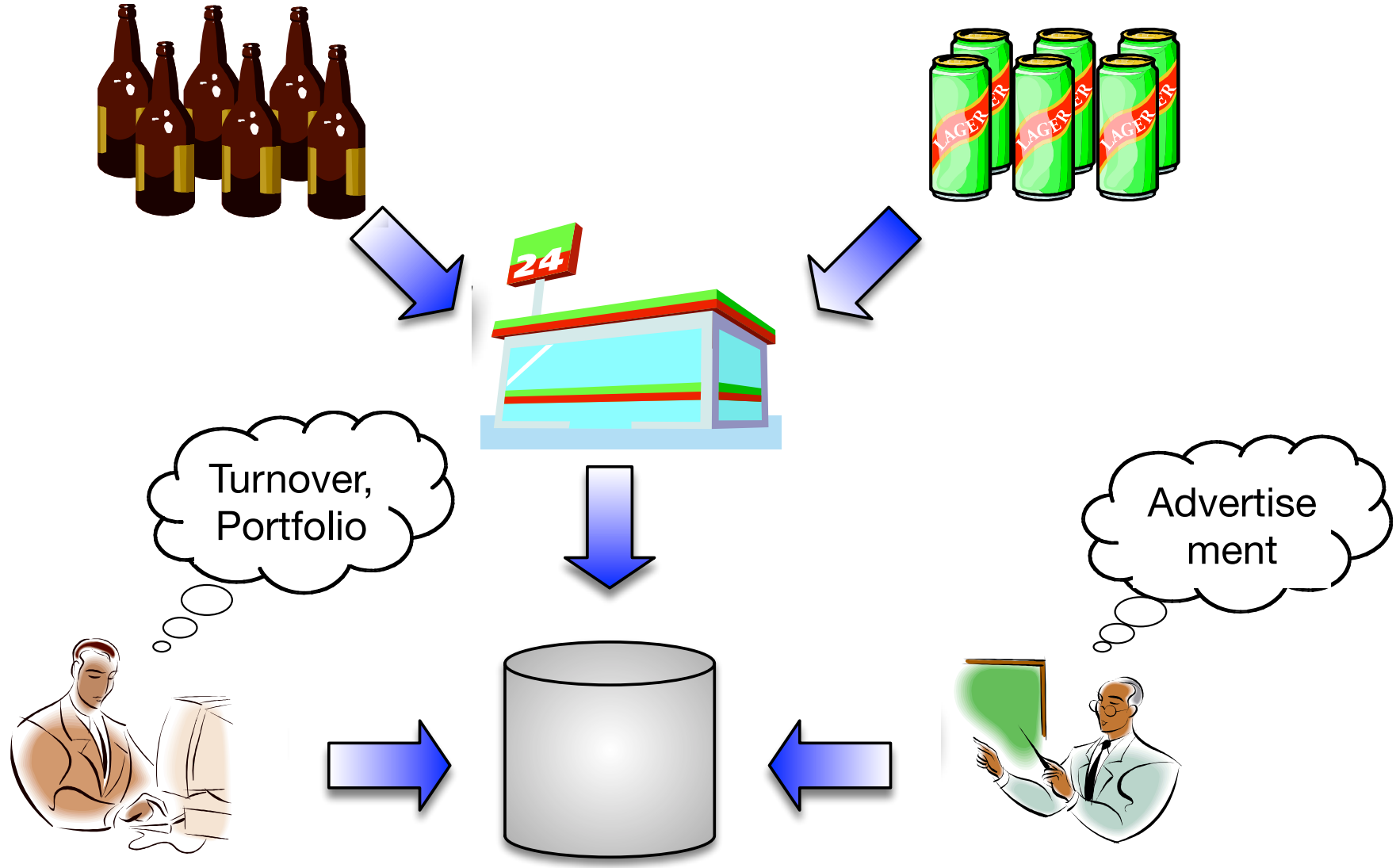
Part I

Introduction

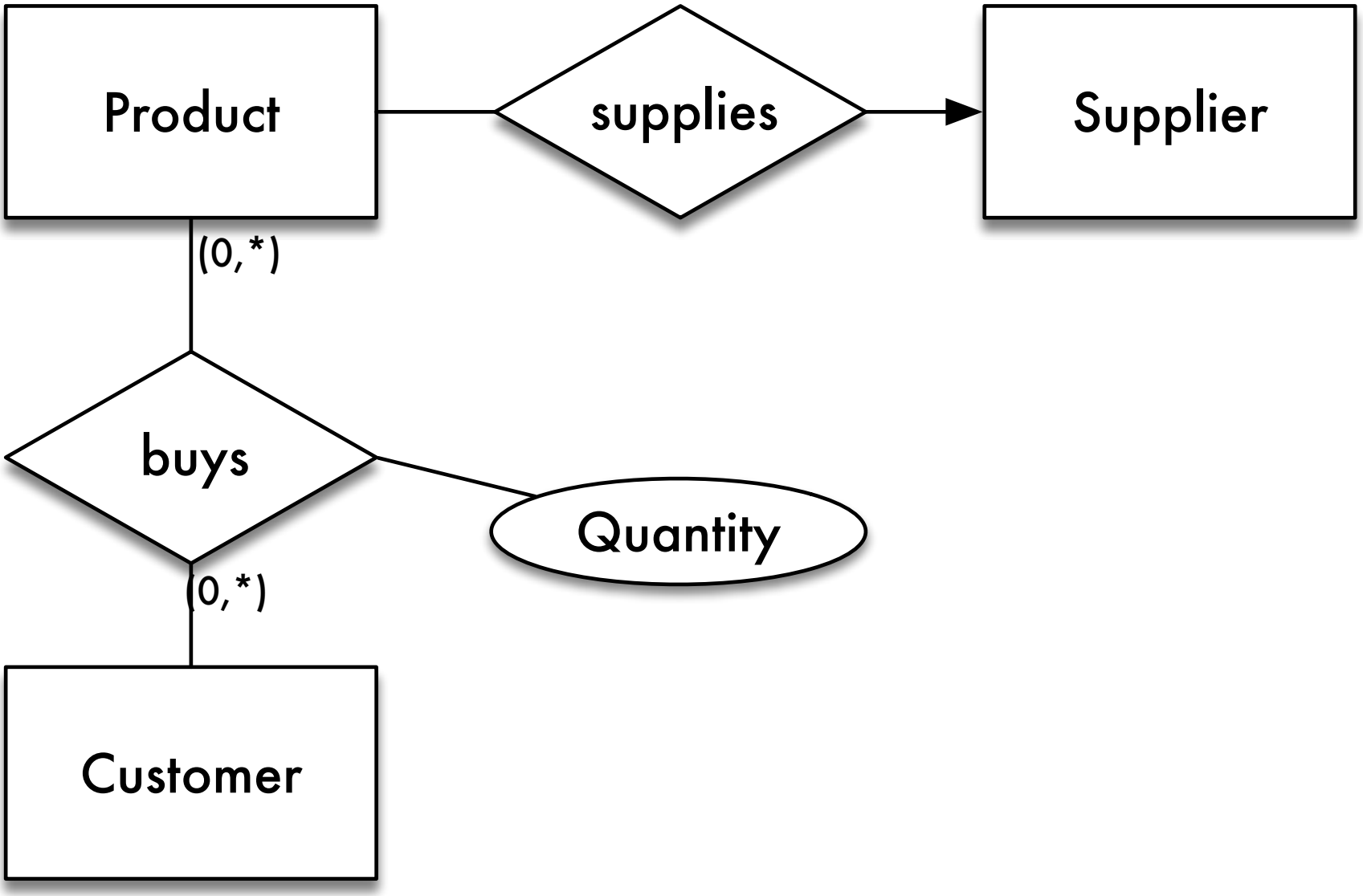
Introduction & Basic Terms

Motivation

Scenario: Beverage Market



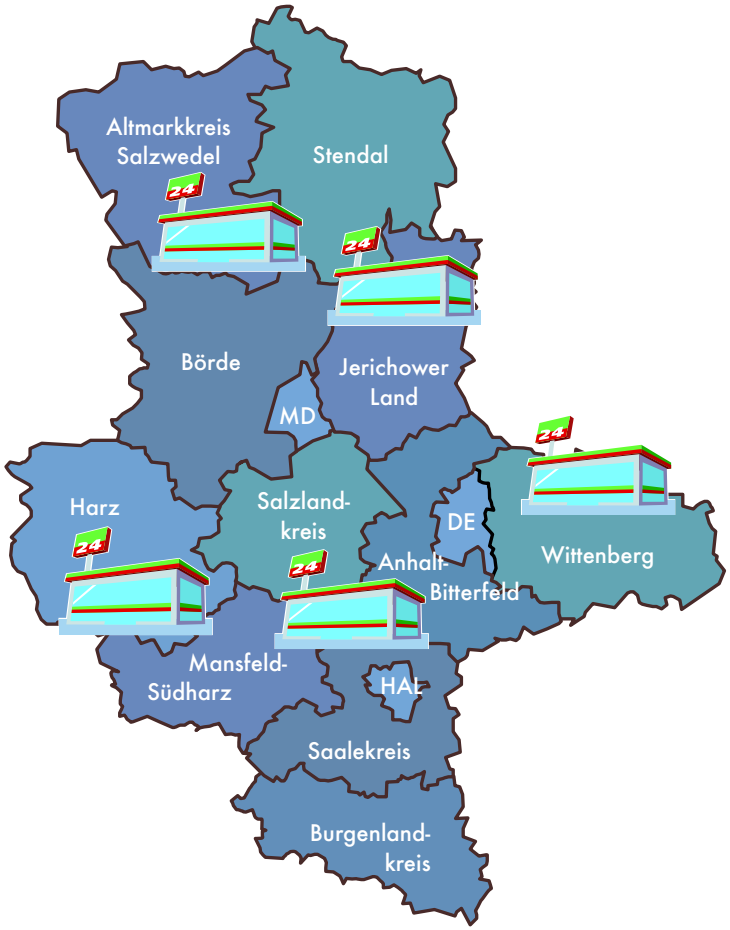
DB-Schema



DB-Usage

- Queries:
  - How many bottles of coke were sold in the last months?
