Dimension Table Design Techniques

Ana Voicu

@ana_voicu

Dimension Table Techniques

What is specific to a dimension table?

Technical guidelines for designing a table
- Working with keys
  - Surrogate key
  - Business key
- Populating the dimension's attributes with data

Examples of common dimensions
- Date dimension
- Product dimension

What Is Specific to a Dimension Table?

Example: Sources for the Product Dimension

Administration system

Inventory changes
- Normalized model
- Products, departments, categories, subcategories

Software product for storing recipes

Used by the cook for daily recipes
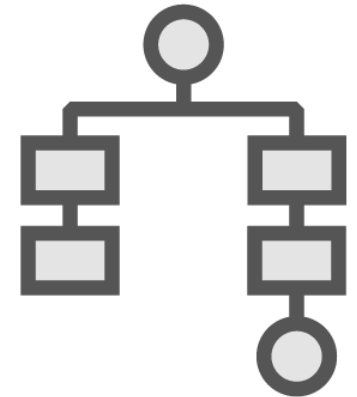
POS used by staff

Take and fulfill orders

Flat file with information from an old accounting system

The system is not in use anymore
- But kept for historic information

Product dimension

Requirements of a Dimension Table

Administration system

Software product for storing recipes

POS used by staff

Flat file with information from an old accounting system

How will you build a table meeting these requirements?
- Dimension table can be linked to fact tables
- Keep track of the source system for data
- Data is consistent and readable
- Keep history of data changes from the operational system

What Is Specific to a Dimension Table?

A dimension table is different than a relational table

- The normalization rules don't apply
- Not optimized for data modification aspects
- Integrates in a consistent way multiple source data
- Data is easy to read by the user

Surrogate Keys

Relationship between Fact and Dimension

One-to-many relationship

Dim Product
PK
Attribute1
Attribute2
Attribute3
Attribute4
Attribute5
...

Fact Sales
PK
FK product
FK date
...
Fact1
Fact2
...

Relationship components

- 1:M example
  - 1 product (chocolate chip cookie)
  - 2 sales (transactions)

- Keys involved
  - Primary key of dimension table
  - Foreign key from fact table

- It's important to create it correctly
  - Plays an important role in the BI solution
  - Should be unique across all data sources for the dimension

Creating the Primary Key

Bad practice

Use the PK from the source systems

Can lead to overlapping rows

Source system 1

Id	Name	...
3	Chocolate chip cookie	...

Source system 2

Id	Name	...
3	Potato	...

Good practice

Generate an artificial key
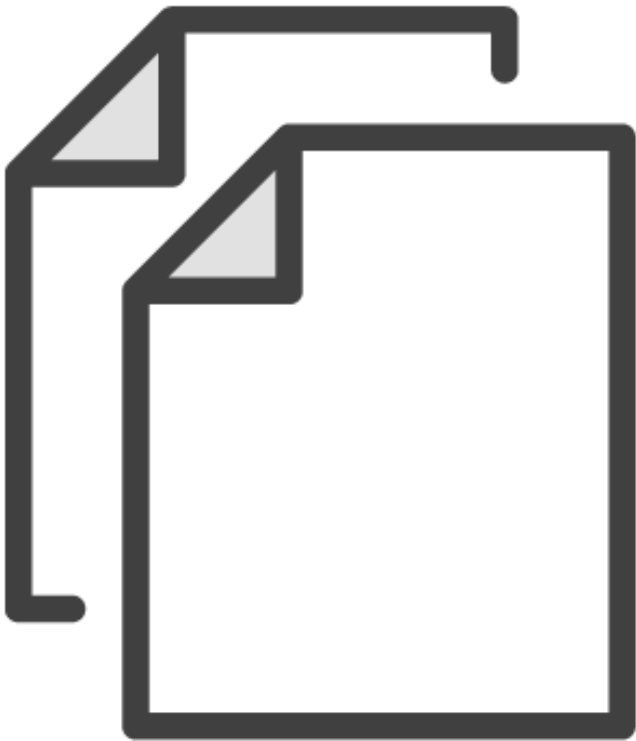
- Simple integer column

