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

DOI: 10.25045/jpit.v08.i2.07

ANALYZING BIG VOLUME MEDICAL DATA: CURRENT PROBLEMS AND PROSPECTS

Big Data is becoming a promising area in the field of healthcare. It is used to improve the results of the analysis of large volumes of data sets and to reduce the costs. Increase in data volume and demand for ad hoc analysis of data created one of the biggest problems of Big Data called Big Data analysis. The article deals with actual problems of a large analysis of data generated in the field of health, and explores the main characteristics of the data. At the same time, it defines the various opportunities, advantages and characteristics of the data in the area of health, and provides a number of recommendations.

Keyword: *data warehouse, big data, big data analytics, biometric data, evidence-based medicine, genomic analytics, MapReduce, Hadoop.*

Introduction

Today, we are living in the next development era of the Internet technologies. The devices are being developed and with the help of transmitters they collect information about the people and the world that surrounds us. Gartner predicts that, by 2020, 30 billion devices will be connected to the Internet, most of which will besports bracelets, refrigerators with the Internet access and other things, rather than the communication devices such as smartphones and computers. These devices will collect data and share with people to take care of them [1].

In this regard, the idea of *Big Data (BD)* has recently been concerned everyone. At present, the term Big Data is used in all fields working with large-scale data, such as medicine, economics, banking, manufacturing, marketing, telecommunications, web analytics, and so forth. In the upcoming years, Big Data revolution is expected to happen in medicine, criminalistics, urban infrastructure and other spheres. Thus, the analysis of the current situation in the health field shows that active research is being carried out in the field of portable devices, which are expected to become massive in the future. These portable devices will include not only watches, glasses, lenses, and various body plastics, but also electronic mini-devices under the skin. Experts note that the experiments in this field are very slow due to the lack of volunteers and the high cost of the experiments. Therefore, unfortunately, the number of people with chronic and other illnesses is gradually increasing [2]. For example, such internal transmitters can solve diabetes problems. Thus, people can monitor their condition in real-time mode and control the level of glucose in blood through the existing application.

The introduction and development of such transmitters generates a large volume of data streams not only for operational processing but also for subsequent analysis. In addition, there are a number of BD problems that need to be solved. The first problem is that the personal data and.in general, all information from customers should be provided to the companies for processing. The point is that the information, in this case, is provided free of charge, analyzed, processed by a computer; a new product is generated and sold back to people. The legal regulations should be introduced to use the personal data in this way.

In the healthcare, it is rather more complicated. For example, not all people are ready to give their genomes in the examination. Because, people are unaware of what it is and what outcome may be. At the same time, there is a medical ethics and physician secret that urge the physicians to keep their patients' illness secret. Thus, companies should collect and store the statistical data that serves for the interests of specific customers. The development of information methods for the analysis of all patients' data may cause unsatisfactory results. It should be noted that the specification of portable devices involves a large number of unstructured information flows, which facilitates its processing.

The companies may realize certain implementations to improve the condition of the patient by using this information. One of the pressing problems is to obtain knowledge and to present the results by analyzing large volume of data collected in the healthcare field. Given this crucial fact, the research and study of existing problems BD analytics in health is essential.

"Big data" sources in healthcare

Healthcare has been historically generating the BD volume. At present, most of the data is stored in paper; there is a tendency to digitize large volumes of data. In addition to the costs reduction, BD offers a wide range of health services, including clinical decision support, disease monitoring and health management of the population for improving the quality of health care [3, 4].

In the field of health care, BD is characterized by the different types of data and their velocity, rather than their volume. BD includes written records and prescriptions of a physician, medical visualization, laboratory, pharmacy, insurance and other administration data, data on patient’s electronic health records, the data transmitted by the donors, the data obtained from the social media, including tweets, blogs, Facebook and other platforms, including the data about the immediate medical aid for certain patients, the latest news and articles in medical journals. The data on the health and well-being of the patient constitutes the BD in the field of health [5, 6].

