Processing of Medical Different Types of Data Using Hadoop and Java MapReduce

Nataliya Boyko [a], Nazar Tkachuk [a]

[a] *Lviv Polytechnic National University, Profesorska Street 1, Lviv, 79013, Ukraine*

Abstract

This article shows the analysis of sample data of different types using Java MapReduce on the Hadoop platform. The Java programming language and the Java MapReduce API are used to work on large amounts of data ("Big Data") that have different formats and structures. So, the task was to process the medical data and get a single source file. The result of the program was saved in the HDFS file system. These source data can then be saved to the NTFS file system using Sqoop or the files can be copied manually to the system for further processing.

Keywords 1

Data processing, Hadoop, Java Map/Reduce, Heterogeneous data processing, MapReduce, Big data, Data Analysis, HDFS, multiple input.

1. Introduction

Big data is a term for data that does not fit the regular segment. Big Data technology handles data so large that traditional methods and approaches cannot be applied to it, data is too large to be hosted on a single user cluster (server), too unstructured to fit in a row column or structured database, or too progressive flow to fit in a permanent data warehouse. Although size is the most important part, there is actually a more problematic aspect of big data - the lack of structure. [1] Big data is used when conventional applications of modern technology do not allow users to quickly, cost-effectively solve problems arising from data processing.

The purpose of this article is to show the processing of different types of data using Hadoop and Java MapReduce, the task - to process the data and get a single source file [3-5].

Traditional methods of analyzing data that work with structured data of small volume (usually information up to several terabytes) are ineffective for processing different types of large data due to their size and atypical structure, which is not clearly defined and prepared for computer perception.

Traditional data analysis is the work with data in order to properly organize them, interpret them with the help of analytical and statistical tools, search for useful information for making rational decisions. This data analysis does not allow to adequately analyze large amounts of data. Big data analytics is the same job, but with large data. Comparison of big data analytics with traditional analytics is given in Table 1 [2].

IDDM'2020: 3rd International Conference on Informatics & Data-Driven Medicine, November 19–21, 2020, Växjö, Sweden
EMAIL: nataliya.i.boyko@lpnu.ua (N. Boyko); ntv3331998@gmail.com (N. Tkachuk)
ORCID: 0000-0002-6962-9363(N. Boyko); 0000-0003-2344-4934 (N. Tkachuk)

Table 1.
Traditional data analysis vs. Big data analysis

	Traditional analytics	Big data analysis
Data sources	Homogeneous sources that provide only structured and consistent data	Heterogeneous sources that provide structured, unstructured / semi-structured and streaming data
Data storage	Isolated own servers	Cloud hosting in a public / private / hybrid cloud
Database technology	Relational data stores (row column data stores)	NoSQL data warehouses (unstructured)
Data Processing	Centralized architecture	Distributed architecture
Analytics	According to previously collected data (static data)	The need for real-time analytics (streaming data)

2. Analysis of scientific sources and literature

The problem of processing various types of data (sensory numerical, text documents, graphs, etc.) in order to form on their basis operational solutions arose during World War II and was actively developed for use in nuclear projects, missile control, navigation, combat control [4].

Processing and analysis of such different types of data is used to model the development of events and situations, as well as in decision support systems. The study of this problem was started by von Neumann, developed by IBM, scientists of the school Lebedeva SO (specialized computer), Glushkova VM (systems analysis, conflict game theory, problem-oriented systems of modeling and data processing), which led to the development of block programming languages, decision support systems [6, 9].

However, the change in the class of research - from operational to analytical, the emergence of new types of data, the need for rapid access to them, led to increased interest in the problem of integration and data processing to improve the quality of management decisions. The highest peak of research activity in the field of integration occurs in the 90s of the XX century. and nowadays due to the rapid development of Business Intelligence methods and increasing the capabilities of data warehouses (increasing the amount of stored data, the availability of analytical data processing procedures - OLAP) [7]. A feature of modern research is the analysis of not only data types (descriptions), but also semantics. Particularly active development of tools for the rapid collection of various types of data, loading them into the data warehouse, analysis and forecasting is observed in the fields of energy and administrative management, the oil and gas sector [10].

3. Methods and tools

3.1 Theoretical analysis

The information technology industry is developing in the analysis of mainly structured data, as the database is a recommended method of storage, processing and analysis of structured data, as the database model is object-oriented [15].

