Description and formation of the database perimeter for systematisation and storage of multi-structured data

A A Nechitaylo[1], O I Vasilchuk[2] and A A Gnutova[1]

[1]Samara National Research University, Moskovskoe Shosse, 34A, Samara, Russia, 443086
[2]Volga Region State University of Service, Gagarin st., 4, Togliatti, Russia, 445677

e-mail: alik.51@mail.ru, lola_red@mail.ru

Abstract. For storage of big data, as a rule, relational databases are used. For multilateral research and analysis of the processes occurring in large economic systems, financiers, economists and other technical specialists use graphs with actual names of enterprises, cities, regions, etc. to move from the physical names of the studied regions to the corresponding parameters of relational databases.

1. Introduction

"Big data" envisages the process of managing and analysing large amounts of data, which began to develop rapidly in the world since 2011, while data analysis tools began to receive information from more diversely structured sources, which is caused by the widespread introduction of digital technologies in various fields (business, medicine, entertainment, etc.). Thus, in particular, according to the Forecast of the socio-economic development of the Russian Federation for the period up to 2036, "the health care system will operate within a single digital circuit based on a unified state health information system (EGISZ), which will enable us to collect, store, process ("Big data") and analyse large amounts of information [2]. One of the final goals of this work includes the processing and intellectual analysis of big data parallel computations to create decision-making systems in real time. To solve such problems, it is necessary to determine not only the relationships (algorithms, models, etc.) of the final goal with the means to achieve it and the existing constraints, but also the forms for describing and forming the database perimeter.

2. Formulation of the problem

The task of synthesising rational schemes for choosing alternatives and evaluating their quality is to choose the best (optimal) one from the set of competing strategies for solving a certain problem, based on the analysis of the conditions and consequences of its implementation. A significant addition to what has been said is that by conditions we mean not some fixed picture of today, but also conditions that can arise during the implementation of the strategy. Accepting well-grounded optimal solutions is impossible without the steady and efficient acquisition of reliable large data arrays [8].

Taking into account the above and taking into account recent trends, in the near future the main sources of information will be Internet of Things (IoT), social media, meteorological data, GPS signals from vehicles, location data of mobile network subscribers, Google Trends, search sites work and other alternative sources of information.

3. Experimental research

The authors conducted a study on the Internet about the availability of programmes working with "big data" in the Russian-speaking community. The study showed that large users (such as Sberbank, Pyaterochka, etc.) are developing such services for their own purposes [10]. As for small business, we have not even identified the formulation of tasks, which determines the relevance of the goal of this work.

In Russia, the Central Bank of the Russian Federation and the Federal Tax Service of the Russian Federation give particular attention to systematisation and storage of multi-structured data. In this regard, the business has to solve a number of systemic and technological issues that prevent the implementation of big data analysis in everyday practice. Among these issues is the lack of strategies for companies to use the methods and data of big data analysis, the lack of modern technological solutions, and the lack of relevant skills and understanding of the key streams of data generation [9].

The study of the problems associated with the implementation of "big data" in the activities of economic entities aimed at ensuring economic security and business development shows that the strengthening of control by the Central Bank of the Russian Federation and the Federal Tax Service of the Russian Federation is directed, first of all, to the formation of a database perimeter for systematisation and storage of multi-structured data of legal entities in a single information space.

