MEST Journal
DOI 10.12709/mest.06.06.01.06

ONTOLOGY OF BIG DATA ANALYTICS

Vasyl Lytvyn
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine

Victoria Vysotska
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine

Oleh Veres
Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine

JEL Category: **D83, L14, L86**

Abstract

This article describes the features of classification methods and technologies analytics Big Data. There are described: a group of methods and technologies, analytics Big Data that are graded according to the functional relationships and formal model of information technology, the problem of the definition of ontology concepts analytics Big Data. Big Data as a technology turns out to be of great practical importance since it enables solving topical issues of everyday life while at the same time constantly creating new ones. Big Data can change the way we live, work and think. Nowadays the ability to store and analyze large volumes and streams of information turns out to be one of the key preconditions for the successful development of global economy. The countries which will master the most effective ways of working with Big Data are thought to face an industrial evolution of new kind. The branch of Big Data consolidates efforts of the organization in terms of storing, processing and analyzing large data sets. Thus, using the formal model developed as well as the results of the critical analysis conducted, an ontology for the analysis of Big Data may be created. Further research will be focused on investigating methods, models, and tools to refine the ontology for the analysis of Big Data and to provide more effective maintenance for the development of structural components for the model of decision support system for Big Data management.

Keywords: *analysis, big data, visualization data, model, data mining, text mining, MapReduce, ontology*

The address of the corresponding author:
Victoria Vysotska
victoria.a.vysotska@lpnu.ua

1 INTRODUCTION

Big Data enables us to distinguish and to understand such relations between the fragments of information which up until recently could have

hardly been grasped (Mayer-Schonberger, 2013). Big Data represents a great amount of new information in the spheres of civil security, global economic models, personal privacy, established moral rules as well as legal relations between people, businesses, and states.

As smart and interrelated devices are getting more and more widespread, the amount of accumulated information is increasing at a rampant pace. About 90% of data within certain domains is stored in an unstructured form with its amount increasing by 50% annually. In order to remain competitive, introduce innovations and bring products and services to the market, one has to be able to analyze this data and to obtain analytical information from it quickly and cost-effectively. When it comes to the analysis of Big Data as well as to other analytical tasks, current decisions cannot provide system response which is quick enough to be of use when working with such tasks, which in its turn decreases customer efficiency and slows down decision-making process (IBM, 2017). Customers are changing and so is the business world. Analyzing sales data only is far from enough nowadays. The idea behind the deployment of an integrated platform for business intelligence and Big Data analysis is to gain better understanding of why, when, what and how when it comes to customers, products and companies (Vysotska, 2007), (Lytvyn, 2017), (Vysotska, 2013), (Vysotska, 2014), (Vysotska, 2016).

Both business methods and the behavior of customers are changing. Customers in their turn are changing as well. To remain competitive, enterprises are eager to find out in the real-time mode when the customers are buying something, where they are buying it and even what are they thinking about before going to the shop or visiting a website. These are the Big Data, Big Data analysis and integrated platform for business intelligence which can provide help in that respect (Mayer-Schonberger, 2013), (IBM, 2017), (Ageev, 2015), (CNews, 2015).

1.1 The Analysis of Recent Researches and Publications

The conventional business practice of a large-scale data analysis is based on the concept of Enterprise Data Warehouse (EDW), which receives queries from business intelligence (BI) software (Cohen, 2009). BI tools support the opportunity of creating reports and interactive interfaces, generalizing data by means of aggregate functions (e.g. to calculate the total amount or average) as well as dividing hierarchical data into groups.

It is generally accepted that thoroughly designed data warehouse plays a major role when it comes to the proper application of information technologies. Designing and developing a detailed scheme of a data warehouse improves both the results and the presentation of all the business processes and thus turns out to be a general principle of the orderly data integration. The resultant database serves as a repository for the characteristic features of critical business functions. Moreover, a database server, where the data warehouse is stored, is usually the main computing device, serving as a central scaling mechanism for the key corporate analytics. The major role of the data warehouse in conceptual and computational aspects makes it a crucially important and expensive source used for generating reports with large amounts of data, these reports being mainly targeted at decision-making authorities. A data warehouse is usually controlled by specially assigned IT workers, who not only maintain the system but also thoroughly monitor access to it so that the authorities could be provided with the high-quality service (Cohen, 2009).

Even though this conventional approach is still being used in a wide variety of situations, many factors promote an entirely different philosophy of large-scale data management within an enterprise. First, storing data turns out to be so cheap nowadays, that even small subgroups within an enterprise are now capable of developing a separate database of an impressive size even on their own budget. The amount of internal corporate large-scale data sources is steadily increasing: large databases nowadays arise even based on single sources of data flow on Web-browsing (click-stream), software logs, email and forum archives etc. The importance of data analysis is now generally recognized. Numerous companies prove that complex data analysis helps not only to reduce expenses but also to increase profits. As a result, large corporations are now shifting towards the approach of collecting and using data within several of their organizational

units. The advantage of such approach lies in the fact that it increases both the efficiency and the culture of data usage. However, it also leads to data decentralization, which is what data warehouses are designed to deal with.

In this ever-changing climate of fragmented large-scale data collection, an approach known as MAD (Magnetic, Agile, Deep data analysis) tends to be the most appropriate one (Cohen, 2009). Its name, which is an acronym, comes from three major features of this environment that distinguish it from conventional data warehouses, namely its being magnetic, agile and deep.

Modern data analysis involves using more and more complex statistical methods which far exceed the limits of roll-up and drill-down of traditional BI methods. Furthermore, when following these algorithms, analysts often have to investigate large sets of data without using instances and samples. The modern data warehouse should, therefore, serve both as a deep data repository and as the mechanism for executing complex algorithms.

Nowadays the need for competent data analysts tends to be a growing one. They often turn out to be highly-qualified statisticians and possess profound knowledge in the software sphere, though they mainly focus on comprehensive data analysis rather than on database management. MAD approach to designing data warehouses and creating the infrastructure of database systems should be used to facilitate their activities. As long as these goals are achieved, new topical problems with regard to the choice of methods and technologies for the analysis of Big Data arise.

Big data is a set of approaches, tools, and methods for processing structured and unstructured data of great scale and variety to obtain results, which can be comprehended by humans, are efficient under the conditions of steady increase of distribution between numerous nodes of the computing system formed in the late 2000s, and serve as an alternative for the conventional database and BI decision management systems (SAS, 2017), (Mitchell, 2014). When it comes to Big data, three major types of tasks may be distinguished Mayer-Schonberger, 2013), (IBM, 2017), (Ageev, 2015), (CNews, 2015), (SAS, 2017), (Mitchell, 2014):

1. Saving and managing data. Hundreds of terabytes or petabytes of data cannot be easily saved and managed by means of traditional relational databases.

2. Unstructured information. The most of Big Data is unstructured.

3. Big Data analysis. How should unstructured information be analyzed? How should one prepare simple reports, create and implement deep predictive models on the basis of Big Data?

Working with the Big Data is not similar to the ordinary process of business analytics, where the result is achieved through the simple addition of known values. When working with the Big Data the result may be achieved in the process of their refinement by means of consecutive modeling: first, the hypothesis is formed, then the statistical, visual or semantic model is built on the basis of which the accuracy of this hypothesis may be evaluated and finally the next hypothesis is put forward. This process requires a researcher either to interpret visual meanings, to create knowledge-based interactive queries or to develop adaptive machine learning algorithms able to achieve the desired result. The lifecycle of such an algorithm may though be a rather short one (IBM, 2017), (SAS, 2017).

Five major approaches to the analysis of Big Data may thus be distinguished (Mitchell, 2014):

- **Discovery** tools may be used throughout the information lifecycle in order to investigate and analyze information obtained from any combination of structured and unstructured sources quickly and intuitively. Like traditional BI systems, such applications enable us to analyze data sources. No up-front modeling is required, and the users may now engage new ideas, come to meaningful conclusions and make reasonable decisions quickly;
- **BI tools** are of crucial importance when it comes to reporting, analysis and efficiency management, first, in terms of transaction data from data warehouses and production information systems. BI applications provide wide opportunities for business analytics and efficiency management;
- **In-Database Analytics** – methods for identifying data patterns and relationships between data. Such methods are used directly within the database – you prevent data

transfer from other analytic servers, which makes information processing considerably faster and decreases the total cost;
- **Hadoop**: it may be applied to preprocess data in order to identify macro trends and to find data elements within the OUTOF-range values. Various organizations use Hadoop as a precursor to the forms of analytics;
- **Decision Management** – predictive modeling, business rules and self-learning aimed at making reasonable decisions based on current context. This type of analysis provides the opportunity for creating decision-making processes in real-time mode.

All these approaches are used to identify hidden relationships.