Unstructured data - from e-mail, images and weblogs to social media messages and sensor data - is growing at an unprecedented rate, so it is not advisable to ignore unstructured data, as effective analytics play a vital role [12].

A striking example of unstructured data is a well-known MS Word word processor. The information in the file can be presented in different ways: the facts can be presented only in the form of text, their tabular presentation can be given, and a diagram illustrating the same question can be given. Finally, the information can be presented in a combined form. Such information is called unstructured, it is the most difficult to automatically process, and its analysis requires human intelligence [2].

Instead, take the simplest database, for example, created in the desktop database management system (DBMS) MS Access. Let it consist of one table. The information contained in it has a rigid structure: the composition of the database record fields (table columns) is determined; each field is assigned a specific name, type and properties; all database records (table rows) have the same composition of zeros, input mask, output format, field validation conditions are the same for all records [18]. All this information is stored in the database together with the contents of the table, ie the database contains not only the information to be stored, but also metadata (information about the information). This information is called structured, which is best suited for automated processing.

Apache Hadoop MapReduce and Apache Spark technologies are the leaders in creating a software platform for the organization of distributed processing of large amounts of data. Compared to Hadoop, Spark provides 100 times more performance when processing data in memory and 10 times more when storing data on disks. Spark stores information on your computer's memory, while Hadoop stores it on disk, providing a higher level of security [14].

Business differentiation technologies such as Hadoop help to universally store and process unstructured data for analysis. [3] Versatility is the storage and processing of data in various ways. In fig. 1 shows the four main stages of data processing in Hadoop.

The first step describes importing data into Hadoop from various resources, such as relational database systems or local files. The second stage is processing. At this stage, the data is stored and processed. The information is stored in a distributed HDFS file system. Hadoop MapReduce performs data processing. The third stage is analysis. In the fourth stage, users can analyze the data obtained [16].

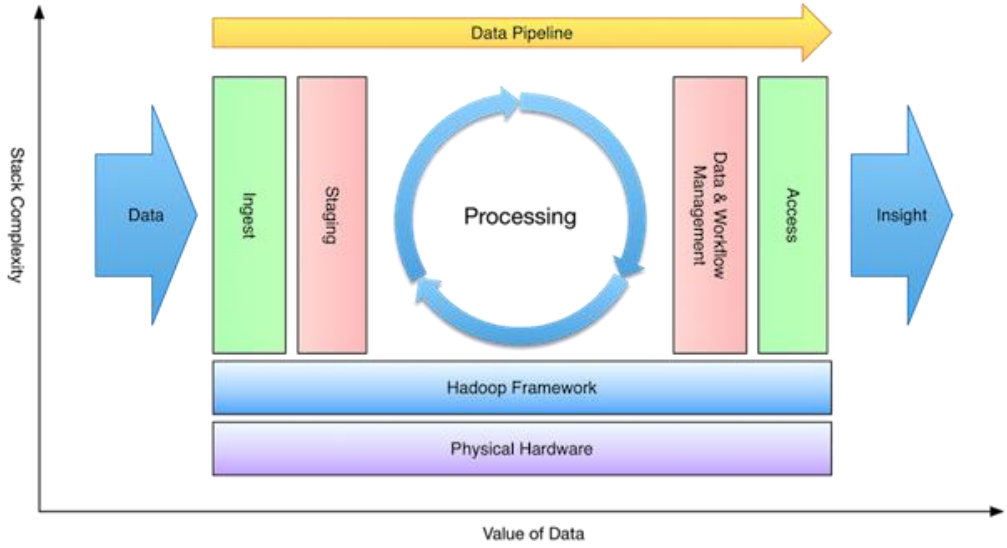

Figure 1: Data processing scheme in Hadoop

3.2 MapReduce

MapReduce is a programming model and corresponding implementation for processing and generating large data sets. Many real world problems are expressed in this model. Programs written in this functional style are automatically parallelized (Hadoop-platforms) and run on a large cluster of machines. A cluster is several independent computers that share and work as one system [17]. The runtime system takes care of the details of the division of input data, scheduling the execution of the program on multiple machines, handling machine failures and managing the necessary between machine communication. This allows programmers who have no experience working with parallel and distributed systems to easily use the resources of a large distributed system [4, 18].

