Normalization of Database Tables

Database Tables and Normalization

- **Normalization is a process for assigning attributes to entities. It reduces data redundancies and helps eliminate the data anomalies.**
- **Normalization works through a series of stages called normal forms:**
  - **First normal form (1NF)**
  - **Second normal form (2NF)**
  - **Third normal form (3NF)**
  - **Fourth normal form (4NF)**
- ***The highest level of normalization is not always desirable*.**

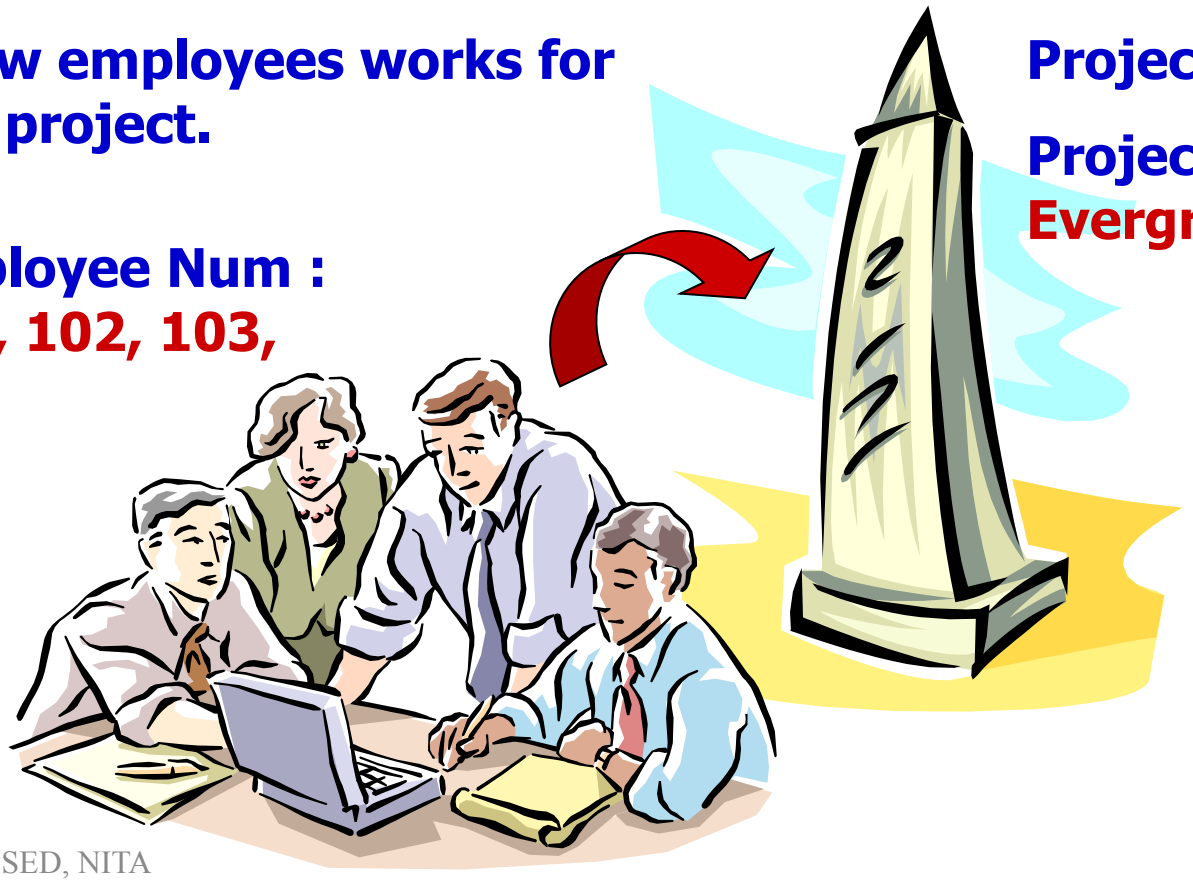
Database Tables and Normalization

- **The Need for Normalization**
  - **Case of a Construction Company**
    - **Building project -- Project number, Name, Employees assigned to the project.**
    - **Employee -- Employee number, Name, Job classification**
    - **The company charges its clients by billing the hours spent on each project. The hourly billing rate is dependent on the employee's position.**
    - **Periodically, a report is generated.**

Scenario

A few employees works for one project.

Employee Num : 101, 102, 103, 105



Project Num : 15

Project Name : Evergreen

Sample Form

Project Num : 15

Project Name : Evergreen

Emp Num	Emp Name	Job Class	Chr Hours	Hrs Billed	Total
101					
102					
103					
105					

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	$3,134.70
		112	Darlene M. Smithson	DSS Analyst	$45.95	44.0	$2,021.80
				Subtotal			**$7,171.47**
22		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		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**

Note: * indicates project leader

Table Structure Matches the Report Format

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

FIGURE 5.1 A TABLE WHOSE STRUCTURE MATCHES THE REPORT FORMAT

Database Tables and Normalization

- **Problems with the Figure**
  - **The project number is intended to be a primary key, but it contains nulls.**
  - **The table displays data redundancies.**
  - **The table entries invite data inconsistencies.**
  - **The data redundancies yield the following anomalies:**
    - **Update anomalies.**
    - **Addition anomalies.**
    - **Deletion anomalies.**

