

Introduction to Data Warehousing (DWH)

DHBW – Fakultät Technik-Informatik, Stuttgart, Spring 2025

Dr. Hermann Völlinger, Mathematics & IT Architecture

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

www.dhbw-stuttgart.de

General Remarks to Lecture DWH (1/2)

- Our first lecture starts on Friday, 14.2.2025, 2:00pm - 5:15pm (4 lecture hours + 15 minutes break). The remaining 8 lectures will also start at 2:00 pm. Last lecture on Thursday, 17.04.2025. Link to timetable: <https://rapla.dhbw.de/rapla/calendar?user=doelker@verwaltung.ba-stuttgart.de&file=22D>
- If an online meeting need to be done, 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 (> 170 documents and technical papers in four DWH categories) you will find it in Moodle/”Supporting Information for DWH Lecture” [Kurs: T3INF4304_3_22D](#)

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 usage of ChatGPT (i.e. bwGPT) for creating program code (i.e. for Jupyter Notebook) is permitted if at least two lines of commentary are inserted per line of generated code.
- The grading of the lecture DWH is done by a Program Design-work (groupwork with two members, English, ~15-20 pages, see in my DHBW-Homepage the list of topics, deadline 26.04.2025). It is recommended to think about the topics already during the lecture period. The grade is part of the DHBW Bachelor certificate.

Informatik		Kurs D	
Stunden- und Prüfungsplan Jahrgang 2022		Anteil	Termin
Modulname	Prüfung		
Software Engineering II	Programmwurf		
Advanced SWE		100% Programmwurf	04.04.25
Big Data Architectures	Kombinierte Prüfung		
IT Architekturen		50% Referat	5. Sem.
Verteilte Systeme		50% Seminararbeit	5. Sem.
IT Sicherheit	Klausur (90 Min)		
IT Sicherheit		100% Klausur (90 Min)	5. Sem.
Datenbanken II	Kombinierte Prüfung		
Aktuelle DB Architekturen und Technologien		30% Seminararbeit	28.03.25
Data Warehouse		20% Referat	vorlesungsintegriert
Data Science	Kombinierte Prüfung		
Grundlagen Data Science		50% Programmwurf	26.04.25
Semantic Web		50% Programmwurf	5. Sem.
Künstliche Intelligenz und Maschinelles Lernen	Kombinierte Prüfung		
Grundlagen der Künstlichen Intelligenz		30% Hausarbeit	5. Sem.
Maschinelles Lernen		20% Referat	5. Sem.
Mensch Maschine Interaktion	Kombinierte Prüfung		
Interaktive Systeme		30% Programmwurf	5. Sem.
(Barrierefreiheit)		20% Seminararbeit	5. Sem.
Integrationsseminar Digitalisierung		50% Referat	16.04.25
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 Program Design papers in the area of Data Warehouse / Business Intelligence

1. **Team Size/Effort/Pages:** group work (2 members); ~10-12 hours/~15-20 pages
2. **Language/Deadline/Details:** English/26.04.2025/ [Examination – Program-Design](#)
3. **Evaluation:** Dr. Hermann Völlinger (send to Moodle-TINF22D: [Kurs: T3INF4304_3_22D](#))
4. **References:** As a source of information and further references to the respective themes DWxx see the instructions in the lecture and look on the papers in Moodle: [Kurs: T3INF4304_3_22D](#)

No	Topic	Details	Students/Points (max=100)
DW-PD01	KG tool - Healthcare and Life Sciences (ex. E1.4b)	Give a program design and detailed description of the creation of a KG tool in the area of Healthcare and Life Sciences (Modeling biological pathways, drug discovery, and patient data...). See also exercise E1.4b	- / -: x
DW-PD02	First Experiences with KNIME Analytics (ex. E1.5)	First Experiences with KNIME Analytics Platform: Install the tool and report about your first experiences and insights. 1. What can be done with the tool? 2. Features for Data-Management (DM)? PD**: Show a concrete simple KNIME DM example/program? 3. Features for Analytics and Data Science (DC)? PD**: Show a concrete simple KNIME DC example/program? Information source is the KNIME Homepage KNIME Open for Innovation and the three mentioned documents in the lesson DW01 (see lesson notes). See also exercise E1.5	- / -: x
DW-PD03	Data Manipulation & Aggregation in KNIME	Rebuild the KNIME Workflow (use given solution) for Data Manipulation & Aggregation and give technical explanations	: x

Content: Introduction to Data Warehousing (DWH)

Goal: Introduction, Architecture and Basic Concepts

1. *DW01 - Introduction to DWH & Business Intelligence (BI) (Friday, 14.02.25)*
2. *DW02 - DWH Architecture (Virtual, 1-Tier, 2-Tier), Advantages & Disadvantages (Fri., 28.02.25)*
3. *DW03 - Overview about DBMS (i.e. Relational Databases) (Fri., 07.03.25)*
4. *DW04 - Introduction to Basics of SQL & Examples (Fri., 07.03.25)*
5. *DW05 – Multi-Dimensional Data Modeling (MDDM), (Fri., 14.03.25)*
6. *DW06 - ETL – Reference Architecture (Introduction) (Fri., 21.03.25)*
7. *DW07 - ETL – Data Population Techniques, Tool Examples (Fri., 28.03.25)*
8. *DW08 – Descriptive Analysis: relational OLAP & multidimensional OLAP Structures (Thursday, 03.04.25)*
9. *DW09 - Advanced Analysis I: Data Mining: Introduction and First Methods (Thu., 10.04.25)*
10. *DW10 – Ad. Analysis II: DM Methods and Tool Examples (Thu., 17.04.25)*

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; SS2025; <http://www.dhbw-stuttgart.de/~hvoellin/>
17. [HVö-2]: *Hermann Völlinger and Other: Exercises & Solutions of the Lecture 'Introduction to Data Warehousing'*; DHBW Stuttgart; SS2025 <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: Skript "Math+ MaschLernen(ML)" im Github HVoellinger* https://github.com/HVoellinger/Mathematische-Grundlagen_von_ML
21. [DHBW-Moodle]: *DHBW-Moodle for TINF22D: 'Directory of supporting Information for the DWH Lecture'*; [Kurs: T3INF4304_3_22D](#): More than 170 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: 68 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: 28 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: 33 Papers/Documents
- *New technology for most of the students.*

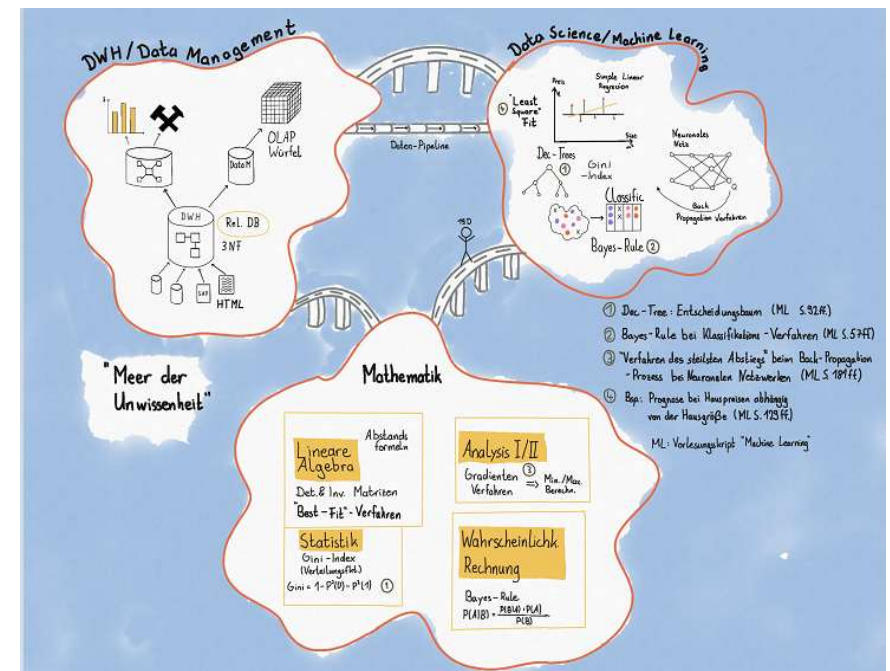
4. Category 4: Descriptive – & Advanced Analytics

- Lessons: DW08 – DW10
- Literature: 8, 13 -16, 19-22; in Moodle: 52 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) ; see: [HVö-5] .



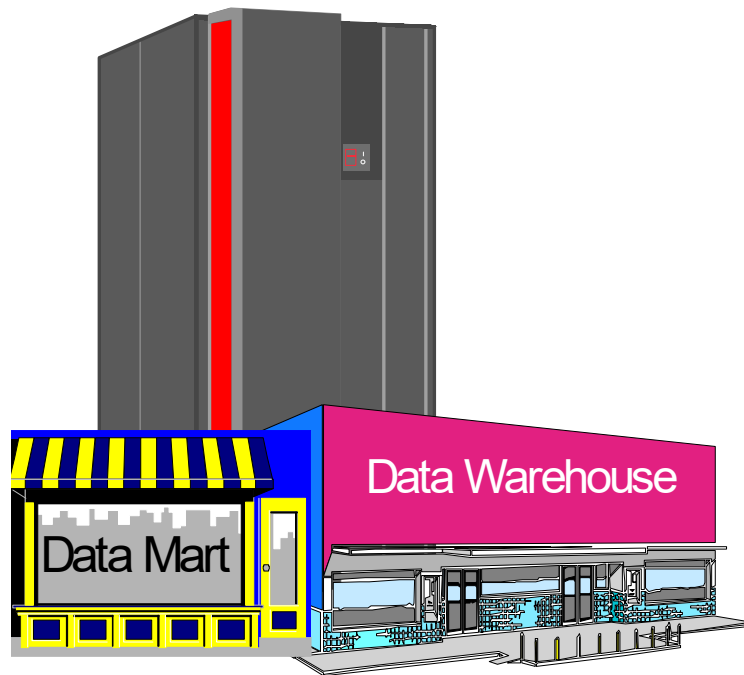
Category 1: Introduction & Architecture of DWH

Category 2: Databases and Data Modeling

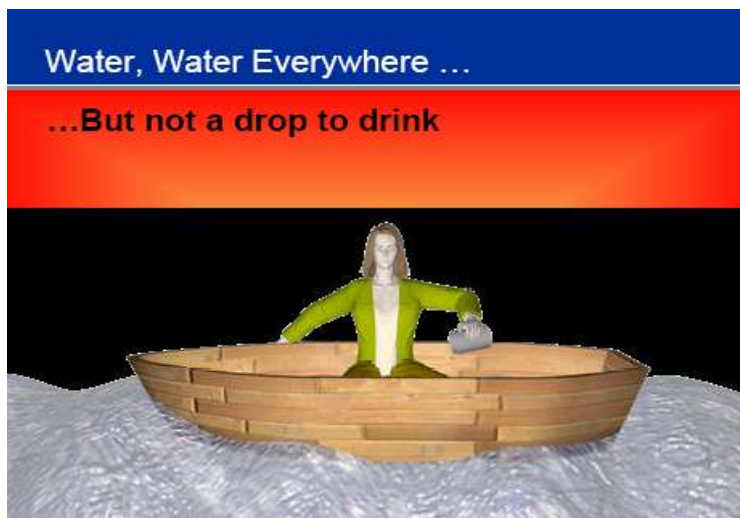
Category 3: ETL: Architecture & Technology

Category 4: Descriptive – & Advanced Analytics

DW01- Introduction to DWH & BI

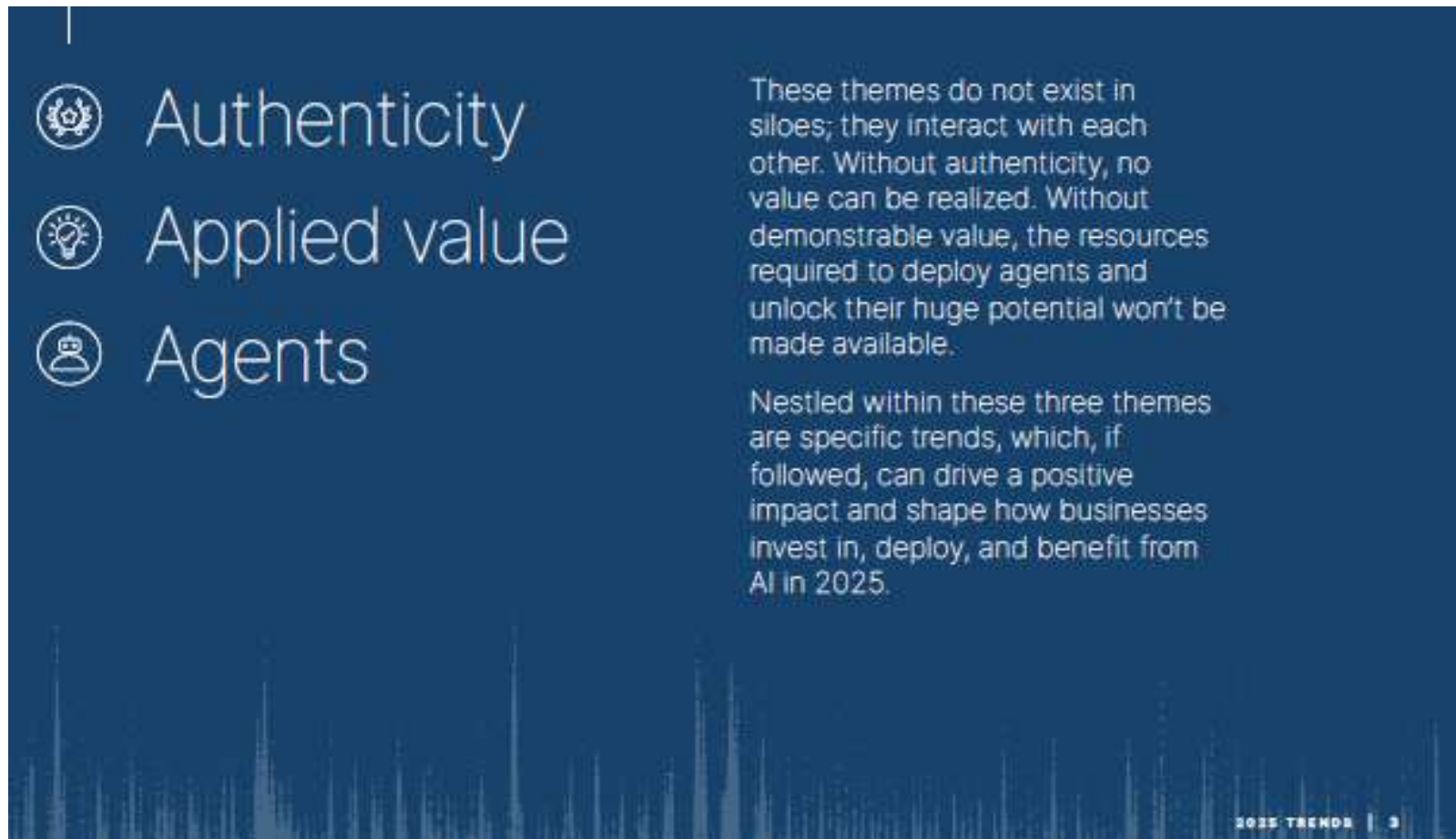


Motivation - What is Business Intelligence (BI) – the Problem



Qlik - Data Trends to Support AI

See paper [Qlik-BI_Data_Trends_2025](#) in [DHBW-Moodle]



The slide features a dark blue background with a subtle pattern of vertical lines at the bottom. On the left, three circular icons are arranged vertically: a gear for 'Authenticity', a lightbulb for 'Applied value', and a robot head for 'Agents'. To the right of these icons, there is a block of text explaining the interconnected nature of these themes and their impact on AI in 2025.

- Authenticity
- Applied value
- Agents

These themes do not exist in siloes; they interact with each other. Without authenticity, no value can be realized. Without demonstrable value, the resources required to deploy agents and unlock their huge potential won't be made available.

Nestled within these three themes are specific trends, which, if followed, can drive a positive impact and shape how businesses invest in, deploy, and benefit from AI in 2025.

2025 TRENDS | 3

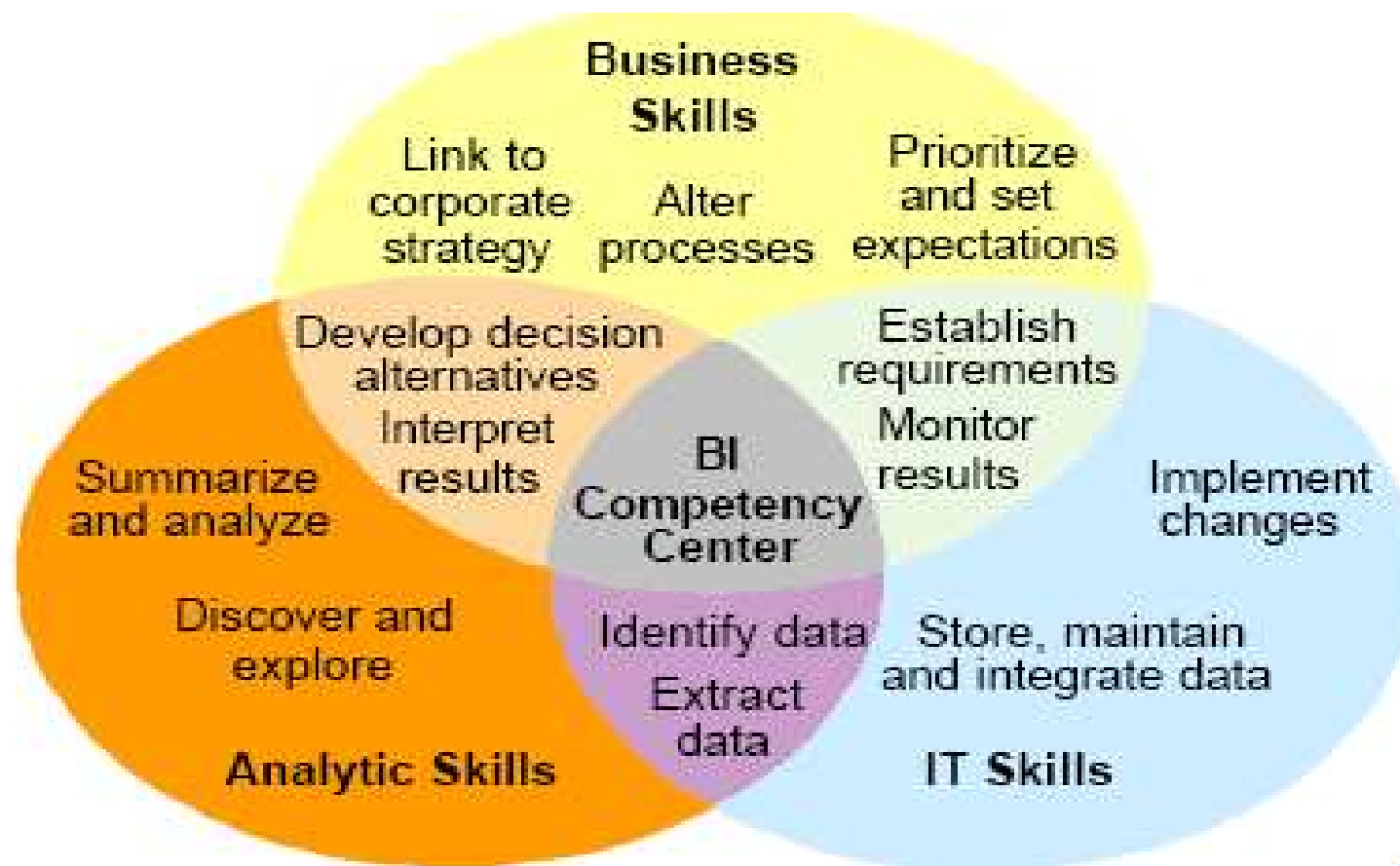
First Definition: What is BI & BI Mission



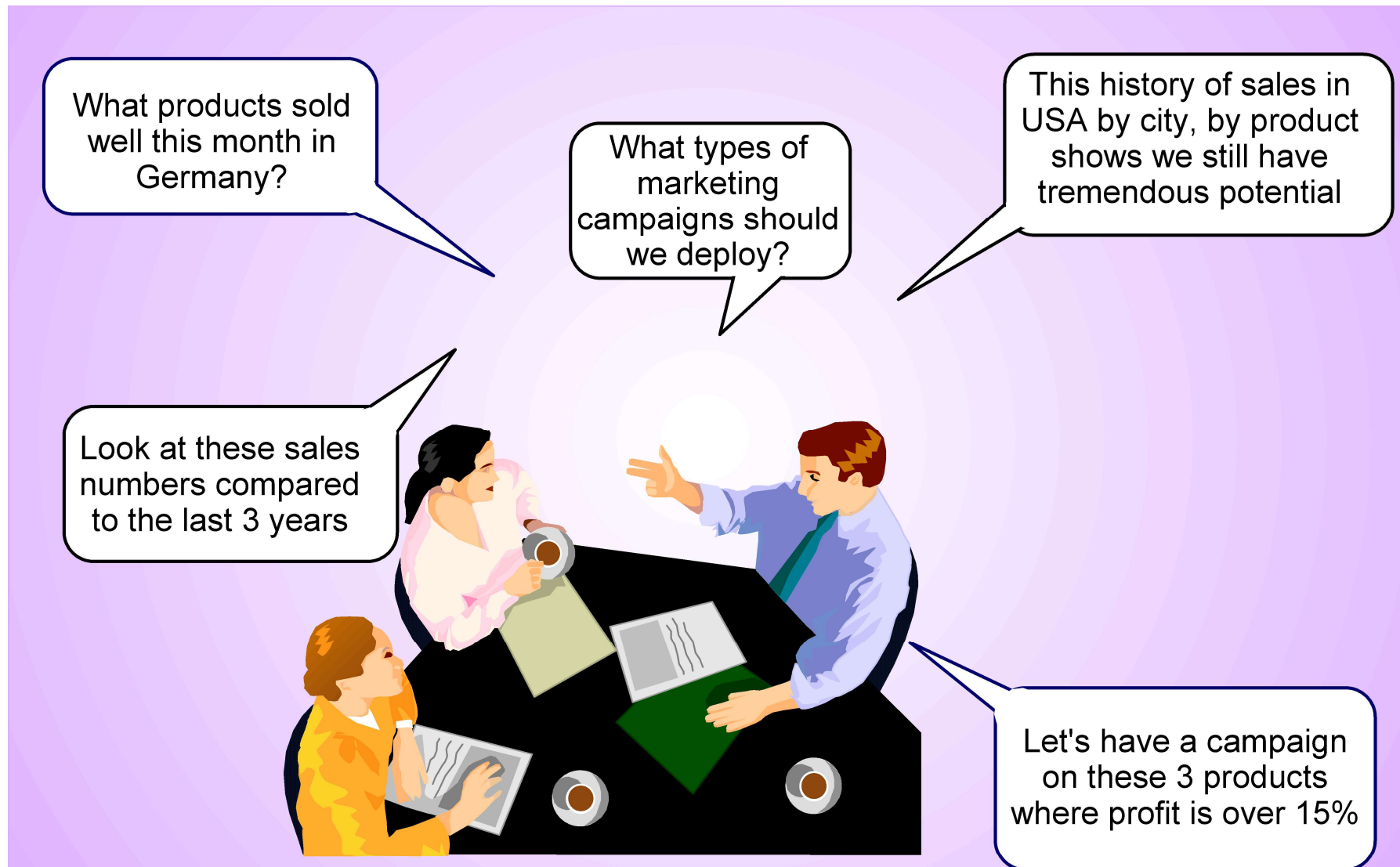
What is BI – the three Pillars of BI



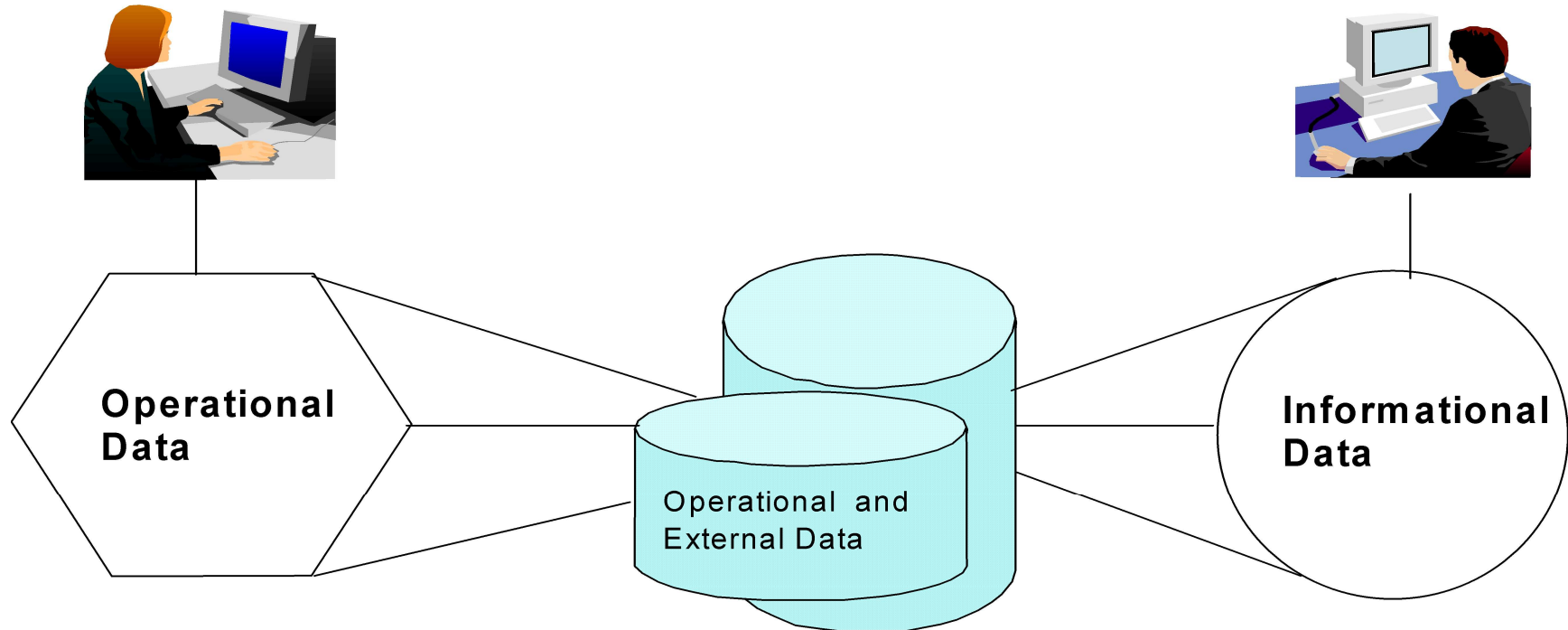
What is BI – the BI Competency Center



BI - Getting the Answers you Need



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 a metadata catalogue**

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

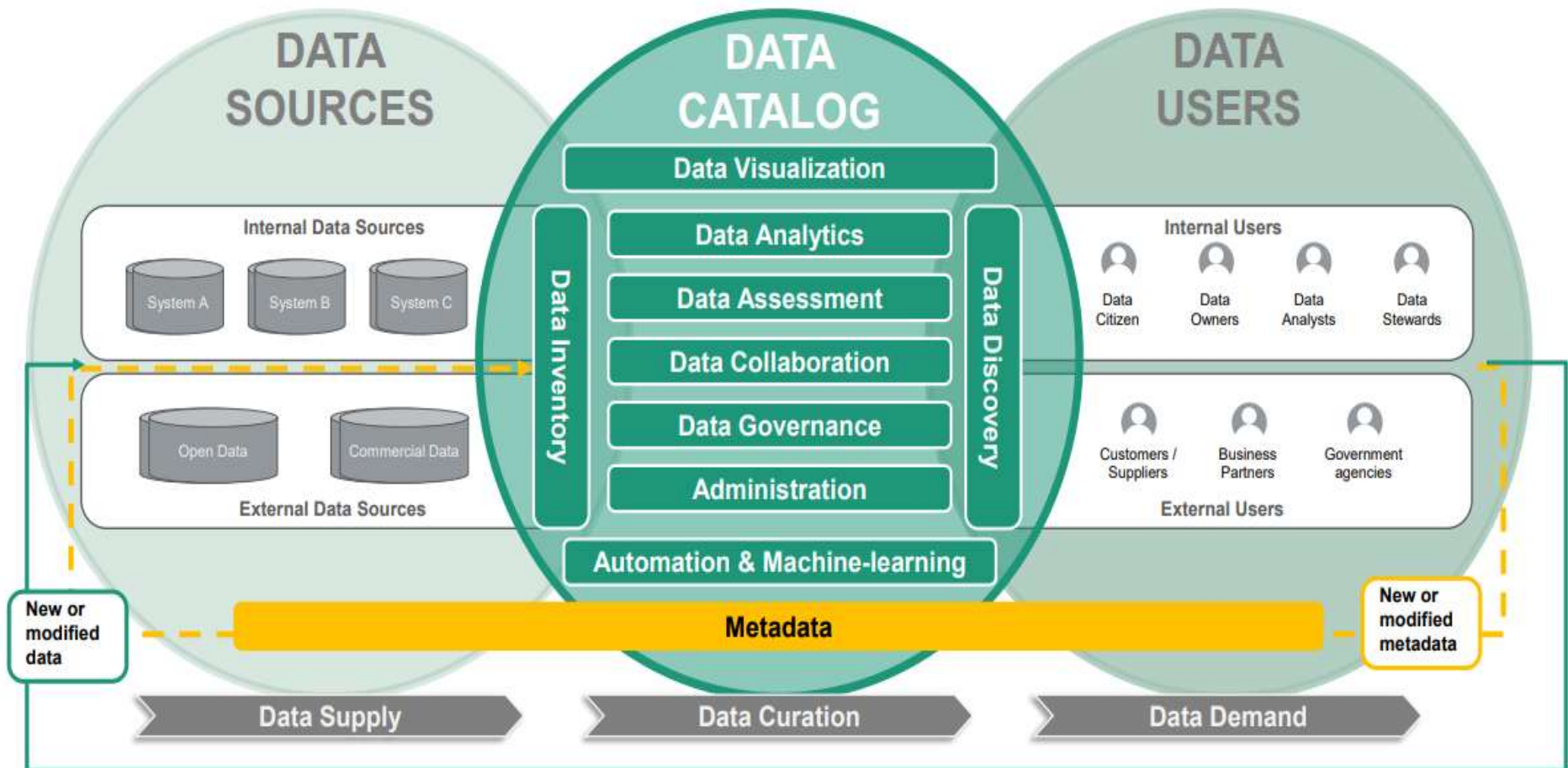
The need for an Information Catalog (Metadata)

- Finding & Understanding the Data



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 Korte¹, Martin Fädler², Markus Spiekermann¹, Prof. Christine Legner², Prof. Boris Otto¹

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)

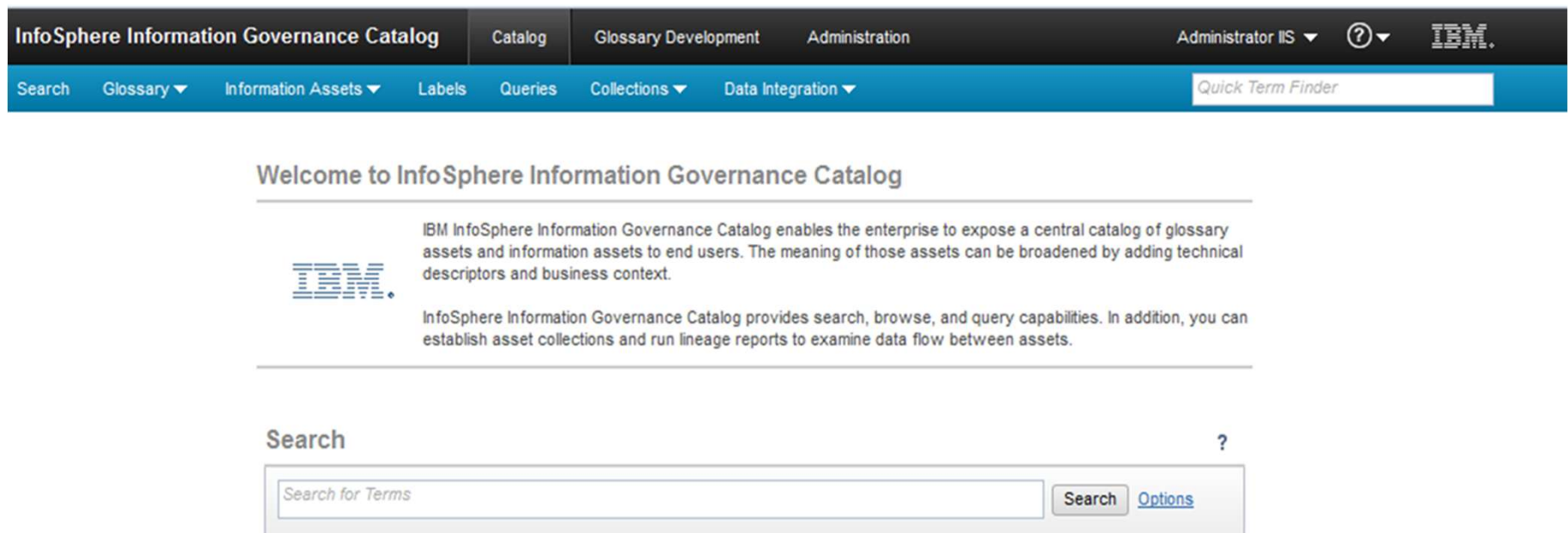
Tobias Korte¹, Martin Fädler², Markus Spiekermann¹,
Prof. Christine Legner², Prof. Boris Otto¹

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

Fraunhofer ISST CDQ

Demo: IBM Information Governance Catalog (IGC)

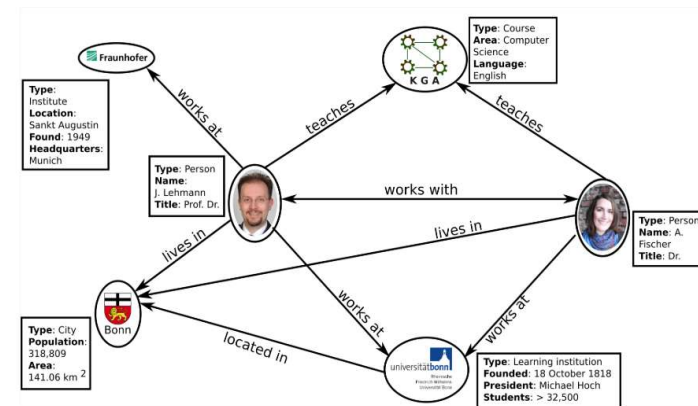
- Allows you to understand where information came from and where it is used
- A key enabler to regulatory compliance and the IBM Data Governance Maturity Model
- Cross-tool reporting on:
 - Data movement and lineage
 - Business meaning
 - Impact of changes
 - Dependencies
 - Data lineage for BI Reports



The screenshot displays the IBM InfoSphere Information Governance Catalog web application. The top navigation bar includes links for 'InfoSphere Information Governance Catalog', 'Catalog', 'Glossary Development', and 'Administration'. On the right, it shows the user role 'Administrator IIS' and the IBM logo. Below this, a secondary navigation bar contains links for 'Search', 'Glossary', 'Information Assets', 'Labels', 'Queries', 'Collections', and 'Data Integration', along with a 'Quick Term Finder' search box. The main content area features a 'Welcome to InfoSphere Information Governance Catalog' message, followed by an IBM logo and a paragraph explaining the catalog's purpose: '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.' Below this, another paragraph states: '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.' At the bottom, there is a 'Search' section with a search input field labeled 'Search for Terms', a 'Search' button, and a link to 'Options'.

Definition/Usage of Knowledge Graphs (KG)

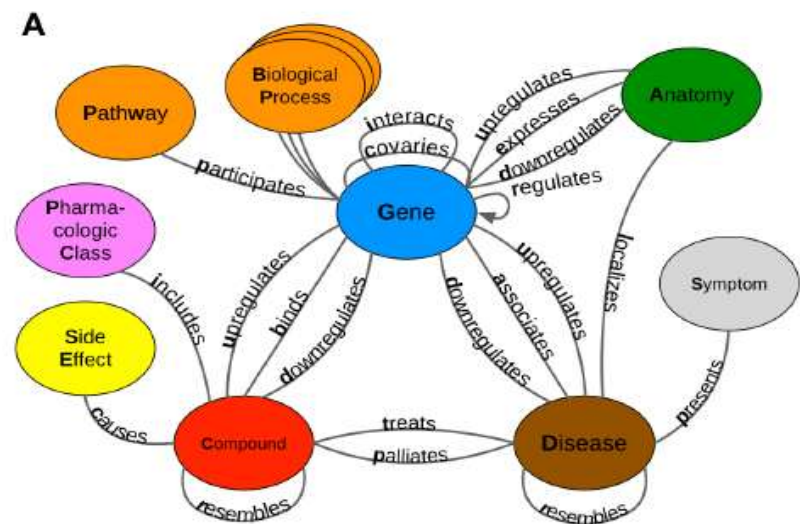
A **knowledge graph (KG) tool** is a software or framework designed to create, manage, and query knowledge graphs. A knowledge graph is a structured representation of information where entities (nodes) and their relationships (edges) are explicitly modeled, often using graph-based structures. These tools enable organizations to organize and analyze complex relationships between data, enabling insights, reasoning, and decision-making.



Example of a KG for Healthcare and Life Sciences:

The cost of developing a new therapeutic drug has been estimated at 1.4 billion dollars, the process typically takes 15 years from lead compound to market. Hetnets (short for heterogeneous networks) are networks with multiple types of nodes and relationships. In this study, we developed a hetnet (Hetionet v1.0) by integrating knowledge and experimental findings from decades of biomedical research spanning millions of publications.

See: [DHBW-Moodle]: “KA4HealthCare-Examples” and KA4HealthCare-Exam_Hetionet”.



Examples / Use Cases of Knowledge Graphs

1. Graph Databases:

- **Neo4j**: A leading property graph database with powerful visualization and querying capabilities.
- **TigerGraph**: Focused on large-scale analytics and real-time processing.
- **ArangoDB**: Multi-model database supporting graph data alongside other types.

2. RDF*-Based Tools:

- **Apache Jena**: A framework for building semantic web and linked data applications.
- **GraphDB**: RDF-based graph database optimized for semantic knowledge graphs.
- **Blazegraph**: High-performance graph database supporting SPARQL.

3. Enterprise Solutions:

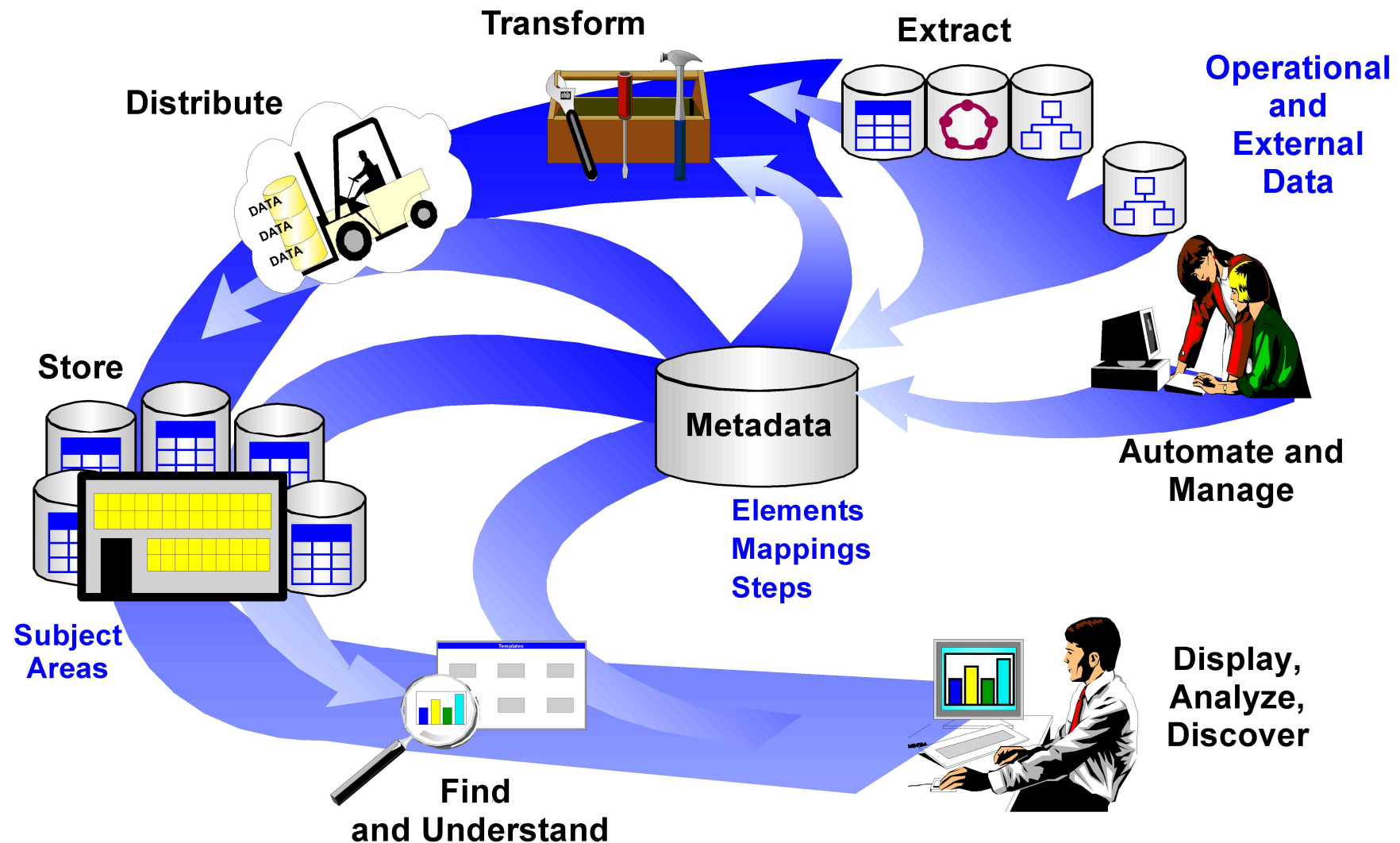
- **Anzo by Cambridge Semantics**: A scalable knowledge graph platform for enterprise data.
- **Stardog**: Combines a graph database with reasoning and integration capabilities.
- **Ontotext**: Focused on enterprise knowledge graphs with RDF and OWL support.

4. Visualization and Query Tools:

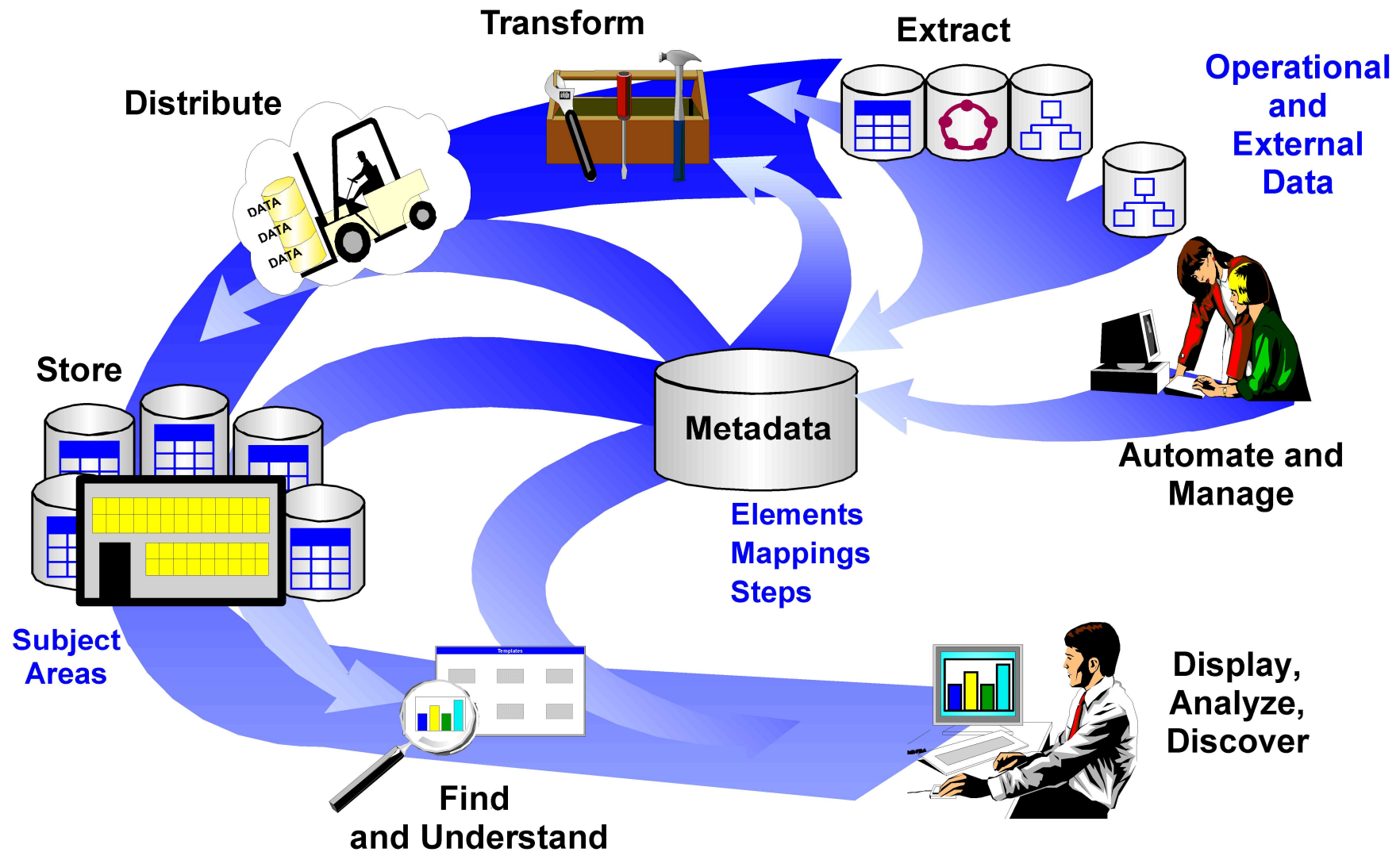
- **GraphX (Apache Spark)**: A distributed graph processing library.
- **Linkurious**: A visualization tool for exploring graph data.

* Resource Description Framework

Turning Data into Information (Part1)

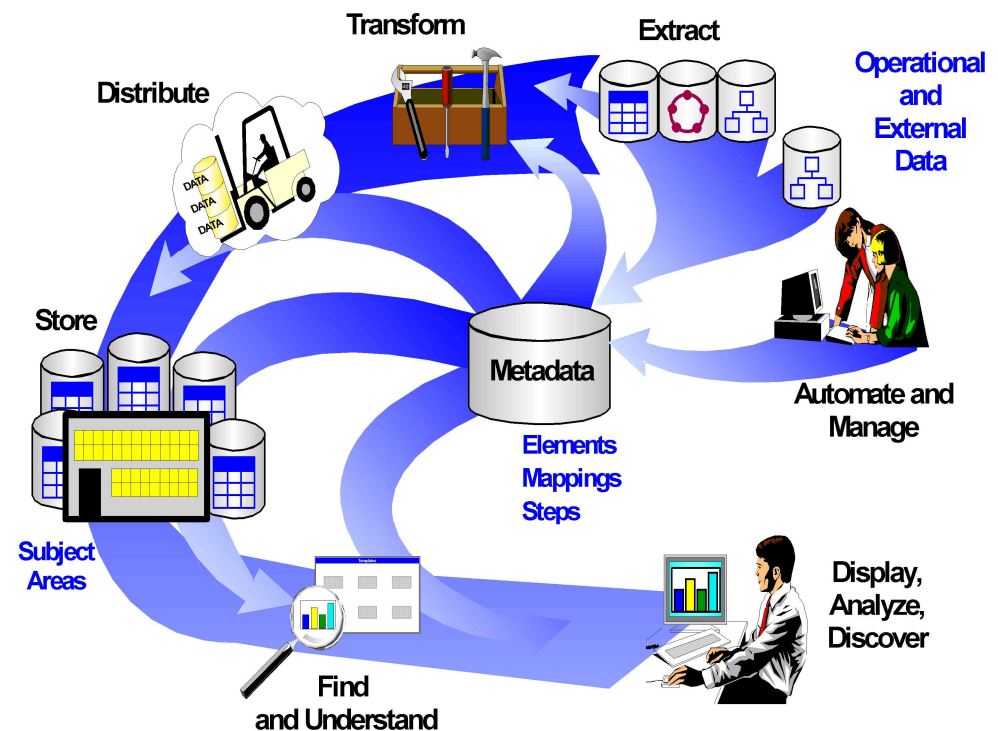


Turning Data into Information (Part2)



What is a Data Warehouse ?

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

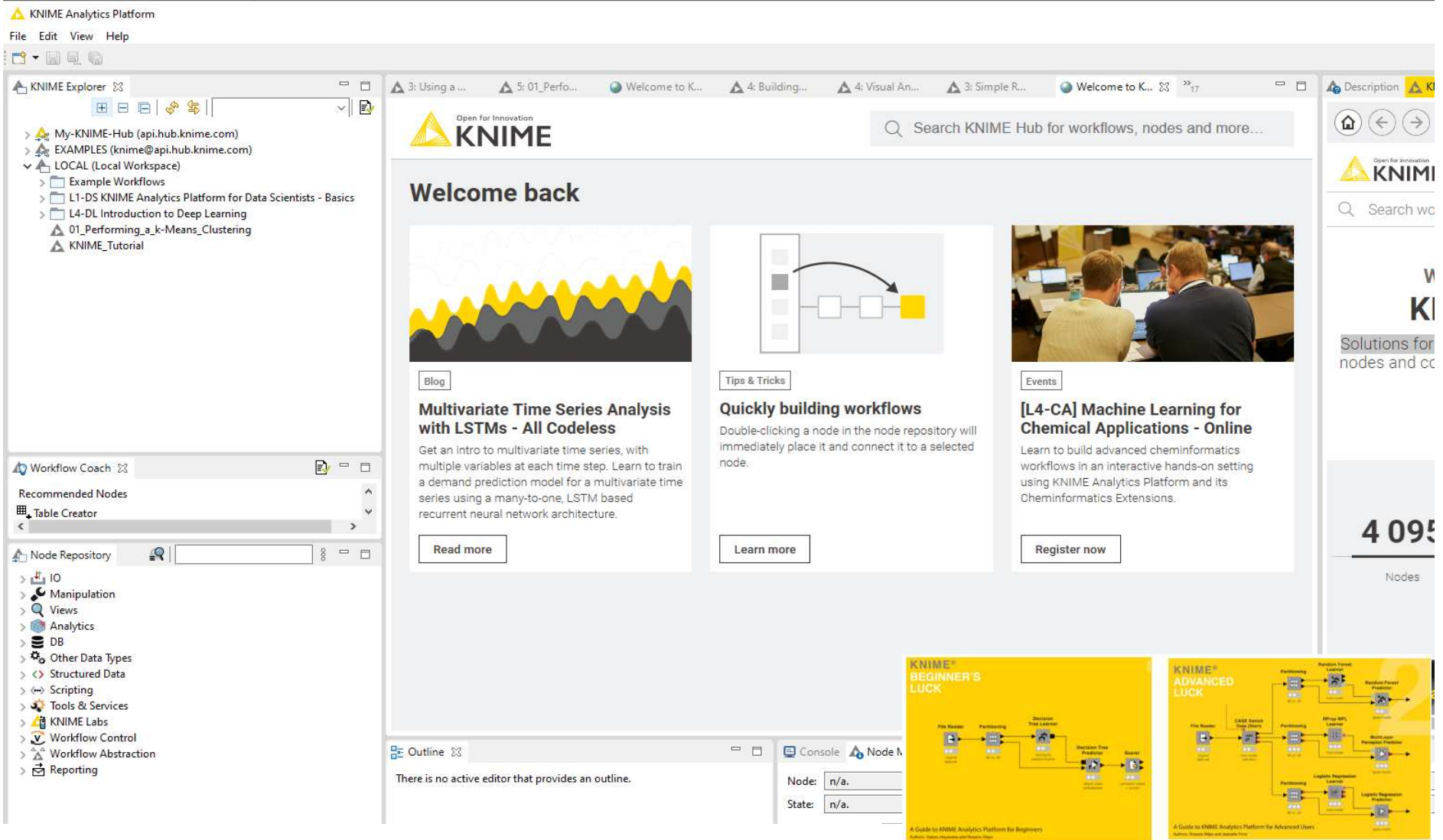


Goal: Turning Data into Information !

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



The screenshot displays the KNIME Analytics Platform interface. On the left, the **KNIME Explorer** shows a project structure with folders like 'My-KNIME-Hub', 'EXAMPLES', and 'LOCAL (Local Workspace)'. Below it, the **Workflow Coach** recommends nodes such as 'Table Creator'. The **Node Repository** lists various node categories like IO, Manipulation, Views, Analytics, DB, and Scripting.

The main workspace features a **Welcome back** message and three featured articles:

- 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](#)
- Quickly building workflows**: Double-clicking a node in the node repository will immediately place it and connect it to a selected node. [Learn more](#)
- [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](#)

At the bottom, there are two guides: **KNIME® BEGINNER'S LUCK** and **KNIME® ADVANCED LUCK**, both providing step-by-step workflows for new and experienced users.

First Exercise for DW01

Exercise E1.1 (SW*): Investigate the BI-Data Trends in 2025.

*Prepare and present the results of the e-book “**BARC-BI_Data_Trend_Monitor-25**”. See more details under Moodle group [Kurs: T3INF4304_3_22D](#). 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 data trends to support the new digital requirements and build the data management for Artificial Intelligence (AI) solutions.*

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

Second Exercise for DW01

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

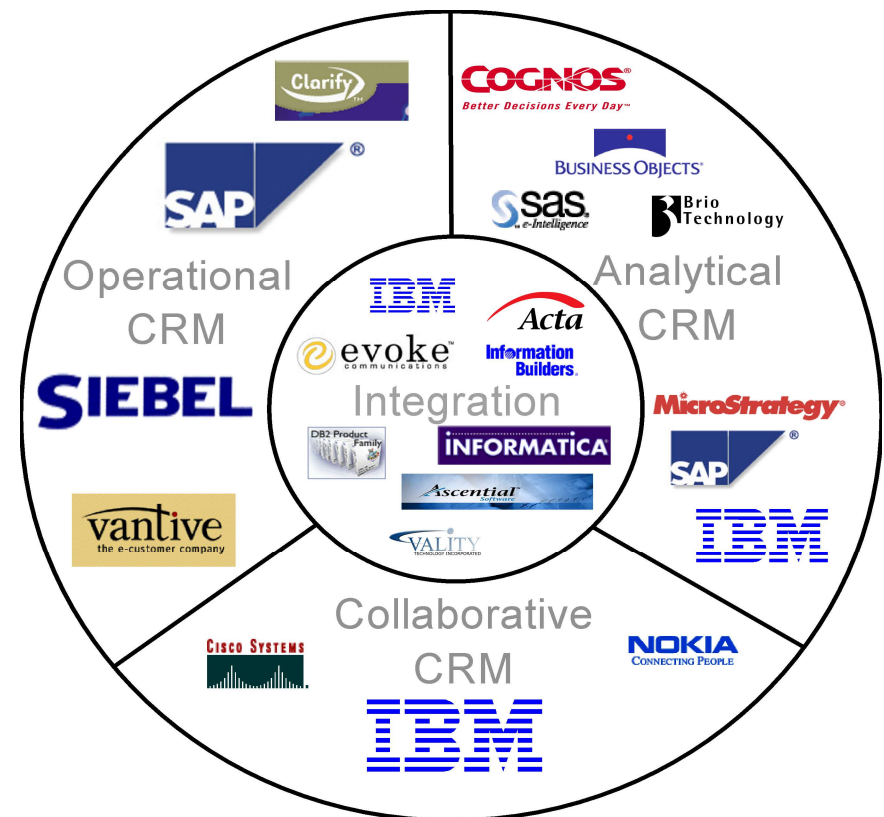
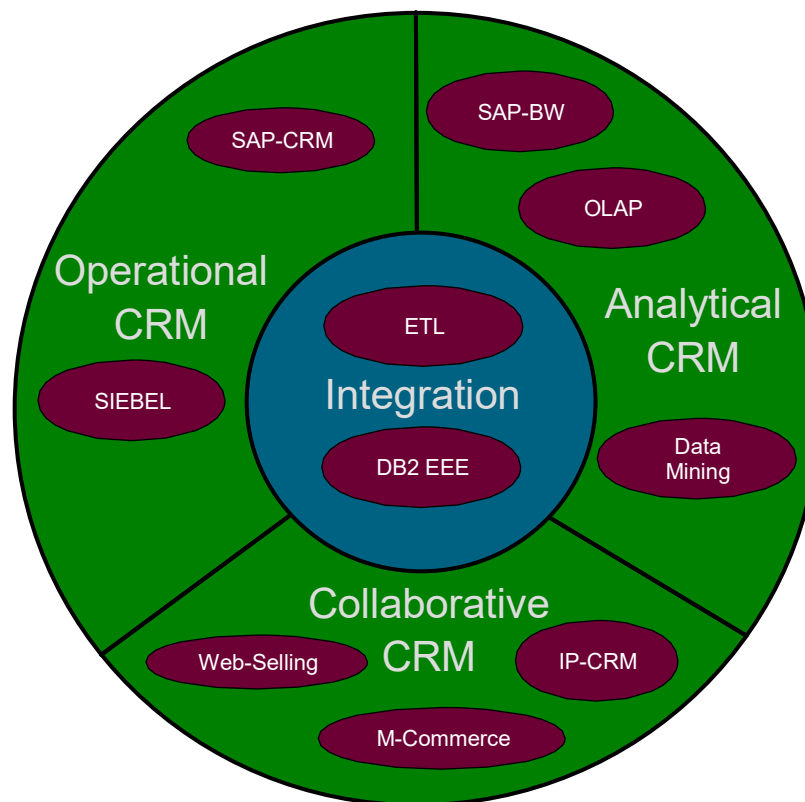
Prepare a report and present it next week; duration = 30 minutes (10 min for each area). Information sources are newspaper or magazine articles or books (see literature list). 3 students.

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)

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

Hints to E1.2: CRM Categories and Tools



Third Exercise for DW01

Exercise 1.3(SW*): 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].

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

Fourth Exercise for DW01

Exercise 1.4a: Examine the differences between IGC and a Knowledge Graph (KG) tool and give a detailed description of the KG tools in one of the two categories (“functionality and “strengths and weaknesses”)

1. *What is the overlap (i.e. what is common) of the two tools?*
2. *What is the distinction of the tools*

Exercise 1.4b (PD**): Give a program design & detailed description of the creation of a KG tool in the area of Healthcare and Life Sciences. (i.e. modeling biological pathways, drug discovery, and patient data.);

See YouTube video "Building Knowledge Graphs in 10 Steps":

<https://www.youtube.com/watch?v=lUc0woFX16M>

See also the description of Google- /IBM HealthCare KG in [DHBW-Moodle], Category1: “KG4HealthCare-Google_Info/IBM_Info.pdf”; “Warum FHIR_ – HL7Deutschland.pdf”

PD**: For the Program Design investigate this KG tool in more detail

Fifth Exercise for DW01

Exercise 1.5 (PD**): 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](https://www.knime.com) and the three mentioned documents in the lesson DW01 (see lesson notes).

