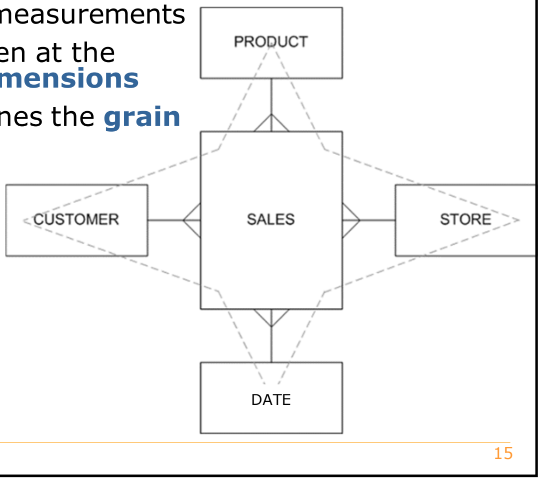
# Bases de Dados e Armazéns de Dados

Departamento de Engenharia Informática (DEI/ISEP)
Paulo Oliveira
pjo@isep.ipp.pt

# Dimensional Data Modeling

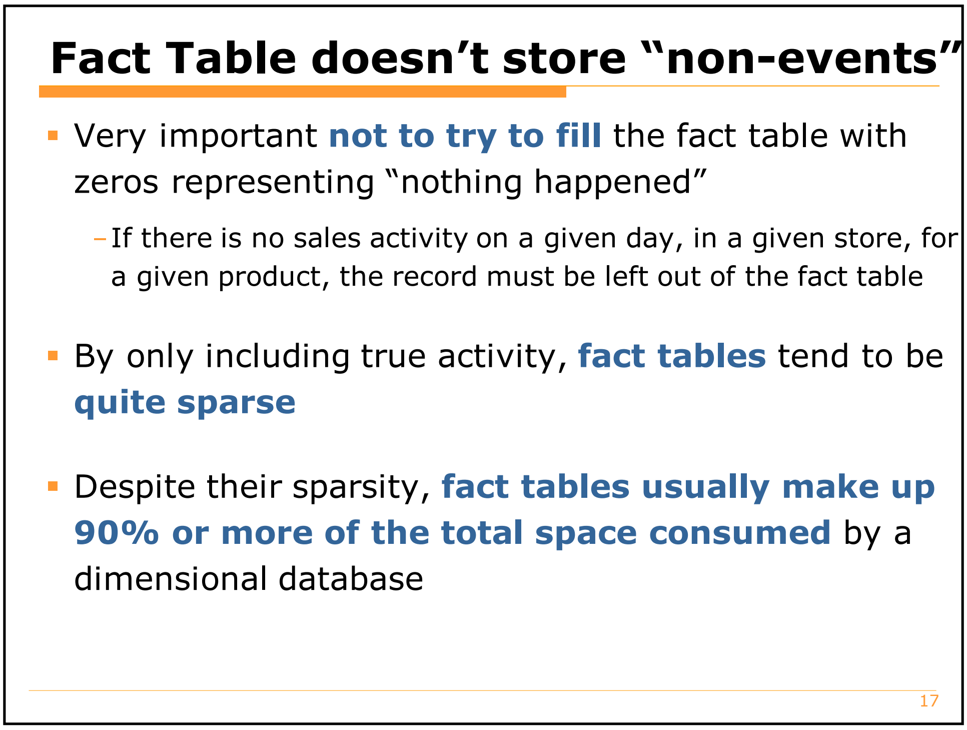
# Dimensional Modeling

- Build around the **numerical measurements** of the business
  - Fact tables contain **measurements**
  - Dimension tables contain the **context** surrounding measurements
  - Measurements are taken at the **intersection of all dimensions**
  - List of dimensions defines the **grain of the fact table**



# Fact Table

- Is the **primary table** in a dimensional model
- Holds the **measurements** of the business
- Composed by a **set of foreign keys** that connect to the dimension tables
- Its primary key is made up by **the set or a subset of the foreign keys**
- Role of a **normalized n-ary associative entity**
- All measurements in a fact table must be at the **same grain**

