarter gives information for each month an so on.

The opposite is **drill up.**



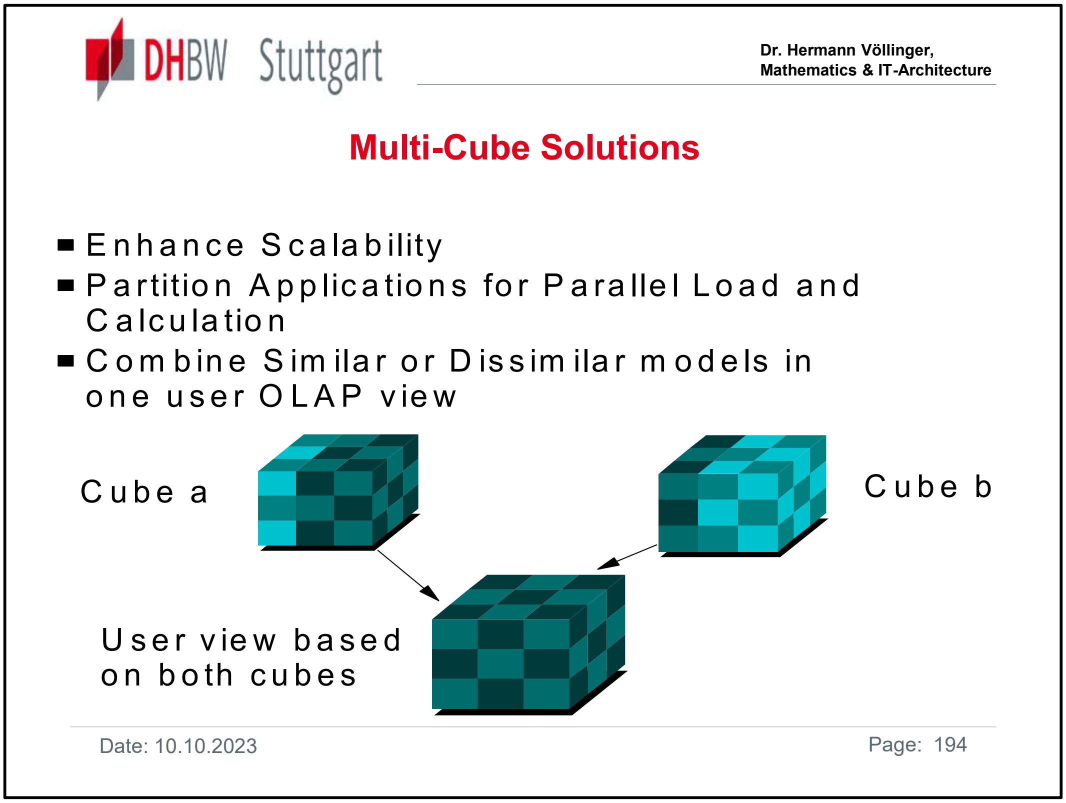
Slice and Dice

Change row, column
and page dimensions

		Bud	Act	Bud	Act
1997	East				
	West				
1996	East				
	West				

		1994	1995	1996	1997
East	Food				
	Drink				
West	Food				
	Drink				

Slice and Dice is another concept you may hear when discussing OLAP.
It is used to describe how multidimensional data can be shown in one way and then looked at in another way.
Effectively swapping page, column and row dimensions.
The action of moving an item from one position to another, for example moving Year displayed as a column dimension to display as a row dimension, is called **Pivot**



Multi-Cube Solutions

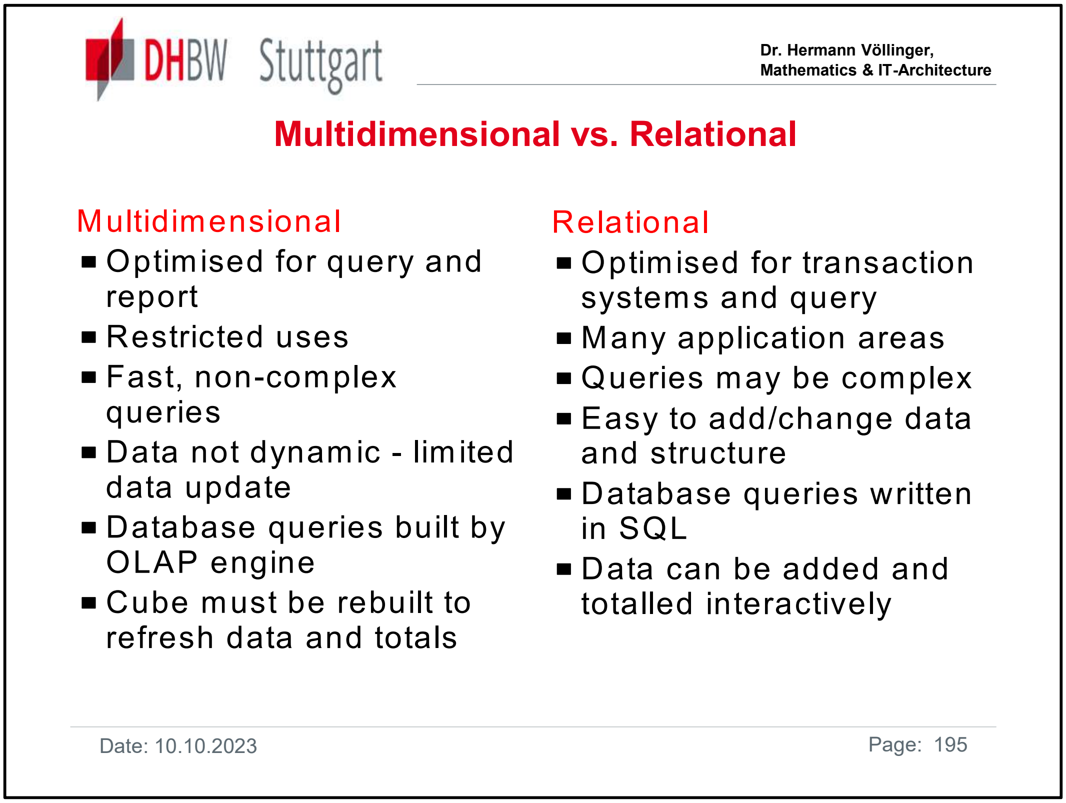
- Enhance Scalability
- Partition Applications for Parallel Load and Calculation
- Combine Similar or Dissimilar models in one user OLAP view

Cube a

Cube b

User view based on both cubes

Some OLAP solutions might use more than one cube of data.
For example, different product groups, countries, manufacturing plants or operating companies might have their own distinct cube.
These may be integrated for reporting at a higher level.
DB2 OLAP Server has facilities to partition applications across multiple cubes.



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Multidimensional vs. Relational

Multidimensional

- Optimised for query and report
- Restricted uses
- Fast, non-complex queries
- Data not dynamic - limited data update
- Database queries built by OLAP engine
- Cube must be rebuilt to refresh data and totals

Relational

- Optimised for transaction systems and query
- Many application areas
- Queries may be complex
- Easy to add/change data and structure
- Database queries written in SQL
- Data can be added and totalled interactively

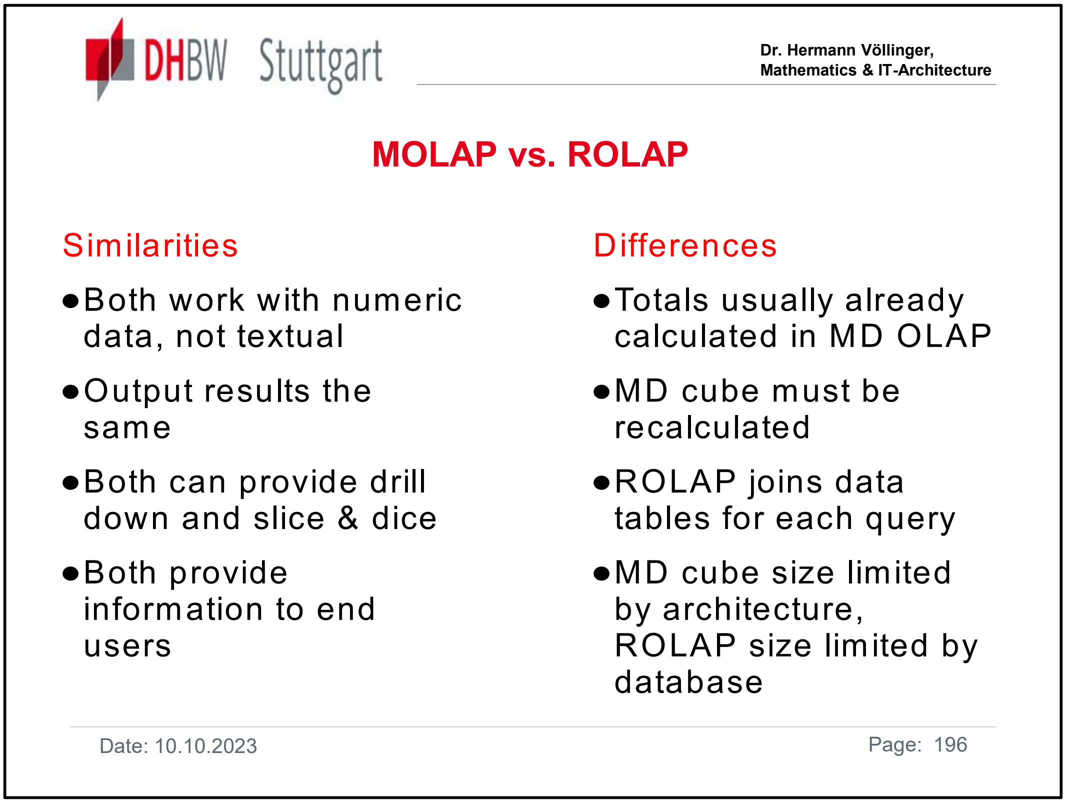
Date: 10.10.2023 Page: 195

OLAP tools do not indicate how the data is actually stored. Given that, it’s not surprising that there are multiple ways to store the data, including storing the data in a dedicated multidimensional database (also referred to as **MOLAP** or MDD). Examples include Arbors Software’s Essbase and Oracle Express Server. In a MOLAP environment, multi-dimensional "cubes" (often greater than 3 dimensions) are pre-calculated on the middle tier from one or more source systems and stored in a proprietary format. End-user queries are run against the cubes and not the underlying databases / RDBMS's.

The other choice involves storing the data in relational databases and having an OLAP tool work directly against the data, referred to as relational OLAP (also referred to as **ROLAP** or RDBMS).

Examples include MicroStrategy’s DSS server and related products, Informix’s Informix-MetaCube, Information Advantage’s Decision Suite, and Platinum Technologies’ Plantinum InfoBeacon. Some also include Red Brick’s Warehouse in this category, but it isn’t really an OLAP tool. Rather, it is a relational database optimised for performing the types of operations that ROLAP tools need.

In a ROLAP environment, SQL queries are sent to a database on an RDBMS, the result is returned to a middle-tier server for final cross-tabbing and formatting, and then sent to the client that initiated the query.



MOLAP vs. ROLAP

Similarities

- Both work with numeric data, not textual
- Output results the same
- Both can provide drill down and slice & dice
- Both provide information to end users

Differences

- Totals usually already calculated in MD OLAP
- MD cube must be recalculated
- ROLAP joins data tables for each query
- MD cube size limited by architecture, ROLAP size limited by database

The slide shows the most important **similarities and differences** between MOLAP and ROLAP concept:

Usually, a scalable, parallel database is used for the large, atomic. Organizationally structured data warehouse, and subsets or summarized data from the warehouse are extracted and replicated to proprietary MDDs.

Because MDD vendors have enabled drill-through features, when a user reaches the limit of what is actually stored in the MDD and seeks more detail data, he/she can drill through to the detail stored in the enterprise database. However, the drill through functionality usually requires creating views for every possible query.

As relational database vendors incorporate sophisticated analytical multidimensional features into their core database technology, the resulting capacity for higher performance saleability and parallelism will enable more sophisticated analysis.

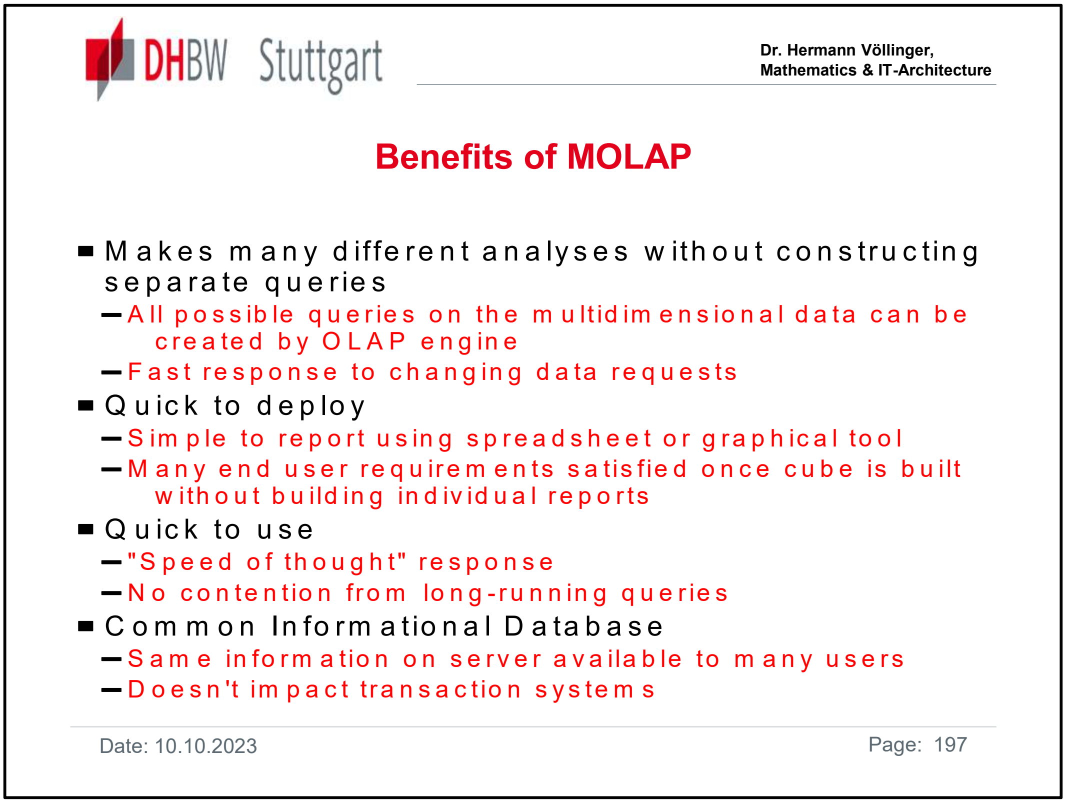
Proprietary database and non-integrated relational OLAP query tool vendors will find it difficult to compete with this integrated ROLAP solution.

Both storage methods have strengths and weaknesses -- the weaknesses, however, are being rapidly addressed by the respective vendors.

Currently, data warehouses are predominantly built using RDBMSs. If you have a warehouse built on a relational database and you want to perform OLAP analysis against it, ROLAP is a natural fit.

This isn’t to say that MDDs can’t be a part of your data warehouse solution. It’s just that MDDs aren’t currently well-suited for large volumes of data (10-50GB is fine, but anything over 50GB is stretching their capabilities).

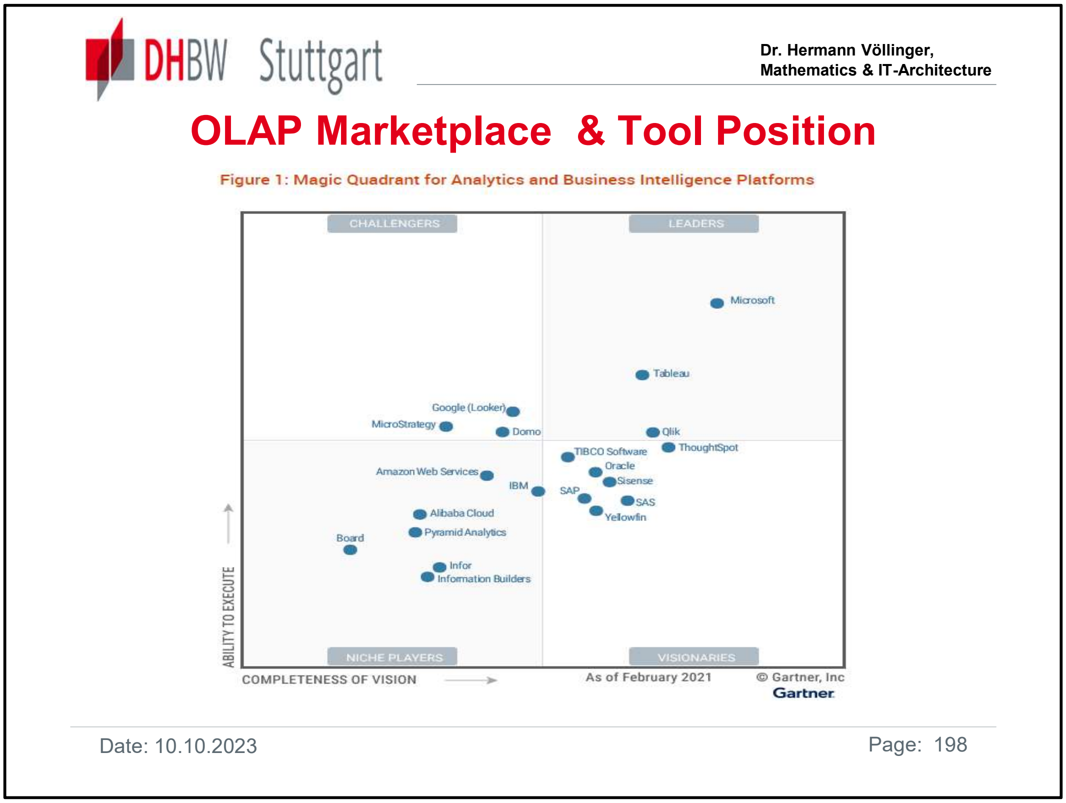
If your really want the functionality benefits that come with MDD, consider sub-setting the data into smaller MDD-based data marts.



Benefits of MOLAP

- Makes many different analyses without constructing separate queries
  - All possible queries on the multidimensional data can be created by OLAP engine
  - Fast response to changing data requests
- Quick to deploy
  - Simple to report using spreadsheet or graphical tool
  - Many end user requirements satisfied once cube is built without building individual reports
- Quick to use
  - "Speed of thought" response
  - No contention from long-running queries
- Common Informational Database
  - Same information on server available to many users
  - Doesn't impact transaction systems

Although there is an initial effort in building a cube and calculating
any stored totals, the main benefit is that once it is made, everything
else is easy and quick.
All possible reports can on any selection and any combination without
Developing new reports extractions, the queries are very fast and it is easy
to add new users.



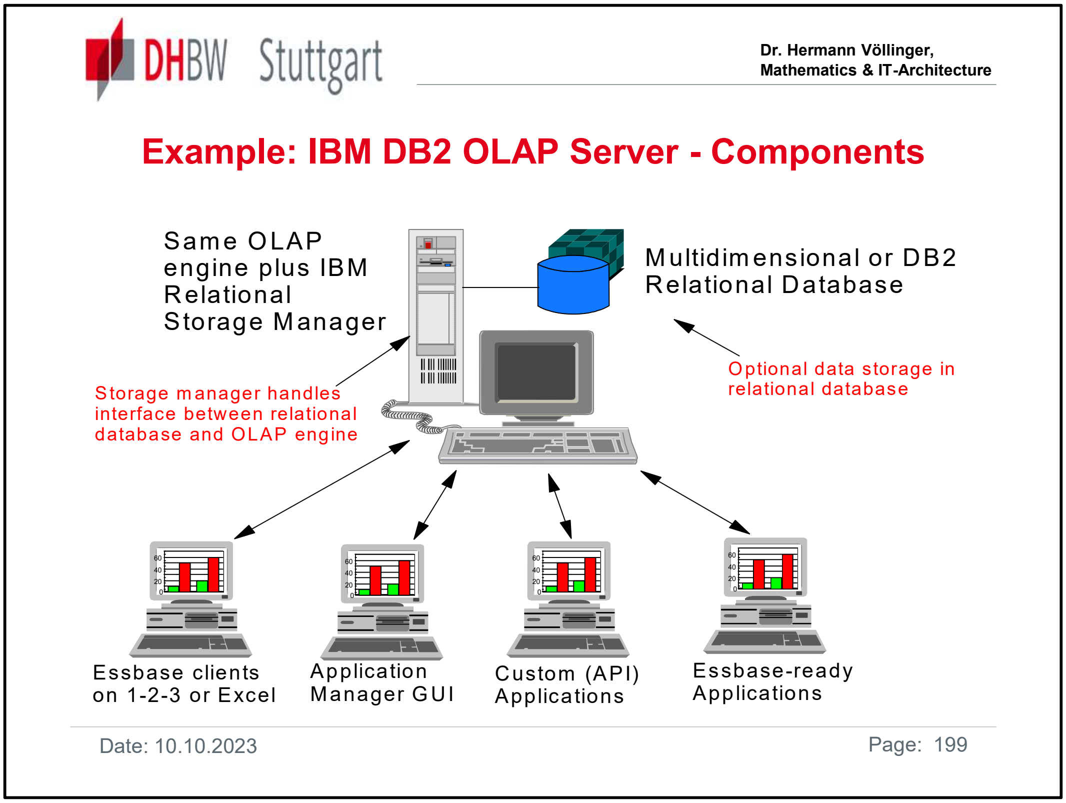
OLAP Marketplace & Tool Position

Figure 1: Magic Quadrant for Analytics and Business Intelligence Platforms

See also in [DHBW-Moodle] the document: **Gartner_Quadrant-OLAP_Tools-Feb2021.pdf :** Analytics and business intelligence (ABI) platforms are characterized by easy-to-use functionality that supports a full analytic workflow — from data preparation to visual exploration and insight generation — with an emphasis on self-service usage and augmented user assistance.

Vendors in the ABI market range from startups backed by venture capital funds to large technology firms. The vast majority of new spending in this market is on cloud deployments, and major cloud platform players are present in the market. In many cases, ABI platforms are entry points for wider sets of cloud data management capabilities offered by these cloud vendors, examples being Microsoft Azure Synapse Analytics and IBM Cloud Pak for Data.

ABI platforms are no longer differentiated by their data visualization capabilities, which are now commoditized. All vendors can build interactive key performance indicator (KPI) dashboards using common chart forms (bar/column, line/area, scatter, pie and geographic maps) and drawing on a wide range of data sources. Differentiation has shifted to how well platforms support augmented analytics. Augmentation utilizes machine learning (ML) and artificial intelligence (AI)-assisted data preparation, insight generation and insight explanation to help business people and data analysts explore and analyze data more effectively than they could manually. …



Example: IBM DB2 OLAP Server - Components

Same OLAP engine plus IBM Relational Storage Manager

Multidimensional or DB2 Relational Database

Storage manager handles interface between relational database and OLAP engine

Optional data storage in relational database

Essbase clients on 1-2-3 or Excel

Application Manager GUI

Custom (API) Applications

Essbase-ready Applications

The Hyperion Essbase product family includes the following feature sets:

- **Hyperion Essbase Application Manager**

A graphical environment for developing and maintaining Hyperion Essbase applications. Tasks include building outlines and dimensions, performing data loads and calculations, and defining security access.

- **Hyperion Essbase OLAP Server**

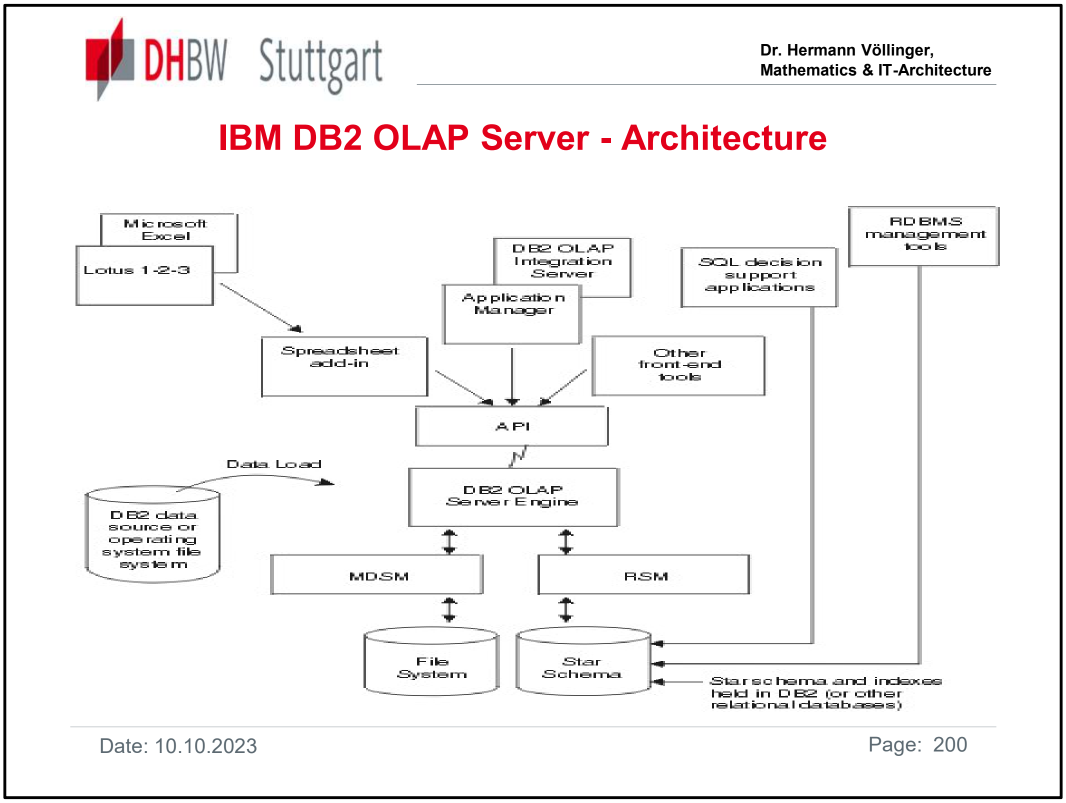
A multidimensional database for storing data with an unlimited number of dimensions, such as time, accounts, region, channel, or product. The Hyperion Essbase server manages analytical data models, data storage, calculations, and data security.

- **Hyperion Essbase Spreadsheet Add-in**

Desktop software enables analysis of the data stored in the Hyperion Essbase server. Hyperion Essbase Spreadsheet Add-in is seamlessly integrated with Microsoft Excel or Lotus 1-2-3 spreadsheets.

- **Hyperion Essbase application tools**

A suite of tools for extending Hyperion Essbase applications. These tools include Hyperion Essbase Currency Conversion, Hyperion Essbase SQL Interface, Hyperion Essbase Spreadsheet Toolkit, and Hyperion Essbase API.



IBM DB2 OLAP Server - Architecture

Microsoft Excel
Lotus 1-2-3
DB2 OLAP Integration Server
Application Manager
SQL decision support applications
RDBMS management tools
Spreadsheet add-in
Other front-end tools
API
Data Load
DB2 OLAP Server Engine
DB2 data source or operating system file system
MDSM
RSM
File System
Star Schema
Star schema and indexes held in DB2 (or other relational databases)

The above figure shows the main components in the **DB2 OLAP Server** environment.

The IBM DB2 OLAP Server is an online analytical processing (OLAP) product that you can use to create a wide range of multidimensional planning, analysis, and reporting applications.

DB2 OLAP Server is based on the OLAP technology that was developed by Hyperion Solutions Corporation.

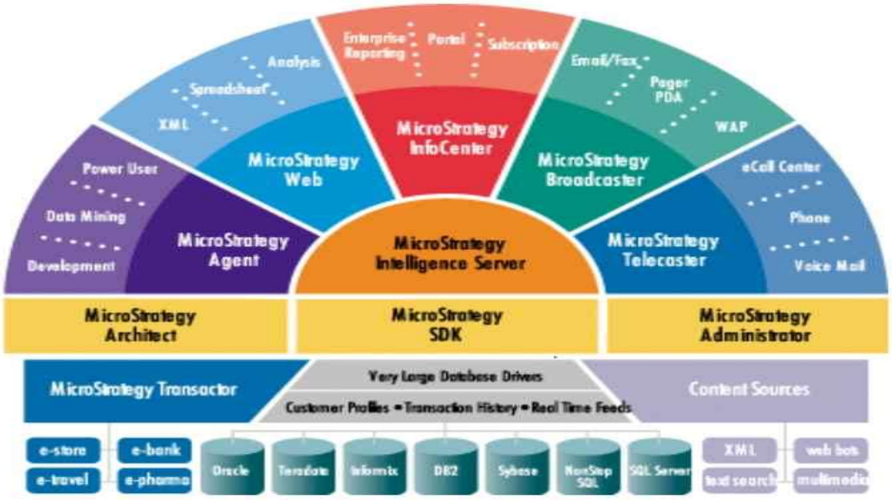
DB2 OLAP Server includes all of the capabilities of Hyperion Essbase.

In addition, it offers the option of storing multidimensional databases as sets of relational tables.

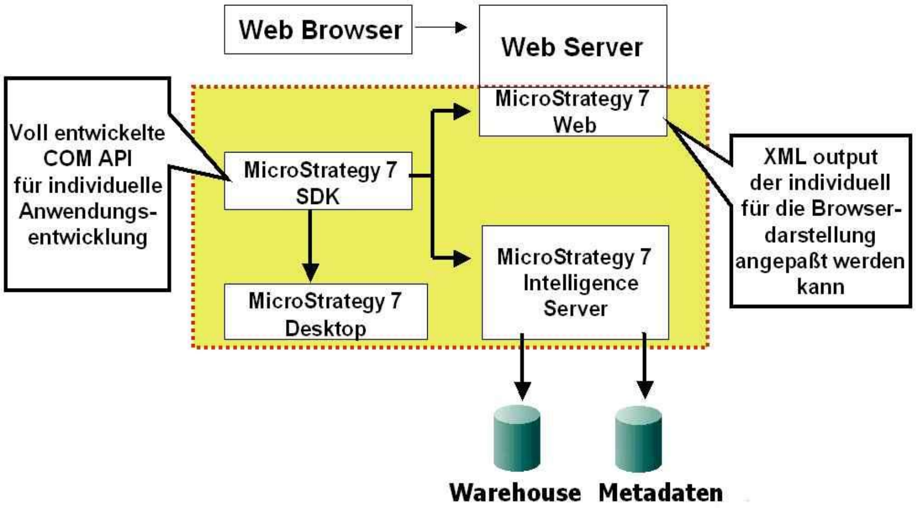
Regardless of the storage management option that you choose, you can use the Essbase Application Manager and Essbase commands to create an Essbase application and its associated databases.

You can also use over 70 Essbase-ready tools provided by independent software vendors that can access multidimensional databases transparently.

