OPTIMIZATION OF THE NUMBER OF DATABASES IN THE BIG DATA PROCESSING

A. R. Akhatov, A. Renavikar*, A. E. Rashidov, F. M. Nazarov

Samarkand State University,
140101, Samarkand, Uzbekistan
*NeARTech Solution,
411033, Pune, India

DOI: 10.24412/2073-0667-2023-1-33-47
EDN: QBRKTM

Today, many organizations and companies increasingly need to use Big Data in order to increase their income, strengthen competitiveness, and study the interests of customers. However, most approaches to real-time processing and analysis of Big Data are based on the cooperation of several servers. In turn, the use of multiple servers limits the possibilities of many organizations and companies due to cost, management and other parameters. This research paper presents an approach for real-time processing and analysis of Big Data on a single server based on a distributed computing engine, and it is based on research that the approach leads to efficiency in terms of cost, reliability, integrity, network independence, and manageability. Also, in order to improve the efficiency of the approach, the methodology of optimizing the number of databases on a single server was developed. This methodology uses MinMaxScaler, StandardScaler, RobustScaler, MaxAbsScaler, QuantileTransformer PowerTransformer scaling functions together with Machine Learning Linear Regression, Random Forest Regression, Multiple Linear Regression, Polynomial Regression, Lasso Regression algorithms. The obtained results were analyzed and the effectiveness of the regression algorithm and scaling function was determined for the experimental data.

Key words: Big Data, Real Time Processing, Single Server Distributed Computing Engine, Architecture, Machine Learning, Regression Algorithms, Scaling.

References

1. Alabdullah B., Beloff N., White M. Rise of Big Data — Issues and Challenges. 2018 // 21st Saudi Computer Society National Computer Conference (NCC) 25–26 April 2018, DOI: 10.1109/NCG.2018.8593166.

2. Big Data — Global Market Trajectory and Analytics. Global Industry Analysts. Inc., 2020.

3. Technology and Media, Big Data Analytics Market, Report ID: FBI 106179, Jul, 2022.

4. Amonov M. T.: The Importance of Small Business in a Market Economy // Academic Journal of Digital Economics and Stability, 2021. V. 7. P. 61–68.

5. Akhatov A. R., Rashidov A. E. Big Data va unig turli sohalardagi tadbiqi // Descendants of Muhammad Al-Khwarizmi, 2021. N 4 (18). P. 135—44.

6. Sassi I., Anter S., Bekkhoucha A. Fast Parallel Constrained Viterbi Algorithm for Big Data wi Applications to Financial Time Series // International Conference on Robot Systems and Applications, ICRSA 9 April 2021, P. 50–55. DOI: 10.1145/3467691.3467697.

7. Alaeddine B., Nabil H., Habiba Ch. Parallel processing using big data and machine learning techniques for intrusion detection // IAES International Journal of Artificial Intelligence (IJ-AI), September 2020. V. 9. N 3. P. 553–560. DOI: 10.11591/ijai.v9.i3.pp553-560.

8. Akhatov A. R., Nazarov F. M., Rashidov A. E. Increasing data reliability by using bigdata parallelization mechanisms // ICISCT 2021: Applications, Trends and Opportunities, 3-5.11.2021, DOI: 10.1109/ICISCT52966.2021.9670387.

9. Landset S., Khoshgoftaar T. M., Richter A. N., Hasanin T. A survey of open source tools for machine learning wi big data in the Hadoop ecosystem // Journal of Big Data (2015). 2:24, DOI: 10.1186/s40537-015-0032-1.

10. Oussous A., Benjelloun F.-Z., Lahcen A. A., Belfkih S. Big Data technologies: A survey // Journal of King Saud University — Computer and Information Sciences2018. N 30. P. 431–448. DOI: 10.1016/j.jksuci.2017.06.001.

11. Tang B., Chen Z., Hefferman G., Wei T., He H., Yang Q. A Hierarchical Distributed Fog Computing Architecture for Big Data Analysis in Smart Cities // ASE BigData and SocialInformatics, ASE BD and SI 2015, DOI: 10.1145/2818869.2818898.

12. Chen P., Chun-Yang Z. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data // Information Sciences, 10 August 2014. V. 275. P. 314–347. 10.1016/j.ins.2014.01.015.

13. Kunanets N., Vasiuta O., Boiko N. Advanced Technologies of Big Data Research in Distributed Information Systems // International Scientific and Technical Conference on Computer Sciences and Information Technologies, September 2019. P. 71–76. DOI: 10.1109/STC-CSIT.2019.8929756.

14. Smeliansky R. L. Model of Distributed Computing System Operation wi Time // Programming and Computer Software, 2013. V. 39. N 5. P. 233–241. DOI: 10.1134/S0361768813050046.

15. Akhatov A., Nazarov F., Rashidov A. Mechanisms of information reliability in big data and blockchain technologies // ICISCT 2021: Applications, Trends and Opportunities, 3–5.11.2021, DOI: 10.1109/ICISCT52966.2021.9670052.

16. B. M. Alom, Henskens F., Hannaford M. Query Processing and Optimization in Distributed Database Systems // IJCSNS International Journal of Computer Science and Network Security, Sept. 2009. V. 9. N 9. P. 143–152.

