Rasim M. Aliguliyev[1], Makrufa Sh. Hajirahimova[2], Aybaniz S. Aliyeva[3]
[1,2,3]Institute of Information Technology of ANAS, Baku, Azerbaijan
[1]secretary@iit.ab.az, [2]makrufa@science.az, [3]aybeniz63@rambler.ru

DOI: 10.25045/jpis.v07.i2.04

CURRENT SCIENTIFIC AND THEORETICAL PROBLEMS OF BIG DATA

A very large size of digital data has been generated in the world in view of the technical and technological development in the recent decade. As a result, a notion of "big data" has emerged; nowadays, it has become an important topic that is broadly discussed in newspaper and journal articles, blogs and etc. Alongside with creating new prospects for the modern society, "Big data" has also brought about some problems for researchers. Unlike the mass media and the business sector, the notion of "big data" is reviewed as a scientific-research object in this article, and a short comment on "big data" concept is provided. The main factors underlying its development as a research direction are presented. Moreover, current scientific and theoretical issues in the focus of researchers are also analyzed.

Keywords*: big data, big data analytics, audio analytics, video analytics, social media analytics, visualization, security.*

Introduction

Starting from the beginning of the 21st century, digital data generated by technologies – computers, mobile phones, Internet, sensor networks, artificial satellites of the Earth, cosmic telescopes, cloud computations and etc. experiences an exponential growth annually. This situation is characterized as an "information explosion" in the society. In this sense, if 5 exabyte of data was generated in total during the period from the existence of humanity till 2003, this indicator constituted 2,7 zetabytes in 2012; this figure is expected to increase by 40% in each next year and reach 44 zetabytes by 2020 [1]. With the rapid increase of digital data, a notion of "big data" has emerged reflecting the new age in the processing, storage and use of data [2]. This notion is intended to specify the mass of big data which cannot be processed by the current management methods and intellectual analysis tools in terms of volume and complexity [3].

As a phenomenal event, *big data* has attracted each segment of the society in the short period of time. It is due to the reason that big data (BD) has a large potential of revolutionary changes in management and business, large profit generation in enterprises, development and realization of scientific ideas in various fields [3,4]. The data analysis, and the extraction of knowledge and useful information from those play an important role in the realization of new scientific discoveries, in well-founded decision-making in organizations, national security and medicine. However, alongside with giving an impetus for the economic development of the society, the BD has also posed several problems to scientific community and created new research paradigms [5-11].

In order to embody the potential of BD, several technical and technological problems must be resolved first. Several problems (computation, comprehensiveness, storage, incorrect correlations and etc.) emanating from the characteristic features (large volume, high velocity, variety) exist which require new scientific point of view, attitudes, modelling, mathematical methods, optimization tools and etc. These problems demand more effective statistical and intellectual analysis tools, architecture attitudes providing comprehensiveness, unified infrastructure and tools from researchers in order to obtain the needed information and knowledge. In this regard, the investigation of the most topical scientific and theoretical problems of the *big data* is of large importance.

Big data concept

In general, when the history of data development is reviewed, it can be seen that a technology used in data management analysis – "data base" management systems have emerged in the 70's of the past century [12]. It can be said that the concept of the "data base" was established ever since. However, mainframes (general-purpose universal electronic-computation machines) were not able to maintain the adequacy for the storage and processing as a result of the increase in data volume. In subsequent years, "parallel data base system" was proposed for the solution of the problem [13]. The architecture of these systems is based on the use of clusters (each processor in cluster consists of the processor, memory and the disc). It is worth mentioning that the "parallel data base system" was quite popular till the end of 90's of the past century. However, with the increase of the varieties of Internet services, the storage and processing of big data has also increased. Fundamental changes in computation architecture and expandable processing mechanisms are required in the solution of problem. Encountered the BD problems, one of the giants of information technologies (IT) sector *Google* corporation have generated *File System Google* [14] and *MapReduce* [15] software and hardware platform for the data management and data analysis at the Internet scale. Open code *Apache Hadoop* and *Hadoop Distributed File System* [16, 17] program software, also *NoSQL* data base was developed which have established the *big data* technologies. This technology is considered as the most accurate choice for the storage and management of large-scale data. It is because this technology is more effective in technology clustering issues, especially in the evaluation of the rating of web-pages. At the same time, it has also defeated the drawbacks of data warehousing systems, and enabled the collection of required information by using more complex analytical tools [11].

Several research works are available regarding the history and the review of the *big data* term [2, 3, 5–9, 18-20]. Investigations show that the big data is one of the terms with known history. Such that, the term of big data was firstly used by John Mashey, the expert on computer sciences of *Silicon Graphics* computer enterprise in 1998. This term is later encountered in 2000 – in a research work published in academic environment by Francis Diebold, a professor of the University of Pennsylvania and one of the leading researchers of the big data term [2]. His further work reckons that the term is firmly established, and has become a rigorous research direction rather than an event or phenomenon [18]. Notwithstanding, the term has gained popularity after an academic article published by Clifford Lynch, the professor of the University of Berkeley [21].

