NORMALIZATION

CONTENT

- Basic concepts
  - Why is normalization needed?
  - Redundancy and possible implications
  - The problem with NULL and how this should be dealt with
- Functional dependencies
  - What is a functional dependency?
  - Data modeling with functional dependencies – Bottom up modeling
  - Mapping between functional dependencies and ER
- Normalization
  - How to identify superkeys and candidate keys
  - Normal forms – The foundation for relational databases

BASIC CONCEPTS

- Why is normalization needed?
  - Modeling with ER and EER is not sufficient for all situations
    - Redesign of legacy systems
    - Simplification of generated tables by removing superfluous tables or attributes
    - Adding controlled redundancy as part of tuning process
    - ER and EER do not model all situations precisely due to default BCNF

BASIC CONCEPTS

HÖGSKOLAN I SKÖVDE

- There are a number of key types used in relational theory
  - Superkey
    - An attribute or set of attributes that which uniquely identifies a tuple (record, row) in a relation such as two identical tuples may never have identical values for all attributes in key (every key is unique, i.e. key integrity)
  - Candidate key
    - A minimal super key from which we cannot remove any attributes without losing the uniqueness constraint
  - Primary key
    - The selected candidate key (there might be several candidate keys)
  - Foreign key
    - An attribute or set of attributes having the same values as the primary key of another relation or the same relation (recursive relations)
    - The foreign key and the primary key it refers to do not have to have the same names (roles are possible to use, e.g. employee and manager)
    - A foreign key may or may not enforce the NOT NULL constraint

Bild 4

BASIC RELATIONAL THEORY

HÖGSKOLAN I SKÖVDE

- A relation must fulfill a set of basic properties
  - Unique name
  - Atomic values
  - Unique names within a relation
  - Domain of a allowed values
  - Set of attributes is unordered
  - Every tuple must be unique
  - Tuples are unordered

Bild 5

BASIC CONCEPTS

HÖGSKOLAN I SKÖVDE

- Redundancy is unnecessary repeated storage of data
  - Redundancy may potentially increase storage needs
  - Redundancy may potentially cause anomalies which may cause the database to enter an invalid state → the basic integrity constraints cannot be enforced
    - Insertion anomalies
    - Deletion anomalies
    - Modification anomalies
  - Managed redundancy may be introduced for efficiency reasons but should be handled very carefully

Bild 6

BASIC CONCEPTS

- Spurious tuples are one form of redundancy which occur when tables are joined but the information from the original tables cannot be recovered completely

CAR

LPNO
ABC123
FGH756
DCE321

PERSON

SSN	Name	Adress
560712-4444	John Stevens	Filvägen 5
750609-3333	John Stevens	Clustervägen 2
481125-9999	Carl Long	Diskvägen 2

OWNSCAR1

Name	LPNO
John Stevens	ABC123
John Stevens	CDE331
Carl Long	ABC123

Bild 7

BASIC CONCEPTS

PERSON*OWNSCAR1

560712-4444	John Stevens	Filvägen 5	ABC123
560712-4444	John Stevens	Filvägen 5	CDE331
750609-3333	John Stevens	Clustervägen 2	ABC123
750609-3333	John Stevens	Clustervägen 2	CDE331
481125-9999	Carl Long	Diskvägen 2	ABC123

Bild 8

BASIC CONCEPTS

CAR

LPNO
ABC123
FGH756
DCE321

PERSON

SSN	Name	Adress
560712-4444	John Stevens	Filvägen 5
750609-3333	John Stevens	Clustervägen 2
481125-9999	Carl Long	Diskvägen 2

OWNSCAR2

SSN	LPNO
560712-4444	ABC123
560712-4444	CDE331
481125-9999	ABC123

PERSON*OWNSCAR2

560712-4444	John Stevens	Filvägen 5	ABC123
560712-4444	John Stevens	Filvägen 5	CDE331
481125-9999	Carl Long	Diskvägen 2	ABC123

Bild 9

BASIC CONCEPTS

- Spurious tuples
  - Avoid spurious tuples by always joining tables on attributes forming a primary-foreign key relationship (as you been taught so far in this course)
- Redundancy in general
  - Design relational schemas so that anomalies are avoided. If this is not possible carefully note potential breakdowns so that they can be managed. Formal normalization methods may be used to eliminate the possibility of these anomalies ever occurring but there are many old databases....with a lot of fixes and patches out there....

