Journal of Theoretical and Applied Information Technology
15th July 2020. Vol.98. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE SOLUTION TO THE PROBLEM OF PROCESSING BIG DATA USING THE EXAMPLE OF ASSESSING THE SOLVENCY OF BORROWERS

GULNAR BALAKAYEVA, DAUREN DARKENBAYEV

Al-Farabi Kazakh National University
Almaty, Kazakhstan
E-mail: gulnardtsa@gmail.com, dauren.kadyrovich@gmail.com

ABSTRACT

This article provides a literature review and comparative analysis of methods for solving the problem of building a credit scoring model; gives definitions of the concepts of large volumes of data (Big Data); and provides an overview of existing tools for processing and storing large volumes of data. The main problems and tasks of building credit scoring are identified. The general statement of the problem is presented. Analysis of the actual problems of assessing bank credit risk, and predicting the credit worthiness of the borrower, etc. is given. The mathematical model of mortgage lending based on the processing of large amounts of data is studied. This article discusses various technologies, including forecasting using modern technologies. This contributes to the storage of big data, as well as the passage of a parallel process. We consider the problems that arise when working with big data, and identify the need for further research, to include the use of big data processing methods for real business processes in organizations that are faced with the need to process large amounts of data. In addition, further analysis of the problems associated with modeling the processing of big data is identified.

Keywords: *Big Data, NoSQL, Database, Map Reduce, Hadoop, Model, Scoring Technology*

1. INTRODUCTION

To analyze and improve business process management and turn information flows into smart digital resources is a large-scale big data problem. The concept of big data does not have a strict generally accepted definition. Usually, big data refers to the process of constant accumulation of the most diverse types of predominantly unstructured data [1]. This concept characterizes the totality of data growing exponentially, which are large, unprocessed, and unstructured for analysis by relational database methods. Whether terabytes or petabytes - the exact amount is not as important as understanding where the data ends and how it can be used. The term “big data” was coined by Clifford Lynch, editor of Nature magazine, who prepared a special issue of the journal in September 3, 2008 with the topic “How can technologies that open up the possibilities of working with large volumes of data affect the future of science?” In this issue materials were collected about the phenomenon of explosive growth in volume and variety of processed data and technological prospects in the paradigm of a possible leap "from quantity to quality" [2].

The value of big data as a source of information is obvious. Examples of tasks that can be solved by analyzing information flows of big data are include:

- Forecast of customer churn rates based on analysis of data from call centers, technical support services and website traffic;
- creating predictive models
- real-time fraud detection
- risk analysis
- construction of situational rooms
- Operational analytical processing, etc. [3].

The transition from analog to digital formats has increased the growth of business data and this is taking on ever more varied forms. According to an IDC study, 1 trillion gigabytes of data were produced in 2010, which was due to the use of several billion mobile phones, tens of billions of social media publications, and an ever-increasing number of network-connected sensors used in automobiles, utility meters, freight containers, outdoor equipment for shops, trading terminals, and many other devices [4]. In most cases, it is generally accepted throughout the world that data analysis involves the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

processing of huge amounts of data, in fact, the amount of data does not matter much in the case of unstructured and diverse data. To date, there are no universal methods or algorithms for processing and analyzing large amounts of data which would be suitable for all possible cases and situations. Each time it is required to obtain knowledge from raw data, it is necessary to develop a particular algorithm and apply appropriate methods for each task separately. Specialists around the world are studying how the development of methods and technologies for processing and analyzing large amounts of data can affect the future of enterprises. All possible data accumulated may pose a problem with the storage of that data, but nevertheless can bring huge profits when properly processed and analyzed. By 2020 (the current year), it has been predicted each inhabitant of the Earth, including the elderly and children, will each have 5,200 GB of data associated with them. Only 15% of this data will be recorded in the cloud (Digital Universe by Lucas Mearian forecast).

It is estimated that the total amount of data stored worldwide will double each year. IDC estimates that by 2020, only 33% of this data will contain information that will be of value when analyzed. By 2020, the full amount of data that humanity will possess will be 35 ST. A typical passenger liner generates 20 terabytes of data for each engine per hour. In one flight from New York to Los Angeles, a Boeing 737 generates 240 terabytes of data. If we take into account that about 30,000 flights (USA) are carried out per day, the data volumes quickly tend to petabytes [5]. Usually an organization with more than 1,000 employees has about 200 TB of data. When dealing with Big Data problems, we need to keep in mind that users need a solution that can easily be integrated into the existing data centre infrastructure and provides all three stages of information processing: collection, its organization, and analysis. Thus, the modern solution for Big Data is not an isolated system, but a complex organism of systems, each of which performs its own tasks and easily integrates with others.