Healthcare is one of the areas in which large amounts of various data with high velocity, such as figures (analysis data, health records), video (ultrasound examination), photography (tomography, radiography), electro-signals (electroencephalogram, electrocardiography) and others are collected. It is very difficult to simply combine, process and analyze them. It is difficult, even impossible, to process electronic health data with traditional software. The main problem is that the most of the information comes from different sources in different formats and uses different indexing schemes. They are not easily managed by the traditional data management tools and methods [7].

As noted, BD in healthcare comes from various sources and in various formats at high pace. Internal sources include electronic health records, clinical decision support, while external sources are geographically distributed government sources, laboratories, pharmacies, insurance companies, and so forth. BD sources and types in health are classified in the following way [8, 9]:

- data generated by *machine-to-machine (M2M)*, including remote transmitters, meter readings, and other vital devices;
- *biometric data*, e.g., fingerprints, genetics, scanned retina, x-rays, other health descriptions, blood pressure, pulse indexes, and other similar data;
- *data generated by humans*, e.g., physician’s records, e-mails and paper documents, and structured and partially structured data;
- *BD agreements*, e.g., health requirements, other accessible partially structured and unstructured account records;
- *web and social media data* refer to the interactions found on Facebook, Twitter, LinkedIn, blogs and other similar projects (including health service schedules for websites, smartphone apps).

It should be noted that the data collected in private databases and warehouses are now becoming commercial. Emerging Big Data market can be divided into several sectors:

- Using Big Data to enhance personal business analytics and effectiveness;
- Selling Big Data to external agents;
- Collecting Big Data by the agents to resell them to different agents;
- Presenting analytical services to the clients providing access to their Big Data;

The informatization of healthcare is related to the use of electronic health records of patients, information management system of the clinics and the introduction of Internet services. Studies show that more than 15% of US doctors has completely shifted into electronic records in clinics and do not fill out paper documents. Note that the USA in the leading country for the development and implementation of information technology (IT). Remaining part partially uses blank paper and cards.

Clinical management of such electronic health through information systems generates large volumes of data flows. In order to be able to analyze this BD flows, it is necessary to register all possible parameters, to completely digitize all the records, and to gather the information from a variety of clinics into a single database. The process of digitization of all the data into carriers is called *datafication* [10].

Problems of BD analysis in the field of health

The conceptual framework for the BD analysis in health care is as in the traditional health informatics. The main difference is how to handle the process. A typical health analytics can be accomplished through business-analytics tools built on independent systems such as personal computers or laptops. Here, large volumes are distributed and processed on several intersections according to their designation. The concept of distributed processing has been available for many decades. However, the novelty is its use in the analysis of large volumes of data sets. Thus, today, health professionals use the BD warehouses for health decisions. Data processing stages in data warehouse (DW) include data *collection, cleaning, download, analysis,* and finally, presentation of the results of the analysis. Each of these stages performs special operations on the data [11]. It should be noted that, if DW technologies were applied for BD analysis, it would be necessary to consider not only algorithms analysis, but also all the stages of the work with data.

Collecting, managing and storing hundreds of terabytes of health BD and extracting useful knowledge from them with available methodologies or tools is a serious problem. Working with structured and unstructured information, deeper data mining and visualization of the analysis results are the key issues of BD analytics. The increase in the volume of data and the need for their real-time analysis has led to the emergence of Big Data Analytics - one of the main problems of BD. To bring unstructured data into tables or graphs, the specific IT-tools for Big Data analytics (e.g. R programming language or Hadoop etc.) should be applied [12].

Big Data problems are mainly related to real-time processing, searching, classifying and analyzing the large-scale and hastily generating information. Regardless of the application areas, there are general characteristics to describe BD, which are divided into three basic categories: volume, velocity and variety. In English-language sources, it is mentioned as "3V". The convergence of these parameters helps to identify BD and to distinguish it from other data [13, 14].

