ISSN 2278-3091

Volume 9, No.1, January – February 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse102912020.pdf

https://doi.org/10.30534/ijatcse/2020/102912020

Big Data Management Challenges

Sabrina Šuman [1], Patrizia Poščić [2], Maja Gligora Marković [3]

[1] Polytechnic of Rijeka, Business department, Croatia, ssuman@veleri.hr

[2] University of Rijeka, Department of Informatics, Croatia, patrizia@inf.uniri.hr

[3] University of Rijeka, Faculty of Medicine, Department of Medical Informatics, Croatia, majagm@medri.uniri.hr

ABSTRACT

The emergence of new data types in big data era implicates the need to analyse and exploit them to gain valuable business insight. Traditional platforms cannot fully meet the analytical needs of the company if support for an unstructured data type is needed. This paper gives an overview and synthesis of areas related to big data technologies, with a series of guidelines for adopting the appropriate software, storage structure, and efficient deployment for big data management. A broad data management context is presented through a conceptual model of business performance management in a modern data management era.

Key words: Big data, Big data management, Hadoop, Big data processing, Data lake

1. INTRODUCTION

The ICT field is extremely dynamic, with the frequent emergence of new technologies. In the last decade there have been major changes in data management with the emergence of new types and sources of data summed up in the big data concept. Big data are complex, layered, large amount of data, and big data technologies are usually considered in comparison to standard 3V: Volume, Variety (of many different data types from different sources), Velocity, and additional 2V: Veracity (Data Quality) and Value (Value for Business). Big data can also be structured, but they are primarily semi-structured and unstructured data types. It is primarily the size and data diversity that creates new analytics approaches [1]. It imposes special ways of retrieving, transforming and preparing, storing and analysing [2] [3]. Utilizing big data technology has unlimited potential for improving both personal life and business competitiveness [4] [5]. At the same time, a large amount of data types make it difficult to find the right values from data [6], and data management in big data is extremely complex. The value of big data lies in the way that information obtained through analytical processing are used (for example to reduce costs, reduce time in business processes and respond to queries, faster and better development of a new product, eliminate failures, errors and failures, better customer relationships, risk assessment and making better quality decisions). Problems related to big data are most commonly related to storage, processing, and management in general [7][8]. There are also issues related to data ownership, privacy and security (see for security issues and algorithms in [9]), quality (a large amount of data from different sources should be readily available for analysis in a short time) and timeliness (a large amount of data, longer analysis, streaming issues analyses) [10]. In big data, there are also new storage structures (e.g. data lakes, distributed file systems, non-relational databases, etc.), needs for new competencies of existing experts and new profiles of experts (big data analyst, big data engineer, big data architect and so on) [3].

In order to maintain and increase the company's ability to use technology for successful decision-making, it is necessary to build an innovative management platform with all new data types, new methods of storage, processing and application of intelligent methods (knowledge and information retrieval, pattern recognition, machine learning, optimization methods, etc.) [11]. As Russom's research [12] shows, employees are aware of changes and issues in big data, and 82% of them (225 of them) believe that data in their company evolves, in terms of diversity in structure, type, sources, the way they are managed and how they are used in business (20% claim to be drastic, while 62% claim to be moderate). In the same survey, one of the major issues highlighted the incompatibility of large data types and structures with relational databases (68% of respondents). This requires a revision of the data management strategy and good information regarding the potentials and disadvantages of big data technology, big data tools, platforms and ways of implementation. The purpose of this paper is to provide arguments and guidelines for the adoption of an appropriate big data management strategy, selection of software tools, storage structure and efficient deployment, through an overview of the field of big data technologies. The aim of the paper is to provide a synthetic overview of the area relevant to the establishment of modern data architecture. After the introduction, a review of concepts important for understanding the wider area of modern data architecture is given - big data technology from the aspect of data types, storage types, analytical processing, and overall big data management strategy. The concept of warehousing modernization, the Hadoop ecosystem and the data lake were discussed in particular. In the results and discussion section, examples of modern data architecture are given as well as the synthetic representation of the processing phase of the big data with the description of each phase with the examples of tools used in phases.

