Normalization in Databases

What is Normalization?

- Unnormalized data exists in flat files

- Normalization is the process of moving data into related tables

- This is usually done by running action queries (Make Table and Append queries)….unless you’re starting from scratch – then do it right the first time!

Why Normalize Tables?

- Save typing of repetitive data
- Increase flexibility to query, sort, summarize, and group data (Simpler to manipulate data!)
- Avoid frequent restructuring of tables and other objects to accommodate new data
- Reduce disk space

A Typical Spreadsheet File

Emp No	Employee Name	Time Card No	Time Card Date	Dept No	Dept Name
10	Thomas Arquette	106	11/02/2002	20	Marketing
10	Thomas Arquette	106	11/02/2002	20	Marketing
10	Thomas Arquette	106	11/02/2002	20	Marketing
10	Thomas Arquette	115	11/09/2002	20	Marketing
99	Janice Smitty			10	Accounting
500	Alan Cook	107	11/02/2002	50	Shipping
500	Alan Cook	107	11/02/2002	50	Shipping
700	Ernest Gold	108	11/02/2002	50	Shipping
700	Ernest Gold	116	11/09/2002	50	Shipping
700	Ernest Gold	116	11/09/2002	50	Shipping

Employee, Department, and Time Card Data in Three Tables

Table: Employees

EmpNo	EmpFirstName	EmpLastName	DeptNo
10	Thomas	Arquette	20
500	Alan	Cook	50
700	Ernest	Gold	50
99	Janice	Smitty	10

Table: Departments

DeptNo	DeptName
10	Accounting
20	Marketing
50	Shipping

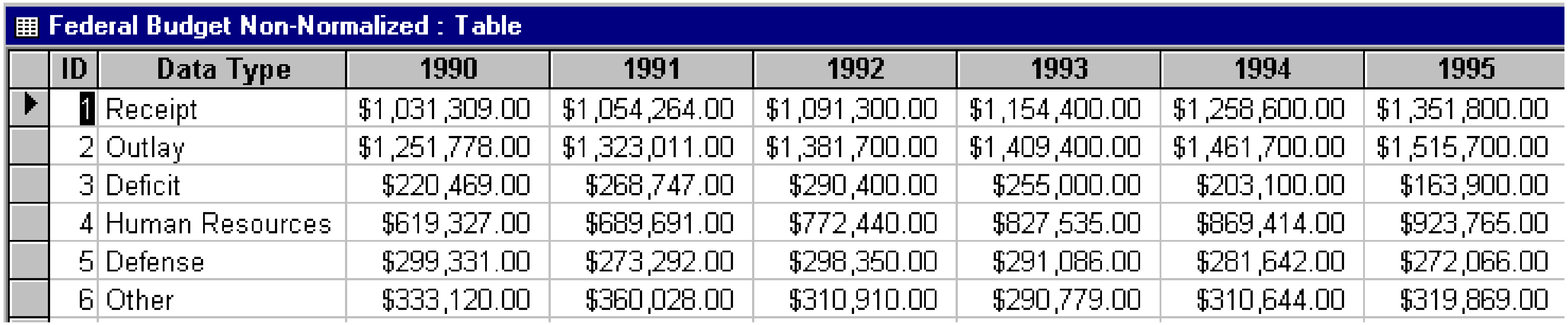
Table: Time Card Data

TimeCardNo	EmpNo	TimeCardDate
106	10	11/02/2002
107	500	11/02/2002
108	700	11/02/2002
115	10	11/09/2002
116	700	11/09/2002

