Normalization

- “… evaluating and correcting table structures to minimize data redundancies, thereby reducing the likelihood of data anomalies.”
- Determinant – attribute value determines another attribute value
  - Do not confuse with derived attributes
  - Given the attribute value, we can get the attributes for the matching instance
- Dependent – attribute value is determined by another attribute
  - Do not confuse with derived attributes
  - Given the determinant value, these are the matching instance’s attributes

Normalization

- Notation:
  - Determinant → Dependent
  - Determinant → Dependent1, Dependent2, …
  - (Determinant1 + Determinant2 + …) → Dependent
  - (Determinant1 + Determinant2 + …) → Dependent1, Dependent2, …
- Examples
  - STU_NUM → STU_LNAME
  - STU_NUM → STU_LNAME, STU_FNAME, STU_GPA
  - (STU_NUM + STU_LNAME) → STU_GPA

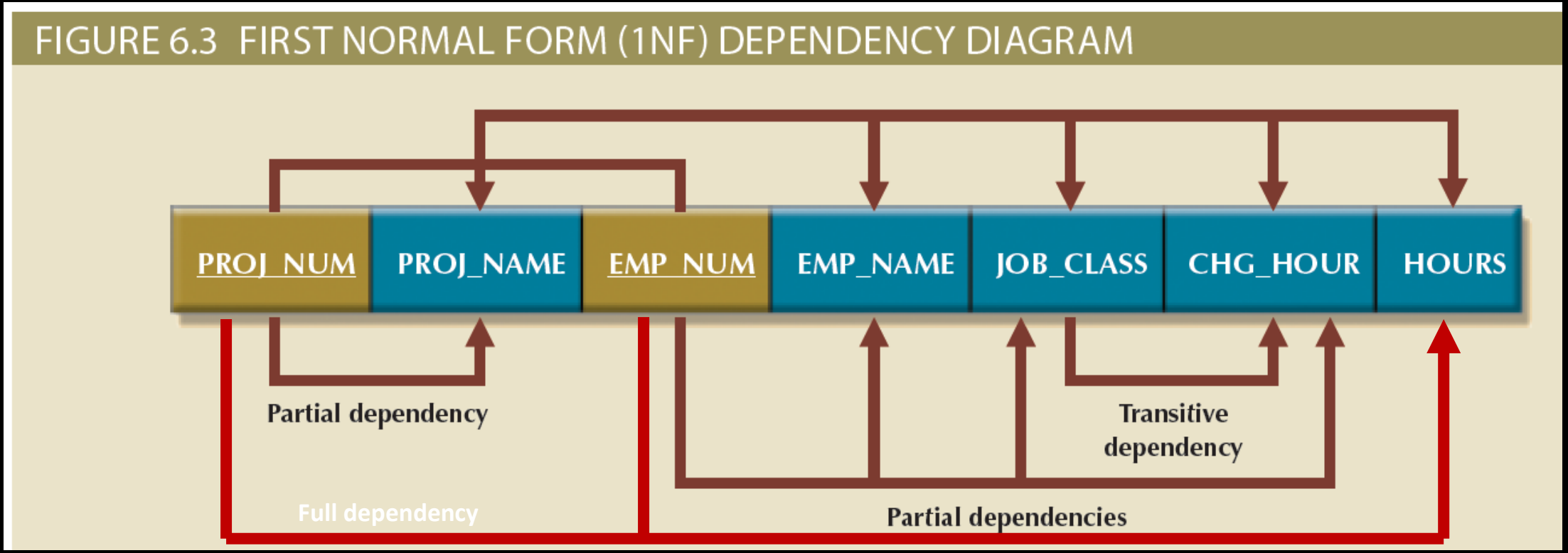
Normalization

- Dependencies describe the determinant-dependent relationship
  - Functional, full functional, partial, and transitive dependencies
- Functional dependence – value of one or more attributes determines the value of one or more other attributes
  - (STU_NUM + STU_LNAME) → STU_FNAME, STU_GPA
- Full functional dependence – entire collection of attributes in the determinant is necessary for the relationship
  - STU_NUM → STU_FNAME, STU_GPA

