DHBW
Duale Hochschule
Baden-Württemberg
Stuttgart

Dr. Hermann Völlinger,
Mathematics & IT-Architecture

Introduction to Data Warehousing (DWH)

DHBW – Fakultät Technik-Informatik, Stuttgart, Autumn 2023

Dr. Hermann Völlinger, Mathematics & IT Architecture

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

www.dhbw-stuttgart.de

Last Update: Thursday, December 14, 2023

General Remarks to Lecture DWH (1/2)

- Our first lecture starts on Tuesday, 10.10.2023, 1:00pm - 4:15pm (4 lecture hours + 15 minutes break). The remaining 8 lectures will all start at 2:00 pm. Last lecture on Tuesday, 12.12.2023.

- We have a total of 9 appointments which are all on Tuesdays except on Tuesday the 21.11.23, because on 22.11.23 the study day of the DHBW took place. The exam week runs from 18.12.23 to 22.12.23. Details: https://rapla.dhbw.de/rapla/calendar?key=YFQc7NlGleuSdybxizoa8NHjLLNjd9D6tjBdAvDwwzXobLEfUIsCXHwYu-Ma7QfggMDkLLj1CsQ-kB7hFJSGjYcYLXE5KV9oTTpcSjsE5apebBNbC_ZjtngvStO4G7YHGryjvwt1kpad5g93Dkdn0A&salt=1046252309

- If an online meeting need to be done (if corona should make this necessary), we will use Zoom (invitation credentials will then be committed in Moodle). Up to now it is planed that all lecture dates are held as face-to-face meetings.

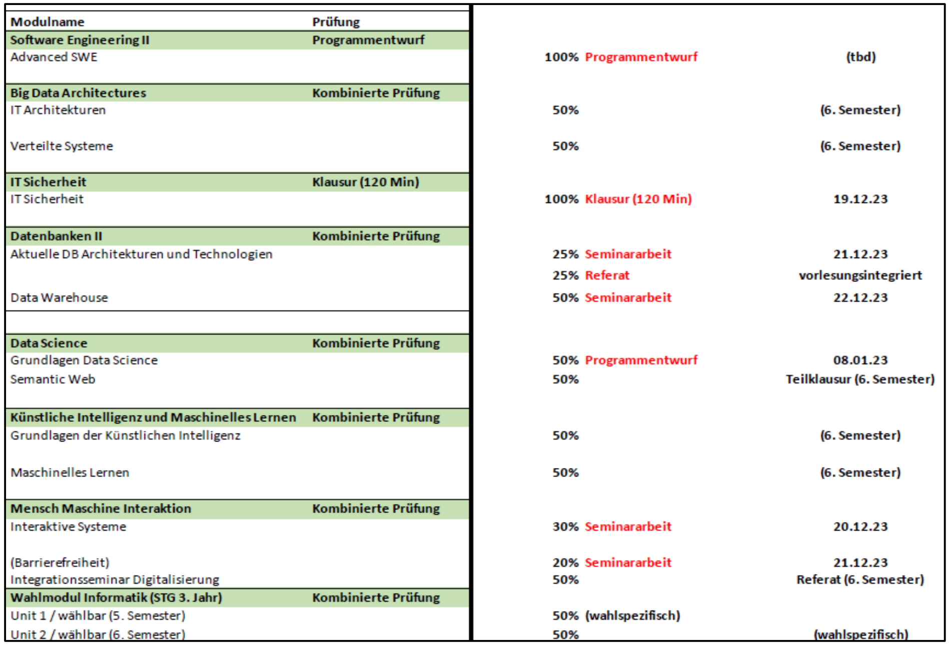
- The lecture script is in **English**, since the common IT language in the area of DWH and Analytics is English. Some dedicated slides are in German, not to loose “Look and Feel” of the slide.

- Lecture information & supporting material (> 140 documents and technical papers in 4 DWH categories) you will find it in Moodle/”Supporting Information for DWH Lecture” Kurs DW 21E:

General Remarks to Lecture DWH (2/2)

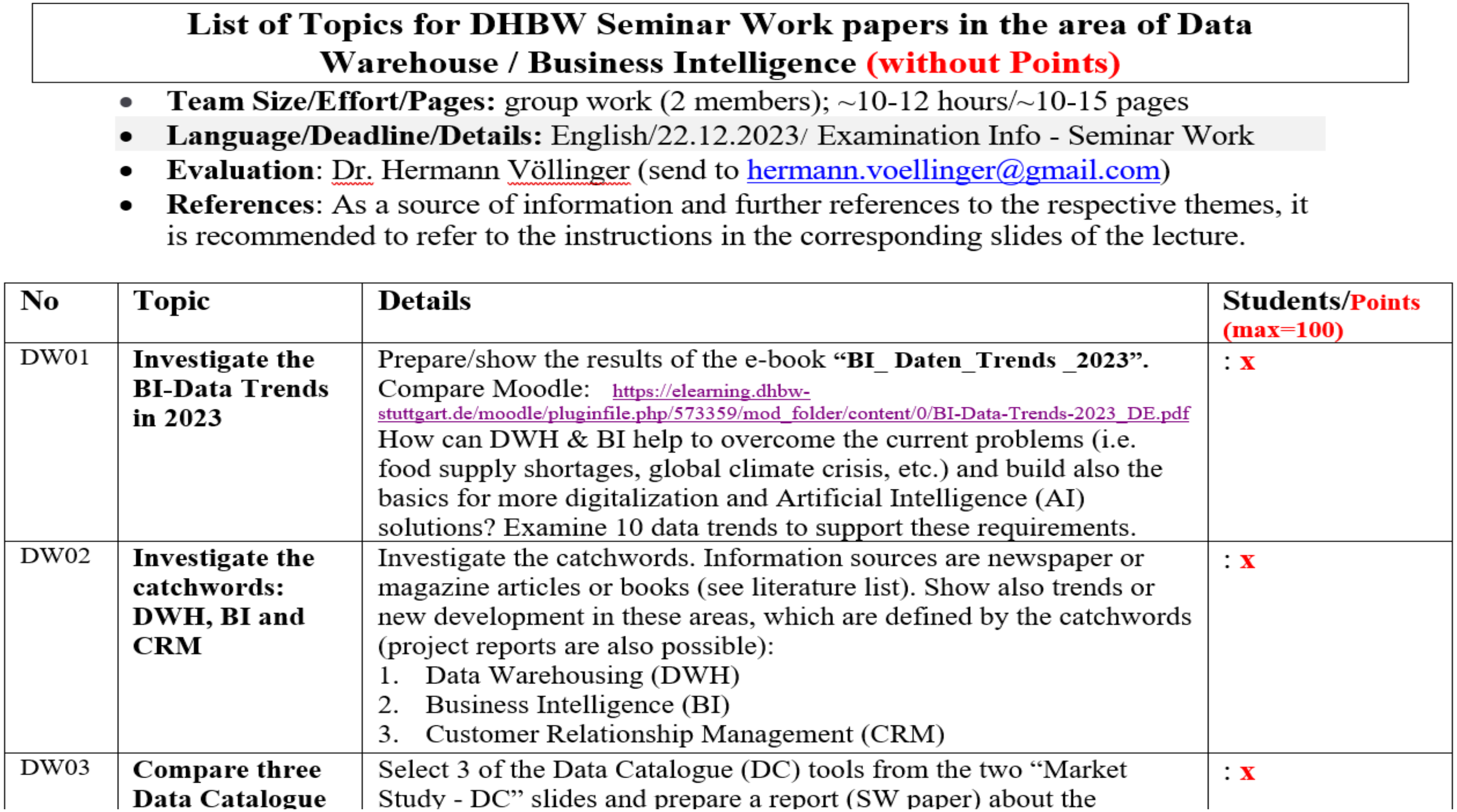
- In exercises everyone should present at least one time his exercise solution. Collection of solutions ("Musterlösungen") together with the lecture script you could find in my DHBW Homepage: http://www.dhbw-stuttgart.de/~hvoellin/ also including sample data for exercises and other information about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

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



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

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



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

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

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

Content: Introduction to Data Warehousing (DWH)

Goal: Introduction, Architecture and Basic Concepts

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

Literature List – Part 1

1. [BD-DWH]: *Barry Devlin 'Data Warehouse....'*, Addison-Wesley, ISBN: 0-201-96425-2

2. **[RK-DWH]: *R. Kimball 'The Data Warehouse Toolkit.'*, John Wiley & Sons, NY 1996, ISBN: 0-471-15337-0**

3. **[AB&HG-DWH]: *Andreas Bauer, Holger Günzel (Hrsg.): 'Data Warehouse Systeme - Architektur, Entwicklung, Anwendung'* DPunkt Verlag Heidelberg 2004, 3. Auflage, ISBN: 978-3-89864-540-9**

4. [RK-DWH/TK]: *R. Kimball and Other: 'The Data Warehouse Lifecycle Toolkit'*, John Wiley & Sons, NY 1998, ISBN: 0-471-25547-5

5. [SE-DWH/BI]: *Stefan Eckrich and Other: 'From Multiplatform Operational Data to Data Warehousing and Business Intelligence'*, IBM Redbook, SG24-5174-00, ISBN: 0-7384-0032-7

6. [VAC&Other-BI/390]: *V. Anavi-Chaput and Other: 'Business Intelligence Architecture on S/390 – Presentation Guide'*, IBM Redbook, SG24-5641-00, ISBN: 0-7384-1752-1

7. **[DM-MD]: *David Marco: 'Building &Managing the Meta Data Repository'*, John Wiley & Sons 2000, ISBN: 0-471-35523-2**

Literature List – Part 2

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

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

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

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

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

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

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

Literature List – Part 3

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

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

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

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

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

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

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

Link between DWH Content and Literature

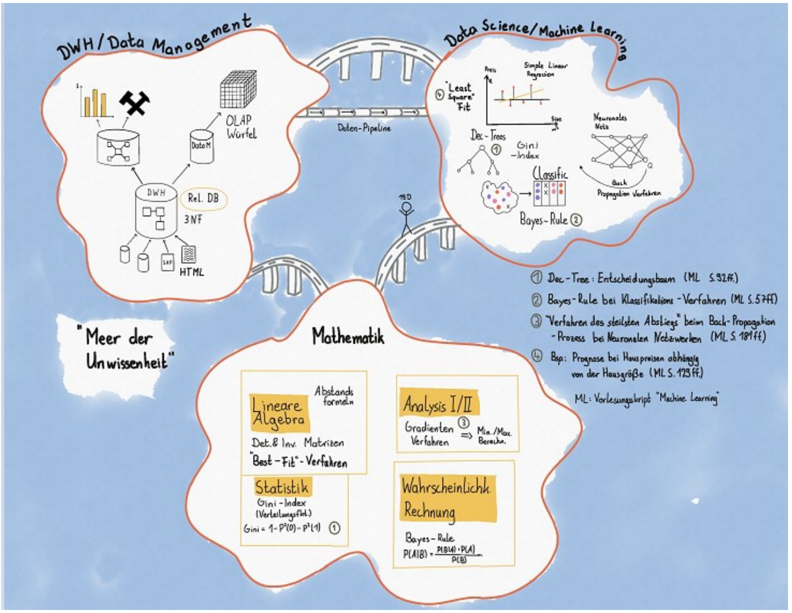
Goal: Sort the 10 Lessons in 4 Categories and connect these with the Literature:

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

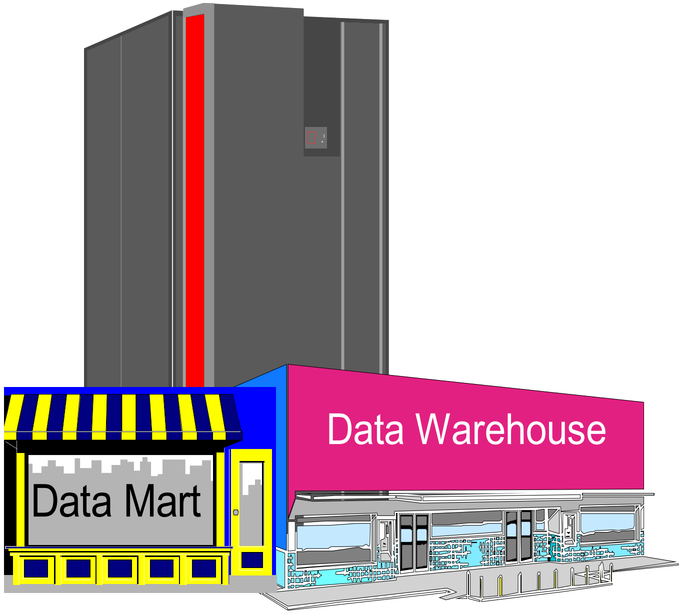
Goals of the Lecture

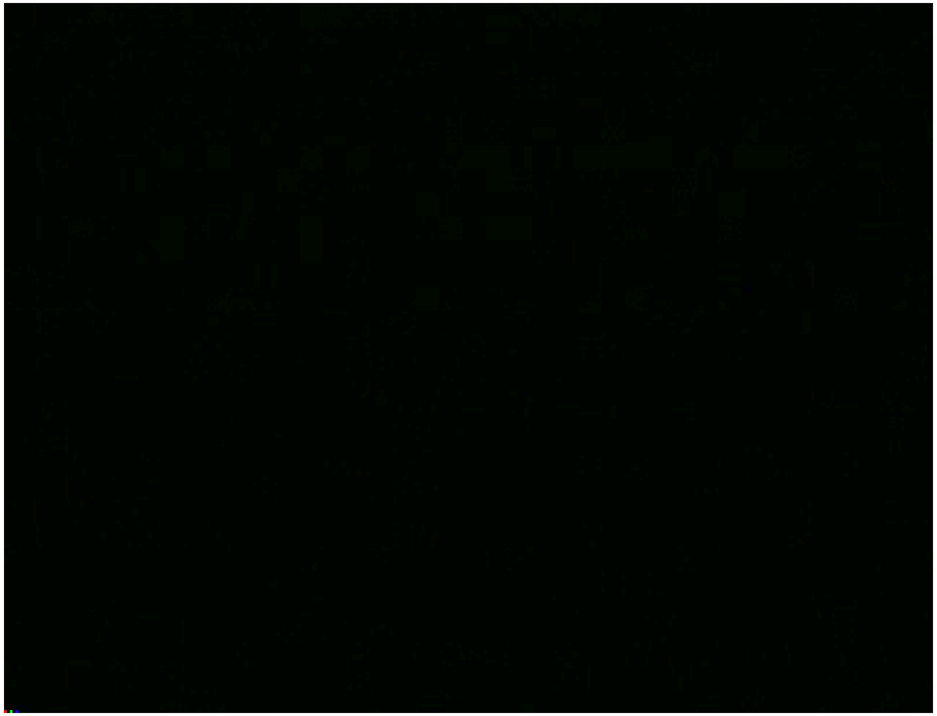
The lecture’s aim is to introduce the concepts of a Data Warehouse (DWH). We learn the most important methods that are used in DWH and they are presented with their essential features. Several references are given to in-depth applications or information through internet-links or further literature. In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

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


DWH / Data Management
OLAP Würfel
DataM
DWH
Rel. DB
3NF
SAP
HTML
Daten-Pipeline
Data Science/Machine Learning
Preis
Simple Linear Regression
④ "Least Square" Fit
Size
Dec - Trees ①
Gini -Index
Neuronales Netz
Classific
Bayes - Rule ②
Back Propagation Verfahren
"Meer der Unwissenheit"
Mathematik
Lineare Algebra
Abstands formeln
Det. & Inv. Matrizen
"Best - Fit" - Verfahren
Statistik
Gini - Index (Verteilungsfkt.)
Gini = 1 - P²(0) - P²(1) ①
Analysis I/II
Gradienten Verfahren ③ ⇒ Min./Max. Berechn.
Wahrscheinlichk. Rechnung
Bayes - Rule
P(A|B) = P(B|A)·P(A) / P(B)
① Dec - Tree : Entscheidungsbaum (ML S.92ff.)
② Bayes - Rule bei Klassifikations - Verfahren (ML S.57ff)
③ "Verfahren des steilsten Abstiegs" beim Back-Propagation - Prozess bei Neuronalen Netzwerken (ML S.181ff)
④ Bsp: Prognose bei Hauspreisen abhängig von der Hausgröße (ML S.129ff.)
ML: Vorlesungsskript "Machine Learning"

DW01- Introduction to DWH & BI





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

Water, Water Everywhere …

…But not a drop to drink

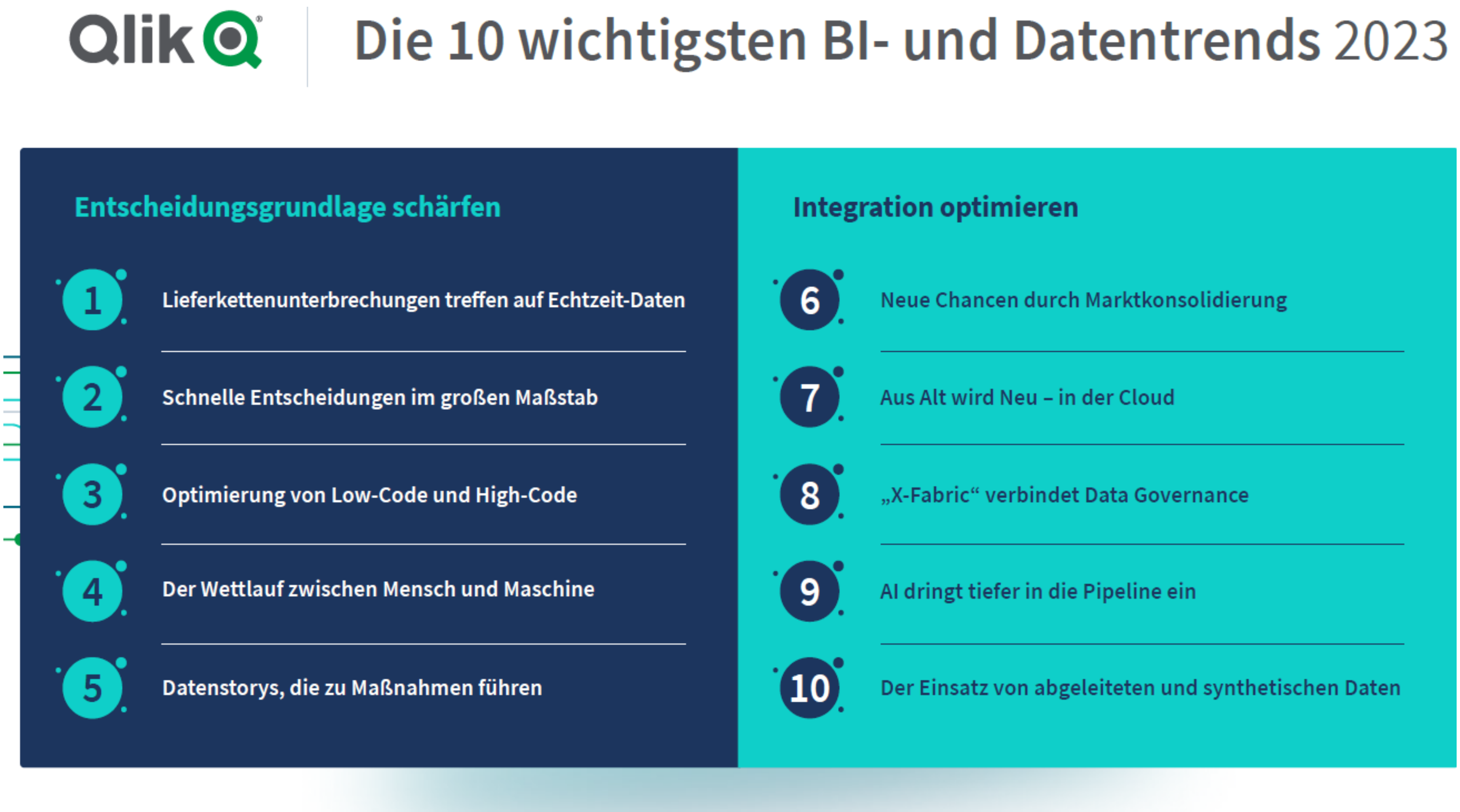
Data, Data Everywhere…

…But none to help me think

Competition

10 Data Trends to support Digitalization

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



Qlik | Die 10 wichtigsten BI- und Datentrends 2023

Entscheidungsgrundlage schärfen

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

Integration optimieren

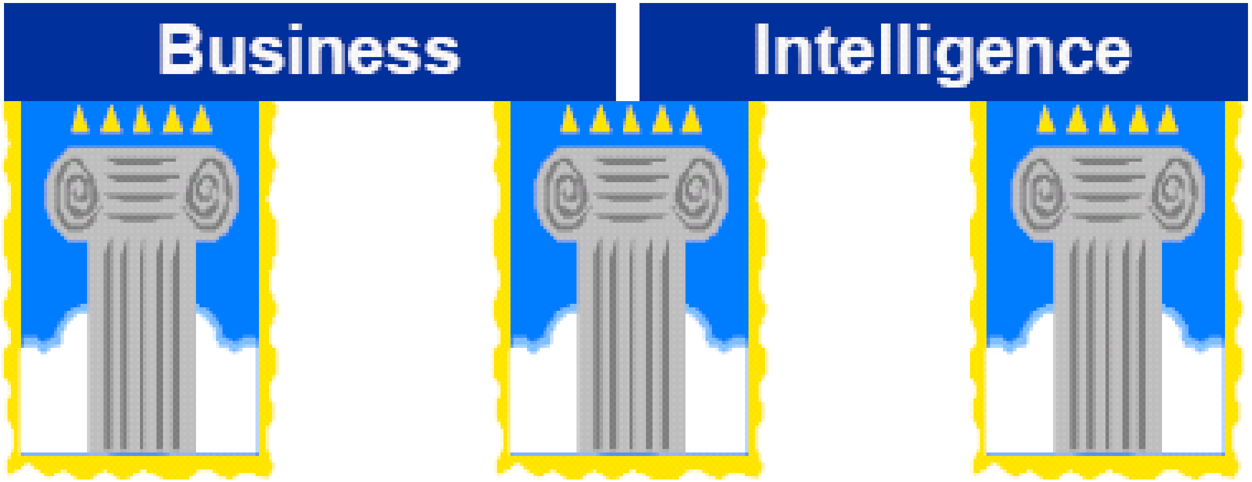
6. Neue Chancen durch Marktkonsolidierung
7. Aus Alt wird Neu – in der Cloud
8. „X-Fabric“ verbindet Data Governance
9. AI dringt tiefer in die Pipeline ein
10. Der Einsatz von abgeleiteten und synthetischen Daten

First Definition: What is BI & BI Mission

Enterprise Applications and Data
Business Intelligence
Insight
Business value,

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

What is BI – the three Pillars of BI



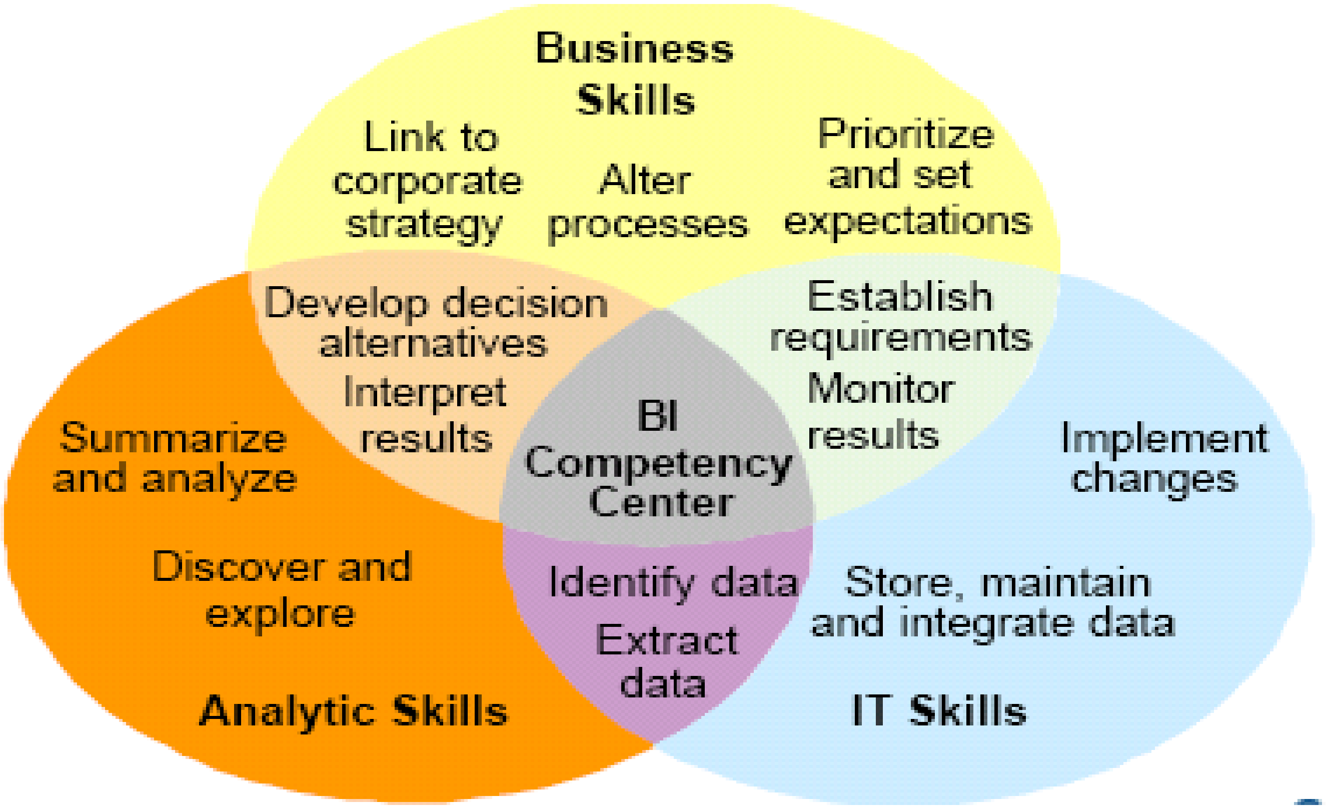
Infrastructure & Tools + Best Practices + Applications

ETL or Data Quality
Database Management
Metadata
Reporting
Ad Hoc Analysis
Data Mining

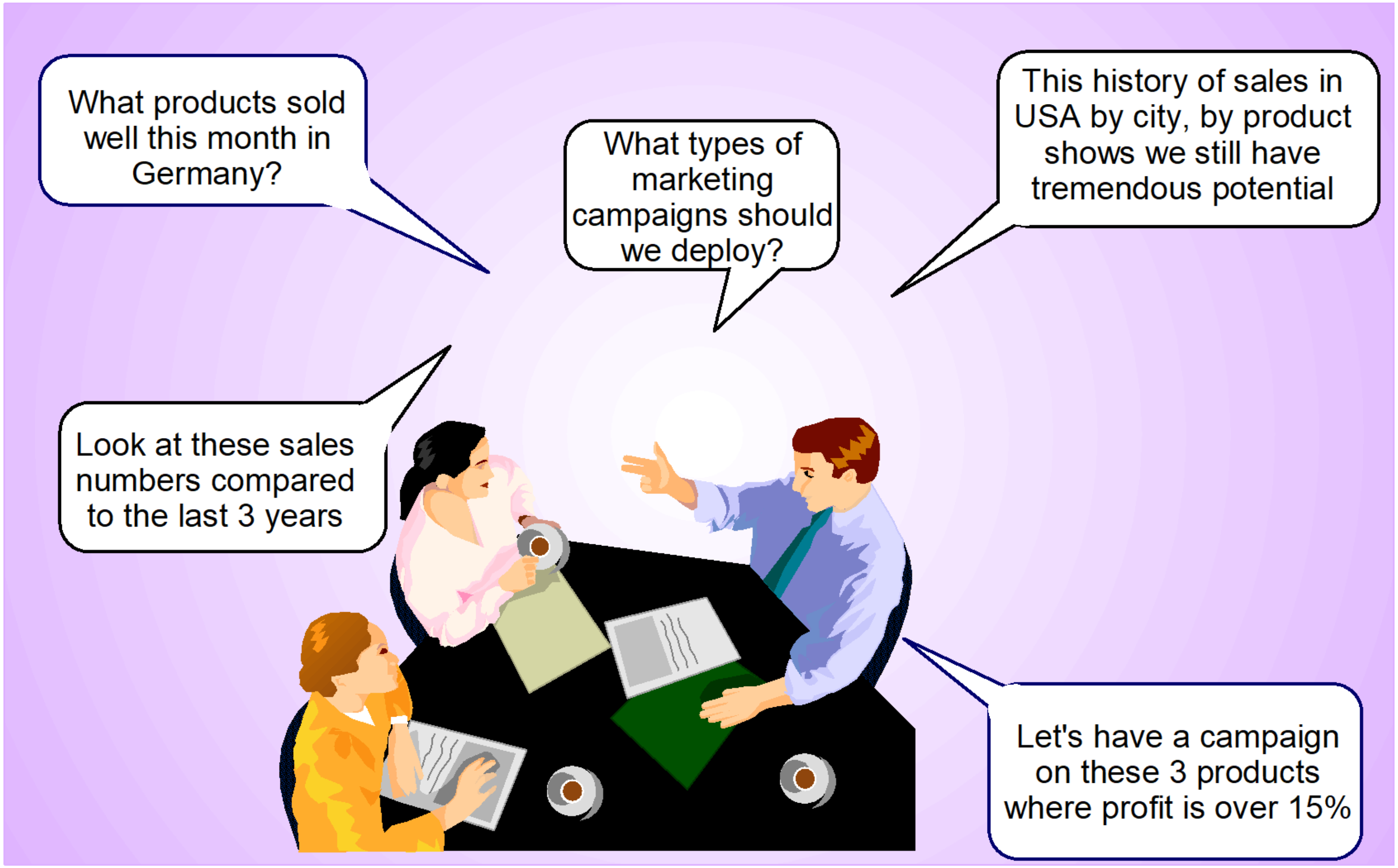
Methodologies
Organization
Management
Culture
Skills

Corporate Performance Management
CRM Analytics
Regulatory Compliance
Predictive Analytics
Profitability Analysis

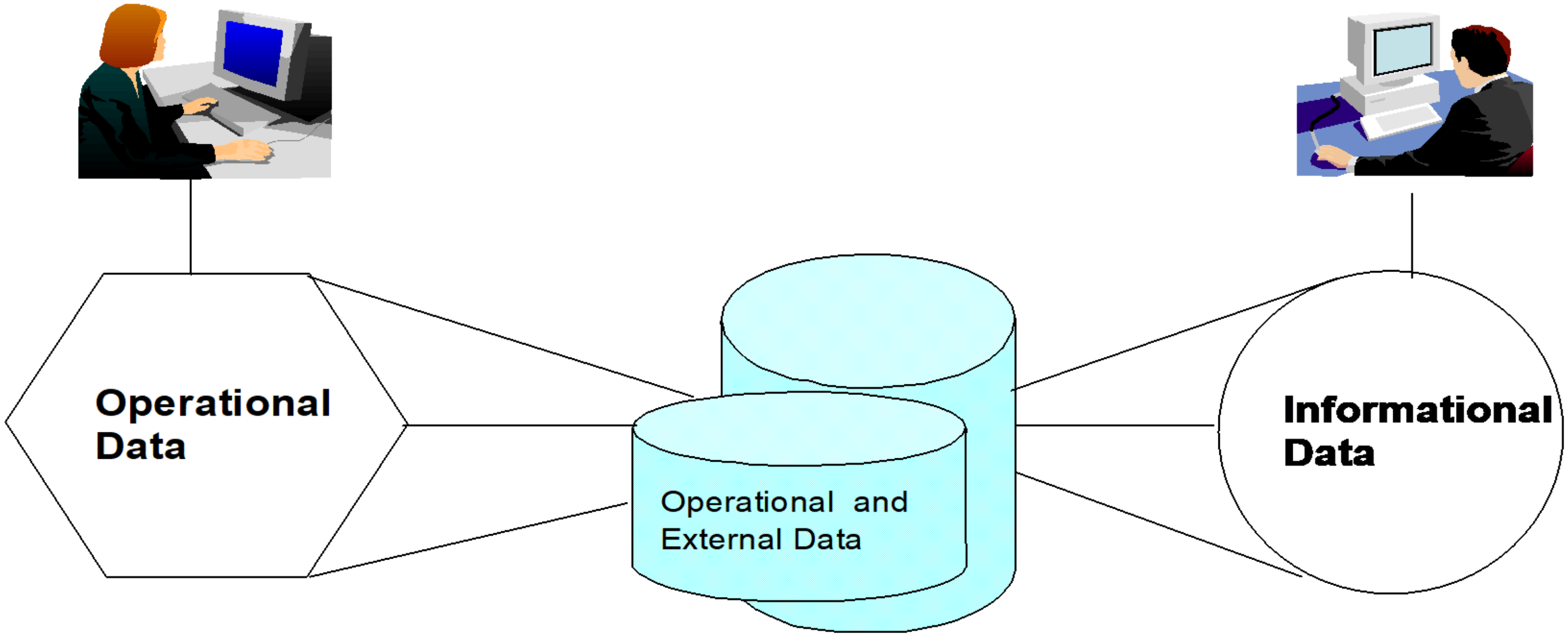
What is BI – the BI Competency Center



BI - Getting the Answers you Need


What products sold well this month in Germany?
What types of marketing campaigns should we deploy?
This history of sales in USA by city, by product shows we still have tremendous potential
Look at these sales numbers compared to the last 3 years
Let's have a campaign on these 3 products where profit is over 15%

Different Data for Different Users



Operational Systems

- Order Entry
- Payroll
- Accounts Receivable
- Personnel

Informational Systems

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

Structuring the Data – Five Data Types

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

Turning Data to Information

- **The need for a warehouse model**

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

- **The need to transform the data**

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

- **The need for an information catalogue**

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

The need for an Information Catalog (Metadata)

- Finding & Understanding the Data

What Data Exists?

Where Is It?

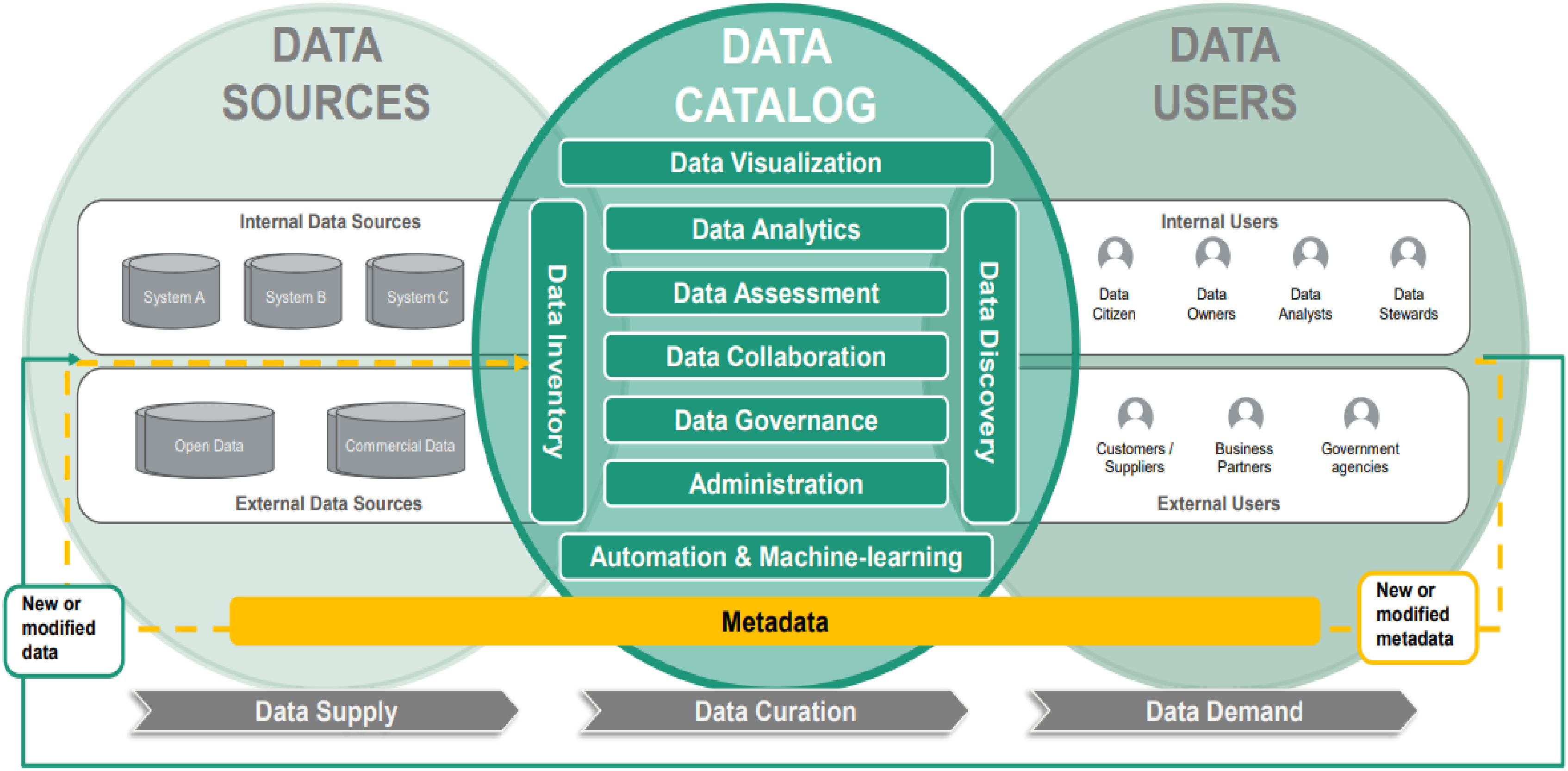
How Can I Get It?

What Format Is It In?

What Does It Mean?

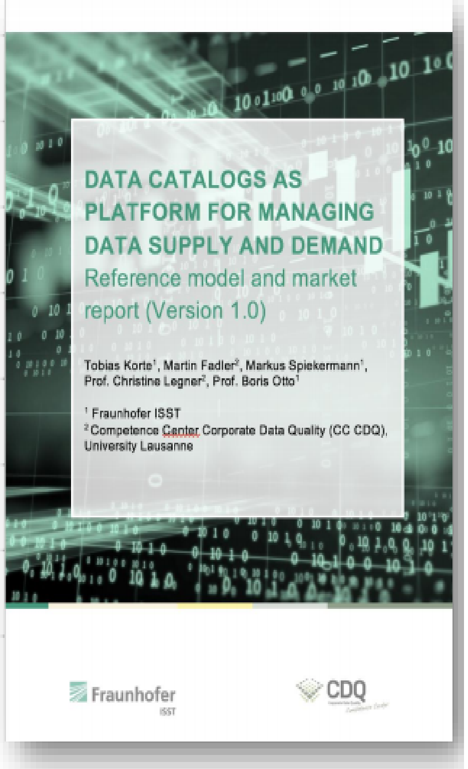
You will learn more about this later

The Data Catalog links Data Supply and Demand



Market Study: Data Catalogues (1/2)

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



DATA CATALOGS AS PLATFORM FOR MANAGING DATA SUPPLY AND DEMAND

Reference model and market report (Version 1.0)

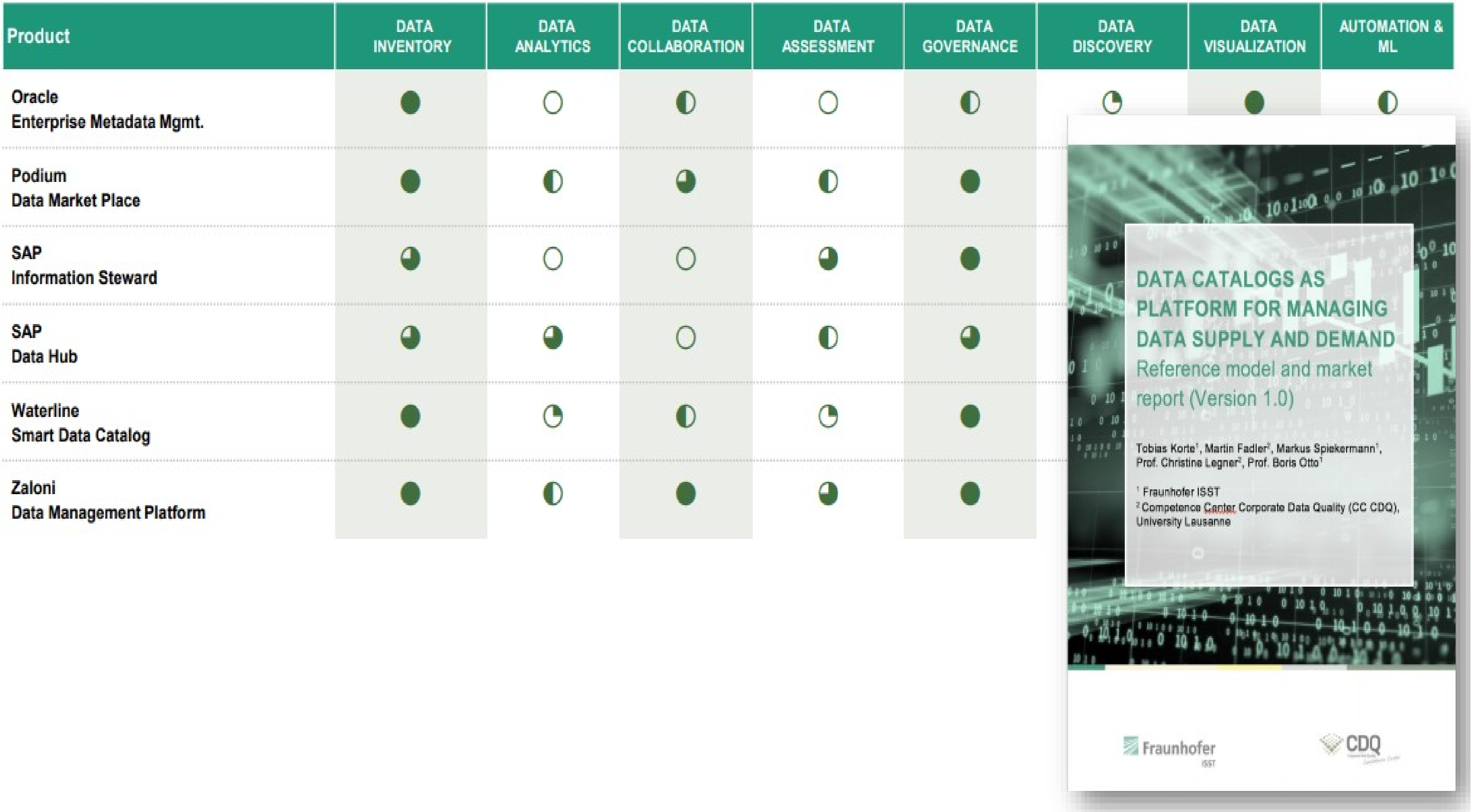
Tobias Korte[1], Martin Fadler[2], Markus Spiekermann[1], Prof. Christine Legner[2], Prof. Boris Otto[1]

[1] Fraunhofer ISST
[2] Competence Center Corporate Data Quality (CC CDQ), University Lausanne

Fraunhofer ISST

CDQ

Market Study: Data Catalogues (2/2)



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

DATA CATALOGS AS PLATFORM FOR MANAGING DATA SUPPLY AND DEMAND

Reference model and market report (Version 1.0)

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

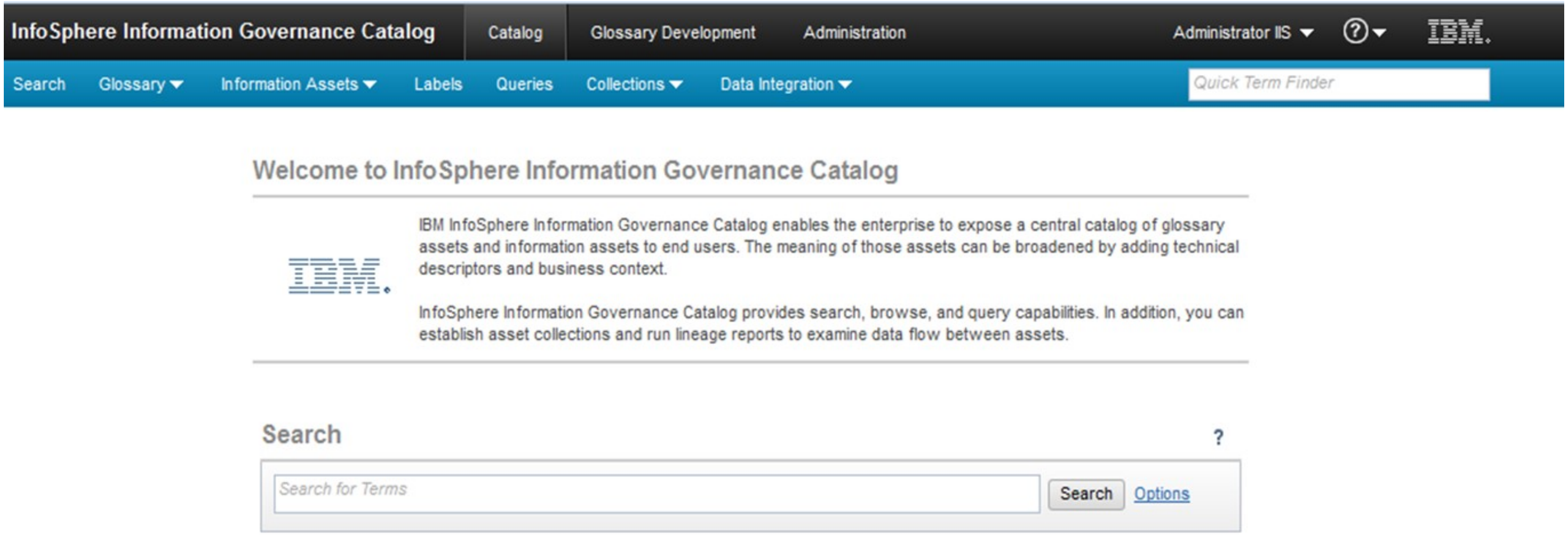
[1] Fraunhofer ISST

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

Demo: IBM Information Governance Catalog (IGC)

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

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

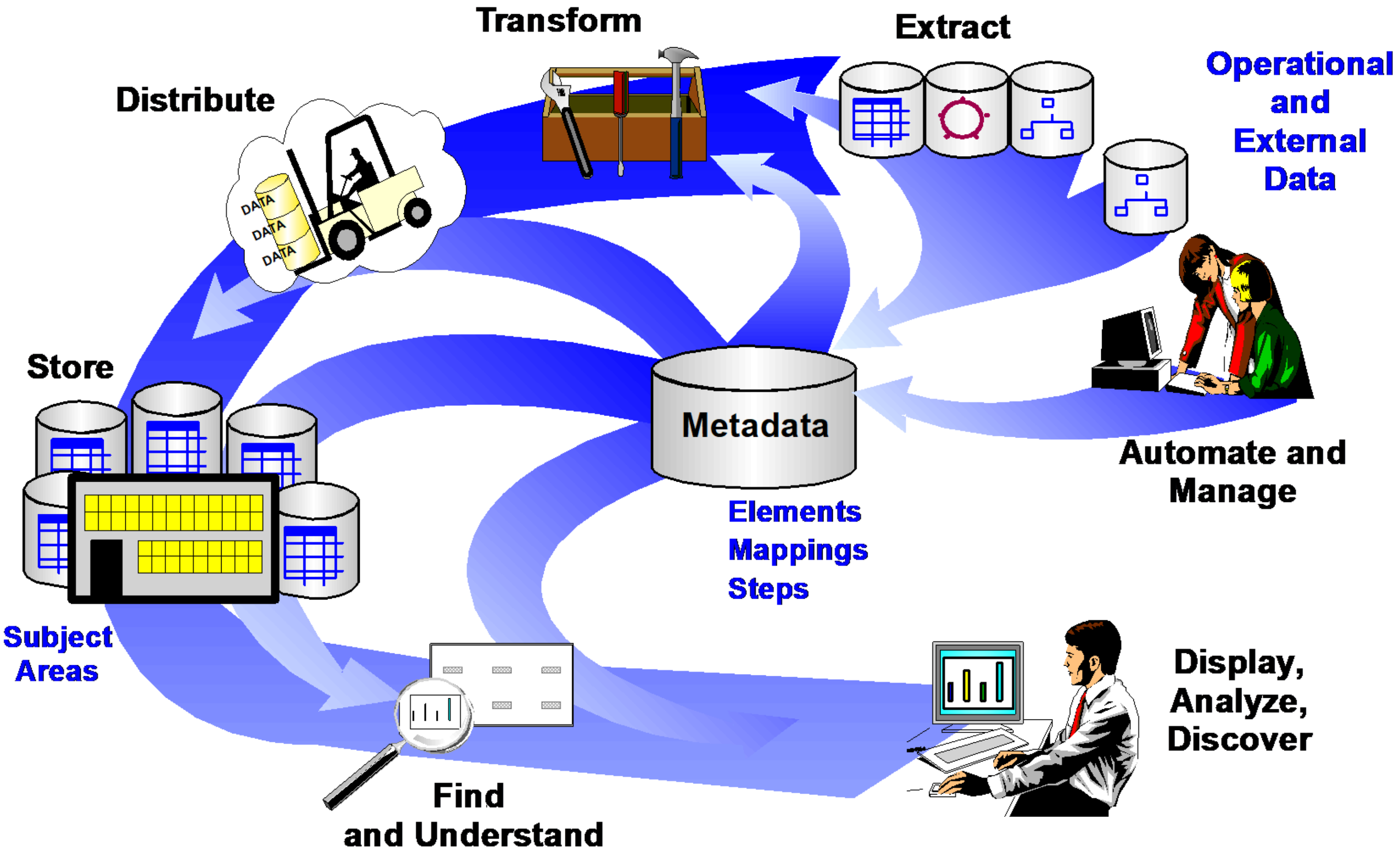


Welcome to InfoSphere Information Governance Catalog

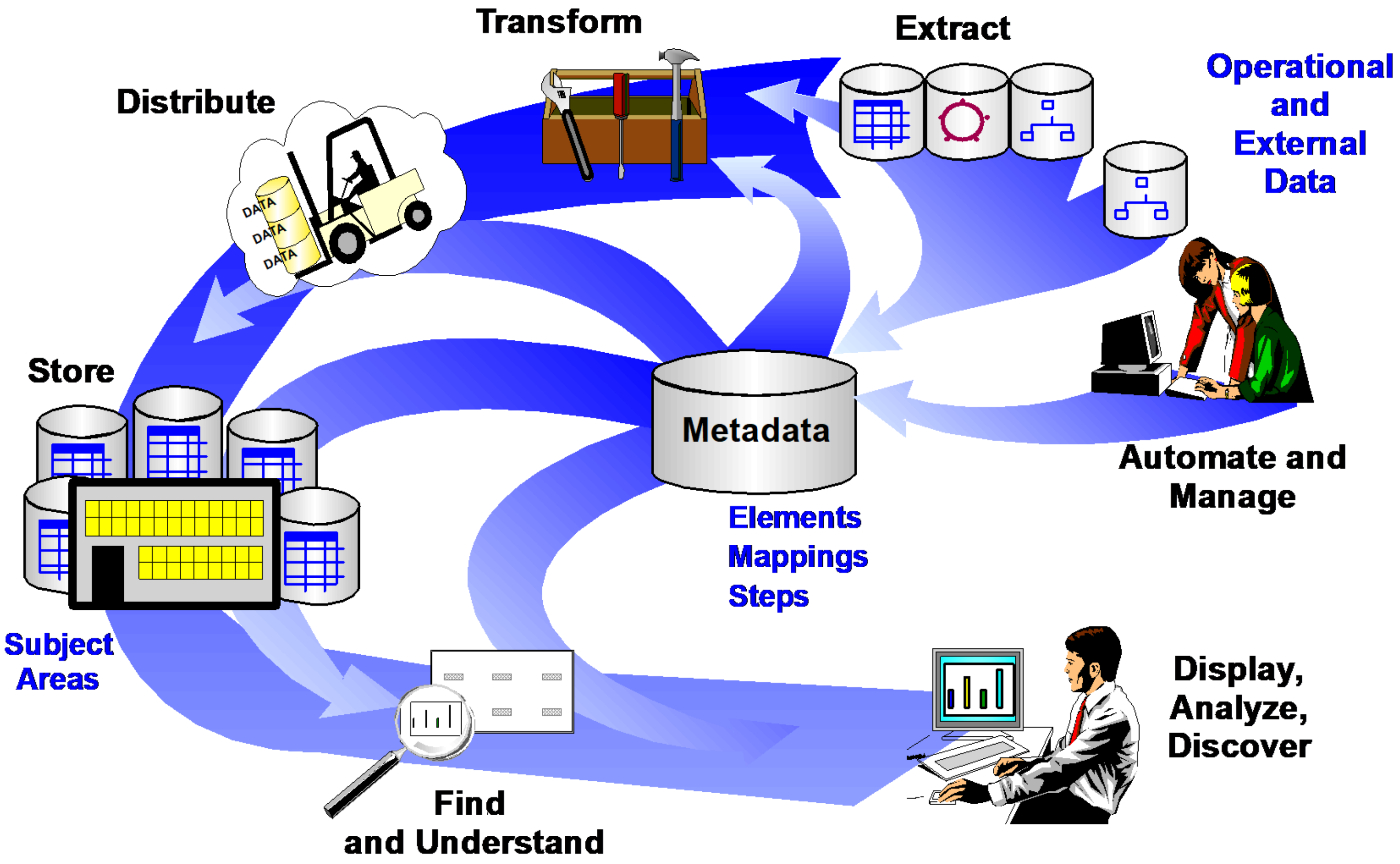
IBM InfoSphere Information Governance Catalog enables the enterprise to expose a central catalog of glossary assets and information assets to end users. The meaning of those assets can be broadened by adding technical descriptors and business context.

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

Turning Data into Information (Part1)

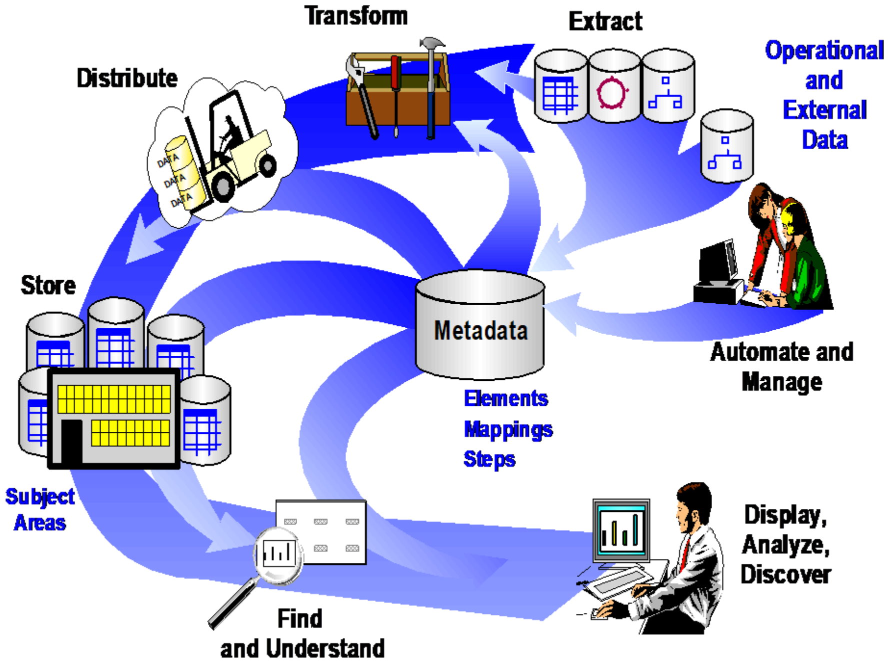


Turning Data into Information (Part2)



What is a Data Warehouse ?

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

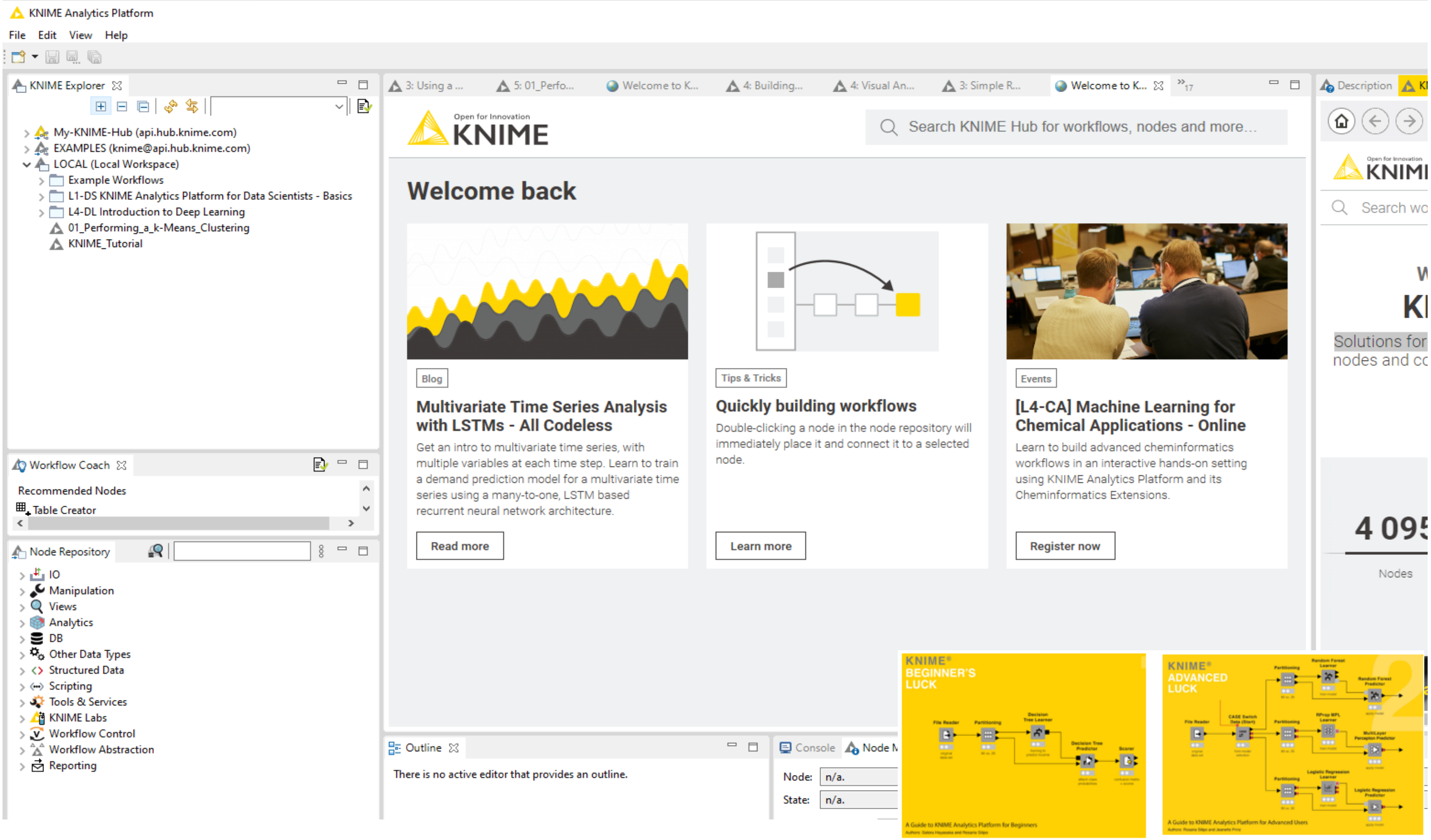


Goal: Turning Data into Information !

Seven Benefits of Data Warehousing

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

Solution Platform for DWH/Analytics: KNIME



KNIME Analytics Platform

File Edit View Help

KNIME Explorer

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

Workflow Coach

Recommended Nodes

Table Creator

Node Repository

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

Tabs: 3: Using a ... | 5: 01_Perfo... | Welcome to K... | 4: Building... | 4: Visual An... | 3: Simple R... | Welcome to K... | »17

Open for Innovation KNIME

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

Welcome back

Blog

Multivariate Time Series Analysis with LSTMs - All Codeless

Get an intro to multivariate time series, with multiple variables at each time step. Learn to train a demand prediction model for a multivariate time series using a many-to-one, LSTM based recurrent neural network architecture.

Read more

Tips & Tricks

Quickly building workflows

Double-clicking a node in the node repository will immediately place it and connect it to a selected node.

Learn more

Events

[L4-CA] Machine Learning for Chemical Applications - Online

Learn to build advanced cheminformatics workflows in an interactive hands-on setting using KNIME Analytics Platform and its Cheminformatics Extensions.

Register now

Outline

There is no active editor that provides an outline.

Console | Node

Node: n/a.

State: n/a.

Description

Search wo

Solutions for

nodes and co

4 095

Nodes

KNIME® BEGINNER'S LUCK
A Guide to KNIME Analytics Platform for Beginners
KNIME® ADVANCED LUCK
A Guide to KNIME Analytics Platform for Advanced Users

First Exercise for DW01

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

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

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

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

Second Exercise for DW01

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

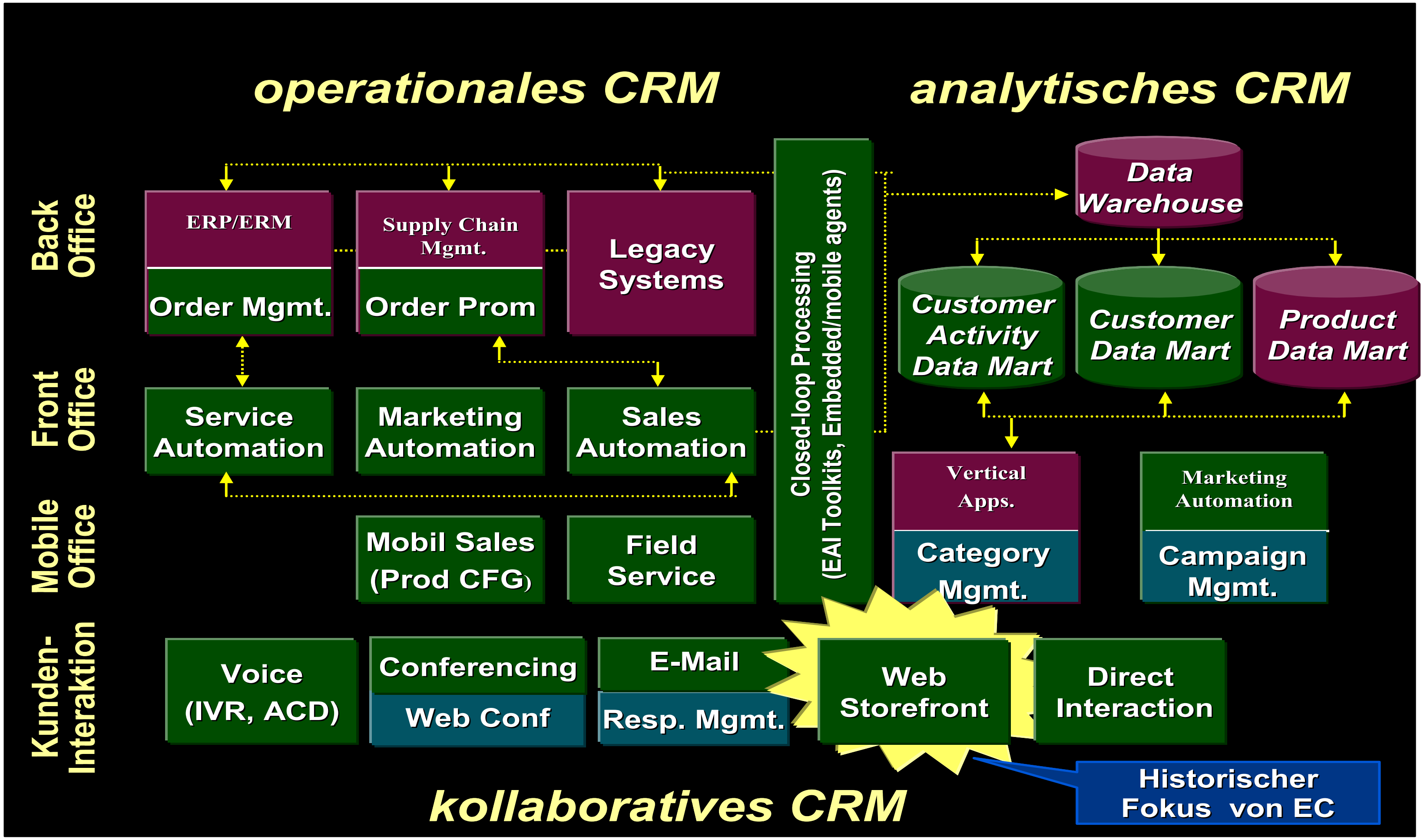
Prepare a report and present it next week; duration = 30 minutes (10 min for each area). Information sources are newspaper or magazine articles or books (see literature list). 3 students.

