Distributed Infrastructure for Big Data Processing in the Transaction Monitoring Systems

M U Sapozhnikova[1], M M Gayanova[1], A M Vulfin[1,2], A V Nikonov[1] and A V Chuykov[1]

[1]Ufa State Aviation Technical University, K. Marks St. 12, Ufa, Russia, 450077

Abstract. To increase the effectiveness of detecting fraudulent bank transactions, the structure of the system is proposed to analyze data of user environment in order to identify potential fraudulent activities. The system for collecting and analyzing information about the user environment allows to accumulate data about the user environment, to mark precedents in manual and automatic modes and build a database of images for classifiers training. It is necessary to implement data collection, storage and access interface for the application of data mining tools. Operation of significant amount of accumulated data requires the use of special tools (frameworks and hardware platforms) for processing large data. In this paper the analysis of the existing software and hardware tools for distributed processing of indefinitely structured data of bank transactions (frameworks: Hadoop, Apache Spark) is presented. The structure and recommendations for the deployment of a hardware and software stand for testing algorithms for detecting financial fraud on the basis of data mining analysis as part of a distributed data processing system for bank transactions based on the selected framework are developed.

1. Introduction

The penetration of information technologies into all spheres of human life creates the basis for the formation of new conditions for the functioning of the market. In these conditions it became possible to develop the digital economy. The key factors of this economy are electronic technologies and services and digital representation of large volume multi-branch data [1, 2]. E-commerce is a significant institution in this branch of economy, penetrating into an increasing number of legal relationships that are emerging in the field of trade in electronic form. There is a rapid growth in the sphere of financial technologies: the introduction of artificial intelligence technologies, machine learning, analysis of large data to improve the efficiency of interaction of all participants in legal relations [3, 4, 5].

An important aspect of the functioning of the digital economy is the provision of information and economic security of business, personal data protection. As a result of the rapid development of financial technologies in the whole world, there has been an increase in fraudulent activities in the electronic environment. According to the Central Bank of Russia for the year 2014, the share of fraudulent transactions in Internet banking was 63%, and in the last 2 years – increased 5.5-fold and accounted for 93% of all crimes related to embezzlement of funds from cardholders' accounts [6, 7].

Nowadays, the application of big data processing technologies and data mining methods (DM) is an important element of the anti-fraud system. For example, introduction of Big Data by HSBC has increased the efficiency of fraudulent incident detection by 10 times [8]. VISA anti-fraud system helps prevent fraudulent payments amounting to $ 2 billion annually [9].

2. Analysis of user environment data as a part of anti-fraud system

The transaction monitoring system (TMS) or anti-fraud system (AFS) is specialized software or hardware-software complex that monitors, detects fraudulent activities, and provides support for decision-making on the detected illegal operation.

The most promising solution for today is the use of technologies to define the user environment in combination with the methods of machine learning. The use of machine learning is a necessary metrics, since a large amount of information about the user environment is collected and the application of rules to this data becomes impossible. Classic methods for detecting fraudulent actions can not sufficiently accurately answer the question: was this action actually performed by the user? There are many ways to obtain illegal access to a user's account: phishing, vishing, pharming, mobile fraud, and other methods related to social engineering techniques [10, 11, 12].

To analyze the large volumes of data collected about the user environment, it is advisable to use approaches based on data mining. The application of the data mining algorithms for solving the problem of fraudulent transaction recognition and analysis of the user environment data is given in [11, 12, 13, 14]. The greatest efficiency is achieved using a combination of various algorithms (stacking-bagging) and the use of large data technologies (Hadoop Processing) [11]. This is explained by the fact that in pure form these algorithms are no longer capable of solving existing problems in view of the increasing volumes of processed data. There is a need to modify these algorithms, combine them to obtain an acceptable result, and also apply technologies capable of processing huge amounts of accumulated information.

The goal of this research is development of the infrastructure for collection and analysis of user environment as a part of anti-fraud system on the basis of big data processing technologies.

To achieve this goal, the following tasks were formulated:

- Development of the structure of the system for collecting and analyzing information about the user environment based on DM techniques.
- Development of a structural and functional scheme of processing user environment data for testing algorithms of detecting financial fraud on the basis of DM as a part of the system of distributed processing of banking transaction data.