ROLAP / MicroStrategy: Components Overview



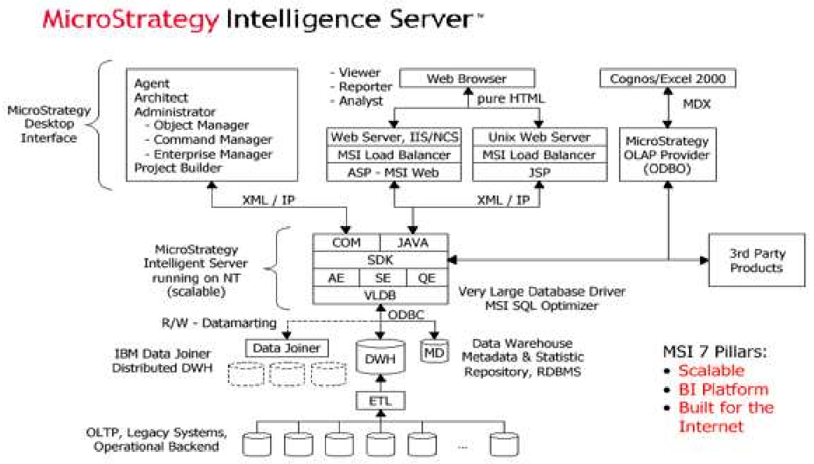
ROLAP Example - MicroStrategy: Analytical Model



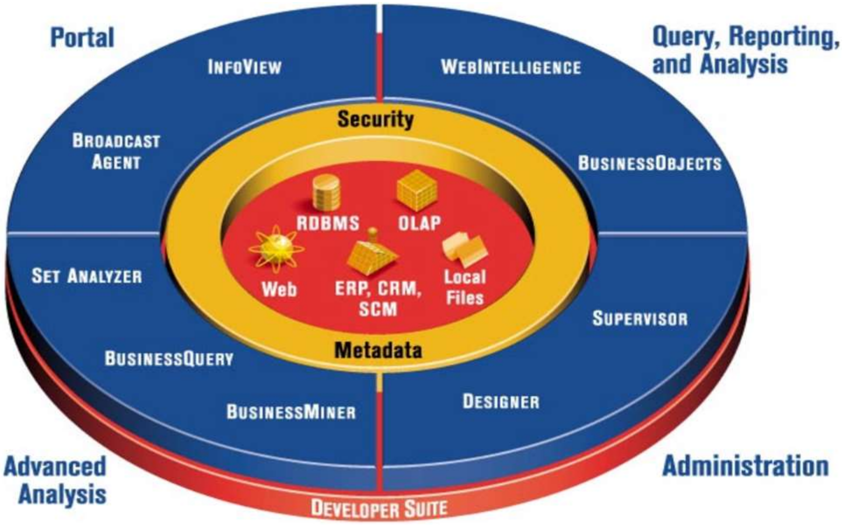
Voll entwickelte COM API für individuelle Anwendungsentwicklung

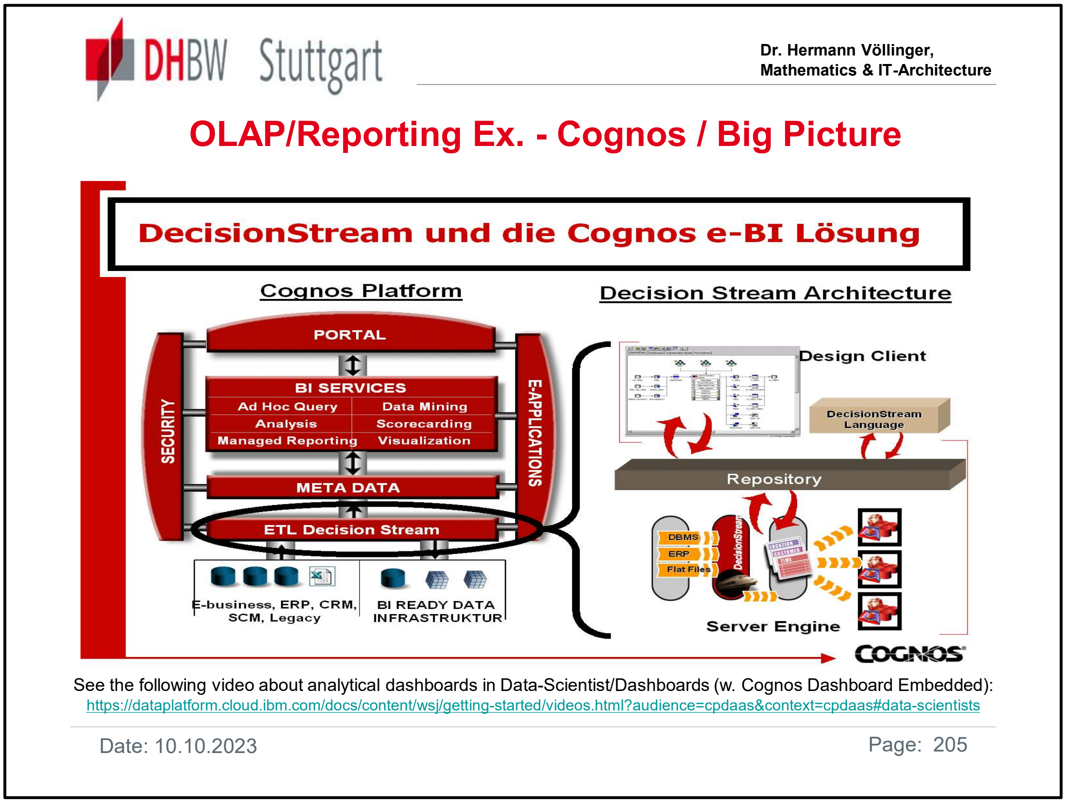
XML output der individuell für die Browserdarstellung angepaßt werden kann

ROLAP Example - MicroStrategy: Big Picture



OLAP/Reporting Ex. - BusinessObject /Big Picture





DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

OLAP/Reporting Ex. - Cognos / Big Picture

See the following video about analytical dashboards in Data-Scientist/Dashboards (w. Cognos Dashboard Embedded): https://dataplatform.cloud.ibm.com/docs/content/wsj/getting-started/videos.html?audience=cpdaas&context=cpdaas#data-scientists

Date: 10.10.2023 Page: 205

Demos with Cognos Analytics 11.1.4:

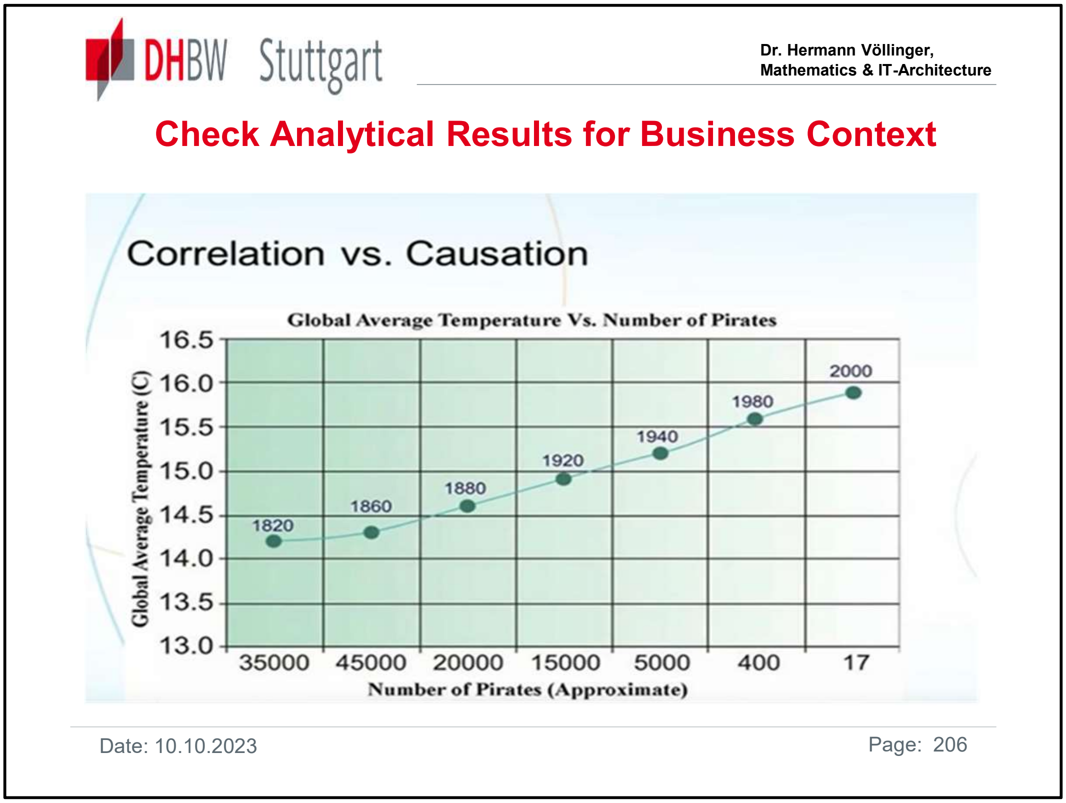
See a YouTube video about an overview of the features of Cognos Analytics: IBM Cognos Analytics – Demos or Cognos Analytics 11.1 Online Training For Beginners - Complete System Tutorial | Data Analysts - YouTube

Cognos Demo about dashboards (a registered IBM Cloud Id is needed!): https://www.ibm.com/cloud/garage/dte/tutorial/ibm-cognos-analytics-1114-getting-started-dashboards

See also in course moodle **[DHBW-Moodle**] under: **'Directory of supporting Information for the DWH Lecture'** the documents:

- Guided Cognos Demo+Tutorial - Short Version.pdf
- Guided Cognos Demo+Tutorial - Long Version.pdf

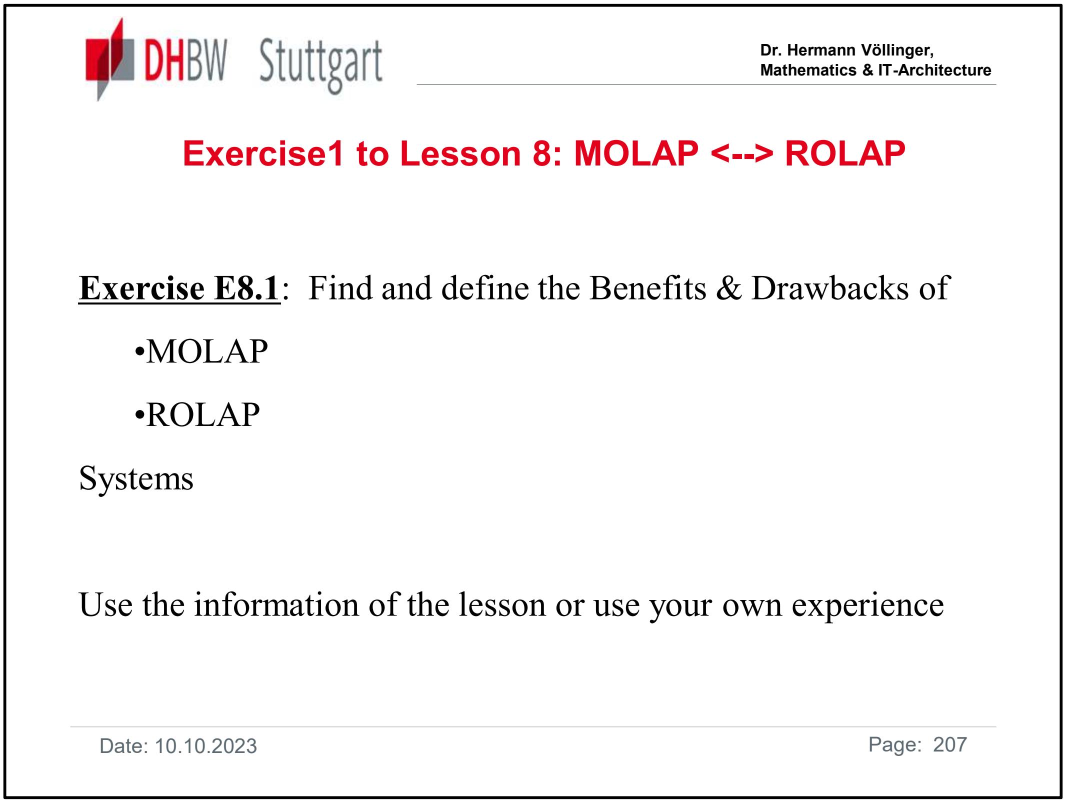
What's next: For more information about Cognos Analytics, see the product page: Cognos Analytics - Overview | IBM or register for a free trial: Sign up for IBM Cognos Analytics



Check Analytical Results for Business Context

Correlation vs. Causation

The diagram (example) suggests that the temperature of the ocean depends on the number of pirates on the oceans. One might come to the conclusion: "Let's increase the number of pirates and the temperature of the oceans will drop again".
It is therefore always to check by business experts, which metrics are only in a correlation and which have a causal business context.



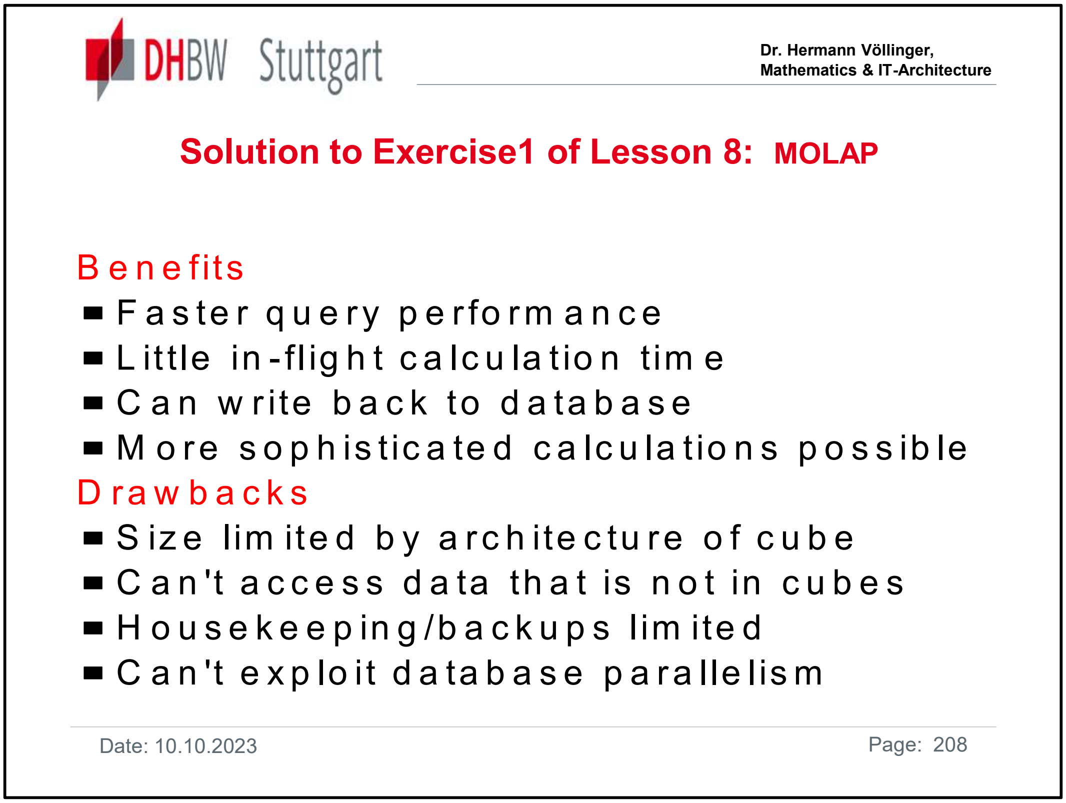
Exercise1 to Lesson 8: MOLAP <--> ROLAP

Exercise E8.1: Find and define the Benefits & Drawbacks of

- MOLAP
- ROLAP

Systems

Use the information of the lesson or use your own experience



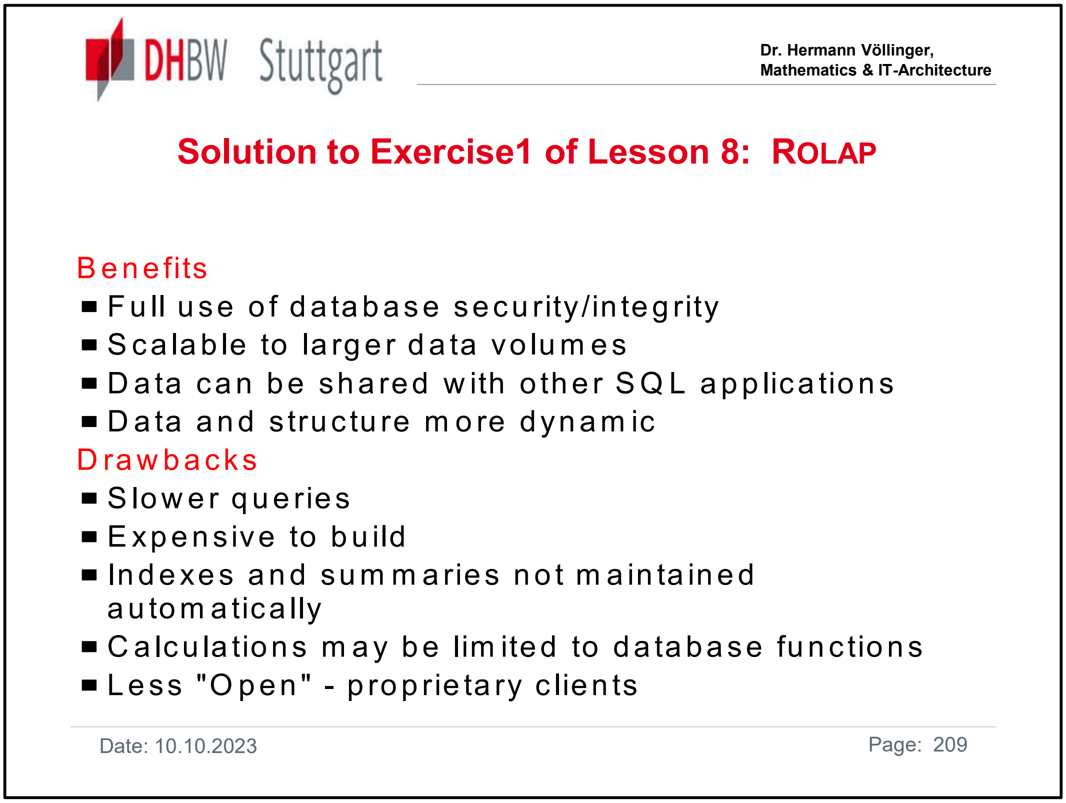
Solution to Exercise1 of Lesson 8: MOLAP

Benefits

- Faster query performance
- Little in-flight calculation time
- Can write back to database
- More sophisticated calculations possible

Drawbacks

- Size limited by architecture of cube
- Can't access data that is not in cubes
- Housekeeping/backups limited
- Can't exploit database parallelism



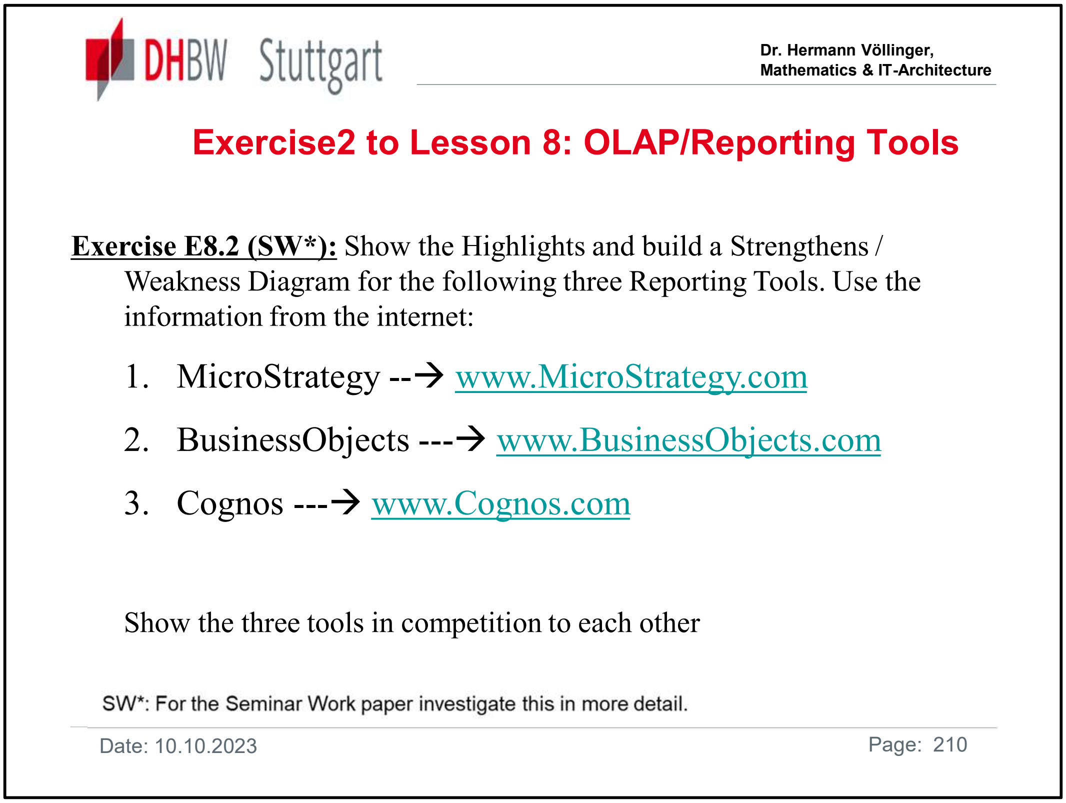
Solution to Exercise1 of Lesson 8: ROLAP

Benefits

- Full use of database security/integrity
- Scalable to larger data volumes
- Data can be shared with other SQL applications
- Data and structure more dynamic

Drawbacks

- Slower queries
- Expensive to build
- Indexes and summaries not maintained automatically
- Calculations may be limited to database functions
- Less "Open" - proprietary clients



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise2 to Lesson 8: OLAP/Reporting Tools

Exercise E8.2 (SW*): Show the Highlights and build a Strengthens / Weakness Diagram for the following three Reporting Tools. Use the information from the internet:

1. MicroStrategy --→ www.MicroStrategy.com
2. BusinessObjects ---→ www.BusinessObjects.com
3. Cognos ---→ www.Cognos.com

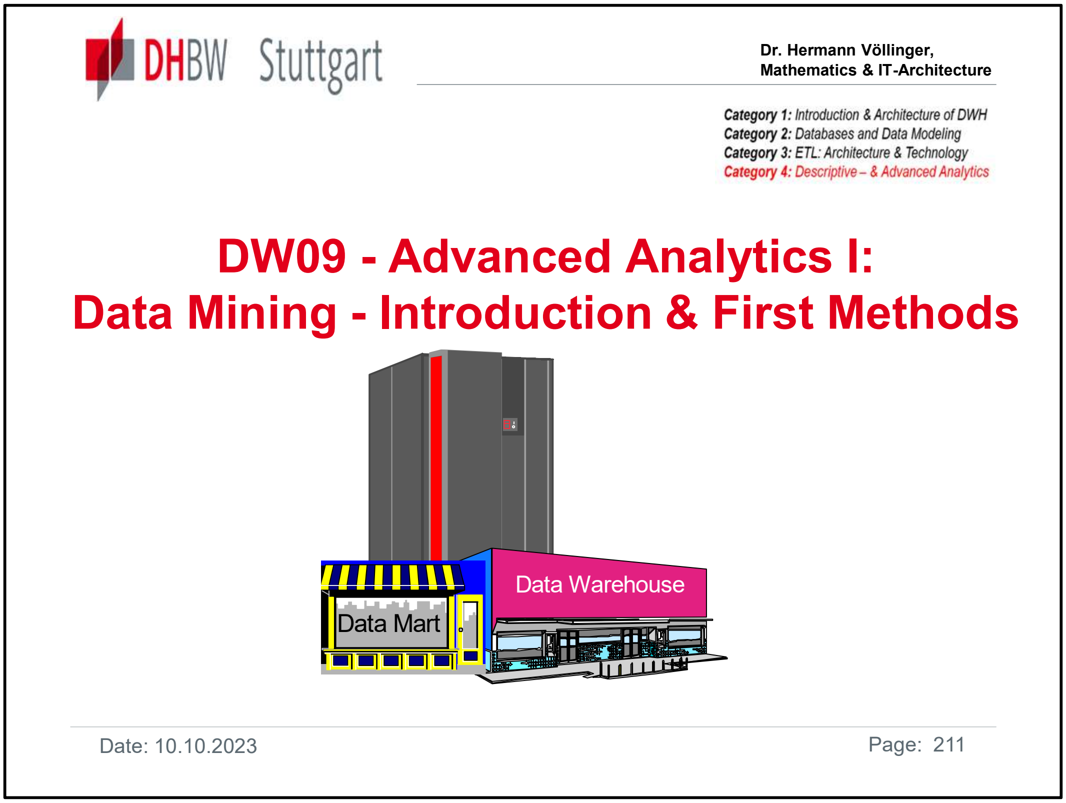
Show the three tools in competition to each other

SW*: For the Seminar Work paper investigate this in more detail.

Date: 10.10.2023 Page: 210

Reporting and analytics are critical components of a Business Intelligence & Analytics based approach to **corporate management (“Unternehmenssteuerung”)**. For more information about such reporting requirements see for example in [DHBW-Moodle] the following TDWI Article:

- TDWI_Artikel-Intelligentes-Reporting.pdf



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Category 1: *Introduction & Architecture of DWH*
Category 2: *Databases and Data Modeling*
Category 3: *ETL: Architecture & Technology*
Category 4: *Descriptive – & Advanced Analytics*

DW09 - Advanced Analytics I: Data Mining - Introduction & First Methods

Date: 10.10.2023

Page: 211

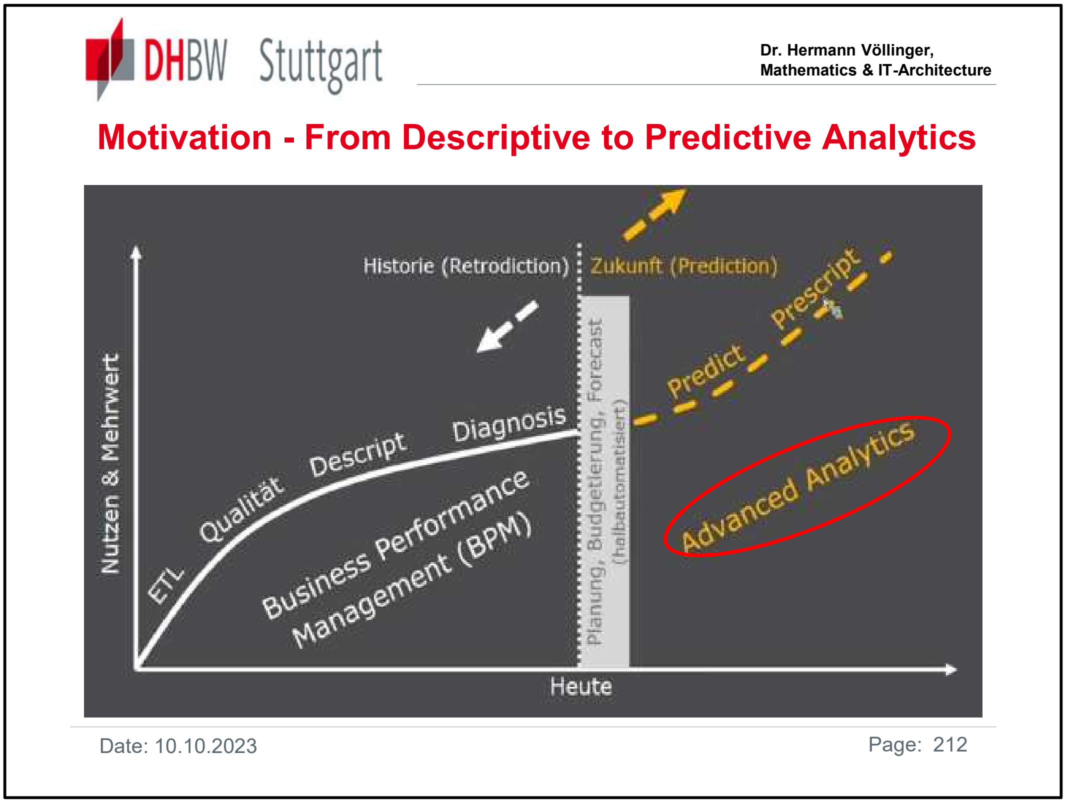
The following two chapters gives an introduction to the methods and techniques of Data Mining as part of the overall architecture of a data warehouse. It explains the architectural ideas behind a Data Mining solution and shows for what industries and applications Data Mining can be used.

Well known applications, which use Data Mining techniques, are:

1. Market Basket Analysis (MBA)
2. Cross Selling
3. Customer Retention
4. Fraud Detection
5. Campaign Management

The following video shows the features of the Dataiku DSS Platform. It shows an **end-to-demo example for the creation of an advanced analytical solution**. In especially you see steps to create a prediction-report “Flight Delays”. The report calculates probabilities for flight delays: https://content.dataiku.com/dataiku9-general/2021-dataiku-demo?utm_campaign=CONTENT+Dataiku+9+March+2021

The following document shows more details about the Dataiku platform. See in [DHBW-Moodle] the document “Dataiku-Platform_Overview.pdf”



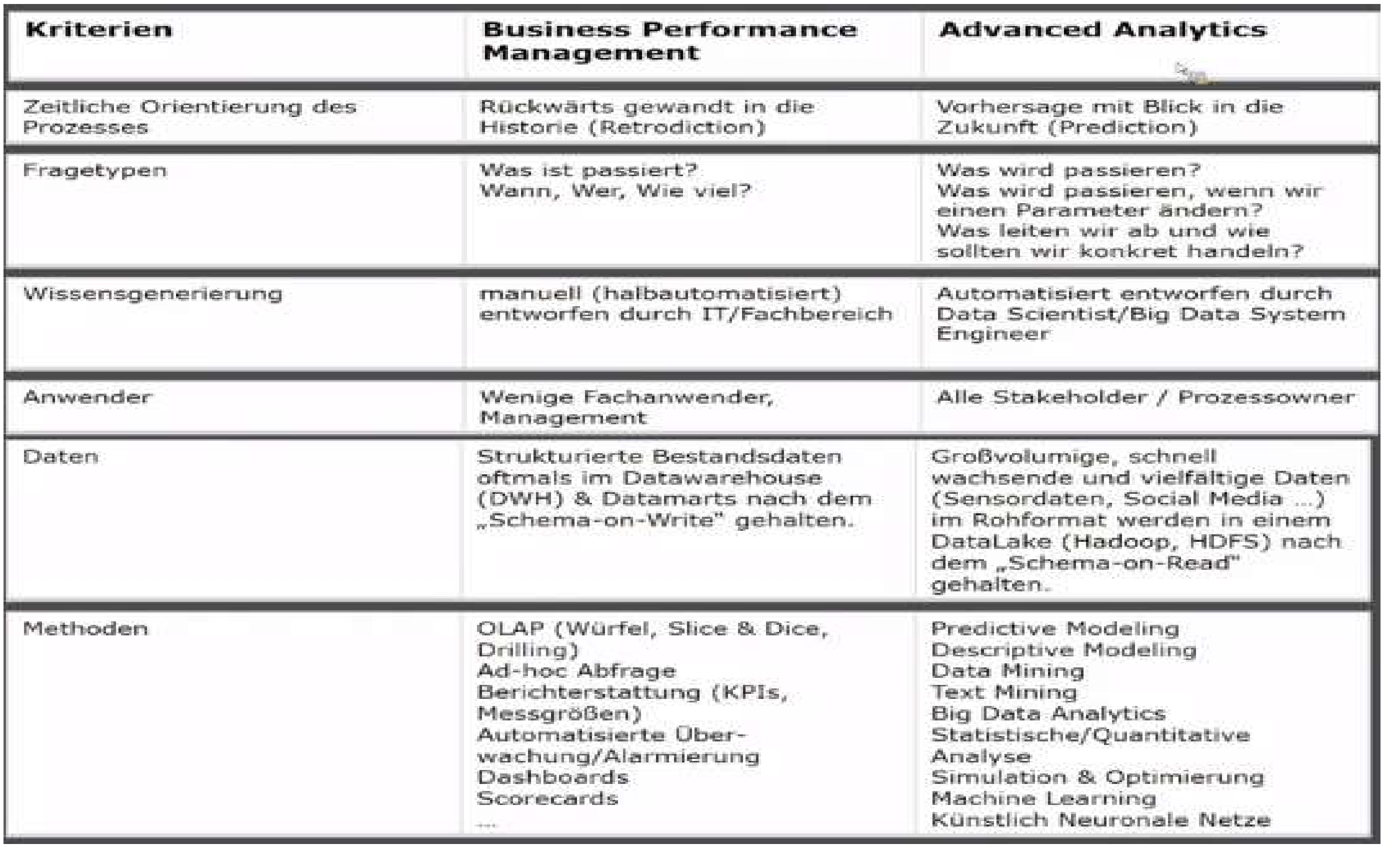
DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

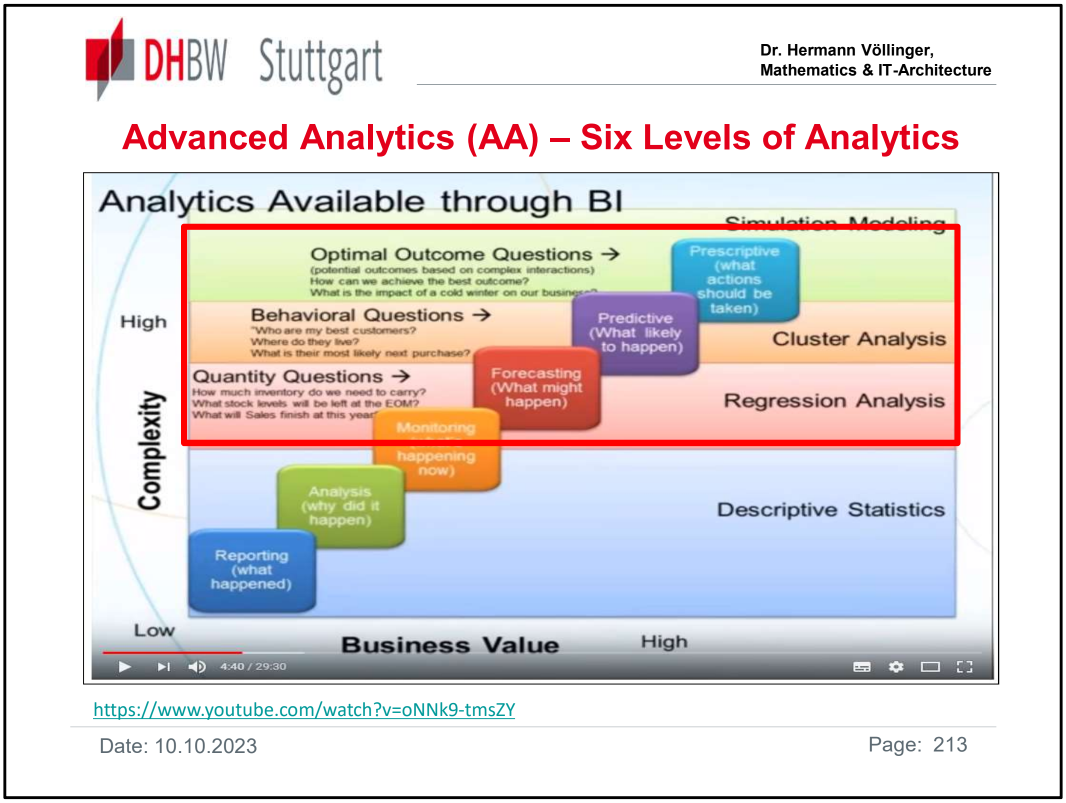
Motivation - From Descriptive to Predictive Analytics

