Science, Engineering & Education, 2, (1), 2017, 50-57

Review of big data and big data mining for adding big value to enterprises

Kamal Al-Barznji, Atanas Atanassov*

University of Chemical Technology and Metallurgy, 8 Kl. Ochridski, 1756 Sofia, Bulgaria

Received 27 September 2017, Accepted 27 October 2017

ABSTRACT

This paper reviews the literature on big data, big data mining algorithms and how the big data adds value to enterprises in the real world. Big data mining for acquiring business intelligence is the main focus in the review. Our paper also covers the need for mining big data besides the rationale of considering big data for comprehensive business intelligence. It throws light into big data used cases, real-time analysis of big data with data integration, turning big data into a big value that helps enterprises to make well-informed decisions to promote business growth, social networking for big data analysis, big graph pattern mining, and other contributions that add big value to organizations.

Keywords: big data, data mining, big data mining, big databases, big value, business intelligence, Hadoop.

INTRODUCTION

Big Data is a new term used to identify the data sets that are of large size and have greater complexity [1]. Big data is defined as the large amount of data which requires new technologies and architectures to make possible to extract value from it by capturing and analysis process [2]. **Big Data Mining** refers to the activity of going through big datasets to look for relevant information. Big Data mining is the capability of extracting useful information from these large datasets or streams of data which were not possible before due to its volume, variety, and velocity. The extracted knowledge is very useful and the mined knowledge is the representation of different types of patterns and each pattern corresponds to knowledge [3]. The goals of big data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between numeral parameters [4].

The process of discovering useful knowledge which can be helpful in Big Data insights from the large stream of data and massive databases referred as Big Data mining, the characteristics of Big Data leads to challenges in Big Data mining [5].

**Correspondence to: Atanas Atanassov, University of Chemical Technology and Metallurgy, 8 Kl. Ochridski, 1756 Sofia, Bulgaria, E-mail: naso@uctm.edu*

MINING BIG DATA FOR BUSINESS INTELLIGENCE

Mining big data can provide comprehensive knowledge or business intelligence which can be used to make well-informed decisions. Enterprises in the real world need such activity on regular basis for business decisions. According to Fania and Miller big data mining can provide richer and deeper insights to enterprises for leveraging operational efficiency and competitive advantage in their chosen business. Intel IT has its wing for analyzing big data that is mostly in the form of unstructured data. Intel took up many projects in 2012 on big data. They include recommendation system, market intelligence, chip design validation, and malware detection. There were situations like "Drowning in Data, Starved for Knowledge". This is changed with the realization of distributed programming frameworks like Hadoop and emergence of cloud computing and its services such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [6].

Need for Big Data Mining

According to Wu et al. [7] big data mining is essential in order to have comprehensive business intelligence. Without considering big data which has features such as volume (data in pet bytes), variety (structured, unstructured and semi-structured data), and velocity (data streaming), flawed business intelligence may result in data mining. The blind men considering a part of elephant concluded what they saw as a hose, wall, tree and rope. Thus it is possible to have biased conclusions when a comprehensive approach is not followed. Big data mining can provide a comprehensive means of considering data to discover latent information from the big data. Such business intelligence can be used to make accurate decisions that result in business growth. Avoiding biased conclusions is very important for enterprises in the real world. The complete truth is revealed when complete data is analyzed. This incident potentially reflects the need for maintaining big data and analyzing the same in the real world [7].

Two Use Cases for Big Data

Intel identified two use cases for big data. They are known as big databases and deep analytics. These two are essential to working with big data. The former is to realize the need for special databases where structured, unstructured and semi-structured data can be stored and retrieved. The latter is the focus of big data analytics that is the discipline used to obtain hidden information or business intelligence from big data [6].

Big Databases: Big database is the database where a huge amount of structured data is stored. Such voluminous data cannot be handled by traditional Relational Database Management System (RDBMS) [6].

Deep Analytics: It is the process of analyzing big data to answer open-ended and complex problems of the real world. Towards this end tools pertaining to big data analytic and data visualization is used for obtaining valuable insights.

