

Data Engineer

Grundlagen Business Intelligence

Allgemein



- **Zielgruppe:** Der Lehrgang richtet sich an Personen mit abgeschlossenem Studium in der Informatik, Wirtschaftsinformatik, Mathematik oder vergleichbarer Qualifikation.
- **Lehrgangsziel:** Sie beherrschen Prozesse rund um die Zusammenführung, Aufbereitung, Anreicherung und Weitergabe von Daten.
- **Teilnahmevoraussetzungen:** Programmierkenntnisse (idealerweise Python) und Erfahrungen mit Datenbanken (SQL) werden vorausgesetzt.
- **Dauer:** 4 Wochen
- **Abschlussprüfung:** Praxisbezogene Projektarbeit mit Abschlusspräsentation
- **Zertifikat:** alfatraining-Zertifikat

Übersicht



Data Engineer



Woche 1	Uhrzeit	Montag	Dienstag	Mittwoch	Donnerstag	Freitag
	8:30 10:00*	Begrüßung, Vorstellungsrunde, Einführung in die Unterrichtsform	Grundlagen Business Intelligence, OLAP, OLTP, Aufgaben eines Data Engineers	Anforderungsmanagement Aufgaben, Ziele und Vorgehensweise in der Anforderungsanalyse	Datenmodellierung, Einführung / Modellierung mit ERM	Einführung/Modellierung in der UML • Klassendiagramme, • Use-Case Analyse• Aktivitätsdiagramme
	10:00 10:15	Pause				
	10:15 11:45*	Einführung in alfaview & digitaler Lernumgebung Vorstellungsrunde	Data Warehousing (DWH): Umgang und Verarbeitung von strukturierten, semi-strukturierten und unstrukturierten Daten	Aufgaben, Ziele und Vorgehensweise in der Anforderungsanalyse	Datenmodellierung, Einführung / Modellierung mit ERM	Datenbanken Grundlagen von Datenbanksystemen
	11:45 11:50	Pause				
	11:50 12:35*	Grundlagen Business Intelligence Anwendungsfelder, Dimensionen einer BI Architektur	Data Warehousing (DWH): Umgang und Verarbeitung von strukturierten, semi-strukturierten und unstrukturierten Daten	Datenmodellierung, Einführung / Modellierung mit ERM	Einführung/Modellierung in der UML • Klassendiagramme, • Use-Case Analyse• Aktivitätsdiagramme	Grundlagen von Datenbanksystemen, Architektur von Datenbankmanagementsystemen
	12:35 13:15	Pause				
	13:15 14:45*	Praktische Umsetzung anhand von Aufgaben/Übungen	Praktische Umsetzung anhand von Aufgaben/Übungen	Praktische Umsetzung anhand von Aufgaben/Übungen	Praktische Umsetzung anhand von Aufgaben/Übungen	Praktische Umsetzung anhand von Aufgaben/Übungen
	14:45 14:50	Pause				
	14:50 15:35*	Praktische Umsetzung anhand von Aufgaben/Übungen	Praktische Umsetzung anhand von Aufgaben/Übungen	Praktische Umsetzung anhand von Aufgaben/Übungen	Praktische Umsetzung anhand von Aufgaben/Übungen	Ausführliche Stellenrecherche, Aktualisierung Jobbörsenprofil

*In Wochen mit Feiertagen Unterricht bis 17:10 Uhr. Kursinhalte des Feiertages verschieben sich entsprechend.

Diese Unterrichtsdokumentation dient der inhaltlichen Orientierung des Kursablaufs. Abweichungen aufgrund von Softwareaktualisierungen oder Arbeitsmarktanforderungen sind möglich.



Übersicht



Grundlagen Business Intelligence (ca. 2 Tage)

Anwendungsfelder, Dimensionen einer BI Architektur
Grundlagen Business Intelligence, OLAP, OLTP, Aufgaben der Data Engineers
Data Warehousing (DWH): Umgang und Verarbeitung von strukturierten, semi-strukturierten und unstrukturierten Daten

Anforderungsmanagement (ca. 2 Tage)

Aufgaben, Ziele und Vorgehensweise in der Anforderungsanalyse
Datenmodellierung, Einführung/Modellierung mit ERM
Einführung/Modellierung in der UML

- Klassendiagramme
- Use-Case Analyse
- Aktivitätsdiagramme

Künstliche Intelligenz (KI) im Arbeitsprozess

Vorstellung von konkreten KI-Technologien im beruflichen Umfeld
Anwendungsmöglichkeiten und Praxis-Übungen

Datenbanken (ca. 3 Tage)

Grundlagen von Datenbanksystemen
Architektur von Datenbankmanagementsystemen
Anwendung RDBMS, Umsetzung Datenmodell in RDBMS, Normalformen
Praktische und theoretische Einführung in SQL
Grenzen von Relationalen Datenbanken, csv, json

Data Warehouse (ca. 4 Tage)

Star Schema
Datenmodellierung
Erstellung Star Schema in RDBMS

Snowflake Schema, Grundlagen, Datenmodellierung

Erstellung Snowflake Schema in RDBMS
Galaxy Schema: Grundlagen, Datenmodellierung
Slowly Changing Dimension Tables Typ 1 bis 5 – Restating, Stacking, Reorganizing, mini Dimension und Typ 5
Einführung in normal, causal, mini und monster, heterogeneous und sub Dimensions
Vergleich von state und transaction oriented Faktentabellen, Density und Storage vom DWH

ETL (ca. 4 Tage)

Data Cleansing

- Null Values
- Aufbereitung von Daten
- Harmonisierung von Daten
- Anwendung von Regular Expressions

Data Understanding

- Datenvalidierung
- Statistische Datenanalyse

Datenschutz, Datensicherheit
Praktischer Aufbau von ETL-Strecken
Data Vault 2.0, Grundlagen, Hubs, Links, Satellites, Hash Key, Hash Diff.
Data Vault Datenmodellierung
Praktischer Aufbau eines Data Vault Modells – Raw Vault, Praktische Umsetzung von Hash-Verfahren

Projektarbeit (ca. 5 Tage)

Zur Vertiefung der gelernten Inhalte
Präsentation der Projektergebnisse

Definitionen



„business intelligence“ = *geschäftliche Erkenntnisse*

Systeme und Prozesse zur systematischen Analyse (intelligence) des eigenen Unternehmens und seines kommerziellen Umfelds

Mit den gewonnenen Erkenntnissen können Unternehmen ihre Geschäftsabläufe, Kunden- und Lieferantenbeziehungen profitabler machen, Kosten senken, Risiken minimieren und die Wertschöpfung vergrößern.

Ziel ist die Gewinnung von Erkenntnissen, die in Hinsicht auf die Unternehmensziele bessere operative, taktische oder strategische Entscheidungen ermöglichen

Aus Daten REAL,PARAMETER :: b = 31.4159265358979323

werden Kennzahlen:

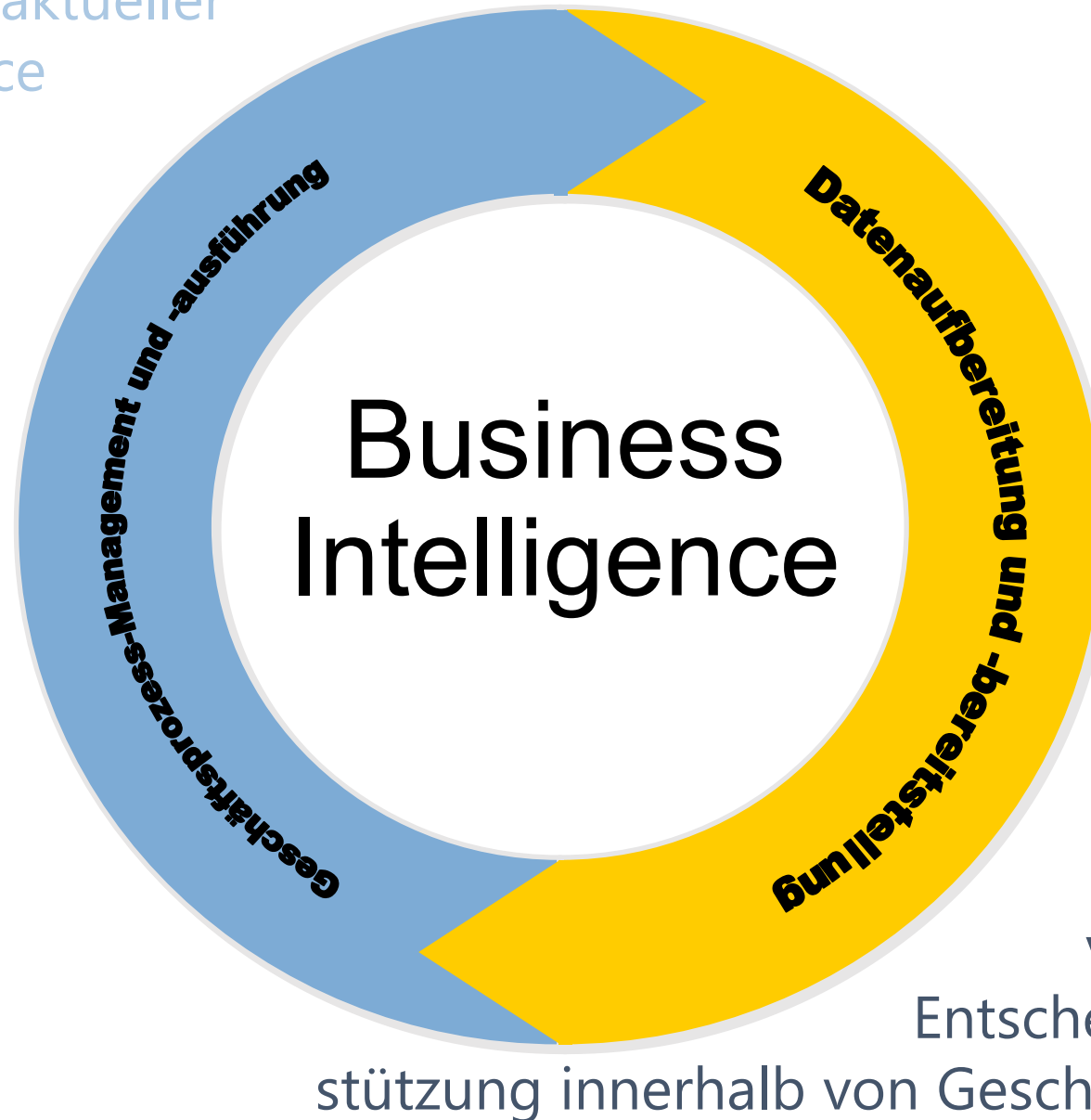
Wie können wir unser strategisches Ziel, 20% mit Produkten, die jünger sind als 5 Jahre, mit unserem Innovationsprozess besser unterstützen?

Schau' Dir die Durchlaufzeit unserer Patentanmeldungen an: Im Durchschnitt sind das 31.4 Tage – hier müssen wir den Prozessablauf beschleunigen !



Ziel:

Transparenz über ganze Geschäftsprozesse
und deren aktueller
Performance



Ziel:

Aufbereitung
von Daten zur
Entscheidungsunter-
stützung innerhalb von Geschäftsprozessen

Why BI

- Today's businesses have access to **more data** than ever before. Companies produce, collect and store vast amounts of data, from customer feedback surveys to manufacturing and delivery statistics. **Business intelligence** is a series of methodologies that puts this data to work, helping businesses become more effective and increase profits. By using these methodologies and specific software analytics, savvy business executives can harness the power of raw data and leverage it to support strategic planning that can help an organization move ahead of the competition.

- <https://www.datapine.com/de/artikel/business-intelligence-bi-system>
- <https://www.softguide.de/software-tipps/business-intelligence-definition>
- https://www.celonis.com/ebook/process-mining-for-dummies/?utm_source=google&utm_medium=cpc&utm_campaign=evergreen&utm_term=process%20mining%20definition&utm_content=en_maxconvvalue_process_mining_rsa3&creative=654280536224&keyword=process%20mining%20definition&matchtype=e&network=g&device=c&_bt=654280536224&_bk=process%20mining%20definition&_bm=e&_bn=g&_bg=140441357211&gad=1&gclid=Cj0KCQjwzdOIBhCNARIsAPMwjbwQ1xmmEPkKAiGsauexCY0IPJg3SKN9G-Laiz0Mim-ELvc35vzhRIAaAixKEALw_wcB

Definition: Business Intelligence

- A broad category of applications and technologies for gathering, providing access to, and analyzing data for the purpose of helping enterprise users make better decisions and reports. The term implies you have a complete understanding of your business. We must have a strong knowledge about all factors of your company including customers, competition, business partners, internal operations, and the economic environment to make effective and good quality business decisions. Business Intelligence allows you to make these kinds of decisions.
- The term BI was used as early as 1996 When Gartner Group said:
- By 2000, Information Democracy will emerge in forward-thinking enterprises, with Business Intelligence information and applications available broadly to employees, consultants, customers, suppliers and the public

<https://www.gartner.com/doc/reprints?id=1-1XYUYQ3I&ct=191219&st=sb>.

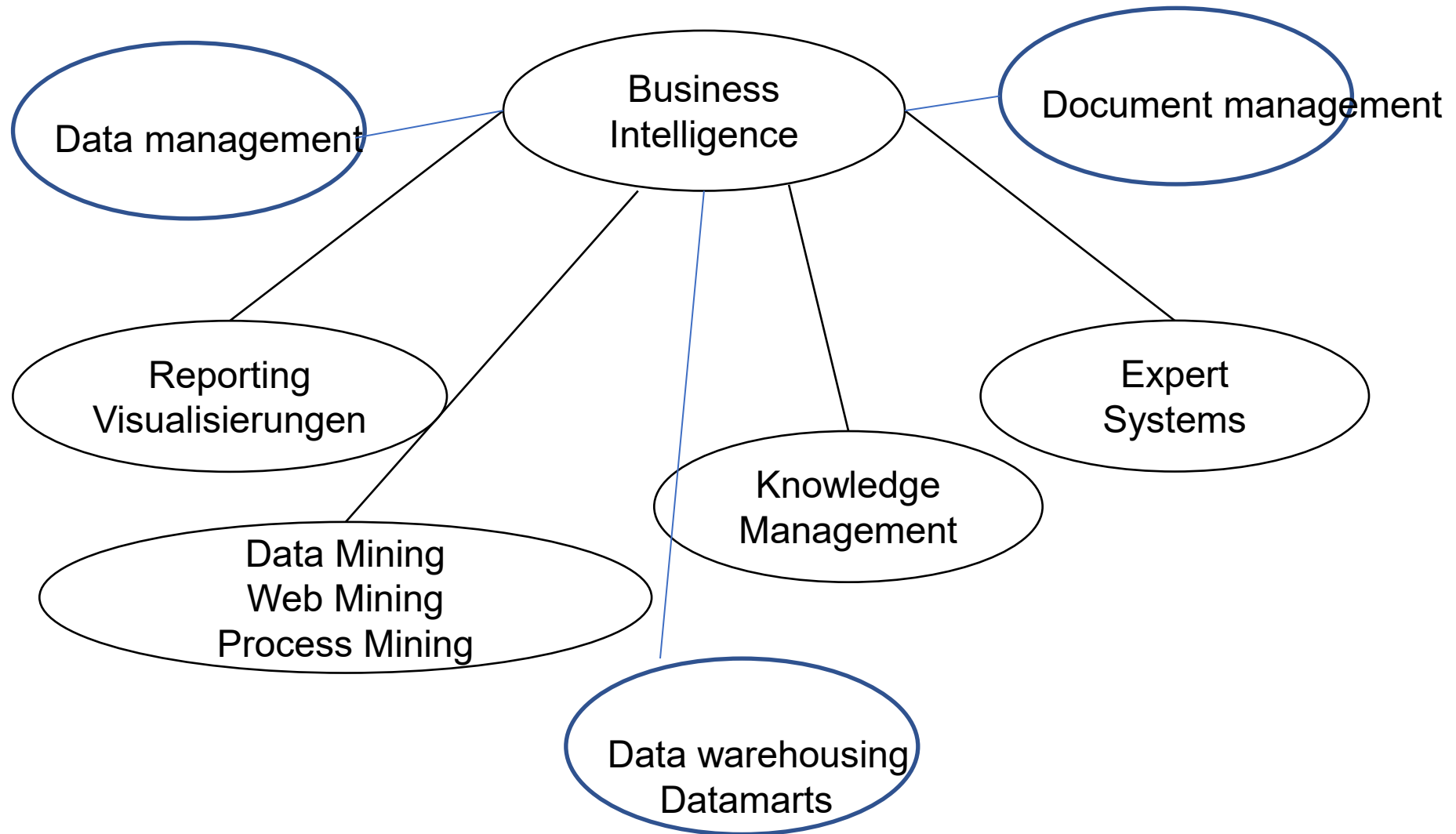
Business Intelligence (BI) encompasses the processes, tools, and technologies required to transform enterprise data into information, and information into knowledge that can be used to enhance decision-making and to create actionable plans that drive effective business activity.

- BI can be used to acquire
 - **Tactical insight** to optimize business processes by identifying trends, anomalies, and behaviors that require management action.
 - **Strategic insight** to align multiple business processes with key business objectives through integrated performance management and analysis.

Why Business Intelligence?

- Better decisions with greater speed and confidence
- Recognize and maximize firm's strengths
- Shorten marketing efforts
- Improve customer relationships
- Align effort with firm strategy
- Improve revenue and profit
- <https://www.gartner.com/doc/reprints?id=1-1XYUYQ3I&ct=191219&st=sb>
- <https://www.gartner.com/doc/reprints?id=1-292LEME3&ct=220209&st=sb>

Business Intelligence



Elements of Business Intelligence

- **Data Gathering**
 - Information capture
- **Analysis**
 - Understanding the context of information
- **Distribution**
 - Timely delivery to the right people who can act on it (Real-Time, 14/7/365, any device)

Begrifflichkeit Big Data / Data Science / BI

- Interpretation
 - *Der Ansatz mit unterschiedlichsten Werkzeugen und auf Basis aller verfügbaren Daten bestehende Geschäftsprozesse zu messen, weiterzuentwickeln und Auffälligkeiten zu erklären.*

- https://de.wikipedia.org/wiki/Big_Data

- Wachsende Anforderungen an Analysen im Bezug auf Geschwindigkeit der Analyse und Verarbeitung von sehr großen Datenmengen.
- Große Anzahl an neuen Produkten und Vermarktungswegen.
- Kundenindividualisierung von Ansprachen und Produkten.
- Übersetzung von Daten in Handlungsempfehlungen.