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

PD**: For the Program Design investigate these KNIME features in more detail

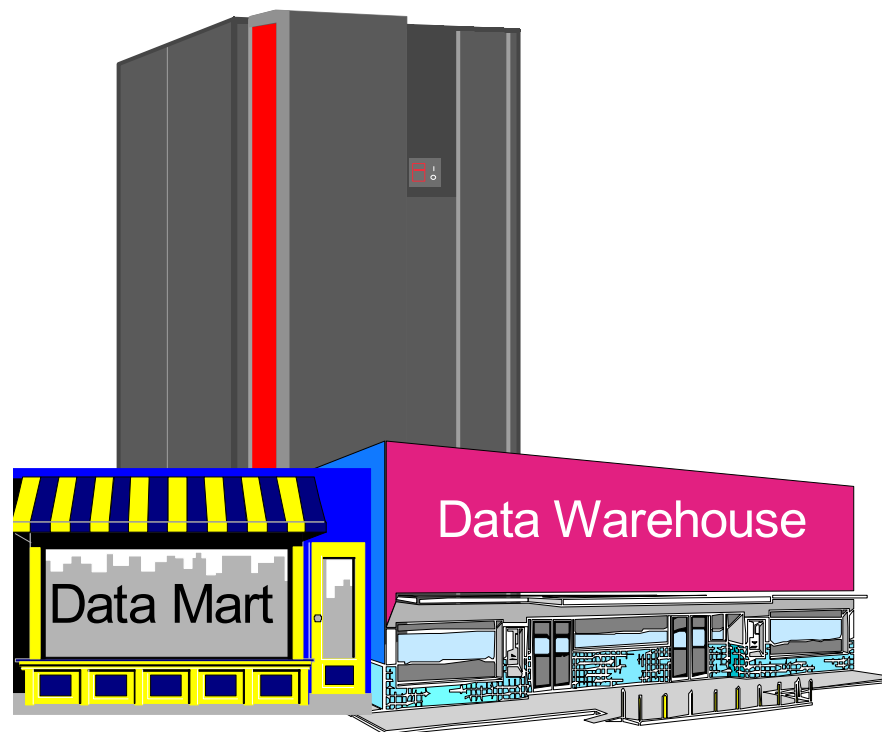
Category 1: Introduction & Architecture of DWH

Category 2: Databases and Data Modeling

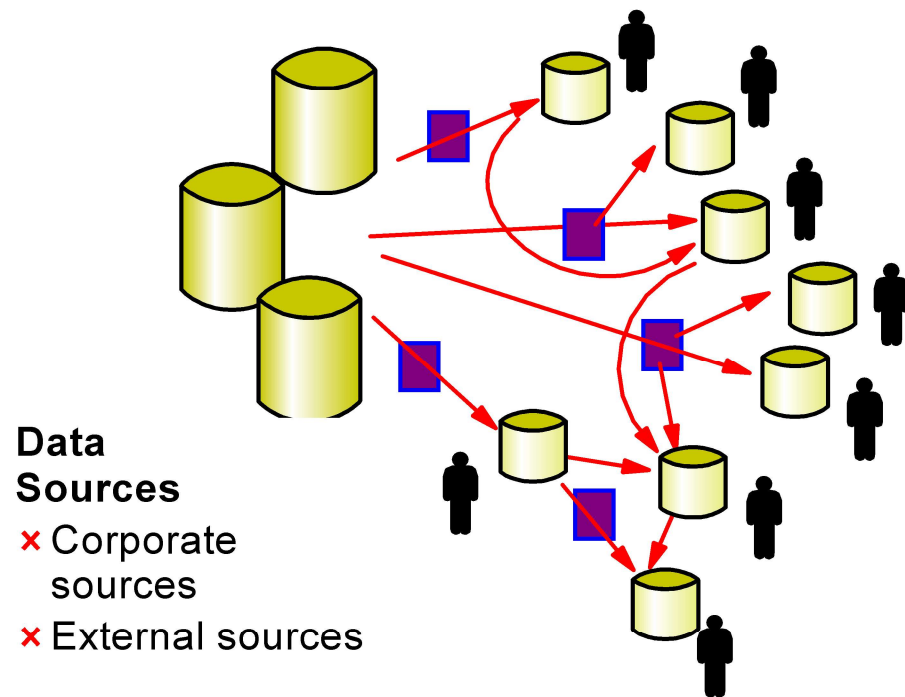
Category 3: ETL: Architecture & Technology

Category 4: Descriptive – & Advanced Analytics

DW02 - Introduction to DWH Architecture



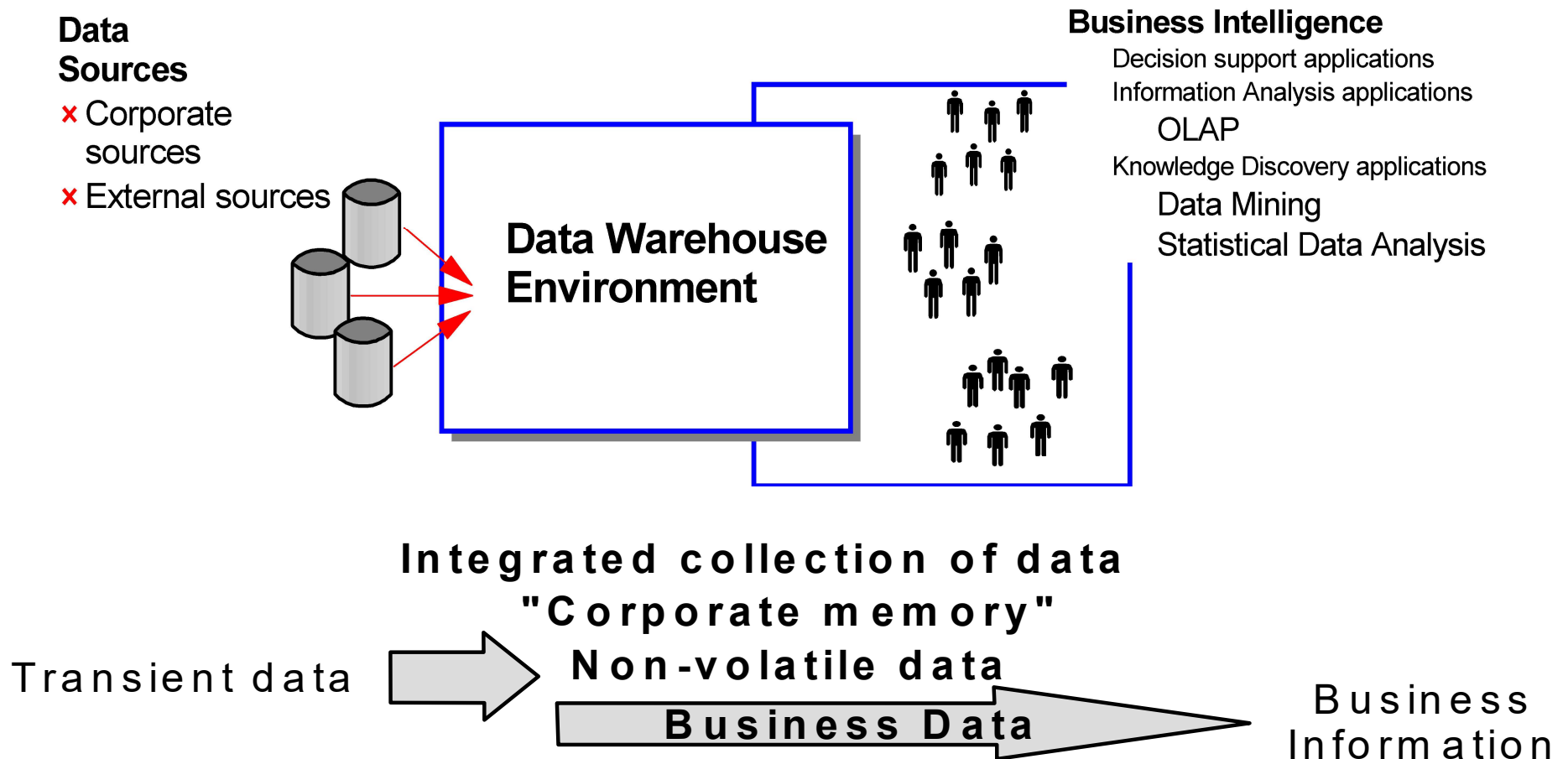
Ad-Hoc Evolving DWH Environments



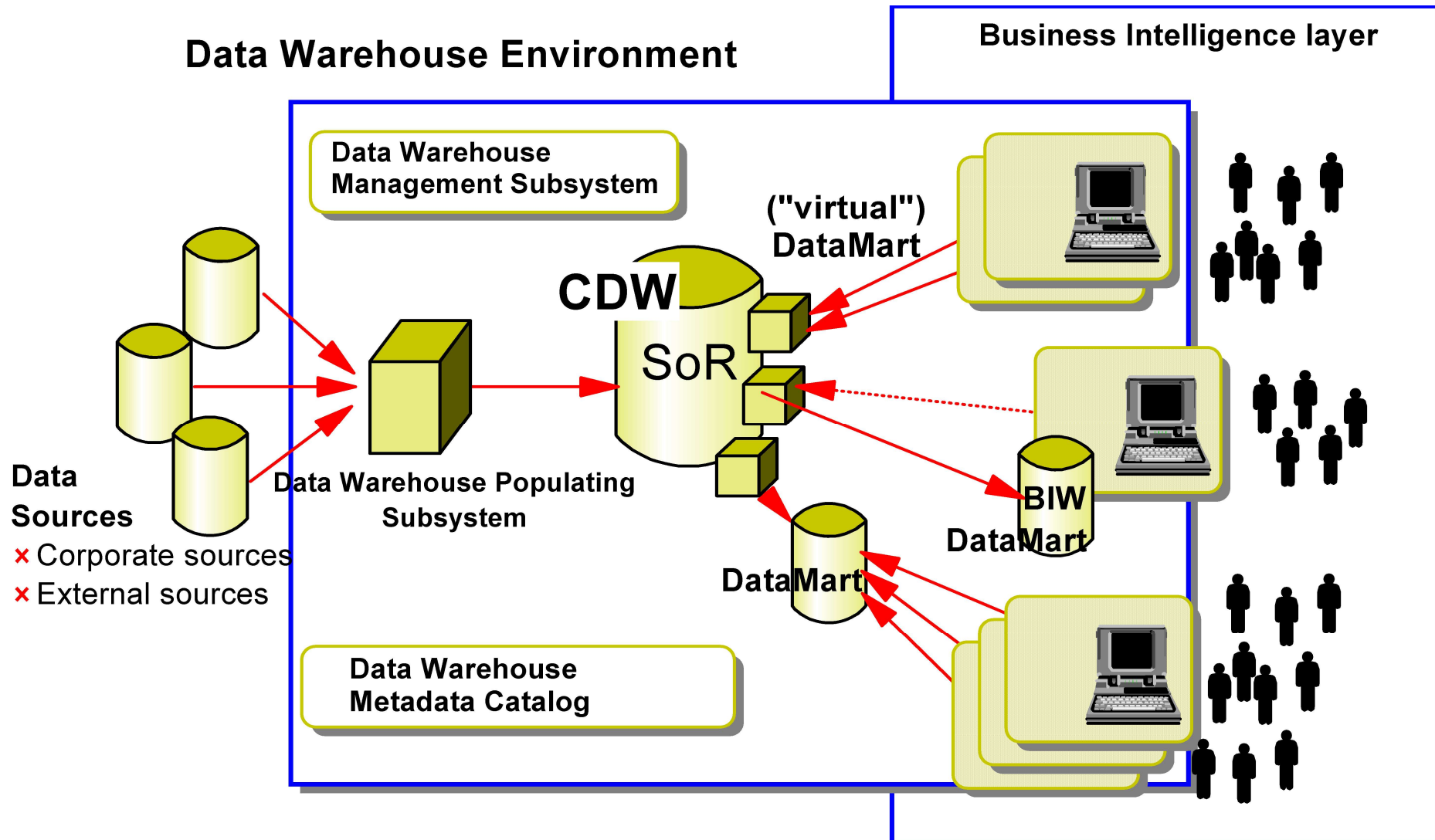
Problems

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

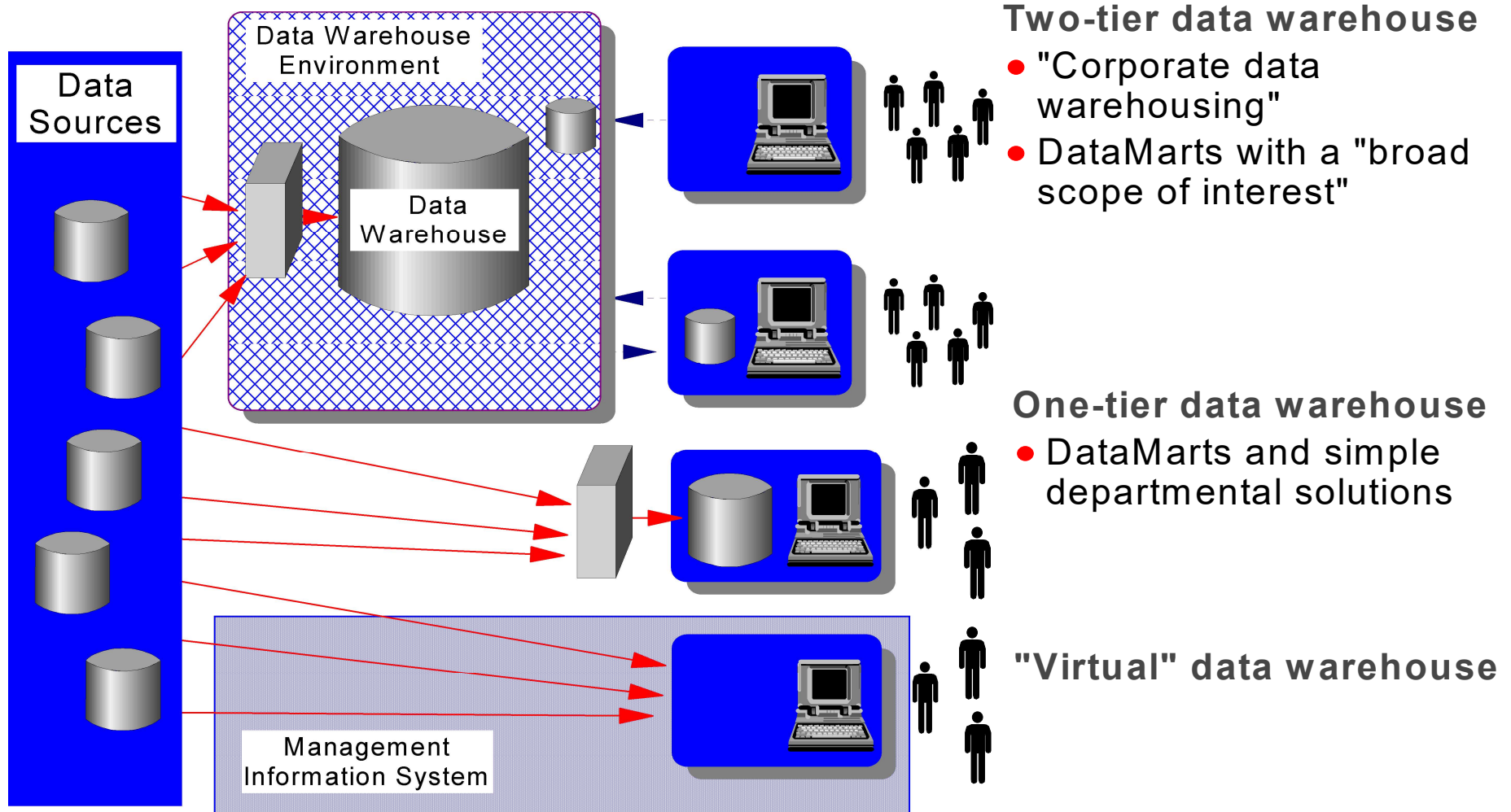
Setting the Scene



Setting the Scene (Cont)



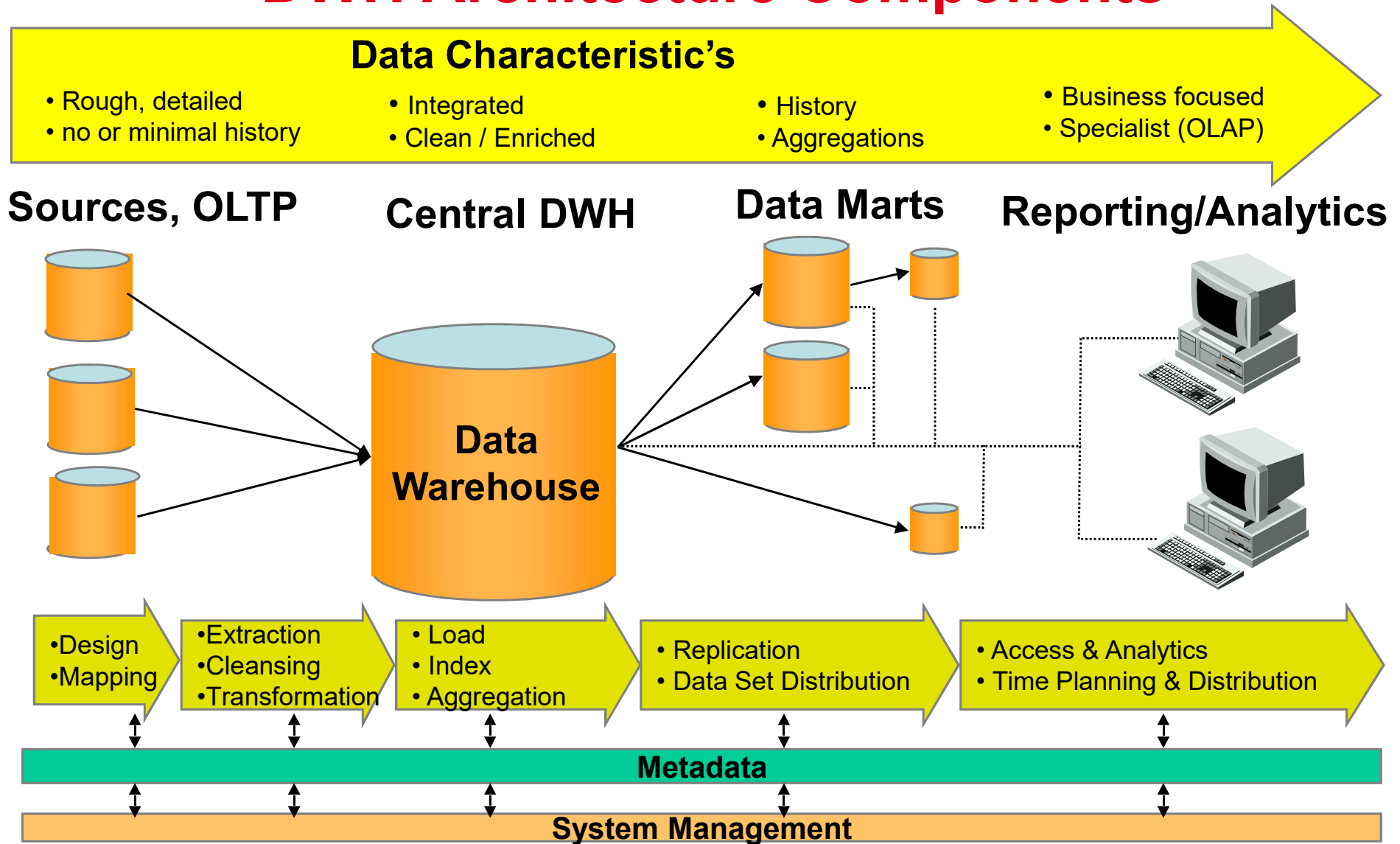
DWH - Possible Approaches



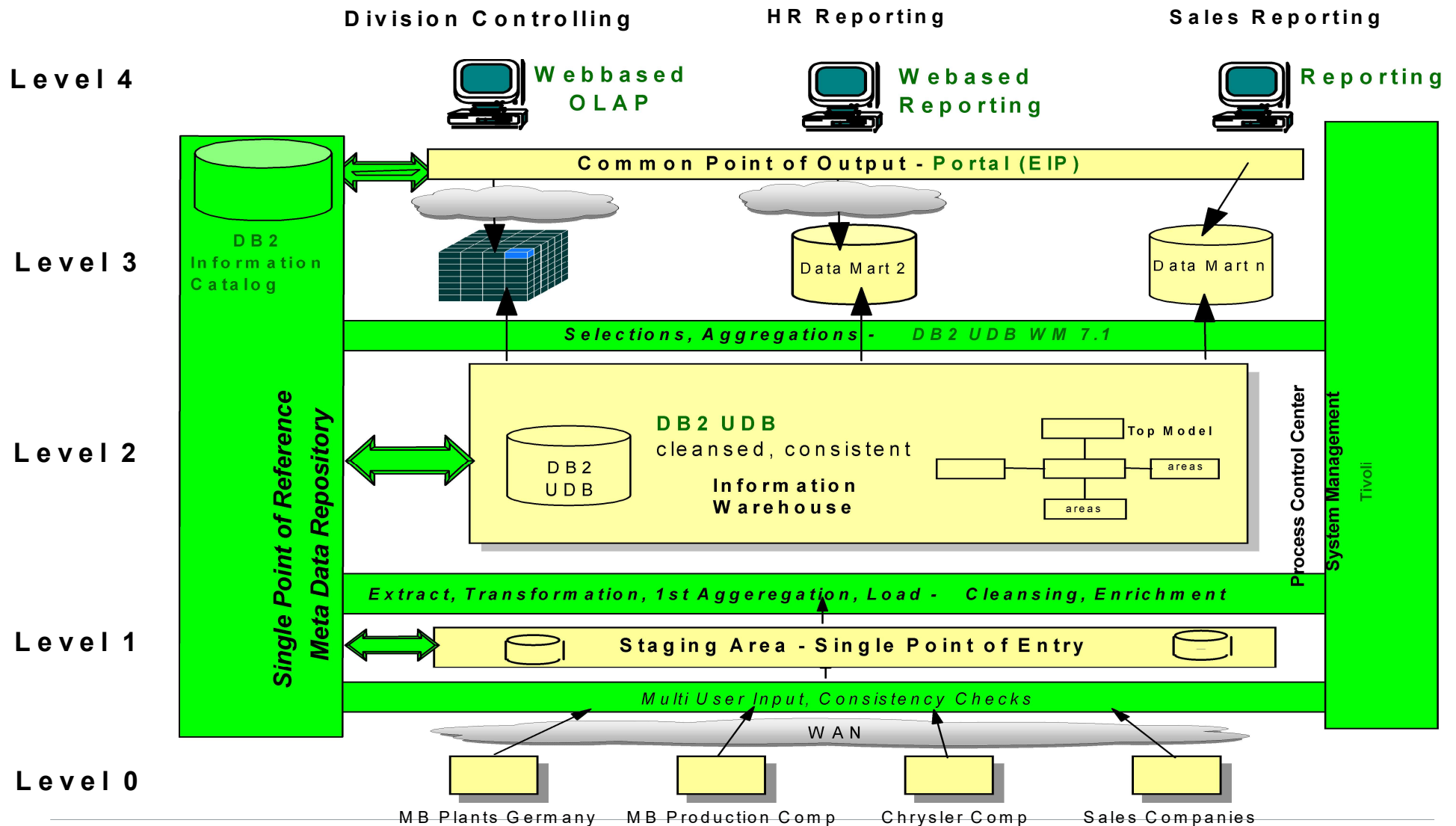
Data Marts or Data Warehouses

- **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**

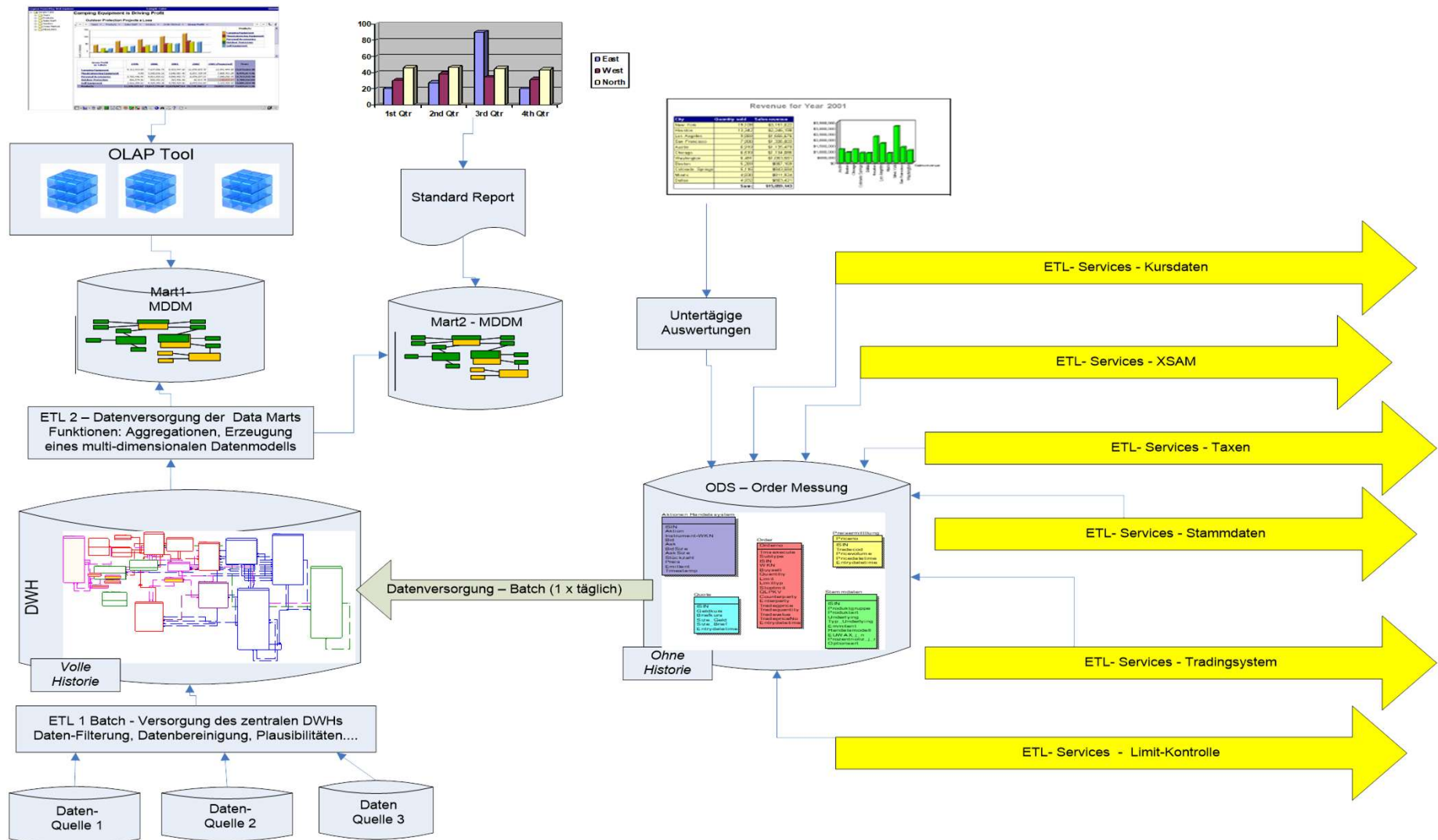
DWH Architecture Components



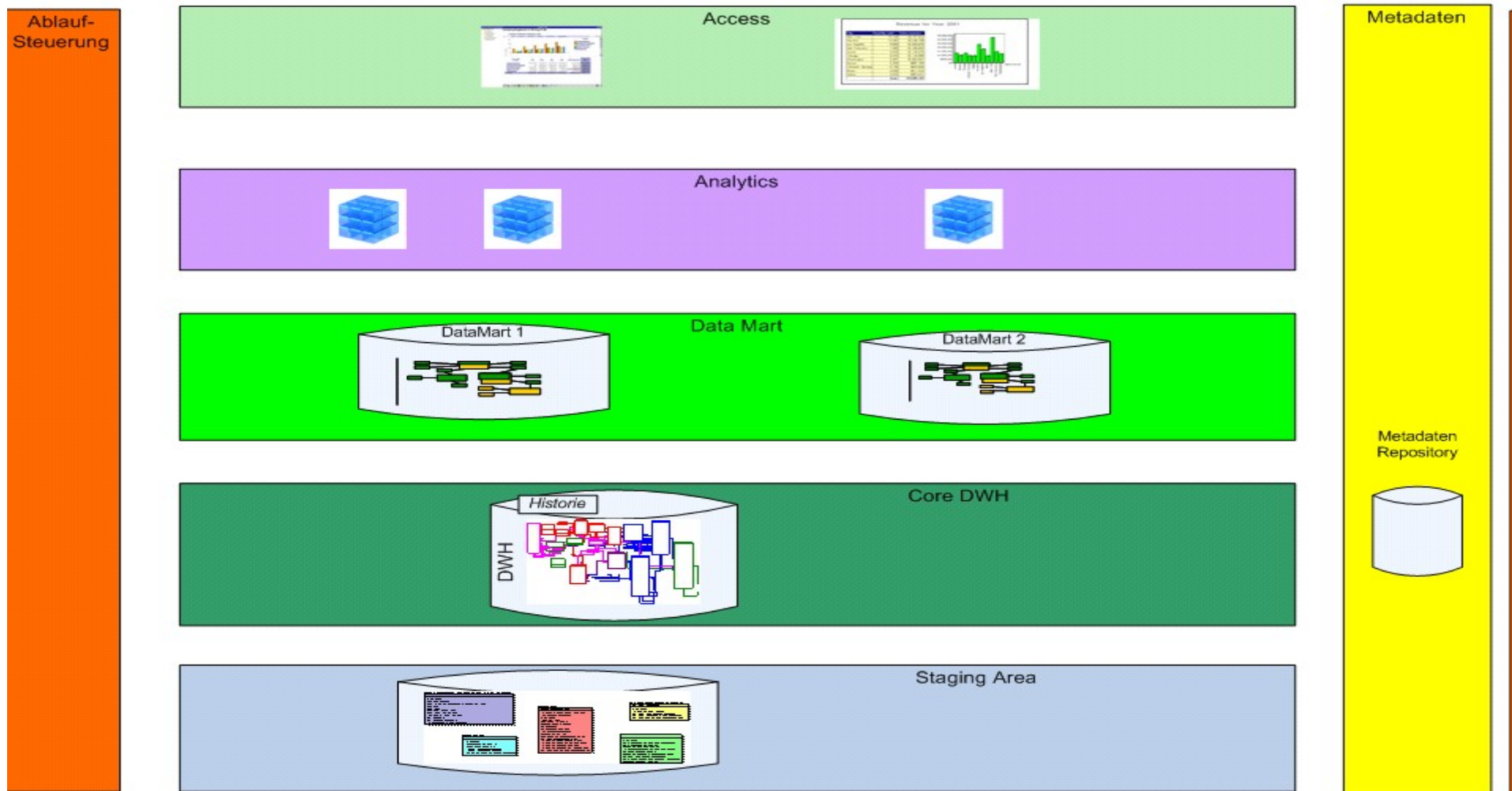
DWH Architecture – ‘Big Picture’ Example



Example of a Financial Market DWH



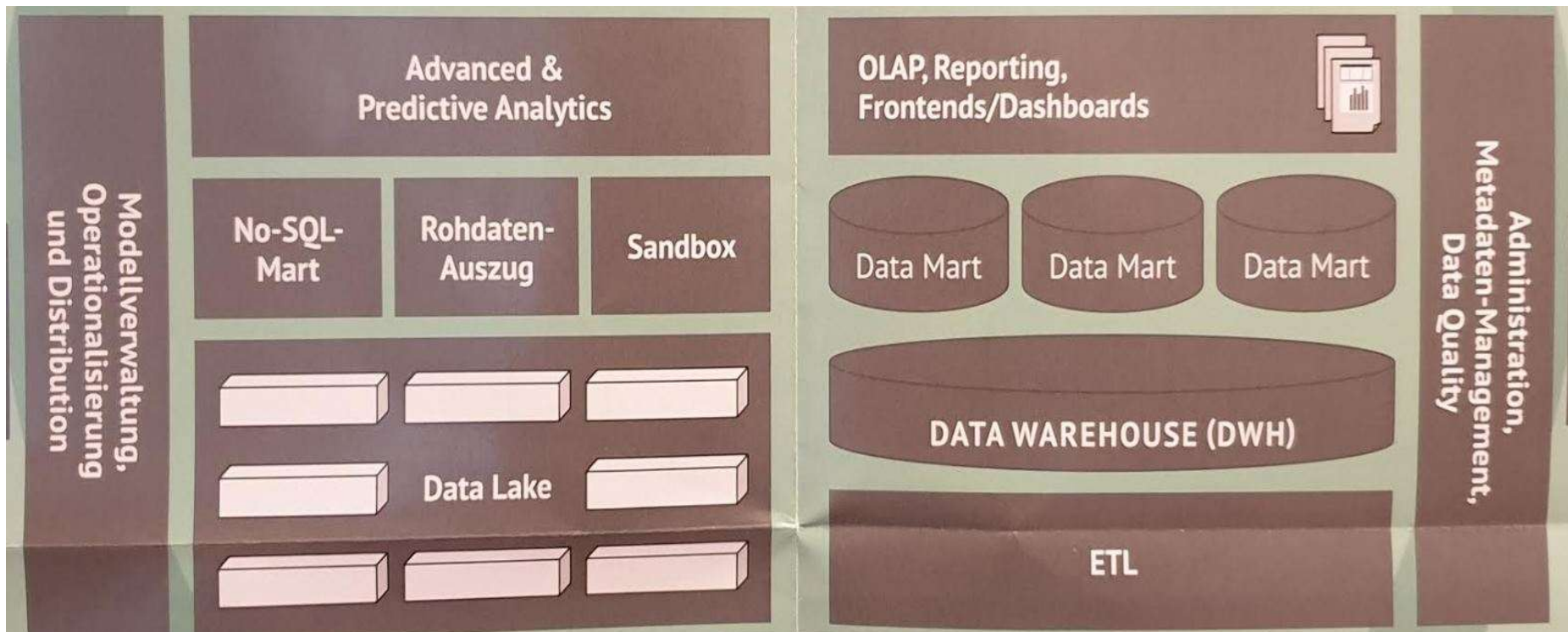
DWH Architecture – Data Layer Concept



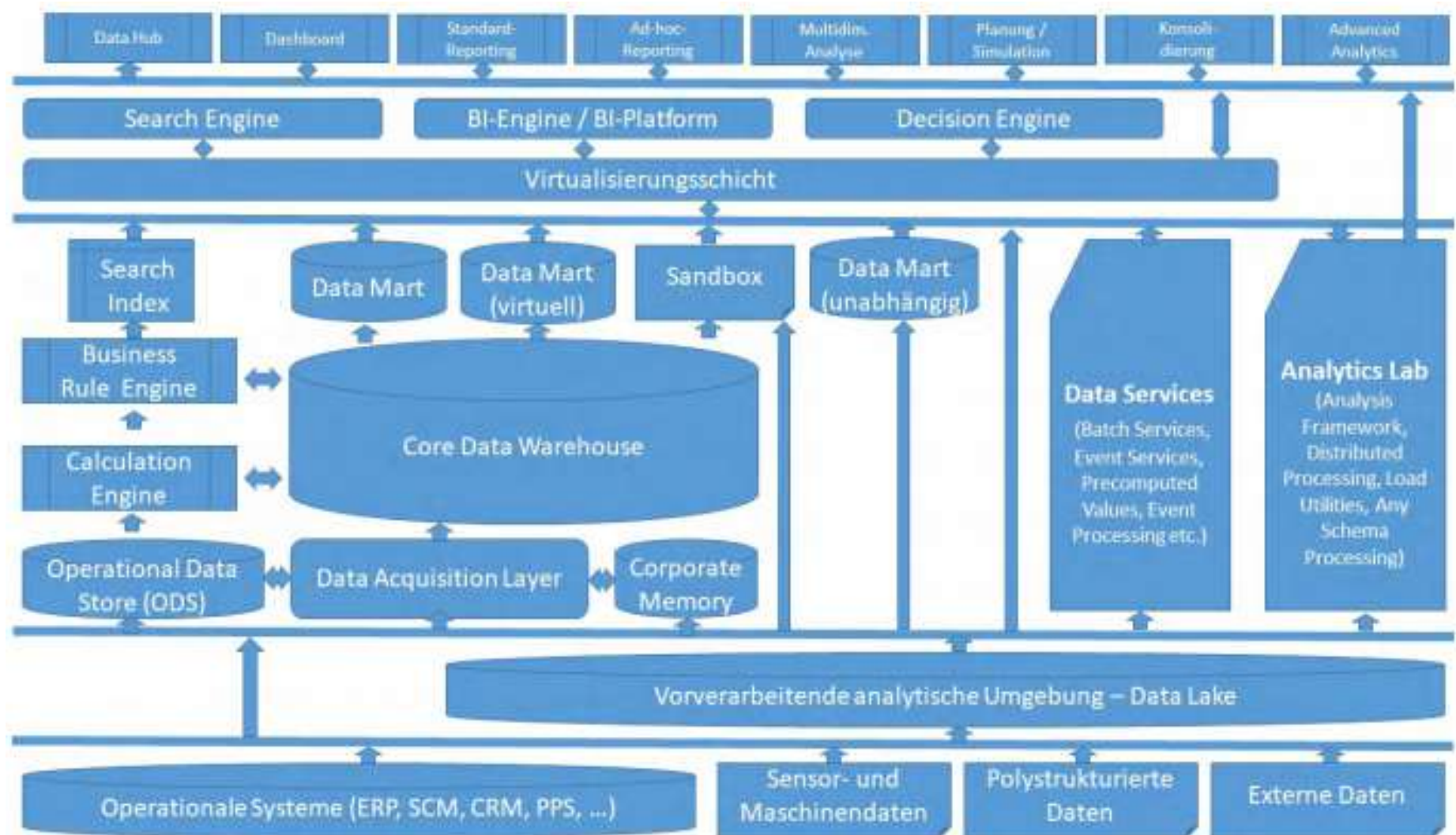
Modern Data Architecture – Big Data & Cloud

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".

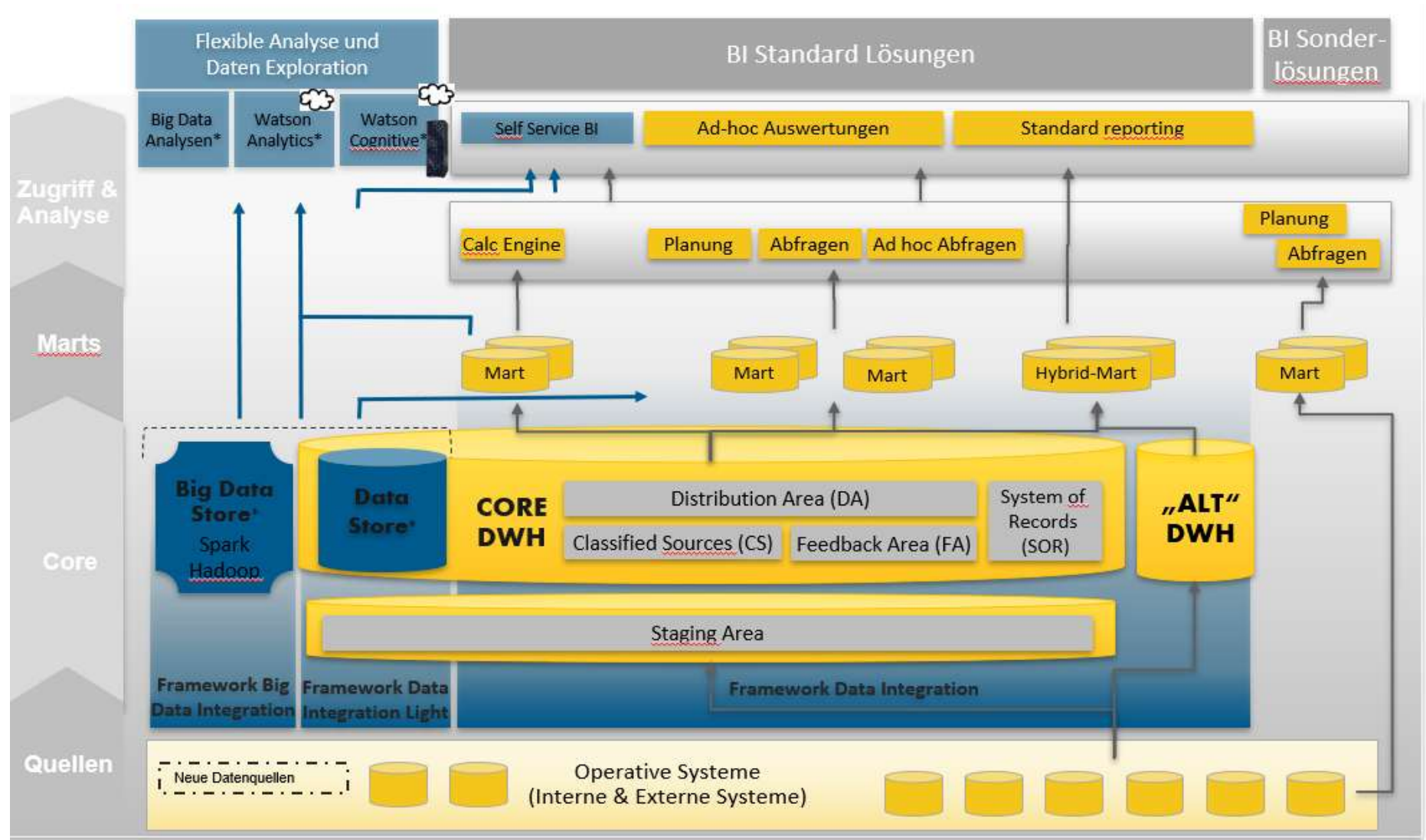
Also advanced analytical processes can be used over the Cloud, i.e. data scientists accessing the data lake data for running predictive analytical jobs and machine learning algorithms.



Modern Data Arch. – Data Lake Integration

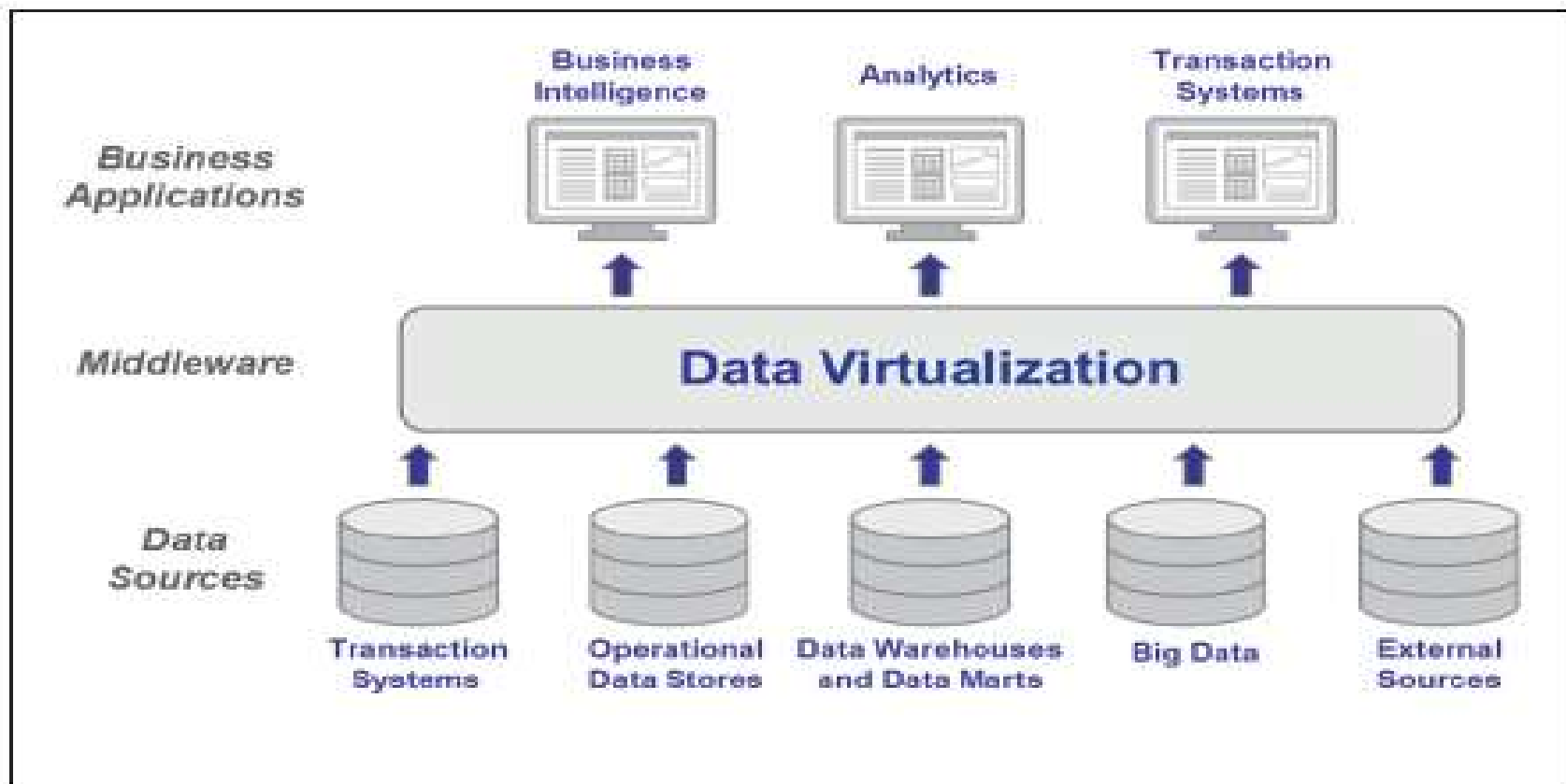


Modern Arch. - Example of a German Insurer



Modern Data Arch. – Virtualization Concept

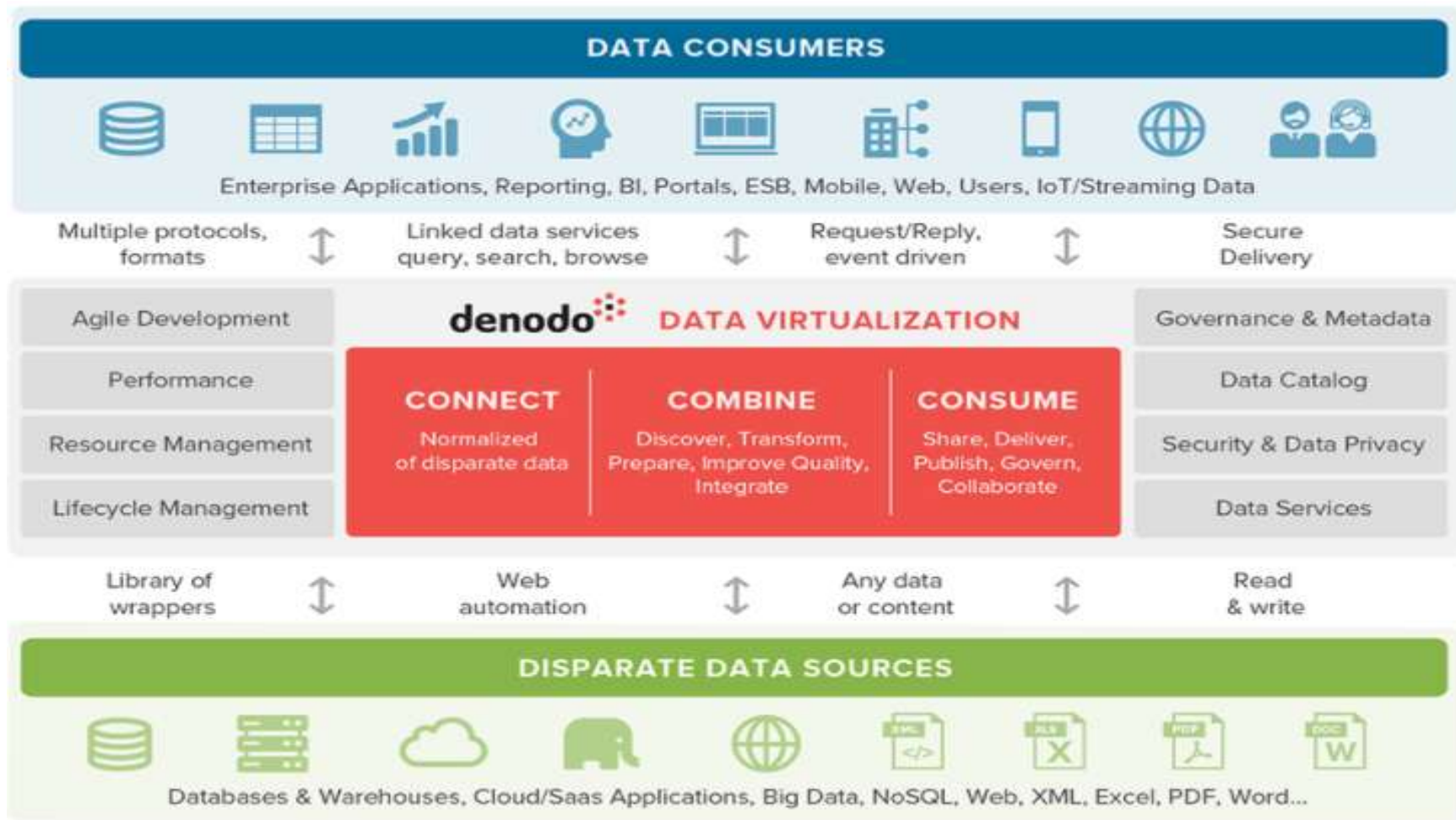
Data virtualization provides a virtual approach to accessing, managing and delivering data without physically replicating it.



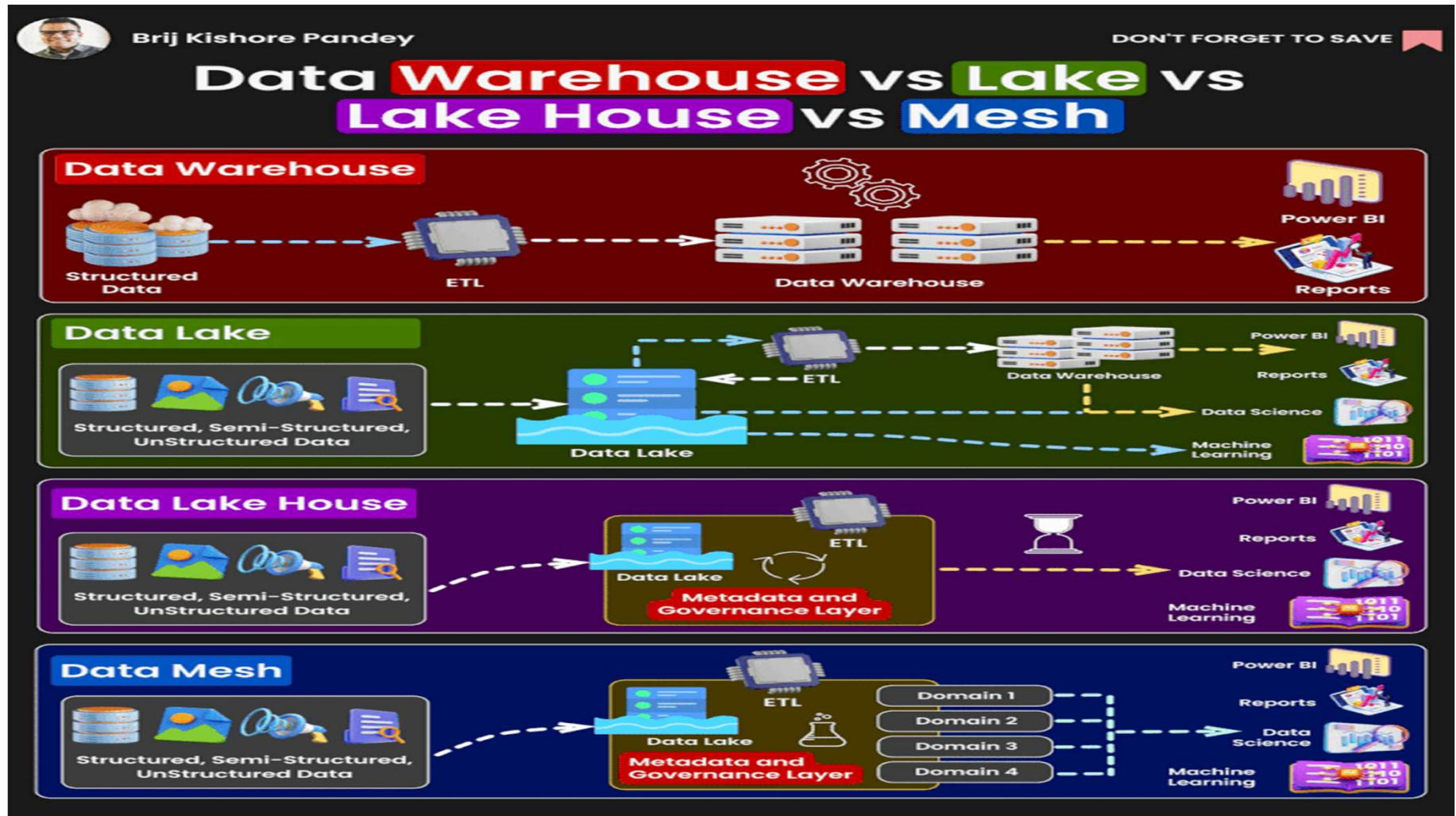
See in [DHBW-Moodle]: eBook_Data_Virtualization_Going_Beyond_Traditional_Data_Integration.pdf

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.



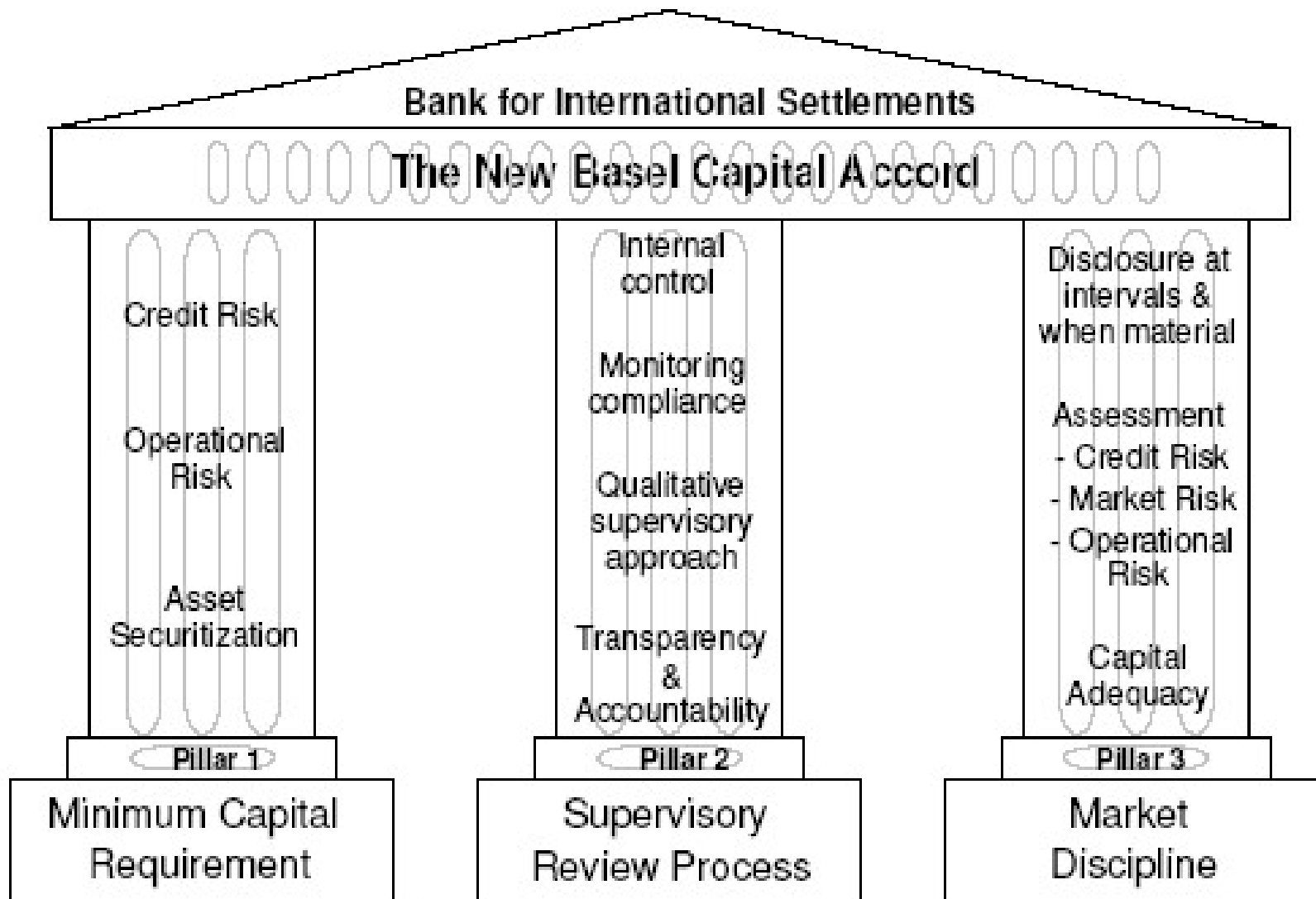
Modern Data Architecture – An Overview



*The Data Visualization Architecture is missing in this picture.

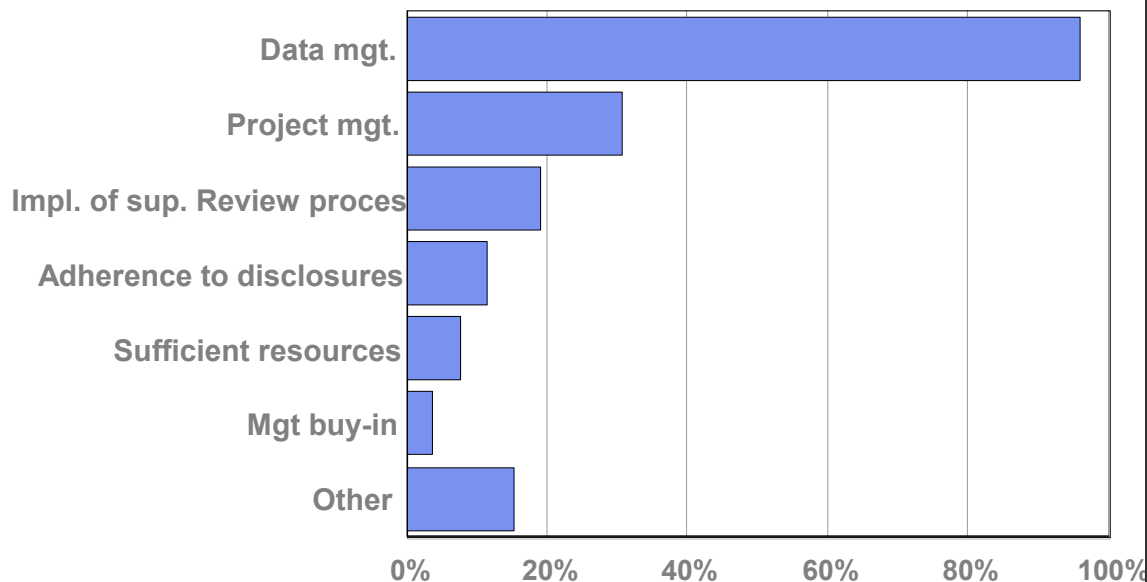
https://github.com/HVoellinger/Data-Warehouse-DWH---Concepts-Applications/blob/main/images/DWH_DLake_DMesh.gif

Use Case I – Basel II (Definition)



Basel II - key challenges – Systems & Data Management

Data Management is the key challenge in meeting Basel II



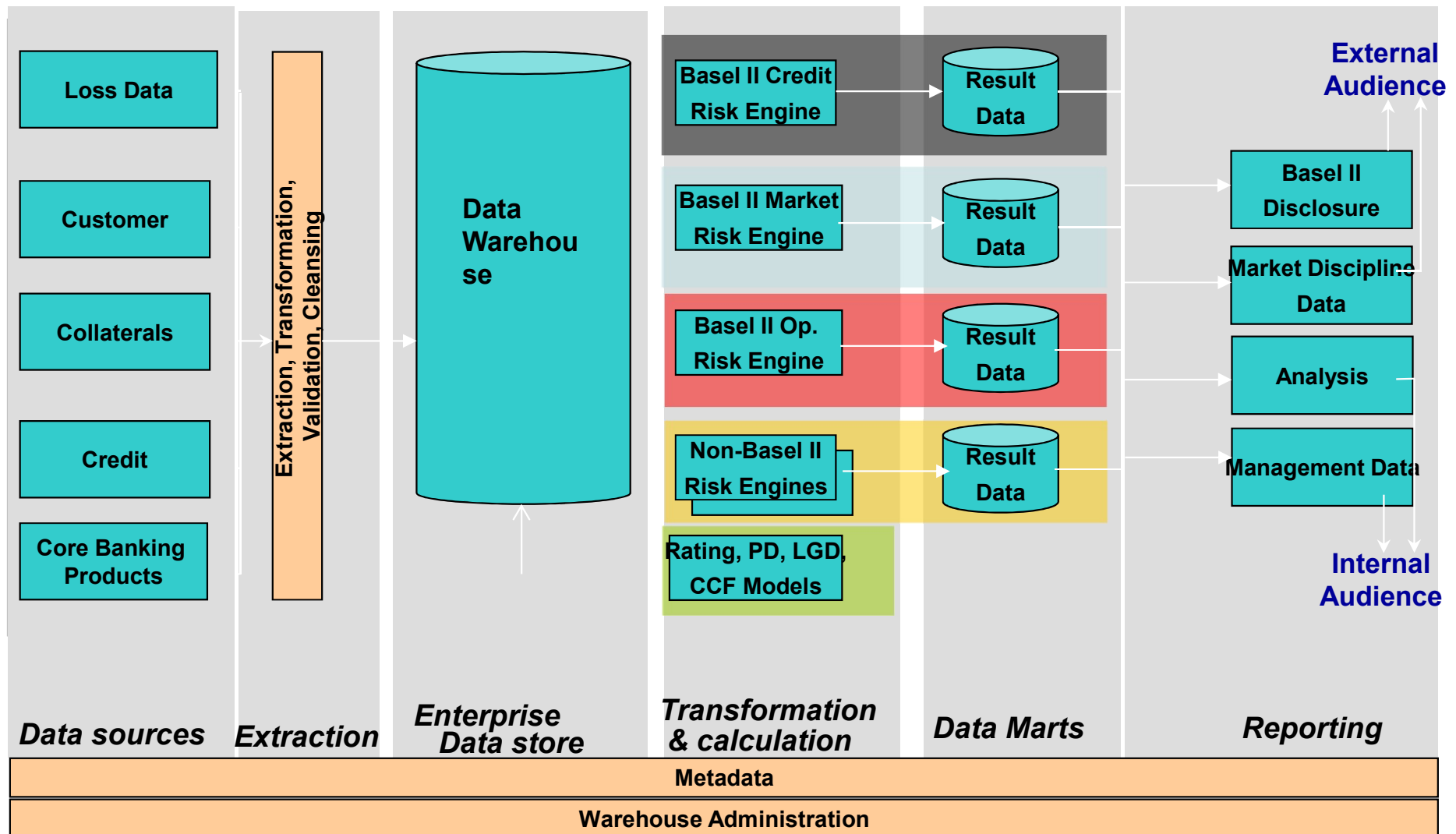
Source: IBM Institute for Business Value analysis, *Banks and Basel II: How Prepared Are They?*, October 2002
interviews with 32 Financial institutions worldwide

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.

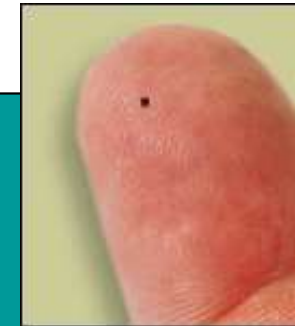
Basel II - 6 Tier Reference Architecture



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



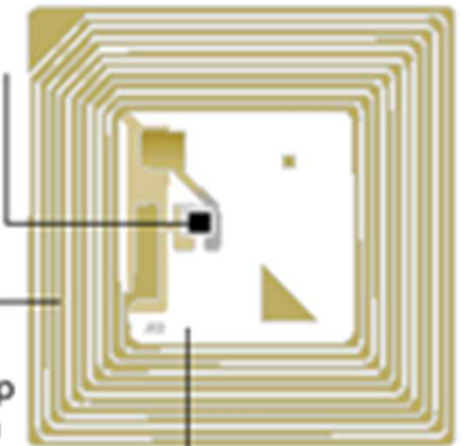
Readers

Reader Interrogators that typically emit a radio signal via an antenna and collect 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:*

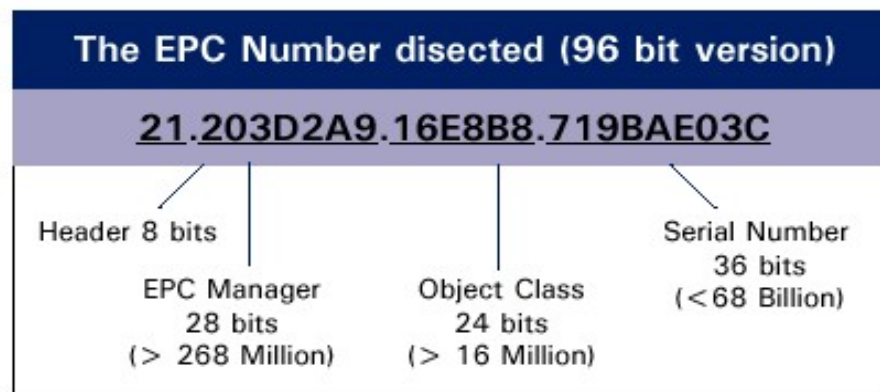
- 1) **Chip:** holds information about the physical object to which the tag is attached
- 2) **Antenna:** transmits information to a reader (e.g., handheld, warehouse portal, store shelf) using radio waves
- 3) **Packaging:** encases the chip and antenna so that tag can be attached to physical object



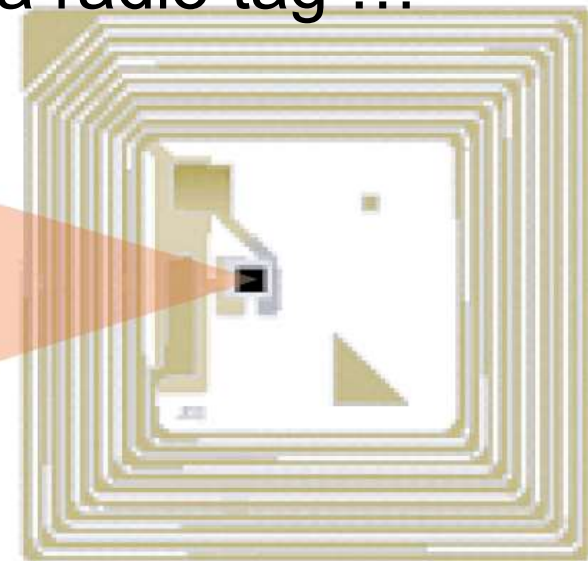
Use Case II - The RFID Numbers

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 ...