17. Fabian P., Alfonsa K. Efficient distributed query processing for autonomous RDF databases // International Conference on Extending Database Technology, EDBT 2012. DOI: 10.1145/2247596.2247640.

18. Ali A., Hamidah I., Izura U. N., Fatimah S. Processing skyline queries in incomplete distributed databases // Journal of Intelligent Information Systems, 2017. N 48. P. 399–420. DOI: 10.1007/s10844-016-0419-2.

19. Reyes-Ortiz J. L., Oneto L., Anguita D. Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf // Procedia Computer Science, 2015. N 53. P. 121–130. DOI: 10.1016/j.procs.2015.07.286.

20. Reis Marco Antonio de Sousa, de Araujo Aleteia Patricia Favacho. ArchaDIA: An Architecture for Big Data as a Service in Private Cloud // CLOSER 2019 — 9th International Conference on Cloud Computing and Sendees Science, P. 187–197, DOI: 10.5220/0007787801870197.

21. Sandhu A. K. Big Data wi Cloud Computing: Discussions and Challenges // Big Data Mining And Analytics, 2022. V. 5. P. 32–40. DOI: 10.26599/BDMA.2021.9020016.

22. Nagarajan R., Thirunavukarasu R. Big Data Analytics in Cloud Computing: Effective Deployment of Data Analytics Tools // IGI Global, 2022, 17 pages, DOI: 10.4018/978-1-6684-3662-2.ch011.

23. Wu C. Research on Clustering Algorithm Based on Big Data Background // Journal of Physics: Conf. 2019. Ser. 1237. P. 22–131. DOI: 10.1088/1742-6596/1237/2/022131.

24. Kurasova O., Marcinkevicius V., Medvedev V., Rapecka A., Stefanovic P. Strategies for Big Data Clustering // IEEE 26th International Conference on Tools wi Artificial Intelligence, 2014. P. 739–747. DOI: 10.1109/ICT AI.2014.115.

25. Garlasu D., Sandulescu V., Halcu I., Neculoiu G., Grigoriu O., Marinescu M., Marinescu V. A Big Data implementation based on Grid Computing // Conference: Roedunet International Conference (RoEduNet), 2013 11th, DOI: 10.1109/RoEduNet.2013.6511732.

26. Yuanyuan J. Smart grid big data processing technology and cloud computing application status quo and challenges // 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), 21–23 January 2022, DOI: 10.1109/ICPECA53709.2022.9719287.

27. Akhatov A. R., Sabharwal M., Nazarov F. M., Rashidov A. E. Application of cryptographic methods to blockchain technology to increase data reliability // 2nd International Conference on Advance Computing and Innovative Technologies in Engineering 2022, 28–29 April, DOI: 10.1109/ICACITE53722.2022.9823674.

28. Bollegala D. Dynamic Feature Scaling for Online Learning of Binary Classifiers // Knowledge-Based Systems, July 2014, DOI: 10.1016/j.knosys.2017.05.010.

ОПТИМИЗАЦИЯ КОЛИЧЕСТВА БАЗ ДАННЫХ ПРИ ОБРАБОТКЕ БОЛЬШИХ ДАННЫХ

А. Р. Ахатов, А. Ренавикар*, А. Э. Рашидов, Ф. М. Назаров

Самаркандский государственный университет,
140101, г. Самарканд, Узбекистан
*NeARTech Solution,
411033, Пуна, Индия

УДК 004.658.4
DOI: 10.24412/2073-0667-2023-1-33-47
EDN: QBRKTM

Сегодня многим организациям и компаниям все чаще необходимо использовать большие данные для увеличения доходов, усиления конкурентоспособности, изучения интересов клиентов. Однако большинство подходов к обработке и анализу больших данных в реальном времени основаны на взаимодействии нескольких серверов. В свою очередь, использование нескольких серверов ограничивает возможности многих организаций и компаний из-за стоимостных, управленческих и других параметров. В этом исследовательском документе представлен подход к обработке и анализу больших данных в режиме реального времени на одном сервере на основе распределенного вычислительного механизма, и он основан на исследованиях, которые приводят к эффективности с точки зрения стоимости, надежности, целостности, независимости от сети и управляемости. Также с целью повышения эффективности подхода была разработана методика оптимизации количества баз данных на одном сервере. В этой методологии используются функции масштабирования MinMaxScaler, StandardScaler, RobustScaler, MaxAbsScaler, QuantileTransformer, PowerTransformer вместе с алгоритмами линейной регрессии машинного обучения, регрессии случайного леса, множественной линейной регрессии, полиномиальной регрессии, регрессии лассо. Полученные результаты были проанализированы и определена эффективность алгоритма регрессии и масштабирующей функции для экспериментальных данных.

Ключевые слова: большие данные, обработка в реальном времени, распределенный вычислительный движок на одном сервере, архитектура, машинное обучение, алгоритмы регрессии, масштабирование.

Introduction. Today, Big Data and Big Data analysis are considered as the basis for the development of science, economy, society, government and all spheres, i.e. a new stage. For this reason, many scientific and practical studies are being conducted on the collection, storage and processing of Big Data. But fundamental works in scientific publications are still insufficient [1].