  - How did the sales of red wine develop in the last year?
  - Who are our top clients?
  - From which supplier do we source most of the boxes?
- Problems
  - Use of external sources (customer database, supplier database, ...)
  - Data with temporal reference

Advanced Scenario



Saxony-Anhalt

Thuringia

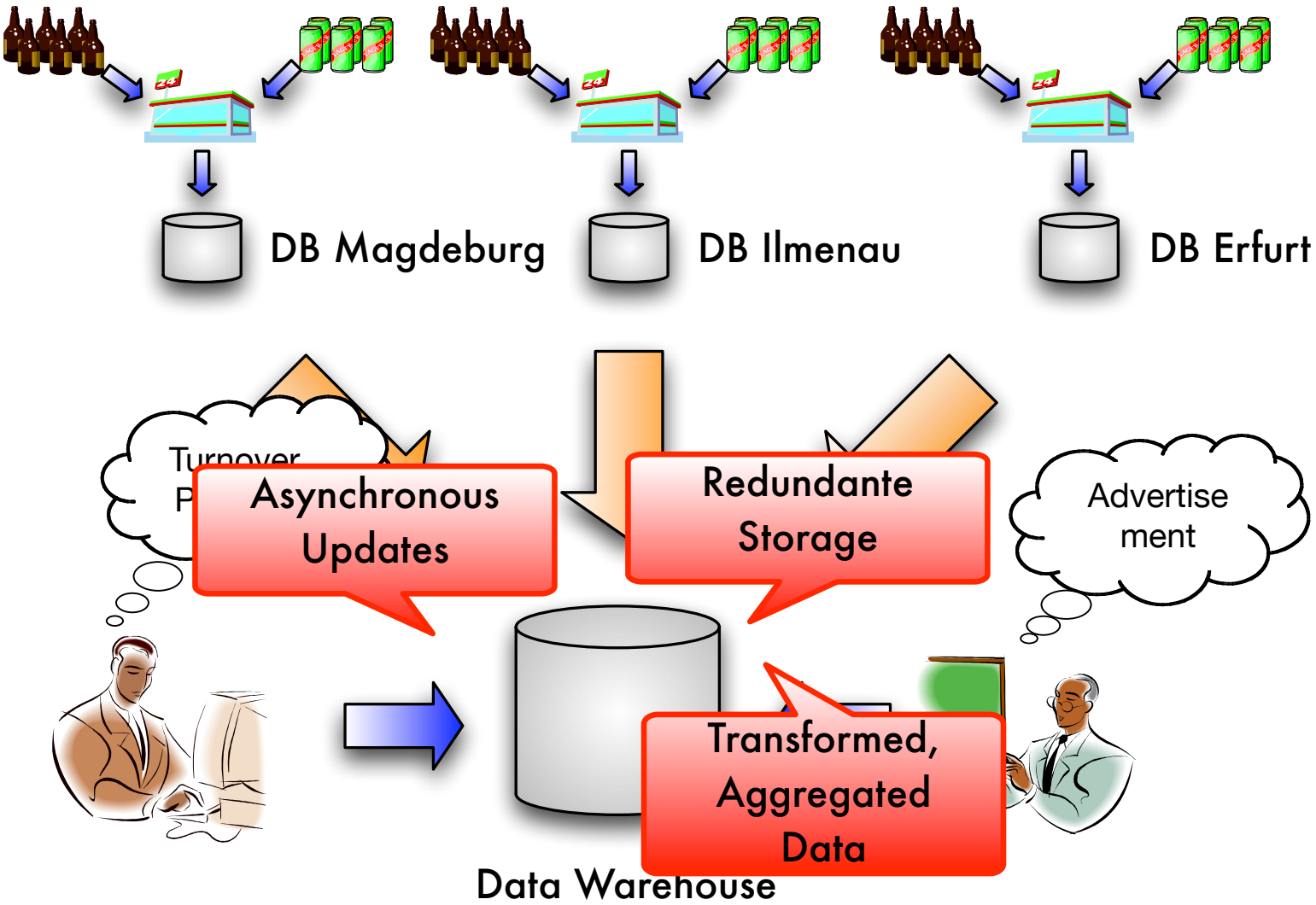
DB-Usage (2)

