CptS 415 Big Data

Design of Relational Data Models

Srini Badri

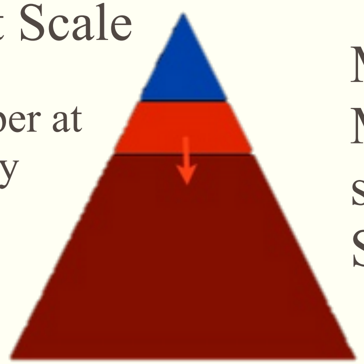
Acknowledgements: Tinghui Wang

Design of relational data models

- Functional dependencies
- The normal forms 1NF, 2NF, 3NF, BCNF

A Big Data Fallacy

- Data model design in the era of Big Data is less important?
    - New high-volume data streams
    - Specialized hardware/software
    - Storage issues coped by hardware appliance
- Fact
    - Most data is physically located in DBMS and new special-purpose appliance
    - Data loads, extract, transform preprocessing operations continue as is
    - Database design for quality assurance



Relational Databases Design

- Relational Database design:
  - The grouping of attributes to form “good” relation schemas
- We have assumed schema R is given
  - Universal Relation: R could have been a single relation containing all attributes that are of interest
  - Normalization breaks R into smaller relations
  - R could have been the results of some ad-hoc design of relations, which we then test/convert to normal form

The Need for Normalization

- Example:
  - Charges its clients by billing hours spent on each contract
  - Hourly billing rate is dependent on employee's position

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

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

- Mixing attributes of multiple entities may cause problems
  - Information is stored redundantly wasting storage
  - Problems with update anomalies
    - Insertion anomalies
    - Deletion anomalies
    - Modification anomalies
- The report may yield different results depending on data anomaly
  - Primary keys
  - Redundancy
  - Possible data inconsistencies
    - E.g. JOB_CLASS: Elect Engineer, Elect. Eng., E. E., El. Eng.

Normalization

- Process of decomposing unsatisfactory “bad” relations by breaking up their attributes into smaller relations
- Normal Form:
  - Condition using keys and functional dependencies (FDs) of a relation to certify whether a relation schema is in a particular normal form
  - 2NF, 3NF, BCNF based on keys and FDs of a relation schema
  - 4NF based on keys, multi-valued dependencies

guarantees

Normal form (1NF)

Second normal form (2NF)

Third normal form (3NF)

Boyce-Codd Normal Form (BCNF)

stronger

Functional Dependencies (FDs)

- FDs are used to specify formal measures of the “goodness” of relational designs
- A set of attributes X functionally determines a set of attributes Y if the value of X determines a unique value Y
- $X \rightarrow Y$ holds if whenever two tuples have the same value for X, they must have the same value for Y on all relation instances

$$\forall t_1, t_2 \in \mathrm{r}(R), \quad \mathrm{t}_1[\mathrm{X}] = \mathrm{t}_2[\mathrm{X}] \implies \mathrm{t}_1[\mathrm{Y}] = \mathrm{t}_2[\mathrm{Y}]$$

- FDs are derived from the real-world constraints on the attributes
  - SSN → ENAME
  - PNUBER → {PNAME, PLOCATION}
  - {SSN, PNUMBER} → HOURS
  - What if LHS is a key?

Function Key (Cont'd)

- $A \rightarrow C$
- $< A,\ B > \ \ \rightarrow C$
- $< A, C > \rightarrow\ D$
- $A \rightarrow\ D$
- $C \rightarrow\ D$

A	B	C	D
1	1	3	2
1	2	2	3
1	3	3	2

- Full/Partial FD
  - If removal of any attribute from X means the FD does not hold any more, it is a full dependency; otherwise it's a Partial dependency
- Transitive FD: $X \rightarrow Y$
  - If there is a set of attributes Z that are neither a primary or candidate key and both $X \rightarrow Z$ and $Z \rightarrow Y$ holds.

