INTRODUCTION TO BUSINESS INTELLIGENCE
LECTURE 2

University of Gdańsk

Agenda

Structure of the data warehouse

- Type of tables
- Type of schemas

Data warehouse architecture

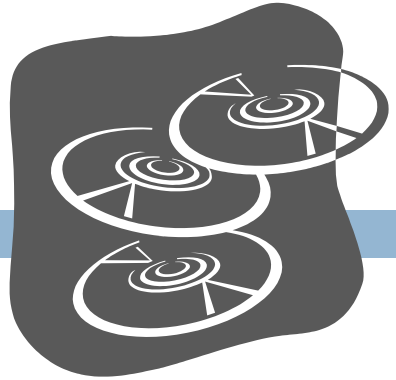
- Type of architectures

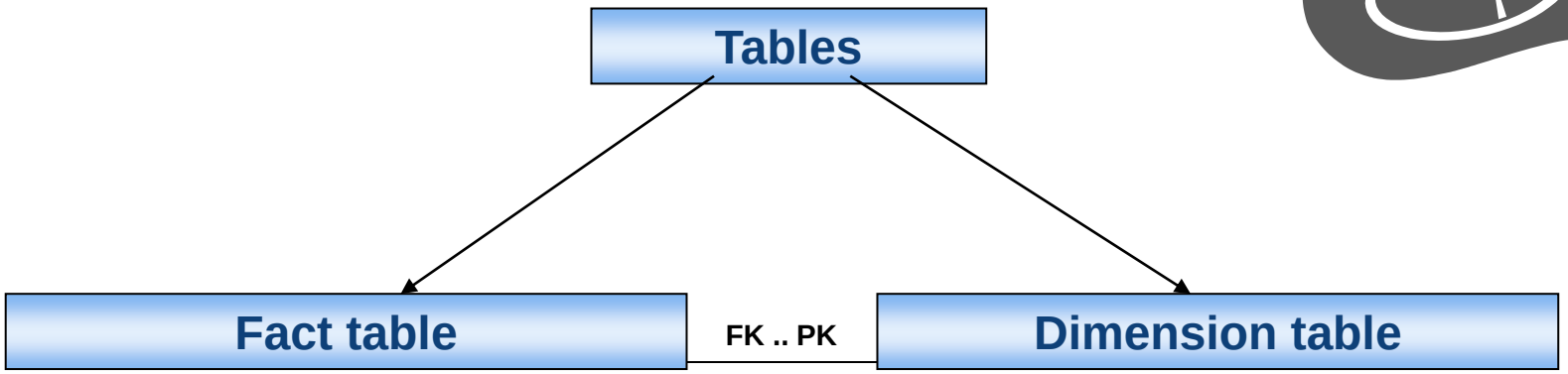
Data types

- Data
- Metadata

Structure of the data warehouse

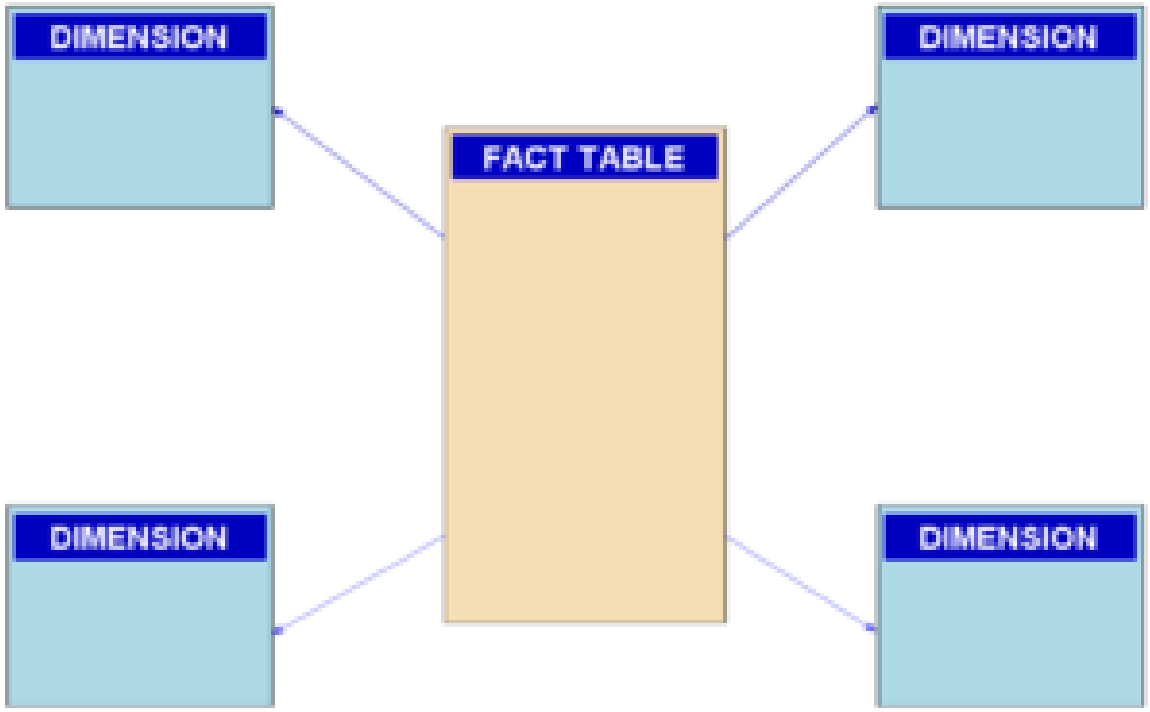
Structure of the data warehouse





Typically there is one fact table and multiple dimension tables.