- Queries
  - Do we sell more beer in Ilmenau than in Erfurt?
  - How much coke was sold in Thuringia in summer?
  - More than water?
- Problem
  - Queries over multiple databases

Solutions

- Variant 1: "Distributed DB"
  - Global query over multiple DBs $\rightarrow$ View with Union
  - Disadvantage: expensive distributed query execution
- Variant 2: "Central DB"'
  - Changes on a central DB
  - Disadvantage: long response times in productive settings

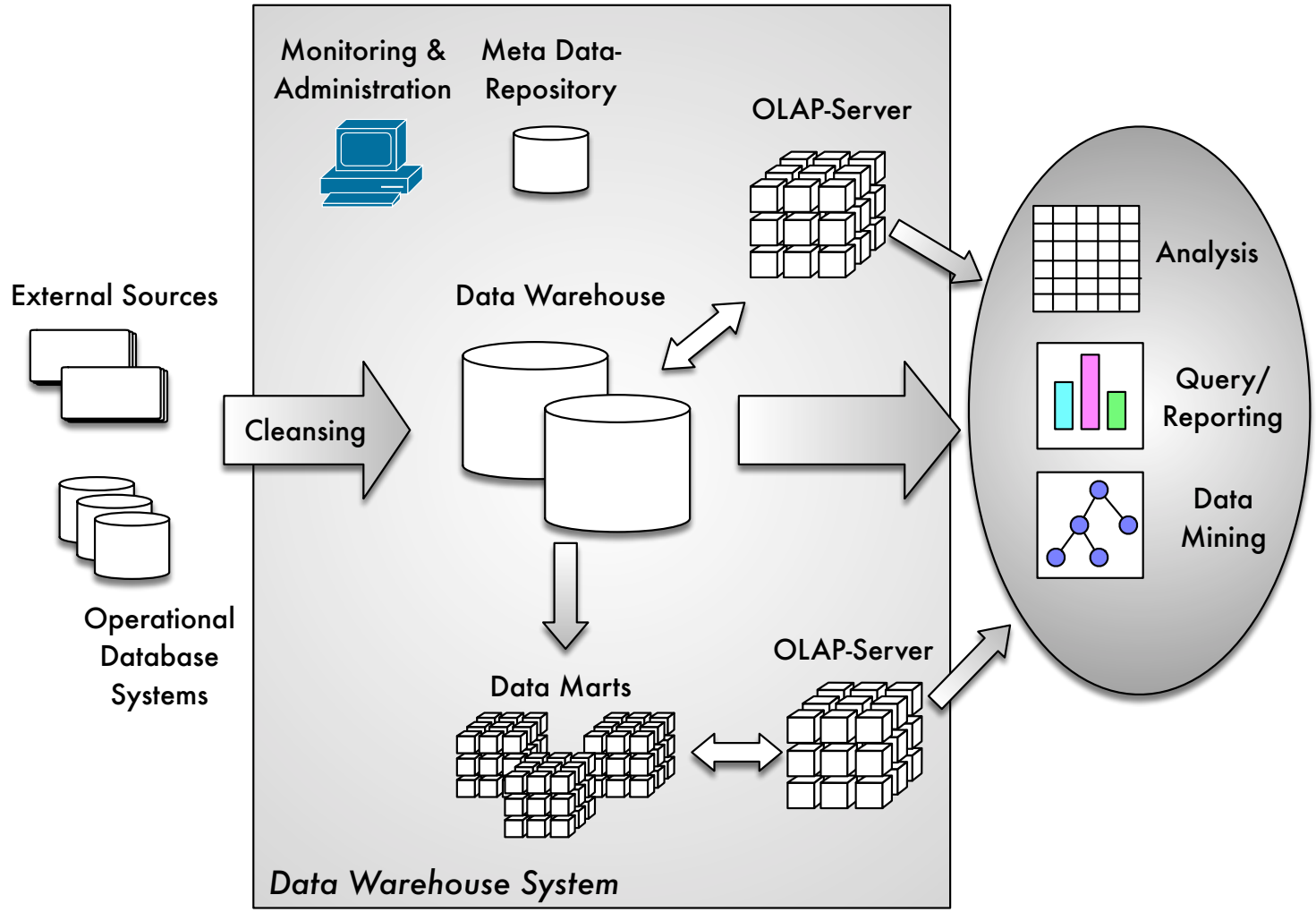
Data Warehouse-Solution



Subject of the Lecture

- Data Warehouse: Collection of data and technologies to support decision making
- Challenge for database technologies
  - Data volume (efficient storage and management, query processing)
  - Data modeling (Time reference, multiple dimensions)
  - Integration of heterogeneous data sources
- Focus
  - Database techniques of Data Warehouses

