BROWN
Computer Science

CS1951A: Data Science

Lecture 2: Database design and SQL

Lorenzo De Stefani

Spring 2022

Outline

- Database design principles
- Why databases?
- Four main phases of database design
- Book of duty
- Entity relation model
- Physical layer

What are databases?

- Data structures meant to store structured data allowing easy access to users
- We want:
  - Scalability: Modern database need to handle efficiently tens of billions of records
  - Integrity: Consistent data, no unwanted repetitions, uniform formatting
  - Ease of update & of access: It must be possible to add, remove, update and access record efficiently and while preserving integrity
  - Allow for concurrent accesses by multiple users

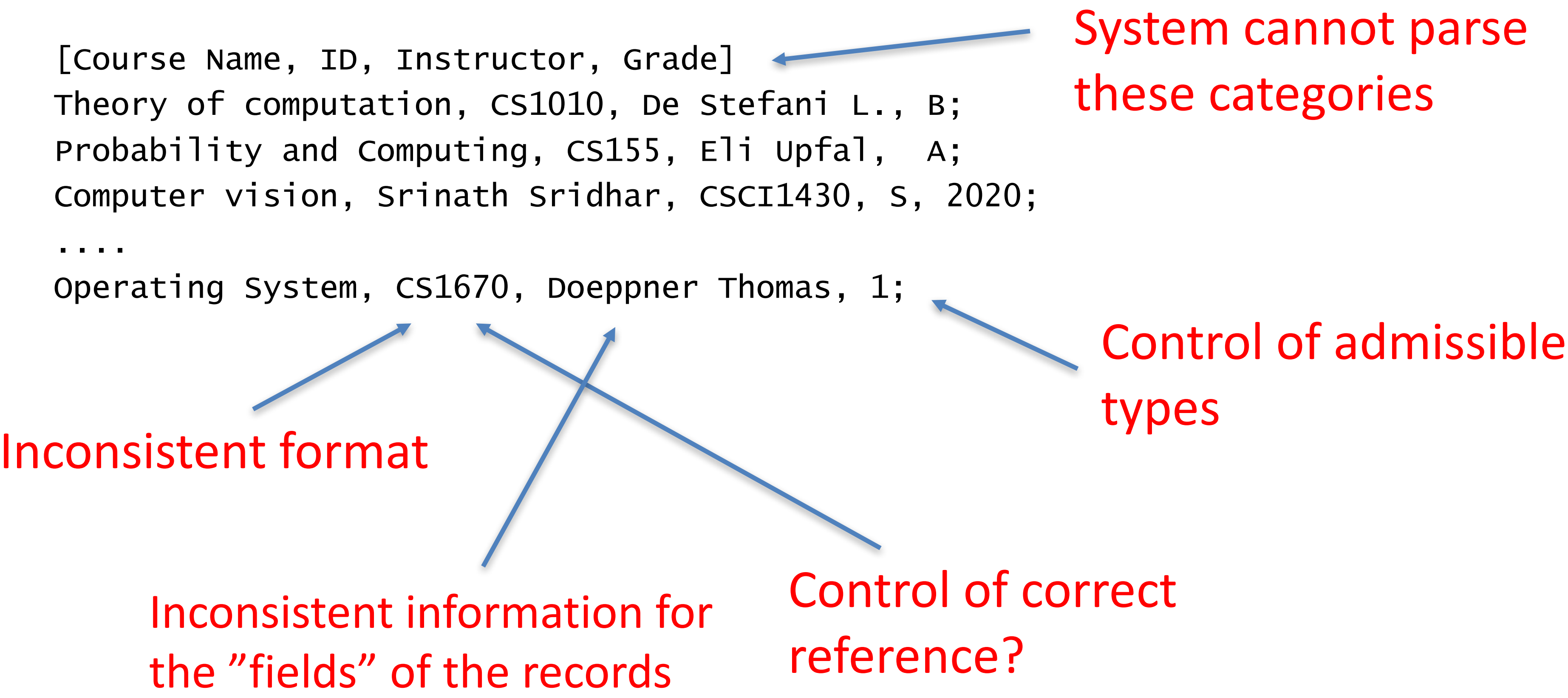
Do we really need databases?

Could we just use plain files?

.... they are so simple!!!

