***Rena T.Gasimova***
Institute of Information Technology of ANAS, Baku, Azerbaijan
rena.gasimova@science.az

DOI: 10.25045/jpit.v07.i1.09

# "BIG DATA" ANALYTICS: AVAILABLE APPROACHES, PROBLEMS AND SOLUTIONS

*Increased volume of data and demand for ad hoc analysis of data leads to the rise of one of the biggest problems of Big Data called Big Data analysis. This article studies the current problems and most frequently used methods of big data analysis and gives some recommendations. The article also investigates the technological stages of Big data processing, and the basic characteristics and features of big data.*

***Keywords:*** *data warehouse, cloud, database management systems, data processing, big data, big data analytics, NoSQL, MapReduce, Hadoop, OLAP*

## Introduction

Data is formally described fact that needs to be converted into information. Until the recent years, whilst talking about data processing, it was referred to an association of algorithmic, logical or statistical operations over relatively small data. However, the convergence of computer technology with the real world is increasing the demand for the conversion of data into information. The amount of data to be processed increases, at the same time, the demand for processing speed expands, which emerges the need for information technology (IT) solutions capable of processing very large volumes of various data in real-time mode.

Logically, IT differs from material technologies, to minimum extent. Raw data becomes the structured information, which is in a rather lucid form and able to turn information into useful knowledge by the power of intellect. In other words, computer reprocesses the raw data, filters useful ones to write in a more suitable form. The joint analysis of various data collected from different sources and acquisition of new knowledge and useful information is a regular technological process. Therefore, common arrangements should be applied to information technology, as this formulates the other technologies. This, first of all, means the growth of amount of recycled raw materials and an increase in the quality of processing. Thus, the quality transition under the title “Big Data”, leading to significant changes in the computer technology, has emerged, and that is, no accidentally, called the industrial revolution.

As time passes, multilateral computer programs are drawing closer to the real world. Increasing the volume of unprocessed data necessitates the creation and implementation of tools that enable an effective solution of Big Data analytics problem along with their real-time analysis. As a result, in recent years, the collection and rapid growth of large-scale information, has begun to attract more attention, both in an academic environment and IT industry. According to the statistical report by Miniwatts Marketing Group analysts, in the first quarter of 2015, more than 3 billion people, 42.4% of the planet’s population accessed the Internet, and the number of mobile subscribers reached 7.1 billion. In 2020, the number of devices connected to the Internet is predicted to be USD 50 billion. The amount of digital information in the world was 2.7 zettabyte in 2012. Three times growth of this volume in 2015 and by 40% increase for each following years has been projected [1–5].

However, the rapid growth of digital information, data diversity, and high-speed transfer gives rise to numerous problems. As mentioned, big data storage, its real-time processing, analysis and management has yielded problems. However, only preliminary studies are being conducted on big data (BD) problem, hence this area has not been fully analyzed yet. Realized studies provide opportunities for thorough exploration of Big Data concept, its essence, classification of different characteristics, BD’s resources, the opportunities of this technology, problems and security issues. Studies show that BD processing and analysis requires perfect analytical technologies and tools [6-8].

Recently, influential international organizations, academic institutions, with the participation of major companies in the field of information technology, have organized numerous conferences, symposiums, seminars, forums dedicated to Big Data issues, which discuss and explore the different aspects of the large-scale data processing. Based on several open-ended questions, they specify future research trends and opportunities in this field. These research trends pave the way for the development the best optimal methods to solve BD problems. Therefore, studying scientific and theoretical problems of this technology is of great importance and relevance.

## Big Data Features

Today, the Big Data is a concept used by giant IT vendors and international research agencies. Note that, while exploring Big Data technology and its distinguishing characteristics, we come across different opinions. According to experts, Big Data is a set of technologies and has the following characteristics [9–16]:

– ***Huge volume***. At present, the volume of BD is characterized by from terabyte the up to zettabyte. Today, even the volume of enterprise-level data no longer surprises anyone.

– ***A large number of data sources***. Obviously, Business Intelligence (BI) includes the methods and tools needed to convert unprocessed information into meaningful and convenient form. Traditional BI has several information sources, while Big Data has tens and thousands of external sources.

– ***Unstructured data***. Traditional Data Warehouse (DW) is created on bases of relational database, and it is a tool designed to search for structured data. In this case, there is a need to develop new tools for addressing the issues such as collection and storage, search, security, and analysis of hundreds of terabytes and Exabyte of different types of unstructured data, which cannot be processed by traditional relational databases. Most Big Data sources include unstructured and sem-structured data. In this case, Big Data approach principally offers more advanced template-based search solutions, which is in turn, implies data storage structure different from relational database.

– ***Dynamically changing data.*** It is characterized by the accumulation of large information *massifs*, which is, now, a big challenge faced by any organization. The great volume creates the storage problem, and requires a large-scale storage and distributed processing. Even the filtered data is becoming increasingly expensive. It is known that storage technologies are expensive and their prices falls more slowly than the development of new sources. In this regard, the organization should exactly determine the duration of the data storage. For example, the organization may require some data for many years, while the others become useless several hours after analysts get what they need.

Recently, the application of “grid” and “cloud computing” technologies in data processing and storage service has almost managed to overcome storage problems. Thus, Big Data solves the similar issues solved by the traditional BI tools, in a broader context, in the stages of volume, source, structure and distribution. As a result, in the technological level, Big Data and BI significantly differ from each other (Table 1) [17–21].

Table 1.

Distinct characteristics of Big Data and Business Intelligence

| Traditional BI | Big Data |
|---|---|
| Single Corporate | Warehouse may be located in the distributed file system |
| Data format to be adapted to the requirements of processing functions | Processing functions to be adapted to a variety of data formats |
| Structured data | Designed to work with both structured data and unstructured data |
| Historical information | Working with the newest information |

| | |
|---|---|
| Data acquisition, transmission and processing according to the consecutive and accurate procedures | Massively Parallel Processing (MPP) |

**Technological phases of Big Data processing**

At present, the majority of the scientific-research works aim at solving technological problems of BD. Main sources of Big Data include sensor and social networks, information transmitters about a wide range of fields, banking transactions, Geographic Information Systems (GIS), Global Positioning System (GPS), scientific experiments, e-mail, digital photos and videos acquired via smartphones, large market companies, large sales centers, domain name servers (DNS) and so on. The efficient use of large-scale data generated by these sources requires proper technologies. These are data collection, storage and analytical processing systems. The following technology groups can be used in different phases [22]:

- Analytical algorithms;
- Parallel programming methods;
- Cloud computing resources;
- Computing systems from PCs to strategic supercomputers;
- Storage systems;
- Networks;
- Various input devices from complex telescopes and tomography, to radio frequency identification (RFID) technologies, and so on.

