Big Data Mining and Knowledge Discovery

Tomas Ruzgas, Kristina Jakubėlienė, Aistė Buivytė
Department of Applied Mathematics
Kaunas University of Technology
Kaunas, Lithuania

***Abstract*— The article dealt with exploration methods and tools for big data. It identifies the challenges encountered in the analysis of big data. Defined notion of big data. describe the technology for big data analysis. Article provides an overview of tools which are designed for big data analytics.**

Keywords*— *big data, analytics platforms

I. Introduction

In the modern world it would be difficult to find areas of human activity, in which data is not collected and analyzed. Data is rapidly increasing by development of new technologies, while increasing the need to analyze the available data [1]. Recently huge amounts of data are constantly generated and stored in the vaults in the various research fields. It is often not only the amount of data is great, but these data are constantly updated and supplemented with the new ones. In addition, there is a very wide variety of data types and sources. Such data is called big data. Processing and analysis of big data are encountered with difficulties in various fields, such as medicine, finance, economics, engineering, etc. Different methods are being developed for big data investigation tasks, such as clustering, classification, statistical and visual analysis. Study of big data is one of the biggest challenges faced by data analytics and researchers so that the methods and tools for routine data analysis, is inappropriate for big data analysis [7], [14]. Visual presentation of big data allows to detect, select and effectively use useful information. The obtained data allows the image to see the data clustering trends and data outliers. This can help in solving the data classification and clustering tasks. The goal is to explore big data research methods, techniques, used tools, in order to identify their strengths and weaknesses.

II. The Data Mining in Era of Big Data

Big data mean those data sets, which become quite difficult or impossible to hadle by using simple data-processing programs and tools because of their size and complex structure [12], [8]. Big data term is used quite often, but it is not always defined correctly [6]. Sometimes big data is characterized only by their volume. Although the word "significant" means the volume and size but big data is described by more than one characteristic [4]:

- Volume - one of the big characteristics, illustrating the size of the data.

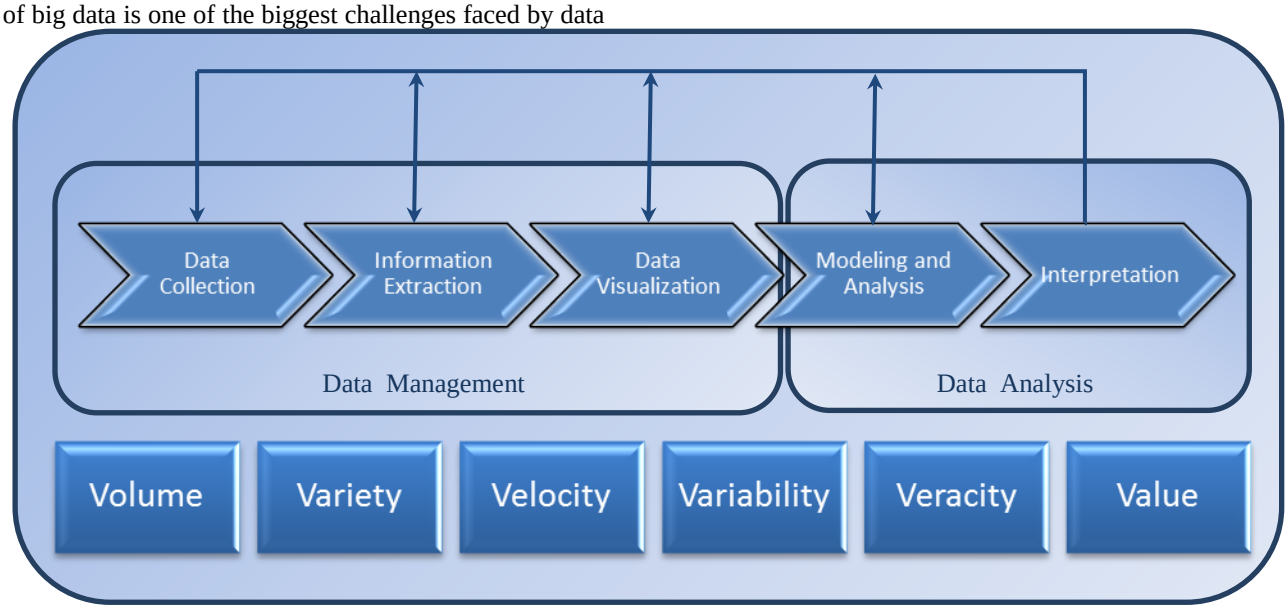

Fig. 1. Big data processing scheme.

