5

Chapter 5

Normalization of Database Tables

Database Systems:
Design, Implementation, and Management,
Sixth Edition, Rob and Coronel

In this chapter, you will learn:

- What normalization is and what role it plays in the database design process
- About the normal forms 1NF, 2NF, 3NF, BCNF, and 4NF
- How normal forms can be transformed from lower normal forms to higher normal forms
- That normalization and ER modeling are used concurrently to produce a good database design
- That some situations require denormalization to generate information efficiently

Database Tables and Normalization

- Normalization
  - Process for evaluating and correcting table structures to minimize data redundancies
    - helps eliminate data anomalies
  - Works through a series of stages called normal forms:
    - Normal form (1NF)
    - Second normal form (2NF)
    - Third normal form (3NF)

Database Tables and Normalization (continued)

- 2NF is better than 1NF; 3NF is better than 2NF
- For most business database design purposes, 3NF is highest we need to go in the normalization process
- Highest level of normalization is not always most desirable

The Need for Normalization

- Example: company that manages building projects
  - Charges its clients by billing hours spent on each contract
  - Hourly billing rate is dependent on employee's position
  - Periodically, a report is generated that contains information displayed in Table 5.1

A Sample Report Layout

TABLE 5.1 A SAMPLE REPORT LAYOUT

PROJ. NUM.	PROJECT NAME	EMPLOYEE NUMBER	EMPLOYEE NAME	JOB CLASS.	CHG/HOUR	HOURS BILLED	TOTAL CHARGE
15	Evergreen	103	June E. Arbough	Elec. Engineer	$84.50	23.8	$2,011.10
		101	John G. News	Database Designer	$105.00	19.4	$2,037.00
		105	Alice K. Johnson *	Database Designer	$105.00	35.7	$3,748.50
		106	William Smithfield	Programmer	$35.75	12.6	$450.45
		102	David H. Senior	Systems Analyst	$96.75	23.8	$2,302.65
				Subtotal			**$10,549.70**
18	Amber Wave	114	Annelise Jones	Applications Designer	$48.10	24.6	$1,183.26
		118	James J. Frommer	General Support	$18.36	45.3	$831.71
		104	Anne K. Ramoras *	Systems Analyst	$96.75	32.4	$3,135.70
		112	Darlene M. Smithson	DSS Analyst	$45.95	44.0	$2,021.80
				Subtotal			**$7,171.47**
22	Rolling Tide	105	Alice K. Johnson	Database Designer	$105.00	64.7	$6,793.50
		104	Anne K. Ramoras	Systems Analyst	$96.75	48.4	$4,682.70
		113	Delbert K. Joenbrood *	Applications Designer	$48.10	23.6	$1,135.16
		111	Geoff B. Wabash	Clerical Support	$26.87	22.0	$591.14
		106	William Smithfield	Programmer	$35.75	12.8	$457.60
				Subtotal			**$13,660.10**
25	Starflight	107	Maria D. Alonzo	Programmer	$35.75	24.6	$879.45
		115	Travis B. Bawangi	Systems Analyst	$96.75	45.8	$4,431.15
		101	John G. News *	Database Designer	$105.00	56.3	$5,911.50
		114	Annelise Jones	Applications Designer	$48.10	33.1	$1,592.11
		108	Ralph B. Washington	Systems Analyst	$96.75	23.6	$2,283.30
		118	James J. Frommer	General Support	$18.36	30.5	$559.98
		112	Darlene M. Smithson	DSS Analyst	$45.95	41.4	$1,902.33
				Subtotal			**$17,559.82**
				Total			**$48,941.09**

* Indicates project leader

A Table in the Report Format

FIGURE 5.1 A TABLE IN THE REPORT FORMAT

Table name: RPT_FORMAT **Database name: Ch05_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
		101	John G. News	Database Designer	$105.00	19.4
		105	Alice K. Johnson *	Database Designer	$105.00	35.7
		106	William Smithfield	Programmer	$35.75	12.6
		102	David H. Senior	Systems Analyst	$96.75	23.8
18	Amber Wave	114	Annelise Jones	Applications Designer	$48.10	24.6
		118	James J. Frommer	General Support	$18.36	45.3
		104	Anne K. Ramoras *	Systems Analyst	$96.75	32.4
		112	Darlene M. Smithson	DSS Analyst	$45.95	44.0
22	Rolling Tide	105	Alice K. Johnson	Database Designer	$105.00	64.7
		104	Anne K. Ramoras	Systems Analyst	$96.75	48.4
		113	Delbert K. Joenbrood *	Applications Designer	$48.10	23.6
		111	Geoff B. Wabash	Clerical Support	$26.87	22.0
		106	William Smithfield	Programmer	$35.75	12.8
25	Starflight	107	Maria D. Alonzo	Programmer	$35.75	24.6
		115	Travis B. Bawangi	Systems Analyst	$96.75	45.8
		101	John G. News *	Database Designer	$105.00	56.3
		114	Annelise Jones	Applications Designer	$48.10	33.1
		108	Ralph B. Washington	Systems Analyst	$96.75	23.6
		118	James J. Frommer	General Support	$18.36	30.5
		112	Darlene M. Smithson	DSS Analyst	$45.95	41.4

The Need for Normalization (continued)

- Structure of data set in Figure 5.1 does not handle data very well
- The table structure *appears* to work; report is generated with ease
- Unfortunately, the report may yield different results, depending on what data anomaly has occurred

Conversion to First Normal Form

- **Repeating group**
  - Derives its name from the fact that a group of multiple (related) entries can exist for any *single* key attribute occurrence
- Relational table must not contain repeating groups
- Normalizing the table structure will reduce these data redundancies
- Normalization is three-step procedure

Step 1: Eliminate the Repeating Groups

- Present data in a tabular format, where each cell has a single value and there are no repeating groups
- Eliminate repeating groups by eliminating nulls, making sure that each repeating group attribute contains an appropriate data value

Data Organization: First Normal Form

FIGURE 5.2 DATA ORGANIZATION: FIRST NORMAL FORM

Table name: DATA_ORG_1NF Database name: Ch05_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
25	Starflight	115	Travis B. Bawangi	Systems Analyst	$96.75	45.8
25	Starflight	101	John G. News *	Database Designer	$105.00	56.3
25	Starflight	114	Annelise Jones	Applications Designer	$48.10	33.1
25	Starflight	108	Ralph B. Washington	Systems Analyst	$96.75	23.6
25	Starflight	118	James J. Frommer	General Support	$18.36	30.5
25	Starflight	112	Darlene M. Smithson	DSS Analyst	$45.95	41.4

Step 2: Identify the Primary Key

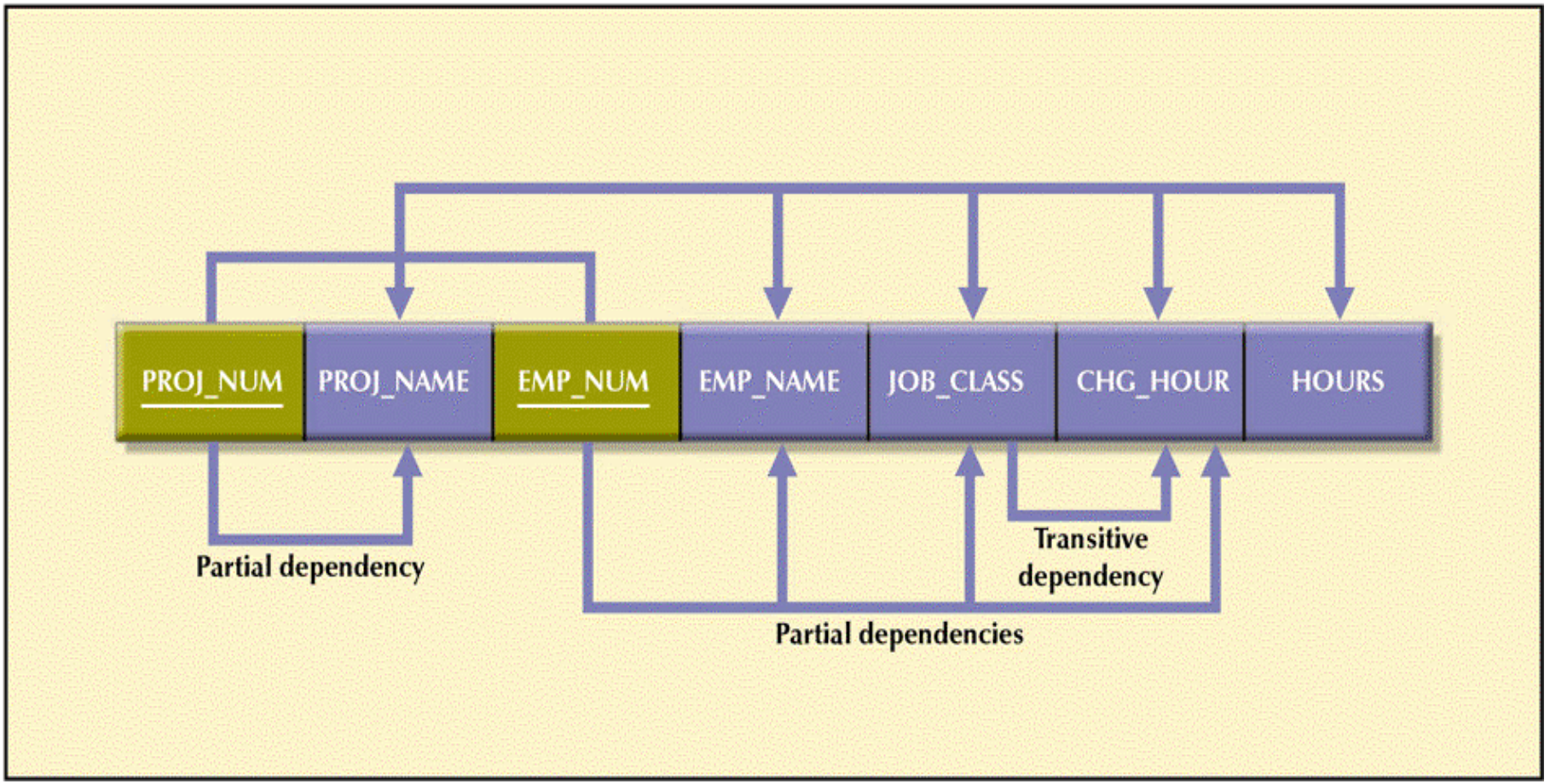
- Primary key must uniquely identify attribute value
- New key must be composed

