Big Data Technologies for Cybersecurity

Sergei A. Petrenko
Information Security Department
Saint Petersburg Electrotechnical University "LETI"
St. Petersburg, Russia
s.petrenko@rambler.ru

Krystina A. Makoveichuk
Department of Informatics and Information Technologies
Vernadsky Crimean Federal University
Yalta, Russia
christin2003@yandex.ru

***Abstract*—The article presents variants for building a cognitive early warning system about a computer attack on the information resources of the Russian Federation on the basis of Big Data technologies. The essence of Big Data technologies is considered in the context of its application in information security. Approaches for stream processing of data based on the CEP model, the MapRedutze modification, the model of actors, the combination of the model of actors and the modification of MapRedutze are analyzed. The architecture of the prototype of the software-hardware complex "Warning-2016", a typical scheme of the hardware implementation of the stand and the technical specification of its equipment are presented.**

Keywords—Big Data; cybersecurity, streaming data processing, actor model, modification of MapReduce, software and hardware complex (SHC).

I. INTRODUCTION

At present, the technology of processing, storage and analysis of large data, Big Data (hereinafter - the technology of Big Data) are becoming increasingly important for the state of critical infrastructure security monitoring of the Russian Federation (electrical networks, pipelines, communication systems, and so forth.) and monitoring the corresponding criteria and indicators of information security and sustainability of functioning in general [3].

Big Data technologies are already being used in a number of cybersecurity applications. For example, in Security information and Event Management systems (SIEM) non-relational database (NoSQL) are used to store logs, messages and security events. In the near future, a qualitative leap in the development of SIEM is expected on the basis of models and methods of forecast analytics. In the known solutions of Red Lambda, Palantir, etc., Big Data technologies are used to build user profiles and social groups in order to detect abnormal behavior [1, 4, 9, 16]. At the same time, the following information sources serve as information sources: corporate mail, CRM system and personnel system, access control system (ACS), as well as various pullers in connected data networks Inetrnet / Intanet and IIoT / IoT, external news tapes, collectors and aggregators in in social networks [24].

The relevance of Big Data technologies is confirmed by the fundamental possibility to conduct "online analysis" of packet and streaming data, to isolate and process significant simple and complex cybersecurity events in real (or quasi-real) time scale, and to generate new useful knowledge for detection and prevention of security incidents. It is significant that Big Data is able to provide proactive security and monitor the impending information security incidents even before they can adversely affect the sustainability of the critical infrastructure [3, 23].

Thus, by Big Data technologies in information security we will understand the technologies of efficient processing of dynamically growing data volumes (structured and unstructured) in heterogeneous Internet / Itranet and IIoT / IoT systems for solving urgent security tasks. The practical significance of Big Data technologies lies in the ability to detect primary and secondary signs of preparation and conduct of computer attacks, the detection of abnormal behavior of controlled objects and subjects, the classification of previously unknown mass and group cyber attacks (including new DDOS and APT), the detection of the traces of computer traces crimes, etc., that is, in all cases when the use of traditional means of information protection (SIEM, IDS / IPS, system of protection from unauthorized access to information, cryptographic information protection facility, antiviruses, etc.) is not very effective [4, 5, 7, 9, 18, 22, 23].

II. COMPARATIVE ANALYSIS OF BIG DATA

Currently, the following approaches are known for streaming data processing [2, 6, 10-15, 19-21] based on:

- Classical CEP model, for example, StreamBase;
- Modification of MapReduce, for example, D-Streams;
- Actor model, for example, Storm, S4 and Zont;
- Combinations of the actor model and the modification of MapReduce, for example, Zont + RTI.

In the first approach, the classic Complex Event Processing (CEP) model is used. The use of CEP allows you to search for "significant" cybersecurity events in a data stream over a certain time interval, perform a correlation analysis of events, and allocate appropriate event patterns that require immediate response.

To automate the process of developing data processing systems based on CEP, a number of tools are proposed (f. e., StreamBase with its own declarative programming languages StreamSQL and EventFlow).

The practice of using CEP has shown that it is optimal for the collection and processing of simple cyber-security events. For example, to extract events from several data streams, aggregate them into complex events, reverse decomposition, etc. However, the implementation of complex logic for handling cybersecurity events is difficult. To solve this

problem, we proposed an approach based on the generalization of MapReduce to the processing of streaming data.