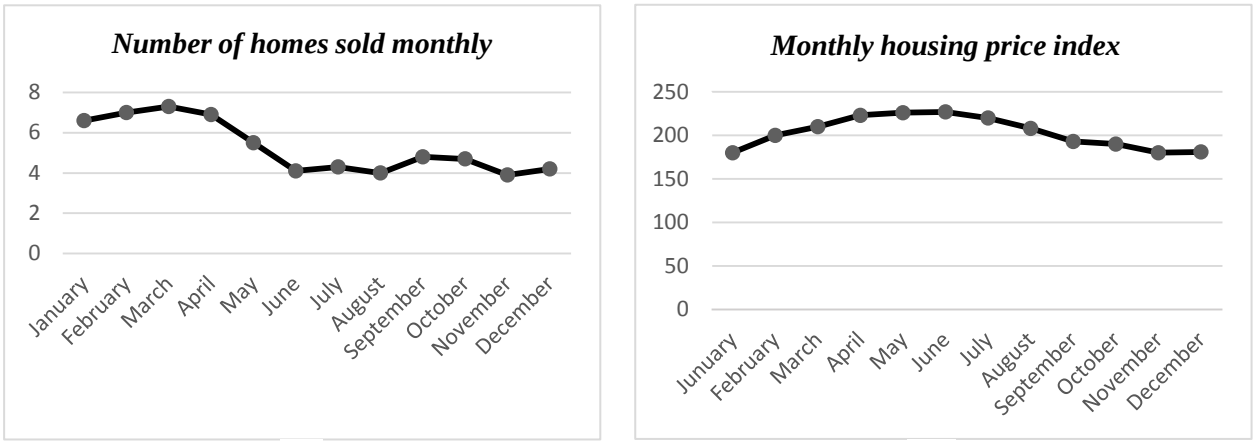
Central banks around the world have created or created departments for working with big data ("big data") in order to better understand the economy that they manage in the hope of one day getting technologies that allow them to monitor the state of the economy in real time. The current global trend is presented in table 1.

Table 1. Activities of the Central Bank of countries to promote "big data".

Region	Description of the activities of the Central Bank of the country to promote "big data"
Russia	The Bank of Russia published the first study on analysis based on "Big data". The report "Evaluation of economic activity based on textual analysis" presents a method for calculating a leading indicator of economic activity in Russia, which is based on daily contextual analysis of news sites using machine learning. A news monitoring system has been created; big data can predict consumer behaviour over a long period of time.
Japan	The Bank of Japan has been using big data since 2013 to analyse economic statistics, which helps the regulator build more accurate forecasts. It is planned to use big data for direct collection of economic data, instead of relying on survey results.
China	The People's Bank of China will more actively use "big data", artificial intelligence and cloud computing to increase its ability to recognise, prevent and reduce inter-industry and intermarket financial risks. In China, big data is interested in the context of tracking consumers and, mainly, to control debtors. One of the main problems of China is the rapid formation of "bubbles" and the tendency of the population to participate in financial pyramids. In May, the local Central Bank announced that it plans to use big data together with artificial intelligence to track such risks.

USA	In monetary policy decision making, the regulator continues to rely on traditional data sets. Economists at the Federal Reserve System (FRS) often use “big data” when studying specific issues, such as spending dynamics after hurricanes. Nevertheless, the Fed sees many shortcomings in big data, especially limited periods of time that cover these supersaturated data sets. This significantly reduces their value for forecasting. In addition, data sets are often produced by private companies that focus on something other than economic analysis. This can make big data less reliable, and the Fed is wary of using it for policy development. However, in individual projects, big data is already used, for example, to analyse consumer and government spending after hurricanes. The problem of big data, according to economists, is the too shallow depth of the sample, which significantly reduces the possibility of analysis. In addition, data is often collected by private companies that pursue their own interests. (Commercial banks: More than 60% of banks in North America believe that big data gives a competitive advantage, more than 90% that the one who copes with big data will win in the future, only 37% of banks have working projects)
Eurozone	The ECB has been exploring big data since 2013. Information on approximately 40 thousand daily transactions in the money market will become the basis of the alternative rate, since traditional benchmarks are becoming unreliable. The regulator has also acquired a large set of pricing data for actual consumer purchases and is exploring ways to measure inflation in real time. ECB analysts track Google Trends to assess unemployment change, and use algorithms to analyse media reports to assess whether the rhetoric of the regulator is viewed as “hawkish” or “pigeon”. However, the ECB remains cautious. Just as there are concerns about fake news that dominates social media, there is a risk that fake news or at least low quality statistics will crowd out better data in public discourse. Information about 40 thousand daily transactions will form the basis of an alternative discount rate. The ECB has also acquired data on the prices of real citizens' purchases and is looking at ways of online scraping to measure real-time inflation.
United Kingdom	The Big Data Board, now called the Data Management Team, has been created, as well as the data laboratory and analytical unit. Bank of England analysts recently used big data to gauge the effects of exchange rate changes. They also created a platform for these trading repositories.
India Singapore Indonesia	India faces security and privacy concerns, so the country's central bank is more concerned about cybersecurity in the context of big data. Singapore has created a Data Analysis Group, whose task is to collect large data, which will be analysed manually, without the use of AI technology. The main task, as in India, is the fight against money laundering and terrorism. The Statistics Department of the Bank of Indonesia explores social networks, news sites and other sources to analyse consumer sentiment. They recently began receiving data from online stores.