Date: 10.10.2023 Page: 212

Business Performance Management (BPM) versus Advanced Analytics (AA):



Kriterien	Business Performance Management	Advanced Analytics
Zeitliche Orientierung des Prozesses	Rückwärts gewandt in die Historie (Retrodiction)	Vorhersage mit Blick in die Zukunft (Prediction)
Fragetypen	Was ist passiert? Wann, Wer, Wie viel?	Was wird passieren? Was wird passieren, wenn wir einen Parameter ändern? Was leiten wir ab und wie sollten wir konkret handeln?
Wissensgenerierung	manuell (halbautomatisiert) entworfen durch IT/Fachbereich	Automatisiert entworfen durch Data Scientist/Big Data System Engineer
Anwender	Wenige Fachanwender, Management	Alle Stakeholder / Prozessowner
Daten	Strukturierte Bestandsdaten oftmals im Datawarehouse (DWH) & Datamarts nach dem „Schema-on-Write" gehalten.	Großvolumige, schnell wachsende und vielfältige Daten (Sensordaten, Social Media ...) im Rohformat werden in einem DataLake (Hadoop, HDFS) nach dem „Schema-on-Read" gehalten.
Methoden	OLAP (Würfel, Slice & Dice, Drilling) Ad-hoc Abfrage Berichterstattung (KPIs, Messgrößen) Automatisierte Überwachung/Alarmierung Dashboards Scorecards ...	Predictive Modeling Descriptive Modeling Data Mining Text Mining Big Data Analytics Statistische/Quantitative Analyse Simulation & Optimierung Machine Learning Künstlich Neuronale Netze



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

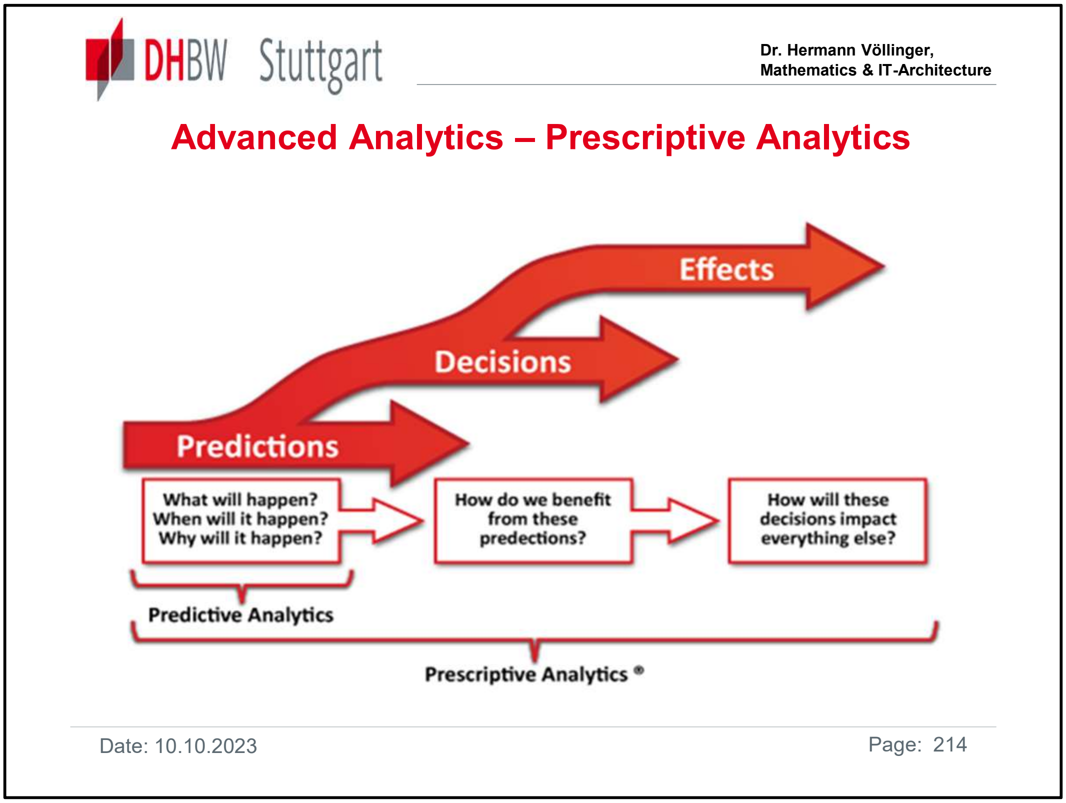
Advanced Analytics (AA) – Six Levels of Analytics

Analytics Available through BI
Simulation Modeling
Optimal Outcome Questions →
(potential outcomes based on complex interactions)
How can we achieve the best outcome?
What is the impact of a cold winter on our business?
Prescriptive (what actions should be taken)
High
Behavioral Questions →
"Who are my best customers?
Where do they live?
What is their most likely next purchase?
Predictive (What likely to happen)
Cluster Analysis
Quantity Questions →
How much inventory do we need to carry?
What stock levels will be left at the EOM?
What will Sales finish at this year?
Forecasting (What might happen)
Regression Analysis
Complexity
Monitoring (what's happening now)
Analysis (why did it happen)
Descriptive Statistics
Reporting (what happened)
Low
Business Value
High
4:40 / 29:30

https://www.youtube.com/watch?v=oNNk9-tmsZY

Date: 10.10.2023 Page: 213

1. Reporting: All companies regularly use reporting tools to report on their current sales and sales figures. These procedures are standard and do not add much new value compared to the competition. Nevertheless, this procedure is proven and sometimes mandatory for the reporting system.
2. **Analysis:** analysis of the departments and the questions why certain numbers have emerged are common today. Here comes the knowledge of the local parameters and good knowledge of the sales field or sales experience to fruition.
3**. Monitoring:** The monitoring also includes a rating and thus offers:• information about what an intervention does, how well it works, and whether it achieves its goals;• Guidelines for future intervention activities;• an important part of accountability to funding agencies and stakeholders.
Monitoring and evaluation plans should be prepared at the beginning of an intervention development process.
4. **Forecasting:** Forecasts are predictions for the future that are based on past and present data and are mostly generated by analyzing trends. A banking example could be the estimate of the probability of a loan application closing at a future point in time. Prediction is a similar but more general term. Both can relate to formal statistical methods using time series, cross-sectional or longitudinal data, or alternatively to less formal assessment methods. Risk and uncertainty are central to forecasts and forecasts. It is generally considered good practice to indicate the degree of uncertainty of the forecasts.
5**. Predictive Analysis:** A prediction is a statement about a future event. A prediction is often, but not always, based on experience or knowledge. Although future events are necessarily uncertain, so that guaranteed accurate information about the future is often not possible, the forecast can be helpful in making plans for possible developments.
6. **Prescriptive Analysis**: The prescriptive analytic termed the "ultimate limit of analytical ability" involves the application of mathematical and computer science and proposes decision-making options to exploit the results of descriptive and predictive analytics.



Advanced Analytics – Prescriptive Analytics

Effects

Decisions

Predictions

What will happen?
When will it happen?
Why will it happen?

How do we benefit
from these
predections?

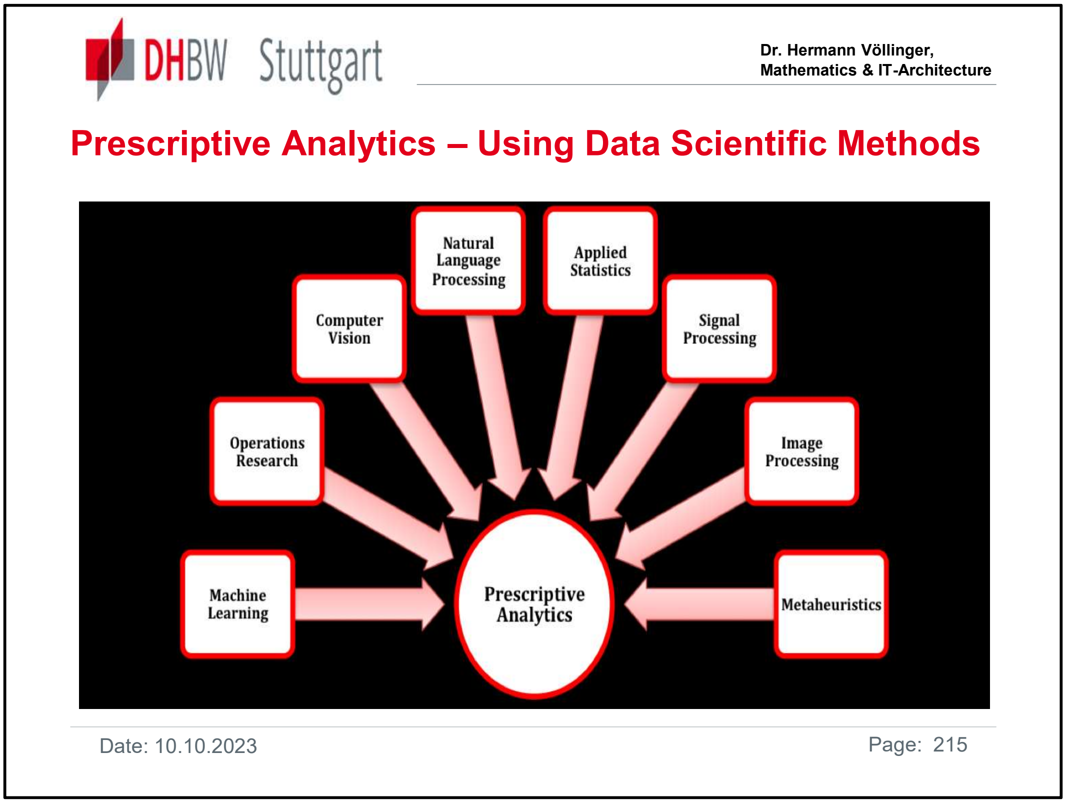
How will these
decisions impact
everything else?

Predictive Analytics

Prescriptive Analytics ®

Prescriptive analytics incorporates both structured and unstructured data and uses a combination of advanced analytic techniques and disciplines to predict, prescribe, and adapt. While the term prescriptive analytics was first coined by IBM and later trademarked by Ayata, the underlying concepts have been around for hundreds of years. The technology behind prescriptive analytics synergistically combines hybrid data, business rules with mathematical models and computational models. The data inputs to prescriptive analytics may come from multiple sources: internal, such as inside a corporation; and external, also known as environmental data. The data may be structured, which includes numbers and categories, as well as unstructured data, such as texts, images, sounds, and videos. Unstructured data differs from structured data in that its format varies widely and cannot be stored in traditional relational databases without significant effort at data transformation.[10] More than 80% of the world's data today is unstructured, according to IBM.

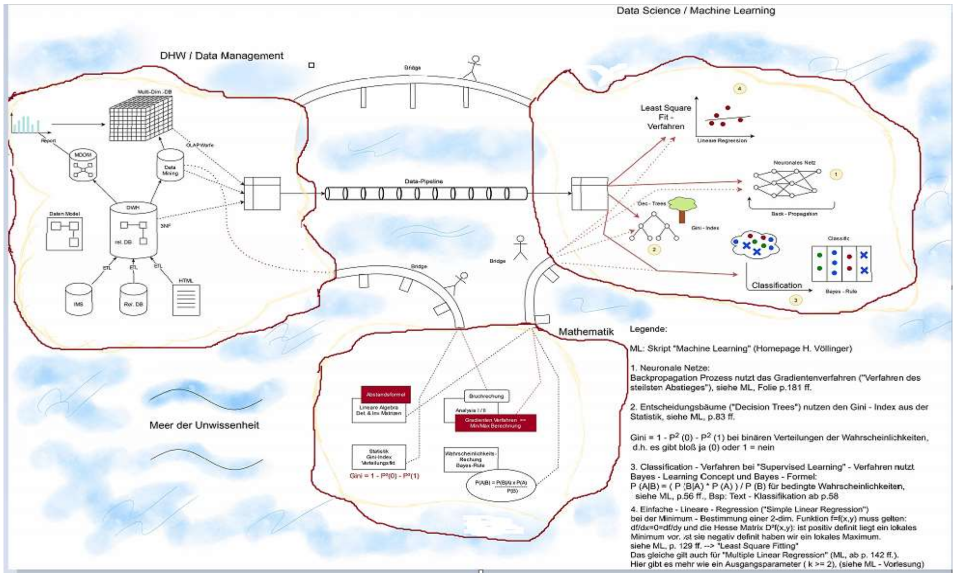
Prescriptive Analytics extends beyond predictive analytics by specifying both the actions necessary to achieve predicted outcomes, and the interrelated effects of each decision



Prescriptive Analytics – Using Data Scientific Methods

In addition to this variety of data types and growing data volume, incoming data can also evolve with respect to velocity, that is, more data being generated at a faster or a variable pace. Business rules define the business process and include objectives constraints, preferences, policies, best practices, and boundaries. Mathematical models and computational models are techniques derived from mathematical sciences, computer science and related disciplines such as applied statistics, machine learning, operations research, natural language processing, computer vision, pattern recognition, image processing, speech recognition, and signal processing. The correct application of all these methods and the verification of their results implies the need for resources on a massive scale including human, computational and temporal for every Prescriptive Analytic project. In order to spare the expense of dozens of people, high performance machines and weeks of work one must consider the reduction of resources and therefore a reduction in the accuracy or reliability of the outcome. The preferable route is a reduction that produces a probabilistic result within acceptable limits.

“Bridges”: DWH/Data M.-Mathematics-ML/Data Science



Legende:

ML: Skript "Machine Learning" (Homepage H. Völlinger)

1. Neuronale Netze:
Backpropagation Prozess nutzt das Gradientenverfahren ("Verfahren des steilsten Abstieges"), siehe ML, Folie p.181 ff.

2. Entscheidungsbäume ("Decision Trees") nutzen den Gini - Index aus der Statistik, siehe ML, p.83 ff.

Gini = 1 - P^2 (0) - P^2 (1) bei binären Verteilungen der Wahrscheinlichkeiten, d.h. es gibt bloß ja (0) oder 1 = nein

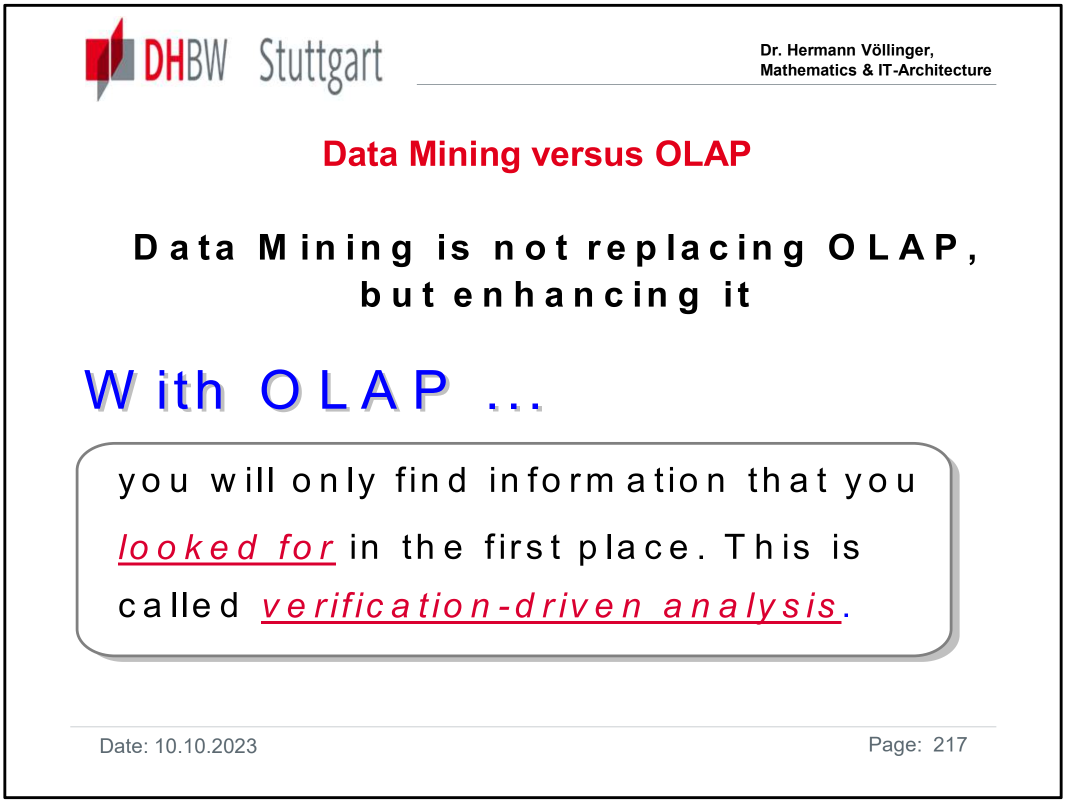
3. Classification - Verfahren bei "Supervised Learning" - Verfahren nutzt Bayes - Learning Concept und Bayes - Formel:
P (A|B) = (P (B|A) * P (A)) / P (B) für bedingte Wahrscheinlichkeiten, siehe ML, p.56 ff., Bsp: Text - Klassifikation ab p.58

4. Einfache - Lineare - Regression ("Simple Linear Regression")
bei der Minimum - Bestimmung einer 2-dim. Funktion f=f(x,y) muss gelten: df/dx=0=df/dy und die Hesse Matrix $D^2f(x,y)$: ist positiv definit liegt ein lokales Minimum vor. ist sie negativ definit haben wir ein lokales Maximum.
siehe ML, p. 129 ff. --> "Least Square Fitting"
Das gleiche gilt auch für "Multiple Linear Regression" (ML, ab p. 142 ff.).
Hier gibt es mehr wie ein Ausgangsparameter (k >= 2), (siehe ML - Vorlesung)

Date: 10.10.2023 Page: 216

Mathematics is a sow of Machine Learning. About 250 years ago, it began to be formalized, but it was only about 100 years ago that modern mathematics was developed.
It is a huge field, with many subfields, such as linear algebra, statistics, etc., and with applications in engineering and physics.
Fortunately, we do not need to know the full breadth and depth of mathematics, to improve our understanding and application of machine learning. The main areas of mathematics that will be of interest for ML, are:

1. Linear Algebra for Machine Learning (ML)
2. Probability and statistical methods for ML
3. Optimization methods for ML
4. Analysis (i.e. "gradient methods") for ML



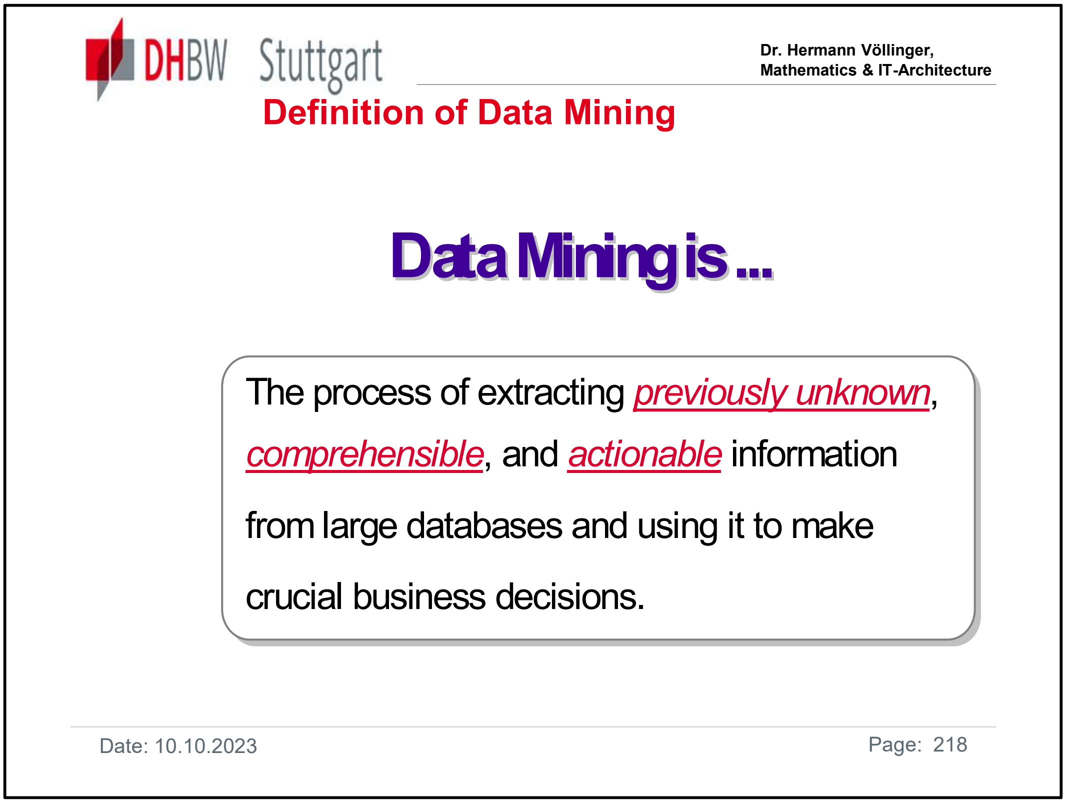
Data Mining versus OLAP

Data Mining is not replacing OLAP, but enhancing it

With OLAP ...

you will only find information that you *looked for* in the first place. This is called *verification-driven analysis*.

OLAP and Data Mining are not replacing each other. Each of them can enhance the other technology.
While OLAP do analysis of existing data and facts, Data Mining creates new knowledge and new information for the decision process.

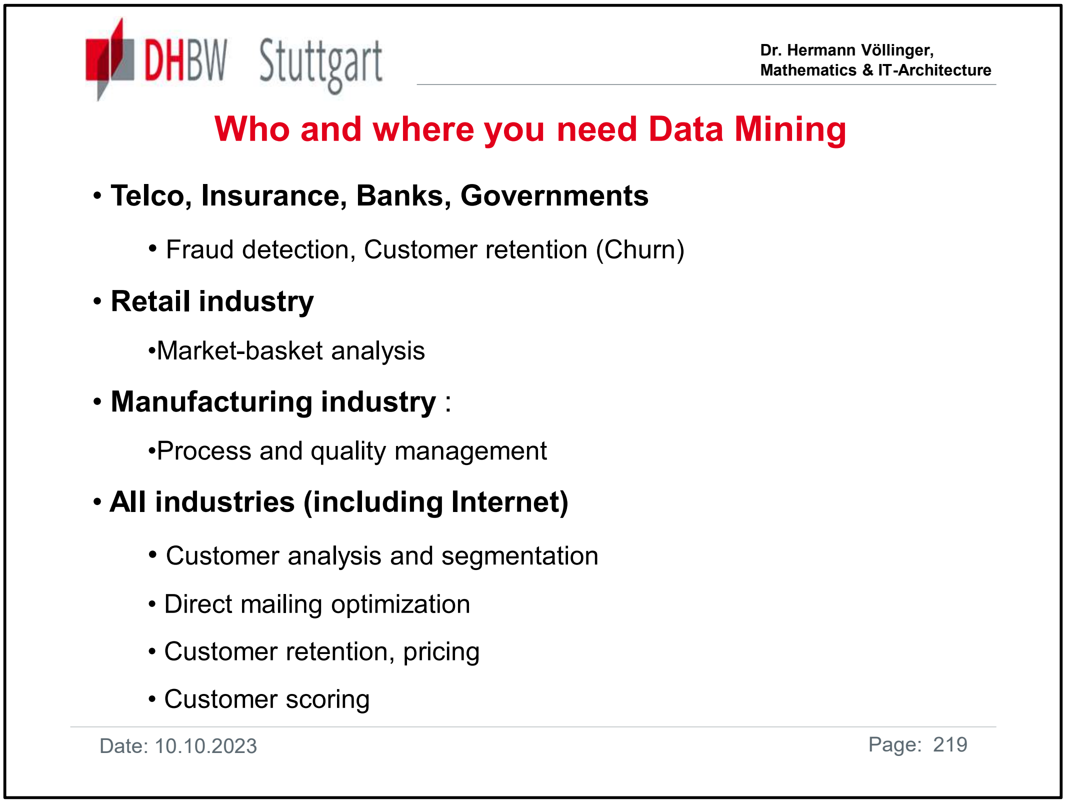


Definition of Data Mining

Data Mining is ...

The process of extracting *previously unknown*, *comprehensible*, and *actionable* information from large databases and using it to make crucial business decisions.

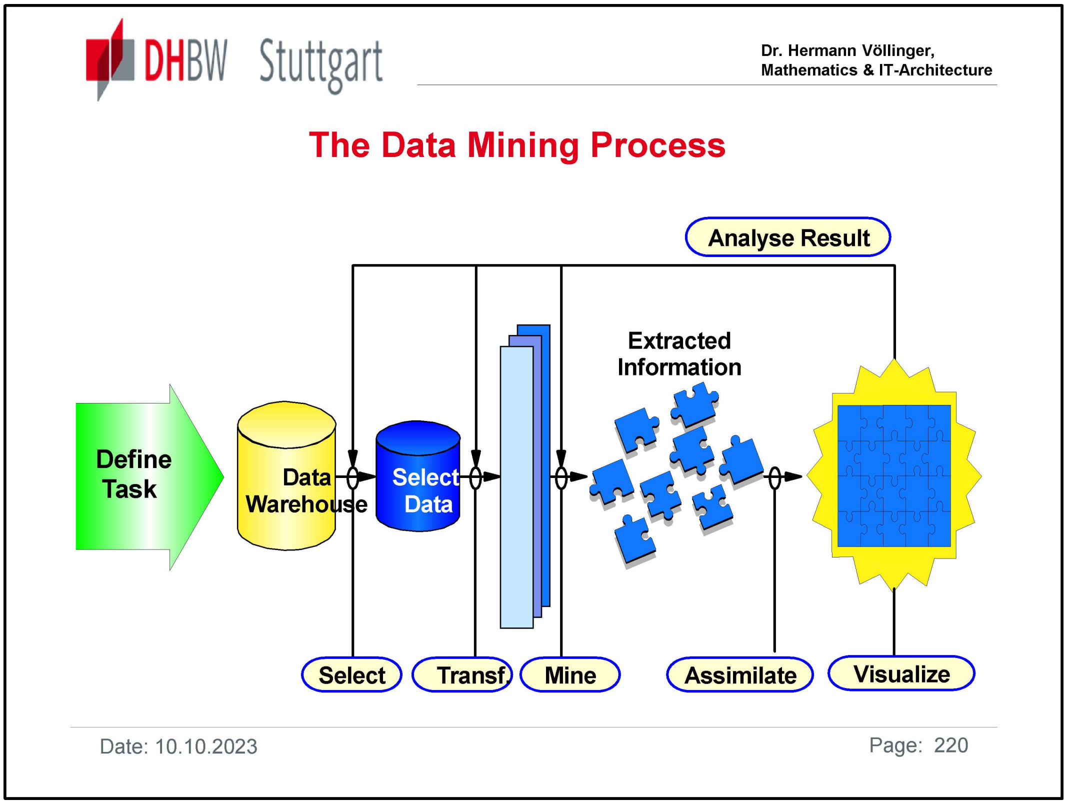
Information technology has developed rapidly over the last three decades. Many organizations store increasingly large volumes of data on their computer systems.
Useful information might be hidden in the data in the form of implicit patterns and connections that are not easy to discern using conventional data queries and statistical calculations.
Data mining is the process of discovering valid, previously unknown, and ultimately comprehensible information from large stores of data.
You can use extracted information to form a prediction or classification model, or to identify similarities between database records. The resulting information can help you make more informed decisions.



Who and where you need Data Mining

- **Telco, Insurance, Banks, Governments**
  - Fraud detection, Customer retention (Churn)
- **Retail industry**
  - •Market-basket analysis
- **Manufacturing industry** :
  - •Process and quality management
- **All industries (including Internet)**
  - Customer analysis and segmentation
  - Direct mailing optimization
  - Customer retention, pricing
  - Customer scoring

Data Mining can be used in all industries.
The slide shows the most used business applications of Data Mining in the different industries.
The list is not complete, but gives a good impression of the importance of Data Mining for the process of getting information (and out of this decisions) out of raw data.



The Data Mining Process

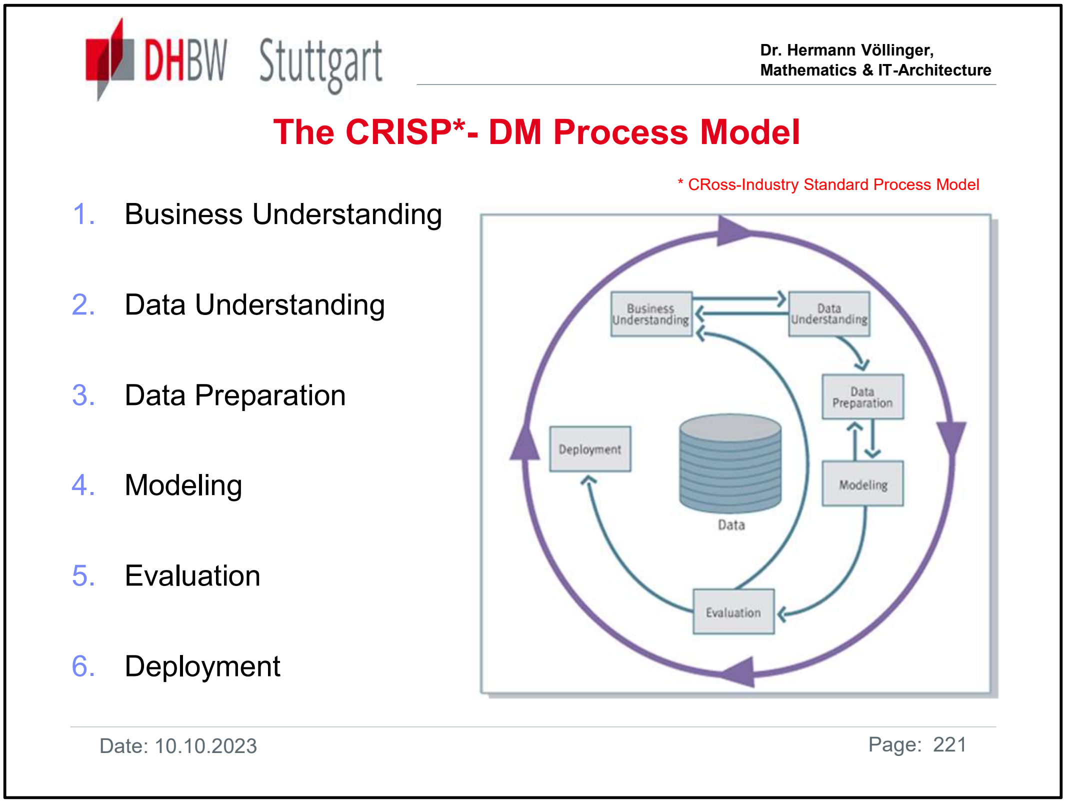
Data mining is an iterative process that typically involves:

- selecting input data,
- transforming it,
- running a mining function
- and interpreting the results.

The Data Mining tool assists you with all the steps in this process. You can apply the functions of the Data Mining tool independently, iteratively, or in combination.

Mining functions use elaborate mathematical techniques to discover hidden patterns in your data. After interpreting the results of your data-mining process, you can modify your selection of data, data processing and statistical functions, or mining parameters to improve and extend your results.

So the Data Mining process is a highly interactive and periodic process. This means in a concrete project you have to refine several times for mining models until you reach a satisfactory and valuable result.