The motivation for this research stems from the current issues of managing new forms of data and concerns all data management phases: from data sources, through purification, analysis, visualization and storage. Also, companies are faced with the problem of choosing the right solutions to manage their data and reviewing possible solutions with a series of guidelines can be beneficial.

2. DATA MANAGEMENT OVERVIEW IN BIG DATA ERA

In this section, we list some traditional and modern elements of data management such as data warehouses, Hadoop framework, Data lake, Spark and Map Reduce

2.1. Data warehouses role in a big data era

Data warehouses are created due to the need to integrate the contents of different databases and other data sources over time and access these data effectively to perform analytical processes. It is a centralized, cleaned and integrated organization of different source data. Today, the following questions can be asked: can data warehouses meet the data management needs in big data? Is data warehouse needed in big data era? Are solutions related to big data management a replacement for data warehouses?

First of all, a big data solution is a technology that involves storage and management of big data, and a data warehouse is architecture. Today, there are cases where some companies may have a big data solution as well as a data warehouse, cases with no big data management solution, but with a data warehouse, and a scenario where the company does not have a data warehouse but has a big data management solution. Each of the scenarios depending on the business activity and business model of the company can or does not have to be successful. While the company needs reliable, consolidated, relevant data for decision-making and management at all levels, it needs also a data warehouse [13]. Yet, the traditional data warehouse provides the basis for reporting and analytics of structured data but probably does not represent the most cost-effective way to store all kinds of data [14]. Data warehouse concept in big data should evolve because it does not solve all issues related to analytics and decision support.

One of the changes in business requirements is the need for analysing unstructured and semi-structured data, performing streaming analytics, network analytics, and other phenomena and needs related to big data. The data are generated in huge quantities, completely incompatible with the relational model, often with lacking ownership, so it is difficult to manage and store them in a rigid structure. Often, a lot of resources were spent on building a warehouse without thinking about data and analytical needs. Also, many implementations were unsuccessful, mostly as a result of bad, rigid designs. More modern and agile design and implementation techniques show greater success, customer satisfaction, and greater return on investment. Such techniques are successful because they allow for flexible upgrades and changes that result from changes in business requirements.

In a response to the need to analyse new types of data in a big data era, many innovative tools and techniques have been developed to store and process this data. The basic concept in developing these tools and techniques was that each company accessed the data in a customized, personalized way that meets the specifics and requirements of the company. Thus, the evolution of data warehousing is needed to adapt and coexist with other analytical solutions that include the use of new data types and new data sources. This does not mean that there will be no need to store and manage structured data in relational structures, but that companies will use different forms of storage, management, and data processing. Using big data and managing them is a completely separate need and additional business potential that cannot replace the need for warehouses and warehousing [13].

2.2. Hadoop framework

The Hadoop framework is the technology that is becoming the standard when it comes to storing and processing large amounts of different data (big data). The Hadoop platform is allows performing the tasks on multiple computers (Cluster). It is optimized for processing large amounts of data. Hadoop architecture consists of:

• Hadoop Common Package – contains Java Archive Files (JARs) and scripts required to run and manage all Hadoop modules;

• MapReduce processing mechanism - a system based on YARN for parallel processing of large data sets;

• Hadoop distributed file system (HDFS) - a distributed, scalable and portable file system which stores large amounts of data (GB and TB) on multiple computers;

• YARN - a central platform for managing operations, security and data through Hadoop clusters [15], [16].

Hadoop is a platform that offers a solution for many limitations of past technologies, such as storage constraints and large volume data processing capabilities [17]. It supports multiple data formats – i.e. structured, semi-structured and unstructured data. It is the open source software, with low implementation costs, and low learning curve. Using NoSQL solutions, such as Apache Cassandra, helps eliminate performance issues, costs, and the availability of big data applications. To manage such data, it is convenient to use Apache Hive, data storage software built on Hadoop. It allows reading, writing and managing large databases stored in distributed storage with SQL support [3].