2 PREVIOUSLY UNSETTLED ASPECTS OF THE GENERAL PROBLEM

Developing the project of corporate Decision Support System for data management inevitably involves dealing with some difficulties regarding Big Data. New approaches to data analysis are to be developed and existing methods are, if necessary, to be expanded. This is where mathematical sciences can make a considerable contribution: a building based on current statistical methods and investigation into the new methods so as to substitute the old ones, which are less applicable, to make analytics truly efficient and, most importantly, to make sure that correct conclusions are being derived from the data available. Before tools for the analysis of Big Data can be developed and used, one should conduct research into the technologies and approaches implemented to deal with difficulties of obtaining relevant knowledge from structured and unstructured sources with the emphasis being made on the application of Big Data information technology.

3 DESCRIPTION OF METHODS AND TECHNOLOGIES FOR BIG DATA ANALYTICS

Taking into consideration the rapid development of businesses, Big Data information technology may be of great help when it comes to preserving the competitiveness of the enterprise and to processing large amounts of structured and unstructured data. Application of methods and technologies for the analysis of Big Data as well as the development of an integrated platform for business analytics also turn out to be especially topical. The objective of this paper is to conduct research into the peculiarities of classification of methods and technologies for the analysis of Big Data while taking into consideration the definition of Big Data as well as the specific features of its practical application.

As an information technology Big Data may be described by the following formal model (Tadviser, 2017), (Inmon, 2014), (Shakhovska, 2014), (Shakhovska, 2015), (Veres, 2015), (Shakhovska, 2016), which looks like:

$$BD = \langle Vol_{BD}, I_p, A_{BD}, T_{BD} \rangle, \quad (1)$$

where: Vol_{BD} – a set of volume types; I_p – a set of data source (information products) types; A_{BD} – a set of techniques for Big Data analysis; T_{BD} – a set of Big Data processing techniques.

Taking into consideration the definition of Big Data, we may thus formulate the key principles for working with it (Inmon, 2014):
- **Horizontal scalability**: since there might be as much data as needed, each system designed for the processing of Big Data should definitely be expandable. If the amount of data is doubled, hardware should also be increased twofold so that the cluster could continue working properly;
- **Fault tolerance**: the principle of horizontal scalability implies that there might be numerous machines within a cluster, which in its turn means that some of these machines will definitely be out of order. Methods for working with Big Data should take into consideration the probability of such cases, so as to get over them with no serious consequences;
- **Data locality**: In large distributed systems data is usually shared among a big number of machines. If data is physically located on one server and processed on another one, the costs for transferring data may even exceed the expenses for the processing itself. That's why the principle of data locality (i.e. data is stored and processed on the same machine) turns out to be one of the key principles of designing Big Dara solutions.

All the modern tools for working with Big Data comply with the above-mentioned principles to a certain extent. Methods, approaches, and paradigms for the development of data processing tools are to be devised so that we could be able to adhere to these principles.

Nowadays A_{BD} ={A*i*} set of various techniques for the analysis of big volumes of data on the basis of tools provided by statistics and information science is available.

The total amount of data is increasing and so is the amount of its internal and external sources. The data itself is becoming more complex and diverse (structured, unstructured, semi-structured) with various indexation schemes (relational, multidimensional, NoSQL) being used, which in its turn proves the need for new analytical tools. Former data processing techniques turn out to be inefficient – *Big Data Analytics* is applicable to large and complex volumes of data, thus the terms *Discovery Analytics* (what analytics discovers) and *Exploratory Analytics* (what analytics explains) are also being used.

Nowadays there tends to be no difference between Big Data and Big Data Analytics terms. Both of them are used to describe the data itself as well as management technologies and methods of analysis (Barsegyan, 2009). Big Data Analytics is a development of the data mining concept since it involves the same tasks, application areas, data sources, methods, and technologies. The years between the introduction of the data mining concept and the beginning of the era of Big Data were marked by dramatic changes in the volumes of data to be analyzed as well as by the development of high-performance computing systems and other new technologies (e.g. MapReduce and its numerous software implementations). The advent of social networks posed new tasks and challenges.

Data mining is a decision support process based on analyzing raw data with the aim to find hidden trends as well as previously unknown, non-trivial, practically applicable knowledge which is available for interpretation and necessary for making new decisions in various spheres of human activity (Barsegyan, 2009), (Paklin, 2009), (Duke, 2001).

Data mining represents a special approach to data analysis. The emphasis is made not only on extracting facts but also on generating hypotheses. Generated hypotheses should be examined by means of ordinary analysis within conventional schemes and/or with the help of experts in the particular subject area.

The mentioned approach involves using traditional analytical tools such as mathematical statistics (regression, correlation, cluster and factor analysis, time series analysis, decision tables etc.) as well as the tools related to artificial intelligence (machine learning, neural networks, genetic algorithms, fuzzy logic etc.).

If data mining approach is supplemented with MapReduce technology and 4V requirement (Volume, Velocity, Variety, Veracity), it will represent the functional relationships of Big Data Analytics (Fig. 1).

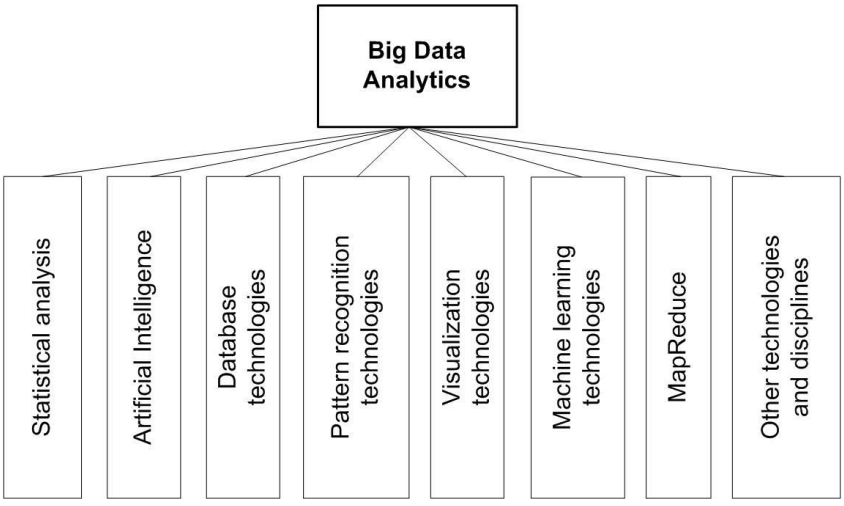

Fig. 1: Functional relationships of Big Data Analytics.

Analysis of large data volumes, as well as the need to understand values in terms of individual behavior, require using processing methods which go far beyond the conventional statistical ones (Barsegyan, 2009).

Methods and techniques of analysis applied to Big Data are also described by McKinsey (Manyika, 2011): Data mining methods, crowdsourcing, data consolidation and integration, machine learning, neural networks, network analysis, optimization (e.g. genetic algorithms), pattern recognition, predictive analysis, simulation modelling, spatial analysis, statistical analysis, analytical data visualization.

In (Manyika, 2011) introduced the following list of methods for the analysis of Big Data analytic methods (in alphabetical order): A/B testing, Association rule learning, Classification, Cluster analysis, Data fusion and data integration, Ensemble learning, Genetic algorithms, Machine learning, Natural Language Processing, Neural networks, Network analysis, Pattern recognition, Predictive modelling, Regression, Sentiment Analysis, Signal Processing, Spatial analysis, Statistics, Supervised and Unsupervised learning), Simulation, Time series analysis and Visualization.

This list can by no means be regarded as a complete one, however, it contains the approaches most widely used within various fields.

Furthermore, some of the techniques mentioned should not necessarily be applied exclusively for the analysis of Big Data, but can rather be successfully used for smaller volumes of data (e.g. A/B testing, regression analysis). It thus becomes obvious, that the larger and the more diversified data set is analyzed, the more accurate and relevant results can be obtained from it.

Let us describe the groups of methods and technologies for the analysis of Big Data classified regarding functional relationships and formal model of the information technology in question – namely Data mining methods, Text mining technologies, MapReduce technology, Data visualization as well as other analytical methods and technologies (Fig. 2).

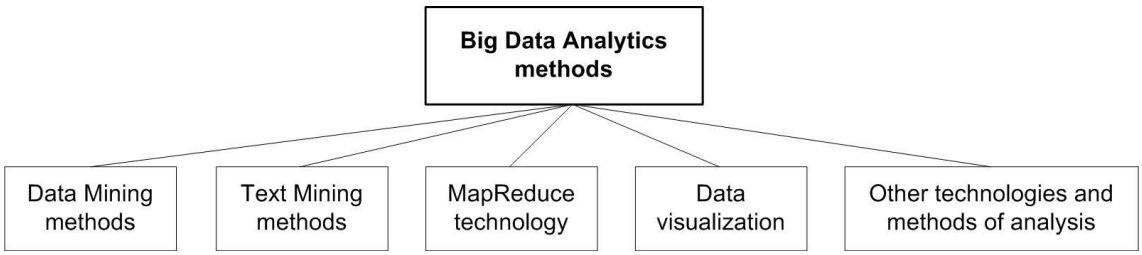

Fig. 2: Groups of Big Data Analytics methods.

