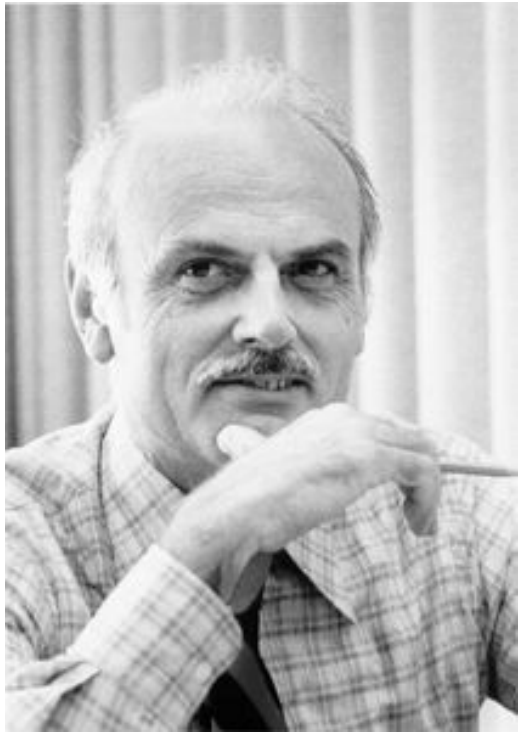

MONTANA
STATE UNIVERSITY

Database Normalization

Designing a Proper Database

Database Normalization

- Structuring database tables such that
  - Redundancy is minimized
  - Data integrity is maximized
- Edgar F Codd: a pioneer in databases
  - Proposed “1st Normal Form” in 1970
  - Went on to propose many more increasingly strict normalized forms



Database Normalization

- Most famous Normal Form is BCNF: Boyce-Codd Normal Form
- We will be focusing on
  - 1st Normal Form
  - 2nd Normal Form
  - 3rd Normal Form & BCNF
- BCNF and 3NF are equivalent in the presence of a single column key (surrogate or synthetic key, most tables today)

UNF (1970)	1NF (1970)	2NF (1971)	3NF (1971)	EKNF (1982)	BCNF (1974)	4NF (1977)	ETNF (2012)	5NF (1979)	DKNF (1981)	6NF (2003)
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓
✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	N/A
✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	N/A
✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	N/A
✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	N/A
✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	N/A
✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	N/A
✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓

Our Database

- Consider the following simple database
- This database is *not* normalized
- Let's fix it

Grades

Grade	Student	Class	Teacher	Satisfied?
B	Joe Smith	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
A	Kelly Chen	CSCI 440	M Wittie	Yes
B	Xerces Orion	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
C	Ted Jacobs	CSCI 440	M Wittie	Yes

1st Normal Form

- To Be in 1NF, there must be a key
- Let’s review the concepts and terminology around keys

Grades

Grade	Student	Class	Teacher	Satisfied?
B	Joe Smith	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
A	Kelly Chen	CSCI 440	M Wittie	Yes
B	Xerces Orion	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
C	Ted Jacobs	CSCI 440	M Wittie	Yes

What Are Keys?

- A key is a set of attributes (columns) that *uniquely* determine a row
- The literature uses the term *superkey*, which makes it sound cool
- In a relation with no duplicates, the set of all columns is a *superkey*
- A *candidate key (aka minimal superkey)* is a key from which no attributes can be removed without causing it to no longer be a key
- The *primary key* is the candidate key used to identify rows in the relation
- In practice, an ID column is typically added to a relation to be the primary key. This is sometimes called a *surrogate* or *synthetic key.*

1st Normal Form

- So, do we have a key in this relation?

Grades

Grade	Student	Class	Teacher	Satisfied?
B	Joe Smith	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
A	Kelly Chen	CSCI 440	M Wittie	Yes
B	Xerces Orion	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
C	Ted Jacobs	CSCI 440	M Wittie	Yes

1st Normal Form

- So, do we have a key in this relation?
- No. There are duplicate rows, so no set of columns uniquely identifies a row.

