Yadigar N. Imamverdiyev
Institute of Information Technology of ANAS, Baku, Azerbaijan
yadigar@lan.ab.az
DOI: 10.25045/jpis.v07.i1.03

BIG PROSPECTS AND PROBLEMS OF BIG DATA TECHNOLOGY

Big Data covers technologies and tools for collecting, processing, analyzing and extracting useful knowledge from structured and unstructured data of large volumes generated at high speed by different sources. Recently, scientific and popular literature promotes Big Data as technology, which opens new perspectives and revolutionary changes in e-government, business, health, science, industry and other fields. In order to determine the true potential of arguments supporting these assertions and to choose the right strategy for Big Data, this paper critically examines essentials, characteristics, basic building components and analytical capabilities of Big Data, and identifies advantages, prospects and existing problems.
Keywords: *Big Data; Big Data analytics; Data Mining; Hadoop; predictive model.*

Introduction

The development and implementation of information and communication technology (ICT) in various sectors of society during the last decade has resulted in the generation of large amounts of data of different formats from multiple sources. According to IDC's evaluations, which deal with the application of information technology market, the volume of the stored data is growing rapidly by 40% per year. the volume of data hit the limit of 1 Zettabyte (ZB) (1 ZB equals to about 1 billion gigabytes) in 2010, and the volume of data reached 2.7 ZB globally in 2012. This figure is predicted to reach 40 ZB by 2020.

E-government, science, economy, transport, communications, trade, tourism, healthcare and other areas generate large volume of data, and the performances of these areas depend more on the effectiveness of data collection, processing and analysis systems.

Large volumes of data generated at high speed by different sources cannot be processed with traditional database technologies. The problem gave an impetus for the innovations in collection, processing and storage of this kind of data. As a result, relevant models and software have been developed for collection, storage, processing and intelligent analysis of the large amounts of data in ICT industry. Computing power, the volume of data storage devices, hardware in the form of all possible sensors, as well as high-speed Internet has reached its mature level for the resolution of such issues.

In the last few years, the rapid increase in the volume and diversity of the processed data, and a number of associated technological solutions has led to the "transition from quantity to quality", which is called Big Data. Experts believe that the trend of Big Data is the driving force of ICT industry [1]. Today, evidently, Big Data is one of the leading fields opening up new perspectives for research in processing and analysis of large volumes of data from various spheres of computer science, and extracting useful knowledge from them.

Today, ICT industry is offering numerous approaches to working with Big Data both for public and business sectors. Large enterprises (banks, telecommunication operators, and retailers) may obtain almost all information about the customers by analyzing the data stored in the customer databases. Through the integration of Big Data and cloud technology, Big Data opens up great opportunities also for small businesses. In spite of all the advantages offered by Big Data technologies, only 0.5% of collected digital data is investigated.

The concept of Big Data

There are various definitions of Big Data concept. It was first defined and described by Meta Group (Gartner company) for its "*3V*" characteristics: *Volume*, *Velocity* and *Variety*. IBM added the fifth *V* based on the data quality: *Veracity*. Referring to the value of Big Data, Oracle added one more *V*: *Value* [2].

Accordingly, more comprehensive definition is called "*5V*": *Volume, Velocity, Variety, Value and Veracity* (Figure 1).

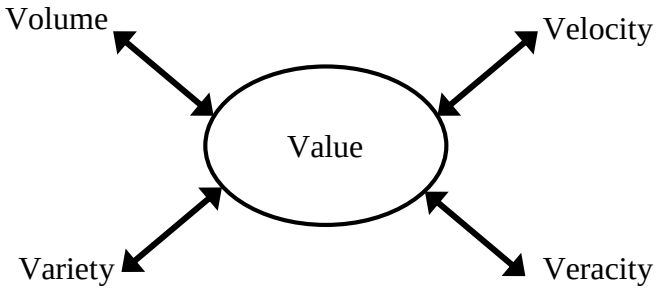

Figure 1. Big Data – 5V [2]

Volume – Big Data unable to be processed by traditional methods due to its large volume.
Velocity - speed of generating and processing new data. Big Data is a phenomenal acceleration of data collection and complication of its processing.
Variety - variety of types of processed data: structured, semi-structured, and unstructured. It is a set of tools that allows working with Big Data irrespectively of its type and volume.
Veracity - quality and source of data: good, bad, indefinite, inconsistent, incomplete and so on.
Value - is of little intensity of value; it is necessary to process very large amounts of valuable data to find necessary information.

