# Data Warehousing

Einleitung und Motivation

Ulf Leser

Wissensmanagement in der Bioinformatik


# Bücher im Internet bestellen

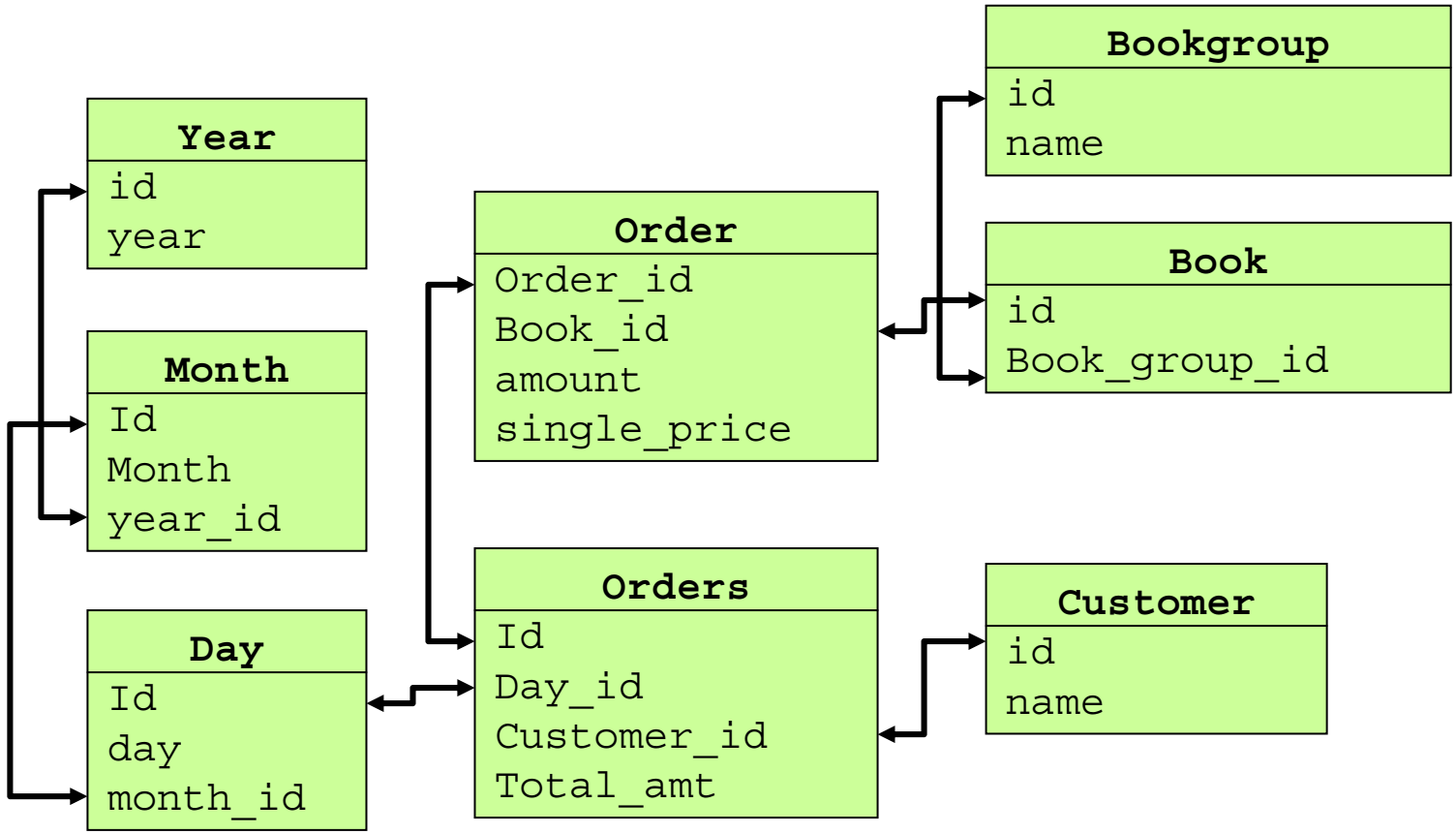



Backup
Durchsatz
Load-
balancing


Portfolio
Umsatz
Werbung

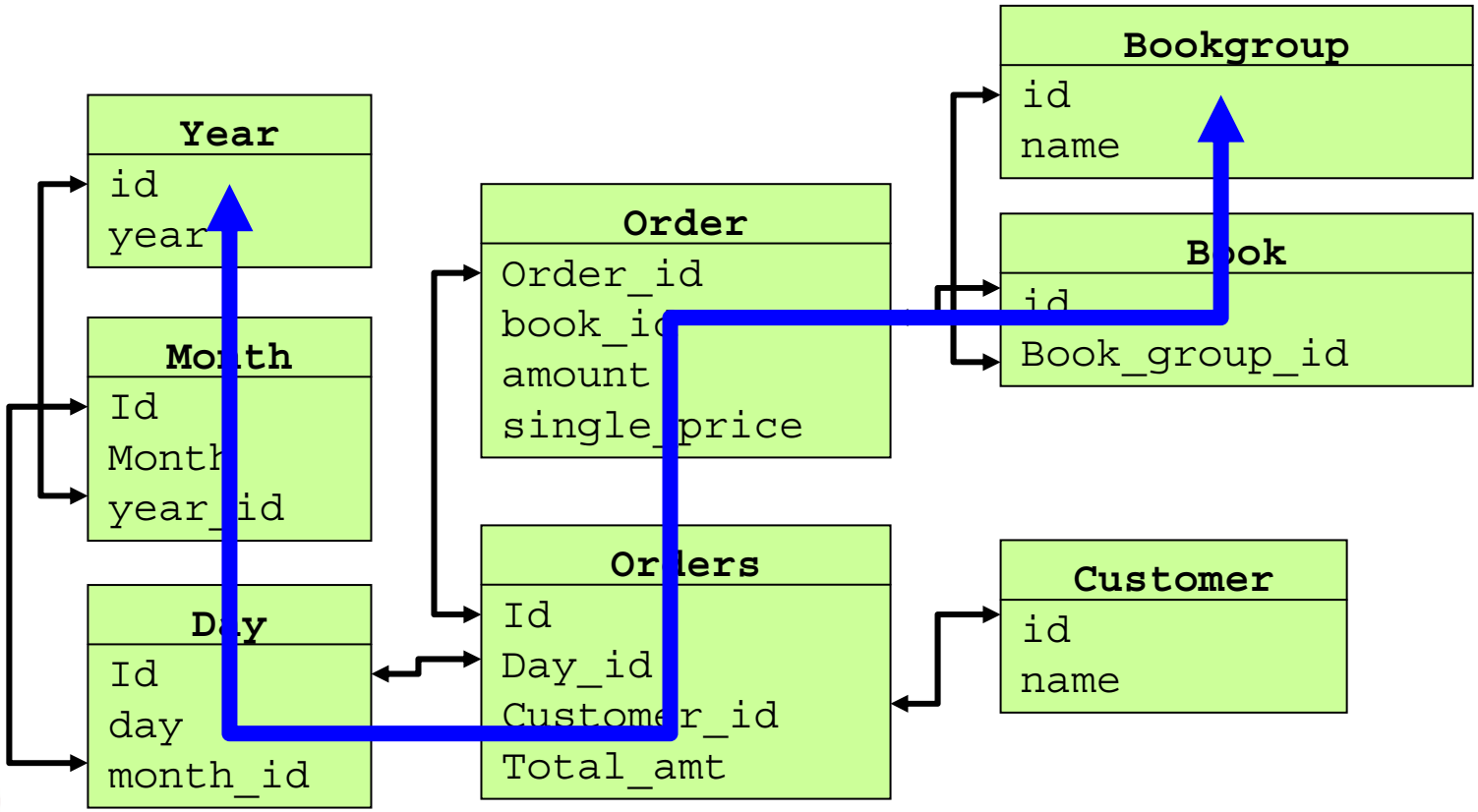
**Daten bank**

# Die Datenbank dazu



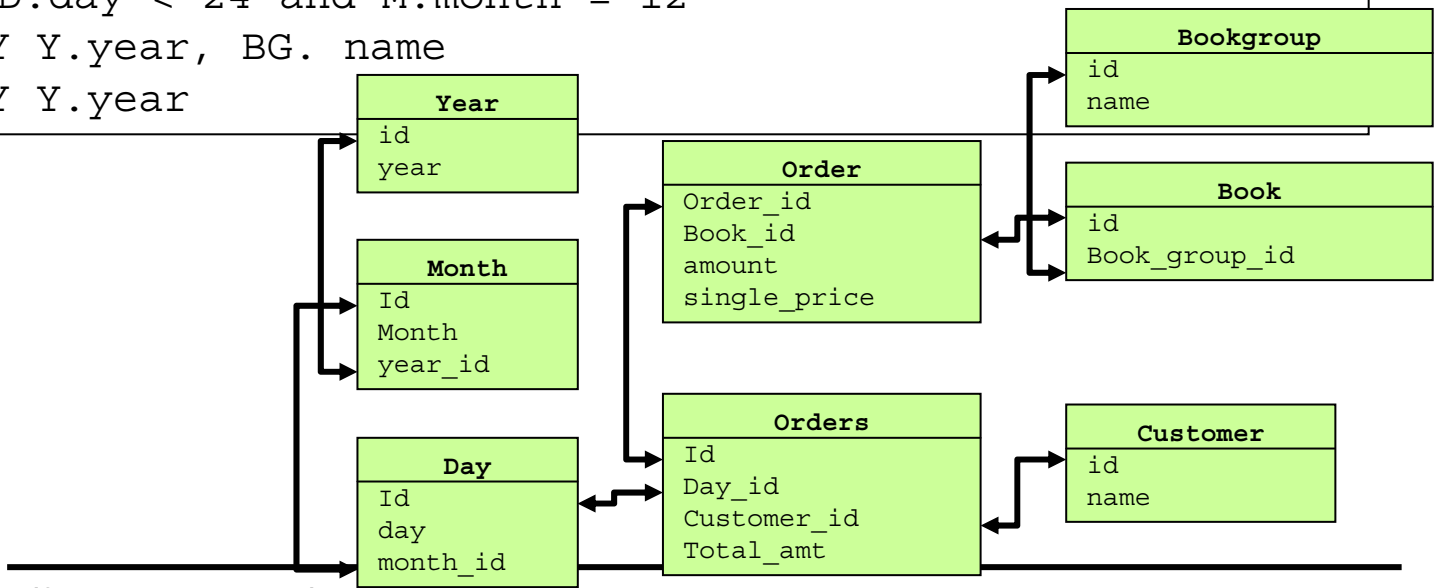
# Fragen eines Marketingleiters

Wie viele Bestellungen haben wir jeweils im Monat vor Weihnachten, aufgeschlüsselt nach Produktgruppen?



# Technisch

```
SELECT  Y.year, BG.name, count(O.id)
FROM    year Y, month M, day D, order O, orders OS,
        book B, bookgroup BG
WHERE   M.year = Y.id and
        M.id = D.month and
        OS.day_id = D.id and
        OS.id = O.order_id and