Grades

Grade	Student	Class	Teacher	Satisfied?
B	Joe Smith	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
A	Kelly Chen	CSCI 440	M Wittie	Yes
B	Xerces Orion	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
C	Ted Jacobs	CSCI 440	M Wittie	Yes

1st Normal Form

- We need to eliminate this duplicate data to get into 1NF
- Most real world databases there will be a surrogate key column, making the database trivially 1NF (Problem?)

Grades

Grade	Student	Class	Teacher	Satisfied?
B	Joe Smith	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
A	Kelly Chen	CSCI 440	M Wittie	Yes
B	Xerces Orion	CSCI 366	C Gross	Yes
C	Ted Jacobs	CSCI 440	M Wittie	Yes

1st Normal Form

- NB: Students can take multiple classes.
- We do not have a single column key here. So, what is our key?

Grades

Grade	Student	Class	Teacher	Satisfied?
B	Joe Smith	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
A	Kelly Chen	CSCI 440	M Wittie	Yes
B	Xerces Orion	CSCI 366	C Gross	Yes
C	Ted Jacobs	CSCI 440	M Wittie	Yes

1st Normal Form

- NB: Students can take multiple classes.
- We do not have a single column key here. So, what is our key?
- Student + Class Uniquely determines a row so {Student, Class} is our key

Grades

Grade	Student	Class	Teacher	Satisfied?
B	Joe Smith	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
A	Kelly Chen	CSCI 440	M Wittie	Yes
B	Xerces Orion	CSCI 366	C Gross	Yes
C	Ted Jacobs	CSCI 440	M Wittie	Yes

2nd Normal Form

- To achieve 2NF, all data must depend on the entire key
- Again, this is trivially true with surrogate keys
  - You can start to see why surrogate keys became a standard

Grades

Grade	Student	Class	Teacher	Satisfied?
B	Joe Smith	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
A	Kelly Chen	CSCI 440	M Wittie	Yes
B	Xerces Orion	CSCI 366	C Gross	Yes
C	Ted Jacobs	CSCI 440	M Wittie	Yes

2nd Normal Form

- We know the key is {Student, Class}
- Is there any data that depends only on part of that key?

Grades

Grade	Student	Class	Teacher	Satisfied?
B	Joe Smith	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
A	Kelly Chen	CSCI 440	M Wittie	Yes
B	Xerces Orion	CSCI 366	C Gross	Yes
C	Ted Jacobs	CSCI 440	M Wittie	Yes

2nd Normal Form

- Teacher depends **only** on Class!
- To fix this, we need to pull Teacher data out to a separate table

Grades

Grade	Student	Class	Teacher	Satisfied?
B	Joe Smith	CSCI 366	C Gross	Yes
A	Marge Liu	CSCI 366	C Gross	Yes
A	Kelly Chen	CSCI 440	M Wittie	Yes
B	Xerces Orion	CSCI 366	C Gross	Yes
C	Ted Jacobs	CSCI 440	M Wittie	Yes

2nd Normal Form

- We are now in 2NF
- Note that `C Gross` and `M Wittie` only appear once
  - Data redundancy has been removed
  - Easier to avoid update errors

Teaching

Class	Teacher
CSCI 366	C Gross
CSCI 440	M Wittie

Grades

Grade	Student	Class	Satisfied?
B	Joe Smith	CSCI 366	Yes
A	Marge Liu	CSCI 366	Yes
A	Kelly Chen	CSCI 440	Yes
B	Xerces Orion	CSCI 366	Yes
C	Ted Jacobs	CSCI 440	Yes

3rd Normal Form

- We can do better! There is still redundant data here!
- 3NF demands that all data depend *only* on the key
- What data here that does not depend on the key?

Teaching

Class	Teacher
CSCI 366	C Gross
CSCI 440	M Wittie

Grades

Grade	Student	Class	Satisfied?
B	Joe Smith	CSCI 366	Yes
A	Marge Liu	CSCI 366	Yes
A	Kelly Chen	CSCI 440	Yes
B	Xerces Orion	CSCI 366	Yes
C	Ted Jacobs	CSCI 440	Yes

3rd Normal Form

- Satisfied does not depend on the key
- Rather, it depends on the Grade column
- OK, so let’s pull that out as well

Teaching

Class	Teacher
CSCI 366	C Gross
CSCI 440	M Wittie

Grades

Grade	Student	Class	Satisfied?
B	Joe Smith	CSCI 366	Yes
A	Marge Liu	CSCI 366	Yes
A	Kelly Chen	CSCI 440	Yes
B	Xerces Orion	CSCI 366	Yes
C	Ted Jacobs	CSCI 440	Yes

3rd Normal Form

- Satisfied does not depend on the key
- Rather, it depends on the Grade column
- OK, so let's pull that out as well!

