Analytical data processing

Module 3. Database operation

Department of Information Systems

Test questions

en:

1. Compare OLTP system and data warehouse
2. Describe OLAP cube operations
3. Compare dimension and fact tables

ru:

1. Сравните OLTP-систему и хранилище данных
2. Опишите операции с кубами OLAP
3. Сравните таблицы измерений и фактов

Content

1. Data Analysis Tools
2. Structure of decision support system
3. The concept of data warehouse
4. OLAP technology
5. Data Warehouse Schema

1. Data Analysis Tools

System classes

Systems focused on operational (transactional) data processing - **OLTP** (On-Line Transaction Processing)

Systems focused on analytical data processing – **DSS** (Decision Support Systems)

Functions

Simultaneous execution of a large number of short transactions from a large number of users.

Data data analysis, domain process modeling, forecasting, finding dependencies between data, “what if:” analysis

Characterized by

Support for a large number of users
The short response time to the request
Relatively short queries
Participating in requests for a small number of tables

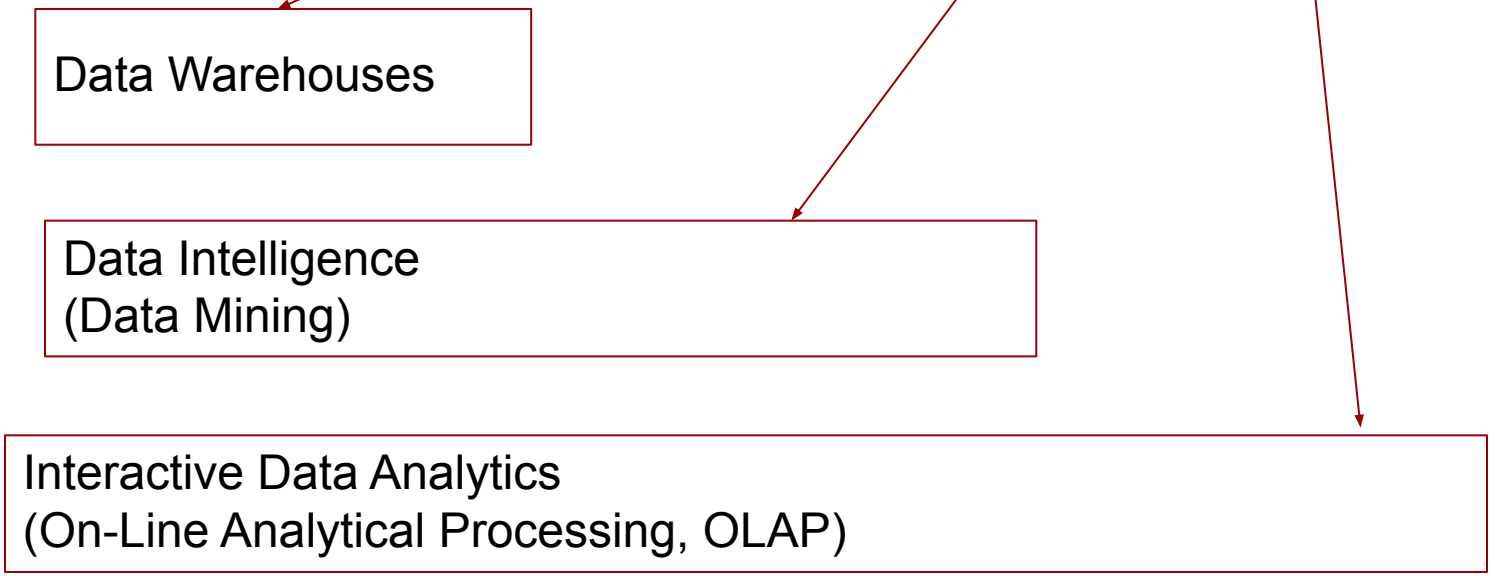
Using large amounts of data
Adding new data to the system is relatively rare, with large blocks;
Data added to the system is usually never deleted.
Before downloading, the data goes through various “cleaning” procedures;.
A small number of users (analysts).

Analytical systems

Static DSS

Dynamic DSS

DSS is a system that has the means of input, storage and analysis of data related to a specific subject area, with the aim of finding solutions.



Main concepts

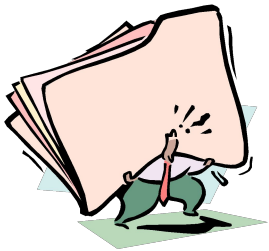
Concept **Data Warehouse** determines the process of collecting, weeding, pre-processing and accumulating data in order to:

1) long-term storage

2) providing the resulting information to users in a convenient way for statistical analysis and analytical reporting.

Concept **Mining** determines the tasks of finding functional and logical patterns in the accumulated information, building models and rules that explain the anomalies found and/or predict the development of certain processes.

Concept **OLAP** comprehensive interactive data processing using methods multidimensional data analysis to support decision-making processes. In theory, OLAP can be used directly to operational data or their exact copies (so as not to interfere with operational users).



By the degree of "intelligence" of data processing highlight three classes of analysis tasks:

1

Information and search. DSS searches for the data you need. A characteristic feature of this analysis is the execution of pre-defined queries;

2

Operational-analytical. DSS groups and summarizes data in any form the analytics needs. Unlike information and search analysis, it is not possible to predict the queries required by the analytics in advance;

3

Intelligent. DSS searches for functional and logical patterns in accumulated data, builds models and rules that explain the patterns found and/or (with a certain probability) predict the development of some processes.