🟦 Primary Key

Another Example of Normalizing

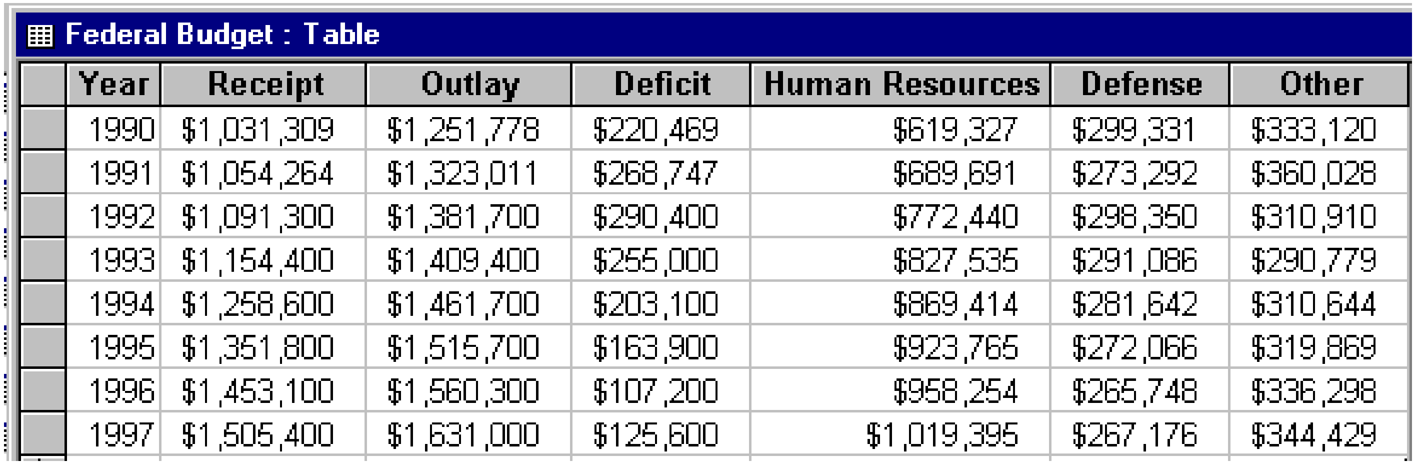
Non-Normalized Table



Federal Budget Non-Normalized : Table

	ID	Data Type	1990	1991	1992	1993	1994	1995
▶	1	Receipt	$1,031,309.00	$1,054,264.00	$1,091,300.00	$1,154,400.00	$1,258,600.00	$1,351,800.00
	2	Outlay	$1,251,778.00	$1,323,011.00	$1,381,700.00	$1,409,400.00	$1,461,700.00	$1,515,700.00
	3	Deficit	$220,469.00	$268,747.00	$290,400.00	$255,000.00	$203,100.00	$163,900.00
	4	Human Resources	$619,327.00	$689,691.00	$772,440.00	$827,535.00	$869,414.00	$923,765.00
	5	Defense	$299,331.00	$273,292.00	$298,350.00	$291,086.00	$281,642.00	$272,066.00
	6	Other	$333,120.00	$360,028.00	$310,910.00	$290,779.00	$310,644.00	$319,869.00

Normalized Table



Federal Budget : Table

	Year	Receipt	Outlay	Deficit	Human Resources	Defense	Other
	1990	$1,031,309	$1,251,778	$220,469	$619,327	$299,331	$333,120
	1991	$1,054,264	$1,323,011	$268,747	$689,691	$273,292	$360,028
	1992	$1,091,300	$1,381,700	$290,400	$772,440	$298,350	$310,910
	1993	$1,154,400	$1,409,400	$255,000	$827,535	$291,086	$290,779
	1994	$1,258,600	$1,461,700	$203,100	$869,414	$281,642	$310,644
	1995	$1,351,800	$1,515,700	$163,900	$923,765	$272,066	$319,869
	1996	$1,453,100	$1,560,300	$107,200	$958,254	$265,748	$336,298
	1997	$1,505,400	$1,631,000	$125,600	$1,019,395	$267,176	$344,429

Types of Normalization

- ***First Normal Form***
  - each field contains the smallest meaningful value
  - the table does not contain repeating groups of fields or repeating data within the same field
    - Create a separate field/table for each set of related data.
    - Identify each set of related data with a primary key

Tables Violating First Normal Form

PART **(Primary Key)**	**WAREHOUSE**
P0010	Warehouse A, Warehouse B, Warehouse C
P0020	Warehouse B, Warehouse D

Really Bad Set-up!

Better, but still flawed!

PART **(Primary Key)**	**WAREHOUSE A**	**WAREHOUSE B**	**WAREHOUSE C**
P0010	Yes	No	Yes
P0020	No	Yes	Yes

Table Conforming to First Normal Form

PART (Primary Key)	WAREHOUSE (Primary Key)	QUANTITY
P0010	Warehouse A	400
P0010	Warehouse B	543
P0010	Warehouse C	329
P0020	Warehouse B	200
P0020	Warehouse D	278

- ***Second Normal Form***
  - usually used in tables with a multiple-field primary key (composite key)
  - each non-key field relates to the entire primary key
  - any field that does not relate to the primary key is placed in a separate table
  - MAIN POINT –
    - eliminate redundant data in a table
    - Create separate tables for sets of values that apply to multiple records