Step 3: Identify all Dependencies

- Dependencies can be depicted with the help of a diagram
- Dependency diagram:
  - Depicts all dependencies found within a given table structure
  - Helpful in getting bird's-eye view of all relationships among a table's attributes
  - Use makes it much less likely that an important dependency will be overlooked

A Dependency Diagram: First Normal Form (1NF)

FIGURE 5.3 A DEPENDENCY DIAGRAM: FIRST NORMAL FORM (1NF)



First Normal Form

- Tabular format in which:
  - All key attributes are defined
  - There are no repeating groups in the table
  - All attributes are dependent on primary key
- All relational tables satisfy 1NF requirements
- Some tables contain partial dependencies
  - Dependencies based on only part of the primary key
  - Sometimes used for performance reasons, but should be used with caution
  - Still subject to data redundancies

Conversion to Second Normal Form

- Relational database design can be improved by converting the database into second normal form (2NF)
- Two steps

Step 1: Identify All Key Components

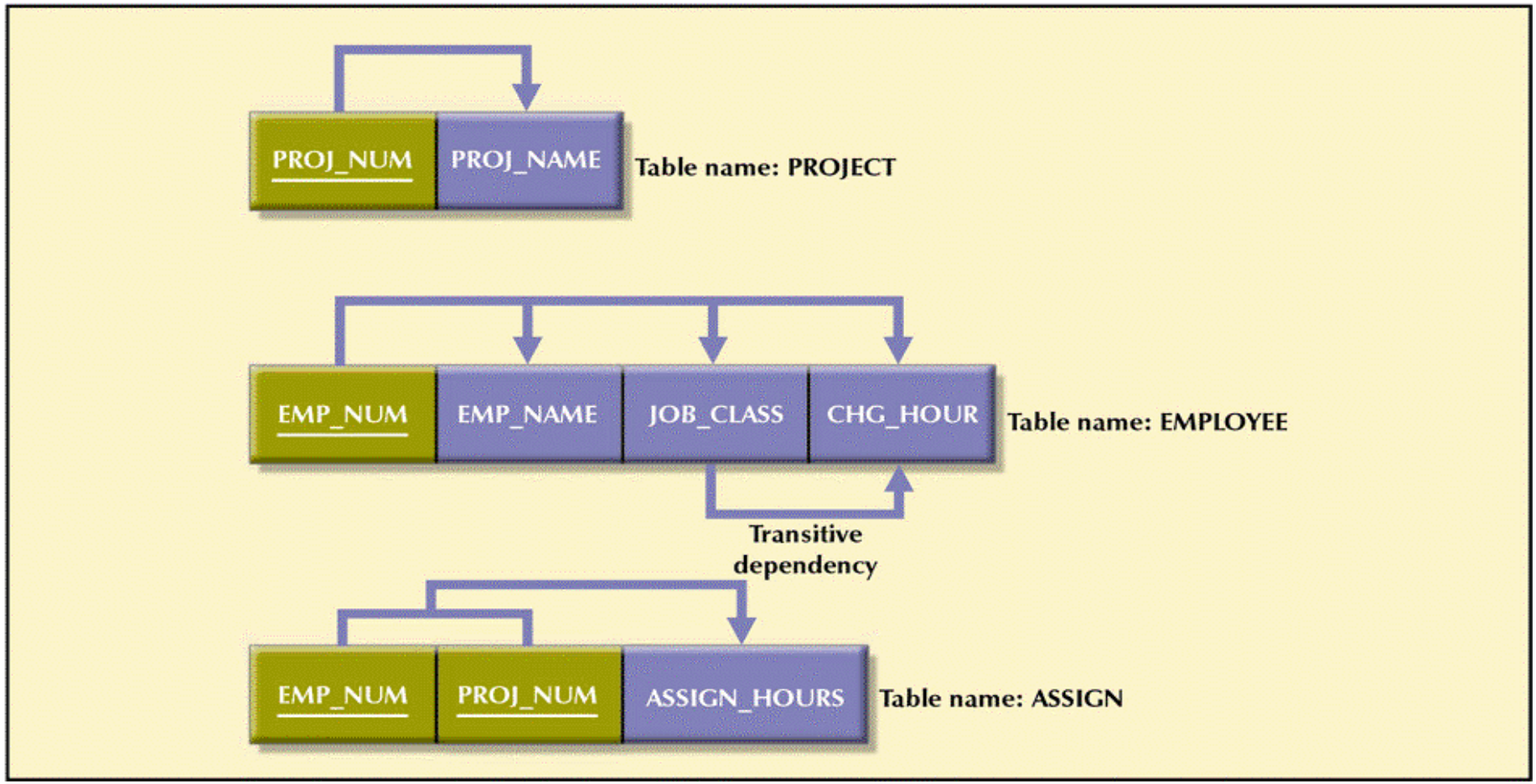
- Write each key component on separate line, and then write the original (composite) key on the last line
- Each component will become the key in a new table

Step 2: Identify the Dependent Attributes

- Determine which attributes are dependent on which other attributes
- At this point, most anomalies have been eliminated

Second Normal Form (2NF) Conversion Results

FIGURE 5.4 SECOND NORMAL FORM (2NF) CONVERSION RESULTS



Second Normal Form

- Table is in second normal form (2NF) if:
  - It is in 1NF and
  - It includes no partial dependencies:
    - No attribute is dependent on only a portion of the primary key

Conversion to Third Normal Form

- Data anomalies created are easily eliminated by completing three steps

Step 1: Identify Each New Determinant

- For every transitive dependency, write its determinant as a PK for a new table
  - **Determinant**
    - Any attribute whose value determines other values within a row