Characteristics of the OLTP system

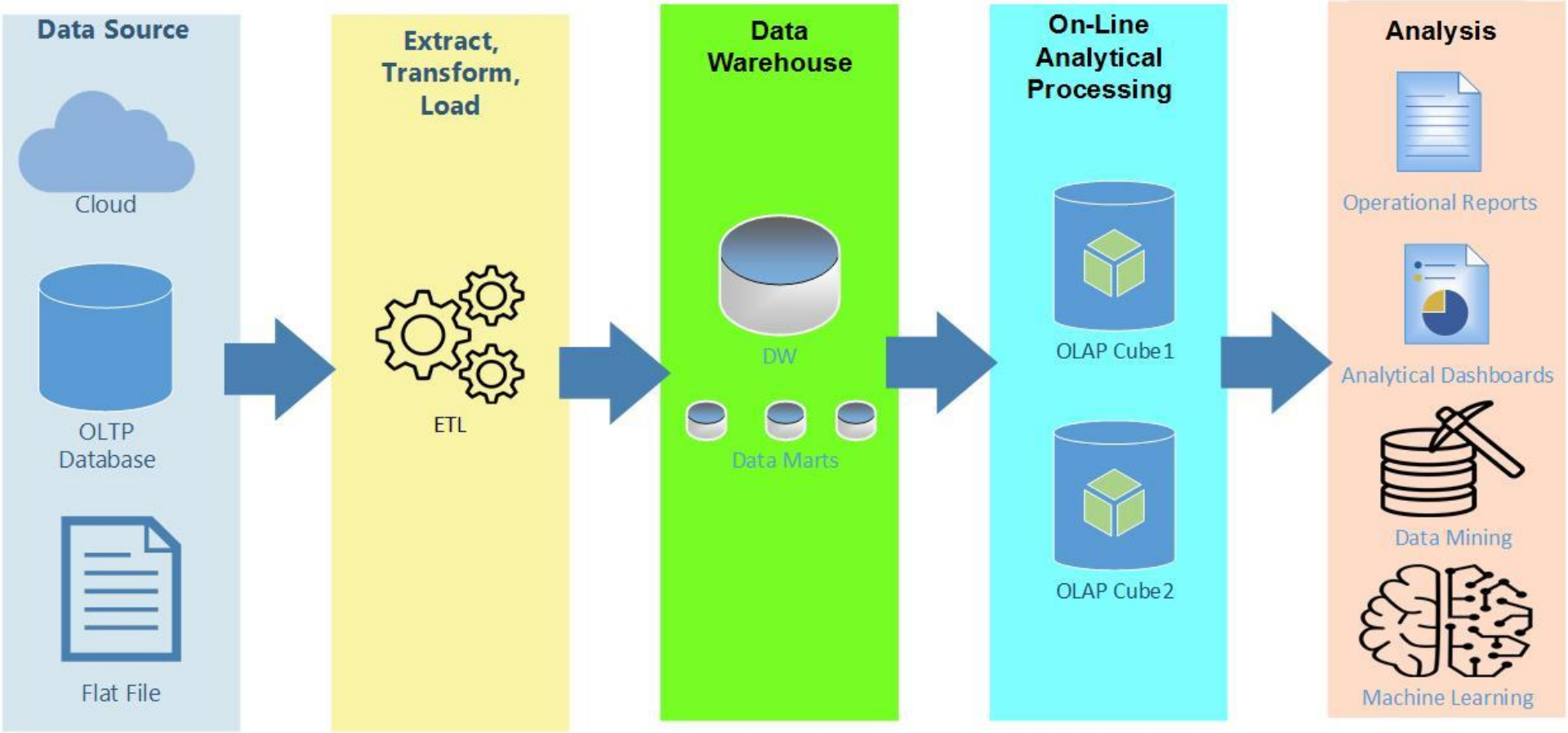
· A large amount of information
· Often different databases for different departments
· Normalized scheme, lack of duplication of information
· Intensive data change
· Transactional mode of operation
· Transactions affect a small amount of data
· Processing current data - snapshot
· Many clients
· Short response time - a few seconds

Characteristics of OLAP Systems

· A large amount of information
· Synchronized information from various databases using common classifiers
· Unnormalized database schema with duplicates
· Data changes rarely. Change through batch download
· Complex ad hoc queries are performed over a large amount of data using groupings and aggregate functions.
· Analysis of time dependencies
· A small number of users - analysts and managers
· Longer response time (but still acceptable) - a few minutes

2. Structure of decision support system

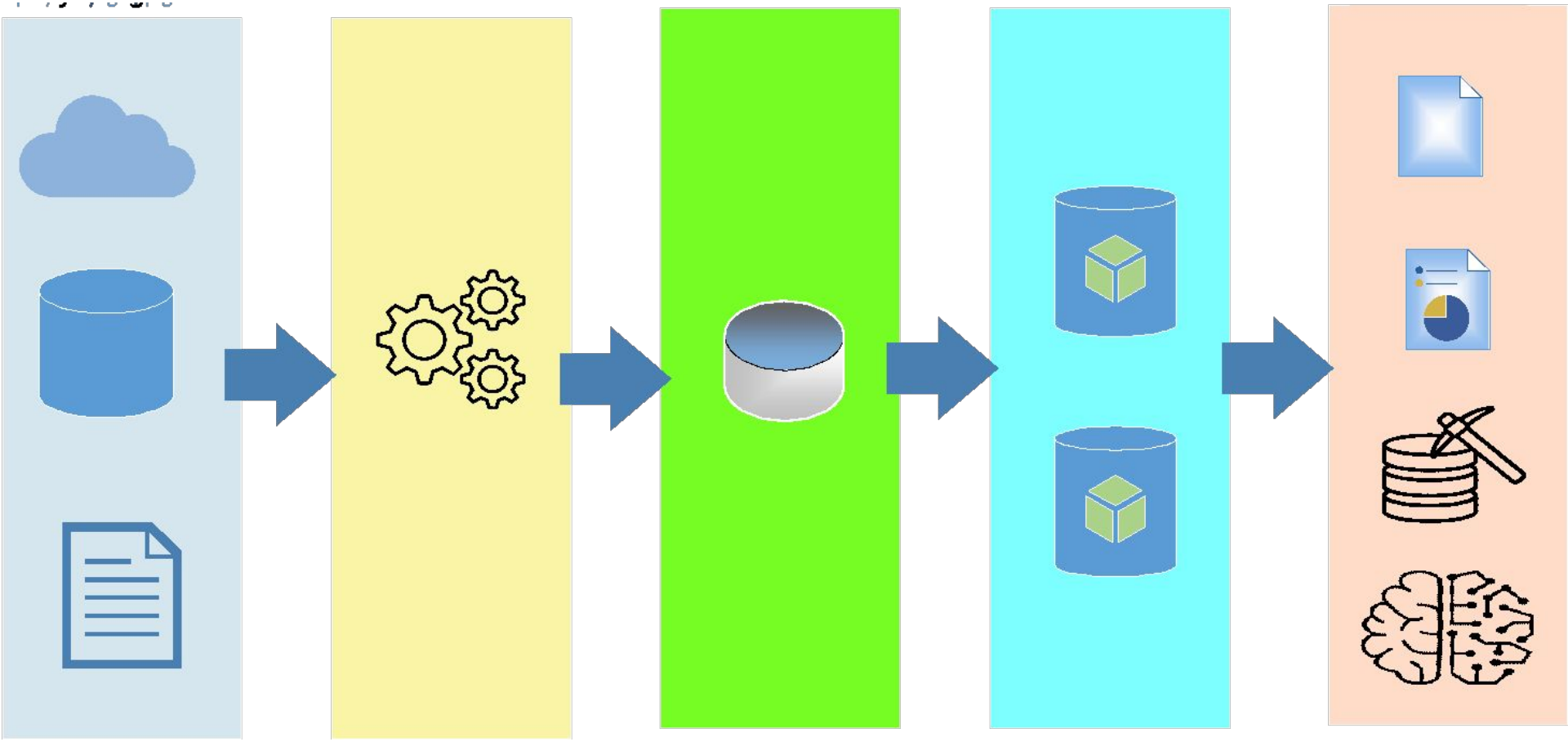
Generalized structure of decision support system based on data warehouse



