Data Warehousing

FS 2020

Dr. Andreas Geppert
geppert@acm.org

What is a Data Warehouse ?

Content

- Introduction
- Typical Application Areas
- Definitions and Terminology
- Outlook and Literature

Motivation

- Enterprises and organizations must make decisions
- Decisions must be made on the basis of facts / information
- Information must be created out of internal and/or external data sources
- Information must be prepared and presented in such a way that the business users can effectively use them for decision making

Motivation (2)

- Required data are distributed over many data sources
- Information must be extracted out of data in database systems
- Data sets are typically very large

➲ Analyses cannot be done using operational systems

- Analyses have challenging functional requirements (analysis logic)
- Analyses have demanding performance requirements
- Visualization and presentation

➲ Data Warehousing and ➲ … business intelligence

Data Warehousing and Business Intelligence: omnipresent technology

"By analyzing customer behaviour over time through the use of Clubcard, Tesco found that in any single store, the top-spending 100 customers were as valuable as the bottom 4,000"

[Humby et al. 2003]

Data Warehousing: omnipresent technology (2)

Data Warehousing: omnipresent technology (3)

- “Wo fahren die meisten schwarz?
- Zürich - Seit einem Jahr hat der Computer das Notizblöckli der VBZ-Kontrolleure ersetzt. Mit den darin gespeicherten Daten will die Züri-Linie künftig deren Einsatzorte bestimmen. Dieses ‘ergebnisorientierte Fahrausweisprüfung’ genannte Programm soll noch dieses Jahr zum Einsatz kommen. ‘Wir sehen so in Zukunft schneller, in welchen Gegenden besonders viele Schwarzfahrer unterwegs sind’, so VBZ-Mann Heinz Illi.”
- 20 Minuten, 9. April 2003

Data Warehousing: omnipresent technology (4)

- “Wie der Kanton Zürich die Integration messen will
- „Zweitens soll eine sogenannte Umfeldanalyse Aussagen über den Integrationsstand der ausländischen Bevölkerung liefern —bis auf die Ebene der einzelnen Gemeinden hinunter. Hierzu werden umfangreiche Daten ausgewertet: über die Herkunft, ... aufenthaltsrechtlichen Status. Weiterhin sollen der Bildungsgrad, die Integration in den Arbeitsmarkt, der sozioökonomische Status und der gesundheitliche Zustand ermittelt werden. ... Ziel ist, jeder Gemeinde eine Analyse über den Ist-Zustand ihrer ausländischen Population liefern zu können ...”
- Tagesanzeiger, 16. Februar 2009

Data Warehousing: omnipresent technology (5)

- " ... Wer glaubt, dass sich eine lange Wartezeit allein am Anrufer-Aufkommen bemisst, irrt. Sie kann auch schlicht auf eine unvorteilhafte Telefonnummer zurückzuführen sein. Denn ob ein Anrufer in der Warteschlange vorn oder hinten lande, entscheide etwa bei einigen großen Mobilfunkgesellschaften der Computer ...Kunden, die über die Nummer identifiziert werden könnten und die in der Unternehmensdatenbank als "gut" klassifiziert seien, kämen schneller dran. "Gut" könnten sie sein, weil sie beispielsweise viel telefonierten und dem Unternehmen entsprechend hohe Umsätze einbrächten. Denkbar ist auch, dass ein Computer anrufende Neukunden automatisch bewertet. Da eine Telefonnummer Rückschlüsse auf den Wohnort des Anrufers ermöglicht, könnten Unternehmen solche Kunden in der Warteschleife nach vorn ziehen, die aus wohlhabenden Gegenden kommen und deshalb als potenziell attraktiv gelten. Technisch ist das jedenfalls kein Problem, und die erforderlichen Daten gibt es zur Genüge, denn Unternehmen können mittlerweile aus einem gigantischen Informationspool schöpfen: ...Bei den Klassifizierungen geht es darum, die Konsumenten zu bewerten und ihr künftiges Verhalten vorauszusagen. ..."
- Süddeutsche Zeitung, 15.07.2005

