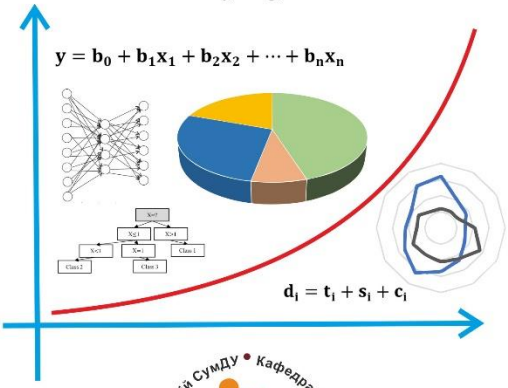
Sumy State University

Department of Psychology, Political Science and Socio-Cultural Technologies

Data analytics

Study guide



Кафедра психології, політології та соціокультурних технологій Сумду

2023

Ministry of Education and Science of Ukraine
Sumy State University

Kotenko O. O.

Data analytics

Study guide

Recommended by the Academic Council of Sumy State University

Sumy
Sumy State University
2023

UDC 001.8:[32+339.5](075.8)
K 77

Reviewers:

Yu. O. Ostapets – Doctor of Political Science, Professor, Dean of the Faculty of Social Sciences of Uzhhorod University;
O. V. Kupenko – Doctor of Education, Associate Professor of the Department of Psychology, Political Science and Sociocultural Technologies of Sumy State University

Recommended for publication
by the Academic Council of Sumy State University
as a study guide
(minutes № 15 of 29.06.2023)

Kotenko O. O.
K 77 Data analytics : study guide / O. O. Kotenko. – Sumy : Sumy State University, 2023. – 155 p.

This study guide is devoted to substantiating the nature, role and importance of data, information, analytical work, explanation of its basic principles within modern information environment, as well as consideration of the main approaches and basic tools while performing the analytical tasks by specialists in the sphere of political analytics as well as of social work.

The study guide can be used to study students in the field of knowledge: 05 – Social and behavioral sciences; 07 – Management and Administration; 23 – Social work; 28 – Public administration and administration; 29 – International relations, of the educational degree "bachelor" within all educational forms (full-time, part-time and distance learning). In addition, a study guide can be useful for students of other specialities and practitioners, whose professional activity is related to work with data.

UDC 001.8:[32+339.5](075.8)

CONTENTS

Introduction

Political, financial and economic activities of modern life are impossible without clear planning, analysis of current activities, assessment of the market situation, forecasting. Today it is extremely hard to imagine an enterprise in which there is permanent work on the accumulation, systematization, processing, storage of various data, which is necessary for the organization of the entrepreneurial activity, as well as ensuring its financial success. The problem of building effective work on data processing is complicated while entering international markets, when the quantitative and qualitative assessment of numerous factors influencing the company's activities is a key aspect of its success or failure, if such work is carried out unprofessionally.

However, bearing in mind the Political technologies and analytics, the authors tried to focus not only on aspects of business analytics, useful only for doing business, but also on various types of relations (economic, social, cultural, scientific) that arise between countries as well as the residents ana non-residents.

At the same time, during preparing the material for this study guide, we also focused on certain technical aspects and software products that are currently used to perform analytical tasks. In our opinion, this will not only expand the potential of future professionals in the aspects of political analytics, but also increase the level of their computer literacy and the culture of using computers, information, databases.

Among other advantages, the “Data analytics” study guide has the next one – it is adapted to the training of those students who are not native English speakers.

This work is a systematization of the knowledge within the “Data analytics” and “Politic analytics” subject fields, is adaptation to the educational process for students, studied at

English-speaking program of specialty 052 – “Political technologies and analytics”, 231 – “Social work”, 292 – “International economic relations” as well as a brief guide to understand the fundamental aspects of analytics.

The author also advises students to contact original sources, links to which will be given below.

Let this study guide will become a reliable guide for students in the fascinating, extremely important and relevant world of analytics.

The main terms

Analytics represents tools and programs to search, analyze, model and deliver information, necessary for decision making.

Data analysis is research related to the computation of a multidimensional data system that has many parameters

Model is an object or description of a situation, a system to substitute (under certain conditions, assumptions, hypotheses) one system (the original system) by another system for the better study of the original or reproduction of its any properties.

Model building is a universal method to obtain, describe and use knowledge. It is used in any professional activity.

Expert is a specialist in the subject area, a professional who, in the course of his or her studies and practical activity, has learned how effectively to solve problems related to a particular subject area.

The analyst is an expert in the field of analysis and modelling.

Hypothesis in the data analysis is often the assumption about the influence of any factor or group of factors on the result.

Replication of knowledge includes a set of methodological and instrumental means for model building, it allows final consumers to use the modelling results in decision-making without the need to understand the methods thanks to which these results are obtained.

Data is information that describes a system, phenomenon, process or object, presented in a certain form and intended for further use.

Unstructured data consists of any form, including text and graphics, multimedia (video, speech, audio). This form of data representation is widely used, for example, on the Internet, and the data is presented to the user as a link by the search engines.

Structured data reflects individual facts of the subject area. Structured data is considered to be well-structured and organized in order to allow certain actions to be applied to them (for example, visual or machine analysis). This is the main form of information representation **in databases.**

Weakly structured data is data for which certain rules and formats are defined but in the most general form.

Access to data is the possibility to allocate an element of data (or a plurality of elements) among other elements on any basis in order to perform certain actions on them.

Continuous data are data which can take any value in a certain interval. The analyst can perform arithmetic operations of addition, subtraction, multiplication, division with continuous data, and they will make sense.

Discrete data – the value of an indication, the total number of which is finite or infinite, but they can be calculated with the help of natural numbers from one to infinity. No arithmetic operations can be made with discrete data or such operations do not make sense.

Business data is rarely accumulated especially to solve tasks of the analysis

Data Mining – the detection in raw data of previously unknown, non-trivial, practically useful and accessible for knowledge interpretation and necessary for decision-making in various spheres of human activity.

Classification means to set up the dependence of a discrete output variable on input variables.

Regression means to set up the dependence of the continuous output variable on the input variables.

Clustering means to group objects (observations, events) based on data describing the properties of objects. Objects within the cluster must **be similar to each other and different from others** that have entered other clusters.

Association means to identify patterns between related events. An example of such regularity is the rule that indicates that event Y follows the event X. **Such rules are called associative**.

Consolidation - a set of methods and procedures aimed at extracting data from different sources, providing the necessary level of their informativeness and quality, converting into a single format, in which they can be loaded into the storage and the other or analytical system.

OLTP-systems – On-Line Transaction Processing of transaction in real time.

DSS-system – decision support information systems. System oriented to the analytical processing of data in order to obtain the knowledge necessary for the development **of management decisions.**

Detailed data – data that comes directly from data sources and corresponds to elementary events registered by OLTP systems.

Aggregated data – is subjected to numerical data (facts), they are calculated based on detailed data. detailed data

Semantic layer – a mechanism that allows the analyst to operate with data through the business terms of the subject area.

Metadata – is necessary to describe the meaning and properties of information in order to understand, use and manage it better.

Data storage – a type of storage system focused on supporting the data analysis process, ensuring integrity, consistency and history, as well as high-speed execution of analytical queries.

Measurements are categorical attributes, names and properties of objects involved in a business process. Measurement values are the names of goods, the supplier and buyers companies names, the people names, cities names, etc.

Facts are data that quantitatively describe a business process, continuous in nature, that is, they can take on an infinite number of values.

ETL-process (Extract, **Transform**, L**oad)** is a set of methods which implement the process to transfer initial data from various sources to an analytic application or data storage that supports it.

Data transformation is a set of methods and algorithms aimed at optimizing the presentation and format of data in terms of tasks and analysis goals.

Data visualization means presentation of data in the form that provides the most efficient user's experience.

Data clearing is a correction of various errors in order to increase the predictive adequacy of models.

Data transformation is a set of methods and algorithms aimed at optimizing the presentation and format of data in terms of tasks and analysis goals.

Quantization – splitting the range for possible values of a numeric attribute into a specified intervals quantity and assigning the intervals with numbers or other labels to the values that fall into them.

Sorting – changing the order of the records at the original data sample in accordance with the algorithm defined by the user.

Merge – combining two tables by fields of the same name or to supplement one table with records from another that are missing in the complemented table.

Normalization – converting a range of changes for numeric feature to another range, more convenient to apply to the data from various analytical algorithms, as well as to reconcile the changing ranges of various signs.

Graphics are lines that represent the relationship between several variables in a certain coordinate system.

Histogram shows the distribution of the dataset inside the sample (for example, the number of bank borrowers by their occupations) in the columns form.

Statistics allows to put forward hypotheses about the behavior of data and their inherent laws, to control the processing results data at different stages of the analytical process.

Scattering diagram is a diagram contains a number of points placed in the Cartesian coordinate system, represent values for two variables. Assigning each variable axis can determine whether there is an interconnection or correlation between these two variables.

Contradictory information is information that does not comply with laws, regulations or reality. First, it is decided what data are necessary to consider as controversial information.

Data pre-processing is a set of methods and algorithms used in an analytical application to prepare data for solving a specific task and bring it into compliance with the requirements determined by the specific nature of the task and methods to solve it.

Data purity is the absence of errors during their input, structural violations, incorrect formats, and other reasons that hinder the data analysis in general.

Data quality closely relates to the specific goals and objectives of the analysis used by models, methods and algorithms.

Clustering – grouping objects based on the proximity of their properties; Each cluster consists of similar objects, and objects in different clusters differ significantly.

Decision tree is a tree-like hierarchical model, where each node checks a certain attribute using a rule.

Artificial neural networks are models that imitate the functioning of the brain during its operation.

Time series is a sequence of observations about parameters changing at a time, which characterize some object or process.

The trend is a time series component that describes the main tendency in the time series and reflects the impact on it of long-term factors that cause smooth and long-term changes in the series.

Seasonal component of time series is a time series component that describes regular changes in its values within a certain period and is a sequence of almost repeated cycles.

Cyclic component of time series is the rise or fall intervals of varying lengths, as well as the different amplitudes for the values contained therein.

1. THE ESSENCE AND ROLE OF ANALYTICS IN INTERNATIONAL POLITICAL AND ECONOMIC RELATIONS

It is known that modern society has entered a new stage in its development, the so-called **informational or post-industrial stage.** It is usually understood that **knowledge and information** are the main motives and productive social force that define both the spiritual and material state of a person, company and state as a whole.

Today the life quality of a person and society, their informational and public security directly depend on the ability to produce, search, analyze, classify, generalize, recognize, process, present information and make decisions. All the above processes belong to **spiritual or ideal**. As for them, there is a direct relationship between the level of a person's spiritual development, his or her abilities to creativity and the amount of information that he or she can absorb and transform. Hence, one can observe the strategic orientation of human evolution as a whole. It can be characterized as a **tendency of transition from material to the spiritual society**

The second trend quite clearly appears and consists in the fact that the transition to the information society results in its structural and material transformation: the information barriers between states, corporations, languages disappear. It means that **globalization** is a natural consequence of the information society development.

It puts forward **new requirements** for the person, to the level of his or her training, to his or her information culture. 10-15 years ago, the information culture of the individual (even in western countries), society, corporation was simply understood as **computer literacy** (possession of computer skills). Today it is the possession of methods and technologies in working with data and knowledge, with databases and information bases.

Ability to work with information becomes the **main condition for employment** while determining the future growth of personality and company (Knafli C.-N., 2015).

Modern society is characterized by one more **important detail:**

- it has come up to the edge when huge massifs of information (in corporations, government departments, social welfare services, etc.) are concentrated in the form of **databases;**
- **the capacity of computer technology**, systems and communication and data transmission **have significantly increased;**
- the Internet has become the only common repository of data and knowledge for humanity;
- a clear need for new opportunities for interaction with data and the means of their processing has been formed;
- the social order has been clearly formed for sufficiently complete and accessible native editions which deal with the information and analytical processing of large data amounts and for the corresponding software tools.

At the same time, in practice, there are no **fundamental materials** regarding such a popular direction of scientific and practical thought. All accumulated knowledge is formed under the influence of **other sciences** (statistics, econometrics, economic and financial analysis, etc.), personal experience of practising analysts and numerous training, which with different levels of success lay out some aspects of analytical activity at the enterprises.

The situation to fill the theoretical material about business analyst in **international economic relations** is even more acute. However, business analyst and business analyst in international economic relations may differ only in local customs and traditions of its implementation, local laws and databases. However, modern business analytics tools are gradually harmonized and brought to some standards.

Only a few years ago, the market for corporate analytics systems in Ukraine was so small that problems within business analytics were discussed in **a narrow circle of professionals** and they were greatly distrusted before the launching process of analytical projects.

However, since companies have collected and stored significant volumes of information, the **situation changed radically**: exponential growth of the market for business analytics (intelligence) systems was started in the country. Reduction of the complexity and cost to implement such systems and examples of successful solutions in the corporate analytical systems contributed to it (Mize Ed., 2017).

Banks, as well as insurance, telecommunication and trading companies, were the first, who have assessed the possibilities of business analytics. They learned to find useful knowledge and to gain benefit from the vast amount of information **collected and stored in corporate databases**. Today, it is obvious that the business cannot exist without analyzing data in a competitive environment.

The term **"business analytics"** does not have a single definition, because it includes a wide range of technologies. In the **broadest sense**, business analytics means "instruments and programs use to search, analyze, research and deliver necessary for decision making information". Business analytics is a **multi-disciplinary area**, located at the junction of:

- information technology;
- databases;
- algorithms of information intellectual processing;
- mathematical statistics;
- methods of visualization.

The decisions are made by managers. T**he tasks of business analysts** are to do everything to make these decisions optimal and timely.

The extremely difficult work of the business analyst is hidden by the **ready-made reports**, which can contain not only numbers in different planes, but also rules, dependencies, trends, predictions (i.e. knowledge).

Analysis methodology.

Data analysis is a broad concept. Dozens of its definitions exist today. In the most general sense, **data analysis is** research related to the computation of a multidimensional data system that has many parameters.

In the data analysis process, the researcher produces actions with the goal to form certain ideas about the nature of the **phenomenon described** by these data. As a rule, various **mathematical methods** are used for data analysis.

Data analysis cannot be considered **only** as information processing after its collection. **Data analysis** is primarily a mean to test hypotheses and to solve the researcher's tasks (Kovtun N., 2015).

The contradiction between the **limited human cognitive abilities** and the Universe infinity forces us to use **models and modelling**, thereby simplifying the study of interesting objects, phenomena and systems.

The word **"model"** means "measure", "resemblance to a certain thing", "method".

Model building is a universal way to study the surrounding world, enabling to:

- detect dependencies;
- predict;
- divide into groups;
- solve many other problems.

The main goal of the modelling consists of the fact that the model has to display the functioning of the modelling system sufficiently well.

A model is an object or description of a situation, a system to substitute (under certain conditions, assumptions,

hypotheses) one system (the original system) by another system for the better study of the original or reproduction of its any properties.

Modelling is a universal method to obtain, describe and use knowledge. It is used in any professional activity.

The models are divided b**y the type into**:

- empirical – obtained on the basis of empirical facts, dependencies;
- theoretical – obtained on the basis of mathematical descriptions, laws;
- mixed, semiempirical – obtained on the basis of empirical dependencies and mathematical descriptions.

Often, **theoretical models emerge from empirical ones**, for example, many laws of physics were originally derived from empirical data.

Example

A great deal of enterprises operates at the market, exchanging goods, raw materials, services, information. If we describe economic laws, the rules of interaction at the market through mathematical relationships, for example, through a system of algebraic equations, where the profit quantities obtained from the interaction between enterprises and the coefficients of the equation are such interactions intensities. Then the mathematical model of the economic system will be obtained, that is an economic and mathematical model of the enterprise's system at the market.

Thus, **data analysis** is closely related to modelling.

Important features of any model are:

- **Simplicity.** The model displays only the essential sides of the object and, at the same time, should be simple to explore or to reproduce.
- **Finiteness.** The model displays the original only in a finite number of its relations, and, in addition, the modelling resources are finite.

- **Approximation.** The reality is displayed by the model roughly or approximately.
- **Adequacy.** The model should successfully describe the simulated system.
- **Integrity.** The model implements a certain system (that is, an integer).
- **Closedness.** The model takes into account and displays a closed system of necessary basic hypotheses, connections and relations.
- **Manageability.** The model must have at least one parameter, the changes of which can simulate the behaviour of the modelling system under different conditions (Provost F., 2013).

Basic approaches to modelling:

1. An **analytical approach** to modelling.

In the traditional sense, the model is the result of **displaying one structure** (studied) **to another** (not well-known).

So, displaying a **physical system** (object) **on a mathematical** (for example, a mathematical apparatus of equations), we obtain a **physical-mathematical model** of the system or a mathematical model of the physical system. Any model is constructed and explored under certain **assumptions, hypotheses**. This is usually done thanks to the **mathematical methods.**

Example.

Let us consider the economic system. The expected demand 's' for the following month (t + 1) is calculated by the formula:

$$s(t+1) = \frac{[s(t)+s(t-1)+s(t-2)]}{3}, \qquad (1)$$

Thus, it is the **average sale**s in the previous three months. It is the simplest mathematical model of the **sales prediction**.

While constructing this model, the following **hypotheses** were adopted:

- firstly, the annual seasonality in sales is absent.
- secondly, the amount of sales is not influenced by any external factors: actions of competitors, macroeconomic situation, etc.

It's easy to use this model: thanks to the sales data for the previous months, we get the prediction for the next month according to the formula.

In the literature, this approach to modelling is **called analytical**. The analytical approach to modelling is based on the fact that the researcher rests on the model in the study of the system (Figure 1).

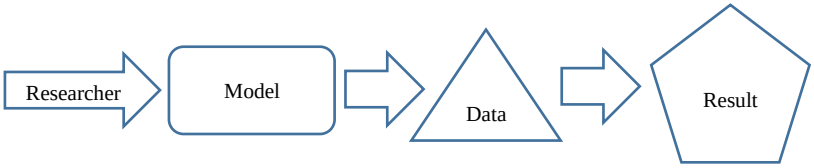


Figure 1 – Structural and logical scheme of the analytical method

In this case, the researcher chooses **the desired model** for one or another reason. Typically, this is:

- theoretical model;
- the law;
- the dependence, which is more often represented in the functional form (for example, an equation that binds the output parameter 'y' to the input parameters of influence x1, x2 ...).

A variation of the input parameters at the output will provide the result, which simulates the behaviour of the system under different conditions.

The result of the simulation **may correspond to reality, and may not be consistent**. In the second case, the researcher has nothing left, **except to choose another model or another**

method to study it. The new model may be **more appropriate to describe** the system under consideration (Seigel E., 2016).

In an analytical approach, the model is not "adapted" to reality, but the researcher is trying to **find the existed analytical model** in such a way that it adequately reflects reality.

The model is always **explored by a certain method** (numerical, qualitative, etc.). Therefore, the choice of simulation method often means to choose a model.

2. Informational approach.

When using the traditional analytical approach in business, problems often arise owing to the discrepancy between the methods of analysis and the reality, which they should reflect. There are difficulties associated with the formalization of business processes. Here, the factors that determine the phenomena **are so varied and numerous**, and their **interconnections are so "intertwined"** that it is almost impossible to create a model that satisfies similar conditions.

The simple imposition of known analytical methods, laws, dependencies on the investigated picture of reality, in this case, will not bring success.

The basis of complexity and weak formalization of business processes lies primarily in the **human factor**, therefore it is difficult to judge the nature of the laws a priori (and sometimes a posteriori, after the implementation of any mathematical method).

With the same success, different models can describe these patterns. Using different methods to solve **one and the same problem** often leads to the **opposite conclusions.** So which method should the researcher choose? The answer to this question can only be based on deep analysis either of the reason for the problem to be solved or the specifications of the used mathematical apparatus.

Therefore, in recent years, an **informational approach** to data-driven modelling has become widespread. **Its main goal**

is to release an analyst from routine operations and possible difficulties in understanding and applying modern mathematical methods.

In the informational approach, the **real object is considered as a "black box"**, which had a number of inputs and outputs, between which certain communications are modelled. In other words, **only the structure of the model is known** (for example, the neural network, linear regression), and the model parameters themselves are "customized" for data describing the behaviour of the object. To correct the model parameters, feedback is used – the model deviates from reality and the process to set up the model often has an iterative (i.e., cyclic) character (Figure 2).

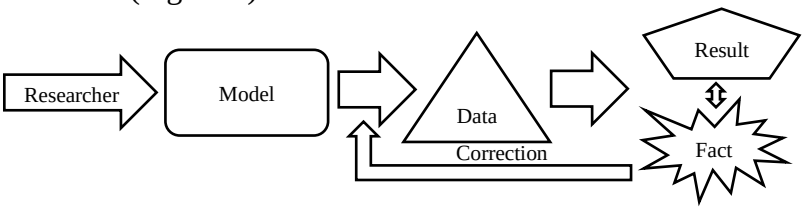


Figure 2 – Structural and logical scheme
of the information method

Thus, according to the **informational approach**, the starting point is the data characterizing the object under study and the model "adjusts" to reality.

That is why in an **analytical approach**, the researcher can choose a model, even without any experimental data characterizing the properties of the system and can start to use it. According to the **informational approach,** it is impossible to construct the model without data since its parameters are completely determined by them.

Example.

In banking risk management, the Durant model to calculate the rating of the borrower's creditworthiness is widely

known. It was spread throughout the 1940s and 1950s. Based on his own experience, Durant developed a point model to estimate the borrower according to his or her property and social parameters (age, sex, profession, etc.). Having overcome certain problems, the borrower was considered creditworthy. This model is an analytical dependence of y=*f*(x), where 'y' is the rating, 'x' is the set of borrower's features.

If a modern bank faces the task to calculate the borrower's rating, it can use the Duran model. *However, will the model, developed in the middle of the last century in the West, be adequate for the modern Ukrainian reality?* Naturally, it will not, because it does not take into account the patterns between the characteristics of Ukrainian borrowers (age, education, income, etc.) and the problem of loans. If the bank accepts its own credit history data and builds a model to calculate its own customer ratings, *then it is likely that it will be operational.*

In the first case, when we took the Durant model, we used an analytical approach. In the second - information; in order to build the model, we needed the data - the credit history of the bank borrowers.

In the first case, when we took the Duran model, we used an analytical approach. *In the second one* – informational: we needed data - credit histories of bank borrowers for the construction of the model.

The models, obtained through the **informational approach,** take into account the specific features of the modelled object, phenomena, unlike the analytical approach. This quality is extremely important for business processes, so the information approach was the basis of most modern industrial technologies and data analysis methods: Knowledge Discovery in Databases, Data Mining, Machine Learning.

However, the **"data models"** concept requires careful consideration to the quality of the output data, since false,

abnormal and noisy data can lead to models and conclusions that have nothing to do with reality.

Therefore:

- data modelling;
- data consolidation;
- cleaning;
- enrichment play an important role in information modelling.

The process of analysis.

There are **three important components:** an expert, a hypothesis, and an analyst in addition to the model in the informational approach to the data analysis.

An expert is a specialist in the subject area, a professional who, in the course of his or her studies and practical activity, has learned how effectively to solve problems related to a particular subject area.

An expert is a **key figure** in the analysis process. Really effective analytical solutions can be obtained not just based on the computer programs but as a result of a combination of the best things that a man and a computer can do.

The expert **proposes hypotheses (assumptions)** and he can review sample data in a variety of ways; builds this or that model to check their reliability.

Example.

A hypothesis in the data analysis is often the assumption about the influence of any factor or group of factors on the result.

For example, when **constructing a sales prediction,** it is assumed that the size of future sales is significantly affected by sales for previous periods and remains in stock.