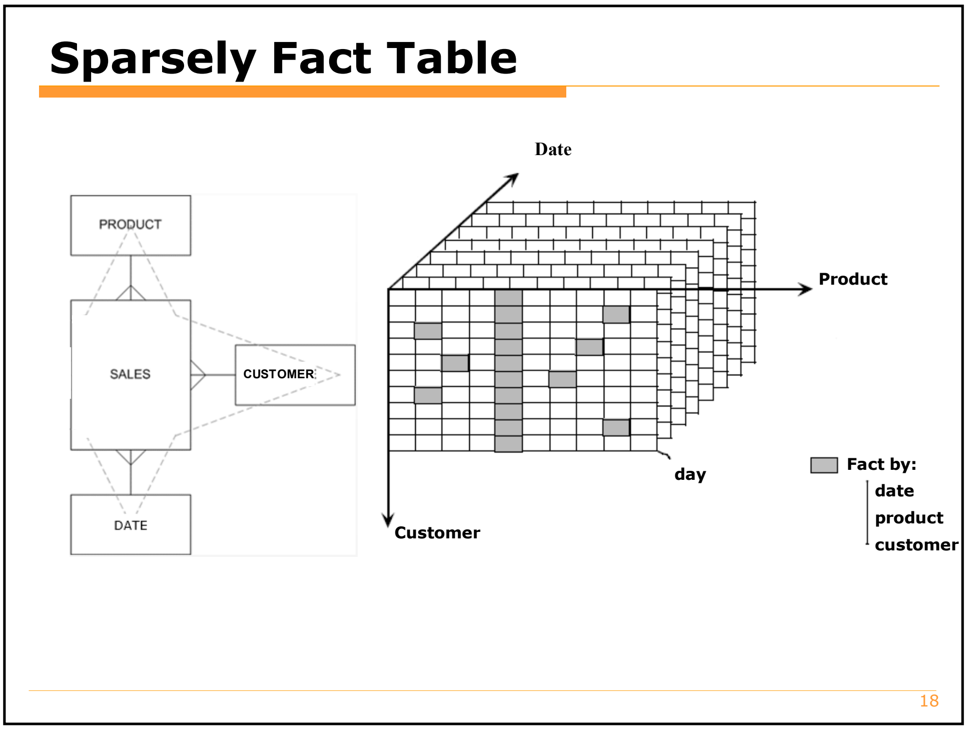


# Fact Table doesn't store "non-events"

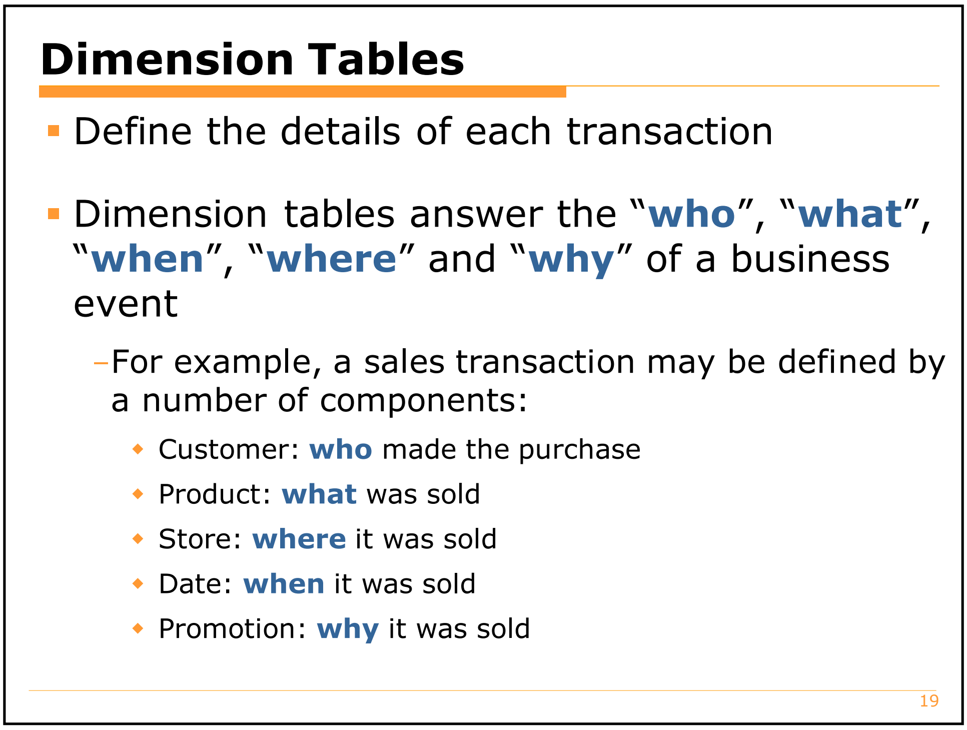
- Very important **not to try to fill** the fact table with zeros representing "nothing happened"
  - If there is no sales activity on a given day, in a given store, for a given product, the record must be left out of the fact table

- By only including true activity, **fact tables** tend to be **quite sparse**

- Despite their sparsity, **fact tables usually make up 90% or more of the total space consumed** by a dimensional database