3.3 Programming Model of MapReduce

The MapReduce method involves the organization of data in the form of lists, which are processed in 3 stages (Fig. 2):

1. Map stage, in which data is processed using the map () function, which is defined by the user. The map function takes a list at the input and returns a set of key-value pairs.
2. Shuffle stage, in which the map function is "parsed by baskets" - each basket corresponds to one key of the map stage.
3. Reduce stage. The reduce function determines the final result for an individual "basket". The set of all values returned by the reduce () function is the final result of the MapReduce task.

All launches of the map (), reduce (), and shuffle () functions work independently and can process medical information in parallel on different machines in the cluster. This operation of the MapReduce method allows you to perform the principle of horizontal scaling.

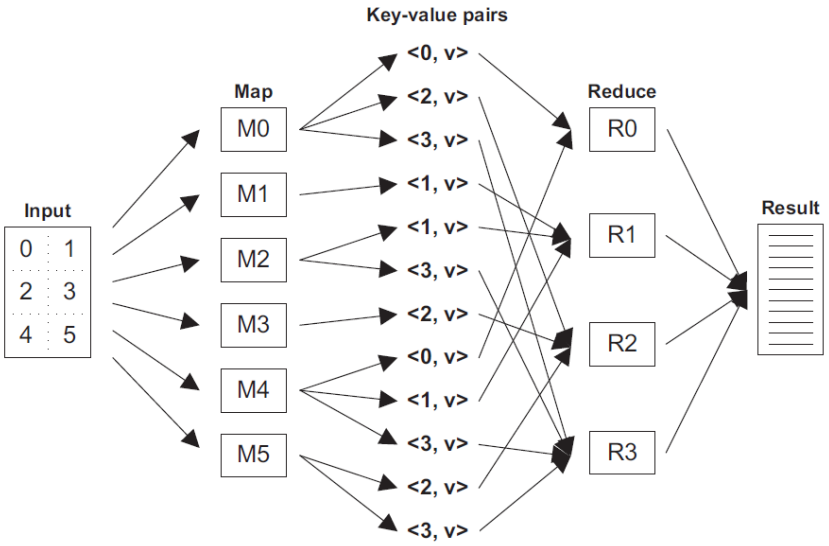

Figure 2: MapReduce data processing model

In the example with the calculation of dividends, the key will be the symbol of a particular medical exchange on which the calculation of the average price. At the Map key stage, the corresponding values from the files are assigned and then the keys are grouped. At the Reduce stage, the necessary calculations are performed on the values, namely the calculation of the average value of dividends for a given year.

4. Experiment

The data in our study are read as input from two dividends.csv and price.csv files containing different structures. The amount of data can reach millions of records. Two MapReduce classes and one Reducer class are used for both files. After completing these tasks, the cluster collects and

truncates the data to generate the corresponding result, and sends it back to the Hadoop server. The evaluation result is very accurate. The result consists of smaller records. Can be displayed on three platforms:

- as a console output
- as an HDFS format
- in an Excel spreadsheet

Sequential data processing algorithm:

1. Place files for processing in the working directory.
2. Specify the path for the HADOOP_HOME variable.
3. Since we have 2 files to process, we will create the appropriate two classes of Maps - ClsPriceMapper and ClsDividendMapper.
4. Let's create a Reducer class - ClsReduce.
5. Let's turn the project into an executable .jar library for easy launch.
6. Processing is started by the hadoop jar command <filename> <input file list> <output file path>. A schematic explanation of the team is presented in Fig. 3.

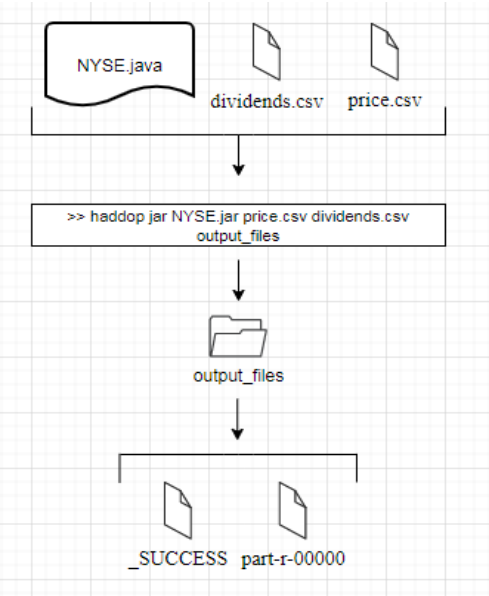

Figure 3: Schematic explanation of the execution command and its parameters

The data processing process is shown in Fig.4. Input using MapReduce classes is sorted and grouped by appropriate keys and represented in the output file.

