Databases I

Etelka Szendrői Dr. (PhD)
associate professor
System and Software Technology Department

szendroi@mik.pte.hu

Objectives

In this lecture, you will learn:

- The difference between data and information
- What a database is, the various types of databases, and why they are valuable assets for decision making
- The importance of database design
- How modern databases evolved from file systems
- About flaws in file system data management
- The main components of the database system
- The main functions of a database management system (DBMS)

Introduction

- Good decisions require good information derived from raw facts
- Data is managed most efficiently when stored in a database
- Databases evolved from computer file systems
- Understanding file system characteristics is important

Why Databases?

- Databases solve many of the problems encountered in data management
  - Used in almost all modern settings involving data management:
    - Business
    - Research
    - Administration
- Important to understand how databases work and interact with other applications

Data vs. Information

- **Data** are raw facts
- **Information** is the result of processing raw data to reveal meaning
- Information **requires context** to reveal meaning
- **Raw data must be formatted** for storage, processing, and presentation
- Data are the foundation of information, which is the bedrock of knowledge

Data vs. Information (cont'd.)

- Data: building blocks of information
- Information produced by processing data
- Information used to reveal meaning in data
- Accurate, relevant, timely information is the key to good decision making
- Good decision making is the key to organizational survival
- Data management: focuses on proper generation, storage, and retrieval of data

Introducing the Databases

- Database: shared, integrated structure that stores a collection of:
  - End-user data: raw facts of interest to end user
  - Metadata: data about data
    - Provides description of data characteristics and relationships in data
    - Complements and expands value of data
- Database management system (DBMS): collection of programs
  - Manages structure and controls access to data
  - Maintains and preserves data

Database Management Systems

- Tasks of DBMS:
  - To create databases
  - To define the content of databases
  - To store data
  - To query data; optimize the query process
  - To modify data
  - To preserve, safe data
  - To encript data
  - To maintain the rights and access

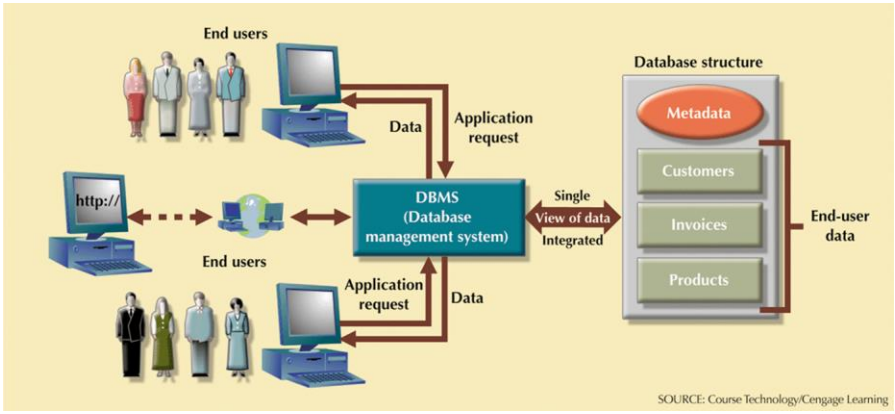
Role and Advantages of the DBMS

- DBMS is the intermediary between the user and the database
  - Database structure stored as file collection
  - Can only access files through the DBMS
- DBMS enables data to be shared
- DBMS integrates many users' views of the data

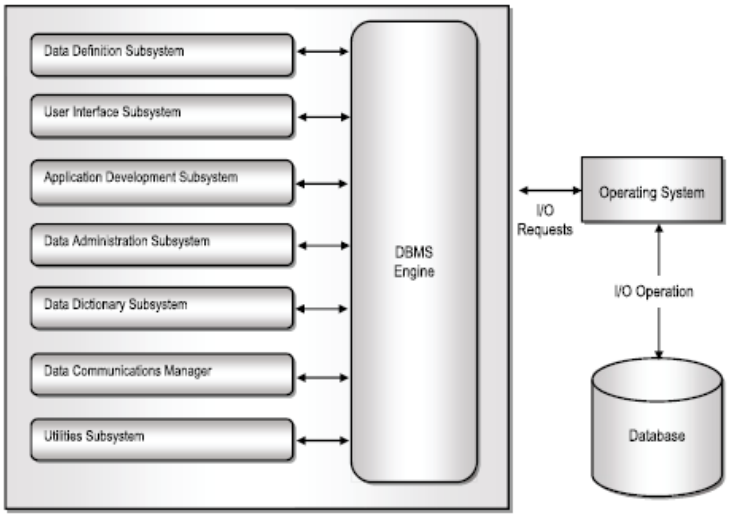
Role and Advantages of the DBMS (cont'd.)

