

Benefits of Normalization

Normalization ensures that data is consistent with minimal redundancy

- More consistent data with fewer anomalies
- Reduced impact when making changes to the database schema

What Is a Database?

- A database is an organized store of data for:
 - Faster retrieval of data
- Databases usually store data in a way that minimizes **redundancy** to achieve:
 - More efficient data storage
 - Reduced data inconsistency

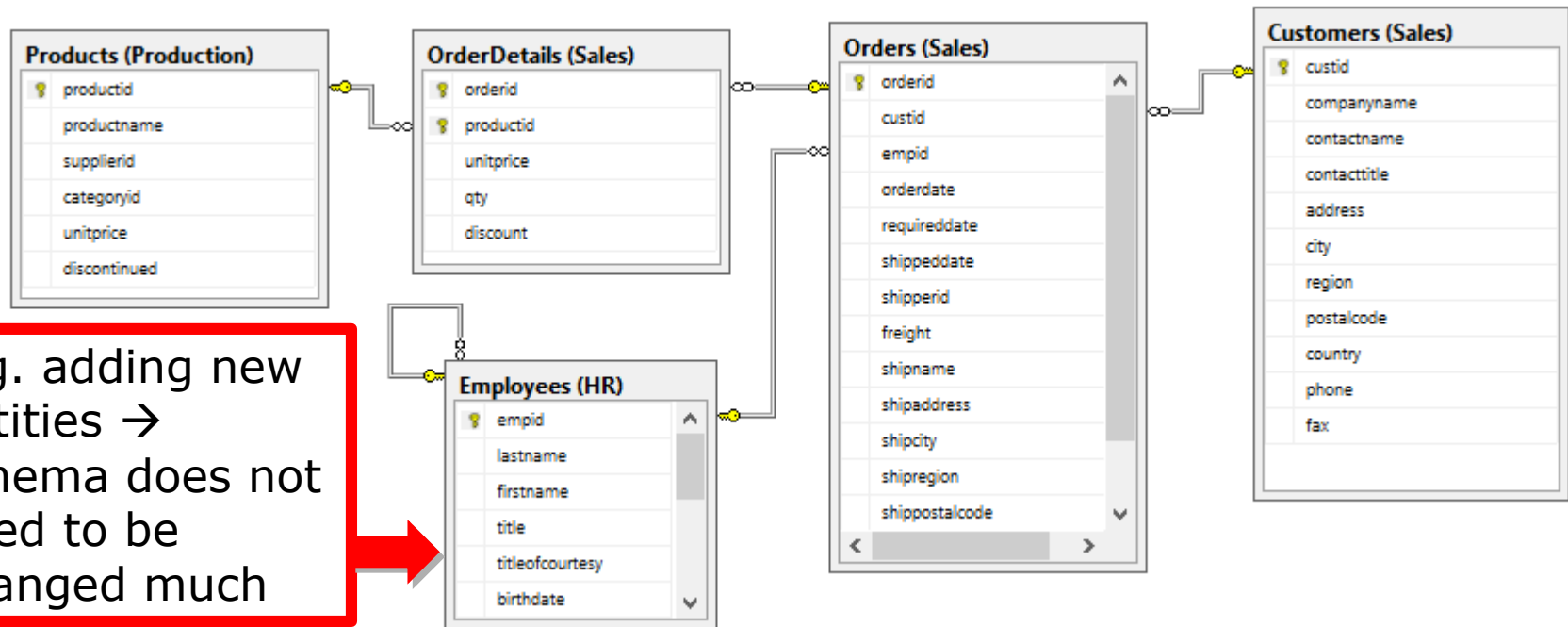
Order ID	Order Date	Customer ID	First Name	Last Name	Country	City
503	2016-03-22	1	Latasha	Navarro	6954 Ranch Rd	Denver
504	2016-03-22	2	Abby	Sai	7074 Spoonwood Court	Seattle
507	2016-03-25	1	Latasha	Navarro	6954 Ranch Rd	Denver
508	2016-03-29	1	Latasha	Navarro	6954 Ranch Rd	Denver