Step 2: Identify the Dependent Attributes

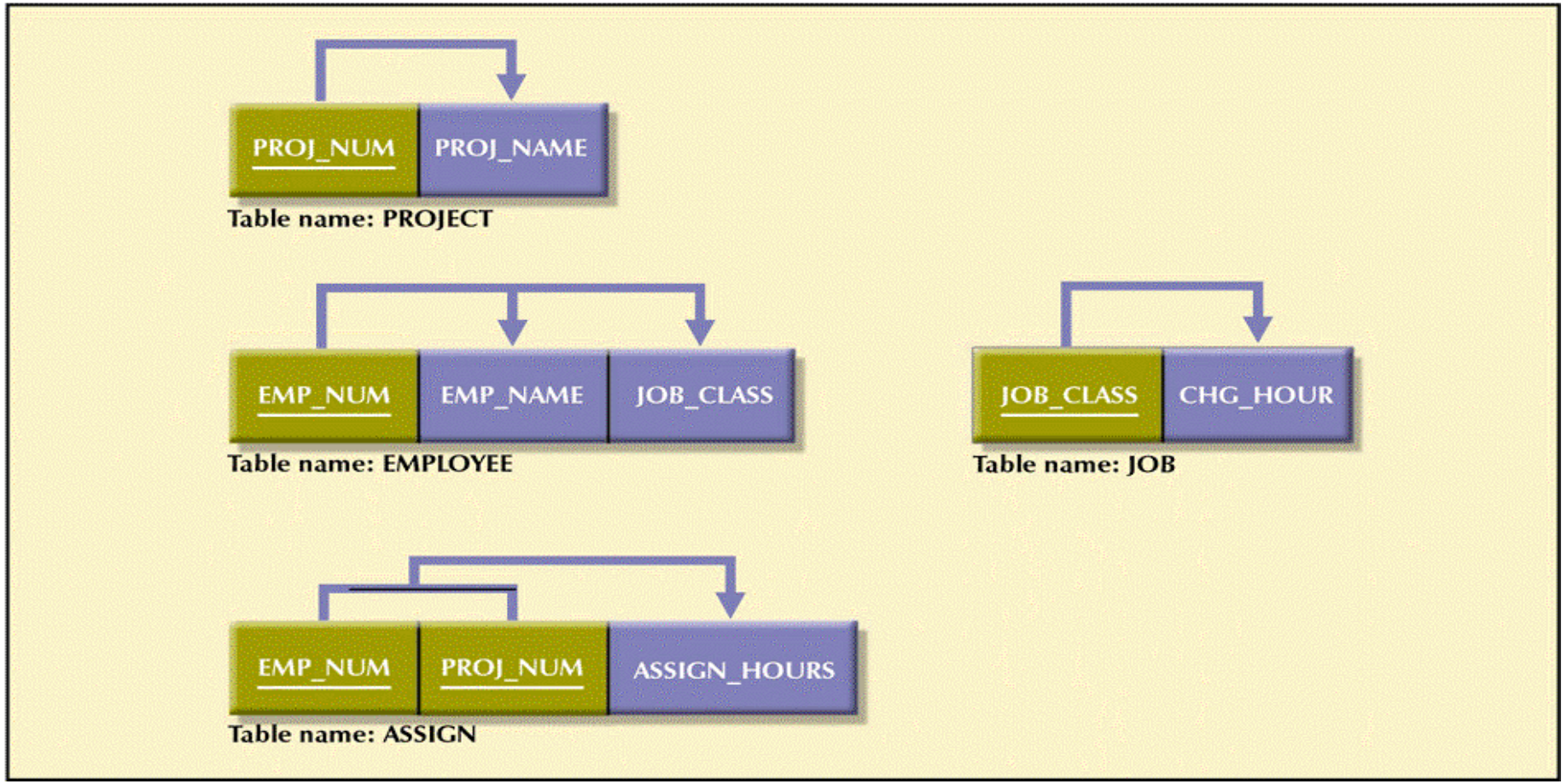
- Identify the attributes dependent on each determinant identified in Step 1 and identify the dependency
- Name the table to reflect its contents and function

Step 3: Remove the Dependent Attributes from Transitive Dependencies

- Eliminate all dependent attributes in transitive relationship(s) from each table that has such a transitive relationship
- Draw a new dependency diagram to show all tables defined in Steps 1–3
- Check new tables and modified tables from Step 3 to make sure that each has a determinant and does not contain inappropriate dependencies

Third Normal Form (3NF) Conversion Results

FIGURE 5.5 THIRD NORMAL FORM (3NF) CONVERSION RESULTS



Third Normal Form

- A table is in third normal form (3NF) if:
  - It is in 2NF and
  - It contains no transitive dependencies

Improving the Design

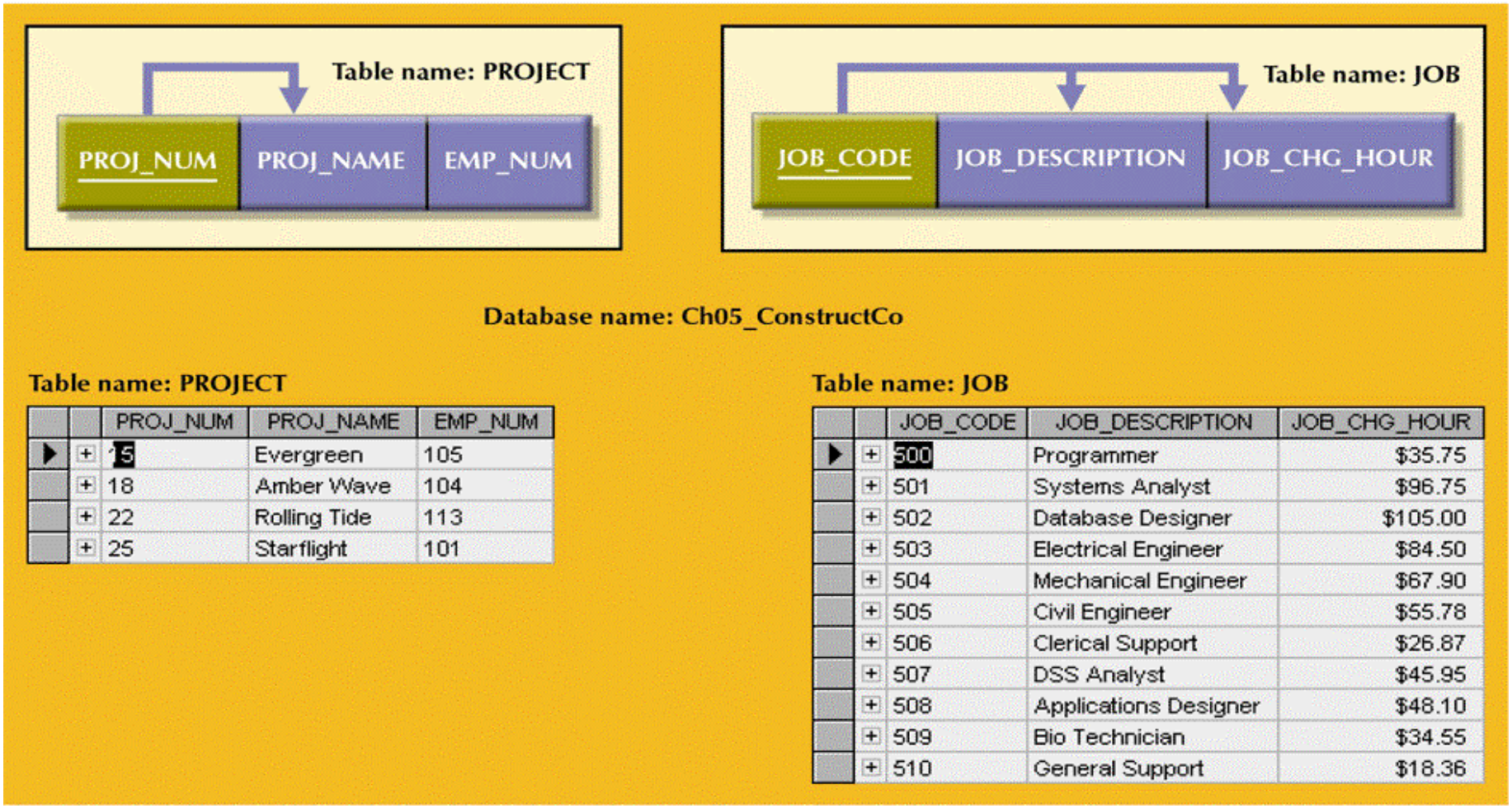
- Table structures are cleaned up to eliminate the troublesome initial partial and transitive dependencies
- Normalization cannot, by itself, be relied on to make good designs
- It is valuable because its use helps eliminate data redundancies

Improving the Design (continued)

- The following changes were made:
  - PK assignment
  - Naming conventions
  - Attribute atomicity
  - Adding attributes
  - Adding relationships
  - Refining PKs
  - Maintaining historical accuracy
  - Using derived attributes