- Advantages of a DBMS:
  - Improved data sharing
  - Improved data security
  - Better data integration
  - Minimized data inconsistency
  - Improved data access
  - Improved decision making
  - Increased end-user productivity

The DBMS manages the interaction between the end user and the database



Functional components of DBMS



DBMS Functions (cont'd.)

- Data storage management
  - DBMS creates and manages complex structures required for data storage
  - Also stores related data entry forms, screen definitions, report definitions, etc.
  - Performance tuning: activities that make the database perform more efficiently
  - DBMS stores the database in multiple physical data files

DBMS Functions (cont'd.)

- Data transformation and presentation
  - DBMS transforms data entered to conform to required data structures
  - DBMS transforms physically retrieved data to conform to user's logical expectations
- Security management
  - DBMS creates a security system that enforces user security and data privacy
  - Security rules determine which users can access the database, which items can be accessed, etc.

DBMS Functions (cont'd.)

- Multiuser access control
  - DBMS uses sophisticated algorithms to ensure concurrent access does not affect integrity
- Backup and recovery management
  - DBMS provides backup and data recovery to ensure data safety and integrity
  - Recovery management deals with recovery of database after a failure
    - Critical to preserving database's integrity

DBMS Functions (cont'd.)

- Data integrity management
  - DBMS promotes and enforces integrity rules
    - Minimizes redundancy
    - Maximizes consistency
  - Data relationships stored in data dictionary used to enforce data integrity
  - Integrity is especially important in transaction-oriented database systems

DBMS Functions (cont'd.)

- Database access languages and application programming interfaces
  - DBMS provides access through a query language
  - Query language is a nonprocedural language
  - Structured Query Language (SQL) is the de facto query language
    - Standard supported by majority of DBMS vendors

DBMS Functions (cont'd.)

- Database communication interfaces
  - Current DBMSs accept end-user requests via multiple different network environments
  - Communications accomplished in several ways:
    - End users generate answers to queries by filling in screen forms through Web browser
    - DBMS automatically publishes predefined reports on a Web site
    - DBMS connects to third-party systems to distribute information via e-mail

The Database System

- Database system: the collection of Database Managements System, databases and user applications which defines and regulates the collection, storage, management, use of data
- Five major parts of a database system:
  - Hardware
  - Software
  - People
  - Procedures
  - Data

Database System (DBS)

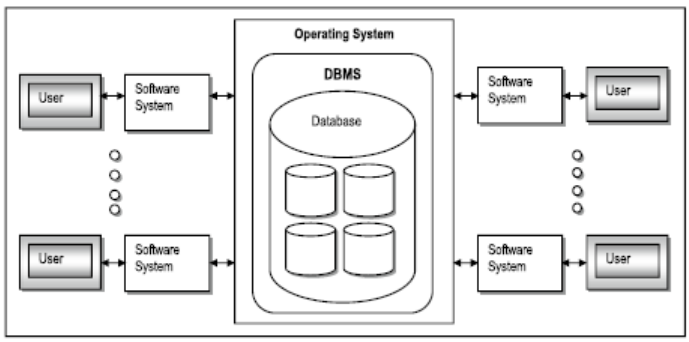
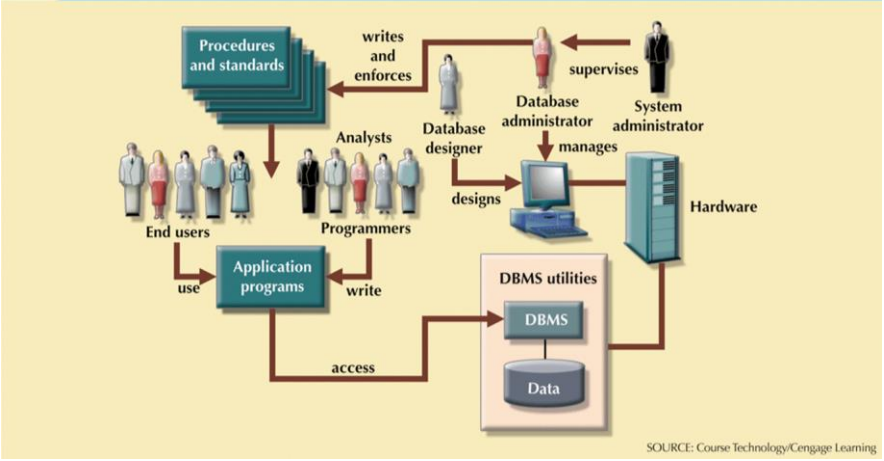