- Automatically incremented

- Doesn't require careful maintenance

In DW terminology, this is the surrogate key

Synonyms for Surrogate Key

Meaningless key

Integer key

Non-natural key

Artificial key

Synthetic key

Advantages of Using a Surrogate Key

Advantages of Using a Surrogate Key

Integrate multiple source systems

Advantages of Using a Surrogate Key

Integrate multiple source systems

- Example: gathering data for the Product dimension in a restaurant or store

Integrate Data from Multiple Sources

Integrate Data from Multiple Sources

Source system 1

ID	Name	Description	UM	Color
123	Milk	Non-fat	L	White
776	Sugar	Brown sugar	KG	Brown

Integrate Data from Multiple Sources

Source system 1

ID	Name	Description	UM	Color
123	Milk	Non-fat	L	White
776	Sugar	Brown sugar	KG	Brown

Source system 2

ID	Name	Description	UM	Color
776	Butter	Non-fat	G	Yellow
778	Sugar	Normal sugar	KG	White

Integrate Data from Multiple Sources

Source system 1

ID	Name	Description	UM	Color
123	Milk	Non-fat	L	White
776	Sugar	Brown sugar	KG	Brown

Source system 2

ID	Name	Description	UM	Color
776	Butter	Non-fat	G	Yellow
778	Sugar	Normal sugar	KG	White

Integrate Data from Multiple Sources

Source system 1

ID	Name	Description	UM	Color
123	Milk	Non-fat	L	White
776	Sugar	Brown sugar	KG	Brown

Source system 2

ID	Name	Description	UM	Color
776	Butter	Non-fat	G	Yellow
778	Sugar	Normal sugar	KG	White

Data warehouse – Product dimension

Dim Key	Original Key	Name	Description	UM	Color
1	123	Milk	Non-fat	L	White
2	776	Sugar	Brown sugar	KG	Brown
3	776	Butter	Non-fat	G	Yellow
4	778	Sugar	Normal sugar	KG	White

Advantages of Using a Surrogate Key

Integrate multiple source systems

Keep track of attribute changes over time

Advantages of Using a Surrogate Key

Integrate multiple source systems

Keep track of attribute changes over time

- Use surrogate keys to handle multiple versions of a row

Keep Track of Attribute Changes over Time

Keep Track of Attribute Changes over Time

Before 09-16-2018

Key	Name	Description	UM	Color
387	Cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink

Keep Track of Attribute Changes over Time

Before 09-16-2018

Key	Name	Description	UM	Color
387	Cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink

After 09-16-2018

Key	Name	Description	UM	Color
387	Super cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink

Keep Track of Attribute Changes over Time

Before 09-16-2018

Key	Name	Description	UM	Color
387	Cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink

After 09-16-2018

Key	Name	Description	UM	Color
387	Super cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink

Keep Track of Attribute Changes over Time

Before 09-16-2018

Key	Name	Description	UM	Color
387	Cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink

Data warehouse – Product dimension

Dim Key	Orig. Key	Name	Description	UM	Color
112	387	Cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink
244	387	Super cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink

After 09-16-2018

Key	Name	Description	UM	Color
387	Super cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink

Keep Track of Attribute Changes over Time

Before 09-16-2018

Key	Name	Description	UM	Color
387	Cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink

Data warehouse – Product dimension

Dim Key	Orig. Key	Name	Description	UM	Color	Valid From	Valid To
112	387	Cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink	01-01-2017	09-16-2018
244	387	Super cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink	09-17-2018	12-31-9999

After 09-16-2018

Key	Name	Description	UM	Color
387	Super cherry toffee	Great cherry taste, sticky, chewy consistency	Piece	Pink

Advantages of Using a Surrogate Key

Integrate multiple source systems

Keep track of attribute changes over time

Protect the data warehouse from operational changes

Protect the Data Warehouse from Operational Changes

Characteristics of an operational system:

- Logs and processes daily activities/current actions
- Some data history is preserved
- Obsolete data is deleted after a period of time
- Codes/IDs can be reassigned to new data