Source: Auto-ID Center



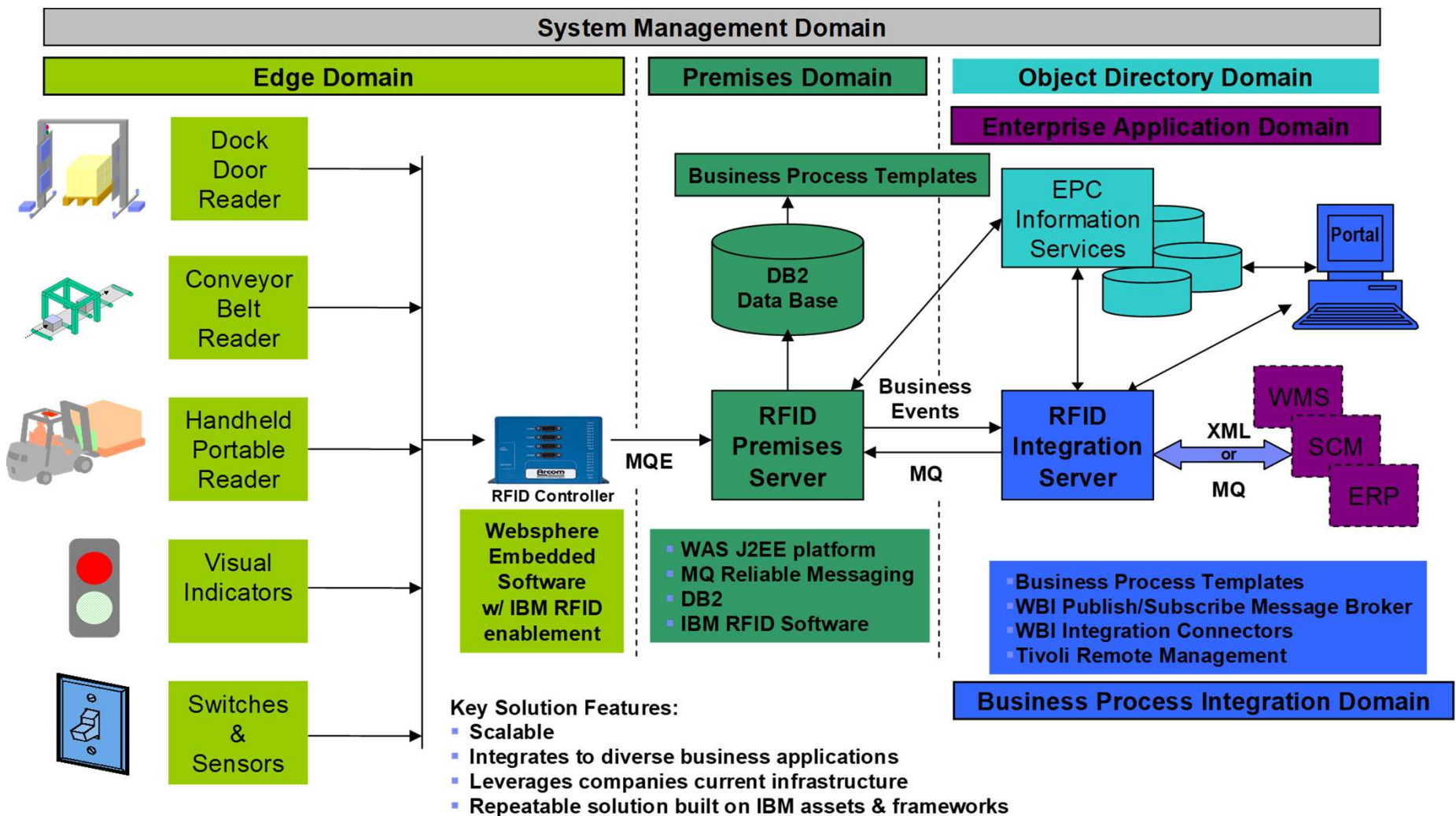
...which together, uniquely identifies an object

The EPC can catalog over 1.3×10^{16} discrete items annually (about the number of grains of rice consumed globally each year). In contrast, the 12 digit UPC barcode can only identify 100,000 products per manufacturer.

Use Case II – The RFID Infrastructure

			
RFID Self-Checkout	Distribution Center Palette Control (DC Exit)	RFID/AutoID Warehouse	EPC RFID Demo
A supermarket scenario similar to the IBM Commercial „Supermarket“	Verify pallet 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

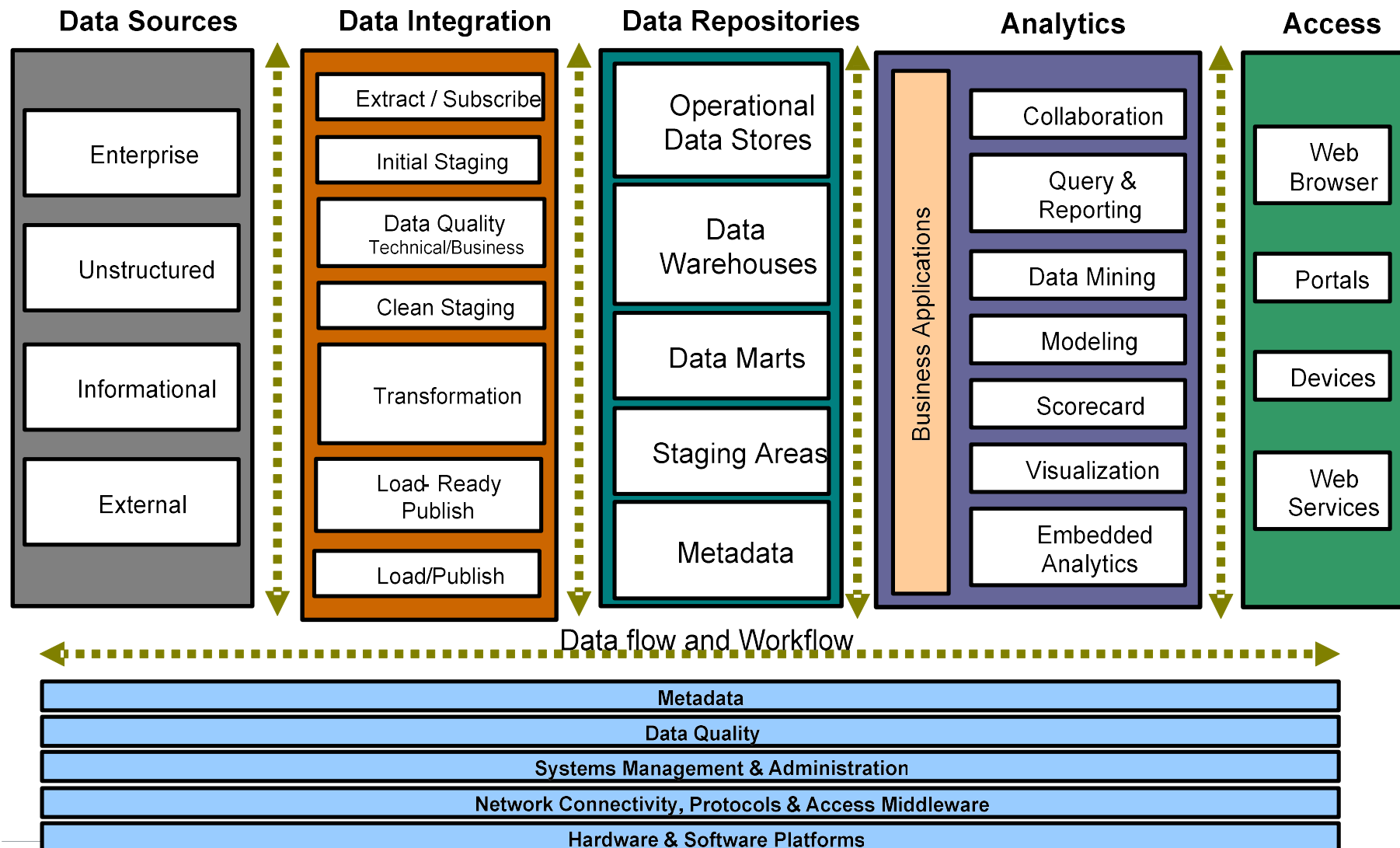


What can go Wrong?

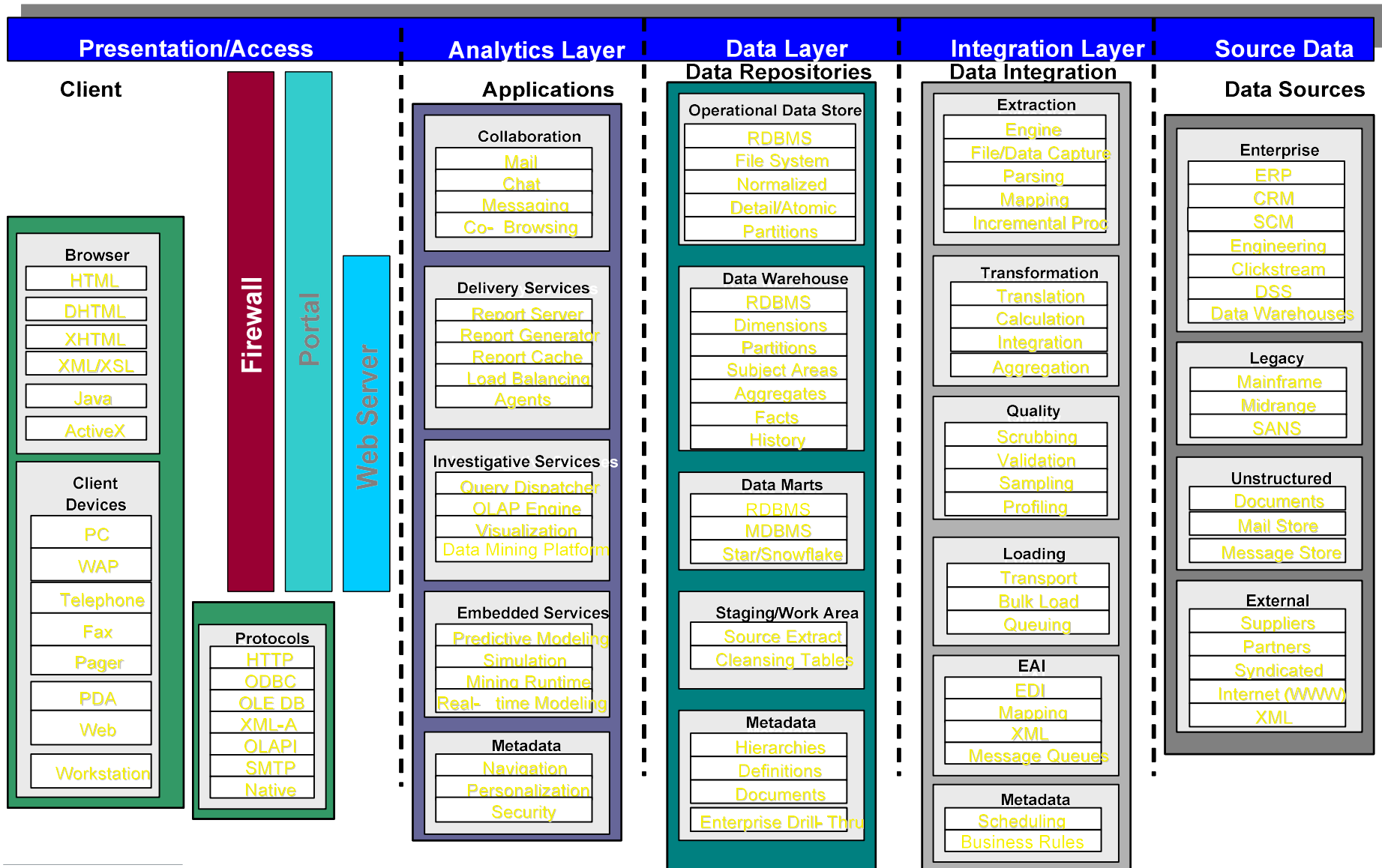
1. **Data Outhouse** - Built too fast; full of dirty, incomplete, out-of-date data; no-one will use it.
2. **Data Basement** - A DW with poor access and/or performance. Not used much.
3. **Data Mausoleum** - Like the basement but built with the finest hardware/software.
4. **Data Shack** - Will soon 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)



IBM DWH Reference Architecture – Details



Exercise 1 to Lesson 2

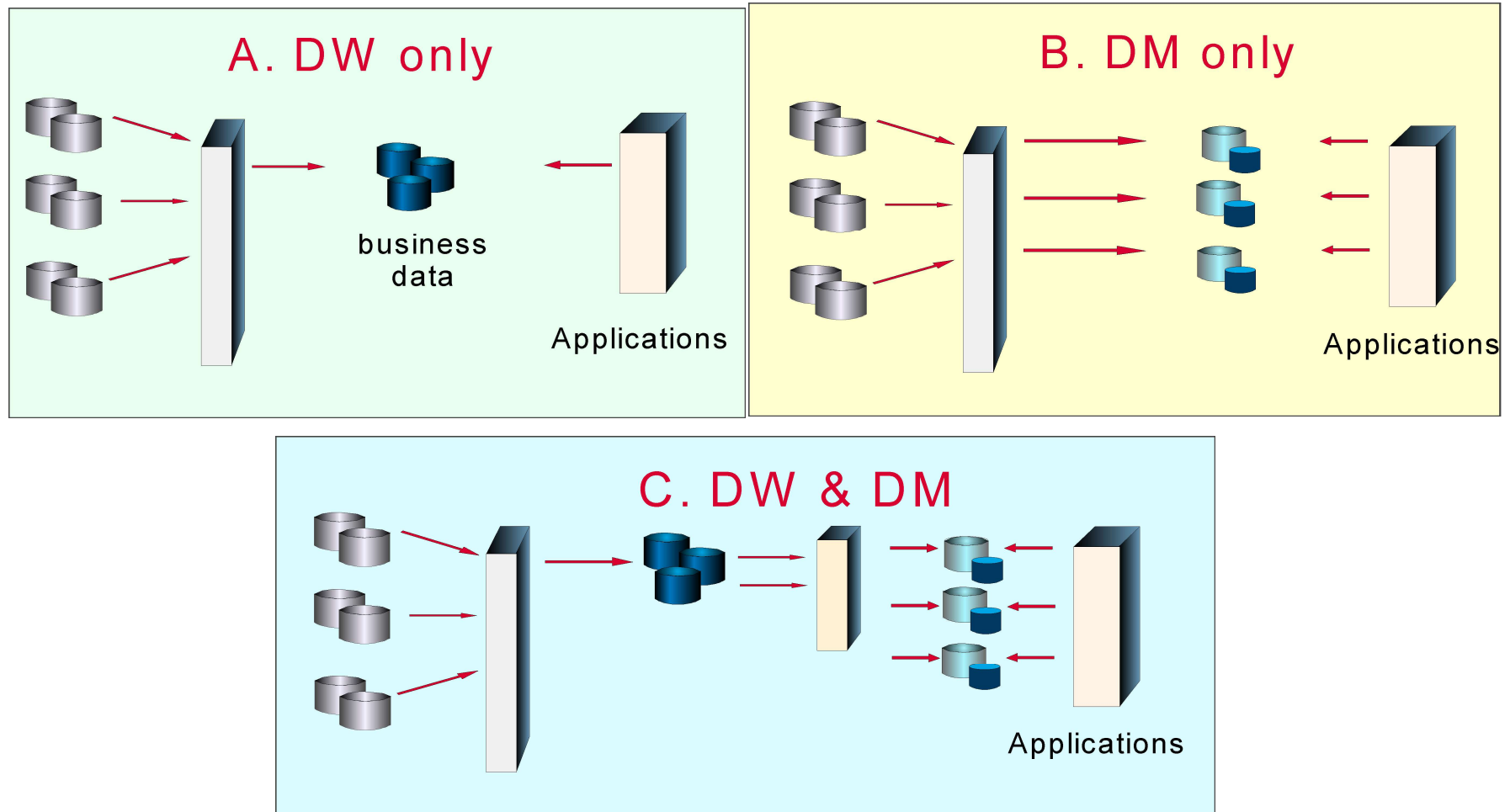
Exercise E2.1 (SW*): Compare the three DWH architectures (DW only, DM only and DW & DM) in the next slide. List the advantages and disadvantages and give a detailed explanation for it. Find also a fourth possible architecture (hint: 'virtual' DWH)

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 .

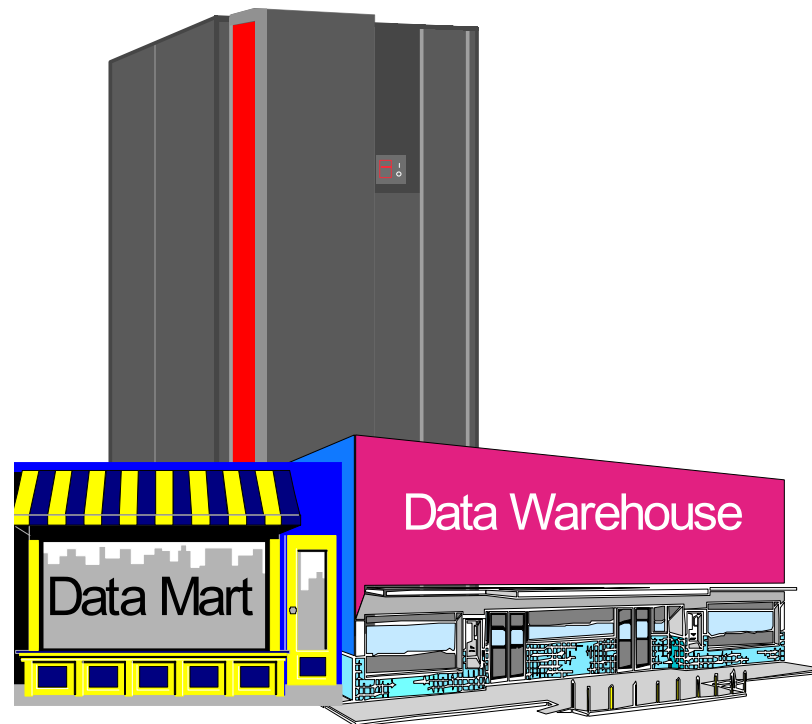
Category 1: Introduction & Architecture of DWH

Category 2: Databases and Data Modeling

Category 3: ETL: Architecture & Technology

Category 4: Descriptive – & Advanced Analytics

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



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

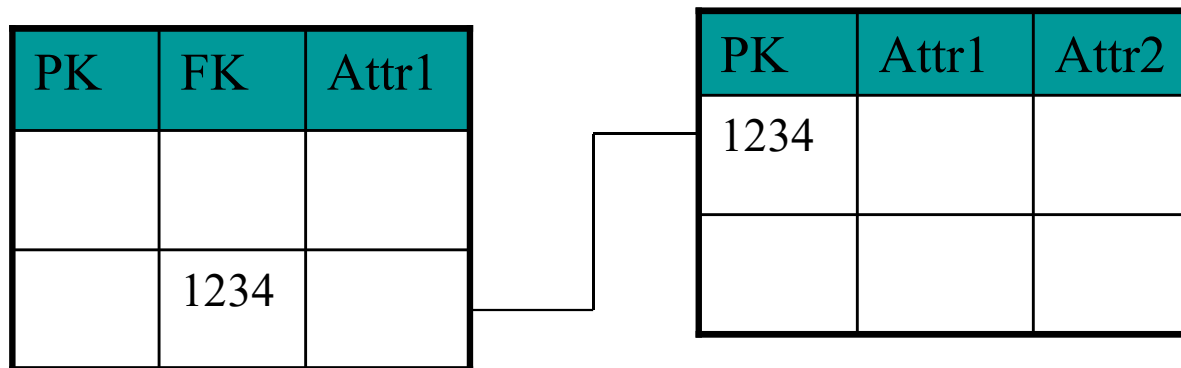
- Organized in the shape of a inverted tree

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

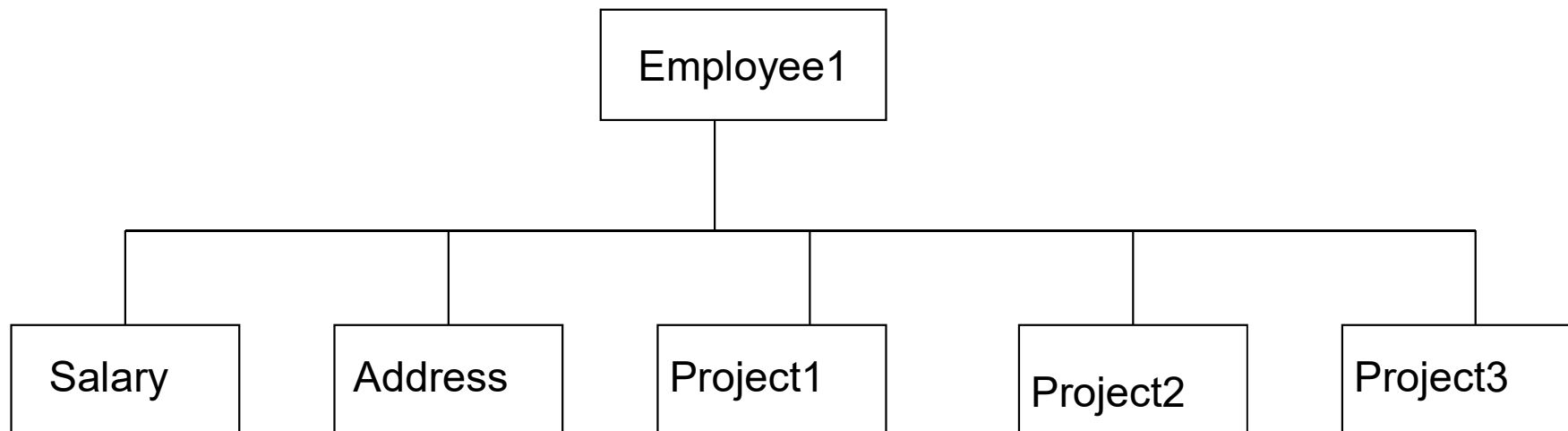


Hierarchical Database Structures

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

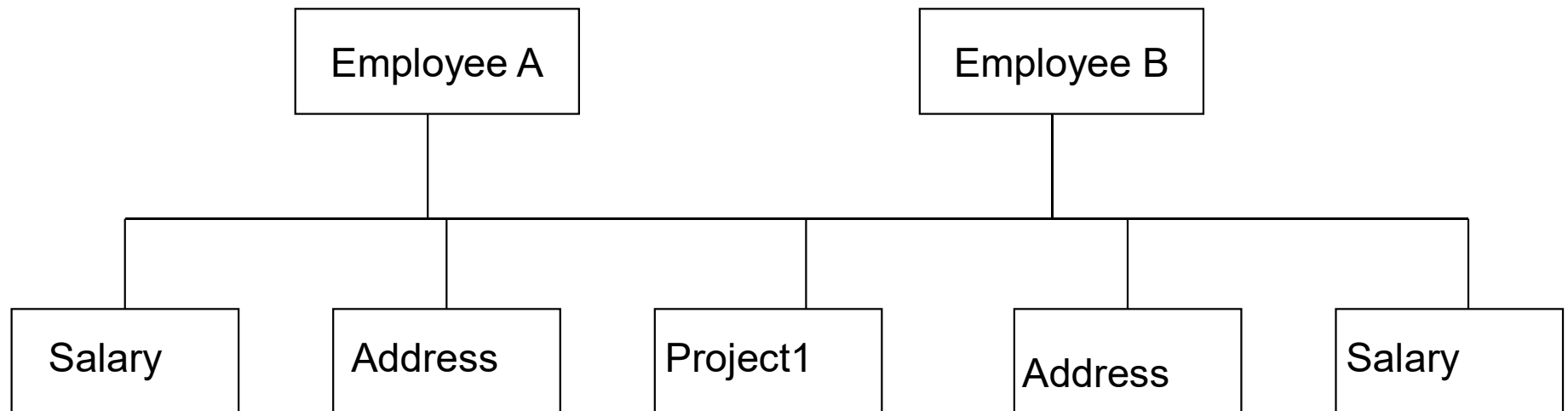
Each record may contain several information parts, for example:

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



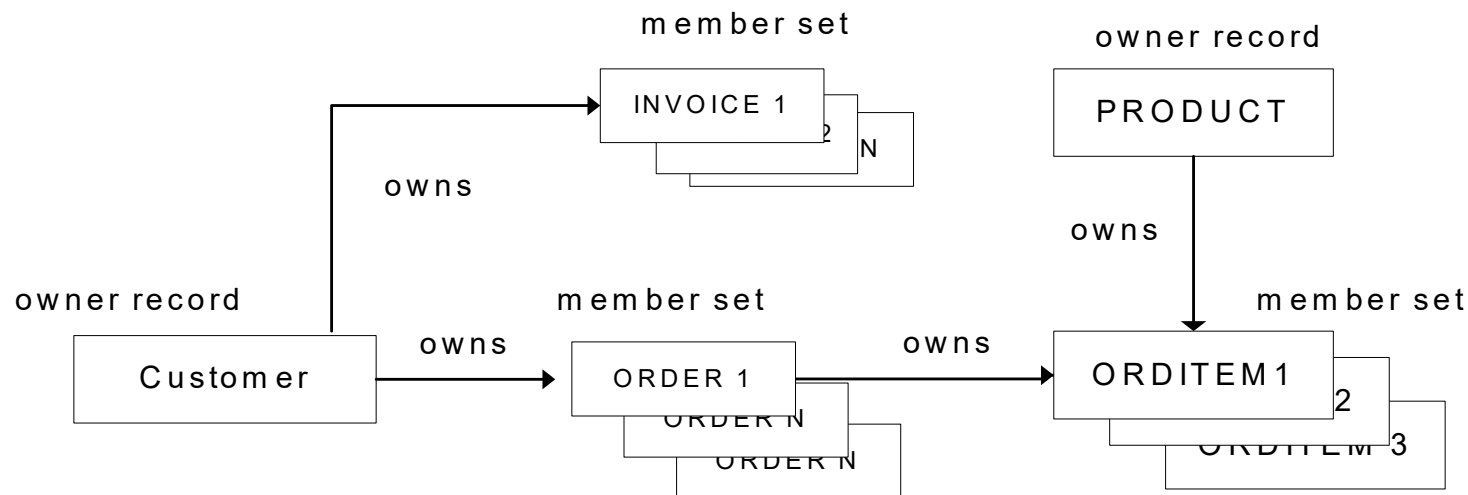
Network Database Structures

- More flexible
- Reduce Redundancy



Example – IDMS Database

- '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 IDMS 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

H y d r o g e n	H	1	1 . 0 0 7 9
H e l i u m	H e	2	4 . 0 0 2 6
L i t h i u m	L i	3	6 . 9 4 1
B e r y l i u m	B e	4	9 . 0 1 2 1 8
B o r o n	B	5	1 0 . 8 1
C a r b o n	C	6	1 2 . 0 1 1
N i t r o g e n	N	7	1 4 . 0 0 6 7
O x y g e n	O	8	1 5 . 9 9 9 4

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 and 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 then 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”:

<u>PO#</u>	<u>SUP#</u>	<u>SupName</u>	<u>Item#</u>	<u>ItemDescription</u>	<u>\$/Unit</u>	<u>Quant</u>
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).

<u>PO#</u>	<u>SUP#</u>	<u>SupName</u>	<u>Item#</u>	<u>ItemDescription</u>	<u>\$/Unit</u>	<u>Quant</u>
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“

<u>PNR</u>	Fach	Prüfer	<u>Student MATNR</u>	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				
<i>CD_ID</i>	Albumtitel	Interpret	<i>Track</i>	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			Lieder		
<i>CD_ID</i>	Albumtitel	Interpret	<i>CD_ID</i>	<i>Track</i>	Titel
4811	Not That Kind	Anastacia	4811	1	Not That Kind
4712	Wish You Were Here	Pink Floyd	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			
<i>CD_ID</i>	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			Künstler	
<i>CD_ID</i>	Albumtitel	Interpret	<i>Interpret</i>	Gründungsjahr
4811	Not That Kind	Anastacia	Anastacia	1999
4713	Bad	Michael Jackson	Michael Jackson	1971
4712	Wish You Were Here	Pink Floyd	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

Exercise / 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

<i>Name</i>	<i>Proj</i>
Haas	AD3100
Thompson	PL2100
Walker	MA2112
Lutz	MA2111

SAMP_STAFF

<i>Name</i>	<i>Job</i>
Haas	<i>PRES</i>
Thompson	<i>MANAGER</i>
Lucchessi	<i>SALESREP</i>
Nicholls	<i>ANALYST</i>

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):

<u>PNR</u>	<u>Fach</u>	<u>Prüfer</u>	<u>Student MATNR</u>	<u>Name</u>	<u>Geb</u>	<u>Adr</u>	<u>Fachbereich</u>	<u>Dekan</u>	<u>Note</u>
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#)

<u>PO#</u>	<u>SUP#</u>	SupName	<u>Item#</u>	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

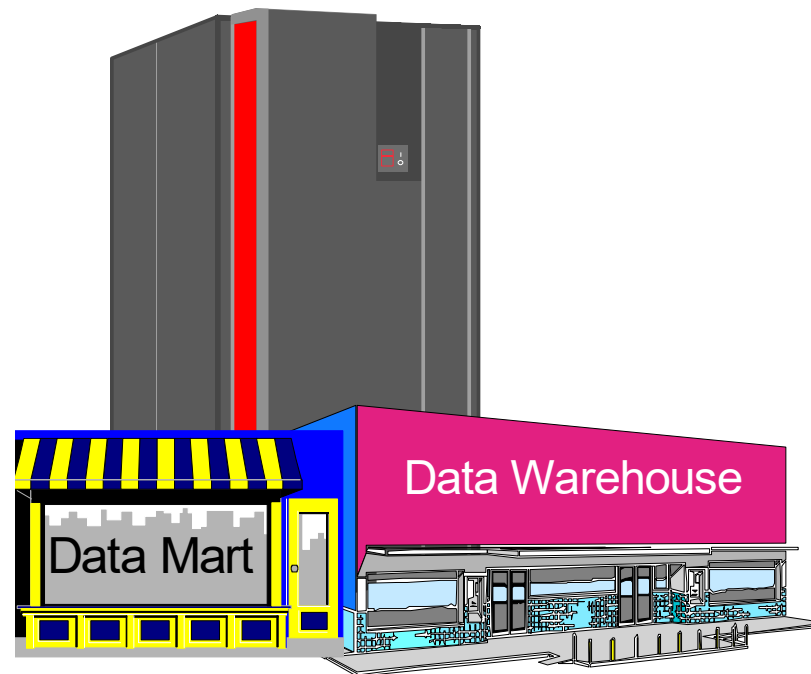
Category 1: Introduction & Architecture of DWH

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,...

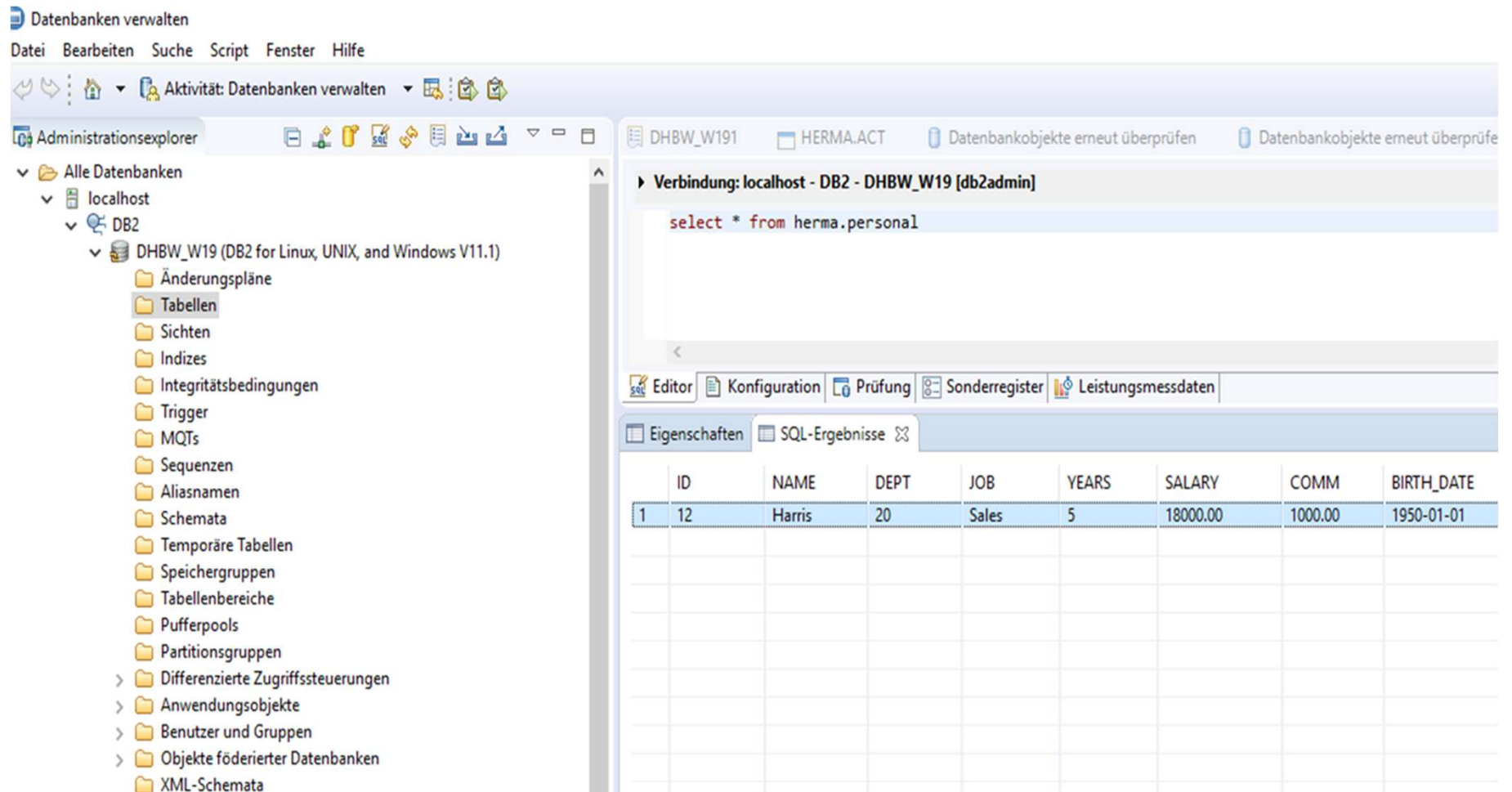
Examples for DCL commands

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.



The screenshot shows the IBM Data Studio interface. On the left, the 'Administration Explorer' tree is expanded to show the 'DHBW_W19 (DB2 for Linux, UNIX, and Windows V11.1)' database. The 'Tabellen' (Tables) folder is selected. The main editor window displays a SQL query: `select * from herma.personal`. Below the query, the 'SQL-Ergebnisse' (SQL Results) tab is active, showing a table with 9 columns: ID, NAME, DEPT, JOB, YEARS, SALARY, COMM, and BIRTH_DATE. The first row of data is highlighted.

	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:

<i>FID</i>	<i>Name</i>
MUC	Muenchen
FRA	Frankfurt
HAN	Hannover
STU	Stuttgart
MAN	Mannheim
BER	Berlin

Flight:

<i>Fno</i>	<i>From</i>	<i>To</i>	<i>Time</i>
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:

Star Tracker

File Edit Aggregates Sequences Comparisons MiniHelp

Time: All Times

Browse Expand

FIELDS GROUPS STUDIES

Date
Day_Number_In_Month
Day_Number_Overall
Day_Of_Week
Fiscal_Period
Holiday_Flag

Promotion: All Promotions

Browse Expand

FIELDS GROUPS STUDIES

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*

Run Report

Product: All Products

Browse Expand

FIELDS GROUPS STUDIES

Brand
Cases_Per_Pallet
Category
Department
Description
Diet_Type

Store: All Stores

Browse Expand

FIELDS GROUPS STUDIES

City
Finance_Services_Type
First_Opened_Date
Floor_Plan_Type
Frozen_Sqft
Grocery_Sqft

No Constraints

	A	B	C	D	E	F	G	H	I
1									
2									
3									
4									
5									
6									
7									

Start | DWH1 | Microsoft Power... | Database Finder | Star Tracker | 18:37

Exercise 2 to Lesson 4 (Part 2)

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

Remark: You can solve this exercise with a database (DB2, Oracle, etc.) or by using an analytical platform, like KNIME.

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

Solution with MS Access SQL Wizard

Microsoft Access

File Edit View Insert Query Tools Window Help

grocery-01 : Database

Objects

- Tables
- Queries
- Forms
- Reports
- Pages
- Macros
- Modules

Groups

- Favorites

Create query in Design view

Create query by using wizard

Query1

Report1

Sales Fact Query

Query1 : Select Query

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:						

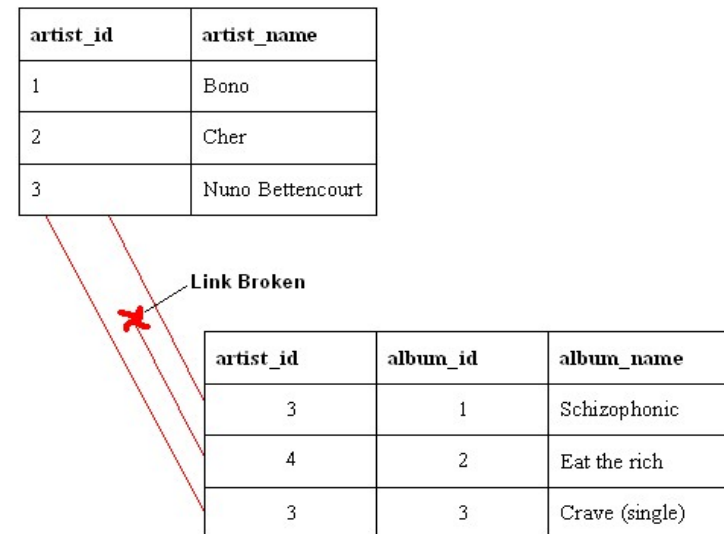
Ready

Exercise 3 to Lesson 4

Advanced Study about concepts in DWH:

Exercise E4.3 (SW*):

*Explain what is “Referential Integrity” (RI)
in a Database?*



Sub-Questions:

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

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

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

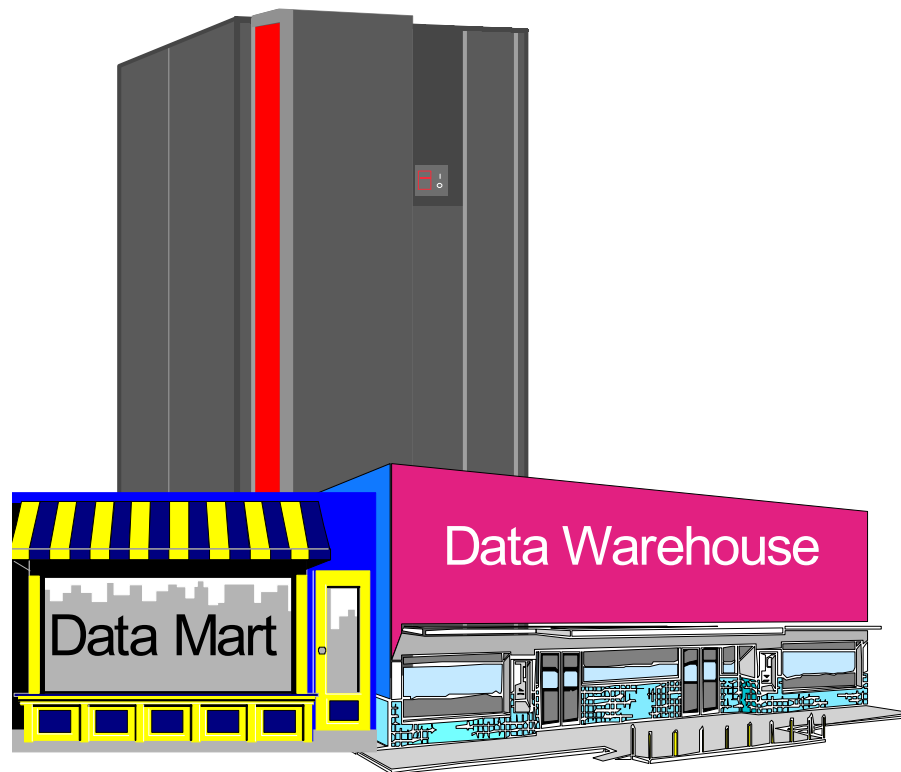
Category 1: Introduction & Architecture of DWH

Category 2: Databases and Data Modeling

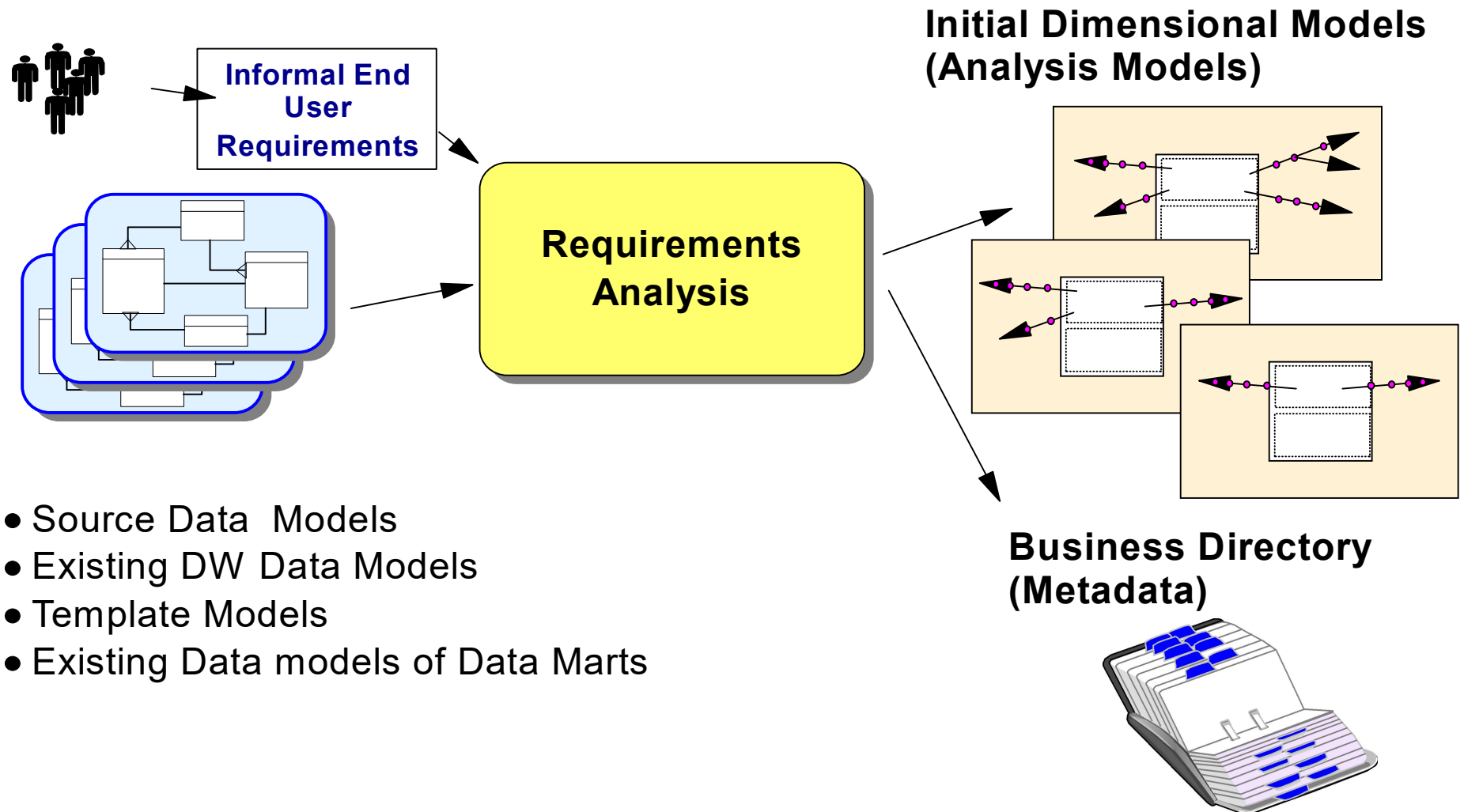
Category 3: ETL: Architecture & Technology

Category 4: Descriptive – & Advanced Analytics

DW05 – Multi-Dimensional Data Modeling



Requirements Analysis- Context



Requirements Analysis - Activities

Informal End-User Requirements

Process-Oriented Requirements

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

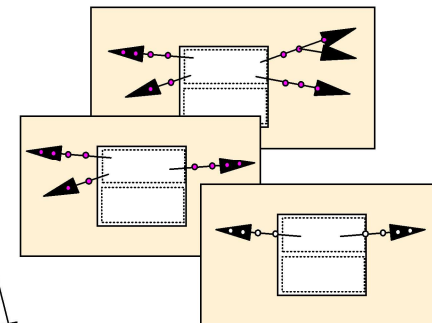
Information Oriented Requirements

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

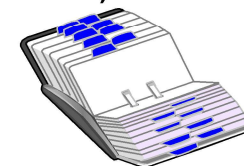
Requirements Analysis

- 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)

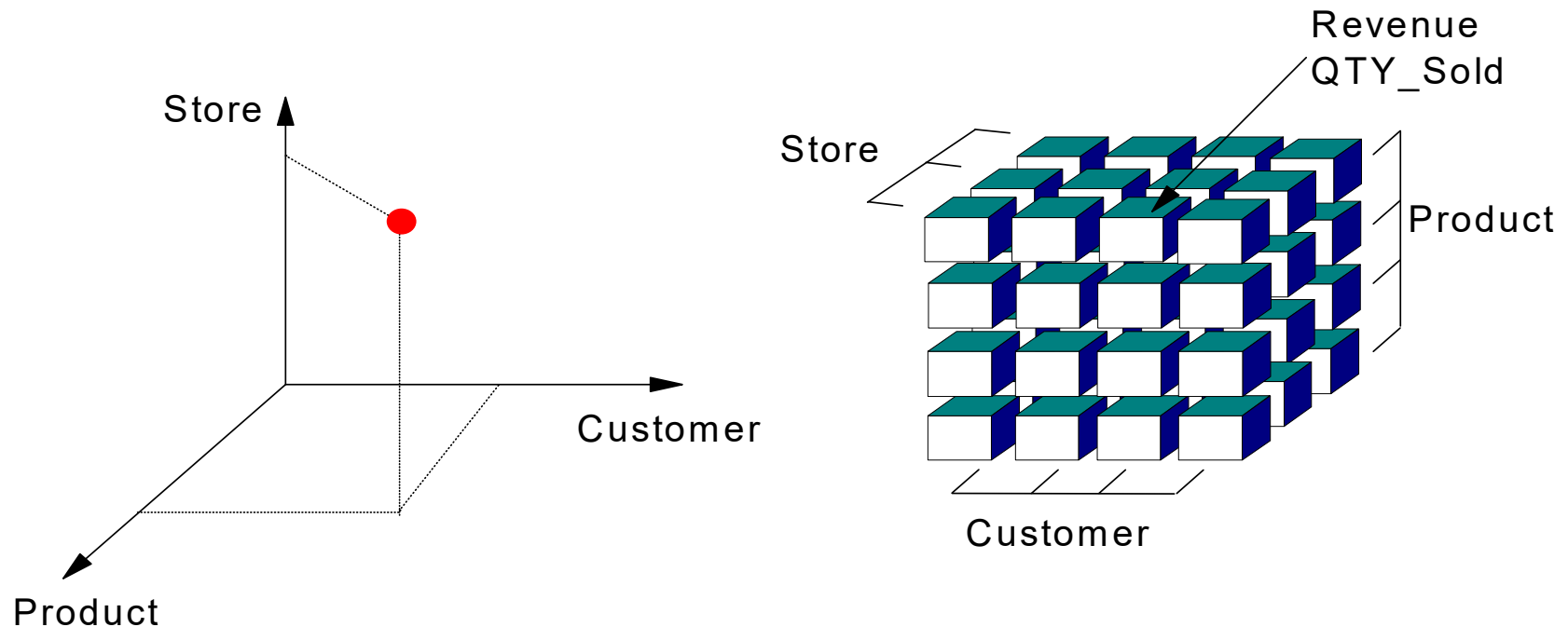


Sample Query

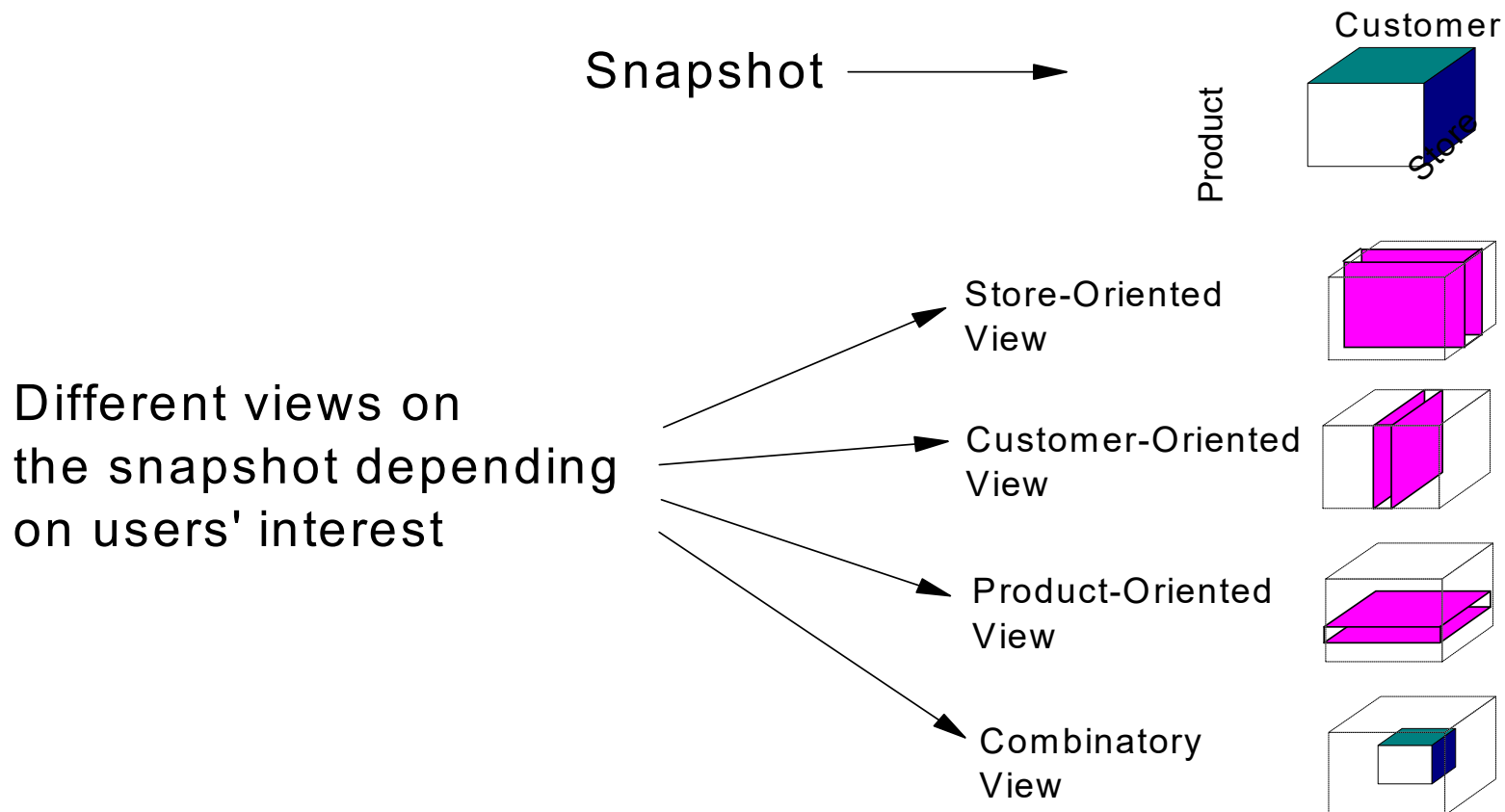
- Query:
"What are the net sales, in terms of revenue (dollars) and quantities of items sold,
Per product,
Per store and sales region,
Per customer and customer sales area,
Per day as well as aggregated over time,
Over the last two weeks?"
- Evaluation entails viewing historical sales figures from multiple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

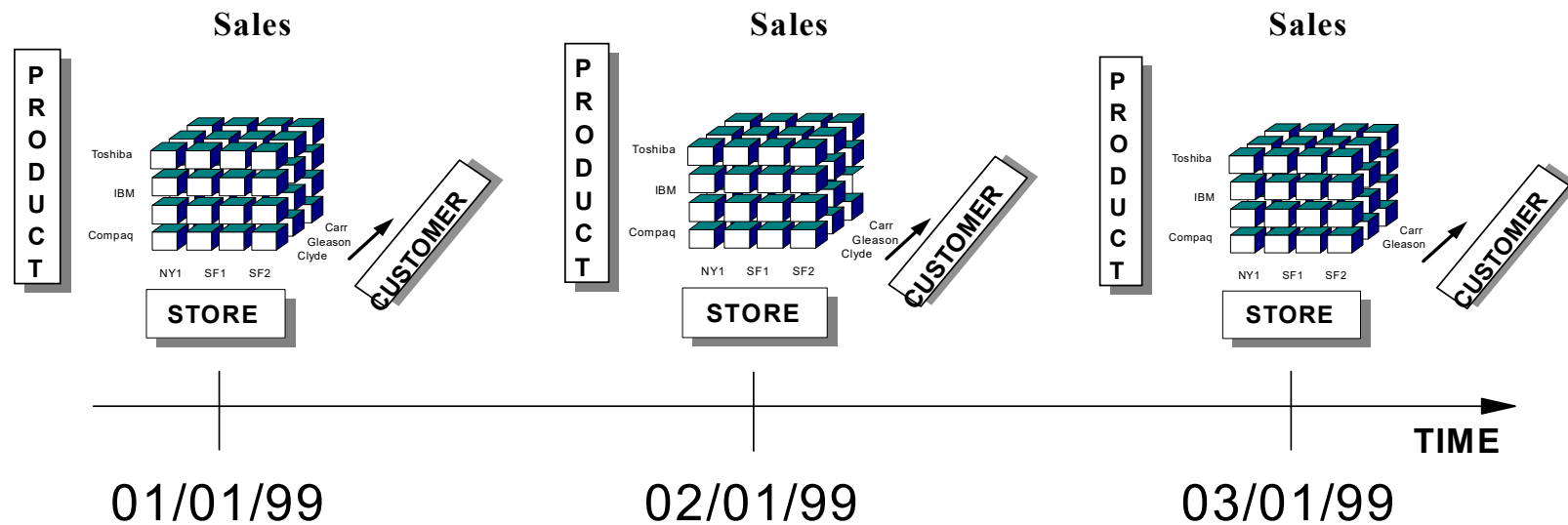


Presentation of the Query as a Cube : Usage



Hypercube Representation

(4th dimension)

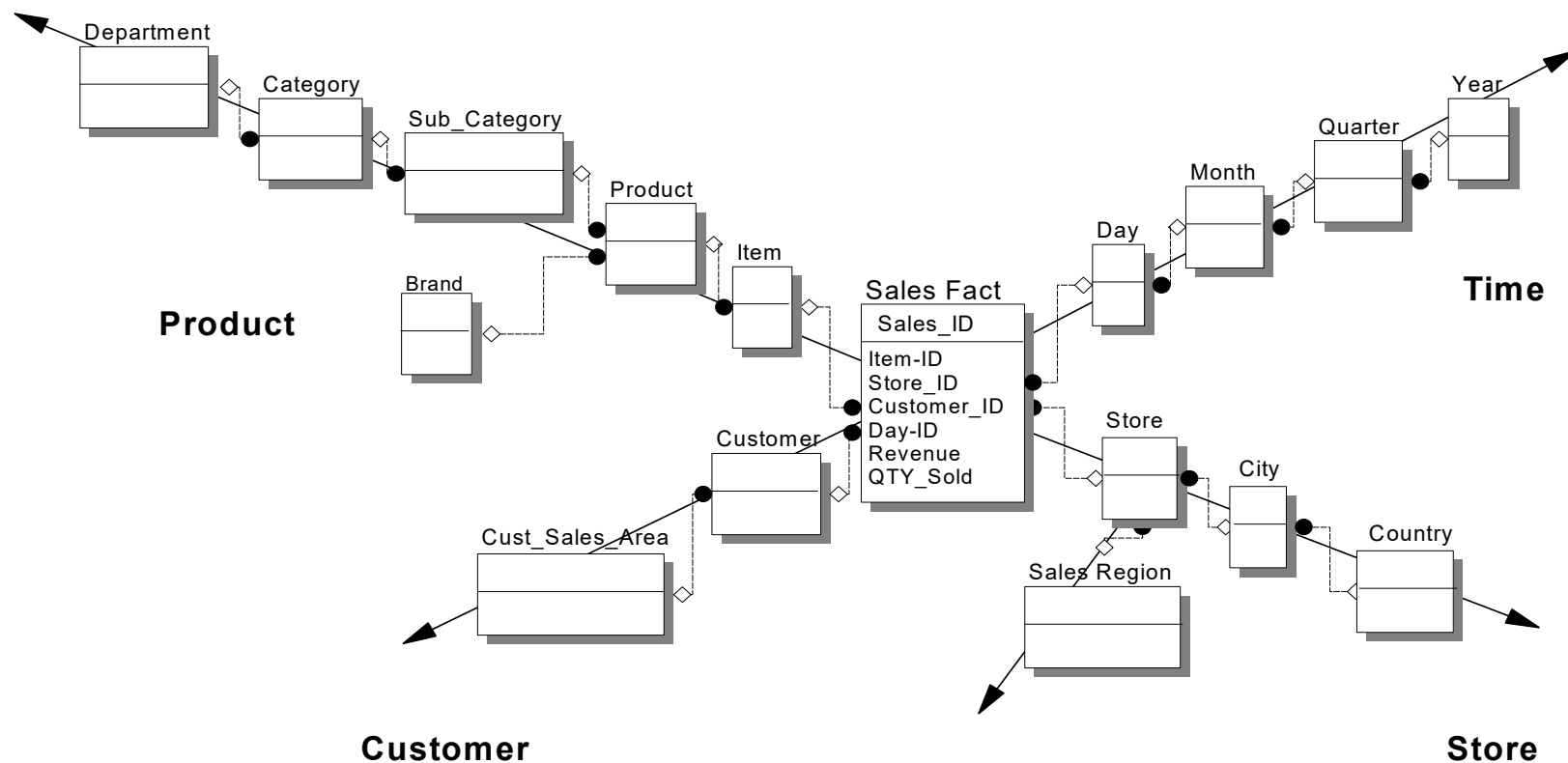


Hypercube:

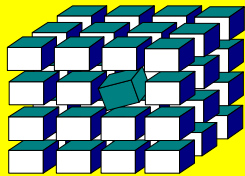
Good visual representation for three dimensions

Difficult to use, when more than four dimensions

Sample Multidimensional Representation Usable for Any Number of Dimensions



The Six Base Concepts of MDDM

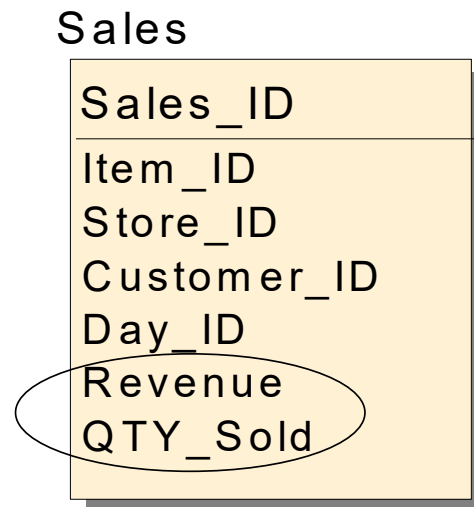


- **Measures**
- **Dimensions**
- **Granularity**
- **Facts**
- **Dimension Hierarchies**
- **Aggregation Levels**

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

Measures



Identify Candidate Measures

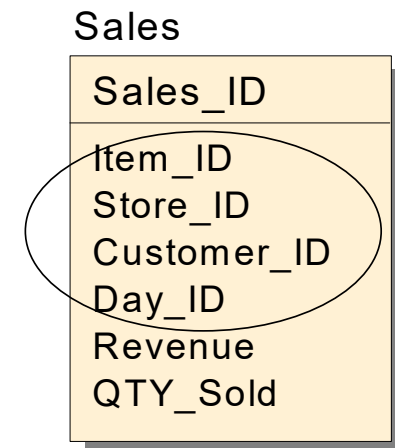
- Query-Oriented Approach
 - Perform a smart, not a mechanical analysis of the available queries
- Candidate Measures are
 - **Numeric, "Continuously" Valued**
 - But not every numeric attribute is a candidate measure
 - Distinguish measures from discrete valued numeric attributes which are part of dimensions
 - **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



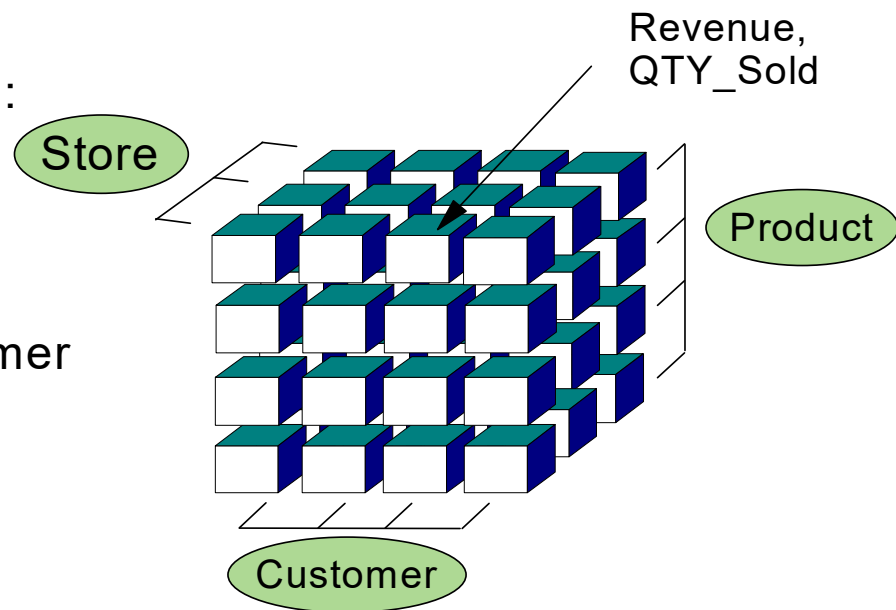
Identify Candidate Dimensions

- 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

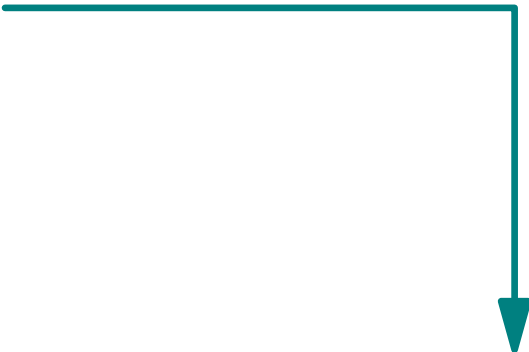
Granularity

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

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

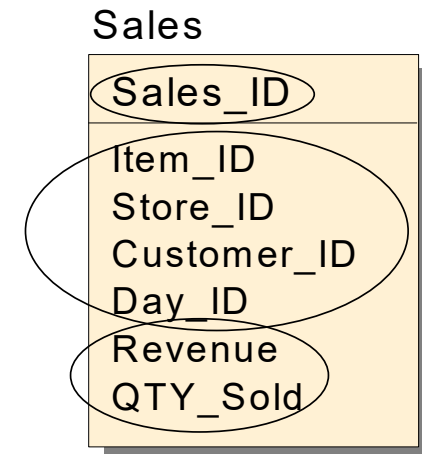


Sales Region1

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

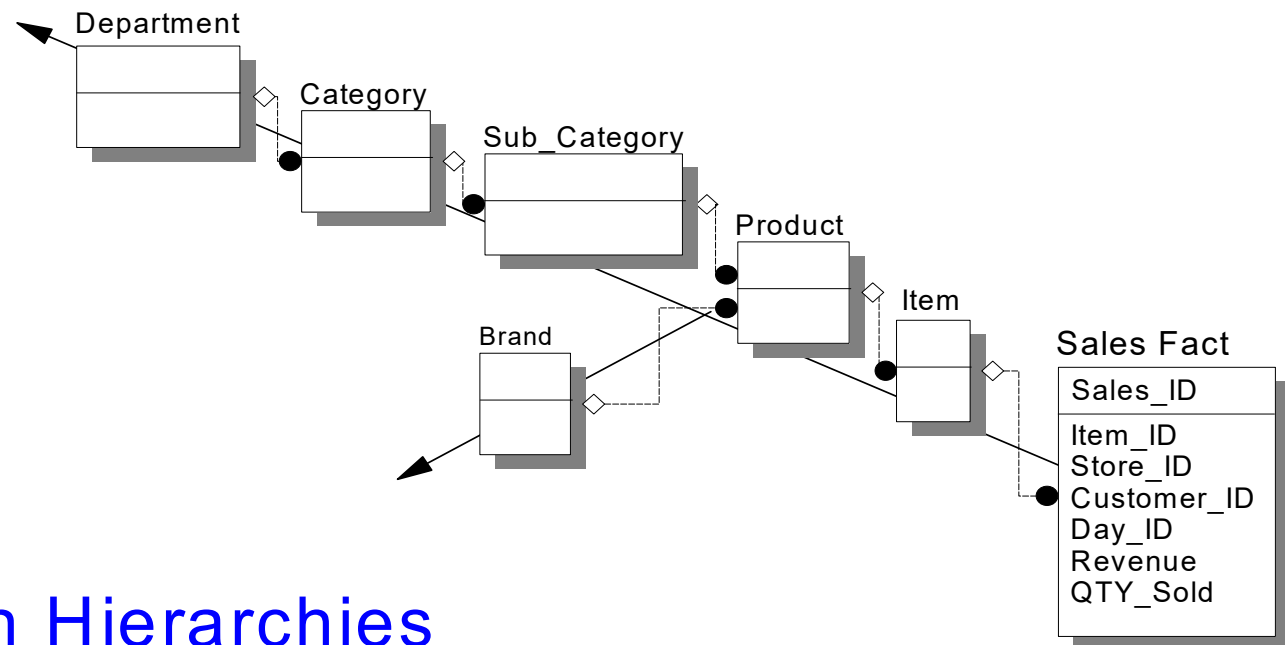
	Dimension 1	Dimension 2	Product	Customer	Store	Time	(...)
Measure 1							
Measure 2							
Revenue			Item	Customer	Store	Day	
Quantity Sold			Item	Customer	Store	Day	
Measure 3							
(...)							