The tools or techniques include Hadoop, Hadoop Distributed File System (HDFS), MapReduce, Hive, HBase, Pig, Mahout, Sqoop, Oozie, and Cassandra. Hadoop is a distributed programming framework which has associated distributed file system named HDFS. It also supports new programming paradigm known as MapReduce which is suitable for processing big data. Hive is Structured Query Language (SQL) for making queries on Hadoop data. HBase is a column-oriented high-speed database that can handle millions of columns with billions of rows. Mahout is a machine learning library available for performing data mining operations on big data. Sqoop is a tool for importing or exporting RDBMS databases. An Ooze is a tool for coordinating complex data processing operations with its workflow environment. Cassandra is used to handling documents. In fact, it is a document-centric database. Massively parallel processing (MPP) data warehousing appliance is used to handle data in terabytes (TB). Such data comes from every second interaction through interactive exploration instantaneous reaction and continuous streaming. The data is transformed and loaded into Hadoop which supports data in pet bytes (PB) that is stored and executed by multiple servers for deep analytics [6].

Patel, Birla and Nair [8], used Hadoop and its MapReduce programming approach for big data processing. They described Hadoop high-level architecture which broadly shows how jobs of Hadoop clients are processed. A Hadoop cluster contains a master node and multiple workers or slave nodes. Job Tracker is responsible for scheduling jobs to a number of task trackers. Task tracker is a node that accepts tasks such as Map, Shuffle, Reduce, etc. A slave node or worker node can act as both data node and task tracker. The name node server contains distributed file system index. There is secondary name node which takes care of backup of the name node. The high-level architecture of Hadoop is shown in Fig. 1.

The data node and task tracker can share a physical node. There is communication between the task tracker and data node. The data node stores and serves blocks of data while the name node maintains a mapping of file blocks to data node slaves. Maitreya and Jhab [9] explored the usage of MapReduce programming approach with Hadoop for big data analysis. According to them, the MapReduce programming can have mappers, reducers, practitioners and combiners. Mappers are used to producing intermediate pairs. Mappers perform intended job. The results are then sorted by the framework. Afterward, the sorted output is given to reducers where the final output is produced. Partitions are used to divide the keyspace before giving to reducers. Combiners are optional and used to have a logical aggregation of data before sorting.

Bertino [10] made a review of challenges and opportunities of big data. They accept that big data processing is related to all components of the society. The rationale behind this is that individuals and organizations are directly or indirectly influenced by big data mining. It is also crucial for research. The identified technical challenges include the acquisition of data, cleaning and information extraction, data representation, data integration, query processing, and interpretation. Katal, Wazid, and Goudar [11] made a review of big data issues and tools used for big data analysis. The technical challenges they found include scalability, fault tolerance, heterogeneous data, and query of data. The tools discussed by them include Hadoop, MapReduce, and HDFS. According to Intel's white paper in [12], Intel has improved its big data framework for better performance. The framework is shown in Fig. 2.

Intel reused many existing open source components without any change. Some of the existing tools are modified by Intel besides having its own proprietary components as part of the big data analysis framework. The tools directly reused by Intel are Hive, HBase, MapReduce, and HDFS. The components that open source in nature and enhanced by Intel are Flume for log connection, Scoop for data exchange, Zoo-Keeper for coordination, Oozie for workflow, Pig for scripting,

