BIG DATA VOLUMES AND SOME APPROACHES TO THE CREATION OF CORPORATE ANALYTICAL SYSTEMS

M. BATURA
Doctor of Engeneering Sciences
Rector of Belarusian State University of Informatics and Radioelectronics, Full Professor

S. DZIK, PhD
Vice-Rector for Education and Students Development Belarusian State University of Informatics and Radioelectronics

I. TSYRELCHUK, PhD
Head of the Department of Information and Computer Systems Design BSUIR

S. BOROVIKOV, PhD
Associate Professor, Department of Information and Computer Systems Design BSUIR

The Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: rector@bsuir.by, tsyrelchuk@bsuir.by, sdick@bsuir.by, bsm@bsuir.by

Abstract. Premises of origin, the main tasks and problems of the technology of Big Data were reviewed in the article. Authors suggest a possible approach to creating a corporate analytical system based on the technology of Big Data.

The system is intended for use with Big Data attributes and prediction value of the target variable. Predictive value of the target variable is recommended to find separately from the data obtained from different sources, and to do the resulting forecast taking into account the reliability of data sources used. Formula obtaining the resulting forecast and calculation of its reliability is present in the article.

From the history of big data.

A term Big Data was first mentioned in September 2008 on the pages of a special issue of the oldest British scientific journal Nature in the article of chief editor Clifford Lynch. This issue of the magazine was dedicated to the explosive growth of global data volumes and their role in science.

According to experts [1] Big Data - it's not just the big data volumes, but also a set of technologies, such as streaming analytics and unstructured data analytics, distributed file systems and databases, massively parallel processing. Some of them are relatively new for mass application in IT, and some received a "second wind" due to the interest in Big Data by virtue of expansion of capability in solving real tasks. Big Data - it is rather a new opportunity of fast handling of big and poorly structured data, which we received with the development of information technologies.

On the one hand, Big Data is a term denoting the boundaries of capability (applicability) of analytical systems that are based on relational Database Management Systems. At the same time this term does not give a clear definition of the boundaries, beyond which there are "big data", this limit can be individual for each company, using the analytical systems, and starts from tens or hundreds of terabytes and beyond. On

the other hand, term "Big Data" denotes the field of application of next generation technologies and architectures that are optimized for processing "big data" both structured and unstructured [1].

Algorithms of Big Data appeared at introduction of the first high-performance fault-tolerant servers (in Russian language the term "mainframe" is used) with significant input-output resources, a large volume of RAM and external memory with sufficient resources for the rapid processing of information, and are suitable for computing and for further analysis.

In fact, Big Data – it is quite conditional and relative concept. The most common definition of it – is a set of data with bigger volume than hard disk of the personal computer and which can not be handled by classical tools used for smaller volumes Some experts consider that is acceptable to call Big Data any flow of data volume of more than 100 GB per day [2].

Scientists say that in the last 2 ... 3 years, the term Big Data has become too widely used, it is used almost everywhere where data flows are mentioned, and as a consequence it was perceived too general and vague.

Big Data - is a phenomenon that is primarily due to changes in the volumes and types of information from a large number of disparate sources: e-mail, all kinds of online systems and web-based applications, video surveillance cameras, etc. Their collection and storage is an important task, but not the only, as there are other challenges: how to ensure the most effective search, analysis and retrieval of data, how to guarantee rapid access to the most important information for the business, etc. Another possible problem is that the data can be represented in different and not compatible formats that greatly complicate working with them. As a result, there is a necessity to develop special solutions for the optimization of such data sets. In practice, for many companies Big Data begins approximately from 50 TB and can reach petabytes, and this indicates that the term has not only qualitative but also quantitative characteristics.

IT-experts express the opinion that the expansion of Big Data and the acceleration of the growth rate has become an objective reality. Every second such sources as social networks, news sites, POS systems, file sharing generate giant volumes of content - and it is only a hundredth part of the suppliers. According to research by IDC Digital Universe, in the coming years the volume of data on the planet will grow up to 40 zettabytes (Figure 1), that is mean 5200 GB for every person living on Earth by 2020 [2].

According to experts about 20 ... 25% of digital data contain "useful" information. However, only about 0.5% of the world data is actually analyzed [3], that fact underlines the importance of technology and a talent to extract hidden patterns and knowledge from all these data.