BD analytics in health is described by three main characteristics. Health-related data are generated for a long time and accumulated uninterruptedly generating BD. The volume of existing health care BD now includes private health records, radiology images, transmitters of clinical trials, human genetic and genome sequence data of the population, and so forth. New BD forms, such as3D visualization, genome and biometric indicators of transmitters also help the exponential growth. Today, achievements in the field of data management, especially virtualization and cloud computing, contribute to the development of the platform for more efficient collection, storage and processing of BD. The data is collected in real time at high velocity. These data must be merged for BD analysis. The uninterrupted flow of new data creates new problems. Although the variety of collected and stored data varies, they are vital for analysis, comparison and decision-making. Recently, the regular monitoring of large amounts of health data has been carried out (e.g. daily diabetes measurement of glucose, measurement of arterial pressure and electrocardiogram diagnosis.). As a result, in many health cases, the real-time permanent data (e.g. monitors in the operating room for anesthesia, monitors on a heart disease bed) save lives. The ability to perform BD analytics on all aspects of healthcare in real time is one of the most urgent issues [15].

Analytical methods of volume, velocity and variety have been developed due to the developing nature of the health data. Only the data collected on electronic health records and other structured data are not analyzed. The majority of Big Data sources includes unstructured and semi-structured data. Working with structured, partially structured and unstructured data makes health data attractive and complicated. Structured data is easy to be stored, analyzed, and manipulated through the machine. The healthcare is one of the areas in which structured and semi-structured data contains the result of the devices and the data received as a result of the digitization of the paper documents. Unstructured data is also generated in healthcare, which are referred to the health files recorded by a nurse and physician, paper recipes, tomography, radiographs, ultrasound examinations, and other images.

Nowadays, new flows of structured and unstructured data in the field of healthcare come from fitness facilities, genetics and genomics, social research, media and other sources. However, a small part of these data can be collected and stored now. The data must be organized to be manipulated with the help of computers and analyzed the useful information in accordance with its content. Health applications need more efficient ways of combining and converting different data, including automated data conversion from unstructured into structured datasets. Potential of BD in the field of health care is to comply the traditional data with new formats both at the individual and population levels. For example, if the manufacturers of pharmaceutical products could integrate the population's clinical data collection with genetic data, producers could then achieve faster access to the best quality of drug treatment compared to the previous years.

Researchers also mention the fourth characteristic of health. Thus, the *veracity* of the data gathered in healthcare is also important. That is, BD, analytics and conclusions are the credible and trustworthy results. How accurate or suspicious decision is made? Of course, the goal is the truth, yet not a reality. For example, data obtained through transmitters is more reliable than that obtained from social media. Veracity encompasses both the scalability and performance of the platform, algorithm, methodology and tools in order to meet BD requirements. Structured and unstructured BD architecture, analytics and tools differ from traditional *Business intelligence (BI)*very much. For example, BD analysis in healthcare is carried out using a parallel computing paradigm in the form of distributed processing of the data on several servers (“nodes”). At the same time, the realization of the data mining models and methods, including statistical methods, visualization methods and algorithms also takes into account BD characteristics. The veracity of the data presented in healthcare has the same problems as in the financial data (especially, from the payer side): the number of the patients, the payer, the compensation code, the validity of the currency, and other veracity settings are also important for healthcare. These include diagnostics, treatments, recipes, procedures, accuracy of results, and so forth.

Data quality issues cause serious concern in healthcare for two reasons. First, the health of people depends on the availability of accurate information, and second, the quality of health data, especially unstructured data, is constantly changing often being wrong (inaccurate translations, poor handwriting in recipes are the most typical examples) [16]. Improving the coordination of medical aid by avoiding mistakes and reducing costs depends on high quality data, achievements in the field of medicines, efficacy and accuracy of the diagnosis. However, the increase in variety and velocity prevents the self-cleaning ability in addition to data analysis and decision-making. The cost of hardware and software is decreasing, therefore, these issues should be solved to maximize the use of the potential of BD analytics in health care.