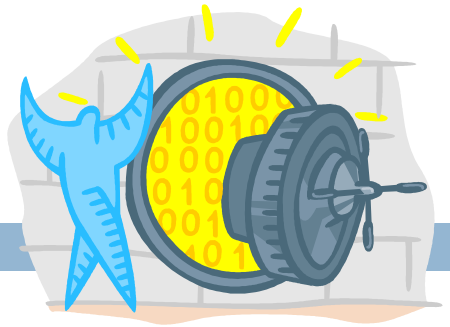
The structure of the data warehouse



Data warehouse structure - the basic principle

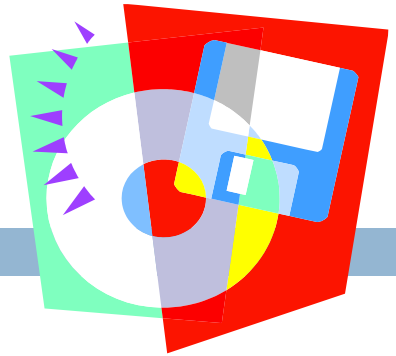
- One of the dimension tables must be TIME, typically containing the following attributes:
  - year
  - quarter
  - month
  - day
- Depending on the needs of data aggregation, there may be more or less of these attributes.

Data warehouse structure - fact table



- Fact tables:
  - even contain hundreds of millions of rows,
  - data is updated on a regular basis,
  - have two types of columns:
  - storing data for later calculations,
  - storing references to dimension tables.
- These tables may contain a mutual relationship.

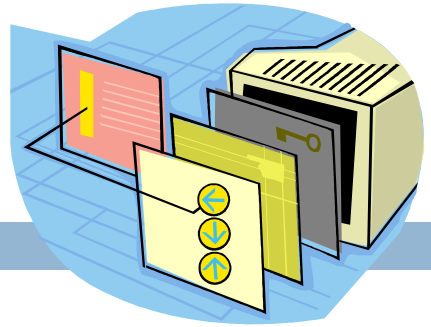
Data warehouse structure - dimension tables



Dimension tables

- describe the events in the fact table,
- can take the form of multidimensional cubes, where each of the dimensions relates to the area of the company's activity,
- they are small in size,
- data rarely changes.

Data warehouse schemas



The most common data warehouse schemas

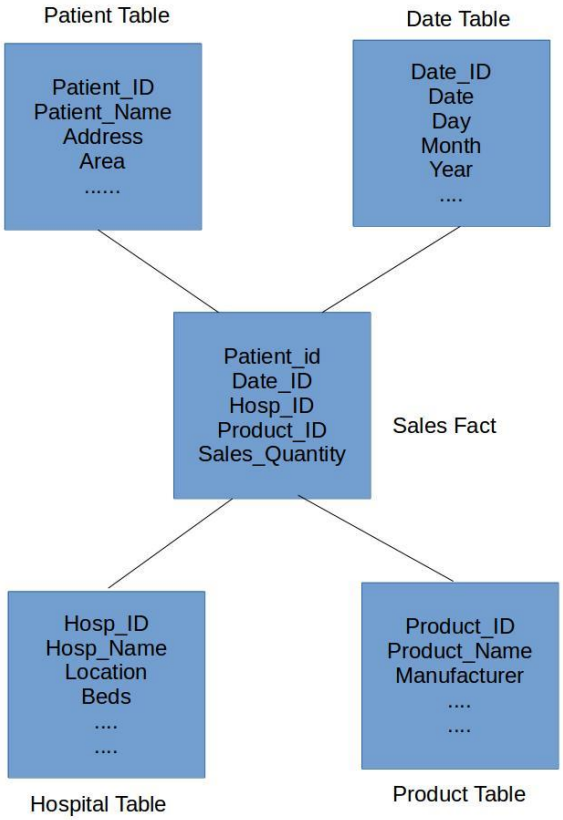
- Star (extended star) schema
- Snowflake schema
- Star constellation schema
- Hybrid schema

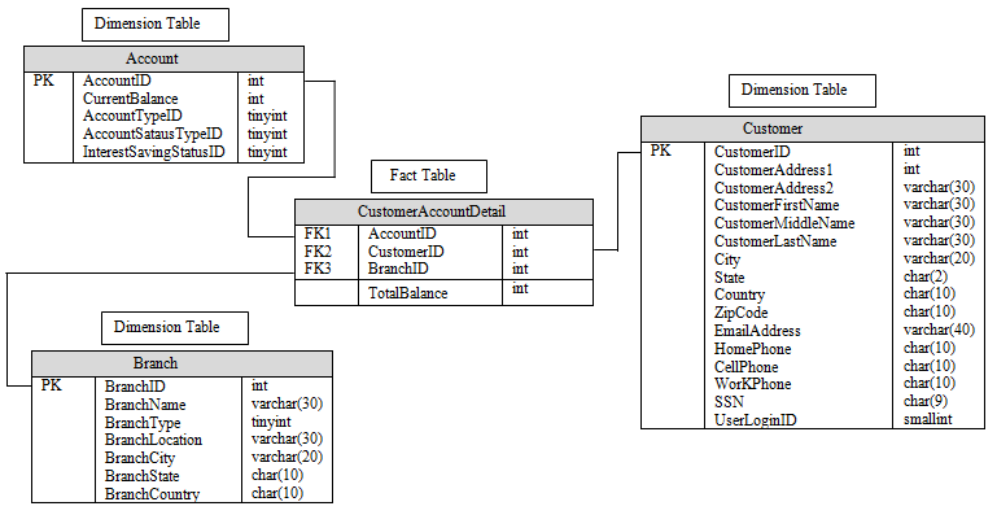
Data warehouse schemas - star schema

Star schema features

- a set of tables in a relational model designed as the basis of a multidimensional model,
- simplicity,
- a small number of tables,
- well-defined connection paths,
- each of the dimensions (region, product, time or other) is directly linked to the fact table (sales, stock level, etc.),
- short response time to inquiries,
- high efficiency of query processing at the expense of data normalization,
- it occurs most often in commercial software.

