VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

Introduction to Data Warehouse

Lecturer : *Dr. Pavle Mogin*

SWEN 432
Advanced Database Design and Implementation

Plan for Introduction to Data Warehouse

- Data Warehouse as an approach to data organization for efficient Decision Support Systems (DSS)
- The basic characteristics of Decision Support Systems
- The structure of a simple Decision Support System
- Main differences between operational and DSS data

  - *Readings:*
    - *Ramakrishnan, Gehrke: "Database Management Systems", Chapter 25, Section 25.1*
    - *S. Chaudhuri, U. Dayal:*

      An Overview of Datawarehousing and OLAP Technology

Origins

- A new approach to organization of Decision Support Systems (DSS) data emerged at the start of nineteen nineties
- This new approach was named the Data Warehouse
- To understand better the Data Warehouse approach to data organization, it is necessary to learn more about DSS and On-line Analytical Processing (OLAP)

Decision Support System (DSS)

- A Decision Support System is a programming system that is aimed to help managers during the process of business decision making
- The main goal of this decision making process is to enhance the success of the company
- The use of decision support systems can be illustrated by the following two questions:
  - What is the ratio of productivity raise in various organizational units of the company during the period of last three years?
  - What is the ratio between product advertising and its selling?
- DSS helps in giving answers to such questions by using historical operational data as the input to mathematical business models

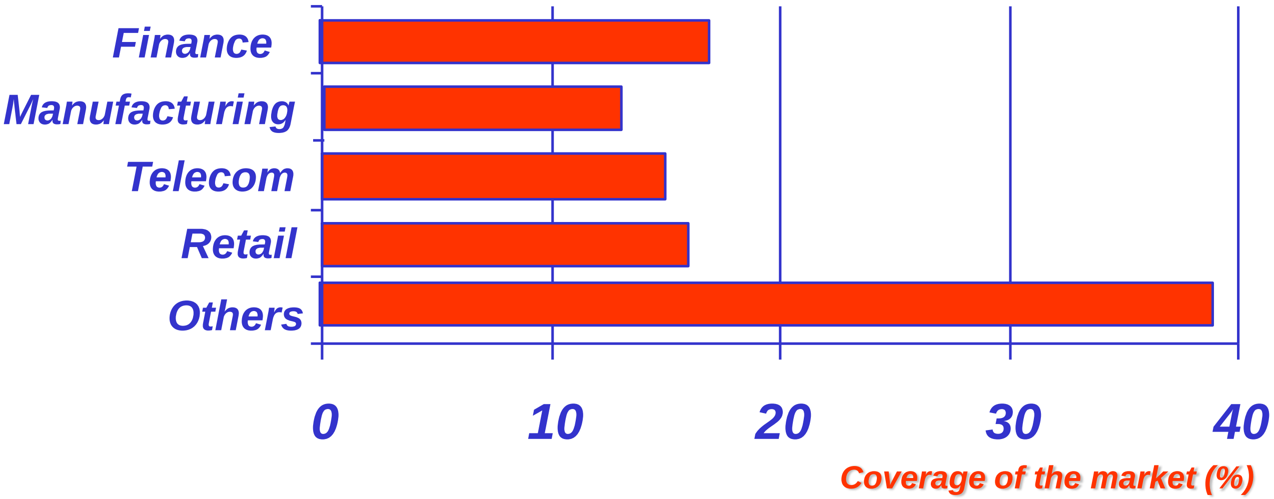
Common Characteristics of DSS

- Decision Support Systems are used in:
  - Manufacturing for order shipment and customer support,
  - Financial services for claims analysis, risk analysis, credit card analysis, and fraud analysis,
  - Retail for user profiling and inventory management,
  - Transportation for fleet management,
  - Telecommunication for call analysis and fraud detection
  - Healthcare for drug consumption analysis and outcome analysis,
  - Utilities for power usage analysis
- Despite the fact that the decision support systems are used in different fields, they posses a common general structure