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

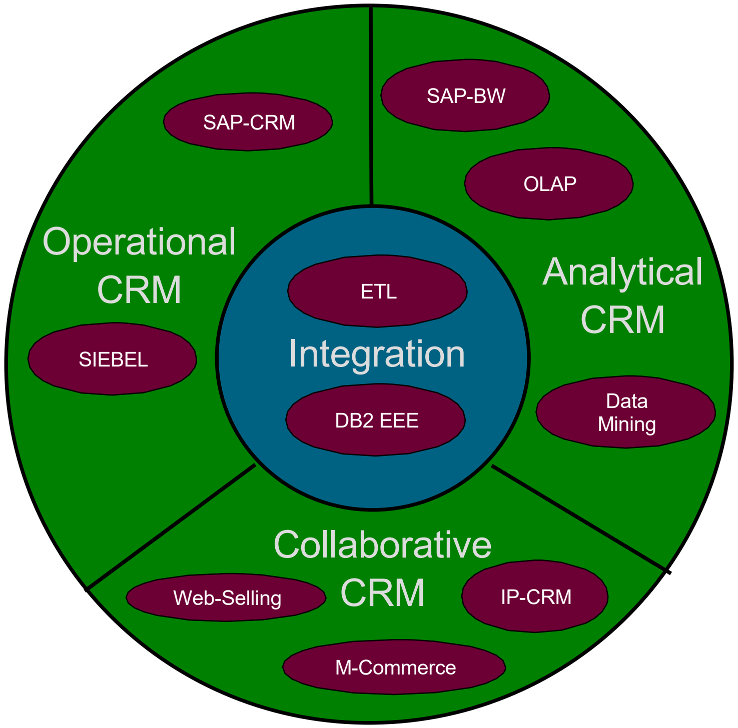
1. Data Warehousing (DWH)
2. Business Intelligence (BI)
3. Customer Relationship Management (CRM)

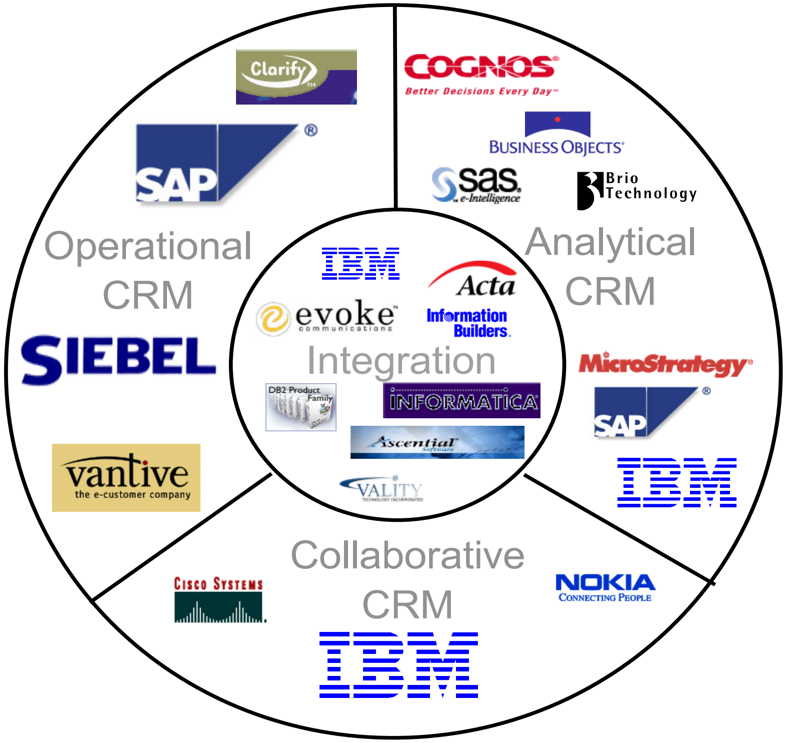
* This exercise is also a task for a Seminar Work (SW)

Hints to E1.2: The BI / CRM Topology



Hints to E1.2: CRM Categories and Tools





Third Exercise for DW01

Exercise 1.3**: Compare two Data Catalogue Tools

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

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

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

Fourth Exercise for DW01

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

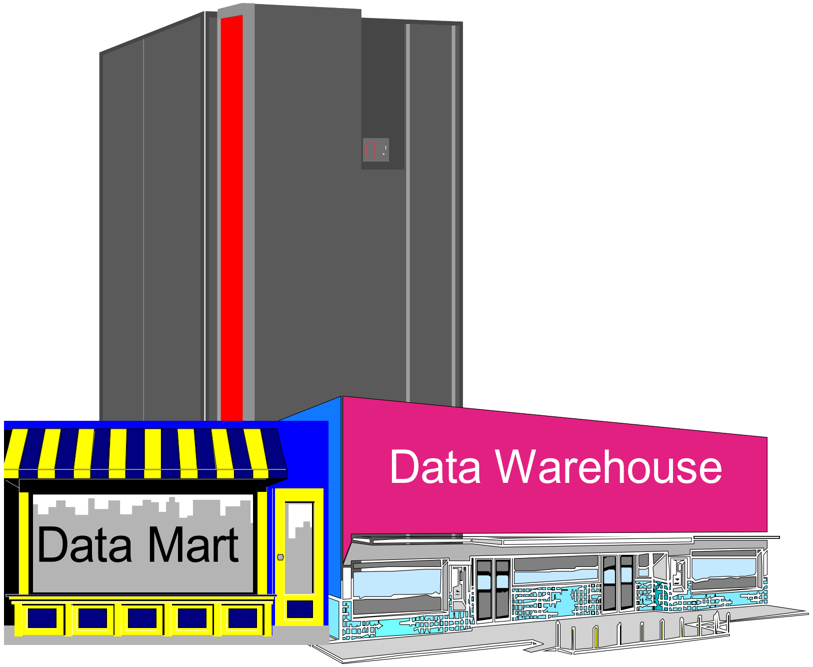
Task: *Install the tool and report about your first experiences and insights. Give answers to the following questions:*

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

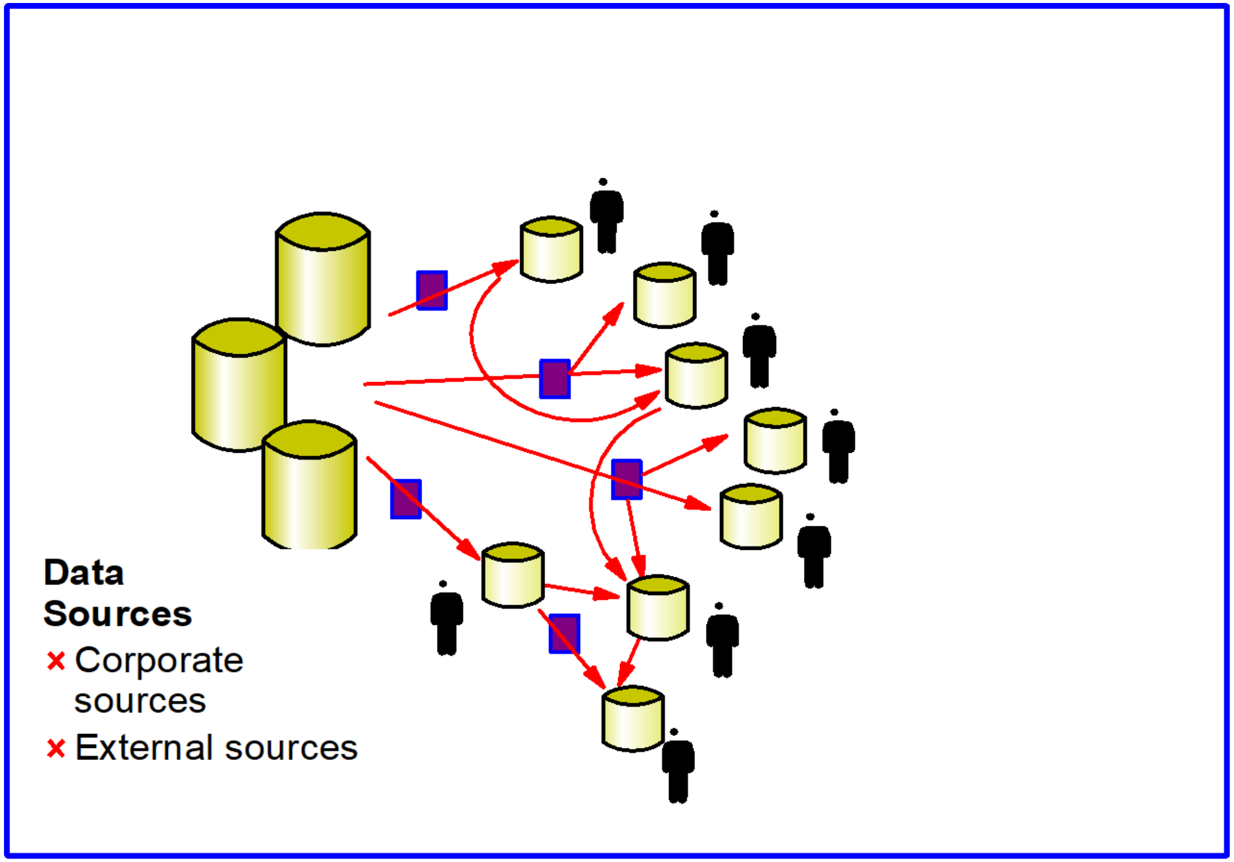
Information source is the KNIME Homepage KNIME | Open for Innovation *and the three mentioned documents in the lesson DW01 (see lesson notes).*

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

DW02 - Introduction to DWH Architecture



Ad-Hoc Evolving DWH Environments



Data Sources

- Corporate sources
- External sources

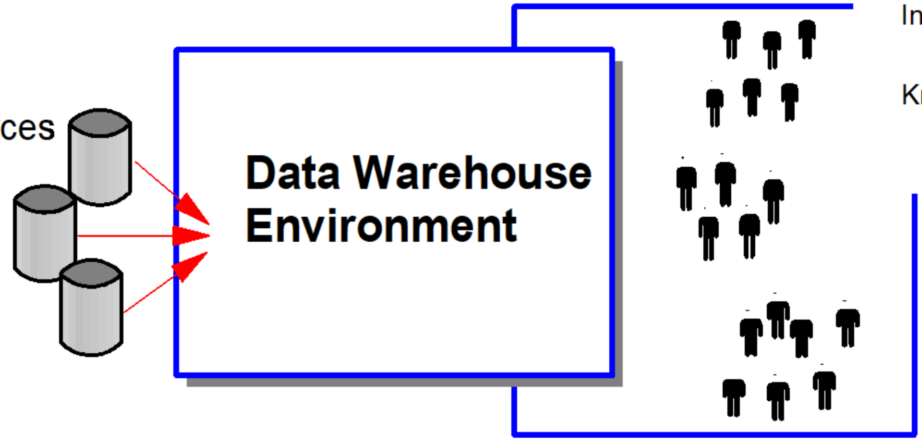
Problems

- Lack of credibility of the data
- Inconsistent information derivation
- Low productivity/High costs
- Complexity

Setting the Scene

Data Sources

- Corporate sources
- External sources



Business Intelligence

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

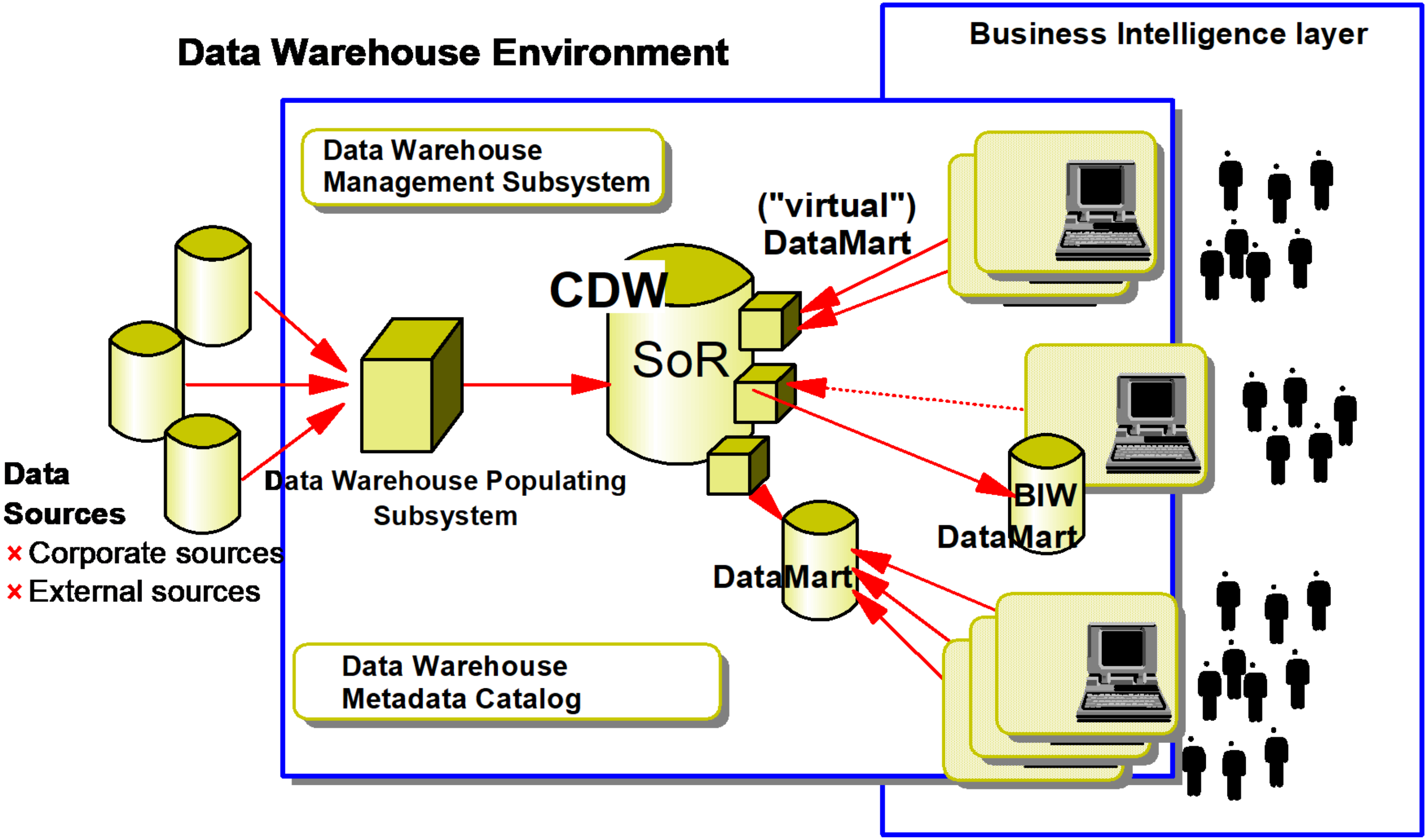
Integrated collection of data
"Corporate memory"
Non-volatile data

Transient data

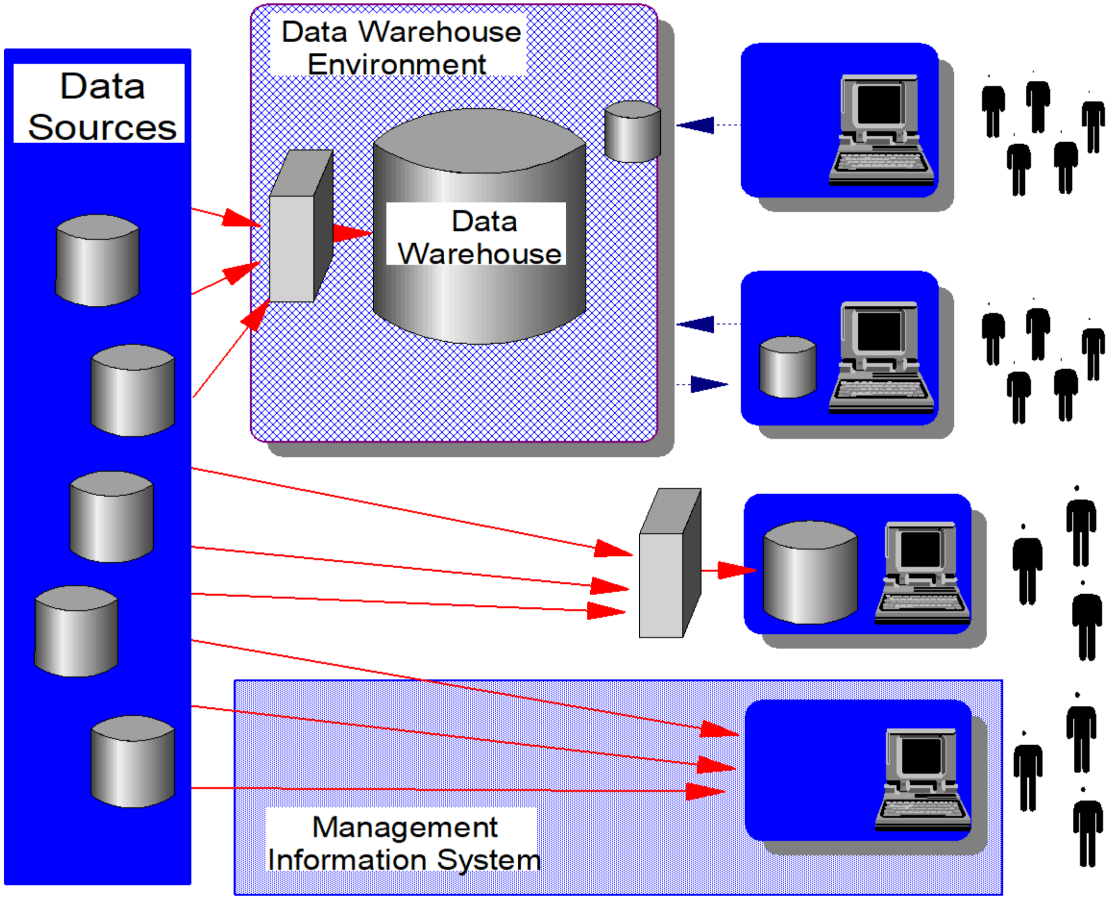
Business Data

Business Information

Setting the Scene (Cont)



DWH - Possible Approaches



Two-tier data warehouse

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

One-tier data warehouse

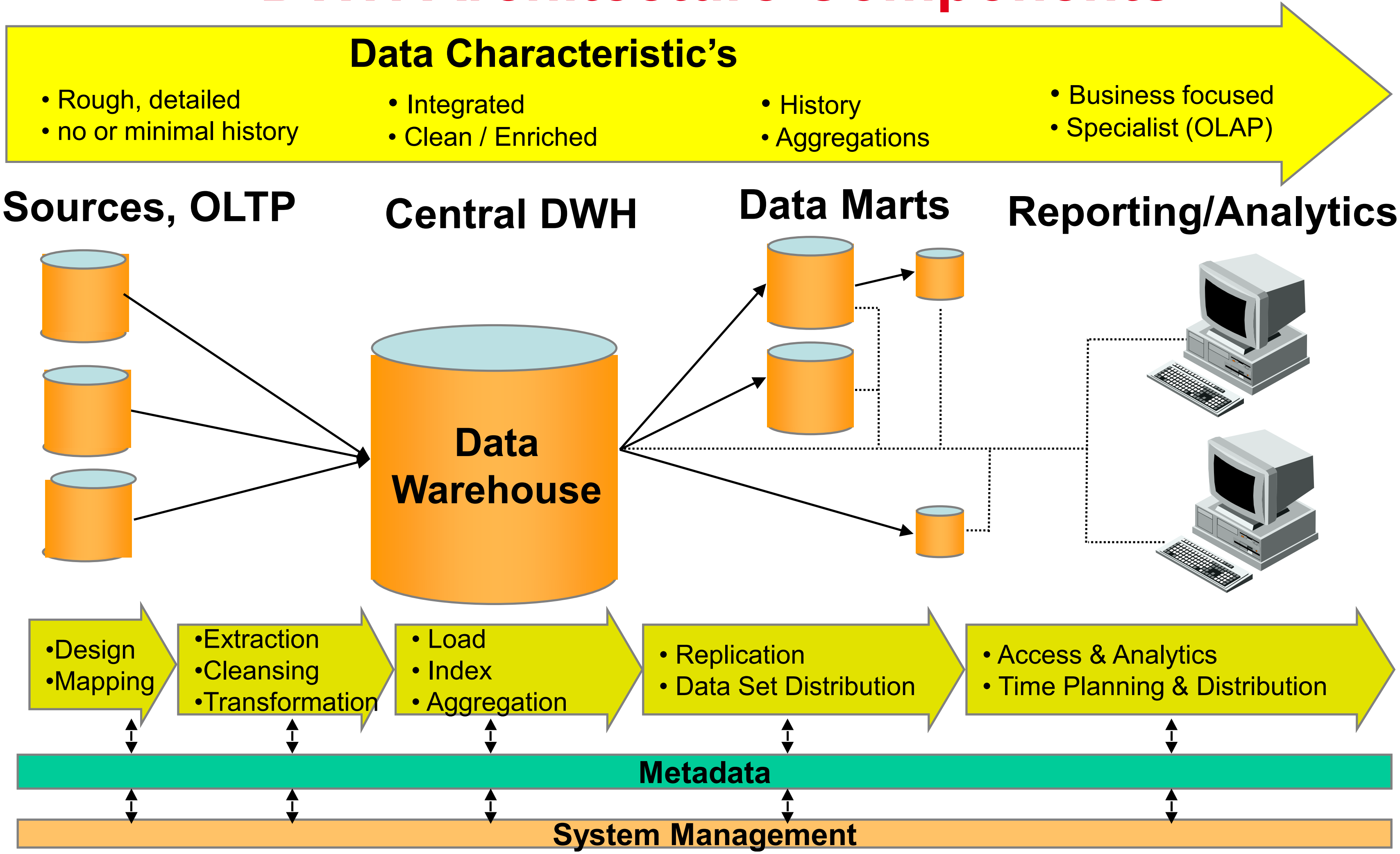
- DataMarts and simple departmental solutions

"Virtual" data warehouse

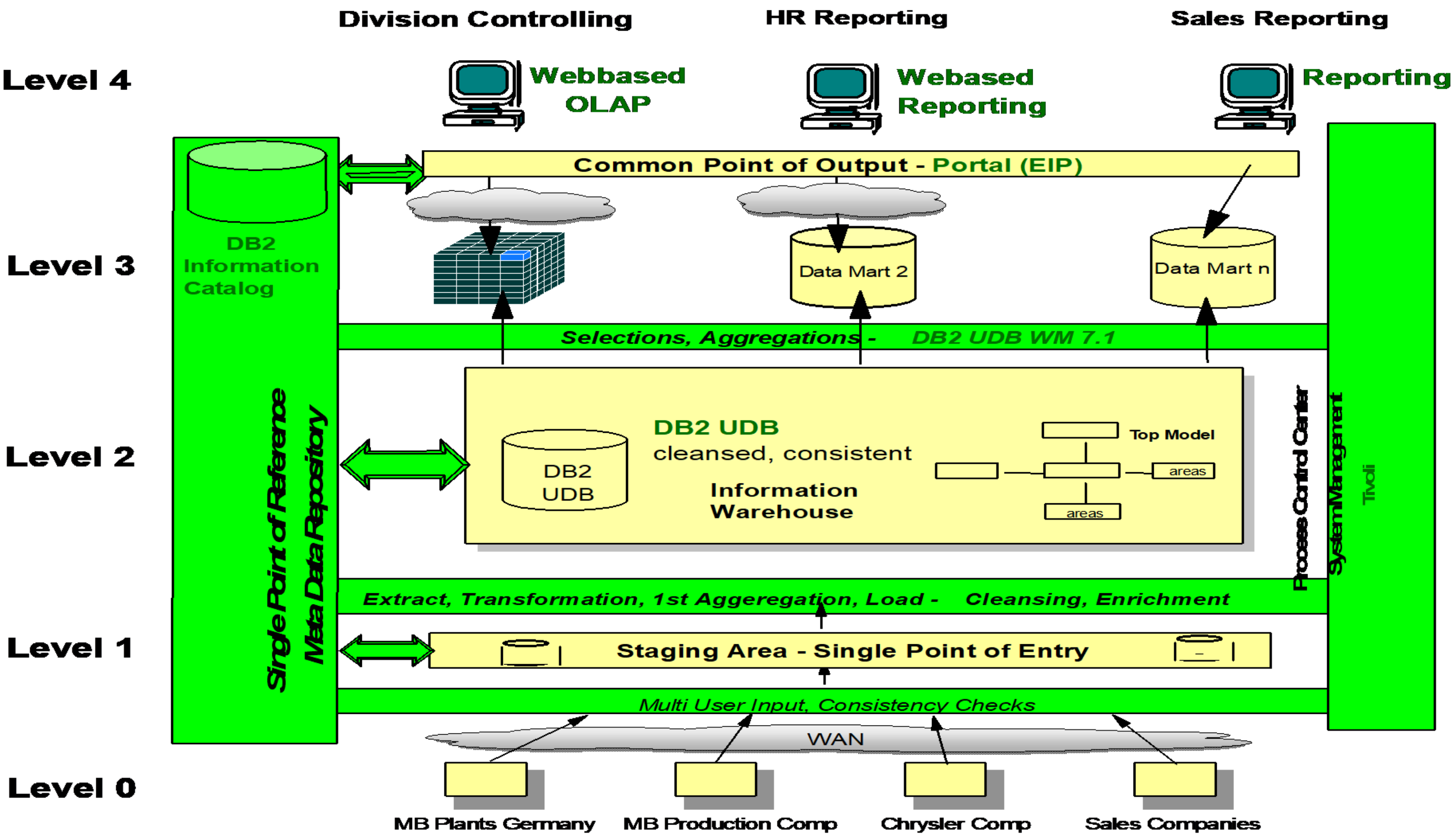
Data Marts or Data Warehouses

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

DWH Architecture Components



DWH Architecture – ‘Big Picture‘ Example



Division Controlling

HR Reporting

Sales Reporting

Level 4

Webbased OLAP

Webased Reporting

Reporting

Common Point of Output - Portal (EIP)

Level 3

DB2 Information Catalog

Data Mart 2

Data Mart n

Selections, Aggregations - DB2 UDB WM 7.1

Level 2

Single Point of Reference

Meta Data Repository

DB2 UDB

DB2 UDB
cleansed, consistent

Information Warehouse

Top Model

areas

areas

Process Control Center

System Management

Tivoli

Extract, Transformation, 1st Aggeregation, Load - Cleansing, Enrichment

Level 1

Staging Area - Single Point of Entry

Multi User Input, Consistency Checks

WAN

Level 0

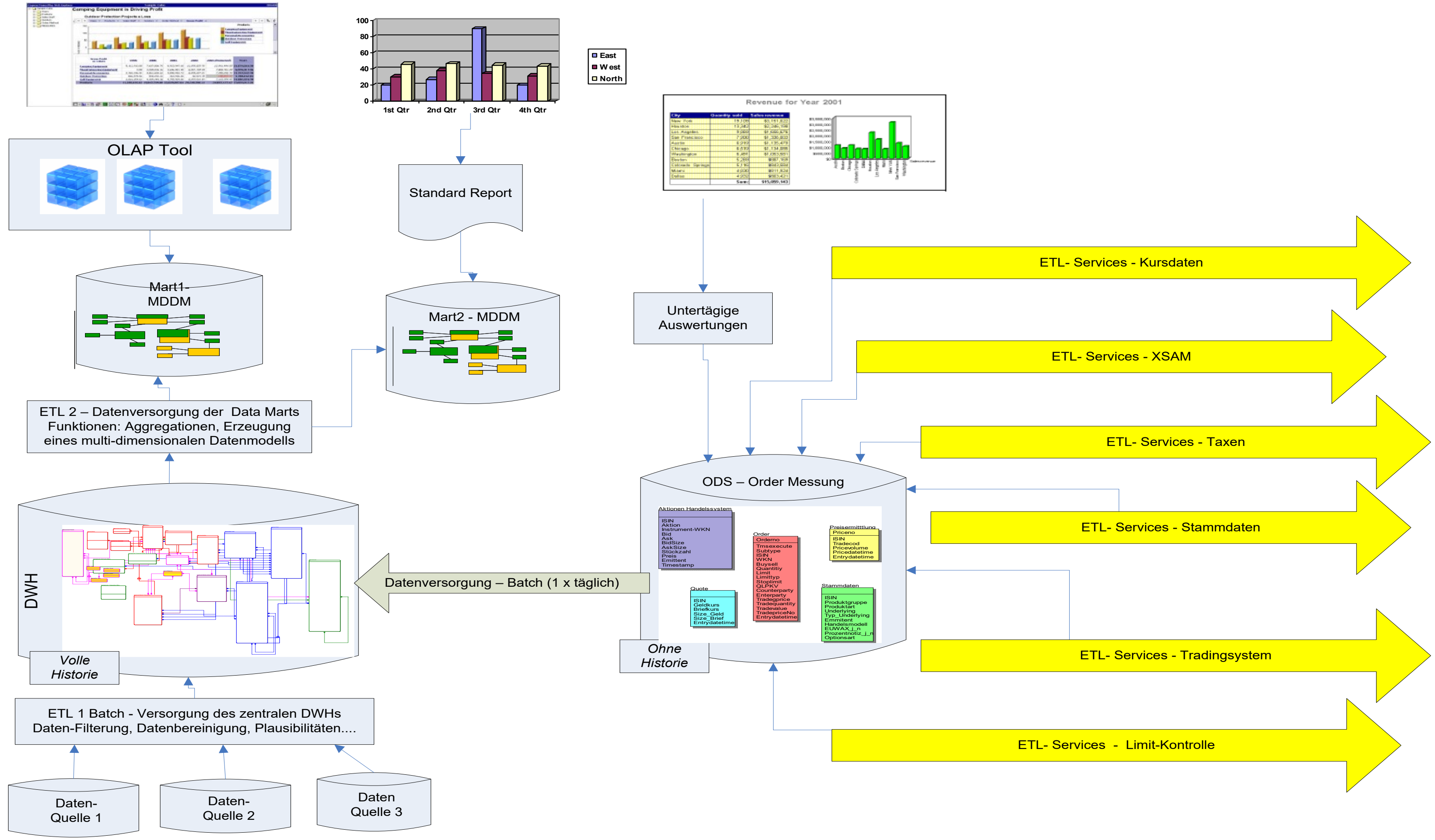
MB Plants Germany

MB Production Comp

Chrysler Comp

Sales Companies

Example of a Financial Market DWH



OLAP Tool

Standard Report

Mart1-MDDM

Mart2 - MDDM

Untertägige Auswertungen

ETL 2 – Datenversorgung der Data Marts
Funktionen: Aggregationen, Erzeugung eines multi-dimensionalen Datenmodells

DWH

Volle Historie

Datenversorgung – Batch (1 x täglich)

ODS – Order Messung

Aktionen Handelssystem: ISIN, Aktion, Instrument-WKN, Bid, Ask, BidSize, AskSize, Stückzahl, Preis, Emittent, Timestamp

Order: Orderno, Tmsexecute, Subtype, ISIN, WKN, Buysell, Quantitiy, Limit, Limittyp, Stoplimit, QLPKV, Counterparty, Enterparty, Tradegprice, Tradequantity, Tradevalue, TradepriceNo, Entrydatetime

Preisermitttlung: Priceno, ISIN, Tradecod, Pricevolume, Pricedatetime, Entrydatetime

Quote: ISIN, Geldkurs, Briefkurs, Size_Geld, Size_Brief, Entrydatetime

Stammdaten: ISIN, Produktgruppe, Produktart, Underlying, Typ_Underlying, Emmitent, Handelsmodell, EUWAX_j_n, Prozentnotiz_j_n, Optionsart

Ohne Historie

ETL- Services - Kursdaten

ETL- Services - XSAM

ETL- Services - Taxen

ETL- Services - Stammdaten

ETL- Services - Tradingsystem

ETL- Services - Limit-Kontrolle

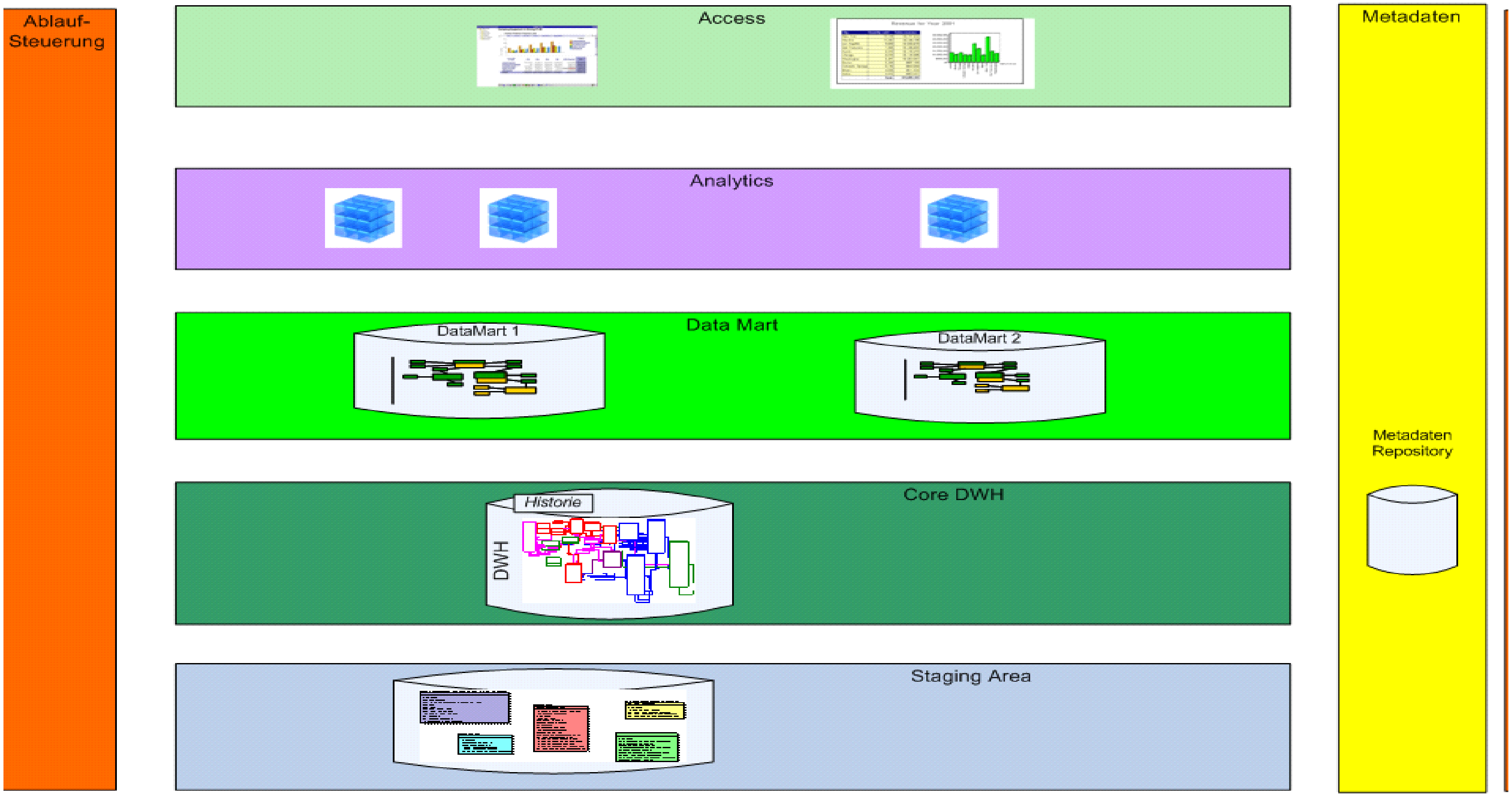
ETL 1 Batch - Versorgung des zentralen DWHs
Daten-Filterung, Datenbereinigung, Plausibilitäten....

Daten-Quelle 1

Daten-Quelle 2

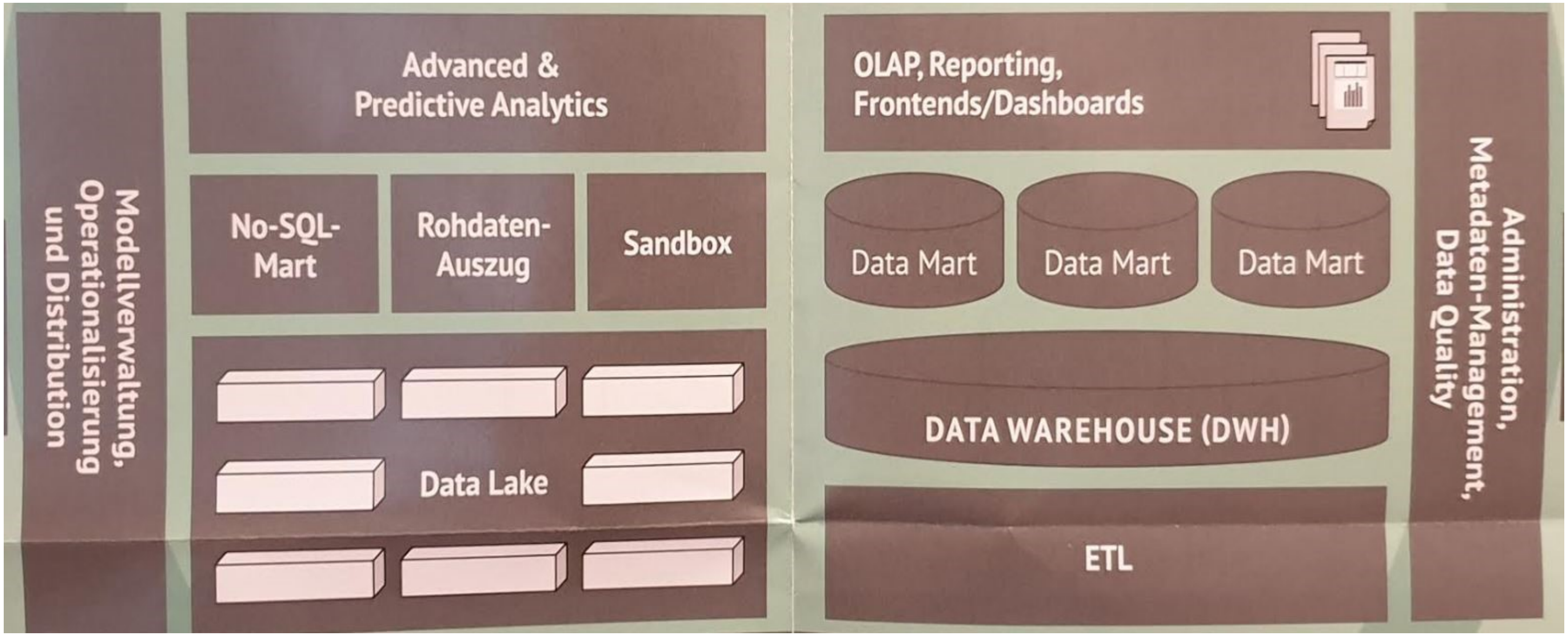
Daten Quelle 3

DWH Architecture – Data Layer Concept

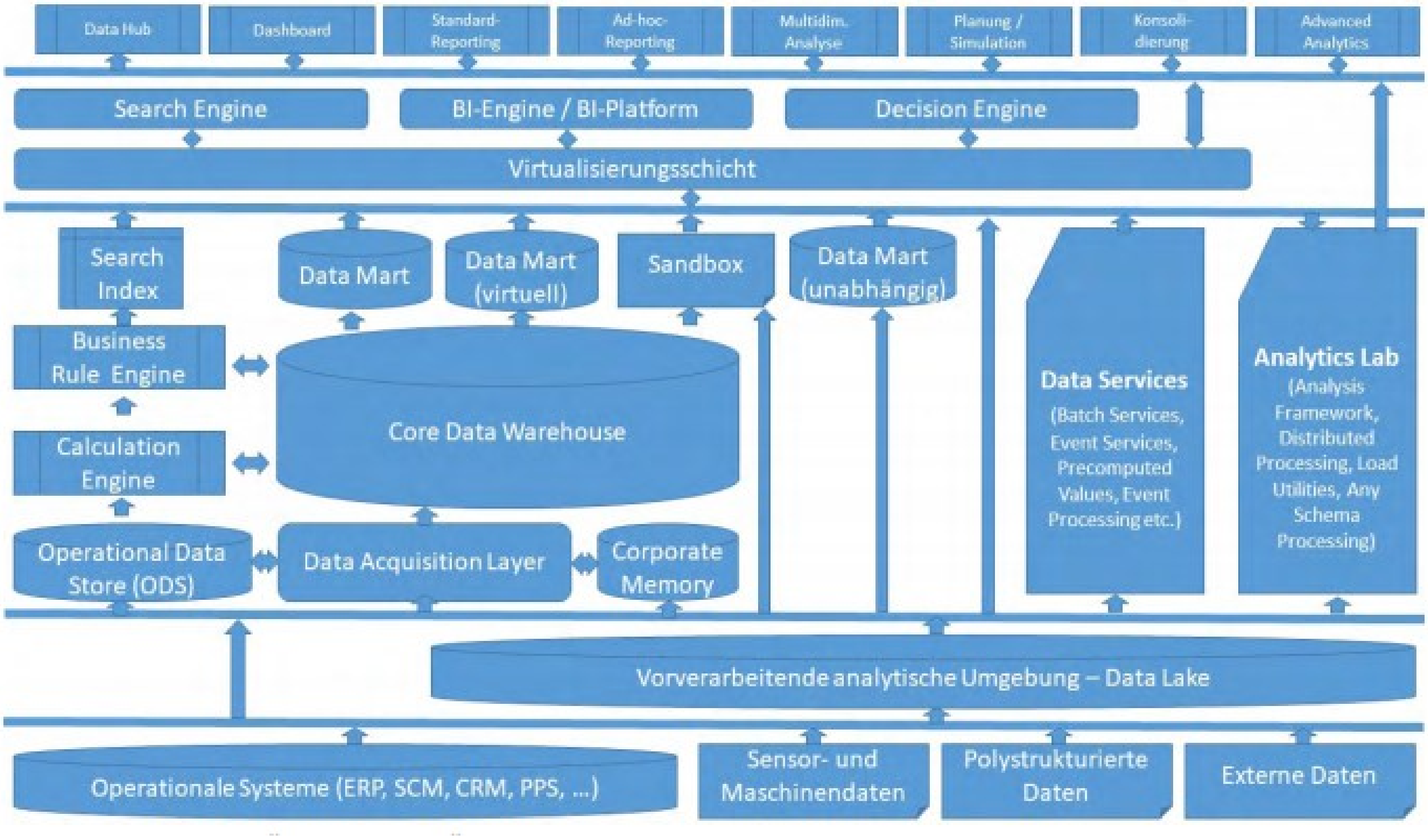


Modern Data Architecture – Big Data & Cloud

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


Modellverwaltung, Operationalisierung und Distribution
Advanced & Predictive Analytics
No-SQL-Mart
Rohdaten-Auszug
Sandbox
Data Lake
OLAP, Reporting, Frontends/Dashboards
Data Mart
Data Mart
Data Mart
DATA WAREHOUSE (DWH)
ETL
Administration, Metadaten-Management, Data Quality

Modern Data Arch. – Data Lake Integration



Data Hub
Dashboard
Standard-Reporting
Ad-hoc-Reporting
Multidim. Analyse
Planung / Simulation
Konsoli-dierung
Advanced Analytics

Search Engine
BI-Engine / BI-Platform
Decision Engine

Virtualisierungsschicht

Search Index
Data Mart
Data Mart (virtuell)
Sandbox
Data Mart (unabhängig)

Business Rule Engine
Calculation Engine
Core Data Warehouse

Data Services
(Batch Services, Event Services, Precomputed Values, Event Processing etc.)

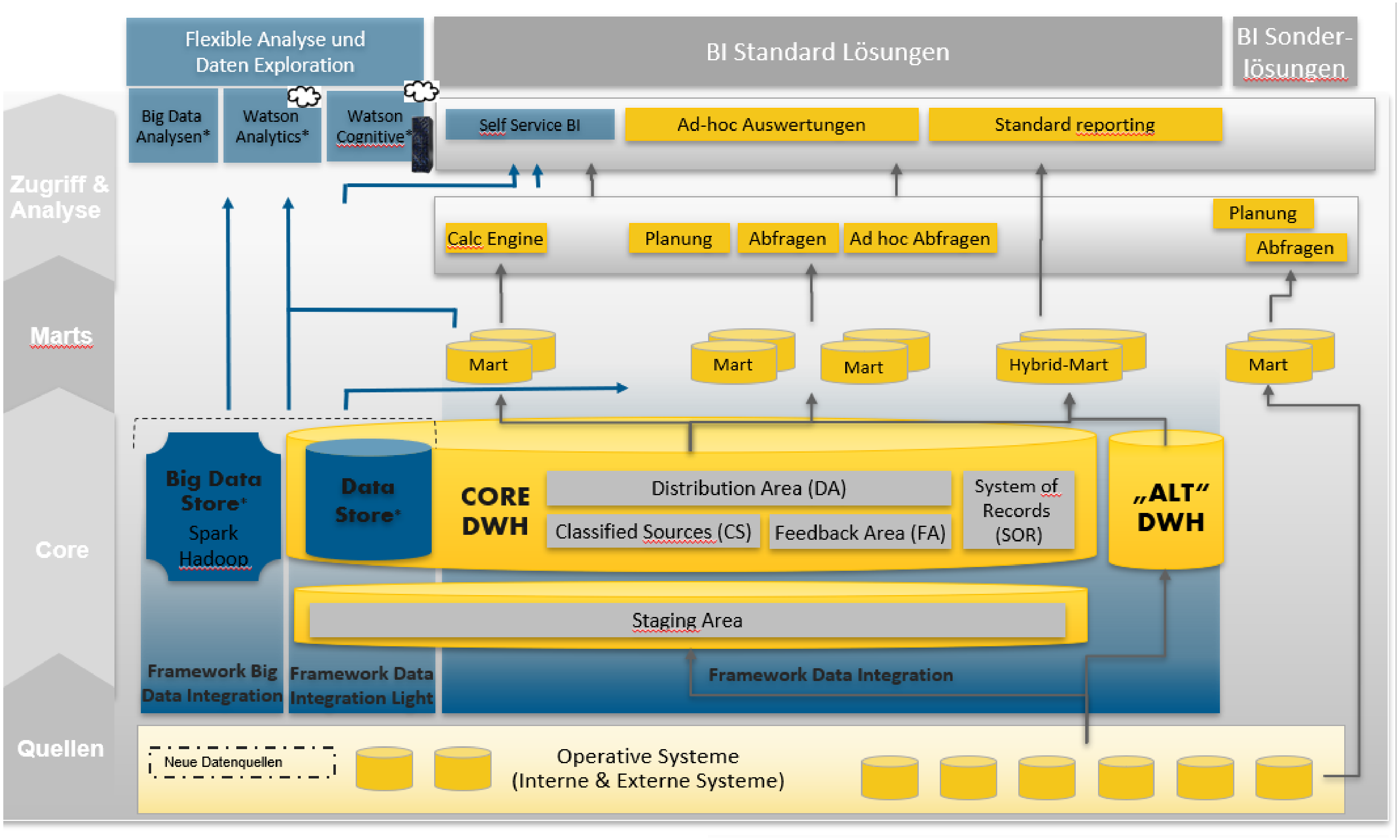
Analytics Lab
(Analysis Framework, Distributed Processing, Load Utilities, Any Schema Processing)

Operational Data Store (ODS)
Data Acquisition Layer
Corporate Memory

Vorverarbeitende analytische Umgebung – Data Lake

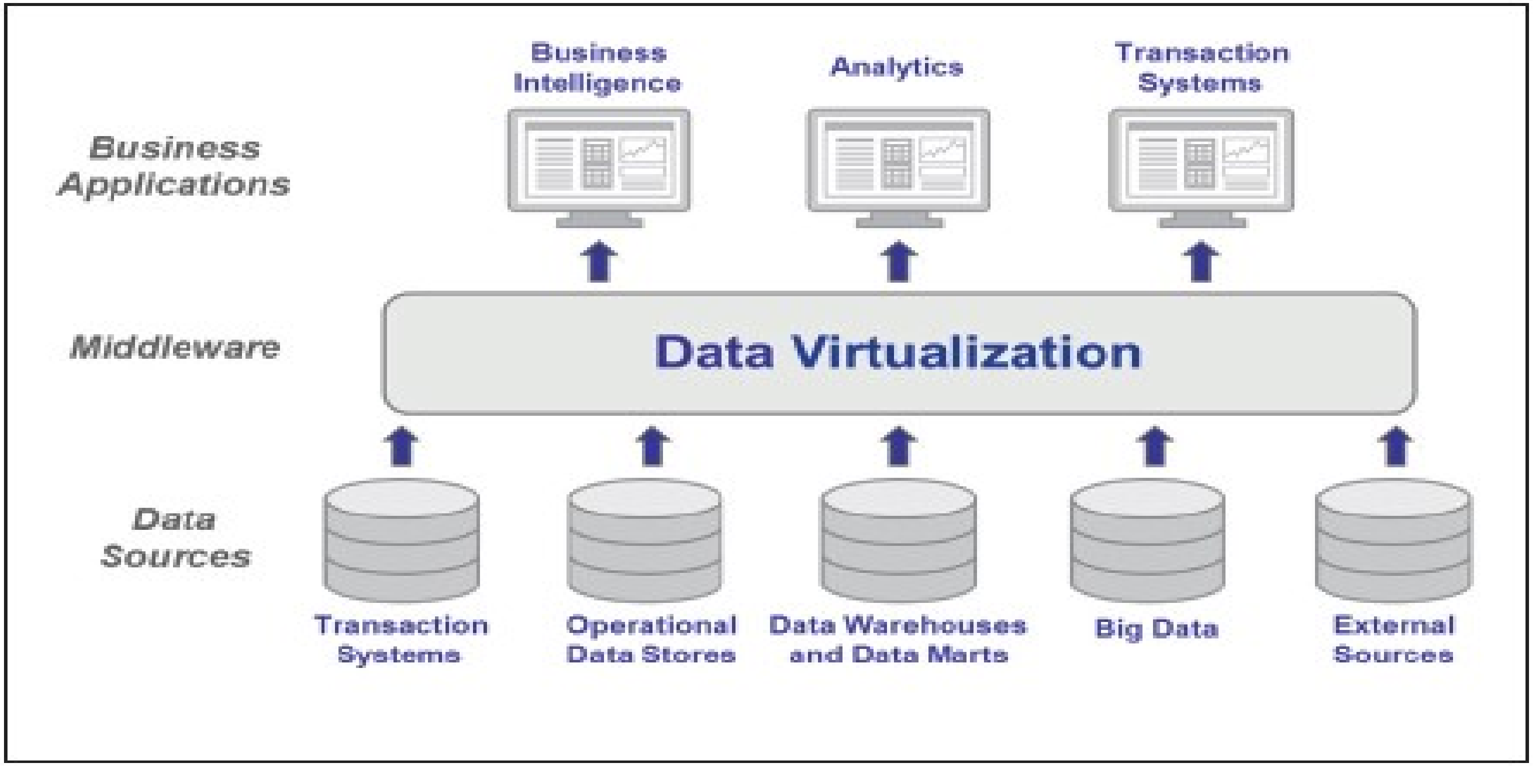
Operationale Systeme (ERP, SCM, CRM, PPS, ...)
Sensor- und Maschinendaten
Polystrukturierte Daten
Externe Daten

Modern Arch. - Example of a German Insurer



Modern Data Arch. – Virtualization Concept

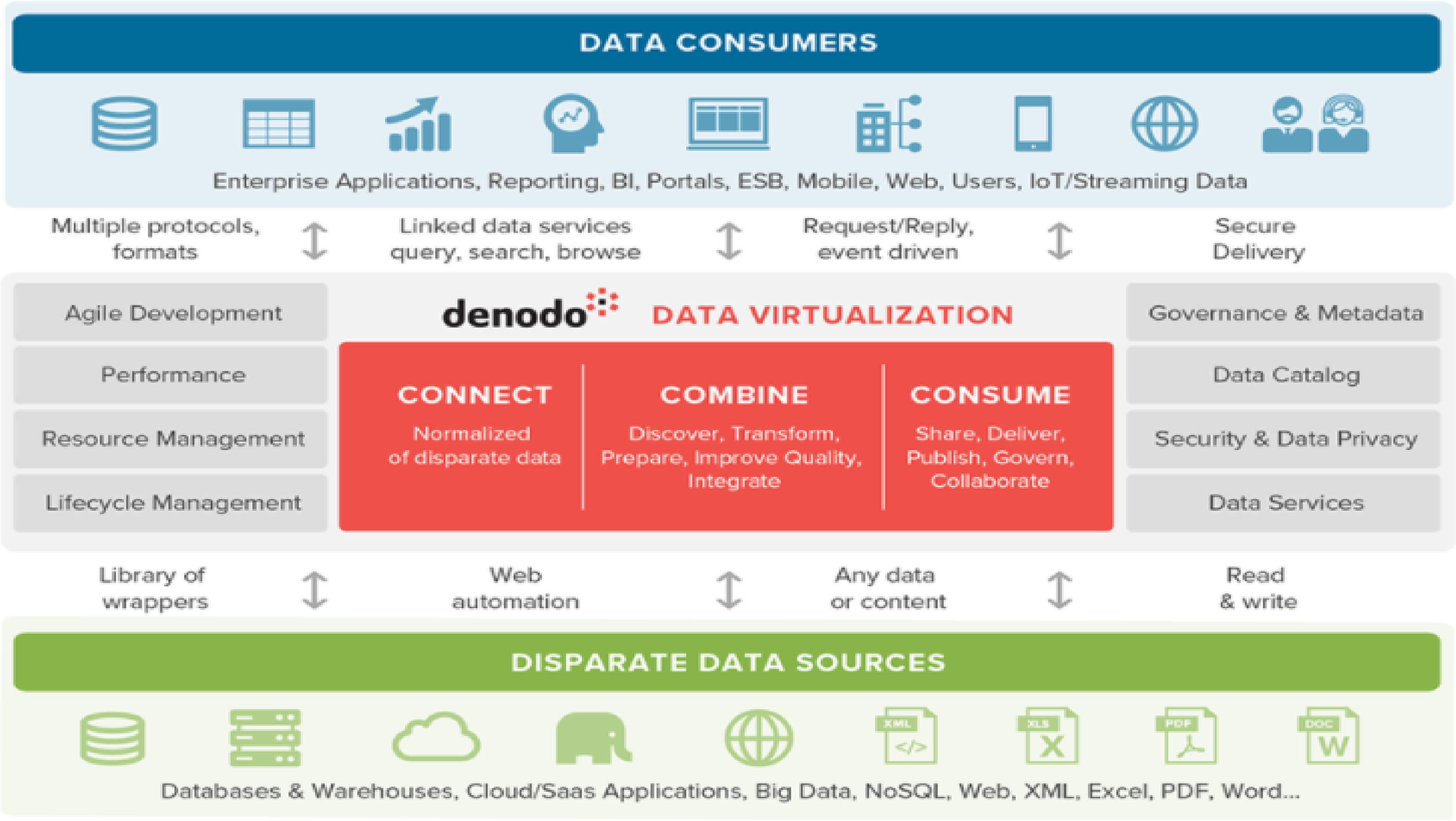
Data virtualization provides a virtual approach to accessing, managing and delivering data without physically replicating it.



See in [DHBW-Moodle]: eBook_Data_Virtualization_Going_Beyond_Traditional_Data_Integration.pdf

Data Virtualization Tool – Denodo Platform

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



DATA CONSUMERS

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

Multiple protocols, formats

Linked data services query, search, browse

Request/Reply, event driven

Secure Delivery

denodo DATA VIRTUALIZATION

Agile Development

Performance

Resource Management

Lifecycle Management

CONNECT — Normalized of disparate data

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

CONSUME — Share, Deliver, Publish, Govern, Collaborate

Governance & Metadata

Data Catalog

Security & Data Privacy

Data Services

Library of wrappers

Web automation

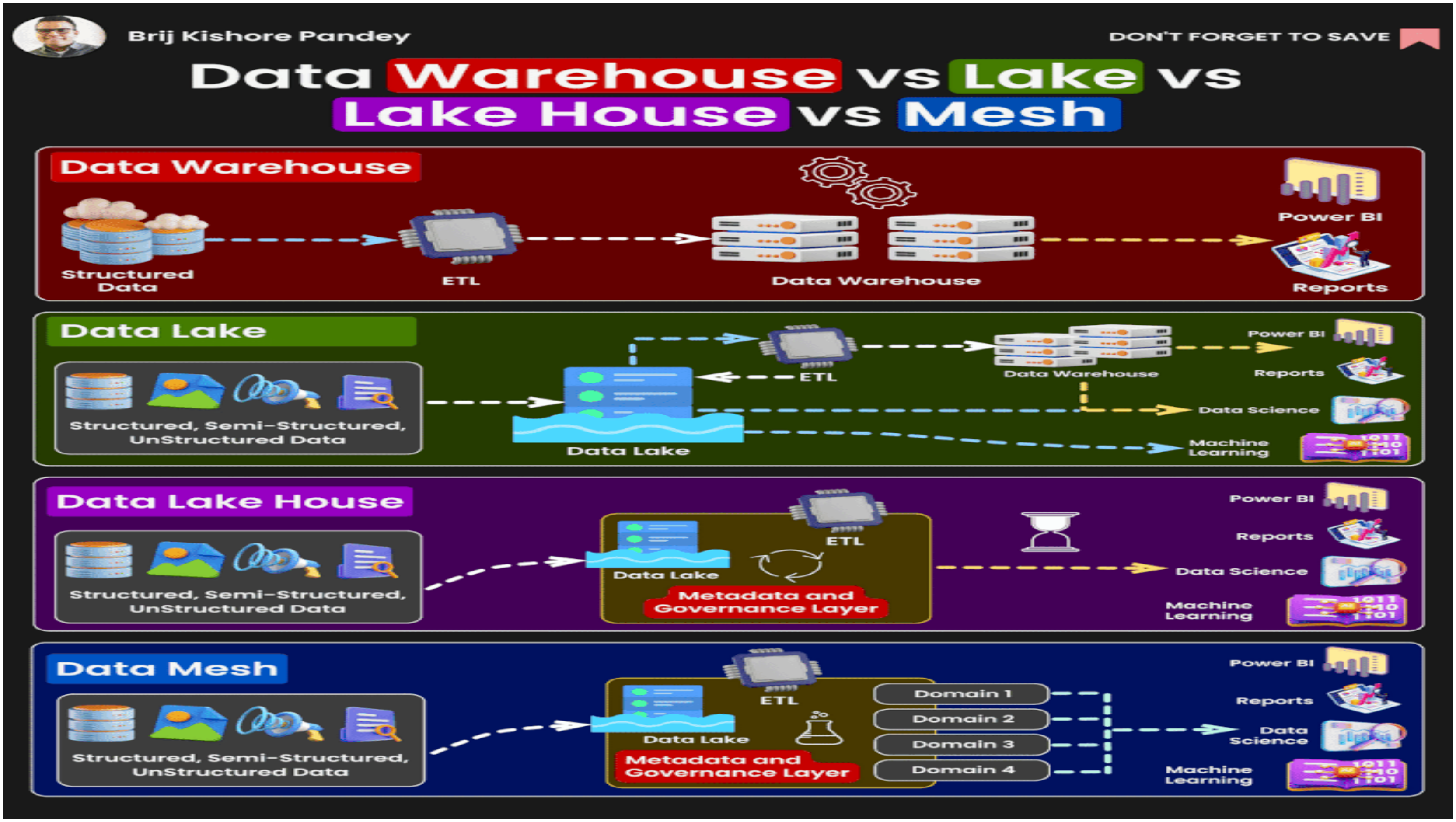
Any data or content

Read & write

DISPARATE DATA SOURCES

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

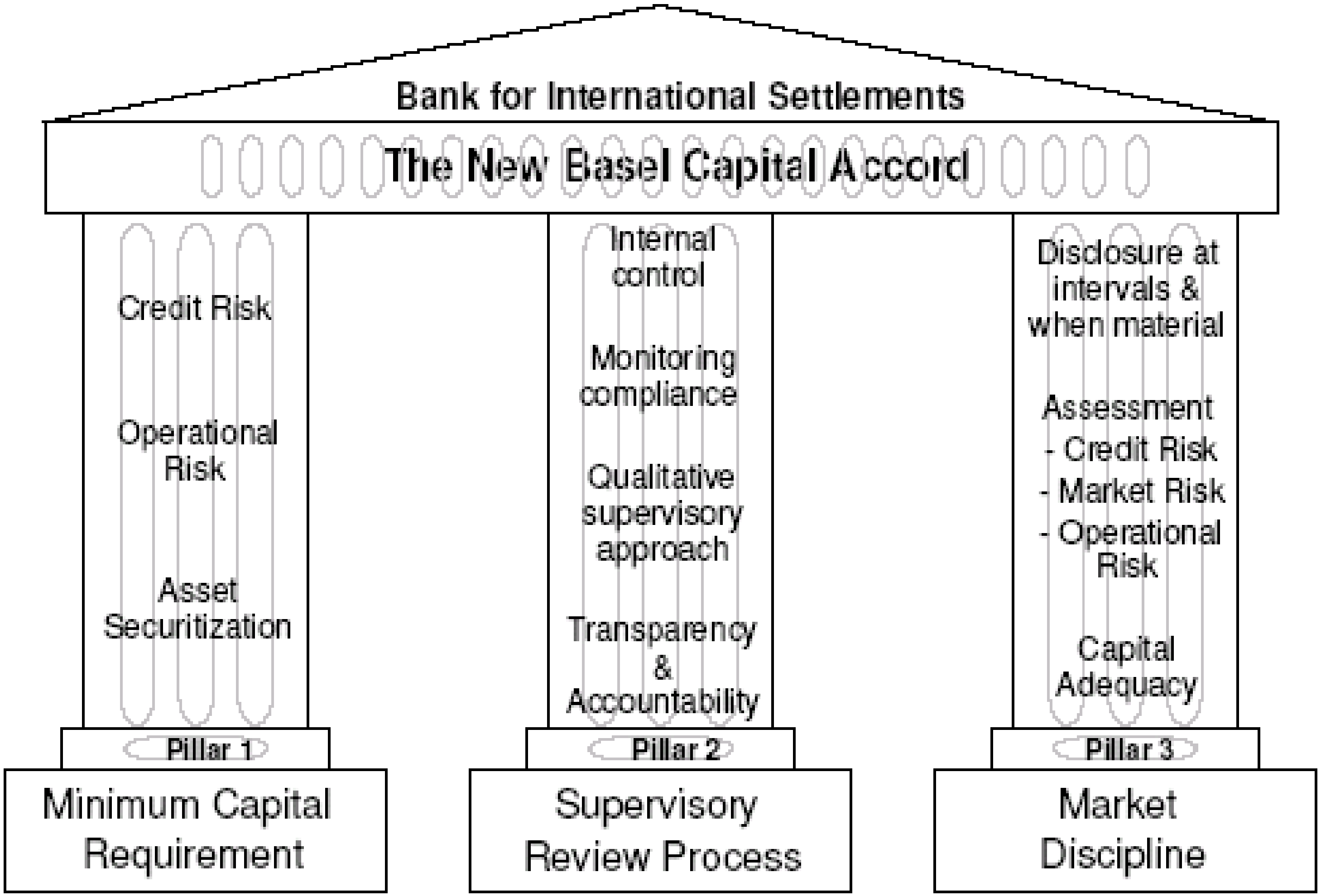
Modern Data Architecture – An Overview



*The Data Visualization Architecture is missing in this picture.

https://github.com/HVoellinger/Data-Warehouse-DWH---Concepts-Applications/blob/main/images/DWH_DLake_DMesh.gif

Use Case I – Basel II (Definition)



Bank for International Settlements

The New Basel Capital Accord

Credit Risk

Operational Risk

Asset Securitization

Pillar 1

Minimum Capital Requirement

Internal control

Monitoring compliance

Qualitative supervisory approach

Transparency & Accountability

Pillar 2

Supervisory Review Process

Disclosure at intervals & when material

Assessment
- Credit Risk
- Market Risk
- Operational Risk

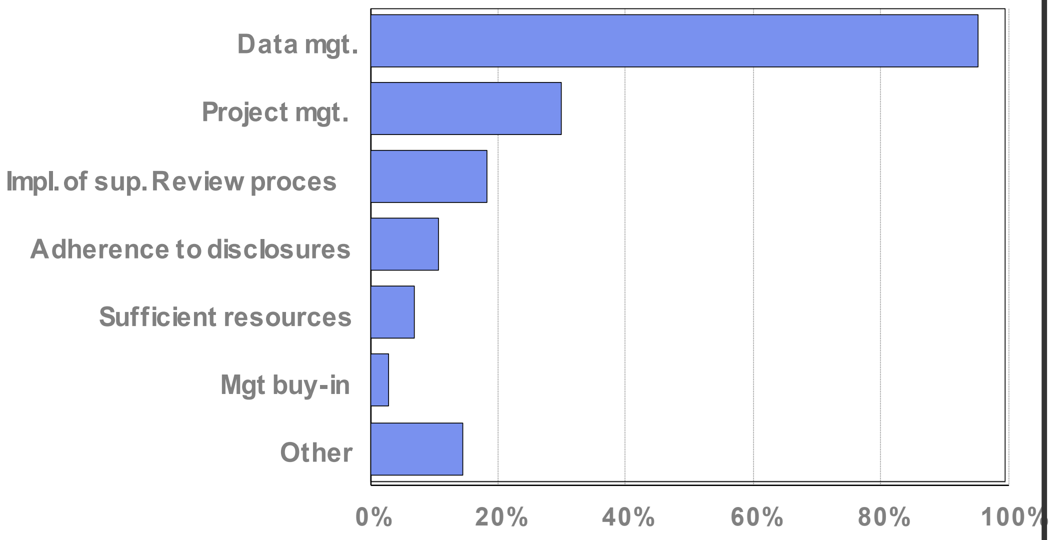
Capital Adequacy

Pillar 3

Market Discipline

Basel II - key challenges – Systems & Data Management

Data Management is the key challenge in meeting Basel II

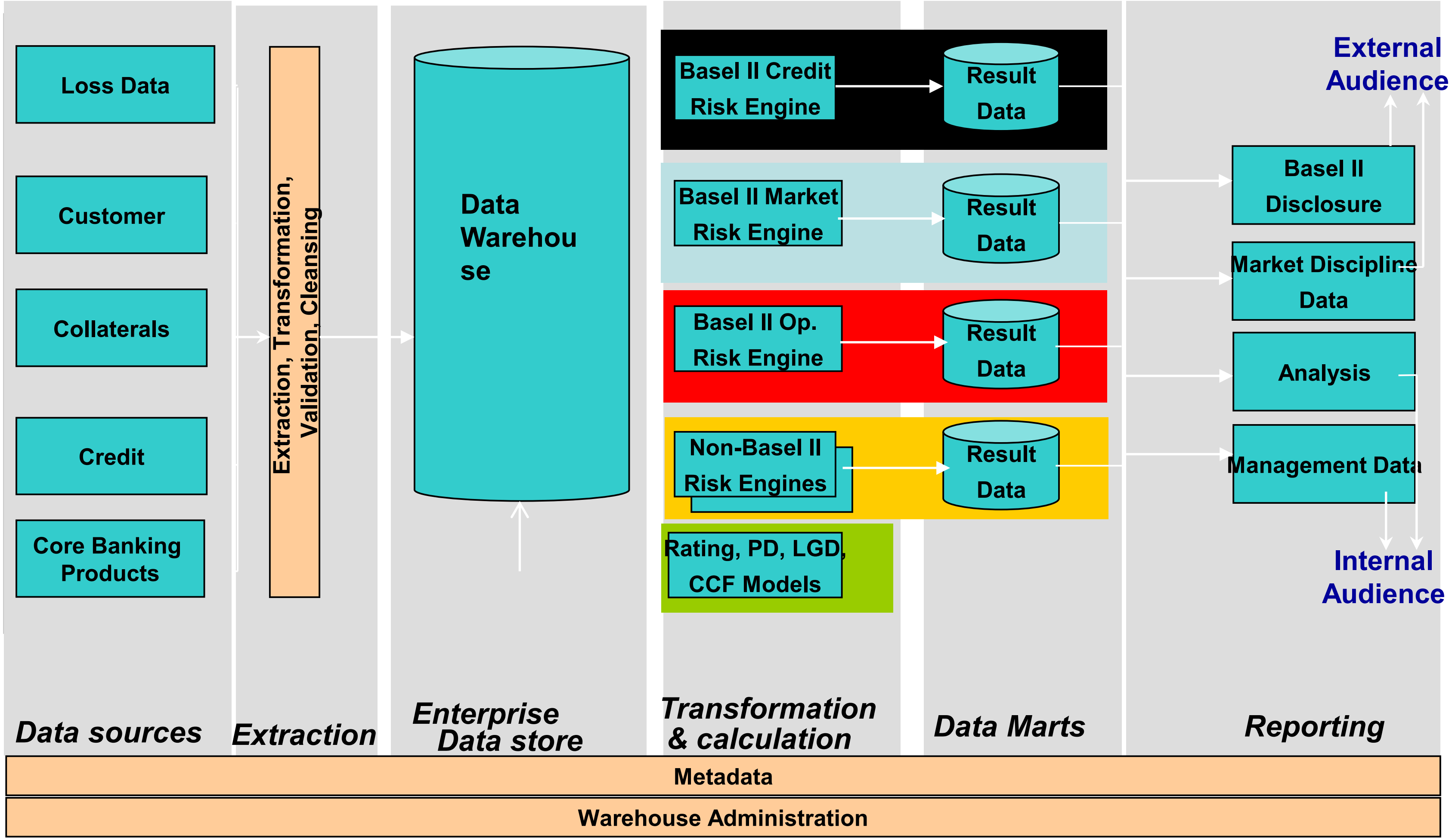


Source: IBM Institute for Business Value analysis, Banks and Basel II: How Prepared Are They?, October 2002 interviews with 32 Financial institutions worldwide

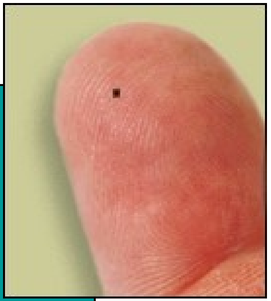
10 Common signs of unstable data foundation

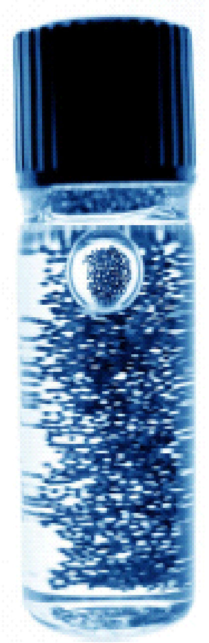
1. There's no single enterprise view of data
2. Inability to gather data for as yet unspecified reporting requirements.
3. Senior Management requests for information require intensive manual effort to respond, and far longer than desired.
4. Multiple databases or spreadsheets storing similar data; no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

Basel II - 6 Tier Reference Architecture



Use Case II – RFID Problem





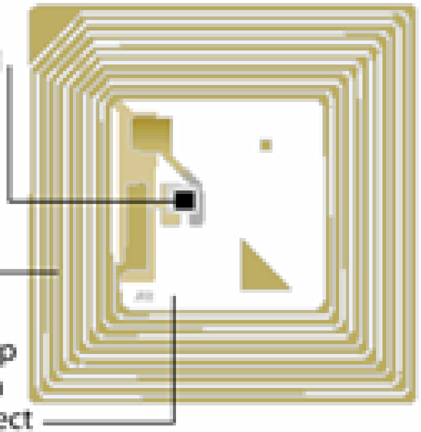
Tags	
Active	Includes a power source to help transmit a signal
Passive	No power to transmit signal; relies on readers
Frequency	Radio wave frequency at which signals are transmitted (Telephone example: 900 Mhz, 2.4 Ghz, 5.8 Ghz)
Data Capacity	Many options, will depend on application
Antenna	Device attached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

Readers

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

Antenna Device attached to a reader which helps transmit radio signals and captures "scan" readings

RFID tags are made up of three parts:*



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

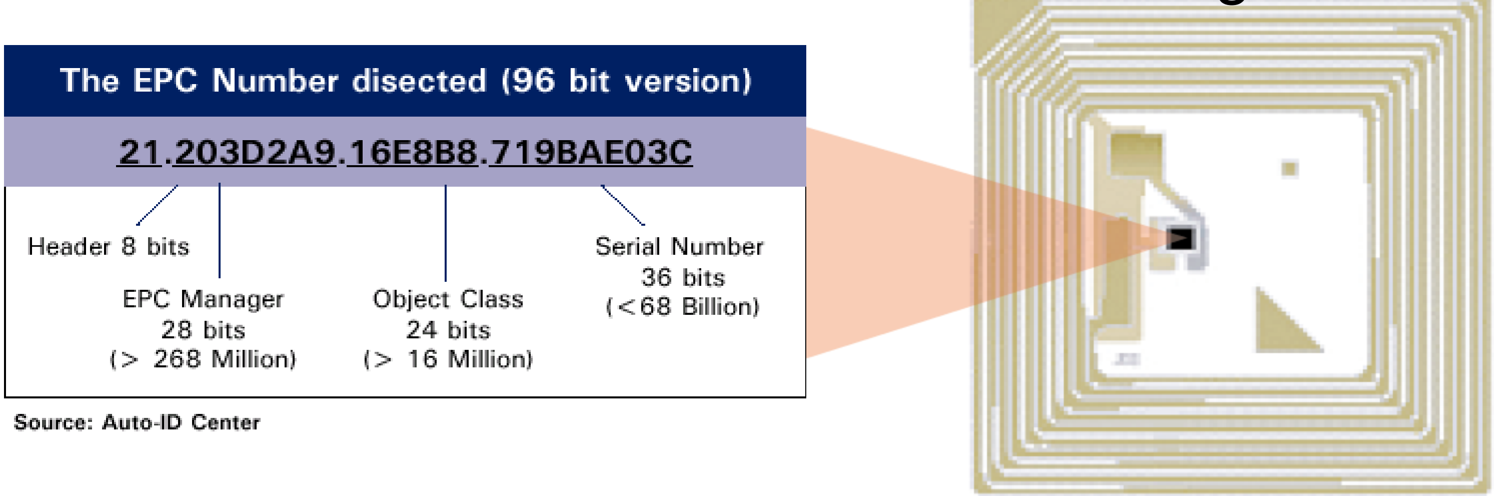
2) **Antenna:** transmits information to a reader (e.g., handheld, warehouse portal, store shelf) using radio waves

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

Use Case II - The RFID Numbers

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

A number .... in a radio tag …



The EPC Number disected (96 bit version)

21.203D2A9.16E8B8.719BAE03C

Header 8 bits

EPC Manager
28 bits
(> 268 Million)

Object Class
24 bits
(> 16 Million)

Serial Number
36 bits
(<68 Billion)

Source: Auto-ID Center

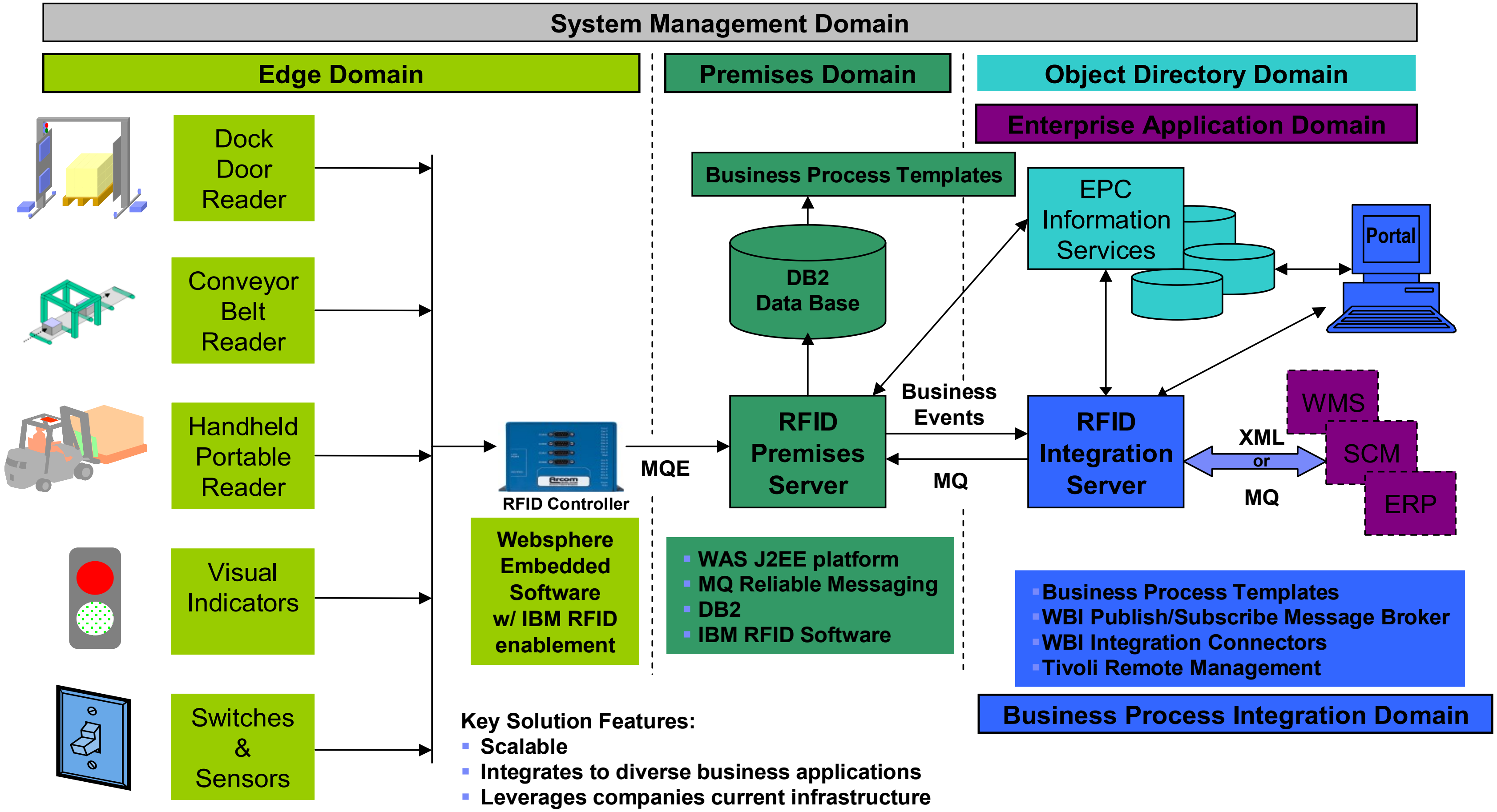
…which together, uniquely identifies an object

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

Use Case II – The RFID Infrastructure

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

Use Case II – RFID Solution with DWH



System Management Domain

Edge Domain

Premises Domain

Object Directory Domain

Enterprise Application Domain

Dock Door Reader

Conveyor Belt Reader

Handheld Portable Reader

Visual Indicators

Switches & Sensors

RFID Controller

Websphere Embedded Software w/ IBM RFID enablement

MQE

Business Process Templates

DB2 Data Base

RFID Premises Server

Business Events

MQ

EPC Information Services

Portal

RFID Integration Server

XML or MQ

WMS

SCM

ERP

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

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

Business Process Integration Domain

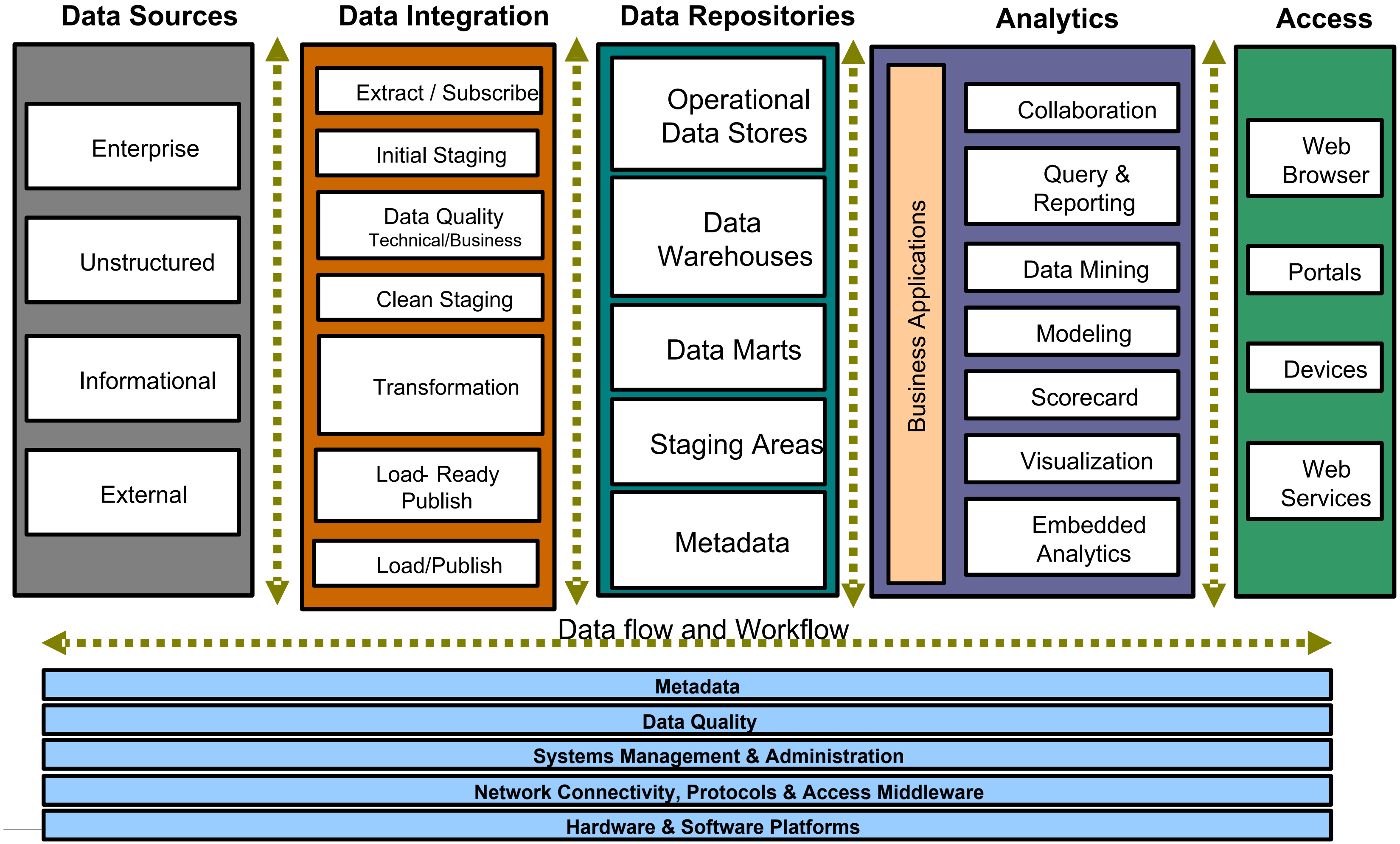
Key Solution Features:

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

What can go Wrong?