Figure 1-1. Simplified Representation of a DBS

The database system environment



The Database System Environment (cont'd.)

- Hardware: all the system's physical devices
- Software: three types of software required
  - Operating system software
  - DBMS software
  - Application programs and utility software

The Database System Environment (cont'd.)

- People: all users of the database system
  - System and database administrators
  - Database designers
  - Systems analysts and programmers
  - End users
- Procedures: instructions and rules that govern the design and use of the database system
- Data: the collection of facts stored in the database

The Database System Environment (cont'd.)

- Database systems are created and managed at different levels of complexity
- Database solutions must be cost-effective as well as tactically and strategically effective
- Database technology already in use affects selection of a database system

Types of Databases

- Databases can be classified according to:
  - Number of users
  - Database location(s)
  - Expected type and extent of use
- Single-user database supports only one user at a time
  - Desktop database: single-user; runs on PC
- Multiuser database supports multiple users at the same time
  - Workgroup and enterprise databases

Types of Databases (cont'd.)

- Centralized database: data located at a single site
- Distributed database: data distributed across several different sites
- Operational database: supports a company's day-to-day operations
  - Transactional or production database
- Data warehouse: stores data used for tactical or strategic decisions

Types of Databases (cont'd.)

- Unstructured data exist in their original state
- Structured data result from formatting
  - Structure applied based on type of processing to be performed
- Semi-structured data have been processed to some extent
- Extensible Markup Language (XML) represents data elements in textual format
  - XML database supports semi-structured XML data

Types of Databases

PRODUCT	NUMBER OF USERS			DATA LOCATION		DATA USAGE		XML
	SINGLE USER	MULTIUSER						
		WORKGROUP	ENTERPRISE	CENTRALIZED	DISTRIBUTED	OPERATIONAL	ANALYTICAL	
MS Access	X	X		X		X		
MS SQL Server	X[3]	X	X	X	X	X	X	X
IBM DB2	X[3]	X	X	X	X	X	X	X
MySQL	X	X	X	X	X	X	X	X
Oracle RDBMS	X[3]	X	X	X	X	X	X	X

[3] Vendor offers single-user/personal DBMS version

Why Database Design Is Important

- Database design focuses on design of database structure used for end-user data
  - Designer must identify database's expected use
- Well-designed database:
  - Facilitates data management
  - Generates accurate and valuable information
- Poorly designed database:
  - Causes difficult-to-trace errors

Managing the Database System: A Shift in Focus

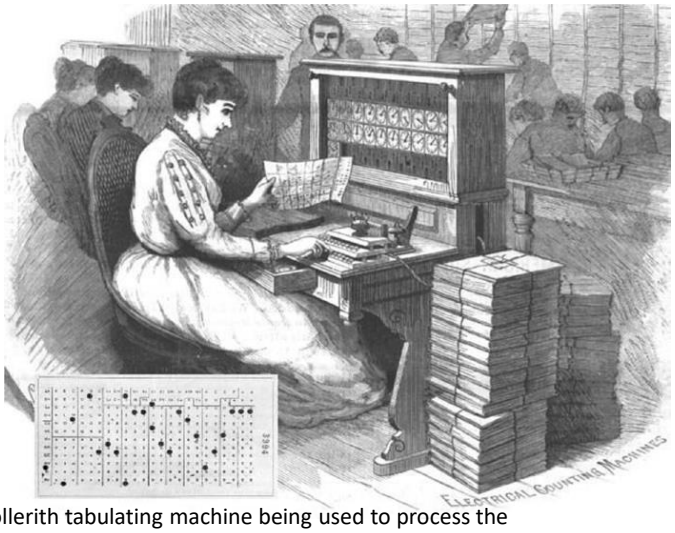
- Database system provides a framework in which strict procedures and standards enforced
  - Role of human changes from programming to managing organization's resources
- Database system enables more sophisticated use of the data
- Data structures created within the database and their relationships determine effectiveness

Managing the Database System: A Shift in Focus (cont'd.)

