Database Systems

History of databases, terms and data models

Bence Molnár

2022.02.21.

BUDAPESTI MŰSZAKI
ÉS GAZDASÁGTUDOMÁNYI EGYETEM
Építőmérnöki Kar - építőmérnöki képzés 1782 óta

Fotogrammetria és Térinformatika Tanszék

Bence Molnár

molnar.bence@emk.bme.hu

E-mail subject prefix: [ABR]

Office: K. 131.

Consultation: Wednesday 15:00-16:00

Agenda

- Homework
- Where do we use databases day-by-day?
- Spreadsheets vs. Databases
- Terms
  - *information, data, data model*
  - *database*
  - *database management system*
- History of databases

Homework stages

- **Specification**: definition of goals and requirements
- **Design**: alphanumerical database
- **Realization**: creating the database based on design, upload of data and creating queries in MS Access

HOMEWORK - SYSTEM

Online interface

- 3 tasks
- Project based, cooperation, continuing colleagues work
- Deadlines enables to deliver the project on time.
- You can communicate at the online interface
- Anonymous communication
- Same level of difficulty
- Tracking

https://abr.fmt.bme.hu/sample#concept

Task 1

Opens: 2nd week

Submission: 4th week, Friday noon

Late submission: 5th week, Friday noon

Project specification, for details, see website:

https://abr.fmt.bme.hu/sample#task1

https://abr.fmt.bme.hu/sample/checklist

Task 2

Opens: 5th week

Submission: 10th week, Friday noon

Late submission: 11th week, Friday noon

Database schema design

https://abr.fmt.bme.hu/sample#task2

https://abr.fmt.bme.hu/sample/checklist

Task 3

Opens: 10th week

Submission: 14th week, Friday noon

Late submission: repetition week, Friday noon

Realization in MS Access

https://abr.fmt.bme.hu/sample#task3

https://abr.fmt.bme.hu/sample/checklist

Homework - requirements

- Fits to your professional field
- Enables to upload of at least 100 records to the database
- Raw, unorganized data at Task 1
- Might be fictitious, generated data, but you can use already existing data: catalogues, statistical data
- Data format, data exchange
- Clear wording

Topics hints

- Daily updated (modified) catalogue, where we need to create a report day-by-day
- We have multiple data sources and they are linked with clear reference on each other
- Each row in dataset is unique, but in some columns we can find repeated values (this enables grouping/categorization of data).

Topic examples

- Storage inventory
  - *Basic components, subject of regular orders. Properties (e.g. amount or price) changes in time.*
- Order management
- Measurements, monitoring
  - *Regular performed with similar circumstances*
- Experiments
- Business systems
- Coworker administration system for work location, workdays and vacations, used devices/units/machines
- Company tool catalogue
- Project administration
- Maintenance (e.g. building, BIM)

HOMEWORK - COMMENTS

- Topic should be a one-time performed design task
- In access, we cannot perform geometric analysis
- Checklist - https://abr.fmt.bme.hu/sample/checklist
- Raw data vs prepared and structured data
- Avoid accents and special characters in filenames
- Use Zip archiving format instead of rar
- 4 entities, 3 analysis, data amount
- Requirements enables to keep everyone's workload balanced

Test

11th week

Written, in person (practice time)

Results will be available at the website

Retake will organized on repetition week

Where do we use databases day-by-day?

BUDAPESTI MŰSZAKI
ÉS GAZDASÁGTUDOMÁNYI EGYETEM
Építőmérnöki Kar - építőmérnöki képzés 1782 óta

Fotogrammetria és Térinformatika Tanszék

Where do we use databases day-by-day?

- Where?
  - *everywhere…*
  - *Online shopping, traveling, phone calls, social media, sporting, etc.*
- Importance
  - *Knight Capital Group (trading software issue, 2012)*
  - …
- Data is the new gold

Where do we use databases day-by-day?

Tons of data

- Every 2 days we create as much information as we did from the beginning of time until 2003.
- Over 90% of all the data in the world was created in the past 2 years
- The total amount of data being captured and stored by industry doubles every 1.2 years
- Every minute we send 204 million emails, generate 1,8 million Facebook likes, send 278 thousand Tweets, and up-load 200 thousand photos to Facebook
- Around 100 hours of video are uploaded to YouTube every minute and it would take you around 15 years to watch every video uploaded by users in one day
- 570 new websites spring into existence every minute of every day

https://www.smartdatacollective.com/big-data-25-facts-everyone-needs-know/

We need to organize this amount of data

- Data is valuable
- We need a structure to find data effectively

When do we use databases?

