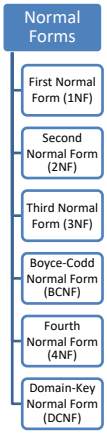
Northeastern University
Khoury College of Computer and Information Sciences

Normalization & Normal Forms

40

Need for Normal Forms



- Relational tables should follow a specific structure that eliminates (or at least reduced) several common anomalies in transaction processing databases when data in inserted, updated, or removed.

Northeastern University
Khoury College of Computer and Information Sciences

41

Normalization vs Design

- Normalization is required when a database is not properly designed through data modeling.
- The most common is the "spreadsheet design" approach that many (novice) database designers use: *lump all data into a single table like a spreadsheet*.
- A database that suffers from *Spreadsheet Syndrome* is subject to numerous data redundancies, data anomalies, and is generally inefficient.

Northeastern University Khoury College of Computer and Information Sciences

42

Do All Databases Need to be Normalized?

➢Short answer: *generally no…*

When a database follows object-oriented design principles it is almost always in BCNF.

If the database starts out as a single (or a few) tables, then the cure for *Spreadsheet Syndrome* is normalization.

Northeastern University Khoury College of Computer and Information Sciences

43

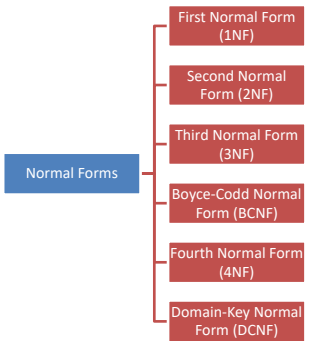
Approaches to Database Design

- ✓ **Top-Down Approach**: Start with a conceptual model based on grouping attributes into their natural entities.
- ✗ **Bottom-Up Approach**: Define tables and their relations and then move attributes between tables until normalized.
- ✗ **Brute-Force Approach**: Put all attributes into a single table and keep breaking tables apart until all are in at least BCNF.

Northeastern University Khoury College of Computer and Information Sciences

44

Which Normal Form is needed?

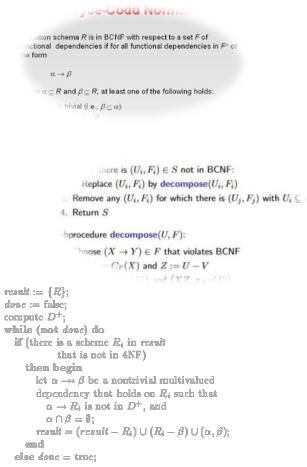


- Generally BCNF (a stronger version of 3NF) is sufficient.
- Well designed conceptual models are generally in BCNF.

Northeastern University Khoury College of Computer and Information Sciences

45

How do I normalize my database?



- You can start normalizing a database by:
  1. Inspection of the tables
  2. Determining functional dependencies and applying normalization algorithms
- Approach (1) is generally “easier” and less theoretic but might sometimes miss a dependency.
- Bottom line: getting a database into normal form is not that hard and is “common sense”.

Northeastern University Khoury College of Computer and Information Sciences

46

Normalization Strategies

- Start with a single table, determine functional dependencies, and decompose tables until all tables are in BCNF.

  and/or

- Define at a set of tables as best as possible and ensure that each is at least in BCNF. Decompose tables as required until all are in BCNF.

Northeastern University Khoury College of Computer and Information Sciences

47

Do all databases need to be normalized and be in at least BCNF?

- Absolutely not…
- Some database are on purpose not normalized because normalization increases the number of tables and increases joins.
- Joins are "costly", *i.e.*, they require memory and compute time and can be "slow".

48

Which databases should be normalized?

Yes
- **Transactional databases**, e.g., those that are updated frequently and hold customer data, should be normalized to avoid certain anomalies.

No
- **Analytical databases** are generally denormalized to increase retrieval speed, but that comes at the price of updates being slow. However, analytical databases are not updated in "real time" but rather periodically, *e.g.*, daily, weekly, monthly.

49

What about those anomalies?

- A database that is not normalized can exhibit certain anomalies when updated, *e.g.*, rows are inserted, deleted, or changed in a table.