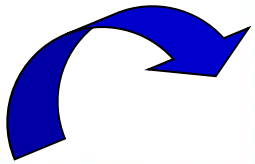
Database Tables and Normalization

- **Conversion to First Normal Form**
  - **A relational table must not contain repeating groups.**
  - **Repeating groups can be eliminated by adding the appropriate entry in at least the primary key column(s).**

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

FIGURE 5.2 THE EVERGREEN DATA

Data Organization: First Normal Form



PROJ_NUM	PROJ_NAME	EMP_NUM	EMP_NAME	JOB_CLASS	CHG_HOUR	HO
15	Evergreen	03	June E. Arbough	Elect. Engineer	$84.50	
		01	John G. News	Database Designer	$105.00	
		05	Alice K. Johnson *	Database Designer	$105.00	
		06	William Smithfield	Programmer	$35.75	
		02	David H. Senior	Systems Analyst	$96.75	
18	Amber Wave	14	Annelise Jones	Applications Designer	$48.10	
		18	James J. Frommer	General Support	$18.36	
		04	Anne K. Ramoras *	Systems Analyst	$96.75	
		12	Darlene M. Smithson	DSS Analyst	$45.95	
22	Rolling Tide	05	Alice K. Johnson	Database Designer	$105.00	
		04	Anne K. Ramoras	Systems Analyst	$96.75	
		13	Delbert K. Joenbrood *	Applications Designer	$48.10	
		11	Geoff B. Wabash	Clerical Support	$26.87	
		06	William Smithfield	Programmer	$35.75	
25	Starflight	07	Maria D. Alonzo	Programmer	$35.75	
		15	Travis B. Bawangi	Systems Analyst	$96.75	
		01	John G. News *	Database Designer	$105.00	
		14	Annelise Jones	Applications Designer	$48.10	33.1
		08	Ralph B. Washington	Systems Analyst	$96.75	23.6
		18	James J. Frommer	General Support	$18.36	30.5
		12	Darlene M. Smithson	DSS Analyst	$45.95	41.4

FIGURE 5.1 A T HES THE REPORT FORMAT

Before

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	12.5
15	Evergreen	102	David H. Senior	Systems Analyst	$96.75	23.9
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	32.1
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	48.9
22	Rolling Tide	113	Delbert K. Joenbrood *	Applications Designer	$48.10	23.6
22	Rolling Tide	111	Geoff B. Wabash	Clerical Support	$26.87	22.5
22	Rolling Tide	106	William Smithfield	Programmer	$35.75	12.1
25	Starflight	107	Maria D. Alonzo	Programmer	$35.75	24.7
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	23.9
25	Starflight	118	James J. Frommer	General Support	$18.36	30.2
25	Starflight	112	Darlene M. Smithson	DSS Analyst	$45.95	41.4

FIGURE 5.3 RST NORMAL FORM

After

First Normal Form (1 NF)

- **1NF Definition**
  - **The term first normal form (1NF) describes the tabular format in which:**
    - **All the key attributes are defined.**
    - **There are no repeating groups in the table.**
    - **All attributes are dependent on the primary /candidate key.**

Dependency Diagram

- **Dependency Diagram**
  - **The primary key components are bold, underlined.**
  - **The arrows above entities indicate all desirable dependencies, i.e., dependencies that are based on PK.**
  - **The arrows below the dependency diagram indicate less desirable dependencies -- partial dependencies and transitive dependencies.**

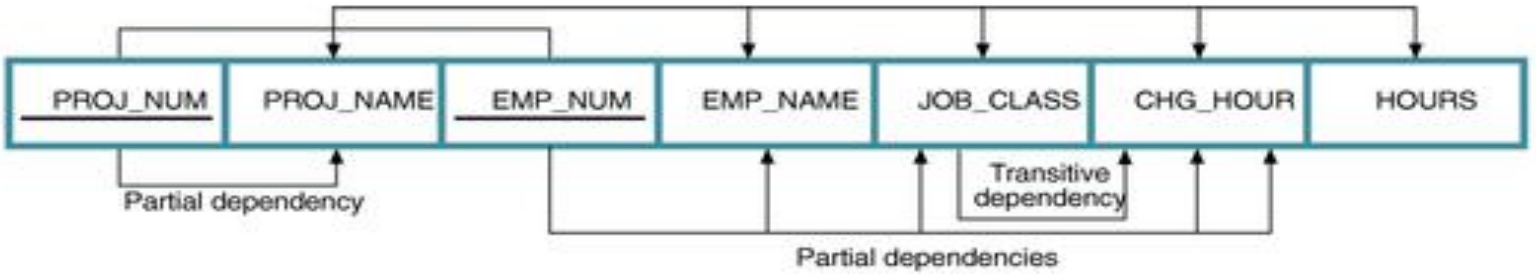


FIGURE 5.4 A DEPENDENCY DIAGRAM: FIRST NORMAL FORM (1NF)

Second Normal Form (2 NF)

- **Conversion to Second Normal Form**
  - **Starting with the 1NF format, the database can be converted into the 2NF format by**
    - **Writing each key component on a separate line, and then writing the original key on the last line and**
    - **Writing the dependent attributes after each new key.**

PROJECT (PROJ_NUM, PROJ_NAME)

EMPLOYEE (EMP_NUM, EMP_NAME, JOB_CLASS, CHG_HOUR)

ASSIGN (PROJ_NUM, EMP_NUM, HOURS)