2. RELATED WORKS

The relevance of the selected study in the world is confirmed by a large number of works and publications. The theoretical aspects of the development and research of Data Mining and credit scoring were studied in the works of Russian scientists, Andrey Pishchulin [6,7], Artem Rumyantsev [8], Sergey Usachev [9], BaseGroup Labs [10,11] and also in the works Foreign scientists Rimmer J. [12] , Rowland J.B. [13], Saar-Tsechansky, M., and F. Provost [14], Berry, M.J.A., and G.S. Linoff [15], Lyn C. Thomas [16], Myers J.H., and Forgy E.W. [17], Churchill G.A., Nevin J.R., Watson R.R. [18], Capon N. [19]. Data Mining methods for scoring development considered in the works of Chung, H.M., and P. Gray [20], E. Mayes [21], N.V. Babina [22], A.A. Zemtsov and T.Yu. Osipova [23], Rastorguev V. [24] and others. The practice of applying various models of credit scoring is described in the writings of Hian Chye Koh and Wei Chin Tan [25].

Today, there are many developments in credit scoring, such as SAS [26], Deductor [10,11], Statistica Data Miner [27], Scorto [28], etc. However, the problems of credit scoring are poorly studied and poorly formalized in existing finished products. Not taking into account changing external factors, as well as not automatically updating the coefficients of the parameters. In most of the final products used, mathematical models are trade secrets and are not disclosed.

2.1 Our contribution to research

When the amount of data for machine analysis began to grow faster than the amount of computational computer resources, existing mathematical methods and models cease to work and must be adapted to the needs of the task. One urgent task is the processing of large amounts of data when storing data on different servers, while collecting and processing data takes a lot of time, and such mathematical calculations as calculating the arithmetic mean value require additional costs.

In this article, we examined the task of processing large amounts of data using the example of building a credit scoring model for the banking system of the Republic of Kazakhstan as a whole. At the same time, the tasks of constructing a credit scoring model with constant weighting coefficients, with weighting coefficients updated in time, the choice of an appropriate algorithm and analysis were solved. We also examined the urgent tasks of assessing bank credit risk, the subject of which are scoring models for predicting the creditworthiness of a borrower, constructed using data mining

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

methods. Credit scoring is chosen as an example, allowing you to show the process of processing large amounts of data.

The aim of the study is to develop a credit scoring system with full automation of all stages of data processing and obtaining final results. Introduction of the obtained scoring model coefficients into the business decision-making process for approving a consumer loan.

As part of the study, a model for processing large amounts of data was developed using the example of building a credit scoring system with automatic updating of weighting coefficients. This model includes automation of all stages of data processing.

In order to solve these problems in the field of research, we applied:

- Obtaining and preparing data, selection of existing data for processing from the existing Database.
- Methods of data cleaning [29], assessing the reliability of information, identifying erroneous and suspicious data: anomalies, duplicates, contradictions, empty values, correcting detected errors.
- Methods of data transformation, transformation used to present data in a format or form optimal from the point of view of the problem being solved.
- Data mining methods, patterns search algorithms: neural networks, decision trees, regression.

We will continue the research and further research results we want to publish in the following scientific papers.

3. EXISTING PRODUCTS FOR STORING AND PROCESSING BIG DATA

For the task of collecting, processing and structuring data, the Oracle Big Data Appliance can be used as a solution - it is a pre-installed Hadoop cluster, Oracle NoSQL Database and provides integration tools for other data warehouses. The task of the Oracle Big Data Appliance is the storage and primary processing of unstructured or partially structured information, demonstrating that Hadoop-based systems work best [30].

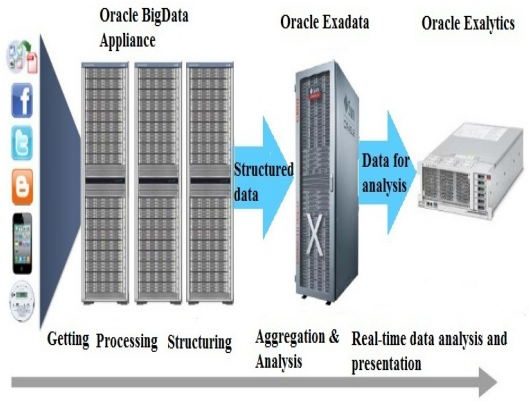

Figure 1: Data storage scheme in Oracle Big Data Appliance [30]

Gartner analysts in their articles [31,32] described the three main characteristics of Big Data, referred to as the three Vs:

- volume - the physical volume of stored data;
- velocity - the speed of data variability and, as a consequence, the subsequent analysis of these changes;
- Variety - the variety of processed data types: both structured and unstructured data.

The Big Data model implemented by the Apache Software Foundation Apache Hadoop [33] is becoming increasingly popular. Hadoop is a software platform for building distributed applications for massively parallel processing (MPP) data. In the Hadoop platform, two main components can be conditionally distinguished:

- Hadoop Distributed File System (HDFS) - a distributed file system that provides high-speed access to application data;
- MapReduce - a software platform for distributed processing of large amounts of data on a computing cluster [34].