The place of Big Data in the world economy can also be seen in the fact that the global size of the Big Data analysis market in 2021 is estimated at 240.56 billion dollars. In 2020, this indicator was equal to 70.5 billion dollars [2]. Also, the Big Data analytics market is projected to grow from 271.83 billion dollars in 2022 to 655.53 billion dollars by 2029 [3]. The main reason for this is that many organizations and companies seek to use Big Data to increase

revenue, retain customers, or improve product quality and become more competitive. Big Data allows enterprises and organizations to gain a deeper understanding of their activities and make strategic decisions in real time. However, not all enterprises and organizations are able to use Big Data analysis solutions. This opinion is especially relevant to the representatives of small businesses, which form the basis of the economy [4].

A number of studies are being conducted by world scientists on Big Data processing and analysis. Such studies include parallel computing mechanisms in real-time processing and analysis of Big Data [5–8], Hadoop ecosystem [9–10], distributed computing systems [11–15], distributed databases [16–18], Blut technologies [19–22], use of Cluster technologies [23–24], Grid system [25–27] can be cited as an example. Distributed computing systems are based on the models, algorithms, methods and approaches proposed as a result of most of the research. Due to the fact that a distributed computing system consists of several independent computing machines that work together, their use for small business representatives has caused an increase in costs. Sometimes these expenses can be more than the expected income. In addition, distributed computing systems require fewer Big Data specialists and experts due to the complexity of designing data security, integrity, and analysis. For these and similar reasons, finding other effective solutions for real-time processing of Big Data remains an urgent research topic.

Based on the literature studied in this research work, and in order to eliminate the above-mentioned shortcomings, a model of using a distributed computing engine on a single server for real-time processing of Big Data is proposed. Also, in order to maintain the effectiveness of the proposed approach, the methods of designing databases based on artificial intelligence will be explained.

1. Materials and Methods.

1.1. *Architecture of real-time data processing on a single server based on a distributed computing mechanism.* Big Data processing and analysis using a distributed computing system is based on the approach of storing data on several servers. In this case, each server has its own memory and operating system, and when requests are made by clients, the servers achieve efficiency by dividing requests or working together. An overview of the architecture of data processing in a distributed computing system is presented in Fig. 1, a.

The Big Data processing and analysis mechanism proposed in this research work is based on the distributed computing system's large-scale data processing and analysis approach. But the proposed approach uses a single server instead of multiple servers. In the process of storing large volumes of data, data is distributed to several databases on a single server based on certain rules. That is, if in distributed computing systems data is distributed to several servers, then in the proposed model it is distributed to the database of a single server. When a client makes requests, the distribution model forwards the requests to the appropriate database. When choosing the necessary database, the criteria used for dividing data into databases are used. During the execution of the request, the search is not performed among all the data on the server, this process is performed only in a certain database. As a result, the query is not performed on all the data has a positive effect on the data processing time. The architecture of data processing on a single server based on the distributed computing mechanism is presented in Figure 1.b.

1.2. *Positive and negative indicators of the approach to data processing on a single server based on a distributed computing mechanism.* The proposed approach can provide positive

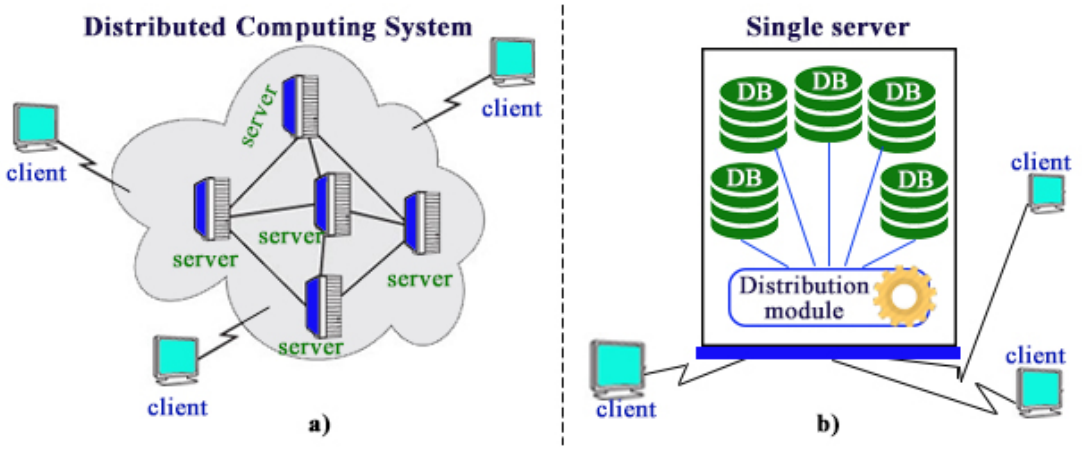

Fig. 1. a) Data processing architecture in a distributed computing system; b) Data processing architecture on a single server based on a distributed computing mechanism

efficiency according to several indicators. This can also be seen in the comparison with the data processing approach in the distributed computing system in Table 1.

As can be seen from Table 1, the proposed Big Data processing and analysis approach on a single server based on the distributed computing mechanism is more effective than the distributed computing systems approaches in terms of cost, controllability, security, integrity, and network reliability. The availability indicator of this approach depends on the constant availability of electric current, protection from external influences, and this problem can be practically solved. The main problem in the approach is the limitation of the amount of data to be stored and processed, that is, the problem of scalability. This is related to the memory resource of a single server, the more memory the server has, the more data it can process and analyze in real time. Based on the proposed approach, servers with several terabytes of memory resources allow real-time processing and analysis of data in the Fast Data class of Big Data. Since Fast Data data processing and analysis can fully meet the needs of small organizations and small business representatives, it is not important to pay attention to the scalability problem of the proposed approach.