Figure 1 below illustrates the use of “big data” by banks to predict US home sales through Google Trends. The technique is based on the fact that people are looking for houses much more immediately before shopping [7].



a) b)

Figure 1. Illustration of changes in prices and the number of homes sold in the United States: a) The number of homes sold monthly; b) monthly price index for housing.

As can be seen, managerial decisions are formed based on the information received and the method of its transfer along the functional units of the organisation. Quality, reliability, timeliness will definitely influence the effectiveness of the management decision.

The age of information technology allows us to form, consolidate, modernise information, and therefore there are problems that lead to an excess of information and a deterioration of its quality [6].

According to experts, the amount of useful information in relation to all the information received will be reduced from year to year. It is believed that by far not all of the data is valuable - according to IDC estimates, by 2020 the share of useful information will be only 35% of the total generated [4].

4. Solution description

In order for the use of information that a manager receives to be effective, it is necessary to determine correctly whether the information obtained is useful and whether it will be important for making management decisions, and only after that choose the right toolkit (algorithms, models, systems, competence, etc.). Experimental comparison of relational and non-relational databases, conducted by the authors, confirms expert assessments of specialists that managing thousands of attributes that are required for economic research in relational databases is inefficient [11].

In this connection, the problem of describing and forming the perimeter of the database for systematizing and storing multi-structured documentary data becomes very relevant for the economy. A schematic representation of this is shown in Figure 2.

Document databases are intuitive to developers, since data at the application level is usually presented as a JSON document. Developers can save data using the same document model that they use in the application code. In a document database, all documents may have the same or different data structure. Each document is self-describing (that is, it contains a schema that can be unique) and does not necessarily depend on any other document. Documents are grouped into “collections,” which are similar in function to tables in relational databases.

For example, a JSON file for describing a book element in a simple book database may look like the following code.

```
[
   {
      "year" : 2013,
      "title" : "Turn It Down, Or Else!",
      "info" : {
         "directors" : [ "Alice Smith", "Bob Jones"],
         "release_date" : "2013-01-18T00:00:00Z",
         "rating" : 6.2,
```

"genres" : ["Comedy", "Drama"],
"image_url": "http://ia.media-imdb.com/images/N/O9ERWAU7FS797AJ7LU8HN09AMUP908RLlo5JF90EWR7LJKQ7@@._V1_SX400_.jpg",
"plot" : "A rock band plays their music at high volumes, annoying the neighbors.",
"actors" : ["David Matthewman", "Jonathan G. Neff"]
}
},
{
"year": 2015,
"title": "The Big New Movie",
"info": {
"plot": "Nothing happens at all.",
"rating": 0
}
}
]