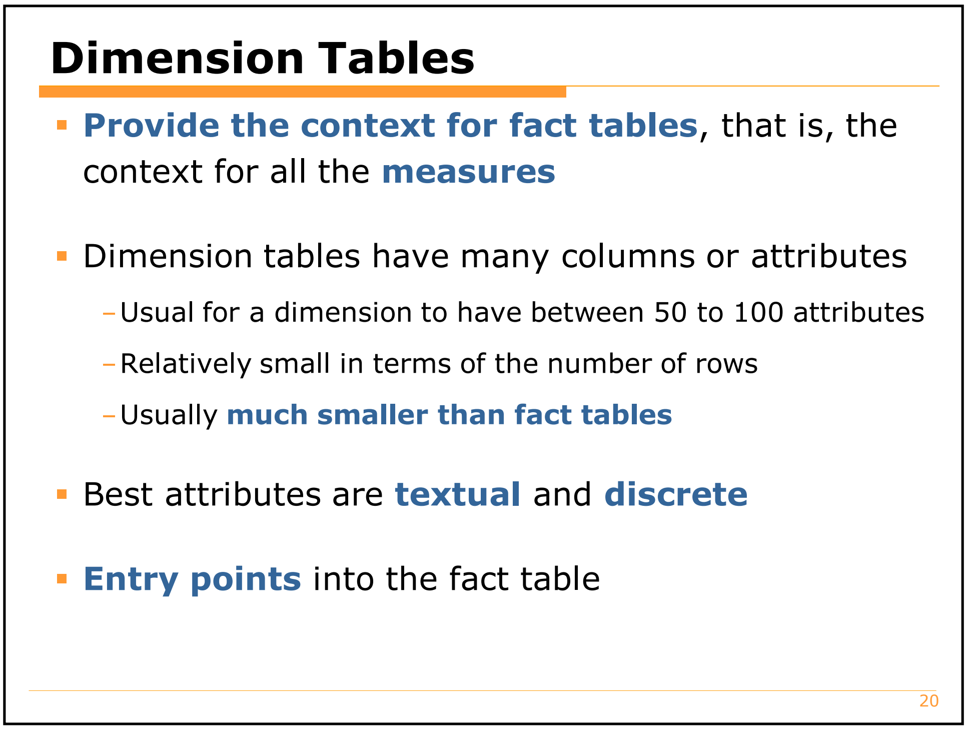


# Sparsely Fact Table



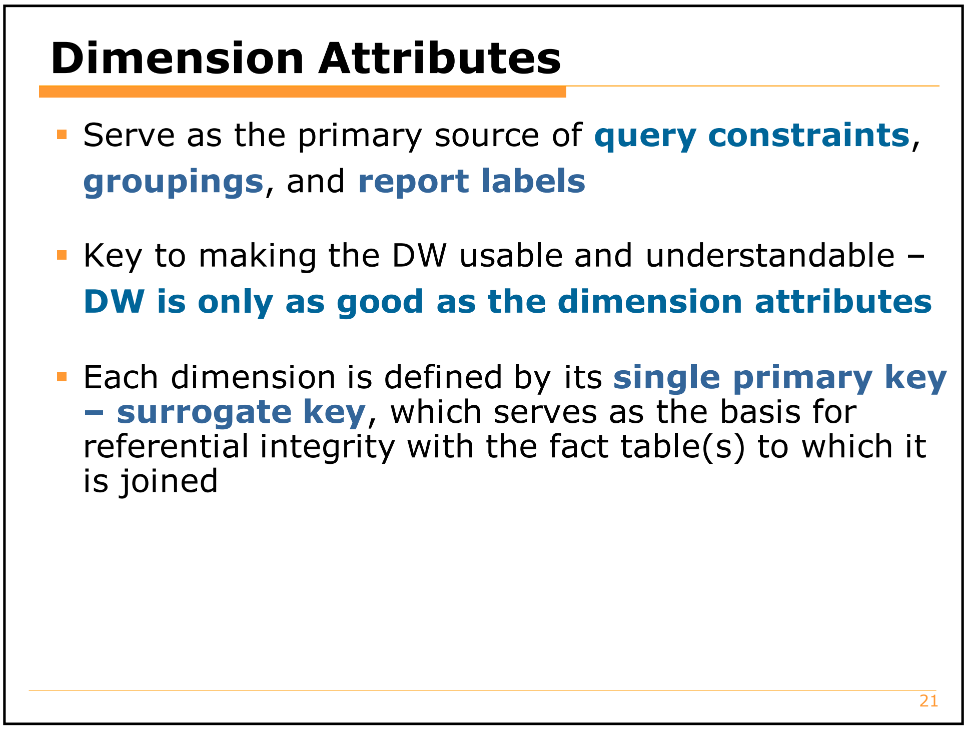
## Dimension Tables

- Define the details of each transaction
- Dimension tables answer the "**who**", "**what**", "**when**", "**where**" and "**why**" of a business event
  - For example, a sales transaction may be defined by a number of components:
    - Customer: **who** made the purchase
    - Product: **what** was sold
    - Store: **where** it was sold
    - Date: **when** it was sold
    - Promotion: **why** it was sold



## Dimension Tables

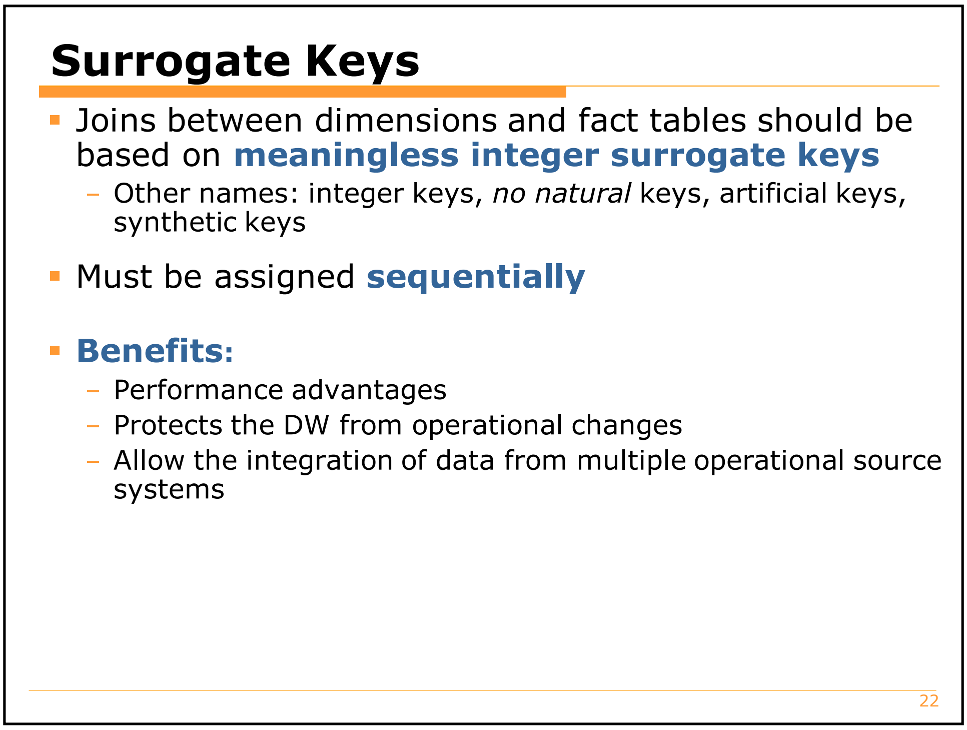
- **Provide the context for fact tables**, that is, the context for all the **measures**
- Dimension tables have many columns or attributes
  - Usual for a dimension to have between 50 to 100 attributes
  - Relatively small in terms of the number of rows
  - Usually **much smaller than fact tables**
- Best attributes are **textual** and **discrete**
- **Entry points** into the fact table