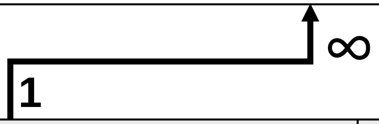
Table Violating Second Normal Form

PART (Primary Key)	WAREHOUSE (Primary Key)	QUANTITY	WAREHOUSE ADDRESS
P0010	Warehouse A	400	1608 New Field Road
P0010	Warehouse B	543	4141 Greenway Drive
P0010	Warehouse C	329	171 Pine Lane
P0020	Warehouse B	200	4141 Greenway Drive
P0020	Warehouse D	278	800 Massey Street

Tables Conforming to Second Normal Form

PART_STOCK TABLE

PART (Primary Key)	WAREHOUSE (Primary Key)	QUANTITY
P0010	Warehouse A	400
P0010	Warehouse B	543
P0010	Warehouse C	329
P0020	Warehouse B	200
P0020	Warehouse D	278



WAREHOUSE TABLE

WAREHOUSE (Primary Key)	WAREHOUSE_ADDRESS
Warehouse A	1608 New Field Road
Warehouse B	4141 Greenway Drive
Warehouse C	171 Pine Lane
Warehouse D	800 Massey Street

- ***Third Normal Form***
  - usually used in tables with a single-field primary key
  - records do not depend on anything other than a table's primary key
  - each non-key field is a fact about the key
  - Values in a record that are not part of that record's key do not belong in the table. In general, any time the contents of a group of fields may apply to more than a single record in the table, consider placing those fields in a separate table.

Table Violating Third Normal Form

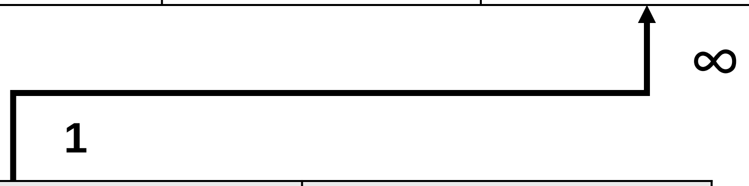
EMPLOYEE_DEPARTMENT TABLE

EMPNO (Primary Key)	**FIRSTNAME**	**LASTNAME**	**WORKDEPT**	**DEPTNAME**
000290	John	Parker	E11	Operations
000320	Ramlal	Mehta	E21	Software Support
000310	Maude	Setright	E11	Operations

Tables Conforming to Third Normal Form

EMPLOYEE TABLE

EMPNO (Primary Key)	FIRSTNAME	LASTNAME	WORKDEPT
000290	John	Parker	E11
000320	Ramlal	Mehta	E21
000310	Maude	Setright	E11



DEPARTMENT TABLE

DEPTNO (Primary Key)	DEPTNAME
E11	Operations
E21	Software Support

Example 1

- Un-normalized Table:

Student#	Advisor#	Advisor	Adv-Room	Class1	Class2	Class3
1022	10	Susan Jones	412	101-07	143-01	159-02
4123	12	Anne Smith	216	101-07	159-02	214-01

- Table in First Normal Form
  - No Repeating Fields
  - Data in Smallest Parts

Student#	Advisor#	AdvisorFName	AdvisorLName	Adv-Room	Class#
1022	10	Susan	Jones	412	101-07
1022	10	Susan	Jones	412	143-01
1022	10	Susan	Jones	412	159-02
4123	12	Anne	Smith	216	101-07
4123	12	Anne	Smith	216	159-02
4123	12	Anne	Smith	216	214-01

- Tables in Second Normal Form
  - Redundant Data Eliminated

Table: Students

Student#	Advisor#	AdvFirstName	AdvLastName	Adv-Room
1022	10	Susan	Jones	412
4123	12	Anne	Smith	216

Table: Registration

Student#	Class#
1022	101-07
1022	143-01
1022	159-02
4123	201-01
4123	211-02
4123	214-01