FACT

Facts

Multidimensional Data Modeling - Base Concepts (5 of 6)

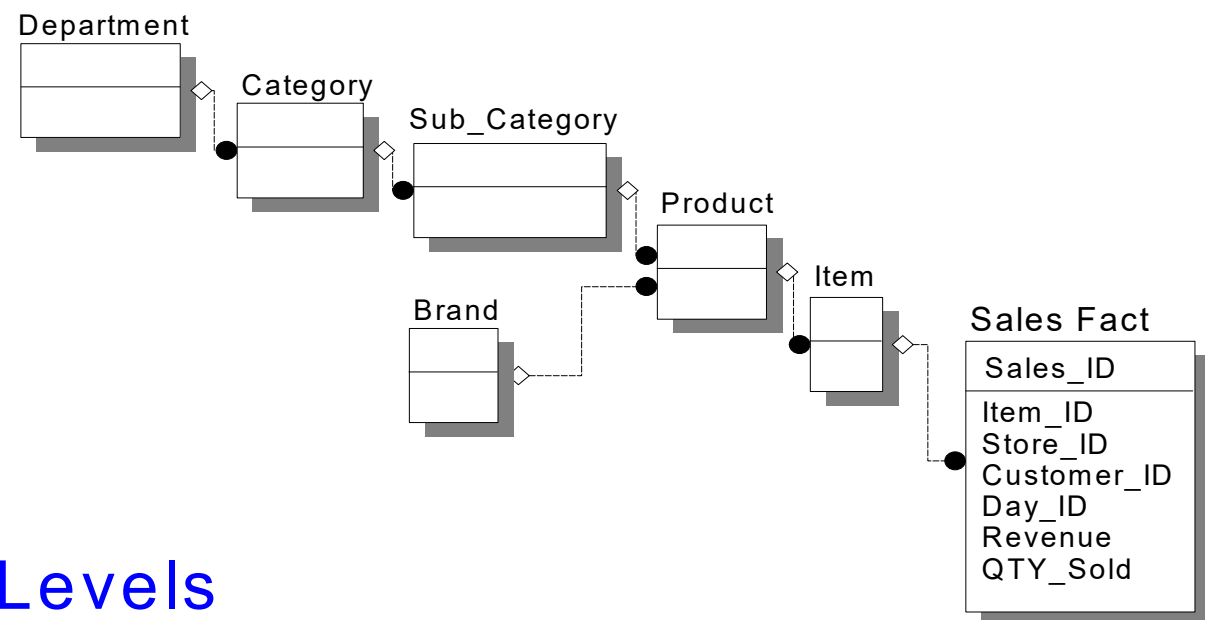
- Dimensions consist of one or more **dimension hierarchies**
- Examples: Hierarchies in the Product Dimension
 - Product Classification Hierarchy ("Merchandising Hierarchy")
 - Branding Hierarchy
 - ...



Dimension Hierarchies

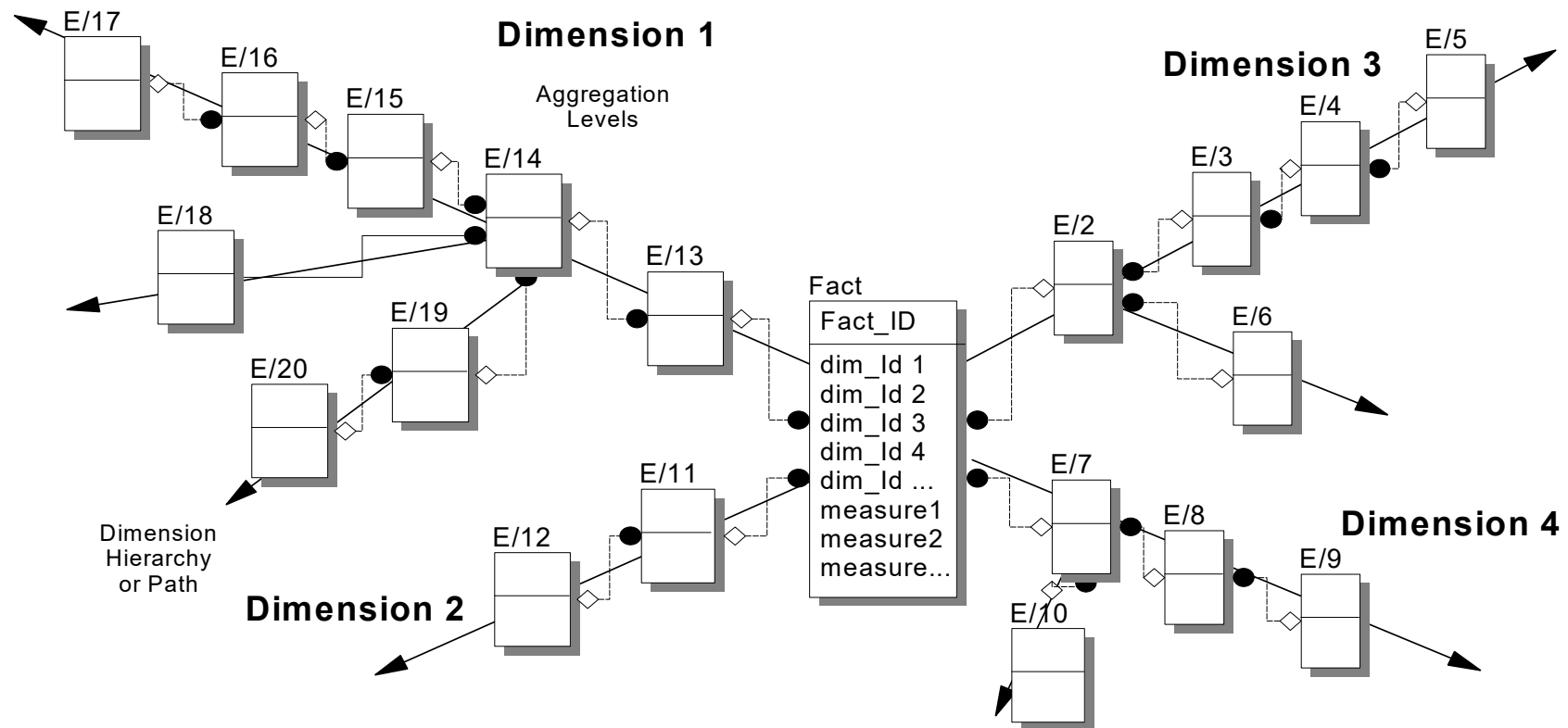
Multidimensional Data Modeling - Base Concepts (6 of 6)

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



Aggregation Levels

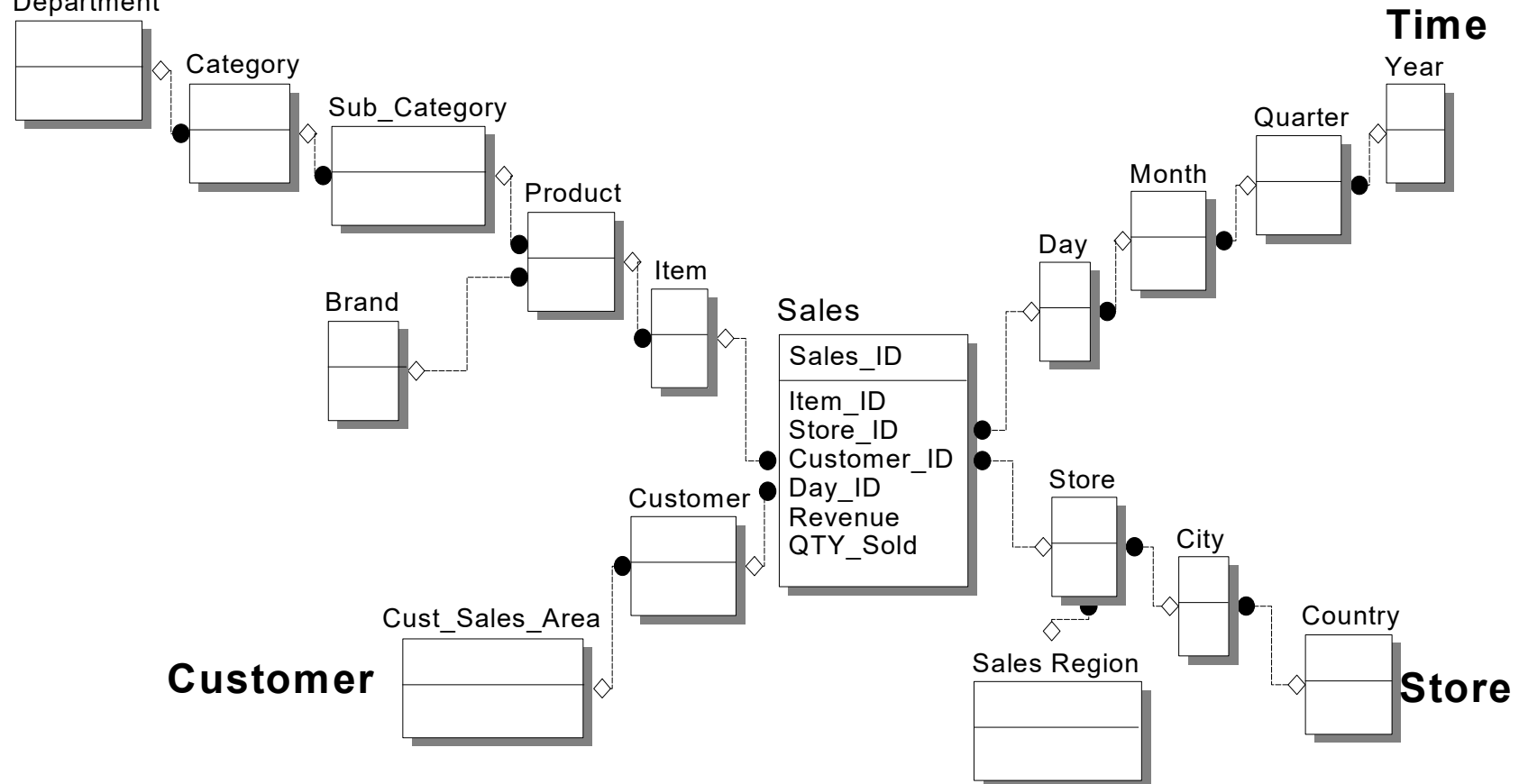
Initial Multidimensional Model - Summary



Initial Multidimensional Model - Example

Product

Department



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

Product
Dimension
Table

ID	NAME	...
1	Products	
2	Skateboards	
3	Bicycles	
4	Tricycles	

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

Demo1: IBM InfoSphere Data Architect (IDA)

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

File Bearbeiten Diagramm Navigieren Suche Projekt Daten Ausführen Fenster Hilfe

Aktivität: Datenmodellierung Tahoma 75%

Lizenzen verwalten

Datenprojektoexplorer

- GLS Logistic DWH
- Henkel-DataVault Reference
- Henkel Future BI POC
 - XML-Schemata
 - Zuordnungen
 - Datendiagramme
 - Andere Dateien
 - Datenmodelle
 - POC 1 STA.dbm
 - POC 1 STA.Idm
 - POC 2 RDV.dbm
 - POC 2 RDV.Idm
 - POC 3 BDV.dbm
 - POC 3 BDV.Idm
 - POC 5 DMA Billing.dbm
 - POC 5 DMA Billing.Idm
 - DM
 - Diagramme
 - DM1 Star Schema
 - <<Domains>>
 - D_BILL_INFO
 - D_COMPANY_CODE
 - D_MAT
 - D_MAT_PLANT
 - Dim Country of Destination
 - Dim Plant
 - Dim Sales Information
 - Dim Time
 - F_BILL_ITEMS

SQL-Scripts

IDA Hands on Lab

Database Connections

- BLUDB
- Derby Sample Connection
- DHBW_W19 (DB2 for Linux, UNIX, ar)
 - DHBW_W19
 - Speicherdiagramme
 - Groups
 - Objekte federierter Datenban
 - Partitionsconnonen

DataMart Lay... KARISMA-BOM.Idm Business Mo... "POC 5 DMA B..." "DM1 Star Sc..."

Diagramm: DM1 Star Schema

Central Fact Table: F_BILL_ITEMS

- Attributes: Net Value DC, Gross Value DC, Net Value Euro, Gross Value Euro
- Dimensions:
 - Dim Country of Destination (FK1)
 - Dim Time (FK1)
 - Dim Sales Information (FK1)
 - Dim Plant (FK1)

Relationships: Billing Items, Company Code, Billing Items, Time, Billing Items, Sales Information, Billing Items, Material, Billing Items, Plant

Eigenschaften

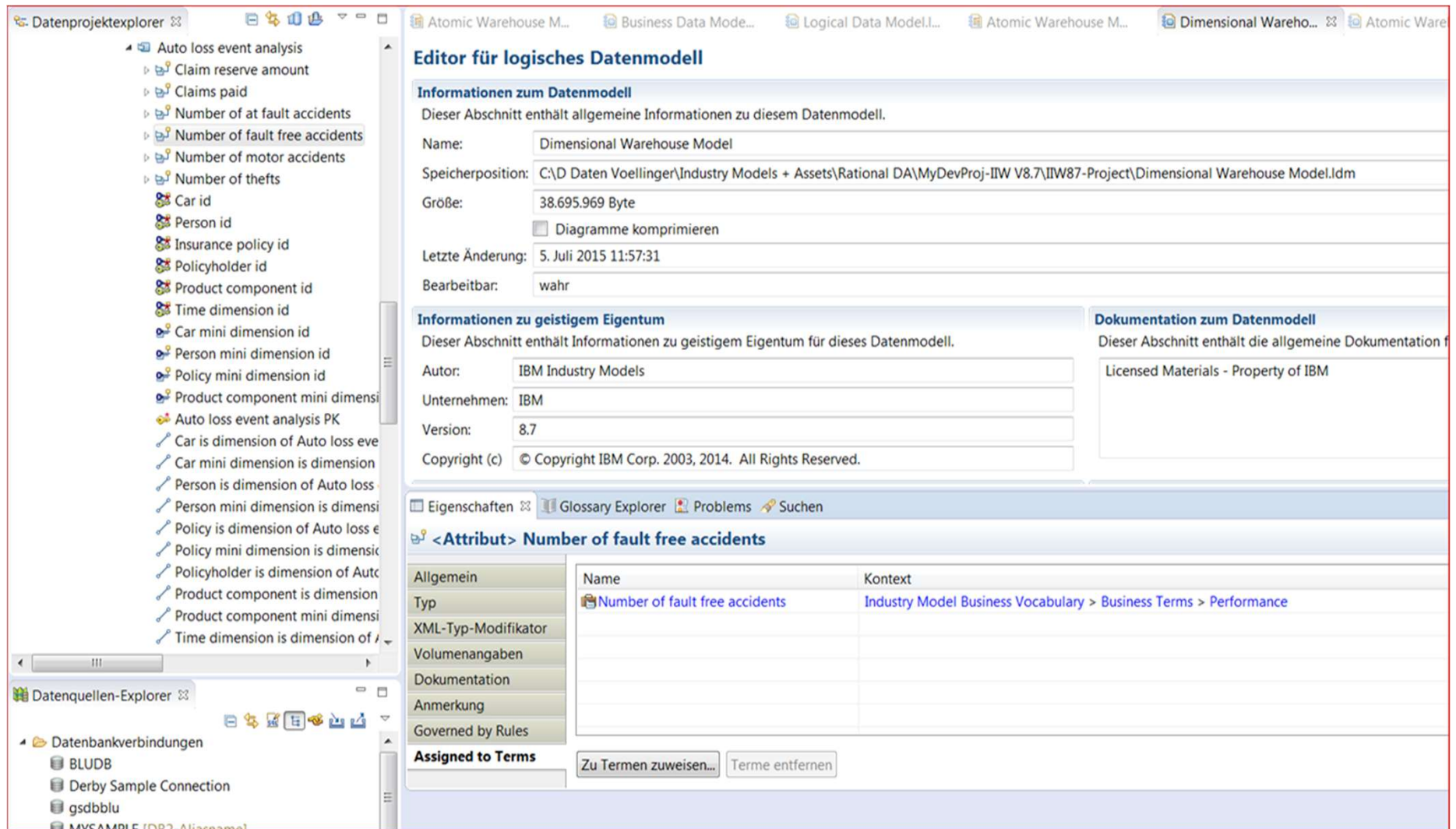
<Entität> F_BILL_ITEMS

Attribut	Name	Primärschlüssel	Ersatzschlüssel	Typ	Länge/Genauigkeit	Anzahl	Erfordert	Alt
Generalisierungssets	F_TIK_BILL_ITEMS	☒	☐	VARCHAR	32		☒	☐
Generalisierungen	Net Value DC	☐	☐	NUMERIC	15	30	☐	☐
Gruppenverweise	Gross Value DC	☐	☐	NUMERIC	15	3	☐	☐
Beziehungen	Net Value Euro	☐	☐	NUMERIC	15	3	☐	☐
Volumenangaben	Gross Value Euro	☐	☐	NUMERIC	15	3	☐	☐
Dokumentation	VBELN	☐	☐	VARCHAR	10		☒	☐
Anmerkung	BUKRS	☐	☐	VARCHAR	4		☒	☐
Erweitert	POSNR	☐	☐	VARCHAR	6		☒	☐
	FKART	☐	☐	VARCHAR	4		☒	☐
	Tech Lauf ID	☐	☐	NUMERIC	19	0	☒	☐
	Tech Liefer ID	☐	☐	NUMERIC	19	0	☒	☐
	Quellsystem	☐	☐	VARCHAR	100		☒	☐
	D_KEY_SALES_INFO	☐	☐	VARCHAR	32		☒	☐
	D_KEY_TIME	☐	☐	VARCHAR	32		☒	☐
	D_KEY_BUKR	☐	☐	VARCHAR	32		☒	☐
	D_KEY_COUNTRY_DE...	☐	☐	VARCHAR	32		☒	☐
	D_KEY_PLANT	☐	☐	VARCHAR	32		☒	☐
	D_KEY_BILL_INFO	☐	☐	VARCHAR	32		☒	☐
	D_KEY_MATERIAL	☐	☐	VARCHAR	32		☒	☐
	D_KEY_MAT_PLANT	☐	☐	VARCHAR	32		☒	☐
	Billing Item Id	☐	☐	CHAR	5		☐	☐

Temporale Attribute:

☐ Periode für Systemzeit ☐ Periode für Geschäftszeit ☐ Geschäftsperiode ohne Überlappung

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.Idm
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 für dieses Datenmodell.

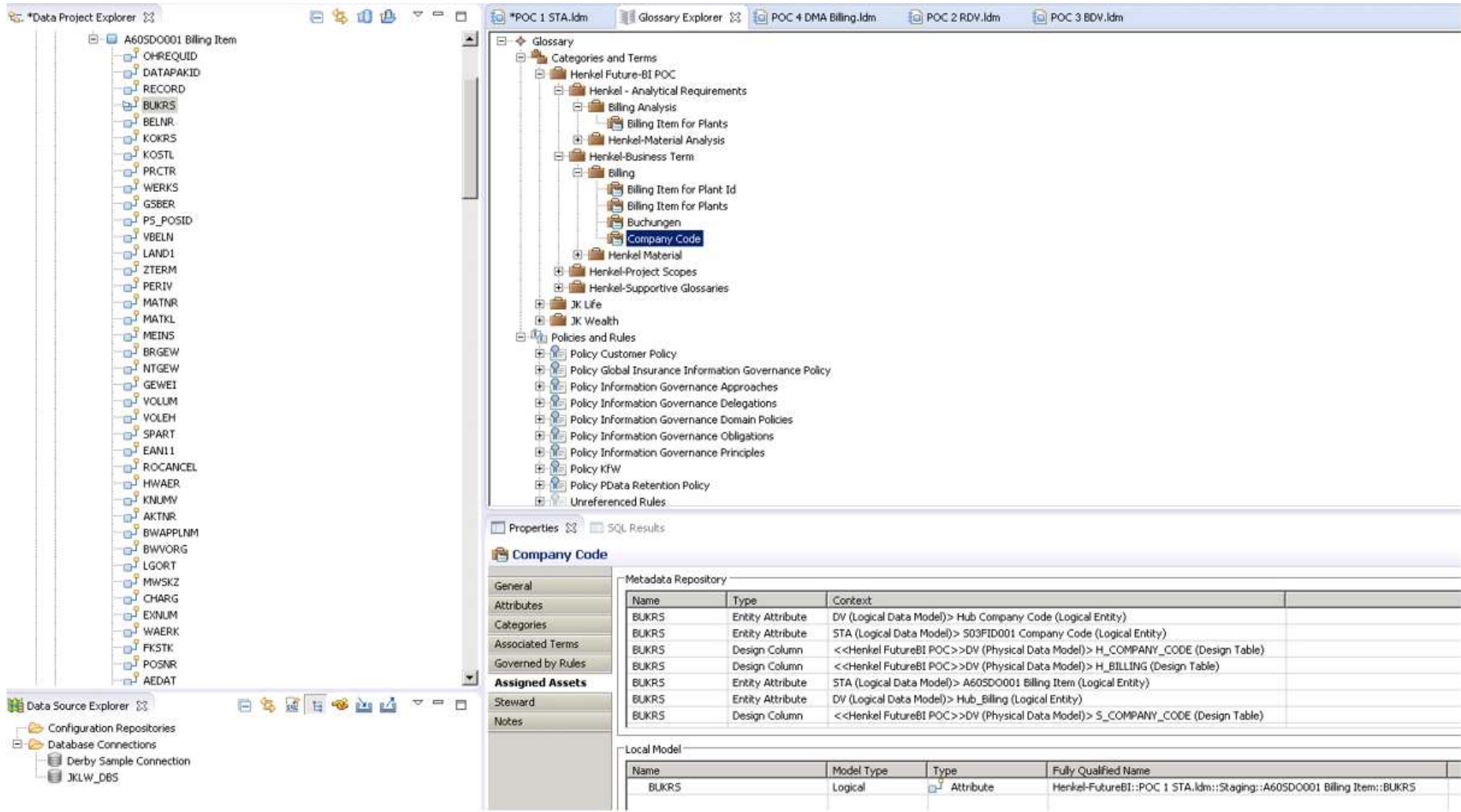
Licensed Materials - Property of IBM

<Attribut> Number of fault free accidents

Allgemein	Name	Kontext
Typ	Number of fault free accidents	Industry Model Business Vocabulary > Business Terms > Performance
XML-Typ-Modifikator		
Volumenangaben		
Dokumentation		
Anmerkung		
Governed by Rules		

Assigned to Terms

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

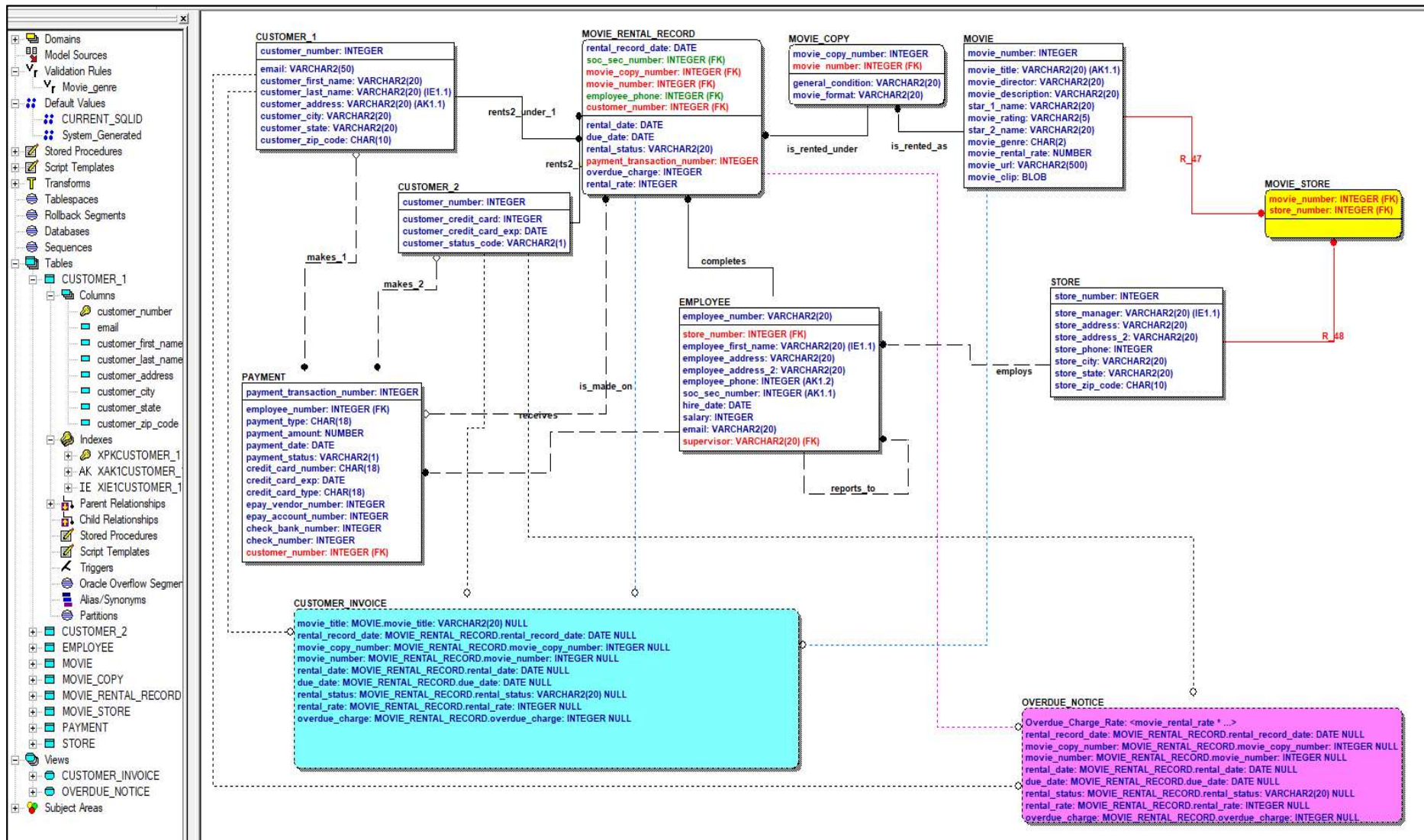


The screenshot displays the IBM Data Architect (IDA) interface. On the left, the 'Data Project Explorer' shows a project structure with a folder 'A60SD0001 Billing Item' containing various data objects. The main pane shows the 'Glossary' view for the project 'POC 1 STA.Idm'. The 'Company Code' term is highlighted under the 'Henkel Future-BI POC' category. The 'Properties' pane on the right shows the 'Company Code' term details, including its 'Assigned Assets'.

Assigned Assets Table:

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) > A60SD0001 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)

The 'Local Model' section at the bottom shows the 'Company Code' term as a 'Logical' entity with an 'Attribute' type, fully qualified name 'Henkel-FutureBI::POC 1 STA.Idm::Staging::A60SD0001 Billing Item::BUKRS'.



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 Redbook 'Data Modeling Techniques for DWH' (see DWH lesson homepage) Chapter 6.3 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 Redbook '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.

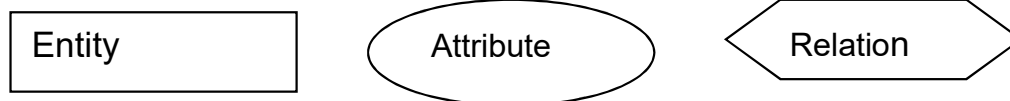
The following objects should be administered by the new ordering system.

- **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

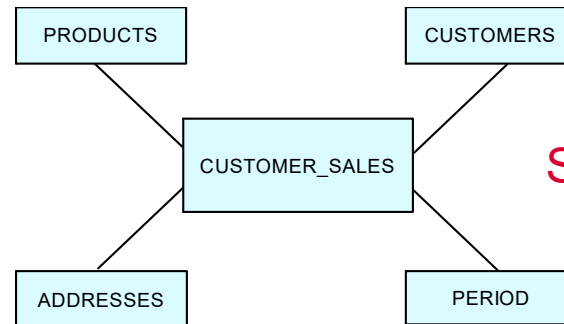
Conditions: Each article can be delivered by one or more suppliers. Each supplier delivers 1 to 10 articles. An order consists of 2 to 10 articles. Each article can only be one time on an order form. But you can order more than one piece of an article. Each order is done by a customer. Customer can have more than one order (no limit).

Good customers will get a 'rabatt'. The number of articles in the store should also be saved. It is 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:



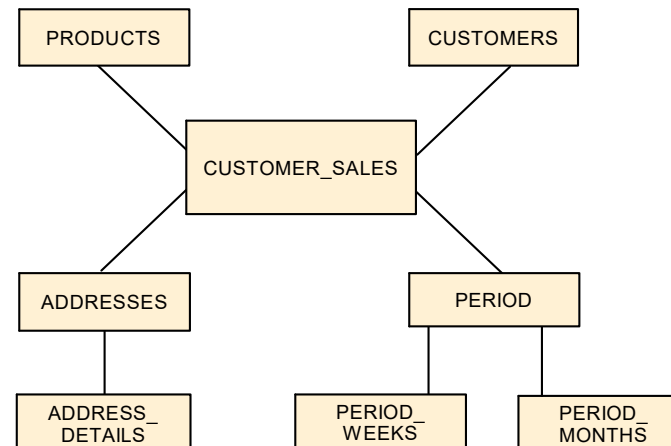
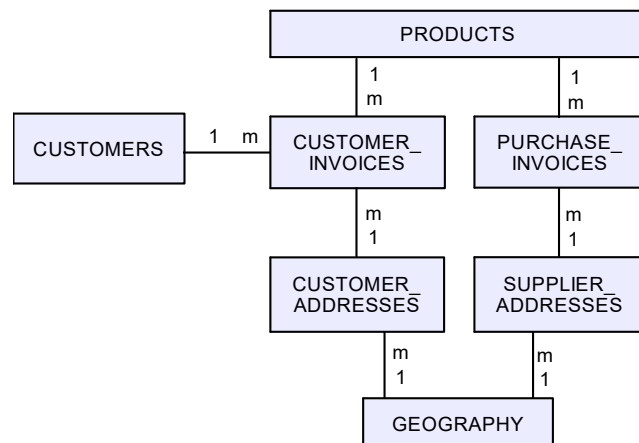
Appendix to MDDM Lesson Exercises



Star Schema



Entity-Relationship



Snowflake Schema



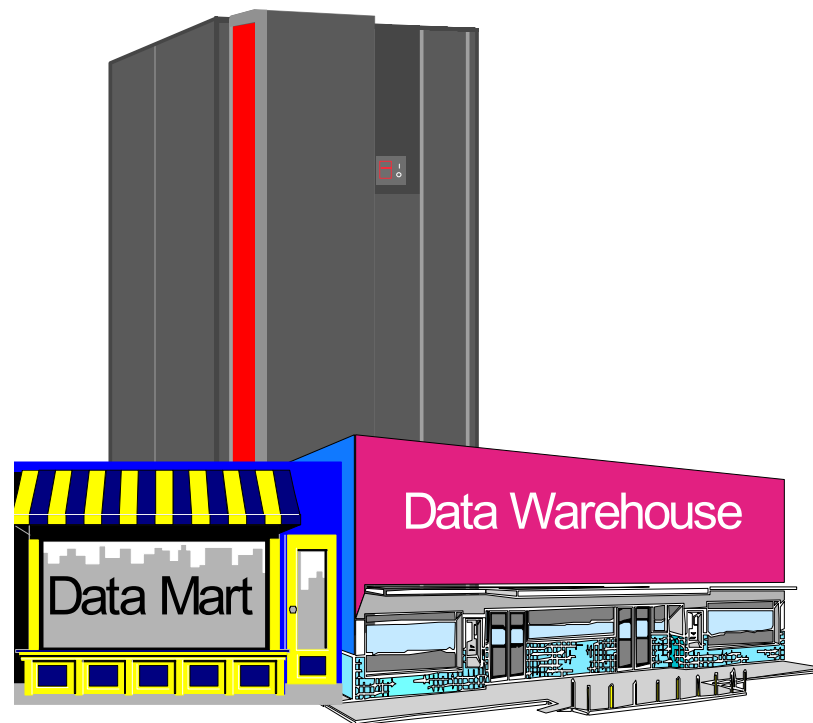
Category 1: Introduction & Architecture of DWH

Category 2: Databases and Data Modeling

Category 3: ETL: Architecture & Technology

Category 4: Descriptive – & Advanced Analytics

DW06 - ETL Reference Architecture

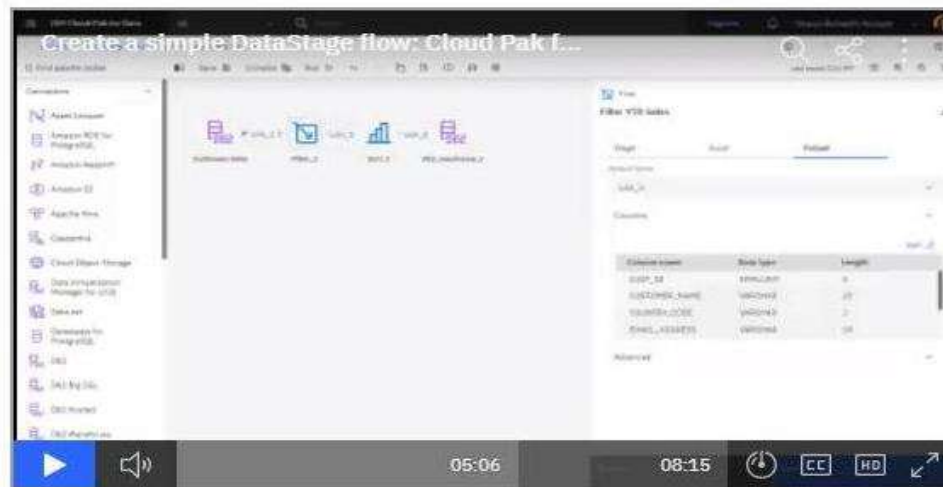


Motivation: Demo (20 Minutes) IBM Cloud Pak for Data - DataStage

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: <http://ibm.biz/cpdaas-videos>
<https://dataplatform.cloud.ibm.com/docs/content/wsj/getting-started/videos.html?audience=cpdaas&context=cpdaas#data-engineers>

Experience shows that ...

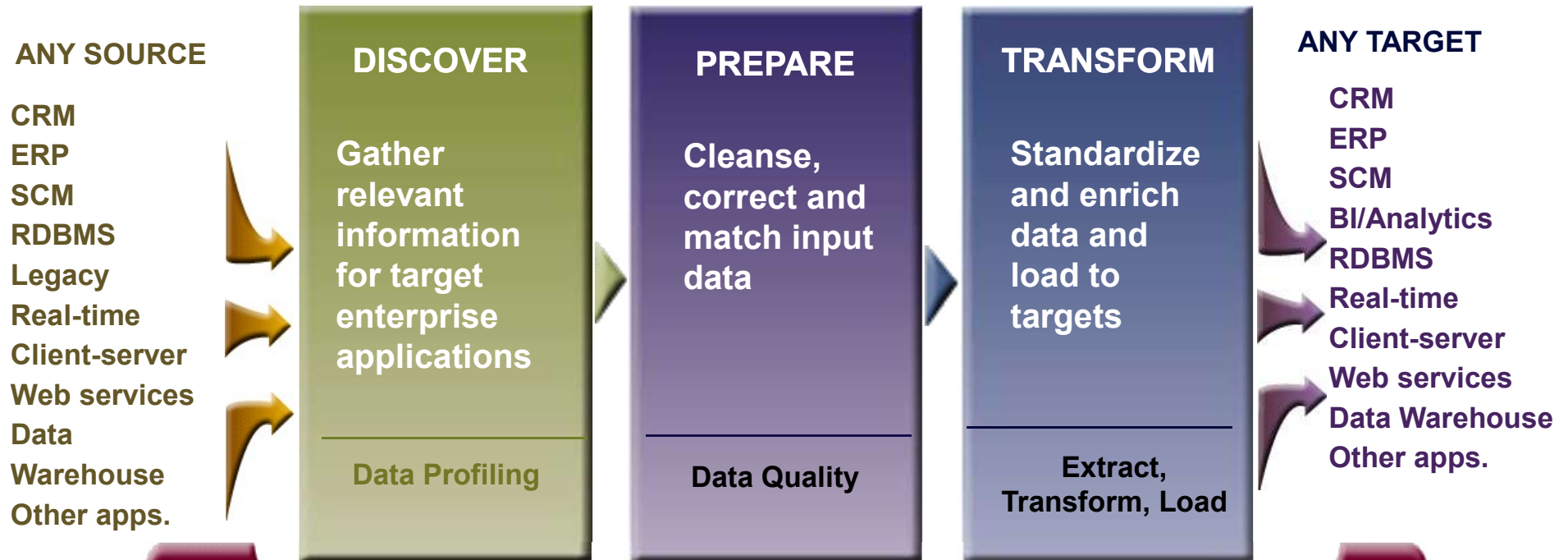
80% of the cost of building and maintaining a Data Warehouse Environment usually relates to the Populating Subsystem ...

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

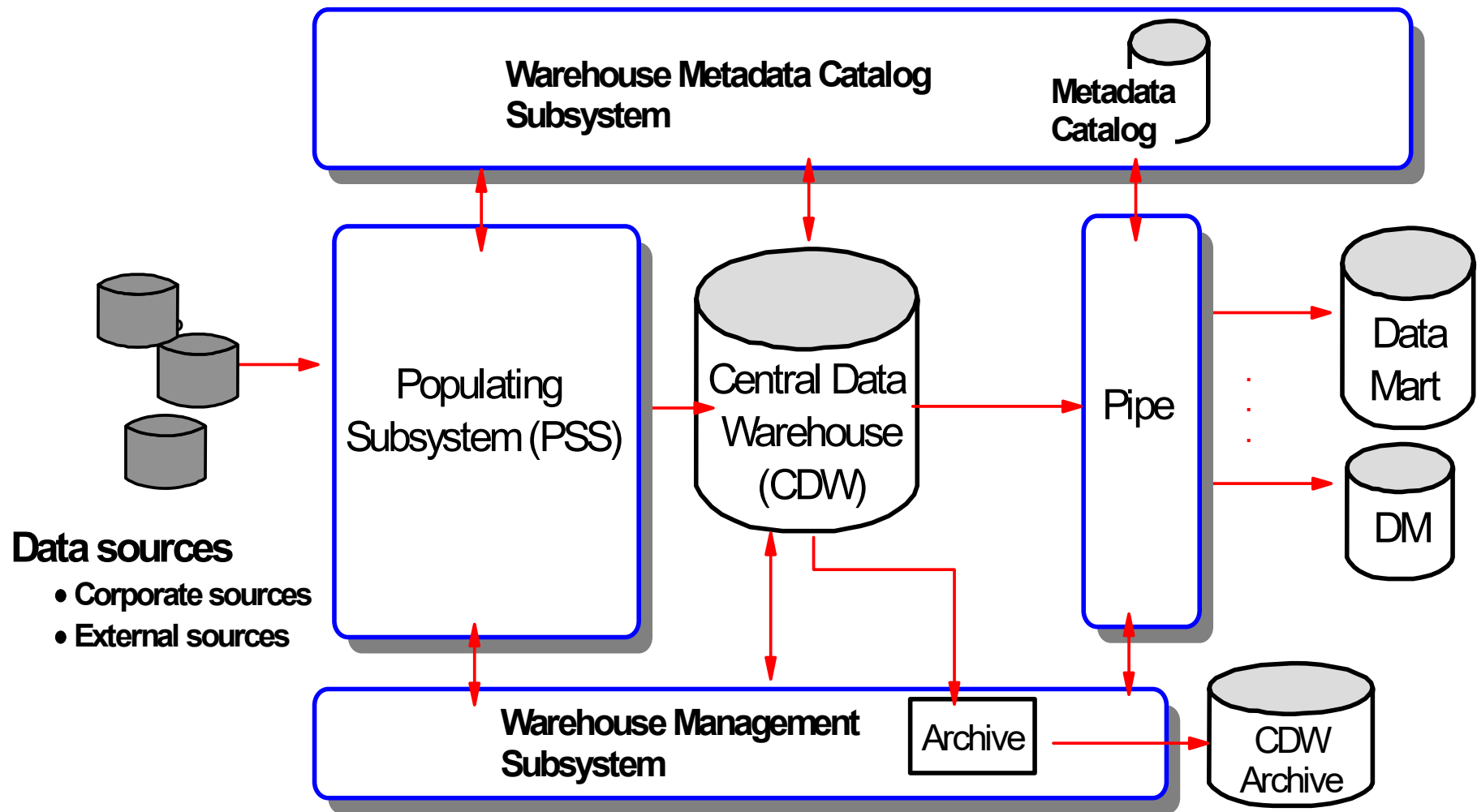


Parallel Execution

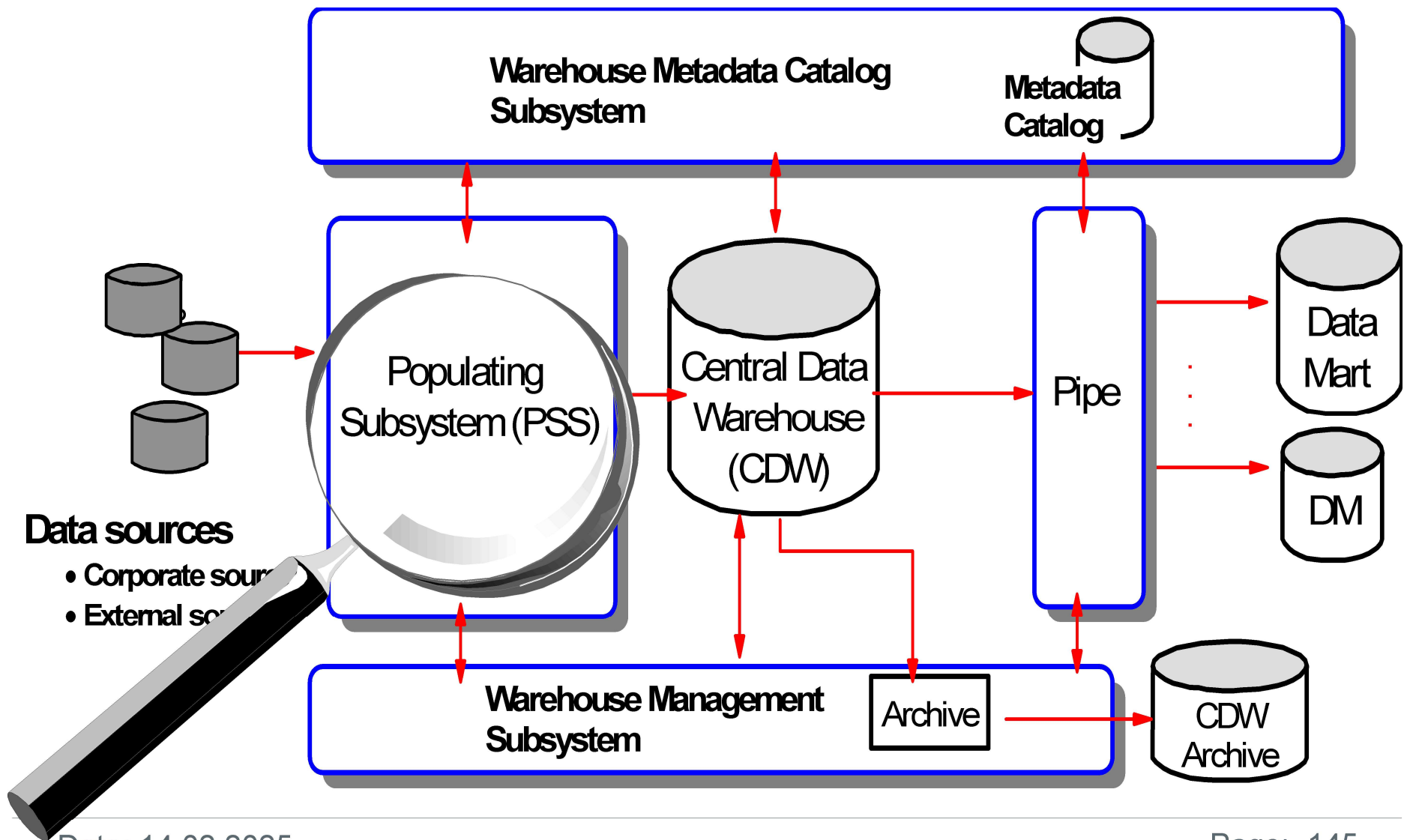
Meta Data Management

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

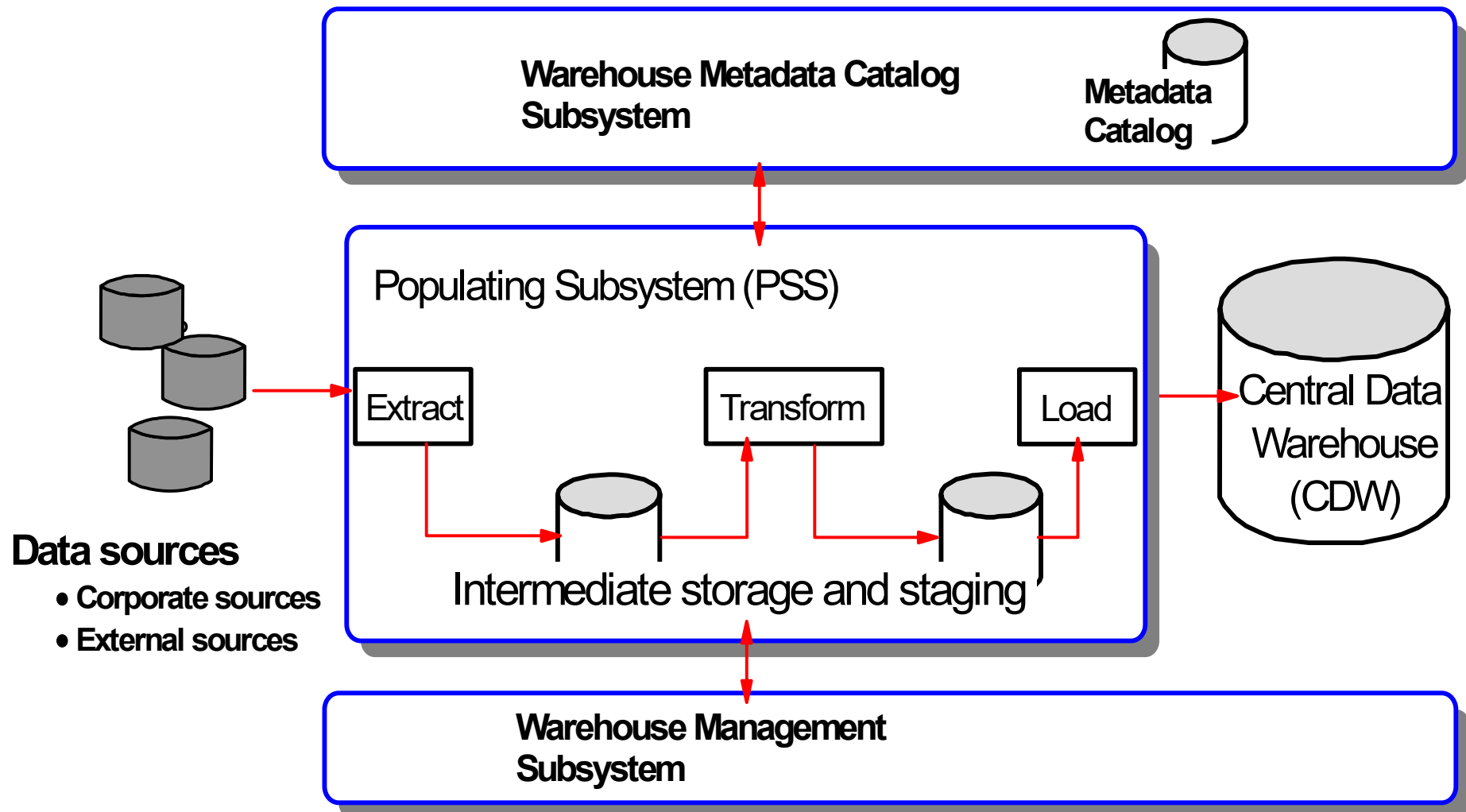
ETL-Reference Architecture - DWH Overview



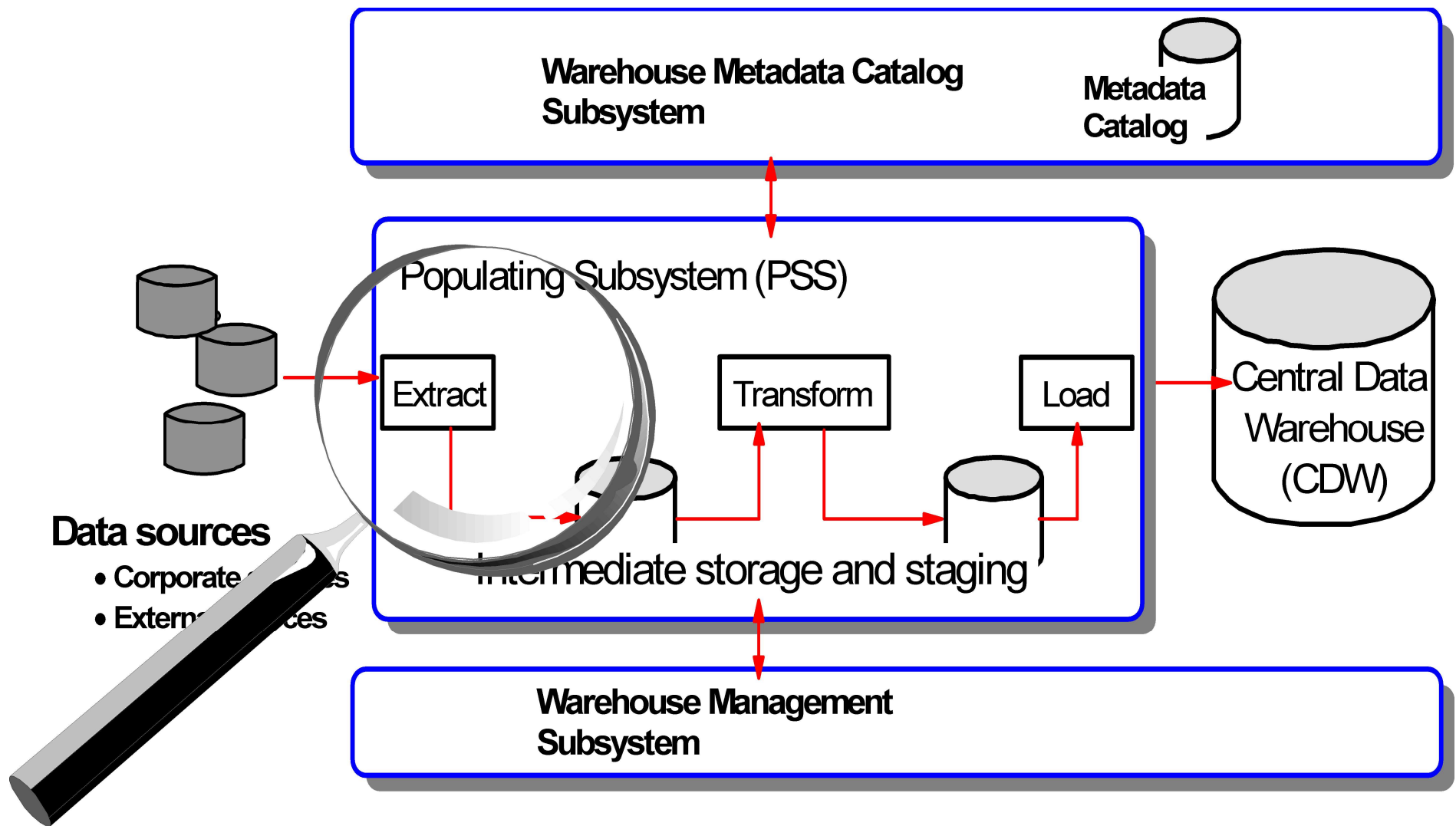
ETL-Reference Architecture - Focus on PSS



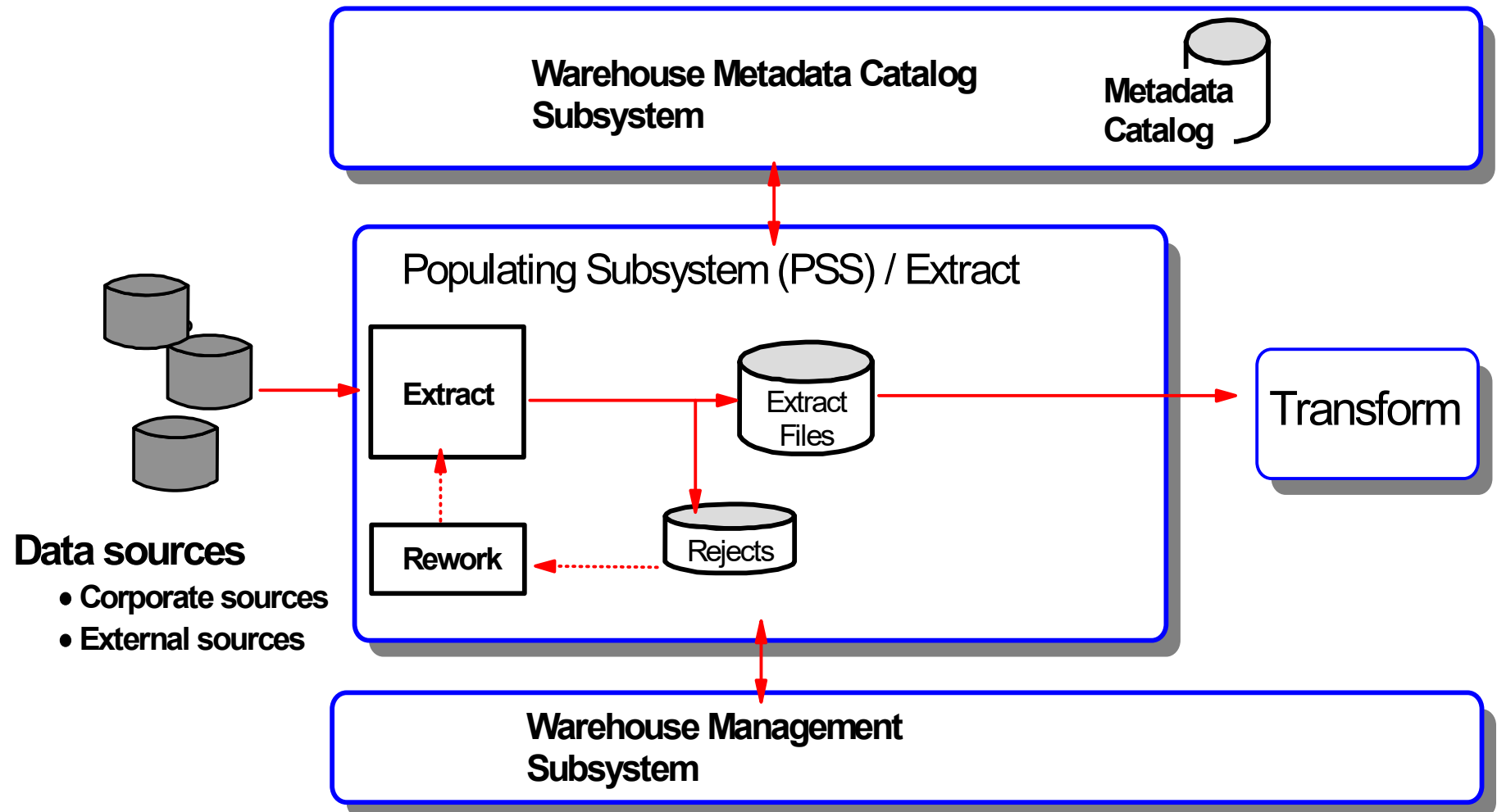
ETL-Reference Architecture – PPS Processes



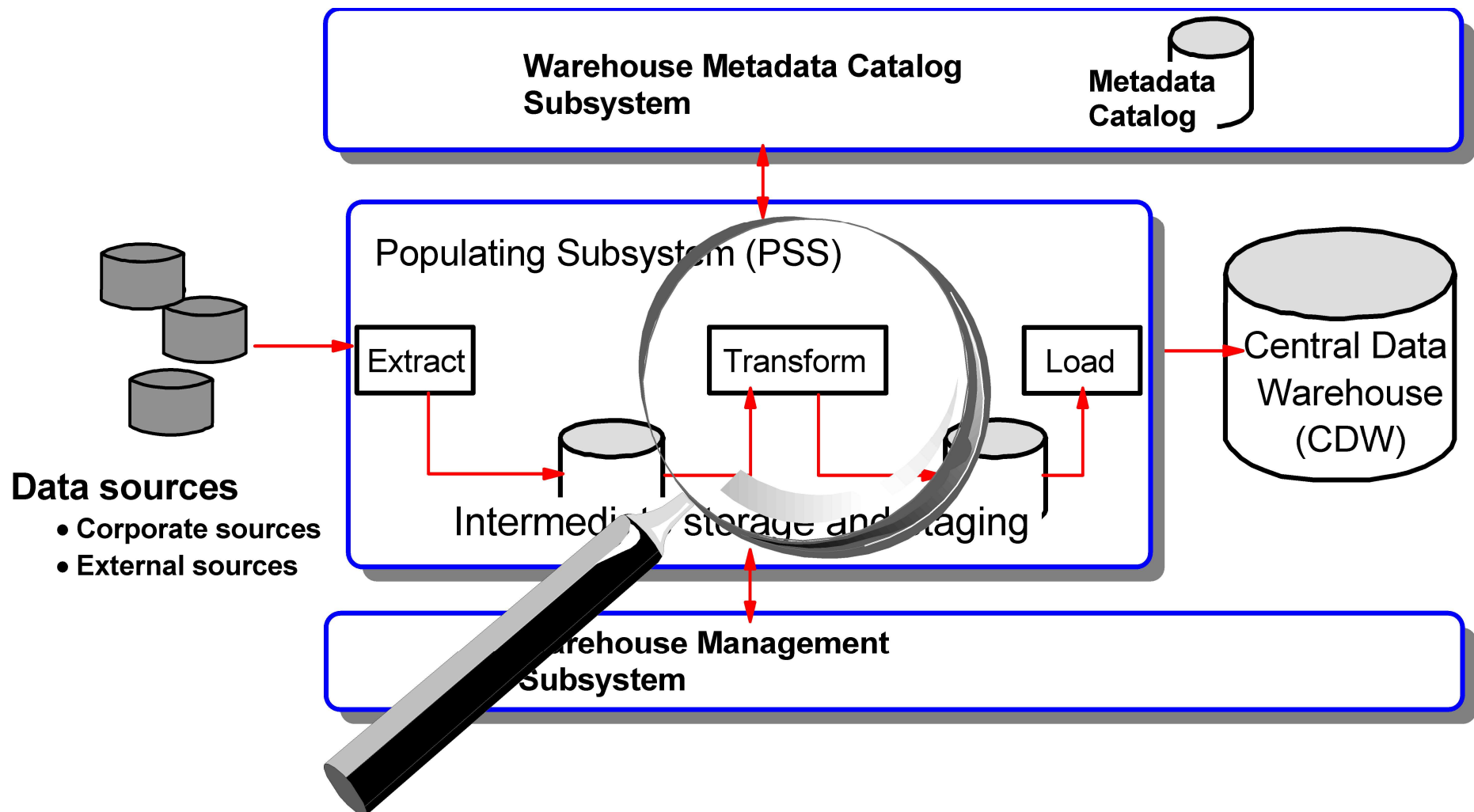
ETL-Reference Architecture – Extract Process



