Expert System for Multidimensional Time Series Anomaly Detection

Boris Palyukh[1], Alexander Vetrov[2]

[1] *Tver State Technical University, A. Nikitin emb.,22, Tver, 170026, Russia*
[2] *Tver State Technical University, A. Nikitin emb.,22, Tver, 170026, Russia*

Abstract

The fundamental problem of creating methods and means of ensuring the stability of technological processes subject to random fluctuations is considered. The existing theoretical and methodological apparatus does not allow industrial enterprises in some cases to effectively solve the problems of safety and process control. The article discusses the issues of creating methods and means of building hybrid expert systems to control the evolution of continuous multi-stage processes under conditions of dynamic uncertainty. The measured values of diagnostic variables coming from sensors form a multidimensional information flow describing the state of the technological process. The constant flow of information in the form of fast, changing, unpredictable and unlimited data flows creates the need for incremental approaches to information processing. Perform calculations on it in real time as soon as it becomes available. A method for detecting anomalies in a multisensory distributed measurement system is proposed. It allows you to combine the measurement results obtained from individual sensors. The Esp System expert system is described, which makes it possible to identify anomalies in the flow of the technological process and determine the source of anomalies, including taking into account the uncertainty of technological information.

Keywords

Hybrid expert systems, anomaly detection, multidimensional time series

1. Introduction

The problem of anomaly detection in complex dynamical systems is a developing area of scientific research and is actively discussed in the literature [1, 2, 3, 4]. There are various approaches to its solution, defined by classes of applied problems. When developing a method for detecting anomalies in continuous multistage production systems, it is necessary to take into account the features of the technological process. The state of the technological process S at each moment of time t is characterized by a set of technological parameters (technological variables) obtained from the sensors of the measurement system. If the technological process proceeds normally, the values of the diagnostic variable fluctuate within the specified limits. This indicates that the random process defined by the diagnostic variable is stationary. The output of the diagnostic variable beyond the limit values indicates a malfunction of the process equipment. However, even with normally operating equipment, there may be rare short-term cases when the values of the diagnostic variable go beyond the regulatory limits. Control over the state of the technological process is carried out on the basis of a sequence of measured values of technological parameters. The moment of transition of the technological process to a non-stationary mode determines the bifurcation point, i.e. the moment of transition of the system to a critical state. The task of process safety management is to prevent the system from going into a critical state. The issues of identifying anomalies in the flow of technological data and determining the sources of their formation are discussed in this article. The paper is structured as follows. Section 2 describes the

The work was carried out with the financial support of the Russian Foundation for Basic Re-search (grant no. 20-07-00199)
Russian Conference on Artificial Intelligence (RCAI-2021), October 11–16, 2021, Taganrog, Russia
EMAIL: pboris@tstu.tver.ru (A. 1); vetrov_48@mail.ru (A. 2)
ORCID: 0000-0001-8064-2852 (A. 1); 0000-0002-5092-5680(A. 2)
© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3044/paper07.pdf

method developed by the authors for detecting anomalies in the multidimensional flow of technological data. In the third section, the possibilities of an expert system to identify the source of the formation of abnormal data are considered. In the section 4, the architecture of software and hardware of a distributed multisensory system for processing primary information is considered.

2. Method for detecting anomalies in multidimensional time series

The measurement system includes *D* sensors of technological equipment. For any sensor *d* and for any fixed time points *t*, the sequence of values of technological parameters $y_1^{(d)}, y_2^{(d)}, \dots, y_t^{(d)}$, is a set of sample values of a separate implementation of a random process. When designing a system, technological restrictions are imposed on the values of these parameters in the form of upper y_d^u and lower y_d^l, boundaries. These parameter constraints define the area of the S_w's operational state. Going beyond these limits means the transition of the system to a critical state of S_k. It is obvious that the transition of a technological system from a normal state to a critical one is possible in the presence of a certain stable trend of movement of diagnostic variables beyond the normative boundaries. The most important characteristics of a random process $Y_t^{(d)} = \{y_1^{(d)}, y_2^{(d)}, \dots, y_t^{(d)}, \dots, y_T^{(d)}\}$ are its average value and autocorrelation function. During normal operation of technological equipment, a random process can be considered ergodic. The mean value and the autocorrelation function can be determined by averaging one sampling function over time. For a non-stationary process, the mean value and the autocorrelation function will depend on time. the harbinger of the transition of the system to a critical state is a violation of the stationary process and going beyond the boundaries of a stable regime. To check the stationarity of a random process, it is necessary to make sure that its characteristics do not change with a time shift. If the process is not stationary, then its characteristics calculated at different time intervals will change significantly. The main characteristics, the analysis of which makes it possible to check the weak stationarity of the process, are the average values and the autocorrelation function. Our proposed method for determining the stationarity of a multidimensional random process consists in analyzing its main characteristics calculated on the basis of sample data for short time series. Consider a measurement system consisting of *D* sensors. The set is a set of values of sample functions of a *D*-dimensional random process at time *t*. We will analyze the state of the system based on studying the sequence of observed values of a time series of length *L*. Preliminary preparation of data for analysis includes the following steps. The time series under study is divided into *N* short intervals, called time windows for all sensors *d*. Denote *i* the number of the *i*-th time window for all sensors *d*. We define the length of the time window as *T*. Let's denote the observation number in the window as *t*. We will first calculate the average values and standard deviations of the "normal process" for a sufficiently long sample for each sensor. Let 's denote these values as