Updating the address:
must update it in all of the locations, to avoid anomalies

Benefits of Normalization

Normalization ensures that data is consistent with minimal redundancy

- More consistent data with fewer anomalies
- Reduced impact when making changes to the database schema
 - Each tables represents one entity
 - The columns are properties that describe that entity

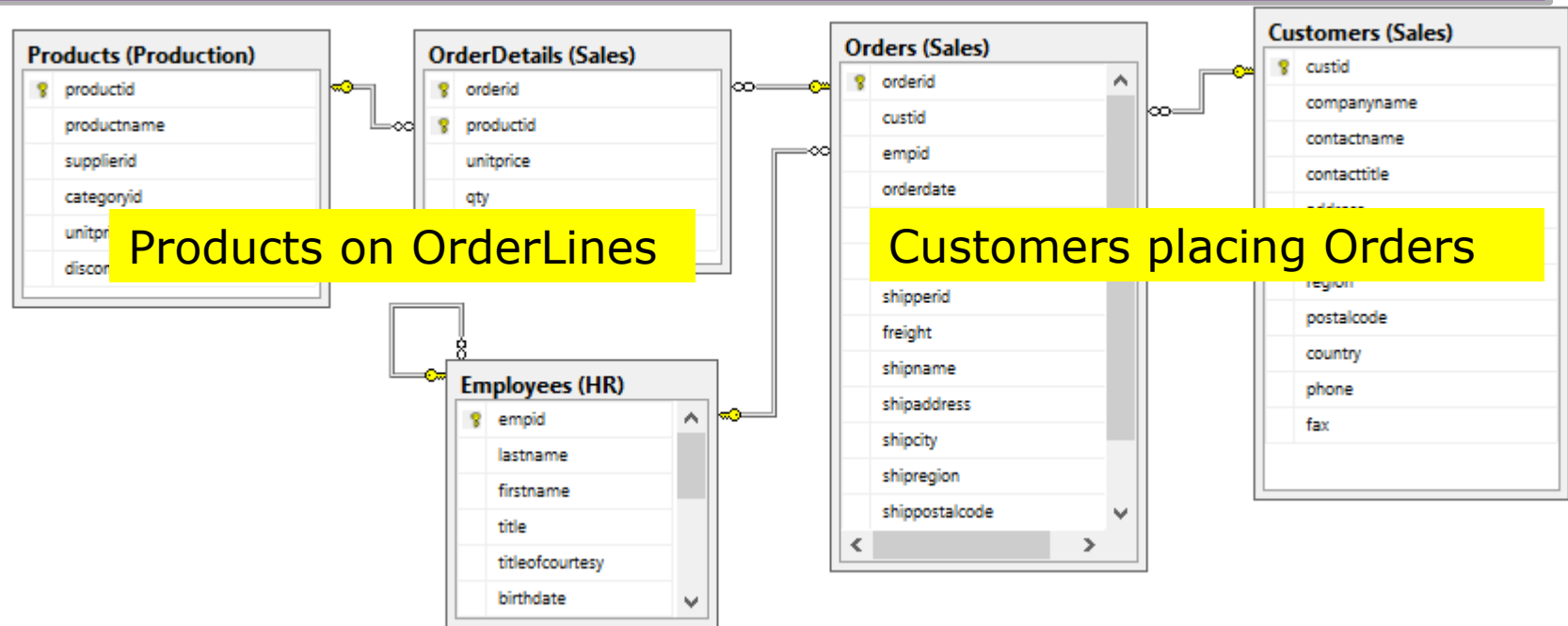


Benefits of Normalization

Normalization ensures that data is consistent with minimal redundancy

- More consistent data with fewer anomalies
- Reduced impact when making changes to the database schema
- An intuitive design that is easy to understand

Normalized database stores data that mirrors real-world processes



Benefits of Normalization

Normalization ensures that data is consistent with minimal redundancy

- More consistent data with fewer anomalies
- Reduced impact when making changes to the database schema
- An intuitive design that is easy to understand
- Improved performance for OLTP workloads
- Databases that require less storage space