ISBN	Book_title	P_ID	Pname	Phone
001-987-760-9	C++	P001	Hills Publications	7134019
001-354-921-1	Ransack	P001	Hills Publications	7134019
001-987-650-5	Differential Calculus	P001	Hills Publications	7134019
002-678-980-4	DBMS	P002	Sunshine Publishers Lt d.	6548909
002-678-880-2	Call Away	P002	Sunshine Publishers Ltd.	6548909
004-765-409-5	UNIX	P003	Bright Publications	7678985
004-765-359-3	Coordinate Geometry	P003	Bright Publications	7678985
003-456-433-6	Introduction to German Language	P004	Paramount Publishing House	9254834
003-456-533-8	Learning French Language	P004	Paramount Publishing House	9254834

Northeastern University Khoury College of Computer and Information Sciences

50

Examples of Potential Anomalies

ISBN	Book_title	P_ID	Pname	Phone
001-987-760-9	C++	P001	Hills Publications	7134019
001-354-921-1	Ransack	P001	Hills Publications	7134019
001-987-650-5	Differential Calculus	P001	Hills Publications	7134019
002-678-980-4	DBMS	P002	Sunshine Publishers Lt d.	6548909
002-678-880-2	Call Away	P002	Sunshine Publishers Ltd.	6548909
004-765-409-5	UNIX	P003	Bright Publications	7678985
004-765-359-3	Coordinate Geometry	P003	Bright Publications	7678985
003-456-433-6	Introduction to German Language	P004	Paramount Publishing House	9254834
003-456-533-8	Learning French Language	P004	Paramount Publishing House	9254834

- What if you need to add a publisher that has not yet published a book or whose book you do not carry?
- What if you remove a book but it's the only book by some publisher?
- What if a publisher's information is updated but not in all places where that publisher appears?

Northeastern University Khoury College of Computer and Information Sciences

51

Before we start normalizing…

- Normalization is a part of relational theory and was first proposed by E. F. Codd in 1970.
- The process of normalization assumes that each relation (table) has a primary key.
- A table without a primary key is not considered to be in first normal form.
- Primary keys may be atomic or compound (composed of several attributes).

Northeastern University Khoury College of Computer and Information Sciences

52

Normal Forms in Plain English…

"The key, the whole key, and nothing but the key, so help me Codd."

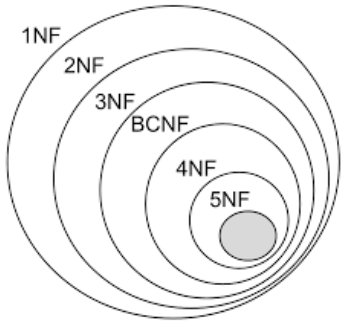
- "The key"
  - *(1NF): tables may not contain repeating groups, which prevent a table from having a primary key*
- "the whole key"
  - *(2NF): every attribute must depend on the entire primary key*
- *"and nothing but the key,"*
  - *(3NF): no dependencies on non-key attributes*

Northeastern University Khoury College of Computer and Information Sciences

53

Normal Forms are Progressive

- Normal forms are progressive, and a higher normal form includes the requirements of a lower normal form.
- For example, for a relation (table) to be in 3NF, it must be in 2NF, and to be in 2NF it must be in 1NF.
- You cannot have a relation in 3NF unless it is also in 1NF and 2NF.
- So, start with the lowest normal form and work upwards.



54

First Normal Form

eid	name	college
4009	Durant	Khoury
3002	Gatterbauer	Khoury
8883	Pflueger	COE, CPS
9994	Spitulnik	Bouve, Khoury
9973	Beauchamp	CSSH
7472	Shah	Khoury
8872	Vitek	COS, Khoury
3311	Offenhuber	CSSH

- This table is in 1NF, if:
  - ❑ every attribute is atomic, *i.e.*, no multi-valued attributes
  - ❑ there are no repeating groups

eid	name
4009	Durant
3002	Gatterbauer
8883	Pflueger
9994	Spitulnik
9973	Beauchamp
7472	Shah
8872	Vitek
3311	Offenhuber

college	eid
Khoury	4009
Khoury	3002
COE	8883
CPS	8883
Bouve	9994
Khoury	9994
CSSH	9973
Khoury	7472
COS	8872
Khoury	8872
CSSH	3311