$$\bar{y}^d = \frac{1}{T}\sum_{t=1}^{T} y_t^d$$

and

$$s_d = \left(\frac{1}{T-1}\sum_{t=1}^{T}(y_t^d - \bar{y}^d)^2\right)^{0,5}.$$

Time windows can be contiguous or have a certain interval. The state of the system at time t is denoted by

$$y_t = [y_t^1, y_t^2, \dots, y_t^d, \dots, y_t^D,$$

where y_1^d represents the value of sensor *d* ($d = \overline{1;D}$). The set of sensor *d* readings in in the *i*-th time window is denoted as

$$y^{(i,d)} = [y_t^{(i,d)}, y_2^{(i,d)}, \dots, y_T^{(i,d)}].$$

At the first stage, we will bring all sensor readings into a dimensionless form, normalizing their values using the expression

$$z_t^{(i,d)} = \frac{y_t^{(i,d)} - \bar{y}^d}{s_d}.$$

The purpose of this transformation is to eliminate the influence of dimension on the analysis result. As a result of normalization, we get the sequence

$$z^{(i,d)} = [z_t^{(i,d)}, z_2^{(i,d)}, \dots, z_t^{(i,d)}, \dots, y_T^{(i,d)}].$$

The second step is to separate the random fluctuations contained in the sample data. To do this, we apply smoothing using a moving average, which can be determined by the formula

$$c_t^{(i,d)} = \frac{z_{t-p}^{(i,d)} + z_{t-p+1}^{(i,d)} + \dots + z_{t+p-1}^{(i,d)} + z_{t+p}^{(i,d)}}{2p+1}.$$

Where $c_t^{(i,d)}$ is the value of the moving average at time *t*; $2p$+1 is the length of the smoothing interval.

As a result of the transformations performed, we get *N* sequences of smoothed normalized values of the original time series for sensor *d*:

$$c^{(i,d)} = [c_t^{(i,d)}, c_2^{(i,d)}, \dots, c_t^{(i,d)}, \dots, c_T^{(i,d)}].$$

In the third step, we calculate the average values of the time series in each window and get a sequence for each sensor

$$\bar{c}^{(d)} = \left[\bar{c}^{(1,d)}, \bar{c}^{(2,d)}, \dots, \bar{c}^{(i,d)}, \dots \bar{c}^{(N,d)}\right].$$

The grouped average values of the smoothed time series for all sensors form a $D \times N$ matrix of the form

$$C_1 = \begin{bmatrix} \bar{c}^{(1,1)} & \dots & \bar{c}^{(N,1)} \\ \dots & \dots & \dots \\ \bar{c}^{(1,D)} & \dots & \bar{c}^{(N,D)} \end{bmatrix}$$

Denote by Ψ_1 the norm of the matrix C_1 so that $\Psi_1 = \|C_1\|$. We calculate its value and perform data transformation as follows.