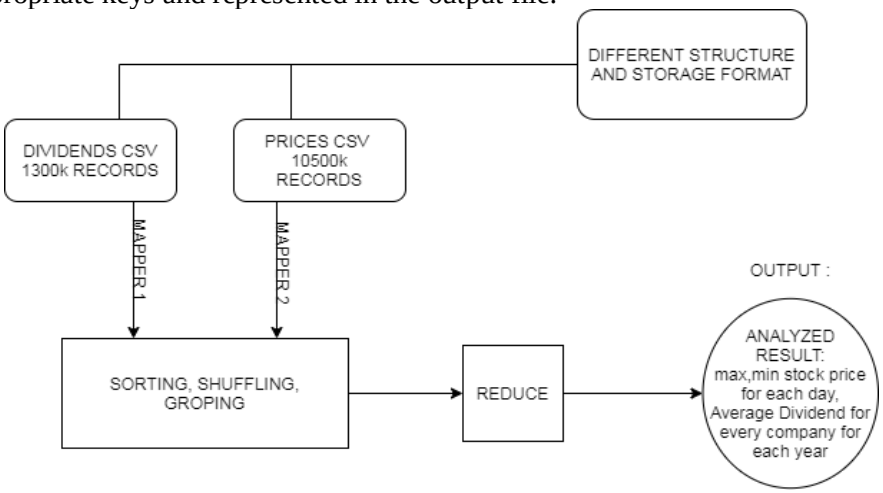

Figure 4: Overview of the process

5. Experiment results

Various types of medical data were successfully processed, and medical information about the result of the command execution appeared on the console. (see Fig. 5). The result window also indicates possible errors that may have occurred during execution. All the necessary calculations were performed in a very short time (compared to traditional data processing methods, a time of 508 ms).

The command to execute is shown in fig. 6, a description of the command structure is given above in the data processing algorithm. After running Java MapReduce, the two files were placed in the source directory (the directory specified in the command on the console). The first file is a _SUCCESS file, which is an indicator of successful program execution. The second file is the part-r-00000 file, which contains the actual source data. In fig. 7 shows the input files and their structure, and in Fig. 8 output file after data processing. In the final file, each line shows the maximum and minimum price of a particular exchange and the calculation of the average share price of any exchange for a particular year.

If the number of gearboxes is deployed more than 1, then the individual files are specified as the source, which are named accordingly (for example, Part-r-00001, etc.). Hadoop provides reliable data display and analysis. This data can be used for further presentation.