Industries Using DW Systems

Industries

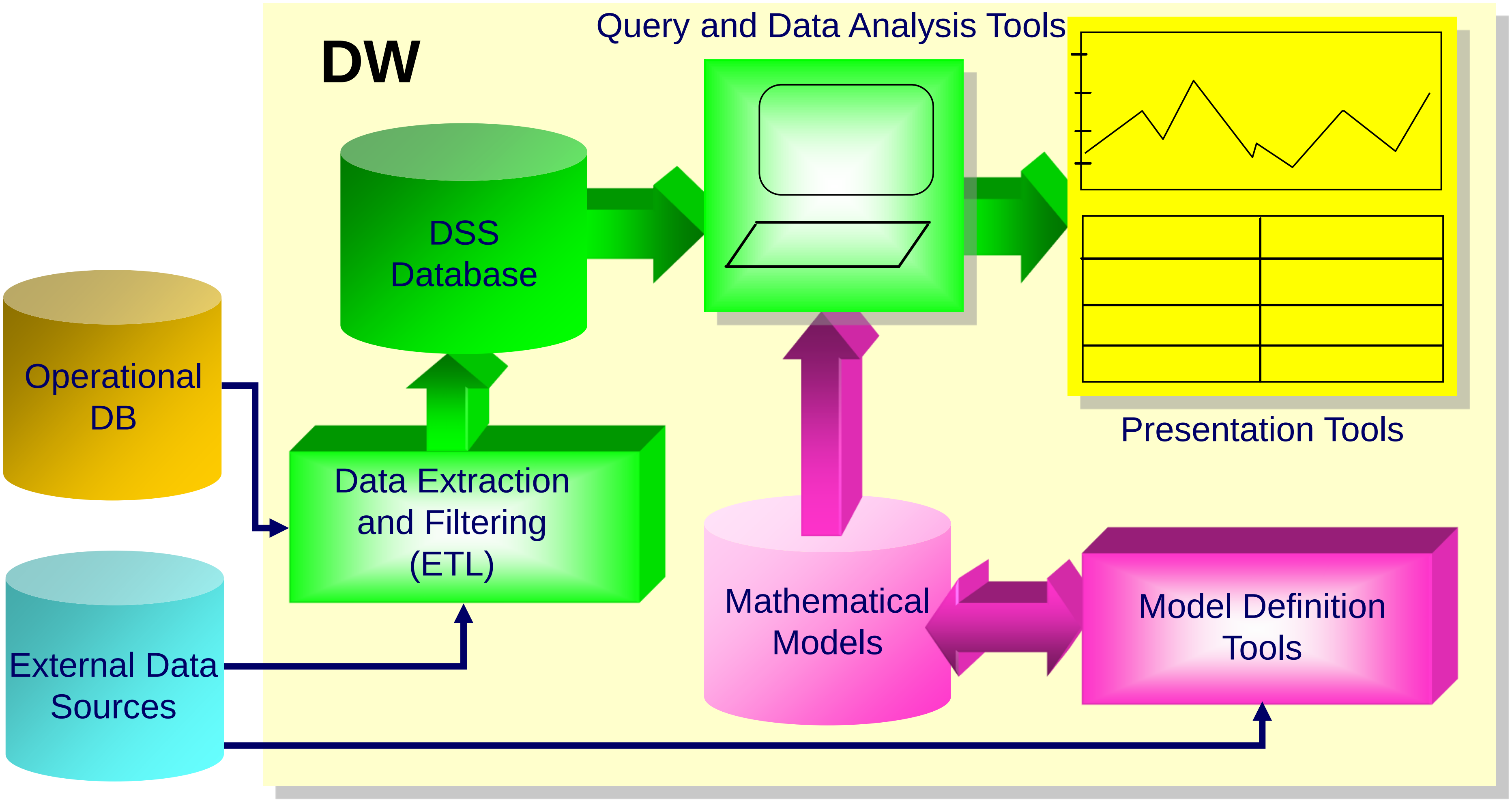
- Airline
- Banking
- Health care
- Investment
- Insurance
- Retail
- Telecommunications
- Manufacturing
- Credit card suppliers
- Clothing distributors



Structure of a DSS

- The structure of a DSS system consists of the following components:
  - A data warehouse data storage
  - A data extraction, transformation and loading component (ETL),
  - Mathematical models,
  - A query tool, and
  - A data presentation component

