

# ***Data warehouse Introduction***

# Decision support systems

- Huge *operational databases* are available in most companies
  - these databases may provide **a large wealth** of useful information
- Decision support systems provide means for
  - in depth analysis of a company's business
  - *faster* and *better* decisions

# Strategic decision support

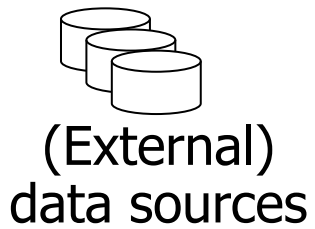
- Demand evolution analysis and forecast
- Critical business areas identification
- Budgeting and management transparency
  - reporting, practices against frauds and money laundering
- Identification and implementation of winning strategies
  - cost reduction and profit increase

# Business Intelligence

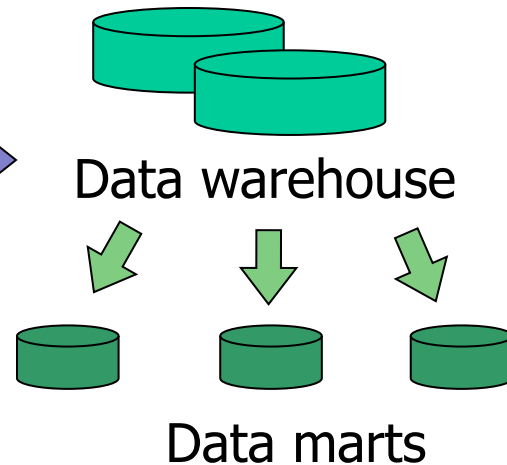
- BI provides support to strategic decision support in companies
- Objective: transforming company data into actionable information
  - at different detail levels
  - for analysis applications
- Users may have heterogeneous needs
- BI requires an appropriate hardware and software infrastructure

# Business intelligence at a glance

## Data sources

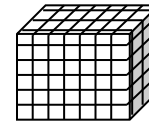


## Data storage



## Data analysis

OLAP Analysis



Data mining

## Result

Data  
visualization

Decision

Knowledge  
management

# Applications

- Manufacturing companies: order management, client support
- Distribution: user profile, stock management
- Financial services: buyer behavior (credit cards)
- Insurance: claim analysis, fraud detection
- Telecommunication: call analysis, churning, fraud detection
- Public service: usage analysis
- Health: service analysis and evaluation

# Type of Information processing

- Transaction processing
- Analytical processing

# Transaction processing

- On Line Transaction Processing (OLTP)
  - Traditional DBMS usage
- Characterized by
  - snapshot of current data values
  - detailed data, relational representation
  - structured, repetitive operations
  - read/write access to few records
  - short transactions
  - isolation, reliability, and integrity are critical (ACID)
  - database size  $\approx$  100MB-GB



# Analytical processing

- On Line Analytical Processing (OLAP)
  - Decision support applications
- Characterized by
  - “historical” data
  - consolidated, integrated data
  - ad hoc applications
  - read access to millions of records
  - complex queries
  - consistency before and after periodical loads
  - database size  $\approx$  100GB-TB

# Data warehouse

- Database devoted to decision support, which is kept *separate* from company operational databases
- Data which is
  - devoted to a specific subject
  - Integrated and consistent
  - time dependent, non volatileused for decision support in a company

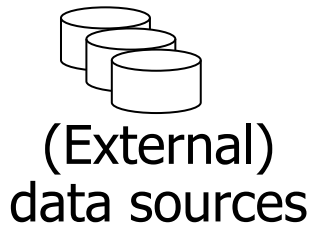
*W. H. Inmon, Building the data warehouse, 1992*

# Why separate data?

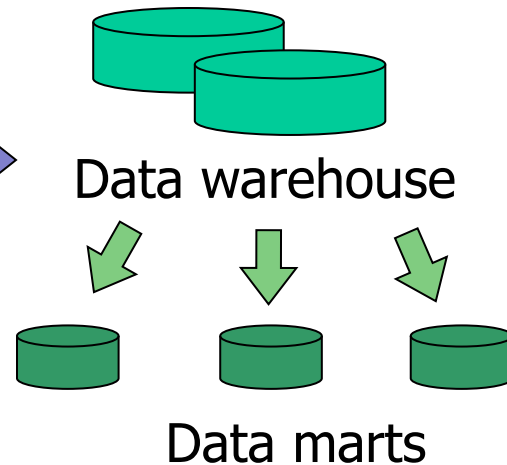
- Performance
  - complex queries reduce performance of operational transaction management
  - different access methods at the physical level
- Data management
  - missing information (e.g., history)
  - data consolidation
  - data quality (inconsistency problems)

