Normalization

Evils of Redundancy

- *When a database schema is poorly designed we get anomalies.*
- *Redundancy* is at the root of several problems associated with relational schemas:

Redundant storage: data is repeated

Update anomalies: need to change in several places

Insertion anomalies: may not be able to add data we want to

Deletion anomalies: may lose data when we don’t want to

Functional Dependencies

- A kind of integrity constraint
  - (hence, part of the schema)
- Finding them is part of the database design

Recall that a **function** returns a single unique value when evaluated on any input.

Functional Dependencies

Table R(.... A_1, A_2, …, A_n… B_1, B_2, …, B_m …)

Functional Dependency:

$$\boxed{A_1, A_2, \ldots, A_n \rightarrow B_1, B_2, \ldots, B_m}$$

Meaning:

If two tuples agree on the attributes

$$\boxed{A_1, A_2, \ldots, A_n}$$

then they must also agree on the attributes

$$\boxed{B_1, B_2, \ldots, B_m}$$

Examples

EmpID	Name	Phone	Position
E0045	Smith	1234	Clerk
E1847	John	9876	Salesrep
E1111	Smith	9876	Salesrep
E9999	Mary	1234	Lawyer

- EmpID → Name, Phone, Position
- Position → Phone
- but Phone $\nrightarrow$ Position

Keys and functional dependencies

- A **superkey** for relation R is a set of attributes A_1, ..., A_n s.t. A_1, ..., A_n functionally determines all attributes of R.
- A **key** is a minimal superkey.

Anomalies

Hourly_emps(ssn, name, lot, rating, hourly_wages, hours_worked)

Suppose hourly wages is determined by rating:

rating → hourly_wages

ssn	name	lot	rating	hourly_wages	hours_worked
29	brutus	48	8	10	40
85	art	22	8	10	30
95	bob	35	5	7	30
96	frodo	35	5	7	32
22	dustin	35	8	10	40

- Redundant storage: association between rating 8 and hourly wages 10 repeated 3 times.
- Update anomalies: hourly_wages updated in first tuple but not second
- Insertion anomalies: must know hourly_wage for rating value
- Deletion anomalies: delete all tuples with certain rating value, lost assoc.

Can null values fix problems?

- Not really.
- Insertion anomaly:
  - What if we know rating and hourly_wages for some rating, but there is no employee with that rating?
  - No. ssn can’t be null.
- Deletion anomaly:
  - If last employee with some rating and hourly_wages value is deleted, replace with nulls?
  - No. ssn can’t be null.

Schema Refinement

- Integrity constraints, in particular *functional dependencies*, can be used to identify schemas with such problems and to suggest refinements.
- Main refinement technique: *decomposition* (replacing ABCD with, say, AB and BCD, or ACD and ABD).
- Decomposition should be used judiciously:
  - Is there reason to decompose a relation?
  - What problems (if any) does the decomposition cause?

Data Anomalies

Persons may have several phones:

Name	SSN	PhoneNumber	City
Fred	123-45-6789	206-555-1234	Seattle
Fred	123-45-6789	206-555-6543	Seattle
Joe	987-65-4321	908-555-2121	Westfield

Anomalies:

- Redundancy = repeat data
- Update anomalies = Fred moves to "Bellevue"
- Deletion anomalies = Joe deletes his phone number: what is his city ?

SSN → Name, City

but not SSN → PhoneNumber

Relation Decomposition

Break the relation into two:

Name	SSN	PhoneNumber	City
Fred	123-45-6789	206-555-1234	Seattle
Fred	123-45-6789	206-555-6543	Seattle
Joe	987-65-4321	908-555-2121	Westfield

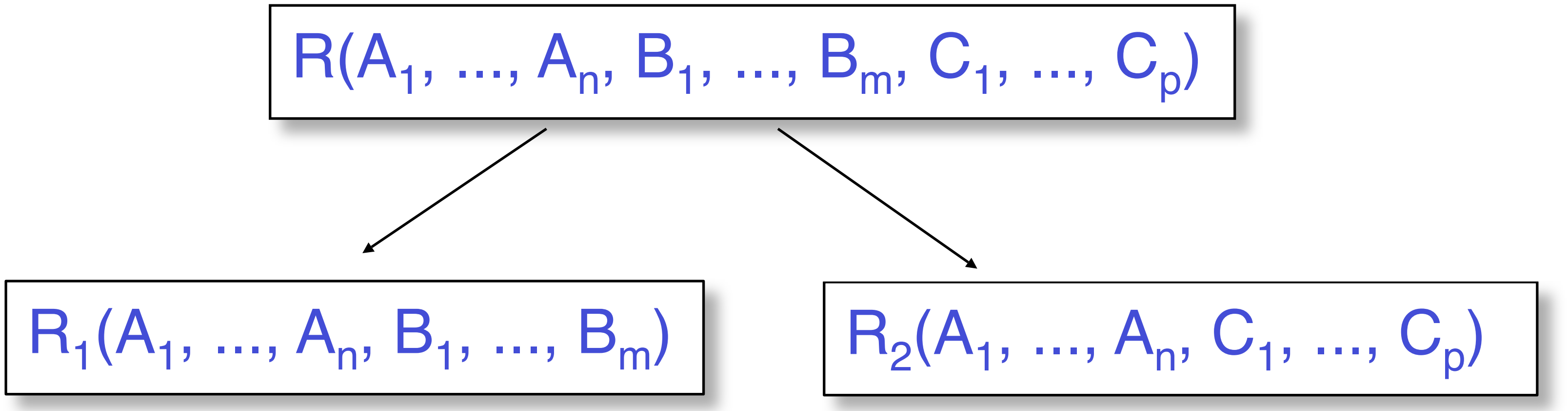
Name	SSN	City
Fred	123-45-6789	Seattle
Joe	987-65-4321	Westfield

SSN	PhoneNumber
123-45-6789	206-555-1234
123-45-6789	206-555-6543
987-65-4321	908-555-2121

Anomalies have gone:

- No more repeated data
- Easy to move Fred to “Bellevue” (how ?)
- Easy to delete all Joe’s phone number (how ?)

Decompositions in General



$R(A_1, \ldots, A_n, B_1, \ldots, B_m, C_1, \ldots, C_p)$

$R_1(A_1, \ldots, A_n, B_1, \ldots, B_m)$

$R_2(A_1, \ldots, A_n, C_1, \ldots, C_p)$

R_1 = projection of R on $A_1, \ldots, A_n, B_1, \ldots, B_m$

R_2 = projection of R on $A_1, \ldots, A_n, C_1, \ldots, C_p$

Lossless Decomposition

- Sometimes it is correct:

Name	Price	Category
Gizmo	19.99	Gadget
OneClick	24.99	Camera
Gizmo	19.99	Camera

Name	Price
Gizmo	19.99
OneClick	24.99

Name	Category
Gizmo	Gadget
OneClick	Camera
Gizmo	Camera

Lossy Decomposition

- Sometimes it is not:

Name	Price	Category
Gizmo	19.99	Gadget
OneClick	24.99	Camera
Gizmo	19.99	Camera

What's wrong ??

Name	Category
Gizmo	Gadget
OneClick	Camera
Gizmo	Camera

Price	Category
19.99	Gadget
24.99	Camera
19.99	Camera