Data Science



- Aufgaben

- Ad-hoc Analysen

- Fragestellungen aus dem täglichen Geschäft
 - Auffälligkeiten

- **Scorewert-Berechnungen**

- Vorhersage des Kundenverhaltens
 - Optimierung der Vermarktung

- **Cockpits**

- Self-service-Angebot für verschiedene Fachbereiche

- **Analytische Projekte**

- Mehrmonatige Projekte zu einzelnen Fragestellungen / Optimierungen

- **KPI**

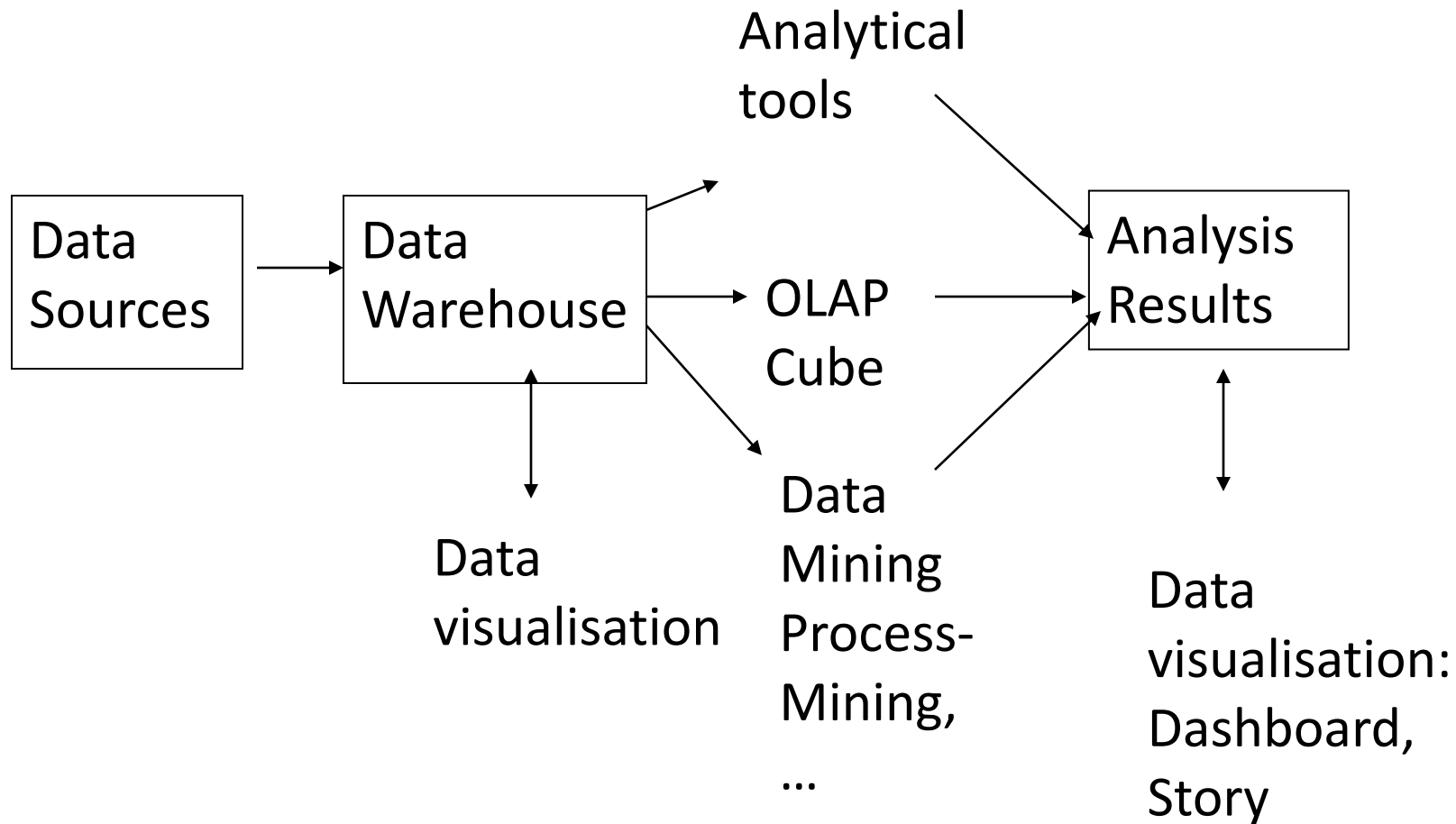
- Beispiele für Anwendungsfälle

- Berechnung von Produktaffinitäten der Kunden.
- Abschätzung von Unternehmensrisiken: Abwanderung und Betrug.
- Prozessoptimierungen: Kategorisierung von Kundenschreibern.
- Analyse und Vorhersage von Produkt- und Branchentrends.
- Optimierungen in der Werbung.
- Analyse von Kundenzufriedenheit.

Data is

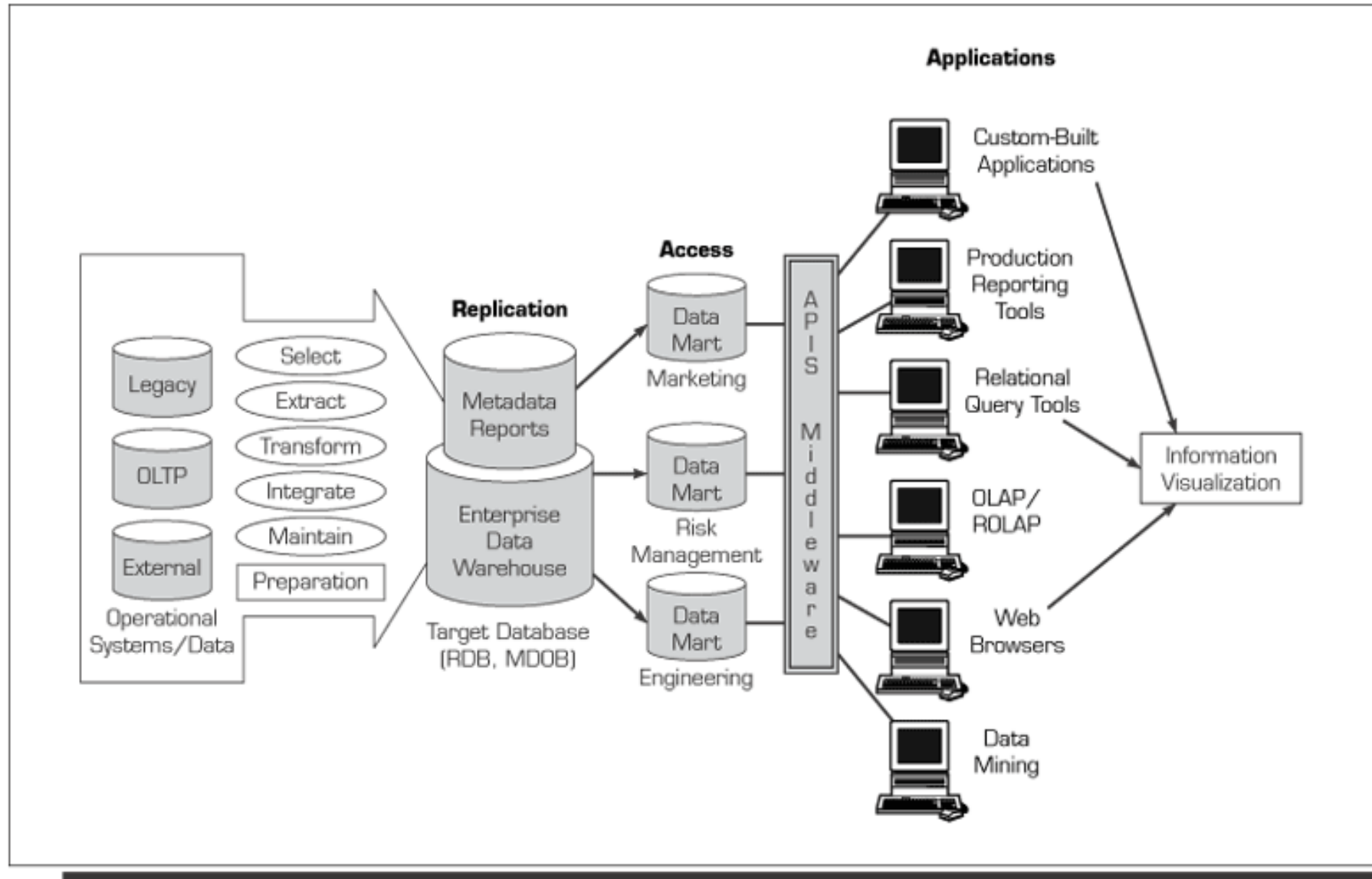
1. Gathered from relevant sources
2. Filtered, and stored
3. Analysed and arranged into meaningful patterns using different tools .
4. Business intelligence is the knowledge gained from that data analysis.

Overview of Business Intelligence



Data → Information → Wissen - Knowledge

Figure 5.2 Data Warehouse Framework and Views



From Turban, Aronson and Liang

Some Questions

- Where does the data come from?
- How can we decide what data is important?
- How can data from different sources be joined together (consolidated and integrated) securely?
- How can data be analysed?
- How can these analyses be viewed?

Where does the data come from?

- Data can be collected manually or automatically.
 - Transaction data e.g. supermarket checkout, bank withdrawal
 - Time studies, questionnaire, observation notes
 - Physical sensors e.g. temperature of a rooms in a house
 - Sensors, scanners, bar codes

How can we decide what data is important?



- Depends what our goals are, the **functional area**(e.g. Sales, HR, marketing..) and what **processes** we are looking at..

Balanced scorecard

Critical success factors

Key performance indicators

Human resources

- employee
- organizational
- departmental

Sales and marketing measures

- products
- customers
- demographics
- promotions
- sales force
- order type

Functional Areas

Finance

- currency standards
- account information
- industry trends

Operations management

- assembly speed
- warehouse stock
- manufacturer and supplier cost
- shift productivity

Wissensarmut im Informationsreichtum



„Eine Informationsschwemme.
Eine Überflutung gar.
Wir erhalten so viele Nachrichten,
ungefragt und aus so vielen Quellen,
in unterschiedlichen Formen und
Konzentrationen, daß Information
zu einer Art Müll wurde.

Quelle: Neil Postman, in: Future 1/99

Wissensarmut im Informationsreichtum



- „Viele Manager kapitulieren vor der kaum mehr zu bewältigenden Daten- und Informationsflut und nehmen das für die Unternehmenssteuerung relevante Geschehen nur noch äußerst selektiv wahr.“
 - Uwe Hannig (Leiter des Ludwigshafener Instituts für Managementinformations-systeme IMIS e. V. sowie des Instituts für Knowledge Management (IKM e. V.) in Zwickau.

Daten - Information - Wissen

- Daten: formatierte, maschinell verarbeitbare Information
- Information: kommunikatives Handeln gemeinsamer Problem- und Wissenskontext (Sender/ Empfänger)
 - Informationsgehalt – semantisch, pragmatisch, explanatorisch, phänomenal
- Wissen: objektiv, subjektiv

Konzepte analytischer Informationssysteme



- Data Warehouse
- On-Line Analytical Processing
- Data Mining

Data Warehouse

- Aufbau eines unternehmensweiten, entscheidungsorientierten Datenpools
- Wirksame Unterstützung unterschiedlicher analytischer Aufgaben

On-Line Analytical Processing

- Werkzeuge zur Entscheidungsunterstützung mit mehrdimensionalem Weltbild
- Kosten- Umsatzgrößen sind nur in Bezug auf Kunden, Regionen usw. aussagefähig
- Bezugsgrößen werden als Dimension gespeichert

Data Mining

- Techniken und Verfahren zum Auffinden von bislang verborgenen Informationen
- Werkzeuge ermöglichen Mustererkennung nach vorheriger Konditionierung
- Erwartet werden Werkzeuge, die handlungsbezogene Empfehlungen geben

Process Mining



- Techniken und Verfahren Darstellung des Workflow im Betrieb

Architektur

- Klassische BI lässt sich in einem 3-Ebenen-Modell darstellen:
 - Datenspeicherung,
 - Informationsgenerierung
 - Zugriff.
-
- Die Abbildung stellt die generelle Architektur eines Business Intelligence Systems auf einem hohen Abstraktionslevel dar.
-
- [Whitepaper_BI_Funktionsweise.pdf](#)

Motivation

Informationen sind ein kritischer Erfolgsfaktor



- Informationsbeschaffung muß nicht das Durchforsten des Papierbestandes sein.
- Informationsbeschaffung läßt sich durch Data Warehouse automatisieren.

Motivation

Nutzung eines Data Warehouse:

- Sofortige und flexible Verfügbarkeit von Berichten, Statistiken und Kennzahlen.
- Information über Zusammenhänge zwischen Markt und Leistungsangebot (Kunden / Produkte und/oder Dienstleistungen).
- Umfassende Information über Geschäftsobjekte und Zusammenhänge.
- Detailinformation über Geschäftsobjekte und deren Entwicklung über die Zeit, über Geschäftsprozesse und deren Kosten und verbrauchte Ressourcen.

Historischer Überblick



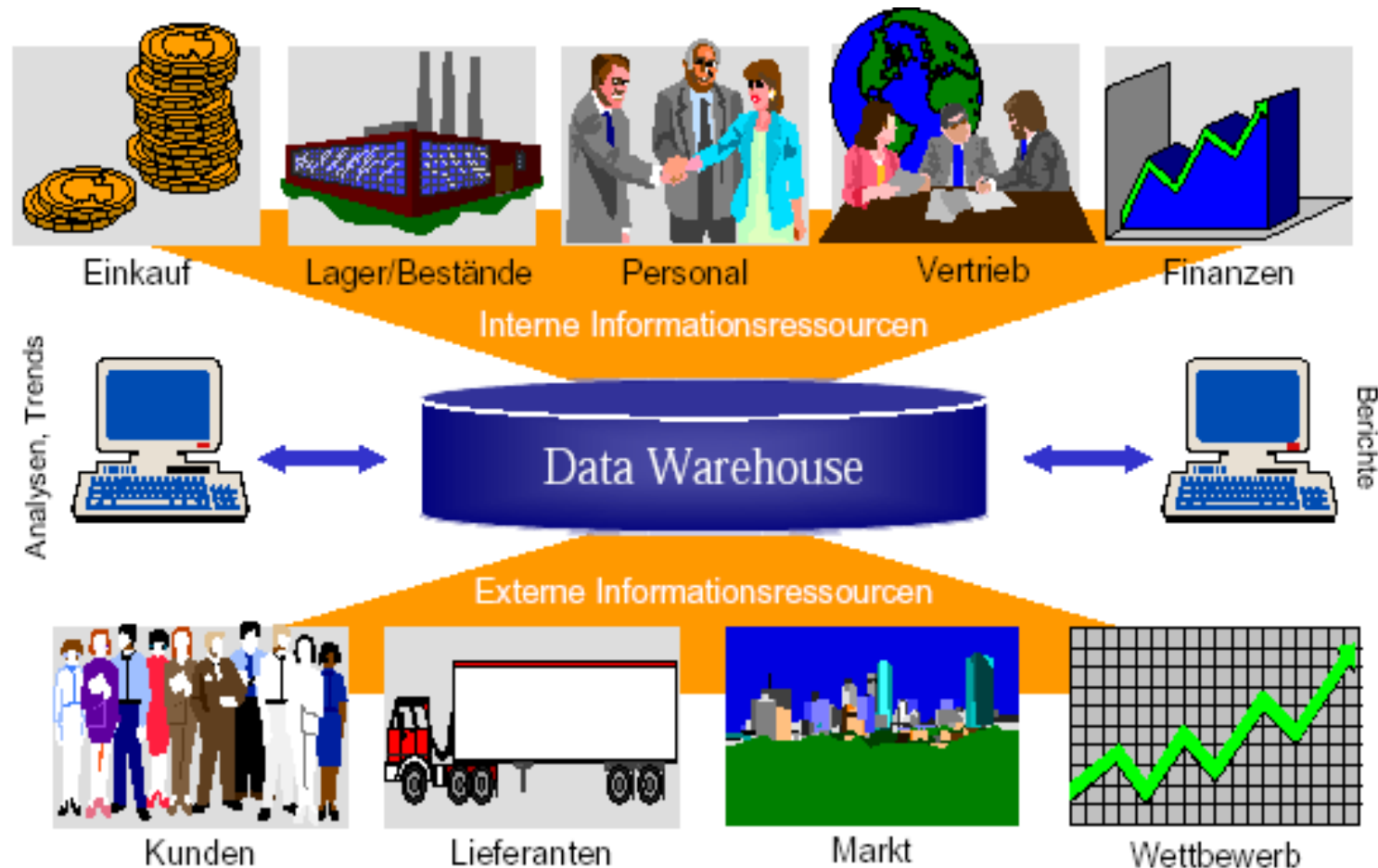
- Vor 1970 setzten sich Manager und andere Entscheidungsträger kaum selbst mit der Datenbeschaffung auseinander, sondern erhielten diese ausgedruckt auf Stapeln von Papier.
- In den Achtziger Jahren kamen erste Datenmodellierungsmethoden auf. Das erlaubte die Anforderungen an die Daten und die dazu benötigten Strukturen formal zu dokumentieren.
- Ende der Achtziger Jahre setzte sich die Unterscheidung zwischen operativen und analytischen Informationssystemen durch.
- Anfangs der Neunziger Jahre erkannte man einerseits, dass die bisher gebräuchlichen Methoden für die Datenbeschaffung nicht robust genug waren, um den zukünftigen Wachstum zu unterstützen.

Definition eines Datawarehouse

Ein Data Warehouse ist eine Sammlung von Technologien zur Entscheidungsunterstützung, die es dem Anwender erlauben soll, schneller bessere Entscheidungen zu treffen.

- Sie erlaubt: effizienten Zugriff auf integrierte Informationen. (meistens heterogenen Informationsquellen)
- Und dient: als unternehmensweite Datenbasis für Managementunterstützungssysteme.

Definition eines Datawarehouse



Vier Merkmale der Datawarehouse

- **subjektorientiert (Themenorientierung)**

Ein Data Warehouse orientiert sich an den wichtigsten Sachverhalten eines Unternehmens. Data Warehouse Daten sind subjektorientiert.(wie z.B. Kunden, Verkäufe, Produkte, Regionen)

- **integriert (Vereinheitlichung)**

In einem Data Warehouse sind alle Daten integriert und zwar ohne Ausnahme. Da diese Daten aus verschiedenen, heterogenen Datenquellen stammen, müssen strukturelle und semantische Unterschiede bereinigt werden, und die Daten müssen entsprechend einem uniformen Datenmodell in Übereinstimmung gebracht werden.

Vier Merkmale der Datawarehouse

- **nicht-volatil (Beständigkeit)**

Der Zugang zu den Daten ist hauptsächlich read-only. Updates werden im Umfeld der operativen Systeme vorgenommen. Änderungen der DW-Daten nur, wenn die geänderten Datenquellen in das DW übernommen werden.

- **variabel bezüglich der Zeit (Zeitorientierung)**

Der Zeithorizont eines DW ist signifikant länger. Er beträgt 5- 10 Jahre. Bei DW-Systemen handelt es sich um historische Daten.

DW-Daten enthalten immer gewisse Zeitelemente wie Jahr, Monat, Tag.

Abgrenzung operative Systeme vs. DW

Operative Systeme (Online Transaction Processing- OLTP):

- Operative Daten repräsentieren den gegenwärtigen Zustand des Unternehmens.
- Werden eingesetzt bei strukturierten Aufgaben, die aus kurzen, vordefinierten Transaktionen bestehen.
- Entworfen, um eine schnelle und effiziente Ausführung einer großen Anzahl von einfachen, vordefinierten read/write Transaktionen zu behandeln.
- Maximierung des Transaktionsdurchsatzes ist das grundlegende Ziel.

