Exercices (+Solutions) to DHBW Lecture Intro2DWH

by

Dr. Hermann Völlinger and Other

Status: 27 October 2022

Goal: Documentation of all Solutions to the Homework/Exercises in the Lecture "Introduction to Data Warehouse (DWH)".

Please send your solutions (if you want) to your lecturer:
hermann.voellinger@gmail.com

Authors of the Solutions: Dr. Hermann Völlinger and Other

Content

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

Exercises (+Solutions) to DHBW Lecture Intro2DWH – Chapter 1

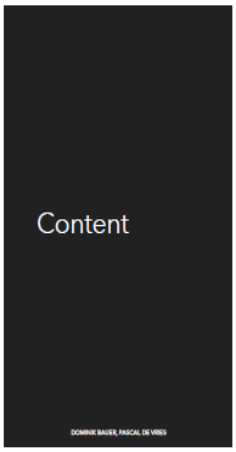
Exercise E1.1*: Investigate the BI-Data Trends in 2021.

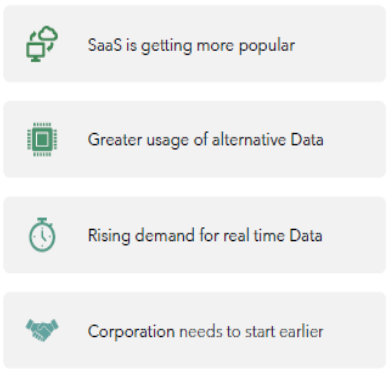
Prepare and present the results of the e-book "BI_ Daten_Trends _2021" TINF18D-DWH: Supporting Material (dhbw-stuttgart.de) in the next exercise session (next week, duration = 20 minutes). 2 students.

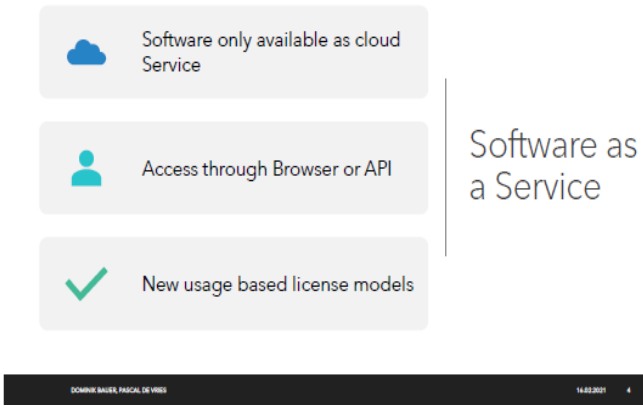
Task: Show how can DWH and BI help to overcome the current problems (i.e. corona pandemic) and build the basics for more digitalization. Examine the ten data trends to support the new digital requirements.

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

Solution:







Impact on digital requirements?

- Availability
- Scalability and elasticity
- Cost
- Complex planning

DOMINIK BAUER, PASCAL DE VRIES 14.02.2021 5

Short Term	Medium Term	Long Term
Immediate switch to more SaaS solutions	First the important data storage systems are migrated to the cloud, then the applications.	By 2022, public cloud services will be the foundation for 90% of innovations in data and analytics.

Development

DOMINIK BAUER, PASCAL DE VRIES 14.02.2021 6

Current Situation

- Small projects are already use SaaS
  - Projects started in the last year
- Corona moved Projects to the cloud
  - Projects that rely on availability
  - Cost's savings

DOMINIK BAUER, PASCAL DE VRIES 14.02.2021 7

Greater usage of alternative Data

WHAT WAS THAT AGAIN?

DOMINIK BAUER, PASCAL DE VRIES 14.02.2021 8

Impact on digital requirements?

- Availability
- Scalability and elasticity
- Cost
- Complex planning

DOMINIK BAUER, PASCAL DE VRIES 14.02.2021 5

Short Term	Medium Term	Long Term
Immediate switch to more SaaS solutions	First the important data storage systems are migrated to the cloud, then the applications.	By 2022, public cloud services will be the foundation for 90% of innovations in data and analytics.

Development

DOMINIK BAUER, PASCAL DE VRIES 14.02.2021 6

Current Situation

- Small projects are already use SaaS
  - Projects started in the last year
- Corona moved Projects to the cloud
  - Projects that rely on availability
  - Cost's savings

DOMINIK BAUER, PASCAL DE VRIES 14.02.2021 7

Greater usage of alternative Data

WHAT WAS THAT AGAIN?

DOMINIK BAUER, PASCAL DE VRIES 14.02.2021 8

Alternative Data

- Usage of external/non-traditional Data
- Finding corelations in the Data
- Extracting new Information form the corelations

Impact on digital requirements?

- Need to make external Data accessible
- Experts to find correlations
- The extracted Information must be trustworthy
- The Information then can be used to gain benefits

Short Term	Medium Term	Long Term
Companies realise the value of alternative Data.	Correlation detection and Information extraction from alternative Data will be a Standard analysis process.	Content analysis with alternative Data will be used in 75% of the fortune 500 Companies to achieve changes and innovations.

Development

Current Situation

- Earlier discovery of Corona Virus
  - Analyse search requests
  - Monitor Traffic around Hospitals
- Predict Corona Hotspots
  - Scientists started analysing wastewater
  - Can predict the number of Infected Persons in the Area
  - Possible use as an early warning system

Rising demand for real time Data

I NEED FRESH DATA. NOW.

Realtime Data

- Usage of external/non-traditional Data
- Finding corelations in the Data
- Extracting new Information form the corelations

Impact on digital requirements?

- Make real time Data accessible
- Make real time Legal requirements for real time data
- Faster response to new information
- Faster / automated analysis

Short Term	Medium Term	Long Term
Capturing data changes faster is critical to success.	The responses based on this data must be at the same pace as the business processes. This is decisive for the step from reactive to pre-active action.	By the end of 2024, 75% of companies will no longer use AI only as a pilot project, but operationally. As a result, data streaming and analysis infrastructures will grow by a factor of 5

Development

Current Situation

- Fast changing Demand
  - Toilet paper
  - Personal protection eqipment
- Fast chaning Supplychains
  - Diffent kinds of Covid-19 in different countries.
  - Fast changing political landscape

DOMINIK BAUER, PASCAL DE VRIES 16.02.2021 17

Corporation needs to start earlier

BUT WHY?

DOMINIK BAUER, PASCAL DE VRIES 16.02.2021 18

What is changing?

CLASSES GET ONLINE

MORE PEOPLE WORK FROM HOME

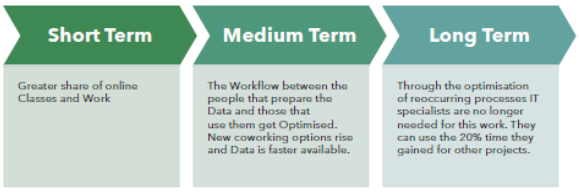
EVERYONE IS CONNECTED ALL THE TIME

DOMINIK BAUER, PASCAL DE VRIES 16.02.2021 19

Results

- There is less time to prepare the Data
- The need for a simple way to access and prepare Data rises
- This must be implemented in Coworking Software
- This would allow to work together more efficient and earlier

DOMINIK BAUER, PASCAL DE VRIES 16.02.2021 20



Development

DOMINIK BAUER, PASCAL DE VRIES 16.02.2021 21

Thank you for Listening

DOMINIK BAUER, PASCAL DE VRIES 16.02.2021 22

Sources

https://elearning.dhbw-stuttgart.de/moodle/pluginfile.php/330418/mod_folder/content/0/BI_Daten_Trends_2021.pdf?forcedownload=1

https://www.aerzteblatt.de/nachrichten/115929/Forscher-ueberpruefen-Abwasser-auf-Coronaviren

DOMINIK BAUER, PASCAL DE VRIES 16.02.2021 23

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

Task: 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.

***Theme**:* Trends or new development in the following areas (project reports are also possible):

1. **Data Warehousing (DWH)**
2. **Business Intelligence (BI)**
3. **Customer Relationship Management (CRM) (operational, analytical, collaborative)**

For Explanation of these 'catchwords' see also the slides of the lesson or search in the internet

Optional: Give an explanation also for the synonyms like: OLAP, OLTP, ETL, ERP, EAI

Solution:

DWH – Data Warehousing:
In vielen Organisationen sammeln sich in den operativen Systemen große, isolierte und meist unterschiedlich formatierte Datenmengen an. Durch Transformation dieser Daten und hinzufügen externer Daten wird es möglich, Informationen integriert im Data Warehouse – eine Art Warenlager für Daten – für Abfragen und weitergehenden Analysen bereitzustellen.

BI – Business Intelligence:
BI ist der Prozess, die angesammelten, rohen, operationalen Daten zu analysieren und sinnvolle Informationen daraus zu extrahieren, um auf Basis dieser integrierten Informationen bessere Geschäftsentscheidungen treffen zu können.

BI ist wenn Geschäftsprozesse anhand der aus dem Data Warehouse gewonnenen Fakten optimiert werden.

CRM – Customer Relationship Management:
CRM steht für kundenorientiertes Handeln, d.h. nicht das Produkt, sondern der Kunde ist Mittelpunkt aller Geschäftsentscheidungen. Durch besseren und individuelleren Service sollen neue Kunden gewonnen und bestehende Kundenkontakte gepflegt werden.

Operatives CRM:
Lösungen zur Automatisierung / Unterstützung von Abwicklungsprozessen mit Kunden (Online Shop, Call Center,…)

Analytisches CRM:
Lösungen, die auf Informationen des Data Warehouse zurückgreifen und auf aufgabenspezifische Analysen (Data Mining) beruhen.

Kollaboratives CRM:
Kommunikationskomponente, die die Interaktion mit dem Kunden ermöglicht.

Gewinnung von Erkenntnissen durch Zusammenarbeit mit dem Kunden. Diese können dann zur Optimierung der Geschäftsprozesse oder Personalisierung der Kundenbeziehung genutzt werden.

OLAP – Online Analytical Processing:

Der Begriff OLAP fasst Technologien, also Methoden, wie auch Tools, zusammen, die die Ad-hoc Analyse multidimensionaler Daten unterstützen. Die Daten können aus dem Data Warehouse, Data Marts oder auch aus operativen Systemen stammen.

(Abgrenzung Data Mining: Suche nach Mustern und bislang unbekannten Zusammenhängen (Neuronale Netze, Warenkorbanalysen,…))

OLTP – Online Transactional Processing:

Operative Softwaresysteme mit deren Transaktionsdaten. Heute analysiert man weniger diese operationalen Daten als vielmehr multidimensionale, navigierbare Daten (OLAP).

ETL – Extraction, Transformation and Loading:

Ein ETL – Tool ist dafür zuständig, um aus den operationalen Daten (real-time-data) gesäuberte und eventuell aggregierte Informationen sowie zusätzliche Metadaten zu erhalten.

ERP – Enterprise Resource Planning:

Unternehmensübergreifende SW-Lösungen, die zur Optimierung von Geschäftsprozessen eingesetzt werden. Dabei handelt es sich um integrierte Lösungen, die den betriebswirtschaftlichen Ablauf in den Bereichen Produktion, Vertrieb, Logistik, Finanzen und Personal steuern und auswerten.

EAI – Enterprise Application Integration:

EAI beschäftigt sich mit der inner- und über-betrieblichen Anwendungsintegration, um einen problemlosen Daten- und Informationsaustausch zu gewährleisten.

Aktuelle Trends:

1) **Explodierendes Datenvolumen**
   - Stärkster Trend
   - Laut Gartner soll 2004 das Datenvolumen 30x so hoch wie 1999 sein.
   - Skalierbarkeit

2) **Integrierte 360° Sicht**
   - Der Kunde soll völlig transparent sein
     ⇒ Trotz verteilter Applikationen soll ein vollständiges Bild des Kunden vorhanden sein. → wichtig für CRM

3) **Komplexe Anfragen und Analysen**
   - Benutzeranforderungen an DWH- / BI- und CRM- Systeme steigen
   - Anfragen nehmen zudem zu

4) **Mehr Endbenutzer**
   - BI- und DWH- Systeme müssen zugänglicher werden
     ⇒ Benutzbarkeit „weniger ist mehr“

5) **Fussion von DWH und CRM**
   - Information (in den DWH's) ist die Basis, um Kunden zu versehen

6) **Active DWH**
   - Wettbewerbsdruck → Daten müssen schnell da sein
   - Aktive DWH sind eng an operationale Systeme gekoppelt → sehr aktuelle Daten + sehr detailliert

7) **Datenansammlungen ('Data Hubs') statt relationaler DBs**
   - Billiger + schneller, aber: kein SQL + nicht für jede Situation

8) **Outsourcing**
   - Zu Anfang Applikationen + Daten; zukünftig auch die Informationshaltung im DWH

9) **Starkes Anwachsen von Datenquellen (z.B. e-Business)**
   - Mehr Daten in unterschiedlichen Plätzen

10) **Re-Engeneering oder sogar Neuaufbau von Business- Systemen (DWH, …)**
    - Kunde war nicht Mittelpunkt oder wurde nicht vollständig betrachtet; Falschplanung (Größe, Geschwindigkeit, …)

Further Solution (SS 2014):

Data Warehouse

Überblick und Trends

Agenda

- Motivation
- Überblick
- DWH Trends
- Zusammenfassung

Agenda

- Motivation
- Überblick
- DWH Trends
- Zusammenfassung

Motivation

- Wachsende heterogene Datenbestände in Unternehmen
- Erschwerte Entscheidungsfindung aufgrund zunehmender Komplexität
- Analyse und Auswertung muss gewährleistet sein

→ DWH unterstützt Lösung dieser Problemstellungen
→ "Turning Data into Information!"

[zitiert aus DWH Skript Dr. Hermann Völlinger]

Agenda

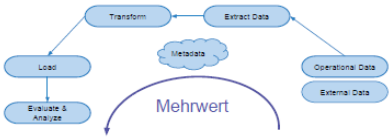
- Motivation
- Überblick
- DWH Trends
- Zusammenfassung

Überblick (I/II)

„A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process."

[Quelle: W.H. Inmon (1996), Seite 33]

Überblick (II/II)



Agenda

- Motivation
- Überblick
- DWH Trends
- Zusammenfassung

Trends (I / III)

1. In-Memory-Datenhaltung
   - Echtzeitanalyse
   - In-DB-Analyse
2. DWH Sicherheit
   - z.B. Banken, Versicherungen, etc.

Trends (II / III)

3. NoSQL DBMS
4. "Datafication" des Unternehmens
   - Mobile Geräte
   - Sensoren (RFID,etc.)
   - Soziale Netzwerke
5. DWH Appliances
   - vorkonfigurierte DWH Lösungen

Trends (III / III)

6. Analytics as a Service
   - dynamische Workloads
   - Anwendung z.B. Rapid Protoyping
7. In-Database Analysen
   - Data-Mining Algorithmen im DBMS
   - Integration von Statistischen Prog.Sprachen (z.B. R)

Agenda

- Motivation
- Überblick
- Trends
- Zusammenfassung

Zusammenfassung

Extraktion / Transformation / Laden von Daten aus verschiedenen Datenquellen mit den Zielen:

- Entscheidungsfindung zu unterstützen
- Datenqualität & -konsistenz sicherzustellen
- Businessprobleme zu lösen
- Einfachen, konsistenten & verständlichen Zugriff auf Daten für alle Beteiligten zu liefern

Business Intelligence

Überblick und Trends

Agenda

- Überblick
- BI Trends
- Zusammenfassung

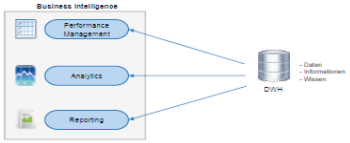
Agenda

- Überblick
- BI Trends
- Zusammenfassung

Agenda

- Überblick
- BI Trends
- Zusammenfassung

Überblick



Trends

Cloud Computing

Visual Data Discovery

Mobile First

Trends

Visual Data Discovery

- Agile Analyse
- Benutzerfreundlich
- Self-Service
- z.B. Tableau

Trends

Cloud Computing

- Marktreife von Cloud
- Infrastrukturelle Vereinfachung
- JasperSoft
  - Amazon-Cloud Host
  - Self-Service BI für Jeden
- Microsoft Power BI, Oracle, ...

Trends

Mobile First

- HTML / Nativ
- Funktionen
- Zugriff / Authorisierung / Sicherheit
- Offline Support
- Nutzer interessiert an Mobile?
- Cognos Mobile

Trends

. . . Weitere Trends, die BI unterstützen

- Informationsgewinn aus unstrukturierten Daten
- Schnellere Analyse von Datenmassen (strukturiert)
- Neue, bessere Arten von Suchen

Agenda

- Überblick
- BI Trends
- Zusammenfassung

Zusammenfassung

Nutzung und Auswertung von verteilten und inhomogenen Daten aus Data Warehouse Anwendungen mit den Zielen:

- Geschäftsabläufe, CRM profitabler machen
- Kosten senken
- Risiken minimieren
- Wertschöpfung vergrößern

→ Erfolgskritisches Wissen über Status, Potentiale und Perspektiven erzeugen

Customer Relationship Management

Überblick und Trends

Agenda

- Was ist CRM?
- Ausblick und Trends
- Zusammenfassung

Agenda

- Was ist CRM?
- Ausblick und Trends
- Zusammenfassung

Was ist CRM?

„Products come and go, but customers remain"
[Rust, Zeithaml, Lemon 2000, S. 6]

Was ist CRM?

→ Ganzheitliche, strategische Ausrichtung aller Geschäftsprozesse am Kunden

→ IT-gestützt

Was ist CRM?

Agenda

- Was ist CRM?
- Ausblick und Trends
- Zusammenfassung

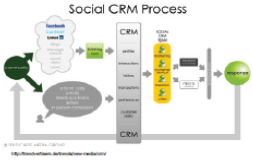
Trendthemen CRM 2014

Social CRM

- Aufbau eigener CRM-Kanäle auf sozialen Plattformen
- Webmonitoring, Data Mining
- Verlagerung der "klassischen" Kommunikation

Social CRM

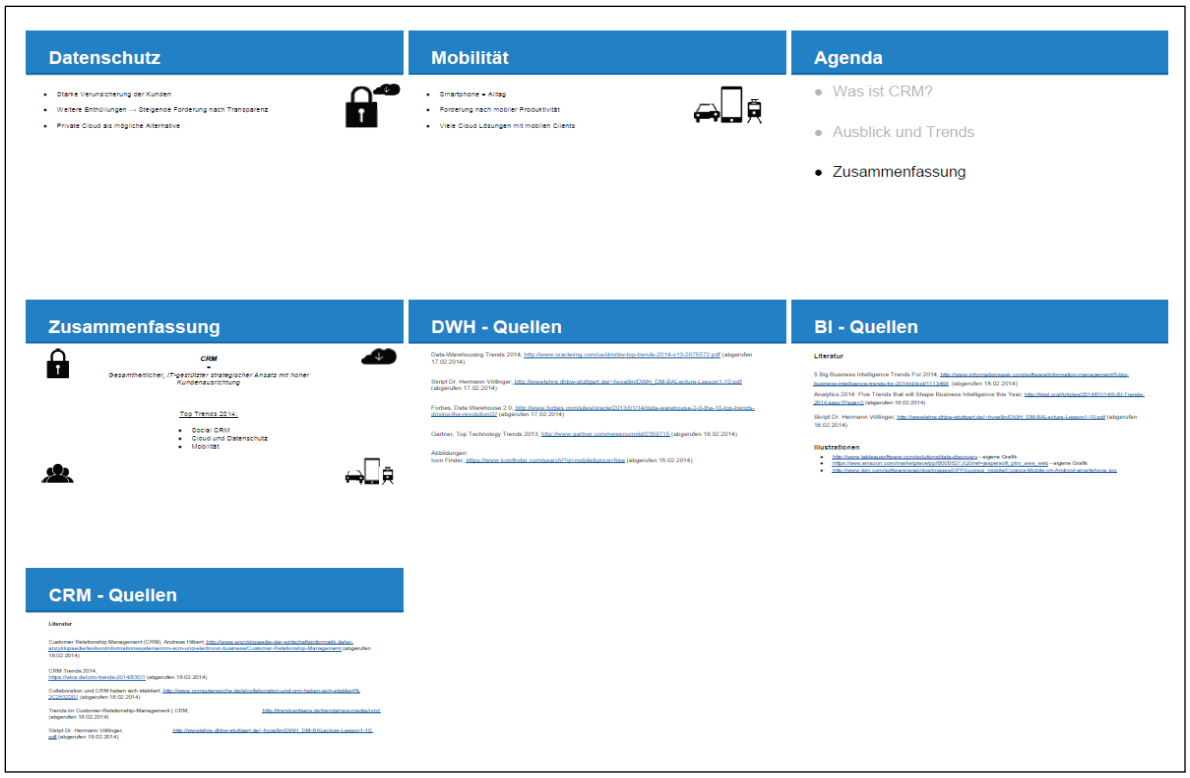
Social CRM Process



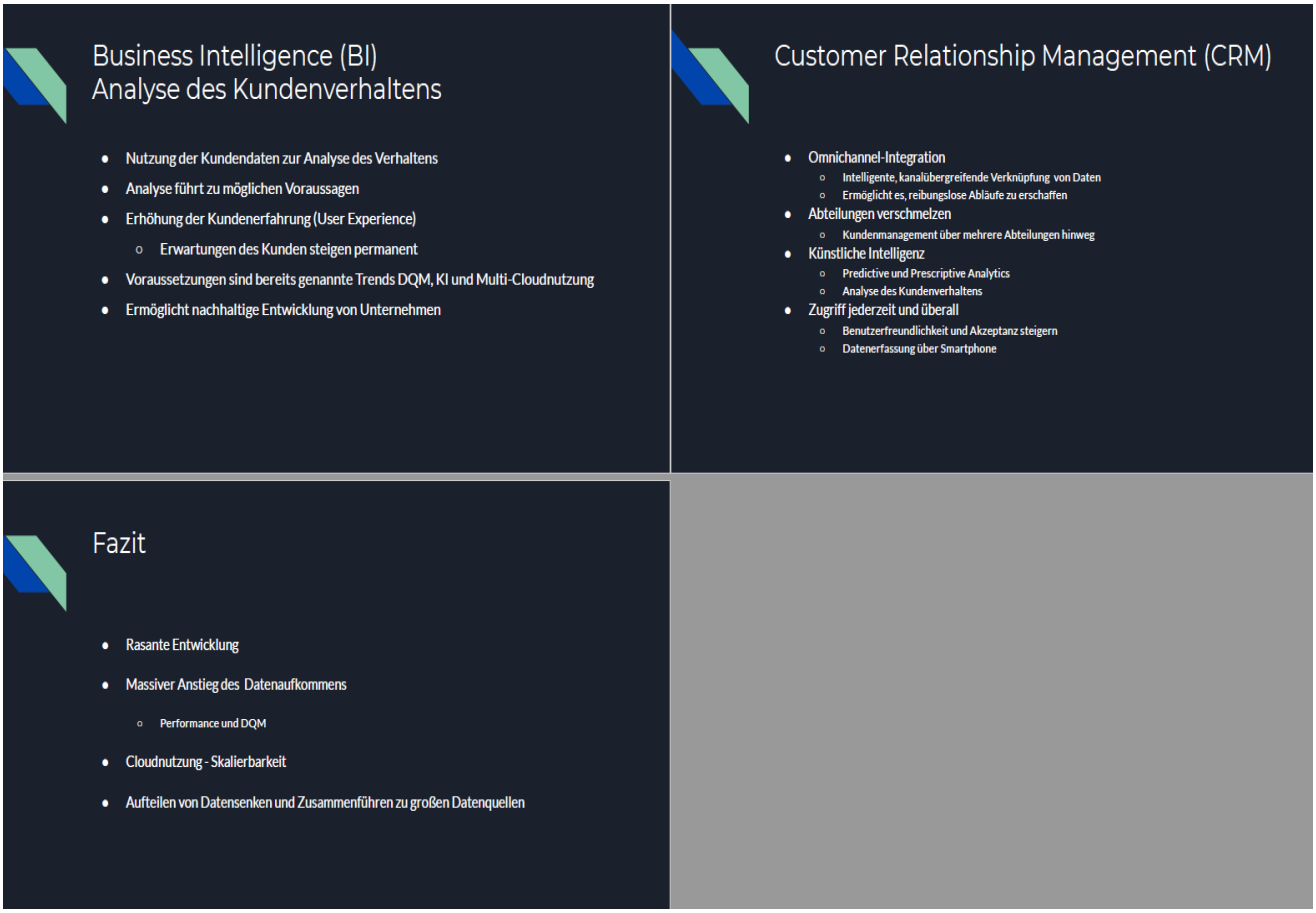
Cloud Computing

- CRM as a Service
- 40% CRM Erlöse in der Cloud

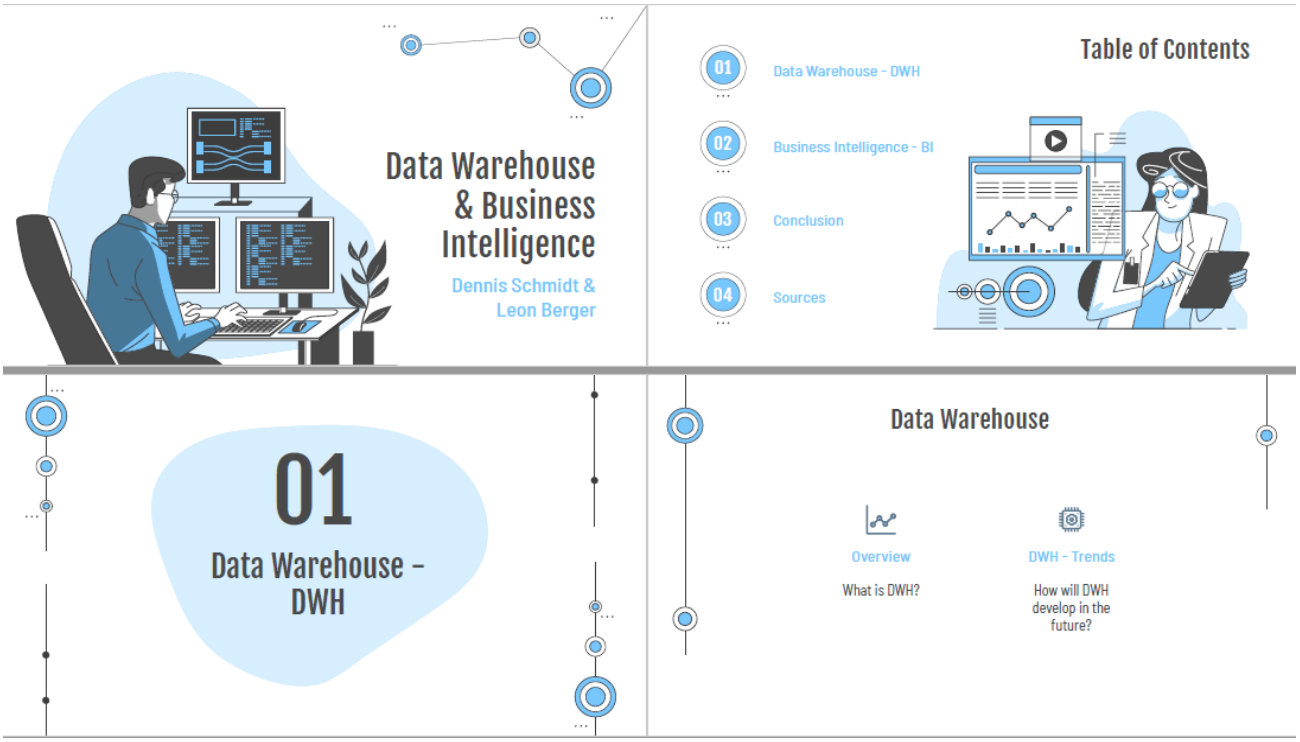
"In kaum einem anderen Bereich der Unternehmensapplikationen hat das SaaS-Modell eine schnellere Verbreitung gefunden als beim Kunden- und Kontakt-Management."



Further Solution (WS 2019):



Further Solution (WS 2021, Leon Berger, Dennis Schmidt):



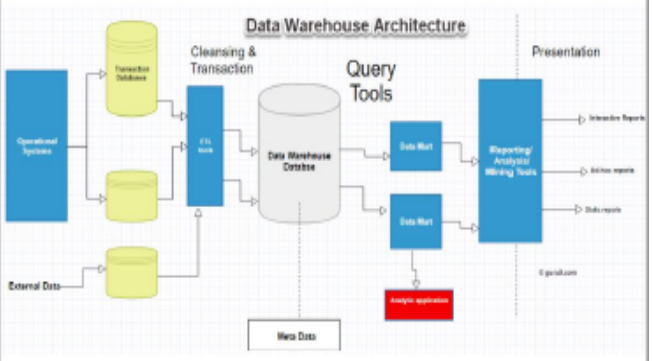
Overview - Structure of DWH

01 Data Acquisition
acquisition and extraction of data from various databases

02 Data Storage
storage of the data in the data warehouse

03 Data Supply
supply of the downstream systems with the required data, provision of data marts

04 Data Evaluation
analyzes and evaluations of the data stocks



-> ETL = Extraction, Transformation and Load

DWH-Trends

DWH - Trends

- DWH-Appliances
- Proof of Concept (POC)
- Data Marts
- Save Data Column by Column
- DWH - Mixed Workloads
- DWH as a Service / Cloud DWH

DWH-Appliances

Compact system from one supplier

- Pre-configured product (Hard- and Software)
- Quick installation
- Good performance is guaranteed
- Easier maintenance and service

Proof of Concept (POC)

Ensure optimal workload in the later data warehouse

- Feasibility study
- Build Prototype
  - As much real data as possible
  - High number of test users

Data Marts

Parts of the workload are moved to individual data storages

- Smaller, function- or department-related data repositories
- The performance of DW applications is optimized

Save Data Column by Column

Faster queries than with a line-by-line storage

- Data is stored in columns and highly compressed
- Providers will also change their pricing models accordingly
  - user-based billing (old)
  - based on the data volume loaded into the database (new)

DWH – Mixed Workloads

Avoidance of high administration costs

- DWH has 6 Workloads (Bulk-Batch-Load, Reporting and OLAP-Analysing, Real-time and Continuous Loads, Data Mining and BI)
- If a DW is delivered with all six workloads the performance must be ensured

DWH as a Service / Cloud DWH

More and more companies are opting for the cloud solution

- Flexibility and scalability
- Easier maintenance and infrastructure
- But: There is still a lack of experience in setting up data models, access regulations for end users and integration strategies

02

Business Intelligence – BI

Business Intelligence

Overview

What is the mission of BI?

BI – Trends

How will BI develop in the future?

BI

Overview

Getting Data from a DWH in the Company and analyse it to get more Data-driven decisions.

→ More knowledge and insight about the business

BI-Trends

BI – Trends

SaaS and Cloud Computing

Artificial Intelligence - AI

Automation

Natural Language Processing - NLP

Data Storytelling & Data Visualization

SaaS & Cloud Computing

More and more companies want their BI solutions in the cloud

- Either self-hosted or as a SaaS (Software as a Service)
- Easier maintenance and infrastructure
- Flexibility and scalability

Example:

- Microsoft Power BI

Artificial Intelligence – AI

AI will be a key part of BI in the future

- Machine learning
- Auto-generated explanation / analytics
- Detect unexpected behaviour

Automation

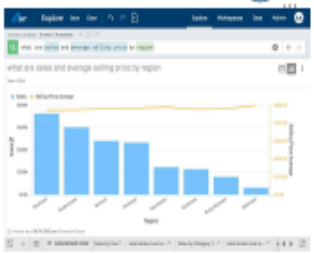
More data = more monotonous work

- Free up resources → less manual data processes
- More time on decision making
- Faster business processes and better workflow

Natural Language Processing – NLP

No need to learn a programming language

- Easier use for non-technical users
- Simplified language → like asking a colleague
- BI tools get more and more popular



Data Storytelling & Data Visualization

The presentation is as important as the data itself

Data Storytelling

- Information into context
- Narrative with dashboards and storyboards

Data Visualization

- Turn Data into graphics and charts
- Easier to understand than text & numbers

03

Conclusion

Conclusion

- ❑ Giant growth in data → more management needed
- ❑ DWH and BI are linked together
- ❑ Both are shifting towards cloud solutions → scalability
- ❑ Future of data storage and data analytics

Thanks!

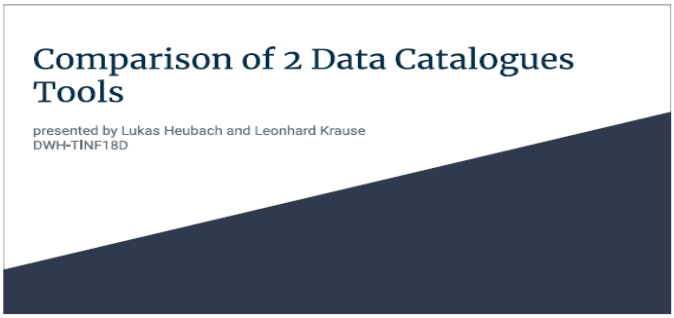
Do you have any questions?

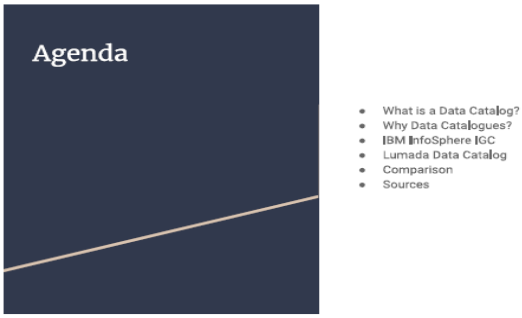
Exercise E1.3*: Compare two Data Catalogue Tools

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).

Hint: 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].

First Solution: Lukas Heubach, Leonhard Krause (WS2020)

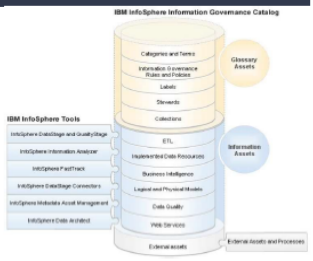




IBM InfoSphere IGC – General

InfoSphere Information Governance Catalog

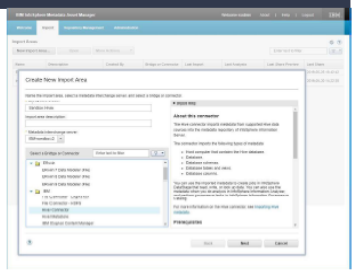
- web-based tool
- Create, manage, share, use business knowledge
- Prices
  - individual offer
- Goals
  - Data -> reliable information
- Can be used in conjunction with other InfoSphere tools



IBM InfoSphere IGC – Functions

Connection of data sources

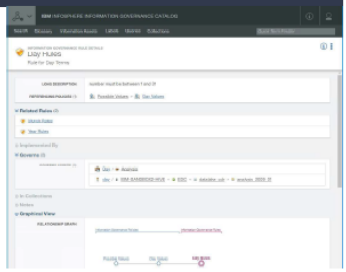
- different types of sources (asset types)
  - (AWS S3, IBM InfoSphere DB2, Oracle ...)
- Import e.g. via Metadata Asset Manager



IBM InfoSphere IGC – Functions

Glossary Assets

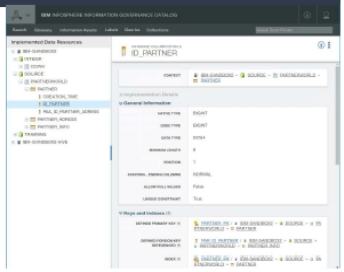
- create/represent complex relationships between assets
- Categories
  - like a folder to structure Glossary Assets
- Terms
  - Word/phrase that describes a characteristic
- IG Rules
  - a natural language description of a criterion that determines whether an information asset meets a business objective.
- IG Policies
  - a natural language description of a subject area
- Labels



IBM InfoSphere IGC – Functions

Information Assets

- Imported records
- Imported metadata
- Display of all included data sources
- Display of all data sets/metadata
- Assign to Glossary Assets



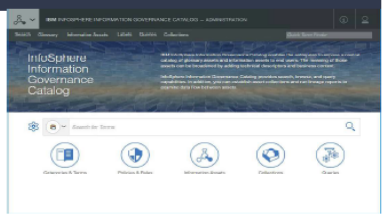
IBM InfoSphere IGC – Functions

Queries

- create your own queries
- for Information Assets
- for Glossary Assets
- result: Table with information

IBM InfoSphere IGC – Who is it for?

- Business Analysts
- Business experts
- Organizations that
  - want to manage a common enterprise vocabulary and governance practices
  - want to leverage the potential of integrated metadata
  - reduce the need for technical training

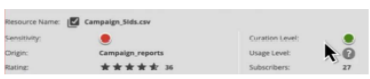


Lumada Data Catalog (Waterline Data Catalog)

- Tool for managing data from diverse sources
- uses machine learning to build data inventory
- patented fingerprinting technology for data
- Management of data lakes
- Goal: analyze large amounts of data automatically
- Price and demo are only available on request

Lumada Data Catalog – Functions

- Data recognition using metadata-based search
- Management of sensitive data
- Patented data fingerprinting
  - data processing based on machine learning
  - automatic recognition of sensitive data
  - enables Google-like search with corporate identifiers
- visualization of data, -origins and -relationships
- Detection of redundant data
- comment, subscription, rating function
- Recording of user data



Lumada Data Catalog – Who is it for?