Regardless the fields applied, some general features pertain to the *big data*. These features can be divided into three main groups by reflecting the main problems of BD: volume, velocity and variety. This is also called "*3V*" in English-speaking sources. These features are broadly commented in several scientific sources [3, 5–9, 20, 22, 23]. The first model enabled to specify the BD and to distinguish it from other data, which was presented by Doung Laney, the analytics expert of *Gartner* enterprise in 2001 [24]. He has forecasted a tendency in electronic commerce: the higher importance and complexity of data management; thereafter, he specified the volume, transmission velocity and the variety of the data as main problems in data management. The considered features constitute the main concept of *big data* technologies in general. This concept reflects the idea of more efficient use, storage and extraction of more valuable information by the analysis of very large-volume data gathered at high velocity and from various sources.

This characteristics are commented in [3] as following:

Volume. Volume is the main feature of BD and the quantitative indicator of the data. At present, this indicator is measured by the volume of terabytes till zetabytes. The volume problem primarily causes a problem of storage which requires large-scale storage and distributed processing.

Velocity. Two cases are considered here. First, new data is generated at high velocity, the existing data are updated and collected. Second, as the volume increases, very high velocity is

required for the processing. The velocity is regarded as a problem of time and interpreted as the capability of existing processing technologies to analyze the data in real time.

Variety. Variety is one of the natural features of BD. The majority of information enters from different sources (e-mail, social networks, web-sites, sensors and etc.) at different formats, and different indexation scheme is applied. It is not an easy task to compile, jointly process and convert them into an appropriate format for the analysis.

Considering the veracity of the data and the value of BD, *IBM* and *Oracle* enterprises have added fourth "V" (*veracity*) and fifth "V" (*value*), respectively.

Veracity. By veracity, the quality of the data (complete, incomplete, conflicting and etc.) is understood. The quality of the data can change at large volumes which may affect the variety and the outcome of analysis.

Value. The data must possess the capability to generate the value. If BD does not generate value, it becomes "information dump". The constant attention to *big data* by business sector is due to its ability to generate value. Hence, this factor is evaluated as a marketing feature [1, 4]. It is because the value of information is determined by how we utilize it.

It must also be mentioned that the number of "v"s is constantly being increased by the experts in recent times.

The factors of formation of *Big data* as a research paradigm

Nowadays, the information abundance called big data is indeed present. However, it must not discourage people; on the contrary, it must be considered as a natural resource. Because, these natural resources possess the comprehensive knowledge, which can give an impetus for scientific discoveries. A new generation of analytical technologies is required for the value generation in the society and the business sector by using these resources at maximum. In this regard, the BD topic has attracted large attention of decision-making persons and politicians in government bodies, as well as the business sector and scientific researchers, and as mentioned, has become a new research direction.

In the first instance, it must be mentioned that various conferences, symposia, seminars and forums are held by notable international organizations, scientific institutions dedicated to different aspects of very large-volume information processing. The main discussion topics of those events are: Big Data architecture, Big Data management, Big Data modelling, Big Data analytics, Big Data toolkits, Big Data open platforms, Big Data as a Service, Big Data in Business Performance Management, Big Data Analytics in e-Government and Society, visualization, security, Big Data algorithms and etc. [3].

The dedication of special issues of scientific and mass journals to the topic of *big data* and the publication of new academic journals elucidating the scientific and theoretical problems of BD starting from 2014 is one of the main factors emphasizing the importance of this field as a rigorous scientific direction.

Fundamental scientific-research is being carried out in popular world scientific centers on the topic of big data collection and processing, storage, architecture, analytics, security, visualization and so forth. [3].

Starting from 2013, "data science" started to be taught as an academic module at bachelor, master and doctoral levels in several leading universities of the world: University of Dandee (Scotland), University of Auckland (New Zealand), University of South California, University of Washington, University of Berkeley, University of New-York, Imperial College London and etc. [6]. Such programs aim to provide a fundamental training such as the computation models for the maximum use of the potential of large-volume data, the mathematical methods for modelling and forecasting, architecture, contemporary programming methods, data collection, storage and analysis. At present, the number of these universities is constantly increasing. The program is being taught in Sabanci University in Turkey, Higher School of Economics in Russia and etc. at

master levels. It is worth mentioning that, the "data scientist" specialty is one of the promising popular specialties. Big data technologies and corresponding fundamental research have become the main scientific research directions of these education centers.

The bibliometric analysis in several leading science databases of the world is one of the main factors emphasizing the development of BD as a research topic [8-11, 25]. In this case, the topic development can be considered in terms of bibliometric indicators such as time, space, scientific fields, the number of published research works, type (book, article, conference proceeding and etc.) and etc. For instance, if only one research work could be found as a result of search on "big data" keyword in 2008, the number of scientific-research works has shown an exponential increase, starting from 2012. First places in geographical distribution of research works are occupied by USA, China, India, Great Britain, South Korea and etc. countries.

The research results carried out in *Google Scholar* and *Springer* databases are similar to [10, 25] as well. The dynamics of research works on years are given in the graph below (Figure 1).