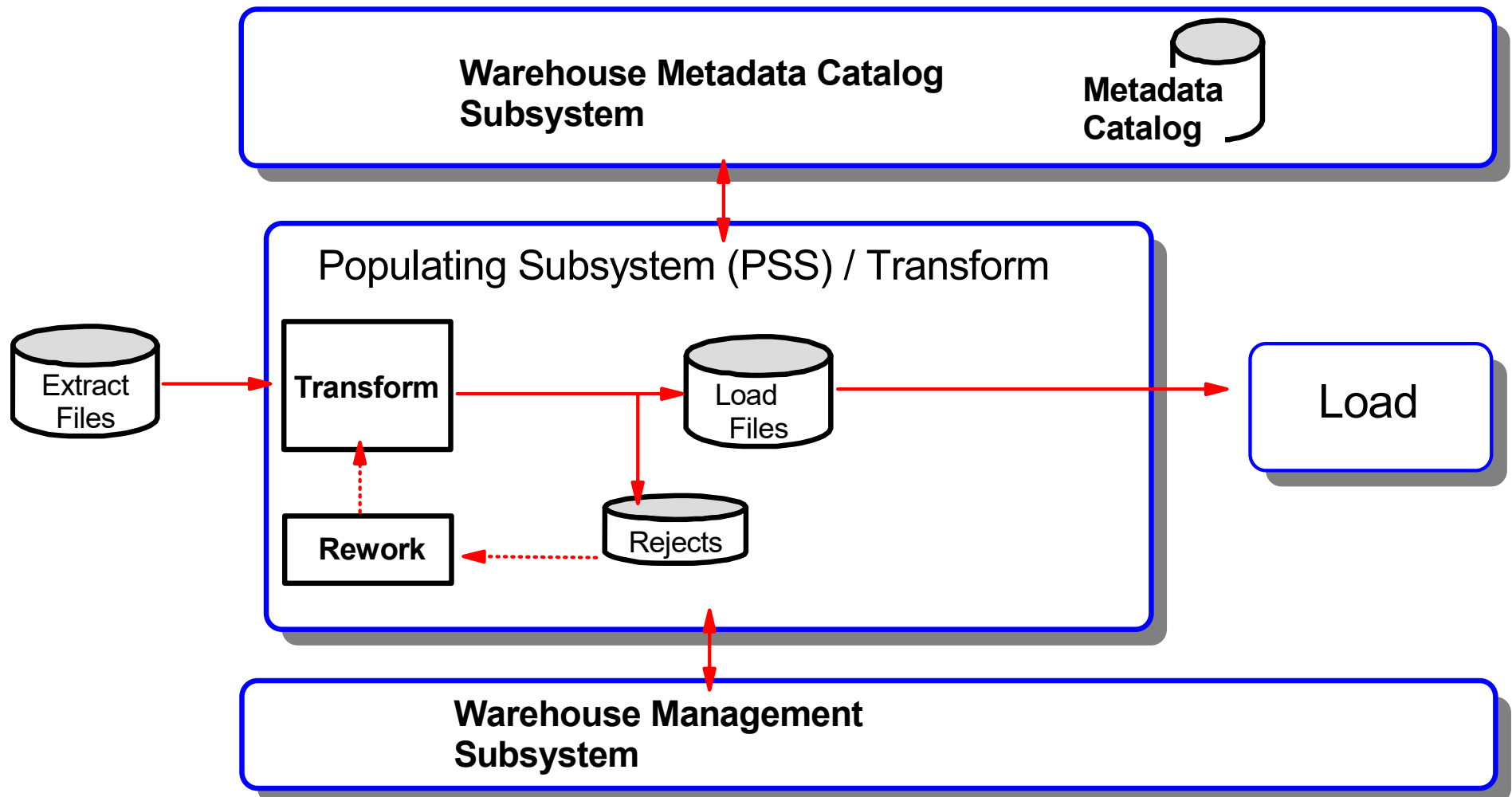
ETL-Reference Architecture - Extract Process



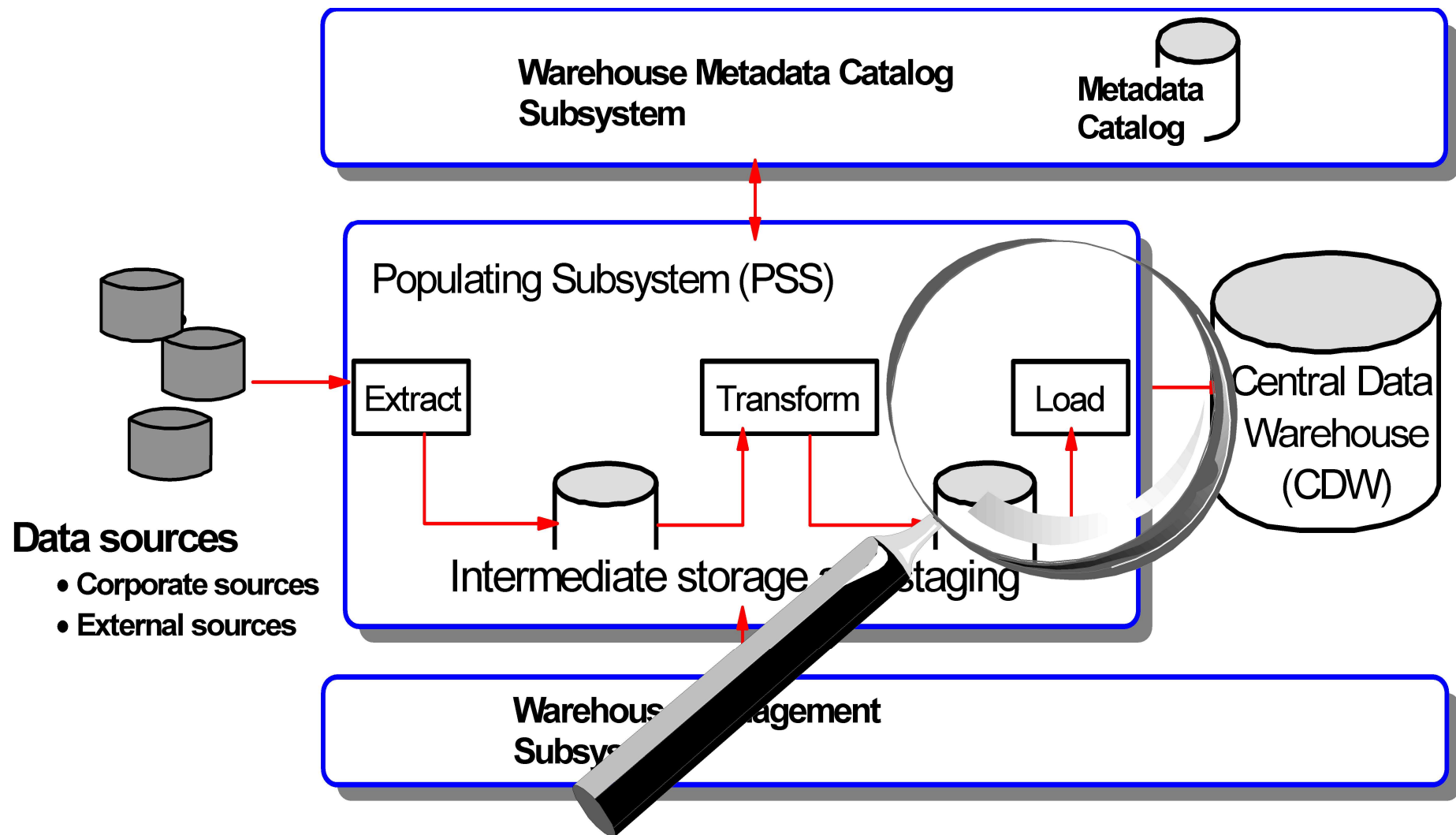
ETL-Reference Architecture–Transform Process



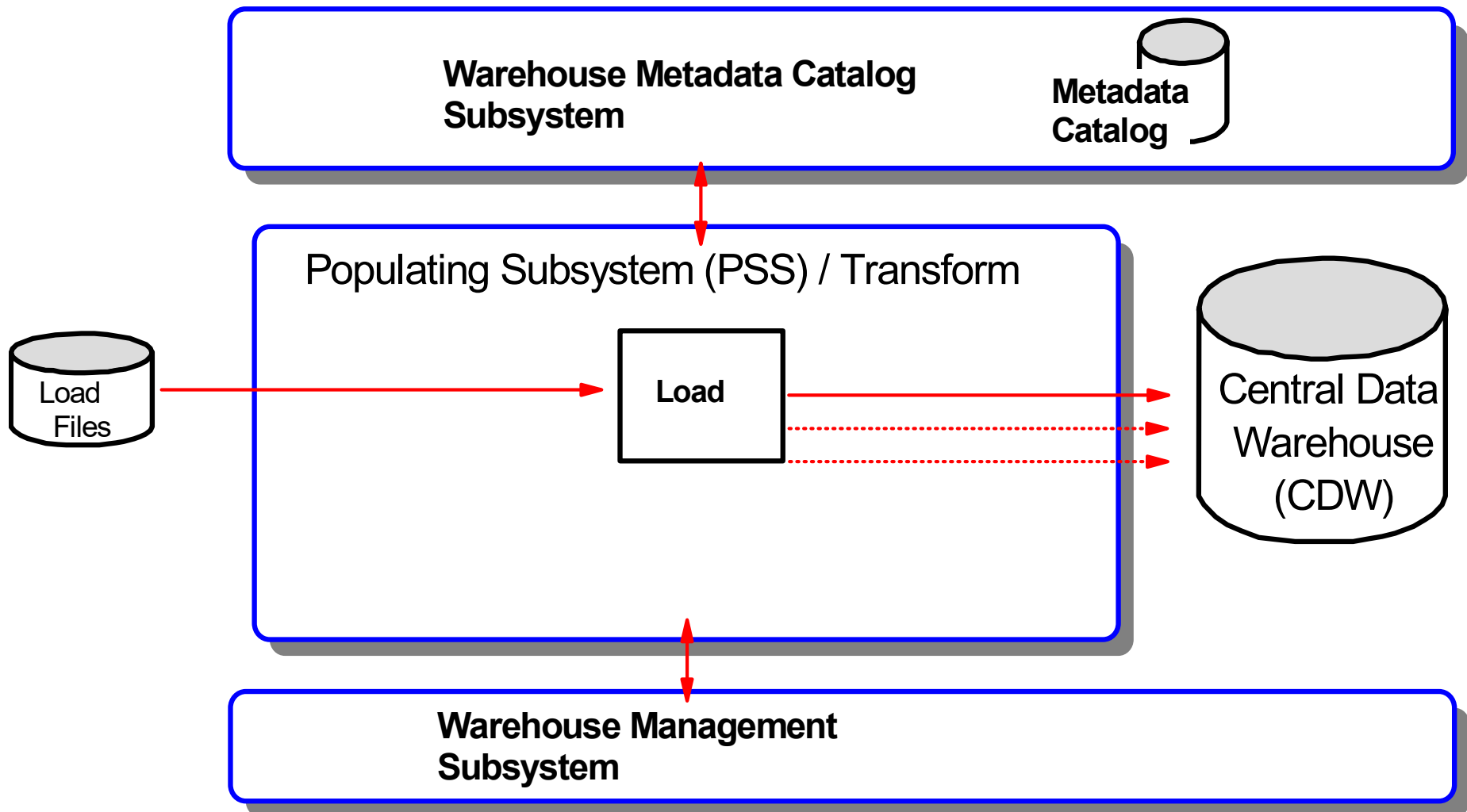
ETL-Reference Architecture-Transform Process



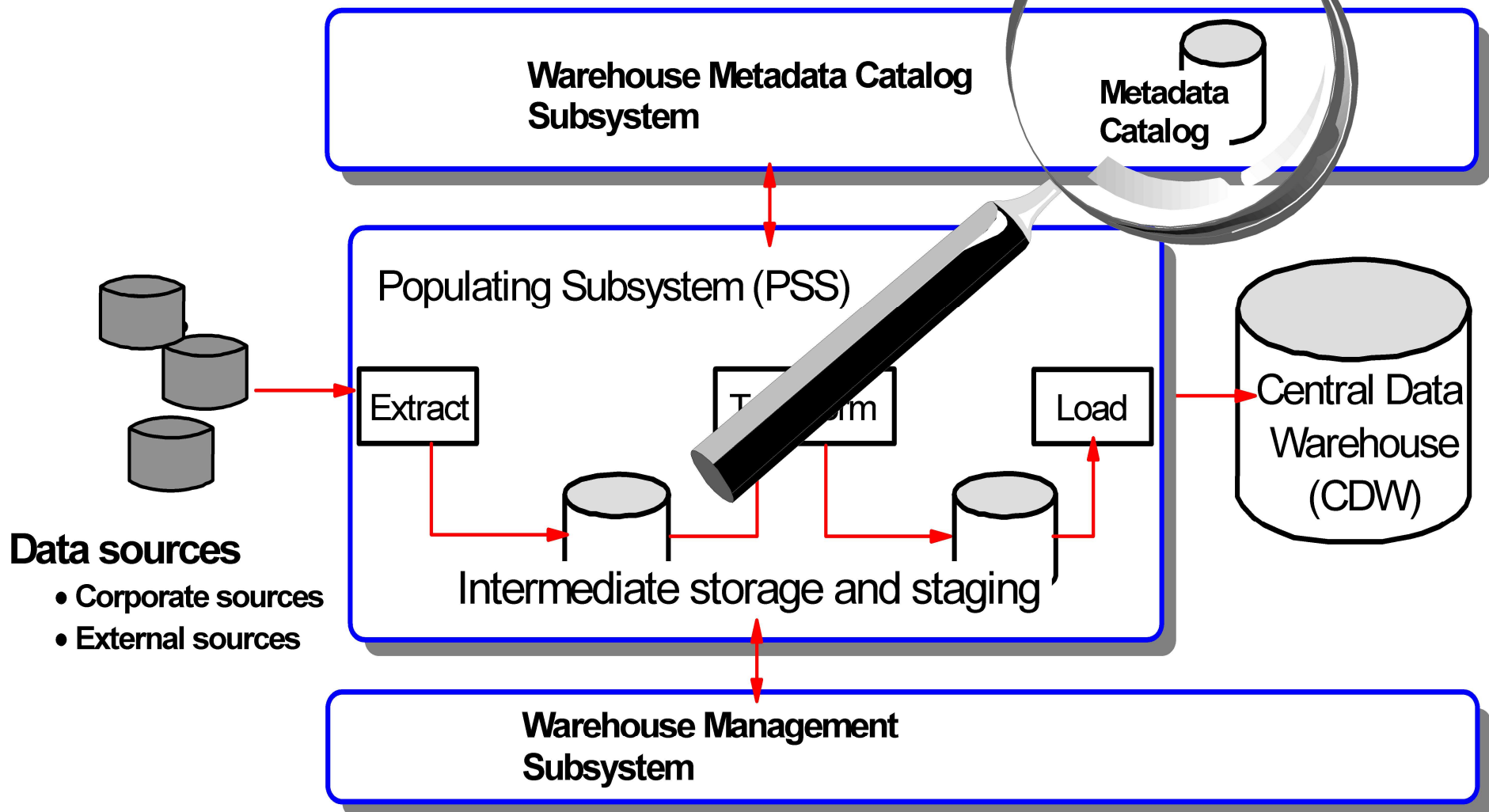
ETL-Reference Architecture – Load Process



ETL-Reference Architecture – Load Process



ETL-Reference Architecture–Metadata Catalog



ETL-Reference Architecture - Metadata Subsystem

• Metadata sources

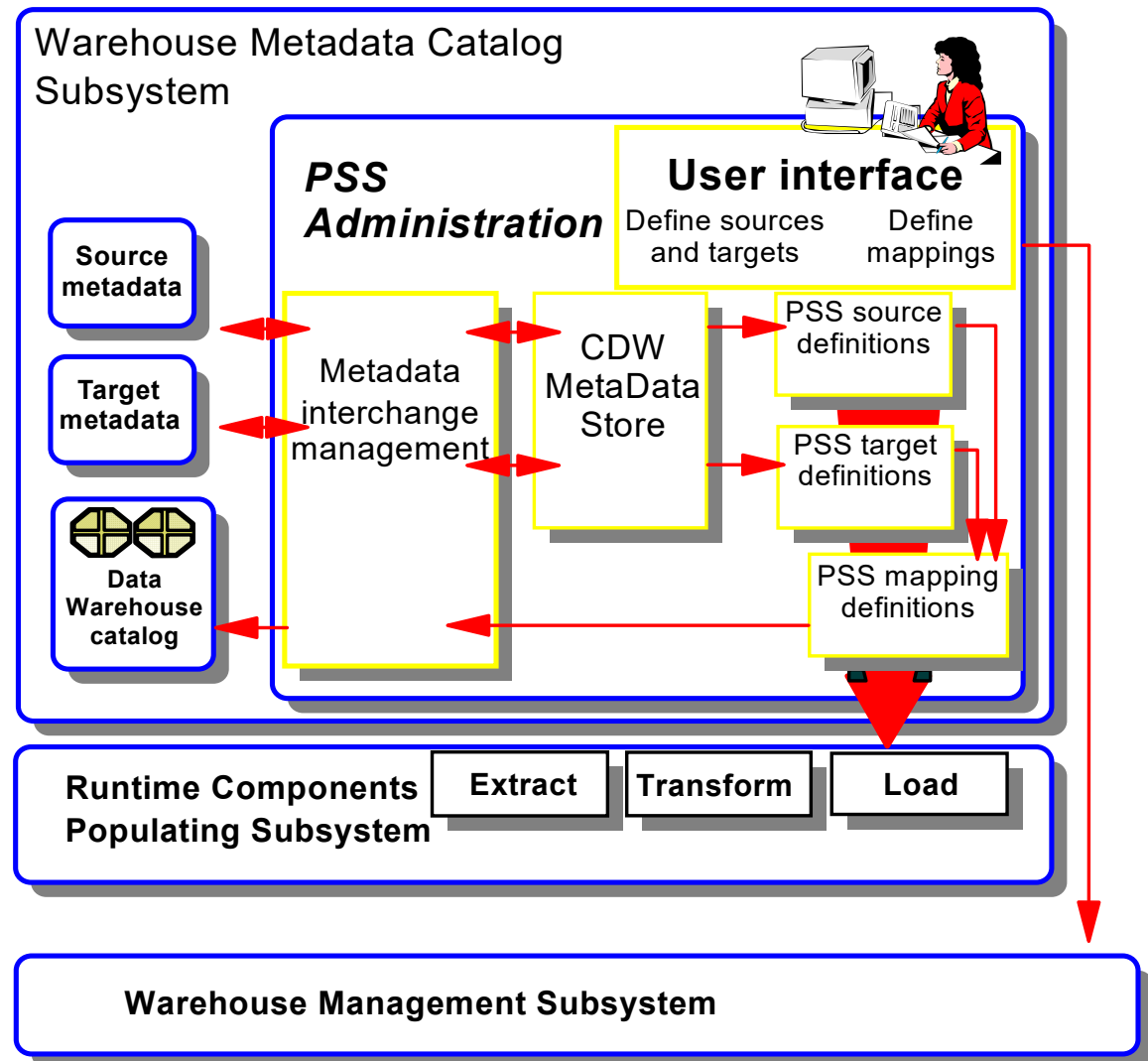
- Data modelling tools
- Database catalogs
- Record definitions in programs
- Populating tools

• Metadata Outputs

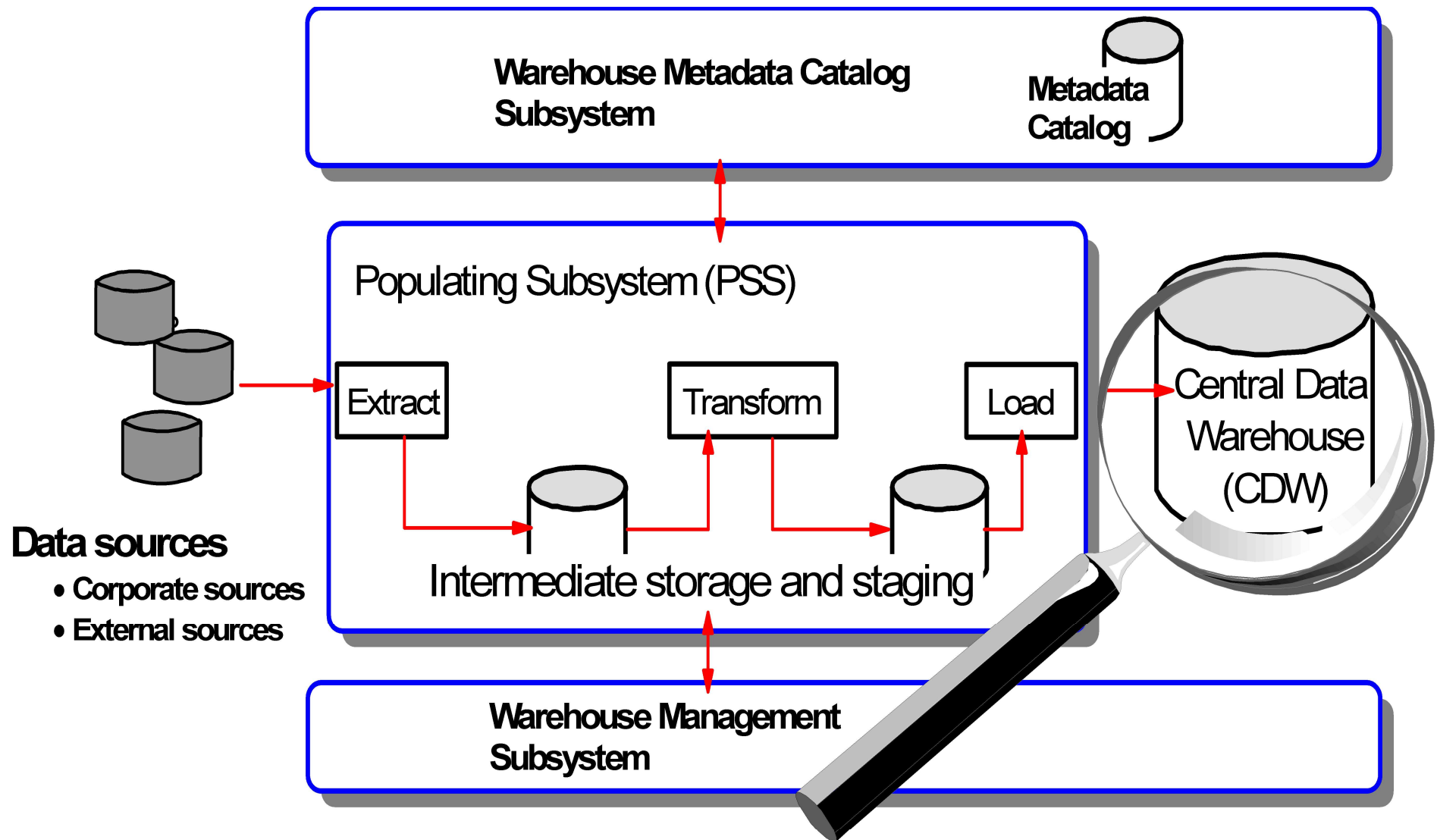
- PSS runtime statistics
- Data Warehouse catalog
- Process management

• Issues

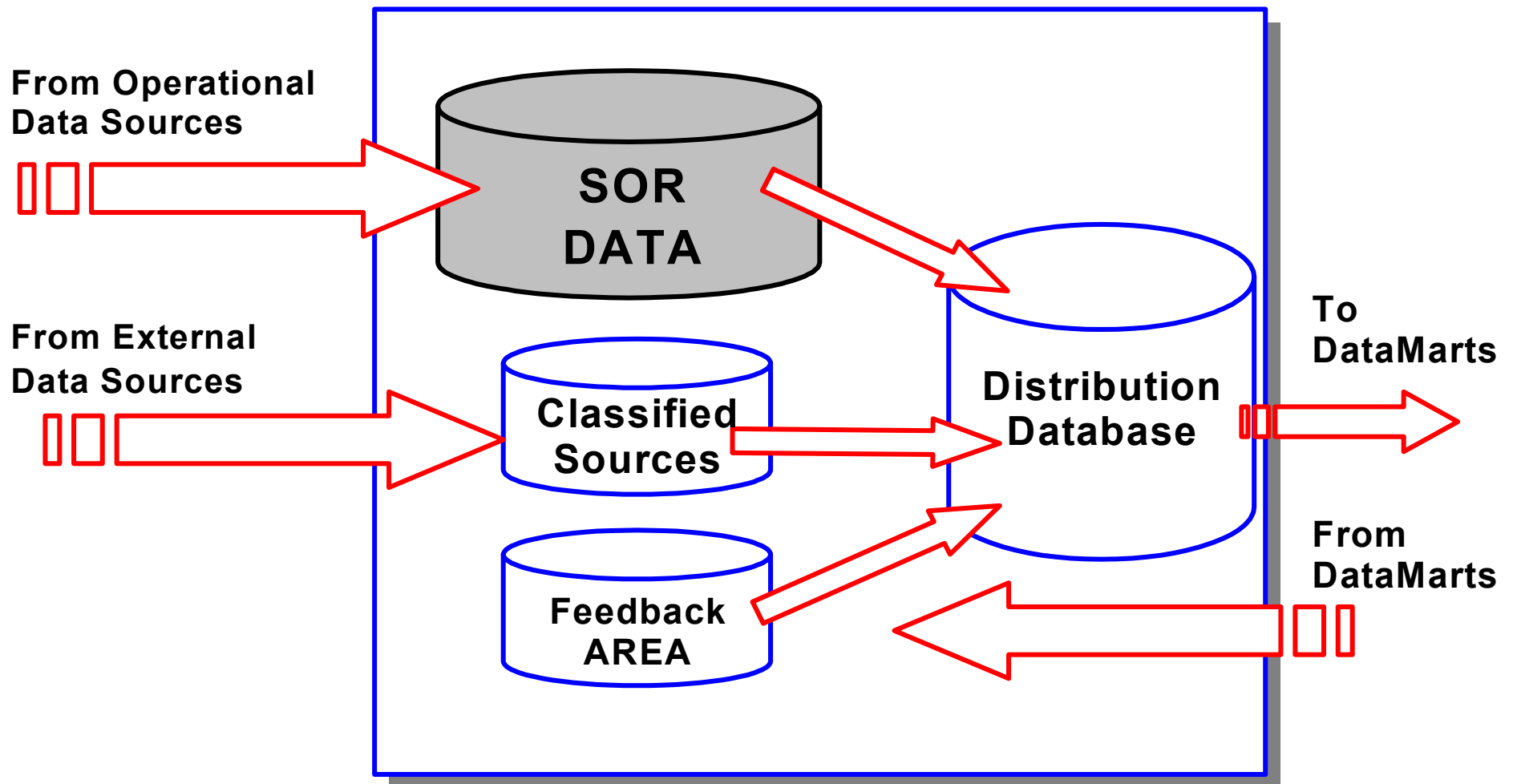
- Metadata access
- Metadata synchronization
 - Metadata Interchange
 - CDW Metadata store
- Today's tools provide little or no support



ETL-Reference Architecture – Central DWH

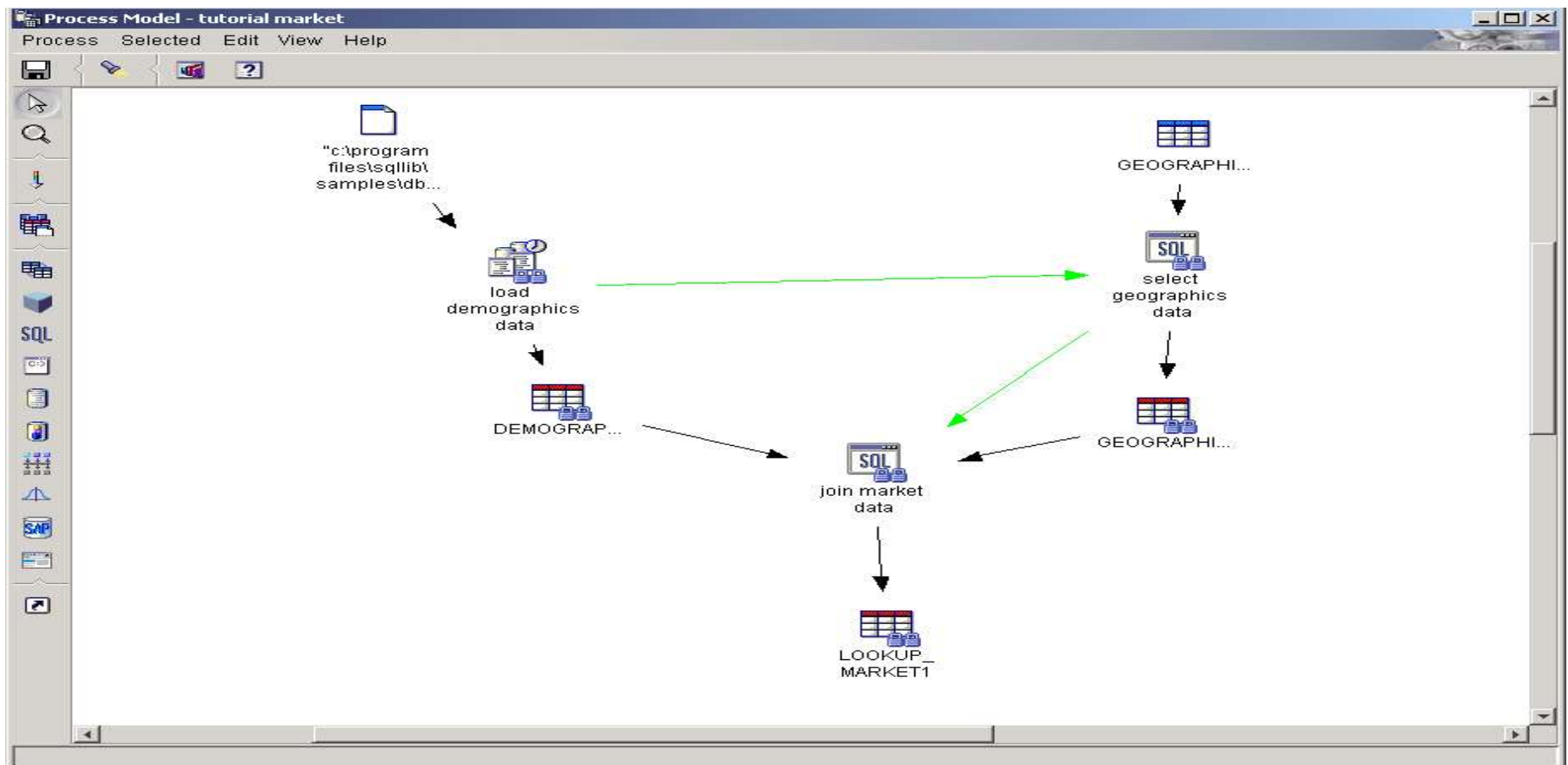


ETL-Reference Architecture - CDW Data Feeds



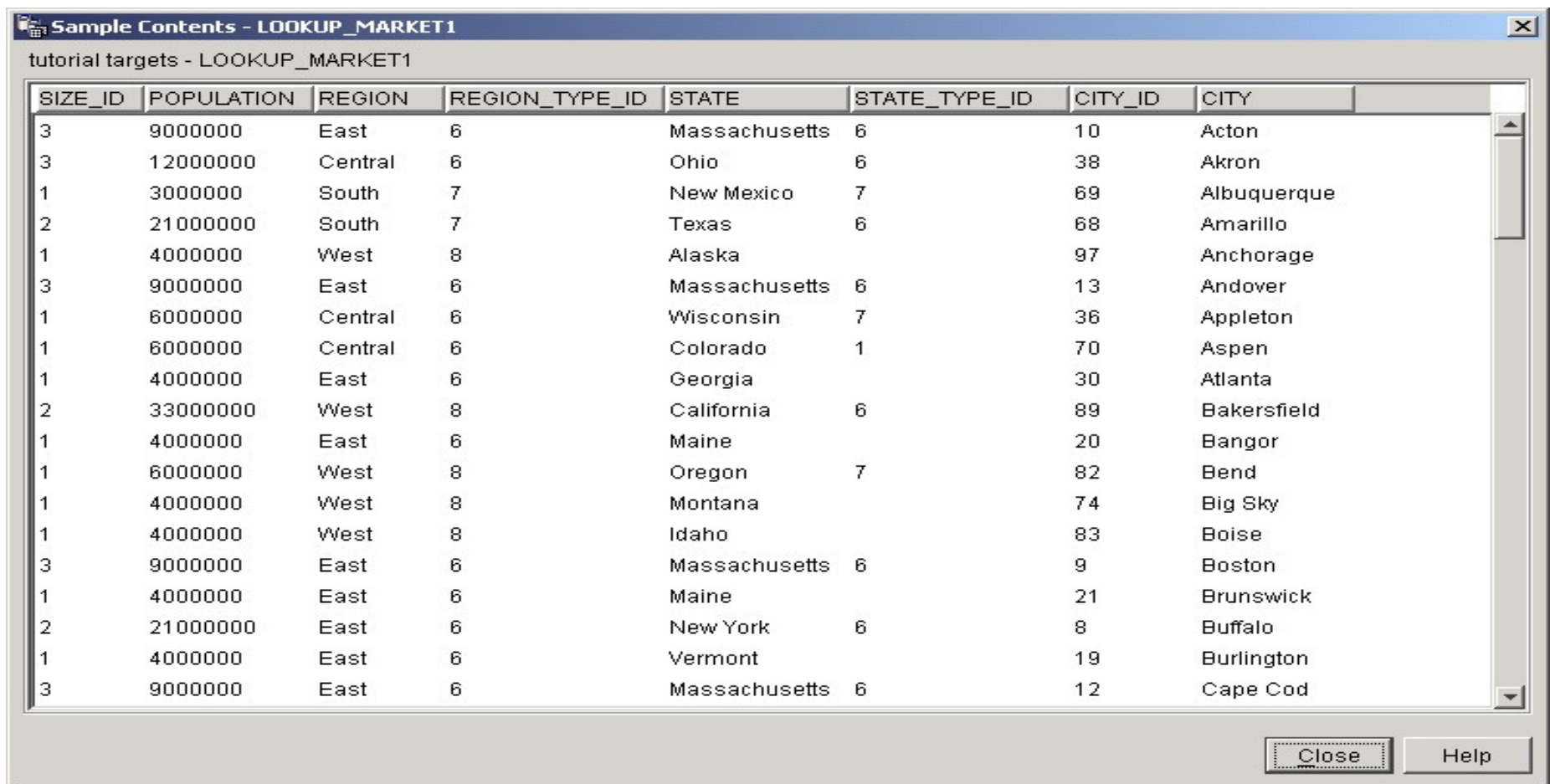
Exercise1 to Lesson 6 - DB2 WH-Manager (Part1)

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:



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....

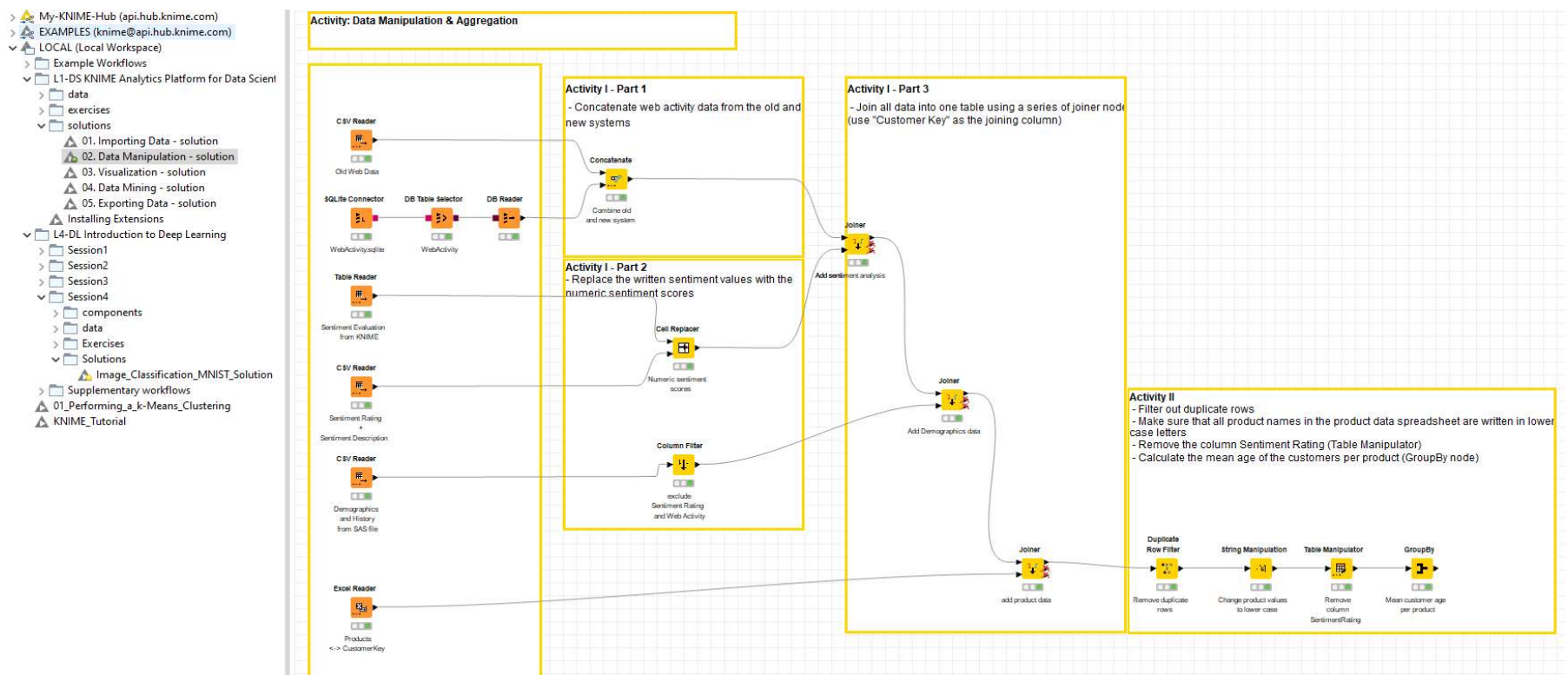
1. **DISCOVER:** Evoke-AXIO (now Informatica), Talend - Open Studio, IBM Infosphere Inform. Server (IIS) – ProfileStage, or ????
2. **PREPARE:** HarteHanks-Trillium, Vality-Integrity, IBM Infosphere Inform. Server (IIS) – QualityStage, or ??????

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 (PD**): 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):



PD**: For the Program Design investigate these KNIME features in more detail

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: : <http://ibm.biz/cpdaas-videos>
<https://dataplatfrom.cloud.ibm.com/docs/content/wsj/getting-started/videos.html?audience=cpdaas&context=cpdaas#data-engineers>

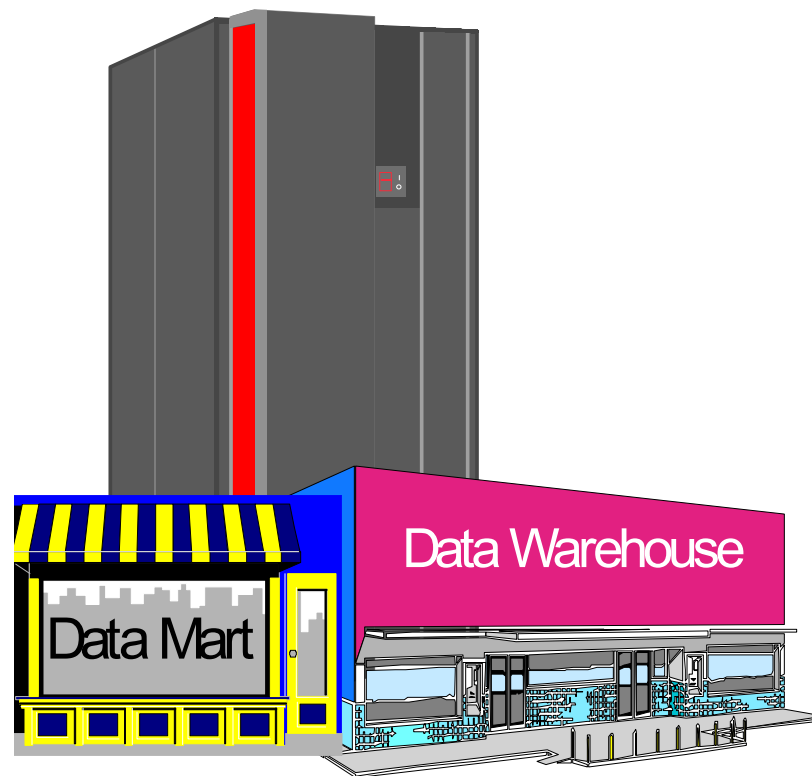
Category 1: Introduction & Architecture of DWH

Category 2: Databases and Data Modeling

Category 3: ETL: Architecture & Technology

Category 4: Descriptive – & Advanced Analytics

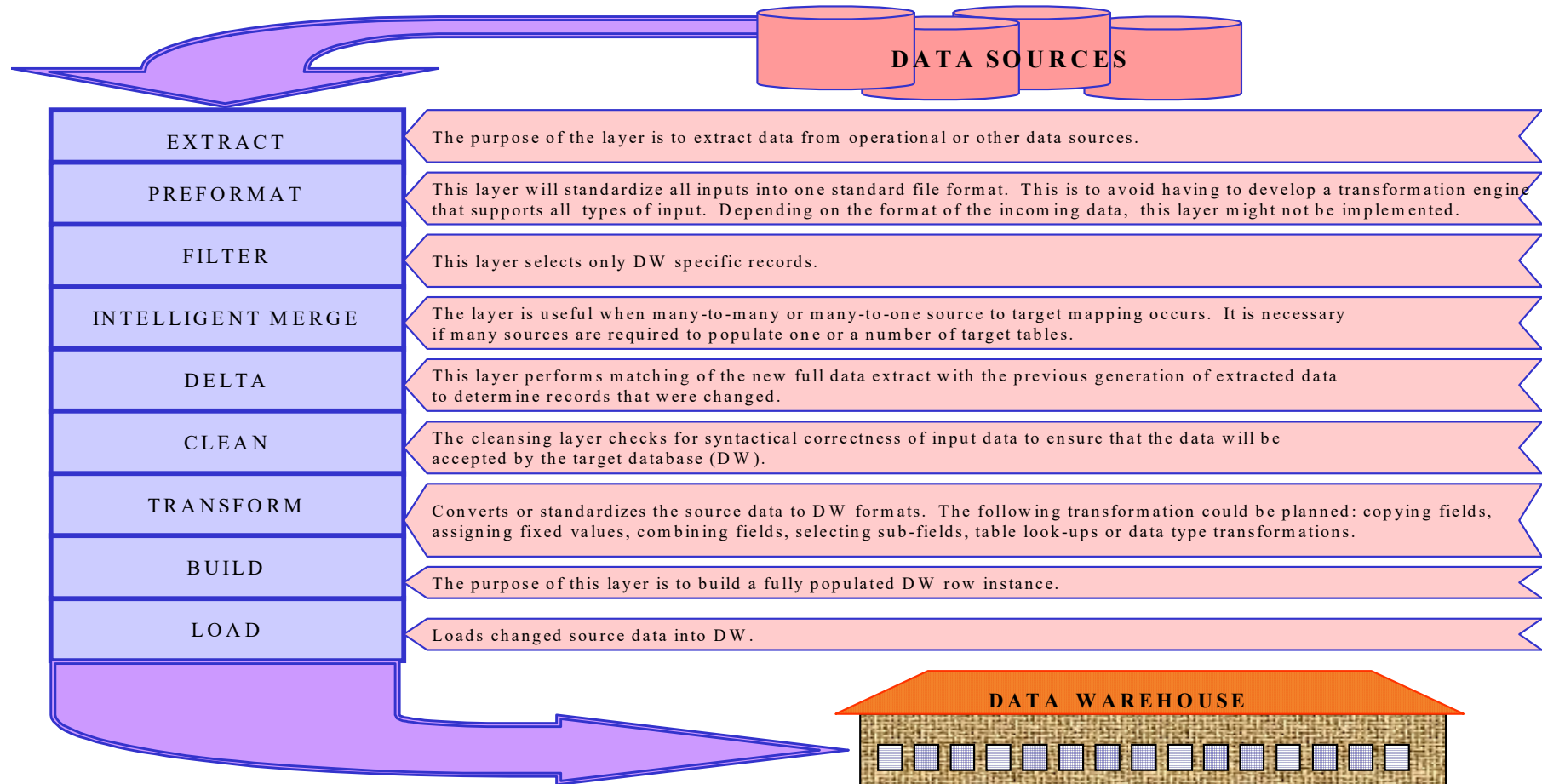
DW07 - ETL Techniques & ETL Tools



5 Highlights to ETL Techniques

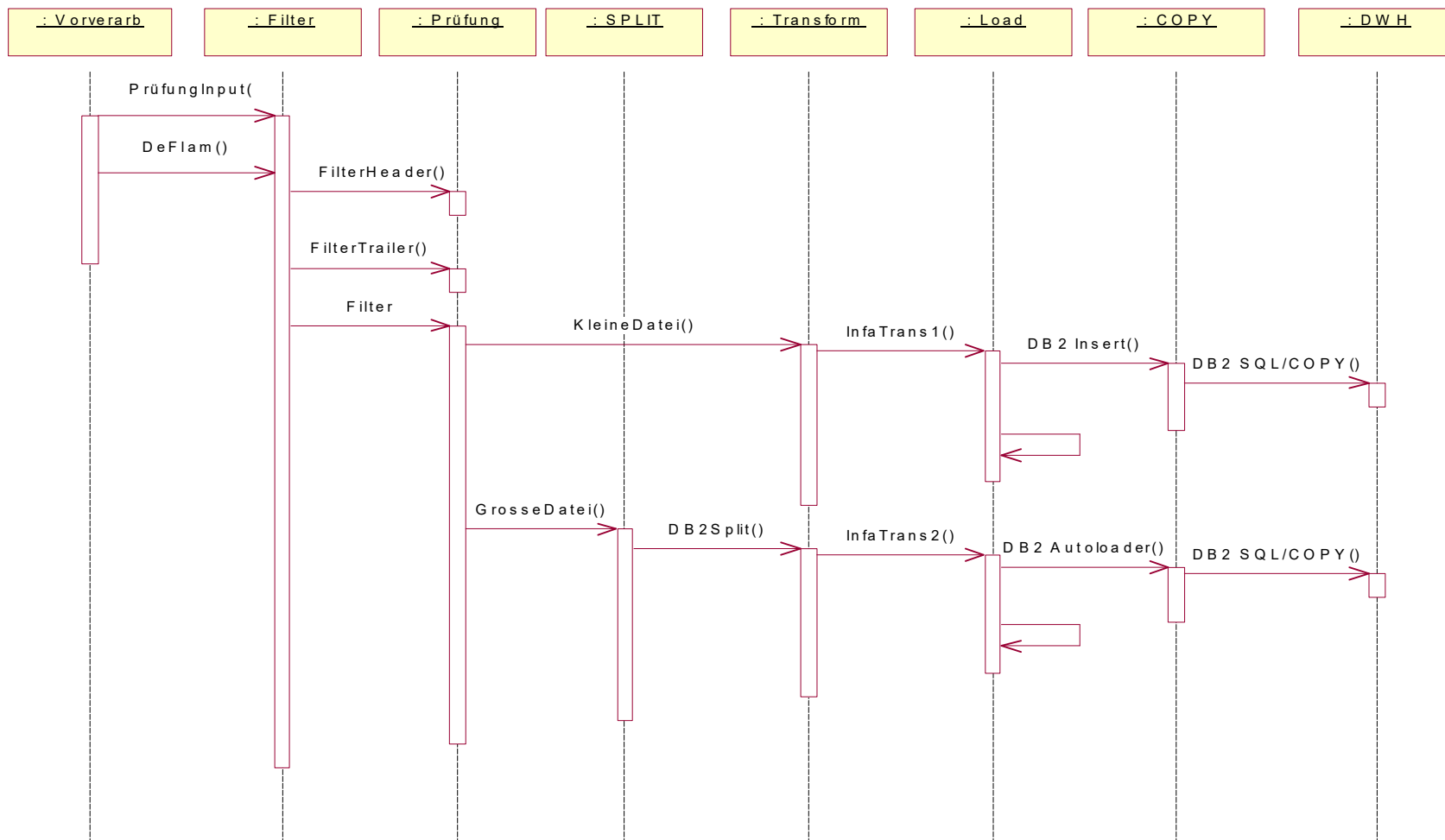
1. ETL Process Layer Concept
2. Framework / Control of Processes
3. Scalability & Parallel Processing
4. Integration of ETL and DB
5. Special ETL Techniques

Generic ETL Process Layers

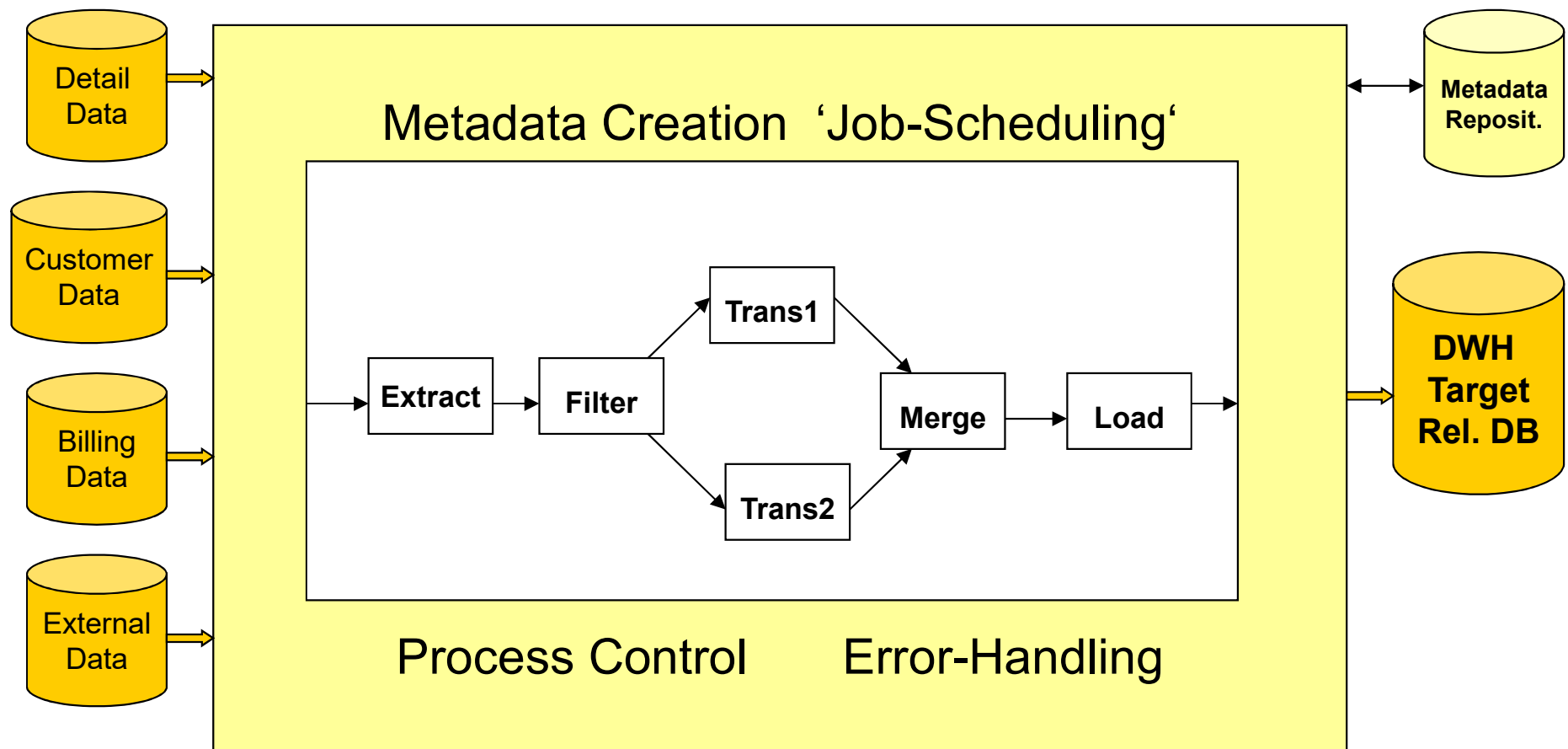


7/27/2001

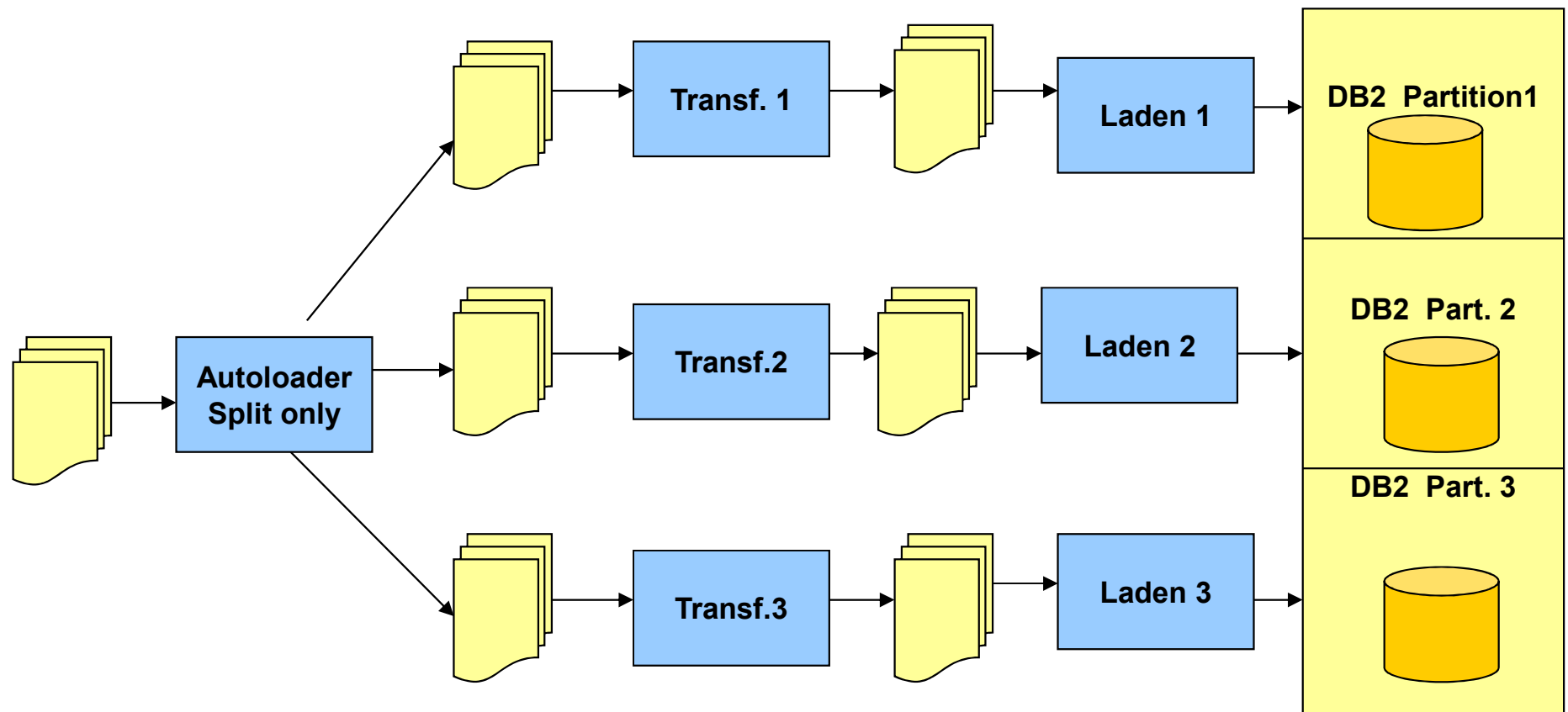
ETL Layer Concept (Example)



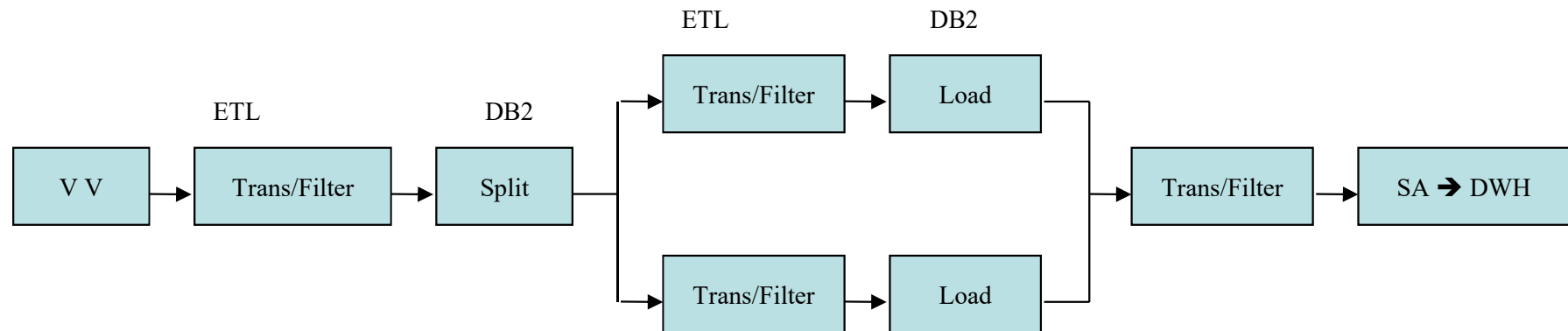
FRAMEWORK / Control of Processes



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

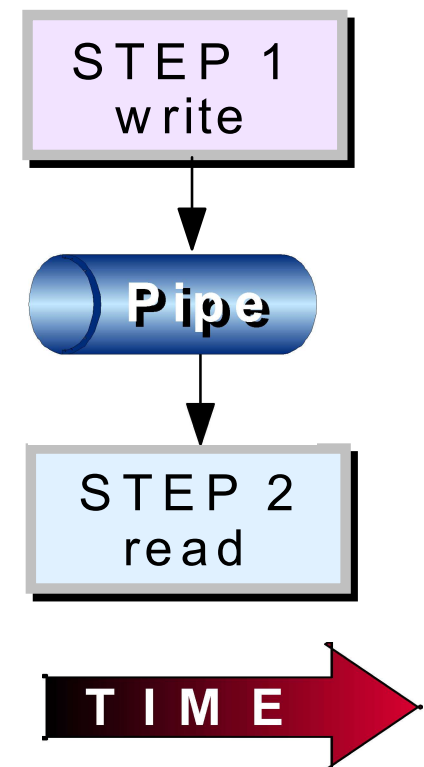
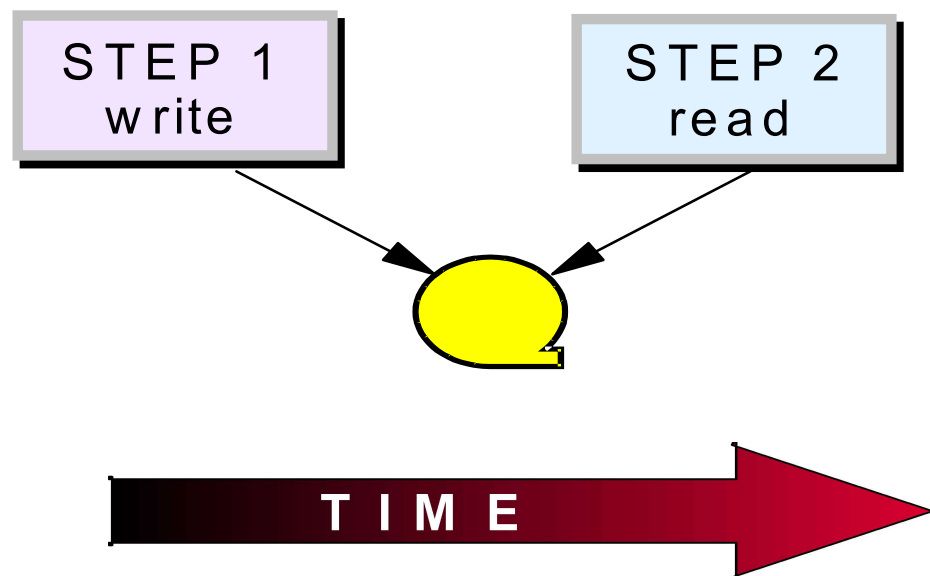
ETL calls DB2 Autoloader (with Split Only)

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



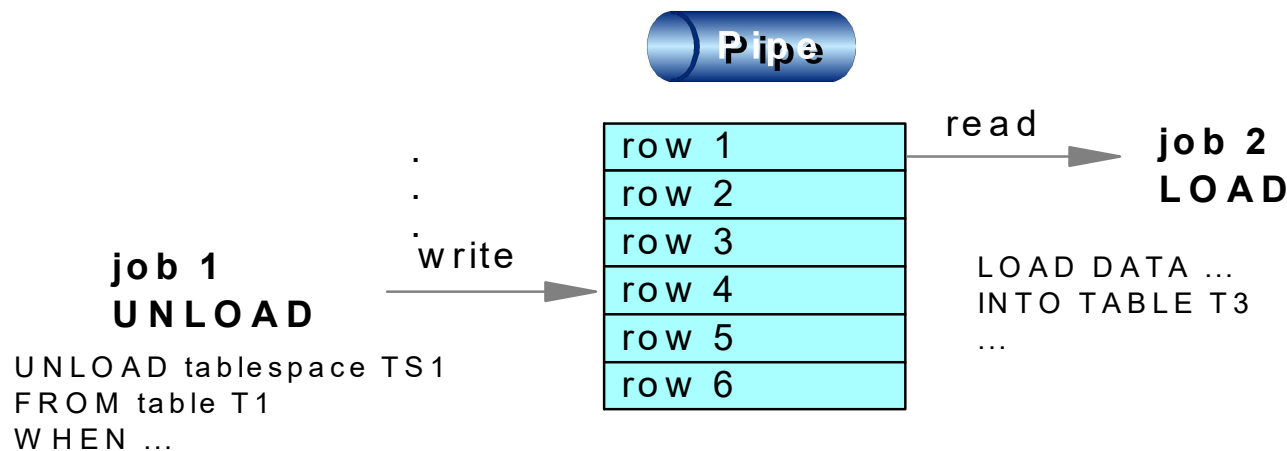
ETL Technique – ‘Piping’ Example

■ UNLOAD

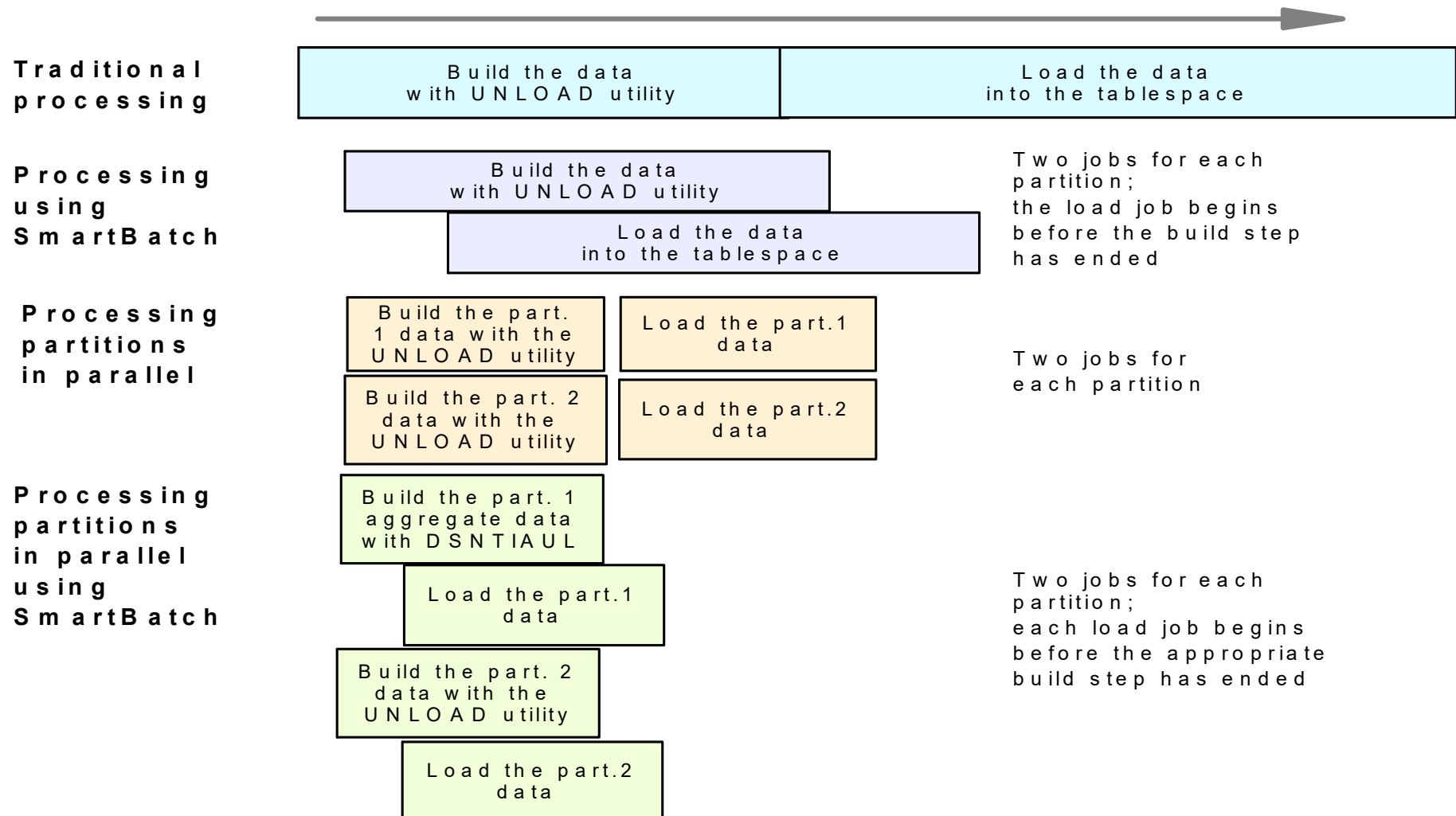
- Provides fast data unload from DB2 table or image copy data set
- Samples rows with selection conditions
- Selects, order and formats fields
- Creates a sequential output that can be used by LOAD

■ LOAD

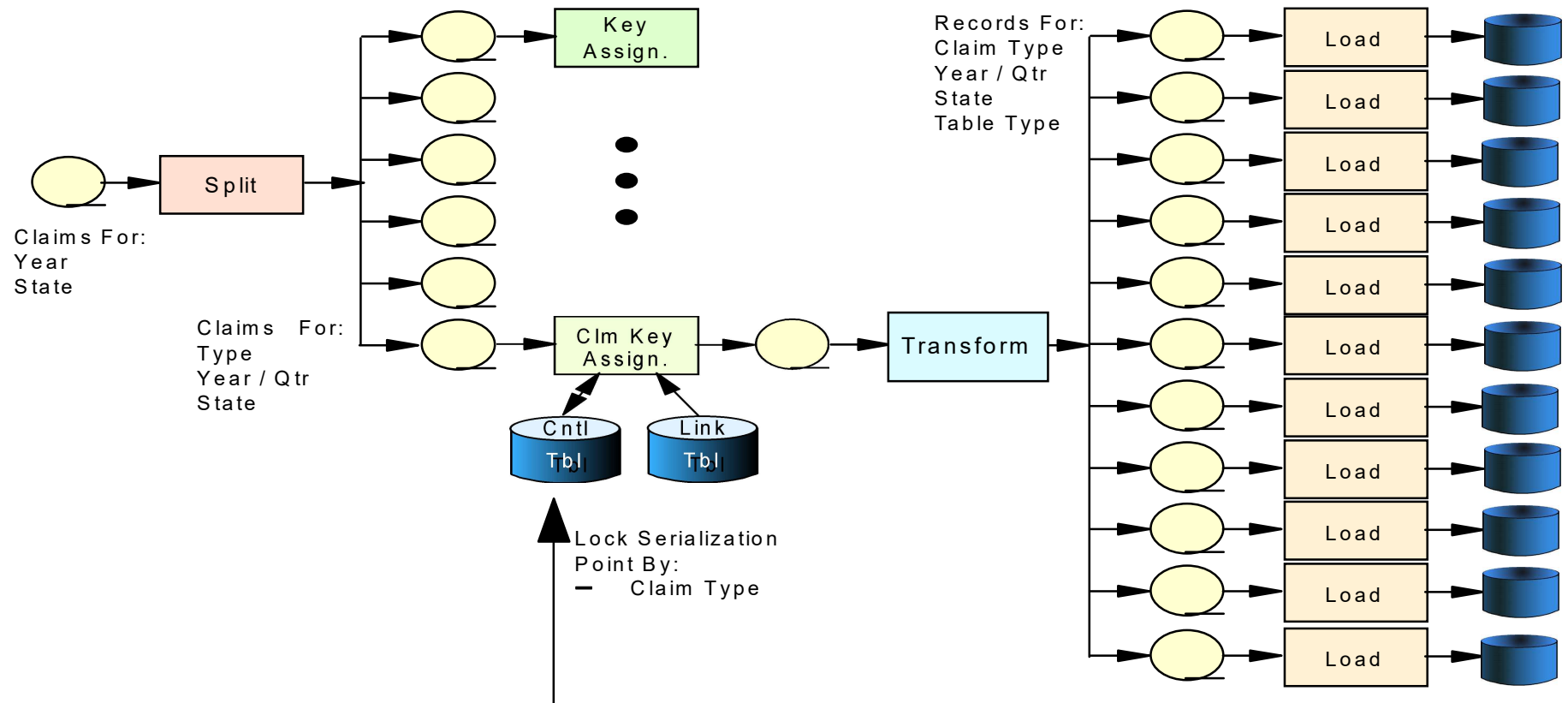
- With SmartBatch, the LOAD job can begin processing the data in the pipe before the UNLOAD job completes.