55

Decomposition is at the heart of normalization

- Normalizing a database means that tables are successively decomposed until all tables are in a desirable normal form at least in BCNF.
- Normalization is there a *recursive process*.
- During decomposition of a relation R, into relations $R_1, R_2, \ldots, R_n$ it must be ensured that each attribute $A_i \in R$ must be in at least one relation schema R_i so that no attributes are lost:
- Formally,

$$R = \cup_{i=1}^{n}(R_i)$$

- **Attribute preservation** property of decomposition.

Northeastern University Khoury College of Computer and Information Sciences

56

Rules for Decomposition

- For small databases, it may be possible for a database designer to decompose a relation schema through inspection.
- For large, complex database schemas, a direct decomposition is difficult as there can be multiple ways in which a relational schema can be decomposed.
- To determine whether a relation schema should be decomposed, and which relation schema may be better than another, a formal methodology is needed.

*Normal forms provide the guidelines and **function dependencies** provide the method by which to decompose.*

Northeastern University Khoury College of Computer and Information Sciences

57

Functional Dependencies

- Functional dependencies (*FD*) are integrity constraints based on the concept of keys.

Definition: Let X and Y be two arbitrary subsets of attributes of a relation schema R. An instance r of R satisfies the functional dependency $X \rightarrow Y$, if and only if for any two tuples $t_1, t_2 \in r \wedge t_1[X] = t_2[X]$, it must be true that $t_1[Y] = t_2[Y]$.

More simply Y is functionally dependent on X, if and only if each value of X in r is associated with it precisely one value of Y.

Northeastern University Khoury College of Computer and Information Sciences

58

Primary Key vs Superkeys

- The definition of FD does not require the set X to be minimal; the minimality condition must be satisfied only for X to be a primary key.

Definition: If $X \rightarrow Y$ holds on R, then X represents a primary key if X is minimal and Y represents the set of all other attributes in the relation.

Definition: X is a superkey of R, if there exists some subset $Z \subset X$ *s.t.* $Z \rightarrow Y$.

Northeastern University Khoury College of Computer and Information Sciences

59

Identifying Functional Dependencies

- What is the primary key?
- What are the functional dependencies?

ISBN	Price	Page_count	R_ID	City	Rating
001-987-760-9	25	800	A002	Atlanta	6
001-987-760-9	25	800	A008	Detroit	7
001-354-921-1	22	200	A006	Albany	7
002-678-980-4	35	860	A003	Los Angeles	5
002-678-980-4	35	860	A001	New York	2
002-678-980-4	35	860	A005	Juneau	7
004-765-409-5	26	550	A003	Los Angeles	4
004-765-359-3	40	650	A007	Austin	3
003-456-433-6	30	500	A010	Virginia Beach	5
001-987-650-5	35	450	A009	Seattle	8
002-678-880-2	25	400	A006	Albany	4
003-456-533-8	30	500	A004	Seattle	9

Northeastern University Khoury College of Computer and Information Sciences

60

What does an *FD* mean?

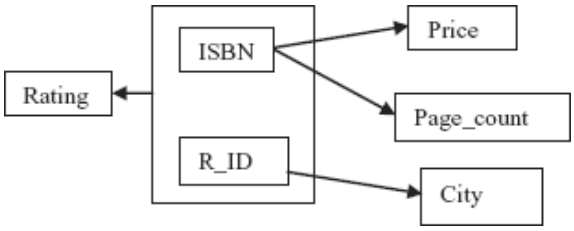
- An FD must hold for all possible and valid instances r of a relation R, regardless whether those instances are already present in R.
- For example, in the table on the previous slide, the FD $Page\_count \rightarrow Price$ holds but it is to be expected that there are books with the same page count that have different prices.
- So, the FD should not be included in the set of valid functional dependencies of R.

Northeastern University Khoury College of Computer and Information Sciences

61

Depicting Functional Dependencies

- Some database designers prefer to use functional dependency diagrams.