In assessing **the potential borrower's creditworthiness**, there is a hypothesis that the

creditworthiness is influenced by the client's social and economic features: age, education, marital status etc.

As a rule, several experts, and an analyst usually participate **in large projects** regarding the creation of applied analytical solutions.

The analyst is an expert in the field of analysis and modelling. An analyst at a sufficient level possesses some data analysis tools and program applications for data analysis, such as Data Mining methods. In addition, the analyst's duties include functions of data systematization, experts' interviewing, coordinating the actions made by all participants involved in the project for data analysis.

The analyst plays the **"connecting link"** between specialists of different levels and areas, that is between experts.

In the course of his or her work he or she:

- accumulates various hypotheses from experts;
- imposes data requirements;
- verifies hypotheses and analyzes the results with experts.

The analyst must have **systemic knowledge** since in addition to the analysis tasks he often takes responsibilities to solve technical issues related to:

- databases;
- their integration with data sources;
- the productivity of databases and sources.

Therefore, in the process of data analysis, the analyst is considered the **main responsible person**. This means that he closely cooperates with field experts.

Despite the fact that there are many analytical tasks, it is possible to distinguish **two main groups of methods** for their solution (Figure 3):

- data extraction and visualization;
- models construction and use.

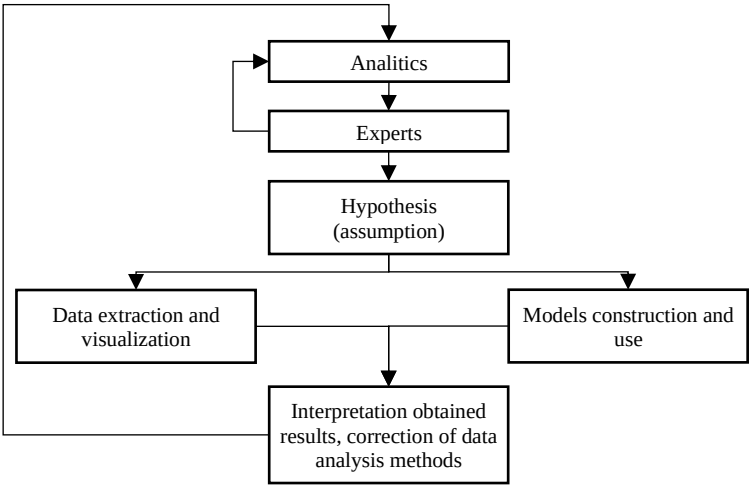


Figure 3 – General data analysis scheme

In order to gain new knowledge about the object or phenomenon in the study, **it is not necessary to construct complex models.** Often it is enough to "look" at the data in the right way to make certain conclusions or to suggest assumptions about the nature of dependencies in the system, to get an answer to questions of interest in a particular situation. **Visualization helps to make it**.

In the questions of visualization, an analyst formulates the request to the information system, extracts the necessary information from different sources and reviews the results. On their basis, he makes the conclusions that are the result of the analysis.

There is a large number of ways **to visualize data:**

- multidimensional cubes (cross-tables and cross-diagrams);
- tables;
- diagrams, histograms;
- maps, projections, slices, etc.

Table 1 is an example of data visualization in the Excel table form. Figure 4 is a graphical and histogram visualization.

Table 1 – Tabular data visualization (abstract data)

Year	Manufacturer	Model	Sales	Minimal price, grn.
2017	Toyota	Camry	345	550000
2018	Toyota	Camry	300	587000
2019	Toyota	Camry	275	653000
2020	Toyota	Camry	250	867000
2017	Toyota	Highlander	94	897000
2018	Toyota	Highlander	90	974000
2019	Toyota	Highlander	83	1256000
2020	Toyota	Highlander	71	1405600
2017	KIA	Rio	545	395000
2018	KIA	Rio	700	415000
2019	KIA	Rio	850	470000
2020	KIA	Rio	903	500000
2017	Nissan	X-Trail	456	527000
2018	Nissan	X-Trail	425	581000
2019	Nissan	X-Trail	390	675000
2020	Nissan	X-Trail	387	723000

In the first case, according to the table, it is difficult to make certain conclusions about the car sales dynamics, since it reflects the sales for several auto producers, several models during 4 years.

In the second variant, representing the data via a combined diagram (diagram + histogram), we can make conclusions about the **dependence between the level of cars sales and their minimum prices dynamics growth.** Having built the combined diagrams or one consolidated diagram, we can conclude that during 4 years the minimum price for all models increased, and the number of units sold, respectively,

decreased. Based on these data, we can take certain price, advertising, managerial **decisions**.

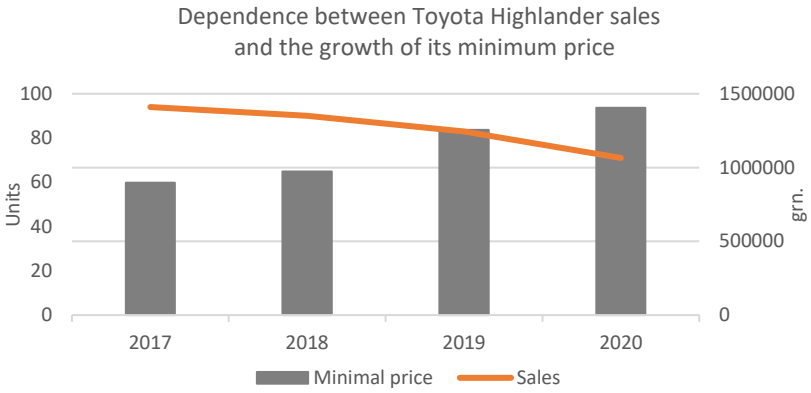


Figure 4 – Analytical data in the form of a chart and histogram (abstract data)

The advantages of visualization include:

- simplicity to create and to implement such systems in practice;
- an ability to apply them practically in any sphere of activity;
- expert's knowledge in the subject area and his or her ability to take into account factors that have a significant impact on business, but which are difficult to formalize by tabular means, are used to the maximum;
- replication of knowledge.

Disadvantages of visualization include:

- the inability of people majority to find complicated and non-trivial dependencies;
- impossibility to separate knowledge from the expert.

Model building is a universal way to study the surrounding world, which enables to detect dependencies,

predict, group, and to solve many other important tasks. The most important thing is that the obtained knowledge can be replicated.

Replication of knowledge includes a set of methodological and instrumental means for model building, it allows final consumers to use the modelling results in decision-making without the need to understand the methods thanks to which these results are obtained (Kovtun N., 2015).

The process to construct models includes several steps:

1. **Formulation of the model building aim.** When constructing a model, one has to consider a task, in this case, as an answer to a question or a range of issues, in which the customer is interested.

Examples of such issues:

- What goods are sold in the country in summer, and which ones are not?
- What colour do middle-aged women buy in this city?
- What nomenclature of goods was sold in the country during the year?

In this case, we are talking about the formation of the model to predict sales, the model to identify associations, etc. This stage is also called **analysis of the problem situation.**

2. **Data preparation and collection.** The information approach to modelling is based on the use of data. Its preparation and systematization is a separate task for analytics.

3. **Selection of the model.** Having collected and systematized data, analytics searches for a model, which explains the available data and lets to get empirically valid answers to the questions of interest. **In the industrial analysis of data**, preference is given to self-learning algorithms, machine learning, Data Mining methods.

If the constructed model **shows favourable results in practice** (for example, during the test operation), it is launched into commercial operation.

Thus, while testing a scoring model that calculates a client's credit rating and makes a decision on granting a loan, each decision can be confirmed by a person – **a credit expert**.

While scoring in **industrial operation**, the human factor is removed –the decision is taken only by the computer.

If the model quality is unsatisfactory, the process of its constructing is repeated.

Model building enables to obtain new knowledge **that cannot be obtained in any other way.** In addition, obtained results make a formal description of a process, therefore these data are subjected to automated processing.

However, the results obtained with the use of models **are extremely sensitive to:**

- data quality;
- the analyst's and expert's knowledge;
- organization of the research process.

In addition, usually, there are cases that **do not fit into any model.**

In practice, approaches are most often combined. For example, data visualization gives an analyst some ideas which he or she tries to test via different models, and the visualization methods are applied to the results.

A full-featured analysis system should not be closed to the use of only one approach or one technique. The mechanisms to visualize and construct models should complement each other. **Maximum returns** can be obtained by combining methods and approaches to data analysis.

Data, describing real objects, processes, and phenomena, **can be represented in different forms, types and classified by essential features.**

Data is information that describes a system, phenomenon, process or object, presented in a certain form and intended for further use.

The following forms of data representation **are distinguished by the structuring degree**:

- unstructured;
- structured;
- weakly structured.

Unstructured data consists of any form, including text and graphics, multimedia (video, speech, audio). This form of data representation is widely used, for example, on the Internet, and the data is presented to the user as a link by the search engines.

Structured data reflects individual facts of the subject area. Structured data is considered to be well-structured and organized in order to allow certain actions to be applied to them (for example, visual or machine analysis). This is the main form of information representation **in databases.**

The particular type of data storage (structured or unstructured) and its organization is associated with **access to them.**

Access to data is the possibility to allocate an element of data (or a plurality of elements) among other elements on any basis in order to perform certain actions on them.

One of the most common types of models for data structuring **is the table**. Within it, all data is ordered into a two-dimensional structure consisting of **columns and rows** (Figure 5).

The cells of such a table **contain data elements:**

- symbols;
- digits;
- logical values.

Unstructured data is unsuitable for processing directly by data analysis methods, so such data is **subjected to special structuring techniques**, and the character of the data during the structuring process can change significantly.

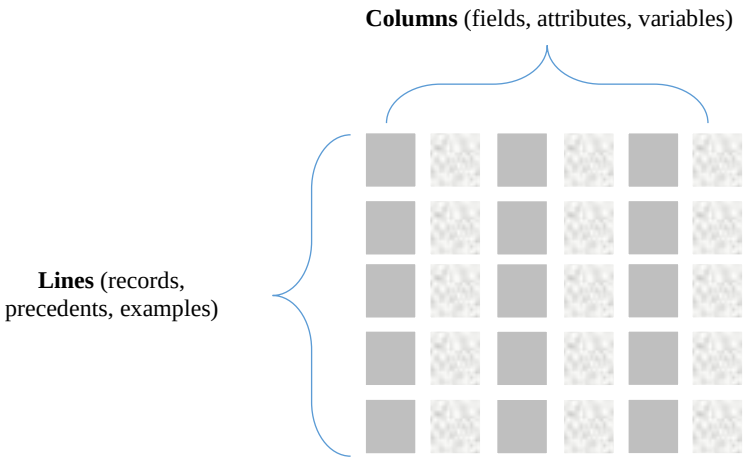


Figure 5 – Tabular form of data structuring

Example

In text analysis (Text Mining), when structuring, a table with a frequency of words repetitions can be generated from the source text, and such a data set will already be processed by methods suitable for structured data.

Weakly structured data is data for which certain rules and formats are defined but in the most general form.

Example

- line with address;
- a line in the price list;
- name and surname;
- place of residence;
- product model;
- the type of goods etc.

Unlike unstructured data, such data **is converted to a structured form with less effort**, but without a conversion procedure, they are also unsuitable for analysis (Figure 6).

Toyota Corolla, 2018, gas engine, 1.6l, automatic transmission, 2WD, City

Attribute	Characteristic
Manufacturer	Toyota
Model	Corolla
Engine	Gas
Engine volume, l.	1,6
Transmission	Automatic
Drive unit	2wd
Equipment	City
Production year	2018

Figure 6 – Transformation of the general information in a structured form (abstract data)

The vast **majority of data analysis methods** can work adequately only with well-structured data presented in tabular form. Therefore, further information will be equitable for structured data.

Structured data is divided into **five types:**

- numeric (quantity of goods, product code, etc.);
- cost (price, discount, etc.);
- essential (last name, name, address, gender, education, etc.);
- logical;
- chronometric (date and time).

Among the essential data, **two subtypes can be distinguished:**

- ordered;
- categorical.

In both these cases, the variable refers to values of the discrete (plural) classes set c_1 ... c_k and describes some qualitative properties for the object. But if in the case of **ordinary data**, these classes can be arranged, then it is impossible to do in the case of a **categorical data type**. When comparing categorical data, only two operations can be applicable: "equal" (=) and "not equal" (≠) (Table 2).

Table 2 – Ordered and categorical data

Flight number	Plane	Plane type	Shipment	Destination
UA4246	Boeing 737-700	Light (Passenger)	Passengers	Seychelles International Airport, Victoria, Seychelles
UA5128	An-225	Heavy (Cargo)	Turbine generator	John Kennedy International Airport, New York, USA

In this case, **only one field** – the plane type is ordered, and the other – categorical one.

It means that the values in the "Plane Type" field **can be ordered from light to heavy**. At the same time, in other fields, making **such an order is senseless**. Values "equal" or "not equal" can only be used for them.

According to the essential features, data can be divided into two main groups:

- continuous data;
- discrete data;

Continuous data are data which can take any value in a certain interval. The analyst can perform arithmetic operations of addition, subtraction, multiplication, division with continuous data, and they will make sense.

Examples of continuous data include age, any cost indicators, quantitative estimates (quantity of goods, sales volume, the weight of shipment).

Discrete data – the value of an indication, the total number of which is finite or infinite, but they can be calculated with the help of natural numbers from one to infinity. No arithmetic operations can be made with discrete data or such operations do not make sense.

Discrete data is all data of **essential and logical type. Numerical data** may also be discrete.

Example.

The field **Product code, which has** values of the whole type, is discrete, since operations of adding, subtracting, multiplying by product code are meaningless.

The correspondence between types and essential features of data is shown in Table 3.

Table 3 – Correspondance of types and essential features of the data

Data types	**Essential characteristics**	
	Continuous data	**Discrete data**
Numeric	Yes	Yes
Cost	Yes	Yes
Essential	–	Yes
Logical	–	Yes
Chronometric	Yes	Yes

It is important for the analyst to understand the **nature of the data** in order to choose adequate methods for their pre-treatment, cleaning and model building.

In relations to the analysis task, **data sets can be:**

- ordered;
- unordered.

In the **ordered dataset**, each column corresponds to one factor, and each line contains ordered events at intervals between the lines. Often time is such a sign (Table 4).

Table 4 – The example of ordered data

Date	Sold units	Unit price
07.01.2020	5	8
08.01.2020	20	8
09.01.2020	17	8
10.01.2020	11	8

In case of an **unordered set**, each column corresponds to the factor and the example (situation, precedent) is entered in each row, respectively, **it is not necessary** mechanically to order the lines (Table 5).

Table 5 – The example of unordered data

№	Last name	Education	Salary
1	Ivanenko	Higher	10500
2	Sidorenko	Secondary	3200
3	Petrenko	Secondary	3700
4	Nesterenko	Higher	9800
5	Nikonenko	Higher	8500
6	Opanasenko	Secondary	5000
7	Kotenko	Higher	11450
8	Andryushchenko	Higher	12500
9	Onishchenko	Secondary	4250
10	Karpenko	Higher	9650

Transactional type is separately distinguished. A **transaction** means several objects or actions that are a logically related unit. This mechanism is used to analyze purchases (checks) in supermarkets. However, generally, any related objects or actions can by transactional (Table 6).

Table 6 – An example Transaction Data

Transaction №	Commodity	Price
102/25	Coca-cola	27
103/21	Milk	10
102/25	Chocolate	21
104/12	Meat	50

The informational approach to analysis is based on **various algorithms to extract regularities** from the data source, the results of which make the models. There are many

such algorithms, but they cannot guarantee a high-quality solution. No method, even very sophisticated ones, will produce a good result **since the quality of the initial data is critically important.** Most often, it causes failure.

Features of data, accumulated in the companies.

Data, accumulated by enterprises and organizations in databases and other sources (so-called business data), **have their own peculiarities:**

1. Business data is rarely accumulated especially to solve tasks of the analysis. Businesses and organizations collect data for commercial purposes:

- accounting;
- financial analysis;
- reporting;
- decision making, etc.

This specificity is the **foundation for the separation** of business data and experimental data, which always accumulate with research aims.

Usually, **the main consumers of business data** are those who make decisions in companies.

2. Business data, as a rule, contain errors, anomalies, contradictions and omissions. This situation arises because companies do not collect data for analysis. Such data has various errors, and thus, its quality is reduced.

3. Extremely large volumes of data. Modern databases contain gigabytes and terabytes of information. For resource-intensive data analysis algorithms, a table with 50.000 entries can be considered as large, so it is important to use **sampling procedures**, to reduce the entries number and selection of informational features, or to use special scalable algorithms that can handle large volumes of information in the construction of models.

The specified features of business data affect both the analysis process and the preparation and systematization of data.

Formalization of data.

When collecting data, it is necessary to **follow the next principles:**

1. To abstract from current information systems and available data. Large volumes of accumulated data absolutely do not say that they are sufficient for analysis in a particular situation. It is necessary to rely on the task and to select data for its solution, **instead of using available information.**

Example

When constructing sales forecast models, a experts survey has shown that the demand for the product is strongly influenced by the colour gamut of the product. An analysis of available data has shown that information on the product colour is not available in the accounting system. This means that the company needs to add this data. Otherwise, analytics should not rely on the high quality of the implemented model.

2. To describe all factors, which potentially affect the process/object being analyzed. The main tool, in this case, is a survey of experts and people who directly know the problem situation. It is necessary to make maximum use of expert knowledge in the subject area and, relying on common sense, to try to collect and systematize the maximum of possible assumptions and hypotheses.

3. To assess expertly the significance of each factor. This assessment is not final, it will be a starting point. In the process of analysis, it may well be clear that the factor, which experts considered as relevant is not such, but insignificant, in their view, the factor can have a powerful effect on the outcome.

4. To determine the way of presenting information – number, date, yes/no, category (that is, **data type)**. It is easy to determine the way to present such data – to formalize it.

Example

Sales in dollars are a certain number. But often it is unclear how to present a factor.

Frequently such problems arise with **qualitative features.**

Example

The volume of sales affects product quality. Quality is a complex concept, but if this indicator is really important then we need to think of a way to formalize it. Quality can be determined by the number of defects per thousand units or evaluated expertly by making several categories – excellent/good/satisfactory/bad.

5. To investigate all easily available factors of impact and to consider the most relevant ones. They can be contained primarily in sources of structured information - accounting systems, databases etc. Obviously, it is impossible to build a qualitative model without them.

6. To assess the complexity and cost of collecting the average and least important by significance factors. Some data are readily available, they can be taken from existing information systems. Nevertheless, there is information that is not easy to collect, such as information about competitors, so we need to estimate what is the cost of collecting such data. **Data collection is not an end.** If the information is easy to obtain, then, accordingly, it is necessary to collect it. If it is difficult, then it is necessary to compare the costs for its collection and systematization with the expected results.

It is possible to consider these principles on the example of data formalization in the situation, **when we need to make a demand prediction.** At the stage of describing the factors (that

affect sales) and making hypotheses, it is useful to compile a table with factors and their significance (Table 7).

Table 7 – Factors that have an impact on sales and their significance

Factor	Expert evaluation of significance on a scale (1-100)
Season	100
Weekday	80
Sales volume in previous weeks	100
Sales volume for the same period last year	95
Advertising company	60
Marketing activities	40
Product quality	30
Brand rating	25
Deviations from the average market price	60
Availability of this product from competitors	15

It is necessary to observe 1-3 positions of data formalization. After that, it is necessary to determine the way in which data should be presented and to estimate the cost of its collection. An additional two columns are added to the table (Table 8). Then analyst can decide what factors to include in the analysis, and which to neglect. **It is obvious that all readily available indicators with a high expert level should be included in the review.** A factor "Quality of Products", for example, can be neglected: according to experts, its significance is low, and the cost to collect such information is relatively high.

Table 8 – Factors that have an impact on sales with expert estimates

Factor	Expert evaluation of significance on a scale (1-100)	The way of submitting data	An expert assessment of the cost for collecting such data
Season	100	numeric	low
Weekday	80	date	low
Sales volume in previous weeks	100	numeric	low
Sales volume for the same period last year	95	numeric	low
Advertising company	60	numeric	average
Marketing activities	40	numeric	average
Product quality	30	high/medium/ low	high
Brand rating	25	known/little-known/unknown	average
Availability of this product from competitors	15	Yes/No	average

Methods of data collection.

There are several methods to collect data, which are necessary for analysis:

1. To receive data from accounting systems. Usually, in accounting systems, there are various mechanisms to construct reports and to export data, so downloading the necessary information from them is often a relatively simple operation.

2. To receive data from other sources of information. Many indicators can describe indirect signs. For example, the evaluation of the real financial situation of residents in a region by the **ratio of goods sold to the poor, middle class and rich people.**

3. To use the open sources (statistical collections, corporate reports, published results of marketing research etc.).

4. To purchase analytical reports from specialized companies. Collected information is usually presented in the form of different tables and summaries that can be successfully applied in the analysis. The cost to obtain such information is often relatively low.

5. To carry out own marketing researches, expert assessments and similar measures on data collection. This data collection option can be quite expensive and labour-intensive.

Informativeness of the data.

One common mistake in collecting data from structured sources **is the desire to take as many features as possible to describe objects for analysis.** At the same time, a preliminary evaluation of the data, which is performed visually using tables and basic statistical information on the data set, significantly helps to determine the informative nature of the features in analysis terms.

Among the **non-informative features**, there are four types (Table 9):

a) signs that contain only one meaning;

b) signs that contain basically the same meaning;

c) signs with unique values;

d) signs between which there is a strong correlation – in this case, one column can be taken for analysis.

Table 9 – Non-informative data features

a)	b)	c)	d)	
Feature	Feature	Flight number	Gender	Characteristics
1	1	UA4022	Male	0
1	1	UA4527	Female	1
1	1	UA5032	Male	0
1	0	UA5537	Male	0
1	1	UA6042	Female	1
1	1	UA6547	Female	1
1	1	UA7052	Female	1

However, signs containing **essentially the same meaning may not always be non-informative, in many cases, it depends on the purpose of the analysis.** For example, when solving a problem regarding the analysis of deviations, such signs can significantly affect the model construction.

Data requirements.

Analytical tools to build models are based on the input data, **so the data that are as close as possible to reality will give an adequate result in the modelling process.** It is important to understand that the model cannot "know" about what is outside of the data collected for analysis.

There are requirements for the minimum amount of data needed to build models based on them. Depending on the presentation of the data and the task to be solved, **these requirements are different:**

1. **For time series that relate to ordered data:**

If a simulated business process (for example, sales) is characterized by **seasonality/cyclicality**, it is necessary to have data for at least one full season/cycle with the possibility of interval variation (weekly, monthly, etc.).

The maximum horizons of prediction depend on the data amount: data for 1.5 years - the prediction is the possible maximum for 1 month; data for 2-3 years - for 2 months.

2. For unordered data:

- The number of examples (precedents) should be significantly more than the factors number.
- It is desirable that the data cover as much as possible real-world situations.
- The proportions of different examples (precedents) should approximate the actual process.

3. Transaction data. It is expedient to analyze the transactions on a large amount of data, otherwise statistically unreasonable rules may be found. The search algorithms for associative links can quickly process huge amounts of data. The approximate relationship between the number of objects and the data volume is:

- 300-500 objects – at least 10 thousand transactions;
- 500-1000 objects – more than 300 thousand transactions.

An informational approach to data analysis has been reflected in such knowledge extraction techniques as Knowledge Discovery in Databases (KDD) and Data Mining. Today, **based on these techniques, most applied analytical solutions in business** and many other areas are created.

Despite the business tasks diversity, almost all of them can be solved by a single method. This technique, which originated in 1989, was called **Knowledge Discovery in Databases** – extraction of knowledge from databases.

It describes **not a particular algorithm or mathematical apparatus, but a sequence of actions** that must be performed to identify useful knowledge.

The methodology does not depend on the subject area; this is **a set of atomic operations,** combining by which, analytics can get the right solution.