- Variety - the characteristic that defines difference of the data types.
- Velocity - the term is understood as the satisfaction of a requirement for updating of data, growth and processing.
- Variability - characteristic used to describe the constant change of data and renewal.
- Veracity - the term is associated with the correctness and accuracy of the data and their analysis.
- Complexity or Value - associated with the ever-growing amounts of data, their variety and the problems arising from the analysis of these data.

Data mining is the main task of big data. The scheme for processing big data is presented in Fig. 1. The main steps of data processing are shown in the picture at the top. The data features which makes processing of data complex and complicated are presented in the bottom. The scheme demonstrates that the data processing is composed of many steps, each of which is encountering some challenges, requiring appropriate solutions.

Storage and processing of big data is different from traditional data analysis methods. When individual computer resources are not enough, it is recommended touse distributed and parallel systems or a distributed data mining. If the analyzed data are divided in some ways, the data research task addresses in parallel computer clusters or Grid. Computer cluster - computers combined in a single network which are able to carry out distributed computing. Grid - this is like a cluster, it is freely available, combined infrastructure, but it consists of separate computing clusters [2]. The main principle of these systems - "divide and win". We try to divide a large task into smaller, much easier for resolving and independent parts, which are carried out in parallel. Intermediate results are

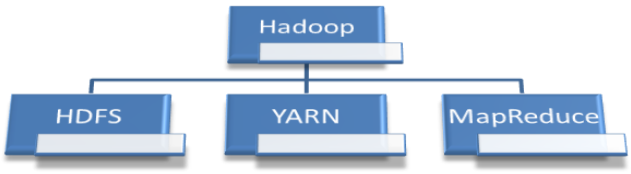

Fig. 2. Apache Hadoop modules.

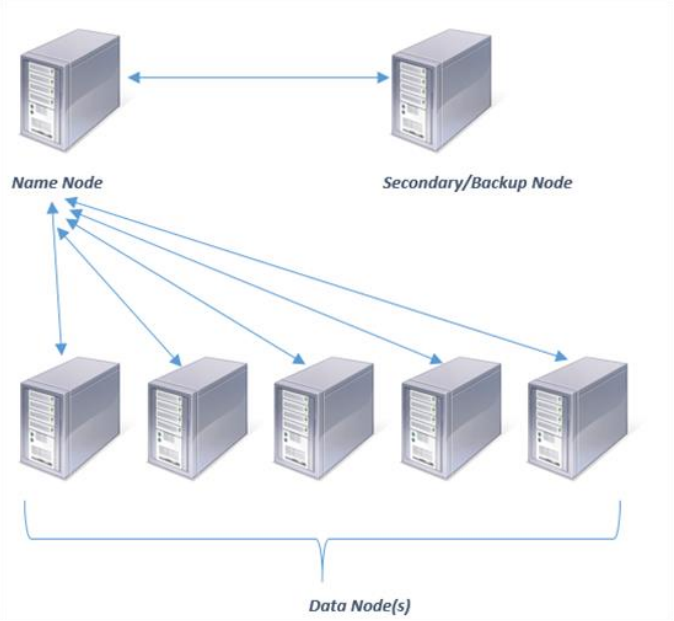

Fig. 3. Data distribution.