Protect the Data Warehouse from Operational Changes

Protect the Data Warehouse from Operational Changes

Before 01-01-2018

Key	Name	Description	UM	Color
432	Chocolate chip cookie	Grandma's cookie recipe	Piece	Brown
776	Sugar	Brown sugar	KG	Brown

Protect the Data Warehouse from Operational Changes

Before 01-01-2018

Key	Name	Description	UM	Color
432	Chocolate chip cookie	Grandma's cookie recipe	Piece	Brown
776	Sugar	Brown sugar	KG	Brown

After 01-01-2018

Key	Name	Description	UM	Color
776	Sugar	Brown sugar	KG	Brown

Protect the Data Warehouse from Operational Changes

Before 01-01-2018

Key	Name	Description	UM	Color
432	Chocolate chip cookie	Grandma's cookie recipe	Piece	Brown
776	Sugar	Brown sugar	KG	Brown

After 01-01-2018

Key	Name	Description	UM	Color
776	Sugar	Brown sugar	KG	Brown

Protect the Data Warehouse from Operational Changes

Before 01-01-2018

Key	Name	Description	UM	Color
432	Chocolate chip cookie	Grandma's cookie recipe	Piece	Brown
776	Sugar	Brown sugar	KG	Brown

After 01-01-2018

Key	Name	Description	UM	Color
776	Sugar	Brown sugar	KG	Brown

Data warehouse – Product dimension

Dim Key	Orig. Key	Name	Description	UM	Color	Valid From	Valid To
83	432	Chocolate chip cookie	Grandma's cookie recipe	Piece	Brown	01-01-1753	01-01-2018
2	776	Sugar	Normal sugar	KG	Brown	01-01-1753	12-31-9999

Advantages of Using a Surrogate Key

Integrate multiple source systems

Keep track of attribute changes over time

Protect the data warehouse from operational changes

Handle null or unknown conditions

Advantages of Using a Surrogate Key

Integrate multiple source systems

Keep track of attribute changes over time

Protect the data warehouse from operational changes

Handle null or unknown conditions
- Empty row technique

Advantages of Using a Surrogate Key

Integrate multiple source systems

Keep track of attribute changes over time

Protect the data warehouse from operational changes

Handle null or unknown conditions

- Empty row technique
- Handle situations when a link does not exist between fact and dimensions

Advantages of Using a Surrogate Key

Integrate multiple source systems

Keep track of attribute changes over time

Protect the data warehouse from operational changes

Handle null or unknown conditions

- Empty row technique
- Handle situations when a link does not exist between fact and dimensions