The Hadoop ecosystem provides a variety of open source and commercial technologies for processing fast, interactive BI queries, data retrieval and research, and sophisticated analytical processes such as testing predictive models [18]. Today, technologies enable users to run queries and analysis within the Hadoop cluster and/or clouds without having to move data to a data warehouse, data marts, or stand-alone BI server [18]. The Hadoop cluster consists of many parallel machines where large data sets are stored and processed. Client computers send tasks to this cloud of computers and get results. Storage and data processing take place within this "cloud" machine. Different users can send computerized tasks to Hadoop from individual clients (their own machines at remote locations from the Hadoop cluster). The linear scalability offered by Hadoop clusters with flexible cloud scalability storage can enable organizations to be agile and flexible in expanding computer power in response to immediate BI analytic needs. Companies can also improve security management - use security procedures that are set on sources, as the preparation and processing takes place where the data is located [18].

2.3. MapReduce

MapReduce is the script frame for applications handling huge amounts of both structured and unstructured data stored in the Hadoop Distributed File System (HDFS). Each MapReduce works in two phases: the map phase which maps the input data to key/value sets and reduce phase which takes key/value pairs and provides a desired output based on applying its own algorithms [19]. MapReduce algorithm can be written in many languages (Java, C ++, or Python) and its tasks are easy to run. MapReduce can handle petabytes of data from HDFS on one cluster at maximum processing speeds. MapReduce takes care of the failures and allows retrieving the redundant copy of the processing data. It moves the processing to data in HDFS, and not vice versa. Processing tasks may appear on a physical node where data is found, significantly contributing to Hadoop's processing speed [20].

2.4. Spark

Apache Spark is a distributed or clustered open source computing platform and a big data processing framework. Apache Spark is a framework for fast processing of a large amount of data that can be used generally for all types of data processing (batch processing, interactive analysis, stream processing, machine learning and graph computing) [21]. Provides a fast-paced computing environment for general purpose and sophisticated analytical capabilities that enable the development of analytic applications written in Java, Scala, Python or R [21]. Spark can run independently, on the Hadoop cluster or in the Mesos environment. Spark can connect at run time on HDFS, Amazon S3, Cassandra, and Apache HBase.

Apache Spark stands out at its speed due to the ability to perform in-memory processing. Spark is more effective for disk processing applications. If linear data processing of large data sets is used, the advantage should be given to Hadoop MapReduce while Spark provides fast performance, real-time analytics, graphing, machine learning and other. In many cases, Spark can surpass Hadoop MapReduce. Also, Spark is fully compatible with the Hadoop Ecosystem [22]. It is suitable to be used by a data scientist or a statistician without or limited knowledge of cluster computing. It is usually enabled by interactive shells similar to those such as MATLAB or R. It is particularly suitable for interactive data mining of large data sets in clusters [23].

2.5. Data lake

It is the central repository for all organization data (without predefined data schema). The goal is to collect all the data before they are potentially lost. It provides data consolidation and a highly customizable analytical approach. The lake consists of a distributed, scalable file system (HDFS or Amazon S3), and one or more dedicated processing and query tools such as Apache Spark, Drill, Impala or Presto [24]. Upgrading the existing data warehouse with a data lake (creating hybrid architecture) is a positive change for businesses. Advantages of increasing data warehouse include:

- Great savings in storage costs - scaling architectures (e.g. Hadoop, AWS S3) can store non-processed data in any format at a much lower cost than data warehouses.
- Significantly accelerating processing - flexible data lake architecture enables faster data loading and parallel processing, resulting in faster instant analytical insight.
- Maximizing efficiency - spending less time on low-value business activities (ETL, for example) and making better use of resources for strategic goals of high business value.
- Getting more valuable business insights, faster and of a larger amount of data (lower storage costs allow you to store more data, leading to more accurate trends, better forecasts, etc.) [25].

Comparison of the classical data warehouse and data lake, based on research data from [25][26] is given in Table 1.

Table 1: Comparison of a classic data warehouse and a data lake.