- Disadvantages of database systems:
  - Increased costs
  - Management complexity
  - Maintaining currency
  - Vendor dependence
  - Frequent upgrade/replacement cycles

History of DBMS

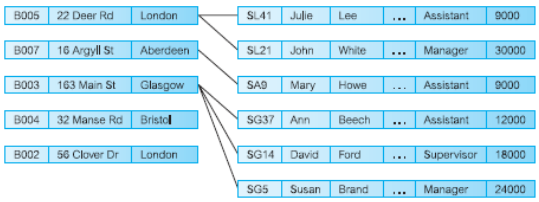
Processing data in 19th Century



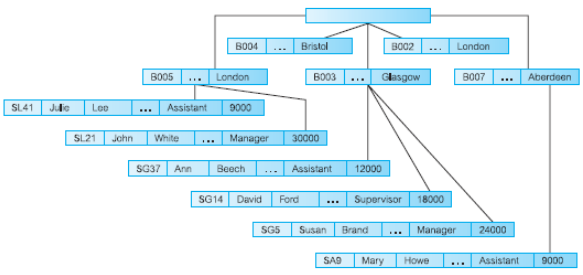
a Hollerith tabulating machine being used to process the U.S. census in 1890.

The first database revolution

Network data model
records appearing as nodes (also called segments) and sets as edges in the graph



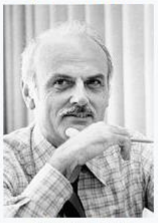
A hierarchical model can be represented as a tree graph, with records appearing as nodes (also called segments) and sets as edges. The hierarchical model allows a node to have only one parent



Second database revolution

- Edgar Codd (IBM), 1970 : *A Relational Model of Data for Large Shared Data Banks*.
- This classic paper contained the core ideas that defined the *relational database model* that became the most significant—almost universal—model for database systems for a generation.

Edgar "Ted" Codd



Born	Edgar Frank Codd 19 August 1923[1][2] Fortuneswell, Dorset, England
Died	18 April 2003 (aged 79) Williams Island, Aventura, Florida, USA

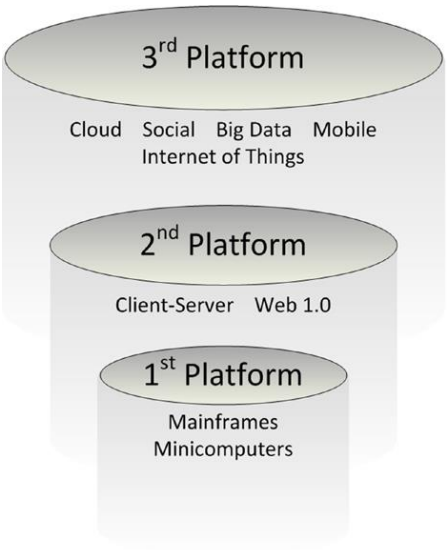
35

The Third Database Revolution

- Google and Hadoop
- By 2005, Google was by far the biggest website in the world
- In 2003, Google revealed details of the distributed file system GFS that formed a foundation for its storage architecture and in 2004 it revealed details of the distributed parallel processing algorithm *MapReduce*, which was used to create World Wide Web indexes. In 2006, Google revealed details about its *BigTable* distributed structured database.
- This concept formed the basis for the *Hadoop* project.