- Data
  - *Large amount*
  - *Existing*
  - *Regularly updated*
  - *Big variability*
  - *Categorized*
- Analysis
  - *Statistics*
  - *Regular reports*
  - *Analytics for decision makers (without expertise)*
  - *Dealing with many factors*
- Usage
  - *Multiple and parallel users*
  - *Multiple application uses as backend*
  - *Online*

When do we avoid using databases?

- Data
  - *We have only predicted (uncertain) data*
  - *Homogenous data*
- Analysis
  - *Single time execution (time wasting)*
  - *Simple to calculate by short formulas based on a number, or two.*

From spreadsheets to Databases

A GENERIC TABLE

Name	Address	Phone number	Educational level	Workplace
John Doe	Budapest	999-9999	Mechanical eng.	Szerszámgyártó Zrt.
Jackie Chan	Cegléd	999-9928	Civil eng.	Út kivitelező Nyrt.
Alba Flores	Budapest	999-9954	Economist	Elszámolok Kft.
Bruce Willis	Budapest	999-5864	Grammar school	Út kivitelező Nyrt.

TABLE TERMS

Column, attribute, property

Name	Address	Phone number	Educational level	Workplace
John Doe	Budapest	999-9999	Mechanical eng.	Szerszámgyártó Zrt.
Jackie Chan	Cegléd	999-9928	Civil eng.	Út kivitelező Nyrt.
Alba Flores	Budapest	999-9954	Economist	Elszámolok Kft.
Bruce Willis	Budapest	999-5864	Grammar school	Út kivitelező Nyrt.

Row, record, tuple

Cell, Field

Pros of spreadsheets

Easy to

- fill cells
- add/remove columns
- add/remove rows
- calculate derived cells
- *some functionalities from databases are now available in spreadsheets (Pivot, Vlookup)*

Difficult with spreadsheets

- Handle big tables
  - *100.000+ rows*
  - *1000+ columns*
- Hard to perform complex searches
- Join datasets
- Enforce consistency
- Special analysis

Databases?

- Recommended if:
  - *High amount of data (billions of records),*
  - *There are lot of attributes,*
  - *Demand of consistency ,*
  - *Complex analysis is required.*
- Once tools are available, even small databases are applied at companies to keep business system homogenous.

Databases

- Data storage
- Structuring data
- Enabling data filtering
- Deliver data to users
- Enabling real-time data upload
- Serving multi users parallelly

Analysis

- Analyzing the entire captured data with statistical methods
  - *Aggregated statistics*
- Parallel serving of multiple applications
- Recurring analysis day-by-day
- Detecting new correlations within different datasources
- Automated analysis, stored procedures

Backend

- Works in the background, all application has some kind of backend database.
- It's intend to be hidden from the users in a different layer and separated infrastructure.

Aims of databases

We would like to derive information from data; therefore

- Data has to be structured
- Data has to be consistent
- Recognize and link relationships between data

Terms

Terms I.

- **Data**: (stored in DB) most valuable. Static: unchanged till manual or automatic intervention. Data itself is meaningless.
- **Information**: derived from set of data by processing. Dynamic: if any of source data changes, derived information changes as well.

Conclusion: data is what we store, information is what we get from data by analysis.

Terms II.

Database model: is an abstract model that organizes elements of data and standardizes how they relate to one another and to the properties of real-world entities. It determines the logical structure of a database. It fundamentally determines in which manner data can be stored, organized and manipulated. (Wikipedia)

Logical model (schema): based on selected database model describes the system of stored data and links in between them.

Database vs. Database management system

Database (DB): data itself.

Database management system (DBMS): software which enables data analysis, manipulation, storage, creation.