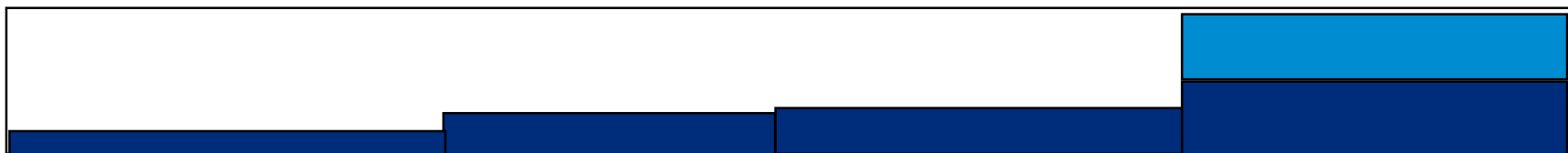
ETL Technique – Compare Runtime



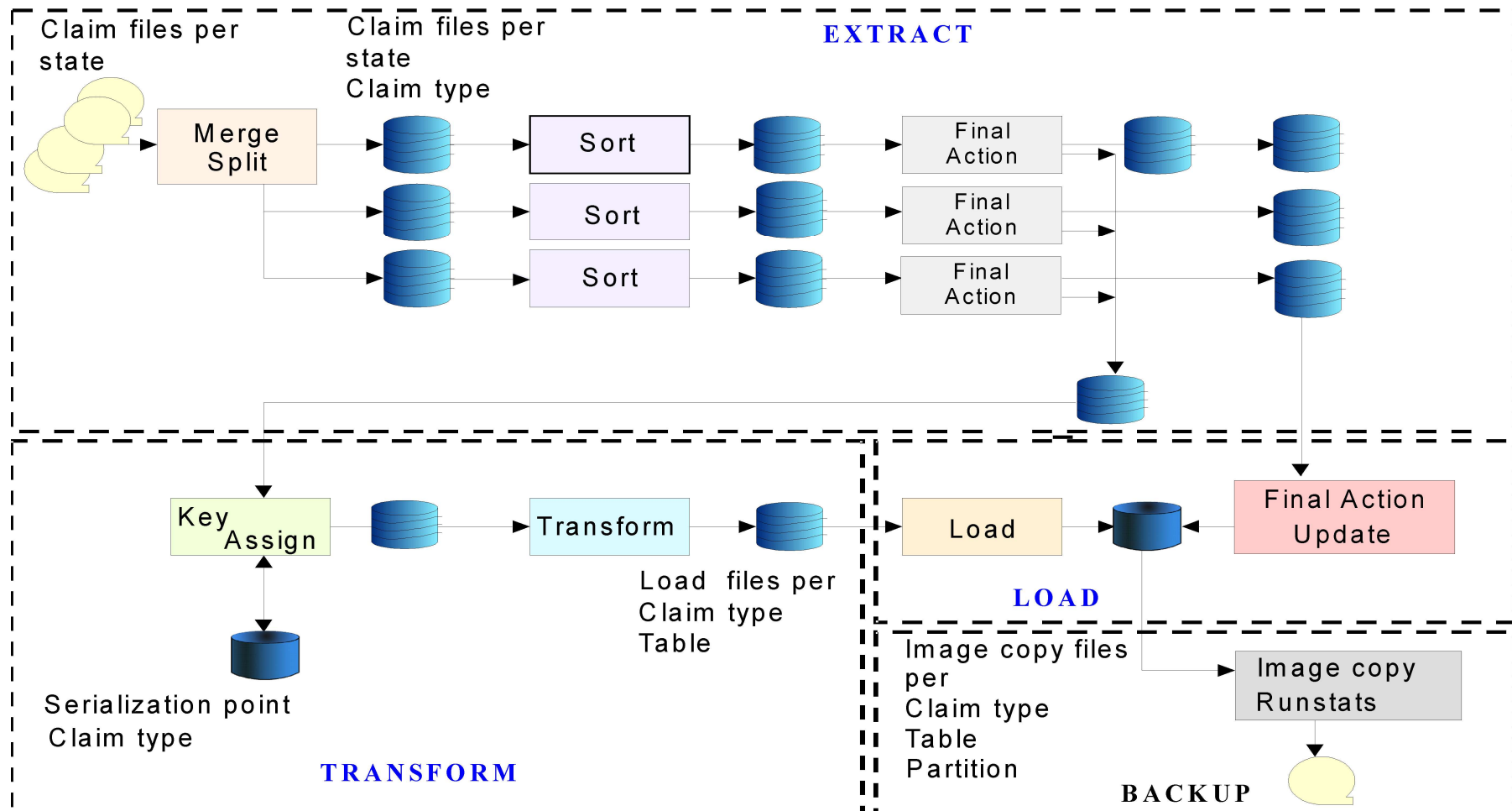
ETL Technique – ‘Sequential Design’



CPU Utilization Approximation

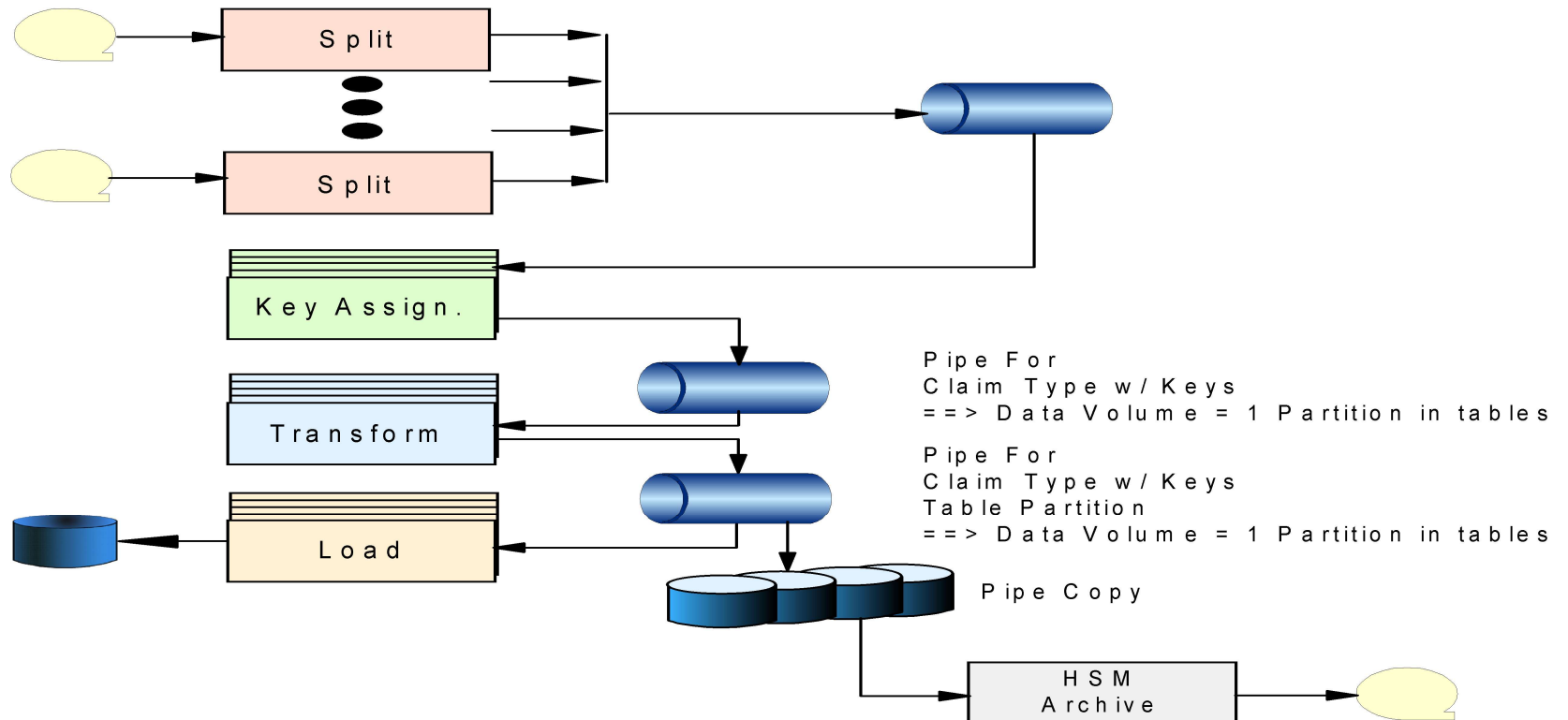


ETL Technique – Sequential Design 2



ETL Technique – ‘Piped Design’

Claims For:
Year
State



CPU Utilization Approximation



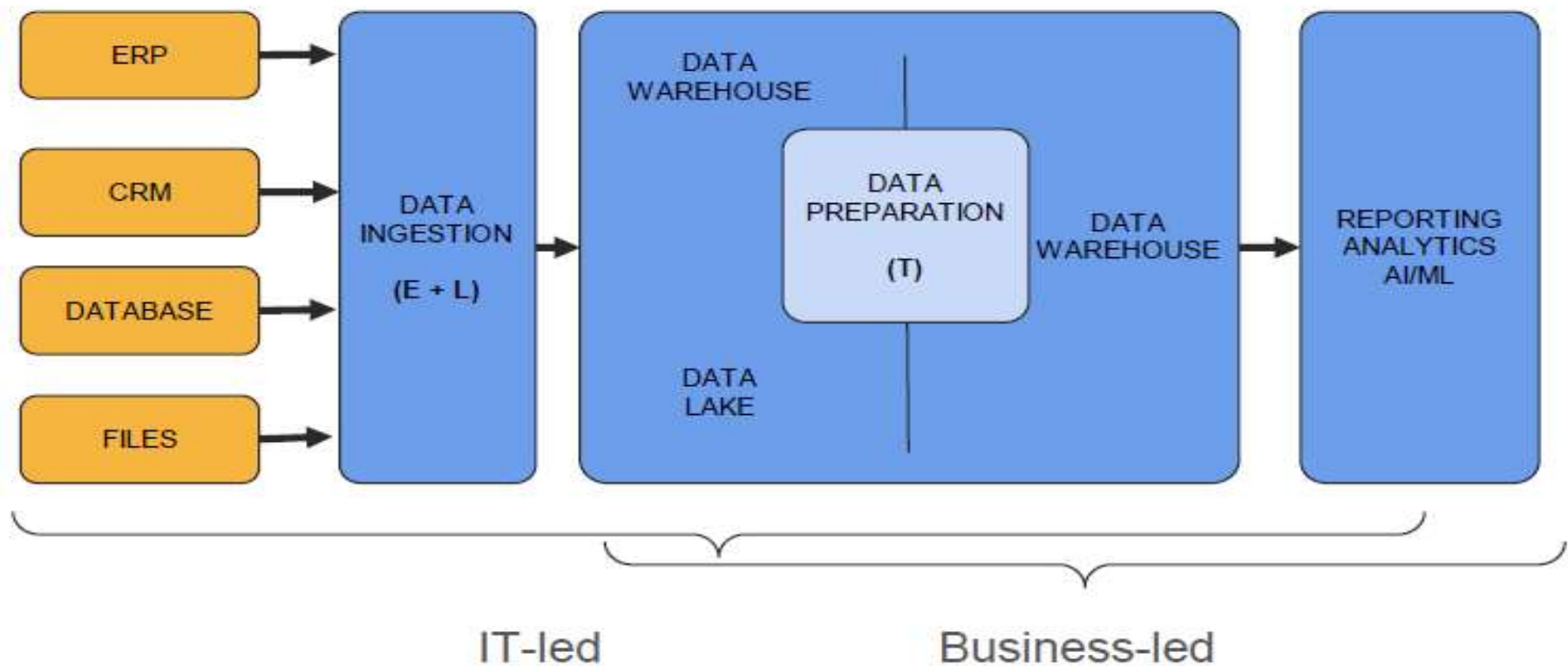
What is the concept of a “Data Fabric”

A **data fabric** in a data warehouse environment is an **architectural framework** that enables seamless access, integration, and management of data across different sources, platforms, and environments (on-premise, cloud, hybrid). It provides a unified, intelligent, and automated data management approach, improving data accessibility, governance, and security across the enterprise.

Comparison: Data Fabric vs. Traditional ETL

Feature	Data Fabric	Traditional ETL
Data Access	Virtual, federated, real-time	Extract, transform, load (batch-based)
Scalability	Highly scalable across cloud & hybrid	Limited scalability
Automation	AI/ML-driven automation	Manual ETL processes
Flexibility	Works across multiple environments	Tightly coupled to specific data warehouses
Speed	Faster insights due to real-time integration	Slower due to batch processing

Modern ELT Stack in a Cloud DWH (AWS)



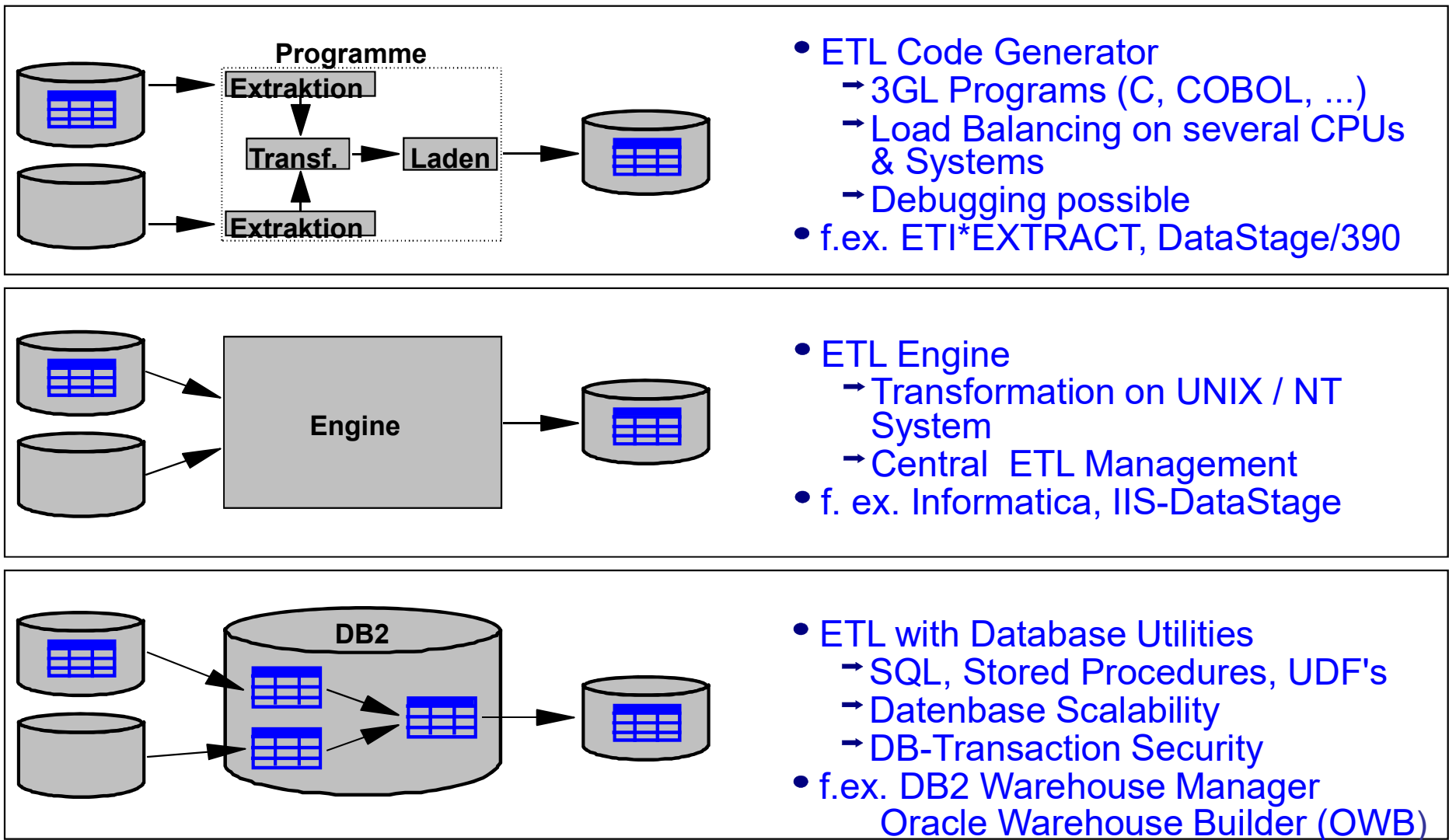
For more information see “ELT-Stack_in_AWS-Cloud-DWH.pdf” in [DHBW-Moodle]

ETL Marketplace & Tools Positions

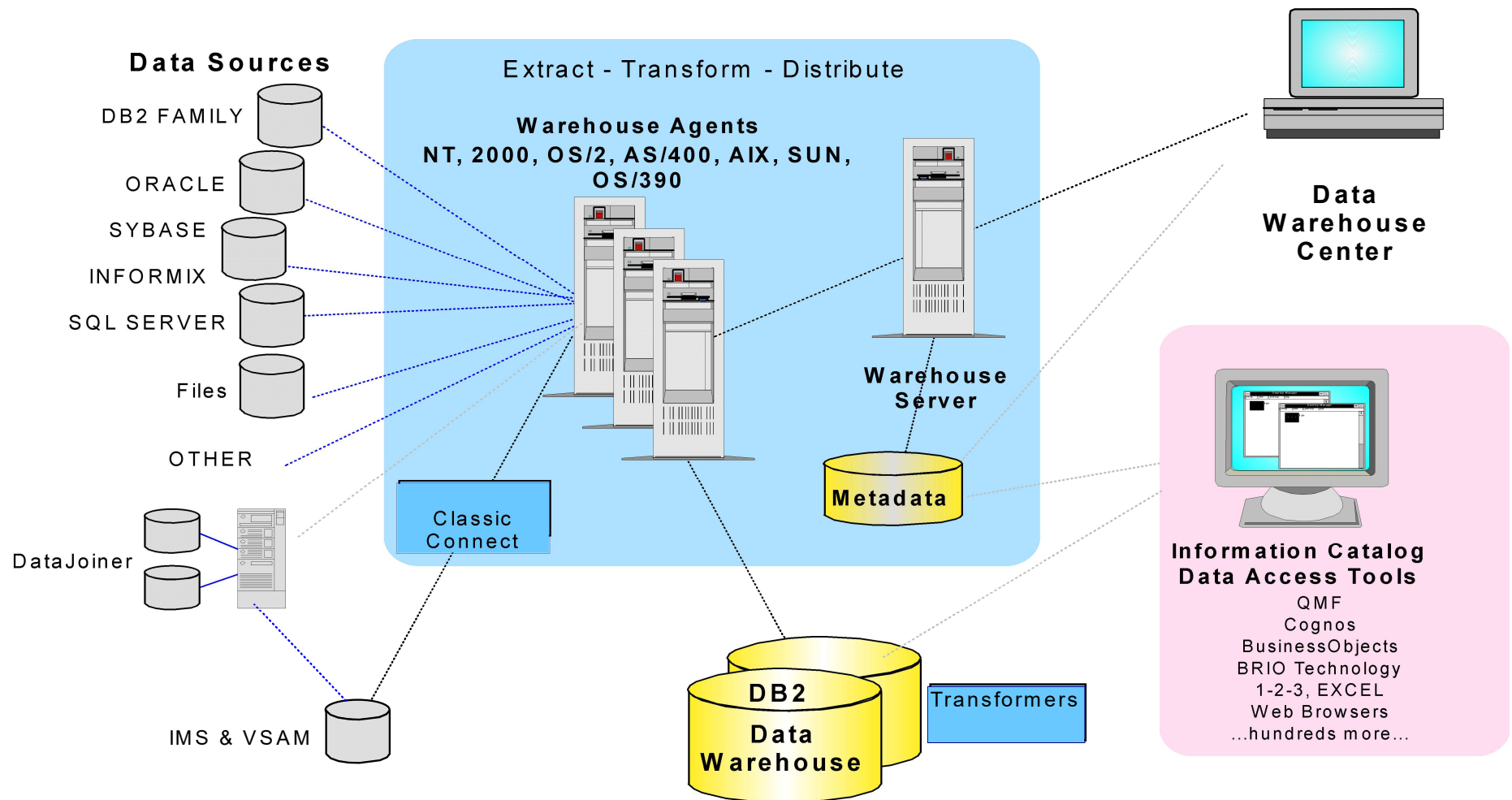
(Source: Gartner “Magic Quadrant for Data Integration Tools (November 2024)”)



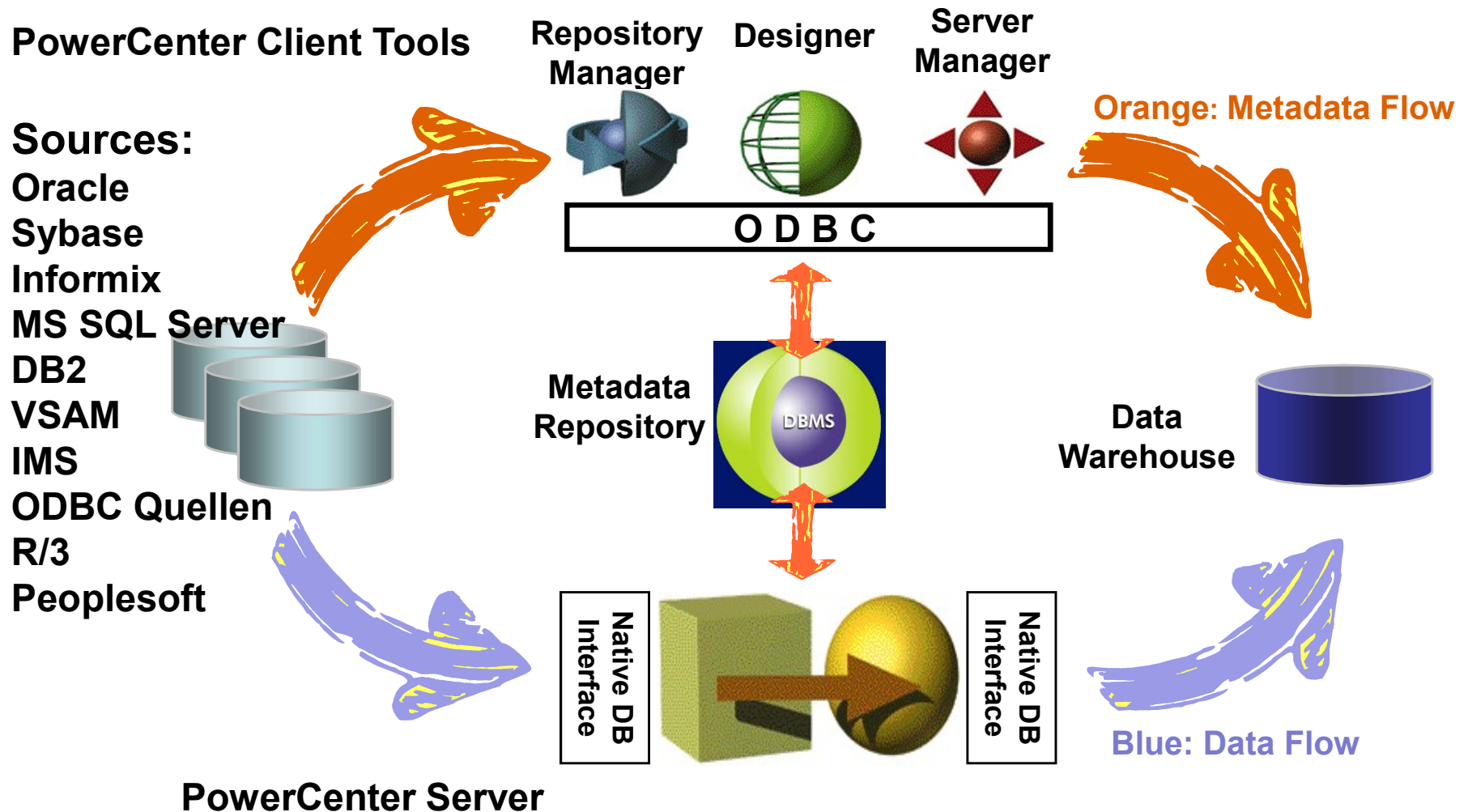
The 3 ETL Tool Architectures



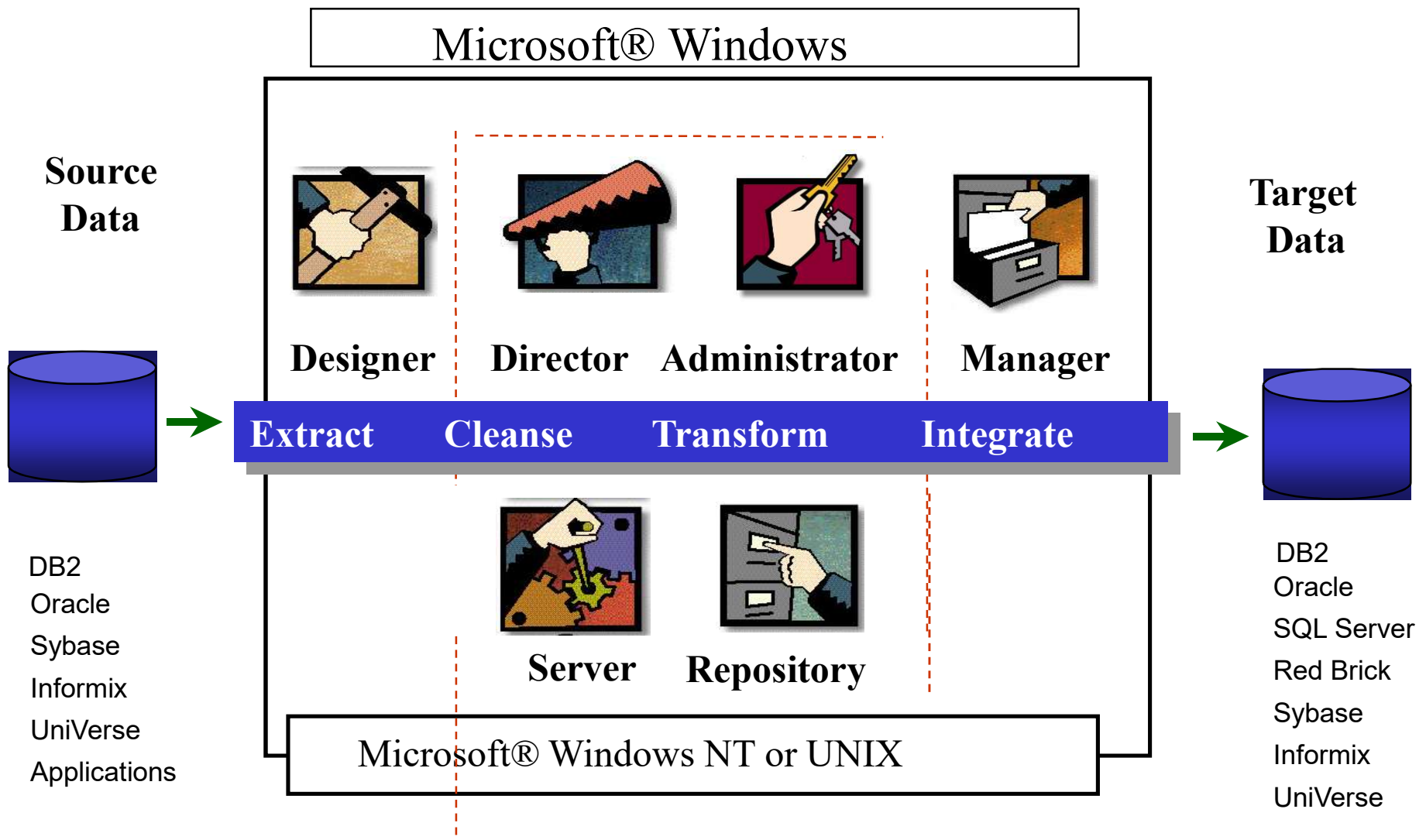
ETL Tool –DB2 Warehouse Manager



ETL Tool – Informatica PowerCenter



ETL Tool – IBM IIS Datastage



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: 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: Data Fabric solutions

Exercise E7.3:

Show the Highlights and build a Strengthens / Weakness Diagram for two of the Data fabric solutions mentioned in the lesson:

- IBM Cloud Pak for Data (see: video)
- Talend Data Fabric
- Informatica Intelligent Data Management Cloud (IDMC)
- Microsoft Azure Purview
- AWS Lake Formation
- Google Dataplex
- Altair RapidMiner (see [DHBW-Moodle], Category 3)

Show the two tools in competition to each other

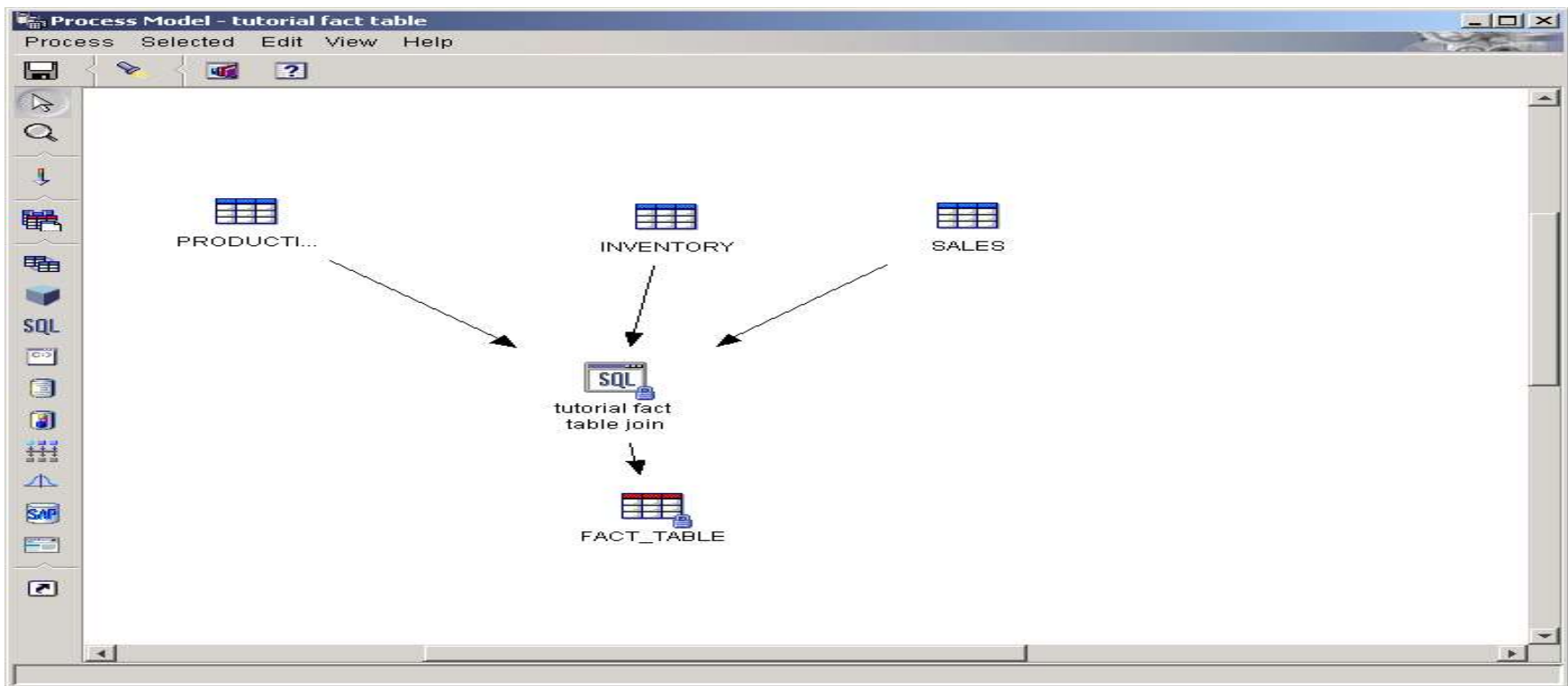
Exercise 4 to Lesson 7: Compare ETL and ELT Approach (AWS Redshift)

Exercise E7.4: 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]

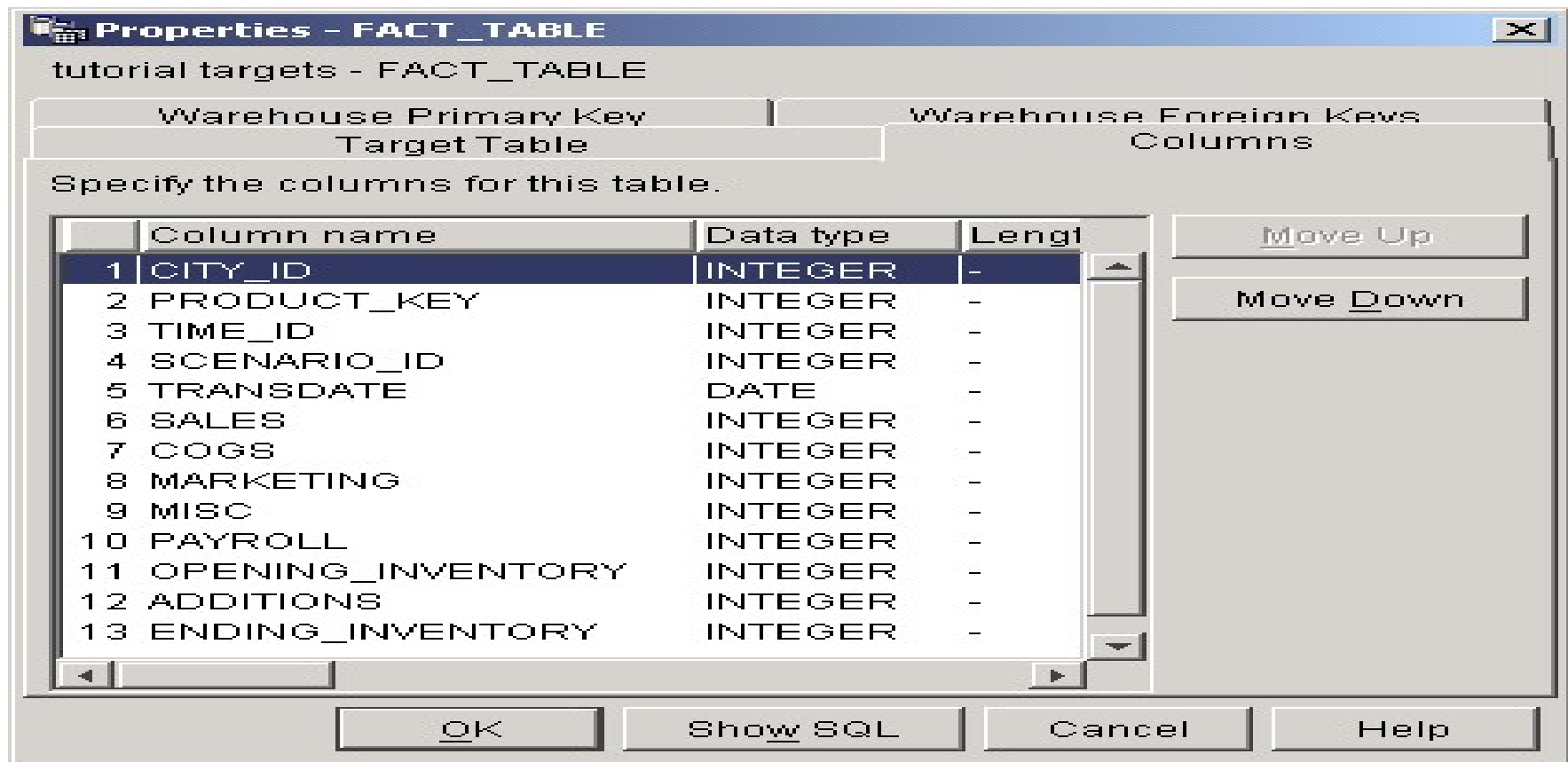
Optional: Exercise 5 to Lesson 7 – SQL Loading of a Fact Table (Part1)

Exercise E7.5: 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 5 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

Warehouse Primary Key | Warehouse Foreign Keys

Target Table | Columns

Specify the columns for this table.

	Column name	Data type	Length
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	-

Move Up

Move Down

OK Show SQL Cancel Help

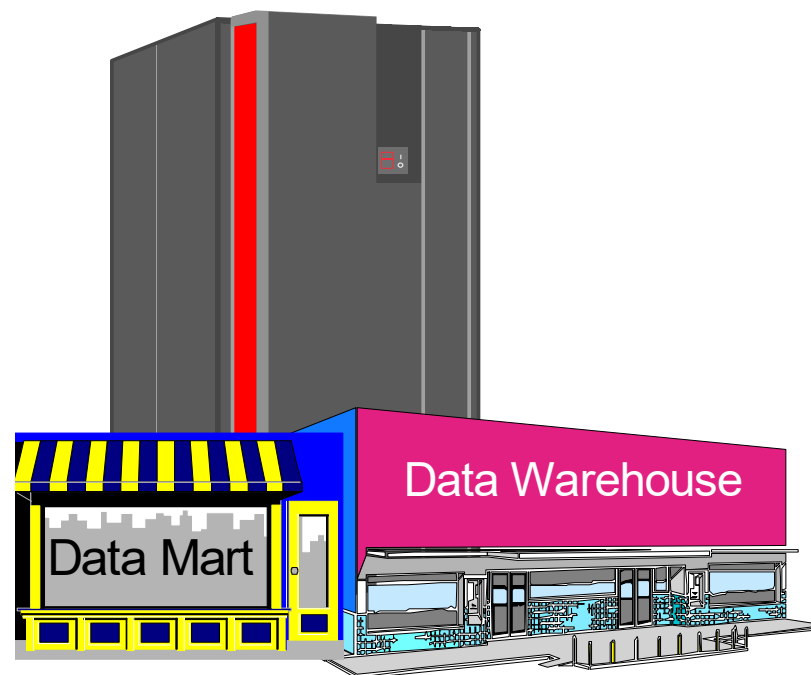
Category 1: Introduction & Architecture of DWH

Category 2: Databases and Data Modeling

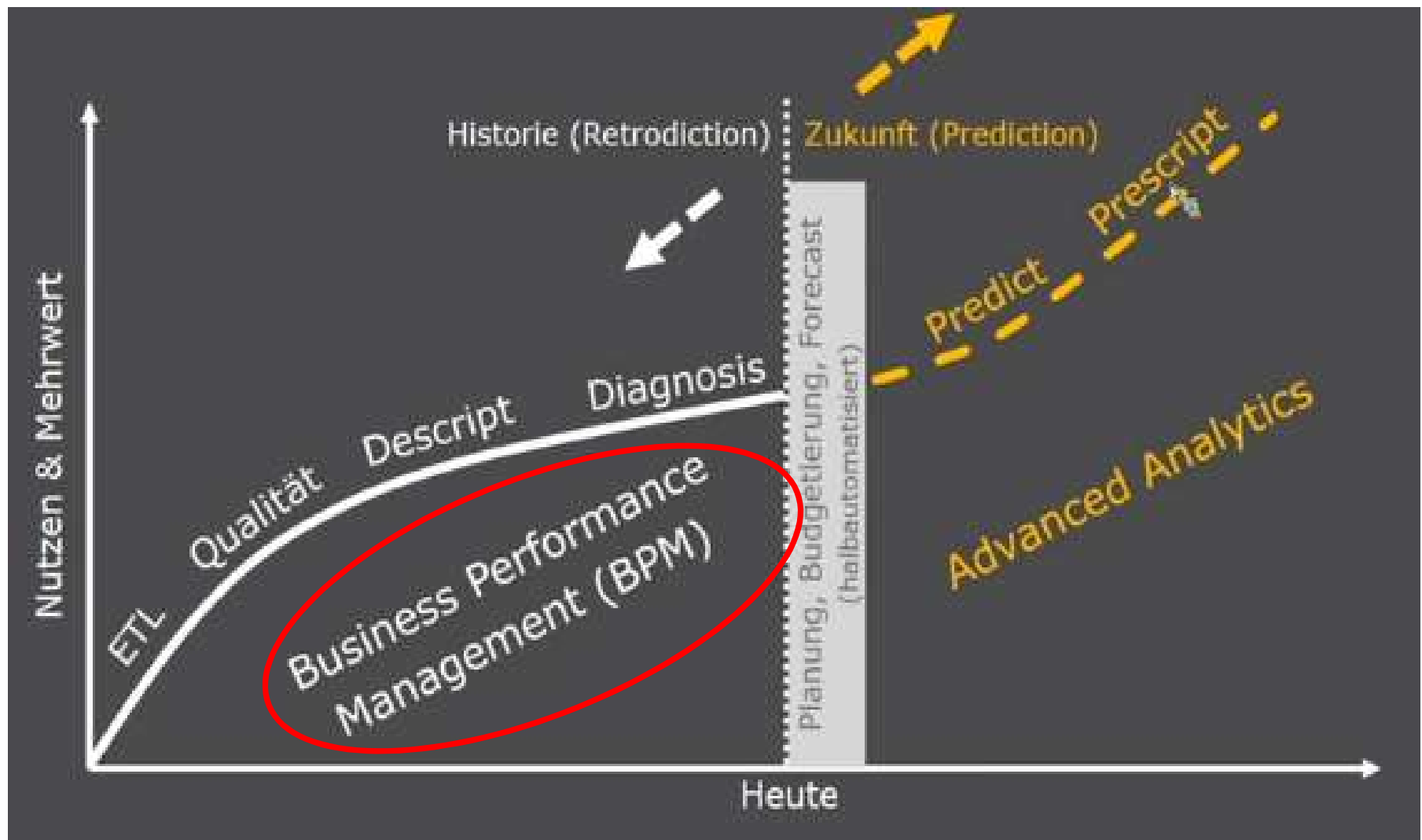
Category 3: ETL: Architecture & Technology

Category 4: Descriptive – & Advanced Analytics

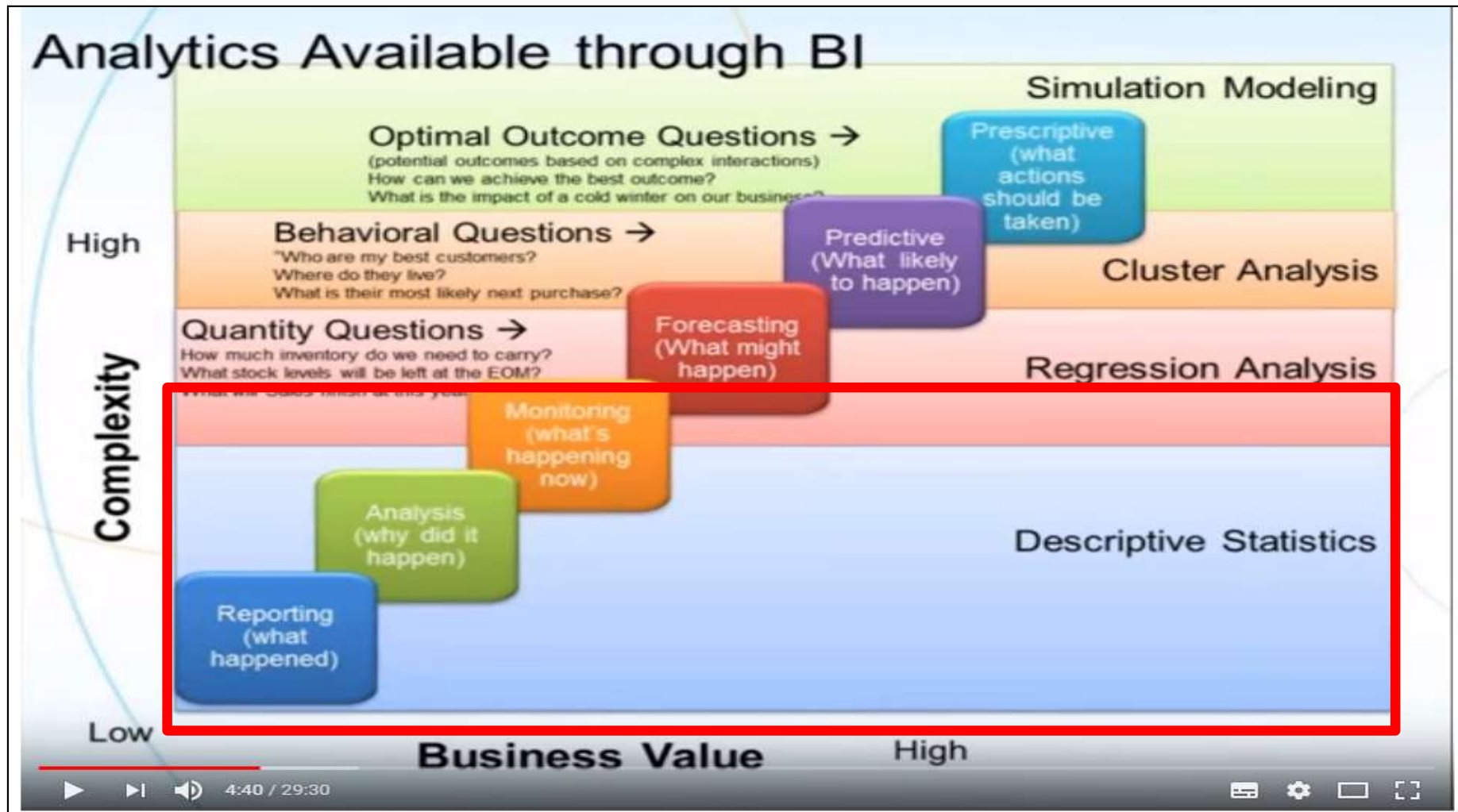
DW08 - Descriptive Analytics: Relational OLAP & Multdim. OLAP



Motivation - From Descriptive to Prescriptive Analytics



Descriptive to Prescriptive – Six Levels of Analytics



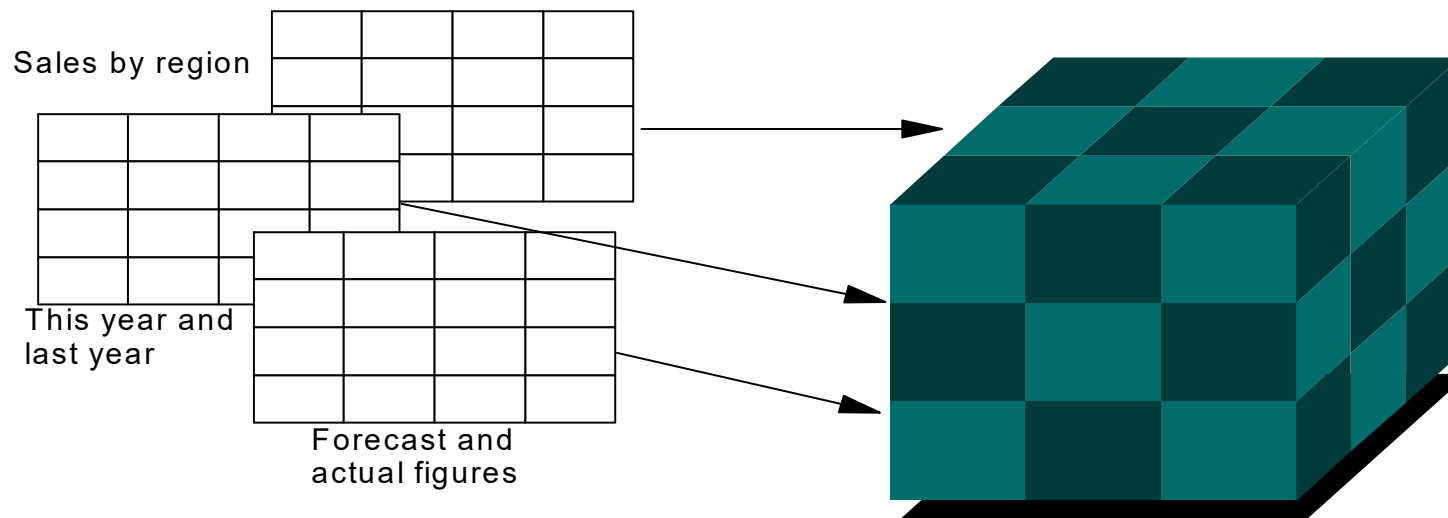
<https://www.youtube.com/watch?v=oNNk9-tmsZY>

Popular Descriptive Method = OLAP: What is OLAP?

- Stands for **OnLine Analytical Processing**
- 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.