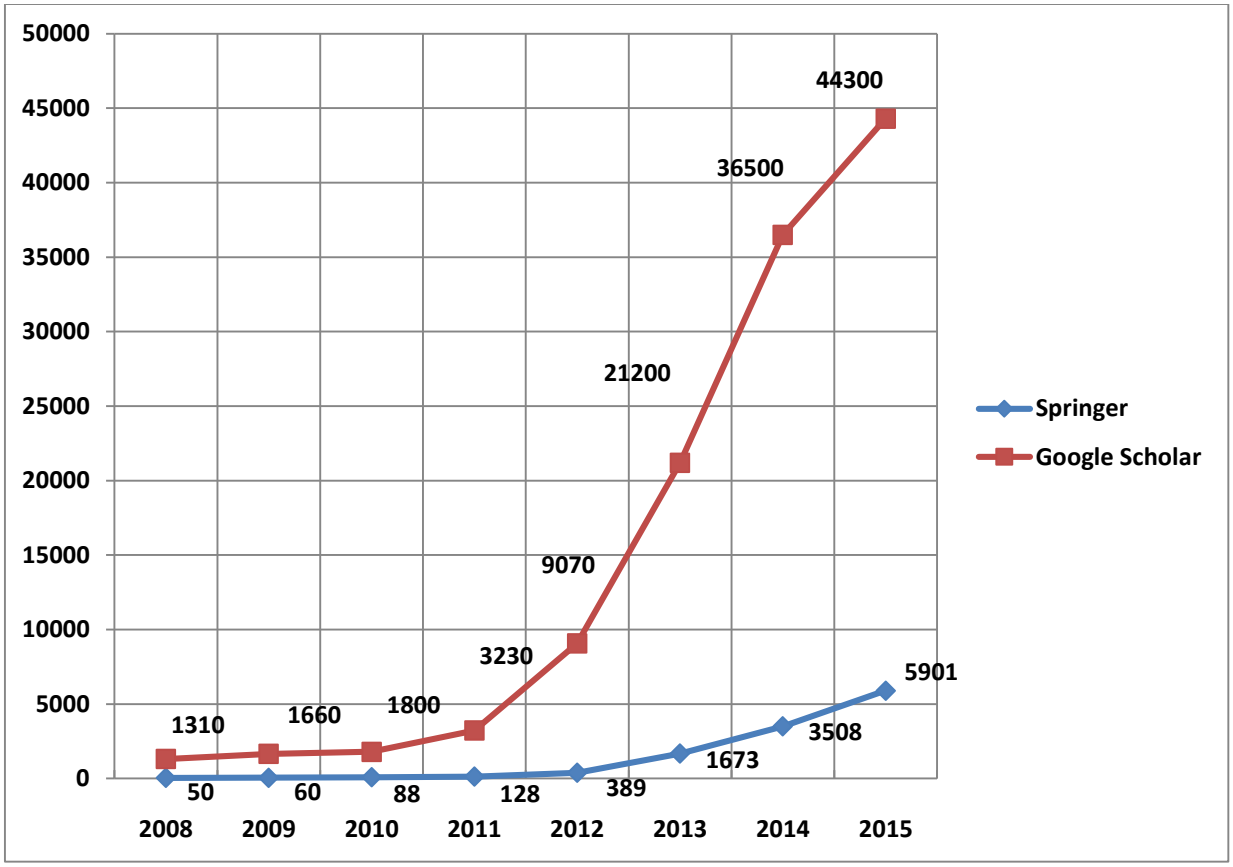

Figure 1. Distribution of research works on *big data* in *Google Scholar* and *Springer* databases

The distribution of works according to the document the type and field of science in *Springer* database are given in Figure 2 and Figure 3 respectively.

As seen from the Graph, the research works related to computer sciences, engineering, management, medicine, mathematics and social sciences occupy a large share in documents. Bibliometric indicators emphasize an increasing importance of the data in various scientific fields and the formation of the *big data* as a scientific direction once again.