Therefore, there is always a need for a new tool to deal with the issues such as the collection and storage, search, security, and analysis of various types of unstructured data. However, the number of architectures and platforms, including the prevalence of open source codes in the available tools, and other issues should be taken into account. Analysis of large volumes of health data, disclosure of confidentiality and associations is crucial for making important decisions, improving people's welfare, rescuing their lives, and solving other issues. Hence, more and more grounded decisions are made in the field of health through the use of BD. Experts specify the following potentials of BD analytics in healthcare [17, 18]:

- analysis of patient’s characteristics to determine the clinical and economically efficient treatment methods (influencing the behavior of the consumer by offering analysis and tools based on the cost and outcome of the treatment);

- use of advanced analytical tools (e.g. segmentation and forecasting modeling) to detect the patients, whose lifestyle has changed;
- large-scale prophylaxis of the disease to detect prognostic measures for prophylaxis support;
- collection and publication of information on medical procedures (assisting the patients during the examination and when the treatment schedule is specified);
- detection, prediction and minimization of the specifics through the use of advanced analytical systems;
- real-time claims solution and so forth.

The volume of data in the field of health is expected to grow sharply in recent years. In addition, the model of health care costs changes. Constructive usage and rewards appear to be important factors in modern health care. Health organization may increase its income, optimize costs, improve and increase financial performance by acquiring existing tools, infrastructure and effective use of BD tools.

Opportunities of BD analytics in healthcare

At present, digitalization, combination and effective use of BD by healthcare organizations provides many advantages. Potential advantages include the opportunity to control the health of a particular individual and of the population, to timely detect diseases for easier and more effective treatment, and to quickly and efficiently detect speculation cases in healthcare. Today, there are many complicated issues that can be solved by BD analysis. Experts refer them to the followings [19, 20]:

- predicting the events or outcomes by assessing large volumes of archive data (e.g. patient’s stay in a health institution);
- identifying the patients who prefers surgery;
- identifying the patients who do not benefit from surgical intervention;
- identifying the patients with complications or risk of medical complications or with the risks of other uncertain diseases;
- identifying the possible pathological conditions and other factors that contribute to the development of the disease and development of risky patients, and so forth.

It is also noted that BD can help to reduce the costs and increase effectiveness in the following three areas:

- ***In clinical operations***, during a comparative study of efficacy to determine economically more effective methods for treating and diagnosing the patients;
- ***In research and projecting***, which imply a forecast model to reduce production costs for research-oriented devices. The use of statistical tools and algorithms to improve clinical trials, including the collection of special patients for good treatment, promotes the rapid release of new treatment methods into the market. As a result, the analysis of the patients and clinical trials is required to define these findings and to detect negative results before the product is launched in the market;
- ***In public health***, which includes the analysis of disease models and the control of the spread of disease to improve public health, the rapid development of precocious vaccines, the conversion of large volumes of data into useful information, and others. This information can be used to identify the needs, provide services, predict and prevent crisis situations, especially in the interest of the public. Additionally, the experts point out that BD analytics in healthcare can contribute to the following areas [21]:
- ***Evidence-based health*** aims to analyze various structured and unstructured, financial, operative, clinical and genome data, to comply the treatment with the results, to predefine the patients with risk of disease and provide more effective service. ***Evidence-based health*** is the concept of organizing medical knowledge based on fundamental scientific

data. In this case, the personal experience and the authority of partners are of secondary significance. The key principle for the adoption of health and managerial decisions is the justification of an objective fact. This section of the medical science is called differently in the literature: *fact-based medicine (FBM), evidence-based decision making (EBDM), fact-based decision making (FBDM), evidence-based healthcare (EBHC)*. The main idea of applying evidence-based medicine to medical practice is to minimize the human factor in the physician's activities [22];