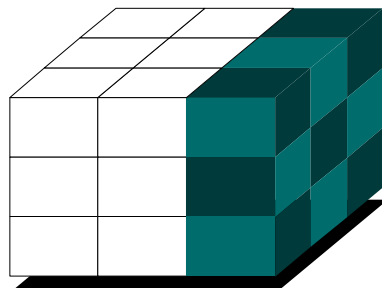


Multidimensional Database

- 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

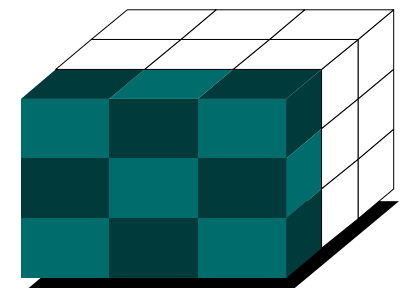
Multidimensional Views

Different selections give different ways of looking at the data

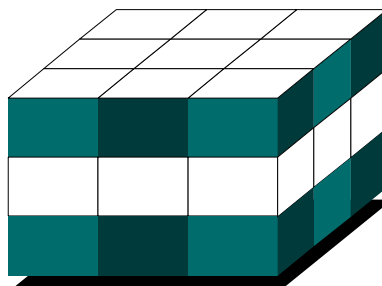


February for all products

		Audio		Jan		Feb		
		Video		Jan		Feb		et
TV			Jan		Feb			
			Actual	Budget	Actual	Budget		
Sales	Paris							
	Moscow							
	London							
	Total							
Costs	Paris							
	Moscow							
	London							
	Total							

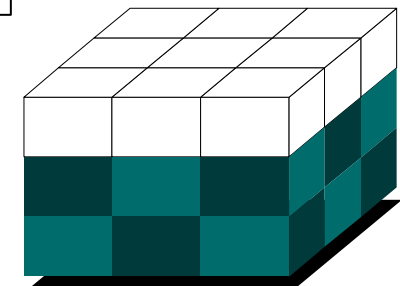


All TV information



Paris Sales and Costs

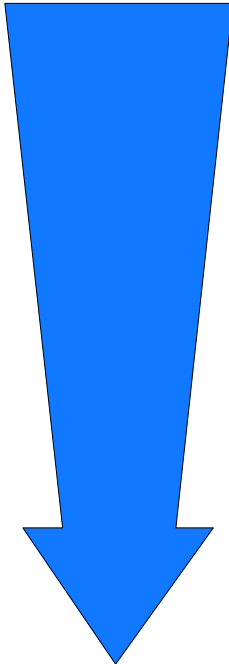
Viewing 5 dimensional database



All cost information

Drill Down

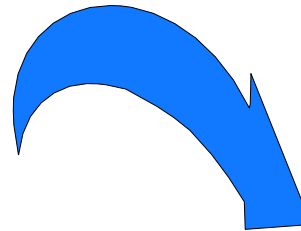
Looks at components in greater detail down same dimension

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

Slice and Dice

Change row, column
and page dimensions

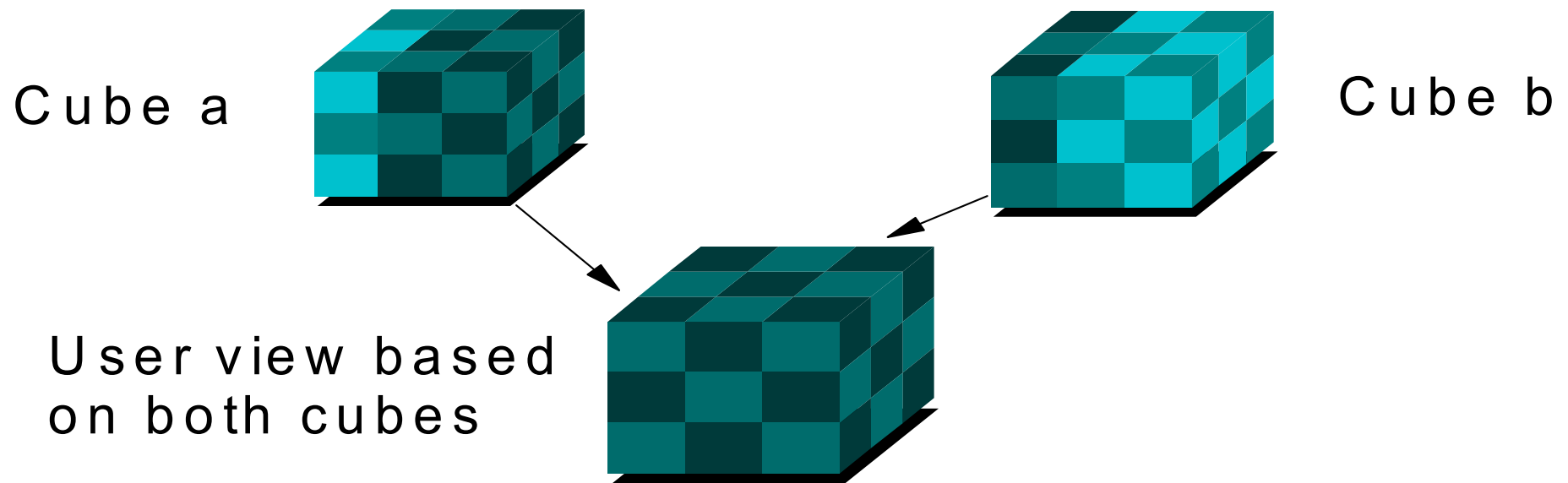
		Bud	Act	Bud	Act
1997	East				
	West				
1996	East				
	West				



		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



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

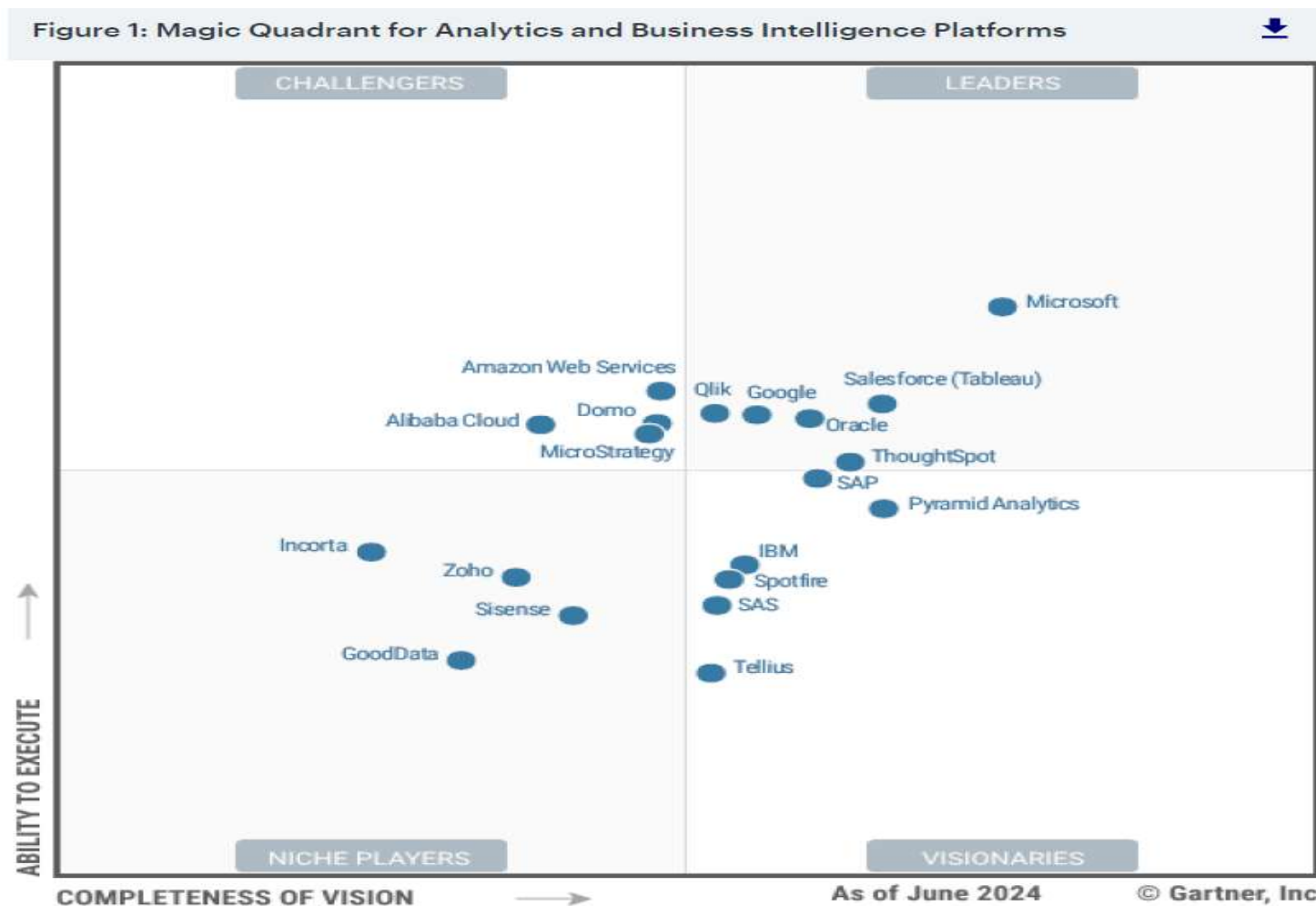
Differences

- 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

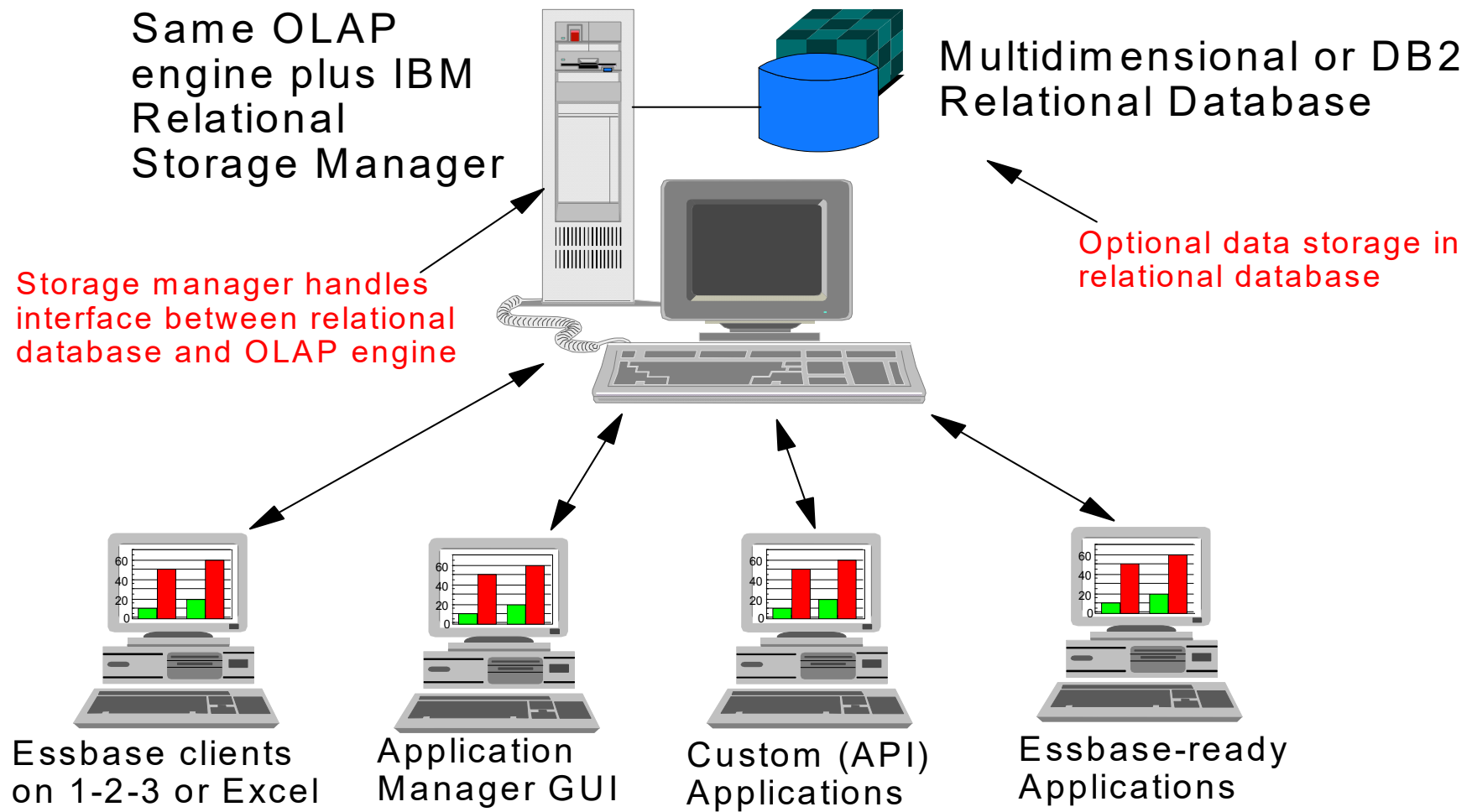
Benefits of MOLAP

- 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

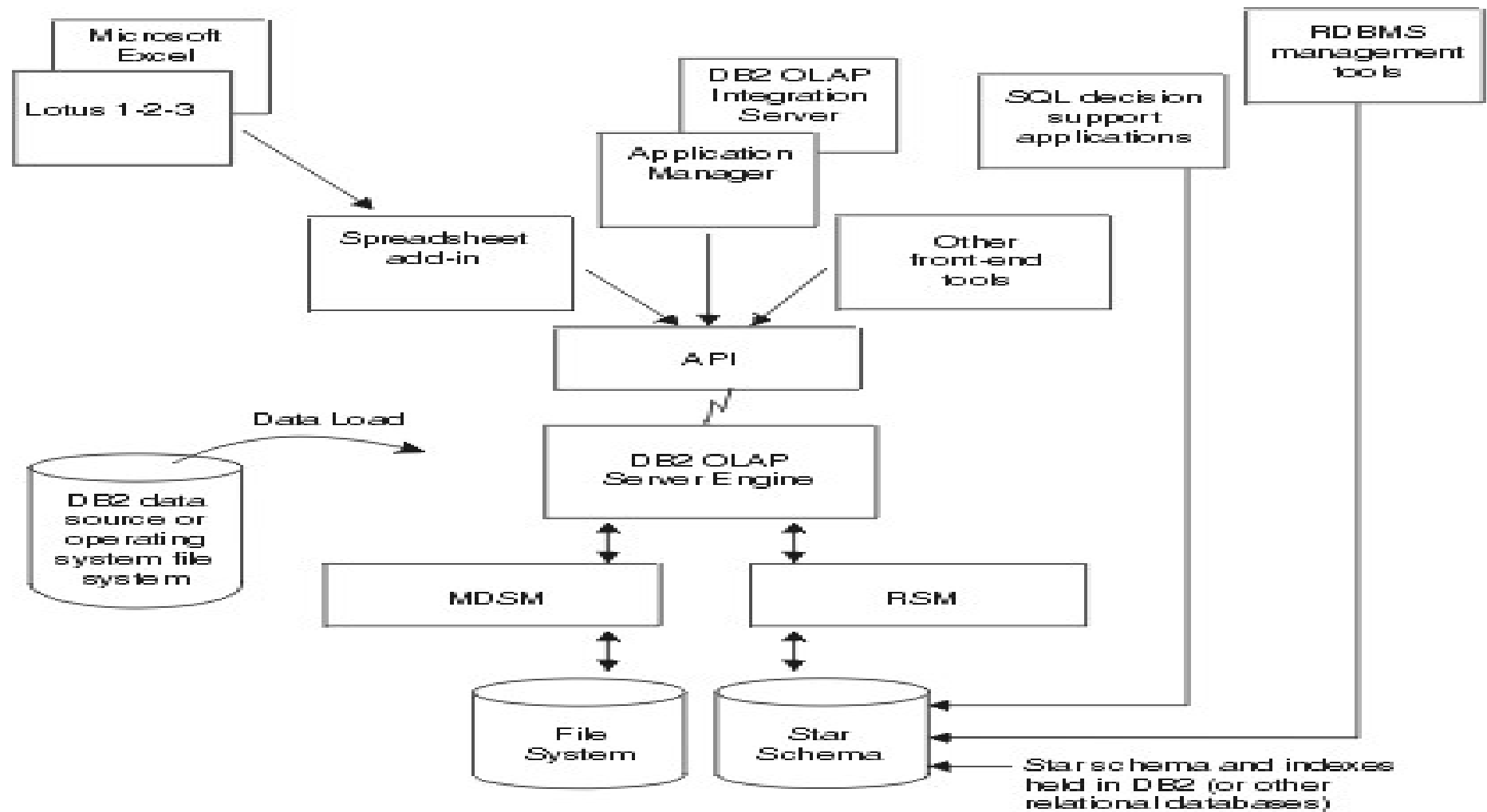
OLAP Marketplace & Tool Position



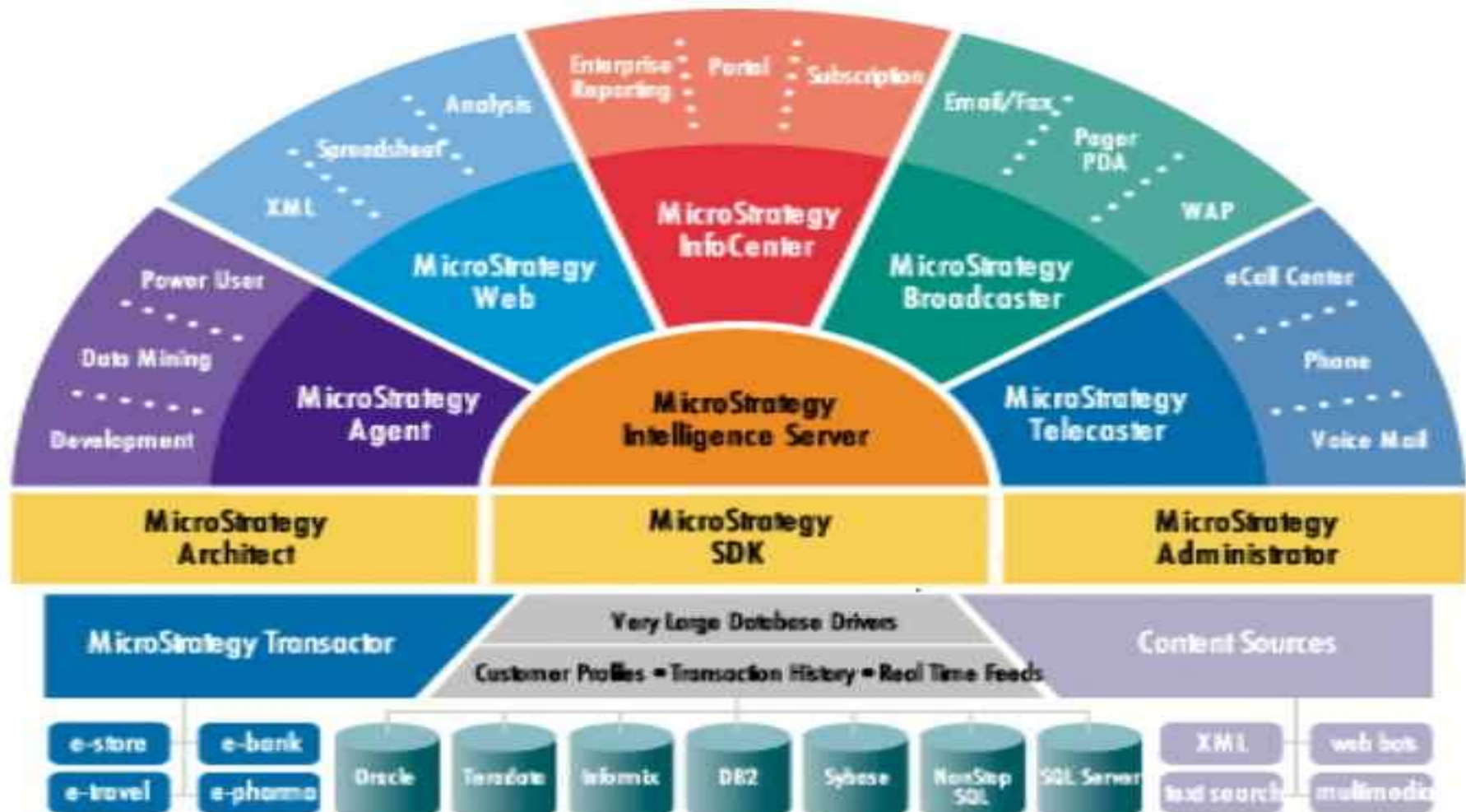
Example: IBM DB2 OLAP Server - Components



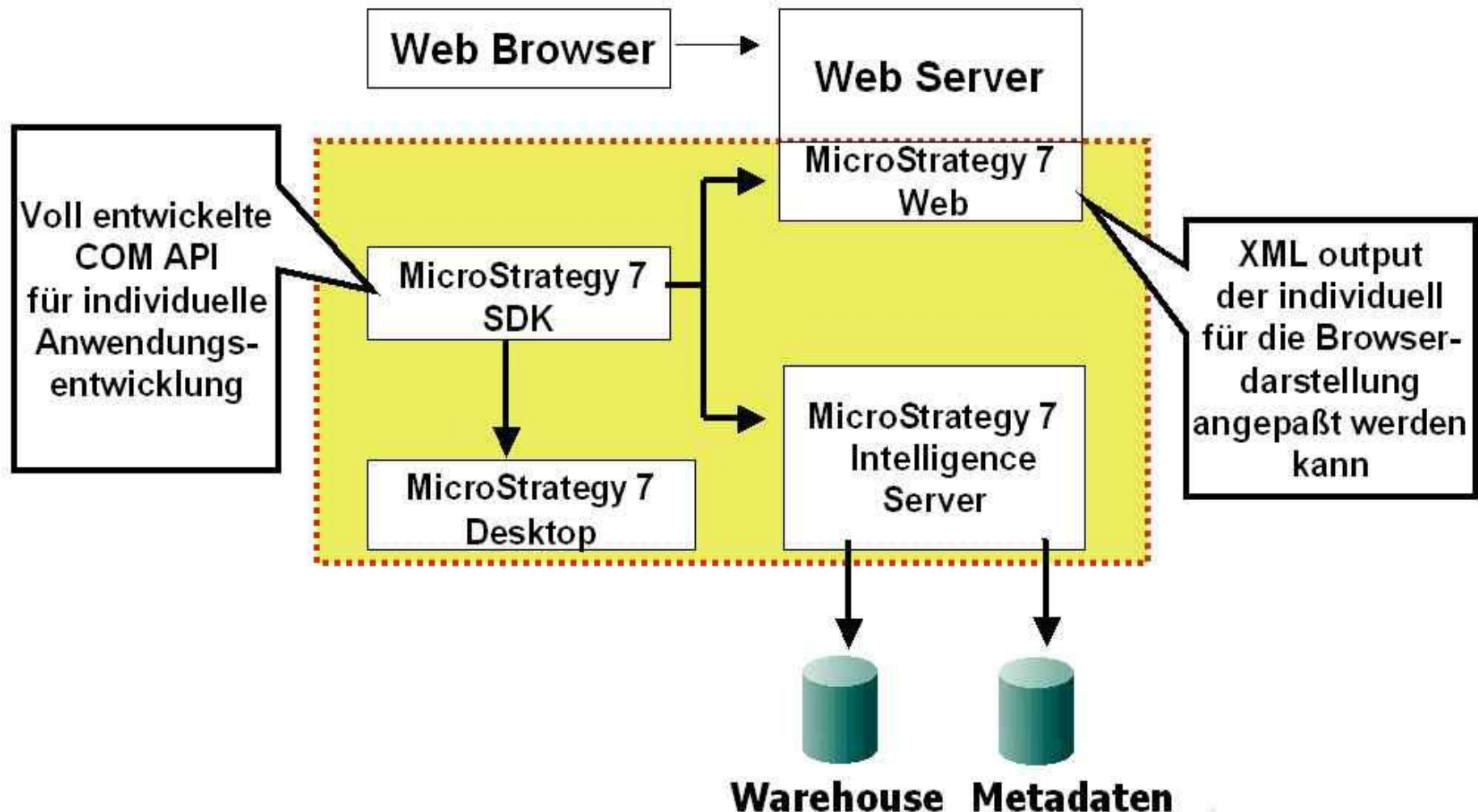
IBM DB2 OLAP Server - Architecture



ROLAP / MicroStrategy: Components Overview

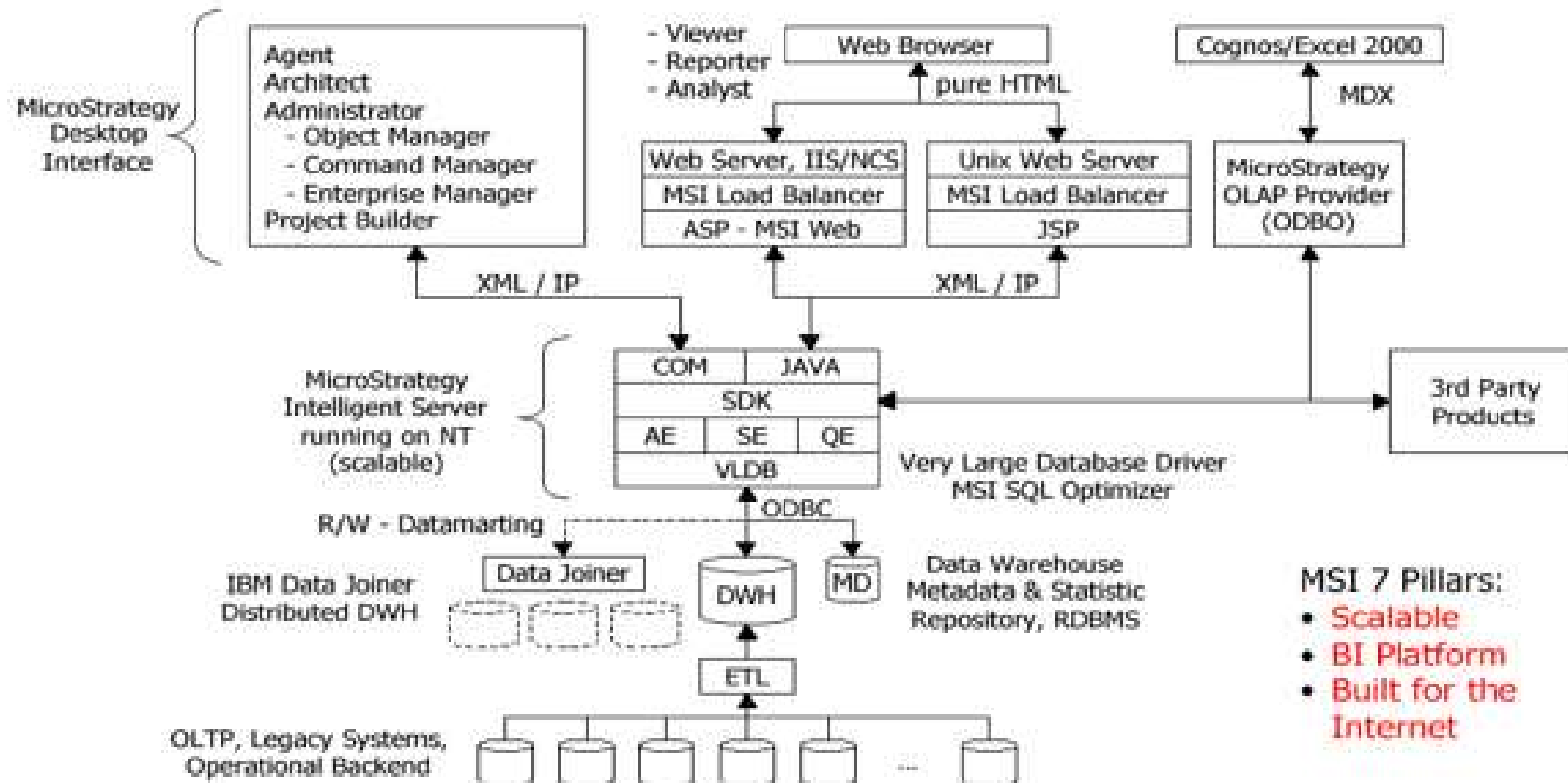


ROLAP Example - MicroStrategy: Analytical Model



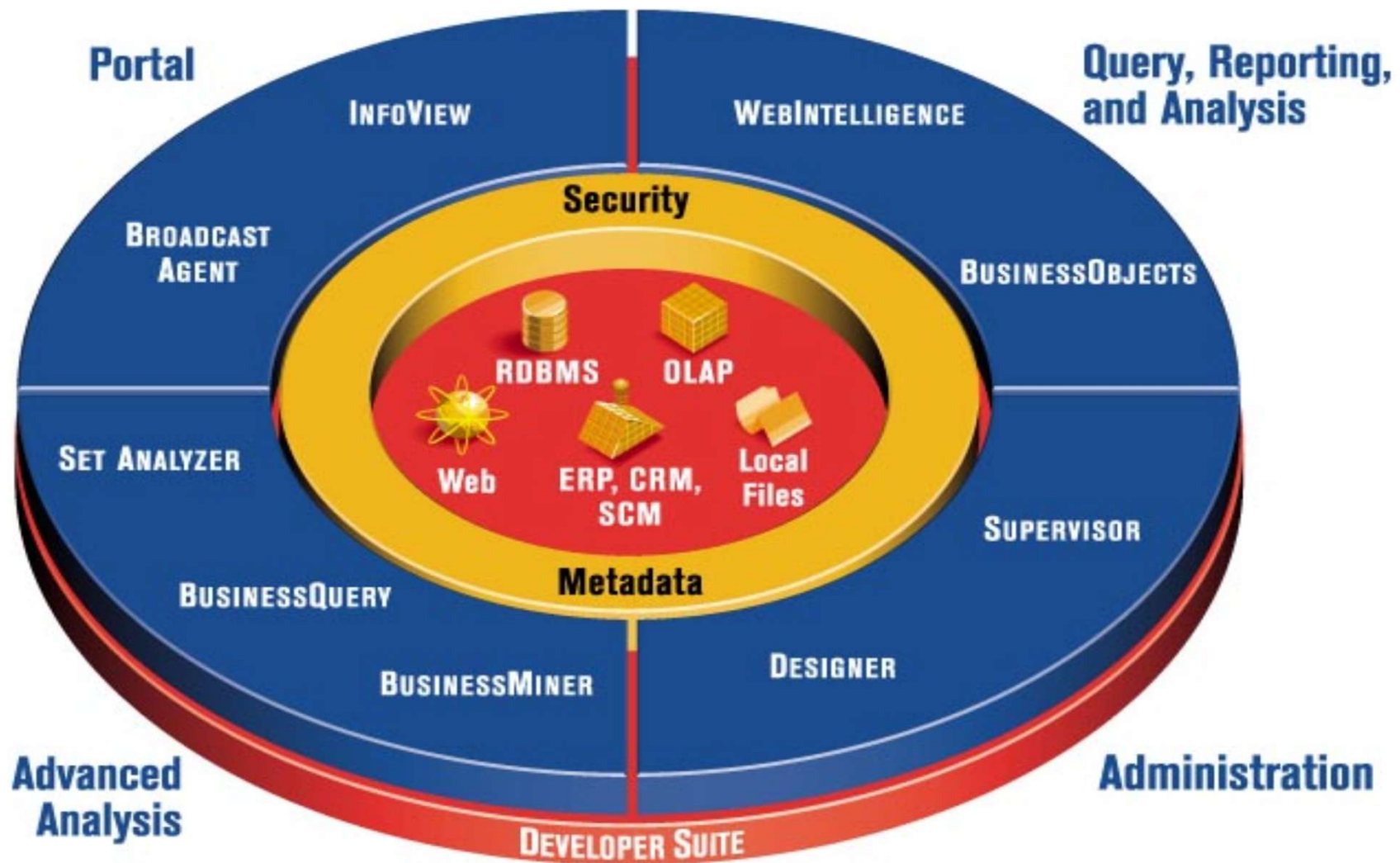
ROLAP Example - MicroStrategy: Big Picture

MicroStrategy Intelligence Server™



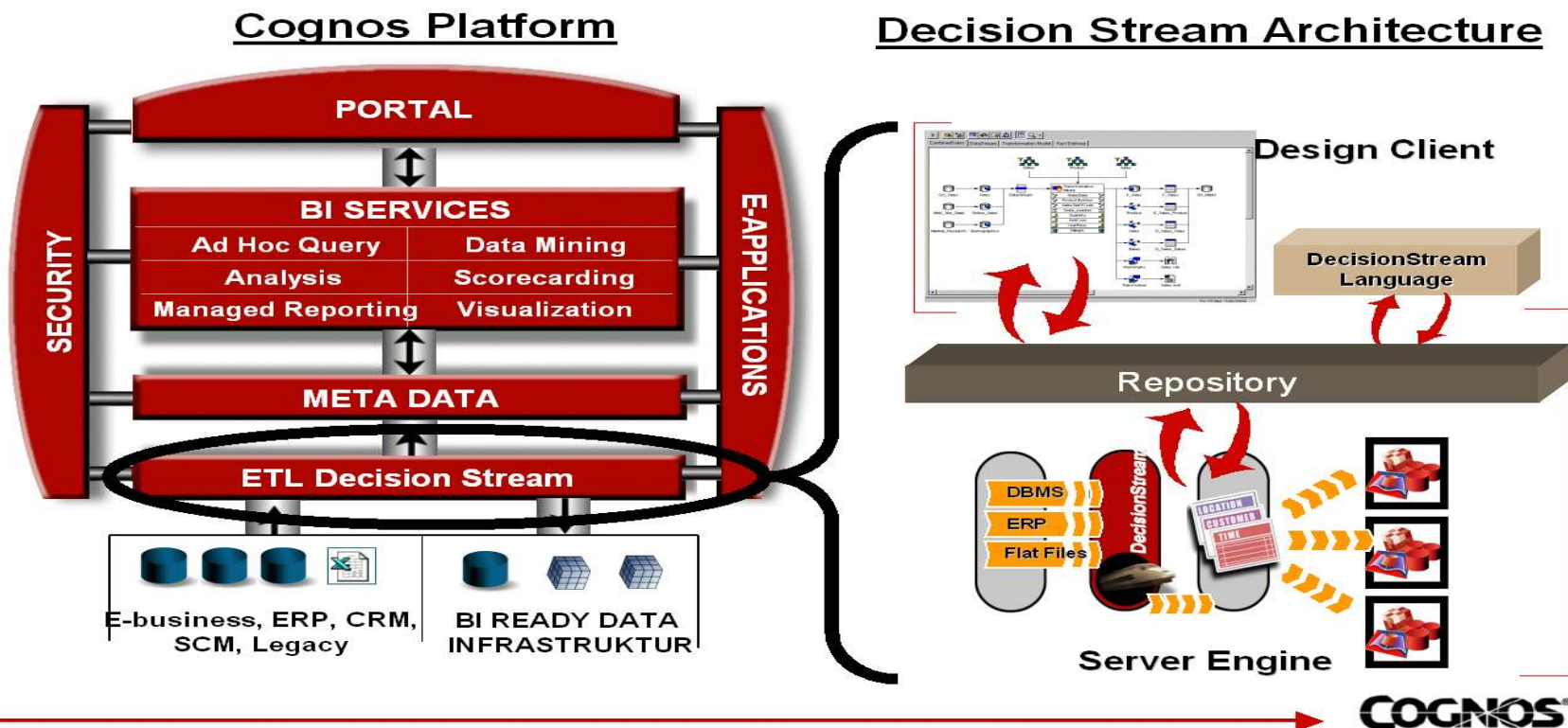
- MSI 7 Pillars:**
- Scalable
 - BI Platform
 - Built for the Internet

OLAP/Reporting Ex. - BusinessObject /Big Picture



OLAP/Reporting Ex. - Cognos / Big Picture

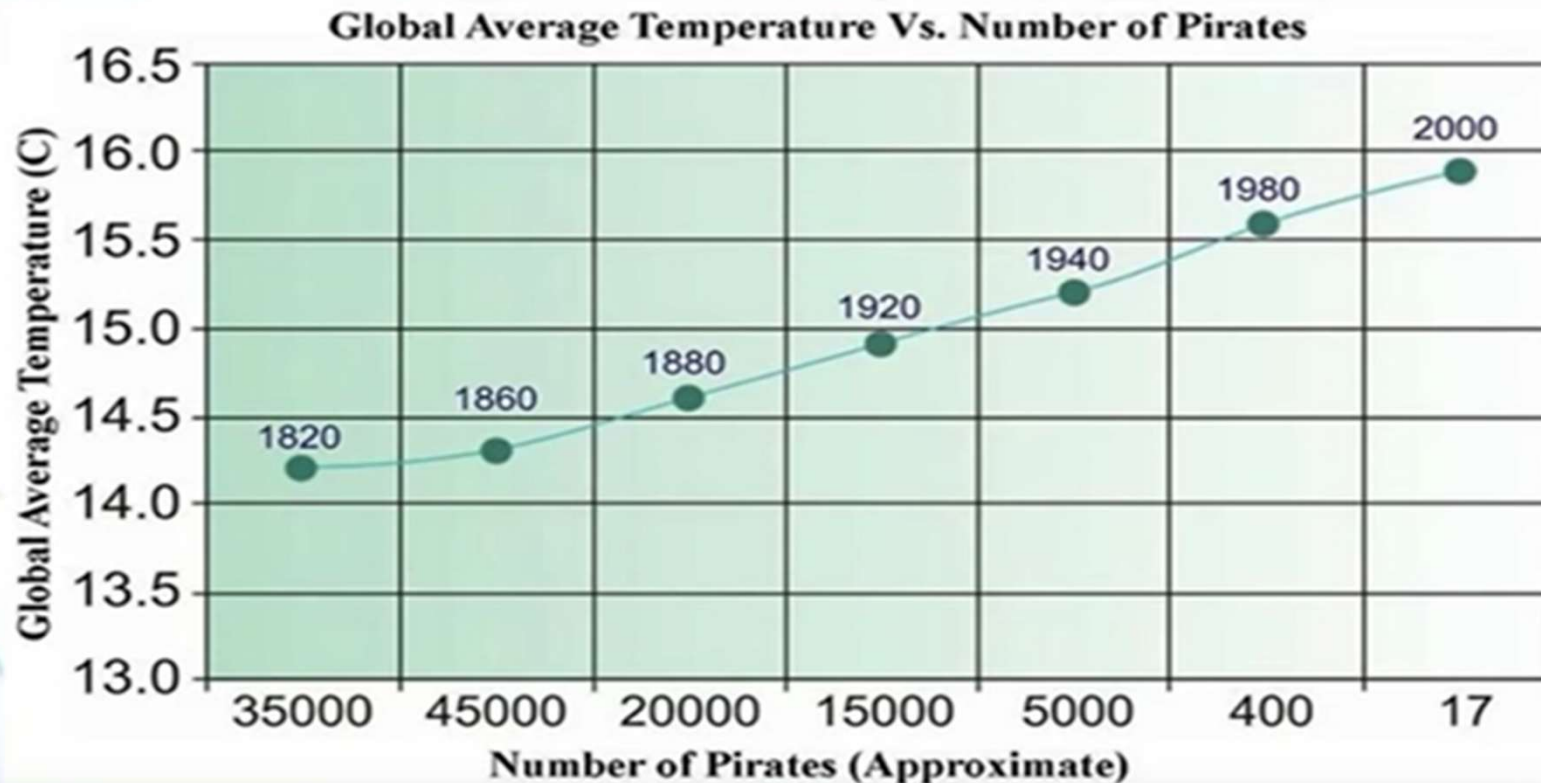
DecisionStream und die Cognos e-BI Lösung



See the following video about analytical dashboards in Data-Scientist/Dashboards (w. Cognos Dashboard Embedded):
<https://datapatform.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.

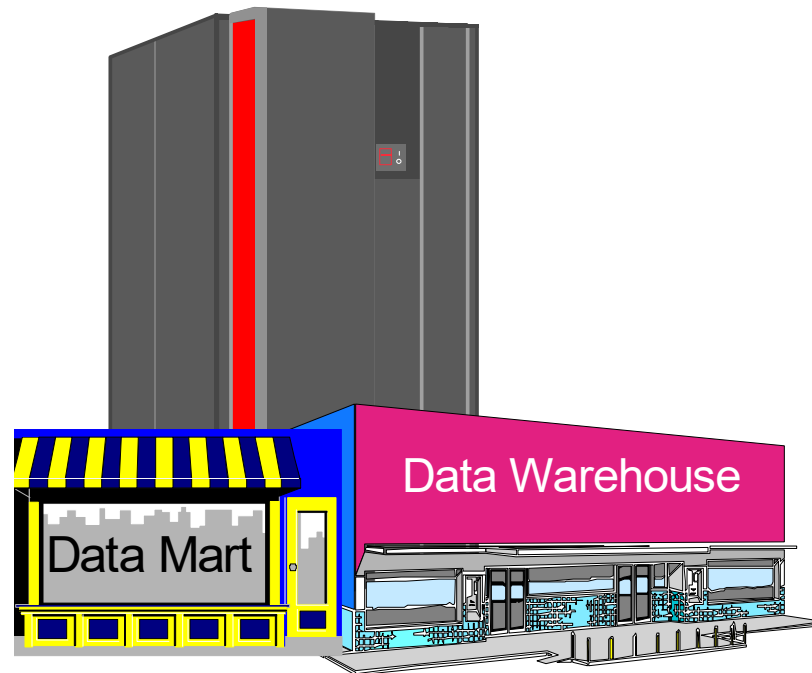
Category 1: Introduction & Architecture of DWH

Category 2: Databases and Data Modeling

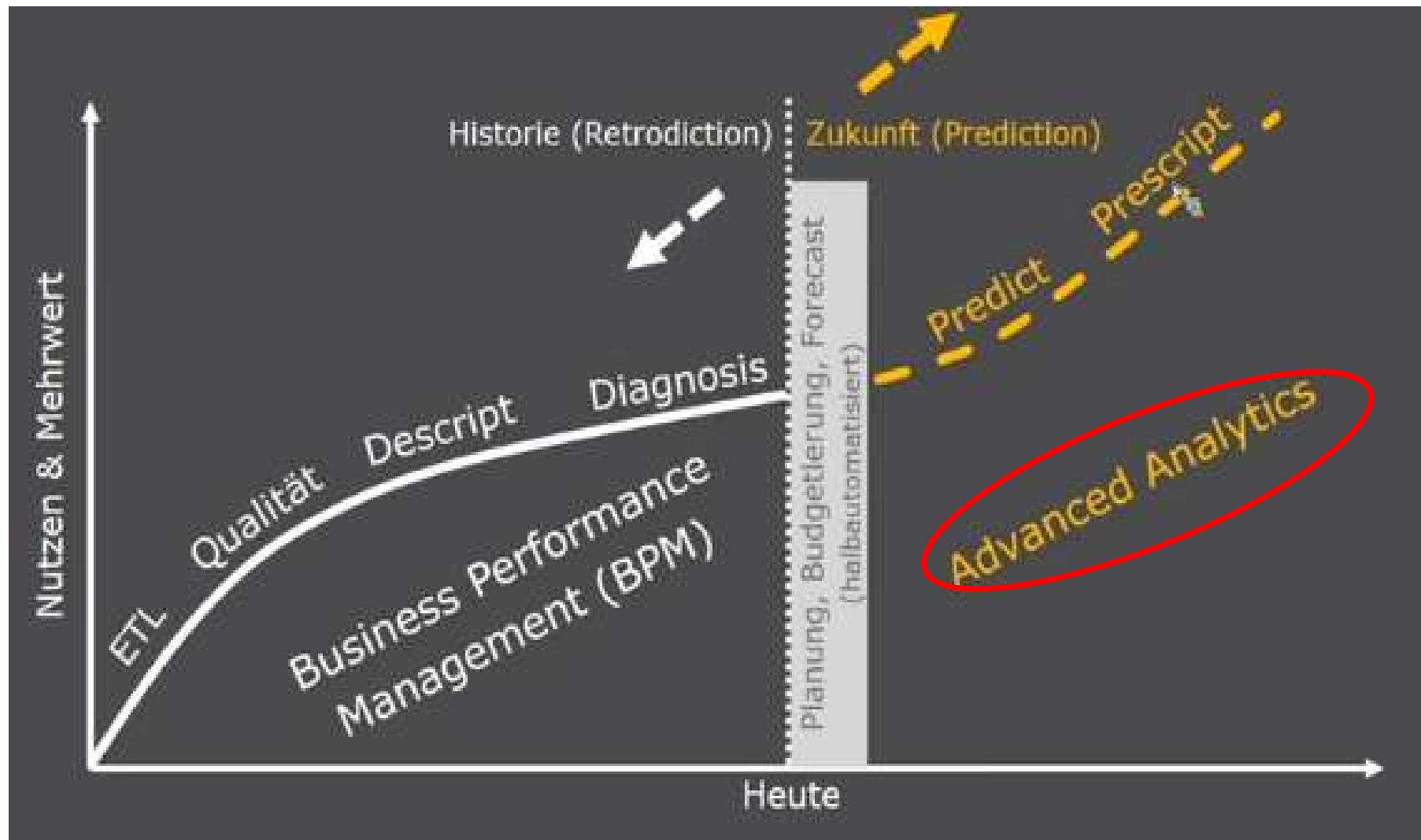
Category 3: ETL: Architecture & Technology

Category 4: Descriptive – & Advanced Analytics

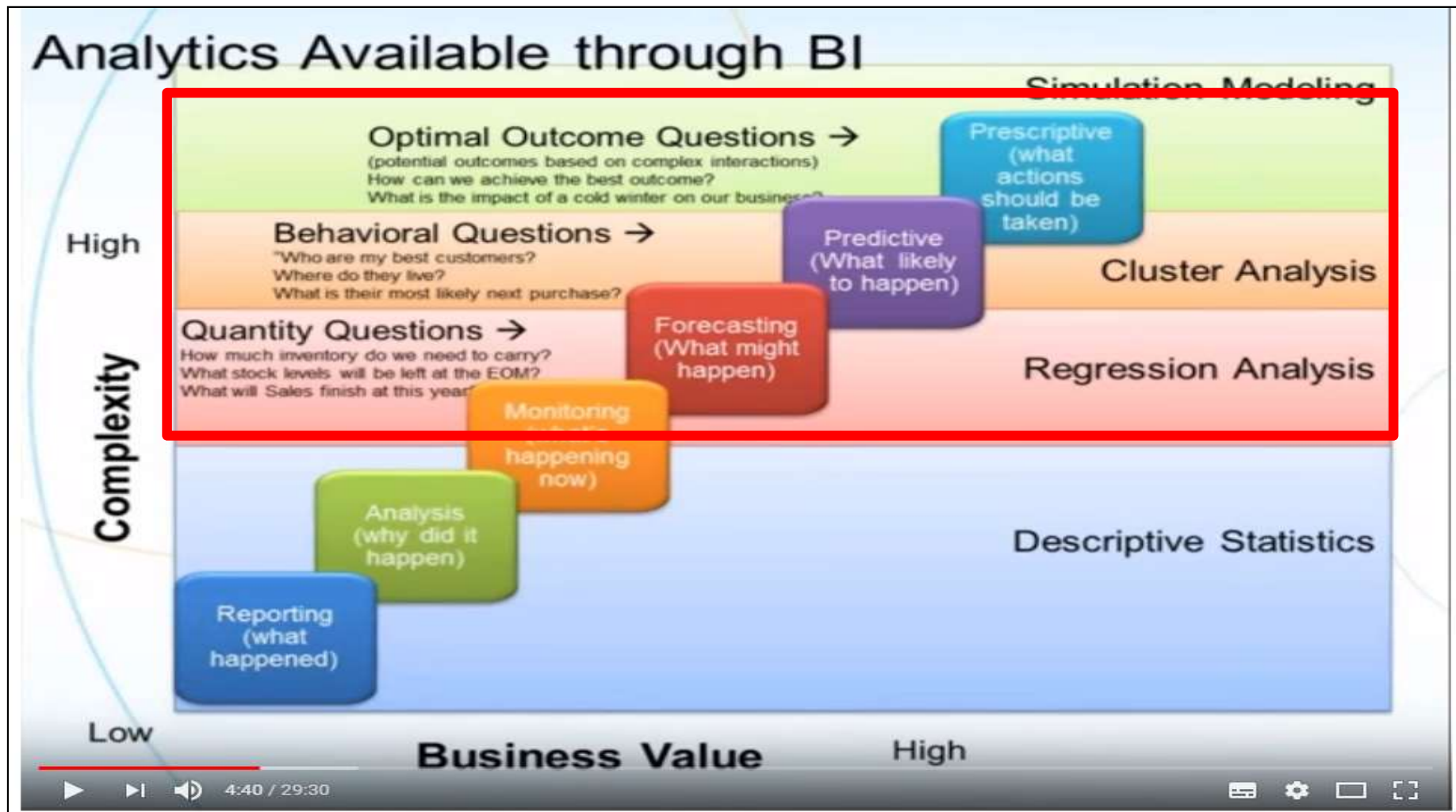
DW09 - Advanced Analytics I: Data Mining - Introduction & First Methods



Motivation - From Descriptive to Prescriptive Analytics

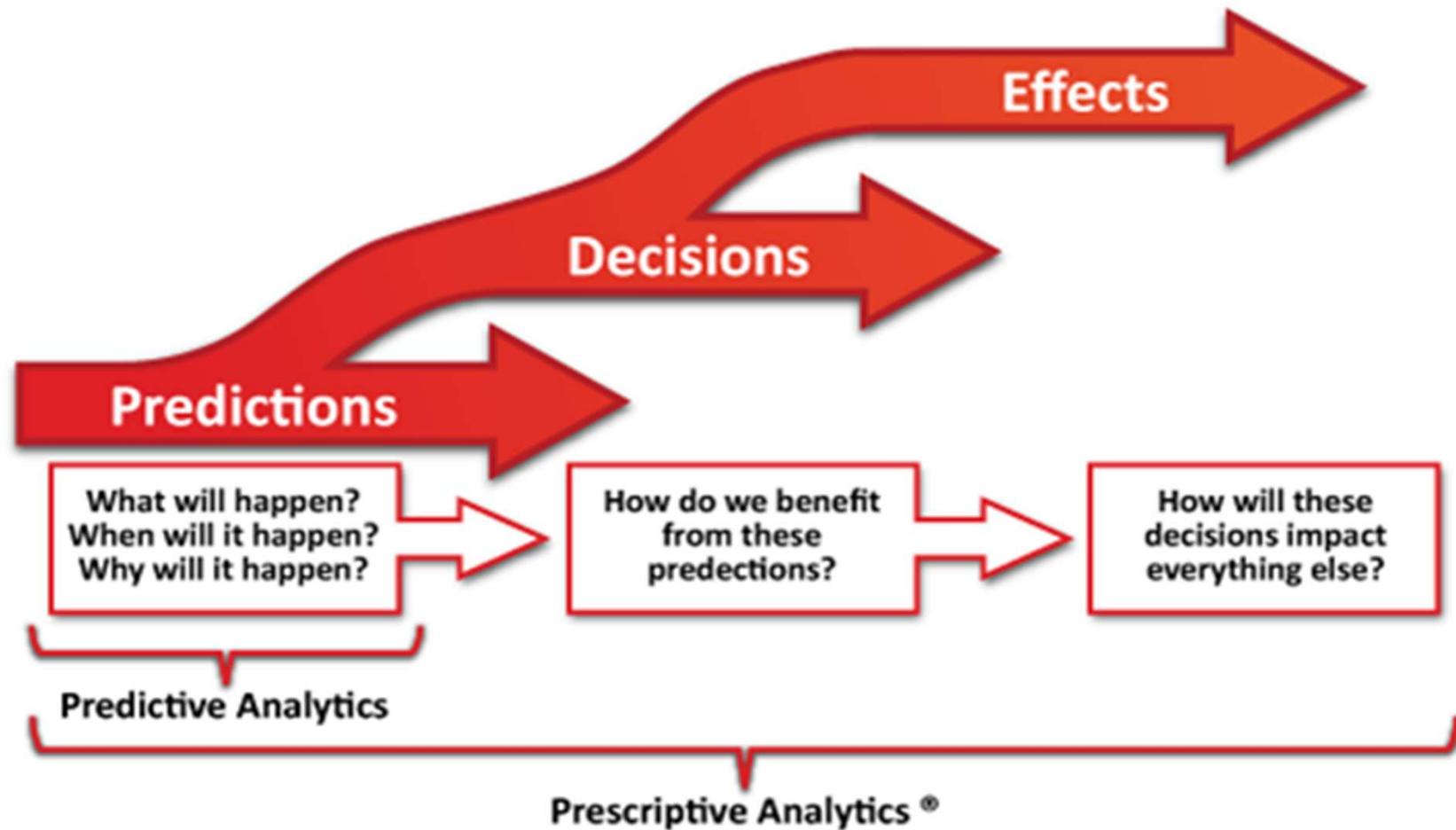


Advanced Analytics (AA) – Six Levels of Analytics

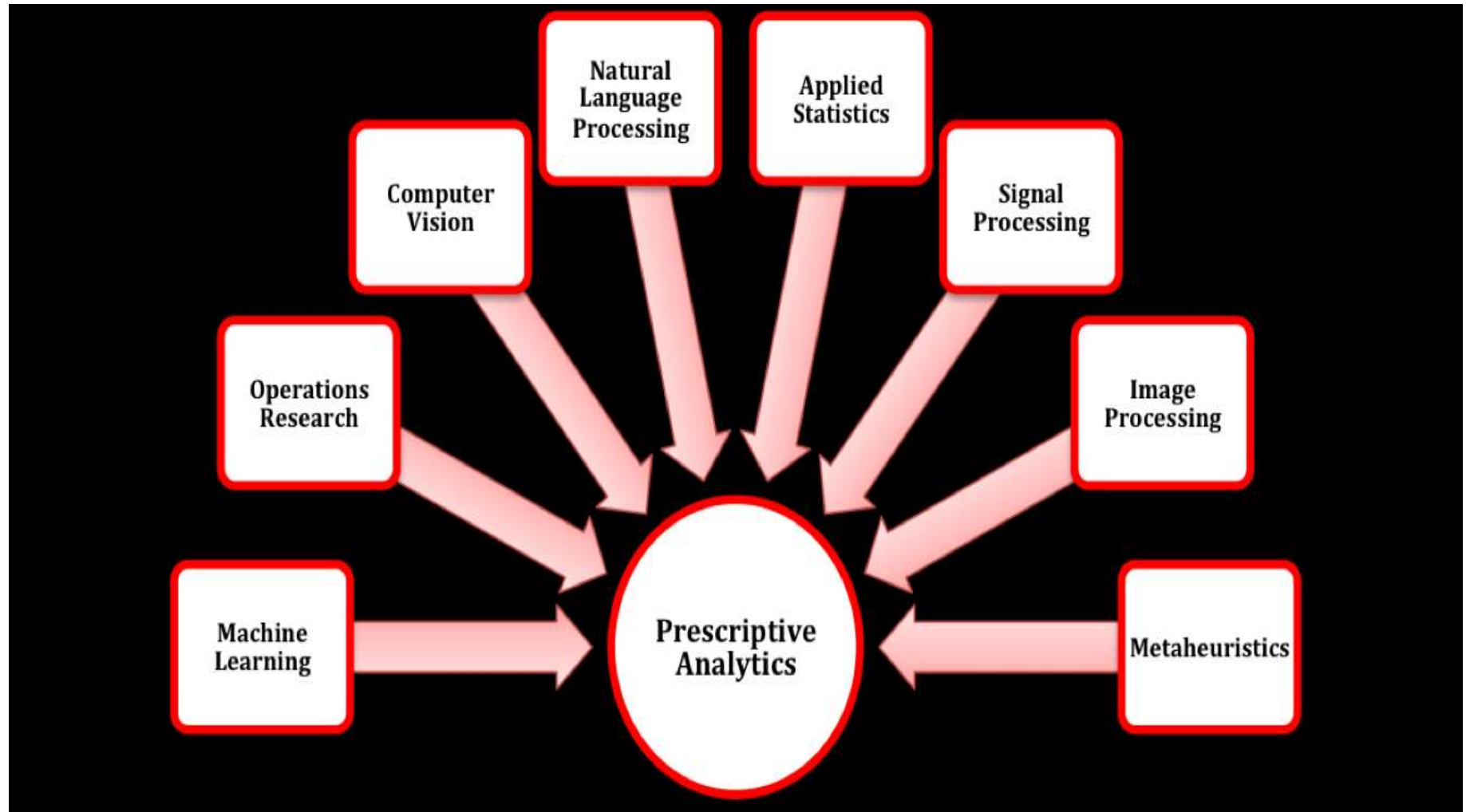


<https://www.youtube.com/watch?v=oNNk9-tmsZY>

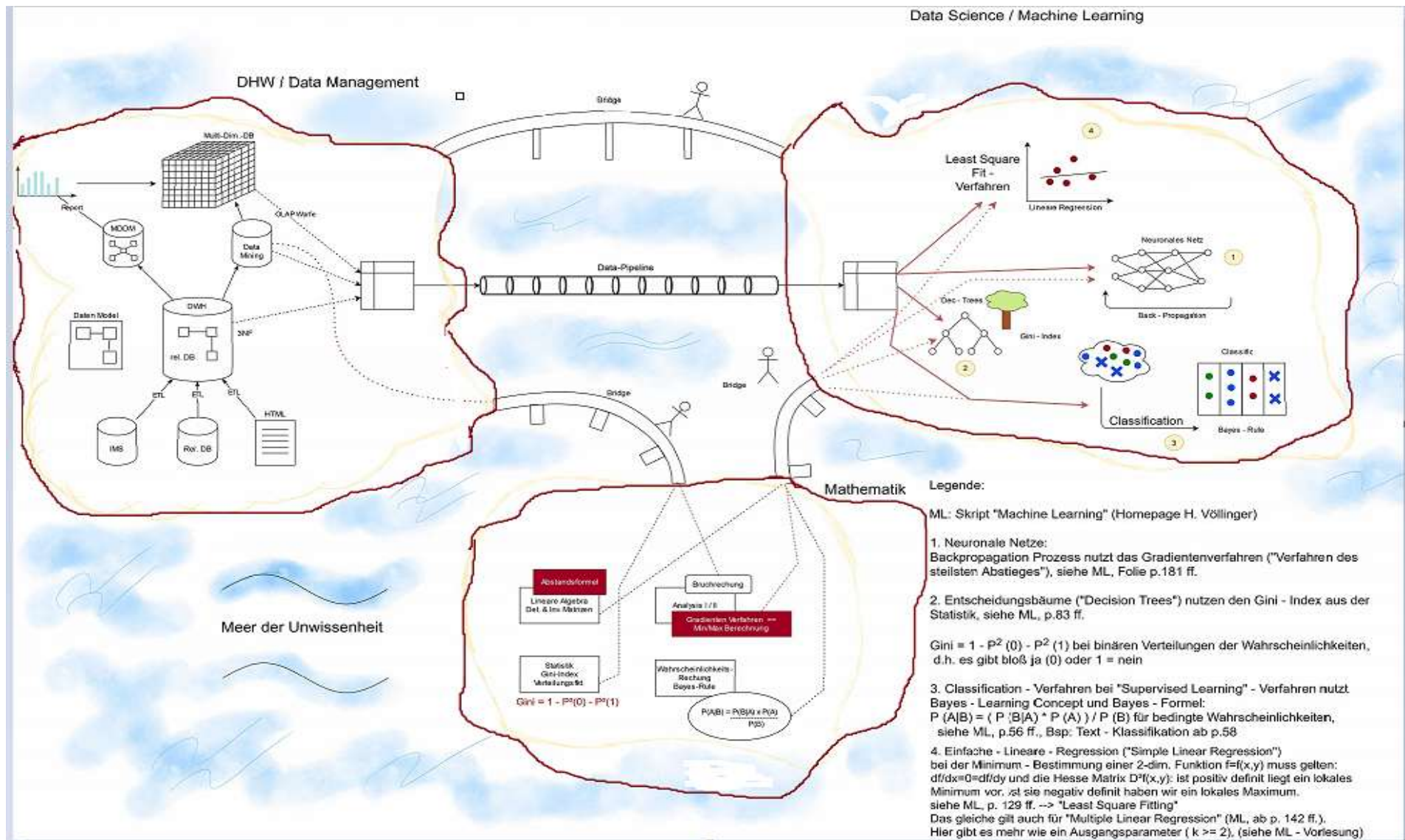
Advanced Analytics – Prescriptive Analytics



Prescriptive Analytics – Using Data Scientific Methods



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



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

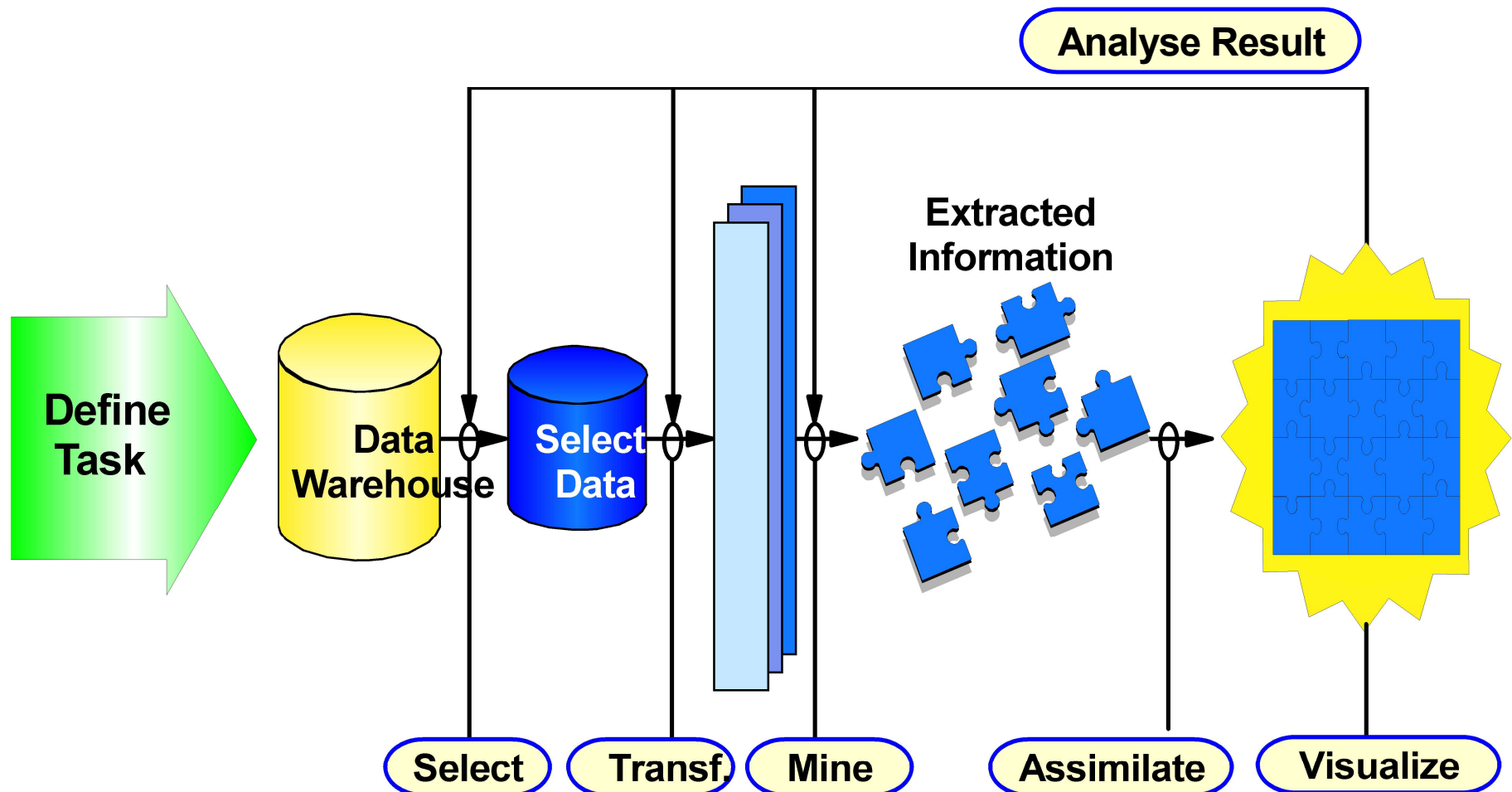
Data Mining is ...