        B.id = O.book_id and
        B.book_group_id = BG.id and
        D.day < 24 and M.month = 12
GROUP BY Y.year, BG. name
ORDER BY Y.year
```



# Ergebnis

```
SELECT   Y.year, BG.name, count(O.id)
FROM     year Y, month M, day D, order O, orders OS,
         book B, bookgroup BG
WHERE    M.year = Y.id and
         M.id = D.month and
         OS.day_id = D.id and
         OS.id = O.order_id and
         B.id = O.book_id and
         B.book_group_id = BG.id and
         D.day < 24 and M.month = 12
...
```

## 6 Joins

- Year: 10 Records
- Month: 120 Records
- Day: 3650 Records
- Orders: 36.000.000
- Order: 72.000.000
- Books: 200.000
- Bookgroups: 100

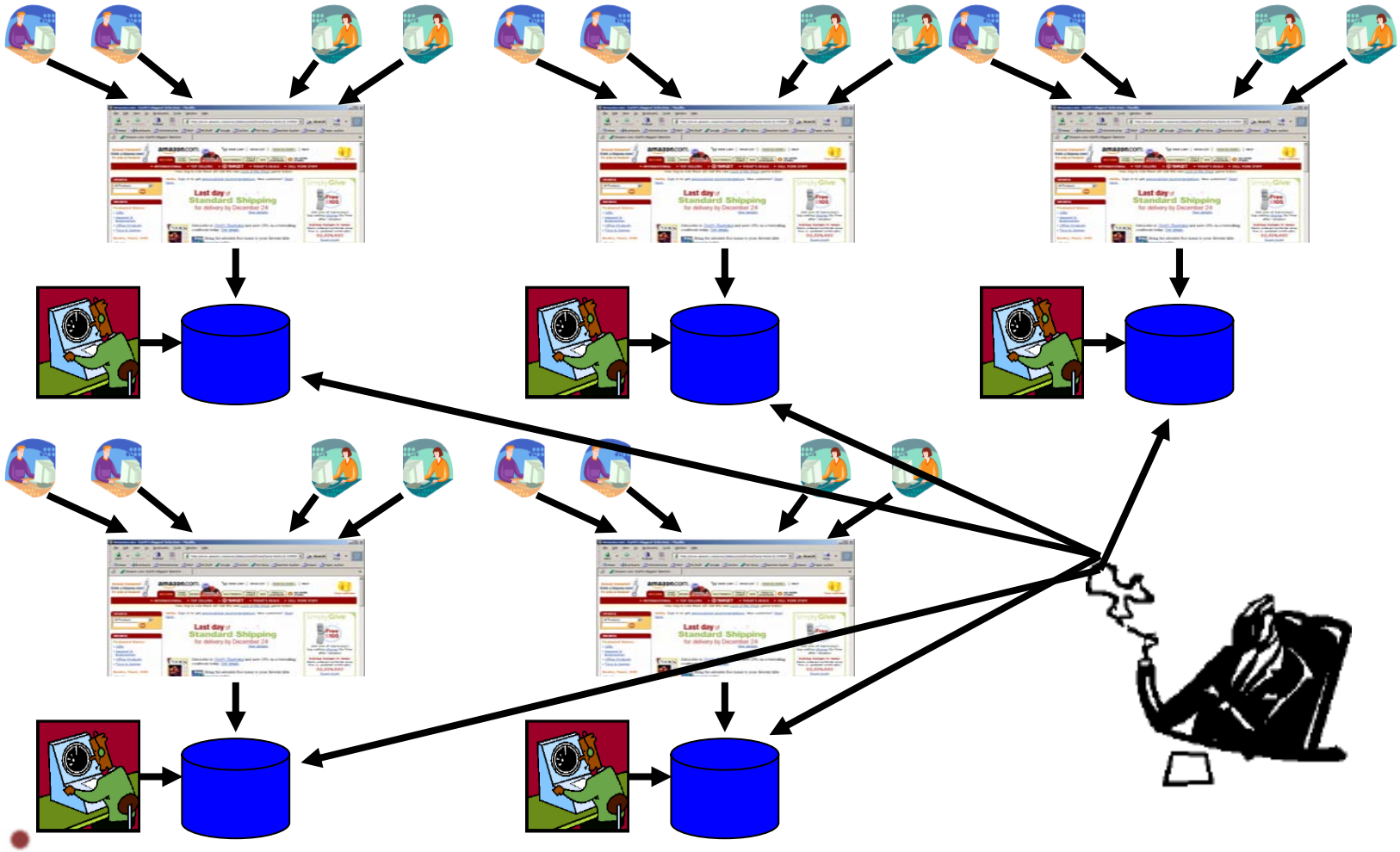
## Problem!