- ***Genome analytics*** is an efficient and cost-effective implementation of gene sequencing and making the genome analysis a part of a routine process of medical aid;
- ***Speculation analysis*** of numerous claims to reduce the wasting and abuse;
- ***Remote monitoring*** is a real-time monitoring of security of continuously changing BD, and at the same time, its collection and analysis for negative predictions of events;
- ***Patient's profile analysis*** intended for the use of advanced analytical tools for the patients' profile (e.g. segmentation and forecasting modeling). That is, effective detection of life-style changing patients, and prophylaxis assisting for the patients with specific risk of developing disease (e.g. diabetes).

Furthermore, the following advantages of BD analytics for health care are highlighted [23]:

- Precisely, determining the patients with acute risk;
- providing reasonable decisions to patients and more effective health management, and ensuring the necessary information for a healthier lifestyle;
- identifying the unsuccessful and expensive procedures, programs and processes;
- reducing the recurrent hospitalization and adjusted treatment plans by detecting environmental and lifestyle factors that increase the risk or create additional effects;
- improving the results by studying vital organs through the monitors installed at home;
- managing the health of the population through detecting the vulnerabilities during the spread of diseases or natural disasters (resulting in a unification of the clinical, financial and operational data, which enables the productive and real-time use of resources) and so forth;

It is known that health decisions require the collection and storage of the patient's personal data. Such BD base has already been created in developed countries (e.g., US). Only professional doctors, health staff, pharmacists, employees of insurance companies, judicial authorities and analysts can access this data. It should be noted that BD should be properly stored in the warehouse thus it can be used for analytics and statistics. At the same time, their third-party ownership should be coded. When comparing the changes in the medical field the experts refer to the Data-driven Medicine. At the same time, this medicine has led to the emergence of the term "Big Data ecosystem". It should be taken into consideration that the Big Data ecosystem is interacting with the society and affects it. The Big Data ecosystem is expanding and receiving the data about the citizens from different fields. This information may include all credit card transactions, log files collected on a personal computer, mobile phone location data, pension fund or insurance company data and many other databases.

Research shows that the BD flows are now attempted to be connected, which brings real benefits and savings to separate sectors of the economy. Now, scientific studies focuses on the scientific experiments rather than the theoretical considerations. Each citizen will get an identification number for the future coding and decoding of their data in various BD bases. The expanding BD ecosystem requires the consolidation of a professional knowledge, since the data storage formats must provide an interoperable environment. Different data from different bases coded by unified method must be suitable for unique analysis.

Experts believe that the recent studies on the application of the Big Data ideology may include the followings [24, 25]:

- attempts to convert all the records to electronic carriers refusing paper carriers;

- optimization of the treatment process through the analysis of BD collected in health data systems;
- provision of data on the population assigned to polyclinics to the pharmaceutical manufacturers for the provision of local markets with the necessary medicines and others;

The rights of patients are related not only to the use of Big Data to improve the quality of life and treatment, but also to the confidentiality and legal regulation of Big Data circulation on the market. Moreover, every citizen must enjoy the right to access, copy and paste his/her personal data free of charge. At present, all the states are developing legal regulations governing the mutual relations of the organizations and the individuals of all ranks, i.e., the founder, owner and consumer of information. The key advantage of Big Data is providing both retrospective and real-time data predictions. In this regard, the relevant central authorities of the country, heads of health institutions, health data systems (HDS) and analytics should collaborate. Experts refer them to follows:

- the use of HDS that expands the composition of the initial records of the physician and modifies their storage structure;
- development of the methods (technologies) that identify the patient’s current status and include them into the databases of health organizations;
- Establishing, enhancing and continuously increasing the computational capacities of regional data processing centers (DPC) and the effectiveness of tools for processing and summarizing the initial records of HDS;
- identification of the development issues of Big Data methods in health care as a priority;