To solve the problem of dealing with large volumes of data, a special kind of NoSQL database was developed [35]. A comparison of the properties of relational databases and NoSQL is presented in Table 1.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

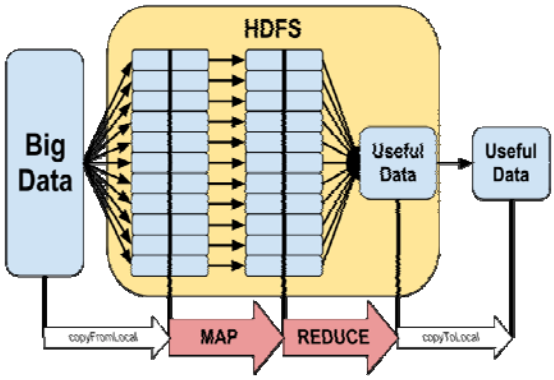

Figure 2: Map of Map-Reduce technology [34]

Table 1: Comparison of the properties of relational databases and NoSQL

Relational databases	NoSQL Databases
Complex data relationships	Very simple relationship
Schemocentricity	Arbitrary circuit; unstructured data
Scalability	Distributed processing
Static memory	Memory scales with computing resources
Universal Properties and Functions	The system is focused on the application and the developer

NoSQL technology (such as Cassandra) is not intended to replace relational databases; rather, it helps solve problems when the amount of data becomes too large. NoSQL often uses clusters of inexpensive standard servers. This solution reduces the cost per gigabyte per second by several times [5].

4. TECHNOLOGIES OF ANALYTICAL PROCESSING OF BIG DATA

Today, many organizations accumulate huge amounts of information. To process this data, it is necessary to apply various methods and approaches, depending on the task and the type of data available. There are no universal methods and technologies for processing large amounts of data and obtaining knowledge from them. Improving productivity during processing is achieved in the following ways.

4.1. Use of special types of equipment, software and hardware systems for storing large amounts of data

There is a fairly wide range of equipment for data storage, some of which use special storage technologies for the convenience of their subsequent processing and analysis. One such technology is based on the use of multiprocessor systems.

Multiprocessing is both a software and a hardware concept. In the latter case it is as the use of several physical processors in a single computer system. In the former case the term refers to the ability of a system to support more than one processor and/or the ability to distribute tasks between them. There are many variations of this concept, and the definition of multiprocessing varies depending on the context, mainly on how the processors are defined (many cores in one chip, many chips in one cluster, many clusters in one system module, etc.) and the programming paradigm [36].

Over the years of development of computer technology multiprocessors have undergone a number of stages in their development. Historically, SIMD (Single Instruction Multiple Data) technology was the first to be mastered. But such architecture tends to be restricted to use within a relatively small domain of problems. Hence interest in SIMD has been largely replaced by a steady increase in interest in MIMD (Multiple Instruction Multiple Data) architectures. MIMD architectures provide great flexibility - with adequate hardware and software support, MIMD can operate as a single-user system, providing high-performance data processing for one application; as a multi-program machine that performs many tasks in parallel; and as some combination of these features. MIMD architectures can take full advantage of modern microprocessor technology based on strict cost/performance considerations. In fact, almost all modern multiprocessor systems are built on the same microprocessors that can be found in personal computers, workstations and small uniprocessor servers [37].

Early computer architectures were based on the SISD (Single Instruction Single Data) model of operation. This implies that a single processor sequentially processes machine instructions one by one; and each machine instruction processes one quantum of data [36]. (There may be more than a single datum involved in an instruction, but the point is that

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the same instruction is not involved in multiple sets of data.)

With SIMD multiprocessors, each processor processes a single instruction stream, each of which can perform parallel calculations on a different quantum of data. This processing has been widely used in the solution of scientific problems, but its use for general, and in particular, business tasks are limited. Programs must be specially written and defined for each task separately in order to be able to use all the proposed features of the architecture.

The third existing multiprocessing architecture - MISD (Multiple Instruction Single Data) - offers the potential for dealing with redundancy, since each processing element perform its own instructions on the same data, reducing the possibility of incorrect results if one of the modules fails. The MISD architecture allows the comparison of computational results in order to detect failures, but otherwise has limited practical use.

The MIMD architecture is general purpose and is suitable for a large number of diverse tasks that implement completely independent and parallel execution of commands relating to various data sets. For this reason and because it is easy to implement, the MIMD paradigm prevails in today's multiprocessing systems. MIMD architectures can be further subdivided into shared memory systems (Figure 3) and distributed memory systems (Figure 4). MIMD software paradigms typically follow these divisions – for example a distributed memory software paradigm is often based on message passing (e.g. MPI – Message Passing Interface) [38]. However it is possible for the software paradigm to be independent of the underlying hardware architecture.