Data warehouse schemas - star schema



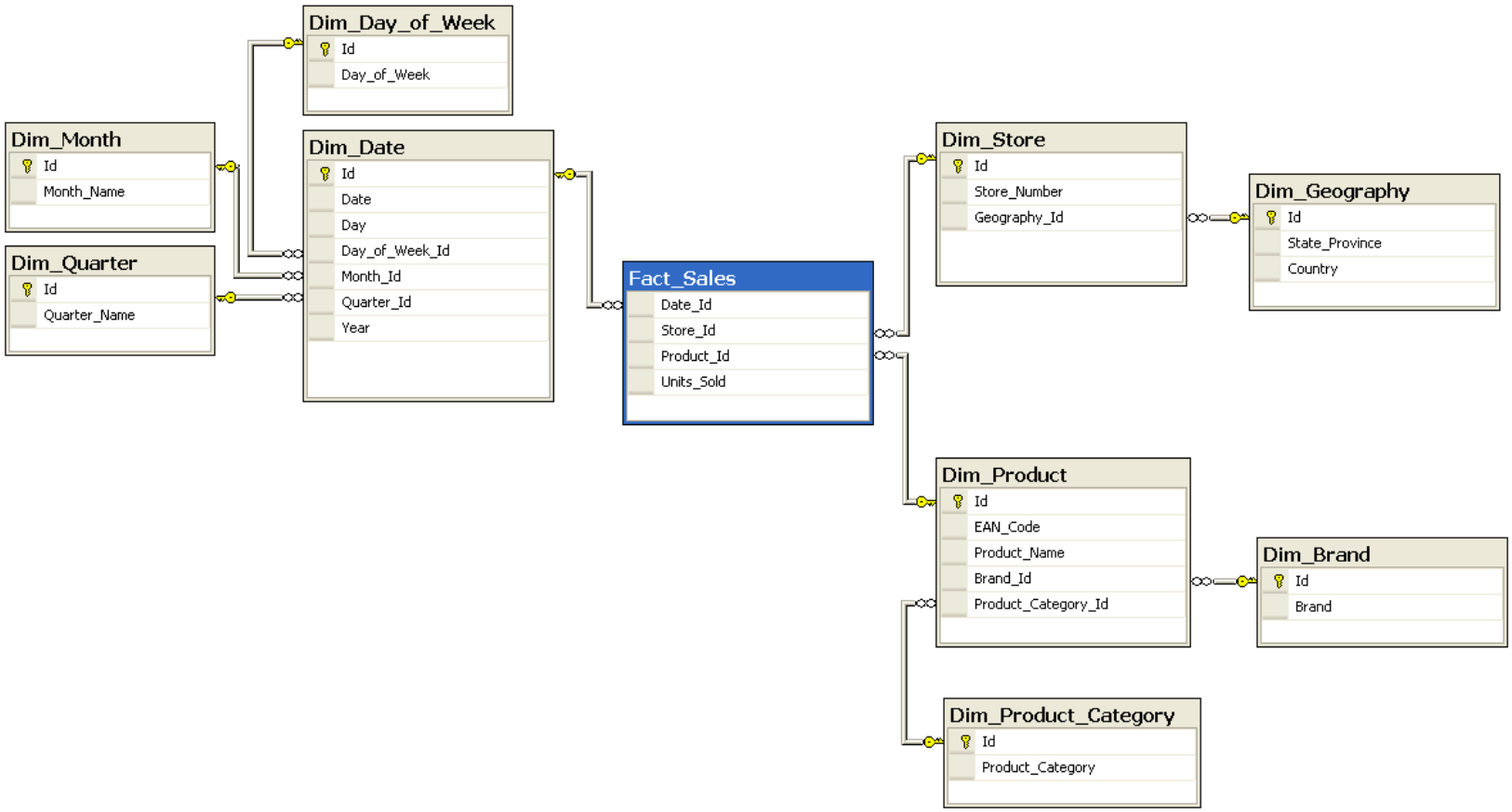


Data warehouse schemas - snowflake schema

Features of the snowflake scheme

- foreign keys can be nested in dimensions,
- dimension tables do not contain denormalized data,
- is created by re-applying the normalization procedure to the star schema dimension tables,
- favors the construction of complex hierarchies of dimensions, making the data model more transparent,
- it is recommended to avoid creating a snowflake schema, if it is not required by the architecture components, the space savings are minimal, while the complexity of the query and reporting process significantly degrades the performance of the data warehouse.