Architecture options for DSS systems

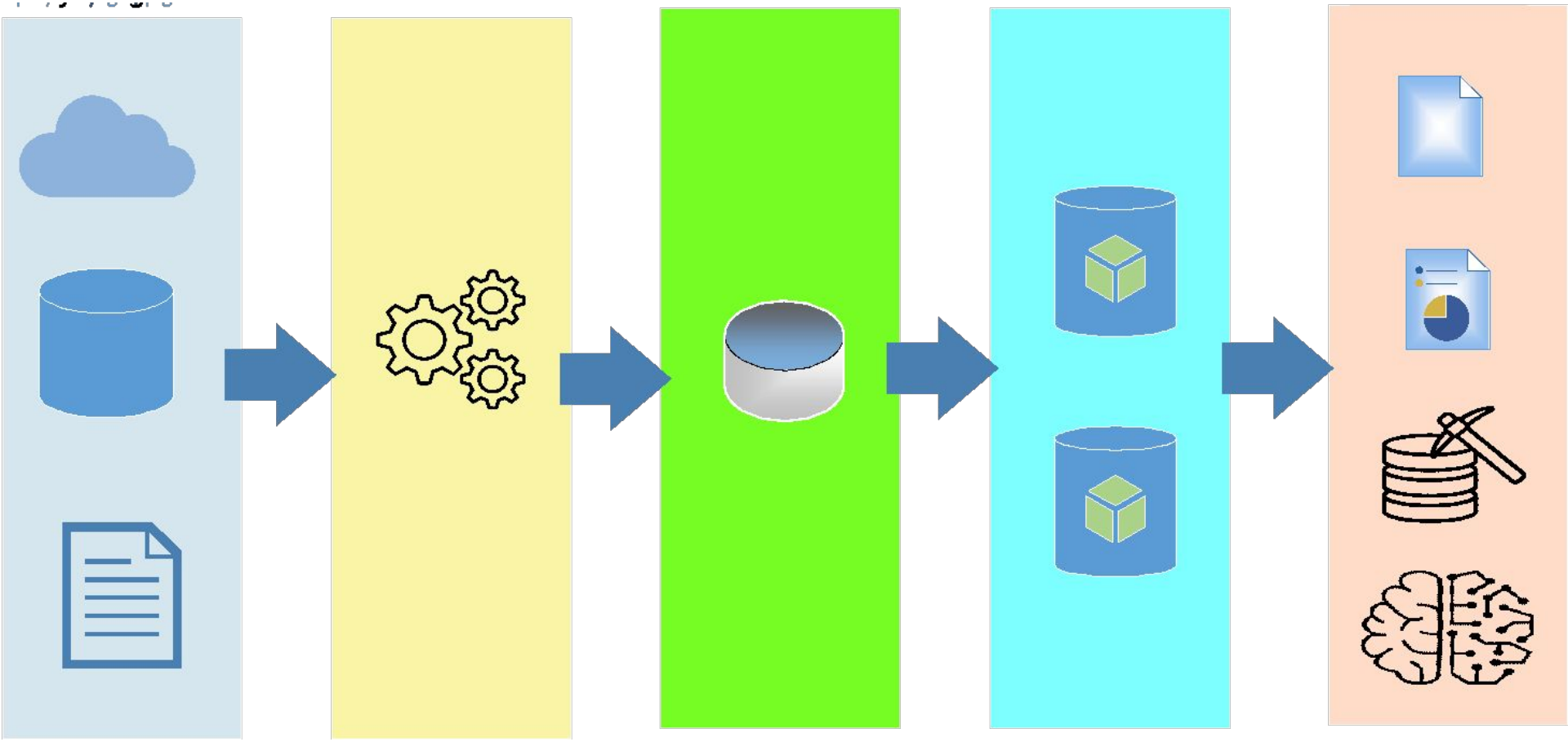
1. DSS systems with physical data warehouse

Data is transferred from various operational data sources to a single repository. The collected data is reduced to a single format, coordinated and summarized. Analytical queries are addressed to the data warehouse.



2. DSS systems with virtual data warehouse

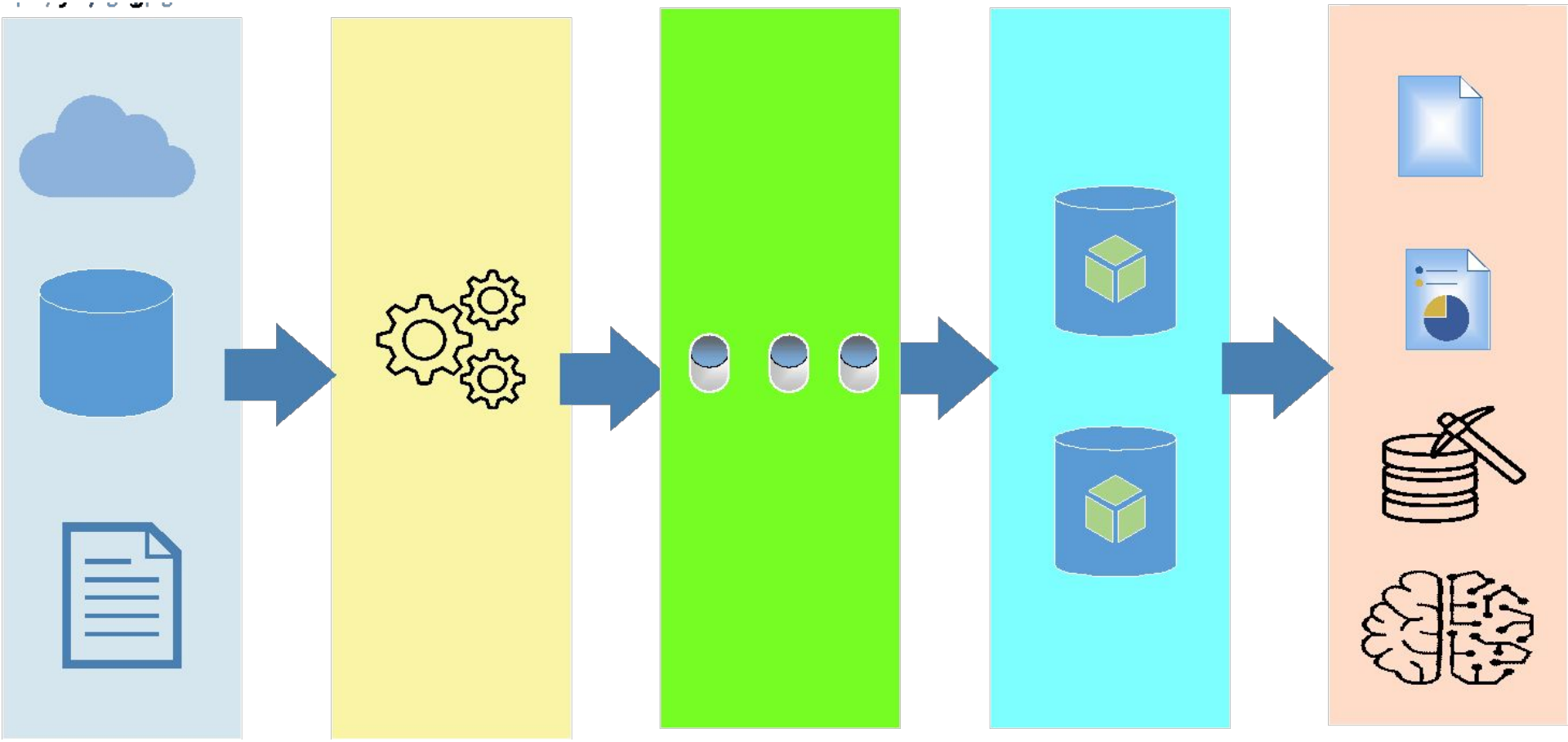
Data from the operational data sources is not copied to a single storage, but is extracted, converted and integrated when performing analytical queries directed directly to the operational data sources.



3. DSS systems with data marts

A data mart is a simplified version of a data warehouse that contains only thematically aggregated data.

The data mart is as close as possible to the end user and contains data thematically oriented towards him.