In the second approach, the D-Streams model of discrete streams is used, in which streaming computations are presented as sets of non-session deterministic batch calculations on small time series intervals. It is significant that such a representation of calculations allowed not only to implement the complex logic of processing cybersecurity events, but also to offer better methods of restoration than traditional replication and backup copying. The fact is that in practice, in computer networks with a large number of nodes (from hundreds or more), failures and "hangs" (or "slow" nodes) inevitably occur, and here the operative data recovery in case of failure or failure is important enough. Since, even a minimum delay of 10-30 seconds can be critical for making the right decision.

It should be stated that, apparently, for the known systems the streaming data Storm, MapReduce Online et al. resiliency reached threshold values. The systems mentioned are based on the model of "long-lived" session operators, which, upon receiving the message, update the internal state and send a new message further. In this case, the system is restored by replicating to a pre-prepared copy of the node or by backing it up in a data stream, meaning "replay" of messages on each new copy of the "fallen" node.

As a result, the use of the replication mechanism results in a costly two or three times the node reservation, and the use of a backup in the data stream is characterized by significant time delays due to the need to wait for the nodes to "update" when the data is re-run through the operators. In addition, none of these approaches can not cope with "hangs". Replication systems use Flu synchronization protocols to coordinate replicas and hangs slow both replicas. When backing up, any "hang" is considered a failure with subsequent costly recovery.

The D-Stream model offers better recovery methods. For example, the Resilient Distributed Datasets (RDD) recovery method, which allows you to restore data directly from memory without having to replicate for several sub-seconds, or a method of parallel restoration of the state of a "lost node" in which, when the node falls "It initiates the" connection "of the workable nodes of the cluster to the" recalculation of the lost "structure of the RDD. Note that in traditional systems of continuous data processing such restoration is impossible due to complex synchronization protocols.

Note that using the D-Streams model requires splitting an array of input data into streams, which inevitably results in the loss of certain events. In addition, in the case of large flows, the data processing system is no longer flexible and scalable. The time the system responds to events slows down and the system moves further away from the real-time mode. To solve these problems was proposed third approach - based on actor model.

In the third approach, the streaming data processing systems are based on the actor model. Here, actors are understood as some primitives of parallel computations. The main advantage of the actors is the ability to store states, including those obtained from historical data, which can be used to highlight significant cyber security events. Among the known solutions of the streaming data based on actor model allocated Storm system (Twitter), and S4 (Yahoo!) and Zont (Moscow Institute of Physics and Technology, MIPT).

The first two solutions, Storm and S4 [8, 17, 21] make it possible to implement the so-called pipeline data processing based on a relatively small number of actors.

The third of the named Zont system can work with a large number of actors. This is true when working with a cloud of sensors, when each sensor is assigned its own actor. To develop distributed resilient systems, it is possible to use the Erlang & RIAK Core development environment. Here, the functional language Erlang (Ericsson) allows you to create programs that can work in a distributed computing environment on several nodes (processors, cores of one processor, cluster of machines), and the open library Riak Core (Basho Technologies) allows to create distributed applications according to Amazon's Dynamo architecture.

Thus, all three systems, Storm, S4 and Zont (MIPT) can be used to process a large data stream from several sources. In this case, Zont is optimally suited for working with a cloud of sensors.

In the fourth approach, the combined advantages of the second and third approaches, which allows the system to create the streaming data in real time based on the combination and modification MapReduce actor model.

III. EXAMPLE OF A SOLUTION BASED ON BIG DATA

Consider the possible options for building a cognitive early warning system about a computer attack on the information resources of the Russian Federation (software and hardware complex, SHC "Warning-2016") on the basis of Big Data technologies.

Variant 1. Implementation of the experimental model of the SHC "Warning-2016" based on HBase.

Here the basis for the proposed solution was the non-relational distributed database HBase, working on top of the HDFS file system (Hadoop Distributed File System).

This database allows you to perform analytical and predictive operations on terabytes of data to assess the threats to cybersecurity and the stability of the critical infrastructure as a whole. It is also possible to prepare in the automated mode appropriate scenarios for detection, neutralization and warning.

The second hypothesis analysis module is designed to handle large amounts of data, respectively, from it require high performance. The module interacts with standard configuration servers and is implemented in C language (via PECL, PHP extensions repository). Special interactive tools based on JavaScript / CSS / DHTML and libraries such as jQuery have been developed to work with the content of the proper provision of cybersecurity.