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

IBM DWH Reference Architecture
(outcome of IBM Unified Method Framework)



Data Sources
- Enterprise
- Unstructured
- Informational
- External

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

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

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

Access
- Web Browser
- Portals
- Devices
- Web Services

Data flow and Workflow

Metadata

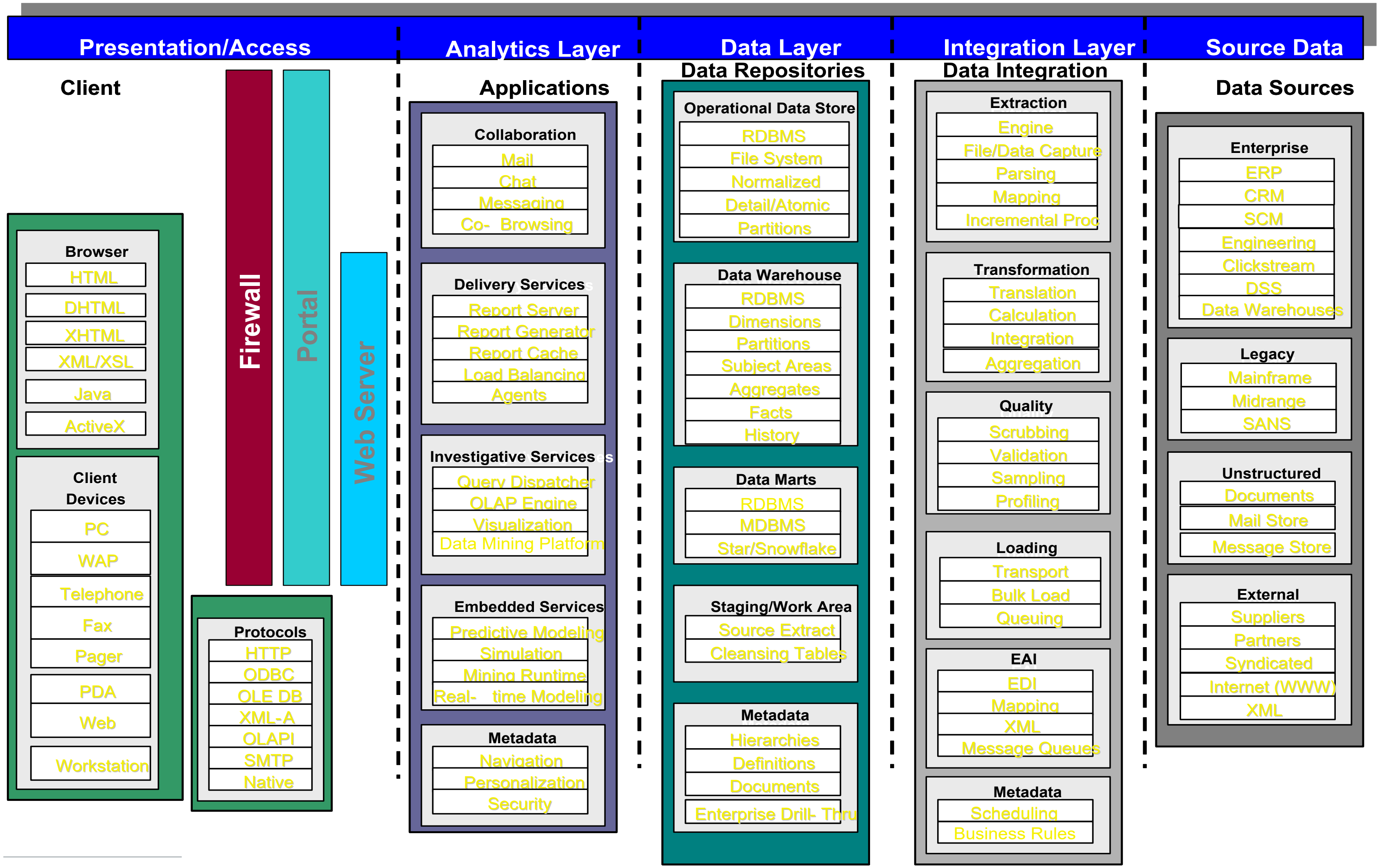
Data Quality

Systems Management & Administration

Network Connectivity, Protocols & Access Middleware

Hardware & Software Platforms

IBM DWH Reference Architecture – Details



Presentation/Access

Client

Browser: HTML, DHTML, XHTML, XML/XSL, Java, ActiveX

Client Devices: PC, WAP, Telephone, Fax, Pager, PDA, Web, Workstation

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

Firewall

Portal

Web Server

Analytics Layer

Applications

Collaboration: Mail, Chat, Messaging, Co- Browsing

Delivery Services: Report Server, Report Generator, Report Cache, Load Balancing, Agents

Investigative Services: Query Dispatcher, OLAP Engine, Visualization, Data Mining Platform

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

Metadata: Navigation, Personalization, Security

Data Layer

Data Repositories

Operational Data Store: RDBMS, File System, Normalized, Detail/Atomic, Partitions

Data Warehouse: RDBMS, Dimensions, Partitions, Subject Areas, Aggregates, Facts, History

Data Marts: RDBMS, MDBMS, Star/Snowflake

Staging/Work Area: Source Extract, Cleansing Tables

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

Integration Layer

Data Integration

Extraction: Engine, File/Data Capture, Parsing, Mapping, Incremental Proc

Transformation: Translation, Calculation, Integration, Aggregation

Quality: Scrubbing, Validation, Sampling, Profiling

Loading: Transport, Bulk Load, Queuing

EAI: EDI, Mapping, XML, Message Queues

Metadata: Scheduling, Business Rules

Source Data

Data Sources

Enterprise: ERP, CRM, SCM, Engineering, Clickstream, DSS, Data Warehouses

Legacy: Mainframe, Midrange, SANS

Unstructured: Documents, Mail Store, Message Store

External: Suppliers, Partners, Syndicated, Internet (WWW), XML

Exercise 1 to Lesson 2

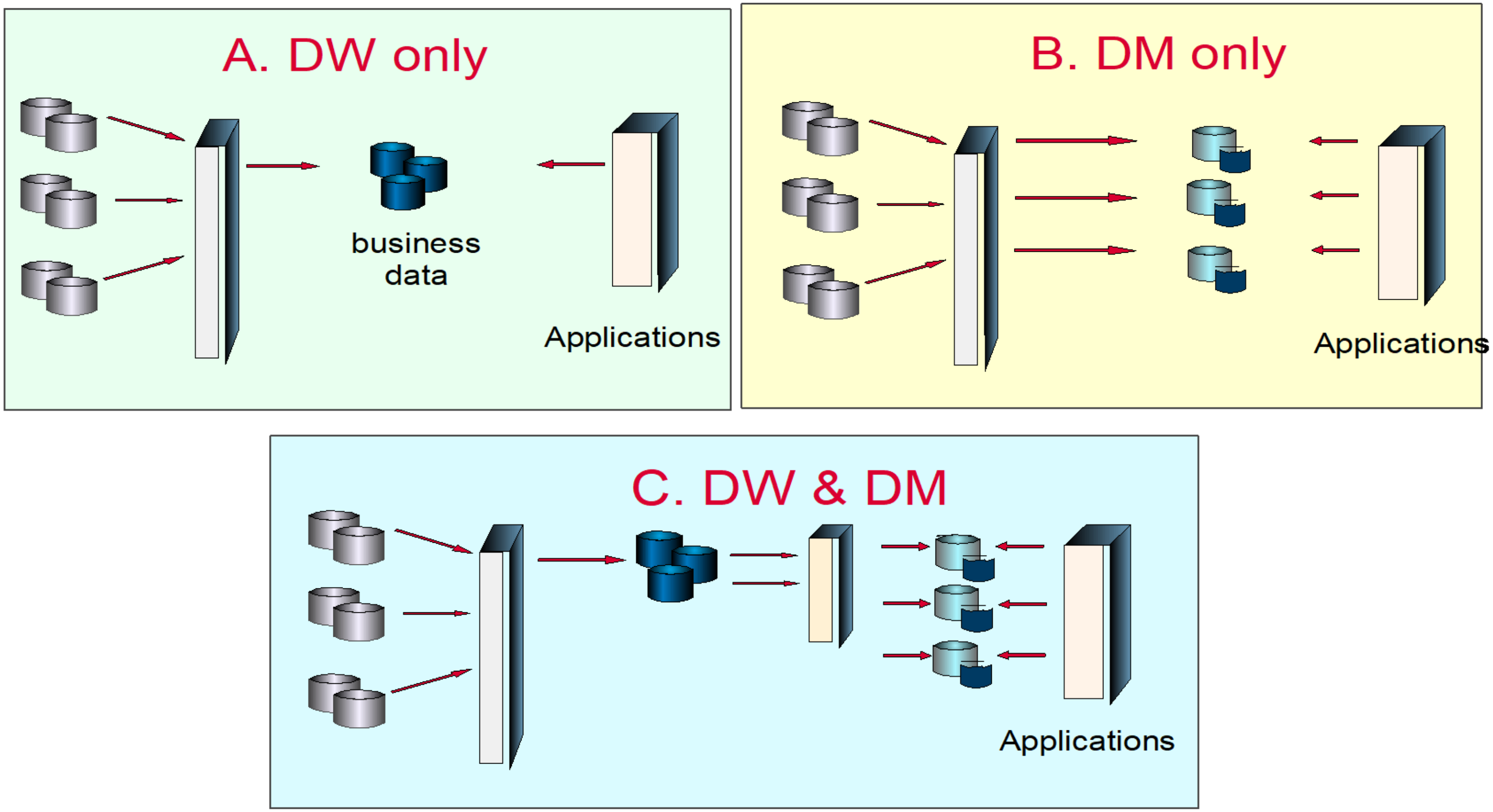
Exercise E2.1 (SW*): *Compare the three DWH architectures (DW only, DM only and DW & DM) in the next slide. List the advantages and disadvantages and give a detailed explanation for it. Find also a fourth possible architecture (hint: 'virtual' DWH)*

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

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

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

Exercise 1 to Lesson 2 (cont.)



Exercise 2 to Lesson 2: Basel II & RFID

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

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

1. Basel II / Basel III
2. RFID

Look also for examples of current projects in Germany

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

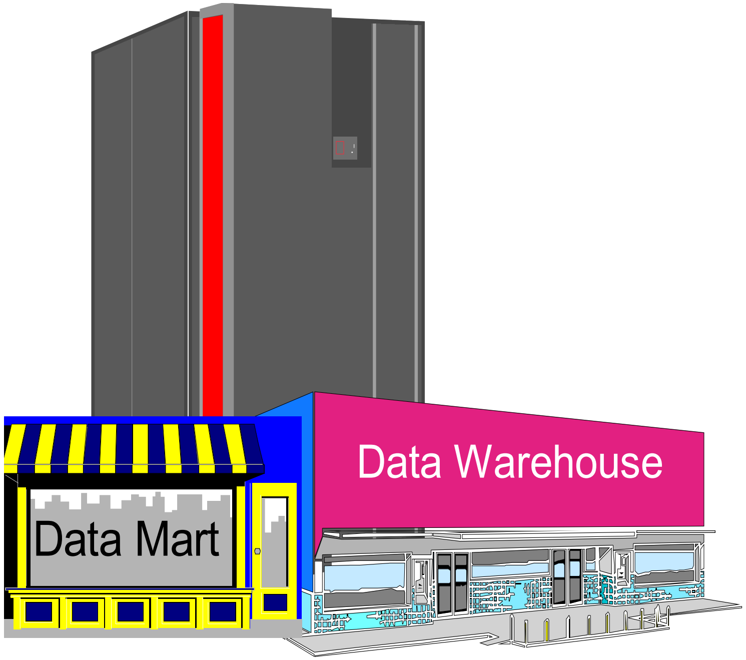
Exercise 3 to Lesson 2: Modern Data Arch.

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

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

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

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



The four Goals of a DBMS

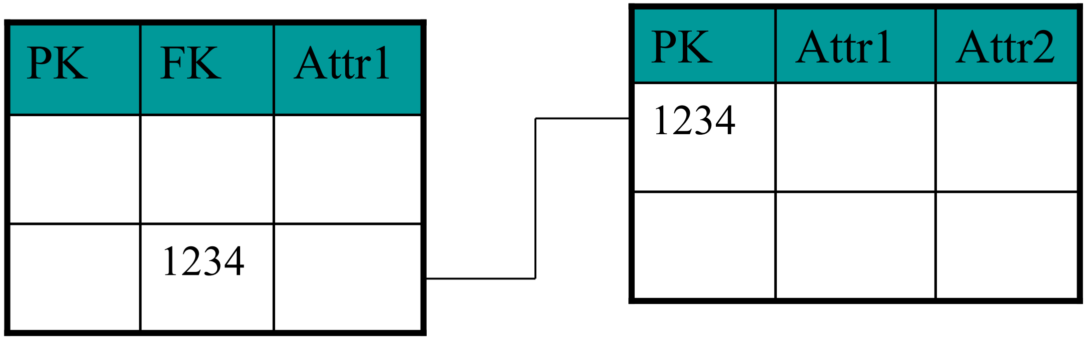
DBMS (Database Management Systems) are designed to achieve the following four main goals:

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

Three traditional Database Structures

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

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



PK	FK	Attr1
	1234	

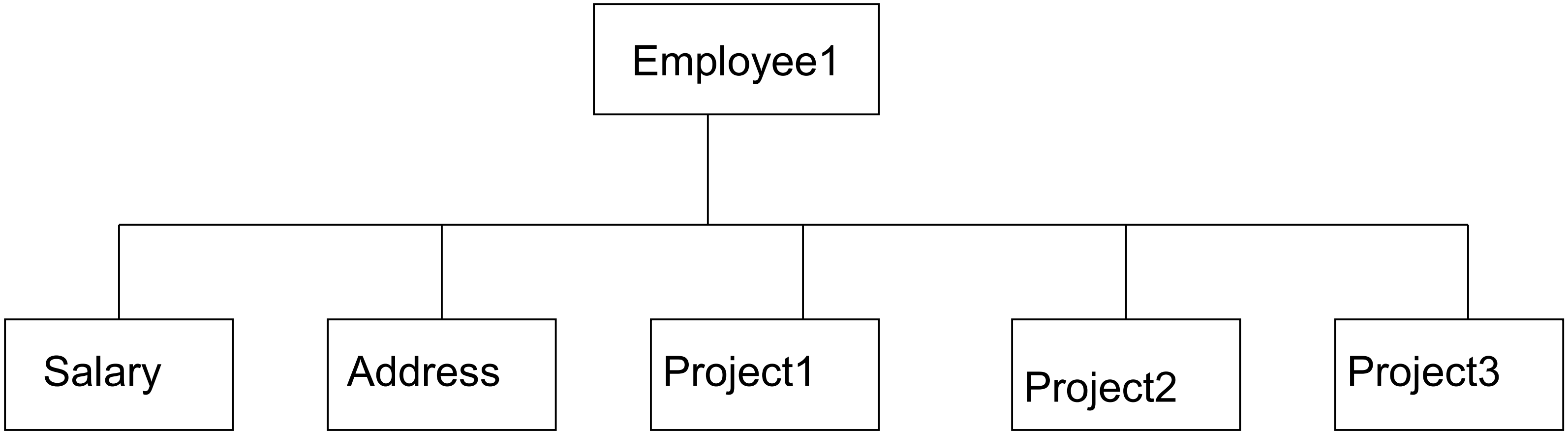
PK	Attr1	Attr2
1234		

Hierarchical Database Structures

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

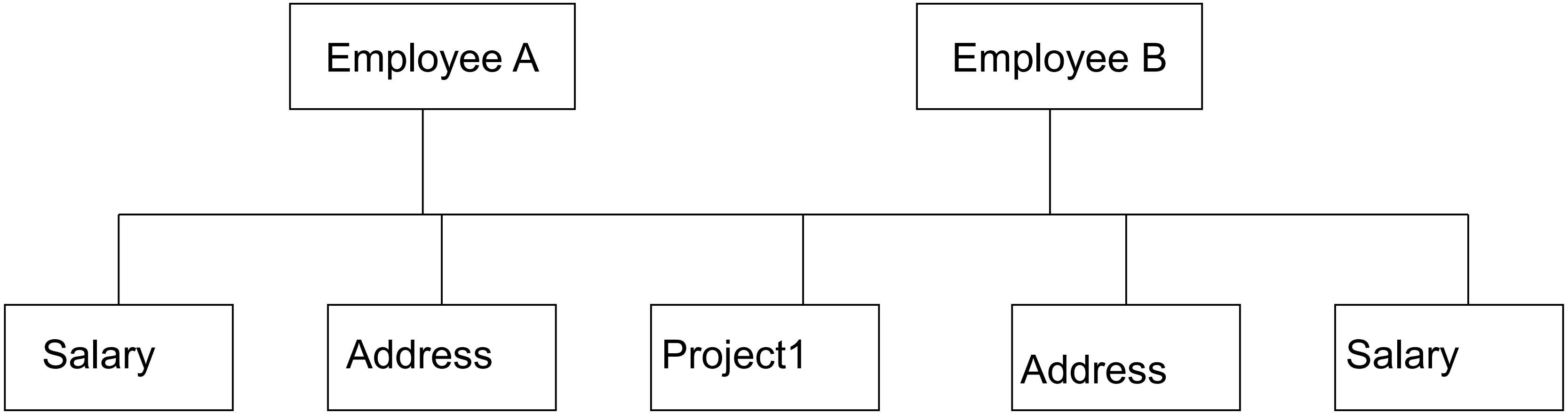
Each record may contain several information parts, for example:

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



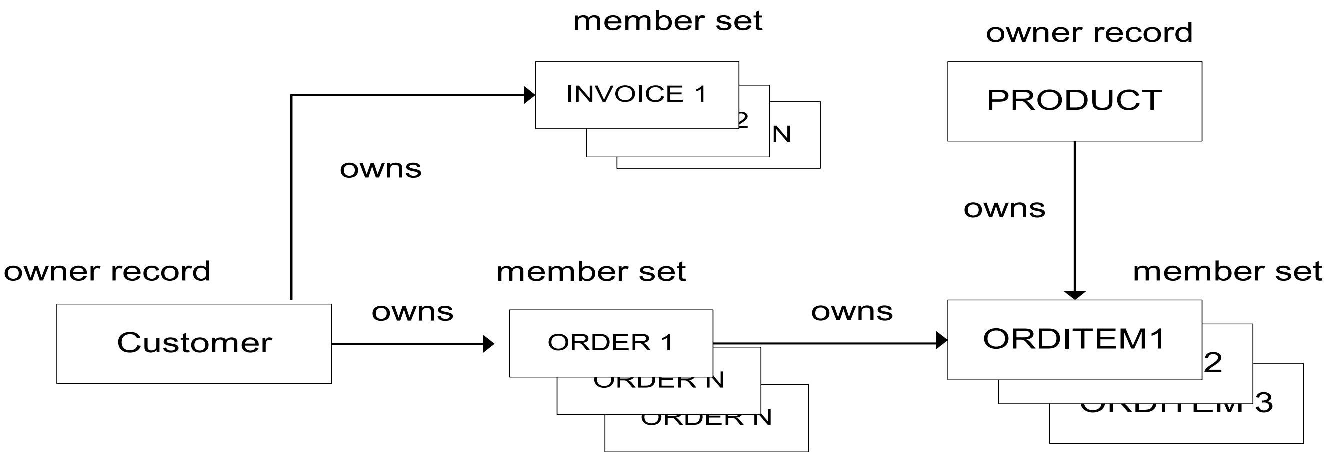
Network Database Structures

- More flexible
- Reduce Redundancy



Example – IDMS Datenbase

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



Migration from IDMS to DB2

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

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

What is a Relational Database

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

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

Relational Database Structures

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

Employee Table

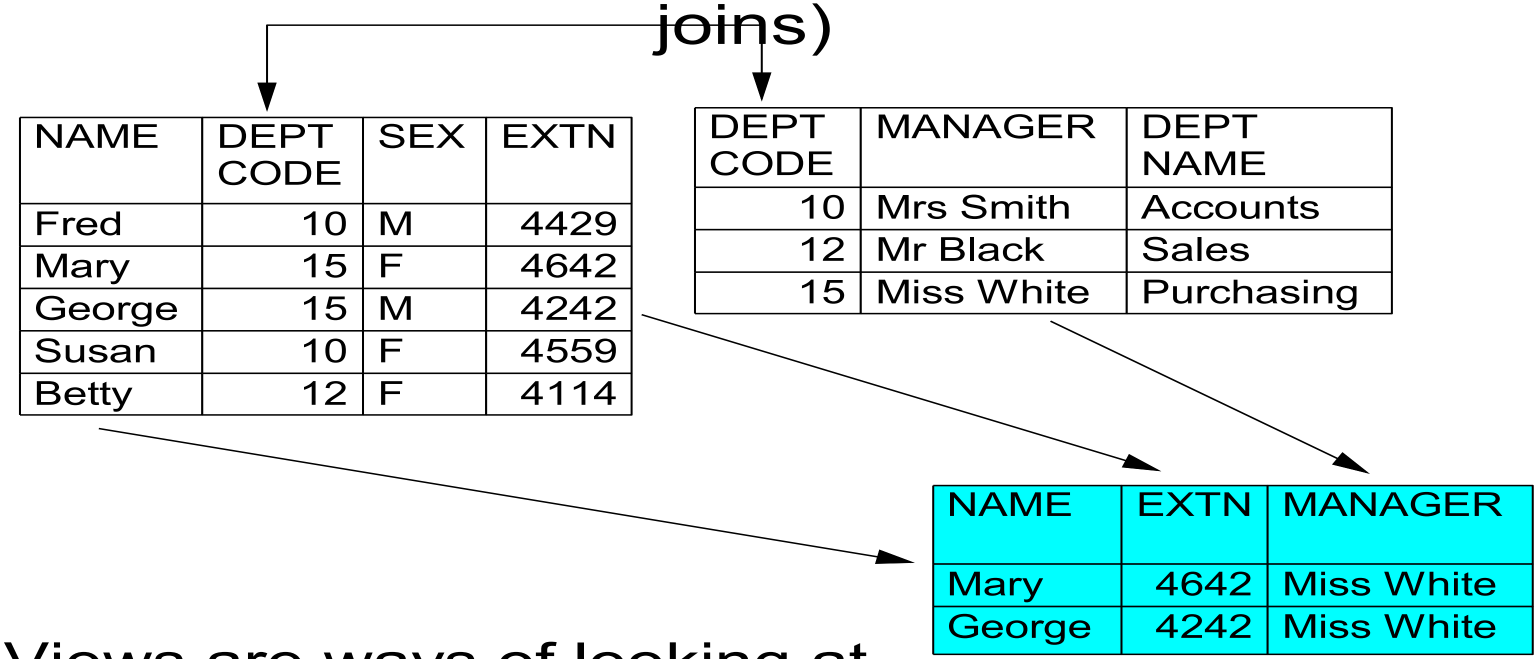
EmpNo	Workdep	Empname	Position
321-412	100	Jones	Programmer
456-673	100	Simpson	Analyst

Project Table

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

Views and Joins

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



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

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

NAME	EXTN	MANAGER
Mary	4642	Miss White
George	4242	Miss White

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

The Database Join Strategies

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

Summary: Relational Database - Features

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

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

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

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

Motivation & Introduction to Normal Forms

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

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

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

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

The First Normal Form (1NF)

Rule:

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

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

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

Original Rule (from Codd):

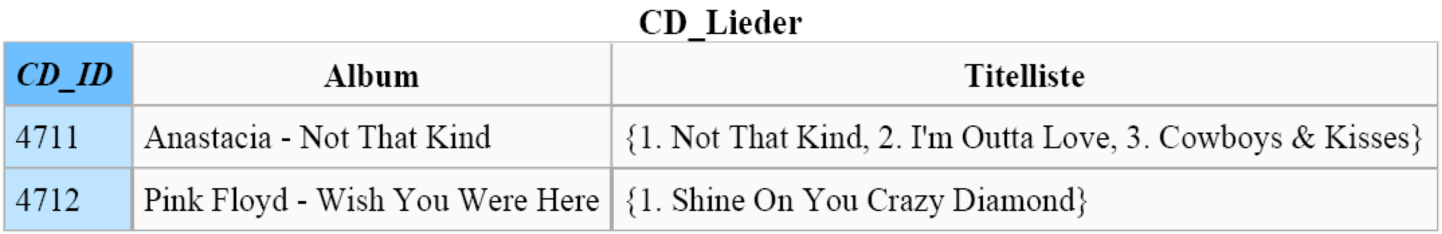
All columns in a relation are only dependent from the key.

Action:

Eliminate repeating values in one atom and repeating groups.

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

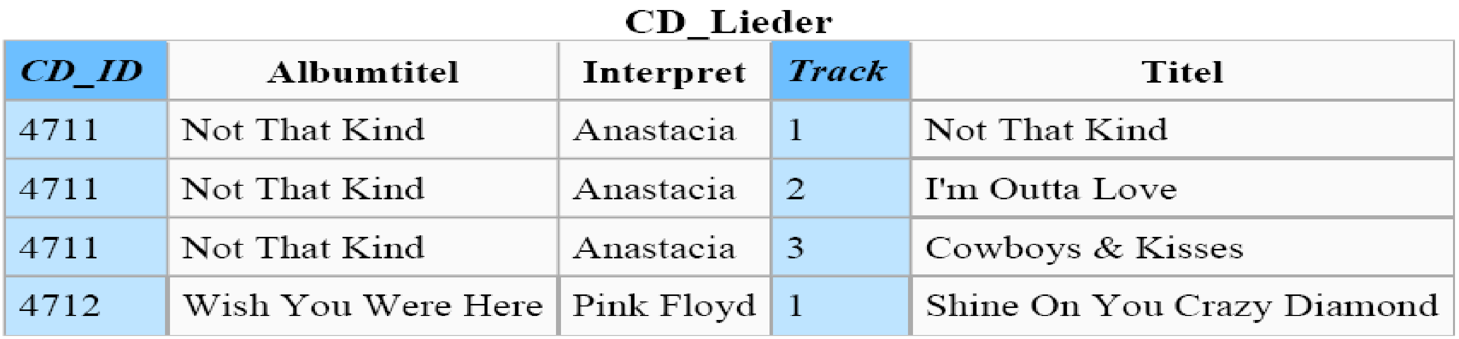
The following table is not in First Normal Form (*examples are from WIKIPEDIA).
The attribute ‘*Album*‘ has information about *Interpret* and *CD Title* ......



CD_Lieder

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

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



CD_Lieder

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

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

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

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

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

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

Example - First Normal Form *(‘Anomalies’)*

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

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

INPUT ‘Anomalien‘

How to insert a student , who never have done an examination?

DELETE ‘Anomalien‘

When you delete the student Pitt, you loose the information about ‘Dekan BWL‘

CHANGE ‘Anomalien‘

When a student changes his address, you have to change the street in several places.

Remark: There is another hidden problem in the data of this table? Any idea?

Second Normal Form (2NF)

Rule:

The table must be in 1NF.

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

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

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

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

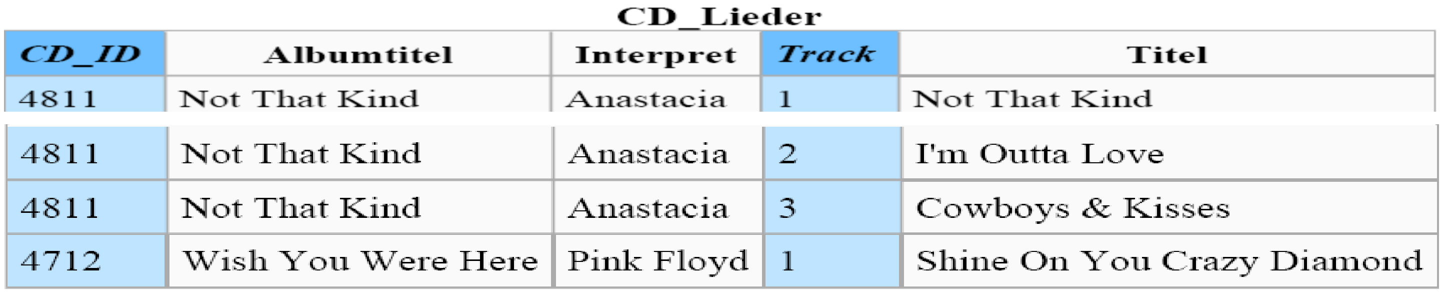
Action:

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

Example for Second Normal Form

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

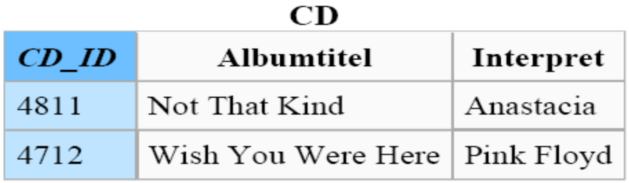
The primary key of the relation exists of the fields *CD_ID* and *Track.* The fields *Albumtitel* and *Interpret are* dependent from the field *CD_ID* but not from the field *Track*.



CD_Lieder

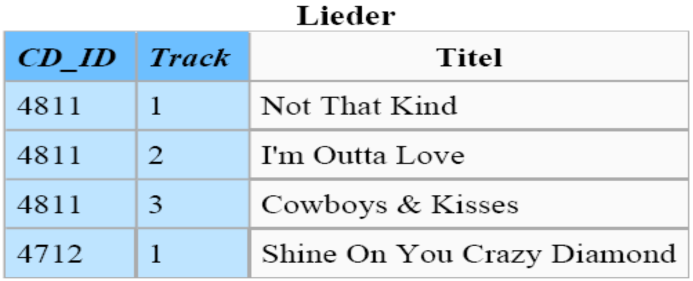
CD_ID	Albumtitel	Interpret	*Track*	Titel
4811	Not That Kind	Anastacia	1	Not That Kind
4811	Not That Kind	Anastacia	2	I'm Outta Love
4811	Not That Kind	Anastacia	3	Cowboys & Kisses
4712	Wish You Were Here	Pink Floyd	1	Shine On You Crazy Diamond

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



CD

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



Lieder

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

Third Normal Form (3NF)

Rule:

The table must be in 2NF.

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

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

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

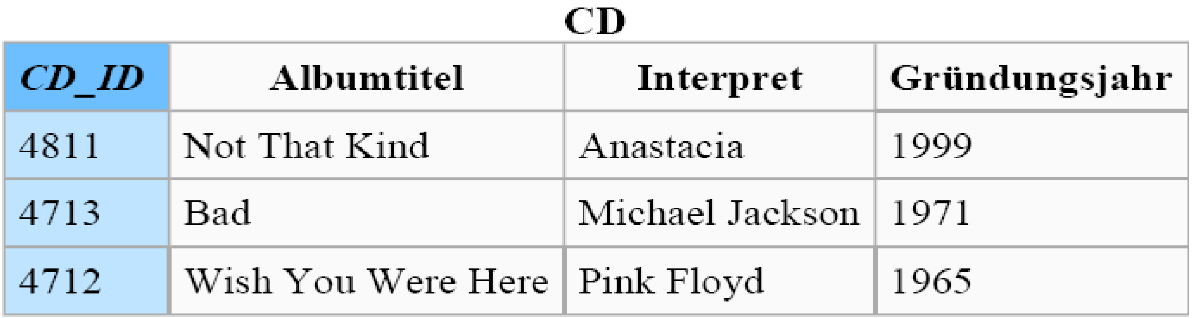
Action:

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

Example for Third Normal Form

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

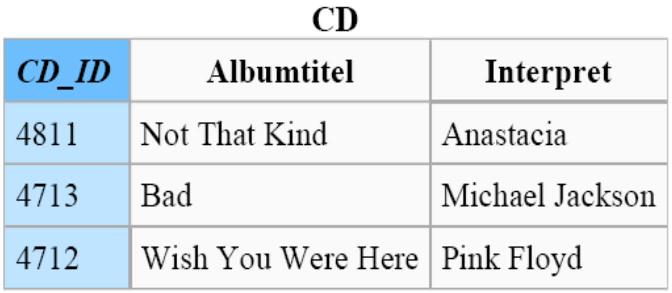
The field *Interpret* of the table CD is dependant from *CD_ID,* but *Gründungsjahr* is also dependant from *Interpret* and therefore transitive dependant from *CD_ID* .



CD

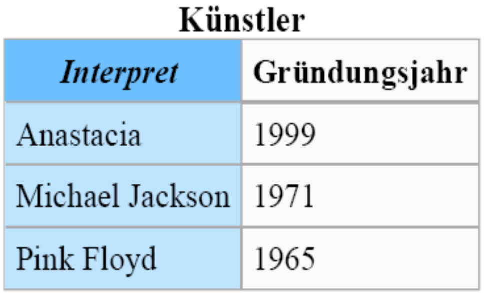
CD_ID	**Albumtitel**	**Interpret**	**Gründungsjahr**
4811	Not That Kind	Anastacia	1999
4713	Bad	Michael Jackson	1971
4712	Wish You Were Here	Pink Floyd	1965

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



CD

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



Künstler

Interpret	**Gründungsjahr**
Anastacia	1999
Michael Jackson	1971
Pink Floyd	1965

Summary – Normal Forms 1NF-3NF

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

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

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

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

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

Normalization Benefits

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

Check your Knowledge about DBMS

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

Exercice / Repetition 1 to Lesson 3 (Optional)

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

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

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

Exercise 2 to Lesson 3

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

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

SAMP_PROJECT

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

SAMP_STAFF

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

Exercise 3 to Lesson 3

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

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

Exercise 4 to Lesson 3

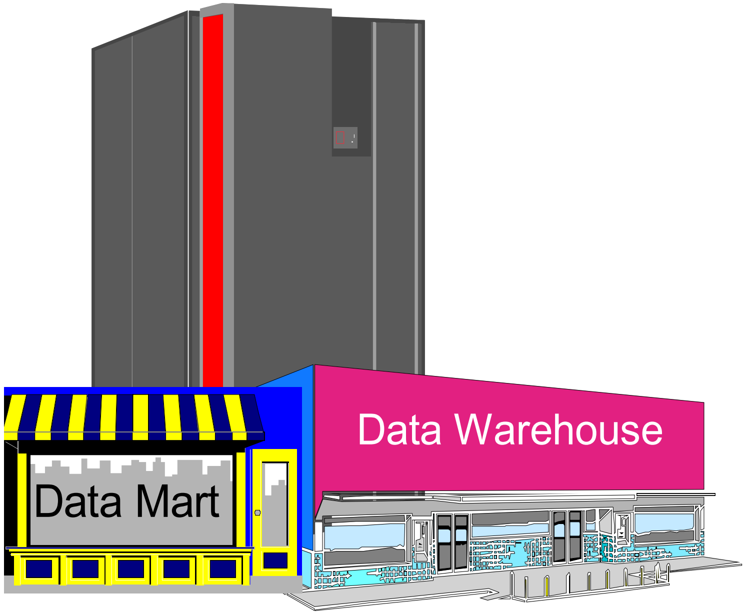
Exercise E3.4: Do the normalization steps 1NF, 2NF and 3NF to the following un-normalized table (show also the immediate results):

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

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

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

DW04 - Introduction to Basics of SQL



Introduction to SQL

SQL is divided into three major categories:

1. **DDL** – Data Definition Language
   - Used to **create**, **modify** or **drop** database objects
2. **DML** – Data Manipulation Language
   - Used to **select**, **insert**, **update** or **delete** database data (records)
3. **DCL** – Data Control Language
   - Used to provide data object access control

Examples of DDL commands

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

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

Examples of DML commands

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

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

Examples of DML commands (Part 2)

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

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

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

- **create view**...

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

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

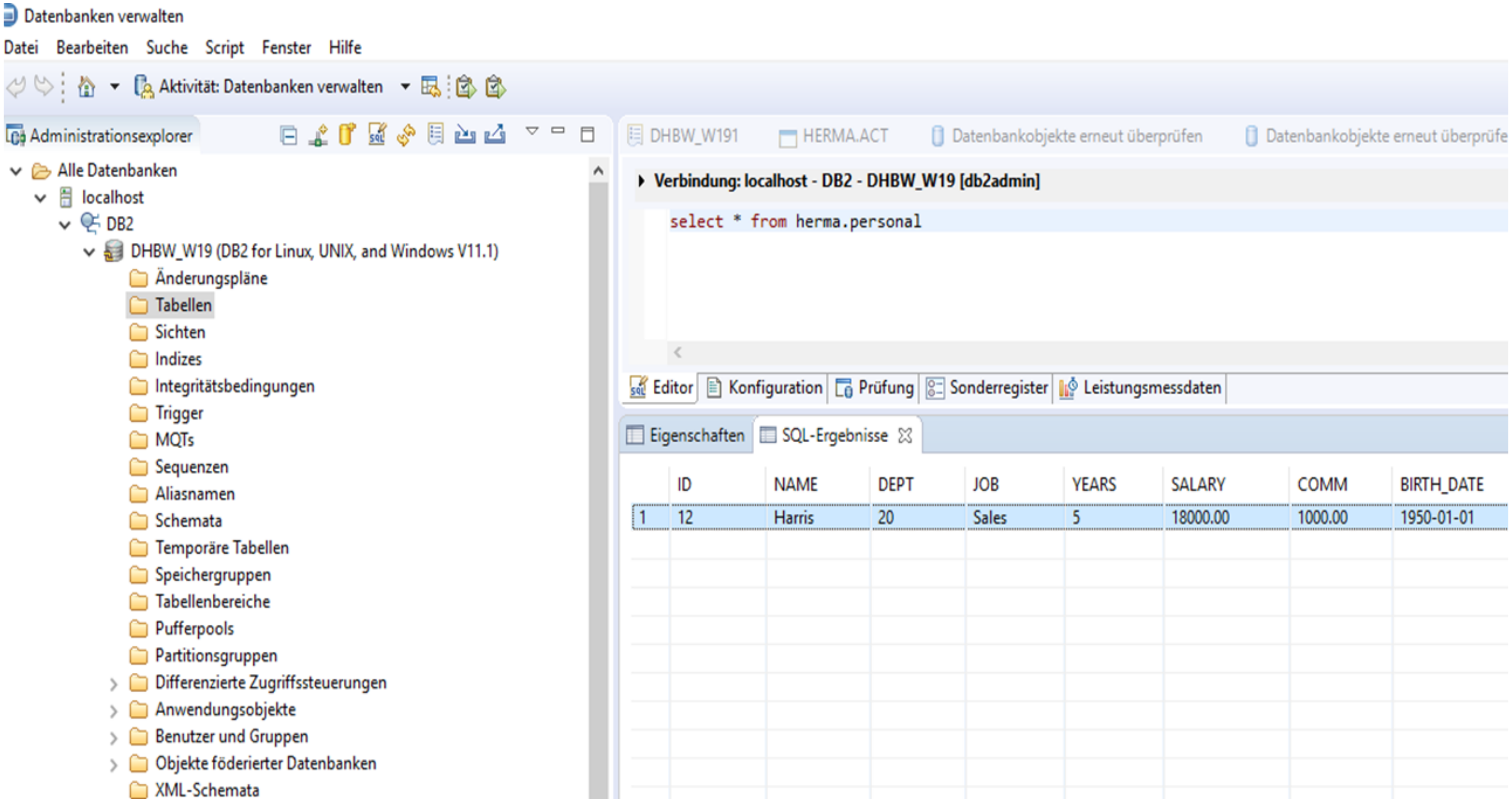
Examples for DCL commands

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

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

Demo with IBM Data Studio

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



Verbindung: localhost - DB2 - DHBW_W19 [db2admin]

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

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

Exercise 1 to Lesson 4

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

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

Optional (difficult)

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

Airport:

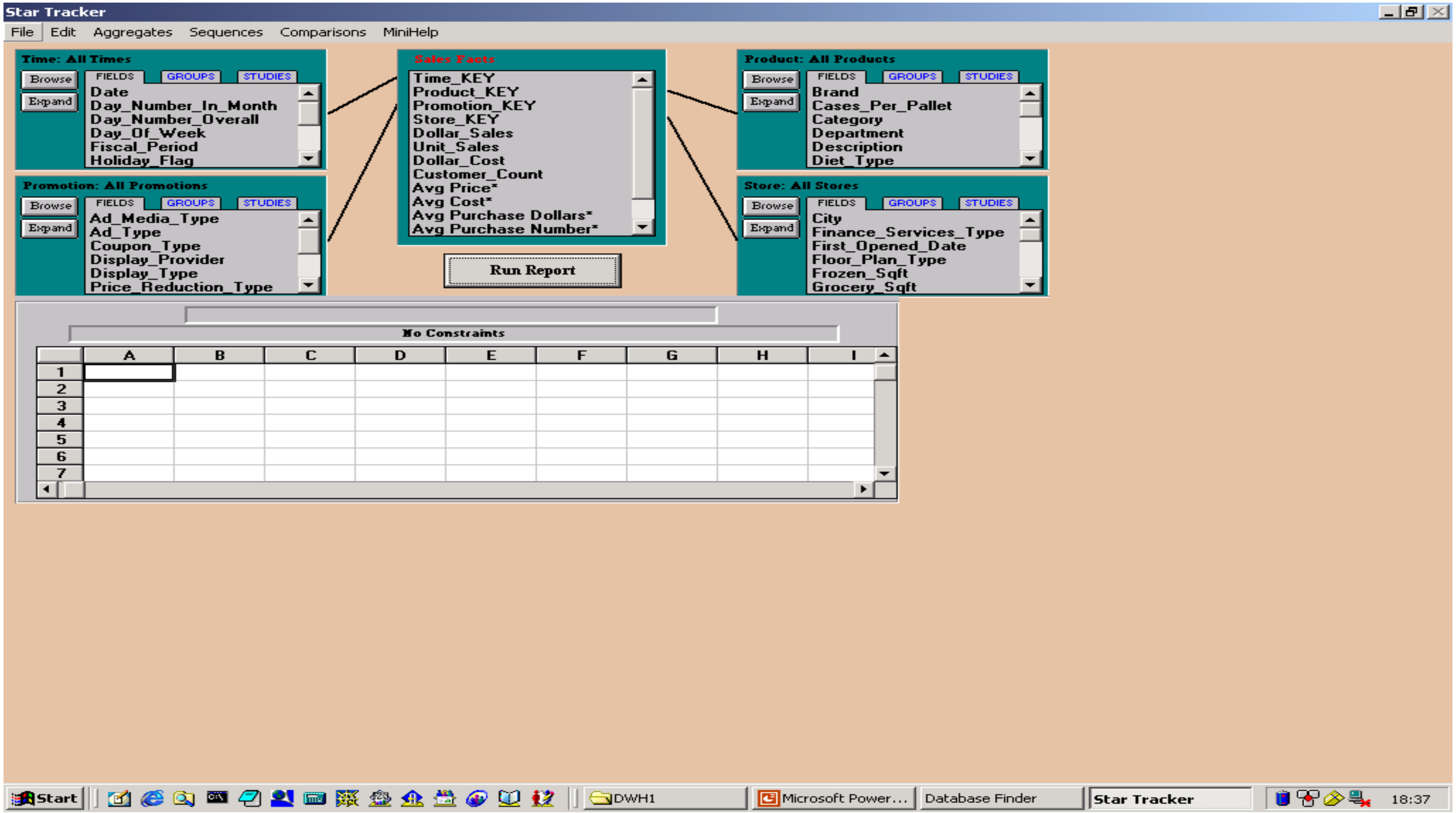
FID	*Name*
MUC	Muenchen
FRA	Frankfurt
HAN	Hannover
STU	Stuttgart
MAN	Mannheim
BER	Berlin

Flight:

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

Exercise 2 to Lesson 4 (First part)

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



Time: All Times
- Date
- Day_Number_In_Month
- Day_Number_Overall
- Day_Of_Week
- Fiscal_Period
- Holiday_Flag

Promotion: All Promotions
- Ad_Media_Type
- Ad_Type
- Coupon_Type
- Display_Provider
- Display_Type
- Price_Reduction_Type

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

Product: All Products
- Brand
- Cases_Per_Pallet
- Category
- Department
- Description
- Diet_Type

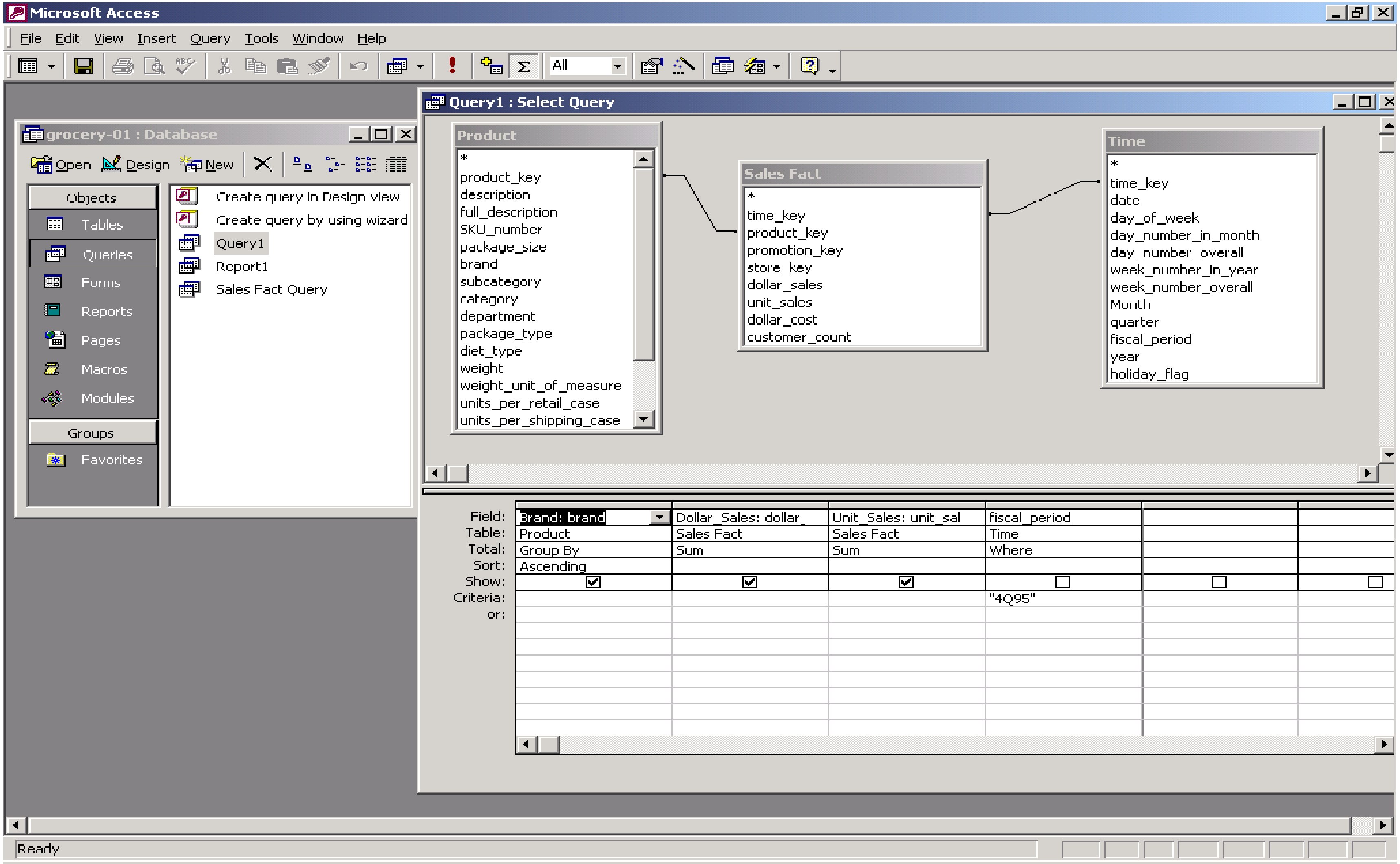
Store: All Stores
- City
- Finance_Services_Type
- First_Opened_Date
- Floor_Plan_Type
- Frozen_Sqft
- Grocery_Sqft

Exercise 2 to Lesson 4 (Part 2)

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

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

Solution with MS Access SQL Wizard



Product

- *
- product_key
- description
- full_description
- SKU_number
- package_size
- brand
- subcategory
- category
- department
- package_type
- diet_type
- weight
- weight_unit_of_measure
- units_per_retail_case
- units_per_shipping_case

Sales Fact

- *
- time_key
- product_key
- promotion_key
- store_key
- dollar_sales
- unit_sales
- dollar_cost
- customer_count

Time

- *
- time_key
- date
- day_of_week
- day_number_in_month
- day_number_overall
- week_number_in_year
- week_number_overall
- Month
- quarter
- fiscal_period
- year
- holiday_flag

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

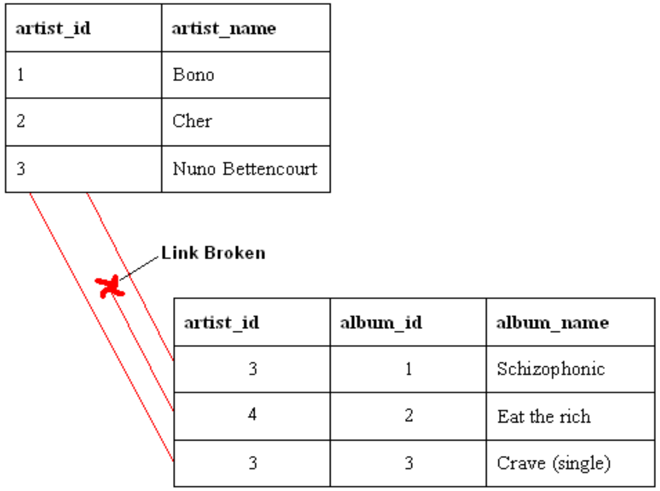
Exercise 3 to Lesson 4

Advanced Study about concepts in DWH:

<u>Exercise E4.3 (SW*):</u>

Explain what is "Referential Integrity" (RI)

in a Database?



artist_id	artist_name
1	Bono
2	Cher
3	Nuno Bettencourt

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

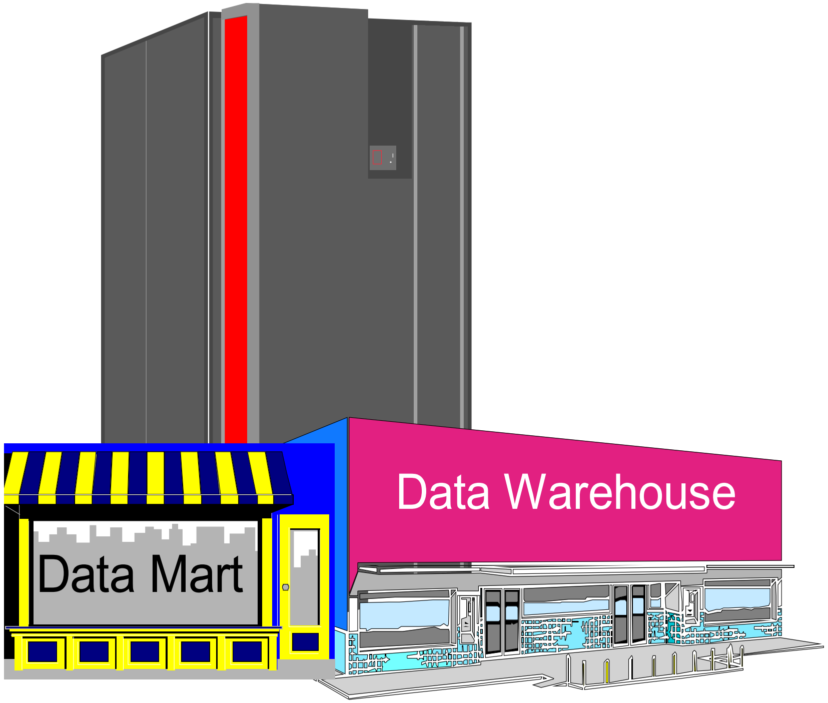
Sub-Questions:

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

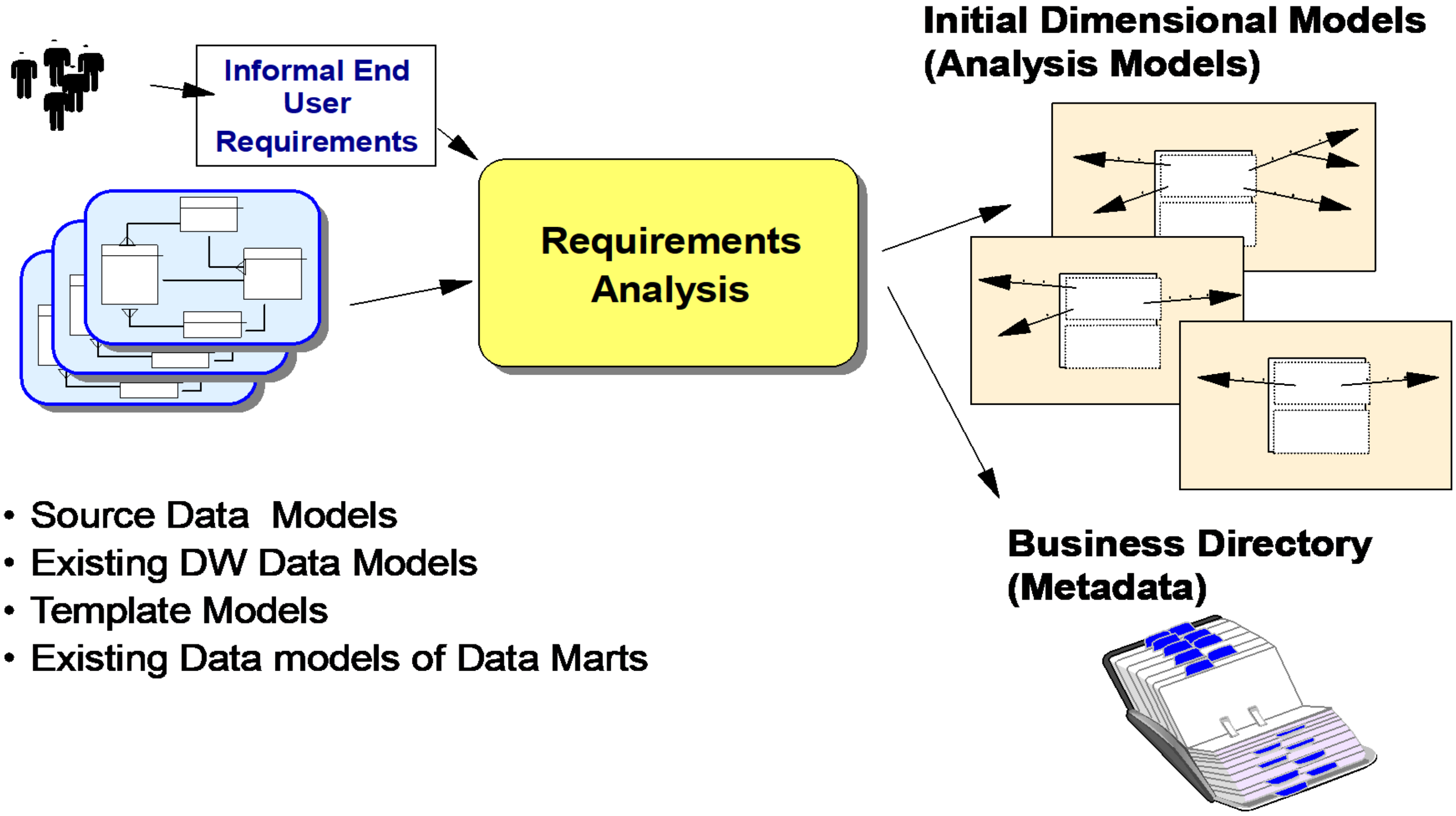
Find explanations and arguments in DWH forums or articles about this theme in the internet or in the literature.

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

DW05 – Multi-Dimensional Data Modeling



Requirements Analysis- Context



Informal End User Requirements

Requirements Analysis

Initial Dimensional Models (Analysis Models)

Business Directory (Metadata)

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

Requirements Analysis - Activities

Informal End-User Requirements

Process-Oriented Requirements

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

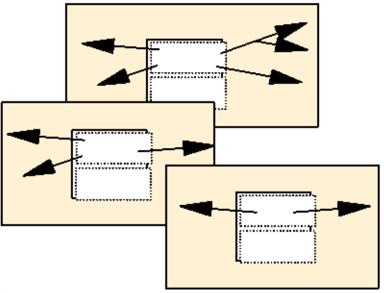
Information Oriented Requirements

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

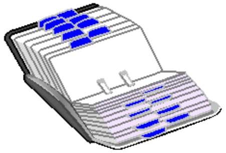
Requirements Analysis

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

Initial Dimensional Models (Analysis Models)



Business Directory (Metadata)

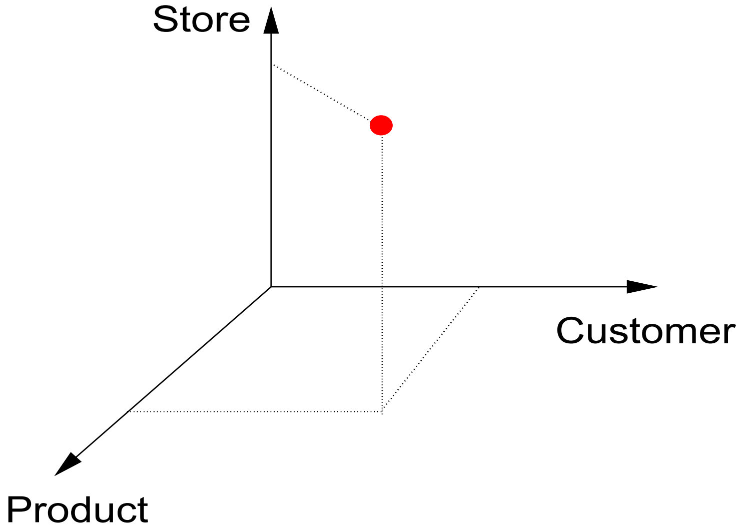


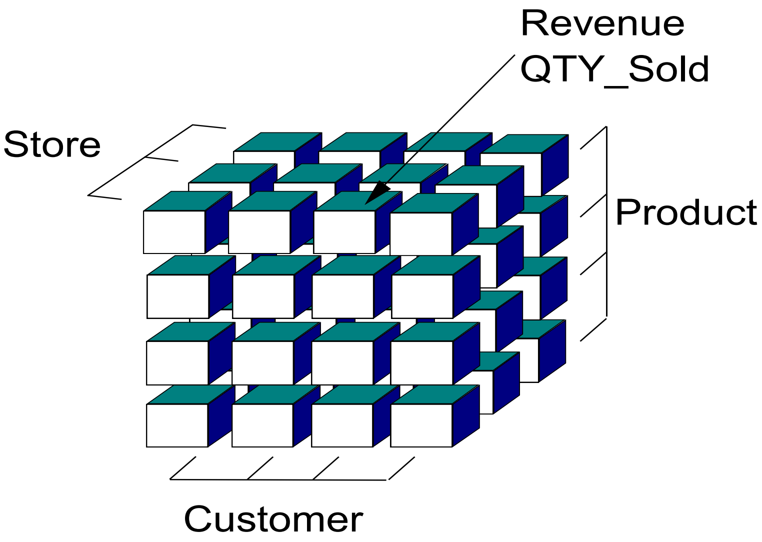
Sample Query

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

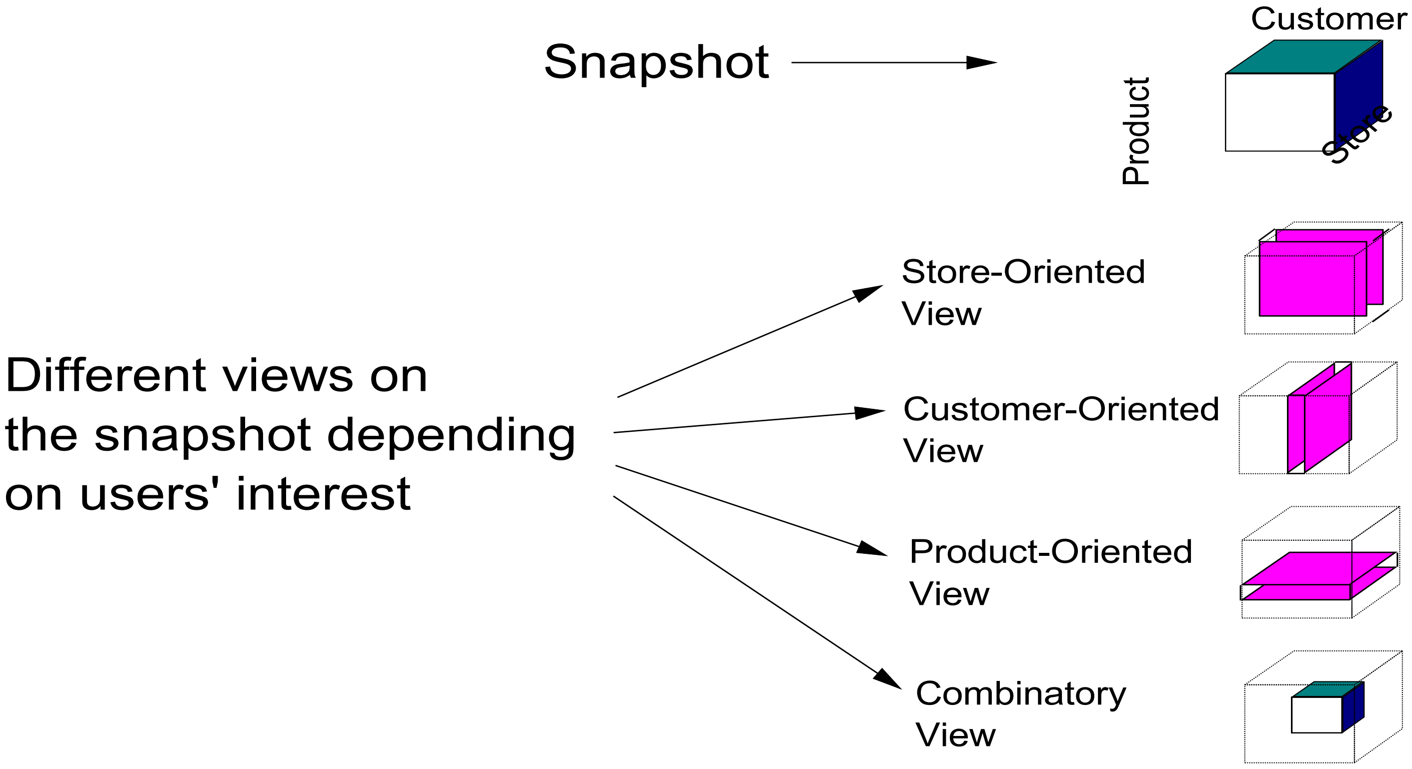
Representation of the Query as a Cube

(3 dimensions)





Presentation of the Query as a Cube : Usage



Snapshot

Different views on the snapshot depending on users' interest

Store-Oriented View

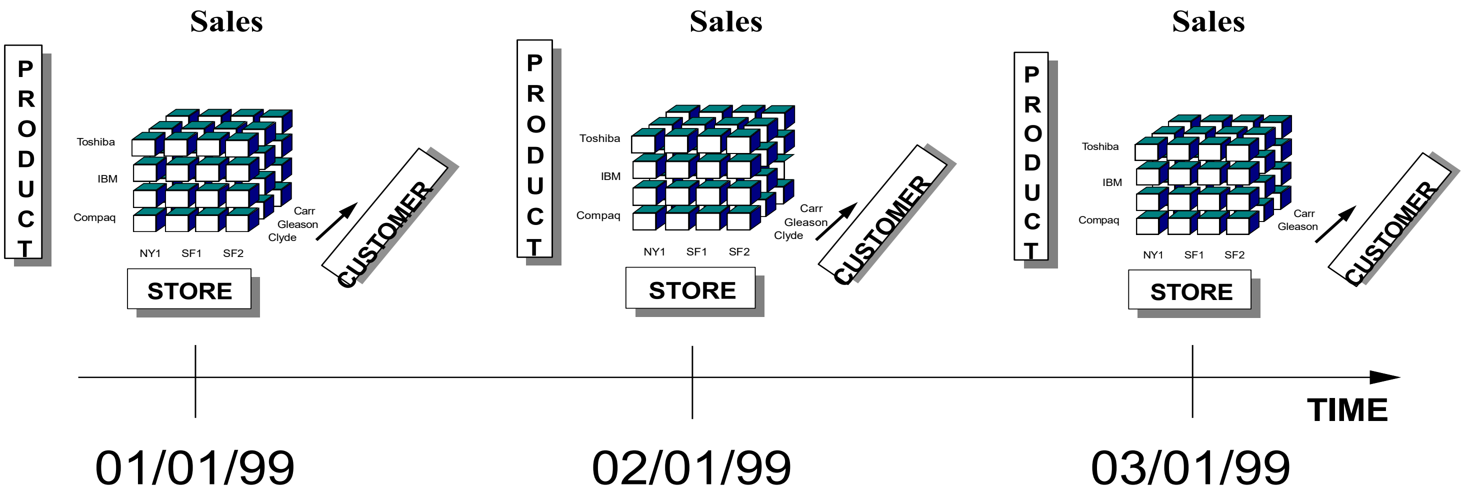
Customer-Oriented View

Product-Oriented View

Combinatory View

Hypercube Representation

(4th dimension)

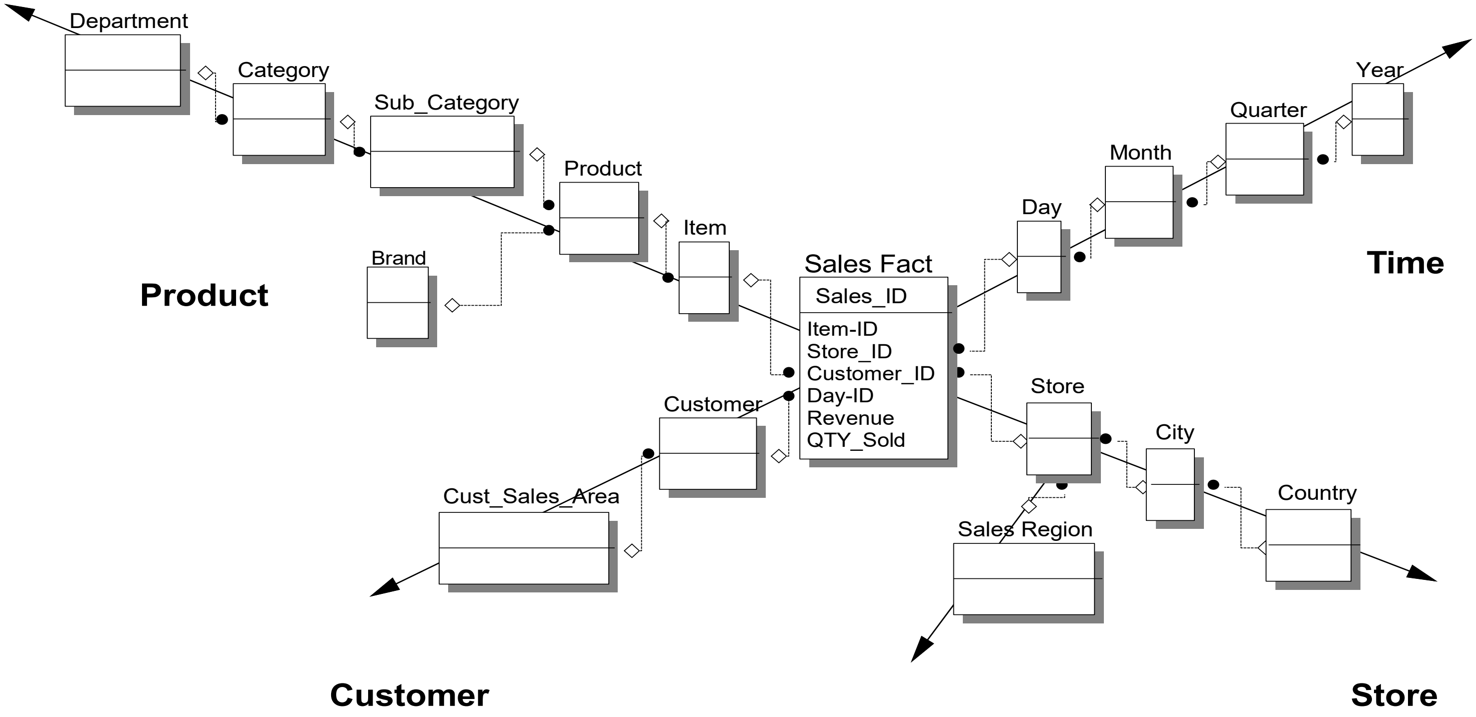


Hypercube:

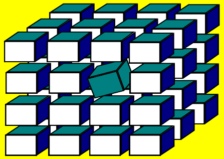
Good visual representation for three dimensions

<u>Difficult to use, when more than four dimensions</u>

Sample Multidimensional Representation Usable for Any Number of Dimensions



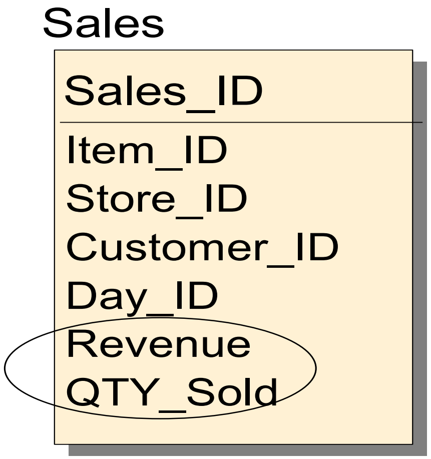
The Six Base Concepts of MDDM



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

Multidimensional Modeling - Base Concepts (1 of 6)

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



Sales

Sales_ID

Item_ID
Store_ID
Customer_ID
Day_ID
Revenue
QTY_Sold

Measures

Identify Candidate Measures

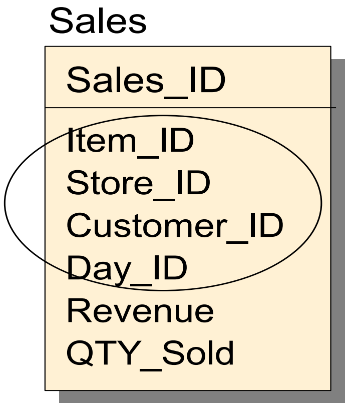
- Query-Oriented Approach
  - Perform a smart, not a mechanical analysis of the available queries
- Candidate Measures are
  - **Numeric, "Continuously" Valued**
    - But not every numeric attribute is a candidate measure
    - Distinguish measures from discrete valued numeric attributes which are part of dimensions
  - **Involved in Aggregation Calculations**
- Examples
  - Revenue (sales query)
  - Quantity sold (sales query)

Measures

Multidimensional
Modeling - Base Concepts (2 of 6)

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

Dimensions



Sales

Sales_ID

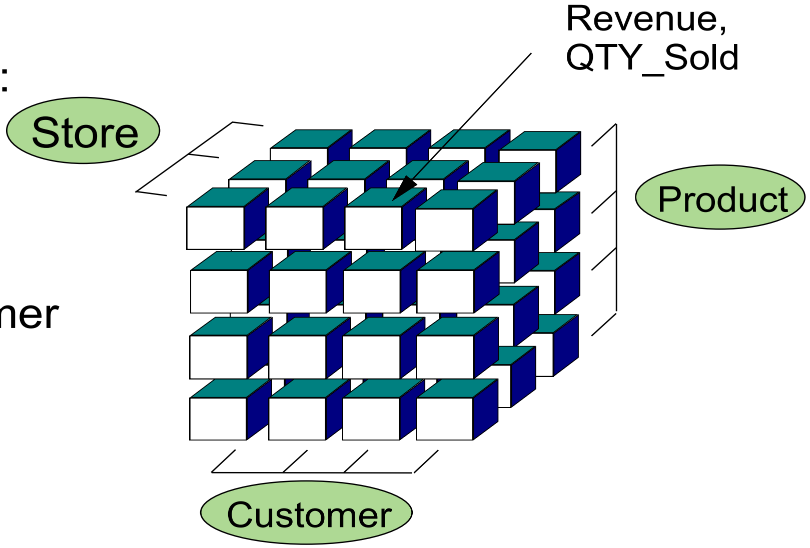
Item_ID
Store_ID
Customer_ID
Day_ID
Revenue
QTY_Sold

Identify Candidate Dimensions

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

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

Dimensions



Modeling - Base Concepts (3 of 6)

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

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

Granularity

About Granularity - Example

Low Granularity Hides Information

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

Sales Region1

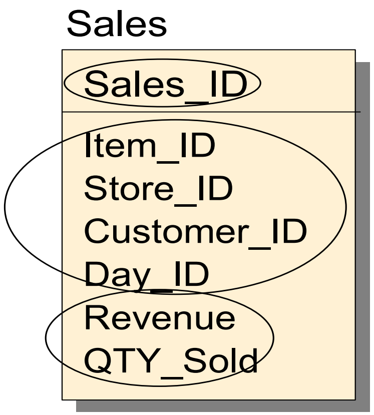
Revenue	1/1	2/1	3/1	4/1
Store1	20	15	35	35
Store2	18	13	5	5
Store3	12	17	14	5
Store4	15	10	21	5

Granularity

Multidimensional Modeling - Base Concepts (4 of 6)

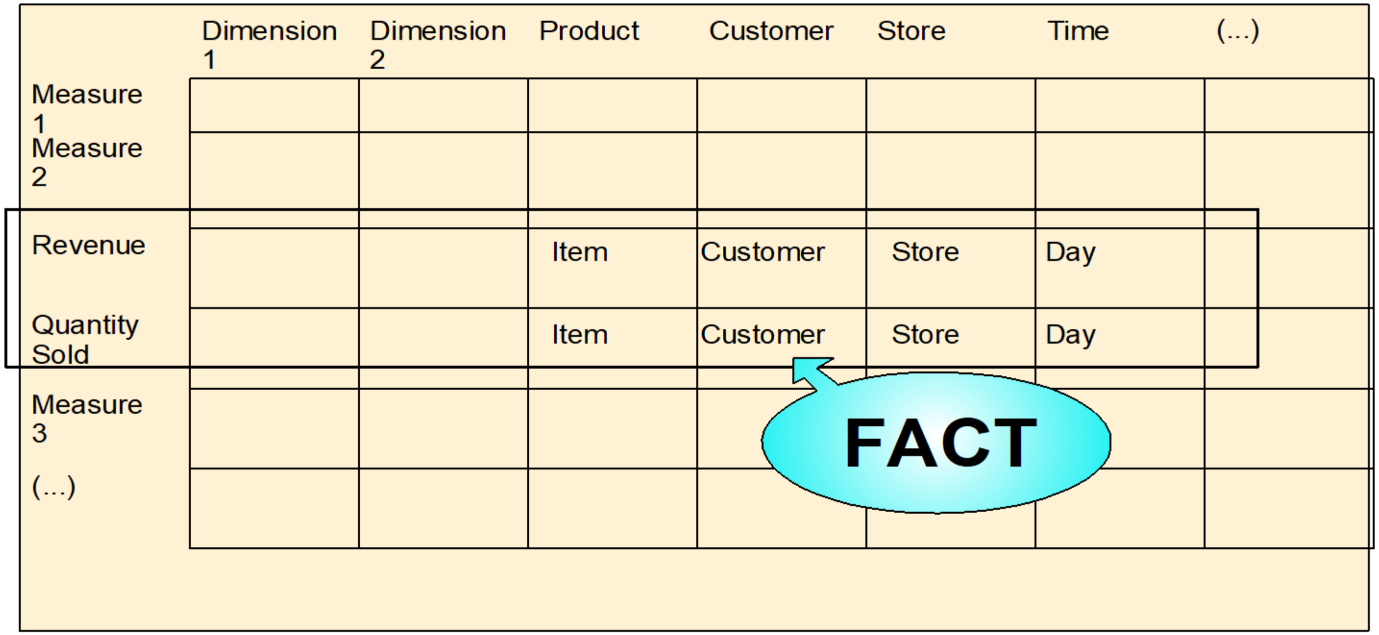
- Fact
  - A fact is a collection of related measures and their associated dimensions, represented by the dimension keys
    - Example: Sales
  - A fact can represent a business object, a business transaction or an event which is used by the information analyst
- Facts contain
  - A Fact Identifier
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

Facts



Identify Candidate Facts

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

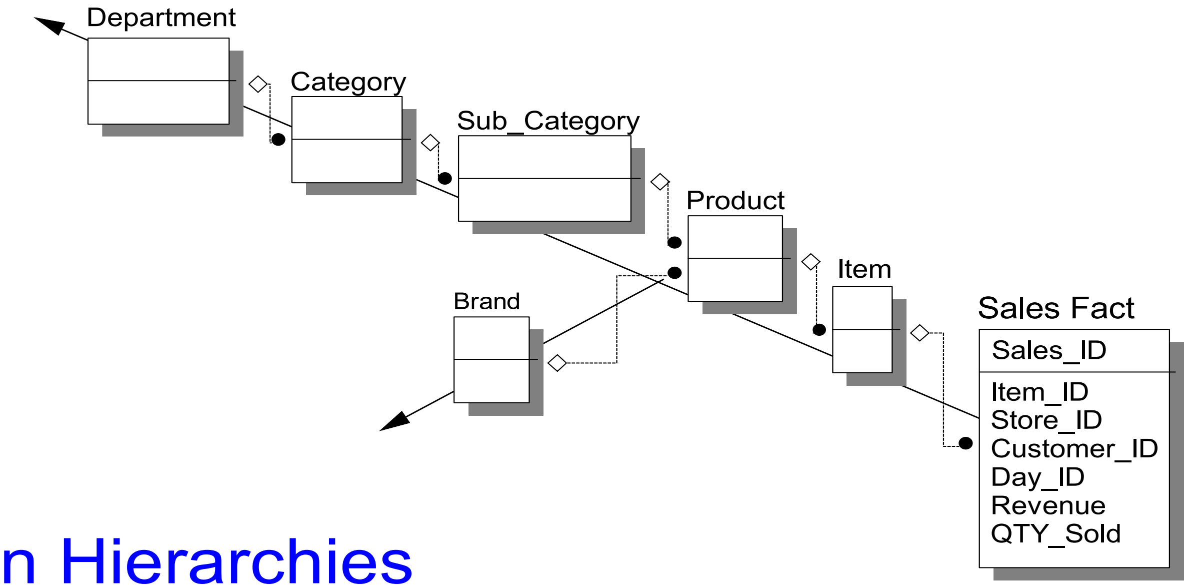


	Dimension 1	Dimension 2	Product	Customer	Store	Time	(...)
Measure 1							
Measure 2							
Revenue			Item	Customer	Store	Day	
Quantity Sold			Item	Customer	Store	Day	
Measure 3							
(...)							