Abgrenzung operative Systeme vs. DW

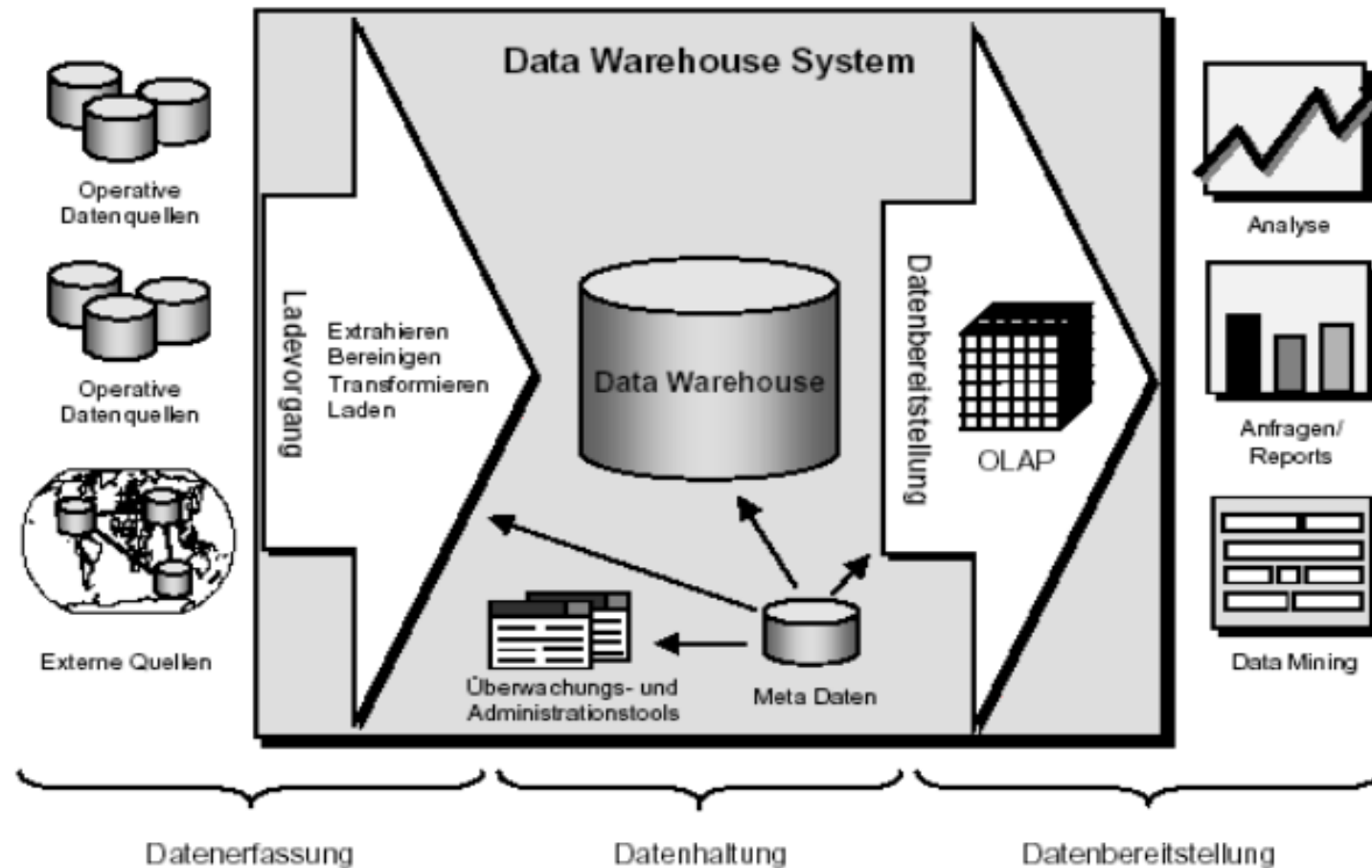
Analytische Systeme (Online Analytical Processing OLAP):

- Werden, basierend auf historischen Daten, für die Unternehmensführung und –kontrolle gebraucht.
- Enthalten konsolidierte, von mehreren operativen Datenbanken integrierte Daten.
- Hauptsächlich entworfen, um die Ausführung von komplexen, ad hoc und meist read-only Anfragen zu unterstützen.
- Anfragedurchsatz und Antwortzeit sind wichtiger als Transaktionsdurchsatz.

Abgrenzung operative Systeme vs. DW

Merkmal	operative Systeme	Data Warehouse
typische Datenstruktur	flache, nicht hierarchische Tabellen	multidimensionale Tabellen
Identifikationskriterien	eindimensional	mehrdimensional
Art der Daten	aktuelle Daten	historische Daten
Datenmanipulation	aktualisierend	analysierend
Betrachtungsebene	detailliert	aggregiert
Datenmenge	eher klein und im Zeitablauf konstant	sehr umfangreich und mit der Zeit wachsend
Zeithorizont	gegenwärtig	historisch, gegenwärtig und zukünftig
typische Transaktionen	kurze Update-Transaktionen	lange Anfrage-Transaktionen
Zugriffe	schnelle Schreib-/Lesezugriffe innerhalb wohlstrukturierter Aufgaben	Lesezugriffe für verschiedene Rechercheaufgaben; Zugriffspfade können nicht vorherbestimmt werden
Datenmodell	keine Redundanzen, da hochgradig normalisiert	Denormalisierung führt zu erheblicher Redundanz
Fokus	Anwendungs- bzw. Prozeßorientierung	Orientierung an unternehmensrelevanten Sachverhalten
Anfrageverhalten	vorhersehbare Anfragen	ad hoc Anfragen
Optimierungsziel	Update-Optimierung	Anfrage-Optimierung

Architektur von DW Systemen



Datenerfassungsebene

Die Datenerfassungsebene umfaßt alle Aufgaben, die mit dem Laden der Daten in das Data Warehouse zusammenhängen, sowohl beim initialen Ladevorgang als auch bei der periodischen Aktualisierung.

Besteht aus einer Vielzahl von Werkzeugen zur :

- Extraktion
- Bereinigung
- Transformation
- Laden der Daten in das Data Warehouse.

Datenhaltungsebene

Die Aufgabe der Datenhaltungsebene ist die Speicherung der Daten im Data Warehouse.

Die Datenhaltung in Data Warehouses muß zwei Hauptanforderungen erfüllen:

- Sehr große Datenmengen zu speichern.
- Schnelle Antwortzeiten auf OLAP-Anfragen möglich sein.

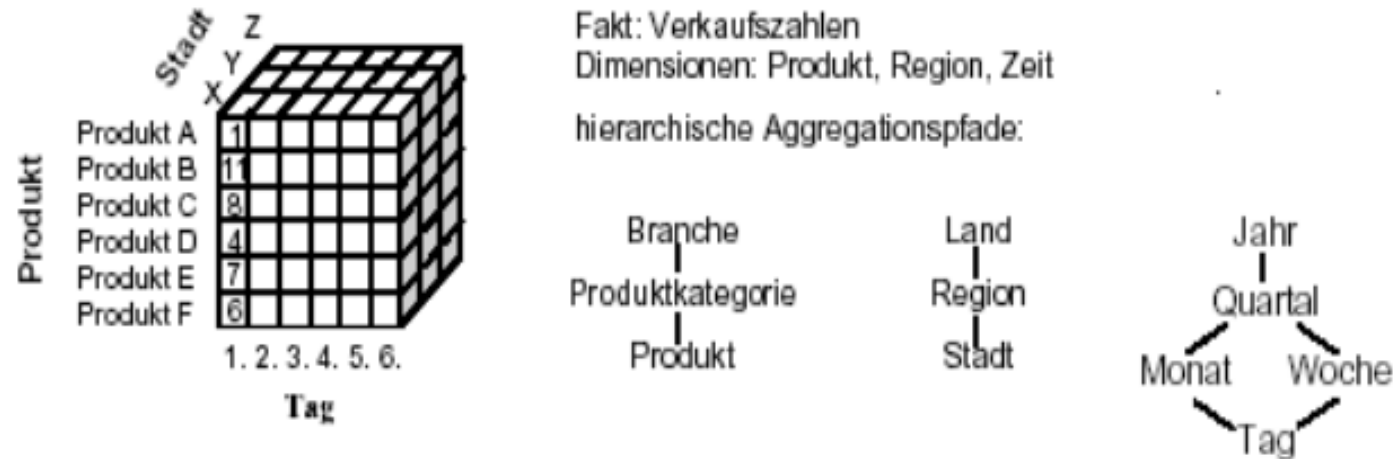
Datenbereitstellungsebene

Die Datenbereitstellungsebene besteht üblicherweise aus einem OLAP-Server, der die Daten des OLAP-Speichers an Front End Tools. (wie z.B. Analyse-, Anfrage)

- Auf diesen Daten muß eine multidimensionale Analyse möglich sein.

Multidimensionales Datenmodell

Das zentrale Objekt einer multidimensionalen Struktur wird *Würfel* (engl. Cube) genannt. Ein Würfel besteht aus einer Menge von orthogonalen Kanten, den *Dimensionen*.



Elemente eines multidimensionalen Modells



- **Kennzahlen** (Variablen)

Sie sind meist quantitative, in numerischer Form vorliegende Werte (wie z.B. Umsatzdaten, kosten, verkaufte Produkte,...)

Betriebswirtschaftliche Variablen sind die eigentlichen Inhalte von OLAP-Würfeln.

- **Dimensionen** (betriebswirtschaftliche Entscheidungsobjekte)

Die Unterteilung von Geschäftsdaten nach verschiedenen Blickwinkeln.

Zeitstruktur: z.B. Tag → Monat → Quartal → Jahr

OLAP (online Analytical Processing)

- Mit OLAP wird eine Datenbanktechnologie bezeichnet, die speziell für Ad hoc (on-line)- Auswertungen mit komplexem (analytical) Charakter entwickelt wurde.
- Diese Technologie ist sehr gut geeignet für Manager, Controller, IT-Profis und Marketingleute. OLAP ermöglicht es Ihnen, Ihre Unternehmens-Informationen aus allen Blickwinkeln zu betrachten.

OLAP- Prinzip

- subjektorientierte, beständige, integrierte und zeitbezogene Daten (Extrakte aus DWH).
- Daten werden in multidimensionaler Sicht (Extrakt aus dem Data Warehouse) dargestellt.
- Unterstützt Manipulation der Daten und der Betrachtungsweise.
- Durch gute Antwortzeiten des Systems (gefordert < 15 Sekunden für OLAP Berichte).
- OLAP verwendet zur Datenhaltung entweder herkömmliche relationale Datenbanken (ORACLE, Informix; Sybase, MS-SQL Server, etc.) oder spezielle multidimensionale Datenbanken (TM1).

OLAP-Funktionen

OLAP wird durch folgende Funktionen charakterisiert

Navigation innerhalb der multidimensionalen Datensicht:

- **Drill-Down:** innerhalb der Dimensionen zu detaillierteren Daten, den Aggregationspfad hinunter wandern.
- **Roll-Up:** innerhalb einer Dimensionen zu mehr aggregierten Daten, den Aggregationspfad hinauf wandern.

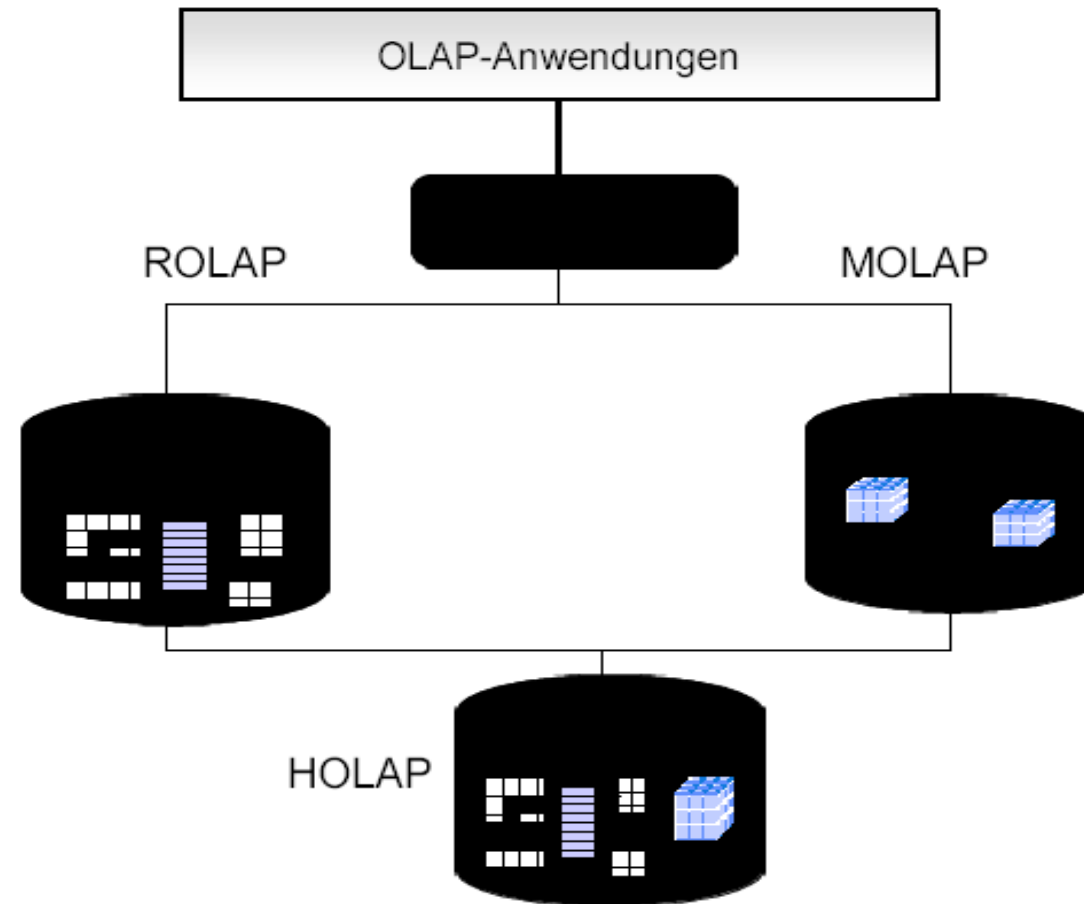
OLAP-Funktionen

- **Pivot:** (*Rotate*) betrachten der Daten mit aus unterschiedlichen Perspektiven, vertauschen der Reihenfolge der Dimensionen, welche dargestellt werden.
- **Slice:** abschneiden einer Scheibe aus den multidimensionalen Würfel.
- **Dice:** ausschneiden eines Teilwürfels.

OLAP- Architekturen

- **ROLAP:** Relational On Line Analytical Processing, relationale Datenspeicherung – Tabellenform.
- **MOLAP:** Multidimensional On Line Analytical Processing, multidimensional Datenspeicherung, n-dimensionaler Würfel.
- **HOLAP:** Hybrid On Line Analytical Processing (HOLAP)
Speicherung eines Teils des DWH's in Form von Würfeln.

OLAP- Architekturen



Einsatzmöglichkeiten von DW-Systemen

- Marketing
- Kundeninformation
- Vertrieb
- Produkt-Analyse
- Personal-wesen
- Buchhaltung

Zukunft der DW- Systeme

- Der Data Warehouse- Markt betrug Ende 1998 bereits 8 Billionen US\$ und der Trend ist nicht nachlassend.
- Im Forschungsbereich gilt er als eines der heißesten Themen.
- Das Data Warehouse ist dynamisch und wird sich kontinuierlich weiterentwickeln.

What is a Data Warehouse?

A data repository that makes operational and other data accessible in a form that is readily acceptable for decision support and other user applications.

Note: A data warehouse is **not** another word for a database. The specific purpose of a data warehouse is to support decisions not operations.

Data warehouses vs operational databases

- an operational database is normalised. Each data item is only held once.
- databases have very fast insert/update performance because only a small amount of data in those tables is affected each time a transaction is processed.
- Older data may be periodically purged from operational systems to improve performance.
- Data warehouses are optimized for speed of data retrieval.
- data in data warehouses may be stored using a dimension-based model.
- To speed data retrieval, data warehouse data are often stored multiple times.
- Data may be held in the data warehouse even after the data has been removed from the operational systems.

How is the data analysed?

Analytics techniques – types of model

- Simulation
- Decision analysis
- Statistics : averages, correlations,
- Linear programming: optimisation
- Queuing theory: “waiting line”analysis
- Network analysis: Maximise flow through a network
e.g. A supply chain
- Multi-criteria decision making: scoring models

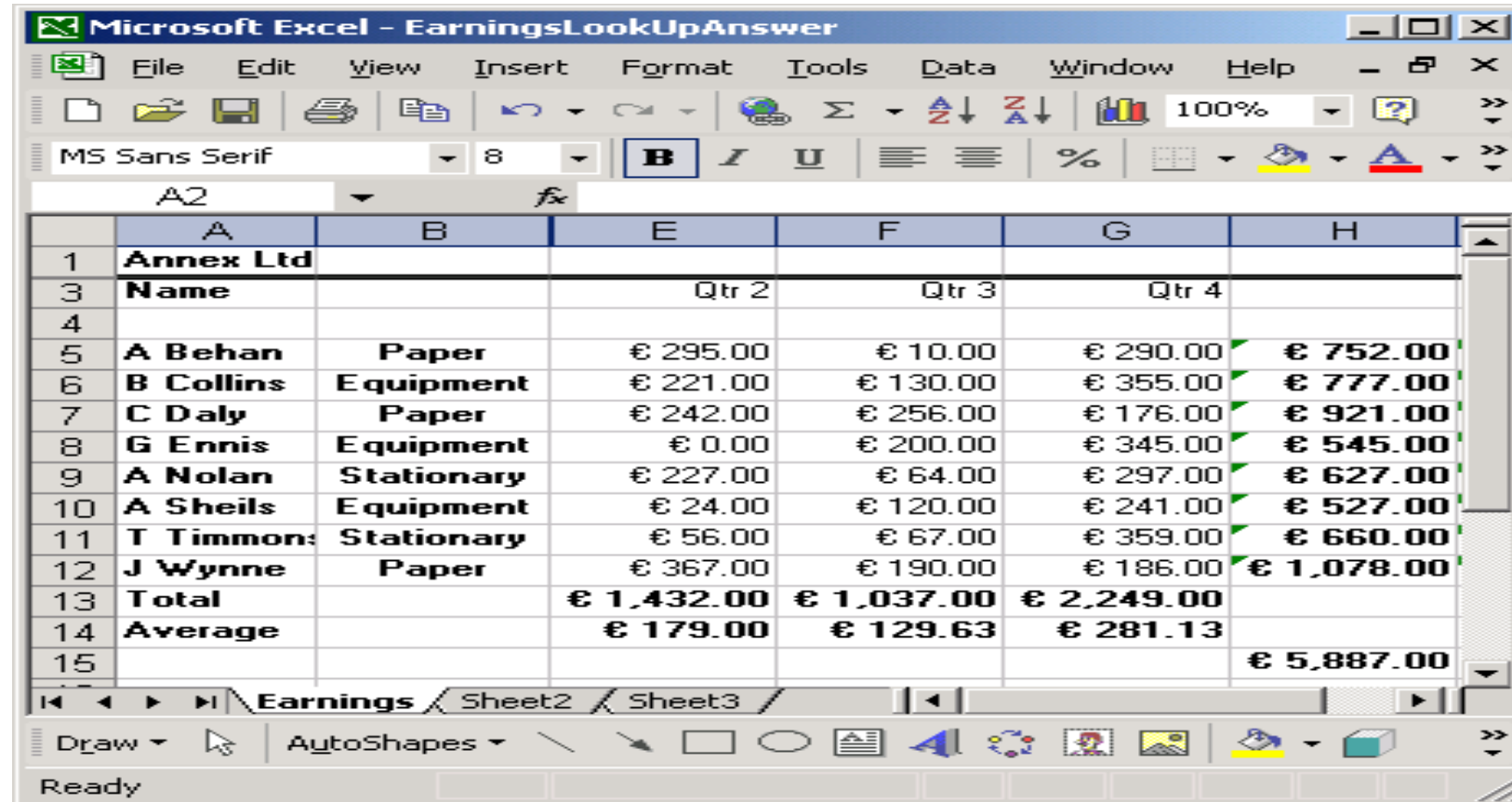
What is Data Mining?