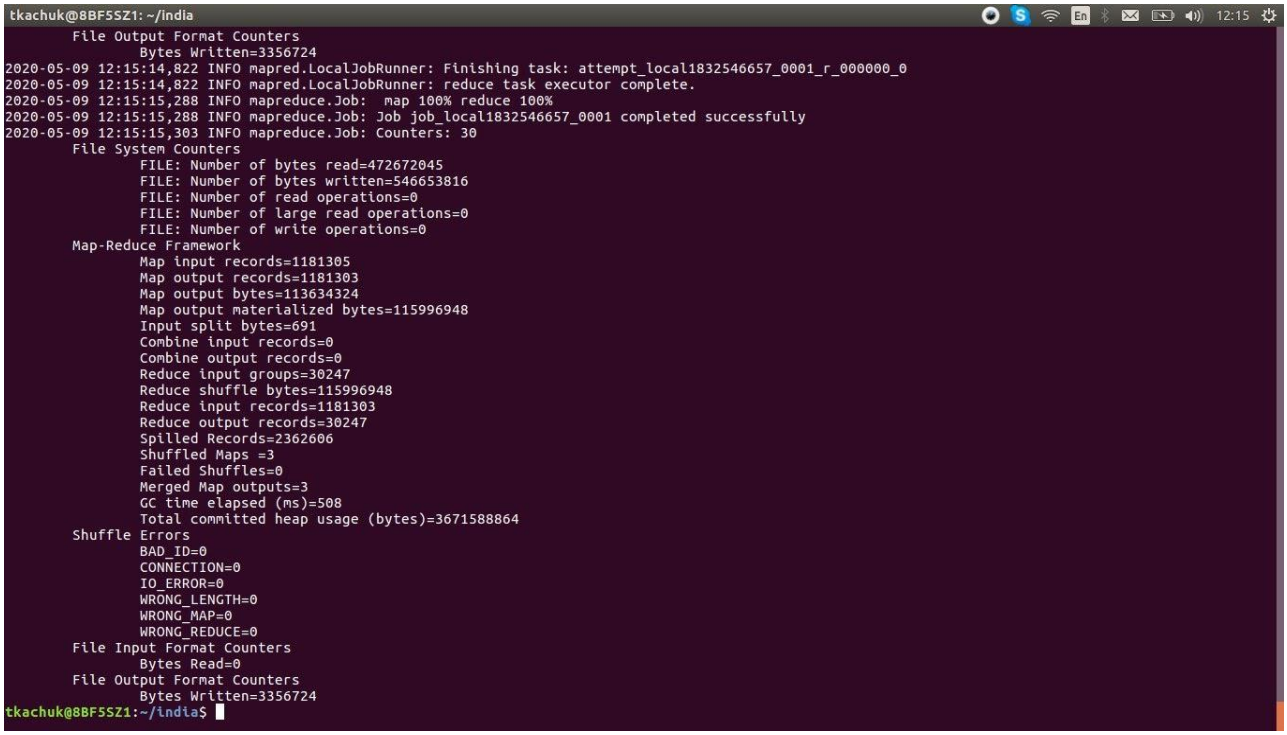

Figure 5: The time required to process the data sample

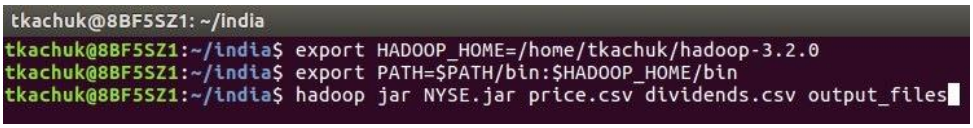

Figure 6: Execution command

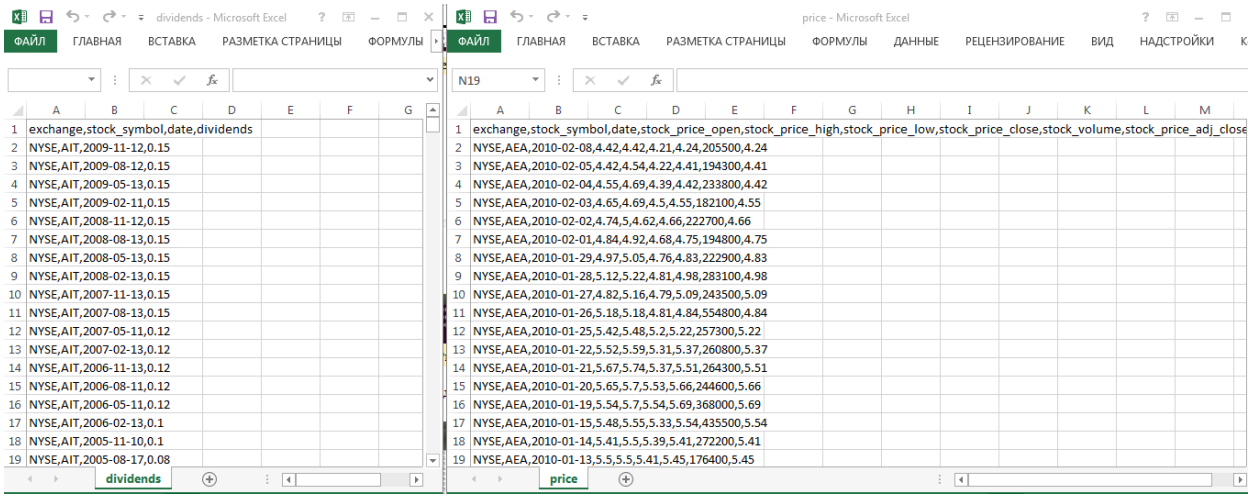

Figure 7: Input files dividends.csv and price.csv

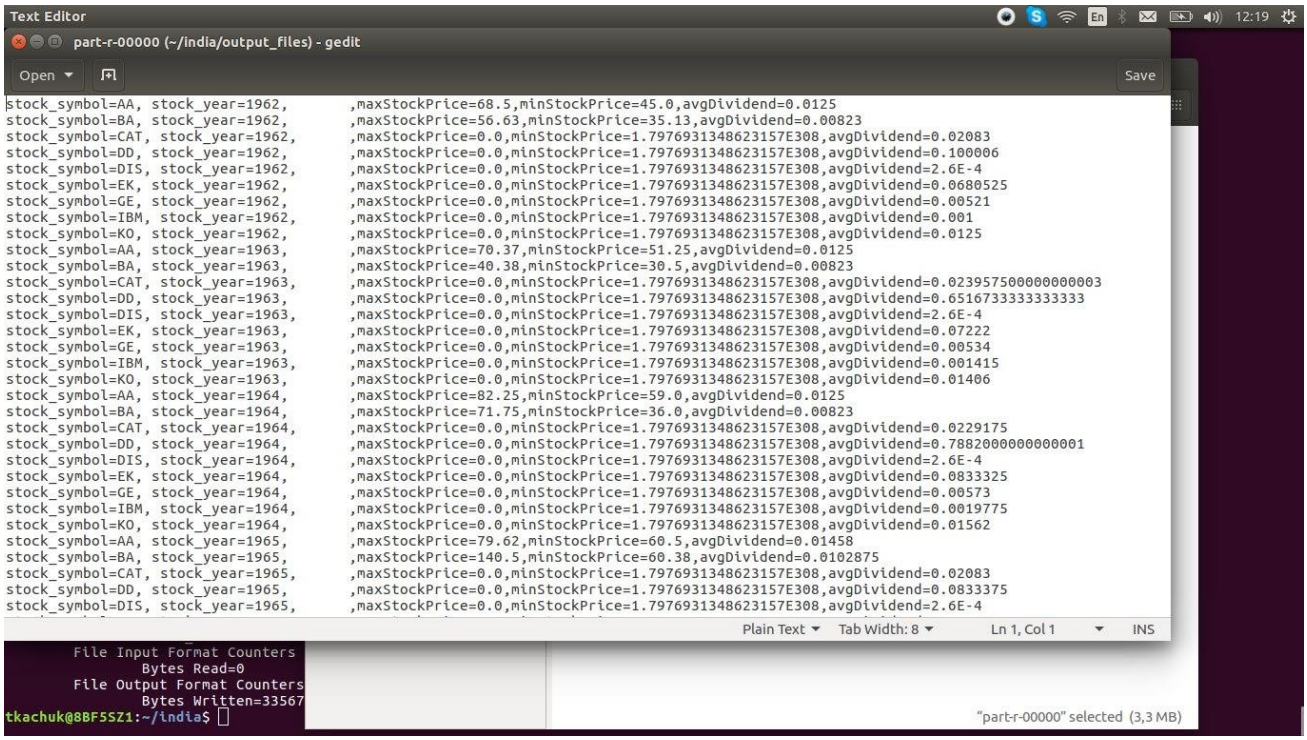

Figure 8: The source file part-r-00000

6. Conclusion

Big data is collected and used by thousands of organizations for research, invention and innovation. They are also used to make money in new ways that were not previously invented or were unattainable. A large database takes into account the complex relationships between samples, models and data sources, along with their development, which change over time and other possible factors. Large, high-performance computing data mining platforms are required. With Big Data technology, you can provide the most relevant and accurate feedback on social perception to better understand our society in real time. [8] The use of Hadoop has increased significantly over the last decade. Almost all well-known companies, such as Yahoo, Google, Amazon, Facebook, IBM and others, prefer to use Hadoop than other technologies, because the processing of unstructured data is easier to do in Hadoop than in others. There are many companies that give up because of the unavailability of big data methods. Currently, more than 50% of Fortune 500 companies use Hadoop. [9, 18]

It is safe to conclude that the processing of big data is less time consuming with Hadoop technology, as compared to the time spent on traditional data analysis. Different types of data can also be received to perform Hadoop data processing. The successful implementation of the Hadoop framework was made in the Linux operating system and the sample data was processed according to our needs using Java Map Reduce.

In the future, this project can be implemented by including various additional frameworks related to Hadoop, such as Pig, Hive and Hbase. The Sqoop platform can also be used to retrieve data from the HDFS format. Also, another type of calculation can be performed on the input data.

7. References

[1] A.K Tung, J. Hou, J. Han, "Spatial clustering in the presence of obstacles", The 17th Intern. conf. on data engineering (ICDE'01), Heidelberg, 2001, pp. 359–367.