Facts

Multidimensional Data Modeling - Base Concepts (5 of 6)

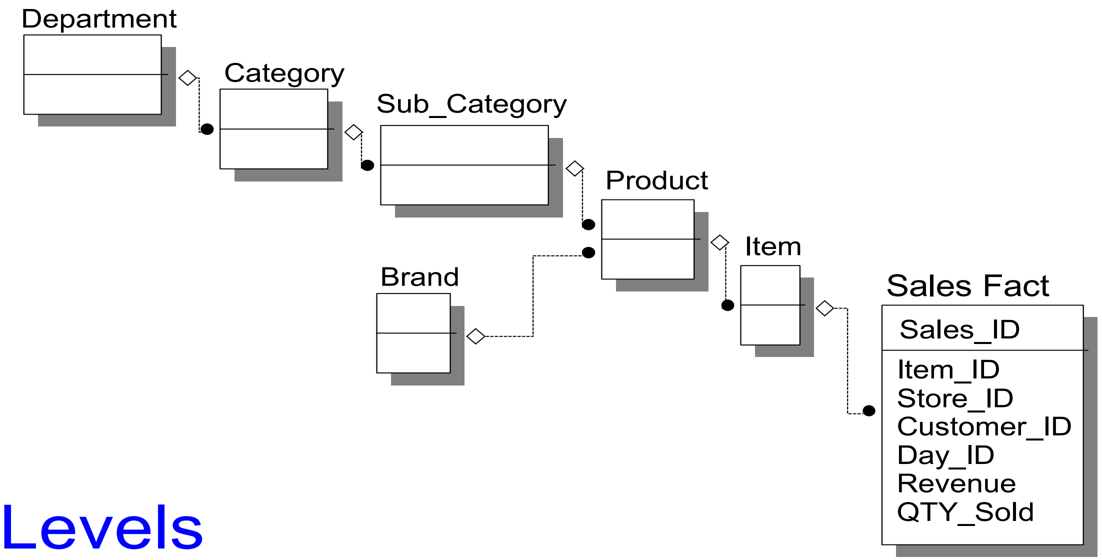
- Dimensions consist of one or more **dimension hierarchies**
- Examples: Hierarchies in the Product Dimension
  - Product Classification Hierarchy ("Merchandising Hierarchy")
  - Branding Hierarchy
  - ...



Dimension Hierarchies

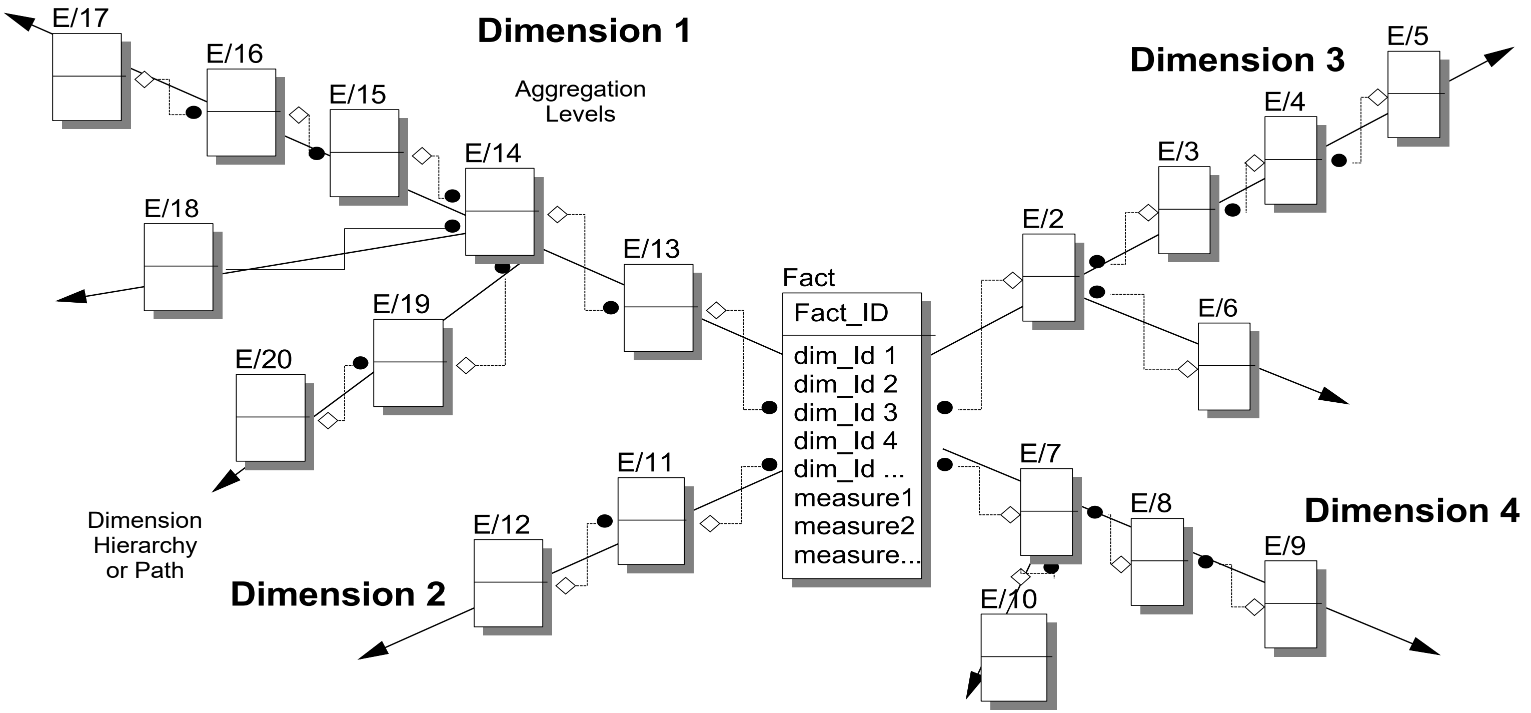
Multidimensional
Data Modeling - Base Concepts (6 of 6)

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

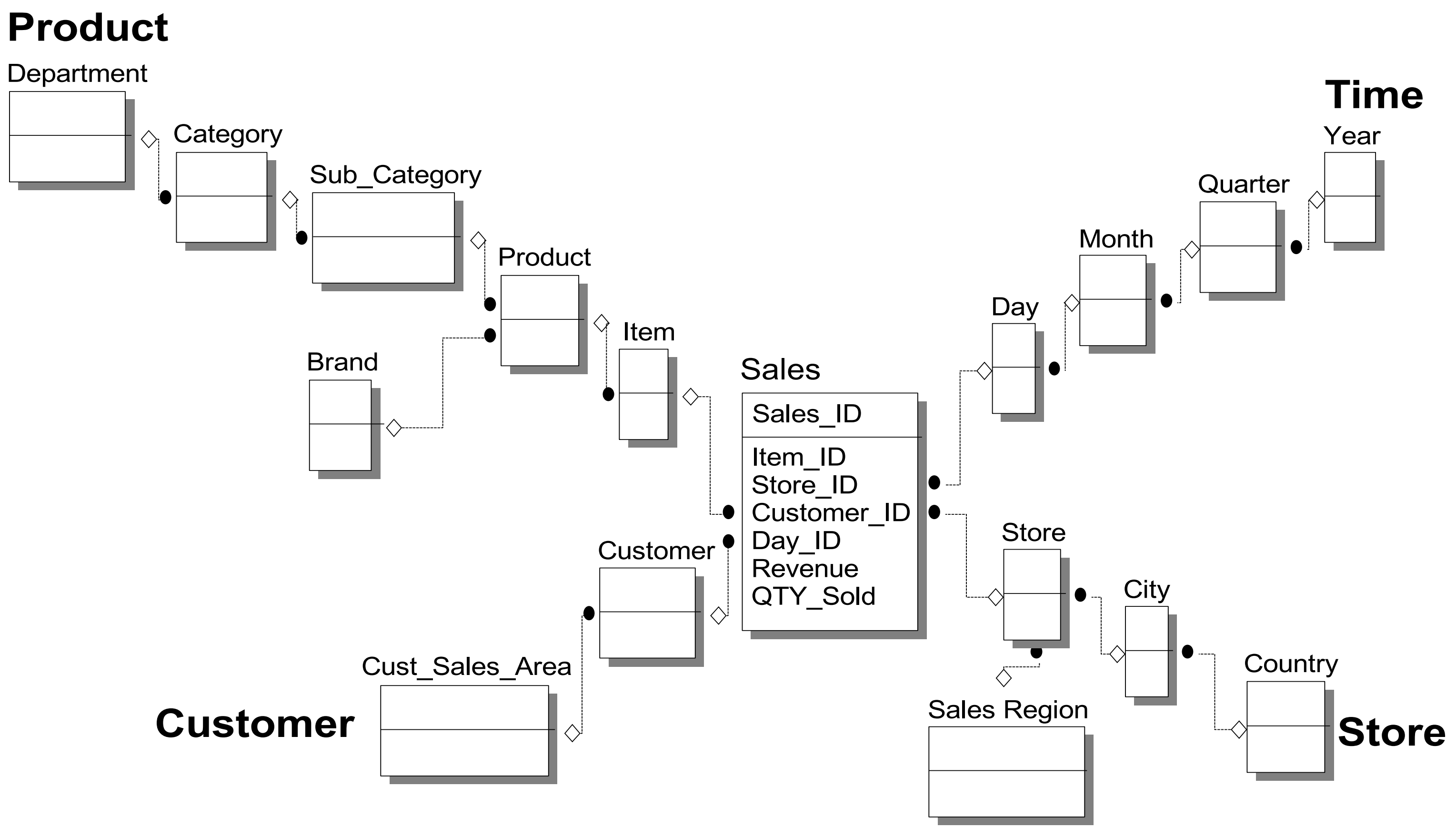


Aggregation Levels

Initial Multidimensional Model - Summary



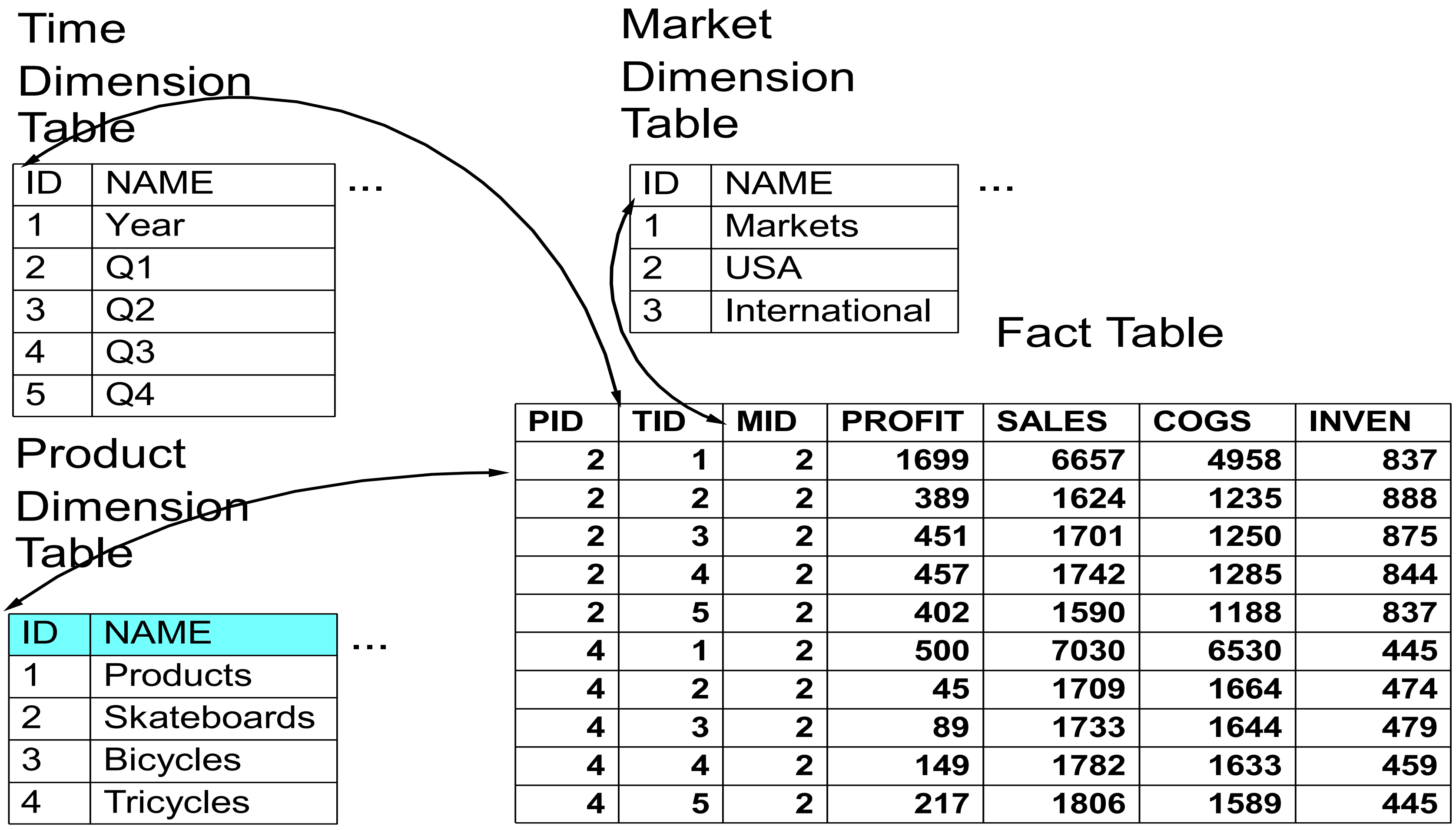
Initial Multidimensional Model - Example



What is a Star Schema ?

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

Star Schema Example



Time Dimension Table

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

…

Market Dimension Table

ID	NAME
1	Markets
2	USA
3	International

…

Fact Table

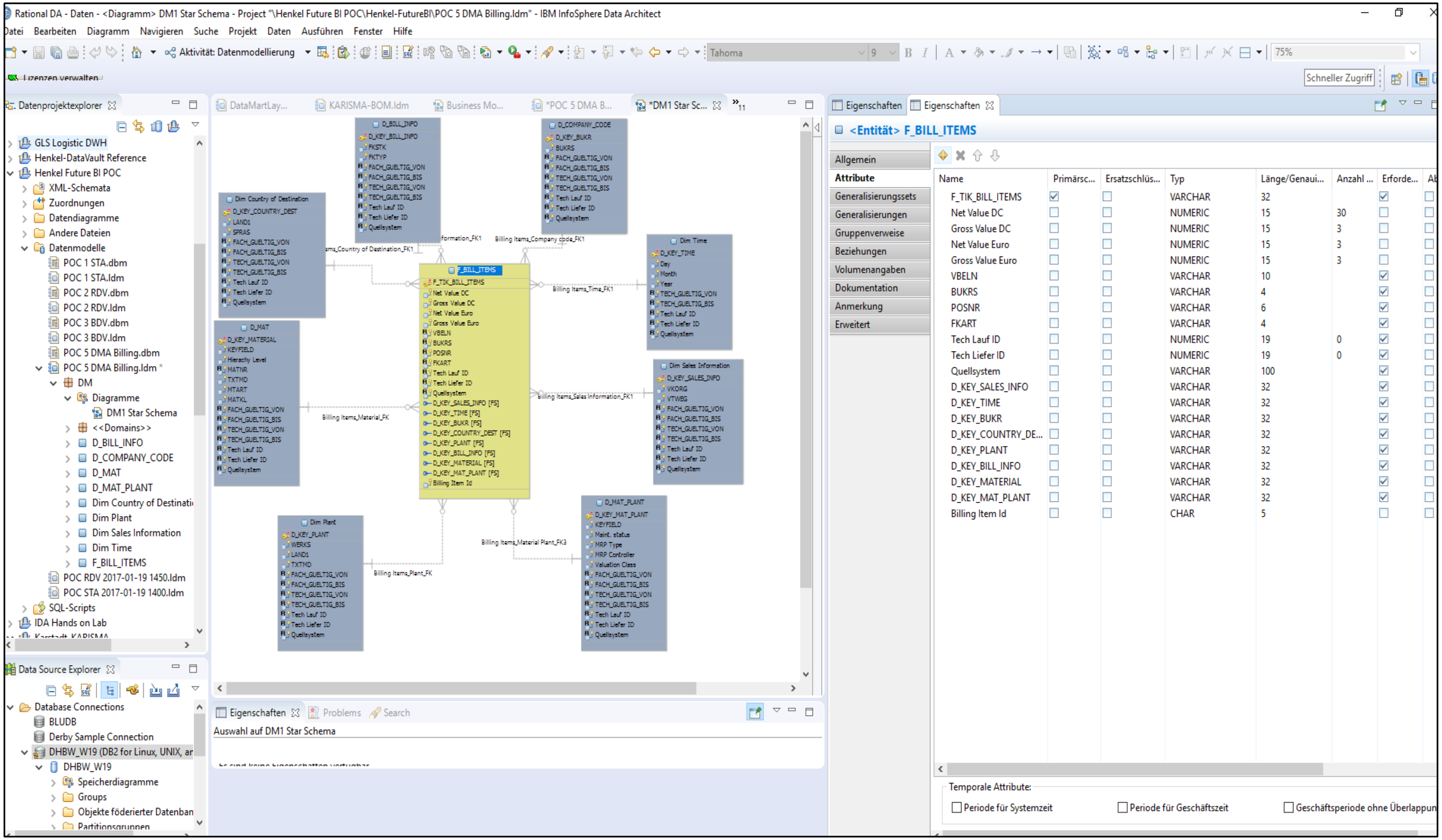
PID	TID	MID	PROFIT	SALES	COGS	INVEN
2	1	2	1699	6657	4958	837
2	2	2	389	1624	1235	888
2	3	2	451	1701	1250	875
2	4	2	457	1742	1285	844
2	5	2	402	1590	1188	837
4	1	2	500	7030	6530	445
4	2	2	45	1709	1664	474
4	3	2	89	1733	1644	479
4	4	2	149	1782	1633	459
4	5	2	217	1806	1589	445

Product Dimension Table

ID	NAME
1	Products
2	Skateboards
3	Bicycles
4	Tricycles

…

Demo1: IBM Infosphere Data Architect (IDA)



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

Auswahl auf DM1 Star Schema

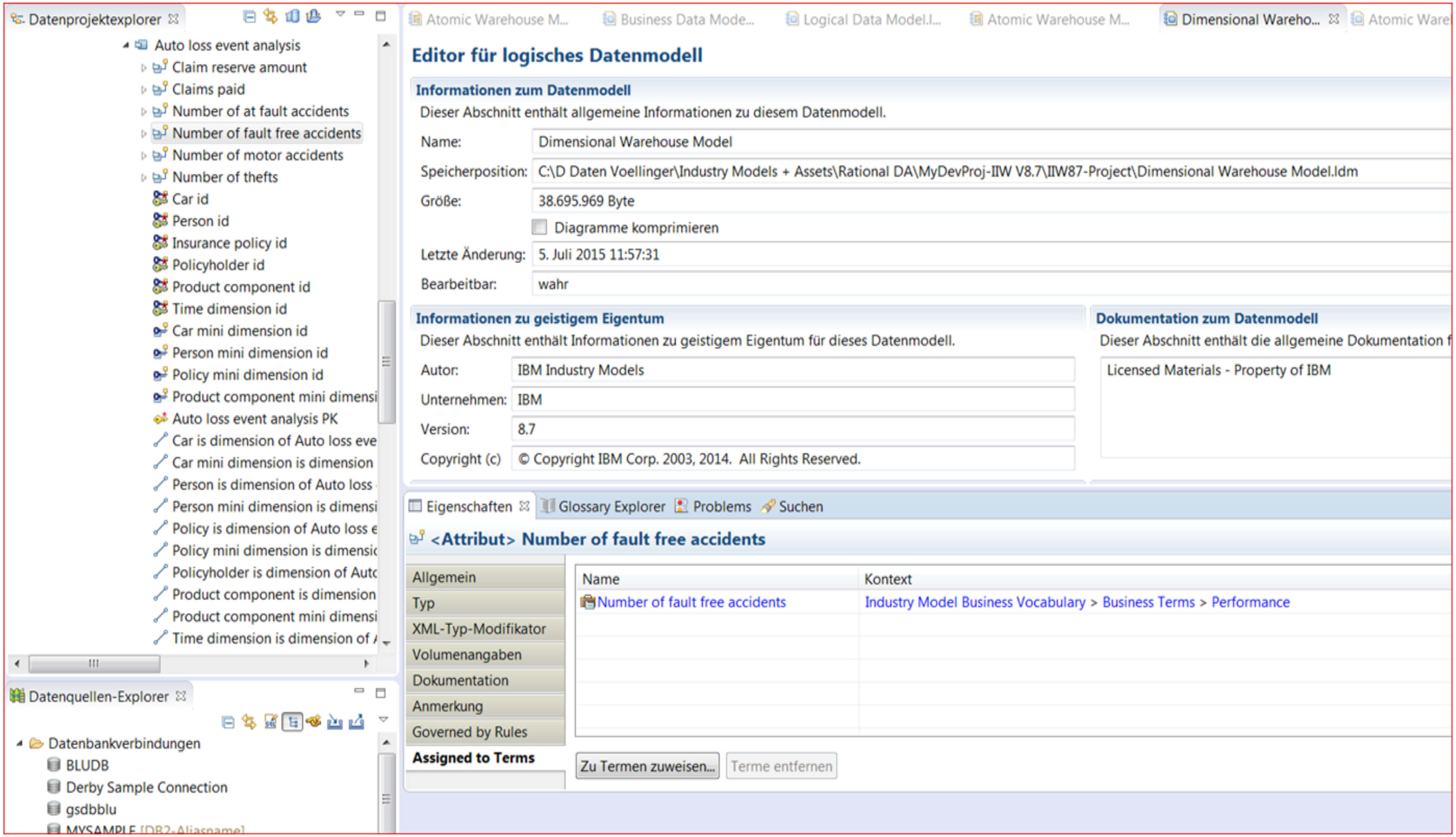
<Entität> F_BILL_ITEMS

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

Temporale Attribute:

□ Periode für Systemzeit □ Periode für Geschäftszeit □ Geschäftsperiode ohne Überlappun

Demo2: Eclipse Plugin “Bridge” of IGC and IDA



Editor für logisches Datenmodell

Informationen zum Datenmodell

Dieser Abschnitt enthält allgemeine Informationen zu diesem Datenmodell.

Name: Dimensional Warehouse Model

Speicherposition: C:\D Daten Voellinger\Industry Models + Assets\Rational DA\MyDevProj-IIW V8.7\IIW87-Project\Dimensional Warehouse Model.ldm

Größe: 38.695.969 Byte

□ Diagramme komprimieren

Letzte Änderung: 5. Juli 2015 11:57:31

Bearbeitbar: wahr

Informationen zu geistigem Eigentum

Dieser Abschnitt enthält Informationen zu geistigem Eigentum für dieses Datenmodell.

Autor: IBM Industry Models

Unternehmen: IBM

Version: 8.7

Copyright (c) © Copyright IBM Corp. 2003, 2014. All Rights Reserved.

Dokumentation zum Datenmodell

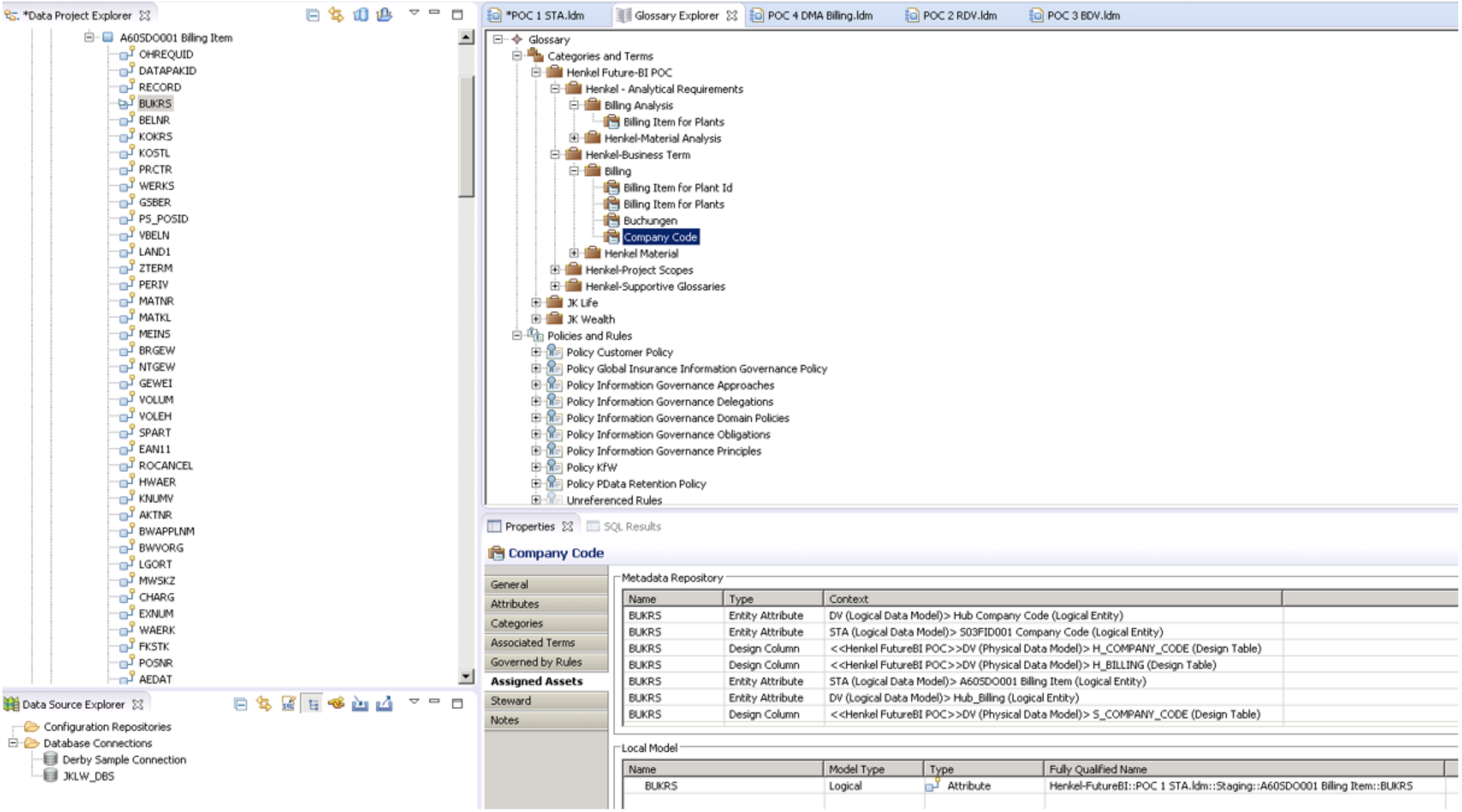
Dieser Abschnitt enthält die allgemeine Dokumentation f

Licensed Materials - Property of IBM

<Attribut> Number of fault free accidents

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

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



- A60SDO001 Billing Item
  - OHREQUID
  - DATAPAKID
  - RECORD
  - BUKRS
  - BELNR
  - KOKRS
  - KOSTL
  - PRCTR
  - WERKS
  - GSBER
  - PS_POSID
  - VBELN
  - LAND1
  - ZTERM
  - PERIV
  - MATNR
  - MATKL
  - MEINS
  - BRGEW
  - NTGEW
  - GEWEI
  - VOLUM
  - VOLEH
  - SPART
  - EAN11
  - ROCANCEL
  - HWAER
  - KNUMV
  - AKTNR
  - BWAPPLNM
  - BWVORG
  - LGORT
  - MWSKZ
  - CHARG
  - EXNUM
  - WAERK
  - FKSTK
  - POSNR
  - AEDAT

Data Source Explorer

- Configuration Repositories
- Database Connections
  - Derby Sample Connection
  - JKLW_DBS

Glossary Explorer

- Glossary
  - Categories and Terms
    - Henkel Future-BI POC
      - Henkel - Analytical Requirements
        - Billing Analysis
          - Billing Item for Plants
        - Henkel-Material Analysis
      - Henkel-Business Term
        - Billing
          - Billing Item for Plant Id
          - Billing Item for Plants
          - Buchungen
          - Company Code
        - Henkel Material
      - Henkel-Project Scopes
      - Henkel-Supportive Glossaries
    - JK Life
    - JK Wealth
  - Policies and Rules
    - Policy Customer Policy
    - Policy Global Insurance Information Governance Policy
    - Policy Information Governance Approaches
    - Policy Information Governance Delegations
    - Policy Information Governance Domain Policies
    - Policy Information Governance Obligations
    - Policy Information Governance Principles
    - Policy KfW
    - Policy PData Retention Policy
    - Unreferenced Rules

Company Code

- General
- Attributes
- Categories
- Associated Terms
- Governed by Rules
- **Assigned Assets**
- Steward
- Notes

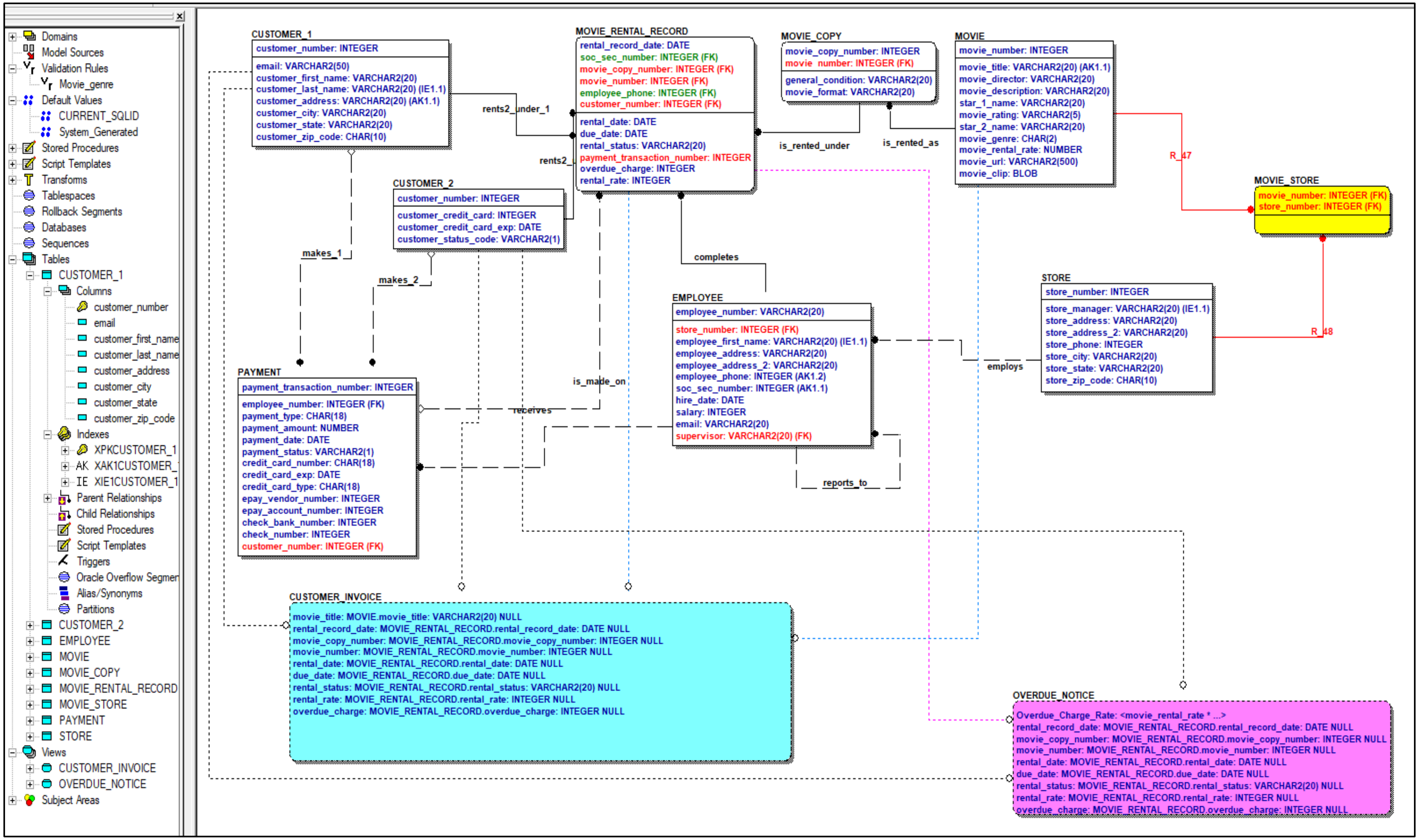
Metadata Repository

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

Local Model

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

Demo2: erwin Data Modeler (eDM)



- Domains
- Model Sources
- Validation Rules
  - Movie_genre
- Default Values
  - CURRENT_SQLID
  - System_Generated
- Stored Procedures
- Script Templates
- Transforms
- Tablespaces
- Rollback Segments
- Databases
- Sequences
- Tables
  - CUSTOMER_1
    - Columns
      - customer_number
      - email
      - customer_first_name
      - customer_last_name
      - customer_address
      - customer_city
      - customer_state
      - customer_zip_code
    - Indexes
      - XPKCUSTOMER_1
      - AK XAK1CUSTOMER_
      - IE XIE1CUSTOMER_1
    - Parent Relationships
    - Child Relationships
    - Stored Procedures
    - Script Templates
    - Triggers
    - Oracle Overflow Segmer
    - Alias/Synonyms
    - Partitions
  - CUSTOMER_2
  - EMPLOYEE
  - MOVIE
  - MOVIE_COPY
  - MOVIE_RENTAL_RECORD
  - MOVIE_STORE
  - PAYMENT
  - STORE
- Views
  - CUSTOMER_INVOICE
  - OVERDUE_NOTICE
- Subject Areas

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

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

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

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

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

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

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

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

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

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

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

Exercise 1 to Lesson 5

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

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

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

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

Exercise 2 to Lesson 5

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

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

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

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

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

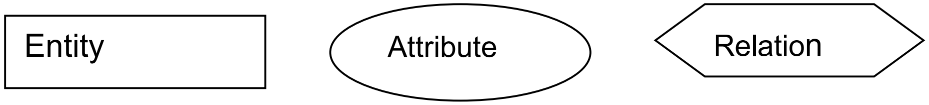
Exercise 3 to Lesson 5

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

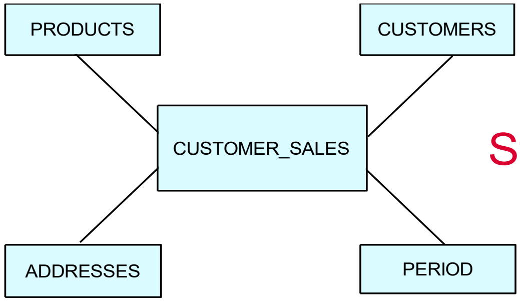
The following objects should be administered by the new ordering system.

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

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

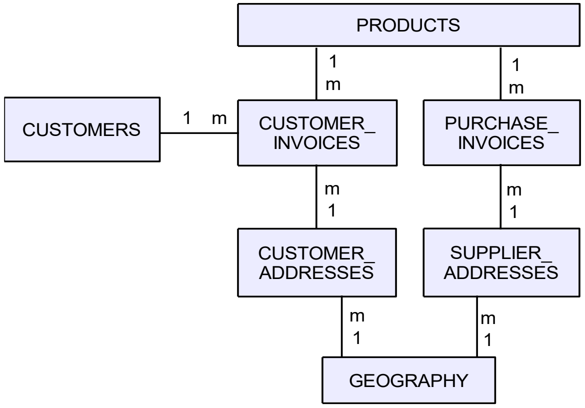


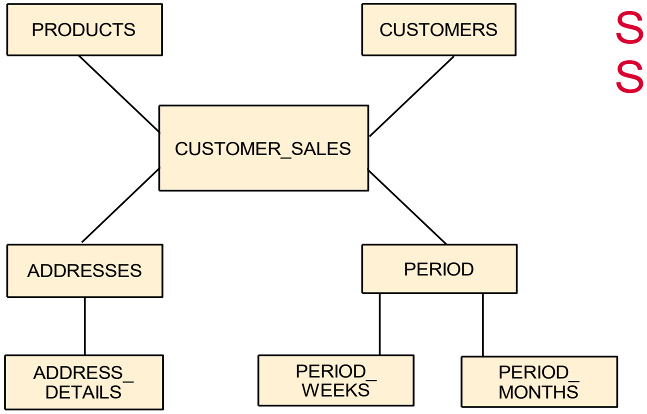
Appendix to MDDM Lesson Exercises



Star Schema

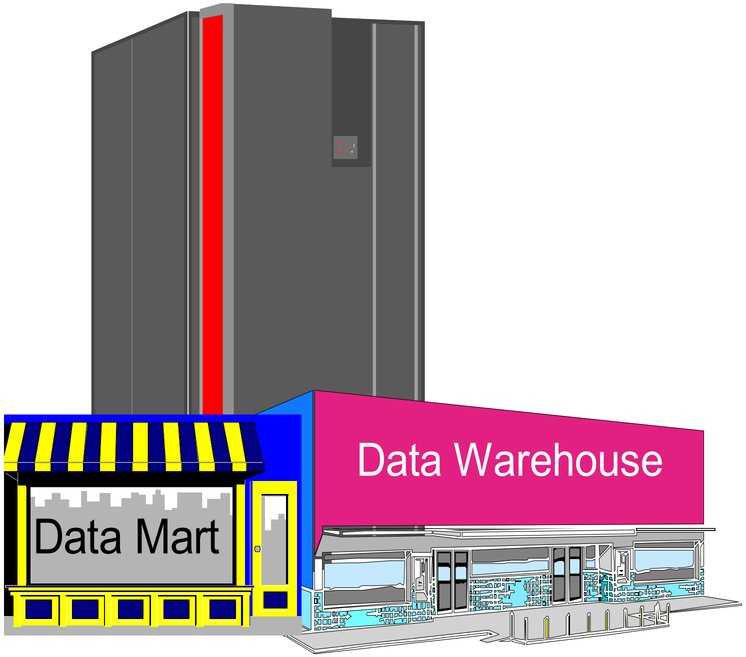
Entity-Relationship



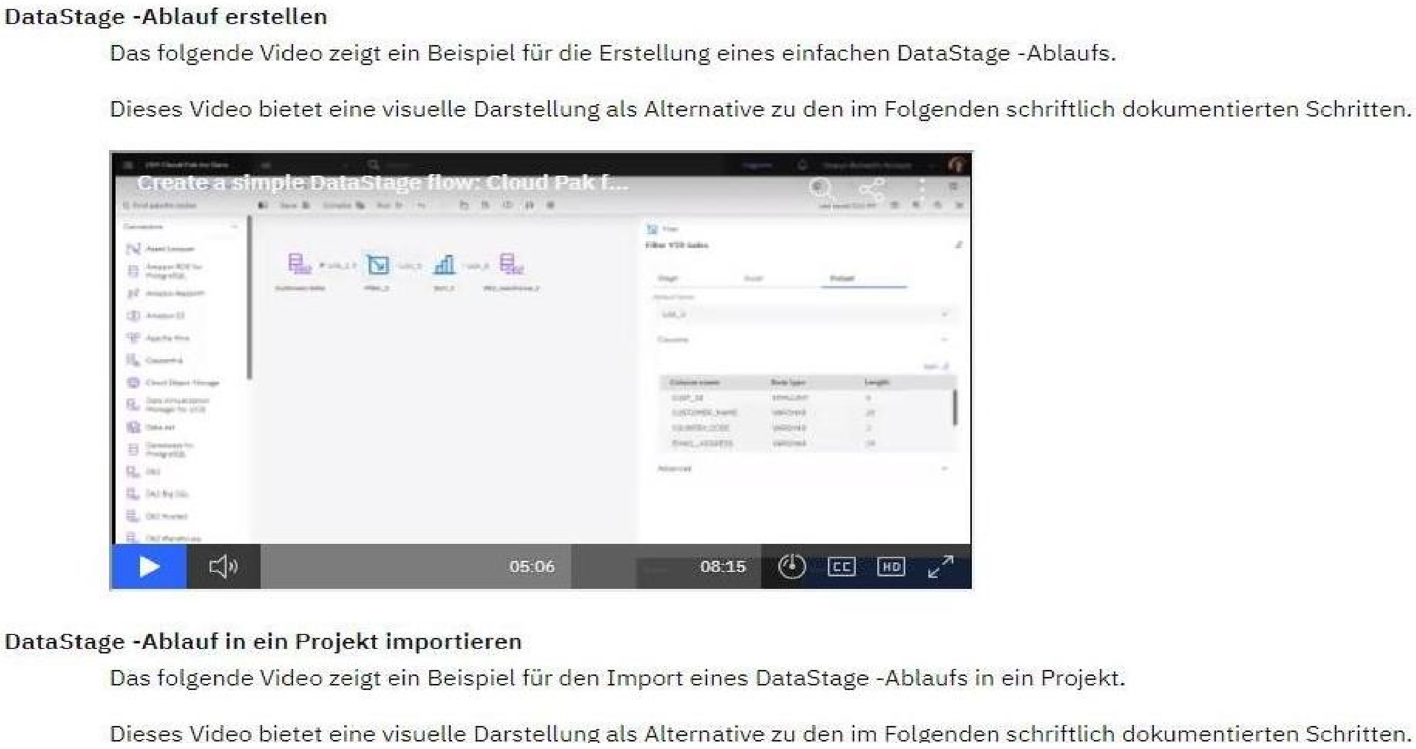


Snowflake Schema

DW06 - ETL Reference Architecture



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



DataStage -Ablauf erstellen

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

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

DataStage -Ablauf in ein Projekt importieren

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

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

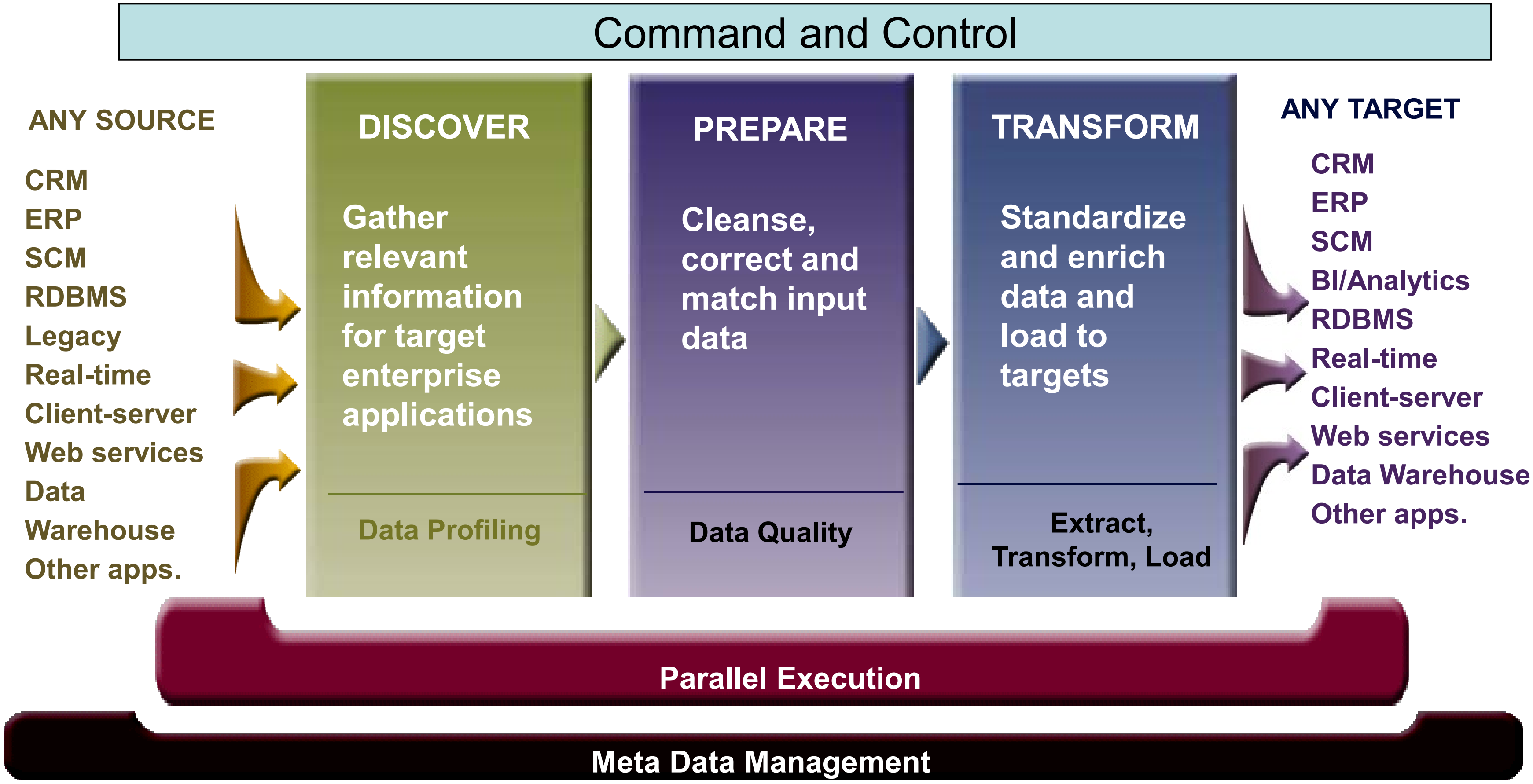
Remark: You can see the video also without being connected to IBM Cloud:
https://dataplatform.cloud.ibm.com/docs/content/wsj/getting-started/videos.html?audience=cpdaas&context=cpdaas#data-engineers

Experience shows that …

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

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

3 Steps for a successful Data Population Strategy



Command and Control

ANY SOURCE

CRM
ERP
SCM
RDBMS
Legacy
Real-time
Client-server
Web services
Data Warehouse
Other apps.

DISCOVER

Gather relevant information for target enterprise applications

Data Profiling

PREPARE

Cleanse, correct and match input data

Data Quality

TRANSFORM

Standardize and enrich data and load to targets

Extract, Transform, Load

ANY TARGET

CRM
ERP
SCM
BI/Analytics
RDBMS
Real-time
Client-server
Web services
Data Warehouse
Other apps.

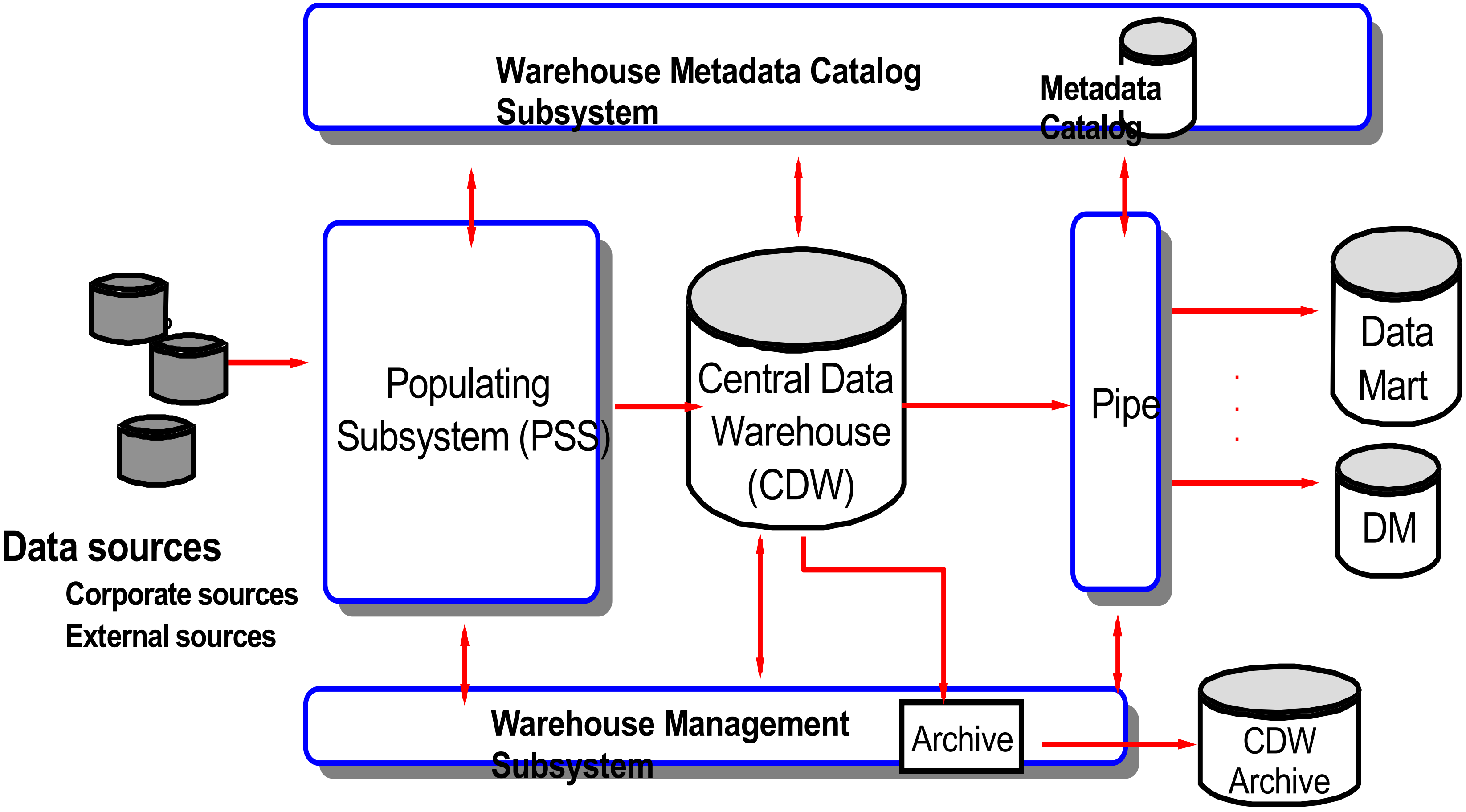
Parallel Execution

Meta Data Management

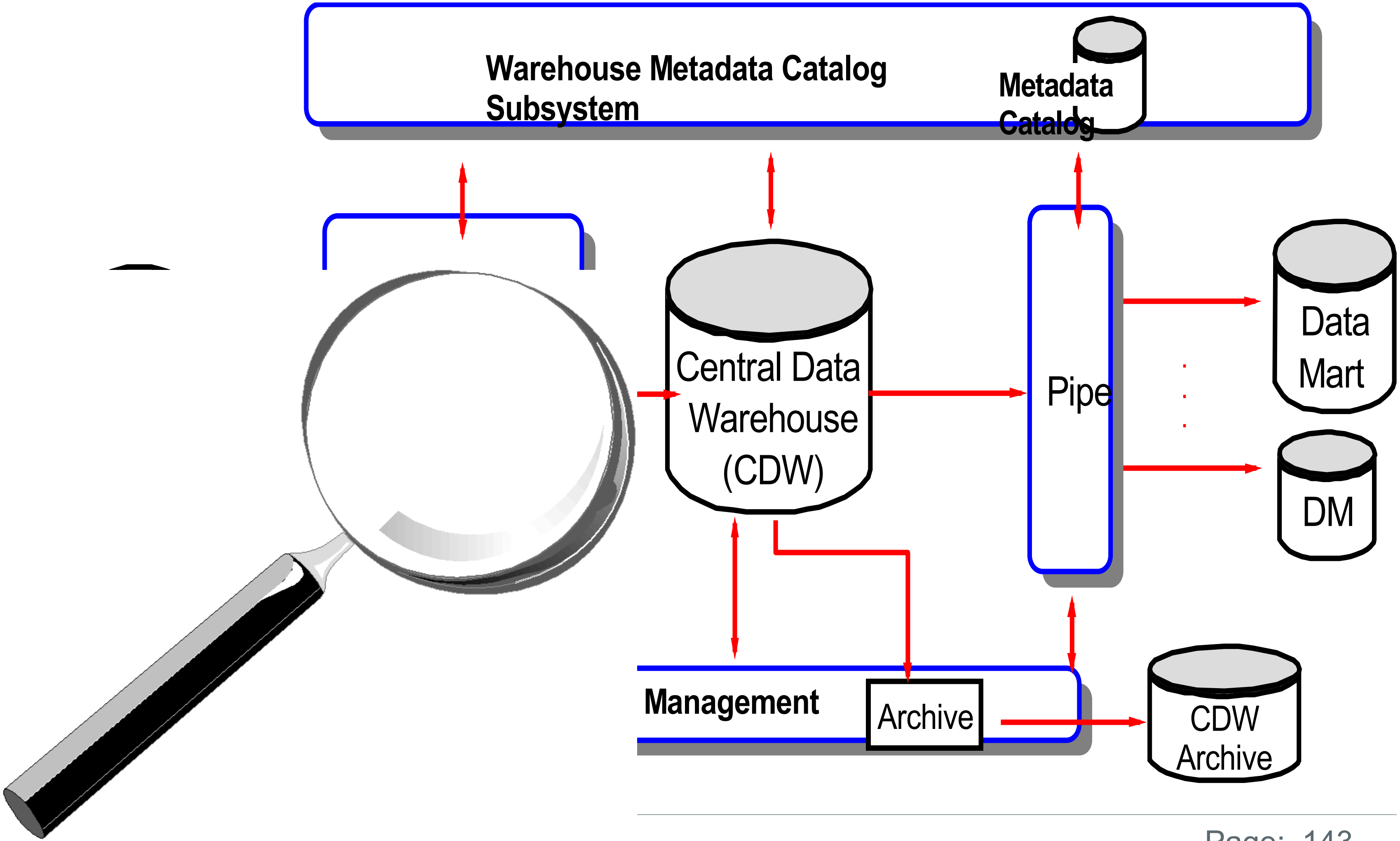
Tools: Informatica Axon DQ (formerly Evoke-AXIO) Precisely-Trillium (formerly HarteHanks) Informatica - PowerCenter

IBM Infosphere Inform. Server (IIS)-ProfileStage IIS - QualityStage IIS - DataStage

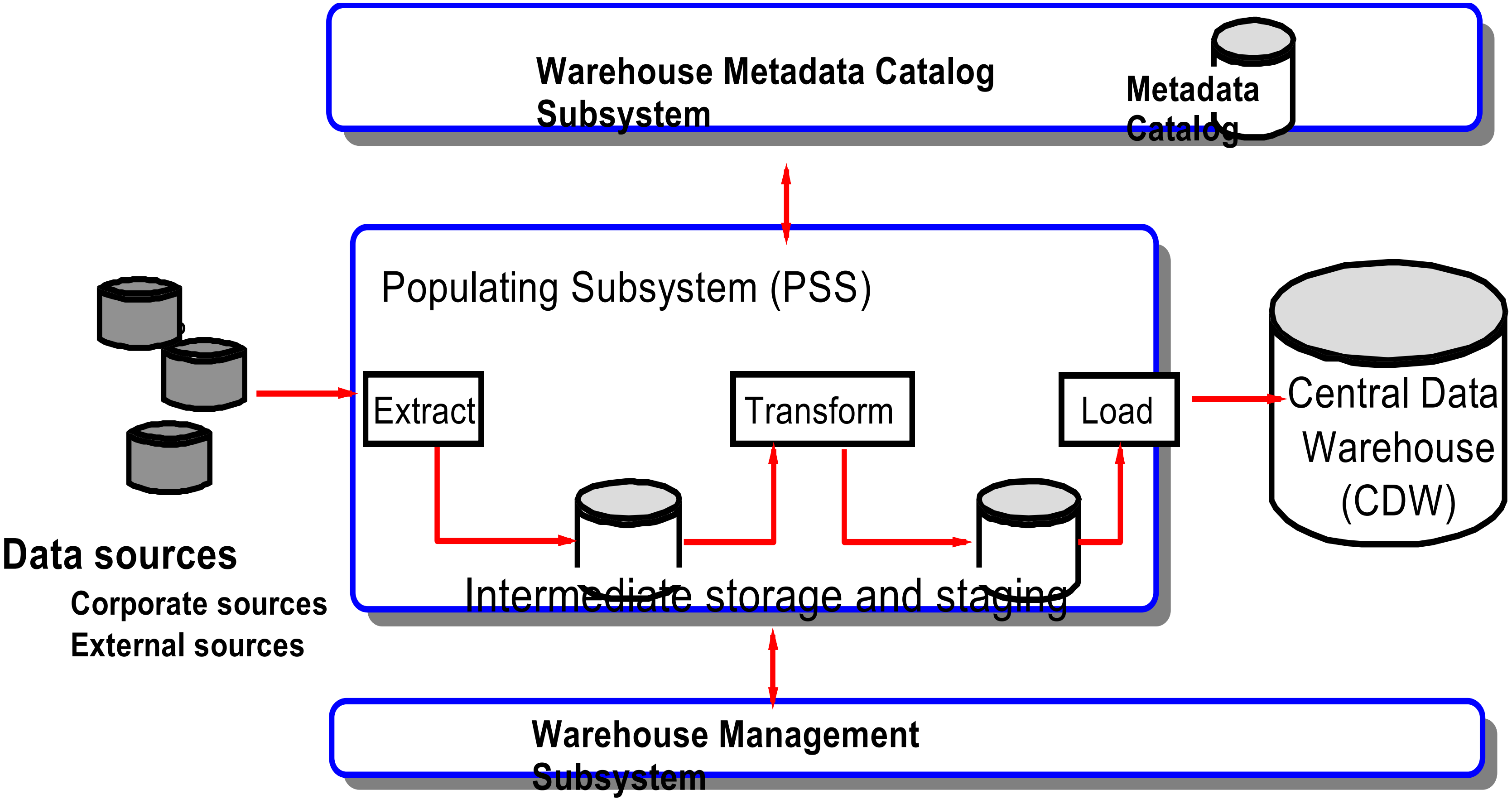
ETL-Reference Architecture - DWH Overview


Warehouse Metadata Catalog Subsystem
Metadata Catalog
Populating Subsystem (PSS)
Central Data Warehouse (CDW)
Pipe
Data Mart
DM
Data sources
Corporate sources
External sources
Warehouse Management Subsystem
Archive
CDW Archive

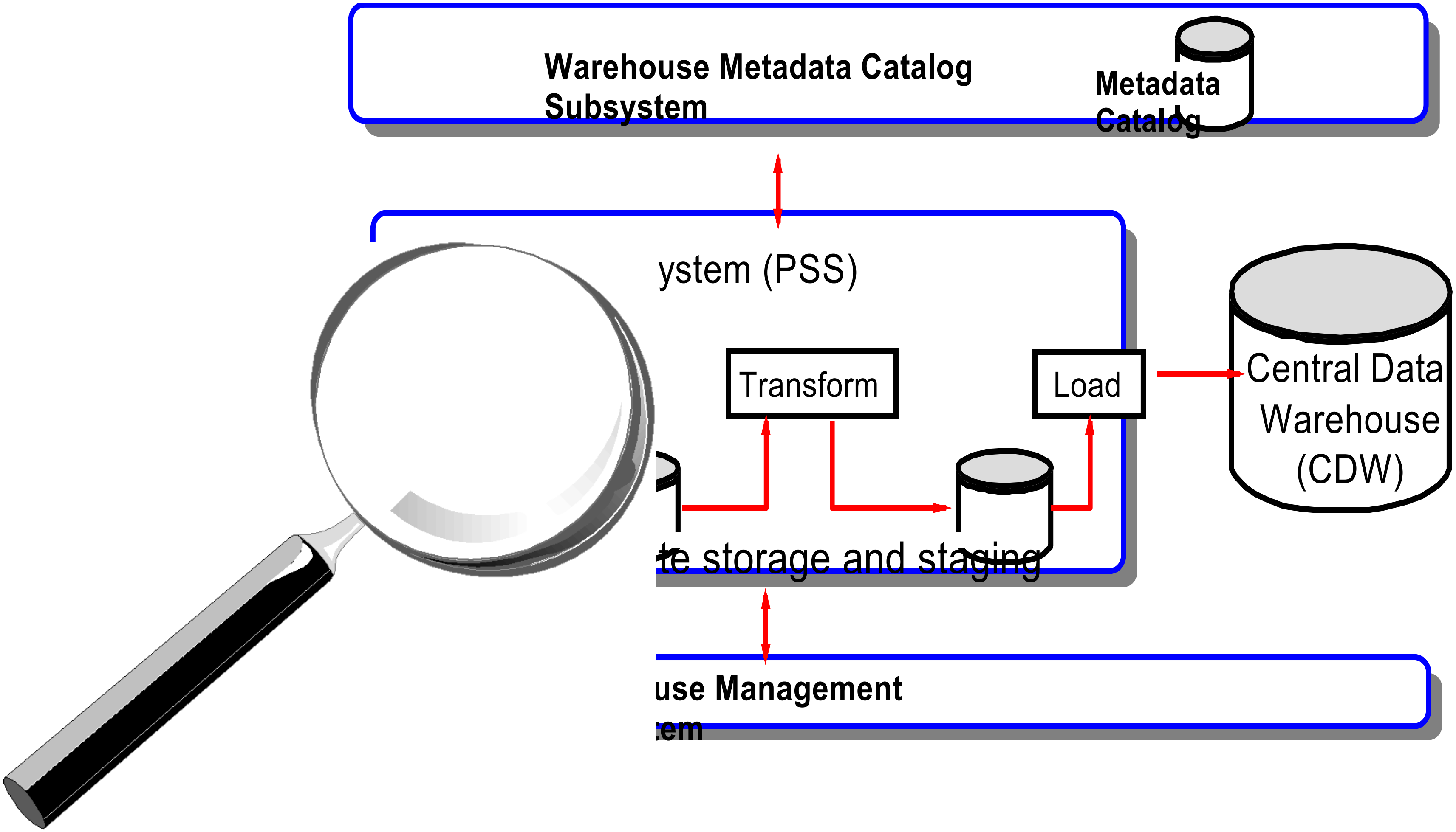
ETL-Reference Architecture - Focus on PSS


Warehouse Metadata Catalog Subsystem
Metadata Catalog
Central Data Warehouse (CDW)
Pipe
Data Mart
DM
Management
Archive
CDW Archive

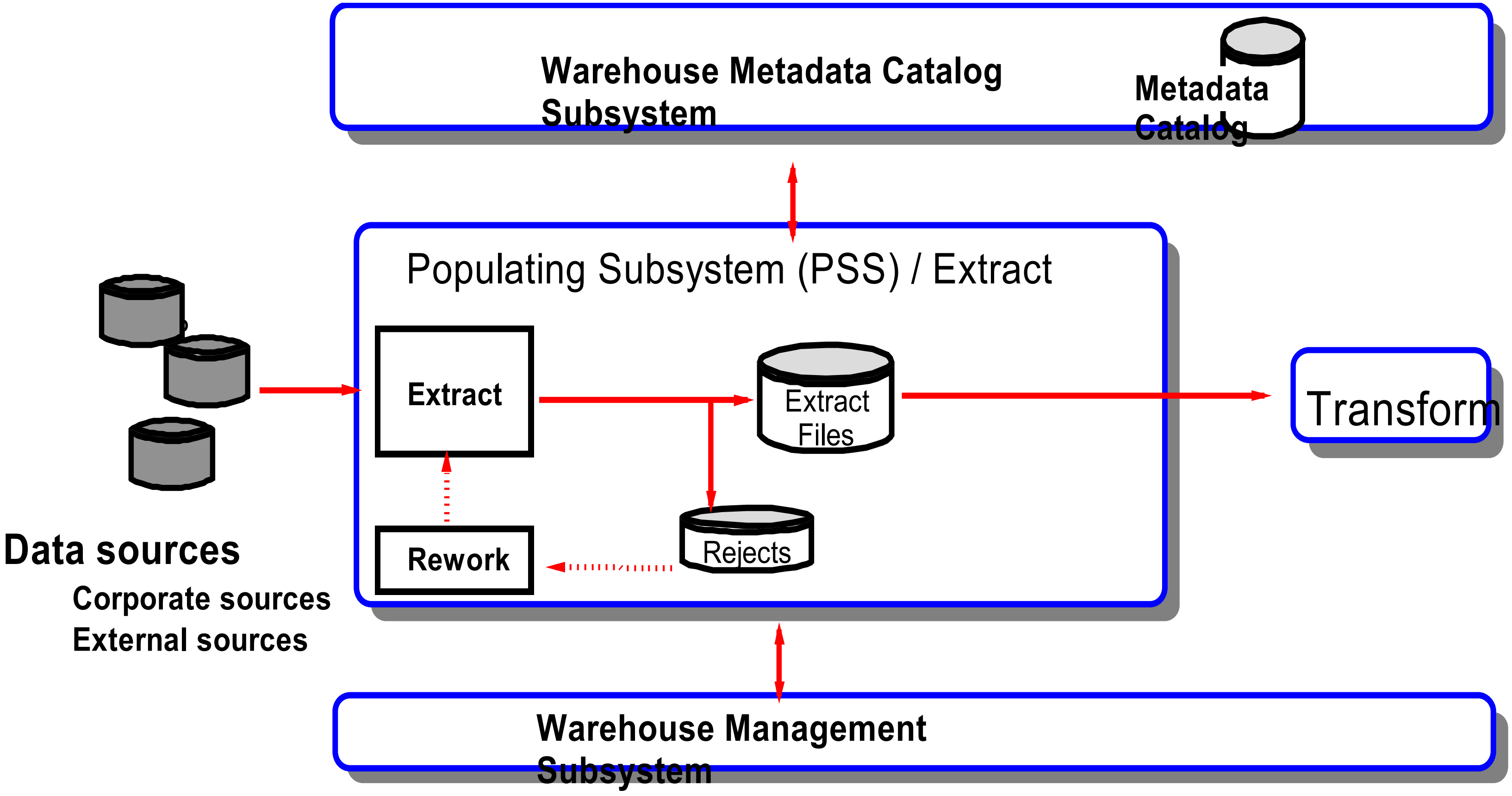
ETL-Reference Architecture – PPS Processes


Warehouse Metadata Catalog Subsystem
Metadata Catalog
Populating Subsystem (PSS)
Extract
Transform
Load
Intermediate storage and staging
Central Data Warehouse (CDW)
Data sources
Corporate sources
External sources
Warehouse Management Subsystem

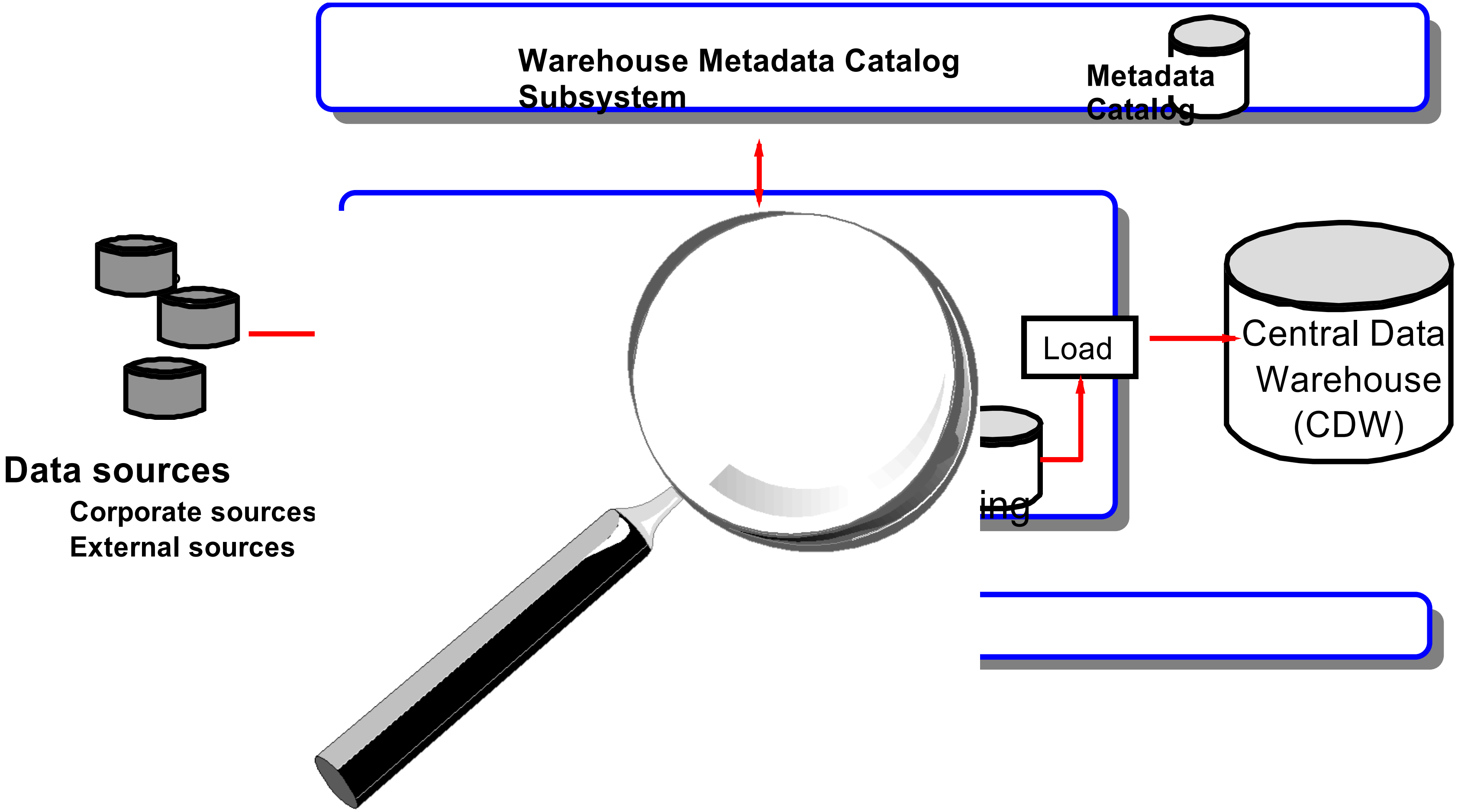
ETL-Reference Architecture – Extract Process