Data Warehousing: omnipresent technology (6)

- " ... The only problem is that identifying pregnant customers is harder than it sounds. … He ran test after test, analyzing the data, and before long some useful patterns emerged... As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy. …"
- «How Companies Learn Your Secrets», The New York Times Sunday Magazine, 19.02.2012 http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1

© The New York Times,
http://graphics8.nytimes.com/images/2012/02/19/magazine/19cover2/19cover2-articleInline-v2.jpg

Content

1. Introduction
2. Typical Application Areas
3. Definitions and Terminology
4. Outlook and Literature

Application Areas: Sales

- Products, customers, sales transactions, stores, suppliers
  - Sales, possibly in stores, possibly to known customers
- Goal: optimize sales
- Use analyses for:
  - Optimization of sales, product offerings
  - Identification of best sellers and slow sellers
  - Identification of product trends
  - Effectiveness of promotions

Examples: Sales, Trading

- Best Buy: large chain selling electronics etc.
- Applications: business, vendor & retail business performance management
- DWH contained every individual sales transaction from all stores
- Information about customers, customer behavior (buying patterns), product preferences, returns, warranty cases
- Wal-Mart: largest retailer world-wide (mainly in the US and Mexico)
- DWH contained 70 TB of data (already in 2001)
- Information and analyses regarding
  - products (sales, inventory)
  - stores
  - vendors
  - etc.

Examples: (electronic) Business

- Similar to conventional trade, but without stores
- Impact of web-based business:
  - Customers are known
    - Login, IP-address
  - Behavior in the Web can be monitored, e.g., via web server logging
- Clickstream mining
- Product analysis
- Customer analysis
- Cross-selling, ad placement

Example: Amazon

- DWH contained 25 TB of data in 2005, + 100% p.a.
- 2000 queries per day
- Data about customers, inventories and stocks, orders, products, supply chains, pricing, clickstream, …

Größeres Bild

Live at Fillmore East [Doppel-CD]
von Jimi Hendrix
(6 Kundenrezensionen)

Preis: EUR 14,97 Kostenlose Lieferung ab EUR
versandkostenfrei. Details

Auf Lager.
Verkauf und Versand durch Amazon.de. Geschenkverpa

Lieferung bis Dienstag, 23. Februar: Siehe Details.
27 neu ab EUR 12,67 4 gebraucht ab EUR 8,05

Amazon.de Verkaufsrang: Nr. 12.598 in Musik (Die Bestseller Musik)
Beliebt in dieser Kategorie:
Nr. 91 in Musik > Rock > **Psychedelic Rock**

Examples: Marketing

- Data about customers (current, past, potential)
- Data about promotions, marketing campaigns, returns, …
- Customer Relationship Analytics

- goal:
  - Optimally targeted advertisements (for potential customers)
  - Cross-Selling
  - Customer segmentation
- Use analyses for
  - Customer value calculation
  - Customer segmentation
  - Identification of potential customer needs
  - Campaign planning
  - Evaluation of campaigns (feedback loop)

Examples: Legal & Compliance

- Data
  - Typically industry-specific (banks, insurance companies, ...)
  - Regulatory rules (anti-money laundering, insider trading, «nachrichtenlose Vermögen», Basel II and III, ...)
- Goals:
  - Implement controls (e.g., detect potential money laundering)
  - Document compliance for regulator

Examples: Quality Assurance

- Data about products (goods or services), components, returns, warranty cases
- goal:
  - Optimize customer satisfaction
  - Competitive advantage because of better quality
  - Cost reduction through reduction of returns
- Analysis for:
  - Identification of quality problems (error cases, error causes)
  - Pricing based on quality characteristics

Obere Mittelklasse/Oberklasse

Rang	*Marke und Modell*
1	Audi A6
2	BMW 5er
3	Mercedes E-Klasse
4	Volvo S60/S70/S80/V70
5	Opel Signum

Quelle: TCS Pannenstatistik 2008, http://www.tcs.ch/main/de/home/der_tcs/presse/mitteilungen/pannenstatistik_2008.RightColumn.0001.CtxLinkDownloadFile1.pdf/Pannen2008_Uebersicht_Fahrzeugklassen.pdf