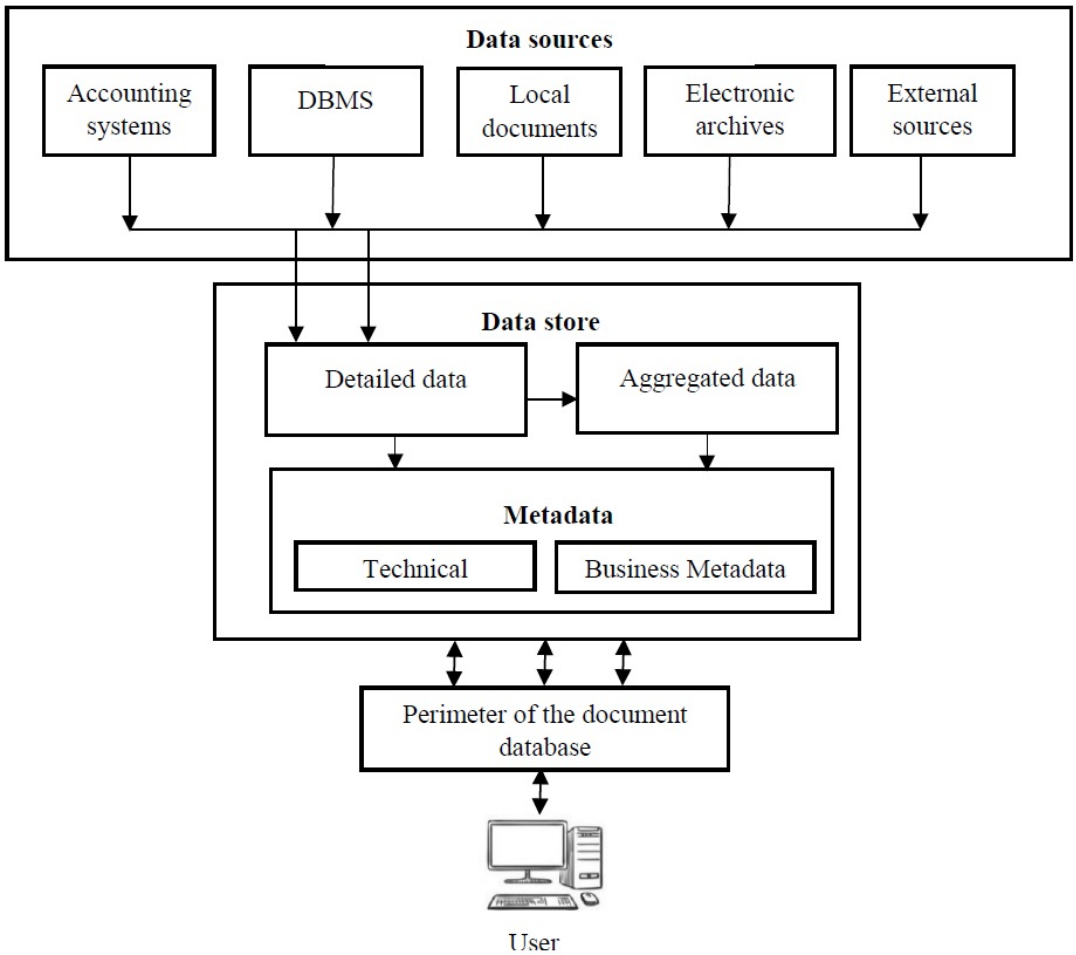


Figure 2. Illustration of the perimeter structure of the database on the systematization and storage of multi-structured data in the economy.

П when using a document database, each entity monitored by the application can be stored as a separate document. The document database allows the developer to conveniently update the application as requirements change. In addition, if you need to change the data model, then only the documents affected by this change need to be updated. To make changes there is no need to update the

schema and interrupt the database. When using a document database, the attributes of each transaction can be described in one document, which simplifies management and improves reading speed. Changing the attributes of one transaction will not affect other transactions.

Analysis of popular document databases: Amazon DocumentDB (compatible with MongoDB), Amazon DynamoDB, MongoDB, and Couchbase, based on literature and expert opinions, showed the promise of using MongoDB Documentation AWs to solve economic problems using the AWS MongoDB Quick solution, Start (also available in PDF format) for deploying a MongoDB cluster in the AWS cloud.

In solving the problems of organising modern production, it is necessary to take into account an increasing number of factors of different natures, which are the subject of research in various fields of knowledge. Under these conditions, one person cannot decide on the choice of factors influencing the achievement of a chain, and cannot determine the essential interrelationships between goals and means; in the formation and analysis of the decision-making model, there should be involved development teams consisting of specialists from various fields of knowledge, between whom interaction and mutual understanding should be organised; and the problem of decision making becomes a problem of collective choice of goals, criteria, means and options for achieving the goal, i.e. the problem of collective decision-making based on modern methods of processing big data. This leads to the fact that the formulation of the problem becomes a problem itself, for the solution of which it is necessary to develop special approaches, techniques, methods. In such cases, it is necessary to determine the scope of the decision-making problem (problem situation); identify the factors influencing its decision; choose techniques and methods that allow you to formulate or set the task so that the decision was made.

If it is possible to obtain an expression (algorithm, methodology, etc.) connecting the goal with the means, then the problem is almost always solved. These expressions can represent not only simple relations, similar to those considered, but also more complex, composite criteria (indicators) of additive or multiplicative form. Of course, in this case, computational difficulties may arise, which, if overcome, may require recourse to the formulation of the problem. However, the obtained formalised representation of the task allows us to apply further formalised methods for analysing the problem situation.

