W4111 – Introduction to Databases
Section 002, Spring 2021

Lecture 3: ER, Relational, SQL (III)

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

W4111 – Introduction to Databases
Section 002, Spring 2021

Lecture 3: ER, Relational, SQL (III)

We will start in a couple of minutes.

Contents

Contents

- Introduction: comments, questions and answers
- Relational Model, Relational Algebra (Completion)
- Intermediate SQL
- Intermediate ER Modeling and SQL Realization
- Class Project/Homework Directions/Discussion

I will cover in recitation.

Questions, Answers, Comments

Homework 1 – “Getting out over the skis”

- Change in Spring ‘21 course approach based on student feedback:
  - I previously expected students to self-learn SQL, and other concepts.
    - Lots of material in book, tutorials, etc.
    - Listening to some present slides is not a good approach.
    - You learn by practice, doing, ...
  - My lecture focus was on topics that you cannot easily learn from books, tutorials, ...
  - End Fall ’20 student feedback expressed dissatisfaction with the approach. Lecture material was not covering homework material.
  - I am pivoting to better cover SQL, core concepts, etc.
  - We did not adequately factor the change in direction into HW 1.

- I go through homework problems in recitations.
  - Any code that I write is a “hint” or “directional guidance.”
  - Please do not expect to simply transcribe what I type, expect it to work, and then ask me to fix it for you if it does not.

Piazza Questions

- Switch to Jupyter Notebook

Result Tables; Derived Tables

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

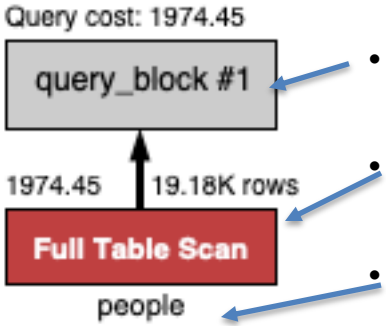
The SQL SELECT Statement

The SELECT statement is used to select data from a database.

The data returned is stored in a result table, called the result-set.

SELECT Syntax

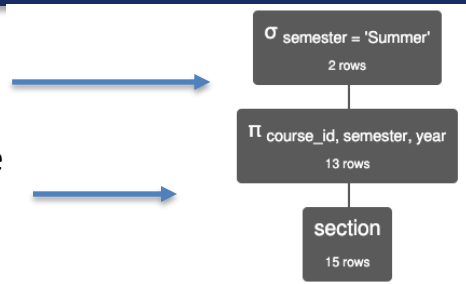
```
SELECT column1, column2, ...
FROM table_name;
```



- Result Table
- Operation
- Base Table

Not exactly what the diagram means but it explains the concepts.

- Result Table
- Derived Table
- Base Table

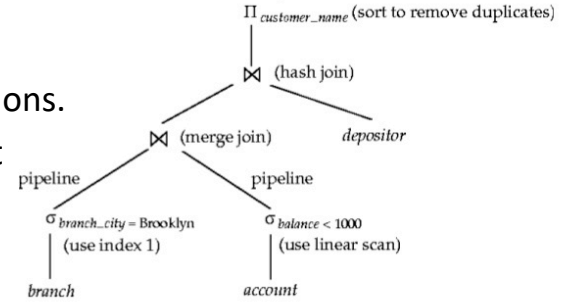


σ semester = 'Summer' (π course_id, semester, year (section))

Query evaluation plan

- An **evaluation plan** defines exactly what algorithm is used for each operation, and how the execution of the operations is coordinated

- Nodes represent relational operations.
- Arrows represent flow of tables.



The Relational Model

The Relational Data Model (Completed)

Relational Algebra

Relational Algebra

- A procedural language consisting of a set of operations that take one or two relations as input and produce a new relation as their result.
- Six basic operators
  - ~~select: σ~~
  - ~~project: Π~~
  - union: $\cup$
  - set difference: –
  - Cartesian product: x
  - rename: ρ
  - intersect
  - $\leftarrow$ assignment

Last Lecture

Composition of Relational Operations

- The result of a relational-algebra operation is relation and therefore of relational-algebra operations can be composed together into a **relational-algebra expression**.
- Consider the query -- Find the names of all instructors in the Physics department.

$$\Pi_{name}(\sigma_{dept\_name = \text{“Physics”}}(instructor))$$

- Instead of giving the name of a relation as the argument of the projection operation, we give an expression that evaluates to a relation.

Why highlight composition before completing coverage of the operations?

- The next core relational operator most people consider is JOIN.
- The definition of JOIN relies on operation composition.

Cartesian-Product Operation

- The Cartesian-product operation (denoted by X) allows us to combine information from any two relations.
- Example: the Cartesian product of the relations *instructor* and t*eaches* is written as:

  instructor X *teaches*

- We construct a tuple of the result out of each possible pair of tuples: one from the *instructor* relation and one from the *teaches* relation (see next slide)
- Since the instructor *ID* appears in both relations we distinguish between these attribute by attaching to the attribute the name of the relation from which the attribute originally came.
  - *instructor.ID*
  - *teaches.ID*

The *instructor* x *teaches* table

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
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…

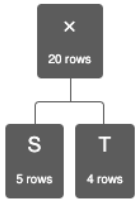
Simpler Example

T

T.b	T.d
'a'	100
'd'	200
'f'	400
'g'	120

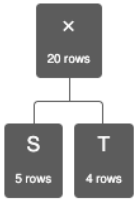
S

S.b	S.d
'a'	100
'b'	300
'c'	400
'd'	200
'e'	150



S × T

S.b	S.d	T.b	T.d
'a'	100	'a'	100
'a'	100	'd'	200
'a'	100	'f'	400
'a'	100	'g'	120
'b'	300	'a'	100
'b'	300	'd'	200
'b'	300	'f'	400
'b'	300	'g'	120
'c'	400	'a'	100
'c'	400	'd'	200



S × T

S.b	S.d	T.b	T.d
'c'	400	'f'	400
'c'	400	'g'	120
'd'	200	'a'	100
'd'	200	'd'	200
'd'	200	'f'	400
'd'	200	'g'	120
'e'	150	'a'	100
'e'	150	'd'	200
'e'	150	'f'	400
'e'	150	'g'	120

- Assume we have two tables
  - S has two columns, 5 rows.
  - T has two columns, 4 rows.
- S x T has
  - 4 columns.
  - 20 rows.
- Cartesian product does not come up a lot in applications.
- There are cases in optimization in which:
  - You want to generate all possible combinations.
  - Score, rate, rank etc. to determine best choices.

Join Operation

- The Cartesian-Product

  instructor X *teaches*

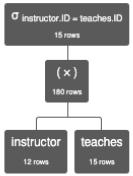
  associates every tuple of instructor with every tuple of teaches.
  - Most of the resulting rows have information about instructors who did NOT teach a particular course.
- To get only those tuples of "*instructor* X *teaches* " that pertain to instructors and the courses that they taught, we write:

  $\sigma_{instructor.id = teaches.id}$ (*instructor* x *teaches*))
  - We get only those tuples of "*instructor* X *teaches*" that pertain to instructors and the courses that they taught.
- The result of this expression, shown in the next slide

A fundamental definition:

- $\sigma_{instructor.ID=teaches.ID}$ (instructor × teaches) = instructor ⋈ teaches
- ⋈ is the JOIN operations.

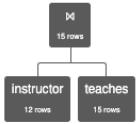
JOIN Definition



σ instructor.ID = teaches.ID (instructor × teaches)

instructor.ID	instructor.name	instructor.dept_name	instructor.salary	teaches.ID	teaches.course_id	teaches.sec_id	teaches.semester	teaches.year
10101	'Srinivasan'	'Comp. Sci.'	65000	10101	'CS-101'	1	'Fall'	2009
10101	'Srinivasan'	'Comp. Sci.'	65000	10101	'CS-315'	1	'Spring'	2010
10101	'Srinivasan'	'Comp. Sci.'	65000	10101	'CS-347'	1	'Fall'	2009
12121	'Wu'	'Finance'	90000	12121	'FIN-201'	1	'Spring'	2010
15151	'Mozart'	'Music'	40000	15151	'MU-199'	1	'Spring'	2010
22222	'Einstein'	'Physics'	95000	22222	'PHY-101'	1	'Fall'	2009
32343	'El Said'	'History'	60000	32343	'HIS-351'	1	'Spring'	2010
45565	'Katz'	'Comp. Sci.'	75000	45565	'CS-101'	1	'Spring'	2010
45565	'Katz'	'Comp. Sci.'	75000	45565	'CS-319'	1	'Spring'	2010
76766	'Crick'	'Biology'	72000	76766	'BIO-101'	1	'Summer'	2009

σ instructor.ID=teaches.ID (instructor × teaches)



instructor ⋈ teaches

instructor.ID	instructor.name	instructor.dept_name	instructor.salary	teaches.course_id	teaches.sec_id	teaches.semester	teaches.year
10101	'Srinivasan'	'Comp. Sci.'	65000	'CS-101'	1	'Fall'	2009
10101	'Srinivasan'	'Comp. Sci.'	65000	'CS-315'	1	'Spring'	2010
10101	'Srinivasan'	'Comp. Sci.'	65000	'CS-347'	1	'Fall'	2009
12121	'Wu'	'Finance'	90000	'FIN-201'	1	'Spring'	2010
15151	'Mozart'	'Music'	40000	'MU-199'	1	'Spring'	2010
22222	'Einstein'	'Physics'	95000	'PHY-101'	1	'Fall'	2009
32343	'El Said'	'History'	60000	'HIS-351'	1	'Spring'	2010
45565	'Katz'	'Comp. Sci.'	75000	'CS-101'	1	'Spring'	2010
45565	'Katz'	'Comp. Sci.'	75000	'CS-319'	1	'Spring'	2010
76766	'Crick'	'Biology'	72000	'BIO-101'	1	'Summer'	2009

instructor ⋈ teaches

Join Operation (Cont.)

- The table corresponding to:

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

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

Join Operation (Cont.)

- The **join** operation allows us to combine a select operation and a Cartesian-Product operation into a single operation.
- Consider relations *r* (*R*) and *s* (*S*)
- Let "theta" be a predicate on attributes in the schema R "union" S. The join operation r $\bowtie_\theta$ s is defined as follows:

$$r \bowtie_\theta s = \sigma_\theta (r \times s)$$

- Thus

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

- Can equivalently be written as

$$instructor \bowtie_{Instructor.id = teaches.id} teaches.$$

Union Operation

- The union operation allows us to combine two relations
- Notation: $r \cup s$
- For $r \cup s$ to be valid.
  1. *r, s* must have the *same* **arity** (same number of attributes)
  2. The attribute domains must be **compatible** (example: 2nd column of *r* deals with the same type of values as does the 2nd column of *s*)
- Example: to find all courses taught in the Fall 2017 semester, or in the Spring 2018 semester, or in both

$$\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2017}(section)) \cup \Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2018}(section))$$

Union Operation (Cont.)

- Result of:

$$\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2017}(section)) \cup \Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2018}(section))$$

course_id
CS-101
CS-315
CS-319
CS-347
FIN-201
HIS-351
MU-199
PHY-101

Note: The preloaded dataset on the RelaX calculator is different from the most recent data referenced in the book. It is from a previous edition.