Big Data is a new generation of technologies and architectures obtaining the greatest value from collecting, processing, and analyzing large volumes of data generated at high speed by different sources, and providing veracity of automatic quality control [3].

Big Data volume. The concept of Big Data is often described as much more than a few terabytes of data. In particular cases, some data warehouses may grow up to thousands of terabytes, that is, a petabyte (1,000 terabytes = 1 petabyte). The data more than a petabyte is measured by Exabyte (Eb), for example, the data collected on the Internet in 2009 is estimated to be 500 Eb. The metrics and proper volumes of data are given in the Table 1.
I can be noted that the following data classification is known:

- *Large data*: 1000 megabytes (1 GB) up to a few 100 GB;
- *Very large data*: 1000 GB (1 terabyte) up to a few terabytes;
- *Big Data*: a few terabytes up to hundreds of terabytes;
- *Extremely Big Data*: 1000 terabytes up to 10,000 terabytes (1 petabyte up to 10 petabytes).

Table 1. Data metrics

Metrics	**Volume, byte**	**Metrics**	**Volume, byte**
Kilobyte (Kb)	10^3	Exabyte (Eb)	10^{18}
Megabyte (Mb)	10^6	Zettabyte (Zb)	10^{21}
Gigabytes (Gb)	10^9	Yottabyte (Yb)	10^{24}
Terabyte (Tb)	10^{12}	Brontobyte (Bb)	10^{27}
Petabyte (Pb)	10^{15}	Geopbyte (Gb)	10^{30}

Big Data Sources. Social networks, log-files of web sites, and scientific data (in astronomy, physics, human genome, meteorology, biochemistry, biology) are well-known *Big Data* sources. 15-20% of data is generated by the "Internet of Things", as well as numerous phones, tablets and other devices. The data generated by the "Internet of Things" is predicted to reach 40% of the total data by 2020.

Modern medical technology also generates large amounts of data associated with the medical assistance (images, video, real-time monitoring).

Manufacturing sectors, such as power plants sometimes generate continuous data streams every minute and even every second, for tens of thousands of parameters. For several years, the application of "*Smart Grid*" technology has been measuring electricity consumption of households every minute or every second.

Big Data Velocity. Since the volume and variety of data changes, its generation velocity also changes. Today, data generation velocity makes it impossible to be processed by traditional systems. About 7 thousand petabytes of new data is generated daily; only 10% of it is structured, at the same time, this rate is constantly decreasing.

There are many organizations generating big data at great velocity. Each day, *Twitter* generates about 5 GB data a minute or 7 TB a day, *Facebook* - 7 GB a minute or 10 TB a day. *YouTube* claims that 24 hours of video is uploaded every minute.

History of Big Data. While most of the debate is related to the business applications of *Big Data*, in fact, the term appeared in a corporate environment, and offered in the scientific article, and it is one of the very few terms, the history of creation of which is precise [4]. The journal "Nature" published its special issue, on September 3, 2008, dedicated to the question "How technologies working with large amounts of data can affect the future of science?" The special issue concluded the discussions on the role of data science, in general, and e-science, in particular. Clifford Lynch, the journal editor, introduced the term *Big Data* analogous to the metaphors in Business English as "big oil" and "big ore".

However, the concept of *Big Data* is not new, it appeared in the time of mainframes and scientific calculations associated with them. It is known that scientific calculations are always distinguished by their complexity, and often occur in association with the need for large-scale data processing.

Big Data industry has emerged from the need to process large amounts of data of many companies, as the traditional methods would no longer work. For example, in some sources, *Google* processes 24 Pb (24 million gigabytes) of information a day. The price of supercomputers, managing such volumes, is too expensive for most companies, and they are looking for its alternatives. One of such ideas was connecting a large number of common computers in a network, and distributing the calculations among them. The problem was frequent accidents in such systems. The program repeating these calculations in different areas of the network solved this problem. Due to this program, the failure of one of the elements did not affect the final result [5].