As a data store, MySQL is used - Percona Server (version 5.6) with the XtraDB engine. DB servers are integrated into a multi-master cluster using the Galera Cluster. For balancing the database servers, haproxy is used. Redis (version 2.8) is used to implement task queues, as well as for data caching.

As a web server, nginx is used. Involved PHP-FPM with APC enabled. For balancing HTTP requests, DNS (multiple A-records) is used. To develop special client applications running Apple iOS programming languages are used: Objective C, C ++, and Apple iOS SDK based on Cocoa Touch, CoreData, UIKit. The above applications are compatible with devices running Apple iOS version 9 and above.

To develop applications running Android OS, the native Google SDK is used. The applications are compatible with devices running Android OS version 4.1 and higher. Software development for the web platform is carried out using PHP and JavaScript.The experimental sample SHC "Warning-2016" is deployed in Saint Petersburg Electrotechnical University "LETI" on the platform DigitalOcean and contains in its composition:

- 3 servers for production stage;
- 1 server for testing stage.

In addition, CloudFlare is used - to increase the speed of the service (through the use of CDN) and protection from DoS attacks. The carried out load testing of the experimental model of the SHC "Warning 2016" indicates the viability of the proposed technical solution [17].

Variant 2. Implementation of the prototype of the SHC "Warning-2016" on the basis of the telematics platform Zont (MIPT)The possible architecture of the experimental layout of the SHC "Warning-2016" based on Zont is presented (see Fig. 1). Here the basis of the proposed solution was the telematics platform Zont, which allows creating resiliency scalable cloud systems for streaming large data. Table 1 describes the modules of the experimental layout of the SHC "Warning-2016" based on Zont.

It is significant that Zont has its own specialized storage, built on the basis of Riak Core technology, for storing and retrieving archived data that represent time series. The backend for the repository is LevelDB, which is developed by Google. LevelDB is an embedded KV database specifically designed for use as a backend in the construction of specialized databases, providing operations for writing, searching and sequential data viewing.

The key advantages of the database are high speed of data recording, predictable speed of data search by key and high speed of sequential reading. Also worth noting the following important characteristics of LevelDB for the storage of time series:

a) Use of the LSM tree model, which allows providing high resistance to failures and failures;

b) Organization of data storage in an ordered form.

TABLE I. DESCRIPTION OF THE DATA PROCESSING MODULES

Name of the element (name in the figure)	Functionality of the element
TCP-Balancer	Network load balancing between cluster members.
TCP session manager (Socket server process)	- Separate control of the network connection for each individual sensor; - Preliminary check of integrity of the incoming data.
Transactional buffer (TX buffer)	- Buffering input data to optimize performance at peak loads; - Calling the data parsing procedure from the sensors; - Redirect data to the corresponding FSM.
The state machine sensor (Sensor FSM)	- Dedicated FSM process for each sensor; - Logical data processing from the sensor; - Track events within a single detector; - Saving the processed data to the database;
NoSql Distributed KV Storage	- Saving the processed data; - Increasing the reliability of writing and reading data by repeatedly storing data on different nodes.
Analytical data processing module	- Analytical data processing, tracking complex events; - Generate reports based on the content of the database.
Module for interaction with client applications	- Interaction with external clients using REST and WebService.

At the same time, geo-index support based on geo-hash technology was added to store the input data of mobile sensors. This index is specifically designed to store a large array of information about the spatial position of point objects, in particular, moving sensors.

As a result, it allowed to obtain better performance indicators compared to such well-known universal solutions as Postgres GiST and MongoDB 2dsphere. The hardware implementation of the prototype SHC "Warning-2016" is a cluster of general-purpose servers connected by a network.

The hardware implementation of the prototype SHC "Warning-2016" is a cluster of general-purpose servers connected by a network (Table 2).

A typical scheme of the hardware implementation of the stand is presented (see Fig. 2).