Inference Rules for FDs

- Given a set of FDs F, we can infer additional FDs that hold whenever the FDs in F holds.
- Armstrong's inference rules
  - Reflexive
    if $Y \subseteq X$, then $X \rightarrow Y$
  - Argumentation
    if $X \rightarrow Y$, then $X \cup Z \rightarrow Y \cup Z$ (or sometime written as $XZ \rightarrow YZ$)
  - Transitive
    if $X \rightarrow Y$ and $\mathrm{Y} \rightarrow Z$, then $X \rightarrow Z$

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
- Disallow composite attributes, multi-valued attributes and nested relations
- 1NF deals with the “shape” of the tables

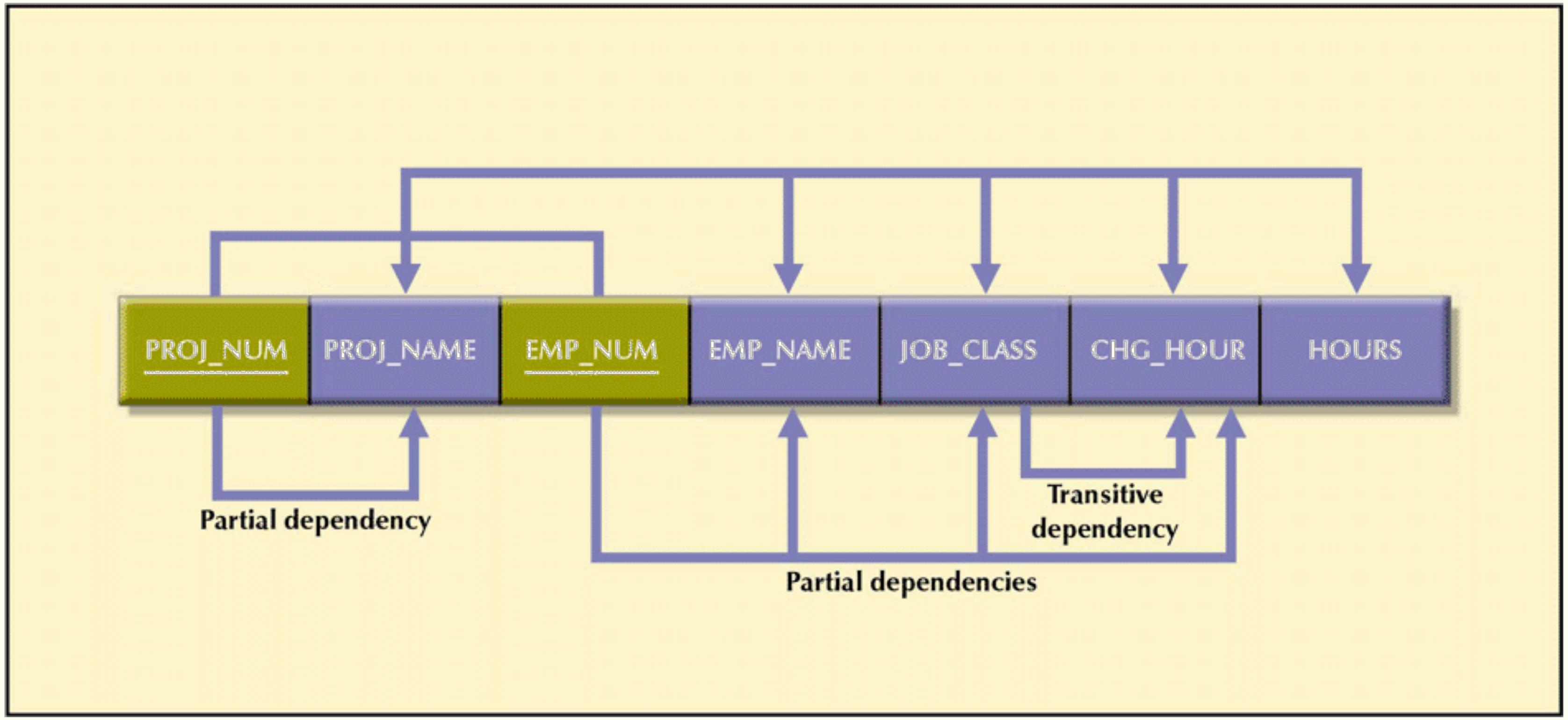
1NF Normalization

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

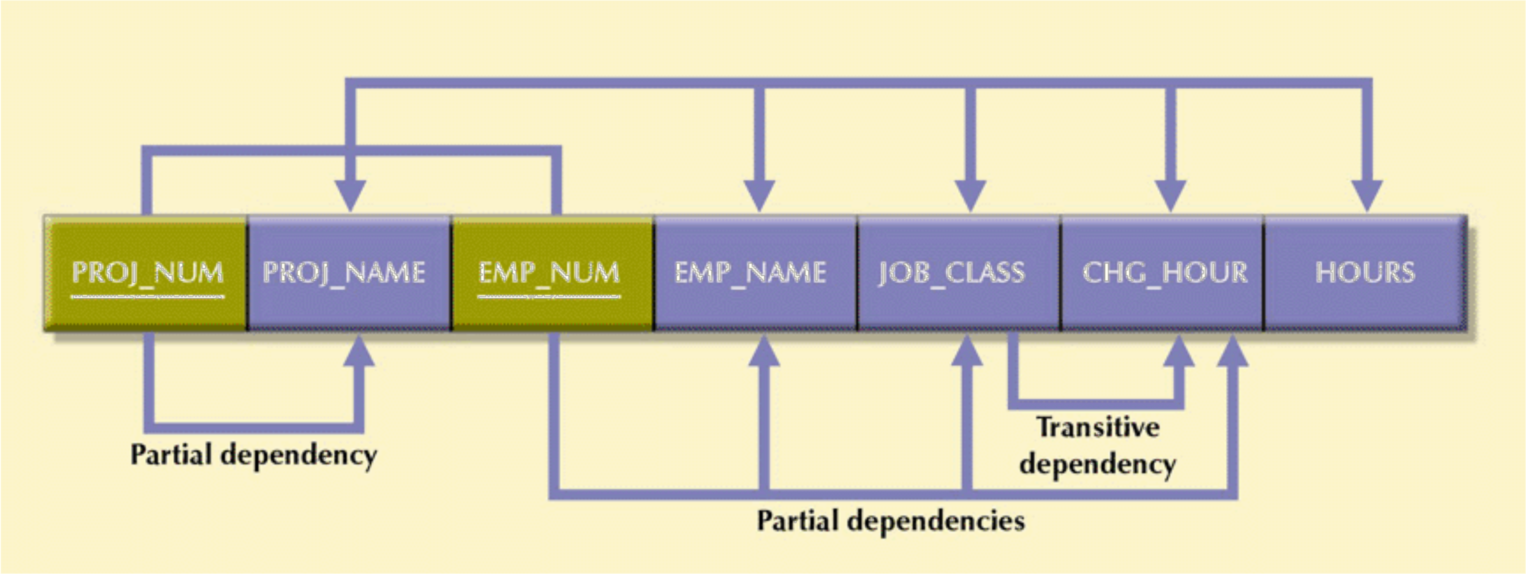
FD Diagram

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



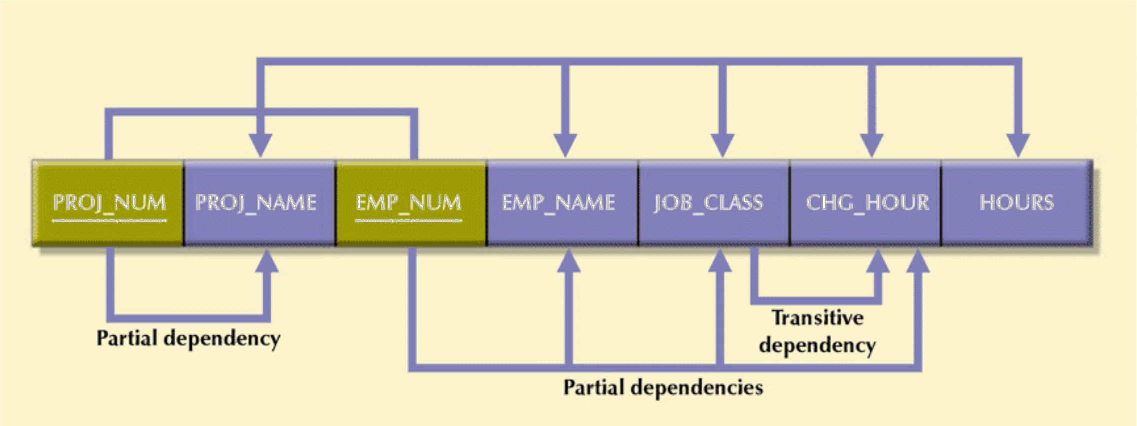
Second Norm Form (2NF)

- Table is in 2NF if
  - It is in 1NF
  - It includes no partial dependencies
    - No attribute is dependent on only a portion of the primary key
    - Every attribute A not in PK is fully functionally dependent on PK
- 2NF deals with the relationship between non-key and key attributes



