Data and Datasets

AI 4 Business

Co-funded by the European Union

Outline

1. Basic Terminology
2. The Database Approach
3. Using a Database System

Co-funded by the European Union

Basic Terminology

Co-funded by the European Union

Most Basic Terminology

Data: known facts that can be recorded
and that have implicit meaning

Database: collection of related data (logically coherent)

- Represents some aspects of the real world (**miniworld**)
- Built for a specific purpose

Examples of large databases

- Amazon.com's product data
- Data collection underlying Webreg

Co-funded by
the European Union

4

Example of a Database

COURSE

Course_name	Course_number	Credit_hours	Department
Intro to Computer Science	CS1310	4	CS
Data Structures	CS3320	4	CS
Discrete Mathematics	MATH2410	3	MATH
Database	CS3380	3	CS

SECTION

Section_identifier	Course_number	Semester	Year	Instructor
85	MATH2410	Fall	04	King
92	CS1310	Fall	04	Anderson
102	CS3320	Spring	05	Knuth
112	MATH2410	Fall	05	Chang
119	CS1310	Fall	05	Anderson
135	CS3380	Fall	05	Stone

GRADE_REPORT

Student_number	Section_identifier	Grade
17	112	B
17	119	C
8	85	A
8	92	A
8	102	B
8	135	A

PREREQUISITE

Course_number	Prerequisite_number
CS3380	CS3320
CS3380	MATH2410
CS3320	CS1310

Co-funded by the European Union

Terminology (cont'd)

Database management system (DBMS)

- Collection of computer programs
- Enables users to create and maintain a database (DB)
- Supports concurrent access to a database by multiple users and programs
- Protects the DB against unauthorized access and manipulation
- Provides means to evolve the DB as requirements change