The General Structure of a DW



DSS Database

- DSS database contains business data
- Business data represent a sequence of states of the company in different moments of time
- These data does not represent mere copy of an operational database
- Business data are produced by aggregating and restructuring operational data

Operational versus DSS Data

- Operational and DSS data are used for different purposes
- Operational data are contained in a third normal form relational database
- Normalization enhances effective database updating, but degrades query performance
- Further, an operational database system is optimized to support execution of a large number of short transactions in a concurrent, or parallel environment

Operational versus DSS Data

- Operational and DSS data differ according to:
  - Time interval,
  - Granularity,
  - Dimensionality,
  - Way of using (update versus querying), and
  - Volume

Time Interval

- Operational data pertain to current, elementary business documents or events like:
  - Purchase order,
  - Invoice
  - Change of stock, or
  - Number of certain items on stock
- All these data have short life span
  - Very fast they become obsolete and archived (made off-line)
  - Deleted (or overwritten) from on-line databases
- For the business decision making are different data needed

Time Interval

- Examples of DSS data:
  - Sales data of a particular product (or a group of products) during the last month, last year, or even during last five years,
  - Total sale amounts during a given time interval,
  - Relationship between purchase, expenses, and profit for product groups, during a time period
- DSS data are historical, they span a long time interval

More about Time Difference

- Operational data, most often, contain only the last updated value
  - It is understood to be “real-time”,
  - It even may not have an explicite reference to a time interval,
  - If old data images are archived, they are rarely kept on-line after the end of a fiscal year
- DSS data have explicitly designated time interval
  - Sales for the last week
  - Sales in November 2014
  - Expenses for last three years
- Very often, DSS data contain aggregates of the same type but over different time intervals

Granularity

- Operational data pertain to individual transactions
- DSS data are produced by aggregating operational data
- Besides, the same operational data can be aggregated on several levels for the purpose of decision making

Granularity

- Operational data have fine granularity
- DSS data have coarser granularity
- Example:
  - When a data analyst analyses sales according to:
    - Geographic districts,

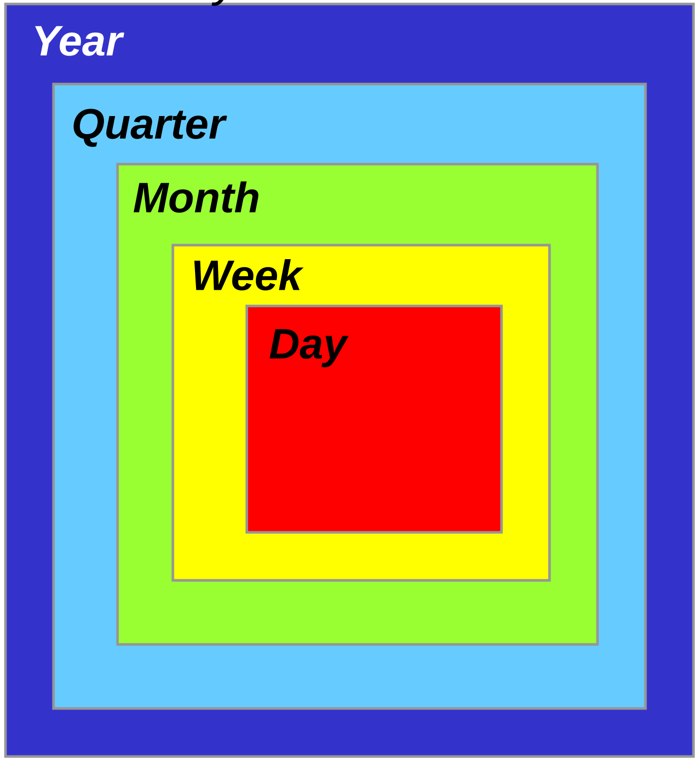
    he/she may need sales data pertaining to:
    - Cities in districts, and even
    - Shops in cities,

    but he / she may not be interested in individual purchases
  - Individual purchases define the granularity of an operational database