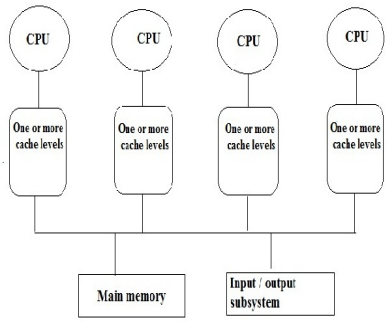

Figure 3: Typical architecture of a multiprocessor system with shared memory

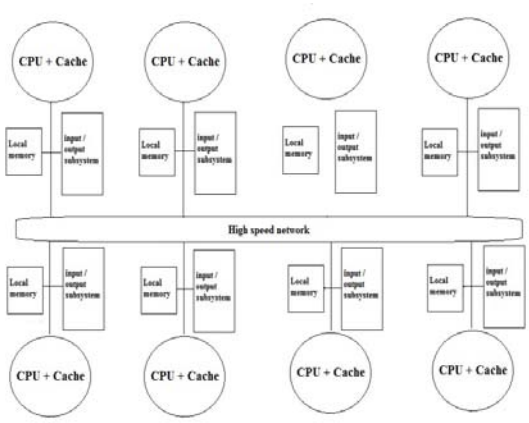

Figure 4: Typical architecture of a distributed memory machine

4.2 New types of databases

A separate problem is the task of working with large amounts of information within databases. As the amount of information grows, there are increasing problems with the required volumes on existing hard drives (although this part of the problem is relatively easy to solve), and, more importantly, with access times to the necessary data. Sophisticated caches can be used, but this, in the end, will not be enough. It is possible to partition a database by placing each class of information in its own database.

As the volume of the data bank grows, the speed of the system decreases significantly. One way to reduce data access time is to place the database in RAM. This technique can result in a gain in speed up to 100 times [5]. In-memory databases - IMDB - databases that use the computer's RAM to store data, i.e. RAM, is the main data warehouse in such systems. Since the cost of RAM decreases every day, using it as a primary storage device becomes profitable to increase the speed of data processing.

There are new types of databases for working with large volumes of data, such as databases with built-in analytics. Today, almost all databases use this concept [39]. However, Teradata developers [40] were the first to develop embedded analytics in the databases.

Also, one of the types of database is column data storages. In recent years, a number of column database systems have appeared, including MonetDB [41,42] and C-Store [43]. The developers of these systems claim that their

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

approach provides performance gains of orders of magnitude for some workloads, especially for analytical workloads with a large number of data read requests, similar to those found in data warehouse applications [44].

4.3 Analytical platform

Unlike a DBMS with a set of data mining algorithms, analytical platforms were initially oriented towards data analysis and designed to create ready-made analytical solutions.

The analytical platform is an information and analytical system, as well as a specialized software solution that contains all the tools necessary for carrying out the process of extracting patterns from "raw" data. The process of extracting patterns from the entire data array is carried out by means of consolidating information in a single source (storage data), data extraction, transformation, data mining algorithms, visualization, shifting of simple and complex methods, and models [45]. Figure 5 shows an example of the analytical platform operation scheme.

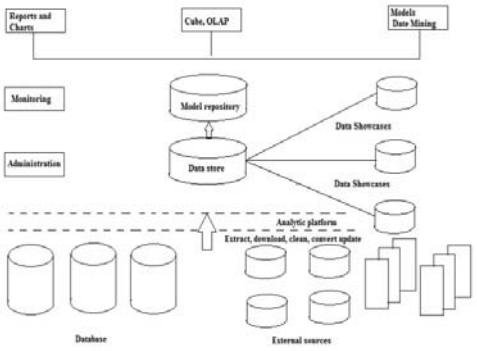

Figure 5: An example of a typical system diagram based on an analytical platform

5. DATA PROCESSING ALGORITHMS

For the correct processing of large volumes of data there is no unified method or algorithm. Accordingly, each task implies its own implementation algorithm, data analysis algorithm. Algorithms for processing large amounts of data are many, however, for certain purposes it is necessary to develop new special-purpose algorithms. Actual tasks of assessing bank credit risk, the subject of which are scoring models for predicting the creditworthiness of a borrower, constructed using data mining methods. Credit scoring is chosen as an example, allowing you to show the process of processing large amounts of data.

5.1 Credit scoring as an example of processing large amounts of data

With the growth of the credit market, the problem of credit risk arises. In this regard, it becomes necessary to assess the creditworthiness of borrowers in order to minimise possible losses. The types of credit risk are presented in Figure 6.

Allocation credit risk
Portfolio Credit Risk
Transactional Credit Risk

Figure 6: Types of credit risk

In accordance with Figure 6, transactional credit risk is the basic level of credit risk. When considering macroeconomic indicators, or economic, social, or demographic factors, there may be a risk of insolvency of borrowers, as the risk of changes in the income of the borrower increases. This situation is a transactional risk. Large groups of loans can be combined into one "big loan" on the principle of similarity, the same products, the same security, etc. This group will be called a portfolio. The need to combine loans into one group is due to the need to reduce management costs. Accordingly, as a result of forming one portfolio, it is assumed that the portfolio can be managed as one big loan. But then such a “big loan” should be characterized by parameters that allow it to assess its inherent risk - the so-called portfolio risk. The portfolio includes loans subject to the influence of the same risk factors, among which there are both economic (for example, the state of demand in industry) and social (for example, the level of household income) factors. The third level of the hierarchy, allocation credit risk, is the risk arising from the distribution of the bank’s assets by industry, region of presence, and bank products. Different dynamics of development and different conditions of regional economies, sectors, and, for example, demand for different types of bank loans, determine the variability in the quality of loan portfolios formed by the bank [46].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Credit scoring is a method for evaluating and managing transactional credit risk. It is a mathematical or statistical model with which, based on the credit history of customers who have already used the services of the bank, the latter tries to determine what is the likelihood that the client will repay the loan in a timely manner, or, to put it another way, to default on it, i.e. this is a diagnosis of the probability of bankruptcy of a potential borrower when considering the issue of its lending [47].