Atomic operations are operations that are either executed in full or not executed at all; an operation that cannot be partially performed and partially not performed.

KDD includes the next steps:

- data preparation;
- choice of informative features;
- cleaning;
- building models;
- postprocessing;
- interpretation of the results obtained.

The core of this process **is Data Mining**, which enables to identify patterns and knowledge.

Data collection. The first step in the analysis is to get the initial sample. Models are built on the basis of the selected data.

At this stage, **active participation of experts is needed** to promote hypotheses and to select the factors that influence the process under analysis. It is better that data have been already collected and consolidated.

It is extremely necessary to get convenient mechanisms of sample preparation:

- requests to databases;
- data filtering;
- sampling.

Often, as a source, it is recommended to use a **specialized data repositories** that consolidate all the information needed for analysis.

Data clearing. Real data for analysis do not **often have sufficient quality.** The necessity for prior processing during the data analysis occurs regardless of technologies and algorithms which are used with this purpose. Moreover, this task can gain an independent value in areas that are not directly related to data analysis.

Data cleaning tasks include:

- fill in gaps;
- suppression of abnormal values;
- anti-aliasing (smoothing);
- exclusion of duplicates and contradictions, etc.

Data transformation. This step is required for those methods in which the initial data **should be presented in some definite form.** It should be noted that various analysis algorithms require data prepared in a special way.

Example

For certain prediction, it is necessary to convert the time series using a sliding window method or to calculate aggregate metrics.

The instruments of data transformation include:

- sliding window;
- bringing types;
- allocation of time intervals;
- quantization (splitting the range of values for a certain indicator into a finite number of levels and rounding up these values to the nearest levels);
- sorting;
- grouping etc.

A sliding window converts a sequence of row values to a table where neighbouring records are represented as adjacent data fields (a window – since only a certain continuous array of data is allocated, sliding – since this window "moves" throughout the sample) (Marchenko O, 2017).

The need for such a table often occurs during constructing models, analyzing and forecasting time series when it is necessary to input the value of several adjacent indices from the output data set.

The values in one of the recording fields will be **for the current count,** while others are shifted from the current **count**

to "the future" or "to the past". Thus, the conversion of a sliding window has two parameters:

- the depth of immersion – the number of "past" counts falling into the window;
- prediction horizon – the number of "future" counts.

Example

We have a history of sales for six months by months (Table 10).

Table 10 – Example of sales history

The first day of the month	Sales volume (thousand dollars)
01.01.2022	1000
01.02.2022	1160
01.03.2022	1210
01.04.2022	1130
01.05.2022	1250
01.06.2022	1300

Incomplete entries will be formed, i.e. records containing empty values for missing past or future counts **for boundary** (limit relative to the sample beginning and end) window values.

The conversion algorithm enables to exclude such records from the sample (then, for several limit counts, the records will not be formed) or to include them (then records are generated for all existing counts, but some of them will be incomplete) (Mize Ed., 2017).

If we set the depth of immersion 2 and the forecast horizon 1, then we obtain the following table with incomplete records (Table 11) or with full records (Table 12).

Table 11 – Data representation in the form of a sliding window with incomplete records

The first day of the month	Sales volume 2 months ago	Sales volume last month	Current sales	Expected sales next month
–	–	–	–	1000
01.01.2022	–	–	1000	1160
01.02.2022	–	1000	1160	1210
01.03.2022	1000	1160	1210	1130
01.04.2022	1160	1210	1130	1250
01.05.2022	1210	1130	1250	1300
01.06.2022	1130	1250	1300	
–	1250	1300	–	–
–	1300	–	–	–

Data Mining. Directly at this stage models are being built.

Interpretation. In the case where the extracted dependencies and templates are not transparent to the user, there must exist post-processing methods that allow them to be interpreted.

Table 12 – Data representation in the form of a sliding window with complete records

The first day of the month	Sales volume 2 months ago	Sales volume last month	Current sales	Expected sales next month
01.03.2022	1000	1160	1210	1130
01.04.2022	1160	1210	1130	1250
01.05.2022	1210	1130	1250	1300

To assess the quality of the model obtained, **both formal methods and the analyst's knowledge should be used.**

The analyst can tell how applicable the model is to real data. The constructed models are, in fact, formalized expert knowledge, and therefore, **they can be replicated.** The knowledge found should be applicable to new data with some degree of certainty.

Data Mining – the detection in raw data of previously unknown, non-trivial, practically useful and accessible for knowledge interpretation and necessary for decision-making in various spheres of human activity (Provost F., 2013).

Dependencies and templates found during the application of Data Mining methods should **be non-trivial** and **previously unknown**, for example, information on average sales is not such one. Knowledge should describe **new relationships** between properties, predict the values of some characteristics based on others.

Often, **KDD is identified with Data Mining**. However, it is more correct to consider Data Mining as the **step of the KDD process**.

Data Mining is not one method, but a large combination of different methods to detect knowledge. There are several conditional classifications of Data Mining Tasks. There are four basic classes of tasks:

1. **Classification** means to set up the dependence of a discrete output variable on input variables.
2. **Regression** means to set up the dependence of the continuous output variable on the input variables.
3. **Clustering** means to group objects (observations, events) based on data describing the properties of objects. Objects within the cluster must **be similar to each other and different from others** that have entered other clusters.

4. **Association** means to identify patterns between related events. An example of such regularity is the rule that indicates that event Y follows the event X. **Such rules are called associative**. For the first time this problem was proposed in order to find typical shopping patterns made in supermarkets, so sometimes it is called **a market basket analysis**. If we are interested in the sequence of events, we can talk about **successive patterns** – the establishment of patterns between time-bound events. An example of such regularity is the rule that indicates that event Y will be followed by the event X after the time t.

In addition to these tasks:

- deviation detection,
- link analysis,
- feature selection

Although these tasks border with **data cleaning and visualization.**

The **task of classification** differs from the regression task by the fact that there is a variable of a discrete type, so-called **a class label** in the classification. The solution of the classification problem lies in the definition of an object **class by its features**, and the set of classes to which an object can be related **is known in advance.**

In the **regression** problem, the output variable is a continuous set of real numbers, for example, the sum of sales. The problem of regression is solved, in particular, through **prediction of time series** based on historical data.

Clustering differs from the classification in the fact that an output variable is not required, and the clusters quantity in which it is necessary to group the entire **set of data may be unknown**. The clustering output is not a ready-made answer (for example, "bad"/"satisfactory"/"good"). Clusters are groups of similar objects. Clustering only indicates the **similarity of objects, and nothing more.** In order to explain the formed clusters, it is necessary to interpret them additionally (Figure 7).

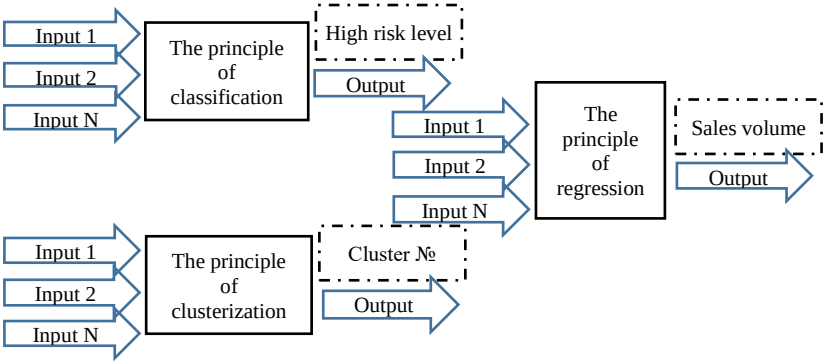


Figure 7 – Classification, regression and clustering principles

The most **famous ways to apply these tasks** in the economy.

Classification is used if classes are known in advance, for example, when assigning a new product to a particular commodity group, assigning a customer to a category (in case of credit, to one of the risk groups).

Regression is used to establish dependencies between factors. For example, in the prediction task, the dependent quantity is sales volumes, and the factors, affecting it, can be previous sales volumes, currency exchange rates, competitors' activity, etc. Alternatively, for example, when lending to individuals, the probability of a loan repayment depends on the customer's personal features, the scope of his or her activities, the availability of the property.

Clustering can be used to segment clients and to build up their profiles. Owing to sufficiently large number of customers, it becomes difficult to approach each individually, so it is convenient to combine them into groups – segments with homogeneous features. Several groups of features, for example, can select segments by area of activity or geographical location. After clustering, you can find out which segments are most active, which ones bring the greatest profit, and highlight the

characteristics features for them. The effectiveness of working with clients is enhanced by taking into account their personal preferences.

Associative rules help to identify jointly acquired goods. It can be useful for more convenient placement of goods on the shelves, stimulating sales. In such a case, the person who bought a pack of spaghetti, will not forget to buy a bottle of sauce.

Successive patterns can be used to plan sales or to provide services. They are similar to associative rules, but in the analysis, a time indicator is added, that is, the sequence of operations is important.

In order to solve the above tasks, various methods and algorithms are used in Data Mining. Due to the fact that Data Mining is developing at the intersection of such disciplines as mathematics, statistics, information theory, machine learning, database theory, programming, parallel computing, it is quite natural that most algorithms and methods of Data Mining were developed on the basis of the approaches used in these subjects.

In general, it is not important **which algorithm will be used to solve the problem**, the main thing is to have a solution method for each class of tasks. Nowadays, **methods of machine learning** are the most widespread in Data Mining:

- decision trees;
- neural networks;
- associative rules etc.

Machine learning is an extensive subsection within the artificial intelligence studies methods to construct algorithms that can be trained on data.

The general formulation of the **training problem is as follows**: there are many objects (situations) and many possible answers (responses, reactions). There is some unknown dependence, between the answers and the objects. Only a finite precedent set is known – pairs of the form "object-response". It is called a **training sample**. Based on these data, it is required to

find a **dependency, that is, to construct a model capable to give an enough accurate answer for any object.** To measure the accuracy of answers, a **quality criterion** is introduced.

At the same time, the **overwhelming majority of business tasks are based on the KDD process.**

Data Mining algorithms solve some popular business tasks (Figure 8).

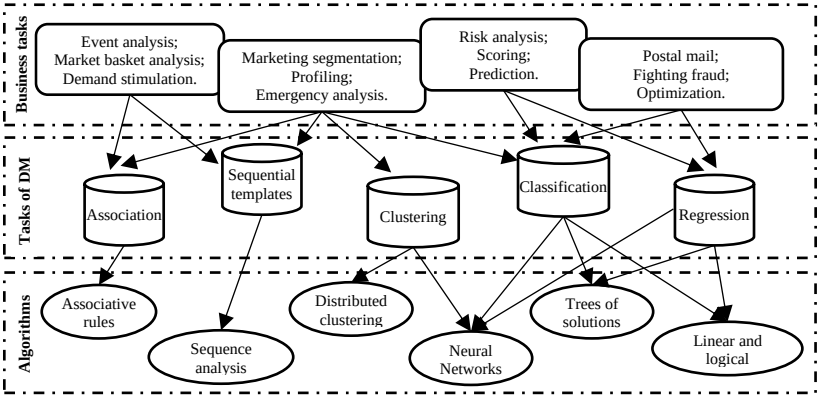


Figure 8 – Popular business tasks that are solved by Data Mining algorithms.

2. DATA CONSOLIDATION

The knowledge value and reliability, obtained as a result from the intellectual analysis of business data, depend not only on the effectiveness of the analytical methods and algorithms, which are used but also on how correctly **the initial data are selected and prepared for the analysis** (Otenko I., 2015)**.**

Usually, managers – heads in business-analysis project from the beginning have to face the following situation:

- Firstly, the data on the enterprise are in various sources and a wide variety of formats or types - in separate files of office documents (Excel, Word, plain text files), in

accounting systems (M.e.doc, etc.) in databases (Oracle, Access, dBase, etc.).

- Secondly, the data may be redundant or, conversely, insufficient.
- Thirdly, the data are "uncleaned", i.e., they contain factors that prevent their proper processing and analysis (omissions, anomalous values, duplicates and contradictions).

Therefore, before proceeding with the analysis of data, it is necessary to **perform a number of procedures,** the purpose of which is to bring the data to an acceptable quality and informality level and to organize their integrated storage in structures that ensure their integrity, consistency, high speed and flexibility of analytical queries.

Consolidation - a set of methods and procedures aimed at extracting data from different sources, providing the necessary level of their informativeness and quality, converting into a single format, in which they can be loaded into the storage and the other or analytical system.

The main tasks of data consolidation:

- selection of data sources;
- development of a consolidation strategy;
- assessment of data quality;
- enrichment;
- cleaning;
- transferring data to the repository.

First, sources, containing data that can relate to a solved task, are selected, **then** the type of sources and the methods to organize access to them are determined. In this regard, there are **three main approaches to organize storage data:**

1. **Data stored in separate (local) files**, for example, in text files with delimiters, Word documents, Excel tables, etc. Such a source can be any file, the data, which are organized in the form of columns and records. Columns must be typed, that

is, they must have one content type, for example, only text or only digital.

The advantage of such sources is that they can be created and edited with the help of simple and popular office applications, the work of which does not require special staff.

The disadvantages include the fact that they are not always optimal in terms of the access speed to them, the presentation compactness and support their structural integrity. For example, there is no guarantee that a table processor user will not place data of different types (digital and textual) in one column, which in the future necessarily leads to problems when they are processed in the analytical program (Seigel E., 2016).

2. Databases of different DBMSs, such as Oracle, SQL Server, Firebird, dBase, FoxPro, Access, etc. Database files better maintain the integrity of data structures because the type and properties of their fields are rigidly asked when constructing tables. However, in order to create and to administer a database, the company needs specialists with a higher level of training (than to work with popular office applications).

3. Specialized data storages (DS) are the most important solution since their structure and functioning are specially optimized for work with an analytical platform. Most DSs provide high-speed data exchange with analytical applications, automatically ensure the integrity and consistency of data.

Another important task to be solved within the consolidation framework is to assess the data quality in terms of its suitability for processing through various analytical algorithms and methods.

In most cases, the initial data is pretty "dirty", that is, they contain factors that do not allow them to be properly analyzed, detect hidden structures and laws, establish communication between data elements, and perform other actions that may be required to obtain an analytical solution.

These factors include:

- input errors;
- omissions;
- abnormal values;
- noises;
- contradictions, etc.

Therefore, before proceeding with data analysis, it is necessary to evaluate their quality and compliance with the **requirements of the analytical platform.** If in the quality assessment process factors that do not allow the correct application of data by certain analytical methods will be identified, appropriate **data purification is necessary.**

Data Clearing – methods and procedures set aimed at eliminating the causes of impeding correct processing: anomalies, omissions, duplicates, contradictions, noise, etc.

Another operation that may be needed when consolidating data **is its enrichment.**

Enrichment is the process of completing data with some information that lets to increase the efficiency to solve analytic tasks.

It should be used in cases where the data do not contain sufficient information for a satisfactory solution to a specific analysis task. The enrichment of data enables to increase their information saturation and, as a consequence, the significance to solve the analytical problem

Generalized consolidation process diagram.

The place for consolidation in the overall process of data analysis can be presented in the form of a structural scheme (Figure 9).

The consolidation procedure **is based on the ETL process** (extraction, transformation, loading). The ETL process solves the problem of the data extracting from different sources types, converting them to a form suitable to store in a particular structure and to load into the appropriate database or data

storage. If the analyst has doubts about the quality and informativeness of the data source, then he or she can use the procedures for assessing their quality, purification or enrichment, which are also integral parts of the **consolidation process data.**

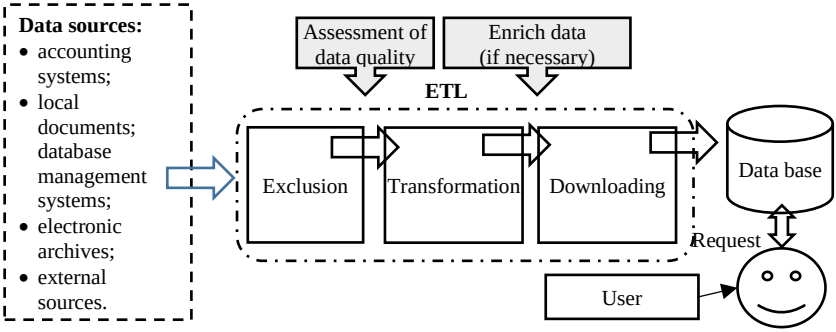


Figure 9 – The process of data consolidating

By the mid-80s. XX century, **the first stage to equip business and government structures with computer facilities** has almost completely ended and a period of rapid development for information systems has begun. The basic aim was **to organize the collection and storage of the large arrays of various business and service information.**

These were mainly **corporate systems** designed for the rapid processing of information, which served accounting, information archives, telephone networks, document registration, banking operations, etc.

When **personal computers were invented,** such systems became available to many small and medium-sized companies, enterprises and organizations (Marchenko O, 2017).

The online information processing systems are **called OLTP** (On-Line Transaction Processing - online, that is, in real time, transaction processing) (Figure 10).

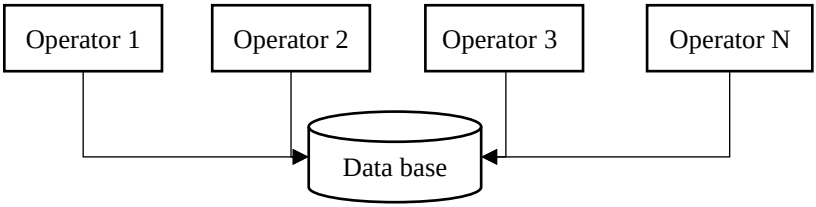


Figure 10 – The general scheme of OLTP operation

A typical example to use OLTP-systems is mass customer service, such as booking airline tickets or paying for telephone companies. Each of these situations has **two common properties:**

- an extremely large number of clients;
- continuous receipt of information.

When booking tickets from numerous points of sale, there is a continuous flow of information about tickets that have already been sold, which are introduced from their salespeople. In the same database information about vacancies is formed. From this task standpoint, the transaction includes **a set of actions such as:**

- operator's request for availability of seats for one flight or another;
- database feedback with the provision of relevant information;
- entering the information about the customer, the number of the ordered place and the paid amount (the availability of other official support information is possible);
- transfer of new information to the database and making appropriate changes to it;
- transfer to the operator’s confirmation that the operation was successful.

Such transactions are executed thousands of times a day in hundreds of sales points. Obviously, the main **priority, in this**

case, is to provide a minimum response time with the maximum boot of the system.

Consider the **characteristic features** of this process, which in one way or another are inherent in all OLTP-systems:

- Queries and reports are fully regulated. The operator cannot formulate his or her own request to clarify or to analyze any information.
- As soon as the flight is completed, customer service information is meaningless, becomes out of the question and is to be deleted after a certain time (ie historical data is not supported).
- Operations are performed over the data with the maximum level of detail, that is, for each client separately

The situation radically changes when the management of the **airline decides to study passenger traffic in order, for example, to optimize them.** Such studies may be a reaction to the information that recently there have been frequent cases of tickets shortage for certain routes at many sale points, which allows us to make an assumption about the advisability to organize additional flights (Mize Ed., 2017).

However, at **least three things are required to conduct such research:**

- Firstly, we need data on ticket sales for a long period (several months or years).
- Secondly, the data should not contain contradictions, omissions, anomalous values and other factors that will not allow the correct analysis to be performed.
- Thirdly, additional information is needed in the business environment: about competitors, market trends, fuel prices, etc.

Obviously, a typical OLTP system cannot provide any of the above. Thus, having understood these problems, there is an awareness of the need **to use more advanced analysis-oriented data storage systems.**

The basic requirement for an OLTP system is the rapid maintenance of relatively simple queries generated by a large number of users, while the waiting time for a typical request should not exceed a few seconds. Subsequently, **large volumes of data** began to accumulate in such systems – documents, information about banking operations, customer information, agreements, provided services etc.

Soon, there was an understanding that **data collection is not an end in itself.** The collected information may be very useful in the management of the organization, the search for ways to improve the activities and obtaining with this competitive advantage. But this requires systems that would perform **not only the simplest actions on the data:** counting amounts, average, maximum and minimum values.

There was a need for **information systems** that would allow for in-depth analytical processing, for which it is necessary to solve such tasks as:

- search for hidden structures and patterns in data arrays;
- search for conclusions of such patterns;
- strategic and operational planning;
- formation of unregulated queries;
- making decisions and forecasting their consequences.

Understanding the **benefits that intelligence analysis can provide** has led to the introduction of a new systems class – **decision support information systems**. That is, the information DSS, oriented to the analytical processing of data in order to obtain the knowledge necessary for the development **of management decisions** (Provost F., 2013).

Additional incentives to improve these systems are factors such as:

- reduce the cost of high-performance computers;
- reduce the cost to store large volumes of information;

- the emergence of processing large data amounts;
- development of appropriate mathematical methods.

The basis for working with such a DSS is the requests that are addressed to it by the user (the decision maker – the manager, expert or analyst).

At the same time, requests that are **permissible in traditional systems** of data processing are rather primitive. For example, for a bank, it could be a request like: "How much money does the customer save on his account?"; "How much money the client averagely spent during the month?"

Obviously, the value of information received with **such a request is insignificant.**

At the same time, the **analytical system** can respond to much more complex queries, for example: "Determine the average time between getting the invoice and organizing the payment by each category of customers".

In the process of developing **the information analysis systems and the methodology for their application**, it was found that for effective functioning, such systems should be organized in a slightly different way than that used in OLTP systems. This is due to the **following reasons:**

- It is necessary to process large data sets from various sources to perform complex analytical queries,
- To perform queries related to trend analysis, forecasting of processes that are time-consuming, historical data accumulated over a sufficiently long period are needed, which is not provided by conventional OLTP systems.
- The data, used for the analysis and maintenance of analytical queries, are different from that used in conventional OLTP systems. In analytical processing, preference is given not to detailed data, but to generalized (aggregated) data. Obviously, to analyze the sales of a large supermarket, it is not very interesting information about individual purchases, but rather

information about sales during certain time intervals (for example, a week or a month).

In this regard, it is possible to distinguish a number of fundamental differences between DSS and OLTP-systems. (Table 13).

Table 13 – Principle differences of DSS and OLTP-systems

Feature	OLTP- system	DSS
Purpose to use data	Quick search, simple processing algorithms	Analytical processing in order to search for hidden patterns, building forecasts and models, etc.
The level of generalization (detail) data	Detailed	Both detailed and generalized (aggregated)
Data quality requirements	Incorrect data may be incorrect (registration errors, input, etc.)	Errors in the data are not allowed, because they can lead to incorrect operation of analytical algorithms.
Data storage format	Data can be stored in different formats depending on the application in which it was created.	Data is stored and processed in a uniform format.
Data storage time	As a rule, no more than a year (within the reporting period)	Years, decades
Data change	Data can be added, changed and deleted.	Only replenishment is allowed; previously added data should not be changed, which ensures their chronology

Continuation of Table 13

Periodicity updates	Often, but in small amounts	Rarely, but in large volumes.
Data access	Access to all current (operational) data should be provided.	Access to historical (that is, accumulated over a sufficiently long period of time) data should be provided in accordance with their chronology.
The nature of the requests	Standard, pre-configured	Ad hoc, generated by the analyst "on the fly" depending on the required analysis
Request time	Few seconds	Up to a few minutes

Obviously, the requirements for DSS and OLTP-systems **differ significantly.**

Therefore, the DSS uses specialized **databases called data repositories or storages** (DRs or DSs).

Data storages are focused on analytical processing and meet the requirements for systems supporting decision-making.