Big Data became more popular after McKinsey consulting company announced report on "*Big data*: The next frontier for innovation, competition, and productivity" in June 2011. The report estimated the potential of Big Data market in billions of dollars [1]. At present, according to the potential of *Big Data*, it is generally accepted at least as the second sector of ICT industry.

***Big Data* analytics tools.** Tools and technologies for collection, management, analysis and visualization of large-scale data are available in several areas, including statistical analysis, computer science, applied mathematics and economics. Some of them were first used to work with small data, and later with large-scale data successfully; while the others were formed from scientific issues and managed by the companies (first of all, Google, Amazon, Yahoo, Facebook, etc.) aimed at working with large amount of data.

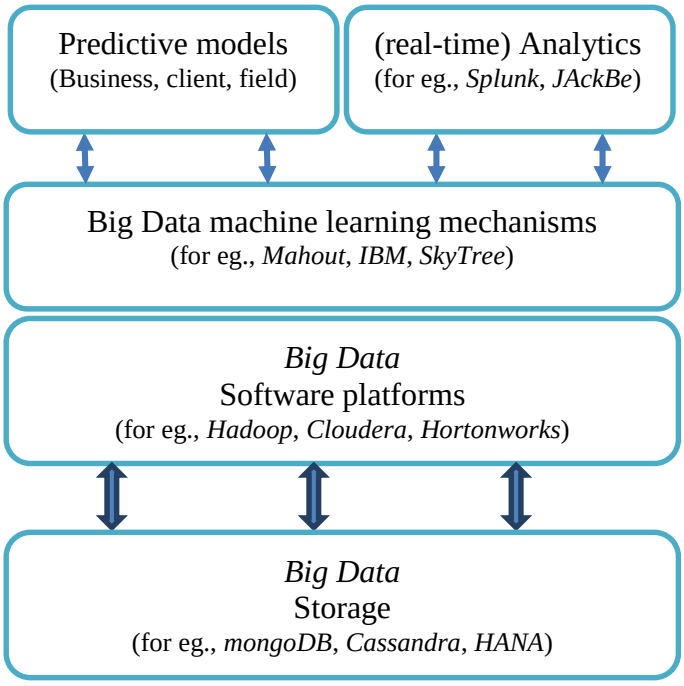

Figure 2. Big Data ecosystem

Big Data ecosystem can be defined as in Figure 2. *Big data* storage and software platforms included into the ecosystem comprise technological base of *Big Data*, which provides data collection from different sources, its storage and management. *Big Data* analytics tools are built on *Data Mining*, *Machine Learning*, and *Text Mining* [6-8].

The problems of *Big Data* ecosystem can be divided into three groups:

1. **Data Storage and Management** – hundreds of terabytes or petabytes of volume does not allow storing and processing the data by traditional relational databases.
2. **Unstructured Data Processing**- most of *Big Data* is unstructured, as text, video, audio, images, multimedia and so on. How to organize unstructured data processing and analysis?
3. ***Big Data* Analysis** - statistical analysis, Data Mining, machine learning, simulation models, optimization methods, data visualization, aggregation, integration and other methods are used for Big Data Analysis. Moreover, Predictive analytics is distinguished as a separate direction [9, 10].

Unstructured Data Processing. Unstructured Data is characterized by a number of signs, which complicate the data processing with standard analytics tools, and at the same time, which is a unique potential for the extraction of new knowledge. First, that is its *variety*. Second, it is not *ambiguous* - a data set may have different meanings depending on the context, language and cultural characteristics. Third, it is *dynamic* - data structure and values vary over time. Moreover, unstructured data often bear subjective and emotional character.

Defining the ontology (structure) of the case study is the first step in structuring the unstructured data. Ontology comprises the scheme of description of the field of study, and the rules of relating the data to this subject. As the scheme it includes the concepts, as essences, attributes the essences, and relations. Relations include the attributes that reflect service information, emotional shade of relations, an object of relation, method of relations and so on. Criteria for concepts, attributes and relationship are specified with the rules of relating the data from the unstructured data streams to this or any other subject field.