36

“Three platforms” model corresponds to three waves of database technology



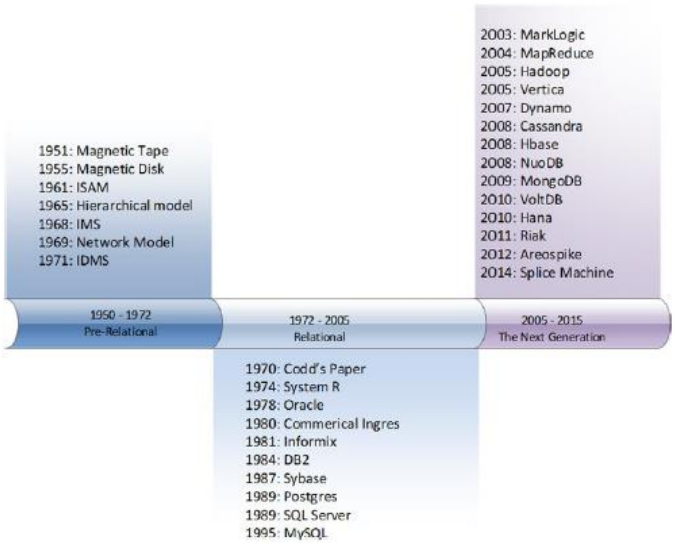
Relational Database
NoSQL
NewSQL
Big Data platforms

Relational Database

Hierarchical Database
Network Database
ISAM files

37

Timeline of major database releases and innovations



38

Forrás: Guy Harrison: Next Generation Databases, 2015, Apress

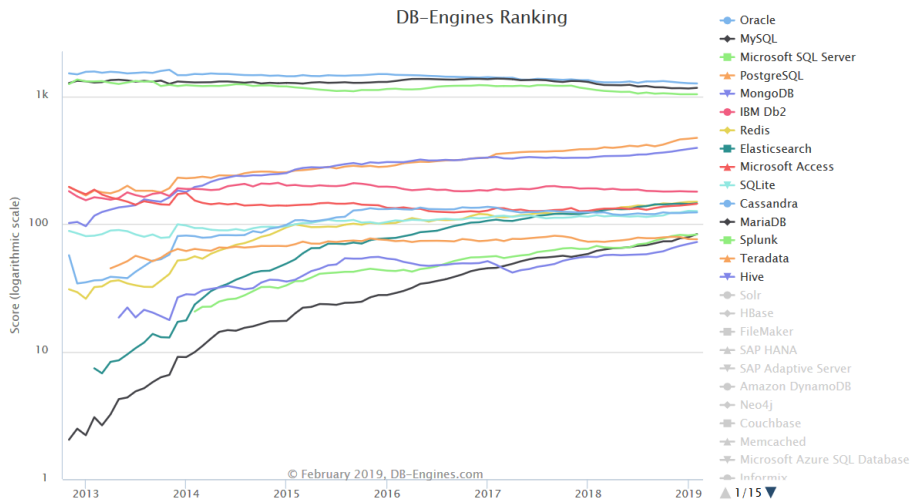
DB-Engines Ranking

343 systems in ranking, February 2019

Rank Feb 2019	Rank Jan 2019	Rank Feb 2018	DBMS	Database Model	Score Feb 2019	Score Jan 2019	Score Feb 2018
1.	1.	1.	Oracle	Relational DBMS	1264.02	-4.82	-39.26
2.	2.	2.	MySQL	Relational DBMS	1167.29	+13.02	-85.18
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1040.05	-0.21	-81.98
4.	4.	4.	PostgreSQL	Relational DBMS	473.56	+7.45	+85.18
5.	5.	5.	MongoDB	Document store	395.09	+7.91	+58.67
6.	6.	6.	IBM Db2	Relational DBMS	179.42	-0.43	-10.55
7.	7.	↑ 8.	Redis	Key-value store	149.45	+0.43	+22.43
8.	8.	↑ 9.	Elasticsearch	Search engine	145.25	+1.81	+19.93
9.	9.	↓ 7.	Microsoft Access	Relational DBMS	144.02	+2.41	+13.95
10.	10.	↑ 11.	SQLite	Relational DBMS	126.17	-0.63	+8.89
11.	11.	↓ 10.	Cassandra	Wide column store	123.37	+0.39	+0.59
12.	↑ 13.	↑ 17.	MariaDB	Relational DBMS	83.42	+4.60	+21.77
13.	↓ 12.	13.	Splunk	Search engine	82.81	+1.39	+15.55
14.	14.	↓ 12.	Teradata	Relational DBMS	75.97	-0.22	+2.98
15.	15.	↑ 18.	Hive	Relational DBMS	72.29	+2.38	+17.23
16.	16.	↓ 14.	Solr	Search engine	60.96	-0.52	-2.91
17.	17.	↓ 16.	HBase	Wide column store	60.28	-0.12	-1.43
18.	18.	↑ 19.	FileMaker	Relational DBMS	57.79	+0.64	+3.43
19.	19.	↑ 20.	SAP HANA	Relational DBMS	56.55	-0.09	+9.19
20.	↑ 21.	↓ 15.	SAP Adaptive Server	Relational DBMS	55.75	+0.71	-7.74
21.	↓ 20.	21.	Amazon DynamoDB	Multi-model	54.95	-0.15	+15.07
22.	22.	22.	Neo4j	Graph DBMS	47.86	+1.06	+8.04