Examples: Inventory Management

- Stock data (inventories, products, suppliers, etc.)
- goal:
  - Optimal stock management
  - Minimal inventory levels without delivery shortages
- Analyses
  - Inventory management
  - Design of supply, ordering, and stocking processes
  - Capacity analysis and planning
- Example: Wal-Mart

Examples: Financial Services

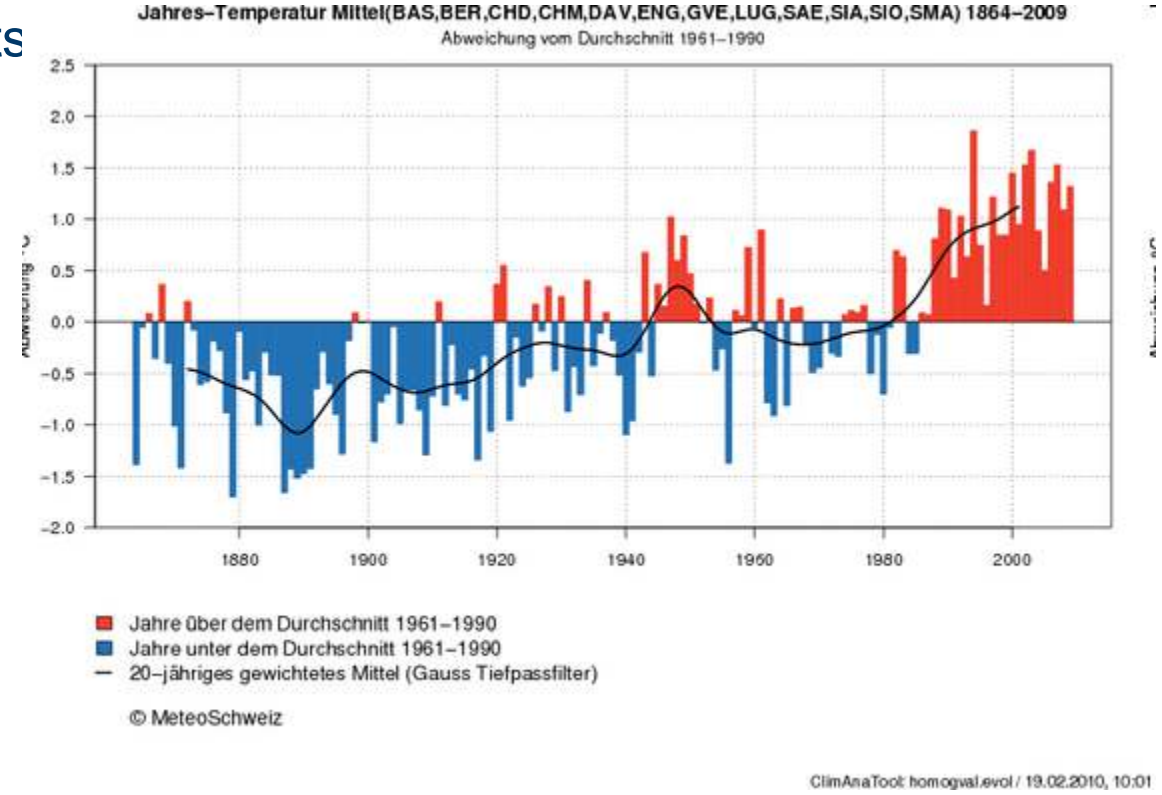
- Data about customers, accounts, payments, assets, loans, …
- Goals:
  - Productivity optimization
  - Risk management (market risk, credit risk)
  - Customer relationship management
  - compliance
- Analysis:
  - Risk calculation, analysis, and monitoring
  - Performance monitoring
  - Customer relationship analytics
  - Regulatory reporting
    - Basel II and III, national regulatory authorities, stock exchanges