The higher levels are, the higher proximity of the data with the knowledge increases. Analytics completes the process of converting the data into knowledge. Lower levels in the list are based on tested scientific and engineering solutions, while the top levels are underdeveloped as they are new and do not have enough scientific support. Experts directly include analytical systems, data cloud storage service, Hadoop and MapReduce technologies and NoSQL distributed database management systems (DBMS) and others on massively parallel processing (MPP) in to the IT category for Big Data. Apache Software Foundation project Hadoop is an open access system of MapReduce model and more widespread and main platform for large-scale (petabyte) data processing and analysis in distributed computing environment. Hadoop is comprised of two main components: Hadoop MapReduce and the Hadoop Distributed File System (HDFS). MapReduce deals with parallel computing, while HDFS tackles data and distributed file system management [23–27].

New devices that provide human-computer interaction may cause the next wave of Big data. They are e-paper, tactile feedback technology (haptics), various video visors, as well as special memory products and things, open-access systems like Wikipedia and so on. Obviously, in a broad sense, technology is necessary knowledge capital required for production of goods and services. Though in applied sense, it is a method needed to control the transformation of energy and information, processing and re-processing of materials, and assembly of finished products, quality and management during the process of production. Machines and mechanisms in the technological sequence transform initial raw material into a product ready for consumption. The current revolution approves more obvious “data technologies” refusing amorphous and non-specific “information technology”. In the start of the technological sequence, the raw data, and in the end, the data ready for use should be included. In this aspect, modern technologies should provide the conversion of the data into new knowledge or the acquisition of knowledge [28, 29].

One of the main distinguishing features of data technologies is that they are working for human beings. Except auto-installed systems, all other computer systems ultimately serve for generating data used by people; only these systems are able to convert data first into information and then into knowledge. The controversy of such technology is that big data problem arises when the volume of data is constantly increasing in the input of technology sequence. Disproportion between the input and outlet of the volume of data essentially determines the main direction of the development of data technologies: the input flow should be controlled without losing any included data, and then all significant data should be selected and presented in an understandable. The current economic crisis hasn't stopped the development of exaflood-related fields (followed by petabyte – exaflood equals to 1018 bytes, combination of exabyte and flood). But, it accelerated the development of Complex Event Processing (CEP), which plays the role of unimportant data filter. This group technology solves one of the key issues associated with BD, i.e., it eases the transition from raw data to obtaining new, secret and useful information [30].

Today, the most advanced methods are provided for the extraction of useful information from large-scale data; technological tools are developed to set up an effective problem solution. Experts distinguish seven main stages of the technological sequence of big data processing, regardless of its field of application [31–33]:

- ***Data collection.*** Raw data may be collected from data warehouses, facility transmitters, and network resources. This stage is more traditional from the engineering point of view, but the type of data is also to be distinguished, for example, text data should not be confused with the numeral data.
- ***Syntactic and grammatical analysis of data***. It implements data structuring, distribution by their categories, and the analysis on several levels. Low-level includes physical signal processing, compressed files disclosure and decryption. Byte-level includes extracting text, media and other files, and the grammatical and structural analysis on the text-level.
- ***Filtration***. It requires only useful data, reduces the input flow without requesting data analysis methods, such as complex event processing tools.
- ***Data acquisition***. It includes statistical and other methods of data acquisition, and provides data solution in a suitable mathematical context.
- ***Presentation***. Data requires better presentation (diagrams, lists, trees, etc.). Data visualization method varies ranging from simple tables and graphics to complex two and three-dimensional images.
- ***Presentation improvements***. It edits data presentation forms.
- ***Interaction***. It develops data manipulation tricks and more visual presentation methods that provides active processing performance.

Data collection and analysis takes place through traditional technologies; filtration and acquisition is a subject of a new science, and data presentation and specification is included into the field of graphic design.

## Big Data analytics

Collection, management, storage of hundreds of terabyte and Exabyte of data through available methodologies or tools, and extracting useful information from them is a serious problem. The main issues of the Big Data Analytics are working with both structured and unstructured data, conducting more in-depth analysis, and the visualization of the of analysis results. Increasing data volume and the need for its real-time analysis has led to the emergence of Big Data Analytics (BDA) [28, 34–35].

Big Data Analytics is the process of studying large-scale data collection of different types. In other words, it detects hidden patterns of data, its unknown correlations and other useful business information. Analytical data can lead to more efficient marketing, new revenue opportunities, improving the quality of service for customers, and increasing the efficiency of the work, and competitive and other business advantages. The distinguished features of this direction such as the large volume, velocity and complexity require proper technologies. Therefore, available major manufacturers in the field of Big Data Analytics offer exceptional software and hardware systems: SAP HANA, Oracle Big Data Appliance, Oracle Exadata Database Machine, Oracle Exalytics Business Intelligence Machine, the Teradata Extreme Performance Appliance, NetApp E-Series Storage Technology IBM Netezza Data Appliance, EMC Greenplum, Vertic

Analytics Platform based on HP Converged Infrastructure. In addition, small and start-up companies also offer software and hardware tools for efficient data processing, such as Cloudera, DataStax, Northscale, Splunk, Palantir, Factual, Kognitio, Datameer, TellApart, Paraccel, Hortonworks [36–38].

The data is processed to obtain information, the volume of which must be so large that people can convert it into knowledge. Volume is the most important characteristic of big data. Three groups of BD are distinguished according to its volume [22, 39–41]:

- Fast Data - measured in terabytes;
- Big Analytics - measured in petabytes;
- Deep Insight - measured in exabytes and zettabytes.

The groups differ from each other not only by the volume of data, but also in their processing quality. The volume of data reflected in the statistics urges the experts to develop new methods and tools in this field once again.

For the Fast Data, processing does not involve the acquisition of new knowledge, and its results are coordinated with a prior knowledge; it controls how these and other processes are operating, at the same time, it allows us watching the processes in details, confirming or denying any hypothesis. Only a small part of the available technology works for Fast Data solutions, such as Greenplum, Netezza, Oracle Exadata, Teradata, Verica type DBMS and others, which are working with warehouses. The velocity of these technologies is growing simultaneously with the increasing volume of data.

The issues solved with Big Analytics tools differ quantitatively and qualitatively very much. The proper technologies convert the data into new knowledge, by helping the acquisition of new knowledge and useful information. In other words, decision-making does not consider artificial intelligence technology; instead, analytical system is based on the “teacher training” principle, and its all analytical potential are applied during the learning process. Classic examples of such analytics are MATLAB, SAS, Revolution R, Apache Hive, Apache Mahout SciPy [42–44].