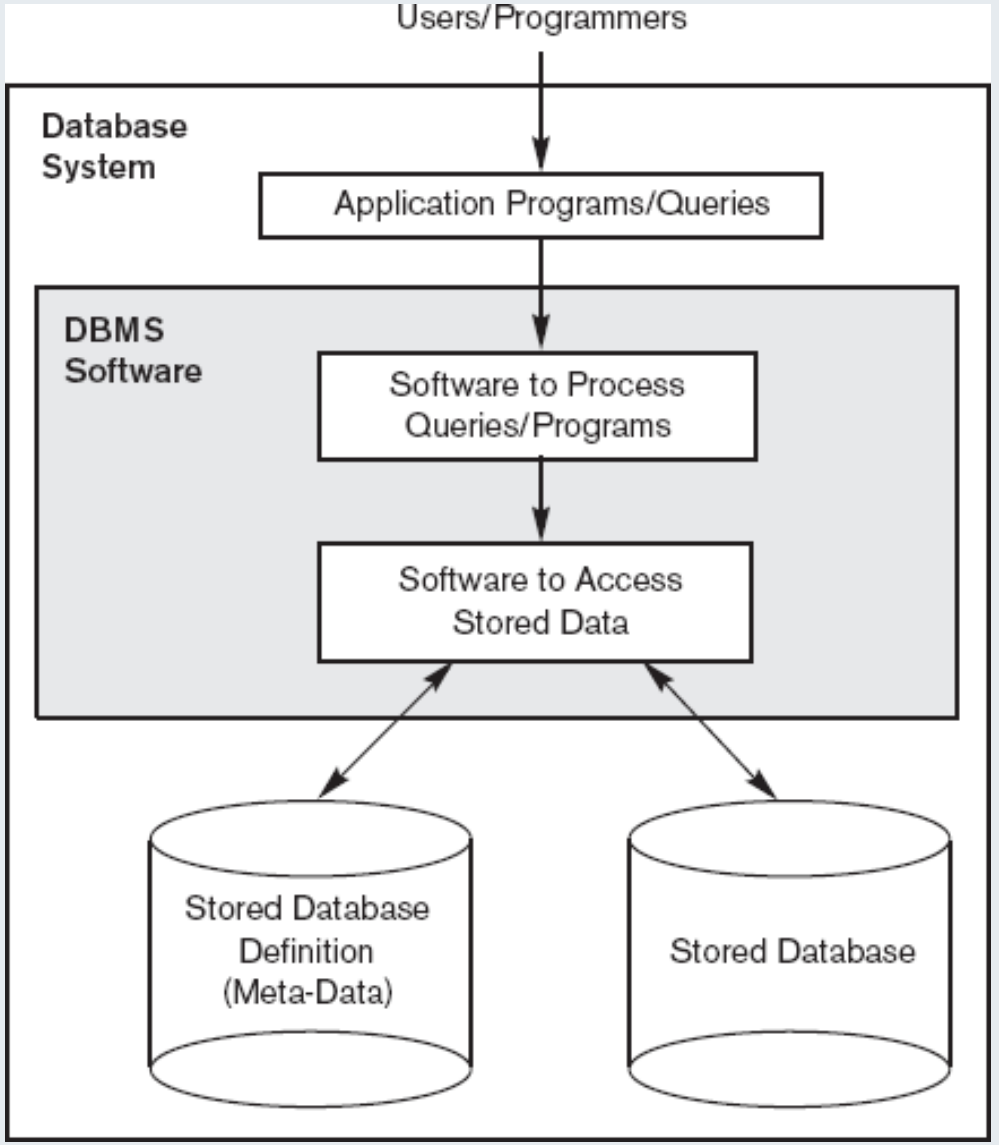
Examples of database management systems

- IBM's DB2, Microsoft's Access, Microsoft's SQL Server, Oracle, SAP's SQL Anywhere, MySQL, PostgreSQL

Co-funded by the European Union

6

Database System



Co-funded by the European Union

The Database Approach

Co-funded by
the European Union

Pre-DBMS Data Management

Used traditional **file processing**

- Each user defines and implements the files needed for a specific software application

https://www.goodfreephotos.com/albums/other-photos/boxes-and-boxes-moving-storage.jpg

As the application base grows

- many shared files
- a multitude of file structures
- a need to exchange data among applications

Co-funded by the European Union

Problems of Pre-DBMS Data Management


https://cdn.pixabay.com/photo/2014/06/01/22/26/clutter-360058_960_720.jpg

- Redundancy: multiple copies
- Inconsistency: independent updates
- Inaccuracy: concurrent updates
- Incompatibility: multiple formats
- Insecurity: proliferation
- Inauditability: poor chain of responsibility
- Inflexibility: changes are difficult to apply

Co-funded by the European Union

Database Approach

- Eventually recognized that data is a critical corporate asset (along with capital and personnel)
  - Need to manage the data in a more systematic manner

- Database approach: Use a *single repository* to maintain data that is defined once and accessed by various users
  - Addresses the aforementioned problems

https://cdn.pixabay.com/photo/2017/06/12/04/21/database-2394312_960_720.jpg

Co-funded by the European Union

Characteristics of the Database Approach

- Programs isolated from data through **abstraction**
  - DBMS does not expose details of how (or where) data is stored or how operations are implemented
  - Programs refer to an abstract model of the data, rather than data storage details
  - Data structures and storage organization can be changed without having to change the application programs
- Support of multiple **views** of the data
  - Different users may see different views of the database,which contain only the data of interest to these users
- Multi-user **transaction** processing
  - Encapsulates sequence of operations to behave atomically
    - e.g., transferring funds

Co-funded by the European Union

12

Characteristics of the Database Approach

- Data is **self-describing**
  - Database system contains a *database catalog* with meta-data that describes structure and constraints of the database(s)
  - Database catalog used by DBMS, and by DB users who need information about DB structure
  - Example:

RELATIONS

Relation_name	No_of_columns
STUDENT	4
COURSE	4
SECTION	5
GRADE_REPORT	3
PREREQUISITE	2

COLUMNS

Column_name	Data_type	Belongs_to_relation
Name	Character (30)	STUDENT
Student_number	Character (4)	STUDENT
Class	Integer (1)	STUDENT
Major	Major_type	STUDENT
Course_name	Character (10)	COURSE
Course_number	XXXXNNNN	COURSE
….	….	…..
….	….	…..
….	….	…..
Prerequisite_number	XXXXNNNN	PREREQUISITE

Example from "Fundamentals of Database Systems" by Elmasri and Navathe, Addison Wesley.