[2] C. Boehm, K. Kailing, H. Kriegel, P. Kroeger, "Density connected clus-tering with local subspace preferences" IEEE Computer Society [Proc. of the 4th IEEE Intern. conf. on data mining, Los Alamitos, 2004, pp. 27–34].

[3] D. Guo, D.J. Peuquet, M. Gahegan, "ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata", vol. 3, N. 7, Geoinfor-matica, 2003, pp. 229–253.

[4] D. Harel, Y. Koren, Clustering spatial data using random walks, Proc. of the 7th ACM SIGKDD Intern. conf. on knowledge discovery and data mining, San Francisco, California, 200, pp. 281–286.

[5] N. Boyko, M. Kuba, L. Mochurad, S. Montenegro "Fractal Distribution of Medical Data in Neural Network", The 2 nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019), Volume 1. Lviv, Ukraine, November 11-13, 2019, pp. 307-318.

[6] D.J. Peuquet, "Representations of space and time", N. Y.: Guilford Press, 2002.

[7] H.-Y. Kang, B.-J. Lim, K.-J. Li, "P2P Spatial query processing by Delaunay triangulation", Lecture notes in computer science, vol. 3428,Springer/Heidelberg, 2005, pp. 136–150.

[8] M. Ankerst, M. Ester, Kriegel H.-P. "Towards an effective cooperation of the user and the computer for classification" [Proc. of the 6th ACM SIGKDD Intern. conf. on knowledge discovery and data mining, Boston, Massachusetts, USA, 2000, pp. 179–188].

[9] O. Veres, N. Shakhovska, "Elements of the formal model big date", The 11th Intern. conf. Perspective Technologies and Methods in MEMS Design (MEMSTEH), Polyana, 2015, pp. 81-83

[10] N. Boyko, O. Pylypiv, Yu. Peleshchak, Yu. Kryvenchuk, J. Campos "Automated Document Analysis for Quick Personal Health Record Creation" The 2 nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019), Volume 1. Lviv, Ukraine, November 11-13, 2019, pp. 208-221.C. Zhang, Y. Murayama, "Testing local spatial autocorrelation using", vol. 14, Intern. J. of Geogr. Inform. Science, 2000, pp. 681–692.

[11] R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, "Automatic sub-space clustering of high dimensional data", vol. 11(1), Data mining knowledge discovery, 2005, pp. 5–33.

[12] V. Estivill-Castro, I. Lee, "Amoeba: Hierarchical clustering based on spatial proximity using Delaunay diagram" [9th Intern. Symp. on spatial data handling, Beijing, China, 2000, pp. 26–41].

[13] P. Vitynskyi, R. Tkachenko, I. Izonin and H. Kutucu, "Hybridization of the SGTM Neural-Like Structure Through Inputs Polynomial Extension," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2018, pp. 386-391, doi: 10.1109/DSMP.2018.8478456.

[14] N. Boyko , L. Mochurad, I. Andrusiak, Yu. Drevnytskyi "Organizational and Legal Aspects of Managing the Process of Recognition of Objects in the Image", Proceedings of the International Workshop on Cyber Hygiene (CybHyg-2019) co-located with 1st International Conference on Cyber Hygiene and Conflict Management in Global Information Networks (CyberConf 2019), Kyiv, Ukraine, November 30, 2019, pp. 571-592.

[15] N. Boyko, N. Shakhovska " Prospects for Using Cloud Data Warehouses in Information Systems", 2018 in IEEE 13th International scientific and technical conference on computer

sciences and information technologies (CSIT), vol. 2, DOI: 10.1109/STC-CSIT.2018.8526745

[16] I. Turton, S. Openshaw, C. Brunsdon “Testing spacetime and more complex hyperspace geographical analysis tools”, Innovations in GIS 7, London: Taylor & Francis, 2000, pp. 87–100.

[17] C. Aggarwal, P. Yu “Finding generalized projected clusters in high dimensional spaces”, ACM SIGMOD Intern. conf. on management of data, 2000, pp. 70–81.

[18] C.M. Procopiuc, M. Jones, P.K. Agarwal, T.M. Murali, T.M. “A Monte Carlo algorithm for fast projective clustering”, ACM SIGMOD Intern. conf. on management of data, Madison, Wisconsin, USA, 2002, pp. 418–427.