The main methods and technologies for developing scoring models for the banking risk management system, which include algorithms for processing large amounts of data, are the tasks of the bank. These tasks are solved by optimizing the flow of applications within the bank's divisions and building the correct and adequate scoring model using existing technologies and developing new algorithms. Automation of this process plays an important role. One of the areas of assessment of a potential borrower, individual or legal entity, before deciding on a loan to them, along with a security check and assessment of financial position is a scoring assessment.

Scoring is a method of classifying the entire sample of interest into various groups. According to Figure 7, in loan products these are groups of "bad" and "good" customers. It is not known in advance whether a client will repay the loan or not, but other facts are known that will help to determine which group the client should be assigned to. In statistics, the ideas of classifying a population into groups were developed by Fisher in 1936 using plants as an example [48]. The same technique was first applied by David Duran in 1941 to classify loans as "bad" and "good", i.e. were there any delays in repayment of the loan or not [49].

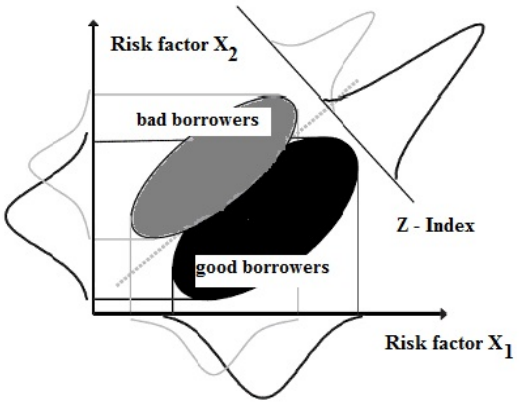

Figure 7: Geometric interpretation of credit scoring

At the outbreak of the Second World War began, many credit analysts of banks were called up to the front and financial organisations were faced with the need to automate decision-making. Analysts, and other experts, urgently prescribed the rules by which they are guided to approve or refuse a loan. This was a prototype of future expert systems. Fair Issac was the very first scoring consulting firm to be established in the city of San Francisco in the 1950s [50]. According to some studies, after the introduction of scoring systems, the level of bad debt was reduced to 50% [51,52].

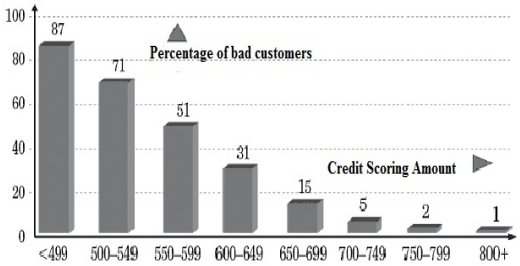

Figure 8: Source Fair Isaac and Co. [53]

Figure 8 illustrates the relationship between the score of a borrower and the probability of default on a loan. This shows that the higher the scoring scores of the borrower, the lower the probability of loan default and the lower the share of transactional risk for this borrower.

Credit bureaus contain the following types of data:

- socio-demographic characteristics;
- court decisions (in the case of the transfer of cases of claiming loan debt to the court);
- bankruptcy information
- Data on individual borrowers received from credit organizations - a bank can receive information about customers of other banks only if it provides similar information [54].

Credit scoring involves. Obtaining a credit indicator of a potential borrower based on some of their characteristics, primarily those contained in the borrower's profile. In contrast behavior scoring is a dynamic assessment of the expected behavior of a client to repay a loan, based on data on the history of transactions related to their accounts and used, in particular,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to prevent the occurrence of debt. In addition, there is collection scoring, designed to select priority "bad" borrowers who are in arrears, and areas of work to collect their debt, as well as scoring among people applying for a loan. The latter type of scoring in domestic practice is often referred to as the reliability check of a potential borrower [55].

Despite the fact that scoring solutions are available on the market today, a number of problems hinder their widespread use in the banking environment. Automated banking systems and scoring solutions exist separately and are poorly integrated with each other. Consumer lending is a system among which scoring is given an important, but not the only, role. In the classic version, it includes the following elements: the interface for remote filling out of questionnaires; a document workflow scheme; scoring; workplaces for security officers and a credit inspector; and automatic generation of a package of documents and integration with an accounting banking system. It is the implementation of all the links in this chain that supports the creation of an effective credit scoring solution, but not a separately implemented scoring technology. In addition, the cross-cutting nature of the business processes that occur during the processing of borrowers' applications leads to the fact that the decision-making time for applications is highly dependent on the interaction of bank departments. Therefore, the deployment of a credit scoring solution using a systematic approach, the re-engineering of business processes, and the process approach to management are a complex and urgent task.