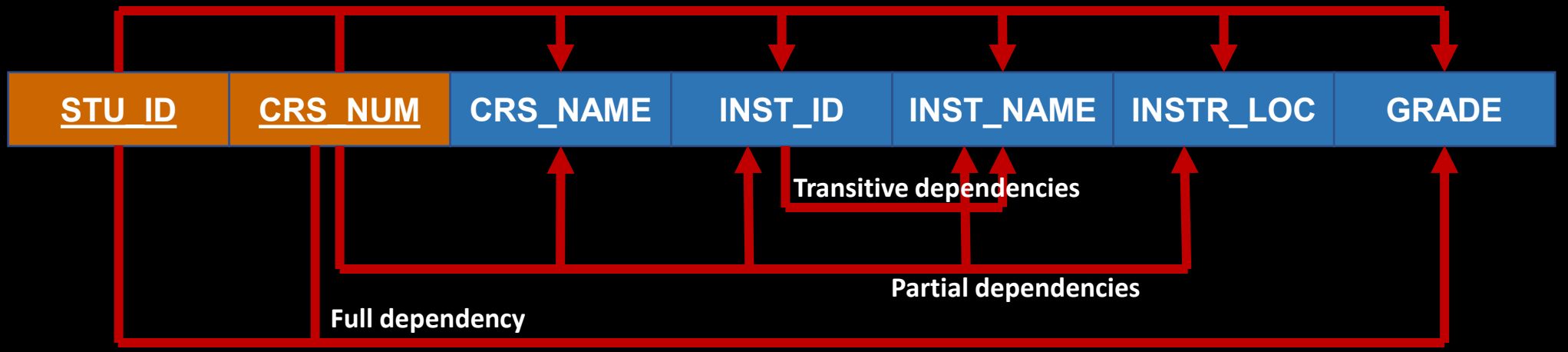
Normalization

- (PROJ_NUM + EMP_NUM) → PROJ_NAME, EMP_NAME, JOB_CLASS, CHG_HOUR, HOURS
- Partial dependence – the dependent is determinable by one of the determinants when there are multiple determinants
  - PROJ_NUM → PROJ_NAME
  - EMP_NUM → EMP_NAME, JOB_CLASS, CHG_HOUR
- Transitive dependence – the dependent can be determined by attribute(s) not associated with the determinant attribute(s)
  - JOB_CLASS → CHG_HOUR

Normalization



Normalization



Normalization

NORMAL FORMS		
NORMAL FORM	**CHARACTERISTIC**	**SECTION**
First normal form (1NF)	Table format, no repeating groups, and PK identified	6-3a
Second normal form (2NF)	1NF and no partial dependencies	6-3b
Third normal form (3NF)	2NF and no transitive dependencies	6-3c
Boyce-Codd normal form (BCNF)	Every determinant is a candidate key (special case of 3NF)	6-6a
Fourth normal form (4NF)	3NF and no independent multivalued dependencies	6-6b

First Normal Form

Table format, no repeating groups, PK and dependencies identified

1. Organize the data as a table
2. Eliminate repeating groups
3. Identify the primary key
4. Identify all the dependencies

ID	FirstName	LastName	Phone1	Phone2
123	Robert	Ingram	555-861-2025	
456	Jane	Wright	555-403-1659	555-776-4100
789	Maria	Fernandez	555-808-9633	

ID	FirstName	LastName	Phone
123	Robert	Ingram	555-861-2025
456	Jane	Wright	555-403-1659 555-776-4100
789	Maria	LastName	555-808-9633

ID	FirstName	Surname	Phone
123	Robert	Ingram	555-861-2025
456	Jane	Wright	555-403-1659
456	Jane	Wright	555-776-4100
789	Maria	Fernandez	555-808-9633