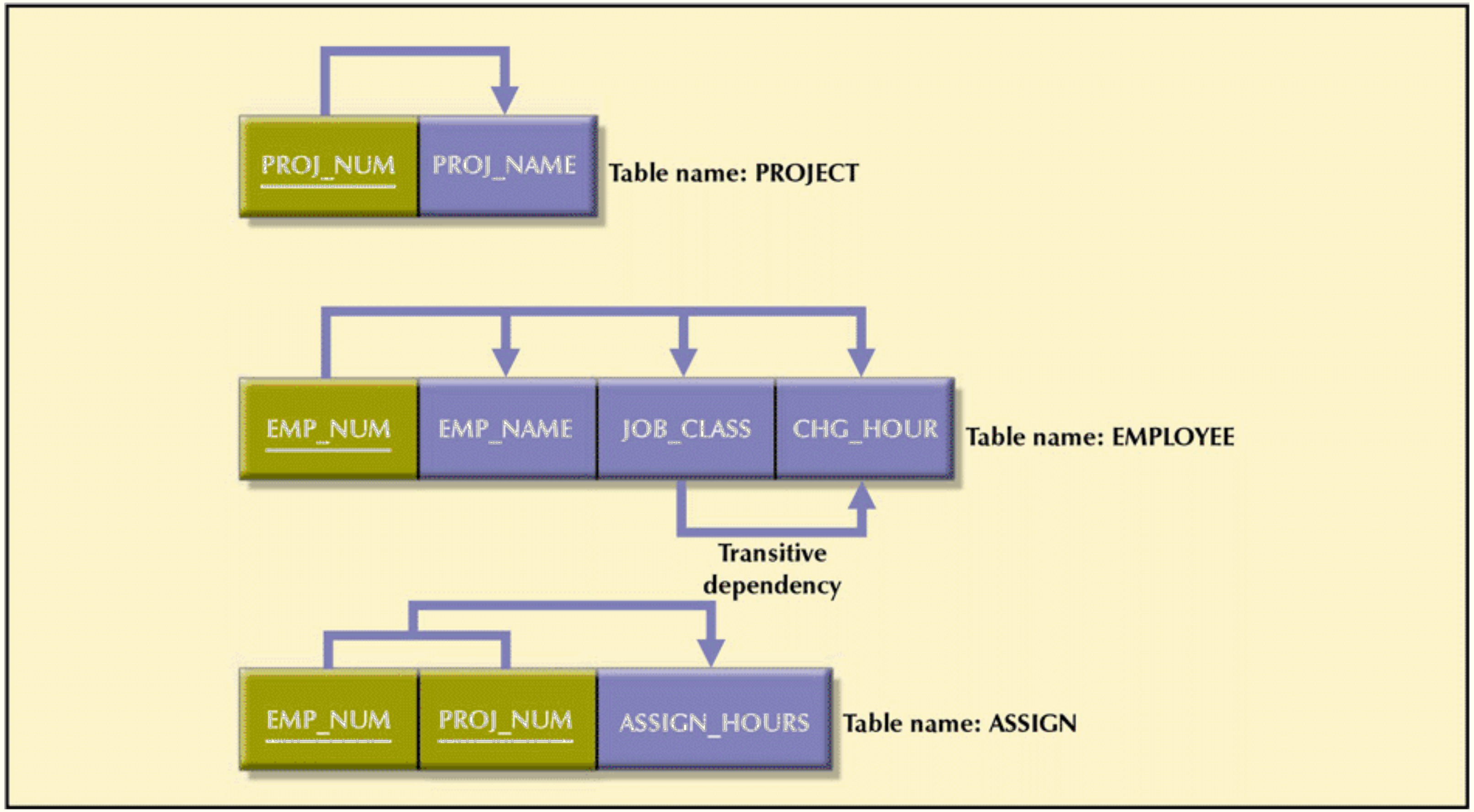
Conversion 1NF to 2NF

- Step 1: Write each key attribute 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: Determine which attributes are dependent on which other attributes (remove anomalies)