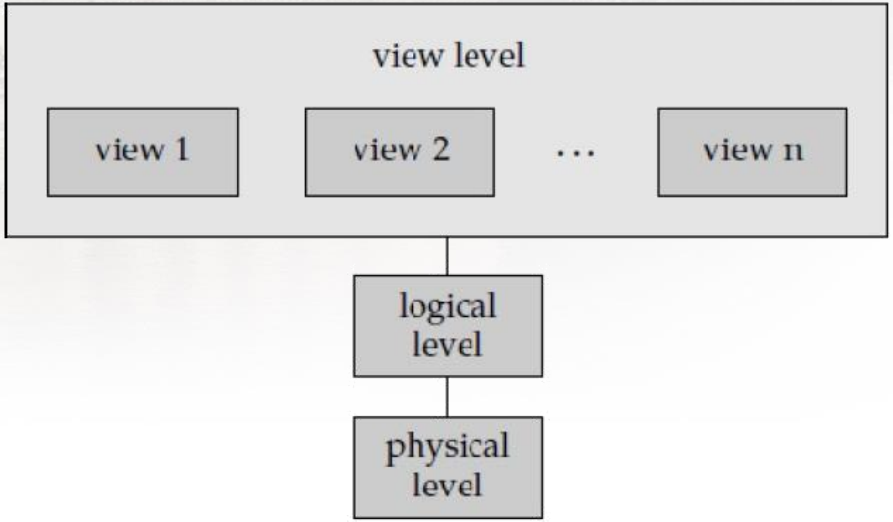
In practice we create databases with DBMSes. DBMS uses a given database model, so DBMS selection is done by considering database model as well.

DBMSes

- **MS Access** (commercial, Microsoft, relational, easy to learn, graphical interface, included in Office, mostly single user, low complexity)
- **SQL Server** (commercial, Microsoft, relational, network access, supports transactions, high complexity)
- **PostgreSQL** (Open Source, community developed, relational, network access, supports transactions, high complexity)
- **MySQL** (GPL, commercial, Oracle, relational, network access, mostly for serving traditional web pages, medium complexity)
- **MariaDB** (Open Source, community developed, relational, network access, mostly for serving traditional web pages, medium complexity)
- **Oracle** (commercial, Oracle, relational, network access, crucial systems like banks and Neptun use it, supports transactions, high complexity)
- **SQLite** (Open Source, Richard Hipp, mostly single users, Mobil apps and browsers use it, low complexity)
- **CouchDB** (Open Source, Apache, NoSQL, network access, modern webpages use it, low complexity)
- **MongoDB** (Open Source, MongoDB Inc, NoSQL, network access, modern webpages use it, low complexity)
- **Cassandra** (Open Source, Apache, NoSQL, network access, modern webpages use it, low complexity)

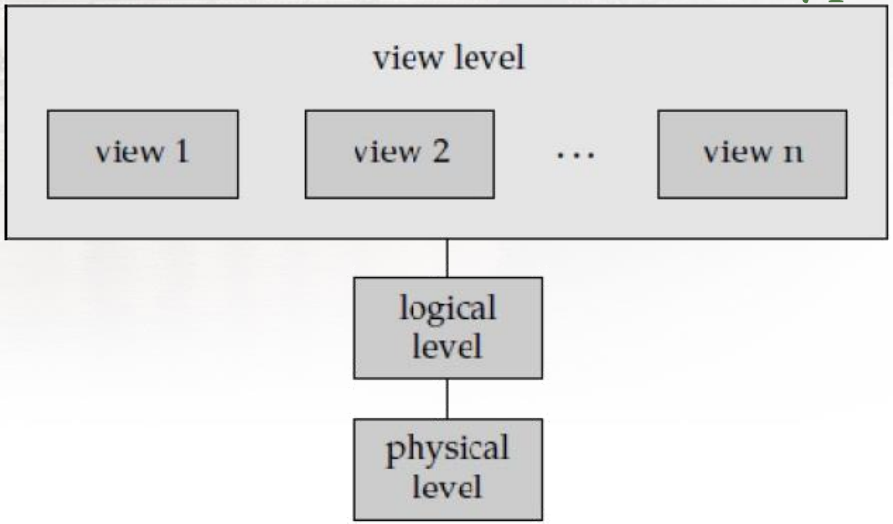
Database levels - physical

The lowest level of abstraction describes how a system actually stores and access data on hard drive. The physical level describes complex low-level data structures in detail.