First Normal Form

Table name: RPT_FORMAT **Database name: Ch06_ConstructCo**

PROJ_NUM	PROJECT_NAME	EMP_NUMBER	EMP_NAME	JOB_CLASS	CHARGE_HOUR	HOURS_BILLED
15	Evergreen	103,101,105, 106, 102	June E. Arbough, John G. News, Alice K. Johnson *, William Smithfield, David H. Senior	Elec. Engineer, Database Designer, Database Designer, Programmer, System Analyst	85.5, 105., 105., 35.75, 98.75	23.8, 19.4, 35.7, 12.6, 23.8
18	Amber Wave	114, 118, 104, 112	Annelise Jones, James J. Frommer, Anne K. Ramoras *, Darlene M. Smithson	Applications Designer, General Support, Systems Analyst, DSS Analyst	48.1, 18.36, 96.75, 45.95	25.6, 45.3, 32.4, 45.
22	Rolling Tide	105, 104, 113, 111, 106	Alice K. Johnson, Anne K. Ramoras, Delbert K. Joenbrood *, Geoff B. Wabash, William Smithfield	DB Designer, Systems Analyst, Applications Designer, Clerical Support, Programmer	105., 96.75, 48.1, 26.87, 35.75	65.7, 48.4, 23.6, 22., 12.8
25	Star Light	107, 115, 101, 114, 108, 118, 112	Maria D. Alonzo, Travis B. Bawangi, John G. News *, Annelise Jones, Ralph B. Washington, James J. Frommer, Darlene M. Smithson	Programmer, Systems Analyst, Database Design, Applications Designer, Systems Analyst, General Support, DSS Analyst	35.75, 96.75, 105., 48.1, 96.75, 18.36, 45.95	25.6, 45.8, 56.3, 33.1, 23.6, 30.5, 41.4

First Normal Form

Table name: DATA_ORG_1NF **Database name: Ch06_ConstructCo**

PROJ_NUM	PROJ_NAME	EMP_NUM	EMP_NAME	JOB_CLASS	CHG_HOUR	HOURS
15	Evergreen	103	June E. Arbough	Elect. Engineer	84.50	23.8
15	Evergreen	101	John G. News	Database Designer	105.00	19.4
15	Evergreen	105	Alice K. Johnson *	Database Designer	105.00	35.7
15	Evergreen	106	William Smithfield	Programmer	35.75	12.6
15	Evergreen	102	David H. Senior	Systems Analyst	96.75	23.8
18	Amber Wave	114	Annelise Jones	Applications Designer	48.10	24.6
18	Amber Wave	118	James J. Frommer	General Support	18.36	45.3
18	Amber Wave	104	Anne K. Ramoras *	Systems Analyst	96.75	32.4
18	Amber Wave	112	Darlene M. Smithson	DSS Analyst	45.95	44.0
22	Rolling Tide	105	Alice K. Johnson	Database Designer	105.00	64.7
22	Rolling Tide	104	Anne K. Ramoras	Systems Analyst	96.75	48.4
22	Rolling Tide	113	Delbert K. Joenbrood *	Applications Designer	48.10	23.6
22	Rolling Tide	111	Geoff B. Wabash	Clerical Support	26.87	22.0
22	Rolling Tide	106	William Smithfield	Programmer	35.75	12.8
25	Starflight	107	Maria D. Alonzo	Programmer	35.75	24.6

First Normal Form

- Repeating data != repeating groups
  - The phone numbers are a repeating group
  - ID and names are repeating data
- Multivalued attributes are often considered a repeating group
  - A list of values in one attribute field
  - Columns that are essentially the same type of information, e.g., multiple phones
- Resist the desire to decompose the table because we do that for 2NF

ID	First Name	Surname	Phone
123	Robert	Ingram	555-861-2025
456	Jane	Wright	555-403-1659
456	Jane	Wright	555-776-4100
789	Maria	Fernandez	555-808-9633

- What is the PK now that we removed the repeating group?