- Schwierig zu optimieren (Join-Order)
- Ja nach Ausführungsplan riesige Zwischenergebnisse
- Ähnliche Anfragen – ähnlich riesige Zwischenergebnisse

# In Wahrheit ...

- Es gibt noch:
  - Amazon.de
  - Amazon.fr
  - Amazon.it
  - ...
- Verteilte Ausführung
  - Count über Union mehrerer gleicher Anfragen in unterschiedlichen Datenbanken

# In Wahrheit ...



# Technisch

```
CREATE VIEW christmas AS
   SELECT      Y.year year, BG.name name, count(O.id) ocount
   FROM        DE.year Y, DE.month M, DE.day D, DE.order O, ...
   WHERE       M.year = Y.id and
   ...
   GROUP BY    Y.year, BG.name
   ORDER BY    Y.year
UNION ALL
   SELECT      Y.year year, BG.name name, count(O.id) ocount
   FROM        EN.year Y, EN.month M, EN.day D, EN.order O, ...
   WHERE       M.year = Y.id and
   ...
```

```
SELECT      year, name, count(ocount)
FROM        christmas
GROUP BY    year, name
ORDER BY    year
```

# Probleme

- Count über Union über verteilte Datenbanken?

➢ Heterogenitätsproblem

  - Quellen werden Schemata verändern
  - Länderspezifischer Eigenheiten (MWST, Versandkosten, Sonderaktionen, ...)
  - Viele verborgene Änderungen

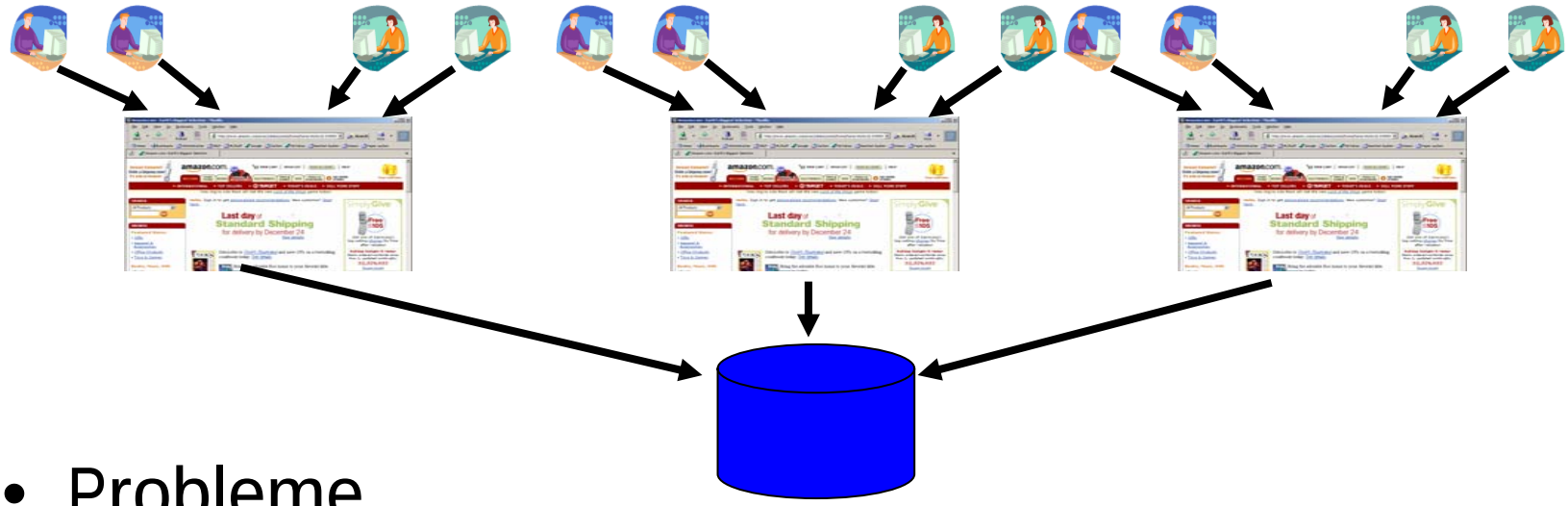
- Berechnung der Zwischenergebnisse bei jeder Anfrage?

➢ Datenmengenproblem

  - Historische Sicht - Datenmengen wachsen weiter
  - Operative Systeme brauchen eigentlich die historischen Daten nicht
  - Erfordert evt. Transport großer Datenmengen durchs Netz

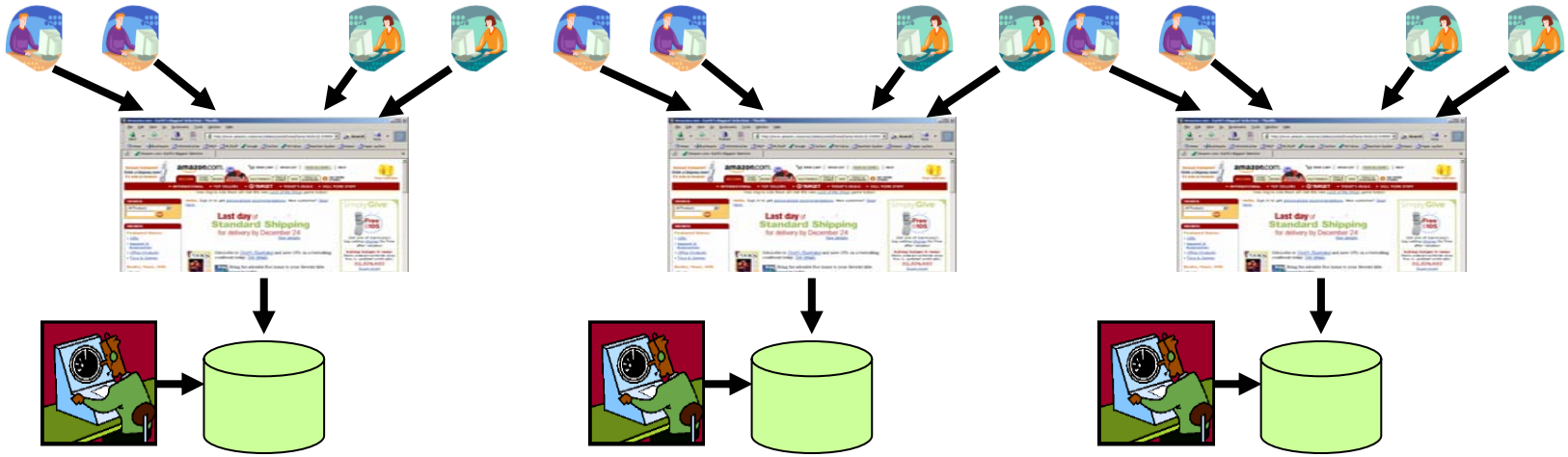
# Lösung Heterogenitätsproblem?