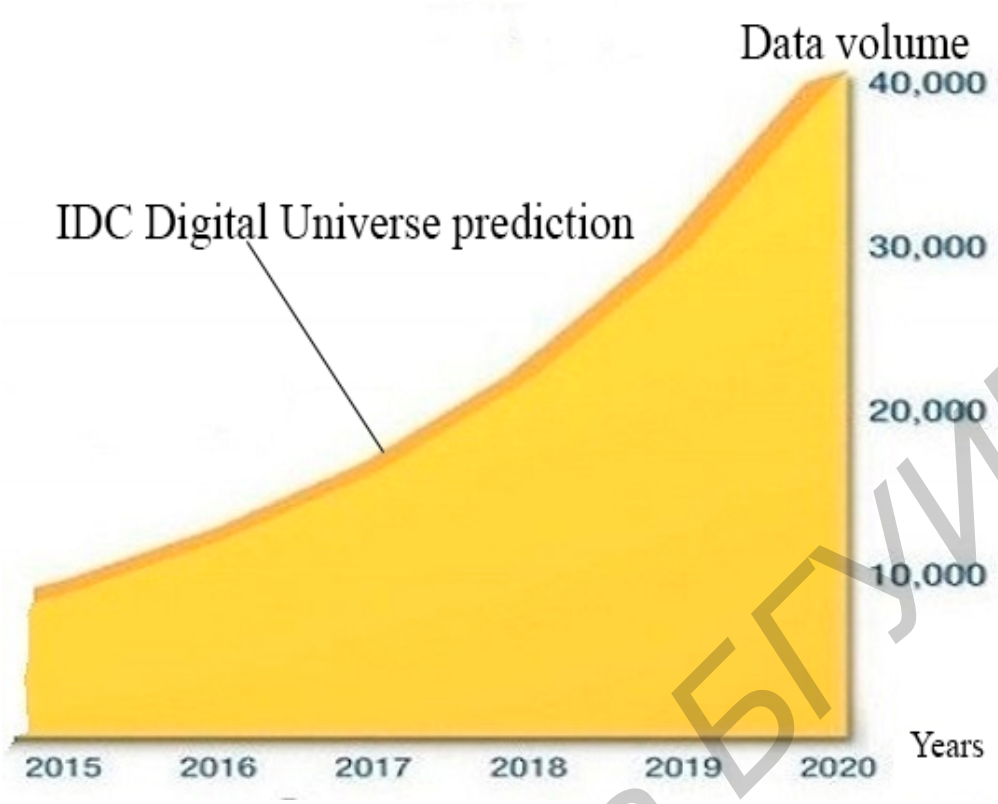

Fig. 1. Forecast of Big Data growth by 2020

In general, a Big Data processing technology can be reduced to solving the three types of tasks [3–5]:

– storage and management of hundreds of terabytes or petabytes of data volume, that can not be effectively used by conventional relational databases;

– structuring of fragmented content: texts, photos, video, audio and all other types of data;

– Big Data analysis and implementation of various methods of unstructured data processing, creation of a variety of analytical reports and forecasts.

Application of Big Data means all areas of work with the huge volume of the most exorbitant data, constantly updated and scattered over various sources. The goal is to ensure maximum efficiency of the companies, organizations, agencies, introduction of new products and the growth of competitiveness.

Problems of Big Data.

Problems of Big Data system can be summarized in three main groups related to volume, unstructuredness and high speed of data processing. As experts say - the three V: Volume, Variety and Velocity [2].

Signs of Big Data.

Volume: really big (although the size depends on the resources available for processing).

Variety: loosely structured and heterogeneous.

Velocity: processing should be very quick (and the results must be obtained

quickly when talking about online services).

Storage of large volumes of data requires special conditions, and it is a question of space and possibilities.

The problem of heterogeneity and unstructuredness arises by reason of different sources, formats and quality. To combine data and for effective data processing, not only work on bringing them to suitable form is required, but also certain analytical tools (systems).

There is also the problem of the limit of the data "size". It is difficult to establish the limit, and therefore it is difficult to predict what technologies and financial investments are required for further development. However, for certain volumes of data (terabytes, for example) existing and actively developing processing tools are already used.

Should be noted that the selection of data for processing and analysis algorithms may be large problem because there is no understanding of what data should be collected and stored, and which can be ignored.

Efficiency of activity of product manufacturing companies and provision of services (commercial, educational, health, etc.) can be described by a target variable, one or more. In many cases of predicting tasks with the use of Big Data target variable can be predicted on the basis of a set of features. Moreover this set should be exhaustive in terms of predicting the target variable for the future point in time with high reliability and minimum of subsequent risks for the company or organization.