Dependency Diagram

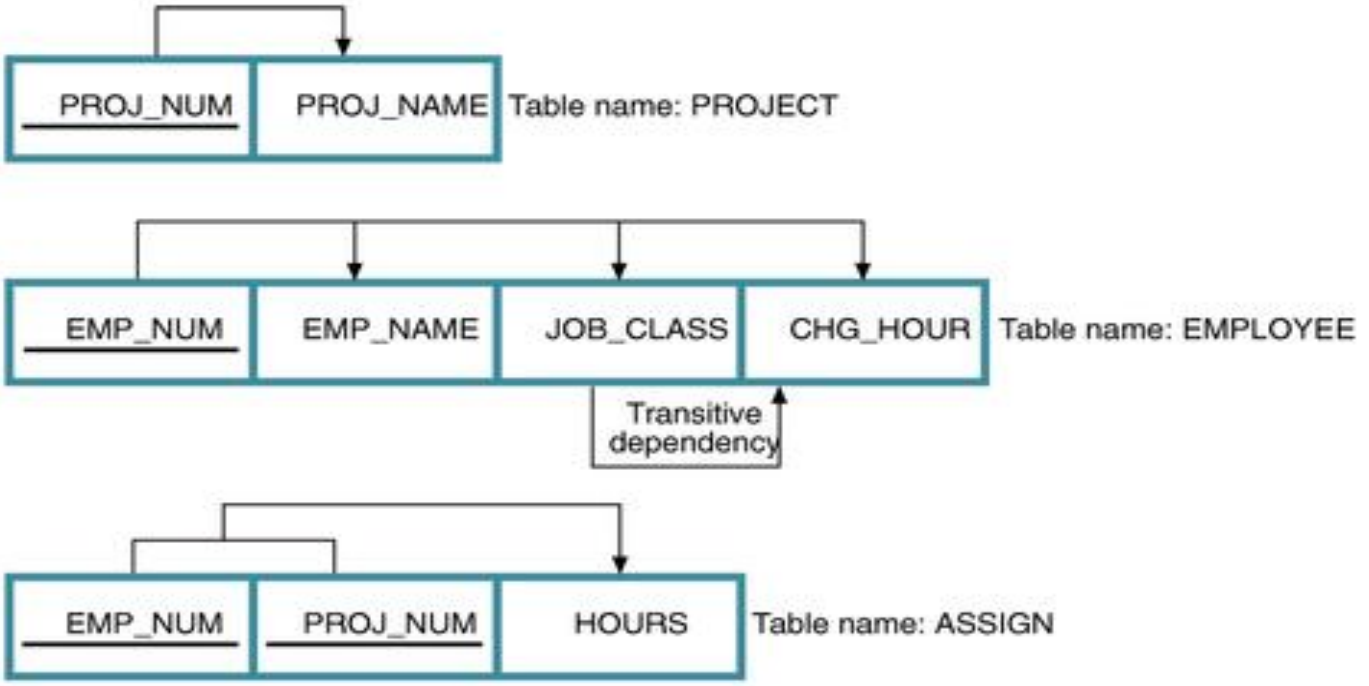


FIGURE 5.5 SECOND NORMAL FORM (2NF) CONVERSION RESULTS

Second Normal Form (2 NF)

A table is in 2NF if:

- **It is in 1NF and**
- **It includes no partial dependencies; that is, no attribute is dependent on only a portion of the primary key.**

  (It is still possible for a table in 2NF to exhibit transitive dependency; that is, one or more attributes may be functionally dependent on nonkey attributes.)

Third Normal Form (3 NF)

- Conversion to Third Normal Form
  - **Create a separate table with attributes in a transitive functional dependence relationship.**

PROJECT (PROJ_NUM, PROJ_NAME)
ASSIGN (PROJ_NUM, EMP_NUM, HOURS)
EMPLOYEE (EMP_NUM, EMP_NAME, JOB_CLASS)
JOB (JOB_CLASS, CHG_HOUR)

Third Normal Form (3 NF)

- **3NF Definition**
  - **A table is in 3NF if:**
    - **It is in 2NF and**
    - **It contains no transitive dependencies.**