Set-Intersection Operation

- The set-intersection operation allows us to find tuples that are in both the input relations.
- Notation: $r \cap s$
- Assume:
  - r, s have the *same arity*
  - attributes of r and s are compatible
- Example: Find the set of all courses taught in both the Fall 2017 and the Spring 2018 semesters.

$$\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2017}(section)) \cap \Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2018}(section))$$

  - Result

course_id
CS-101

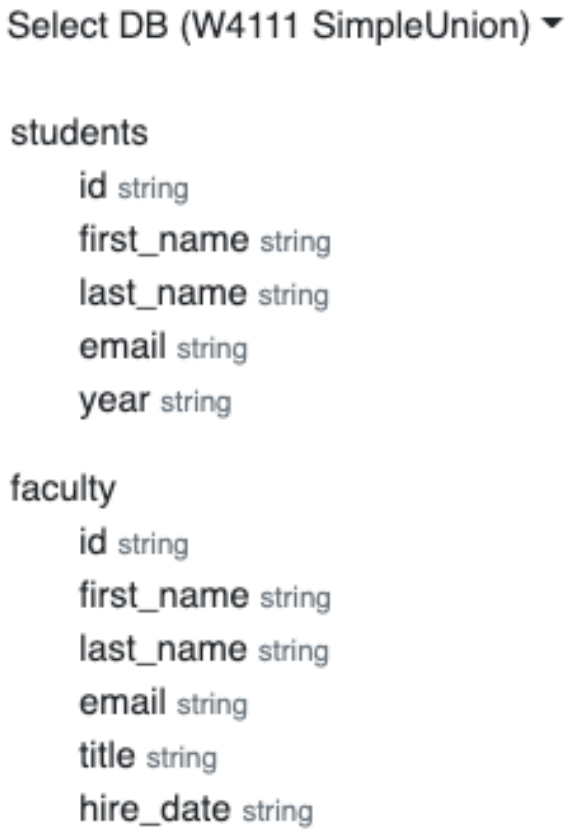
Set Difference Operation

- The set-difference operation allows us to find tuples that are in one relation but are not in another.
- Notation $r - s$
- Set differences must be taken between **compatible** relations.
  - r and s must have the same arity
  - attribute domains of r and s must be compatible
- Example: to find all courses taught in the Fall 2017 semester, but not in the Spring 2018 semester

$$\Pi_{course\_id}(\sigma_{semester="Fall" \wedge year=2017}(section)) - \Pi_{course\_id}(\sigma_{semester="Spring" \wedge year=2018}(section))$$

course_id
CS-347
PHY-101

Same “arity”



- Same “arity”
  - Same number of columns.
  - Compatible types.
  - The i-th column in each table is from a compatible domain.
    - Student 5th column is “year.”
    - Faculty 5th column is “title”
    - Both are strings but combing them does not make sense.
- You can shape two incompatible tables using *project operations.* For example
  - π first_name,last_name,email (students)
    ∩
    π first_name,last_name,email (faculty)
  - π last_name,email, title←'Student' (students)
    ∪
    π last_name,email, title (faculty)

The Assignment Operation

- It is convenient at times to write a relational-algebra expression by assigning parts of it to temporary relation variables.
- The assignment operation is denoted by ← and works like assignment in a programming language.
- Example: Find all instructor in the "Physics" and Music department.

$$Physics \leftarrow \sigma_{dept\_name="Physics"}(instructor)$$

$$Music \leftarrow \sigma_{dept\_name="Music"}(instructor)$$

$$Physics \cup Music$$

- With the assignment operation, a query can be written as a sequential program consisting of a series of assignments followed by an expression whose value is displayed as the result of the query.

The Rename Operation

- The results of relational-algebra expressions do not have a name that we can use to refer to them. The rename operator, ρ, is provided for that purpose
- The expression:

  $\rho_x(E)$

  returns the result of expression E under the name x
- Another form of the rename operation:

  $\rho_{x(A1,A2,\ldots An)}(E)$

Note: Assignment and rename can act a little wonky when using the calculator.

Equivalent Queries

- There is more than one way to write a query in relational algebra.
- Example: Find information about courses taught by instructors in the Physics department with salary greater than 90,000
- Query 1

$$\sigma_{dept\_name=\text{“Physics”} \wedge salary > 90{,}000}(instructor)$$

- Query 2

$$\sigma_{dept\_name=\text{“Physics”}}(\sigma_{salary > 90.000}(instructor))$$

- The two queries are not identical; they are, however, equivalent -- they give the same result on any database.

Equivalent Queries

- There is more than one way to write a query in relational algebra.
- Example: Find information about courses taught by instructors in the Physics department
- Query 1

$$\sigma_{dept\_name=\text{“Physics”}}(instructor \bowtie_{instructor.ID = teaches.ID} teaches)$$

- Query 2

$$(\sigma_{dept\_name=\text{“Physics”}}(instructor)) \bowtie_{instructor.ID = teaches.ID} teaches$$

- The two queries are not identical; they are, however, equivalent -- they give the same result on any database.

What are all those other Symbols?

- τ order by
- γ group by
- ¬ negation
- ÷ set division
- ⋈ natural join, theta-join
- ⟕ left outer join
- ⟖ right outer join
- ⟗ full outer join
- ⋉ left semi join
- ⋊ right semi join
- ▷ anti-join

- Some of these are pretty obscure
  - Division
  - Anti-Join
  - Left semi-join
  - Right semi-join
- Most SQL engines do not support them.
  - You can implement them using combinations of JOIN, SELECT, WHERE, ...
  - But, I cannot every remember using them in applications I have developed.
- Outer JOIN is very useful, but less common. We will cover.
- There are also some “patterns” or “terms”
  - Equijoin
  - Non-equi join
  - Natural join
  - Theta join
  - ... ...
- I may ask you to define these terms on some exams because they may be common internships/job interview questions.

SQL (II)

Outline

- ~~Overview of The SQL Query Language~~
- ~~SQL Data Definition~~
- ~~Basic Query Structure of SQL Queries~~
- Additional Basic Operations
- Set Operations
- ~~Null Values~~
- Aggregate Functions
- Nested Subqueries
- Modification of the Database

Covered Last Lecture

Additional Operations

We have to discuss types/domains before talking about some of the other material.

Data Types and Functions

Domain Types in SQL

- **char(n).** Fixed length character string, with user-specified length *n.*
- **varchar(n).** Variable length character strings, with user-specified maximum length *n.*
- **int.** Integer (a finite subset of the integers that is machine-dependent).
- **smallint.** Small integer (a machine-dependent subset of the integer domain type).
- **numeric(p,d).** Fixed point number, with user-specified precision of *p* digits, with *d* digits to the right of decimal point. (ex., **numeric**(3,1), allows 44.5 to be stores exactly, but not 444.5 or 0.32)
- **real, double precision.** Floating point and double-precision floating point numbers, with machine-dependent precision.
- **float(n).** Floating point number, with user-specified precision of at least *n* digits.
- More are covered in Chapter 4.

Built-in Data Types in SQL

- **date:** Dates, containing a (4 digit) year, month and date
  - Example: **date** '2005-7-27'
- **time:** Time of day, in hours, minutes and seconds.
  - Example: **time** '09:00:30' **time** '09:00:30.75'
- **timestamp:** date plus time of day
  - Example: **timestamp** '2005-7-27 09:00:30.75'
- **interval:** period of time
  - Example: interval '1' day
  - Subtracting a date/time/timestamp value from another gives an interval value
  - Interval values can be added to date/time/timestamp values

Note:

- All implementations of SQL support a common core, but ...
- The have proprietary extensions: operators on types, additional types.

Large-Object Types

- Large objects (photos, videos, CAD files, etc.) are stored as a *large object*:
  - **blob**: binary large object -- object is a large collection of uninterpreted binary data (whose interpretation is left to an application outside of the database system)
  - **clob**: character large object -- object is a large collection of character data
- When a query returns a large object, a pointer is returned rather than the large object itself.

Note:

- CLOBs and BLOBs are very uncommon.
- Today applications keep documents, images, CAD files, etc. in “block/pbject storage.”
- The database contains just the hyperlinks to the files/content.
- We will see some examples when we talk about web applications.

MySQL Cheat Sheet

https://websitesetup.org/wp-content/uploads/2020/04/MySQL-Cheat-Sheet-websitesetup.org.pdf
Pretty good, concise place to find information about MySQL.

https://dev.mysql.com/doc/refman/8.0/en/data-types.html

Data Types

Data types indicate what type of information you can store in a particular column of your table. MySQL has three main categories of data types:

- Numeric
- Text
- Date/time

These lists are not complete.

Blob and Text Data Types

BLOB binary range enables you to store larger amounts of text data. The maximum length of a BLOB is **65,535 ($2^{16} - 1$)** bytes. BLOB values are stored using a 2-byte length prefix.

NB: Since text data can get long, always double-check that you do not exceed the maximum lengths. The system will typically generate a warning if you go beyond the limit. But if nonspace characters get truncated, you may just receive an error without a warning.

- **TINYBLOB** – sets the maximum column length at 255 ($2^8 - 1$) bytes. TINYBLOB values are stored using a 1-byte length prefix.
- **MEDIUMBLOB** – sets the maximum column length at 16,777,215 ($2^{24} - 1$) bytes. MEDIUMBLOB values are stored using a 3-byte length prefix.
- **LONGBLOB** – sets the maximum column length at 4,294,967,295 or 4GB ($2^{32} - 1$) bytes. LONGBLOB values are stored using a 4-byte length prefix.

  Note: The max length will also depend on the maximum packet size that you configure in the client/server protocol, plus available memory.

If unsigned, the column will expand to hold the data up till a certain upper boundary range.

- **BIT[(M)]** — specify a bit-value type. **M** stands for the number of bits per value, ranging from 1 to 64. The default is 1 if no T specified.
- **ZEROFILL** — auto-add UNSIGNED attribute to the column. Deprecated since the MySQL 8.0.17 version.
- **TINYINT(M)** — the smallest integer with a range of -128 to 127.
  - **TINYINT(M) [UNSIGNED]** — the range is 0 to 255.
  - **BOOL, BOOLEAN** — synonyms for TINYINT(1)
- **SMALLINT(M)** — small integer with a range of -32768 and 32767.
  - **SMALLINT(M) [UNSIGNED]** — the range is 0 to 65535.
- **MEDIUMINT(M)** — medium integer with a range of -8388608 to 8388607.
  - **MEDIUMINT(M) [UNSIGNED]** — the range is 0 to 16777215.
- **INT(M) and INTEGER (M)** — normal range integer with a range of -2147483648 to 2147483647.
  - **INT(M)[UNSIGNED] and INTEGER (M)[UNSIGNED]** — the range is 0 to 4294967295.
- **BIGINT(M)** — the largest integer with a range of -9223372036854775808 to 9223372036854775807.
  - **BIGINT(M) [UNSIGNED]** — the range is 0 to 844674073709551615.
- **DECIMAL (M, D)** — store a double value as a string. **M** specifies the total number of digits. **D** stands for the number of digits after the decimal point. Handy for storing currency values.
  - Max number of M is 65. If omitted, the default M value is 10.
  - Max number of D is 30. If omitted, the default D is 0.
- **FLOAT (M, D)** — record an approximate number with a floating decimal point. The support for FLOAT is removed as of MySQL 8.0.17 and above.
  - Permissible values ranges are -3.402823466E+38 to -1.175494351E-38, 0, and 1.175494351E-38 to 3.402823466E+38.