- Tables in Third Normal Form
  - – Data Not Dependent On Key is Eliminated

Table: Advisors

Advisor#	AdvFirstName	AdvLastName	Adv-Room
10	Susan	Jones	412
12	Anne	Smith	216

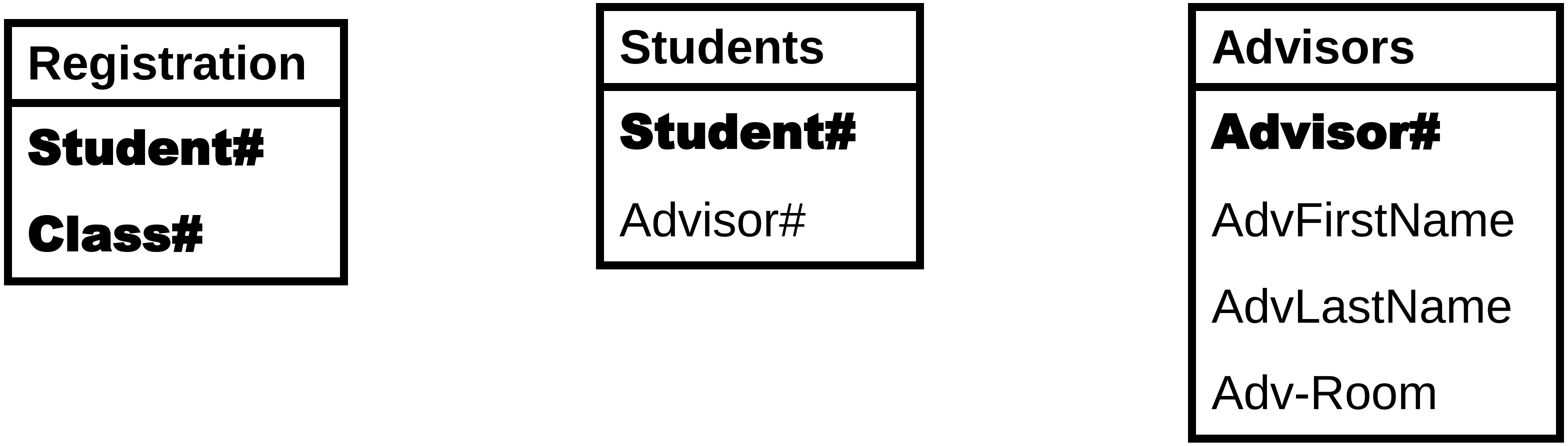
Table: Students

Student#	Advisor#	StudentFName	StudentLName
1022	10	Jane	Mayo
4123	12	Mark	Baker

Table: Registration

Student#	Class#
1022	101-07
1022	143-01
1022	159-02
4123	201-01
4123	211-02
4123	214-01

Relationships for Example 1



Registration

Student#

Class#

Students

Student#

Advisor#

Advisors

Advisor#

AdvFirstName

AdvLastName

Adv-Room

Example 2

- Un-normalized Table:

EmpID	Name	Dept Code	Dept Name	Proj 1	Time Proj 1	Proj 2	Time Proj 2	Proj 3	Time Proj 3
EN1-26	Sean Breen	TW	Technical Writing	30-T3	25%	30-TC	40%	31-T3	30%
EN1-33	Amy Guya	TW	Technical Writing	30-T3	50%	30-TC	35%	31-T3	60%
EN1-36	Liz Roslyn	AC	Accounting	35-TC	90%				

Table in First Normal Form

EmpID	Project Number	Time on Project	Last Name	First Name	Dept Code	Dept Name
EN1-26	30-T3	25%	Breen	Sean	TW	Technical Writing
EN1-26	30-TC	40%	Breen	Sean	TW	Technical Writing
EN1-26	31-T3	30%	Breen	Sean	TW	Technical Writing
EN1-33	30-T3	50%	Guya	Amy	TW	Technical Writing
EN1-33	30-TC	35%	Guya	Amy	TW	Technical Writing
EN1-33	31-T3	60%	Guya	Amy	TW	Technical Writing
EN1-36	35-TC	90%	Roslyn	Liz	AC	Accounting

Tables in Second Normal Form

Table: Employees and Projects

EmpID	Project Number	Time on Project
EN1-26	30-T3	25%
EN1-26	30-T3	40%
EN1-26	31-T3	30%
EN1-33	30-T3	50%
EN1-33	30-TC	35%
EN1-33	31-T3	60%
EN1-36	35-TC	90%

Table: Employees

EmpID	Last Name	First Name	Dept Code	Dept Name
EN1-26	Breen	Sean	TW	Technical Writing
EN1-33	Guya	Amy	TW	Technical Writing
EN1-36	Roslyn	Liz	AC	Accounting