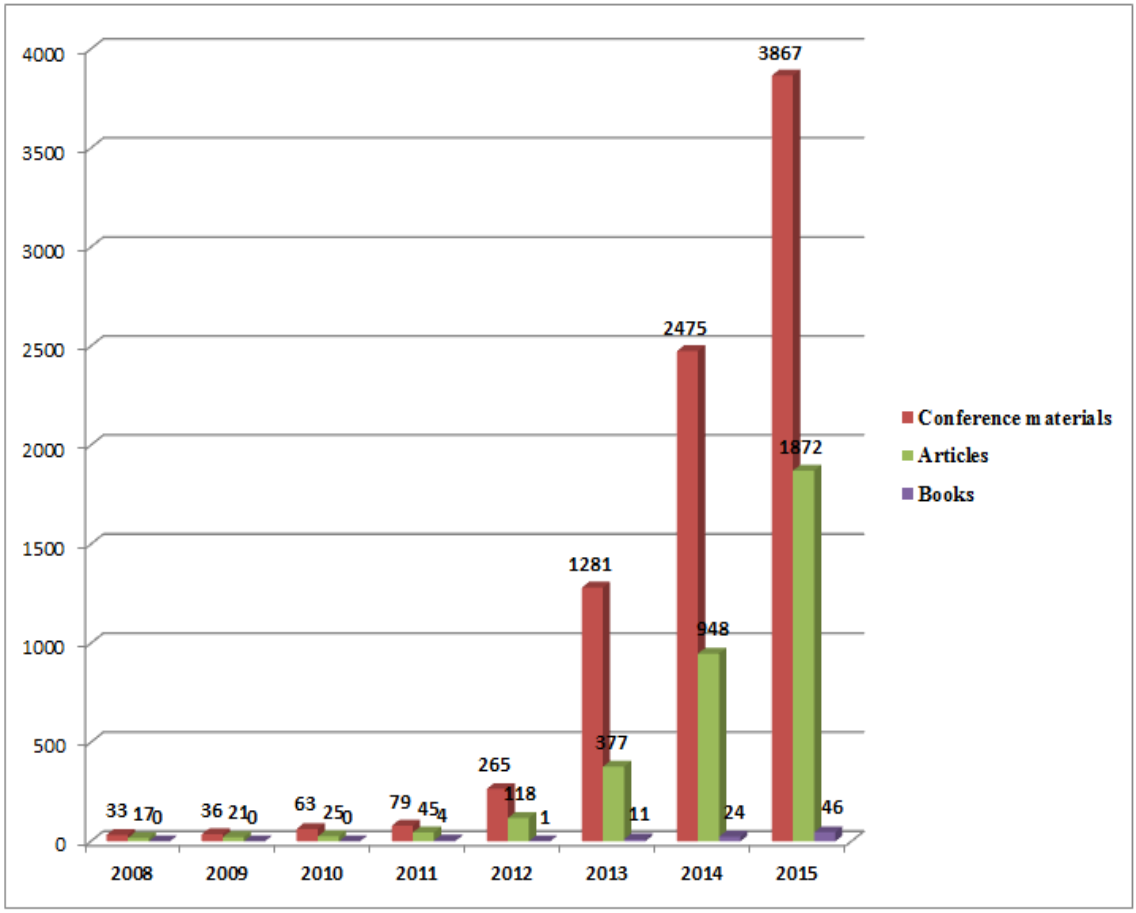

Figure 2. Distribution of research works according to the type of documents on *big data* in *Springer* database