In order for the proposed approach to provide the expected efficiency, it is necessary to correctly organize the work of the distribution module in the server and correctly design the databases. The work of the distribution module is considered to be correctly organized when it is ensured that the incoming data is distributed as evenly as possible to the database based on a certain rule. Distribution of data based on a specific rule helps to determine which database to find them in the process of processing. Equal distribution of data helps to equalize the time required for operations on the same databases and, as a result, to minimize the total time of system operation.

1.3. *Methodology for optimizing the number of databases in the approach of data processing on a single server based on a distributed computing mechanism.* The efficiency of large data processing on a single server based on a distributed computing mechanism depends on the number of databases in the server. On the one hand, in the proposed approach time efficiency is achieved by increasing the number of databases, on the other hand, the increase of the database, which is not proportional to the volume of data, has a negative effect on the time indicator. Because the total number of databases is directly proportional to the time spent

Table 1

Evaluation of data processing approaches on a single server based on a distributed computing mechanism and in a distributed computing system according to various indicators

Indicators	An approach to data processing on a single server based on a distributed computing mechanism	Data processing architecture in distributed computing systems
Cost	building an infrastructure with a single computer is several times cheaper than a distributed computing system that includes several servers	It is expensive due to the fact that it contains several servers and requires special methods and tools
Manageability	A single server architecture requires fewer specialists and labor management, and the implementation of processes does not depend on other servers	since the processes involved depend on several servers, the management process is complex and requires special specialists
Safety	it is more possible to install security controls on a single server than to ensure the security of several servers	a head on a single server can threaten the security of the entire head system
Availability	availability is low compared to distributed computing systems	Due to the fact that data can be replicated on several servers, availability is high due to the fact that the downtime of one server is hidden by another server
Integrity	integrity is easy to maintain because the same data is stored in a single database on a single server	integrity is difficult to maintain because the same data is replicated on multiple servers
Network dependency	insensitive to network problems such as network failures, latency, quality of service and bandwidth overload	It is sensitive to any network failures and delays due to the fact that several servers are connected through communication channels and are required to work cooperatively through these communication channels
Scalability	the size of the data is limited by the size of the server's memory resource	data volume increase is solved by adding new servers to the system
Parallelism	only internal parallelism can exist	High performance parallelism can be achieved through multiple servers

on the distribution module. Therefore, in order to achieve high efficiency in the approach, it is necessary to determine the optimal number of databases. In the study, Machine Learning algorithms were used to determine the optimal number of databases suitable for this data for real-time data processing. That is, based on the data collected as a result of the experiment, the optimal number of the database was predicted using various Machine Learning algorithms.

Optimization of the number of databases in the approach of processing big data on a single server based on a distributed computing engine is carried out based on the following methodology:

Step 1. Experimental tests are conducted on several servers individually based on the proposed approach. In this case, the same set of data is distributed to different number of databases and the results of the processing process are recorded.

Step 2. The data collected as a result of mining is scaled for Machine Learning processing.

Step 3. Based on the scaled data, the optimal number of the database is predicted using various Machine Learning algorithms.

Step 4. Prediction errors of Machine Learning algorithms are determined and optimal algorithm is selected.

Table 2

Proportionality of the data collected as a result of the experiment to the time variable, correlation coefficients

Variables	Correlation coefficient	
	Based on all the information collected	Collected only on the basis of big data
d.tuple	0.591772	0.536529
d.base	−0.295736	−0.412476
volume.MB	0.558165	0.487971
RAM.MB	−0.116031	−0.247784
RAM.s.MHz	−0.102711	−0.241887
hard.d.r.s.MB/s	−0.108392	−0.172707
CPU.MHz	−0.124257	−0.247309
CPU.Core	−0.087359	−0.210822
Cashe.L1.MB	−0.087359	−0.210822
Cashe.L2.MB	−0.087359	−0.210822
Cashe.L3.MB	−0.093424	−0.219544

Step 1. During the research and experimental tests, the following information was collected:
✓d.tuple — total number of tuples in databases;
✓d.base — the number of databases where data is distributed on the server;
✓volume.MB — the total volume of data in the database (Mbayt);
✓hard.d.r.s.MB/s — the speed of reading data from the hard disk (Mbayt/sekund)
✓RAM.MB — RAM capacity (Mbayt);
✓RAM.s.MHz — RAM frequency (MHz);
✓CPU.MHz — processor frequency (MHz);
✓CPU.core — number of processor cores;
✓cashe.L1.MB — the size of the memory of 1st cache (Mbayt);
✓cashe.L2.MB — the size of the memory of 2nd cache (Mbayt);
✓cashe.L3.MB — the size of the memory of 3rd cache. (Mbayt);

d.tuple, d.base, time.secund and volume.MB can be used to determine the optimal number of databases for a single server. All the collected information is used to determine the optimal number of databases when the server parameters change.