After determining the ontology of the subject field, several tools can be applied to the structured data such as search, classification, visualization, analysis, forecasting, pattern detection, identification of emotional shades and data extraction.

Unstructured Data Mining is relatively newer field of scientific research rather than Text Mining [11, 12]. Key research works in the field of Text Mining mainly focus on texts classification, clustering, summarization, feature extraction, question answering, thematic indexing, keyword searching, sentiment analysis and Opinion Mining [8, 10, 12].
Today, many leading software manufacturers offer Text Mining products such as:

- *Intelligent Miner for Text (IBM);*
- *TextAnalyst, PolyAnalyst (Megaputer);*
- *Text Miner (SAS);*
- *SemioMap (Semio Corp.);*
- *Oracle Text (Oracle);*
- *Knowledge Server (Autonomy).*

Predictive analytics is a set of statistical analysis, data analysis and game theory methods used in the analysis of present and past data or events to predict the future data or events.

Predictive is close to *Data Mining*, since the predictive analytics partly used similar methods. The main point of Predictive analytics is to define a predictor or predictors (parameters affecting the predicted event). For example, when determining insurance coverage, insurance companies review predictors such as age and driving experience. Predictive analytical model comprises majority of predictors. This model probably predicts the future of the reviewed event.
The most famous example of the usage of Predictive analytics is the introduction of scoring models to assess the solvency of the customer when issuing credit in the bank. Nonetheless, predictive analytics is applied in a wide range of fields, however, it is mostly needed in banking and financial services working with end consumers, insurance, pharmaceutical, public sector, telecommunications and information technology, and retail sales.

Eric Siegel shows ten most widespread scopes of application of predictive analytics in his book "*Predictive Analytics*" [13]: direct marketing; predictive advertising orientation; detection of fraudulent schemes; investment risk management; maintaining customers; recommendation services; education; political campaigns; decision-making systems of medicine; insurance and mortgage loan.

Building Big Data Models. Building exact Big Data models is often challenging. There are *Map-Reduce* realizers for parallel processing of large volumes of data of Different Data Mining and Machine Learning algorithms. However, the final model obtained from the processing of large amounts of data, is really hard to be considered exact.

In fact, building small data models is more profitable. One of the approaches to *Big Data* analysis includes the usage of data for segmenting and clustering entire volume of data, and then, building multiple models from the obtained small segments and clusters, and finally, forecasting on the appropriate models. In the case of limit, separate models of any person can be built in the data warehouse to predict future purchases of the customers.

Accordingly, analytics platform supporting Big Data should be capable to handle hundreds, or even thousands of models, and, if necessary, to configure them again [14].

Big Data technologies

Distributed file systems. *Big Data* (terabytes, petabytes) can be stored and systemized in the distributed file systems. Using standard hardware and open source software (e.g. *Hadoop*) in distributed file system management can relatively ease the realization of safe data warehouse [15].
Google File System, Lustre (Linux and Clusters or Lustre File System, LFS), IBM's General Parallel File System (GPFS), Hadoop Distributed File System (HDFS) and many other solutions support *Big Data* distributed file system.

Google File System is mostly designed for search, while *HDFS* is more suitable for analytical applications [16]. However, there are such application areas that none of these tools does not fully meet the requirements. The tools are required distributed data requests, and the data may be

distributed geographically and among different warehouses. In this case, the data to minimize costs and avoid unnecessary migration data and used in a multi-level system, you need to keep close to the ground.

New types of databases. Effected by a sharp increase in the volume of data, some motion starts to be noticed in database management systems (DBMS), which is considered a stable area, and it is self-evident in the appearance of *NoSQL* and *NewSQL* [17, 18].

NoSQL (*not only SQL* or *no SQL*) is a notion that conveys some approaches and projects aimed at the realization of database models significantly differing from the requests to the data via SQL used in traditional relational databases (appeared in 2009).