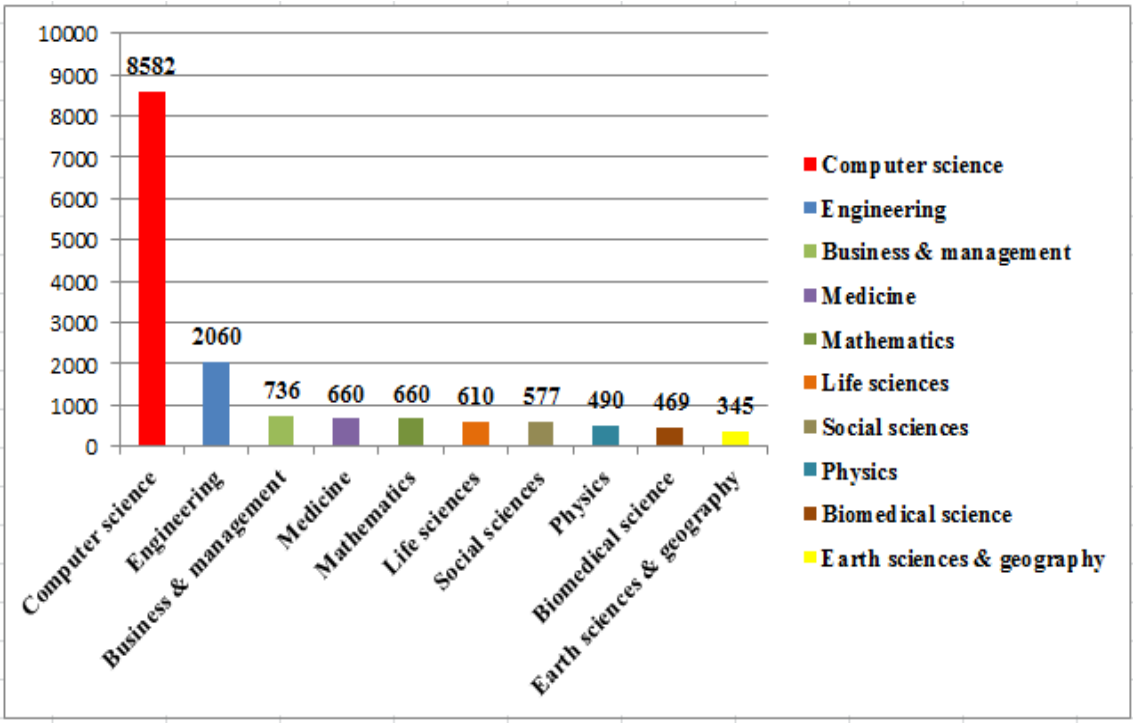

Figure 3. Distribution of research works on *big data* in *Springer* database

In general, big data has encouraged the revolutionary changes in the methodology of scientific research, scientific thoughts and methods. The extraction of new knowledge from data has established scientific investigations based on the intensive processing of this resource [6, 11]. The Turing Award winner Jim Gray has noted four scientific paradigms. Hence, the science was empirical thousand years ago, theoretical in hundreds centuries thereafter, and computational in recent decades; as the fourth paradigm, he has named the science based on the intensive processing of the data – the electronic science distinguished from computational science [26]. J. Gray reckons that the systematic solution of the majority of the complex global challenges, encountered by the humanity, can only be proposed by the fourth paradigm of the scientific research. According to this paradigm, the fields of science based on the intensive processing of the data has started to develop, whereas the scientific investigations become to be dependent on the data acting as a main source of the modern scientific discoveries [26]. The development of the fourth paradigm is stimulated by the large-volume data; various scientific fields are managed with the data at high level [9]. For example, fields of science such as astronomy [27], social evaluations [28], bioinformatics [29] computational biology, meteorology and etc. are based on the large-volume data processing. The interdisciplinary field named as "data science" gains its positions gradually [6, 30, 31]. The research topic of this field is the big data oriented to the extraction of generalized knowledge from the data. The data science is interlinked with several disciplines including informatics, mathematics, social sciences, network science, economics and etc. Hence, alongside with mathematics, statistics and informatics, the fundamentals of BD processing and analytics are constituted by various methods and theories, including the probability theory, image recognition, neural networks, intellectual analysis [32,33], machine learning [34], signal processing, natural language processing, forecast analysis, visualization [35,36], optimization, statistical methods [37], programming, engineering, the modelling of uncertainties, high productivity computations and etc.

Scientific and theoretical problems of *Big data*

Surely, the achievements of scientific research conducted in relation with the revelation of mysteries, surrounding the humanity, is a result of the efficient application of BD capabilities [9]. However, the revelation of knowledge unknown by BD and the decision-making, based on such knowledge, is a complex issue in terms of the data organization and processing – a new paradigm called "*big data computing*" [11]. This new paradigm comprises the methods and models of large-scale storage, processing and computation. That is, new scientific approaches and sophisticated analysis tools are required for the solution of the problems such as compilation and management, storage, security, search, analysis (generation of analytical reports and visualization, forecasting) and etc. of texts, images, audio, video, and other types of unstructured information with volume of hundreds of terabytes which carry exceptionally useful information and not processed by common relational databases [9].

In [38], the problems pertaining to BD are divided into three groups:

1. *Data issues* are related to the volume, velocity, variety and veracity. Thus, the data volume is increasing in comparison with past periods and existing tools are not capable to process them. Data may be in different format (text, sensor, audio, video, graph and etc.). Data is generated as constant flow and it is needed to extract useful information from them in real time. The incompatibility in databases may complicate the process of data processing and management. The quality of data may change and this change may impact the outcome of the analysis.
2. *Processing problem* includes the collection of valid information for analysis and search for solution compatibilities from various sources. Moreover, this includes the analysis and the presentation of access, that is, the visualization of the results in most appropriate manner comprehensible for a person.

3. *Data management* is the last of the proposed classifications. The management problem is mainly related to the preparation of data for analysis, storage, collection, the provision of data privacy and security, that is, the management of data lifecycle in general. It is the provision of correct use of data. This is regulated by information security policies and rules considered at national and international levels.

Problems of big data analysis. In general, the rapid increase in data volume and the demand for their analysis in real-time regime has led to the establishment of Big Data Analytics which is considered one of the main problems of BD. This is the process of detection of hidden regularities, unknown correlations and other useful information in large-volume data for the optimal decision-making. While *Big data Analytics* is applied to larger and more complex massives, *Discovery Analytics* and *Exploratory Analytics* terms are used alongside. Regardless of how it is named, the essence of analytics is not altered – to establish a feedback providing the information on different processes for decision-making people [3].

Three types of BD analytics are mentioned in [39]: big data descriptive analytics, big data predictive analytics, and big data prescriptive analytics.

Big data descriptive analytics is engaged in questions "what happened?", "why has it happened?".

Big Data predicative analytics answers the question "what will happen?".

Big Data prescriptive analytics not only shows "what will happen", "when will it happen", but also "why will it happen".

In general, BD analytics is one of the main problems of BD. This problem is related to the features, existing analysis models and methods and the limitations of data processing systems [9]. In [6, 40], the integration of various types of data, the volume, scaling, security, and incompatibility of data are indicated as main problems of analytics. Certainly, complex data analysis methods also exist. However, the majority of traditional methods of data analysis are not capable to dynamically adapt to different situations and to scale; those do not operate in parallel computation environment. The analysis of unstructured information such as text and audio, the collection of information from various sources and their integration is a serious problem. That is, the majority of existing methods are not sufficient for large and complex data [41]. It is known that an analysis allows to find the correlation among different parameters, features, events and etc., to classify, to prepare analytical reports and forecast in this basis. From this point of view, modern technologies must allow to convert the information into new knowledge or obtain the knowledge. The storage, processing and the analysis of BD require architectural attitudes and unified infrastructures providing the computation power and scalability. Such big data mining methods must be applied, which are able, in the first instance, to detect the changes in data. In short, there is a need for the conduction of practical and theoretical research for the creation of distributed version of existing models and methods or the development of new ones.

Main directions of BD analytics are associated with text, video, audio and social media analytics [8].