3. System structure for collecting and analyzing user environment information

Technologies of remote banking service (internet banking) for accessing accounts and operations through a web browser do not require installation of the client part software and have become very widespread [15, 16]. The user makes certain manipulations in a web browser that interacts with the Frontend server interface. Frontend server generates a set of data about the user environment and transfers data about the user's actions about transaction initializing to the Backend server of remote banking system (RB) and then to the automated banking system (ABS) for the calculations [15, 16, 17]. Backend server transfers transaction data and collected data about the user's environment for analysis to the anti-fraud system. If the legitimacy of the transaction is recognized, data is transferred to the ABS server, otherwise the Backend server refuses user to perform the transaction.

Anti-fraud system evaluates risk of the current transaction and, in case of exceeding a certain threshold value, triggers additional mechanisms for verifying the legitimacy of the transaction [18]:

- automated ways of additional transaction authentication
  - SMS / push-notification;
  - request to answer the test questions;
- manual ways of additional transaction authentication
  - a phone call from a security specialist to the user.

In this architecture, user session management module (USMM) is the main element of the anti-fraud system. The module analyzes the transaction data and the user environment data (UED) collected by the client-side script. The implemented script forms a set of data about the user environment [19]. The generalized structure of collected UEDs is the following:

- Color depth;
- Document size;
- Screen size;
- Time zone offset;
- Fonts;
- Plugins;
- IP-address;
- Number of processor cores;
- UserAgent.

The module’s main task is to classify the current session and its transaction (legitimate transaction or fraudulent actions) based on the composition of the analysis methods: signature and automatic. If the module's estimated legitimacy of the transaction is below the threshold, an additional mechanism for authenticating the user is triggered.

The signature analysis module allows the use an expert knowledge and their formalization in the form of a system of production rules ”IF-TO”. The main task of this module is to classify UED and/or existing data about banking transaction in order to detect fraudulent activities. A special feature of the module is the use of a unified signature database based on the system of production rules, which allows to integrate mechanism for explaining the decision in an understandable form for security expert into the system. Initially, the database of the signature analysis module contains typical templates of fraudulent and legitimate transactions and UEDs in appropriate cases. Replenishment of the signature database is possible both in manual mode through the interface of the manual analysis module, and with the help of the automatic analysis modules which detects signatures in the new data processing mode. Analysis of the accumulated data under control of the analyst makes it possible to identify new production rules in an automatic mode based on the DM technologies.

Task of the automatic analysis module is automatic data classification based on the DM methods, for example, using a neural network classifier trained on a database of committed transactions marked by an expert [11]. At the stage of data preparation for DM algorithms application, the following tasks must be solved:

- features selection – selection of the most significant features for the adoption of a classification decision.
- features transformation – filling gaps in data, removing emissions and filtering noise components.
- features extraction – transformation of the selected features into a new feature space for feeding to the classifier input;

In the previous works [11, 12] the original feature space included 40 parameters. After evaluating available parameter values and its distributions, 12 parameters were deleted from the original data. As a result of expert analysis of the feature space transformation results, new nonlinear features were added – combinations of the initial features, characterizing possible combinations of some available parameters.

The core of automatic analysis module is the neural network – the “black box” – which allows to assign the current vector of features extracted from the data collected about the user environment to one of the previously defined classes. The following classes are proposed:

- User system is under remote control;
- The user system uses action anonymization mechanisms;
- User system does not contain any suspicious elements.

Results of the signature and automatic analysis modules are comparing. If the verdicts do not coincide, then an anti-fraud system expert can be brought in for manual analysis of the precedent. While the accumulation of data on transactions and UED occurs, the neural network classifier is after-trained. While group of use cases, the parameters of which do not fit into the current scheme of “IF-TO” signature rules, are being formed, new rules are being extracted, replenishing the existing signature database. If new signatures are added by the expert, then the current base of marked use cases is analyzed in order to update the class labels and retrain the neural network classifier.

Manual analysis module is designed to correct the markup of the existing database of the user environment and transactions in order to form a training sample for the neural network classifier. The module allows expert to analyze the system solution for each of the use cases and correct it in case of erroneous operations.

The AFS control modules allow to monitor the basic performance of AFS, analyze the log of the system and debug the interaction of the signature and automatic modules.