The Groups of methods and technologies for the analysis of Big Data can be formally presented as five sets, which looks like:

$$A_{BD} = \langle M_{Data\ Mining}, M_{Visualization}, T_{Text\ Mining}, T_{MapReduce}, T_{other} \rangle, \quad (2)$$

where:

$M_{Data\ Mining}$ — a set of data mining methods;

$M_{Visualization}$ – methods of graphic representation of the analsyis of Big Data;

$T_{Text\ Mining}$ – Text mining technologies;

$T_{MapReduce}$ – MapReduce technology;

T_{other}. – specific methods and technologies for the analysis of Big Data.

3.1 Data mining methods

Data Mining – is a process of detecting hidden relationships and general trends between the variables within large volumes of raw (unprocessed) data.

The following tasks may be solved by means of data mining methods and technologies (Barsegyan, 2009), (Paklin, 2009), (Duke, 2001), (Zhuravlev, 2006), (Zinovev, 2000), (Chubukova, 2006), (Sitnik, 2007), (Witten, 2011):

1. *Classification* – the process of identifying the features characteristic for classes – groups of objects within the set of data under research. These features allow determining whether any new object belongs to the class in question. The

method of Nearest Neighbor and k-Nearest Neighbour as well as Bayesian Networks, induction of decision trees and neural networks are generally used to solve the classification tasks.

2. *Clustering* – the process of dividing objects into groups.

3. *Association* – the process of finding general trends for the related events within a data set. Apriori algorithm is the one most generally applied for association rule learning.

4. *Sequence (sequential association)* allows specifying temporal dependencies between the transactions. Both sequence and association are mostly represented in the same way, but the former aims at defining relationships between the events related in time, which means that it is characterized by a high probability of the chain of such events.

5. *Forecasting:* future values of indexes are estimated on the basis of characteristic features of collected data. Neural networks, as well as the methods of mathematical statistics, are generally used to solve the forecasting tasks.

6. *Deviation Detection, deviation analysis,* and *outlier analysis* – the process of identifying unusual patterns as well as detecting and analyzing data which is the most different from the rest of the dataset.

7. *Estimation* comes down to predicting the continuous values of attributes.

8. *Link analysis* – the task of identifying dependencies within a data set.

9. *Visualization (Graph Mining):* graphic images of data are developed. Graphic methods representing the general trends of a data set are commonly used to solve the visualization tasks.

10. *Summarization* – the process of describing particular groups of objects by means of the data set under analysis.

Data mining is a set of techniques allowing to detect consumer groups most receptive to the promoted product or service, to define the characteristic features of the most successful workers as well as to develop consumer behavior model.

Data mining methods are generally divided into two broad categories – supervised learning and unsupervised learning (Barsegyan, 2009), (Paklin, 2009), (Duke, 2001).

According to another classification, the whole variety of data mining methods can be divided into two groups – statistical and cybernetic methods (Fig. 3). The following classification scheme is based on various approaches to mathematical models of learning (Barsegyan, 2009), (Paklin, 2009), (Duke, 2001), (Zhuravlev, 2006), (Zinovev, 2000), (Chubukova, 2006), (Sitnik, 2007), (Witten, 2011).

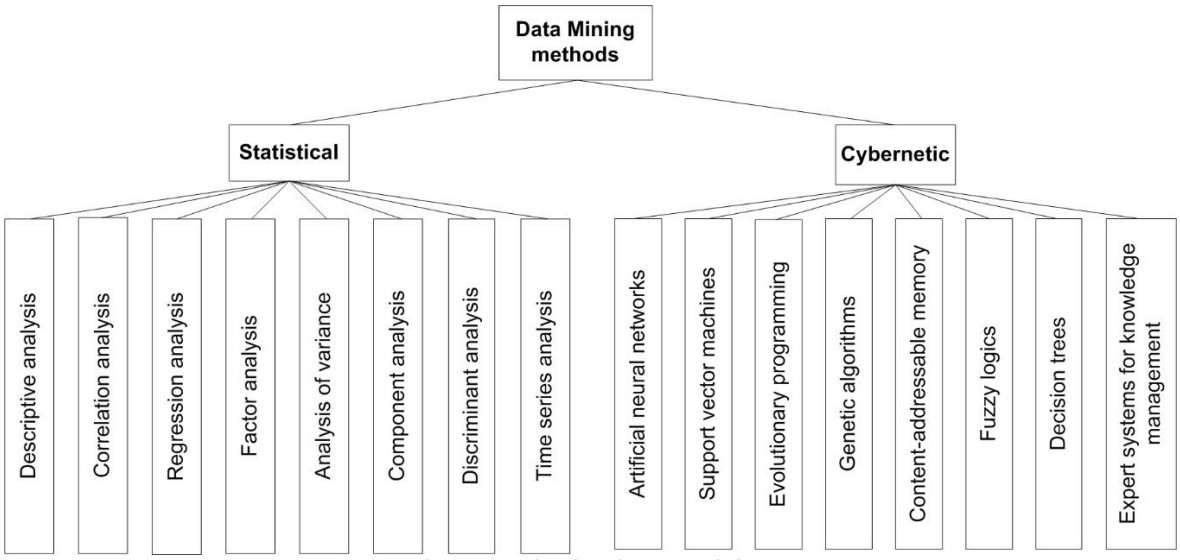

Fig. 3: Methods of Data Mining.

Statistical methods of Data Mining. These methods include conducting primary analysis into the nature of statistical data (verification of stationery hypothesis, normality hypothesis,

independence hypothesis and homogeneity hypothesis; evaluating the type of distribution function and its parameters) and multidimensional statistical analysis (linear and nonlinear discriminant analysis, correlation analysis), identifying relationships and regular patterns (linear and nonlinear regression analysis, cluster analysis, component analysis, factor analysis) as well as developing dynamic models and making forecasts on the basis of time series. Statistical methods of data mining are divided into four groups: descriptive analysis and description of the initial data; relationship analysis (correlation and regression analysis, factor analysis, analysis of variance); multidimensional statistical analysis (component analysis, discriminant analysis, multidimensional regression analysis, canonical correlation analysis); time series analysis (dynamic models and forecasting).

Cybernetic methods of *Data Mining.* This group includes the following methods: evolutionary programming, content-addressable memory (finding analogs and prototypes); fuzzy logic; decision trees; expert systems for knowledge management; artificial neural networks (pattern recognition, clustering, forecasting); genetic algorithms (optimization). Let us describe those of the mentioned methods which are most applicable for the analysis of Big Data (Barsegyan, 2009), (Paklin, 2009), (Duke, 2001), (Zhuravlev, 2006), (Zinovev, 2000), (Chubukova, 2006), (Sitnik, 2007), (Witten, 2011), (Marr, 2015), (Einav, 2013), (Vanyashin, 2013), (Serov, 2012), (Ronen, 2014), (Aflalo, 2013), (Gadepally, 2014), (Weinstein, 2014), (Paklin, 2013).

Association Rule Learning – a group of techniques for identifying relationships (i.e. association rules) between variables in large data sets. Association Rule Learning is a method for detecting curious correlations between variables in large databases. **Association analysis** is carried out in terms of market basket analysis.

Classification is a set of methods allowing to predict the behavior of consumers within a particular market sector (making decisions about purchases, outflows, consumption rates etc.).

Statistical classification is a method for defining to which category a new observation belongs. It requires preparing a set of properly identified observations or historical data. Statistical classification can be used e.g. to automatically assign documents to particular categories, to classify objects into groups, to develop profiles for the students taking an online course as well as for the purposes of focused hiring.

Decision trees is one of the most popular methods for solving classification and forecasting tasks. In its simplest form decision tree is a technique for representing rules in a sequential hierarchical structure. Answers to the certain range of 'Yes'/'No' questions constitute the basis for this structure. Algorithms for creating decision trees usually involve two basic stages – *tree building* and *tree pruning.* On the stage of tree building, the task for choosing decomposition criteria is usually solved and the process of learning is terminated (in case this was predefined by an algorithm). The stage of tree pruning mostly involves cutting some branches of the tree. Method of decision trees is generally known as a 'naive' approach.

Cluster analysis is a method for classifying objects into groups by means of identifying previously unknown general features. Market segmentation is an example of cluster analysis.

Girvan-Newman algorithm within the MLP method (Markov Cluster Algorithm) is used to solve the clustering tasks for graphs.

'Dynamic Quantum Clustering Methodology' implementing the paradigm 'let data speak for itself' was developed for the analysis of multidimensional Big Data (Weinstein, 2014).

DQC method (like many other methods used for analysis of Big Data) 'works' with no prior knowledge about such 'structures', their type and topologies, which can be hidden within data and discovered only after the application of the method in question. DQC method also 'works' perfectly well with multidimensional data with the duration of analysis being linearly dependent on the size of data set.

Regression is a set of statistical methods for identifying regular patterns between a dependent variable and one or several independent variables. It is very often used for making predictions and forecasting. Regression analysis on its basic level involves working with an independent variable (e.g. background music) so as to find out how it affects a dependent variable

(e.g. time spent in the shop). It defines how the values of a dependent variable are changing when an independent variable is changing itself. It functions best for continuous quantitative data such as weight, speed or age. Regression analysis is used to determine how customer satisfaction influences customer loyalty, how the number of received support calls depends on the weather forecast from the previous day, how the neighborhood and the size of the house affect its price and how good are your chances to find the love of your life on the dating website.