First Normal Form

- PROJECT(**PROJ_NUM**, **EMP_NUM**, PROJ_NAME, EMP_NAME, JOB_CLASS, CHG_HOURS, HOURS)
- Full: (PROJ_NUM + EMP_NUM) → HOURS
- Partial: PROJ_NUM → PROJ_NAME
- EMP_NUM → EMP_NAME, JOB_CLASS, CHG_HOUR
- Transitive: JOB_CLASS → CHG_HOUR

PROJ_NUM	PROJ_NAME	EMP_NUM	EMP_NAME	JOB_CLASS	CHG_HOUR	HOURS
15	Evergreen	103	June E. Arbough	Elect. Engineer	84.50	23.8
15	Evergreen	101	John G. News	Database Designer	105.00	19.4
15	Evergreen	105	Alice K. Johnson *	Database Designer	105.00	35.7
15	Evergreen	106	William Smithfield	Programmer	35.75	12.6
15	Evergreen	102	David H. Senior	Systems Analyst	96.75	23.8
18	Amber Wave	114	Annelise Jones	Applications Designer	48.10	24.6

Second Normal Form

- 1NF and no partial dependencies

1. Convert into 1NF
2. Make new tables to eliminate partial dependencies
3. Reassign corresponding dependent attributes

PROJECT(**PROJ_NUM**, **EMP_NUM**, PROJ_NAME, EMP_NAME, JOB_CLASS, CHG_HOURS, HOURS)