Hadoop technology for BD processing and analysis in health

Experts include analytical systems, data cloud services, MapReduce technologies and NoSQL distributed databases management systems (DBMS) in *Massively Parallel Processing (MPP)* platforms to the Big Data Category. Apache Software Foundation project - Hadoop is more widespread technology, a key platform for the processing and analysis of BD (of Petabytes scale) in a distributed computing environment. It is the open access MapReduce model. Hadoop consists of two main components: Hadoop MapReduce and Hadoop Distributed File System (HDFS). MapReduce is based on parallel computations, while the HDFS is based on data management [27-29].

At present, open access Hadoop for Big Data processing is available to every user free-of-charge. The open access platforms, such as Hadoop, MapReduce have also created favorable conditions for the application of BD analytics in the field of healthcare. Traditional health analytics tools are very comfortable and transparent. While the algorithms and models of traditional analytics are similar, user interfaces of BD completely differ. But on the other hand, BD analysis tools and programming are very complex and require a variety of habits. Hadoop can play a dual organizational role of data and analytics tools. It offers great potential and enables enterprises to use the data that was difficult to manage and analyze so far. Hadoopalso enables the processing of various structured and generally unstructured BD. However, Hadoop system can be difficult to built, set and managed. Moreover, it is not easy to find the users capable to work with this system [30].

At the same time, these platforms are open access systems, hence their usage is not comfortable. Perhaps, for these reasons, organizations are not ready for the full implementation of Hadoop. The additional platform and the tools’ ecosystem support the Hadoop distributed platform. The disadvantage is the lack of technical support and minimal security. In the field of health, this is, of course, a major deficiency. Furthermore, as noted above, these platforms and tools require the end users to have great and distinctive programming skills in the field of health. At the same time, many issues, including management, ownership rights, confidentiality, security and standards should be solved taking into account the fact that BD is a novelty in the field of health. The advanced methods and existing analytical technologies may give better treatment consequences being applied in a large number of available health records. These results can also be applied in the future to preserve the health of the population. Private and group data can help each physician to appoint a more suitable treatment option for a particular patient and use them in the decision-making process.

From bioinformatics to personalized health

The genome, which is a separate field of molecular genetics, is rapidly developing. Today, many companies are analyzing the methods for the genome mining. They believe that, over the next five years, the personalized medicine will become common and widely spread. All these achievements are possible due to the development of BD processing methods. Big Data technology is a new generation of technology and architecture in the study of the automated processing methods of various types of health data. The use of BD allows obtaining knowledge and delivering the results due to the prompt analysis at a short time. The specialized algorithms used in Big Data technologies reduce the computing burden by ten times, and the new web platforms provide access to the petabytes of genetic data.

Nowadays, there are many approaches to the development of personalized medicine, but ultimately, they allow physicians to advise the treatment of diseases based on the genetic data of the patients. The main problem in the personalized medicine is the development of software to compare genetic data of as many people as possible for scientific and practical purposes. Collecting the data of the patients who have undergone routine medical examination is essential for the realization of the software. There may be both individuals with various diseases and healthy volunteers among them. Comparative analysis of millions of records can help detect the details of pathological processes and the role of genetic disorders in their formation. Correlation models can also be used to analyze various genomic records and traditional data on the condition of their patients' health. Experts believe that it is possible to predict the behavior of each individual in different situations by adding genetic data to his/her characteristic and provide more effective medical aid during the illness.

Evidently, this method is also useful for addressing public health issues. Medical genetics and BD analysis will play an important role in the treatment of the diseases such as various forms of cancer and insulin-dependent diabetes, and others. The complete human genome sequencing was first performed for practical purposes at the *Translational Genomics Research Institute*, Arizona, USA [31]. Prior to the availability of this opportunity, in complex clinical situations, it was limited to the search for certain mutations. As the cost of decoding the genome declines, the method becomes more accessible and bioinformatics is deeply penetrating into the healthcare. Decoding of the human gene could be achieved 10 years, while this period is reduced to a week now.