Granularity

- The data aggregation level is defined by the “finest” granularity of context data

Granularity Levels of Time



Year

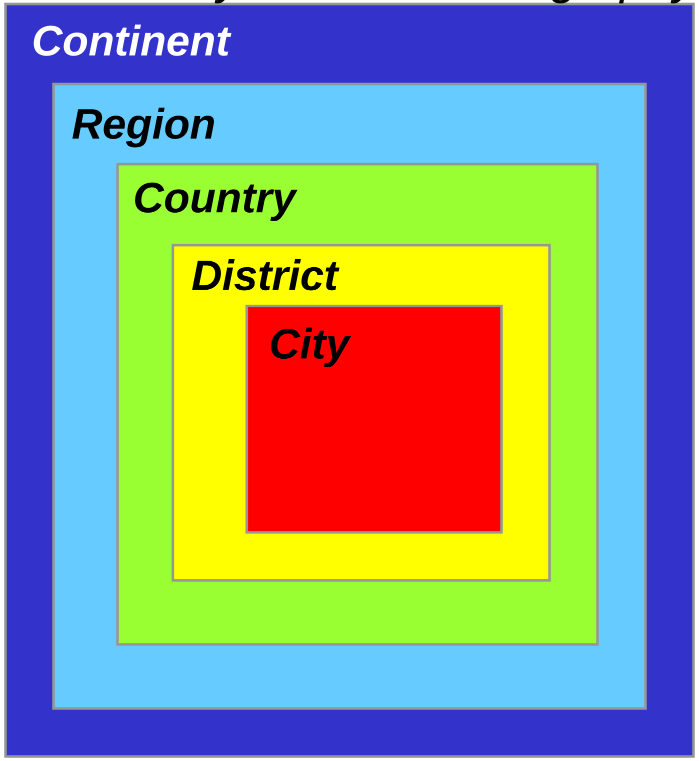
Quarter

Month

Week

Day

Granularity Levels of Geography



Continent

Region

Country

District

City

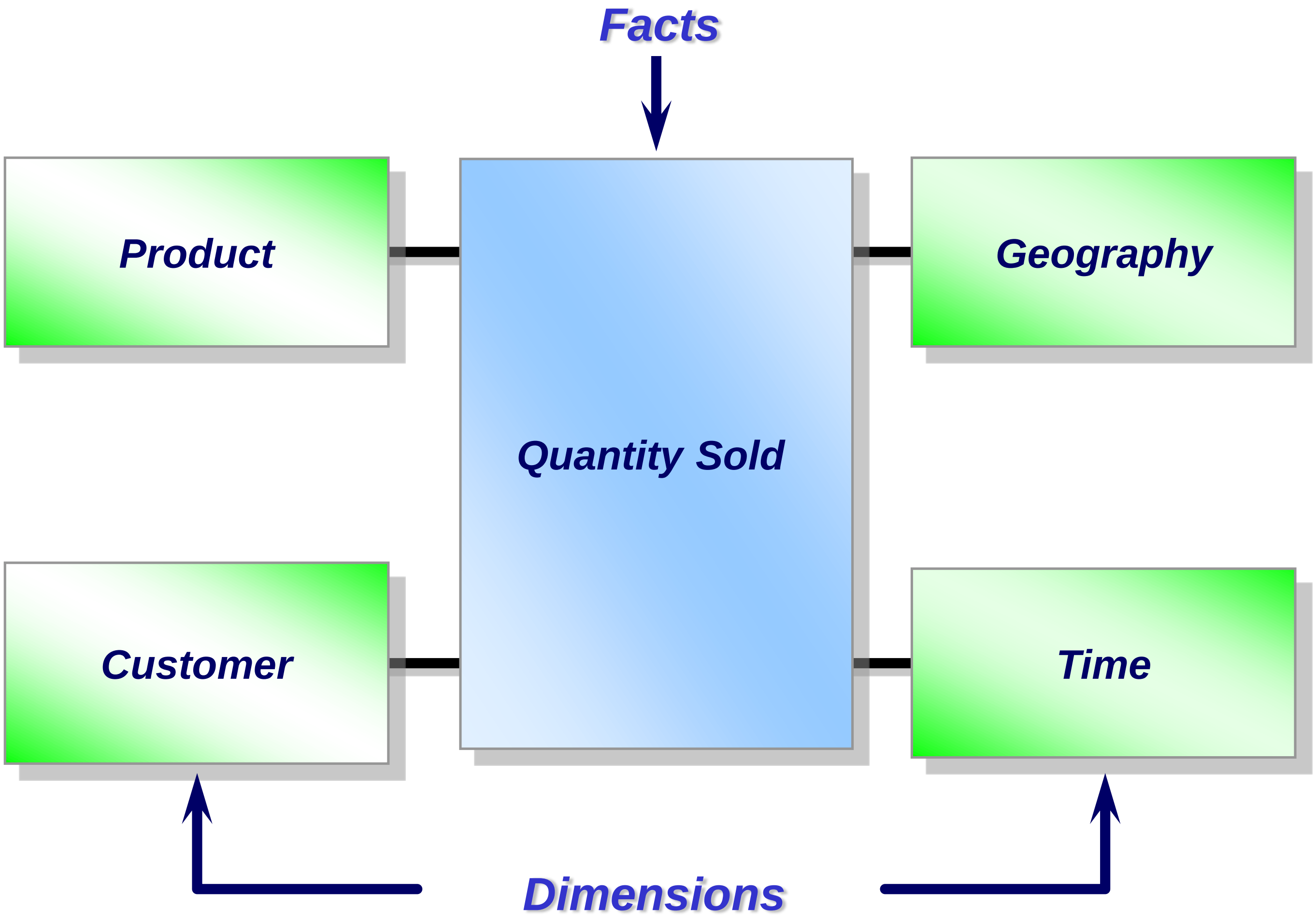
Dimensionality

- Multidimensionality is the most specific DSS data property
- Multidimensionality is the consequence of the fact that for the business decision making it is necessary to bond and to put into the relationship individual and aggregated data about different entities
- Dimensions give context to business measures
- Operational data contain information about dimensions, but this information is there implicit
- DSS data have explicit dimensions

Dimensionality