**Zentrale Datenbank**



- Probleme
  - Zweigstellen schreiben übers Netz
  - Schlechter Durchsatz
  - Lange Antwortzeiten im operativen Betrieb

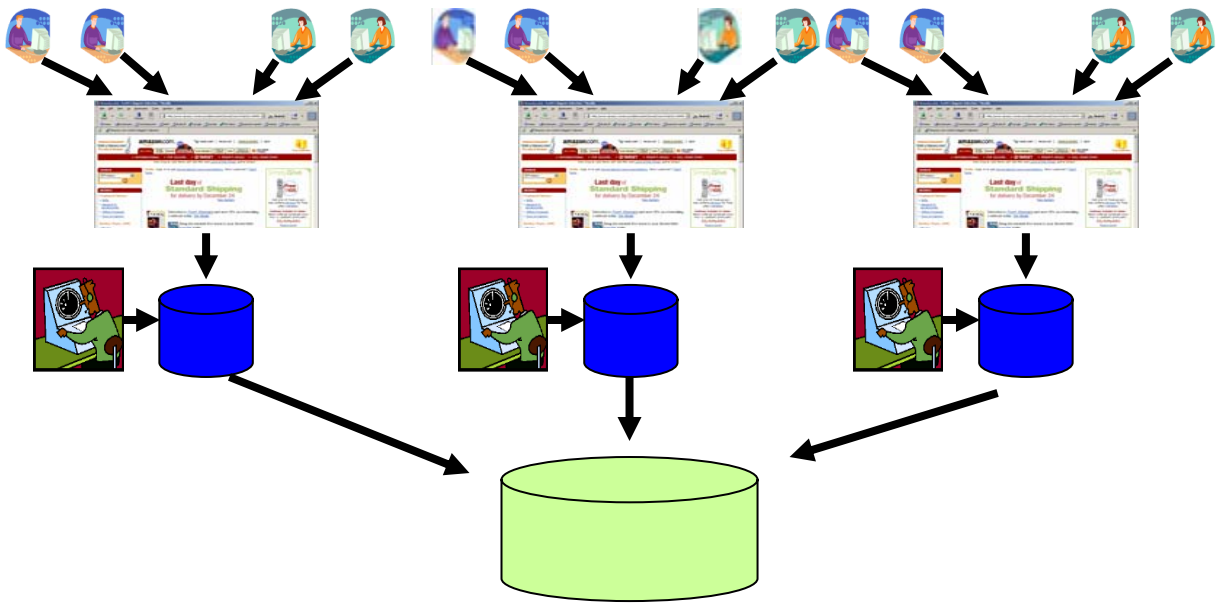
# Datenmengenproblem



- Probleme
  - Schnelle lokale Anfragen, Verteilungsproblem bleibt
  - Jeder lesende / schreibende Zugriff erfolgt auf eine Tabelle mit 72 Mill. Records
  - Lange Antwortzeiten im operativen Betrieb

# Tatsächliche Lösung

## Aufbau eines Data Warehouse



- Redundante Datenhaltung
- Spezielle Modellierung
- Transformierte und präselektierte Daten
- Asynchrone Aktualisierung

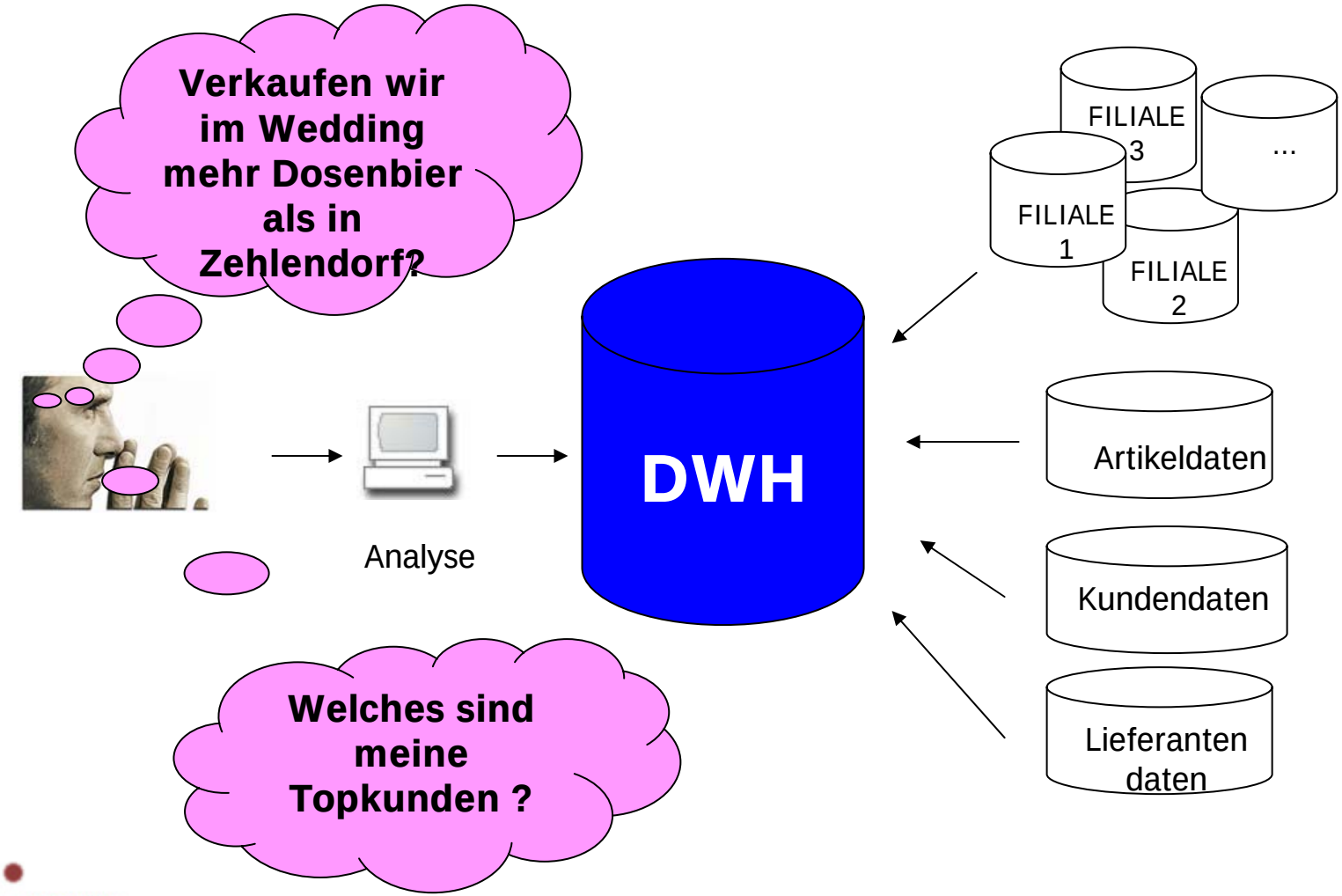
# Inhalt dieser Vorlesung

- Definition und Abgrenzung
- Geschichte & Statistische Datenbanken
- Sichtweisen
  - Betriebswirtschaft / Informatik
  - Benutzer / Entwickler
- Einige Anwendungen
- Grosse Datenmengen und TPC-H