.... perhaps too simple to ensure our goals!

Reason 1: Data consistency



```
[Course Name, ID, Instructor, Grade]
Theory of computation, CS1010, De Stefani L., B;
Probability and Computing, CS155, Eli Upfal,  A;
Computer vision, Srinath Sridhar, CSCI1430, S, 2020;
....
Operating System, CS1670, Doeppner Thomas, 1;
```

System cannot parse these categories

Control of admissible types

Inconsistent format

Inconsistent information for the ”fields” of the records

Control of correct reference?

Reason 2: Scalability

Modern Data Base Management Systems (DBMSs) need to handle billions of records stored using hundreds of terabytes of data (and growing)

- We need optimized implementations on single computing nodes
- Single node implementations are not efficient
- Data must be distributed over many (100s-1000s) of nodes managed by (DBMSs)

Reason 3: Data Access

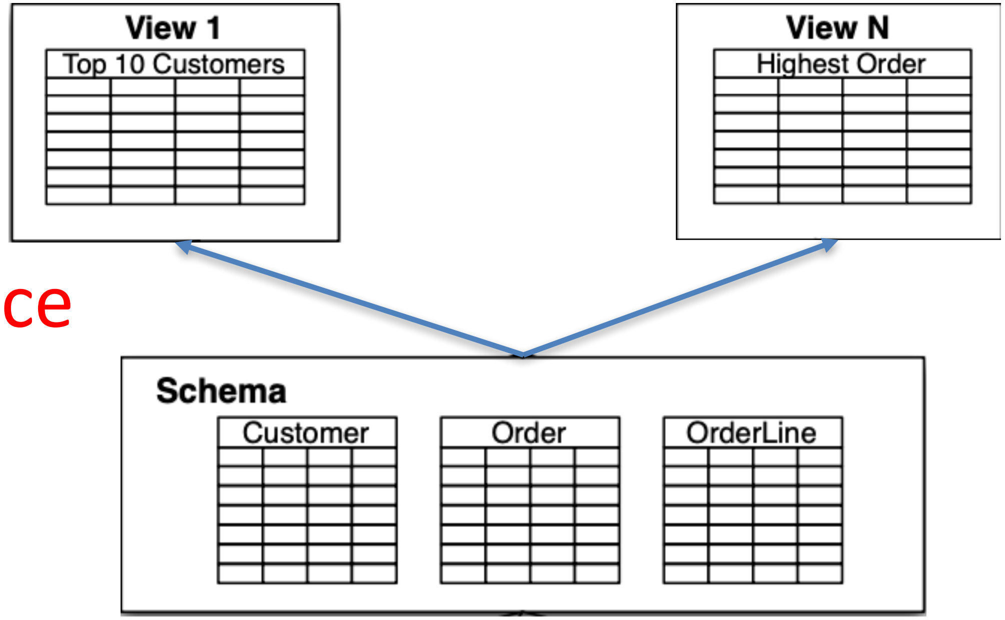
```
[Course Name, ID, Instructor, Grade]
Theory of computation, CS1010, De Stefani L., B;
Probability and Computing, CS155, Eli Upfal,  A;
Computer vision, Srinath Sridhar, CSCI1430, S, 2020;
....
Operating System, CS1670, Doeppner Thomas, 1;
Data Science, CS1951A, Lorenzo De Stefani, B
```

Query: “Find all courses taught by Lorenzo De Stefani”

- Practicality issues: we have to design a program to parse the file and retrieve the information
- Efficiency issues: we need to read the entire file to answer the query