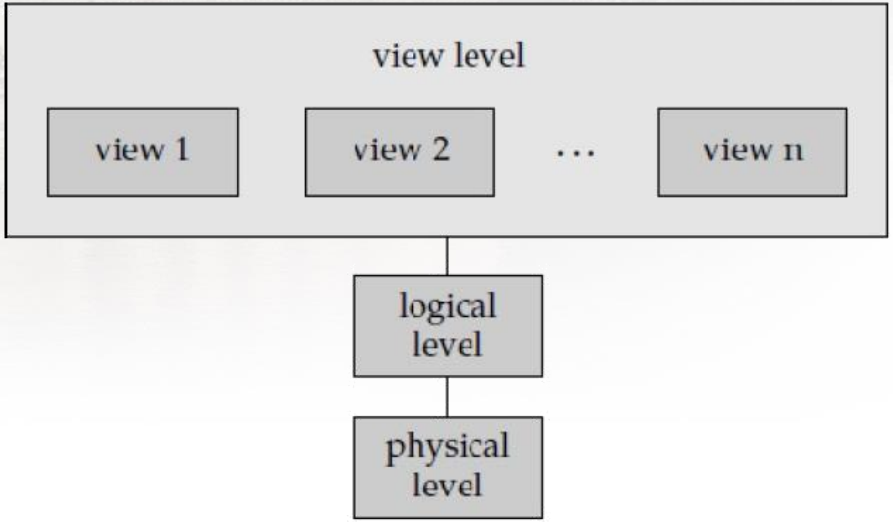
Database levels - logical

The logical level of abstraction describes what data the database stores, and what relationships exist among those data. This means the table definitions like attributes, attribute types and table connections.



Database levels - view

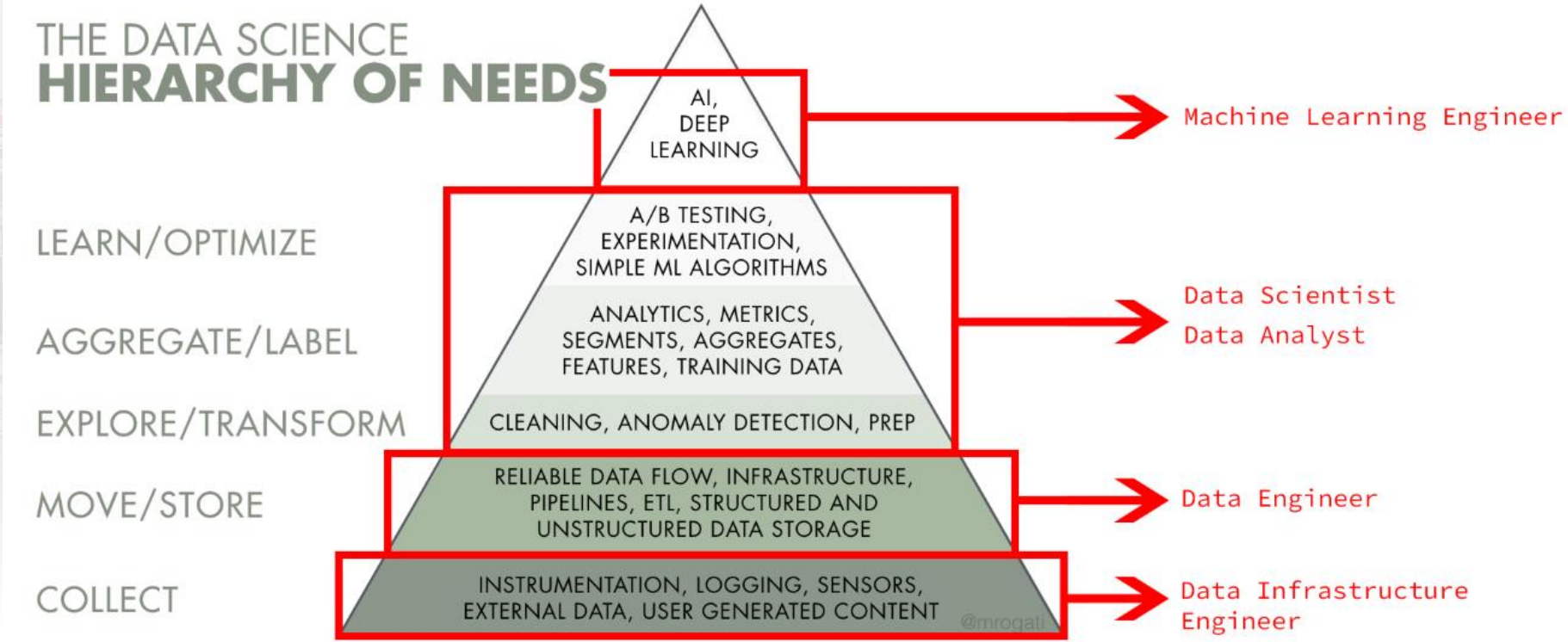
Views are data representations through the defined data model. Type and number of views are determined by the specification and demands.