Overview



[nach Chaudhuri&Dayal 1997]

Applications

Business applications

- Information provisioning
  - Data and information as a basis for decisions (e.g., metrics)
  - Influence on a future operating result and on the handling of business processes
  - User: Manager, head of department, professionals
  - Types of provisioning:
    - Query Approaches: free definable queries and reports (individual solution strategy)
    - Reporting: Access to pre-defined reports (fixed range of solutions)
    - Editorially prepared, personalized information
    - Domain-specific data views
    - Pre-calculated metrics (e.g., by Data Mining Algorithms)

Business applications (2)

- Analyse
  - Detailed Analysis of the data to detect changes
  - Scenario techniques (What-If-Analyzes)
  - Users: Specialists (e.g., Controlling, Marketing)
- Planning
  - Support through explorative data analysis
  - Aggregation of individual plans
  - Forecasting methods (e.g., statistical seasonal models)
- Campaign Management
  - Support of strategic campaigns
  - Customer analysis, portfolio- and risk analysis

Scientific and Technical Applications

- Scientific Applications
  - Statistical and Scientific Databases $\rightarrow$ technical roots of the DW
  - Example: Projekt Earth Observing System (Climate- and Environmental Research)
    - Approx. 1,9 TB meteorological data
    - Preprocessing and Analysis (statistical, Data Mining)
- Technical Applications
  - Public sector: DW with environmental or geopgraphical data (e.g., water analyzes)

Example Use Case

- Wal-Mart (www.wal-mart.com)
- Market leader in American retail
- Business-wide Data Warehouse
  - Size: approx. 300 TB (2003), 480 TB (2004), today: approx. 12 PB
  - Daily around 25.000 DW-Queries
  - High level of detail (daily evaluation of article sales, warehouse stock, customer behavior)
  - Basis for customer basket analysis, customer classification . . .

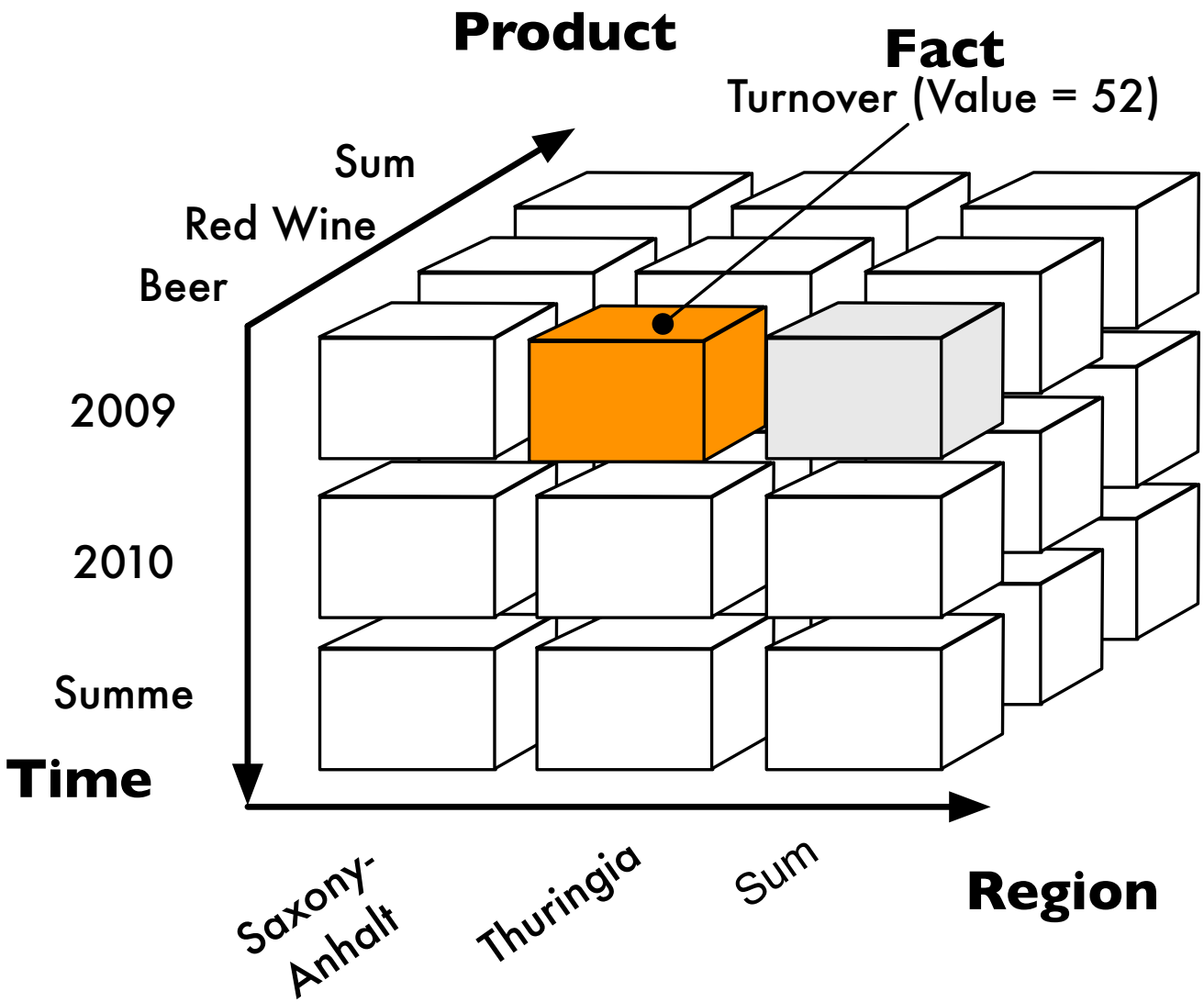
Questions and Tasks (Example)

- Checking the assortment of goods to identify slow sellers or bestsellers
- Location analysis to estimate the profitability of branch offices
- Investigation and prognosis of marketing actions
- Evaluation of customer surveys, complaints regarding certain products etc.
- analysis of the stock
- Shopping cart analysis with the help of cash register data (financial transactions)

Example of a Query

Which sales have been made in the years 2009 and 2010 in the product segments beer and red wine in the federal states Saxony-Anhalt und Thuringia?