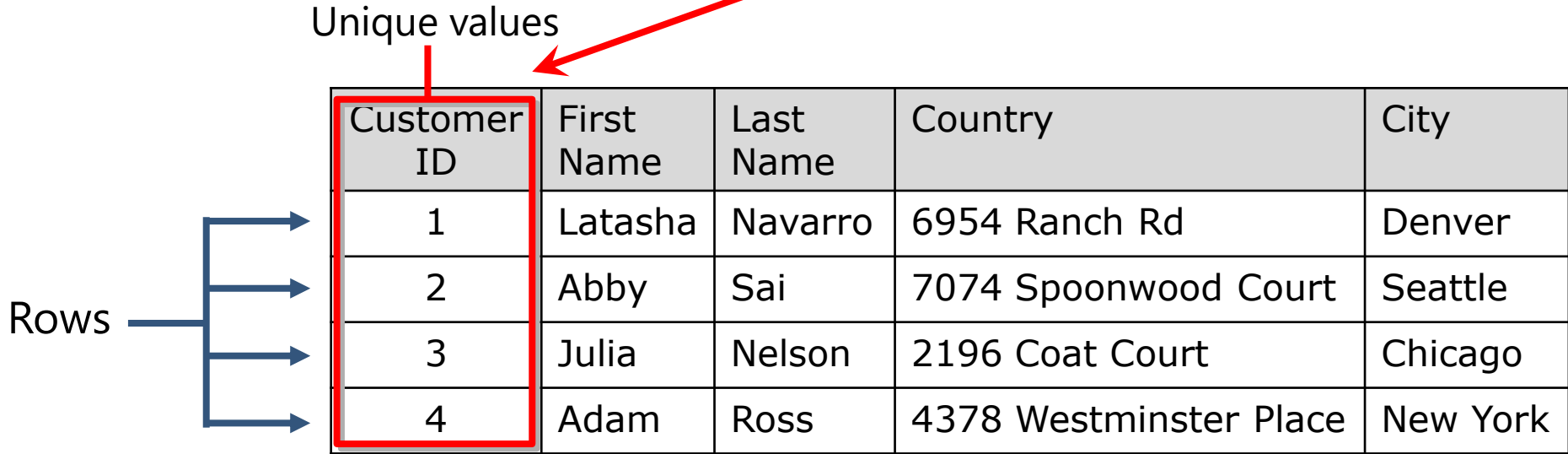
Relational Database Fundamentals

Storing data in tables

Column with unique values:
To unambiguously identify each row in the table
Add a **primary key** constraint to the column

Unique values

Rows



Customer ID	First Name	Last Name	Country	City
1	Latasha	Navarro	6954 Ranch Rd	Denver
2	Abby	Sai	7074 Spoonwood Court	Seattle
3	Julia	Nelson	2196 Coat Court	Chicago
4	Adam	Ross	4378 Westminster Place	New York

A primary key constraint:

- Prevents the addition of any data value to the column if that value already exists in the column
- Compound key:
a primary key that includes more than one column

Keys - Candidate keys

- Candidate key
 - Every column (or combination of columns) that contains unique data values
 - Can potentially serve as the table's primary key
 - N.B.: Candidates should not be configured to allow NULL values

Keys - Surrogate keys

- Surrogate key
 - A column that can be created to serve as a candidate key
 - Has no meaning, in the real world, outside the context of the database
 - Often simply contain integer values that the database generates automatically

Keys

- A primary key uniquely identifies every row in a table
- Candidate keys are the potential primary keys for a table
- Surrogate keys are candidate keys that you can create when there are no other suitable candidate keys

Identifying Candidate Keys

```
SELECT *  
FROM Production.Product;  
GO
```

ProductID	Name	ProductNumber	Color	SafetyStockLevel	ListPrice	Size
1	Adjustable Race	AR-5381	NULL	1000	0,00	NULL
2	Bearing Ball	BA-8327	NULL	1000	0,00	NULL
3	BB Ball Bearing	BE-2349	NULL	800	0,00	NULL
4	Headset Ball Bearings	BE-2908	NULL	800	0,00	NULL
316	Blade	BL-2036	NULL	800	0,00	NULL
317	LL Crankarm	CA-5965	Black	500	0,00	NULL
318	ML Crankarm	CA-6738	Black	500	0,00	NULL
319	HL Crankarm	CA-7457	Black	500	0,00	NULL
320	Chainring Bolts	CB-2903	Silver	1000	0,00	NULL
321	Chainring Nut	CN-6137	Silver	1000	0,00	NULL
322	Chainring	CR-7833	Black	1000	0,00	NULL
323	Crown Race	CR-9981	NULL	1000	0,00	NULL
324	Chain Stays	CS-2812	NULL	1000	0,00	NULL
325	Decal 1	DC-8732	NULL	1000	0,00	NULL