BASIC CONCEPTS

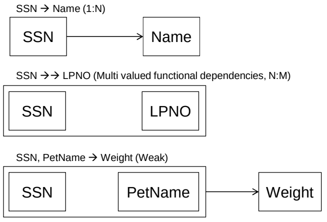
- NULL values
  - NULL is NOT the same as 0!!!
  - NULL is an empty cell in a relation
- NULL values often cause interpreting difficulties since several interpretations are possible:
  - The value is unknown
  - The value is known but has not been inserted yet
  - The value is known but is not applicable in a specific tuple
- NULLs cause problems with aggregates, such as COUNT()
- NULLs waste storage space
- Always try to avoid relations with many NULLs. In special cases (e.g. >90% NULLs), create an extra table with NULL column and primary key

FUNCTIONAL DEPENDENCIES

- The theory on which relational databases are built
- Have a lot in common with mathematical functions and are based on the idea that each value has exactly one corresponding value, e.g. SSN → Name
  - SSN gives name
  - SSN is a determinant for name (math talk)
  - SSN has a name. A name may have more than one SSN
  - This is similar to a 1:N relationship in ER, however, functional dependencies only look at attributes and not entities
- Functional dependencies can bee used for different purposes
  - Modeling technique (bottom up)
  - Formal theory used for analysis and normalization
  - Complement to other models, such as ER