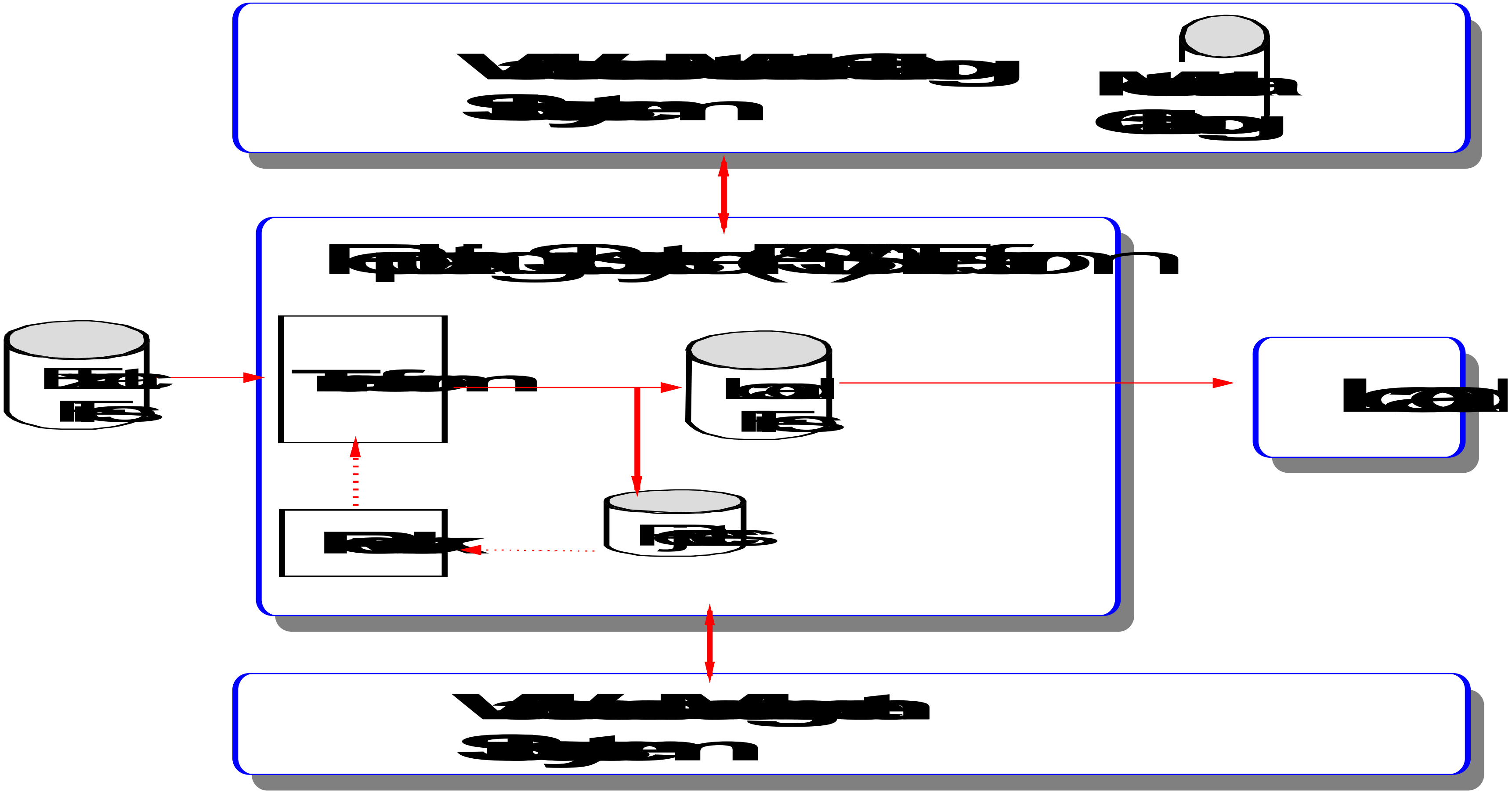
ETL-Reference Architecture - Extract Process



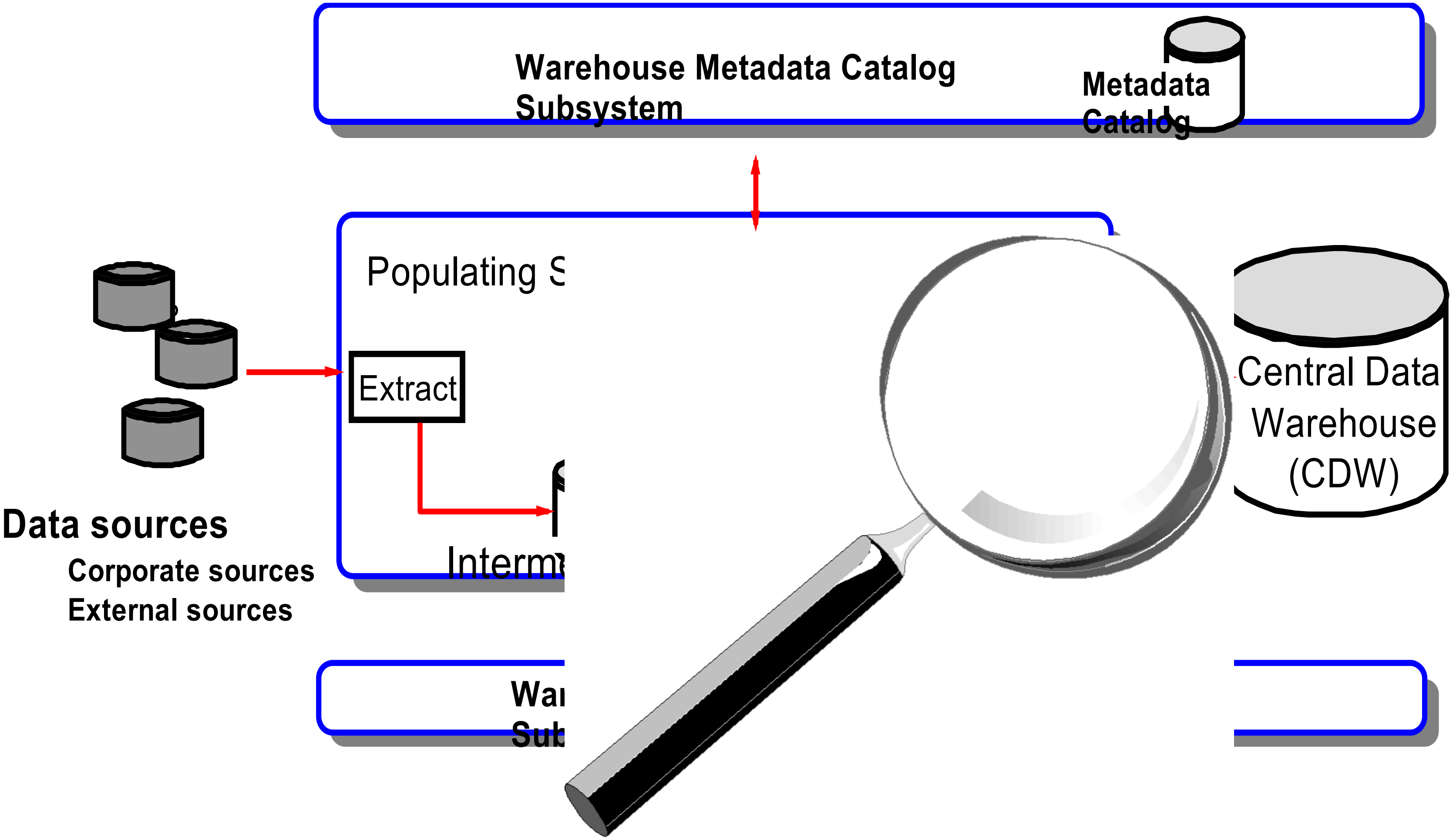
ETL-Reference Architecture–Transform Process



ETL-Reference Architecture-Transform Process



ETL-Reference Architecture – Load Process



Warehouse Metadata Catalog
Subsystem

Metadata
Catalog

Populating S

Extract

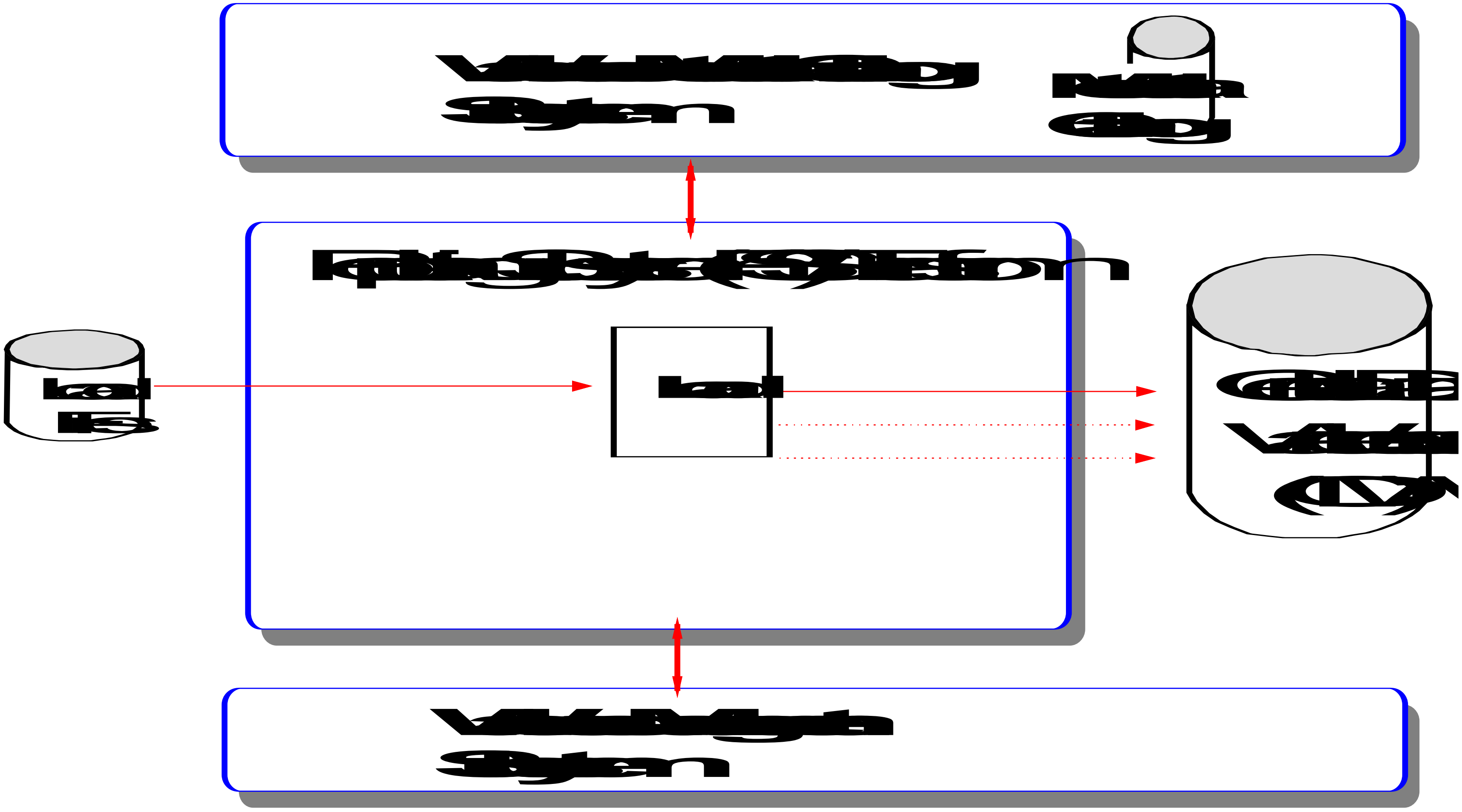
Interme

Central Data
Warehouse
(CDW)

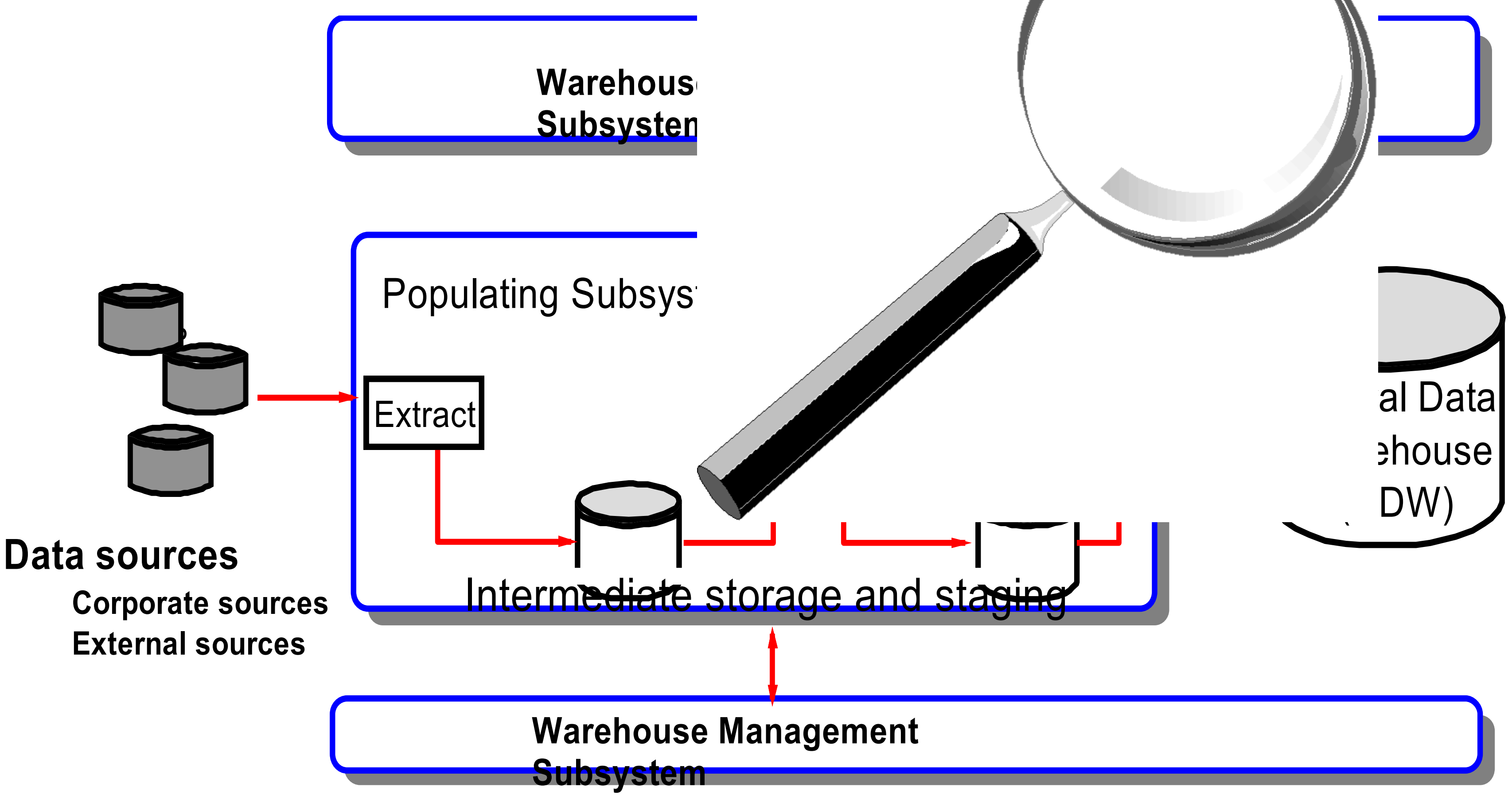
Data sources
Corporate sources
External sources

Wa
Sub

ETL-Reference Architecture – Load Process



ETL-Reference Archit log



Data sources

Corporate sources
External sources

ETL-Reference Architecture - Metadata Subsystem

Metadata sources

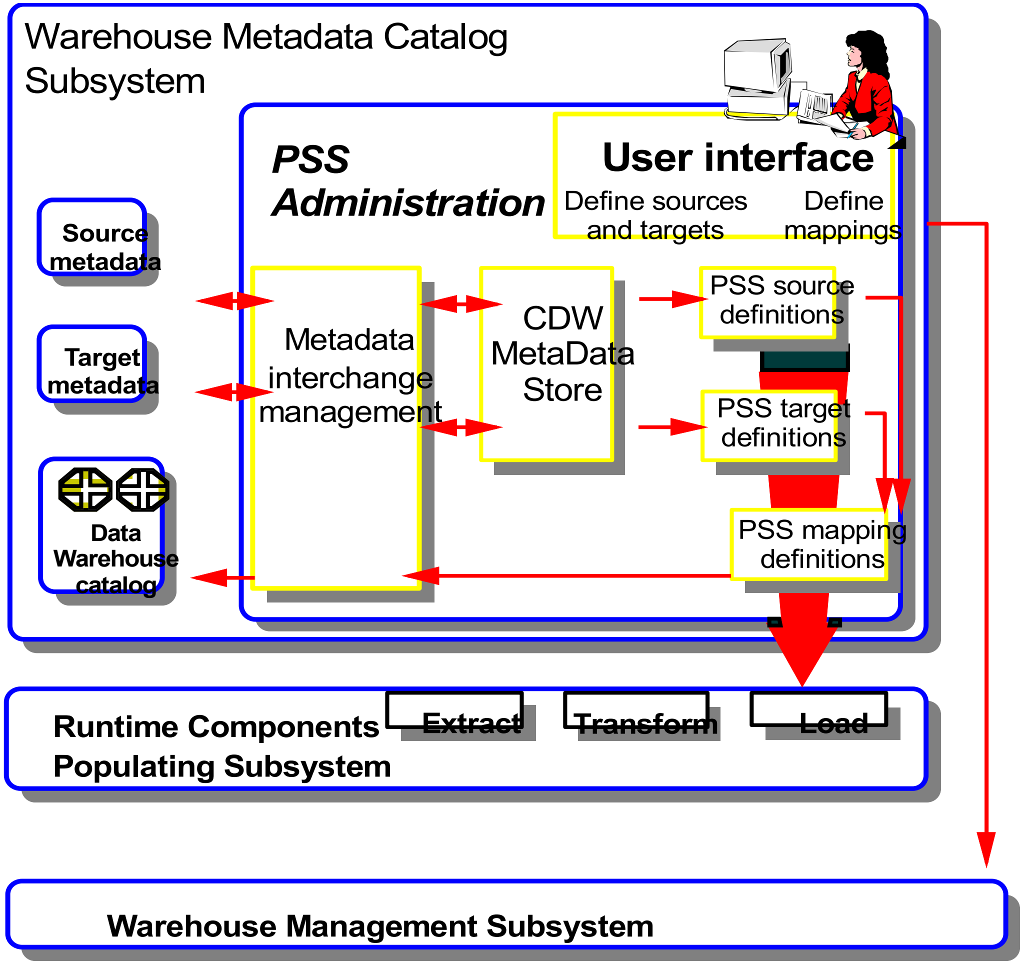
- Data modelling tools
- Database catalogs
- Record definitions in programs
- Populating tools

Metadata Outputs

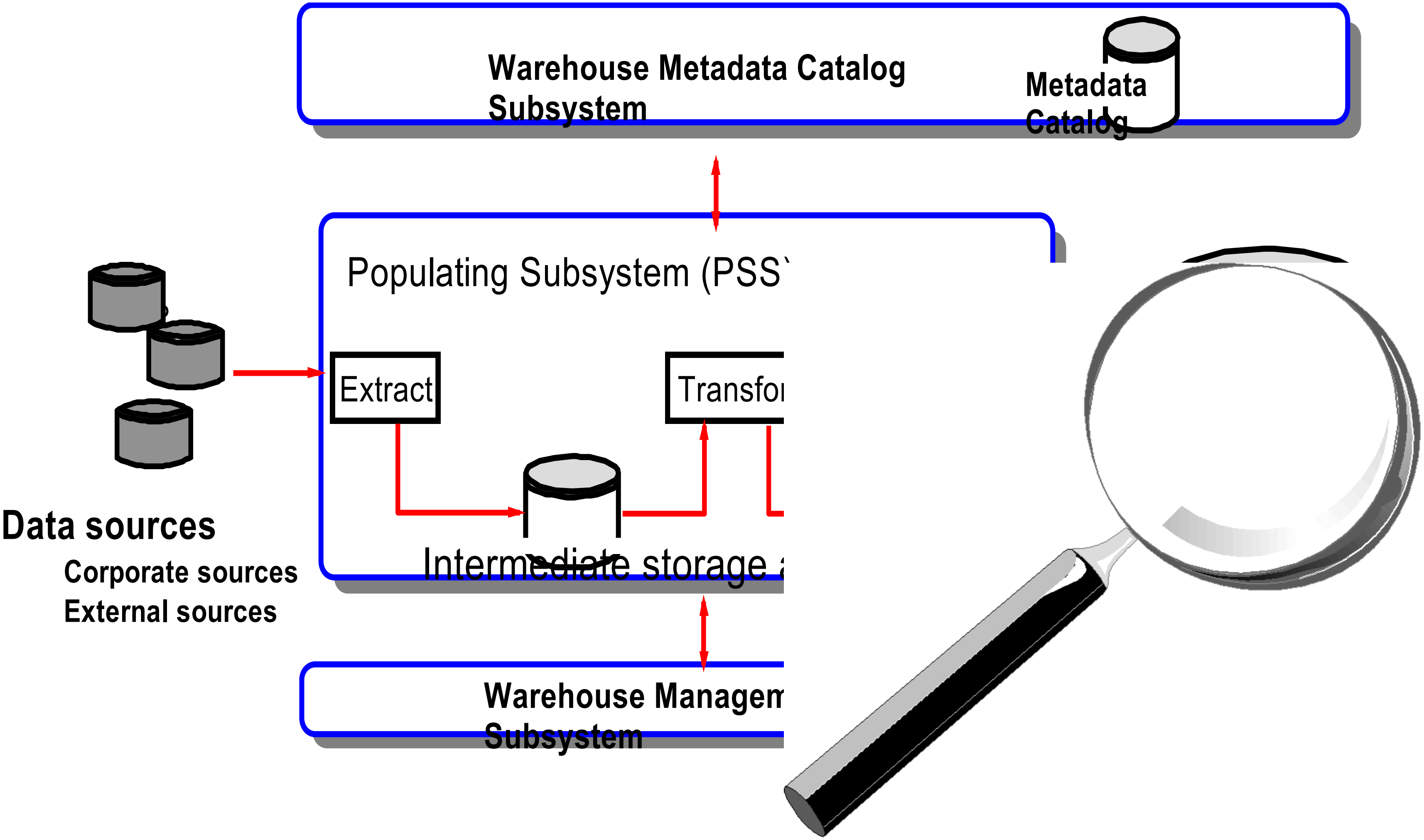
- PSS runtime statistics
- Data Warehouse catalog
- Process management

Issues

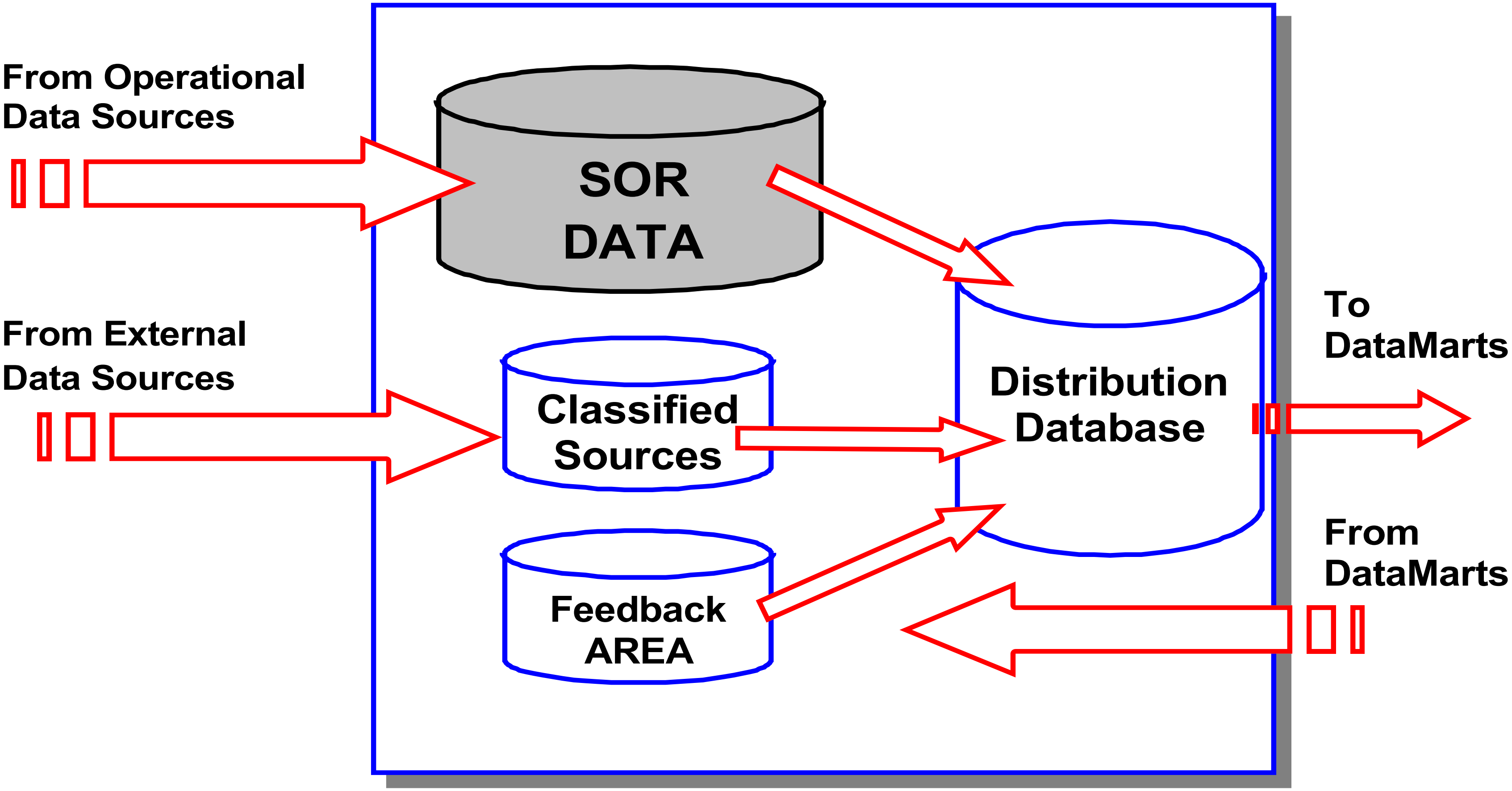
- Metadata access
- Metadata synchronization
  - Metadata Interchange
  - CDW Metadata store
- Today's tools provide little or no support



ETL-Reference Architecture – Central DWH

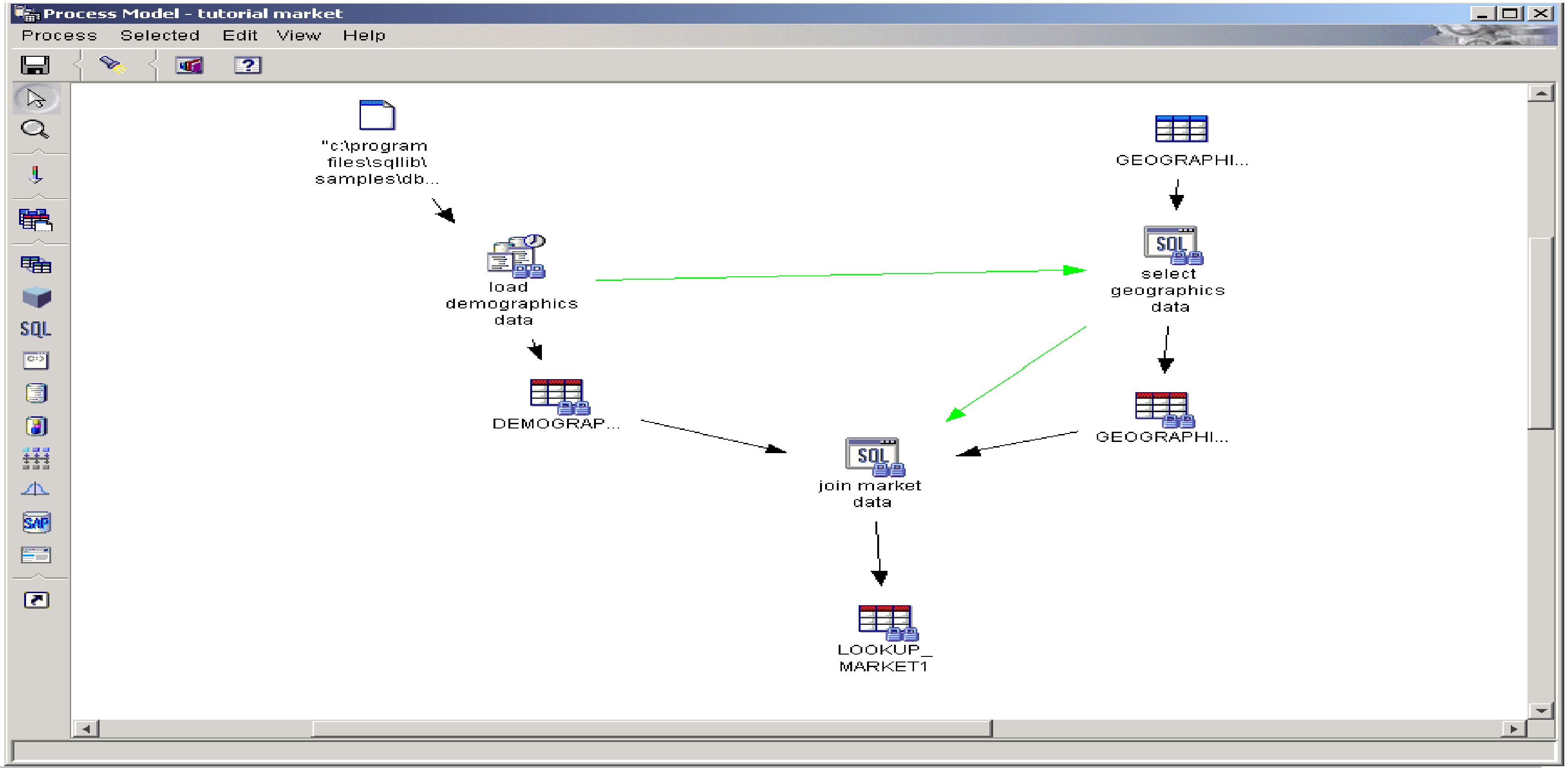


ETL-Reference Architecture - CDW Data Feeds



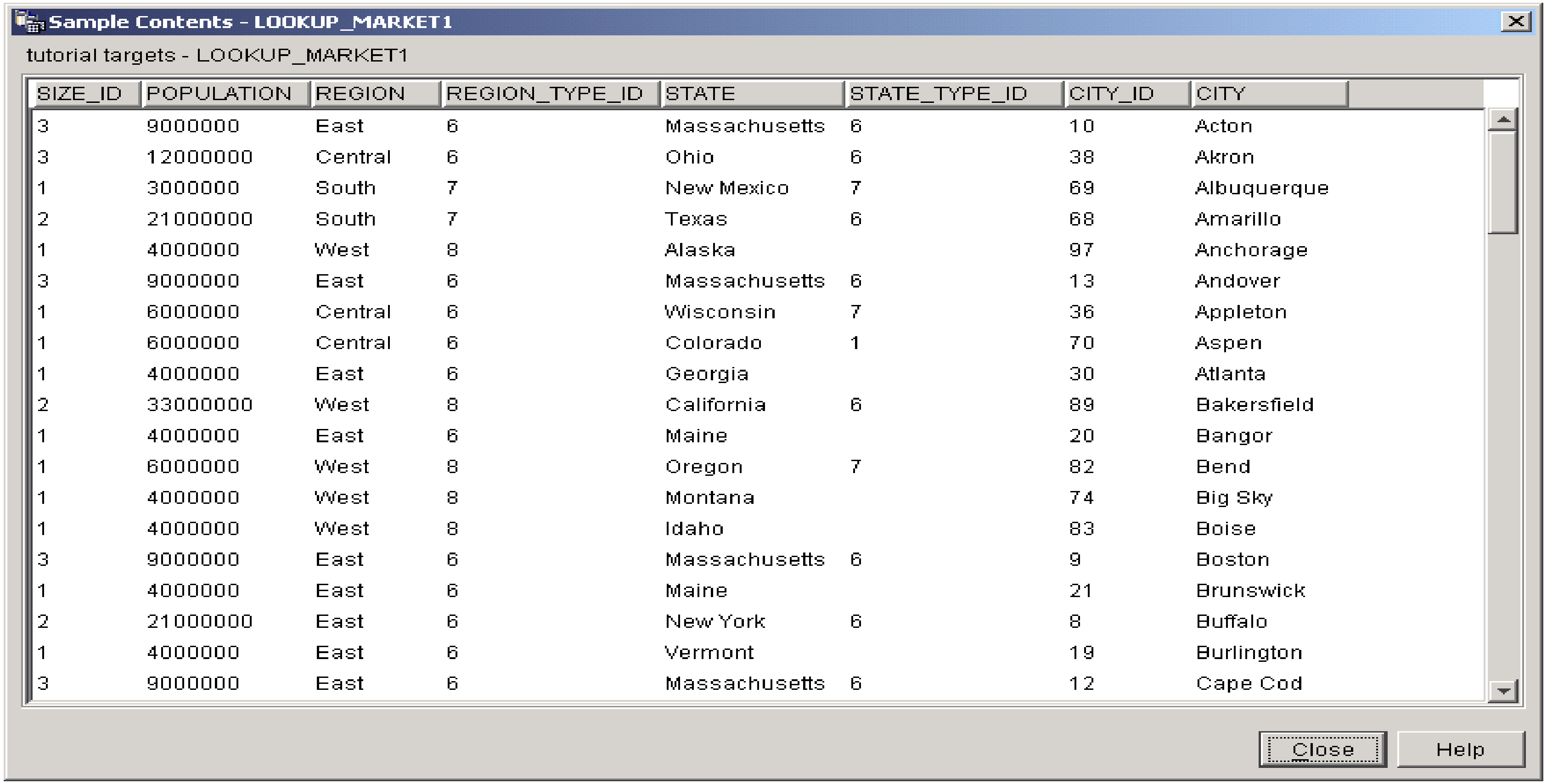
Exercise1 to Lesson 6 - DB2 WH-Manager (Part1)

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



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

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



tutorial targets - LOOKUP_MARKET1

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

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

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

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

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

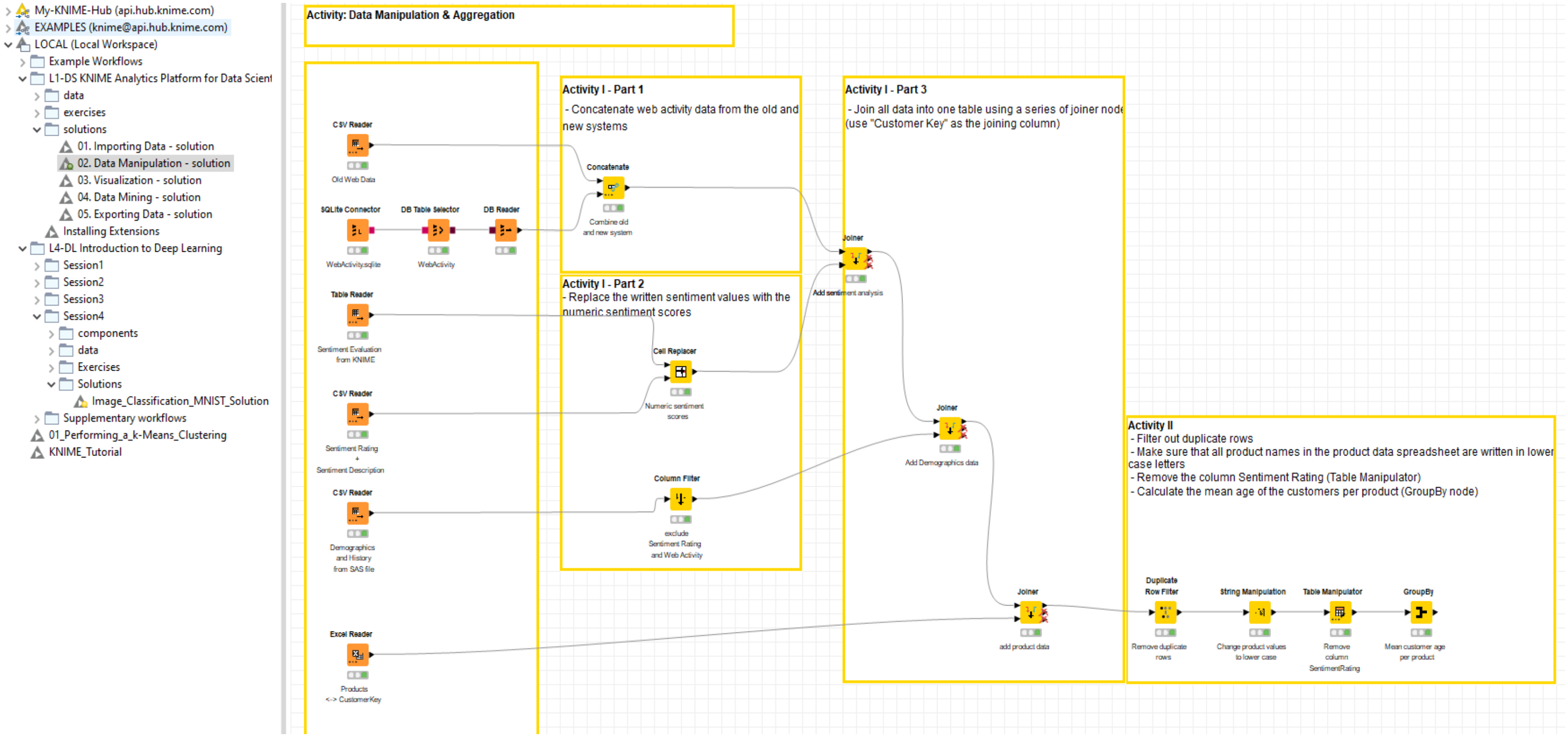
1. **DISCOVER**: Evoke-AXIO (now Informatica), Talend - Open Studio, IBM Infosphere Inform. Sever (IIS) – ProfileStage, or ????
2. **PREPARE**: HarteHanks-Trillium, Vality-Integrity, IBM Infosphere Inform. Server (IIS) – QualityStage, or ??????

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

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

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

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


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

Activity: Data Manipulation & Aggregation

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

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

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

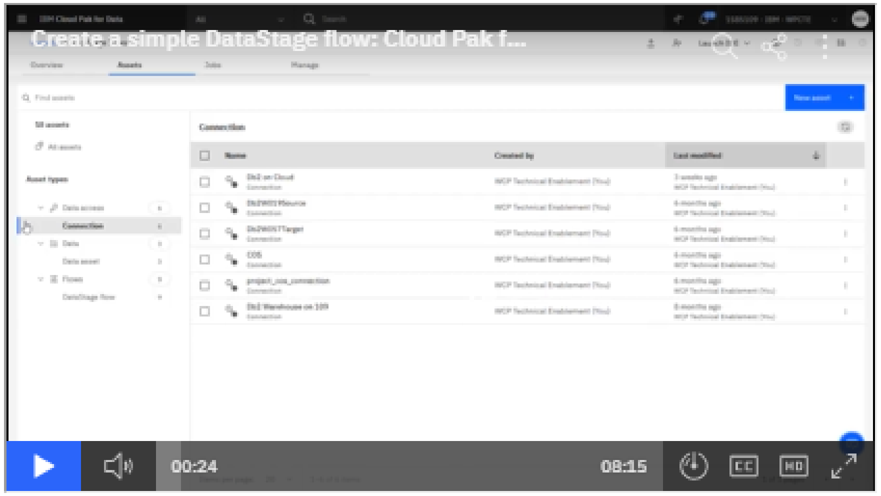
Activity II
- Filter out duplicate rows
- Make sure that all product names in the product data spreadsheet are written in lower case letters
- Remove the column Sentiment Rating (Table Manipulator)
- Calculate the mean age of the customers per product (GroupBy node)

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

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

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

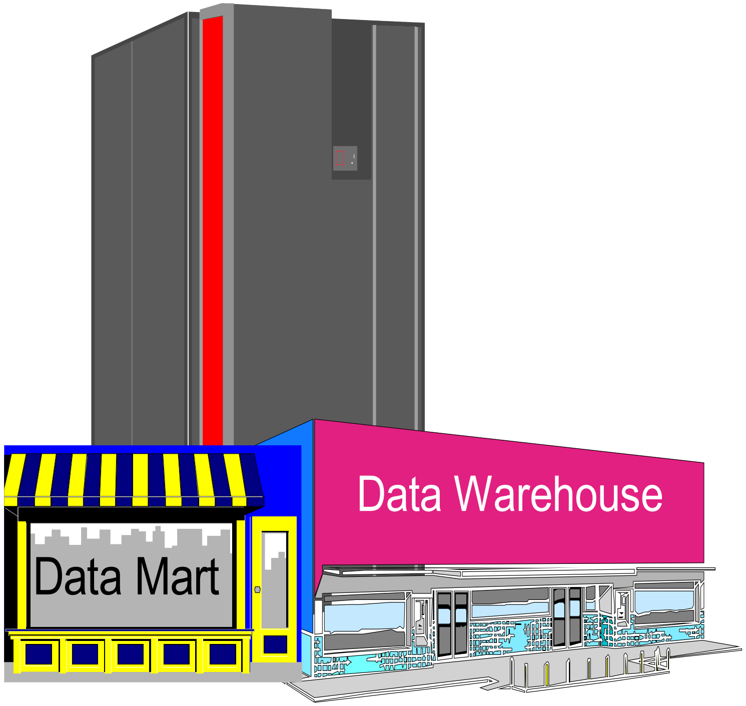
Part1: Look on the short videos about “Creation of simple DataStage flow”. Rebuild these mappings in your own environment.



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

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

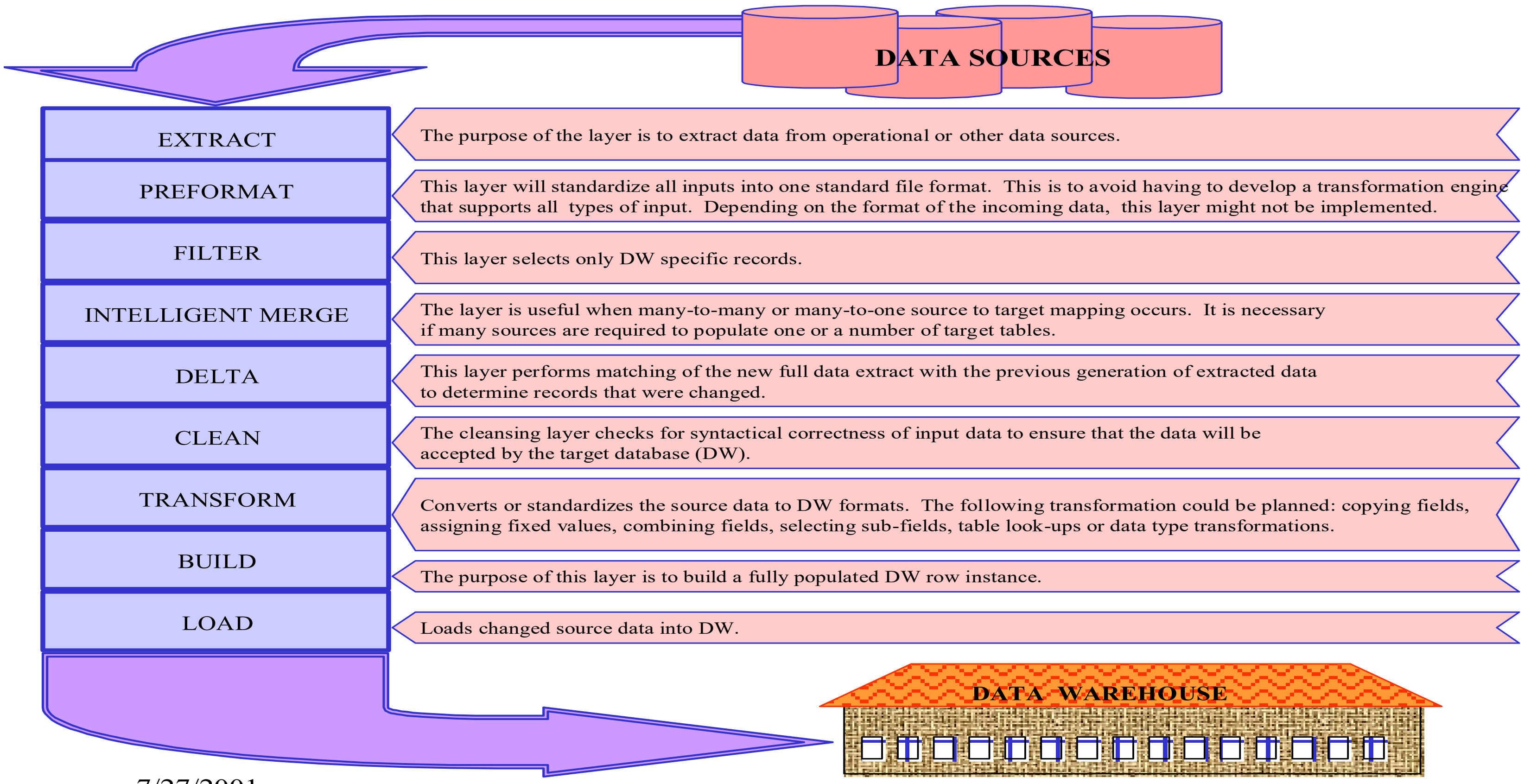
DW07 - ETL Techniques & ETL Tools



5 Highlights to ETL Techniques

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

Generic ETL Process Layers



DATA SOURCES

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

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

FILTER: This layer selects only DW specific records.

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

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

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

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

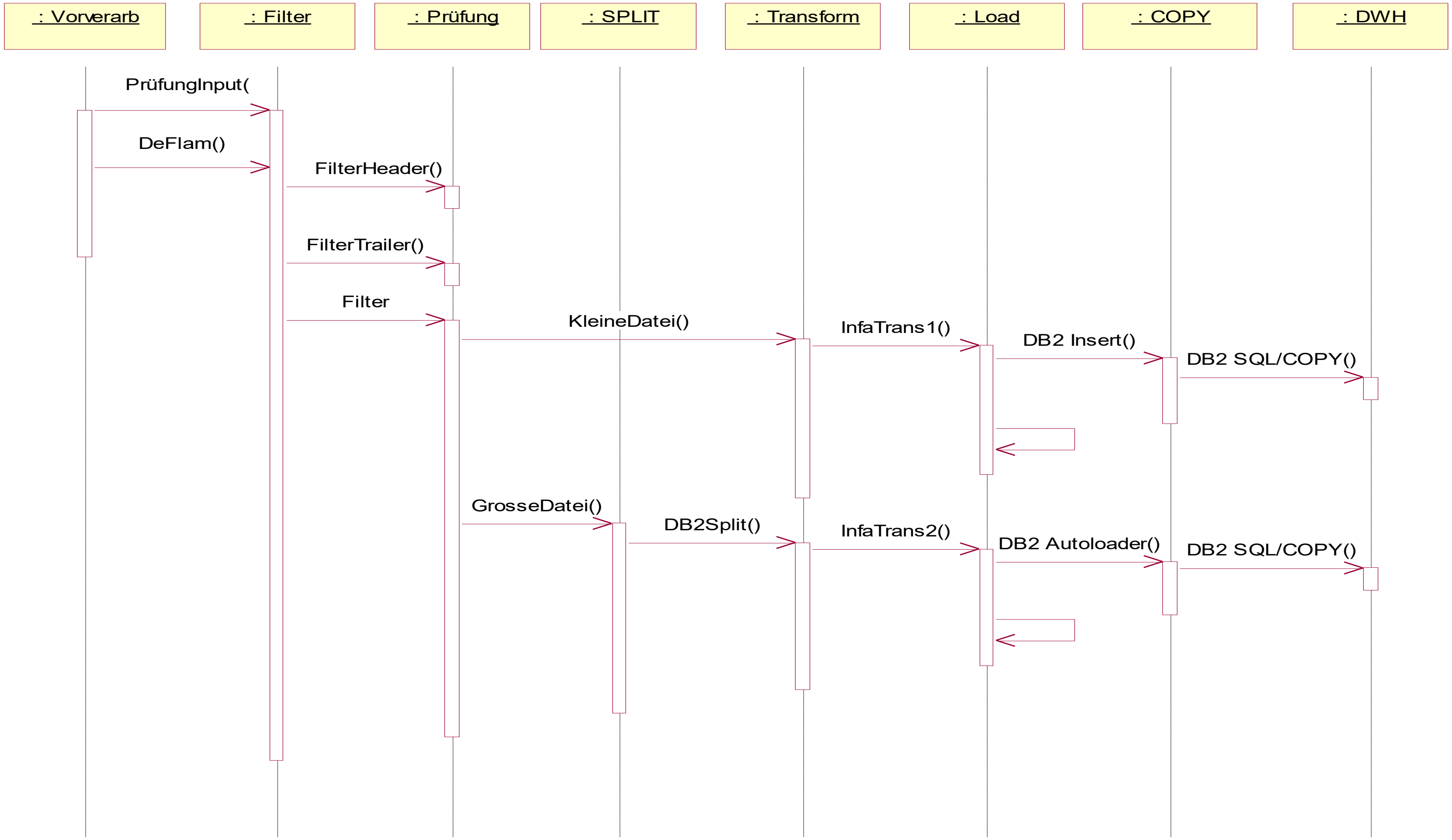
BUILD: The purpose of this layer is to build a fully populated DW row instance.

LOAD: Loads changed source data into DW.

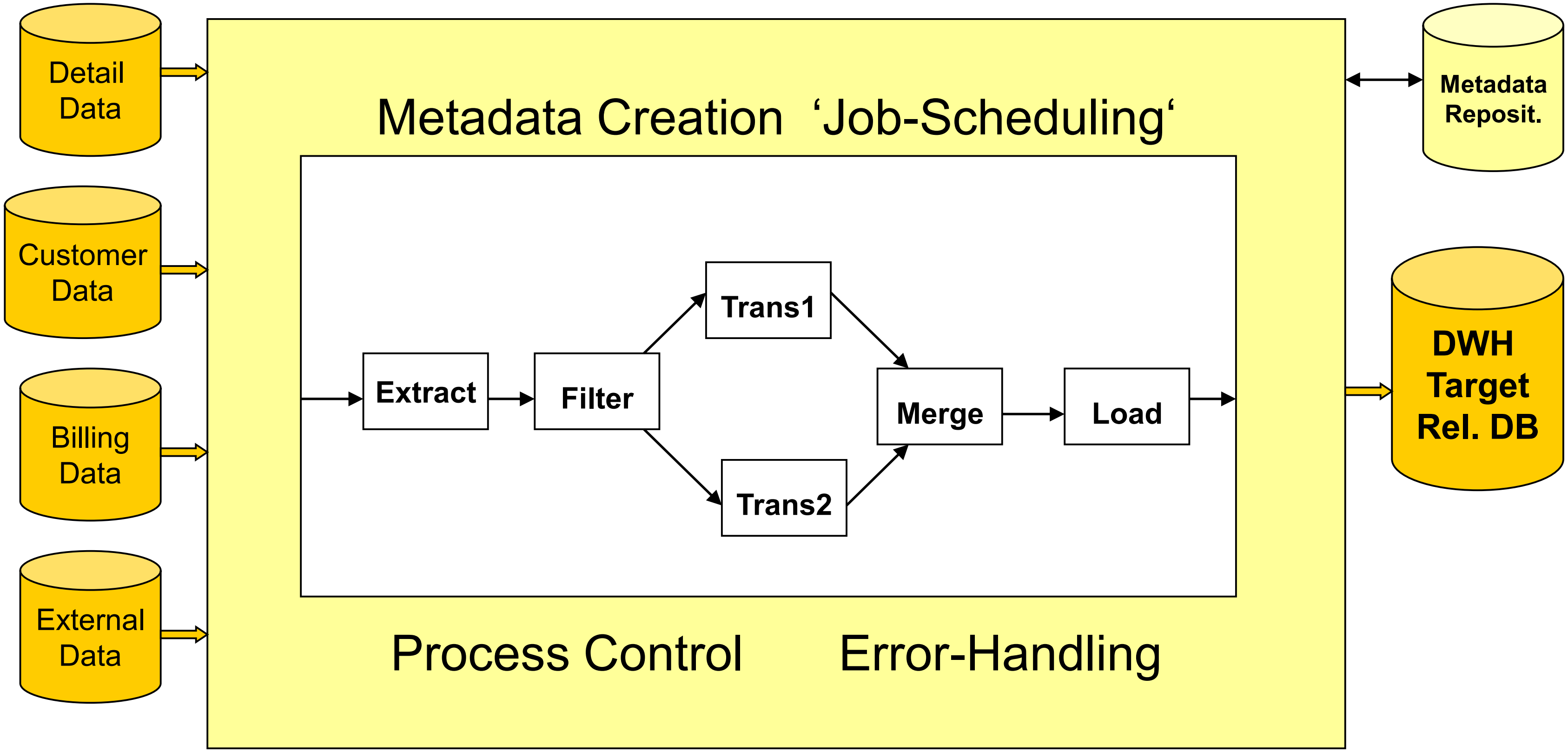
DATA WAREHOUSE

7/27/2001

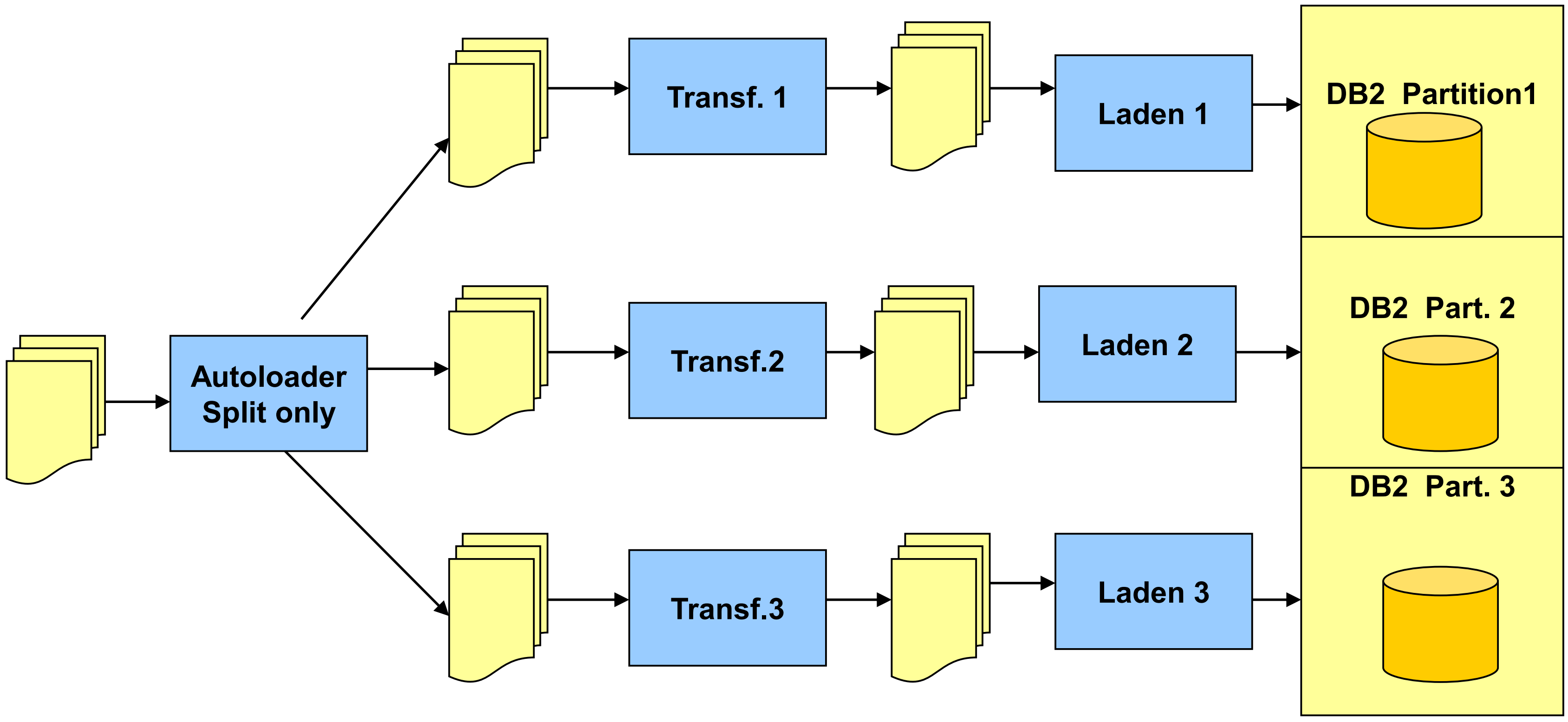
ETL Layer Concept (Example)



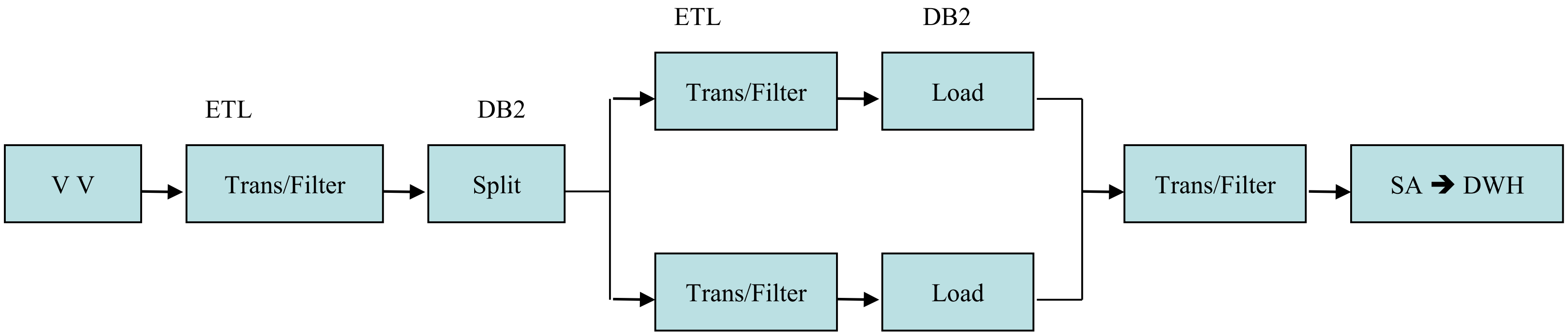
FRAMEWORK / Control of Processes



Scalability & Parallel Processing



Integration of ETL & Database (Variante N)



Parallel Transformation, dependent from DB2 partitions (db2split)

Performance: dependent from ETL & DB2 Load

Piping versus temp. Flat Files

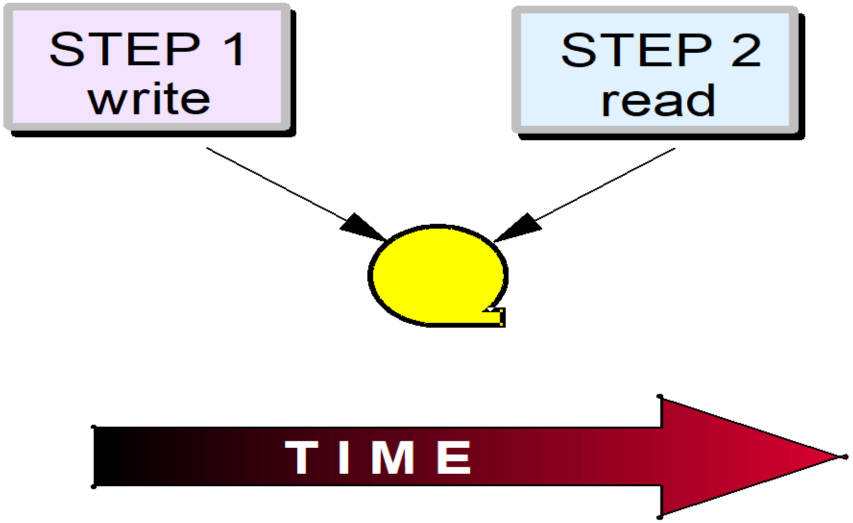
ETL calls DB2 Autoloader (with Split Only)

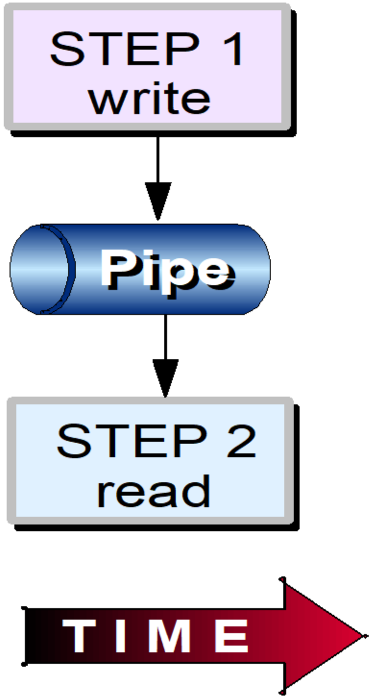
Special ETL Techniques

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

ETL Technique – ‘Piping‘

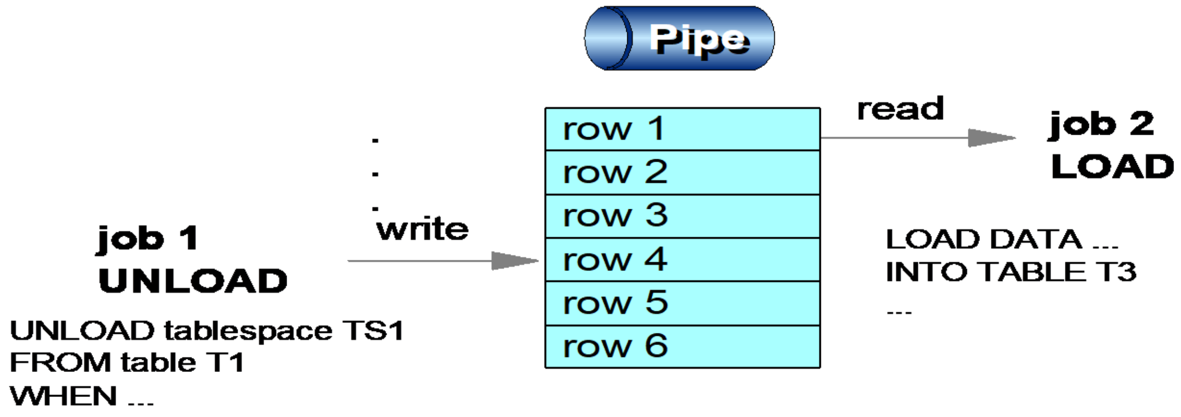
- Manage workload, optimize data flow between parallel tasks
- Reduce I/Os





ETL Technique – ‘Piping‘ Example

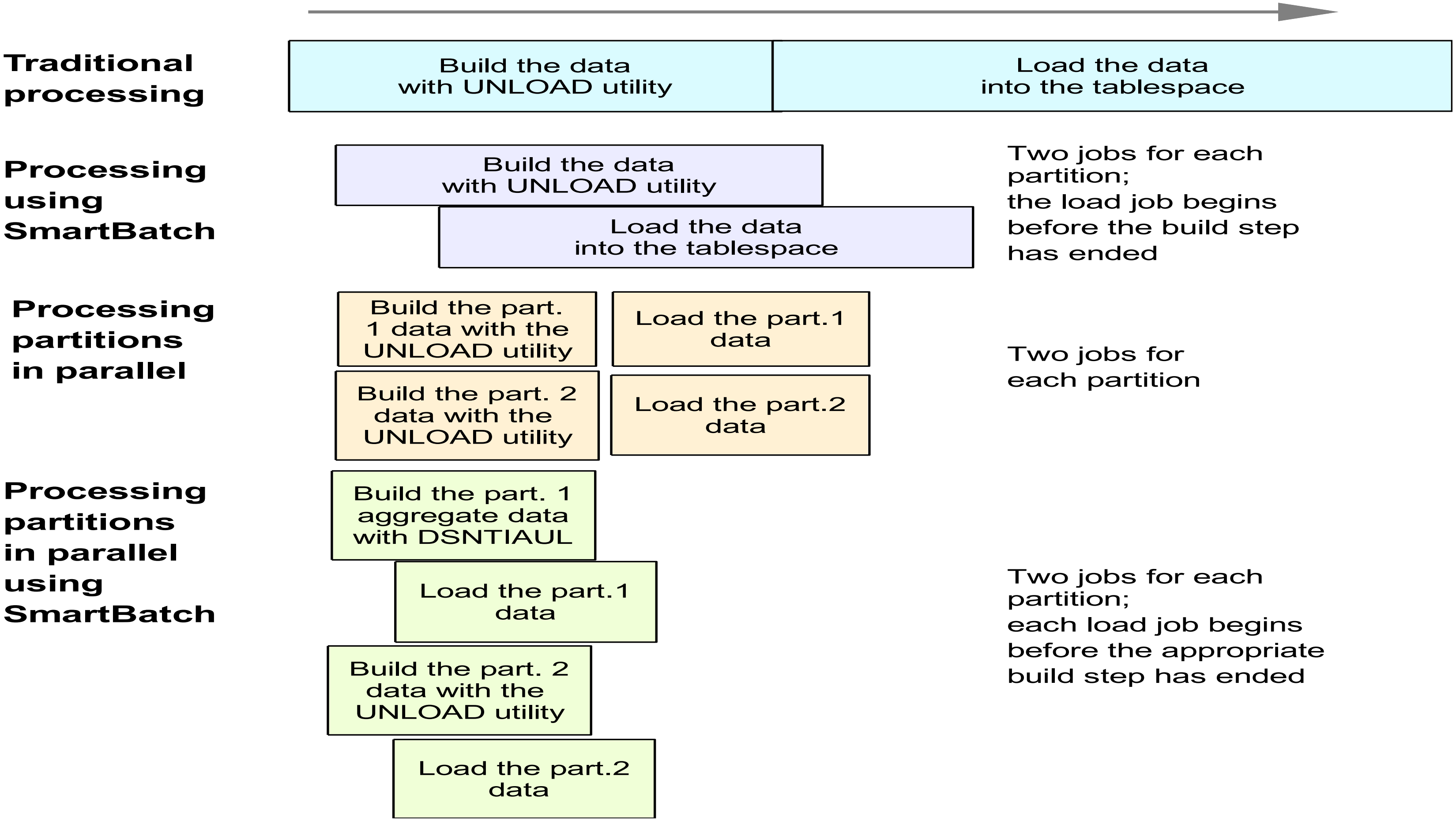
- UNLOAD
  - Provides fast data unload from DB2 table or image copy data set
  - Samples rows with selection conditions
  - Selects, order and formats fields
  - Creates a sequential output that can be used by LOAD
- LOAD
  - With SmartBatch, the LOAD job can begin processing the data in the pipe before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



UNLOAD tablespace TS1
FROM table T1
WHEN ...