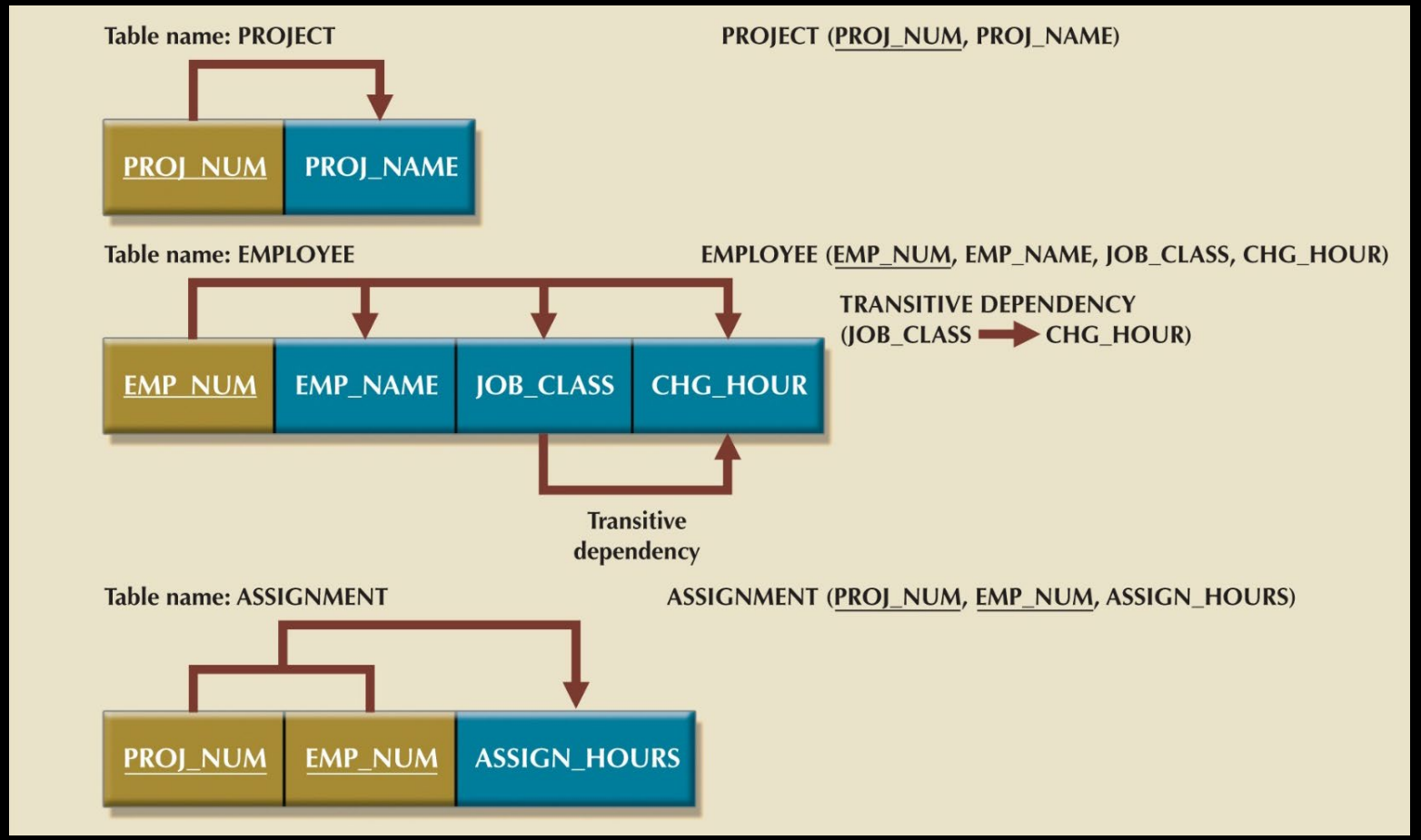
Full: (PROJ_NUM + EMP_NUM) → HOURS

Partial: PROJ_NUM → PROJ_NAME

EMP_NUM → EMP_NAME, JOB_CLASS, CHG_HOUR

Transitive: JOB_CLASS → CHG_HOUR

Second Normal Form



Second Normal Form

Partial: PROJ_NUM → PROJ_NAME
EMP_NUM → EMP_NAME, JOB_CLASS, CHG_HOUR

PROJECT(**PROJ_NUM**, PROJ_NAME)
Full: PROJ_NUM → PROJ_NAME

EMPLOYEE(**EMP_NUM**, EMP_NAME, JOB_CLASS, CHG_HOUR)
Full: EMP_NUM → EMP_NAME, JOB_CLASS, CHG_HOUR
Transitive: JOB_CLASS → CHG_HOUR

ASSIGNMENT(***PROJ_NUM***, ***EMP_NUM***, ASSIGN_HOURS)
Full: PROJ_NUM + EMP_NUM → HOURS

Third Normal Form

- 2NF and no transitive dependencies

1. Convert to 2NF
2. Make new tables to eliminate transitive dependencies
3. Reassign corresponding dependent attributes

EMPLOYEE(**EMP_NUM**, EMP_NAME, JOB_CLASS, CHG_HOUR)

Full: EMP_NUM → EMP_NAME, JOB_CLASS, CHG_HOUR

Transitive: JOB_CLASS → CHG_HOUR

Third Normal Form

PROJECT(**PROJ_NUM**, PROJ_NAME)
Full: PROJ_NUM → PROJ_NAME

EMPLOYEE(**EMP_NUM**, EMP_NAME, ***JOB_CLASS***)
Full: EMP_NUM → EMP_NAME, JOB_CLASS

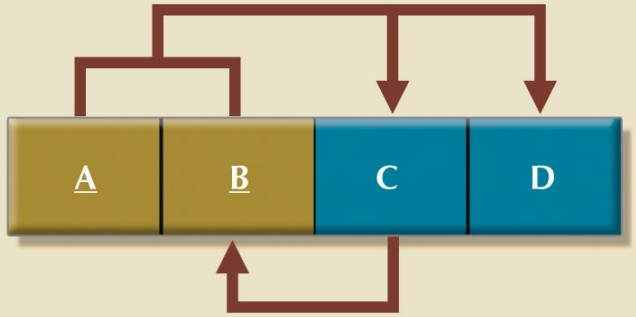
JOB(**JOB_CLASS**, CHG_HOUR)
Full: JOB_CLASS → CHG_HOUR

ASSIGNMENT(***PROJ_NUM***, ***EMP_NUM***, ASSIGN_HOURS)
Full: (PROJ_NUM + EMP_NUM) → HOURS

Boyce-Codd Normal Form

- Every determinant is a candidate key (special case of 3NF)