Tables in Third Normal Form

Table: Employees_and_Projects

EmpID	Project Number	Time on Project
EN1-26	30-T3	25%
EN1-26	30-T3	40%
EN1-26	31-T3	30%
EN1-33	30-T3	50%
EN1-33	30-TC	35%
EN1-33	31-T3	60%
EN1-36	35-TC	90%

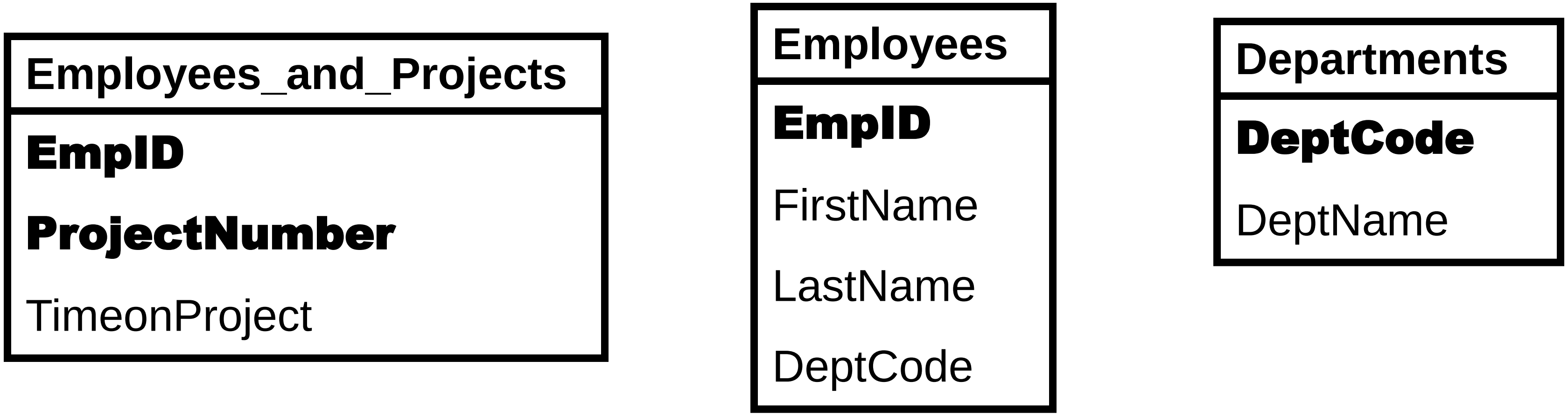
Table: Employees

EmpID	Last Name	First Name	Dept Code
EN1-26	Breen	Sean	TW
EN1-33	Guya	Amy	TW
EN1-36	Roslyn	Liz	AC

Table: Departments

Dept Code	Dept Name
TW	Technical Writing
AC	Accounting

Relationships for Example 2



Employees_and_Projects

EmpID

ProjectNumber

TimeonProject

Employees

EmpID

FirstName

LastName

DeptCode

Departments

DeptCode

DeptName

Example 3

- Un-normalized Table:

EmpID	**Name**	**Manager**	**Dept**	**Sector**	**Spouse/Children**
285	Carl Carlson	Smithers	Engineering	6G	
365	Lenny	Smithers	Marketing	8G	
458	Homer Simpson	Mr. Burns	Safety	7G	Marge, Bart, Lisa, Maggie

Table in First Normal Form

Fields contain smallest meaningful values

EmpID	FName	LName	Manager	Dept	Sector	Spouse	Child1	Child2	Child3
285	Carl	Carlson	Smithers	Eng.	6G				
365	Lenny		Smithers	Marketing	8G				
458	Homer	Simpson	Mr. Burns	Safety	7G	Marge	Bart	Lisa	Maggie

Table in First Normal Form

No more repeated fields

EmpID	FName	LName	Manager	Department	Sector	Dependent
285	Carl	Carlson	Smithers	Engineering	6G	
365	Lenny		Smithers	Marketing	8G	
458	Homer	Simpson	Mr. Burns	Safety	7G	Marge
458	Homer	Simpson	Mr. Burns	Safety	7G	Bart
458	Homer	Simpson	Mr. Burns	Safety	7G	Lisa
458	Homer	Simpson	Mr. Burns	Safety	7G	Maggie

Second/Third Normal Form
Remove Repeated Data From Table
Step 1

EmpID	**FName**	**LName**	**Manager**	**Department**	**Sector**
285	Carl	Carlson	Smithers	Engineering	6G
365	Lenny		Smithers	Marketing	8G
458	Homer	Simpson	Mr. Burns	Safety	7G

