Normalization

Amos Azaria

Dependencies

- An attribute (or set of attributes), B, is said to be dependent of another attribute (or set of attributes), A, if there exists a relation (function) such that $A \rightarrow B$.
- In other words, if given A, it is not possible for an entry to have two different values for B, we say that $A \rightarrow B$.
- For example, A = student ID, B=student first name.
- This dependency is also called functional dependency (B is functionally dependent of A).

Dependencies

- Obviously, for every B such that B $\subseteq$ A, we have that A → B.
  - E.g.: A = stFirstName, stLastName. B = stFirstName

A	B	Dependency?
{Street, City, State}	Zip code	A→B
Day of week	Date = {Day, Month, Year}	B→A
First Name	Last Name	None
{University, Department}	DepartmentHeadId	A→B and B→A

id	age	gender	degree	firstName	lastName
111	21	1	1	Chaya	Glass
444	23	0	1	Moti	Cohen
222	28	1	3	Tal	Negev
333	24	0	1	Gadi	Golan

Keys

…s, Minir…

id	name	lecturer	year	semste
10	Introduction to intro.	Knows Nothing	2020	1
20	Calculus	Tamar Ezra	2021	1
30	Algebra	Shay Mann	2022	1
35	Calculus	Adel Smith	2022	1
40	Advanced Program...	David Gol	2022	2

- A minimal set of attributes that determines an entry. That is, all other attributes are dependent on the key E.g.:
  - Student id in student table.
  - Course table: {id} or {name, year, semester}.

Minimal set: removal of any attribute from the set, will no longer determine the entry.

- Is student name a key?
  - No (there may be multiple students with the same name)
- What would be a key for the grades table?
  - Student id + course id

courseId	studentId	grade	passed
20	111	43	0
20	222	85	1
30	111	90	1
30	444	95	1
40	222	67	1
40	333	40	0

Keys (cont.)

- A single table can have more than one set of keys (both being minimal), e.g.:
  - R(university, department, depHeadId)
    - depHeadId
    - {university, department}

Assuming every department has a single head, and a person can be a department head of a single department in a single university.

Prime / Non-Prime

- Prime attributes are attributes that are part of some candidate-key.
- Similarly, non-prime attributes are attributes that are not part of any candidate-key.

Super-Key

- **Any** set of attributes that determines an entry.
  - E.g. the whole set of attributes.
- Same as candidate key, just without the minimal requirement.

Normalization

- What is the problem with the following relation?

Heavy redundancy.
What happens when we update student's address? And what if we delete all grades of a student?

		udentLast	Courses
		gasi	4244, 3423, 6734
956	Ta	Atiya	4244, 5437

Multiple values for a single attribute. How can we get all students in 3423?

StudentId	StudentFirst	StudentLast	Address	CourseId	Grade
542	Yossi	Agasi	Harambam 45,	4244	87
542					65
956	Tamar		Herzeliya	4244	86
542	Yossi	Agasi	Harambam 45, Ariel	6734	80

Every table should hold a single "idea" or "theme".

1NF (=Normalized Form)

- Every attribute must hold a single atomic value (searchability)

StudentId	StudentFirst	StudentLast	Courses
542	Yossi	Agasi	4244
956	Tamar	Atiya	4244
754	Gabbi	Matar	4325
327	Shay	Shalom	5324
542	Yossi	Agasi	3423
542	Yossi	Agasi	6734
956	Tamar	Atiya	5437
754	Gabbi	Matar	6543
754	Gabbi	Matar	564

2NF

- Table must be in 1NF
- Non-prime attributes do not depend on a (strict/proper) subset of a candidate key.

But StudentFirst, StudentLast and Address depend only on StudentId

What is the key?

StudentId+CourseId

StudentId	StudentFirst	StudentLast	Address	CourseId	Grade
542	Yossi	Agasi	Harambam 45, Ariel	4244	87
542	Yossi	Agasi	Harambam 45, Ariel	3423	65
956	Tamar	Atiya	Hadekel 12, Herzeliya	4244	86
542	Yossi	Agasi	Harambam 45, Ariel	6734	80

Fixing Table to Become 2NF

- In order to correct a relation that is not in 2NF, we split the information into 2 tables:

StudentId	StudentFirst	StudentLast	Address
542	Yossi	Agasi	Harambam 45, Ariel
956	Tamar	Atiya	Hadekel 12, Herzeliya

StudentId	CourseId	Grade
542	4244	87
542	3423	65
956	4244	86
542	6734	80

956	Tamar	Atiya	Hadekel 12, Herzeliya	4244	86
542	Yossi	Agasi	Harambam 45, Ariel	6734	80

Note that the new tables have 20 cells in total, while the original table had 24 cells. The new tables have 105 characters (combined) while the old table had 143.

2NF (cont.)

- Given: R(author, bookId, #pages)
- What is the candidate key?
  - {author, bookId}
- Is it in 2NF?
  - No:
    - bookId $\rightarrow$ #pages
    - {bookId} isn't a key
- How to fix?
  - Split to R1(author, bookId) and R2(bookId, #pages)

3NF

City depends on Zip. (also Zip depends on {Address, City})

- Table must be in 2NF
- Non-prime attributes cannot depend on any set that isn't a super-key (transitive dependency)

except trivial

StudentId	StudentFirst	StudentLast	Address	City	Zip
542	Yossi	Agasi	Harambam 45	Ariel	40743
956	Tamar	Atiya	Hadekel 12	Herzeliya	65475

StudentId	StudentFirst	StudentLast	Address	Zip
542	Yossi	Agasi	Harambam 45	40743
956	Tamar	Atiya	Hadekel 12	65475

Zip	City
40743	Ariel
65475	Herzeliya

StId	StFirst	StLast	Address	City
542	Yossi	Agasi	Harambam 45	Ariel
956	Tamar	Atiya	Hadekel 12	Herzeliya

Address	City	Zip
Harambam 45	Ariel	40743
Hadekel 12	Herzeliya	65475

3NF (cont.)