FIGURE 6.8 A TABLE THAT IS IN 3NF BUT NOT IN BCNF



- There are no partial or transitive dependencies in this example!
- (A + B) → C, D The chosen primary key determines the others
- C → B But C can determine B, and B can determine C
- (A + C) → B, D Therefore (A + C) can also be a primary key

Boyce-Codd Normal Form

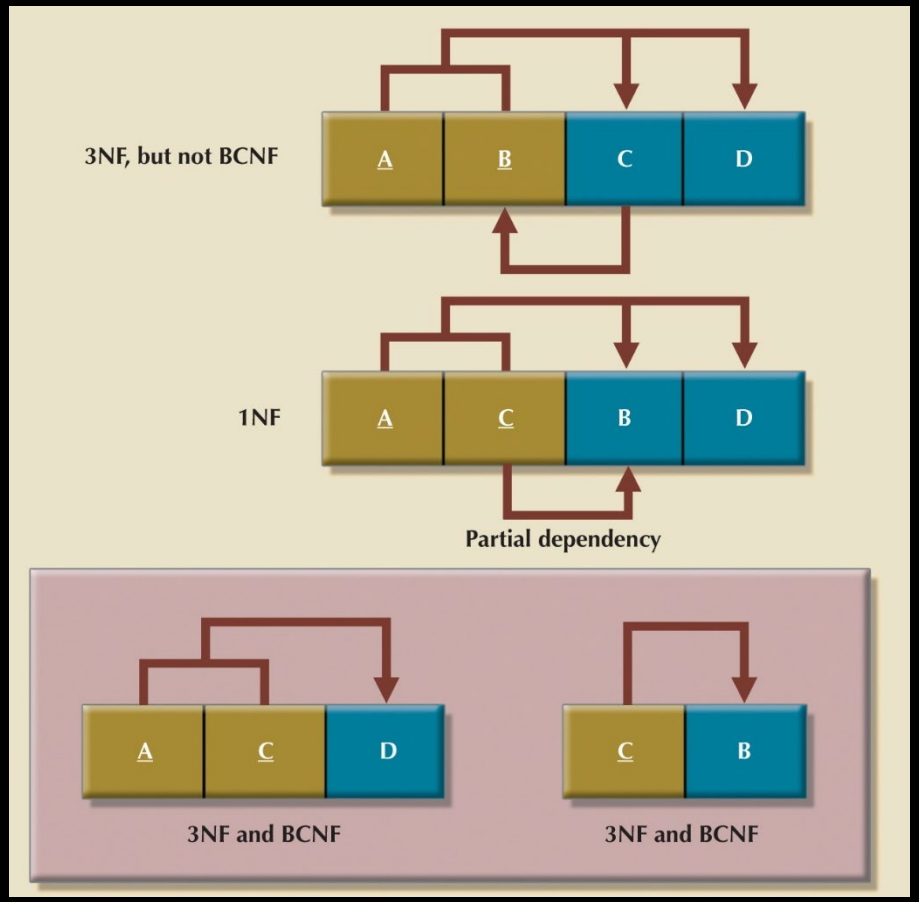
- If we use A and C as the PK
- (A + C) → B, D becomes the partial dependency of C → B
- To be in 2NF and 3NF, the partial dependency needs to be resolved

TABLE1(**A**, ***C***, D)

Full: A + C → D

TABLE2(**C**, B)