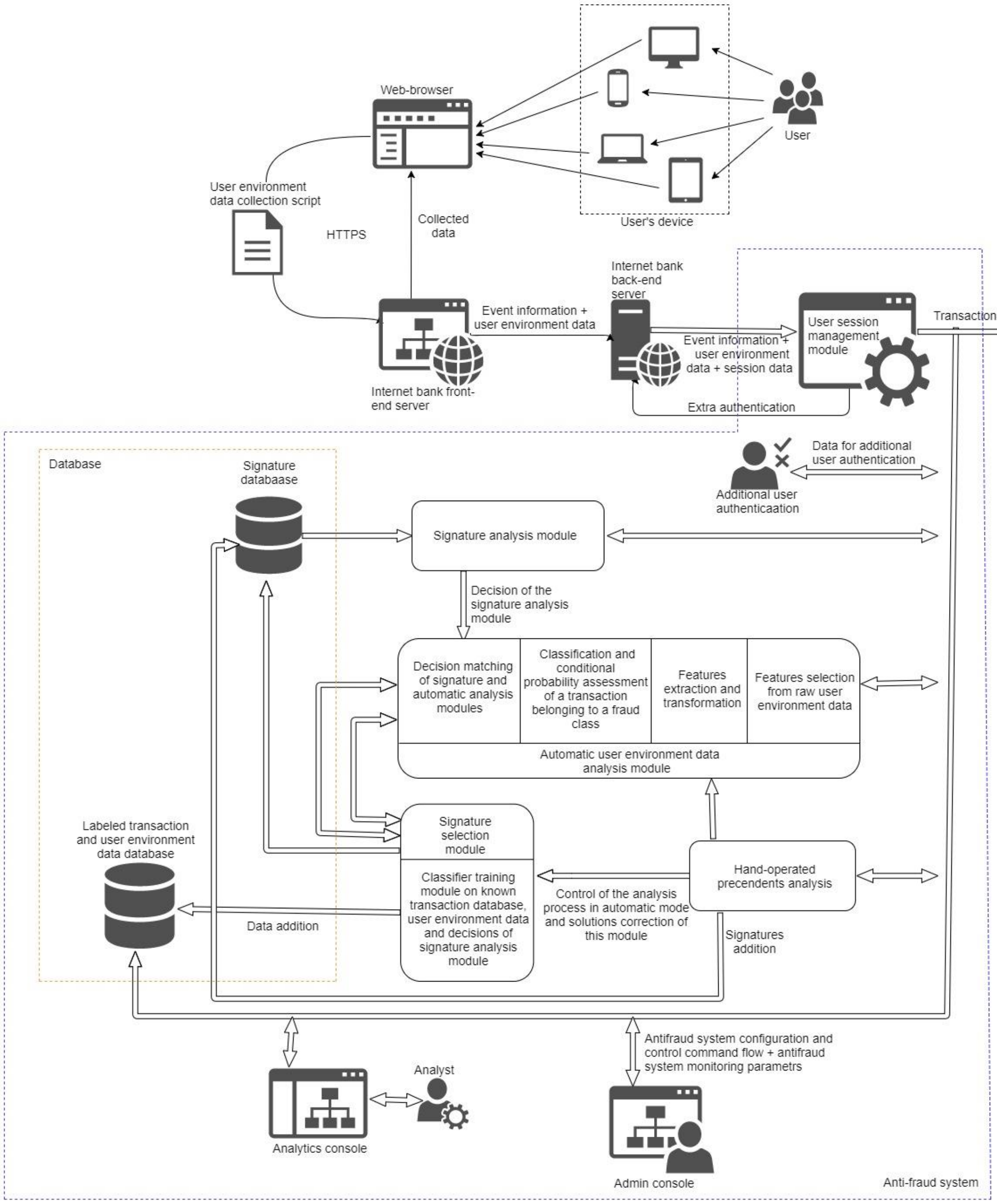
Web-browser
User
User's device
User environment data collection script
HTTPS
Collected data
Internet bank back-end server
Automated bankig system
Event information + user environment data
Event information + user environment data + session data
User session management module
Transaction
Extra authentication
Internet bank front-end server
Database
Signature databaase
Data for additional user authentication
Additional user authenticaation
Signature analysis module
Decision of the signature analysis module
Decision matching of signature and automatic analysis modules
Classification and conditional probability assessment of a transaction belonging to a fraud class
Features extraction and transformation
Features selection from raw user environment data
Automatic user environment data analysis module
Labeled transaction and user environment data database
Signature selection module
Classifier training module on known transaction database, user environment data and decisions of signature analysis module
Hand-operated precendents analysis
Control of the analysis process in automatic mode and solutions correction of this module
Data addition
Signatures addition
Antifraud system configuration and control command flow + antifraud system monitoring parametrs
Analyst
Analytics console
Admin console
Anti-fraud system

Figure 1. Structure of the user environment data and transaction analysis system within the AFS.