Co-funded by the European Union

Using a Database System

Digital 4 Business
Evolving your digital future

Co-funded by the European Union

Defining a Database

Specifying the data types, structures, and constraints of the data to be stored

Uses a *Data Definition Language (DDL)*

Meta-data: Database definition or descriptive information

- Stored by the DBMS in a *database catalog* or *data dictionary*

Phases for designing a database:

- **Requirements specification and analysis**
- **Conceptual design**
  - e.g., using the *Entity-Relationship model*
- **Logical design**
  - e.g., using the *relational model*
- **Physical design**

Co-funded by the European Union

15

Database System Design Process

Two main activities:

- **Database design** focuses on defining the database
- **Application design** focuses on the programs and interfaces that access the database (out of scope of this lecture)

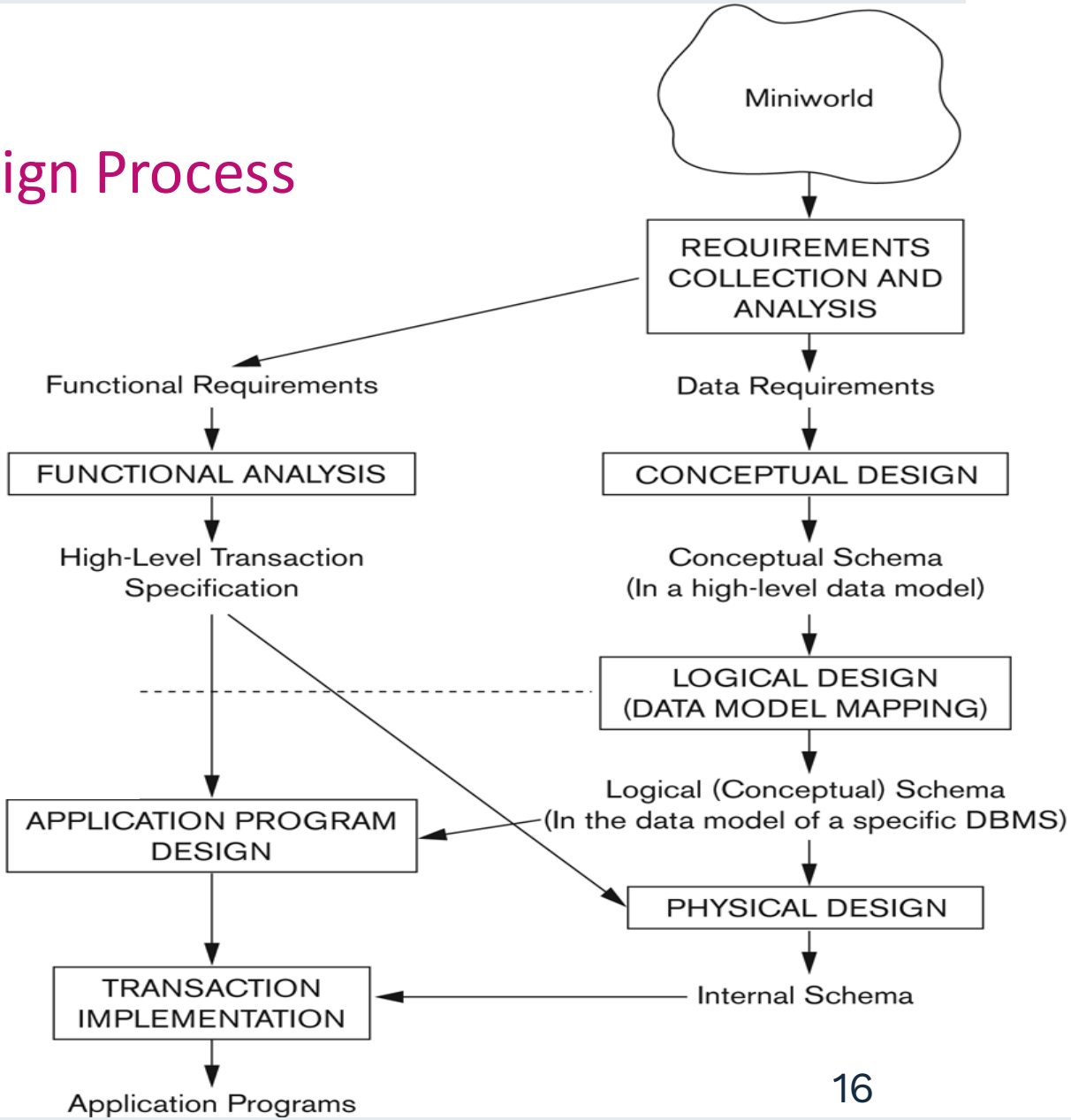


Figure from "Fundamentals of Database Systems" by Elmasri and Navathe, Addison Wesley.

Co-funded by the European Union

Example of Data Requirements

A taxi company needs to model their activities.

There are two types of employees in the company: drivers and operators. For drivers it is interesting to know the date of issue and type of the driving license, and the date of issue of the taxi driver's certificate. For all employees it is interesting to know their personal number, address and the available phone numbers.