# Beispielszenario

- Ein beliebiges Handelshaus : Spar, Extra, Kaufhof, ...
- Physikalische Datenverteilung
  - Viele Niederlassungen (bis zu mehrere tausend)
  - Noch mehr Registerkassen
- Aber: Zentrale Planung, Beschaffung, Verteilung
  - Was wird wo und wie oft verkauft?
  - Was muss wann wohin geliefert werden?
    - Bedenke: Verderbliche Waren
- ... nur möglich, wenn
  - Zentrale Übersicht über Umsätze
  - Integration mit Lieferanten / Produktdaten

# Handels - DWH


Verkaufen wir im Wedding mehr Dosenbier als in Zehlendorf?
Analyse
DWH
FILIALE 3
...
FILIALE 1
FILIALE 2
Artikeldaten
Kundendaten
Lieferanten daten
Welches sind meine Topkunden ?

# DWH Datenquellen

- Lieferantendatenbanken
  - Produktinformationen: Packungsgrößen, Farben, ...
  - Lieferbedingungen, Rabatte, Lieferzeiten, ...
- Personaldatenbank
  - Zuordnung Kassenbuchung auf Mitarbeiter
  - Stundenabrechnung, Prämien
- Kundendatenbank
  - Kundenklassen; Premium, Normal, soziale Brennpunkte, ...
  - Persönliche Vorlieben & Historie
    - Kundenkarten, TESCO
- Weitere Vertriebswege
  - Internet, Katalogbestellung, Verkaufsclubs, ...

# Definition DWH

- *A DWH is a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management's decision* [Inm96]
  - Subj-orient.: Verkäufe, Personen, Produkte, etc.
  - Integrated: Erstellt aus vielen Quellen
  - Non-Volatile: Hält Daten unverändert über die Zeit
  - Time-Variant: Vergleich von Daten über die Zeit
  - Decisions: Wichtige Daten rein, unwichtige raus

# Geschichte von DWH

- Managementinformationssysteme (MIS),
  Decision Support Systeme (DSS),
  Executive Information Systeme (EIS)
  - Seit den 60er Jahren
  - Feste, programmierte Reports
  - Redundante Datenhaltung sehr teuer – dadurch langsam, oft Downtimes für operative Systeme
  - Datenzugriff sehr schwierig: fehlende Vernetzung, Schnittstellen, Anwendungsneutralität nicht gegeben
  - Analysemethoden ungenügend (Data Mining)
  - Schwierig zu bedienen

➢ Teuer und unflexibel

➢ Schattendasein

# Erfolg von DWH

- Top-Thema seit Mitte der 90er Jahre
- Voraussetzungen
  - Extreme Verbilligung von Plattenspeicherplatz
  - Relationale Modellierung: Anwendungsneutral
  - Graphische Benutzeroberflächen und Terminals
  - IT in allen Unternehmensbereichen (SAP R/3)
  - Vernetzung und DB Standardisierung (SQL)
- Aber
  - Vision der vollständigen Integration scheitert (immer wieder aufs neue)
- Soziale versus technische Aspekte

# Abgrenzung

- Parallele Datenbanken
  - Blendet Heterogenität aus
  - Ausfallsicherheit und Performanceverbesserung
  - Technik zur Realisierung eines DWH
- Verteilte Datenbanken
  - Gewollte Verteilung als Mittel zur Lastverteilung
  - Keine physische Integration der Daten (Materialisierung)
- Föderierte Datenbanken
  - Höhere Autonomie und Heterogenität
  - Verteilung bleibt erhalten

# Sichtweisen auf DWH Projekte

- Betriebswirtschaftliche Sichtweise
- Informatiker Sichtweise

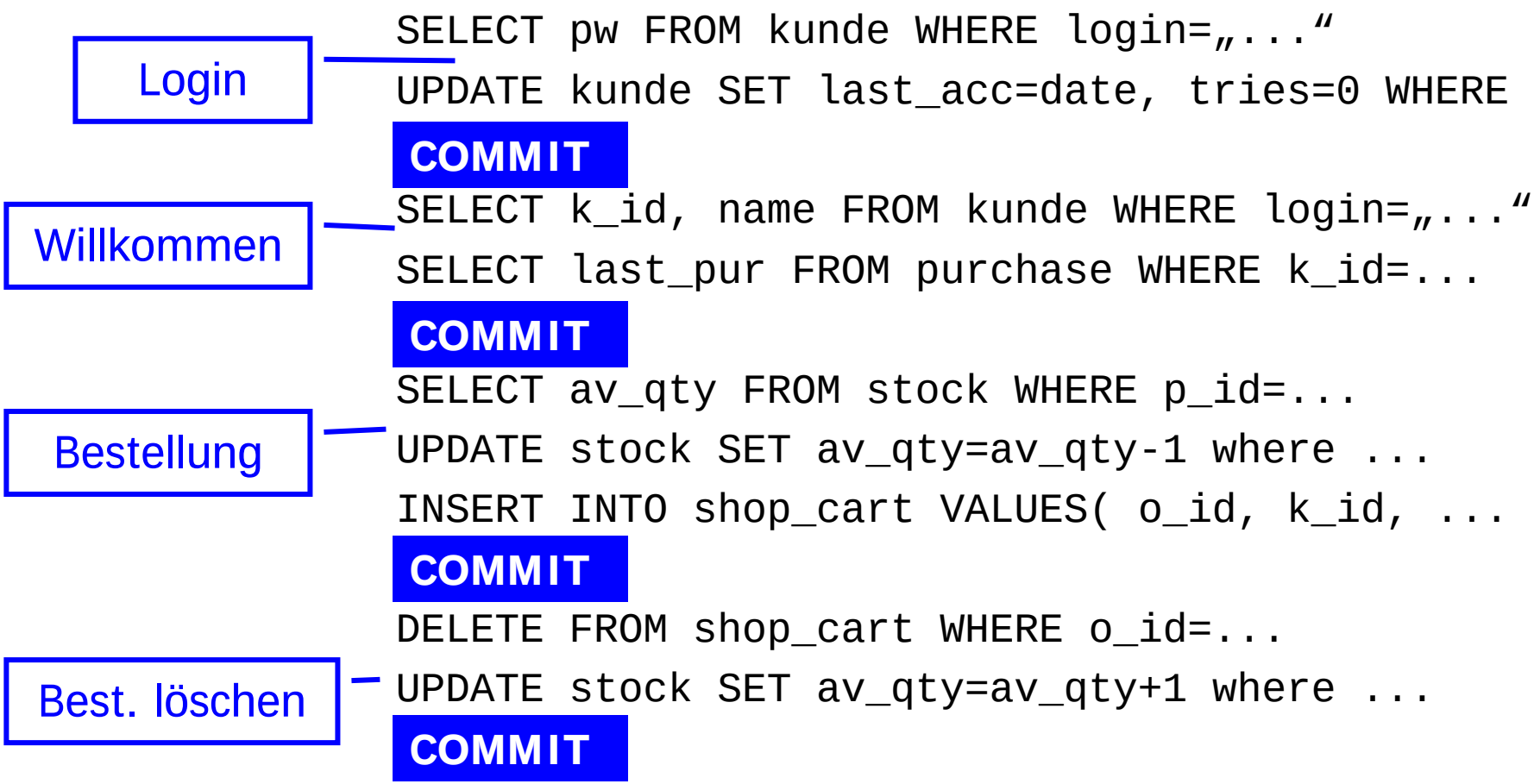
# Betriebswirtschaftssicht

- Ein DWH
  - Ermöglicht viele neue Fragen
  - Verbessert viele Antworten erheblich
- ... durch ...
  - Zugriff auf integrierte Daten
    - Ohne DWH nur manuell (Programmierung) möglich
  - Übergreifende Analysen
    - Inhaltliche Verknüpfung der Daten
  - Bessere Datenqualität
    - Fehlerminimierung, Ergänzung, Plausibilitätschecks
    - Entfernung von Fehlern hier verkraftbar, aber nicht in den operativen Systemen
  - Anreicherung mit externen Daten
    - Externe Kundenprofile, geographische Daten
    - Würde operative Systeme überlasten