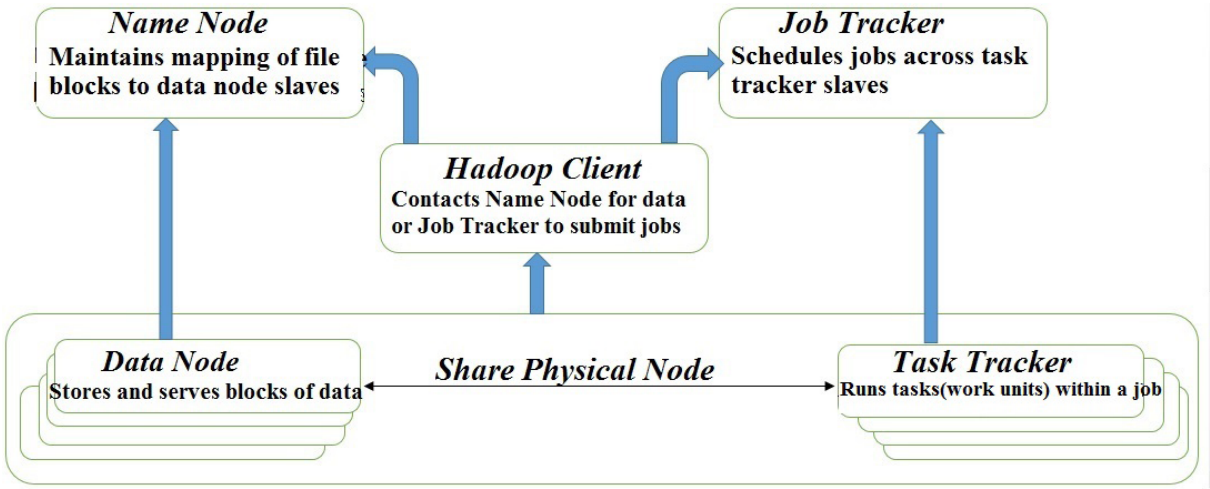

Fig.1. High-Level Architecture of Hadoop [8].

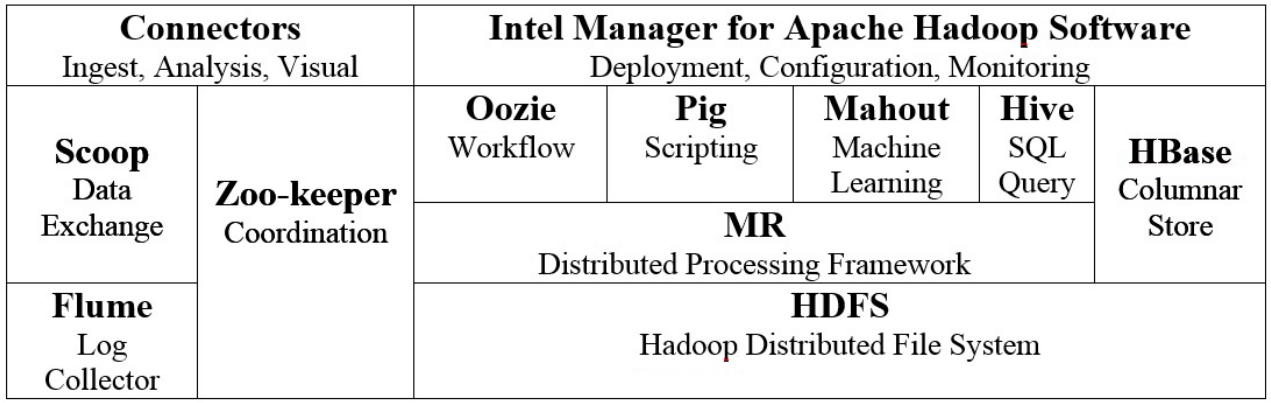

Fig. 2. Intel's Big Data Framework Based on Apache Hadoop [12].

and Mahout for machine learning. Intel's own components include connectors used for ingest, analysis and visualization and Intel manager for apache Hadoop software which is used for deployment, configuration and monitoring [12].

Khare [13] discussed relational databases and their limitations in the context of cloud computing and emerging big data processing. Not only SQL (NoSQL) is the term associated with big data. It does mean that big data refers to different kinds of data and not limited to structured data stored in relational databases. NoSQL data databases are designed to have rapid access to data which is in the form of key-value pairs. Their opinion is that when big data is not used by any organization, that organization may experiences a drop in profit margins, losses in market share, less competitive advantage, and missed opportunities in business. They also creat that with big data processing, value creation to an enterprise is possible.

Bechini, Marcelloni, Segatori [14] focused on the associative classification of big data using MapReduce paradigm. Associative classifiers are used to solve classification problems in the context of data mining. They used FP-Growth algorithm for generating classification association rules (CARs). Before mining CARs, they performed discretization of big data. As part of discretization, they had two steps namely parallel bin generation and parallel cut-point generation.