One of the first partners of personalized medicine, the first pioneer in genomic sequencing, the founder of the DNAnexus company and the Institute for Genome Research, California, is Craig Venter who suggested the joint use of genetic data through the cloud service. He is also the creator of the Human Longevity biotechnology company. The employees of DNAnexus believe that such online services are the basic for further development of genomics and a more effective adaptation method for the needs of personalized medicine [32]. Researchers also reliably share with the data access right when building the DNAnexus cloud services and securely collaborate with the third parties, such as universities, clinical laboratories, hospitals and statistical centers. It complies with the *HIPAA* (US Health Insurance Portability and Accountability Act), which provides the right of physicians' secrecy and the protection of personal data [33-35].

Conclusion

Health professionals make a justified decision through the complex technology, using clinical and other data warehouses. BD analytics is predicted to be rapidly and widely introduced in all areas of the health care system. To this end, a number of problems in this area should be solved. Many issues of BD analytics, such as privacy, security, standards and management systems, including the use of cutting-edge tools and technologies constantly draw attention. At present, BD analysis and applications in the field of healthcare is at an early stage of development. Nevertheless, the rapid development of platforms and tools can accelerate this development process.

Analytical platforms in health care should support key functions for data processing. The platform evaluation criteria include availability, accessibility, simplicity of use, extensibility, manipulation capability in various detailing levels, confidentiality and security, quality assurance and so forth. Most currently applied open source platforms have advantages and disadvantages. In the field of health care, BD analysis should be organized so that the management would be comfortable and transparent. The real-time BD analysis is a key requirement in the healthcare. The gap between the collection and processing of data should be resolved. Uninterrupted data collection and cleaning should be reviewed and important administrative issues such as the management and standards should also be considered. Health data is not always standardized, but often fragmented and generated through outdated IT systems. This big problem is necessary to be solved.

References

1. 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
2. Laura B. Madsen Data-Driven Healthcare: How Analytics and BI are Transforming the Industry, Publisher: John Wiley & Sons, Inc.,2014, 224 p.
3. Wullianallur R. Data Mining in Health Care. Healthcare Informatics: Improving Efficiency and Productivity, CRC Press, 2010, Taylor & Francis, pp.211–224. www.crcnetbase.com.
4. Wullianallur R., Viju R. Big data analytics in healthcare: promise and potential, Raghupathi and Raghupathi; licensee BioMed Central Ltd. Health Information Science and Systems, 2014, vol.2, no.3, pp.2–10.online resource,www.hissjournal.com.
5. Bill F. The taming of large data. How to extract knowledge from data arrays using deep analytics, trans. from English. Andrei Baranov, Moscow: Mann, Ivanov and Ferber, 2014, 352 p.
6. Bian J., Topaloglu U., Yu F. Towards Large-scale Twitter Mining for Drug-related Adverse Events / Proceedings of the 2012 international workshop on Smart health and wellbeing (SHB'12), New York, USA, 2012, pp.25–32.
7. Mayer-Schonberger V., Kukier K. Big data. A revolution that will change how we live, work and think, trln.from English. Inna Gaiduk, Moscow: Mann, Ivanov and Ferber, 2013, 240 p.
8. 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.
9. Institute for Health Technology Transformation (IHTT): Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry, 2013. http://c4fd63cb482ce6861463-bc6183f1c18e748a49b87a25911a0555.r93.cf2.rackcdn.com/iHT2_BigData_2013.pdf
10. Ayankoya K., Calitz A., Greyling J. Intrinsic Relations between Data Science, Big Data, Business Analytics and Datafication / Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference (SAICSIT 2014), New York, USA, 2014, pp.192.
11. Gasimova R.T. Conceptual basis for the creation of a knowledge base of domain names // News of Baku University, Physics and Mathematics series, No 4, 2010, pp.95–102.