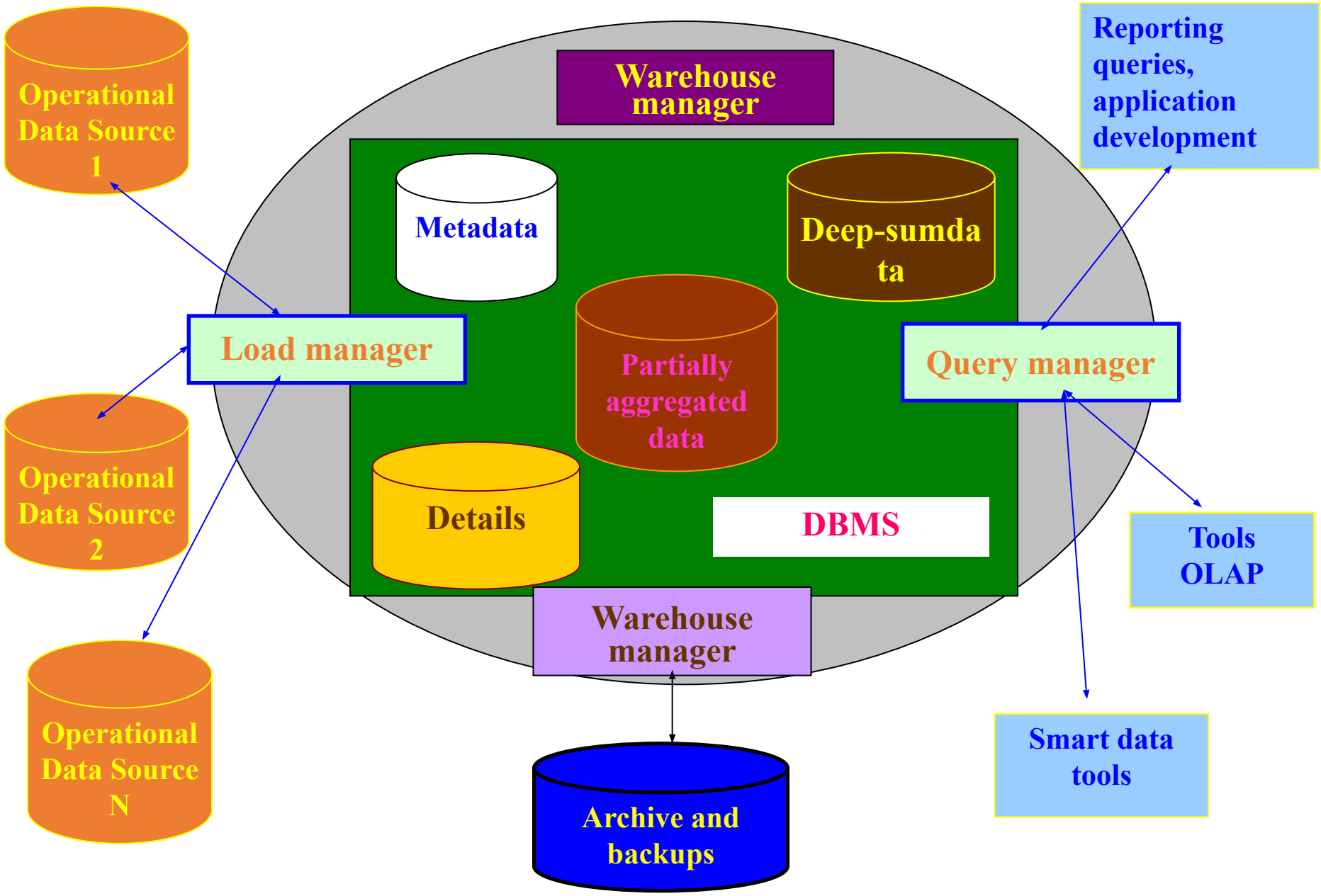
3. THE CONCEPT OF DATA WAREHOUSE

Data warehouse (Bill Inmon's definition) - is a subject-oriented, integrated, time-variant and non-volatile collection of **data** in support of management's decision making process.

Basic data warehouse requirements

- Focusing on a subject area
- Integration and internal consistency
- Maintaining a high rate of data from warehouse
- The ability to receive and compare data slices
- Non-volatile
- Time-variant
- The completeness and reliability of stored data
- Supporting a quality data refill process

Typical data warehouse architecture



Managers

The **load manager** performs operations related to the extraction and loading of data into the data warehouse.

The **warehouse manager** performs operations related to the management of information placed in the data warehouse:

- data consistency analysis;
- Creating Indexes and Views for Base Tables
- data denormalization (if necessary);
- generalization of data (if necessary);
- storage and archiving of backups.

The **query manager** performs operations related to managing user queries.

Detailed data. They correspond to elementary events recorded by OLTP systems (sales, experiments, etc.) and are divided into **Dimensions** (data needed to describe events: cities, events, people, etc.) and **Facts** (data reflecting the essence of the event: the number of goods sold, the results of experiments, etc.).

Aggregated data

Category data in data warehouse

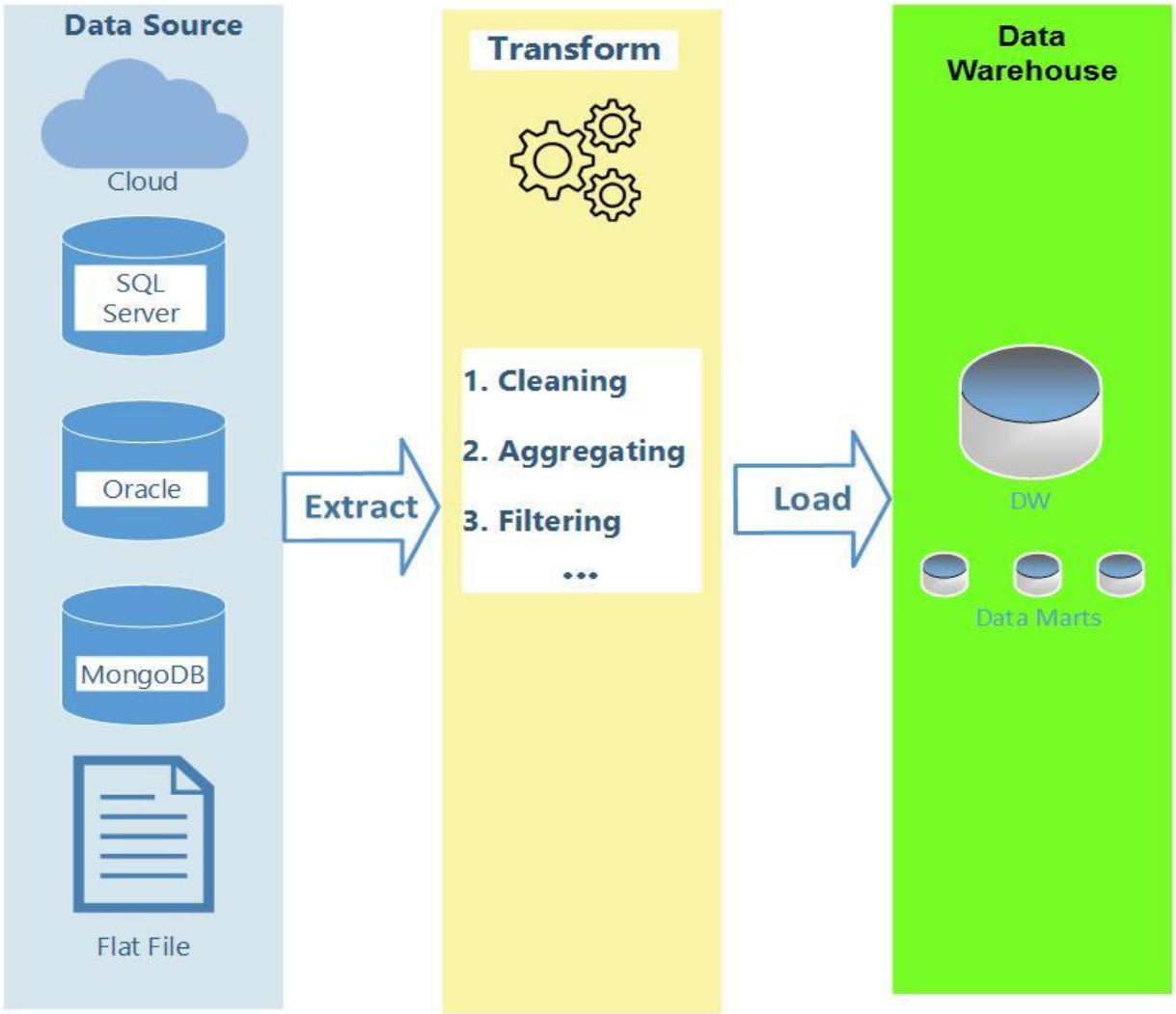
- *Additive* - numerical actual data that can be summarized across all dimensions;
- *Semi-addive* - numerical actual data that can only be summed up over certain measurements; (except for time: account balance, average