Reason 4: Data independence

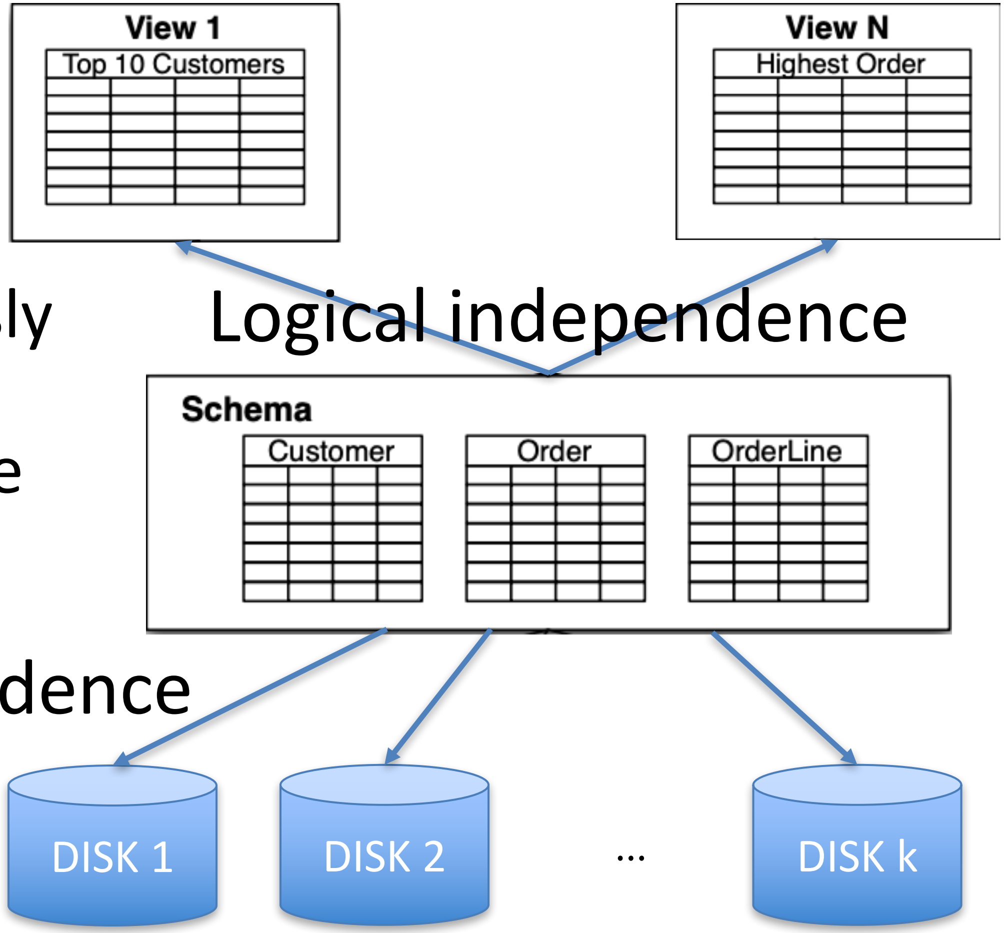
Logical independence



DBMSs allow to easily present the data in specific representations (view) selected by a given query

Reason 4: Data independence

DBMSs allow to seamlessly integrate data store in different nodes or storage



View 1

Top 10 Customers

View N

Highest Order

Logical independence

Schema

Customer

Order

OrderLine

Physical independence

DISK 1

DISK 2

…

DISK k

Reason 5: Concurrent access

Love Me Tender
UNCOOKED
Bell&Evans
AIR CHILLED
BREADED
CHICKEN BREAST
TENDERS
FLAVOR AND MOISTURE ASSURED WITH UP TO 6% MARINADE
WHOLE BREAST MEAT TENDERS
Raised Without Antibiotics
NET WT 12 oz (340g)
KEEP FROZEN

Bell & Evans, Frozen Breaded Chicken Tender (Raised without Antibiotics), 12 oz

Brand: Bell & Evans

★★★★☆ 1,936 ratings

Price: $6.99 ($6.99 / Count)

Get $100 off instantly: Pay $0.00 ~~$6.99~~ upon approval for the Amazon Prime Rewards Visa Card. No annual fee.

This item is non-returnable

Brand	Bell & Evans
Ingredients	Chicken Breast Meat Tenders, Marinated In Water And Sea Salt, Breaded With Unbleached Wheat Flour, Water, Cane Sugar, Dried Yeast, Sea Salt, Black Pepper, Paprika. See more
Weight	0.45 Pounds
Protein	20 Grams
Serving Description	4 ounces (113 grams)

About this item

- Marinated in water and sea salt Breaded with organic gluten free crumb
- Flash fried in organic expeller pressed soybean oil to set breading
- They are 100% pure breast tender meat, lightly breaded with gluten free breading and flash fried for your convenience

Report incorrect product information.