(504 row(s) affected)

Identifying Candidate Keys

```
SELECT DISTINCT ProductID
FROM Production.Product;      (504 row(s) affected)
GO
```

```
SELECT DISTINCT Name
FROM Production.Product;      (504 row(s) affected)
GO
```

```
SELECT DISTINCT ProductNumber
FROM Production.Product;      (504 row(s) affected)
GO
```

```
SELECT DISTINCT Color
FROM Production.Product;      (10 row(s) affected)
GO
```

Color
NULL
Black
Blue
Grey
Multi
Red
Silver
Silver/Black
White
Yellow

Goals of Normalization

- To ensure uniqueness
- To ensure that non-key values in tables have a dependency on the key

Non-key values are the values that are present in a table but which do not form part of the primary key

Example Social Security Number (SSN) is primary key
The values in the other columns are non-key values

SSN	Employee Name	Date Of Birth	Job Title	Phone Number	Department	Salary
134969118	Miller	1970-01-23	Sales Represent ative	555-0412	Sales	56,000
811994146	Margheim	1986-06-05	Marketing Assistant	555-0189	Marketing	48,000

Dependencies

Dependency on the key: means that the values of non-key attributes are determined by the key itself.

Dependencies

SSN	Employee Name	Date Of Birth	Job Title	Phone Number	Department	Salary
134969118	Miller	1970-01-23	Sales Representative	555-0412	Sales	56,000
811994146	Margheim	1986-06-05	Marketing Assistant	555-0189	Marketing	48,000

Every time you look up a particular SSN, it will always return the same values for the other columns.

The SSN value in a row **determines** the values in the columns Employee Name, Date of Birth, Job Title, Phone Number, and Department

notation $A \rightarrow C$ the value A determines the value C

The following dependencies exist:

Functional Dependency

- $SSN \rightarrow$ Employee Name
- $SSN \rightarrow$ Date Of Birth
- $SSN \rightarrow$ Job Title
- $SSN \rightarrow$ Phone Number
- $SSN \rightarrow$ Department

Summarizing:

$SSN \rightarrow$ (Employee Name, Date Of Birth, Job Title, Phone Number, Department)

Dependencies

SSN	Employee Name	Date Of Birth	Job Title	Phone Number	Department	Salary
134969118	Miller	1970-01-23	Sales Representative	555-0412	Sales	56,000
811994146	Margheim	1986-06-05	Marketing Assistant	555-0189	Marketing	48,000

Every time you look up a particular Job Title (or Employee Name), it will NOT always return the same value for the SSN column.

The value of Salary is not directly dependent on the key value of SSN, but only **indirectly** dependent on it.

notation SSN → Job Title → Salary

Not the value of SSN, but the value of Job Title determines the value of Salary,

- Job Title → SSN

Is not a valid dependency

- Employee Name → SSN

Is not a valid dependency

- SSN → Salary

Is not a valid dependency

Transitive Dependency

Dependencies

SSN	Employee Name	Date Of Birth	Job Title	Phone Number	Department	Salary
134969118	Miller	1970-01-23	Sales Representative	555-0412	Sales	56,000
811994146	Margheim	1986-06-05	Marketing Assistant	555-0189	Marketing	48,000

Every time you look up a particular Job Title (or Employee Name), it will NOT always return the same value for the SSN column.