Time series analysis comprises a set of methods borrowed from statistics and digital signal processing, which are used for the analysis of data sequences recurring in the course of time. Its obvious applications include tracking securities market and patient disease rates.

Outlier analysis is applied for fraud detection, personality marketing as well as for medical analysis.

Machine learning is a branch of informatics (also known as 'artificial intelligence', this being its historical name), which is aimed at developing self-learning algorithms on the basis of empirical data analysis.

'Machine learning' as a branch appeared after the science of neural networks split into the learning methods for neural networks and types of network architecture and network topology. It also comprises the methods of mathematical statistics. Basic network types, namely perceptron and multilayer perceptron (as well as their modifications) can perform both supervised and unsupervised learning as well as reinforcement learning and active learning. However, the majority of statistical methods, as well as some neural networks, belong to a single type of learning only. That's why the methods of machine learning are classified in accordance with learning types. Though when it comes to neural networks, these are the learning algorithms that are classified:

- *supervised learning:* each precedent is associated with a 'situation – appropriate decision' pair; error correction and backpropagation methods are applied;
- *unsupervised learning:* each precedent is associated with a particular 'situation' only; objects are to be united in clusters on the basis of data on their pairwise comparison, and/or the dimensionality of data is to be reduced; alpha-system reinforcement, gamma-system reinforcement, the method of nearest neighbors;
- *reinforcement learning:* each precedent is associated with a pair 'situation – decision made';
- *active learning:* a learner algorithm itself assigns future situation for which the correct answer will be known for further investigation;
- *semi-supervised learning:* some precedents are associated with a 'situation – appropriate decision' pair, whereas others – with a particular 'situation' only;
- *transduction learning* – semi-supervised learning when the prediction is to be made only for the precedents from the test set;
- *multi-task learning* involves learning a group of interrelated tasks simultaneously. A certain 'Situation – appropriate decision' pair is associated with each of these groups;
- *multi-instant learning:* precedents are divided into groups. One precedent within each group (however, it is not known which one exactly) is associated with a pair 'situation – appropriate decision', whereas all the other precedents are associated with a particular 'situation' only.

Machine learning comprises software able to extract knowledge from data. This allows computers to learn without being directly programmed, and to focus on making predictions on the basis of familiar features, extracted from the sets of training data.

Nowadays machine learning is used to detect spam and ham email messages, to obtain information with regard to customer benefits as well as to make suggestions based on this information, to identify better content so as to attract prospective customers, to assess the probability of winning the case and to establish legal norms for the delivered bills.

Supervised and unsupervised learning. A set of techniques based on the technologies of machine learning which allow identifying functional relationships within the data sets under research. Unsupervised learning in terms of its basic features turns out to be quite similar to cluster analysis.

Ensemble learning. This method involves applying a set of predictive models which improves the quality of forecasting.

Evolution Analysis. Genetic algorithms. Genetic algorithms were inspired by the nature of evolutionary processes – that is, by the mechanisms of mutation, inheritance and natural selection. These mechanisms are used to 'evolutionize' useful solutions to problems which require optimization. This technique involves representing possible solutions as 'chromosomes', which can combine with each other and mutate. Like in the process of natural evolution this is the most adapted creature that survives.

Genetic algorithms are used to solve various tasks – arranging the schedule for doctors in an emergency department; creating combinations of the most appropriate materials and engineering techniques needed to design cost-effective cars; generating artificial creative content - e.g. jokes or puns, forecasting stock market index by means of time series analysis.

Neural networks represent a class of models which function by analogy to the human brain and after completing the learning stage with training data sets are used to solve various tasks of the data analysis.

Neural networks serve as a model of biological neural networks of the brain with homogenous elements (artificial neurons) imitating neurons.

The neural network is represented in the form of weighted directed graphs, where artificial neurons serve as nodes and synaptic links between them – as edges.

Neural networks are used to solve the following tasks: automating pattern recognition processes, making predictions regarding enterprise performance metrics, performing medical diagnostics, forecasting, adaptive management, expert systems development, associative memory implementation, digital and analog signal processing, synthesis and identification of electronic systems.

Neural networks may, for instance, also be used to predict sales volumes and financial market rates, to recognize signals and to develop self-learning systems.

3.2 Data visualization

Data visualization comprises methods used to convey the results of Big Data analysis in graphical form – as diagrams or as animation in order to make these results easier to understand and to interpret. Data visualization involves representing information in the form of pictures, graphs, schemas and diagrams using interactive features and animation both for the results and as an input data for further analysis (Paklin, 2013).

Illustrative representation of the results of Big Data analysis turns out to be of major importance for their further interpretation (Zelazny, 2004), (Roem, 2014), (Tafti, 2014), (Yau, 2013), (Iliinsky, 2011), (Krum, 2014), (Tukey, 1981), (Alper, 2006). Human perception is limited, and the scientists still continue their research aimed at improving modern methods of presenting data in form of images, diagrams or animation. The following methods are among the most advanced ones in terms of data visualization:

- *tag cloud.* Each element within a tag cloud is associated with a certain weighted index which correlates with the font size. In the process of text analysis value of the weighted index directly depends on the usage (citation) frequency of a single word or word combination. It enables a reader to get a general idea of an arbitrarily big text or text sets in a considerably short period of time;
- *clustergram* is a visualization technique applied during cluster analysis. It displays how certain elements of a data set correlate with clusters when the quantities of clusters are being changed. Determining the optimal number of clusters turns out to be the essential component of cluster analysis;
- *history flow:* it allows tracking changes in the document which is being worked on by many authors simultaneously. The time is measured along the horizontal axis, whereas the contribution by each of the authors (the amount of text entered) – along with the vertical one. Each author is associated with a certain color on the diagram;
- *spatial information flow:* this diagram helps tracking spatial distribution of information. The brighter the line – the more data is transferred in a unit of time.

3.3 Text mining technologies

Statistical and linguistic analysis, as well as artificial intelligence methods, serve as a basis for **Text mining** technology. This technology is used for conducting analysis, providing search and navigation in unstructured texts (Barsegyan, 2009), (Statsoft, 2017), (Lande, 2017), (Barsegyan, 2007), (Linyuchev, 2007), (Pleskach, 2011). Application of the Text mining technology allows users to obtain new knowledge.

Text mining technologies represent a set of methods aimed at deriving knowledge from the texts by means of modern information computer systems allowing to detect regular patterns, which enable users to obtain useful data and new knowledge.

Text mining like the majority of other cognitive technologies is an algorithmic detection of previously unknown relationships and correlations already present within the text data.

The approaches and the methodology of data extraction analysis (e.g. classification and clustering methods) are widely used in terms of Text mining technology. Text mining in its turn provides new opportunities – automatic text annotation and phenomenon (i.e. concepts, facts) detection.

One of the important tasks of Text mining technology consists in extracting characteristic elements and features from the text in order to use them as the metadata, keywords, and annotation for the document in question. Another important task involves defining the category within the given text classification scheme to which this document belongs. Text mining technologies provide a new level of semantic document search.

The capabilities of modern Text Mining systems are used to solve the tasks of identifying templates within a document, automatically 'popping' data and dividing it by profiles, creating document overview etc.

Text Mining is a tool allowing to analyze large amounts of data to identify trends, templates, and interrelations, which can turn out to be especially helpful when making strategic decisions. The key idea behind the Text Mining technology consists in enabling analysts to work with large amounts of input data by means of the automated knowledge extraction process. Key methods of Text Mining technology involve classification, clustering, development of semantic networks, relationship, event and fact extraction, feature extraction, summarization, automatic annotation, question answering, thematic indexing, keyword searching, development and maintenance of taxonomies and thesauri. Visualization as a technique for processing structured digital data is also of major importance for Text Mining technology. Visualization is used both as a method of content representation and as a navigation mechanism which can be applied when processing documents or document classes.

Content analysis may serve as an example of successful application of Text Mining technology. Content analysis stands for both qualitative and quantitative systematic processing, evaluation and interpretation of text form and content.

Content-analysis is characterized by a strict procedure and is known to provide well-grounded conclusions. This method is based upon text quantification and further interpretation of the obtained results. The subject of content analysis comprises both problems of social reality which are described or vice versa hidden in the documents and inner characteristic features of the research object itself (Lande, 2017). One reason for the popularity of this method is that it allows measuring human behavior (if we consider verbal behavior to be one of its forms). Unlike surveys content analysis assesses not what people say they have done or will do, but what they actually did.

Two major types of content analysis – qualitative and quantitative analysis – can be distinguished. The task of calculating the frequency of certain topics, words or symbols within a text can be solved by means of quantitative content analysis. Qualitative analysis is applied to single out unusual expressions and linguistic intonations while still placing great value on the content of the message itself. As the automation tools and texts in electronic form become more and more widespread, the methods of content analysis for large volumes of data develop rapidly.