The CRISP*- DM Process Model

* CRoss-Industry Standard Process Model

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Business Understanding: This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

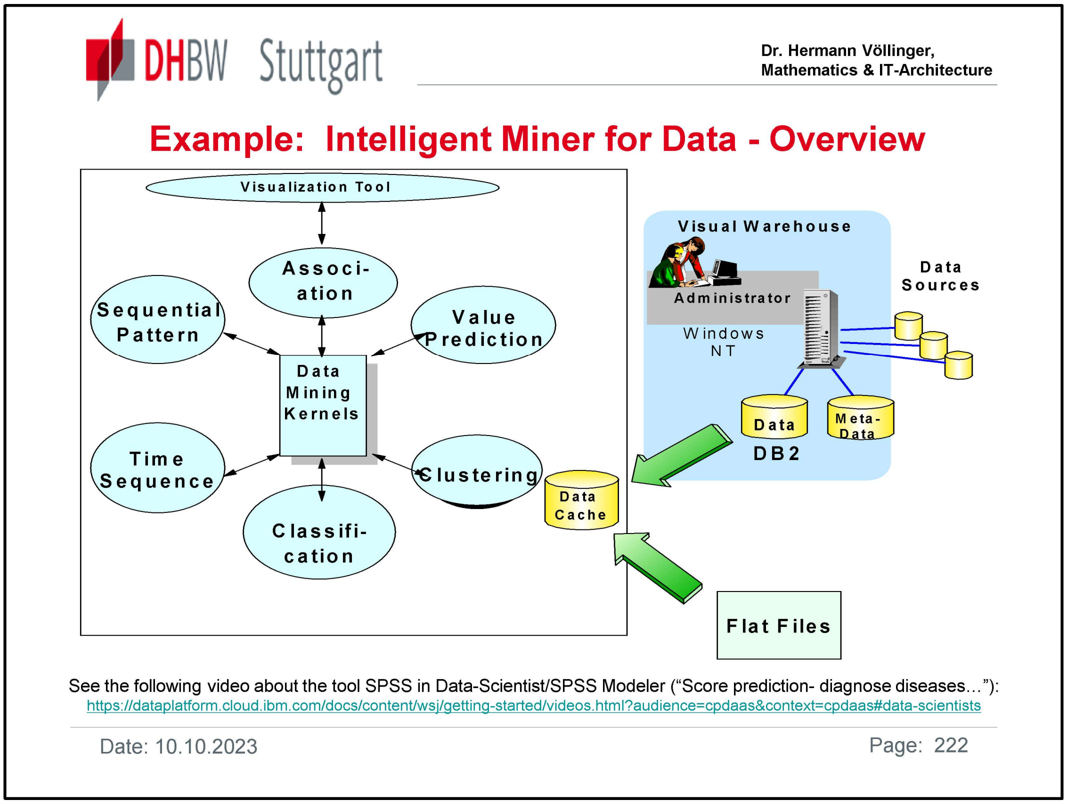
Data Understanding: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Data Preparation: The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

Modeling: In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

Evaluation: At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment: Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions will need to be carried out in order to actually make use of the created models.



Example: Intelligent Miner for Data - Overview

See the following video about the tool SPSS in Data-Scientist/SPSS Modeler ("Score prediction- diagnose diseases…"): https://dataplatform.cloud.ibm.com/docs/content/wsj/getting-started/videos.html?audience=cpdaas&context=cpdaas#data-scientists

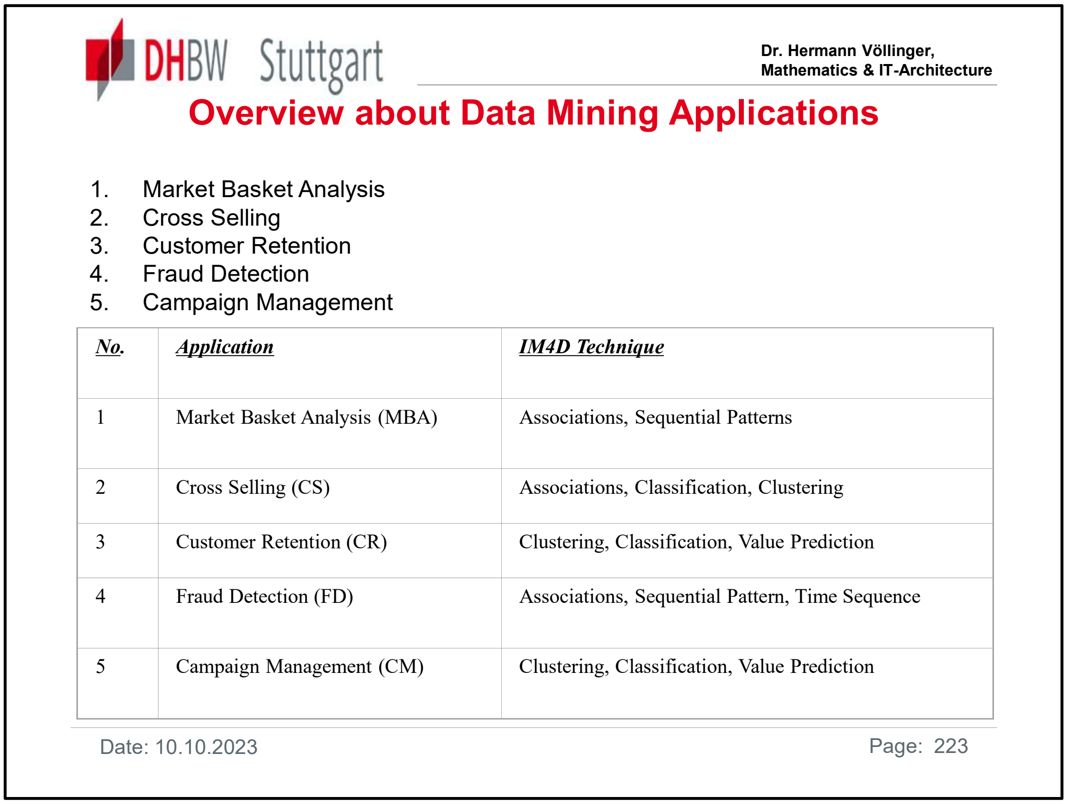
We look here as an example on the IBM solution, which is *Intelligent Miner for Data (IM4D).*

Intelligent Miner's mining functions use innovative techniques to efficiently discover previously unknown patterns in your business data.

The following mining functions are available:

- Associations mining function
- Clustering mining functions
- Sequential Patterns mining function
- Time Sequences mining function
- Classification mining functions
- Prediction mining functions

For clustering, classification, and prediction, different methods are available. Each of these methods is suited to a different class of business problems**.**



Overview about Data Mining Applications

1. Market Basket Analysis
2. Cross Selling
3. Customer Retention
4. Fraud Detection
5. Campaign Management

No.	***Application***	***IM4D Technique***
1	Market Basket Analysis (MBA)	Associations, Sequential Patterns
2	Cross Selling (CS)	Associations, Classification, Clustering
3	Customer Retention (CR)	Clustering, Classification, Value Prediction
4	Fraud Detection (FD)	Associations, Sequential Pattern, Time Sequence
5	Campaign Management (CM)	Clustering, Classification, Value Prediction

The **best-known applications**, which uses Data Mining techniques are:

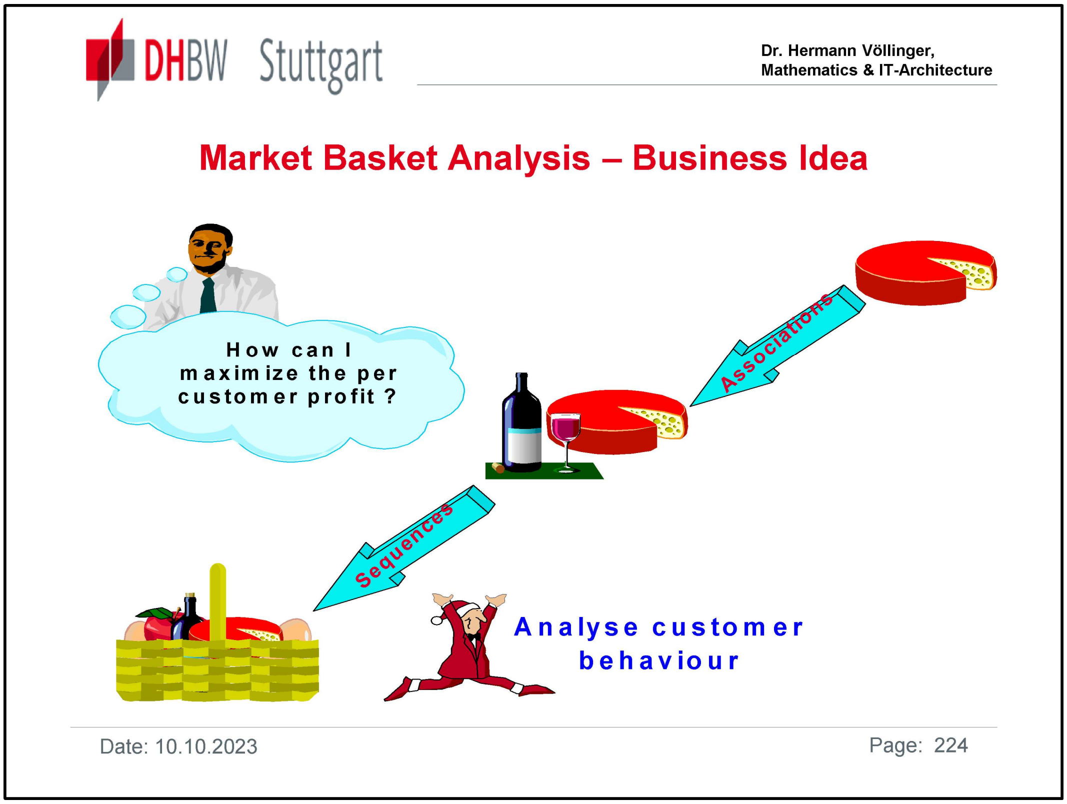
1. Market Basket Analysis (MBA)
2. Cross Selling (CS)
3. Customer Retention (CR)
4. Fraud Detection (FD)
5. Campaign Management (CM)

The **link between the DM applications and IM4D methods** is given with the following list:

1. Market Basket Analysis (MBA) ---→ Associations, Sequential Patterns
2. Cross Selling ---→ Associations, Classification, Clustering
3. Customer Retention →Clustering, Classification, Value Prediction
4. Fraud Detection ---→Associations, Sequential Pattern, Time Sequence
5. Campaign Management ---→ Value Prediction, Classification, Clustering

Examples where this DM applications are **used in real-life:**

1. Market Basket Analysis (MBA) -→ Retail: How products are placed in the shelf in the store. Goods which are often sold together are placed together in the shelf. Special prices for bundles of goods.
2. Cross Selling CR -→ Automobile companies also offers now Financial Services (Mercedes Benz Bank, VW Bank). Coffee shops also offer clothes other goods for the kitchen or household. Telephone shops offer other articles together with a handy, like a handy envelope (leather).
3. CR -→ 'Miles & More' at Lufthansa or 'Happy Digits' at D. Telekom. Gold Status or Premium Status for customer cards.
4. FD -→ Insider Trading (stock exchange). Fraud Detection at Automobile-Insurance. Fraud detection with EC- or Master-Card.
5. CM -→ Campaigns to prevent silent attrition of valuable customers, i.e. special campaigns for special customers which offer better prices or tariff models.



Market Basket Analysis – Business Idea

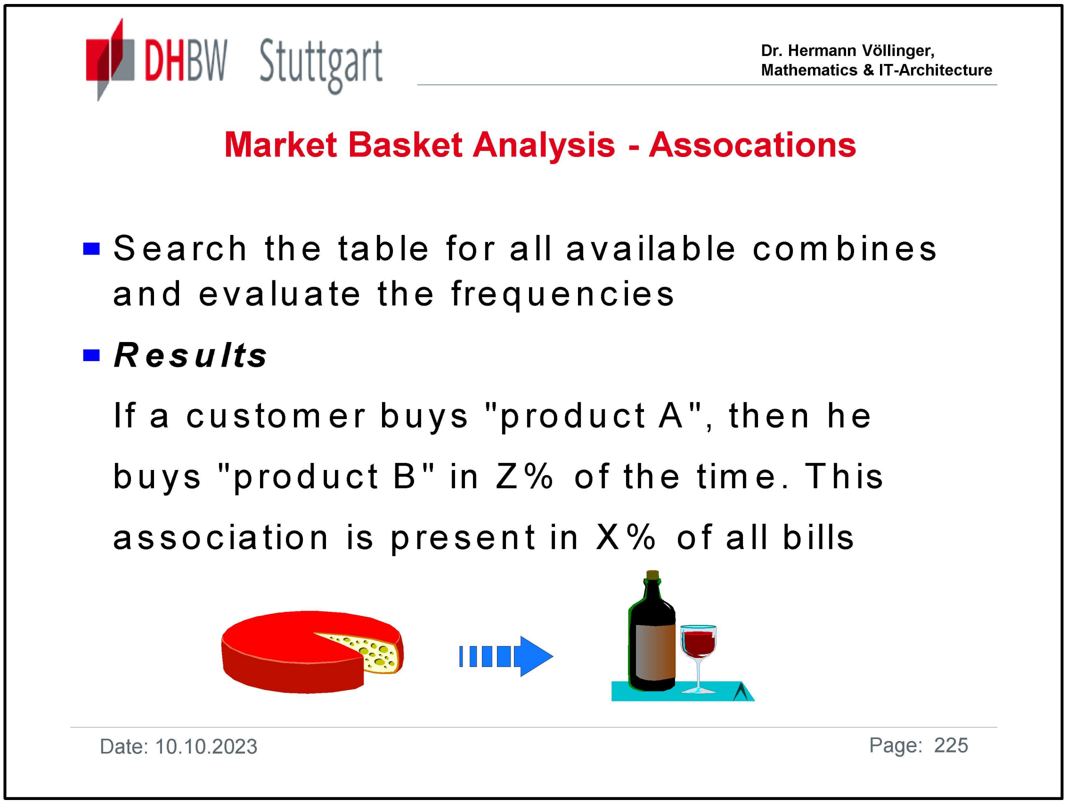
How can I maximize the per customer profit ?

Associations

Sequences

Analyse customer behaviour

The first method is **Market Basket Analysis (MBA).**
It shows the result for a retail scenario.
In the slide above one sees an example for the DM concept of Associations & Sequences.
Associations: This is possibly the best known mining technique, thanks to the well-known, but untrue, "diapers (nappies) and beer" anecdote.
As some sets of items will never appear together, and some only infrequently, statistics are used to calculate how often they occur together and decide how confident we can be that this is a real relationship.

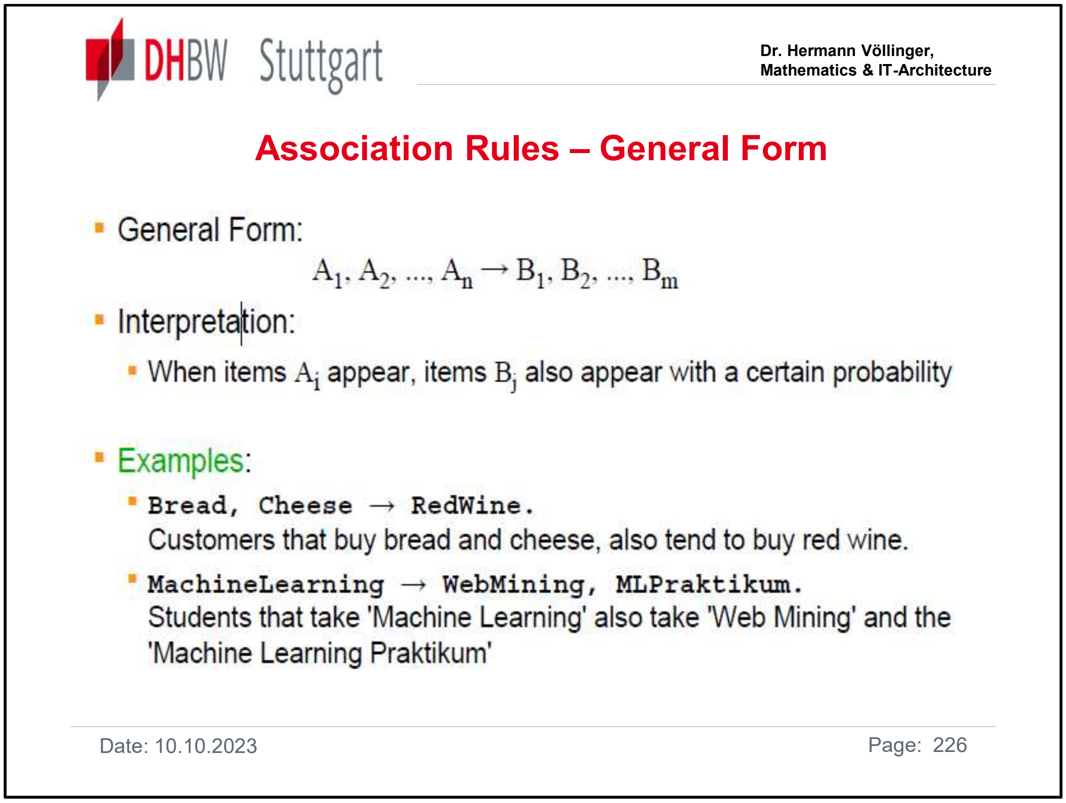


Market Basket Analysis - Assocations

- Search the table for all available combines and evaluate the frequencies
- ***Results***

  If a customer buys "product A", then he buys "product B" in Z% of the time. This association is present in X% of all bills

The slide shows how the **Associations** concept is working.
Associations: A relationship implies if one item is present, then the other will likely be present too. This is shown in the 4 concepts of Support, Confidence, Lift and Type reported by IM and the logic of this is fairly easy to understand.
1. Support represents the percentage of transactions in which items appear together. It shows how relevant this is compared to the total number of transactions.
2. Confidence gives the percentage of transactions that contain the first item and also the Second (the probability that the second item will be there). This defines the strength of the pattern, so if 90% of the time when a customer bought X they also bought Y, then X implies Y with 90% confidence).
3. Lift is the probability of both parts of the association occurring compared to how likely on of the parts is of occurring. (actual confidence factor divided by the expected confidence). A high lift means that the connection between the items is stronger. An item which is not expected to occur very often, which always occurs with another item has a strong lift, whereas an item which occurs nearly all the time probably does not have an association and will lead to a weak lift.
4. Type identifies where the lift is statistically significant, meaning that the appearance of one item does imply the appearance of another.
Hierarchies, called **Taxonomies,** can be set up to find associations at different levels. For example there may be no association with product A and a specific cola drink, but there may be one for all cola products or all soft drinks. The taxonomy effectively acts like an alias for the lower level items (children of the item in the hierarchy).



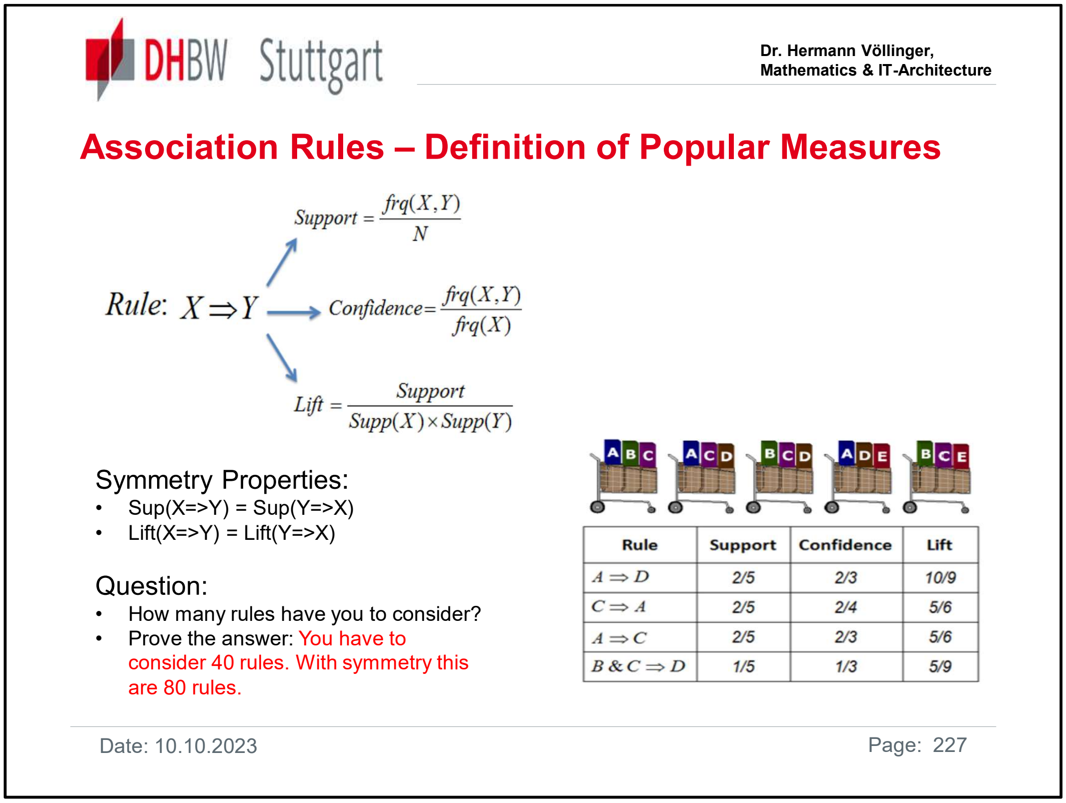
Association Rules – General Form

- General Form:

$$A_1, A_2, \dots, A_n \rightarrow B_1, B_2, \dots, B_m$$

- Interpretation:
  - When items A_i appear, items B_j also appear with a certain probability

- Examples:
  - **`Bread, Cheese → RedWine.`**
    Customers that buy bread and cheese, also tend to buy red wine.
  - **`MachineLearning → WebMining, MLPraktikum.`**
    Students that take 'Machine Learning' also take 'Web Mining' and the 'Machine Learning Praktikum'



Association Rules – Definition of Popular Measures

$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Symmetry Properties:

- Sup(X=>Y) = Sup(Y=>X)
- Lift(X=>Y) = Lift(Y=>X)

Question:

- How many rules have you to consider?
- Prove the answer: You have to consider 40 rules. With symmetry this are 80 rules.

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \,\&\, C \Rightarrow D$	1/5	1/3	5/9

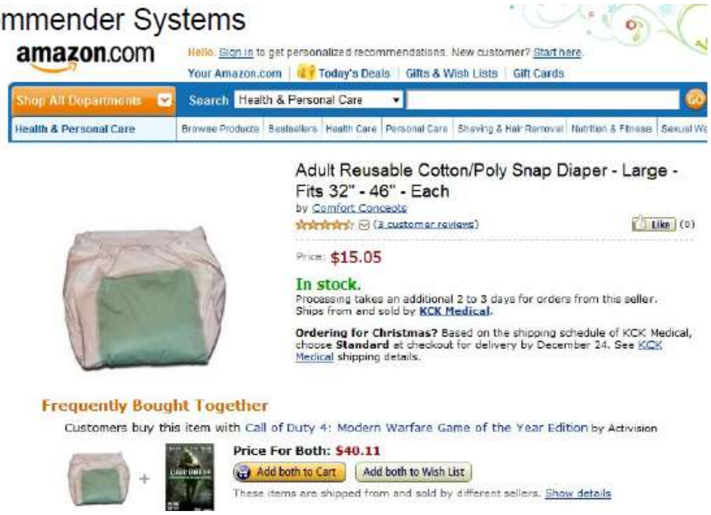
N=5
Support (A=>D):= frq(A,D)/5=2/5
Support (C=>A):= frq(C,D)/5=2/5
Support (A=>C):= frq(A,C)/5=2/5
Support (B&C=>D):= frq(B&C,D)/5=1/5

Confidence(A=>D):=frq(A,D)/frq(A)=(2/5)/(3/5)=2/3
Confidence(C=>A):=frq(C,A)/frq(C)=(2/5)/(4/5)=2/4=1/2
Confidence(A=>C):=frq(A,C)/frq(A)=(2/5)/(3/5)=2/3
Confidence(B&C=>D):=frq(B&C,D)/frq(B&C)=(1/5)/(3/5)=1/3

Lift(A=>D):=Sup(A=>D)/(Sup(A)*Sup(D))=(2/5)/(3/5*3/5)=(2/5)/(9/25)=2/(9/5)=10/9
Lift(C=>A):=Sup(C=>A)/(Sup(C)*Sup(A))=(2/5)/(4/5*3/5)=(2/5)/(12/25)=2/(12/5)=10/12=5/6
Lift(A=>C):=Sup(A=>C)/(Sup(A)*Sup(C))=(2/5)/(3/5*4/5)=(2/5)/(12/25)=2/(12/5)=10/12=5/6
Lift(B&C=>D):=Sup(B&C=>D)/(Sup(B&C)*Sup(D))=(1/5)/(3/5*3/5)=(1/5)/(9/25)=1/(9/5)=5/9

Association Rules – Example of predictive MBA

- Recommender Systems



Hello. Sign in to get personalized recommendations. New customer? Start here.

Adult Reusable Cotton/Poly Snap Diaper - Large - Fits 32" - 46" - Each

by Comfort Concepts

(3 customer reviews)

Price: $15.05

In stock.

Processing takes an additional 2 to 3 days for orders from this seller.
Ships from and sold by KCK Medical.

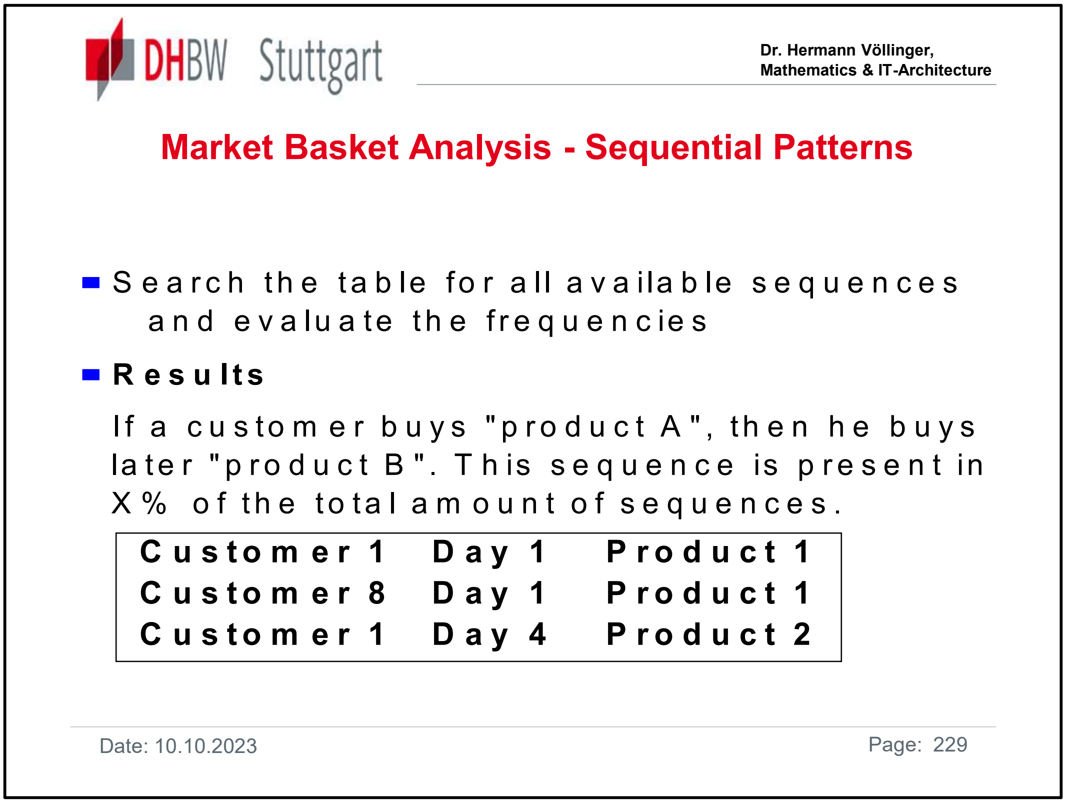
Ordering for Christmas? Based on the shipping schedule of KCK Medical, choose **Standard** at checkout for delivery by December 24. See KCK Medical shipping details.

Frequently Bought Together

Customers buy this item with Call of Duty 4: Modern Warfare Game of the Year Edition by Activision

Price For Both: $40.11

These items are shipped from and sold by different sellers. Show details



Market Basket Analysis - Sequential Patterns

- Search the table for all available sequences and evaluate the frequencies
- **Results**

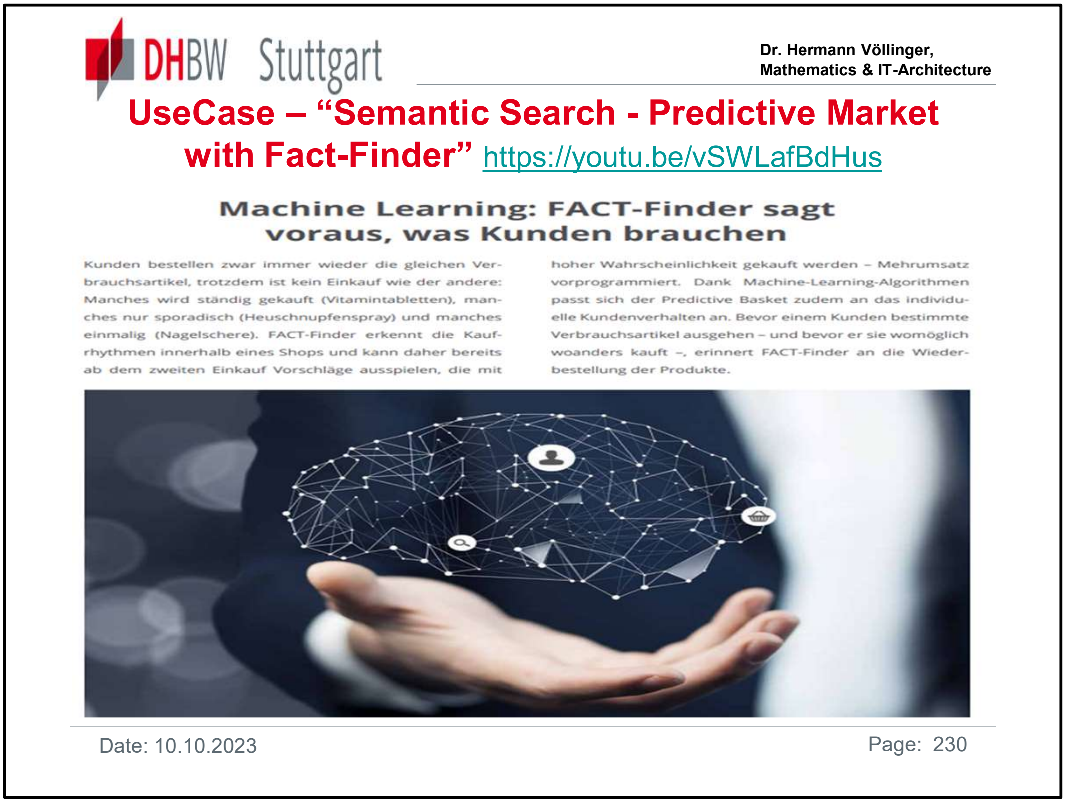
If a customer buys "product A", then he buys later "product B". This sequence is present in X % of the total amount of sequences.

Customer 1	**Day 1**	**Product 1**
Customer 8	**Day 1**	**Product 1**
Customer 1	**Day 4**	**Product 2**

The slide shows how the **Sequential Pattern** concept is working.
It shows the result for a retail scenario.

Sequential Patterns

Sequential Patterns are similar to associations, but are based on behavior over time, not a snapshot of activity at one point in time. Associations are centered on transactions (A+B happened together), Sequential Patterns on the sequence (A happened, then B, then C).