Full: C → B

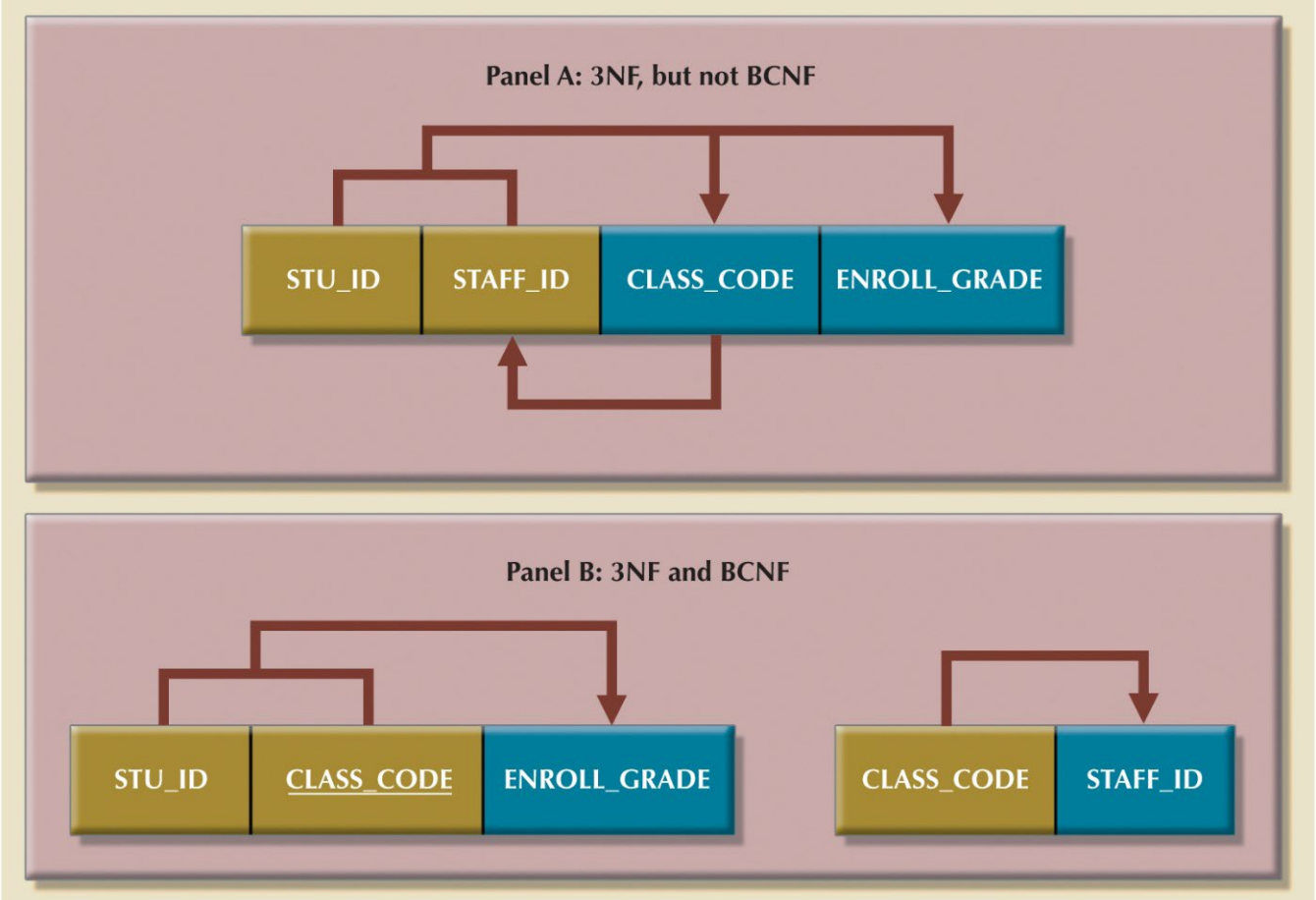


Boyce-Codd Normal Form

SAMPLE DATA FOR A BCNF CONVERSION			
STU_ID	STAFF_ID	CLASS_CODE	ENROLL_GRADE
125	25	21334	A
125	20	32456	C
135	20	28458	B
144	25	27563	C
144	20	32456	B

- (STU_ID + STAFF_ID) → CLASS_CODE, ENROLL_GRADE
- CLASS_CODE → STAFF_ID because each class only has one staff member assigned to

Boyce-Codd Normal Form



Fourth Normal Form

- 3NF and no independent multivalued dependencies
  - Multivalued dependency
    - Independent facts that vary separately for the same key
    - One key determines multiple values of attribute, and those values are independent of another multivalued attribute
  - Not to be confused with multivalued attributes
    - 1NF eliminates lists inside attributes (several degrees, several phones, etc.).
    - 4NF eliminates unrelated lists about the same entity
  - Rule: “A table should not store two or more independent multivalued facts about the same entity.”
    - All attributes must be dependent on the PK, but must also be independent of each other
    - No row may contain two or more multivalued facts about the entity