Deep Insight considers the use of unsupervised learning and modern methods of analytics, as well as various visualization methods. It may reveal a prior knowledge and arrangements.

In terms of quality, Big Data Analytics programs require not only new technologies, but also new thinking. Analytics is considered separately from the initial data development tools, visualization and other technologies that present the results. Even the views of the Data Warehousing Institute toward analytics are different. The organization reports that currently 38% of the organizations are exploring the opportunities of Advanced Analytics in the management practice, while 50% of them are planning this within the next three years. It should be noted that such interest for this area is justified by bringing business argument. Thus, more advanced management system is required in new businesses conditions. Its creation starts up with the establishment of contacts that are decision support systems, and as a result, automate decision-making will be available in the future. The development of automated control systems of technological objects is not a new problem. This issue has long been existed in the database, distributed database, and creation of the architecture for joint use of resources has also contributed to the solution of this issue [45].

New analysis tools are required since there are many data sources, as well as, due to their variety of formats (structured, unstructured, semi-structured), and the use of different indexation schemes (relational, multidimensional, noSQL). Accordingly, large-scale data collection, processing and shaping for analysis becomes very difficult. Traditional methods are no longer valuable for working with data. Since Big Data Analytics is applied to much larger and more complex massifs, Discovery Analytics and Exploratory Analytics terms are also used. Regardless of the names, they are essentially the same - requiring feedback to provide decision-makers with information about the various processes [41, 46–48].

These days, modern IT factors: big data, analytics and cloud technologies are impossible to be imagined apart from each other. Increased demand for multilevel storage systems, the availability of cloud technologies in real terms has also increased the interest in BD analytics. Cloud technologies are one of the extremely successful approaches to the complex computing; the work with BD is not possible without cloud storage and cloud computing. The large volume of digital information are managed through "cloud" services as IaaS (Infrastructure as a service), PaaS (Platform as a Service), and SaaS (Software as a service) and stored in a centralized way. Giant IT companies pay considerable attention to the very scaling aspects and multi-level data storage in new generation storage systems [50, 51].

As practice shows that, today, it is required to download many systems for the performance of analytical issues. Nevertheless, business requires all services, applications and data to be always accessible. In addition, currently, the need for the results of analytical studies is very high, because of the accurate, literate and timely analytical processes that significantly increase the business productivity.

***Big data analysis issues.*** Today, any approaches and methods are attempted for the development of DW technologies for BD analysis. In addition, some features of traditional operations may contradict to the specific BD processing. Significant difference between operational and analytical processing issues was observed in the beginning of the database technology development. The term of DW – "Data Warehouse" was first proposed by Bill Inmon in the 70s, but the interest in this technology rose 20 years later, when the demand for such systems increased and the necessary computing power became available [52].

Data processing stages in DW consists of data *acquisition, cleaning, overloading, analysis*, and finally the *presentation* of the analysis results. Each of these stages implements special data operations. It should be noted that, if the application of DW is attempted for BD analysis, then, not only the analysis of algorithms, but also all stages should be considered.

***Data acquisition***. DW is aimed at cleaning, coordination, integration of data collected from separate processing systems, and shaping it for analysis (Figure 1). DW provides that information is extracted from the operational database, converted, checked and only then. Downloaded to the system. Hence, data development technology in DW comprises three associated stages [53]:

1. Data Acquisition.
2. Data Cleaning.
3. Data Consolidation.

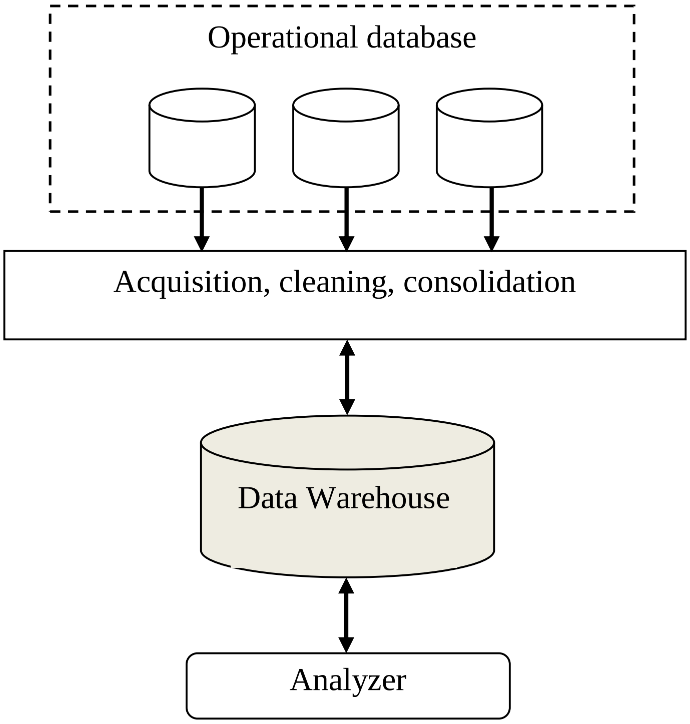

Figure 1. Data acquisition in data warehouse in traditional analysis

The above-mentioned transactions are carried out periodically. However, while working with BD, periodicity is not always possible in order to provide data analysis include from various sources. The period between the data acquisition and availability may be shorter than that needed for the implementation of operations, when establishing DW. For example, the monitoring carried out to detect the information quickly spread on social networks and to identify its sources, to define active users, and to detect negative expressions or the facts of confidential information leakage [54].

Obviously, all these incidents should be identified and neutralized more quickly. However, an informal data description is available, the processing of which (initially not distinguished with high speed) requires text-mining algorithms. For example, in social networks monitoring expressions, comments, evaluations, photographs of users and so on are presented. Apparently, popular social networks can be monitored with multiple users. Moreover, it is clear that, high number of users and their high sensitivity implies collecting and processing of large volumes of informally described data [55]. During the initial processing of data in DW (for example, looking for inconsistencies) the structure (data), previously collected in the warehouse, is considered to be used, which is also hard to be fulfilled while working with BD. In other words, the problem is that this data is always distributed to be suitable not only for analysis, but also for gathering. For example, if it comes to telecommunications systems, the data is collected from the regional servers (Figure 2).