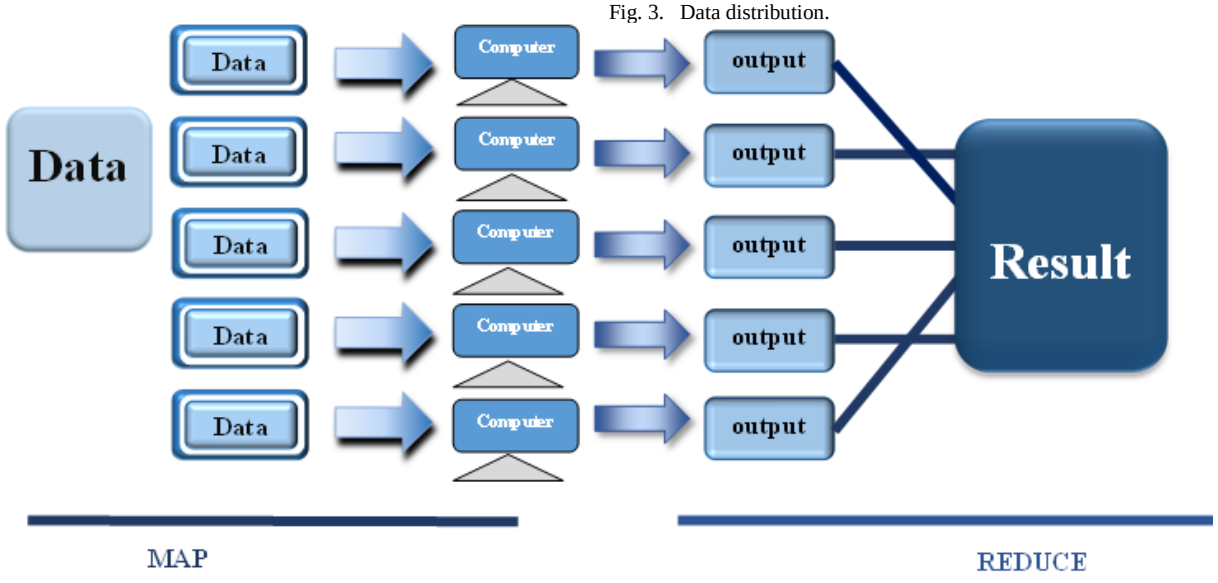

Fig. 4. MapReduce processing scheme.

combined and thenthe final result is obtained. One of characteristic example of such systems - Apache Hadoop software.

Apache Hadoop - it's open-source software, which is designed for a scalable big data distributed computing. Apache Hadoop software library - this is a system that allows to process big data sets in the computer clusters using a simple programming model. Apache Hadoop includes three modules, which are shown in Fig. 2.

Hadoop Distributed File System, HDFS) logically separates the file systems data and metadata. Compared with traditional distributed file system, HDFS has two important advantages: (1) a high error tolerance level - unlike traditional distributed file systems, which use data protection mechanisms, HDFS keeps copies of data in multiple data nodes, which allows for detected error to recover data from other data nodes; (2) the use of big data volumes - Hadoop clusters can place data sets whose size can be up to petabytes (PB).

MapReduce programming framework consists of two main steps: Map (mapping process) and Reduce (reduction process). In the Mapping process the main node is the input node, which is divided into several smaller units. Smaller units can also be split into smaller units in which the data analysis tasks are executed in parallel, and intermediate results are obtained (Fig. 3). In the process reduction step, intermediate results are combined and thus the final data analysis result is obtained (Fig. 4) [5].

III. Analytics of Big Data

One of the most important stages of the big data processing is their analytics. At this stage, it can not be such errors:

- Do not analyze all the data. It should be selected only those data which are needed for the data analytics. Excess data which do not have significant impact for the analysis should not be selected.
- Do not analyze "false" data. It has to be used only closely interconnected data. It must be considered which the data have the largest impact for the ongoing investigation.
- Introduce organized data. It is important to choose a method for data analysis: matrix, graphs, diagrams, maps, and so on. Data is convenient to provide already grouped, sorted by size, importance, etc., use color for categories, or for marking clusters [10].

Gartner Inc. advanced analytics is defined as the analysis of all types of data using complicated quantitative techniques (For example, statistics, descriptive and predictive analysis of data, modeling and optimization) in order to perform the insights of traditional business intelligence techniques - such as queries and reports - hardly reveals.

Quadrants descriptors [3]:

- *Leaders* are software vendors that are firmly established itself in the market. They can also affect the growth of the market and direction. Most organizations consider these leaders as suitable suppliers. They should not be the only vendors evaluated, but at least two are likely to be included in the typical shortlist of five to eight vendors.
- *Challengers* are classified into one of two categories. They can be long-term competitors who want to restore their vision, to keep up with changes in the market and become more influential. Or, they may be well-known vendors in adjacent markets which are associated with this market and have solutions which can reasonably be regarded as appropriate by their clients. If these suppliers demonstrate its ability to influence the market in the long term then they can become leaders.
- *Visionaries* are usually smaller suppliers who embody trends which form or will form the market. They make it possible for some organizations to jump over generation of technologies, or provide a convincing ability which provides a competitive advantage as a supplement or substitute to the existing solutions. When visionaries matured over time and have demonstrated their ability, they may eventually become as leaders.
- *Niche Players* fall into one of the two categories. Some of them are awaiting visionaries because they have a vision, but they try to make it more convincing or develop consolidation of continuing innovations what makes them as visionaries. Others are waiting Challenge - often it is the suppliers from neighboring markets which are still mature decisions in this area; their products and achievements are not always strong enough to be a safe choice for their existing customers (Challengers features), but they can become Challengers, if they continue to develop their products and demonstrate success.

According to the world's leading of information technologies research and consulting company Gartner, Inc. In 2015 the assessment made by Magic Quadrant for Advanced Analytics Platforms [3] SAS is recognized as the absolute leader in this field in the world. The assessment results are shown (Fig. 5) (Note: The best position is closest to the upper right corner). SAS Institute Inc. is located in Cary, North Carolina, U.S. With more than 40 thousand customers and the largest consumer and partner ecosystem, SAS is the most common choice among organizations which have advanced intelligence environment. SAS is anchored in the sectors of banking, insurance, business services and government.