MySQL Cheat Sheet

https://websitesetup.org/wp-content/uploads/2020/04/MySQL-Cheat-Sheet-websitesetup.org.pdf
Pretty good, concise place to find information about MySQL.

https://dev.mysql.com/doc/refman/8.0/en/data-types.html

Text Storage Formats

- **CHAR** — specifies the max number of non-binary characters you can store. The range is from 0 to 255.
- **VARCHAR** — store variable-length non-binary strings. The maximum number of characters you can store is 65,535 (equal to the max row size).
  - VARCHAR values are stored as a 1-byte or 2-byte length prefix plus data, unlike CHAR values.
- **BYNARY** — store binary data in the form of byte strings. Similar to CHAR.
- **VARBYNARY** — store binary data of variable length in the form of byte strings. Similar to VARCHAR.
- **ENUM** — store permitted text values that you enumerated in the column specification when creating a table.
  - ENUM columns can contain a maximum of 65,535 distinct elements and have > 255 unique element list definitions among its ENUM.
- **SET** — another way to store several text values that were chosen from a predefined list of values.
  - SET column can contain a maximum of 64 distinct members and have > 255 unique element list definitions among its SET.

Date and Time Data Types

As the name implies, this data type lets you store the time data in different formats.

- **DATE** — use it for values with a date part only. MySQL displays DATE values in the 'YYYY-MM-DD' format.
  - Supported data range is '1000-01-01' to '9999-12-31'.
- **DATETIME** — record values that have both date and time parts. The display format is 'YYYY-MM-DD hh:mm:ss'.
  - Supported data range is '1000-01-01 00:00:00' to '9999-12-31 23:59:59'.
- **TIMESTAMP** — add more precision to record values that have both date and time parts, up till microseconds in UTC.
  - Supported data range is '1970-01-01 00:00:01' UTC to '2038-01-19 03:14:07' UTC.
- **TIME** — record just time values in either 'hh:mm:ss' or 'hhh:mm:ss' format. The latter can represent elapsed time and time intervals.
  - Supported data range is '-838:59:59' to '838:59:59'.
- **YEAR** — use this 1-byte type used to store year values.
  - A 4-digit format displays YEAR values as 0000, with a range between 1901 to 2155.
  - A 2-digit format displays YEAR values as 00. The accepted range is '0' to '99' and MySQL will convert YEAR values in the ranges 2000 to 2069 and 1970 to 1999.

These lists are not complete for the basic groups, and there are some advanced types (JSON, Spatial).
You learn all of this by practicing, trial and error, etc.

String Operations

- SQL includes a string-matching operator for comparisons on character strings. The operator **like** uses patterns that are described using two special characters:
  - percent (%). The % character matches any substring.
  - underscore (_). The _ character matches any character.
- Find the names of all instructors whose name includes the substring “dar”.

```
select name
from instructor
where name like '%dar%'
```

- Match the string “100%”

```
like '100 \%'  escape  '\'
```

in that above we use backslash (\) as the escape character.

String Operations (Cont.)

- Patterns are case sensitive.
- Pattern matching examples:
  - 'Intro%' matches any string beginning with “Intro”.
  - '%Comp%' matches any string containing “Comp” as a substring.
  - '_ _ _' matches any string of exactly three characters.
  - '_ _ _ %' matches any string of at least three characters.
- SQL supports a variety of string operations such as
  - concatenation (using “||”)
  - converting from upper to lower case (and vice versa)
  - finding string length, extracting substrings, etc.

Operations and Functions

MySQL Type Conversion

BINARY 'string'
CAST (expression AS datatype)
CONVERT (expression, datatype)

MySQL Grouping Functions

AVG	MAX
BIT_AND	STD
BIT_OR	STDDEV
COUNT	SUM
GROUP_CONCAT	VARIANCE
MIN	

MySQL Mathematical Functions

ABS	COS
SIGN	SIN
MOD	TAN
FLOOR	ACOS
CEILING	ASIN
ROUND	ATAN, ATAN2
DIV	COT
EXP	RAND
LN	LEAST
LOG, LOG2, LOG10	GREATEST
POW	DEGREES
POWER	RADIANS
SQRT	TRUNCATE
PI	

MySQL String Functions

ASCII	SUBSTRING
ORD	MID
CONV	SUBSTRING_INDEX
BIN	LTRIM
OCT	RTRIM
HEX	TRIM
CHAR	SOUNDEX
CONCAT	SPACE
CONCAT_WS	REPLACE
LENGTH	REPEAT
CHAR_LENGTH	REVERSE
BIT_LENGTH	INSERT
LOCATE	ELT
INSTR	FIELD
LPAD	LCASE
RPAD	UCASE
LEFT	LOAD_FILE
RIGHT	QUOTE

MySQL Date and Time Functions

DAYOFWEEK	DATE_SUB
WEEKDAY	ADDDATE
DAYOFMONTH	SUBDATE
DAYOFYEAR	EXTRACT
MONTH	TO_DAYS
DAYNAME	FROM_DAYS
MONTHNAME	DATE_FORMAT
QUARTER	TIME_FORMAT
WEEK	CURRENT_DATE
YEAR	CURRENT_TIME
YEARWEEK	NOW
HOUR	SYSDATE
MINUTE	UNIX_TIMESTAMP
SECOND	FROM_UNIXTIME
PERIOD_ADD	SEC_TO_TIME
PERIOD_DIFF	TIME_TO_SEC
DATE_ADD	

MySQL Control Flow Functions

IF	NULLIF
IFNULL	

Almost all programming languages have similar function libraries.

Operations and Functions

- How do you learn all of these functions?
- Well, “How do I get to Carnegie Hall?”
- “Practice.”

- Switch to notebook.

Summary

- SQL Data Types:
  - There is a common, standard core set of data types that all SQL implementations support.
  - Most SQL implementations have additional types and extensions.
- Functions on Data Types:
  - There is a common, standard core set of functions for each data type that all SQL implementations support.
  - Most SQL implementations have additional functions and extensions.
- I do not expect you to memorize the types, functions, etc.
  You can look them up, play with them, ... On homework assignments and exams.
- We saw some interesting additional features that we will cover later in this lecture:
  - GROUP BY
  - ORDER BY

Order By

Ordering the Display of Tuples

- List in alphabetic order the names of all instructors

```
select distinct name
from   instructor
    order by name
```

- We may specify **desc** for descending order or **asc** for ascending order, for each attribute; ascending order is the default.
  - Example: **order by** *name* **desc**
- Can sort on multiple attributes
  - Example: **order by** *dept_name, name*

Show notebook for order by example.

INSERT, UPDATE, DELETE

Updates to tables

- **Insert**
  - **insert into** *instructor* **values** ('10211', 'Smith', 'Biology', 66000);
- **Delete**
  - Remove all tuples from the *student* relation
    - **delete from** *student*
- **Drop Table**
  - **drop table** *r*
- **Alter**
  - **alter table** *r* **add** *A D*
    - where *A* is the name of the attribute to be added to relation *r* and *D* is the domain of *A.*
    - All exiting tuples in the relation are assigned *null* as the value for the new attribute.
  - **alter table** *r* **drop** *A*
    - where *A* is the name of an attribute of relation *r*
    - Dropping of attributes not supported by many databases.

Modification of the Database

- Deletion of tuples from a given relation.
- Insertion of new tuples into a given relation
- Updating of values in some tuples in a given relation

Deletion

- Delete all instructors

  delete from *instructor*

- Delete all instructors from the Finance department

  delete from *instructor*
  where *dept_name*= 'Finance';

- *Delete all tuples in the instructor relation for those instructors associated with a department located in the Watson building.*

  delete from *instructor*
  where *dept name* **in** (**select** *dept name*
  from *department*
  where *building* = 'Watson');

Deletion (Cont.)

- Delete all instructors whose salary is less than the average salary of instructors

```
delete from instructor
where salary < (select avg (salary)
                from instructor);
```

  - Problem: as we delete tuples from *instructor*, the average salary changes
  - Solution used in SQL:
    1. First, compute **avg** (salary) and find all tuples to delete
    2. Next, delete all tuples found above (without recomputing **avg** or retesting the tuples)

Insertion

- Add a new tuple to *course*

  insert into *course*
  values ('CS-437', 'Database Systems', 'Comp. Sci.', 4);

- or equivalently

  insert into *course* (*course_id*, *title*, *dept_name*, *credits*)
  values ('CS-437', 'Database Systems', 'Comp. Sci.', 4);

- Add a new tuple to *student* with *tot_creds* set to null

  insert into *student*
  values ('3003', 'Green', 'Finance', *null*);

Insertion (Cont.)

- Make each student in the Music department who has earned more than 144 credit hours an instructor in the Music department with a salary of $18,000.

```
insert into instructor
    select ID, name, dept_name, 18000
    from   student
    where  dept_name = 'Music' and total_cred > 144;
```

- The **select from where** statement is evaluated fully before any of its results are inserted into the relation.

Otherwise queries like

```
insert into table1 select * from table1
```

would cause problem

Updates

- Give a 5% salary raise to all instructors

```
update instructor
  set salary = salary * 1.05
```

- Give a 5% salary raise to those instructors who earn less than 70000

```
update instructor
  set salary = salary * 1.05
 where salary < 70000;
```

- Give a 5% salary raise to instructors whose salary is less than average

```
update instructor
set salary = salary * 1.05
where salary < (select avg (salary)
                from instructor);
```

Updates (Cont.)

- Increase salaries of instructors whose salary is over $100,000 by 3%, and all others by a 5%
  - Write two **update** statements:

```
update instructor
   set salary = salary * 1.03
   where salary > 100000;
update instructor
    set salary = salary * 1.05
    where salary <= 100000;
```

  - The order is important
  - Can be done better using the **case** statement (next slide)

Case Statement for Conditional Updates

- Same query as before but with case statement

```
update instructor
set salary = case
                when salary <= 100000 then salary * 1.05
                else salary * 1.03
             end
```

Updates with Scalar Subqueries

- Recompute and update tot_creds value for all students

```
update student S
set tot_cred = (select sum(credits)
                from takes, course
                where takes.course_id = course.course_id  and
                      S.ID= takes.ID.and
                            takes.grade <> 'F' and
                       takes.grade is not null);
```

- Sets *tot_creds* to null for students who have not taken any course
- Instead of **sum**(*credits*), use:

```
case
    when sum(credits) is not null then sum(credits)
    else 0
 end
```

Summary

- INSERT, UPDATE and DELETE are pretty straightforward.
- UPDATE and DELETE are very similar to SELECT
  - WHERE clause specifies which rows are affected.
  - The SELECT choose the columns to return.
  - The SET clause chooses and changes columns.
  - DELETE just removes the specified rows.
- INSERT, UPDATE and DELETE changes must not violate constraints, e.g.
  - INSERT a row that causes a duplicate key.
  - DELETE a referenced (target) foreign key.
  - UPDATE columns that create a duplicate key.
  - INSERT values do not include all NOT NULL columns.
- I am not going to do example now, but you have seen and will see me do examples in the context of larger examples.