Add a new window for all sensors and delete the first one. We repeat steps 1-3 sequentially and get a sequence of values Ψ_1, Ψ_2, ..., Ψ_l, ..., Ψ_L, where *L* is the number of repeated data updates, which is determined experimentally. We check the resulting sequence for the presence of a trend using a modification of the Foster-Stewart criterion. To do this, we determine the values

$$u_k = \begin{cases} 1 & if\ \ \Psi_k > \Psi_{k-1}, \Psi_k - 2, \dots, \Psi_1 \\ 0 & else \end{cases}$$

and

$$v_k = \begin{cases} 1 & if\ \ \Psi_k < \Psi_{k-1}, \Psi_k - 2, \dots, \Psi_1 \\ 0 & else \end{cases}$$

To test the hypothesis about the presence of a trend in the data, we calculate statistics

$$W = \sum_{k=2}^{L} (u_k - v_k)$$

In the absence of a trend, the normalized value of statistics

$$t_W = \frac{W}{\hat{\sigma}_w},$$

where

$$\hat{\sigma}_W = (2 \sum_{k=2}^{L} \frac{1}{k})^{0,5}$$

is approximately described by the Student's distribution with degrees of freedom *df* = *L*.

As noted above, to check the stationarity of the process, it is necessary to check the absence of a trend not only of average values, but also of autocorrelation functions. Assuming that the mean value of the square or variance of the process under study is stationary, and the autocorrelation function is also stationary. The problem of identifying the stationarity of the autocorrelation function is reduced to the problem of checking the stationarity of the mean value of the square of the magnitude z - z^2. The testing procedure is in many ways similar to the procedure for checking the stationarity of the average *z* value. Step 1 remains unchanged, step 2 is skipped, and in the third step we calculate the average values of the square of the magnitude *z* using the expression

$$\bar{z}^2 = \frac{1}{T-1}\sum_{t=1}^{T} z_t^2$$

The obtained average values of the square of the magnitude z for each sensor window d form a sequence

$$\bar{z}_d^2 = [\bar{z}_{1d}^2, \bar{z}_{2d}^2, \ldots, \bar{z}_{id}^2, \ldots, \bar{z}_{Nd}^2]$$

The grouped average values of the smoothed time series for all sensors form a D × N matrix of the form

$$\bar{z}_1^2 = \begin{bmatrix} \bar{z}_{11}^2 & \ldots & \bar{z}_{N1}^2 \\ \ldots & \ldots & \ldots \\ \bar{z}_{1D}^2 & \ldots & \bar{z}_{ND}^2 \end{bmatrix}$$

Let Θ_1 denote the norm of the matrix $\bar{z}_1^2$ so that $\Theta_1 = \|\bar{z}_1^2\|$. We calculate its value and perform data transformation as follows. Add a new window for all sensors and delete the first one. We repeat steps 1, 3 sequentially and obtain a series of values $\Theta_1, \Theta_2, \ldots, \Theta_l, \ldots, \Theta_L$, where L is the number of repeated data updates.

Using a modification of the Foster-Steward criterion, we check the resulting sequence for the presence of a trend.

The considered method of detecting anomalies of multidimensional time series is built into an expert system.

3. ExpSystem communication

To support decision-making on safety management of continuous multistage technological processes, the authors have developed an expert system ExpSystem [5, 6]. It allows you to combine the results of expert assessment and quantitative analysis of the state of the technological process to control the safety of production in conditions of uncertainty. Figure 1 shows a three-step procedure for determining the bifurcation point, identifying technological circuits responsible for disrupting the normal operation of the system and finding faulty equipment.

