

UMD DATA605 - Big Data Systems

Relational DBs

SQL Intro

SQL tutorial

Relational DB internals

Dr. GP Saggese

gsaggese@umd.edu

with thanks to Prof.

Alan Sussman

Amol Deshpande

UMD DATA605 - Big Data Systems

Relational DBs

SQL Intro

SQL tutorial

Relational DB internals

Silberschatz: Chap 2

Relational Model: Overview

- Introduced by Ted Codd (late 60's, early 70's)
- First prototypes:
 - Ingres Project at Berkeley (1970-1985):
 - Ingres (INteractive Graphics REtrieval System) -> PostgreSQL (=Post Ingres)
 - Postgres -> Illustra -> Informix -> IBM
 - IBM System R (1970) -> Oracle, IBM DB2
- Relational data model contributions
 - Formal semantics
 - Data independence: separation of logical and physical data models
 - Declarative query languages (e.g., SQL)
 - Query optimization
- Key to commercial success

Relational Model: Key Definitions

- A relational DB consists of a collection of **tables** (aka **relations**)
 - Each table has a unique name and a schema
- Each **row** (aka **tuple**, **record**) in a table represents a relationship among a set of values
- Each **element** of the row corresponds to a **column** (aka **attribute**)
 - Each element in a column is atomic (e.g., a phone number is a single object and not a sequence of numbers)
 - A valid value is `NULL` representing a value that is unknown or doesn't exist (e.g., someone not having a phone number)
- E.g., *instructor* and *course* relations
- **Relation schema**
 - A list of attributes and their domains
 - Like type definition in programming languages
 - E.g., the domain of *salary* is integers ≥ 0
- **Instance of relation**
 - A particular instantiation of a relation with actual values
 - Will change over time

<i>ID</i>	<i>name</i>	<i>dept_name</i>	<i>salary</i>
10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	60000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	62000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

instructor relation

<i>course_id</i>	<i>title</i>	<i>dept_name</i>	<i>credits</i>
BIO-101	Intro. to Biology	Biology	4
BIO-301	Genetics	Biology	4
BIO-399	Computational Biology	Biology	3
CS-101	Intro. to Computer Science	Comp. Sci.	4
CS-190	Game Design	Comp. Sci.	4
CS-315	Robotics	Comp. Sci.	3
CS-319	Image Processing	Comp. Sci.	3
CS-347	Database System Concepts	Comp. Sci.	3
EE-181	Intro. to Digital Systems	Elec. Eng.	3
FIN-201	Investment Banking	Finance	3
HIS-351	World History	History	3
MU-199	Music Video Production	Music	3
PHY-101	Physical Principles	Physics	4

course relation

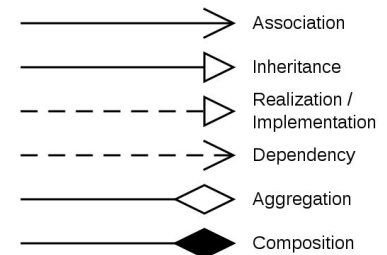
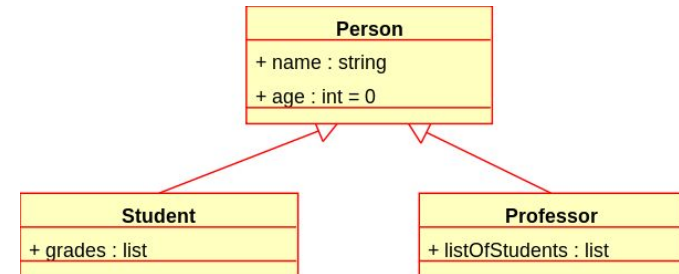
UML Class Diagram

- UML class diagram

- Used in OOP and DB design
- UML = Unified Modeling Language
- Diagram showing classes, attributes, methods, and relationships