Data warehouse schemas – snowflake schema (normalization)

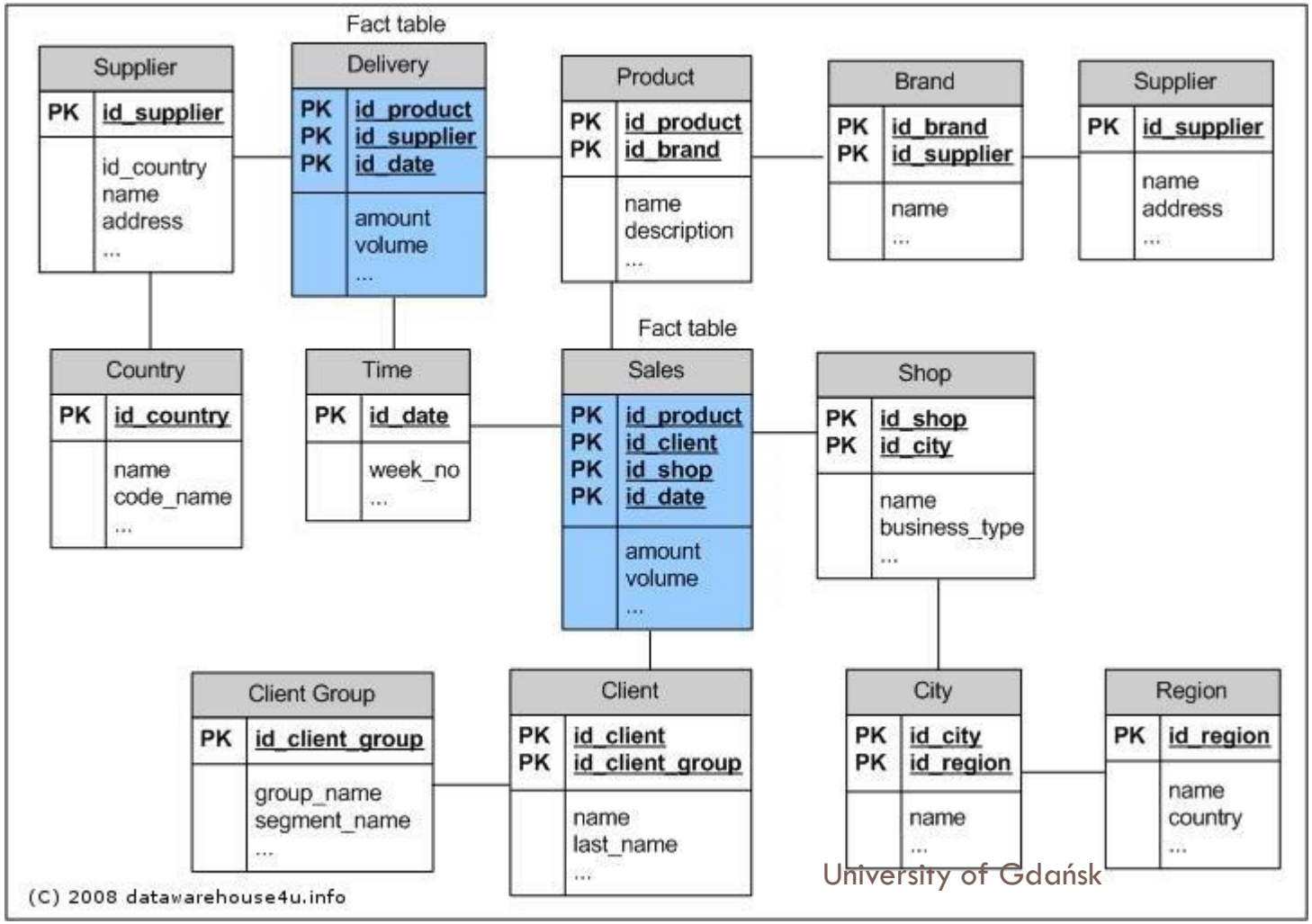


Data warehouse schemas – star constellation schema

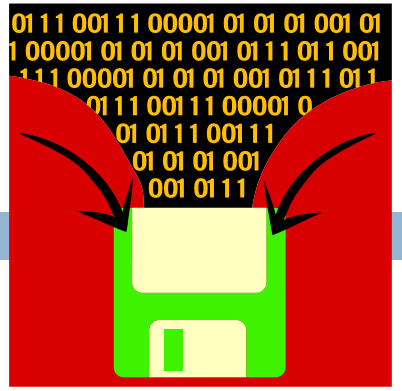
Features of a schema called fact constellation schema

- alternative solution to star schema,
- structurally similar to the star diagram,
- the difference is the presence of multiple fact tables,
- that share dimension tables,
- it is used for more complex applications that require the presence of many fact tables,
- a more complicated form results from different variants for individual types of aggregation,
- dimension tables remain of this size,
- as with the simple star diagram.

Data warehouse schemas – star constellation schema



Data warehouse schemas - data aggregation

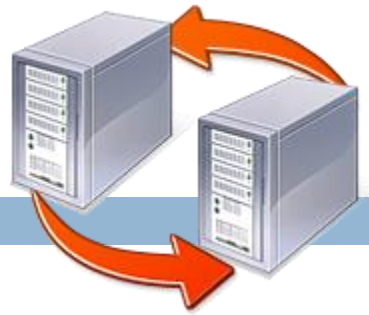


Aggregation:

- to improve query performance,
- reduction of the number of lines.

Most of the reports prepared are not based on very detailed but general data.