- **Data mining is a** capability to support the recognition of previously unknown but potentially useful relationships within large databases/ data warehouses.
- Basically software to analyse data and spot patterns.

Visualising Data

- Digital images- These can be still or animated.
- Maps e.g. Geographic Information Systems
- Multidimensions - (OLAP)
- Tables and graphs
- Virtual reality
- Dashboards

A Table

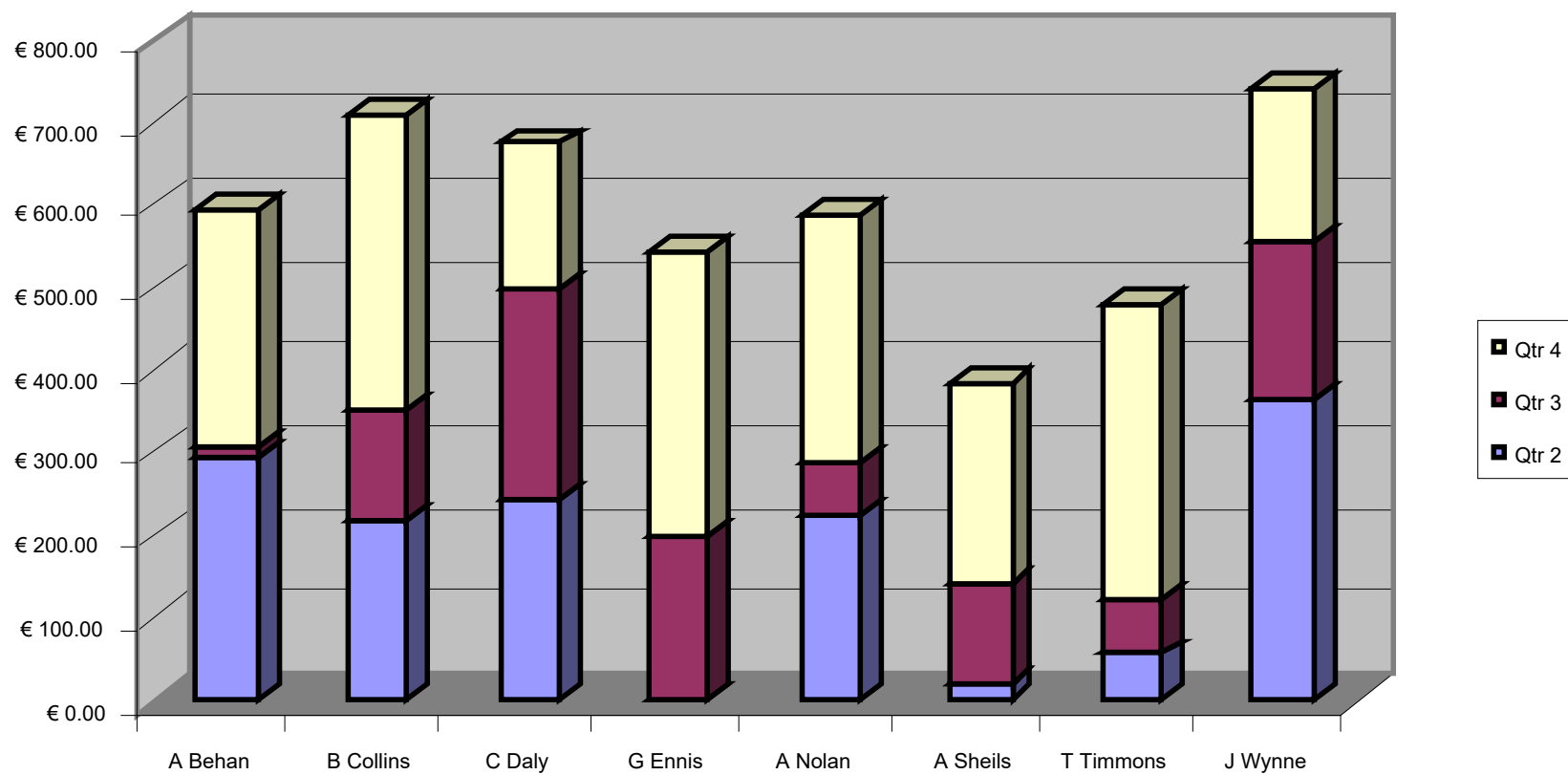


The screenshot shows a Microsoft Excel window titled "Microsoft Excel - EarningsLookUpAnswer". The interface includes a menu bar (File, Edit, View, Insert, Format, Tools, Data, Window, Help), a toolbar with various icons, and a formatting toolbar with options like font face (MS Sans Serif), size (8), bold (B), italic (I), underline (U), and alignment. The active cell is A2. The spreadsheet contains a table with the following data:

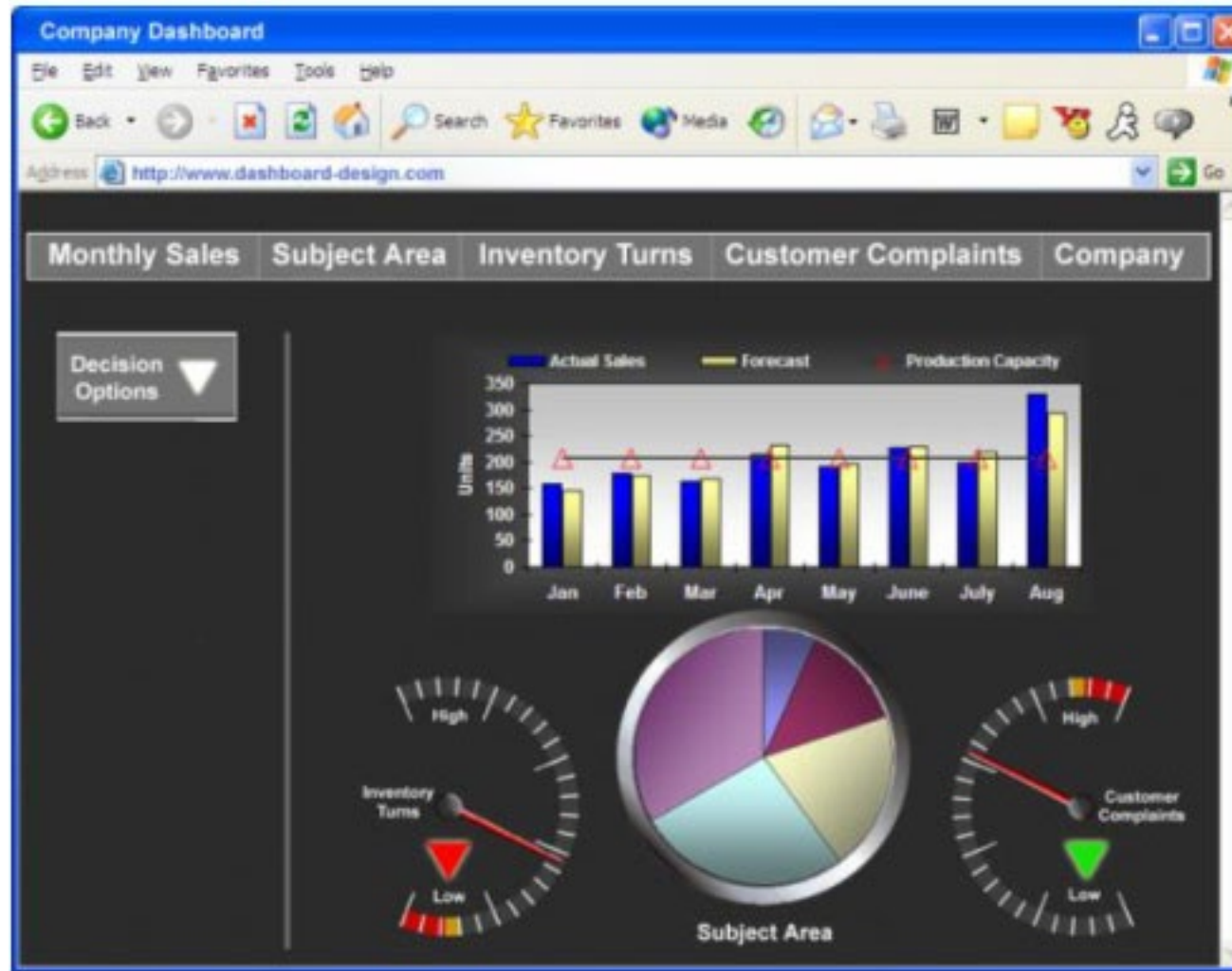
	A	B	E	F	G	H
1	Annex Ltd					
3	Name		Qtr 2	Qtr 3	Qtr 4	
4						
5	A Behan	Paper	€ 295.00	€ 10.00	€ 290.00	€ 752.00
6	B Collins	Equipment	€ 221.00	€ 130.00	€ 355.00	€ 777.00
7	C Daly	Paper	€ 242.00	€ 256.00	€ 176.00	€ 921.00
8	G Ennis	Equipment	€ 0.00	€ 200.00	€ 345.00	€ 545.00
9	A Nolan	Stationary	€ 227.00	€ 64.00	€ 297.00	€ 627.00
10	A Sheils	Equipment	€ 24.00	€ 120.00	€ 241.00	€ 527.00
11	T Timmons	Stationary	€ 56.00	€ 67.00	€ 359.00	€ 660.00
12	J Wynne	Paper	€ 367.00	€ 190.00	€ 186.00	€ 1,078.00
13	Total		€ 1,432.00	€ 1,037.00	€ 2,249.00	
14	Average		€ 179.00	€ 129.63	€ 281.13	
15						€ 5,887.00

The status bar at the bottom indicates "Ready".

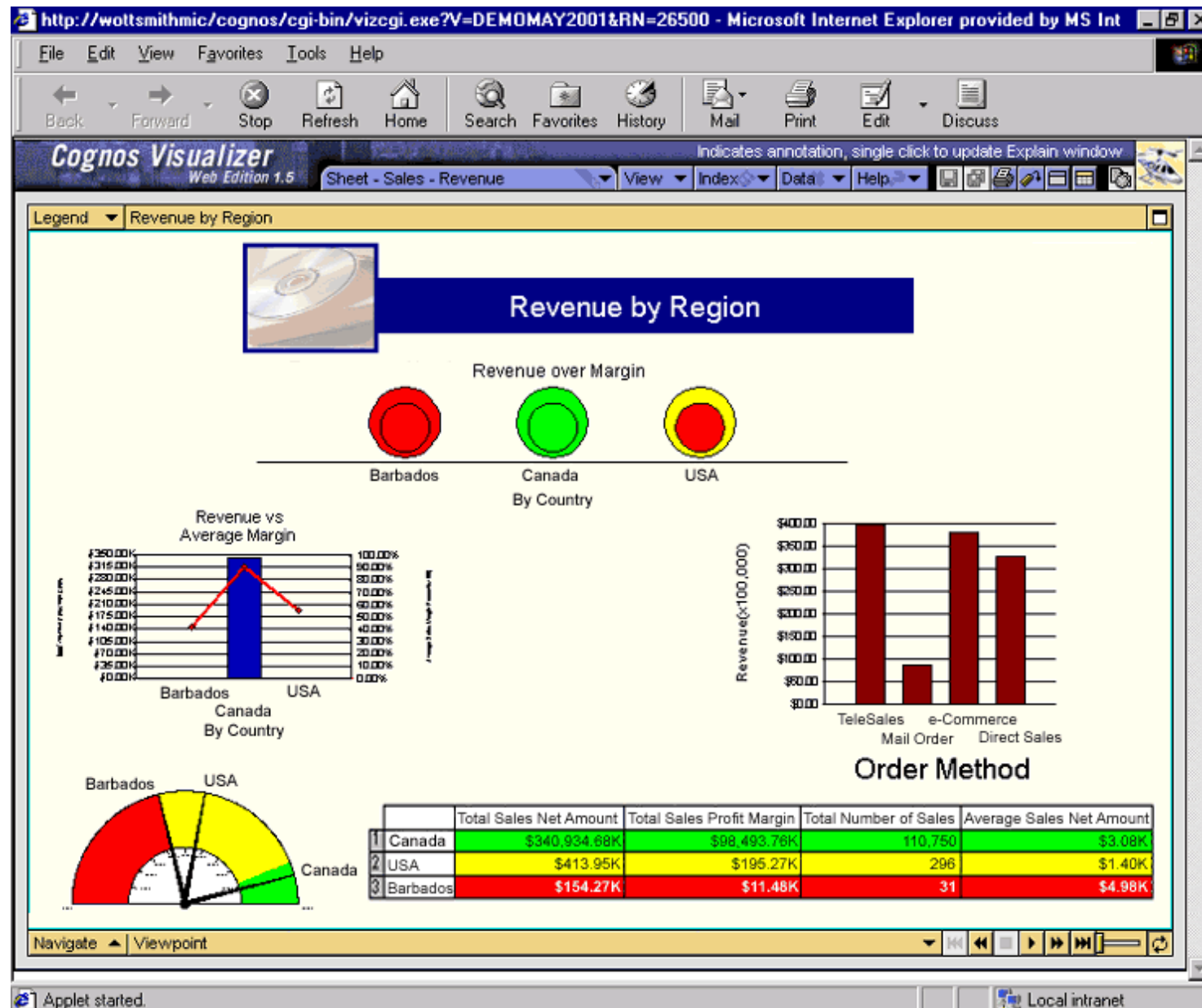
A Chart



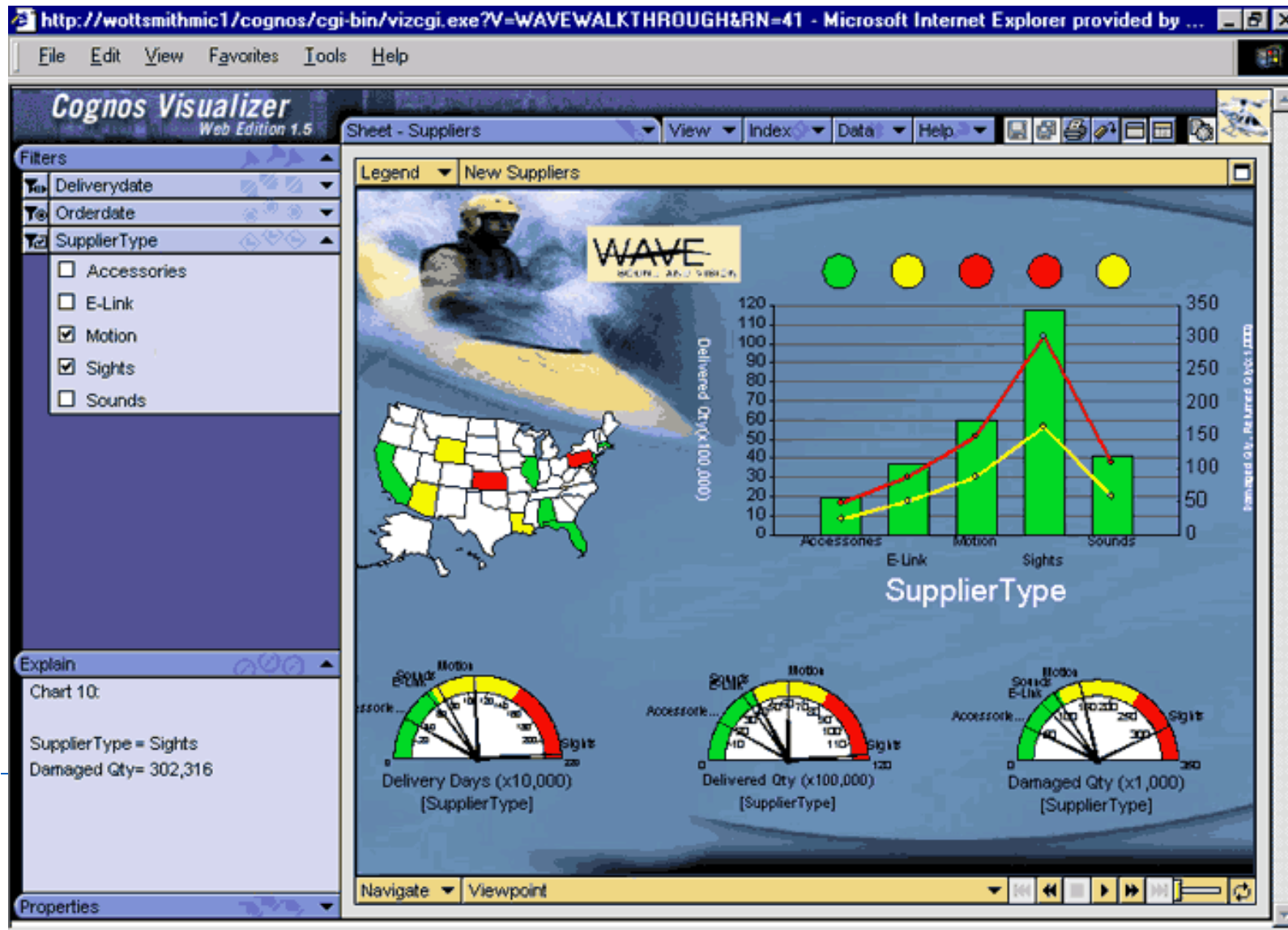
Dashboards



Taken from <http://gbr.pepperdine.edu/034/bis.html>



multiple, synchronized chart types



A visualization with multiple displays showing a Supplier scorecard in conjunction with a geographical display.

