HOW TO MASTER BIG DATA

Justyna Stasieńko

Abstract. *Information is the value that is now of critical importance in building competitive advantage. However, the amount and variety of data becomes a challenge for business and IT. Both possess a large amount of information, its accelerating growth and expectations of more effective management of these data still give birth to new technical problems and system. When BI proved to be insufficient to process the data in the shortest possible time, there is Big Data, which became the best current method of sourcing, collection and analysis of data. Using Big Data can be found as an answer for questions: How to respond to the changing expectations of customers and employees? How to improve models and business strategies? How to keep up with the next innovation, emerging at a dizzying pace? And finally, how to change your business from the inside?*

Keywords: *Business Intelligence, Business Discovery, information, analysis, QlickView, Natural Analytics*

ACM Classification Keywords: *K.6 MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS - K.6.0 General* Economics

Introduction

Data analysis, or even so called data analytics, is not a new idea. Technological development caused the increase of data amount and the ways of analysing them. It is available for these organizations that have a suitable amount of data. In the past the knowledge of organization activity was enough because new solutions, such as ERP systems or operational reporting, were used in order to have an advantage over the competition. After that the emphasis was put on processing the data by analysing them what resulted in development of Business Intelligence (BI) systems. They provide an easy access to information, its analysis and sharing within the whole organization and business surrounding [Stasieńko, 2011a)],[Stasieńko, 2011b)]. They make it possible to see the whole business of the organization. Their aim is to support the effective management of the enterprise and planning the business by delivering the proper information. The group of such tools includes the systems of information resources management, reporting and analysing tools and the solutions which help to manage the efficiency. Analysing the non-structural data is a new trend. IT appears in 'text-mining' and in the analysis of the image of social media which expanded their knowledge, which is essential for making right decisions.

Personalizing the products and services is an additional element causing the increase of the amount of information. The environment created in this way includes a big amount of data resulting from processes complexity, personalizing offers and the tendency to the offers to the groups of clients (even the smallest ones). The structure and the contents of the data which is going to be analysed come from the unknown source. Another characteristic feature is a great variability of economic models, in which many organizations function. The existing analytical and reporting data environments are able to support this process but their dynamics is not enough for business workers. That is how 'Big Data' started to exist.

Big Data – the evolution of unlimited data processing

Data and its aims were always essential for the organization. A lot of tools were created which made it easier to gather and process the data. 'Big Data' made it possible to make analysis quicker and more accurately by using the data coming from different sources. 'Big Data' is a new and large sector which helps the business to find the point in large data collections. The data may come from different sources e.g. purchase and sale transactions, public data bases, video files digital photos GPS signals from mobile phones, social networks etc. The aim of 'Big Data' is to process different types of data. Data collections such as 'Big Data' are based on four 'V': volume, variety, velocity and veracity [Płoszajski, 2013]. Volume which stands for the amount is characterised by large data beginning from petabytes collections. It is estimated that about 2,5 eksabytes of information are created every day. It means that 90% of all data in the world have been created during last two years. Variety refers to different types of data and files for which the traditional data bases are not suitably adapted e.g. sound and video files, documents, text links, network loggings and geolocation data. Velocity refers to the speed of an update and the use of data essential for creating its value. Veracity – the credibility of the information used for making decisions. Using 'Big Data' is most profitable for telecommunication companies, banks, insurance companies, Internet services: Google, Facebook as well as administration and medicine. Thanks to Big Data the bank is able to predict if the particular client divorces or has children or if they listen to rap. It all may have an influence on their credibility. Insurance companies can check if their client likes extreme sports and they can change their offers. It is up to the particular company if it uses Big Data properly or not, or if it decides to improve the cooperation with the client and minimise unnecessary costs and mistakes. Internet services, which offer various services, may also use the data. It is crucial to pay attention to the regulation and to what the user allows accepting the regulation. It is worth to be conscious that the law does not allow to be free with the clients' data. It is still being discussed if it is possible to collect and protect the clients' data and at the same time not to limit the process of its analysis. The new model gathers all data that are available and its processing is not expensive and it is used for building large data bases. After that it is possible to ask questions but not necessarily. The existing methods allow to analyse and to search for various data bases in order to find an unexpected correlation. The method used by Google is a good example [Anderson, 2008]. Google algorithms offer the word that are written the most often. Google does not use a dictionary for that but it uses the recorded answers for the previous questions asked in the past. Google suggests the word that was used the most often. The same method is used to translate texts. It searches for the sentences that have already been translated. It is an example of the machine being able to learn. In Big Data the mechanism of learning machines will become the main constituent of the business models. Learning algorithms allow the companies to follow the changing market conditions and to keep the clients. They make it also possible to find new trends. It seems that 'the revolution of infinite computing' has just started. It is the result of three trends: exponential increase of measures of computer performance, the access to them and their decrease in prices. Nowadays, the data processing is the cheapest resource used for solving the management problems. Thanks to the scalability of clouds computing it is possible to connect the powers of many computers in order to face different challenges. Processing large amount of data makes value by making the information clear and more available, creating and gathering more information about transactions in digital form for better examining all activities' effectiveness, creating more precise clients niches as well as products and services which are better adjusted to them, supporting the development of products and services of other generations and carrying out controlled experiments [Płoszajski, 2013].