- Some examples:
  - The analysis of a product sales to a customer during last six month has three dimensions
    - These are: customer, product and time
  - The analysis of a product sales to some customers in the cities of a district during last six month has four dimensions
    - These are: customer, product, space, and time
- By the rule, the time is always one of the DSS data dimensions

Dimensionality



Technical Characteristics

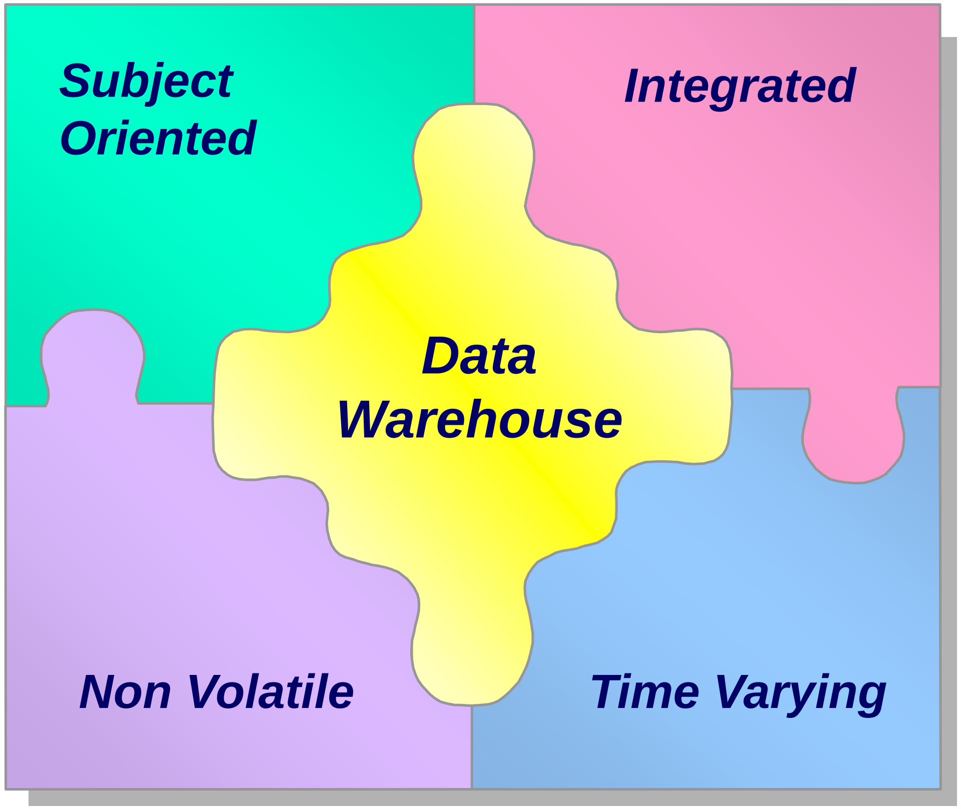
CHARACTERISTICS	OPERATIONAL	DSS
Schema	3NF or BCNF	1NF
Transaction Type	Mainly Update	Mainly Query
Frequency of Updates	High	Low (Periodic Refreshing)
Frequency of Queries	Small to Medium	High
Volume of Data Queried	Small	Large
Query Complexity	Small to Medium	Very large
Data Volume	MByte - GByte	GByte – TByte