- Benefits
  - SAS's product stack is the widest in the industry. It is most closely rivaled in terms of the range of analytic techniques available by the open-source programming environment R.
  - The strength of SAS's user community and high product scores contribute to a high level of customer

loyalty. Customers frequently praised SAS for its training programs.

- Customers identified the performance, scalability, stability and reliability of the platform as reasons for choosing SAS.

- Disadvantages
  - SAS's product stack includes multiple products with similar capabilities (for example, predictive modeling). Customers identified issues with the integration of, and interoperability between, product offerings and with the variety of UIs.
  - The complexity of the learning curve for SAS's product line is intended for nonexpert users.
  - High license fees and a complex licensing structure remain concerns for SAS users and are the main reason for organizations switching to competitors' solutions. Most of the surveyed customers thought that SAS's platform enabled them to achieve their business objectives but just a few thought that it delivered good value for money.

Fig. 5. Magic quadrant for advanced analytics platforms. Source: Gartner.

IBM is not far behind from SAS. IBM is located in Armonk, New York, U.S. He developed himself so that his analytics predictive options are available for different types of users, or the skill levels. The best-known products and solutions are SPSS Statistics and SPSS Modeler. IBM provides a wide range of analytics challenges which are associated with customers, operations, material wealth and risks.

- Benefits
  - IBM has demonstrated considerable corporate commitment to this market. It is well-known and its products' capabilities are broadly understood by prospective customers.
  - IBM's has a large user base and user community which helps it to hire experienced analytics and should increase the market's understanding of its offerings.
  - IBM's vision is strong, as shown by its integration of analytics into business-user-friendly tools, and the recent introduction of SPSS in the cloud, for example. However, additional details on the relationship between Watson Analytics (released in December 2014) and IBM's other analytics products must be forthcoming to clarify its road map.

- Disadvantages
  - IBM's product stack and individual offerings can make it difficult to use in settings where the required functionality spans many discrete product offerings. Improvements have been made, including the new integrated solutions called Predictive Customer Intelligence and Counter Fraud Management, to address this issue in specific domains, but a lack of integration among products remains a problem for the broad platform.
  - IBM must continue the progress it made in 2014 by addressing customers' continuing concerns about pricing structure and value for money with changes like the introduction of line-of-business-driven pricing models (number of customers, number of claims scored and so on) and Watson Analytics as a freemium offering.
  - Although some customers expressed high levels of satisfaction with IBM, its overall satisfaction rating was lower than average. Customers pointed to shortcomings in IBM's account management, poor or missing documentation, insufficient training, weak technical support, complexity of installation, and poor inclusion of feedback into the product development road map.

Other well-known international IT market research and consulting company is International Data Corporation (IDC) in his recent report [13], in which an advanced analytical forecasting software market is considered. It indicates that the software SAS has the largest share of the market, ie., 33.3%. IBM is in the second place, its market share is twice lower than SAS (15.8%) (see in Fig. 6).

SAS software is a "from one hand". In order to strengthen its position in the past few years, the large companies such as IBM, Oracle and SAP, have bought smaller companies that have been successful in analytics and / or operational intelligence areas. At the moment, SAS stayed the only one independent company in which data integration, business intelligence, analytics and business solutions products are from

a single supplier. SAS products use a common platform, which consist of the common system service and metadata. All metadata is stored in a common repository for metadata.

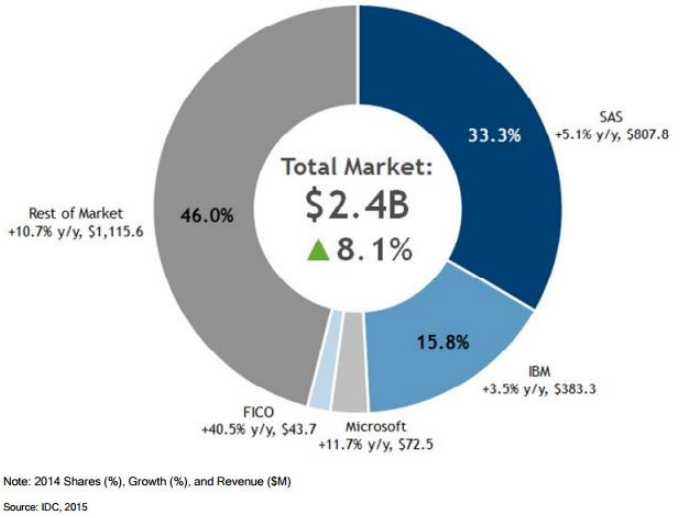

Fig. 6. Advanced analytics software share. Source: IDC

Many advanced tools for data mining are available either as open source or commercial software. They cover a wide range of software products, from comfortable problem independent data mining suites, to business centered data warehouses with integrated data mining capabilities, to early research prototypes for newly developed methods. Different types of tools vary in many different charakteristics, such as possible data structures, implemented tasks and methods, import and export capabilities, platforms and license policies are variable. Recent tools are able to handle large datasets with single features, time series, and even unstructured data like texts as powerful and generalized mining tools for multidimensional datasets [9].