62

Minimal Functional Dependencies

- The functional dependencies hold:

ISBN → Price
ISBN → Page_count
R_ID → City
{R_ID, ISBN} → City
{R_ID, ISBN} → {City, Rating}
{R_ID, ISBN} → ISBN
{R_ID, ISBN} → {R_ID, ISBN, City, Rating}

- To reduce complexity of constraint checking we need $FD' \subset FD$, *s.t.* all functional dependencies that hold in FD also hold in FD'.

63

Trivial Dependencies

- A functional dependency is said to be **trivial** if it is satisfied by all relations.

Definition: A function dependency $A \rightarrow B$ is trivial, if and only if $B \subseteq A$.

- For example, $ISBN \rightarrow ISBN$ and $\{R\_ID, ISBN\} \rightarrow ISBN$ are trivial dependencies.
- All functional dependencies that are not trivial are non-trivial.

Northeastern University Khoury College of Computer and Information Sciences

64

Armstrong's Axioms and Rules

Armstrong's Axioms

Reflexivity Rule: If $B \subseteq A$, then $A \rightarrow B$ holds.

Augmentation Rule: If $A \rightarrow B$, then $AC \rightarrow BC$ holds.

Transitivity Rule: If both $A \rightarrow B$ and $B \rightarrow C$, then $A \rightarrow C$ holds.

Rules Derived from Axioms

Decomposition Rule: If $A \rightarrow BC$, then $A \rightarrow B$ and $A \rightarrow C$ holds.

Union Rule: If $A \rightarrow B$ and $A \rightarrow C$ holds, then $A \rightarrow BC$ holds.

Pseudotransitivity Rule: If both $A \rightarrow B$ and $BC \rightarrow D$ holds, then $AC \rightarrow D$ holds.

Northeastern University Khoury College of Computer and Information Sciences

65

Definition: 1NF

Definition: A relation R is in 1NF if all attributes of R are single valued.

While multi-valued attributes are allowed in an object-oriented conceptual model, they should be avoided in a logical model.

Many modern non-relational databases support multi-valued attributes and therefore should not be "normalized" out in the conceptual model.

While relational databases can support multi-valued attributes through separated text (*e.g.*, *valueA, valueB*), they should be avoided as substring search is expense.

Northeastern University Khoury College of Computer and Information Sciences

66

Definition: 2NF

Definition: A relation R is in 2NF if it is in 1NF and if all non-prime attributes of R are fully functionally dependent in the primary key.

An attribute Y of a relation schema R is **fully functional dependent** on attribute X $(X \rightarrow Y)$, if there $\nexists A : A \subset X$ such that $A \rightarrow Y$.

This means that if you remove any attribute from X that it can no longer uniquely determine Y.

Likewise, a dependency $X \rightarrow Y$ is a **partial dependency** if $\{\exists A : A \subset X \wedge A \rightarrow Y\}$.

Northeastern University Khoury College of Computer and Information Sciences

67

Example: 2NF Decomposition

ISBN	Price	Page_count	P_ID	R_ID	Rating
001-987-760-9	25	800	P001	A002	6
001-987-760-9	25	800	P001	A008	7
001-354-921-1	22	200	P001	A006	7
002-678-980-4	35	860	P002	A001	2
002-678-980-4	35	860	P002	A003	5
002-678-980-4	35	860	P002	A005	7
004-765-409-5	26	550	P003	A003	4
004-765-359-3	40	650	P003	A007	3
003-456-433-6	30	500	P004	A010	5
001-987-650-5	35	450	P001	A009	8
002-678-880-2	25	400	P002	A006	4
003-456-533-8	30	500	P004	A004	9

BOOK

ISBN	Price	Page_count	P_ID
001-987-760-9	25	800	P001
001-354-921-1	22	200	P001
002-678-980-4	35	860	P002
004-765-409-5	26	550	P003
004-765-359-3	40	650	P003
003-456-433-6	30	500	P004
001-987-650-5	35	450	P001
002-678-880-2	25	400	P002
003-456-533-8	30	500	P004

REVIEW

ISBN	R_ID	Rating
001-987-760-9	A002	6
001-987-760-9	A008	7
001-354-921-1	A006	7
002-678-980-4	A001	2
002-678-980-4	A003	5
002-678-980-4	A005	7
004-765-409-5	A003	4
004-765-359-3	A007	3
003-456-433-6	A010	5
001-987-650-5	A009	8
002-678-880-2	A006	4
003-456-533-8	A004	9

BOOK{ISBN,Price, Page_count, P_ID}
REVIEW {ISBN,R_ID, Rating}

Neither BOOK nor REVIEW contain any partial dependencies. Each of the non-prime attributes of each relation is fully functional dependent on their primary keys ISBN and (ISBN, R_ID).