As the result of this revolution the company's resources became the part of its informational system. They are able to collect and process the data, to communicate, to cooperate with others, and even to adapt or react automatically to the changes in the surrounding. They may be called 'intelligent' resources, which improve the quality of processes and will create new business models. The aim is to analyse all transactions, the clients' interactions in order to shorten the waiting time for the data and to make decisions in the real time.

In 2012 Google company introduced a free product called Analytics Content Experiments. Its role is to check the content of websites by measuring, testing and optimizing them at the same time. It makes it possible to test each version of the particular website and to decide which of them is the best. Internet giants (Amazon, Google, eBay) have already used it for some time. Nowadays, technology is available also for small companies. What is more, it is also possible now to automate completely the process of making decisions without the participation of a human being.

The problem of privacy in Big Data

Using data is well-known and important in economy, nevertheless Big Data gives more possibilities. Many companies process, analyse and visualize the data taken legally from various resource. Most often the data come from own legal resources of the company. For example, banks use all the information about their clients' accounts. They possess all the information about their payments, their shopping and their transfers to the account etc. Bank gathers all this data straight away and make it available to the clients after logging on. The organizations draw conclusions and create an outline of a particular situation, or a person who helps in their activity. It can be said that Big Data is a process based on using the data and not only on collecting it. The way of collecting it seems illegal for the clients. The world of technology causes that an average Internet user leaves a lot of their traces. On one hand it is dangerous but on the other hand it gives many possibilities of personalizing the clients, or the whole systems of managing the relations with them.

Thanks to the availability of more and more data about the clients' behavior, the activities concentrated on the client make it possible to deliver the value for the client with increasing the effectiveness of the company at the same time. Better analytics can change the assessment of activities and show how to avoid these with the smallest cots returns. The proper implementation of the clients' segmentation will constitute a tool for changing the processes within the organization in order to reach the expected goals.

Big Data allows to forecast the clients' behavior and as a result to adjust the offer to their needs, optimization of logistic and marketing activities. It also make it possible to gain knowledge about the competition. It requires the analysis of data coming from various resources e.g. the rival's prices, the amount of sold products, the clients' preferences and customer loyalty program[6].

Using Big Data, the bank workers may estimate the credibility of borrowers. They collect and analyse the information about the clients from social services, any systems of marketing information, the clients' data bases or by installing cameras and microphones in institutions. It makes the clients afraid and they associate Big Data with the invigilation and collecting data which are used for their use without any scruples. If some organizations, such as employment services, ZUS, IRS, worked with each other in correlation, it would be possible to gain information who was unemployed, who worked, who paid the taxes and how many times they have changed their employee. This knowledge would help to manage the human and financial resources better. There are also some units in big cities (e.g. New York) which work on Big Data. They improve the effective management of the city: the traffic, security services, reactions in critical situations and

[6] Amazon- the biggest market's player - using these solutions.

make it easier for the habitants to make decisions connected for example with buying the flat. They can check the buildings in every respect: redecoration, technical state etc. before they buy it or start to live in it. Checking the frequency of choosing the given words on a particular topic (e.g. a flu epidemic) it is easier to forecast something than by gathering the data from other sources (e.g. the data coming from hospital reports).

The constituents of Big Data

Gathering the data only does not guarantee a success. The data must be used skillfully. This process is supported by BI tools, which are able to aggregate the data from different sources (Oracle) or the programs analysing the data structures (MongoDB, Cassandra, Hadoop). Many of new services, which allow to collect and analyse the data, work in cloud computing. It allows to save the costs connected with creating own facility. Nevertheless, a well-qualified computing staff is indispensable here to operate the software and analysts who will create the models of analysis. Building the models and algorithms is the last part of Big Data analytics. Big Data systems used for analysing large volumes enrich the existing databases with new functions available for the users. Their germs exist in organizations in form of scattered repositories created directly by analysts. They should be provide with high performance computing solutions in order to be profitable. They can help to utilize their capability to create new business solutions. IT may successfully deliver this quality creating an unique bridge between technology and business by extending the existing data bases.

Nowadays, IT plays an important role in the organization because it takes part in building a competitive advantage by creating and adjusting the computer systems. Creating the computer system based on Big Data conception should be done taking into account the following rules. The analytic system should:

- response quickly to the questions and make analysis improving the analysts and designers' job,
- be effective and flexible as far as delivering large amount of data goes as well as enriching and making connections between them (data explorer),
- possess some analytical functions used by business analysts,
- make it possible to create interactive analyses delivered also to portable devices (designer analysts).