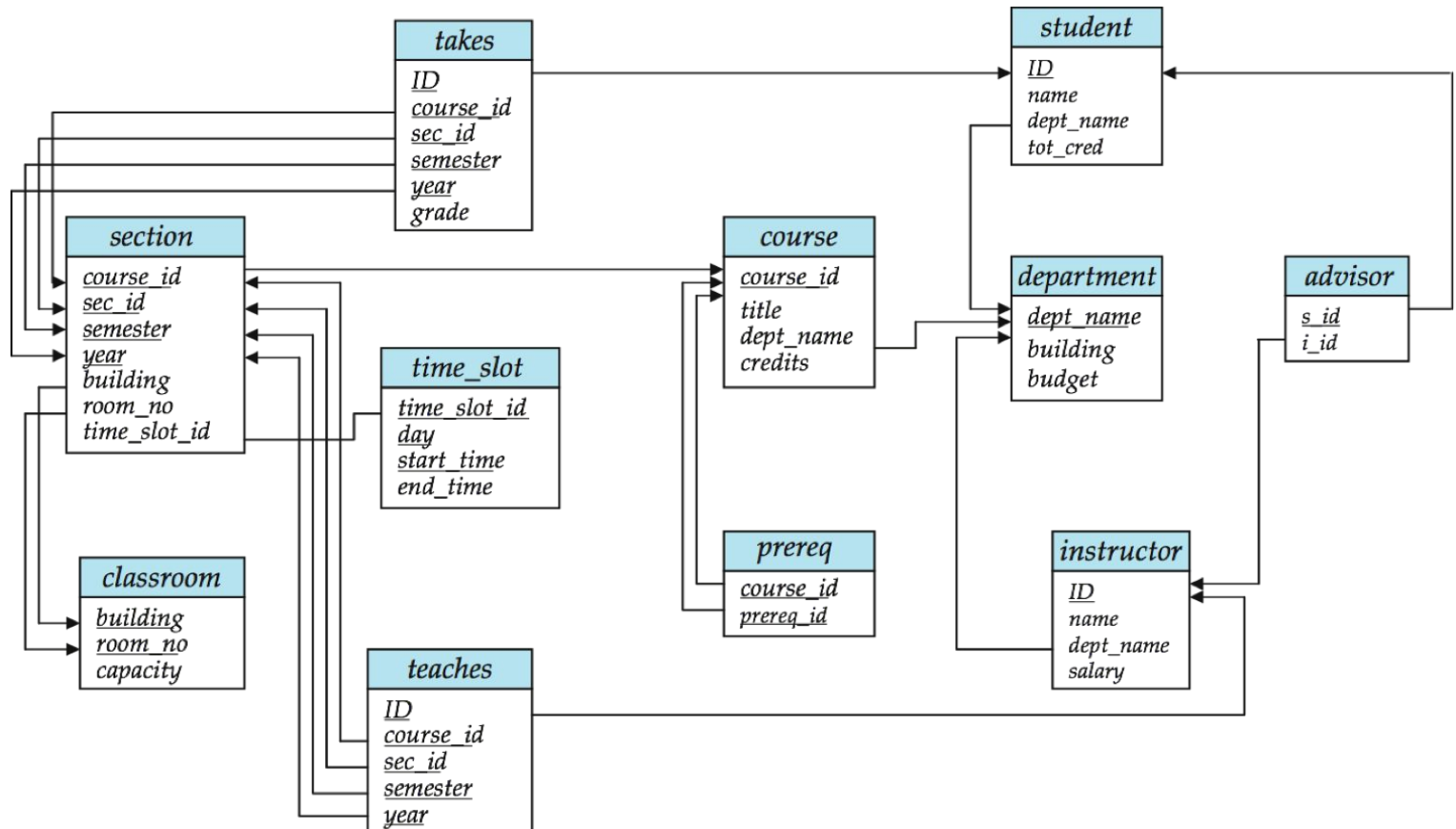
- Schema diagram

- Each box is a relation
- Attributes are listed inside the box
- Primary keys
- Foreign key constraints



Example: University DB

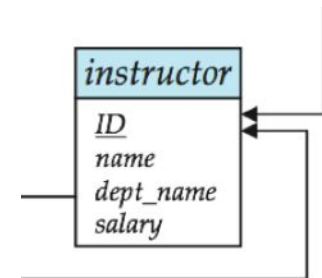
- Schema diagram of a DB representing University
- Each box is a table / relation
- Column / fields / attributes are listed inside the box
- Primary keys: underlined fields
- Foreign key constraints: arrows between boxes



Primary Key

- R is the set of attributes of a relation r
- K is a **superkey** of R if values for K are sufficient to identify a unique tuple of each possible relation $r(R)$
 - E.g., $\{ID\}$ and $\{ID, name\}$ are both superkeys of *instructor*
 - $\{name\}$ is not a superkey (or a key) of *instructor*
- Superkey K is a **candidate key** if K is minimal
 - E.g., $\{ID\}$ is a candidate key for *instructor*
- One of the candidate keys is selected to be the **primary key**
 - Typically one that is small and immutable (doesn't change often)
 - Primary keys typically are the first in the schema and highlighted / underlined
 - Would SSN be a primary key?
- **Primary key constraint**: tuples in the relation can't have the same primary key

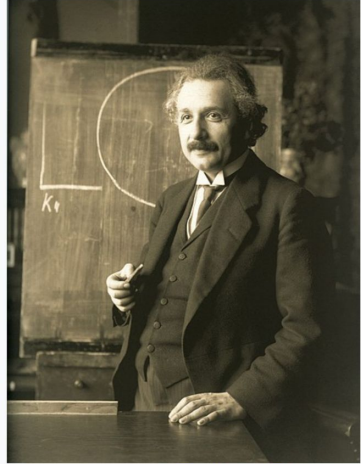
<i>ID</i>	<i>name</i>	<i>dept_name</i>	<i>salary</i>
10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	60000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	62000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000



Exercise: What are Primary Keys?