Love Me Tender
UNCOOKED
Bell&Evans
AIR CHILLED
BREADED
CHICKEN BREAST
TENDERS
FLAVOR AND MOISTURE ASSURED WITH UP TO 6% MARINADE
WHOLE BREAST MEAT TENDERS
Raised Without Antibiotics
NET WT 12 oz (340g)
KEEP FROZEN

Bell & Evans, Frozen Breaded Chicken Tender (Raised without Antibiotics), 12 oz

Brand: Bell & Evans

★★★★☆ 1,936 ratings

Price: $6.99 ($6.99 / Count)

Get $100 off instantly: Pay $0.00 ~~$6.99~~ upon approval for the Amazon Prime Rewards Visa Card. No annual fee.

This item is non-returnable

Brand Bell & Evans

Ingredients Chicken Breast Meat Tenders, Marinated In Water And Sea Salt, Breaded With Unbleached Wheat Flour, Water, Cane Sugar, Dried Yeast, Sea Salt, Black Pepper, Paprika.

See more

Weight 0.45 Pounds

Protein 20 Grams

Serving Description 4 ounces (113 grams)

About this item

- Marinated in water and sea salt Breaded with organic gluten free crumb
- Flash fried in organic expeller pressed soybean oil to set breading
- They are 100% pure breast tender meat, lightly breaded with gluten free breading and flash fried for your convenience

Report incorrect product information.

Stock 1

Stock 1

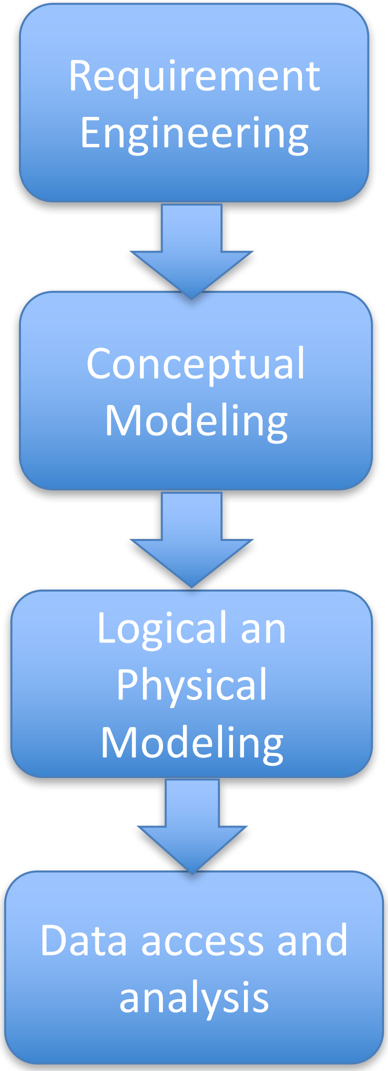
Stock 0

Stock 0

Product ID	Product Name	Stock
0473902	Chicken Tendies	1
…	…….	……

DBMS ensure correctness while allowing concurrent access to multiple users

Databases for Data Scientists



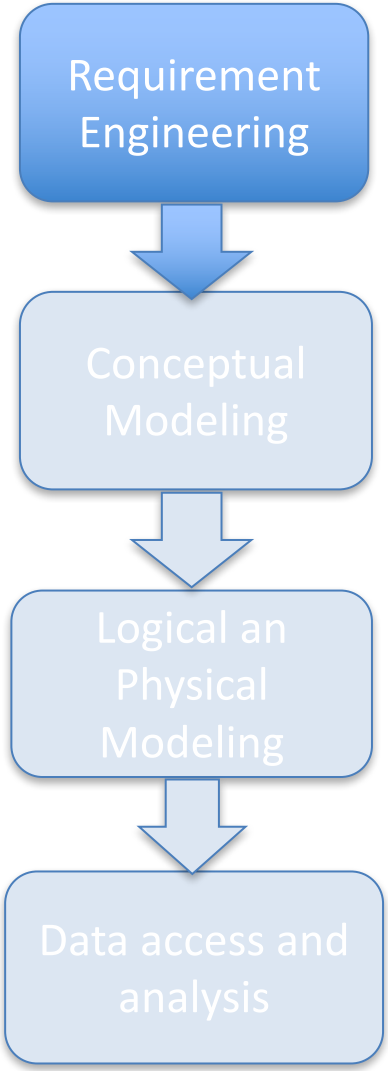
- "Book of duty"
- Understand and model the "world" of interest