Taking into account technological solutions, there are tools available on the market which possess the data in different way:

- in-memory providing better quality and shorter time of calculating;
- in-database optimizing the use of the equipment power for analysing and processing the data;
- grid computing – processing the data by many computers at the same time.

The solutions described below are based on processing the data in-memory.

Big data is focused on complicated algorithms used for processing the data in huge processing centers by using highly efficient network servers. Their main role is to solve complicated calculation problems of academia, public administration as well as companies and corporations. People creating such algorithms are called the data researcher. The classic algorithmic model search through all the data in order to find connections between them. Business users prefer asking questions adhoc in order to take proper business decisions.

Fig.1 shows the process of data flow from the source to the form prepared on the angle of analyses and presentations. The raw data comes from various sources. Big Data possess the technical data from devices (e.g. web logs, server logs), transactional data (data coming from sales systems) as well as data from cloud

computing (data from social services). Such information often is unstructured (series of pictures or signs) or partly structured (logs with timestamps, IP addresses or other more detailed information). Big Data definition suggests that this data is high volume (given in terabytes or petabytes), high growth (given in gigabytes or petabytes) and it has a very high level of localization dispersion (many different databases and applications generating the data in their own formats).

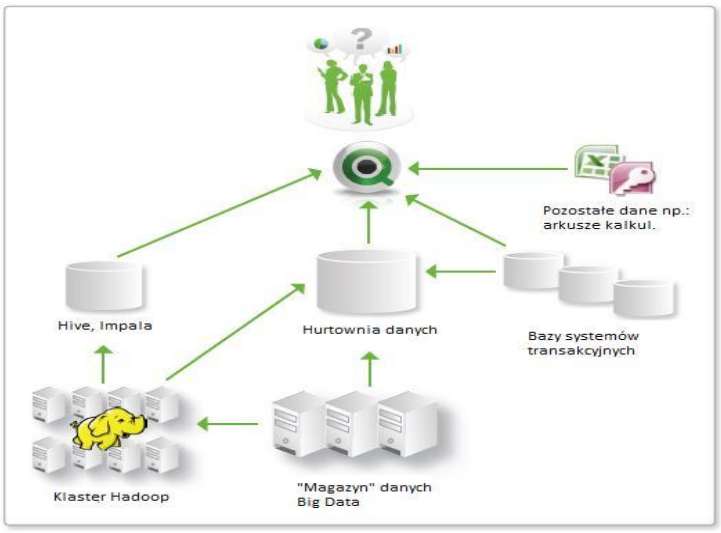


Fig.1. The data flow from the source to the analytical system using QlikView system. Source: [Pastuszek, 2013]

During the first stage of processing, if the costs of storage are high, the data is copied to Hadoop. It makes it possible to process, manipulate and aggregate the data at the same time. It is the first stage of interpreting the raw data. In the following processes the organizations use data warehouses as a central repository of unstructured data used in analyses. The data in warehouses come from Storage Area Network or Network Attached Storage) as well as from Hadoop clusters. The data in warehouses is organized and as a result it is easier to use it. Data analysis is the last stage while taking the right decisions. The current tools used for analysing should integrate the data from different sources (e.g. QlikView) regardless of what their format and origin are.

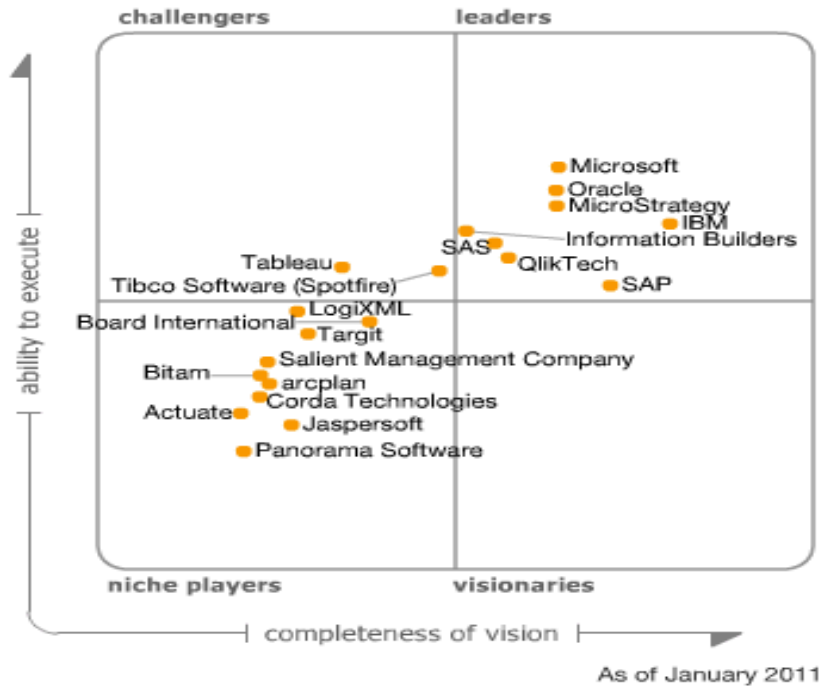
The business users require more and more as far as the access to the data, its search and analysis go without using complicated tools. Two main ways of using Big Data sources can be distinguished. The first on is when the most important data is in Big Data repository or in one repository.