12. Gasimova R.T. Big data analytics: existing approaches, problems and solutions // Problems of Information Technology, 2016, No1, pp.75–93.
13. Clifford L. Big data: How do your data grow? // Nature, 2008, vol.455, pp. 28–29.
14. Alguliyev RM, Hacirahimova M.Sh. Big data phenomenon: problems and opportunities // Problems of Information Technology, 2014, No2, pp.3–16.
15. Andreas H. Biomedical Informatics 2014: Discovering Knowledge in Big Data, Publisher: Springer International Publishing AG., 1st. Edition, 2014, 606 p.
16. Babu S., Herodotou H. Massively Parallel Databases and MapReduce Systems, Foundations and Trends in Databases, 2013, vol.5, no.1, pp.1–104,.
17. Deng X., Donghui W. Big data and predictive modeling topics in healthcare / Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '15), New York, USA, 2015, p.677.
18. Intel: Big Data Analytics, 2012, http://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insights-peer-research-report.pdf
19. Jimeng S., Chandan K. Big data analytics for healthcare / Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13), New York, USA, 2013, pp.1525.
20. Varun C., Sukumar Sreenivas R., Schryver Jack C. Knowledge discovery from massive healthcare claims data / Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13), New York, USA, pp.1312–1320, 2013.
21. Davenport T.H., Dwight McNeill. Analytics in Healthcare and the Life Sciences: Strategies, Implementation Methods, and Best Practices, Publisher: Pearson Education, USA, 2013, 352 p.
22. Nambiar U., Niranjan T. Data management & analytics for healthcare (DARE 2013) / Proceedings of the 22nd ACM international conference on Information & Knowledge Management (CIKM '13)New York, USA, 2013, pp.2565–2566.
23. Korolyuk I.P. Medical Informatics. Textbook, 2 ed, Samara: "OFORT", "SamGMU", 2012, 244 p. http://www.samsmu.ru/files/smu/chairs/radiology/med_inf.pdf
24. Yamakami T. Inter-service revisit analysis of three user groups using intra-day behavior in the mobile clickstream / Proceedings of the 2009 International Conference on Hybrid Information Technology (ICHIT '09), New York, USA,2009, pp.340–344.
25. Kolesnichenko O.Yu., Smorodin G.N. Big Data: Social Challenges / Abstracts of the V Sociological Grushin Conference "Big Sociology: Expanding the Data Space", Proceedings of the Conference, M: VTsIOM, 2015, pp.26–29.
26. Schmarzo B. Big Data MBA: Driving Business Strategies with Data Science, Publisher: John Wiley & Sons, Ing. 1st. Edition, 2015, 312 p.
27. Hadoop Distributed File System. http://hadoop.apache.org/docs
28. Vignesh P. Big Data Analytics with R and Hadoop, Publisher: Packt Publishing Ltd, 2013, pp.238.
29. Chuck L. Hadoop in Action, Publisher: DMK Press, 2012, 424 p.
30. Ohlhorst Frank J. Big Data Analytics: Turning Big Data into Big Money, Publisher: John Wiley & Sons Inc, 2013, 176 p.
31. Translational Genomics Research Institute, Arizona, USA, https://www.tgen.org/
32. DNAnexus, Providing cloud solutions for the global genomics industry, USA, https://www.dnanexus.com
33. Human Longevity, Inc., San Diego, California, USA, http://www.humanlongevity.com/about/j-craig-venter
34. Department of Health & Human Services (HHS) USA, http://www.hhs.gov/about/ index.html
35. Assunção M.D., Rodrigo N., Bianchi S., Netto Marco A.S., Rajkumar B. Big Data computing and clouds // Journal of Parallel and Distributed Computing, 2015, vol.79, pp.3–15.