Of open source data mining tools that have been examined in [11], KNIME is the package that would be recommended to those who are highly skilled. The software is simply very robust with built-in features and with additional functionality that can be obtained from third-party libraries. Based on the analysis, Weka would be considered a very close second to KNIME because of its many built-in features that require no programming or coding knowledge. In comparison, Rapid Miner and Orange would be considered appropriate for advanced users, particularly those in the hard sciences, because of the additional programming skills that are needed, and the limited visualization support that is provided. It can be concluded from above tables that though data mining is the basic concept to all tool yet, Rapid miner is the only tool which is independent of language limitation and has statistical and predictive analysis capabilities, So it can be easily used and implemented on any system, moreover it integrates maximum algorithms of other mentioned tools.

CONCLUSIONS

This article is an overview of big data inquiry tools. After analyzing the nowadays popular data mining research tools and their proposed methods, it can be said that the world's most advanced analytics leader is a SAS software. Each tool has its own advantages and disadvantages. The existing variety of analytics tools provides to choose the right tool for the data researchers, taking into account the specifics of the data and get the desired result.

REFERENCES

[1] G. Dzemyda, O. Kurasova, and J. Zilinskas, "Multidimensional data visualization", New York: Springer-Verlag, p. 252, 2013.

[2] I. Foster, "Grid technologies and applications: architecture and achievements", Astronomical Data Analysis Software and Systems XI, ASP Conference Proceedings, Vol. 281, pp. 11-16, 2002.

[3] G. Herschel, A. Linden, and L. Kart, "Magic quadrant for advanced analytics platforms", 2015 <http://www.gartner.com/technology/reprints.do?id=1-2A881DN&ct=150219&st=sb> [2016-11-30].

[4] M. Hilbert, "Digital technology and social change", Open Online Course at the University of California, 2015.

[5] T. Huang , L. Lan, X. Fang, P. An, J. Min,, and F. Wang, "Promises and challenges of big data computing in health science", Big Data Research, vol. 2(1), pp. 2–11, 2015.

[6] H.V. Jagadish, "Big data and science: myths and reality", Big Data Research, vol. 2(2), pp. 49–52, 2015.

[7] X. Jin, B.V. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research", Big Data Research, vol. 2(2), pp. 59–64, 2015.

[8] O. Kurasova, V. Marcinkevisius, V. Medvedev, A. Rapecka, and P. Stefanovic, "Strategies for big data clustering". Proceedings of IEEE 26th International Conference on Tools with Artificial Intelligence, ICTAI, pp. 740–747, 2014.

[9] R. Mikut and M. Reischl, "Data mining tools", Data mining and knowledge discovery, vol, 1(5), pp. 431-443, 2011.

[10] S. Mittelstadt, A. Stoffel, D. Keim, "Methods for compensating contrast effects in information visualization", Computer Graphics Forum, vol. 33(3), pp. 231–240, 2014.

[11] K. Rangra and K.L. Bansal, "Comparative study of data mining tools", International journal of advanced research in computer science and software engineering, vol. 4(6), pp. 216-223, 2014.

[12] C. Sherman, "What's the big deal about big data?" Online Searcher 38.2. ProQuest Central, pp. 10–17, 2014.

[13] A. Woodward and D. Vasset, "International data corporation, report: worldwide advanced and predictive analytics software market shares, 2014: the rise of the long tail", 2015 <http://www.sas.com/content/dam/SAS/en_us/doc/analystreport/idc-apa-software-market-shares-108013.pdf> [2016-11-30].

[14] L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstadt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. Keim, "Visual analytics for the big data era – a comparative review of state-of-the-art commercial systems", Proceedings of IEEE Symposium on Visual Analytics Science and Technology, pp. 173–182, 2012.

[15] Hadoop <https://hadoop.apache.org> [2016-11-30].

[16] IBM <http://www.ibm.com> [2016-11-30].

[17] International Data Corporation < https://www.idc.com > [2016-11-30].

[18] SAS Institute Inc. <http://www.sas.com> [2016-11-30].