2NF Results

FIGURE 5.4 SECOND NORMAL FORM (2NF) CONVERSION RESULTS

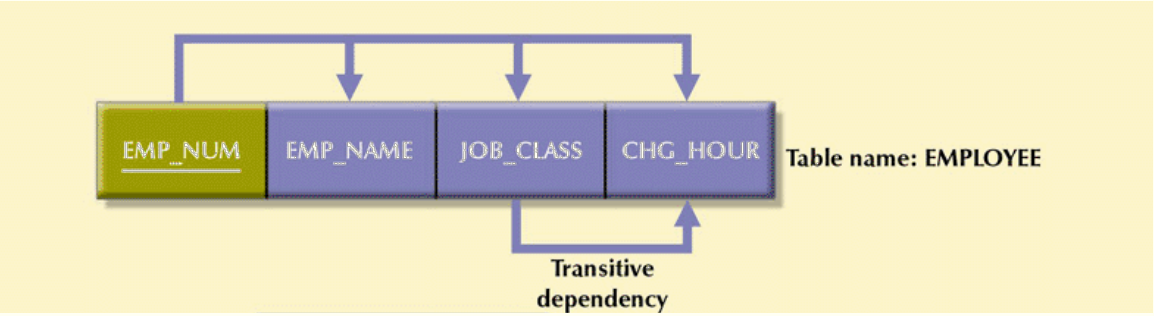


Third Normal Form (3NF)