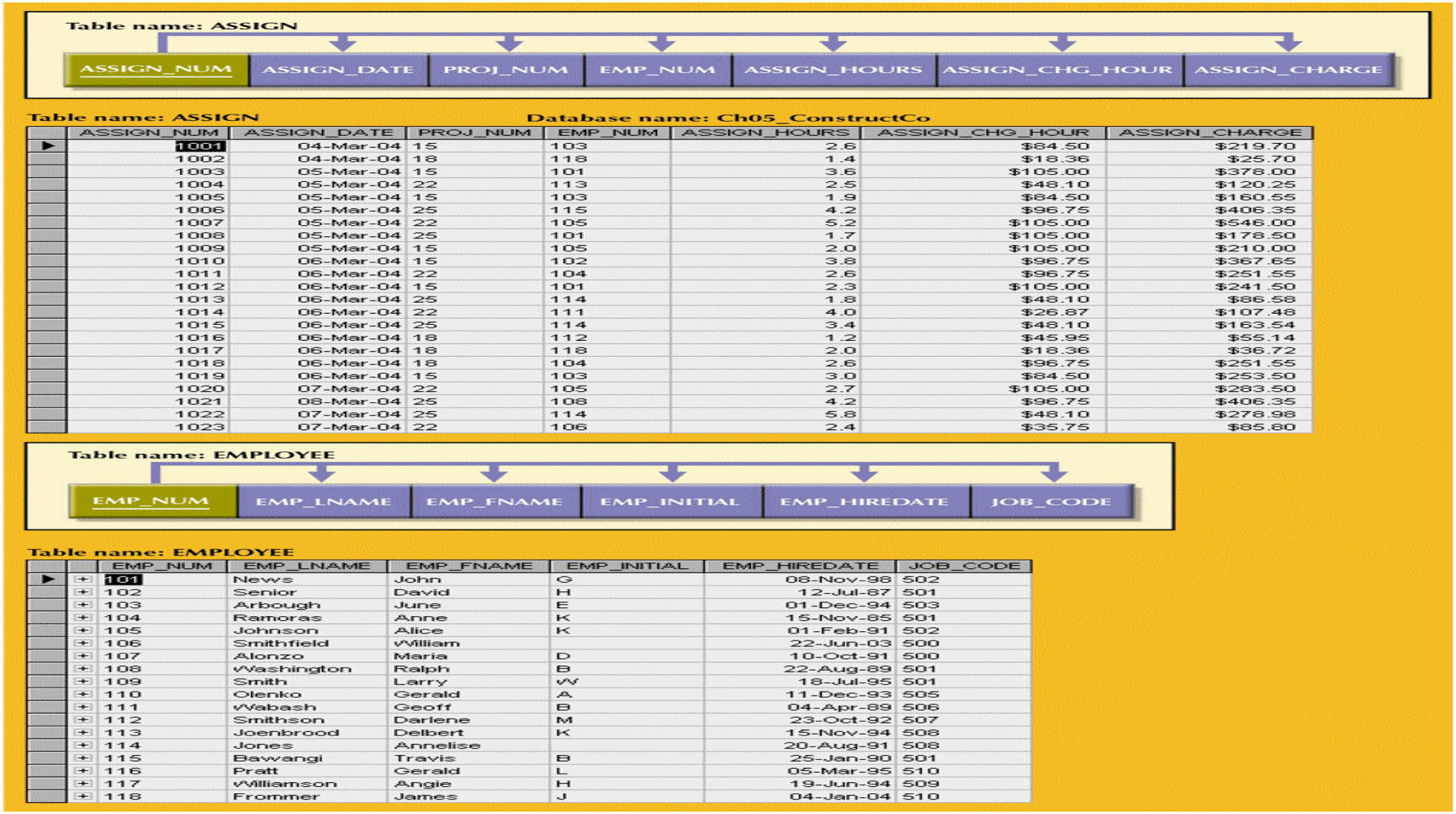
The Completed Database

FIGURE 5.6 THE COMPLETED DATABASE



The Completed Database (continued)

FIGURE 5.6 THE COMPLETED DATABASE (CONTINUED)



Limitations on System-Assigned Keys

- System-assigned primary key may not prevent confusing entries
- Data entries in Table 5.2 are inappropriate because they duplicate existing records
  - Yet there has been no violation of either entity integrity or referential integrity

Duplicate Entries in the JOB Table

TABLE 5.2 DUPLICATE ENTRIES IN THE JOB TABLE

JOB_CODE	JOB_DESCRIPTION	JOB_CHG_HOUR
511	Programmer	$35.75
512	Programmer	$35.75

The Boyce-Codd Normal Form (BCNF)

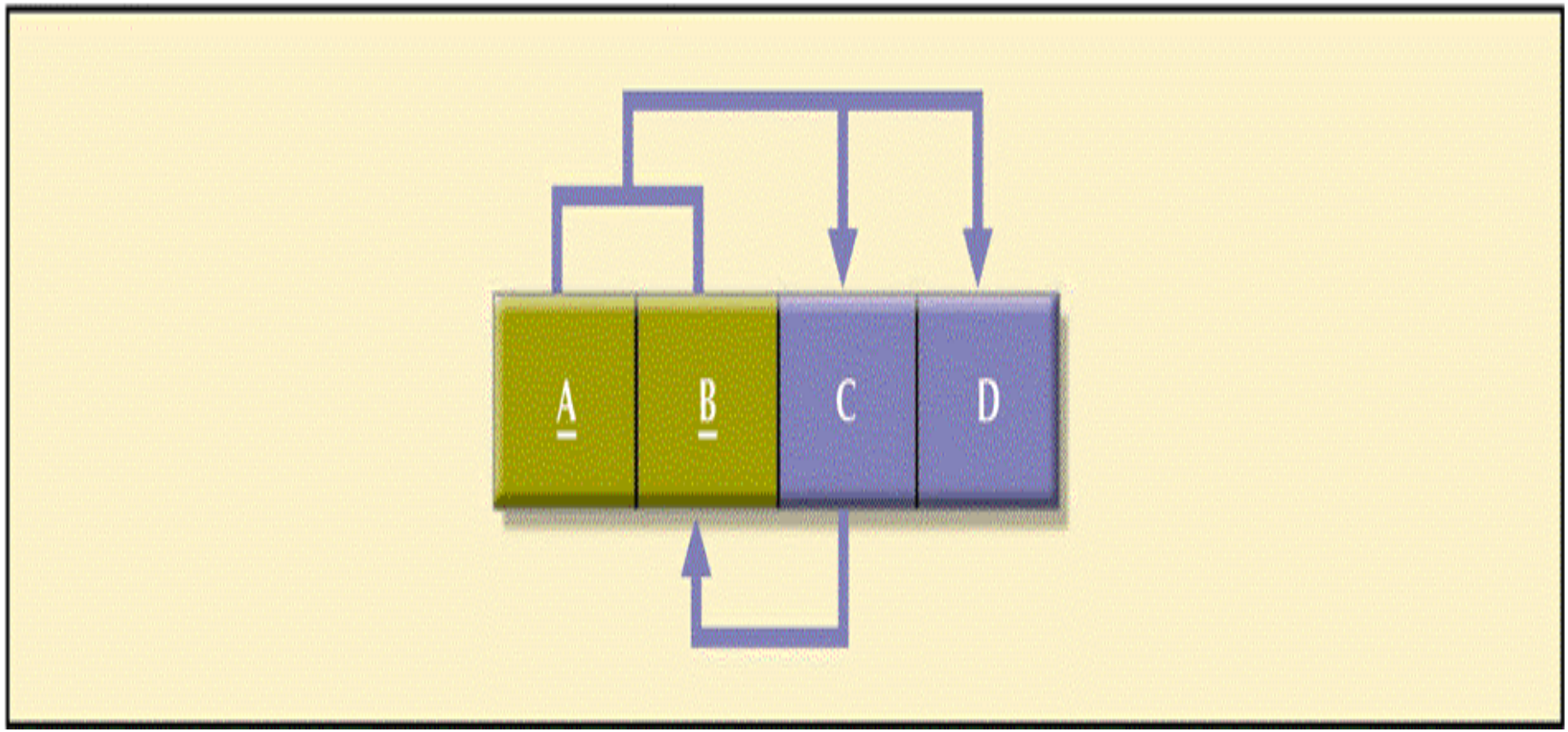
- Every determinant in the table is a candidate key
  - Has same characteristics as primary key, but for some reason, not chosen to be primary key
- If a table contains only one candidate key, the 3NF and the BCNF are equivalent
- BCNF can be violated only if the table contains more than one candidate key

The Boyce-Codd Normal Form (BCNF) (continued)

- Most designers consider the Boyce-Codd normal form (BCNF) as a special case of 3NF
- A table is in 3NF if it is in 2NF and there are no transitive dependencies
- A table can be in 3NF and not be in BCNF
  - A transitive dependency exists when one nonprime attribute is dependent on another nonprime attribute
  - A nonkey attribute is the determinant of a key attribute

