Lecture 8A (Monday, Feb. 26)

Logistics

- Quiz next week
- HW2 due Wednesday
- HW3 out this week
- Almost done w/ HW2 grading

Today: normalization.

Normalization

- Many ways to break down tables
- We have tried to use reasonably good design in examples so far
- Standardized ways of avoiding certain types of design flaws

Why learn these?

- Identify specific types of flaws, have a way to talk about them
- Standardized recipes for fixing them
- Know what you're doing when you break them

Data can be represented in *normal forms*.

Definitions

- **Primary key** - we've seen this.
- **Key** - a set of attributes that uniquely identify members of the relation.
- **Candidate key** - a key that is not the primary key
- **Prime attribute** - an attribute that is part of a key
- **Non-prime attribute** - an attribute that is not part of a key

First Normal Form

- Textbook: a table is in 1NF if each attribute may only contain a single, atomic value.
- Similar definition: each entry (row/column intersection) must have exactly one value. No lists, no structs.
    - can think of null as violating this, but we won't.
- My informal definition: the table is a *table*.

SQL doesn't let you violate 1NF (easily). PostgreSQL allows you to by declaring a column to be of an array or JSON type, but we will not be using it.

Second Normal Form

- Textbook: a table is in 2NF if no non-key column is dependent on only a portion of the primary key.
- Must first be in 1NF. Normal forms are hierarchical — if a table is in 2NF, it must be in 1NF too.
- Simplified: All non-key columns are applicable to the entire primary key.
- My informal: each row describes the whole entity

The primary key *identifies* the object that each row is about. The remaining fields *describe* that object.

2NF says that each field needs to describe the entire object, not just a part of it.
Why do we want this?

- Reduce redundancy — record the information for each entity once
- Simplified edits, deletes, etc.

If the primary key is not compound, 2NF is automatically achieved.

Third Normal Form

- TExtbook: 3NF if no nonprime attribute is transitively dependent on the primary key.
- Other textbook: 3NF if the only determinants are candidate keys.
- Simplified: if the value of some field fixes the values of other fields, then the table is not in 3NF.
- My informal: each row describes just the entity, not related entities

Be careful to decompose on meaningful keys. We like synthetic primary keys.
Why?

- More redundancy reduction
- Easy updates

However, be careful!

- Price lists - what price did the user actually pay? (discounts, price changes, etc.)

Boyce-Codd Normal Form

Like 3NF, except that no functional dependencies are allowed in key attributes either.

Interlude: Synthetic vs. Natural Primary Keys

- Seems ‘natural’ to use natural primary keys
- When single-valued (non-composite), makes sense
    - part numbers
- But what about collisions?
    - different part vendors reusing numbers, oops!
    - composite key: vendor + part
- Synthetic makes editing easier
- Sometimes natural keys aren't
    - Names

Fourth Normal Form

- Textbook definition is difficult and not very obvious
- Basic idea: when you break out a table to make 1NF, you need a separate table for each group that is being broken.
- If an entity has a one-to-many (or many-to-many) with multiple things, each thing needs a separate table for linking.
- Example: the publisher records book sales
- Why? Coupling, and making well-defined entities.
- My informal: each row describes one thing, not two or more

Fifth Normal Form

- We do not worry about this one so much.
- Split out data basically as much as possible
- Gets rid of a lot of NULLs, but they still come back, but replaces them with outer joins.

General Guidance

- Naming is hard, but it is a good guide.
    - If there really isn't a good name, is it the right thing? E.g. 4NF violations.
- Ask questions about the data model
    - What are the things?
    - What are their relationships?
    - What is the data tracked?

Going forward

Always use 4NF in this class.