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

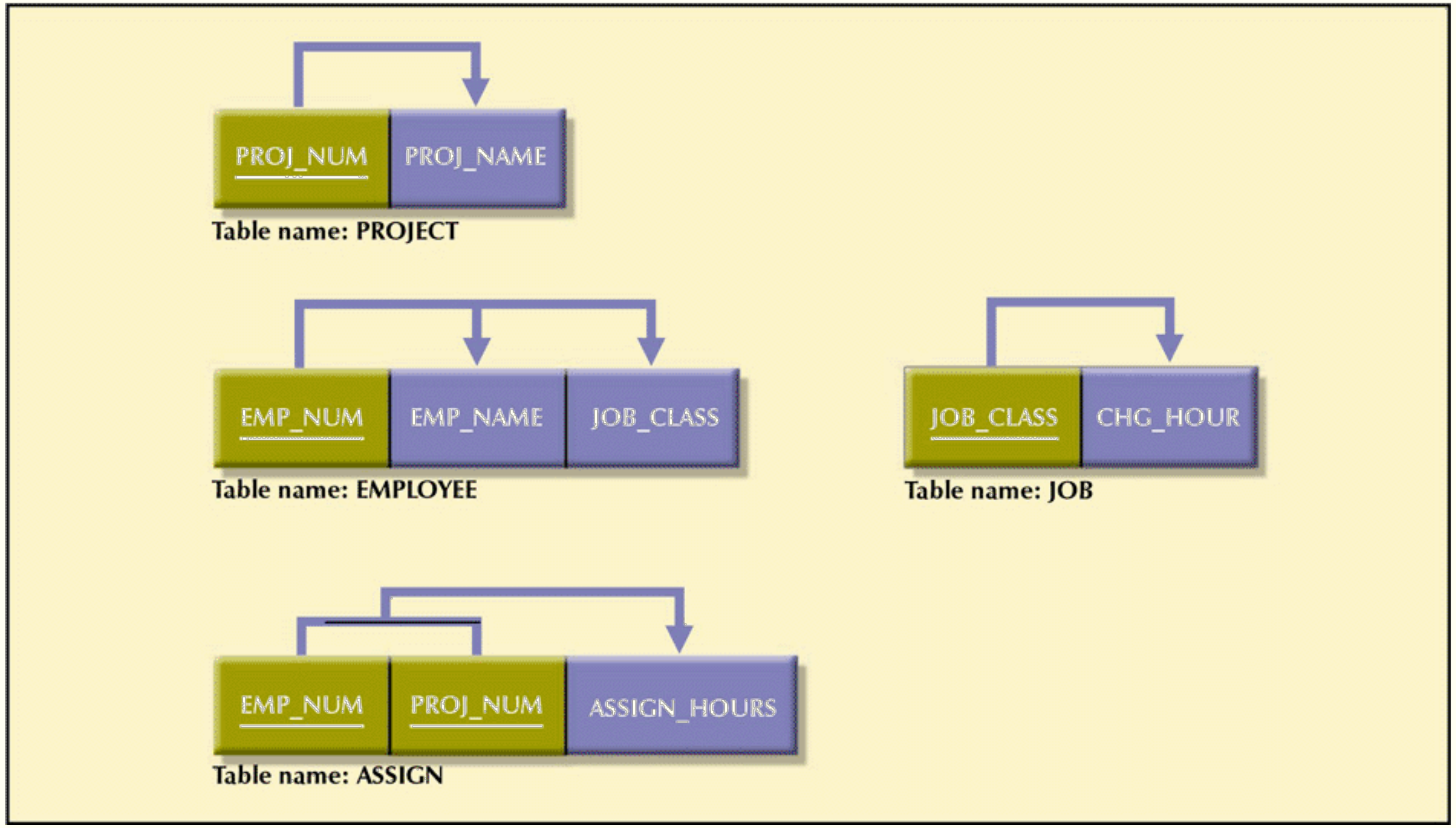
Conversion to 3NF

- Step 1: Find new Fact
  - For every transitive dependency $X \to Y$, write fact Z as a PK for a new table where $X \to Z$ and $Z \to Y$
- Step 2: Identify the dependent attributes
  - Identify the attributes dependent on each Z identified in previous step and find the dependency
  - Name the table to reflect its contents and function
- Step 3: Remove $X \to Y$ from original table



3NF Results

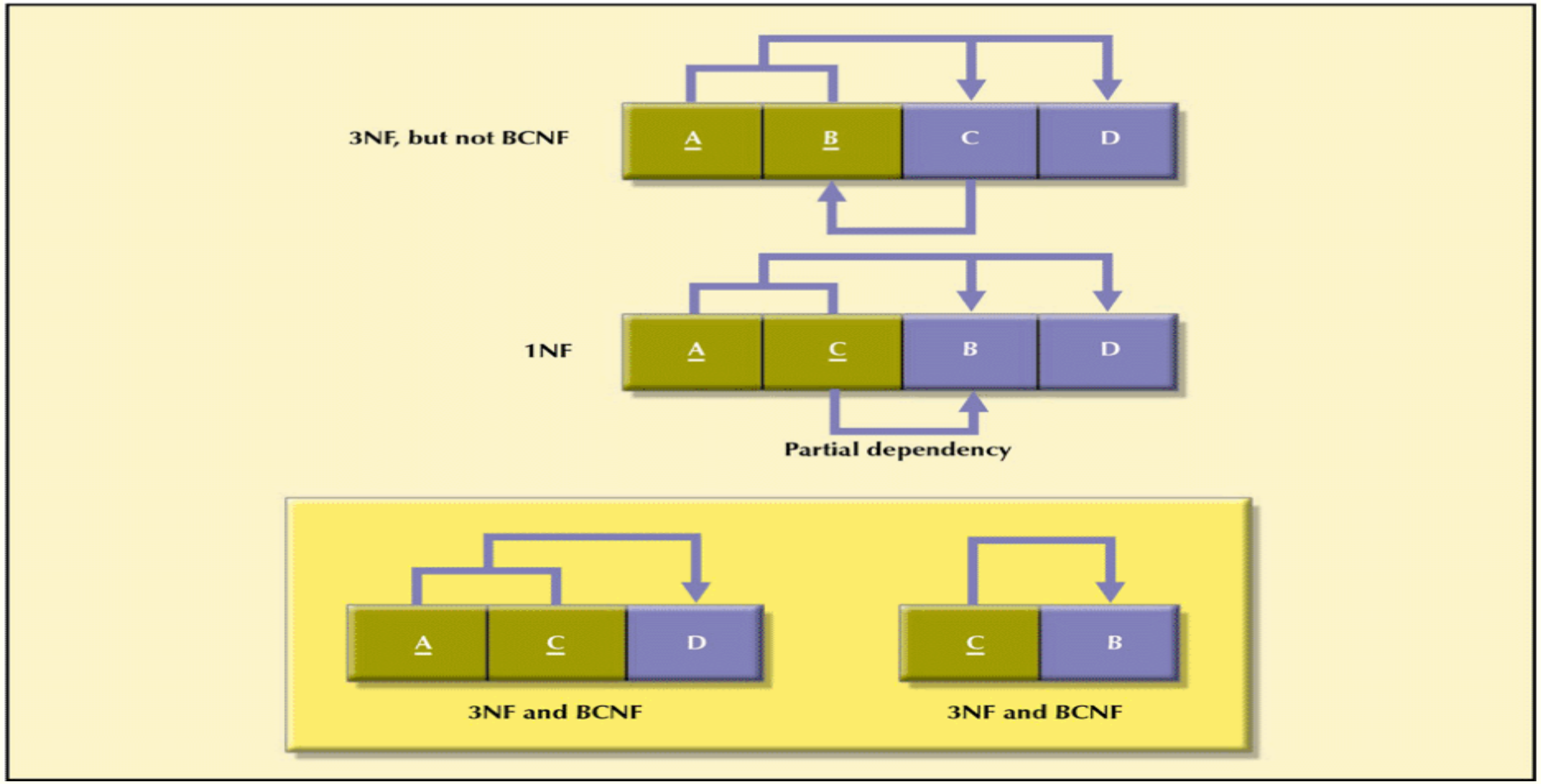
FIGURE 5.5 THIRD NORMAL FORM (3NF) CONVERSION RESULTS