Step 2. It is very important to scale the data before transferring it to Machine Learning [28]. The following data scaling methods were used in the study:

1) MinMaxScaler. This scaling method is calculated by formula (1):

$$x_{ijnew} = \frac{x_{ij} - x_{jmin}}{x_{jmax} - x_{jmin}} \tag{1}$$

here x_{ij}— is the variable at the intersection of the i row and the j column, x_{jmin}— is the smallest of the variables in the h column, x_{jmax}— is the largest of the variables in the j column. The data collected using the *MinMaxScaler* formula is brought to the interval $[0, 1]$.

2) MaxAbsScaler. This scaling method is calculated by formula (2):

$$x_{ijnew} = \frac{x_{ij}}{max(abs(x_j))} \tag{2}$$

Table 3

Error indicators depending on the scaling functions of the experimentally applied Machine Learning algorithms. *degree — represents the degree of variables in the Polynomial Regression algorithm

Type of Algorithms	Type of errors	MinMaxScaler	StandartScaler	RobustScaler	MaxAbsScaler	QuantileTransformer	PowerTransformer method="box-cox"	PowerTransformer method="yeo-johnson"
Linear Regression	MAE	0.1030	0.1030	0.1030	0.1030	5.4786	4.4256	4.1186
	RMSE	0.1561	0.1561	0.1561	0.1561	6.8508	5.8557	5.6301
Random Forest Regression	MAE	0.0771	0.0705	0.0829	0.0833	0.0700	0.0600	0.0743
	RMSE	0.2614	0.2547	0.3072	0.3068	0.1770	0.2103	0.2650
Multiple Linear Regression	MAE	0.1750	0.6957	0.3500	0.1714	0.1009	0.4354	0.2882
	RMSE	0.2168	0.8617	0.4335	0.2123	0.1324	0.5528	0.3752
Polinomial Regression		degree = 5	degree = 5	degree = 6	degree = 5	degree = 5	degree = 2	degree = 2
	MAE	0.0490	0.1943	0.0602	0.0480	0.0275	0.4473	0.3393
	RMSE	0.0834	0.3315	0.1200	0.0817	0.0421	0.5689	0.4090
Lasso Regression	MAE	0.2053	0.8160	0.4101	0.2011	0.2794	0.9946	0.8123
	RMSE	0.2515	1.0000	0.5026	0.2464	0.3223	1.0769	0.8722

here $max(abs(x_j))-$ is the largest absolute value in column j. Data values are scaled to the interval [−1,1] using the MaxAbsScaler formula.

3) StandartScaler. The StandardScaler scaling method is based on formula (3):

$$x_{ijnew} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{3}$$

here μ_j- is the arimetic mean value of the variables in the j column, σ_j- is the mean square deviation of the variables in the j column, determined using formula (4):

$$\sigma_j = \sqrt{\frac{\sum(x_{ij} - \mu_j)^2}{N_j}} \tag{4}$$

here N_j- is the number of variables in column j. The defaultScaler formula scales data values to a specific range, not [0 : 1]

4) RobustScaler. The RobustScaler scaling method is based on formula (5):

$$x_{ijnew} = \frac{x_{ij} - Q_{2j}}{Q_{3j} - Q_{1j}} \tag{5}$$

here Q_{2j} is the median of the variables in the j column, Q_{1j} — is the median between the median and minimum of the ranked variables in the j column, Q_{3j} is the median between the median and the maximum of the ranked variables in the j column. It is more efficient to use RobustScaler wi a large minimum and maximum range.

In addition to the scaling methods mentioned above, scaling methods such as Quantile Transformer Scaler and Power Transformer Scaler were used during the research. The main goal is to choose the scaling method that has the best result.

Step 3. Linear Regression, Random Forest Regression, Multiple Linear Regression, Polynomial Regression and Lasso Regression algorithms of Machine Learning were used to optimize the number of databases in the approach of data processing on a single server based on a distributed computing engine. It is known that Linear Regression is based on the formula (6).

$$f(x) = a_0 + a_1 \cdot x \tag{6}$$

here $x-$ is a known value, the goal of the Linear Regression algorithm is to find a_0 and a_1 that determine a given $f(x)$ the least error for a given value of x, and $f(x)$ for the next incoming values of x is to find the prediction value of $\hat{y}$.

The Linear Regression algorithm may not provide high accuracy because the optimal number of databases in the study depends on several variables. In this case, the Multiple Linear Regression algorim is more effective. The Multiple Linear Regression algorithm is based on formula (7).

$$f(x) = a_0 \cdot x_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + ... + a_n \cdot x_n \tag{7}$$

here $x_0, x_1,..., x_n-$ are the input variables and $x_0 = 1$. $a_0, a_1,..., a_n-$ coefficients to be found.

If the formula 7 is adapted to the rows of a certain data set, it comes to the equality in the form of (8):

$$\begin{aligned}
\hat{y}^{(1)} &= a_0 \cdot x_0^{(1)} + a_1 \cdot x_1^{(1)} + a_2 \cdot x_2^{(1)} + a_3 \cdot x_3^{(1)} + ... + a_n \cdot x_n^{(1)} \\
\hat{y}^{(2)} &= a_0 \cdot x_0^{(2)} + a_1 \cdot x_1^{(2)} + a_2 \cdot x_2^{(2)} + a_3 \cdot x_3^{(2)} + ... + a_n \cdot x_n^{(2)} \\
\hat{y}^{(3)} &= a_0 \cdot x_0^{(3)} + a_1 \cdot x_1^{(3)} + a_2 \cdot x_2^{(3)} + a_3 \cdot x_3^{(3)} + ... + a_n \cdot x_n^{(3)} \\
&\vdots \\
\hat{y}^{(m)} &= a_0 \cdot x_0^{(m)} + a_1 \cdot x_1^{(m)} + a_2 \cdot x_2^{(m)} + a_3 \cdot x_3^{(m)} + ... + a_n \cdot x_n^{(m)}
\end{aligned} \tag{8}$$