Data Warehouse

- A database that has:
  - Historical data,
  - Aggregated data with varying granularity,
  - Explicitly designed dimensions,
  - Design that is optimized for efficient answering DSS queries

  is called a Data Warehouse
- The main characteristics of the Data Warehouse approach are:
  - Integration,
  - Orientation towards subjects
  - Dependency on time intervals, and
  - (relative) Non volatility

Properties of a DW System



Subject Oriented

Integrated

Data Warehouse

Non Volatile

Time Varying

Integration

- Integration pertains to the fact that a Data Warehouse represents a centralized database
- This database contains data of all organizational parts of a company in a standard format
- Very often, the operational database of a company is partitioned across the organizational units, and each individual database has different standards and data formats, even a different DBMS

Orientation to Subjects

- Orientation to subjects is a consequence of the fact that different company functions use the same Data Warehouse
- Possible subjects:
  - Sales,
  - Marketing,
  - Finance,
  - Transport,
  - Production
- Each subject contains data that are of an interest to the function considered
- Different subjects may have some data in common

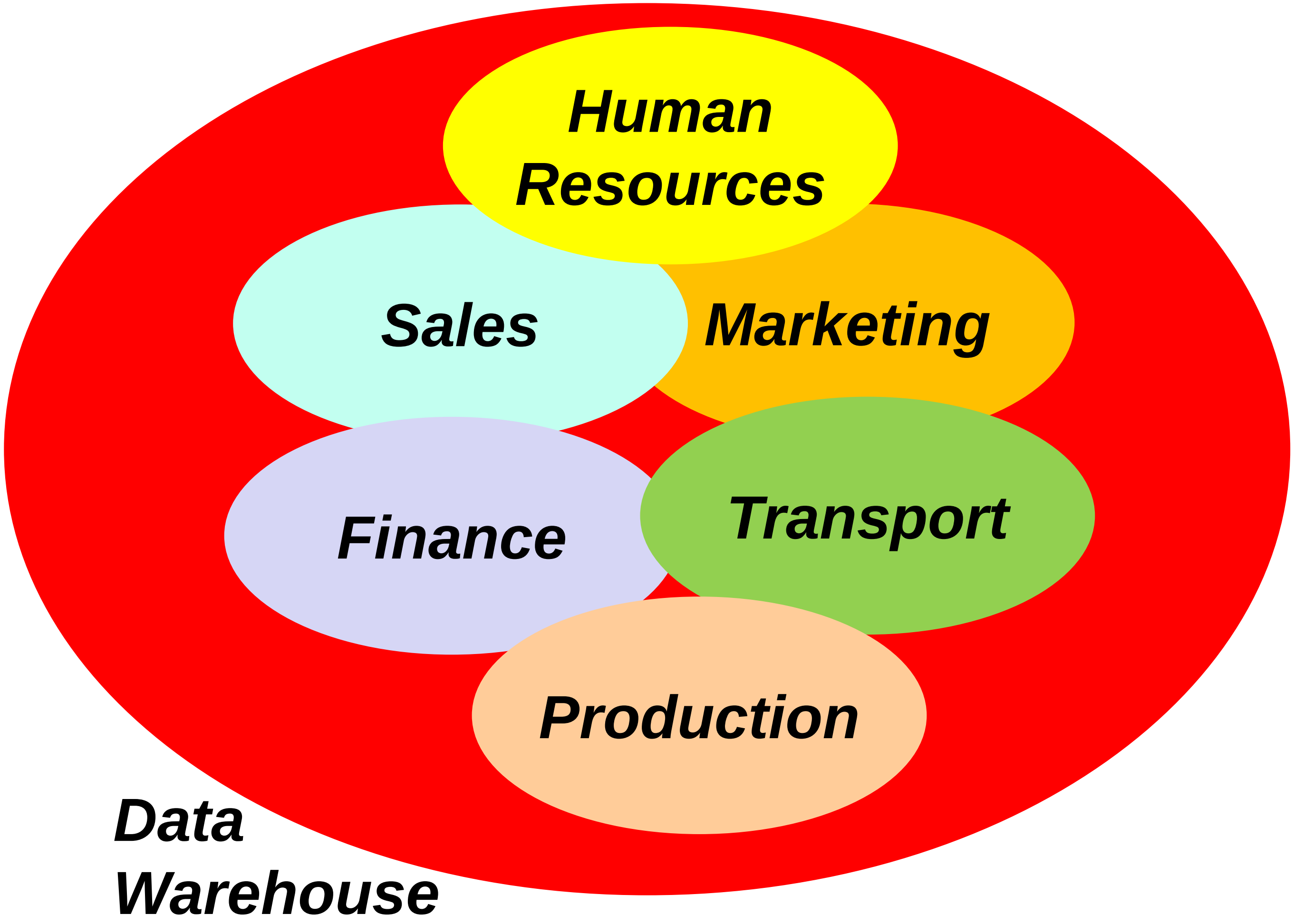
Subjects

- Sales subject contains data about:
  - products,
  - customers,
  - organizational units,
  - geography (districts, cities)
- But customers data also pertain to:
  - marketing,
  - finance, and
  - transport

Similarly, besides to sales, product data pertain to:

- marketing, and
- production

Subjects as DW Components



Human Resources

Sales

Marketing

Finance

Transport

Production

Data Warehouse

Time Dependency

- Dependency on time has two aspects:
  - First, DSS data contain time as a dimension
    - Aggregated data regarding business performance during previous time intervals
  - The second aspect – new data are periodically added to a Data Warehouse