- Conceptual DB design
- Entity Relations (ER) method

- Logical design (schema, table names, data types)
- Physical design (index, hints, memory organization)

- Asking and answering questions (queries)
- Extract information form the DBMS (views)
- SQL and relational algebra

Databases for Data Scientists



- “Book of duty”
- Understand and model the “world” of interest

- Conceptual DB design
- Entity Relations (ER) method

- Logical design (schema, table names, data types)
- Physical design (index, hints, memory organization)

- Asking and answering questions (queries)
- Extract information form the DBMS (views)
- SQL and relational algebra

Book of duty

A description of the population of the database system and the desired mean of access/interaction

- Description can be informal but should be detailed
- Describe information requirements
  - What are the items in the populations
    - Eg., items for sale, records of sale, entries in a transcript
  - Which are the concepts that should be represented?
    - E.g., items, storage facilities, students, courses
  - What are the attributes of the concepts
    - E.g., price, color, availability, grade
  - What are the domains of attributes of objects?
    - E.g., letters, integer numbers, dates
  - How are objects identified/referenced?
    - E.g., BannerID, DOI, SSN
  - Are there relationships between concepts? What is their nature?
    - E.g., authorship, manufacturer, distributor, publisher,...

Book of duty

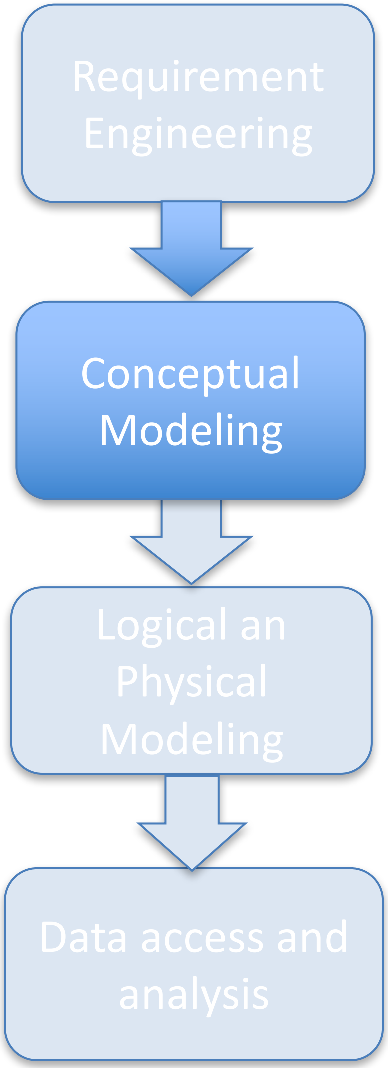
- Describe processing requirements
  - Cardinalities: how many items is the system expected to manage?
    - E.g. # students university database, # items online shop, # number movies on a streaming platform;
    - Estimates rather than exact values: meaningful as guidelines
  - Distributions
    - E.g., grade distributions in a class, number of order request through the day
  - Workload
    - Read/write frequency
  - Priorities and service level agreements
    - Are there different tiers of users?
    - What guarantees on the service should be ensured?
    - Privacy of users and records

Practice time

Come up with a example Book of Duty for the records of students in the CS department

- Concepts
- Attributes:
- Types of data:
- Identifiers:
- Relationships between concepts:
- Cardinalities
- Workload
- Priorities and service level agreements

Databases for Data Scientists



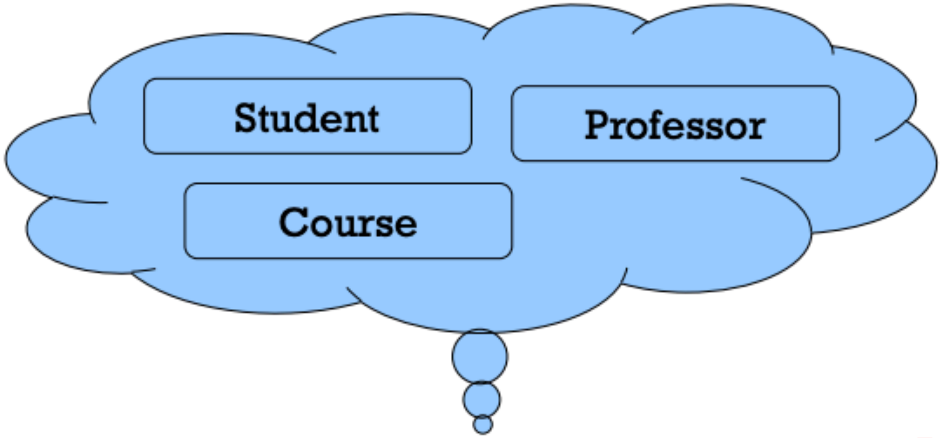
- “Book of duty”
- Understand and model the “world” of interest

