Databases

SQL and relational data

INFO 3401

University of Colorado
Boulder

Databases

- “A **database** is an organized collection of data, generally stored and accessed electronically from a computer system”
  - https://en.wikipedia.org/wiki/Database
- Databases are very fundamental tools for organizing and storing electronic data
- Data scientists, data analysts and software developers use databases all day, every day.

Databases

- Offer a SQL API (that will make more sense later)
- Often guarantee ACID properties (that will make more sense later)

  https://en.wikipedia.org/wiki/ACID

Databases consist of tables

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	OrangePizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

Tables have records (=rows)

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	OrangePizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

One record

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	OrangePizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

Tables have fields (=columns)

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	OrangePizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

The primary key uniquely identifies records

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	OrangePizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

OrangePizza joined on a tablet and gave a score of 4

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	OrangePizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

Why not just pandas?

- In 3401 have seen how to explore data in-memory using pandas
  - This discussion of tabular data sure seems a lot like pandas.
  - Do we even need this?
- A few things to consider
  - What happens if we want to update data?
  - What happens if multiple people want to analyze the same data?
  - What happens if we have more data than can fit in memory?
  - What happens if we have data that only some people are allowed to see or update?
- If these kinds of questions become more important, you usually end up needing a database.

Why “relational” data?

- Why “relational” databases?
  - Stay tuned

Recap

- To recap, this table shows 4 records
- It also has 4 fields: `primary_key, user_ID, user_modality, satisfaction_score`

In-class activity in groups

See Canvas link for today

We interact with databases via SQL statements

- The main statements are `SELECT`, `INSERT`, `UPDATE`, `JOIN`
- That will make more sense as we keep going

This is a table called `users`

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	Orangepizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

SELECT user_id from users

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	Orangepizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

SELECT user_id, user_modality from users

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	Orangepizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

SELECT * from users

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	Orangepizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

`WHERE` clauses

SELECT * from users where user_id="Jeff41"

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	Orangepizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

SELECT user_modality from users where user_id="Jeff41"

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	Orangepizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4

SELECT primary_key, user_ID from users where user_modality="tablet"

primary_key	user_ID	user_modality	satisfaction_score
1	Jeff41	PC	2
2	Orangepizza	tablet	5
3	User4958	tablet	4
4	Cookbook_Revolution	phone	4