ABSTRACT

thesis of Darkenbayev Dauren
on the topic: **«Numerical modeling and software development for processing large amounts of data»**
submitted for the degree of Doctor of Philosophy (PhD) in the specialty
6D075100 – « Computer Science, Computer Engineering and Management»

Relevance of the research topic. Data growth rates have increased significantly in the past decade. Research has shown that over the past two decades, the amount of data has increased approximately tenfold every two years - this exceeded Moore's Law, which doubles the power of processors. About thirty thousand gigabytes of data are accumulated every second, and their processing requires an increase in the efficiency of data processing. Uploading videos, photos and letters of users on social networks leads to the accumulation of a large amount of data, including unstructured ones. This leads to the need to work with big data of different formats, which must be prepared in a certain way for further work in order to obtain the results of modeling and calculations. In connection with the above, the research on processing big data, the development of a model and algorithms for the solution, carried out in the dissertation work, are very relevant. Undoubtedly, every year information flows will increase and in this regard, it is relevant to solve the problems of storing and processing large amounts of data, and in addition, the relevance of the dissertation topic is due to the growing digitalization, the increasing transition to professional activities online in many areas of modern society.

The dissertation work contains the results of research on the development of a model for a big data processing system, analysis and forecasting using Data Mining and machine learning methods in solving the problem of mortgage lending, specifically for analyzing, forecasting and determining the solvency of individuals receiving a mortgage loan. The dissertation deals with the solution of one of the urgent problems of the banking system - mortgage lending. The main problem lies in predicting the solvency of mortgage borrowers for a long time using the method of data mining. The main task is to implement the Big Data processing process based on the developed system that determines the solvency of mortgage borrowers. At present, the pace of long-term mortgage lending is growing annually, as a result of this, the timeliness of the research carried out in the dissertation work on the development of a model of the mortgage lending system, which clearly predicts the solvency of anyone who wants to get housing and makes the appropriate decisions, is very relevant.

The choice of the problem of mortgage lending is due to the fact that mortgage lending programs are currently being implemented in the Republic of Kazakhstan, which, in turn, requires the development of a system for analyzing, determining and predicting the solvency of mortgage loan recipients for a long time. Due to the large amount of data on consumers, their attributes, it is necessary to process a large amount of data. In the dissertation work, an effective software package for

mortgage lending was developed, which provides for modeling with a weight coefficient, which is updated in accordance with time using neural network algorithms, which greatly simplifies the work on data analysis. For example, the implementation of the developed software package by financial lending organizations significantly simplifies issues such as quality of service, drafting new programs, management, security, etc.

Many credit institutions, in particular banks and microcredit organizations, use an automated system to decide whether or not to provide loans to their clients, the system calculates a credit rating based on the characteristics of the client and gives a positive or negative result, respectively. This system is implemented by requesting the central credit bureau, checking the borrower's details and checking the credit history. It should be noted that it is impossible to get a long-term mortgage loan for a client who was unable to repay the previous loan on time due to various circumstances, which in turn will limit the possibility of obtaining housing. In connection with the above, in the dissertation work, a model of the mortgage lending system has been developed, which quite accurately predicts the solvency of any person who wants to have a home and allows making a decision on granting a loan. It is obvious that the processing of each client's data requires modern technologies and high-quality technical equipment. The system developed in the dissertation work allows you to solve many problems in the issuance of a mortgage, in particular, the borrower's data from the central database will be processed in real time and an appropriate decision will be made. The challenge is to quickly process borrower data and create a system that changes depending on the macroeconomic environment. There is only one solution to this problem - the creation of an automated system of mortgage lending based on the results of processing, analysis and comparison of the data we have. Banks keep millions of customer records and process new customer data every year. This, in turn, will greatly contribute to the creation of a new system model rich in real data. Applying for a mortgage can be done remotely and you can find out the decision of a financial organization without leaving your home, which will significantly help save time for many clients and is important in the current pandemic.

The annual growth of competition between various financial institutions dictates the need for fast processing of customer data and making appropriate decisions in the shortest possible time. Existing devices and systems do not meet the requirements of financial institutions. In this regard, the research carried out in the dissertation on modeling big data processing, creating a mortgage lending system is very relevant.