The value of Salary is not directly dependent on the key value of SSN, but only **indirectly** dependent on it.

notation SSN → Job Title → Salary

Not the value of SSN, but the value of Job Title determines the value of Salary,

- Job Title → SSN

Is not a valid dependency

- Employee Name → SSN

Is not a valid dependency

- SSN → Salary

Is not a valid dependency

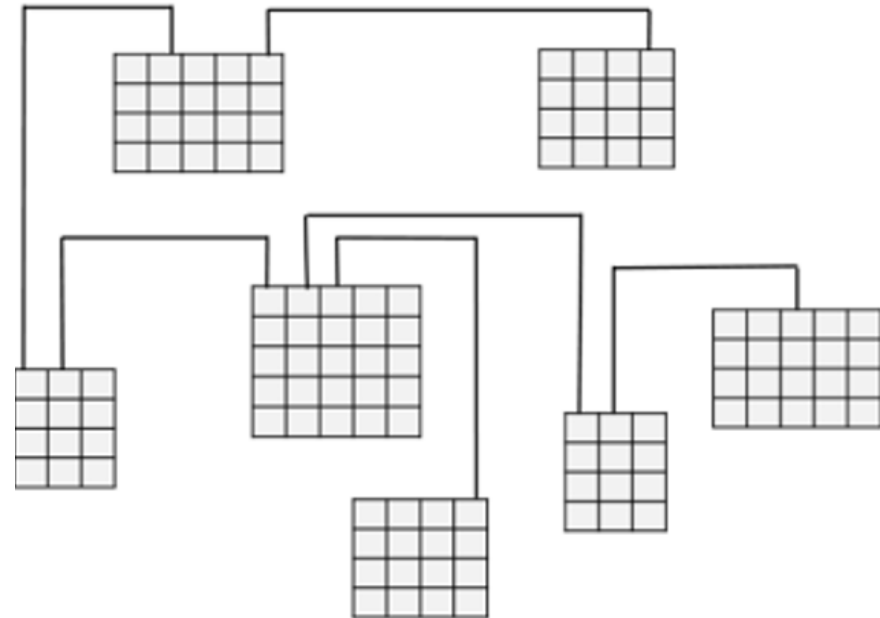
Transitive Dependency

If you remove or change the intermediate value (Job Title), the relationship between the other values (SSN, Salary) is no longer valid: inconsistent data

Introduction to Normalization

- OLTP (online transactional processing)
 - Frequent inserts and updates
- OLTP databases typically perform better when redundant data is minimized
 - Inserting new order:
updating only the tables that contain order data

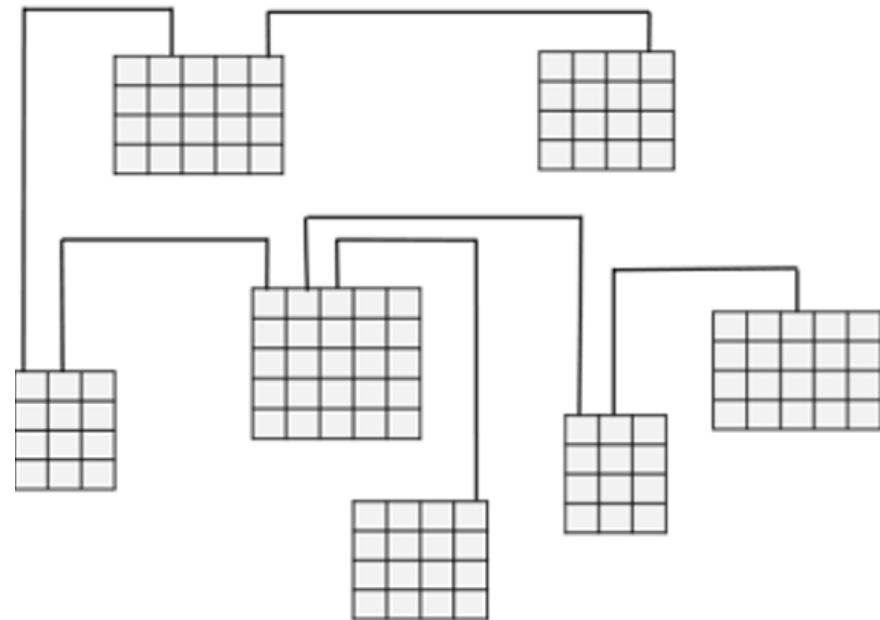
Normalization is the process of reducing or minimizing redundant data in a database



Introduction to Normalization

- Each normalization-step
 - Addresses specific types of data redundancy issues
 - Encompasses the previous one (cumulative steps)
- The steps are called first normal form (1NF), second normal form (2NF), and so on

Normalization is the process of reducing or minimizing redundant data in a database