The stand includes the following components:

- Erlang Virtual Server (Erlang Node Server) to implement the distributed cloud platform module;
- Erlang Virtual Server (Erlang Node test Server) to implement test module, including sensor network emulator;

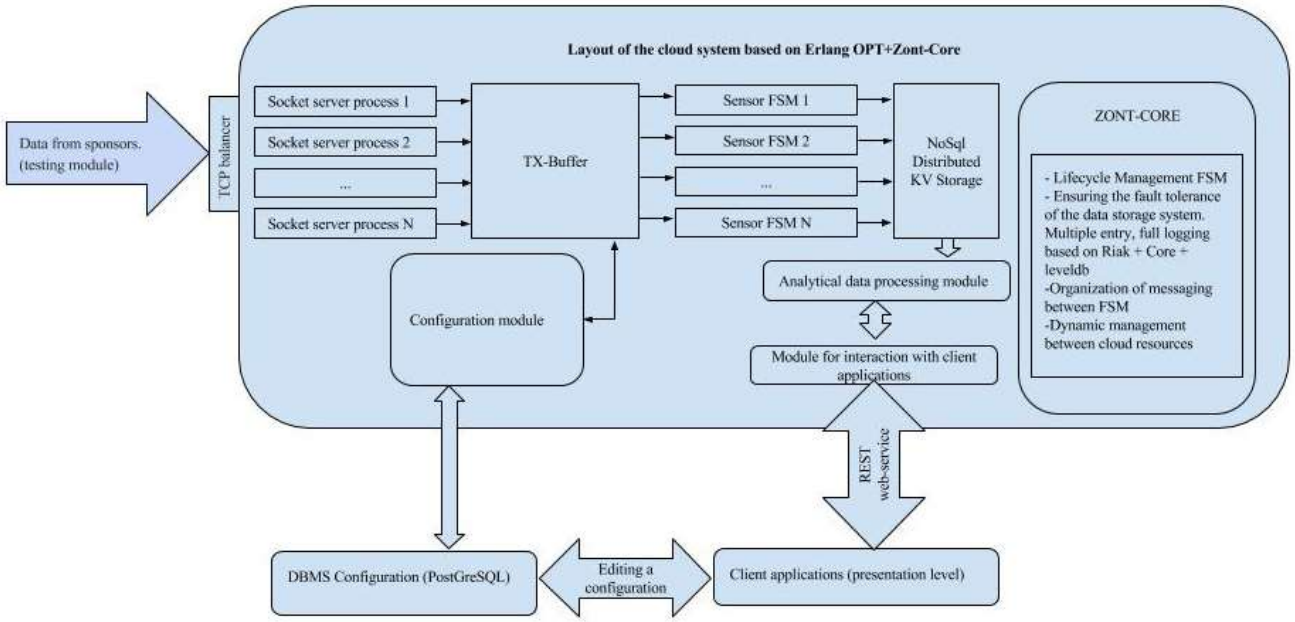

Fig. 1. Possible architecture of the SHC "Warning-2016" based on Zont

• Auxiliary software server of the Platform, which contains the JBoss application server and PostgreSQL DBMS.

TABLE II. TECHNICAL SPECIFICATION OF STAND EQUIPMENT

Equipment	Standard server cluster	High-performance cluster "IScalare"
The number of servers allocated for the layout of the Platform	6	200
CPU	2x Intel Xeon E5-26206-core 2,0 ГГц	2 x Intel Xeon E5-2690 (8-core) 2,9 ГГц
Memory	32 Гб	64 Гб
Hard disk drives	3Тб	600Гб (SATA)
Network Security	Ethernet 1Gb	InfiniBand – QDR (40Гбит/с)

The system and application software of the stand included:

• Linux OS CentOS 6.4;
• Erlang R160B2 as the execution environment of the distributed machine Erlang;
• HAProxy - as a proxy server,
• JBoss - as an application server;
• PostgreSQL - for storing metadata, etc.

Preliminary tests of the experimental design of the SHC "Warning of 2016" showed its ability to act [17].