NoSQL is a new databases type: non-relational, distributed, open source, and horizontally scalable. Hash-tables, trees, and other data structures can be used for the description of the scheme during the application of *NoSQL* solutions.

Supporters of the concept *NoSQL* note that this concept does not completely deny any relational model and *SQL* language. The project points from the fact that *SQL* is an important tool, however it cannot be universal. One of the problems of the relational databases is associated with its poor functioning with large amounts of data. The goal of the project is to expand the capabilities of database where *SQL* is not flexible.

Effective cluster solutions. Currently, parallel database technologies are widespread. This technology provides the requests of the processor set to the integrated databases. In turn, it increases the rate of transactions, supports the functioning of multiple users simultaneously, and accelerates the implementation of complex requests.

SNA *(Shared Nothing Architecture)* – an architecture, the resources of which are not distributed, is better scaled and gets increasingly popular. *SNA* is a distributed independent computing architecture, in which each node has its memory, disk array, and input and output devices. In such architecture, each node is independent and does not share anything with other nodes of the network. Each *SNA* node has a special interaction protocol with other nodes and fulfills its own problems. The efficiency of such systems can be increased by adding processors, operational memory, disk memory to each node, or by increasing the number of such nodes.

Big Data and cloud technologies. Cloud technology, first of all, is a flexible approach that provides efficiency, scalability, migration and expansion for *Big Data* analysis [19]. Cloud environment enhances the effectiveness of the requests to data and offers flexible majority of resources to enable processing of large amounts of data. This solves the problem of ensuring sufficient amount of computing resources for the storage and processing of large-scaled data. Data is placed in several areas of cloud, which allows placing it close to the user, reducing time and increasing productivity.

Virtual company *Pivotal Initiative* has been designed to deliberately support integration of *Big Data* and cloud technologies, and it includes companies such as *Pivotal Labs, Greenplum, vFabric, Cloud Foundry, Spring* and *Cetas.* That is combining *PaaS* and *Big Data* analytics solutions in a unified structure. In this alliance, *VMware* products are responsible for the combination of infrastructure with *PaaS, Greenplum* systems - analytics, *Pivotal* - product lines, and for the creation of general commercial solution.

***Hadoop ecosystem*.** Today, Hadoop ecosystem (Figure 3) is identical to Big Data. In *Hadoop,* MapReduce technology is implemented, which provides automatic data paralleling and processing in the clusters (created by Doug Cutting and Mike Cafarella, in 2005, and *Hadoop* is derived from the name of the elephant toy of Cutting's little son). Most of the *Hadoop* components are open-source software developed within various *Apache* projects [16, 20].

A brief description of some components included into *Hadoop* ecosystem is shown below:

