Normalization

2024-02-13

Normalization is going to feature heavily on midterm.

Initial Table - Short Stories

Bad version, not normalized at all.

Table:Publications

Author	Title	Genre	Book	PageLocation	CopyrightYear	ISBN
Tookey,Skoda						

Normalization

Data only appears once in any given table.

First Normal Form

Good link explaining the different forms * No repeating groups * Only one datapoint in a element * Relationshipt: N:M (n to many)
* Can you have more than one story in a book? no? Then Story defines a book, put book ID in Story table * Can you have a story in more than 1 book? No? Then book defines a story, put storyID in book table

```
*  many to many
    * Need a separate relational table
    * Book doesn't define story, story doesn't define book.

        |StoryID|BookID|
        |---|---|
```

Table: Author

Author	AuthorID	AuthorName	Phone	PenName
1	Tookey			
1	Skoda			

Table: Write

AuthorID	StoryID
1	1
2	1

Table: Publications

StoryID	Title	Page
1	My Story	102

Table:Book

BookID	Title	Copyright	ISBN
1	Story Book	2022	123-345-567

Table: Story_Book

StoryID	BookID
1	1

Table: Genre

GenreID	GenreName
1	Biography

Table: Story_Genre

GenreID	StoryID
1	1

2nd Normal Form

- if A and B determines C, and A determines D
  - C depends on A&B
  - D depends on a part of A&B <-**BAD**
    - Violation of 2nd normal form
    - **cannot depend on part of a multi-attribute key**

3rd Normal Form

- A determines B, which determines C <-**BAD**>
  - Violation of 3rd normal form
  - **No transitive dependance**

4th Normal Form

- key defines a group
- don't need to know this right now

Projection Join normal form problem

- Splitting a group, you can't rejoin to create original example

Midterm Review

- March 12-ish?, during lab session
- Breakout room, by yourself, screensharing.

1. given a design,
    a. what tables are entities, how do you know
    b. Which are relationshipes
    c. which are attributes?

Questions about queries, writing queries, explain concepts. FD - File determinancy (arrows, what determines what)

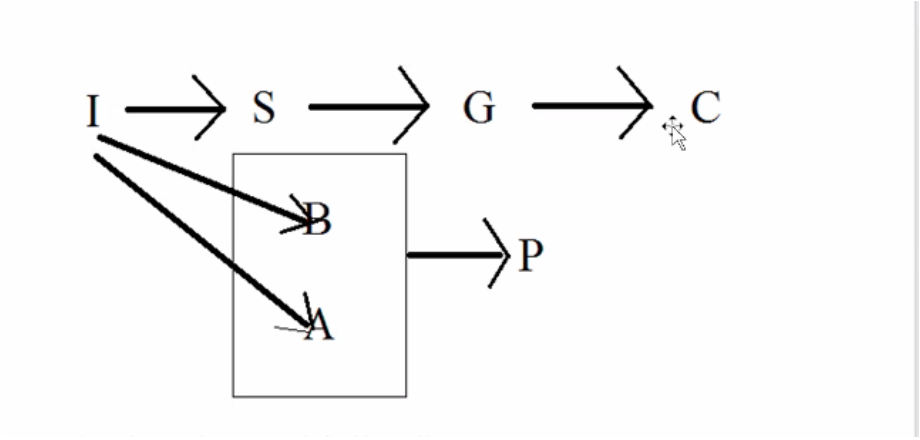

Figure 1: Midterm FD Example

2. Look for inconsistency

- Starting with B&A determine p
    - any time you see b1 and a1 , it determines p1 (based on first row).
    - do you see any b1+a1 not equaling p1?
- i determines s, so:
    - does any i values tie to a different s value? i1->s1, any i1s that don't equal s1?

3. Given the FD diagram above and data below, look for an inconsistency. Explain what the inconsistency is, or why there isn't one.

I	B	A	P	S	G	C
i1	b1	a1	p1	s1	g1	c1
i2	b2	a2	p2	s1	g1	c1
i3	b3	a3	p1	s2	g2	c2
i4	b4	a4	p2	s2	g2	c2
i5	b5	a5	p5	s1	g2	c2
i6	b6	a6	p7	s2	g2	c2

Figure 2: Midterm table example

- s determines g, so:
  - same as above
- G determines c
  - same as above
- This table isn't in 3rd normal form
  - no transitive dependence.
    - We'd need to split up a lot on the table
    - b, a, and p would need to be in a different table.
    - b and a would still be in original table.
  - example design

B	A	P
b1	a1	p1
b2	a2	p2
b3	a3	p1
b4	a4	p2
b5	a5	p5
b6	a6	p7

S	G
s1	g1
s2	g2

S	G
g1	c1
g2	c2

I	A	B	S
i1	a1	b1	s1
i2	a2	b2	s1
i3	a3	b3	s2
i4	a4	b4	s2
i5	a5	b5	s1
i6	a6	b5	s2

  - In this example, we have 7 attributes, and 6 rows = 42 elements.
  - In the normalized version we have `3 * 6 = 18, 2*2-4, 2*2=4, 3*6=18 (sum=44)`, sometimes has less storage space requirements.
    * Also saves on redundancy, especially in first normal form.

3. Create an entity diagram for a situation.
4. Access formatting:
   - answer what is gained by splitting tables in a particular way
   - or how would the data look in the split tables.

- When it gets closer to the midterm, go through these and practice.

Project:

When done designing tables, send in to check for normalization.