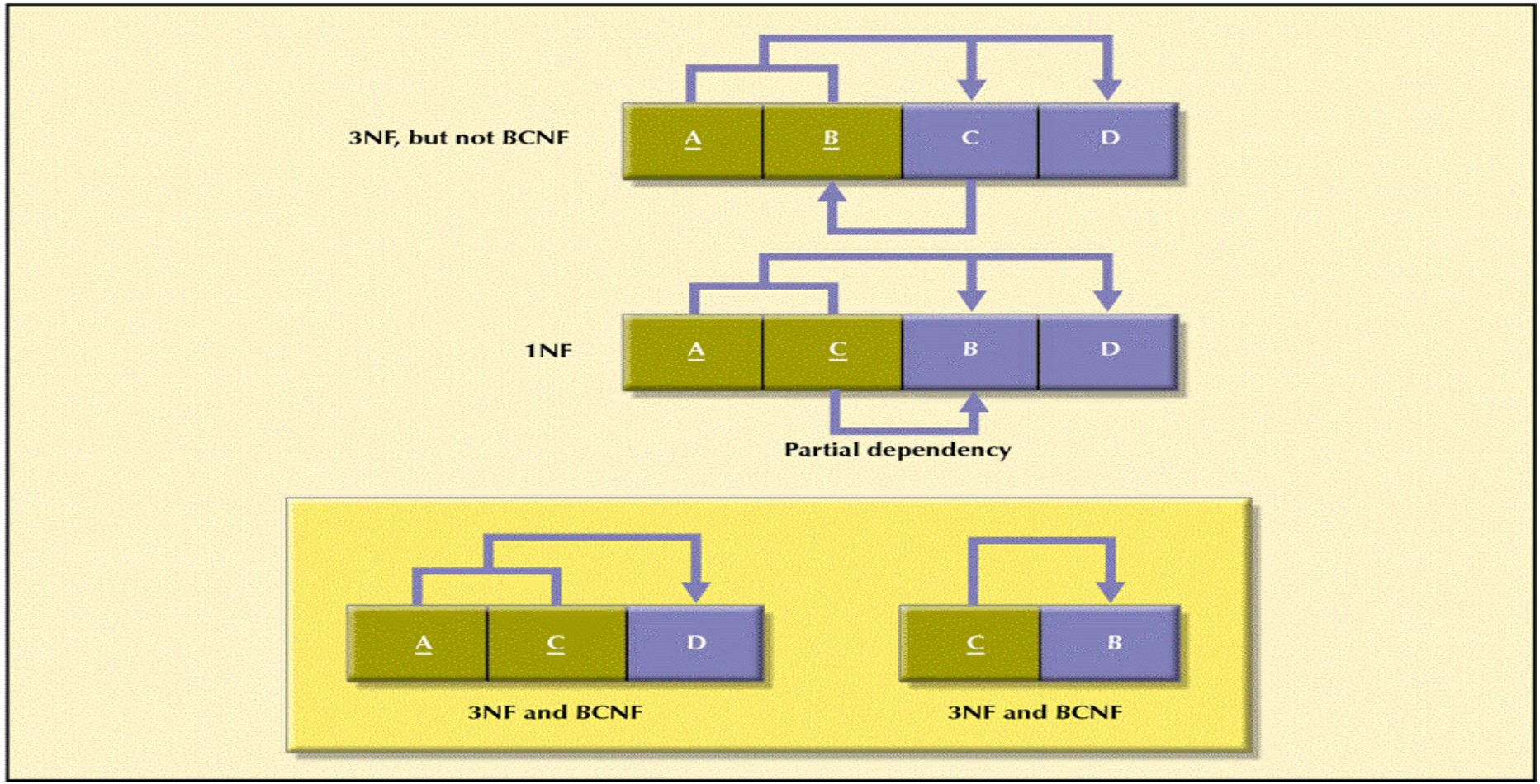
A Table That is in 3NF but not in BCNF

FIGURE 5.7 A TABLE THAT IS IN 3NF BUT NOT IN BCNF



Decomposition to BCNF

FIGURE 5.8 DECOMPOSITION TO BCNF



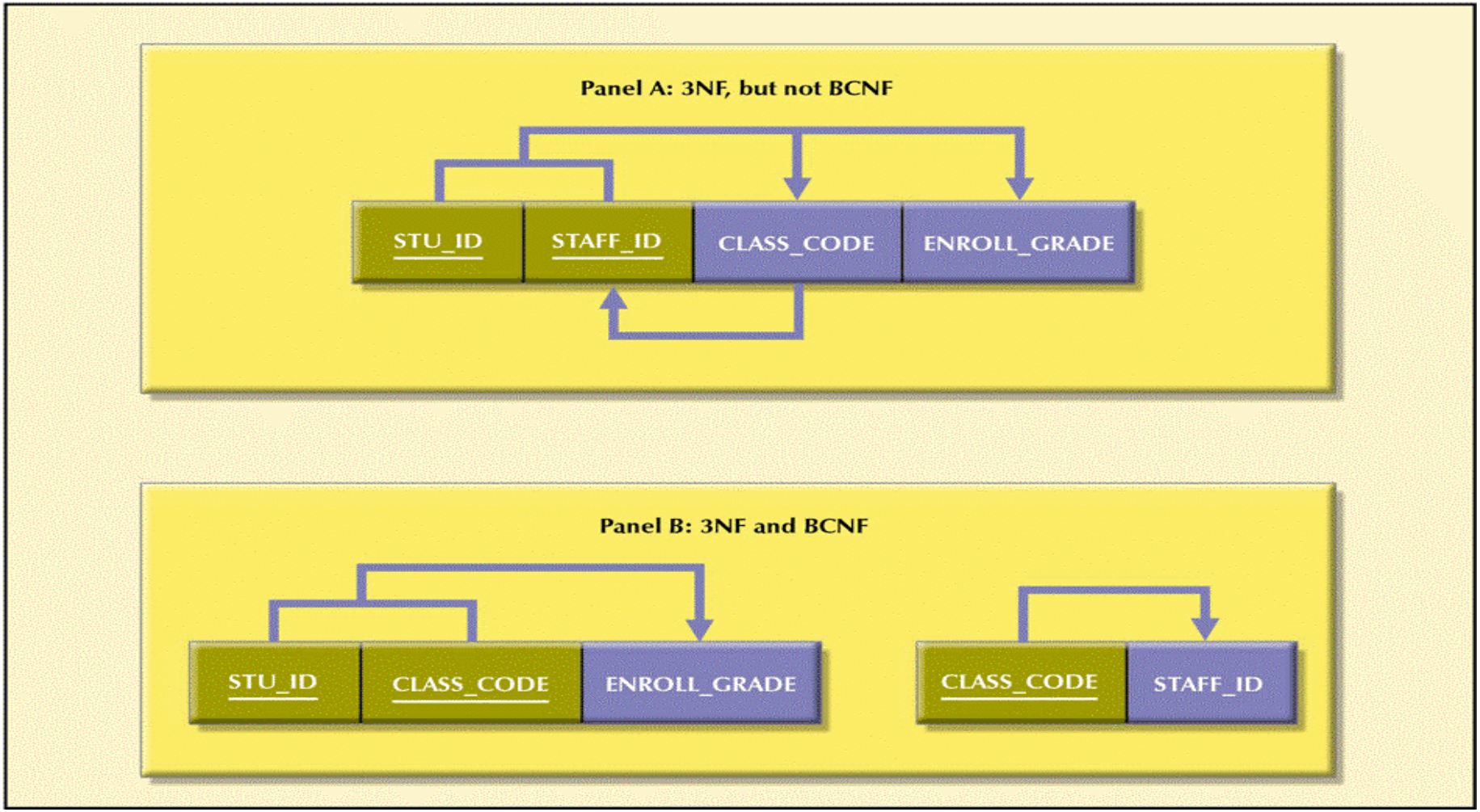
Sample Data for a BCNF Conversion

TABLE 5.3 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

Another BCNF Decomposition

FIGURE 5.9 ANOTHER BCNF DECOMPOSITION



Normalization and Database Design

- Normalization should be part of design process
- Make sure that proposed entities meet required normal form before table structures are created
- Many real-world databases have been improperly designed or burdened with anomalies if improperly modified during course of time
- You may be asked to redesign and modify existing databases

Normalization and Database Design (continued)