Thus, structure of the system for collection and analysis of user environment as a part of anti-fraud system is presented in figure 1. The key element of the system is data mining module. Analysis algorithms should be applicable in the context of “Big Data” (a set of approaches, tools and methods for processing structured and unstructured data of huge volumes and significant diversity for obtaining human-readable results that are effective under conditions of continuous growth, distribution over multiple nodes of the computer network) [20].

4. Designing a structural and functional scheme for processing “big data” of the user environment as a part of a distributed data processing system for bank transactions

The implementation of algorithms for detecting financial fraud based on data mining techniques of banking transactions as part of a distributed processing system of banking transactions requires the solution of a few tasks related to the design and deployment of an appropriate infrastructure for storing and processing accumulated data.

To date, there are many tools for the distributed processing of banking transaction data (frameworks: Hadoop, Apache Spark, ClickHouse, ElasticSearch, Splunk Free) [21, 22, 23, 24, 25]. The proposed structure of the distributed processing system of banking transaction data is presented in the figure 2.

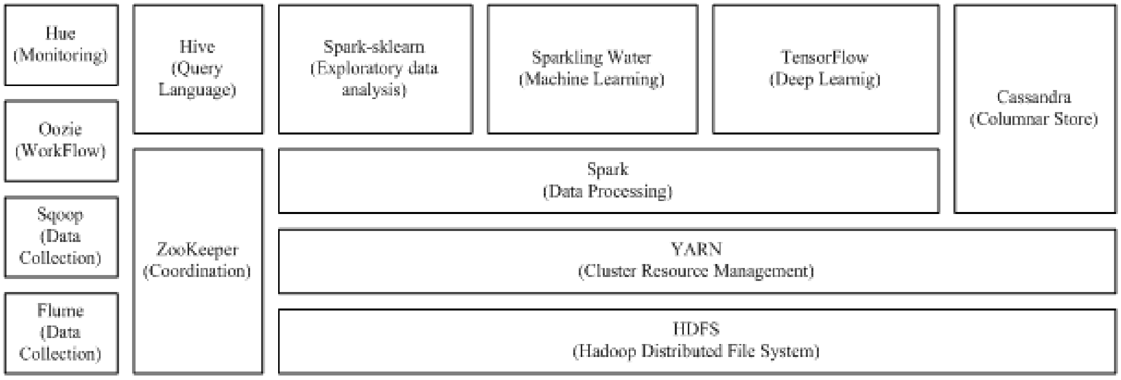
Hue (Monitoring)
Oozie (WorkFlow)
Sqoop (Data Collection)
Flume (Data Collection)
Hive (Query Language)
ZooKeeper (Coordination)
Spark-sklearn (Exploratory data analysis)
Sparkling Water (Machine Learning)
TensorFlow (Deep Learnig)
Cassandra (Columnar Store)
Spark (Data Processing)
YARN (Cluster Resource Management)
HDFS (Hadoop Distributed File System)

Figure 2. Hadoop cluster structure

Main element of the distributed processing system of banking transaction data is a distributed file system. Nowadays the most popular distributed file system is HDFS [21].

Next element of the big data processing system is the distributed programming and machine learning infrastructure. The core of this element is Spark, a cluster computing infrastructure similar in MapReduce [21]. The structure of this infrastructure includes the machine learning tool MLLib, which allows to implement the DM tools of the accumulated data.

To directly store the accumulated data, it is proposed to use solutions from the New-SQL family [26].

A detailed description of the additional elements and their functions in the distributed data processing system for bank transactions is presented in the table 1.