## Dimension Attributes

- Serve as the primary source of **query constraints**, **groupings**, and **report labels**
- Key to making the DW usable and understandable – **DW is only as good as the dimension attributes**
- Each dimension is defined by its **single primary key – surrogate key**, which serves as the basis for referential integrity with the fact table(s) to which it is joined

21



## Surrogate Keys

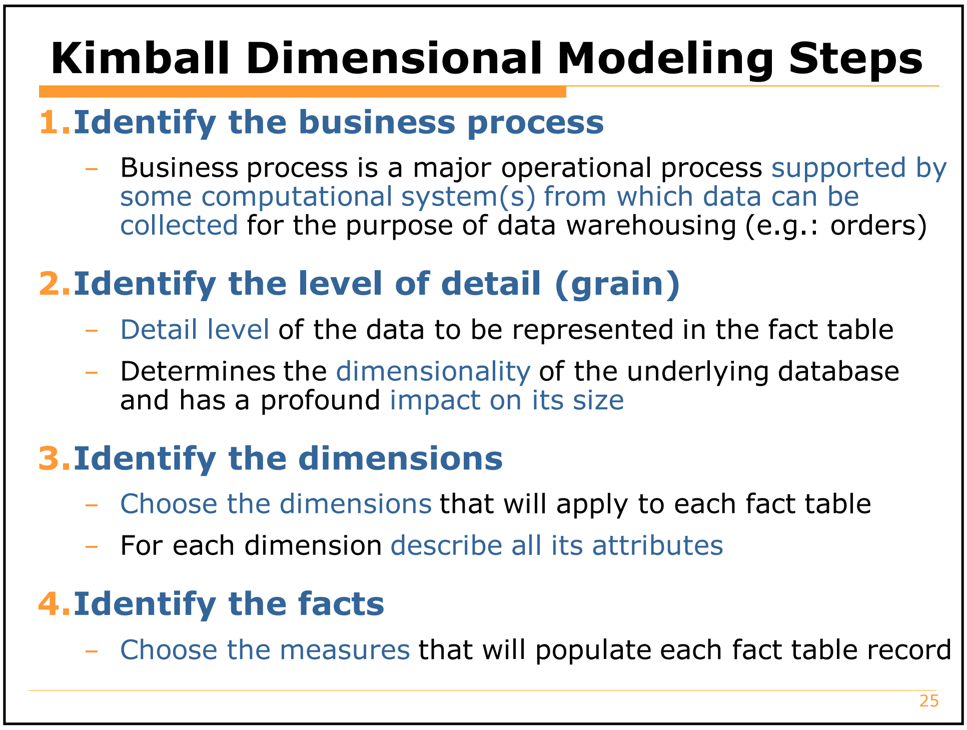
- Joins between dimensions and fact tables should be based on **meaningless integer surrogate keys**
  - Other names: integer keys, *no natural* keys, artificial keys, synthetic keys
- Must be assigned **sequentially**
- **Benefits:**
  - Performance advantages
  - Protects the DW from operational changes
  - Allow the integration of data from multiple operational source systems

22

# The Grocery Store

## Grocery Store Business - Brief Description

- The business has **500 large grocery stores** spread over the country. Each store is divided by **departments** such as grocery, frozen foods, dairy, meat, bakery, floral, drugs,… Each store has roughly **60000 individual products** (called **Stock Keeping Units** – **SKUs**) on its shelves. About 40000 SKUs come from outside manufactures and have bar codes called **Universal Product Codes** – **UPCs**.
- The remaining 20000 SKUs come from departments like the meat, bakery, or floral departments and don’t have UPC codes. Nevertheless, as a grocery store, these products **also have SKU numbers** assigned to them.
- At the grocery store, management is concerned with the sales of the products as well as maximizing the profit at each store. The most significant management decisions have to do with **pricing**, **promotions** and **good visibility of promotions**.



## Kimball Dimensional Modeling Steps

### 1. Identify the business process

- Business process is a major operational process supported by some computational system(s) from which data can be collected for the purpose of data warehousing (e.g.: orders)