Data warehouse schemas - hybrid schema



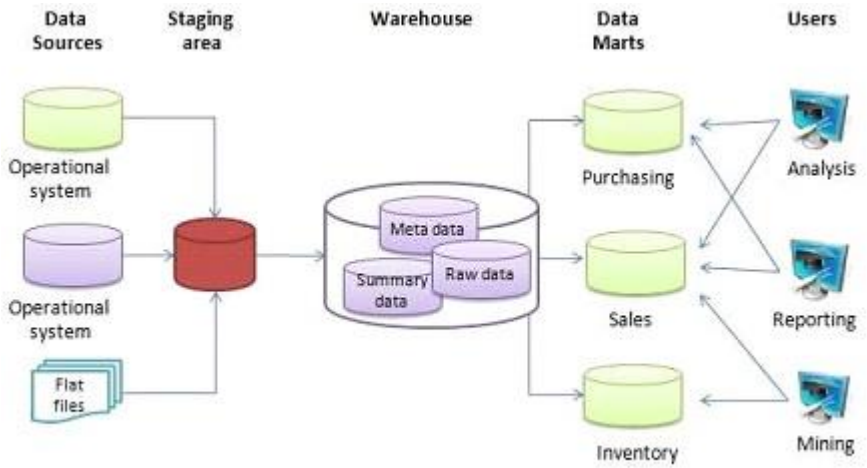
Hybrid schema

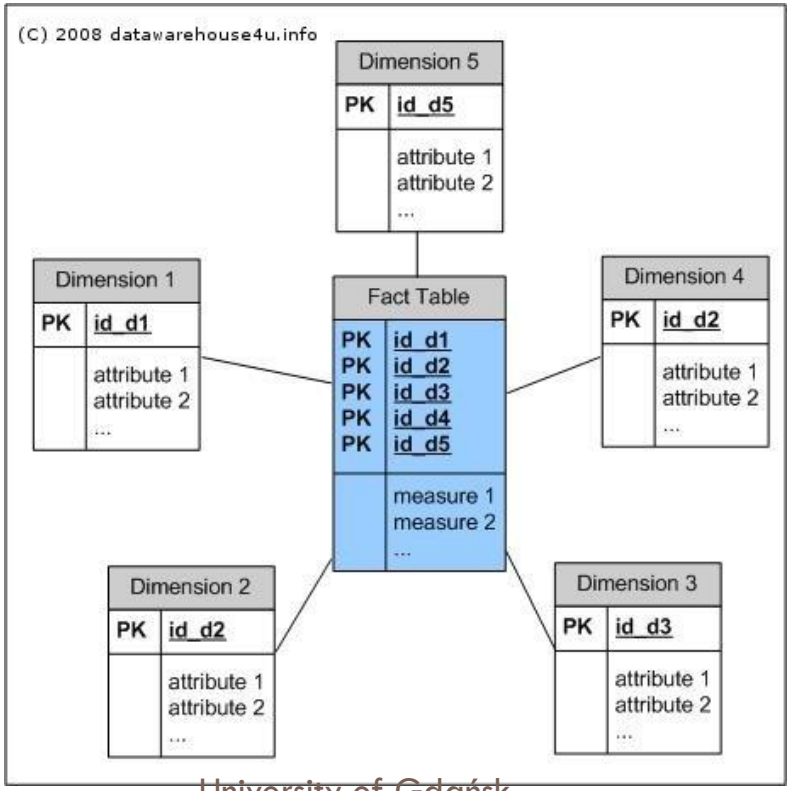
- is created from the combination of existing solutions,
- most often it is a combination of denormalized star schemas with normalized snowflake schemas,
- some dimensions may appear in both forms,
- to satisfy a variety of query requirements.

What will be in the fact table and what will be in the dimension table?

- Product name
- Product description
- Product size
- Product price
- The total number of units of the product sold
- Total value received for the goods sold
- The name of the region where the goods were sold
- The code of the region in which the goods were sold
- Year of sale
- Product type
- Unit of measure

Data warehouse example - summary



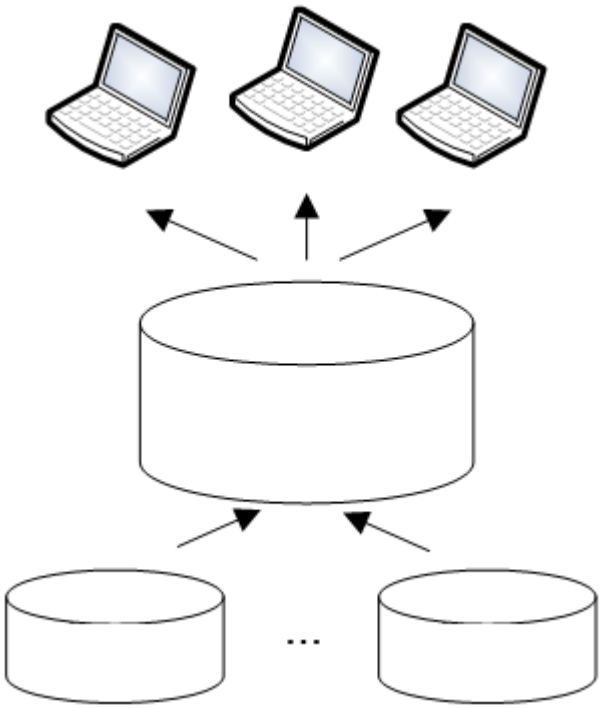


Data warehouse architectures

Centralized architecture

Characteristics:

All data used for analysis in the organization is stored in one data warehouse



Cons:

- lower efficiency compared to distributed systems.