- Given: R1(studentId, courseId, grade, passed), [R2(courseId, passingGrade)]
- What is the candidate key in R1?
  - {studentId, courseId}
- What are the (non trivial) dependencies?
  - {studentId, courseId} → grade
  - {studentId, courseId} → passed
  - {courseId, grade} → passed
- Is it in 2NF?
  - Yes.
    - No attribute (including 'passed') is dependent on a subset of a key.
- Is it in 3NF?
  - No:
    - passed is non-prime
    - {courseId, grade} → passed
    - {courseId, grade} isn't a superkey

Boyce and Codd Normal Form (BCNF)

- BCNF is sometimes referred to as 3.5NF.
- Table must be in 3NF.
- For any two sets, X, Y, ($Y \not\subseteq X$) such that $X \rightarrow Y$, X is a super-key.
- Note that while, Y may (by definition) be a group, we can assume that Y is a single attribute.
- Note that if Y is prime, but $X \rightarrow Y$, and X is not a super-key, while the table might be in 3NF, it is not in BCNF.
- True or false?
  - Any 3NF table with a single candidate key is also in BCNF.

True: if Y is non-prime then from 3NF we get that X is a super-key. If Y is prime, assume by contradiction that X is not a super-key, if we replace Y with X in Y's candidate key (and minimize) we get a second candidate key (since $Y \not\subseteq X$)

BCNF (3.5NF) example

Street	City	Zip
Gilboa 32	Ariel	40726
Hatamar 12	Jerusalem	33673
Goren 45	Haifa	88645

- Dependencies:
  - Zip → City
  - {Street, City} → Zip
- Candidate-Keys:
  - {Street, Zip}
  - {Street, City}
- Super-Keys:
  - {Street, City, Zip}
- 3NF?
  - Yes (all attributes are prime)
- BCNF?
  - No, Zip → City, but {Zip} is not a super-key!

1-3.5NF

- The data depends on the key (1NF), the whole key (2NF) and nothing but the key (3NF + 3.5NF)

“No, I do not think ‘The truth, the whole truth, and nothing but the truth’ is overkill.”

BCNF (cont.)

- Look at the following table used in a mobile company:
- R(mobilePhoneNum, simSerialNumber, callDateTime)
- (Assume that once a phone number is burnt into a SIM card it can't be changed)
- Is it BCNF?
- Dependencies:
  - simSerialNumber → mobilePhoneNum
  - {callDateTime, mobilePhoneNum} → simSerialNumber
- Candidate-keys:
  - {callDateTime, simSerialNumber }
  - {callDateTime, mobilePhoneNum}
- simSerialNumber → mobilePhoneNum, but {simSerialNumber} is not a super-key.

4NF

- Look at the following table:

StudentId	Department	SportTeam
111	CS	Soccer
111	Biology	Soccer
111	CS	Baseball
111	Biology	Baseball
222	Biology	Basketball
222	Biology	Soccer
333	CS	Basketball

- The key is:
  - {studentId, department, sportTeam}
- It doesn't violate NF 1-3.5
- But still it seems wrong:
  - What happens if 111 joins another sprotTeam?
  - What happens if 222 joins another department?

4NF (cont.)

- 4NF requires BCNF + no multivalued dependencies.
- A multivalued dependency occurs when the presence of one or more rows in a table implies the presence of one or more other rows in that same table.
- That is, from observing some *rows,* one can deduce the presence of other rows.
- In our e… e department are inde… e multivalued depend…
- We writ…
  - stude…
  - stude…
- Every table should hold a single "idea" or "theme"!

StudentId	Department	SportTeam
111	CS	Soccer
111	Biology	Soccer
111	CS	Baseball
111	Biology	Baseball

Multivalued Dependency (Formal Definition)

- Multidependency is a condition on the existence of rows (entries / tuples / entities) in the relation.
- Given two sets of attributes, A, and B, we say that A multidetermines B (A -->> B) if:
  - Let C = R \ (A U B) (that is, all the rest of the attributes)
  - Given rows x and y, such that:
    - x[A] = y[A] and
    - x[B] ≠ y[B] and
    - x[C] ≠ y[C]
  - Entails that, there exists a row z, such that:
    - z[A] = x[A] (= y[A]) and
    - z[B] = x[B] and
    - z[C] = y[C]

	A	B	C
x	a1	b1	c1
y	a1	b2	c2
z	a1	b1	c2
w			

5NF

- 5NF is related to situations in which some rules are applied on the rows of the table. In such situations if the table can be decomposed into smaller tables by removing redundant data, the table is not in 5NF.
- "Only in rare situations does a 4NF table not conform to the higher normal form 5NF. These are situations in which a complex real-world constraint governing the valid combinations of attribute values in the 4NF table is not implicit in the structure of that table." (Wikipedia)
- Therefore, we won't be dealing with 5NF.

Question

- What NF is the following relation:
  - R(A,B,C,D)
  - $\{A,B\} \rightarrow D$
  - $\{A,D\} \rightarrow C$
- What are the candidate key(s)?
- Is it in 2NF?
- Is it in 3NF?

If we have an attribute that appears only on the right of the dependency list, what may we conclude?

If we have an attribute that does not appear on the right of the dependency list, what may we conclude?