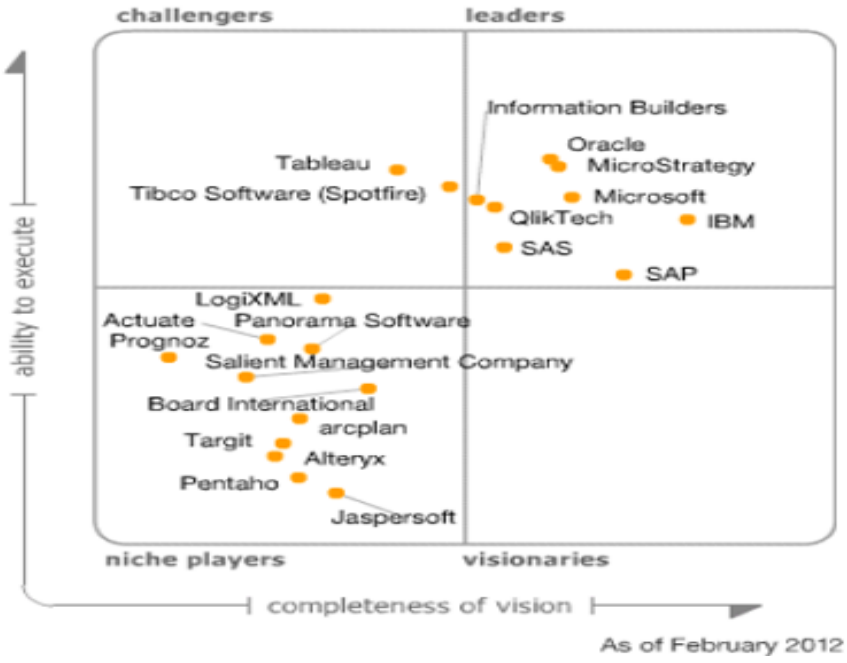
Big Data as a BI tool

Most of software potent have such a software which is used for carrying out the business analyses [Stasieńko, 2012]. So far, the market has been dominated by tools based on traditional solutions. Nowadays, the potent in IT (Microsoft, IBM, SAP, Oracle) try to include in-memory solutions in their offers. They usually create the hybrid tools which are to combination of BI and BI in-memory. On the other hand, BI products offered by these potent are getting less important as the solutions treated as the standards of analytic platforms in organizations.

The evolution of computer systems is strictly connected with the changes of consumers' behaviors e.g. these connected with using portable devices, Internet services, online shops etc. That is why, there is a need to use BI solutions in everyday life. A good example is a Social Media BI platform which process the

data left in the Internet by the Internet users. Fig.2 shows that for the last 4 years some of the companies such as Microsoft lost their leader position, and others such as QlikView and Tableau improved their situation. QlikView takes the first place among other suppliers responsible for carrying out the business analyses on their own. It has been four years since it is in the bottom square (Fig.2) among the leaders creating the software Business Analytic. Tableau has moved from one quarter of the BI magic square to the other.