Result (Cube)


Product
Fact
Turnover (Value = 52)
Sum
Red Wine
Beer
2009
2010
Summe
Time
Saxony-Anhalt
Thuringia
Sum
Region

Result (2-dim. Cube Illustration)

Sales		Beer	Red Wine	**Sum**
2009	Saxony-Anhalt	45	32	77
	Thuringia	52	21	73
	Sum	97	53	150
2010	Saxony-Anhalt	60	37	97
	Thuringia	58	20	78
	Sum	118	57	175

Distinction

Aspects of Data Warehouses

- Integration
  - Unification of data from different, mostly heterogeneous sources
  - Overcoming heterogeneity on different levels (system, schema, data)
- Analysis
  - Provision of data in a form desired by the user (related to decision area)
  - Requires preselection, time reference, aggregation

Short Transaction (OLTP)

Customer					
ID	name	first name	postal code	city	street
4711	Saake	Gunter	01234	Somewhere	On the Hill 3
42	Sattler	K.	12345	Here	Route 18
0800	Köppen	Veit	60701	There	Pathway 9A

```
SELECT first name, last name
FROM Customer
WHERE id = 0800
```

Result

First Name	Last Name
Veit	Köppen

Long-running Transaction (OLAP)

```
SELECT DISTINCT ROW  Time.Dimension AS Year,
                     Product.dimension AS Articles,
                     AVG(Fact.Turnover) AS Average turnover,
                     Place.Dimension AS Sales area

FROM (Product group INNER JOIN Product ON Product group.
      [Group No] = Product.[Group ID]) INNER JOIN
      ((((Product INNER JOIN [Fact.Turnover] ON Product.[Item no.]
      = [Fact.Turnover].[Article no.])  INNER JOIN Order ON
      [Fact.Turnover].[Order-No]= Order.[Order-ID]) INNER JOIN
      Time.Dimension ON Orders.[Order-ID] =
      Time.Dimension [Order-ID]) INNER JOIN Place.Dimension ON
      Order.[Order-ID] = Place.Dimension.[Order-ID]) ON
      Product group.[Group No.]  = Product.[Group ID]

GROUP BY Product.Dimension.Group.Name, Place.Dimension.State,
         Time.Dimension.Year;
```

Distinction to OLTP

- Classical operational information systems
  → Online Transactional Processing (OLTP)
  - Data collection and management
  - Processing under responsibility of the respective department
  - Transactional processing: short read/write accesses to few data records
- Data Warehouse
  → Online Analytical Processing (OLAP)
  - Analysis in the center
  - Long-running read transactions on many data sets
  - Integration, consolidation and aggregation of data

Distinction to OLTP: Queries

	OLTP	**OLAP**
Focus	read, write, modify, delete	read, periodic insert
Transaction duration and type	short read/write transactions	long-lasting read transactions
Query structure	simple structured	complex
Volume of a query	few records	many records
Data model	query flexible	analysis-oriented

Distinction to OLTP: Data

	OLTP	**OLAP**
Data sources	usually one	more
Properties	non-derivative, current, autonomous, dynamic	derived / consolidated, historicized, integrated, stable
Data volume	MByte . . . GByte	GByte . . . TByte . . . PByte
Accesses	single tuple access	table access (column by column)

Distinction to OLTP: Users

	OLTP	**OLAP**
User type	Input/Output by employee or application software	Manager, Controller, Analyst
User number	many	few (up to a few hundred)
Response time	msecs . . . secs	secs . . . min

Definition: DBMS Techniques

- Parallel Databases
  - Technique for the implementation of a DWH
- Distributed databases
  - Usually no redundant data management
  - Distribution as a means of load distribution
  - No content integration/consolidation of data
- Federated Databases
  - Greater autonomy and heterogeneity
  - No specific analytical purpose
  - No read access optimization

Term: Data Warehouse

Data Warehouse: Definition

A Data Warehouse is a **subject-oriented**, **integrated**, **non-volatile**, and **time variant** collection of data in support of managements decisions.

(W.H. Inmon 1996)

Data Warehouse: Characteristics

- Subject-oriented:
  - Purpose is to support cross-divisional evaluation possibilities for different domains
  - Centralized provision of data on business objects (topics)
- Integrated database:
  - Processing of data from several different (internal and external) data sources (e.g., operational DB or web)
- Non-volatile database:
  - Stable, persistent database
  - Data in the DW are generally no longer removed or changed
- Time-related data (time-variant):
  - Comparison of data over time possible (time series analysis)
  - Storage over a longer period of time

Further Terms

- Data Warehousing
  - Data Warehouse process, i.e., all steps of data retrieval (extraction, transformation, loading), saving and analysis
- Data Mart
  - External (partial) view of the data warehouse
  - By copying
  - Application-specific
- OLAP (Online Analytical Processing)
  - Explorative, interactive analysis based on the conceptual data model
- Business Intelligence
  - Data Warehousing + Reporting + Analysis (OLAP, Data Mining); also automatically generated reports in companies

Division of operative and analytical systems: Reasons

- Response time behavior: Analysis on operational source data systems $\rightarrow$ bad performance
- Historization of company data
  Long term storage of data $\rightarrow$ Time series analysis
- access to data independent of operative data sources (availability, integration problems)
- Unification of the data format in DW
- Guarantee of the data quality in DW

History: Roots

- 60s: Executive Information Systems (EIS)
  - Qualitative information supply for decision makers
  - Small, condensed extracts of the operative data stock
  - Preparation in the form of static reports
  - Mainframe
- 80s: Management Information Systems (MIS)
  - Mostly static report generators
  - Introduction of hierarchical levels for evaluation of key figures (Roll-Up, Drill-Down)
  - Client-server architectures, GUI (Windows, Apple)

History

- 1992: Introduction of the data warehouse concept by W.H. Inmon
  - Redundant storage of data, detached from source systems
  - Limitation of the data to analysis purposes
- 1993: Definition of the term OLAP by E.F. Codd
  - Dynamic, multidimensional analysis
- Other areas of influence
  - Dissemination of business process oriented transaction systems (SAP R/3) $\rightarrow$ Provision of decision relevant Information
  - Data Mining
  - WWW (Web-enabled Data Warehouse etc.)