# Business intelligence at a glance

## Data sources

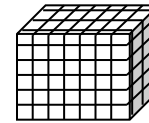


## Data storage



## Data analysis

OLAP Analysis



Data mining

## Result

Data  
visualization

Decision

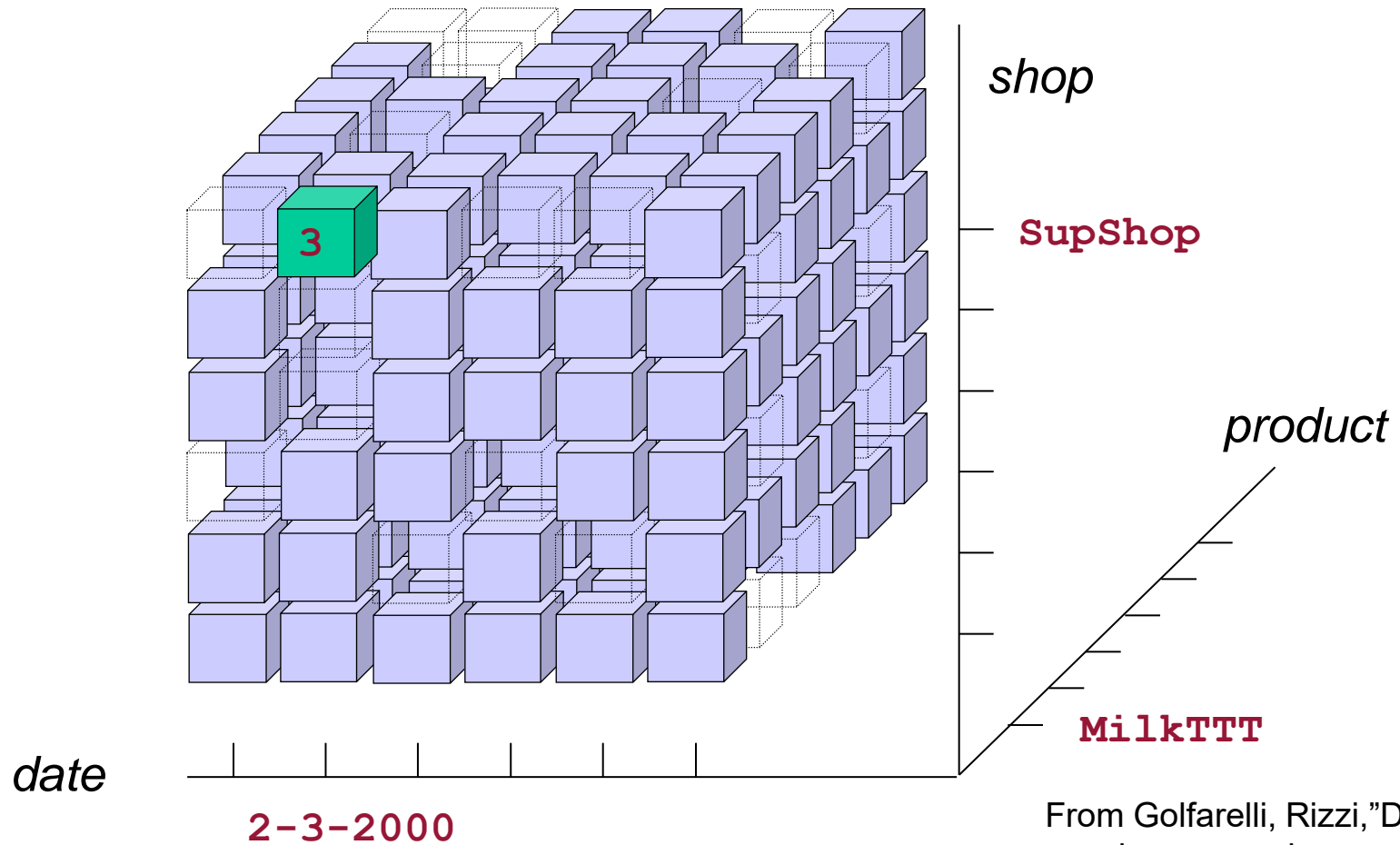
Knowledge  
management

# *Data model*

# Multidimensional representation

- Data are represented as an (hyper)cube with three or more dimensions
- Measures on which analysis is performed: cells at dimension intersection
- Data warehouse for tracking sales in a supermarket chain:
  - dimensions: product, shop, time
  - measures: sold quantity, sold amount, ...