EmpID	Dependent
458	Marge
458	Bart
458	Lisa
458	Maggie

Tables in Second Normal Form

Removed Repeated Data From Table
Step 2

EmpID	FName	LName	ManagerID	Dept	Sector
285	Carl	Carlson	2	Engineering	6G
365	Lenny		2	Marketing	8G
458	Homer	Simpson	1	Safety	7G

EmpID	Dependent
458	Marge
458	Bart
458	Lisa
458	Maggie

ManagerID	Manager
1	Mr. Burns
2	Smithers

Tables in Third Normal Form

Employees Table

EmpID	FName	LName	DeptCode
285	Carl	Carlson	EN
365	Lenny		MK
458	Homer	Simpson	SF

Manager Table

ManagerID	Manager
1	Mr. Burns
2	Smithers

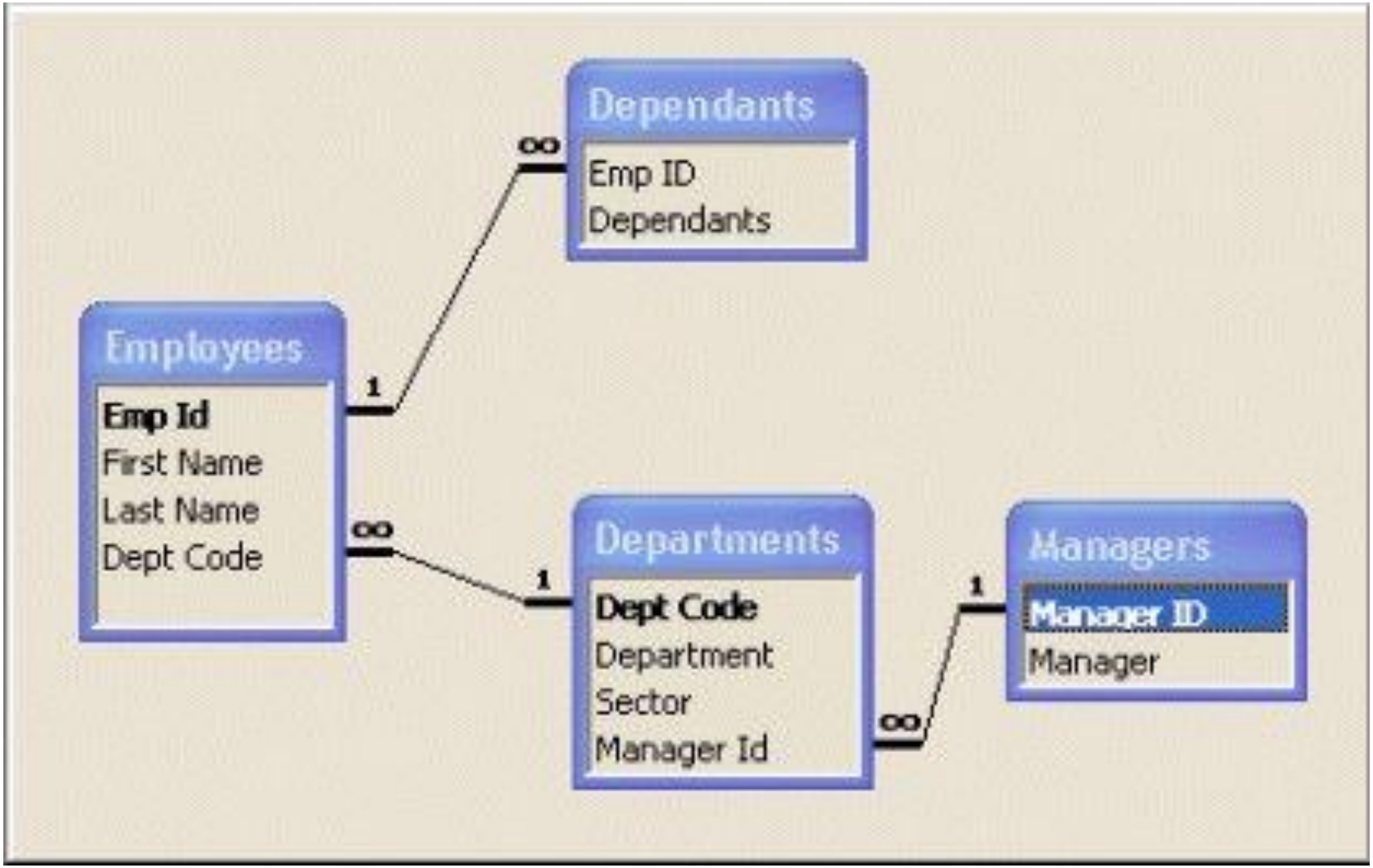
Dependents Table

EmpID	Dependent
458	Marge
458	Bart
458	Lisa
458	Maggie

Department Table

DeptCode	Department	Sector	ManagerID
EN	Engineering	6G	2
MK	Marketing	8G	2
SF	Safety	7G	1