Benefits:

- simplification of access to data resulting from the standardization of the model used,
- creating and maintaining a central database is much easier than in the case of a distributed system.

Application:

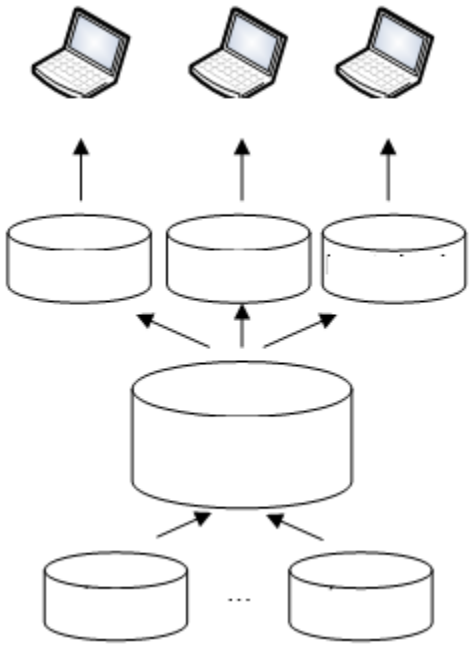
Architecture in a centralized form should be used in those organizations where operational activity is also centralized.

The use of a decentralized architecture brings benefits only in the case of distributed operational data processing.

Federation architecture

Characteristics:

Data that is logically homogeneous but physically stored in different databases located on one or more computer systems.



Cons:

- in this form, the global data warehouse is a purely virtual creation,

- the use of a decentralized architecture brings benefits only in the case of distributed operational data processing.

Benefits:

- as they contain much smaller amounts of data, their data can be presented and analyzed locally at different levels of detail.

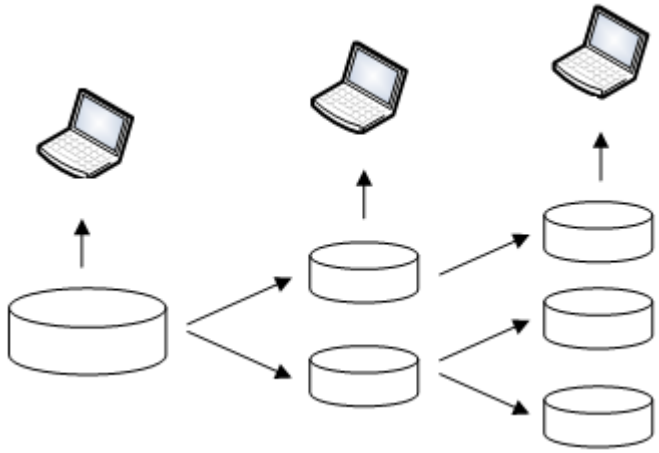
Application:

As part of a local, thematic data warehouse, data specific to a specific department of a given organization are stored.

Layered architecture

Characteristics:

The assumption that the data warehouse is a real physical database.



Cons:

- expansion of the data warehouse requires changes at many levels of computer systems.

Benefits:

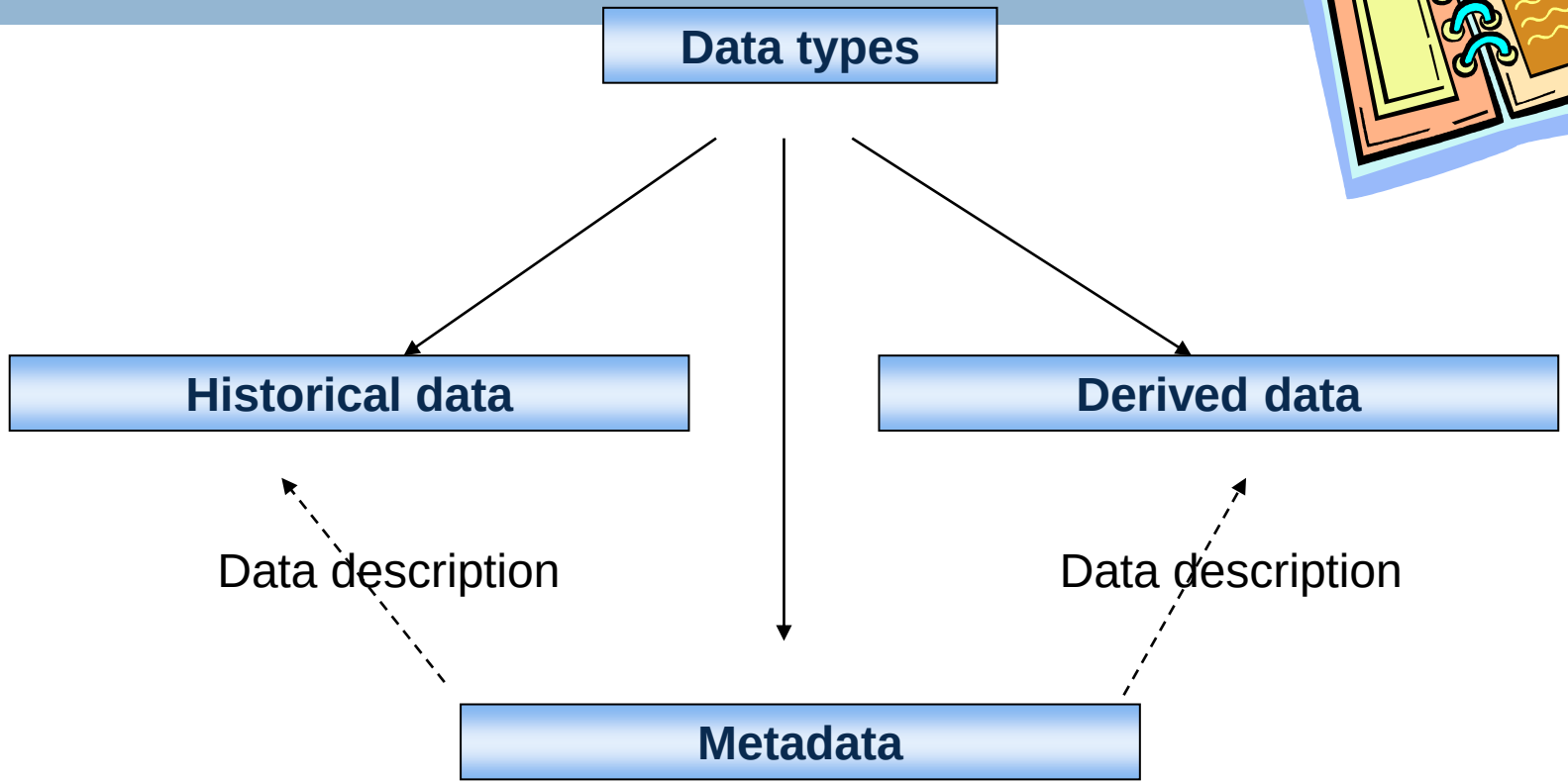
- shorter response time of the data warehouse, because the physical data is located closer to the user,
- reducing the size of the searched database.