### 2. Identify the level of detail (grain)

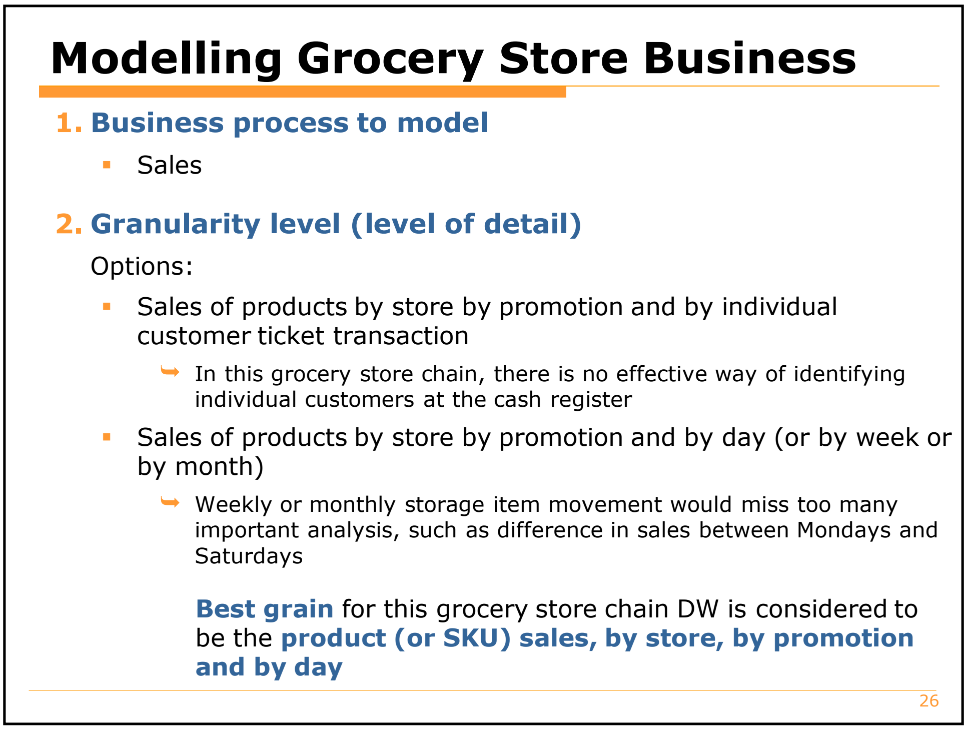
- Detail level of the data to be represented in the fact table
- Determines the dimensionality of the underlying database and has a profound impact on its size

### 3. Identify the dimensions

- Choose the dimensions that will apply to each fact table
- For each dimension describe all its attributes

### 4. Identify the facts

- Choose the measures that will populate each fact table record



## Modelling Grocery Store Business

### 1. Business process to model

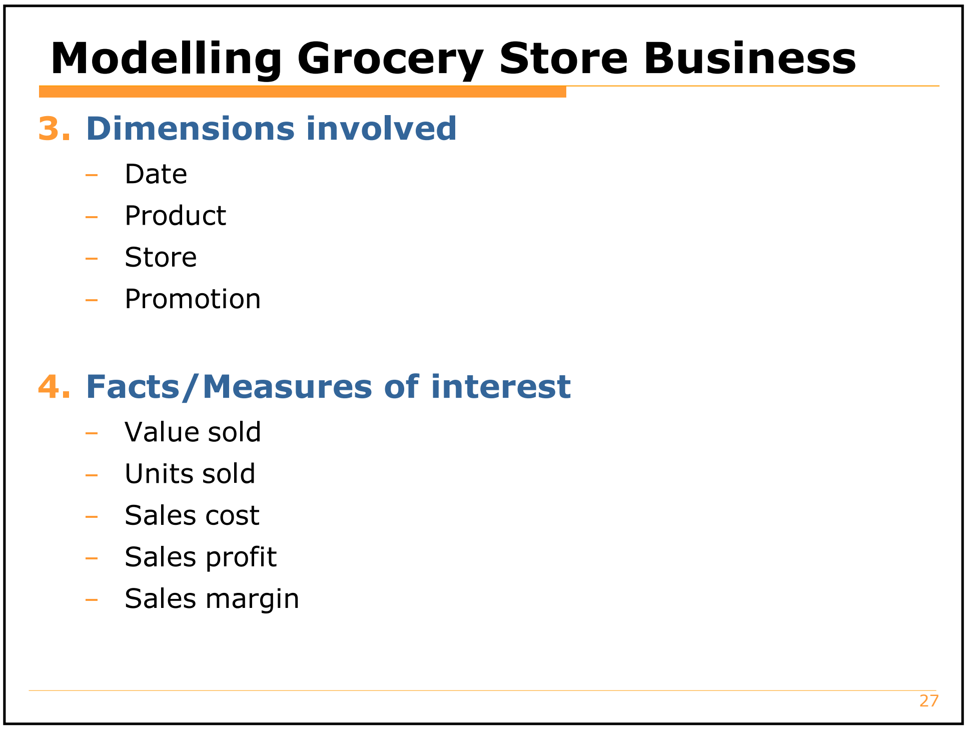
- Sales

### 2. Granularity level (level of detail)

Options:

- Sales of products by store by promotion and by individual customer ticket transaction
  - ➥ In this grocery store chain, there is no effective way of identifying individual customers at the cash register
- Sales of products by store by promotion and by day (or by week or by month)
  - ➥ Weekly or monthly storage item movement would miss too many important analysis, such as difference in sales between Mondays and Saturdays

**Best grain** for this grocery store chain DW is considered to be the **product (or SKU) sales, by store, by promotion and by day**