- Companies with large amounts of data
- Data analysts
- Fast and efficient compliance with data protection regulations
- Management of sensitive data

Comparison

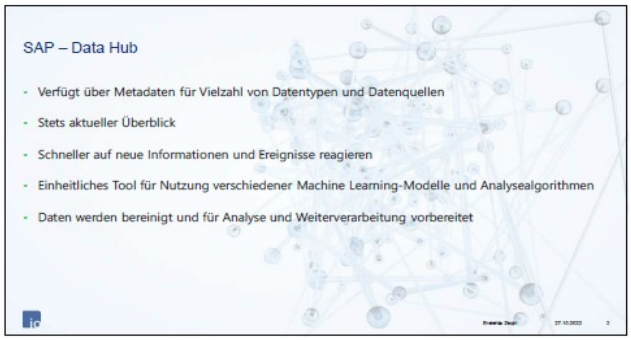
Category	IBM InfoSphere IGC	Lumada Data Catalogue
Similarities	• Data Catalog Tools • individual price offer • use of different data sources	
User Interface	Web Tool	no specification (possibly desktop application)
Focus	data management	data analysis
AI	No	AI-based data analysis and structuring
Product Portfolio	IBM InfoSphere Family	Lumada Data Services
"Social-Media" Tools	No	Yes (comments, subscription, rating)

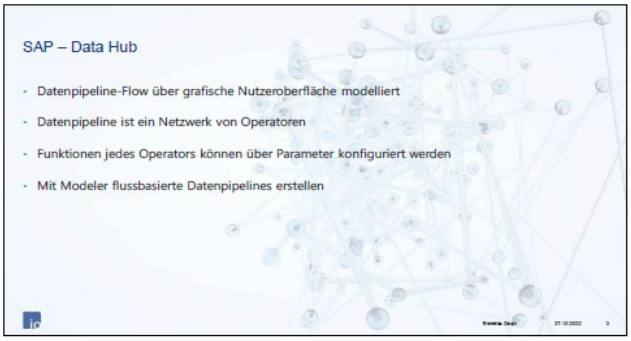
Sources

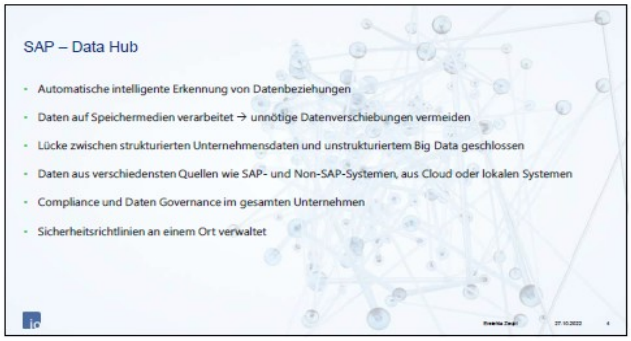
- https://www.hitachivantara.com/de-de/products/data-management-analytics/lumada-data-services/lumada-data-catalog.html (retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.hitachivantara.com/en-us/products/data-management-analytics/lumada-data-catalog.html (retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.hitachivantara.com/en-us/pdfd/datasheet/lumada-data-lake-datasheet.pdf (abgerufen am: 13.02.2021 at 14:32 Uhr)
- https://www.hitachivantara.com/de-de/pdfd/datasheet/lumada-data-catalog-datasheet-de.pdf (retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.talend.com/de/resources/what-is-data-catalog/(retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.cc-cdq.ch/data-catalogs (retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.ibm.com/de-de/marketplace/information-governance-catalog (retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.saracus.com/blog/der-ibm-infosphere-information-governance-catalog/ (retrieved on: 13.02.2021 at 14:32 Uhr)

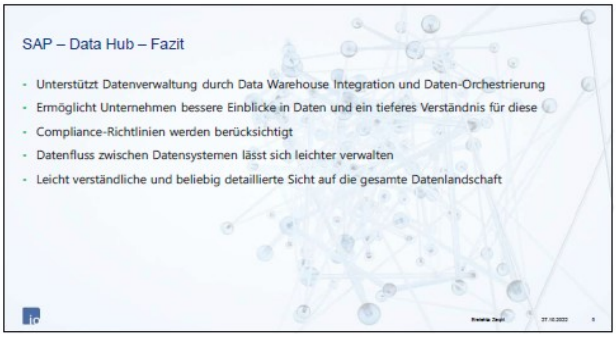
THANK YOU FOR YOUR ATTENTION!

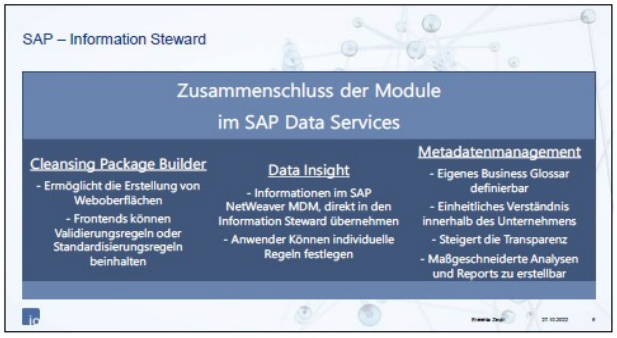
Second Solution: Erelehta Zeqiri (WS2022)











Exercise 1.4: First Experiences with KNIME Analytics Platform

Task: Install the tool and report about your first experiences. 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).

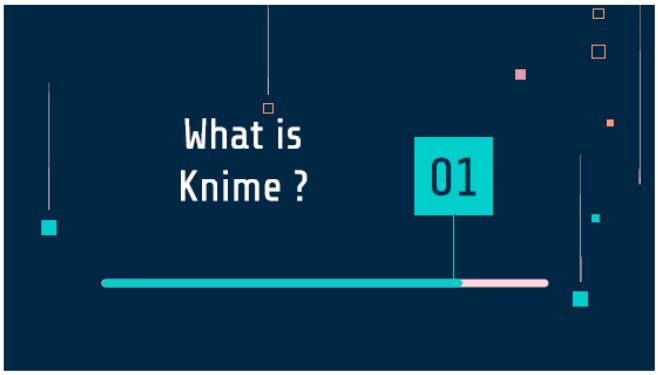
Hint: The installation of KMIME is described in the "KNIME-BeginnersGuide.pdf". The document can be found in the first category of the "Supporting Information for DWH Lecture" in the Course-Moodle: Kurs: T3INF4304_3_Data Warehouse (dhbw-stuttgart.de)

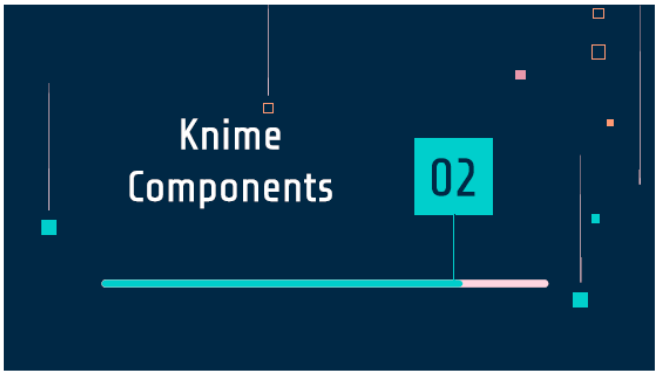
Download KNIME Analytics Platform

- 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

Solution: Copyright: Creative Commons (CC) license by Felix Grohme, Finn Markwitz (WS2021) - "Bridge the gap between data science and business":





Knime Features:

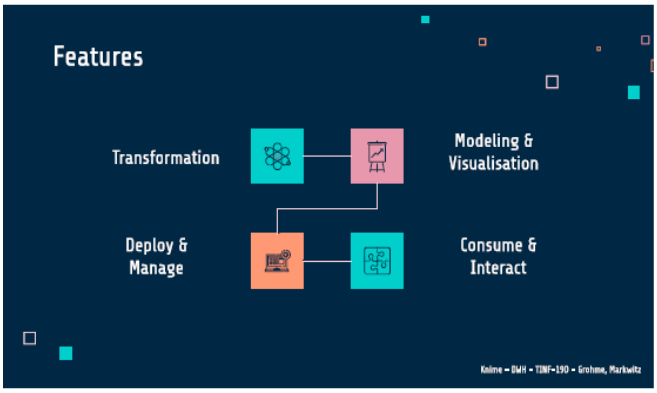
Blend & Transform:
- Access data from different sources (e.g Databases, Files, etc.)
- Merging of data from different data sources (adapting data if necessary)
- Prefabricated interfaces for various DBs and DWHs
- Interfaces are extensible
- Documentation of executed steps for better traceability

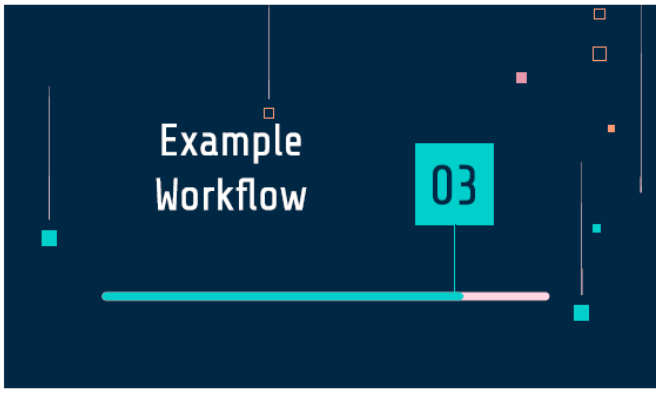
Model & Visualize:
- Allows to combine data with context -> different visualization possibilities
- Apply different tools via Knime -> Tensorflow, H2O, R and Python
- Create high quality data models -> data is accessible and easy to find -> visual documentation through framework

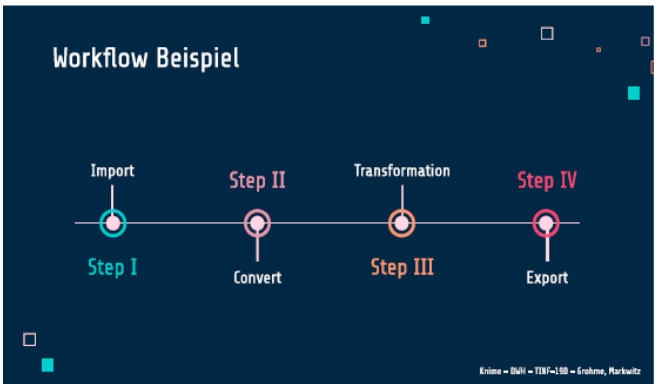
Deploy & Manage:
- Create interfaces to make data available in other systems
- Integrated rights system -> who is allowed to access which data
- Persons with little knowledge can map processes thanks to workflow editor

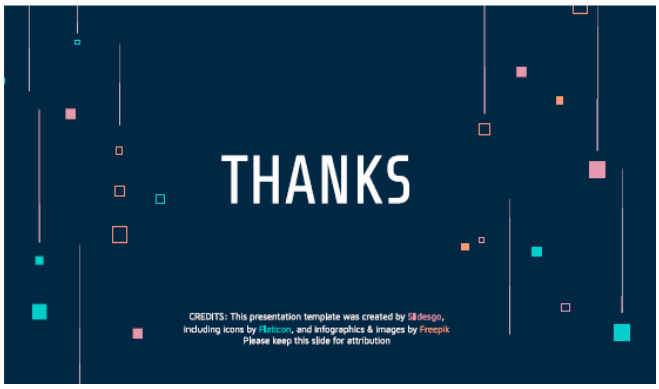
Consume & Interact:

- Allows the easy creation of reports (diagrams, Excel)
- Security for sensitive data in the form of encryption, versioning, logging, etc.





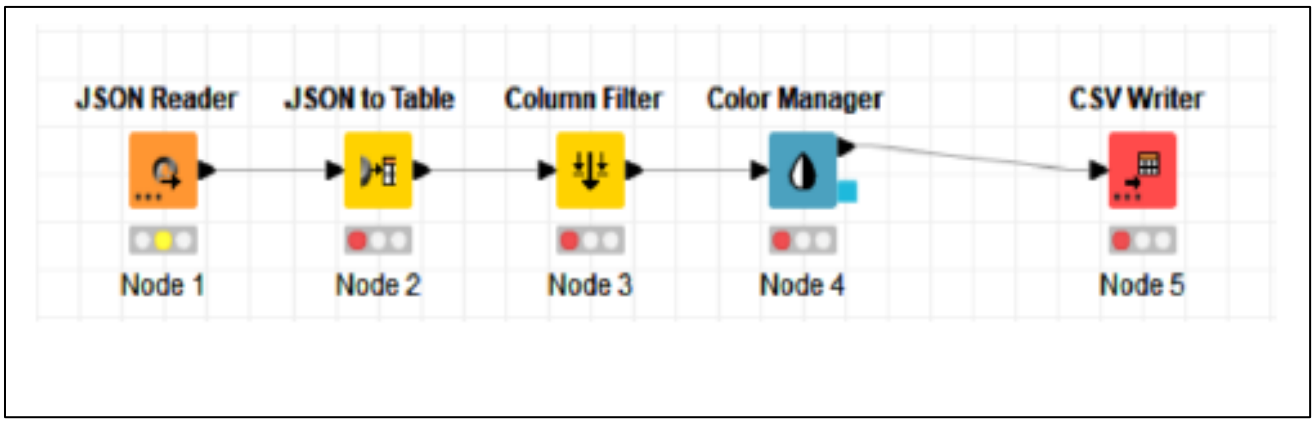




How does Knime work:

Knime uses so-called workflows to create a process. This allows people who do not have the required skills in data science but have expertise of the economic process, or vice versa, to easily create a data-workflow which can be implemented within a production environment.

An easy to understand example is shown in the following text:



First, we import data using a JSON-Reader Node, since KNIME holds the processed data of each node in the context of the node, this is where the now imported dataset is present.
This allows the user to view each step of the workflow and recap which node transforms the data in which way. After importing the JSON data we're telling the import node to only represent the data matching a given JSON-Path.

This can be achieved via the "dotwalk"-annotation within the configuration of the JSON-Reader node. The JSON-Reader node is able to automatically convert the JSON-Array into a table using one row for each monster.
When passing the table containing the objects into a JSON to Table node. This node takes the properties of the JSON-objects in the rows and maps them to new columns.
Taking a look at the generated table, we can see a good overview of the monsters. Each property is now sorted into a new column. If we take a look at the resulting table we see a lot of columns with no values. This can happen since KNIME maps the JSON-object with all the values for all the objects. Since we only want the important properties, we sort out the important columns using a Column Filter. This node allows us to remove or even merge, certain columns from the table.
The transformation results in a table containing only the wanted columns.
Let's say the use case ends here and our company wants to use the now corrected dataset within a third-party-software, we could for example export it into a CSV-File to make it available for further usage.

Exercises (+Solutions) to DHBW Lecture Intro2DWH – Chapter 2

Exercise E2.1*: Compare 3 DWH Architectures

Task: 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					
....					

Solution:

Datawarehousing 5. Semester, IT00PMB, Marcel Petz
Projekt: Comparison of different DWH Architectures
Dokument: Comparison Matrix, 25.10.02
Seite: 1/ 1

Criteria	Datawarehouse only	Datamarts only	Datawarehouse & Datamarts	"Virtual" Datawarehouse	Description
Costs of Implementation	+	+	++	-	Costs of implementation describes the complexity of implementing a datawarehouse and its components. Complexity depends on general architecture (e.g. networked environment, database based, amount of hardware needed,...) of the datawarehouse.
Costs of administration	+	-	0	(--)	Costs of administration describes the costs for maintaining and run the datawarehouse.
Average data age	0	0	-	++	How old is the information presented to the frontend users of a datawarehouse
Performance		+	+	--	The Performance of the whole system and all involved components. Which architecture has a good performance.
Flexibility	-	0	+	++	Flexibility describes the ability of changing datastructures or parts of the datawarehouse programs. High flexibility means, that it is easy to make changes to the data structures and client applications.
Implementation-time	-	-	--	+	Describes the time-to-market. The time from begin of implementation until the system is activated for production use. Positive values mean short implementation time.
Data Consistency	+	+	+	--	
Quality of informations					Quality of data stored in the datawarehouse depend strongly on the quality of the ETL process. It is not possible to draw general conclusions at this point
History	++	++	++	--	The ability to look at certain points or periods of time in the past und gather information about it.

Implementation costs

The implementation of a Data Warehouse with Data Marts is the most expensive solution, because it is necessary to build the system including connections between Data Warehouse and its Data Marts.
It is also necessary to build a second ETL which manages the preparation of data for the Data Marts.
In case of implementing Data Marts or a Data Warehouse only, the ETL is only implemented once. The costs may be almost the same in building one of these systems. The Data Marts only require a little more hardware and network connections to the data sources. But due to the fact, that building the ETL is the most expensive part, these costs may be relatively low.
The virtual Data Warehouse may have the lowest implementation costs, because e.g. existing applications and infrastructure is used.

Administration costs

The **Data Warehouse only solution** offers the best effort in minimizing the administration costs, due to the centralized design of the system. In this solution it is only necessary to manage a central system. Normally the client management is no problem, if using web technology or a centralized
client deployment, which should be a standard in all mid-size to big enterprises. A central Backup can cover the whole data of the Data Warehouse.
The solution with **Data Marts only** are more expensive, because of its decentralized design. There are higher costs in cases of product updates or maintaining the online connections, you also have to backup each Data Mart for itself, depending on his physical location.
Also the process of filling a single Data Mart is critical. Errors during update may cause loss of data. In case of an error during an update, the system administration must react at once.
Data Marts with a central Data Warehouse are more efficient, because all necessary data is stored in a single place. When an error during an update of a Data Mart occurs, this is normally no problem, because the data is not lost and can be recovered directly from the Data Warehouse. It may also be possible to recover a whole Data Mart out of the Data Warehouse.
Virtual Data Warehouses administration costs depend on the quality of the implementation. Problems with connections to the online data sources may cause user to ask for support, even if the problem was caused by a broken online connection or a failure in the online data source. End-users may not be able to realize whether the data source or the application on their computer cause a problem.

Average data age

The virtual Data Warehouse represents the most actual data, because the application directly connects to the data sources and fetches its information online. The retrieved information is always up to date.
Information provided by Data Mart only or Data Warehouse only solutions are collected to specific time. Generally, each day by night. These times can vary from hourly to monthly or even longer. The selected period depends on the cost of the process retrieving and checking the information.
A solution with one central Data Warehouse and additional Data Marts houses less actual data then Data Warehouse only. The data of the Data Warehouse must be converted and copied to the Data Marts, which is time consuming.

Performance

A virtual Data Warehouse has the poorest performance all over. All data is retrieved during runtime directly from the data sources. Before data can be used, it must be converted for presentation. Therefore, a huge amount of time is spent by retrieval and converting of data.
The Data Marts host information, which are already optimized for the client applications. All data s stored in an optimal state in the database. Special indexes in the databases speed up information retrieval.

Implementation Time
The implementation of a Data Warehouse with its Data Marts takes the longest time, because complex networks and transformations must be created. Creating Data Warehouse only or Data Marts only should take almost the same amount of time. Most time is normally spent on creating the ETL (about 80%), so the differences between Data Warehouse only and Data Marts only should not differ much.
Implementing a Virtual Data Warehouse can be done very fast because of its simple structure. It is not necessary to build a central database with all connectors.

Data Consistency
When using Data Warehouse or Data Mart technology a maximum consistency of data is achieved.
All provided information is checked for validity and consistency. A virtual Data Warehouse may have problems with data consistency because all data is retrieved at runtime. When data organization on sources changes, the consistency of new data may be consistent, but older data may not be represented in its current model.

Flexibility
The highest flexibility has a virtual data warehouse. It is possible to change the data preparation process very easy because only the clients are directly involved. There are nearly no components, which depend on each other.
In Data Warehouse only solution flexibility is poor, because there may exist different types of clients that depend on the data model of the Data Warehouse. If it would be necessary to change a particular part of the data model intensive testing for compatibility with existing applications must be done, or even the client applications have to be updated.
A solution with Data Marts, with or without a central Data Warehouse has medium flexibility due that client applications normally uses Data Marts as their point of information. In case of a change in the central Data Warehouse or the data sources, it is only necessary to update the process of filling the Data Marts.
In case of change in the Data Marts only the depending, client applications are involved and not all client applications.

Data Consistency
Data consistency is poor in a virtual Data Warehouse. But it also depends on the quality of the process, which gathers information from the sources.
Data Warehouses and Data Marts have very good data consistency because the information stored in their databases have been checked during the ETL process.

Quality of information
The quality of information hardly depends on the quality of the data population process (ETL process) and how good the information is processed and filtered before stored in the Data Warehouse or presented to a user. Therefore, it is not possible to give a concrete statement.

History
A virtual Data Warehouse has no history at all, because the values or information are retrieved at runtime. In this architecture it is not possible to store a history because no central database is present.
The other architectures provide a central point to store this information. The history provides a basis for analysing business process and their efforts, because it is possible to compare actual information with information of the past.

Second Solution (SS2021):

COMPARE 3 DWH ARCHITECTURES

LUKAS SCHULT, LUANA JUHL

COMPARISON TABLE

	DW Only	DM Only	DW & DM	Virtual DW	Explanation
Administration expenses	+	-	0	-	Run & Manage/Maintain cost
Implementation	-	+	--	++	Cost and time needed to implement the Architecture including all components and necessary hardware
Performance	0	+	+	+	Speed at which Data can be accessed
Size	>100 GB	<100 GB			

COMPARISON TABLE

	DW Only	DM Only	DW & DM	Virtual DW	Explanation
Flexibility	-	+	0	++	Adaptability in response to changes
History	+	+	+	--	Storage of historical data used to determine data trends
Security	-	+	0	+	Management of access to data stored in the data model
Data Quality & Consistency	+	-	+	-	

IMPLEMENTATION

DW Only	DM Only	DW & DM	Virtual DW
• Big amounts of data ➢ Large amounts of resources necessary ➢ Implementation process more complex • Months – Years • High cost	• Less data ➢ Implementation more simple • Days - Months • Cost efficient	• DW & DM need to be implemented + Connections • Long implementation time • Very high cost	• Simple structure ➢ Small complexity • Little implementation time • Low cost

FLEXIBILITY

DW Only	DM Only	DW & DM	Virtual DW
• Defined by various domains ➢ Adapting to changes more difficult	• Defined by single subject matter • Small data model ➢ Changes easy & quick	←	• Defined for varying formats and structures ➢ Changes easy & quick

DATA QUALITY & CONSISTENCY

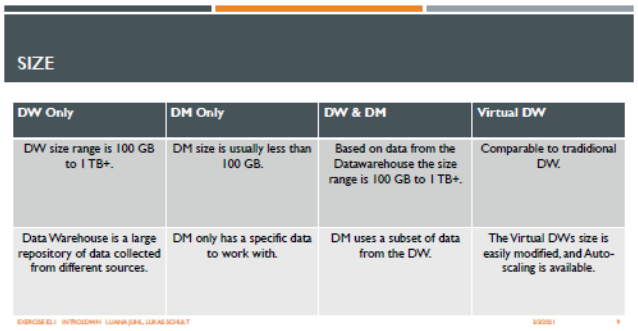
DW Only	DM Only	DW & DM	Virtual DW
• High • Data conversion into common format ➢ No discrepancies • "Single source of truth"	• Redundant Data in various DMs	• Uniform data format	• Data quality logic manually created

SECURITY

DW Only	DM Only	DW & DM	Virtual DW
• Central repository ➢ Data access not limited	• Separate repositories ➢ Data access limited	• ←	• Security permissions defined in meta data ➢ Data access controlled

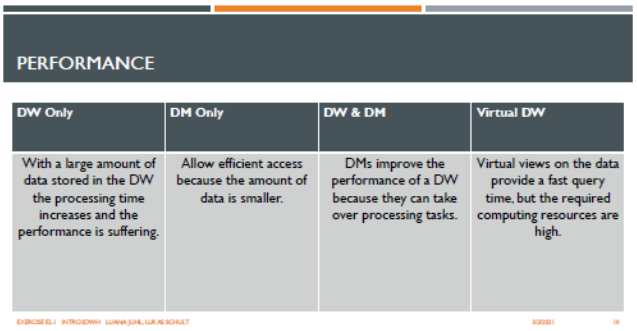
ADMINISTRATION EXPENSES

DW Only	DM Only	DW & DM	Virtual DW
It uses a centralized system design which simplifies Management and Backup capabilities.	Decentralized processes like backup, updates have to be done for each DM. In combination with a high error risk, this results in higher costs than DW.	Data is centralized, which allows safer management of DMs. When used together, the costs level out with savings from DW and more expenses from DMs.	Requires Views to the underlying Databases to be managed. Stacking Views on Views can also require extensive computing resources. These factors make the architecture very expensive.



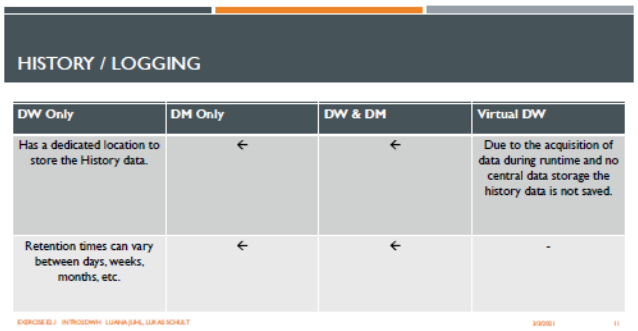
SIZE

DW Only	DM Only	DW & DM	Virtual DW
DW size range is 100 GB to 1 TB+.	DM size is usually less than 100 GB.	Based on data from the Datawarehouse the size range is 100 GB to 1 TB+.	Comparable to tradidional DW.
Data Warehouse is a large repository of data collected from different sources.	DM only has a specific data to work with.	DM uses a subset of data from the DW.	The Virtual DWs size is easily modified, and Auto-scaling is available.



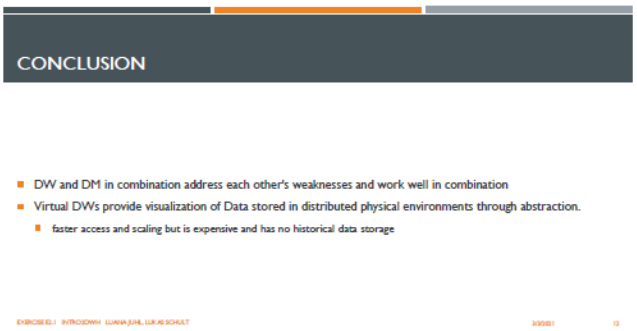
PERFORMANCE

DW Only	DM Only	DW & DM	Virtual DW
With a large amount of data stored in the DW the processing time increases and the performance is suffering.	Allow efficient access because the amount of data is smaller.	DMs improve the performance of a DW because they can take over processing tasks.	Virtual views on the data provide a fast query time, but the required computing resources are high.



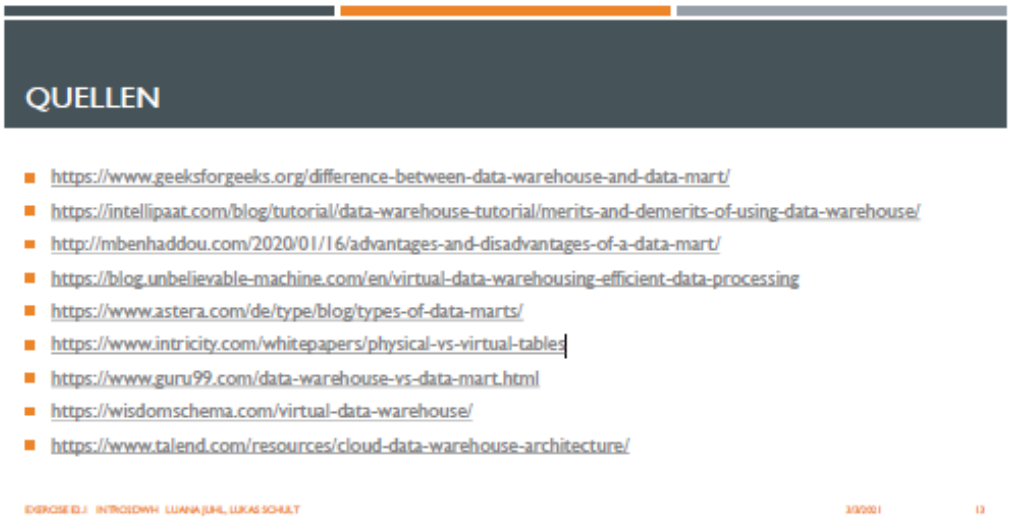
HISTORY / LOGGING

DW Only	DM Only	DW & DM	Virtual DW
Has a dedicated location to store the History data.	←	←	Due to the acquisition of data during runtime and no central data storage the history data is not saved.
Retention times can vary between days, weeks, months, etc.	←	←	-



CONCLUSION

- DW and DM in combination address each other's weaknesses and work well in combination
- Virtual DWs provide visualization of Data stored in distributed physical environments through abstraction.
  - faster access and scaling but is expensive and has no historical data storage



QUELLEN

- https://www.geeksforgeeks.org/difference-between-data-warehouse-and-data-mart/
- https://intellipaat.com/blog/tutorial/data-warehouse-tutorial/merits-and-demerits-of-using-data-warehouse/
- http://mbenhaddou.com/2020/01/16/advantages-and-disadvantages-of-a-data-mart/
- https://blog.unbelievable-machine.com/en/virtual-data-warehousing-efficient-data-processing
- https://www.astera.com/de/type/blog/types-of-data-marts/
- https://www.intricity.com/whitepapers/physical-vs-virtual-tables
- https://www.guru99.com/data-warehouse-vs-data-mart.html
- https://wisdomschema.com/virtual-data-warehouse/
- https://www.talend.com/resources/cloud-data-warehouse-architecture/

Third Solution (WS2021):


Data Warehouse
Architectures
Max Petershans, Fabian Regele

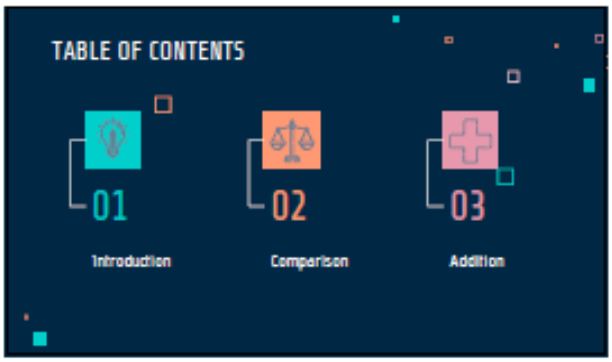
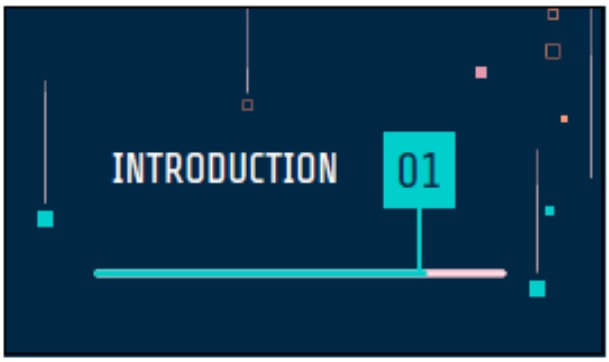
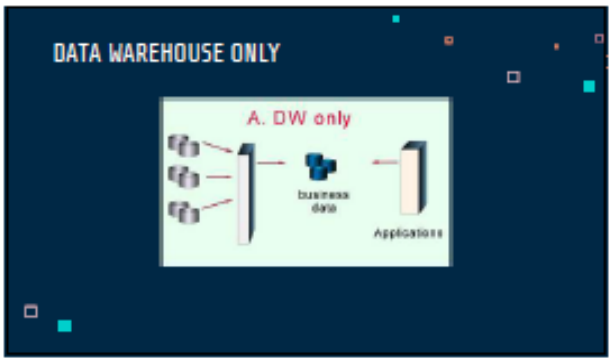
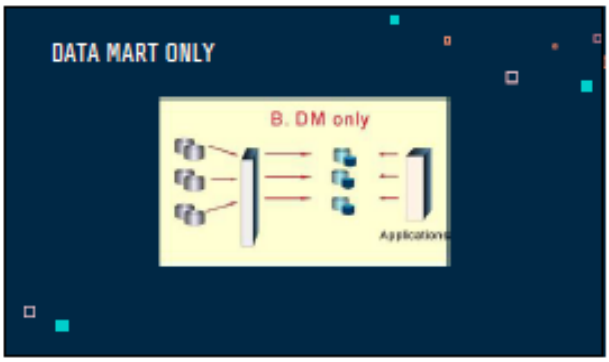
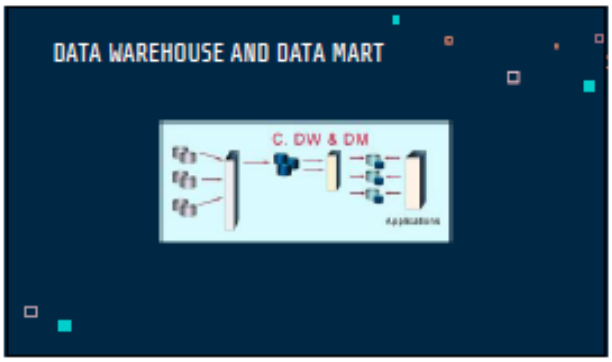


TABLE OF CONTENTS
01
Introduction
02
Comparison
03
Addition


INTRODUCTION
01


DATA WAREHOUSE ONLY
A. DW only
business data
Applications


DATA MART ONLY
B. DM only
Applications


DATA WAREHOUSE AND DATA MART
C. DW & DM

COMPARISON 02

COMPARISON

	DW Only	DM Only	DW & DM	Explanation
Cost of Implementation	-	-	–	Cost and time to implement the architecture
Cost of Administration	+	-	0	Cost of maintaining and run the warehouse
Performance	0	+	+	The performance of the architecture

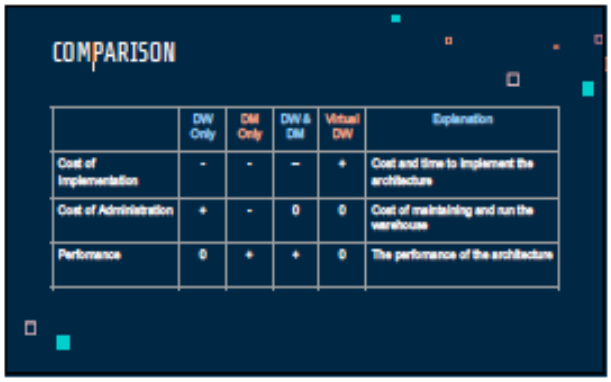
COMPARISON

	DW Only	DM Only	DW & DM	Explanation
Flexibility	-	+	0	Changing datastructure
Quality	+	0	+	Quality or Redundancy of data
Security	-	+	0	Access control to data

Addition 03

VIRTUAL DATA WAREHOUSE

VIRTUAL DATA WAREHOUSE



Exercise E2.2*: Basel II and RFID

Task: *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*

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

1. Basel II
2. RFID

Look also for examples of current projects in Germany

Solution:

IBM

Basel II

Michael Illiger, Stefan Tietz, Steve Gebhardt, Thomas Dürr

Agenda

- **Warum Basel-Abkommen?**
- **Überblick Basel I + II**
- **Basel II Roadmap**
- **Basel II und Data Warehousing**
- **Tools**
- **Ausblick**

Warum Basel-Abkommen?

- **Risiko: Kreditausfall**
- **Geringe Eigenkapitalquote**
- **Keine einheitlichen Rating-Richtlinien**
- **→ Basel I (1988)**

Basel I

- **8% der Kreditsumme durch Eigenkapital abdecken**
- **Kunden-Rating anhand interner Prüfungen**
- **Grundlage: Bilanzen + bisherige Kreditwürdigkeit**
- **→ falsche Anreizsetzung, unabgedeckte Risiken**

Kirch-Krise, Bankenkrise in Japan

→ Basel II

Basel II

- **Kundenrating intern und extern**
- **Reservebildung je nach Kreditrisiko**
- **Aufteilung in Qualitative und Quantitative Risikofaktoren**
- **erweiterte Offenlegung der Finanzsituation in Banken**
- ***Eigenkapital* = *Kreditsumme* x *Risikogewicht* x *Kapitalquote***

Basel II - Säulenmodell

Quelle: KfW

Basel II Roadmap

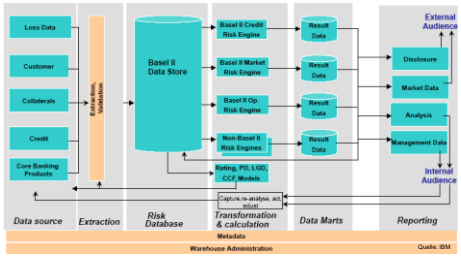
- **2003:** **Aufnahme von Basel II in die Strategie der Institute**
- **2004:** **Aufbau der DWH-Infrastruktur**
- **2005:** **Datensammlung + Auswertungsstrategie**
- **2006:** **... Parallel-Lauf von Basel I + II**
- **2007:** **Basel II wird bindend**

Basel II und Data Warehousing

- **grosse Datenmengen zur Analyse**
- **DWH werden benötigt von:**
  - Banken → Kunden-Rating
  - Rating-Agenturen → Service zur Verfügung stellen
  - Unternehmen → optimale Finanzsituation verringert Kreditkosten

Tools

- **IBM-Lösung → Banking Data Warehouse (BDW)**



Ausblick

- **Vorbereitungen laufen seit 2003**
- **Start: 01.01.2007**
- **verändert Verhältnis zwischen Kapitalgeber und Kapitalnehmer**
- **umfangreicher Absatzmarkt für DWH-Services entsteht (weltweit!)**

Eine weitere Lösung zu Basel2 und DWH ist wie folgt:

Basel II & DWH

Christian Schäfer, 28.10.2005

Agenda

- Sicherung der Stabilität im Finanzsektor
- Eigenkapitalvereinbarung von 1988 (Basel I)
- Von Basel I zu Basel II
- Gründe für Basel II
- Rating von Krediten nach Basel II
- Die Bank will uns kennen lernen
- Auswirkungen von Basel II
- Herausforderungen an Data Warehouse Systeme

Sicherung der Stabilität im Finanzsektor

- Umgang mit Kredit-, Markt, Liquiditäts- und anderen Risiken ist Aufgabe und Geschäftszweck von Kreditinstituten/Banken

Probleme:

- Der freie Umgang mit Risiken darf nicht zu Instabilitäten im Finanzsektor führen
- Freier Umgang bei der Sicherung von Risiken führt zu unterschiedlichen Wettbewerbsbedingungen im Bankwesen

Sicherung der Stabilität im Finanzsektor

Lösung:

- Sicherung einer angemessenen Eigenkapitalausstattung der Banken
- Schaffung einheitlicher internationaler Wettbewerbsbedingungen

Eigenkapitalvereinbarung von 1988 (Basel I)

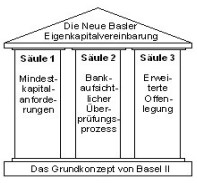
- Richtlinie zur Förderung und Sicherung eines funktionierenden Bankwesens.