Data Mining with Big Data

Bifet [1] explored mining of big data in real time. Real-time big data analysis is needed when data is being streamed from different sources. As many organizations produce data in real time, there is every possibility that the data is not at rest. Such data grows dynamically. Processing such data needs a framework that can realize the processing of such data. Special issue on mining big data (2014) was intended to be explored big data mining process on two related domains namely healthcare and biomedicine.

Wu et al. [7] proposed a theory known as HACE for characterizing features of big data and processing of the same. They combined both data-driven and demand-driven approaches in order to have an intelligent model for data mining. Their big data processing framework is having a big data mining platform besides other things such as mining complex and dynamic data, local learning and model fusion, mining process, big data privacy and information sharing, and big data applications. The framework is shown in Fig. 3.

Big data processing framework provides the required phases that are used to have big data mining. The big data mining platform can help in discovering business intelligence from big data. Such business intelligence when interpreted and used results in organizational growth [7]. Siddaraju, Sowmya, Rashmi and Rahul [15] explored MapReduce framework for analyzing big data. Intel [16] presented a low-cost big data solution by reusing existing open source tools and applications. They used a third party solution for massively parallel processing platform, made in-house development for predictive analytics engine, enhanced Apache Hadoop software.

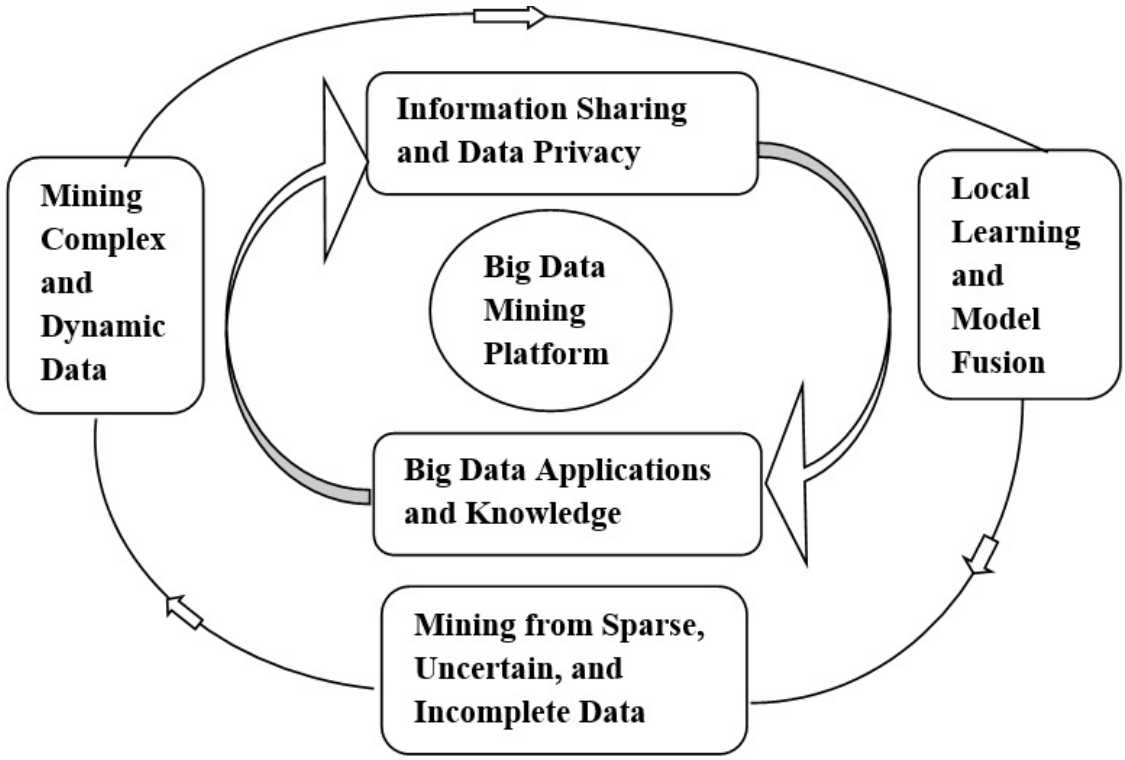

Fig. 3. Big Data Processing Framework.