Northeastern University Khoury College of Computer and Information Sciences

68

Definition: BCNF

Definition: A relation R is in BCNF if it is in 2NF and if and only if for every functional dependency F: $X \rightarrow Y$, where X is a subset of attributes of R at least one the following is true:

- $X \rightarrow Y$ is a trivial functional dependency, *i.e.*, $Y \subseteq X$
- X is a superkey for R

In simple terms, BCNF means, that for a dependency $X \rightarrow Y$, X cannot be a **non-prime attribute**, if Y is a **prime attribute**.

Northeastern University Khoury College of Computer and Information Sciences

69

Example: BCNF Decomposition

ISBN	Price	Page_count	P_ID	R_ID	Rating
001-987-760-9	25	800	P001	A002	6
001-987-760-9	25	800	P001	A008	7
001-354-921-1	22	200	P001	A006	7
002-678-980-4	35	860	P002	A001	2
002-678-980-4	35	860	P002	A003	5
002-678-980-4	35	860	P002	A005	7
004-765-409-5	26	550	P003	A003	4
004-765-359-3	40	650	P003	A007	3
003-456-433-6	30	500	P004	A010	5
001-987-650-5	35	450	P001	A009	8
002-678-880-2	25	400	P002	A006	4
003-456-533-8	30	500	P004	A004	9

BOOK

ISBN	Price	Page_count	P_ID
001-987-760-9	25	800	P001
001-354-921-1	22	200	P001
002-678-980-4	35	860	P002
004-765-409-5	26	550	P003
004-765-359-3	40	650	P003
003-456-433-6	30	500	P004
001-987-650-5	35	450	P001
002-678-880-2	25	400	P002
003-456-533-8	30	500	P004

REVIEW

ISBN	R_ID	Rating
001-987-760-9	A002	6
001-987-760-9	A008	7
001-354-921-1	A006	7
002-678-980-4	A001	2
002-678-980-4	A003	5
002-678-980-4	A005	7
004-765-409-5	A003	4
004-765-359-3	A007	3
003-456-433-6	A010	5
001-987-650-5	A009	8
002-678-880-2	A006	4
003-456-533-8	A004	9

BOOK{ISBN,Price, Page_count, P_ID}
REVIEW {ISBN,R_ID, Rating}

Neither BOOK nor REVIEW contain any partial dependencies. Each of the non-prime attributes of each relation is fully functional dependent on their primary keys ISBN and (ISBN, R_ID).

Northeastern University Khoury College of Computer and Information Sciences

70

Evaluating a Table

studentID	subject	instructor
101	Java	Durant, Kathleen
101	C++	Annunziato, Jose
102	Java	Heisenberg, Fred
103	C#	Derbinsky, Nate
104	Java	Durant, Kathleen
102	Python	Derbinsky, Nate
103	Python	Derbinsky, Nate

- What is the primary key?
- What are the functional dependencies?
- In which normal form is the above table?
- Can it be further normalized?

Northeastern University Khoury College of Computer and Information Sciences

71

Evaluating Normal Forms

What is the highest normal form of the relation $R(A, B, C, D, E)$ with a set of functional dependencies as follows?

$$\{ B \rightarrow E, BC \rightarrow D, AC \rightarrow BE \}$$

72

Another BCNF Example

Consider the following relationship : **R (A,B,C,D)**

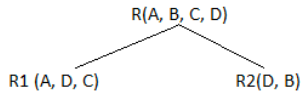
and following dependencies :

A -> BCD
BC -> AD
D -> B

Above relationship is already in 3rd NF. Keys are **A** and **BC**.