From here, it is possible to extract the matrix of predictions — $\hat{Y}$ and the matrix of variables — X (9) and the matrix of coefficients — A (10).

$$\hat{Y} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix}, X = \begin{bmatrix} x_0^{(1)}, & x_1^{(1)}, & x_2^{(1)}, & x_3^{(1)}, & ... & x_n^{(1)} \\ x_0^{(2)}, & x_1^{(2)}, & x_2^{(2)}, & x_3^{(2)}, & ... & x_n^{(2)} \\ \vdots & & & & & \\ x_0^{(m)}, & x_1^{(m)}, & x_2^{(m)}, & x_3^{(m)}, & ... & x_n^{(m)} \end{bmatrix} \tag{9}$$

$$A = \begin{bmatrix} a_0, & a_1, & a_2, & ... & a_n \end{bmatrix}. \tag{10}$$

According to the matrix transposition property, formula (10) can be converted into form (11):

$$A^T = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \tag{11}$$

In the case of using formulas 9 and 11, formula (8) can be changed to form (12):

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} = \begin{bmatrix} x_0^{(1)}, & x_1^{(1)}, & x_2^{(1)}, & x_3^{(1)}, & ... & x_n^{(1)} \\ x_0^{(2)}, & x_1^{(2)}, & x_2^{(2)}, & x_3^{(2)}, & ... & x_n^{(2)} \\ \vdots & & & & & \\ x_0^{(m)}, & x_1^{(m)}, & x_2^{(m)}, & x_3^{(m)}, & ... & x_n^{(m)} \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \tag{12}$$

The formula of Multiple Linear Regression can be expressed by matrices in the following short form (13):

$$\hat{Y} = X \times A^T \tag{13}$$

The relationship between the data does not always represent a linear function. Often, the elements of the data set are arranged very irregularly. At such times, using the Polynomial Regression algorithm for prediction can be highly effective.

Step 4. In this study, the most widely used formulas of mean absolute error (MAE) and root mean square error (RMSE) were used to evaluate the errors of Machine Learning algorithms.

2. Result. When determining the proportionality of the data collected as a result of the experiment conducted in the study, the correlation coefficients shown in Table 2 were determined.

From the data in Table 2, it can be seen that the absolute value of the correlation coefficients of all variables is significantly different from zero. This means that all variables are proportional to the time variable. Another conclusion from this experiment is that data proportionality over time is more robust in large data ($d.tuple > 1000000$). Since the main goal of the research is to increase the efficiency of large-volume data processing, in the continuation of the research, those tuples that satisfy the $d.tuple > 1000000$ conditions were used. One of the main results obtained from the research methodology is the evaluation of algorithms for determining the optimal value of the number of databases in the approach of data processing on a single server based on a distributed computing mechanism. Table 3 shows the determined error indicators depending on the scaling functions of the Machine Learning algorithms evaluated in the study.

As can be seen from Table 3, the best result, that is, the smallest mean absolute error ($MAE = 0.0275$) and root mean square error ($RMSE = 0.0421$) was achieved by the Polynomial Regression algorithm using the QuantileTransformer scaling function. These results are very good for research work. Because $MAE = 0.0275$ means finding the optimal number of the database in accuracy. This means finding the optimal number of databases wiin 1 error at most, or no error at all.

3. Discussion. The fact that the absolute value of the correlation coefficient of the studied data is higher than zero does not always indicate the correlation of these variables. In this study, $CPU.Core, Cashe.L2.MB$ and $Cashe.L3.MB$ are also independent of the time variable. Because in the proposed approach, the tasks are not distributed to the processor cores. $Cashe.L2.MB$ and $Cashe.L3.MB$ represent the size of the cache memory that stores the data previously accessed. The results of the time indicator obtained in the study were collected as a result of a single reference to certain information. In addition, since $volume.MB$ and $d.tuple$ variables are linearly related, it is sufficient to use only one of them in the study. One of the most important results obtained in the study is the change of error values using different scaling functions of Machine Learning algorithms. Many literatures suggest using scaling functions MinMaxScaler or StandardScaler depending on whether the data set obeys Gaussian distribution or not. But this research also shows that choosing the best scaling functions is a matter of testing them.

Conclusion. In this study, a Big Data processing approach on a single server was proposed based on a distributed computing mechanism. The studied literature and the conducted experiments can conclude that the proposed approach enables real-time processing of large volumes of data on a single server, leading to efficiency in terms of cost, reliability, integrity, network independence and manageability. will come. This helps to meet the need for large-scale data processing and analysis in many fields.