Wichtigstes Merkmal:

- Eigenkapitalunterlegung seitens der Bank muss mindestens 8% der Kreditsumme betragen

Von Basel I zu Basel II

- Kritik an Basel I
  - Risiken der Kreditvergabe werden unzureichend abgebildet
  - Neue Finanzierungsmöglichkeiten werden nicht berücksichtigt
- Verbesserungen durch Basel II
  - Internes oder externes Rating von Kreditrisiken
  - Marktrisiko (Branche etc.)
  - Operationelles Risiko (Ausfall Von Mitarbeiter, Systeme etc.)
  - Bankaufsichtlicher Überprüfungsprozess
  - Erweiterte Offenlegung (Selbstkontrolle des Marktes)



Rating von Krediten nach Basel II

- Bewertung von Kreditnehmern durch unabhängige Ratingagenturen wie z.B. Standard & Poor's (externes Rating)
- Bewertung von Kreditnehmern durch den Kreditgeber selbst (auch IRB-Ansatz genannt). Genehmigungspflichtig durch Bankenaufsicht
  - Harte Faktoren
    - Branchenabhängiges Risiko
    - Geschäftsprofil und Wettbewerbsposition
    - Ertragskraft
    - Kapitalanlagen- und Ausstattung
    - Liquidität
  - Weiche Faktoren
    - Unternehmensstrategie
    - Personalpolitik
    - Managementqualitäten
    - Organisationsstrukturen

Erhebung großer Datenmengen

Die Bank will uns kennen lernen

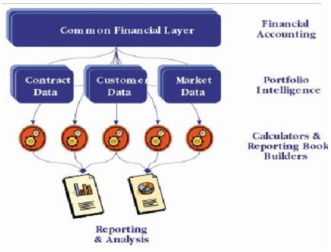
- Datenerhebung zur Berechnung des erwarteten Kreditausfalls (Expected Loss)
  - Adresse
  - Transaktionen
  - Laufende Kredite
  - Liquidität
  - Eigenkapital
  - ...
  - -> Differenzierte Kreditbewertung
  - -> Berechnung der Eigenkapitalquote

Auswirkungen von Basel II

Herausforderungen für Data Warehouse Systeme

- Daten der Kunden als Basis für Analysen und Statistiken
- DWH als Möglichkeit zur Steuerung der Qualität des Kreditportfolios
- Transaktionen, Kontostände, laufende Kredite zur schnellen Bewertung von Risiken
- Erkennen von externen Einflussfaktoren (Marktgeschehen, etc.)
- Daten müssen Konsistent sein und mindestens eine Historie von 3 Jahren aufweisen

DWH Konzept



Roadmap Basel II

Monat	Jahr	Ereignis
Juli	1988	Veröffentlichung der Baseler Eigenkapitalvereinbarung (Basel I)
Ende	1992	Inkrafttreten von Basel I
Januar	1996	Baseler Marktrisikopapier
Juni	1999	Erstes Konsultationspapier zur Neufassung der Eigenkapitalvereinbarung (Basel II)
Januar	2001	Zweites Konsultationspapier zu Basel II
Dezember	2001	Änderung des ursprünglich vorgesehenen Zeitplans für die Fertigstellung des neuen Akkords.
Mai	2003	Drittes Konsultationspapier zu Basel II
Mitte	2004	Veröffentlichung der neuen Eigenkapitalvereinbarung
Ende	2006	Inkrafttreten von Basel II

Quellen

- http://de.wikipedia.org/wiki/Basel_II
- http://www.bundesbank.de/bankenaufsicht
- http://www.foerderland.de/353.0.html
- http://www.basel-ii.info/
- http://www.hp.com/

<u>Eine weitere Lösung (dritte Lsg.) zu Basel2 und DWH finden Sie in der folgenden Darstellung:</u>

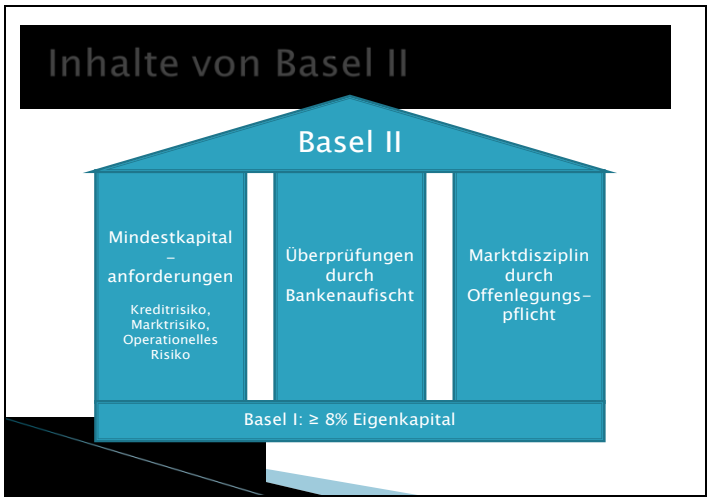
Basel I: Kreditvergabepraxis limitiert durch Verknüpfung mit Eigenkapital
Vergabe von Krediten an Kunde mit mäßiger Bonität -> höhere Zinssätze

1974: Zusammenbruch Herrstatt-Bank

- Devisenspekulationen

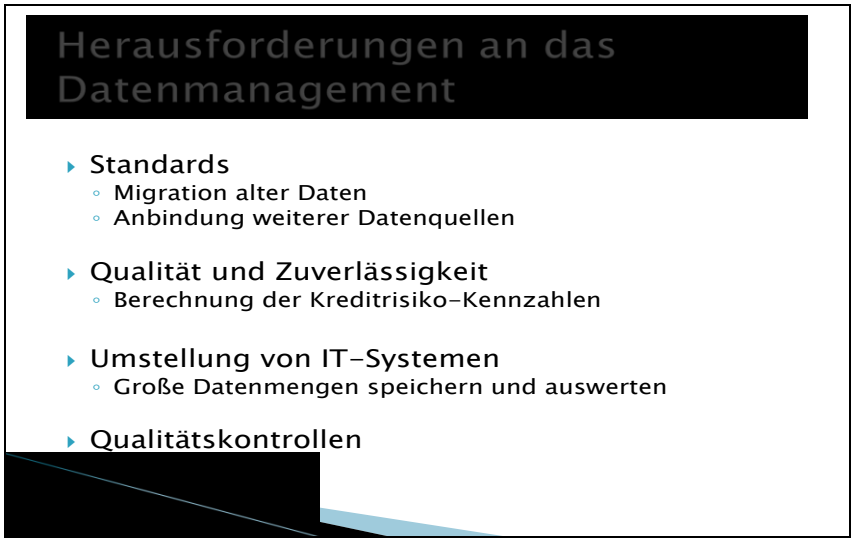
1988: Eigenkapitalvereinbarung „Basel I“

- Kreditvergabepraxis



Basel II: nur Mindestkapital basierend auf Kredit- und Marktrisiken
Marktdisziplin: Verhalten, Öffentlichkeit über Kapital & Risiko zu informieren -> günstige Bedingung bei Beschaffung Fremdkapitals

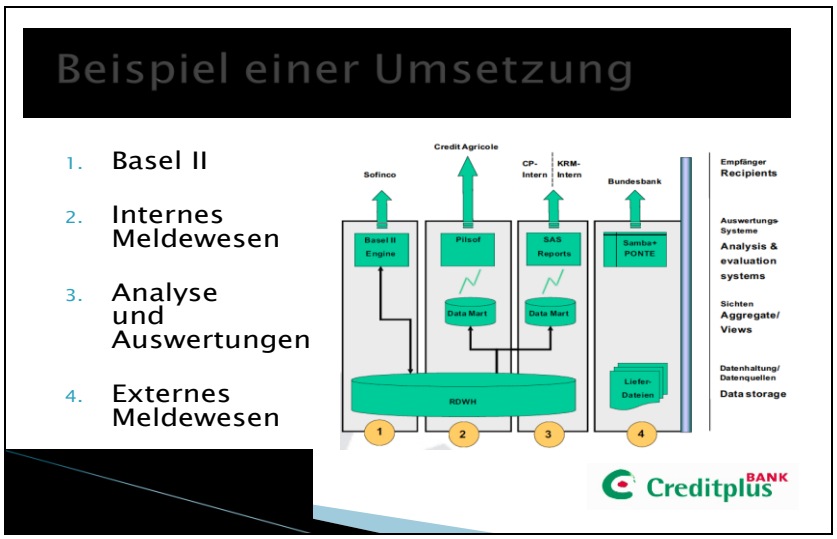
http://www.bundesbank.de/bankenaufsicht/bankenaufsicht_basel.php



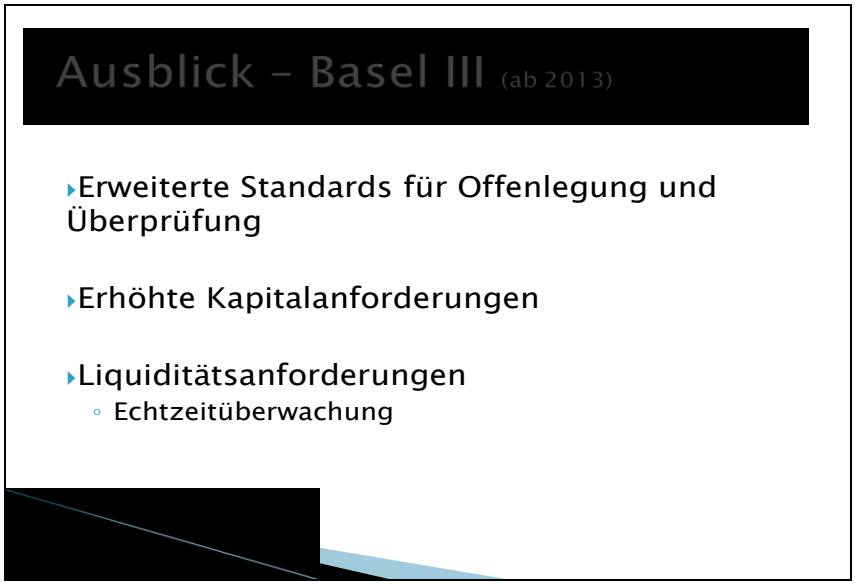
DM strategy: Risk International
http://db.riskwaters.com/data/Risk__free_article_/basel.pdf

PD: Ausfallwahrscheinlichkeit, Verlustquote bei Ausfall, Höhe bei Ausfall -> erwarteter Verlust

http://www.it-observer.com/data-management-challenges-basel-ii-readiness.html
http://www.facebook.com/topic.php?uid=25192258947&topic=5725&_fb_noscript=1



CreditBank Plus AG, Stuttgart
www.information-works.de



http://www.finextra.com/community/fullblog.aspx?blogid=4988
frei verfügbare Anlagen hoher Qualität halten, welche auch in Krisenzeiten verkäuflich, Echtzeit -> data quality challenge

http://www.information-management.com/news/data_risk_management_Basel-10018723-1.html
http://www.pwc.lu/en/risk-management/docs/pwc-basel-III-a-risk-management-perspective.pdf

An additional presentation about RFID & DWH:

R F I D

Radio Frequency Identifikation

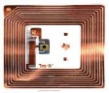
Stefan Baudy, Max Nagel, Andreas Bitzer

Agenda

- Was ist RFID
- Anwendungsgebiete
- RFID & Data-Warehouse
- Ausblick

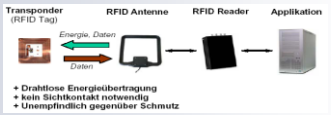
Was ist R F I D

- Kontaktlose Kommunikation über elektromagnetische Wellen
- Sillicon-Chip mit gespeicherter ID
- Abruf von Lesegerät über Aussenden von Wellen
- Chip sendet ID zurück
- Empfänger leitet Information weiter



Was ist RFID → Anwendungsgebiete → RFID & Data Warehouse → Ausblick

R F I D - Funktionalität



Was ist RFID → Anwendungsgebiete → RFID & Data Warehouse → Ausblick

RFID & Data-Warehouse

- Anforderungen an ein DWH
  - hohe Anzahl gleichzeitiger Transaktionen
  - extrem hohe Datenmengen
  - kurze Antwortzeiten
- Edge-Computing
- Dezentale Speicherung der Daten

Was ist RFID → Anwendungsgebiete → **RFID & Data Warehouse** → Ausblick

Anwendungsgebiete RFID

- Barcode: Ersatz, Erweiterung
  - Inventarüberwachung
  - Automatische Lagersysteme
- Sicherheitssysteme
  - Zugangskontrolle
  - Diebstahlschutz
  - Gepäckkontrolle

Was ist RFID → **Anwendungsgebiete** → RFID & Data Warehouse → Ausblick

RFID & Data-Warehouse

- Anforderungen an ein DWH
  - hohe Anzahl gleichzeitiger Transaktionen
  - extrem hohe Datenmengen
  - kurze Antwortzeiten
- Edge-Computing
- Dezentale Speicherung der Daten

Was ist RFID → Anwendungsgebiete → **RFID & Data Warehouse** → Ausblick

Ausblick

- Standards?
- Kosten vs. Nutzen (Barcodeersatz)
- Nutzen vs. Ausspionieren d. Kunden
- Höchst politisches Thema

Was ist RFID → Anwendungsgebiete → RFID & Data Warehouse → **Ausblick**

One further Solution:

RFID

Radio Frequency Identification

Von Friederike Mey

Agenda

- RFID
  - Einsatzmöglichkeiten
  - Funktionsweise
  - Komponenten
  - Herausforderungen
- EPC Global
- Beispielprojekt

Was ist RFID?

- Wird zur Identifikation von Gegenständen und Personen benutzt
- RFID funktioniert mit Hilfe von Radiowellen

Einsatzmöglichkeiten

Produkt Authentifizierung

Viehbestand

Straßenbenutzungs-gebühren, Parkplätze

Marathon

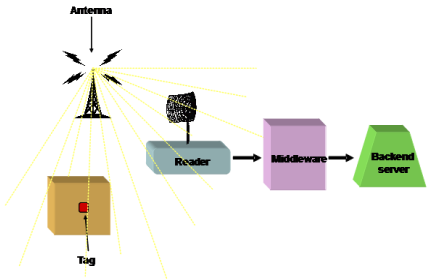
Einsatzmöglichkeiten

Gebäudekontrollen, Sicherheit

Veranstaltungen, Tickets

Warenhaus, Lieferkette, Logistik

Funktionsweise



Komponenten

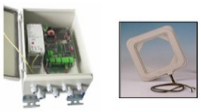
Tag

- Besteht aus einem elektronischen Chip, Speicher, manchmal eigene Energiequellen und einer Antenne

Komponenten

Reader

- Aktiver Teil
- Kann teilweise auch den Tag beschreiben



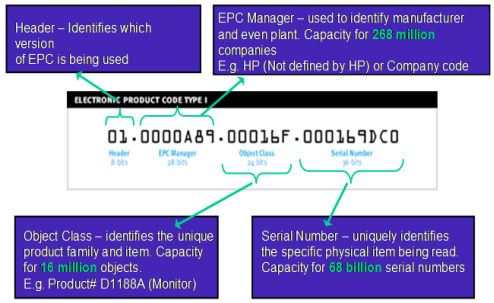
Herausforderungen

- Probleme bei speziellen Materialien (Flüssigkeiten, Metalle)
- Frequenzbereich
- Hohe Kosten
- Datenschutz

EPC Global

- Organisation um Standarts zu setzen
- Beispiele:
  - Electronic Product Code (EPC)

EPC Global



Komponenten

Middleware

- Säubert und filtert die Daten

Printer

- Druckt die Tags und schreibt Daten auf den Chip

EPC Global

- Organisation um Standarts zu setzen
- Beispiele:
  - Electronic Product Code (EPC)
  - Savant
  - Object Name Service (ONS)
  - Physical Markup Language (PML)

Projekte

- Metro Future Store in Rheinberg
  - RFID Innovation Center in Neuss

RFID & Data Warehouse

- Riesige Datenmengen
- Hohe Anzahl gleichzeitiger Transaktionen
- Aufbereitung und Bereitstellung der Daten notwendig
- Kurze Antwortzeiten

Quellen

- Internet
  - www.future-store.org
  - www.hporaclectc.com
  - www.oracle.de
  - www.epcglobal.de
- Buch
  - RFID Handbuch von Klaus Finkenzeller

A further presentation to Basel II/III & DWH (SS2021):

BASEL I…N

EXCERCISE

DATA WAREHOUSE LECTURE

DR. HERMANN VOELLINGER

NICK GRÖßNER

TINF18D

24.02.2021

TOPICS

- The Problem: Basel and the financial crisis
  - Why?
  - The I, II, III of Basel
- What does that mean for me as a big bank?
- Data Mart for Basel
  - What data do we need?
  - Where do we get it from?
  - Is it necessary to merge data?
- Sources

BASEL AND THE CRISIS

- Basel I (1974, 1988):
  - German Herstatt-Bank goes bankrupt
  - Basler Ausschuss für Bankenaufsicht gets founded
  - Equity of 8% necessary when giving credit to someone
- Basel II (2004, 2007):
  - Calculation method was too crude
  - Provisions now stronger tied to the actual credit risk
- Basel III (2009, 2014):
  - 2008 happened…
  - More core capital needed
  - Securities must now consist of a mix of assets
  - Capital conservation buffer, …
- Basel IV (wrsl. 2023):
  - Even more core capital needed

EFFECTS AND OBLIGATIONS

- Calculate size of credit risk (RWA)
  - Information about the client
  - Information about the own financial situation
  - 2 ways of calculation
    - KSA - standardized
    - IRBA – intern
- Credit assessment
  - Much stricter
  - Rating-System for Companies
- We need DATA, DATA, DATA!

Die drei Säulen von Basel II

WHY DATA WARHOUSE?

- Comprehensive picture of customer hast to be made
  - Data privacy?
  - Controlling access
- Big volume of data from different sources required

EIN DATA MART FÜR BASEL II

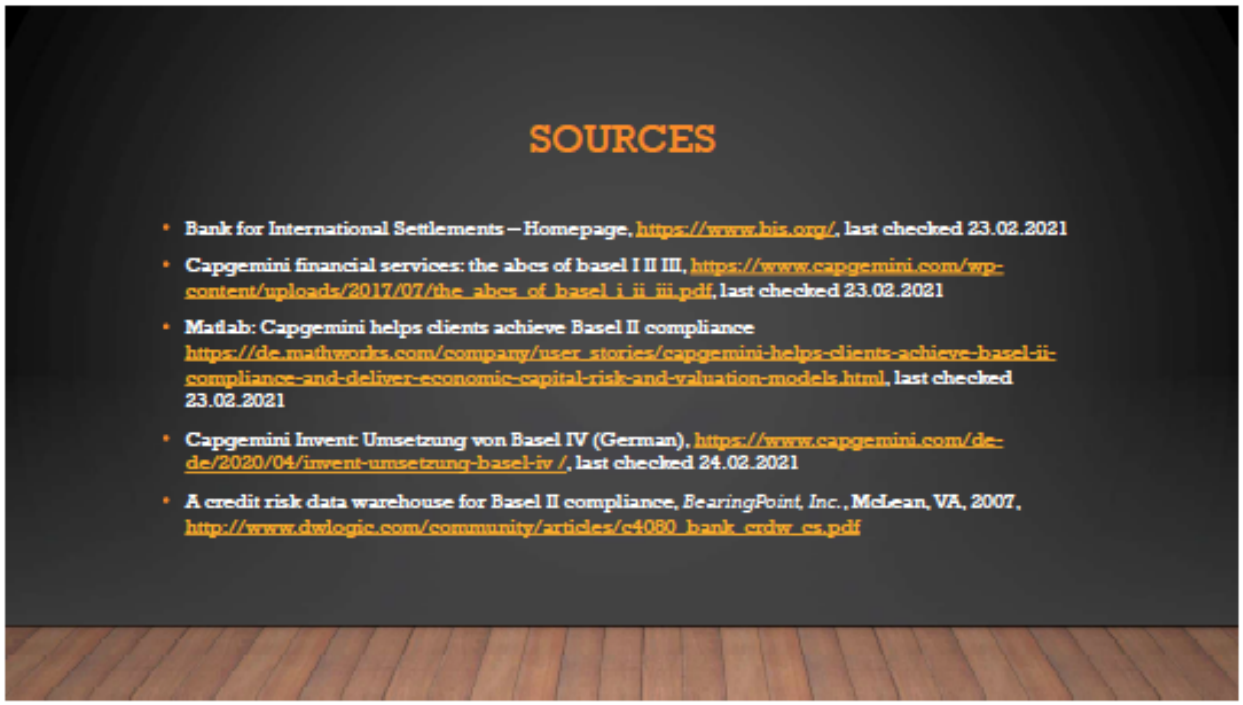
- Bank should structure their data according to data warehouse priciples
- Creation of a Data Mart containing relevant data
  - Credit history
  - Assets (client and bank)
  - Current credit
  - Information from rating companies
  - …

EIN DATA MART FÜR BASEL II

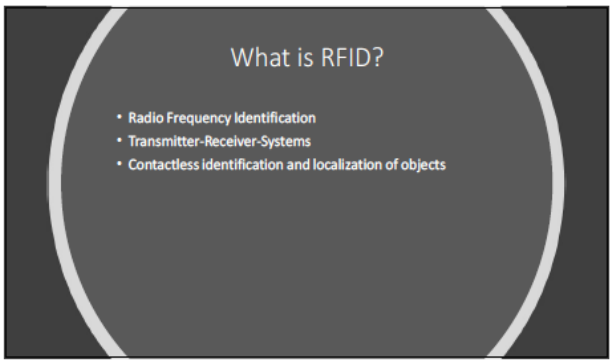
- Source-to-target-mapping
- ETL-Phase – how to unify data?
- Interfaces for the Systems
- Physical data model
- Testing!

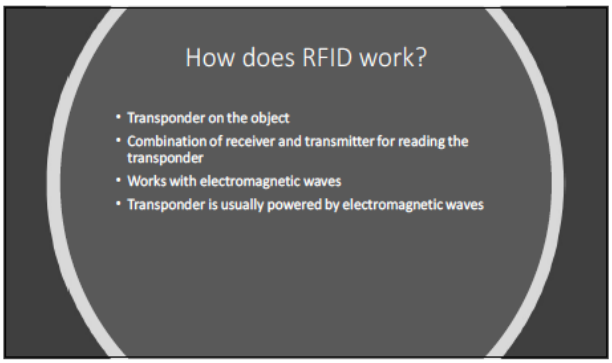
BASEL III AND IV

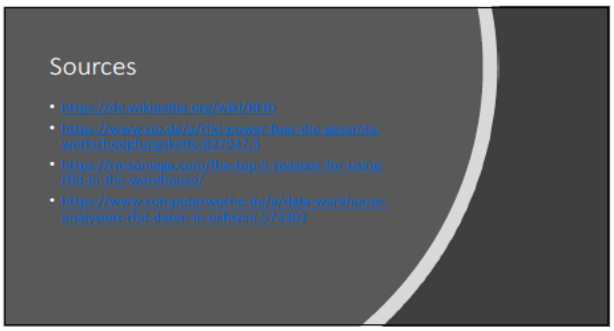
- New methods of data analysis needed
- For data mining and machine learning we need well structured data
- -> Data warehouse is a perfect foundation



A further presentation to RFID & DWH (SS2021):







A further presentation to Basel II/III & RFID (WS2021):

DIE FOLGE

Gründung des
"Basler Ausschuss für Bankenaufsicht"

- Mitglieder: G10-Staaten
- Sitz: An der "Bank für internationalen Zahlungsausgleich" in Basel -> Basler Ausschuss

BASEL I – 1988

Grundlegende Regelung für Eigenkapitalquote der Banken bei Kreditvergabe im Jahre 1988.
(... 14 Jahre nach der Gründung des Basler Ausschusses)

Ziel: Verringerung des Kreditrisikos (Durch fehlende Rückzahlungen)

BASEL II – 2004

RISIKOMANAGEMENT

OPERATIONELLES RISIKO

VALUE-AT-RISK MARKTRISIKO

QUANTIFIZIERUNG DES RISIKOS

KUNDENDATEN

Adresse, Alter, Familienstatus, Liquidität, Eigenkapital, Anlagen, Laufende Kredite

TRANSAKTIONSDATEN

Getätigte Überweisungen, Ausgestellte Kredite, Eigenkapitalquote

DATENQUELLEN

BENCHMARKING

Vorgänge bei anderen Banken, Zinskurse bei Konkurrenten, Werte Entwicklungen, Kundenbindung, Sonderleistungen

MARKTGESCHEHEN

Anlagenentwicklungen, Wirtschaftsnews aus In- & Ausland, Regularieren im In- & Ausland, Krisen, Politische Entwicklungen

BONITÄT

DURCH DWH

RADIO FREQUENCY IDENTIFICATION

Roman Luithle

RFID

- Identifikation von Personen und Gegenständen
- Größere Reichweite durch neue Technologien
  -> mehr Use Cases
- Sicherheitskontrolle, Lagersysteme, Veranstaltungen
- Speicherung kleiner Menge an Daten

Exercises (+Solutions) to DHBW Lecture Intro2DWH – Chapter 3

Exercise E3.1: Overview about 4 Database Types

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 BA database lesson 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

Solution to 3.1.1 - Non-relational databases (IMS, VSAM …):

Datenmodell: Die zur Beschreibung von Daten und deren Beziehungen untereinander auf logischer Ebene zur Verfügung stehenden Datenstrukturen bezeichnet man zusammenfassend als **Datenmodell**.

Dient zur formalen Beschreibung des konzeptionellen (bzw. logischen) Schemas und der externen Schemata mit Hilfe entsprechender Datendefinitionssprachen.

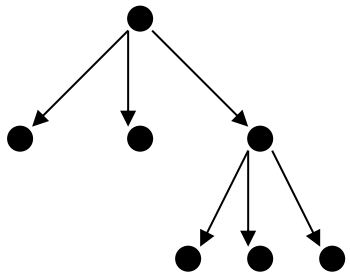
Das Hierarchische Datenmodell – HDM

- primär können nur hierarchisch-baumartige Beziehungen von Objekttypen dargestellt werden.
- Reale Beziehungen sind oft von netzwerkartiger Struktur, sodass Erweiterungen des Datenmodells erforderlich sind => z.B. bei IMS

Strukturelemente:

- Objekttypen
- Hierarchische unbenannte Beziehungen (Kanten haben keine Bezeichnungen)

Ergebnis: Baumstruktur



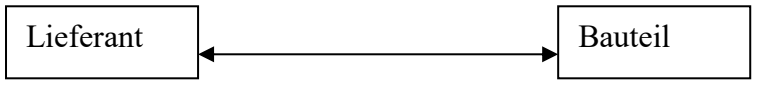
Wurzelbaum-Typ (Hierarchie-Typ) stellt Objekttypen und deren Beziehungen zueinander dar.

Hierarchische Datenbank ist eine Menge von disjunkten Wurzelbaum-(Hierarchie)-Typen.

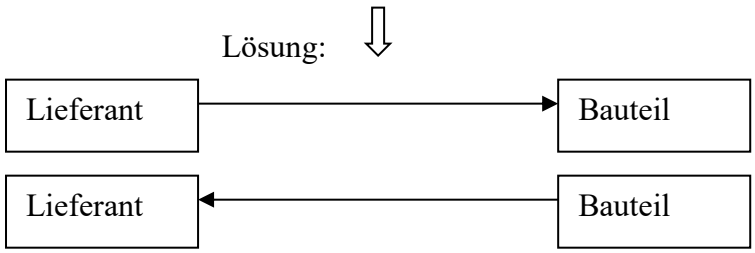
Im hierarchischen Modell ist jedes Wurzel-Objekt über einen Primärschlüssel erreichbar, alle anderen Objekte gemäß der hierarchischen Ordnung. Der Zugriff auf Datenobjekte erfolgt also entlang den logischen Zugriffspfaden (durch Kanten dargestellt). Dies setzt seitens des Anwenders eine genaue Kenntnis der DB-Struktur voraus und bedingt eine prozedurale Beschreibung des Zugriffs. Man spricht bildlich von einem **Navigieren** durch die Datenbank.

Darstellung von Strukturen im HDM:

In einem (strengen) HDM können netzwerkartige Strukturen nicht dargestellt werden. Eine n:m Beziehung, wie z.B. die zwischen Bauteilen und Lieferanten, kann nur durch zwei getrennt Hierarchie-Typen dargestellt werden
=> Redundanz!
Bsp:



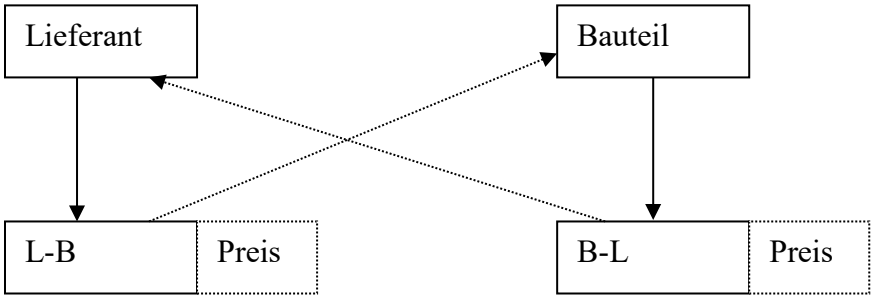
Geht nicht!



Problem: Lieferanten und Bauteile sind mehrfach gespeichert.

Problemlösung: **Pairing**

Abweichend vom strengen HDM werden zusätzliche logische Zugriffe eingeführt, damit n:m Beziehungen dargestellt werden können.



Lieferanten und Bauteile sind nun nur einfach vorhanden.

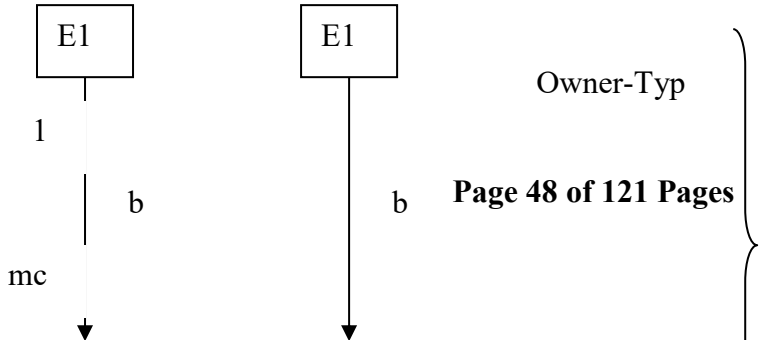
Problem: Preise, die als Attribute bei zusätzlich eingeführten Objekttypen B-L und L-B gespeichert werden, sind immer noch redundant.

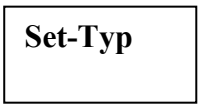
IMS Information Management System:
- kennt keine genaue Unterscheidung zwischen den 3 Schemas (extern, konzeptionell. intern)
- logische Datenmodellierung und physikalische Datenorganisation ineinander verwoben
- Datendefinition erfolgt mit Hilfe der Sprache DL / I
- hierarchische Strukturen können über logische Zeiger miteinander verkettet werden
- Anwender-Sichten können definiert werden

1. Das Netzwerkmodell

- „Erweiterung“ des HDM um netzwerkartige Beziehungen

Strukturelemente:
- Objekttypen
- hierarchische Beziehungen (1:mc), die als Set-Typen bezeichnet werden



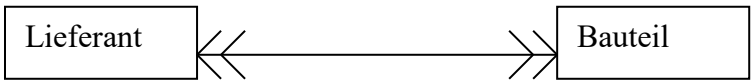


Member-Typ

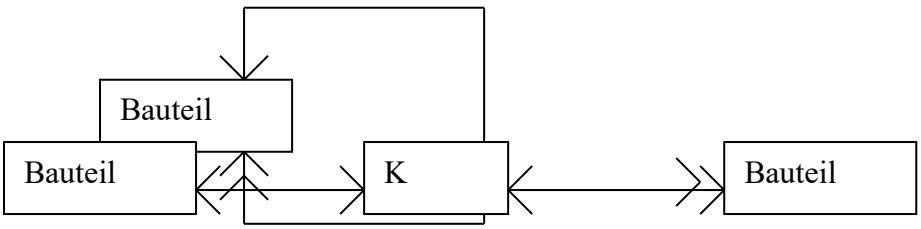
In einem Set-Typ gibt es genau 1 Owner.
1 Owner kann viele Members haben (0 ..*)
1 Owner kann Member sein (in einem anderen Set-Typ), 1 Member kann auch Owner sein

Darstellung von Strukturen im NDM:

1:m ist trivial
m:n durch K ett-Objekt-Typ (link entity type)



Objekt-Typen können auch mit sich selbst in Beziehung stehen, z.B. kann ein Bauteil ein Bauteil eines anderen Bauteils sein.



VSAM: Virtual Storage Access Method

Virtuel: Hardware-Unabhängigkeit, d.h. bei der Dateiorganisation wird primär kein Bezug auf die physische Speicherorganisation (z.B. Zylinder und Spuren der Magnetplatte) genommen.

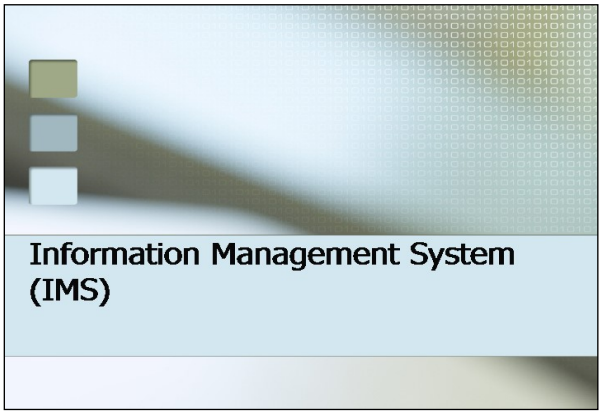
Die auch den B- und B^+-Bäumen zugrunde liegenden Prinzipien, nämlich

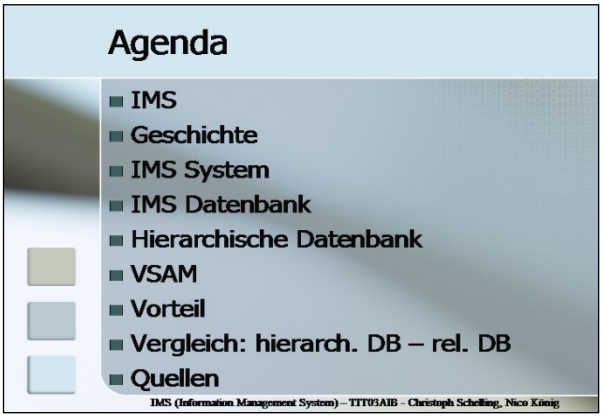
- in Speicherbereichen fester Größe (Knoten) verteilten freien Speicherplatz zur Aufnahme einzufügender Datenobjekte vorzusehen

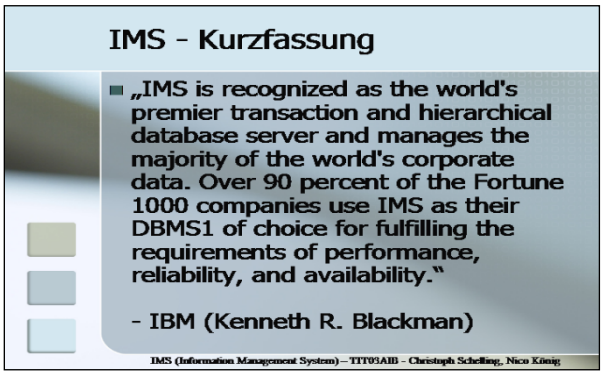
- durch „Zell-Teilung“ (cellular splitting) neuen Speicherplatz zu schaffe, falls der Platz beim Einfügen nicht ausreicht,

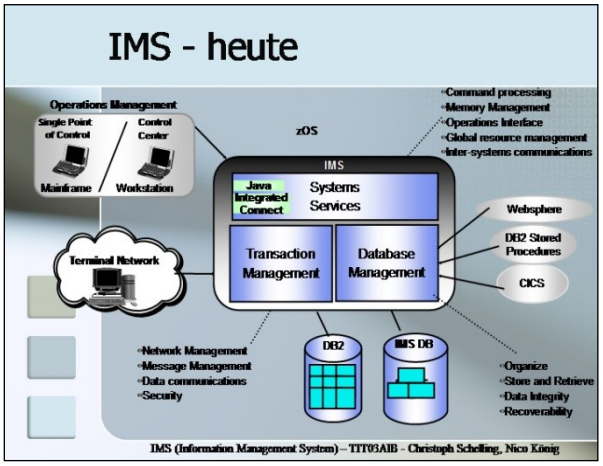
werden hier auch auf die Speicherung der Datensätze selbst (Primärdaten) angewendet und als Index ein B^+-Baum verwendet, dessen Blätter gekettet sind, so dass eine logisch fortlaufende Verarbeitung nach aufsteigenden und absteigenden Schlüsselwerten und auch der (quasi-) direkte Zugriff möglich ist.