# Multidimensional representation



From Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

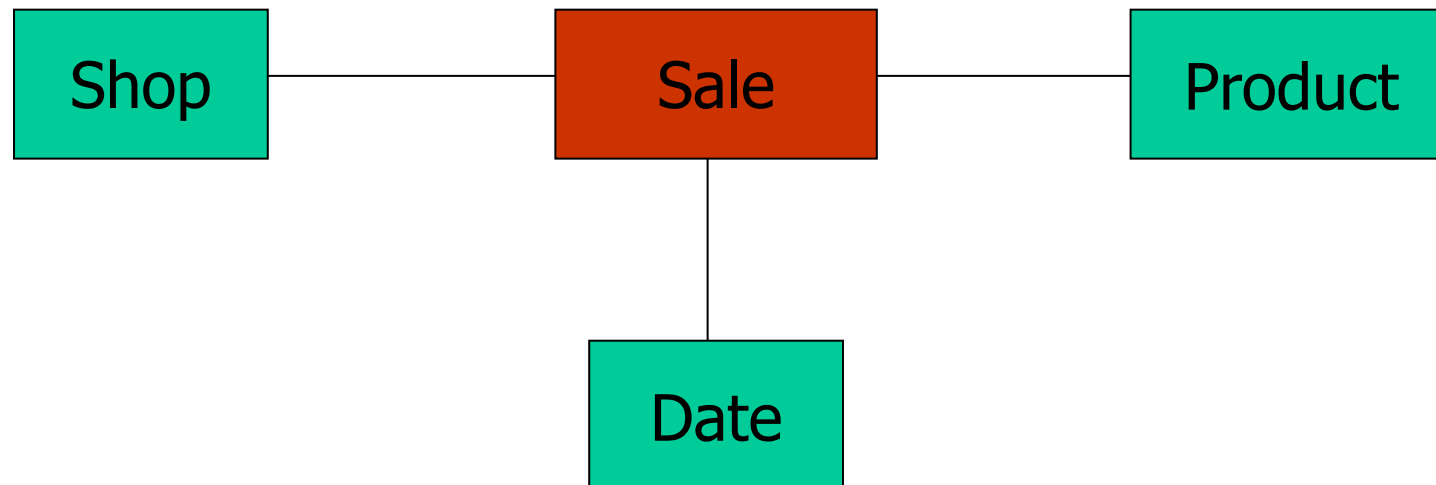
# Relational representation: star model

- Numerical measures stored in the *fact table*
  - attribute domain is numeric
- *Dimensions* describe the context of each measure in the fact table
  - characterized by many descriptive attributes



# Example

Data warehouse for tracking sales in a supermarket chain



# Data warehouse size

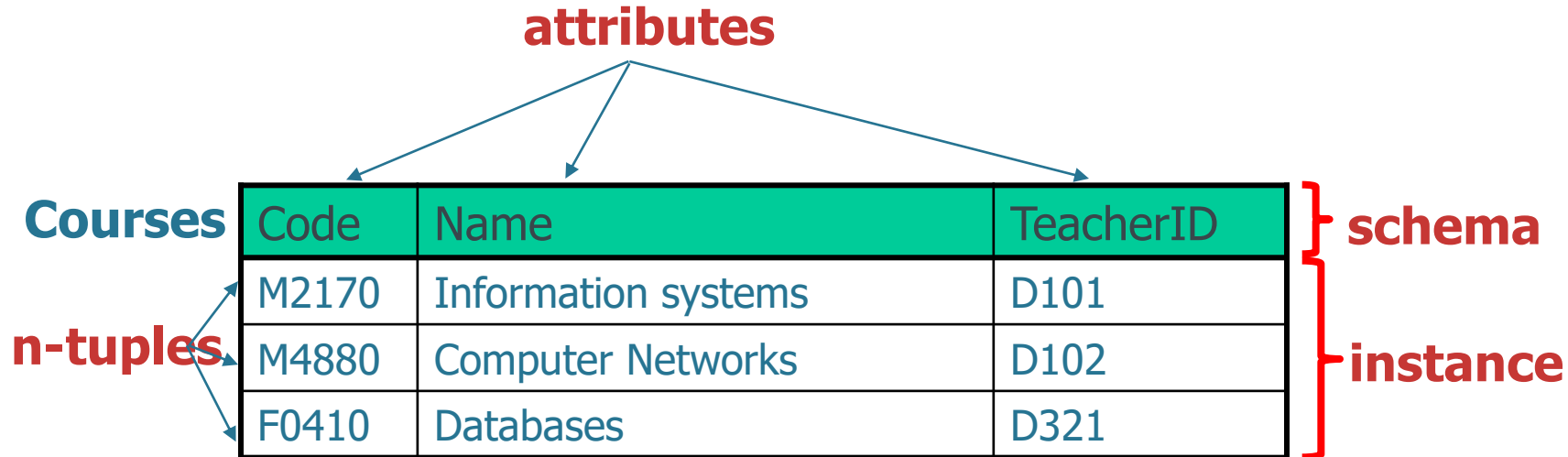
- Time dimension: 2 years x 365 days
- Shop dimension: 300 shops
- Product dimension: 30.000 products, of which 3.000 sold every day in every shop
- Number of rows in the fact table:

$$730 \times 300 \times 3000 = 657 \text{ millions}$$

⇒ Size of the fact table  $\approx$  21GB

# Data representation in relational model

- Proposed by E. F. Codd in 1970 to support higher abstract levels compared to the previous models
  - Data independence
- Made available in commercial DBMSs in 1981,
  - Today it is the main model exploited in commercial DBMSs



# NOSQL data representation

- A database is a set of collections
- Each collection contains a set of documents
- Each document is described by a list of key-value fields and each field can hold any data type
- Documents from the same collection can be heterogeneous
- Since the data representation is schema-less it not required to define the schema of the documents a-priori and objects of the same collections can be characterized by different fields

<b>Relational database</b>	<b>NOSQL database</b>
Table	Collection
Row	Document
Column	Field

# Example of Document Data

- Records are stored into Documents
  - field-value pairs
  - similar to JSON objects
  - may be nested

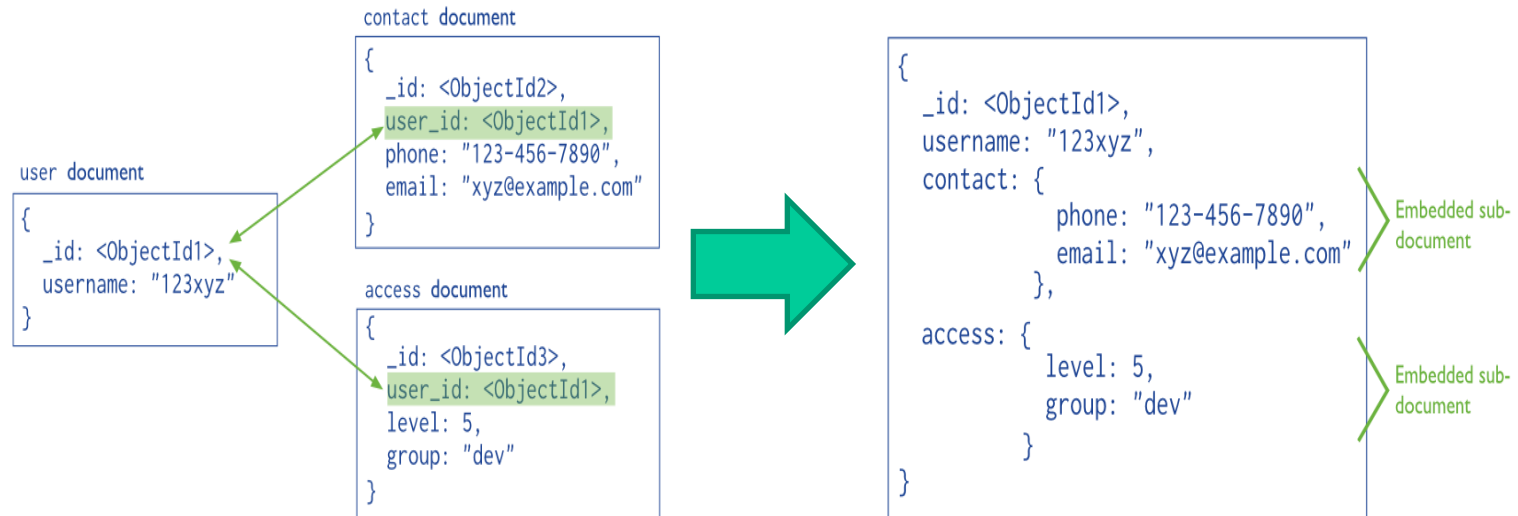
```
{
  _id: <ObjectId>,
  username: "123xyz",
  contact: {
    phone: "123-456-7890",
    email: "xyz@example.com"
  },
  access: {
    level: 5,
    group: "dev"
  }
}
```