Relationships for Example 3



Example 4

Table Violating 1st Normal Form

Rep ID	Representative	Client 1	Time 1	Client 2	Time 2	Client 3	Time 3
TS-89	Gilroy Gladstone	US Corp.	14 hrs	Taggarts	26 hrs	Kilroy Inc.	9 hrs
RK-56	Mary Mayhem	Italiana	67 hrs	Linkers	2 hrs		

Table in 1st Normal Form

Rep ID	Rep First Name	Rep Last Name	Client ID*	Client	Time With Client
TS-89	Gilroy	Gladstone	978	US Corp	14 hrs
TS-89	Gilroy	Gladstone	665	Taggarts	26 hrs
TS-89	Gilroy	Gladstone	782	Kilroy Inc.	9 hrs
RK-56	Mary	Mayhem	221	Italiana	67 hrs
RK-56	Mary	Mayhem	982	Linkers	2 hrs

Tables in 2nd and 3rd Normal Form

Rep ID*	First Name	Last Name
TS-89	Gilroy	Gladstone
RK-56	Mary	Mayhem

Rep ID*	Client ID*	Time With Client
TS-89	978	14 hrs
TS-89	665	26 hrs
TS-89	782	9 hrs
RK-56	221	67 hrs
RK-56	982	2 hrs
RK-56	665	4 hrs

Client ID*	Client Name
978	US Corp
665	Taggarts
782	Kilroy Inc.
221	Italiana
982	Linkers

This example comes from a tutorial from
http://www.devhood.com/tutorials/tutorial_details.aspx?tutorial_id=95
and
http://www.devhood.com/tutorials/tutorial_details.aspx?tutorial_id=104
Please check them out, as they are very well done.

Example 5

Table in 1st Normal Form

SupplierID	Status	City	PartID	Quantity
S1	20	London	P1	300
S1	20	London	P2	200
S2	10	Paris	P1	300
S2	10	Paris	P2	400
S3	10	Paris	P2	200
S4	20	London	P2	200
S4	20	London	P4	300

Although this table is in 1NF it contains redundant data. For example, information about the supplier's location and the location's status have to be repeated for every part supplied. Redundancy causes what are called *update anomalies*. Update anomalies are problems that arise when information is inserted, deleted, or updated. For example, the following anomalies could occur in this table:

INSERT. The fact that a certain supplier (s5) is located in a particular city (Athens) cannot be added until they supplied a part.
DELETE. If a row is deleted, then not only is the information about quantity and part lost but also information about the supplier.
UPDATE. If supplier s1 moved from London to New York, then two rows would have to be updated with this new information.

Tables in 2NF

Suppliers

SupplierID	Status	City
S1	20	London
S2	10	Paris
S3	10	Paris
S4	20	London
S5	30	Athens