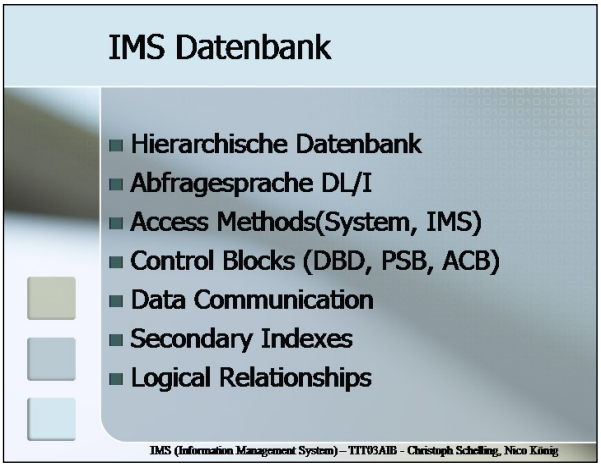
Eine weitere Lösung (2. Lösung):





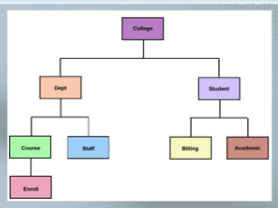






Hierarchische Datenbank

- Beispiel Struktur:

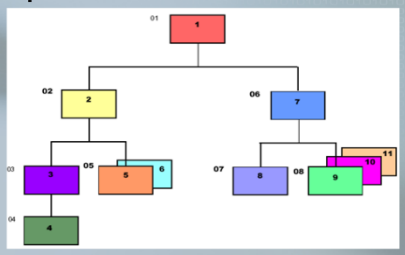


- Segment:

Prefix			Data		
Segment code	Delete byte	counters and pointers	Size field	seq. (key) field	data length varies,
1 byte	1 byte	4 bytes per element	2 bytes		based on a minimum and maximum size

Hierarchische Datenbank

- Suchpfad:



01... = Typ, 1... = Reihenfolge der Suche

VSAM – Virtual Storage Access Method

- Zugriffsmethode auf Dateien in auf IBM Großrechnersystemen (z/OS)
- keine Rücksicht auf physikalische Eigenschaften der Speichermedien mehr nötig
- Speicherung in Cluster

VSAM – Virtual Storage Access Method

- Clusterformen
  - KSDS (Key sequential DataSet):
    - Datenzugriff über einen Index oder sequentiell
  - ESDS (Entry sequential DataSet):
    - sequentiell Zugriff
  - RRDS (Relative Record DataSet):
    - Der Zugriff mit Hilfe von logischen Satznummern

VSAM – Virtual Storage Access Method

- Master-Catalog:
  - jedes System, das VSAM nutzt hat genau einen Master Katalog
  - enthält Informationen zu Datensätzen und Strukturen, um die VSAM Operationen zu steuern
- User Catalog:
  - enthält Einträge zu Anwendungsspezifischen Daten
  - Informationen zur Beschreibung des User Catalog sind im Master Catalog gespeichert

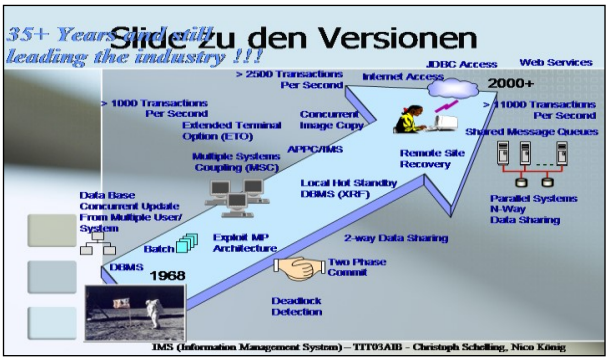
Vergleich hier. DB - rel.DB

- Vorteile:
  - Leistungsstärker (keine techn. Metadaten Verwaltung)
  - Sehr große Datenmengen mit vielen Transaktionen effizienter verwalten
  - Komplexe Abfragen schnell abrufbar
- Nachteile:
  - Hohe Komplexität der Entwicklung
  - Lange Entwicklungsdauer

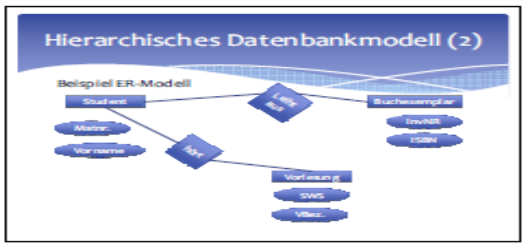
Quellen

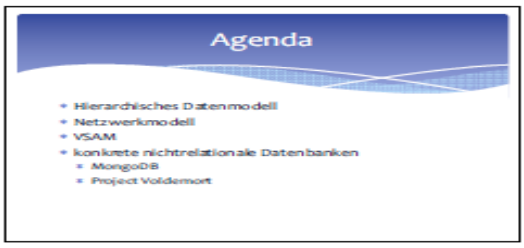
- IBM Webseiten
  www.ibm.com/ims
- DBAzine.com
  http://www.dbazine.com/ofinterest/oi-articles/ims1
- Charles Babbage Institute
  http://www.cbi.umn.edu/shp/entries/ims.html

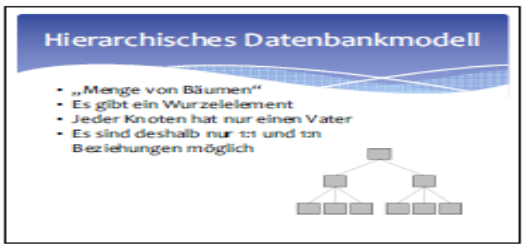
Slide zu den Versionen

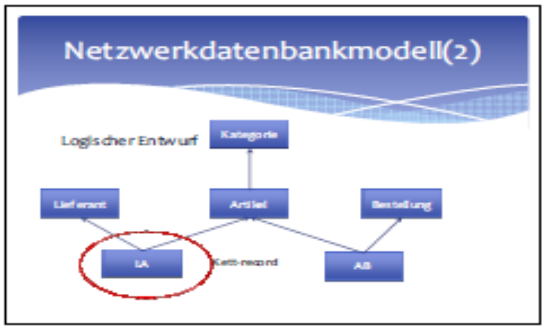


Eine dritte Lösung zu 3.1.1 finden Sie hier:

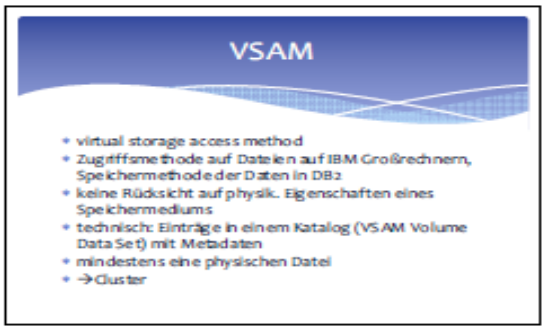




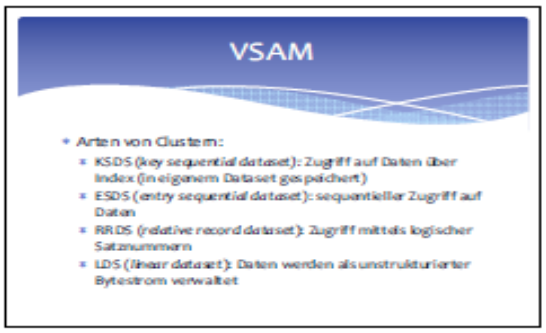


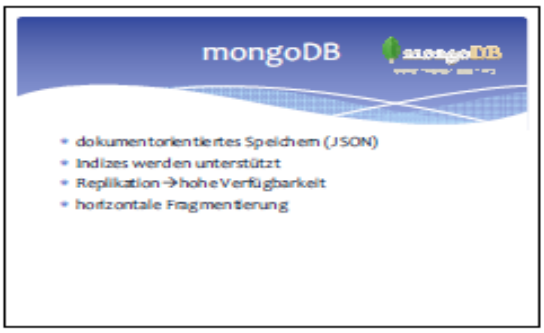

Netzwerkdatenbankmodell(2)
Logischer Entwurf
Kategorie
Lieferant
Artikel
Bestellung
LA
Sets-record
AB


VSAM
Master-Catalog:
jedes System, das VSAM verwendet, hat genau einen
Informationen zu Datensätzen und Strukturen, um VSAM-Operationen zu steuern
User-Catalog:
Einträge zu Anwendungsspezifischen Daten
Informationen zur Beschreibung dieses sind in Master-Catalog


VSAM
virtual storage access method
Zugriffsmethode auf Dateien auf IBM Großrechnern, Speichermethode der Daten in DB2
keine Rücksicht auf physik. Eigenschaften eines Speichermediums
technisch: Einträge in einem Katalog (VSAM Volume Data Set) mit Metadaten
mindestens eine physischen Datei
→Cluster


VSAM
Vorteile:
sehr große Datenmengen mit vielen Transaktionen verwaltbar
komplexe Abfragen schnell aufrufbar
Nachteile:
Hohe Komplexität der Entwicklung
Lange Entwicklungsdauer


VSAM
Arten von Clustern:
KSDS (key sequential dataset): Zugriff auf Daten über Index (in eigenem Dataset gespeichert)
ESDS (entry sequential dataset): sequentieller Zugriff auf Daten
RRDS (relative record dataset): Zugriff mittels logischer Satznummern
LDS (linear dataset): Daten werden als unstrukturierter Bytestrom verwaltet


mongoDB
dokumentorientiertes Speichern (JSON)
Indizes werden unterstützt
Replikation →hohe Verfügbarkeit
horizontale Fragmentierung

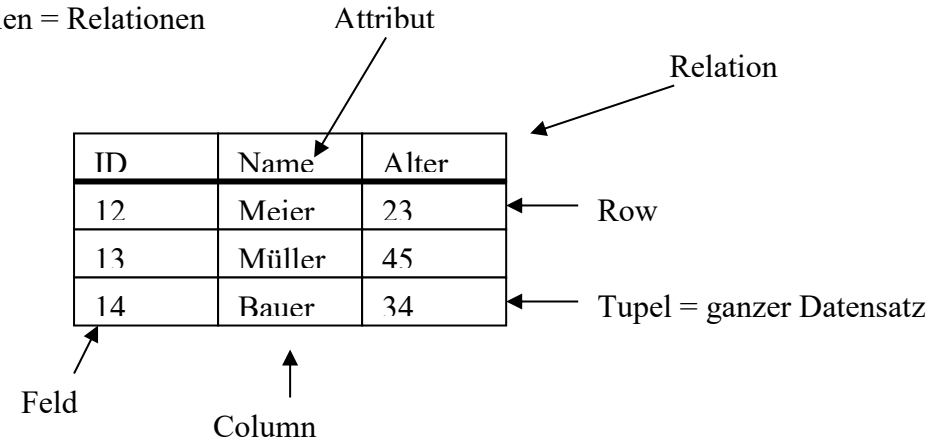

Project Voldemort
key-value storage system
Daten werden automatisch über mehrere Server repliziert / verteilt
Daten werden versioniert um Integrität zu gewährleisten
3 Operationen: value=get(key), put(key,value),delete(key)
effiziente Operationen→vorhersehbare Performance

Solution to 3.1.2 - Relationale Datenbanken:

→ Relation = Beziehung/Abhängigkeit von Objekten und Daten zueinander

→ Definition:
- rel. DB-Modell 1970 von Codd
- Datenspeicherung in Tabellen (Relationen) mit einer festen Anzahl an Spalten und einer flexiblen Anzahl an Zeilen
- Durch das Verteilen der Informationen auf einzelne Tabellen werden Redundanzen vermieden.
- Mit Schlüsselfeldern können Verknüpfungen zw. den Tabellen erstellt werden.

→ Tabellen = Relationen



→ Eine Menge von miteinander verbundenen Relationen bildet eine Datenbank.

→ In einer Tabelle gibt es keine zwei Tupel, die für alle Attribute die gleichen Werte haben.

→ Schlüssel = identifizierende Attributmenge

→ Primärschlüssel
= eine Spalte der Tabelle, durch deren Werte jeder Datensatz der Tabelle eindeutig identifiziert wird.
Der Wert eines Primärschlüsselfeldes einer Tabelle darf nicht doppelt vorkommen.
Jede Tabelle kann nur einen Primärschlüssel haben.
Er kann sich aus mehreren Datenfeldern zusammensetzen und darf nicht leer sein.

→ Fremdschlüssel
= eine Spalte einer Tabelle, deren Werte auf den Primärschlüssel einer anderen Tabelle verweisen.
Eine Tabelle kann mehrere Fremdschlüssel enthalten.
Er kann aus mehreren Feldern der Tabelle bestehen, er kann leer sein.
Für jeden Wert eines Fremdschlüssels muss es einen entsprechenden Wert im Primärschlüssel der korrespondierenden Tabelle geben (Integrität)

→ Basisoperationen: (siehe SQL-Anweisungen)
- Selektion
- Verbund

- Projektion

→ Weitere Regeln der relationalen Datenbank:

- Transaktionen müssen entweder vollständig durchgeführt werden oder, bei einem Abbruch, vollständig zurückgesetzt werden.
- Der Zugriff auf die Daten durch den Benutzer muss unabhängig davon sein, wie die Daten gespeichert wurden oder wie physikalisch auf sie zugegriffen wird.
- Ändert der Datenbankverwalter die physikalische Struktur, darf der Anwender davon nichts mitbekommen.

Solution to 3.1.3- SQL Basics:

Compare standard books about SQL language

Solution to 3.1.4 - Normalization:

Ziel von Normalformen

- Update-Anomalien innerhalb einer Relation vermeiden
- Update-Anomalien: Redundanzen in Datenbanken, die einerseits unnötigen Speicherplatz verbrauchen und andererseits dazu führen, dass sich Änderungsoperationen nur schwer umsetzen lassen (Änderung bei allen Vorkommen einer Information)
- Ziel: Redundanzen entfernen, die aufgrund von funktionalen Abhängigkeiten innerhalb einer Relation entstehen

Abhängigkeiten

a) funktional abhängig
zu einer Attributkombination von A gibt es genau eine Attributkombination von B
B ist funktional abhängig von A: A -> B

b) voll funktional abhängig
A und B als Attributkombination der gleichen Relation R
B ist voll funktional abhängig von A, wenn es von der gesamten Attributkombination von A funktional abhängt, aber nicht schon von einem Teil: A => B

c) transitiv abhängig
B ist abhängig von A und C ist abhängig von B: A -> B -> C
C darf dabei nicht Schlüsselattribut sein und nicht in B vorkommen

Anomalien:

Prüfungsgeschehen

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

Einfüge-Anomalien
Wo fügt man in dieser Relation einen Studenten ein, der noch nicht an einer Prüfung teilgenommen hat?

a) Lösch-Anomalien
Mit Löschung des Studenten Pitt, geht auch die Information über den Dekan vom Dachbereich BWL verloren.

b) Änderungs-Anomalien
Zieht ein Student um, der an mehreren Prüfungen teilgenommen hat, so muß die Adressänderung in mehreren Tupeln vollzogen werden

Exercise E3.2: Build Join Strategies

Build all join strategies for the following tables SAMP_PROJECT and SAMP_STAFF: i.e.

1. *Cross Product*
2. *Inner Join*
3. *Outer Join*
    a. *Left Outer Join*
    b. *Right Outer Join*
    c. *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***

Solution:

See lesson notes.

Exercise E3.3: Example of a Normalization

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

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

Solution:

Erste Normalform

- Nur atomare Attribute, also Elemente von Standard-Datentypen und nicht Listen, Tabellen oder ähnliche komplexe Strukturen

Prüfungsgeschehen

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

Bsp. enthält eine weitere Relation
1. Lösung: jede Zeile um die ersten drei Attribute erweitern, dann entstehen aber Redundanzen
2. Lösung: Auslagerung in eine neue Tabelle Prüfung

PNR	Fach	Prüfer
3	Elektronik	Richter
4	Informatik	Schwinn
5	TMS	Müller

Prüfling

PNR	MATNR	Name	Geb	Adr	Fachbereich	Dekan	Note
3	123456	Meier	010203	Weg 1	Informatik	Wutz	1
3	124538	Schulz	050678	Str 1	Informatik	Wutz	2
4	245633	Kunz	021279	Gas. 2	Informatik	Wutz	1
4	124538	Schulz	050678	Str 1	Informatik	Wutz	1
5	856214	Schmidt	120178	Str 2	Informatik	Wutz	3
5	369852	Pitt	140677	Gas. 1	BWL	Butz	1

Beide Relationen sind nun in 1. NF

Zweite Normalform

Ziel: aufgrund von funktionalen Abhängigkeiten Redundanzen entdecken
Erlaubt keine partiellen Abhängigkeiten zwischen Schlüsseln des Relationen Schemas und weiteren Attributen (jedes Nicht-Primärattribut muss also voll funktional abhängig sein von jedem Schlüsselattribut der Relation)

Prüfling

PNR	MATNR	Name	Geb	Adr	Fachbereich	Dekan	Note
3	123456	Meier	010203	Weg 1	Informatik	Wutz	1
3	124538	Schulz	050678	Str 1	Informatik	Wutz	2
4	245633	Kunz	021279	Gas. 2	Informatik	Wutz	1
4	124538	Schulz	050678	Str 1	Informatik	Wutz	1
5	856214	Schmidt	120178	Str 2	Informatik	Wutz	3
5	369852	Pitt	140677	Gas. 1	BWL	Butz	1

Erkennbar: Daten des Studenten (Name, Geb, Adr, Fachbereich, Dekan) hängen nur von MATNR ab und nicht von PNR, ist somit nicht voll funktional abhängig
Erzeugung der zweiten Normalform durch Elimination der rechten Seite der partiellen Abhängigkeit und Kopie der linken Seite

Student

MATNR	Name	Geb	Adr	Fachbereich	Dekan
123456	Meier	010203	Weg 1	Informatik	Wutz
124538	Schulz	050678	Str 1	Informatik	Wutz
245633	Kunz	021279	Gas. 2	Informatik	Wutz
124538	Schulz	050678	Str 1	Informatik	Wutz
856214	Schmidt	120178	Str 2	Informatik	Wutz
369852	Pitt	140677	Gas. 1	BWL	Butz

Prüfungsergebnis

PNR	MATNR	Note
3	123456	1
3	124538	2
4	245633	1
4	124538	1
5	856214	3
5	369852	1

- Eine Relation R ist in 2. NF, wenn sie in 1.NF ist und jedes Nicht-Primärattribut von R voll von jedem Schlüssel in R abhängt (also keine Attribute des Schlüssels unwesentlich ist)
- Problem der Anomalien noch nicht beseitigt
  Einfüge-A.: Fachbereichsdaten nicht ohne eingeschriebenen Studenten speicherbar
  Lösch-A: Fachbereichsdaten verschwinden mit Löschen des letzten Studenten
  Änderungs-A: Wechsel des Dekans muss an mehreren Stellen vollzogen werden

Dritte Normalform

- 3. NF: keine transitiven Abhängigkeiten

Student

MATNR	Name	Geb	Adr	Fachbereich	Dekan
123456	Meier	010203	Weg 1	Informatik	Wutz
124538	Schulz	050678	Str 1	Informatik	Wutz
245633	Kunz	021279	Gas. 2	Informatik	Wutz
124538	Schulz	050678	Str 1	Informatik	Wutz
856214	Schmidt	120178	Str 2	Informatik	Wutz
369852	Pitt	140677	Gas. 1	BWL	Butz

- transitive Abhängigkeit: Dekan ist von Fachbereich abhängig, da es zu jedem Fachbereich genau einen Dekan gibt (demnach ist Dekan transitiv abhängig von MATNR)
- Eliminieren von transitiven Abhängigkeiten: Auslagerung der abhängigen Attribute in eine neue Relation

Fachbereich

Fachbereich	Dekan
Informatik	Wutz
BWL	Butz

Student

MATNR	Name	Geb	Adr	Fachbereich
123456	Meier	010203	Weg 1	Informatik
124538	Schulz	050678	Str 1	Informatik
245633	Kunz	021279	Gas. 2	Informatik
124538	Schulz	050678	Str 1	Informatik
856214	Schmidt	120178	Str 2	Informatik
369852	Pitt	140677	Gas. 1	BWL

Exercise E3.4: Example of a Normalization

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

Solution to 3.4:

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

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

.....

Exercises (+Solutions) to DHBW Lecture Intro2DWH – Chapter 4

Exercise E4.1: Create SQL Queries

Given the two tables:

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

Define the right SQL such that:

1. you get a list of airports which have no incoming flights (no arrivals) (6 points)
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 (6 points)
3. insert a new flight from BER to HAN at 17:30 with FNo 471 (4 points)
4. Change FlightTime of Fno=181 to 10:35 (4 points)

Optional (difficult) –10 points:

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

Solution:

Ad 1.:

```
select fid, name from airport
where fid not in
 (select distinct to from flight)
```

Ad 2.:

```
create view Flights_to_Munich2
as select f.Fno as FNr, a.name as Dep_Airp, f.time as DepT  from flight f, airport
a
where f.to='MUC' and a.fid=f.from
```

Ad3.:

```
insert into flight
values (471,'BER','HAN','17.30.00')
```

Ad4.:

```
update flight
set time = '10.35.00'
```

```
where Fno=181
```

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:

```
**********************************************************************
db2 => select name as Departure_ Airport, count (*) as Departure_Count from airpo
rt, flight where fid=from group by name union select name as Departure_Airpo
rt, 0 as Departure_Count from airport where not exists (select * from flight whe
re from=fid) order by departure_count

DEPARTURE_AIRPORT    DEPARTURE_COUNT
---------------------------  ----------------------------
Berlin                                0
Frankfurt                             0
Hannover                              1
Mannheim                              1
Stuttgart                             1
Muenchen                              2

  6 record(s) selected.
**********************************************************************
```

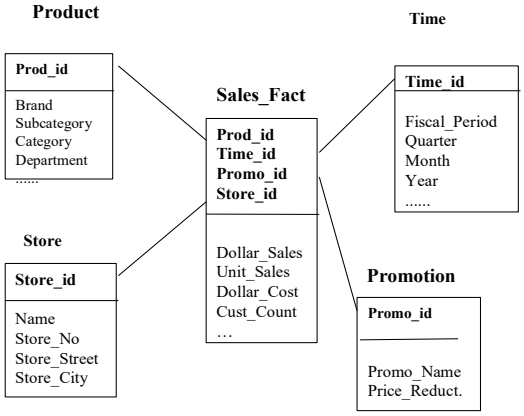
Here is also a **second solution** (which is shorter) and gives the same results as above by **Stefan Seufert**:

```
SELECT Name as Departure_Airport, count (Flight.From) as  Departure_Count
FROM Airport LEFT OUTER JOIN Flight ON Airport.FID = Flight.From
GROUP BY Name
ORDER BY Departure_Count
```

The idea is, that count(Field) in contradiction to count(*) only count the fields which are not NULL. Since the attribute in the count function is from the flight table, only the flights which have departures are counted, all other get the 0 value.

Exercise E4.2: Build SQL for a STAR Schema

Consider the following Star Schema:



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	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**':

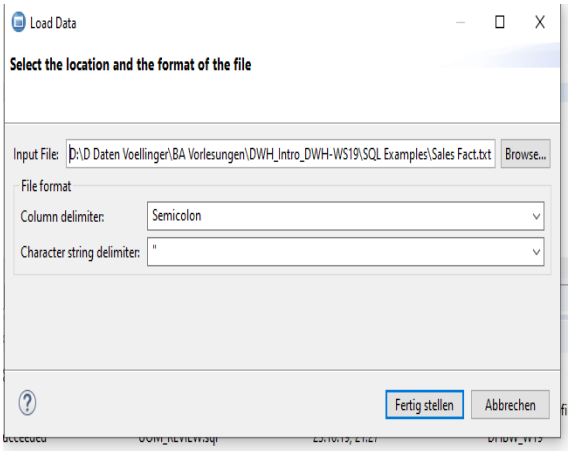
SELECT Product.brand **AS** Brand, Sum([Sales Fact].dollar_sales) **AS** Dollar_Sales,Sum([Sales Fact].unit_sales) **AS** Unit_Sales
FROM ([Sales Fact]
INNER JOIN [Time] **ON** [Sales Fact].time_key = Time.time_key)
INNER JOIN Product **ON** [Sales Fact].product_key = Product.product_key
WHERE (((Time.fiscal_period)="4Q95"))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

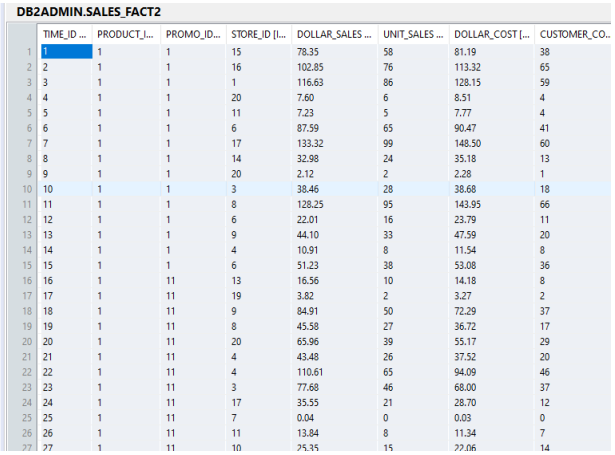
Solution with Standard SQL(for example with DB2) by loading the data (flat files) into DB2:

First connect to database "Grocery". Then create the necessary tables and load the data from flat Files (*.txt Files) into the corresponding tables:

```
CREATE TABLE "DB2ADMIN"."SALES_FACT" (
                "TIME_ID" INTEGER,
                "PRODUCT_ID" INTEGER,
                "PROMO_ID" INTEGER,
                "STORE_ID" INTEGER,
                "DOLLAR_SALES" DECIMAL(7 , 2),
                "UNIT_SALES" INTEGER,
                "DOLLAR_COST" DECIMAL(7 , 2),
                "CUSTOMER_COUNT" INTEGER
        )
        ORGANIZE BY ROW
        DATA CAPTURE NONE
        IN "USERSPACE1"
        COMPRESS NO;
```

Load the data from the Sales_Fact.txt file by using the "Load Data" feature of the table DB2ADMIN.Sales_Fact in the GROCERY database:





Do the same for the four dimension-tables: "Product", "Time", "Store" and "Promotion".

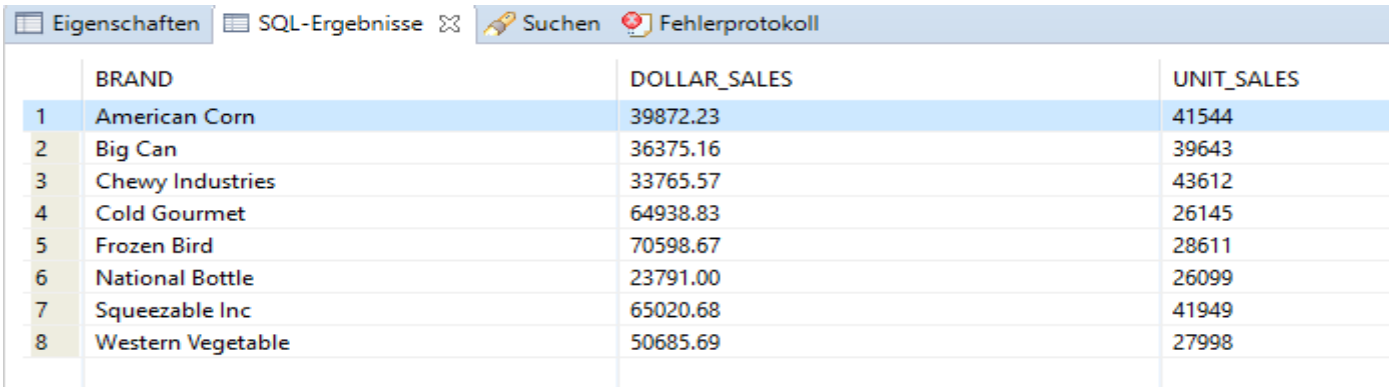
```
CREATE TABLE "DB2ADMIN"."TIME" ("TIME_ID" INTEGER,
            "DATE" varchar(20),"DAY_IN_WEEK" varchar(12),
```

```
              "DAY_NUMBER_IN_MONTH" Double,
              "DAY_NUMBER_OVERALL" Double,
              "WEEK_NUMBER_IN_YEAR" Double,
              "WEEK_NUMBER_OVERALL" Double,
               "MONTH" Double, "QUARTER" int,
              "FISCAL_PERIOD" varchar(4),"YEAR" int,
             "HOLIDAY_FLAG" varchar(1))
       ORGANIZE BY ROW
       DATA CAPTURE NONE
       IN "USERSPACE1"
       COMPRESS NO;

CREATE TABLE "DB2ADMIN"."PRODUCT" ("PRODUCT_ID" INTEGER,
              "DESCRIPTION" varchar(20),"FULL_DESCRIPTION" varchar(30),
              "SKU_NUMBER" decimal(12,0),"PACKAGE_SIZE" varchar(8),
              "BRAND" varchar(20),"SUBCATEGORY" varchar(20), "CATEGORY"
varchar(15),
              "DEPARTMENT" varchar(15),"PACKAGE_TYPE" varchar(12),"DIET_TYPE"
varchar(10),
               "WEIGHT" decimal(5,2),"WEIGHT_UNIT_OF_MEASURE" varchar(2),
              "UNITS_PER_RETAIL_CASE" int,"UNITS_PER_SHIPPING_CASE" int,
              "CASES_PER_PALLET" int, "SHELF_WIDTH_CM" decimal(8,4),
               "SHELF_HEIGHT_CM" decimal(8,4),"SHELF_DEPTH_CM" decimal(8,4))
       ORGANIZE BY ROW
       DATA CAPTURE NONE
       IN "USERSPACE1"
       COMPRESS NO;
```

Finally run the SQL to produce the result for the quarter “4Q95”:

```
SELECT p.BRAND AS Brand, Sum(s.DOLLAR_SALES) AS Dollar_Sales, Sum(s.UNIT_SALES) AS
Unit_Sales
FROM "DB2ADMIN"."SALES_FACT" s, "DB2ADMIN"."PRODUCT" p, "DB2ADMIN"."TIME" t
WHERE p.PRODUCT_ID = s.PRODUCT_ID
      AND s.TIME_ID = t.TIME_ID
      AND t."FISCAL_PERIOD" = '4Q95'
GROUP BY p.BRAND
ORDER BY p.BRAND;
```



	BRAND	DOLLAR_SALES	UNIT_SALES
1	American Corn	39872.23	41544
2	Big Can	36375.16	39643
3	Chewy Industries	33765.57	43612
4	Cold Gourmet	64938.83	26145
5	Frozen Bird	70598.67	28611
6	National Bottle	23791.00	26099
7	Squeezable Inc	65020.68	41949
8	Western Vegetable	50685.69	27998

Alternative:

```
SELECT p.BRAND AS Brand, Sum(s.DOLLAR_SALES) AS Dollar_Sales, Sum(s.UNIT_SALES) AS
Unit_Sales
FROM "DB2ADMIN"."SALES_FACT" s, "DB2ADMIN"."PRODUCT" p, "DB2ADMIN"."TIME" t
WHERE p.PRODUCT_ID = s.PRODUCT_ID
      AND s.TIME_ID = t.TIME_ID
      AND t.QUARTER = 4
```

```
      AND t.YEAR = 1995
GROUP BY p.BRAND
ORDER BY p.BRAND;
```

Eigenschaften | SQL-Ergebnisse | Suchen | Fehlerprotokoll

	BRAND	DOLLAR_SALES	UNIT_SALES
1	American Corn	39872.23	41544
2	Big Can	36375.16	39643
3	Chewy Industries	33765.57	43612
4	Cold Gourmet	64938.83	26145
5	Frozen Bird	70598.67	28611
6	National Bottle	23791.00	26099
7	Squeezable Inc	65020.68	41949
8	Western Vegetable	50685.69	27998

Finally run the SQL to produce the result for the both quarters “4Q95” and “4Q96”:

```
SELECT p.BRAND AS Brand, Sum(s.DOLLAR_SALES) AS Dollar_Sales, Sum(s.UNIT_SALES) AS
Unit_Sales
FROM "DB2ADMIN"."SALES_FACT" s, "DB2ADMIN"."PRODUCT" p, "DB2ADMIN"."TIME" t
WHERE p.PRODUCT_ID = s.PRODUCT_ID
      AND s.TIME_ID = t.TIME_ID
      AND (t."FISCAL_PERIOD" = '4Q95' OR t."FISCAL_PERIOD" = '4Q94')
GROUP BY p.BRAND
ORDER BY p.BRAND;
```

Alternative:
You just omit the selection of a special quarter. In addition, you can create a View with name “Sales_Per_Brand”:

```
Create View "DB2ADMIN"."Sales_Per_Brand" AS
SELECT p.BRAND AS Brand, Sum(s.DOLLAR_SALES) AS Dollar_Sales, Sum(s.UNIT_SALES) AS
Unit_Sales
FROM "DB2ADMIN"."SALES_FACT" s, "DB2ADMIN"."PRODUCT" p, "DB2ADMIN"."TIME" t
WHERE p.PRODUCT_ID = s.PRODUCT_ID
      AND s.TIME_ID = t.TIME_ID
GROUP BY p.BRAND;
```

Remark: You have also to omit “ORDER BY” not to get an error in DB2. Nevertheless, the result is ordered automatically by the brand name. See resulting view:

DB2ADMIN.Sales_Per_Brand

	BRAND [VARCHAR(20 OCTETS)]	DOLLAR_SALES [DECIMAL(31 , 2)]	UNIT_SALES [INTEGER]
1	American Corn	84361.00	82117
2	Big Can	73730.29	80474
3	Chewy Industries	65646.03	84850
4	Cold Gourmet	135002.09	53548
5	Frozen Bird	140953.10	56070
6	National Bottle	49418.31	54309
7	Squeezable Inc	129494.04	83219
8	Western Vegetable	102798.73	56133

```
Create View "DB2ADMIN"."Sales_Per_Brand1" AS
SELECT p.BRAND AS Brand, Sum(s.DOLLAR_SALES) AS Dollar_Sales, Sum(s.UNIT_SALES) AS
Unit_Sales
FROM "DB2ADMIN"."SALES_FACT" s, "DB2ADMIN"."PRODUCT" p, "DB2ADMIN"."TIME" t
WHERE p.PRODUCT_ID = s.PRODUCT_ID
      AND s.TIME_ID = t.TIME_ID
      AND (t."FISCAL_PERIOD" = '4Q95' OR t."FISCAL_PERIOD" = '4Q94')
GROUP BY p.BRAND;
```

DB2ADMIN.Sales_Per_Brand1

	BRAND [VARCHAR(20 OCTETS)]	DOLLAR_SALES [DECIMAL(31 , 2)]	UNIT_SALES [INTEGER]
1	American Corn	84361.00	82117
2	Big Can	73730.29	80474
3	Chewy Industries	65646.03	84850
4	Cold Gourmet	135002.09	53548
5	Frozen Bird	140953.10	56070
6	National Bottle	49418.31	54309
7	Squeezable Inc	129494.04	83219
8	Western Vegetable	102798.73	56133

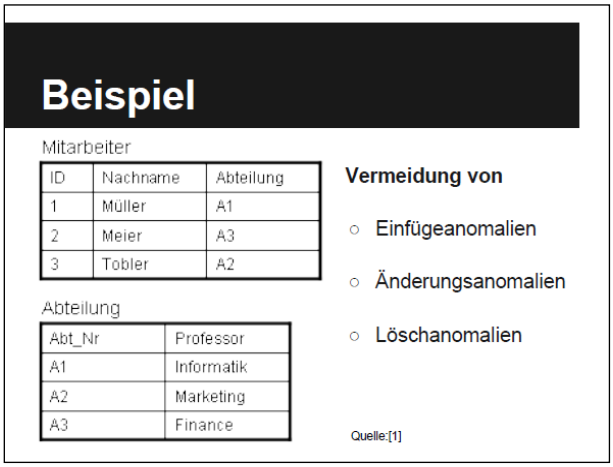
Exercise E4.3*: Advanced Study about Referential Integrity

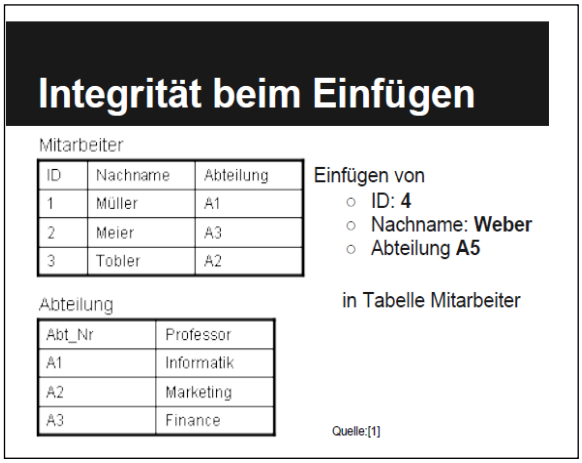
Explain: What is "Referential Integrity" (RI) in a Database?
Sub-Questions:

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.

First SOLUTION:





Gründe für RI im DWH

- Datenkonsistenz
- Änderungen müssen nur an einer Stelle durchgeführt werden.
- Integritätsprobleme werden verhindert

Gründe gegen RI im DWH

- Längere Zugriffszeiten / Viel Overhead
- Zu große Datenmengen für konventionelle RI-Methoden aus DBMS
- RI Prüfung wird meist vor dem Laden der Daten ins DWH durchgeführt
- Im DWH werden keine Updates durchgeführt

Quelle: [2]

Mögliche Lösung

- Bounded Referential Integrity Lösungsvorschlag von Bill Inmon
- Ähnelt stark der klassischen RI
  - Nur Teilmengen werden und keine ganzen Tabellen werden geprüft
  - Prüfung findet erst nach dem Laden der Daten ins DWH statt

Quelle: [2]

Quellen

[1] http://www.gitta.info/LogicModelin/de/html/DBIntegrity_Ref_Integ.html

[2]http://social.technet.microsoft.com/Forums/it-IT/sqlserverit/thread/0b51568a-7d19-4afc-87a5-928b9ecd4a6b

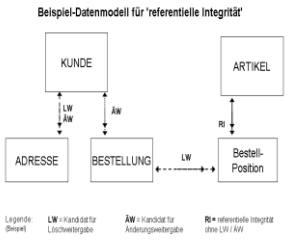
Second SOLUTION:

REFERENTIELLE INTEGRITÄT

- Sicherung der Datenintegrität bei RDB
- Datensätze dürfen nur auf existierende Datensätze verweisen

FRANCOIS TWEER-ROLLER & MARCO ROSIN 12/17/2013 11

BEISPIEL



FRANCOIS TWEER-ROLLER & MARCO ROSIN 12/17/2013 12

Third Solution:

Definition
"Über referentielle Integrität werden in einem DBMS die Beziehungen zwischen Datenobjekten kontrolliert".
Vorteile

- Steigerung der Datenqualität: Referenzielle Integrität hilft Fehler zu vermeiden.
- Schnellere Entwicklung: Referenzielle Integrität muss nicht in jeder Applikation neu implementiert werden.
- Weniger Fehler: Einmal definierte referenzielle Integritätsbedingungen gelten für alle Applikationen derselben Datenbank
- Konsistentere Applikationen: Referenzielle Integrität ist für alle Applikationen, die auf dieselbe Datenbank zugreifen gleich.

Nachteile

- Löschproblematik aufgrund von Integrität
- Temporäres außer Kraft setzen der RI für großen Datenimport.

Referenzielle Integrität in einem DWH

- Daten müssen im DWH nicht 100%ig konsistent sein.
- Durch Import von großen Datenmengen ist die Kontrolle der Integrität zu aufwendig
- Inkonsistente Daten können in keinen konsistenten Zustand gebracht werden.

Meiner Meinung nach ist die Realisierung von der referentiellen Integrität möglich, aber mit viel Aufwand und Kosten verbunden.

Fourth Solution (SS2021):

Referential Integrity (RI)

Von: Arkan Abdel
Leonard Faix

Agande

- What is Referential Integrity (RI) in a Database?
- What means RI in a Data Warehouse?
- Should one have RI in a DWH or not?
- pro and cons of Referential Integrity (RI)

Referential Integrity (RI) in a Database?

- the relational data in database tables has to be universally configurable
- keys that reference elements of other tables need to be connected to those other fields
- not separately
- prevents errors

What means RI in a Data Warehouse

- Referential Integrity in the data warehouse is a form of data integrity
- Relational databases break the storage of data down into elements
- data would get dropped (If it is not implemented properly)

Integrity Constraints :

- Impose restrictions on allowable data,beyond those
- Imposed by structure and types

Referential integrity

. Integrity of references
. No dangling pointers

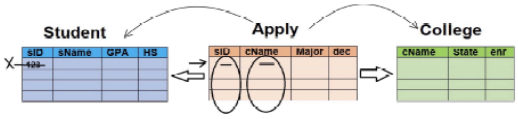
Referential integrity from R.A to S.B
Each value in column A of table R must appear in column B of table S

- A is called the „foreign Key" (foreign key constraints)
- B is usually required to be the primary key for table S or at least unique
- Multi-attribute foreign keys are allowed

Referential Integrity Enforcement (R.A to S.B)

- Potentially violating modifications :
- Insert into R
- Delete from S
- Update R.A
- Update S.B

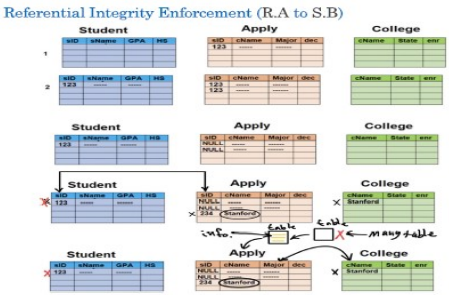
If violation → error



Referential Integrity Enforcement (R.A to S.B)

Special actions:

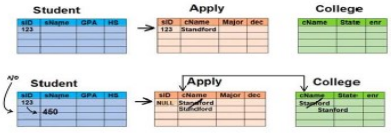
- Delete from S → Error
  Restrict (default), set Null, Cascade
- Update S.B



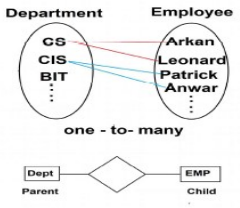
Referential Integrity Enforcement (R.A to S.B)

Special actions:

- Delete from S → Error
  Restrict (default), set Null, Cascade
- Update S.B
  Restrict (default), set Null, Cascade



Beispiel :



EMP

EID	First	Last	Salary	Dno
10	Arkan	Abdel	42000	1
5	Leonard	Faix	60000	2
4	Patrick	Foucks	65000	1
7	Anwar	Adial	7000	2
8	jon	mark	55000	3

DEP.

DID	Dname	Location
1	CS	floor 1
2	CIS	floor 2
3	BIT	floor 3

Die Problemlösung :

Create table Dept (DID char(1) Primary Key, Dname varchar(20) , Location varchar(20) ,

on delete Cascade

on update Cascade);

+ We can create our own Data Type :

Create type TypeName as varchar(30) ;

create table Emp (ID.....

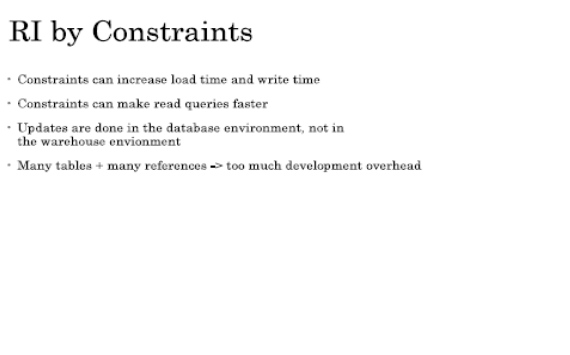
Last TypeName.....