3.4 Other technologies and research techniques

Let us describe several research methods and disciplines related to Big Data technology (Zhuravlev, 2006), (Zinovev, 2000), (Chubukova, 2006), (Sitnik, 2007), (Witten, 2011), (Marr, 2015), (Einav, 2013), (Vanyashin, 2013), (Serov, 2012), (Ronen, 2014), (Aflalo, 2013), (Gadepally, 2014), (Weinstein, 2014), (Paklin, 2013).

A/B testing (*Split testing*) is a method of marketing research which involves consistently comparing control set with other sets. We can thus define the optimal combination of various factors to achieve, for instance, the best customer response to a certain marketing offer. Big Data allows to run a great number of iterations and consequently obtain statistically correct results. This method is also applied when optimizing Web-pages for certain purpose.

Natural language processing is a set of techniques borrowed from linguistics and information science, which are used for natural language recognition.

Sentiment analysis. Techniques for natural language recognition serve as a basis for the methods of sentiment analysis. By means of this method, specific messages related to the certain object (e.g. an item of consumer goods) may be singled out from the general information flow. The polarity (positive or negative coloring) of the expression, as well as its emotional degree, are then being evaluated.

Sentiment analysis allows researchers to define the attitude of speakers or authors to a certain topic. It is therefore widely used, for instance, to improve the quality of service in the hotel chain by analyzing comments of the guests, to arrange services and benefits in accordance with what customers ask for, to identify consumers influenced by social media.

Network analysis comprises a set of techniques used for the analysis of links between the nodes of the network. When applied to social networks, it allows examining relationships between certain users, companies, and communities.

Social network analysis was first used in the telecommunications industry and was then eagerly adopted by sociologists to investigate interpersonal relationships. Nowadays it is applied to analyze relations between people in various spheres of human life, especially in business activities. Nodes represent people in the network whereas links represent relationships between the individuals. Social network analysis is also applied to solve the following tasks:

- investigating how people from different backgrounds establish connections with the strangers;
- measuring the degree to which a individual can influence the group and estimating the importance of such influence;
- defining the minimum number of direct links needed to connect two people;
- understanding the complex social structure of the client database.

Optimization represents a set of quantitative methods used for redesigning complex systems and processes in order improve one or several parameters. It can be of great help when making strategic decisions, for instance, when creating a product line to be brought to the market or conducting investment analysis.

Pattern recognition comprises a set of techniques together with the elements of self-learning which may be used to predict the models of customer behavior.

Predictive modeling is a set of techniques allowing to create a mathematical model of previously specified course of events or another probable scenario. Analysis of a CRM-system database aimed at identifying conditions which may prompt users to switch to another provider can serve as an example of predictive modeling.

Signal processing comprises a set of techniques borrowed from radio technology, which aim at recognizing signal in the background noise as well as analyzing it.

Spatial analysis is a set of techniques partly borrowed from statistics which involve using topological, geometric and geographical information for data analysis. In such cases, these are the geo-information systems that serve as a source of Big Data.

Statistics is a science of collecting, organizing and interpreting data as well as creating questionnaires and conducting experiments. Statistical methods are very often applied to

analyze judgments about the relations between various events.

Modelling the behavior of complex systems is widely used to make predictions, forecast and process various scenarios while planning.

Crowdsourcing is a method for collecting data from many sources. Crowdsourcing involves categorizing and enriching data by a wide indefinite circle of people to use their creative abilities, knowledge, and experience through the application of information and communication technologies.

Data Fusion and Data Integration comprise a set of techniques allowing to integrate heterogeneous data from various sources to conduct the deep analysis. Combination of these techniques provides an opportunity for analyzing user comments on social networks and comparing them with current sales results.

3.5 MapReduce Technology

The technology of distributed file systems made it possible to develop and maintain data warehouses hundreds of terabytes or petabytes in size [48]. When hundreds of terabytes or petabytes of data are being analyzed, data cannot be transferred to any other location for further analysis. The process of transferring data to a separate server or parallel processing server is very likely to take a lot of time and to require too large traffic. Analytical calculations should rather be carried out in the place physically close to where the data is stored. Distributed systems for data processing index and store data on several (and even several thousands) hard drives and servers instead of saving it within the single file structure. A map containing information about the location of certain data is then being created.

Hadoop is one the most famous systems which uses the mentioned approach.

To process data within a distributed file system, one should perform low-level calculations such as addition, aggregation etc. in its physical location within the file system. Then the map of executed algorithms is created so that the local results can be tracked. These results can further be reduced. Such approach and template for executing computable algorithms are therefore known as **MapReduce** (Stonebraker, 2010), (Berezin, 2013), (Lebedenko, 2013), (Pavlo, 2009), (Asash, 2015). MapReduce is a framework for calculating certain sets of distributed tasks by using a large number of computers (nodes) which constitute a cluster. Data saved both within the file system (unstructured data) and in the databases (structured data) can thus be processed.

Broadly speaking, MapReduce has only two main steps: On Map-step the input data is being preprocessed. One of the computers (master node) receives the input tasks, divides them into smaller parts and distributes them to other computers (worker nodes) for further processing. On Reduce-step preprocessed data is being rolled up. Master node receives responses from the worker nodes, which then serve as a basis for the final result – solution of the input task. Users apply Map function to process all the key/value pairs and generate a set of intermediate key/value pairs, whereas Reduce function is used for combining all the intermediate values, associated by the same intermediate key. Another approach to MapReduce technology involves following 5 steps of parallel and distributed calculations, namely preparing input data for Map() function; executing Map() function set by the user, ‘shuffling’ output of the Map() function to Reduce processors; executing Reduce() function set by the user and generating final result.

MapReduce is a model of distributed calculations, introduced by Google which is used to make parallel calculations on large (several petabytes) datasets within computer clusters Stonebraker, 2010), (Berezin, 2013), (Lebedenko, 2013), (Pavlo, 2009), (Asash, 2015).

As far as the realization is concerned, an analytical platform for the analysis of Big Data should be able to apply modern MapReduce technology. In fact, however, analysis of Big Data rarely involves making statistical conclusions for all the data. The importance of Big Data consists in the ability to divide data into ‘microsegments’ and create a great number of models for small observation groups by means of data mining and predictive modeling.

Numerous practical tasks may be implemented within this programming model. Various tools for data aggregation in a distributed file system are available, which allows conducting the analytical process quite easily.

4 ONTOLOGY FOR BIG DATA ANALYSIS

Suggested description of methods and technologies for the analysis of Big Data allows us to create an ontology in terms of METHONTOLOGY (Gavrilova, 2000), (Gavrilova, 2001), (Gavrilova, 2003), (Lytvyn, 2011), (Veres, 2015), (Lytvyn, 2017), (Vysotska, 2013), (Vysotska, 2014), (Vysotska, 2016) approach, which represents the process of iterative development. According to METHONTOLOGY methodology, term glossary contains all the terms (concepts, their instances, attributes, actions) important for the analysis of Big Data, as well as their natural language descriptions.

Term glossary of ontology for the analysis of Big Data contains previously mentioned terms which can be semantically divided into the following groups: task structure (technology analytics groups, connections), data covering the task (methods applied for each of the groups) and results of calculations (recommendations on applying Big Data to increase the efficiency of the decisions made). Ontology for Big Data analysis developed by means of Protégé-OWL ontology editor is displayed in Fig. 4 to Fig. 8.

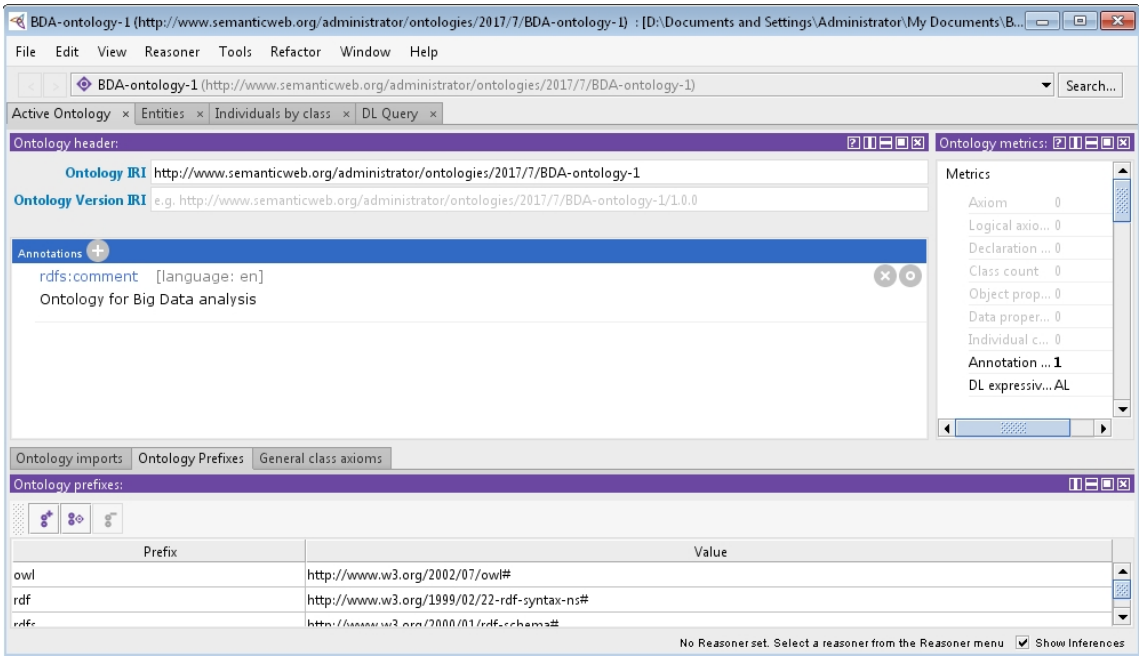

Fig. 4: Annotation the ontology for Big Data analysis.