[Return to Document](#)

Tableau visualization white paper

[whitepaper_visual-analysis-guidebook_0.pdf](#)

Summary



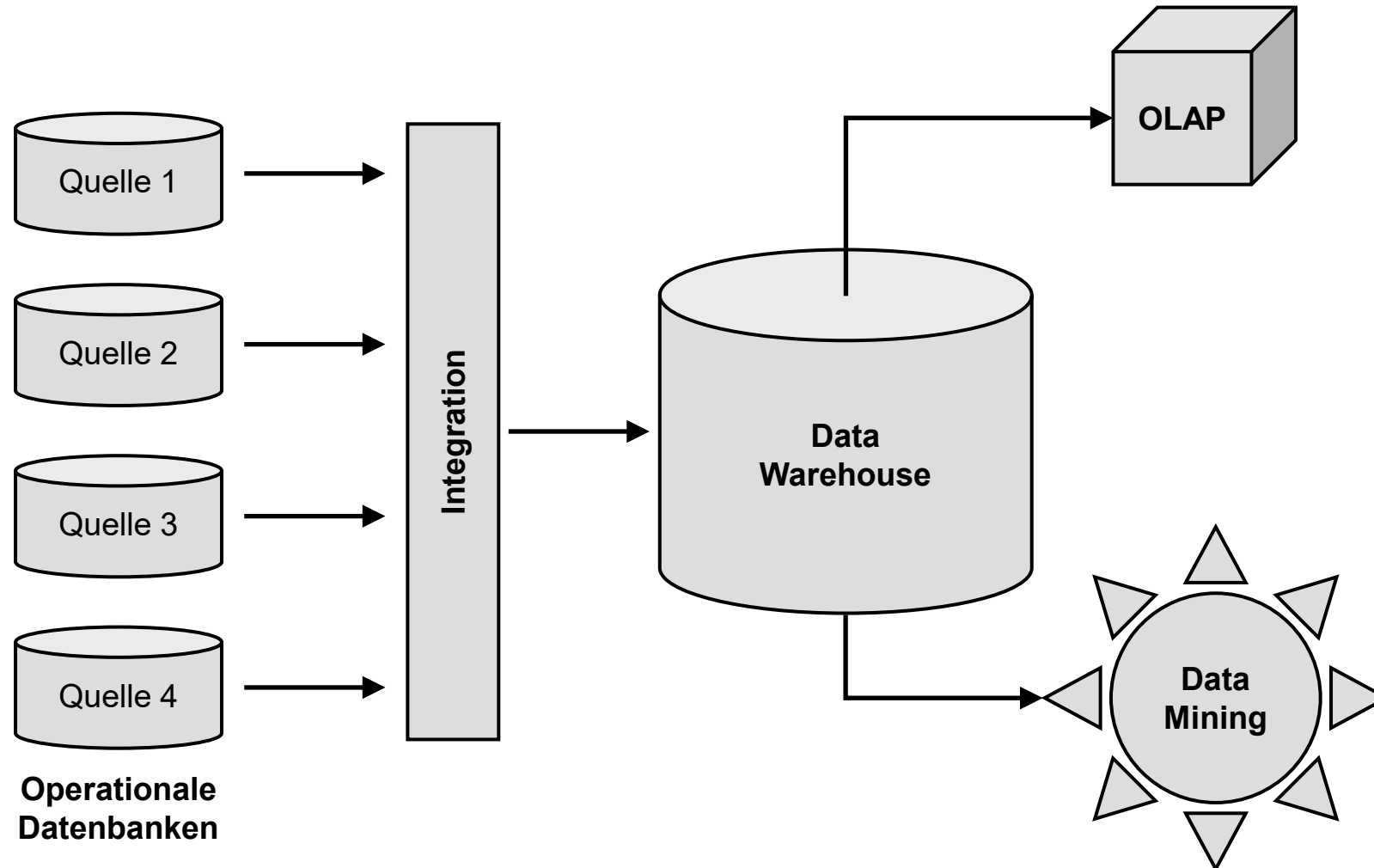
- Decision support involves data and models
- BI involves acquiring data and information from a wide variety of sources and utilising them in decision-making. Data is
 - Gathered, selected
 - Consolidated and integrated -> data warehouse
 - Analysed in different ways (analytic techniques)
 - Results are Visualised

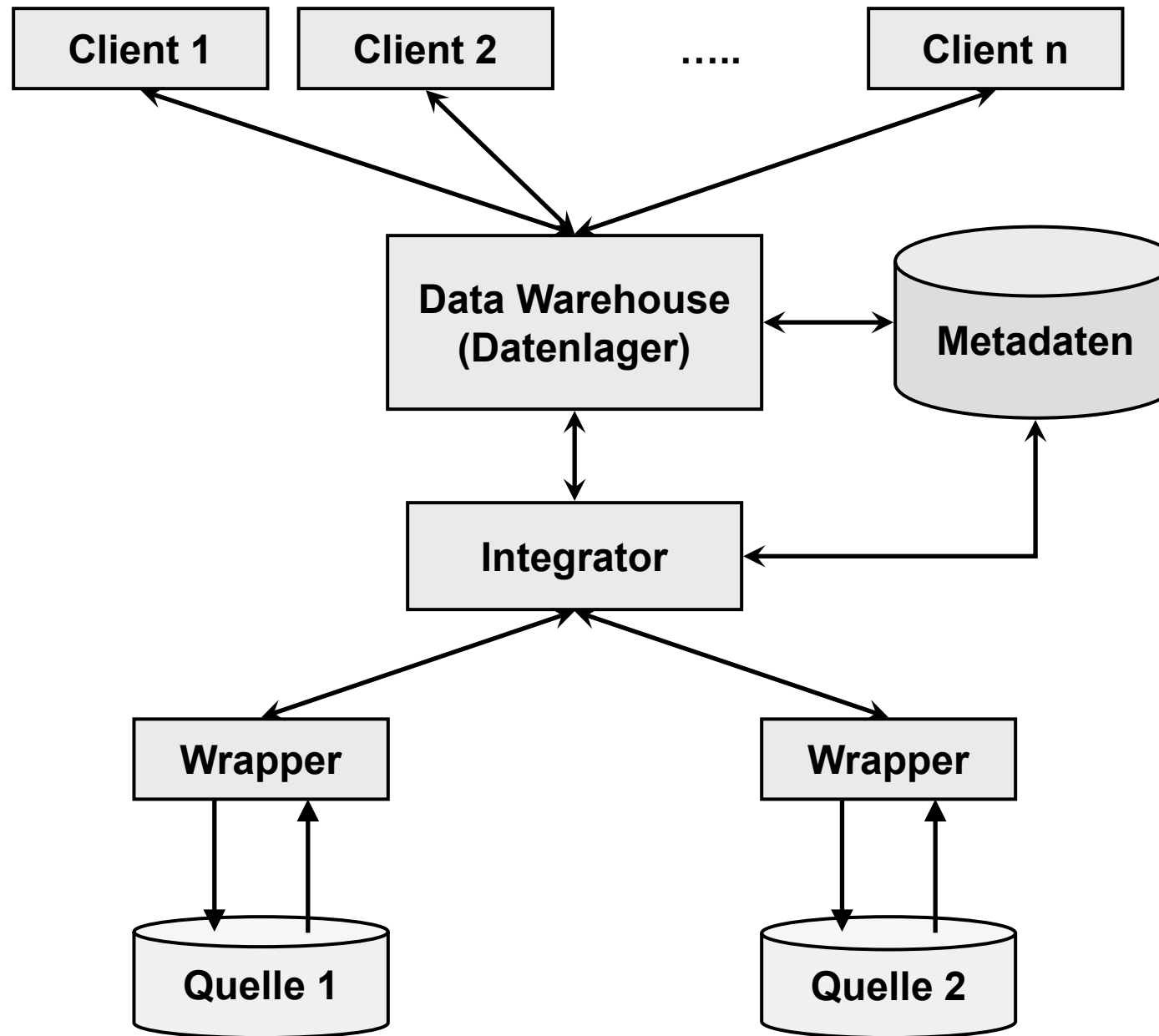
We need to Understand

- Data issues – data quality
- Where data comes from
- How data is stored: data warehouses
- How data is analysed
- Tools to do this.
- Limitations of the computer
- Our own blind spots (if this is possible)!

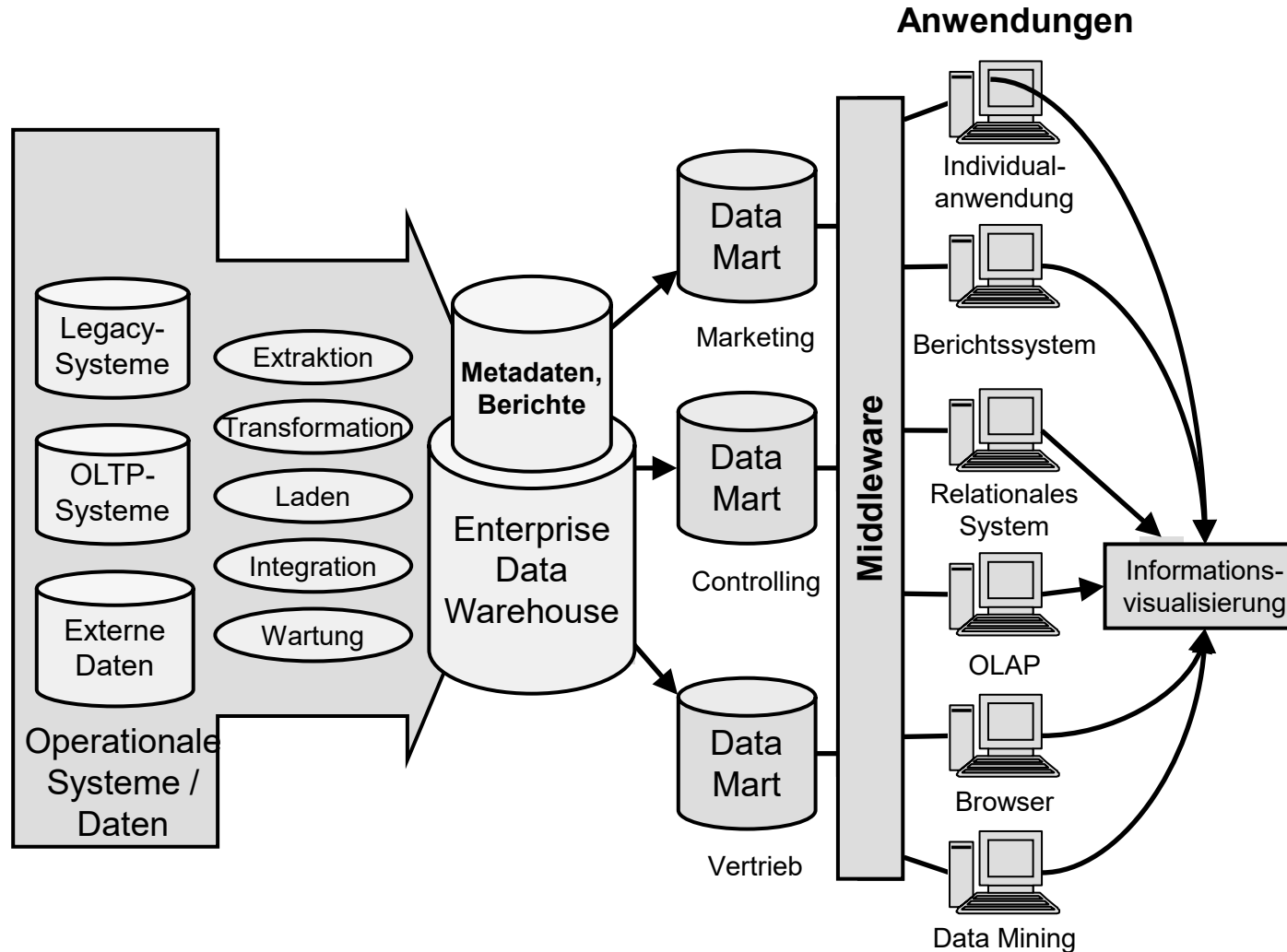
ETL

15.1 Data-Warehouse-Szenario mit Anwendungen

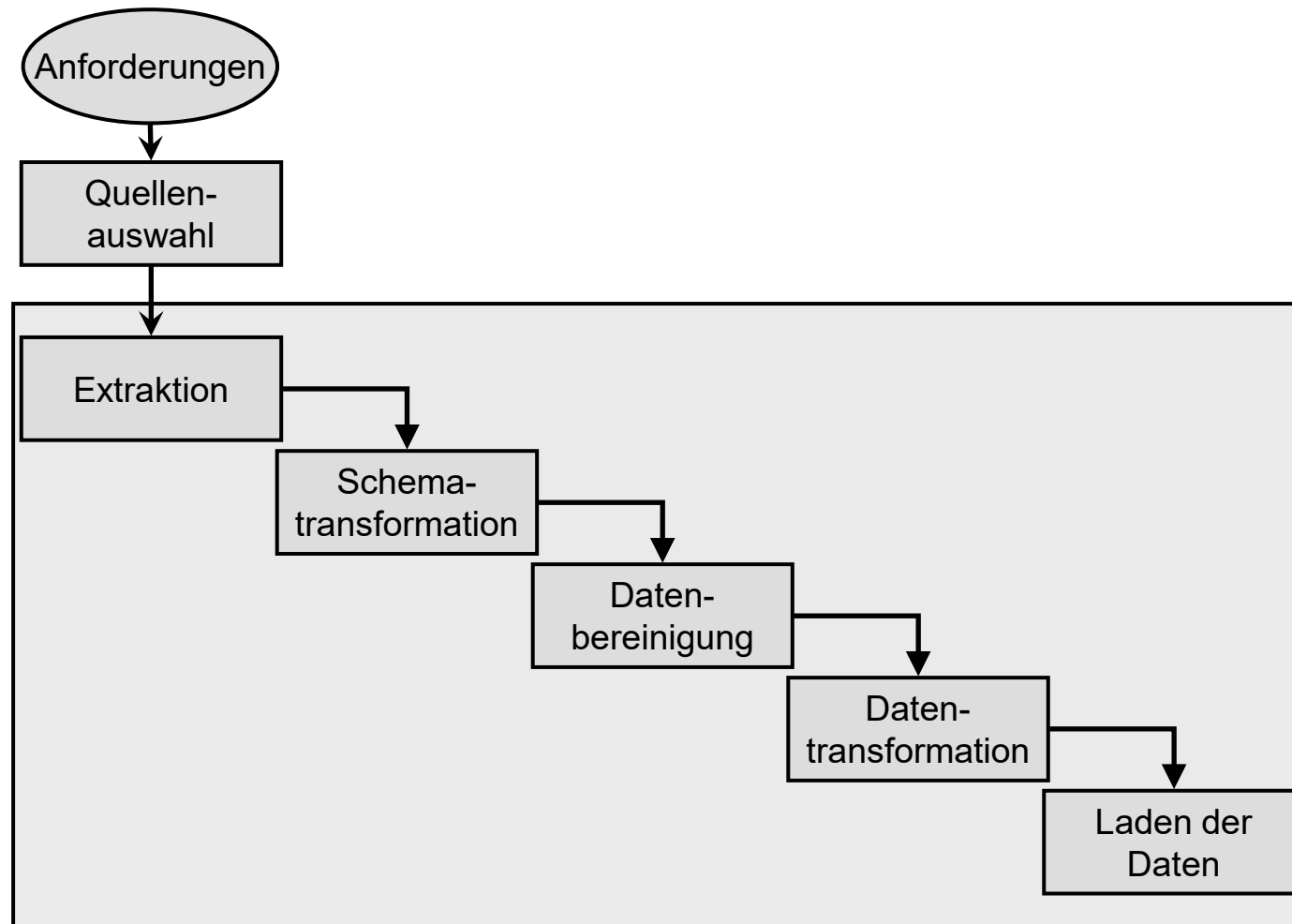




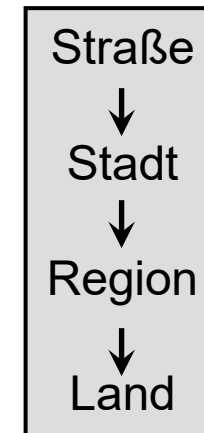
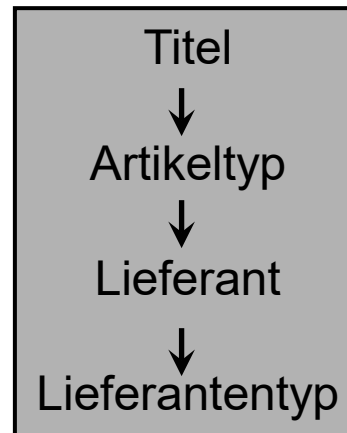
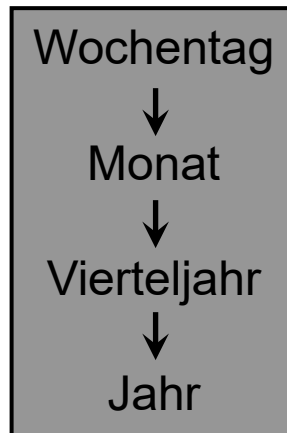
15.3 Allgemeine Enterprise-Data-Warehouse-Architektur