Examples: Telecom

- Data about networks, lines, calls, customers, outages, …
- Analyses for
  - Customer behavior
  - Churn analysis
  - fraud detection
  - Network utilization (capacity management and planning)
- Example: France Telekom
- ca. 180 billion calls (Call Detail Records, CDRs, 2003)
- today, typical telco data warehouses are in the petabyte range
- Supporting fraud detection, analysis of network traffic, customer service

Examples: Climatology and Meteorology

- Data about measurements of meteorological parameters
- goals:
  - Weather forecast (Meteorology)
  - Climatic trends
- Analyses for
  - Optimization of weather forecast
  - Understanding of climate models
  - Understanding climate change



Quelle: http://www.meteoschweiz.admin.ch/web/de/klima/klima_heute/trends_schweiz.html

Examples: Environmental Science

- Measurements of environmental data, e.g. pollution
- goals:
  - understand environmental trends, e.g. impact of pollution
- Analyses
  - Pollution in the ground, water, air
  - Location-based
  - Time-dependent

Examples: Medical, Biological and Sociological Sciences

- Data about the population, patients, diseases
- Goals:
  - Understanding of habits and lifestyles, trends, diseases, effectiveness of treatments
- Analyses e.g. for
  - Identification of risk factors
  - Regional distribution of diseases
  - Changes in health state over time
- Classical application area for statistical analysis
- "When we weren't in his office, working out our tabulations, curves, and correlation charts, we were off on the road, collecting data, because, as Prok said, over and over, you could never have enough data." [T.C. Boyle, The Inner Circle]

Examples: Public Administration, Electronic Government

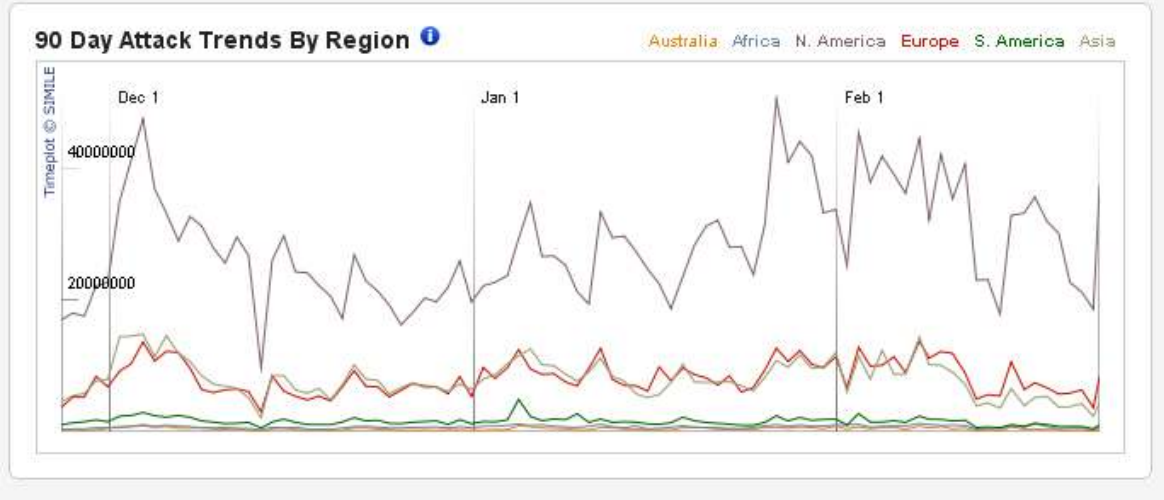
- Data
  - demographic data
  - Public services
- Goals:
  - Planning (kind of capacity management and planning), depending on demographic trends and population needs
  - Adequate spend of public means
- Analyses for
  - Identification of demographic facts and trends
  - Fraud detection
  - Performance management
  - Infrastructure planning

Examples: Technical Data Warehouses

- Data
  - Technical data (inventories, capacity, tickets, …)
  - Logs
  - Measurements and sensor data
- Goals:
  - Planning, resource optimization
  - Optimization of IT processes
  - Meeting agreed service levels
- Analyses:
  - Performance analysis
  - capacity management
  - SLA-Reporting