Key features of the data storages concept

Currently, an unambiguous definition of data storage does not exist, due to the fact that a large number of different architectures and data storage technologies have been developed, and the storage facilities are used to solve a variety of wide tasks. Each author puts his or her vision of the issue into this concept. Summarizing the requirements for DSS, we can give the **following definition** of a data storage, which does not claim to be complete and unambiguous, but allows you to understand the basic idea,

Data storage – a type of storage system focused on supporting the data analysis process, ensuring integrity,

consistency and history, as well as high-speed execution of analytical queries (Seigel E., 2016).

The most important element of the data storage is the **semantic layer** – a mechanism that allows the analyst to operate with data through the business terms of the subject area. The semantic layer **gives the user the opportunity to focus** on the analysis and not think about the mechanisms for obtaining data.

Typical data storage is significantly different from conventional storage systems. The main difference is the purpose of use.

Example.

Sales registration and issuance the appropriate relevant documentation is the task for **OLTP-system** level using ordinary relational DBMS. Analyzing the dynamics for sales and demand over several years, which makes it possible to work out a strategy for the development of a company and work plan with suppliers and customers, is most conveniently done with the support of **a data storage.**

Another important difference is the dynamics of change in data. Databases in the **OLTP-systems** are characterized by a very high dynamics of changing the records due to the daily work generated by a large number of users (where, by the way, there is a high probability for the occurrence of contradictions, errors, data integrity problems, etc.). As for the **data storage**, the data from it is not deleted, and the replenishment occurs in accordance with a specific regulation (once an hour, day, week, at a certain time).

Basic requirements for data storage.

In order for a data storage to perform functions, which are appropriate to its main task – **to support the data analysis process** – it must meet the requirements formulated by R. Kimball, one of the authors for the data storage concept:

- the high speed of receiving data from the storage;
- automatic support for internal data consistency;

- the ability to obtain and compare data slices;
- availability of convenient tools to view data in the repository;
- ensuring the integrity and reliability of the stored data.

In order to meet all the listed requirements, to build and to run DS, as a rule, **not just one application is used, but a system** that includes several software products. Some of them constitute the actual data storage system, others are means of viewing, retrieving, loading etc.

In recent decades, DS technology has developed rapidly. **Dozens of companies offer their DS solutions on the market, and thousands of organizations are already using this powerful tool to support analytical projects.**

It is believed that the DS concept was firstly investigated by the technical director in Prism Solutions Bill Inmona, who in the early 1990s published a number of works that became fundamental studies to further research in the analytical systems field.

The basis of the DS concept is the **following provisions:**

- Integration and reconciliation of data from various sources, such as conventional systems of operational processing, databases, accounting systems, office documents, electronic archives, located both inside the enterprise and in the external environment;
- The division of data sets used by transaction execution systems and DSSs.

Inmona gave such **a definition for DS** – subject-oriented, integrated, unchanged and supporting a chronology data set, designed to provide managerial decisions.

Under objective orientation, in this case, means that the DS should be developed taking into account the specifics of the subject area but not the analytical applications with which it is intended to be used. **The DS structure** should reflect the views

of the analyst on the information according to which he has to work (Mishenina N., 2014).

Integration means that it should be possible to download information from sources that support various data formats and created in different applications:

- accounting systems;
- databases;
- spreadsheets and other office applications that support data structuring (for example, delimited text files).

The immutability principle implies that, unlike the usual systems of operative data processing, the data after download should not be subject to any changes, except for the new data addition.

Chronology support means keeping track of the records order, for which the key attributes of Date and Time are entered into the DS structure. In addition, if you physically arrange the records in chronological order, for example, in the order of Date attribute growth, you can reduce the time the analytic queries run.

Using the DS concept in DSS and data analysis helps to achieve **such goals as:**

- timely provision of analysts and managers with all the information necessary for the reasonable and high-quality managerial decisions development;
- creation of the unified model of data representation in the organization;
- creation of an integrated data source that provides convenient access to various types of information and ensures that identical responses are obtained for the same requests from various analytical applications.

Tasks, which are solved by DS.

The process of DS development is very laborious, some organizations spend several months and even years on it, as well as investing significant financial resources.

The main tasks that need to be addressed during the DS development are:

- selection of a storage structure that provides high-speed queries execution and minimization of the RAM amount;
- initial filling and subsequent replenishment of the storage;
- providing a unified method for working with heterogeneous data and creating a user-friendly interface.

The range of tasks for intellectual data processing is very wide, and the tasks themselves vary significantly in terms of complexity.

The data is extracted from various sources and loaded into the DS, which contains both actual data presented in accordance with a model and metadata.

Detailed and aggregated data.

Data in DS is stored both in **detailed and in aggregated form**.

The data in a detailed form comes directly from data sources and corresponds to elementary events registered by OLTP systems.

Such data can be daily sales, the number of products produced, etc. **These are indivisible values, an attempt to further detail which deprives them of their logical meaning.**

Many analysis tasks (for example, forecasting) **require the use of data with a certain generalization degree.**

For example, sales by day can provide a very uneven series of data, making it difficult to identify characteristic periods, patterns, or trends. However, if we summarize these data within a week or a month and take the sum, average, maximum and minimum values for the corresponding period, the resulting series may be more informative.

The process of summarizing the detailed data is called aggregation, and the largest data themselves are called aggregated (sometimes aggregates). Typically, the aggregation

is subjected to numerical data (facts), they are calculated and contained in DS along with detailed data.

The term "metadata" (from the Greek - Meta and Latin - Data) literally translates as "data about the data." Metadata in a broad sense is necessary to describe the meaning and properties of information in order to understand, use and manage it better. Anyone who read books or used a library, dealt with metadata to some degree (Seigel E., 2016).

It is well known that in any book, besides the text, there is a significant amount of additional information. Its purpose is:

- firstly, to help the reader to quickly familiarize with the contents of the book and to comprehend it,
- secondly, to describe the structure of the book for the more efficient search of the necessary information.

There are such elements as annotation, comments, glossary, notes, in. To solve the first problem, To search for the necessary information, a table of contents, titles of chapters, paragraphs and sections, page numbers, footers, subject indexes, etc. are used.

In addition, the reader may need information about the authors or the publisher.

All this information, which is not part of the book, but serves to increase the efficiency of working with it, is metadata. In the library, metadata are used to search for publications and track their movements, for example, systematic or alphabetical catalogues that use book names, authors' names, year of publication, etc. Thus, metadata have great importance when working with various types of information.

From the IT technologies standpoint, metadata include any information, which are necessary for the analysis, design, construction, implementation and application of a computer information system. One of the main purposes in metadata case is to increase search efficiency. Search queries that use metadata

make it possible to perform complex filtering operations and data selection.

If we consider the "metadata" concept in the context of DS technology, then it can be defined as follows.

Metadata include a high-level means to reflect the information model and to describe the data structure that is used in DS. Metadata should contain a description of the data structure in the warehouse and data structures within the imported sources. Metadata are stored separately in the so-called metadata repository.

A bright example of metadata in social networks is **the hashtag.**

Metadata is a key factor in the success of DS development and implementation. They contain all the information necessary to obtain, convert and download data from various sources, as well as for further use and interpretation of the data contained in the DS.

There are **two levels of metadata:**

- technical (administrative). The technical level contains the metadata needed to maintain the vault function (data download and usage statistics, data model description, etc.);
- business metadata provide to the user the opportunity to focus on the analysis process, not on the technical aspects of working with the repository; they include business terms and definitions that are used to operating the user.

In fact, **business metadata is a description of the subject area** for which an analytic system or DS is created. Experts and analysts should be actively involved in the formation of business metadata, which will subsequently use the system to receive analytical reports.

Business metadata describe the objects of the subject area, the information contained in the DS, – the attributes of the objects and their possible values, the corresponding fields in the tables, etc. Business metadata form **the so-called semantic**

layer. The user operates his or her domain terms close to him: product, client, sales, purchases, etc. The semantic layer translates business terms into low-level requests for data in the storage.

Development and construction **of corporate DS** is an expensive and time-consuming task. The success of the DS implementation **largely depends on the:**

- business processes' informatization level in a company;
- established information flows;
- volume and structure of the data used;
- requirements for the speed of queries;
- the frequency to update the storage;
- nature of the analytical tasks being solved, etc.

Several storage architectures are currently being developed – relational, multidimensional, hybrid, and virtual in order t**o bring the DS to the conditions and to the specifics of a particular organization**.

- **Relational DSs** use the classical relational model, which is peculiar for operational recording OLTP systems. The data are stored in relational tables, but form special structures that emulate a multidimensional representation of the data. This technology is abbreviated as ROLAP - Relational OLAP (relational online analytical processing).
- **Multidimensional DSs** implement multidimensional data representation at the physical level in the form of multidimensional cubes. This technology is called MOLAP - Multidimensional OLAP.
- **Hybrid DSs** combine the properties of both relational and multidimensional data models. In hybrid DS, detailed data are stored in relational tables, and aggregate in multidimensional cubes. Such technology for building DS is called HOLAP - Hybrid OLAP.

- **Virtual DSs** are not data storages in the usual sense. In such systems, work is carried out with **individual data sources**, but it also emulates the work of a regular DS. In other words, the data are not physically consolidated but are collected directly in the course of the query execution.

In addition, all DSs can be divided into **single-platform and cross-platform**. Single-platform DSs are built on the basis of only one DBMS, while **cross-platform** DSs can be built on the basis of several DBMS.

The main purpose of multidimensional data storages (MOLAP) is to support systems that focus on analytical data processing, since such storages do a better job of executing complex non-reglamentated queries.

The multidimensional data model underlying the multidimensional data storages construction is based on the concept of **multidimensional cubes, or hypercubes**. They are ordered multidimensional arrays, which are also often called OLAP-cubes (abbreviation OLAP stands for On-Line Analytical Processing).

The essence of multidimensional data representation is as follows. Most real business processes are described **by a variety of indicators, properties, attributes, etc.**

For example, to describe the sales process, **you need information about** the names of the goods or their groups, the supplier and the buyer, the city where the sales were made, as well as prices, quantities of goods sold and total amounts.

In addition, to track the process over time, an attribute **such as a date should be introduced.**

If you collect all this information in a table, it will be **difficult for visual analysis and reflection.** Moreover, it may turn out to be redundant: if, for example, the same product was sold on the same day in different cities, then it would be necessary to repeat the same «city-product» correspondence several times with different amounts and quantities. All this can

finally confuse anyone who tries to **extract useful information from such a table** in order to analyze the current state of sales and find ways to optimize the trading process.

These problems arise for one simple reason: multidimensional data are stored in a flat table.

Measurements and facts - basic concepts of the multidimensional data model.

The basis of multidimensional data presentation is their division into two groups - **measurements and facts.**

Measurements are categorical attributes, names and properties of objects involved in a business process. Measurement values are the names of goods, the supplier and buyers companies names, the people names, cities names, etc. **Measurements can be numeric** if a category (for example, the name of a product) corresponds to a numeric code, but in any case, it is **discrete data**, that is, taking values from a limited set. **The measurements qualitatively describe the business process being investigated.**

Facts are data that quantitatively describe a business process, continuous in nature, that is, they can take on an infinite number of values. Examples of facts are the products price, their quantity, the sales or purchases amount, the employee's salary, the loan amount, insurance compensation, etc.

Multidimensional Cube Structure.

A multidimensional cube can be considered **as a coordinate system** axes of which are measurements, for example, **Date, Product, Buyer**. Along the axes, measurement values will be deferred – dates, goods names, purchasers names, individuals names, etc.

In such a system, each set of **measurement values** (for example, "date-product-customer") will correspond to a cell in which can be placed the numerical indicators (that is, **facts**) associated with this set. Thus, an unambiguous connection will

be established between the objects of the business process and their numerical characteristics.

The principle of a multidimensional cube organization is illustrated in Figure 11.

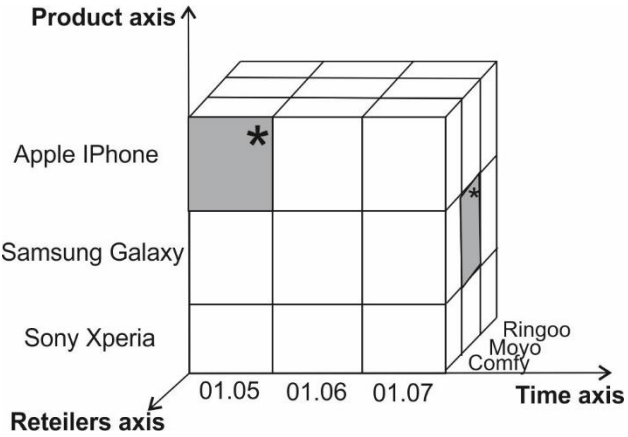


Figure 11 – The principle to organize a multidimensional cube

In this case, here is a multidimensional view of the **Date, Product and Buyer dimensions**. Facts, in this case, may be Price, Quantity, Amount. Then the selected segments will contain information about how many goods, for what amount and at what price the specified firms purchased.

Thus, the information in the multidimensional **data storage is logically coherent.** These are not just sets of string and numeric values, which in the case of a relational model need to be obtained from different tables, but whole structures like "who were the consumers; what were sold; what quantity of good were sold at a specific point in time".

The benefits of a multi-dimensional approach:

- Data presentation in the multidimensional cubes form is more obvious than the totality of normalized tables in the

relational model, which structure is understood only by the DB administrator.

- The possibilities of building analytical queries to a system using MDS are wider.
- In some cases, the multidimensional model use can significantly reduce the duration of the search in MDS, ensuring the execution of analytical queries **almost in real time.** This is due to the fact that the aggregated data is calculated in advance and stored in multidimensional cubes along with detailed ones, so you do not need to waste time on calculating aggregates when the query is executed (Mize Ed., 2017).

Disadvantages of this approach:

- a large amount of memory is required for its implementation.
- the multidimensional structure is more difficult to modify; if necessary, to integrate one more dimension, you must perform a physical rebuilding of a multidimensional cube.

Based on this, we can conclude that the use of storage systems based on data multidimensional presentation, is **advisable only in cases when the amount of the used data is relatively small, and the multidimensional model has stable measurements set.**

In the early 1970s thanks to simplicity and flexibility, the relational model became dominant, and relational DBMS became the de facto industrial standard.

Relational database is a set of relations containing all the information that should be stored in the database. Physically, it is expressed in the way that information is stored in the form of two-dimensional tables connected with key fields.

The use of the relational model in creating DSs in some cases makes it possible to gain **advantages over multidimensional technology**, especially in terms of the efficiency for work with large data arrays and the computer memory usage.

The RDS technology is based on the principle that **measurements are stored in flat tables in the same way as in conventional relational DBMS, and the facts (aggregated data) are stored in separate special tables of the same database**.

In this case, the **facts table is the basis for its associated dimension tables.** It contains quantitative features of objects and events, totality of which is supposed to be analyzed further (Knafli C.-N., 2015).

RDS construction schemes.

At the logical level, there are two schemes to construct RDS – **the "star" and "snowflake".**

When using **the star scheme**, the facts table is central, so with it, all dimension tables are associated. Thus, information about each dimension is located in a separate table, which simplifies their viewing, and the scheme itself makes it logically transparent and understandable to the user (Figure 12).

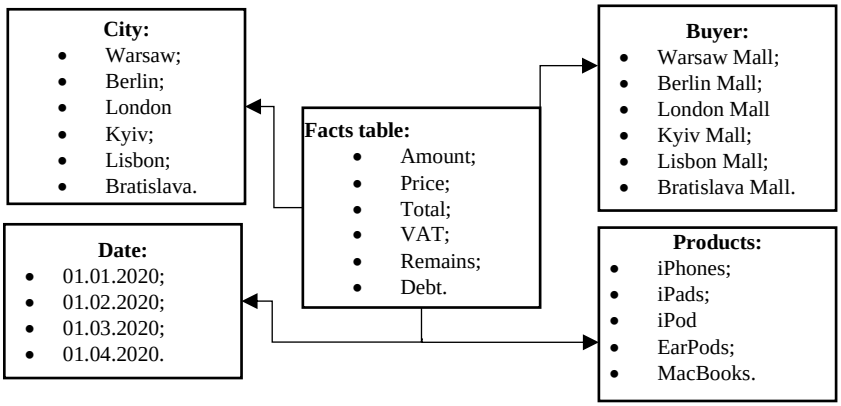


Figure 12 – An example of constructing the RDS by the "star" scheme

However, the placement of all information about the measurement in one table **is not always justified.**

For example, if the sold goods are grouped (hierarchy takes place),

it is necessary to show in one or another way which group each product belongs to, **it will lead to multiple repetitions of group names.** This will not only intensify in redundancy but also increase the probability of contradictions (if, for example, the same product is mistakenly assigned to different groups).

Therefore, **the "snowflake"** as a modification of the star scheme was developed for more effective work with hierarchical measurements.

The main feature of the snowflake scheme is that information about one dimension can be stored in several related tables. That is if at least one of the dimension tables has one or several other dimension tables associated with it, in this case, the "snowball" scheme will be applied (Figure 13).

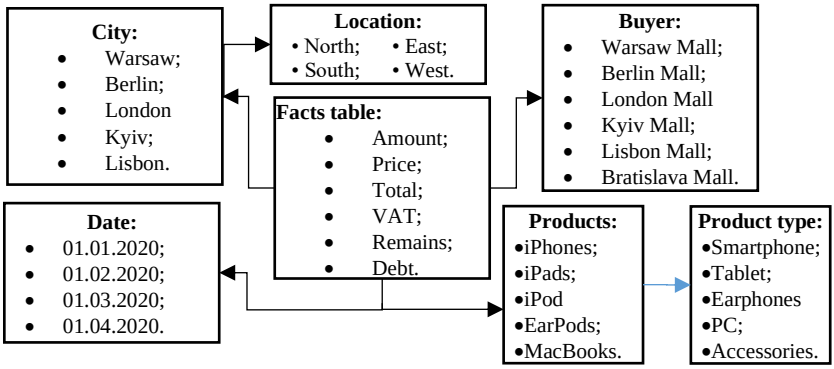


Figure 13 – An example of constructing the RDS by the "snowflake" scheme

The main functional difference between the snowflake scheme and the star scheme is the **ability to work with hierarchical levels** that determine the degree of data detail. In the above tables, the snowflake scheme lets to work with data at the maximum level of detail, for example, with each product

separately, or use a generalized view of product groups **with appropriate facts aggregation.**

The choice of the constructing RDS scheme depends on the data collection and processing mechanisms. Each of the schemes has its advantages and disadvantages, which, however, can manifest themselves to a greater or lesser degree depending **on the characteristics by which the DS is functioning as a whole** (Knafli C.-N., 2015).

Advantages and disadvantages of RDS

The main advantages of RDS are:

- almost unlimited amount of stored data;
- since relational DBMSs underlie the construction of many operational processing systems (OLTP), which are usually the main data sources for DS, the relational model use simplifies the process of loading and integrating data into the repository;
- when adding new data dimensions, there is no need to perform a complex physical reorganization of the repository, unlike, for example, from multidimensional DSs;
- provides a high level of data protection and a wide range for access rights.

The main disadvantage within relational data storage is that when using a high level of data aggregation and measurement hierarchy in such data stores, **aggregate tables begin to multiply.** As a result, the query execution speed of the relational storage **slows down.**

At the same time, in multidimensional repositories, where data are stored in the form of multidimensional cubes, **this problem practically does not arise** and in most cases, it is possible to achieve a higher speed of query execution.

Thus, the choice of the relational model in constructing DS is appropriate **in the following cases:**

- The amount of stored data is significant (multidimensional DSs become ineffective).

- The measurement hierarchy is simple (in other words, some aggregated data).
- Requires frequent changes in data dimension. When using the relational model, you can limit yourself to adding new tables, and for the multidimensional model, you will have to perform a complex restructuring of the physical structure of the repository.

It is obvious that multidimensional and relational data storages models have **advantages and disadvantages.** For example, a multidimensional model enables to get an answer to a query relatively quickly but does not guarantee efficient managing huge data amount as a relational model does.

It would be logical to use such a DS model, which **would be a combination** of relational and multidimensional models and would allow combining high-performance feature of a multidimensional model, and the ability to store arbitrarily large data arrays inherent in the relational model. A model that **combines the principles** of relational and multidimensional models, called a hybrid, or HOLAP (Hybrid OLAP).

The main principle of HDS construction is that detailed data are stored in a relational structure (ROLAP), which allows storing large data amounts, and aggregated data is stored in multidimensional (MOLAP) that provides to increase the speed of query execution (since analytical queries do not require the calculation of aggregates) (Figure 14).

Supermarket serves thousands of customers every day. In this case, the maximum detailing level for the registered data is the purchase by one check, which indicates the total purchase amount, the name or code of the goods and the value of each product.

Operational information, which includes the detailed data, is consolidated in the relational structure of the DS.

From the standpoint of analysis, general data is more interesting, for example, in product groups, departments, or

some time intervals. Therefore, the original detailed data is aggregated and the calculated aggregates are stored in the multidimensional structure of the hybrid DS (Seigel E., 2016).

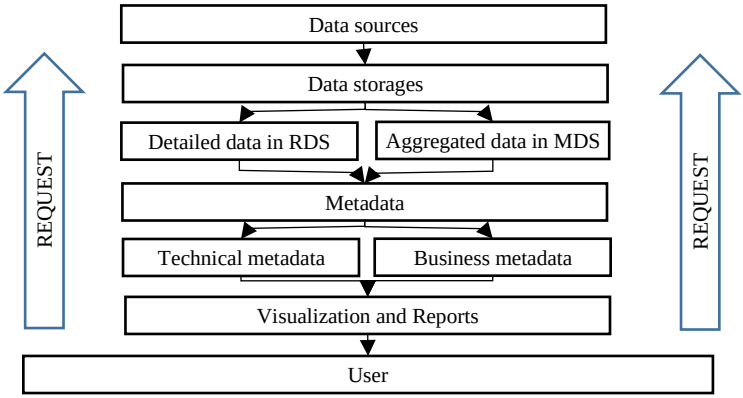


Figure 14 – The principle of building a hybrid data storages

If the data coming from the OLTP system has a large amount (a few thousands of records per day or more) and a high degree of detail and mainly generalized data are used for the analysis, the hybrid storage architecture turns out to be the most suitable.

The disadvantage of the hybrid model is the difficulty of administering the DS through the more complex rules for its replenishment since it is necessary to coordinate the changes both in the relational and in the multidimensional structures. However, there are a number of advantages that make the hybrid DS model pretty attractive:

- Storing data in a relational structure makes it more systemically independent, which is especially important when using economic information (indicators) in the management of an enterprise.
- The relational structure forms stable and consistent anchor points for multidimensional storage.

- Since relational storage maintains the relevance and correctness of data, it provides a very reliable transport layer for delivering information to multidimensional storage.

With all the positive aspects of the data store as a separate consolidated source, there are **situations where this idea does not work.** In fact, the night load of data, collected during the day from the OLTP-systems in the DS is common practice.

Such a regulation **allows to reduce the load on the OLTP system** during the working day, that is, during its active use. **Supplement the DS more than once a day does not make sense.** However, this state of affairs does not provide the opportunity t**o analyze information during the working day** as it arrives. Sometimes it is extremely critical and in such a situation, it is advisable to search for an alternative to the traditional (physical) DS.

In addition, the inevitable problem with the DS use in business **analytics is redundancy.** It reduces the efficiency of using disk space and server memory, and with very high volumes of stored and processed information, it can cause a decrease in performance, an increase in waiting time for response to a request and even result in a complete failure of the system. **Redundancy is, to varying degrees, characteristic of both relational and multidimensional repositories.**

The situation is aggravated by the fact that DSs store historical information and implement the **principle of data invariability**. That is, unlike conventional operational processing systems (OLTP systems), **where only actual data are stored**, and the data that have lost relevance are deleted, the DS can **only be updated with new data**, and historical data are not deleted. In addition, it is often necessary to store large amounts of aggregated data. Taken together, these factors can lead to an “explosive” increase in the volume of DS (Mize Ed., 2017).