- Conceptual DB design
- Entity Relations (ER) method

- Logical design (schema, table names, data types)
- Physical design (index, hints, memory organization)

- Asking and answering questions (queries)
- Extract information form the DBMS (views)
- SQL and relational algebra

Conceptual modeling

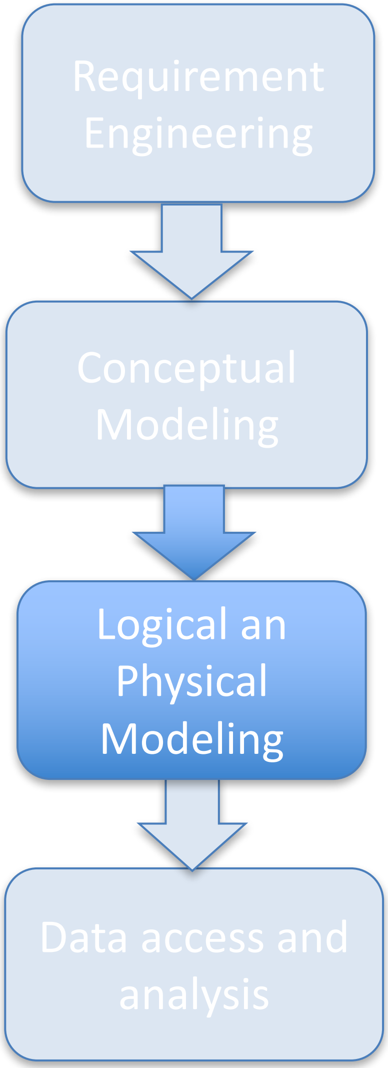


Book of Duty

Identifying Attributes

Entity

Databases for Data Scientists



Requirement Engineering

Conceptual Modeling

Logical an Physical Modeling

Data access and analysis

- "Book of duty"
- Understand and model the "world" of interest

- Conceptual DB design
- Entity Relations (ER) method

- Logical design (schema, table names, data types)
- Physical design (index, hints, memory organization)

- Asking and answering questions (queries)
- Extract information form the DBMS (views)
- SQL and relational algebra

Logical and Physical costraint

Student ID	Name
0473902	Jack
9408545	Adam
7576463	Sumiko

Table: Student

Student ID	Class ID
0473902	CS101
9408545	CS145
7576463	CS019

Table: Attendance

Instructor	Class ID
De Stefani	CS101
Upfal	CS145
Krishamurti	CS019

Table: Teaching

Logical Design

- Table / column names
- Data types
- Constraints
- …

Data Definition: Data Types

- Numeric: INT, FLOAT, REAL, DOUBLE
- Character Strings: CHAR(n), VARCHAR(n), CLOB(size)
  - CHAR is fixed with, VARCHAR is not
  - CLOB(2MB) for large objects e.g. documents/web pages
- Bit Strings: BIT(n), BIT VARYING(n), BLOB
  - BLOB(20MB) e.g. for images
- Boolean
- Dates: DATE, TIME, TIMESTAMP, TIME WITH TIME ZONE
- Opportune choice of data type leads to improved performance and better memory utilization