Teaching

Class	Teacher
CSCI 366	C Gross
CSCI 440	M Wittie

Satisfied

Grade	Satisfied?
A	Yes
B	Yes
C	Yes
D	No
F	No

Grades

Grade	Student	Class
B	Joe Smith	CSCI 366
A	Marge Liu	CSCI 366
A	Kelly Chen	CSCI 440
B	Xerces Orion	CSCI 366
C	Ted Jacobs	CSCI 440

3rd Normal Form

- We now have a database in 3NF
- It is also in BCNF
- 3NF typically satisfies BCNF, especially with surrogate keys

Teaching

Class	Teacher
CSCI 366	C Gross
CSCI 440	M Wittie

Satisfied

Grade	Satisfied?
A	Yes
B	Yes
C	Yes
D	No
F	No

Grades

Grade	Student	Class
B	Joe Smith	CSCI 366
A	Marge Liu	CSCI 366
A	Kelly Chen	CSCI 440
B	Xerces Orion	CSCI 366
C	Ted Jacobs	CSCI 440

3rd Normal Form

- What have we accomplished?
- Data redundancy has been minimized
- Update complexity has been minimized
  - E.g. it is easy to change “Satisfied” criteria now

Teaching

Class	Teacher
CSCI 366	C Gross
CSCI 440	M Wittie

Satisfied

Grade	Satisfied?
A	Yes
B	Yes
C	Yes
D	No
F	No

Grades

Grade	Student	Class
B	Joe Smith	CSCI 366
A	Marge Liu	CSCI 366
A	Kelly Chen	CSCI 440
B	Xerces Orion	CSCI 366
C	Ted Jacobs	CSCI 440

Normal Form Summary

- Each non-key column in a relation depends on
  - The key (1NF)
  - The whole key (2NF)
  - And nothing but the key (3NF/BCNF)
  - *So help me Cobb* ;)
- In the presence of a surrogate key, things become pretty obvious
  - In industry, there is *always* a surrogate key
- What's The General Principle?

Normal Form Summary

- Each non-key column in a relation depends on
  - The key (1NF)
  - The whole key (2NF)
  - And nothing but the key (3NF/BCNF)
  - *So help me Cobb* ;)
- In the presence of a surrogate key, things become pretty obvious
  - In industry, there is *always* a surrogate key
- What's The General Principle?

Don't Repeat Yourself! (DRY)

Denormalizing

- We’ve talked about how to normalize a database
- Would you ever want to *de*normalize a database?

Denormalizing

- Yep! (Careful talking with the DBA though!)
  - Performance is the biggest reason to denormalize data
  - Say you wanted to find all students who didn't pass a class. The denormalized table would be *faster* to work with
  - No need to merge two tables together, just a simple filter

Grades

Grade	Student	Class	Satisfied?
B	Joe Smith	CSCI 366	Yes
A	Marge Liu	CSCI 366	Yes
A	Kelly Chen	CSCI 440	Yes
B	Xerces Orion	CSCI 366	Yes
C	Ted Jacobs	CSCI 440	Yes

Denormalizing

- Denormalization is basically caching at the database level
- *“There are only two hard things in Computer Science: cache invalidation and naming things.” -- Phil Karlton*
- Be careful, but judicious denormalization can be a big win!

Grades

Grade	Student	Class	Satisfied?
B	Joe Smith	CSCI 366	Yes
A	Marge Liu	CSCI 366	Yes
A	Kelly Chen	CSCI 440	Yes
B	Xerces Orion	CSCI 366	Yes
C	Ted Jacobs	CSCI 440	Yes


MONTANA
STATE UNIVERSITY