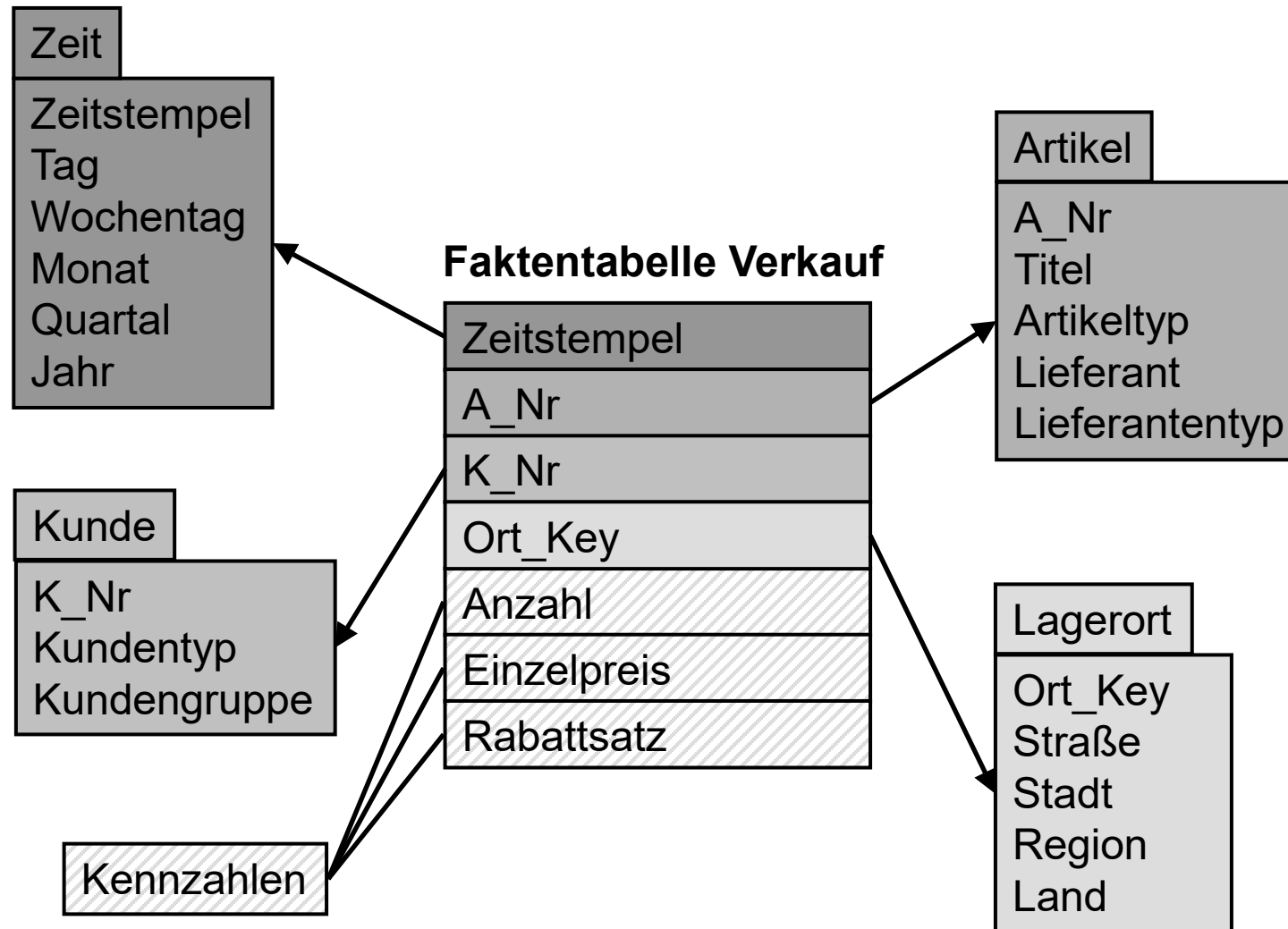
15.4 ETL-Prozess in der Übersicht



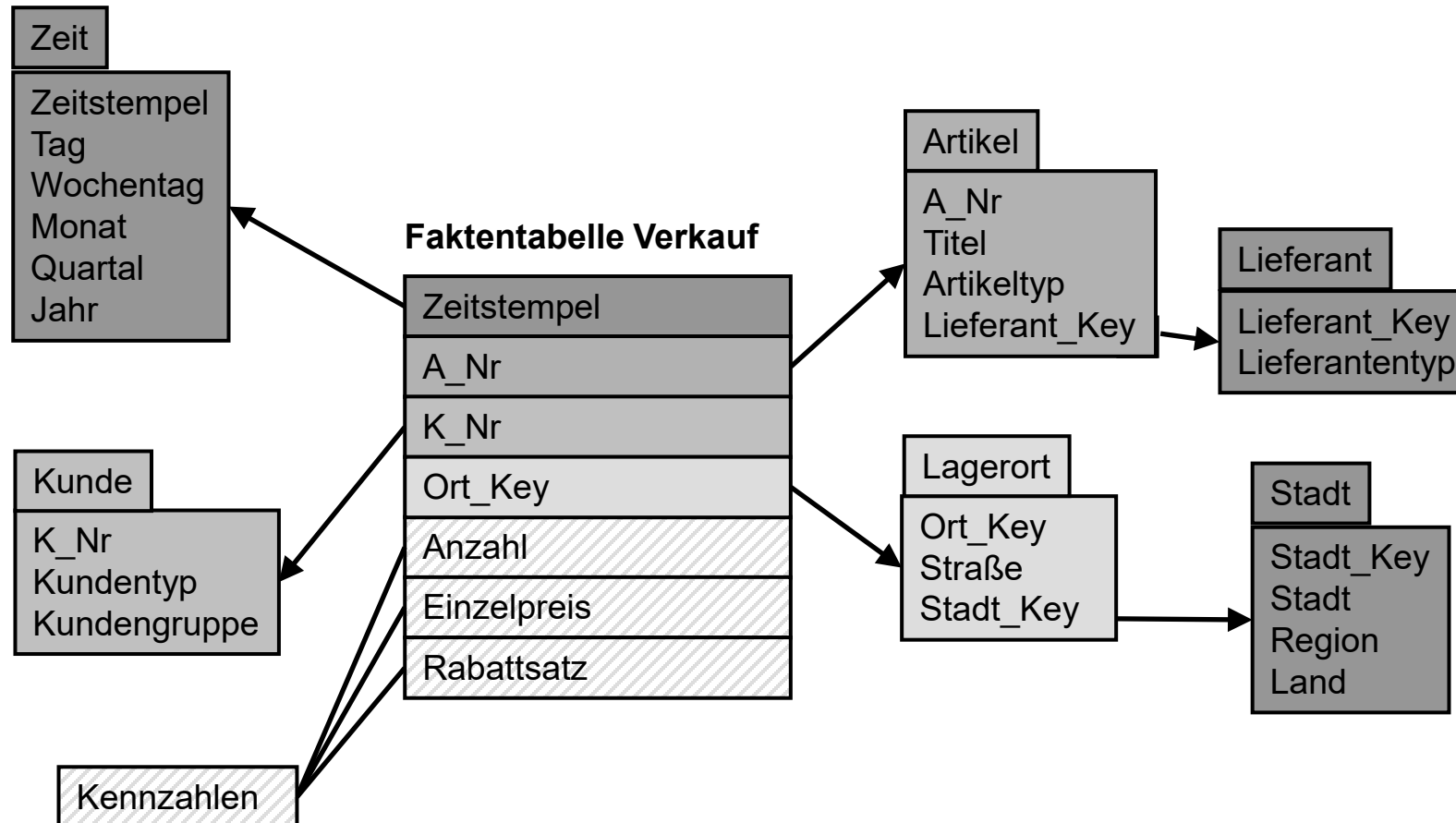
15.5 Attribuhierarchien für einzelne Dimensionen



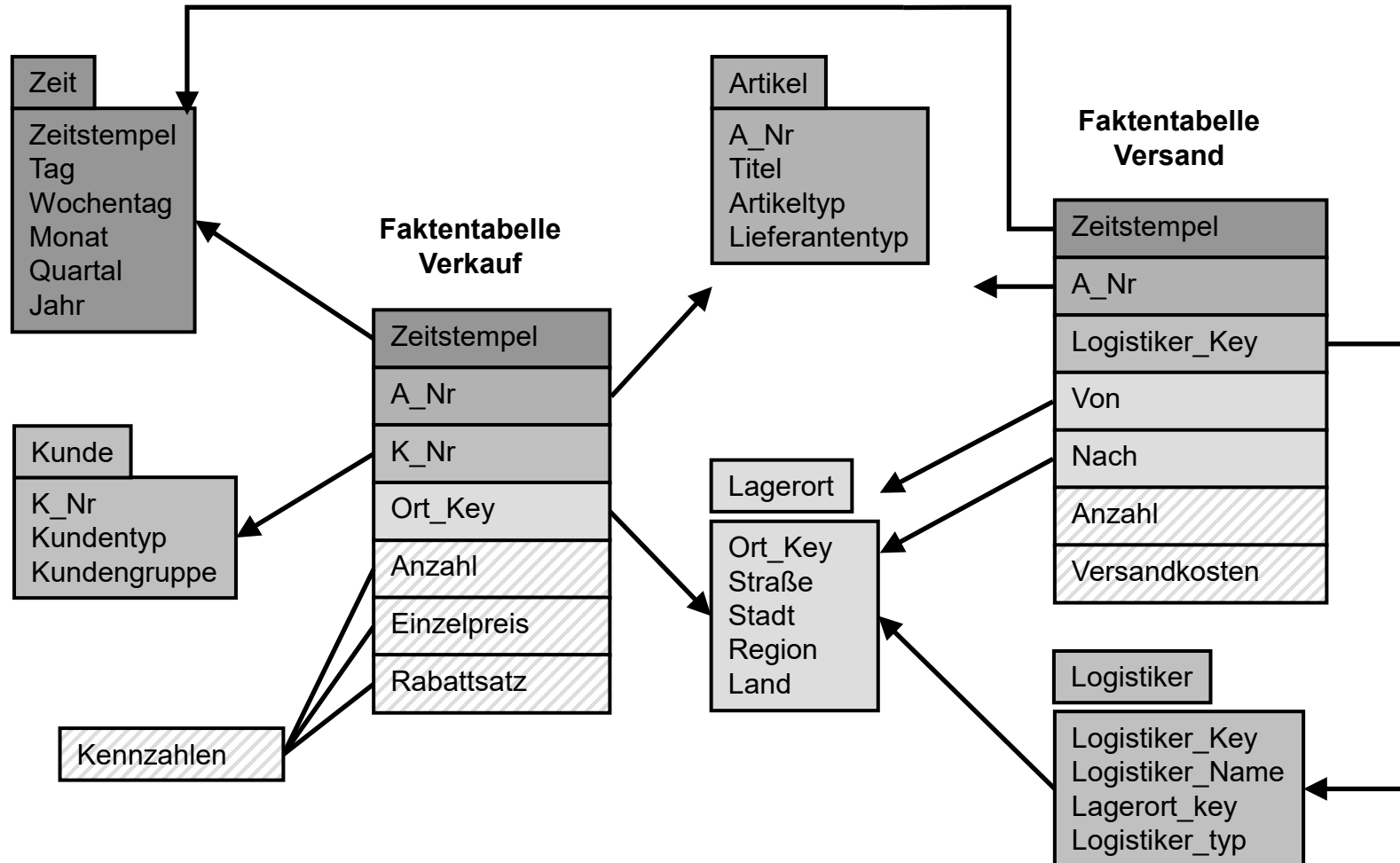
15.6 Sternschema für relationales OLAP



15.7 Schneeflockenschema für relationales OLAP



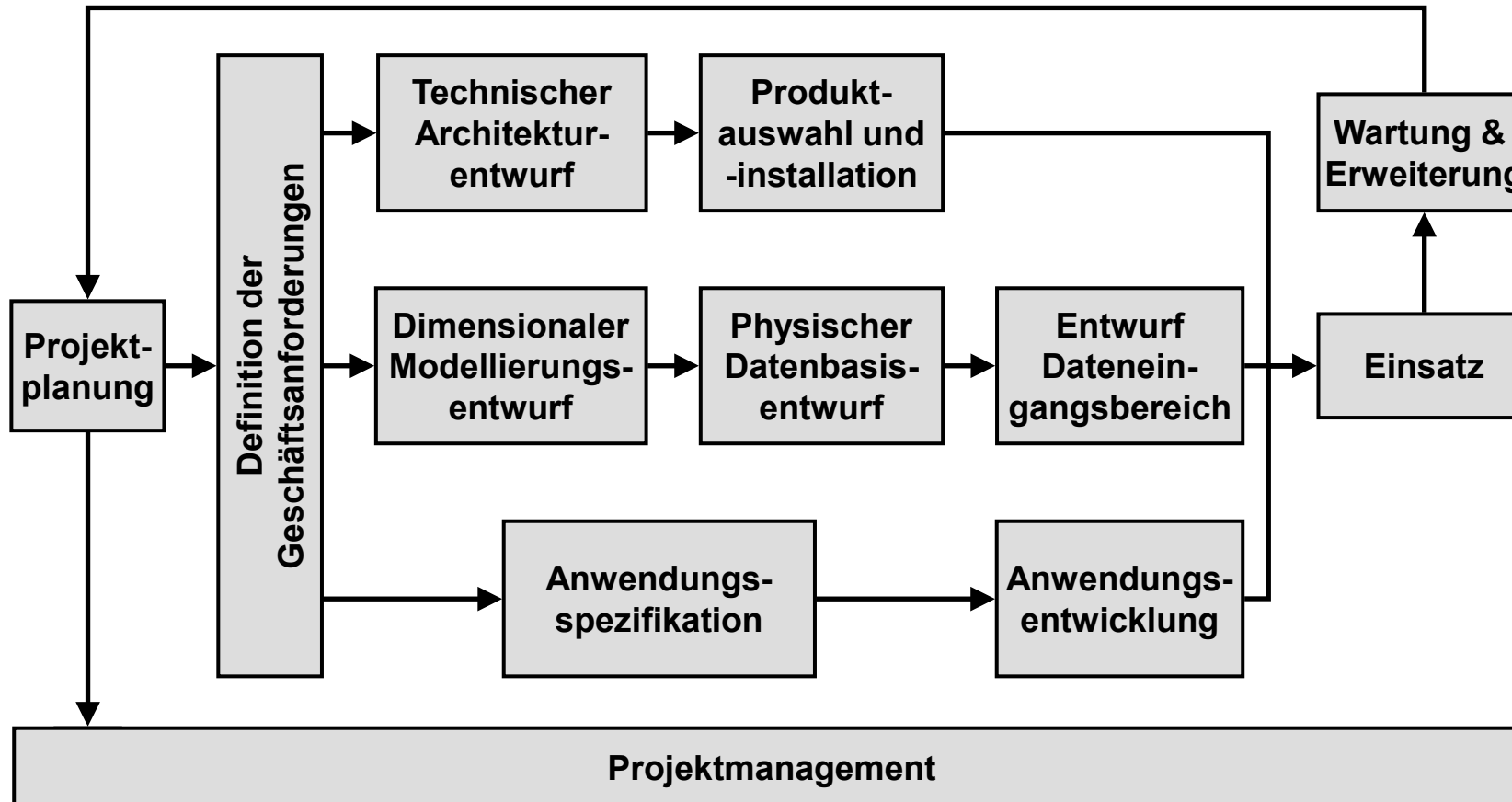
15.8 Constellation-Schema für relationales OLAP



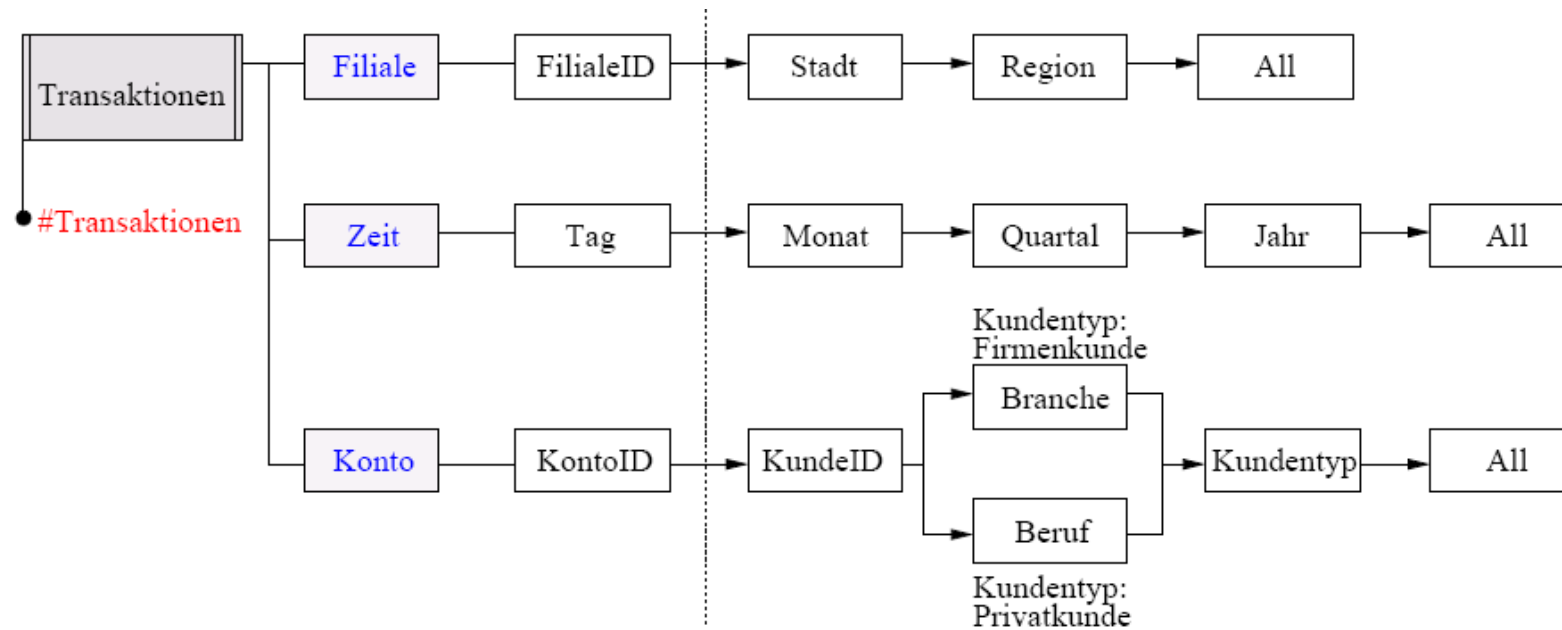
15.9 Beispiel einer Faktentabelle

<i>Sales</i>			
<i>Model</i>	<i>Year</i>	<i>Color</i>	<i>Sold</i>
Chevy	1990	red	5
Chevy	1990	white	87
Chevy	1990	blue	62
Chevy	1991	red	54
Chevy	1991	white	95
Chevy	1991	blue	49
Chevy	1992	red	31
Chevy	1992	white	54
Chevy	1992	blue	71
Ford	1990	red	64
Ford	1990	white	62
Ford	1990	blue	63
Ford	1991	red	52
Ford	1991	white	9
Ford	1991	blue	55
Ford	1992	red	27
Ford	1992	white	62
Ford	1992	blue	39

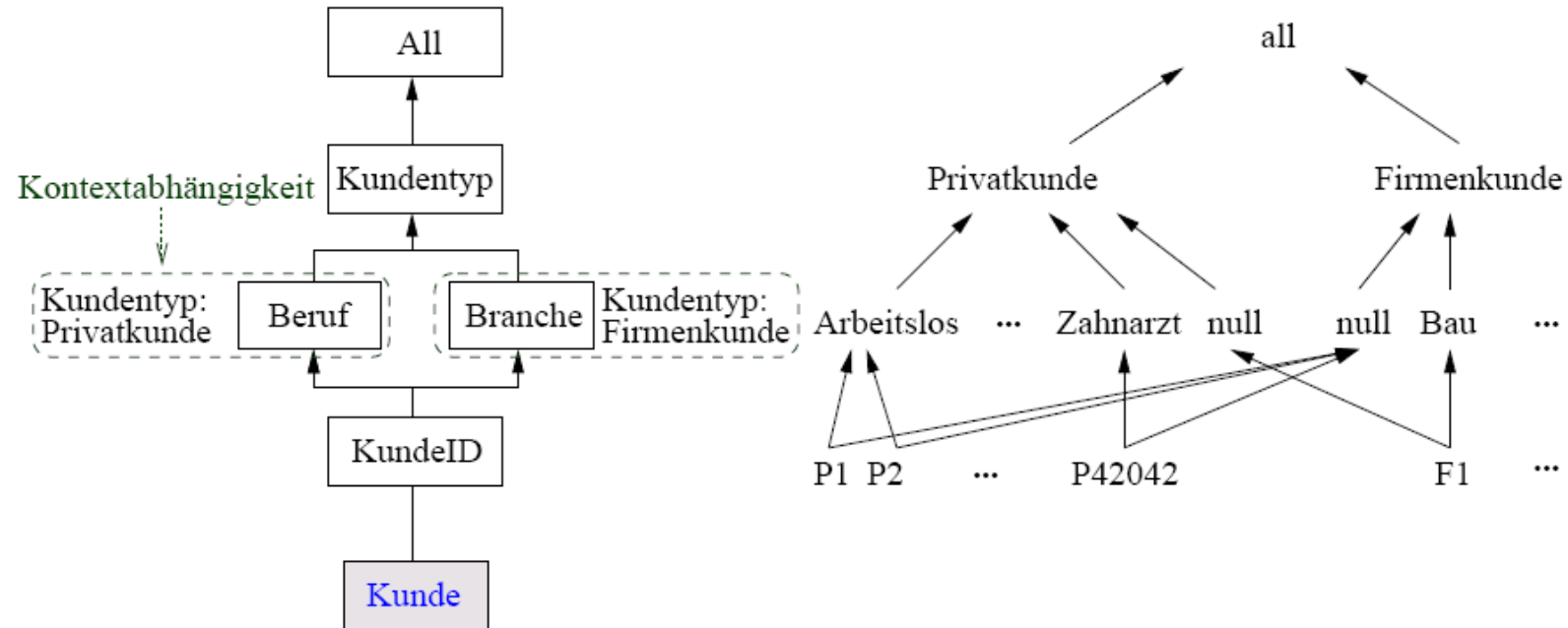
15.11 Datenlager-Lebenszyklus (nach Lehner)