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

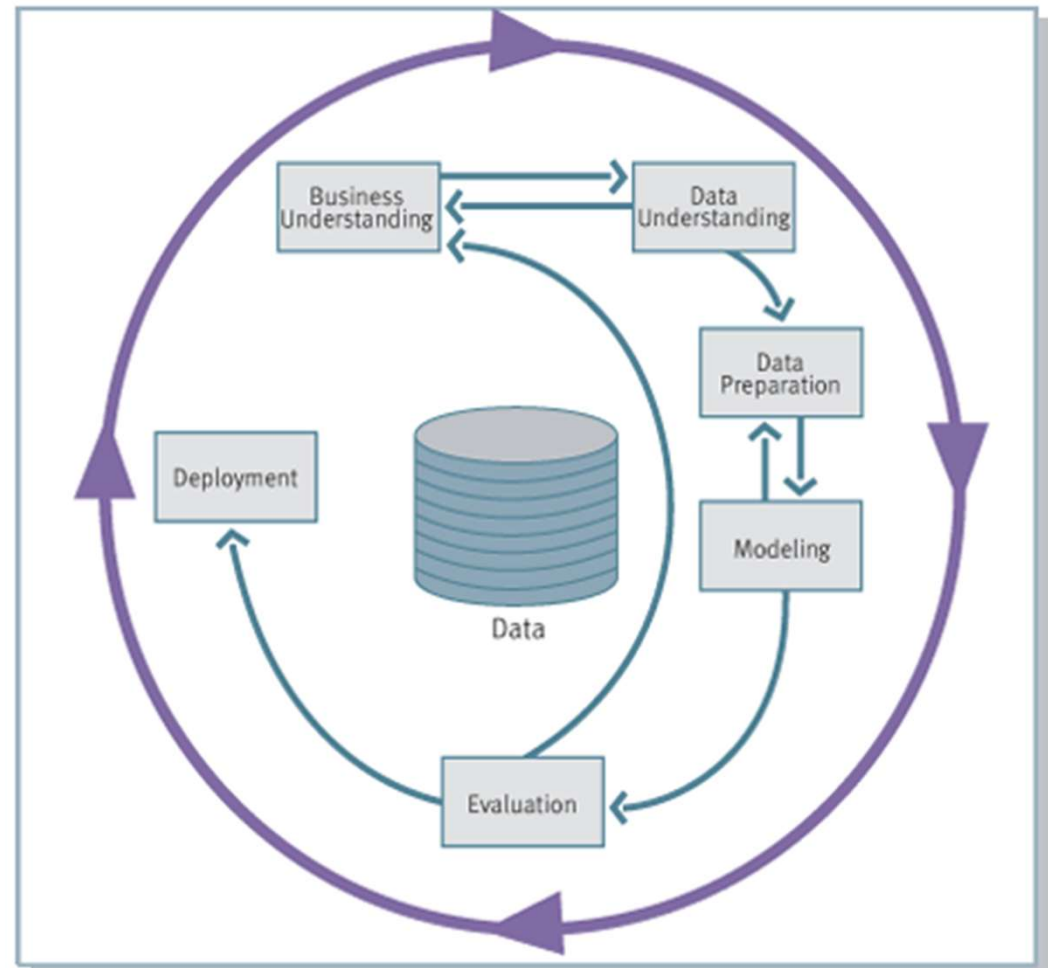
The Data Mining Process



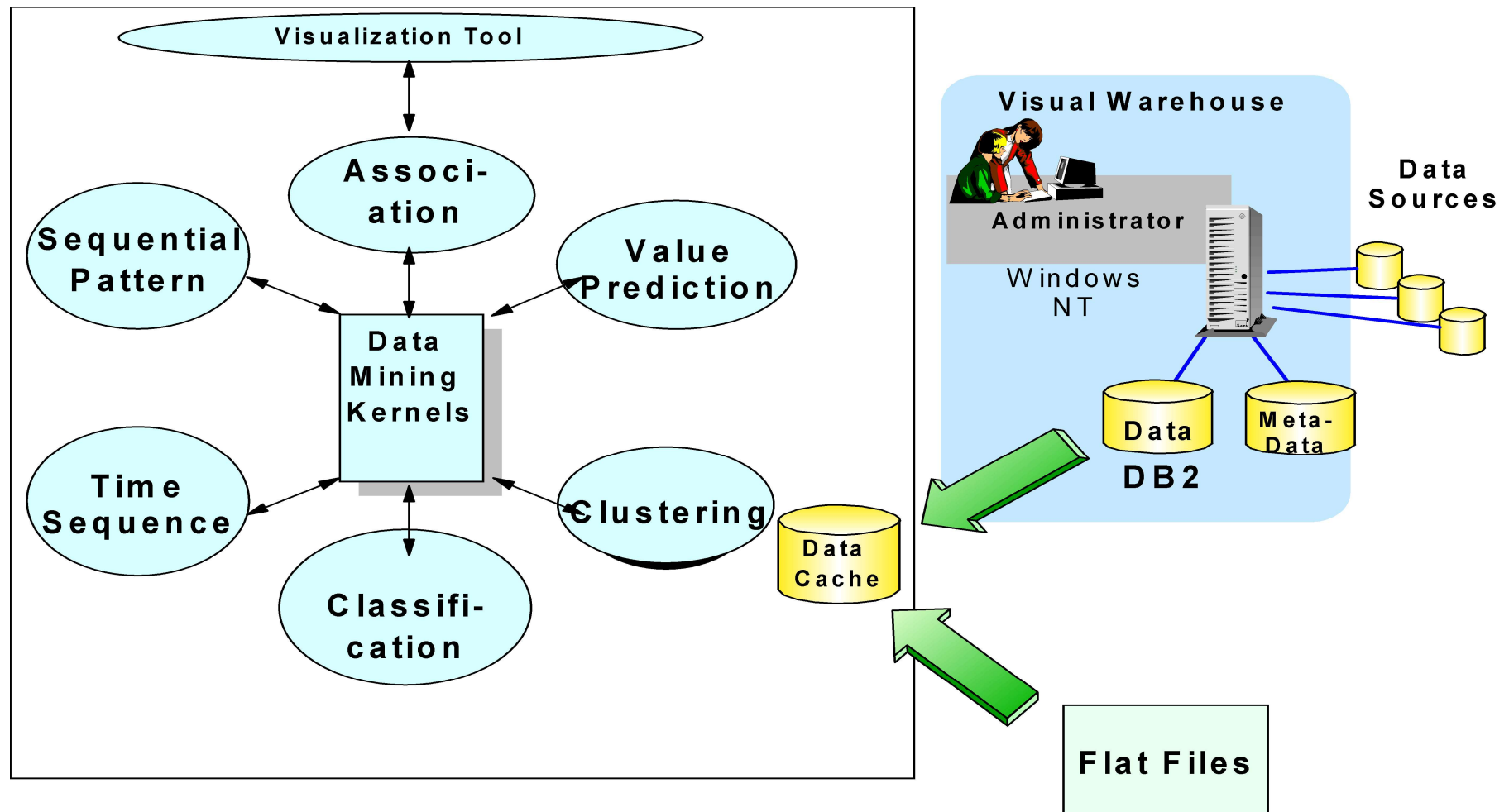
The CRISP*- DM Process Model

* 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



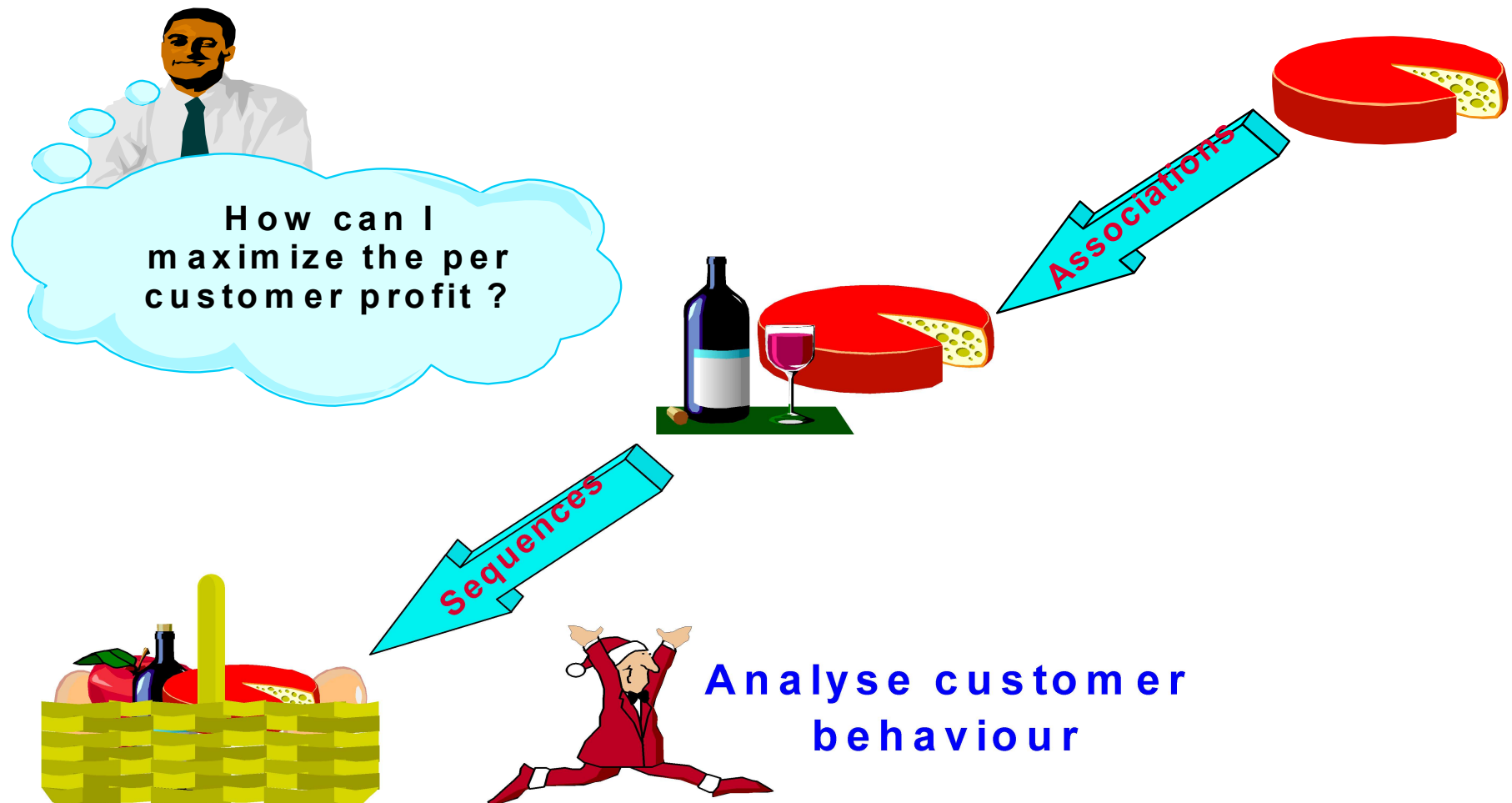
See the following video about the tool SPSS in Data-Scientist/SPSS Modeler ("Score prediction- diagnose diseases..."):
<https://datapatform.cloud.ibm.com/docs/content/wsj/getting-started/videos.html?audience=cpdaas&context=cpdaas#data-scientists>

Overview about Data Mining Applications

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

<u>No.</u>	<u>Application</u>	<u>IM4D Technique</u>
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

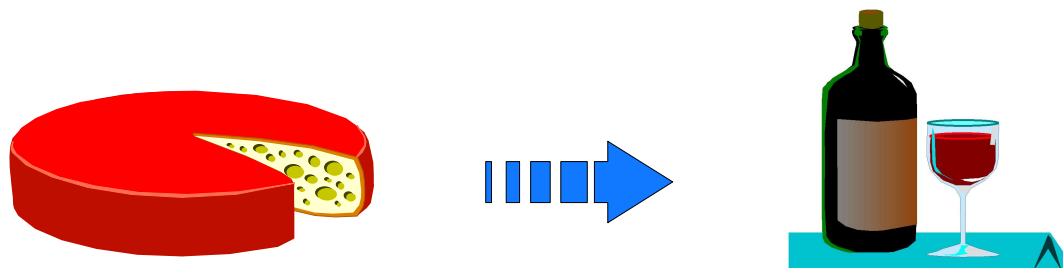


Market Basket Analysis - Associations

- 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:

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

- Interpretation:

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

- Examples:

- Bread, Cheese $\rightarrow$ RedWine.**

Customers that buy bread and cheese, also tend to buy red wine.

- MachineLearning $\rightarrow$ WebMining, MLPraktikum.**

Students that take 'Machine Learning' also take 'Web Mining' and the 'Machine Learning Praktikum'

Association Rules – Definition of Popular Measures

Rule: $X \Rightarrow Y$

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

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

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

Symmetry Properties:

- $\text{Sup}(X \Rightarrow Y) = \text{Sup}(Y \Rightarrow X)$
- $\text{Lift}(X \Rightarrow Y) = \text{Lift}(Y \Rightarrow X)$

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



The screenshot shows the Amazon.com interface for a product page. At the top, the Amazon logo and navigation links are visible. The search bar shows 'Health & Personal Care'. The product title is 'Adult Reusable Cotton/Poly Snap Diaper - Large - Fits 32" - 46" - Each' by 'Comfort Concepts'. The price is \$15.05, and it is marked 'In stock'. Below the product image, there is a section 'Frequently Bought Together' showing the diaper and a 'Call of Duty 4: Modern Warfare Game of the Year Edition' by Activision. The price for both is \$40.11. The page also includes customer reviews and shipping information.

amazon.com Hello. Sign in to get personalized recommendations. New customer? [Start here](#).
Your Amazon.com | [Today's Deals](#) | [Gifts & Wish Lists](#) | [Gift Cards](#)

Shop All Departments [Health & Personal Care](#) Search [Health & Personal Care](#) [Browse Products](#) [Bestsellers](#) [Health Care](#) [Personal Care](#) [Shaving & Hair Removal](#) [Nutrition & Fitness](#) [Sexual We](#)

Adult Reusable Cotton/Poly Snap Diaper - Large - Fits 32" - 46" - Each
by [Comfort Concepts](#)
★★★★★ (3 customer reviews) [Like](#) (0)

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
[Add both to Cart](#) [Add both to Wish List](#)
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.

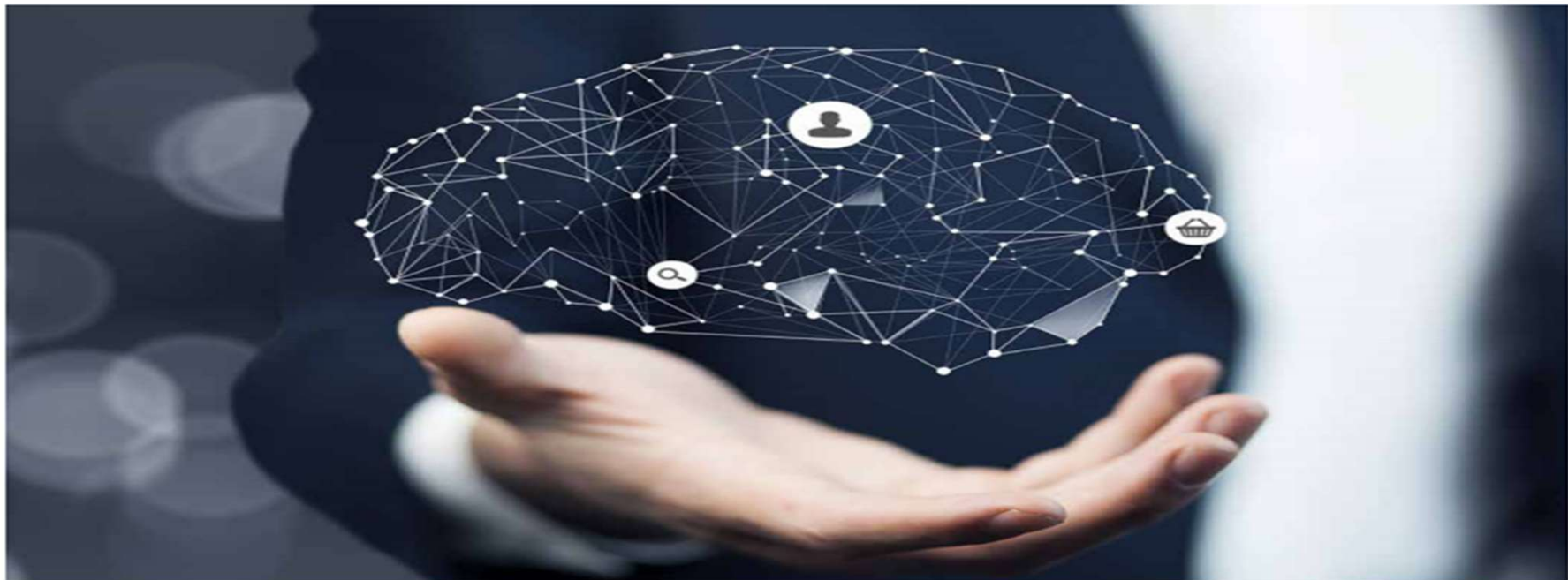
C u s t o m e r 1	D a y 1	P r o d u c t 1
C u s t o m e r 8	D a y 1	P r o d u c t 1
C u s t o m e r 1	D a y 4	P r o d u c t 2

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 (Vitamin-tabletten), 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 } $\rightarrow$ { 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:

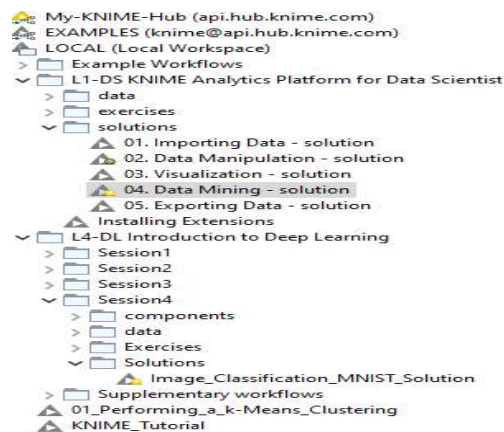
<https://youtu.be/vSWLafBdHus>

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

Exercise 5 to Lesson 9: Performing KNIME DM-Basics

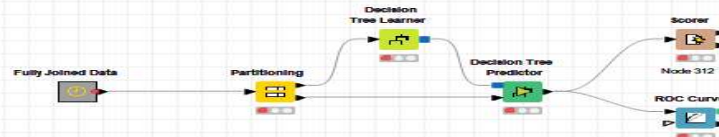
Exercise E9.5 (SW*, PD**): Run a KNIME-Basics Data Mining solution

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



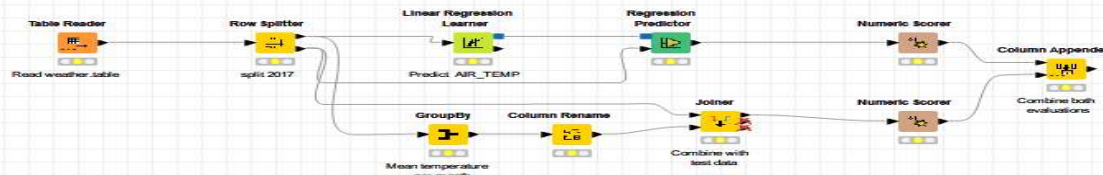
Activity I: Decision Trees

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



Activity II: Linear Regression

- Read *weather.tabledata*
- 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 Regression model



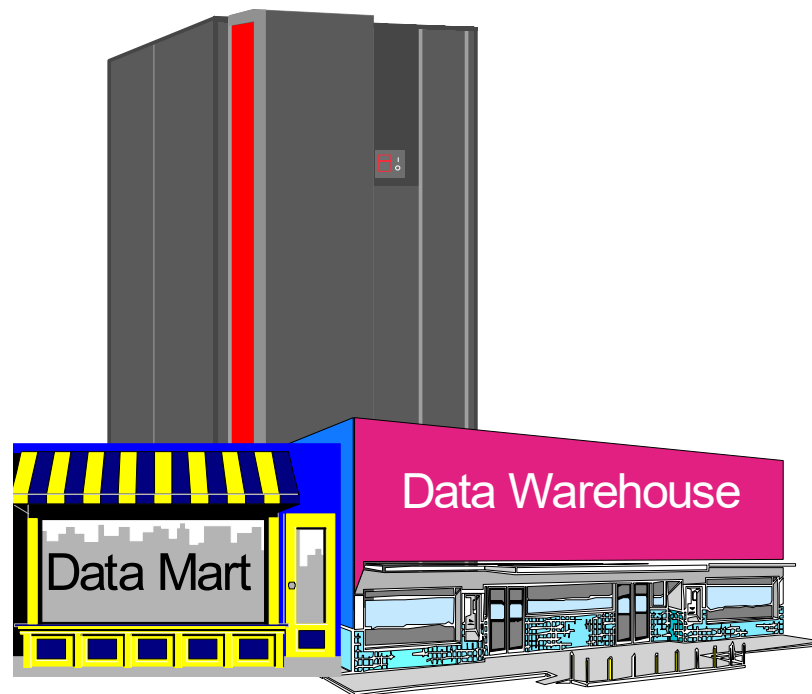
Activity III: k-Means

- Read *location_data.tabledata*
- 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

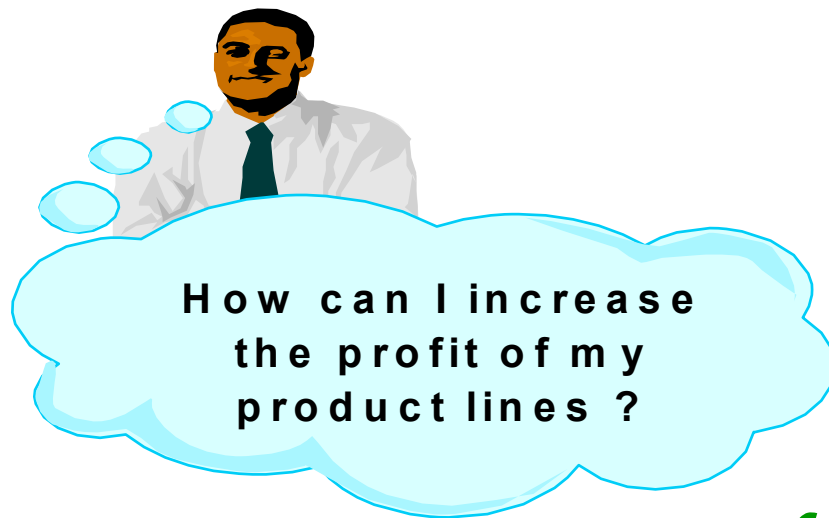
SW*: In the Seminar Work paper investigate this in more detail.

PD**: For the Program Design investigate these KNIME features in more detail

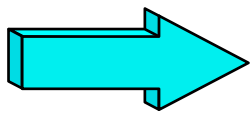
DW10 - Advanced Analytics II: Data Mining – Methods & Tools



Cross Selling – Business Idea



Associations
Segmentation



Increase Customer Loyalty

Cross Selling - 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 portfolios*

- Use IM Associations or Sequential Patterns

Cross Selling - Goals

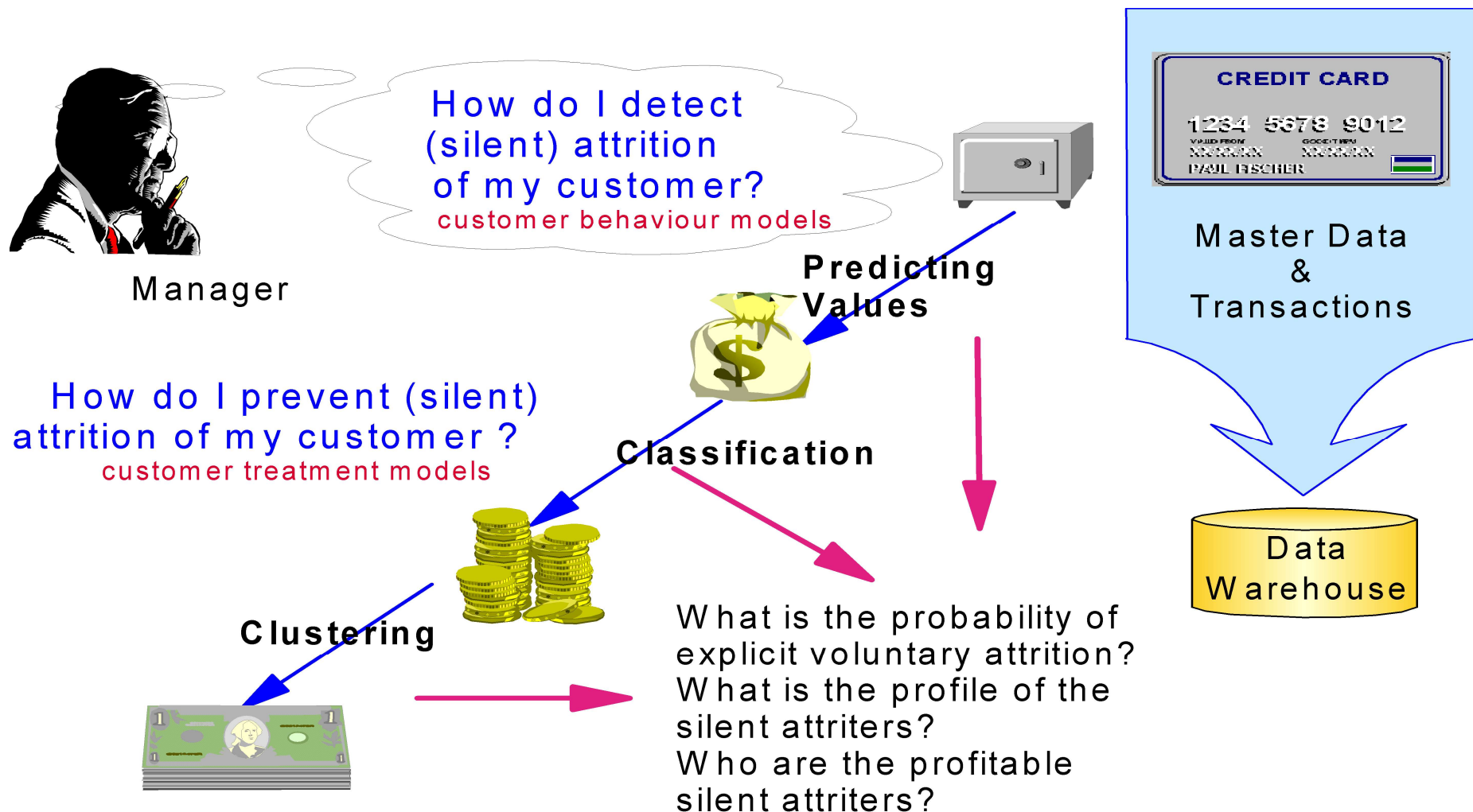


➤ *Goal:*

- 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



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

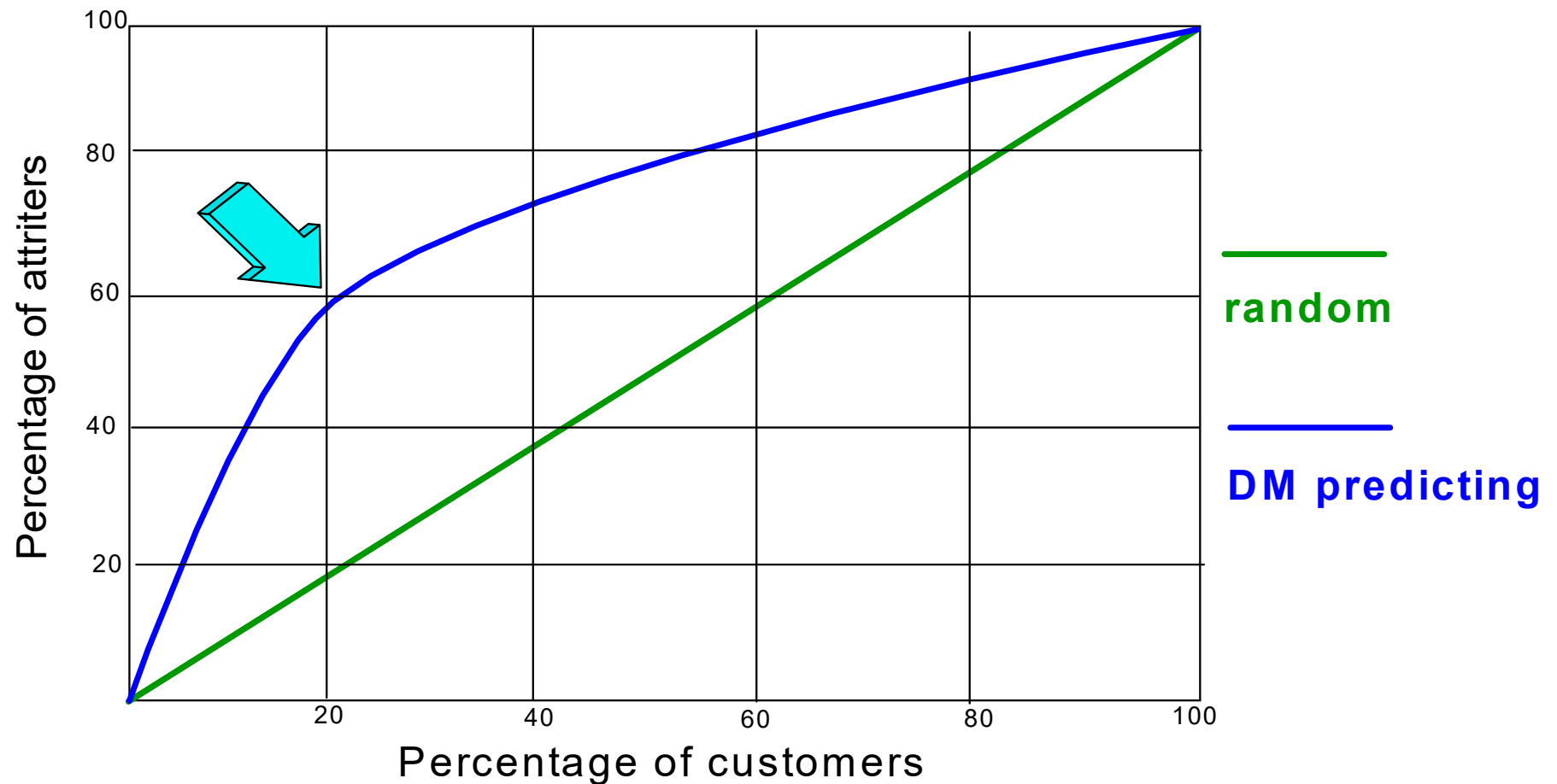
➤ ***Result Analysis***

➤ ***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

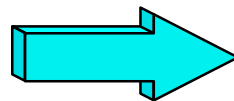
Campaign Management – Business Idea



Can I be more efficient
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

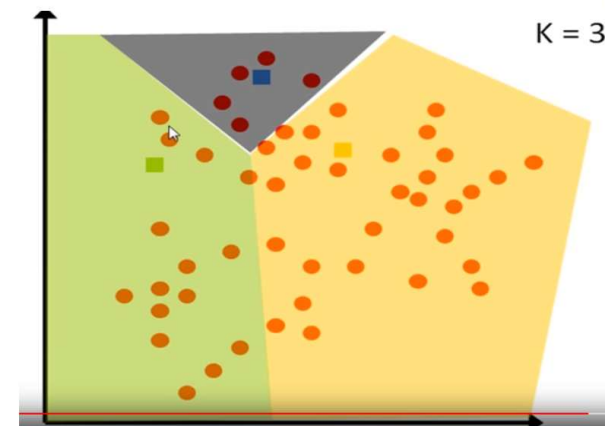
- *Choose the interesting segments*
- *Start the Campaign on a sample of people - adapt message to profile*
- *Analyse deeply the campaign results*
 - Build a model to explain why some replied and some did not
 - Use a scoring method
 - IM RBF Prediction
 - IM Neural Prediction
 - IM Tree/Neural Classification

Data Mining Method: K-Means-Clustering Algorithm

K-Means Learning Algorithm:

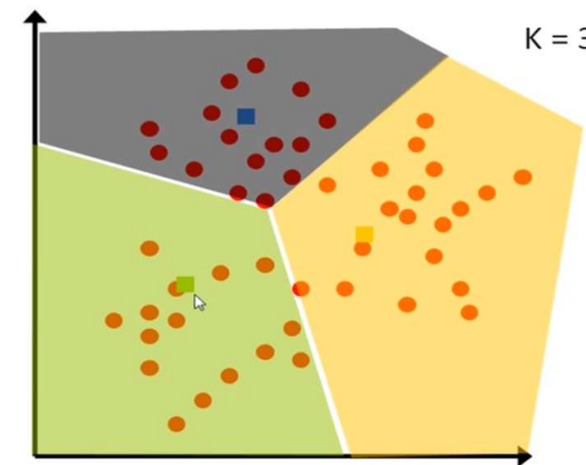
1. Define an initial (random) solution as vectors of means
 $\mathbf{m}(t=0) = [\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_K]^T$
2. Classify each input data according to $\mathbf{m}(t)$
3. Use the classification obtained in step 2 to recompute the vectors of means $\mathbf{m}(t+1)$
4. Update $t =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

K-Means – Initial Cluster Model



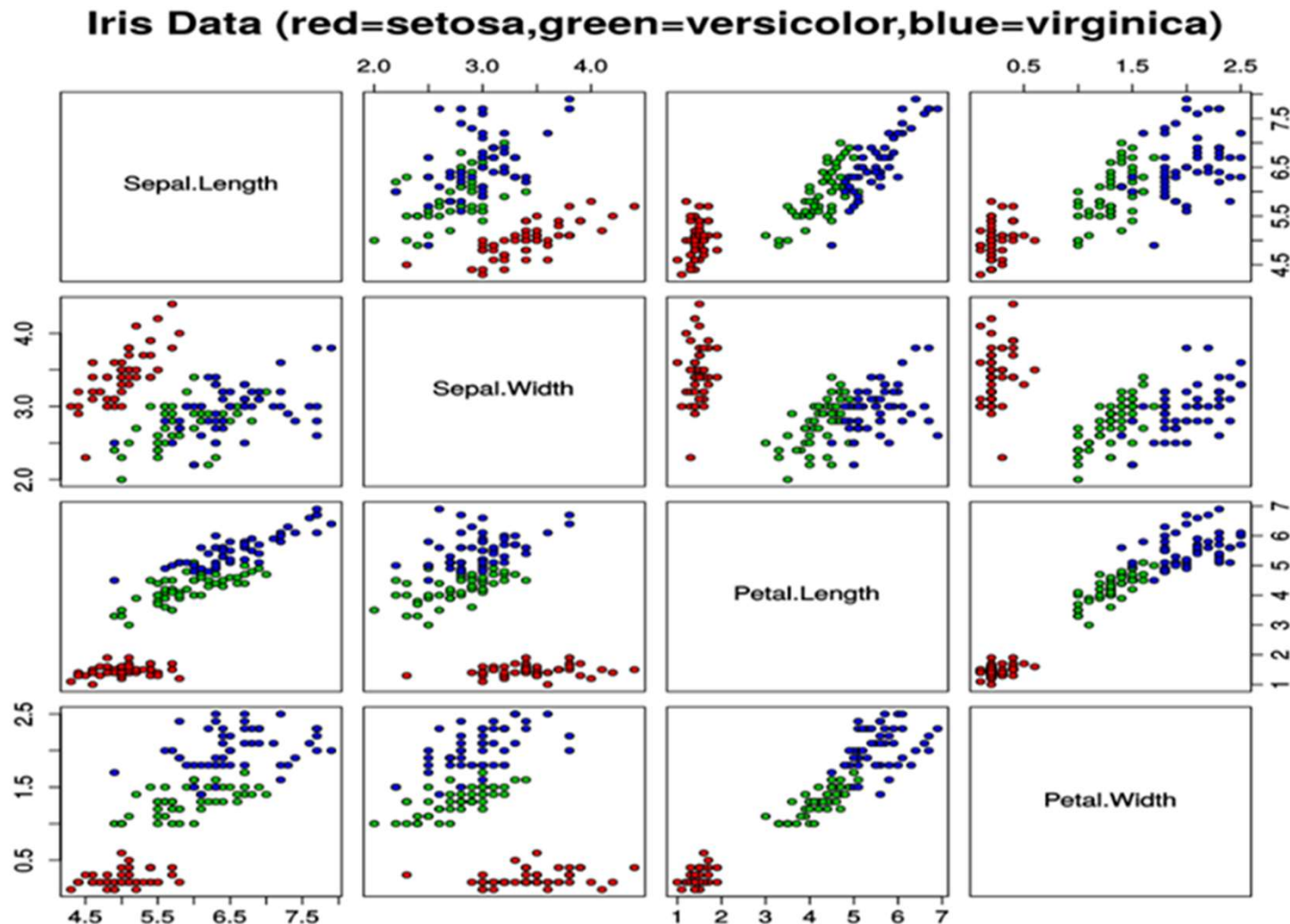
Repetition –
Data Mining

K-Means – Improve Initial Model



Repetition –
Data Mining

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



Iris Setosa



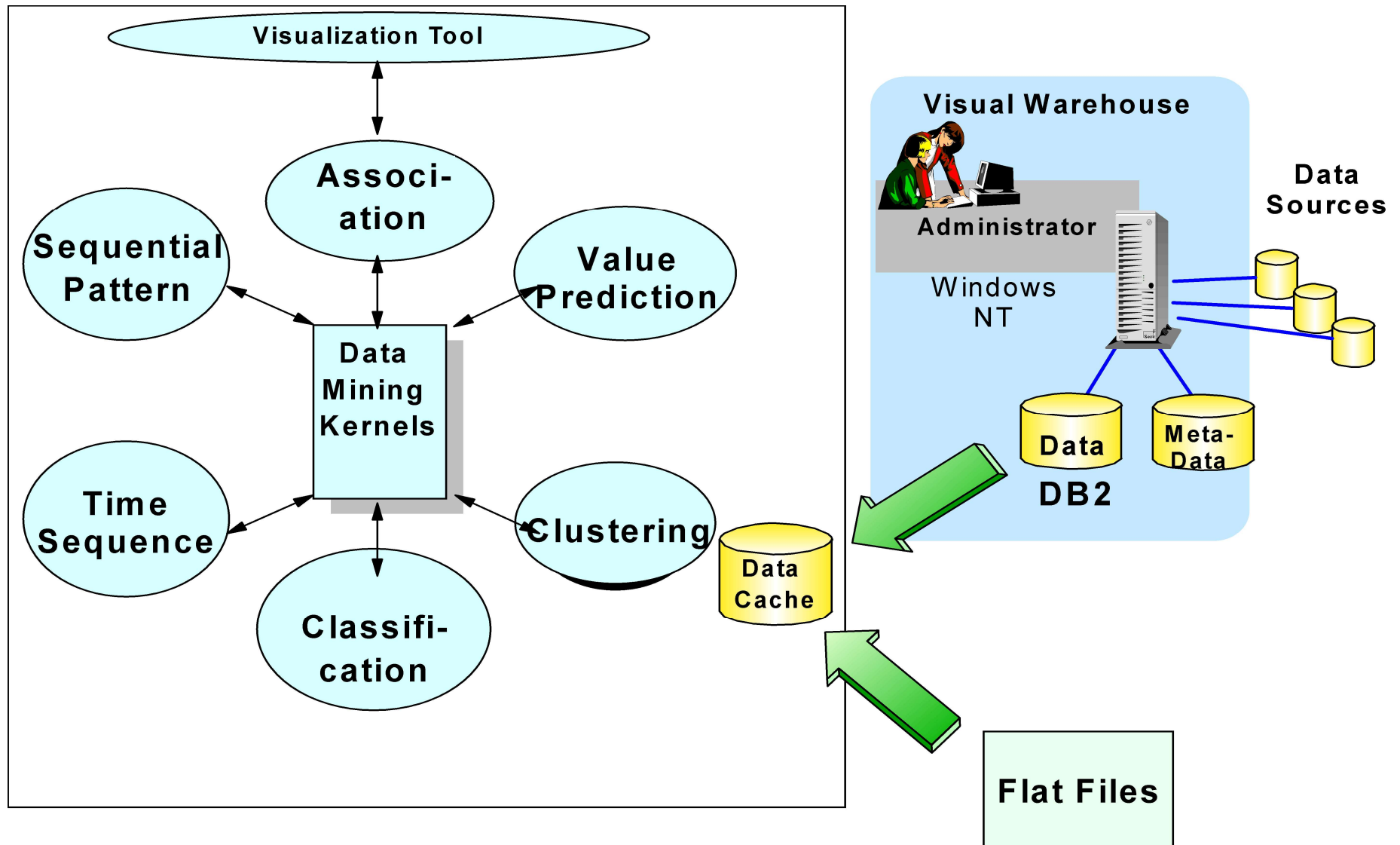
Iris Versicolor



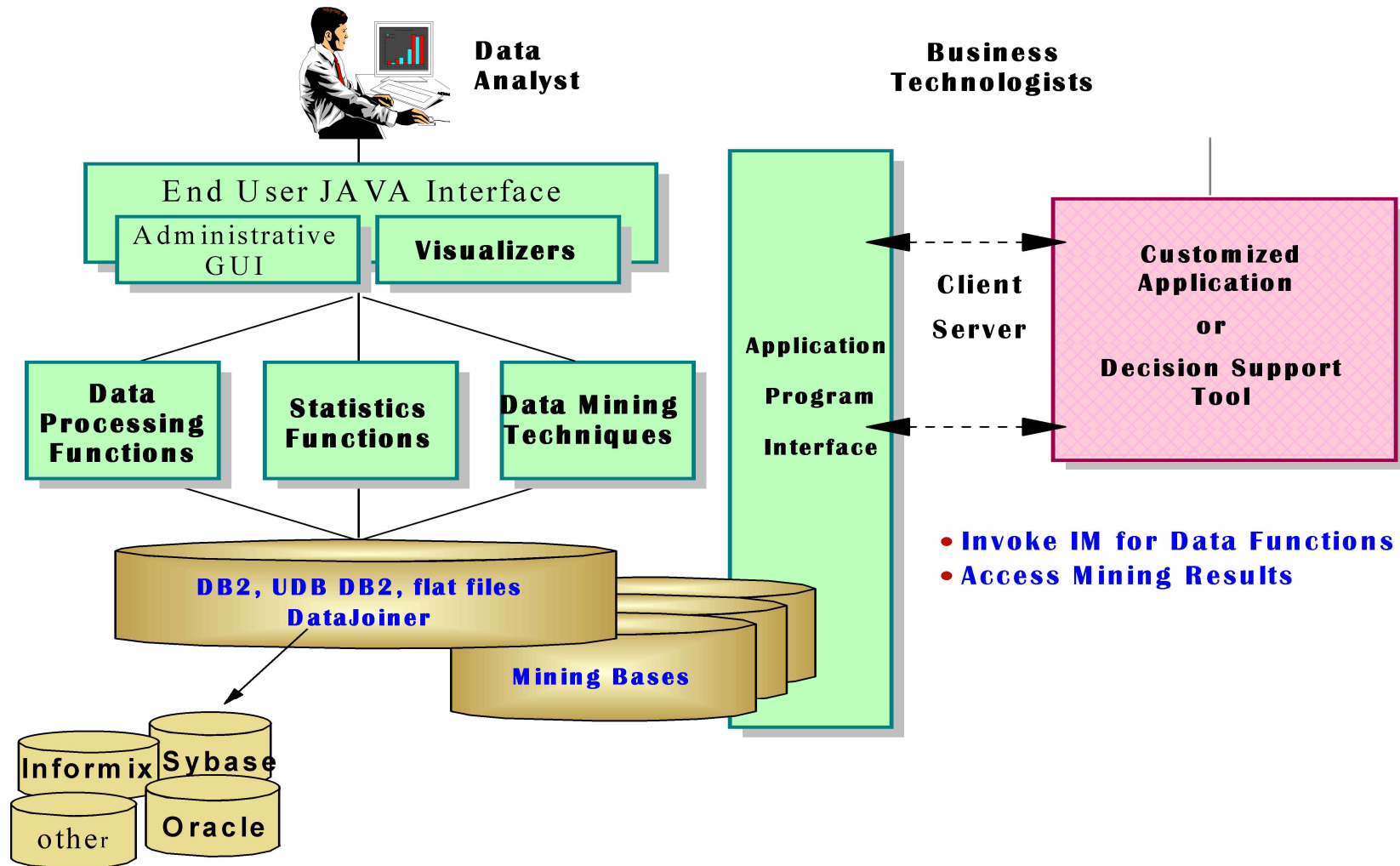
Iris Virginica

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

IM for Data - Overview



IM for Data – Tool Architecture

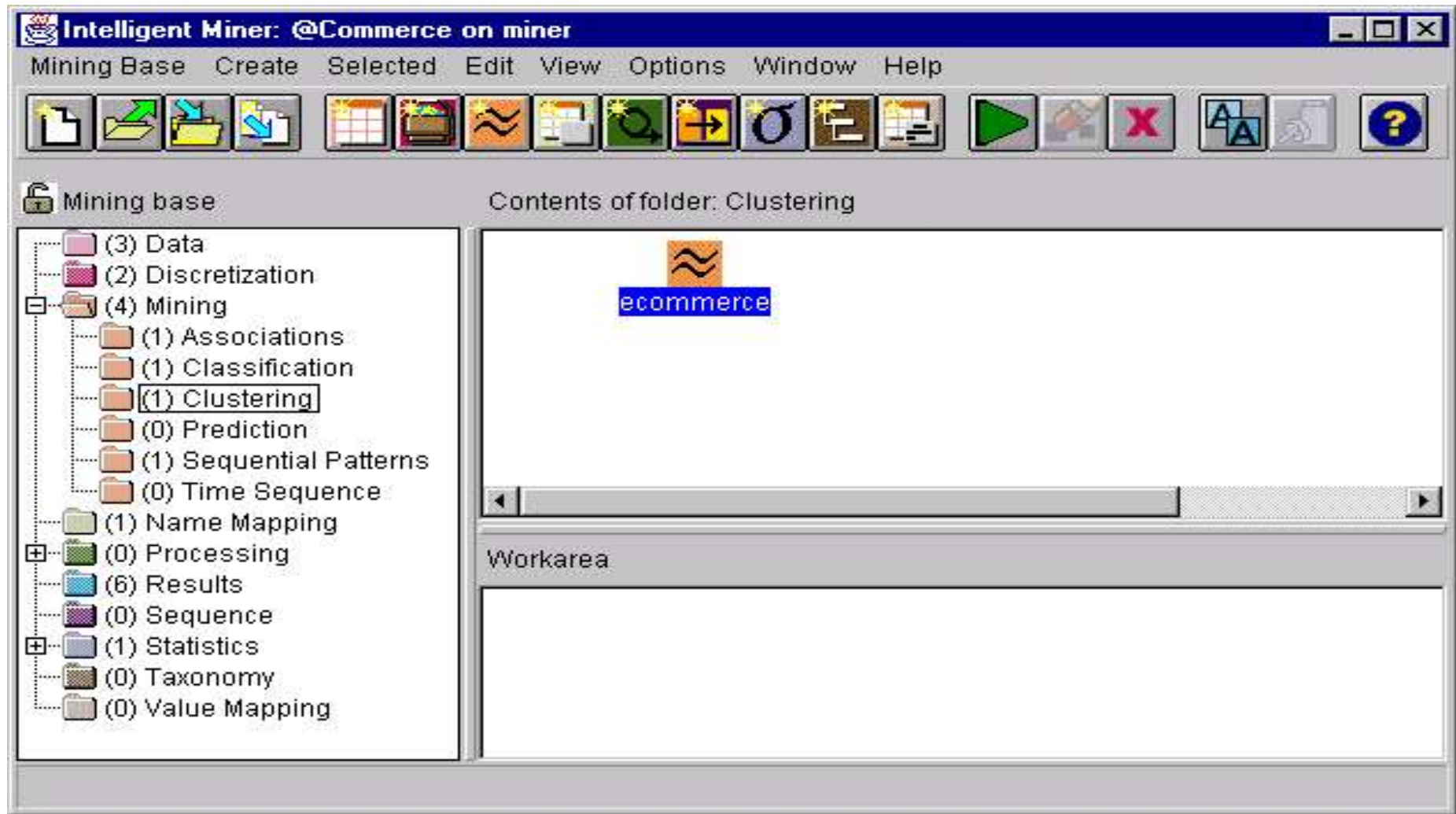


IBM IM for Data - Life Demo Overview

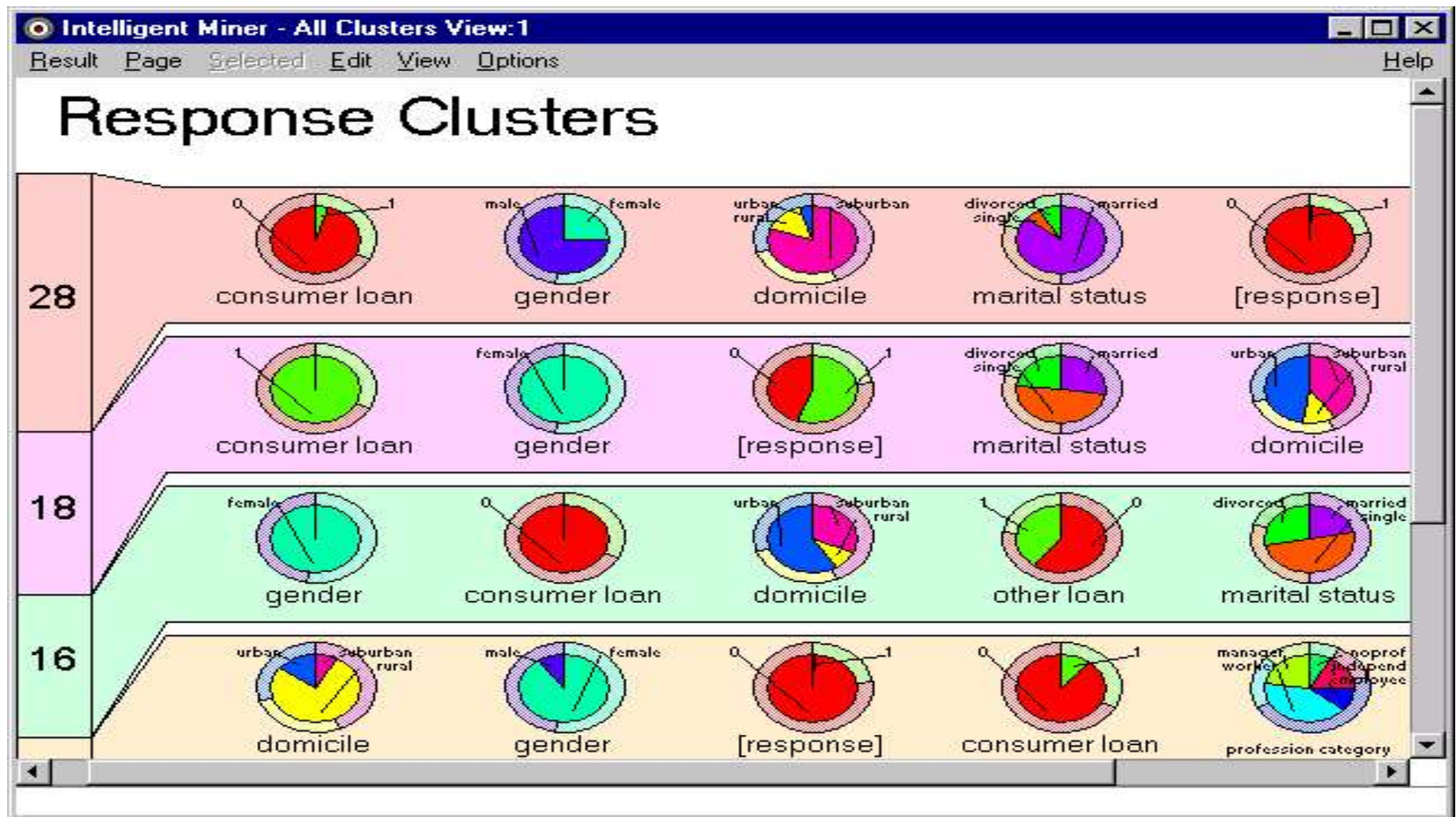
The demo will demonstrate the five phases of data mining tasks:

- 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



IBM Intelligent Miner for Data - Life Demo 2



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

Exercise E10.1 (SW*): Give detailed descriptions of two of the leading tools in the quadrant. See “Gartner_Quadrant-DS+ML_Tools-Apr2024.pdf” in [DHBW-Moodle]



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

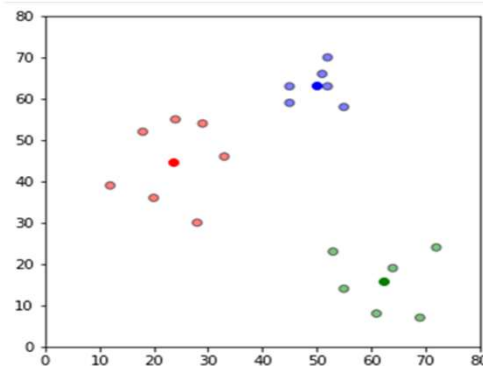
Exercise E10.3 (PD**): Create a K-Means Clustering in Python

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)

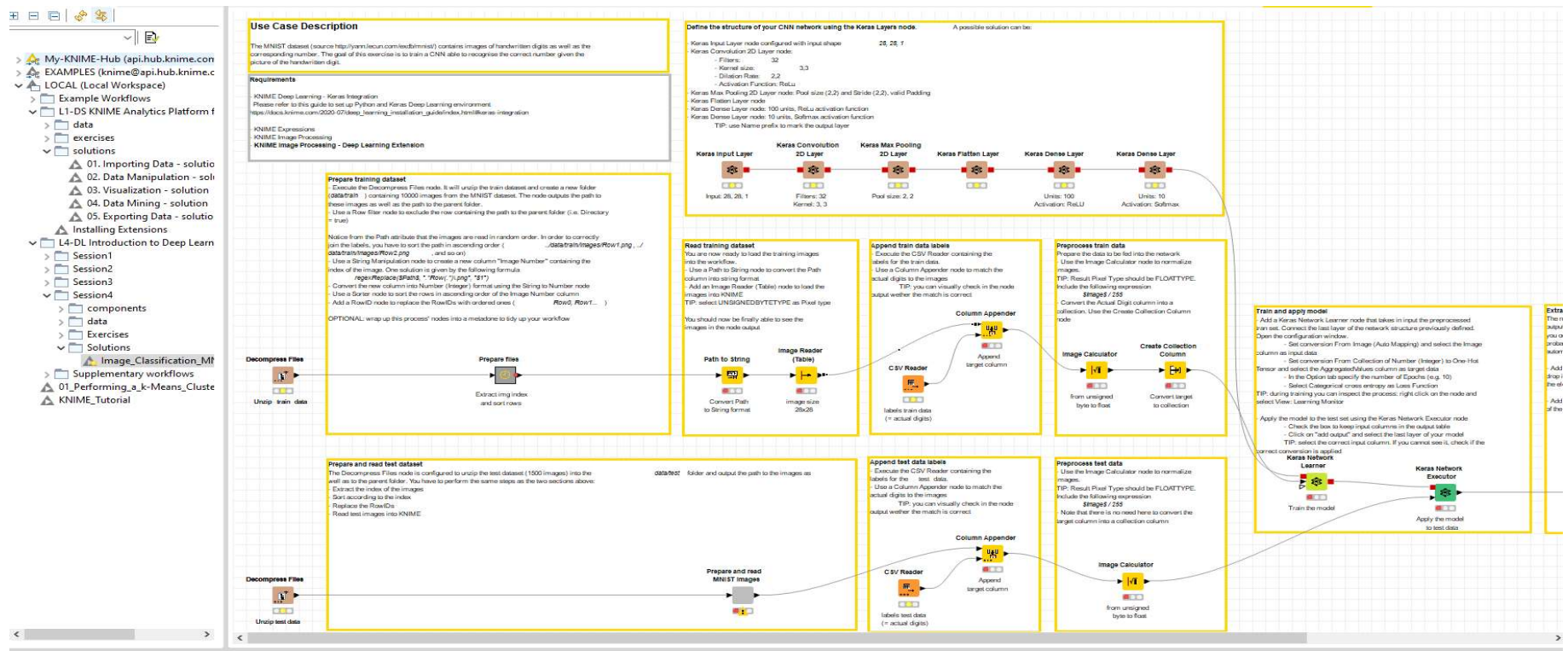


PD**: For the Program Design investigate these Python pgm in more detail

Exercise 4 to Lesson 10: KNIME Image-Classification

Exercise E10.4 (SW*, PD**): 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):



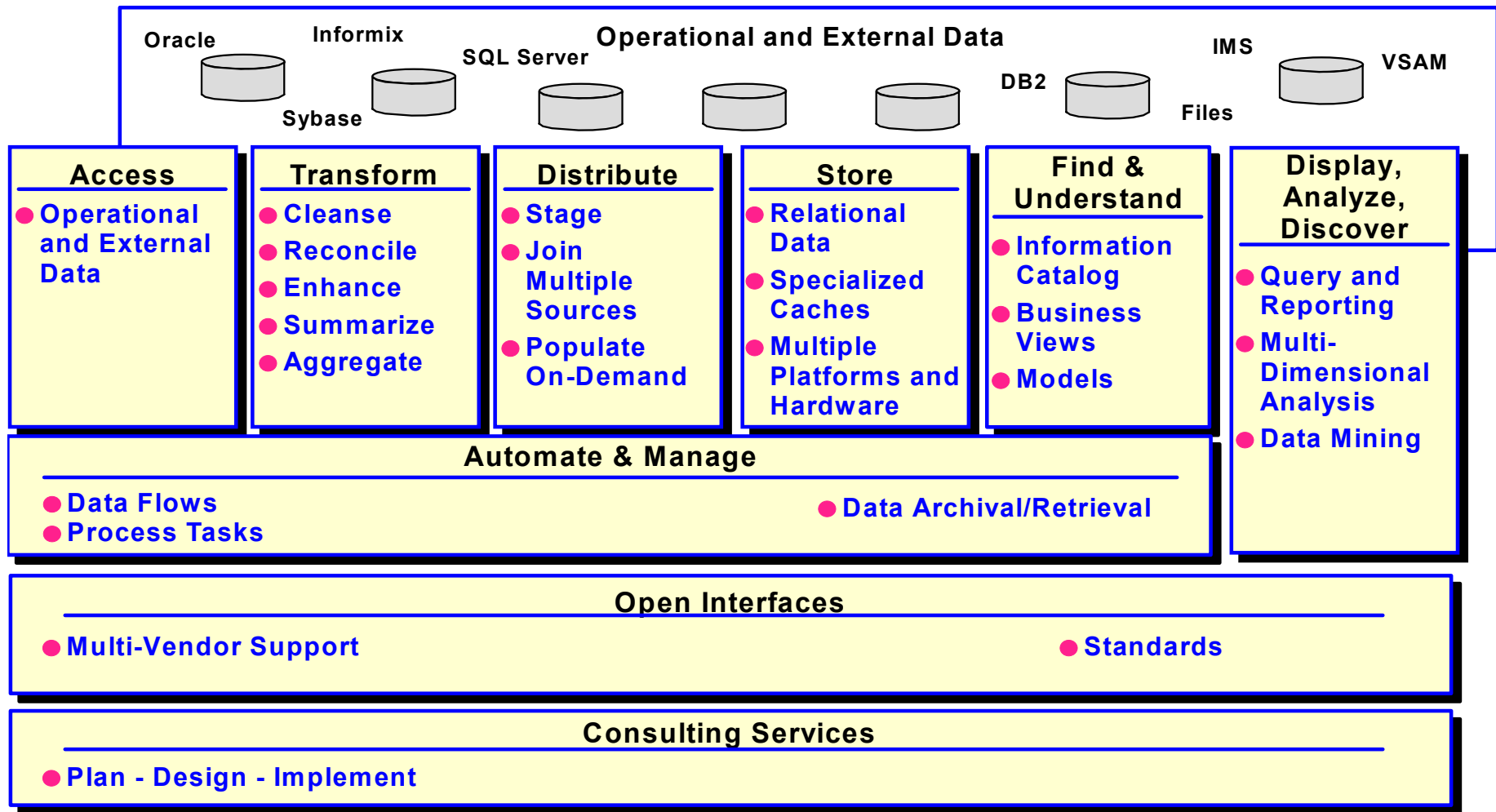
SW*: In the Seminar Work paper investigate this in more detail.

PD**: For the Program Design investigate these KNIME features in more detail

Anhang

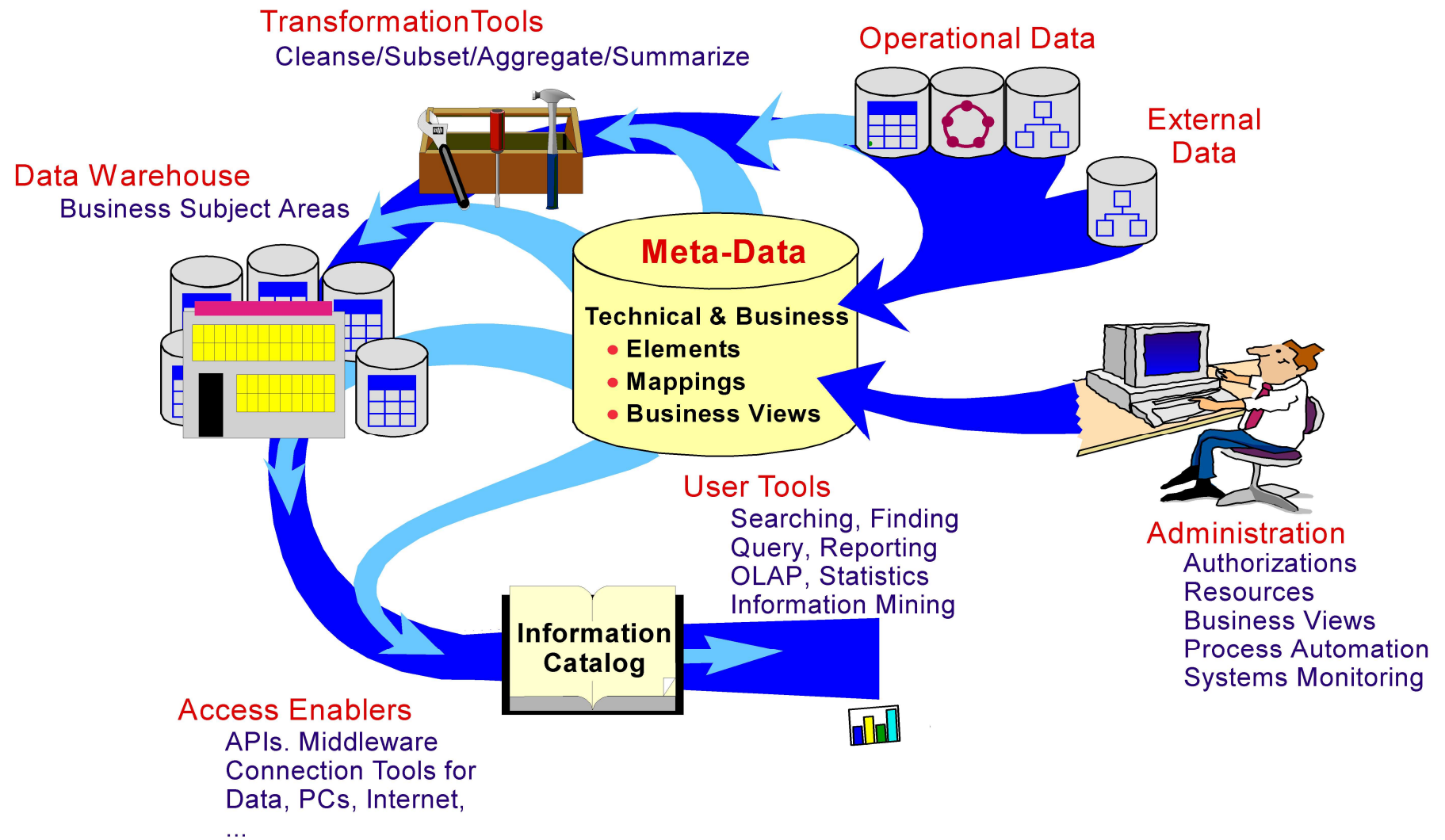
BACKUP Slides

Components of a Data Warehouse

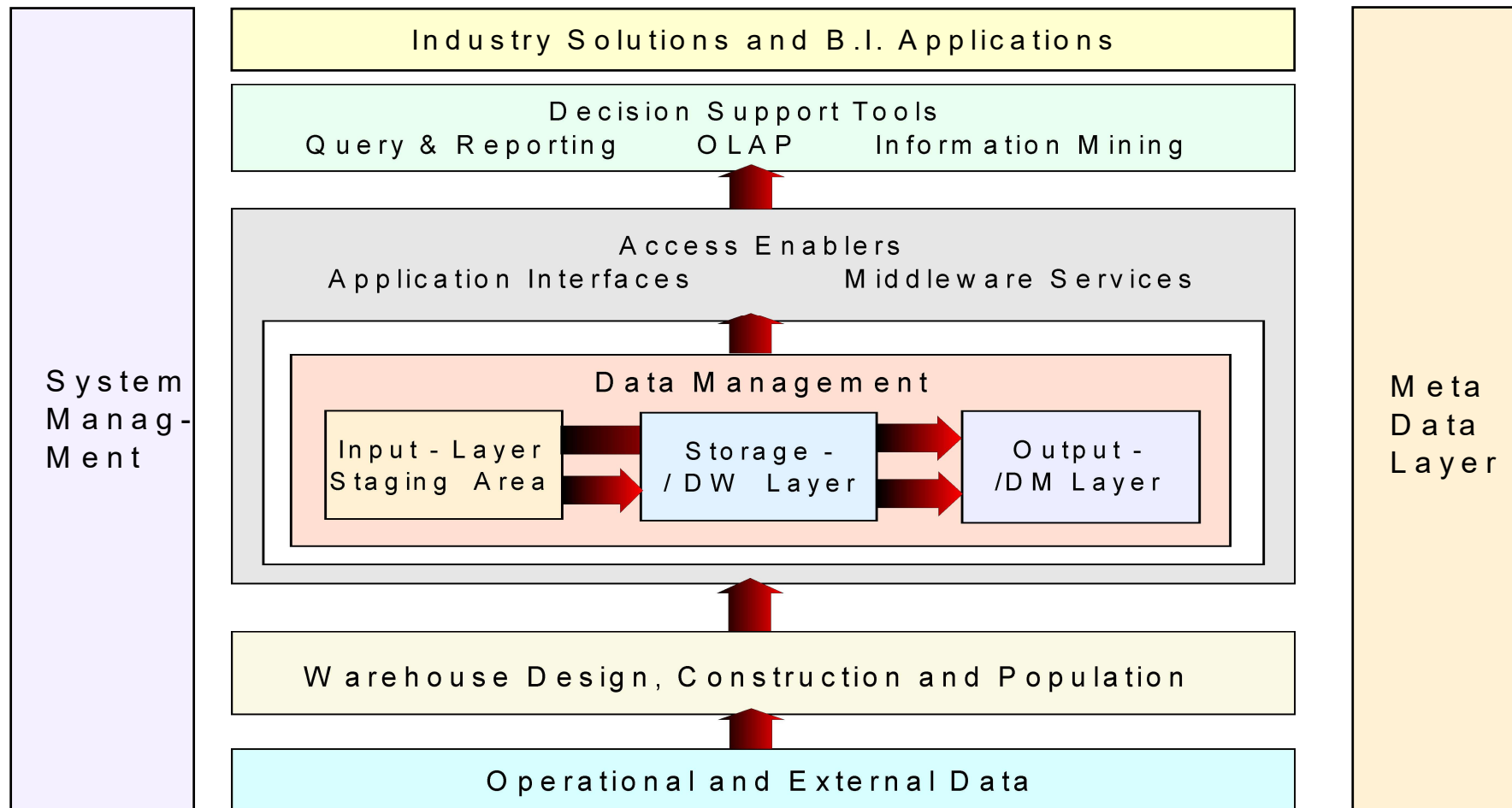


Enabling the Solution

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*