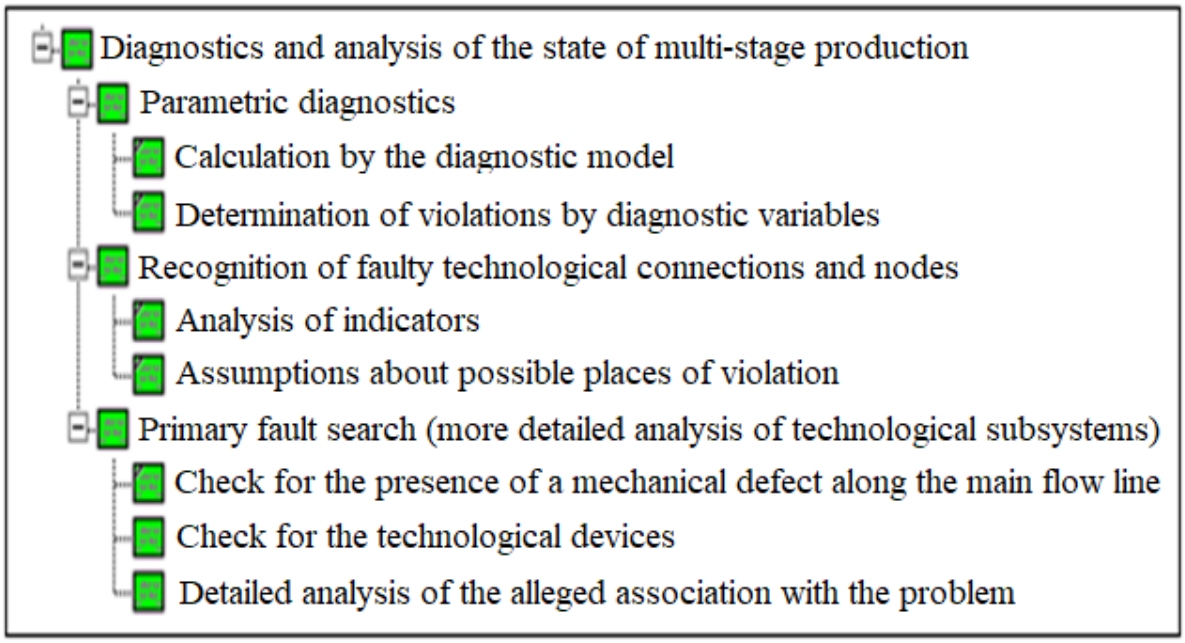

Figure 1: Diagram of a three-step diagnostic

At the first stage of the procedure, parametric diagnostics is carried out, based on continuous monitoring of data coming from sensors of technological equipment. The values of diagnostic variables measured at a discrete time t reflect the dynamics of changes in the state of the technological process.

The analysis of the current values of diagnostic variables in accordance with the procedure described in section 2 serves as the basis for identifying bifurcations of the technological process. As a result of the analysis, diagnostic variables are determined whose values go beyond technological limitations. Figure 2 shows an example of detecting abnormal values of diagnostic variables.