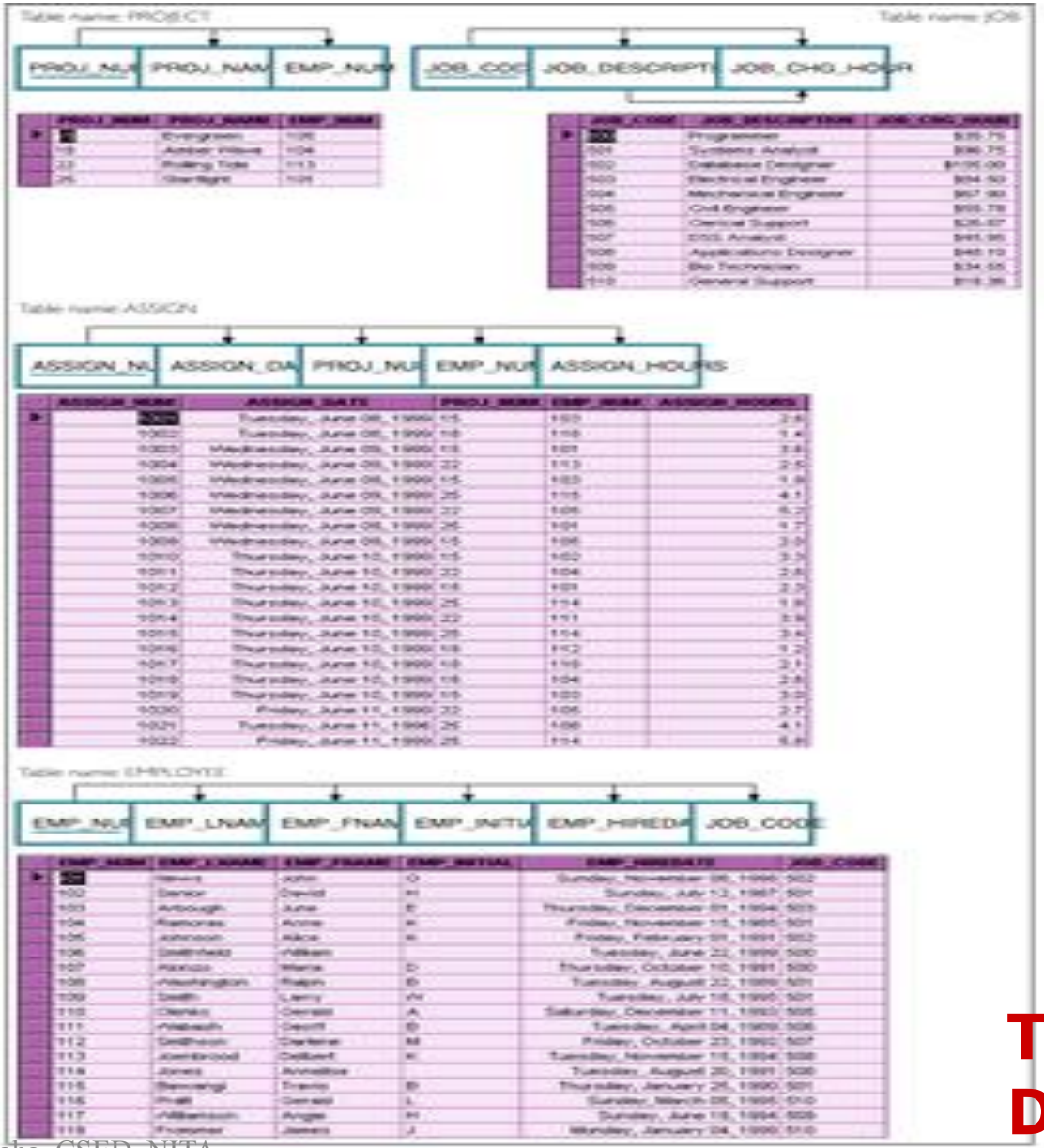


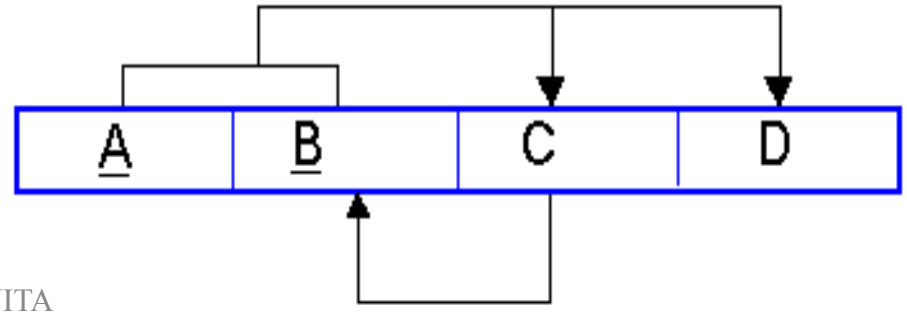
FIGURE 5.6 THE COMPLETED DATABASE

The Completed Database

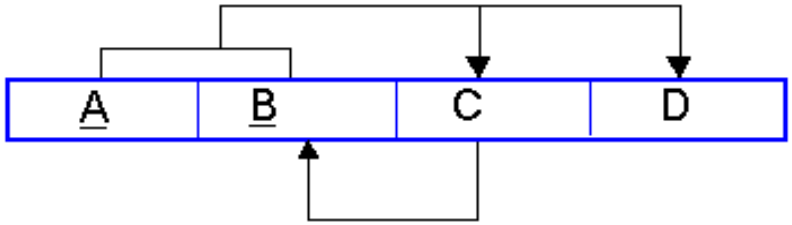
BCNF

- **Boyce Codd Normal Form (BCNF)** is considered a special condition of third Normal form. **A table is in BCNF if every determinant is a candidate key.** A table can be in 3NF but not in BCNF. This occurs when a non key attribute is a determinant of a key attribute.

- The dependency diagram may look like the one below



BCNF



- The table is in 3NF. A and B are the keys and C and D depend on both A and B.
- The table is not in BCNF because a dependency exists between C and B. In other words if we know the value of C we can determine the value of B.
- We can also show the dependencies as

  A B } → C D

  C } → B

Example

S_Num	T_Code	Subject_id	Exam Date
123599	FIT104	01764	2nd March
123599	PIT305	01765	12th April
123599	PIT107	01789	2nd May
346700	FIT104	01764	3rd March
346700	PIT305	01765	7th May

An example table from the University Database might be as follows:

If we know the **Student Number** and **Teacher Code** we know the subject ID the student is in. We also know the exam date.