Embedded sub-document

Embedded sub-document

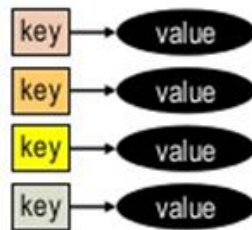
# Example of Document Data

- Relations among documents are inefficient, and leads to de-normalization
  - Object(ID) reference, with **no native join**

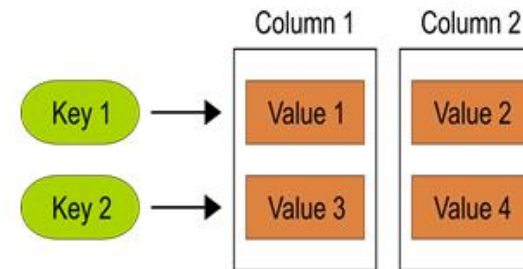


# Types of NoSQL databases

## Key-Value



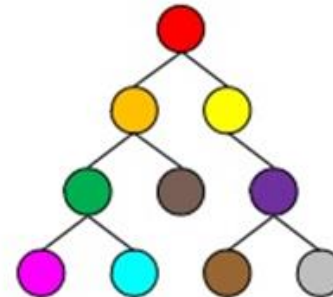
## Column-Family



## Graph



## Document



# *Data analysis*



# Data analysis tools

- OLAP analysis: complex aggregate function computation
  - support to different types of aggregate functions (e.g., moving average, top ten)
- Data analysis by means of data mining techniques
  - various analysis types
  - significant algorithmic contribution

# Key Performance Indicator (KPI)

- KPIs are measurable values that demonstrates how effectively a company is achieving key business objectives.
- They are used to periodically assess at multiple levels the performance of organizations and their success at reaching targets
  - high-level KPIs may focus on the overall performance of the business
  - low-level KPIs may focus on processes in specifics areas/departments (e.g, sales, marketing, HR).
- One of the most important aspects of KPIs is that they are a form of communication.
- Example KPIs: Days to deliver an order, number of new customers acquired, employee satisfaction, ...

# Data analysis tools

- Presentation
  - separate activity: data returned by a query may be rendered by means of different presentation tools
- Motivation search
  - Data exploration by means of progressive, “incremental” refinements (e.g., drill down)