Text analytics – extracts the knowledge by detecting the previously unknown relations and correlations from natural language texts with the help of methods pertaining to *data mining* class [32, 33]. The classification and clustering, information extraction, summarization are the main problems solved by text mining. Essentially, text mining employs information search algorithms, as well as machine learning algorithms of linguistic and statistical methods for more comprehensive text analysis [8, 42].

Video analytics – comprises the monitoring and the analysis of video flow, and the extraction of useful information; it is a serious problem from the point of view of big data. Video-information is the main form of digital information and observations, and has a large volume. The analytics of this type of recordings is at the initial stage in comparison with other sorts of

data analysis. The analysis of video-information is one of the most complex issues posed to researchers. The main problem here is the information loss occurring as a result of loss in frame frequency and the precision of images [8].

Audio analytics – is the method of analysis of unstructured audio data and the extraction of useful information from those. These methods allow to improve the quality of services provided to customers, and to control the conduction of realization issues such as privacy and security. Its main problems are related to speech recognition, noise and etc. [8].

Social media analytics – carries out structured and unstructured data analysis of channels; it is used for forecasting the user behavior. The content in social media is usually large-volume, noisy and dynamic. Hence, the problems mentioned in text, audio and video analytics, as well as BD transmission also pertain to social media analytics [8, 40, 43, and 44].

Architecture problems of big data. At present, no broadly accepted architecture is present for BD analytics. The main functional components of big data architecture include data extraction, stream processing, information extraction, data quality/uncertainty management, data integration, data analysis, data distribution, data storage, metadata management, data lifecycle management and privacy. [19, 45] presents the review of existing 'etalon' architectures and platforms for large-volume data analysis such as *Hadoop*, *MapReduce*, *NoSQL* and etc. For now, it is not specified how the optimal architecture of analytical systems must look like for the simultaneous processing of retrospective and real time data.

Problems of big data processing and storage. *Package* provides the package processing*; Stream Computing* provides the analytical processing of the data regularly updated in real time regime and allows for forecasting, more rapid analysis and decision-making [9]; *Data-intensive computing*: is oriented to parallel computations in the processing of terabytes and petabytes of data. The necessity of data processing in petabytes has led to the emergence of *data–intensive computing* approach, which reckons that "not the computations, but the data is the greater wealth" [3, 9, 46].

Big data has radically altered the methods of data processing and storage, storage devices, storage architecture and data access mechanisms. The storage devices must be capable of providing the accessibility and operative analysis of large-volume data [9]. At present, *DAS* (*Direct Attach Storage*), *NAS* (*Network Attached Storage*), *SAN* (*Storage Area Networks*), *HSM* (*Hierarchical Storage Management*), *ILM* (*Information Life-cycle Management*) technologies are broadly used for the solution of the storage problem which are capable to carry out the transmission of information among devices [3, 6]. However, as for the level of large-scale distribution systems, the drawbacks and limitations of these storage architectures of data emerge. Recently, the application of *grid* and *cloud computing* technologies, which carry out the expansion of the memory capacity of devices, clustering and virtualization of computation and memory resources has almost eliminated the problems in storage [3, 9]. The cloud computing is one of the exceptionally successful approaches in storing and performing the big data computations. On the other hand, cloud storage creates a problem of data security in terms of the control over data completeness [9]. The evaluation and optimization of energy efficiency of BD processing and storage is of rigorous scientific-research importance [7, 8].

Problems of big data visualization. One of the main issues in BD analysis is the representation of results – visualization. Overall, the data visualization is one of the most simple and natural methods in data processing and analysis. This method allows a person to get familiar with presented information instantly and to make optimal decisions by correctly assessing the results [36, 35].

The visualization of large-size and large-volume data is not an easy task. Considering the features of BD, the process of visualization of such data encounters the following problems [35]:

- ***visual noise***. This problem occurs due to the excessive interconnectedness among the objects in dataset. By noise, the shrinking and loss of separate objects rather than the

deterioration and distortion of data is understood. This complicates the extraction of useful information from total view and additional processing is needed.

- ***large image reception***. Human brain is capable of receiving a visual view at some extent. Although the level of perception of graphic data visualization is higher than table visualization, some limitations exist. Hence, after a certain level, a human loses the capability to extract additional information from already loaded visual data. Surely, the methods of visualization are limited with the capabilities of technical devices providing the output view of data. No matter how advanced equipment we use, we encounter the limitations of human perception. That is, the methods of data visualization are not solely limited with the capabilities of devices, but also with physical perception of human. The data filtering – shrinking approach is used in order solve the mentioned problem.
- ***Information loss***. This is the problem emanating from the solution of perception of visual noise and large image. The attitudes applied in solution of mentioned problems reduce the data used at the end, but also lead to the emergence of the problem of information loss. It is because the methods of shrinkage of visual information carries out the aggregation and filtering of data according to one or several criteria based on the similarity of objects. These approaches may mislead analysts, and thus rather important and interesting issues may be ignored. Moreover, the process of data aggregation for obtaining the accurate and necessary information may require substantial time and computation resources.
- ***high performance requirements***. Graphical analysis is not only confined with the static visualization of images, but dynamic visualization is used as well; in this case, a subtle problem occurs in static visualization. There emerges a need for process performance at a certain velocity of visualization. It is due to the reason that, substantial time and computation resources are required for the filtering of large-volume data during the analysis process.
- ***high rate of image change***. As seen from subtitle, this problem is related to the high rate of change of images. That is, a person simply cannot react to the rapid change of data or their intensity on the screen during observation. The reduction in the rate of changing images is not able to provide the desirable effectiveness of the process. However, the speed of human reaction creates certain problems in this process.