Example

S_Num	T_Code	Subject_id	Exam Date
123599	FIT104	01764	2nd March
123599	PIT305	01765	12th April
123599	PIT107	01789	2nd May
346700	FIT104	01764	3rd March
346700	PIT305	01765	7th May

The table is not in BCNF as if we know the subject ID we know who the teacher is. Assume each subject can only have one teacher!

Subject-Id→T_Code

A non key attribute (**Subject-Id**) is a **determinant**.

Example

S_Num	T_Code	Subject_id	Exam Date
123599	FIT104	01764	2nd March
123599	PIT305	01765	12th April
123599	PIT107	01789	2nd May
346700	FIT104	01764	3rd March
346700	PIT305	01765	7th May

If we look at the table we can see a combination of T_Code and subject_id is repeated several times. Eg FIT104 and 01764.

Converting to BCNF

- The situation is resolved by following the steps below

  →The determinant, subj_id becomes part of the key and the dependant attribute T_Code, becomes a non key attribute. So the Dependency diagram is now

  S_Num, subj_id } → T_Code, exam Date

  →There are problems with this structure as T_Code is may not dependant on any part of the key . This violates the rules for 2NF

Converting to BCNF

→ So the table needs to be divided with the partial dependency becoming a new table. The dependencies would then be

- S_Num, Subj_id } → exam Date
- Subj_id } → T_Code

StudentExam

S_Num	Subj_id	Exam Date
123599	01764	2nd March
123599	01765	12th April
123599	01789	2nd May
346700	01764	3rd March
346700	01765	7th May

Subj_teacher

Subj_id	T_Code
01764	FIT104
01765	PIT305
01789	PIT107

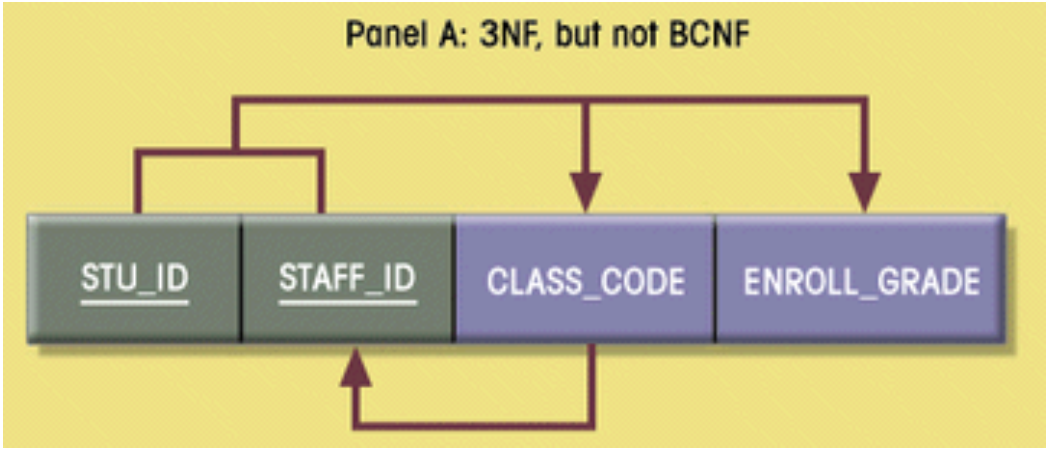
The original table is divided into two new tables. Each is in 3NF and in BCNF.

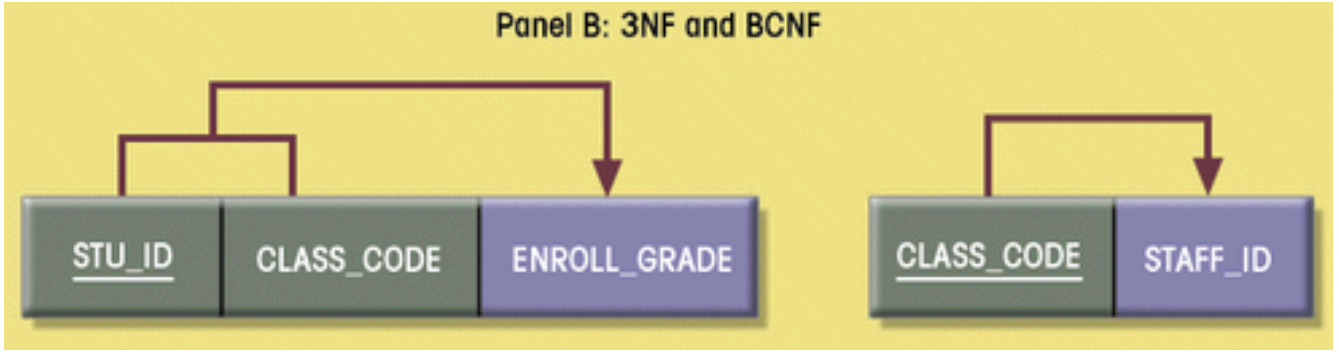
Sample Data for a BCNF Conversion

TABLE 5.2 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

Decomposition into BCNF





BCNF Definition

- **BCNF Definition**
  - **A table is in BCNF if every determinant in that table is a candidate key. If a table contains only one candidate key, 3NF and BCNF are equivalent.**