History

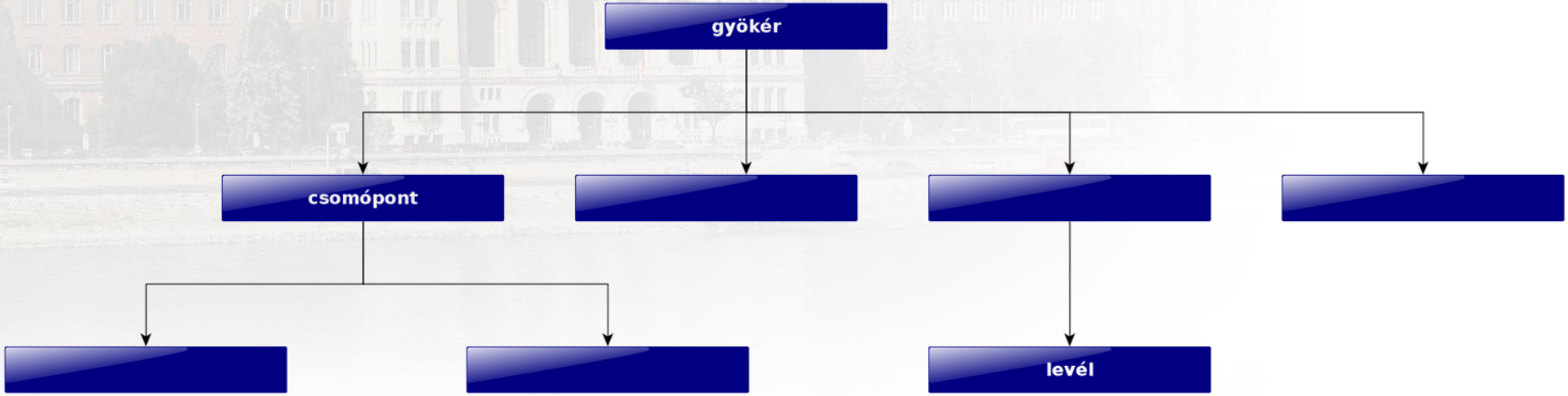
1960's (first official appearance 1962, Oxford dictionary: data-base)

History is introduced through the connected database model.


THE DATA SCIENCE
HIERARCHY OF NEEDS
AI,
DEEP
LEARNING
A/B TESTING,
EXPERIMENTATION,
SIMPLE ML ALGORITHMS
ANALYTICS, METRICS,
SEGMENTS, AGGREGATES,
FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE,
PIPELINES, ETL, STRUCTURED AND
UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT
@mrogati
LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
Machine Learning Engineer
Data Scientist
Data Analyst
Data Engineer
Data Infrastructure
Engineer