- Marital status
 - Married(person1_ssn, person2_ssn, date_married, date_divorced)
- Bank account
 - Account(cust_ssn, account_number, cust_name, balance, cust_address)
- Research assistantship at UMD
 - RA(student_id, project_id, supervisor_id, appt_time, appt_start_date, appt_end_date)
- Information typically found on Wikipedia
 - Person(Name, Born, Died, Citizenship, Education, ...)
- Info about US President on Wikipedia
 - President(name, start_date, end_date, vice_president, preceded_by, succeeded_by)
- Tour de France: Historical Rider Participation Information
 - Rider(Name, Born, Team-name, Coach, Sponsor, Year)

Albert Einstein



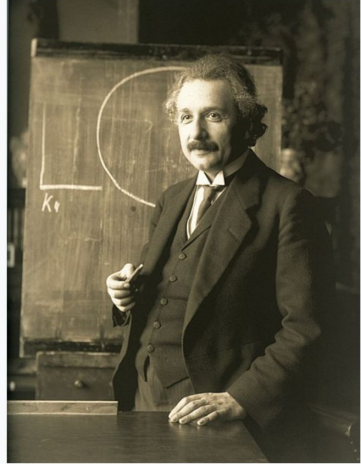
Einstein in 1921, by [Ferdinand Schmutzer](#)

Born	14 March 1879 Ulm, Germany
Died	18 April 1955 (aged 76) Princeton, New Jersey, U.S.
Citizenship	Full list [show]
Education	Federal polytechnic school in Zurich (Federal teaching diploma, 1900) University of Zurich (PhD, 1905)
Known for	General relativity Special relativity Photoelectric effect E=mc² (Mass–energy equivalence) E=hf (Planck–Einstein relation) Theory of Brownian motion

Exercise: What are Primary Keys?

- Marital status
 - Married(**person1_ssn**, **person2_ssn**, date_married, date_divorced)
- Bank account
 - Account(cust_ssn, **account_number**, cust_name, balance, cust_address)
- Research assistantship at UMD
 - RA(**student_id**, **project_id**, supervisor_id, appt_time, **appt_start_date**, appt_end_date)
- Information typically found on Wikipedia
 - Person(**Name**, **Born**, **Died**, **Citizenship**, **Education**, ...)
- Info about US President on Wikipedia
 - President(**name**, **start_date**, end_date, vice_president, preceded_by, succeeded_by)
- Tour de France: Historical Rider Participation Information
 - Rider(**Name**, **Born**, **Team-name**, Coach, Sponsor, **Year**)

Albert Einstein

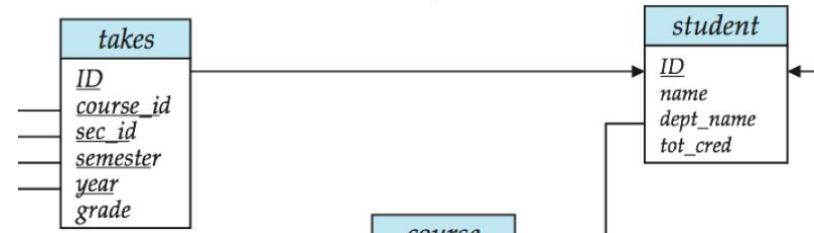


Einstein in 1921, by [Ferdinand Schmutzer](#)

Born	14 March 1879 Ulm, Germany
Died	18 April 1955 (aged 76) Princeton, New Jersey, U.S.
Citizenship	Full list [show]
Education	Federal polytechnic school in Zurich (Federal teaching diploma, 1900) University of Zurich (PhD, 1905)
Known for	General relativity Special relativity Photoelectric effect E=mc² (Mass–energy equivalence) E=hf (Planck–Einstein relation) Theory of Brownian motion

Foreign Key

- **Foreign key**: primary key of a relation that appears in another relation
 - E.g., {ID} from *student* appears in *takes*, *advisor*
 - *takes* is the “referencing relation” (with the foreign key)
 - *student* called “referenced relation” (with the primary key)
 - Typically shown by an arrow from referencing to referenced relation
- **Foreign key constraint**: the tuple corresponding to that primary key must exist, e.g.,
 - Tuple: ('student101', 'DATA605') in *takes*
 - But no tuple corresponding to 'student101' in *student*
 - Aka referential integrity constraint



Relational Algebra

- **Relation:** set of tuples (without replication)
- **Relational algebra:** set operations that take one or more relations as input and produce a new relation, e.g.,
 - unary relation: select, project, rename
 - binary relation: union, Cartesian product, set difference
- **Selection:** select tuples that satisfy a given predicate
 - E.g., select tuples of *instructor* where `dept_name = "Physics"
- **Projection:** return tuples with a subset of attributes
 - E.g., project tuples of *instructor* with only (name, salary)
- **Set operations:** union, intersection, set_difference of relations
 - Need to be compatible (i.e., have the same attributes and names)

ID	name	dept_name	salary
10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	60000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	62000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

ID	name	dept_name	salary
22222	Einstein	Physics	95000
33456	Gold	Physics	87000

$\sigma_{dept_name = "Physics"}(instructor)$

ID	name	salary
10101	Srinivasan	65000
12121	Wu	90000
15151	Mozart	40000
22222	Einstein	95000
32343	El Said	60000
33456	Gold	87000
45565	Katz	75000
58583	Califieri	62000
76543	Singh	80000
76766	Crick	72000
83821	Brandt	92000
98345	Kim	80000

$\Pi_{ID, name, salary}(instructor)$

Relational Algebra

- **Cartesian product:** combine information from two relations
 - E.g., *`instructor`* x *`teaches`*
 - *`instructor`* = (ID, name, dept_name) and *`teaches`* = (ID, course_id, sec_id, semester, year)
 - Gives *`(instructor.ID, instructor.name, instructor.dept_name, teaches.ID, ...)`*

ID	name	dept_name	salary
10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	60000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	62000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