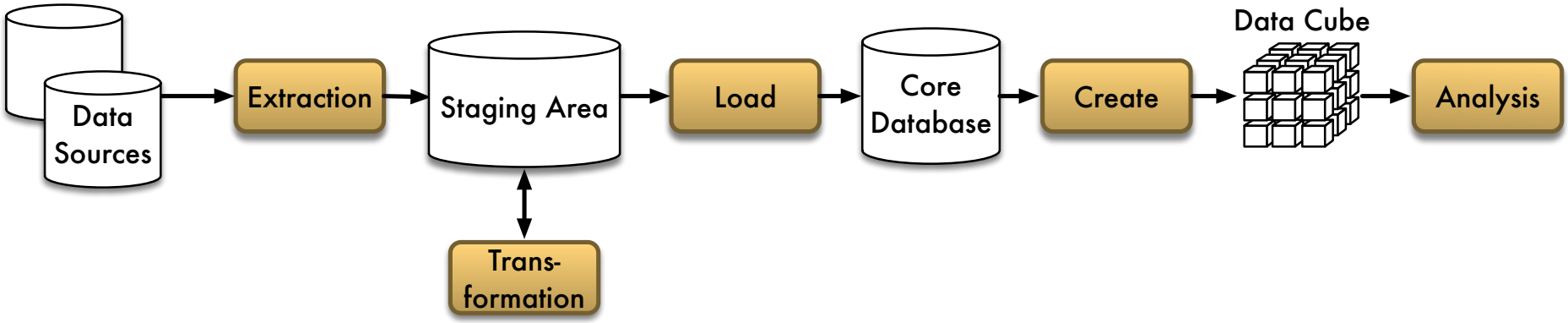
Topics

Lecture: Objectives

- Transfer of knowledge about database techniques for building and implementing data warehouses
- Application of known DB techniques (see lecture "Database systems")
  - Data Modeling
  - Query languages and processing
- DW-specific techniques
  - Multidimensional data modeling
  - Special querying techniques
  - Index structures
  - Materialized views
  - Fields of application: business intelligence

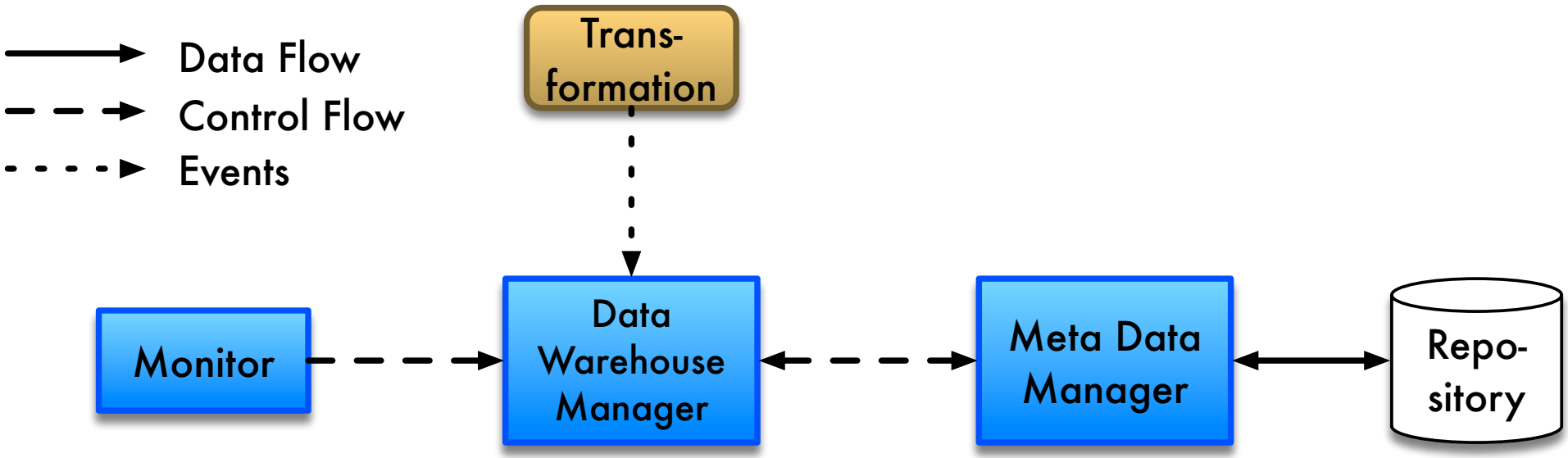
DW Architecture

- Components of DW and their tasks
- Databases
  - Data sources: Origin of the data
  - Staging Area: temporary database for transformation
  - Data Warehouse: physical database for analysis
  - Repository: Database with metadata



DW Architecture: Components

- Data-Warehouse-Manager: central control and management
- Monitors: Monitoring sources for changes
- Extractors: Selection and transport of data from sources in Data cleansing area
- Transformers: Standardization and data cleansing
- Loading components: Loading the transformed data into the DW
- Analysis components: Analysis and presentation of data



Multidimensional Data Model

- Data model to support the analysis
  - Facts and dimensions
  - Classification scheme
  - Cube
- Operations: Pivoting, Roll-Up, Drill-Down, Drill-Across, Slice and Dice
- Notations for conceptual modeling
- Relational implementation
  - Star Scheme, Snowflake Scheme

ETL Process

- Process of extraction, transformation and loading
- Extraction of data from sources:
  - Operational databases,
  - Web,
  - Files, etc.
- Loading data into the DWH
- Aspects of data quality
  - Term
  - Problems
  - Data Cleaning

Index and Memory Structures

- Classification
- Repetition
  - B-tree and B+-tree
- Multidimensional index structures
  - R-tree
  - UB-Tree
  - Bitmap Index
  - Comparison
- Other forms
- Multidimensional Storage

Queries to Data Warehouses

- Grouping and Aggregation
- Supergroups, CUBE
- OLAP functions from SQL:2003
- Multidimensional extensions of query languages: MDX