UseCase – “Semantic Search - Predictive Market with Fact-Finder” https://youtu.be/vSWLafBdHus

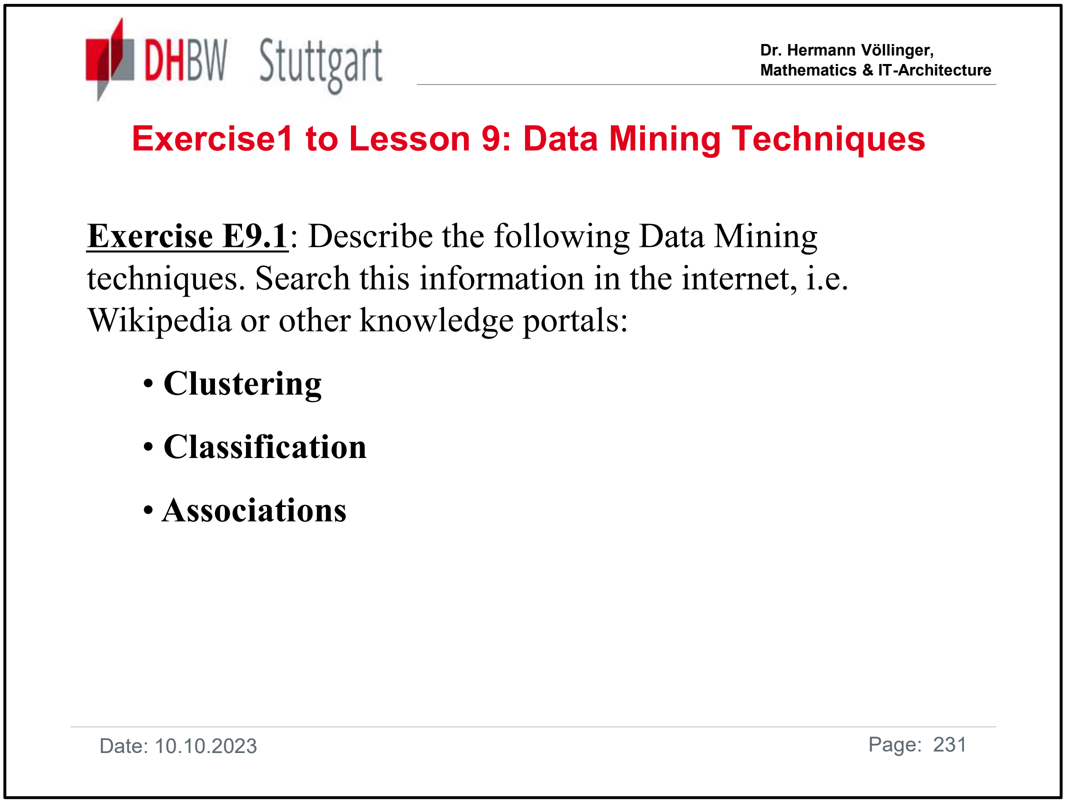
Machine Learning: FACT-Finder sagt voraus, was Kunden brauchen

Kunden bestellen zwar immer wieder die gleichen Verbrauchsartikel, trotzdem ist kein Einkauf wie der andere: Manches wird ständig gekauft (Vitamintabletten), manches nur sporadisch (Heuschnupfenspray) und manches einmalig (Nagelschere). FACT-Finder erkennt die Kaufrhythmen innerhalb eines Shops und kann daher bereits ab dem zweiten Einkauf Vorschläge ausspielen, die mit hoher Wahrscheinlichkeit gekauft werden – Mehrumsatz vorprogrammiert. Dank Machine-Learning-Algorithmen passt sich der Predictive Basket zudem an das individuelle Kundenverhalten an. Bevor einem Kunden bestimmte Verbrauchsartikel ausgehen – und bevor er sie womöglich woanders kauft –, erinnert FACT-Finder an die Wiederbestellung der Produkte.

Maximize re-orders. With the Predictive Basket. Online shopping without rummaging and thinking: The Predictive Basket by FACT-Finder shows yours customers are likely to buy those products that they are likely to buy in their current session become.

Whether click, search query or purchase - with every interaction in the shop your customers leave traces in the form of data. With our tracking interface, this data is captured by your FACT-Finder and used for shop optimization. But what exactly does FACT-Finder use the tracking data for? And what does shop tracking bring to your customers and you as a user? In this post you will find the answers. Your box office hits automatically move into the focus of the customers, based on the shop tracking data. FACT-Finder learns which products are most popular with your customers - that is, which ones they most click on, add to the shopping cart and buy. This knowledge can be incorporated into the sorting of your search results - by activating the Automatic Search Optimization. Your bestsellers will automatically move up to the top positions in the result ranking over time. This increases the purchase probability. Because the higher the relevant products appear, the faster they will catch your customers. See also Fact-Finder Webinar: Dr. Holger Schmidt **– „Wie Plattformen den eCommerce disrupten“:**

https://www.youtube.com/watch?v=9T9sOxRB9qg&feature=youtu.be



Exercise1 to Lesson 9: Data Mining Techniques

Exercise E9.1: Describe the following Data Mining techniques. Search this information in the internet, i.e. Wikipedia or other knowledge portals:

- **Clustering**
- **Classification**
- **Associations**

Clustering: Clustering is used to segment a database into subsets, the clusters, with the members of each cluster having similar properties. IM for Data can perform clustering by using either a statistical clustering algorithm (Demographic Clustering) or a neural network algorithm (Kohonen Clustering), depending on the type of the input data set. The neural clustering algorithm requires the user to specify the number of clusters required; the statistical clustering algorithm automatically determines the "natural" number of clusters.

When clustering is performed there are no preconceived notions of what patterns exist within the data; it is a discovery process. The results of the clustering process can be visualized to determine the composition of each cluster. Visualization graphically presents the statistical distributions of the characteristics of those records that compose the cluster in comparison with the data set as a whole. Tabular output is also provided to enable further analysis.

In addition to producing graphical and tabular output, a "cluster model" is also generated (Training Mode). It is also possible to generate a user-defined table, which can include selected information from the input records, together with the cluster number of the segment to which the record has been assigned. The output table can also include details on the next nearest cluster and a measure of the confidence in the degree of matching to the nearest and next nearest clusters for each record (Test Mode). An Application Mode is also provided, in which new data records are assigned to clusters and an output table generated. In the commercial environment clustering is used in the areas of cross-marketing, cross-selling, customizing marketing plans for different customer types, deciding on media approach, understanding shopping goals, etc.

Classification: Classification is the process of automatically creating a model of classes from a set of records that contain class labels. The induced model consists of patterns, essentially generalizations over the records that are useful for distinguishing the classes. Once a model is induced, it can be used to automatically predict the class of other unclassified records. IM for Data has two classification algorithms, a tree induction algorithm (modified CART regression tree) and a neural network algorithm (back propagation), to compute the classes.

The tree and neural network algorithms develop arbitrary accuracy. While neural networks often produce the most accurate classifications, trees are easy to understand and modify and the model developed can be expressed as a set of decision rules. Commercial applications of classification include credit card scoring, ranking of customers for directed mailing, and attrition prediction. One of the main uses of the tree algorithm is to determine the rules that describe the differences between the clusters generated by the clustering algorithm. This is achieved by taking the output table from the clustering algorithm and constructing the decision tree using the cluster label as the class.

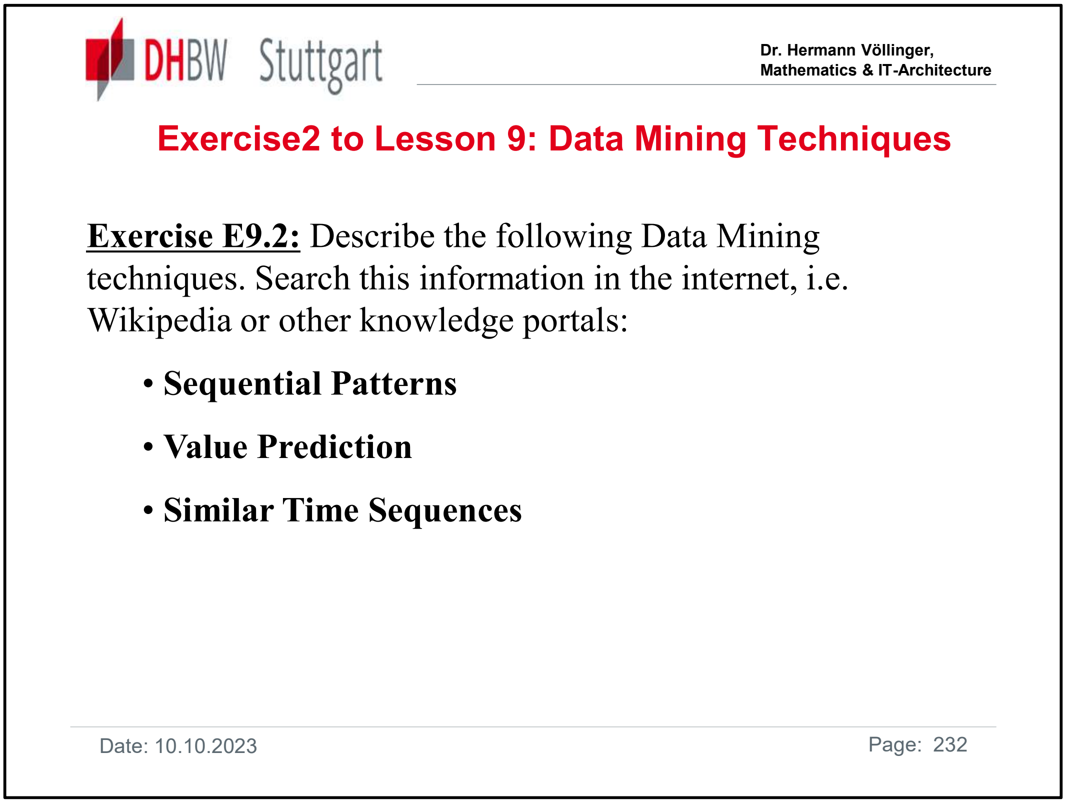
Associations: The association algorithm, developed at the IBM Almaden Research Center in San Jose, California, compares lists of records to determine if common patterns occur across the different lists. In a typical commercial application the algorithm looks for patterns such as whether, when a customer buys paint, they also buy paintbrushes. More specifically, it assigns probabilities; for example, if a customer buys paint, there is a 20% chance that they will buy a paintbrush. The advantage of this approach is that it compares all possible associations. It also finds multiple associations, for example, if a customer buys paint and paint brushes, there is a 40% chance they will also buy paint thinner.

When the algorithm runs, it potentially creates hundreds or thousands of such rules. The user can however select a subset of rules that have either higher confidence levels (a high likelihood of B given A) or support levels (the percent of transactions in the database that follow the rule) or high lift (the ratio of measured to expected confidence for a rule). It is up to the user to read the rules and decide if the rules are:

- Chance correlations (for example, paint and hair rollers were on sale the same day and therefore were correlated by chance).
- Known correlations (for example, the paint and paint brush correlation is something that would have been known).
- Unknown but trivial correlations (for example, red gloss paint and red non gloss paint correlation may be something unknown and is unimportant to know).
- Unknown and important correlations (for example, paint and basketballs, which may be something previously unknown and very useful in both organization of advertising and product placement within the store).

Association discovery is used in market basket analysis, item placement planning, promotional sales planning, and so forth.

The association algorithm also includes the capability to include a taxonomy for the items in the lists (for example, paint and a paintbrush are hardware) and the algorithm will discover associations across the taxonomy (for example, there is a 50% confidence that customers who buy hardware also buy soft furnishing).



Exercise2 to Lesson 9: Data Mining Techniques

Exercise E9.2: Describe the following Data Mining techniques. Search this information in the internet, i.e. Wikipedia or other knowledge portals:

- **Sequential Patterns**
- **Value Prediction**
- **Similar Time Sequences**

Sequential Patterns

The purpose of discovering sequential patterns is to find predictable patterns of behavior over a period of time. This means that a certain behavior at a given time is likely to produce another behavior or a sequence of behaviors within a certain time frame. The rule generation method is a variation of the association technique. It analyzes the shopping behavior of customers, for example, over time. Instead of looking at 10,000 purchases, the algorithm looks at 10,000 sets of purchases.

These sets are, for example, lists of purchases from a sequence of shopping trips by a single customer. As a typical commercial example, one set of lists may be the purchases of computer:

- Computer in December
- Computer games and joy-stick in January
- Additional computer memory and larger hard drive in March

If this sequence, possibly with different time scales but the same order, were repeated across a number of customers, then the sequential association algorithm would typically return a rule, such as:

If following the purchase of a computer, the customer purchases computer games, then there is a 30% chance that extra computer memory will be purchased in a subsequent visit to the store.

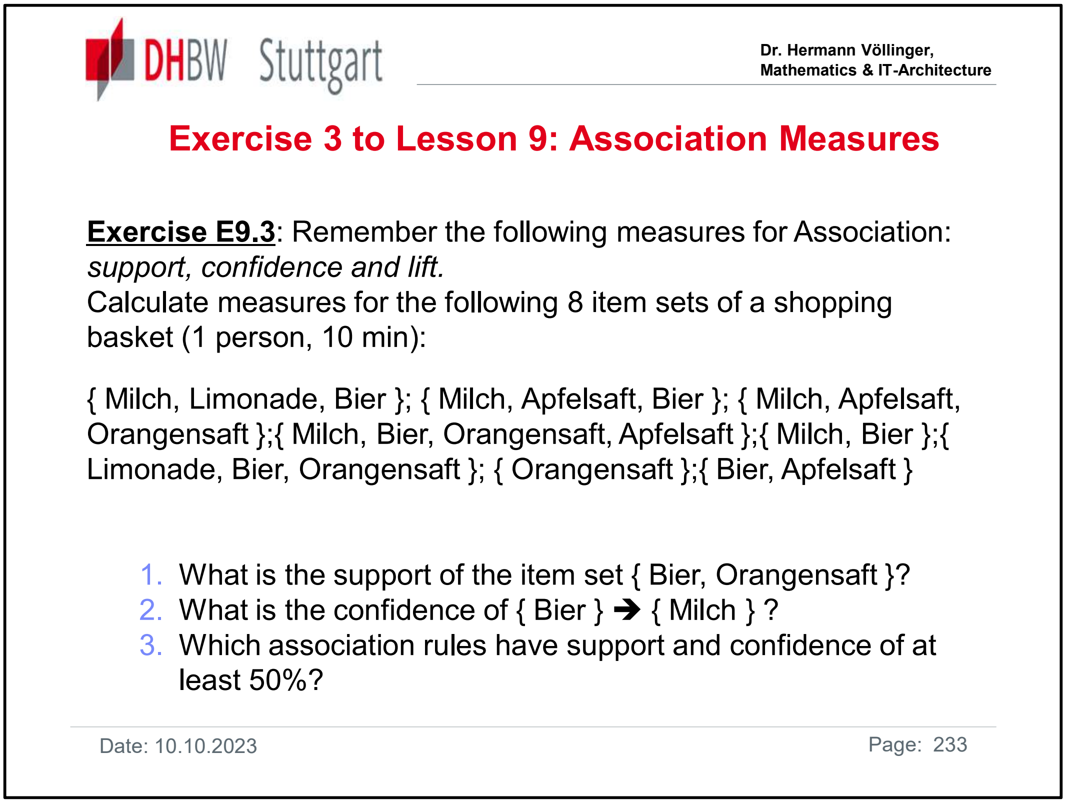
The algorithm also includes the capability to define minimum and maximum time periods between the items in the lists. This would, for example, enable the above rule to include the statement that computer memory will be purchased no earlier than one month and within three months of the purchase of the computer games. Sequential pattern detection can therefore be used to discover associations over time. This is especially useful in commercial applications, such as direct marketing, or the design special advertising supplements, and so on.

Value prediction

Value prediction is similar to classification; the goal is to build a data model as a generalization of the records. However, the difference is that the target is not a class membership but a continuous value or ranking. IM for Data has two prediction algorithms: a neural network algorithm and a Radial Basis Functions (RBF) algorithm. The radial basis function is particularly efficient and is appropriate for value prediction with very large data sets.

Similar time sequences

The purpose of this process is to discover all occurrences of similar subsequences in a database of time sequences. Given a database of time sequences, the goal is to find sequences similar to a given one or find all occurrences of similar sequences. The powerful alternatives afforded by multiple methods are enhanced by the fact that several of the methods are supported by more than one mining technique. Multiple techniques are often used in combination to address a specific business problem.



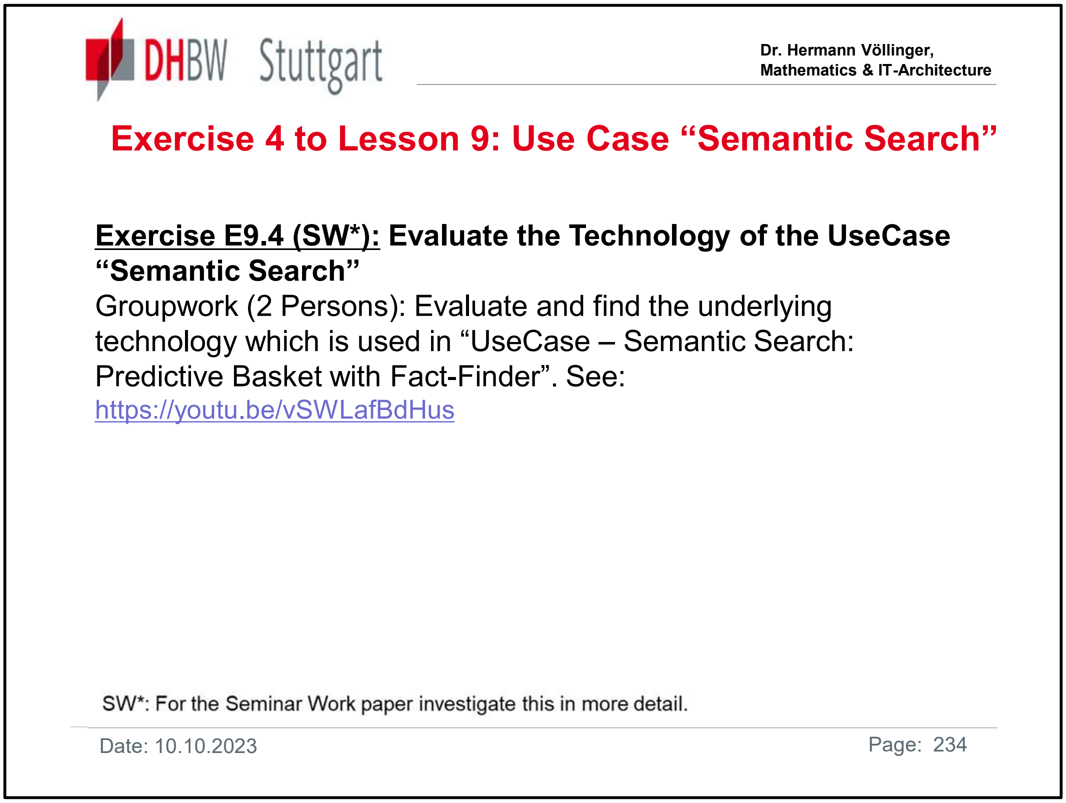
Exercise 3 to Lesson 9: Association Measures

Exercise E9.3: Remember the following measures for Association: *support, confidence and lift.*
Calculate measures for the following 8 item sets of a shopping basket (1 person, 10 min):

{ Milch, Limonade, Bier }; { Milch, Apfelsaft, Bier }; { Milch, Apfelsaft, Orangensaft };{ Milch, Bier, Orangensaft, Apfelsaft };{ Milch, Bier };{ Limonade, Bier, Orangensaft }; { Orangensaft };{ Bier, Apfelsaft }

1. What is the support of the item set { Bier, Orangensaft }?
2. What is the confidence of { Bier } ➔ { Milch } ?
3. Which association rules have support and confidence of at least 50%?

See the solution in „Exercises&Solutions-Intro2DWH.pdf“



Exercise 4 to Lesson 9: Use Case “Semantic Search”

Exercise E9.4 (SW*): Evaluate the Technology of the UseCase “Semantic Search”

Groupwork (2 Persons): Evaluate and find the underlying technology which is used in “UseCase – Semantic Search: Predictive Basket with Fact-Finder”. See: https://youtu.be/vSWLafBdHus

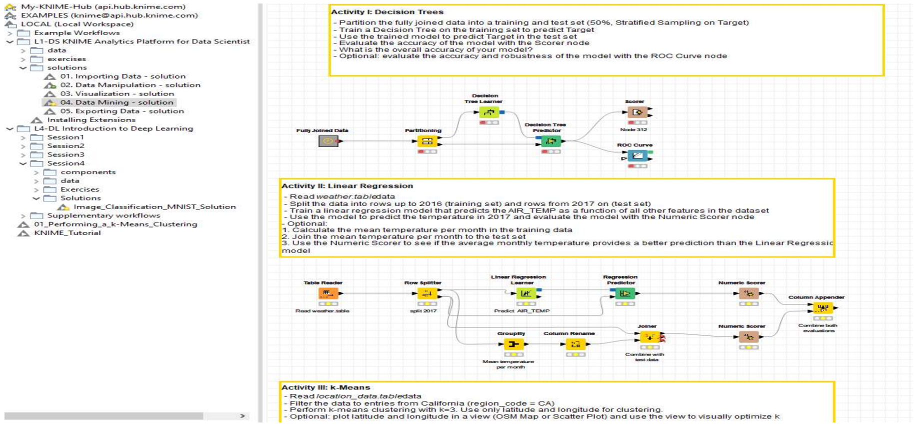
SW*: For the Seminar Work paper investigate this in more detail.

See the solution in „Exercises&Solutions-Intro2DWH.pdf“

Exercise 5 to Lesson 9: Performing KNIME DM-Basics

<u>Exercise E9.5 (SW*):</u> Run a KNIME-Basics Data Mining solution

Homework for 2 Persons: KNIME-Basics Workflow (use given solution) for one of the 3 KNIME solutions and give a technical explanation to the solution steps (see image below).



Activity I: Decision Trees

- Partition the fully joined data into a training and test set (50%, Stratified Sampling on Target)
- Train a Decision Tree on the training set to predict Target
- Use the trained model to predict Target in the test set
- Evaluate the accuracy of the model with the Scorer node
- What is the overall accuracy of your model?
- Optional: evaluate the accuracy and robustness of the model with the ROC Curve node

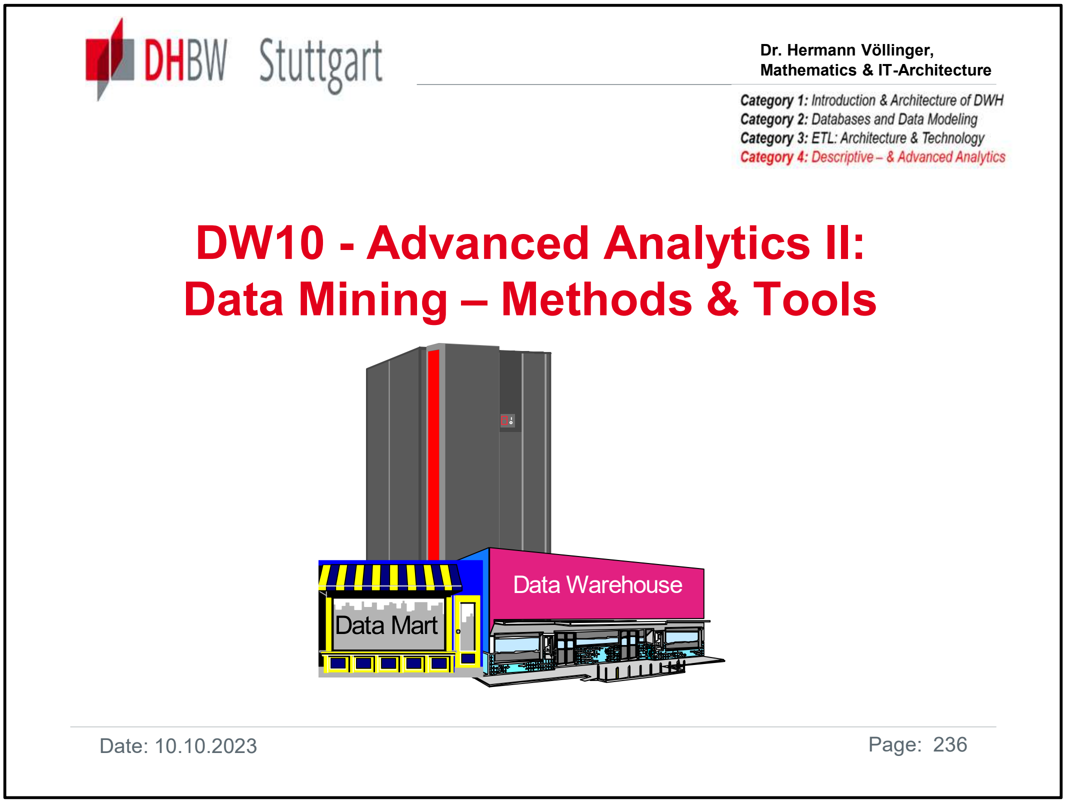
Activity II: Linear Regression

- Read *weather.table*data
- Split the data into rows up to 2016 (training set) and rows from 2017 on (test set)
- Train a linear regression model that predicts the AIR_TEMP as a function of all other features in the dataset
- Use the model to predict the temperature in 2017 and evaluate the model with the Numeric Scorer node
- Optional:
1. Calculate the mean temperature per month in the training data
2. Join the mean temperature per month to the test set
3. Use the Numeric Scorer to see if the average monthly temperature provides a better prediction than the Linear Regressic model

Activity III: k-Means

- Read *location_data.table*data
- Filter the data to entries from California (region_code = CA)
- Perform k-means clustering with k=3. Use only latitude and longitude for clustering.
- Optional: plot latitude and longitude in a view (OSM Map or Scatter Plot) and use the view to visually optimize k

SW*: In the Seminar Work paper investigate this in more detail.



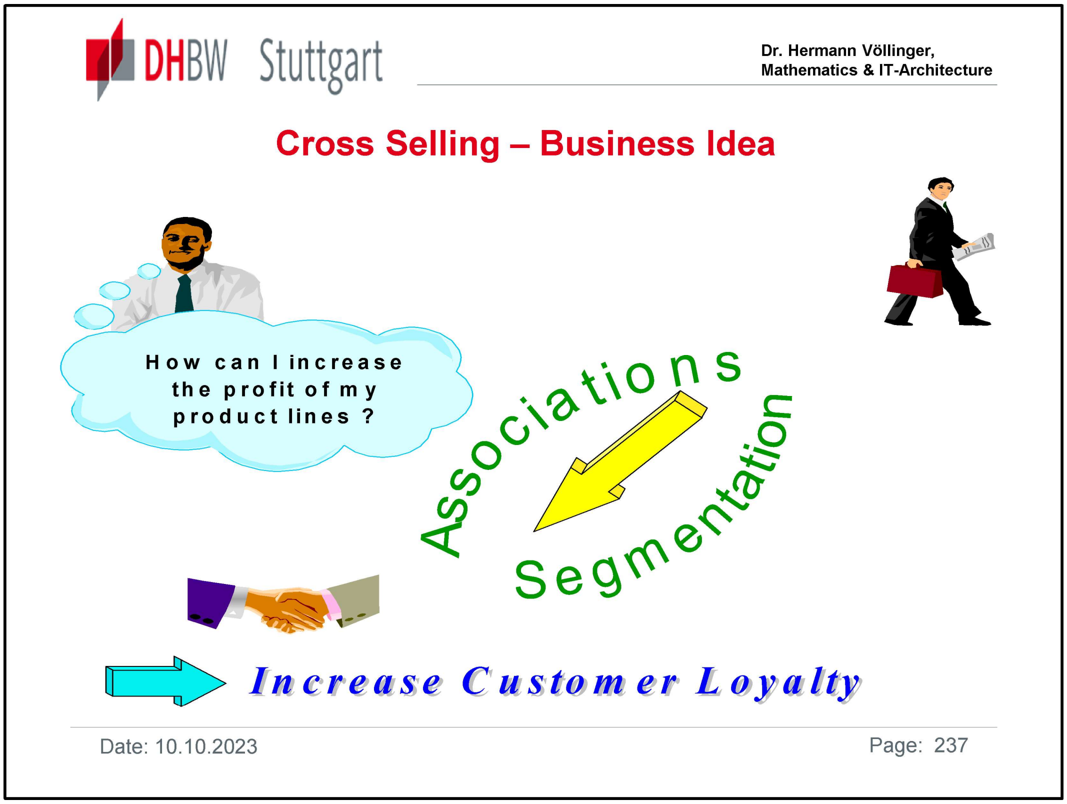
DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Category 1: *Introduction & Architecture of DWH*
Category 2: *Databases and Data Modeling*
Category 3: *ETL: Architecture & Technology*
Category 4: *Descriptive – & Advanced Analytics*

DW10 - Advanced Analytics II: Data Mining – Methods & Tools

Date: 10.10.2023 Page: 236



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Cross Selling – Business Idea

How can I increase the profit of my product lines ?

Associations

Segmentation

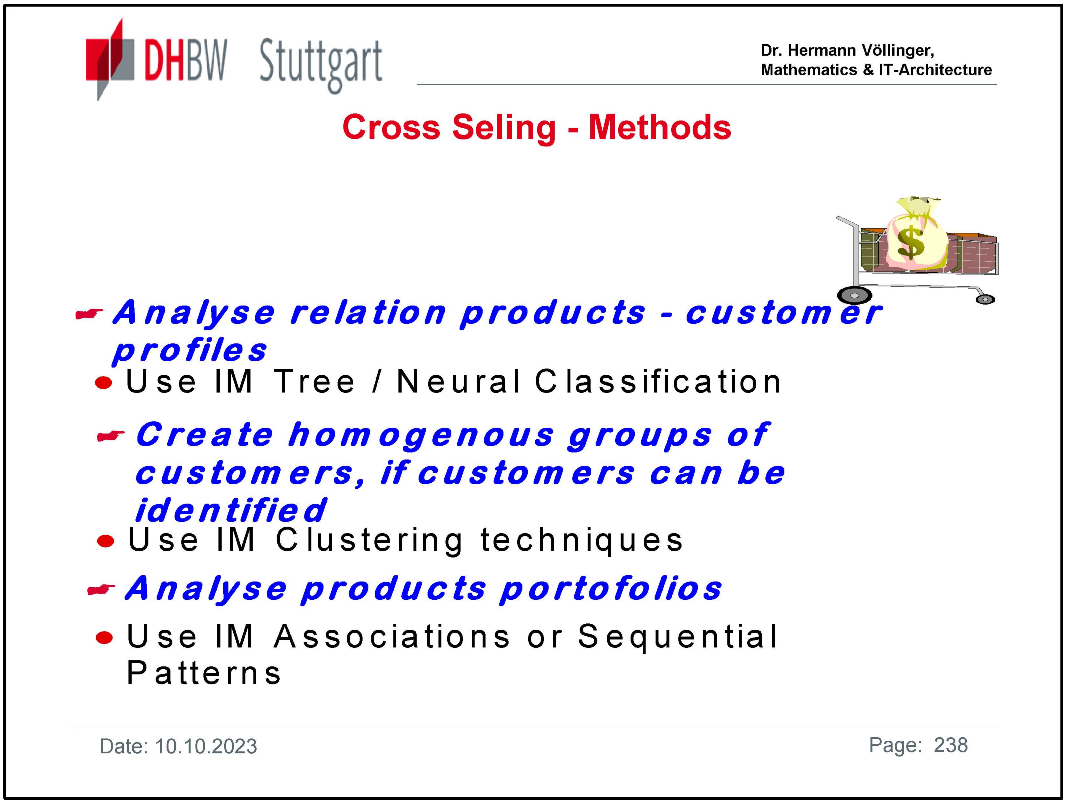
Increase Customer Loyalty

Date: 10.10.2023 Page: 237

Associations : The idea is to determine which things go together (e.g., retail chains). But it can also be used to identify cross-selling opportunities and to design an attractive package or grouping of products and services.

Example:

- Policyholders who hold auto insurance also hold term life insurance product.