ID	course_id	sec_id	semester	year
10101	CS-101	1	Fall	2017
10101	CS-315	1	Spring	2018
10101	CS-347	1	Fall	2017
12121	FIN-201	1	Spring	2018
15151	MU-199	1	Spring	2018
22222	PHY-101	1	Fall	2017
32343	HIS-351	1	Spring	2018
45565	CS-101	1	Spring	2018
45565	CS-319	1	Spring	2018
76766	BIO-101	1	Summer	2017
76766	BIO-301	1	Summer	2018
83821	CS-190	1	Spring	2017
83821	CS-190	2	Spring	2017
83821	CS-319	2	Spring	2018
98345	EE-181	1	Spring	2017

instructor.ID	name	dept_name	salary	teaches.ID	course_id	sec_id	semester	year
10101	Srinivasan	Comp. Sci.	65000	10101	CS-101	1	Fall	2017
10101	Srinivasan	Comp. Sci.	65000	10101	CS-315	1	Spring	2018
10101	Srinivasan	Comp. Sci.	65000	10101	CS-347	1	Fall	2017
10101	Srinivasan	Comp. Sci.	65000	12121	FIN-201	1	Spring	2018
10101	Srinivasan	Comp. Sci.	65000	15151	MU-199	1	Spring	2018
10101	Srinivasan	Comp. Sci.	65000	22222	PHY-101	1	Fall	2017
...
...
12121	Wu	Finance	90000	10101	CS-101	1	Fall	2017
12121	Wu	Finance	90000	10101	CS-315	1	Spring	2018
12121	Wu	Finance	90000	10101	CS-347	1	Fall	2017
12121	Wu	Finance	90000	12121	FIN-201	1	Spring	2018
12121	Wu	Finance	90000	15151	MU-199	1	Spring	2018
12121	Wu	Finance	90000	22222	PHY-101	1	Fall	2017
...
...
15151	Mozart	Music	40000	10101	CS-101	1	Fall	2017
15151	Mozart	Music	40000	10101	CS-315	1	Spring	2018
15151	Mozart	Music	40000	10101	CS-347	1	Fall	2017
15151	Mozart	Music	40000	12121	FIN-201	1	Spring	2018
15151	Mozart	Music	40000	15151	MU-199	1	Spring	2018
15151	Mozart	Music	40000	22222	PHY-101	1	Fall	2017
...
...
22222	Einstein	Physics	95000	10101	CS-101	1	Fall	2017
22222	Einstein	Physics	95000	10101	CS-315	1	Spring	2018
22222	Einstein	Physics	95000	10101	CS-347	1	Fall	2017
22222	Einstein	Physics	95000	12121	FIN-201	1	Spring	2018
22222	Einstein	Physics	95000	15151	MU-199	1	Spring	2018
22222	Einstein	Physics	95000	22222	PHY-101	1	Fall	2017
...
...

instructor × *teaches*.

Relational Algebra

- **Join:** Cartesian-product and then a selection based on equality between two fields
 - E.g., instructor.ID = teaches.ID

ID	name	dept_name	salary
10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	60000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	62000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

ID	course_id	sec_id	semester	year
10101	CS-101	1	Fall	2017
10101	CS-315	1	Spring	2018
10101	CS-347	1	Fall	2017
12121	FIN-201	1	Spring	2018
15151	MU-199	1	Spring	2018
22222	PHY-101	1	Fall	2017
32343	HIS-351	1	Spring	2018
45565	CS-101	1	Spring	2018
45565	CS-319	1	Spring	2018
76766	BIO-101	1	Summer	2017
76766	BIO-301	1	Summer	2018
83821	CS-190	1	Spring	2017
83821	CS-190	2	Spring	2017
83821	CS-319	2	Spring	2018
98345	EE-181	1	Spring	2017

instructor.ID	name	dept_name	salary	teaches.ID	course_id	sec_id	semester	year
10101	Srinivasan	Comp. Sci.	65000	10101	CS-101	1	Fall	2017
10101	Srinivasan	Comp. Sci.	65000	10101	CS-315	1	Spring	2018
10101	Srinivasan	Comp. Sci.	65000	10101	CS-347	1	Fall	2017
12121	Wu	Finance	90000	12121	FIN-201	1	Spring	2018
15151	Mozart	Music	40000	15151	MU-199	1	Spring	2018
22222	Einstein	Physics	95000	22222	PHY-101	1	Fall	2017
32343	El Said	History	60000	32343	HIS-351	1	Spring	2018
45565	Katz	Comp. Sci.	75000	45565	CS-101	1	Spring	2018
45565	Katz	Comp. Sci.	75000	45565	CS-319	1	Spring	2018
76766	Crick	Biology	72000	76766	BIO-101	1	Summer	2017
76766	Crick	Biology	72000	76766	BIO-301	1	Summer	2018
83821	Brandt	Comp. Sci.	92000	83821	CS-190	1	Spring	2017
83821	Brandt	Comp. Sci.	92000	83821	CS-190	2	Spring	2017
83821	Brandt	Comp. Sci.	92000	83821	CS-319	2	Spring	2018
98345	Kim	Elec. Eng.	80000	98345	EE-181	1	Spring	2017

$\sigma_{\text{instructor.ID} = \text{teaches.ID}}(\text{instructor} \times \text{teaches})$