Query processing and optimization

- Calculation of grouping and cubes
- Star-Joins
- Further optimization aspects

Materialized Views

- Materialized view: in advance calculated section from a fact table
- Use: Request substitution
- Selection: Determination of the redundant data
  - Static vs. dynamic selection procedure
  - Semantic caching
- Maintenance and updating

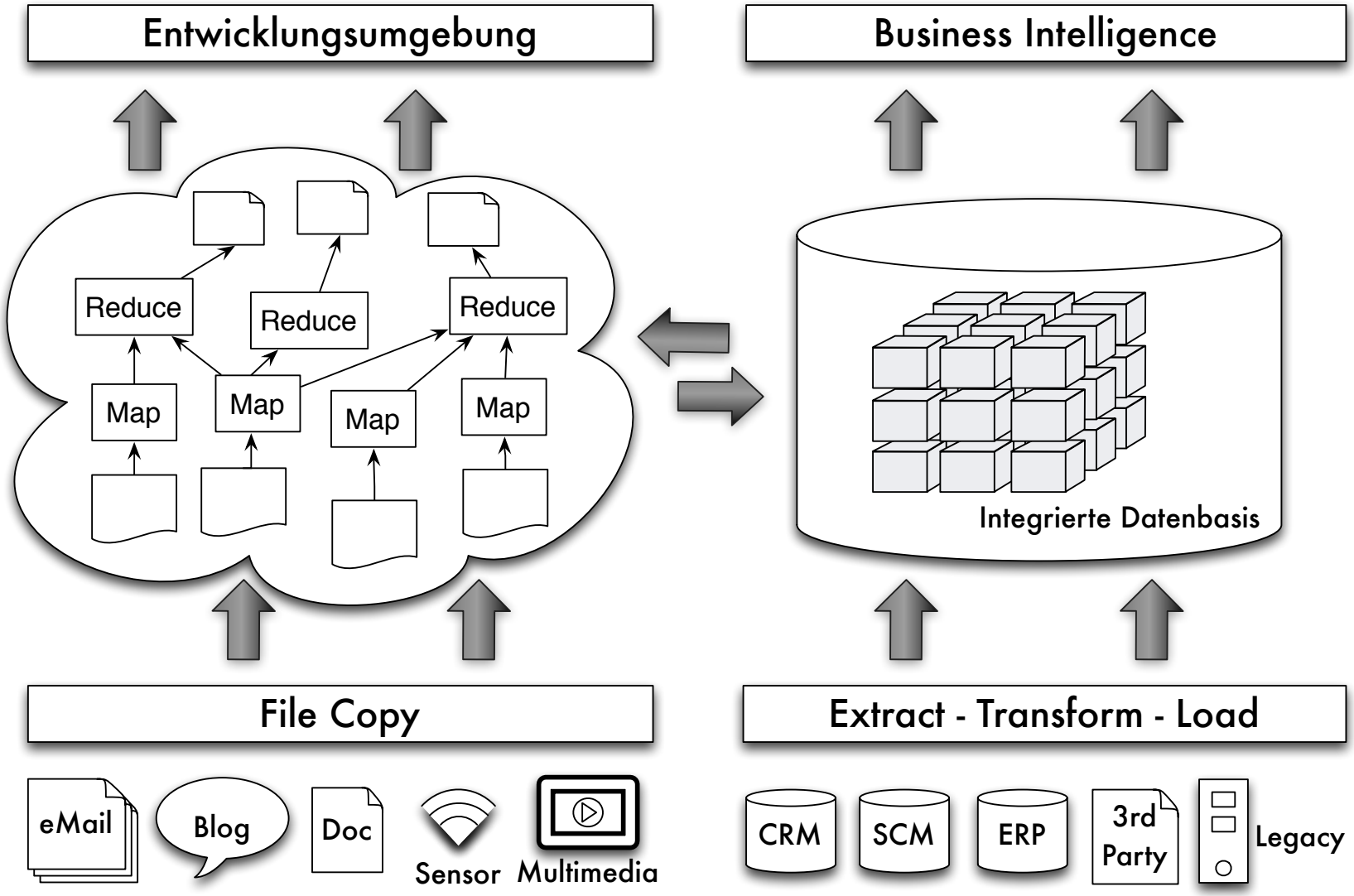
Applications for Data Warehouses

- Reporting
- Data Exploration
  - classification
  - shopping cart analysis
  - forecast
- Application scenarios

Big Data: 5 V's

- Volume - very high data volume (doubling every 2 years)
- Variety - structured and unstructured data
- Velocity - from batch to real-time
- Veracity - trust in data
- Value - value of (business) data