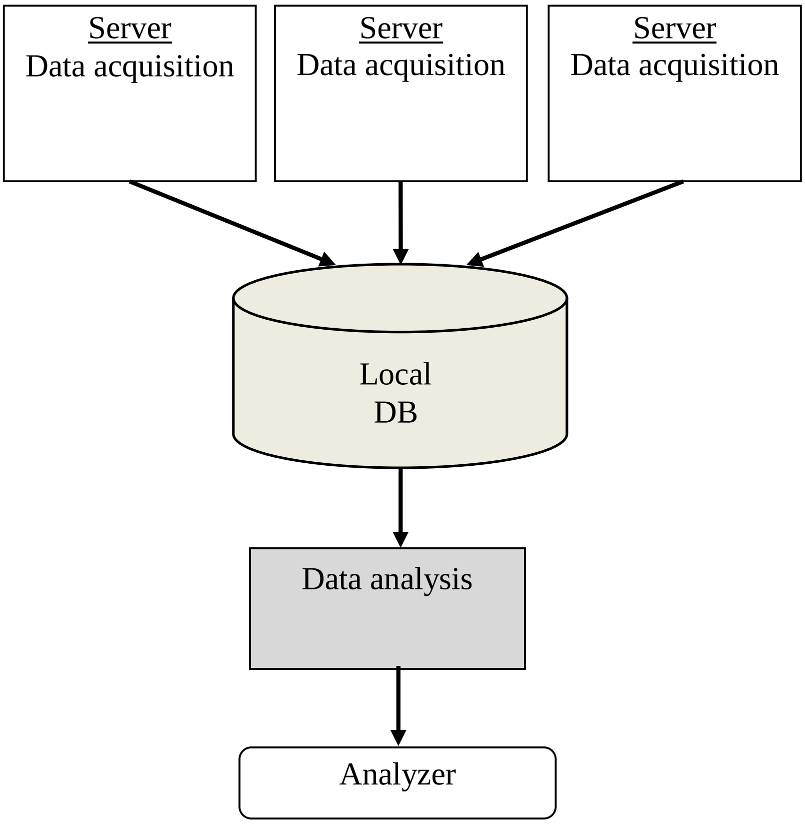

Figure 2. Big Data Analysis

Thus, the analysis has shown that all data in the traditional DW always goes through a single logical blocks, where they are converted, checked, cleaned, loaded, and the period of implementation of these operations is rarely important for the rest of the system. However, a single block is impossible during BD processing. It should be noted that issues of intensive input data flow are not many thus far, but acquisition, cleaning, conversion and loading block can be implemented in the form logically integrated and distributed system.

***Big data analysis.*** Operational analysis of large-scale data for various purposes has always concerned the managers. The analysis of the scientific and technical literature, the applied products, as well as other analytical systems still do not fully meet all the requirements. The companies as Gartner, McKinsey Global Institute, IDC (International Data Corporation) and others define the BD as a growing, and dynamically developing field. McKinsey Global Institute presents analysis methods applicable to the BD in its analytical report [5]:

- *Data Mining Methods* – association rule learning, classification, cluster analysis, regression analysis etc. .;
- *online analytical processing* (OLAP);
- *crowdsourcing* - data processing through a group of people involved by the social contract without labor relations;
- *data fusion and integration* - technical tools to integrate data from different sources in order to provide deeper analysis. For example, digital signal processing and natural language processing (including tonal analysis);
- *machine learning* – includes with and without teacher learning, as well as Ensemble learning, statistical analysis based on base models for complex predictions or constituent models (if a model is a component of a larger one);
- *artificial neural networks, network analysis, optimization*, including *genetic algorithms*;
- *recognition of images*;
- *predictive analytics*;
- *imitation modeling*;
- *spatial analysis* – the methods that use topological, geometric, and geographical data;
- *statistical analysis*, as A/B testing and time series analysis;
- *analytical data visualization* – obtaining results, further analysis, and presenting the initial data using figures, diagrams, interactive features and animation.

All classic B-trees and available methods are appropriate for multi-dimensional operations on BD. Today, traditional DWs offer nearly identical sets of tools for data analysis [56-58]. Besides, the products such as SAP HANA, Greenplum Chorus, Oracle Exalytics, Oracle Exadata, Aster Data nCluster are also available, which enable the operation of the listed methods in BD, and realize analytical processing of terabytes of data. In addition, there are hardware and software tools effectively processing terabytes and exabytes of data based on relational data management systems offered by Netezza, Teradata, Greenplum and other traditional companies. To understand the potential of such solutions, it is important to review the methods based on these algorithms, and to analyze their possible ways for parallel computing, which is the key for BD processing. In this case, the main characteristic parameters for BD (network interaction intensity, volume, velocity, etc.) should be considered, but not the specific distributed data processing technology. The number of input/output operations and efficiency of the developed indices are important factors affecting the pace of work of any DBMS. As the first step in solving the issue when working with BD, it is recommended to classify the data to be processed by its format, the type of analysis, applied the processing methods, as well as by the sources of the data to be submitted, loaded, processed and analyzed by the system.

OLAP (On-line Analytical Processing) – is on-line operational data, analytics, multi-dimensional processing to support management decisions. The method aims at building multidimensional cube and obtaining its different cross sections. By organizing the analytical data as multidimensional cube, we obtain its simple and meaningful model. The results often include a table, which contains aggregate indicators (number, average, minimum or maximum value, etc.) in the fields. OLAP system is comprised by four stages:

- Data acquisition;
- Data storage;

- Loading analytic subsystems in multidimensional cubes;
- Data description.

Depending on the realization, that is where the data to be analyzed is stored, the following types of multidimensional data analysis systems are distinguished [53]:

- MOLAP - Multidimensional OLAP. Both the data and the multidimensional aggregates are stored in a database.
- ROLAP - Relational OLAP. Data is stored in their previous relational databases, and the aggregates are stored in the special working tables developed in the same databases.
- HOLAP - Hybrid OLAP. Data is stored in relational databases, and the aggregates in the multidimensional ones.
- DOLAP - Desktop OLAP. Designed for multi-dimensional analysis supporting not the local multiuser mode, and used in small-scale data processing.

ROLAP is the most common transparent and well-studied systems based on relational DBMS. Internal structures of MOLAP and HOLAP systems are usually more closed, and related to specific commercial know-how products. MOLAP describes the data in the multidimensional model, whereas, internally, it uses “star” and “snowflake” schemes from ROLAP. In terms of DBMS, ROLAP is a usual relational database, and therefore supports all operations. However, there are a number of shortcomings. For example, the phase of data submission is impossible to be controlled, and the optimum structure is unavailable in order to collect statistics and to store the indices. It is impossible to optimize the placement of the disc to ensure high data input/output rate, and the mean aggregate value cashing is unattainable. Finally, in-depth statistical analysis cannot be conducted and the optimal solution plan is not practical due to the demand for high velocity during the performance of analytical requests. ROLAP uses relational request optimizations, which does not take into account the multidimensionality of the database. MOLAP does not have such drawbacks, and therefore it enables high analysis rate [59].

