

Aleksandra WOŹNIAK

APPLICATION OF BIG DATA IN POLAND AND IN THE WORLD

Summary. The results of a survey conducted by the author are presented, in order to compare the Big Data tools currently used for the analysis of distributed data about the consumer between Polish and foreign companies, and to check what data is being analysed. Enterprises in Poland usually analyse data coming from their internal systems, while foreign companies examine data from mobile applications and geographical location.

Keywords: Big Data, Business Intelligence, Hadoop, Internet of Things

ZASTOSOWANIE BIG DATA W POLSCE I NA ŚWIECIE

Streszczenie. Zaprezentowano wyniki badania przeprowadzonego przez autorkę, które miało na celu m.in. porównanie stosowanych obecnie narzędzi Big Data do analizy rozproszonych danych o konsumencie pomiędzy polskimi i zagranicznymi firmami, sprawdzenie jakie dane są obecnie poddawane analizie. Firmy w Polsce najczęściej analizują dane pochodzące ze swoich systemów wewnętrznych, zagraniczne – dane pochodzące z aplikacji mobilnych i położenie geograficzne.

Słowa kluczowe: Big Data, Business Intelligence, Hadoop, Internet Rzeczy

1. Introduction

Big Data is a term used to refer to large, variable, fast flowing and diverse collections of data in which processing and analysis is difficult, but can also be valuable, because it can lead to new knowledge [1]. This concept has been gaining in importance for several years, as more and more data is captured and collected, and with cheap commodity hardware respectively juxtaposed, allows the increase in the volume of data processed [2]. In addition, open source software becomes popular, to which everyone has free access [3, 4] and thanks to which one

can analyze such data. Companies are slowly starting to profit from this practical implementation of Big Data solutions in their operational activity.

2. Applications of conducted Big Data analyses

Research accompanying the post diploma thesis of Aleksandra Woźniak entitled „Review of the Big Data tools for analysing distributed data about the consumer” was conducted in the second quarter of 2015 year. The goal of the author was to compare Big Data tools, currently used for the analysis of distributed data about the consumer between Polish and foreign companies, to see whether (and where) differences occur between those two groups or not. Such an analysis has never been conducted, so far, in Poland, nor abroad.

The survey was designed to gather information about the Big Data tools currently used for the analysis of dispersed data about the consumer in Polish and foreign companies and institutions, to check what sort of data is currently being analysed and what and how analysis companies intend to carry this out in the near future. The survey was sent to over 100 companies in Poland and 50 abroad. Responses were received from 55 Polish respondents and 19 foreign.

In the study, among others, enterprises and institutions, such as: PKN Orlen, PZU, T-Mobile, Polkomtel, Play, Virtual Poland, Onet, Interia, Allegro, Alior Bank, Procter & Gamble, the Ministry of Administration and Digitization, Facebook, Twitter, LinkedIn, eBay, Amazon.com, Alibaba, Apple, Motorola, Vodafone and Tesco, participated.

2.1. Current applications of conducted Big Data analyses

Enterprises in Poland usually analyse data coming from their internal systems. On the other hand, foreign companies examine data from mobile applications and geographical location. In the near future the two groups intend to implement analytical solutions, mainly in order to better understand the needs of their consumers.

The survey shows that the most frequently analysed data in companies operating in Poland are those derived from the internal systems, then followed by – user behavior on company websites, as well as data from mobile applications. The areas least taken into consideration, are posts on forums, the data generated by the Internet of Things (IoT) and email correspondence. Whilst, abroad, the focus is put on the analysis of: data from mobile applications, geographical location, data from internal systems of companies, user behavior on company websites and Internet of Things (Table 1).