Aggregate Functions

Aggregate Functions

- These functions operate on the multiset of values of a column of a relation, and return a value

  avg: average value
  min: minimum value
  max: maximum value
  sum: sum of values
  count: number of values

Note: Some database implementations have additional aggregate functions.

Aggregate Functions – Group By

- Find the average salary of instructors in each department
  - **select** *dept_name*, **avg** (*salary*) **as** *avg_salary*
    from *instructor*
    group by *dept_name*;

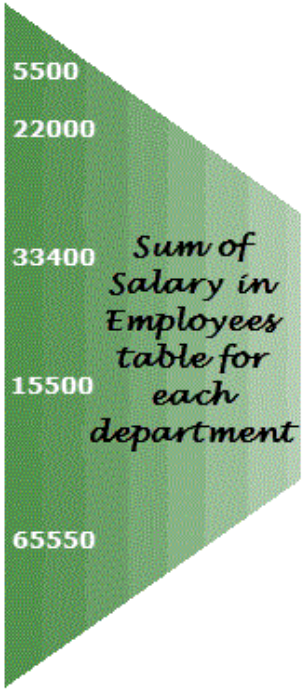
ID	*name*	*dept_name*	*salary*
76766	Crick	Biology	72000
45565	Katz	Comp. Sci.	75000
10101	Srinivasan	Comp. Sci.	65000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000
12121	Wu	Finance	90000
76543	Singh	Finance	80000
32343	El Said	History	60000
58583	Califieri	History	62000
15151	Mozart	Music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

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

Another View

Employees

DEPARTMENT_ID	SALARY
10	5500
20	15000
20	7000
30	12000
30	5100
30	4900
30	5800
30	5600
40	7500
40	8000
50	9000
50	8500
50	9500
50	8500
50	10500
50	10000
50	9500



DEPARTMENT_ID	SUM(SALARY)
10	5500
20	22000
30	33400
40	15500
50	65550

- GROUP BY column list
  - Forms partitions containing multiple rows.
  - All rows in a partition have the same values for the GROUP BY columns.
- The aggregate functions
  - Merge the non-group by attributes, which may differ from row to row.
  - Into a single value for each attribute.
- The result is one row per distinct set of GROUP BY values.
- There may be multiple non-GROUP BY COLUMNS, each with its own aggregate function.
- You can use HAVING in place of WHERE on the GROUP BY result.

Aggregate Functions Examples

- Find the average salary of instructors in the Computer Science department
  - **select avg** (*salary*)
    from *instructor*
    where *dept_name*= 'Comp. Sci.';
- Find the total number of instructors who teach a course in the Spring 2018 semester
  - **select count** (**distinct** *ID*)
    from *teaches*
    where *semester* = 'Spring' **and** *year* = 2018;
- Find the number of tuples in the *course* relation
  - **select count** (*)
    from *course*;

Aggregation (Cont.)

- Attributes in **select** clause outside of aggregate functions must appear in **group by** list
  - /* erroneous query */
    select *dept_name*, *ID*, **avg** (*salary*)
    from *instructor*
    group by *dept_name*;

Aggregate Functions – Having Clause

- Find the names and average salaries of all departments whose average salary is greater than 42000

  select *dept_name*, **avg** (*salary*) **as** *avg_salary*
  from *instructor*
  group by *dept_name*
  having avg (*salary*) > 42000;

- Note: predicates in the **having** clause are applied after the formation of groups whereas predicates in the **where** clause are applied before forming groups

Example

- Switch to notebook.
- Career batting statistics.

JOIN

Joined Relations

- **Join operations** take two relations and return as a result another relation.
- A join operation is a Cartesian product which requires that tuples in the two relations match (under some condition). It also specifies the attributes that are present in the result of the join
- The join operations are typically used as subquery expressions in the **from** clause
- Three types of joins:
  - Natural join
  - Inner join
  - Outer join

Notes:

- You will also hear terms like equi-join, non-equi-join, theta join, semi-join, … …
- I ask for definitions on exams, but you can just look them up.

Natural Join in SQL

- Natural join matches tuples with the same values for all common attributes, and retains only one copy of each common column.
- List the names of instructors along with the course ID of the courses that they taught
  - **select** *name*, *course_id*
    from *students, takes*
    where *student.ID = takes.ID*;
- Same query in SQL with “natural join” construct
  - **select** *name*, *course_id*
    from *student* **natural join** *takes*;

Natural Join in SQL (Cont.)

- The **from** clause can have multiple relations combined using natural join:

  select $A_1, A_2, \ldots A_n$
  from r_1 **natural join** r_2 **natural join .. natural join** r_n
  where P ;

Student Relation

ID	*name*	*dept_name*	*tot_cred*
00128	Zhang	Comp. Sci.	102
12345	Shankar	Comp. Sci.	32
19991	Brandt	History	80
23121	Chavez	Finance	110
44553	Peltier	Physics	56
45678	Levy	Physics	46
54321	Williams	Comp. Sci.	54
55739	Sanchez	Music	38
70557	Snow	Physics	0
76543	Brown	Comp. Sci.	58
76653	Aoi	Elec. Eng.	60
98765	Bourikas	Elec. Eng.	98
98988	Tanaka	Biology	120

Takes Relation

ID	*course_id*	*sec_id*	*semester*	*year*	*grade*
00128	CS-101	1	Fall	2017	A
00128	CS-347	1	Fall	2017	A-
12345	CS-101	1	Fall	2017	C
12345	CS-190	2	Spring	2017	A
12345	CS-315	1	Spring	2018	A
12345	CS-347	1	Fall	2017	A
19991	HIS-351	1	Spring	2018	B
23121	FIN-201	1	Spring	2018	C+
44553	PHY-101	1	Fall	2017	B-
45678	CS-101	1	Fall	2017	F
45678	CS-101	1	Spring	2018	B+
45678	CS-319	1	Spring	2018	B
54321	CS-101	1	Fall	2017	A-
54321	CS-190	2	Spring	2017	B+
55739	MU-199	1	Spring	2018	A-
76543	CS-101	1	Fall	2017	A
76543	CS-319	2	Spring	2018	A
76653	EE-181	1	Spring	2017	C
98765	CS-101	1	Fall	2017	C-
98765	CS-315	1	Spring	2018	B
98988	BIO-101	1	Summer	2017	A
98988	BIO-301	1	Summer	2018	*null*

student **natural join** *takes*

ID	*name*	*dept_name*	*tot_cred*	*course_id*	*sec_id*	*semester*	*year*	*grade*
00128	Zhang	Comp. Sci.	102	CS-101	1	Fall	2017	A
00128	Zhang	Comp. Sci.	102	CS-347	1	Fall	2017	A-
12345	Shankar	Comp. Sci.	32	CS-101	1	Fall	2017	C
12345	Shankar	Comp. Sci.	32	CS-190	2	Spring	2017	A
12345	Shankar	Comp. Sci.	32	CS-315	1	Spring	2018	A
12345	Shankar	Comp. Sci.	32	CS-347	1	Fall	2017	A
19991	Brandt	History	80	HIS-351	1	Spring	2018	B
23121	Chavez	Finance	110	FIN-201	1	Spring	2018	C+
44553	Peltier	Physics	56	PHY-101	1	Fall	2017	B-
45678	Levy	Physics	46	CS-101	1	Fall	2017	F
45678	Levy	Physics	46	CS-101	1	Spring	2018	B+
45678	Levy	Physics	46	CS-319	1	Spring	2018	B
54321	Williams	Comp. Sci.	54	CS-101	1	Fall	2017	A-
54321	Williams	Comp. Sci.	54	CS-190	2	Spring	2017	B+
55739	Sanchez	Music	38	MU-199	1	Spring	2018	A-
76543	Brown	Comp. Sci.	58	CS-101	1	Fall	2017	A
76543	Brown	Comp. Sci.	58	CS-319	2	Spring	2018	A
76653	Aoi	Elec. Eng.	60	EE-181	1	Spring	2017	C
98765	Bourikas	Elec. Eng.	98	CS-101	1	Fall	2017	C-
98765	Bourikas	Elec. Eng.	98	CS-315	1	Spring	2018	B
98988	Tanaka	Biology	120	BIO-101	1	Summer	2017	A
98988	Tanaka	Biology	120	BIO-301	1	Summer	2018	*null*

Dangerous in Natural Join

- Beware of unrelated attributes with same name which get equated incorrectly
- Example -- List the names of students instructors along with the titles of courses that they have taken
  - Correct version

    select *name*, *title*
    from *student* **natural join** *takes*, *course*
    where *takes.course_id* = *course.course_id*;

  - Incorrect version

    select *name*, *title*
    from *student* **natural join** *takes* **natural join** *course*;

    - This query omits all (student name, course title) pairs where the student takes a course in a department other than the student's own department.
    - The correct version (above), correctly outputs such pairs.

Natural Join with Using Clause

- To avoid the danger of equating attributes erroneously, we can use the "**using**" construct that allows us to specify exactly which columns should be equated.
- Query example

  select *name*, *title*
  from (*student* **natural join** *takes*) **join** *course* **using** (*course_id*)

Join Condition

- The **on** condition allows a general predicate over the relations being joined
- This predicate is written like a **where** clause predicate except for the use of the keyword **on**
- Query example

  select *
  from *student* **join** *takes* **on** *student_ID* = *takes_ID*

  - The **on** condition above specifies that a tuple from *student* matches a tuple from *takes* if their *ID* values are equal.
- Equivalent to:

  select *
  from *student , takes*
  where *student_ID* = *takes_ID*

Join Condition (Cont.)

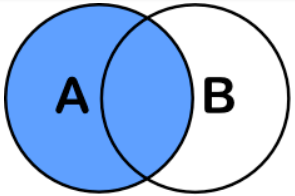
- The **on** condition allows a general predicate over the relations being joined.
- This predicate is written like a **where** clause predicate except for the use of the keyword **on**.
- Query example

  select *
  from *student* **join** *takes* **on** *student_ID = takes_ID*

  - The **on** condition above specifies that a tuple from *student* matches a tuple from *takes* if their *ID* values are equal.
- Equivalent to:

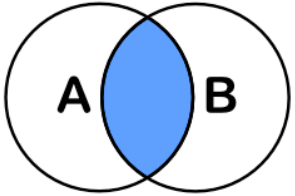
  select *
  from *student , takes*
  where *student_ID = takes_ID*

One Way to Think About Joins

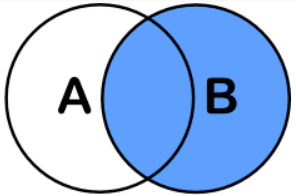


SELECT <auswahl>
FROM tabelleA A
LEFT JOIN tabelleB B
ON A.key = B.key

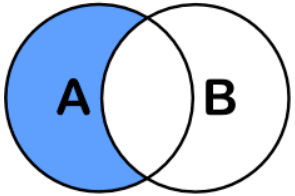
CHEATSHEET
SQL
JOINS



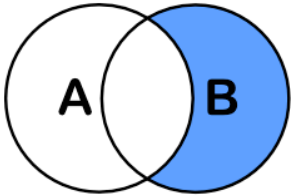
SELECT <auswahl>
FROM tabelleA A
INNER JOIN tabelleB B
ON A.key = B.key