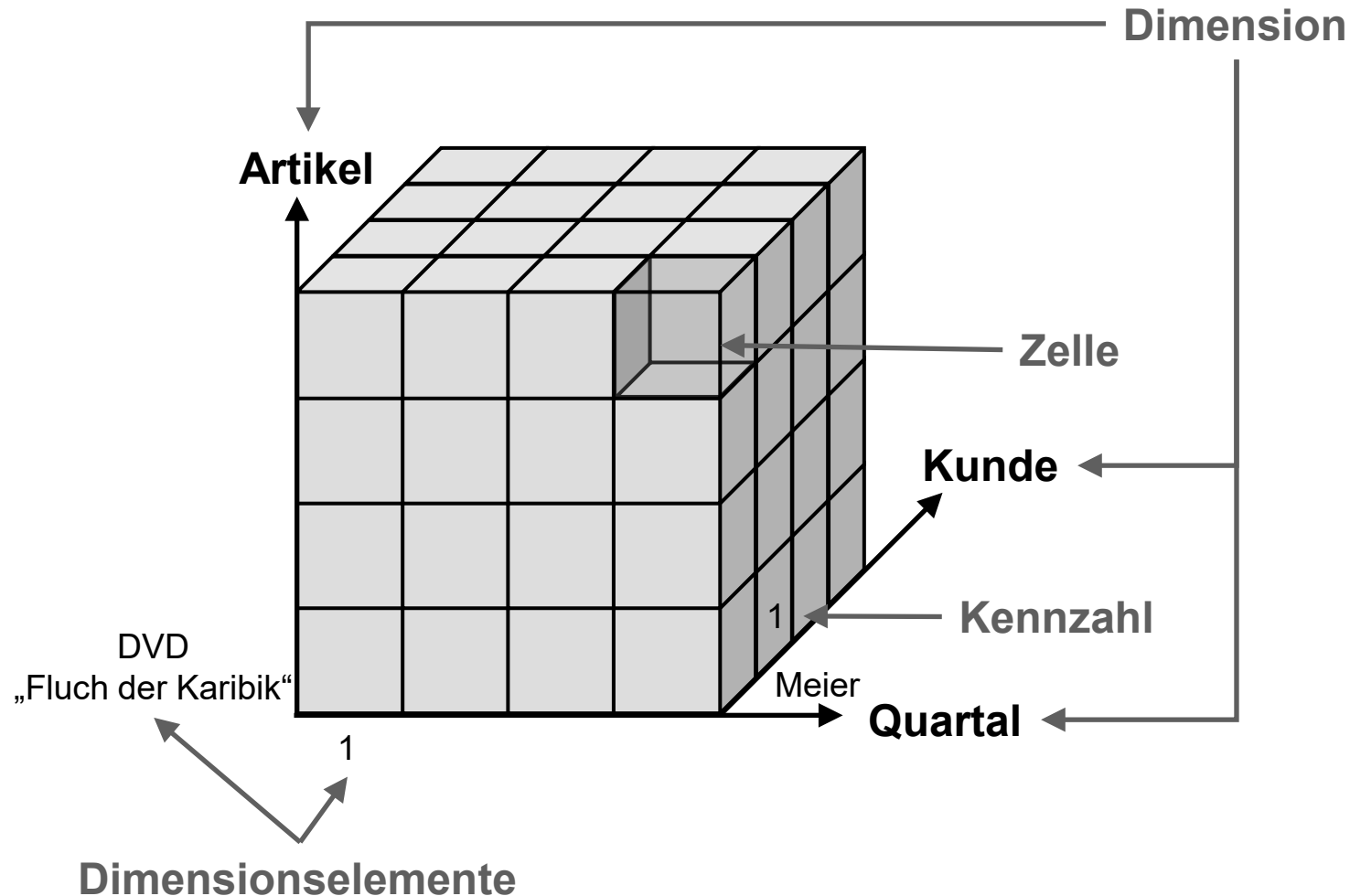
15.12 Exemplarisches Faktenschema



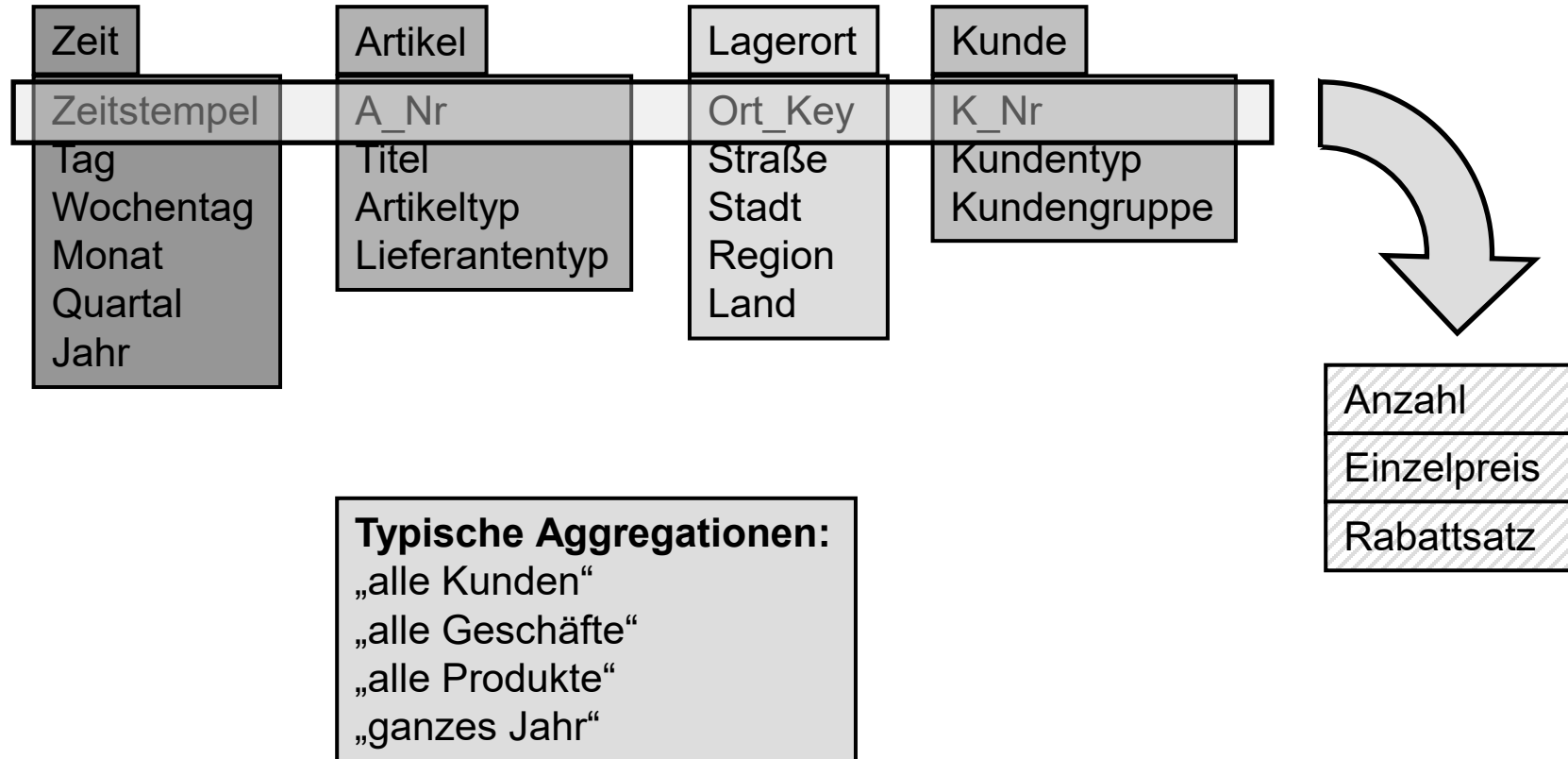
15.13 Dimensionsschema Kunde mit Instanz



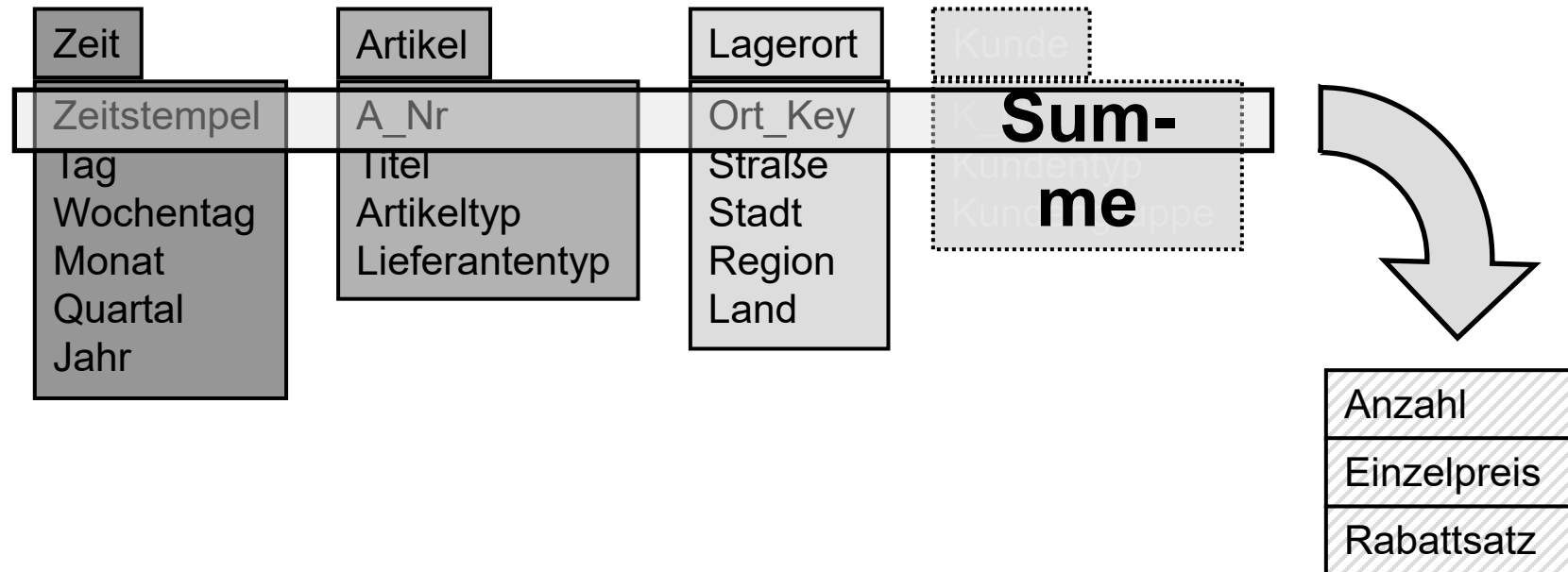
15.14 Die Bestandteile eines (3D-) Datenwürfels



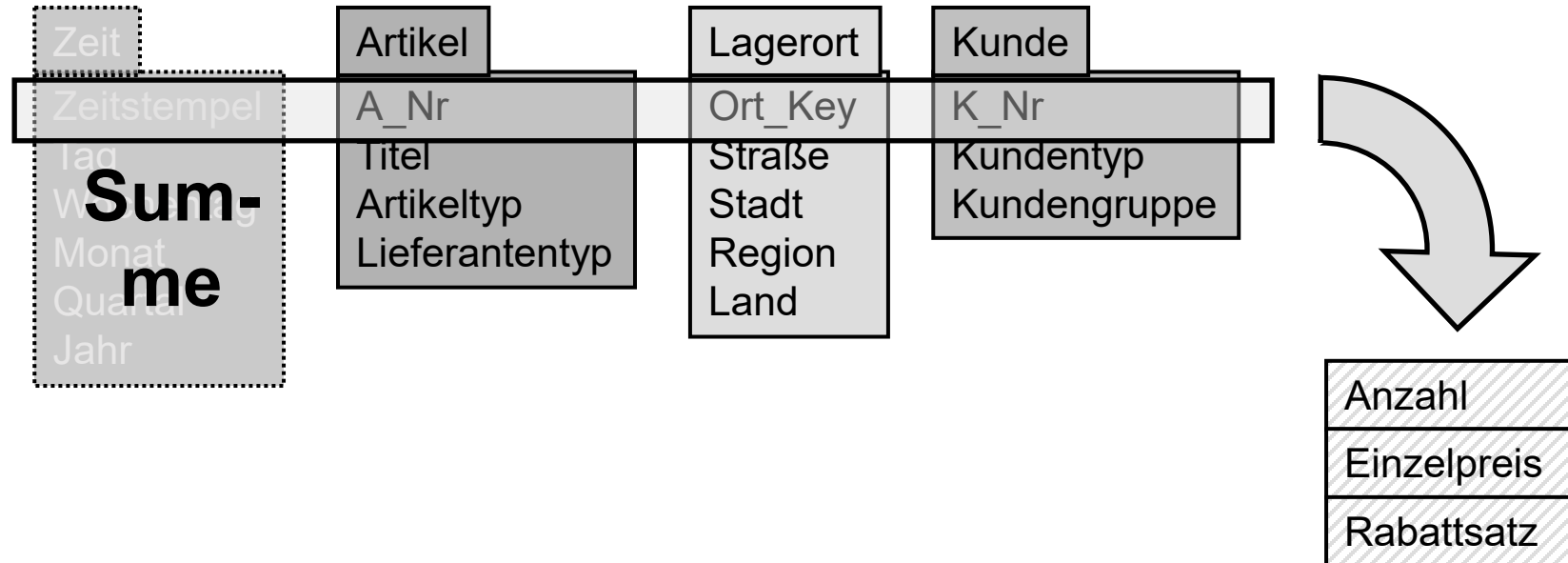
15.15 Alternative Darstellung eines Data Cubes



15.16 Elimination der Dimension Kunde



15.17 Elimination der Dimension Zeit



15.18 Beispiel einer Faktentabelle

<i>Sales</i>			
<i>Model</i>	<i>Year</i>	<i>Color</i>	<i>Sold</i>
Chevy	1990	red	5
Chevy	1990	white	87
Chevy	1990	blue	62
Chevy	1991	red	54
Chevy	1991	white	95
Chevy	1991	blue	49
Chevy	1992	red	31
Chevy	1992	white	54
Chevy	1992	blue	71
Ford	1990	red	64
Ford	1990	white	62
Ford	1990	blue	63
Ford	1991	red	52
Ford	1991	white	9
Ford	1991	blue	55
Ford	1992	red	27
Ford	1992	white	62
Ford	1992	blue	39

15.19 Ergebnis einer Cube-Anwendung

<i>Cube Table</i>			
<i>Model</i>	<i>Year</i>	<i>Color</i>	<i>Sales</i>
Chevy	1990	red	5
Chevy	1990	white	87
Chevy	1990	blue	62
Chevy	1990	All	154
Chevy	1991	red	54
Chevy	1991	white	95
Chevy	1991	blue	49
Chevy	1991	All	198
Chevy	1992	red	31
Chevy	1992	white	54
Chevy	1992	blue	71
Chevy	1992	All	156
Chevy	All	red	90
Chevy	All	white	236
Chevy	All	blue	182
Chevy	All	All	508
Ford	1990	red	64
Ford	1990	white	62
Ford	1990	blue	63
Ford	1990	All	189
Ford	1991	red	52
Ford	1991	white	9
Ford	1991	blue	55
Ford	1991	All	116

<i>Cube Table (Cont'd)</i>			
<i>Model</i>	<i>Year</i>	<i>Color</i>	<i>Sales</i>
Ford	1992	red	27
Ford	1992	white	62
Ford	1992	blue	39
Ford	1992	All	128
Ford	All	red	143
Ford	All	white	133
Ford	All	blue	157
Ford	All	All	433
All	1990	red	69
All	1990	white	149
All	1990	blue	125
All	1990	All	343
All	1991	red	106
All	1991	white	104
All	1991	blue	104
All	1991	All	314
All	1992	red	58
All	1992	white	116
All	1992	blue	110
All	1992	All	284
All	All	red	233
All	All	white	369
All	All	blue	339
All	All	All	941

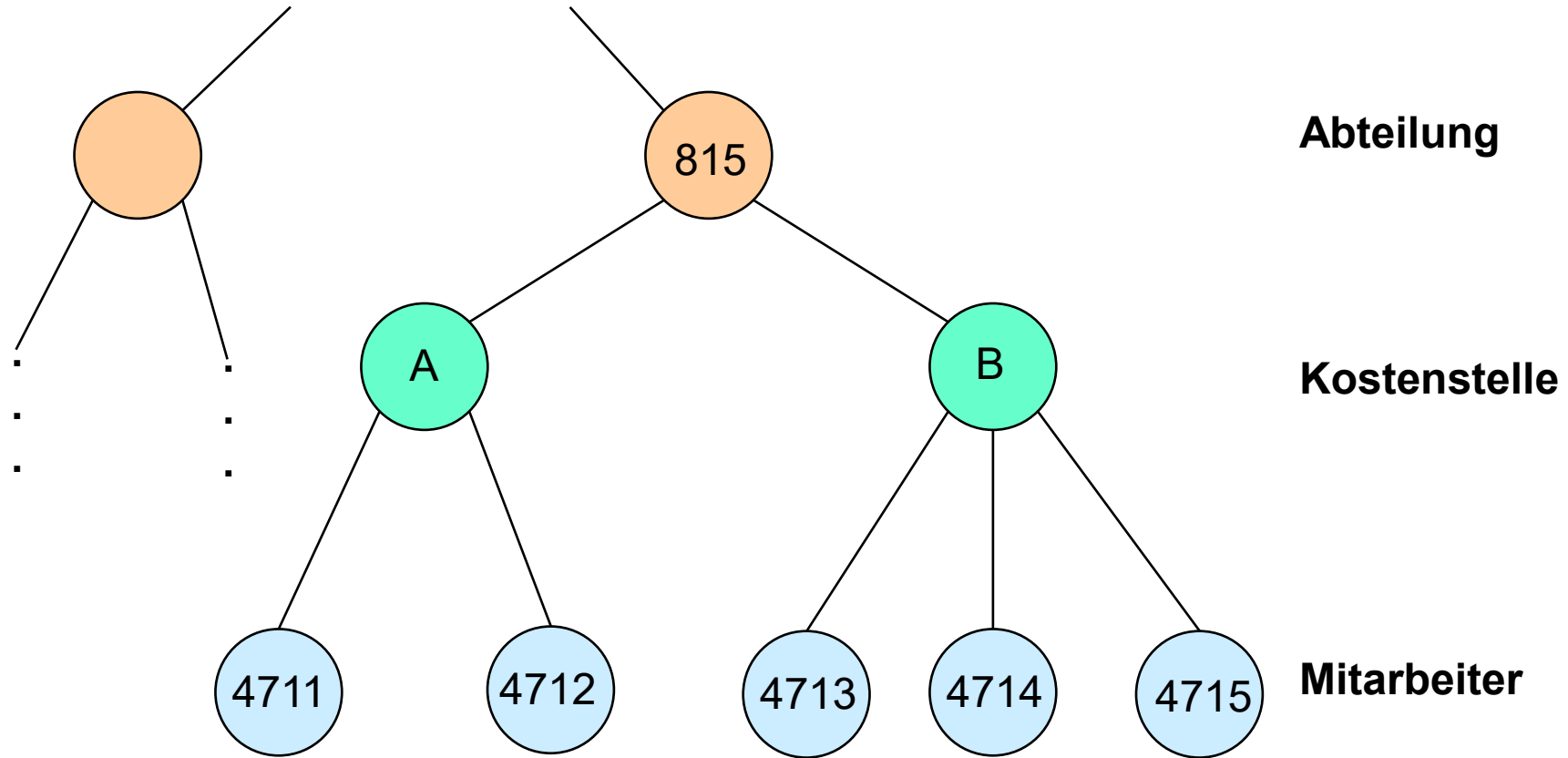
15.20 Beispiel einer Zensustabelle

<i>Name</i>	<i>Ort</i>	<i>Land- kreis</i>	<i>Bundes- staat</i>	<i>Geburts- datum</i>	<i>Ge- schlecht</i>	<i>Ein- kommen</i>
Joe	Miami	Dade	FL	8/20/55	M	32100
Chen	Miami	Dade	FL	6/05/57	M	40200
Bob	Hialeah	Dade	FL	3/21/57	M	33500
Karen	Hialeah	Dade	FL	8/23/55	F	43900
Jim	—	Dade	FL	10/24/56	M	29600
Joan	—	Dade	FL	11/15/56	F	36300
Dave	Orlando	Orange	FL	9/25/57	M	38000
Linda	Orlando	Orange	FL	5/13/55	F	46700
Jeff	Taft	Orange	FL	2/08/57	M	32600
Pat	Taft	Orange	FL	10/30/57	F	26500
Sam	Baytown	Harris	TX	3/02/55	M	28500
Bill	Baytown	Harris	TX	12/21/56	M	32800
Mary	Houston	Harris	TX	—	F	44700
Susan	Houston	Harris	TX	4/30/55	F	—
Alex	Houston	Harris	TX	7/11/57	M	30900
John	Austin	Travis	TX	1/06/56	M	38400
Fred	Austin	Travis	TX	10/25/56	M	42500
Anne	—	Travis	TX	8/17/55	F	34800

SLOWLY CHANGED DIMENSIONS

Worum geht es?