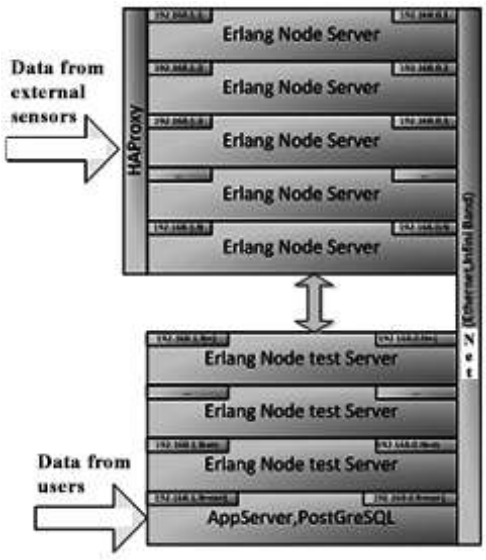

Fig. 2. Software and hardware implementation of the stand

IV. CONCLUSIONS

The obtained positive experience of using Big Data for solving information security problems testifies to the expediency of choosing solutions based on the actor model and Map Reduce technology over the distributed cloud KV data warehouse. At the same time, solutions based on CEP turned out to be less demanding for memory. They allow storing data in a single "window" of events, but they demanded considerable computing resources when analyzing such "windows". And based on the actor model solutions were less demanding of computer resources, but more demanding of memory due to the need to duplicate data for each event / object. Accordingly, the solutions based on the modification of MapReduce took an intermediate position.

In our opinion, the technology Big Data to radically change the situation in the following areas of information security:

- Proactive management of cybersecurity incidents;
- Early detection, prevention and elimination of the consequences of computer attacks;
- Predictive network monitoring of cybersecurity;
- Authentication, user authorization and identity management;
- Preventing computer crime and fraud;
- Information security risk management;
- Compliance with regulatory requirements, etc.

In this case, the first results should be expected precisely in the proactive management of incidents of cybersecurity and early warning of computer attacks. Note that the known results of foreign developers of information protection tools confirm this assumption. For example, in 2015-2016. RSA and IBM announced plans to create a new generation of security management centers (Security Operations Center, SOC) [17], the so-called Intelligence-Driven Security Operations Center, iSOC.

REFERENCES

[1] Armstrong T. G., Ponnekanti V., Borthakur D., Callaghan M. Linkbench: A database benchmark based on the facebook social graph. [Proc. In: 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD '13], New York, 2013. ACM, pp. 1185–1196. DOI: 10.1145/2463676.2465296.

[2] Babcock B., Babu S., Datar M., Motwani R., Widom J. Models and issues in data stream systems, in: 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '02, ACM, New York, USA, 2002, pp. 1–16. DOI:10.1145/543613.543615.

[3] Barabanov A., Markov A., Tsirlov V. Procedure for Substantiated Development of Measures to Design Secure Software for Automated Process Control Systems. In Proceedings of the 12th International Siberian Conference on Control and Communications (Moscow, Russia, May 12-14, 2016). SIBCON 2016. IEEE, 7491660, 1-4. DOI: 10.1109/SIBCON.2016.7491660.

[4] Biryukov D. N., Lomako A. G., Rostovtsev Yu. G. The appearance of anti-cyber systems to prevent the risks of cyber-threat [*Proc. SPIIRAN*]. 2015, V. 39, pp. 5 - 25. DOI: http://dx.doi.org/10.15622/sp.39.1

[5] Borovsky A.S., Ryapolova E.I. Building a model of the protection system in cloud technologies based on a multi-agent approach with the use of automatic model. Voprosy kiberbezopasnosti [Cybersecurity issues]. 2017. No. 4 (22), pp. 10-20. DOI: 10.21681/2311-3456-2017-4-10-20.

[6] Brian F. Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, and Russell Sears. Benchmarking cloud serving systems with ycsb. [Proc. 1st ACM Symposium on Cloud Computing, SoCC '10], New York, 2010, ACM, pp. 143–154. DOI: 10.1145/1807128.1807152.

[7] Gai K., Qiu M., Zhao H. Cost-aware multimedia data allocation for heterogeneous memory using genetic algorithm in cloud computing, IEEE Transactions on Cloud Computing PP (99) (2016) 1–1. DOI:10.1109/TCC.2016.2594172.

[8] Gedik B., Ozsema H., Oztürk O. Pipelined fission for stream programs with dynamic selectivity and partitioned state, Journal of Parallel and Distributed Computing 96 (2016) 106–120. DOI: http://dx.doi.org/10.1016/j.jpdc.2016.05.003.

[9] Ghazal A., Rabl T., Hu M., Raab F., Poess M., Crolotte A., Jacobsen H.-A. Bigbench: Towards an industry standard benchmark for big data analytics. [Proc. ACM SIGMOD International Conference on Management of Data, SIGMOD '13], New York, 2013. ACM, pp. 1197–1208. DOI: 10.1145/2463676.2463712.

[10] Golab L., Ozsu M. T. Issues in data stream management, SIG-MOD Record 32 (2) (2003) 5–14. DOI:10.1145/776985.776986.