The degree of study of the research topic. The relevance of the selected research in the world is confirmed by the studies of foreign and domestic authors. Similar studies were carried out in the works of scientists from far abroad Chung, H.M., Joyce Jackson, Srinivasan V., KimYong, Henley W. E., Desai V. S., Conway D. G., Crook J. Russian scientists N.V. Babina, A.A. Zemtsova, T.Yu. Osipov, V. Rastorguev, domestic scientists Kalimoldayev M.N., Amirgaliyev E.N., Balakayeva G.T., Mamyrbayev O.Zh. and etc.

The purpose of the dissertation work: Development of a model and algorithms for processing big data, analysis and forecasting in the implementation of the applied problem of mortgage lending.

Research objectives:

1. Review of data processing methods and systems;
2. Development of algorithms and models for processing large amounts of data based on Data Mining methods: linear regression, logistic regression, multilayer neural network;
3. Evaluation of the quality of work of big data processing systems, data testing;
4. Estimation and forecasting of the solvency of individuals based on the processing of unstructured data;

Object of study. Development of a big data processing system for analyzing and predicting solvency in mortgage lending.

Subject of study: Methods and algorithms for processing BigData.

Research methods: BigData theory and technologies, Data Mining methods: linear regression, logistic regression, multilayer neural networks. NoSQL technologies, software development design.

Scientific novelty of the work:

1. Algorithms for processing unstructured data have been developed.
2. A computer model for processing large unstructured data has been developed;
3. Machine learning algorithms have been modified in accordance with the format of the problem being solved, the solvency of individuals receiving mortgage loans has been predicted;

The theoretical and practical significance of the work. The results obtained can be used in theory and practice to automate the work of financial institutions for mortgage lending. During a global pandemic, mortgage borrowers can remotely, online apply for an apartment and receive a decision from a financial institution. Thus, the research carried out in the dissertation work makes it possible to analyze and effectively decide whether or not to provide a mortgage loan to borrowers for a long time. The developed system for processing big data can be used not only to calculate the solvency of citizens, but also in other areas, for example, for the diagnosis of diseases in the field of medicine, in geoinformatics, in the field of education and other fields.

The main provisions for the defense. Based on modern technologies (NoSQL, MongoDB), a model of a system for processing large unstructured data using DataMining methods has been developed. Machine learning algorithms have been modified to determine the solvency of individuals. Analyzed the data of individuals registered in the database and predicted the solvency of individuals who received mortgage loans. It has been proven that the use of multilayer neural networks in processing large unstructured data is highly efficient.

Personal contribution of the researcher. The applicant independently solved all the problems of the dissertation work. The dissertation has studied and used modern technologies of Data Mining, MongoDB, etc., a numerical model of a system for processing large unstructured data has been developed, a software package has been created for analysis and forecasting in mortgage lending.

Volume and structure of work. The dissertation consists of an introduction, three sections, a conclusion and a list of references. The total volume of the thesis: 101 pages of written text, including 34 figures, 10 tables, bibliography from 83 sources, 3 annexes.

The introduction identified the relevance of the work and showed the problems associated with the topic. The idea of work, the purpose and objectives of the research, scientific novelty and practical value of the research, research methods are described.

In the first chapter of the dissertation, in order to determine the solvency of individuals who receive mortgage loans, an overview of scientific works and a detailed definition of Big Data are provided. An overview of the devices used to process and store large amounts of data was also presented. Mongo DB was chosen as a single database for storing unstructured data and its implementation was considered.

The second chapter examines methods for creating a model of a system that determines the solvency of individuals who receive long-term mortgage loans. The main stages are considered in detail. It was shown how Data Mining methods were used to create a model of the system. The main tasks and difficulties of creating a system for determining the solvency of individuals are identified. General problems of the thesis are described and methods of solution are proposed. An analysis was carried out to check the quality of the new system model.

In the third chapter, modified algorithms of the system model based on Data Mining and machine learning methods are implemented, the results of the development of a software package are presented. The experimental part of the thesis is presented, the results are presented in the form of tables, graphs and screen shots.

In the conclusion, the main results and conclusions of the dissertation research are presented.