LOAD DATA ...
INTO TABLE T3

ETL Technique – Compare Runtime



Traditional processing

Build the data with UNLOAD utility

Load the data into the tablespace

Processing using SmartBatch

Build the data with UNLOAD utility

Load the data into the tablespace

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

Processing partitions in parallel

Build the part. 1 data with the UNLOAD utility

Load the part.1 data

Build the part. 2 data with the UNLOAD utility

Load the part.2 data

Two jobs for each partition

Processing partitions in parallel using SmartBatch

Build the part. 1 aggregate data with DSNTIAUL

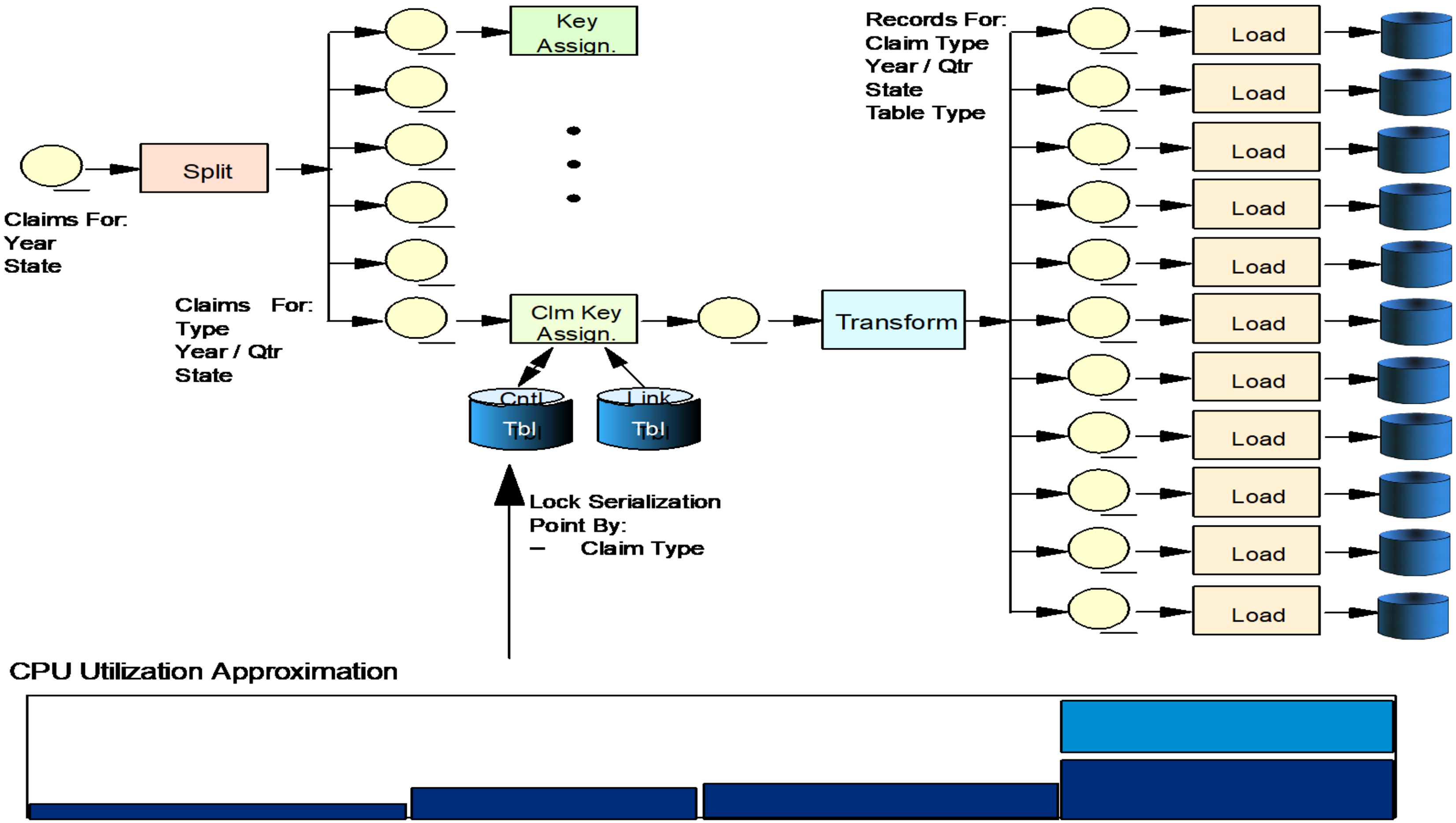
Load the part.1 data

Build the part. 2 data with the UNLOAD utility

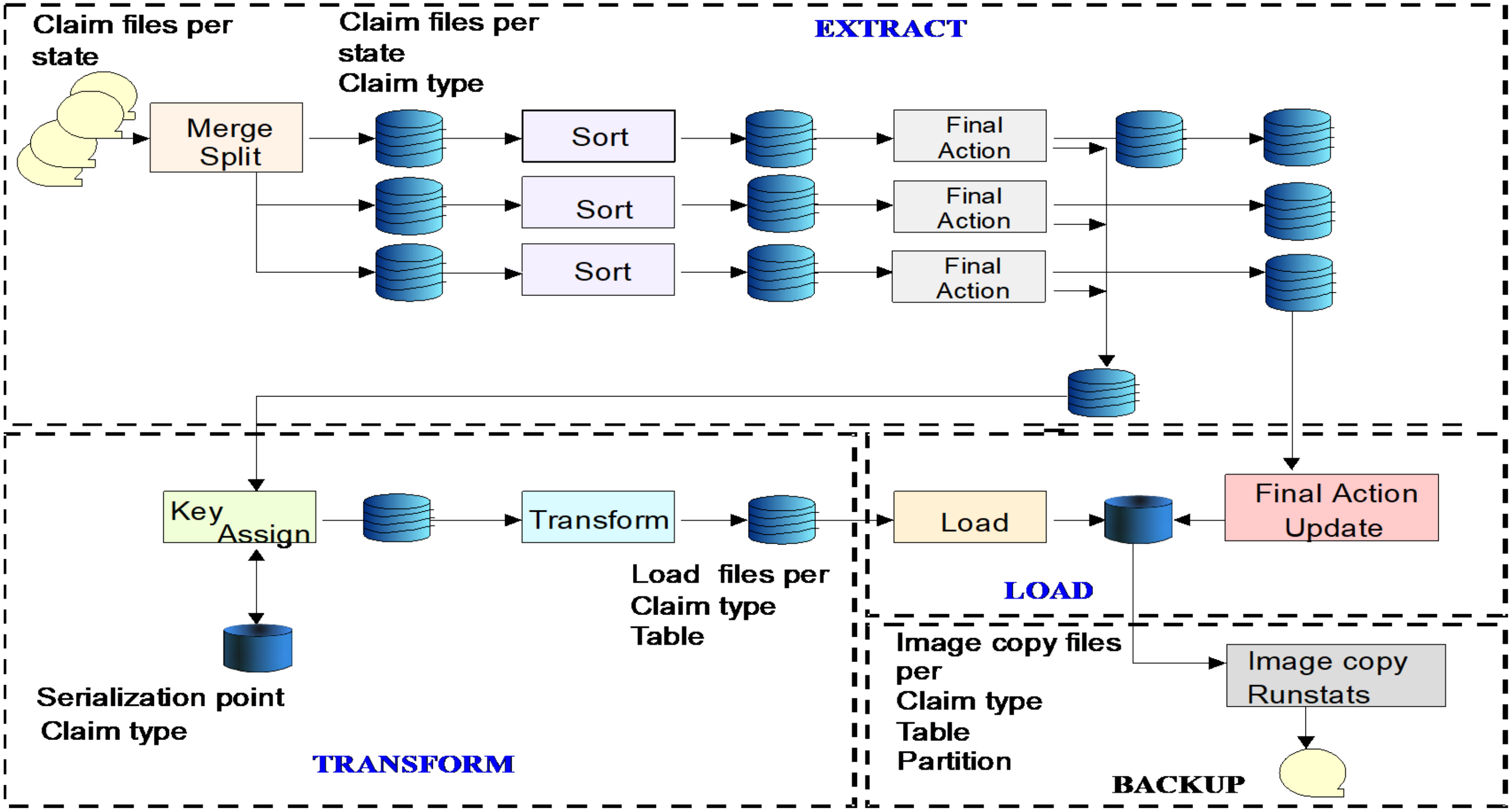
Load the part.2 data

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

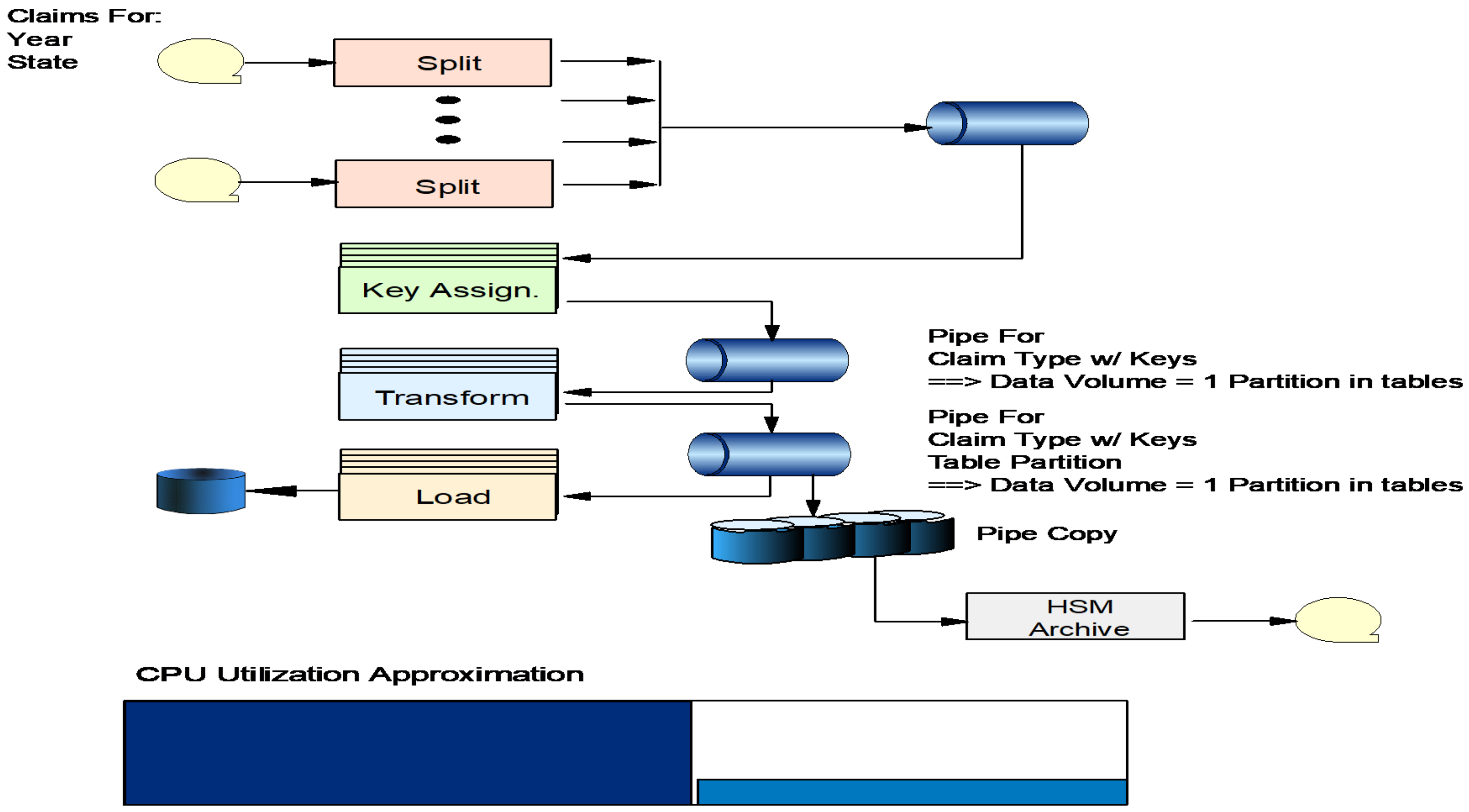
ETL Technique – ‘Sequential Design‘



ETL Technique – Sequential Design 2



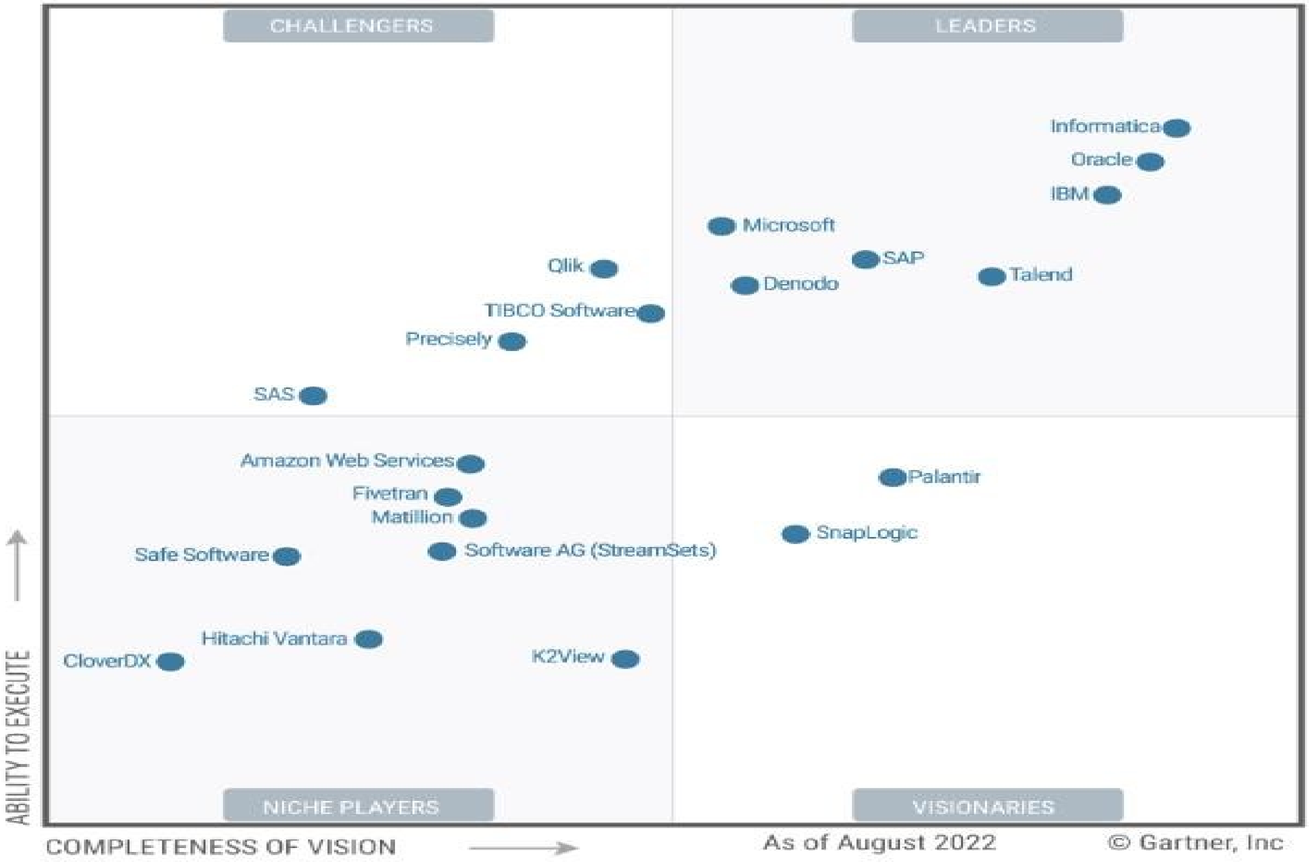
ETL Technique – ‘Piped Design‘



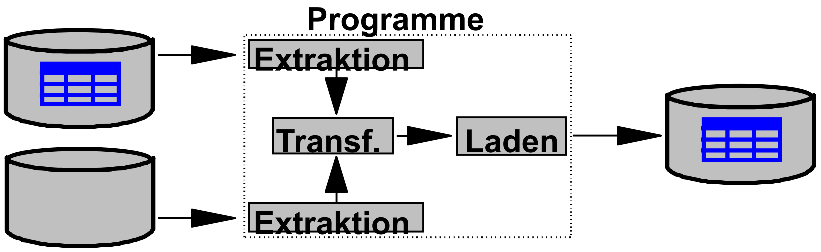
ETL Marketplace & Tools Positions

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

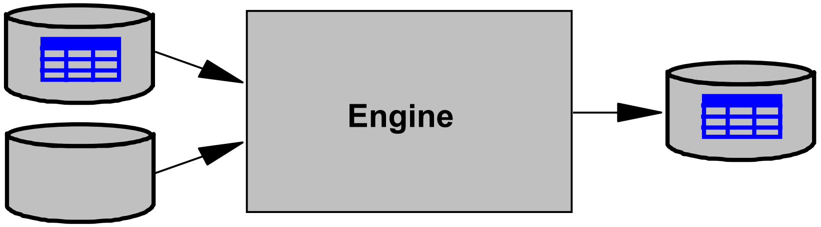
Figure 1: Magic Quadrant for Data Integration Tools



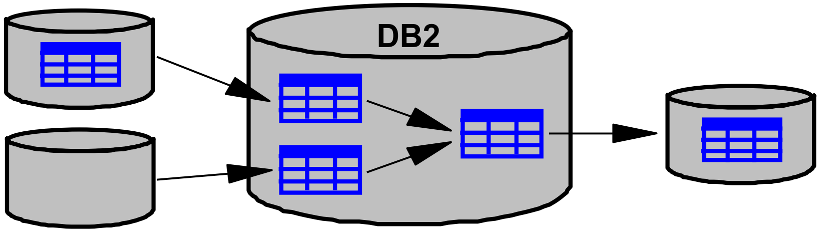
The 3 ETL Tool Architectures



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

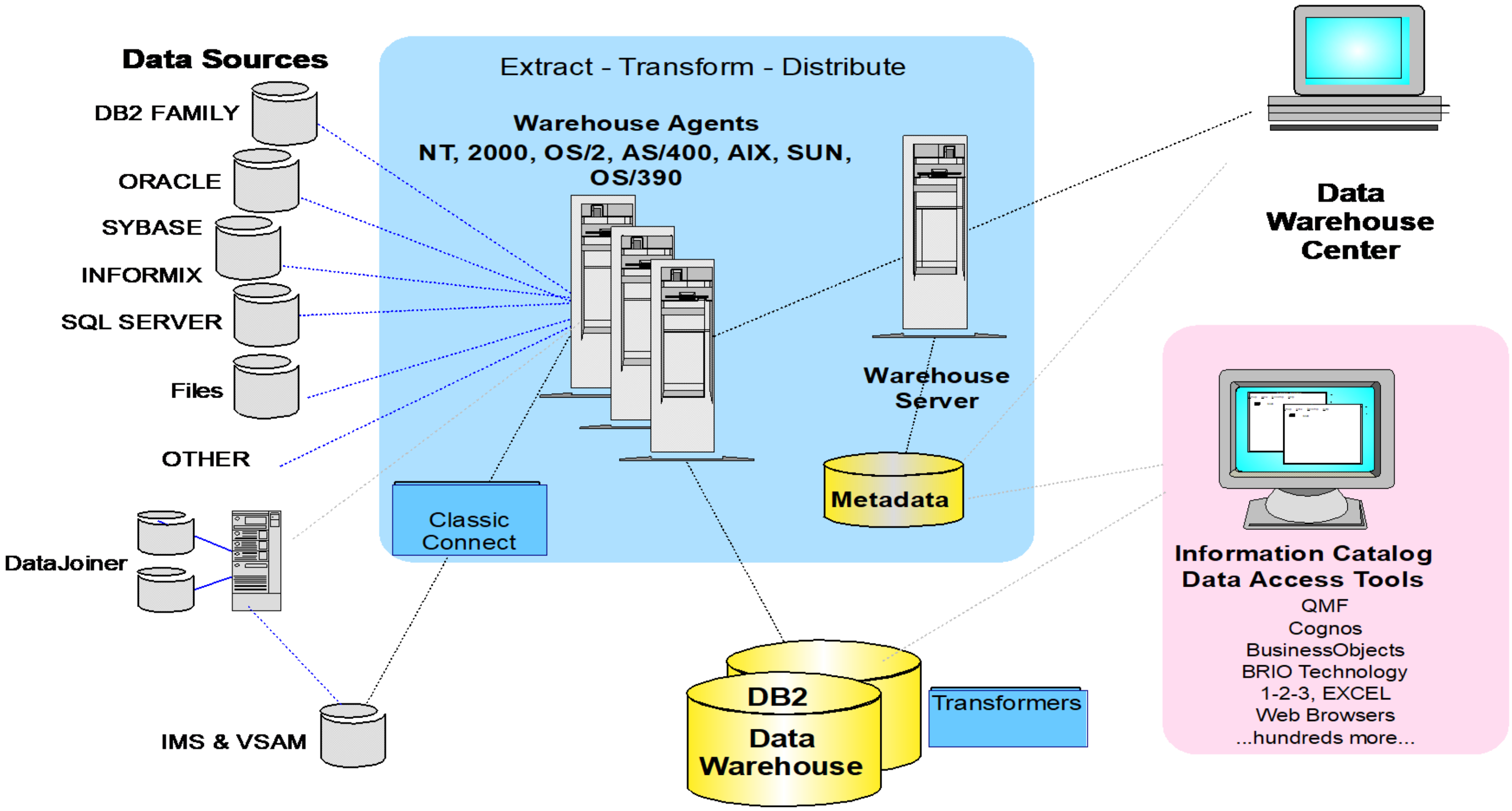


- ETL Engine
  - Transformation on UNIX / NT System
  - Central ETL Management
- f. ex. Informatica, IIS-DataStage

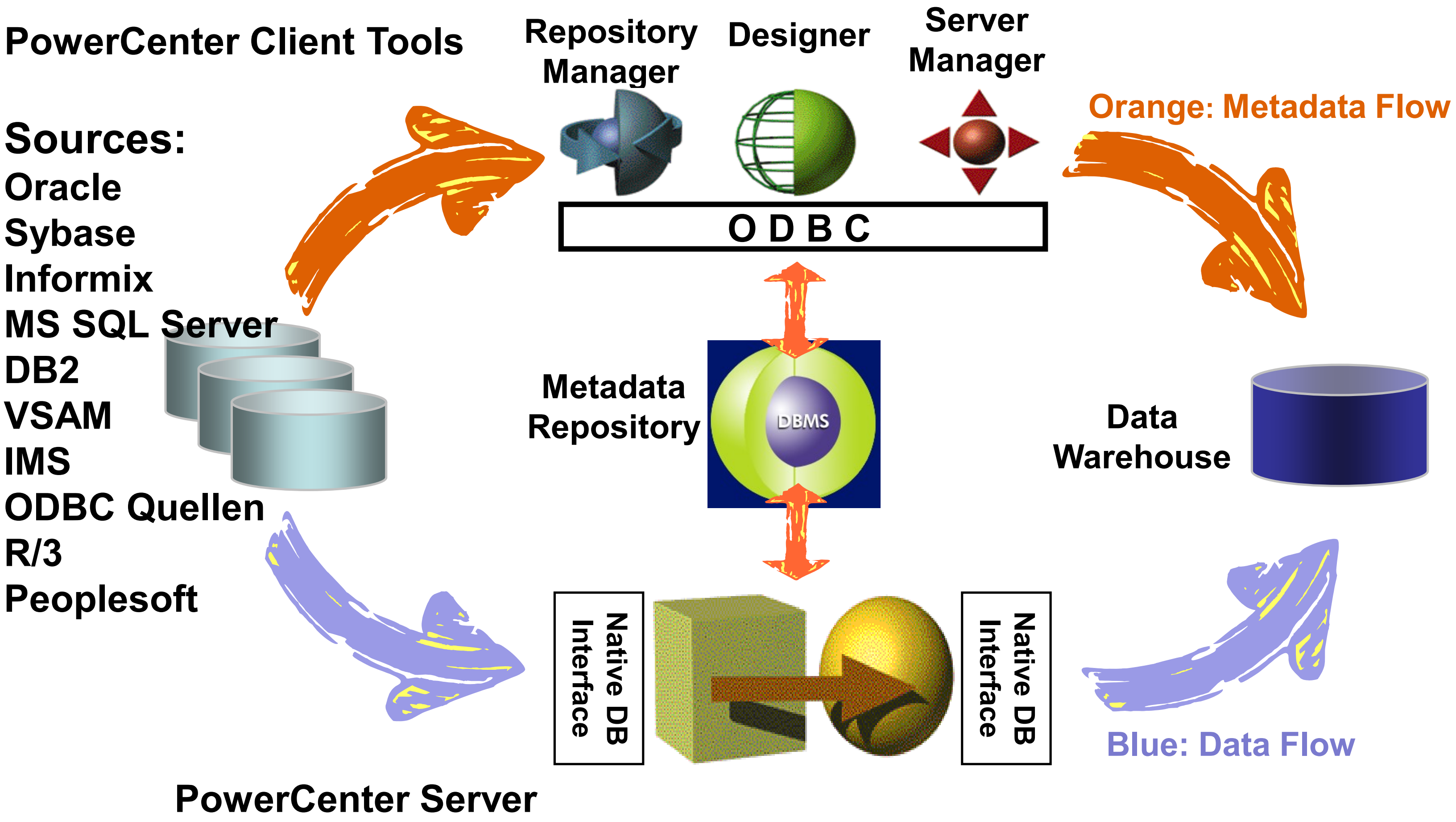


- ETL with Database Utilities
  - SQL, Stored Procedures, UDF's
  - Datenbase Scalability
  - DB-Transaction Security
- f.ex. DB2 Warehouse Manager
  Oracle Warehouse Builder (OWB)

ETL Tool –DB2 Warehouse Manager



ETL Tool – Informatica PowerCenter



PowerCenter Client Tools

Repository Manager

Designer

Server Manager

Orange: Metadata Flow

O D B C

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

Metadata Repository

Data Warehouse

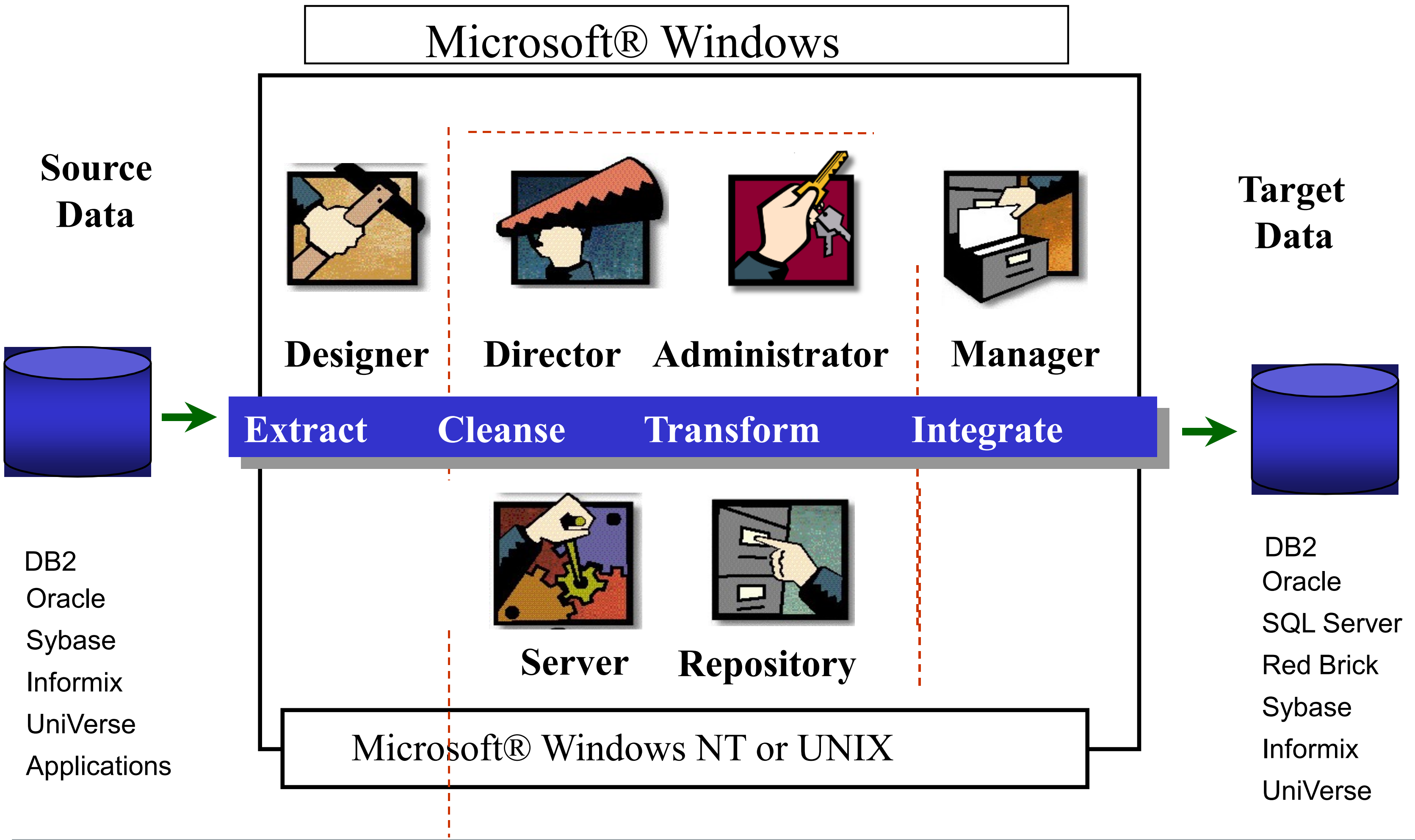
Native DB Interface

Native DB Interface

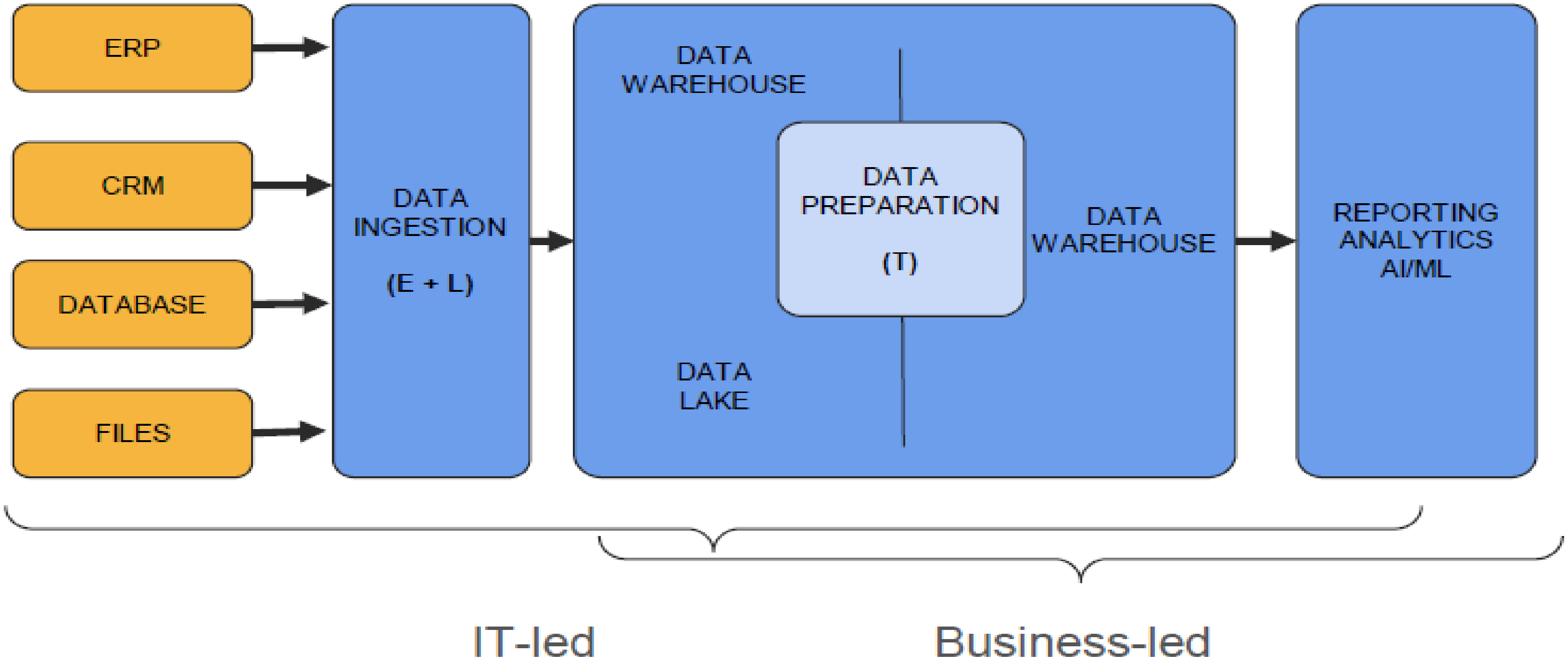
Blue: Data Flow

PowerCenter Server

ETL Tool – IBM IIS Datastage



Modern ELT Stack in a Cloud DWH (AWS)



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

Exercise 1 to Lesson 7: ETL Tool Evaluation

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

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

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

Show the three tools in competition to each other

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

Exercise 2 to Lesson 7: Demo of Datastage

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

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

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

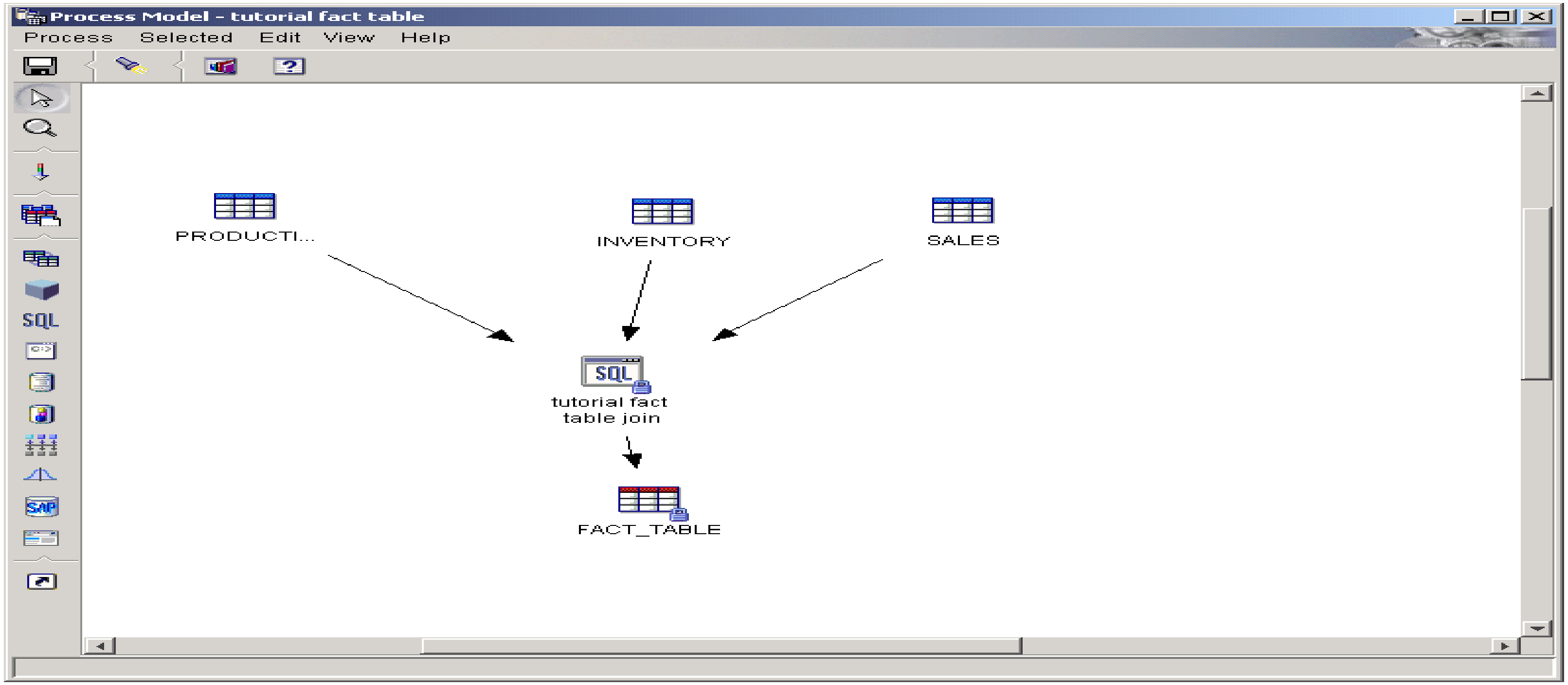
Exercise 3 to Lesson 7: Compare ETL and ELT Approach (AWS Redshift)

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

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

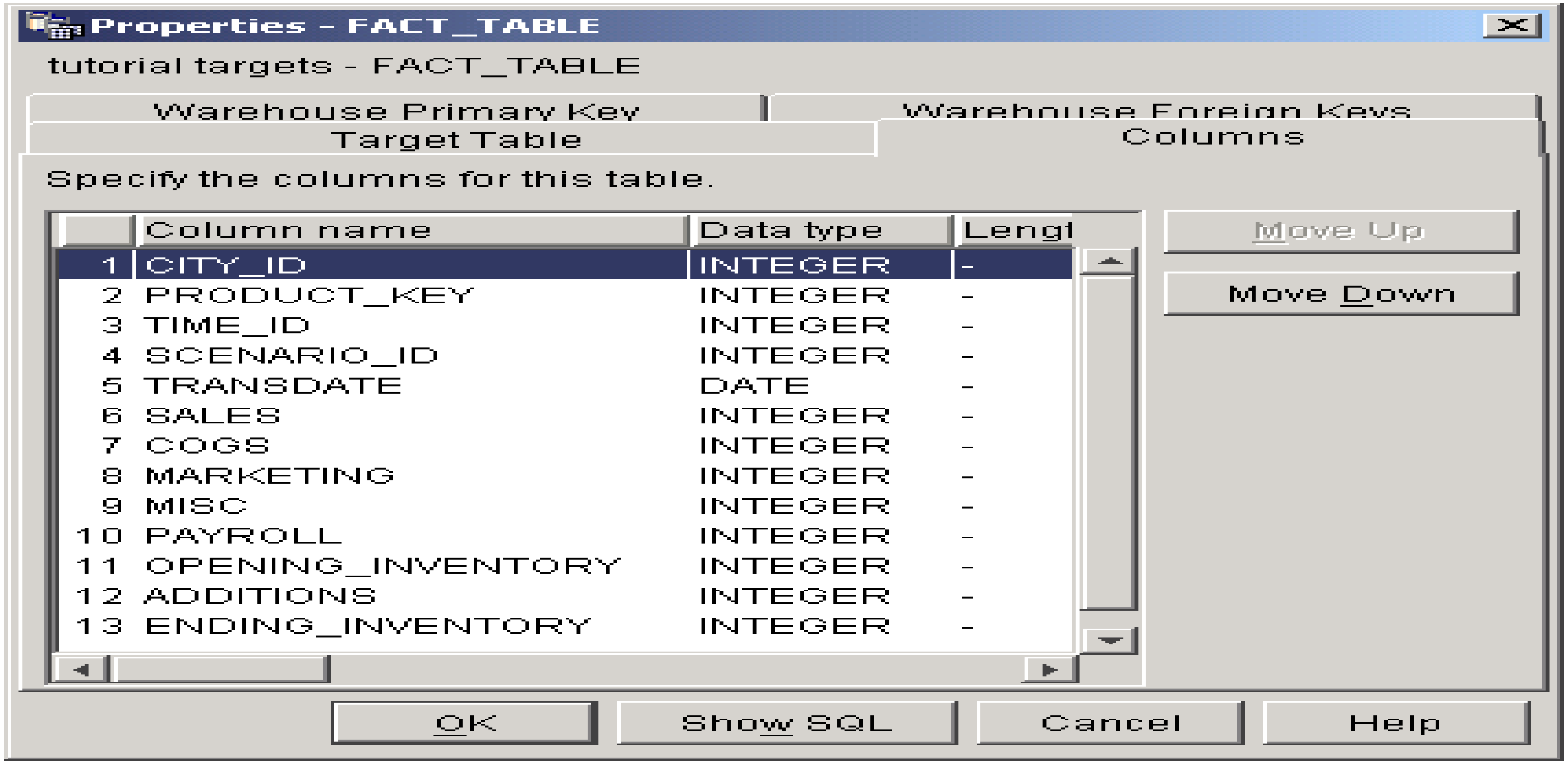
Optional: Exercise 4 to Lesson 7 – SQL Loading of a Fact Table (Part1)

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



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

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



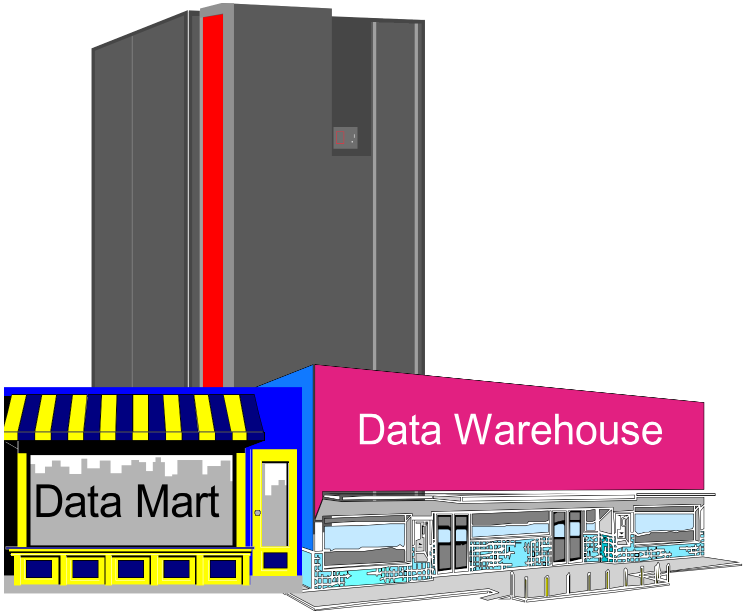
Properties - FACT_TABLE

tutorial targets - FACT_TABLE

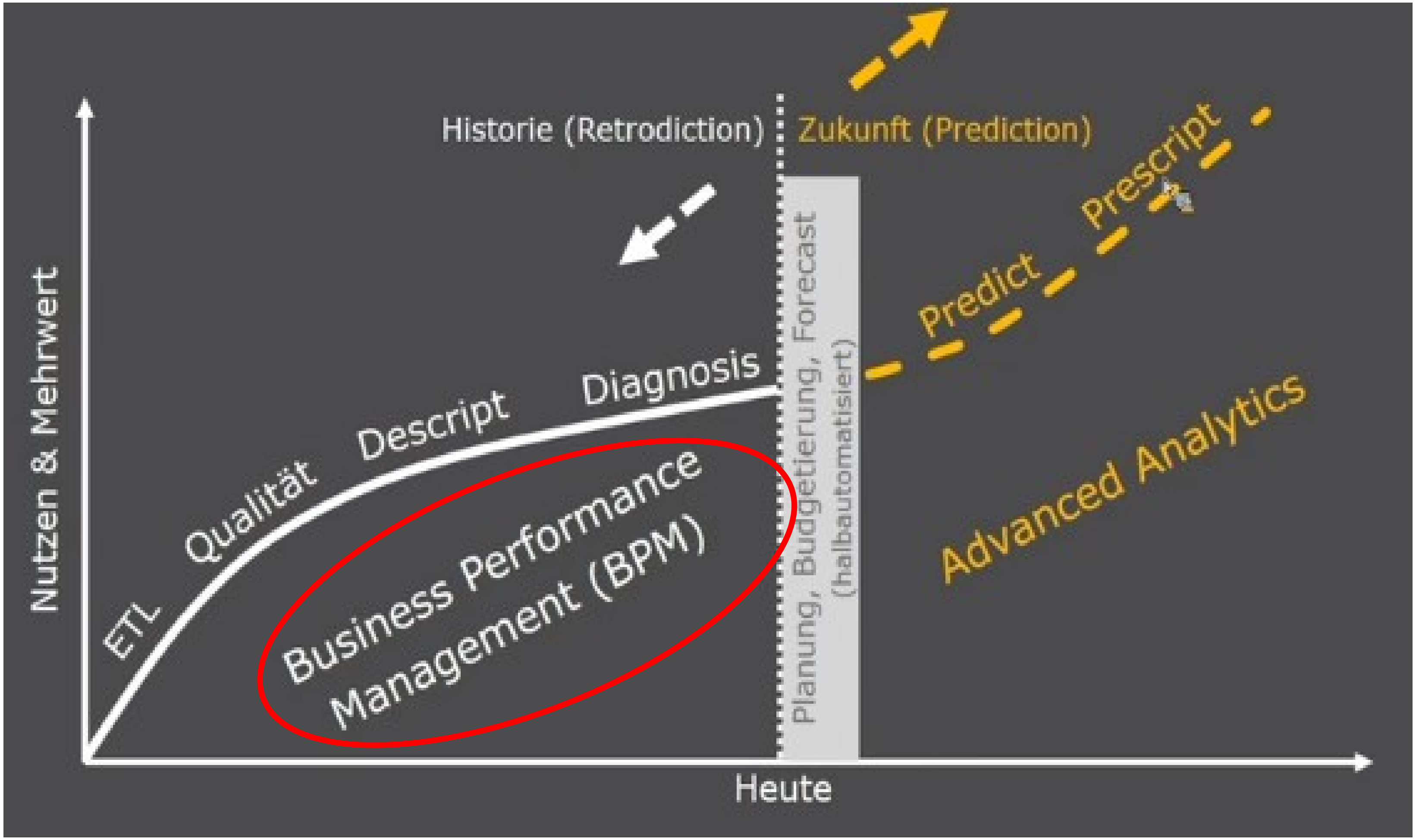
Specify the columns for this table.

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

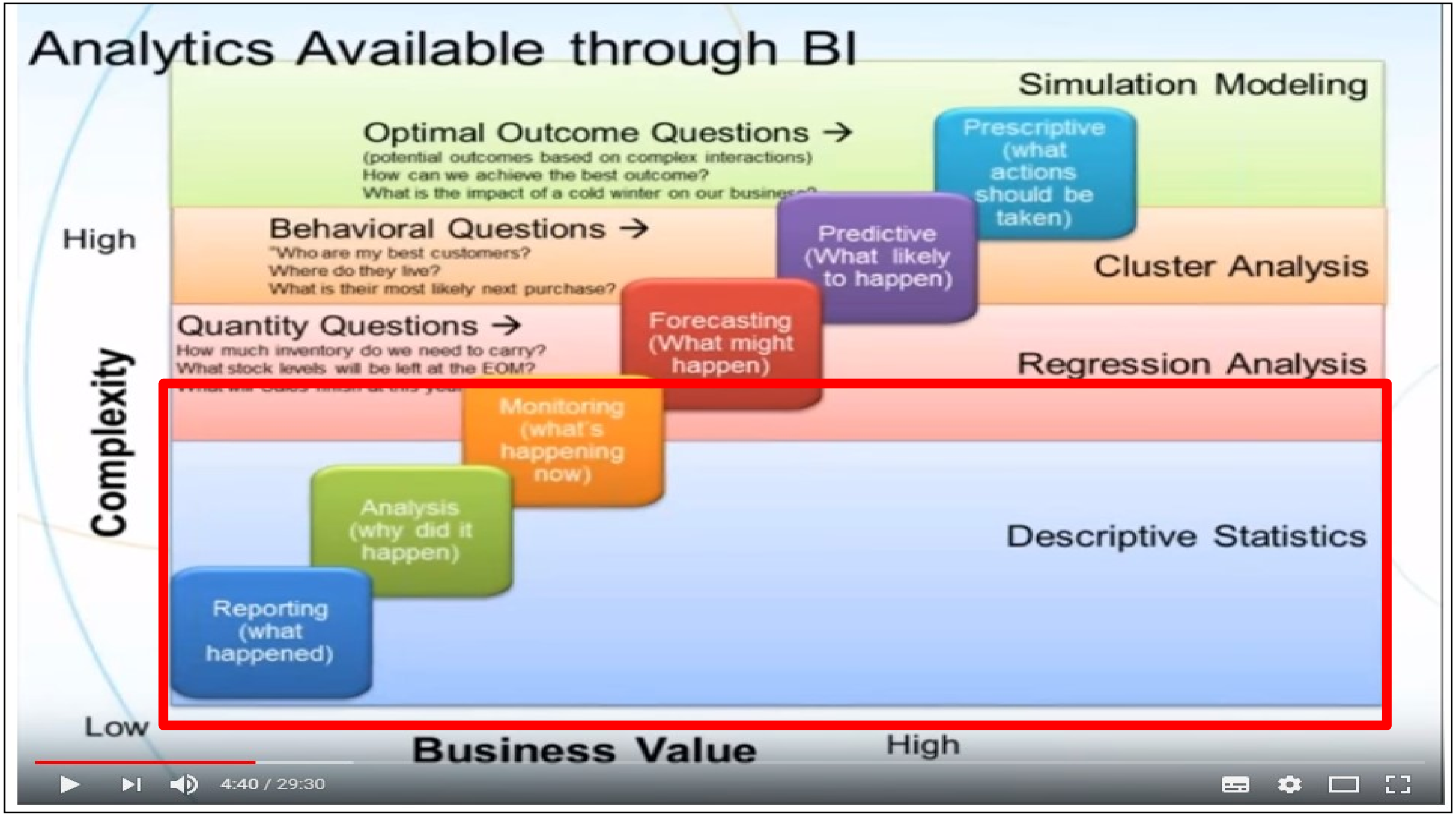
DW08 - Descriptive Analytics: Relational OLAP & Multdim. OLAP



Motivation - From Descriptive to Predictive Analytics



Descriptive Analytics (DA) – Six Levels of Analytics



Analytics Available through BI

Optimal Outcome Questions →
(potential outcomes based on complex interactions)
How can we achieve the best outcome?
What is the impact of a cold winter on our busine

Behavioral Questions →
"Who are my best customers?
Where do they live?
What is their most likely next purchase?

Quantity Questions →
How much inventory do we need to carry?
What stock levels will be left at the EOM?

Simulation Modeling
Prescriptive (what actions should be taken)
High
Predictive (What likely to happen)
Cluster Analysis
Forecasting (What might happen)
Regression Analysis
Complexity
Monitoring (what's happening now)
Analysis (why did it happen)
Descriptive Statistics
Reporting (what happened)
Low
Business Value
High
4:40 / 29:30

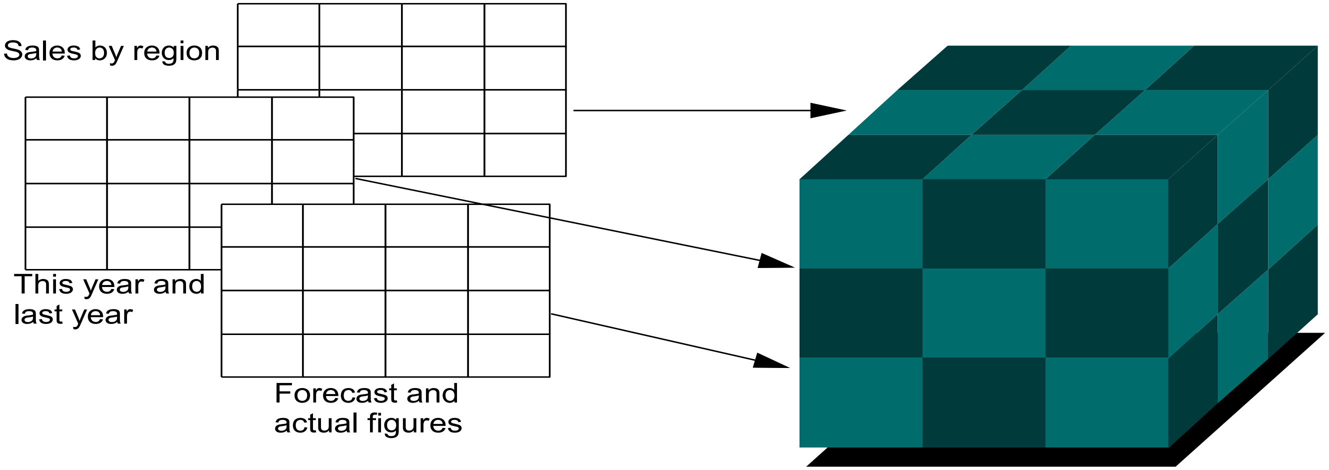
https://www.youtube.com/watch?v=oNNk9-tmsZY

Popular Descriptive Method = OLAP: What is OLAP?

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

What is Multidimensionality

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

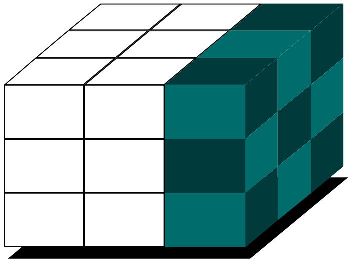


Multidimensional Database

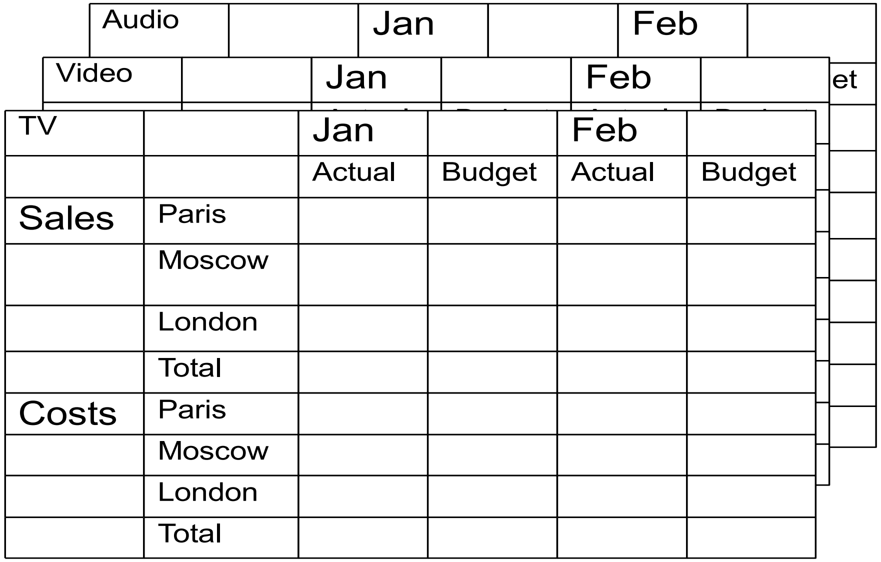
- A database specially designed to handle the organisation of data in multiple dimensions!
- Holds data cells in blocks that can be quickly built into a virtual cube depending on the query it is satisfying
- Optimised to handle large amounts of numeric data
  - Index of descriptive names held separately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

Multidimensional Views

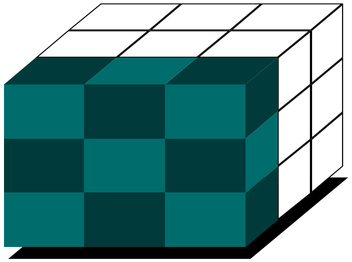
Different selections give different ways of looking at the data



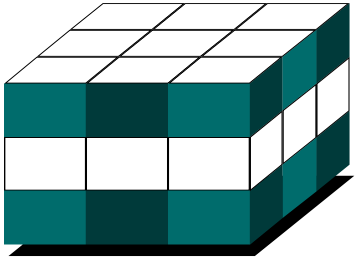
February for all products



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

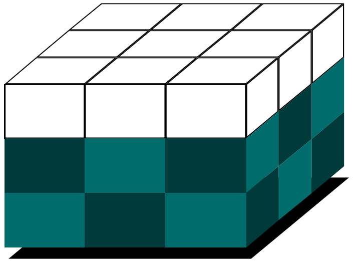


All TV information



Paris Sales and Costs

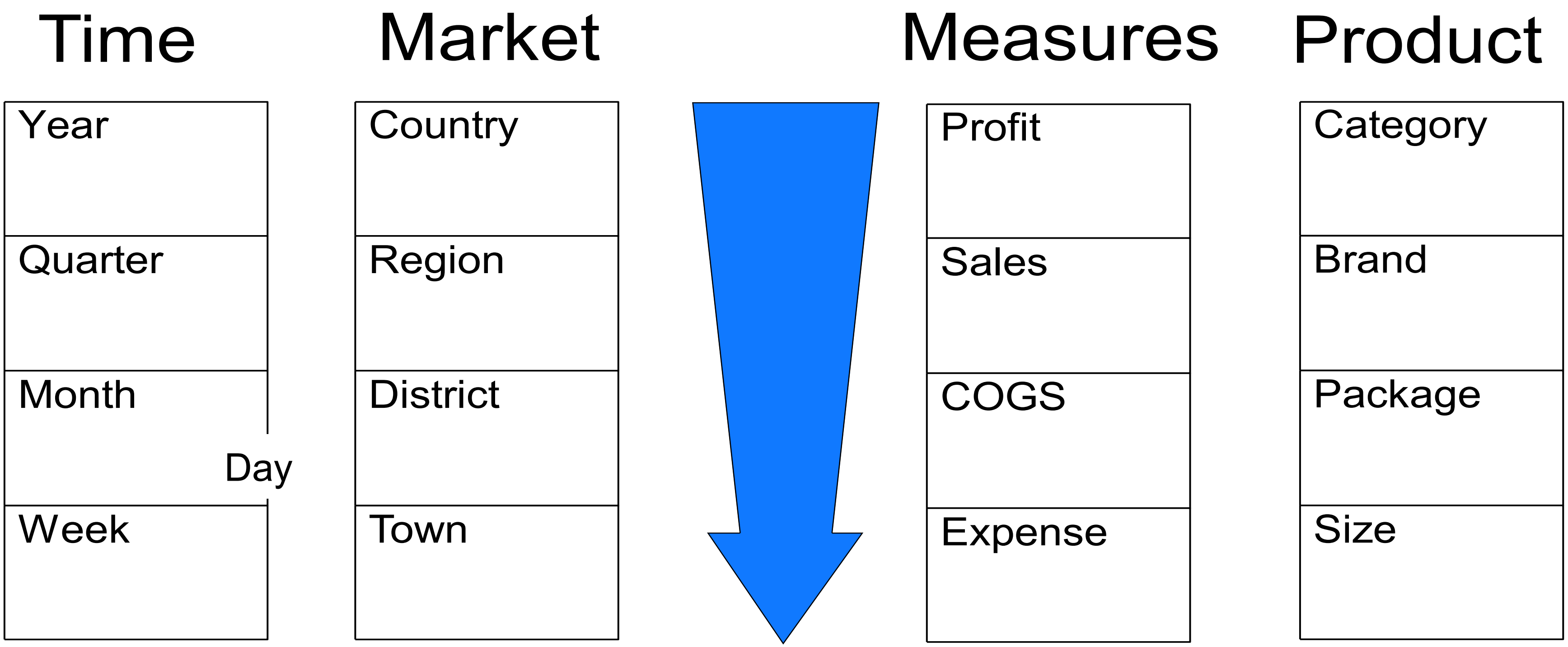
Viewing 5 dimensional database



All cost information

Drill Down

Looks at components in greater detail down same dimension

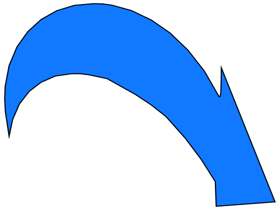


Time	Market	Measures	Product
Year	Country	Profit	Category
Quarter	Region	Sales	Brand
Month Day	District	COGS	Package
Week	Town	Expense	Size

Slice and Dice

Change row, column and page dimensions

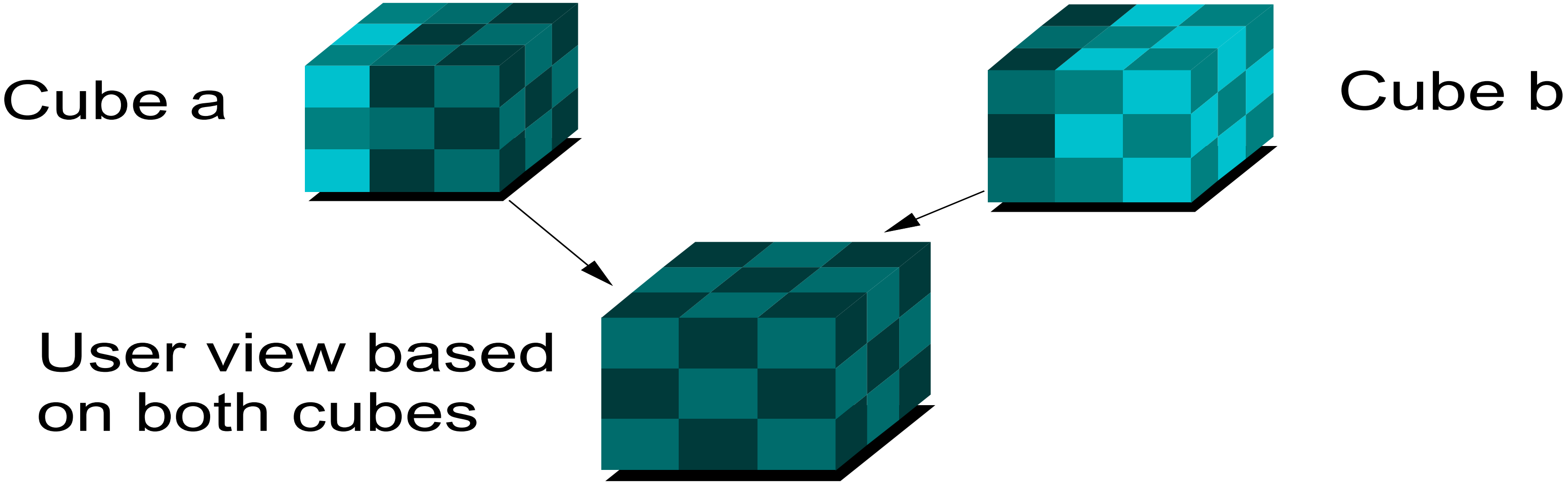
		Bud	Act	Bud	Act
1997	East				
	West				
1996	East				
	West				



		1994	1995	1996	1997
East	Food				
	Drink				
West	Food				
	Drink				

Multi-Cube Solutions

- Enhance Scalability
- Partition Applications for Parallel Load and Calculation
- Combine Similar or Dissimilar models in one user OLAP view



Cube a

Cube b

User view based on both cubes

Multidimensional vs. Relational

Multidimensional

- Optimised for query and report
- Restricted uses
- Fast, non-complex queries
- Data not dynamic - limited data update
- Database queries built by OLAP engine
- Cube must be rebuilt to refresh data and totals

Relational

- Optimised for transaction systems and query
- Many application areas
- Queries may be complex
- Easy to add/change data and structure
- Database queries written in SQL
- Data can be added and totalled interactively

MOLAP vs. ROLAP

Similarities

- Both work with numeric data, not textual
- Output results the same
- Both can provide drill down and slice & dice
- Both provide information to end users

Differences

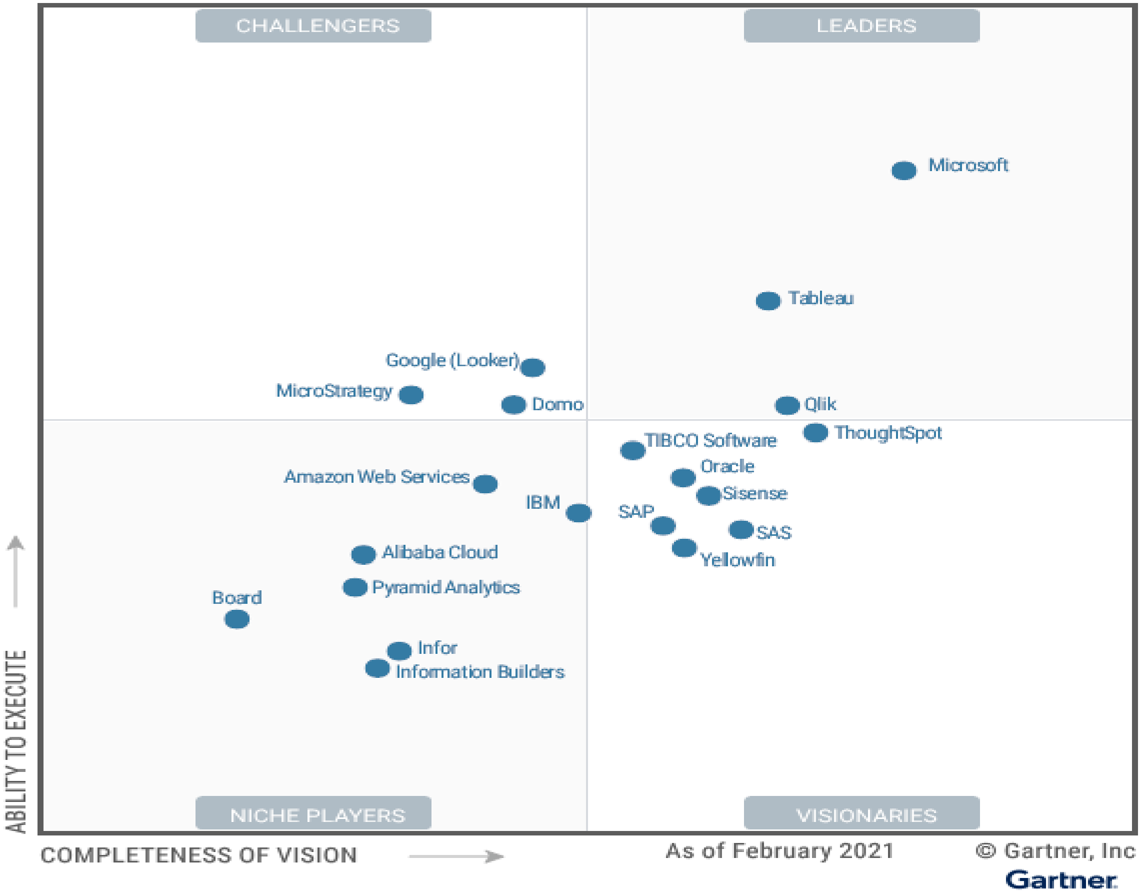
- Totals usually already calculated in MD OLAP
- MD cube must be recalculated
- ROLAP joins data tables for each query
- MD cube size limited by architecture, ROLAP size limited by database

Benefits of MOLAP

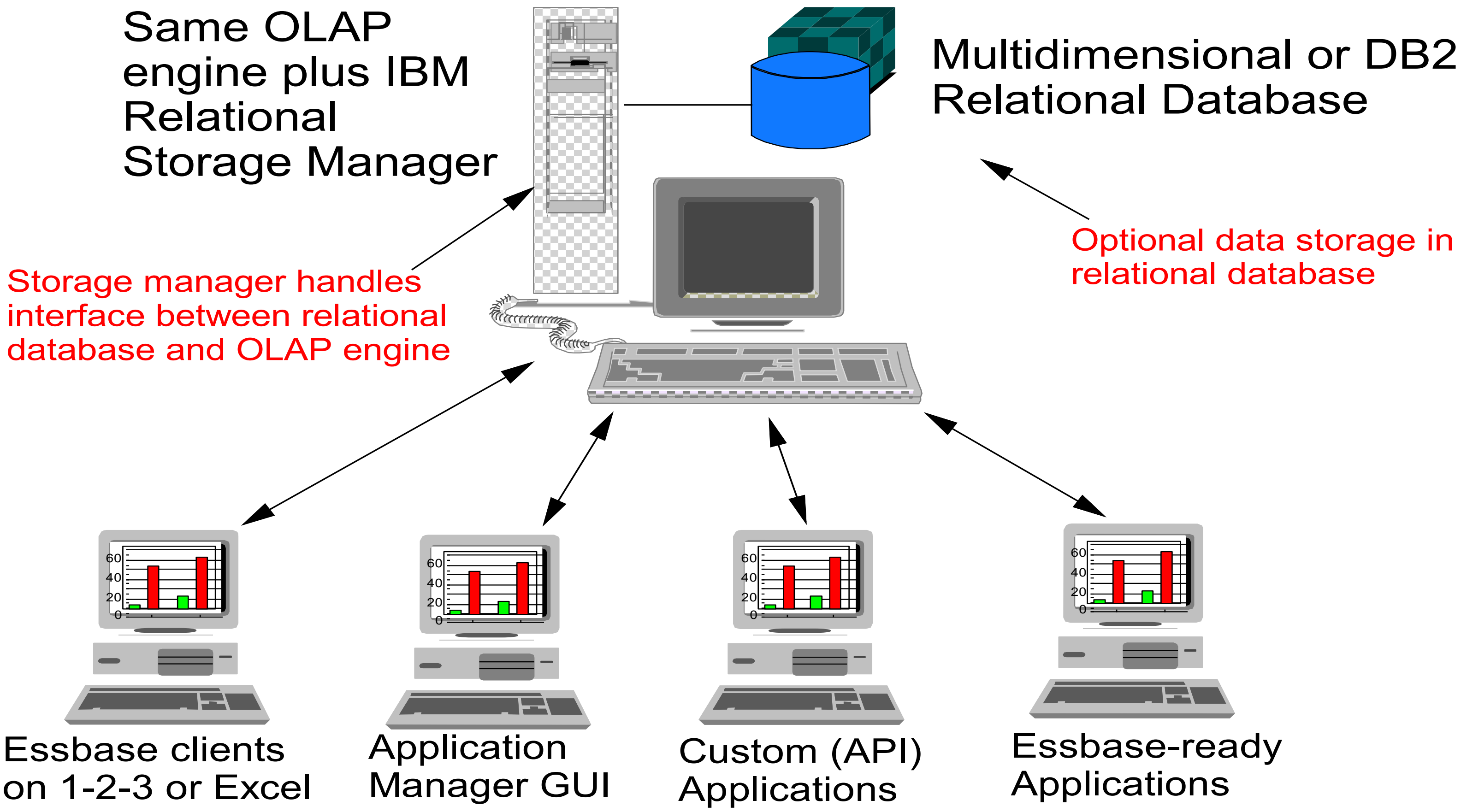
- Makes many different analyses without constructing separate queries
  - All possible queries on the multidimensional data can be created by OLAP engine
  - Fast response to changing data requests
- Quick to deploy
  - Simple to report using spreadsheet or graphical tool
  - Many end user requirements satisfied once cube is built without building individual reports
- Quick to use
  - "Speed of thought" response
  - No contention from long-running queries
- Common Informational Database
  - Same information on server available to many users
  - Doesn't impact transaction systems

OLAP Marketplace & Tool Position

Figure 1: Magic Quadrant for Analytics and Business Intelligence Platforms



Example: IBM DB2 OLAP Server - Components



Same OLAP engine plus IBM Relational Storage Manager

Multidimensional or DB2 Relational Database

Storage manager handles interface between relational database and OLAP engine

Optional data storage in relational database

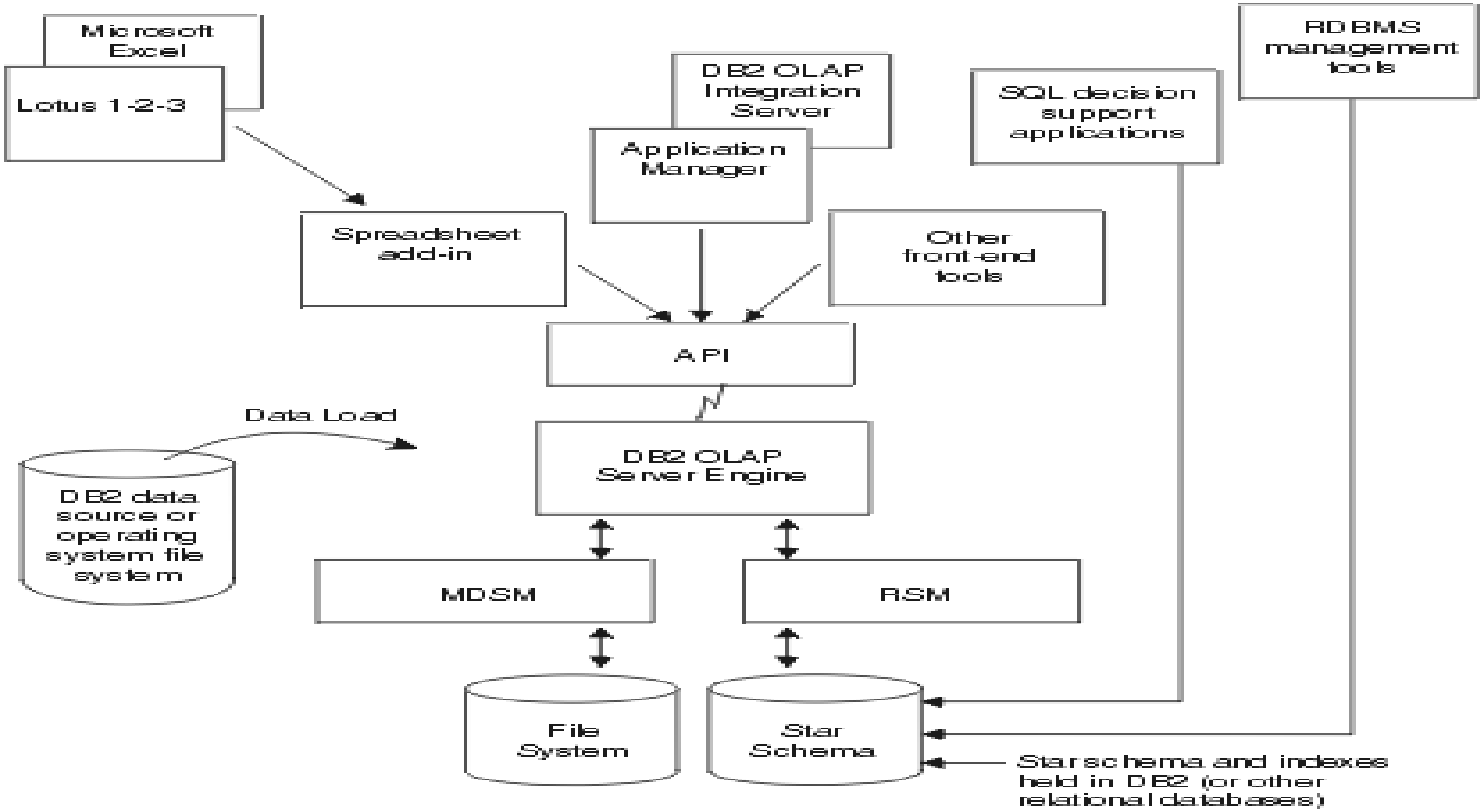
Essbase clients on 1-2-3 or Excel

Application Manager GUI

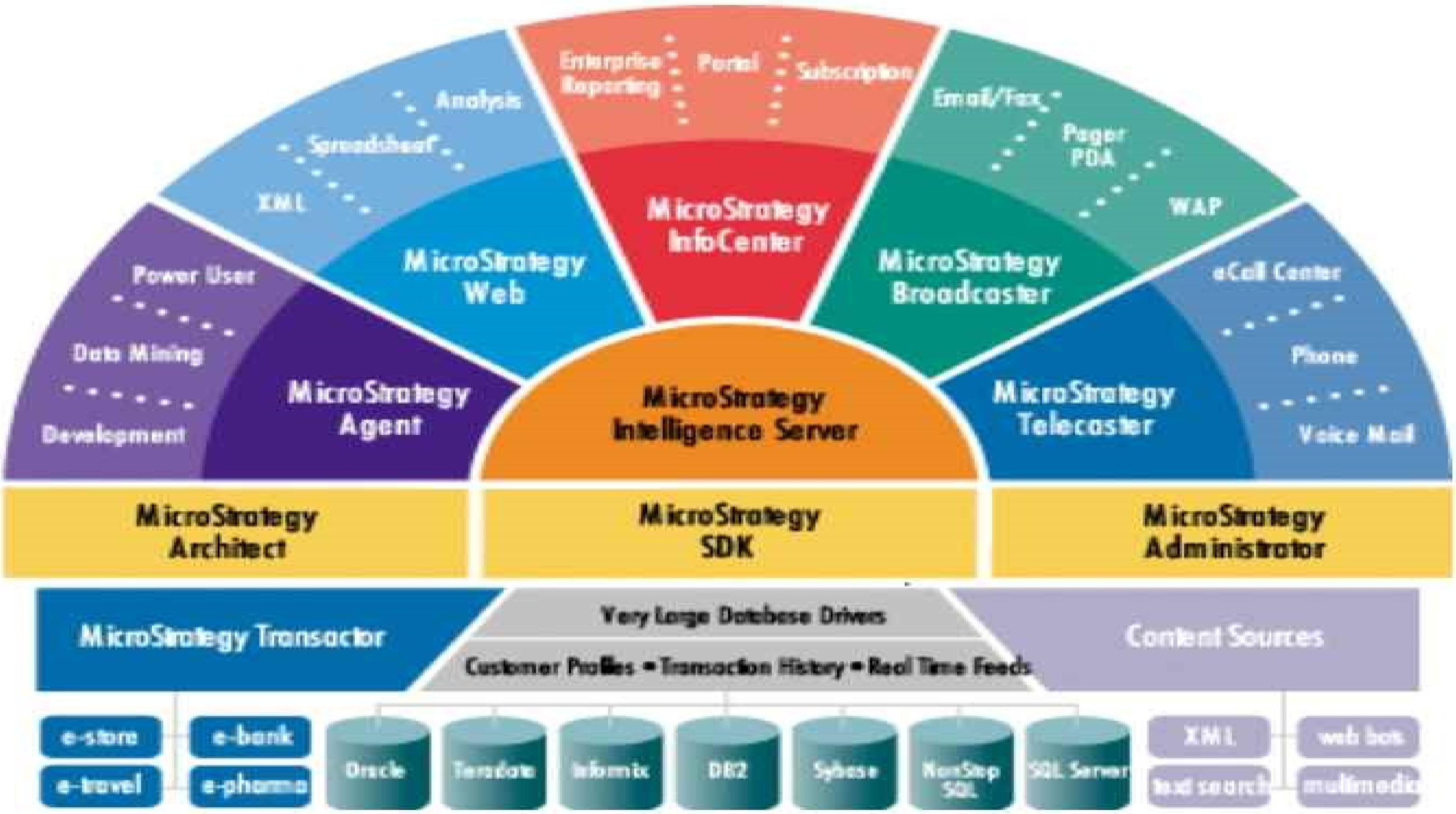
Custom (API) Applications

Essbase-ready Applications

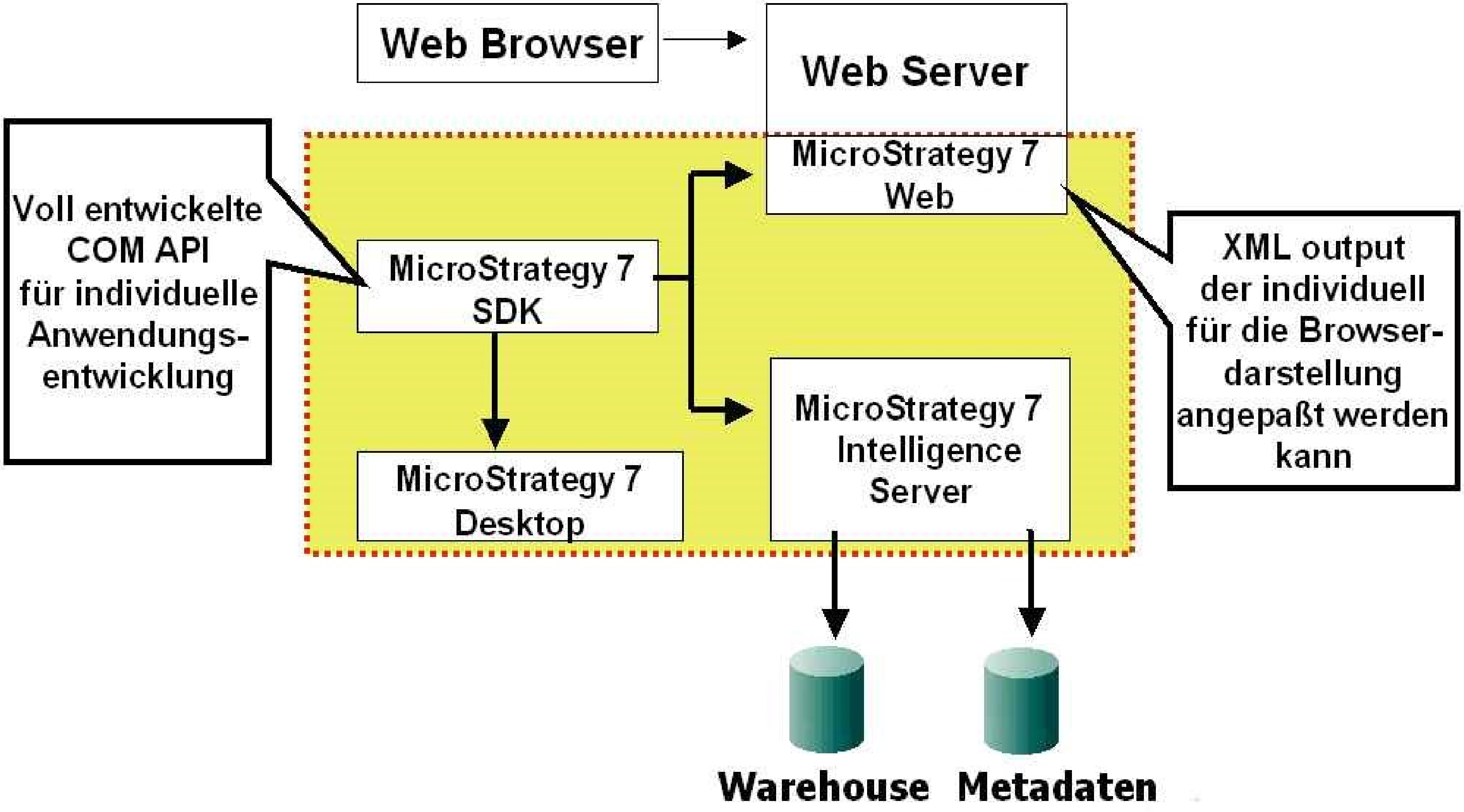
IBM DB2 OLAP Server - Architecture


Microsoft Excel
Lotus 1-2-3
DB2 OLAP Integration Server
Application Manager
SQL decision support applications
RDBMS management tools
Spreadsheet add-in
Other front-end tools
API
Data Load
DB2 OLAP Server Engine
DB2 data source or operating system file system
MDSM
RSM
File System
Star Schema
Star schema and indexes held in DB2 (or other relational databases)