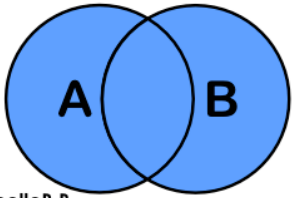
SELECT <auswahl>
FROM tabelleA A
RIGHT JOIN tabelleB B
ON A.key = B.key



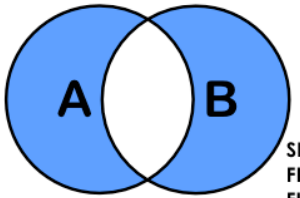
SELECT <auswahl>
FROM tabelleA A
LEFT JOIN tabelleB B
ON A.key = B.key
WHERE B.key IS NULL



SELECT <auswahl>
FROM tabelleA A
RIGHT JOIN tabelleB B
ON A.key = B.key
WHERE A.key IS NULL



SELECT <auswahl>
FROM tabelleA A
FULL OUTER JOIN tabelleB B
ON A.key = B.key

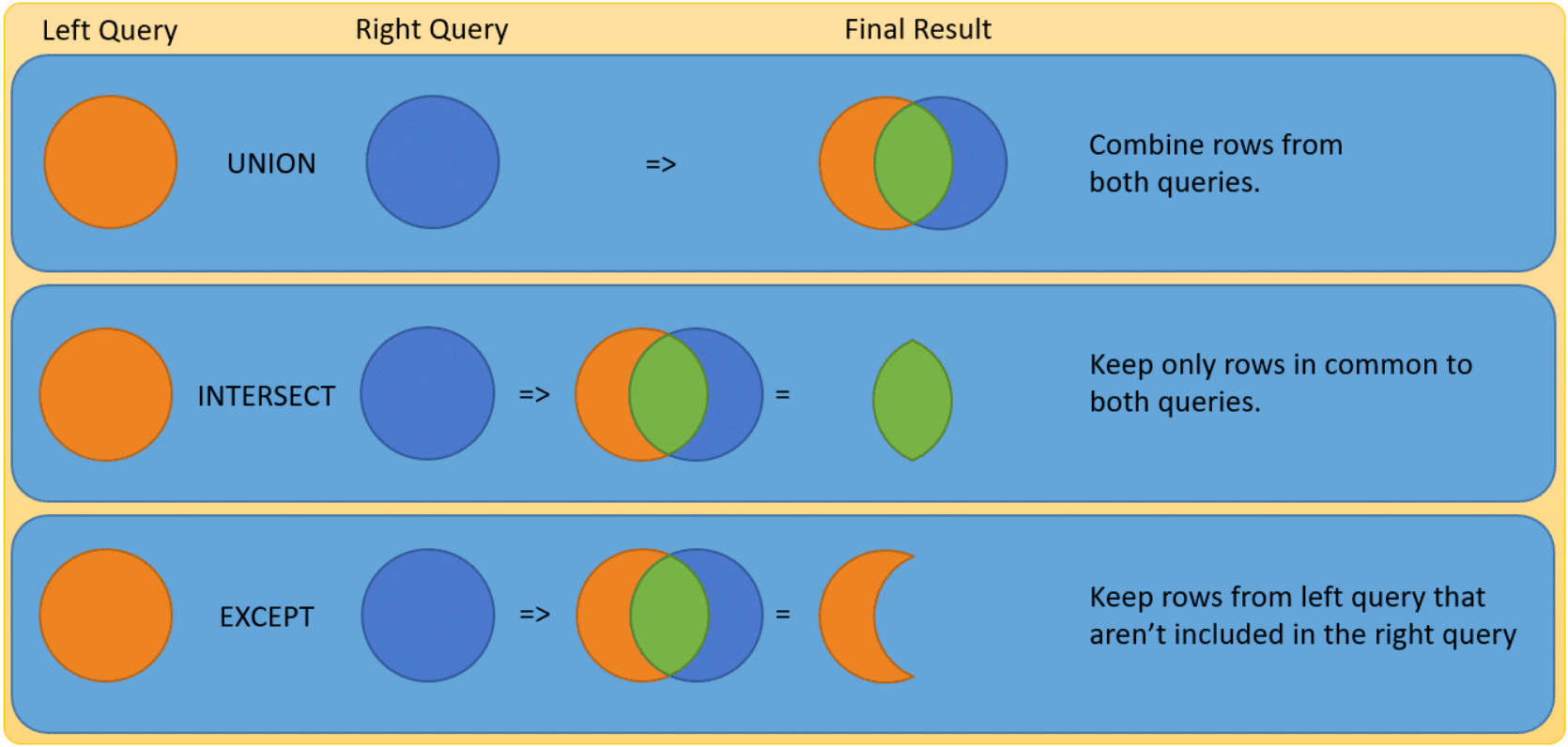


SELECT <auswahl>
FROM tabelleA A
FULL OUTER JOIN tabelleB B
ON A.key = B.key
WHERE A.key IS NULL
OR B.key IS NULL

Set Operations

Set Operations

Visual Explanation of UNION, INTERSECT, and EXCEPT operators



Set Operations (Cont.)

- Set operations **union**, **intersect**, and **except**
  - Each of the above operations automatically eliminates duplicates
- To retain all duplicates use the
  - **union all**,
  - **intersect all**
  - **except all**.

NOTE:

- SELECT implementing a project can have duplicate rows. If you do not want duplicates, you must use the DISTINCT key word.
- UNION behaves the other way. It removes duplicates. If you want to keep the duplicates, you have to select UNION ALL.
- Some SQL engines do not implement INTERSECT and/or EXCEPT, you can implement the function with subqueries, which we will cover soon.

Set Operations

- Find courses that ran in Fall 2017 or in Spring 2018

  (**select** *course_id* **from** *section* **where** *sem* = 'Fall' **and** *year* = 2017)
  union
  (**select** *course_id* **from** *section* **where** *sem* = 'Spring' **and** *year* = 2018)

- Find courses that ran in Fall 2017 and in Spring 2018

  (**select** *course_id* **from** *section* **where** *sem* = 'Fall' **and** *year* = 2017)
  intersect
  (**select** *course_id* **from** *section* **where** *sem* = 'Spring' **and** *year* = 2018)

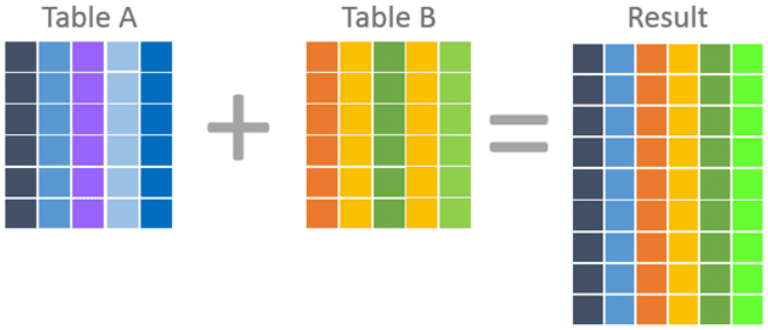
- Find courses that ran in Fall 2017 but not in Spring 2018

  (**select** *course_id* **from** *section* **where** *sem* = 'Fall' **and** *year* = 2017)
  except
  (**select** *course_id* **from** *section* **where** *sem* = 'Spring' **and** *year* = 2018)

Let’s Practice – Go to the Notebook

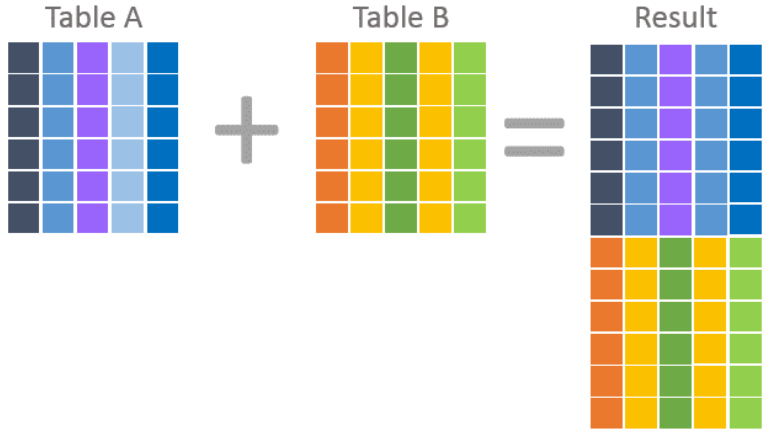
JOIN and UNION – A Final Word

Here is a visual depiction of a join. Table A and B's columns are combined into a single result.



Joins Combine Columns

Now compare the above depiction with that of a union. In a union, each row within the result is from one table OR the other. In a union, columns aren't combined to create results, rows are combined.



Unions Combine Rows

- UNION vs JOIN can be confusing. Basically,
  - JOIN puts the tables together "side by side."
  - Union puts the tables together "one on top of the oter."

Let's Practice

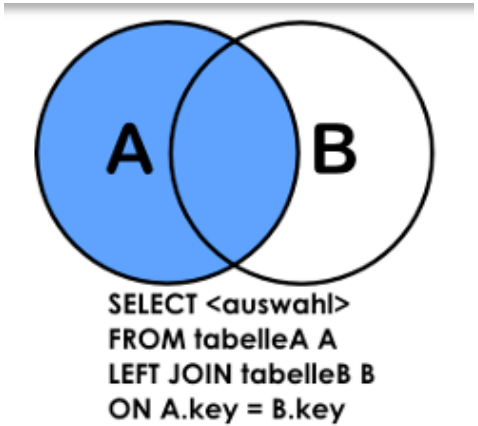
You Guessed It.

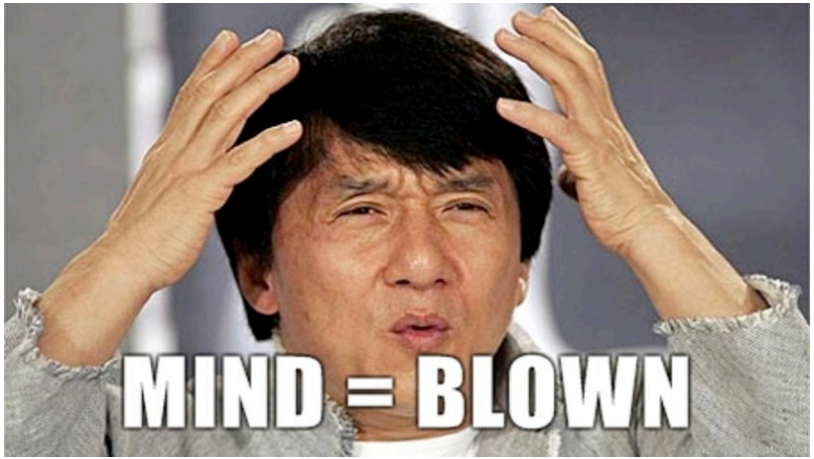
Over to the Notebook.

It is Baseball and Game of Thrones Time!