Relational Algebra

- **Query:** combination of relational algebra operations
- **Assignment:** assign parts of relational algebra to temporary relation variables
 - A query can be written as a sequential program
- **Equivalent queries:** two queries that give the same result on any DB instance
 - Some formulation can be more efficient than others

$$\Pi_{course.id} (\sigma_{semester = \text{"Fall"} \wedge year = 2017} (section))$$
$$\begin{aligned} courses_fall_2017 &\leftarrow \Pi_{course.id} (\sigma_{semester = \text{"Fall"} \wedge year = 2017} (section)) \\ courses_spring_2018 &\leftarrow \Pi_{course.id} (\sigma_{semester = \text{"Spring"} \wedge year = 2018} (section)) \\ &courses_fall_2017 \cap courses_spring_2018 \end{aligned}$$

SQL Overview

- **Relational Algebra**: mathematical description of a language to manipulate Relations (aka Tables)
- **SQL**
 - Programming language to describe and transform data in a Relational DB
 - Originally called Sequel
 - Name was changed to Structured Query Language
- Data definition language (DDL)
 - Define schema of the data (e.g., tables, attributes, indices)
 - Specify integrity constraints (e.g., primary key, foreign key, not null)
- Data modification language (DML)
 - Modify the data in tables
 - Insert
 - Update
- Query data
- Authorization
 - Specify security constraints
- Define views
- Control transactions
 - E.g., specify beginning and end, control isolation level

SQL Overview

Data description language

```
CREATE TABLE <name> ( <field> <domain>, ... )
```

Data modification language

```
INSERT INTO <name> (<field names>) VALUES (<field values>)
```

```
DELETE FROM <name> WHERE <condition>
```

```
UPDATE <name> SET <field name> = <value> WHERE <condition>
```

Query language




```
SELECT <fields> FROM <name> WHERE <condition>
```


Jupyter Notebook Tutorial

- [Jupyter tutorial dir](#)
- [Readme](#)
- [Notebook to execute / study](#)

SQL Tutorial

- [SQL tutorial dir](#)
- [Readme](#)
 - Explains how to run the tutorial
- Three notebooks in tutorial_university
- **How to learn from a tutorial**
 - Reset the notebook
 - Execute each cell one at the time
 - Understand what each cell does
 - Look at the output
 - Change the code
 - Play with it
 - Build your mental model
 - Ideally create a new file and retype (!) everything

	sql_basics.ipynb
	sql_joins.ipynb
	sql_nulls_and_unknown.ipynb

Create Table

```
CREATE TABLE r
  (A_1 D_1,
   A_2 D_2,
   ...
   A_n D_n,
   IntegrityConstraint_1,
   IntegrityConstraint_n);
```

- where:
 - r is name of *table* (aka *relation*)
 - A_i name of *attribute* (aka *field*, *column*)
 - D_i domain of attribute A_i
- Constraints:
 - Primary key
 - Need to be all non-null and unique
 - **PRIMARY KEY (A_j1, A_j2, ..., A_jn)**
 - Foreign key
 - Values of attributes for any tuple in current relation must correspond to values of the primary key attributes of some tuple in relation s
 - **FOREIGN KEY (A_k1, A_k2, ..., A_kn) REFERENCES s**
 - Not null
 - Specify that null value is not allowed for that attribute
 - **A_i D_i NOT NULL**
- SQL will prevent changes to the DB that violate any integrity constraint

Select

```
SELECT A_1, A_2, ..., A_n  
      FROM r_1, r_2, ..., r_m  
      WHERE P;
```

- SELECT: select the attributes to list (i.e., projection)
- FROM: list of tables to be accessed
 - Define a Cartesian product of the tables
 - The query is going to be optimized to avoid to enumerate tuples that will be eliminated
- WHERE: predicate involving attributes of the relations in the FROM clause (i.e., selection)
- In SELECT or WHERE clauses, might need to use the table names as prefix to qualify the attribute name
 - E.g., instructor.ID vs teaches.ID
- A SELECT statement can be expressed in terms of relational algebra
 - Cartesian product -> selection -> projection
 - Difference: SQL allows duplicated values, relational algebra works with mathematical sets

Null values

- An arithmetic operation with NULL yields NULL
- Comparison with NULL
 - $1 < \text{NULL}$?
 - What about $\text{NOT}(1 < \text{NULL})$?
 - SQL yields UNKNOWN when comparing with NULL value
 - There are 3 logical values: True, False, Unknown
- Boolean operators
 - Can be extended according to common sense, e.g.,
 - $\text{True AND Unknown} = \text{Unknown}$
 - $\text{False AND Unknown} = \text{False}$
- In a WHERE clause, if the result is Unknown it's not included

Group by Query

- The attributes in GROUP BY are used to form groups
 - Tuples with the same value on all attributes are placed in one group
- Any attribute that is not in the GROUP BY can appear in the SELECT clause only as argument of aggregate function

ID	name	dept_name	salary
76766	Crick	Biology	72000
45565	Katz	Comp. Sci.	75000
10101	Srinivasan	Comp. Sci.	65000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000
12121	Wu	Finance	90000
76543	Singh	Finance	80000
32343	El Said	History	60000
58583	Califieri	History	62000
15151	Mozart	Music	40000
33456	Gold	Physics	87000
22222	Einstein	Physics	95000

```
SELECT dept_name, AVG(salary)
FROM instructor
GROUP BY dept_name;
```