It is necessary to develop new methods and technologies, as well as the qualified personnel for the elimination of such problems.

Security problems of big data. Overall, data security is considered to be an important issue at any time. New challenges have appeared for the information security with the emergence of BD. These problems may be approached in two aspects: 1) the application of big data analytics for information security; 2) information security in big data analytics [47, 48]. Both aspects are one of the topical problems posed to researchers. Such that, the application of big data technologies has demonstrated the obsoleteness and inadequacy of existing security models applied 15 years ago for today. The more digitalized the information and the more information added, it is more accessible and the number of users is higher. As a result, the incidents related to information security emerge such as the theft, distortion of information and network hack, the capture of personal information easily by malicious subjects.

The analysis of individual information without the consent of those individuals is unacceptable both ethically and legally; it is a serious problem in terms of security and privacy [49].

The features of BD, such as the variety and velocity, have exacerbated the security and privacy problems. First, the size of big data is excessively large in relation to existing security approaches. This, in turn, complicates the measures in security field. On the other hand, big data is stored in distributed form and network threats may increase such threats [9]. The large-scale "cloud" infrastructure, the variety of data sources, collection of stream information and the

migration of large-volume information into "clouds" have also revealed caveats in security systems. Therefore, traditional security mechanisms are not sufficient in situations of BD expansion. At the same time, data stream requires quite flexible and rapid security solutions [9, 47-49]. The problems constituting the main directions of scientific research, such as the detection of cyber-attacks; information systems security, management of security risks, information risk assessment and so on, are one of the important problems challenging the researchers in the solution of problems similar to BD [47].

Conclusion

As a valuable source of knowledge, *big data* has become the most debated topic and new multidisciplinary scientific research direction in the field of information technologies by attracting the state, business and scientific communities throughout the world. The main directions of scientific research works are constituted by science-intensive problems emanating from natural features, such as volume, velocity and the variety; the latter form the basis for the big data concept. The problem groups include various issues starting with the data collection and ending with the presentation of results to user. For this reason, the problems must reach their scientific solution related to the processing and management, storage, security, search, analysis and etc. of the text, images, audio-video and other type of the unstructured information in terabytes and exabytes. The development of methods such as data mining-class methods which are more effective in problem solution (associative rules, regression, classification, clustering and etc.), artificial neural networks, machine learning, optimization, as well as genetic algorithms, image recognition, predictive analytics, imitation modelling, statistical analysis, and the visualization of analytical data, are one of the main issues posed to researchers.

References

1. The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, November 2014, http://www.emc.com/leadership/digital-universe/index.htm
2. Diebold F. Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting / Discussion Read to the Eighth World Congress of the Econometric Society, Cambridge: Cambridge University Press, 2000, pp. 115-122.
3. Aliguliyev R.M., Hajirahimova M.S. "Big data" phenomenon: problems and prospects // Information Technologies problems, 2014. №2, pp.3-16.
4. Big data: The next frontier for innovation, competition, and productivity. Analyst report, McKinsey Global Institute, May 2011, http://www.mckinsey.com
5. Fan J., Han F. & Liu H. Challenges of Big Data analysis // National Science Review, 2014, vol. 1, no. 2, pp. 293–314.
6. Chen P.L., Zhang C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data // Information Sciences, 2014, vol. 275, pp.314–347.
7. Chen M, Mao S, Liu Y. Big data survey // Mobile Networks and Applications, 2014, vol.19, no.2, pp.171–209.
8. Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics //International Journal of Information Management, 2015, vol. 35, pp. 137–144.
9. Jina X., Benjamin W. Waha, Chenga X., Wanga Y. Significance and Challenges of Big Data Research, 2015, vol.2, no.2, pp. 59–64.
10. Halevi G. The Evolution of Big Data as a Research and Scientific Topic // Research Trends, 2012, no.30, pp.3-6.
11. Raghavendra K. et al. The anatomy of big data computing // Software: practice and experience, 2016, no. 46, pp.79–105.