# *Data visualization*

# Informative Dashboard

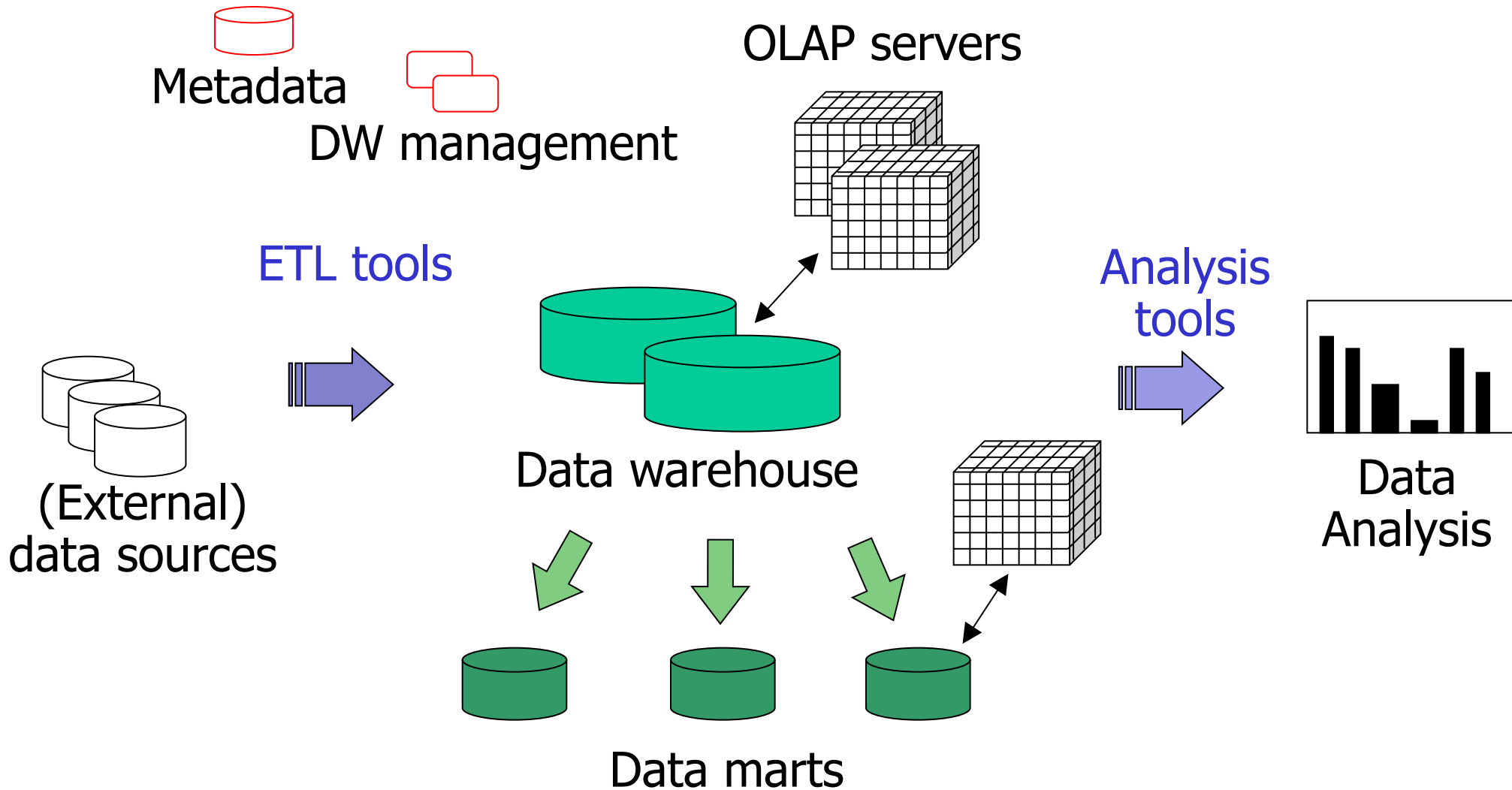
- A dashboard is a user interface that organizes and presents information in a way that is easy to read.
- It is a visual display of the most important information needed to achieve one or more objectives
- Dashboard are small and concise to allow monitoring relevant phenomena at a glance
- Visual Business Intelligence for enlightening analysis and communication

# *Data warehouse architectures*

# Data warehouse architectures

- Separation between transactional computing and data analysis
  - avoid one level architectures
- Architectures characterized by two or more levels
  - separate to a different extent data incoming into the data warehouse from analyzed data
  - more scalable

# Data warehouse: architecture





# Data warehouse and data mart

*Company data warehouse:* it contains *all* the information on the company business

- extensive functional modelling process
- design and implementation require a long time

*Data mart:* departmental information subset focused on a given subject

- two architectures
  - dependent, fed by the company data warehouse
  - independent, fed directly by the sources
- faster implementation
- requires careful design, to avoid subsequent data mart integration problems

# Servers for Data Warehouses

- ROLAP (Relational OLAP) server
  - extended relational DBMS
    - compact representation for sparse data
  - SQL extensions for aggregate computation
  - specialized access methods which implement efficient OLAP data access
- MOLAP (Multidimensional OLAP) server
  - data represented in proprietary (multidimensional) matrix format
    - sparse data require compression
  - special OLAP primitives
- HOLAP (Hybrid OLAP) server
- NOSQL architectures

# Extraction, Transformation and Loading (ETL)

- Prepares data to be loaded into the data warehouse
  - data extraction from (OLTP and external) sources
  - data cleaning
  - data transformation
  - data loading
- Performed
  - when the DW is first loaded
  - during periodical DW refresh