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Cross Seling - Methods

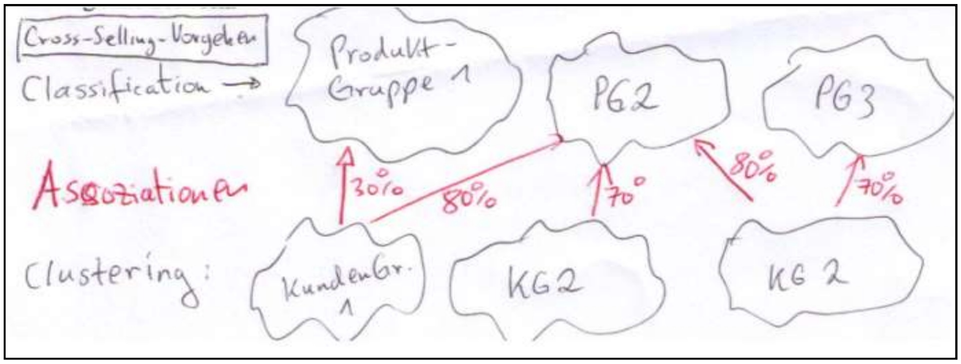
- ***Analyse relation products - customer profiles***
  - Use IM Tree / Neural Classification
- ***Create homogenous groups of customers, if customers can be identified***
  - Use IM Clustering techniques
- ***Analyse products portofolios***
  - Use IM Associations or Sequential Patterns

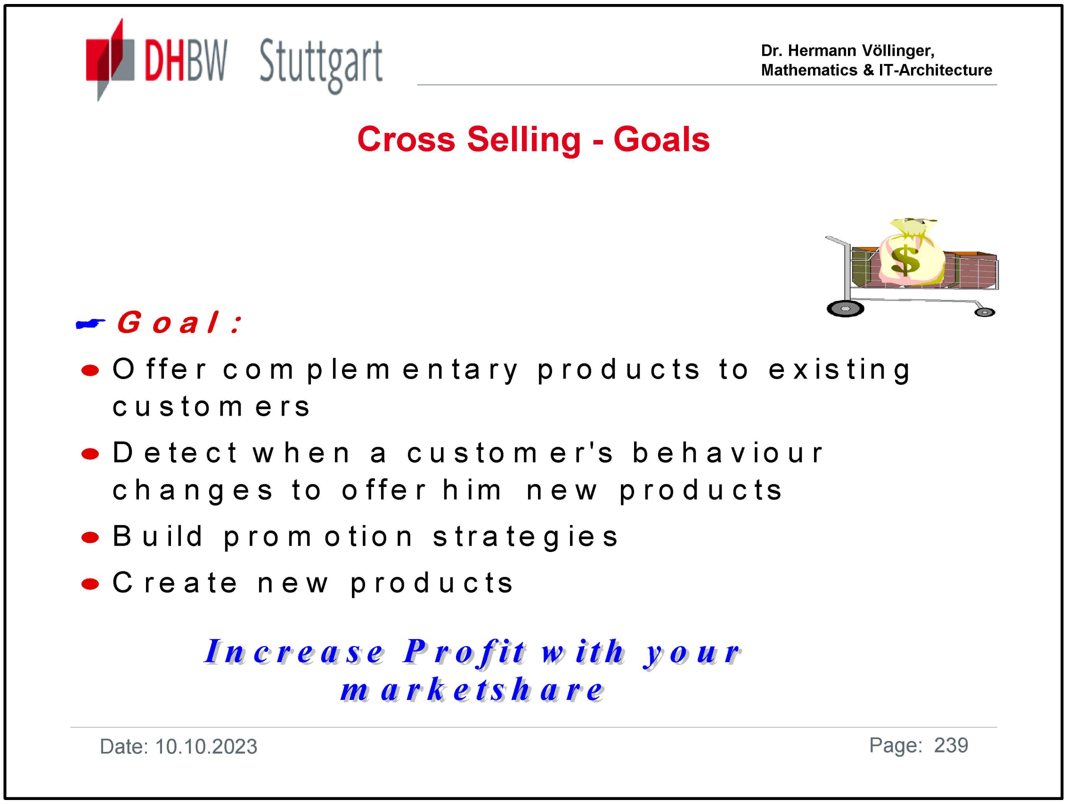
Date: 10.10.2023 Page: 238

IM Tree: **Tree Classification** is similar to clustering, in that it assigns records to groups, but different in that it assigns them to predefined groups, *Customer will stay/Customer will go* for example. It uses known data with a known outcome and decides what determined which group the record was in, i.e. the characteristics of each group. This can be used on unknown data to predict which group a new record is likely to belong to. There are various methods tree classification, three well-known names are CHisquare-Automatic-Interaction-Detection(CHAID), Classification And Regression Trees (CART) and Quick, Unbiased, Efficient Statistical Tree (QUEST). CHAID will split into two or more child nodes, CART and QUEST only into two child nodes.

Neural Classification: This, as you would expect, does the same job as Tree classification, but in a different way. The model is not a tree, but a network of 'neurons', the connection between each neuron has a strength or 'weight'. Each variable will be connected to some neurons and will have a weighting effect on whether the connection ends up as one outcome or another. The different values of the connection weights will give a different likely outcome depending on the path through the network.

Motivation for “Cross Selling”:





Cross Selling - Goals

Goal:

- Offer complementary products to existing customers
- Detect when a customer's behaviour changes to offer him new products
- Build promotion strategies
- Create new products

Increase Profit with your marketshare

IM Clustering: **Demographic Clustering & Neural Clustering**

Clustering, takes records and fits them into groups (clusters), which should be as similar as possible within the group and as dissimilar to other groups as possible.

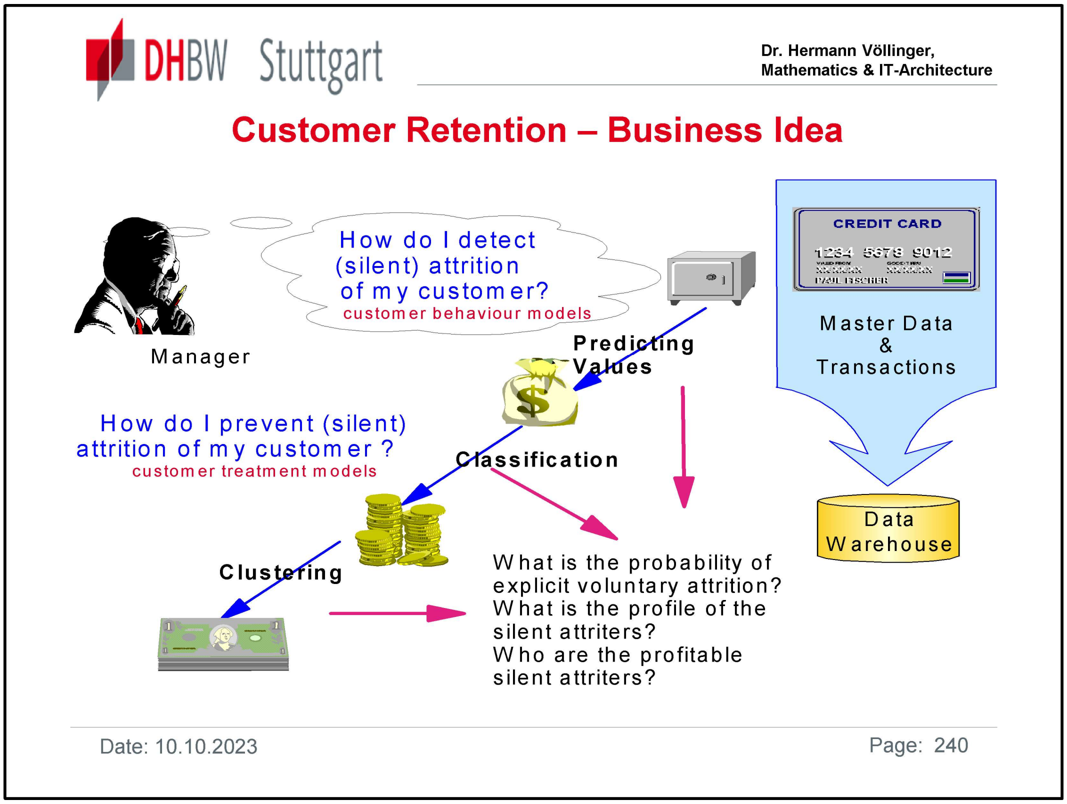
Demographic clustering uses something called the **Condorcet Criterion**, which it tries to maximize, to determine the cluster to which a record fits the best. This is easy for one variable, say sex, each record is either a boy or a girl, but when you have many variables, it is trickier to say how different a record is, but is similar enough to be in an existing group, or if it should be the first member of a new group.

For each variable, IM will class categorical variables as similar if they are the same, so sex=M is only similar to another variable where sex is also M. For numerical variables IM has to decide how far apart they can be, and still be classed as similar.

Neural clustering does the same job as Demographic clustering, but in a different way. IM uses a technique called Kohonen Feature Maps which use a process called self-organization to group similar input records together.

This is trickier to understand and more compute intensive than Demographic clustering.

Kohonen Feature Maps rely on a network of clusters (fixed number) and IM has to decide which cluster a record belongs to. This is done by some fairly serious statistics.



Customer Retention – Business Idea

Manager

How do I detect (silent) attrition of my customer?
customer behaviour models

Predicting Values

Classification

Clustering

How do I prevent (silent) attrition of my customer ?
customer treatment models

CREDIT CARD
1234 5678 9012
PAUL FISCHER

Master Data & Transactions

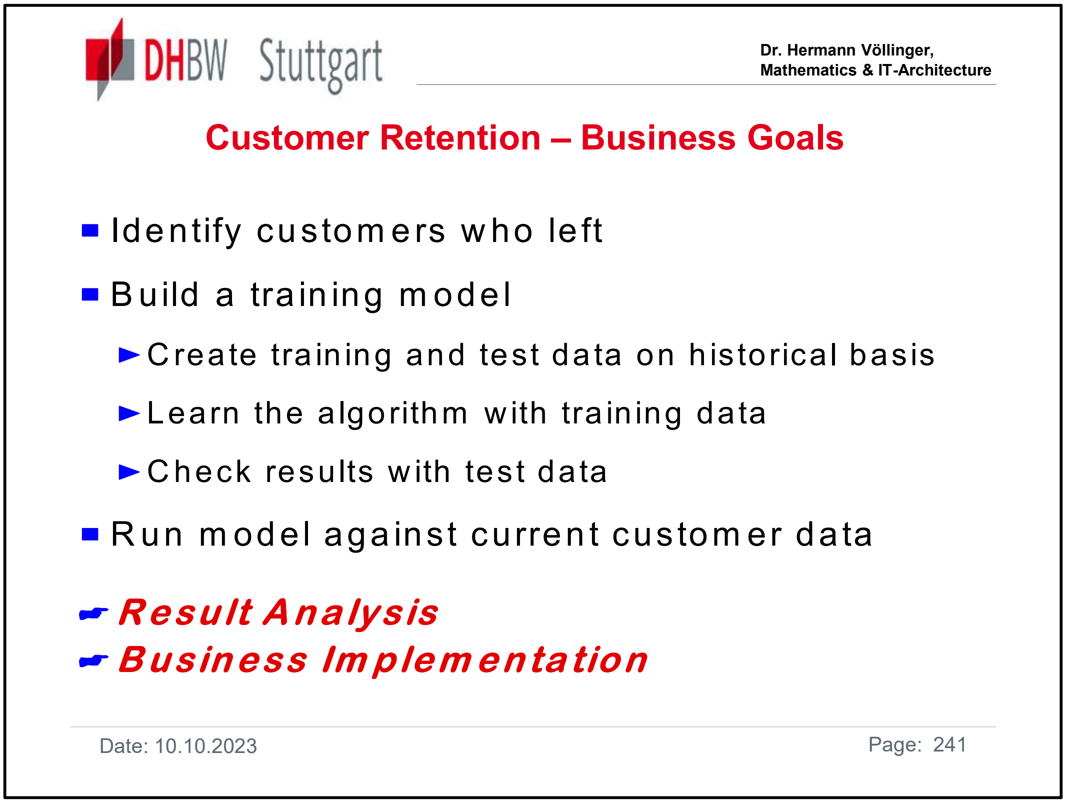
Data Warehouse

What is the probability of explicit voluntary attrition?
What is the profile of the silent attriters?
Who are the profitable silent attriters?

Prediction is the same as classification or estimation except that the records are classified according to some predicted future behavior or estimated future value. In a prediction task, the only way to check the accuracy of the classification is to wait & see.
Historical data is used to build a model that explains the current observed behavior.
Example: Predicting which customers will leave within the next 6 months

Classification consists of examining the features of a newly presented object (e.g., record in a database) and assigning it to one of a predefined set of classes on the basis of a model developed through training on pre-classified examples. In a case of records in a database, classification consists of updating each record by filling in a field with a class code. It deals with discrete values: Yes/No, Accept/Reject, ...

Examples of classification tasks include:
- classifying credit applicants as low, medium and high risk;
- spotting fraudulent insurance claims.

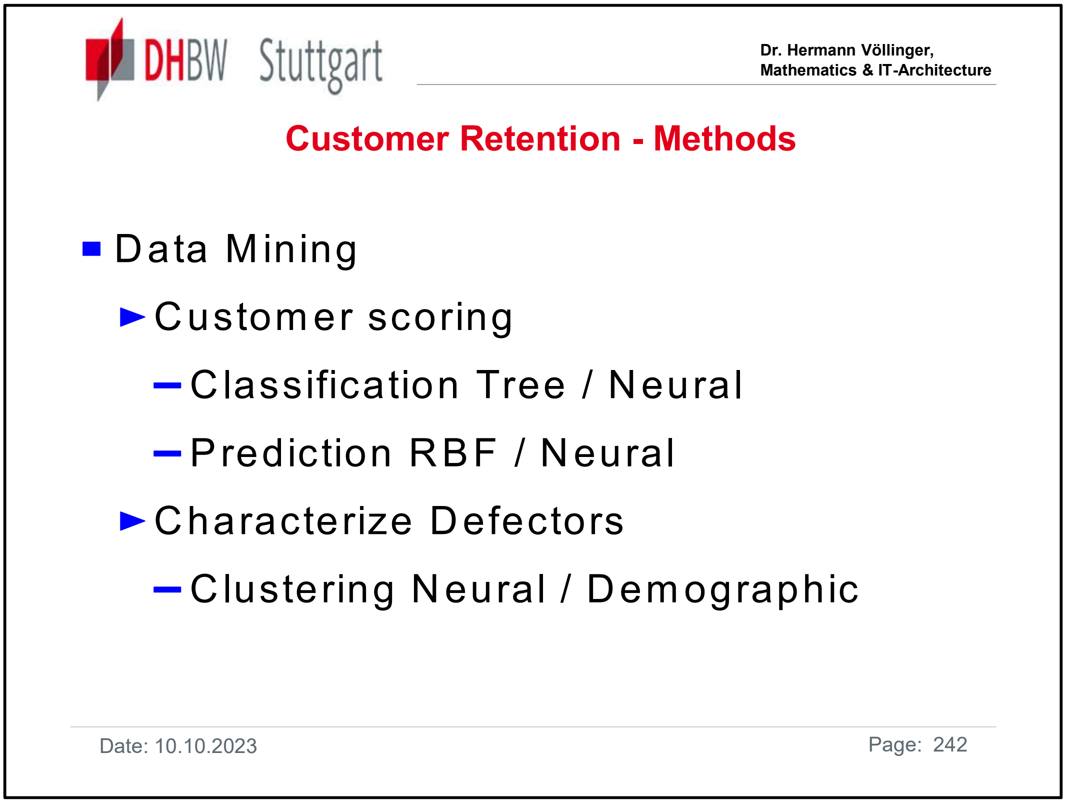


Customer Retention – Business Goals

- Identify customers who left
- Build a training model
  - ► Create training and test data on historical basis
  - ► Learn the algorithm with training data
  - ► Check results with test data
- Run model against current customer data

☛ ***Result Analysis***

☛ ***Business Implementation***



Customer Retention - Methods

- Data Mining
  - Customer scoring
    - Classification Tree / Neural
    - Prediction RBF / Neural
  - Characterize Defectors
    - Clustering Neural / Demographic

Value Prediction Radial Basis Functions does not try to fit a line to all values, but looks for groups of similar values. Radial Basis Functions represent the functions of the distance (or radius) from a particular point. RBF tries to find regions in the input space where the outputs are all similar and creates an RBF center for each group that predicts the average output of the group.

RBF is sophisticated and copes with out of line values (**noise**) by leaving a percentage of the data out and testing its model at the end of each pass against this 'held-out' data. If doing an extra pass did not help on this held-out data the radial basis functions stops.

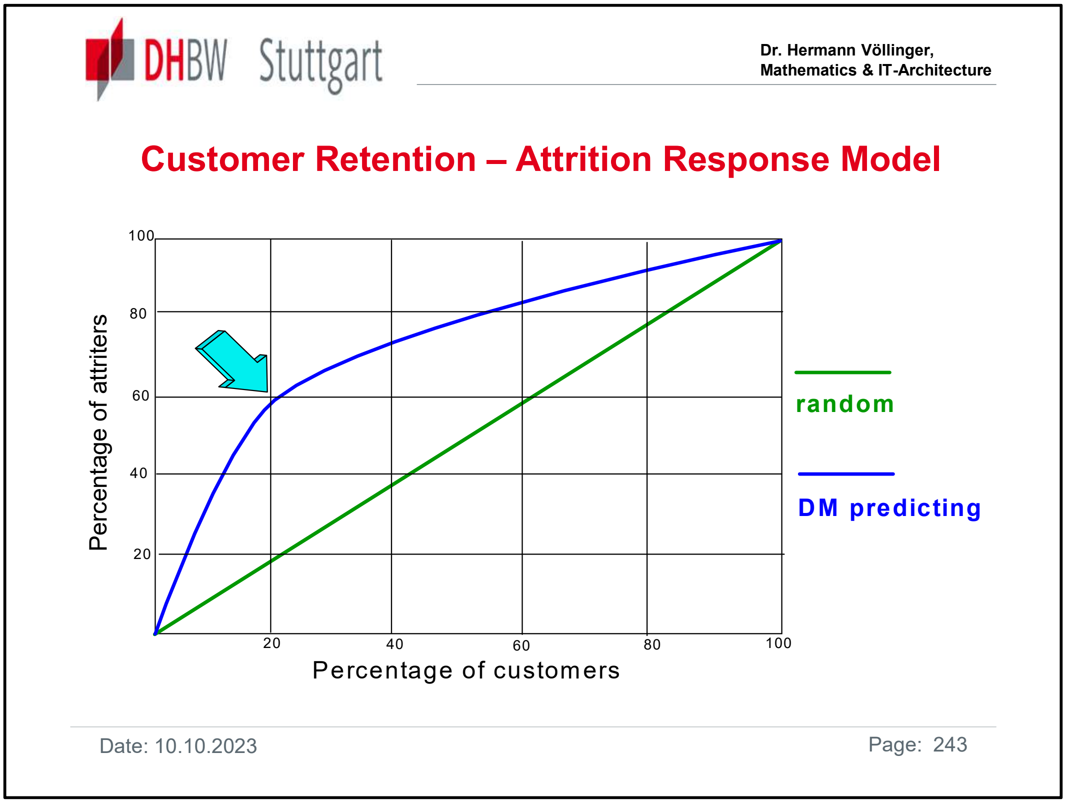
Value Prediction Neural Networks

The process for Neural Networks Value Prediction is similar to that for Classification. (Neural Value Prediction and RBF are fairly close too, but RBF is faster.)

As with RBF, Neural Value Prediction determines the value of a variable, based on known values of other variables.

If the value to be predicted is from a numeric variable, the result will also be a numeric, such as quantity or monetary value. If the value to be predicted is for a character variable, then the result is a number representing the probability of a specific value of the variable occurring.

Neural Value Prediction can predict values and time-series (multiple values, as a linear regression can predict a time series if you extend the line.) As with Neural Classification, it uses a back propagation algorithm, to see if it has passed the point where further improvements can be made.

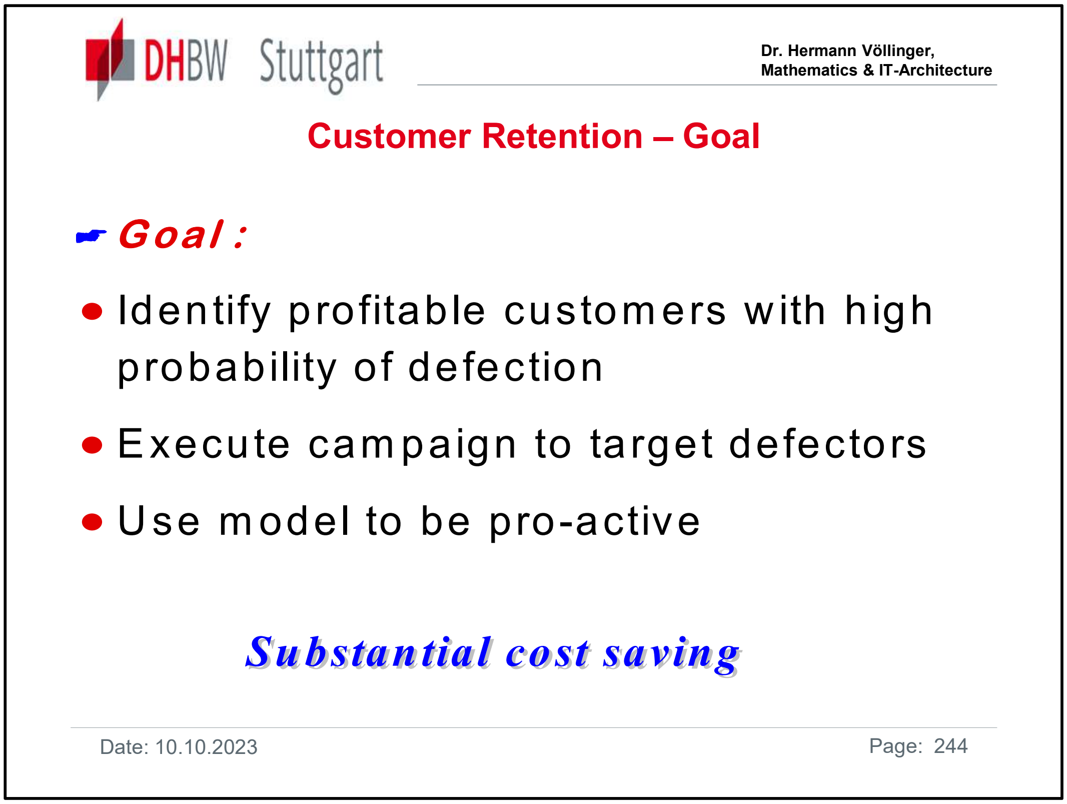

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Customer Retention – Attrition Response Model

Date: 10.10.2023

Page: 243

Consider the following attrition model, which shows a high volume of customers which a candidates to leave the enterprise.
The goal of Data Mining is to reduce this number to an average number, which is shown in the random curve.



DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Customer Retention – Goal

☛ ***Goal :***

- Identify profitable customers with high probability of defection
- Execute campaign to target defectors
- Use model to be pro-active

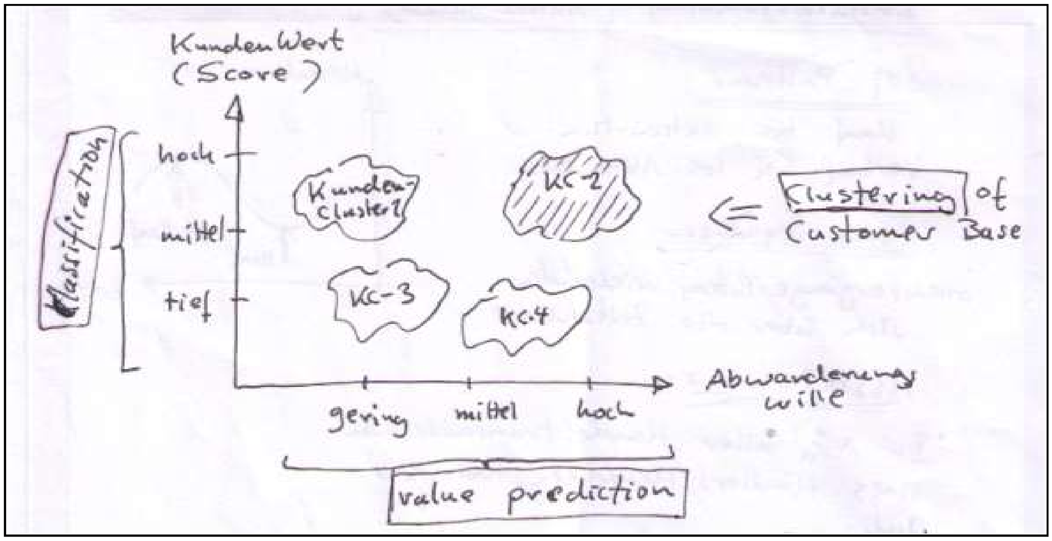
Substantial cost saving

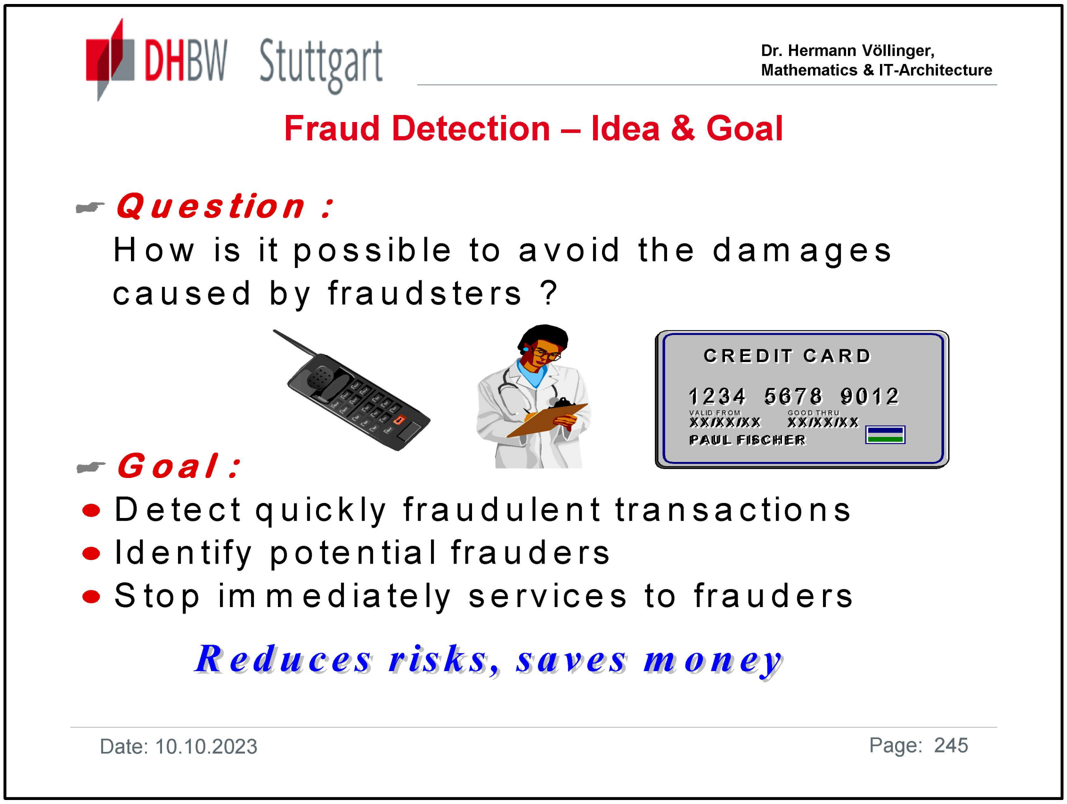
Date: 10.10.2023 Page: 244

The following picture gives a motivation to “Customer Retention”. In a first step you run Clustering to get Customer Clusters CC1, CC2, CC3 and CC4. The customers in each cluster have the two properties:

1. Value of Customer (“Classification”)
2. Probability of Defection =“Abwanderungswille” (“Value Prediction”).

Now place the customer clusters (CC) in this two axes. You will then see CCs will are valuable to start special marketing activities with them, to prevent these customer to go away as customers. See picture:





Fraud Detection – Idea & Goal

Question :

How is it possible to avoid the damages caused by fraudsters ?

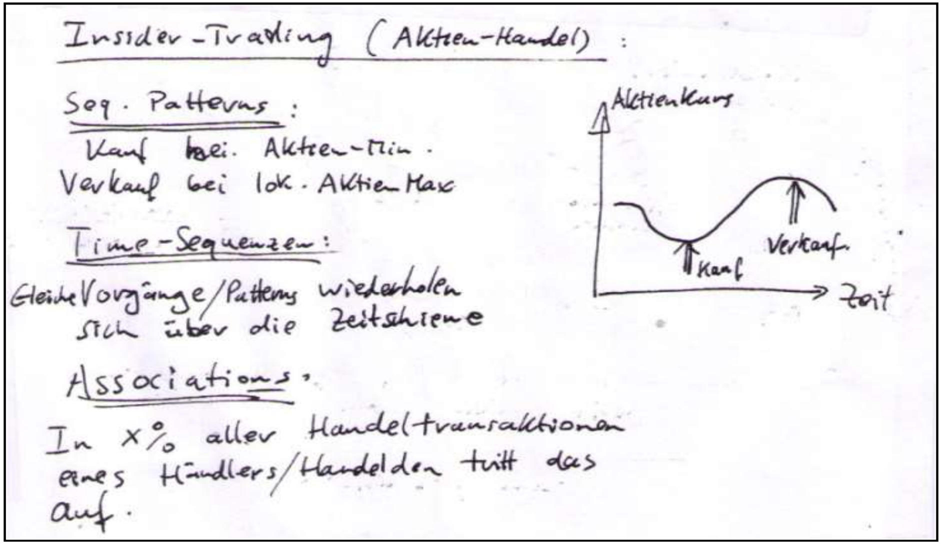
Goal :

- Detect quickly fraudulent transactions
- Identify potential frauders
- Stop immediately services to frauders

Reduces risks, saves money

The following picture gives a motivation to “Fraud Detection”. You use the 3 methods:
1. Sequential patterns
2. Time Sequences
3. Associations

See the example for “Insider Trading”:



Insider-Trading (Aktien-Handel) :

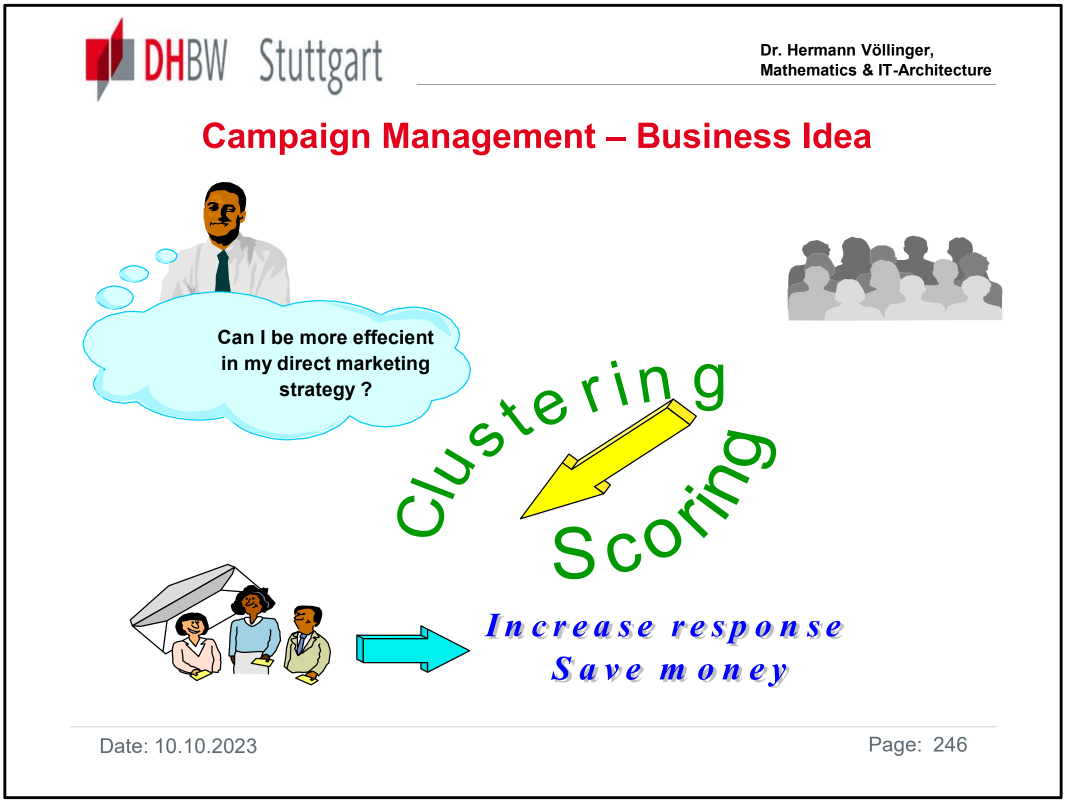
Seq. Patterns :
Kauf bei. Aktien-Min.
Verkauf bei lok. Aktien-Max

Time-Sequenzen:
Gleiche Vorgänge/Patterns wiederholen sich über die Zeitschiene

Associations:
In x% aller Handeltransaktionen eines Händlers/Handelnden tritt das auf.

Aktienkurs
Kauf
Verkauf
Zeit

Further examples of “Fraud Detection”:
- Fraud with credit cards
- Terroristic activities
- Crime cases
- Fraud in insurance. i.e. car insurance: examine all information/data about claims (look for pattern in a time sequences. Repeating = association)



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Campaign Management – Business Idea

Can I be more effecient in my direct marketing strategy ?