# Informatikersicht

- Operative Systeme
  - Viele Benutzer
  - Kurze Transaktionen, einfache Queries
  - Echtzeitanforderungen
  - Kurzes Gedächtnis
  - Beispiel: Kassensysteme, Bankautomaten
  - OLTP (Online Transaction Processing)
- DWH
  - Wenige(r) Benutzer
  - Komplexe Queries mit langer Laufzeit, nur lesend
  - Zeitlich eher unkritisch
  - Historische Daten
  - Beispiel: Sortimentplanung, Kapazitätsplanung
  - OLAP (Online Analytical Processing)

# OLTP Beispiel

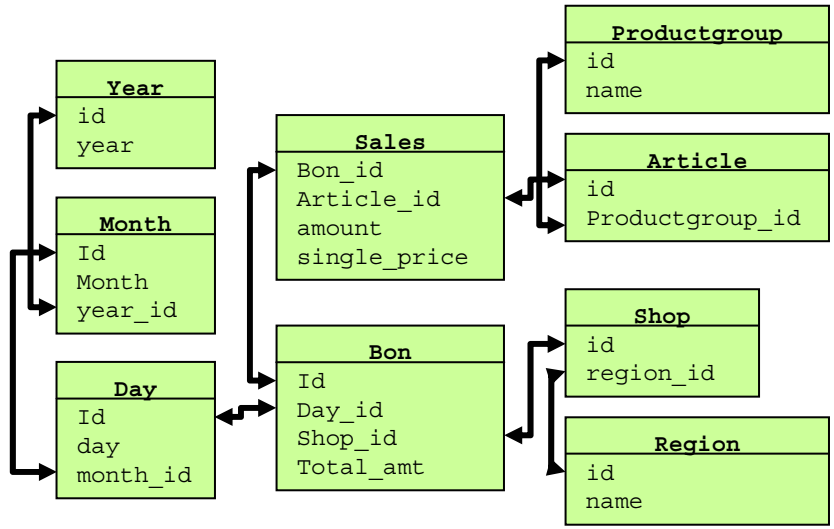


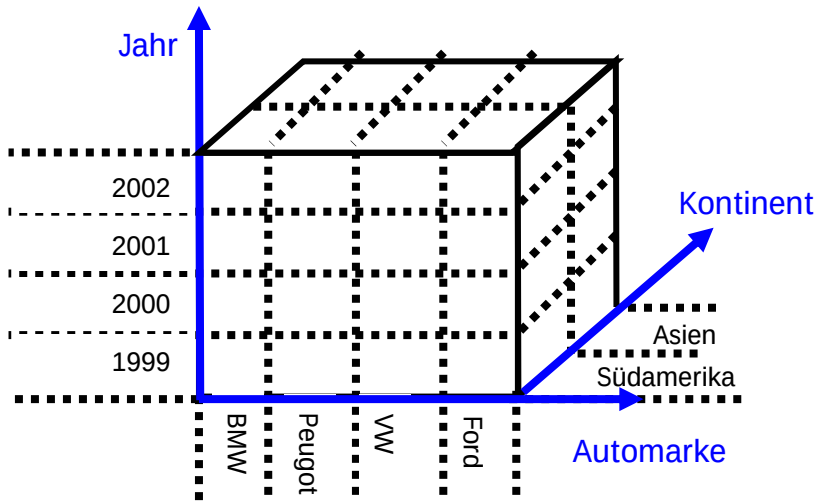
# OLAP Beispiel

- Welche Produkte hatten im letzten Jahr im Bereich Bamberg einen Umsatzrückgang um mehr als 10%?
  - Welche Produktgruppen sind davon betroffen?
  - Welche Lieferanten haben diese Produkte?
- Welche Kunden haben über die letzten 5 Jahre eine Bestellung über 50 Euro innerhalb von 4 Wochen nach einem persönlichen Anschreiben aufgegeben?
  - Wie hoch waren die Bestellungen im Schnitt?
  - Wie hoch waren die Bestellungen im Vergleich zu den durchschnittlich. Bestellungen des jew. Kunden in einem vergleichbaren Zeitraum?
  - Lohnen sich Mailing-Aktionen?
- Haben Zweigstellen einen höheren Umsatz, die gemeinsam gekaufte Produkte zusammen stellen ?
  - Welche Produkte werden überhaupt zusammen gekauft – und wo?

# Modellierung im DWH

- Modellierung in operativen Systemen: Normalisierung
- Modellierung in DWH: Dimensionen und Fakten



# OLAP versus OLTP

Ausfall kostet Millionen.
Hochverfügbarkeit:
24x7x52

Ausfall ärgerlich

| | OLTP | OLAP |
|---|---|---|
| Typische Operationen | Insert, Update, Delete, Select | Select<br>Bulk-Inserts |
| Transaktionen | Viele und kurz | Lesetransaktionen |
| Typische Anfragen | Einfache Queries, Primärschlüsselzugriff, Schnelle Abfolgen von Selects/inserts/updates/deletes | Komplexe Queries: Aggregate, Gruppierung, Subselects, etc.<br>Bereichsanfragen über mehrere Attribute |
| Daten pro Operation | Wenige Tupel | Mega-/ Gigabyte |
| Datenmenge in DB | Gigabyte | Terabyte |
| Eigenschaften der Daten | Rohdaten, häufige Änderungen | Abgeleitete Daten, historisch & stabil |
| Erwartete Antwortzeiten | Echtzeit bis wenige Sek. | Minuten |
| Modellierung | Anwendungsorientiert | Themenorientiert |
| Typische Benutzer | Sachbearbeiter | Management |

- Anwendungsgebiete

# Wal Mart

- Unternehmensweites Data Warehouse
  - Verkaufzahlen, Lagerhaltung, Filialen
- Größe: ca. 300 Terabyte (4/03)
- Größe: ca. 480 Terabyte (11/04, New York Times)
  - 100.000.000 Kunden pro Woche, 3.600 Geschäfte
  - Datenbank: Teradata, NCR
  - Keine Paybackinformation – aber viele Kreditkarten
- Täglich bis zu 20.000 DW-Anfragen
  - Die meisten Analysen laufen auf Shop-, nicht auf Kundenebene
  - Überprüfung des Warensortiments zur Erkennung von Ladenhütern oder Verkaufsschlagern
  - Standortanalyse, Rentabilität von Niederlassungen
  - Untersuchung der Wirksamkeit von Marketing-Aktionen
  - Analyse / Planung der Zwischenlager