The concept **of virtual data storage (VDS) lets t**o overcome the above problems. It is based on the principle that the data from local sources, the external environment, databases and accounting systems are not **physically consolidated into a single DS**, but retrieved, converted and integrated directly when the **request is executed in the computer's RAM.** In fact, requests are addressed directly to data sources.

Requests for VDS are always carried out using the **semantic layer.**

A virtual data repository is a system that works with disparate data sources and emulates the work of the physical data storage by unloading, converting and integrating data directly in the query execution process.

The advantages of this approach are obvious.

- There is an opportunity to analyze data in the OLTP system immediately after they arrive without waiting for the download to the storage.
- The volume of the required disk and RAM is minimized since there is no need to store historical data and numerous aggregated data for different levels of information aggregation .
- The presence of a developed semantic layer in the VDS allows the analyst to completely ignore the problems associated with the process to extract data from various sources and focus on solving data analysis problems.

When working with the VDS, the user, in fact, deals with the **"illusion - immitation of the data storage"**. **Virtuality assumes** that the VDS exists only as long as the corresponding application is running. As soon as it quits, the virtual storage ceases to exist.

The concept of the VDS **has a number of disadvantages** compared with the DS, where information is consolidated physically:

- The load on the OLTP system is increasing, because, in addition to regular users, analysts with ad hoc queries turn to it. As a result, the performance of the OLTP system decreases.
- data sources from which information is requested in the VDS may be inaccessible if they are accessed over the network or if their localization has been changed. Temporary unavailability of at least one source may cause the impossibility to fulfil the request or the distortion of the information provided on it.
- There is no automatic support for data integrity and consistency, some fragments of documents may be lost, etc.
- Data in sources are stored in various formats and encodings, which can lead to errors in their processing and distort information received in response to a request.
- Due to the possible inconsistency between the points of data replenishment sources and due to the lack of support in their chronology for the same query, different data can be obtained at different points in time.
- It is practically impossible to work with data accumulated over a long period of time since only those data that are in the sources at a particular point in time are available in the VDS.

The most important VDS feature is that they deal with data within **a certain period of relevance**, working directly with sources that contain operational data. This is because OLTP systems do not store historical data. Therefore, if historical data play an important role in the analysis, then it is preferable to use varieties of DS with physical data consolidation. Thus, the VDS should be used in systems focused on the analysis of operational information that is relevant only for a limited period.

Thus, the use of VDS is useful for enterprises that do not have **the technical means and qualified personnel to support physical DSs.**

ETL is a set of methods which implement the process to transfer initial data from various sources to an analytic application or data storage that supports it.

The main goals and objectives of the ETL process.

ETL applications extract information from one or more sources, convert it to a format supported by the storage and processing system that is the recipient of the data, and then load the converted information into it.

Initially, ETL systems were used to transfer information from earlier versions of various information systems to new ones. Currently, ETL-systems are increasingly used precisely to consolidate data for the purpose of their further analysis. Obviously, since DSs can be built on the basis of various data models (multidimensional, relational, hybrid), the ETL process should be developed taking into account **all the features** used in the DS model.

In addition, it is desirable that the **ETL system should be universal**, that is, it can extract and transfer data of as many types and formats as possible.

Regardless of the peculiarities of construction and functioning, the ETL system should ensure the implementation for the three main stages of the data transfer process (ETL process):

- **Extract data.** At this stage, data are extracted from one or more sources and prepared for conversion. It should be noted that for the correct representation of data after they are downloaded to the DS, from the sources should be removed not only the data but also information describing their structure from which the metadata for the repository will be formed.
- **Transform data.** Formats conversion and data encoding, as well as their generalization and cleaning, are performed.
- **Data Download** – copying the converted data to the appropriate storage system.

Data movement in the ETL process can be divided into a sequence of procedures:

1. **Extract.** The data is extracted from the sources and loaded into the intermediate area.

2. **Search for errors.** The data is checked for compliance with the specifications and the possibility of subsequent loading into DS.

3. **Transformation.** The data is grouped and converted to the form corresponding to the structure of DS.

4. **Distribution.** The data are divided into several streams in accordance with the peculiarities for the organization of their loading process into DS.

5. **Insert.** Data is uploaded to the repository.

Thus, from the standpoint of the ETL process, the DS architecture can be represented as **three components** such as:

- **data source** - contains structured data in the form of a separate table, a set of tables or just a file, in which data, for example, is arranged in typed columns separated from each other by certain separator characters;
- **intermediate area** - contains auxiliary tables, created temporarily and exclusively for the organization of the unloading process;
- **recipient of data** – DS or just a database in which the extracted data should be placed.

The movement of data from source to destination is called **a data flow.** Requirements for organizing data flows are described by the analyst.

ETL is often reviewed as a simple subsystem for transferring data from various sources to a centralized repository. As for the data storages, strictly speaking, they are not related to the solution of any particular analytical task. **The task of the storage** is to provide reliable and fast access to data, to maintain their chronology, integrity and consistency.

Therefore, at first glance, **ETL is separated from the actual data analysis.** However, such a look at ETL will not provide the benefits that can be obtained if it is considered as an integral part of the analytical process.

An experienced analyst knows the features and nature of data circulating at different levels in the corporate information system, the frequency and update, the "pollution" degree, the level of their significance, etc. This enables him to use a combination of different methods for data converting into ETL to achieve a sufficient level of aggregation and data quality (Kovtun N., 2015).

Data Aggregation in ETL.

As a rule, operational data processing systems (OLTP systems), accounting systems, files of various DBMS, local files of individual users, etc. act as data sources for storages. The common feature of all these sources is that they **contain data with a maximum degree of detail** – information about daily sales or even about each sale fact separately, about servicing each client, etc.

It is commonly believed that such a detailed reproduction of events in the investigated business process is only in the favour, since the data are never superfluous and the more they are collected, the more accurate the results of the analysis will be.

In practice, this is not entirely true. Elementary events that make up a business process, such as serving a single client, executing one order, etc., also **called atomically** (that is, indivisible), by their nature, are random variables that are prone to the influence of **many different random factors** – from weather to customer' mood. Therefore, information about each individual event in the business process is virtually worthless. Indeed, based on sales information **in one day**, one can not conclude on all the peculiarities of trade. Similarly, it is not

possible to develop a **customer strategy** based on the behaviour of one client.

In other words, for a reliable description the subject area, the **use the maximum detailed data is not always appropriate,** so the most interesting for analysis are data aggregated for a certain time interval, for a group of customers, goods, etc. Such large numbers are **called aggregated** (sometimes aggregates), and the process of their calculation is aggregation.

As an aggregation result, a large number of records about each event in the business process is replaced by a relatively small number of records containing aggregated values.

Example.

Instead of the information about each from the 365 daily sales per year, as a result of aggregation, 52 records will be stored with a summary by week, 12 by month or 1 by year.

If the purpose of the analysis is to develop a sales forecast, then for a short-term operational forecast, it is sufficient to use data by week, and for a long-term strategic forecast, by months or even years.

In fact, when aggregating, **several records are combined into one with the calculation of an aggregated value based on the values for each record.**

When calculating aggregates, **several methods** can be used (Figure 15).

Among them:

- **Average** – for data located within the interval in which they are generalized, the average value is calculated. Then one containing their average value replaces all records from this interval.
- **Amount** – aggregated records are replaced by one, which indicates the sum of aggregated values.
- **Maximum** – the resultant record contains the maximum value of all united ones.

- **Minimum** – the resultant record contains the minimum value of all united ones.
- **The number of unique values** – the aggregation result is the number of unique values that appear in the cells from the same fields. Therefore, for a field containing information about the client's profession, this aggregation method will show how many times a particular profession appeared on the list. **For example,** if in the 25 entries in the field the occupation was the value of the System Analyst, and in 50 - the Manager, then as a result of aggregation, we get to number 2.
- **Quantity** – the aggregation result is the number of entries contained in the field. In the above example (the client's occupation) with this aggregation variant, we get 75.
- **Median** - calculates the median of aggregated values.

Date	Price	Quantity	Amount
01.01.2019	120	57	6840
01.02.2019	119	50	5950
01.03.2019	125	67	8375
01.04.2019	124	70	8680
01.05.2019	130	65	8450
01.06.2019	128	70	8960
01.07.2019	136	45	6120
01.08.2019	136	61	8296
01.09.2019	135	69	9315
01.10.2019	138	65	8970
01.11.2019	145	62	8990
01.12.2019	142	56	7952

Date	Average price	Average quantity	Max price	Min price	Median by amount
01.01-01.12	131.50	61.41667	9315	5950	8412.5

Figure 15 – An example of using the basic methods for calculating aggregates

The median represents the ordinal statistics, which is calculated as follows. An aggregated values set, such as sales by days of the week, is sorted in ascending order. **Then the median will be** the central element of the ordered set if this set contains an **odd values number** or the average of two central elements **if the elements number is even.**

For example, let the sales for every day during the week were {100, 120, 115, 119, 107, 131, 102}. Then, to determine the median, you need to build these values for growth: {100, 102, 107, 115, 119, 120, 131}. The value of the central element of the resulting sequence, **that is, 115, will be median.** If the sale was carried out only 6 days a week (Sunday is a weekend), then a sequence of even numbers {100, 107, 115, 119, 120, 131} will be obtained. In this case, the median will be (115+119)/2=117.

The logical question is: **should all data be aggregated indiscriminately across all possible levels of generalization or should this be carefully considered?** To answer, it is necessary to study the most probable directions of using data in the DS. However, if the storage is at the stage of development and implementation and the method for its use is not yet fully developed, then it is difficult to do so.

From all possible aggregation options, the analyst should choose the most significant ones from the standpoint of the planned areas for making analysis, and discard the others. Obviously, the analyst can opt out the aggregates that have a small number of subordinate aggregated values (for example, aggregating monthly sales per quarter), since they are easy to calculate in the analysis process. Alternatively, on the contrary, the analyst can opt out the aggregates with a maximum degree of detail (for example, aggregating daily sales). The second option is most preferable, since at first glance it offers significant savings since the number of goods and sellers multiplies the number of days per year. However, if sales data is sparse, that

is, not every product is sold daily, then the savings can be quite insignificant.

The choice of the necessary aggregates is always determined by the peculiarities of the business. It should be remembered that the units required for analysis can also be calculated directly during the execution of an analytical query to the DS, although thereby the time for its execution will increase slightly. Such an approach allows, for example, abandoning the aggregation of rarely used data (Provost F., 2013).

Thus, choosing the right data aggregation strategy in the ELT is a complex and controversial task. The increase in the aggregates number in the DS leads to an increase in its size and complexity of the data structure. Reducing the aggregates number in the DS can lead to the need to calculate them in the process of performing analytical queries, which will increase the waiting time for the user. Therefore, it is necessary to provide a reasonable compromise between these factors. **There is also a simpler rule defining an aggregation strategy:** create only those units that are very likely to be needed when analyzing data.

3 TRANSFORMATION, VISUALIZATION, CLEARING AND DATA PROCESSING

In addition to preprocessing, the purpose of which is to bring data into conformity with certain quality criteria, the process of preparing data for analysis usually involves another step - **transformation.**

Each sample of the initial data that is loaded into the analytical application is characterized by some properties set that can affect the models' efficiency and reduce the reliability of the analysis results. Even if the data is cleared of such factors that impair their quality, such as duplicates, contradictions, noises, anomalous values, omissions, etc., they may still do not correspond to the methods and purposes of the analysis. This is

due not to the content of the data, but to their presentation and internal organization (Seigel E., 2016).

A paradox may arise: the data is completely correct from the quality standpoint, and their information content is quite enough to solve an analytical problem, but the presentation and organization of data makes it difficult to analyze or even make it impossible.

Data can be separated, not ordered, presented in formats with which one or another algorithm does not work. Data transformation, that is, their conversion to a specific representation, format, or view, optimal from the position of the problem being solved, is designed to solve this task.

Data cleansing deals only with their content, the transformation is the process to optimize their presentation and organization from the standpoint of a particular analysis method. Data transformation depends on the tasks, algorithms and objectives of the analysis. That is, for one task, some types of transformation may be needed, and for another, others.

The inconsistency between data presentation and organization with the methodology and objectives of the analysis can manifest itself both at the model level and at the individual analysis tasks level.

The inconsistency at the model level means that even if the data in their current presentation is suitable for use in the neural network, it is quite possible that their use with other Data Mining models, such as decision trees, will be difficult or impossible.

With regard to the level of tasks,a to predict, it is enough to have one-dimensional observations series for the predicted parameter of the process or object under study, while solving the **classification problem,** requires a sample containing several attributes of the objects being classified. **If** the output data must be numeric **in order to solve the regression problem**, then in the **classification** – categorical. At

the same time, **all three Data Mining tasks:** classification, regression and forecasting – can be solved within the framework of one model (for example, neural network).

Data transformation is a very broad concept that **has no well-defined boundaries.** In various areas of data processing, the term "transformation" is sometimes extended to any manipulation with data, regardless of their goals and methods. However, in the context of **business analytics** data transformation has very specific goals and objectives, and uses a stable set of methods.

Data transformation is a set of methods and algorithms aimed at optimizing the presentation and format of data in terms of tasks and analysis goals.

Data transformation **is not intended to change** the information content of the data. **Its task is to present** this information in such a way that it can be used most effectively.

Data transformation is not necessarily associated **only with an analytical application** – in one or another form it is performed at all stages of the analytical process:

- in recording systems (OLTP systems);
- in the process of transferring and loading data into the DS (ETL process);
- directly during data preparation to the analysis in the analytical application.

This distribution of the transformation process is due to the fact that at each stage **it has different goals.**

In the systems of **primary data registration (OLTP-systems)** transformation can be performed to ensure the data technical and logical compatibility, its preparation for the extraction, transfer to the data storage, etc.

For example, addresses are often entered in one line. At the same time, an address individual components that are in text format (street, city) and numerical (house number, office) can be of interest for analysis. With the help of transformation, you can

distribute the corresponding elements into separate fields and convert them to the desired format.

Another example. Different registration systems (barcode scanner, cash register, etc.) can produce the same type of data in different formats. So, in one case, a dot can be used as a separator for the integer and fractional parts of numbers, and in the other cases - a comma. Attempting to insert such data into one field of the database can lead to serious technical problems, **and the transformation,** that is, conversion to a single format, will help to resolve this issue.

The main goals of data transformation at the ETT process are to bring them in line with the data model used in the repository, the implementation of correct consolidation and the actual loading into the repository.

The basic data transformation methods are (Mishenina N., 2014)**:**

1. **Ordered data transformation.** It allows optimizing the presentation of such data in order to provide further analysis, for example, solving the task of forecasting a time series or grouping over a time period.
2. **Quantization.** Allows splitting the range for possible values of a numeric attribute into a specified intervals quantity and assigning the intervals with numbers or other labels to the values that fall into them.
3. **Sorting.** Allows changing the order of the records at the original data sample in accordance with the algorithm defined by the user. In some cases, sorting allows simplifying the visual analysis for the sample, quickly determine the largest and smallest values of attributes, etc.
4. **Merge**. Allows to combine two tables by fields of the same name or to supplement one table with records from another that are missing in the complemented table. A merge is used when the information in the analyzed data sample needs to be supplemented with information from another sample. When

combined, all other records are added to the original selection records. In the case of additions to the original sample, only those data that were missing in the original are added. **The merge operation is one of the ways to enrich the data:** if the sample does not contain enough data for analysis, then it can be supplemented with missing information from another sample.

5. **Grouping and ungrouping.** Very often, the information of interest to the analyst in the table is "diluted" with extraneous data, fragmented, scattered across individual fields and records. **Using grouping,** the analyst can summarize the necessary information, combine it into the minimum required number of fields and values. Usually, provide the ability to perform the reverse operations – ungrouping.

6. **Set up a dataset.** Allows changing the names, types, labels and destination fields of the original data sample. For example, if a field containing numeric information in a data source has a date type for some reason, the values of this field cannot be processed as numbers.

7. **Tabular substitution of values.** Allows the replacement of values in the original data sample based on the so-called lookup table. The lookup table contains "original value - new value" pairs. Each value in the data sample is checked for compliance with the original value of the lookup table, and if such a match is found, then the value of the choice is changed to the corresponding new value from the lookup table. This is a very convenient way to automatically adjusting the values.

8. **Calculated values.** Sometimes analysis requires information that is not explicitly available in the data source, but can be obtained from calculations over existing values. For example, if the price and quantity of the goods are known, the amount can be calculated as their derivative. For these purposes, a calculator is included in the analytical application, which allows to perform various calculations on the data of the initial sample. Since the analyzed data can be in different types (string,

numeric, date/time, logical), the calculation mechanism must support the work not only with numeric data, but also with other data types.

9. **Normalization.** Normalization allows to convert a range of changes for numeric feature to another range, more convenient to apply to the data from various analytical algorithms, as well as to reconcile the changing ranges of various signs. Frequently used is the reduction to one, when the entire available range of data is "compressed" into the interval [0; 1] or [-1; 1]. It is especially important to make correct data normalization in Data Mining algorithms, which are based on measuring the distance between the vectors of objects in the multidimensional feature space (for example, clustering).

Many **analytical tasks**, such as forecasting, analyzing sales, the demand dynamics, the state of business objects and other lengthy processes, are associated with the processing of time-dependent data. Such data is called **ordered, or time series**. During the processing of time series, special data preparation is required to optimize their presentation for solving certain analytical tasks (forecasting, classifying objects' states, identifying patterns that explain the dynamics of business processes, etc.) **for all possible date and time intervals.**

A time series includes a sequence of **observations** on the parameters (features) state of the objects or processes under study. If the observations contain one trait, then the series is **one-dimensional**, and if two or more - **multi-dimensional**. Since the values of the time series are determined only at fixed points in time, the so-called samples, the sequence of its values can be represented as follows:

$$\mathbf{X} = \{\mathbf{x}(\mathbf{1}), \mathbf{x}(\mathbf{2}) \ldots \mathbf{x}(\mathbf{n})\} \qquad (2)$$

where:

x (n) – the last value of the considered time sequence. In this case, time counts are understood to be equidistant from each other.

A time series is always a time-ordered data. However, the opposite is not always true: not all ordered data can form a time series. For example, when taking readings from a well, features are measured with a certain depth step. These are also ordered data, but they cannot be called a time series since it is not time that is fixed, but depth.

The goal of time series transformation is not to change their content, but to present information in such a way that ensures maximum efficiency in solving a specific analysis task.

Two main transformation types are most often used in the preparation of time series for analysis:

- **A sliding window** is used to solve the predicted problems and to classify the business objects states in order to convert a sequence of series values into a table, which can be used to build models or some other processing.
- **Transforming the date and time** means to bring the date and time to the most convenient form for visual analysis and processing of the time series. In this case, the results of the date conversion are no longer values like Date/Time type and can be processed as ordinary numbers and strings.

One of the most important components in business analytics is **visualization** – the presentation of data in the form that provides the most efficient user's experience. The way of visualization should reflect the data behaviour, the information contained in it, trends, patterns, as completely as possible,.

Choosing **the visualization method** depends on the nature of the data being studied and on the analysis task, as well as on the user's preferences.

Many people associate **visualization only with the interpretation**, assessment of the quality and reliability of the

analysis results. However, this is fundamentally wrong. **Visualization is necessary to apply at all stages of the analytical process without exception**. In practice, in the process of the data analysis, the user continuously works with various visualizers.

Currently, business analytics uses several dozen of visualization methods. The features and nature of the data, the specifics of the problem being solved and, finally, the user's preferences determine the method choice.

The main methods of data visualization include (Seigel E., 2016):

1. **Tabular and graphical methods.** As a rule, tables are applied when the user needs to work with individual data values, make changes, control data formats, omissions, inconsistencies, etc. Graphical methods enable to see better the general nature of the data – patterns, trends, periodic changes. In addition, graphical methods **more effectively compare the data**: it is enough to build graphs of the two studied processes in one coordinate system to assess the degree of their similarities and differences.

2. One-dimensional and multidimensional. One-dimensional visualizers provide information about only one dimension of the data, while multidimensional – about two or more. If the chart shows the dependence within the number of sales on the date, then it will be one-dimensional because it will display only one **dimension** – the Date, the values of which will correspond to the **fact** – Price. If dates and names of goods give information on sales, then **another dimension** appears – Goods, and then a multidimensional visualizer is used to correctly present the data. Popular multidimensional visualizers: OLAP-cube, multidimensional diagram, Kohonen map, etc.

3. General-purpose and specialized. General-purpose visualization methods do not relate to any particular type of analysis task or data type and can be used at any stage of the

analytical process. This is a kind of typical visualizers: graphs and charts, histograms and their varieties, statistical characteristics, etc. At the same time, there are a number of tasks, specificity of which requires the specialized visualizers' usage. For example, Kohonen maps are specifically designed to visualize clustering results, classification matrices are used mainly to test the consistency of classification models, and the correctness of the regression models is estimated using scatter diagrams.

General-purpose visualizers:

- graphics;
- charts;
- histograms;
- statistics.

Graphics.

Graphics are lines that represent the relationship between several variables in a certain coordinate system. The line on the chart consists of points set, each position is determined by the values of the dependent and independent variable (variables). The Cartesian coordinate system (X, Y, Z) is most often used. A polar coordinate system (z, θ) can also be used, where the position of the point on the coordinate plane depends on the distance to the origin of the coordinate z and angle θ.

Despite their relative simplicity, graphs **are effective means of visual data analysis**, because it is often with the construction of the graph that the work with the data begins. Using graphs, we estimate the degree of data smoothness, the presence of noise in them, anomalous emissions and gaps. They are especially useful when analyzing **time series**. Sometimes one glance at a graph is enough to reveal the presence of a trend, a seasonal component, to assess the influence degree of a random component on the process being studied.

To build the graph, it is enough to set the table values of the dependent and independent variable, mark the corresponding points on the coordinate plane and connect them with lines. The lines connecting the graph nodes can be straight or smoothed (Figure 16).

In many cases, smooth graphs are more convenient than straight ones, for visual perception and more correctly reflect real business processes, **which also often change smoothly**. Sometimes the points on which the graph is built are not connected at all, in this case, the traffic is **called a point graph.**

If it is required to present several data series on a graph, then several lines are drawn in one **coordinate system.** However, for its creation, it is necessary that all the rows displayed on the chart have the same units of measure and can be represented on the same scale.

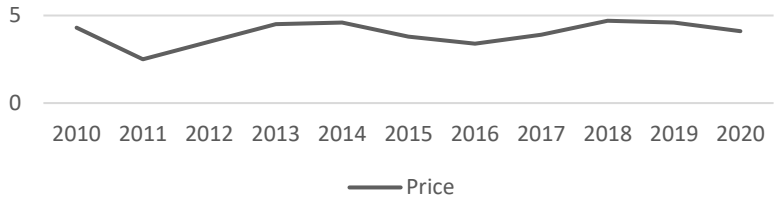


Figure 16 – The simple graph

If the values in the data series **lie in different ranges** and differ by several orders of magnitude, then displaying these series on the same graph can cause certain difficulties. Suppose you need to compare the dynamics of sales for several products. The prices of various goods, even in one group, can vary tens or hundreds of times (Kovtun N., 2015).

In this case, resorting to such a technique as the **compression of the values range** are displayed on the graph. For example, this can be done using logarithms or normalization.

In some cases, another solution to this problem is to build a graph **with two axes OY.**

Charts.

Using the graph it is most convenient to display **continuous (numerical) values** since it is possible to get a sufficient number of points to construct it. If the analyst deals with categorical (discrete) values, then a more suitable means of visualization **is a diagram (chart).**

There is no fundamental difference between the concepts of "graph" and "chart". Traditionally graphs are understood as the representation of dependencies as lines, while values in a diagram are displayed using a wide variety of objects and shapes. As a rule, categories are plotted along the horizontal axis OX, and values along the vertical OY.