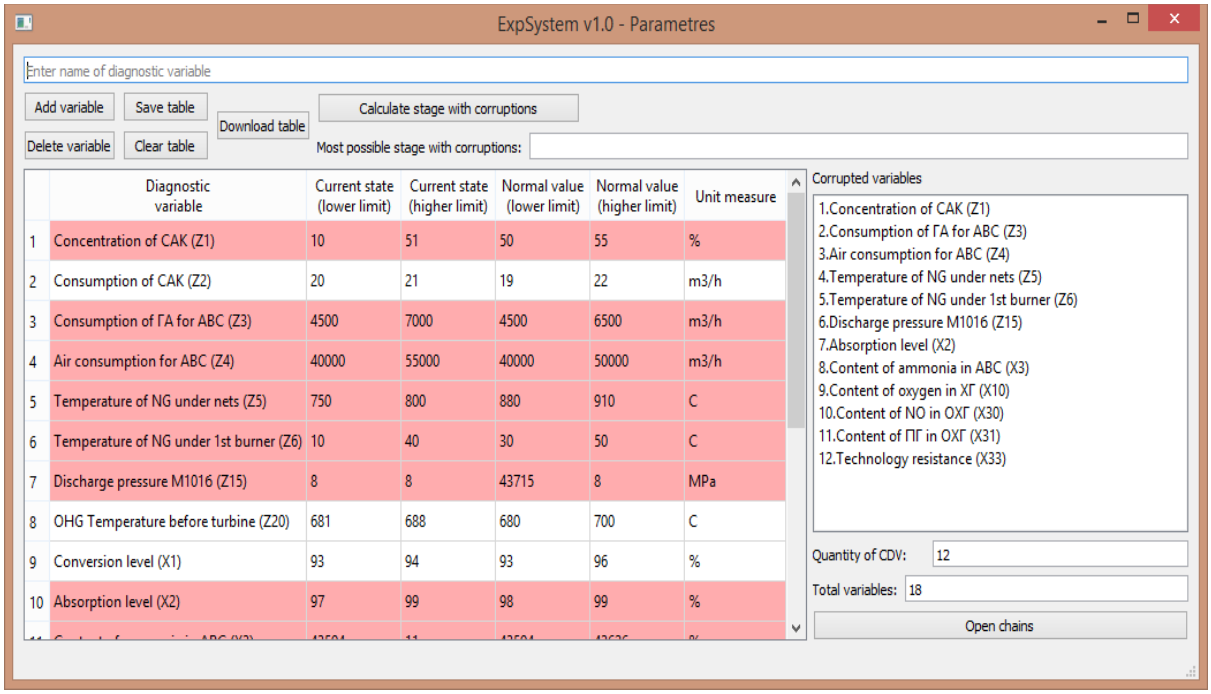

Figure 2: Abnormal values of technological parameters are shown in red

If the beginning of the transition of the technological process to a non-stationary mode is recorded at the first stage, the expert system proceeds to the second stage. The task of the second stage is to identify technological chains (stages) of continuous production, which are possible sources of critical condition formation. To do this, a diagnostic matrix is used, which determines the probability of a connection between diagnostic variables and the stages of the technological process. Using the Dempster-Schafer trust function allows determining the overall probability measure on subsets of faulty process chains [7, 8]. It allows you to combine individual evidence and determine the probabilistic characteristics of the violation of the technological process for its individual stages. The third stage of the procedure for diagnostics and analysis of the state of multistage production is the search for faulty devices in the process chain identified at the second stage. For this purpose, a knowledge base is used, which is formed by specialists in accordance with technological regulations. Troubleshooting is carried out by the operator through a dialogue with the expert system.

4. Multisensor system for processing primary information

The multidimensional information flow in the control system of continuous multi-stage technological processes is formed by a set of sensors that register the values of technological parameters in real time. The high frequency of parameter values removal required by the technological regulations leads to the need to transmit and process large volumes of data at high speed. Taking into account the limited capabilities of the equipment, we can talk about distributed data processing on a certain set of application servers (Server App1) of the first stage. Each of the Server App1 can process only a certain portion of the readings of a part of the sensors. After processing the data, each Server App1 must transfer its part of the data to the database server (Server BD) and the main server (Server APP2) where the final analysis of the data received from the Server App1 of the first stage should be performed. As such devices, programmable external interface cards (PEIC) were used to read and transmit sensor readings with a low-level GPIO interface (General Purpose Input/Outputs), to the outputs of which all kinds of

sensors can be connected. At the same time, it is possible to use both sensors with a single-wire bus for data transmission (1-Wire) and with a multi-wire bus (I2C, SPI standard).

The presented architecture assumes the following interaction of components. The process of obtaining sensor readings consists of two cycles – a data measurement cycle and a data capture cycle. All commands are initiated by the PEIC master device, which, having received the sensor readings in the form of electrical pulses, converts them into numerical form and sends them to Server App1 over the Modbus RTU network for further processing. The data from the sensors is received at certain intervals to the PEIC device and transmitted over the network to the first-stage Server App. The latter install applications that represent OS services. Thus, the processing time for primary data is minimized. The Server App of the second stage performs the final processing of data and provides them for analysis using the ExpSystem client application to the expert technologist's terminal over the HTTP protocol. A Firewall is enabled on the Server App server of the second stage, which, if necessary, will block unauthorized access to the server. In addition to Server App servers and PEIC clients, the system includes a MySQL database server (DB server) for data storage. All equipment is integrated into a Modbus RTU network using routers.

The presented architecture makes it possible to analyze technological data (sensor readings) in a timely manner and, if necessary, to act properly, in case of a malfunction, on the technological process to an expert technologist in real time.

For control systems of continuous multi-stage technological processes, monitoring and control are possible only on the basis of data streams coming from sensors. Streaming data processing has an advantage in terms of data processing time due to real-time operation, if possible, transfer already processed up-to-date data for further use and in terms of uniformity of the load on software and hardware. The ExpSystem used a distributed computing model [10, 11, 12] used for parallel computing over very large datasets in computer clusters.

The streaming data processing technology [13, 14] developed for the ExpSystem requires parallelization. The readings of each sensor are stored in the device's own memory. The sensor communicates over the 1-Wire bus. All processes on the bus are controlled by a central microprocessor. Initialization and reading commands are evaluated in a few microseconds. To parallelize the data flow, sensors are grouped together. To distinguish them, the leading device of the group uses a 64-bit serial code unique for each sensor. On each master device of the PEIC, the algorithm of interaction of the devices with sensors is launched, described below.