First Normal Form

- Tables should not contain repeating data groups

**Employees
table not
normalized**

SSN	Employee Name	Job Title	Phone Number 1	Phone Number 2
134969118	Miller	Sales Representative	555-0619	555-0412
811994146	Margheim	Marketing Assistant	555-0124	555-0189

**Repeating data
group**

- Rows should not contain repeating groups within a column:
Data in a column should be atomic (one value per row per column)

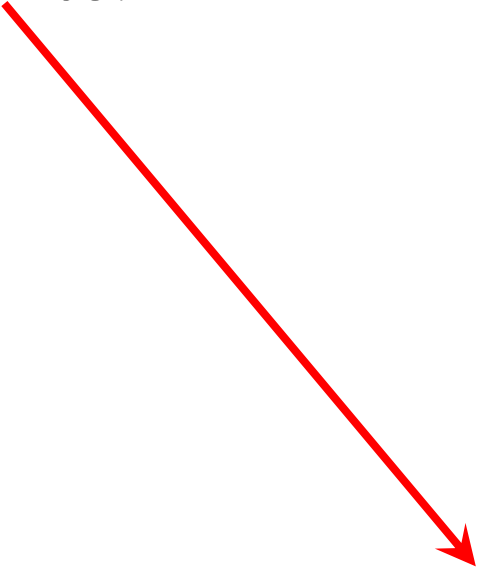
SSN	Employee Name	Job Title	Phone Number
134969118	Miller	Sales Representative	555-0619, 555-0412
811994146	Margheim	Marketing Assistant	555-0124, 555-0189

First Normal Form

- Tables should not contain repeating data groups

Assessment checks:

- Are there any columns that contain data that includes separator values, such as commas?



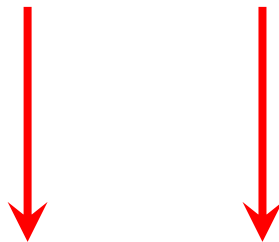
SSN		Phone Number
134969118		555-0619, 555-0412
811994146		555-0124, 555-0189

First Normal Form

- Tables should not contain repeating data groups

Assessment checks:

- Are there any columns that contain data that includes separator values, such as commas?
- Are there any columns with similar names, perhaps differentiated by numbers at the end, such as Address1, Address2, and so on?



SSN		Phone Number 1	Phone Number 2
134969118		555-0619	555-0412
811994146		555-0124	555-0189

SSN		Phone Number
134969118		555-0619, 555-0412
811994146		555-0124, 555-0189

First Normal Form

- Tables should not contain repeating data groups

Assessment checks:

- Are there any columns that contain data that includes separator values, such as commas?
- Are there any columns with similar names, perhaps differentiated by numbers at the end, such as Address1, Address2, and so on?
- Is there a column or set of columns that uniquely identifies each row?

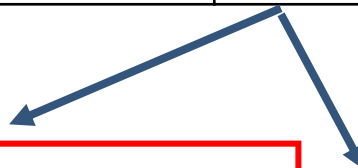
SSN		Phone Number 1	Phone Number 2
134969118		555-0619	555-0412
811994146		555-0124	555-0189

SSN		Phone Number
134969118		555-0619, 555-0412
811994146		555-0124, 555-0189

First Normal Form

**Employees
table not
normalized**

SSN	Employee Name	Job Title	Phone Number 1	Phone Number 2
134969118	Miller	Sales Represent ative	555-0619	555-0412
811994146	Margheim	Marketing Assistant	555-0124	555-0189



SSN	Employee Name	Job Title
134969118	Miller	Sales Represent ative
811994146	Margheim	Marketing Assistant

**Employees table,
normalized to first
normal form**

SSN	Phone Number
134969118	555-0619
134969118	555-0412
811994146	555-0124
811994146	555-0189
111969468	555-0189