The simplest and most frequently used diagrams are **bar diagrams.** The value of each category on them is represented as a column, the height of which is proportional to the corresponding value. A variation of the bar diagram is a linear diagram, which differs from the bar chart by the axis position: the axis of the categories is laid out vertically and the axis of the values horizontally (Figure 17).

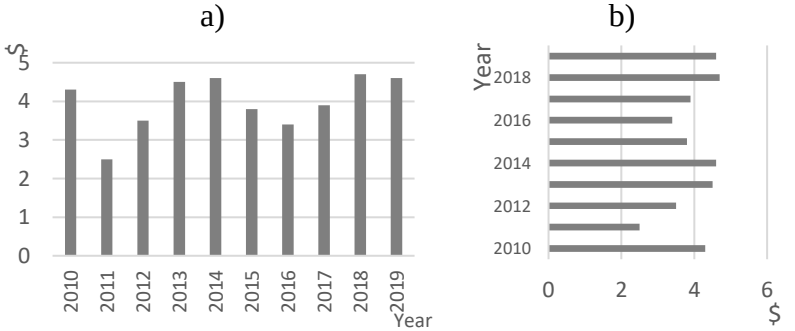


Figure 17 – The example of bar a) and linear b) diagrams

Another common type of chart is a **circular chart** (Figure 18). It is very convenient to use it if it is necessary to show the

share that a particular value contributes to the total amount. This share can be expressed both in absolute units and in percentages (for example, the percentage of ownership in the authorized capital).

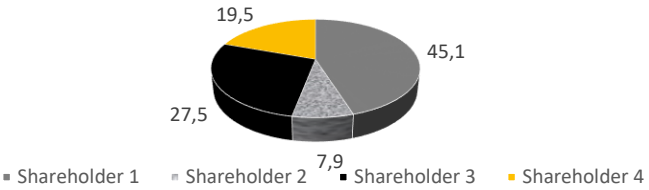


Figure 18 – The example of a circular (pie) chart

A high level of visibility is provided **by petal diagrams.** They resent each category of data as a separate axis (petal), which shows the corresponding value. Lines connect values on all axes. The presence of several series in the diagram allows to compare, for example, the changes in the sales structure dynamics in groups of goods by months (Figure 19).

The histogram shows the distribution of the dataset inside the sample (for example, the number of bank borrowers by their occupations) in the columns form.

Histograms are widely used in statistics to determine **the most probable values** that can acquire a certain value, as well as to determine the distribution law to which a random variable is subjected.

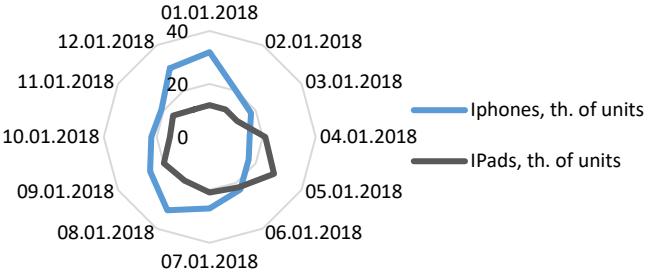


Figure 19 – The example of petal chart

The histogram is constructed in the following way. Let the investigated indicator be the daily sales in retail shops during a month. At the same time, the minimum observable value was 0 thousand UAH, and the maximum – 100 thousand UAH.

In this case, it is possible to break the range of value change by 5 subranges by 20 thousand UAH, and then calculate how many times the value of sales falls into one or another subrange (Table 14).

Table 14. Input data for histogram

Range, ths. UAH	0-20	20-40	40-60	60-80	80-100
Number of matches	1	3	9	3	2

Based on the data of this table it is possible to build **a histogram (Figure 20).**

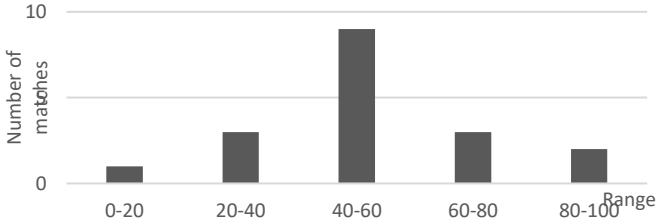


Figure 20 – The example of a histogram

Sometimes **a normalized histogram (Figure 21)** is used, which makes it possible to operate not with the values of observations, but with their probabilities. In order to do this, each element of the histogram is divided by the number of observations, that is, in this case, 31 (the number of days in a month).

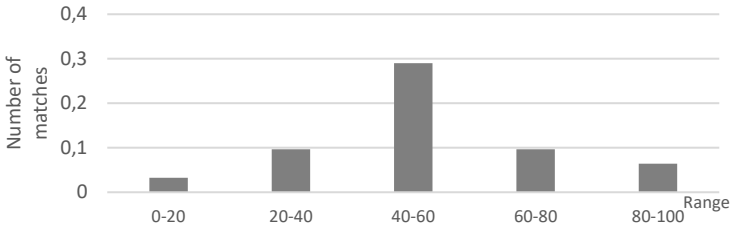


Figure 21 – The example of a normalized histogram

Statistics.

Another common visualization tool that is part of most analytic programs is information on the statistical characteristics of the sample. They are usually given **in tabular form.** Statistical characteristics allow to put forward hypotheses about the behaviour of data and their inherent laws, to control the processing results data at different stages of the analytical process.

Statistics generally includes the **following features:**

Minimum and maximum let to determine the range of changes in values. Knowledge of the minimum and maximum values gives an opportunity to see if the value lies in the range allowed for application in one or another analytical model, choose the correct method for comparing values etc.

The mean value and the mathematical expectation make it possible to put forward the hypotheses about the most probable values that can be taken by the investigated quantity.

The standard (square root) deviation and dispersion show the distribution degree for the amount of value relative to the average. These features allow to evaluate the smoothness of data rows, the noise in them, control the degree of data smoothing in the preprocessing process etc.

Distribution indicates the correspondence of the sample to some statistical distribution (normal, uniform, exponential, etc.). The distribution of the investigated value enables to

explain the peculiarities of its behaviour, for example, to determine which values are most probable.

In addition, the set of defined statistical characteristics **may include** the median, the coefficients of asymmetry and excess, etc. In addition, the statistic visualizer often contains the number of unique values for the discrete values and the number of missing values found in the sample.

Scattering diagram.

In the scattering diagrams, a number of points placed in the Cartesian coordinate system, represent values for two variables. Assigning each variable axis can determine whether there is an interconnection or correlation between these two variables.

The patterns on the scattering diagrams let to determine the different types of correlations. Including:

- positive (both values increase);
- negative (one value increases, while the second decreases);
- zero (no correlation);
- linear;
- exponential;
- horseshoe-shaped.

The correlation force is determined by the extent to which the point on the graph is closely spaced. Points that are far removed from the common cluster of points are called emissions.

The graph can use **lines or curves** that show the distance of each point to a certain "ideal value". These lines are referred to as "line of maximum compliance" or "general direction line" (Mize Ed., 2017).

Scatter diagrams are used when it is necessary to assess the **existence and effect strength of one variable to another.**

At the same time, the **correlation is not a causal dependence.** Therefore, factors that have not been taken into account may have an influence on it (Figure 22).

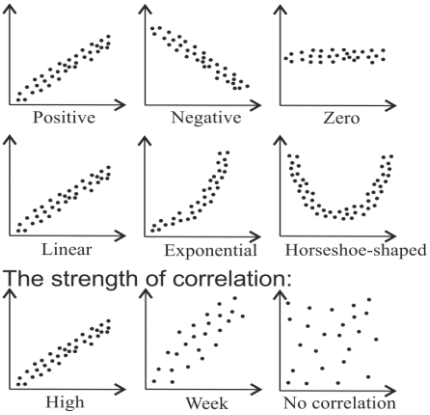


Figure 22 – Scattering diagram, type and strength of correlation

Retrospective prediction.

A special case of the regression problem is **time series prediction.** To make a prediction, a regression model is built. It is based on the past values of the series and calculates the predicted values. It has a certain set of parameters that allow to obtain a prediction taking into account the behaviour of a series in the past (Figure 23).

The last two points (sales for 2021 and 2022) are a prognosis.

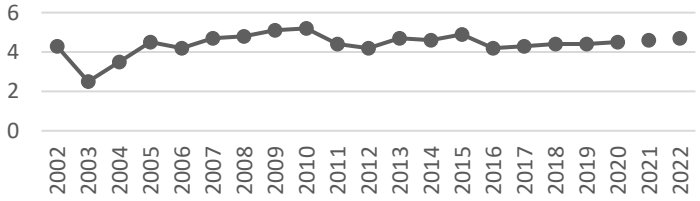


Figure 23 – Prediction graph

To test the quality of the predictive model, a special visualizer is used – **retrospective prediction.**

To build a retrospective prediction, it is necessary to select some set of data from the past to use them as initial. Then ,the predicted model with the given parameters is used. The model generates a set of predictive values, which are then compared with the data that actually occurred in the past. If in the result of this comparison, appears that between values predicted and real data there is a big difference, the predictive model requires adjustment.

The data clearing. When creating data storages, rather limited attention, as a rule, is given **to the clearing** of information that enters it. It is believed that the larger the size of the repository is better for analytical purposes. At the same time, analysts consider this practice to be harmful and hopeless on the one hand, and on the other hand, the best way to turn data storage into **a repository of meaningless data.**

Therefore, it is necessary to clear the data, because the information is always heterogeneous and accumulates from different sources. The **variety of information-gathering point** makes the cleaning process **particularly relevant.**

Largely, **errors always occur**, and it is impossible to remove them completely. Moreover, sometimes there are cases when data analysis with them is more effective than spending financial resources and time to clear data arrays from them. However, in the general case, it is necessary **to reduce the number of errors to an acceptable level** in any way**.**

In addition, **the psychological aspect of the problem** should be considered**.** If the analyst is not sure about the data that is received from the data storage, then he will try not to use them, but to receive data from other sources. In such a case, the question of **the expediency** for such data source arises (Marchenko O, 2017).

The most common mistakes in this context are the **discrepancy between types, differences in input formats and encodings.** For example, cases where information comes from different sources, in which different protocols are used to refer to the same fact. A typical example of such a mistake is the designation of **a person's gender.** In some sources it is denoted as M or F, in others it is as 1/0, in the third one it is True/False. Such problems today are quite effectively solved by **setting rules to re-encode and bring the types.**

For analytics, problems of a higher order are important, that is, those that are not solved by such elementary methods.

There are many variants of this kind of mistakes. **There are also errors that are peculiar only for a specific subject area or a particular task.**

Let us consider the **following errors that are common to any tasks:**

- Contradictory information;
- Omissions in data;
- Abnormal values;
- Noises;
- Error entering data;
- Duplication.

Contradictory information is information that does not comply with laws, regulations or reality. First, it is decided what data are necessary to consider as controversial information (Mishenina N., 2014).

Anomalous values are such values, which quite strongly detract from the general trend. Often, there is a situation where any anomaly is perceived as absolutely normal value. It is due to the fact that the prediction tools cannot understand the nature of the processes. In this case, the prediction of future values will be rather distorted. Therefore, accidental failure or success will be considered as a regularity.

Data breaks are the type of error when the data in the filling fields are missed or partially unavailable. This problem is considered very serious for most DSs. Most of the predicting methods are based on the assumption that the data are received by a constant flow.

Noise forms data indicators of which are much higher or lower than optimal values.

Incompatibility of data formats is considered data of the same type with different presentation formats.

Data entry errors are one of the most common because a person inputs them. Input errors are the type of error when data contains missed, unnecessary or distorted data.

Duplication is a duplicate of data. Repeating of different data is the most common mistake when working with data that is recorded in the DS.

There are methods to solve each of these problems. Of course, errors can be handled manually, but it is difficult to do with large volumes of data. Therefore, analytical applications, as a rule, offer variants of the decision for these tasks **in an automatic mode** with the minimum analysts' participation.

Therefore, the **data clearing essence** is a correction of various errors in order to increase the predictive adequacy of models.

Despite the multi-step data clearing procedure (in OLTP systems, in the ETL process, in SD), even at the stage of their direct analysis, **there are serious problems that** impede effective and correct data analysis. These problems are not necessarily related to data contamination (Knafli C.-N., 2015).

In terms of analysis the **"data quality" and "purity" – are not identical.** If under the data **purity** is the absence of errors during their input, structural violations, incorrect formats, and other reasons that hinder the data analysis in general, then the **data quality** closely relates to the specific goals and

objectives of the analysis used by models, methods and algorithms.

It means that data that can be considered **qualitative** from particular analytical standpoint (for example, forecasting), will be **inappropriate** for another solution (for example, a classification). To construct a prediction, it may be quite sufficient to observe the development of the investigated process in a certain time interval, and for the classification of the object, most likely, it will be necessary for its comprehensive description. **Obviously, even the criteria for data quality for these tasks may be different.**

In other words, it is impossible to consider the data qualitative or not until it won't be associated with a specific analytical task. Therefore, as soon as the data sample is loaded into the analytical application in order to solve a specific task, the analyst should evaluate the degree of their compliance with the requirements, which will let to obtain a high-quality analytical solution. If the data **does not meet these requirements**, the analyst must apply a procedure to improve the level of quality or abandon such an analysis.

The pre-processing success depends **on following factors:**

- the number of problems "inherited" from the previous data lifecycle stages (OLTP-systems, ETL-process, DS, etc.);
- a set of processing tools provided by the analytical platform;
- the analyst's ability to use these tools to bring the data into compliance with the requirements of a specific analysis task.

In order to prepare data to solve a particular analytical problem, a complex of tools is used, which is called **"pre-processing of data".**

Data pre-processing is a set of methods and algorithms used in an analytical application to prepare data for solving a

specific task and bring it into compliance with the requirements determined by the specific nature of the task and methods to solve it.

Data cleaning in an analytical application is only one aspect of pre-processing, although these two processes **are often identified**. However, cleaning is not synonymous with preprocessing. Moreover, if there are no problems in the data loaded into the analytical application that require cleaning, or their influence on the quality of the solution is assessed as minimal, then the data may not be cleaned during the pre-processing process.

At the same time, the pre-processing is carried out in any case.

4. INTRODUCTION TO DATA MINING

Data mining is the process linked with the finding of knowledge in raw data, previously unknown, non-trivial, practically useful and accessible for interpretation necessary for making decisions in various spheres of human activity.

The essence and purpose of Data Mining technology can be described as follows: it is a technology that is designed to search unobvious, objective and useful patterns in large volumes of data.

Unobvious means that the found patterns are not detected by standard methods of information processing or expert way.

Objective means that the observed patterns will fully correspond to reality, in contrast to expert opinion, which is always subjective.

Practically useful – this means that the conclusions have a specific value, which can be used in practice.

Traditional data analysis methods (statistical methods) and OLAP are mainly **focused on testing pre-formulated**

hypotheses (verification-driven data mining) and "rough" analysis, which forms the basis of operational analytical data processing (OnLine Analytical Processing, OLAP), while one of the main Data Mining provisions is to search **the non-obvious patterns** (Kovtun N., 2015)**.**

Data mining tools can find such **patterns independently and automatically build hypotheses about interconnections.** Since the formulation of the hypothesis about interconnections is the most difficult task, Data Mining advantage compared to other analysis methods is obvious.

The basic concepts of Data Mining (Marchenko O, 2017)**:**

Generic and specific notion. Generic is a divisible concept, which division elements are types of this generic, incompatible with each other, that is, those that do not have common elements. An example of the notions division: depending on the energy source, power plants (generic) are divided into (types) of hydroelectric power stations, solar power stations, geothermal, wind and thermal (nuclear power plants are referred to thermal power plants).

Data is a raw material provided by data source and used by consumers to generate information based on this data.

The object is a set of attributes. An object is also known as a record, a case, an example, a table row, etc.

Attribute is a property that characterizes the object. For example the colour of the human eyes, water temperature, etc. An attribute is also called a variable, a table field, a measurement, a characteristic.

General population – the whole set of investigated objects that interest the researcher.

The sample is a part of the general population, selected in a certain way for the purpose of research and obtaining conclusions about the properties and characteristics of the general population.

Parameters - numerical features of the general population.

Statistics - numerical features of the sample.

The hypothesis is a partially substantiated pattern that serves either for the connection between different empirical facts or to explain the fact or a group of facts. **An example of a hypothesis:** there is a relationship between life expectancy and quality of food. In this case, the purpose of the study may be to explain the changes in a particular variable, in this case – the duration of life. Let us suppose there is a hypothesis that the dependent variable (life expectancy) varies under the influence of some reasons (food quality, life style, place of residence, etc.), which in this case are independent variables. However, the variable **is not initially dependent or independent**. It becomes such one after the formulation of a specific hypothesis. The dependent variable in one hypothesis may be independent in another hypothesis.

Measurement is the process to assign numbers to the features for the objects which are studied according to a certain rule. In the data preparing process, the object is not measured, but its features do.

Scale – the rule according to which objects are assigned numbers. There are five types of measurement scales: nominal, ordinal, interval, relative and dichotomous.

Nominal scale – a scale containing only categories; the data in it can not be ordered, no arithmetic operations can be made with them. The nominal scale includes names, categories, names for the classification and sorting of objects or observations on a certain attribute. **An example of such a scale:** profession, the city of residence, marital status. For this scale, only such operations are used: equal (=), not equal (≠).

The ordinal scale is the scale in which numbers are assigned to objects with the aim to indicate the relative position of objects, but not the magnitude of the differences between

them. The scale of measurements makes it possible to rank the values of variables. The scale of measurements makes it possible to rank the values of variables. Measurements in the ordinal scale only contain information about the quantities, but do not allow to answer the questions " for what value one fact is greater than another," or "for how much it is less than another." **An example of such a scale:** the place (1, 2, 3rd) that the team received at the competitions, the student's number in the ranking of academic performance (1st, 23rd, etc.), while it is unknown how successful one student is, in comparison with another, only his number in a rating is known. This scale allows only the following operations: equal (=), not equal (≠), greater than (>), less than (<).

The interval scale – the scale, the difference between the values of which can be calculated, but their relationship does not make logical sense. This scale allows to find the difference between two values, has the properties of nominal and ordinal scales, and also lets to determine the quantitative change of the sign. **An example of such a scale:** the water temperature in the sea in the morning is 19 degrees, in the evening it is 24, i.e. the evening is 5 degrees higher, but it cannot be said that it is 1.26 times higher. **The nominal and ordinal scales are discrete, and the interval scale is continuous,** it allows making accurate measurements of the feature and performing some arithmetic operations, like addition, subtraction. This scale allows only the following operations: equal (=), not equal (≠), greater than (>), less than (<), addition (+) and subtraction (-).

A ratio scale is a scale in which there is a certain starting point and the possible relationship between the values of the scale. **An example of such a scale**: the weight of the postal item (4 kg and 3 kg). The first is 1.33 times heavier. The price for potatoes in the supermarket is 1.2 times higher than the price at the market. **Relative and interval scales are numerical.** This scale allows only the following operations: equal (=), not equal

(≠), greater (>), less (<), addition operations (+) and subtraction (-), multiplication (*) and division (/).

Dichotomous scale is a scale containing only two categories. **An example of such a scale:** gender (male and female): Yes/No; 1/0.

Variable data is data that changes its values in the process of solving a problem.

Permanent data are data that retain their values in the process of solving a problem (mathematical constants, coordinates of fixed objects) and do not depend on external factors.

Conditionally constant data are data that can sometimes change their values, but these changes do not depend on the process of solving the problem but are determined by external factors.

The data for the period characterize a certain period of time. **An example of data for a period can be** a company's profit per month, an average temperature per month.

Point data represents the value of a variable at a particular point in time. **An example of point data:** the balance on the account on the first day of the month, the temperature at eight in the morning.

Clustering is one of the tasks of Data Mining, and a **cluster** is a group of similar objects.

Clustering - 1) grouping objects based on the proximity of their properties; Each cluster consists of similar objects, and objects in different clusters differ significantly; 2) a procedure that assigns a cluster label y € Y to any object x € X.

Clustering is used when **there is no a priori information about the classes** to which the objects of the studied data set can be attributed, or when the number of objects is large, which makes it difficult to manually analyze them.

The formulation of the clustering problem is complex and ambiguous, **since:**

- the optimal number of clusters is generally unknown;
- the choice of the measure that describes “similarity” or proximity by the properties among objects, as well as the criterion to define the clustering quality, is often subjective.

Figure 24 shows an example of clustering for objects that are described by **two numerical features**, so objects can be easily represented on a plane. Unfortunately, in real applications, there is a significant number of signs for objects. Therefore, the use of such a presentation method cannot be used.

The clustering problem has been known for a long time and specialists in various scientific fields operate **on a number of other terms** – taxonomy, segmentation, grouping, and self-organization. In Data Mining, the term "clustering" is used.

The goals of clustering in Data Mining may be different and depend on the **specific problem** being solved. These include:

- **Data studying.** Breaking up a set of objects into groups helps to reveal internal patterns, increase the visibility of data presentation, put forward new hypotheses, understand how informative are the properties within the particular objects.
- **Simplifying the analysis.** With the help of clustering, the analyst can simplify further data processing and model building: each cluster is processed individually, and a model is created for each cluster separately. In this sense, clustering can be considered as a **preparatory stage** before solving other Data Mining tasks: classification, regression, association, sequential patterns.
- **Data compression.** In the case when the data is large, clustering reduces the amount of stored data, leaving one, the most typical representative within each cluster.
- **Prediction.** Clusters are used not only for the compact representation of the objects but also for the recognition of new ones. Every new object belongs to the cluster, joining which meets the criterion of clustering quality in the best way. It means

that an analyst can predict the behaviour of the object, assuming the behaviour of other cluster objects (Mize Ed., 2017).

- **Detection of anomalies.** Clustering is used to highlight non-typical objects. This task is also called outermost detection. Clusters (groups), in which an extremely small number of objects falls (one or two), in this case, represent a particular interest.

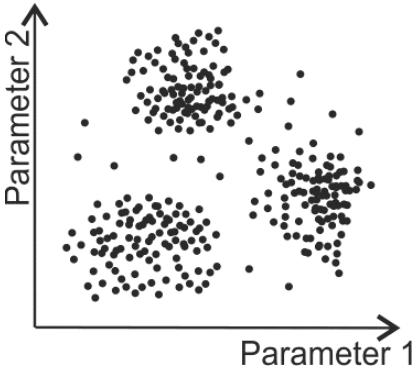


Figure 24 – Example of clustering

The clustering problem has been known for a long time and specialists in various scientific fields operate **on a number of other terms** – taxonomy, segmentation, grouping, and self-organization. In Data Mining, the term "clustering" is used.

The goals of clustering in Data Mining may be different and depend on the **specific problem** being solved. These include:

- **Data studying.** Breaking up a set of objects into groups helps to reveal internal patterns, increase the visibility of data presentation, put forward new hypotheses, understand how informative are the properties within the particular objects.
- **Simplifying the analysis.** With the help of clustering, the analyst can simplify further data processing and model building: each cluster is processed individually, and a model is

created for each cluster separately. In this sense, clustering can be considered as a **preparatory stage** before solving other Data Mining tasks: classification, regression, association, sequential patterns.

- **Data compression.** In the case when the data is large, clustering reduces the amount of stored data, leaving one, the most typical representative within each cluster.
- **Prediction.** Clusters are used not only for the compact representation of the objects but also for the recognition of new ones. Every new object belongs to the cluster, joining which meets the criterion of clustering quality in the best way. It means that an analyst can predict the behaviour of the object, assuming the behaviour of other cluster objects (Mize Ed., 2017).
- **Detection of anomalies.** Clustering is used to highlight non-typical objects. This task is also called outermost detection. Clusters (groups), in which an extremely small number of objects falls (one or two), in this case, represent a particular interest.