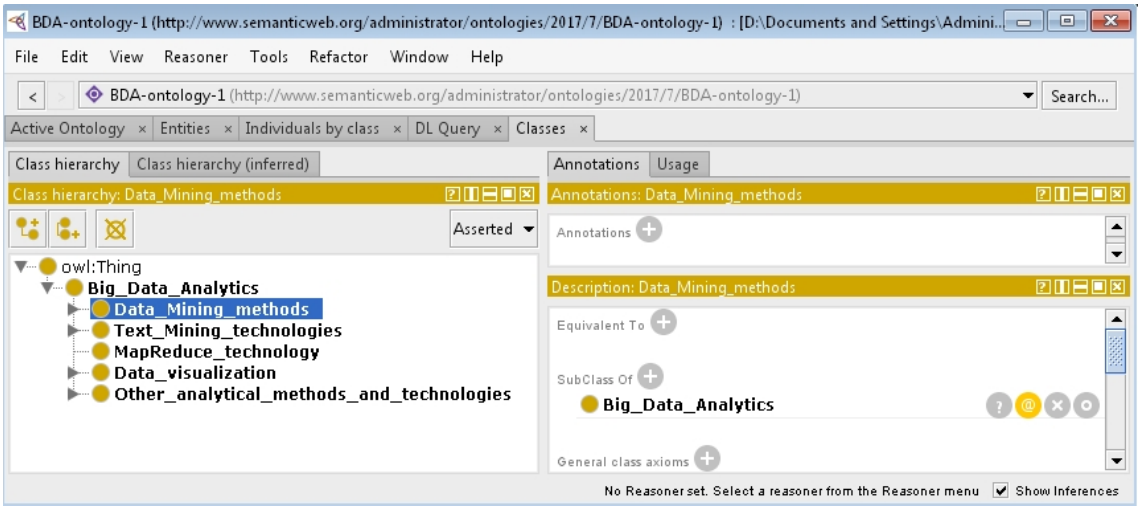

Fig. 5: Subclasses of class "Big Data analysis".

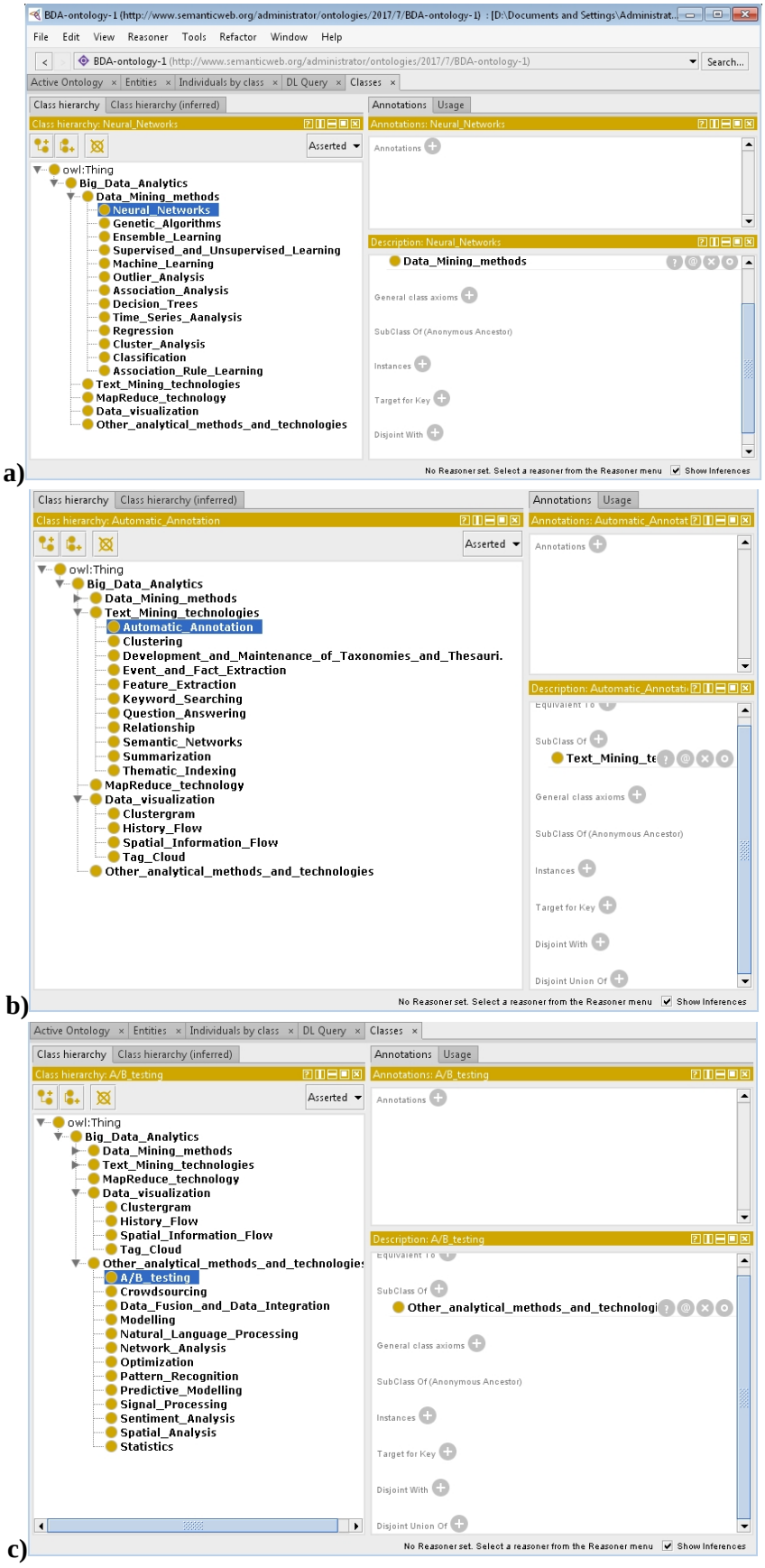

Fig. 6. Subclasses of class a) "Data mining methods";
b) "Text mining technologies"; c) "Other technologies and research techniques".

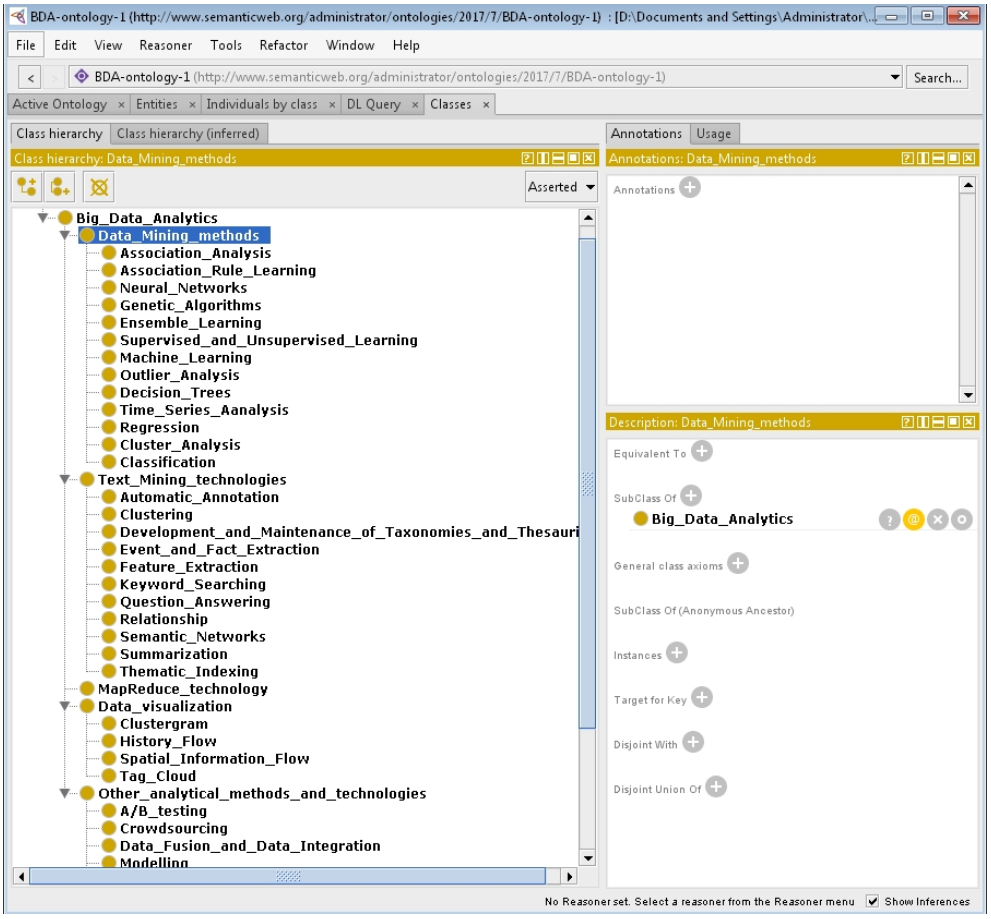

Fig. 7: Hierarchy of classes within the ontology for Big Data analysis.