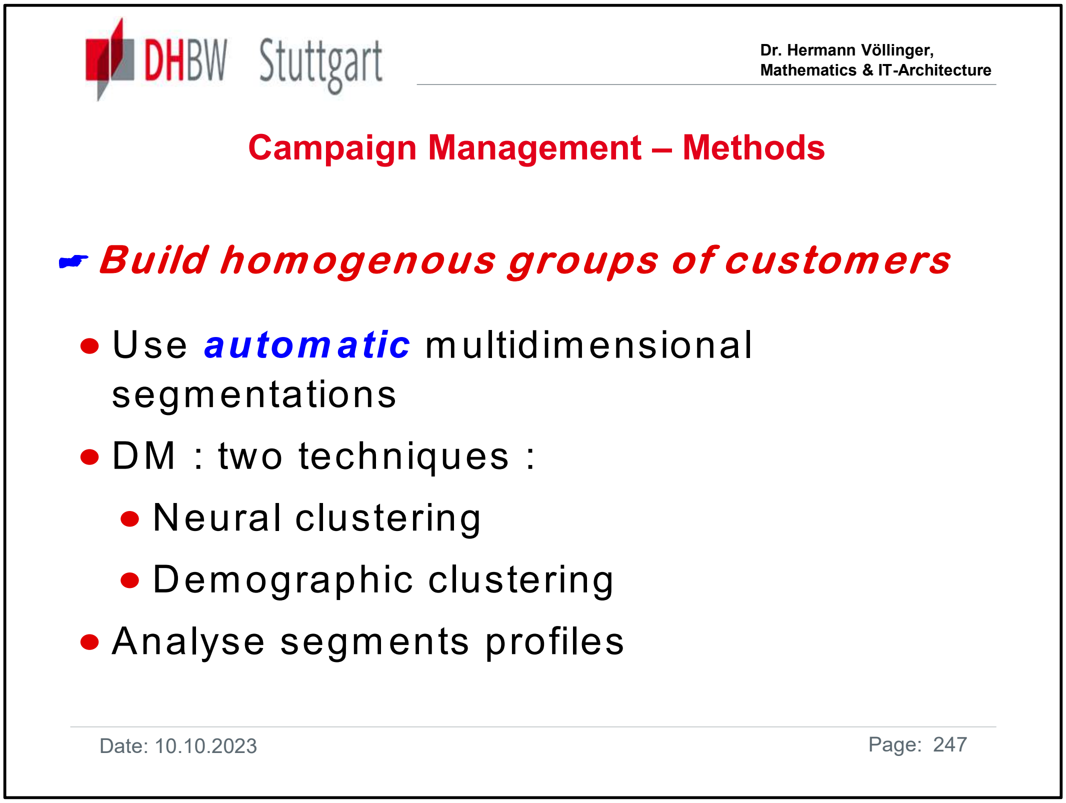
Clustering

Scoring

Increase response

Save money

The next data mining application is “Campaign Management”. The goal is to improve the success of marketing actions (i.e. marketing campaigns).
We will use for this clustering and also scoring methods (see next slides).

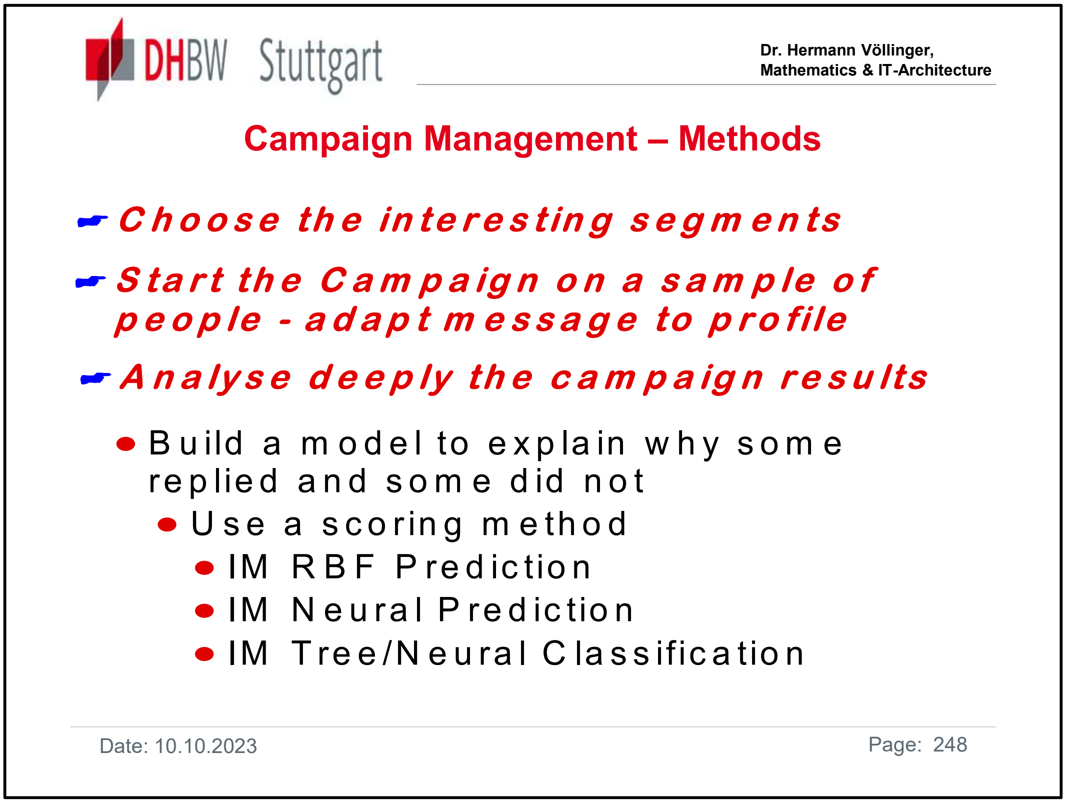


Campaign Management – Methods

☛ ***Build homogenous groups of customers***

- Use ***automatic*** multidimensional segmentations
- DM : two techniques :
  - Neural clustering
  - Demographic clustering
- Analyse segments profiles

First Method: Build customer segments – “Clustering”

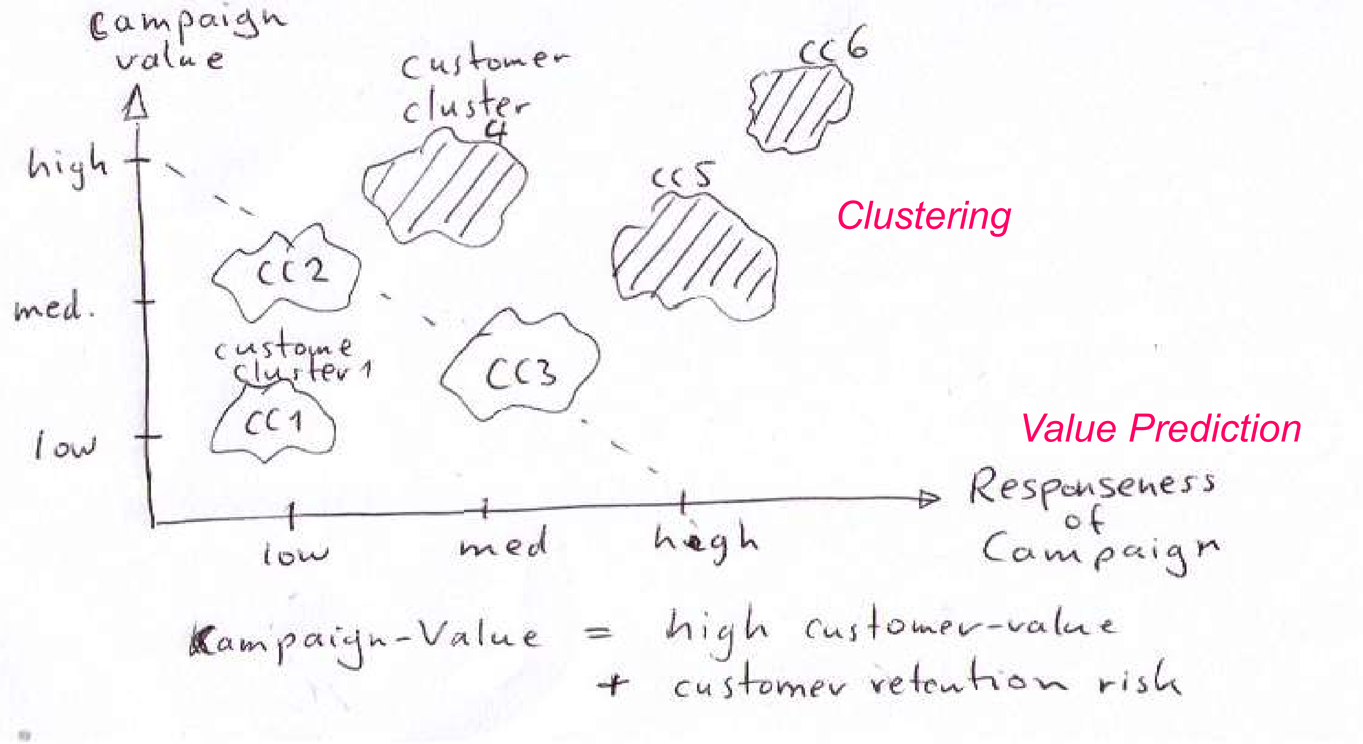


Campaign Management – Methods

- ***Choose the interesting segments***
- ***Start the Campaign on a sample of people - adapt message to profile***
- ***Analyse deeply the campaign results***
  - Build a model to explain why some replied and some did not
    - Use a scoring method
      - IM RBF Prediction
      - IM Neural Prediction
      - IM Tree/Neural Classification

For a motivation for “Campaign Management” see the following picture:

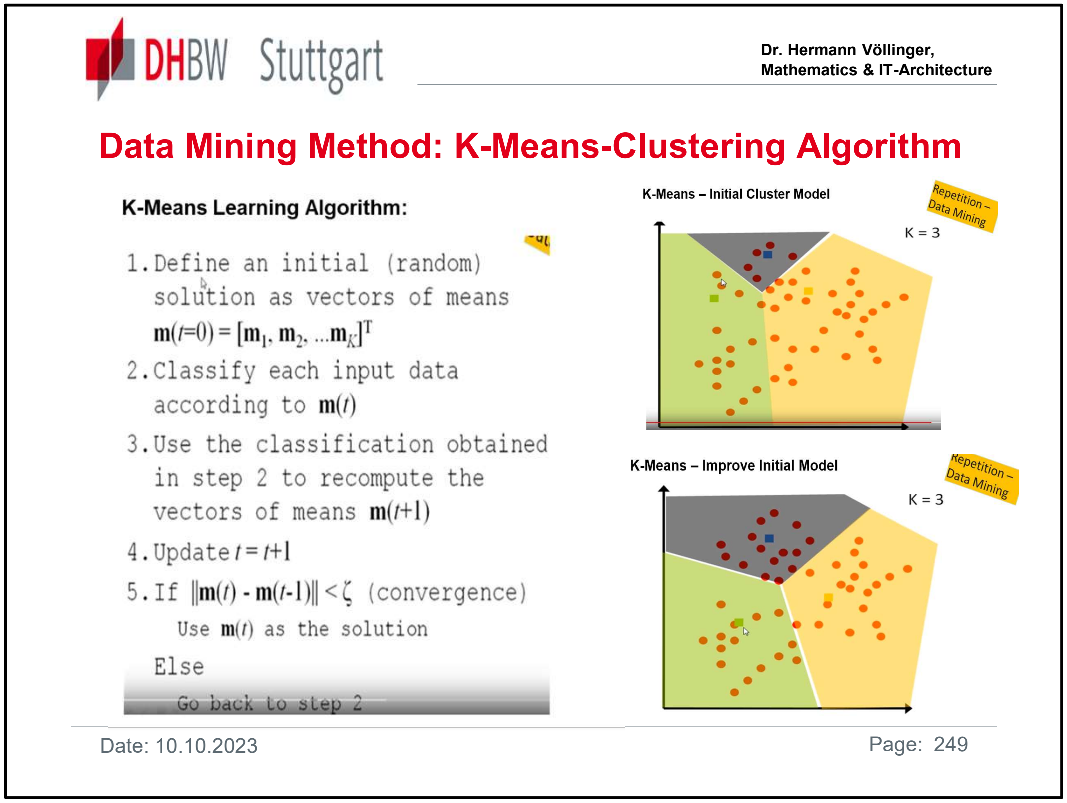
Classification (“Cust. Scoring”)
+ Value Prediction (“Retention Risk”)



Clustering

Value Prediction

Kampaign-Value = high customer-value + customer retention risk



Data Mining Method: K-Means-Clustering Algorithm

K-Means Learning Algorithm:

```
1. Define an initial (random)
   solution as vectors of means
   $\mathbf{m}(t=0) = [\mathbf{m}_1, \mathbf{m}_2, \ldots \mathbf{m}_K]^T$
2. Classify each input data
   according to $\mathbf{m}(t)$
3. Use the classification obtained
   in step 2 to recompute the
   vectors of means $\mathbf{m}(t+1)$
4. Update $t = t+1$
5. If $\|\mathbf{m}(t) - \mathbf{m}(t-1)\| < \zeta$ (convergence)
     Use $\mathbf{m}(t)$ as the solution
   Else
     Go back to step 2
```

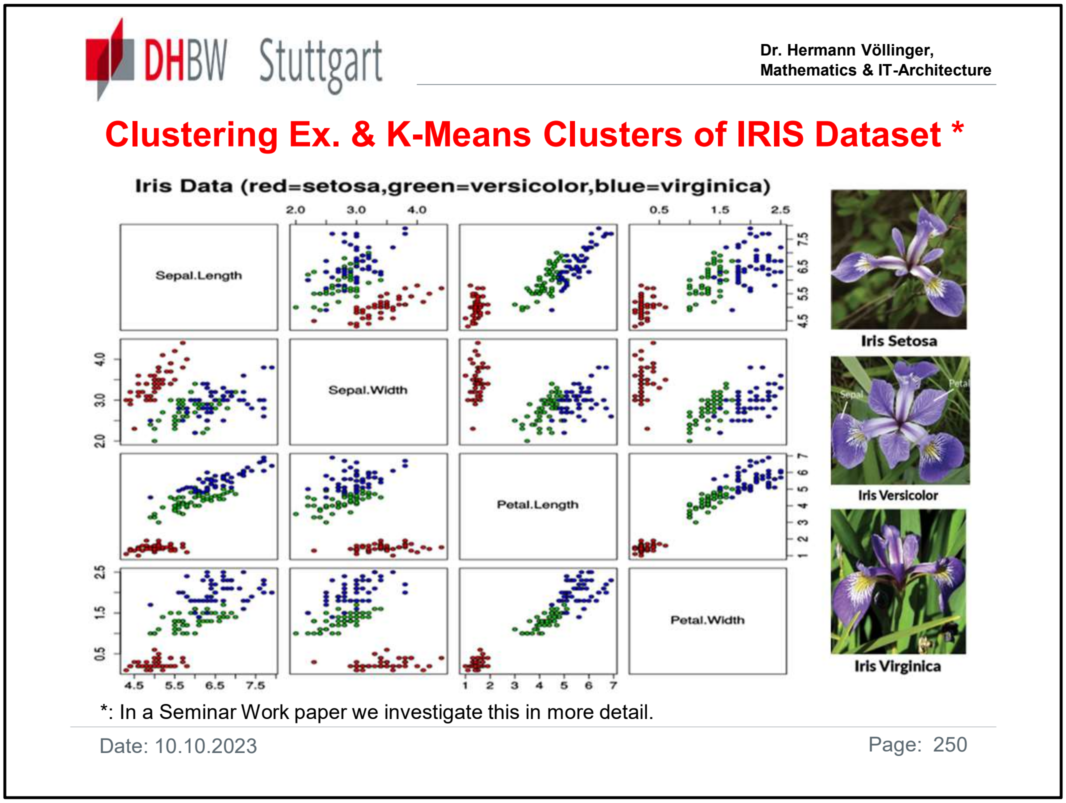
K-Means – Initial Cluster Model
Repetition – Data Mining
K = 3
K-Means – Improve Initial Model
Repetition – Data Mining
K = 3

K-means Clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

A cluster refers to a collection of data points aggregated together because of certain similarities. You'll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

In other words, the **K-means algorithm identifies k number of centroids**, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid. More details:

K-means algorithm: Let's see the steps on how the K-means machine learning algorithm works using the Python. We'll use the Scikit-learn library and some random data to illustrate a K-means clustering ….See more details under: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

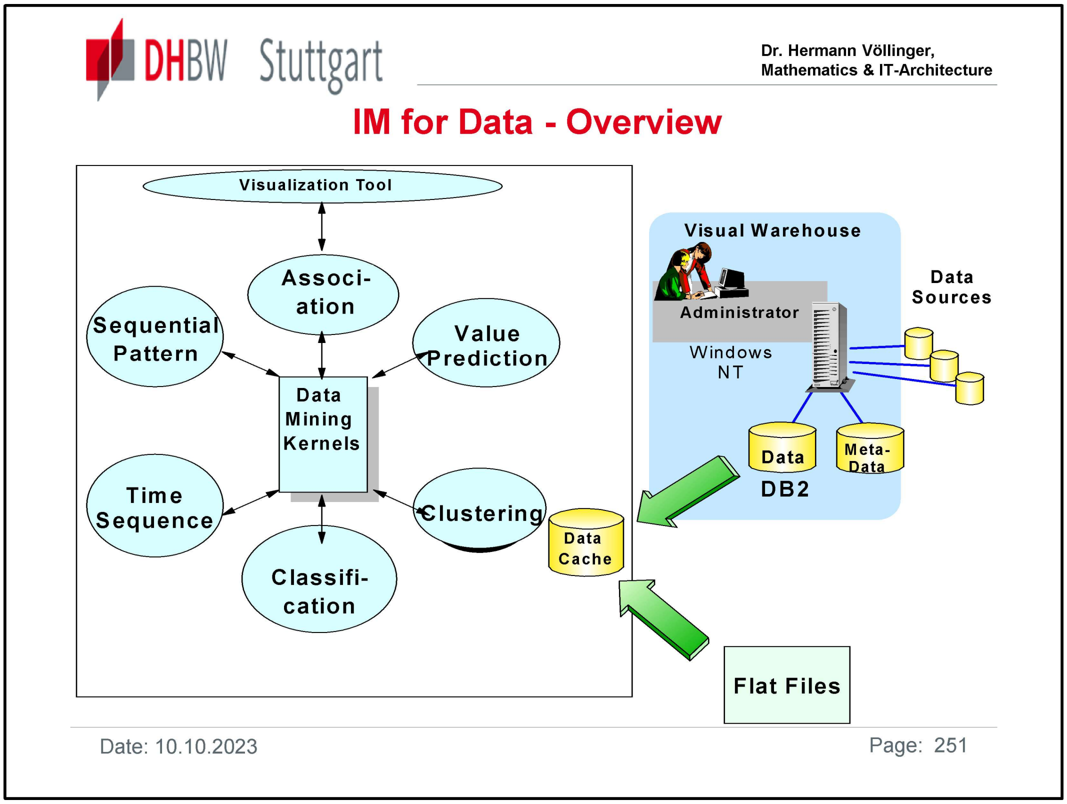


Clustering Ex. & K-Means Clusters of IRIS Dataset *

*: In a Seminar Work paper we investigate this in more detail.

How the K-means algorithm works: To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids It halts creating and optimizing clusters when either: The centroids have stabilized — there is no change in their values because the clustering has been successful. The defined number of iterations has been achieved.

K-means Clusters of IRIS Dataset: The Iris dataset contains the data for 50 flowers from each of the 3 species - Setosa, Versicolor and Virginica.
http://www.lac.inpe.br/~rafael.santos/Docs/CAP394/WholeStory-Iris.html
The data gives the measurements in centimeters of the variables sepal length and width and petal length and width for each of the flowers. Goal of the study is to perform exploratory analysis on the data and build a K-means clustering model to cluster them into groups. Here we have assumed we do not have the species column to form clusters and then used it to check our model performance. Since we are not using the species column we have an unsupervised learning method.
Develop a Python program by using the Scikit-learn library can bee see under:
https://github.com/bhattbhavesh91/k_means_iris_dataset/blob/master/K_in_K_means_Clustering.ipynb



IM for Data - Overview

Visualization Tool

Association

Sequential Pattern

Value Prediction

Data Mining Kernels

Time Sequence

Clustering

Classification

Data Cache

Visual Warehouse

Administrator

Windows NT

Data Sources

Data

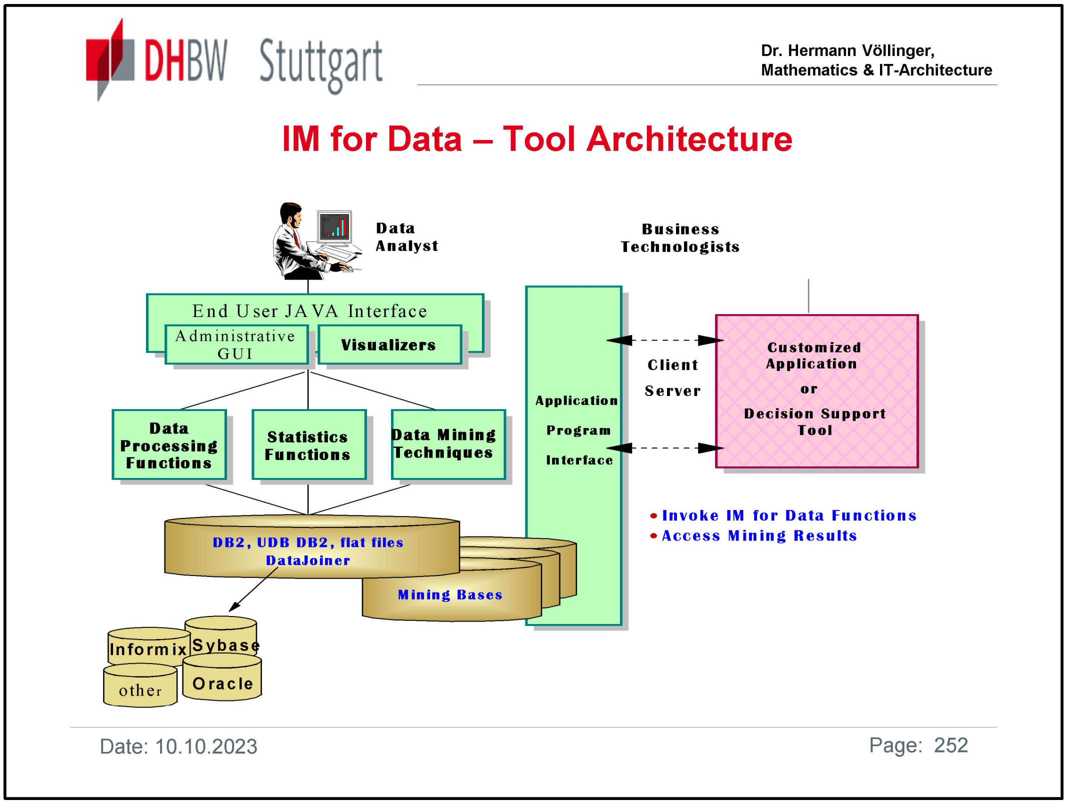
Meta-Data

DB2

Flat Files

IM4D consists of the following six basis Data Mining techniques:

1. Association
2. Sequential Pattern
3. Value Prediction
4. Clustering
5. Classification
6. Time Sequence



IM for Data – Tool Architecture

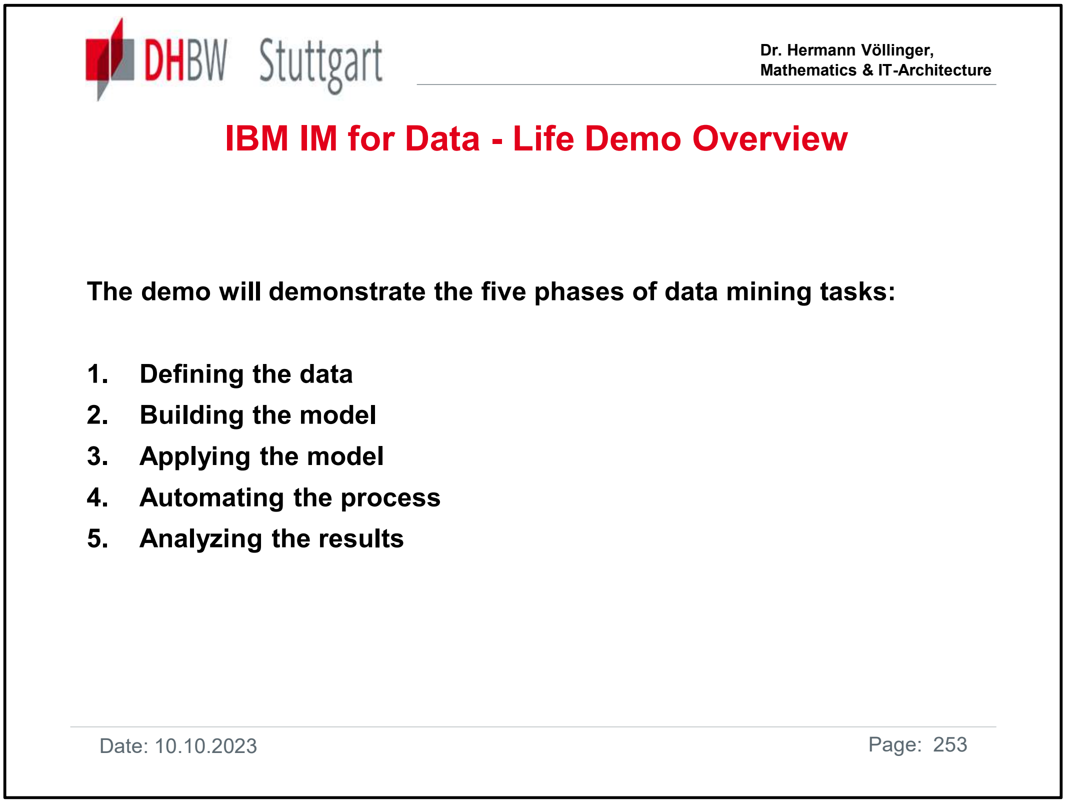
- Invoke IM for Data Functions
- Access Mining Results

Next we will see a demo of Im4Data:

The business problem

Imagine that you work for a bank that sells several products, including Regular Checking, Premier Checking, and Exclusive Checking accounts and option packages for each account. The bank already knows that Premier Checking is their most profitable product, and wants to increase the number of customers who have this type of checking account. The marketing department wants to identify
different groups based on demographic data, such as age and income, within the Premier Checking customers so that the department can prepare different ad campaigns for each of the groups. Additionally, the department wants to identify customers who are not currently Premier Checking customers who have similar demographics to the customers who are Premier Checking customers.

You have obtained some customer data from corporate headquarters to solve this business problem. This data is named **banking.txt**. It contains information about customers from all branches of the bank. You can use the Intelligent Miner to mine this data and provide demographic information to the marketing department. Your customer data includes information about customers who already have the Premier Checking account, so you can use the Demographic Clustering mining function to
identify different groups, based on demographic data, among customers who already have Premier Checking.



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

IBM IM for Data - Life Demo Overview

The demo will demonstrate the five phases of data mining tasks:

1. **Defining the data**
2. **Building the model**
3. **Applying the model**
4. **Automating the process**
5. **Analyzing the results**

This Intelligent Miner demo consists of an abbreviated data mining scenario with five phases: Defining data, building a model, applying the model, automating the process, and analyzing the results. By following the steps in this tutorial, you will learn how to use the Intelligent Miner wizards to define data objects, run mining functions, and view results in the Intelligent Miner.

Defining the data

Define a data object that points to a flat file containing your customer data file banking.txt. The data object will be named **Customers**.

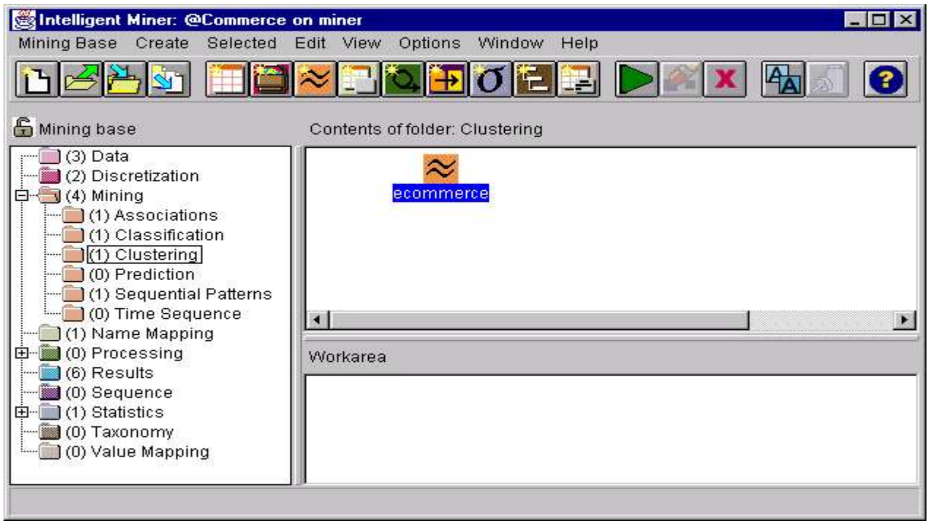
You must specify which properties of your customers are contained in the data, their data types, and the columns in the flat file that they occupy.

The Intelligent Miner data objects simply point to the location of your data, so that the Intelligent Miner can process this data. You will not actually be changing the contents of the banking.txt file.

Building the model

Define a Demographic Clustering settings object named **Build model**. This settings object uses the Customers data object as the input data. It runs in clustering mode, and produces a results object named **Model**. This model contains information that describes the clusters identified during the mining run.

IBM Intelligent Miner for Data - Life Demo

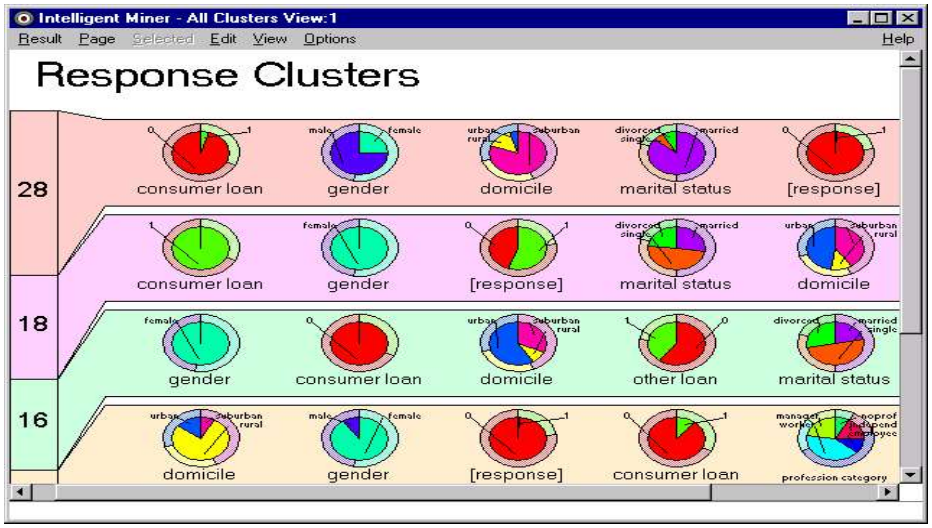


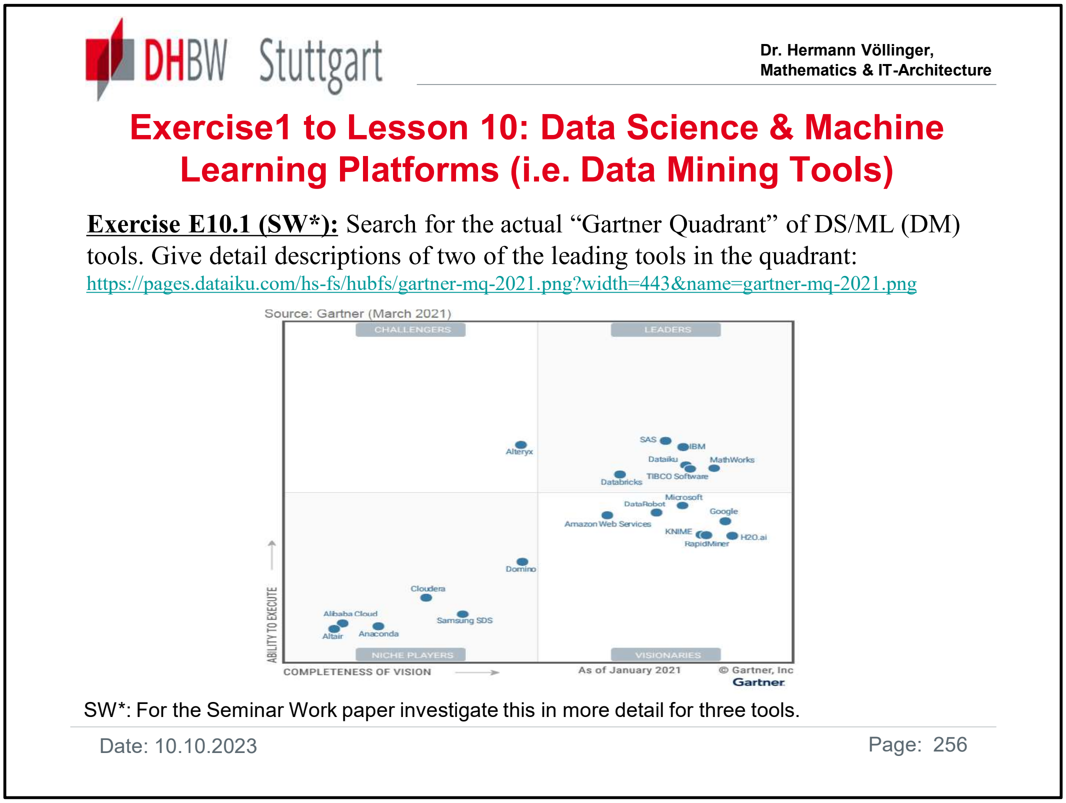
Applying the model:
Define a Demographic Clustering settings object named **Apply model**. This settings object uses the Customers data object as the input data. It runs in application mode using the Model results object and produces an output data object named **Scored customers** and a flat file named **scored.tx**t. This output file identifies the subgroup associated with a customer record.