Backup

Left JOIN



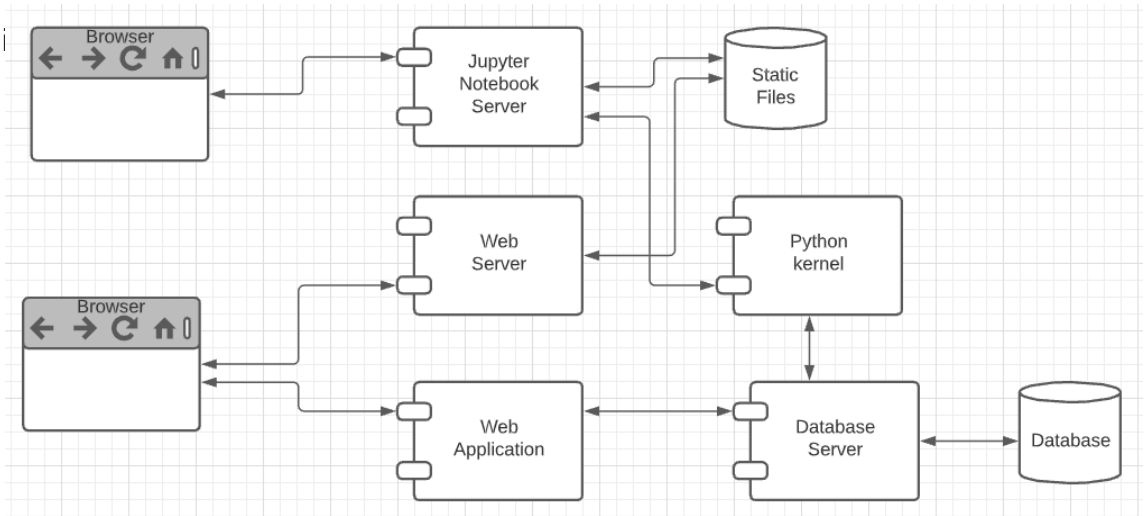


- The easiest way to think about this is ...
  - Do a “normal” JOIN that produces all of the pairs of matching rows.
  - There MAY be some some rows in A that did not match anything.
  - For each of those rows,
    - Put one row in the JOIN result
    - That has the column values from A and NULL values for the columns from B.
- Right JOIN is exactly the same, except swap the logic for A and B.
- A FULL OUTER JOIN is the UNION of a left join and a right join. Some SQL engines do not support it and you must explicitly do the left, right and union.

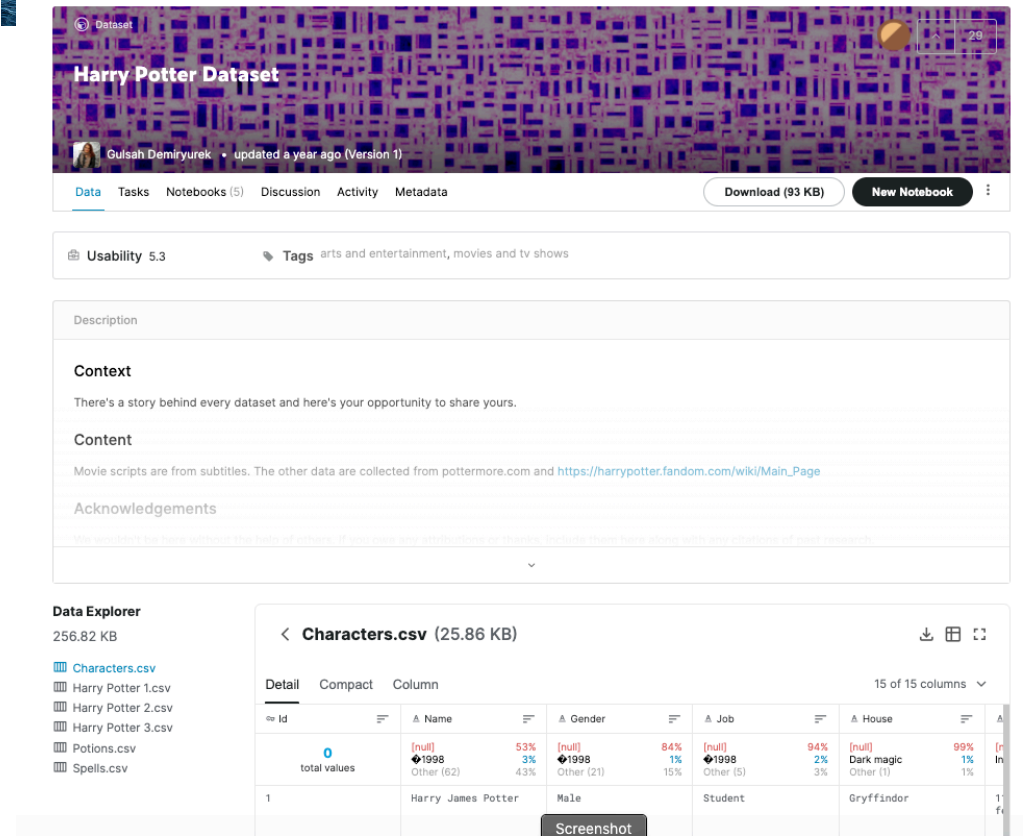
Today's Task #1

Harry Potter Character Information Application

- Dataset:
  - https://www.kaggle.com/gulsahdemiryurek/harry-potter-dataset
  - Six “comma separated value (CSV)” files. We will start with *Characters.csv.*
- Tasks: (Both tasks start with *importing and engineering the data.)*
  - Build an interactive web application for *create-retrieve-update-delete* entries. The application must enforce *semantic integrity.*
  - Build a Jupyter Notebook that does some i
- Elements:
  - Content/applications in a browser.
  - Jupyter Notebook Server is both:
    - A web server
    - A (specialized) web application server.
  - Web server delivers HTML, images, ... to the browser.
  - Web application server executes an
    - Application server framework (Flask)
    - The Python code implementing the app.
  - The Jupyter Notebook Server runs code cells in a Python kernel.
  - I am mostly doing to ignore the browser and content.



The Data – Characters.csv



15 columns:

- Id
- Name
- Gender
- Job
- House
- Wand
- Patronus
- Species
- Blood status
- Hair colour
- Eye colour
- Loyalty
- Skills
- Birth
- Death

- Downloaded the file.
- The columns are separated by ';'
- Despite being "Comma" Separated Values, other separators are common.

- Loaded using DataGrip into MySQL
  - Connect
  - Create Schema
  - Use Schema
  - Use the data load tool.

But First, Let's Look at the Data

There are some weird characters, no pun intended.

- 9º" Chestnut dragon heartstring
- Dumbledore's Army |Hogwarts School of Witchcraft and Wizardry
- 9Ω" Fir dragon heartstring
- Professor†of†Transfiguration†| Head of Gryffindor

- How did this happen?
- What do we do?

	Female	Professor†of†Transfiguration†| Head of Gryffindor	Gryffindor	9Ω" Fir dragon heartstring
	Female		Gryffindor	Unknown
	Male	Head of the†Misuse of Muggle Artefacts Office	Gryffindor	Unknown
	Male	Defence Against the Dark Arts(1991-1992)	Ravenclaw	9" Alder unicorn hair bendy
	Female	Student	Ravenclaw	Unknown
	Female	Student	Ravenclaw	Unknown
	Male	Defence Against the Dark Arts(1992-1993)	Ravenclaw	9" Cherry dragon heartstring
	Male	Professor of Charms	Head of Ravenclaw	Ravenclaw	Unknown
	Female	Professor†of†Divination	Ravenclaw	9 Ωî hazel unicorn hair core
	Male	Wandmaker	Ravenclaw	12æ" Hornbeam dragon heartstring
tle)	Female	Student	Ravenclaw	Unknown
	Female	Student	Ravenclaw	Unknown
	Male	Student	Ravenclaw	Unknown
	Female	Student	Ravenclaw	Unknown
	Male	Student	Ravenclaw	Unknown
	Male	Student	Ravenclaw	Unknown
	Male	Professor of Potions	Head of Slytherin	Slytherin	Unknown
	Male	Student	Slytherin	10" Hawthorn unicorn hair
	Male	Student	Slytherin	Unknown
	Male	Student	Slytherin	Unknown
	Female		Slytherin	12æ" Walnut dragon heartstring
	Female	Professor†of†Defence Against the Dark Arts†| Department of Magical Law Enforcement	Slytherin	8" Birch dragon heartstring
		Professor of Potions	Slytherin	10º" Cedar dragon heartstring fairly flexible
	Male	School Governor	Slytherin	Elm and dragon heartstring
	Female		Slytherin	Unknown
	Male		Slytherin	Unknown
	Female	Student	Slytherin	Unknown

Analysis (I)

- Some of the issues are character encoding/set issues:
  - “In computing, data storage, and data transmission, character encoding is used to represent a repertoire of characters by some kind of encoding system that assigns a number to each character for digital representation.” (https://en.wikipedia.org/wiki/Character_encoding)
  - “A character set is a collection of characters that might be used by multiple languages.” (https://en.wikipedia.org/wiki/Character_encoding)
- This will be a common issue when dealing with string data that you import to build your database.
- This is especially true if the data comes from “data scraping”.
  - “Data scraping is a technique in which a computer program extracts data from human-readable output coming from another program.” (https://en.wikipedia.org/wiki/Data_scraping)
  - There are several types of scraping: screen scraping, web scraping, ...
- That is what happened here. The data comes from web scraping a wiki page.

Content

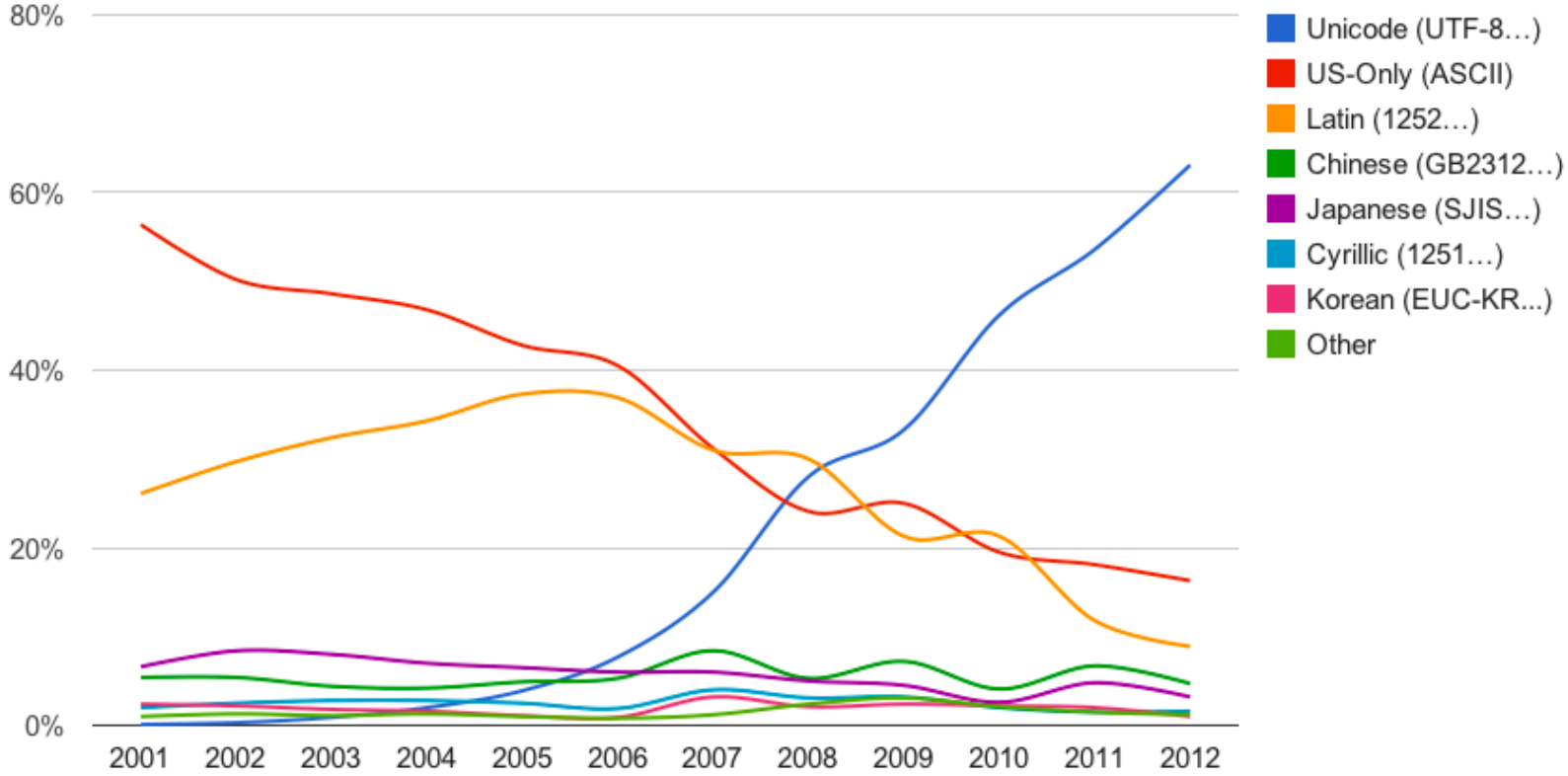
Movie scripts are from subtitles. The other data are collected from pottermore.com and https://harrypotter.fandom.com/wiki/Main_Page