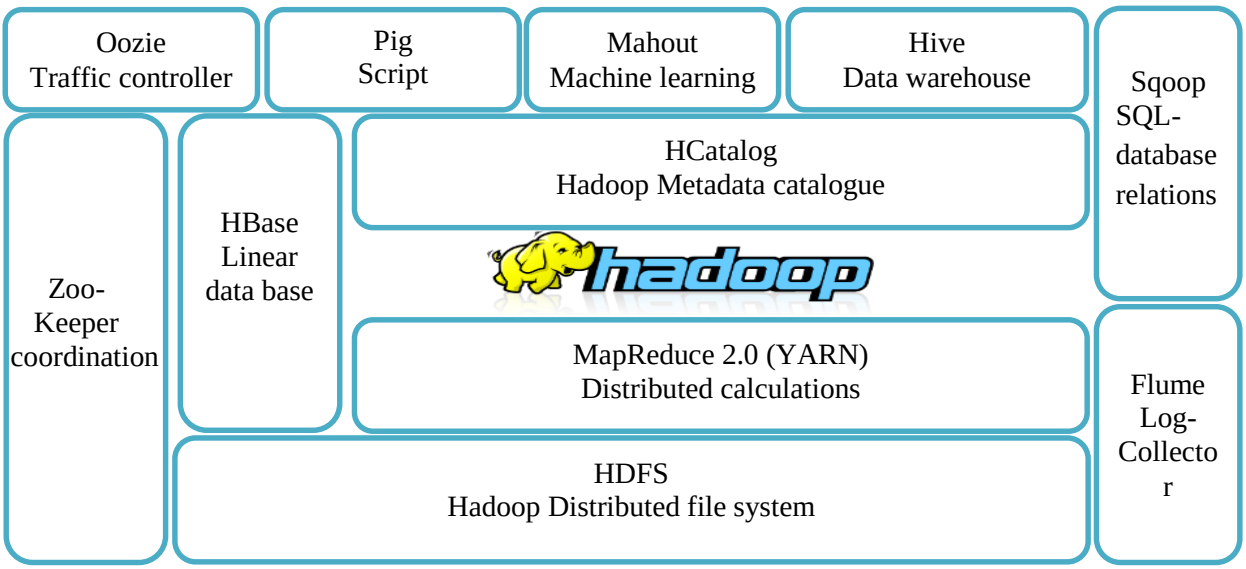

Figure 3. Hadoop ecosystem

HDFS *(Hadoop Distributed File System)* - is a core of *Hadoop*, and it is a distributed file system for storage and management of the data warehouse with the storage capacity from several petabytes up to terabytes. *HDFS* breaks input data into blocks and locates each block in the places in server sets allocated for them. *TCP/IP* level is used for communication. *HDFS* fault resistant, the failure of one of the components does not affect the work of the system. *HDFS* enables applied programs to work with hundreds of odes and petabytes of data. For example, *HDFS* system manages about 40 Petabytes of data in *Yahoo!*.

MapReduce – performs distributed computing model (in *Java*) offered by *Google* in 2004 for the parallel processing of large volumes (petabytes) of data in computer clusters. *MapReduce's* function consists of two steps: *Map* and *Reduce*. In *Map*, introduced data is preliminarily processed. For this, one of the computers (*master node*) receives preliminary data of the issue, breaks it into sections and allocates it among working nodes of distributed file systems for pre-processing. In *Reduce*, master node collects and combines preliminarily processed data, and forms the problem solution [21].

Pig - top-level language to evaluate large amounts of data. *Apache Pig* component was designed for the generation and implementation of the teams on *Big Data* sets. The main feature of *Pig* is paralleling, which helps to manage large data sets. *Pig* component consists of the compiler, which generates a sequence of *MapReduce*, and "*Pig Latin*" language. *Hadoop* supports the handling of requests similar to *SQL* in the distributed databases.

Hive – data storage infrastructure, used to request to large amounts of data located in Hadoop file system via SQL, and it fully supports *MapReduce*. Another feature of *Hive* is supporting the indexes as bit-map indexes to speed up queries. *Apache Hive*, created by *Facebook*, is now used and developed by several companies.

HCatalog - provides management services for the tables and storages of the data created in *Hadoop*. *HCatalog* offers features such as a mechanism for common schemes and data-types, table abstraction, and supports sustainable functioning of *Hadoop* components, such as *Pig, MapReduce, Streaming* and *Hive*.

HBase (*Hadoop DataBase*) – distributed and linear database (stems from *Google's BigTable*), it uses *HDFS* for storage purposes. It controls both calculations in package mode through the use of *MapReduce*, and *random reads*.

Zookeeper - another important component of *Hadoop* ecosystem. Its main function is to provide the storage of coordination details, naming, distributed synchronization and group

services, which are important for a variety of distributed systems. In fact, *HBase's* funtioning depends on *ZooKeeper*.

Mahout - software for machine learning, which includes basic algorithms as classification, clustering, and collaborative recommendation and packet filtering. The basic algorithms are realized with *Map/Reduce* paradigm in top-level of *Hadoop*, however, it can also be used beyond *Hadoop* as a software library targeted at linear algebra and statistics.

Sqoop and ***Flume*** included into the ecosystem is used to transfer data to *Hadoop*-clusters and vice versa.

Hadoop is often used together with standard technologies for storage and processing of data, and sometimes, innovative solutions such as ***Storm***, ***Dremel***, and ***Drill*** are added. In addition, practically all the major manufacturers of business analytics products add functional capabilities in order to request to the data constantly stored in *Hadoop*-clusters. The list of components can be expanded many times, because more and more companies join the market with their products related to *Hadoop* in this or other terms.