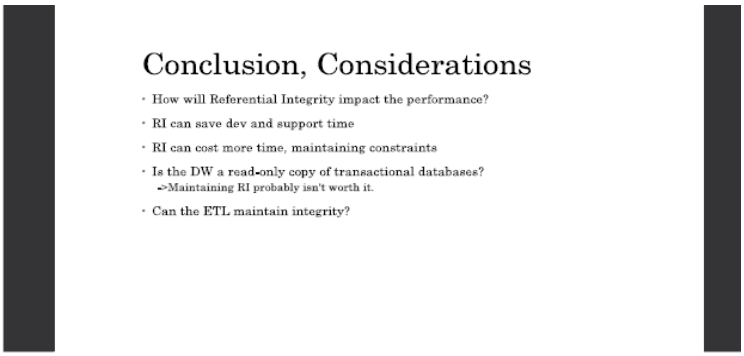
Why should I enforce RI?

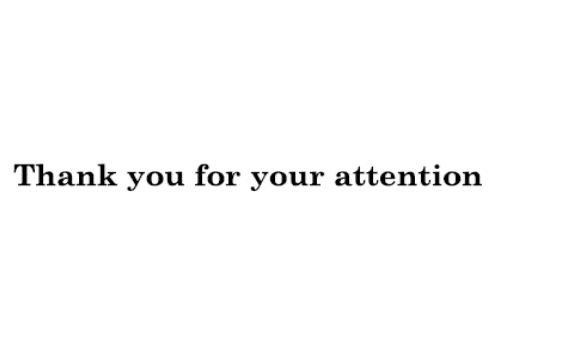
- Ensuring that relationships between rows of data exist and are used as they are defined.
- User can trust data and rely on relationships

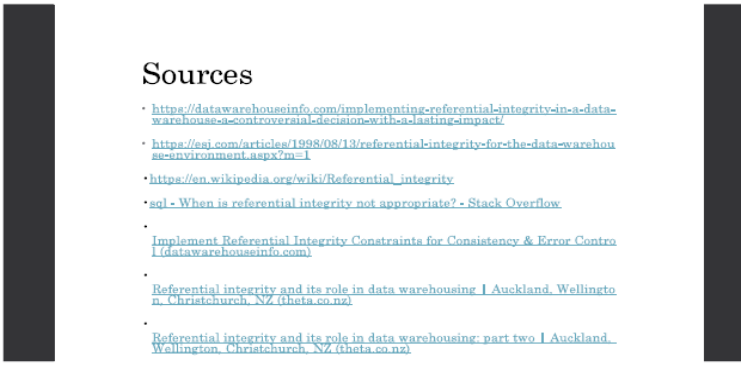
Referential Integrity in Data Warehouse

- Referential integrity is a decision, not a standard practice. It depends on the data
- ETL can ensure RI -> need strong control over ETL
- Constraints can ensure RI
  - Foreign Key...









Exercises (+Solutions) to DHBW Lecture Intro2DWH – Chapter 5

Exercise E5.1: Compare ER and MDDM

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/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 --

Solution:

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.
MDDM Model: Compare as examples the Star - and Snowflake schemas, which are explained in the next solution (5.2)

Exercise E5.2*: Compare Star and SNOWFLAKE

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 (compare exercise 5.1):

Solution:

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

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, de-normalized 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.

Expert Meaning (from DM Review):

First, let's describe them.

A star schema is a dimensional structure in which a single fact is surrounded by a single circle of dimensions; any dimension that is multileveled is flattened out into a single dimension. The star schema is designed for direct support of queries that have an inherent dimension-fact structure.

A snowflake is also a structure in which a single fact is surrounded by a single circle of dimensions; however, in any dimension that is multileveled, at least one dimension structure is kept separate. The snowflake schema is designed for flexible querying across more complex dimension relationships. The snowflake schema is suitable for many-to-many and one-to-many relationships among related dimension levels. However, and this is significant, the snowflake schema is *required* for many-to-many fact-dimension relationships. A good example is customer and policy in insurance. A customer can have many policies and a policy can cover many customers.

The primary justification for using the star is performance and understandability. The simplicity of the star has been one of its attractions. While the star is generally considered to be the better performing structure, that is not always the case. In general, one should select a star as first choice where feasible. However, there are some conspicuous exceptions. The remainder of this response will address these situations.

First, some technologies such a MicroStrategy require a snowflake and others like Cognos require the star. This is significant.

Second, some queries naturally lend themselves to a breakdown into fact and dimension. Not all do. Where they do, a star is generally a better choice.

Third, there are some business requirements that just cannot be represented in a star. The relationship between customer and account in banking, and customer and policy in Insurance, cannot be represented in a pure star because the relationship across these is many-to-many. You really do not have any reasonable choice but to use a snowflake solution. There are many other examples of this. The world is not a star and cannot be force fit into it.

Fourth, a snowflake should be used wherever you need greater flexibility in the interrelationship across dimension levels and components. The main advantage of a snowflake is greater flexibility in the data.

Fifth, let us take the typical example of Order data in the DW. Dimensional designer would not bat an eyelash in collapsing the Order Header into the Order Item. However, consider this. Say there are 25 attributes common to the Order and that belong to the Order Header. You sell consumer products. A typical delivery can average 50 products. So you have 25 attributes with a ratio of 1:50. In this case, it would be grossly cumbersome to collapse the header data into the Line Item data as in a star. In a huge fact table you would be introducing a lot of redundancy more than say 2 billion rows in a fact table. By the way, the Walmart model, which is one of the most famous of all time, does not collapse Order Header into Order Item. However, if you are a video store, with few attributes describing the transaction, and an average ratio of 1:2, it would be best to collapse the two.

Sixth, take the example of changing dimensions. Say your dimension, Employee, consists of some data that does not change (or if it does you do not care, i.e., Type 1) and some data that does change (Type 2). Say also that there are some important relationships to the employee data that does not change (always getting its current value only), and not to the changeable data. The dimensional modeler would always collapse the two creating a Slowly Changing Dimension, Type 2. This means that the Type 1 is absorbed into the Type 2. In some cases I have worked on, it has caused more trouble than it was worth to collapse in this way. It was far better to split the dimension into Employee (type 1) and Employee History (type 2). Thereby, in such more complex history situations, a snowflake can be better.

Seventh, whether the star schema is more understandable than the snowflake is entire subjective. I have personally worked on several data warehouse where the user community complained that in the star, because everything was flattened out, they could not understand the hierarchy of the dimensions. This was particularly the case when the dimension had many columns.

Finally, it would be nice to quit the theorizing and run some tests. So I did. I took a data model with a wide customer dimension and ran it as a star and as a snowflake. The customer dimension had many attributes. We used about 150MM rows. I split the customer dimension into three tables, related 1:1:1. The result was that the snowflake performed faster. Why? Because with the wide dimension, the DBMS could fit fewer rows into a page. DBMSs read by pre-fetching data and with the wide rows it could pre-fetch less each time than with the skinnier rows. If you do this make sure you split the table based on data usage. Put data into each piece of the 1:1:1 that is used together.

What is the point of all this? I think it is unwise to pre-determine what is the best solution. A number of important factors come into play and these need to be considered. I have worked to provide some of that thought-process in this response.

Second Solution (SS2021):

E5.2

Star vs. Snowflake

DWH Presentation
L. Katzmaier & M. Marhofen

Structure

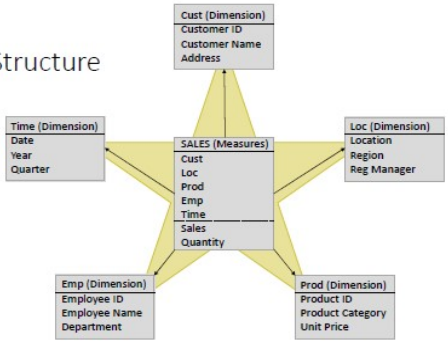
- STAR
  - Model
  - Structure
  - Join
  - Advantages & Disadvantages
- SNOWFLAKE
  - Model
  - Structure
  - Join
  - Advantages & Disadvantages
- COMPARISON

Star Schema

Star – Model

- Application area
  - multidimensional data structures in relational databases
  - Analytical applications
- Attempt to minimize number of tables
- Measures: express important relationships in a quantitatively measurable and condensed form
- Dimensions: enable different views of the measures

Star – Structure

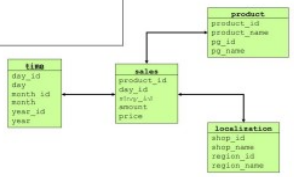


Star – Join

Source of the example: https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/ws_1718/vl_dwhdm/05_rolap_molap.pdf (page 11 and 14)

```
SELECT    L.shop_name, T.year, sum(amount*price)
FROM      sales S, product P, time T, localization L
WHERE     P.pg_name="Wasser" AND
          P.product_id = S.product_id AND
          T.day_id = S.day_id AND
          L.shop_id = S.shop_id
group by  L.shop_name, T.year
```

- 3 Joins
- Number of joins independent of aggregation paths in the request
- Number of joins increases linearly with the number of dimensions in the request.



Star – Advantages & Disadvantages

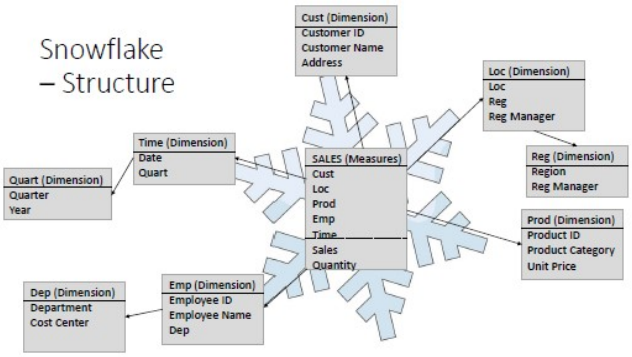
- Advantages
  - Intuitive data model
  - Simpler queries
  - Simplified Business Reporting Logic
  - Optimizes navigation
- Disadvantages
  - Poor response behavior with large dimension tables
  - Lack of flexibility
  - Data integrity
  - Large number of redundancies

Snowflake – Schema

Snowflake – Model

- Normalized version of star schema
  - Fact tables remain
  - Storing of dimensions in normal form
- Use cases:
  - Large amount of data
  - Many users with different scenarios

Snowflake – Structure



Snowflake – Join

- Relations are defined with foreign keys
- Often many join-operations are necessary
- Joins put more load on the performance
- Advantage due to non-redundancy might be lost

Snowflake – Advantages & Disadvantages

- Advantages
  - More structured data
  - Less redundancy, improved integrity
  - Needs less storage
  - Easier maintenance
  - Performance improvements due to less data being retrieved
- Disadvantages
  - Harder to design
  - More complex join-operations
  - More joins result in worse performance
  - Requires higher skills

	STAR	SNOWFLAKE
Database design	Simple	Complex
Hierarchies	Stores the hierarchies for the dimensions	Separated into different tables
Joins	Single join to create relation between dimension and fact table	Many joins required to fetch data
Dimension tables	Dimension tables surround fact tables	Every fact table surrounded by a dimension table, which is surrounded by more dimension tables
Data Redundancy	High-level	Low-level
Cube Processing	Fast	Might be slower (due to complex join)
Data in dimension tables	Single one contains aggregated data	Data split into different tables
Amount of retrieved data in query	All data is retrieved	Only the requested tables are retrieved

Thank you for your Attention

DWH Presentation
L. Katzmaier & M. Marhofen

Sources

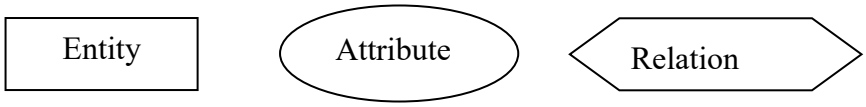
- https://www.datenbanken-verstehen.de/data-warehouse/data-warehouse-grundlagen/data-warehouse-datenmodell/star-schema/
- https://www.tecchannel.de/a/business-intelligence-teil-3-datenmodellierung-relationale-und-multidimensionale-modelle,1744994,10
- https://www.geeksforgeeks.org/star-schema-in-data-warehouse-modeling/
- https://hackr.io/blog/star-schema-in-data-warehouse
- https://www.ibm.com/support/knowledgecenter/en/SSEPEK_11.0.0/perf/src/tpc/db2z_starschemaaccess.html
- https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/ws_1718/vl_dwhdm/05_rolap_molap.pdf
- https://dbs.uni-leipzig.de/file/DWH_WS0910_Kap3_Modellierung.pdf
- https://www.vertabelo.com/blog/data-warehouse-modeling-the-snowflake-schema/?utm_source=dbweekly&utm_medium=email
- https://www.upsolver.com/blog/difference-between-star-schema-and-snowflake-schema
- Buch datenmodellierung (isbn: 3-89842-535-5)
- Sein Skript (S. 120)

Exercise E5.3: Build a Logical Data Model

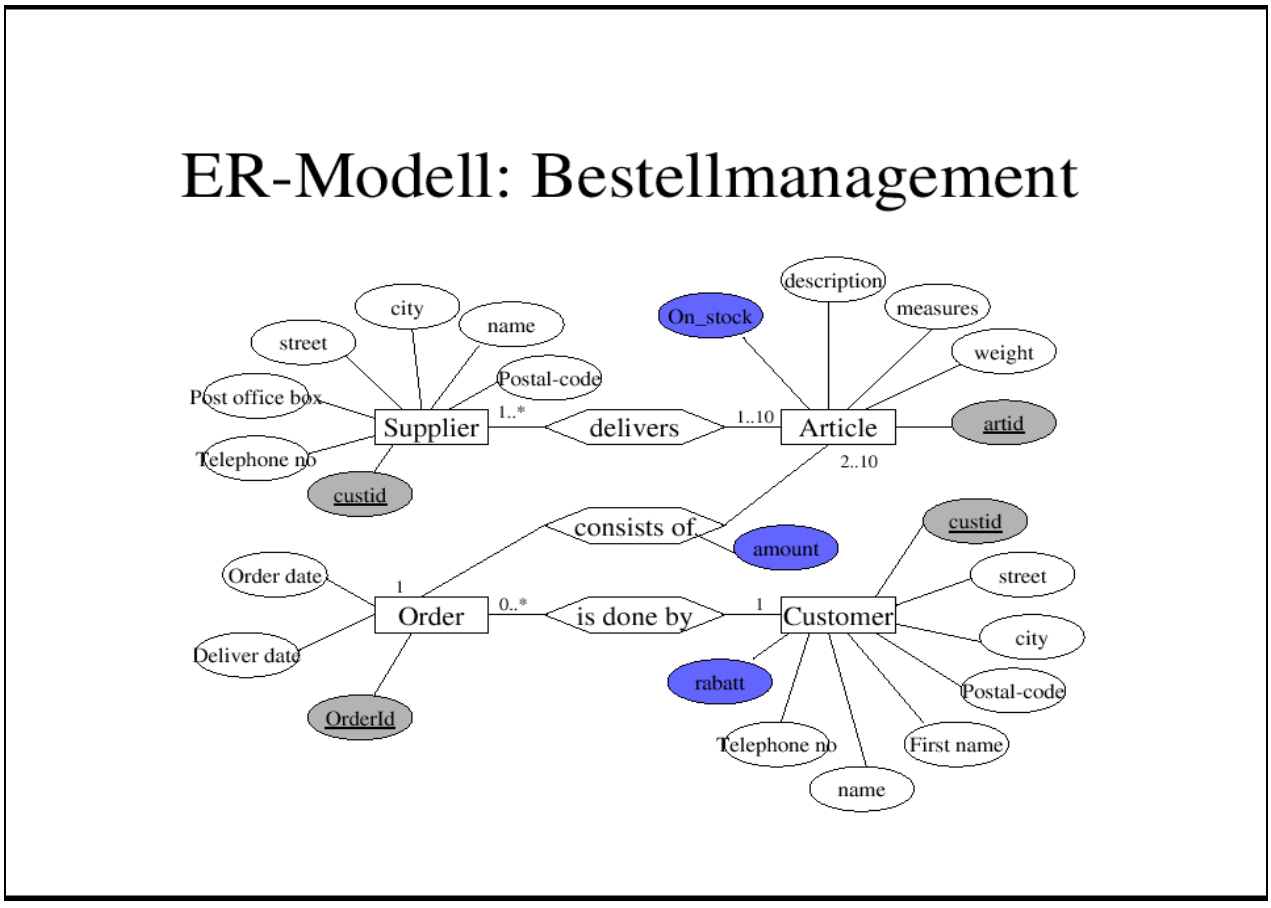
An enterprise wants to build up an ordering system.
The following objects should be administered by the new ordering system.

- **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

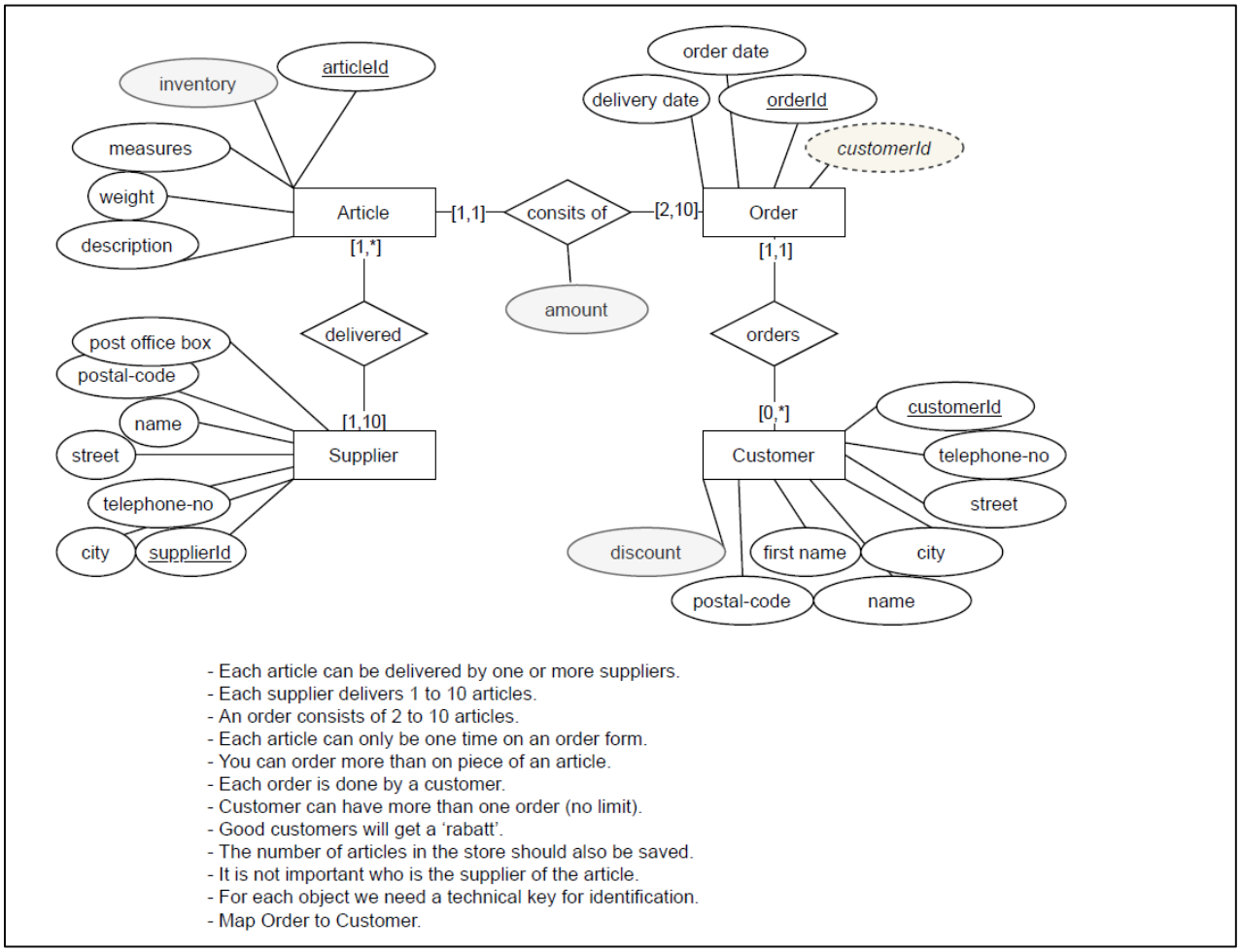
Conditions: Each article can be delivered by one or more suppliers. Each supplier delivers 1 to 10 articles. An order consists of 2 to 10 articles. Each article can only be one time on an order form. But you can order more than on piece of an article. Each order is done by a customer. Customer can have 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 a Logical ER model. Model the necessary objects and the relations between them. Define the attributes and the keys. Use the following notation:



First Solution:



Second Solution (M. Haug, A. Riess, WS2021):

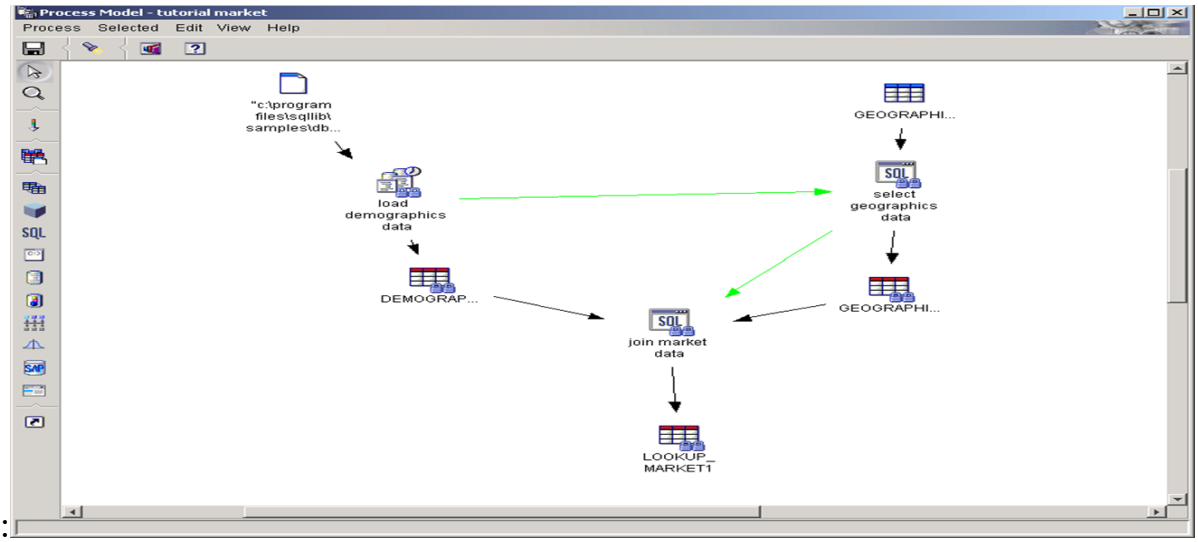


Exercises (+Solutions) to DHBW Lecture Intro2DWH – Chapter 6

Exercise E6.1: ETL: SQL Loading of a Lookup Table

Define the underlying SQL for the loading of Lookup_Market table:

:



Solution:

……

Exercise E6.2*: Discover and Prepare

In the lecture to this chapter we have seen 3 steps: “Discover”, “Prepare” and ”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. Server (IIS) – ProfileStage, or ????
2. **PREPARE**: HarteHanks-Trillium, Vality-Integrity, IBM Infosphere Inform. Server (IIS) – QualityStage, or ??????

Solution (SS2021):

EXAMPLE TOOLS

FOR DATA DISCOVERY & PREPARATION

STRUCTURE

Discover

Informatica
Talend

Prepare

Alteryx
Altair

DISCOVER

INFORMATICA

Made up of various applications

Data Catalog
Data Preparation
Cloud Data Integration
Cloud Data Quality
Cloud Data Ingestion
...

Flexible Pricing depending applications and the scale

INFORMATICA - ADVANTAGES

CLAIRE ML engine

Indendent

Can be used across any cloud ecosystem

Mass Ingestion Tool

Multiple data sources

Easy exploration using the data catalog

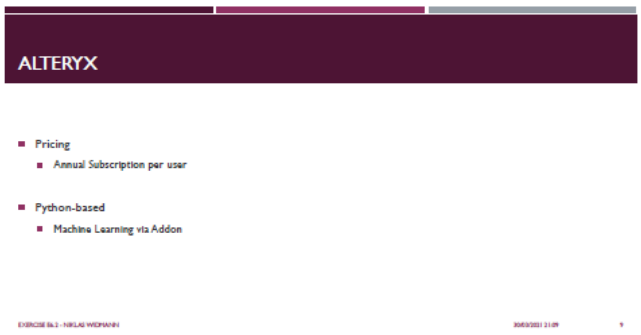
TALEND

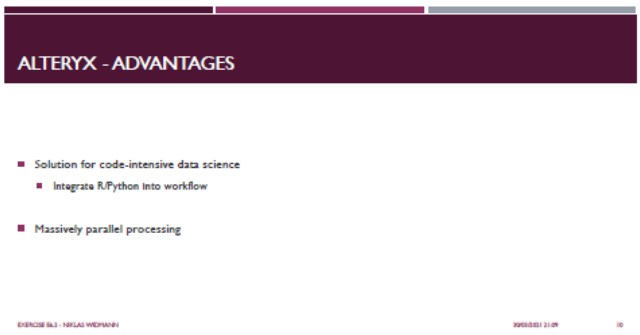
- Pricing
  - Free desktop version
  - Commercial cloud version
- Talend Data Fabric
  - API & Application Integration
  - Data Integration

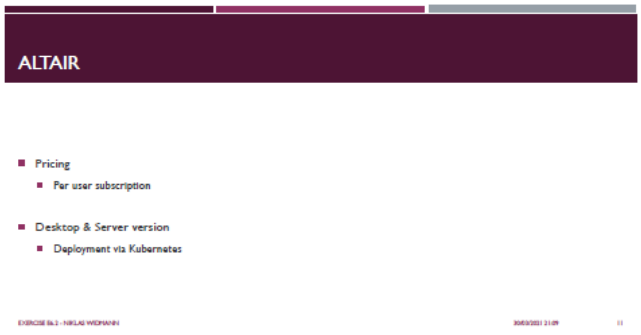
TALEND - ADVANTAGES

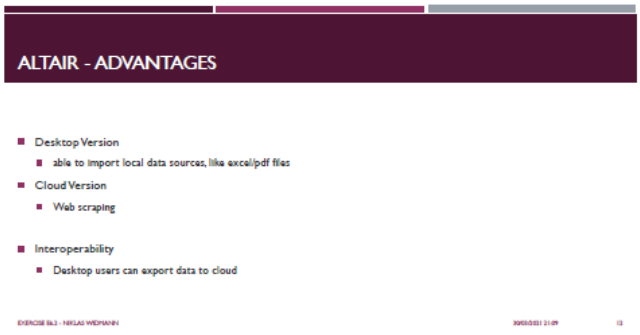
- Free Version
- Many connectivity options
- Apache Beam allows dataflows into
  - Any data storage
  - Cloud or on-premise storage

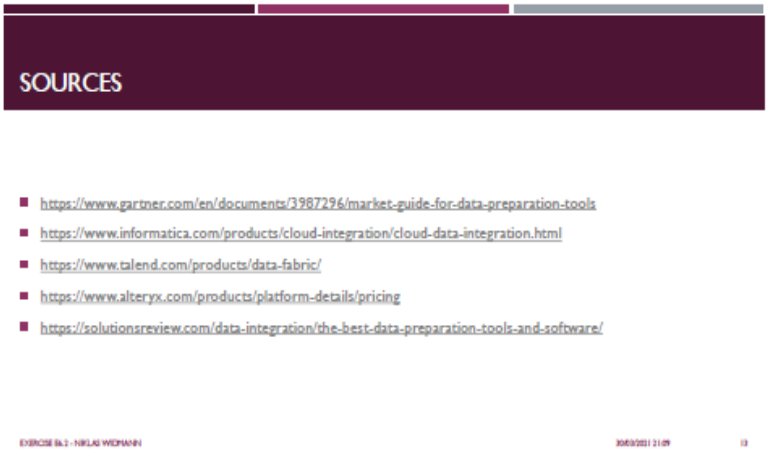
PREPARE



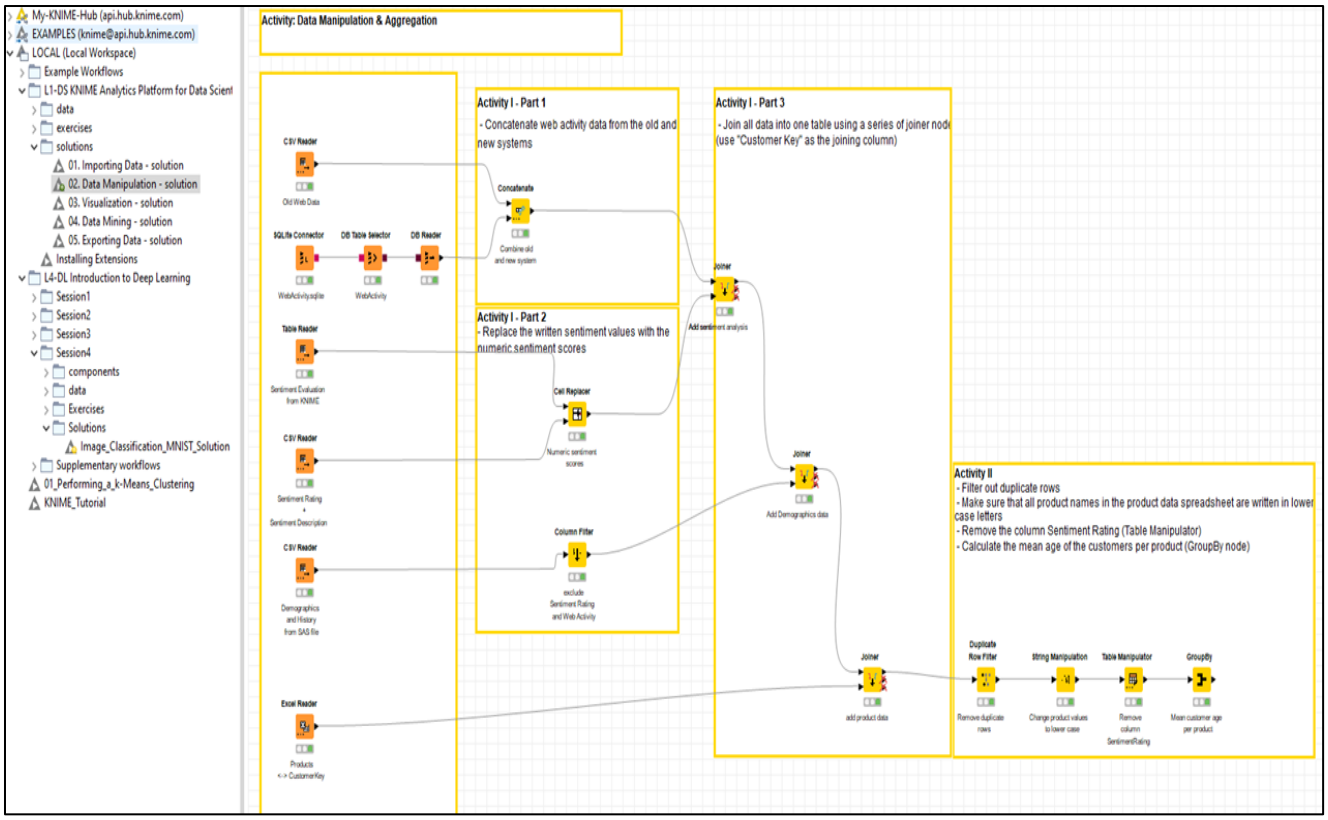








<u>Exercise E6.3</u>: Data Manipulation and Aggregation using KNIME Platform
Homework for 2 Persons: Rebuild the KNIME Workflow (use given solution) for Data Manipulation & Aggregation and give technical explanations to the solution steps.
<u>Hint:</u> Follow the instructions given in the KNIME workflow "KNIME Analytics Platform for Data Scientists – Basics (02. Data Manipulation -solution)" - see image below:



Solution: WS2021

Workflow « Data Manipulation and Aggregation » :

The Data Manipulation and Aggregation workflow analyses data on customers of investment products are analysed.
First, sample data from KNIME is read in (which can be found with the installation under Example Workflow- The Data- Misc). The data comes from different sources e.g. CSV, Excel and from a SQLite DB.
Reading: The top CSV Reader loads data to the First Web Activity from customers. Each Reader node needs a path to a file in the appropriate format. This can be customized under the configurations.
The SQLite Connector below establishes a connection to a SQLite file and with the DB Table Selector the table Web Activity is selected. The DB Reader loads
the data into a KNIME Table. The database and the CSV file provide data about customers and their First Web activity. The CSV file provides e.g. older customer data and in the database the database contains the newer data.
Activity 1 - Part 1: In the next node Concatenate the two data sources are merged in the column Customer Key, because both tables have this column. In the setting of the node is set to concatenate only the columns that have the same name. The node now contains the IDs of all customers.
Import: In the next part, a customer sentiment table is read in. The column Sentiment Analysis here consists of strings. This is to be mapped by integer values, which are mapped in the CSV file below. Each string value is then mapped to a number.
Activity 1 - Part 2: In the Cell Replacer node, the cells of a column can be replaced. The node needs an input table and a translation table. In the example here the upper

table is the input table and the CSV file provides the translation. The Cell Replacer adds according to the configuration, the Cell Replacer adds a new Sentiment Rating column with the appropriate integer values to the input table.
However, it can also be specified that the string values are replaced by the numbers in the same column without creating a new column.
Import: Next, customer data e.g. income, gender etc. is read in from a CSV file. file.
Activity 1 - Part 2: With the next node Column Filter one or more columns can be filtered out of a table. from a table. Here we want to exclude the columns Sentiment Rating and Web Activity should be excluded.
Import: Finally, an Excel table is read in with the Excel Reader. In it you can see which customer has purchased which investment product. In the next step, these individual data are merged.

In the 3rd part of the workflow, join operations take place. We know that with the help of this to merge different tables. For this purpose, KNIME offers the possibility to configure the Join- operation in KNIME (Knode --→ Config)
Joiner Settings: you can select one/several of 3 options (include in output):
- Matching rows (inner Join): only the rows are included in the common table, that are included in both original tables.
- Left unmatched rows (left outer join): adds additionally the columns from the left table that are not that are not present in the right table. In the right table the missing data is columns in the right table.
- Right unmatched rows (right outer join): adds the columns from the right table that are not present in the table that are not present in the left table. In the left table missing data will be columns in the left table.
Column selection: there is a possibility to select the desired columns (manually, with the RegEx or by the data type).
The first joiner merges the tables from Part 1 and Part 2, the second joiner uses the result of the first joiner. Joiner uses the result of the 1st Joiner and the data, which results after the filtering in the Part 2 are used. The 3rd Joiner composes its result from the data of the 2nd Joiner and from the Excel Reader, which includes the "Products" table.
After that the data flows into the next phase, where the received data can be manipulated again.
The "Duplicate Row Filter" node identifies duplicate rows. I can either remove all duplicate rows from the input table and keep only the unique and selected rows, or I can or provide the rows with additional information about their duplication status. In the configurations it is possible to specify in which columns it should search for duplicates.
Using the String Manipulation node you can perform different operations on strings. This can also be set in configurations.
The Table Manipulator node allows to perform multiple column transformations on any number of input tables, such as renaming, filtering, rearranging and type change of the input columns. If there is more than one input table, the node concatenates all input rows into a single result table. If the input tables contain the same row identifier, the node can either create a new row identifier or use the index of the input index of the input table to the original row identifier of the corresponding input table.
The last node "GroupBy" groups the rows of a table according to the unique values in the selected group columns. For each unique set of values of the selected group column, one row is created. The remaining columns are aggregated based on the specified aggregation settings.