Boyce-Codd Normal Form (BCNF)

- A relation schema R is in BCNF, a.k.a. 3.5NF, if whenever an FD $X \to A$ holds in R, then X is a super-key of R

FIGURE 5.8 Decomposition to BCNF

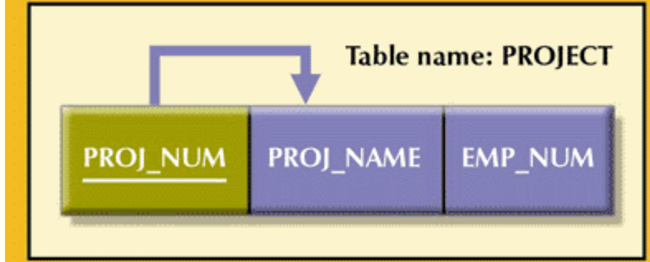


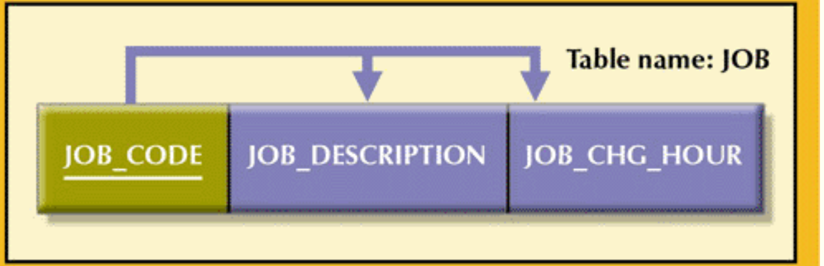
Normalization

- Each normal form is strictly stronger than the previous one
  - Every 2NF relation is in 1NF
  - Every 3NF relation is in 2NF
  - Every BCNF relation is in 3NF
- There exist relations that in 3NF but not in BCNF
- The goal is to have each relation in BCNF (or 3NF)

The Completed Database

FIGURE 5.6 THE COMPLETED DATABASE





Database name: Ch05_ConstructCo

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: 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

The Completed Database

FIGURE 5.6 THE COMPLETED DATABASE (CONTINUED)

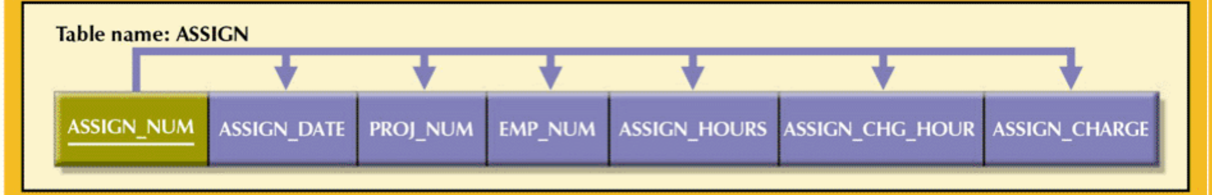


Table name: ASSIGN

Database name: Ch05_ConstructCo

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

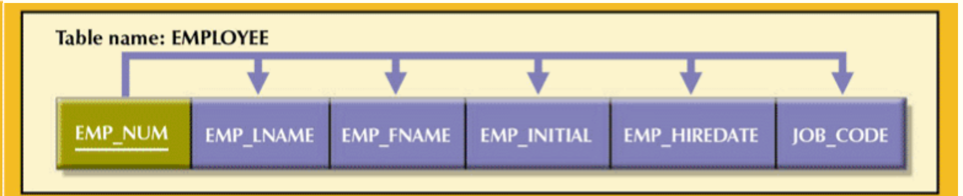


Table name: EMPLOYEE

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