Metadata. Information about data stored in the data warehouse. Metadata should answer the following questions:

What Describe objects in the domain stored in the data store;
Who Describe the categories of users used data and their access rights;
Where Describe the location of servers, workstations, ID storage sites and the distribution of data between them;
When Describe the time it takes to perform different data operations.
Why Describe the reasons for the performance of certain operations over the data.

ETL

ETL (**E**xtract, **T**ransform and **L**oad) is defined as a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system.



ETL tasks

Extracting the data from source systems (Cloud, SQL Server, Oracle, MongoDB, Flat files …), data from different source systems is converted into one consolidated data warehouse format which is ready for transformation processing.

Transforming the data may involve the following tasks:

- applying **business rules** (so-called derivations, e.g., calculating new measures and dimensions),
- **cleaning** (e.g., mapping NULL to 0 or "Male" to "M" and "Female" to "F" etc.),
- **filtering** (e.g., selecting only certain columns to load),
- **splitting** a column into multiple columns and vice versa,
- **joining** together data from multiple sources (e.g., lookup, merge),
- **transposing** rows and columns,
- **applying** any kind of simple or complex data validation (e.g., if the first 3 columns in a row are empty then reject the row from processing)

Loading the data into a data warehouse or data repository other reporting applications.

4. OLAP technology

OLAP Definition

OLAP (Online analytical processing), is an approach to answer multi-dimensional analytical (MDA) queries swiftly in computing (Edward Codd).

He also formulated 12 OLAP principles, which were later redesigned into a so-called test **FASMI**:

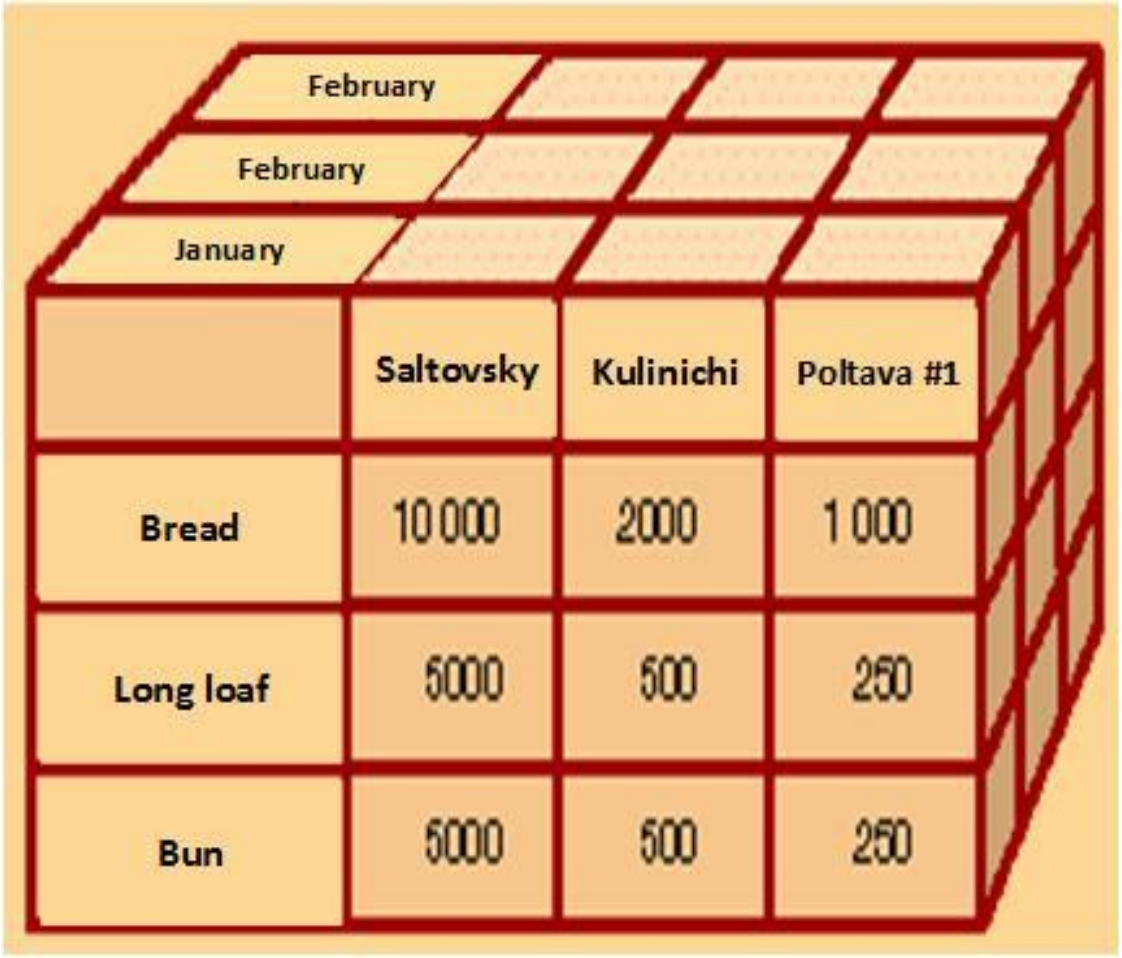
- **Fast**. Analysis should be performed equally quickly on all aspects of the information. An acceptable response time of 5 s or less.
- **Analysis**. It should be possible to carry out the main types of numerical and statistical analysis, predetermined by the application developer or arbitrarily determined by the user.
- **Shared**. Many users must have access to data, and access to sensitive information must be controlled.
- **Multidimensional**. Is the main, most essential characteristic of OLAP.
- **Information**. The application should be able to access any information you need, regardless of its volume and storage location.