Technical Data Warehouses: Security

- Data
  - Requests and logs
  - User data
  - Access rules
- Goals:
  - Compliance
  - Enforce the need-to-know principle
  - Defense against attacks
- Analyses:
  - Analysis of browser logs
  - Identification of access roles (Role Mining)
  - Auditing and traceability



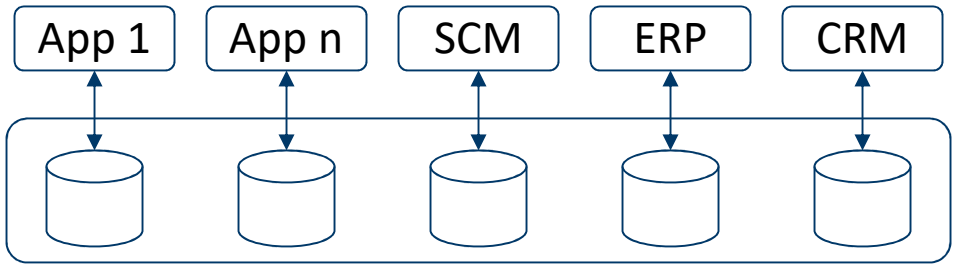
Quelle: http://www.symantec.com/business/security_response/index.jsp

Content

- Introduction
- Typical Application Areas
- **Definitions and Terminology**
- Outlook and Literature

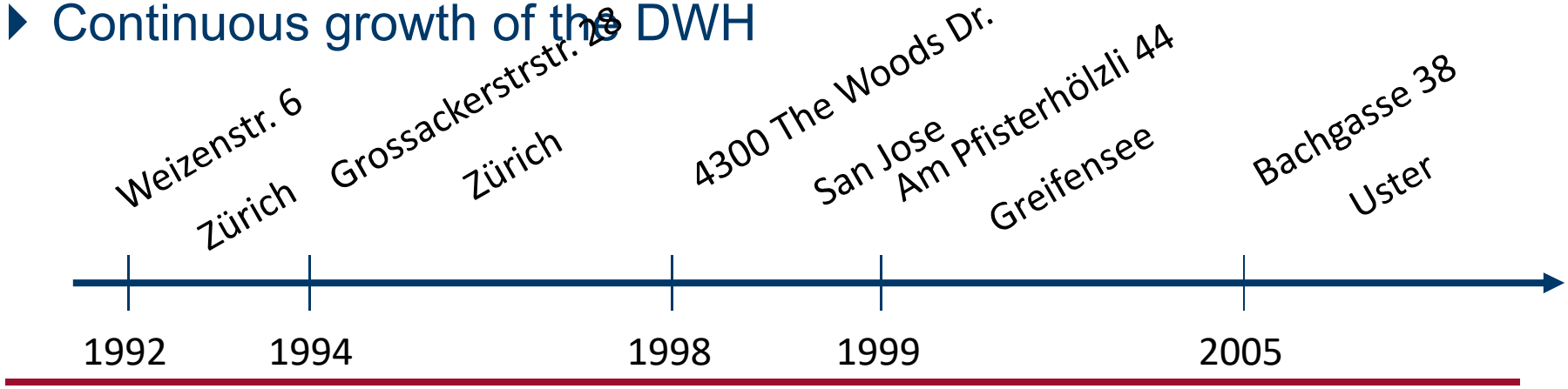
Integration

- Inclusion of all relevant data
- Required data are typically distributed over many data sources, especially in large organizations
- Internal and external data sources
- Data sources are typically heterogeneous
  - Because of mergers and acquisitions, or lack of central data management
- DWH must provide integrated view of all the relevant data
  - “schema integration"
  - Data integration