Exercises (+Solutions) to DHBW Lecture Intro2DWH – Chapter 7

Exercise E7.1*: Compare 3 ETL Tools

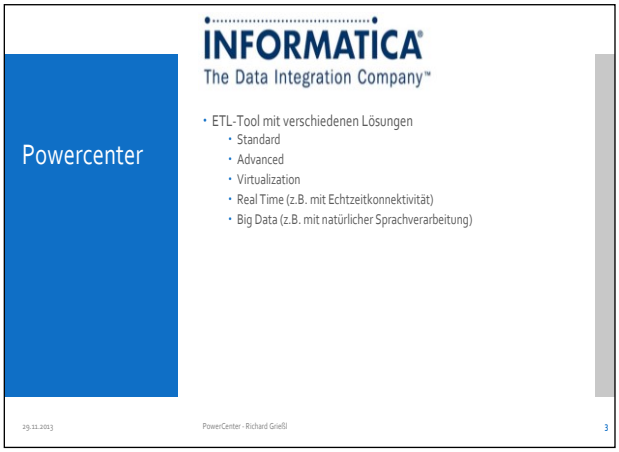
Show the Highlights and build a Strengthens/Weakness Diagram for the following three ETL Tools. Use the information from the internet:

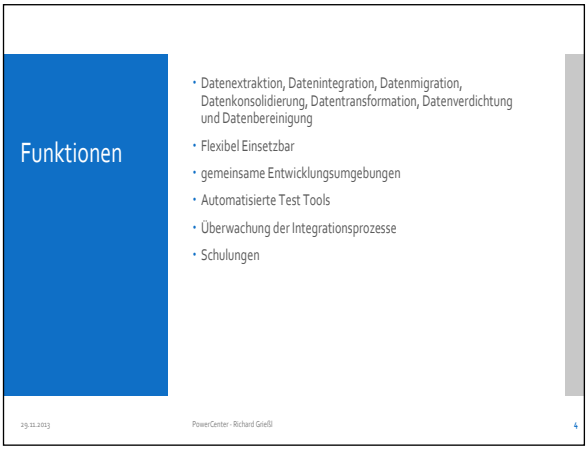
1. Informatica – PowerCenter --→ www.informatica.com
2. IBM - Infosphere Inform. Server - DataStage ---→ https://www.ibm.com/us-en/marketplace/datastage?loc=de-de
3. Oracle – Warehouse Builder (OWB) --→

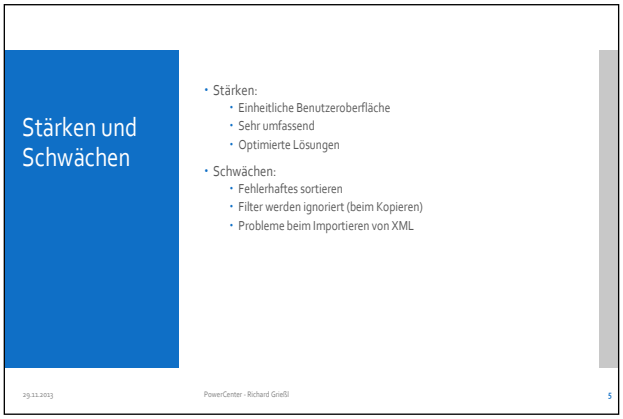
https://docs.oracle.com/cd/B28359_01/owb.111/b31278/concept_overview.htm#WBDOD10100

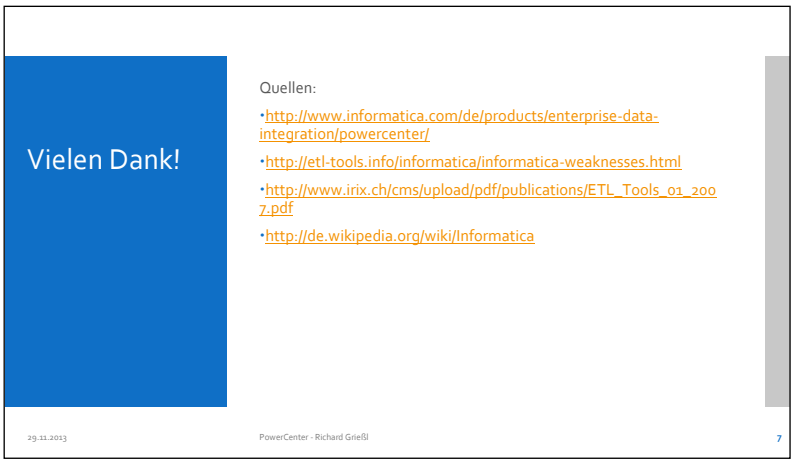
Show the three tools in competition to each other

Solution - Informatica:

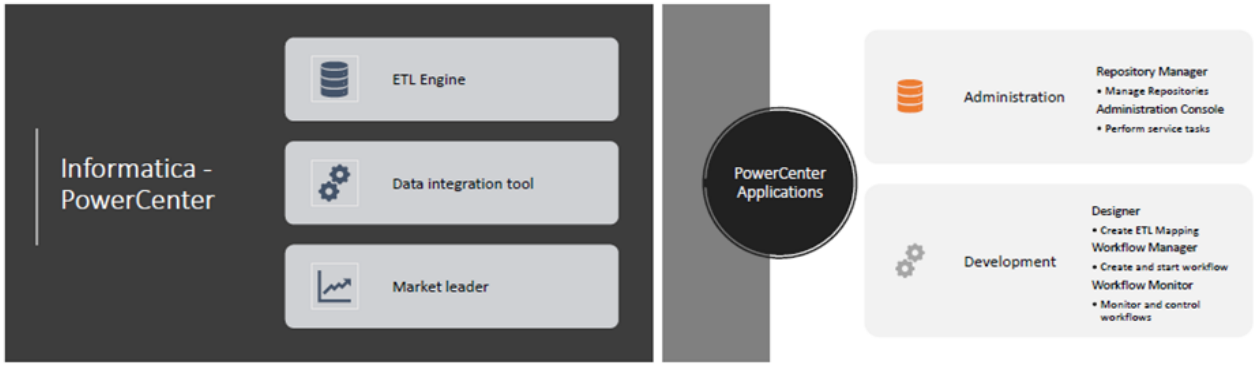








Second Solution (SS2021):



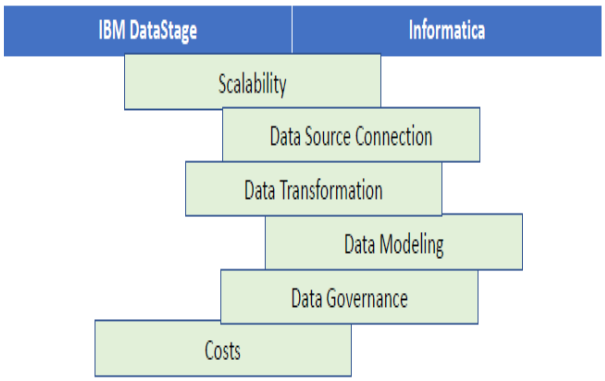
Informatica - PowerCenter

Pros	Cons
Data migration from multiple sources	Licensing cost - no open-source version
Very extensive	More in-depth training
Good performance even during processing of large amount of data	Multiple interfaces

IBM DataStage

Pros	Cons
Speed Workload Execution	Wrong Use of Algorithm costs more Time
Deploy Anywhere	Development must include Connectors
Modernize Data Integration	
Reduce Costs	
AI services	
In-Flight Data	

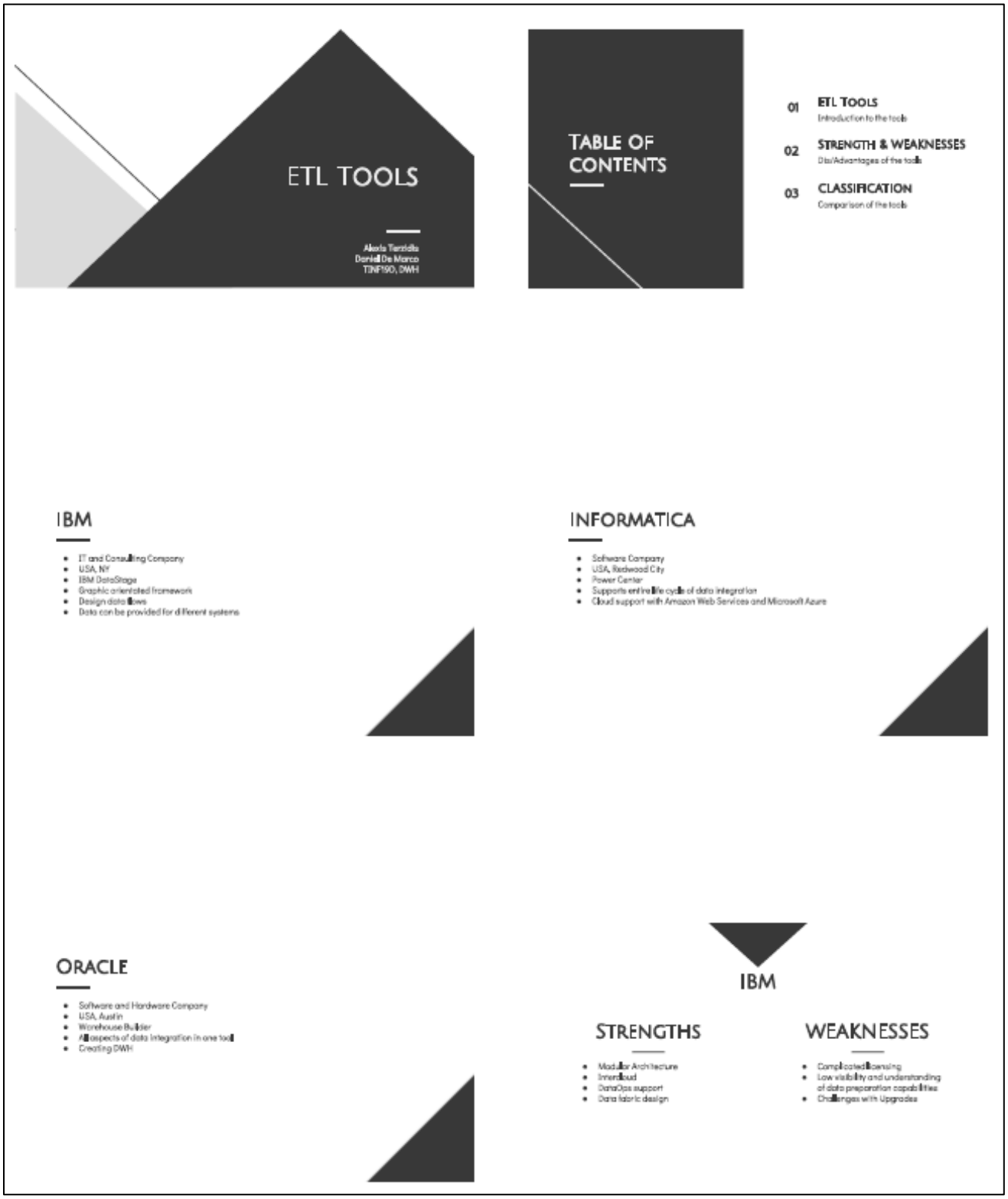
Informatica vs. IBM DataStage

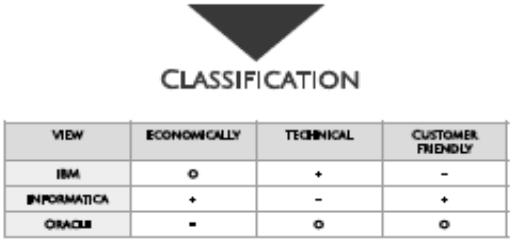


Sources

- https://www.youtube.com/watch?v=3scD3llibJA
- https://www.trustradius.com/products/informatica-powercenter/reviews?o=recent&qs=pros-and-cons
- https://docs.informatica.com/data-integration/powercenter/10-5/getting-started/preface.html
- https://www.ibm.com/products/infosphere-datastage?loc=de-de

Third Solution (WS2021):





Exercise E7.2: Demo of Datastage

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!):

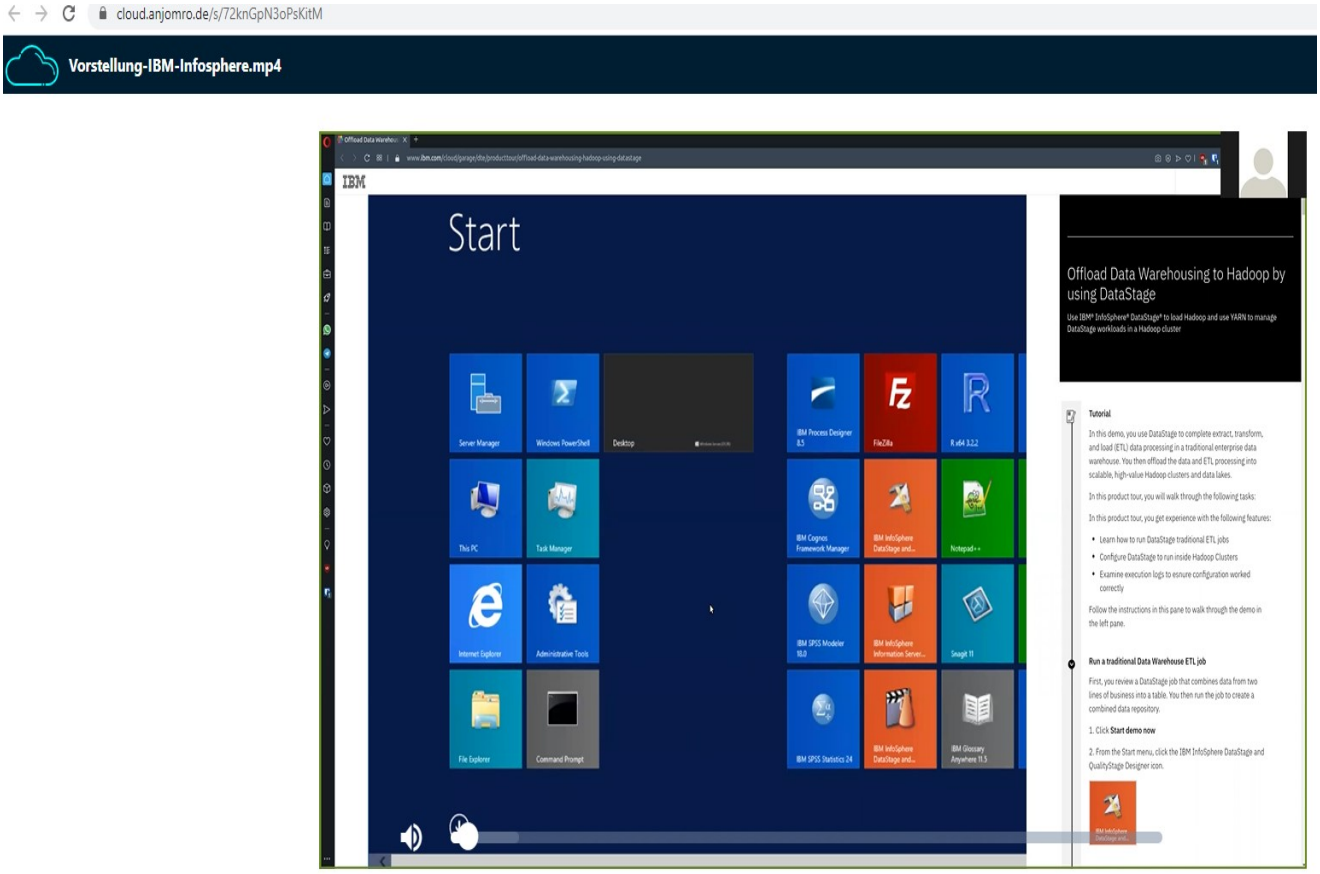
https://www.ibm.com/cloud/garage/dte/producttour/offload-data-warehousing-hadoop-using-datastage

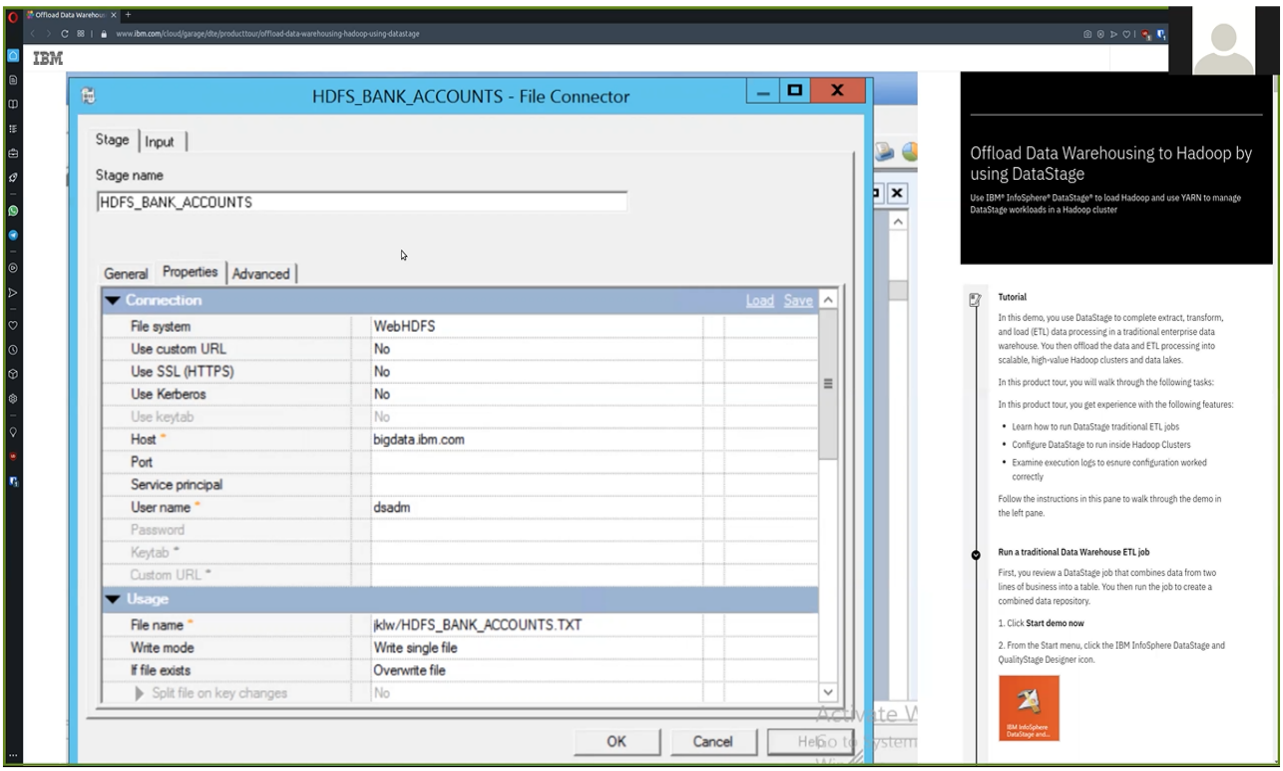
Explain each step in the demo with your own words….

Solution (SS2021):

See the execution of this demo in the IBM Cloud in the following video:
https://cloud.anjomro.de/s/72knGpN3oPsKitM

…





….

Exercise E7.3: Compare ETL and ELT Approach

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]

Solution (WS 2021):

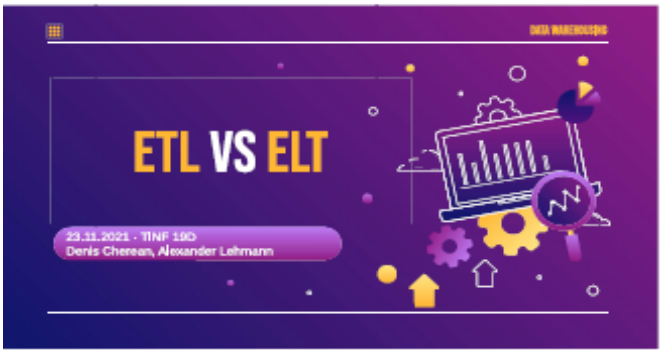
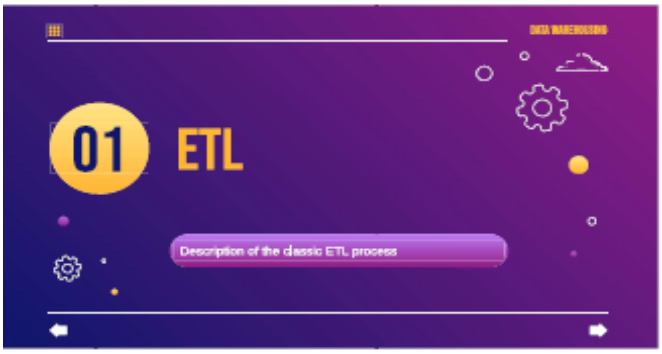

ETL VS ELT
23.11.2021 - TINF 19D
Denis Cherean, Alexander Lehmann


TABLE OF CONTENT
01
ETL
Description of the classic ETL process
02
ELT
Description of the new ELT process
03
COMPARISON, PROS & CONS
Comparison of ETL vs. ELT, pros and cons
04
CONCLUSION & OUTLOOK
A conclusion and outlook for the future

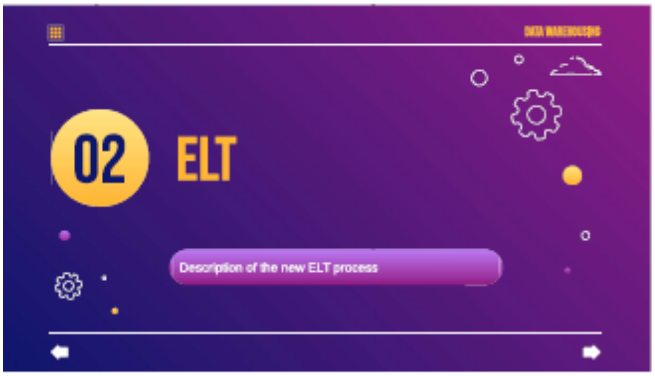

01
ETL
Description of the classic ETL process

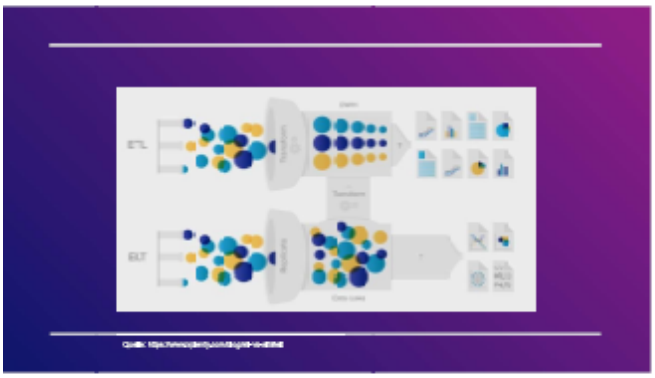

ETL
Raw data is:
extracted from third-party systems
cleaned / enriched and transformed by predefined processes
loaded into the target system (e.g. DWH)
Target:
Get rid of all (currently) irrelevant data
Keep smallest amount of required, structured data in target system
Time consuming process from extraction to loading
Analysis
of predefined questions
efficient and stable
Data protection oriented
transformation before loading
"Old school" technology
Mature tools and systems
Data engineers with know how available

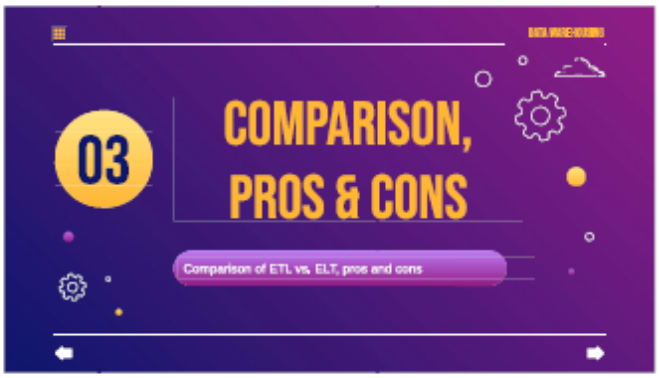

SO WHY
ELT?

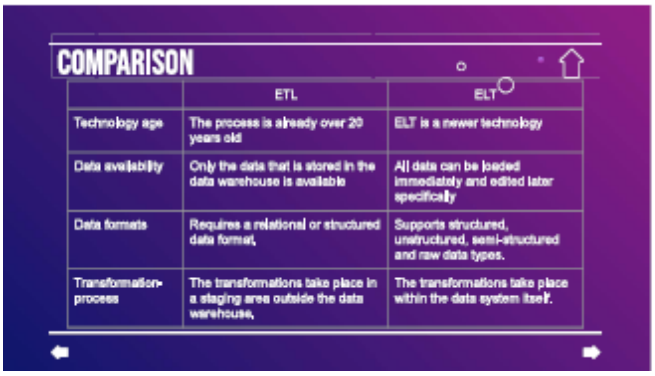

"Traditionally, data was extracted, transformed, then loaded – ETL, in short – into a data warehouse. For ETL, complex transformation pipelines were built at the data source. However, cloud data warehouses have finally made it cost-effective to store all of a company's data in a central location; we no longer need to transform data before we load it into a data warehouse. Transformation can be done when running analytics in a data warehouse."
Martin Casado, Andreessen Horowitz

……..

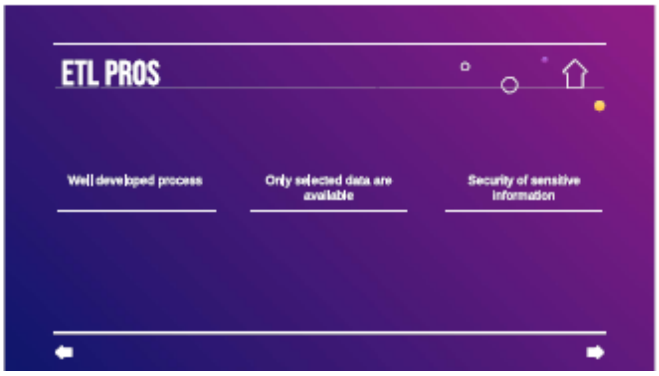


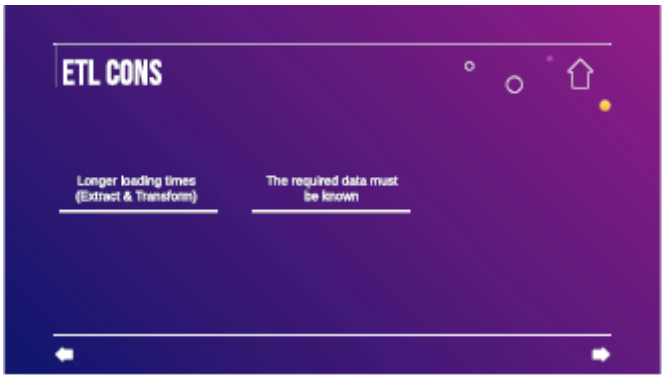


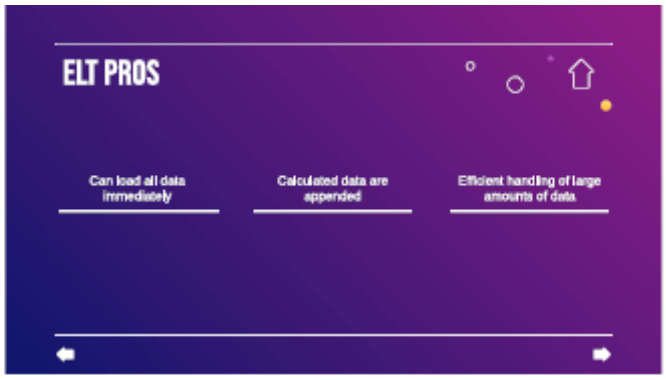


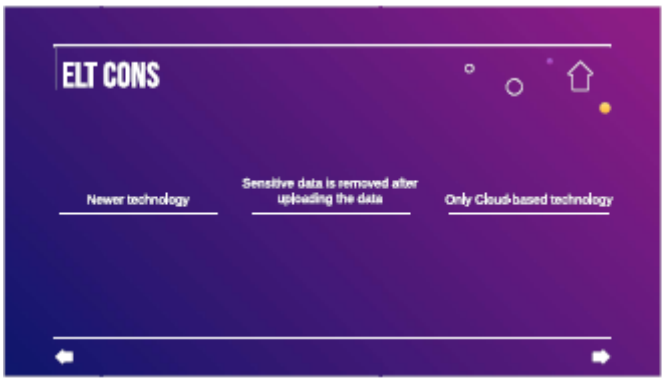


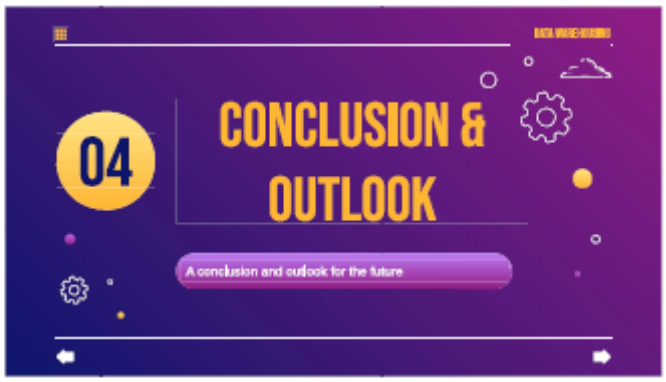
	ETL	ELT
Technology age	The process is already over 20 years old	ELT is a newer technology
Data availability	Only the data that is stored in the data warehouse is available	All data can be loaded immediately and edited later specifically
Data formats	Requires a relational or structured data format.	Supports structured, unstructured, semi-structured and raw data types.
Transformation-process	The transformations take place in a staging area outside the data warehouse.	The transformations take place within the data system itself.

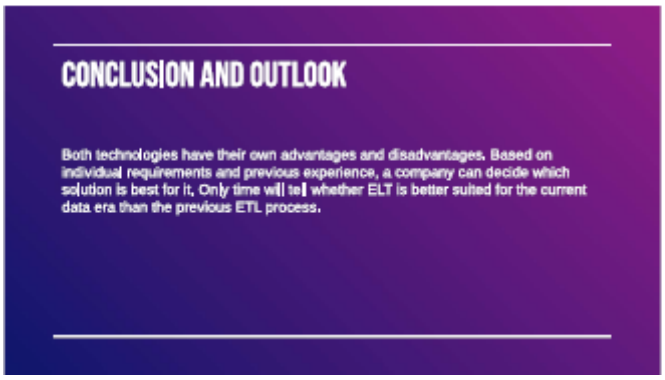


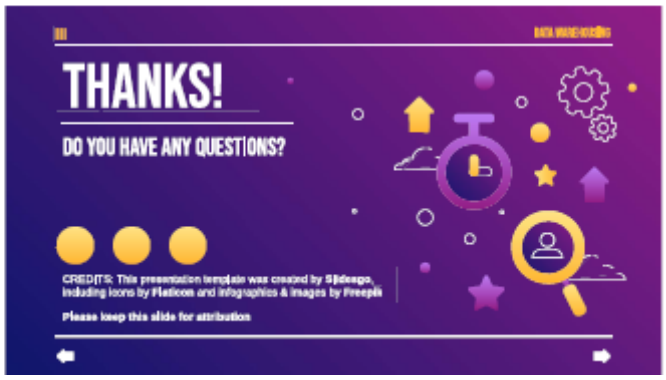






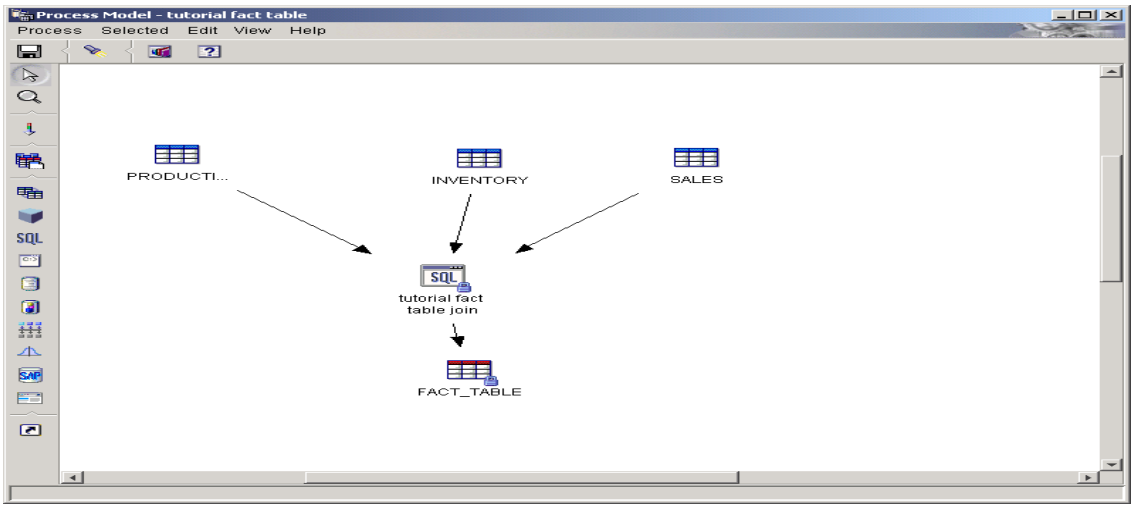




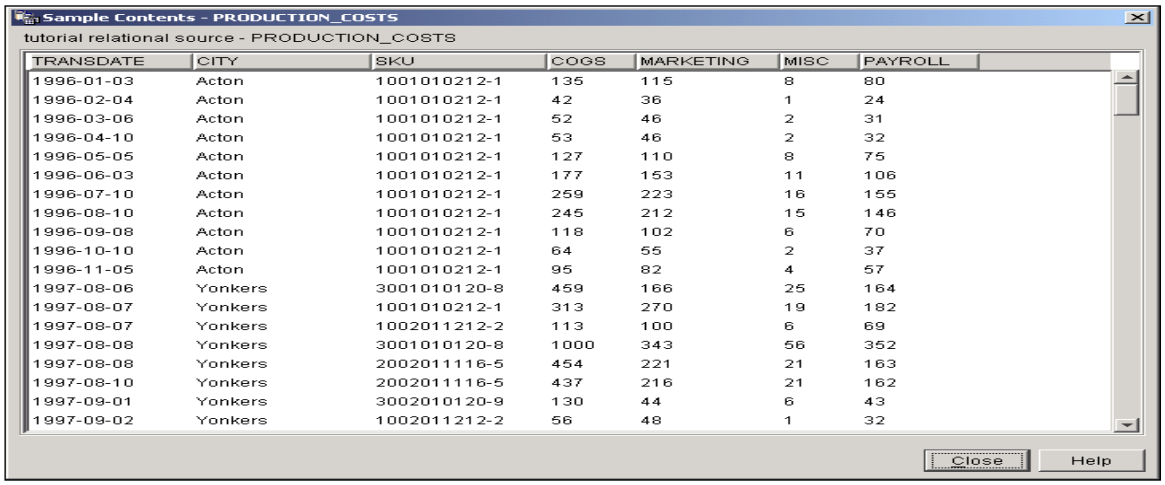


<u>Exercise E7.4:</u> ETL : SQL Loading of a Fact Table

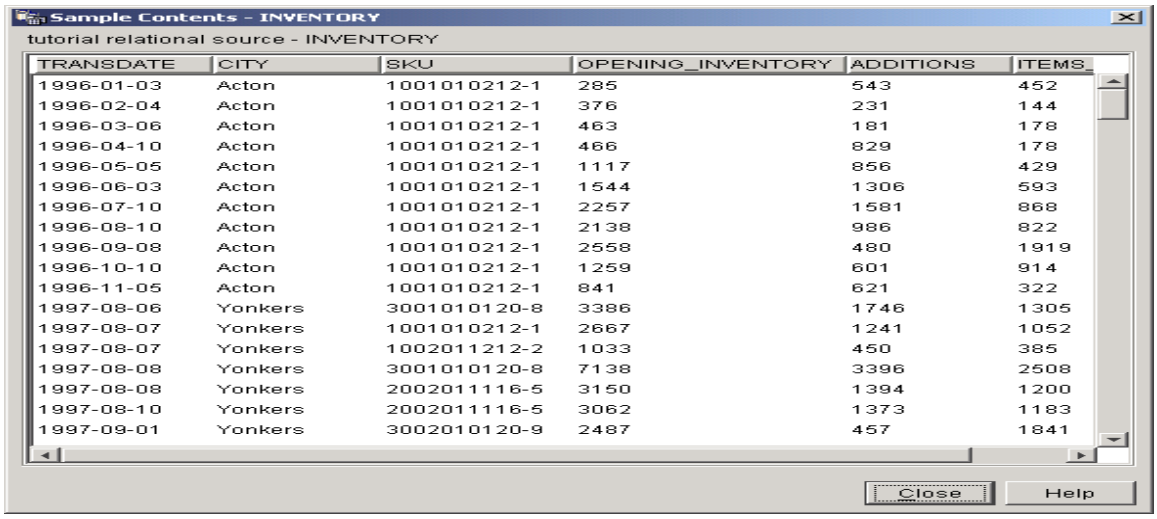
Define the underlying SQL for the following loading of the_ Fact “FACT-TABLE” from the three tables: “PRODUCTIO_COSTS”, “INVENTORY” & “SALES”.



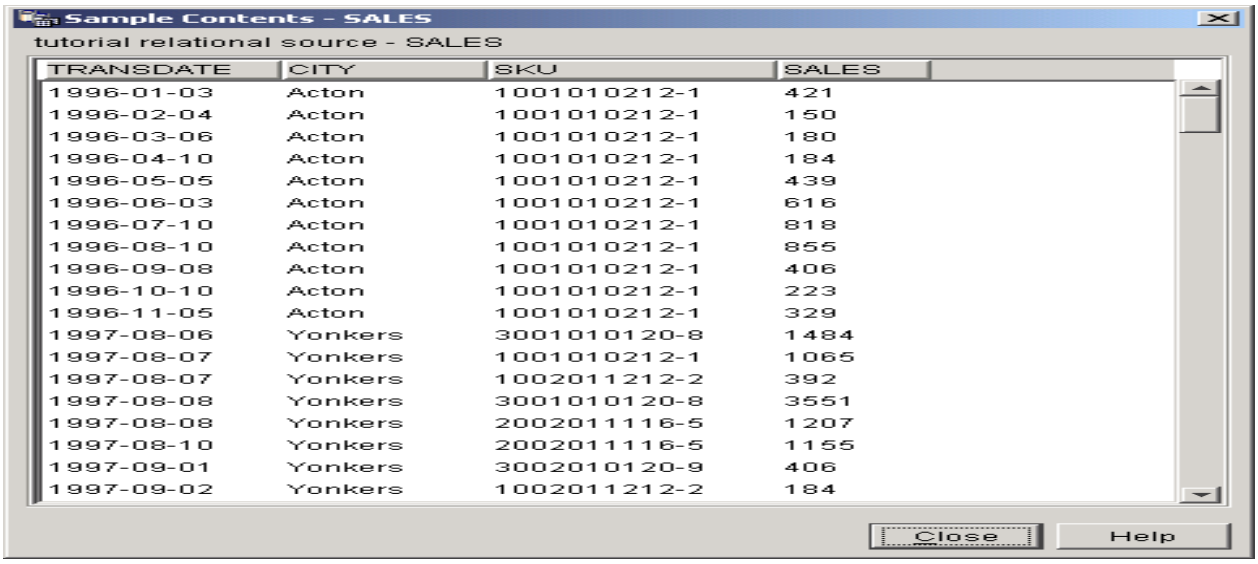
The content of the three input-tables are seen here:



TRANSDATE	CITY	SKU	COGS	MARKETING	MISC	PAYROLL
1996-01-03	Acton	1001010212-1	135	115	8	80
1996-02-04	Acton	1001010212-1	42	36	1	24
1996-03-06	Acton	1001010212-1	52	46	2	31
1996-04-10	Acton	1001010212-1	53	46	2	32
1996-05-05	Acton	1001010212-1	127	110	8	75
1996-06-03	Acton	1001010212-1	177	153	11	106
1996-07-10	Acton	1001010212-1	259	223	16	155
1996-08-10	Acton	1001010212-1	245	212	15	146
1996-09-08	Acton	1001010212-1	118	102	6	70
1996-10-10	Acton	1001010212-1	64	55	2	37
1996-11-05	Acton	1001010212-1	95	82	4	57
1997-08-06	Yonkers	3001010120-8	459	166	25	164
1997-08-07	Yonkers	1001010212-1	313	270	19	182
1997-08-07	Yonkers	1002011212-2	113	100	6	69
1997-08-08	Yonkers	3001010120-8	1000	343	56	352
1997-08-08	Yonkers	2002011116-5	454	221	21	163
1997-08-10	Yonkers	2002011116-5	437	216	21	162
1997-09-01	Yonkers	3002010120-9	130	44	6	43
1997-09-02	Yonkers	1002011212-2	56	48	1	32



TRANSDATE	CITY	SKU	OPENING_INVENTORY	ADDITIONS	ITEMS
1996-01-03	Acton	1001010212-1	285	543	452
1996-02-04	Acton	1001010212-1	376	231	144
1996-03-06	Acton	1001010212-1	463	181	178
1996-04-10	Acton	1001010212-1	466	829	178
1996-05-05	Acton	1001010212-1	1117	856	429
1996-06-03	Acton	1001010212-1	1544	1306	593
1996-07-10	Acton	1001010212-1	2257	1581	868
1996-08-10	Acton	1001010212-1	2138	986	822
1996-09-08	Acton	1001010212-1	2558	480	1919
1996-10-10	Acton	1001010212-1	1259	601	914
1996-11-05	Acton	1001010212-1	841	621	322
1997-08-06	Yonkers	3001010120-8	3386	1746	1305
1997-08-07	Yonkers	1001010212-1	2667	1241	1052
1997-08-07	Yonkers	1002011212-2	1033	450	385
1997-08-08	Yonkers	3001010120-8	7138	3396	2508
1997-08-08	Yonkers	2002011116-5	3150	1394	1200
1997-08-10	Yonkers	2002011116-5	3062	1373	1183
1997-09-01	Yonkers	3002010120-9	2487	457	1841