OLAP = Multidimensional View = Cube

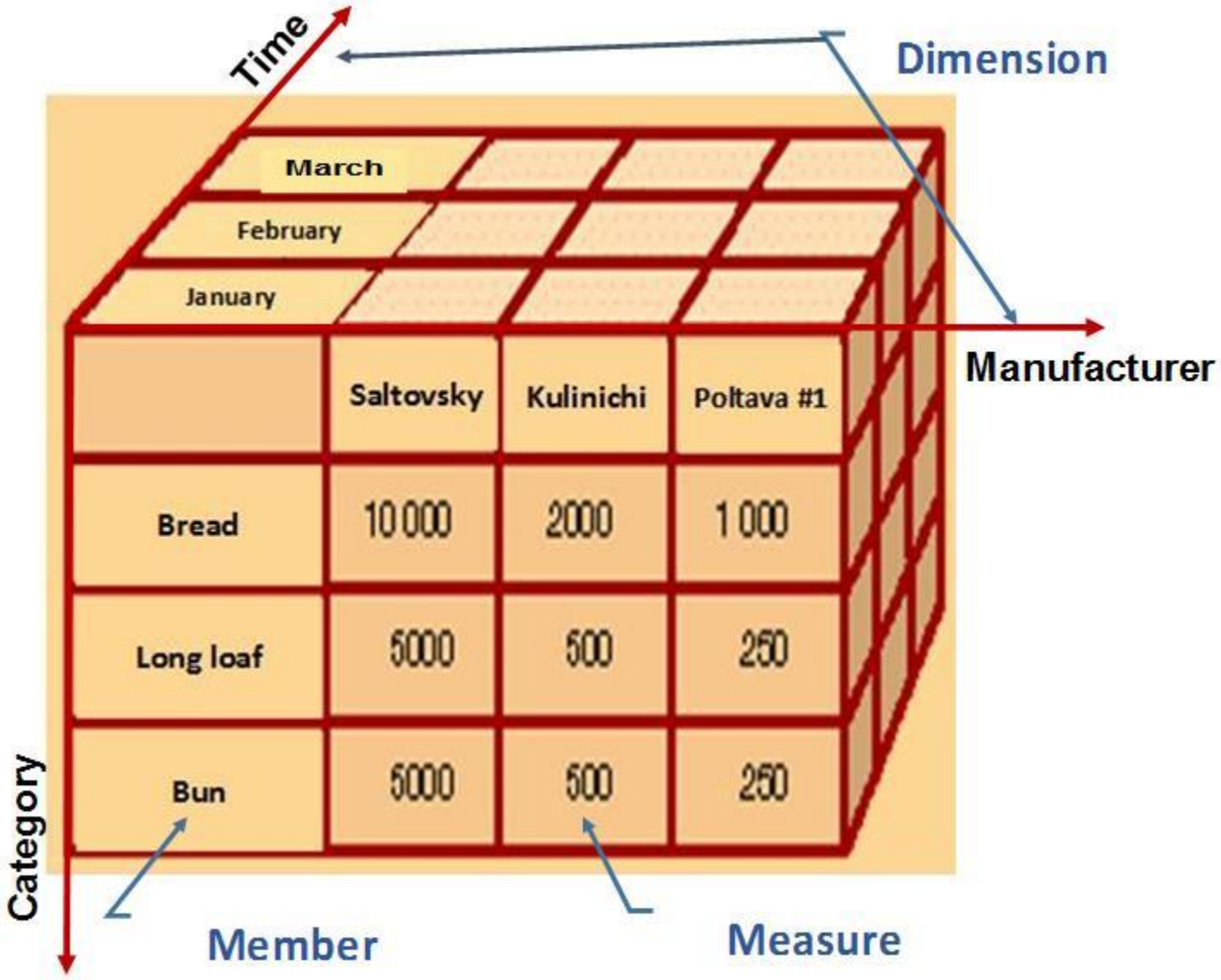
OLAP technology presents data for analysis in the form of multidimensional (and, therefore, non-relational) data sets called **multidimensional cubes** (hypercube, metacube, fact cube), whose axes contain parameters (dimensions), and the cells - aggregate data depending on them.

Moreover, a hypercube is a **conceptual logical model** of data organization, and not a physical implementation of their storage, since such data can also be stored in **relational tables** ("relational databases were, are and will be the most suitable technology for storing corporate data" - E. Codd).

A three-dimensional cube where **sales amounts** are used as **facts**, and **time**, **product category** and **manufacturer** are used as measurements defined at different levels of grouping: products are grouped by category, sales time data is by month, and manufacturers are not grouped.



Basic OLAP Cube Data Structures



Basic OLAP Cube Data Structures

Dimension is a metadata element that describes the main economic indicators of an enterprise (*product categories, manufacturers, cities, time periods,* …).

Member is a single data item within dimension (*Bread, Kulinichi, January,* …).

Measure is a numeric values that users want analyse (*How much bread was sold from Kulinichi in January? **2000** pcs*).

Hierarchy is a set of parent-child relationships, typically where a parent member summarizes its children. Parent elements can further be aggregated as the children of another parent.
***Time**: Date – Month – Quarter – Year;*
***Product**: Product – Category – Industry;*
***Location**: Office – City – Region – Country*.

Level is a position in a hierarchy (*Month* in *Time*, *Category* in *Product*).

OLAP cube operations

Slice is the act of picking a rectangular subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension.

Category = Bread

	Saltovsky	Kulinichi	Poltava #1
January	5000	500	250
February	500	800	2000
March	3000	1000	500

Dice is an "extension" of the slice operation as it allows users to extract a subcube by selecting values for several dimensions.

Roll-up is synonym for "consolidation" or "aggregation." The Roll-up operation can be performed in 2 ways:

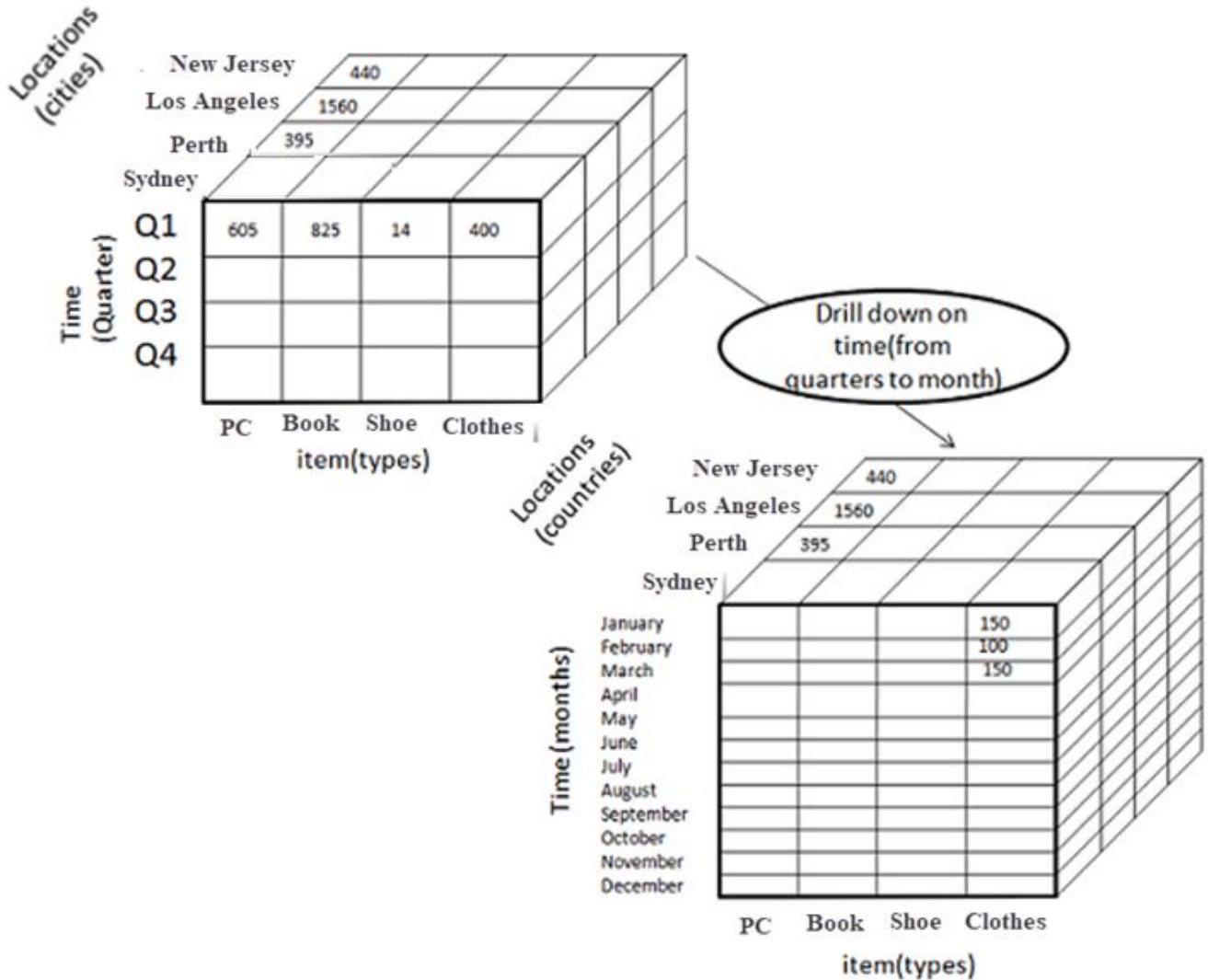
1. Reducing dimensions
2. Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level.