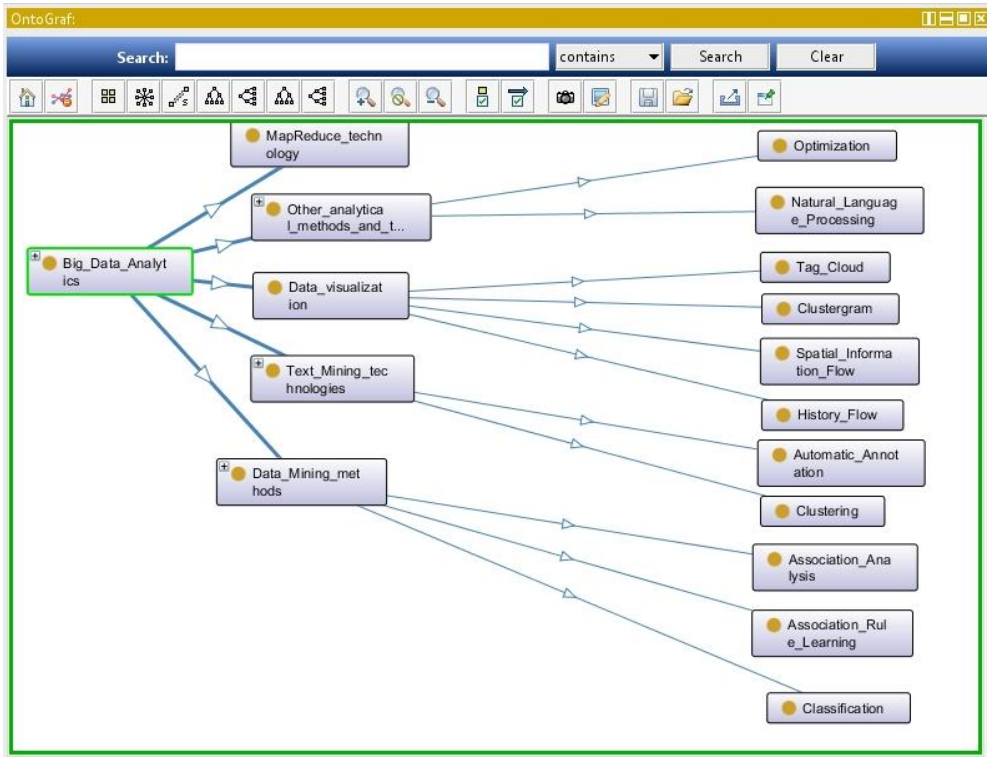

Fig. 8: The structure of the ontology for Big Data analysis in the form of a graph.

5 CONCLUSION

Big Data as a technology turns out to be of great practical importance since it enables solving topical issues of everyday life while at the same time constantly creating new ones. Big Data can change the way we live, work and think.

Nowadays the ability to store and analyze large volumes and streams of information turns out to be one of the key preconditions for the successful development of global economy. The countries which will master the most effective ways of working with Big Data are thought to face an industrial evolution of new kind. The branch of Big Data consolidates efforts of the organization in terms of storing, processing and analyzing large data sets.

One of the most common mistakes with regard to big amounts of data is a naïve expectation that buying large computer infrastructure will definitely be beneficial for the business. However, both information technologies, informatics, and mathematics should go hand-in-hand. Infrastructure is without any doubt extremely necessary, however, in order to benefit from Big Data, one should be able to adopt more complex methods of its analysis.

As a result of this research, formal model of Big Data information technology was used to explain the classification of methods and techniques for the analysis of Big Data into groups. In order to achieve the desired goal and with regard to functional relationships and the formal model of an information technology in question, it was suggested that the methods available should be classified into data mining methods, Text Mining technologies, MapReduce technology, Data visualization and other methods and techniques of analysis. Characteristic features of methods and techniques belonging to each group were described taking into consideration the definition of Big Data.

Thus, using the formal model developed as well as the results of the critical analysis conducted, an ontology for the analysis of Big Data may be created.

Further research will be focused on investigating methods, models, and tools in order to refine the ontology for the analysis of Big Data and to provide more effective maintenance for the development of structural components for the model of decision support system for Big Data management.

WORKS CITED

Aflalo, Y., & Kimmel, R. (2013). *Spectral multidimensional scaling*. PNAS, vol. 110, no. 45, November 5, Retrieved from: http://www.cs.technion.ac.il/ ~ron/PAPERS/Journal/AflaloKimmelPNAS2013.pdf.

Ageev, A., (2015). *Analysts have warned about the dangers of Big Data*. Retrieved from: http://bigdata.cnews.ru/news/top/2015-10-23_eksperty_predosteregayut_ot_nepravilnogo_obrashcheniya.

Alper, C., Brown, K., & Wagner, G.R. (2006). *New Software for Visualizing the Past*. Present and Future, DSSResources.COM, Retrieved from: http://dssresources.com/papers/features/alperbrown&wagner/alperbrown& wagner09212006.html.

Asash. (2015). *Big Data from A to Ya. Part 1: Principles of working with large data, the MapReduce paradigm*. Retrieved from: https://habrahabr.ru/company/dca/blog/267361/.

Asash. (2015). *Big Data from A to Ya. Part 3: Methods and strategies for developing MapReduce applications*. Retrieved from: https://habrahabr.ru/company/dca/blog/270453/.

Barsegyan, A.A., Kupriyanov, M.S., Kholod, I.I., Tess, M.D., & Elizarov, S.I. (2009). *Analysis of data and processes*. BHV-Petersburg, 512 p.

Barsegyan, A.A., Kupriyanov, M.S., Kholod, I.I., Tess, M.D., & Elizarov, S.I. (2009). *Analysis of data and processes*. BHV-Petersburg, St. Petersburg, 512 p.

Barsegyan, A.A., Kupriyanov, M.S., Stepanenko, V.V., & Kholod, I.I. (2007) *Data Analysis Technologies. Data Mining, Visual Mining, Text Mining, OLAP*. BHV-Petersburg, St. Petersburg, 384 p.

Berezin, A. (2013). *Map-Reduce on the example of MongoDB*. Retrieved from: https://habrahabr.ru/post/184130/.

Chubukova, I.A. (2006). *Data Mining: A Tutorial*. Internet University of Information Technologies: BINOM: Laboratory of Knowledge, Moscow, 382 p.

CNews. (2015). *Named the causes braking big data market*. Retrieved from: http://bigdata.cnews.ru/news/top/2015-11-20_analitiki_otsenili_tempy_rosta_mirovogo_rynka.

Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M., & Welton, C. (2009). *MAD Skills: New Analysis Practices for Big Data*. Proceedings of the VLDB'09 Conference, Lyon, France, August 24-28. Retrieved from: http://citforum.ru/database/articles/mad_skills/.

Duke, V., & Samoylenko, A. (2001). *Data Mining: training course*. St. Petersburg, 368 p.

Einav, L., & Levin, J. (2013). *The Data Revolution and Economic Analysis*. NBER Working Paper, No. 19035, Retrieved from: http://www.nber.org/chapters/c12942.pdf.

Gadepally, V., & Kepner, J. (2014). *Big Data Dimensional Analysis*. arXiv:1408.0517v1. Retrieved from: https://arxiv.org/pdf/1408.0517v1.pdf.

Gavrilova, T.A. (2001). *Ontology for the study of knowledge engineering*. Proceedings of the International Scientific and Practical Conference KDS-2001.

Gavrilova, T.A. (2003). *Ontological approach to knowledge management in the development of corporate information systems*. News of Artificial Intelligence, №2, pp.24-30.

Gavrilova, T.A., & Khoroshevsky, V.F. (2000). *Intelligent Systems Knowledge Base*. Piter, St. Petersburg, 384 p.

IBM. (2017). *Big Data and analytics*. Retrieved from: http://www-03.ibm.com/systems/ru/technicalcomputing/bigdata.html.

Iliinsky, N., & Steele, J. (2011). *Designing Data Visualizations*. Sebastopol : O'Reilly, 110 p.

Inmon, W.H. (2014). *Big Data – getting it right: A checklist to evaluate your environment, DSSResources.COM.* Retrieved from: http://dssresources.com/papers/features/ inmon/inmon01162014.htm.

Krum, R. (2014). *Cool infographics: effective communication with data visualization and design*. Indianapolis: Wiley, 348 p.

Lande, D. (2017). *Deep text analysis technology for effective analysis of text data*. Retrieved from: http://visti.net/~dwl/art/dz/.