Confidence level and validation results. The results of the study were discussed at scientific seminars of the Department of Informatics of the KazNU al-Farabi and were reported at the following international conferences:

– XIV Miedzynarodowej naukowi-praktycznej konferencji «Naukowa przestrzeń Europy – 2018» (Прага, Чехия);

– Студенттер және жас ғалымдардың «Фараби әлемі» атты халықаралық ғылыми конференциясы (2018, Алматы, Қазақстан);

– Международная конференция «Актуальные проблемы вычислительной и прикладной математики», «Марчуковские чтения – 2019» (Академгородок, Новосибирск, Россия);

– II Халықаралық ғылыми-пратикалық интернет конференциясы «Заманауи зерттеулердің өзекті мәселелері» (2019, Нұр-Сұлтан, Қазақстан);

– Ф.К.Бойконың 100 жылдығына арналған «Ф.К.Бойко I мерейтойлық оқулары» атты халықаралық ғылыми-техникалық конференциясы (2020, Павлодар, Қазақстан)

13 articles were published on the topic of the dissertation, the copyright certificate and act of implementation were received:

1. Даркенбаев Д.Қ. Big Data. Үлкен көлемді деректермен жұмыс істеу қағидалары // ҚазҰПУ хабаршысы. – 2017. -№ 3 (59). – Б. 211-214.
2. Balakayeva G.T., Darkenbayev D.K., Chris Phillips. Investigation of technologies of processing of Big Data // Internation Journal of Mathematics and Physics. – 2017. – Vol.8. No.2. – P.13-18.
3. Balakayeva G.T., Darkenbayev D.K. Modeling the processing of a large amount of data// Al-Farabi Kazakh National University. Journal of Mathematics, Mechanics and Computer Science. – 2018. -Vol.1(97). – P.120 – 126.
4. Балақаева Г.Т., Даркенбаев Д.Қ. Үлкен өлшемді деректерді өңдеу үдерісін моделдеу // ҚазҰПУ хабаршысы. – 2018. -№ 1(61). – Б. 248-252.
5. Darkenbayev D.K. Numerical solution of the regression model for analysis and processing of Big Data//Vestnik KazNRTU. – 2018. – № 6(130).–P.132 – 139.
6. Balakayeva G.T., Darkenbayev D.K. Correlation and regression analysis for Big Data processing // Vestnik KazNRTU. – 2019. – № 1(131). – P.338 – 345.
7. Balakayeva G.T., Chris Phillips, Darkenbayev D.K., Turdaliyev M. Using NoSQL for processing unstructured Big Data // News of the National Academy of sciences of the Republic of Kazakhstan. – 2019. –Vol.6.No.438. – P. 12 – 21.
8. G. Balakayeva, D. Darkenbayev. The solution to the problem of processing Big Data using the example of assessing the solvency of borrowers // Journal of Theoritical and Applied Information Technology. – 2020. – Vol.98. No13.– P. 2659-2670. (Scopus).
9. Darkenbayev D.K. Increasing the efficiency of processing large-size data using Big SQL technology//Materialy XIV Miedzynarodowej naukowi-praktycznej konferencji, «Naukowa przestrzeń Europy - 2018».– Vol.10. – P. 50-55.
10. Даркенбаев Д.К. Повышение эффективности и применение новых технологий для обработки больших объемов данных //V Международные Фарабиевские чтения. – Алматы, 2018. – С. 215.
11. D. K. Darkenbayev, G. T.Balakayeva. Modeling big data processing using regression analysis // Марчуковские научные чтения – 2019. – Академгородок, Новосибирск, Россия. – С. 135.
12. Даркенбаев Д.Қ. Үлкен көлемді деректерді сақтау және талдау әдістері//Заманауи зерттеулердің өзекті мәселелері» II Халықаралық ғылыми –практикалық интернет конференциясы. – Нұр-Сұлтан, 2019. – Б.120-124.
13. Darkenbayev D.K. Building a linear regression model for processing Big Data in the definition of solvency of citizens// Материалы международной научно-технической конференции «I юбилейные чтения Бойко Ф. К», посвященной 100-летию Бойко Ф. К. – Павлодар, 2020. – С. 23-29.

The copyright certificate and act of implementation were received::

1. Computer program "Processing large amounts of data using NoSQL technology and neural networks" copyright certificate No. 8459 dated February 28, 2020.
2. Act on the implementation of the results of dissertation work.