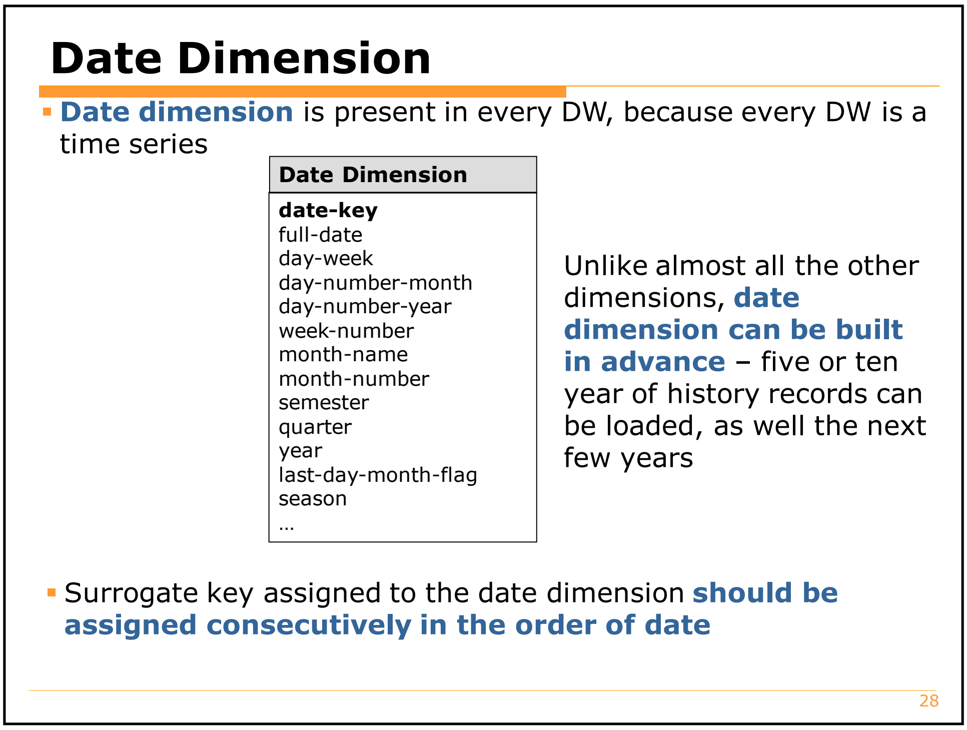
# Modelling Grocery Store Business

## 3. Dimensions involved

- Date
- Product
- Store
- Promotion

## 4. Facts/Measures of interest

- Value sold
- Units sold
- Sales cost
- Sales profit
- Sales margin



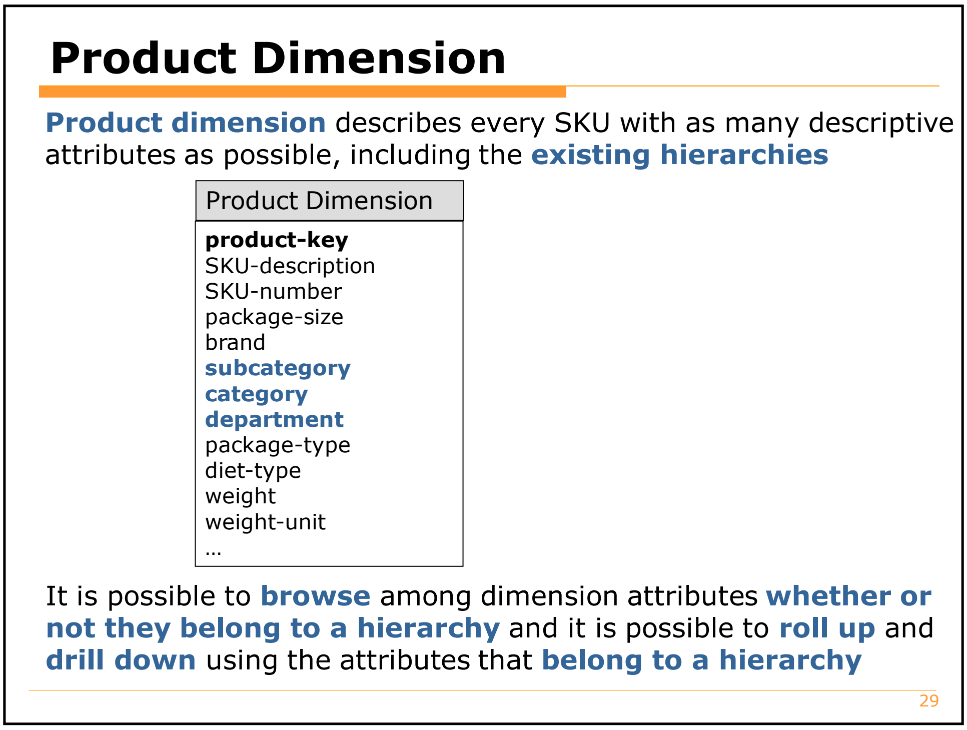
# Date Dimension

- **Date dimension** is present in every DW, because every DW is a time series

| Date Dimension |
|---|
| **date-key**<br>full-date<br>day-week<br>day-number-month<br>day-number-year<br>week-number<br>month-name<br>month-number<br>semester<br>quarter<br>year<br>last-day-month-flag<br>season<br>… |

Unlike almost all the other dimensions, **date dimension can be built in advance** – five or ten year of history records can be loaded, as well the next few years

- Surrogate key assigned to the date dimension **should be assigned consecutively in the order of date**