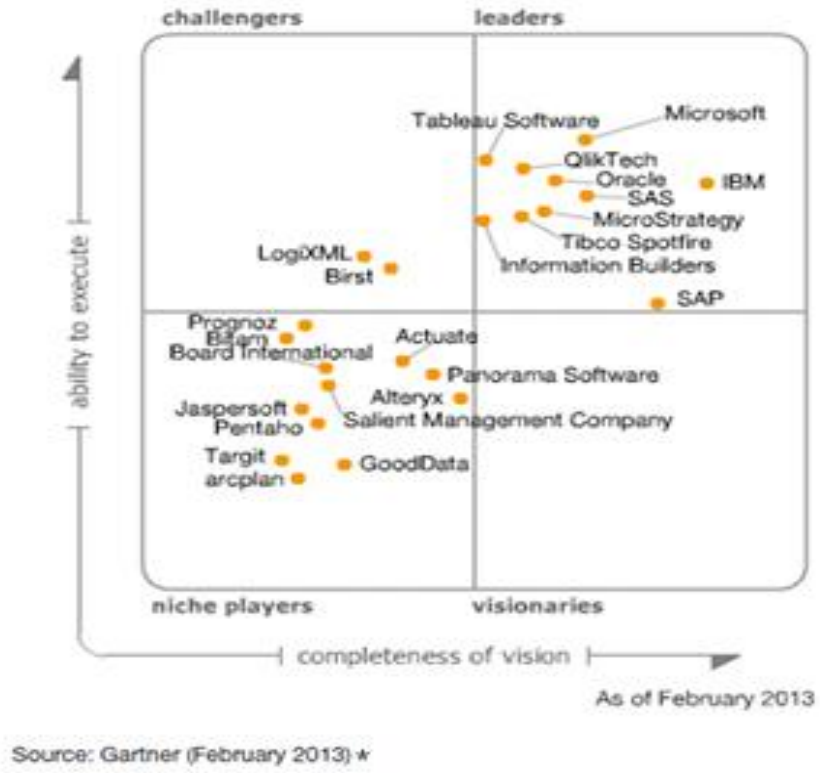


Fig.2. The magic square for BI and the analytical platforms according to Gartner[7] *:*
rok 2011 [http://bi.pl/publications/art/59-magic-quadrant-gartnera-dla-platform-business-intelligence-na-2011-rok]
rok 2011 [http://bi.pl/publications/art/59-magic-quadrant-gartnera-dla-platform-business-intelligence-na-2012-rok]
rok 2013 [source: Gartner (February 2013); http://www.sybico.pl/index.php/raport-gartnera-magiczny-kwadrant-dla-business-intelligence-i-platform-analitycznych/]
rok 2014 [source: Gartner (February 2014); http://eliasgagas.com/2014/02/25/business-intelligence-tools-magic-quadrant-by-gartner/

[7] Gartner's report shows a cross-sectional picture of the activities of suppliers in a given market segment to help end users make the best decision when choosing a partner or supplier of services or products.

QlikView – the pioneer of business analyses

QlikView is a program used to analyse data easier. It also plays an important role in Big Data implementation. It provides a quick and elastic system of presenting the data and the ability to integrate the data from different sources (Hadoop repositories, data warehouses, local data bases, spreadsheet) in one integrated solution. The most important thing is that at proper time a particular person receives proper information [Stasieńko, 2012].

QlikView offers the servicing of large amount of data. Thanks to that the client may get the best factor cost in relation to data volumes and the speed of processing it. Despite of a huge progress in technology of creating hard discs, the capacity and the access time of RAM memory are still better than other discs. That is why, if it is necessary to have 'on demand' analysis the ideal situation is to load the data into the memory. Analytical application QlikView may work on a particular data volume with information granulation which is indispensable for providing a proper level of analyses accuracy [Stasieńko, 2011], [Stasieńko, 2012]. It is possible thanks to the data compression equipment, multi-tiered architecture, the servers service and incremental loading of the data. The data loaded into the memory by QlikView is compressed up to 90% and thanks to the advanced algorithms of combining and searching the data, its processing is very effective. The solutions used recently make it possible to use great amount of memory in order to develop in-memory systems. Multi-core processors and servers make it possible to increase the computational power in a very cheap way. The servers may be divided according to their tasks. The server of the lowest level is responsible for extracting the data from the source systems. The server responsible for processing the data use the extracted data, load and process it. In QlikView there is a possibility to use the cluster of servers in order to process the data. It can also be configured in such a way which allows to take only the data which has changed from the last time. The second variant, as far as the amount of data goes, is QlikView Direct Discovery. It is a hybrid solution which combines processing the data in memory with the data loaded in currently. That is why it is intended to process large amount of data. It makes it possible to ask the BI sources without the complicated process of extracting, transforming and loading the data. The hybrid QlikView Direct Discovery attitude, makes it easier for the users to have an access to the information at the same time they need it. The users may look through the information and do not notice the difference between the data in-memory and in Big Data repository. Direct Discovery efficiency is strictly connected with the efficiency of the repository.

Another product of QlikView – QlikView.Next is an implementation of BI future. It is a complete platform of Business Analytic surrounded by a full ecosystem of people. The software and services which will satisfy the requirements of the modern organization in the realization of its strategic vision. The substructure of the platform is the conception Natural Analytic. It is based on a natural ability of human's brain to process the complex information. It uses the intuitive way of processing the information. It allows to examine the complicated data and discovering the relations between them. It is based on pairing and comparing. Natural Analytics helps to move from one data element to creating the connections with other elements. It gives potential answers intuitively, discovering new and unexpected connections between the data thanks to the cooperation and Data Dialogs. It initiates interactive discussions which refer to the data and processes in the real time. Data Dialogs helps to reach an agreement between in situations when many people work together. Thanks to Natural Analytic, BI is a real social tool. It also helps to show the connections between the data and to discover the opposed points of view while making decisions.

QlikView.Next is a new platform with a huge, associative mechanism of searching. The first clients may use the platform since 2013 but with the limited availability.

Tableau – the platform of the future

Tableau is a quite new tool which seems to be the leader on Business Intelligence market nowadays. The tools, created and developed by Tableau Software - the American company, are often called the precursor of new trends in BI solutions.

Tableau allows to quick reporting and visualizing the data from many sources. It is done by using 'in-memory' technology, which gives the possibility to analyse huge amount of data very quickly with the access to other 'on-line' sources. Tableau allows to create the data visualization automatically which are helpful while searching interesting information with the use of many dimensions (Fig. 3 and Fig. 4). This attitude, many areas of data can be analysed at the same time.