Decision making is a scientific direction that began to take shape in the middle of the last century. The task of this direction is the synthesis of rational schemes for choosing alternatives and evaluating their quality, which consists of choosing the best (optimal) one from the set of competing strategies for solving a certain problem. A significant addition to the last phrase is that the terms are understood not as some frozen picture of “today”, but also those conditions that may arise during the implementation of the strategy.

This scientific direction is distinguished by the fact that the choice of the optimality criterion must be approached creatively. According to this approach, the optimality criterion is not a kind of extremum of a function of one variable, but is an area of multidimensional feature space in which some particular parameters may be non-optimal. It is implied that we are talking about the fact that not all particular utility functions are considered as equilibrium, but as a hierarchically ordered system of utility functions with different weights (the choice of which, along with the choice of the functions themselves, is actually the content of the decision-making process).

Thus, in order to make a decision, it is necessary to obtain an expression associating the goal with the means of achieving it using the input criteria for assessing the attainability of the goal and evaluating the means. If such an expression is obtained, then the problem is solved.

5. Conclusion

In the classical theory of decision making, the central question is associated with the axioms of "rational" choice. As a result, when referring to the methods of the classical theory of decision making, the choice is reduced to binary preference relations. However, the classical rational bases of choice are not universal, but represent only a limited part of the grounds on which reasonable and natural decision-making mechanisms can be built. In order to simplify the construction and interaction of

these mechanisms (algorithms, techniques, etc.) for different sectors of the national economy, it is advisable to build typical perimeters (possibly interfaces) of big data collection and storage bases.

The number and complexity of such problems, for which it is impossible to get the performance criterion in analytical form immediately, but as the degree of development of civilisation increases the price of the wrong decision also increases. For problems of decision making, as a rule, a combination of qualitative and quantitative methods is characteristic. Decision-making in industrial control systems is often associated with a lack of time: it is better not to make the best decision, but in the required time, because otherwise the best solution may no longer be needed. Therefore, the decision often has to be taken in the context of incomplete information (its uncertainty or deficit), and it is necessary to ensure that the most relevant decision-making information and the most objective preferences underlying the decision-making can be determined as quickly as possible.

6. References

[1] Central banks use big data to form financial policy URL: http://www.vestifinance.ru/articles/95398 (20.12.2018)
[2] Forecast of the socio-economic development of the Russian Federation for the period up to 2036 URL: http://economy.gov.ru/wps/wcm/connect/9e711dab-fec8-4623-a3b1-33060a39859d/prognoz2036.pdf?MOD=AJPERES&CACHEID=9e711dab-fec8-4623-a3b1-33060a39859d (15.11.2018)
[3] How big is the internet? URL: https://geektimes.ru/company/asus/blog/275032/ (25.10.2018)
[4] How Central Banks Are Using Big Data to Help Shape Policy URL: https://www.bloomberg.com/news/articles/2017-12-18/central-banks-are-turning-to-big-data-to-help-them-craft-policy (15.11.2018)
[5] The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2022293 (05.11.2018)
[6] The problems of making effective management decisions URL: https://cyberleninka.ru/article/n/problemy-prinyatiya-effektivnogo-upravlencheskogo-resheniya (25.10.2018)
[7] Vashko T A 2011 Information duplication technology as a means of improving the quality of decision making *Problems of the modern economy* **4** 137-141
[8] Kazanskiy N L 2017 Efficiency of deep integration between a research university and an academic institute *Procedia Engineering* **201** 817-831 DOI: 10.1016/j.proeng.2017.09.604
[9] Kazanskiy N L, Protsenko V I and Serafimovich P G 2017 Performance analysis of real-time face detection system based on stream data mining frameworks *Procedia Engineering* **201** 806-816 DOI: 10.1016/j.proeng.2017.09.602
[10] Protsenko V I, Seraphimovich P G, Popov S B and Kazanskiy N L 2016 Software and hardware infrastructure for data stream processing *CEUR Workshop Proceedings* **1638** 782-787 DOI: 10.18287/1613-0073-2016-1638-782-787
[11] Kazanskiy N L, Protsenko V I and Serafimovich P G 2014 Comparison of system performance for streaming data analysis in image processing tasks by sliding window *Computer Optics* **38(4)** 804-810