Currently, for corporate analytical systems using Big Data characteristics of the data are very important. They are the degree of importance or influence (hereinafter Value) and the degree of reliability of the data (hereinafter Veracity). Data are received from many different sources: e-mail, all kinds of online systems and web-based application, cash and bank terminals, security cameras, etc. Big Data, obtained from different sources, have, on the one hand, a different value, and on the other hand, different veracity of the prediction for procedure of business process. The question is how to create a corporate analytical system for predicting the target variable in terms acceptable to the Company and take into account different values and reliability of Big Data.

A possible approach to creating a corporate analytical system.

The proposed approach to the creation of analytical system involves a solution for a number of tasks that can be divided into the following main stages:

1. Clarification of the target variable, which is the most important for the business activity of the company.

2. Determining the value and veracity of Big Data, derived from a variety of sources and related activities of the company

Performing this task is ensured by the definition or specification of the numerical characteristics that describe the value and the veracity of the data that will be used to find features and formation of predictive assessment of the target variable. For example, it is obvious that the data received from POS, are more valuable and reliable than the data taken from the forums in social networks.

Heuristic prediction methods or a priori known information that can be trusted with high probability can be used during performing this step.

Should be noted that during heuristic prediction analysts prepare the forecast on the basis of subjective weighting of a combination of factors, usually most of factors are qualitative in nature. The prediction result in this case depends largely on the experience and intuition of analysts.

In the simplest case, for heuristic forecast the group of specialists and experts in the amount of 7 ... 12 people is formed from among the analysts, IT-specialists, professionals in production and marketing and provision of services from the company. Formed group of experts invited to give a numerical estimate to the value and / or veracity, for example, 10 or 100-point scale, on the basis of available quantitative and qualitative information about Big Data, obtained from different sources. Moreover, the evaluation should be given by each expert independently of other experts. The resulting predictive evaluation v_{pr} of values and / or veracity of the data obtained from different sources, are finding by averaging the estimates made by different experts. Good results are obtained by averaging, that takes into account the qualifications, experience and expert intuition. In this case, we recommend the expression [6]

$$v_{pr} = \frac{\sum_{i=1}^{n} \alpha_i v_i}{\sum_{i=1}^{n} \alpha_i}, \tag{1}$$

where v_i – quantitive assessment of characteristics v (value or veracity), made by i-th expert; α_i – weighting factor of i-th expert, established for him depending on his skills, experience, intuition, etc.

Factor α_i characterizes the degree of confidence to the expert. The values α_i (i = 1, 2, …, n) can also be determined by the expert survey with the use of the expression (1) by analogy with getting assessments v_{pr}.

3. Processing of unstructured data from different sources and which was taken into account in step 2.

When performing this stage programming model Google MapReduce is suitable, it allows to solve the problem of sorting and data grouping [7]. With its help, for example, it is convenient to organize a counter of occurrence required words in a large file (Term-vector construction) or the number of frequency of references to a specified address, to calculate the volume of all web-pages from all the URL-address of a particular host or to create a list of all addresses containing the required data, etc. Google MapReduce is a powerful architecture that provides:

– automated parallelization of data from a vast array over a plurality of processing units that perform procedures Map/Reduce;

– efficient loading balance of computing nodes, which do not gives them to be idle or be overloaded;

– technology of fault-tolerant work, envisaging the fact that at performance of

general task some of processing nodes can be damaged or can stop processing data by some other reason.

Google MapReduce, on the one hand, provides the user with data processing procedures, and on the other makes the parallelization of the processing transparent on a powerful cluster Google (a group of computers, united by high-speed links and representing a unified hardware resource by user's view).

Most likely that when performing this step it is necessary to use Hadoop (Apache Software Foundation project). The main purpose of Hadoop is to provide data processing control on multiple servers and their synchronization, but only at the expense of the software, removing the cluster and hardware management [8, 9].

4. Search (on the results of phase 3) of a set of values, considered as features, and identifying those of them that have a noticeable effect on the target variable.

To perform this step, well-designed Data Mining methods are suitable, they include also statistical methods: correlation and regression analysis, factor analysis, variance analysis, component analysis, discriminant analysis, time series analysis, etc. [10, 11]. Data Mining methods can detect in the data previously unknown laws. The laws can be non-trivial, practically useful and accessible for interpretation and are necessary for making decisions about predictive assessment of the target variable in this sphere of human activity (business, education, insurance, etc.).