39

Source: db-engines.com

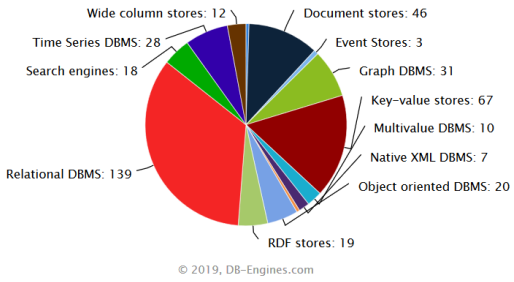


40

DBMS popularity according to the database model

DBMS popularity broken down by database model

Number of systems per category, February 2019

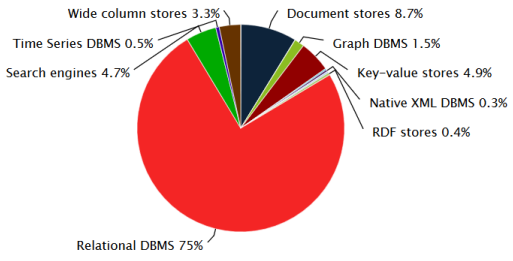


DB-Engines lists 343 different database management systems, which are classified according to their database model (e.g. relational DBMS, key-value stores etc.).
This pie-chart shows the number of systems in each category. Some of the systems belong to more than one category.

41

The popularity of data model categories in percent

Ranking scores per category in percent, February 2019



This chart shows the popularity of each category. It is calculated with the popularity (i.e. the ranking scores) of all individual systems per category. The sum of all ranking scores is 100%.

42

Popularity rank of commercial and open source systems

The top 5 commercial systems, February 2019

Rank	System	Score	Overall Rank
1.	Oracle	1264	1.
2.	Microsoft SQL Server	1040	3.
3.	IBM Db2	179	6.
4.	Microsoft Access	144	9.
5.	Splunk	83	13.

The top 5 open source systems, February 2019

Rank	System	Score	Overall Rank
1.	MySQL	1167	2.
2.	PostgreSQL	474	4.
3.	MongoDB	395	5.
4.	Redis	149	7.
5.	Elasticsearch	145	8.

43

Summary

- Data are raw facts
- Information is the result of processing data to reveal its meaning
- Accurate, relevant, and timely information is the key to good decision making
- Data are usually stored in a database
- DBMS implements a database and manages its contents

Summary (cont'd.)

- Metadata is data about data
- Database design defines the database structure
  - Well-designed database facilitates data management and generates valuable information
  - Poorly designed database leads to bad decision making and organizational failure
- Databases evolved from manual and computerized file systems

Summary (cont'd.)

- In a file system, data stored in independent files
  - Each requires its own management program
- Some limitations of file system data management:
  - Requires extensive programming
  - System administration is complex and difficult
  - Changing existing structures is difficult
  - Security features are likely inadequate
  - Independent files tend to contain redundant data
    - Structural and data dependency problems

Summary (cont'd.)

- Database management systems were developed to address file system's inherent weaknesses
- DBMS present database to end user as single repository
  - Promotes data sharing
  - Eliminates islands of information
- DBMS enforces data integrity, eliminates redundancy, and promotes security

Thank you for your attention!