ROLAP / MicroStrategy: Components Overview



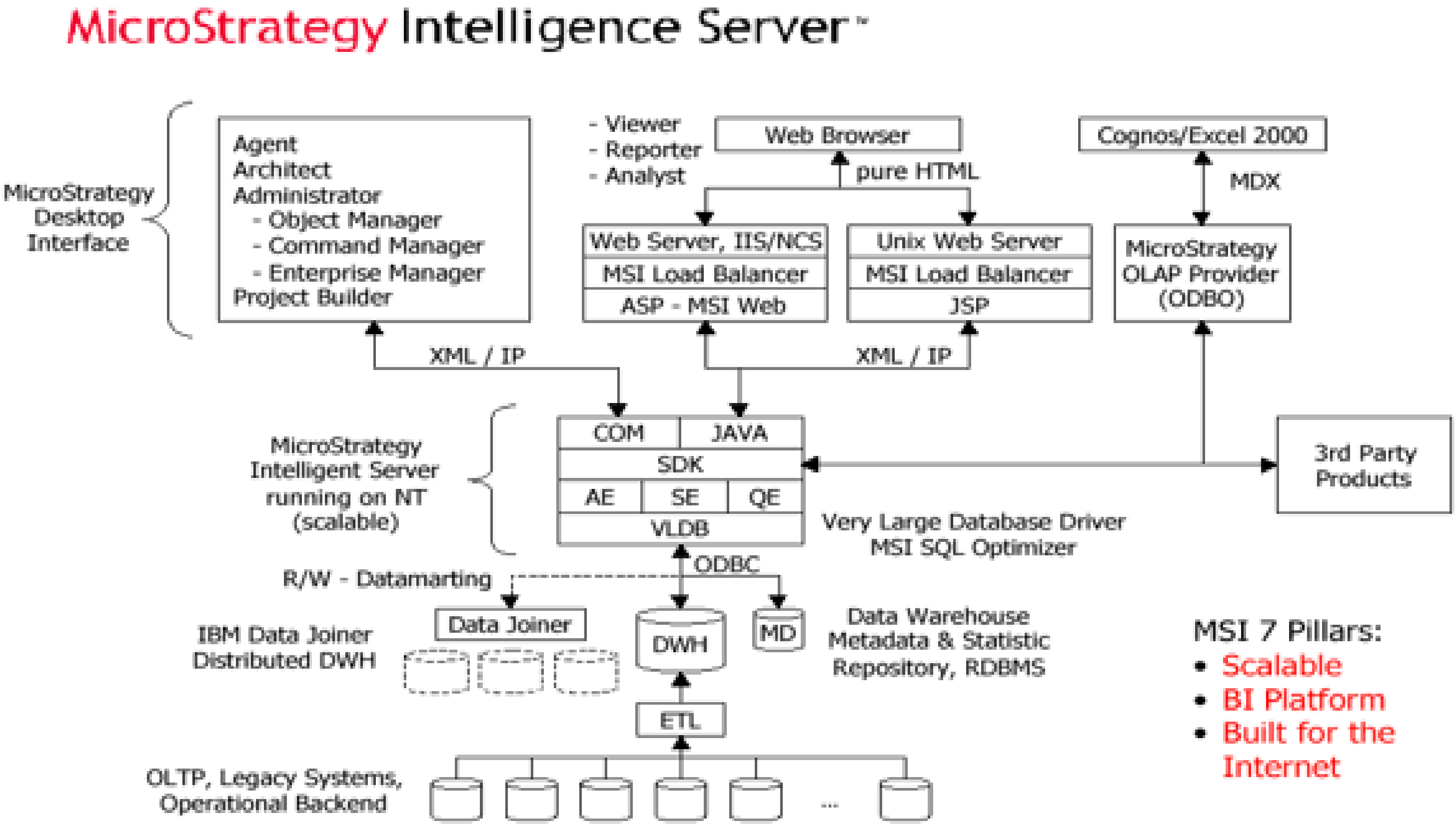
ROLAP Example - MicroStrategy: Analytical Model



Voll entwickelte COM API für individuelle Anwendungsentwicklung

XML output der individuell für die Browserdarstellung angepaßt werden kann

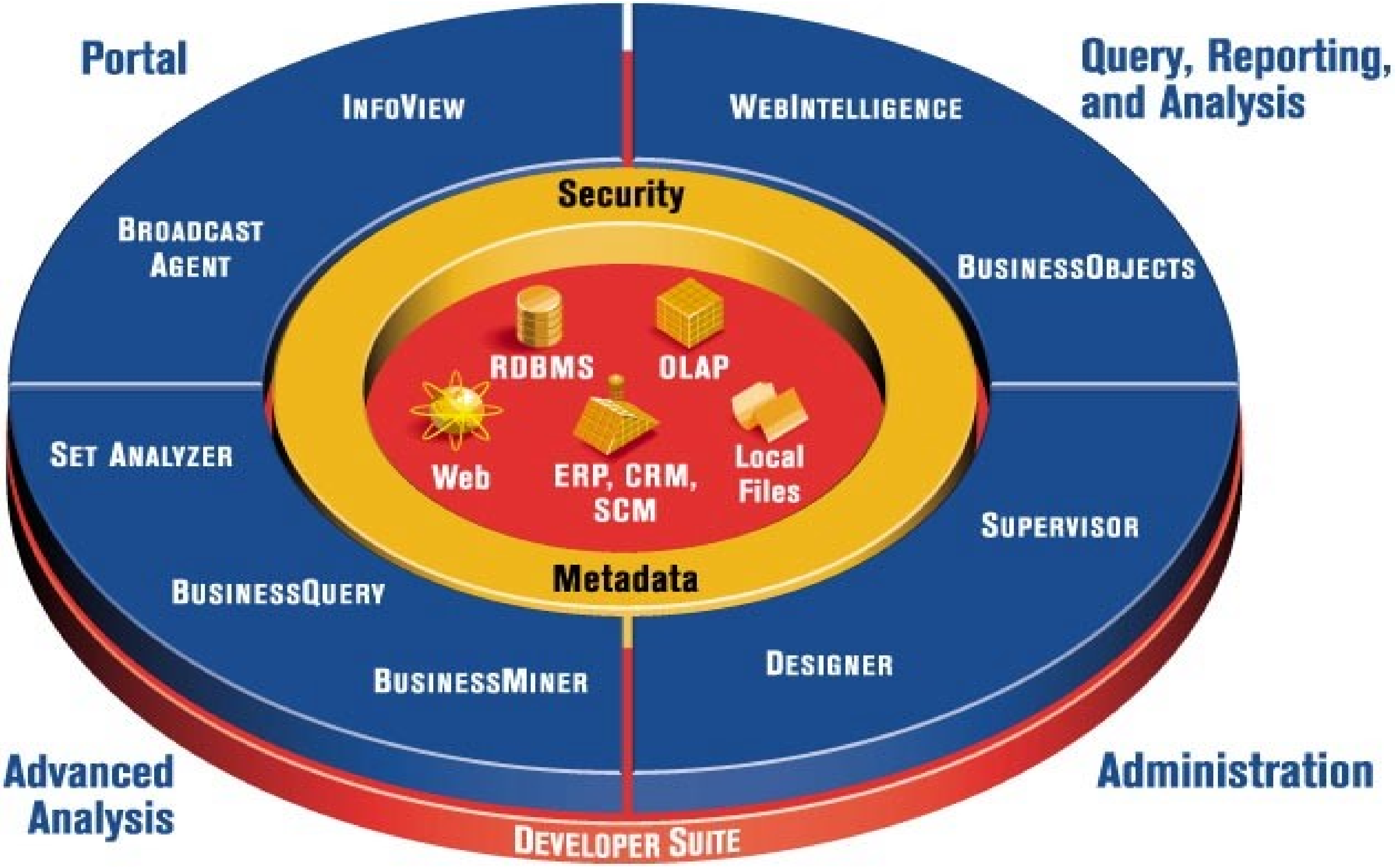
ROLAP Example - MicroStrategy: Big Picture



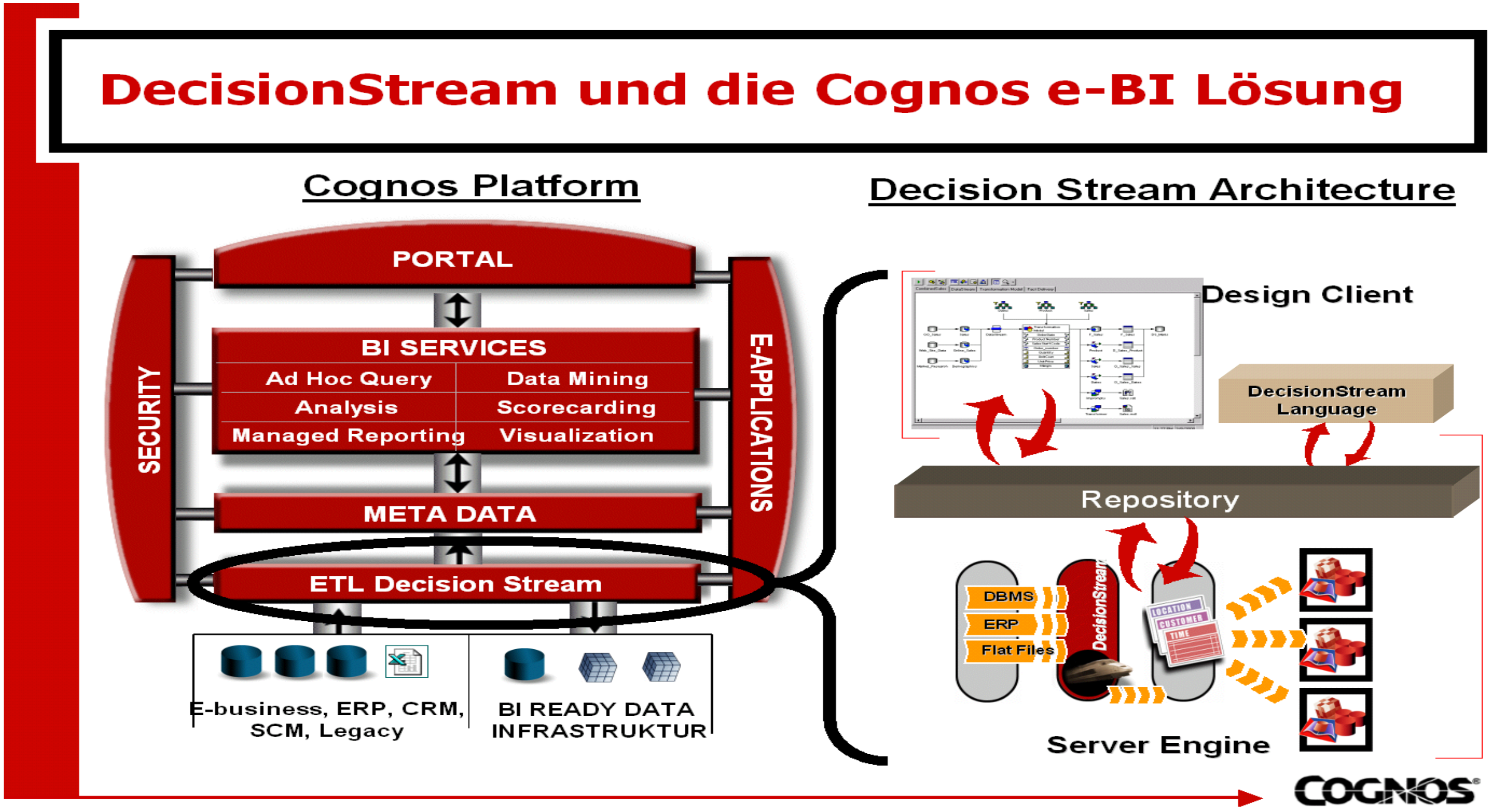
MSI 7 Pillars:

- Scalable
- BI Platform
- Built for the Internet

OLAP/Reporting Ex. - BusinessObject /Big Picture

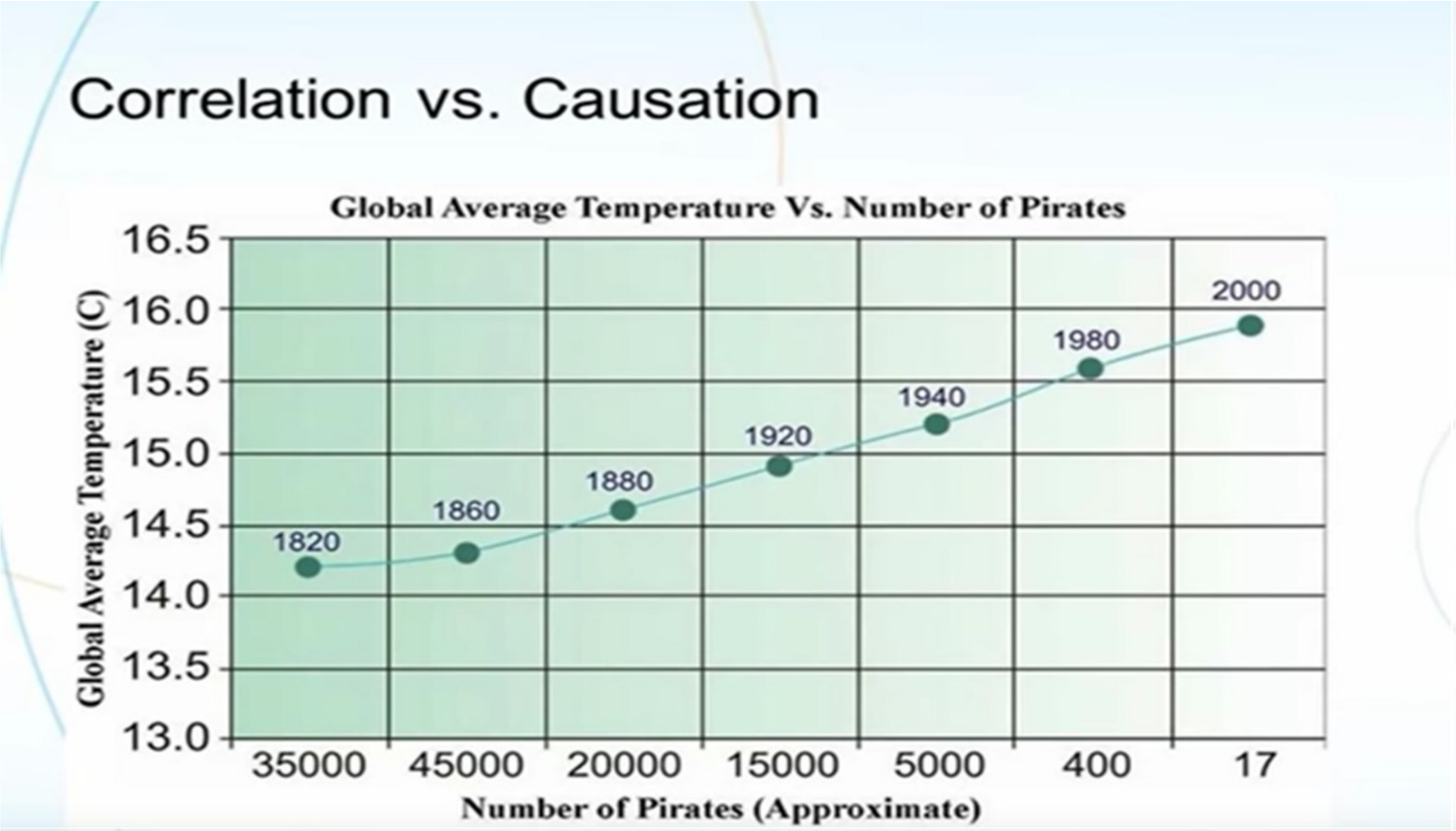


OLAP/Reporting Ex. - Cognos / Big Picture



See the following video about analytical dashboards in Data-Scientist/Dashboards (w. Cognos Dashboard Embedded):
https://dataplatform.cloud.ibm.com/docs/content/wsj/getting-started/videos.html?audience=cpdaas&context=cpdaas#data-scientists

Check Analytical Results for Business Context



Correlation vs. Causation

Exercise1 to Lesson 8: MOLAP <--> ROLAP

Exercise E8.1: Find and define the Benefits & Drawbacks of

- MOLAP
- ROLAP

Systems

Use the information of the lesson or use your own experience

Solution to Exercise1 of Lesson 8: MOLAP

Benefits

- Faster query performance
- Little in-flight calculation time
- Can write back to database
- More sophisticated calculations possible

Drawbacks

- Size limited by architecture of cube
- Can't access data that is not in cubes
- Housekeeping/backups limited
- Can't exploit database parallelism

Solution to Exercise1 of Lesson 8: ROLAP

Benefits

- Full use of database security/integrity
- Scalable to larger data volumes
- Data can be shared with other SQL applications
- Data and structure more dynamic

Drawbacks

- Slower queries
- Expensive to build
- Indexes and summaries not maintained automatically
- Calculations may be limited to database functions
- Less "Open" - proprietary clients

Exercise2 to Lesson 8: OLAP/Reporting Tools

Exercise E8.2 (SW*): Show the Highlights and build a Strengthens / Weakness Diagram for the following three Reporting Tools. Use the information from the internet:

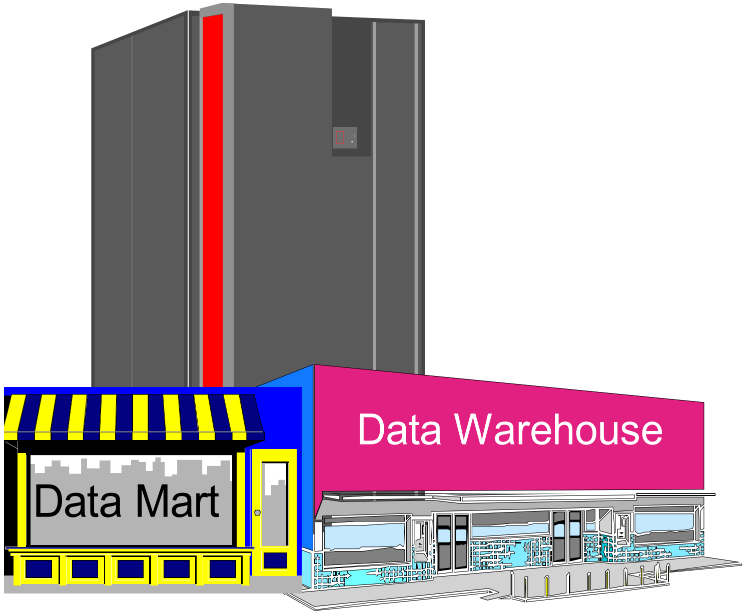
1. MicroStrategy --→ www.MicroStrategy.com
2. BusinessObjects ---→ www.BusinessObjects.com
3. Cognos ---→ www.Cognos.com

Show the three tools in competition to each other

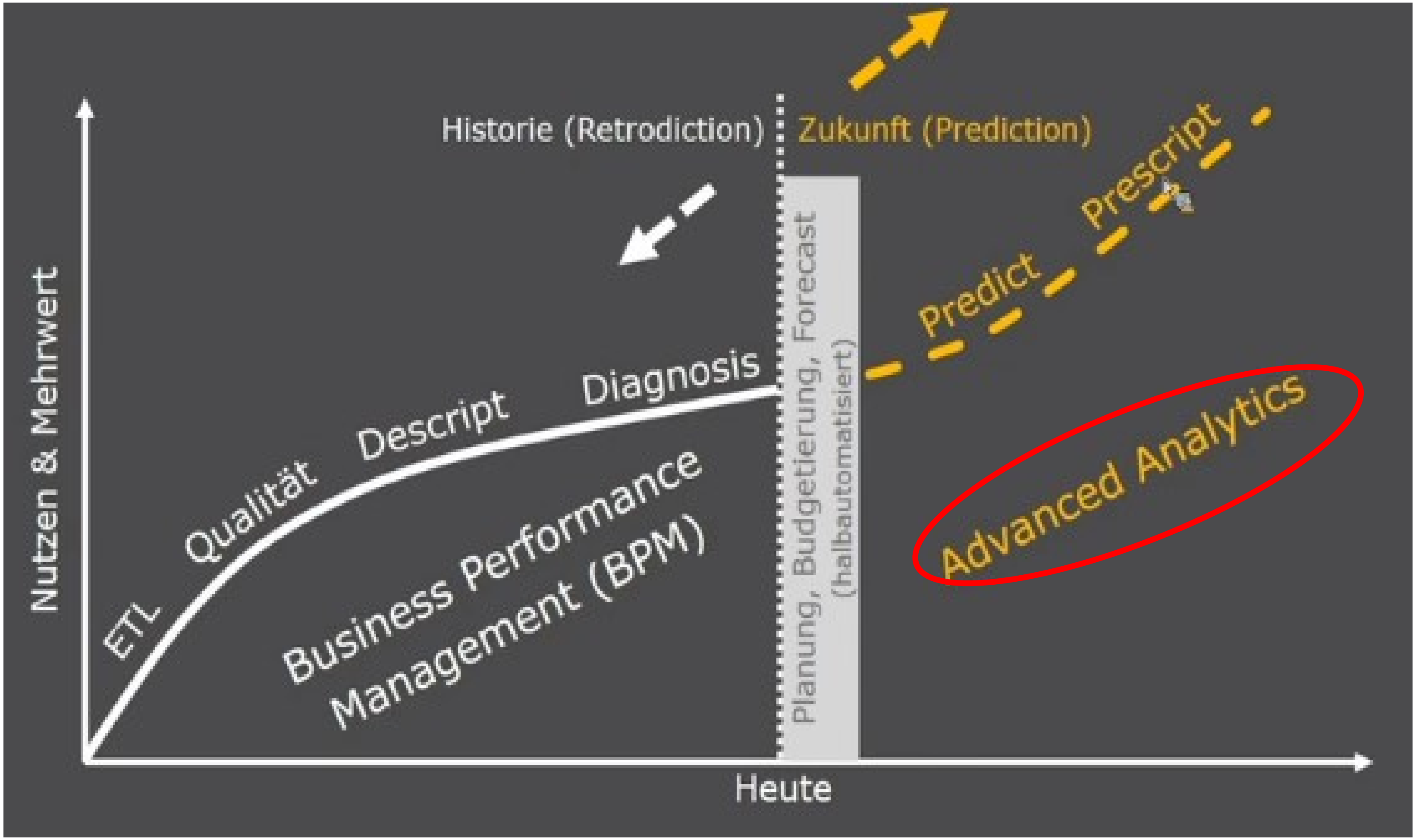
SW*: For the Seminar Work paper investigate this in more detail.

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

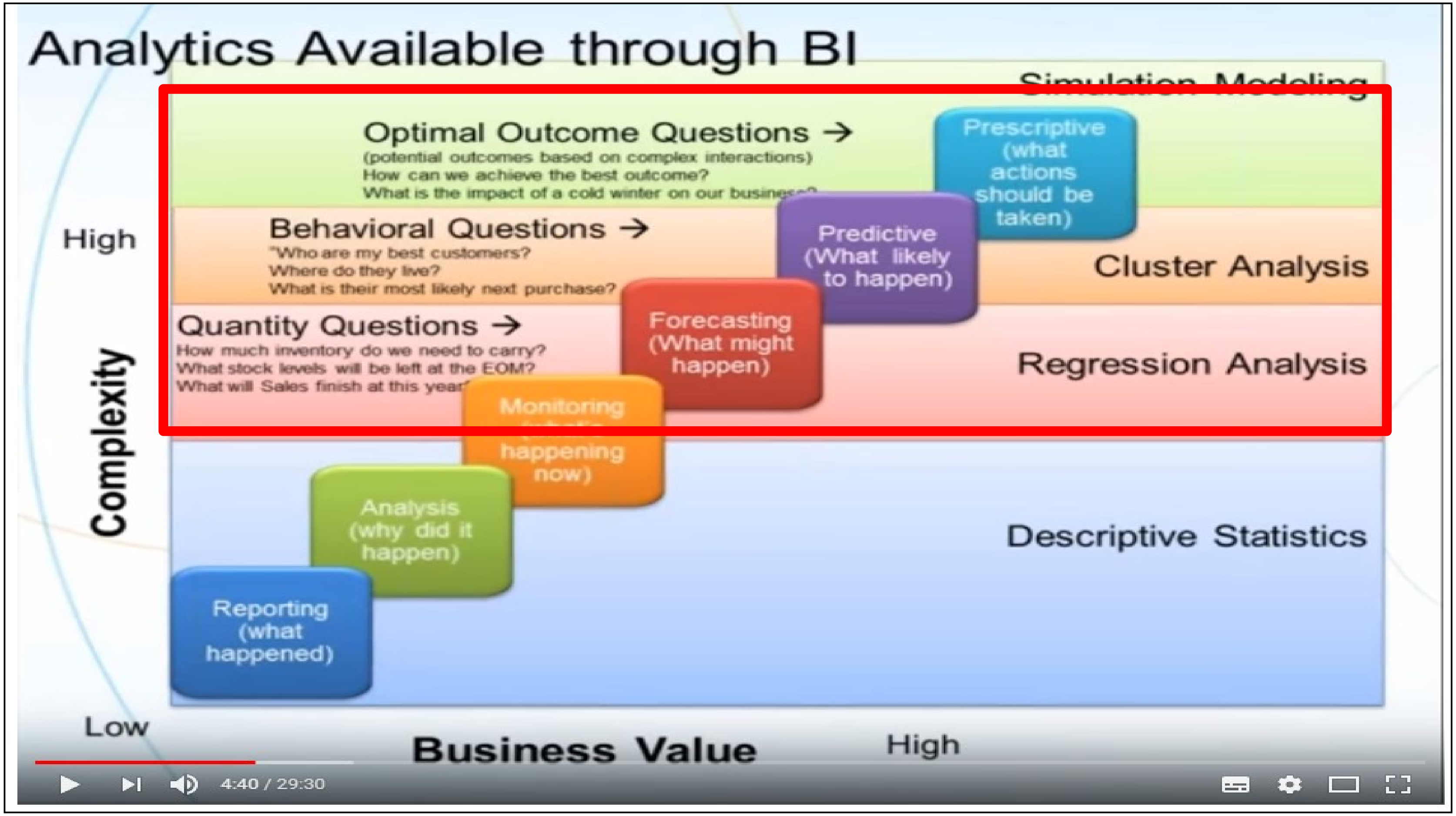
DW09 - Advanced Analytics I: Data Mining - Introduction & First Methods



Motivation - From Descriptive to Predictive Analytics



Advanced Analytics (AA) – Six Levels of Analytics



Analytics Available through BI

Optimal Outcome Questions →
(potential outcomes based on complex interactions)
How can we achieve the best outcome?
What is the impact of a cold winter on our busine

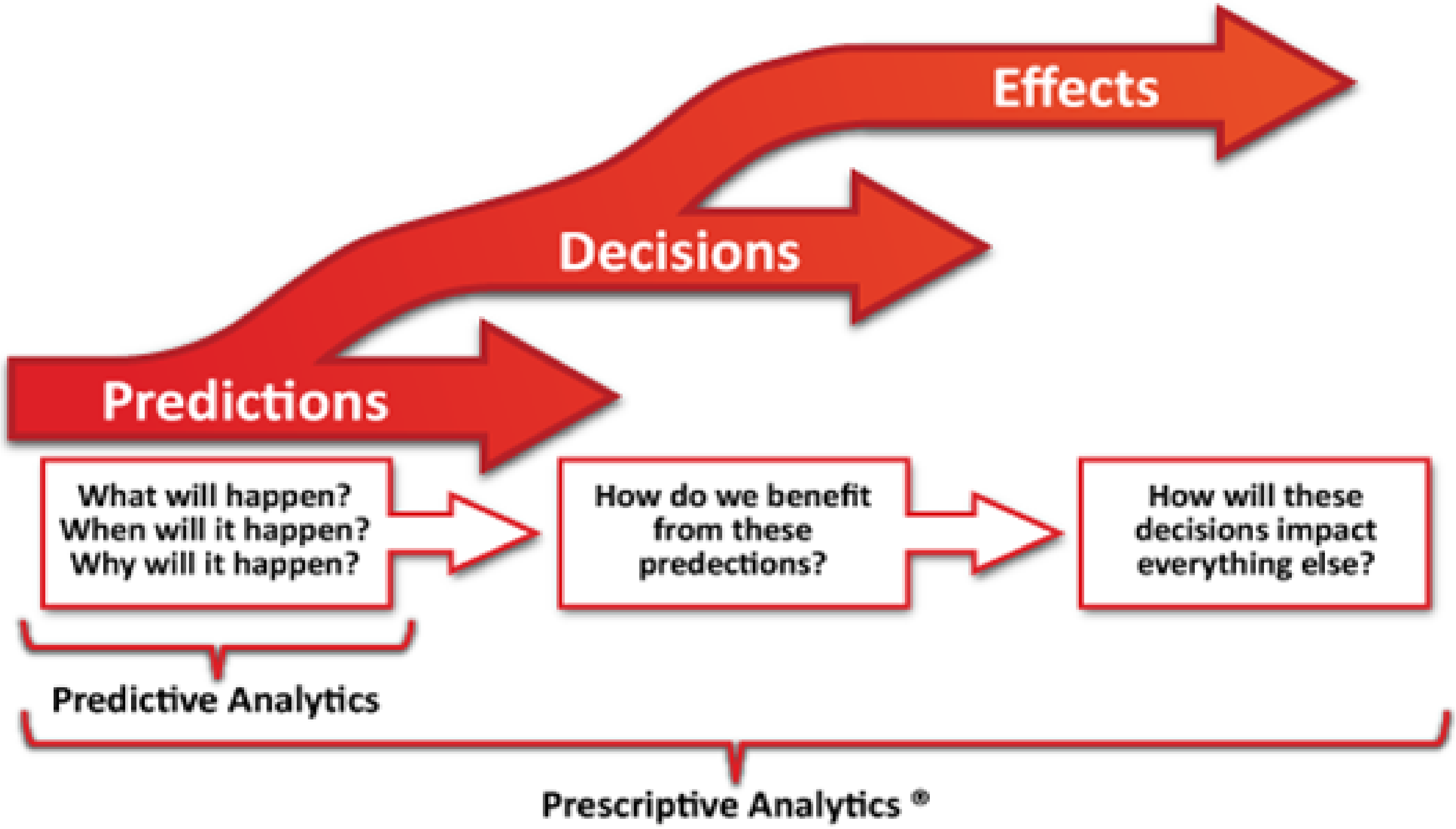
Behavioral Questions →
"Who are my best customers?
Where do they live?
What is their most likely next purchase?

Quantity Questions →
How much inventory do we need to carry?
What stock levels will be left at the EOM?
What will Sales finish at this year

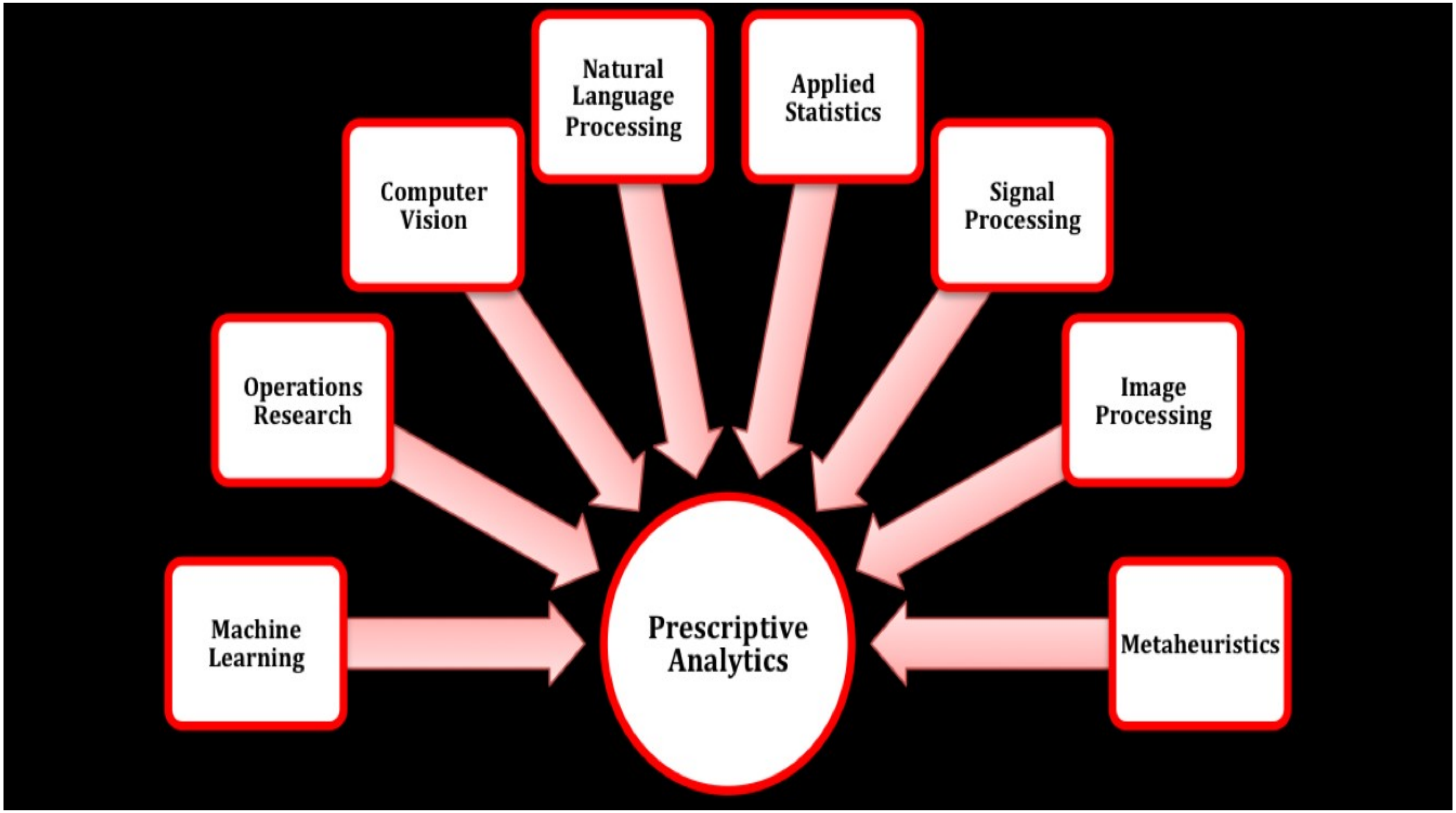
Simulation Modeling
Prescriptive (what actions should be taken)
Predictive (What likely to happen)
Cluster Analysis
Forecasting (What might happen)
Regression Analysis
Monitoring (what's happening now)
Analysis (why did it happen)
Reporting (what happened)
Descriptive Statistics
High
Complexity
Low
Business Value
High
4:40 / 29:30

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

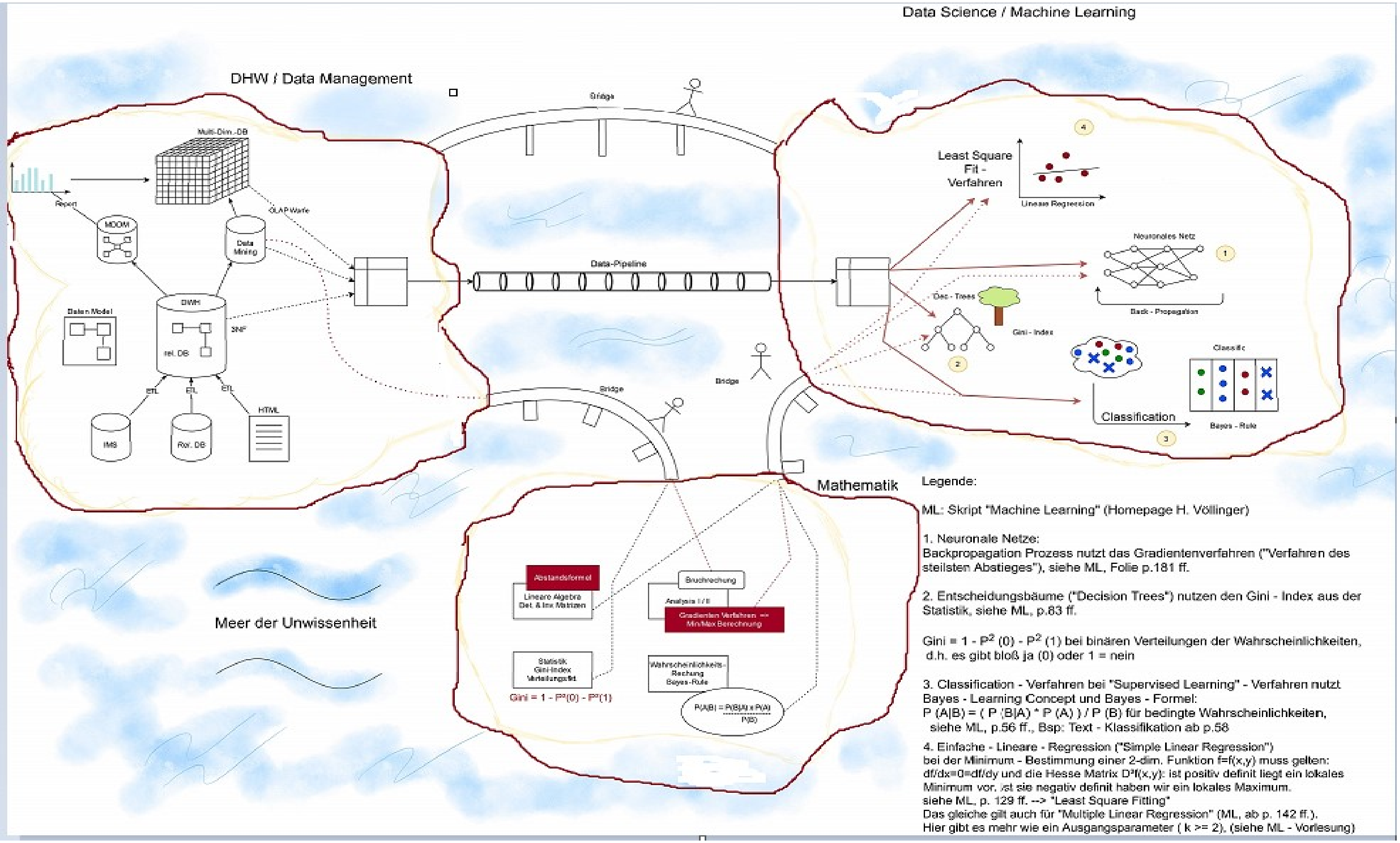
Advanced Analytics – Prescriptive Analytics



Prescriptive Analytics – Using Data Scientific Methods



“Bridges”: DWH/Data M.-Mathematics-ML/Data Science



Legende:

ML: Skript "Machine Learning" (Homepage H. Völlinger)

1. Neuronale Netze:
Backpropagation Prozess nutzt das Gradientenverfahren ("Verfahren des steilsten Abstieges"), siehe ML, Folie p.181 ff.

2. Entscheidungsbäume ("Decision Trees") nutzen den Gini - Index aus der Statistik, siehe ML, p.83 ff.

Gini = 1 - P^2 (0) - P^2 (1) bei binären Verteilungen der Wahrscheinlichkeiten, d.h. es gibt bloß ja (0) oder 1 = nein

3. Classification - Verfahren bei "Supervised Learning" - Verfahren nutzt Bayes - Learning Concept und Bayes - Formel:
P (A|B) = (P (B|A) * P (A)) / P (B) für bedingte Wahrscheinlichkeiten, siehe ML, p.56 ff., Bsp: Text - Klassifikation ab p.58

4. Einfache - Lineare - Regression ("Simple Linear Regression")
bei der Minimum - Bestimmung einer 2-dim. Funktion f=f(x,y) muss gelten:
df/dx=0=df/dy und die Hesse Matrix D²f(x,y): ist positiv definit liegt ein lokales Minimum vor. Ist sie negativ definit haben wir ein lokales Maximum.
siehe ML, p. 129 ff. --> "Least Square Fitting"
Das gleiche gilt auch für "Multiple Linear Regression" (ML, ab p. 142 ff.).
Hier gibt es mehr wie ein Ausgangsparameter (k >= 2), (siehe ML - Vorlesung)

Data Mining versus OLAP

Data Mining is not replacing OLAP, but enhancing it

With OLAP ...

you will only find information that you *looked for* in the first place. This is called *verification-driven analysis*.

Definition of Data Mining

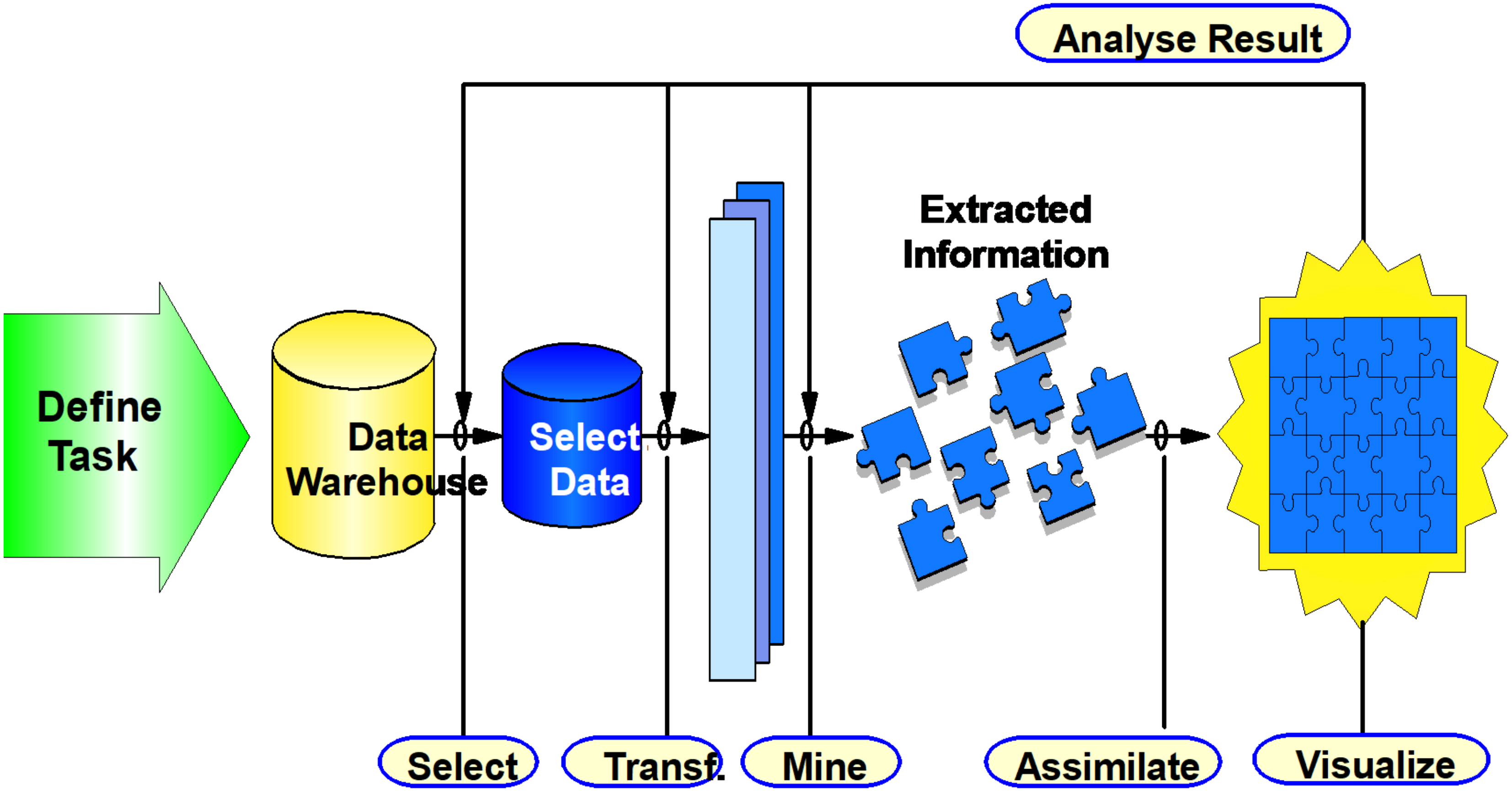
Data Mining is ...

The process of extracting *previously unknown*, *comprehensible*, and *actionable* information from large databases and using it to make crucial business decisions.

Who and where you need Data Mining

- **Telco, Insurance, Banks, Governments**
  - Fraud detection, Customer retention (Churn)
- **Retail industry**
  - Market-basket analysis
- **Manufacturing industry** :
  - Process and quality management
- **All industries (including Internet)**
  - Customer analysis and segmentation
  - Direct mailing optimization
  - Customer retention, pricing
  - Customer scoring

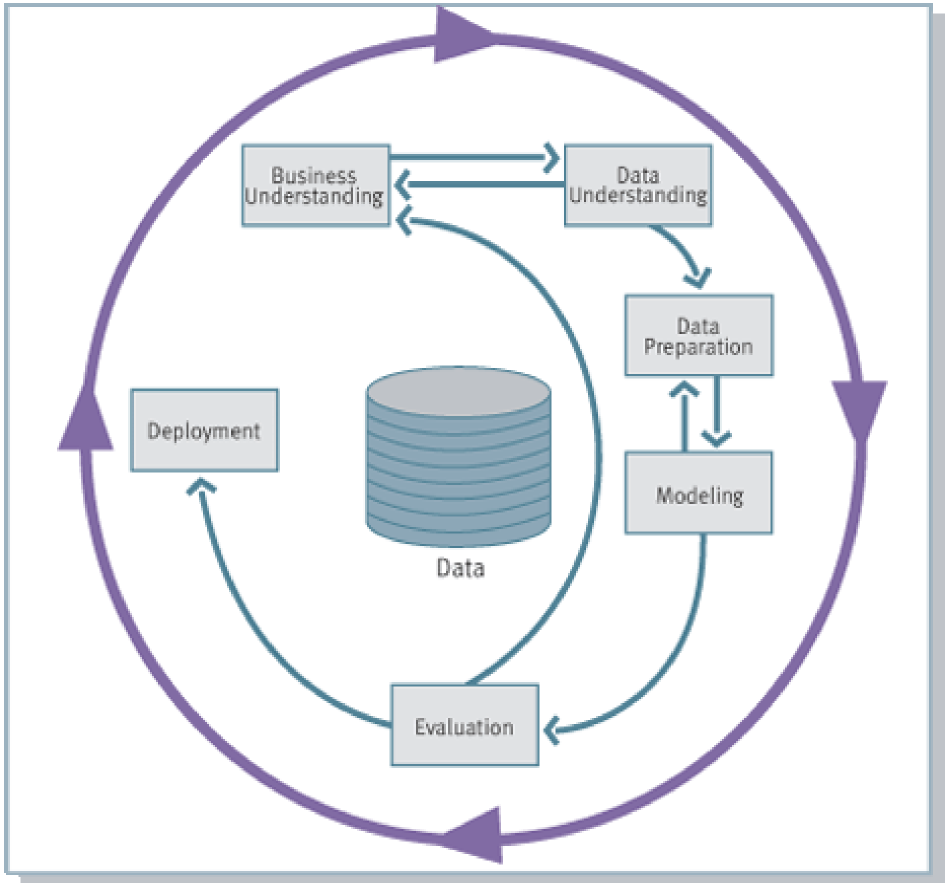
The Data Mining Process



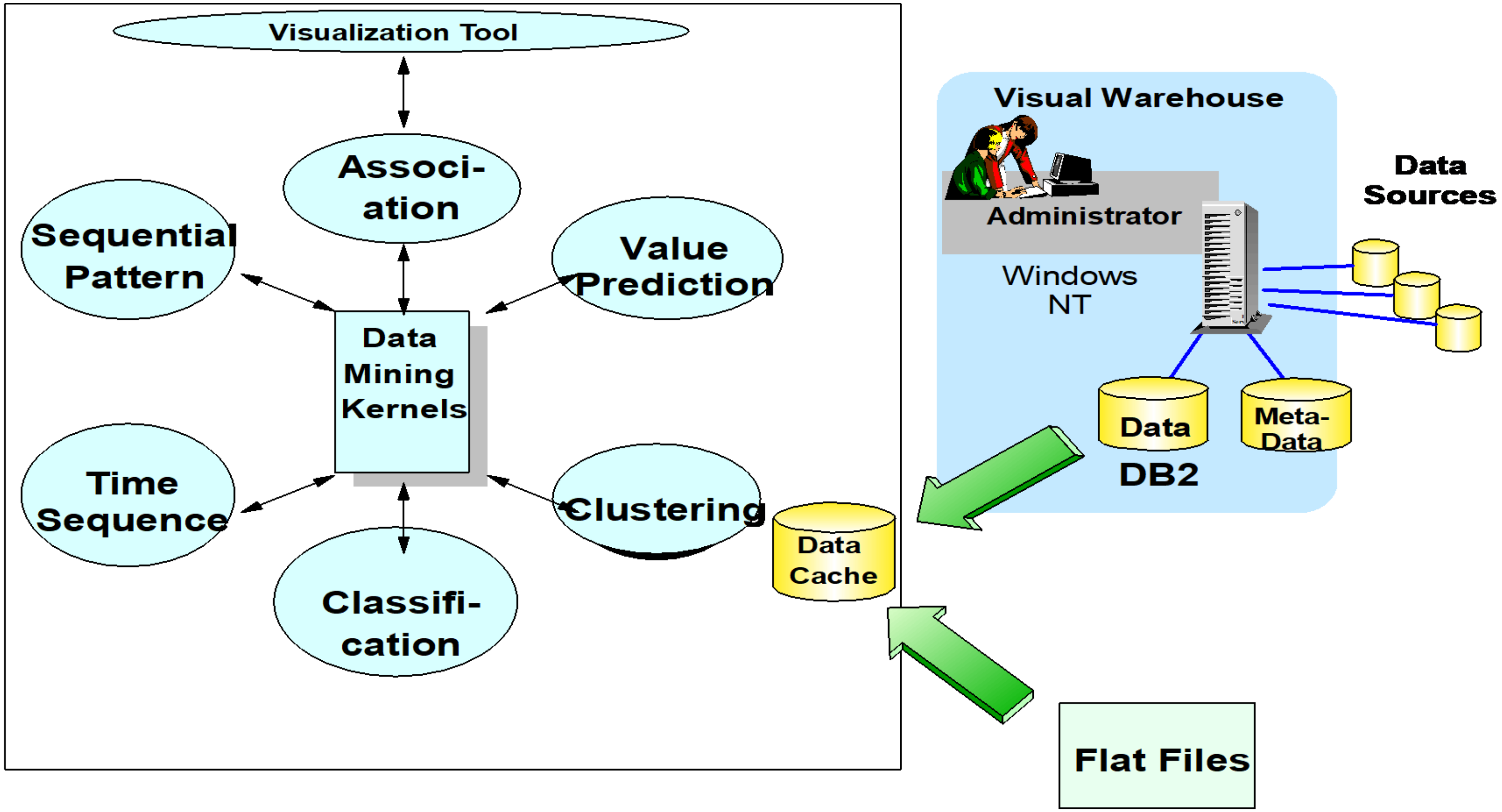
The CRISP*- DM Process Model

* CRoss-Industry Standard Process Model

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment



Example: Intelligent Miner for Data - Overview



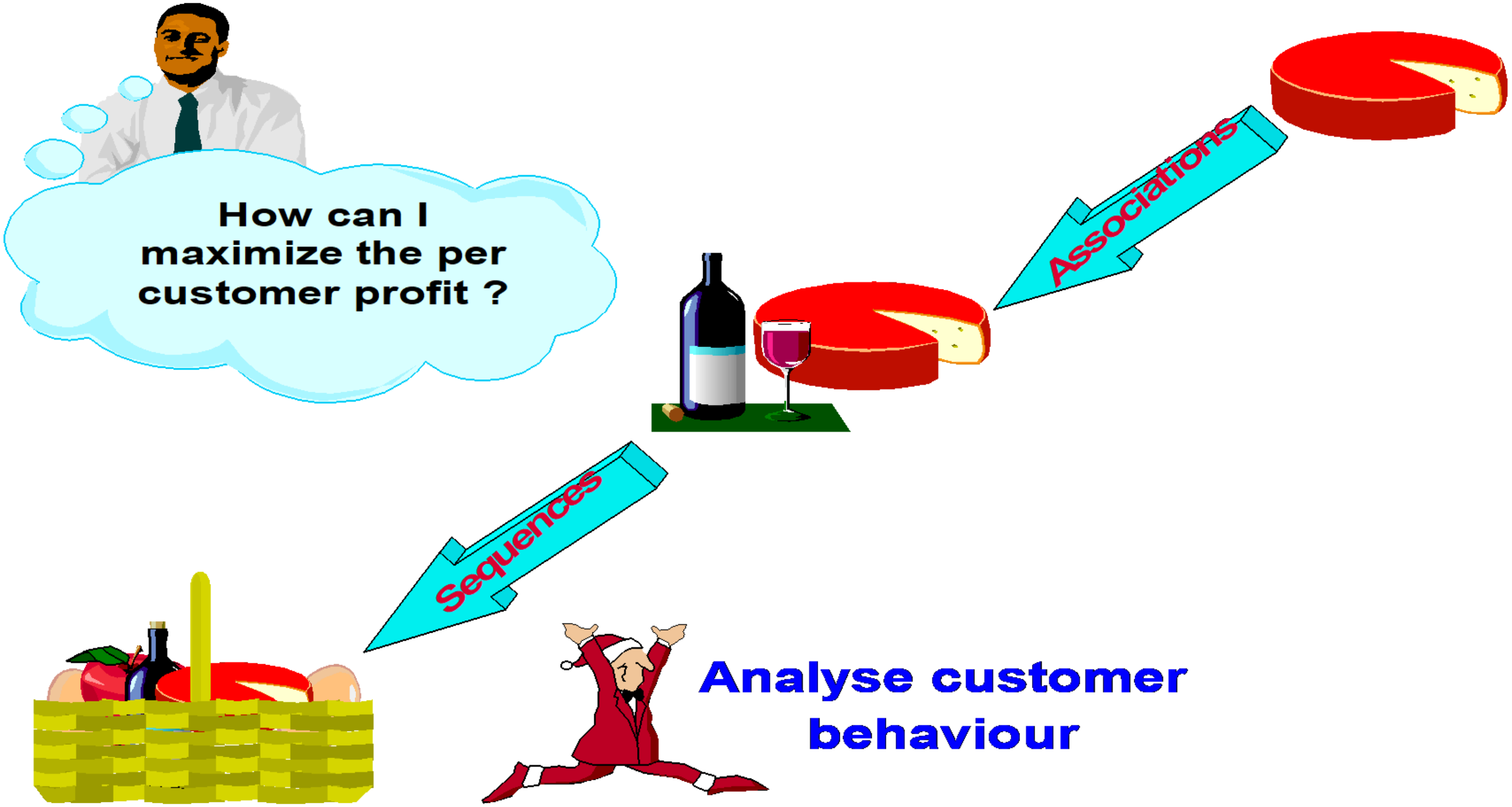
See the following video about the tool SPSS in Data-Scientist/SPSS Modeler (“Score prediction- diagnose diseases…”): https://dataplatform.cloud.ibm.com/docs/content/wsj/getting-started/videos.html?audience=cpdaas&context=cpdaas#data-scientists

Overview about Data Mining Applications

1. Market Basket Analysis
2. Cross Selling
3. Customer Retention
4. Fraud Detection
5. Campaign Management

No.	***Application***	***IM4D Technique***
1	Market Basket Analysis (MBA)	Associations, Sequential Patterns
2	Cross Selling (CS)	Associations, Classification, Clustering
3	Customer Retention (CR)	Clustering, Classification, Value Prediction
4	Fraud Detection (FD)	Associations, Sequential Pattern, Time Sequence
5	Campaign Management (CM)	Clustering, Classification, Value Prediction

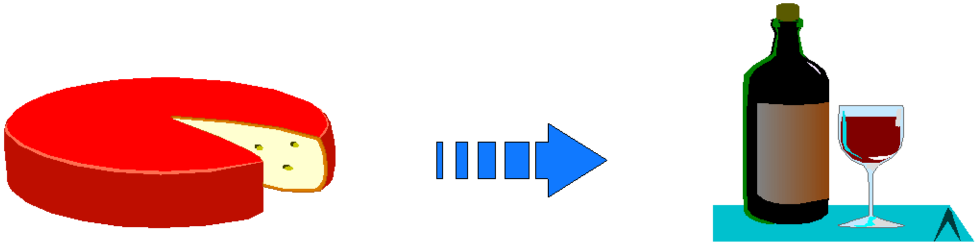
Market Basket Analysis – Business Idea


How can I maximize the per customer profit ?
Associations
Sequences
Analyse customer behaviour

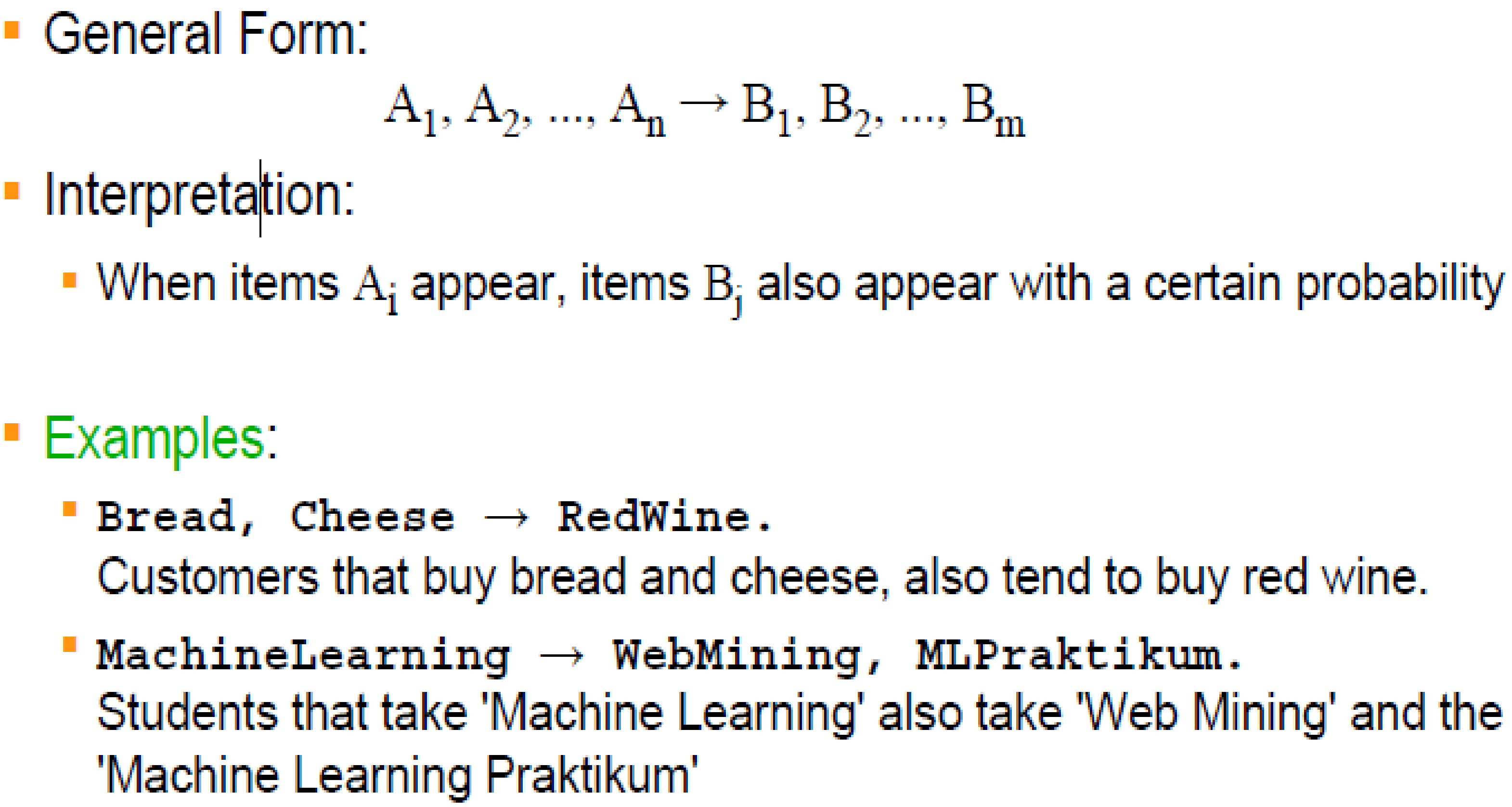
Market Basket Analysis - Assocations

- Search the table for all available combines and evaluate the frequencies
- ***Results***

  If a customer buys "product A", then he buys "product B" in Z% of the time. This association is present in X% of all bills



Association Rules – General Form



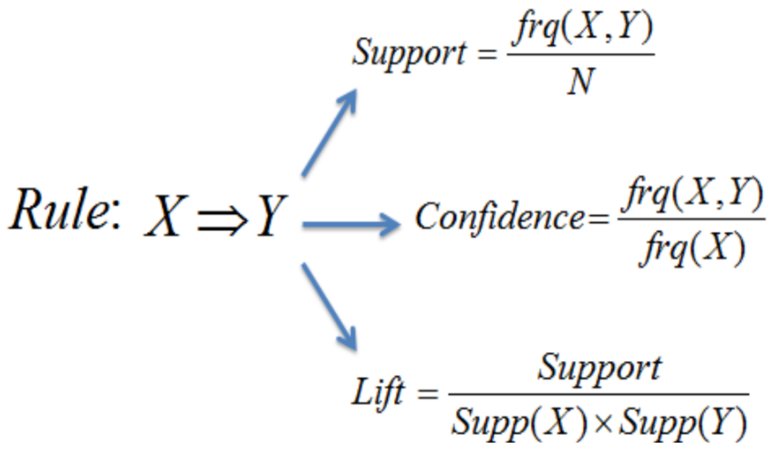
- General Form:

  $$A_1, A_2, \ldots, A_n \rightarrow B_1, B_2, \ldots, B_m$$

- Interpretation:
  - When items A_i appear, items B_j also appear with a certain probability

- Examples:
  - `Bread, Cheese → RedWine.`
    Customers that buy bread and cheese, also tend to buy red wine.
  - `MachineLearning → WebMining, MLPraktikum.`
    Students that take 'Machine Learning' also take 'Web Mining' and the 'Machine Learning Praktikum'

Association Rules – Definition of Popular Measures



$$Rule: \; X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

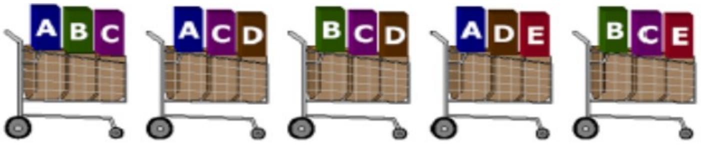
$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

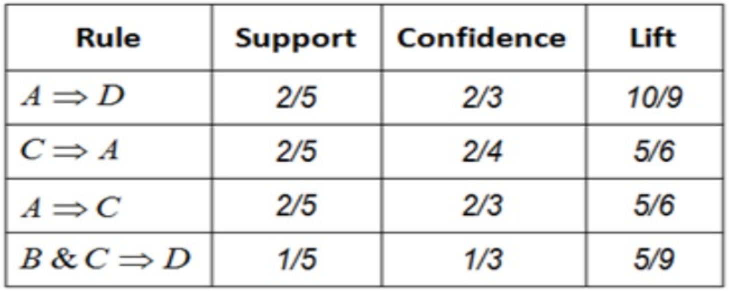
Symmetry Properties:

- Sup(X=>Y) = Sup(Y=>X)
- Lift(X=>Y) = Lift(Y=>X)

Question:

- How many rules have you to consider?
- Prove the answer: You have to consider 40 rules. With symmetry this are 80 rules.

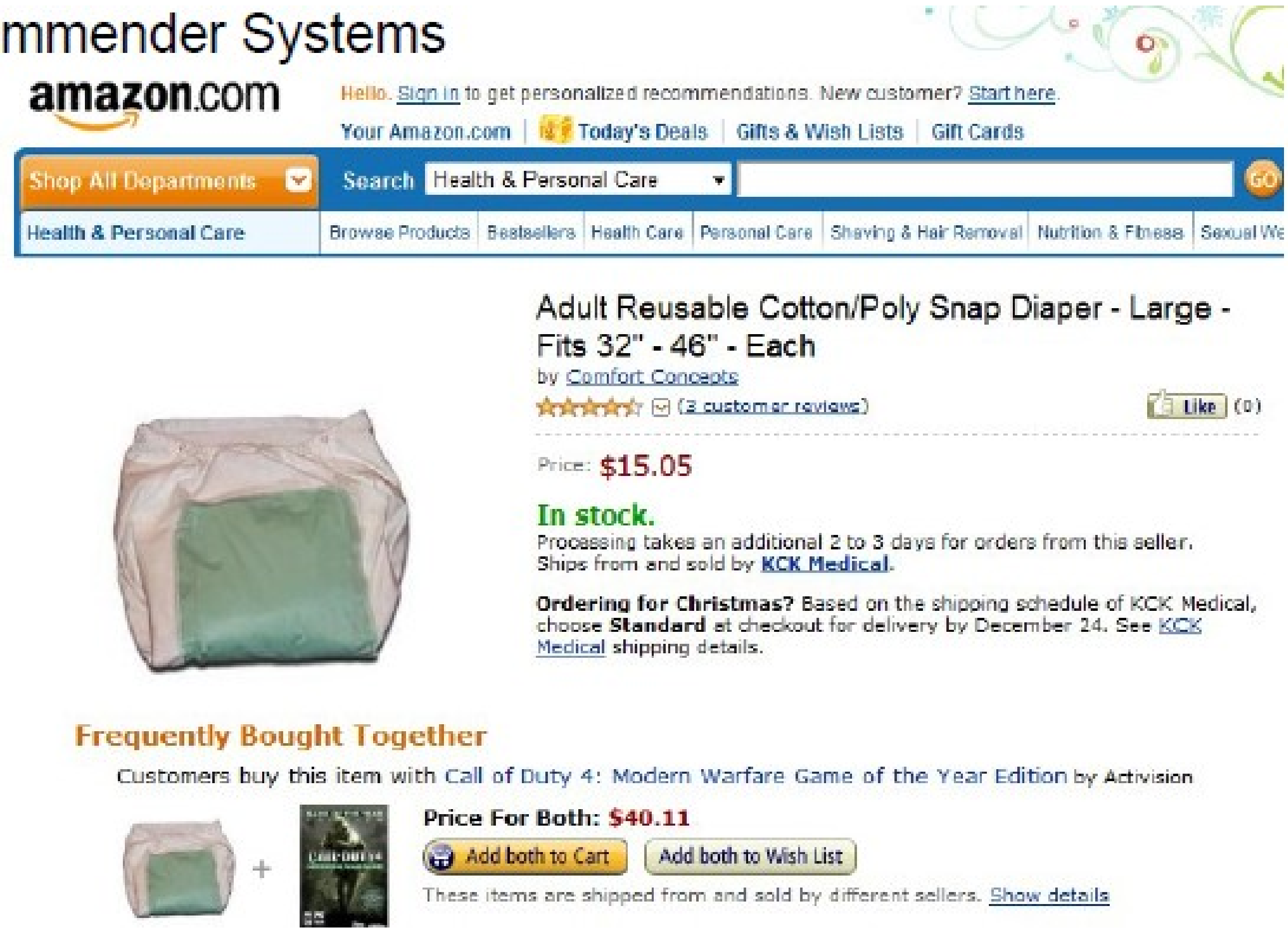




Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Association Rules – Example of predictive MBA

- Recommender Systems



Hello. Sign in to get personalized recommendations. New customer? Start here.

Adult Reusable Cotton/Poly Snap Diaper - Large - Fits 32" - 46" - Each

by Comfort Concepts

(3 customer reviews)

Price: $15.05

In stock.

Processing takes an additional 2 to 3 days for orders from this seller.
Ships from and sold by KCK Medical.

Ordering for Christmas? Based on the shipping schedule of KCK Medical, choose **Standard** at checkout for delivery by December 24. See KCK Medical shipping details.

Frequently Bought Together

Customers buy this item with Call of Duty 4: Modern Warfare Game of the Year Edition by Activision

Price For Both: $40.11

These items are shipped from and sold by different sellers. Show details

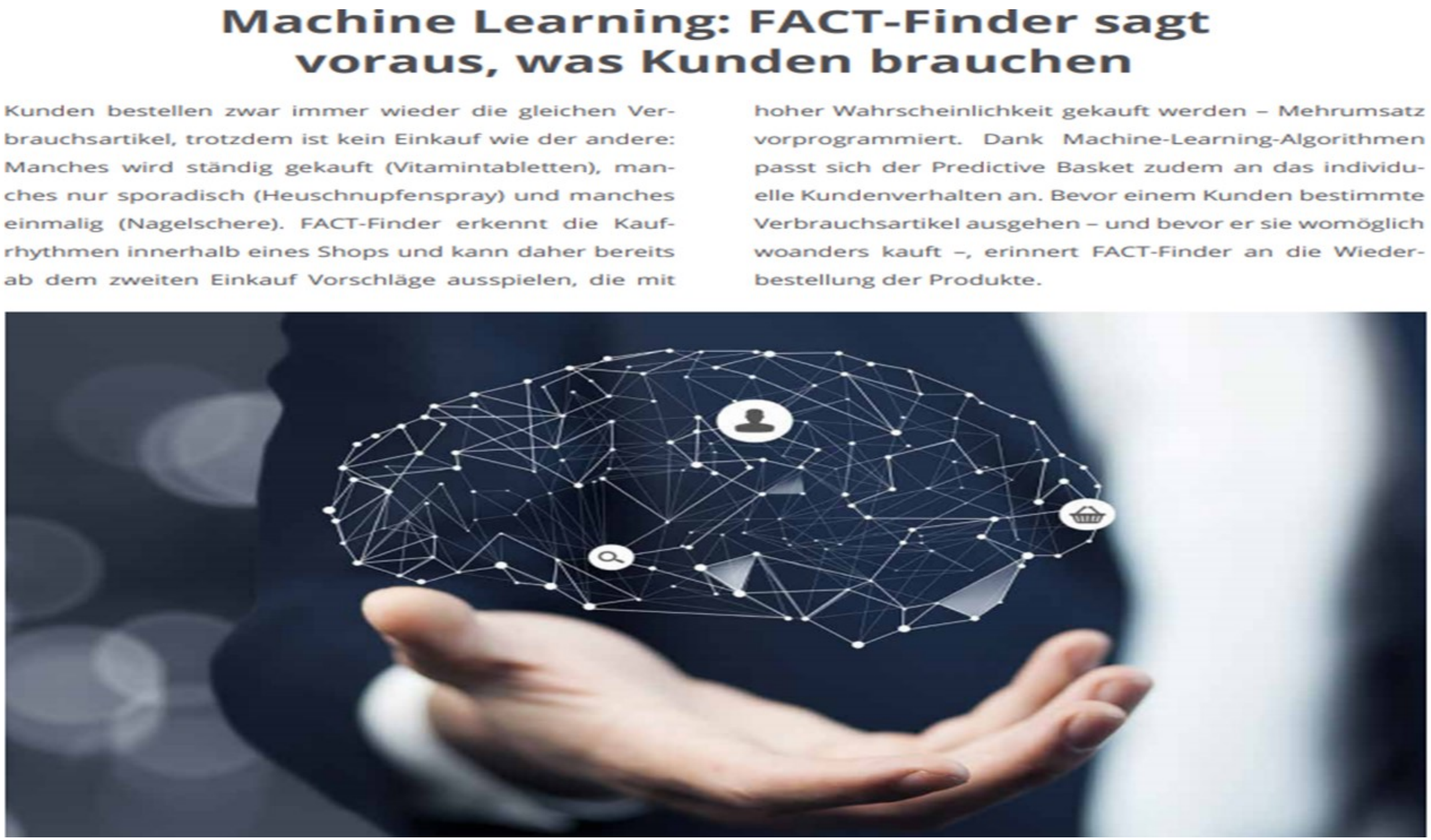
Market Basket Analysis - Sequential Patterns

- Search the table for all available sequences and evaluate the frequencies
- **Results**

If a customer buys "product A", then he buys later "product B". This sequence is present in X% of the total amount of sequences.

Customer 1	**Day 1**	**Product 1**
Customer 8	**Day 1**	**Product 1**
Customer 1	**Day 4**	**Product 2**

UseCase – “Semantic Search - Predictive Market with Fact-Finder” https://youtu.be/vSWLafBdHus



Machine Learning: FACT-Finder sagt voraus, was Kunden brauchen

Kunden bestellen zwar immer wieder die gleichen Verbrauchsartikel, trotzdem ist kein Einkauf wie der andere: Manches wird ständig gekauft (Vitamintabletten), manches nur sporadisch (Heuschnupfenspray) und manches einmalig (Nagelschere). FACT-Finder erkennt die Kaufrhythmen innerhalb eines Shops und kann daher bereits ab dem zweiten Einkauf Vorschläge ausspielen, die mit hoher Wahrscheinlichkeit gekauft werden – Mehrumsatz vorprogrammiert. Dank Machine-Learning-Algorithmen passt sich der Predictive Basket zudem an das individuelle Kundenverhalten an. Bevor einem Kunden bestimmte Verbrauchsartikel ausgehen – und bevor er sie womöglich woanders kauft –, erinnert FACT-Finder an die Wiederbestellung der Produkte.