Fourth Normal Form

- An employee has many volunteer services
- An employee has many assignments
- Tables V1 and V2 as solutions:
  - May possibly have many nulls
  - Has no viable PK because of nulls
- Table V3 as solutions:
  - Has a PK, but requires all the attributes
  - Technically in 3NF, but contains redundancies

Table name: VOLUNTEER_V1

EMP_NUM	ORG_CODE	ASSIGN_NUM
10123	RC	1
10123	UW	3
10123		4

Table name: VOLUNTEER_V2

EMP_NUM	ORG_CODE	ASSIGN_NUM
10123	RC	
10123	UW	
10123		1
10123		3
10123		4

Table name: VOLUNTEER_V3

EMP_NUM	ORG_CODE	ASSIGN_NUM
10123	RC	1
10123	RC	3
10123	UW	4

Fourth Normal Form

Table name: PROJECT

PROJ_CODE	PROJ_NAME	PROJ_BUDGET
1	BeThere	1023245.00
2	BlueMoon	20198608.00
3	GreenThumb	3234456.00
4	GoFast	5674000.00
5	GoSlow	1002500.00

Table name: ASSIGNMENT

ASSIGN_NUM	EMP_NUM	PROJ_CODE
1	10123	1
2	10121	2
3	10123	3
4	10123	4
5	10121	1
6	10124	2
7	10124	3
8	10124	5

Table name: EMPLOYEE

EMP_NUM	EMP_LNAME
10121	Rogers
10122	O'Leery
10123	Panera
10124	Johnson

Table name: ORGANIZATION

ORG_CODE	ORG_NAME
RC	Red Cross
UW	United Way
WF	Wildlife Fund

Table name: SERVICE_V1

EMP_NUM	ORG_CODE
10123	RC
10123	UW
10123	WF

Fourth Normal Form

- PersonSkillLanguage(ID, Skill, Language)
  - A person can have many skills and speak many languages
  - Skills and languages are unrelated to each other
- Solution:
  - PersonSkill(**ID**, Skill)
  - PersonLanguage(**ID**, Language)

Fifth Normal Form

- “Jason, why do you want to question reality and existence?”
- BCNF and 4NF and it cannot be decomposed into smaller tables without losing information
  - Break it down into the smallest possible tables while ensuring you can reconstruct the original table without any loss and have no unnecessary repetition of data.
- Intuition
  - 4NF: Do not mix independent facts
  - 5NF: Do not store facts that only exist because of a specific combination of 3 things
- Appears when
  - Relationships involve three or more entities
  - No single pairwise relationship fully captures the information

Fifth Normal Form

- Project-Join Normal Form
  - Projection – splitting a table
  - Join – reconstruct the table
  - Every valid join dependency is a consequence of the candidate keys.
  - If a table can only be correctly reconstructed by joining three or more tables, and that rule is not enforced by keys, then it is not in 5NF.
- Dependencies
  - Functional dependency: determined by keys
  - Multivalued dependency: determined by independent facts
  - Join dependency: determined by how relations must be recombined

Fifth Normal Form

- A join dependency exists when a table can be losslessly reconstructed by joining multiple projections of that table, and the need for that reconstruction is not captured by functional or multivalued dependencies alone.
- A join dependency describes a rule where a relation’s meaning is preserved only when it is reconstructed by joining two or more smaller relations.
- A join dependency exists when the correctness of a table depends on how multiple tables are joined together.

Fifth Normal Form

- SupplierPartProject(Supplier, Part, Project)
  - A supplier can supply a part — SupplierPart(Supplier, Part)
  - A part can be used in a project — PartProject(Part, Project)
  - A supplier can work on a project — SupplierProject(Supplier, Project)
  - But not every combination is valid
    - A supplier supplies a part, but does not work on the project
- “5NF is about relationships that only make sense when three things are true at the same time.”