During the BD analysis, selection of MOLAP, ROLAP, HOLAP technologies depends on the frequency of the database updating. In terms of parallel computing of processing, at first glance, any multi-dimensional cube can be “cut” by the sections of one of the measures and distributed among multiple servers. For example, the cube may be distinguished by time intervals (by years and months), by the location (each server is responsible for its region) and etc. It is known that, not only one, but also numerous servers are responsible for the implementation of a multidimensional request in cube division, and then, the results are stored as a whole. For instance, if a user request for statistics by the country within a given period, and if the data is distributed by several regional OLAP-servers, then each server replies separately and as a result, all the data is collected into one place. If the data is distributed by the time criteria, then a single server is loaded when implementing the examined request. Such a situation gives rise to the following problems [46]:

- It becomes difficult to predetermine optimal data distribution by servers;
- It becomes unknown which data and servers are needed for a part of the analytical requests.

In BD, it means that available approaches to multidimensional analysis are scaled well and they enable the acquisition of distributed data (Figure 2). Hence, each server may collect independent information, clean and download it on a local database.

***Clustering method***. It is known that, Clustering is very vital for data analysis and data mining. Clustering is the mathematical model of multidimensional analysis, and it bases on the sets of indicators characterizing several objects and break them into clusters, where the objects included in each cluster become the same and similar compared to the others included in another cluster. Clustering is based on the distance between the objects’ parameters specified in figures.

Cluster analysis also includes the concept of metrics for quantity measurement of similarity. Thus, the metrics option is one of the most important conditions for clustering analysis (objects’ proximity). The similarity or difference between the classified objects is determined by the metric distance between them (two units submitted, in the exit, their similarity rate becomes zero for fully identical units). In cluster analysis, various distance measures between the objects can be used. Finding distance measures and weight coefficients of the classified variables is a very important phase of cluster analysis. At present, clustering is often the first step taken in the data analysis. The algorithms for clustering methods can be classified in accurate and inaccurate, hierarchical and non-hierarchical, iterative and other methods [56, 60–61].

The problem of BD's clustering is that available algorithms provide the possibility of direct request for any information from the initial data unit. It means that it is impossible to find which units are required in advance. The initial data, in turn, can be distributed on different servers, in this case, each cluster is not necessarily stored on a server. If the distribution of data on the servers is set for clustering algorithm (i.e., the algorithm finds that the data is located in the distributed virtual memory), then it will lead to the migration of large-scale data from one server into another. In this case, the solution to this problem will be as follows: in each server one algorithm is enabled, which is operating with the data of only that server, and in the end, the parameters of found clusters and the weights measured depending on the number of elements within the cluster are given. Later, obtained data is collected in the central server and meta-clustering is implemented, neighboring cluster group is selected by taking into account the weights. This method is universal, and provides a good parallel processing and can be used by other clustering algorithms. Nonetheless, it requires precise research, real data testing and comparison of the obtained results with other local methods [46].

Apparently, there is no single universal clustering algorithm. The use of any algorithm requires understanding its advantages and shortcomings, and the consideration of the nature of the better working data and its scalability. Studies show that, most clustering methods are not so useful for the BD analysis, and additional studies should be conducted. Clustering algorithms improvement ways should be continually investigated [62].

***Classification method***. The classic classification problem is similar to regression issue, which is the establishment and use of dependence of a variable on others. For example, if there is a database of top-level domain zones values, we can set up the system of rules that provides the approximate value of the new domain name to be obtained appropriate to the zone in advance. The classification differs from the regression with the fact that time series analysis is not conducted and the values submitted in the entry can not be regulated. Recently, many classification methods have been developed (Bayes functions, solutions trees, neural networks, etc.), each of which has a well-developed scientific theory (self-study systems, and supervised learning methods) [63–64].

Research shows that the classification methods are based on the same scheme. Initially, the algorithm is studied within a relatively small selection, and then, the obtained rules are applied in the rest of the options. In the first phase, the data massifs can be transferred to a server without the parallel computing for the functioning of classical learning algorithm. However, in the second phase, the data can be processed independently. In this case, the system of rules resulting from self-study is transferred to each server, and the data massifs stored on this server are transmitted through this system. The attained results may be stored on the server or sent for further processing. Thus, in the learning phase, BD is not processed. Consequently, there is no any option of such volume developed to study the systems; in the classification phase, of the separate parts of the data are processed independently. This means that the current classification methods are appropriate for working with BD [46].

***Search for patterns***. Analysis shows that search algorithms for association rules are more commonly used in the automated detection of new patterns from large-scaled data. Search algorithms for association rules have broad areas of application. They are successfully implemented in trade, medicine, Web Mining, Text Mining, census data analysis, forecasting and analysis of telecommunications equipment failures and so forth. Association rules determine the correlation between the different characteristics. In other words, association rules enable to find correlations between the related events. In association rules, the purpose of the analysis is to establish the following dependencies: if there is a set (collection) of a few elements *X* in the transaction, then it can be concluded that another set of elements *Y* should also be created in this transaction. The establishment of such interdependencies allows to find very simple and intuitive rules.

Search algorithms for association rules have been defined for finding all the rules *X* and *Y*. In addition, support and accuracy coefficients of these rules should be greater than some pre-set limit values, which are called minsupport and minconfidence, respectively. From analytic and practical point of view, the accuracy and support coefficients of these rules are of great importance, but balance among these indicators is the practical matter of question. Finding associative rules comprises two sub-problems [65]:

1. Finding a set of all the elements that provide minsupport limit. Such element sets are most common.
2. Generating the rules from the set of elements obtained with the accuracy, that provide minconfidence limit according to the Paragraph 1.

Association rule:

- (first) to be important, that is a set of  elements *X* and *Y* of the researched data should come across quite frequently;
- to be precise, that is the share of the transactions comprising the set Y in the transactions comprising the set X should be high;
- to be interesting, that is the presence of the set X in the transaction should increase the probability of the presence of the set *Y* in the same transaction.

Pattern search problem is solved with the help of Apriori algorithm. Obviously, in terms of BD processing, the main operation is the calculation of aggregation function, which is accomplished through multi-dimensional analysis. Another important point of Apriori algorithm, which prevents BD processing, is the ability to work with the set of information features. Nevertheless, we should take into account that the number of reviewed collection depends on the number of information features, but not on the conceptual data model and their volume. The modification of Apriori algorithm for the BD processing is based on the method of aggregation function calculation, and therefore, parallel computing algorithms of multidimensional analysis can be used here.