12. Codd E. F. A Relational Model of Data for Large Shared Databanks // Communication ACM, 1970, vol.13, no.6, pp. 377-387.
13. DeWitt D., Gray J. Parallel database systems: the future of high performance database systems // Communication ACM, 1992, vol. 35, no.6, pp. 85–98.
14. Ghemawat S., Gobioff H. and Leung S.T. The Google File System / Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, New York, USA, October 2003, pp. 29–43.
15. Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters / Proceedings of the Sixth Symposium on Operating System Design and Implementation, volume 6 of OSDI '04, Berkeley, CA, USA, 2004, pp.137–150.
16. Hadoop MapReduce, http://www.hadoop.apache.org/docs/stable/mapred_tutorial.html
17. Hadoop Distributed File System, http://www.hadoop.apache.org/docs
18. Diebold F. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012, http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_ Big_Data.pdf
19. Maier M. Towards a Big Data Reference Architecture, 2013, http://www.win.tue.nl/~gfletche/Maier_MSc_thesis.pdf
20. Imamverdiyev Y.N. Broad perspectives and problems of big data technologies // Information society problems, 2016, №1, pp. 23-34.
21. Clifford L. Big data: How do your data grow? // Nature, 2008, vol.455, pp. 28–29.
22. Kaisler S. et al. Money W. Big Data: issues and challenges moving forward / Proceedings of the 46th Hawaii International Conference on System Sciences, 2013, pp. 995–1004.
23. Tole A.A, et.all. Big Data challenges // Database Systems Journal, 2013, vol. 4, no. 3, pp. 31–40.
24. Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety. Technical report, META Group, Inc (now Gartner, Inc.), February 2001, http://www.blogs.gartner.com/doug-laney/files/2012/01
25. Alguliyev R.M., Ismayilova N.T. Bibliometric Analysis of Big Data Research / "Big data: capabilities, multidisciplinary problems and perspectives" First Republican scientific-practical conference proceedings, Baku, 25 February, 2016. Pp. 58-60.
26. Hey T., Tansley S., Tolle K. (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corporation, 2009, 287 p.
27. Zhang Y., Zhao Y. Astronomy in the Big Data Era // Data Science Journal, 2015, vol. 14, no.11, pp.1-9.
28. Hays R. R., Daker-W. G. The care.data consensus? A qualitative analysis of opinions expressed on Twitter // BMC Public Health, 2015, vol. 15, no. 838, pp. 2-13.
29. Greene C.S., Tan J., Ung M., Moore J.H., and Cheng C. Big data bioinformatics // Journal of Cellular Physiology, 2014, vol.229, no.12, pp.1896–1900.
30. Wu Z. From Big Data to Data Science: A Multi-disciplinary Perspective // Big Data Research, 2014, vol. 1, p.1.
31. Jagadish H.V. Big Data and Science: Myths and Reality // Big Data Research, 2015, vol. 2, no 2, pp. 49–52.
32. Wu X., Zhu X., Wu G.Q., Ding W. Data mining with bigdata // IEEE Transactionson Knowledge and Data Engineering, 2014, vol.26, no. 1, pp. 97–107.
33. Alguliev R., Aliguliyev R., Hajirahimova M. Multi-document summarization model based on integer linear programming // Intelligent Control and Automation, 2010, vol.1, no.1, pp.105-111.
34. Omar Y. Al-J. et al. Efficient Machine Learning for Big Data: A Review // Big Data Research, 2015, vol. 2, no. 3, pp. 87–93.
35. Gorodov E., Gubarev V. Analytical Rewiew of Data Visalization Methods in Application to Big Data // Journal of Electrical and Computer Engineering, 2013, 1-7 p.

36. Olshannikova E. et.all.Visualizing Big Data with augmented and virtual reality: challenges and research agenda//Journal of Big Data, 2015, vol. 2, pp.2-22.
37. Hajirahimova M.Sh., Aliyeva A.S., Review of statistical analysis methods of high-dimensional data // Eastern-European Journal of Enterprise Technologies, Kharkov, 2015, no 5, pp. 23-30.
38. Akerkar R. Big Data computing. Boca Raton, FL: CRC Press, Taylor&Francis Group, 2013, 562 p.
39. Sun Z., Pambel F., Wang F. Incorporating big data analytics into enterprise information systems, Lecture Notes in Computer Science, 2015, vol. 9357, pp. 300-309.
40. Kambatla K., Kollias G., Kumar V., Grama A. Trends big data analytics// Parallel and Distributed Computing, 2014, vol.74, no.7, pp. 2561-2573.
41. Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, Big data analytics: a survey // Journal of Big Data, 2015, 2(21), 1-32.
42. Jiang J. Information extraction from text. In C. C. Aggarwal, & C. Zhai (Eds.), Mining text data (pp. 11–41). United States: Springer, 2012.
43. Kalampokis E., Tambouris E. and Tarabanis, K. Understanding the Predictive Power of Social Media // Internet Research, 2013, vol. 23, no. 5, pp. 544–559.
44. Barbier, G., & Liu, H. Data mining in social media. In C. C. Aggarwal (Ed.), Social network data analytics, United States: Springer, 2011. pp. 327–352.
45. Pekka P., Pakkala D. Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems // Big Data Research, 2015, vol. 2, no 4, pp.166-186.
46. Klemenkov P.A., Kuznetsov S.D. Big data: contemporary approaches to storage and processing // The Institute of System programming materials RAS, vol. 23, 2012, pp. 143-156.
47. Alguliyev R., Imamverdiyev Y. Big Data: Big promises for information security / Proceedings of the IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, Kazakhstan, 15-17 October, 2014, pp. 216–219.
48. Hajirahimova M. "Big data" technologies and the problems of information security // Information technologies problems, 2016, №1, pp. 49-56.
49. Lei X., Chunxiao J., Jian W., Jian Y., Yong R., Information Security in Big Data: Privacy and Data Mining // IEEE Access, 2014, vol.2, pp.1149–1176.