Character Encodings/Character Sets

Common character encodings [edit]

- ISO 646
  - ASCII
- EBCDIC
- ISO 8859:
  - ISO 8859-1 Western Europe
  - ISO 8859-2 Western and Central Europe
  - ISO 8859-3 Western Europe and South European (Turkish, Maltese plus Esperanto)
  - ISO 8859-4 Western Europe and Baltic countries (Lithuania, Estonia, Latvia and Lapp)
  - ISO 8859-5 Cyrillic alphabet
  - ISO 8859-6 Arabic
  - ISO 8859-7 Greek
  - ISO 8859-8 Hebrew
  - ISO 8859-9 Western Europe with amended Turkish character set
  - ISO 8859-10 Western Europe with rationalised character set for Nordic languages, including complete Icelandic set
  - ISO 8859-11 Thai
  - ISO 8859-13 Baltic languages plus Polish
  - ISO 8859-14 Celtic languages (Irish Gaelic, Scottish, Welsh)
  - ISO 8859-15 Added the Euro sign and other rationalisations to ISO 8859-1
  - ISO 8859-16 Central, Eastern and Southern European languages (Albanian, Bosnian, Croatian, Hungarian, Polish, Romanian, Serbian and Slovenian, but also French, German, Italian and Irish Gaelic)
- CP437, CP720, CP737, CP850, CP852, CP855, CP857, CP858, CP860, CP861, CP862, CP863, CP865, CP866, CP869, CP872
- MS-Windows character sets:
  - Windows-1250 for Central European languages that use Latin script, (Polish, Czech, Slovak, Hungarian, Slovene, Serbian, Croatian, Bosnian, Romanian and Albanian)
  - Windows-1251 for Cyrillic alphabets
  - Windows-1252 for Western languages
  - Windows-1253 for Greek
  - Windows-1254 for Turkish
  - Windows-1255 for Hebrew
  - Windows-1256 for Arabic
  - Windows-1257 for Baltic languages
  - Windows-1258 for Vietnamese
- Mac OS Roman
- KOI8-R, KOI8-U, KOI7
- MIK
- ISCII
- TSCII
- VISCII
- JIS X 0208 is a widely deployed standard for Japanese character encoding that has several encoding forms.
  - Shift JIS (Microsoft Code page 932 is a dialect of Shift_JIS)
  - EUC-JP
  - ISO-2022-JP
- JIS X 0213 is an extended version of JIS X 0208.
  - Shift_JIS-2004
  - EUC-JIS-2004
  - ISO-2022-JP-2004
- Chinese Guobiao
  - GB 2312
  - GBK (Microsoft Code page 936)
  - GB 18030
- Taiwan Big5 (a more famous variant is Microsoft Code page 950)
  - Hong Kong HKSCS
- Korean
  - KS X 1001 is a Korean double-byte character encoding standard
  - EUC-KR
  - ISO-2022-KR
- Unicode (and subsets thereof, such as the 16-bit 'Basic Multilingual Plane')
  - UTF-8
  - UTF-16
  - UTF-32
- ANSEL or ISO/IEC 6937

Character Set Usage over Time



https://www.w3.org/International/questions/qa-who-uses-unicode

But, ...

- Consider **Job** column entry: “Professor†of†Transfiguration†l Head of Gryffindor”
- There are two odd characters: “†” and “l”
- How should I correct/interpret the strings? My interpretation is:
  - Replace “†” with “ “.
  - The “l” means that the person has two jobs:
    - Professor of Transfiguration
    - Head of Gryffindor
  - The **Job** column is a **multivalued** attribute.
  - The character “l” indicates that a text string is one string containing multiple values for an attribute.
- Can I always replace non-l, odd characters with “ “?
  - ‘9Ω" Fir dragon heartstring’ and ‘9º" Chestnut dragon heartstring’ would wind up having an extra space. I want 9”, not 9 “.
  - The cleanup is getting complicated.

The Cleanup is Rule-Based

- In this dataset, I can use a set of rules for data clean up:
  - Two characters separated by † maps to two characters separated by “ “ (space).
  - A number and “ separated by an odd character maps the the number and “
  - Strings of the form s1 | s2 | s3 maps to [s1, s2, s3]
  - etc.
- Well, I could write a custom function for each rule,
  - But what happens when I find I need other rules.
  - It would be nice if there were some kind of tools that helps.
- A **regular expression** is a sequence of characters that define a search pattern. Usually, such patterns are used by string-searching algorithms for "find" or "find and replace" operations on strings, or for input validation. It is a technique developed in theoretical computer science and formal language theory.

Multivalued Attributes

- How about strings of the form s1 | s2 | s3 maps to [s1, s2, s3]. Is this cool?
- All the data entity sets we have seen have had simple types/value attributes

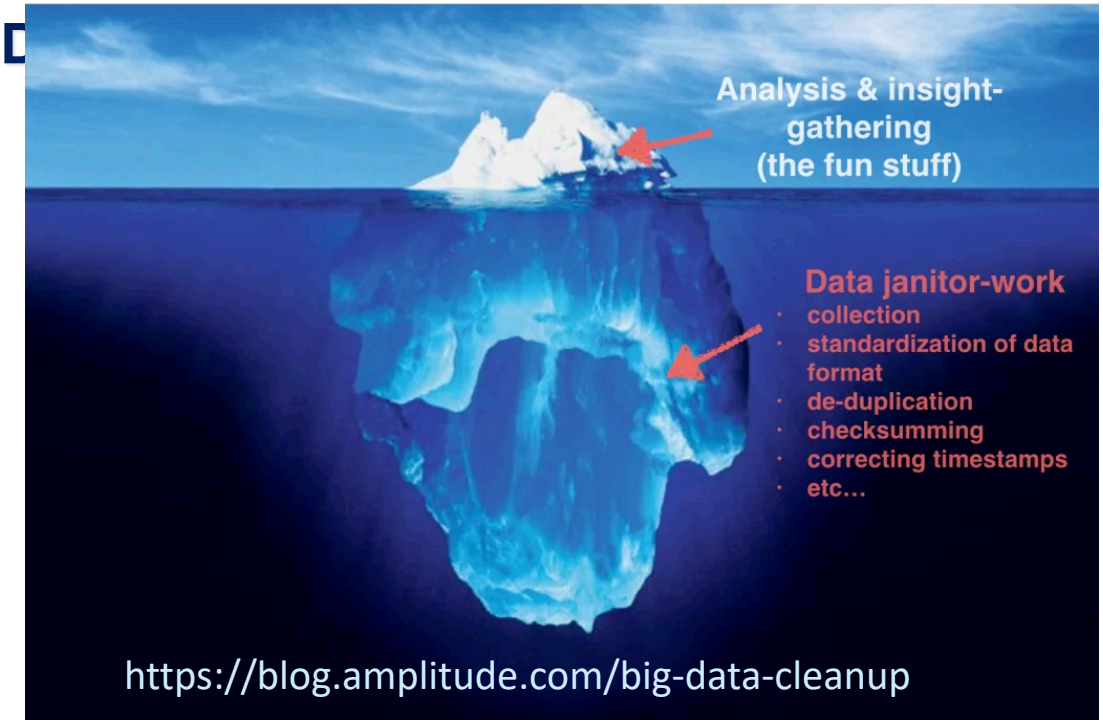
del,

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

(a) The *instructor* table

course_id	*title*	*dept_name*	*credits*
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

Figure 2.2 The *course* relation.



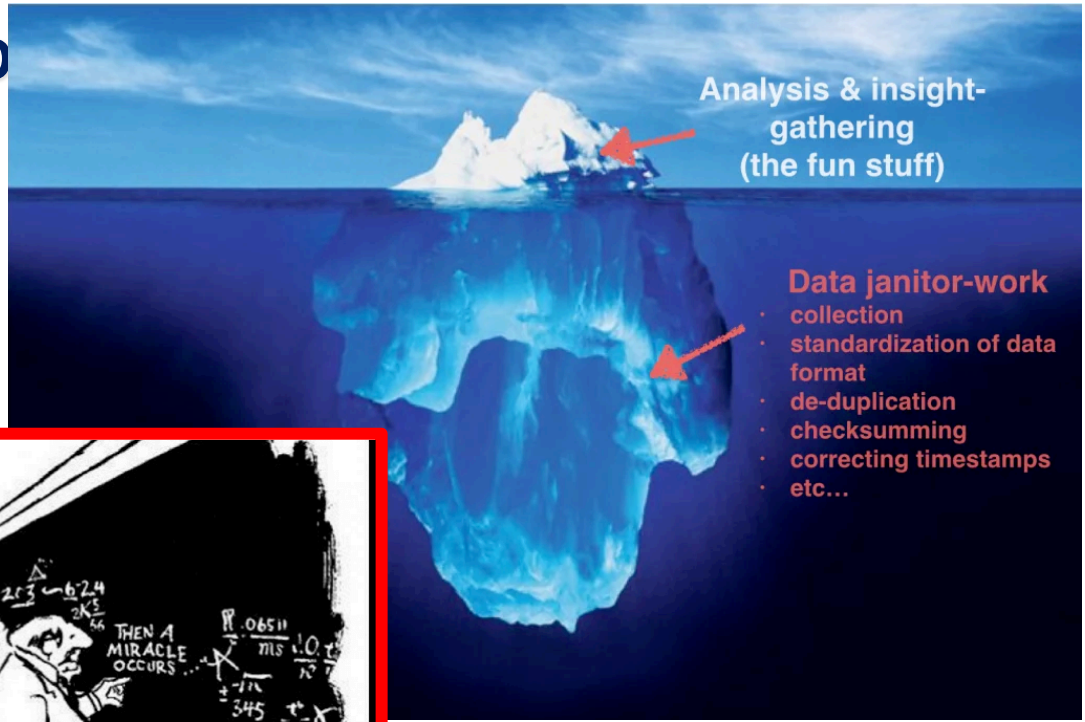
Database and data science classes **love to teach the fun stuff**
queries, data modeling, machine learning, spooky math, algorithms, ...