The subject of this study is the means of constructing a mathematical model of the risks of a credit institution (scoring model) in an integrated banking system, together with an analysis of the development of a scoring system for consumer lending, which includes all of these elements.

Scoring models are used in economic practice when assessing the creditworthiness of individuals and legal entities, the risk of bankruptcy and solving other problems. In general terms, the mathematical scoring model is as follows:

$$S = p_1 x_1 + p_2 x_2 + \ldots p_n x_n \quad (1)$$

where S is the value of the generalized assessment of the object; $\overline{x_1}, \overline{x_2}, \ldots \overline{x_n}$ are normalized values of factors affecting the analyzed characteristics of the evaluated object; and $p_1, p_2, \ldots p_n$ are weights characterizing the significance of the relevant factors for experts.

The simplest implementation of the scoring model is a weighted sum of various characteristics of the borrower, the resulting integral indicator is then compared with a selected threshold value, on the basis of which a decision is made as to whether or not to issue a loan. This simplest implementation well reflects the essence of scoring assessment - the division of borrowers into two groups: to whom to give loans, and to whom - not [56].

Thus, the essence of the scoring model is quite simple. However, a number of difficulties lie behind this apparent simplicity. The first problem of scoring is the difficulty in determining which characteristics should be included in the model and which weights should correspond to them. The quality of the final assessment and, ultimately, the effectiveness of risk assessment and the profitability of the loan portfolio depend to a large extent on the choice of source data. There are several approaches to this problem - the classic approach is, of course, a training sample of clients that are already known, they have proven themselves to be good borrowers or not. Sample size is not a problem in Western countries, but in Kazakhstan to develop a truly effective system we need historical data on loans issued. For this we need to give out loans. And for this we need a scoring system. We thus have a vicious circle, for the normal functioning of the scoring system it is necessary to have a certain sample size, based on which we can draw a conclusion about the creditworthiness of the borrower. But this sample, in turn, is taken just from the scoring system. Therefore, it is necessary to spend some time and, through trial and error, begin to accumulate information which will form the basis of a scoring system adapted to reality. At the moment, a large amount of customer data has been accumulated in credit bureau and this information can then be used to obtain samples for scoring, in the absence of the proper amount of information in the bank database itself [56].

Theoretically, the development of a truly effective scoring system requires a significant sample of borrowers over several years with already known results (either repaid or did not repay the loan) and “training” of the system in this sample, but many banks do not yet have such statistics. Using “ready-made” Western

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

systems, in the light of the foregoing, is not an acceptable option [57,58].

Table 5 - Example Of A Scoring Card [59]

Title	№	Values	Score
Is there a housing	1	Yes	5
	2	No	1
	3	Other, not specified	-7
Age	1	52 years and more	5
	2	46 to 51	15
	3	36 to 45	3
	4	32 to 35	1

The introduction of scoring systems in the practice of banks is necessary both for the banks themselves in terms of confidence in the repayment of the loan by the borrower, and for borrowers, for whom the scoring system will significantly reduce the time for the bank to make a decision on granting a loan.

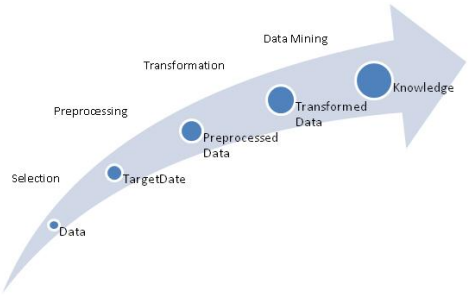

Figure 9: Data Processing Steps [60]

As shown in Figure 9, when building a scoring model, it is necessary to implement automation of the following steps.

- Obtain data from the database.
- Data preprocessing, i.e. removal of unnecessary information, empty fields and so on.
- Conversion - representation of test fields in a numerical format.
- Data mining - finding useful information from a huge amount of "raw" data that is not of interest in its original form [60].
- Interpretation / assessment - obtaining ready-made scoring model coefficients for use in consumer lending [61].

The stage of data preprocessing/cleaning involves the analysis of sharply distinguished observations, the removal of logical errors, filling in gaps in the data, etc. Data modeling - specification, selection and evaluation of models, in particular, the addition of random errors distributed according to various laws for stable estimation of model parameters, etc [61]. Post-processing, in particular scoring - checking the stability of built models on new data, visualization, etc. [62].

It should be noted that the scoring system solves not only the analysis and assessment of the creditworthiness of the borrower, but also a number of production issues that are currently becoming increasingly important for participants in the retail lending market. Namely:

- Increase in information flows
- The need to reduce decision time
- The requirements of an individual approach to each client
- Automation of decision making
- Reduced lab our costs
- Quick adaptation to changing market conditions.

The development of algorithms for constructing a scoring model with automated updating of calculation weight coefficients using data mining methods includes logistic regression and neural networks. One of the problems that arises when building credit scoring is the change in models over time. The key assumption for any mining simulation is that "the past can predict the future." In credit scoring, this means that the characteristics of past applicants, who are subsequently classified as “good” or “bad” lenders, can be used to predict the creditworthiness of new applicants. Sometimes, the tendency toward the distribution of characteristics changes over time so quickly that it requires constant updating of the credit scoring model in order to remain relevant [63].