Table 1. Hadoop ecosystem applications

Function	Tools name	Short description
Store	HDFS[19] Cassandra [24]	Distributed file system NoSql database management system
Cluster resource management	YARN [19]	Operating system for big data application
Data processing	Spark [20]	Engine for big data processing
Machine learning	Spark-sklearn [25]	Scikit-learn python library integrated in Apache Spark for exploratory data analysis
	Sparkling Water [26]	H2O library integrated in Apache Spark for machine learning in Hadoop system
	TensorFlow [27]	TensorFlow library integrated in Apache Spark for deep learning in Hadoop system
Coordination	Zookeeper [28]	Application for maintaining configuration information, naming and etc
Data Access	Hive [29]	Application for data summarization, SQL-like query, and analysis
Data Collection	Sqoop [30]	Application for transferring data between relational

		databases and Hadoop
	Flume [31]	Application for transferring data between relational databases and Hadoop
WorkFlow	Oozie [32]	Application for collecting, aggregating, and moving of unstructured data
Monitoring	Hue [33]	Web interface to monitor Hadoop system

The use of three machine learning libraries is due to the need of rapid prototyping of the developed algorithms, debugging on small amounts of data and the possibility of importing the developed models.

The hardware-software stand structure for testing the algorithms of detecting financial fraud based on DM techniques in the distributed data processing system for bank transactions based on the selected big data processing stack is shown on figure 3.