[WesS00] Paul Westerman: „Data Warehousing: Using the Wal-Mart Model“, Morgan Kaufman Pub, 2000

# SBC communications

- Amerikanisches Telekommunikationsunternehmen
- 57.000.000 Telefon / DSL Kunden
- Teradata Datenbank
- „Derzeit größtes DWH der Welt“ (4/2004)
  - Ca. 360 Terabyte
  - 12.000 Tabellen (?)
  - 300.000 Logins pro Tag
  - Massiv parallele 500 Prozessor-Maschine
- Teradata
  - Erste Terabyte Warehouse: 1990
  - Testsysteme laufen heute (5/2004) im Petabyte Bereich
  - Erwarten das im kommerziellen Betrieb ab 2006

# CRM: Customer Relationship Management

- Vertriebswege, Vertriebsorganisationen
- Kundenkäufe, Präferenzen, Kontaktdaten, Reklamationen, Bonität
- Wer sind Premiumkunden -> Sonderbehandlung
- Beratungssysteme, Personalisierung , Mailings

Amazon
- Customers who bought X also bought Y
- Personalisierte Empfehlungen
- Neuerscheinungen

# Controlling & Kostenrechnung

- Kostenstellen, Kostenträger, Kostenarten
- Gewinn-Verlustrechnung
- Bilanzierung
- Vollkostenrechnung
- Plan – Ist Zahlen, Szenarios
- Kennzahlensysteme
- Balanced Scorecard

# Weitere Anwendungsgebiete

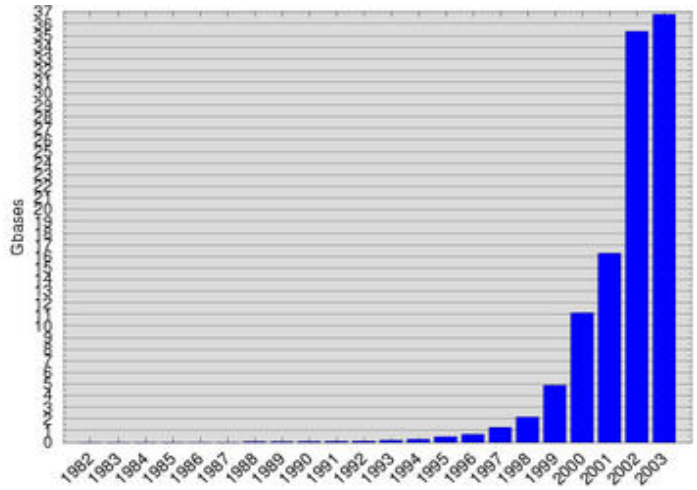
- Logistik
  - Flottenmanagement
  - Disposition
  - Tracking
- Finanzen
  - Kreditkartenanalyse
  - Risikoanalysen
  - Fraud Detection
- Health Care
  - Studienüberwachung
  - Wirkstoffanalyse

# Was sind große Datenbanken ?

- British Telecom
  - 20 Terabyte
  - >100 Milliarden Records
  - Datenbank: Oracle 8i
- Deutsche Telecom
  - 100 Terabyte
  - IBM, DB2
- Walmart
  - „Wal-Mart … will expand its data-warehouse system, which handles 50,000 queries a week, from 7.5 to 24 terabytes.“ [Info. Week, 1997]
  - „At 70 terabytes and growing …“ [Wes, 2000]
  - „… 300 Terabyte Datenbank“ [Computerzeitung, 16/2003]
  - NCR Teradata

# Was sind große Datenbanken –2-

- Beispiel: Google
  - [ct, 1/2003, Heise Verlag]
    10 Terabyte Rohdaten
    15.000 PC, 1 Petabyte (Redundanz)
  - [Technology Review, 4/2004, Heise Verlag]
    >100.000 PC, 4 Petabyte
- Genbank (Humanes Genom)
  - Release 104, Flatfile:
    Ca. 100 Gigabyte
  - Aber eindrucksvolles
    Wachstum ...




Quelle: EMBL, Stand 17.2.2003

Home | Video | Kurzberichte | Wirtschaft | Politik | High-Tech | Wissenschaft | Unterhaltung | Lifestyle | Vermischtes

Dienstag 12. Oktober 2004, 13:35 Uhr

## Data-Warehouse speichert eine Bio. Datenreihen

155 Terabyte Rohdaten auf 55 Terabyte-Speicher

Dublin (pte) - Der kalifornische Anbieter von IT-Infrastrukturen und mobiler Softwar[illegible]ybase http://www.sybase.com und Sun Microsystems http://www.sun.com haben n[illegible] eigenen Angaben das derzeit weltweit größte Data-Warehouse aufgebaut. Die analytische Datenbank Syb[illegible]e-IQ und iForce-Enterprise-Data-Warenhouse-Referenzarchitektur von Sun komprimi[illegible]en insgesamt 155 Terabyte Rohdaten auf weniger als 55 Terabyte Speicherdaten.

"Unternehmen sehen sich heute explodierenden Datenmengen gegenüber, de[illegible] Quellen von Online-Transaktionen bis zu RFID-Übertragungen reichen", kommentiert Francois [illegible]h, President von InfoSizing. "Mit der EDW-Referenzarchitektur könne sie nun nicht nur diese Menge bewältigen, sond[illegible] bei Bedarf auch dynamisch skalieren und Speicherkosten rapide senken", so Raab.

Mit der CMT-Technologie von Sun (Chip-Multithreading) und Sybase-IQ steigt die Abfrage- und Ladegeschwindigkeit der Daten laut Sybase auch dann nicht, wenn die Query-Ausgaberaten um das Fünfache ansteigen. Damit sollen Unternehmen Entscheidungen in Sekundenschnelle treffen können, wie es etwa für Finanztransaktionen notwendig ist. Weiters können mit der neuen Lösung laut Sybase die Speicherkosten gegenüber Wettbewerbsprodukten um bis zu 90 Prozent reduziert werden.

Mit einer Mio. Datenreihen kann das neue Data-Warehouse von Sybase und Sun rein rechnerisch genügend Daten aufnehmen, um die Historie aller Handelstransaktionen an allen Börsen der Welt zu verfolgen oder alle Kredit- und Debit-Transaktionen weltweit in den letzen sieben Jahren zu dokumentieren. Das Warehouse besteht aus Sun-Fire-Servern, Sun-StorEdge-Speichersystemen und der hochskalierbaren analytischen Engine Sybase-IQ.