Table 1

The data (structured and unstructured) analysed companies in Poland and abroad (the percentage of each response given by the group of respondents)

| Data being analysed | In Poland | Abroad |
|---------------------------------|------------------|---------------|
| The company's internal systems | 24% | 15% |
| User behaviour on web sites | 16% | 13% |
| Data from mobile applications | 14% | 17% |
| Geographical location | 11% | 17% |
| Sources of customer acquisition | 11% | 8% |
| Behaviour in social media | 6% | 9% |
| Interviews with consultants | 6% | 1% |
| Other | 4% | 4% |
| E-mail correspondence | 3% | 5% |
| Entries in the forums | 2% | 3% |
| Internet of Things | 2% | 9% |

Types of data being analysed in Polish enterprises vary, depending on the sector. For example, in the telecommunications industry usually geographic location is being analysed, then the data from internal systems and data about user behavior on the company's websites. In IT / E-commerce, usually, the user's behavior on the company's websites is being analysed, then the data from internal systems and mobile applications. In the sector of media and advertising, as well as in FMCG (Fast Moving Consumer Goods), most analysed are data from mobile applications and geolocation. The energy as well as the banking and insurance sectors in Poland, usually take into account data from their own internal systems.

For comparison, foreign telecommunications companies taking part in the study place emphasis primarily on the analysis of data generated by the Internet of Things, and only then – geolocation data and data from mobile applications. In contrast, the companies derived from the IT / E-commerce sectors examine data from mobile applications, geographical location, as well as user behavior on company's websites.

2.2. Planned applications of conducted Big Data analyses

The essential objectives and nature of conducted research now and in the near future will not change, much. Currently, companies operating in Poland use the Big Data analytics, primarily in order to better understand customer needs and build their value (Fig. 1). In the near future there will be a change of emphasis between those two issues. Soon, the most frequently performed analysis, will be associated with building customer value, and later, those aimed at a better understanding of their needs. The same change of priorities will be also among the foreign participants (Table 2).

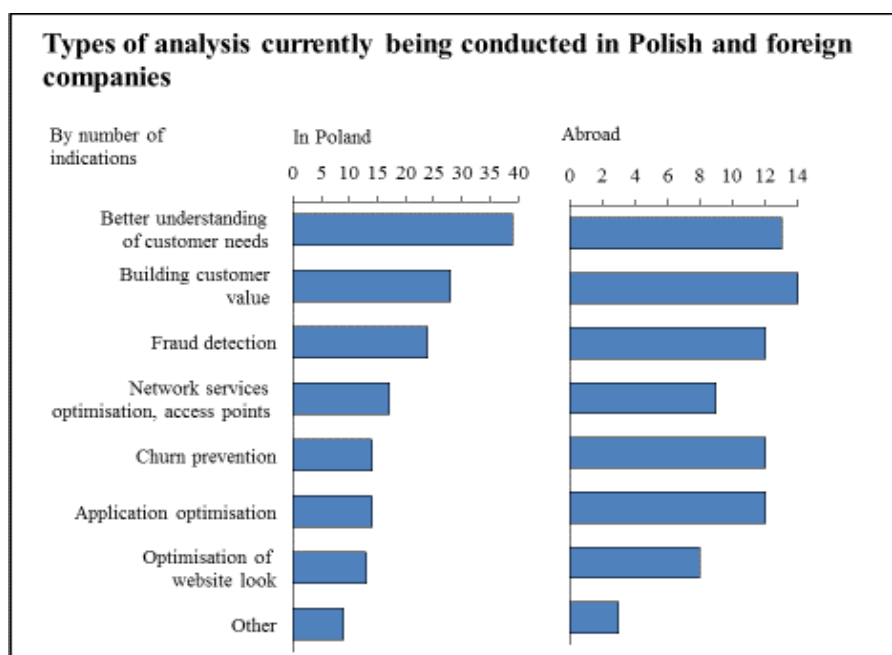


Fig. 1. The types of analysis currently being conducted in Polish and foreign companies
Rys. 1. Rodzaje analiz prowadzonych obecnie w polskich i zagranicznych przedsiębiorstwach

Table 2

The types of analysis planned to be conducted in Polish and foreign companies

| Planned data analysis | In Poland | Abroad |
|--|-----------|--------|
| Building customer value | 20% | 18% |
| Better understanding of customer needs | 19% | 22% |
| Churn prevention | 13% | 16% |
| Application optimisation | 12% | 14% |
| Fraud detection | 10% | 14% |
| Network services optimisation, access points | 9% | 8% |
| Optimisation of web site look | 9% | 8% |
| Other | 8% | 2% |

Although the types of analysis that Polish companies will carry out in the near future, may vary depending on the sectors in the economy, most of them will focus on issues related to building customer value. Other priorities will dominate in the sectors of: FMCG (preventing customer churn) and public administration (application optimisation and fraud detection).