Some challenges of *Big Data*

One of the problems of *Big Data* applications is related to the evaluation of the effectiveness of *Big Data* projects. Two sources of the effectiveness of such projects exist: 1) It reduces time and costs of large data analysis, and provides quick preparation of information for timely decision making; 2) Applying *Big Data* technologies provides individualization of e-services. Along with this, despite the promises for announced economic efficiency and expediency, it is very difficult to calculate the economic efficiency of *Big Data* projects.

The second problem is related to the training of specialists, who are capable to apply *Big Data* technology in various fields. Such specialists, on one hand, should master mathematical statistics, data analysis, and machine learning, and they should have programming skills, should be able to work with hardware-software suits offered by *IBM, Oracle, HP, SAS, SAP* and other companies. On the other hand, they should have skills to formulate the problems of a specific area, where *Big Data* technologies are applied. They should know methods, scenarios and algorithms of the subject field, and should be able to form requirements for the functional characteristics of hardware and software systems, which implement *Big Data* technologies.

At the same time, abusing new technologies can also be disappointing. For example, worthless corrections can be detected as the result of Big Data analysis. Harvard University professor David Leinweber proves that the revenues of the companies included in the S & P 500 can be predicted according to high accuracy based on the volume of butter production in Bangladesh [22]. Many interesting problems can be studied and solved with the use of sparse data sets, as well.

***"Small Data"* movement**. Another important revolution lies behind the success of *Big Data*, namely *Small Data*. Due to the rapid fall in the costs of data storage, its collection and processing is democratizing. On the contrary, ecosystem of centralized, distributed data and knowledge has the biggest potential in the age of technology:

- More work can be performed in a single computer even in the organizations such as *Microsoft* and *Yahoo!* For example, in *Microsoft*, the average job size equals 14 GB and 80% of work is less than 1 TB. average job size in *Yahoo!* is approximately 12 GB [23].
- Surveys show that the work in *Facebook* is subject to the exponential distribution law, and slight work is dominating [24]. At least in 90% of the work, the size of input data is less than 100 GB. Precise study of *Hadoop* loading in *Facebook* revealed that, very small portion of work reaches terabyte and more, and in most cases, input and output data averages around megabytes and gigabytes [25].

Protection of personal information. *Big Data* technologies pose serious problems in terms of privacy. Various socio-economic actors collect plenty information about users, such as the user's behavior in web site and social networks, behavior and relationships of other persons related

to the user, their trade behavior and so on. Furthermore, emotional communication shades are also analyzed. Along with what the user writes in social media, it is analyzed how they are written. In a word, necessary or unnecessary information about the potential user is collected, and his/her overall profile is created.

Thus, based on modern surveillance technology and *Big Data* analysis, the possibility of tracking a person throughout his/her life poses a critical dilemma between the attempts to protect the integrity of personal life and information needs of the society [26].

Conclusion

At present, public interest in *Big Data* is at its peak. Thus far, it is difficult to say whether *Big Data* technology will become a necessity in the life of people as the personal computer and the Internet in the near future, but it is obvious and no one doubts that it is changing our lives and business environment.

Big Data technologies have a great potential, and it is likely to seriously affect numerous areas. Changing the architecture of corporate information is not sufficient. It is required to change the work of almost all structural units of the organization. Gradually changing the process of data analysis becomes an integral part of business processes. To take advantage of the great potential of *Big Data* technologies, the development of smart *Big Data* strategy targeted at large-scale data management and analysis is of particular importance.