6. CONCLUSION

This article contains the main results of a literature review. The main role and importance of large volumes of data in the modern world is determined, a review of modern technologies for processing large amounts of data is carried out. Considerable attention has been paid to credit scoring as an example of processing large amounts of data. The main components of credit risk have been described, as an introduction to the concept of credit scoring, which is the main tool to minimize the transaction risk of financial institutions. Also, some description has been given of modern products and tools currently known and used to solve processing problems. The task of building credit scoring, feature

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

analysis and classification methods have been discussed in detail. Also covered has been the general statement of the problem of constructing an automated complex of a credit scoring system with automatic updating of weighting coefficients.

REFERENCES:

[1] V.A. Rubanov "Between management standards and the element of information", Technological Forecast. 2010.No 3. pp.7-14.

[2] Vinit Yadav Processing BigData with Azure HDInsight, Building Real-Word BigData Systems on Azure HDInsight Using the Hadoop Ecosystem ISBN-13(pbk):978-1-4842-2868-5, DOI 10.1007/978-1-4842-2869-2. 2017.

[3] K.K. Nurlybayeva, G.T. Balakayeva, Algorithmization of the process of building scoring models, *Bulletin of KazNTU. Series Engineering.* Almaty, No. 6 (106), 2014. pp. 195-200.

[4] Big Data Big Opportunity //http://www.oracle.com. 28.01.2012.

[5] Yu. A. Semenov, Large amounts of data (big data), http://book.itep.ru. 04/21/2013.

[6] Pishchulin A. Credit scoring. Not everything is so scary. - What is scoring, credit scoring, scoring model. The main types of scoring: Application-scoring, Fraud-scoring, Collection-scoring // http://www.webcitation.org. 05/13/2008.

[7] Pishchulin A. National features of credit scoring // Banking. - 2008. - No. 2.-pp.91-97.

[8] Rumyantsev A. Scoring systems: science helps business // Financial Director. 2006. - No. 7.-p.1148.

[9] Usachev S. Credit scoring: desktop or enterprise solutions // Banks and Technologies. – 2008. - No. 04. - pp.50-54.

[10] BaseGroup Labs Lending to individuals // http: //www.basegroup.ru.14.03.2012.

[11] BaseGroup Labs A comprehensive scoring solution. // http: //www.basegroup.ru. 14.03.2012.

[12]. Rimmer J. Contemporary changes in credit scoring // Credit Control. - 2005.- No. 26 (4). -pp. 56-60.

[13] Rowland J. B. Confidently evaluate small businesses with credit scoring // Business Credit. -2003.- pp.26-31.

[14] Saar-Tsechansky M., ProvostF. Active sampling for class probability estimation and ranking // Machine Learning. 2004. No. 54 (2). -pp. 153-178.

[15] Berry M.J.A., Linoff G.S. Mastering Data Mining: The Art and Science of Customer Relationship Management – N.-Y .: John Wiley and Sons, 2000.– p.512.

[16] Thomas Lyn C. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers // International Journal of Forecasting. - 2000. No. 16. -pp. 149-172.

[17] Myers J. H., Forgy E. W. The development of numerical credit evaluation systems // Journal of American Statistical Association.-1963.- No. 58.- pp.799-806.

[18] Churchill G. A., Nevin J. R., Watson R. R. The role of credit scoring in the loan decision // Credit World. - 1977. - pp.6-10.

[19] Capon N. Credit scoring systems: A Critical analysis // J. Marketing. –1982. - No. 46.-pp.82-91.

[20] Chung, H. M., Gray P. Data mining // Journal of Management Information Systems. - 1999.-No.16 (1). - pp.11-13.

[21] Meise. Credit Scoring Guide .- Minsk: Grevtsov Publisher. -2008 .- p.464 .

[22] Babina N.V. Scoring as a method of assessing credit risk of consumer lending // Finance and credit. - 2007.-No 3. - pp.30-36.

[23] Zemtsov A.A., Osipova T. Yu. Credit scoring. Indirect method of assessing household wealth // Bulletin of Tomsk State University. - 2008. - No. 2. – pp.11-13.

[24] Highlights: Unique Features of Statistica Data Miner // http://www.statsoft.com. 02/01/2014.

[25] Rastorguev V. DataMining technology for data analysis in credit scoring methods // Banking Technologies. - 2003. - No. 11. - pp.14-18.

[26] Usachev S. Credit scoring: desktop or enterprise solutions // Banks and Technologies. – 2008. - No. 04. - pp.50-54.

[27] Scoring in a modern bank: problems and their solutions // http://www.scorto.ru. 06/20/2009.

[28]. Methods for cleaning and enriching personal data // http: //www.basegroup.ru. 20.06.2009.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[29] Ershov A.V. Methods of data cleaning used in building databases // Scientific session MEPHI-2008. -No 11. pp.73-75.

[30] C. Artyomov. Big Data: new opportunities for a growing business, Jet Info systems http://www.pcweek.ru. 08/20/2008.