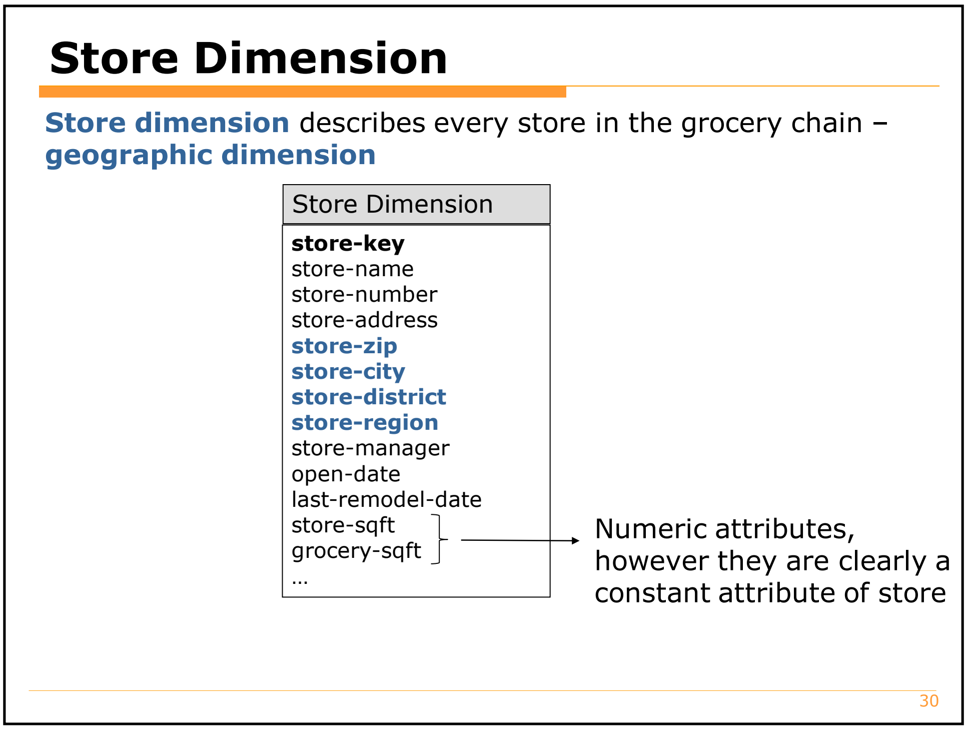


## Product Dimension

**Product dimension** describes every SKU with as many descriptive attributes as possible, including the **existing hierarchies**

| Product Dimension |
|---|
| **product-key** |
| SKU-description |
| SKU-number |
| package-size |
| brand |
| **subcategory** |
| **category** |
| **department** |
| package-type |
| diet-type |
| weight |
| weight-unit |
| … |

It is possible to **browse** among dimension attributes **whether or not they belong to a hierarchy** and it is possible to **roll up** and **drill down** using the attributes that **belong to a hierarchy**

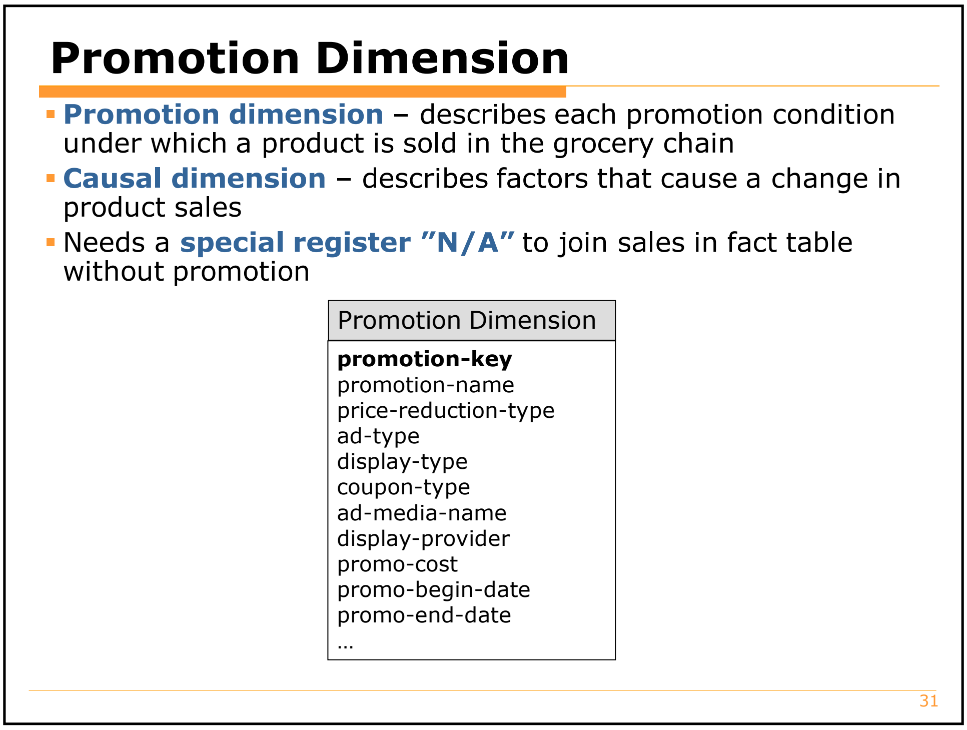


## Store Dimension

**Store dimension** describes every store in the grocery chain – **geographic dimension**

| Store Dimension |
|---|
| **store-key** |
| store-name |
| store-number |
| store-address |
| **store-zip** |
| **store-city** |
| **store-district** |
| **store-region** |
| store-manager |
| open-date |
| last-remodel-date |
| store-sqft |
| grocery-sqft |
| … |

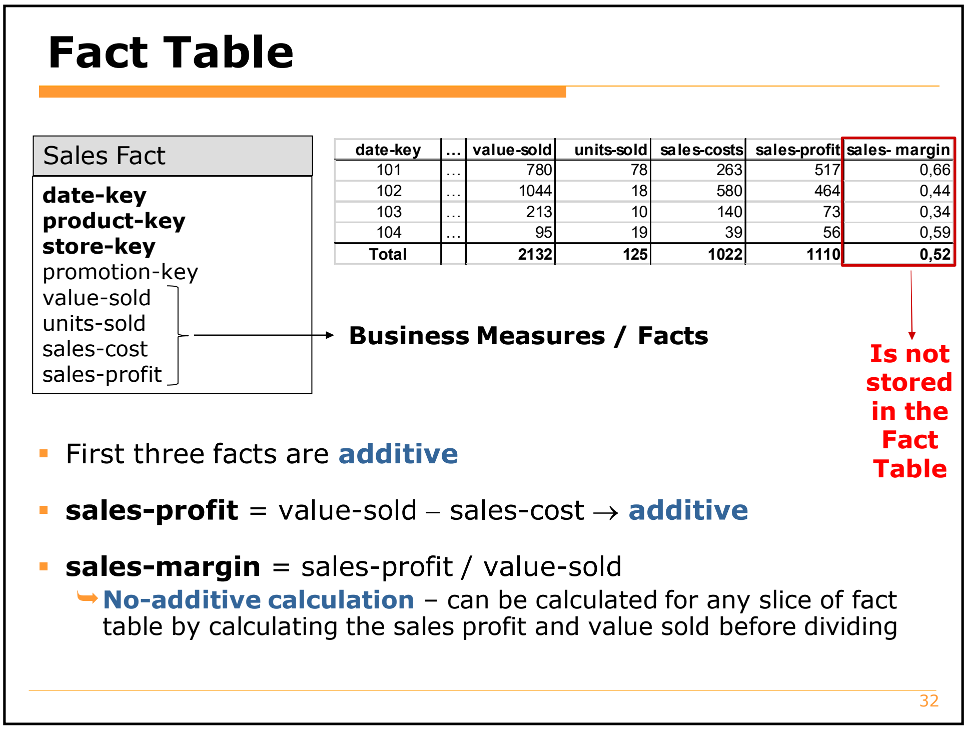
Numeric attributes, however they are clearly a constant attribute of store