Normalization and Database Design

- **Attribute ASSIGN_HOUR is assigned to the composite entity ASSIGN.**
- **"Manages" relationship is created between EMPLOYEE and PROJECT.**

PROJECT (PROJ_NUM, PROJ_NAME, EMP_NUM)
EMPLOYEE (EMP_NUM, EMP_LNAME, EMP_FNAME, EMP_INITIAL, EMP_HIREDATE, JOB_CODE)
JOB (JOB_CODE, JOB_DESCRIPTION, JOB_CHG_HOUR)
ASSIGN (ASSIGN_NUM, ASSIGN_DATE, PROJ_NUM, EMP_NUM, ASSIGN_HOURS)

Relational Schema

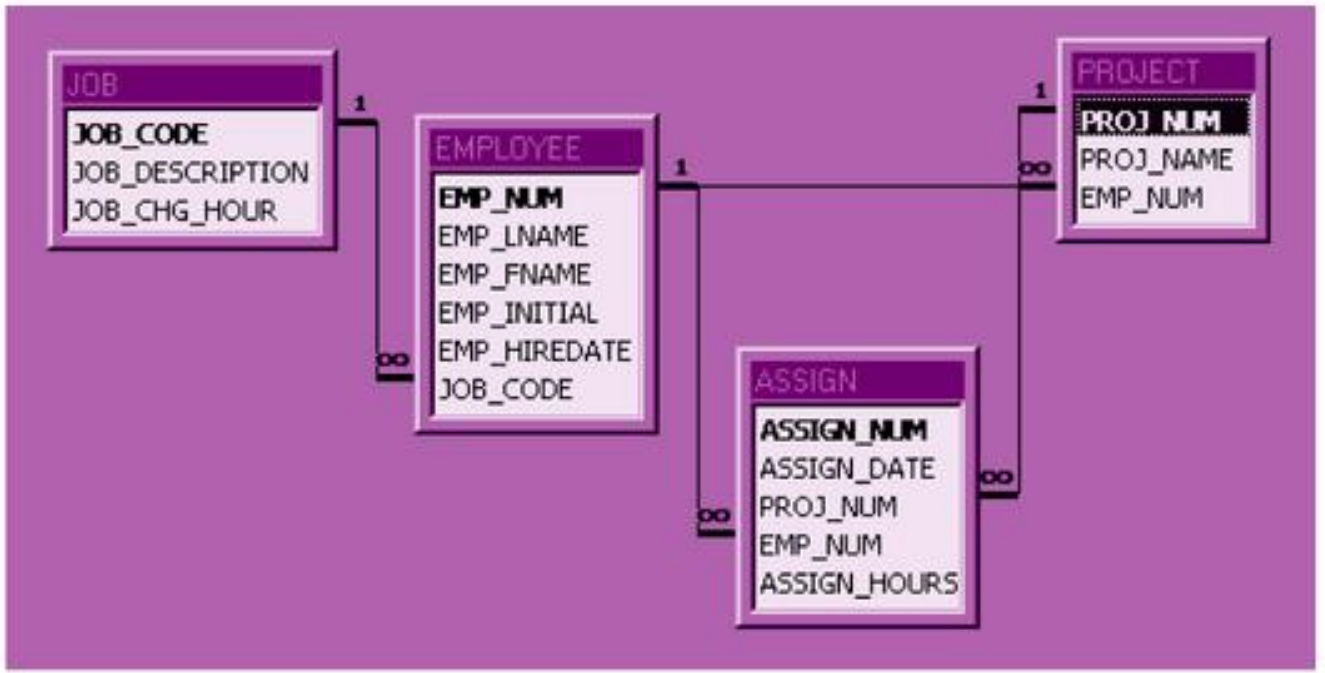


FIGURE 5.13 THE RELATIONAL SCHEMA FOR THE CONTRACTING COMPANY

Higher-Level Normal Forms

- **4NF Definition**
  - **A table is in 4NF if it is in 3NF and has no multiple sets of multivalued dependencies.**

Table name: FIG5_15A

EMP_NUM	EMP_SERVICE
1023	Red Cross
1023	United Way

Table name: FIG5_15B

EMP_NUM	EMP_ASSIGN
1023	1
1023	5
1023	12

FIGURE 5.15 A SET OF TABLES IN 4NF

5th Normalization

- **Fifth normal form** (**5NF**), also known as **project-join normal form** (**PJ/NF**) is a level of database normalization designed to reduce redundancy in relational databases recording multi-valued facts by isolating semantically related multiple relationships.
- A table is said to be in the 5NF if and only if every non-trivial join dependency in it is implied by the candidate keys.