Exercise1 to Lesson 9: Data Mining Techniques

Exercise E9.1: Describe the following Data Mining techniques. Search this information in the internet, i.e. Wikipedia or other knowledge portals:

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

Exercise2 to Lesson 9: Data Mining Techniques

Exercise E9.2: Describe the following Data Mining techniques. Search this information in the internet, i.e. Wikipedia or other knowledge portals:

- **Sequential Patterns**
- **Value Prediction**
- **Similar Time Sequences**

Exercise 3 to Lesson 9: Association Measures

Exercise E9.3: Remember the following measures for Association: *support, confidence and lift.*
Calculate measures for the following 8 item sets of a shopping basket (1 person, 10 min):

{ Milch, Limonade, Bier }; { Milch, Apfelsaft, Bier }; { Milch, Apfelsaft, Orangensaft };{ Milch, Bier, Orangensaft, Apfelsaft };{ Milch, Bier };{ Limonade, Bier, Orangensaft }; { Orangensaft };{ Bier, Apfelsaft }

1. What is the support of the item set { Bier, Orangensaft }?
2. What is the confidence of { Bier } ➔ { Milch } ?
3. Which association rules have support and confidence of at least 50%?

Exercise 4 to Lesson 9: Use Case “Semantic Search”

Exercise E9.4 (SW*): Evaluate the Technology of the UseCase “Semantic Search”

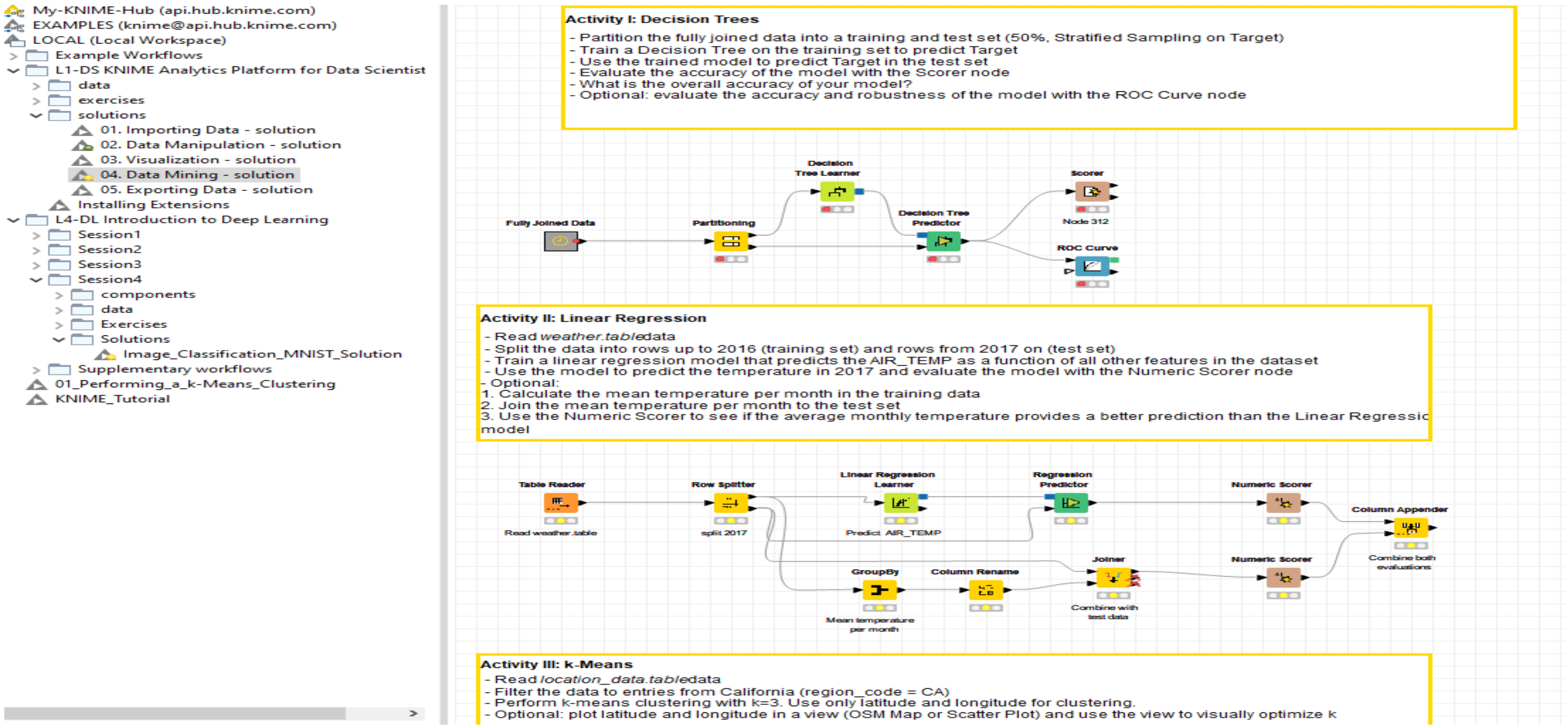
Groupwork (2 Persons): Evaluate and find the underlying technology which is used in “UseCase – Semantic Search: Predictive Basket with Fact-Finder”. See: https://youtu.be/vSWLafBdHus

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

Exercise 5 to Lesson 9: Performing KNIME DM-Basics

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

Homework for 2 Persons: KNIME-Basics Workflow (use given solution) for one of the 3 KNIME solutions and give a technical explanation to the solution steps (see image below).



My-KNIME-Hub (api.hub.knime.com)
EXAMPLES (knime@api.hub.knime.com)
LOCAL (Local Workspace)
- Example Workflows
- L1-DS KNIME Analytics Platform for Data Scientist
  - data
  - exercises
  - solutions
    - 01. Importing Data - solution
    - 02. Data Manipulation - solution
    - 03. Visualization - solution
    - 04. Data Mining - solution
    - 05. Exporting Data - solution
  - Installing Extensions
- L4-DL Introduction to Deep Learning
  - Session1
  - Session2
  - Session3
  - Session4
    - components
    - data
    - Exercises
    - Solutions
      - Image_Classification_MNIST_Solution
  - Supplementary workflows
- 01_Performing_a_k-Means_Clustering
- KNIME_Tutorial

Activity I: Decision Trees

- Partition the fully joined data into a training and test set (50%, Stratified Sampling on Target)
- Train a Decision Tree on the training set to predict Target
- Use the trained model to predict Target in the test set
- Evaluate the accuracy of the model with the Scorer node
- What is the overall accuracy of your model?
- Optional: evaluate the accuracy and robustness of the model with the ROC Curve node

Activity II: Linear Regression

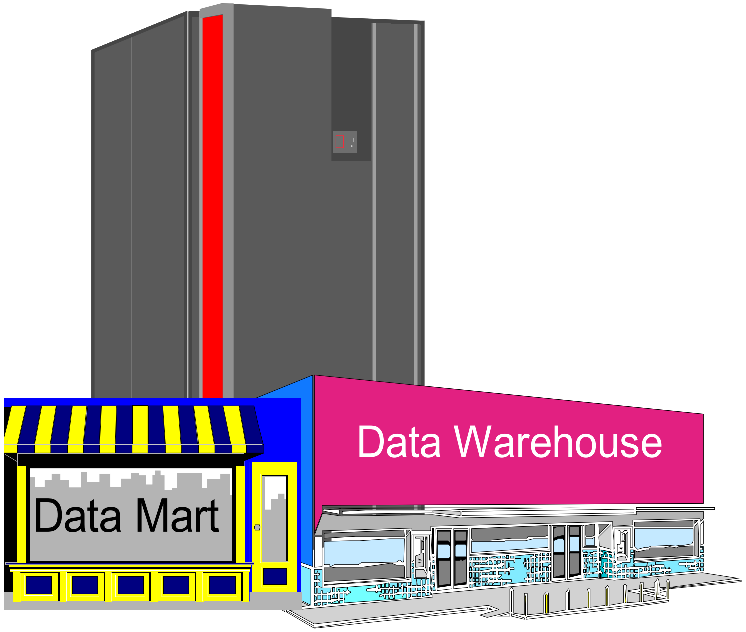
- Read *weather.table*data
- Split the data into rows up to 2016 (training set) and rows from 2017 on (test set)
- Train a linear regression model that predicts the AIR_TEMP as a function of all other features in the dataset
- Use the model to predict the temperature in 2017 and evaluate the model with the Numeric Scorer node
- Optional:
1. Calculate the mean temperature per month in the training data
2. Join the mean temperature per month to the test set
3. Use the Numeric Scorer to see if the average monthly temperature provides a better prediction than the Linear Regressic model

Activity III: k-Means

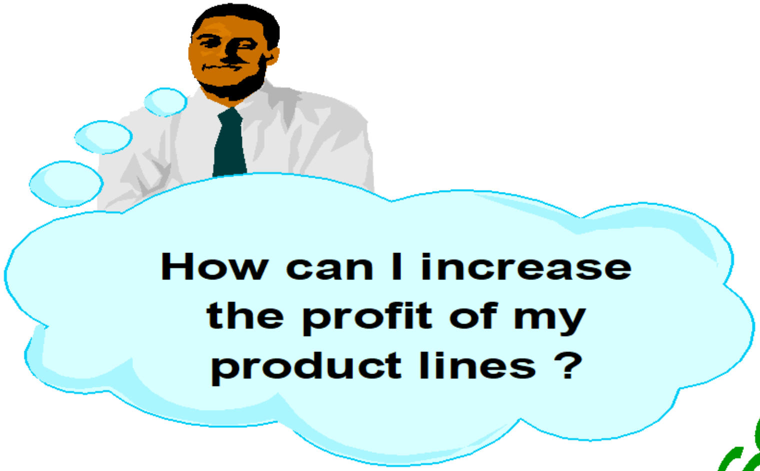
- Read *location_data.table*data
- Filter the data to entries from California (region_code = CA)
- Perform k-means clustering with k=3. Use only latitude and longitude for clustering.
- Optional: plot latitude and longitude in a view (OSM Map or Scatter Plot) and use the view to visually optimize k

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

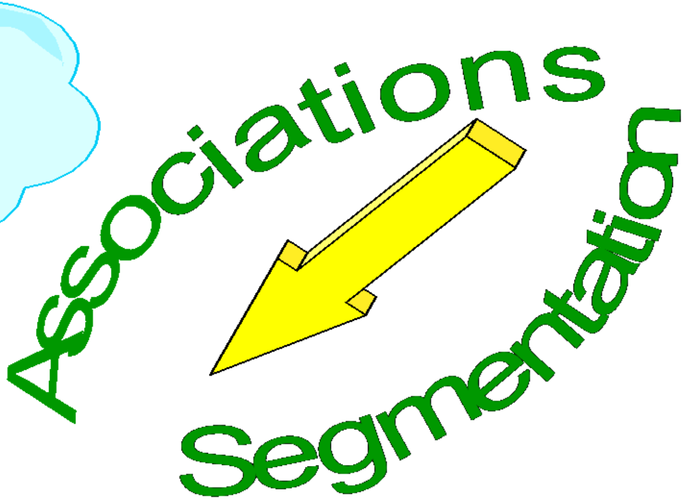
DW10 - Advanced Analytics II: Data Mining – Methods & Tools



Cross Selling – Business Idea



How can I increase the profit of my product lines ?



Increase Customer Loyalty

Cross Seling - Methods

☛ ***Analyse relation products - customer profiles***

- Use IM Tree / Neural Classification

☛ ***Create homogenous groups of customers, if customers can be identified***

- Use IM Clustering techniques

☛ ***Analyse products portofolios***

- Use IM Associations or Sequential Patterns

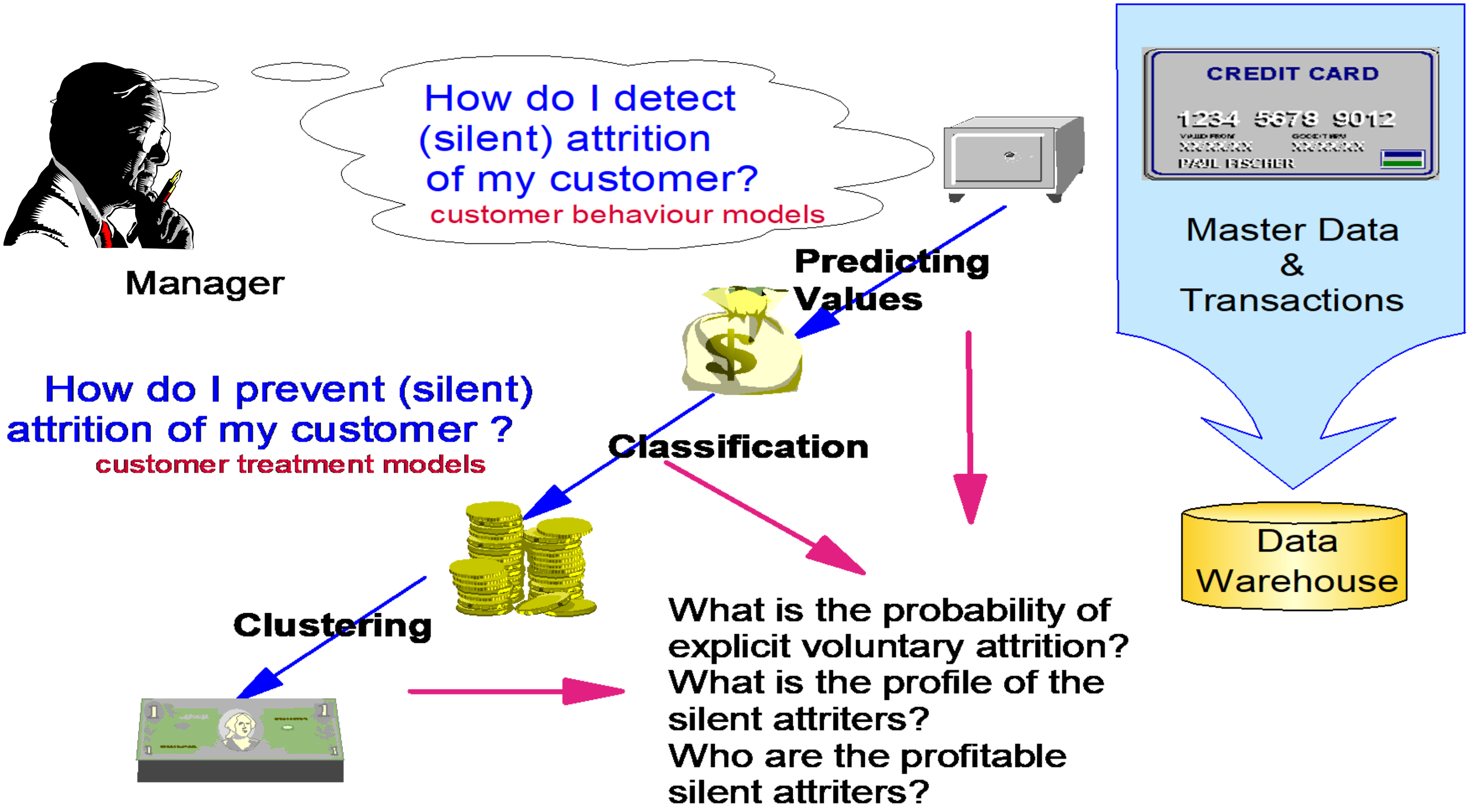
Cross Selling - Goals

☛ *Goal :*

- Offer complementary products to existing customers
- Detect when a customer's behaviour changes to offer him new products
- Build promotion strategies
- Create new products

Increase Profit with your marketshare

Customer Retention – Business Idea



Manager

How do I detect (silent) attrition of my customer?

customer behaviour models

CREDIT CARD

1234 5678 9012

PAUL FISCHER

Master Data & Transactions

Data Warehouse

Predicting Values

Classification

Clustering

How do I prevent (silent) attrition of my customer ?

customer treatment models

What is the probability of explicit voluntary attrition?
What is the profile of the silent attriters?
Who are the profitable silent attriters?

Customer Retention – Business Goals

- Identify customers who left
- Build a training model
  - ► Create training and test data on historical basis
  - ► Learn the algorithm with training data
  - ► Check results with test data
- Run model against current customer data

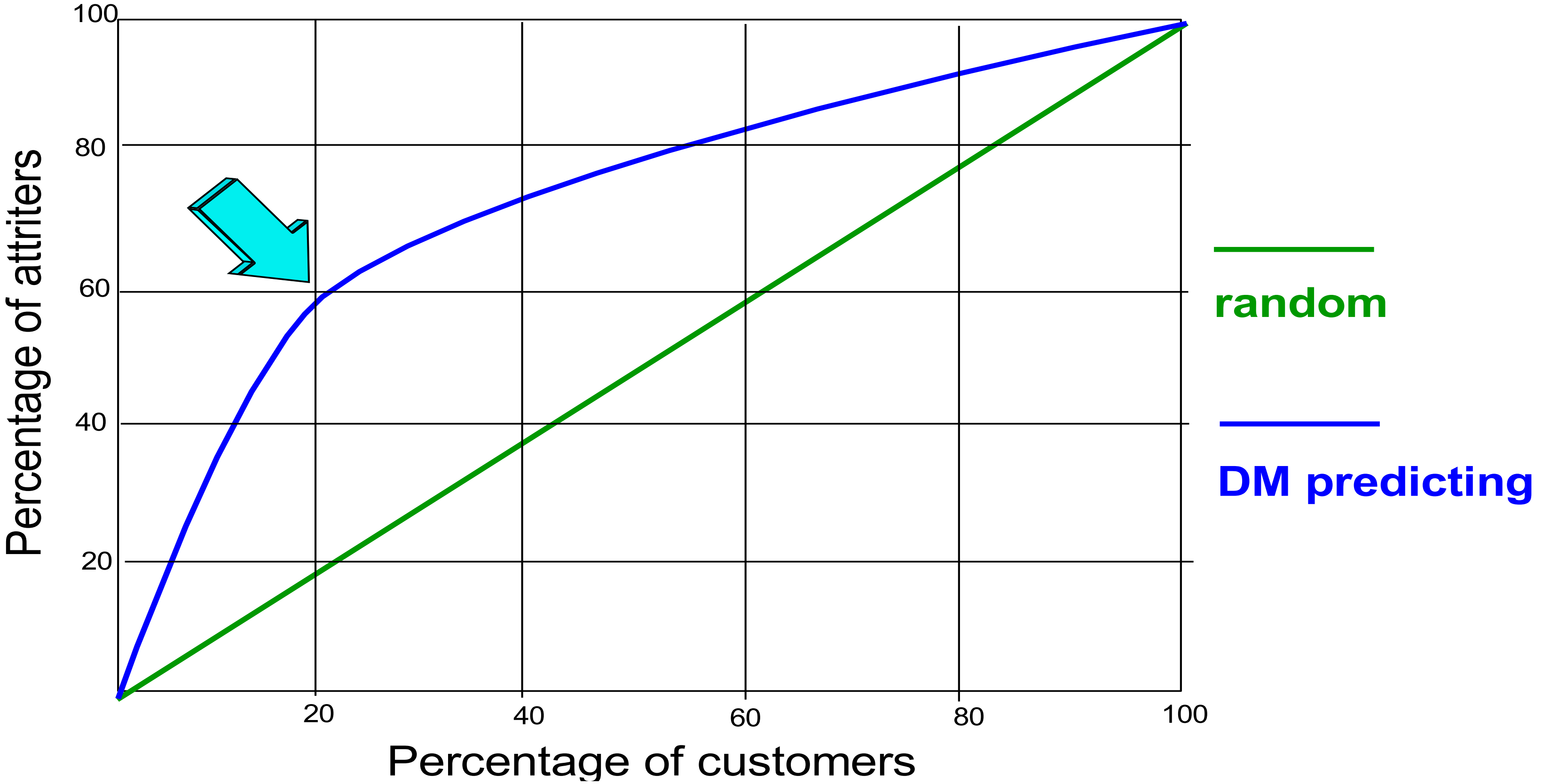
☛ ***Result Analysis***

☛ ***Business Implementation***

Customer Retention - Methods

- Data Mining
  - Customer scoring
    - Classification Tree / Neural
    - Prediction RBF / Neural
  - Characterize Defectors
    - Clustering Neural / Demographic

Customer Retention – Attrition Response Model



Customer Retention – Goal

☛ ***Goal :***

- Identify profitable customers with high probability of defection
- Execute campaign to target defectors
- Use model to be pro-active

Substantial cost saving

Fraud Detection – Idea & Goal

☛ ***Question :***

How is it possible to avoid the damages caused by fraudsters ?

☛ ***Goal :***

- Detect quickly fraudulent transactions
- Identify potential frauders
- Stop immediately services to frauders

Reduces risks, saves money

Campaign Management – Business Idea

Can I be more effecient in my direct marketing strategy ?

Clustering

Scoring

Increase response

Save money

Campaign Management – Methods

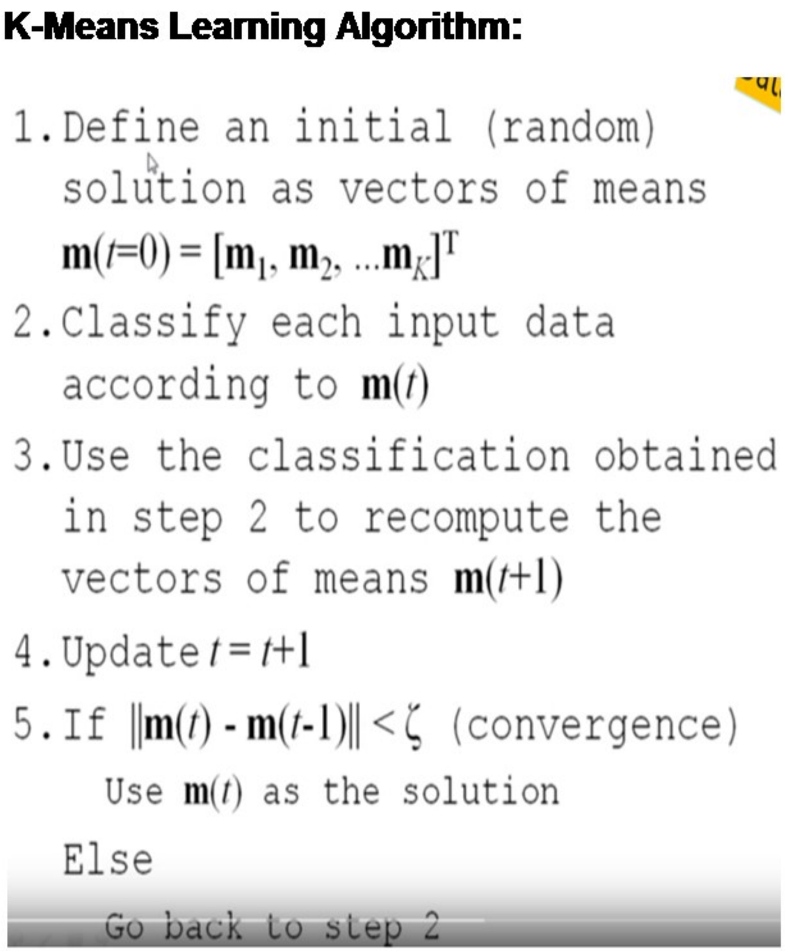
☛ ***Build homogenous groups of customers***

- Use ***automatic*** multidimensional segmentations
- DM : two techniques :
  - Neural clustering
  - Demographic clustering
- Analyse segments profiles

Campaign Management – Methods

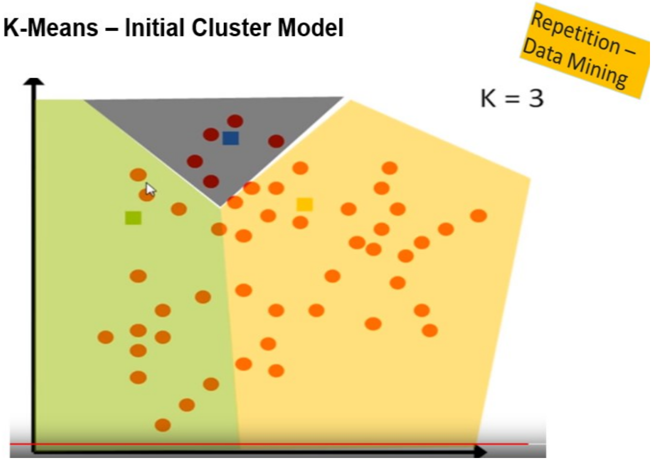
- ***Choose the interesting segments***
- ***Start the Campaign on a sample of people - adapt message to profile***
- ***Analyse deeply the campaign results***
  - Build a model to explain why some replied and some did not
    - Use a scoring method
      - IM RBF Prediction
      - IM Neural Prediction
      - IM Tree/Neural Classification

Data Mining Method: K-Means-Clustering Algorithm

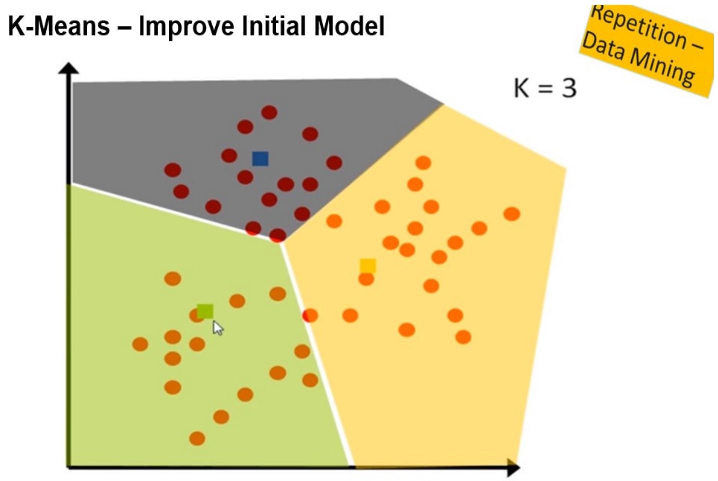


K-Means Learning Algorithm:

```
1. Define an initial (random)
   solution as vectors of means
   $\mathbf{m}(t=0) = [\mathbf{m}_1, \mathbf{m}_2, ...\mathbf{m}_K]^T$
2. Classify each input data
   according to $\mathbf{m}(t)$
3. Use the classification obtained
   in step 2 to recompute the
   vectors of means $\mathbf{m}(t+1)$
4. Update $t = t+1$
5. If $\|\mathbf{m}(t) - \mathbf{m}(t-1)\| < \zeta$ (convergence)
     Use $\mathbf{m}(t)$ as the solution
   Else
     Go back to step 2
```

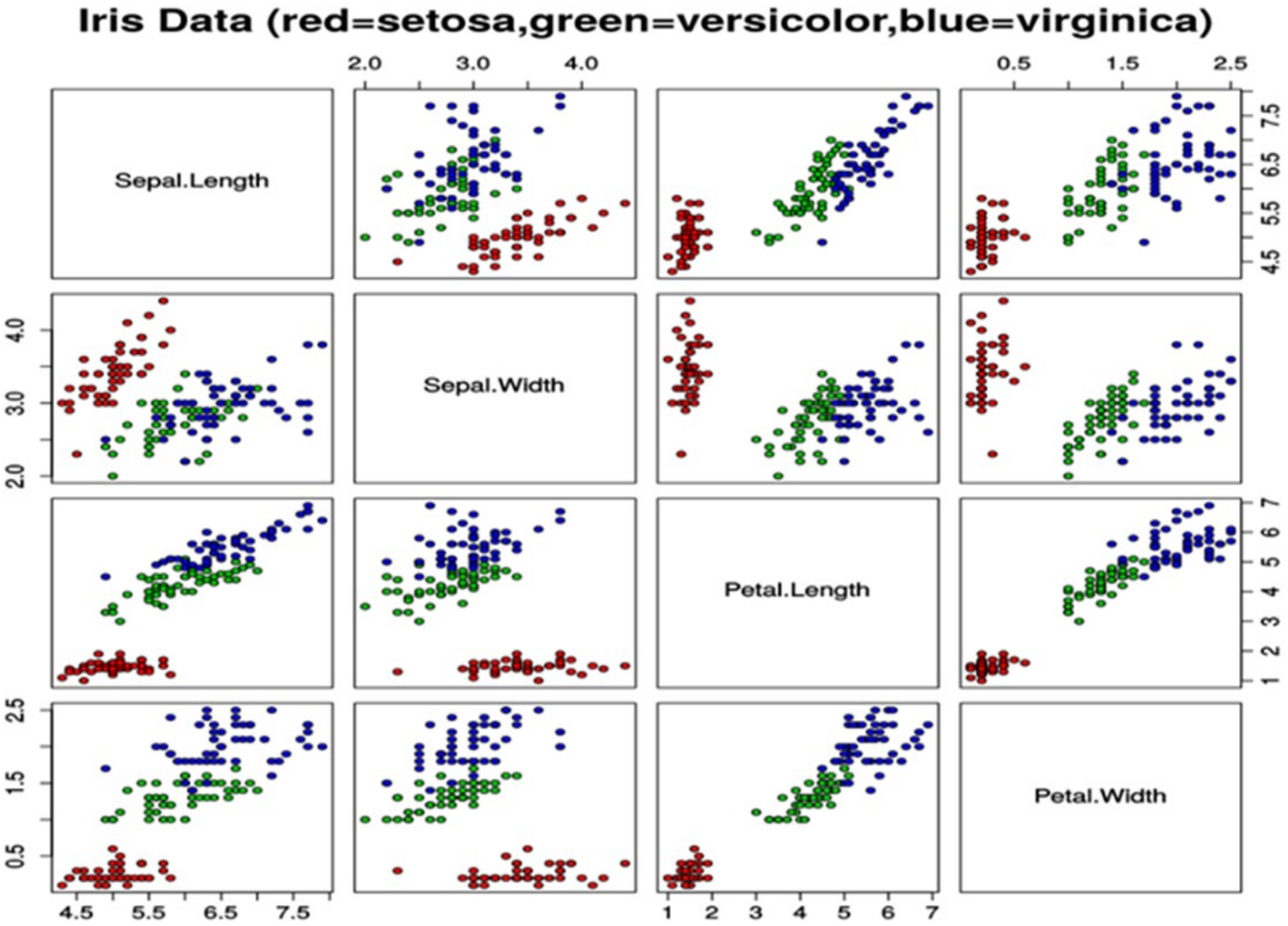


K-Means – Initial Cluster Model



K-Means – Improve Initial Model

Clustering Ex. & K-Means Clusters of IRIS Dataset *



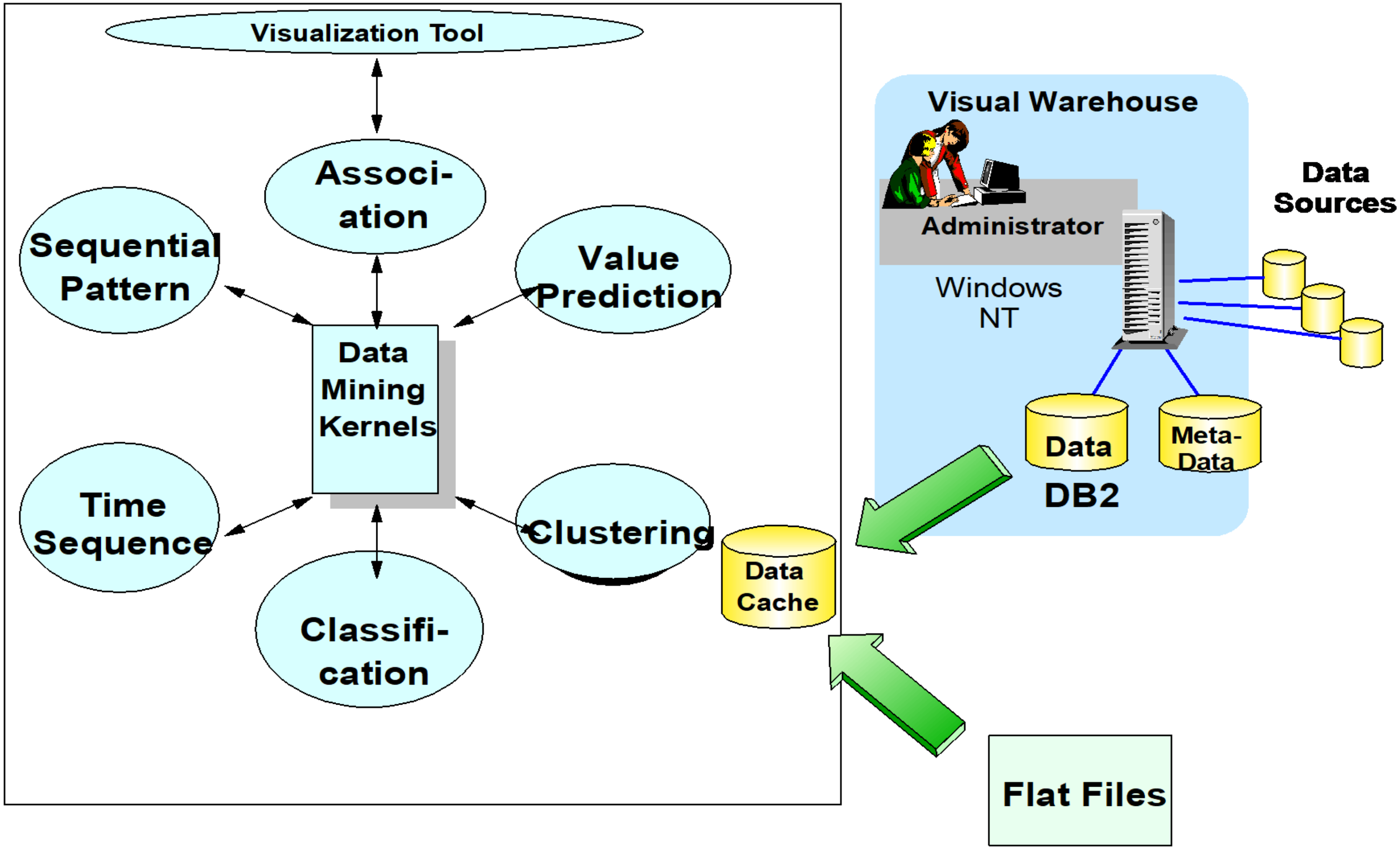
Iris Setosa

Iris Versicolor

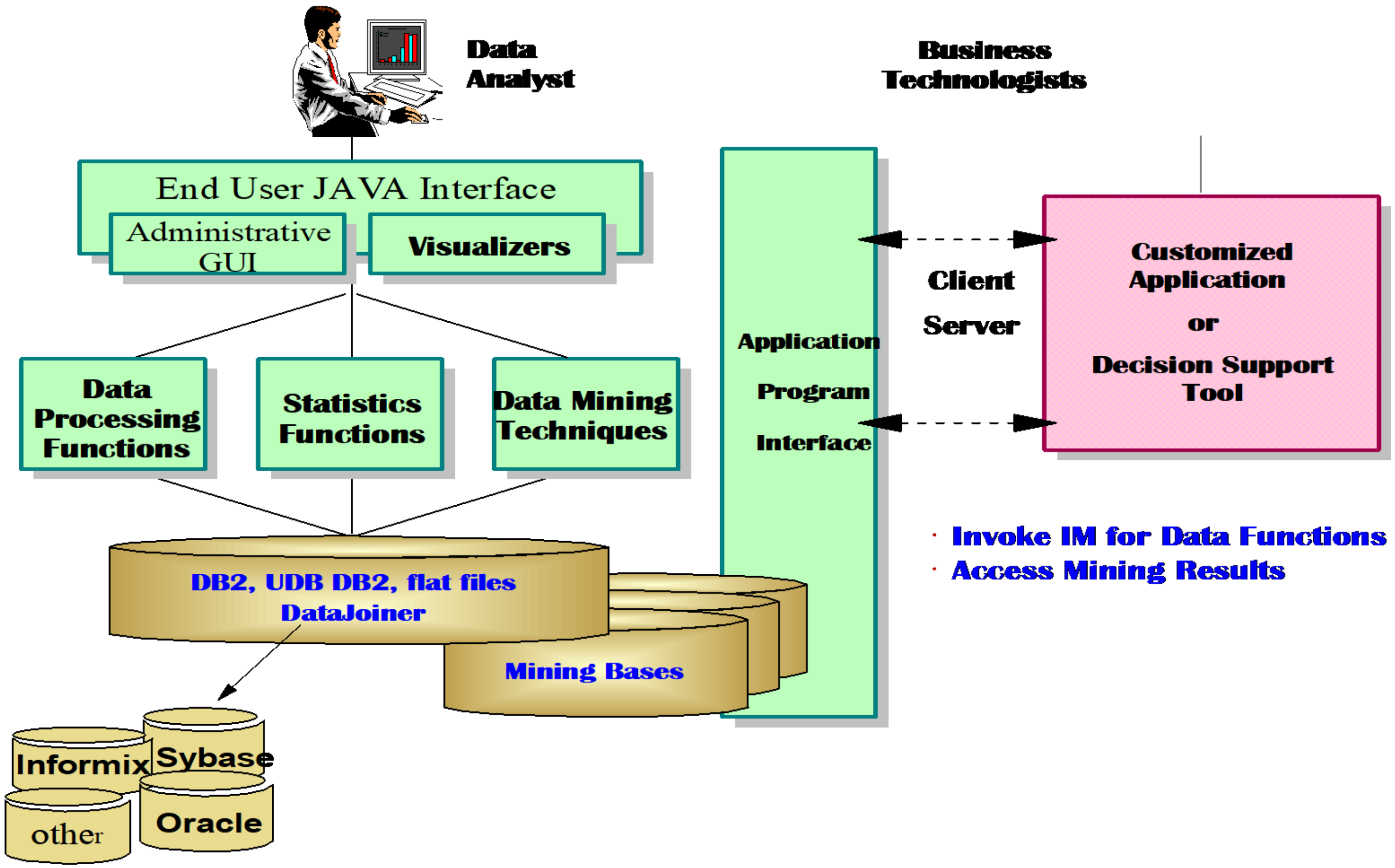
Iris Virginica

*: In a Seminar Work paper we investigate this in more detail.

IM for Data - Overview



IM for Data – Tool Architecture



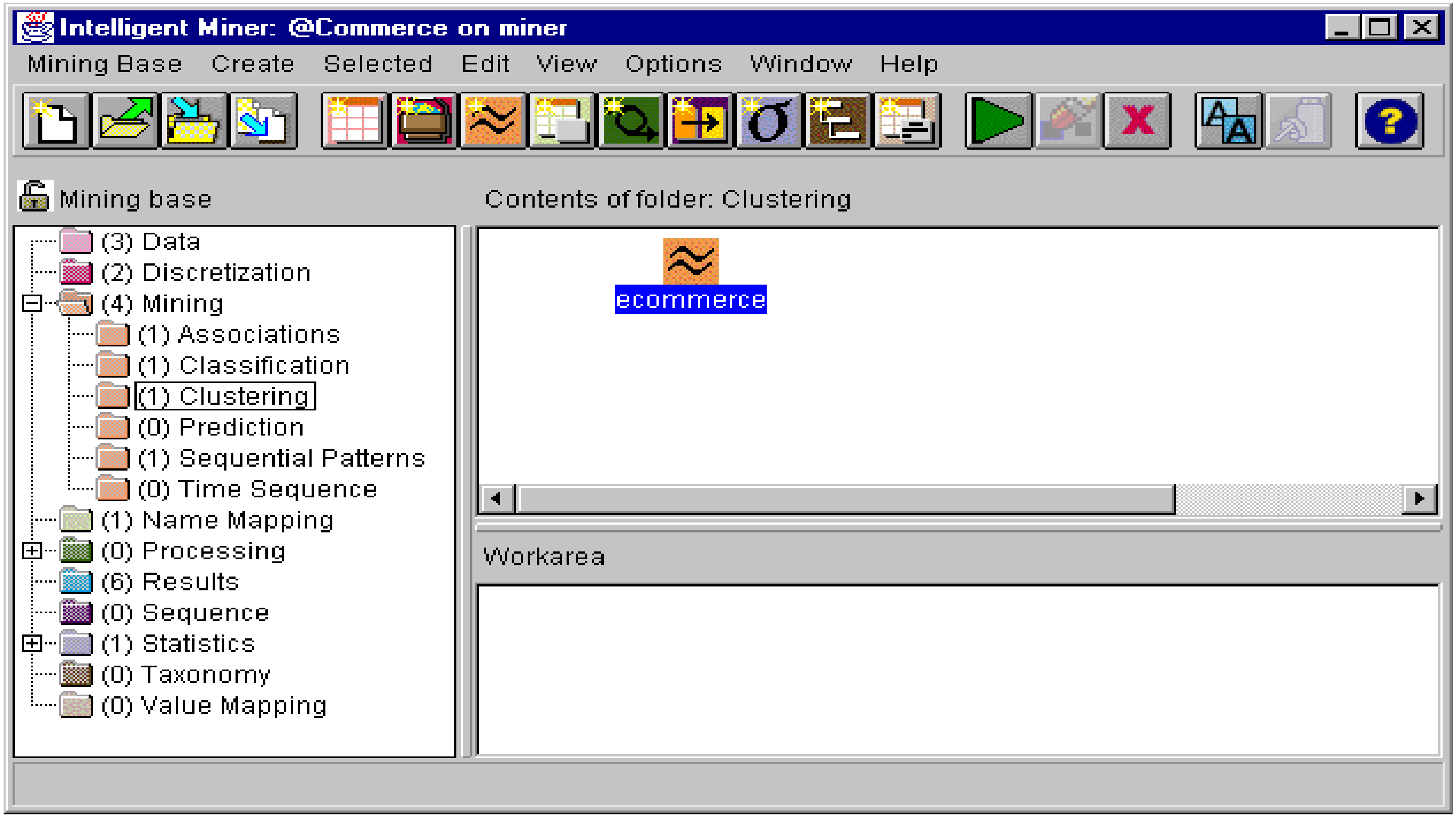
- Invoke IM for Data Functions
- Access Mining Results

IBM IM for Data - Life Demo Overview

The demo will demonstrate the five phases of data mining tasks:

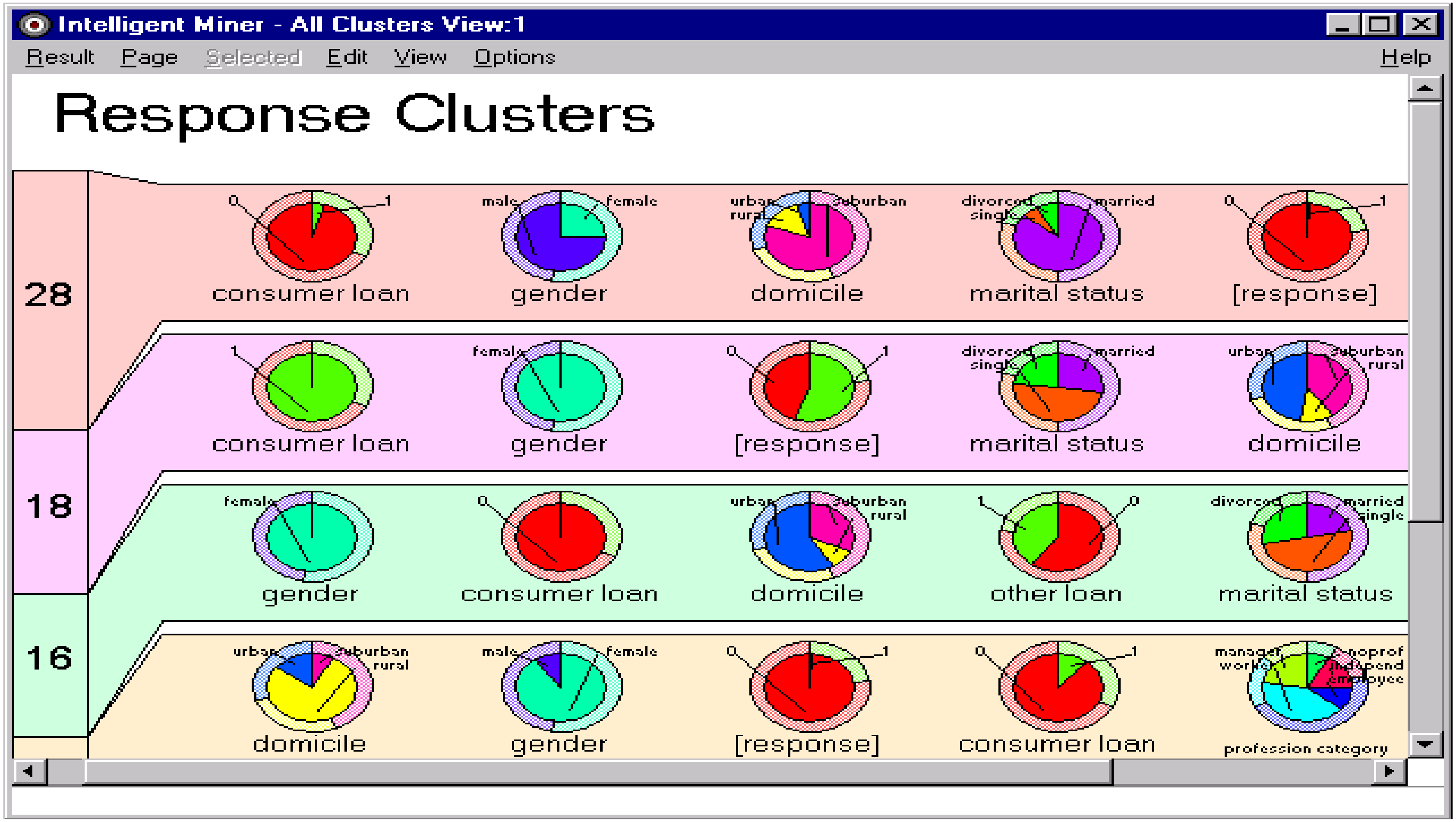
1. **Defining the data**
2. **Building the model**
3. **Applying the model**
4. **Automating the process**
5. **Analyzing the results**

IBM Intelligent Miner for Data - Life Demo



- (3) Data
- (2) Discretization
- (4) Mining
  - (1) Associations
  - (1) Classification
  - (1) Clustering
  - (0) Prediction
  - (1) Sequential Patterns
  - (0) Time Sequence
- (1) Name Mapping
- (0) Processing
- (6) Results
- (0) Sequence
- (1) Statistics
- (0) Taxonomy
- (0) Value Mapping

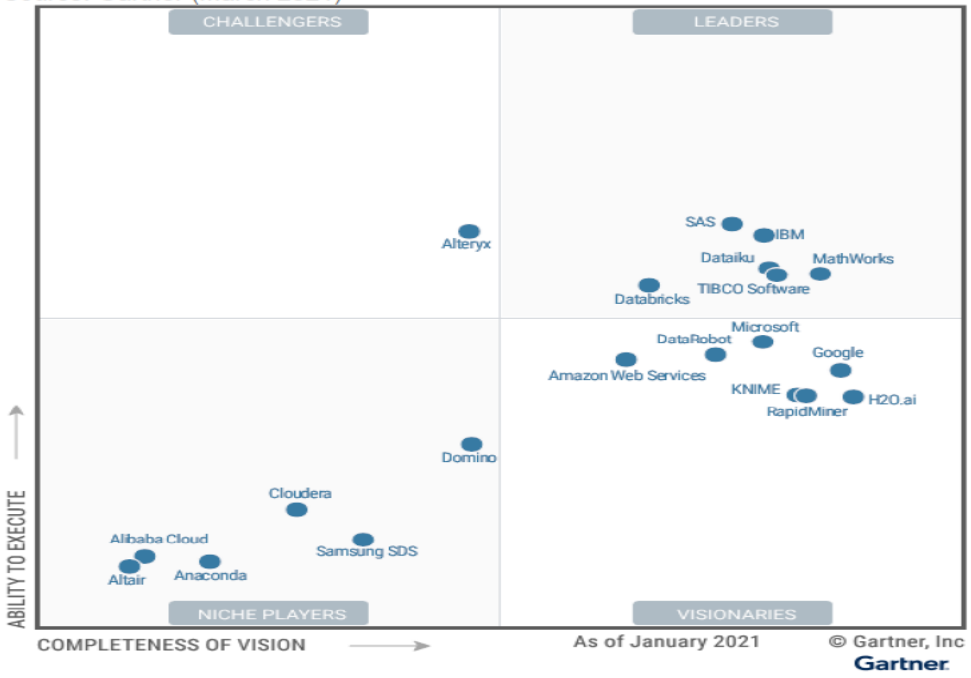
IBM Intelligent Miner for Data - Life Demo 2



Exercise1 to Lesson 10: Data Science & Machine Learning Platforms (i.e. Data Mining Tools)

<u>**Exercise E10.1 (SW*):**</u> Search for the actual “Gartner Quadrant” of DS/ML (DM) tools. Give detail descriptions of two of the leading tools in the quadrant: https://pages.dataiku.com/hs-fs/hubfs/gartner-mq-2021.png?width=443&name=gartner-mq-2021.png

Source: Gartner (March 2021)

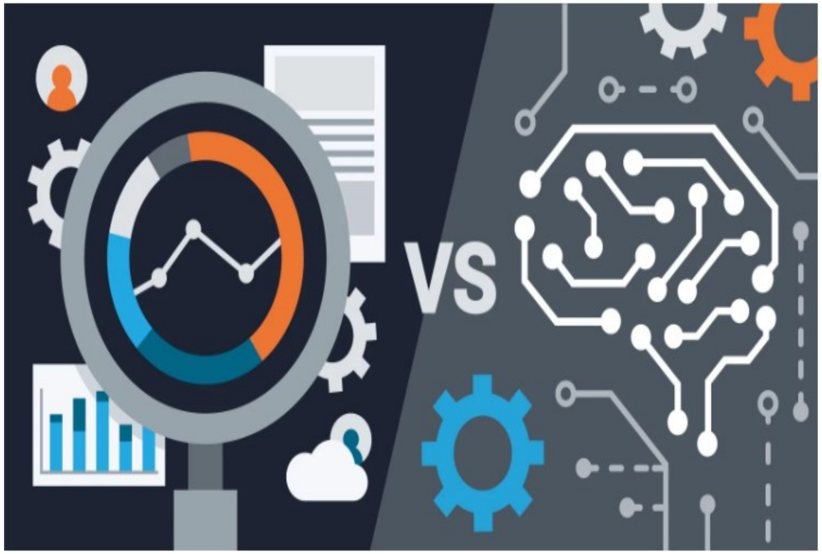


SW*: For the Seminar Work paper investigate this in more detail for three tools.

Exercise2 to Lesson 10: Advanced Analytics (AA) versus Artificial Intelligence (AI)

Exercise E10.2 (SW*): Advanced Analytics vs. Artificial Intelligence.

Look for example on the blog: https://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?

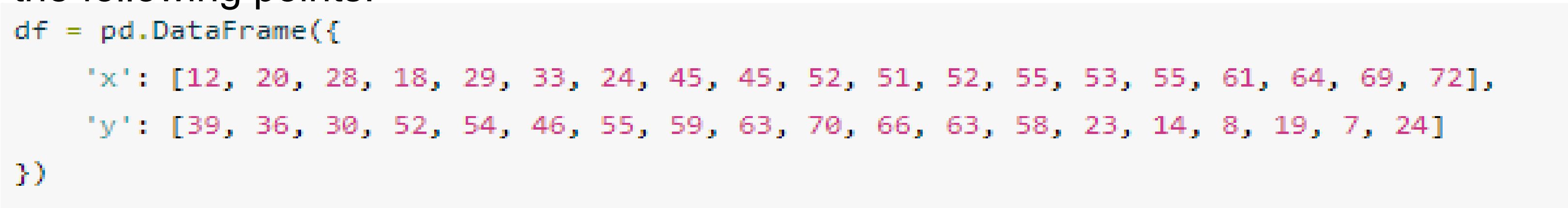


SW*: In the Seminar Work paper investigate this in more detail.

Exercise3 to Lesson 10: K-Means Clustering in Python

<u>Exercise E10.3:</u> Create a K-Means Clustering in Python

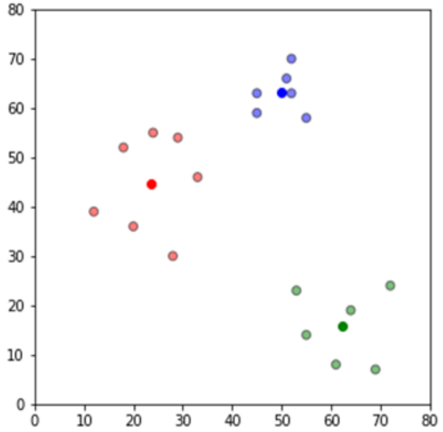
Homework for 2 Persons: Create a python algorithm (in Jupyter Notebook) which clusters the following points:



```
df = pd.DataFrame({
    'x': [12, 20, 28, 18, 29, 33, 24, 45, 45, 52, 51, 52, 55, 53, 55, 61, 64, 69, 72],
    'y': [39, 36, 30, 52, 54, 46, 55, 59, 63, 70, 66, 63, 58, 23, 14, 8, 19, 7, 24]
})
```

Following the description of: https://benalexkeen.com/k-means-clustering-in-python/ to come to 3 clear clusters with 3 means at the centre of these clusters:

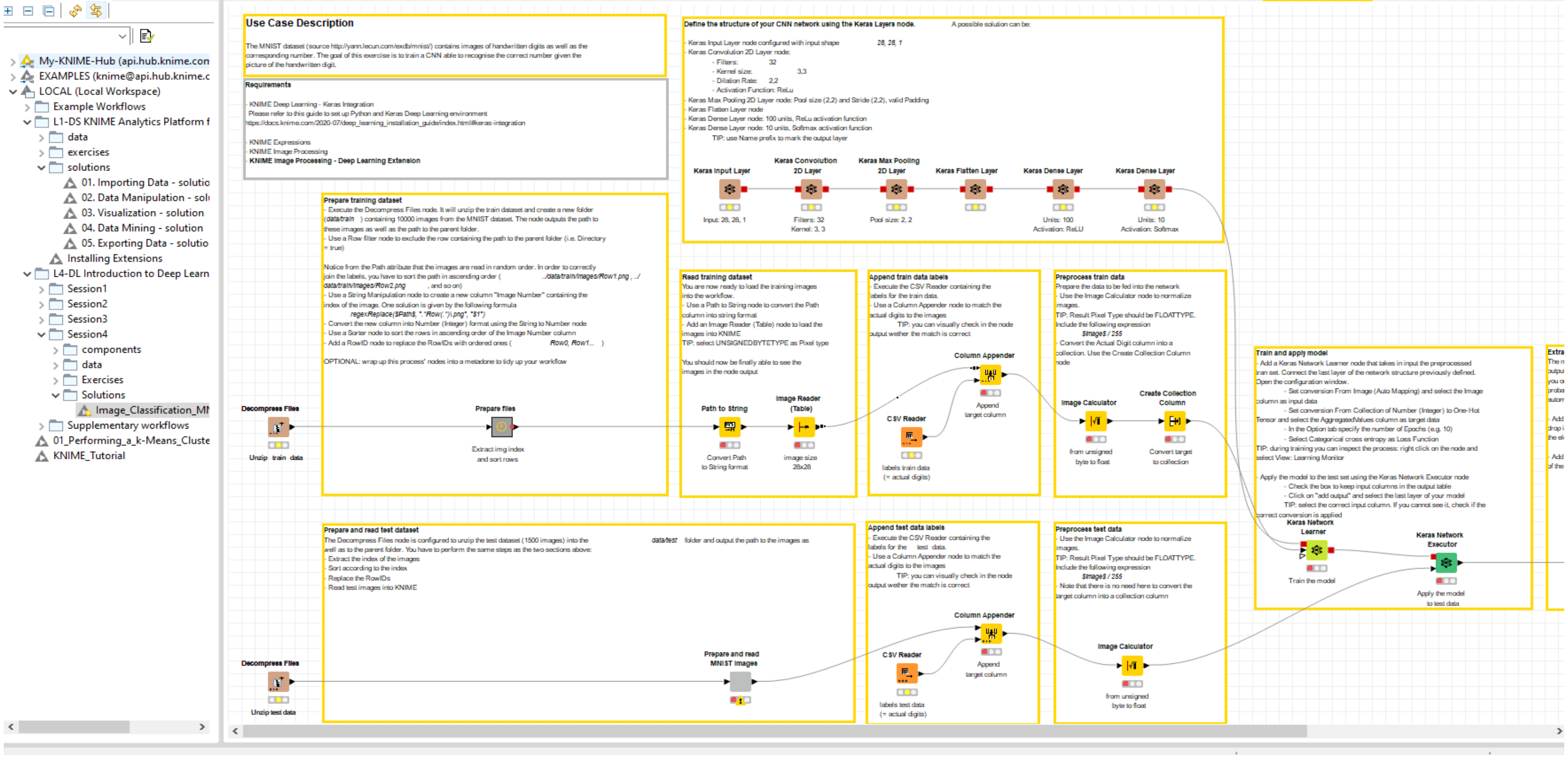
We'll do this manually first (1 person), then show how it's done using scikit-learn (1 person)



Exercise 4 to Lesson 10: KNIME Image-Classification

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

Homework for 2 Persons: Rebuild the KNIME Workflow (use given solution) for Image-Classification and give technical explanations to the solution steps (see image below):



My-KNIME-Hub (api.hub.knime.con
EXAMPLES (knime@api.hub.knime.c
LOCAL (Local Workspace)
Example Workflows
L1-DS KNIME Analytics Platform f
data
exercises
solutions
01. Importing Data - solutio
02. Data Manipulation - sol
03. Visualization - solution
04. Data Mining - solution
05. Exporting Data - solutio
Installing Extensions
L4-DL Introduction to Deep Learn
Session1
Session2
Session3
Session4
components
data
Exercises
Solutions
Image_Classification_MN
Supplementary workflows
01_Performing_a_k-Means_Cluste
KNIME_Tutorial

Use Case Description

The MNIST dataset (source http://yann.lecun.com/exdb/mnist/) contains images of handwritten digits as well as the corresponding number. The goal of this exercise is to train a CNN able to recognise the correct number given the picture of the handwritten digit.

Requirements

- KNIME Deep Learning - Keras Integration
  Please refer to this guide to set up Python and Keras Deep Learning environment
  https://docs.knime.com/2020-07/deep_learning_installation_guide/index.html#keras-integration

- KNIME Expressions
- KNIME Image Processing
- **KNIME Image Processing - Deep Learning Extension**

Define the structure of your CNN network using the Keras Layers node. A possible solution can be:

- Keras Input Layer node configured with input shape 28, 28, 1
- Keras Convolution 2D Layer node:
  - Filters: 32
  - Kernel size: 3,3
  - Dilation Rate: 2,2
  - Activation Function: ReLu
- Keras Max Pooling 2D Layer node: Pool size (2,2) and Stride (2,2), valid Padding
- Keras Flatten Layer node
- Keras Dense Layer node: 100 units, ReLu activation function
- Keras Dense Layer node: 10 units, Softmax activation function
  TIP: use Name prefix to mark the output layer

Prepare training dataset

- Execute the Decompress Files node. It will unzip the train dataset and create a new folder (*data/train*) containing 10000 images from the MNIST dataset. The node outputs the path to these images as well as the path to the parent folder.
- Use a Row filter node to exclude the row containing the path to the parent folder (i.e. Directory = true)

Notice from the Path attribute that the images are read in random order. In order to correctly join the labels, you have to sort the path in ascending order (*../data/train/images/Row1.png* , *../data/train/images/Row2.png* , and so on)
- Use a String Manipulation node to create a new column "Image Number" containing the index of the image. One solution is given by the following formula
  *regexReplace($Path$, ".*Row(.*)\.png", "$1")*
- Convert the new column into Number (Integer) format using the String to Number node
- Use a Sorter node to sort the rows in ascending order of the Image Number column
- Add a RowID node to replace the RowIDs with ordered ones (*Row0, Row1...*)

OPTIONAL: wrap up this process' nodes into a metadone to tidy up your workflow

Read training dataset

You are now ready to load the training images into the workflow.
- Use a Path to String node to convert the Path column into string format
- Add an Image Reader (Table) node to load the images into KNIME
TIP: select UNSIGNEDBYTETYPE as Pixel type

You should now be finally able to see the images in the node output

Append train data labels

- Execute the CSV Reader containing the labels for the train data.
- Use a Column Appender node to match the actual digits to the images
  TIP: you can visually check in the node output wether the match is correct

Preprocess train data

Prepare the data to be fed into the network
- Use the Image Calculator node to normalize images.
TIP: Result Pixel Type should be FLOATTYPE. Include the following expression
  $Image$ / 255
- Convert the Actual Digit column into a collection. Use the Create Collection Column node

Train and apply model

- Add a Keras Network Learner node that takes in input the preprocessed tran set. Connect the last layer of the network structure previously defined. Open the configuration window.
  - Set conversion From Image (Auto Mapping) and select the Image column as input data
  - Set conversion From Collection of Number (Integer) to One-Hot Tensor and select the AggregatedValues column as target data
  - In the Option tab specify the number of Epochs (e.g. 10)
  - Select Categorical cross entropy as Loss Function
TIP: during training you can inspect the process: right click on the node and select View: Learning Monitor

- Apply the model to the test set using the Keras Network Executor node
  - Check the box to keep input columns in the output table
  - Click on "add output" and select the last layer of your model
  TIP: select the correct input column. If you cannot see it, check if the correct conversion is applied

Extra
The n
outpu
you c
proba
autom

- Add
drop i
the el

- Add
of the

Prepare and read test dataset

The Decompress Files node is configured to unzip the test dataset (1500 images) into the *data/test* folder and output the path to the images as well as to the parent folder. You have to perform the same steps as the two sections above:
- Extract the index of the images
- Sort according to the index
- Replace the RowIDs
- Read test images into KNIME

Append test data labels

- Execute the CSV Reader containing the labels for the test data.
- Use a Column Appender node to match the actual digits to the images
  TIP: you can visually check in the node output wether the match is correct

Preprocess test data

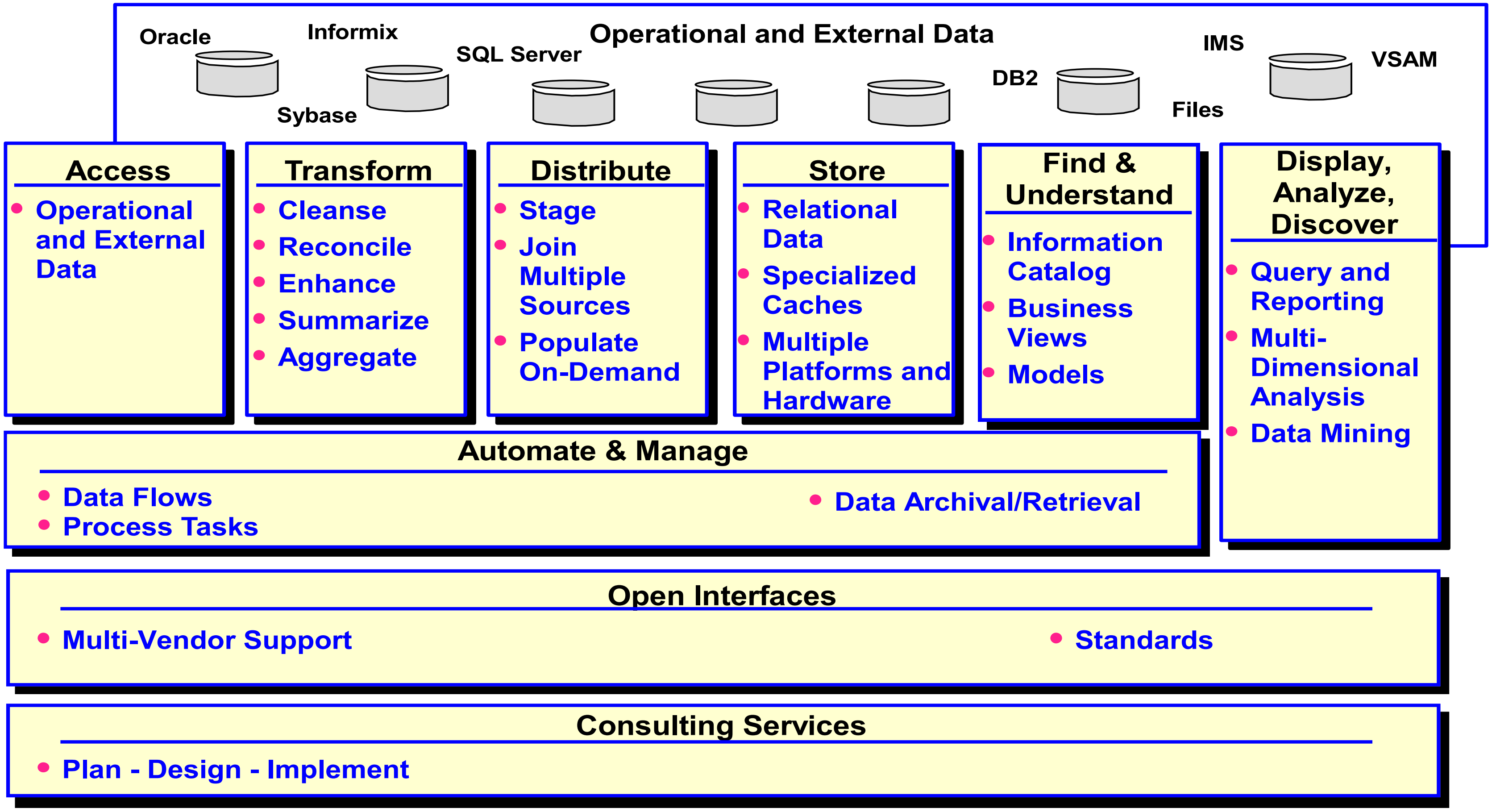
- Use the Image Calculator node to normalize images.
TIP: Result Pixel Type should be FLOATTYPE. Include the following expression
  $Image$ / 255
- Note that there is no need here to convert the target column into a collection column

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

Anhang

BACKUP
Slides

Components of a Data Warehouse



Operational and External Data

Access
- Operational and External Data

Transform
- Cleanse
- Reconcile
- Enhance
- Summarize
- Aggregate

Distribute
- Stage
- Join Multiple Sources
- Populate On-Demand

Store
- Relational Data
- Specialized Caches
- Multiple Platforms and Hardware

Find & Understand
- Information Catalog
- Business Views
- Models

Display, Analyze, Discover
- Query and Reporting
- Multi-Dimensional Analysis
- Data Mining

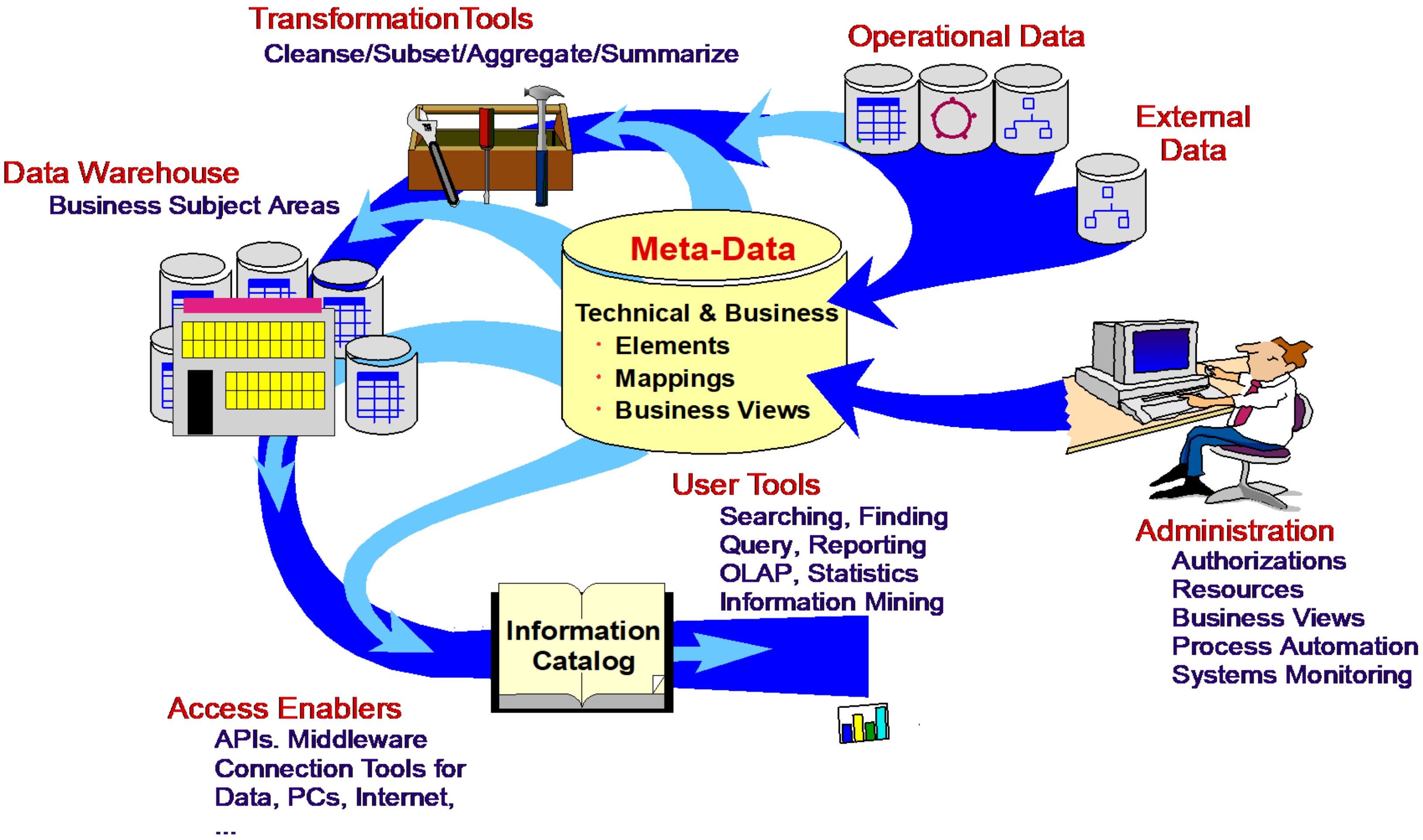
Automate & Manage
- Data Flows
- Process Tasks
- Data Archival/Retrieval

Open Interfaces
- Multi-Vendor Support
- Standards

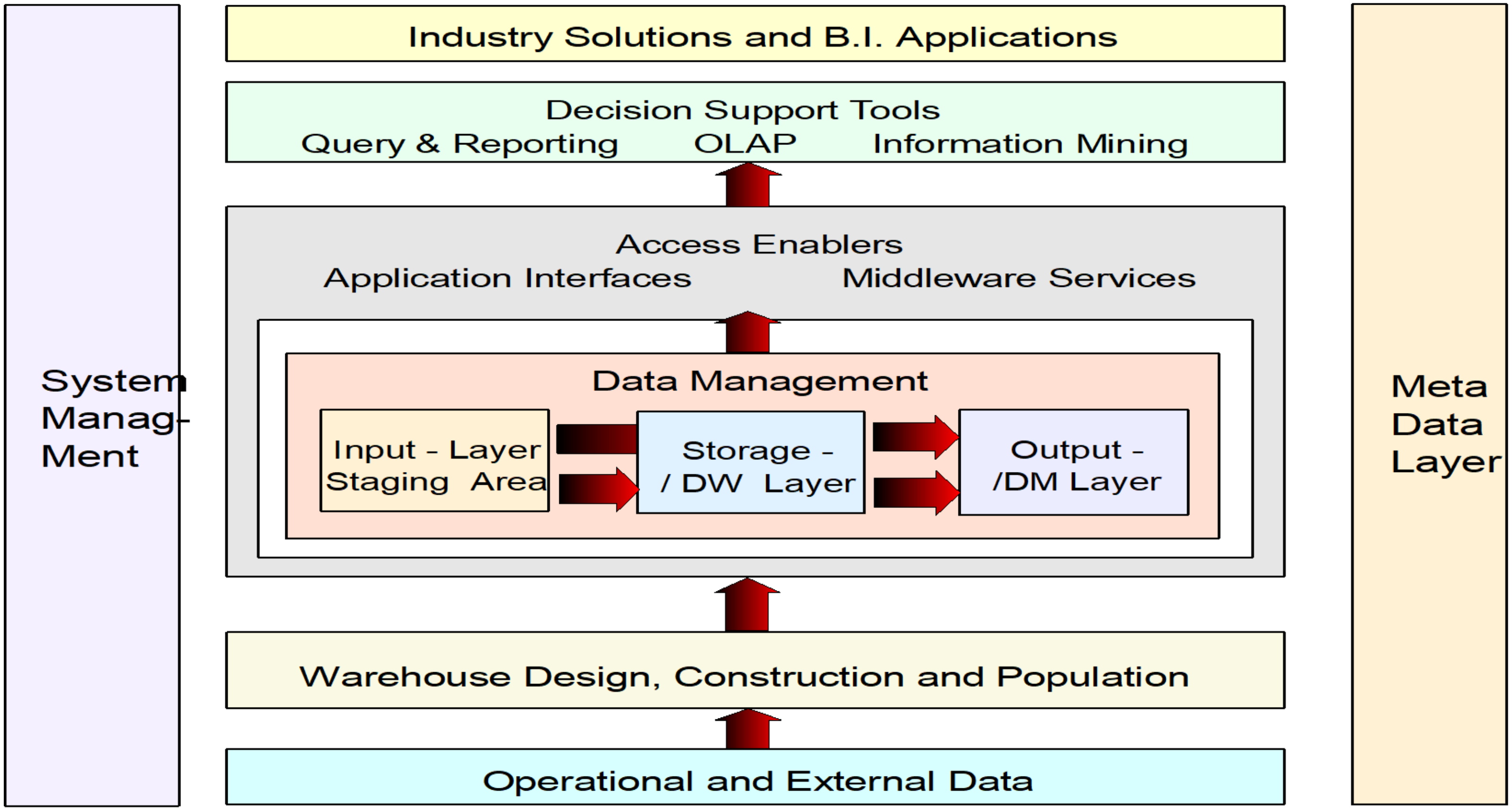
Consulting Services
- Plan - Design - Implement

Enabling the Solution

DWH Architecture – Processes



Process Layers of the DWH



DWH Lecture Categories

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

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

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

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