Scenario for Null or Unknown Conditions
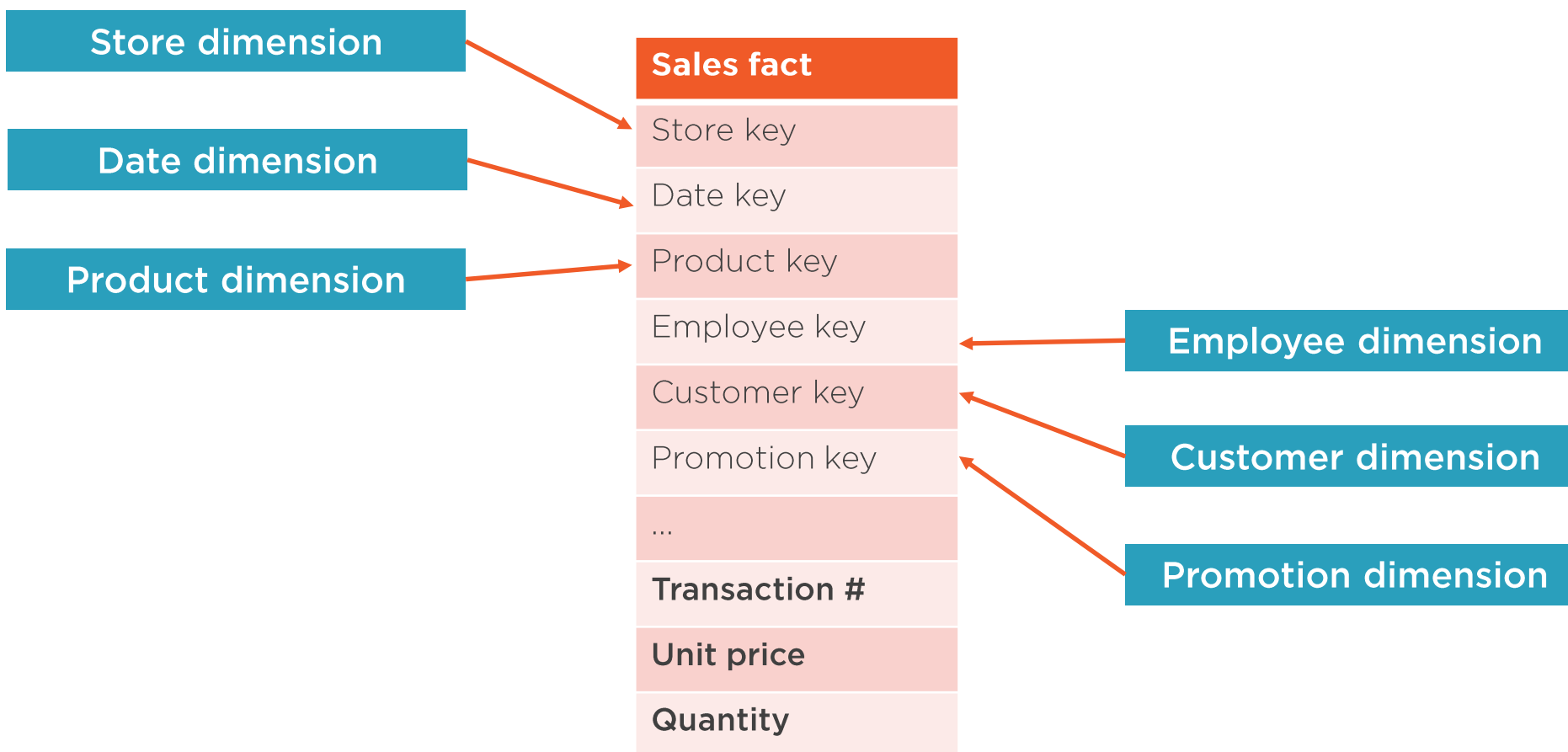
Scenario for Null or Unknown Conditions
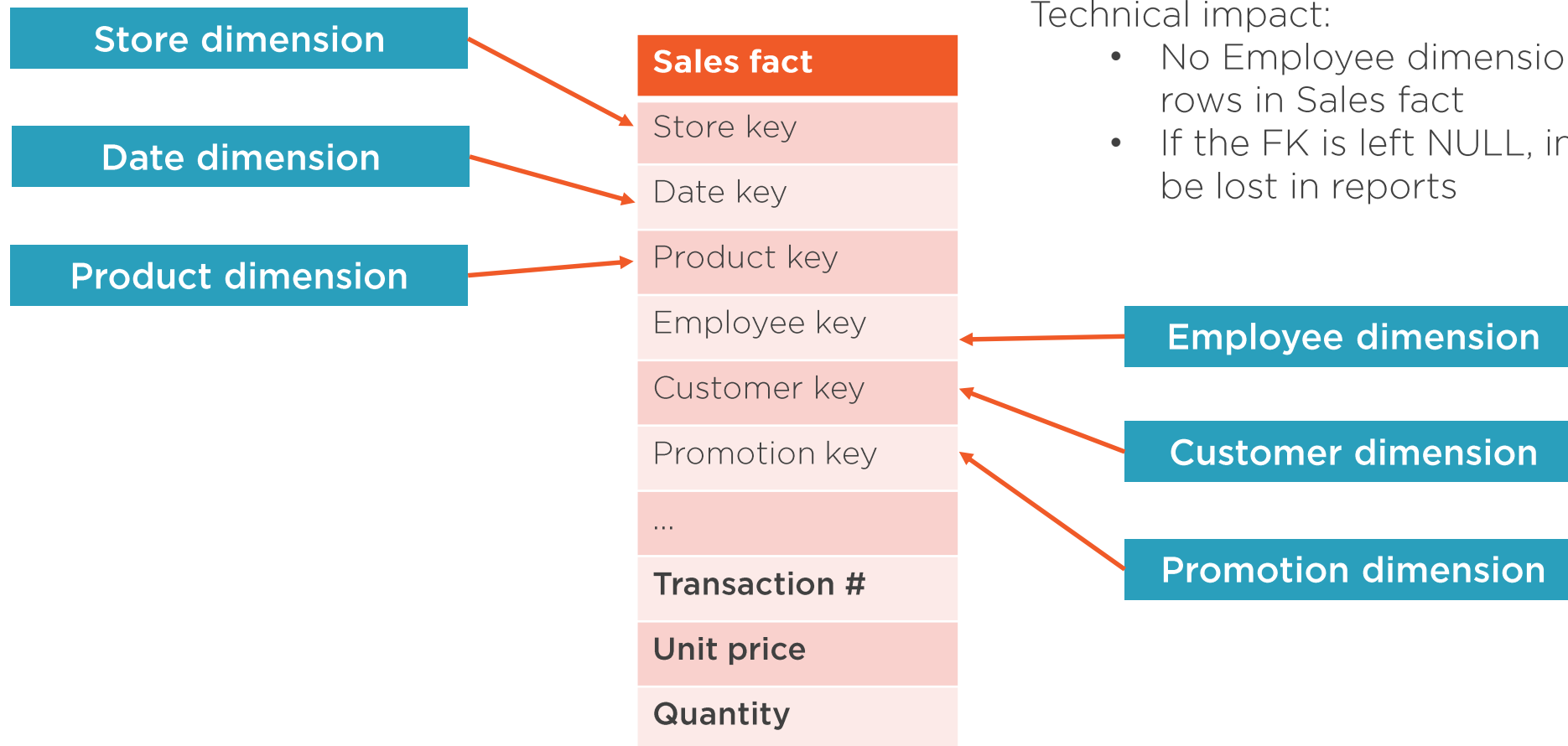
Self-checkout kiosks were introduced
Customers pay by themselves
Employees are not involved for this transaction
Technical impact:
- No Employee dimension key for these rows in Sales fact
- If the FK is left NULL, information may be lost in reports