Siguenza-Guzman et al. [17] made a review of data mining applications. They explored methods pertaining to clustering, classification, association, and regression. With this, different kinds of analysis such as service analysis, usage analysis, quality analysis, and collection analysis are made. Vitolo et al. [18] provided an overview of web-based technologies used for handling environmental big data. Their research resulted in understanding how big data related to weather can provide valuable information to the public. Ozkose, Ari, and Gencerb [19] explored two more characteristics of big data such as value and veracity. Veracity refers to the accuracy of data while the value refers to the result of big data that can help enterprises to get benefited.

Real-Time Analytics with Big Data Integration

Intel, Gupta, S. Shilpi [20] has integrated SAP HANA for smart data access with its big data ecosystem. Data analysis, data mining and data reporting are associated with big data processing framework. The OLAP data is integrated with big data analysis for real-time acquisition of business intelligence. Weblogs, call logs, and sensor logs are used as big data in the optimized Intel's framework. ETL is the process which has operations like Extract, Transform and Load is used by the ETL process. The data is related to weather, market and location data. SAP HANA is used to optimize data relocation, acceleration, and query federation. It is also used to have caching, hot replication and proxy tables. SAP business objects and SAP data services are also used in the framework. The framework has open source components with some Intel optimization. There are some components that are subjected to extensive optimizations [20].

TURNING BIG DATA INTO BIG VALUE

Converting big data into big value is crucial for any enterprise in the real world. Intel white paper [21] has been doing research on the big data analytics. They proposed a practical strategy for adding big value to big businesses using big data mining. Also, many enterprises got game-changing benefits with big data processing. The mobile users and social networking users are

contributing to the exponential growth of data. In addition to this Internet of Things (IoT) is also contributing to the growth of data. When such data is not analyzed, enterprises lose business opportunities. There is structured, unstructured and semi-structured data which is part of big data. There is an increase in the big data and its mining operations faster from 2006 through 2020. This indicates that there is the big value being added to enterprises when big data is subjected to obtaining comprehensive business intelligence. There are three usage models such as ETL, interactive queries and predictive analysis [21].

BIG DATA vs. SOCIAL NETWORKING VULNERABILITY

Mansour [22] made a review of big data and its related issues with respect to social networking. He thinks that big data can lead to social networking vulnerability. This is due to potential abuse of private information and the spread of malicious content across accounts related to users of social networking applications. As big data mining can bring about comprehensive business intelligence, there is the ever possibility of misusing sensitive data available in social networking applications. Thus malicious people can indulge into social engineering activities for monetary and other gains.

HOW BIG DATA ADDS VALUE TO ENTERPRISES?

According to Ningyuxin and Liyueling [23] adding big value to enterprises is possible by understanding opportunities provided by big data. By obtaining comprehensive business intelligence with big data mining value can be added in the areas such as business administration, industrial structures, public resources, government, innovation, and transmission. Deep changes in business administration are possible by obtaining the value to make promotions in marketing. Customer evaluation and understanding customer behavior can help organizations grow faster. Moreover, the accuracy of decision making is improved using big data analysis. Big data analysis can help to improve future industrial structures. The exponential growth of data can help in understanding business intelligence and make required optimizations for betterment in industries. Public services can be improved by analyzing big data as it can help in making real-time analysis and making decisions on the fly. Big data mining can help governments to understand ground realities and make expert decisions. Public information sharing and getting rid of redundant investments is possible in public sector. Big data mining can promote knowledge