Also, factors affecting the effectiveness of the proposed approach were highlighted in the study. One of these factors is the number of databases on one server. MinMaxScaler, StandardScaler, RobustScaler, MaxAbsScaler, QuantileTransformer PowerTransformer scaling functions and Machine Learning Linear Regression, Random Forest Regression, Multiple Linear Regression, Polynomial Regression and Lasso Regression algorithms were used to increase efficiency and determine the optimal number of databases. As a result of the experiment, the best performance was achieved in the Polynomial Regression algorithm and the QuantileTransformer scaling function. According to it, the smallest average absolute error, $MAE = 0.0275$ and root mean square error, $RMSE = 0.0421$.

References

1. Alabdullah B., Beloff N., White M. Rise of Big Data — Issues and Challenges. 2018 // 21st Saudi Computer Society National Computer Conference (NCC) 25–26 April 2018, DOI: 10.1109/NCG.2018.8593166.

2. Big Data — Global Market Trajectory and Analytics. Global Industry Analysts. Inc., 2020.

3. Technology and Media, Big Data Analytics Market, Report ID: FBI 106179, Jul, 2022.

4. Amonov M. T.: The Importance of Small Business in a Market Economy // Academic Journal of Digital Economics and Stability, 2021. V. 7. P. 61–68.

5. Akhatov A. R., Rashidov A. E. Big Data va unig turli sohalardagi tadbiqi // Descendants of Muhammad Al-Khwarizmi, 2021. N 4 (18). P. 135—44.

6. Sassi I., Anter S., Bekkhoucha A. Fast Parallel Constrained Viterbi Algorithm for Big Data wi Applications to Financial Time Series // International Conference on Robot Systems and Applications, ICRSA 9 April 2021, P. 50–55. DOI: 10.1145/3467691.3467697.

7. Alaeddine B., Nabil H., Habiba Ch. Parallel processing using big data and machine learning techniques for intrusion detection // IAES International Journal of Artificial Intelligence (IJ-AI), September 2020. V. 9. N 3. P. 553–560. DOI: 10.11591/ijai.v9.i3.pp553-560.

8. Akhatov A. R., Nazarov F. M., Rashidov A. E. Increasing data reliability by using bigdata parallelization mechanisms // ICISCT 2021: Applications, Trends and Opportunities, 3-5.11.2021, DOI: 10.1109/ICISCT52966.2021.9670387.

9. Landset S., Khoshgoftaar T. M., Richter A. N., Hasanin T. A survey of open source tools for machine learning wi big data in the Hadoop ecosystem // Journal of Big Data (2015). 2:24, DOI: 10.1186/s40537-015-0032-1.

10. Oussous A., Benjelloun F.-Z., Lahcen A. A., Belfkih S. Big Data technologies: A survey // Journal of King Saud University — Computer and Information Sciences2018. N 30. P. 431–448. DOI: 10.1016/j.jksuci.2017.06.001.

11. Tang B., Chen Z., Hefferman G., Wei T., He H., Yang Q. A Hierarchical Distributed Fog Computing Architecture for Big Data Analysis in Smart Cities // ASE BigData and SocialInformatics, ASE BD and SI 2015, DOI: 10.1145/2818869.2818898.

12. Chen P., Chun-Yang Z. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data // Information Sciences, 10 August 2014. V. 275. P. 314–347. 10.1016/j.ins.2014.01.015.

13. Kunanets N., Vasiuta O., Boiko N. Advanced Technologies of Big Data Research in Distributed Information Systems // International Scientific and Technical Conference on Computer Sciences and Information Technologies, September 2019. P. 71–76. DOI: 10.1109/STC-CSIT.2019.8929756.

14. Smeliansky R. L. Model of Distributed Computing System Operation wi Time // Programming and Computer Software, 2013. V. 39. N 5. P. 233–241. DOI: 10.1134/S0361768813050046.

15. Akhatov A., Nazarov F., Rashidov A. Mechanisms of information reliability in big data and blockchain technologies // ICISCT 2021: Applications, Trends and Opportunities, 3–5.11.2021, DOI: 10.1109/ICISCT52966.2021.9670052.

16. B. M. Alom, Henskens F., Hannaford M. Query Processing and Optimization in Distributed Database Systems // IJCSNS International Journal of Computer Science and Network Security, Sept. 2009. V. 9. N 9. P. 143–152.

17. Fabian P., Alfonsa K. Efficient distributed query processing for autonomous RDF databases // International Conference on Extending Database Technology, EDBT 2012. DOI: 10.1145/2247596.2247640.

18. Ali A., Hamidah I., Izura U. N., Fatimah S. Processing skyline queries in incomplete distributed databases // Journal of Intelligent Information Systems, 2017. N 48. P. 399–420. DOI: 10.1007/s10844-016-0419-2.

19. Reyes-Ortiz J. L., Oneto L., Anguita D. Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf // Procedia Computer Science, 2015. N 53. P. 121–130. DOI: 10.1016/j.procs.2015.07.286.

20. Reis Marco Antonio de Sousa, de Araujo Aleteia Patricia Favacho. ArchaDIA: An Architecture for Big Data as a Service in Private Cloud // CLOSER 2019 — 9th International Conference on Cloud Computing and Sendees Science, P. 187–197, DOI: 10.5220/0007787801870197.

21. Sandhu A. K. Big Data wi Cloud Computing: Discussions and Challenges // Big Data Mining And Analytics, 2022. V. 5. P. 32–40. DOI: 10.26599/BDMA.2021.9020016.