Store dimension

Date dimension

Product dimension

Sales fact
Store key
Date key
Product key
Employee key
Customer key
Promotion key
...
Transaction #
Unit price
Quantity

Employee dimension

Customer dimension

Promotion dimension

Handle Null or Unknown Conditions

Adding the "empty row" to each dimension table

Key	Name	Birthday	Gender	PhoneNr.
1	Unknown	01-01-1753	Unknown	Unknown
32	Alice Cohen	03-05-1989	Female	Unknown

Handle Null or Unknown Conditions

Adding the "empty row" to each dimension table

Key	Name	Birthday	Gender	PhoneNr.
1	Unknown	01-01-1753	Unknown	Unknown
32	Alice Cohen	03-05-1989	Female	Unknown

Data warehouse – Sales fact table

Fact Key	Product Key	Customer Key	Employee Key	Transaction Nr	Amt.
...	432	10	32	#1050	$5
...	776	12	1	#2367	$3

Advantages of Using a Surrogate Key

Integrate multiple source systems

Keep track of attribute changes over time

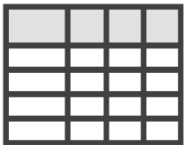Protect the data warehouse from operational changes

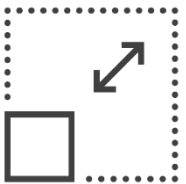Handle null or unknown conditions

Improve performance

Improve Performance

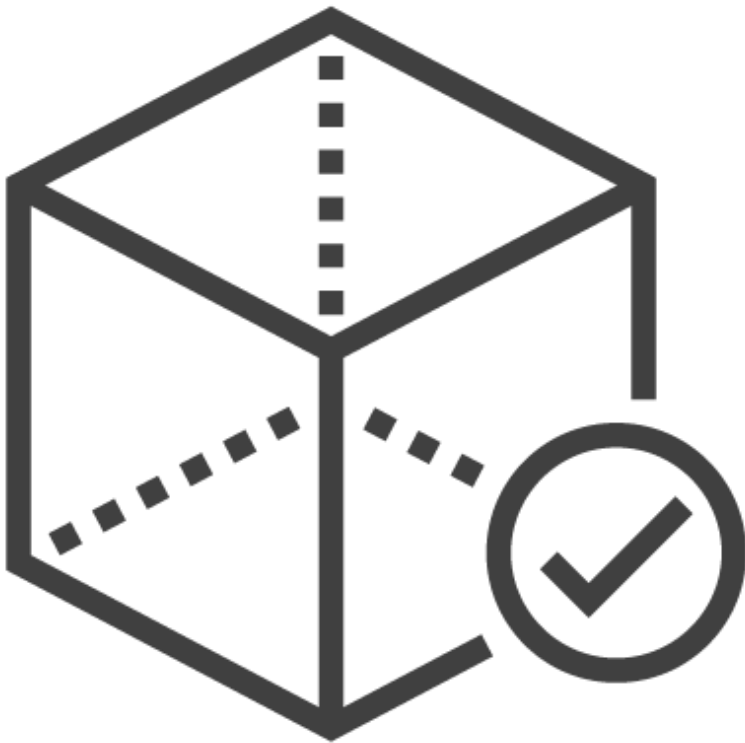The surrogate key should be a small number, preferably an integer

You need to anticipate the growth of your dimension table (a 4-byte int can handle ~2 billion positive values and ~4 billion values in total)

The smaller the surrogate key, the smaller the fact table

Dimensions of a Fact Table

Fact tables

- Have FKs to a lot of dimension tables
- Grow very fast
- Every byte counts!

FK length	Fact table growth	Fact table storage
X bytes	1 billion rows	Y GB
X + 1 bytes	1 billion rows	Y + 1 GB

Advantages of Using a Surrogate Key

Integrate multiple source systems

Keep track of attribute changes over time

Protect the data warehouse from operational changes

Handle null or unknown conditions

Improve performance

Business Keys

Business Keys

The primary key of the source system

Also known as

- Natural keys

- Production keys

- Operational keys

Business Keys Properties

Attribute in the dimension table

Business Keys Properties

Attribute in the dimension table

Can be prefixed with a source system's code

(POS|432 or CSV|635)

Business Keys Properties

Attribute in the dimension table

Can be prefixed with a source system's code

(POS|432 or CSV|635)

Can cause duplicated items in the dimensions

Business Keys Properties

Attribute in the dimension table

Can be prefixed with a source system's code

(POS|432 or CSV|635)

Can cause duplicated items in the dimensions

If the BK is composed of meaningful codes, it should be split and each code stored in separate columns

Keys in a Dimension Table

Surrogate key

- Primary key, no business value

Business key

- Original primary from the source system

Other Dimensional Design Considerations

Characteristics of a Dimension Table

Companion of a fact table

Used to describe the business in clear terms

Design implications

- Store descriptive words instead of codes
- Replace flags or indicators with descriptions
- Replace null values with meaningful words

Use Descriptions Instead of Codes

Each code stored should be accompanied by a descriptive decode

Each attribute should be easy to interpret by people

Example of Using Descriptions Instead of Codes

Country of import for the most sold products

# Sold products	Name	Imported from
5000	Zywiec Beer	PL
4856	Pan Beer	HR
3991	Ursus Beer	RO
2674	Mort Subite Beer	BE

Id	Name	Imported from
5000	Zywiec Beer	Poland
4856	Pan Beer	Croatia
3991	Ursus Beer	Romania
2674	Mort Subite Beer	Belgium