***Regression analysis***. Regression analysis is a method of statistical analysis of the impact of one or several independent variables on others. It aims at exploring the relationship between one dependent variable and one or several independent variables. It is necessary to explain the behavior of independent variables, which are systematically affected by dependent ones, and to isolate them from random effects.

Regression is building a parametric function, which describes the change of the numerical quantity in the given period. This function is built based on known data, then used to find the further values this quantity in advance. The sequences of “time-value” couple, such as the sales rate of a particular products in the specific region, is submitted to the method entry, which describes the status of the quantity in the given conditions. Function parameters that describe the state of the investigated quantity are obtained in the output.

Regardless of the type of the used parametric function, the values of its parameters are chosen in the same way. The common difference between the values achieved by the function with the current values of the parameters (i.e., given in the method entry) is calculated with the observed values of the quantity. Then, to reduce the difference, it is determined how the parameters are regulated. These operations are repeated until the sum difference reaches its minimum value or its subsequent decline.

In terms of data processing, calculating available difference and defining parameters are the basic operations in regression analysis. If the first operation is carried out in parallel clearly (the sum is partially calculated in separate servers, and then concentrated on the central server), the second part will be difficult. In general, well-known mathematical fact is used to correct the weights: function of some parameter increases toward gradient, and declines in an opposite direction against the gradient. In turn, the calculation of the gradient involves the calculation of special derivative of the function by each parameter, which is reduced to the discrete differensialization based on calculation of the sum of the weights. Consequently, the values of the parameters are defined and reduced to the sum to be performed in parallel. If regression analysis is reduced to the calculation of the sum of the weights, then it has the same application degree as the multidimensional analysis when working with BD. In other words, the systems working with regression analysis may be completely scaled and may operate in terms of distributed data collection. Thus, since the available algorithms for data analysis are capable to implement parallel computing, they potentially can be used for the analysis of BD [46, 66].

General characteristics describing Big Data, huge volume, high-velocity and variety of data encounter BD analysis and make it difficult to obtain new knowledge and useful information. These characteristics reflect the basic problems of big data and urge the analytics architecture scaling. Currently, no single tool is available to cope with these requirements. As a result, many hybrid architecture has been developed. Advanced Analytics Platform was developed in the course of practical experiments and provides the necessary capabilities for working with data.

***Visualization problems of the analysis results***. Investigations show that the majority of the scientific studies dedicated to the BD analysis mainly focus on the analysis itself and processing of findings is overlooked. It is considered that available methods are likely to be applied as the generation of reports, as well as the establishment of various types of charts or graphics. However, existing methods cannot be applied to review the results of analysis due to following reasons [46]:

- Large volume of submitted data generates a large number of results to be analyzed in the end. Obviously, many statistical regularities are able to overcome errors. However, only the main regularities are not enough for decision making. To achieve maximum efficiency in decision making in BD analysis, very little different regularities and trends should also be considered. Otherwise, processing of the most various data flows will not make sense at all;
- Conceptual model of the final information becomes very difficult. A typical report of DW does not include more parameters (e.g., time interval, region, etc.) than that, because the report becomes artificial, composed of zeros and empty rows. However, this is not the case for BD.

In the analysis issues, as the volume of preliminary data increases, appropriate analysis methods are used first in the level of simple search and data review, then in the level of multidimensional and statistical analysis, and after that in the level of Data Mining. However, the more preliminary data increases in volume, the more output information is obtained in the Data Mining level. In other words, before, several report sheets had to be studied for decision making, yet it is not the case for BD. Thus, the decision maker challenges with another problem, which is selection of the most important and significant information. The problem can be solved with the use of automated tools and by means of reorganization [67].

Automated tools allows us choosing more important reports, for example, such reports are chosen during the monitoring of the sales dynamics, so that a sharp change of indicators is observed compared to previous years. However, such methods are not always applied. Thus, the change in the indicators can be simply attributed to external factors unknown to the system, such as the fall in oil prices can be explained by the increase in the price of cars, and it does not mean that such sales volume should be planned for the next period. Extracting the most important information from the large number of reports is based on the re-organization of the work. In this case, individuals are responsible for review of the reports and formulation of the summaries sent to the decision-makers (Figure 3).

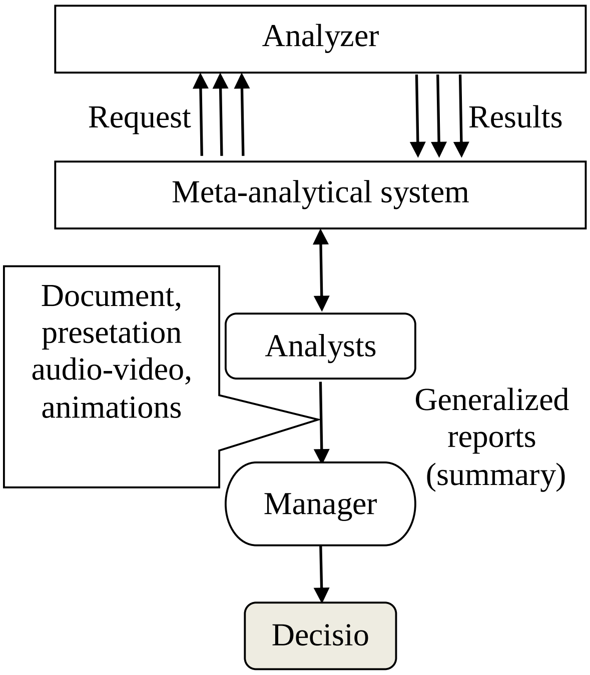

Figure 3. Reorganization of Big Data analysis

The generation of such summaries considerably differs from the available tools for report generation for the following reasons [46]:

- Summary may include different types of information. In other words, the document may comprise everything from sales indicators, price growth rate, as well as the changes in the staff list and even photos;
- All summaries are unique and may not be in common format;
- Summary is always developed for a specific case for a particular person, as a result, is takes into account the specificities of the circumstances and characteristics of the person. This may be printed document, any presentation, audio or video material.

In short, a summary should be conducive to obtaining the information as quickly as possible and visually for decision-making. Above-mentioned requirements make up the concept of automated meta-analytical systems, which enable the visualization of the results of the analysis, and based on them, the generation of documents and presentations for only review, as well as montage of audio and video materials, and the creation of Flash animations. However, these methods are not as simple as it seems at first glance; they focus on the choice and maximum visual presentation of variety of information. Manual solution of this problem creates a huge burden for staff; at least partially automated performance requires interdisciplinary research for finding methods to present the information in the summary more effectively. Currently, such systems may seem unnecessary. On the other hand, as the number of practical issues associated with BD analysis increases, and there will be a need for the preparation of reports in a very swift and easy fashion. Otherwise, decision-making staff will “sink” in the ocean of reports and analytical results.