[11] Jamshidi P., Casale G. An Uncertainty-Aware Approach to Optimal Configuration of Stream Processing Systems [Proc. In: MASCOTS 2016]. DOI: 10.5281/zenodo.56238.

[12] Jayashree M., Zahoor S. U. H. Beyond Batch Process: A BigData processing Platform based on Memory Computing and Streaming Data // International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization). 2016. - V. 5, I. 10, pp. 1783-1789. DOI:10.15680/IJIRSET.2016.0510013.

[13] Kuhlenkamp J., Klems M., Ross O. Benchmarking scalability and elasticity of distributed database systems. [Proc. VLDB Endow.], 2014, v. 7(12), pp. 1219–1230. DOI: 10.14778/2732977.2732995.

[14] Lachhab F,., Bakhouya M., Ouladsine R., Essaaidi M. Performance evaluation of CEP engines for stream data processing. [Proc. 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech)], Marrakech, 2016. DOI: 10.1109/CloudTech.2016.7847726.

[15] Malewicz G., Austern M. H., Bik A. J. C., James C. Dehnert, Horn I., Leiser N., Czajkowski G. Pregel: A system for large-scale graph processing - "abstract". 2009, pp. 6–6. DOI: 10.1145/1582716.1582723.

[16] Petrenko S.A., Makoveichuk K.A., Chetyrbok P.V., Petrenko A.S. About Readiness for Digital Economy. In Proceedings of the 2017 IEEE II International Conference on Control in Technical Systems, IEEE, CTS, 2017, pp. 96–99. DOI: 10.1109/CTSYS.2017.8109498.

[17] Petrenko S. A., Stupin D. D. Natsional'naya sistema rannego preduprezhdeniya o komp'yuternom napadenii [National system of advance computer attacks alerting]. Innopolis, Afina Publ., 2017. 440 p.

[18] Skatkov A., Shevchenko V. Expansion of reference model for the cloud computing environment in the concept of large-scale scientific researches. Trudy ISP RAN/Proc. ISP RAS, vol. 27, issue 6, 2015, pp. 285-306 (in Russian). DOI:10.15514/ISPRAS-2015-27(6)-18.

[19] Surekha D., Swamy G., Venkatramaphanikumar S. Real time streaming data storage and processing using storm and analytics with Hive. [Proc. International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)], Ramanathapuram, 2016. DOI: 10.1109/ICACCCT.2016.7831712.

[20] Tang Y., Gedik B. Autopipelining for data stream processing, IEEE Transactions on Parallel and Distributed Systems 24 (12) (2013) 2344–2354. DOI:10.1109/TPDS.2012.333.

[21] Tasharofi S., Dinges P., Johnson R. E. Why Do Scala Developers Mix the Actor Model with other Concurrency Models? Proc. In: Castagna G. (eds) Conference proceedings of the 27th European Conference on Object-Oriented Programming, ECOOP 2013, Montpellier, France, July 1-5. Lecture Notes in Computer Science, Berlin, Heidelberg, Springer, vol 7920. DOI: 10.1007/978-3-642-39038-8_13.

[22] Vorobiev E.G., Petrenko S.A., Kovaleva I.V., Abrosimov I.K. Organization of the entrusted calculations in crucial objects of informatization under uncertainty. In Proceedings of the 20th IEEE International Conference on Soft Computing and Measurements (24-26 May 2017, St. Petersburg, Russia). SCM 2017, 2017, pp. 299 - 300. DOI: 10.1109/SCM.2017.7970566.

[23] Vorobiev E.G., Petrenko S.A., Kovaleva I.V., Abrosimov I.K. Analysis of computer security incidents using fuzzy logic. In Proceedings of the 20th IEEE International Conference on Soft Computing and Measurements (24-26 May 2017, St. Petersburg, Russia). SCM 2017, 2017, pp. 369 - 371. DOI: 10.1109/SCM.2017.7970587.

[24] Petrenko A.S., Petrenko S.A., Makoveichuk K.A., Chetyrbok P.V. The IIoT/IoT device control model based on narrow-band IoT (NB-IoT). In Proceedings of the the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (29 Jan.-1 Feb. 2018, Moscow and St. Petersburg, Russia) EIConRus, IEEE, 2018, pp. 950-953. DOI: 10.1109/EIConRus.2018.8317246.