Codes are not easy to interpret by everyone

Decodes can be retrieved from the operational system as well

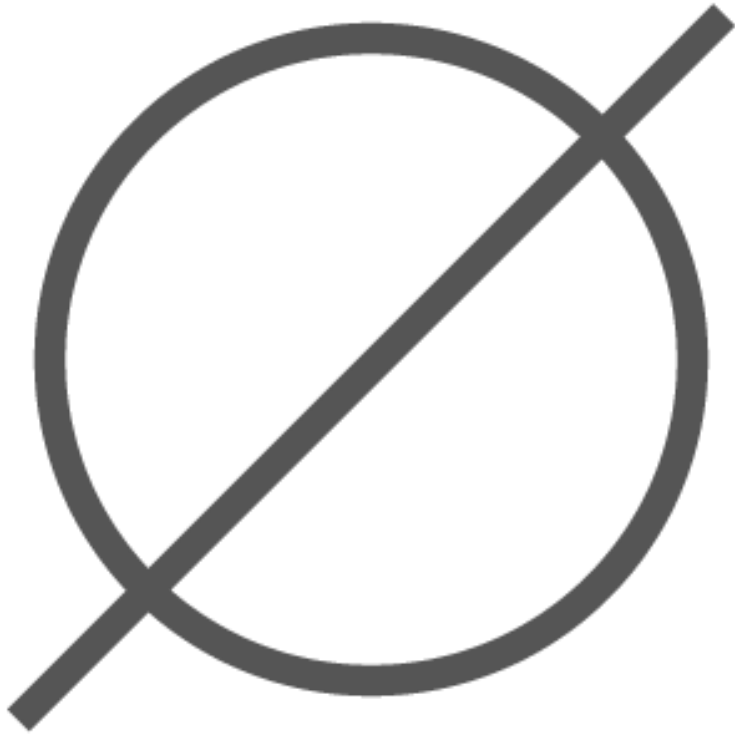# Use Textual Attributes Instead of Flags and Indicators

Flags and operational indicators should be supplemented with meaningful text words

Examples from the Date dimension:

Column	Meaningful name	Indicator
Weekday indicator	Weekday/Weekend	1/0
Holiday indicator	Holiday/Non-holiday	true/false

Replace Null Values in Dimensions

Replace null values with meaningful words, like

- Unknown
- Not applicable

Null values in data warehouse tables have unexpected behaviors

Nulls are not present in drop-down menus, so data may be missing from reports

Demo

Creating the Date dimension
- What is a Date dimension?
- Why is it important?

What Is a Date Dimension?

A table that stores dates

Is populated with a large number of dates (10 years, 100 years)

As a dimension table, it is not large

It is loaded once, not periodically

Why Is It Important?

Is present in almost all data warehouses

Helps performing time analysis

- Products sold this year vs. last year
- Products sold per month compared to last year
- Per week, per weekend, per holiday..