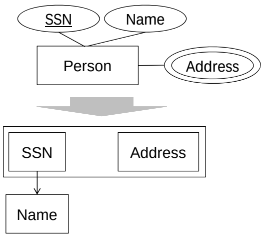
HÖGSKOLAN I SKÖVDE

FUNCTIONAL DEPENDENCIES – HOW TO DRAW FUNCTIONAL DEPENDENCIES



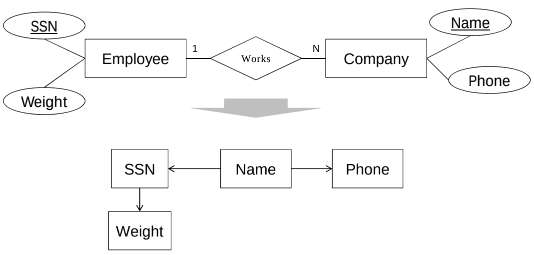
HÖGSKOLAN I SKÖVDE

FUNCTIONAL DEPENDENCIES – HOW TO DRAW FUNCTIONAL DEPENDENCIES



HÖGSKOLAN I SKÖVDE

FUNCTIONAL DEPENDENCIES – HOW TO DRAW FUNCTIONAL DEPENDENCIES



HÖGSKOLAN I SKÖVDE

FUNCTIONAL DEPENDENCIES – HOW TO DRAW FUNCTIONAL DEPENDENCIES

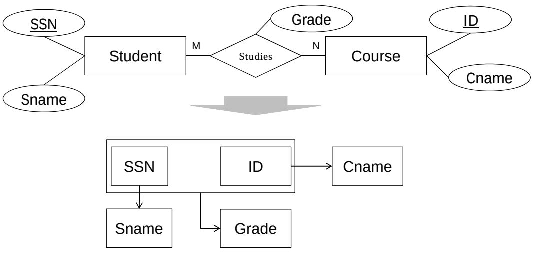


Bild 16

HÖGSKOLAN I SKÖVDE

FUNCTIONAL DEPENDENCIES – HOW TO DRAW FUNCTIONAL DEPENDENCIES

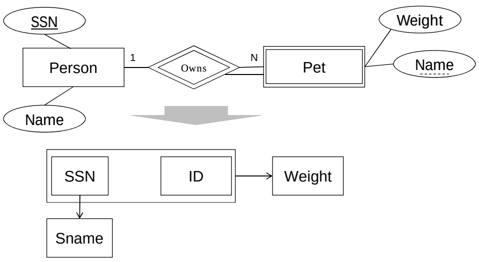


Bild 17

HÖGSKOLAN I SKÖVDE

FUNCTIONAL DEPENDENCIES – BOTTOM UP MODELING

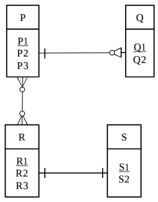
- A person is identified by SSN. Each person also has the properties name, address, and phone. A person may own several cars.
- A car is identified by license plate number and also has the properties color and brand. A car can be owned by one person only. A car may have one or more special items, such as a spoiler, dices in the front rear etc.
- Fast food meal is identified by name and also has the properties burger, drink, and toy.
- A feast occurs now and then. A feast is identified by the date the feast occurred along with the person having the feast, the fast food meal involved, and the car in which the feast occurred.
- Each feast also has a grade regarding the quality of the feast.

Bild 18

PRACTICE ASSIGNMENT

HÖGSKOLAN I SKÖVDE

1. Create a corresponding model in functional dependencies of the IE model to the right
2. What cannot be preserved when modeling with functional dependencies that is expressed in the IE model?



NORMALIZATION

HÖGSKOLAN I SKÖVDE

- The process of analyzing relational schemas based on their functional dependencies and primary keys to achieve
  - Minimized redundancy
  - Minimize the possibilities for anomalies to occur
- Based on stepwise rules used to decompose relations into more restrictive forms
- Starts with one big relation and usually ends up in a number of smaller relations connected via foreign keys

UNF → 1NF → 2NF → 3NF → BCNF → 4NF

One big table → Many small tables

NORMALIZATION

HÖGSKOLAN I SKÖVDE

- UNF (Un-normalized form)
  - A relation consisting of a set of attributes and in which an attribute may itself be a set, i.e. non atomic attributes
  - A relation missing a *VALID* key

SSN	Name	Address	Phone
661213-4552	Sture Fors Ove Sundby	Mosvägen 11	0735689741
690910-4256	Snaff Barkma	Grottstigen 23	0735698741
NULL	Benny Bly	Avenyn 2	0702987256
760902-8163	Lisa Lax	Porjus 4	0719532146
781116-4796	Pat Mesh	Köttstigen 5	0706541289

NORMALIZATION– HOW TO FIND A CANDIDATE KEY

HÖGSKOLAN I SKÖVDE

- In order to be in 1NF a valid key has to exist. This key (candidate key) may be identified using the X^+ algorithm
- Consider the following relation R in UNF along with the set of functional dependencies F covered by R:

  R(A, B, C, D, E, F)
  $F=\{A \rightarrow D, DE \rightarrow F, F \rightarrow C, B \rightarrow E\}$

- Try to identify a possible super key and then test it by using the X^+ algorithm

  $AB^0=AB$
  $AB^1=ABDE$
  $AB^2=ABDEF$
  $AB^3=ABDEFC$
  $AB^4=AB^3=AB^+$

- We know that AB is a super key for R, but is AB also a candidate key?

Bild 22

NORMALIZATION

HÖGSKOLAN I SKÖVDE

- 1NF (First normal form) is present if:
  - All attributes are atomic
  - A valid super key is present
- 2nf (Second normal form) is present if:
  - The relation is in 1NF
  - All attributes in the relation are fully functionally dependent of the WHOLE key in the relation
    - If the key has more than one attribute no non key attribute should be dependent of only a part of the key (partial dependency)
    - A relation having a valid key consisting of one attribute only is always in 2NF
    - If partial dependency exists, decompose the relation into smaller relations where non key attributes are dependent of the whole key of the relation

Bild 23

NORMALIZATION

HÖGSKOLAN I SKÖVDE

- 3NF (Third normal form)
  - A relation fulfilling the requirements for 1NF and 2NF
  - No non key attribute in the relation should be determined by another non key attribute (transitivity)
  - If transitivity exist, decompose the relation into smaller relations where each non key attribute is determined by a key attribute