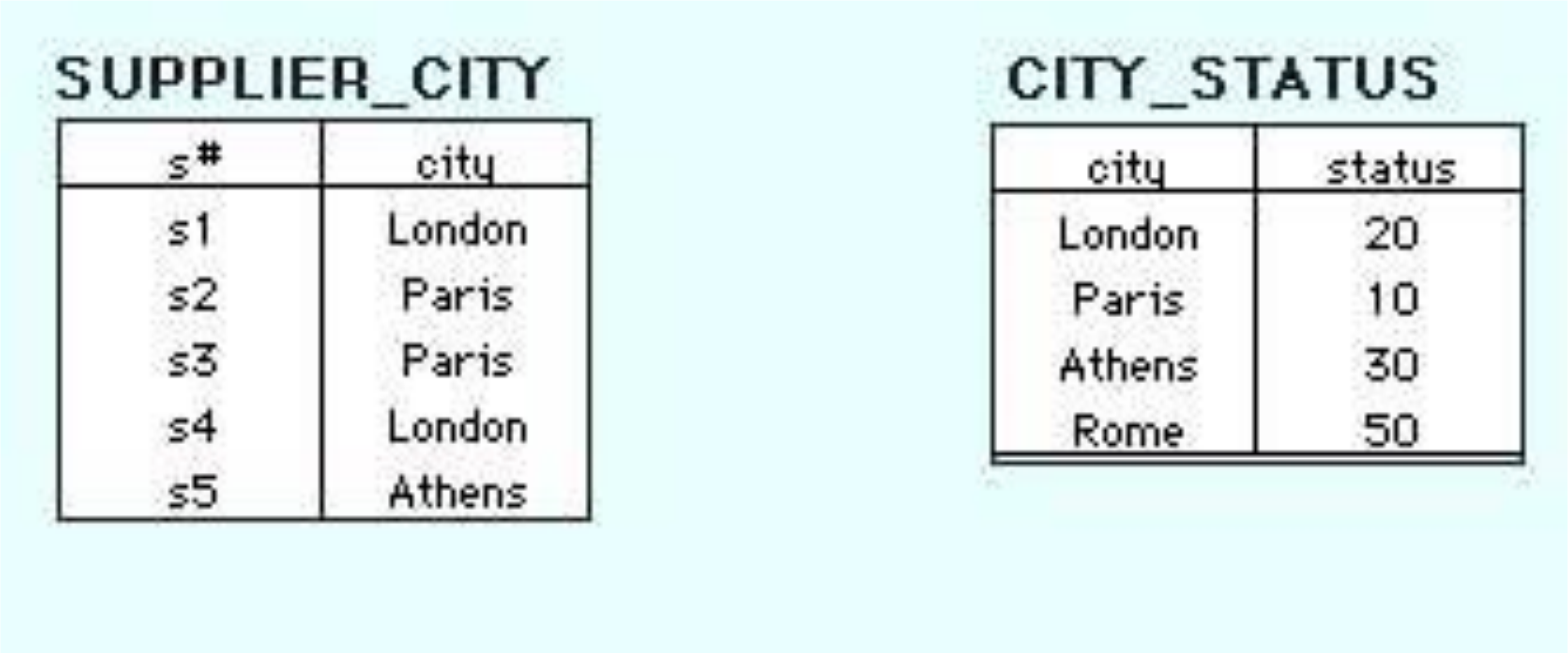
Parts

SupplierID	PartID	Quantity
S1	P1	300
S1	P2	200
S2	P1	300
S2	P2	400
S3	P2	200
S4	P4	300
S4	P5	400

Tables in 2NF but not in 3NF still contain modification anomalies. In the example of Suppliers, they are:

INSERT. The fact that a particular city has a certain status (Rome has a status of 50) cannot be inserted until there is a supplier in the city.
DELETE. Deleting any row in SUPPLIER destroys the status information about the city as well as the association between supplier and city.

Tables in 3NF



SUPPLIER_CITY

s#	city
s1	London
s2	Paris
s3	Paris
s4	London
s5	Athens

CITY_STATUS

city	status
London	20
Paris	10
Athens	30
Rome	50

Advantages of Third Normal Form
The advantage of having relational tables in 3NF is that it eliminates redundant data which in turn saves space and reduces manipulation anomalies. For example, the improvements to our sample database are:

INSERT. Facts about the status of a city, Rome has a status of 50, can be added even though there is not supplier in that city. Likewise, facts about new suppliers can be added even though they have not yet supplied parts.
DELETE. Information about parts supplied can be deleted without destroying information about a supplier or a city.
UPDATE. Changing the location of a supplier or the status of a city requires modifying only one row.

Additional Notes About Example 3

- Going to extremes can create too many tables which in turn can make it difficult to manage your data. The key to developing an efficient database is to determine your needs.
- A postal carrier may need an Address field broken down into smaller fields for sorting and grouping purposes, but do you?
- Another good example is Example 3 - leaving the Dept Code field in our completed table design. If you also wanted to track information such as pay rate, health insurance, etc., then a new table that contains company related data for the employee would be necessary. If all you need is to track the department an employee belongs to then leaving it in the Employees table is fine.

In Summary

- If you type a data value more than once then consider placing the field in another table.
- Consider your sorting and grouping needs. If you need to sort or group on a portion of a field, then the field is not broken down into its smallest meaningful value.
- If you have multiple groups of fields, such as several telephone numbers, then consider eliminating those fields and turning them into records in another table. Think vertically—not horizontally!