22. Nagarajan R., Thirunavukarasu R. Big Data Analytics in Cloud Computing: Effective Deployment of Data Analytics Tools // IGI Global, 2022, 17 pages, DOI: 10.4018/978-1-6684-3662-2.ch011.

23. Wu C. Research on Clustering Algorithm Based on Big Data Background // Journal of Physics: Conf. 2019. Ser. 1237. P. 22–131. DOI: 10.1088/1742-6596/1237/2/022131.

24. Kurasova O., Marcinkevicius V., Medvedev V., Rapecka A., Stefanovic P. Strategies for Big Data Clustering // IEEE 26th International Conference on Tools wi Artificial Intelligence, 2014. P. 739–747. DOI: 10.1109/ICT AI.2014.115.

25. Garlasu D., Sandulescu V., Halcu I., Neculoiu G., Grigoriu O., Marinescu M., Marinescu V. A Big Data implementation based on Grid Computing // Conference: Roedunet International Conference (RoEduNet), 2013 11th, DOI: 10.1109/RoEduNet.2013.6511732.

26. Yuanyuan J. Smart grid big data processing technology and cloud computing application status quo and challenges // 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), 21–23 January 2022, DOI: 10.1109/ICPECA53709.2022.9719287.

27. Akhatov A. R., Sabharwal M., Nazarov F. M., Rashidov A. E. Application of cryptographic methods to blockchain technology to increase data reliability // 2nd International Conference on Advance Computing and Innovative Technologies in Engineering 2022, 28–29 April, DOI: 10.1109/ICACITE53722.2022.9823674.

28. Bollegala D. Dynamic Feature Scaling for Online Learning of Binary Classifiers // Knowledge-Based Systems, July 2014, DOI: 10.1016/j.knosys.2017.05.010.

Ахатов Акмаль Рустамович — доктор технических наук, профессор, проректор по международному сотрудничеству Самаркандского государственного университета имени Шарофа Рашидова по специальности 05.01.02 — «Системный анализ, управление и обработ-

ка информации». E-mail: akmalar@rambler.ru, тел: +998902716418. Количество опубликованных научных работ — более 100. Область научных интересов: обработка данных, повышение и обеспечение достоверности этого процесса, автоматизация процессов принятия решений и управления при обработке информации, системы искусственного интеллекта и большие данные.

Akmal Rustamovich Akhatov — doctor of technical sciences, professor, vice-rector for international cooperation of Samarkand State University named after Sharof Rashidov, specialty 05.01.02 — "Systematic analysis, management and information processing". E-mail: akmalar@rambler.ru, phone: +998902716418. The number of published scientific works is more than 100. Field of scientific interests: data processing, increasing and ensuring the reliability of this process, automation of decision-making and management processes in information processing, artificial intelligence systems and Big Data.

Ашвини Ренавикар — доктор, профессор, директор научно-исследовательского центра NeARTech Solution (Индия), зарубежный научный руководитель докторантов Самаркандского государственного университета имени Шарофа Рашидова. E-mail: ashwini20.renavikar@gmail.com, тел: +919890247127. Количество опубликованных научных работ — около 90. Область научных интересов: искусственный интеллект, машинное обучение, информационная безопасность, большие данные, Agile software development и JIRA.

Ashwini Renavikar — doctor of technical sciences, professor, head of NeARTech Solution scientific research center (India), foreign scientific advisor to doctoral students at Samarkand State University named after Sharof Rashidov. E-mail: ashwini20.renavikar@gmail.com, phone: +919890247127. The number of published scientific works is about 90. Field of scientific interests: Artificial intelligence, Machine Learning, information security, Big Data, Agile software development and JIRA.

Рашидов Акбар Эргаш угли — аспирант Самаркандского государственного университета имени Шарофа Рашидова. E-mail: researcher.are@gmail.com, телефон: +998941816422. Количество опубликованных научных работ — более 10. Область научных интересов: большие данные, технологии распределенных вычислений, оптимизация систем управления данными, искусственный интеллект, технологии программирования, веб-технологии.

Akbar Ergash o'g'li Rashidov — a doctoral student of Samarkand State University named after Sharof Rashidov. E-mail: researcher.are@gmail.com, phone: +998941816422. The number of published scientific works is more than 10. Field of scientific interests: Big Data, distributed computing technologies, optimization of data management systems, artificial intelligence, programming technologies, web technologies.

Назаров Файзулло Махмадиярович — кандидат технических наук, доцент, заведующий кафедрой искусственного интеллекта и информационных систем Самаркандского государственного университета имени Шарофа Рашидова по специальности 05.01.02 — «Системный анализ, управление и обработка информации». E-mail: fayzullo-samsu@mail.ru, тел: +998990636901. Количество опубликованных научных работ более 50. Область научных интересов: искусственный интеллект, информационная безопасность, технология Block Chain, технологии программирования, Big Data технологии, технологии распределенных вычислений.

Nazarov Faizullo Makhmadiyarovich — PhD, associate professor, head of the Department of Artificial Intelligence and Information Systems of Samarkand State

University named after Sharof Rashidov, specialty 05.01.02 — "Systematic analysis, management and information processing". E-mail: fayzullo-samsu@mail.ru, phone: +998990636901. The number of published scientific works is more than 50. Field of scientific interests: Artificial intelligence, information security, Block Chain technology, programming technologies, Big Data technologies, distributed computing technologies.

Дата поступления — 15.11.2022