[31] L. Doug, 3D Data Management: Controlling Data Volume, Velocity and Variety, Meta Delta, 2001. pp. 949-951.

[32] C. Pettey, L. Goasduff Gartner Says Solving Big Data Challenge Involves More Than Just Managing Volumes of Data, http://www.gartner.com.27.06.2011.

[33] http://hadoop.apache.org

[34] D. Petukhov Big Data. Problem and solutions, http://www.codeinstinct.pro. 08/11/2012.

[35] http://www.nosql-database.org

[36] A.B. Shipunov, and E.M. Baldin, Data analysis with R, http://www.inp.nsk.su. 05.01.2008.

[37] E.S. Tanenbaum, Modern operating systems,
https: //ru.wikipedia.org. 25.07.2014.

[38] https://www.mpi-forum.org/

[39] V.Z. Schnitman, S,D. Kuznetsov, Multiprocessor systems, http://citforum.ru 03.03.2015.

[40] A. Belov, What are RAID arrays and why are they needed, http://sonikelf.ru. 11/13/2010.

[41] A. Obukhov, In-Memory. Database in random access memory, http://ecm-journal.ru. 04.04.2014.

[42] B. Franks, The Taming of Big Data: How to Extract Knowledge from Arrays of Information Using Deep Analytics. M.: Mann, Ivanov and Ferber, 2014. –180 p.

[43] P. Boncz, M. Zukowski, N. Nes, MonetDB/X100: Hyper-pipelining query execution // Proceedings of conference CIDR. 2005.

[44] P.A. Boncz, M.L., Kersten M.I.L. primitives for querying a fragmented world, VLDB Journal. №8 (2), 1999, pp. 101–119.

[45] M. Stonebraker, A. Batkin, X. Chen, M. Cherniack, M. Ferreira, E. Lau, A. Lin, S. R. Madden, E. J. O'Neil, P.E. O'Neil, A. Rasin, and N. Tran, S.B. Zdonik, C-Store: A Column-Oriented DBMS, VLDB Journal, 2005, pp. 553-564.

[46] D.J. Abadi, S. Madden and N. Hachem, Column Stores vs. Row Stores: How Different Are They Really? // Proceedings of the ACM SIGMOD International Conference on Management of Dat, Vancouver, 2008.

[47] D. Gavrilov, Analytical platform - what is it? // http://www.abc.org.ru. 12/28/2006.

[48] D. Evgueni Solojentsev, Scenario Logic and Probalilistic Management of Risk in Business and Engineering. Springer Science Business Media Inc 2005. Printed in USA springeronline.com

[49] HAH Abdou, J. Pointon, Credit scoring, statistical techniques and evaluation criteria: Are view of t h e literature This version is available at: http://usir.salfor d.ac.uk/id/e print/16 5 1 8/ 2011

[50] Nicola Jentzsch Financial Privacy, An International Comparison of Credit Reporting Systems Second Edition Springer-Verlag Berlin Heidelberg 2007.

[51] V.N. Cherkashenko, Risk management of lending to small and medium-sized businesses //http://bankir.ru. 07/31/2012

[52] G.A. Churchill, J.R. Nevin, and R.R. Watson The role of credit scoring in the loan decision, CreditWorld, 1977.

[53] Donncha Marron //Consumer Credit in the United States, A Sociological Perspective from the 19th Century to the Present// 2009/ DOI https://doi.org/10.1057/9780230101517

[54] G. Andreeva, Scoring as a method of assessing credit risk, Banking Technologies, http://www.cfin.ru. 07/05/2002.

[55] A.V. Panyukov, E.S. Budina Using a credit scoring system to organize the retail lending process, 2009, pp. 41-47.

[56] E. Mazin, Banks go to the regions, Business, No. 99, 2005.

[57] V.A. Burakova, Problems of applying scoring in Russian banking practice, http://bizness-gruppa.ru. 01/28/2009.

[58] K.K. Nurlybayeva and G.T. Balakayeva, Analysis of large volumes of data for decision making, *Abstracts of the International Conference of Students and Young Scientists*. Almaty, 2014, pp. 133.

[59] Sidney J. Blatt, Q. Richard Ford Plenum press, New York, 1994.

[60] Tell me who your borrower is and I'll tell you who you are, http://www.credits.ru. 11/05/2003.

[61] K.K. Nurlybayeva, Modeling the processing of large amounts of data for decision making, *Abstracts of the X International Scientific Conference of Alumni, Students,*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Undergraduates and Young Scientists, Astana, 2014, pp. 88-90.

[62] I. Chubukova Data Mining lecture course. Lecture 5: Objectives of Data Mining. Classification and clustering, http://www.intuit.ru. 06/05/2006.

[63] G.T. Balakayeva, C. Phillips, D.K. Darkenbayev, M. Turdaliyev, Using NoSQL for processing unstructured BigData, News of the Republic of Kazakhstan *series of Geology and Technical sciences* ISSN 2224-5278 Volume 6, Number 438 (2019), pp.12 – 21