The company owns a number of cars. For each car there is a need to know its type, year of manufacturing, number of places in the car and date of the last service.

The company wants to have a record of car trips. A taxi may be picked on a street or ordered through an operator who assigns the order to a certain driver and a car. Departure and destination addresses together with times should also be recorded.

Co-funded by the European Union

17

Another Example

Movie database: information concerning movies, actors, awards

Data records

- Film
- Person
- Role
- Honors

Define structure of each type of data record by specifying **data elements** to include and **data type** for each element

- String (sequence of alphabetic characters)
- Numeric (integer or real)
- Date (year or year-month-day)
- Monetary amount
- etc.

Co-funded by the European Union

18

Using a Database

Populating a DB: Inserting data to reflect the miniworld

- e.g., store data to represent each film, actor, role, director, etc

Film

title	genre	year	director	runtime	budget	gross
The Company Men	drama	2010	John Wells	104	15,000,000	4,439,063
Lincoln	biography	2012	Steven Spielberg	150	65,000,000	181,408,467
War Horse	drama	2011	Steven Spielberg	146	66,000,000	79,883,359
Argo	drama	2012	Ben Affleck	120	44,500,000	135,178,251
Fire Sale	comedy	1977	Alan Arkin	88	1,500,000	0

Person

name	birth	city
Ben Affleck	1972	Berkeley
Alan Arkin	1934	New York
Tommy Lee Jones	1946	San Saba
John Wells	1957	Alexandria
Steven Spielberg	1946	Cincinnati
Daniel Day-Lewis	1957	Greenwich

Honors

movie	award	category	winner
Lincoln	Critic's Choice	actor	Daniel Day-Lewis
Argo	Critic's Choice	director	Ben Affleck
Lincoln	Screen Actors Guild	supporting actor	Tommy Lee Jones
Lincoln	Screen Actors Guild	actor	Daniel Day-Lewis
Lincoln	Critic's Choice	screenplay	Tony Kushner
Argo	Screen Actors Guild	cast	Argo
War Horse	BMI Flim	music	John Williams

Role

actor	movie	persona
Ben Affleck	Argo	Tony Mendez
Alan Arkin	Argo	Lester Siegel
Ben Affleck	The Company Men	Bobby Walker
Tommy Lee Jones	The Company Men	Gene McClary
Tommy Lee Jones	Lincoln	Thaddeus Stevens
Alan Arkin	Fire Sale	Ezra Fikus
Daniel Day-Lewis	Lincoln	Abraham Lincoln

Co-funded by the European Union

Using a Database (cont'd)

Populating a DB: Inserting data to reflect the miniworld

Query: Interaction causing some data to be retrieved

- Uses a *Query Language*

Examples of queries:

- List the cast of characters for *Lincoln*.
- Who directed a *drama* in *2012*?
- Who directed a film in which he or she also played a role?
- What awards were won by *War Horse*?

Co-funded by the European Union

Using a Database (cont'd)

Populating a DB: Inserting data to reflect the miniworld

Query: Interaction causing some data to be retrieved

- Uses a *Query Language*

Manipulating a DB

- Querying and updating the DB to understand/reflect miniworld
- Generating reports
- Uses a *Data Manipulation Language (DML)*

Examples of updates:

- Record that *Argo* won a Golden Globe award for best picture.
- Add another $395,533 to the gross earnings for *Lincoln*.
- Change the birthplace for *Daniel Day-Lewis* to *London*.
- Delete all data about the movie *Fire Sale* from the database.

Co-funded by the European Union

Using a Database (cont'd)

Populating a DB: Inserting data to reflect the miniworld

Query: Interaction causing some data to be retrieved

- Uses a *Query Language*

Manipulating a DB

- Querying and updating the DB to understand/reflect miniworld
- Generating reports
- Uses a *Data Manipulation Language (DML)*

Application program

- Accesses DB by sending queries and updates to DBMS

Co-funded by the European Union

Reorganizing a Database

Changes the metadata rather than the data

More drastic than data updates

- May require massive changes to the data
- May require changes to some application programs

Uses the *Data Definition Language (DDL)* again

Examples:

- Move *director* from FILM to a separate relation DIRECTOR with columns for *person* and *movie*
- Change *birth* from *yyyy* to *yyyy/mm/dd*
- Split name in PERSON to separate *surname* from *given names*.
- Include data element *movieID* in FILM (to accommodate remakes and other duplications of film title); update other relations accordingly

Co-funded by the European Union

23