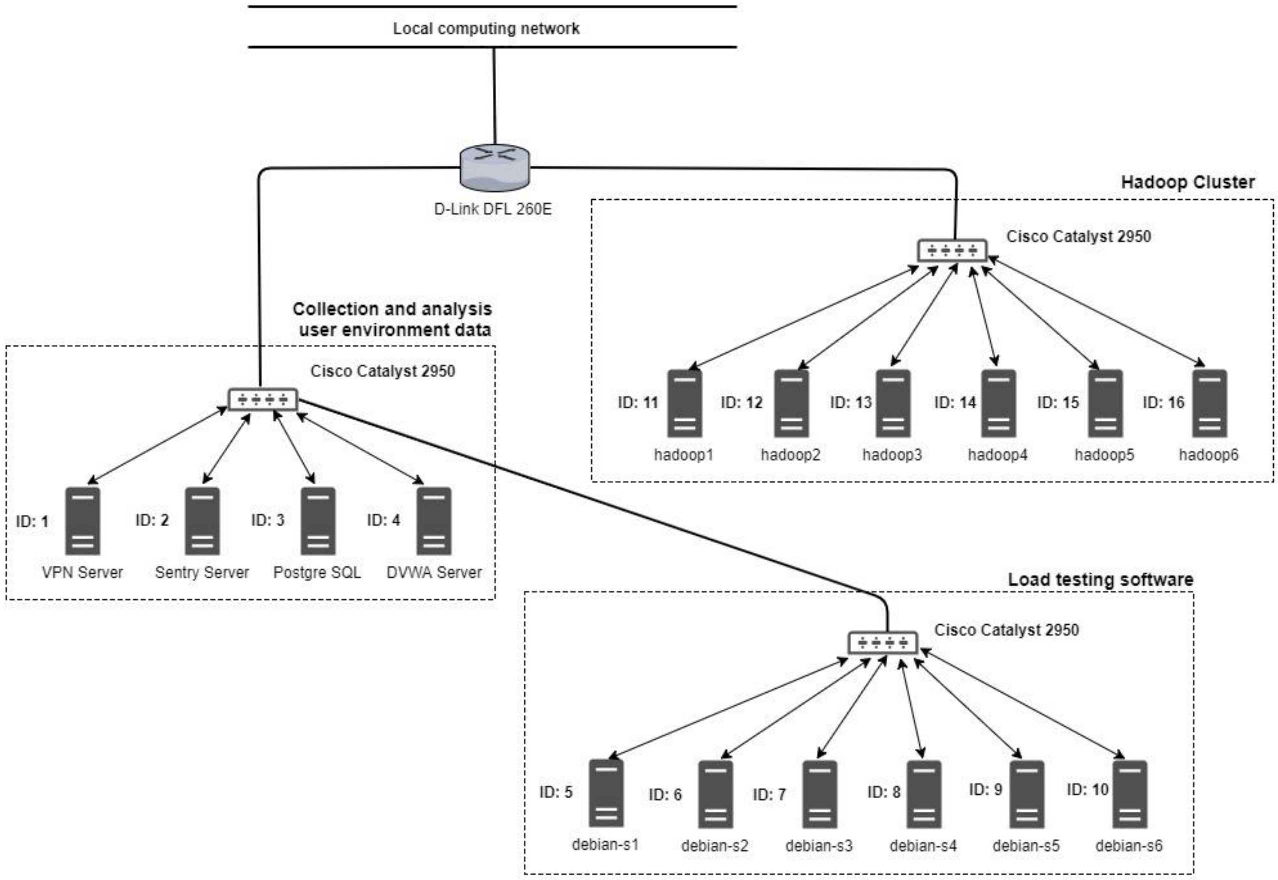

Figure 3. The hardware-software stand structure testing algorithms for detecting financial fraud.

The data collection and analysis module consists of:

- Sentry software suite [36] for collection client-side script logs;
- Gitlab service [37] for organization joint work on the implemented analysis algorithms source code;
- DVWA (Damn Vulnerable Web Application) for testing the script of user environment data collection.

The load testing module is designed to automate collection of the user environment database.

For distributed processing of bank transaction data, the Hadoop cluster is used, on which the software is deployed from Table 1. Typical configuration of the used server machine pool is shown in Table 2.

Table 2. Server settings.

ID	Sector	Configuration	OS
1	Collection and analysis user environment data	2x3.4 GHz / 4GB	debian 8.2
2		2x3.4 GHz / 2GB	debian 8.2
3		2x3.0 GHz / 1GB D	ubuntu 16.04
4		1x2.4 GHz / 1GB	ubuntu 16.04
5	Load testing software	2x3.0 GHz / 3GB	debian 9.2
6		2x3.0 GHz / 3GB	debian 9.2
7		2x3.0 GHz / 4GB	debian 9.2
8		2x3.0 GHz / 3GB	debian 9.2
9		2x3.0 GHz / 4GB	debian 9.2
10		2x3.0 GHz / 3GB	debian 9.2
11	Hadoop Cluster	2x3.4 GHz / 12GB	ubuntu 14.04
12		2x3.4 GHz / 12GB	ubuntu 14.04
13		2x3.4 GHz / 8GB	ubuntu 14.04
14		2x3.4 GHz / 6GB	ubuntu 14.04
15		2x3.2 GHz / 6GB	ubuntu 14.04
16		2x3.0 GHz / 6GB	ubuntu 14.04

5. Conclusion

The main problem of improving the TMS efficiency is the insufficient amount of registered parameters transferred from an online banking client side to a processing center, and the imperfection of signature analysis methods and algorithms because of low adaptability and configuration flexibility.

Nowadays the most promising solution is the use of technologies for determining user environment in combination with the methods of machine learning within TMS. The use of machine learning is an indispensable criterion, because big amount of user environment information being collected and application of the rules to this data becomes difficult. Algorithms for analysis should be applicable in the context of “big data”.

In the paper the infrastructure for collection and analysis of user environment as a part of anti-fraud system on the basis of big data processing technologies is proposed.

6. References

[1] Teoh C S and Mahmood A K 2017 National cyber security strategies for digital economy *Int. Conf. on Research and Innovation in Information Systems* 1-6

[2] Crabtree A 2016 Enabling the new economic actor: Personal data regulation and the digital economy *Proc. IEEE Int. Conf. on Cloud Engineering Workshops* 124-129

[3] Tung H H, Cheng C C, Chen Y Y and Chen Y F 2016 Binary Classification and Data Analysisfor Modeling Calendar Anomalies in Financial Markets *7th Int. Conf. on Cloud Computing and Big Data* 116-121

[4] Trelewicz J Q 2017 Big Data and Big Money: The Role of Data in the Financial Sector *IT Prof.* **19(3)** 8-10

[5] Luvizan S S, Nascimento P T and Yu A 2017 Big Data for innovation: The case of credit evaluation using mobile data analyzed by innovation ecosystem lens *Proc. Portland Int. Conf. on Management of Engineering and Technology: Technology Management For Social Innovation* 925-936

[6] Sapozhnikova M U, Gayanova M M, Nikonov A V and Vulfin A M 2017 Data mining algorithms of bank transaction as a part of antifraud system *Proc. Conf. Information Technologies for Intelligent Decision Making Support* **2** 143-149