# ETL process

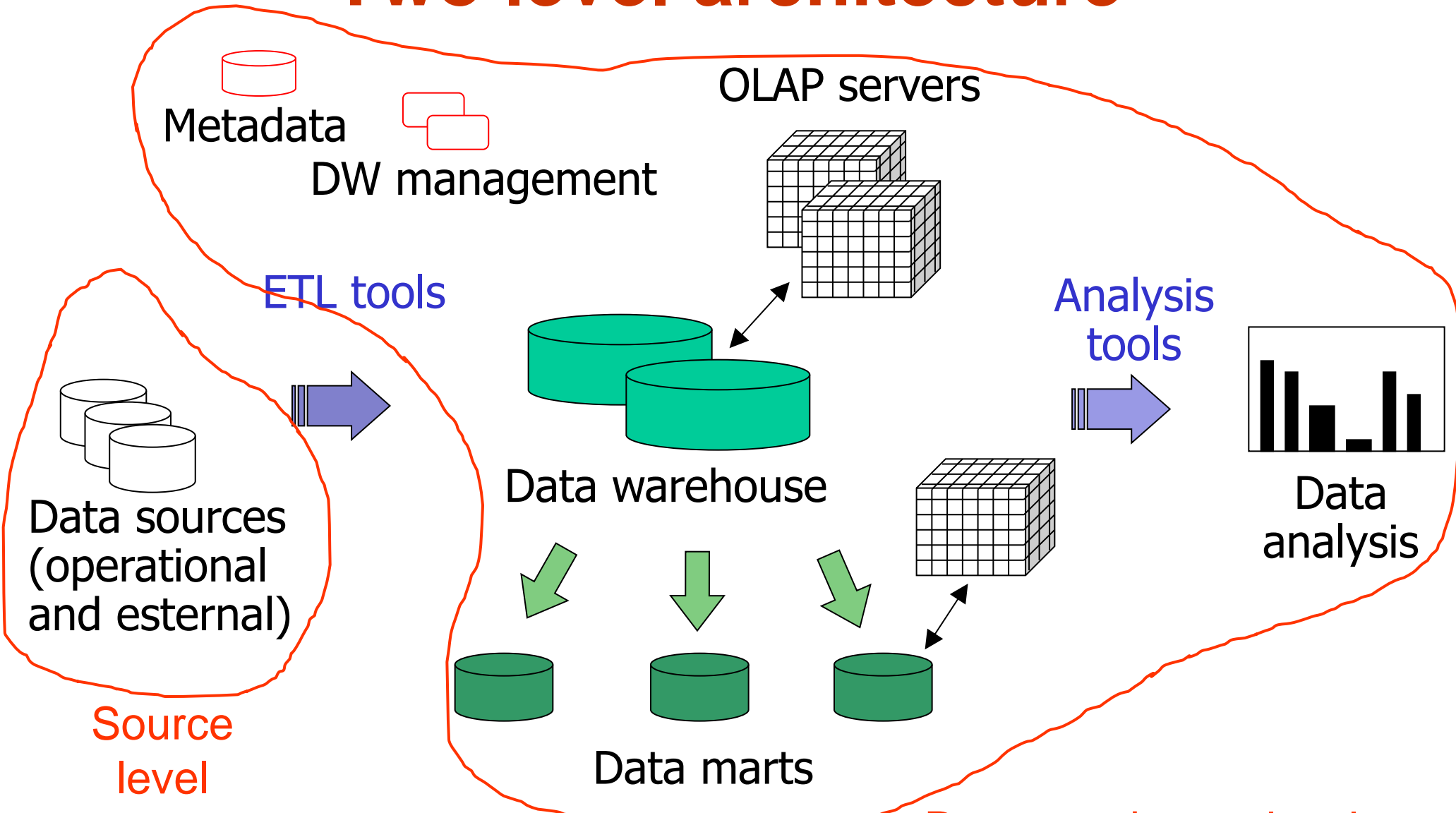
- *Data extraction*: data acquisition from sources
- *Data cleaning*: techniques for improving data quality (correctness and consistency)
- *Data transformation*: data conversion from operational format to data warehouse format
- *Data loading*: update propagation to the data warehouse

# Metadata

metadata = data about data

- Different types of metadata:
  - for data transformation and loading: describe data sources and needed transformation operations
    - Useful using a common notation to represent data sources and data after transformation
    - CWMI (Common Warehouse Metadata Initiative): standard proposed by OMG to exchange data between DW tools and repository of metadata in heterogenous and distributed environments
  - for data management: describe the structure of the data in the data warehouse
    - also for materialized view
  - for query management: data on query structure and to monitor query execution
    - SQL code for the query
    - execution plan
    - memory and CPU usage