One of the most important appointments of Data Mining methods is a visual representation of calculation results (visualization) that allows using Data Mining toolkit by people who have no special mathematical skills. At the same time, the use of statistical methods of data analysis requires a good command of the theory of probability and mathematical statistics.

5. Getting the predictive assessment of the target variable, and generating the reports.

Forecasting is one of the tasks of Data Mining at the same time one of the key points at making decisions. Forecasting is a common and relevant task in many areas of human activity. As a result of forecasting the risk of making incorrect, unreasonable or subjective decisions decreases.

In the most general terms, forecasting procedure involves the following tasks:

- choice of a forecasting model;
- getting predictive assessment (forecast) of the target variable;
- definition of the veracity of the forecast.

It is proposed to get predicting assessment of the target variable (denoted by y) separately from the data recieved from different sources. As a result, the values y_1, y_1, …, y_m will be determined, where m – the total number of data sources taken into account. The resulting assessment y_{pr}, based on which the company will make business decisions, should be determined by taking into account the values y_1, …, y_m, which was found on the data obtained from m sources.

To determine the predictive value y_{pr} next expression is suitable:

$$y_{pr} = \frac{c_1 y_1 + c_2 y_2 + ... + c_m y_m}{c_1 + c_2 + ... + c_m}, \quad (2)$$

where c_j – factor, describing data worth, obtained from a j-th source (j = 1, 2, …, m).

Factors c_j (j = 1, 2, …, m), used in expression (2), are real numbers obtained, for example, from the 10- or 100-point scale (see. step 2).

6. Calculation of veracity of predictive assessment of the target variable y_{pr}.

Under the assumption of independence of reliability of data obtained from m different sources, the next formula should be used:

$$R_{pr} = 1 - (1 - r_1)(1 - r_2) \dots (1 - r_m), \quad (3)$$

where R_{pr} – veracity of assessment y_{pr}, expressed by probability; r_j – veracity of data, obtained from j-th source, expressed by probability (j = 1, 2, …, m).

7. Determination of risk due to the adoption of business decisions on the basis of the forecast of the target variable y_{pr} as a result of the work of analysis system, using Big Data.

risk calculation should be carried out taking into account the probability of no confidence to the forecast y_{pr}, determined by the difference (1 – R_{pr}), and size of possible losses, caused by this probability.

Литература

[1]. Kuznetsov, S.: The term Big Data hides all sorts of things [Electronic resource]. – Access mode : http://www.iksmedia.ru/articles/5033748-SKuzneczov-Pod-terminom-Big-Data.html – Access date 06.05.2016.

[2]. What is Big Data in marketing: problems, algorithms, methods of analysis [Electronic resource] – Access mode : http:// lpgenerator.ru/blog/2015/11/17/chto-takoe-big-data-bolshie-dannye-v-marketinge-problemy-algoritmy-metody-analiza/ – Access date 06.05.2016.

[3]. Working with Big Data: the main areas and opportunities [Electronic resource]. – Access mode : marketing.spb.ru›lib-research/methods/Big_Data.htm – Access date 06.05.2016.

[4]. Mayer-Schönberger, V. Big Data: A Revolution That Will Transform How We Live, Work, and Think / V. Mayer-Schönberger, K. Cukier; 2014 – 240 p.

[5]. Franks, B. Taming the big data. Finding Opportunities in Huge Data Streams with Advanced Analytics / B. Franks; 2014 – 352 p.

[6]. Sergey M. Borovikov. Theoretical bases of design, technology and reliability: textbook for universities' engineering specialties/ S.M. Borovikov. – Minsk : PRO Design, 1998. – 336 p.

[7]. Dean, J. MapReduce: Simplied Data Processing on Large Clusters / J. Dean, S. Ghemawat. Google, 2004.

[8]. Prajapati, V. Big Data Analytics with R and Hadoop / V. Prajapati. – Birmingham : Packt Publishing, 2013. – 238 p.

[9]. Shiva Achari hadoop Essentials / A. Shiva. – Birmingham : Packt Publishing Limited, 2015. – 194 p.

[10]. Data Analysis Technology / A. A. Barsegyan [and etc.] ; 3rd edition. St. Petersburg : BHV-Petersburg, 2009. – 512 p.

[11]. Paklin, N. B. Business analytics: from data to knowledge / N. B. Paklin, V. I. Oreshkov. – St. Petersburg : Piter, 2013. – 704 p.