Time Dependency

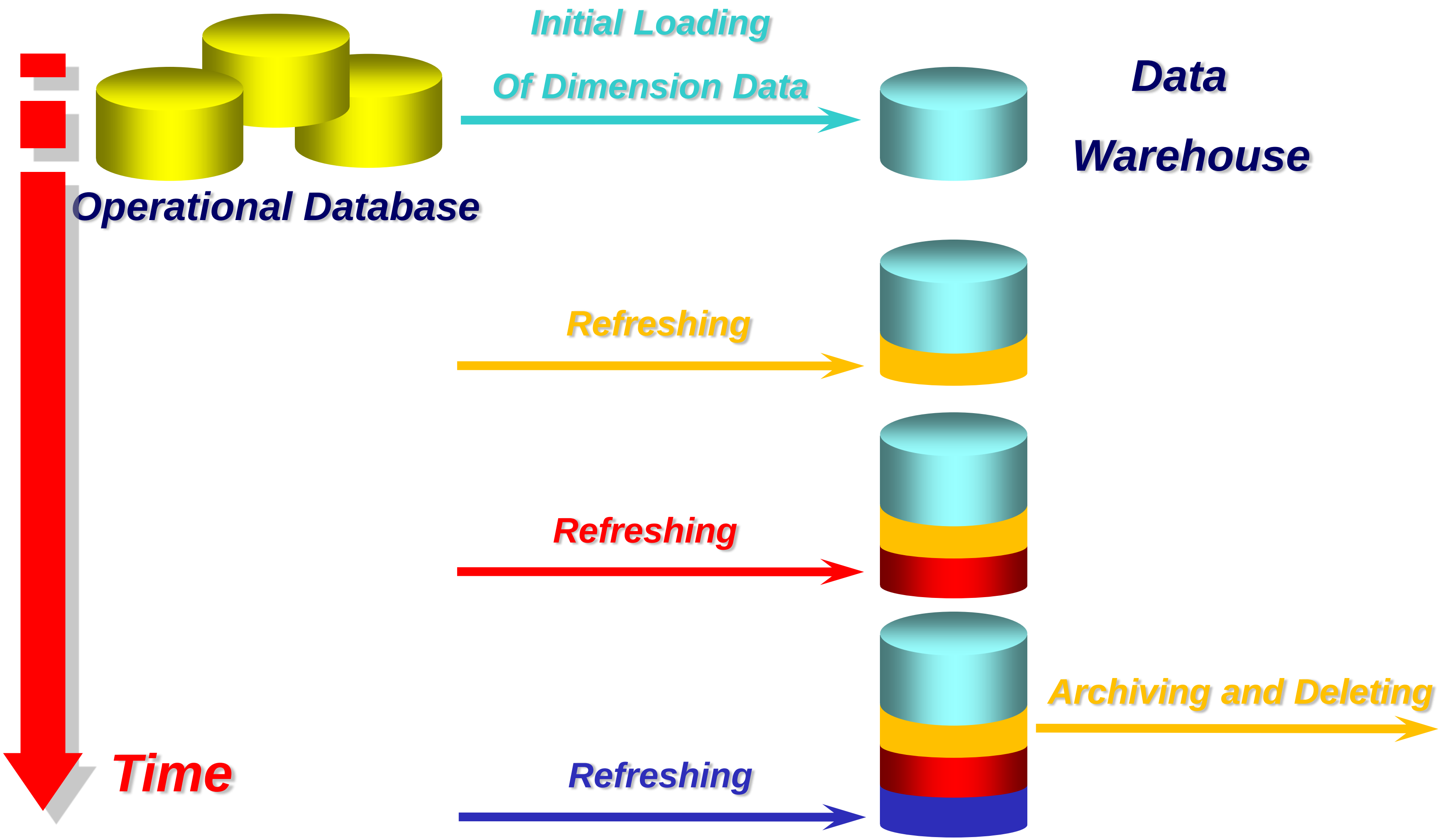
- The periodical addition of data to a Data Warehouse is called refreshing
- During refreshing, all time dependent aggregations are computed again
- Example:
  - If the sales data for the last time period are loaded into a Data Warehouse, aggregated sales data for:
    - Products,
    - Customers, and
    - Shops

    have to be recalculated

Non Volatility

- Non volatility pertains to the fact that a Data Warehouse data are (rarely) deleted and relatively rarely updated
- There are mainly only new data added into a Data Warehouse
- This way a Data Warehouse only grows during a given period of time
- So, a Data Warehouse DBMS has to manage a many TByte database
- In practice, a Data Warehouse is often “re-synchronized” to keep its volume manageable
  - Re-synchronization implies that a Data Warehouse stores data for a fixed number of time periods
  - After a new period is added, the oldest is deleted

Non Volatility



Summary

- Modern decision support systems are used in almost every field of human activities to enhance making decisions
- Structure of a DSS consists of (at least) the following components:
  - Data extraction and filtering
  - Data Warehouse
  - Query and analysis tool, and
  - Presentation tool
- Content of a Data Warehouse is extracted from operational databases and external sources
- Data Warehouse data significantly differ from operational data

Summary

Differences between DSS and operational data from the point of view of a data analyst

Characteristics	Operational Data	DSS Data
CURRENCY	Real Time	Historical
GRANULARITY	Fine	Course
AGRREGATION	Seldom	Frequent
DIMENSIONALITY	Implicit	Explicit
VOLUME	MB to GB	GB to TB