Bild 24

NORMALIZATION

- Non trivial multivalued functional dependencies
  - Represents a dependency between attributes which are not related
  - Se for example the relationship between Person, Sport and Course
    - A person may study several courses and a course may be studied by several persons, SSN→→ID (N:M)
    - A person may practice several sports and a sport may be practiced by several persons, SSN→→SportName (N:M)
  - If one single relation is used to illustrate these dependencies, i.e. SSN→→ID|SportName (<u>SSN, ID, SportName</u>), massive redundancy will be the result
  - Interferes with the laws of integrity
- Non trivial multivalued dependencies shall be avoided

Bild 25

NORMALIZATION

- 4NF (Fourth normal form)
  - A relation fulfilling the requirements for 1NF, 2NF, and 3NF
  - No non trivial multivalued functional dependencies may exist in the relation
  - If so, split the relation in a semantically consistent way so that no non trivial multivalued dependencies exist

Bild 26

NORMALIZATION

- BCNF (Boyce Codd Normal Form)
  - BCNF requires 1NF, 2NF, and 3NF but is not as restrictive as 4NF
  - A relation is in BCNF when every determining attribute also is the super key for the relation
    - 3NF: No non key attribute in the relation should be determined by another non key attribute (transitivity)
    - BCNF: No attribute at all, i.e. key attributes or non key attributes, should be determined by a non key attribute
    - Formally spoken: A relation R is in BNCF if whenever a non trivial functional dependency X→A holds in R, then X is a super key for R. The only difference between 3NF and BCNF is that 3NF allows A to be prime, i.e. only if X→A holds in a relation R with X not being a super key AND A being a prime attribute will R be in 3NF but not in BCNF

Bild 27

NORMALIZATION – PRACTICE EXAMPLE 1

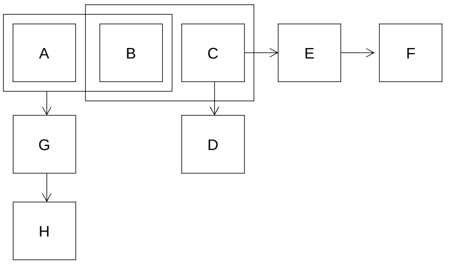


Bild 28

NORMALIZATION – PRACTICE EXAMPLE 2

Explain which normal form is applicable for the following relations and their corresponding sets of functional dependencies:

a) R(A, B, C, D) F={A→B, A→C, C→D}
b) R(A, B, C, D) F={A→B, A→CD, D→A}
c) R(A, B, C, D) F={A→B, A→CD}
d) R(A, B, C, D) F={A→B, B→CD}
e) R(A, B, C, D) F={A→B, B→CD}

Bild 29

NORMALIZATION – PRACTICE EXAMPLE 2

Explain which normal form is applicable for the following relations and their corresponding sets of functional dependencies:

a) R(A, B, C, D) F={A→B, A→C, C→D}
b) R(A, B, C, D) F={A→B, A→CD, D→A}
c) R(A, B, C, D) F={A→B, A→CD}
d) R(A, B, C, D) F={A→B, B→CD}
e) R(A, B, C, D) F={A→B, B→CD}

Bild 30

NORMALIZATION – PRACTICE EXAMPLE 3

Överför följande domänbeskrivning till en modell som visar vilka funktionella beroenden som finns mellan de attribut som är relevanta. Identifiera därefter en kandidatnyckel och överför modellen till relationer i 1NF, 2NF, 3NF och 4NF.

Person: En person identifieras av personnummer och har även egenskaperna namn och adress. En person kan äga flera bilar och gilla flera maträtter. För varje maträtt en person gillar skall en kommentar kunna lagras.

Bil: En bil identifieras av registreringsnummer och har även egenskaperna färg, märke och typ. En bil kan endast ägas av en person.

Maträtt: En maträtt identifieras av sitt namn och tillhör en viss kategori, exempelvis asiatiskt, rawfood och vegetarisk. En maträtt kan gillas av flera personer och innehåller flera ingredienser.

Ingrediens: En ingrediens identifieras av sitt namn och har även egenskaperna pris och sort. En ingrediens kan ingå i flera olika maträtter.