3. Big Data technologies oriented to their needs

Big Data solutions are very popular in the area of marketing and sales (almost 40% of the responses of the surveyed companies in Poland and abroad). Less frequently noted, however, were used in finance and controlling (20% – 30% of respondents) and security (20%).

The survey results indicate that for the analysis of distributed data about the consumer, companies (both Polish and foreign) use both traditional Business Intelligence solutions (usu-

ally: Oracle, Microsoft, SAS), and various Big Data technologies (like: Hadoop, Cassandra, MongoDB). They use both commercial solutions, as well as open source (eg. R, Python).

Traditional commercial solutions are usually applied to the process of obtaining data ETL (Extract, Transform and Load), for Business Intelligence platforms, large data warehouses and data management.

Big Data tools – mostly based on open source software – are used generally used to collect large volumes of data and NoSQL (Not Only Structured Query Language) / NewSQL databases, data analysis, Business Intelligence and data visualization.

The largest companies (eg. Facebook, Alibaba etc) often develop their own Big Data solutions, based mostly on the available open source tools. This type of trend is becoming clearer – it is supported by the APIs (Application Program Interface) accessibility and engaged community of users and developers.

The vast majority of respondents (76% of Polish and 79% of foreign respondents) are satisfied with their existing tools and software to analyse Big Data. At the same time the majority of respondents participating in the survey – both Polish (75%) and foreign (95%) – declared that the Big Data solutions used in their organizations have evolved over the past few years. The reason being – at that time better Big Data tools and solutions appeared on the market (57% of Polish respondents and 56% from abroad) or less expensive tools of this type (37% and 28%) appeared. Currently, more than half of the respondents (59% of companies operating in Poland and 47% abroad) do not plan to change their tools for the analysis of large data volumes, currently used in their organizations.

4. Big Data changing the business

Most respondents see the increase in trends in the Big Data area. Several major that were mentioned in the study include: Big Data analytics in the cloud, the increasing popularity of NoSQL databases, SQL on Hadoop, real-time analytics and in-memory (of the server), “data lake”, graph databases, machine learning, “deep learning” or the Internet of Things.

Although Hadoop was designed to work on physical machines, currently there are more and more available technologies for data processing in the cloud. It seems that in the future Big Data solutions will be hosted hybridally – locally and in the cloud, which was indicated by the respondents in the survey [5]. Apache Spark is one of the open source tools that allows rapid data analysis in real-time in-memory technology. More and more often, in order to increase the flexibility of their analytics environments, companies develop Hadoop adding other solutions to it, eg. launching the Apache Spark platform and NoSQL databases [5].

Technology “deep learning” (based on a set of neural network techniques of machine learning) and machine learning, which is still at the development stage, are starting to show the enormous business potential, both in Polish and foreign companies [6].

Respondents in both groups note that, of increasing importance, is also the use of graph representation in the relationship between the data in data mining [7, 8]. An important role in the development of Big Data has begun to be played out in the Internet of Things. Foreign companies (eg. in the telecommunications industry) are already analysing the data generated by the IoT.

5. Conclusions

According to results of the survey, Polish enterprises usually analyse data coming from their internal systems, while foreign companies examine data from mobile applications and geographical location. In the near future, the two groups intend to implement analytical solutions, mainly in order to better understand their customer’s needs. For the analysis of distributed data about the consumer, companies use both traditional Business Intelligence tools (by Oracle, Microsoft, SAS), and various technologies Big Data (like Hadoop, Cassandra, MongoDB), commercial and open sources (eg. R, Python). Tools and software dedicated Big Data vary between sectors, both in Poland and abroad. In the majority of companies surveyed in Poland and abroad, used tools / software for Big Data have changed in the past few years, because of other and better and cheaper tools / software that have come onto the market. More than half of the respondents of both groups are satisfied with the current Big Data tools / software, they use.