Algorithm:
Cycle
    The PEIC device generates a sensor initialization command
    If the sensors respond, they simultaneously form presence pulses, then
        The PEIC device generates a command Reading the sensor memory readings
        The PEIC device detects sensors for reading readings, leaving all sensors active
        The PEIC device reads the sensor memory readings
        If the readings of the read parameter exceed or below a certain value, then
            The PEIC device sends the marked sensor readings to the first stage Server App with the "failure" key about a possible system failure
        Otherwise
            The PEIC device sends sensor readings to the Server App of the first stage in the order they are received with the "normal" key
        All_if
        The PEIC device deactivates the sensors
    All_if
All_cycle

Thus, the main task of the leading PEIC devices is to survey sensors and transmit their readings in digital form to communication channels. In order to ensure the distribution of calculations across several master devices of the PEIC, they are combined into a cluster. Data from each cluster is transmitted to the application servers over a computer network. The first stage server (Server App1) receives this data in real-time, and performs their initial operational processing. In this case, the marked data is delivered to Server App1 first. Then the processed data is transferred to the main Server App 2. The proposed scheme for processing streaming data in a multisensory system allowed minimizing processing time

during data analysis. The models and tools for processing streaming data described in this paper have been pre-tested and have shown their effectiveness in processing streaming data from 30 sensors with a data capture rate of 2 minutes. Further research is related to the development of reliable methods for determining bifurcation points under partial uncertainty conditions.

The proposed approach to detecting anomalies in continuous multi-stage technological processes is implemented in the ExpSystem expert system developed by the authors. The method of predicting critical states of the technological process is the initial stage of a sequential analysis of the state of operation of the technological complex, starting from the localization of the source of problems to the level of primary malfunction. Tests conducted on model data confirmed the predictive ability of the proposed method. Further development consists in improving the proposed method in terms of increasing its performance and reducing errors of the first and second kind.

5. Acknowledgements

The authors consider it their pleasant duty to thank for the financial support of the RFBR (project No. 20-07-00199).

6. References

[1] Korbicz, J., Kościelny, J. M. (Eds.) Modeling, Diagnostics and Process Control Implementation in the DiaSter System Springer-Verlag Berlin Heidelberg. (2011).

[2] Smith, C.L., Borgonovo, E. Decision Making During Nuclear Power Plant Incidents – A New Approach to the Evaluation of Precursors Events. Risk Analysis. 2007.

[3] Phimister, J. R. Bier, V. M. KunreutheR, H. C.) Accident Precursor Analysis and Management: Reducing Technological Risk ThroughDiligence. National Academy Press, Washington, DC, 2004

[4] Matthews, C. A practical guide to engineering failure investigation. Wiley 1998.

[5] Palyukh B., Merkuryev S., Vetrov A., Shabanov B. and Sotnikov A. Methods for Forecasting Critical States of the Technological Process in the Evolutionary Management of Continuous Multi-Stage Production / 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). (2021), DOI: 10.1109/ElConRus51938.2021.9396111

[6] Shabanov B., Sotnikov A., Palyukh B., Vetrov A., Alexandrova D. Expert System for Managing Policy of Technological Security in Uncertainty Conditions: Architectural, Algorithmic, and Computing Aspects // Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus) January 28-30. (2019)

[7] Dempster A. P. Belief functions in the 21st century: A statistical perspective // Proceedings of Institute for Operations Research and Management Science Annual meeting, Springer-Verlag Berlin Heidelberg. (2008)

[8] Ronald R. Yager, Liping Liu, Classic Works of the Dempster-Shafer Theory of Belief Functions. Springer Distributed data processing. Technology "client-server" / StudFiles(2019). URL: https://studfiles.net/preview/2484532/page:19/.

[9] Eileen McNulty. Understanding Big Data: The Seven V's. Dataconomy (2014). - https://dataconomy.com/2014/05/seven-vs-big-data.

[10] Memeti Suejb, Pllana Sabri. HSTREAM: A Directive-Based Language Extension for Heterogeneous Stream Computing / IEEE. arXiv:1809.09387. (2018). DOI:10.1109/CSE.2018.00026.

[11] Namiot Dmitry. On Big Data Stream Processing / International Journal of Open Information Technologies. ISSN: 2307-8162 vol. 3, no. 8. (2015)

[12] RasmussenU. Stream processing using grammars and regular expressions / DIKU, Department of Computer Science University of Copenhagen, Denmark, 2017.arXiv:1704.08820v1 [cs.FL] 28 Apr 2017. (2017)

[13] Reniers G.L., Dullaert, W., Ale, B. J. M., Soudan, K. Developing an external domino accident prevention framework: Hazwim. Journal of Loss Prevention in the Process Industries. 2005.

[14] Typical Programmer 2019. URL: http://typicalprogrammer.com/relational-database-expersts-jump-the-mapreduce-shark (last accessed: 21.01.2019). (2019)