De-normalization

- **Normalization is only one of many database design goals.**
- **Normalization purity is often difficult to sustain in the modern database environment. The conflict between design efficiency, information requirements, and processing speed are often resolved through compromises that include de-normalization.**

SUMMARY

The Initial 1NF Structure

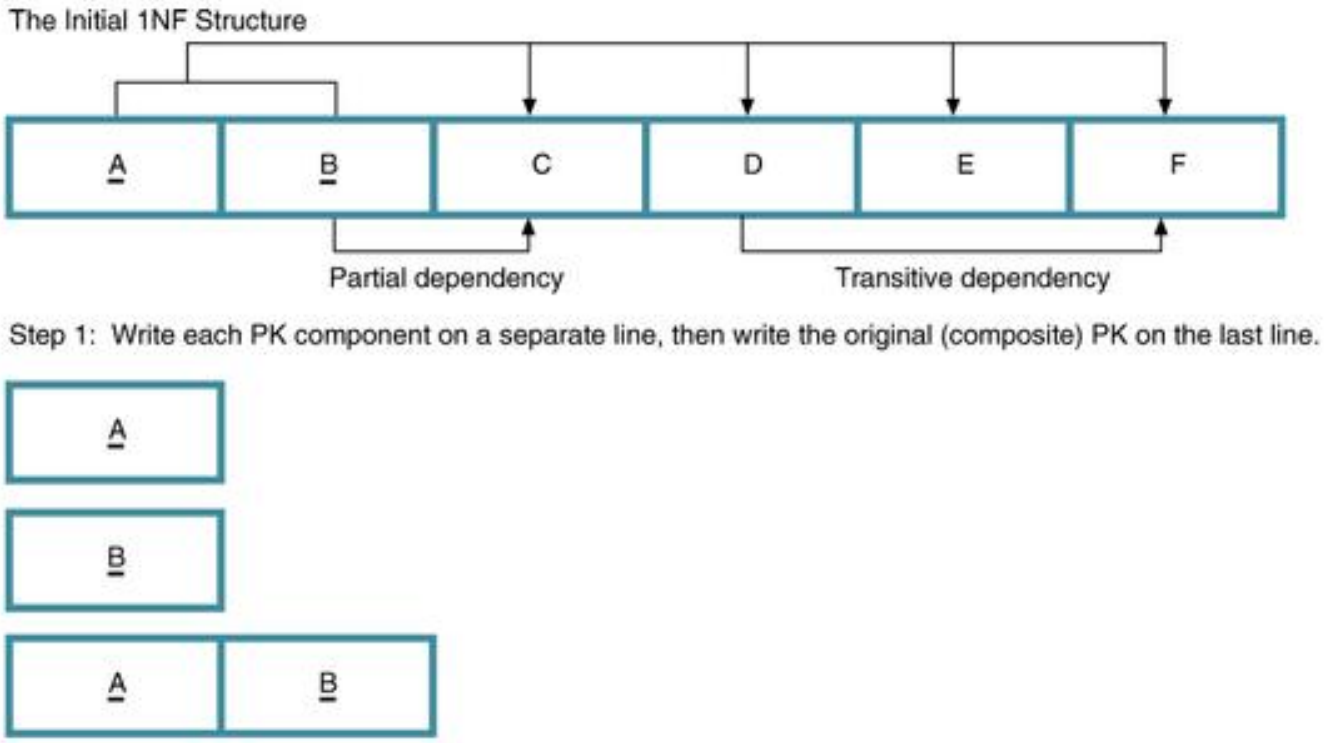


FIGURE 5.16 THE INITIAL 1NF STRUCTURE

Identifying the Possible PK Attributes

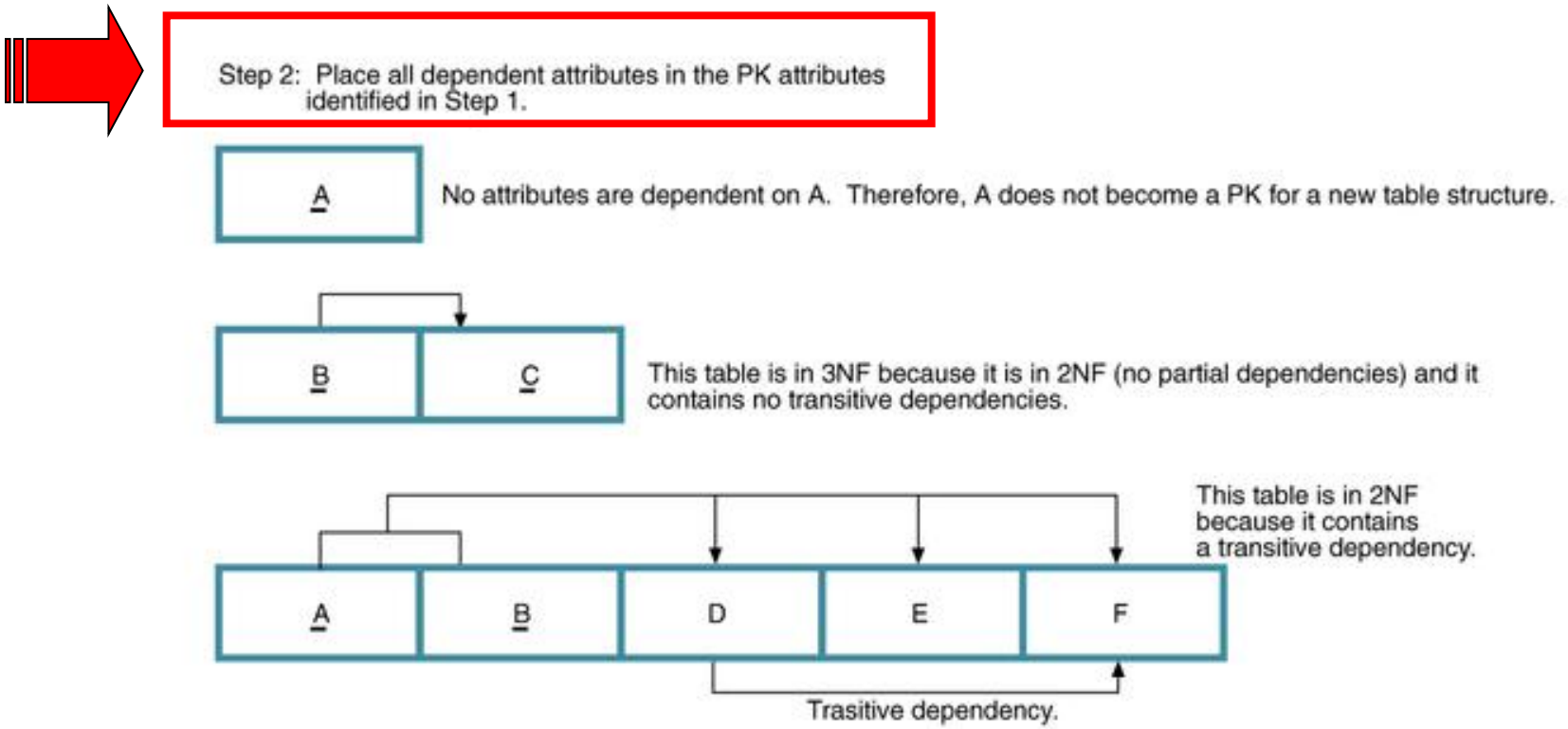


FIGURE 5.17 IDENTIFYING THE POSSIBLE PK ATTRIBUTES

Table Structures Based On The Selected PKs

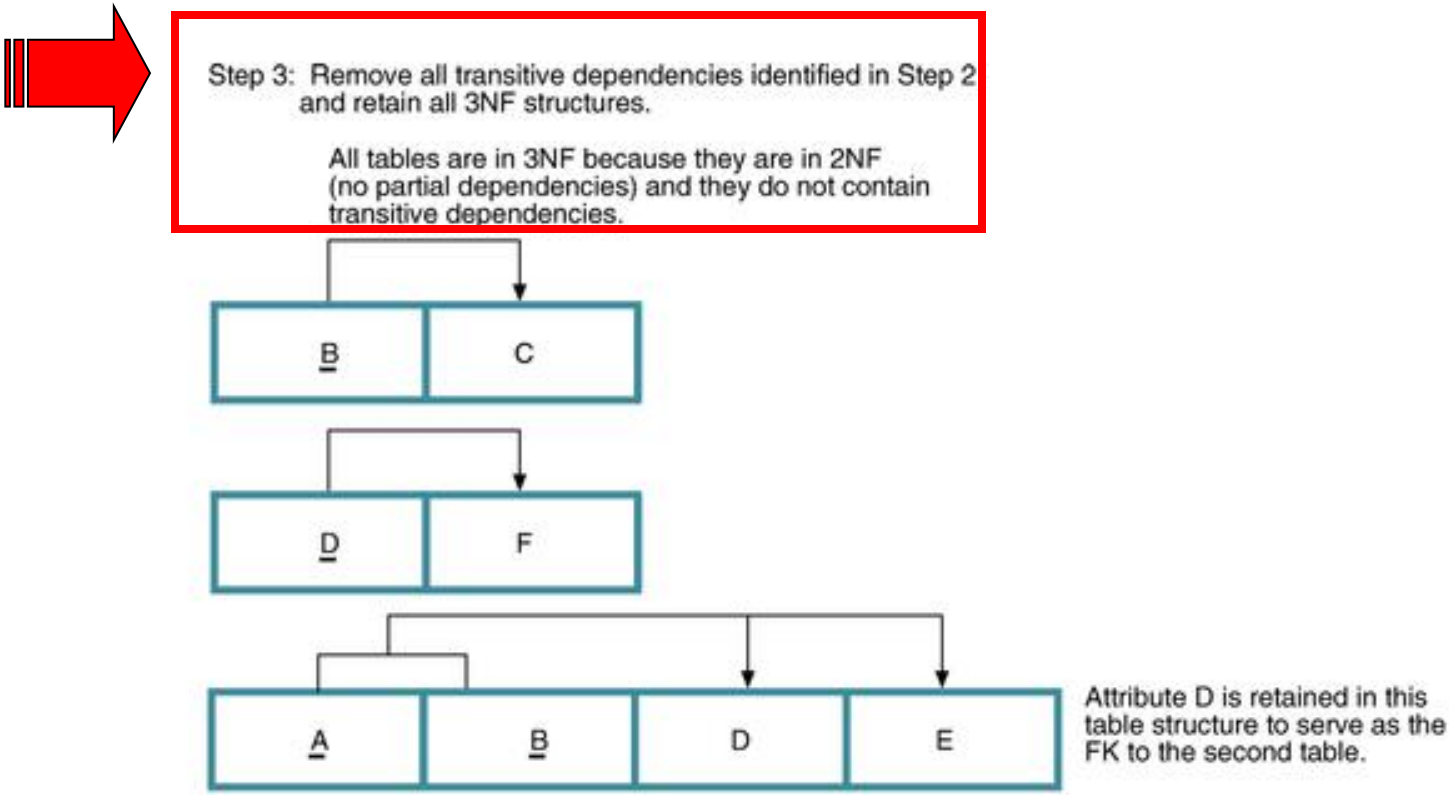


FIGURE 5.18 TABLE STRUCTURES BASED ON THE SELECTED PKS