BIBLIOGRAPHY

1. Haddad J.: How to construct a big data strategy, 14.05.2014, <http://www.techradar.com/news/world-of-tech/management/how-to-construct-a-big-data-strategy-1248021>, (27.04.2015).
2. Narzędzia i trendy Big Data, Big Data Summit, 25.03.2015, <http://bigdatasummit.pl/pdf2015/Prezentacja2.pdf>, (27.04.2015).
3. <http://blog.evolpe.pl/open-source-przyszlosc-oprogramowania/>, (27.04.2015).
4. Ching-Yung L.: Overview of Big Data Analytics, 04.09.2014, <http://www.ee.columbia.edu/~cylin/course/bigdata/EECS6893-BigDataAnalytics-Lecture1.pdf>, (27.04.2015).

5. Steliński A.: Osiem najważniejszych trendów big data, 29.10.2014, http://www.computerworld.pl/news/399634_1/Osiem.najwazniejszych.trendow.big.data.html, (27.04.2015)
6. Reszka P.: Top 7 trendy w Big Data 2015, https://docs.google.com/a/astrafox.pl/file/d/0B_fdWDPfGa5jZDhaV2s1MEpwRHc/edit?pli=1, (27.04.2015).
7. Yuhanna N.: Making Sense of Graph Database, <http://www.cs.utexas.edu/~cannata/dbms/Class%20Notes/09%20objectivityforresterwebinarslidesyuhannaclarkfinal-140717171322-phpapp01.pdf>, (27.04.2015).
8. Ching-Yung L.: Linked Big Data — Graph Computing (I), 30.10.2014, <http://www.ee.columbia.edu/~cylin/course/bigdata/EECS6893-BigDataAnalytics-Lecture9.pdf>, (27.04.2015).

Omówienie

Autorka przeprowadziła badanie w ubiegłym roku, które miało na celu porównanie stosowanych obecnie narzędzi Big Data do analizy rozproszonych danych o konsumencie między polskimi i zagranicznymi firmami, sprawdzenie, czy (i gdzie) występują różnice pomiędzy tymi dwiema grupami, jakie dane są poddawane analizie i jakie analizy firmy zamierzają przeprowadzić w tym zakresie w niedalekiej przyszłości. Ankietę wysłano do ponad 100 firm w Polsce i 50 za granicą. Działające w Polsce przedsiębiorstwa najczęściej analizują dane pochodzące ze swoich systemów wewnętrznych. Z kolei firmy zagraniczne – dane pochodzące z aplikacji mobilnych i położenie geograficzne. W najbliższym czasie obie grupy zamierzają wdrażać rozwiązania analityczne, głównie w celu lepszego zrozumienia potrzeb konsumentów. Do analizy danych rozproszonych o konsumencie przedsiębiorstwa używają zarówno tradycyjnych rozwiązań Business Intelligence (Oracle, Microsoft, SAS), jak i różnych technologii Big Data (Hadoop, Cassandra, MongoDB), komercyjnych, i open source (np. R, Python). Stosowane narzędzia i oprogramowanie dedykowane Big Data różnią się w przypadku poszczególnych branż. W większości pytaných przedsiębiorstw w Polsce i zagranicznych stosowane narzędzia/oprogramowania Big Data zmieniały się w ciągu ostatnich kilku lat, ponieważ pojawiły się inne lepsze i tańsze narzędzia/oprogramowanie. Ponad połowa ankietowanych osób jest zadowolona z obecnych narzędzi/oprogramowania Big Data.

Address

Aleksandra WOŹNIAK: Al. K.E.N. 88/54, 02-777 Warsaw, Poland, olwen@interia.pl.