Hence, in the functional dependency, **A -> BCD**, A is the super key.
in second relation, **BC -> AD**, BC is also a key.
but in, **D -> B**, D is not a key.

Hence we can break our relationship R into two relationships **R1** and **R2**.



Breaking, table into two tables, one with A, D and C while the other with D and B.

From https://www.studytonight.com/dbms/boyce-codd-normal-form.php

73

Intractability of BCNF Verification

- Determining whether a database schema that is in 3NF is also in BCNF was shown to be *NP-Complete* by Beeri and Bernstein (1979).

Beeri, C., & Bernstein, P. A. (1979). Computational problems related to the design of normal form relational schemas. *ACM Transactions on Database Systems (TODS)*, *4*(1), 30-59.

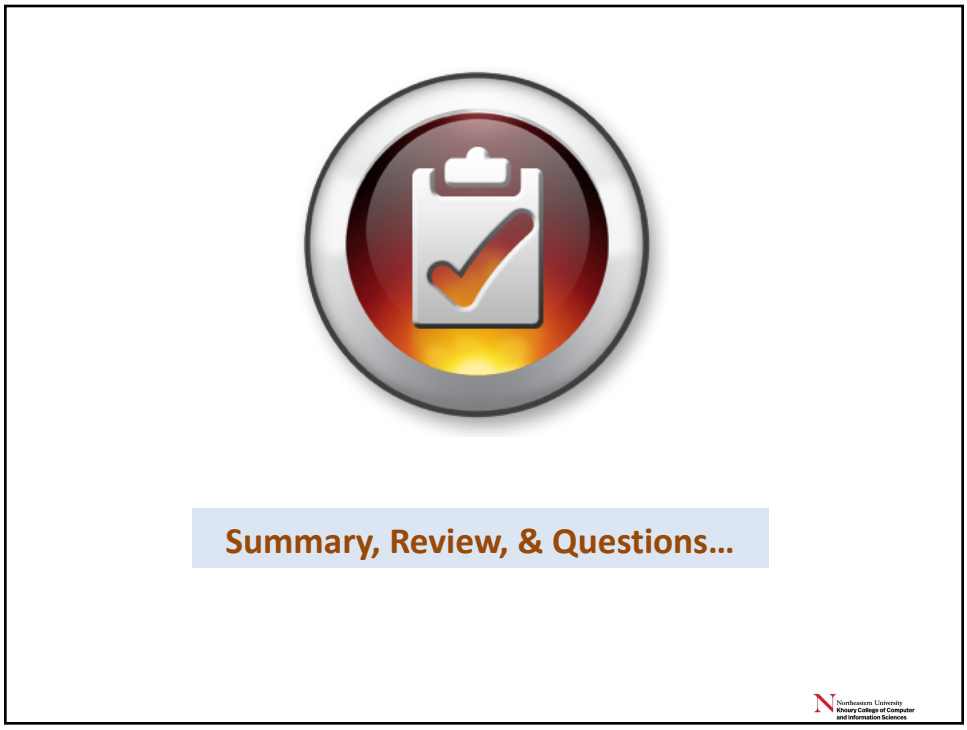
74

History

- In Date (2005), Chris Date states that "*since that definition predated Boyce and Codd's own definition by some three years, it seems to me that BCNF ought by rights to be called Heath Normal Form*."
- Perhaps this is, once again, *Stigler's law of eponymy* in action. Just think about Bayes' Theorem which should really be called the Price-Bayes Theorem.

Codd, E. F. (2002). A relational model of data for large shared data banks. In *Software pioneers* (pp. 263-294). Springer, Berlin, Heidelberg.
Codd, E. F. (1975). Recent Investigations in Relational Data Base Systems.
Heath, I. J. (1971, November). Unacceptable file operations in a relational data base.
In *Proceedings of the 1971 ACM SIGFIDET (now SIGMOD) Workshop on Data Description, Access and Control* (pp. 19-33).
Date, C. (2005). *Database in depth: relational theory for practitioners*. " O'Reilly Media, Inc.".

75


Summary, Review, & Questions…
Northeastern University
Khoury College of Computer and Information Sciences

76