▷ Diesen Artikel per E-Mail versenden

Anzeige



**Weitere Themen**

- Viren - Vorsicht vor gefährlichen Dateien
- Internet - Neues aus dem WWW

**High-Tech Extra**

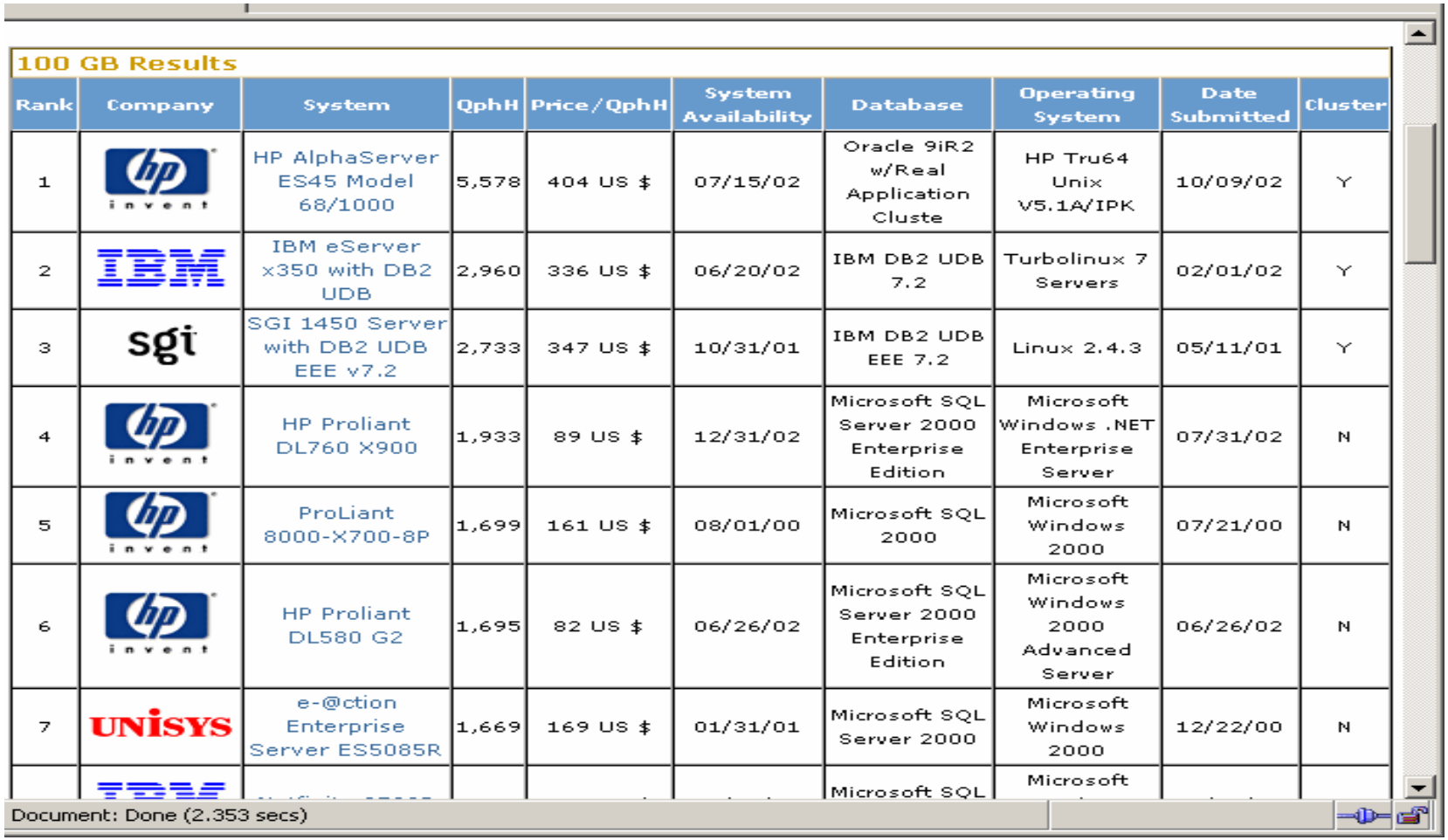
- Themen des Tages
- Pressemitteilungen

Done

# TPC Benchmarks

- Vergleich der Leistungsfähigkeit von Datenbanken
  - TPC-C OLTP Benchmark
  - TPC-H Ad-Hoc Decision Support (variable Anfragen)
  - TPC-R Reporting Decision Support (feste Anfragen)
  - TPC-W eCommerce Transaktionsverarbeitung
  - TPC-D Abgelöst durch H und R
- Vorgegebene Schemata (Lieferwesen)
- Schema-, Anfrage- und Datengeneratoren
- Unterschiedliche DB-Größen
  - TPC-H: 100 GB - 300 GB - 1 TB - 3 TB

# TPC-H Ergebnisse 100 GB



100 GB Results

| Rank | Company | System | QphH | Price/QphH | System Availability | Database | Operating System | Date Submitted | Cluster |
|---|---|---|---|---|---|---|---|---|---|
| 1 | hp invent | HP AlphaServer ES45 Model 68/1000 | 5,578 | 404 US $ | 07/15/02 | Oracle 9iR2 w/Real Application Cluste | HP Tru64 Unix V5.1A/IPK | 10/09/02 | Y |
| 2 | IBM | IBM eServer x350 with DB2 UDB | 2,960 | 336 US $ | 06/20/02 | IBM DB2 UDB 7.2 | Turbolinux 7 Servers | 02/01/02 | Y |
| 3 | sgi | SGI 1450 Server with DB2 UDB EEE v7.2 | 2,733 | 347 US $ | 10/31/01 | IBM DB2 UDB EEE 7.2 | Linux 2.4.3 | 05/11/01 | Y |
| 4 | hp invent | HP Proliant DL760 X900 | 1,933 | 89 US $ | 12/31/02 | Microsoft SQL Server 2000 Enterprise Edition | Microsoft Windows .NET Enterprise Server | 07/31/02 | N |
| 5 | hp invent | ProLiant 8000-X700-8P | 1,699 | 161 US $ | 08/01/00 | Microsoft SQL 2000 | Microsoft Windows 2000 | 07/21/00 | N |
| 6 | hp invent | HP Proliant DL580 G2 | 1,695 | 82 US $ | 06/26/02 | Microsoft SQL Server 2000 Enterprise Edition | Microsoft Windows 2000 Advanced Server | 06/26/02 | N |
| 7 | UNISYS | e-@ction Enterprise Server ES5085R | 1,669 | 169 US $ | 01/31/01 | Microsoft SQL Server 2000 | Microsoft Windows 2000 | 12/22/00 | N |
| | IBM | | | | | Microsoft SQL | Microsoft | | |

Document: Done (2.353 secs)

Quelle: http://www.tpc.org, Januar 2003

# TPC-H Ergebnisse 3000 GB

**3,000 GB Results**

| Rank | Company | System | QphH | Price/QphH | System Availability | Database | Operating System | Date Submitted | Cluster |
|---|---|---|---|---|---|---|---|---|---|
| 1 | hp invent | HP 9000 Superdome Enterprise Server | 27,094 | 240 US $ | 10/30/02 | Oracle 9i Database Enterprise Edition v9.2.0.2.0 | HP UX 11.i 64-bit | 10/04/02 | N |
| 2 | Sun microsystems | Sun Fire[TM] 15K Server with Oracle9i R2 | 23,813 | 237 US $ | 10/30/02 | Oracle 9i R2 Enterprise Edition | Sun Solaris 9 | 06/26/02 | N |
| 3 | hp invent | Compaq ProLiant DL760 X900-128P | 21,053 | 291 US $ | 06/20/02 | IBM DB2 UDB 7.2 | Microsoft Windows 2000 Advanced Server | 02/06/02 | Y |
| 4 | Teradata a division of NCR | WorldMark 5250 | 18,803 | 989 US $ | 07/27/01 | Teradata V2R4.1 | MP-RAS 3.02.00 | 10/09/01 | Y |
| 5 | hp invent | HP 9000 Superdome Enterprise Server | 17,908 | 569 US $ | 05/15/02 | Oracle 9i Database Enterprise Edition | HP UX 11.i 64-bit | 01/28/02 | N |

Quelle: http://www.tpc.org