Application:

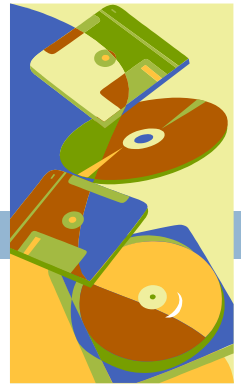
The global warehouse is supplemented by successive levels of local thematic data warehouses, containing copies of the previous layer data or their summaries, without the details present in the federation structure, for example.

Data and metadata

Division of data stored in warehouses



Data breakdown - historical data

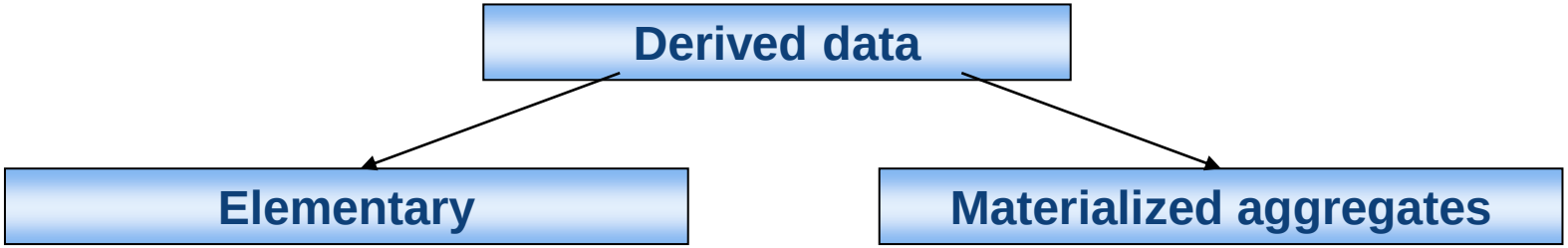


Historical data

- collected over many years of operation of wholesalers,
- placed at the lowest level in the database (fact table),
- they can be archival or transactional data.

Data breakdown - derived data

The data provided is generated from existing data by using mathematical or transforming operations.



- **elementary**, which are a copy of current source data obtained from operational databases and properly processed,

- **materialized aggregates**, which are calculated values in a different cross-section (time, territorial) and at different levels of aggregation (daily, monthly, annual).

Derived data life cycle

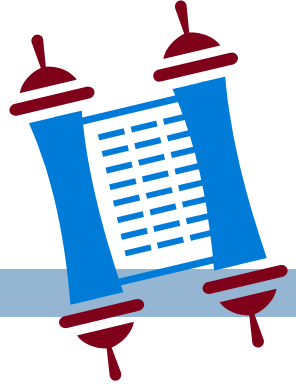
Loading and merging - data is periodically loaded from operational databases, during this process the data is unified

Aggregation - the process of calculating materialized aggregates

Transfer to historical data - elementary data are marked as historical for later analysis aimed at making comparisons of time series