Big Data and Data Warehouse

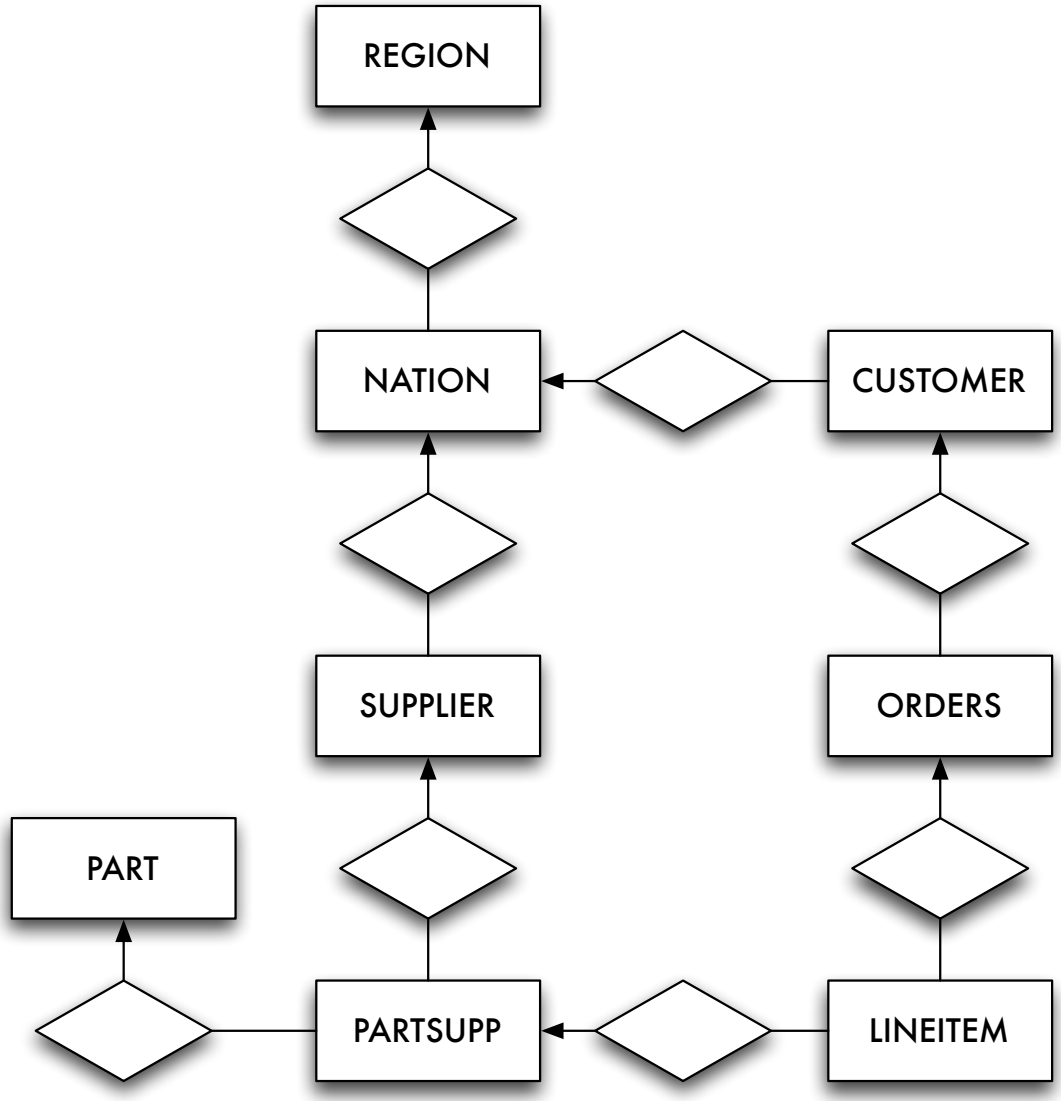


Benchmarks

TPC Benchmarks

- Comparison of the performance of databases (www.tpc.org)
  - TPC-C: OLTP Benchmark
  - TPC-H: Ad-hoc Decision Support (variable parts)
  - TPC-R: Reporting Decision Support (fixed requests)
  - TPC-W: eCommerce transaction processing
- Predefined schemas (supply chain)
- Schema, query and data generators
- Different DB sizes
  - TPC-H: 100 GB - 300 GB - 1 TB - 3 TB - 10 TB

TPC-H: Schema


REGION
NATION
CUSTOMER
SUPPLIER
ORDERS
PART
PARTSUPP
LINEITEM

TPC-H: Queries

```
SELECT c_name, c_custkey,
   o_orderkey, o_orderdate,
   o_total price, SUM (l_quantity)
FROM customer, orders, lineitem
WHERE o_orderkey IN (SELECT l_orderkey
      FROM lineitem
      GROUP BY l_orderkey
      HAVING SUM (l_quantity) > :1)
   AND c_custkey = o_custkey
   AND o_orderkey = l_orderkey
GROUP BY c_name, c_custkey, o_orderkey,
   o_orderdate, o_totalprice
ORDER BY o_totalprice desc, o_orderdate
```

TPC-H: Numbers (10,000 GB) - 2011

10,000 GB Results

Rank	Company	System	QphH	Price/QphH	Watts/KQphH	System Availability	Database	Operating System	Date Submitted	Cluster
1	DELL	Dell PowerEdge R710 using EXASolution 4.0	7,128,255	.53 USD	NR	10/01/11	EXASOL EXASolution 4.0	EXASOL EXACluster OS 4.0	04/05/11	Y
2	IBM	IBM System p 570	343,551	32.89 USD	NR	04/15/08	IBM DB2 Warehouse 9.5	IBM AIX 5L V5.3	10/15/07	Y
3	HP invent	HP Integrity Superdome/Dual-Core Itanium/1.6 GHz	208,457	27.97 USD	NR	09/10/08	Oracle Database 11g Enterprise Edition	HP-UX 11.i v3 64 bit	03/10/08	N
4	IBM	IBM System p5 575 with DB2 UDB 8.2	180,108	47.00 USD	NR	08/30/06	IBM DB2 UDB 8.2	IBM AIX 5L V5.3	07/14/06	Y
5	HP invent	HP Integrity Superdome-DC Itanium2/1.6GHz /64p/128c	171,380	32.91 USD	NR	04/01/07	Oracle Database 10g R2 Enterprise Edt w/Partitioning	HP-UX 11i v3 64 bit	11/30/06	N
6	HP invent	HP Integrity Superdome - Itanium2/1.5 GHz-128p/128	86,282	161.24 USD	NR	04/06/05	Oracle Database 10g Enterprise Edition	HP UX 11.i V2 64 bit	10/07/04	Y
7	UNISYS	Unisys ES7000 Model 7600R Enterprise Server(16s)	80,172	18.95 USD	NR	02/17/09	Microsoft SQL Server 2008 Enterprise x64 Edition	Microsoft Windows Server 2008 Datacenter x64 Edition	02/17/09	N
8	HP invent	HP Integrity Superdome	63,650	38.54 USD	NR	08/30/08	Microsoft SQL Server 2008 Enterprise Edition	Microsoft Windows Server 2008 Intanium based Systems	02/27/08	N
9	HP invent	HP Integrity Superdome - Itanium2/1.5 GHz-64p/64c	49,104	118.13 USD	NR	03/25/04	Oracle Database 10g Enterprise Edition	HP-UX 11.i 64-bit Base OS	01/05/04	N

Products

- OLAP-Tools/Server
  - MS Analysis Services, Hyperion, Cognos
- DW Extensions for RDBMS
  - Oracle11g, IBM DB2, MS SQL Server: SQL extensions, index structures, mat.Views, Bulk-Load/Insert, ...
- BI Accelerator
  - read-optimized DBS solutions: Main memory processing, column oriented data organization, MapReduce techniques, cluster architectures
  - e.g. SAP TREX, Greenplum, Vertica, EXASOL, ...
- ETL-Tools
  - MS Integration Services, Oracle Warehouse Builder, ...