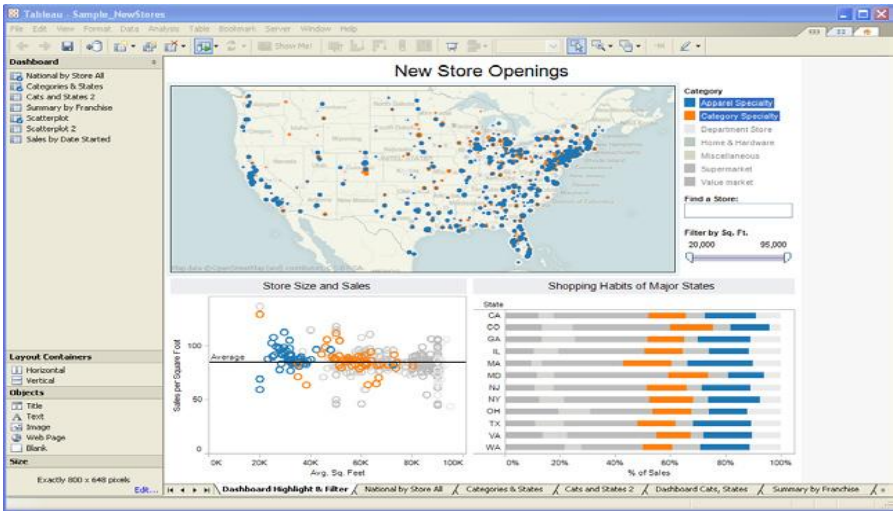


Fig. 3. Data analysis in Tableau application. Source: Own elaboration.

Tableau Business Intelligence took an advantage over other BI solutions as it is presented in Gartner's Report Fig. 2. It was done because of the intuitive solution for the user, not for the programmer, functionality adjusted to the changing needs of the users, easier Ad-hoc analyses with the use of a mouse, an easy data gathering and publishing it in the Internet.

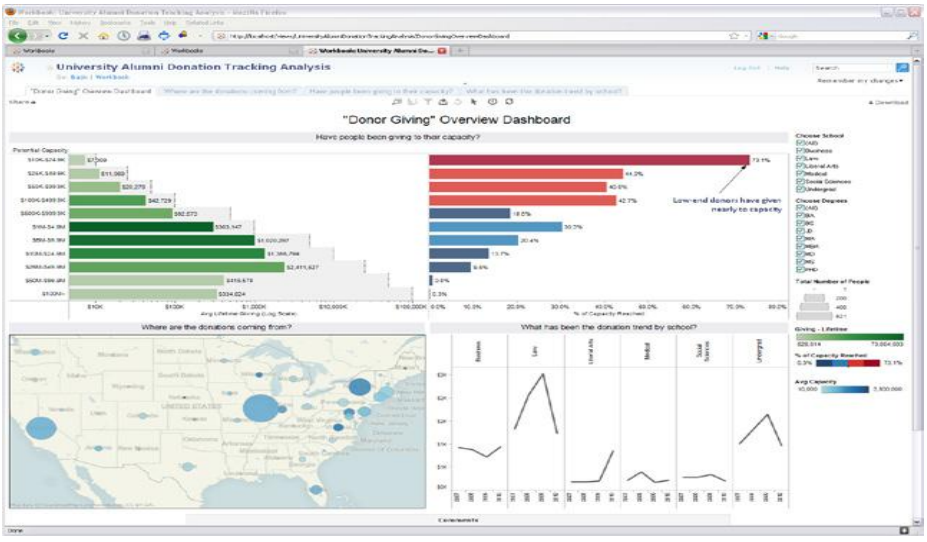
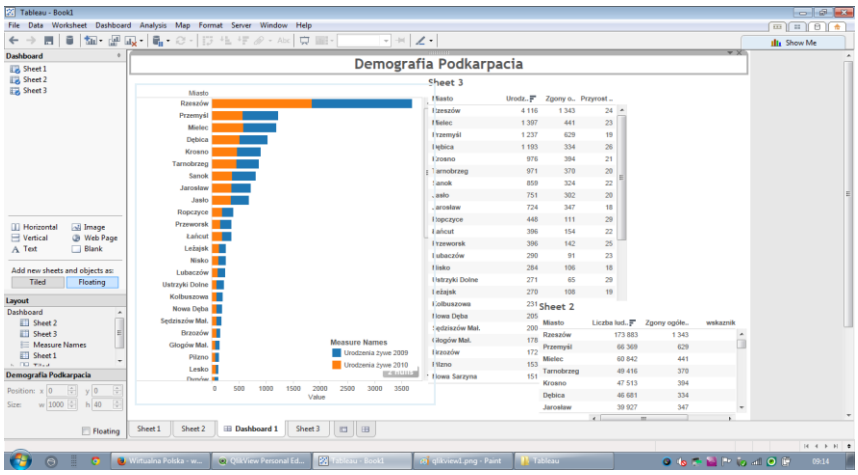


Fig. 4. Publication of data analysis from Tableau application into web browser. Source: Own elaboration.