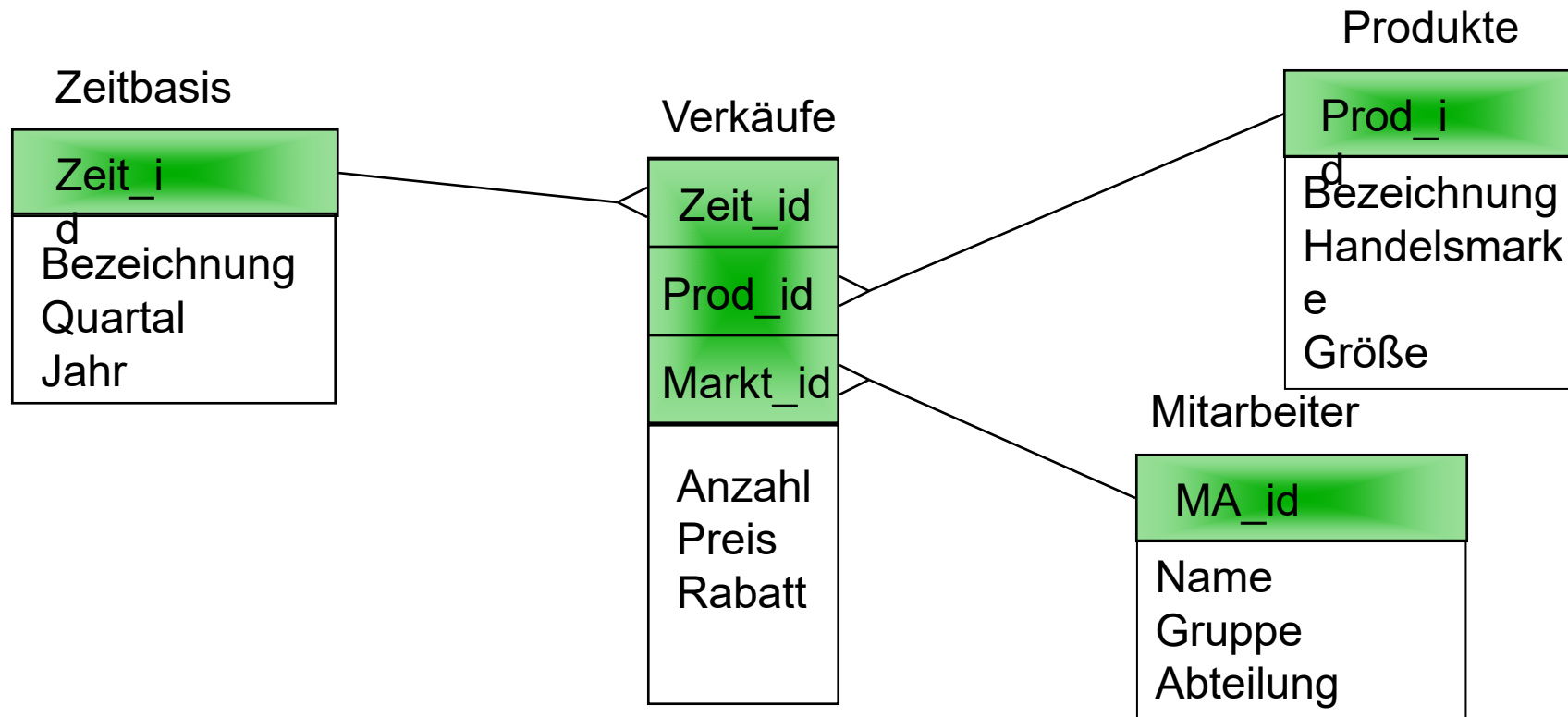
Hierarchien für Drill-Analysen



Worum geht es?

- **Beispiele für Hierarchien**
 - **Kostenstellen / Organigramm**
 - **Vertriebsregionen**
 - **Kundengruppen**
 - **Produktgruppen**
 - **Geografische Hierarchie**
- **Hierarchien strukturieren den (Stamm-) Datenbestand**
- **Hierarchien sind zentraler Bestandteil eines Data Warehouse**

Star Schema Modell

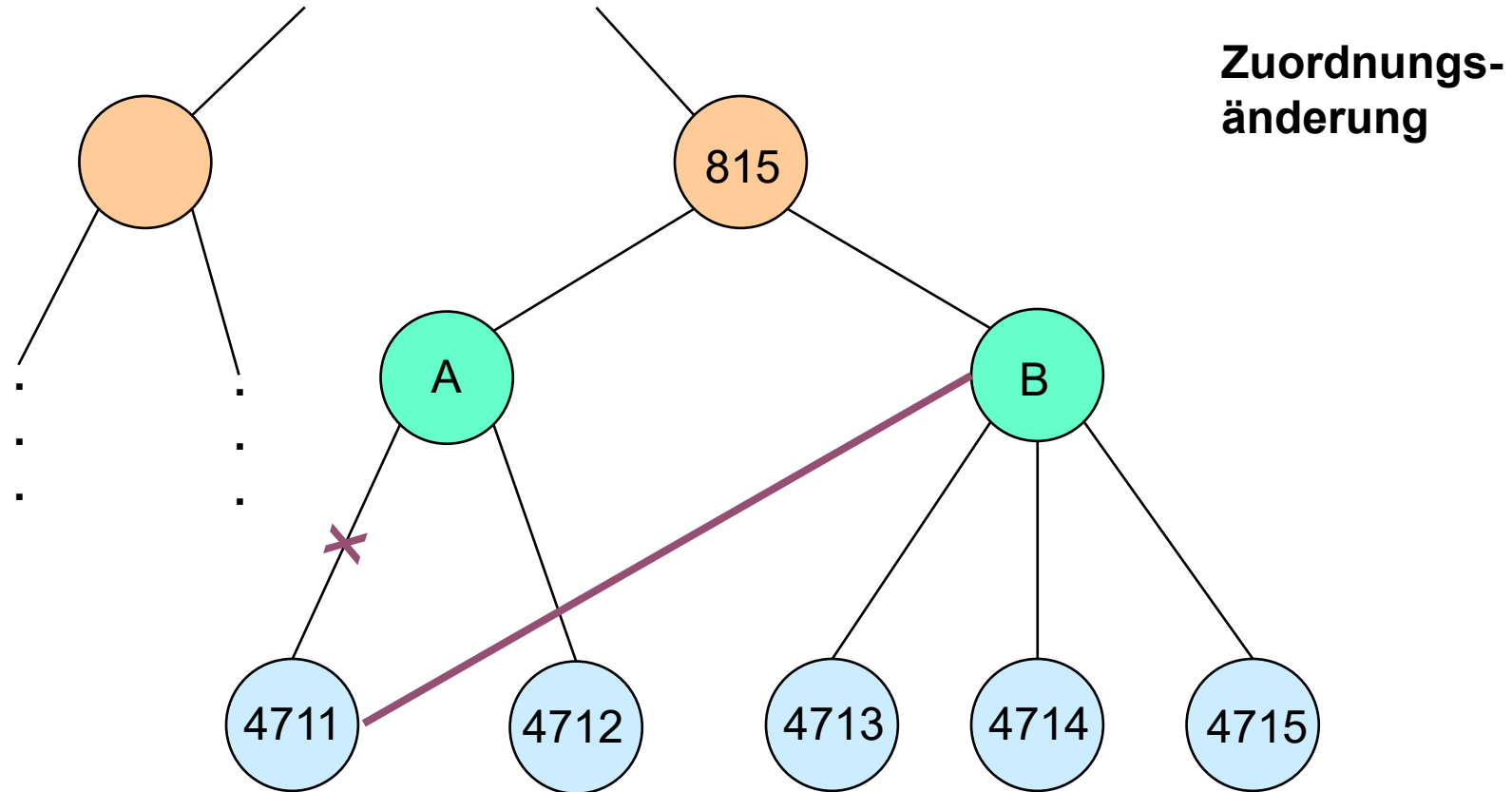


Worum geht es?

Speicherung in Dimensionstabelle

Mitarbeiter Nr.	Name	Eintrittsdatum	Gruppe	Abtlg.
4711	Meier	01.01.1980	A	815
4712	Müller	01.04.1994	A	815
4713	Schmidt	01.12.1992	B	815
4714	Becker	01.01.1997	B	815
4715	Schuster	01.03.2000	B	815

Was ist das Problem ?



Was ist das Problem ?

Auf Datensatzebene

Mitarbeiter Nr.	Name	Eintrittsdatum	Gruppe	Abtlg.
4711	Meier	01.01.1980	A	815
4712	Müller	01.04.1994	A	815
4713	Schmidt	01.12.1992	B	815
4714	Becker	01.01.1997	B	815
4715	Schuster	01.03.2000	B	815

Quelle
(4711, Meier, B,
815, 01.01.1980)

Schnittstelle:
(MA-Nr., Name, Kostenstelle,
Abtlg., Änd.datum)

Slowly Changing Dimensions

(nach Ralph Kimball)



- **Typ 1: Überschreiben**
- **Typ 2: Historisierung durch Versionierung**
 - **Künstliche Schlüssel, evtl. Versionsnummern**
 - **Keine Einschränkung auf Zeit notwendig**
 - **automatische Partitionierung der Vergangenheit**
- **Typ 3: Historisierung durch neues Attribut**
 - **nur Original und aktueller Wert**
 - **Partitionierung nur via Zeiteinschränkung**

Slowly Changing Dimensions



Problem:

- Historische Zuordnung für Auswertungen
- Aktuelle Zuordnung für Vergleichbarkeit

Typkritik:

- Typ 1: nur aktuell
- Typ 2: nur historisiert
- Typ 3: nur zwei Zustände

Lösung: SCD Typ 2 + 3






- **Künstlicher Primärschlüssel**
- **Auszeichnen des Attributs für Produktionsschlüssel**
- **Festlegen der zu historisierenden Attribute**
- **Versionierung bei Zuordnungsänderung**
- **Gültigkeitszeitraum**
- **jeweils Zusatzattribut „Aktuelle Zuordnung“**

Tabellenstruktur

DIM_MITARBEITER



#	*	789	MA_ID
	*	A	MA_NUMMER
	*	A	NAME
	*		GUELTIG_SEIT
	○		EINTRITTSDATUM
	○	A	GRUPPE
	○	A	GRUPPE_AKTUELL
	○	A	ABTEILUNG
	○		GUELTIG_BIS

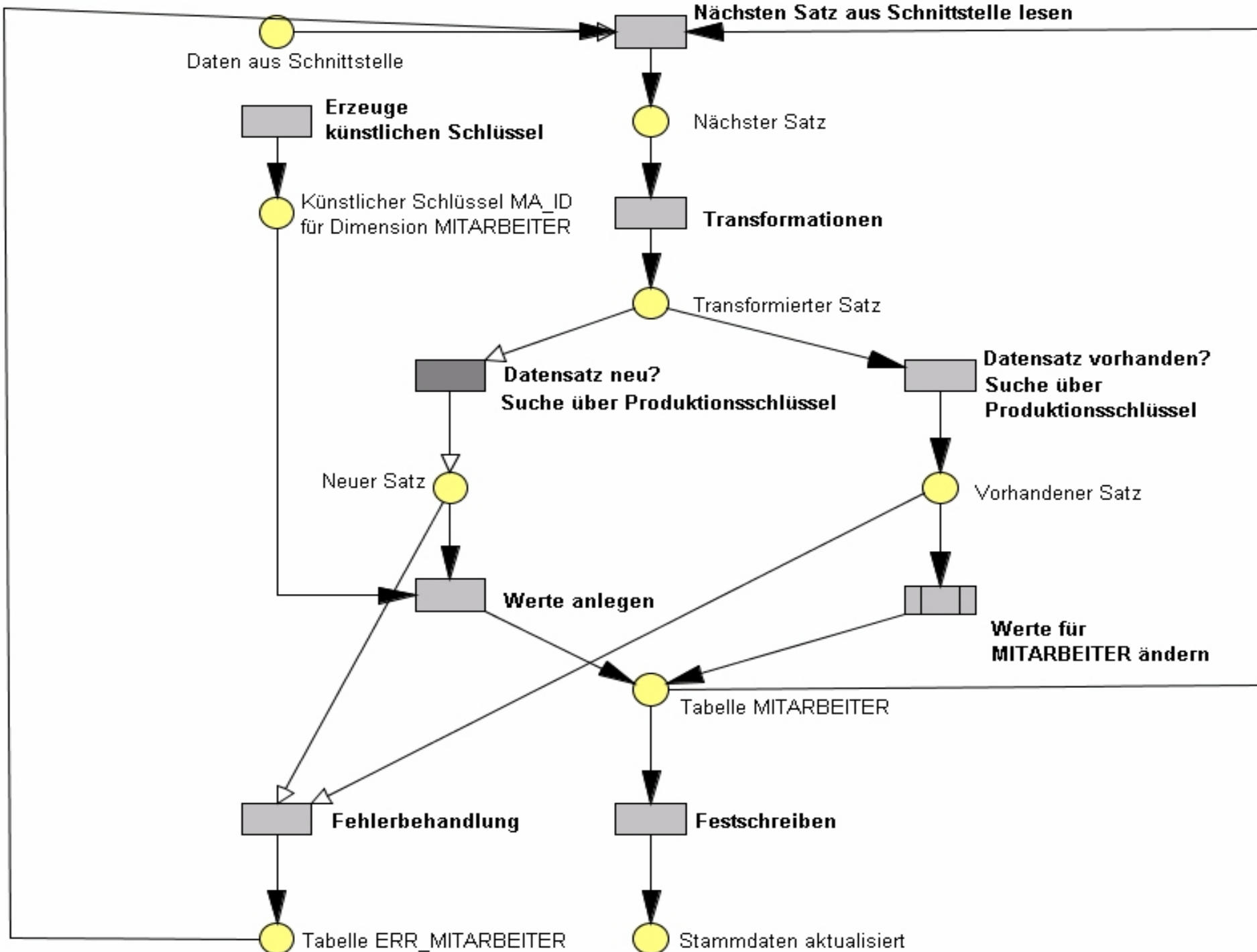
Vorgehen

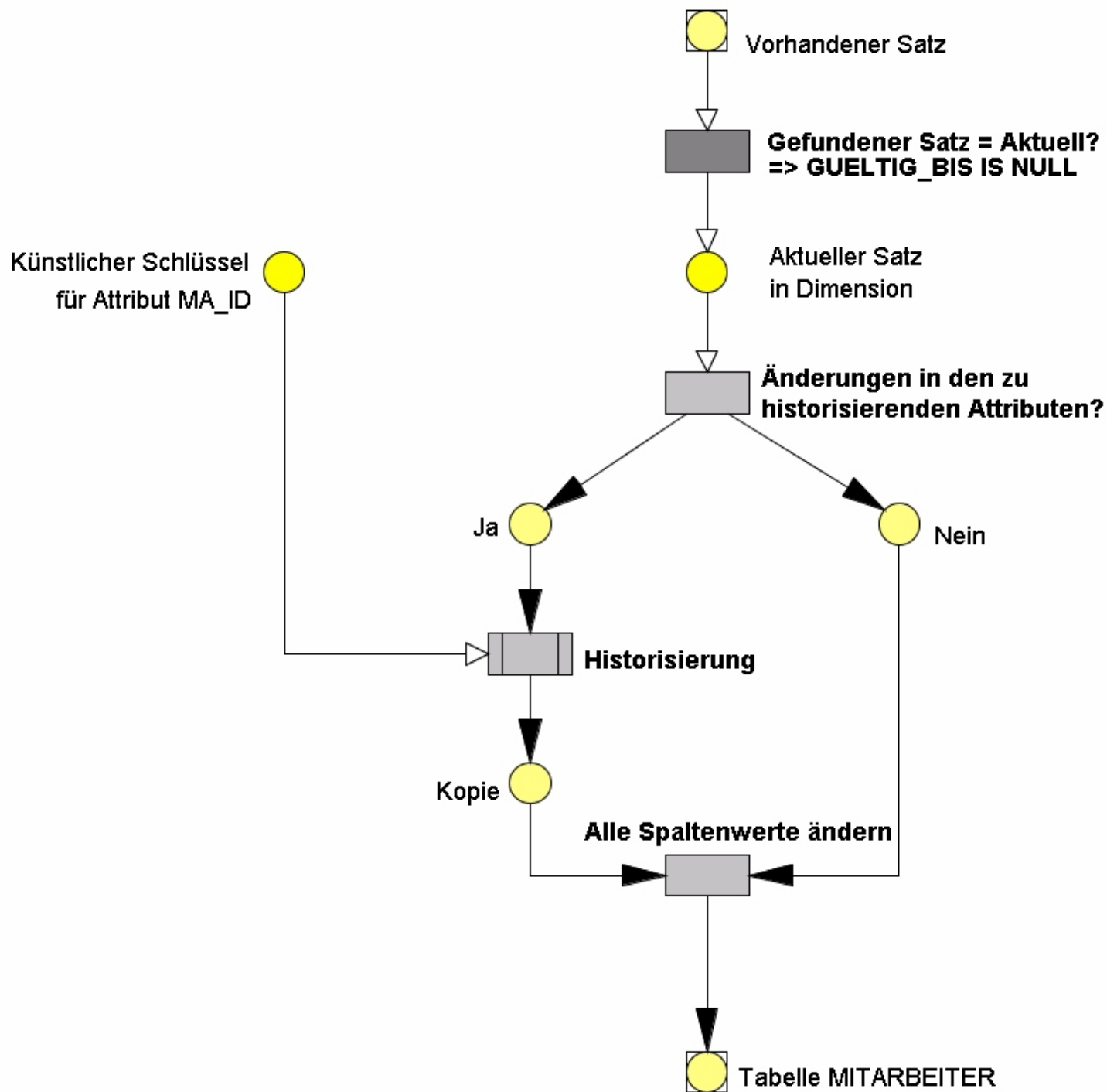
Schnittstelle S:

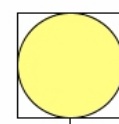
$S = \{ \vec{s} \mid \vec{s} = (\text{MA_NUMMER}, \text{NAME}, \text{GRUPPE}, \text{ABTLG}, \text{AENDDATUM}) \}$

Dimension D:

$D = \{ \vec{d} \mid \vec{d} = (\text{MA_ID}, \text{MA_NUMMER}, \text{NAME}, \text{GRUPPE}, \text{GRUPPE_AKT}, \text{ABTEILUNG}, \text{GUELTIG_SEIT}, \text{GUELTIG_BIS}) \}$



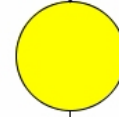




Ursprünglicher Datensatz



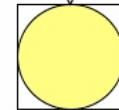
Neuen Datensatz anlegen
Neue künstliche ID;
GUELTIG_VON setzen



Neuer Datensatz angelegt



Vorhandenen Datensatz
GUELTIG_BIS setzen



Kopie

Auswertungen

- Zwei Auswertepfade:
 - aktuelle Zuordnung
 - historische Zuordnung
- Erhöhte Komplexität auf Metaebene
- Erhöhte Anzahl von Auswertungen
- Eindeutige Auswertungen

Zusammenfassung



- Anforderungen der Anwender nicht immer eindeutig
- Sowohl historische als auch aktuelle Zuordnung notwendig
- Auswertepformance wichtig
- Programmierung standardisierbar