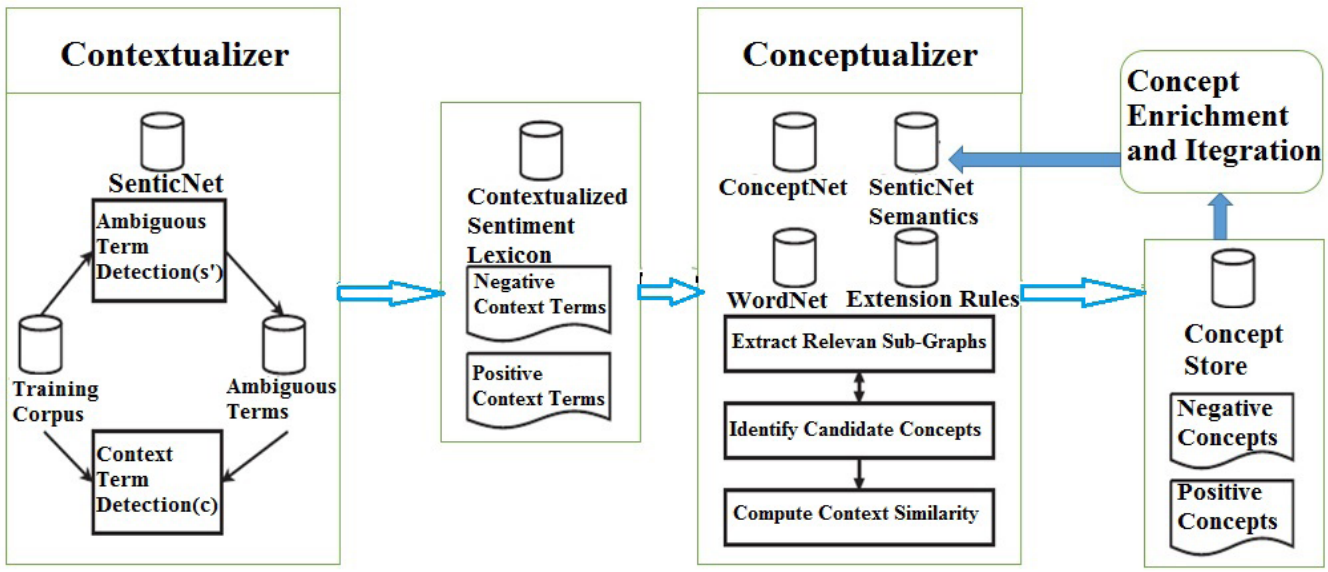

Fig. 4. Overview of Framework for Enriching Knowledge Bases [24].

innovation that leads to transformation of existing systems to have better optimizations.

OPINION MINING WITH BIG DATA

Weichselbraun, Gingl and Scharl [24] proposed a method for making semantic knowledge bases that can be used for semantic and lexical analysis. Such bases can be used for opinion mining or sentiment analysis. Their aim was to capture data from social media and apply a method to enrich knowledge bases. Towards this end, they use training corpus, perform sentiment and context term detection processes in order to identify ambiguous terms. Then they are classified into positive and negative context terms. By using both concept not and the sentence semantics more useful lexical database is generated which has the concept core associated with positive and negative concepts. The framework method is shown in Fig. 4.

The concept enrichment and integration are done in order to have enriched knowledge bases. These are used in the real world to have sentiment analysis or opinion mining. The concept of graphs and subgraphs, identifying candidate concepts and finalizing candidates is done prior to enriching knowledge bases.

CONCLUSIONS

In this paper we reviewed the literature on big data and its mining for adding big value to enterprises. Many insights extracted from the review of the literature are as follows. Big data, when processed, can provide business intelligence that is comprehensive and accurate. Unless big data is considered for mining, the acquired business intelligence is inadequate to make decisions. Stating differently, it results in biased intelligence [25, 26]. From the review of the literature, it is understood that big data processing is done in distributed environment. Hadoop kind of distributed programming frameworks is needed for processing big data. Data centers associated with cloud are generally used to store big data.

The gaps found in the literature are summarized here. Big data processing is relatively new phenomena. Here the algorithms need to consider data with characteristics such as volume, variety, and velocity. Volume refers to the huge amount of data with exponential growth. Variety refers to structured, semi-structured and unstructured data. Velocity refers to continuous streaming of new data gets added to data sources. Considering these insights, it is clear that new algorithms are needed in the area of finding frequent patterns, collaborative filtering for recommendations in order to have knowledge discovered from big data.