Automating the process:
To automate the process you create a sequence object **Target Marketing** containing the Build model settings object and the Apply model settings object. A sequence is an object containing several other objects in a specific sequential order. You can run a sequence, which runs each of the objects within the sequence in the order that you specified. This allows you to combine several mining tasks into one step.

Analyzing the results:
Define a Bivariate Statistics function named **Analyze**. This statistical function analyzes the data object **Scored customers** and an produces an output data object **Target customers**, a flat file **target.txt**, and a result object **Target customer demographics**.

IBM Intelligent Miner for Data - Life Demo 2





DHBW Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise1 to Lesson 10: Data Science & Machine Learning Platforms (i.e. Data Mining Tools)

Exercise E10.1 (SW*): Search for the actual "Gartner Quadrant" of DS/ML (DM) tools. Give detail descriptions of two of the leading tools in the quadrant: https://pages.dataiku.com/hs-fs/hubfs/gartner-mq-2021.png?width=443&name=gartner-mq-2021.png

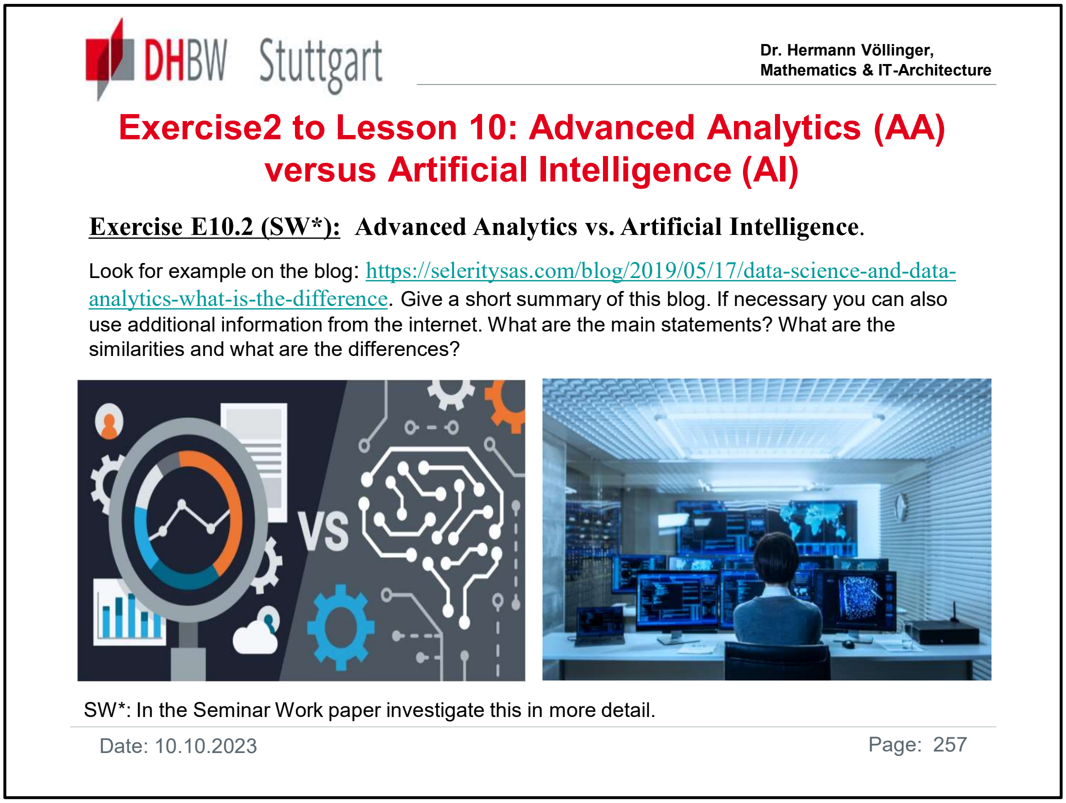
Source: Gartner (March 2021)

SW*: For the Seminar Work paper investigate this in more detail for three tools.

Date: 10.10.2023 Page: 256

This market features a diverse range of vendors: Gartner invited a wide range of data science and ML platform vendors. Users of these platforms, who include data scientists, citizen data scientists and application developers, have different requirements and preferences for user interfaces (UIs) and tools. Expert data scientists prefer to code data science models in Python or R, or to build and run data models in notebooks. Other users are most comfortable building models by using a point-and-click UI to create visual pipelines. Many members of emerging citizen data science communities favor a much more augmented approach that uses ML techniques "behind the scenes" to guide these less expert data scientists through the model building and operationalization process. Over time, expert data scientists may also come to prefer an augmented approach, which would enable them to navigate the model-building and operationalization process more efficiently. Tool and use case diversity is more important than ever. Further Information:

- Losers, and Trends in Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms. https://www.kdnuggets.com/2019/02/gartner-2019-mq-data-science-machine-learning-changes.html
- Forbes: Gartner's 2020 Magic Quadrant For Data Science And Machine Learning ...: https://www.forbes.com/sites/janakirammsv/2020/02/20/gartners-2020-magic-quadrant-for-data-science-and-machine-learning-platforms-has-many-surprises/?sh=329a9153f550
- [DHBW-Moodle] the document "Gartner-Machine_Learning_Platform.pdf": https://elearning.dhbw-stuttgart.de/moodle/mod/folder/view.php?id=187506



Exercise2 to Lesson 10: Advanced Analytics (AA) versus Artificial Intelligence (AI)

Exercise E10.2 (SW*): Advanced Analytics vs. Artificial Intelligence.

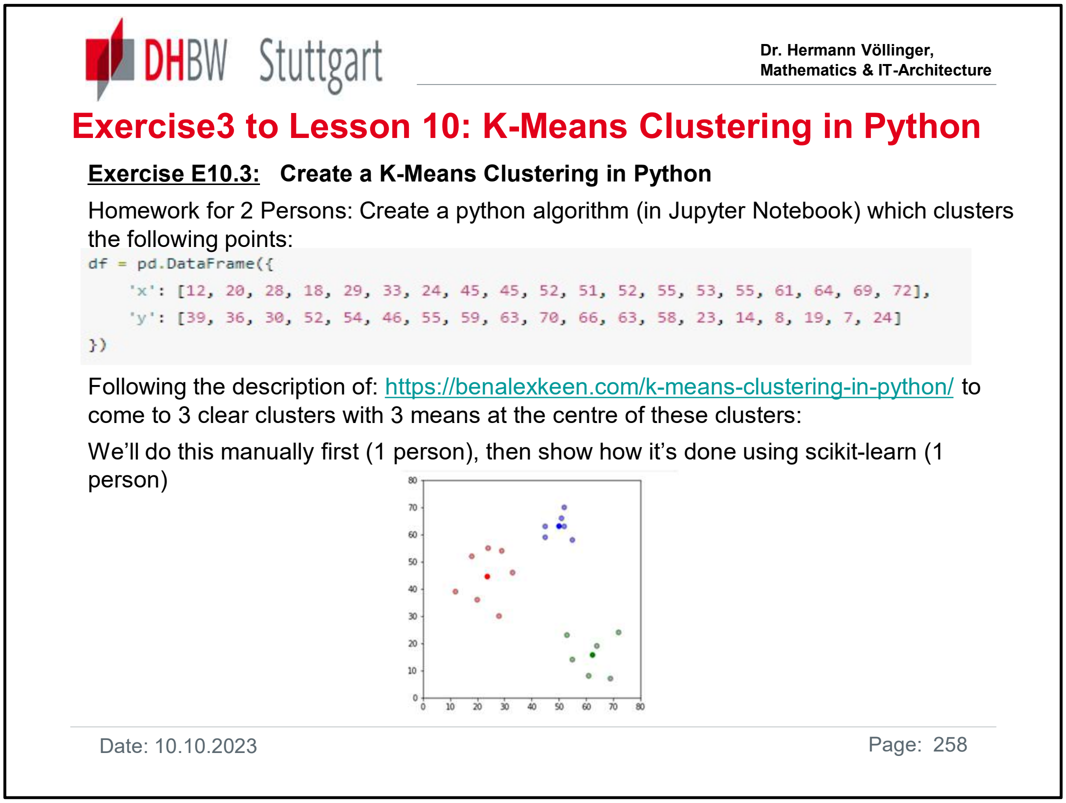
Look for example on the blog: https://selerityas.com/blog/2019/05/17/data-science-and-data-analytics-what-is-the-difference. Give a short summary of this blog. If necessary you can also use additional information from the internet. What are the main statements? What are the similarities and what are the differences?

SW*: In the Seminar Work paper investigate this in more detail.

Many companies today are using **advanced analytics** to gain insights from multiple sources of data, insights that can be used to make predictions and streamline business processes.
The number of companies using **artificial intelligence (AI)** to analyze massive volumes of data is growing at a fast pace.
Companies are using AI to find patterns in data and make predictions. Advanced analytics and AI sound like they are the same thing, but these terms are not synonymous. This post highlights some of the similarities and differences between advanced analytics and AI.

Further information links:

- https://www.forbes.com/sites/forbesagencycouncil/2018/08/01/do-you-know-the-difference-between-data-analytics-and-ai-machine-learning/?sh=3e3e38df5878
- https://hackr.io/blog/data-science-vs-data-analytics
- https://www.northeastern.edu/graduate/blog/data-analytics-vs-data-science/
- So kann Business Intelligence von KI profitieren (bi-scout.com)



Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Exercise3 to Lesson 10: K-Means Clustering in Python

Exercise E10.3: **Create a K-Means Clustering in Python**

Homework for 2 Persons: Create a python algorithm (in Jupyter Notebook) which clusters the following points:

```
df = pd.DataFrame({
    'x': [12, 20, 28, 18, 29, 33, 24, 45, 45, 52, 51, 52, 55, 53, 55, 61, 64, 69, 72],
    'y': [39, 36, 30, 52, 54, 46, 55, 59, 63, 70, 66, 63, 58, 23, 14, 8, 19, 7, 24]
})
```

Following the description of: https://benalexkeen.com/k-means-clustering-in-python/ to come to 3 clear clusters with 3 means at the centre of these clusters:

We'll do this manually first (1 person), then show how it's done using scikit-learn (1 person)

Repeat *K-Means* clustering (DM lesson or internet). Describe and explain the 4 necessary steps to reach the final cluster

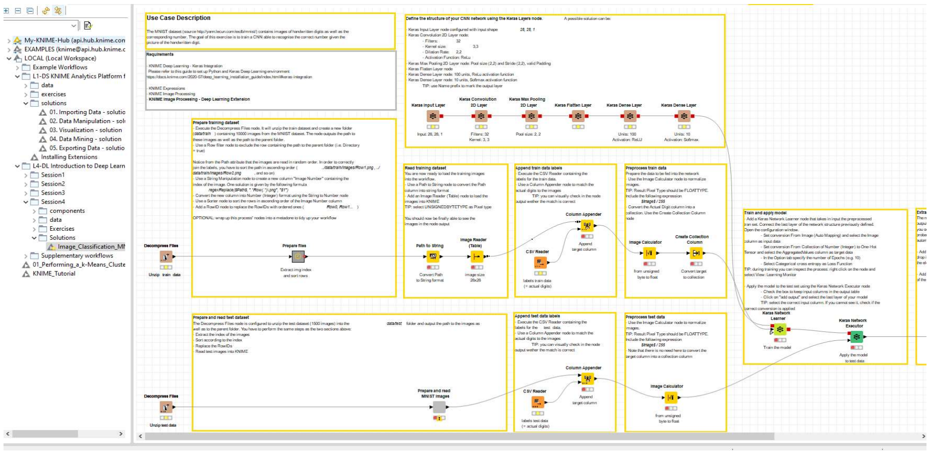
1. The centroids.
2. Assigning the first clusters.
3. Calculating the center of gravity and interacting.
4. The final clusters.

For an example of a solution see in: **[HVö-3]: Homework H3.4* – "Create a K-Means Clustering in Python"** http://www.dhbw-stuttgart.de/~hvoellin/
or in **[HVö-5]: Homework_H3.4_k-Means_Clustering.pdf**
https://github.com/HVoellinger/Lecture-Notes-to-ML-WS2020

Exercise 4 to Lesson 10: KNIME Image-Classification

<u>Exercise E10.4 (SW*)</u>: Image-Classification with MNIST Data using KNIME

Homework for 2 Persons: Rebuild the KNIME Workflow (use given solution) for Image-Classification and give technical explanations to the solution steps (see image below):


My-KNIME-Hub (api.hub.knime.com
EXAMPLES (knime@api.hub.knime.c
LOCAL (Local Workspace)
Example Workflows
L1-DS KNIME Analytics Platform f
data
exercises
solutions
01. Importing Data - solutio
02. Data Manipulation - sol
03. Visualization - solution
04. Data Mining - solution
05. Exporting Data - solutio
Installing Extensions
L4-DL Introduction to Deep Learn
Session1
Session2
Session3
Session4
components
data
Exercises
Solutions
Image_Classification_MN
Supplementary workflows
01_Performing_a_k-Means_Cluste
KNIME_Tutorial

Use Case Description

The MNIST dataset (source http://yann.lecun.com/exdb/mnist/) contains images of handwritten digits as well as the corresponding number. The goal of this exercise is to train a CNN able to recognise the correct number given the picture of the handwritten digit.

Requirements

- KNIME Deep Learning - Keras Integration
 Please refer to this guide to set up Python and Keras Deep Learning environment
https://docs.knime.com/2020-07/deep_learning_installation_guide/index.html#keras-integration

- KNIME Expressions
- KNIME Image Processing
- **KNIME Image Processing - Deep Learning Extension**

Prepare training dataset

- Execute the Decompress Files node. It will unzip the train dataset and create a new folder (data/train) containing 10000 images from the MNIST dataset. The node outputs the path to these images as well as the path to the parent folder.
- Use a Row filter node to exclude the row containing the path to the parent folder (i.e. Directory = true)

Notice from the Path attribute that the images are read in random order. In order to correctly join the labels, you have to sort the path in ascending order (*../data/train/images/Row1.png* , *../data/train/images/Row2.png* , and so on)
- Use a String Manipulation node to create a new column "Image Number" containing the index of the image. One solution is given by the following formula
 *regexReplace($Path$, ".*Row(.*)\.png", "$1")*
- Convert the new column into Number (Integer) format using the String to Number node
- Use a Sorter node to sort the rows in ascending order of the Image Number column
- Add a RowID node to replace the RowIDs with ordered ones (*Row0, Row1...*)

OPTIONAL: wrap up this process' nodes into a metadone to tidy up your workflow

Define the structure of your CNN network using the Keras Layers node. A possible solution can be:

- Keras Input Layer node configured with input shape 28, 28, 1
- Keras Convolution 2D Layer node:
  - Filters: 32
  - Kernel size: 3,3
  - Dilation Rate: 2,2
  - Activation Function: ReLu
- Keras Max Pooling 2D Layer node: Pool size (2,2) and Stride (2,2), valid Padding
- Keras Flatten Layer node
- Keras Dense Layer node: 100 units, ReLu activation function
- Keras Dense Layer node: 10 units, Softmax activation function
 TIP: use Name prefix to mark the output layer

Read training dataset

You are now ready to load the training images into the workflow.
- Use a Path to String node to convert the Path column into string format
- Add an Image Reader (Table) node to load the images into KNIME
TIP: select UNSIGNEDBYTETYPE as Pixel type

You should now be finally able to see the images in the node output

Append train data labels

- Execute the CSV Reader containing the labels for the train data.
- Use a Column Appender node to match the actual digits to the images
 TIP: you can visually check in the node output wether the match is correct

Preprocess train data

Prepare the data to be fed into the network
- Use the Image Calculator node to normalize images.
TIP: Result Pixel Type should be FLOATTYPE.
Include the following expression
 $Image$ / 255
- Convert the Actual Digit column into a collection. Use the Create Collection Column node

Train and apply model

- Add a Keras Network Learner node that takes in input the preprocessed tran set. Connect the last layer of the network structure previously defined. Open the configuration window.
  - Set conversion From Image (Auto Mapping) and select the Image column as input data
  - Set conversion From Collection of Number (Integer) to One-Hot Tensor and select the AggregatedValues column as target data
  - In the Option tab specify the number of Epochs (e.g. 10)
  - Select Categorical cross entropy as Loss Function

TIP: during training you can inspect the process: right click on the node and select View: Learning Monitor

- Apply the model to the test set using the Keras Network Executor node
  - Check the box to keep input columns in the output table
  - Click on "add output" and select the last layer of your model
 TIP: select the correct input column. If you cannot see it, check if the correct conversion is applied

Prepare and read test dataset

The Decompress Files node is configured to unzip the test dataset (1500 images) into the *data/test* folder and output the path to the images as well as to the parent folder. You have to perform the same steps as the two sections above:
- Extract the index of the images
- Sort according to the index
- Replace the RowIDs
- Read test images into KNIME

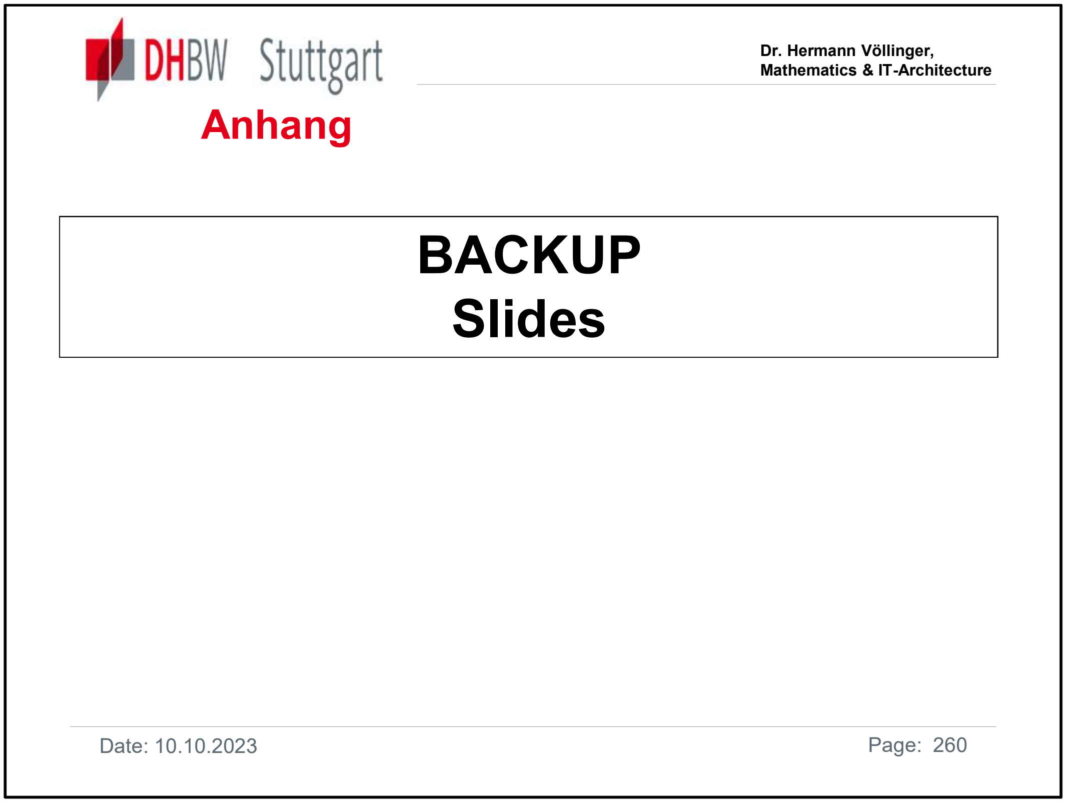
Append test data labels

- Execute the CSV Reader containing the labels for the test data.
- Use a Column Appender node to match the actual digits to the images
 TIP: you can visually check in the node output wether the match is correct

Preprocess test data

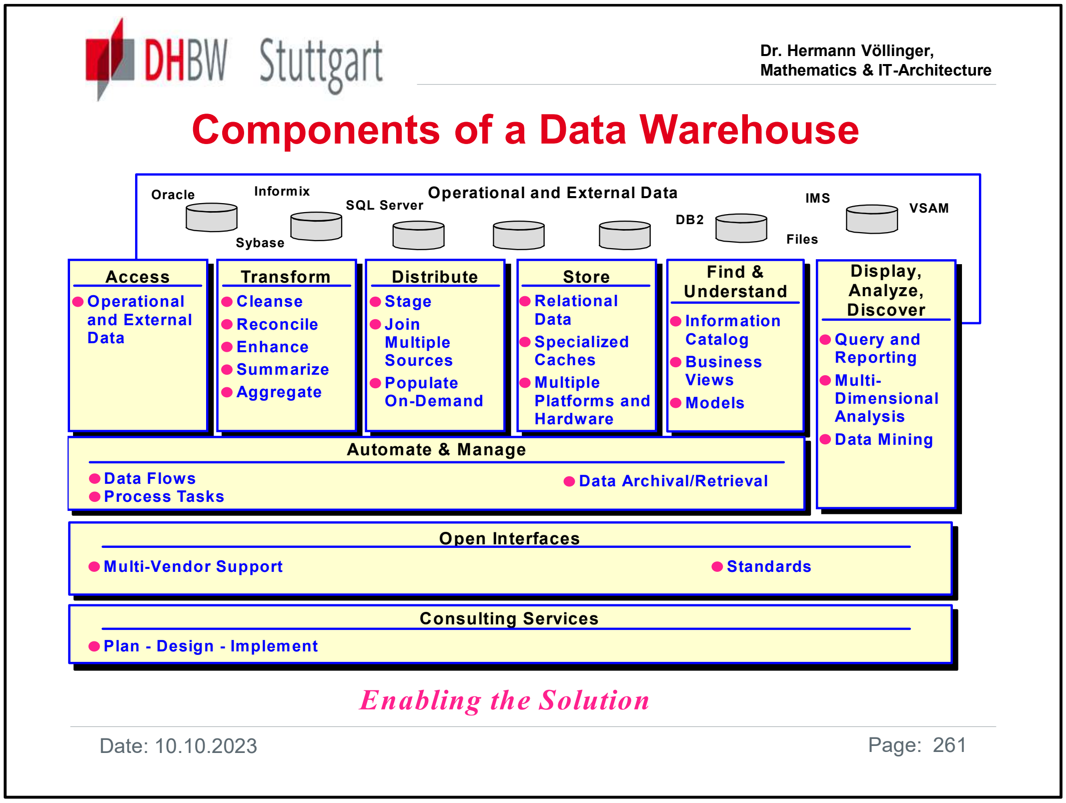
- Use the Image Calculator node to normalize images.
TIP: Result Pixel Type should be FLOATTYPE.
Include the following expression
 $Image$ / 255
- Note that there is no need here to convert the target column into a collection column

SW*: In the Seminar Work paper investigate this in more detail.



Anhang

BACKUP
Slides



Components of a Data Warehouse

Operational and External Data

Access
- Operational and External Data

Transform
- Cleanse
- Reconcile
- Enhance
- Summarize
- Aggregate

Distribute
- Stage
- Join Multiple Sources
- Populate On-Demand

Store
- Relational Data
- Specialized Caches
- Multiple Platforms and Hardware

Find & Understand
- Information Catalog
- Business Views
- Models

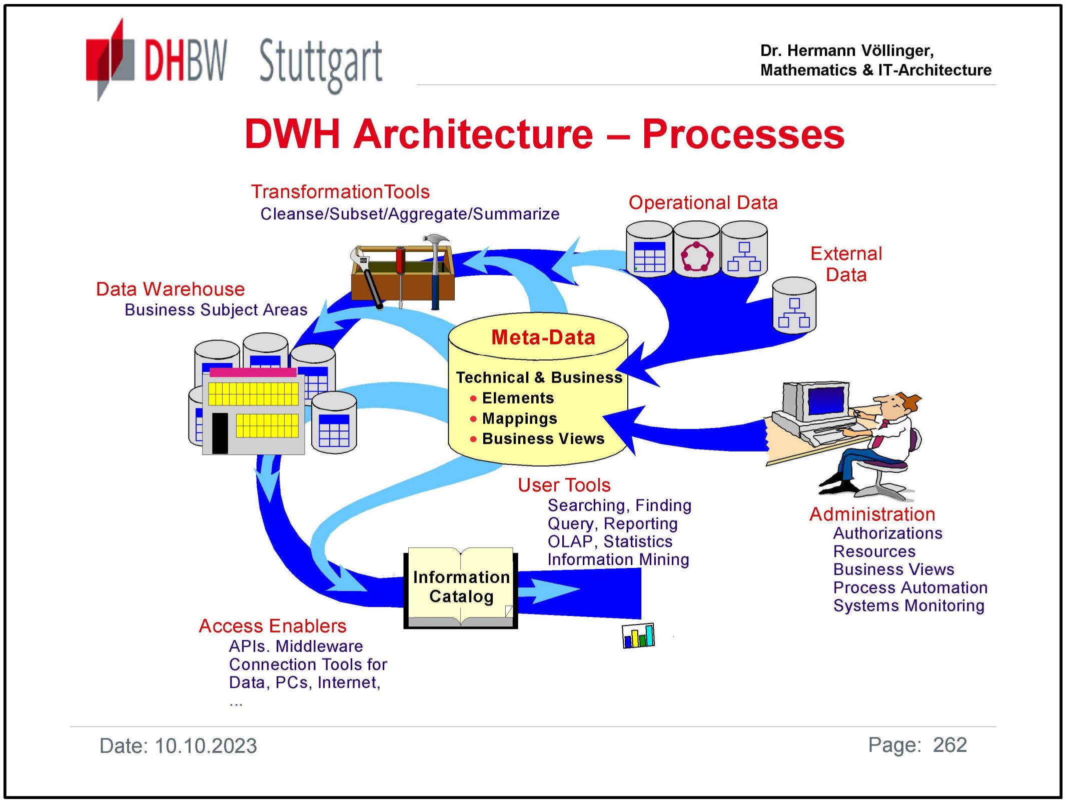
Display, Analyze, Discover
- Query and Reporting
- Multi-Dimensional Analysis
- Data Mining

Automate & Manage
- Data Flows
- Process Tasks
- Data Archival/Retrieval

Open Interfaces
- Multi-Vendor Support
- Standards

Consulting Services
- Plan - Design - Implement

Enabling the Solution



DWH Architecture – Processes

TransformationTools
Cleanse/Subset/Aggregate/Summarize

Operational Data

External
Data

Data Warehouse
Business Subject Areas

Meta-Data

Technical & Business
- **Elements**
- **Mappings**
- **Business Views**

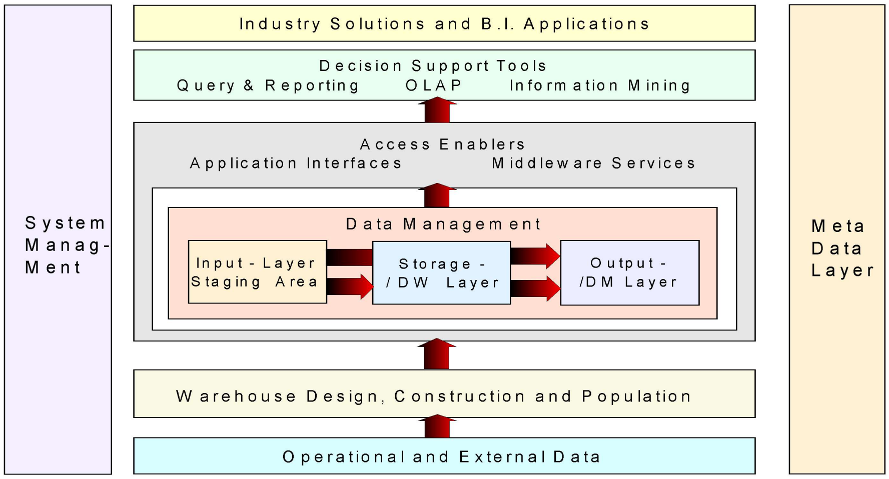
User Tools
Searching, Finding
Query, Reporting
OLAP, Statistics
Information Mining

Administration
Authorizations
Resources
Business Views
Process Automation
Systems Monitoring

Information Catalog

Access Enablers
APIs. Middleware
Connection Tools for
Data, PCs, Internet,
...

Process Layers of the DWH



Industry Solutions and B.I. Applications

Decision Support Tools
Query & Reporting OLAP Information Mining

Access Enablers
Application Interfaces Middleware Services

Data Management

Input - Layer
Staging Area

Storage -
/ DW Layer

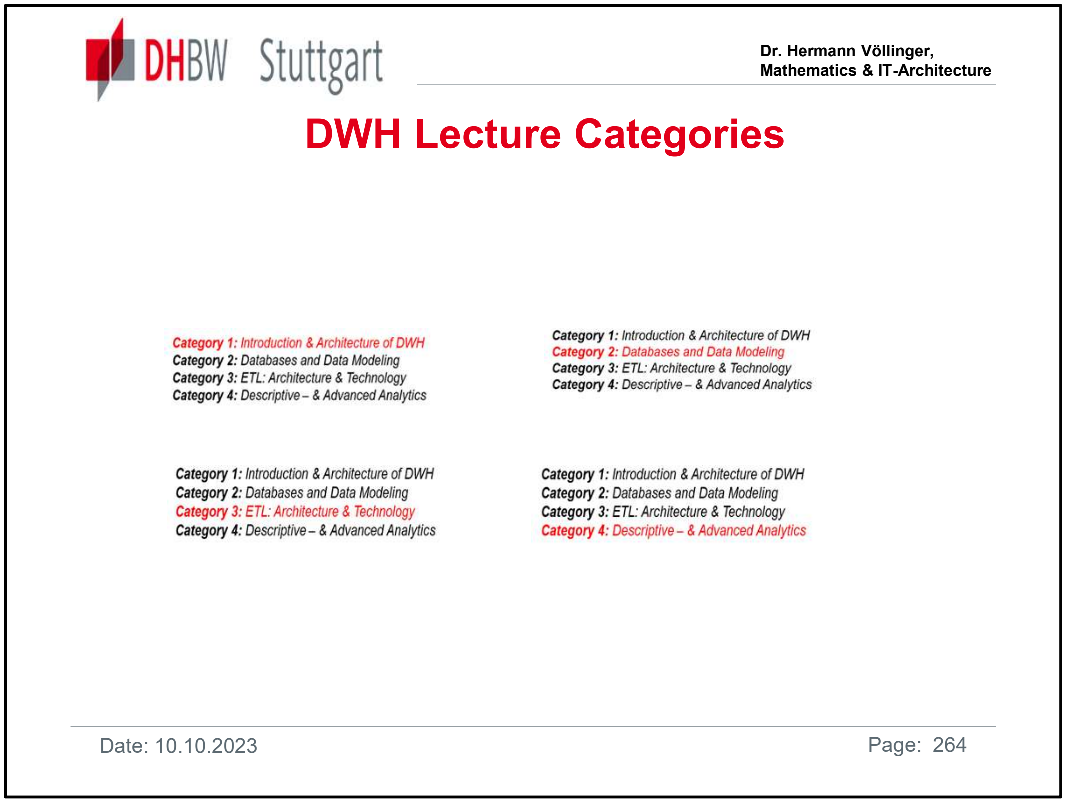
Output -
/DM Layer

Warehouse Design, Construction and Population

Operational and External Data

System
Manag-
Ment

Meta
Data
Layer



DWH Lecture Categories

Category 1: *Introduction & Architecture of DWH*
Category 2: *Databases and Data Modeling*
Category 3: *ETL: Architecture & Technology*
Category 4: *Descriptive – & Advanced Analytics*

Category 1: *Introduction & Architecture of DWH*
Category 2: *Databases and Data Modeling*
Category 3: *ETL: Architecture & Technology*
Category 4: *Descriptive – & Advanced Analytics*

Category 1: *Introduction & Architecture of DWH*
Category 2: *Databases and Data Modeling*
Category 3: *ETL: Architecture & Technology*
Category 4: *Descriptive – & Advanced Analytics*

Category 1: *Introduction & Architecture of DWH*
Category 2: *Databases and Data Modeling*
Category 3: *ETL: Architecture & Technology*
Category 4: *Descriptive – & Advanced Analytics*