Removal - usually performed very rarely or not at all

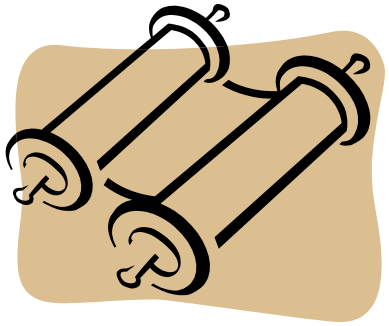
Metadata

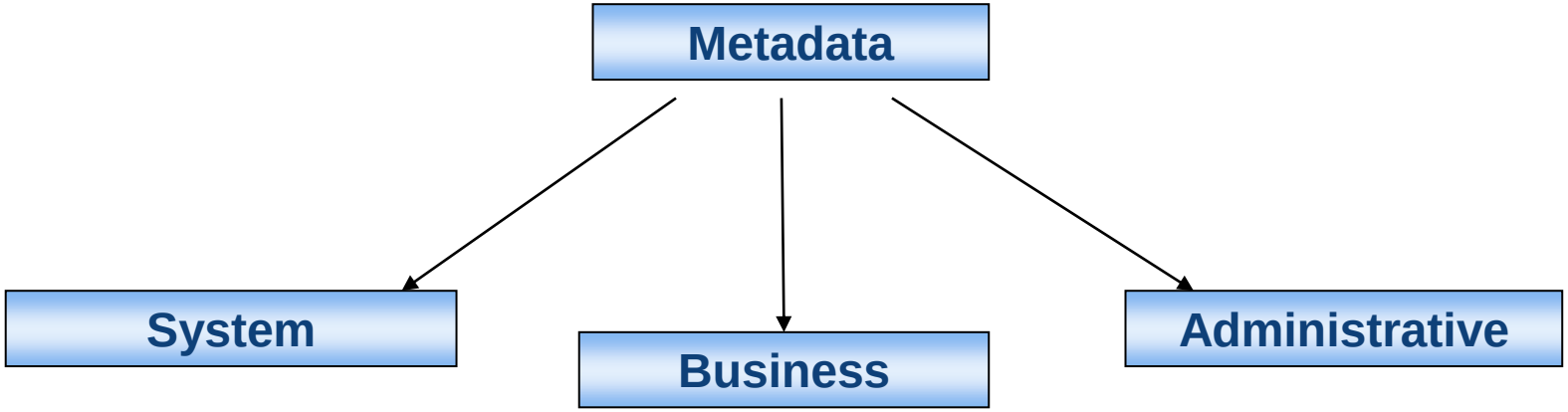


Metadata

- describe the data and diagrams of objects,
- are dictionary information describing the structure of the data warehouse, source databases for the warehouse and the method of calculating aggregate data,
- it is a set of definitions of all data contained in the data warehouse, supplying the warehouse or obtained from it, along with an indication of the places (programs) where these data are used, are used by applications to count and validate data stored in the warehouse.

Metadata





Metadata

Warehouse metadata can be divided into three types:

system - otherwise known as navigation, they describe what types of data are in the system and enable the product to operate properly; they are not changed by the warehouse users as they are used by the system or administrator;

administrative *- or transformational, describing all details of data management, such as data update schedule, joins or division of tables, changes made, retrospective system used, or the source system from which the item was extracted;*

business *- also known as business meaning metadata, targeted directly at users; allow you to find and understand the data contained in the warehouse; often include a search engine that allows you to browse data based on keywords.*

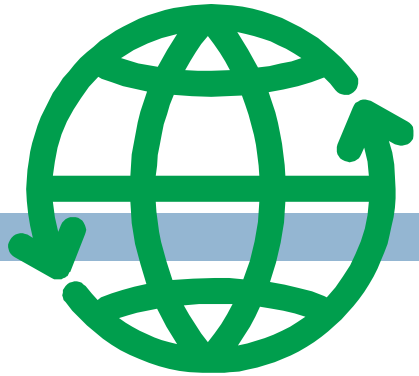
Metadata life cycle

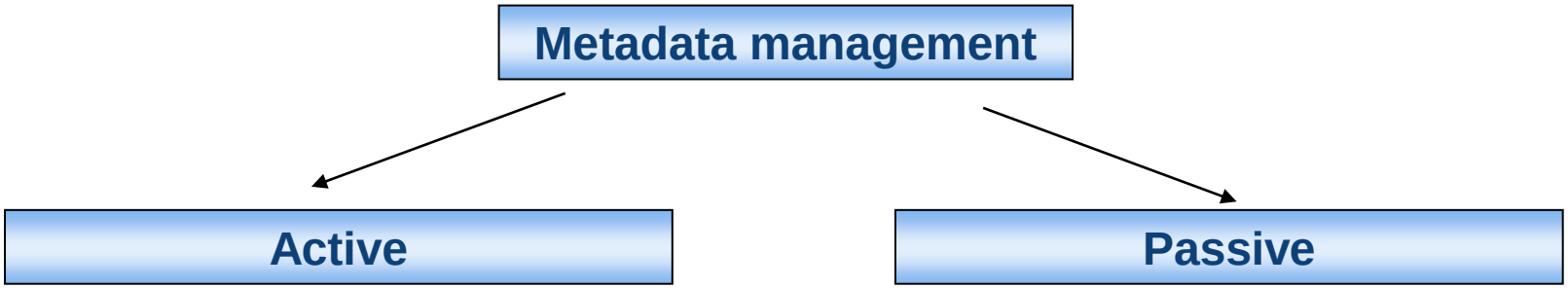
Collecting *- identifying metadata and uploading it to a central repository*

Maintain *- synchronization of metadata automatically with a change of architecture*

Deployment *- delivering metadata to users in the right form through the right tools*

Metadata management





Active - metadata management is part of the data warehouse operation.

Passive - metadata management systems are separate products, placed outside the data warehouse system.

Metadata database

The metadata database may additionally contain

- data dictionaries containing database definitions and relationships between data elements,
- information on data flow, including direction and frequency of the influx of new portions,
- information about the data transformation performed during carrying them,
- the version numbers of the metadata stored,
- information about modifications,
- data usage statistics (data profile),
- names given to individual fields in the database,
- user rights to access data.

Question...

- Which of the following is not a data warehouse schema?
  - Star schema
  - Snowflake schema
  - Extended star schema
  - Solaris schema