-- Error.

```
SELECT dept_name, salary
FROM instructor
GROUP BY dept_name;
```

dept_name	avg_salary
Biology	72000
Comp. Sci.	77333
Elec. Eng.	80000
Finance	85000
History	61000
Music	40000
Physics	91000

- salary is not in GROUP BY so it must be in an aggregate function

Having

- State a condition that applies to groups instead of tuples (like WHERE)
- Any attribute in the HAVING clause must appear in the GROUP BY clause
- E.g., find departments with avg salary of instructors > \$42k

```
SELECT dept_name, AVG(salary) AS avg_salary
FROM instructor
GROUP BY dept_name
HAVING AVG(salary) > 42000;
```

- How does it work
 - FROM is evaluated to create a relation
 - (optional) WHERE is used to filter
 - GROUP BY collects tuples into groups
 - (optional) HAVING is applied to each group and groups are filtered
 - SELECT generates tuples of the results, applying aggregate functions to get a single result for each group

Nested subqueries

- SQL allows to use the result of a query in another query
 - E.g., one can use a subquery returning only one attribute (aka scalar subquery) in any place a value is used
 - E.g., use the result of a query for set membership in the WHERE clause
 - E.g., use the result of a query in a FROM clause

```
SELECT tmp.dept_name, tmp.avg_salary
      FROM (SELECT dept_name, AVG(salary) AS
avg_salary
            FROM instructor
            GROUP BY dept_name) AS tmp
WHERE avg_salary > 42000
```

dept_name	avg_salary
Finance	85000.000000000000
History	61000.000000000000
Physics	91000.000000000000
Comp. Sci.	77333.333333333333
Biology	72000.000000000000
Elec. Eng.	80000.000000000000

With

- WITH clause allows to define a temporary relation containing the results of a subquery
- It can be equivalent to a nested subqueries, but clearer
- Find department with the maximum budget.

```
WITH max_budget(value) as (  
    SELECT MAX(budget) FROM department)  
SELECT department.dept_name, budget  
    FROM department, max_budget  
    WHERE department.budget = max_budget.value
```

dept_name	budget
Finance	120000.00

Delete

- One can delete tuples using a query returning entire rows of a table

DELETE FROM r WHERE p

where:

- r is a relation
- P is a predicate
- Remove all tuples (but not the table)

DELETE FROM instructor

Insert

To insert data into a relation we can specify tuples to insert

- Tuples

```
INSERT INTO course VALUES ('DATA-605', 'Big data systems', 'Comp. Sci.', 4)
INSERT INTO course(course_id, title, dept_name, credits) VALUES ('DATA-605',
'Big data systems', 'Comp. Sci.', 4)
```

- Query whose results is a set of tuples

```
INSERT INTO instructor (
    SELECT ID, name, dept_name, 18000
    FROM student
    WHERE dept_name = 'Music' AND tot_cred > 144)
```

- Nested queries are evaluated and then inserted so this doesn't create infinite workload

```
INSERT INTO student (SELECT * FROM student)
```

- Many DB have bulk loader utilities to insert a large set of tuples into a relation, reading from formatted text files
 - This is much faster than INSERT statements

Update

- SQL can change a value in a tuple without changing all the other values
- E.g., increase salary of all instructors by 5%

UPDATE instructor

SET salary = salary * 1.05

- E.g., conditionally

UPDATE instructor

SET salary = salary * 1.05

WHERE salary < 70000

- Nesting is allowed

UPDATE instructor

SET salary = salary * 1.05

WHERE salary < (SELECT AVG(salary) FROM instructor)

UMD DATA605 - Big Data Systems

Movie Database Example (Optional)