The examples of practical clustering application are given in Table 15.

During clustering, the question often arises: what is meant by **"similarity of properties"**. The terms "similarity", "proximity" can be understood differently, therefore, depending on which version of the estimation the proximity the analyst chooses, he will receive one or another clustering result.

In Data Mining, the common means of assessing the proximity between objects is the **metric** or the setting distance method. The problem of choosing one or another metric in models is always an acute problem facing the analyst. **The most popular metrics** are the Euclidean distance (a simple distance between two points) and the Manhattan distance (the distance between two points is equal to the sum of the modulus for their coordinate differences).

It is important to understand that clustering itself **does not yield any results of the analysis**. To obtain an effect, it is necessary to carry out a meaningful interpretation of each cluster.

For interpretation, the analyst examines in detail each cluster: its statistical features, the distribution in values of the object properties in the cluster, evaluates the **cluster power** – the number of objects that fall into it.

Table 15 – Examples of practical clustering application

No.	Field of use	Clustering problem	Description
1	Retail trade	simplifying of the analysis	Constructing of associative rules in retail stores that detect products that are acquired jointly led to difficult results with a large number of rules. Thanks to clustering, all buyers were split into several segments, and associative rules were detected in each segment separately. It enabled to break the task into a subtask and to find associative rules for each segment of customers separately.
2	Banking	data studying, simplifying the analysis	The sales department in the commercial bank operating in the retail lending market aims to study the profiles of potential customers who apply for consumer loans. The cluster named "Youth" was rather small – working students and young people under the age of 23. It turned out that the bank does not have offices in those areas of the city where the universities are concentrated. This fact was taken into account by the sales service and the Bank's business development service
		prognostication	Investigating bank customers who took auto loans, with the help of clustering tools, isolated a cluster, which included men aged 23 to 28 who live in the Sumy region for less than a year. They were united by the fact that almost all of them had a long overdue loan. Most likely, these are young people who have overestimated their capabilities, or crooks. This information has been taken into account by the bank in the course of development scoring cards

Continuation of Table 15

3	Telecommunica-tions	data studying	The database analysis, which includes the clients of a large cellular network, allowed the allocation of several clusters. A small cluster was the one that includes elderly people who are actively calling only in the spring and summer. Largely, these are pensioners who have their own private plots and much of the warmer months live outside the city. To increase the number of this cluster, an appropriate tariff plan was developed that would suit the subscribers from this group
4	Insurance	detection anomalies	The insurance company made clustering of its clients, who have been insured against accidents, revealed a small volume cluster, which contained the same doctors' names; the number of insurance payments also varied slightly. The test showed that in 90% of such cases there was a secret deal with the doctor.
5	Public services	data studying	During the study of the Migration Service Database, which contained information about people who moved from village to city (age, education, marital status, etc.), using a clustering algorithm, several segments were distinguished, the features of which allowed them to be meaningfully interpreted.

Classification and regression are one of the most important tasks of the data analysis.

Both the classification and the regression model are regularities between input and output variables. But if the input and output variables are **continuous** – we have a **regression** problem. If the source variable is single and it is **discrete** (the label of the class), then we are talking about the **classification** tasks.

The classification and regression tasks arise practically in all areas of human activity. **Classification** is used in particular:

In banking, the classification is used to solve scoring tasks, to determine the clients' credit rating in order to minimize the issuing loans risks, detecting fraud with credit cards.

In retail trade, the classification can be used to select groups of buyers with certain preferences that will more fully satisfy demand and adapt to its changes, forecast market conditions.

Speaking about **regression**, it can be used in solving the same problems, but in a slightly different setting.

For example, when determining a credit rating, all bank customers can be divided into three classes: high, medium and low. If the analyst puts the same problem as a regression problem, then as an initial variable he or she can choose an estimate of a **loan repayment probability**. Like the probability of any event, it will vary from 0 to 1. High probability of repayment (0.8-1) corresponds to a high credit rating, low probability (0.1-0.3) – low, intermediate values (0,3- 0.7) – medium.

When formulating the problem of both regression and classification, **the main goal is kept** – minimizing the risks of lending.

Classification of data can be considered as a process, which **includes two stages.**

At the first stage (Figure 25), a model is constructed that describes a predefined set of classes or categories. The model is based on the analysis of numerous data records that contain features (attributes), characterized objects and the corresponding class label. Such a dataset is called a **training sample.** Records may also have the name: observations, examples, precedents, or objects.

Since the class label for each sample in the training sample is predefined, the construction of the classification model is often referred to as **learning with the teacher.** In the

process of learning, rules are formed according to which the object is assigned to a particular class (Marchenko O, 2017).

At the second stage, the model is used to classify new, previously unknown objects and observations. Before this, the **accuracy of the constructed classification model is assessed.** One of the easiest ways to do this is to use **a test set.**

The client	Age	Income	Credit rating
Petrenko	22	low	low
Ivanenko	21	low	low
Andrienko	35	high	high
Nesterenko	41	medium	high
Kotenko	30	high	high
Nepovertaenko	43	medium	high

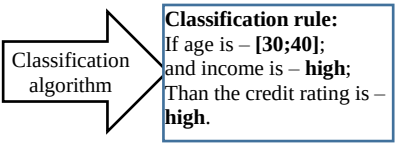


Figure 25 – The first stage of classification

From the available data set, a subset of examples is randomly selected (5-10% of the total volume), which are not used to adjust the model parameters but are fed to its input to verify learning outcomes. The specified class of each test case is compared with the class to which the example was assigned by the model. The model accuracy, assessed using a test set, **will be determined by the percentage of correctly classified test cases.**

If it is established that the **model has acceptable precision and sufficient generalization,** it can be used to classify new observations or objects for which the class is still unknown.

Basic classification methods:

In the context of business analytics, all methods can be divided into **statistical and machine learning methods.**

Statistical methods are based on mathematical statistics. These include, in particular:

- linear regression;
- logistic regression;
- Bayesian classification, etc.

Machine learning methods use learning-based models, such as:

- decision trees;
- decision rules;
- neural networks;
- method k-nearest neighbours (a simple nonparametric classification method, where the distances (usually Euclidean) used to classify objects within the space of properties are counted among all other objects. The objects to which the distance is the smallest are selected, and they are allocated in a separate class).

The task of **linear regression** is to find the coefficients of the linear regression equation, which has the next view (Knafli C.-N., 2015):

$$\mathbf{y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n,} \qquad (3)$$

where:
y - output (dependent) variable model;
x_1, x_2 ... x_n - input (independent) variables;
b - coefficients of linear regression, also called model parameters (b_0 - free coefficient).

The linear regression task is to select the coefficients of the equation in such a way that allows regression model to form the **desired output value** y for the given **input vector** X=(x_1, x_2 ... x_n).

Within the most popular applications of linear regression is a **prediction.** In this case, the input variables of the model **X** are observations from the past (predictors), and **y** is the predicted value.

Despite its versatility, the linear regression model is not always suitable for the qualitative prediction of the dependent variable. When building a linear regression model to solve a

problem, there are usually no restrictions on the values of the dependent variable. Nevertheless, in practice, such restrictions can be very significant.

Example.

The output variable can be **categorical or binary.** In such cases, it is necessary to use various special regression modifications, one of which is **logistic regression,** designed to predict a dependent variable that takes values in the range from 0 to 1. This situation is typical for problems for estimating the probability of an event based on the independent variables values. In addition, logistic regression is used to solve problems **of binary classification**, in which the output variable can take only two values - 0 or 1, "Yes" or "No", etc.

Bayesian classification.

The Bayesian approach combines a group of classification algorithms based on the maximum a posteriori (conditional) probability: for an object, using the Bayes formula, the posterior probability of belonging to each class is determined and the class for which it is maximal is selected.

A special place in this area is occupied by a simple **Bayesian classification**, which is based on the assumption about the independence of the signs describing the objects being classified. This assumption greatly simplifies the problem, because instead of a complex procedure to estimate a multidimensional probability density, it is necessary to estimate several one-dimensional ones (Marchenko O, 2017).

Decision trees.

Decision Trees (Classification Trees) is a popular classification technique in which decision rules are extracted directly from the source data in the learning process.

A decision tree is a tree-like hierarchical model, where each node checks a certain attribute using a rule. According to the results of the check, two or more child nodes are formed, into

which objects fall and for which the values of this attribute satisfy (or do not satisfy) the rule in the parent node.

Each leaf node (leaf) contains objects belonging to the same class (Figure 25).

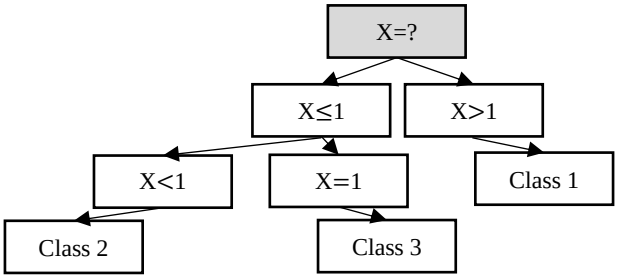


Figure 25 – The example of a decision tree

The classic algorithm to build decision trees uses a divide-and-conquer strategy. Starting from the root (parent) node, where all the training examples are present, they are divided into two or more subsets based on the attribute values selected in accordance with the division criterion (rule). For each of the received subsets, a child node is created. Then the branching process is repeated for each child node until one of the conditions for stopping the algorithm is fulfilled (Marchenko O, 2017).

Today, a large number **of decision tree generation algorithms** have been developed. They differ in the way they select attributes for splitting at each node, the conditions of stopping, and the method of **simplifying the constructed tree**.

Simplification of the tree is that after its construction, those nodes are removed, the rules in which have low value because they relate to a small number of examples. Simplification allows making the decision tree more compact.

Artificial Neural Networks

Neural networks, or artificial neural networks, are models that imitate the functioning of the brain during the operation. The neural network consists of the simplest computational elements – artificial neurons interconnected with each other.

Each neuron has several inputs and one output connection. Each input connection has a weight, on which the signal (synapse) arriving on it from the output of another neuron is multiplied. Each neuron performs the simplest transformation – weighted summation of its inputs (Figure 26).

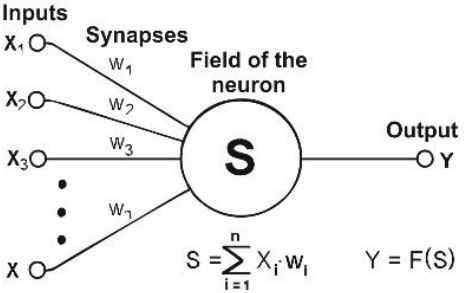


Figure 26 – Artificial neuron

In neural networks, neurons are combined into layers, while the neurons outputs from the previous layer are the inputs for the neurons from the next layer. In each layer, neurons perform parallel data processing (Figure 27) (Provost F., 2013).

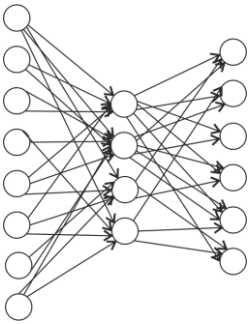


Figure 27 – The example of a neural network

5. TIME SERIES ANALYSIS

Prediction is one of the most in-demand business analytics tasks. Knowing the nature of future events, the analyst can make more informed management decisions, plan activities, develop appropriate sets of measures, effectively allocate resources, etc.

From the data analysis technologies standpoint, **forecasting can be considered as the determination of a certain unknown quantity from a related values set.** Therefore, forecasting is performed using Data Mining tasks such as regression, classification and clustering (Mize Ed., 2017).

Sales, deliveries, orders are the processes distributed in time. The data collected and used to develop forecasts, most often represent **the time series**, that is, they describe the development of a business process over time. Consequently, forecasting in the areas of sales, demand, management of material stocks and flows is usually associated with the **time series** analysis.

All forecasting methods can be divided into three large groups – **formalized, heuristic and complex.**

Formalized methods let to obtain quantitative indicators as predictions describing the state of a certain object or process. It is assumed that the analyzed object or process has the property of inertia, that is, in the future, it will continue to evolve in accordance with the same laws that developed in the past and exist in the present. The disadvantage of formalized methods consists in the fact that they can only use historical data that are within the evolutionary cycle for the development of an object or process in order to predict. Therefore, **such methods are suitable only** for operational and short-term predictions.

Formalized methods include:

- extrapolation and regression methods;

- methods of mathematical statistics;
- factor analysis, etc.

Heuristic methods are based on the use of expert estimates. An expert (a group of experts), based on his or her knowledge in the subject area and practical experience, is able to predict qualitative changes in the behaviour of the object or process under study.

These **methods are especially useful** in cases where the behaviour of objects and processes for which a prediction is required is characterized by a large unevenness. If formalized methods due to their inherent limitations are used for operational and short-term forecasts, then heuristic methods are more often used for medium-term and long-term ones.

Complex (comprehensive) prediction uses a combination of a formalized approach with expert assessments, which in some cases enables to achieve the best result.

The conditions for a market economy make it i**mpossible to manage a business without prediction effectively.** The success of the enterprise's activity will depend on how accurate and timely the forecast will be, as well as on its compliance with the tasks set (Marchenko O, 2017).

Forecasediction is a very broad concept. In most cases, it is associated with foresight in time, with the prediction of further developments. In this context, prediction is also understood in business analytics systems.

The analyst often has to deal with data, representing the history of changing various objects in time. Such data are called **time series data.** They accumulate the greatest interest in terms of analysis tasks setting and especially forecasting.

Time series is a sequence of observations about parameters changing at a time, which characterize some object or process. Strictly speaking, **each process is continuous in time**, that is, some values of the parameters within this process exist at any given time.

For example, if an analyst requests the bank about the current exchange rate, the bank workers will never say that currently there is no exchange rate. Perhaps the new course was set a minute ago, maybe in an hour it will change, but at the moment some of its value certainly exists.

For analysis tasks, you do not need to know the value of object parameters at any given time. The **timings in this context have** a special interest – fixed values in some, usually equidistant moments of time (Mishenina N., 2014).

Samples can be taken at **different intervals:** in a minute, hour, day, week, month, or year. This depends on how detailed the process is to be analyzed. In problems of time series analysis, analyst deals with discrete time, when each observation of a parameter forms a timing (time count).

The analyst should not be confused with such concepts **as continuous or discrete series and continuous or discrete data:**

- a continuous series is a process whose values are known at any given time;
- discrete series – a process whose values are known only at predetermined time points (timings);
- continuous data are data that can take many values. Only numerical data can be continuous;
- discrete data can accept a limited set of predefined values (categories).

All-time counters are numbered in order of magnitude. In this case, the time series will be represented in the following form: $X = \{x_1, x_2 \ldots x_n\}$ (Figure 28).

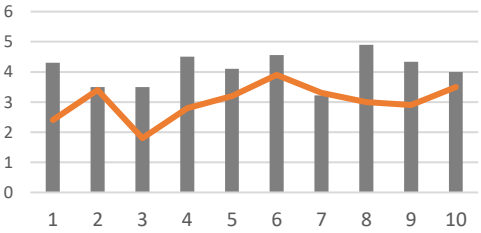


Figure 28 – An example of a discrete and continuous series

Time series may be one-dimensional and multidimensional. **The one-dimensional series** contain an observation of the change for only one parameter within the investigated process or an object, **and multidimensional** – in two parameters or more. A three-dimensional time series containing the observation of the three parameters – X, Y, Z – of the process T can be written as follows:

$$F = \{(x1, y1, z1); (x2, y2, z2) \dots (xn, yn, zn)\} \qquad (4)$$

The **time series** values are obtained by registering the corresponding parameter of the investigated process at certain time intervals. At the same time, depending on the nature of the data and the tasks being solved, either the current value (for example, the temperature or the exchange rate) or the sum of values accumulated over a certain time interval (for example, the sales per day amount, the clients per week number, etc.) is recorded (Provost F., 2013).

Goals and objectives of time series analysis.

When studying a time series, the analyst should make conclusions about the process nature and regularities, which are described by this series, based on a certain segment of finite length series. Most often during the analysis of time series the **following tasks are solved:**

- **description of** the features and patterns of the series. Based on this description, the properties of the relevant business processes can be identified;
- **modelling** – building a model of the process under study;
- **forecasting** – a prediction the future values of the time series;
- **management –** knowing the time series properties, analytics can develop methods of influence on the relevant business processes to manage them.

Two components of the time series - the natural (deterministic) and random (stochastic) can be distinguished.

The regular (deterministic) component of the time series is a values sequence, which elements can be calculated according to a certain function. The time series regular component reflects the influence of known factors and quantities.

The random (stochastic) component of the time series is a values sequence, which is the result of the influence on the investigated process by random factors. The random component and its influence on the time series can be estimated only by statistical methods.

Time series models.

Observed values of the time series are the interaction result between the deterministic and random components. There are **two types of such interaction** (Seigel E., 2016)**.**

1. **Additive** – the time series value is obtained based on the addition of deterministic and random components;

Additive time series model:

$$x_i = d_i + p_i \qquad (5)$$

where: i =1...n – serial number of the time series.

2. **Multiplicative** - the values of the time series are obtained as a result of the multiplication of deterministic and random components.

Multiplicative time series model:

$$x_i = d_i \times p_i \qquad (6)$$

Time series components.

The number of various processes in the economy, management, business, social and public sphere is very large, and the behaviour of the time series describing these processes may vary significantly. Therefore, to describe the time series behaviour, three components were introduced, a kind of **typical structures** that can be distinguished in a time series – trend, a seasonal component and a cyclical component.

Given these components, the **deterministic component** of the series can be written in the form:

$$\mathbf{d_i = t_i + s_i + c_i}, \qquad (7)$$

where:

t_i – trend;

s_i – seasonal component;

c_i – cyclic component;

i=1...n – serial number of the time series.

The trend is the most important component of the time series. The analysis of time series most often begins from the trend allocation.

The trend is a time series component that describes the main tendency in the time series and reflects the impact on it of long-term factors that cause smooth and long-term changes in the series.

Among the factors influencing economic and business processes are fast-acting and slow-acting.

Fast-acting factors such as natural disaster, stock market collapse, etc., may change the situation in a few days or even hours.

Slow-acting factors can change the situation for several months or even years. Therefore, if there is a gradual outflow of labour from the region, then it becomes scarce and, in order to attract skilled staff, employers are forced to raise wages that affect the price of goods and services. However, this process can last for years.

In addition, although for some time intervals these figures may fluctuate (for example, under the influence of other factors), the overall trend will be maintained. **Thus, trend lets to identify these patterns in the data series.**

In order to understand the trend's nature, it is usually enough to look at the graph of the time series. At the same time, different models are used to describe the trend, **the most popular of which are** (Otenko I., 2015)**:**

1. **Simple linear model:**

$$y = ax + b, \qquad (8)$$

where:
x – the number of the time period in a row (for example, the number of the month, quarter, day);
y – the sequence of values we analyze (for example, monthly sales.);
b – the point of intersection with the y-axis in the graph (the minimum level);
a – the value at which the next value of the time series increases.

If a> 0, then **the dynamics is positive.** If a <0, then the trend **dynamics is negative.**

Despite its simplicity, a linear trend model is often useful in solving real analysis tasks, since a large number of business processes **are linear in nature.**

2. The polynomial model is used to describe the values of the time series, which alternately grow and fall. Polynoms are used to analyze a large set of unstable data (for example, selling seasonal goods).

Polynom is a power function:

$$y = a + b_i \times i + b_2 \times i^2 + \cdots + b_n \times i^2, \qquad (9)$$

In most real problems, the degree of a polynomial does not exceed 5.

3. Exponential model:

$$y = \exp(a + b \times i), \qquad (10)$$

It is used in cases where the process is characterized **by a uniform increase in growth rates.**

4. Logistic model is constructed on the basis of S-shaped lines – sigmoids. They are quite vividly describing processes with a steady growth rate.

Seasonal component.

Many processes are characterized by **repeatability over time,** and the frequency of such repetitions can vary over a very wide range. Obviously, the trend is not suitable to describe such periodic changes that are present in the time series. Therefore, another component is introduced, called seasonality, or seasonal component.

Seasonal component is a time series component that describes regular changes in its values within a certain period and is a sequence of almost repeated cycles.

The seasonal component can be tied to a specific calendar time interval: day, week, quarter, month or year – or to any event that does not directly correlate with specific calendar intervals.

Cyclic component.

The time series often contain rather smooth and obvious **for the random component** changes.

At the same time, such changes can not be attributed either **to the trend**, because they are not long enough, nor **to the seasonal component** since they are not regular. Similar changes are a cyclic component of the time series. It occupies an intermediate position between the deterministic and random components of the time series (Mishenina N., 2014).

The cyclic time series component is the rise or fall intervals of varying lengths, as well as the different amplitudes for the values contained therein.

The study of the cyclic component is often useful for prediction, **especially of the short-term one.**

Main predicting principals.

In modern business analytics, **predicting models** are the main tool for further planning. The accuracy and reliability of the prediction depend on how well its model is adequate to the conditions in which the company operates, how fully it takes into account external and internal factors affecting certain business processes.

When solving the problem of predicting the time series values, that describe the dynamics of a certain business process, the input values are observations for the process development in the past, and the output values are the predicted values in the future. In this case, the time intervals of past observations and the time intervals for which a forecast is required should be consistent with each other.

For example, analyst needs to get a sales forecast for the next week. Observations on which the forecast will be based should also be taken in a week. If the database shows the sales history by day, he can get the data by week using **the appropriate aggregation. A training set is constructed by converting a time series using a sliding window.**

In addition, the observations number on the development history of the process in the past, on based on which the forecast is built, **must be greater** than the number of predicted intervals. In other words, if we want to get a forecast for the week, then we must take observations in the past few weeks.

Basic predicting models.

"Naive" predicting model

The “naive” model assumes that the last period of the predicted time series describes the future dynamics in the best way. In such models, the prediction is usually a fairly simple function of observations for the predicted value in the recent past:

$$y(t) = y(t + 1), \qquad (11)$$

where:
y(t) – the last observed data;
y(t+1) – the prediction.

This model assumes that in the next period the situation will be better than in the past. This model does not take into account: the patterns of change in the phenomenon under study; the influence of random factors; the seasonal component and trends. Therefore, the prediction accuracy of using this model **is rather conditional.**

At the same time, the "naive" model, due to its simplicity, **can be modified to take into account the seasonal component and trends.** However, such a modification requires the construction of separate predicting models and therefore involves additional consolidation and data analysis.

Extrapolation.

Extrapolation is an attempt to extend the pattern of a function behaviour from the interval in which its values are known, beyond its limits. In other words, if the values of the

function f(x) are known in a certain interval [x_0; x_n], the purpose of extrapolation is to determine the most likely value at x_{n+1}.

Extrapolation is applicable **only in cases** when the function f(x) (and, accordingly, the time series described with its help) is sufficiently stable and not subjected to sudden changes. If this requirement is not met, most likely, the behaviour of the function at different intervals will obey different laws.

The essential factors that determine the extrapolation effectiveness, **is the reliability of the data underlying the analysis.**

Extrapolation of trends has been widely used **in normative prediction**. In particular, this method helps to established whether it is possible, using existing technologies, to achieve specified production or other business indicators.

The most popular extrapolation method at present **is exponential smoothing.** Its main principle is to consider the forecast for all observations, but with exponentially decreasing weights. The method allows taking into account the seasonal fluctuations of the series and predicting the behaviour of the trend component.

Prediction by the medium and variable mean.