# Promotion Dimension

- **Promotion dimension** – describes each promotion condition under which a product is sold in the grocery chain
- **Causal dimension** – describes factors that cause a change in product sales
- Needs a **special register ”N/A”** to join sales in fact table without promotion

| Promotion Dimension |
|---|
| **promotion-key**<br>promotion-name<br>price-reduction-type<br>ad-type<br>display-type<br>coupon-type<br>ad-media-name<br>display-provider<br>promo-cost<br>promo-begin-date<br>promo-end-date<br>… |



# Fact Table

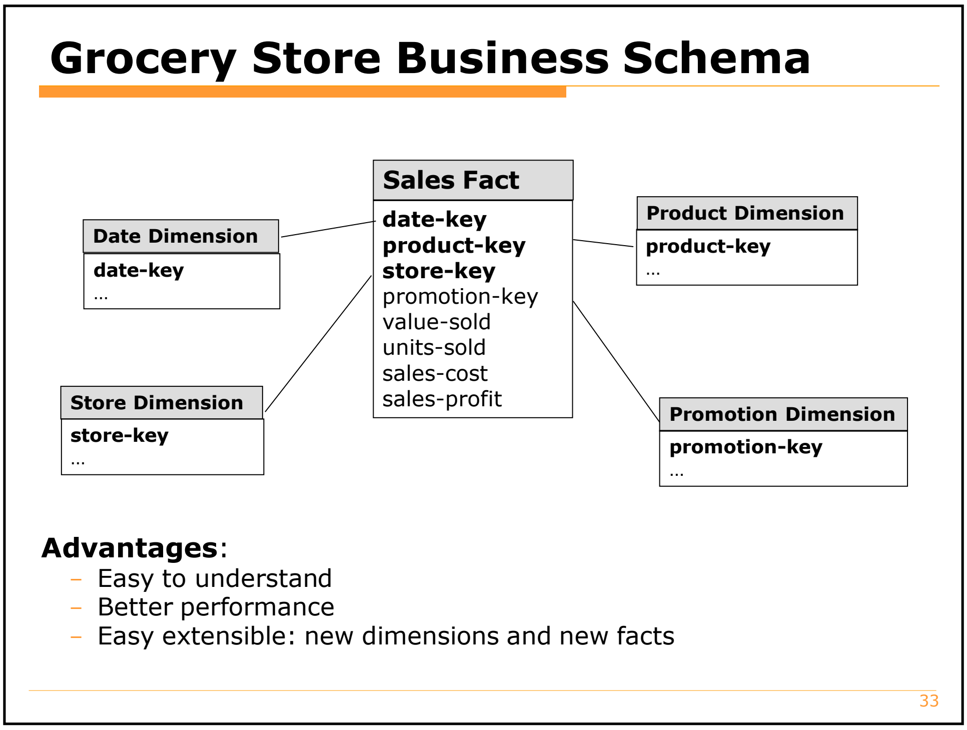
| Sales Fact |
|---|
| **date-key**<br>**product-key**<br>**store-key**<br>promotion-key<br>value-sold<br>units-sold<br>sales-cost<br>sales-profit |

→ **Business Measures / Facts**

| date-key | … | value-sold | units-sold | sales-costs | sales-profit | sales- margin |
|---|---|---|---|---|---|---|
| 101 | … | 780 | 78 | 263 | 517 | 0,66 |
| 102 | … | 1044 | 18 | 580 | 464 | 0,44 |
| 103 | … | 213 | 10 | 140 | 73 | 0,34 |
| 104 | … | 95 | 19 | 39 | 56 | 0,59 |
| **Total** | | **2132** | **125** | **1022** | **1110** | **0,52** |

**Is not stored in the Fact Table**

- First three facts are **additive**
- **sales-profit** = value-sold – sales-cost → **additive**
- **sales-margin** = sales-profit / value-sold
  - **No-additive calculation** – can be calculated for any slice of fact table by calculating the sales profit and value sold before dividing



# Grocery Store Business Schema

**Sales Fact**
**date-key**
**product-key**
**store-key**
promotion-key
value-sold
units-sold
sales-cost
sales-profit

**Date Dimension**
**date-key**
…

**Product Dimension**
**product-key**
…

**Store Dimension**
**store-key**
…

**Promotion Dimension**
**promotion-key**
…

**Advantages**:

- Easy to understand
- Better performance
- Easy extensible: new dimensions and new facts