There are many benefits for the company because of the implementation of Tableau Business Intelligence. First of all, the most important thing is the coherent and comprehensive information about what is going on in the company and about many ways of presenting the information from various data sources as well as from the scattered systems in the company. The identification of the previous unknown data is done in a very short time. All users are able to create and make available the analyses. The information is presented in a visual form (graphics, navigation desktops, reports). What is more, Tableau BI shortens the time and resources needed for reports and analytic works. It reduces the costs of supporting and implementing and it also minimize the dependence from IT. This application has low requirements as far as the equipment goes but it has a very high ROI coefficient.

Both applications are comparable as far as their use goes. They use in-memory to process large amounts of data. They are intuitive in the use. Both applications show only this data which are interesting for the client. In Tableau the Business Intelligence platform is based on web browser. Both applications do not require data warehouses. Fig.5 and Fig.6 show the visualization of the same data according to the same question. They load the data, which will be used by the client, intuitively into the application. The document format does not matter here. The interesting data is marked easily and quickly. The advantage of QlikView is the way of starting off the application. Those who use the spreadsheets may find Tableau even better. What is more, there is also the possibility to present the analyses in the Internet via dashboards.

a)



b)

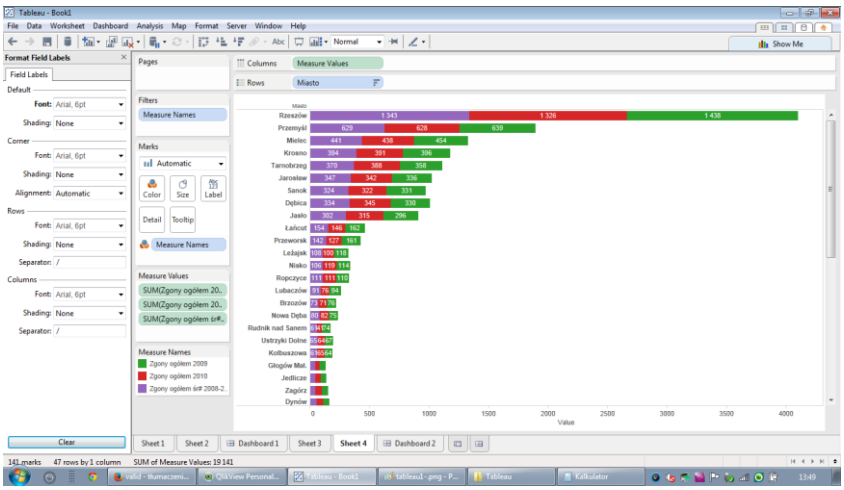


Fig.5. Presentation of data analysis in a dashboard (a) spreadsheet (b) Tableau application Source: Own elaboration.

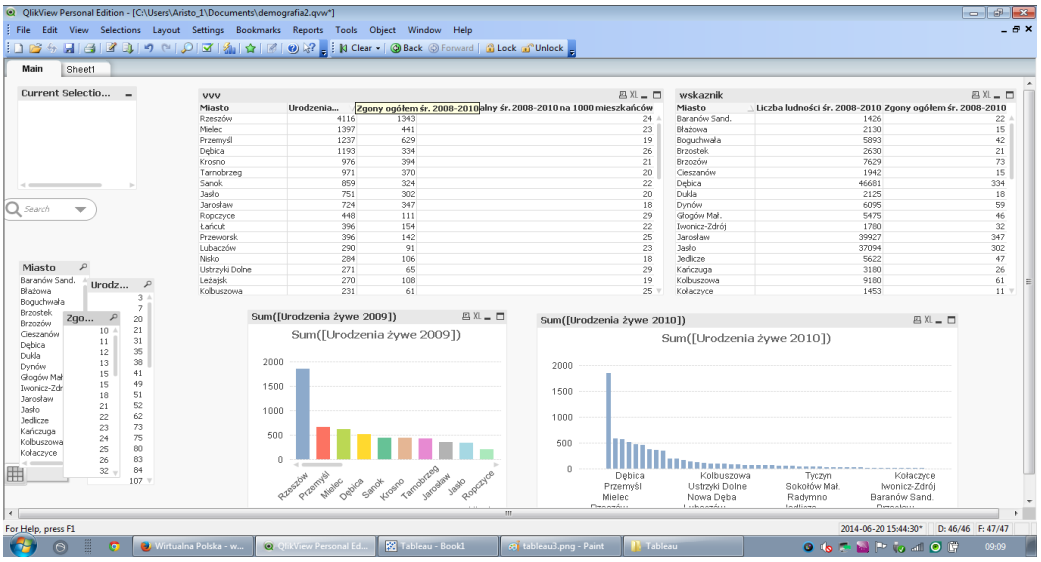


Fig.6. The presentation of data in one window of QlikView application. Source: Own elaboration.