Lebedenko, E. (2013). *Google MapReduce technology: divide and conquer*. Retrieved from: http://www.computerra.ru/82659/mapreduce/.

Linyuchev, P. (2007). *Text Mining: modern technologies on information mines*. PC Week, RE №6 (564), February 27 - March 5, Retrieved from: https://www.pcweek.ru/idea/article/detail.php?ID=82081.

Lytvyn, V., Vysotska, V., Veres, O., Rishnyak, I., & Rishnyak, H. (2017). *Classification Methods of Text Documents Using Ontology Based Approach*. Advances in Intelligent Systems and Computing, Springer International Publishing, pp.229-240. DOI: 10.1007/978-3-319-45991-2_15

Lytvyn, V.V. (2011). *Knowledge Base of Intelligent Decision Support Systems*: monograph, Lviv Polytechnic Publishing House, Lviv, 240 p.

Manyika, James et al. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, June, 156 p.

Marr, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance*. John Wiley & Sons Ltd, 256 p.

Mayer-Schonberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. John Murray Publishers, UK, ISBN 1848547927 9781848547926.

Mitchell, R.L. (2014). *8 big trends in big data analytics*. Computerworld, OCT 23, Retrieved from: http://www.computerworld.com/article/2690856/big-data/8-big-trends-in-big-data-analytics.html.

Paklin, N.B., & Oreshkov, V.I. (2013). *Business Intelligence: from data to knowledge*. Piter, St. Petersburg, 702 p.

Paklin, N.B., & Oreshkov, V.I., (2009). *Business analysis: from data to knowledge*. St. Petersburg, 624 p.

Pavlo, A., Paulson, E., Rasin, A., Abadi, D.J., DeWitt, D.J., Madden, S., & Stonebraker, M. (2009). *A Comparison of Approaches to Large-Scale Data Analysis*. Proceedings of the 35th SIGMOD International Conference

on Management of Data, Providence, Rhode Island, USA, Retrieved from: http://citforum.ru/database/articles/mr_vs_dbms/2.shtml.

Pleskach, V.L., & Zatonatskaya, T.G. (2011). *Information systems and technologies at enterprises*. Znannya, Kiev, 718 p.

Roem, D. (2014). *The practice of visual thinking. An original method for solving complex problems*. Mann, Ivanov and Ferber, Moscow, 396 p.

Ronen, S., Gonçalves, B., Hu, K.Z., Vespignani, A., Pinker, S., & Hidalgo, C.A. (2014). *Links that speak: The global language network and its association with global fame*. PNAS, Vol. 111, No.52, Retrieved from: http://stevenpinker.com/files/pinker/files/pnas_hildago_et_al_global_ language_network_2014.pdf.

SAS. (2017). History and evolution of big data analytics. Retrieved from: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html.

Serov, D. (2012). *Analytics of "big data" - new perspectives*. Storage News, №1 (49), Retrieved from: http://www.storagenews.ru/49/EMC_BigData _49.pdf.

Shakhovska, N., Veres, O., & Bolubash, Y. (2015). *Big Data Information Technology and Data Space Architecture*. Sensors & Transducers, Vol. 195, No. 12, p. 69-76.

Shakhovska, N., Veres, O., & Hirnyak, M. (2016). Generalized formal model of Big Data, ECONTECHMOD, Vol. 5, No. 2, p. 33–38.

Shakhovska, N., Veres, O., Bolubash, Y., & Bychkovska-Lipinska, L. (2015). *Data space architecture for Big Data managering*. Computer Sciences and Information Technologies, Lviv, p. 184-187.

Shakhovska, N.B., Bolubash, Y.J., & Veres, O.M. (2015). *Big data federated repository model.* CAD Systems in Microelectronics, Lviv, p. 382-384, DOI: 10.1109/CADSM.2015.7230882.

Shakhovska, N.B., Bolubash, Yu.Ja., & Veres, O.M. (2014). Big Data organizing in a distributed environment. Computer Science and Automation, Vol. 2(27), p. 147-155.

Sitnik, V.F., & Krasnyuk, M.T. (2007). *Data Mining*. KNEU, Kiev, 376 p.

Statsoft. (2017). *Text Mining*. Retrieved from: http://statsoft.ru/home/textbook/modules/sttextmin.html#index.

Stonebraker, M., Abadi, D., Dawitt, D.J., Madden, S., Paulson, E., Pavlo, A., & Rasin, A. (2010). *MapReduce and Parallel DBMSs: Friends or Foes?* Communications of the ACM, vol. 53, no. 1, Retrieved from: http://citforum.ru/database/articles/mr_vs_dbms-2/.

Tadviser (2017). *Big Data*. Retrieved from: http://tadviser.ru/a/125096.

Tafti, E. (2014). *Presentation of Information*. Retrieved from: http://envisioninginformation.daiquiri.ru/15.

Tukey, J. (1981). *Analysis of Observation Results: Exploratory Analysis*. Mir, Moscow, 693 p.

Vanyashin, A., Klimentov, A., & Korenkov, V. (2013). *PANDA follows the large data*. Supercomputers, 3 (11), p. 56-61.

Veres, O. (2015). *Ontology Data Cleansing*. Bulletin of the National University of Lviv Polytechnic. Series: Information systems and networks, № 814, 237-245 pp.

Veres, O., & Shakhovska, N. (2015). *Elements of the formal model big date*. International Conference on Perspective Technologies and Methods in MEMS Design, Lviv, p. 81-83.

Vysotska, V., & Chyrun L. (2013). Web Content Processing Method for Electronic Business Systems*. International Journal of Computers & Technology, 12*(2), p. 3211-3220.

Vysotska, V., & Chyrun L. (2014). Life Cycle Model of Commercial Content Processing in Electronic Commerce System. *Computational Problems in Electrical Engineering. Founder and Publisher Lviv Polytechnic National University, 3*(2), p. 118-122.

Vysotska, V., & Chyrun L. (2014). Set-theoretic models and unified methods of information resources processing in e-business systems. *Applied Computer Science journal, 10*(3), pp. 5-22.

Vysotska, V., Chyrun L., Lytvyn, V., & Dosyn, D. (2016). Methods based on ontologies for information resources processing : Monograph. LAP Lambert Academic Publishing. Saarbrucken, Germany.

Vysotska, V., Rishnyak, I., & Chyrun L. (2007). Analysis and evaluation of risks in electronic commerce. CAD Systems in Microelectronics, CADSM '07, 9th International Conference. p. 332-333.

Weinstein, M., Meirer, F., Hume, A., Sciau, Ph., Shaked, G., Hofstetter, R., Persi, E., Mehta, A., & Horn, D. (2014). *Analyzing Big Data with Dynamic Quantum Clustering*. arXiv:1310.2700. Retrieved from: https://arxiv.org/ftp/arxiv/papers/1310/1310.2700.pdf.

Witten, I.H., Frank, E., &Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd Edition, Morgan Kaufmann, 664 p.

Yau, N. (2013). *The art of visualization in business*. How to present complex information with simple images, Mann, Ivanov and Ferber, Moscow, 352 p.

Zelazny, D. (2004). *Speak in the language of diagrams: manual on visual communications for managers*. Institute for Comprehensive Strategic Studies, Moscow, 220 p.

Zhuravlev, J.I., Ryazanov, V.V., & Senko, O.V. (2006). *Recognition. Mathematical methods. Software system. Practical applications*. Phasis, Moscow, 176 p.

Zinovev, A.Y. (2000). *Visualization of multidimensional data*. KSTU, Krasnoyarsk, 180 p.

Received for publication: 11.12.2017
Revision received: 23.12.2017
Accepted for publication: 10.01.2018

How to cite this article?

Style – **APA** *Sixth Edition:*

Lytvyn, V., Vysotska, V., & Veres, O. (2018, Jan 15). Ontology of Big Data Analytics. (Z. Čekerevac, Ed.) *MEST Journal, 6*(1), 41-60. doi:10.12709/mest.06.06.01.06

Style – **Chicago** *Sixteenth Edition:*

Lytvyn, Vasyl, Victoria Vysotska, and Oleh Veres. 2018. "Ontology of Big Data Analytics." Edited by Zoran Čekerevac. *MEST Journal* (MESTE) 6 (1): 41-60. doi:10.12709/mest.06.06.01.06.

Style – **GOST** *Name Sort:*

Lytvyn Vasyl, Vysotska Victoria and Veres Oleh Ontology of Big Data Analytics [Journal] // MEST Journal / ed. Čekerevac Zoran. - Toronto : MESTE, Jan 15, 2018. - 1 : Vol. 6. - pp. 41-60.

Style – **Harvard** *Anglia:*

Lytvyn, V., Vysotska, V. & Veres, O., 2018. Ontology of Big Data Analytics. *MEST Journal,* 15 Jan, 6(1), pp. 41-60.

Style – **ISO 690** *Numerical Reference:*

Ontology of Big Data Analytics. **Lytvyn, Vasyl, Vysotska, Victoria and Veres, Oleh.** [ed.] Zoran Čekerevac. 1, Toronto : MESTE, Jan 15, 2018, MEST Journal, Vol. 6, pp. 41-60.