Example Schema for SQL Queries

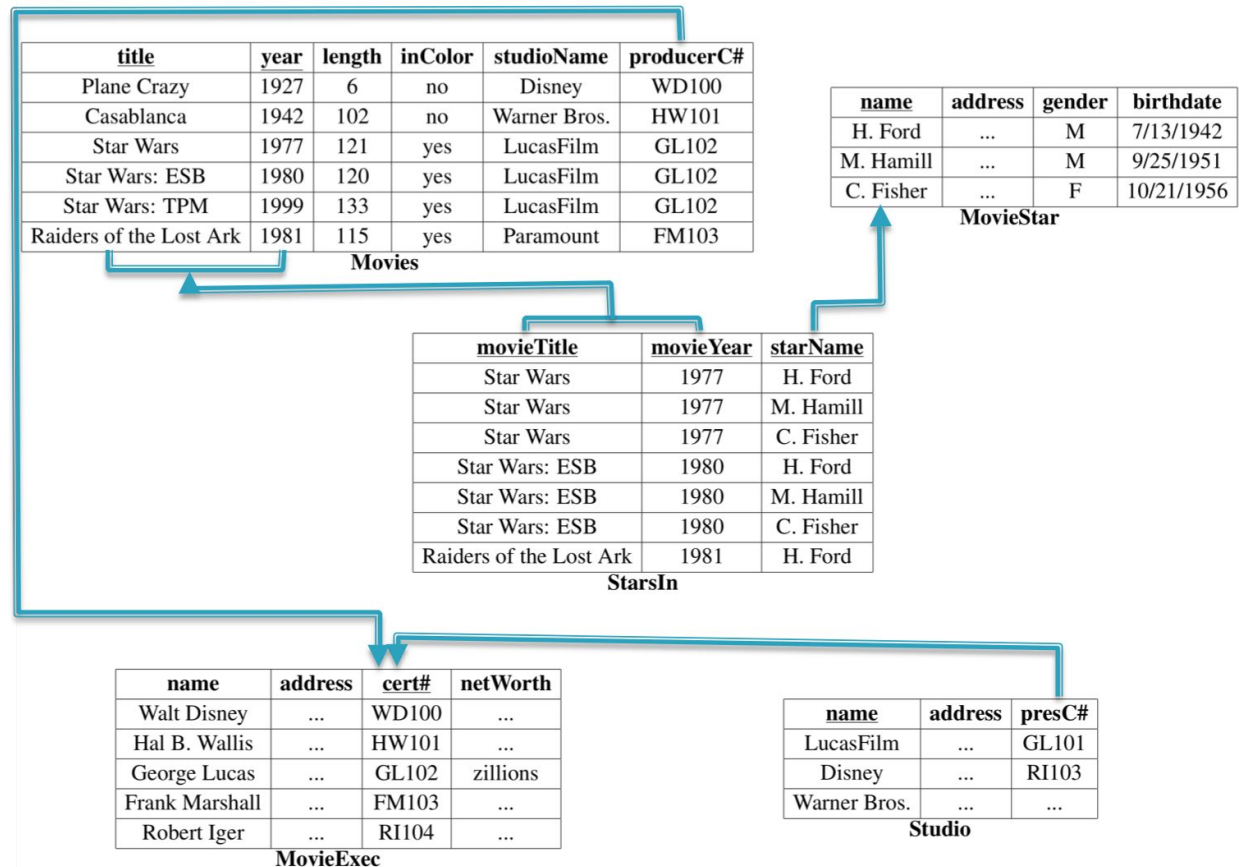
Movie(title, year, length, inColor, studioName, producerC#)

StarsIn(movieTitle, movieYear, starName)

MovieStar(name, address, gender, birthdate)

MovieExec(name, address, cert#, netWorth)

Studio(name, address, presC#)



SQL: Data Definition

- CREATE TABLE

```
CREATE TABLE movieExec (  
    name char(30),  
    address char(100),  
    cert# integer primary key,  
    networth integer);  
  
CREATE TABLE movie (  
    title char(100),  
    year integer,  
    length integer,  
    inColor smallint,  
    studioName char(20),  
    producerC# integer references  
        movieExec(cert#) );
```

- Must define movieExec before movie. Why?

SQL: Data Manipulation

- INSERT

```
INSERT INTO StarsIn values('King Kong', 2005, 'Naomi Watts');  
INSERT INTO StarsIn(starName, movieTitle, movieYear)  
values('Naomi Watts', 'King Kong', 2005);
```

- DELETE

```
DELETE FROM movies WHERE movieYear < 1980;
```

- Syntax is fine, but this command will be rejected. Why?

```
DELETE FROM movies WHERE length < (  
    SELECT avg(length) FROM movies);
```

- Problem: as we delete tuples, the average length changes
- Solution:
 - First, compute avg length and find all tuples to delete
 - Next, delete all tuples found above (without recomputing avg or retesting the tuples)

SQL: Data Manipulation

- UPDATE

- Increase all movieExec netWorth's over \$100,000 by 6%, all other accounts receive 5%

- Write two update statements:

```
UPDATE movieExec SET netWorth = netWorth * 1.06 WHERE netWorth > 100000;  
UPDATE movieExec SET netWorth = netWorth * 1.05 WHERE netWorth <= 100000;
```

- The order is important
- Can be done better using the case statement

```
UPDATE movieExec  
SET netWorth =  
CASE  
    WHEN netWorth > 100000  
        THEN netWorth * 1.06  
    WHEN netWorth <= 100000  
        THEN netWorth * 1.05  
END;
```

SQL Single Table Queries

- Movies produced by Disney in 1990: note the *rename*

```
SELECT m.title, m.year  
FROM movie m  
WHERE m.studioname = 'disney' AND m.year = 1990;
```

- The **SELECT** clause can contain expressions

```
SELECT title || ' (' || to_char(year) || ')' AS titleyear  
SELECT 2014 - year
```

- The **WHERE** clause support a large number of different predicates and combinations thereof

```
year BETWEEN 1990 and 1995  
title LIKE 'star wars%'  
title LIKE 'star wars _'
```

Single Table Queries

- Find distinct movies sorted by title

```
SELECT DISTINCT title
FROM movie
WHERE studioname = 'disney' AND year = 1990
ORDER by title;
```

- Average length of a movie

```
SELECT year, avg(length)
FROM movie
GROUP BY year;
```

- **GROUP BY:** is a very important concept that shows up in many data processing platforms

- What it does:

- Partition the tuples by the group attributes (*year* in this case)
- Do something (*compute avg* in this case) for each group
- Number of resulting tuples == number of groups

Single Table Queries

- Find movie with the maximum length

```
SELECT title, year
FROM movie
where movie.length = (select max(length) from movie);
```

- The smaller "subquery" is called a "nested subquery"

- Find movies with at most 5 stars: an example of a correlated subquery

```
SELECT *
FROM movies m
WHERE 5 >= (SELECT count(*)
            FROM starsIn si
            WHERE si.title = m.title AND
                  si.year = m.year);
```

- The "inner" subquery counts the number of actors for that movie.