References

1. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A.H. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. 2011.
2. Baaziz A., Quoniam L. How to use Big Data technologies to optimize operations in Upstream Petroleum Industry / International Journal of Innovation, 2013, vol. 1, no. 1, pp. 19-29.
3. Feblowitz J. The Big Deal about Big Data in upstream oil and gas. IDC Energy Insights. October 2012.
4. Editorial: Community cleverness required // Nature, 4 September 2008, vol. 455, no. 7209, pp. 1-1. doi:10.1038/455001a
5. Dean J., Ghemawat S. MapReduce: Simplified data processing on large clusters / Proc. of the 6th Conference on Symposium on Opearting Systems Design & Implementation (OSDI'04), 2004, vol. 6, pp. 137-150.
6. Han J., Kamber M., Jian P. Data mining: concepts and techniques. Morgan Kaufmann, 2006.
7. Bishop C.M. Pattern recognition and machine learning. Springer. 2006.
8. Feldman R., Sanger J. The Text Mining Handbook: Advanced approaches in analyzing unstructured data. Cambridge University Press, 2007.
9. Junqué de Fortuny E., Martens D., Provost F. Predictive modelling with Big Data: Is bigger really better? // Big Data, 2013, vol. 1, no. 4, pp. 215-226.
10. Weiss Sh. M., Indurkhya N., Zhang T., Damerau F. Text Mining: Predictive methods for analyzing unstructured information. Springer; 2005, 260 p.
11. Aliguliyev R.M. A new sentence similarity measure and sentence based extractive technique for automatic text summarization // Expert Systems with Applications, vol. 36, no. 4, 2009, pp. 7764–7772.
12. Alguliev R.M., Aliguliyev R.M., Isazade N.R. Multiple documents summarization based on evolutionary optimization algorithm // Expert Systems with Applications, 2013, vol. 40, no. 5, pp. 1675-1689.
13. Siegel E. Predictive Analytics: The power to predict who will click, buy, lie, or die. Wiley; 1st edition. 2013. 320 p.
14. Karthik K., Kollias G., Kumar V., Grama A. Trends in Big Data analytics / Journal of Parallel and Distributed Computing, 2014, vol. 74, no. 7, pp. 2561-2573.

15. White T. Hadoop: The definitive guide. O'Reilly Media, Inc., 2012.
16. Ghemawat S., Gobioff H., Leung S. The Google file system / Proc. of the 19th ACM Symposium on Operating Systems Principles, 2003, pp. 29-43.
17. Anglade T. noSQL Tapes. http://www.nosqltapes.com.
18. Stonebraker M., Madden S., Abadi D. J., Harizopoulos S., Hachem N., Helland P. End of an Architectural Era (It's Time for a Complete Rewrite) / Proc. of the 33rd International Conference on Very Large Data Bases (VLDB '07), 2007, pp. 1150-1160.
19. Agrawal D., Das S., El Abbadi A. Big data and cloud computing: current state and future opportunities / Proc. of the 14th International Conference on Extending Database Technology, 2011, pp. 530-533.
20. Shvachko K., Kuang H., Radia S., Chansler R. The Hadoop distributed file system / IEEE 26th Symposium on Mass Storage Systems and Technologies, 2010, pp. 1-10.
21. Lee K.H., Lee Y.J., Choi H., Chung Y.D., Moon B. Parallel data processing with MapReduce: a survey // ACM SIGMOD Record, 2012, vol. 40, no. 4, pp. 11-20.
22. Leinweber D., Stupid Data Miner tricks: Overfitting the S&P 500 // The Journal of Investing, 2007, vol. 16, no. 1, pp. 15-22.
23. Rowstron A., Narayanan D., Donnelly A., O'Shea G., Douglas A., Nobody ever got fired for using Hadoop on a cluster / Proc. of the Workshop on Hot Topics in Cloud Data Processing (HotCDP), 2012, Article No. 2. doi:10.1145/2169090.2169092
24. Ananthanarayanan G., Ghodsi A., Wang A., Borthakur D., Kandula S., Shenker S., Stoica I. PACMan: Coordinated memory caching for parallel jobs / Proc. of the 9th USENIX Conference on Networked Systems Design and Implementation, 2012, pp. 20.
25. Chen Y., Alspaugh S., Katz R.H. Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads // Proc. of the VLDB Endowment (PVLDB), 2012, vol. 5, no. 12, pp. 1802–1813.
26. Tene O., Polonetsky J. Privacy in the age of big data: A time for big decisions // Stanford Law Review Online, 2012. http://www.stanfordlawreview.org/online/privacy-paradox/big-data