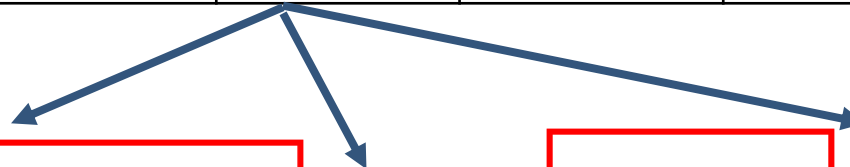
**EmployeesPhone
table**

Also when phone numbers might be shared by employees

First Normal Form

**Employees
table not
normalized**

SSN	Employee Name	Job Title	Phone Number 1	Phone Number 2
134969118	Miller	Sales Represent ative	555-0619	555-0412
811994146	Margheim	Marketing Assistant	555-0124	555-0189



SSN	Employee Name	Job Title
134969118	Miller	Sales Represent ative
811994146	Margheim	Marketing Assistant

**Employees table,
normalized to first
normal form**

SSN	Phone Number ID	Phone Number
134969118	1	555-0619
134969118	2	555-0412
811994146	3	555-0124
811994146	4	555-0189

**EmployeesPhone
table**

**PhoneNumbers
table**

Second Normal Form

2

- Tables should be in first normal form
- All non-key columns in a table should be functionally dependent on the whole of the primary key

Composite
primary key →

SSN	Start Date	Employee Name	Date Of Birth
134969118	2012-03-16	Miller	1970-01-23
811994146	2010-05-02	Margheim	1986-06-05

**Employees
table**

Employees frequently leave and then return to the company

SSN	Start Date	Employee Name	Date Of Birth
134969118	2012-03-16	Miller	1970-01-23
811994146	2010-05-02	Margheim	1986-06-05
134969118	2012-08-01	Miller	1970-01-23
134969118	2013-03-01	Miller	1970-01-23

↑
Redundancy

Second Normal Form

3

- Tables should be in first normal form
- All non-key columns in a table should be functionally dependent on the whole of the primary key

Composite
primary key

SSN	Start Date	Employee Name	Date Of Birth
134969118	2012-03-16	Miller	1970-01-23
811994146	2010-05-02	Margheim	1986-06-05

**Employees
table**

Composite
primary key

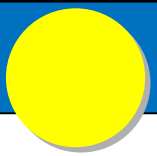
SSN	Employee Name	Date Of Birth
134969118	Miller	1970-01-23
811994146	Margheim	1986-06-05

**Employees table in
second normal form**

SSN	Start Date
134969118	2012-03-16
811994146	2010-05-02

EmployeeDate table

Third Normal Form



- Tables should be in second normal form
- There should be no transitive dependencies between non-key attributes on the primary key

The Salary column is transitively dependent on the primary key, through the Job Title column: SSN → Job Title → Salary

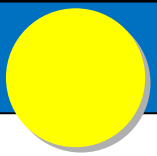


SSN	Employee Name	Date Of Birth	Job Title	Salary
134969118	Miller	1970-01-23	Sales Representative	56,000
811994146	Margheim	1986-06-05	Marketing Assistant	48,000
.....	Marketing Assistant	46,000
.....	Marketing Assistant	49,000

1

There is nothing to prevent the addition of rows to the Employees table that have the same value in the Job Title column, but different values for the Salary column

Third Normal Form



- Tables should be in second normal form
- There should be no transitive dependencies between non-key attributes on the primary key

The Salary column is transitively dependent on the primary key, through the Job Title column: SSN → Job Title → Salary



SSN	Employee Name	Date Of Birth	Job Title	Salary
134969118	Miller	1970-01-23	Sales Representative	56,000
811994146	Margheim	1986-06-05	Marketing Assistant	48,000
.....	Marketing Assistant	48,000
.....	Marketing Assistant	48,000

2

Now the salary for any given job title is stored more than once

Third Normal Form

- Tables should be in second normal form
- There should be no transitive dependencies between non-key attributes on the primary key

The Salary column is transitively dependent on the primary key, through the Job Title column: SSN → Job Title → Salary

SSN	Employee Name	Date Of Birth	Job Title	Salary
134969118	Miller	1970-01-23	Sales Representative	56,000
811994146	Margheim	1986-06-05	Marketing Assistant	48,000

SSN	Employee Name	Date Of Birth	Job Title
134969118	Miller	1970-01-23	Sales Representative
811994146	Margheim	1986-06-05	Marketing Assistant

Employees table in third normal form

Job Title	Salary
Sales Representative	56,000
Marketing Assistant	48,000

TitleSalary table