It should be noted that, in some cases, effective methods used in DW can be applicable for working with BD, and some available algorithms can be adapted for the processing of large-scale distributed data. In addition, serious problems can arise in the visual presentation of the results. Due to the large volume of submitted information, the number of various reports sharply increases in output. New software fundamentally distinct from the report generators used in the traditional warehouses is required.

## Conclusion

In the course of exploring the analysis methods applied in BD, the advantages and disadvantages of each of them have been analyzed. The increase in the volume of data, the need for their real-time analysis requires the creation and implementation of tools that enable an effective solution of Big Data analytics. IT giants mainly use big Data analytics tools and business-analytics technologies, which is due to the reduced use of business-analytics in the enterprises and the difficulty of perception of current analysis methods by business-user. Given this, the analysts of some companies (as Information Builders) offer the product to work with the data from any source in real time, which is deemed easy for using.

Big Data is not a simple idea, but a symbol of the technological revolution. Therefore, philosophically, Big Data lays the foundations of a new cognition for the civilization. The need for the analytical work with BD provokes the rise of new software and hardware platforms, which will entirely change IT industry. Today, the most advanced methods are provided for the analysis of large volumes of data: artificial neural networks – the models based on the principle of biological neural networks and function, predictive analytics, statistics, Natural Language Processing methods, including other methods or crowdsourcing, as A/B testing, sentiment analysis and so on.

Available methods as cloud tags and the latest Clustergram, History Flow and Spatial Information Flow are used for visualization of the results. BD technology supports Google File System, Cassandra, HBase, distributed file systems Lustre and ZFS, MapReduce and Hadoop software constructors and other decisions. According to the experts, BD will mainly transform manufacturing, healthcare, trade, and administrative areas.

Thus, while looking at the global processes and the experience of leading countries, it becomes clear that various models and conceptual approaches are offered for the real-time analytical processing and analysis of the data that automatically and continuously generated from different sources. In the view of this important fact, the international studies associated with BD analytics should be thoroughly explored. Obviously, as the large-scale digital data collection is the major problem faced by all spheres of society, the related research works are of great importance for numerous fields of science. Because, raw and unstructured data is a source of the knowledge that holds the power of making fundamental changes in all spheres of society. In this regard, the development of scientific and theoretical basis of all aspects of Big Data technology is significantly crucial.

## References

1. Miniwatts Marketing Group, Worldwide Internet Market Research, www.miniwatts.com
2. The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. www.emc.com/leadership/digital-universe
3. Worldwide Big Data Technology and Services 2013-2017 Forecast, http://www.idc.com
4. Data Science Central, The online resource for Big Data practitioners, www.datasciencecentral.com
5. Big data: The next frontier for innovation, competition, and productivity. Analyst report, McKinsey Global Institute, May 2011. http://www.mckinsey.com
6. Madden S. From Databases to Big Data // IEEE Internet Computing, 2012, vol.16, no.3, pp.4–6.
7. What is big data? - Bringing big data to the enterprise, 2013. http://www-01.ibm.com
8. Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety. Technical report, META Group, Inc (now Gartner, Inc.), February 2001. http://blogs.gartner.com
9. Clifford L. Big data: How do your data grow? // Nature, 2008, vol.455, pp.28–29.
10. The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. www.emc.com/leadership/digital-universe
11. Wei Fan, Albert Bifet. Mining big data: current status, and forecast to the future // ACM SIGKDD Explorations Newsletter, 2012, vol.14, no.2, pp.1–5.

12. Maté A., Llorens H., Gregorio E. An integrated multidimensional modeling approach to access big data in business intelligence platforms / Proceedings of the 2012 international conference on Advances in Conceptual Modeling (ER'12), Heidelberg, 2012, pp.111–120.
13. Szalay A., Gray J. 2020 Computing: Science in an exponential world // Nature, 2006, vol. 440, pp.413–414.
14. McAfee A., Brynjolfsson E. Big Data: The Management Revolution // Harvard Business Review, 2012, vol.90, no.10, p.60–68.
15. Birke R., Björkqvist M., Chen L. Y., Smirni E., Engbersen T. (Big)data in a virtualized world: volume, velocity, and variety in cloud datacenters / Proceedings of the 12th USENIX conference on File and Storage Technologies (FAST'14), USENIX Association Berkeley, CA, USA, 2014, pp.177–189.
16. Richard Price. Volume, velocity and variety: key challenges for mining large volumes of multimedia information // Proceedings of the 7th Australasian Data Mining Conference (AusDM '08), Australia, 2008, vol.87, p.17.
17. Chiang R.H.L., Goes P., Stohr E.A. Business Intelligence and Analytics Education, and Program Development: A Unique Opportunity for the Information Systems Discipline // ACM Transactions on Management Information Systems (TMIS), 2012, vol.3, no.3, Article 12 (pp.12:1-12:13).
18. Chen H., Chiang R.H. L., Storey V.C. Business intelligence and analytics: from big data to big impact // Journal MIS Quarterly, 2012, vol.36, no.4, pp.1165–1188.
19. Omar El-Gayar, Prem Timsina. Opportunities for Business Intelligence and Big Data Analytics in Evidence Based Medicine / HICSS '14 Proceedings of the 2014 47th Hawaii International Conference on System Sciences( HICSS '14), USA, 2014, pp.749–757.
20. Statchuk C., Iles M., Thomas F. Big data and analytics / Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research (CASCON '13), USA, 2013, pp.341–343.
21. Foster Y., Kesselman C., Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations // Intern. J. of High Performance Computing Applications, 2001, vol.15, no. 3, pp.200–222.
22. Leonid Chernyak. Big Data - A new Theory and Practice // Open Systems, 2011, No 10, pp.18–25.
23. Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters // Communications of the ACM, 2008, vol.5, no.1, pp.107–113.
24. Lee K-H., Lee Y-J., Choi H., Chung Y.D., Moon B. Parallel data processing with MapReduce: a survey // ACM SIGMOD Record, 2011, vol.40, no.4, pp.11–20.
25. Brunozzi Simone. Big Data and NoSQL with Amazon DynamoDB / Proceedings of the 2012 workshop on Management of big data systems (MBDS '12), USA, 2012, pp.41–42.
26. Weiyi Shang, Zhen Ming Jiang, Hadi Hemmati, Bram Adams, Ahmed E. Hassan, Patrick Martin. Assisting developers of big data analytics applications when deploying on hadoop clouds / Proceedings of the 2013 International Conference on Software Engineering (ICSE '13), NJ, USA, 2013, pp.402–411.
27. Chuck Lamb. Hadoop in Action, Publisher: DMK Press, 2012, p.424.
28. Leonid Chernyak. Calculations with a focus on data // Open Systems, 2008, No 8, pp. 36–39.
29. Wu X., Zhu X., Wu G., Ding W. Data Mining with Big Data // Journal IEEE Transactions on Knowledge and Data Engineering, 2014, vol.26, no.1, pp. 97–107.
30. Wang Y.H., Cao K., Zhang X.M. Complex event processing over distributed probabilistic event streams // Computers & Mathematics with Applications, 2013, vol.66, no.10, pp.1808–1821.
31. Leonid Chernyak. Time of Troubles for database // Open Systems, 2012, No 2, pp. 16–21.
32. Leonid Chernyak. What to do with the chaos of data? // Open Systems 2013, No 9, pp. 16–20.
33. Natalia Dubova. Big Data closeup // Open Systems, 2011, No 10, pp. 30–33.