The sum of sales for all products
(Reducing dimensions**):**

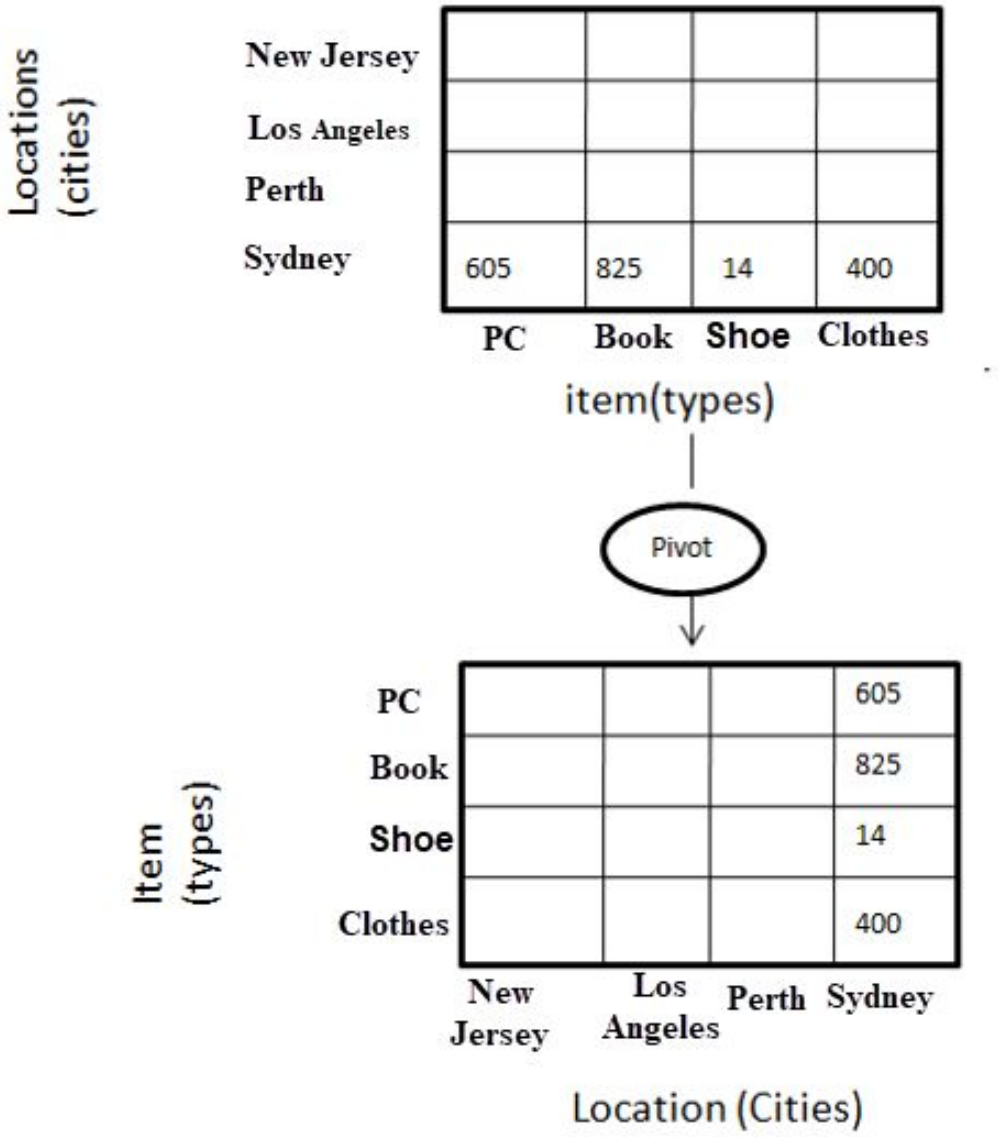
	Saltovsky	Kulinichi	Poltava #1
January	20 000	4000	2000
February	30 000	6000	3000
March	50 000	10 000	5000

In **drill-down** data is fragmented into smaller parts. It is the opposite of the rollup process. It can be done via

1. Moving down the concept hierarchy
2. Increasing a dimension



In **Pivot** (Rotate), you **rotate** the data axes to provide a substitute presentation of data.



Types of OLAP systems

ROLAP (Relational OLAP) is an extended RDBMS along with multidimensional data mapping to perform the standard relational operation.

MOLAP (Multidimensional OLAP) implements operation in multidimensional data.

In **HOLAP** (Hybrid OLAP) approach the **aggregated** totals are stored in a **multidimensional** database while the **detailed** data is stored in the **relational** database. This offers both data efficiency of the ROLAP model and the performance of the MOLAP model.

5. Data Warehouse Schema

Dimensional Model Concept

A **dimensional model** is a data structure technique optimized for Data warehousing tools. The concept of Dimensional Modelling was developed by Ralph Kimball and is consists of "**fact**" and "**dimension**" tables.

A **dimensional model** is designed to read, summarize, analyse numeric information like values, balances, counts, weights, etc. in a data warehouse. In contrast, **relation models** are optimized for addition, updating and deletion of data in a real-time Online Transaction System.

These dimensional and relational models have their unique way of data storage that has specific advantages:

For instance, in the **relational** mode, **normalization** and ER models reduce **redundancy** in data. On the contrary, **dimensional** model arranges data in such a way that it is **easier to retrieve information** and **generate reports**.

Elements of Dimensional Data Model

Fact. Facts are dimensions / metrics or facts of a business process. For a business's sales process, the measurement is the monthly total sales.

Dimension. A dimension provides the context surrounding a business process event. Simply put, they give who, what, where it is a fact. In the Sales business process, for the actual monthly sales volume, the dimensions will be:

- **Who** – Manufacturer
- **Where** – Location
- **What** – Product
- **When –** Time

Attributes. The Attributes are the various **characteristics** of the dimension.

Time dimension: Date, Month, Quarter, Year;

Product dimension: Product, Category, Industry.

Attributes are used to **search**, **filter**, or **classify** facts. Dimension tables contain attributes.