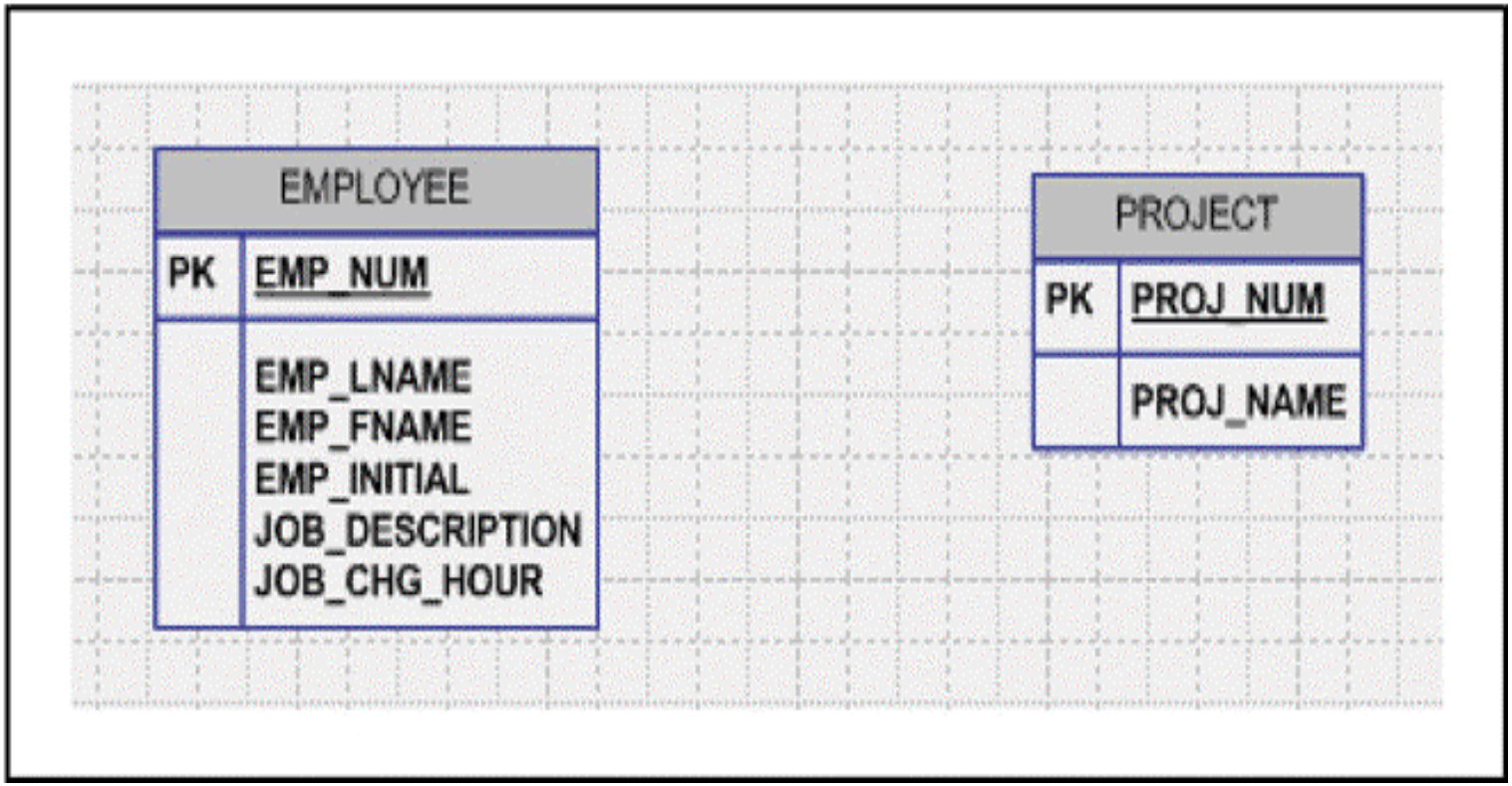
- ER diagram
  - Provides the big picture, or macro view, of an organization's data requirements and operations
  - Created through an iterative process
    - Identifying relevant entities, their attributes and their relationship
    - Use results to identify additional entities and attributes

Normalization and Database Design (continued)

- Normalization procedures
  - Focus on the characteristics of specific entities
  - A micro view of the entities within the ER diagram
- Difficult to separate normalization process from ER modeling process
- Two techniques should be used concurrently

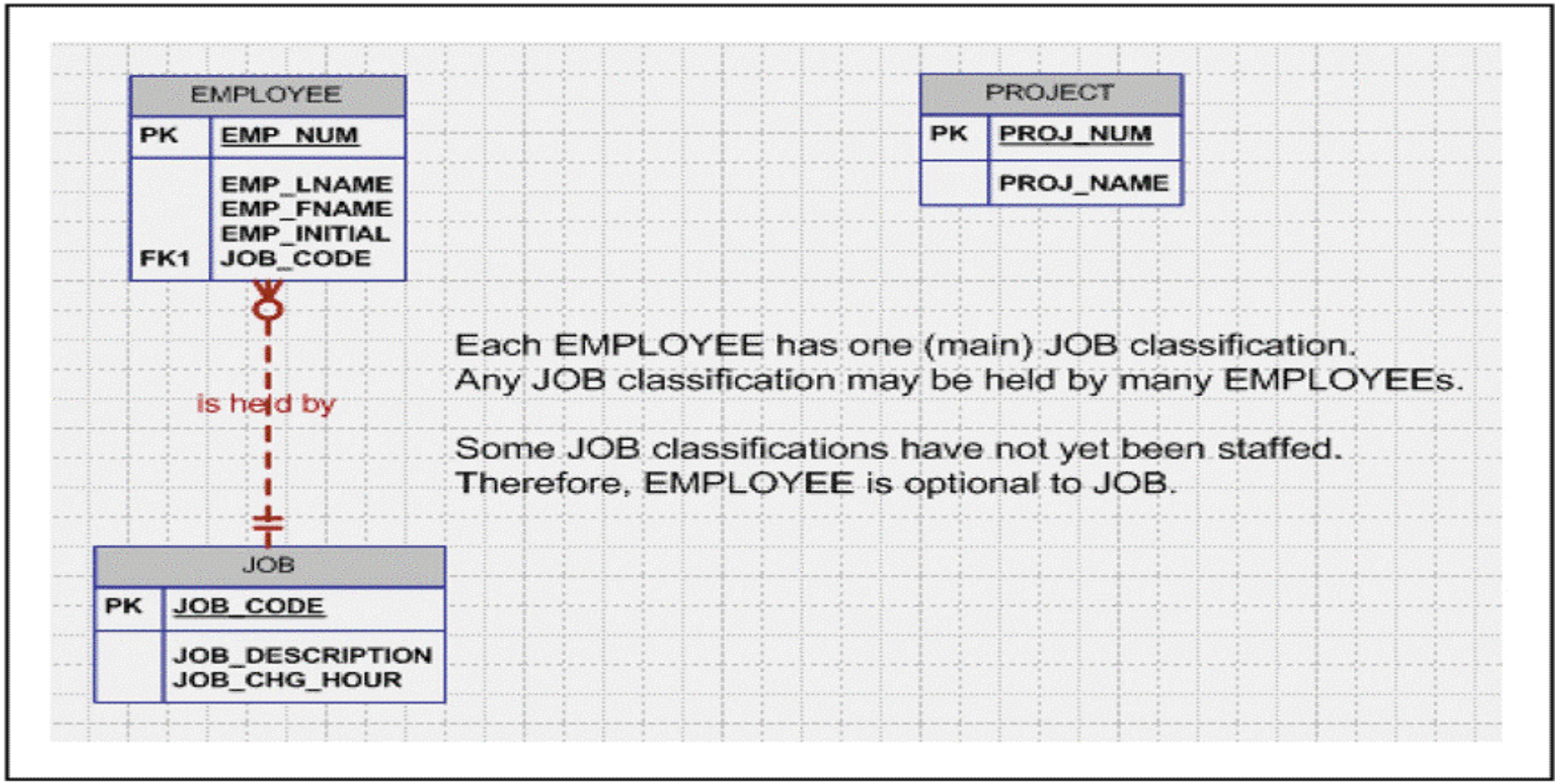
The Initial ERD for a Contracting Company