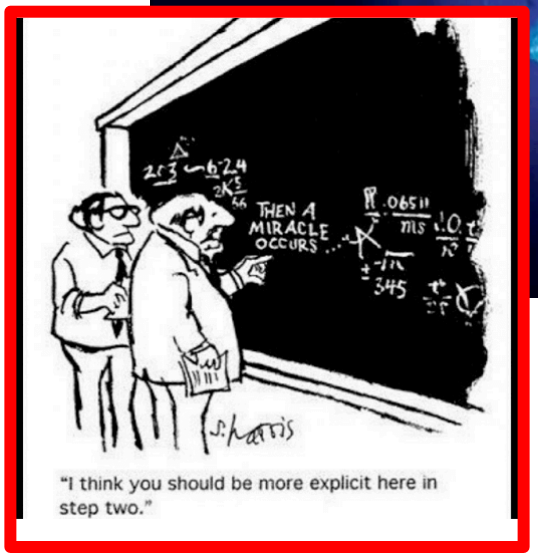
Syllabus Topics

- Relational Foundations
  - Overview (1 lecture)
  - ER Model (2 lectures)
  - Relational Model (4 lectures)
  - Relational Algebra (2 lectures)
  - SQL (5 lectures)
  - Application Programming and Database APIs (1 lecture)
  - Security (2 lectures)
  - Normalization (2 lectures)
  - Overview of Storage and Indexes (1 lecture)
  - Overview of Query Optimization (1 lecture)
  - Overview of Transaction Processing (1 lecture)
- Beyond Relational Foundations
  - NoSQL (1 lecture)
  - Data Preparation and Cleaning (1 lecture)
  - Graphs (1 lecture)
  - Object-Relational Databases (2 lectures)
  - Cloud Databases (1 lecture)

Recommended Syllabus

D

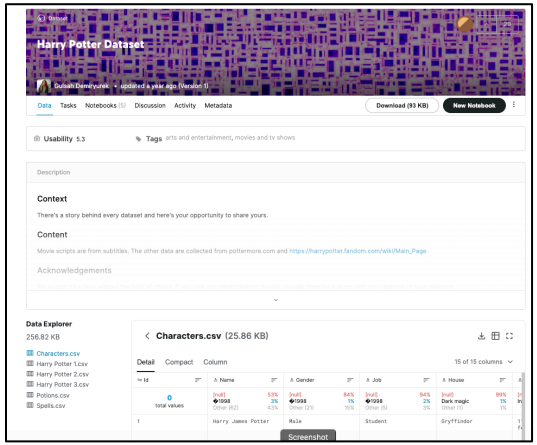




I place more emphasis on data cleansing and refactoring than other sections of W4111.

Next Steps in the Plan

Source Data



Relational Data

attributes (or columns)

tuples (or rows)

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

- We are starting the semester with the *relational datamodel/databases.*
- Before we get too far in our data janitor work,
  we should understand the relational model a little more.

But, Before We Leave ... ...

- What are these regular expressions whereout you speak?
- Many environments come with regular expressions functions for searching and transforming strings.
  - Python: Lib/re.py (https://docs.python.org/3/library/re.html)
  - Java: java.util.regex (https://www.w3school
  - "SNOBOL ("StriNg Oriented and symBOlic languages ... ... having patterns as a first-cl (https://en.wikipedia.org/wiki/SNOBOL)
  - Most SQL databases have some from of re xpression library, but the libraries differ fro roduct to product.
    - This was FYI.
    - Not a core part of W4111.
    - We will play with it a little.

Table 12.14 Regular Expression Functions and Operators

Name	Description
NOT REGEXP	Negation of REGEXP
REGEXP	Whether string matches regular expression
REGEXP_INSTR()	Starting index of substring matching regular expression
REGEXP_LIKE()	Whether string matches regular expression
REGEXP_REPLACE()	Replace substrings matching regular expression
REGEXP_SUBSTR()	Return substring matching regular expression
RLIKE	Whether string matches regular expression


```
In [40]:
1  s = '12æ" Hornbeam dragon heartstring'
2  x = re.sub(r'([0-9]).(")', r'\1\2', s)
3  x

Out[40]: '12" Hornbeam dragon heartstring'
```

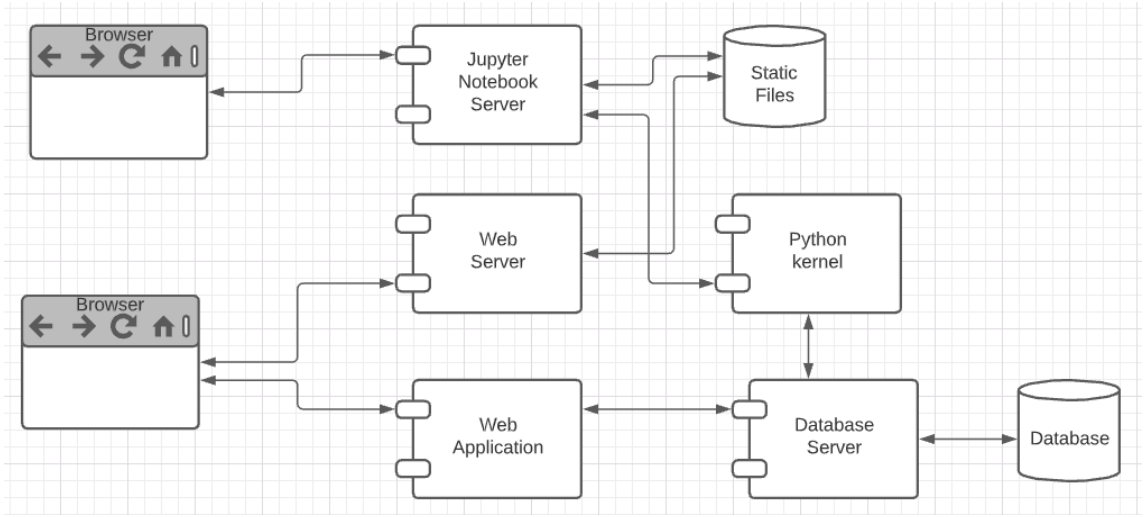
Today's Task #1

I did the set up and provided the tools, but did not do the task.

I will cover in a recitation.

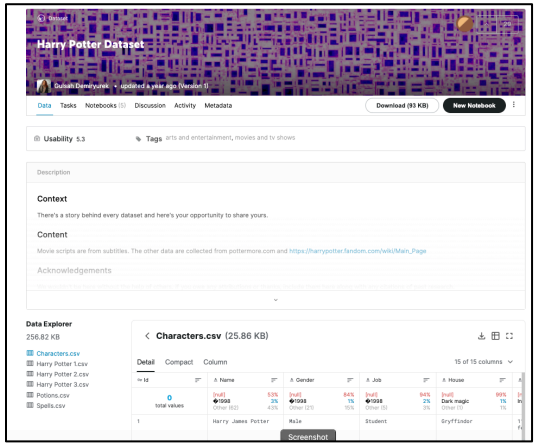
Harry Potter Character Information Application (Reminder)

- Dataset:
  - https://www.kaggle.com/gulsahdemiryurek/harry-potter-dataset
  - Six “comma separated value (CSV)” files. We will start with *Characters.csv.*
- Tasks: (Both tasks start with *importing and engineering the data.)*
  - Build an interactive web application for *create-retrieve-update-delete* entries. The application must enforce *semantic integrity.*
  - Build a Jupyter Notebook that does some interesting visualization and analysis.
- Elements:
  - Content/applications in a browser.
  - Jupyter Notebook Server is both:
    - A web server
    - A (specialized) web application server.
  - Web server delivers HTML, images, ... to the browser.
  - Web application server executes an
    - Application server framework (Flask)
    - The Python code implementing the app.
  - The Jupyter Notebook Server runs code cells in a Python kernel.
  - I am mostly going to ignore the browser and content.



Next Steps in the Plan

Source Data



Relational Data

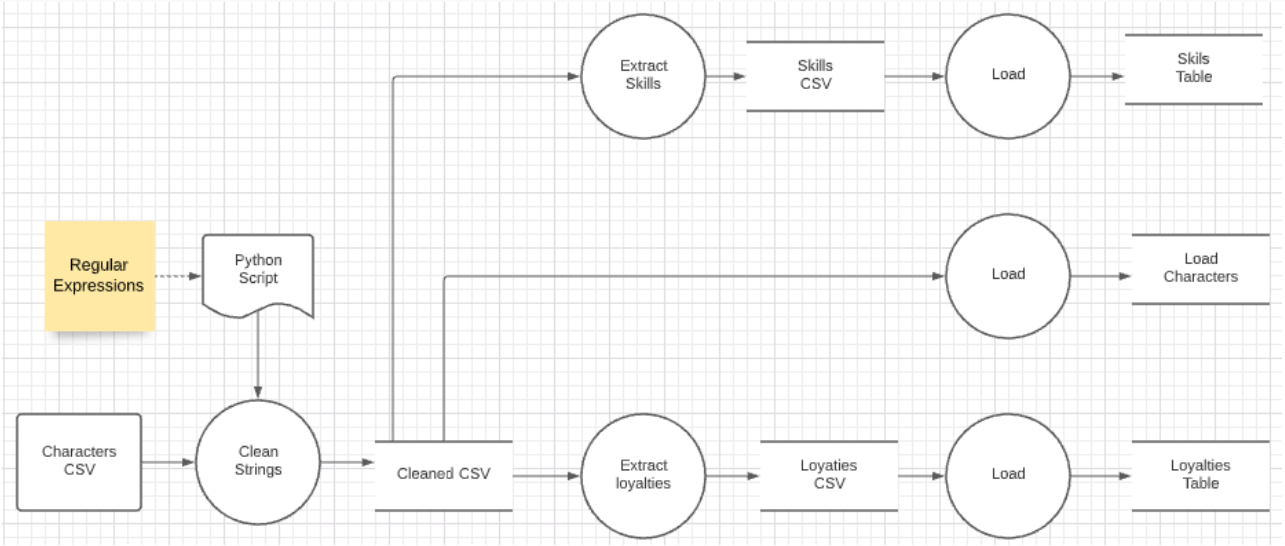
attributes (or columns)

tuples (or rows)

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

- We are starting the semester with the *relational datamodel/databases.*
- Before we get too far in our data janitor work,
  we should understand the relational model a little more.

Data Flow



- I downloaded the Harry Potter character set CSV file.
- I had two cleanup problems.
  - The file contains some weird characters in strings, e.g. “9Ω" Fir dragon heartstring
  - There is a one-to-many relationship in columns marked with delimiters, e.g. Professor†of†Transfiguration†| Head of Gryffindor
- I wrote a little Extract-Transform-Load (ETL) process.
  We will cover ETL in Module IV in the final part of the semester.

The next steps are to clean up the data in the RDB.