Tables in the data warehouse

Fact table. A fact table is a primary table in a dimensional model. It contains

- Measurements/facts
- Foreign key to dimension table

Dimension table. A dimension table contains dimensions of a fact.
They are joined to fact table via a foreign key.
Dimension tables are de-normalized tables.

The **dimension attributes** are the various **columns** in a dimension table.
Dimensions offers **descriptive characteristics** of the facts with the help of their attributes.

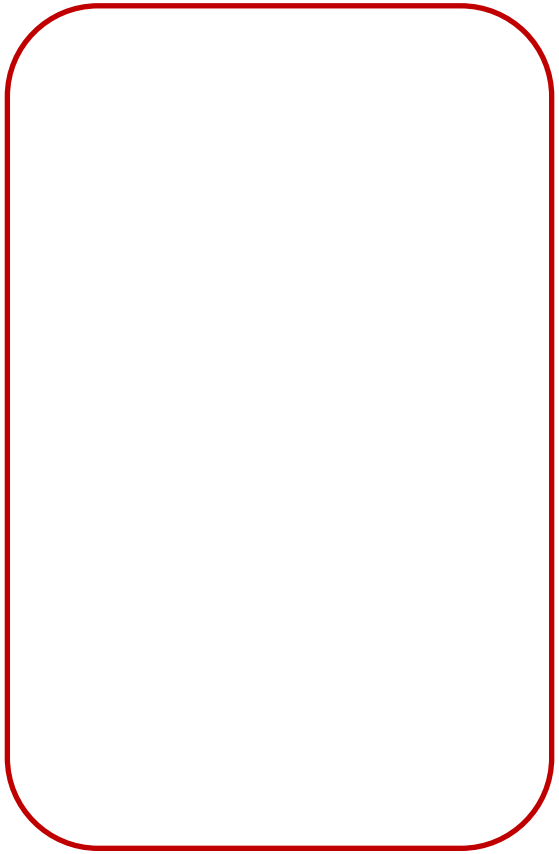
The dimension can also contain one or more **hierarchical relationships**

Schema types

Star Schema. It is called a star schema because diagram resembles a **star**, with **points** radiating from a center. The **center** of the star consists of the **fact** table, and the **points** of the star is **dimension** tables.
The **fact** tables in a star schema which is **third normal form** whereas **dimensional** tables are **de-normalized**.

Snowflake Schema. The snowflake schema is an extension of the star schema. In a snowflake schema, each dimension are **normalized** and connected to more dimension tables.

Galaxy Schema. A Galaxy Schema contains **two or more fact tables** that shares dimension tables. It is also called Fact Constellation Schema. The schema is viewed as a collection of stars hence the name Galaxy Schema.



Difference between database system and data warehouse

Parameter	Database	Data Warehouse
Purpose	Is designed to record	Is designed to analyze
Processing Method	The database uses the Online Transactional Processing (OLTP)	Data warehouse uses Online Analytical Processing (OLAP).
Usage	The database helps to perform fundamental operations for your business	Data warehouse allows you to analyze your business.
Tables and Joins	Tables and joins of a database are complex as they are normalized.	Table and joins are simple in a data warehouse because they are denormalized.
Orientation	Is an application-oriented collection of data	It is a subject-oriented collection of data
Storage limit	Generally limited to a single application	Stores data from any number of applications
Availability	Data is available real-time	Data is refreshed from source systems as and when needed

Difference between OLTP system and data warehouse

Parameter	Database	Data Warehouse
Usage	ER modeling techniques are used for designing.	Data modeling techniques are used for designing.
Technique	Capture data	Analyze data
Data Type	Data stored in the Database is up to date.	Current and Historical Data is stored in Data Warehouse. May not be up to date.
Storage of data	Flat Relational Approach method is used for data storage.	Data Ware House uses dimensional and normalized approach for the data structure. Example: Star and Snowflake schema.
Query Type	Simple transaction queries are used.	Complex queries are used for analysis purpose.
Data Summary	Detailed Data is stored in a database.	It stores highly summarized data.

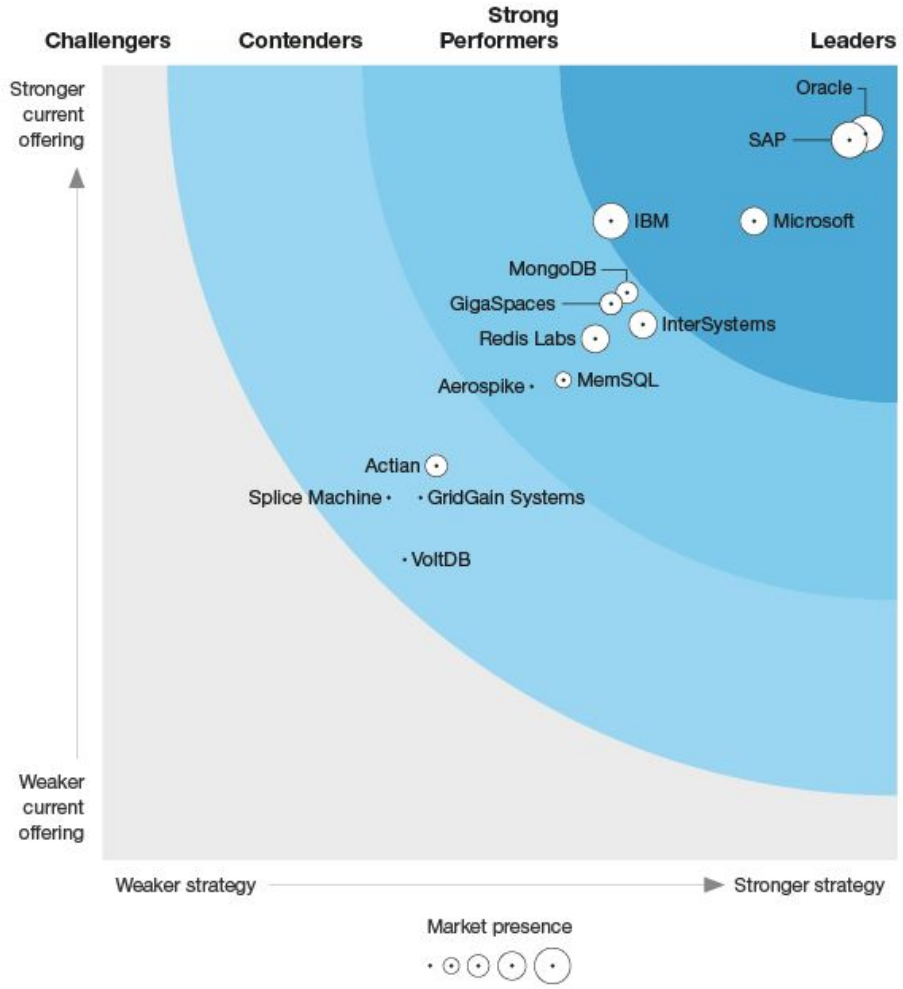
Figure 1. Magic Quadrant for Data Management Solutions for Analytics

Source: Gartner (January 2019)

THE FORRESTER WAVE™

Translytical Data Platforms

Q4 2019



144975

Test questions

en:

1. Compare OLTP system and data warehouse
2. Describe OLAP cube operations
3. Compare dimension and fact tables

ru:

1. Сравните OLTP-систему и хранилище данных
2. Опишите операции с кубами OLAP
3. Сравните таблицы измерений и фактов