[7] Lopez-Rojas E A and Axelsson S 2016 A review of computer simulation for fraud detection research in financial datasets *Proc. of Future Technologies Conf.* 932-935

[8] Dodds L S 2014 Big ideas are coming from using big data (Access mode: https://www.raconteur.net/technology/big-ideas-are-coming-from-using-big-data)

[9] Rosenbush S 2014 Visa Says Big Data Identifies Billions of Dollars in Fraud *CIO J.* (Access mode: https://blogs.wsj.com/cio/2013/03/11/visa-says-big-data-identifies-billions-of-dollars-in-fraud)

[10] Piskunov I 2017 Anti-fraud systems and how it works *J. Securitylab* (Access mode: https://www.securitylab.ru/blog/personal/Informacionnaya_bezopasnost_v_detalyah/ 339929)

[11] Sapozhnikova M U, Gayanova M M, Nikonov A V and Vulfin A M 2017 Data mining technologies in the problem of designing the bank transaction monitoring system *ComputerScience and Information Technologies* 45-56

[12] Sapozhnikova M U, Gayanova M M, Nikonov A V, Vulfin A M and Kurrenov D V 2017 Anti-fraud system on the basis of data mining technologies Int. *Symp. on Signal Processing and Information Technology* 1-5

[13] Vizilter, Yu V, Gorbatsevich V S, Vorotnikov A V and Kostromov N A 2017 Real-time face identification via CNN and boosted hashing forest *Computer Optics* **41(2)** 254-265 DOI: 10.18287/2412-6179-2017-41-2-254-265

[14] Ivanov A I, Lozhnikov P S and Sulavko A E 2017 Evaluation of signature verification reliability based on artificial neural networks, Bayesian multivariate functional and quadratic forms *Computer Optics* **41(5)** 765-774 DOI: 10.18287/2412-6179-2017-41-5-765-774

[15] Abbad M, Abed J M and Abbad M 2012 The Development of E-Banking in Developing Countries in the Middle East. *J. Financ. Account. Manag*. **3** 107-123

[16] Jarrett J E 2016 Internet Banking Development *J. Entrep. Organ. Manag*. **5** 2-5

[17] *Global Mass Payments, AP Software, B2B Payments* (Access mode: https://tipalti.com/)

[18] Fedotenko M 2017 How banks are protected: explaining the structure and the principles of a bank antifraud system *J. Hacker* (Access mode: https://xakep.ru/2017/04/21/antifrod-1/)

[19] Cao Y, Li S and Wijmans E 2017 (Cross-) Browser Fingerprinting via OS and Hardware Level Features *Proc. Network and Distributed System Security Symp.* 1-15

[20] *Big Data* (Access mode: https://en.wikipedia.org/wiki/Big_data/)

[21] *Apache Hadoop* (Access mode: http://hadoop.apache.org/)

[22] *Apache Spark* (Access mode: https://spark.apache.org/)

[23] *Click House* (Access mode: https://clickhouse.yandex/)

[24] *Elastic search* (Access mode: https://www.elastic.co/products/elasticsearch/)

[25] *Splunk* (Access mode: https://www.splunk.com/)

[26] *Apache Cassandra* (Access mode: http://cassandra.apache.org/)

[27] *Spark-sklearn* (Access mode: https://github.com/databricks/spark-sklearn/)

[28] *Sparkling Water* (Access mode: https://www.h2o.ai/sparkling-water/)

[29] *Tensor Flow On Spark* (Access mode: github.com/yahoo/TensorFlowOnSpark/)

[30] *Apache Zookeeper* (Access mode: https://zookeeper.apache.org/)

[31] *Apache Hive* (Access mode: https://hive.apache.org/)

[32] *Apache Sqoop* (Access mode: http://sqoop.apache.org/)

[33] *Apache Flume* (Access mode: https://flume.apache.org/)

[34] *Apache Oozie* (Access mode: http://oozie.apache.org/)

[35] *Cloudera Hue* (Access mode: https://github.com/cloudera/hue/)

[36] *Sentry* (Access mode: https://sentry.io/welcome/)

[37] *GitLab* (Access mode: https://about.gitlab.com/)

Acknowledgments

Work was supported by the Russian Foundation for Basic Research, research № 17–48–020095.