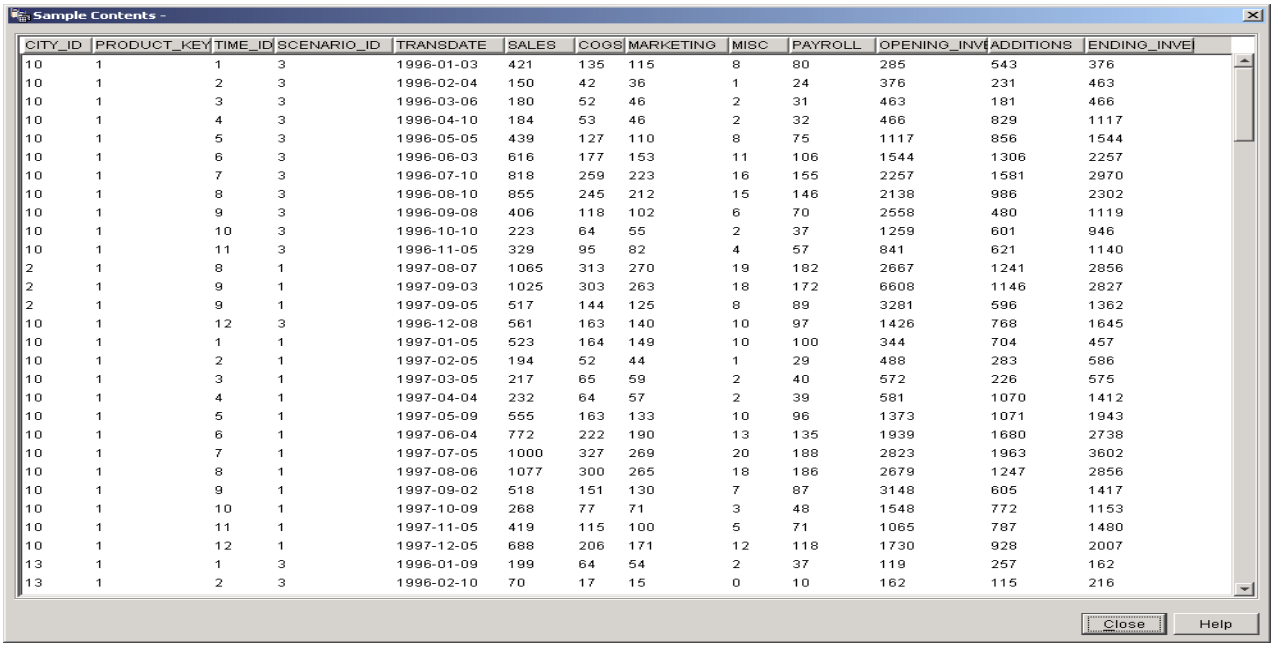
Sample Contents – SALES

tutorial relational source - SALES

TRANSDATE	CITY	SKU	SALES
1996-01-03	Acton	1001010212-1	421
1996-02-04	Acton	1001010212-1	150
1996-03-06	Acton	1001010212-1	180
1996-04-10	Acton	1001010212-1	184
1996-05-05	Acton	1001010212-1	439
1996-06-03	Acton	1001010212-1	616
1996-07-10	Acton	1001010212-1	818
1996-08-10	Acton	1001010212-1	855
1996-09-08	Acton	1001010212-1	406
1996-10-10	Acton	1001010212-1	223
1996-11-05	Acton	1001010212-1	329
1997-08-06	Yonkers	3001010120-8	1484
1997-08-07	Yonkers	1001010212-1	1065
1997-08-07	Yonkers	1002011212-2	392
1997-08-08	Yonkers	3001010120-8	3551
1997-08-08	Yonkers	2002011116-5	1207
1997-08-10	Yonkers	2002011116-5	1155
1997-09-01	Yonkers	3002010120-9	406
1997-09-02	Yonkers	1002011212-2	184

Close Help

write a SQL script, s.t. you get the following content of the target table:



Sample Contents -

CITY_ID	PRODUCT_KEY	TIME_ID	SCENARIO_ID	TRANSDATE	SALES	COGS	MARKETING	MISC	PAYROLL	OPENING_INVE	ADDITIONS	ENDING_INVE
10	1	1	3	1996-01-03	421	135	115	8	80	285	543	376
10	1	2	3	1996-02-04	150	42	36	1	24	376	231	463
10	1	3	3	1996-03-06	180	52	46	2	31	463	181	466
10	1	4	3	1996-04-10	184	53	46	2	32	466	829	1117
10	1	5	3	1996-05-05	439	127	110	8	75	1117	856	1544
10	1	6	3	1996-06-03	616	177	153	11	106	1544	1306	2257
10	1	7	3	1996-07-10	818	259	223	16	155	2257	1581	2970
10	1	8	3	1996-08-10	855	245	212	15	146	2138	986	2302
10	1	9	3	1996-09-08	406	118	102	6	70	2558	480	1119
10	1	10	3	1996-10-10	223	64	55	2	37	1259	601	946
10	1	11	3	1996-11-05	329	95	82	4	57	841	621	1140
2	1	8	1	1997-08-07	1065	313	270	19	182	2667	1241	2856
2	1	9	1	1997-09-03	1025	303	263	18	172	6608	1146	2827
2	1	9	1	1997-09-05	517	144	125	8	89	3281	596	1362
10	1	12	3	1996-12-08	561	163	140	10	97	1426	768	1645
10	1	1	1	1997-01-05	523	164	149	10	100	344	704	457
10	1	2	1	1997-02-05	194	52	44	1	29	488	283	586
10	1	3	1	1997-03-05	217	65	59	2	40	572	226	575
10	1	4	1	1997-04-04	232	64	57	2	39	581	1070	1412
10	1	5	1	1997-05-09	555	163	133	10	96	1373	1071	1943
10	1	6	1	1997-06-04	772	222	190	13	135	1939	1680	2738
10	1	7	1	1997-07-05	1000	327	269	20	188	2823	1963	3602
10	1	8	1	1997-08-06	1077	300	265	18	186	2679	1247	2856
10	1	9	1	1997-09-02	518	151	130	7	87	3148	605	1417
10	1	10	1	1997-10-09	268	77	71	3	48	1548	772	1153
10	1	11	1	1997-11-05	419	115	100	5	71	1065	787	1480
10	1	12	1	1997-12-05	688	206	171	12	118	1730	928	2007
13	1	1	3	1996-01-09	199	64	54	2	37	119	257	162
13	1	2	3	1996-02-10	70	17	15	0	10	162	115	216

Close Help

We see the following conditions:

1. Map the name of the cities in the sources to a number 1 – 100, define this as **City_Id**
2. Define last digit of SKU in SALES as **Product_Key**
3. Define them Month of Transdate as **Time _Id (**range:01 –12)
4. Def. **Scenario_Id** with cases (Year of Transdate)=1997 as 1,(…Transdate)=1996 as 3 , else 2
5. Fill all columns of target table with the same columns of sources
6. Define new column: **Ending_Inventory** = (Opening_Inv. + Additions) -Items_Sold

Solution:

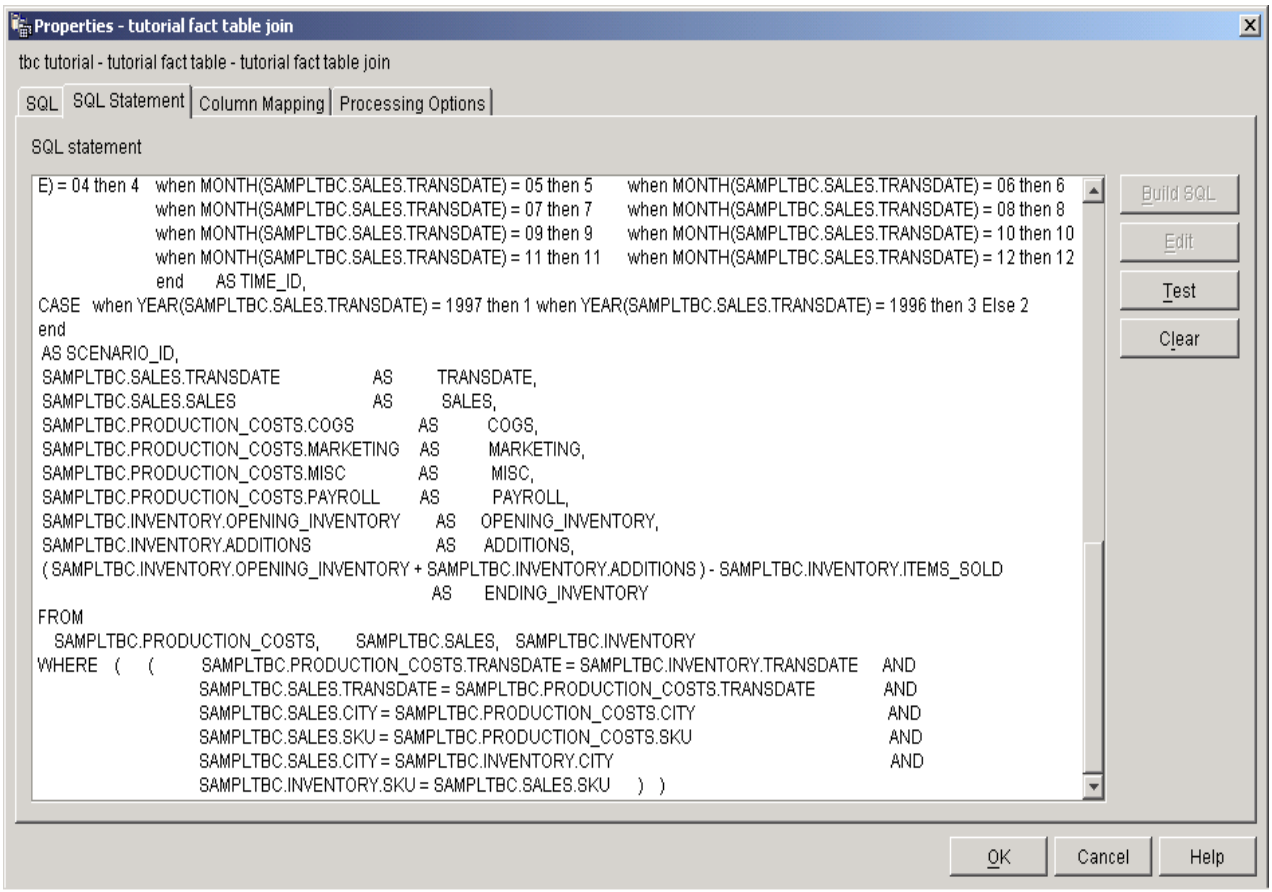
Select
Case SAMPLTBC.sales.city

When 'Manhattan' **then** 1
…..
When 'Maui' **then** 100
End
As City_Id

Substr (SAMPLTBC.sales.SKU,12,1) **as** Product_Key

Case
When Month(SAMPLTBC.sales.transdate) = 01 **then** 1
….
When Month(SAMPLTBC.sales.transdate) = 12 **then** 12
End
As Time_Id

Case
When Year(SAMPLTBC.sales.transdate) = 1997 **then** 1
When Year(SAMPLTBC.sales.transdate) = 1996 **then** 3
Else 2
End
As Scenario_Id
….
See screenshot:



Exercises (+Solutions) to DHBW Lecture Intro2DWH – Chapter 8

Exercise E8.1: Compare MOLAP to ROLAP

Find and define the Benefits & Drawbacks of
- MOLAP
- ROLAP

Systems
Use the information of the lesson or use your own experience

First Solution:

Criteria	ROLAP	MOLAP
Data volume	+ > 50 GB possible: low expansion factor (low aggregation rate)	- Not > 50 GB: expansion factor too big (high aggregation rate)
Dimensions	+ > 10 possible (depends only on DBMS)	- Bad performance for > 10 (due to high aggregation rate)
Query Performance	(+ When querying single tables) - When joining many tables	+ When using high aggregated data (- when using low aggregated data)
Update flexibility	+ Update during operation possible + Fast and flexible	- Cube has to be rebuilt completely each time (partly correct, in depends on calculation rules) - Operation has to be stopped
Query complexity	+ Complex, dynamic queries possible (impact on query performance)	- Only standard queries, that the cube is built for, possible (but combinations are possible)
Usability	- Not intuitive: SQL knowledge necessary	+ Intuitive, easy to handle, no special knowledge required
Price	+ Cheaper; simpler SQL-based front-ends sufficient (but more performance needed)	- Expensive; costly front-end tool necessary

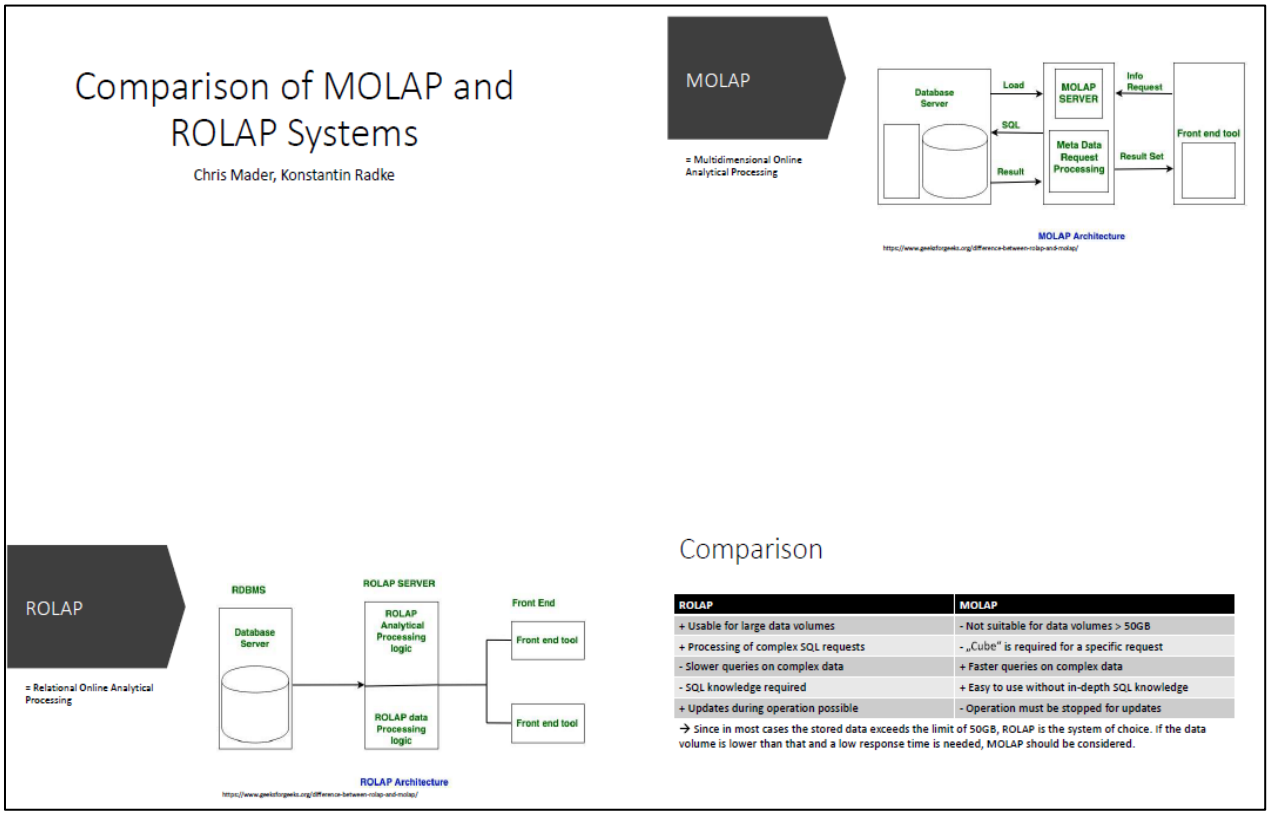
➔ ROLAP is for many criteria superior to MOLAP. As most data marts today are bigger than 50 GB, ROLAP is many cases better choice due to performance and storage reasons.

Second Solution:

MOLAP	ROLAP

-	Erstellen der Cubes aufwendig	+	Komplexere Anfragen möglich (Verwendung von SQL)
+	Schnelle Queries	-	Anwender muss SQL-Kenntnisse haben
-	Zugriff nur auf Daten des Cubes	-	Queries dauern länger, da komplexer
+	Auf Problemstellung angepasste Anfragen möglich	+	Zugriff auf alle Daten in DB
-	Bei Update müssen Cubes neu erstellt werden	+	Update auch während Operationen möglich
-	Nur bei Cube-Erstellung definierte Anfragen möglich		

Third Solution (SS2021):



Exercise E8.2*: Compare 3 Classical Analytics Tools

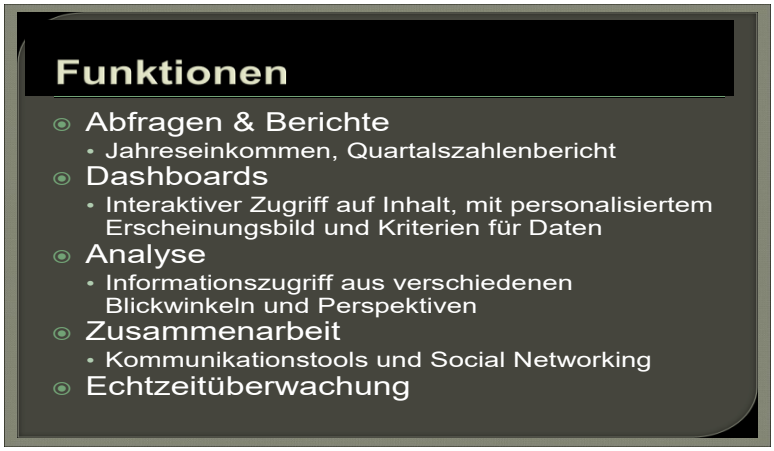
Show the Highlights and build a Strengthens/Weakness Diagram for the following three Reporting Tools.Use the information from the internet:

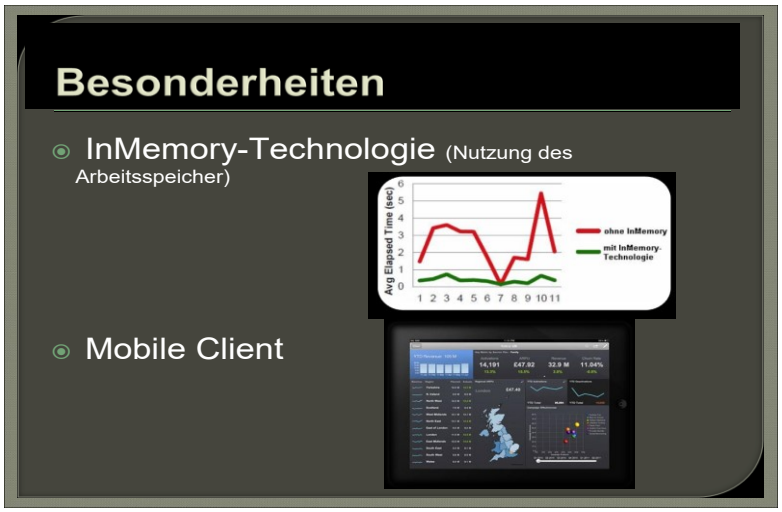
1. **MicroStrategy** ---→ www.MicroStrategy.com

2. **BusinessObjects** ----→ www.BusinessObjects.com
3. **Cognos** --→ www.Cognos.com

Show the three tools in competition to each other.

Solution: ***Presentation of Cognos:***





Second Solution (SS2021):

3 REPORTING TOOLS

JÖRN FRERICHS, DAVID SCHAU

WHAT IS A REPORTING TOOL

- Creating basic, medium and complex Reports
- Uses data from data warehouse
- Filters data
- Decision giver

STRUCTURE

- What is a Reporting Tool
- Microstrategy
- Cognos
- Business-Objects
- Sources

Product	Microstategy	Cognos	Business-Objects
Company	MicroStrategy Incorporated	IBM	SAP
Price	$600 - $1200 per user per month	$15 - $70 per user per month	Starting at $14000 per year
Support	Email, Phone, Live-Support, Training	Email, Phone, Training, Tickets	Email, Phone, Live-Support, Training, Tickets
Supported Languages	English	English, Chinese, German, Spanish, French, Italian	English, Chinese, German, Japanese, Spanish, French, Russian, Italian, Dutch, Polish, Turkish, Swedish
Target Client	Medium, Large Enterprises	Medium, Large Enterprises	Small, Medium Enterprises
Available Devices	Windows, Linux, Mac, Android, **Web-Based**	Windows, Linux, Mac, **Web-Based**	Windows, Mac, **Web-Based**

FEATURES OF MICROSTRATEGY

- Advanced and predictive analytics
- Cloud
- High-performance business intelligence
- Big data solutions
- Software as a service (SaaS)
- Real-time WYSIWYG report design
- Scorecards and dashboards

FEATURES OF BUSINESS-OBJECTS

- Enterprise Business Intelligence Reporting System
- Ad Hoc Querying and Reporting
- Data Visualization and Analytics Applications
- Self-Service Features
- Enterprise-Wide Sharing
- Role-Based Dashboards
- Microsoft Office Integration
- Real-Time Analytics
- Large-Scale Data Analysis

FEATURES OF COGNOS

Devided into 3 segments

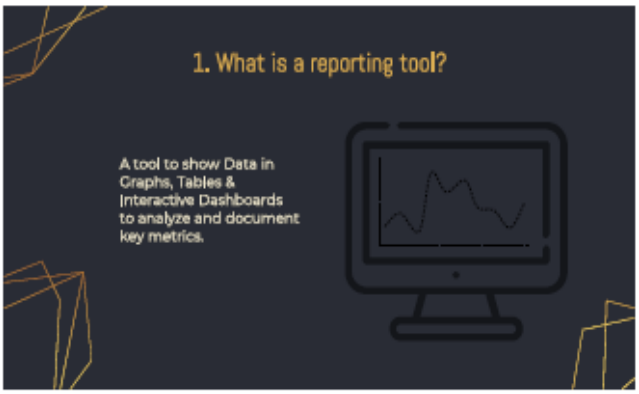
- Analysis Studio
- Query Studio
- Report Studio

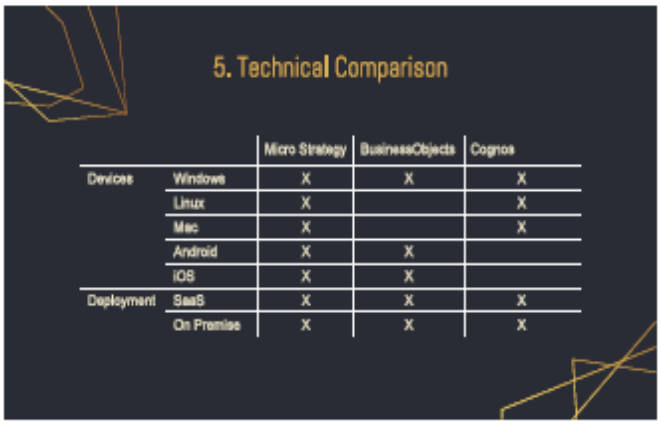
- Data protected with layers of permissions, authentication, and history
- Controls to protect data whether you're creating one report for many or many are creating one report
- Dashboards created using drag and drop on mobile device or desktop
- Smart search
- Scheduling and alerts
- Interactive content available online or offline
- Data models can be automatically generated based on keywords

SOURCES

- Financesonline.com (2021): Compare IBM Cognos vs MicroStrategy 2021 | FinancesOnline. Online verfügbar unter https://comparisons.financesonline.com/ibm-cognos-vs-microstrategy, zuletzt aktualisiert am 13.04.2021, zuletzt geprüft am 13.04.2021.
- SAP BusinessObjects Business Intelligence Reviews and Pricing - 2021 (2021). Online verfügbar unter https://www.capterra.com/p/92075/SAP-BusinessObjects/, zuletzt aktualisiert am 13.04.2021, zuletzt geprüft am 13.04.2021.
- dummies (2021): Querying and Reporting Tools for Data Warehousing - dummies. Online verfügbar unter https://www.dummies.com/programming/big-data/engineering/querying-and-reporting-tools-for-data-warehousing/, zuletzt aktualisiert am 10.04.2021, zuletzt geprüft am 13.04.2021.
- Data Warehouse - ETL & Reporting Tools - Tutorialspoint (2021). Online verfügbar unter https://www.tutorialspoint.com/cognos/data_warehouse_etl_and_reporting_tools.htm, zuletzt aktualisiert am 10.04.2021, zuletzt geprüft am 13.04.2021.

Third Solution (WS2021):





		Micro Strategy	BusinessObjects	Cognos
Devices	Windows	X	X	X
	Linux	X		X
	Mac	X		X
	Android	X	X	
	iOS	X	X	
Deployment	SaaS	X	X	X
	On Premise	X	X	X

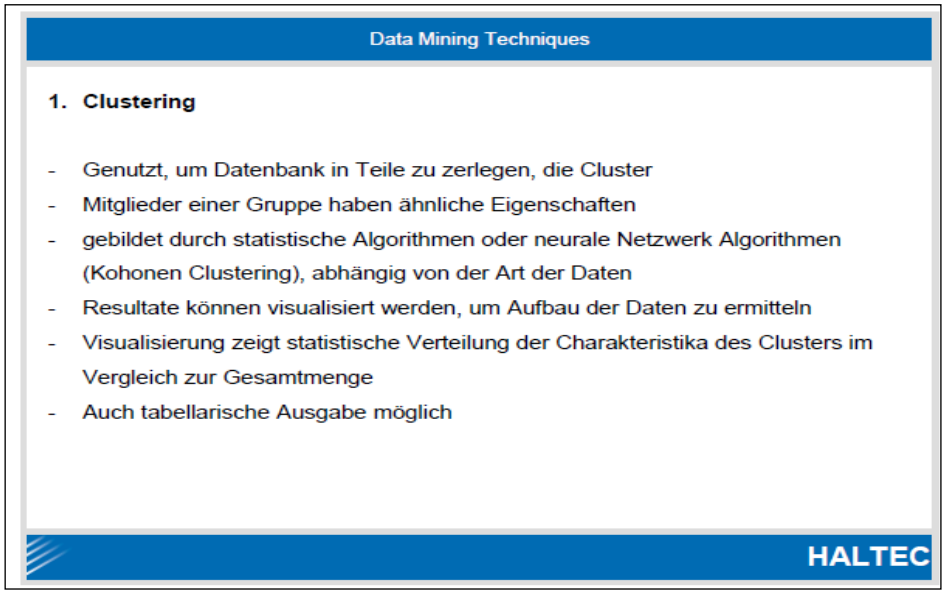
Exercises (+Solutions) to DHBW Lecture Intro2DWH – Chapter 9

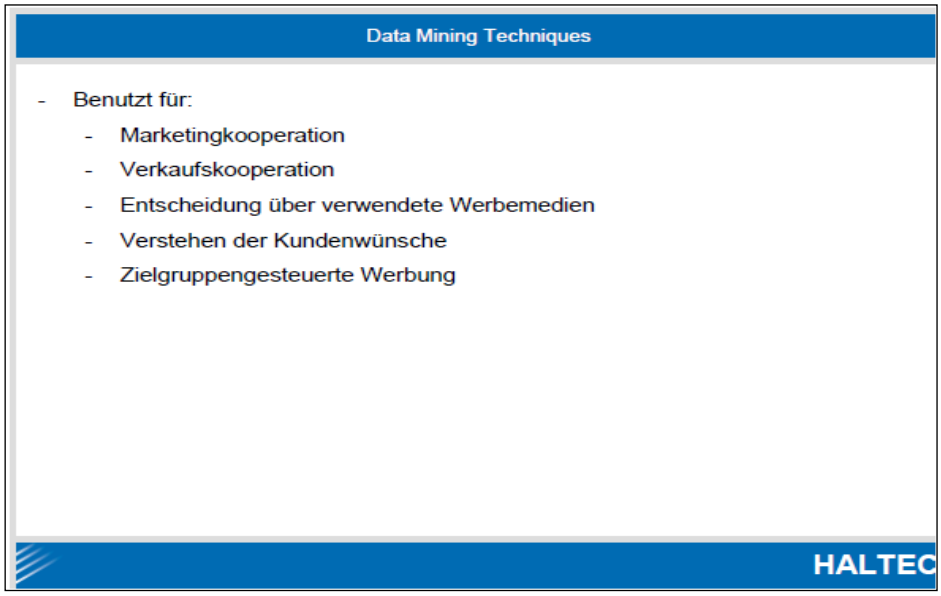
Exercise E9.1: Three Data Mining Methods (Part1)

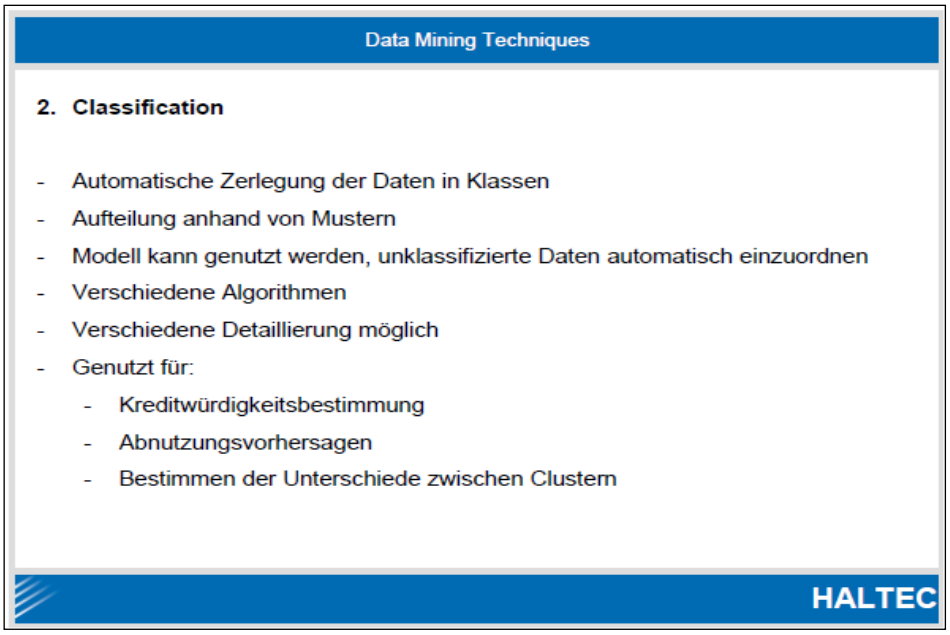
Task: Describe the following Data Mining techniques. Search this information in the internet, i.e. Wikipedia or other knowledge portals:

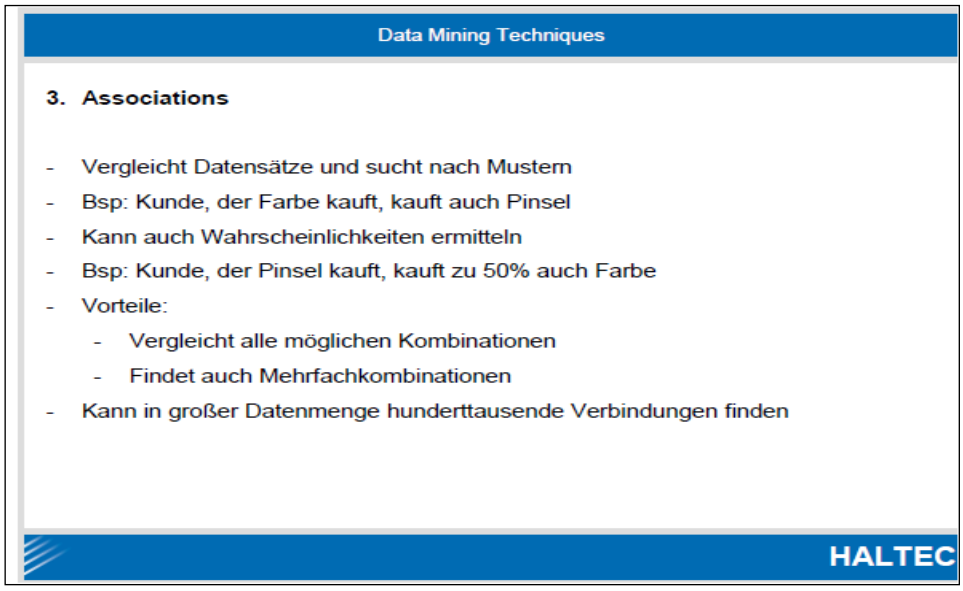
- **Clustering**
- **Classification**
- **Associations**

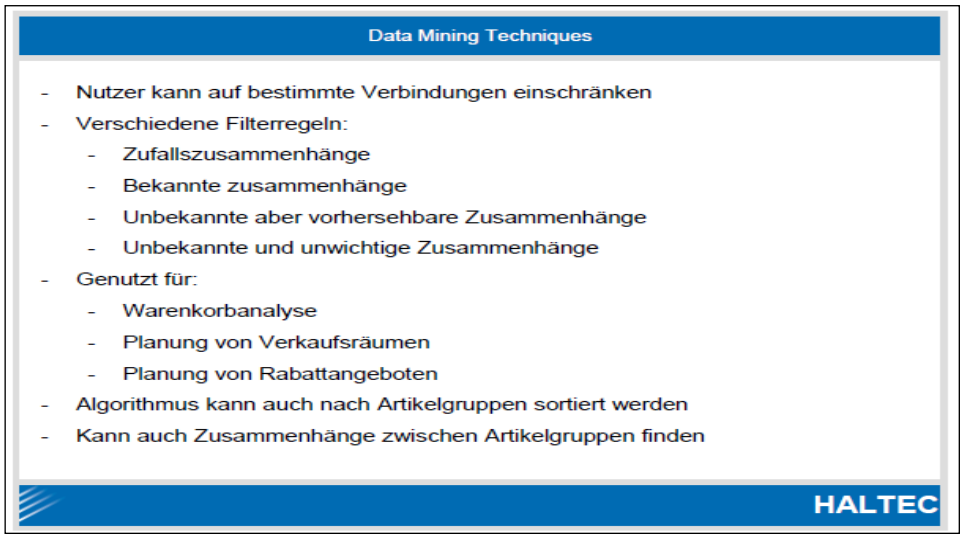
Solution:



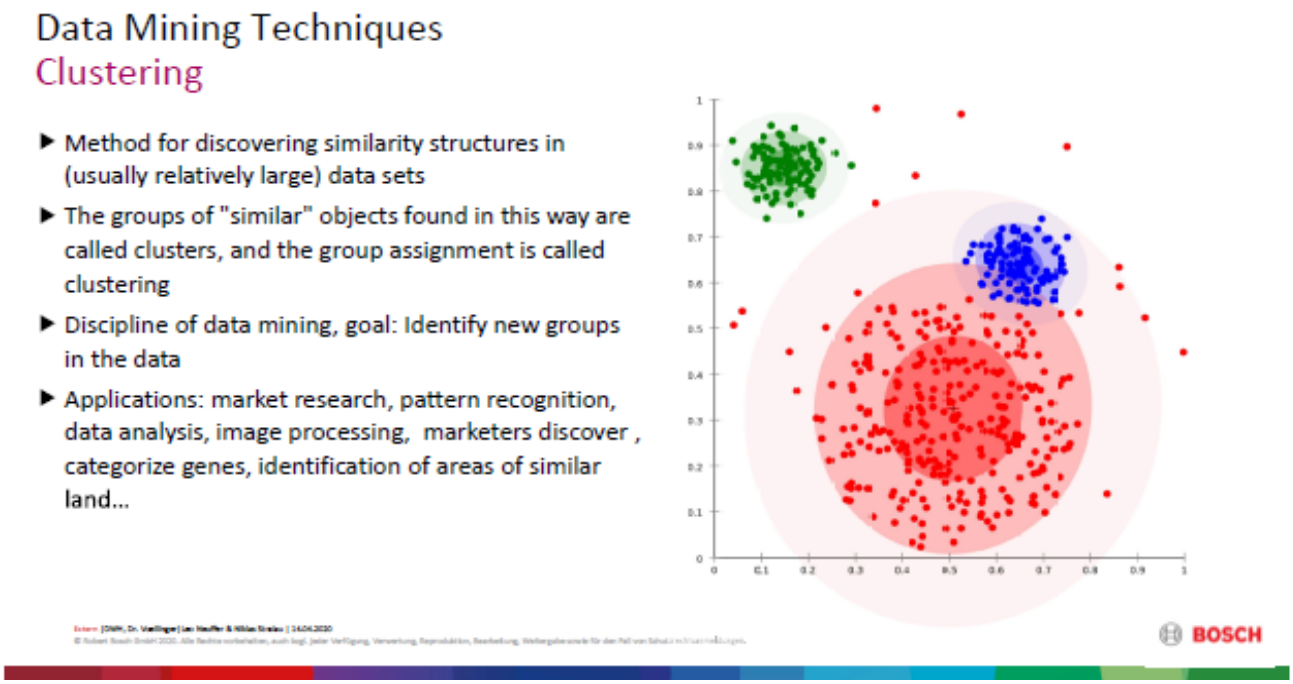



Data Mining Techniques
2. Classification
- Automatische Zerlegung der Daten in Klassen
- Aufteilung anhand von Mustern
- Modell kann genutzt werden, unklassifizierte Daten automatisch einzuordnen
- Verschiedene Algorithmen
- Verschiedene Detaillierung möglich
- Genutzt für:
  - Kreditwürdigkeitsbestimmung
  - Abnutzungsvorhersagen
  - Bestimmen der Unterschiede zwischen Clustern
HALTEC


Data Mining Techniques
3. Associations
- Vergleicht Datensätze und sucht nach Mustern
- Bsp: Kunde, der Farbe kauft, kauft auch Pinsel
- Kann auch Wahrscheinlichkeiten ermitteln
- Bsp: Kunde, der Pinsel kauft, kauft zu 50% auch Farbe
- Vorteile:
  - Vergleicht alle möglichen Kombinationen
  - Findet auch Mehrfachkombinationen
- Kann in großer Datenmenge hunderttausende Verbindungen finden
HALTEC

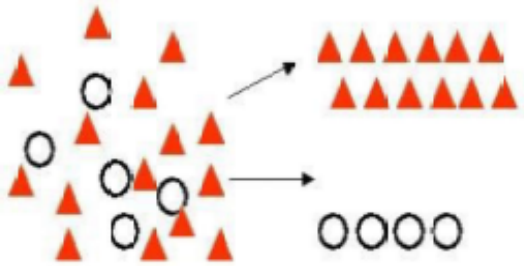

Data Mining Techniques
- Nutzer kann auf bestimmte Verbindungen einschränken
- Verschiedene Filterregeln:
  - Zufallszusammenhänge
  - Bekannte zusammenhänge
  - Unbekannte aber vorhersehbare Zusammenhänge
  - Unbekannte und unwichtige Zusammenhänge
- Genutzt für:
  - Warenkorbanalyse
  - Planung von Verkaufsräumen
  - Planung von Rabattangeboten
- Algorithmus kann auch nach Artikelgruppen sortiert werden
- Kann auch Zusammenhänge zwischen Artikelgruppen finden
HALTEC

Second Solution (SS2021):



Data Mining Techniqes
Classification

- Classification methods, are methods and criteria for the division (classification) of objects or situations into classes
- Many methods can be implemented as algorithms; this is also referred to as machine or automatic classification
- Methods: Decision Tree, K-Nearest Neighbours, Logistic Regression, Neuronal Network



- Applications: Pattern recognition, artificial intelligence, credit rating

BOSCH

Data Mining Techniques
Associations

- Searches for patterns and correlations in the data sets
- These resulting association rules describe correlations between co-occurring things
- Identify items that imply the occurrence of other items within a transaction
- Typical field of application is the interrelationships in purchasing, so-called market basket analysis

For example: In 80 percent of purchases where wine is purchased, bread is also purchased.
Both products occur in 10 percent of purchases

- Other examples: Planning of sales rooms, planning of discount offers
- Advantage: finding enormous numbers of connections in large amounts of data

BOSCH

Data Mining Techniques
Associations – Example Market Basket Analysis (MBA)

Sources

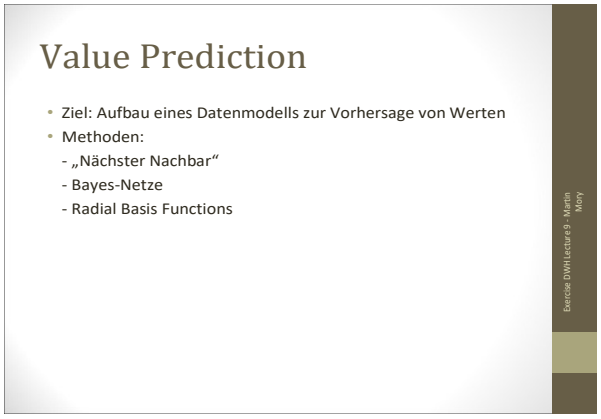
- https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm
- https://de.wikipedia.org/wiki/Clusteranalyse
- https://de.wikipedia.org/wiki/Klassifikationsverfahren
- https://de.wikipedia.org/wiki/Assoziationsanalyse
- H. Voellinger, Lecture DWH at DHBW Stuttgart, DW09 - Advanced Analytics I: Data Mining - Introduction & First Methods, Slide 213

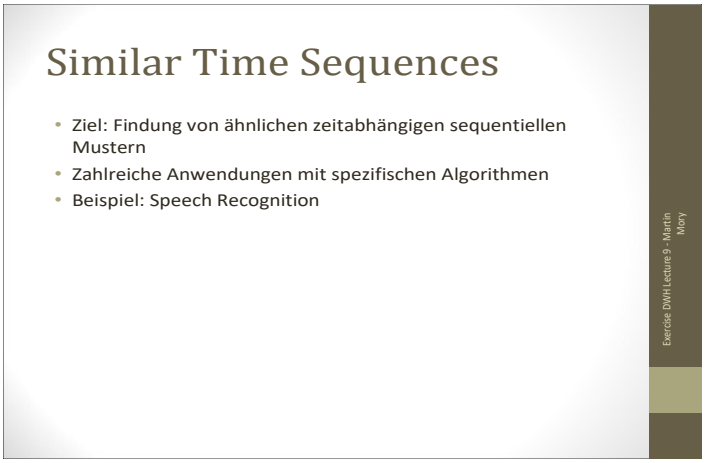
Exercise E9.2: Three Data Mining Methods (Part2)

Task: 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**

Solution:





Exercise E9.3: Measures for Association

Task: 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%?