Conclusion

Development of new tools and analytical methods is connected with the need to analyse the large amount of data which is being created all the time. Nowadays, each organization is obliged to look for new meaning and unexpected correlation in large data bases.

Big Data is not a cure for all the problems. First of all, the business needs must be defined and the first attempts should be made on a small scale. Many organization try to gather large amounts of data while most of them will not be used at all. Money spend on analytical tools and technologies do not guarantee the success. The key to it is the time of answering, asking the right business questions and the proper selection of data for the analysis. Big Data has just started to develop. There are still not enough books about this topic and no one has been translated into Polish. There are no effective ways of overcoming V4 and automatic structuring the files coming from different sources. Big Data is supposed to become a tool which can fulfill all the clients' needs.

The advanced methods of analysis will be used in the future to control the procedures. All institutions will control their workers in the future no matter what kind of organization they will be. It will be done with the use of Business Intelligence systems and the tools for thoroughgoing analysis of large amount of data. Big Data is going to be the most radical change that is going to take place in the future. This software is able to detect the economic abuse, corrupt activities and other things which are not accurate.

Bibliography

[Anderson, 2008] Anderson Ch.: The End of Theory,
http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory [05.2014].

[Januszewski, 2008] Januszewski A. Funkcjonalność Informatycznych Systemów Zarządzania, t.2 Systemy Business Intelligence, PWN, Warszawa, 2008

[Morejjon, 2012] Morejjon M.: Qliktech, IBM provide newview of OLAP.
http://www.crn.com/news/applications-os/18839582/qliktech-ibm-provide-new-view-of-olap.htm;jsessionid=dczRI7rxU7AMP3DdEVVM+g**.ecappj02 [07.2012]

[Nycz, 2008] Nycz M., Smok B. Busienss Intelligence w zarządzaniu. Materiały konferencyjne SWO, Katowice, 2008. http://www.swo.ae.katowice.pl/_pdf/421.pdf [05.2012]

[Pastuszek, 2013] Pastuszek B.: Big Data w QlikView, 2013, http://www.inmemory.bpx.pl/newsy/46-big-data-w-qlikview [04.2014]

[Płoszajski, 2013] Płoszajski P.: Big Data: nowe źródło przewag i wzrostu firm, E-mentor 3(50)/2013; http://www.e-mentor.edu.pl/artykul/index/numer/50/id/1016 [03.2014]

[Stasieńko, 2010] Stasieńko J.: BI in-memory – nowa jakoś systemów Business Intelligence. V Konferencja Naukowa Information Systems in Management – Information Systems in Management IX – Business Intelligence and Knowledge Management, Warszawa 2011, s.88-98

[Stasieńko, 2011 a)] Stasieńko J.: BI in-memory – nowa generacja narzędzi analitycznych, VII Krajowa Konferencja Bazy Danych: Aplikacje i Systemy, Zeszyty Naukowe Politechniki Śląskiej, seria INFORMATYKA, Wydawnictwo Politechniki Śląskiej, Gliwice, s.317-328.

[Stasieńko, 2011 b)] Stasieńko J.: Business Discovery – A new dimension of Business Intelligence. Methods and Instruments of Artificial Intelligence, ITHEA, Rzeszów-Sofia, 2011. s. 141-148

[Stasieńko, 2012] Stasieńko J.: BI – supporting the processes of the organization's knowledge management; 5th International Conference on Intelligent Information and Engineering Systems INFOS, Methods and Instruments of Artificial Intelligence, ITHEA, Rzeszów-Sofia 2012

[Surma 2009] Surma J. Business Intelligence – Systemy Wspomagania Decyzji Biznesowych, PWN, Warszawa, 2009

Netografia

[1] http://www.qlikview.com

[2] http://www.qlikviewaddict.com

[3] http://community.qlikview.com

[4] http://www.businessintelligence.pl

[5]http://www.comarch.pl/centrum-prasowe/aktualnosci/erp/comarch-cdn-xl-bi-start-nowosc-w-ofercie-comarch-erp

[6] http://www.sas.com/offices/europe/poland/actual/press/news2_01_13.html [27.04.2014]

[7] http://www.deloitte.com

[8] http://www.sybico.pl/?p=15 [28.04.2014]

[9] http://www.sybico.pl/ [27.04.2014]

[10] http://bi.pl/publications/art [25.04.2014]

Authors' Information

Justyna Stasieńko – lecturer, The Institute of Technical Engineering, The Bronisław Markiewicz Higher State School of Technology and Economics, Czarnieckiego Street 16, 37-500 Jarosław, Poland; e-mail: justyna.stasienko@pwste.edu.pl

Major Fields of Scientific Research: Management Information Systems, Business information technology