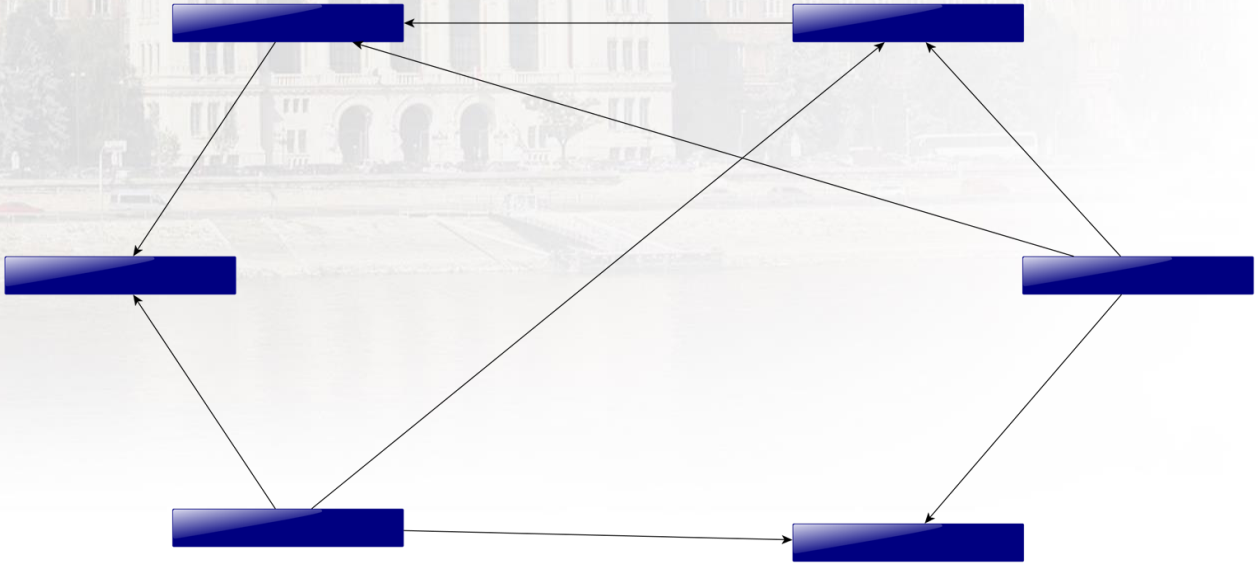
Early database models

Hierarchical model



Early database models

Network model



Relational model

- 1969 Edgar Codd (1980-)
- Flexible, easy to expand
- Widely used
- Easy to overview
- Connections are not defined at model level
- Clear mathematics background: set theory (relational algebra)

Object-Oriented Model

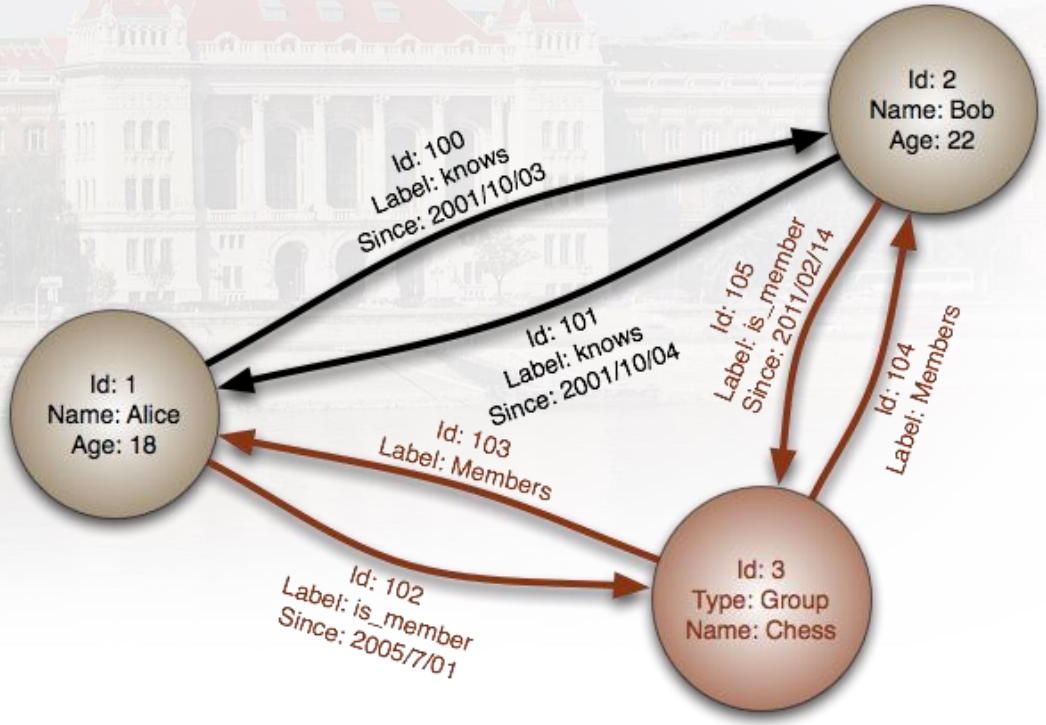
"Not only SQL" - NoSQL

- Started some years ago
- Different methodologies to store data
- Queries are not SQL based

Examples:

- MongoDB, CouchDB, Cassandra
- Document based DBs: semi-structured data, motivated by Web2.0

Graph databases



Relational data model - terms

Relational model - basics

Structural concept:

- table (relation),
- row (record, tuple),
- Attribute/property/column,
- Cell, field
- Special fields (e.g. complex, derive, multi-value).
- Field properties (data type) (NOT NULL, DEFAULT)

Conclusion

- Homework
- Where do we use databases day-by-day?
- Spreadsheets vs. Databases
- Terms
  - *information, data, data model*
  - *database*
  - *database management system*
- History of databases

Resources

http://en.wikipedia.org/wiki/Graph_database
http://en.wikipedia.org/wiki/Database
http://guide.couchdb.org
J. D. Ullman – J. Widom: Database systems
M. J. Hernandez: Database design
http://www.bigonehost.com

Thank you for your attention!

Questions?

BUDAPESTI MŰSZAKI
ÉS GAZDASÁGTUDOMÁNYI EGYETEM
Építőmérnöki Kar - építőmérnöki képzés 1782 óta

Fotogrammetria és Térinformatika Tanszék