Dr. Andreas Behrend

*Foundations of Information Systems*

*Winter Semester 2016/2017*

- I am indebted to Prof. Wolfgang Lehner for providing me slides -
Who for Whom

Lecturer
- PD Dr. Andreas Behrend
  behrend@iai.uni-bonn.de

Exercise
- M.A. Christiane Engels
  christiane.engels@googlemail.com
Classes and Exercises

- 1st semester master students of Life Science Informatics

- Class 2 SWS: Every Tuesday, 10:45 to 12:15 (room Rheinsaal)
- Exercise 2 SWS: Every Wednesday, 9:00 to 10:30 (room Rheinsaal)
- Credits 3, weekly lecture hours (WLH) 4

Lecture Notes

- Available at http://pages.iai.uni-bonn.de/behrend_andreas/lehre/FIM/WS16/
- Check homepage for news!
Basic Rules

Exercises
- You are strongly encouraged to attend the exercise sessions but it is not mandatory to do so.
- The written examination will be strongly influenced by the tasks discussed during the exercises.
- The solutions to the exercises will **not** be published.

Final exam
- Will be announced during the semester
- The written exam will last for 90 minutes.
> Spot the Database!
Databases
What are they good for?
Why Databases?

**Importance of database systems**
- Established research field since 1960
- Key technology just like:
  - operating systems
  - artificial intelligence
  - software engineering
- Vital commercial market
  - Fundamental for every data scientist

**Application scenarios for database systems**
- Almost every business (small/big) needs them
  - Enterprise resources, personnel administration
  - Customer data of banks, insurance companies, real estate
- Efficient storage and processing of large amounts of data
  - Structured and semi-structured
  - Bibliographic, document-text, statistical, multimedia

But how large is large?
Managing Large Amounts of Data

**Petabyte Age**

- **Estimated Size of Data (2015)**
  - Google: 15,000 PB
  - Facebook: 300 PB
  - Ebay: 90 PB

- **Data Processed per Day**
  - Google: 100 PB
  - Ebay: 100 PB
  - NSA: 29 PB
  - Facebook: 600 TB

1024 bytes = 1 KB
1024 KB = 1 MB
1024 MB = 1 GB
1024 GB = 1 TB
1024 TB = 1 PB

KB = Kilobyte
MB = Megabyte
GB = Gigabyte
TB = Terabyte
PB = Petabyte
Challenge

- Assume you got 1 PB data on disk
- Now do some analysis with it!

With 100MB/s disk, sequential search takes

\[ 1 \text{PB} \div 100 \text{ MB} \div 2 = 58 \text{ days} \]

Database systems can be much better but efficient searching is just one reason for using them!


> Attempt of a Definition

**What is a database (DB)?**
- Elmasri/Navathe: “A collection of Related Data” such that it...
  - ... represents a specific part of the real world
  - ... is logical consistent and has a specific meaning

**What is a database system (DBS)?**
- DBS comprises at least
  - An arbitrary number of databases
  - A generic software component called database management system (DBMS)
Principle Scheme

Applications and Users

Clients

Server

DBMS

DB1

DB2

DB3

DBS
Main Tasks of a DBMS

Aside from efficient data access, a DBMS supports:

- Data independence (logical/physical)
- Hardware independency
- Data security (consistency, access control, recovery)
- Concurrent data access (scheduling)
- Efficient data analysis (query language)

Transactions
QueryOptimizer
Data Independence

Problem

- Applications want to change the data structure over time:
  - New fields
  - Other data types
  - Different storage structure
- What happens to other applications still using the original data structure?

Solution

- DBMS manages different views of the data for applications
  - Logical data independence
- Changes in the physical storage layer is hidden for applications
  - Physical data independence

Diagram:

- External schema
  - View 1
  - ... View n
- Logical schema
- Physical schema
- ANSI-SPARC-Architecture
Concurrent Data Access

**Problem**
- Concurrent data access by different clients may lead to dirty data

**Solution**
- Automatic synchronization of clients for creating the Illusion of an
  - exclusive and
  - sequential data access

⇒ Transaction concept!
Data Analysis

Problem

- Moving large amounts of data is infeasible
- How to perform data analysis within the DBS?

Solution

- Provide expressive database languages for deriving analysis results
- Flexible query language which do not interfere logical/physical independence

⇒ Declarative programming languages like SQL

CREATE VIEW cardiac_arrest AS
SELECT patientId, heartRate, bloodPressure
FROM patientData pd
WHERE
  bloodPressure > 80 AND
  heartRate > 120 AND
  NOT EXISTS
    (SELECT *
     FROM patientData
     WHERE patientID=pd.patientID AND
           bloodPressure <= 120 AND
           timestamp>NOW()-timestamp1)
  AND
    (bloodPressure < 0.8 *
     (SELECT max(bloodPressure)
      FROM patientData
      WHERE ...));
Different Databases
What are relational ones?
Why relational Databases?

Relational databases are the foundation of the western civilization.

Bruce Lindsay,
IBM Fellow @ IBM Almaden Research Center
Data Models

Data Model

- Collection of concepts for structuring data
  - Tables, attributes, keys in the relational model
  - Property graphs in graph-oriented models
  - Classes in object-oriented models
- Structuring data according to a given data model leads to a schema

Classification of DBS

- Generic DBS are typically classified with respect to the supported data model
  - Relational databases
  - Object oriented databases
  - XML databases
  - Graph databases
  - ...
The Relational Data Model

- The idea to organize data in tables is quite old and pretty obvious.
- The idea to investigate this representation of data by means of the theory of relations is due to Edgar F. Codd, who proposed this view at the end of the 1960s:

For this pioneering work Codd received the Turing Award in 1982, the „Nobel price of informatics".
Summary

Main Tasks of a DBS

- Data independence (logical/physical)
- Hardware independency
- Data security (consistency, access control, recovery)
- Concurrent data access (scheduling)
- Efficient data analysis (query language)

Further Notions

- Database (DB)
- Database System (DBS)
- Client/Server approach
- Declarative languages
- Data Model
- Database Management System (DBMS)
- Transactions
- Schema
What is in the Lecture?

1. Database Usage
   - Query
   - Programming
   - Design

2. Database Architecture
   - Indexes
   - Transactions
   - Query Processing

3. Database Scaling
   - Scale-out
   - Scale-in
Table of Content (tentative)

- Database Design and Entity Relationship Model
- Relational Data Model and Relational Algebra
- Structured Query Language (SQL)
- DB2 Architecture and Tablespaces
- User-defined Functions, Stored Procedures and Trigger
- JDBC, ODBC and .NET
- SQLJ, Embedded SQL, External Functions and external Stored Procedures
- XML und XQuery
- Query Processing and –optimization
- Transactions and Synchronization
- Backup, Recovery and High Availability
- Monitoring and Access Control
- Database support for Data-Warehousing (OLAP)
> Literature and Copyright

---

**Basic literature**

- J. Hoffer:, M. Prescott, H. Topi: „Modern Database Management” (9th Edition)

---

**Copyright Note**

- All material may contain under consideration of other copyright constraints examples and figures from additional sources
- The (commercial) use of the lecture notes in business not related to the University of Bonn is explicitly not allowed