Solution:

To 1.:

We have 8 market baskets -→Support(Bier=>Orangensaft)=frq(Bier,Orangensaft)/8
We see two baskets which have Bier and Orangensaft together
--→Support = 2/8=1/4 = 25%

To 2.:

We see frq(Bier)=6 und frq(Bier,Milch)=4 -→Conf(Bier=>Milch)=4/6=2/3= 66,7%

To 3.:

To have a support>=50% we need items/products which occur in more than 4 baskets, we see for example Milch is in 5 baskets (#Milch=5), #Bier=6, #Apfelsaft=4 and #Orangensaft=4
Only the 2-pair #(Milch, Bier)=4 has minimum of 4 occurrences. We see this by calculating the Frequency-Matric(frq(X=>Y)) for all tuples (X,Y):

frq(X,Y)	Bier	Milch	A-Saft	O-Saft	Limo
Bier		4	3	2	2
Milch	4		3	2	1
A-Saft	3	3		2	0
O-Saft	2	2	2		1
Limo	2	1	0	1	

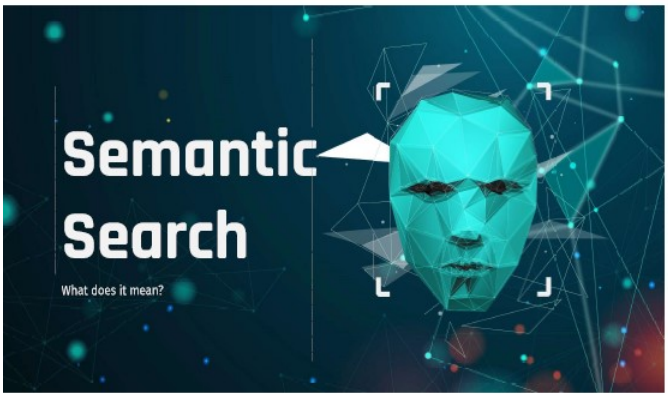
It is easy to see that there are no 3-pairs with a minimum of 4 occurrences.
We see from the above matric, that: Supp(Milch=>Bier)=Supp(Bier=>Milch)4/8=1/2=50%
We now calculate: Conf(Milch=>Bier)=4/#Milch=4/5=80%
From Question 2, we know that Conf(Bier=>Milch)=66,7%

Solution: Only the two association rules (Bier=>Milch) and (Milch=>Bier) have support and confidence >=50%.

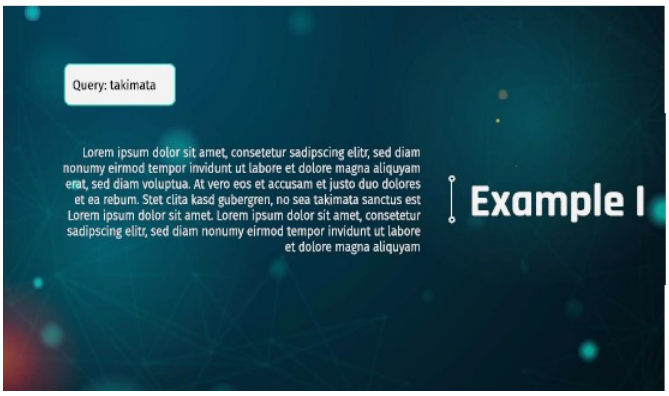
Exercise E9.4*: Evaluate the Technology of the UseCase "Semantic Search"

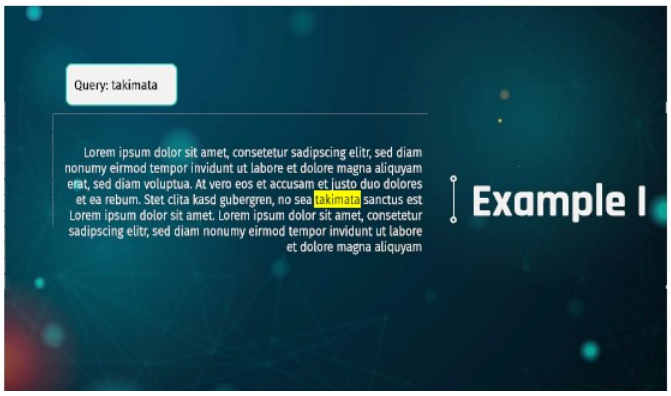
Task: Groupwork (2 Persons): Evaluate and find the underlying technology which is used in "UseCase – Semantic Search: Predictive Basket with Fact-Finder". See: https://youtu.be/vSWLafBdHus

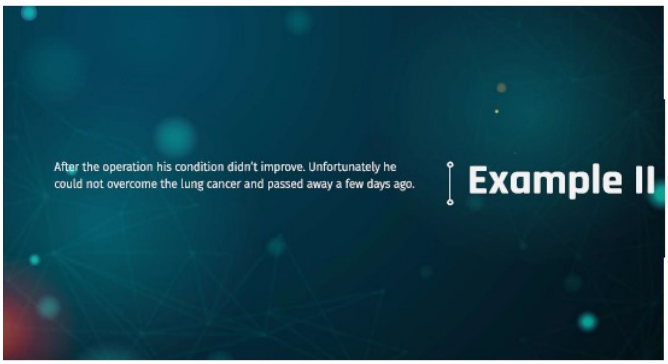
Solution (WS2021):

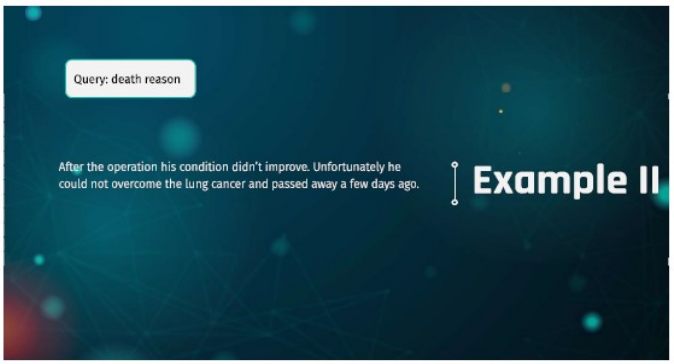

Semantic
Search
What does it mean?

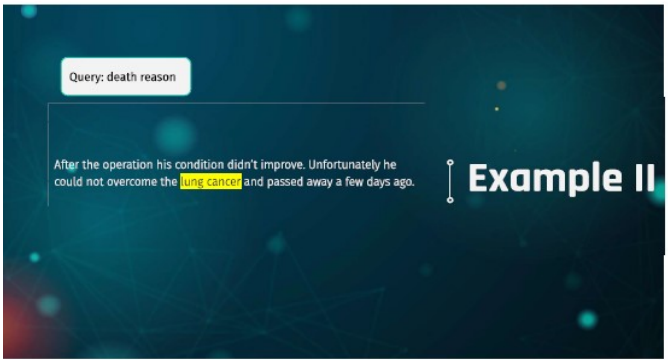

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam
Example I

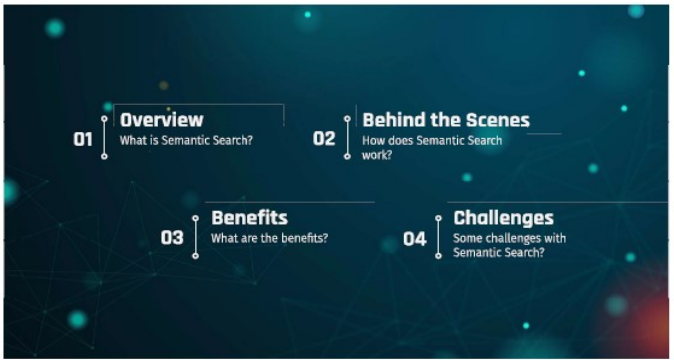

Query: takimata
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam
Example I

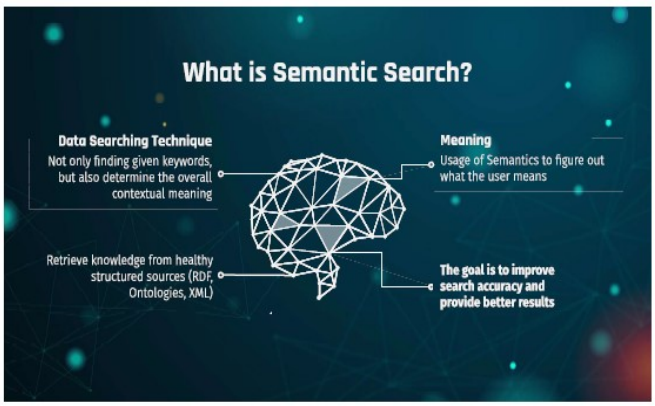

Query: takimata
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam
Example I

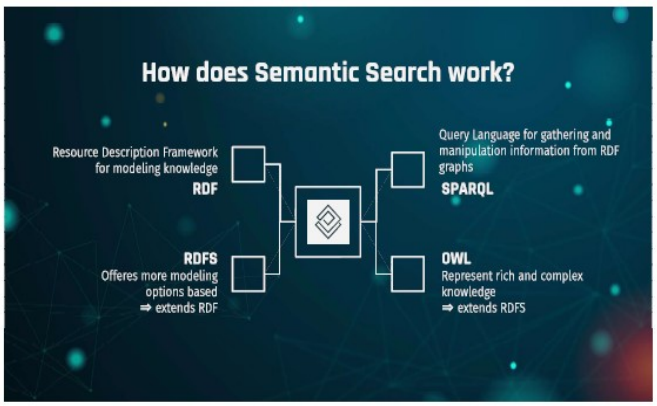

After the operation his condition didn't improve. Unfortunately he could not overcome the lung cancer and passed away a few days ago.
Example II

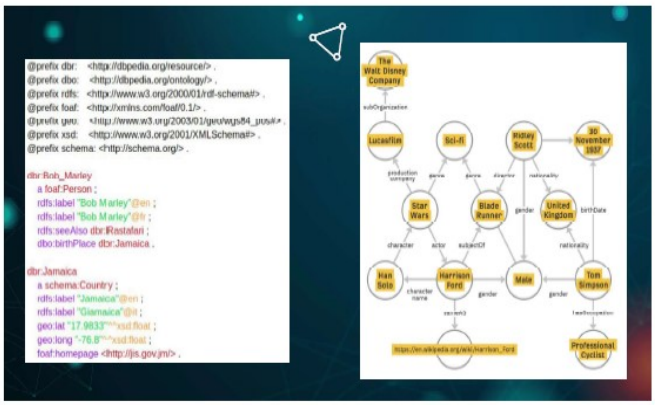

Query: death reason
After the operation his condition didn't improve. Unfortunately he could not overcome the lung cancer and passed away a few days ago.
Example II

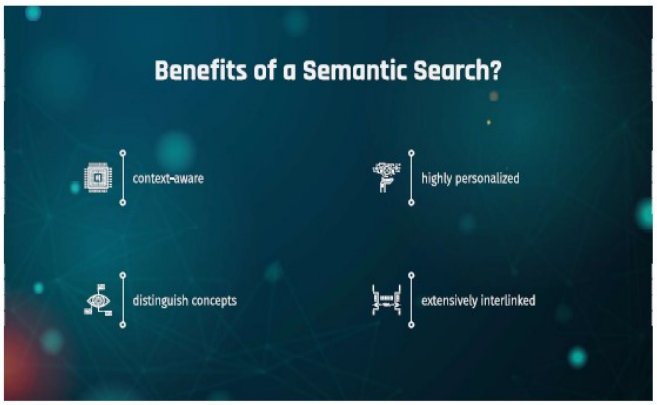

Query: death reason
After the operation his condition didn't improve. Unfortunately he could not overcome the lung cancer and passed away a few days ago.
Example II

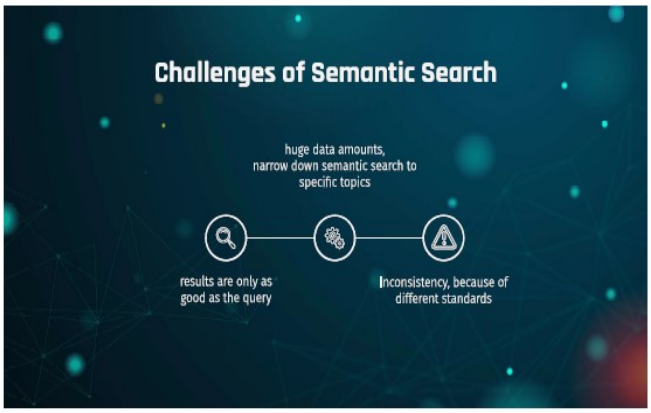

01 Overview
What is Semantic Search?
02 Behind the Scenes
How does Semantic Search work?
03 Benefits
What are the benefits?
04 Challenges
Some challenges with Semantic Search?

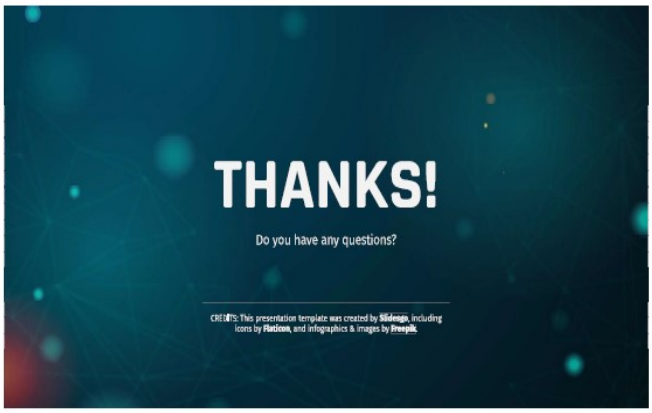








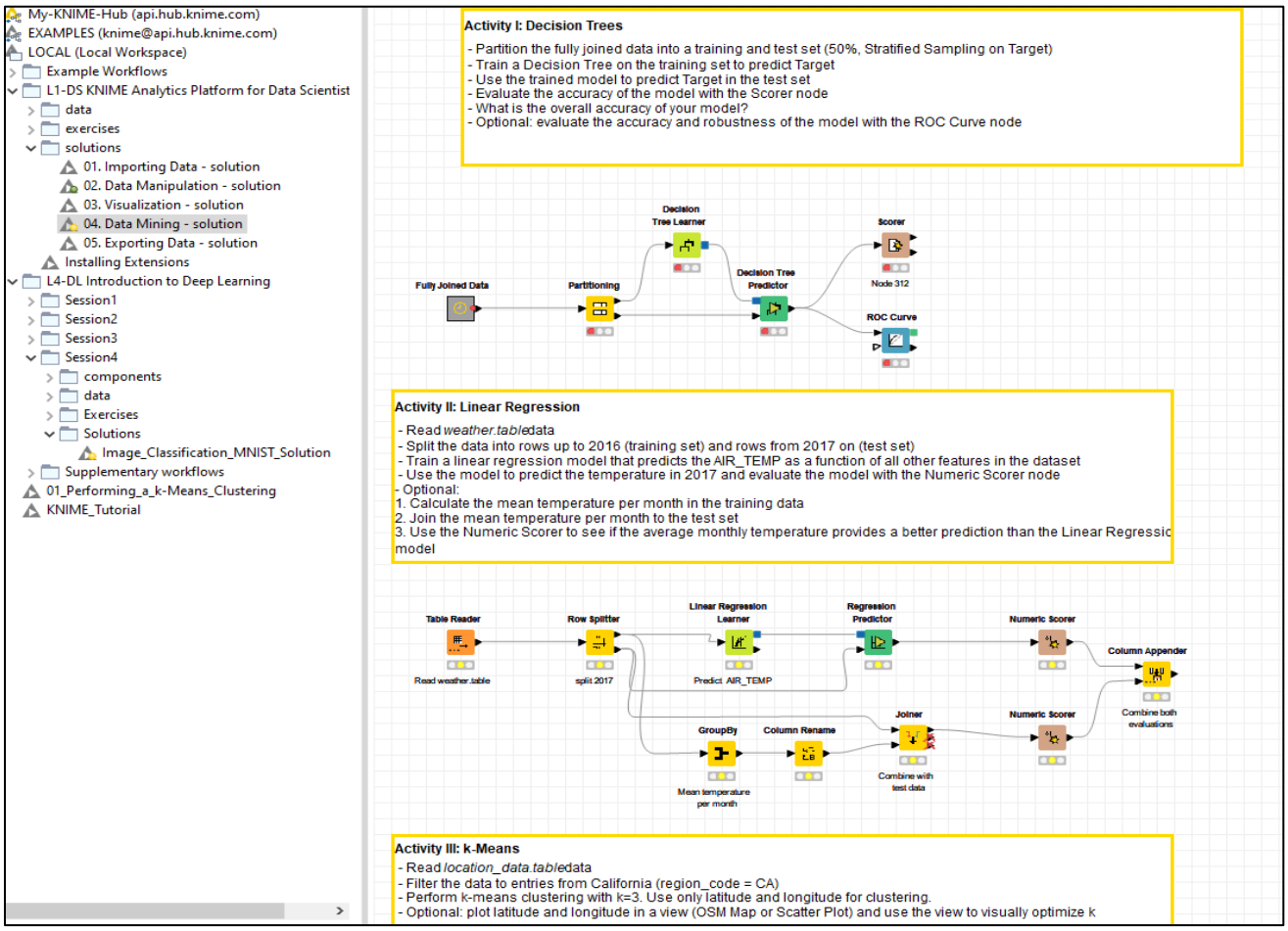




Exercise E9.5*: Run a KNIME-Basics Data Mining solution

Task: 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.

Hint: Follow the instructions given in the KNIME workflow "KNIME Analytics Platform for Data Scientists – Basics (04. Data Mining – solution)" - see image below:



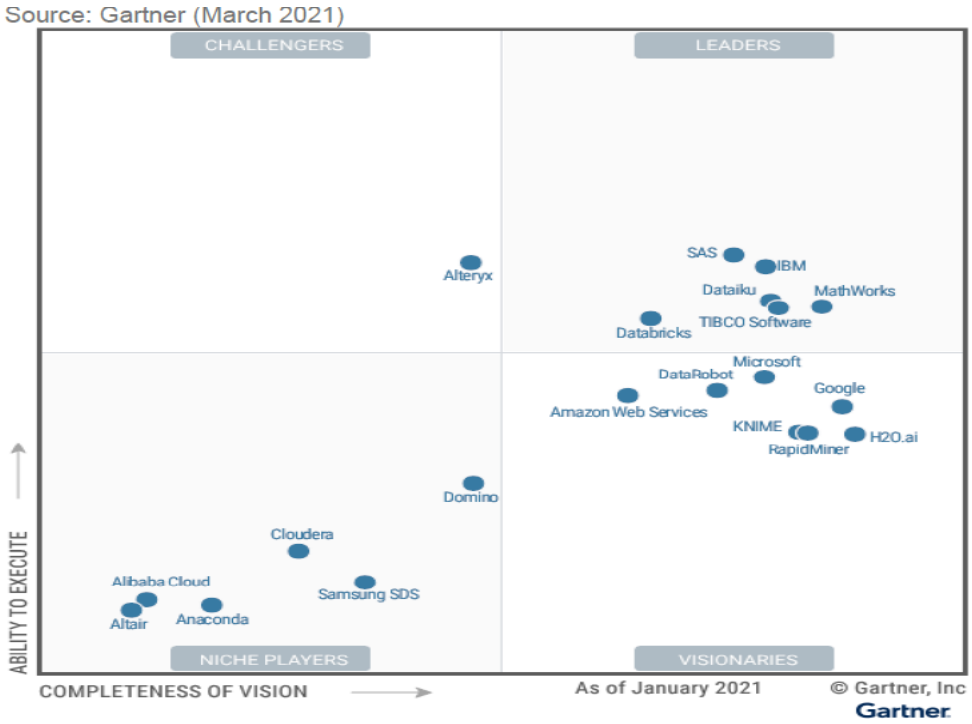
Solution:

……

Exercises (+Solutions) to DHBW Lecture Intro2DWH-Chapter 10

Exercise E10.1*: Compare Data Science/Machine Learning (i.e. DM) Tools

Task: Search for the actual "Gartner Quadrant" of Data Science/Machine Learning (i.e. DM) tools. Give detail descriptions of two of the leading tools in the quadrant:



For further information see in [DHBW-Moodle] the document "Gartner-Machine_Learning_Platform.pdf"

Solution:

……

Exercise E10.2*: Advanced Analytics vs. Artificial Intelligence.

Task: Look for example on the blog: https://seleritysas.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?

Solution:

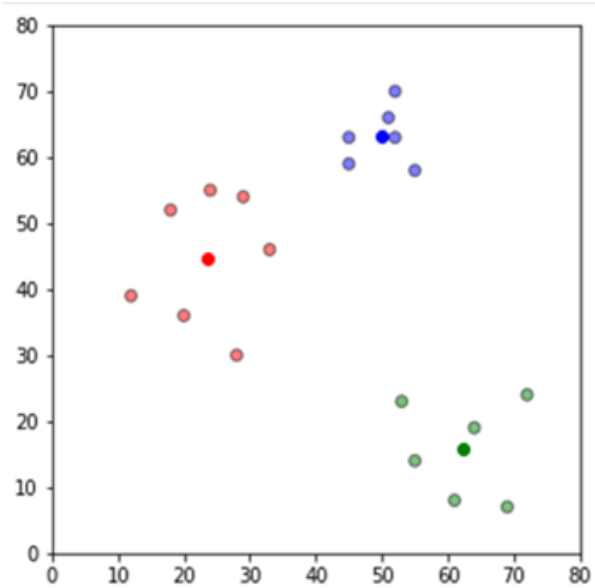
……..

Exercise E10.3*: Create a K-Means Clustering in Python

Task: 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:



Solution:

For a sample solution see: [HVö-5] **Homework_H3.4_k-Means_Clustering.pdf**
https://github.com/HVoellinger/Lecture-Notes-to-ML-WS2020

Second Solution (SS2021):

DHW-Exercise_E10.3-k-Means-Clustering_Example

April 21, 2021

1 Step1: Introduction to DWH-Exercise E10.3 k-Means Clustering

Exercise E10.3 from Exercises to Lesson "DW10-Advanced Analysis II" of the lecture "Introduction to Data Warehousing" at DHBW Stuttgart (SS2021.)

by Luis Seybold and Sven Stahl (DHBW Stuttgart); 19. April 2021 Reviewed and extended by Dr. Hermann Völlinger (DHBW Stuttgart); 21. April 2021

The k-means algorithm of the libraray sklearn.cluster is used to generate k-means clusters for the 19 datapoints of the exercise text. We want 3 clusters, so we set k=3 (= number of centroids). We will import these datapoints (dataset) in Step2 with a 2-dimensional dataframe.

The program is structured in 6 steps:

1. Introduction of the Exercise E10.3
2. Import the packages and classes you need.
3. Provide data to work with and eventually do appropriate transformations
4. Define the Visualization of the data clusters
5. Execute the K-means clustering algorithmus
6. Summary and final remarks

Prerequisites: input data - given in the program; images - all images are located in the directory 'Images/'

```
# Introduction - Print the execution plan
print(" We start the execution of the 6 steps:")
print(" 1.        Introduction of the Exercise E10.3")
print(" 2.        Import the packages and classes you need")
print(" 3.        Provide data to work with and eventually do appropriate␣
 ↪transformations")
print(" 4.        Define the visualization of the data clusters")
print(" 5.        Execute the K-means clustering algorithmus")
print(" 6.        Summary and final remarks")
```

```
 We start the execution of the 6 steps:
 1.     Introduction of the Exercise E10.3
 2.     Import the packages and classes you need
 3.     Provide data to work with and eventually do appropriate transformations
 4.     Define the visualization of the data clusters
 5.     Execute the K-means clustering algorithmus
 6.     Summary and final remarks
```

2 Step2: Import the needed libraries

sklearn - sklearn for the algorithm implementation

pandas - loads the dataset and provids necessary frame details. pandas is also uese for the dataset management

mathplotlip - matplotlib for plotting the results and steps of the algorithm.

pprint - prints the dictionary storage

sys - version information to pythonImport of libraries

[2]:
```
# libraries to import
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import sklearn as sk
import pandas as pd
import matplotlib.pyplot as plt
import pprint

# python version check library
import sys

# to check the time of execution, import function time
import time
import datetime

print(f"This notebook is now launched: {datetime.datetime.now()}")
print()
print("Versions of the used runtime and libraries:")

# print python version, for some imports this version number is viewed as␣
 ↪theirs.
print("python {}".format(sys.version))

# print sklearn
print("sklearn {}".format(sk.__version__))

# print pandas version
print("pandas {}".format(pd.__version__))
```

```
This notebook is now launched: 2021-04-21 18:51:57.970680

Versions of the used runtime and libraries:
```

```
python 3.7.6 (default, Jan  8 2020, 20:23:39) [MSC v.1916 64 bit (AMD64)]
sklearn 0.22.1
pandas 1.0.1
```

3 Step3: Import the data

The data (i.e. 19 datapoints) is given in the program in a 2-dimensional dataframe.

[3]:
```
# Dataset declaration
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]
})
```

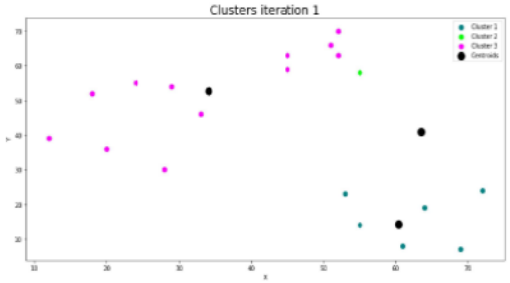
4 Step4: Define the visualization of the k-means clusters

This function is used to plot the output of the k-means algorithm. It plots three clusters with it centers. For this it takes the clusters, the k-means-model and the current iteration that is to be plotted. Dependant to the expected outcome we define the size/shape of the plot. #### Remark: If we use another number k (i.e. k greater than 4) the size and shape of the plot shoud be changed , s.t. all centroids are visible.

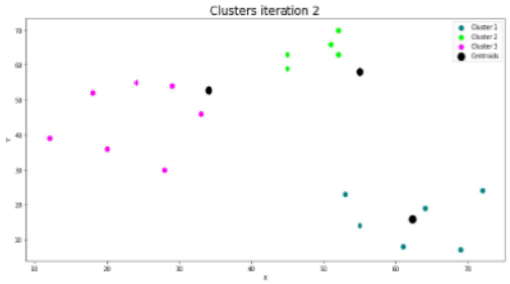
[4]:
```
# plotting function
def plot_clusters(clusters, clf, iteration):
    fig, ax = plt.subplots(figsize=(15,7))
    plt.scatter(x=clusters[clusters['Cluster_Prediction'] == 0]['x'],
                y=clusters[clusters['Cluster_Prediction'] == 0]['y'],
                s=70,edgecolor='teal', linewidth=0.3, c='teal', label='Cluster␣
↪1')


    plt.scatter(x=clusters[clusters['Cluster_Prediction'] == 1]['x'],
                y=clusters[clusters['Cluster_Prediction'] == 1]['y'],
                s=70,edgecolor='lime', linewidth=0.3, c='lime', label='Cluster␣
↪2')


    plt.scatter(x=clusters[clusters['Cluster_Prediction'] == 2]['x'],
                y=clusters[clusters['Cluster_Prediction'] == 2]['y'],
                s=70,edgecolor='magenta', linewidth=0.3, c='magenta',␣
↪label='Cluster 3')
```




```
Initialization complete
start iteration
done sorting
end inner loop
Iteration 0, inertia 5123.666666666667
start iteration
done sorting
end inner loop
Iteration 1, inertia 4593.333333333331
```




```
    plt.scatter(x=clf.cluster_centers_[:, 0], y=clf.cluster_centers_[:, 1], s =␣
↪170, c = 'black', label = 'Centroids',edgecolor='black', linewidth=0.3)
    plt.legend(loc='upper right')
    ax.set_ylabel('Y')
    ax.set_xlabel('X')
    plt.title('Clusters iteration ' + str(iteration), fontsize = 20)
    plt.show()
```

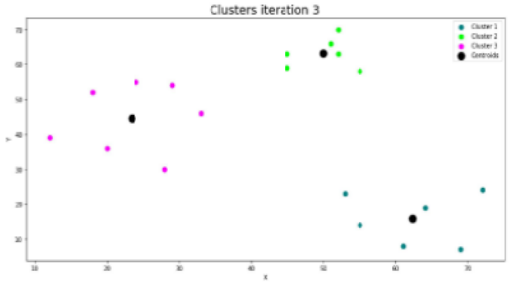
5 Step5: Execute the k-means clustering algorithm

In this cell we calculate and plot iterations of the k-means algorithm. For every step we fit our StandardScaler and the k-means. We choose random initializations, three clusters and cap the maximum iterations to the number of steps. As a result the k-means is only calculated up to the given iteration. We plot the output of the k-means with the above function from Step3. The detailed mode for the k-means alogrithm shows us that the algorithm converges at the fourth iteration. After that we have our final centers. The proof of the convergence of this method is outside the scope of this particular task.

[5]:
```
# plot 4 steps of kmeans
for it in range(1, 6):
    scaler = StandardScaler().fit(df)
    kmeans = KMeans(n_clusters=3, init='random', max_iter=it, n_init=1,␣
↪random_state=10, verbose=1)
    clusters = df.copy()
    clusters['Cluster_Prediction'] = kmeans.fit_predict(df)
    plot_clusters(clusters, kmeans, iteration=it)
```

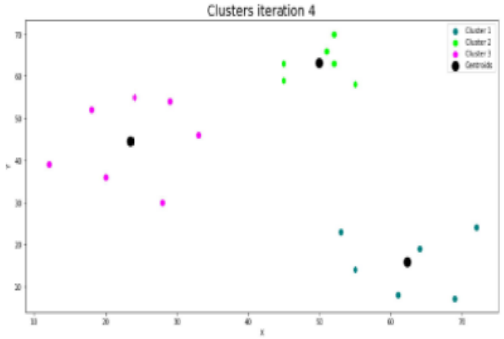
```
Initialization complete
start iteration
done sorting
end inner loop
Iteration 0, inertia 5123.666666666667
```

```
Initialization complete
start iteration
done sorting
end inner loop
Iteration 0, inertia 5123.666666666667
start iteration
done sorting
end inner loop
Iteration 1, inertia 4593.333333333331
start iteration
done sorting
end inner loop
Iteration 2, inertia 1624.4285714285716
```

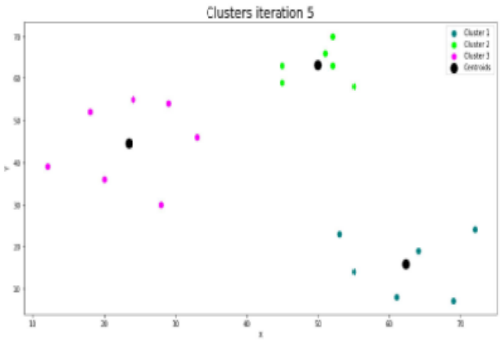



```
Initialization complete
start iteration
done sorting
end inner loop
Iteration 0, inertia 5123.666666666667
start iteration
done sorting
end inner loop
Iteration 1, inertia 4593.333333333331
start iteration
done sorting
end inner loop
Iteration 2, inertia 1624.4285714285716
start iteration
done sorting
```

```
end inner loop
Iteration 3, inertia 1624.4285714285716
center shift 0.000000e+00 within tolerance 3.593213e-02
```




```
Initialization complete
start iteration
done sorting
end inner loop
Iteration 0, inertia 5123.666666666667
start iteration
done sorting
end inner loop
Iteration 1, inertia 4593.333333333331
start iteration
done sorting
end inner loop
Iteration 2, inertia 1624.4285714285716
start iteration
done sorting
end inner loop
Iteration 3, inertia 1624.4285714285716
center shift 0.000000e+00 within tolerance 3.593213e-02
```



6 Step6: Summary and final remarks

Remark1: We stopped the algorithm by defining that only 5 iterations should run. We know by experience that then the centroids are in a stable location for these set of data. Actually the algorithm is stable after 3 iterations (look on the calculated parameter "center shift"). It is also possible to let the algorithm calculate the numbers of iterations by using this parameter.

Remark2: By mathematical methods we should also be able to proof, that the k-means algorithm will convert to stable centroid locations. But this is outside the problem scope of this special exercise.

Remark1 and remark2 will be done in the next version of the program.

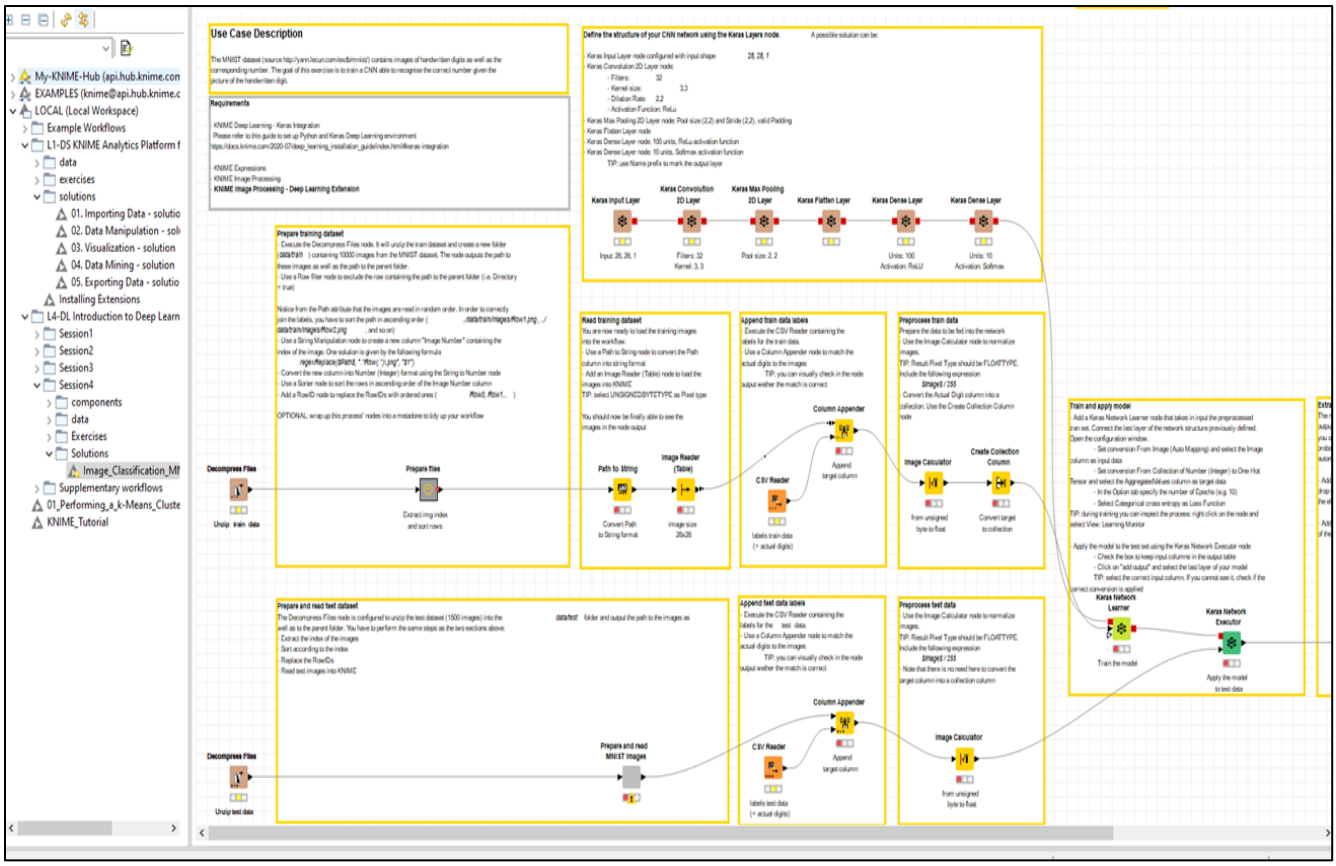
Finally we print the execution-time and execution-date.

[6]:
```
# print current date and time
print("date",time.strftime("%d.%m.%Y %H:%M:%S"))
print ("*** End of Homework-E10.3_K-Means Clustering ***")
```

```
date 21.04.2021 18:51:59
*** End of Homework-E10.3_K-Means Clustering ***
```

Exercise E10.4*: Image-Classification with MNIST Data using KNIME

Task: Homework for 2 Persons: Rebuild the KNIME Workflow (use given solution) for Image-Classification and give technical explanations to the solution steps.
Hint: Follow the instructions given in the KNIME workflow "L4-DL Introduction to Deep Learning/Session4/Solutions (Image Classification MNIST Solution)" - see image below:


Use Case Description
Requirements
Define the structure of your CNN network using the Keras Layers node.
Prepare training dataset
Read training dataset
Append train data labels
Preprocess train data
Train and apply model
Prepare and read test dataset
Append test data labels
Preprocess test data
Decompress Files
Prepare files
Path to String
Image Reader (Table)
CSV Reader
Column Appender
Image Calculator
Create Collection Column
Keras Input Layer
Keras Convolution 2D Layer
Keras Max Pooling 2D Layer
Keras Flatten Layer
Keras Dense Layer
Keras Dense Layer
Keras Network Learner
Keras Network Executor
Train the model
Apply the model to test data
Prepare and read MNIST images

Solution:

……..

References

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**
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 .... 'Intelligent Miner for Data – Applications Guide'*, IBM Redbook, SG24-5252-00, ISBN: 0-7384-1276-7**

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; WS2021; http://www.dhbw-stuttgart.de/~hvoellin/

17. [HVÖ-2]: *Hermann Völlinger and Other: Exercises & Solutions of the Lecture "Introduction to Data Warehousing";* DHBW Stuttgart; WS2021 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 TINF19D: 'Directory of supporting Information for the DWH Lecture';* *Kurs: T3INF4304_3_Data Warehouse (dhbw-stuttgart.de)***