Single Table Queries

- Rank movies by their length

```
SELECT title, year,  
       (SELECT count(*)  
        FROM movies m2  
        WHERE m1.length <= m2.length) AS rank  
FROM movies m1;
```

- Key insight: A movie is ranked 5th if there are exactly 4 movies with longer length.
- Most database systems support some sort of a *rank* keyword for doing this
- The above query doesn't work in presence of ties, etc.

- Set operations

```
SELECT name  
FROM movieExec  
union/intersect/minus  
  SELECT name FROM  
  movieStar
```

Single Table Queries

- Set Comparisons

```
SELECT *  
FROM movies  
WHERE year IN [1990, 1995, 2000];
```

```
SELECT *  
FROM movies  
WHERE year NOT IN (  
    SELECT EXTRACT(year from birthdate)  
    FROM MovieStar  
);
```

Multi-table Queries

- Key:
 - Do a join to get an appropriate table
 - Use the constructs for single-table queries
 - You will get used to doing all at once
- Examples:

```
SELECT title, year, me.name AS producerName  
FROM movies m, movieexec me  
WHERE m.producerC# = me.cert#;
```

Multi-table Queries

- Consider the query:

```
select title, year, producerC#, count(starName)
from movies, starsIn
where title = starsIn.movieTitle and
       year = starsIn.movieYear
group by title, year, producerC#
```

- What about movies with no stars?
- Need to use **outer joins**

```
select title, year, producerC#, count(starName)
from movies left outer join starsIn
on title = starsIn.movieTitle and year = starsIn.movieYear
group by title, year, producerC#
```

- All tuples from 'movies' that have no matches in starsIn are included with NULLs
- So if a tuple (m1, 1990) has no match in starsIn, we get (m1, 1990, NULL) in the result
- The count(starName) works correctly then
- Note: count(*) would not work correctly (NULLs can have unintuitive behavior)

Other SQL Constructs

- Views

```
create view DisneyMovies
```

```
select *
```

```
from movie m
```

```
where m.studioname = 'disney';
```

- Can use it in any place where a table name is used
- Views are used quite extensively to: (1) simplify queries, (2) hide data (by giving users access only to specific views)
- Views may be *materialized* or not

Other SQL Constructs

- NULLs

- Value of any attribute can be NULL
 - Because: value is unknown, or it is not applicable, or hidden, etc.
- Can lead to counterintuitive behavior
- For example, the following query does not return movies where length = NULL

```
select * from movies
where length >= 120 or length <= 120
```
- Aggregate operations can be especially tricky

- Transactions

- A transaction is a sequence of queries and update statements executed as a single unit
- For example, transferring money from one account to another
 - Both the *deduction* from one account and *credit* to the other account should happen, or neither should

- Triggers

- A trigger is a statement that is executed automatically by the system as a side effect of a modification to the database

Other SQL Constructs

- Integrity Constraints

- Predicates on the database that must always hold
- Key Constraints: Specifying something is a primary key or unique

```
CREATE TABLE customer (  
    ssn CHAR(9) PRIMARY KEY,  
    cname CHAR(15), address CHAR(30), city CHAR(10),  
    UNIQUE (cname, address, city));
```

- Attribute constraints: Constraints on the values of attributes

```
bname char(15) not null  
balance int not null, check (balance >= 0)
```

Integrity Constraints

- Referential integrity: prevent dangling tuples

```
CREATE TABLE branch(  
  bname CHAR(15) PRIMARY KEY,  
  ...);  
CREATE TABLE loan(...,  
  FOREIGN KEY bname REFERENCES branch);
```

- Can tell the system what to do if a referenced tuple is being deleted

Integrity Constraints

– Global Constraints

- Single-table

```
CREATE TABLE branch (...,  
    bcity CHAR(15),  
    assets INT,  
    CHECK (NOT(bcity = 'Bkln') OR assets>5M))
```

- Multi-table

```
CREATE ASSERTION loan-constraint  
CHECK (NOT EXISTS (  
    SELECT*  
    FROM loan AS L  
    WHERE NOT EXISTS(  
        SELECT*  
        FROM borrower B, depositor D, account A  
        WHERE B.cname = D.cname AND  
              D.acct_no = A.acct_no AND  
              L.lno= B.lno)))
```

Additional SQL Constructs

- **Select** subquery factoring

- To allow assigning a name to a subquery, then use its result by referencing that name

```
with temp as (  
    select title, avg(length)  
    from movies  
    group by year)  
select count(*) from temp;
```

- Can have multiple subqueries (multiple with clauses)
- Real advantage is when subquery needs to be referenced multiple times in main select
- Helps with complex queries, both for readability and maybe performance (can cache subquery results)

Another SQL Construct

- **Select** having clause

- Used in combination with **group by** to restrict the groups of returned rows to only those where condition evaluates to true

```
select year, count(*)  
  from movies where year > 1980  
  group by year  
  having count(*) > 10;
```

- Difference from **where** clause is that it applies to summarized group records, and where applies to individual records