REFERENCES

1. A. Bifet, Mining Big Data in Real Time, Informatica (Slovenia), 2013, p.15-20.
2. S. Lenka Venkata, A Survey on Challenges and Advantages in Big Data, IJCST, 2015, p.115-119.
3. K.U. Jaseena, J.M. David, Issues, Challenges, and Solutions: Big Data Mining, Compute. Sci. Inf. Technol., 2014, p.131-140.
4. D. Che, M. Safran, Z. Peng, From Big Data to Big Data Mining : Challenges, Issues, and Opportunities, 18th Int. Conf. DASFAA, Springer-Verlag Berlin Heidelb, 2013, p. 1-15.
5. S. Gole, A survey of Big Data in social media using data mining techniques, IEEE Int. Conf. Adv. Comput. Commun. Syst., 2015, p. 5-10.
6. M. Fania, J.D. Miller, Mining Big Data in the Enterprise for Better Business Intelligence, Intel IT Best Practices-Business Intelligence, 2012, p. 1-7.
7. Gong-Qing Wu, Wei Ding, Data Mining with Big Data, IEEE, 2014, p.1-11.
8. Aditya B. Patel, Manashvi Birla, Ushma Nair, Addressing Big Data Problem Using Hadoop and Map Reduce, IEEE., 2012, p.1-5.
9. Seema Maitreya, C.K. Jhab, MapReduce: Simplified Data Analysis of Big Data, Published by Elsevier B.V, 2015, p.563-571.
10. Elisa Bertino, Big Data - Opportunities and Challenges. IEEE, 2013, p.479-480.
11. Avita Katal, Mohammad Wazid, R.H.

Goudar, Big Data: Issues, Challenges, Tools and Good Practices. IEEE, 2013, p.1-6.
12. Intel, Bhasker Allene, Better Performance for Big Data, Intel White paper, 2013, p.1-8.
13. A. Khare, Big Data: Magnification Beyond the Relational Database and Data Mining Exigency of Cloud Computing, IEEE, 2014, p.1-6.
14. A. Bechini, F. Marcelloni, A. Segatori, A MapReduce solution for associative classification of big data, Information Sciences, 2016, p.33-55.
15. Dr. Siddaraju, C.L. Sowmya, K. Rashmi, M. Rahul, Efficient Analysis of Big Data Using Map Reduce Framework, IJRDET, 2014, p 64-68.
16. Ajay Chandramouly, How Intel Implemented a Low - Cost Big Data Solution in Five Weeks, Intel White Paper, 2014, p. 1-12.
17. Lorena Siguenza-Guzman, Victor Saquicela, Elina Avila-Ordóñez, Joos Vandewalle, Dirk Cattrysse, Literature Review of Data Mining Applications in Academic Libraries, Elsevier Inc., 2015, p. 1-12.
18. C. Vitolo, Y. Elkha, D. Reusser, C.J.A. Macleod, W. Buytaert, Web technologies for environmental Big Data. Environmental Modelling and Software, 2015, p.185-198.
19. Hakan Özkösea, Emin Sertac Ari, Cevriye Gencerb, Yesterday, Today and Tomorrow of Big Data, Social and Behavioral Sciences, 2015, p.1042-1050.
20 Intel, Gupta, S., Shilpi, Real-Time Big Data Analytics. Intel White Paper, 2014, p.1-8.
21. Intel, Philippe Botteri, Turn Big Data into Big Value, Intel White Paper, 2013, p.1-8.
22. Romany F. Mansour, Understanding how big data leads to social networking vulnerability, Computers in Human Behavior, 2016, p.1-4.
23. Ningyuxin, Liyueling, How We Could Realize Big Data Value, IMSNA, IEEE, 2013, p.425-427.
24. A. Weichselbraun, S. Gindl, A. Scharl, Enriching semantic knowledge bases for opinion mining in big data applications, Knowledge-Based Systems, Elsevier, 2014, p.78-85.
25. F. Tomova, Application of the MicroStrategy BI platform to calculate claim ratio for the insurance companies, XI-th international conference "Challenges in higher education and research in the 21st century", Sozopol, Bulgaria, 2013.
26. F. Tomova, Research on methods of predicting the state of the Peirce-Smith converters for the purposes of predictive maintenance, Journal of Automatics and Informatics, 2, 2014, 14-21.