34. InfoSphere Platform: Big Data Analytics, 2013, http://www-01.ibm.com/software
35. Jacobs A. The pathologies of big data // Communications of the ACM. 2009, vol.52. no.8, pp. 36–44.
36. Vakhrameev Kirill. Database for Big Data analysis // Open Systems, 2011, No 10, pp. 26–29.
37. Babu S., Herodotou H. Massively Parallel Databases and MapReduce Systems // Foundations and Trends in Databases, 2013, vol.5, no.1, pp.1–104.
38. Vignesh Prajapati, Big Data Analytics with R and Hadoop, Publisher: Packt Publishing Ltd, 2013, pp.238.
39. Leonid Chernyak. A fresh look at Big Data // Open Systems 2013, No 7, pp. 48–51.
40. Krish Krishnan. Data Warehousing in the Age of Big Data. 1st Edition, Morgan Kaufmann Publishers Inc. San Francisco, USA, 2013, pp.370.
41. Bill Franks. Taming big data. How to extract knowledge from data arrays using deep analytics, trans. from English. Andrey Baranov, M .: Mann, Ivanov and Ferber, 2014, p. 352.
42. Big Data - What Is It? 2013, http://www.sas.com/big-data
43. MathWorks, http://www.mathworks.com/discovery/big-data-matlab.html
44. Hadoop Distributed File System. http://hadoop.apache.org/docs
45. Witt D., Gray J. Parallel Database Systems: The Future of High Performance Database Systems // Communications of the ACM, 1992, vol.35, no.6, pp. 85–98.
46. Seleznev K. Problems of Big Data analysis // Open Systems 2012, No7, pp. 25–29.
47. Gudivada V.N., Rao D., Raghavan V.V. NoSQL Systems for Big Data Management / Proceedings of the 2014 IEEE World Congress on Services (SERVICES '14), USA, 2014, pp.190–197.
48. Mayer-Shenberger Victor, Kukier Kenneth. Big data. A revolution that will change the way we live, work and think, trans. from English. Inna Gaydyuk, M .: Mann, Ivanov and Ferber, 2013 p. 240.
49. Kenn Slagter, Ching-Hsien Hsu, Yeh-Ching Chung, Daqiang Zhang. An improved partitioning mechanism for optimizing massive data analysis using MapReduce // The Journal of Supercomputing, 2013, vol.66, no.1, pp.539–555.
50. Alguliyev R.M., Hajirahimova M.S. Big data phenomenon: Challenges and Opportunities// Information Technology, 2014, No 2, pp. 3-16.
51. Marcos D. Assunção, Rodrigo N., Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya. Big Data computing and clouds // Journal of Parallel and Distributed Computing, 2015, vol.79, pp. 3–15.
52. Inmon W. H. "Building the Data Warehouse," 3rd Edition, John Wiley & Sons, Inc., New York, 2002, pp.41.2.
53. Alguliyev R.M., Gasimova R.T., Alakbarova I.Y. About modern decision support concepts // ANAS News, physics and mathematics and technical sciences series, 2005, No 2, pp. 70-75.
54. Tonkin E.L., Pfeiffer H.D. Zombies Walk Among Us: Cross-Platform Data Mining for Event Monitoring / Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW '13), USA, 2013, pp.452–459.
55. Krishna Kumar K.P., Geethakumari G. A taxonomy for modelling and analysis of diffusion of (mis)information in social networks // International Journal of Communication Networks and Distributed Systems, Switzerland, 2014, vol. 3, no. 2, pp.119–143.
56. Alguliev R.M., Gasimova R.T. Identification of Categorical Registration Data of Domain Names in Data Warehouse Construction Task // Intelligent Control and Automation, 2013, vol.4, no.2, pp.227–234.
57. Alguliev R.M., Gasimova R.T. On an approach for intellectual analysis of registration data of domain names // International Journal of Ubiquitous Computing and Internationalization, 2011, vol.3, no.1, pp. 27–30.

58. Ordonez C. Can we analyze big data inside a DBMS? / Proceedings of the sixteenth international workshop on Data warehousing and OLAP (DOLAP '13), USA, 2013, pp. 85–92.
59. Alguliev R.M., Gasimova R.T, Alakbarova I.Y. An approach to performing complex queries based on OLAP technology // Information Technology of Simulation and Control, 2006, No 6, pp.728–731.
60. Gasimova R.T. Conceptual basis for the creation of a knowledge base of domain names // News of Baku University. Physical and Mathematical Sciences Series, 2010, No 4, pp. 95–102.
61. Park H.S., Jun C.H. A simple and fast algorithm for K-medoids clustering // Expert Systems with Applications, 2009, vol.36, no.2, pp.3336–3341.
62. Nevsky I.M., Filippovich A.Y. The technique of adaptive clustering factual data based on the integration of MST algorithms and Fuzzy C-means // Proceedings of the higher educational institutions. Printing problems and publishing industry. M.: Publishing house MSUP, 2009, No 3, pp. 48–61.
63. Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan. Automatic Subspace Clustering of High Dimensional Data // Data Mining and Knowledge Discovery, 2005, vol.11, no.1, pp.5–33.
64. Agrawal R., Imielinski T., Swami A. Mining association rules between sets of items in large databases. / Proceedings of the ACM SIGMOD Conference on Management of Data, Washington D.C., May 1993, pp.207–216.
65. Tsai-Hung Fan, Dennis K. J. Lin, Kuang-Fu Cheng. Regression analysis for massive datasets // Journal Data & Knowledge Engineering, 2007, vol.61, no.3, p. 554–562.
66. Abousalh-Neto N.A., Kazgan S. Big data exploration through visual analytics / Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST '12), USA, 2012, pp. 285–286.
67. Phil Simon. The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions. Publisher: Wiley, 2014, 240 p.