The model represents the usual averaging of the observations set for the predicted series (Mize Ed., 2017):

$$y(t) = \frac{(y(t) + y(t-1) + y(t-2) \ldots y(1))}{t}, \qquad (12)$$

The principle of a simple average model: "Tomorrow the situation will be like it was last time on average." **The advantage** of this approach compared with the "naive" model is obvious: when averaging, sharp changes and data emissions are smoothed. It makes the prediction results more stable to the variability of the series. However, in general, this forecasting

model is as primitive as the "naive", and it has **the same drawbacks.**

In the prediction formula on the basis of the mean, it is assumed that the series is averaged over a sufficiently long time interval (for all observations). In terms of forecast, this is not entirely correct, since the old values at the time series could be formed on other laws and lose relevance. Therefore, new observations from the near past better describe the forecast than the older values of the same series. To improve prediction accuracy, **a moving average can be used:**

$$y(t) = \frac{(y(t) + y(t-1) + y(t-2) \dots y(t-T))}{T+1}, \qquad (13)$$

The meaning of this method is that the model **"sees" only the nearest past** at T readings in time and the prediction is based only on these observations.

The moving average method is quite simple, and its results fairly accurately reflect changes in the main indicators of the previous period. Sometimes it is even **more effective** than methods based on long-term observations.

The method of time series decomposition.

Within the methods for predicting time series **is to determine the factors** that affect each value of the time series. To do this, each component of the time series is allocated, its contribution to the total component is calculated and then based on it, the future time series values are predicted. This method is called **the time series decomposition.**

The term "decomposition" means that the original time series is represented as a composition of components - trend, seasonal and cyclical. To build a forecast, these components are selected from the series, that is, **decomposition** of the series into components. Decomposition methods can be used to build both short-term and long-term forecasts.

In fact, decomposition is a selection of components from the time series and their projection to the future, followed by a combination to obtain a forecast. The method was developed for a long time ago, but today it is used less and less because of its inherent limitations. The problem is that it is very difficult to provide a sufficiently high forecast accuracy for individual components.

Generally, dozens of different forecasting methods have been developed in business analytics and their use depends on the purposes and conditions of the data analysis.

Prediction is extremely in demand in international trade, as well as in inventory management. At the same time, prediction enables to improve business processes at both end sellers and manufacturers. In international trade, forecasting contributes to the optimization of inventories, and in production, it allows to appropriately distribute capacity, plan the products range and quantity.

Particularly relevant is the introduction of predicting systems for enterprises operating in a dynamically developing economic environment. In an environment where markets are emerging, the benefits of prediction **are not only measured in monetary terms.** Those companies that embed it in their management and business planning processes earlier than their competitors gain additional market segments and thus lay the foundation for the future.

Another important aspect of the activities of commercial enterprises is the management of **material stocks and flows.** It is impossible to plan stocks without predicting sales, and the issue price is very high. Non-optimal distribution of inventories can lead to their absence where there is an urgent need for them, or to warehouses overflow in the opposite situation. Costs associated with the movement and storage of goods, can minimize income and deprive the company with a profit. In the ideal case, the product should be delivered with just-in-time

mode (in particular place and precisely when it is needed). Only demand and sales prediction will enable the logistics department to plan the inventory management process to minimize its costs.

Manual work denial and the use of specialized tools for optimizing inventory stocks is relevant **in the following cases:**

- large assortment (from 1000 positions and more);
- systematic updating of the nomenclature;
- complex planning algorithm taking into account demand forecasts, delivery methods, production plans;
- great complexity – the influence of many factors, the need for data exchange with different systems;
- a small number of procurement managers.

Inventory optimization starts **with prediction.** Sales and movements of goods are processes that are distributed in time, so all the time series analysis methods discussed above are applicable here. However, the solution to this problem does not come down to the construction of time series models: it is often a complex, multi-step procedure involving the preliminary stages of data consolidation, cleaning, and preprocessing. In the cleaning and preprocessing processes, those actions are performed that can significantly affect the quality of the prediction and the required reserves calculation, such as editing anomalies, filling gaps, smoothing, etc.

For example, abnormal sales surges or their failures associated with technical problems or the lack of goods in stock, seriously distort the objective picture and do not allow accurately to calculate the need for goods.

At the same time, it should be noted that prediction and Data Mining are used today in all branches of the **manufacturing and non-productive sectors**, in particular:

- international investment;
- international finance;
- international logistics;

- production;
- franchising;
- outsourcing;
- benchmarking;
- HR-technologies;
- PR-technologies;
- international technology and capital transfer;
- international co-operation, etc.

6. THE ESSENCE OF THE MODEL ASSEMBLY

During creation of the machine learning algorithms, developers are faced **such problems as** the computational cost for implementing the algorithm, the transparency of the constructed models for the user, and the results accuracy.

Most researchers focus on improving the classification and prediction accuracy, so the performance of new systems is often considered from this standpoint. It is easy to accept: accuracy plays a crucial role in all machine learning applications and can be easily evaluated, while transparency for the user is subjective. Speaking about the computational costs, it should be noted that with the progress of computing technology, in many cases, they have generally receded into the background.

In the past few years, interest has increased significantly in the issue to improve the training-based models accuracy in Data-mining by creating and aggregating a **set of classifiers.** As a result, new approaches to analysis, that are based **on assemblies of classifiers have appeared,** which are applicable to a wide range of machine learning systems and are based on a theoretical study of composite classifiers' behaviour.

Combination of solutions.

Taking an important decision, an experienced manager not only relies on his or her own knowledge and intuition but

also tries to attract experts in specific subject areas. It is believed that expert conclusions obtained based on the analysis of data related to the solvable problem will allow making the **optimal choice.**

For example, when an enterprise plans to launch a new product, all pros and cons are weighed down, experts are involved in the economics, marketing, production, advertising, etc. areas, and each of them expresses his opinion. Having listened to all, the head comes to **a final decision.**

However, the conclusions drawn by various experts may contradict each other. Therefore, the question arises – how to combine several expert assessments to make the right decision on their basis?

Two versions of experts' assessments are used: **parallel and complementary.** In the **first case**, each expert expresses an opinion on the whole range of problems associated with the task being solved. **In the second**, each expert gives a conclusion on only one aspect of the task. Then the experts' conclusions are complementary, covering the **whole spectrum of problems together.**

Thus, choosing the methods of extracting expert assessments and combining the results, the analyst can find **the best solution.**

A similar situation arises when using models based **on machine learning** - decision trees, neural networks, etc. These models actually play the role of experts and provide explicit or implicit information that is necessary for sound decision-making.

In principle, the analyst can restrict the results obtained **by a single model.** However, as well as experts, the models may be **weak and strong.** If a well-managed model is able to divide classes and makes few classification errors, then such a model can be **considered as strong.** The **weak model**, on the contrary,

does not allow reliable separation of classes or accurate predictions may give a lot of errors.

From this, the question arises: **how to strengthen the weak model**, what to do to improve the efficiency of the classification? A quite logical way out from this situation is trying to apply to the unsuccessful results, made by the first model, another model, which task is to **classify those examples that remain unrecognized.**

If, after that, the results are unsatisfactory, the analyst can apply a third model and so on until a **precise solution is obtained.**

Thus, in order to solve one problem of classification or regression, an analyst can apply several models, while he is interested not in the result of each individual model, but the result that gives the whole models set. Such sets of models **are called model ensembles.**

A models set used together to solve a single task is called an **assembly (committee) of models.**

The purpose to combine models **is to improve (strengthen) the solution** that a separate model provides. It is assumed that the **single model** will never be able to achieve the efficiency that the assembly will provide.

Using assemblies instead of a separate model in most cases improves the solutions quality, but this approach is associated **with some problems,** the main among which are:

- increase in time and computational costs for training several models;
- difficulty in interpreting results;
- an ambiguous choice of methods for combining the results.

Types of assemblies.

The first issue when forming an assembly is the choice of the **basic model.** The assembly as a whole can be considered

as a complicated, complex model consisting of separate (basic) models. In this case, **there are two main options:**

1. The assembly includes basic models **of the same type**, for example, only from decision trees, only from neural networks, etc.

2. The assembly includes models **of different types** – neural networks, decision trees, regression models, etc.

Each approach has its **advantages and disadvantages.** The use of different models gives the classifier additional flexibility. Nevertheless, since the output of one model is used to form the training set for another, additional transformations may be needed **to match the inputs and outputs** in the models (Knafli C.-N., 2015).

The second question is: how to use the training set when building an assembly? There are also two approaches here:

1**. Reselection.** Several sub-sections are selected from the original training set, each of which is used to teach one among the models in the assembly. If the assembly is based on models of different types, then each type will have its own learning algorithm.

2. Use of one training set **to train all models** in the assembly.

The third question concerns the **method to combine the results** published by individual models: what will be considered an output of the assembly in certain variants of the initial data? Usually, the **following methods** of combining are used:

1. **Voting.** Applies to the tasks of classification. Selects the class that was obtained by a simple majority of assembly models.

2. **Weighted vote.** In the assembly, some models can work better and others are worse. Accordingly, the results of some models are more in trust, and the results of others – less. In order to take into account the level of the results reliability, scales (points) can be assigned to models within the assembly.

3. **Averaging (weighted or unweighted).** If the assembly solves the regression problem, then the outputs of its models will be numerical. The assembly output can be defined as a simple average value of outputs among all models.

Researches of model assembly in Data Mining have become relatively recent. However many **different methods and algorithms for the** assembly **formation have been developed**. Among them, the most widespread methods are bagging, boosting and stacking.

Bagging forms a set of classifiers that are combined by voting or averaging. The concept of bagging is based on the technology called **"perturbation and combining"**.

Models that, in the learning process, **adapt their state** according to the training set, such as decision trees and neural networks, are rather unstable. Even minor changes in the training set (replacing or removing of one example) can **lead to significant changes** in the state of the model – in the structure of the decision tree or in the distribution of the weights in the neural network.

The instability of decision trees is largely due **to competing nodes** – nodes that are working about the same. Therefore, even a small change in the data can lead to the fact that the learning process will go to another node and another decision tree will be built.

The instability of models, especially decision trees, is used to create model **assembly using "perturbation and combining"** technology.

Perturbation refers to the introduction of some changes, often random, to training data and the construction of several alternative models on the modified data, followed by the results combination. **For perturbation** the following techniques are used:

- extract samples from the training set. In this case, by sampling from the initial training set, several samples are taken out and each model learns a separate model;
- sampling from the samples, forming within the samples some subgroup;
- adding noise;
- adaptive weighing;
- random selection between competing nodes (breakdowns).

Adding **various elements** to the learning process is often called **randomization.**

The idea of bagging is simple. **First, multiple samples** are formed based on the initial set of data **by random selection.** They contain the same number of examples as the original set.

However, since selection is made by accident, the set of examples in these samples **will be different:** some examples can be selected several times, and others – never. Then, based on each sample, the classifier is constructed and the outputs of all classifiers are combined (aggregated) by voting or simple averaging. It is expected that the result will be much more accurate than any single model built based on the **original data set.**

Thus, bagging includes **the following steps:**

1. The certain number of samples with the same size is extracted from the training set.
2. Based on each sample, a model is being built.
3. The overall results by voting or averaging output models are determined.

Boosting is a bit more complicated than bagging, but in many cases **it works more efficiently.** Like bagging, boosting uses instability of learning algorithms and starts creating an ensemble based on a **single output set.**

However, there are some **fundamental differences** (Mize Ed., 2017)**:**

- if in bagging the models are built in parallel and independently from each other, then in boosting, each new model is built based on the results of previously constructed models, that is, models are created sequentially
- boosting creates new models in such a way that they complement the previously constructed, perform the work that other models could not make in the previous steps.
- all built models, depending on their accuracy, are assigned weights (bagging, recall, uses weighted voting or averaging).

Instead of extracting samples from the initial data set, boosting uses **the weighing of examples** as a "disturbing" (**perturbation)** factor. The weight of each example is determined according to its impact on the classifier's training.

In each iteration, the vector of weights **is configured to reflect** the classifier's efficiency. The final classifier also aggregates the trained classifiers by voting, but now the voice in the classifier is a function of its accuracy.

Thus, the parameters that are configured at each iteration are **the weights of the examples.** Moreover, as more times the example was incorrectly recognized by previous models, as higher its weight. Weight can be considered **as the probability to get an example for the next iteration.**

In the construction **of the first model,** all the examples will be involved, in the construction **of the second**, only those that were incorrectly recognized by the first, in the construction **of the third,** examples that were incorrectly recognized by the two previous models, etc.

Models **will complement** each other – work with those examples that other models have **"abandoned"**. It is known from the theory of machine learning that only difficult to recognize examples push the learning process forward, forcing the model to classify them, **while simple examples are recognized quickly and useless for further training.**

The weights of the examples increase or decrease **depending on the output generated by each new classifier**. As a result, some difficult examples may become even more difficult, and easy ones even easier. After each iteration, the weights reflect how often an example has been classified incorrectly.

The disadvantages of boosting.

The main disadvantage of boosting is that:

- examples with low weights do not fall into the next iteration, so some of the information useful for learning may be lost.
- in the process of boosting, the simplest examples are gradually excluded from participation in training, which leads to the so-called "degeneration" of training samples based on which the latest models are built.
- first of all, the most typical examples are recognized that reflect the subject area well. Samples that are difficult to recognize are often atypical, abnormal, etc.
- training of classifiers at later iterations can be almost entirely conducted on such examples, which ultimately leads to overtraining.

Because of the above, **boosting is more prone to overtraining than bagging.**

Finally, **a common problem of boosting and bagging** is the low transparency of models for the analyst and the difficulty to interpret the results, which is generally peculiar for all model assemblies.

Application of model assemblies to various analysis tasks opens up wide opportunities to increase the models efficiency in Data Mining. Therefore, in the last few years, active research has been carried out in this area, resulting in a large number of methods.

Not all methods work equally well in different situations, and the choice of options allows the analyst to

achieve the best results. For example, boosting and tagging originally arose as methods for increasing the efficiency of classification models, especially tree-based solutions, and could not be directly used for regression models. Therefore, for the solution of regression problems, the boosting variants, called the additive regression, in particular, the LogitBoost algorithm, were developed. Such methods of combining models as selection trees and stacking are also widespread.

Additive regression.

The creation of boosting has given rise to an interest in Data Mining, as it allows to get very high-performance models. There was the construction of so-called **additive regression models** among boosting**.** The term "additive regression" refers to any predicting methods based on combining (summing) contributions from several regression models.

The idea of additive regression is based on **forwarding stepwise additive modelling,** the algorithm of which contains the following steps:

1. First, a regular regression model is built, for example, a regression decision tree.
2. To calculate the errors obtained on the training set, as the difference between the desired and the observed values. These errors are called residuals.
3. To minimize residuals that is performed using the second model (another decision tree), which tries to predict the observed residuals. To do this, the corresponding residues replace the original target values before learning the second model.
4. Since the second model is not ideal either, it will also provide some residuals. Therefore, the process will continue: a third model is built that learns to predict residuals from residuals, etc.

Therefore, the algorithm **based on forwarding stepwise additive modelling minimizes the error** of the ensemble as a whole, not individual models.

The **additive regression** tends to be over-taught since each model added to the assembly makes its tuning more accurate. To determine **the stopping time,** an error is used on the test set selected the number of iterations that minimizes it.

Additive logistic regression.

Additive regression can be used to solve **binary classification problems.** However, since logistic regression is more suitable for this, the additive model needs to be modified accordingly. With the help of logistic transformation, each class and its probability is carried out. After that, the task of regression is solved using the **models' ensemble.** Each step adds a model that maximizes the probability of a particular class.

Stacking.

Stacked generalization, or just stacking, is one way to create component models. Although the method was developed a few years ago, it is still less known than bagging and boosting. It is partly due to the complexity of the theoretical analysis, and partly because the general concept regarding the usage of this method is currently absent – the basic idea can be used **in different variations.**

Unlike bagging and boosting, stacking is not usually used to combine models of the same type, such as a decision trees set. Instead, it applies to models constructed using various learning algorithms.

In addition to the specified algorithms, **selection trees, meta learning, etc.** can also be used.

The list of questions

1. What is the role of model in data analysis? What benefits does the model representation provide to solve analytical problems?
2. Explain the essence of data? What are the types of data? Explain differences between structured, weak structured and unstructured data.
3. What are the main methods to collect data?
4. What is Data mining? What are the main instruments of data mining?
5. What is the role of OLTP and DSS systems in Data Mining?
6. Explain the essence and main features of data storage. What are the main features of data storage? What are the main architectures of data storage?
7. What is ETL process?
8. Explain the essence of transformation, visualization, clearing and data processing. What is their purpose, main instruments and practical using?
9. What is cauterization? How to use it in Data Mining?
10. What is the assembly of models? What are advantages and disadvantages to use them?

The requirements to prepare of an individual task

Requirements to design of an individual task:

1. The total volume of scientific work: 4-5 pages A4.
2. Scientific work is performed by the font size – 12, an interval of 1.5, alignment: justify text.
3. Times New Roman font.
4. The title page shall be made in accordance with the further sample.
5. Tables (if they are in the text) should be made in the following sample:

Table 1.3 – Dynamics of GDP growth in 2010-2019

6. Drawings, pictures, diagrams (if they are in the text) should be made in the following pattern:

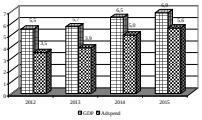


Figure 1.8 – Global advertising market growth forecast

7. Formulas:

$$MV = PQ, \qquad (1)$$

where:
M – volume of money supply,
V – ….

Typical requirements for meaningful content of an individual task:

Chapter 1. Brief theoretical description of the problem under study **(1-1.5 pages):**

- Historical aspect of the problem.
- The essence of the economic phenomenon (basic political, macroeconomic indicators, exchange rates, etc.).
- Establishment of the company, its specialization, success story (if the topic is about the company);
- Creating an online platform (YouTube, Twitch, Instagram, Facebook, etc.)
- Other theoretical aspects.

Chapter 2. Analytical section (3-4 pages):

- Indicators that are researched within the framework of an individual task and selected issues (principle of calculation, economic content);
- Visualization of analytical data - dynamics of macroeconomic and poliical indicators, financial indicators of the company, sociological researches, dynamics of stock exchange indices, fluctuations of exchange rates, shares of ownership within the authorized capital of the enterprise, dynamics of income / profits, dynamics of sales, dynamics of market dynamics. **Visualization can be in tabular or graphical form.**
- Performing the necessary calculations if the data are incomplete but allow aggregates to be calculated.
- Conclusions on the quality of the data obtained to solve specific problems – existence of anomalous data, which data should also be investigated by the company, government, etc.

- Conclusions to the analytical data. What tendencies do they demonstrate, whether there is a seasonal component in them, what conclusions can be drawn from them, an explanation of the reasons that led to this level of indicators being studied.

General conclusion on an individual task.

References

A typical structure is an example. Depending on the topic you've been selected, the content of chapters may vary. However, Chapter 2 should include analytical data in a visualized form.

Separate introduction is not required.

In addition, you need to prepare and protect the work with the appropriate presentation.

Examples of topics and sources of information are presented below.

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SUMY STATE UNIVERSITY
Department of Psychology, Political Science and Socio-Cultural Technologies

INDIVIDUAL TASK

on the subject "Data analytics"

on the topic:
“_____”

Student _____ Course _____ _____
(course number) (signature) (full name)
group _____
(group’s code)

Lecturer: PhD, senior lecturer _____ O.Kotenko
(signature)

Sumy, 20__

Possible topics for individual task:

1. Fuel scandals, their causes, consequences for international political processes
2. The essence and features of financial transactions using blockchain technology.
3. The essence and features of cryptocurrency operations. Political and economic consequences for countries (any and by choice).
4. "You-tube" in the context of an analytical study of a video hosting resource.
5. Monetization in YouTube, its features and data visualization.
6. "Twitch" in the context of an analytical study of a video streaming resource.
7. "Instagram" in the context of an analytical study of the photo and video sharing resource.
8. "Facebook" in the context of social network analytical research.
9. Cloud networks and their role in modern analytics of international political process.
10. The main features of analytics in the framework of trading with benchmark oil brands.
11. Analytics within the framework of the trading by securities on the Nasdaq stock exchange.
12. Forecasting political and economic exchange rate fluctuations based on analytical research.
13. Dynamics of the inflation index and its influence on the main macroeconomic indicators. Political consequences.
14. General characteristics of the political and macroeconomic environment in... (in a certain country – the USA, Germany, etc.)
15. Dynamics of export and import and its impact on the main macroeconomic indicators and political processes.

16. Features of analytical studies within industrial production in... (in a certain country – USA, Germany, etc.)

17. Features of analytical studies within the real estate market in.... (in a certain country – USA, Germany, etc.)

18. Analytical features of the study within the geographical structure of foreign trade in... conclusions and political perspectives. (in a certain country – USA, Germany, etc.)

19. Dynamics of direct foreign investment in... (in a certain country – USA, Germany, etc.)

20. The impact of the pandemic, quarantine measures, military conflicts on political processes in the world.

Sources of information for an individual task:

- Сайт Державної служби статистики України [Electronic resource]. – Access mode : http://www.ukrstat.gov.ua/ (вкладка – статистична інформація).
- Національний банк України [Electronic resource]. – Access mode: https://bank.gov.ua/statistic.
- Bank of England [Electronic resource] – Access mode : https://www.bankofengland.co.uk/statistics.
- Укравтопром [Electronic resource]. – Access mode : http://ukrautoprom.com.ua/en/statistika.
- Economics and Statistics Administration [Electronic resource]. – Access mode : https://www.selectusa.gov/data.
- United Nations Statistics Division [Electronic resource]. – Access mode : https://unstats.un.org/home/ (вкладка – Data).
- International monetary fund [Electronic resource]. – Access mode : https://data.imf.org/?sk=388DFA60-1D26-4ADE-B505-A05A558D9A42.

- The World Bank [Electronic resource]. – Access mode : https://data.worldbank.org.
- HowMuch.net – Understanding Money [Electronic resource]. – Access mode : https://howmuch.net.
- NASDAQ [Electronic resource]. – Access mode : https://www.nasdaq.com/ (вкладка – market activity).
- SocialBlade [Electronic resource]. – Access mode : https://socialblade.com.
- Trackalytics Social Media Statistics [Electronic resource]. – Access mode : https://www.trackalytics.com.

References

1. Knafli Cole Nussbaumer (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals: Wiley.
2. Mize Edward (2017). Data Analytics: The Ultimate Beginner's Guide to Data Analytics: CreateSpace Independent Publishing Platform.
3. Provost Foster (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking 1st Edition: O'Reilly Media.
4. Siegel Eric (2016). Predictive Analytics: Wiley.
5. Ковтун Н. В. et al. (2015). Фінансова статистика. Київ : Київський нац. ун-т ім. Т. Шевченка.
6. Марченко О. О., Россада Т. В. (2017). Актуальні проблеми Data Mining : навчальний посібник для студентів факультету комп'ютерних наук та кібернетики. Київ.
7. Мішеніна Н. В. et al. (2014). Економічний аналіз. Суми : СумДУ.
8. Отенко І. П. et al. (2015) Фінансовий аналіз : навчальний посібник. Харків : ХНЕУ ім. С. Кузнеця.

Електронне навчальне видання

Котенко Олександр Олександрович

АНАЛІТИКА ДАНИХ

Навчальний посібник

(Англійською мовою)

Художнє оформлення обкладинки О. О. Котенка
Комп'ютерне верстання О. О. Котенка

Стиль та орфографія автора збережені.

Формат 60×84/16. Ум. друк. арк. 9,07. Обл.-вид. арк. 8,99.

Видавець і виготовлювач
Сумський державний університет,
вул. Римського-Корсакова, 2, м. Суми, 40007
Свідоцтво суб'єкта видавничої справи ДК № 3062 від 17.12.2007.

Цей навчальний посібник присвячений обґрунтуванню сутності, ролі та значення даних, інформації, аналітичної роботи, поясненню її основних принципів у сучасному інформаційному середовищі, а також розгляду основних підходів та базових інструментів під час виконання аналітичних завдань фахівцями в галузі політичної аналітики, а також соціальної роботи.