FIGURE 5.10 THE INITIAL ERD FOR A CONTRACTING COMPANY



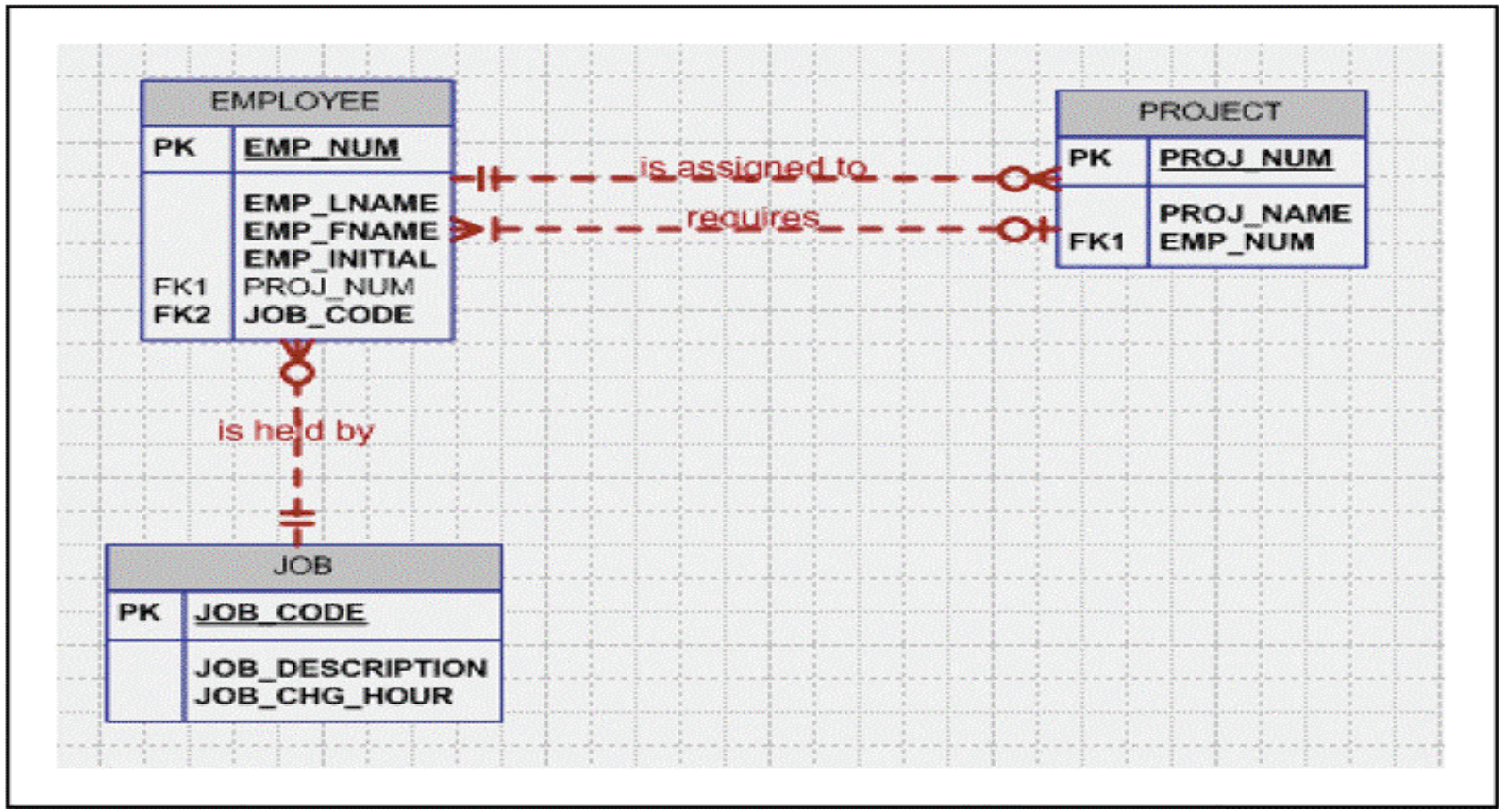
The Modified ERD for a Contracting Company

FIGURE 5.11 THE MODIFIED ERD FOR A CONTRACTING COMPANY



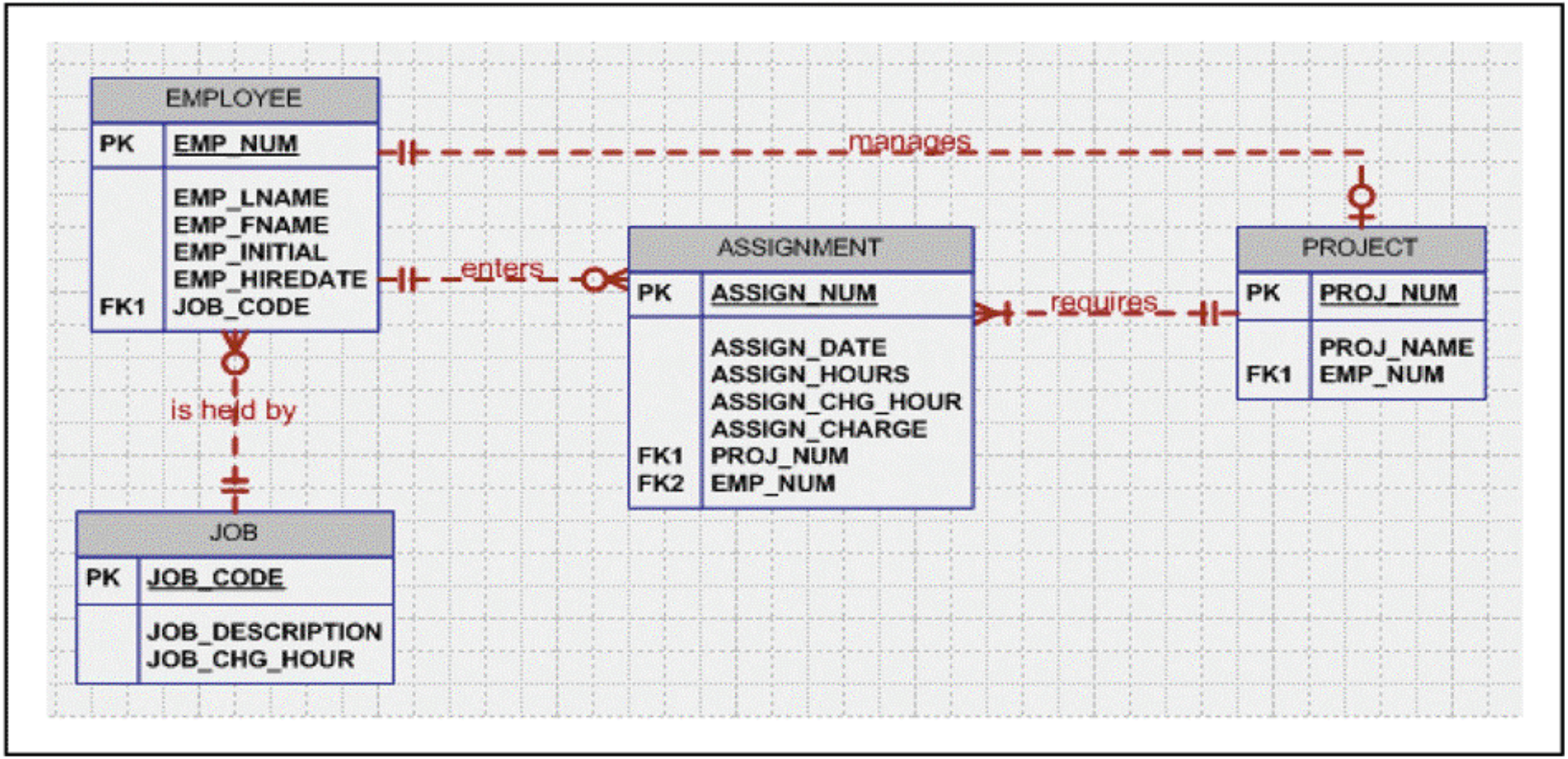
The Incorrect Representation of a M:N Relationship

FIGURE 5.12 THE INCORRECT REPRESENTATION OF A M:N RELATIONSHIP



The Final (Implementable) ERD for a Contracting Company

FIGURE 5.13 THE FINAL (IMPLEMENTABLE) ERD FOR A CONTRACTING COMPANY



The Implemented Database for the Contracting Company

FIGURE 5.14 THE IMPLEMENTED DATABASE FOR THE CONTRACTING COMPANY

Table name: EMPLOYEE — Database name: Ch05_ConstructCo

EMP_NUM	EMP_LNAME	EMP_FNAME	EMP_INITIAL	EMP_HIREDATE	JOB_CODE
101	News	John	G	08-Nov-98	502
102	Senior	David	H	12-Jul-87	501
103	Arbough	June	E	01-Dec-94	503
104	Ramoras	Anne	K	15-Nov-85	501
105	Johnson	Alice	K	01-Feb-91	502
106	Smithfield	William		22-Jun-03	500
107	Alonzo	Maria	D	10-Oct-91	500
108	Washington	Ralph	B	22-Aug-89	501
109	Smith	Larry	W	18-Jul-95	501
110	Olenko	Gerald	A	11-Dec-93	505
111	Wabash	Geoff	B	04-Apr-89	506
112	Smithson	Darlene	M	23-Oct-92	507
113	Joenbrood	Delbert	K	15-Nov-94	508
114	Jones	Annelise		20-Aug-91	508
115	Bawangi	Travis	B	25-Jan-90	501
116	Pratt	Gerald	L	05-Mar-95	510
117	Williamson	Angie	H	19-Jun-94	509
118	Frommer	James	J	04-Jan-04	510