https://www.w3schools.com/sql/sql_datatypes.asp

Logical and Physical constraint

Student ID	Name
0473902	Jack
9408545	Adam
7576463	Sumiko

Table: Student

Student ID	Class ID
0473902	CS101
9408545	CS145
7576463	CS019

Table: Attendance

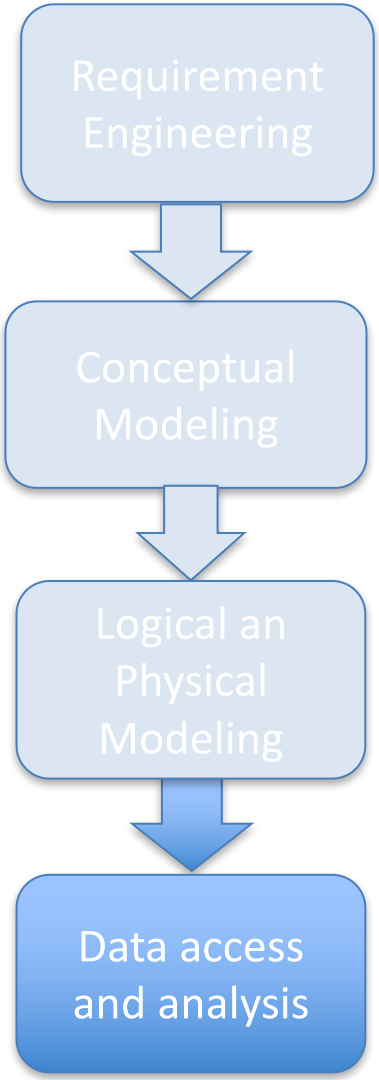
Instructor	Class ID
De Stefani	CS101
Upfal	CS145
Krishamurti	CS019

Table: Teaching

Physical Design

- Indexes to speed up retrieval
- Memory layout
- Compression
- Distribution on multiple machines
- …

Databases for Data Scientists



- “Book of duty”
- Understand and model the “world” of interest

- Conceptual DB design
- Entity Relations (ER) method

- Logical design (schema, table names, data types)
- Physical design (index, hints, memory organization)

- Asking and answering questions (queries)
- Extract information form the DBMS (views)
- SQL and relational algebra

How to ask questions

Data managed in DBMS are accessed by stating of “questions” or queries

- E.g., “How many students attended CS145 in 2019?”

Student ID	Name
0473902	Jack
9408545	Adam
7576463	Sumiko

Table: Student

Student ID	Class ID	AA
0473902	CS101	2020
9408545	CS145	2019
7576463	CS019	2018

Table: Attendance

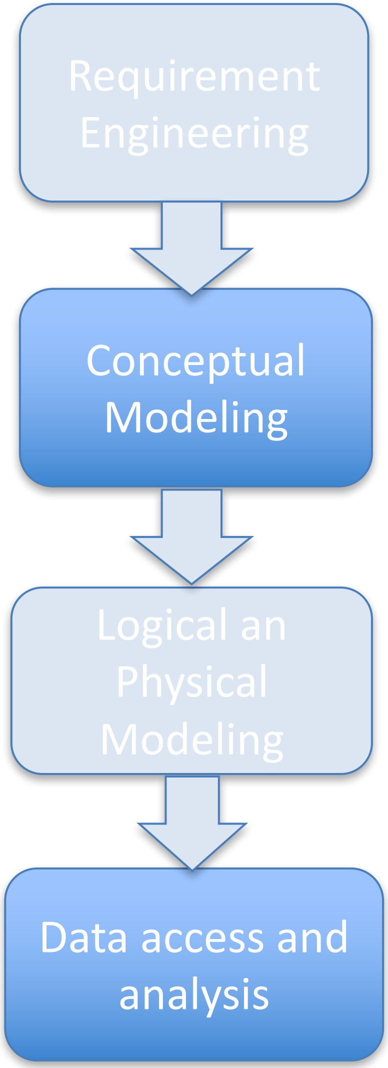
We will consider mostly as SQL queries

SELECT COUNT(*)

FROM Student s, Attendance a WHERE s.StudentID=a.StudentID AND a.CourseID=‘CS145’ AND a.CourseID=‘2019’

- Queries formulated using Relational Algebra

Plan for next time



Requirement Engineering

Conceptual Modeling

Logical an Physical Modeling

Data access and analysis

- “Book of duty”
- Understand and model the “world” of interest

- Conceptual DB design
- Entity Relations (ER) method

- Logical design (schema, table names, data types)
- Physical design (index, hints, memory organization)

- Asking and answering questions (queries)
- Extract information form the DBMS (views)
- SQL and relational algebra