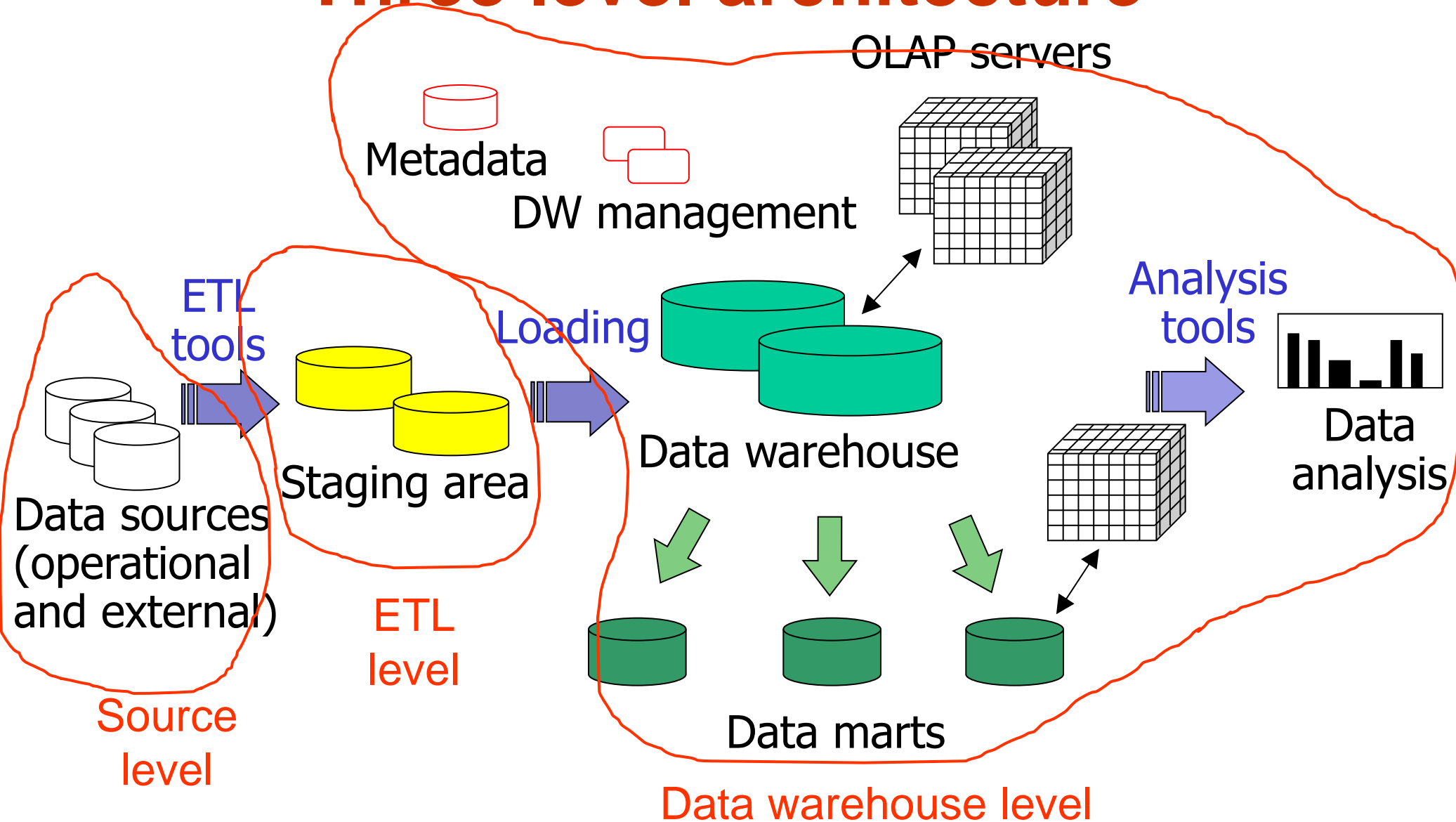
# Two level architecture



# Two level architecture features

- Decoupling between source and DW data
  - management of external (not OLTP) data sources (e.g., text files)
  - data modelling suited for OLAP analysis
  - physical design tailored for OLAP load
- Easy management of different temporal granularity of operational and analytical data
- Partitioning between transactional and analytical load
- “On the fly” data transformation and cleaning (ETL)

# Three level architecture



From Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006



# Three level architecture features

- *Staging area*: buffer area allowing the separation between ET management and data warehouse loading
  - complex transformation and cleaning operations are eased
  - provides an integrated model of business data, still close to OLTP representation
  - sometime denoted as Operational Data Store (ODS)
- Introduces further redundancy
  - more disk space is required for data storage