Table name: JOB

JOB_CODE	JOB_DESCRIPTION	JOB_CHG_HOUR
500	Programmer	$35.75
501	Systems Analyst	$96.75
502	Database Designer	$105.00
503	Electrical Engineer	$84.50
504	Mechanical Engineer	$67.90
505	Civil Engineer	$55.78
506	Clerical Support	$26.87
507	DSS Analyst	$45.95
508	Applications Designer	$48.10
509	Bio Technician	$34.55
510	General Support	$18.36

Table name: PROJECT

PROJ_NUM	PROJ_NAME	EMP_NUM
15	Evergreen	105
18	Amber Wave	104
22	Rolling Tide	113
25	Starflight	101

Table name: ASSIGNMENT

ASSIGN_NUM	ASSIGN_DATE	PROJ_NUM	EMP_NUM	ASSIGN_HOURS	ASSIGN_CHG_HOUR	ASSIGN_CHARGE
1001	04-Mar-04	15	103	2.6	$84.50	$219.70
1002	04-Mar-04	18	118	1.4	$18.36	$25.70
1003	05-Mar-04	15	101	3.6	$105.00	$378.00
1004	05-Mar-04	22	113	2.5	$48.10	$120.25
1005	05-Mar-04	15	103	1.9	$84.50	$160.55
1006	05-Mar-04	25	115	4.2	$96.75	$406.35
1007	05-Mar-04	22	105	5.2	$105.00	$546.00
1008	05-Mar-04	25	101	1.7	$105.00	$178.50
1009	05-Mar-04	15	105	2.0	$105.00	$210.00
1010	06-Mar-04	15	102	3.8	$96.75	$367.65
1011	06-Mar-04	22	104	2.6	$96.75	$251.55
1012	06-Mar-04	15	101	2.3	$105.00	$241.50
1013	06-Mar-04	25	114	1.8	$48.10	$86.58
1014	06-Mar-04	22	111	4.0	$26.87	$107.48
1015	06-Mar-04	25	114	3.4	$48.10	$163.54
1016	06-Mar-04	18	112	1.2	$45.95	$55.14
1017	06-Mar-04	18	118	2.0	$18.36	$36.72
1018	06-Mar-04	18	104	2.6	$96.75	$251.55
1019	06-Mar-04	15	103	3.0	$84.50	$253.50
1020	07-Mar-04	22	105	2.7	$105.00	$283.50
1021	08-Mar-04	25	108	4.2	$96.75	$406.35
1022	07-Mar-04	25	114	5.8	$48.10	$278.98
1023	07-Mar-04	22	106	2.4	$35.75	$85.80

Higher-Level Normal Forms

- In some databases, multiple multivalued attributes exist

Tables with Multivalued Dependencies

FIGURE 5.15 TABLES WITH MULTIVALUED DEPENDENCIES

Database name: Ch05_Service

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
10223		4

Table name: VOLUNTEER_V3

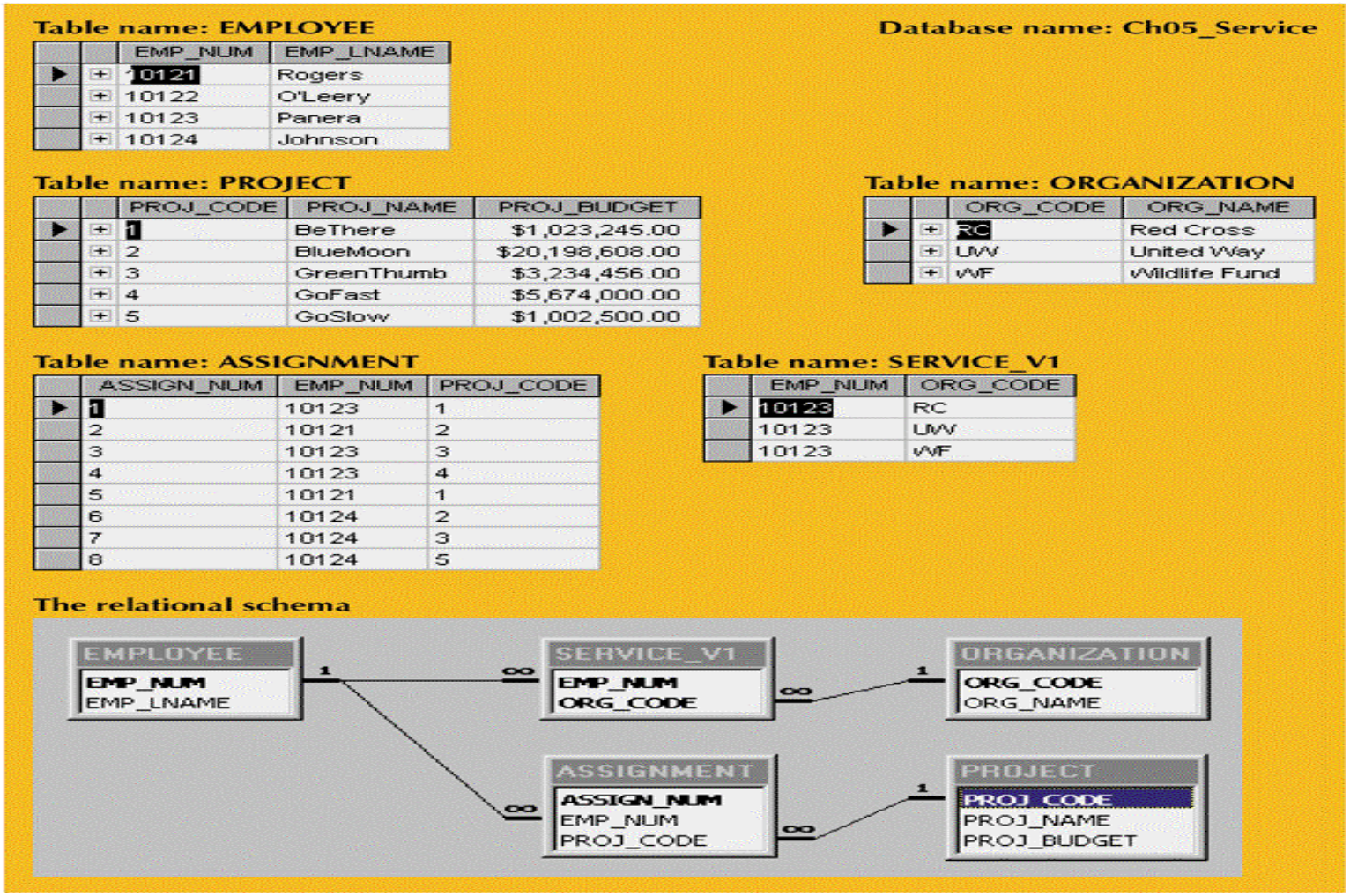
EMP_NUM	ORG_CODE	ASSIGN_NUM
10123	RC	1
10123	RC	3
10123	UW	4

Fourth Normal Form

- Table is in fourth normal form (4NF) if
  - It is in 3NF
  - Has no multiple sets of multivalued dependencies
- 4NF is largely academic if tables conform to the following two rules:
  - All attributes are dependent on primary key but independent of each other
  - No row contains two or more multivalued facts about an entity

A Set of Tables in 4NF

FIGURE 5.16 A SET OF TABLES IN 4NF



Denormalization

- Creation of normalized relations is important database design goal
- Processing requirements should also be a goal
- If tables decomposed to conform to normalization requirements
  - Number of database tables expands

Denormalization (continued)

- Joining larger number of tables takes additional disk input/output (I/O) operations and processing logic
  - Reduces system speed
- Conflicts among design efficiency, information requirements, and processing speed are often resolved through compromises that may include denormalization

Denormalization (continued)

- Unnormalized tables in a production database tend to have these defects:
  - Data updates are less efficient because programs that read and update tables must deal with larger tables
  - Indexing is much more cumbersome
  - Unnormalized tables yield no simple strategies for creating virtual tables known as *views*

Denormalization (continued)

- Use denormalization cautiously
- Understand why—under some circumstances—unnormalized tables are a better choice

Summary

- Normalization is a table design technique aimed at minimizing data redundancies
- First three normal forms (1NF, 2NF, and 3NF) are most commonly encountered
- Normalization is an important part—but only a part—of the design process
- Continue the iterative ER process until all entities and their attributes are defined and all equivalent tables are in 3NF

Summary (continued)

- A table in 3NF may contain multivalued dependencies that produce either numerous null values or redundant data
- It may be necessary to convert a 3NF table to the fourth normal form (4NF) by
  - splitting such a table to remove multivalued dependencies
- Tables are sometimes denormalized to yield less I/O which increases processing speed