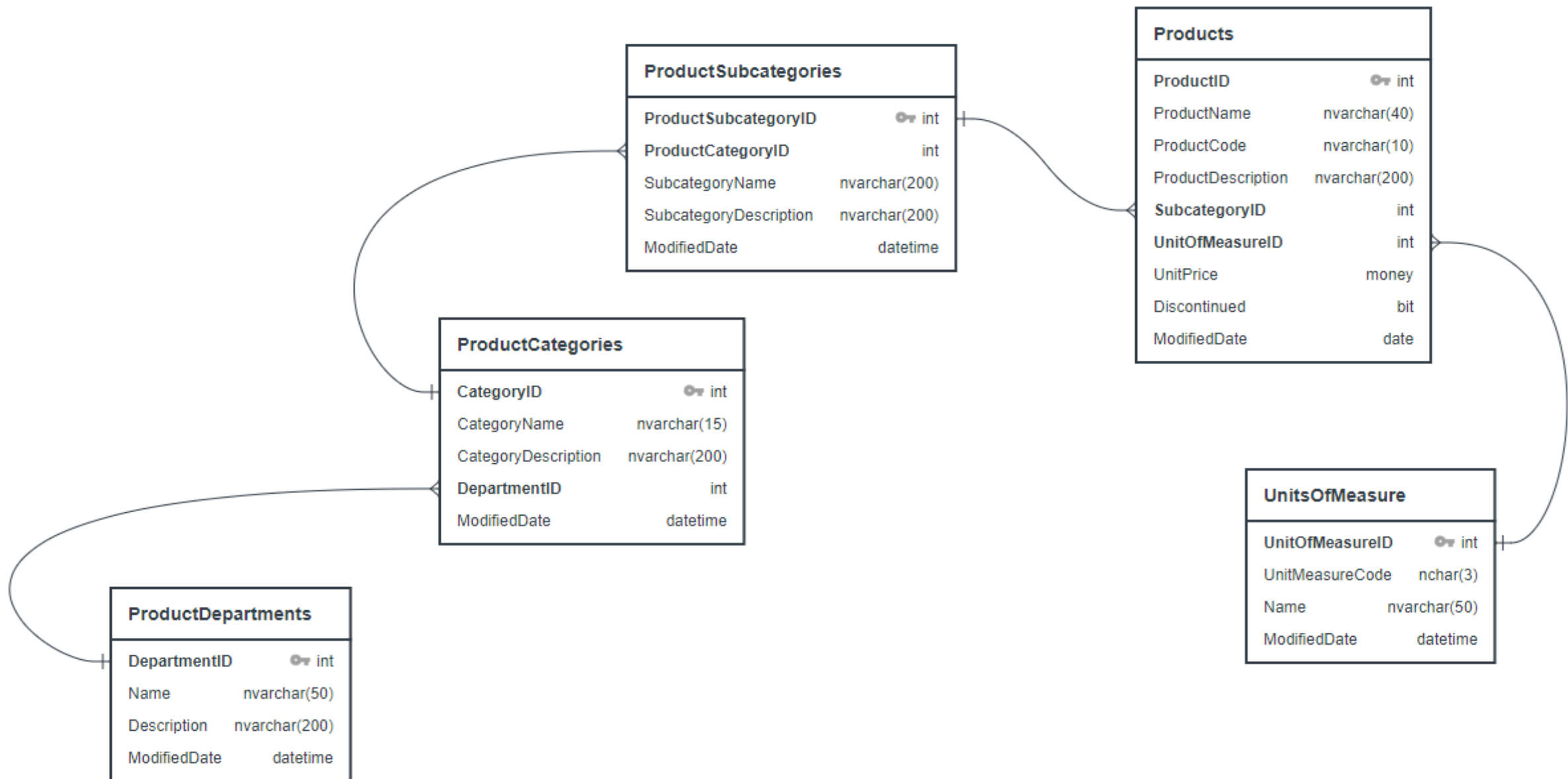**Contains pre-calculated information about dates**

Demo

Creating the Product dimension

- Product dimension is present in many data warehouse projects
- It doesn't have a standard set of attributes

Product-related Tables

ProductSubcategories

ProductSubcategoryID	🔑 int
ProductCategoryID	int
SubcategoryName	nvarchar(200)
SubcategoryDescription	nvarchar(200)
ModifiedDate	datetime

Products

ProductID	🔑 int
ProductName	nvarchar(40)
ProductCode	nvarchar(10)
ProductDescription	nvarchar(200)
SubcategoryID	int
UnitOfMeasureID	int
UnitPrice	money
Discontinued	bit
ModifiedDate	date

ProductCategories

CategoryID	🔑 int
CategoryName	nvarchar(15)
CategoryDescription	nvarchar(200)
DepartmentID	int
ModifiedDate	datetime

UnitsOfMeasure

UnitOfMeasureID	🔑 int
UnitMeasureCode	nchar(3)
Name	nvarchar(50)
ModifiedDate	datetime

ProductDepartments

DepartmentID	🔑 int
Name	nvarchar(50)
Description	nvarchar(200)
ModifiedDate	datetime

Summary

Characteristics of a dimension table

- Is different than a relational table
- Integrates data from multiple sources
- Data is easy to read by the user

Keys in a dimension table

- Surrogate key
- Business key

Guidelines for implementing a dimension table

- Store descriptive words instead of codes
- Replace flags or indicators with descriptions
- Replace null values with meaningful words