Historization

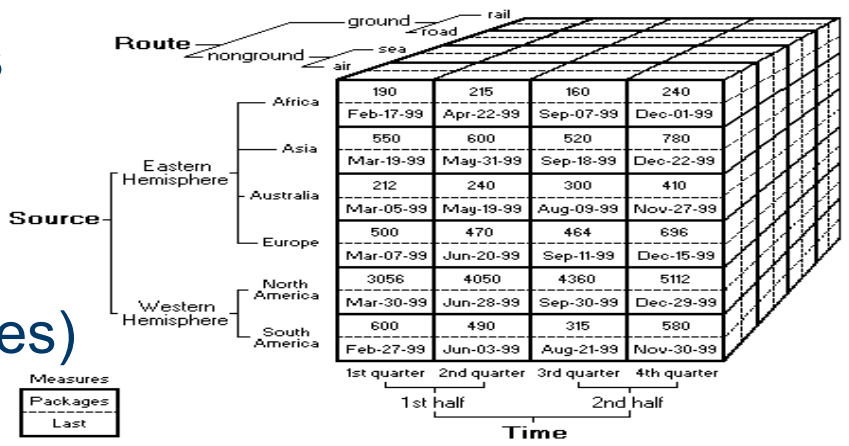
- Data sources (operational systems) often provide only current view
  - Modulo archiving, depending on retention requirements
- But the DWH must support analysis along the time axis
- Data in the DWH must be enriched with time information
- No deletion of old data
- Continuous growth of the DWH



Multi-Dimensionality

- Analysis of data along multiple criteria and perspectives

- Time dimension practically always required
- Spatial dimension (e.g., where do customers live, where are my stores)
- Further dimensions depending on the application
- Customers, products, stores, employees, …



Definition (1)

- What is a Data Warehouse?
- Definition according to Westerman:

"The concept of data warehousing is really quite simple. Data from older systems is **copied** into a new computer system **dedicated entirely to analyzing** that data. Normally, the data warehouse will store a substantial amount of **historical data**. Users of this system are able to **continuously ask or query** it to retrieve data for analysis...

Use your data to provide information to people in your company so that they can **make better, informed decisions faster**"

Definition (2)

- What is a Data Warehouse?
- Inmon‘s definition:

A data warehouse is a

subject-oriented, integrated, non-volatile, and **time-variant**

collection of data

in **support of management‘s decisions**

- „decision support“ might be considered as too narrow
- „in support of satisfying management‘s information needs“ might be more appropriate

Terminology

- Data Warehouse: Collection of data and metadata
- Data Warehouse System: Data + Metadata + Software (≡ database system)
- Data Warehousing: overall process of building and using a data warehouse system

- Online Analytical Processing, OLAP:
  - Special form of analytics
  - Often used as synonym for DWH, delineation against OLTP
- Online Transaction Processing, OLTP:
  - Operational transaction processing (e.g., executing a sales transaction, performing a payment)

Business Intelligence

- "Data analysis, reporting, and query tools can help business users wade through a sea of data to synthesize valuable information from it – today these tools collectively fall into a category called business intelligence“
  [Gartner 2004]
- "Unter Business Intelligence wird ein integrierter, unternehmensspezifischer, IT-basierter Gesamtansatz zur betrieblichen Managementunterstützung verstanden“
  [Kemper & Baars 2006]

Delineation: Access Profile and Queries

Differences …	OLTP	DWH
Operations	read, insert, update, delete	Read, periodical inserts
“transactions”	(very) short	Long
Response times	ms-s	s-m-h
Queries	Simple	complex
Operate on	Single tuples	Tuple ranges, aggregates
Results	Single/few rows	Many rows

Delineation: Data

Differences …	OLTP	DWH
Properties	Original	Derived
	Current	Historized
	Autonomous	Integrated
	Volatile	Non-volatile
Size	Gigabyte	Terabyte and more
Schema	Application-independent	Tailored to analysis

Delineation: Users

Differences …	OLTP	DWH
User profiles	Clerks	Analysts
End-users		Controller
		Managers
		Analysts, “Data Scientists”
Numbers	Many	Few
Access	Via applications	Analytic applications
		Analysis/Query tools

Outline

- **Introduction**
- DWH Architecture
- DWH-Design and multi-dimensional data models
- Extract, Transform, Load (ETL)
- Metadata
- Data Quality
- Analytic Applications and Business Intelligence
- Implementation and Performance
- Security and Privacy (?)