Data warehouse	**comparison of**	**Data lake**
structured, processed, data preparation requires IT department assistance, administrator permissions etc.	data and data preparation	structured, unstructured, in its source (raw) form, provide self-service ad-hoc transformation without administrator permissions
the data were processed before being entered in the data warehouse	processing	the data is in the original format and is processed as needed
on large databases, very expensive for large amounts of data	storage	designed for a large amount of data at a low cost
fixed configuration	flexibility	flexible, possible reconfiguration as needed
mature and steady	security	in development
business professionals	for whom it is intended?	data scientists, data scientists, data engineers
moderate scaling but with high-cost	scaling	large scaling at a low cost
efficiently utilizes storage and processing capabilities but with high cost	cost / efficiency	efficiently utilizes storage and processing capabilities at a low cost
easily manage the quality and security of data	data management	requires an approach to create metadata for raising quality, security and privacy

Russom’s research [12] shows that out of 225 respondents, 90% of them know the concept of data lakes, 24% think that Hadoop data lake has a really high impact on the success of data management strategies in their company, 32% consider the impact is moderate, while 44% of respondents do not even consider this question at all. The users consider the following items (in terms of consequences) as the ones that would benefit the most of data lake implementation (Table 2).

Table 2 data lake's implementation positive impact [12]

Advanced Analytics (data mining, Machine Learning, Complex SQL)	49%
Data Exploration and Knowledge Discovery	49%
Sources of big data for analytics	45%
Data warehouse widening	39%
Data retention for storage	36%
Reducing Cost of Data Storage	34%
A possibility of using the data also for nontechnical user types	24%
To accept unstructured data	21%

As barriers to the implementation of Hadoop data lakes, users have often reported a series of reasons related to data management, security and lack of knowledge and skills related to Hadoop and big data technology.

2.6. Other elements of a modern data architecture

Analytics / Sandbox Environments - This environment almost completely contradicts the easy-to-manage BI / DW environment of a predictable workload that supports classical managerial reporting of a type "what's happened" with business questions. It represents a research environment that have very unpredictable burden and usage patterns. In this environment data scientists should have the freedom of experiment with new data types from different sources, new methods of data transformations, and new analytical models to get new valuable insights from data and build predictable business models. It's too manageable and allows data scientists to use any tools that prefer research, analysis, and analytical modelling [27].

Data Lab - Big data and advanced analytics require different technologies and approaches. The analysis may require data that are not available in the warehouse. Models can be CPU-intensive and create problems for other applications at the same time. There may be conflicts between a warehouse administrator who wants a carefully controlled environment and analysts, especially data scientists who want maximum flexibility. That is why data labs can be created where users can make changes, unlike in original storage tables. The Data Warehouse Administrator creates certain lab owners for specific areas, such as marketing and sales. Each lab owner identifies people who have access to a data lab. The Data Store Administrator in collaboration with the users determines how much the workspace is allocated to each lab and sets the expiration date [27].

3. RESULTS AND DISCUSSION

Based on changes and challenges reported in previous section in area related to modern data management, a visualised and summarized overview and a more detailed insight into the modern data architecture is given in Figure 1. It starts with the general data management cycle (with some modern architecture elements added). Then the possible benefits from introducing modern data platforms' elements, a number of possible strategies and implementation solutions are discussed. At the end of this chapter, detailed big data processing phases are described in detail, along with descriptions of basic processes, and a description of software support, needed for their realization.

3.1. Data management context in big data era

In order to show a wider context related to data management in big data era, a synthetic graphical overview of the performance management cycle of an enterprise is given. A conceptual model, i.e. a framework (Figure 1), has been developed that specifies the components, that is, the areas that participate in the efficient business of the company.

Model is separated into three interconnected areas or subsystems. A subsystem representing internal or external data and/or systems generating such data. A subsystem for "preparing" and storing data (cleaning, consolidating, structuring, aggregating, storing) to serve the company's analytical needs as efficiently as possible, and an analytic subsystem where data is "exploited" for the implementation of previously defined goals. In the last subsystem, there are a number of activities in which a whole spectrum of different tools is used to extract the potential of different types of data.

The need for continuity of running such performance management cycles is stressed, so that the company moderates goals, KPIs, and selects the tools that deliver the best results. Storage methods and analytical tools and methods used are divided into traditional BIs (storage in data warehouses and analytical processes mostly over structured data) and on big data typical storage and analytical processes (data lakes and sandboxes). For each analytical category there is a full range of techniques and tools [28] [29].

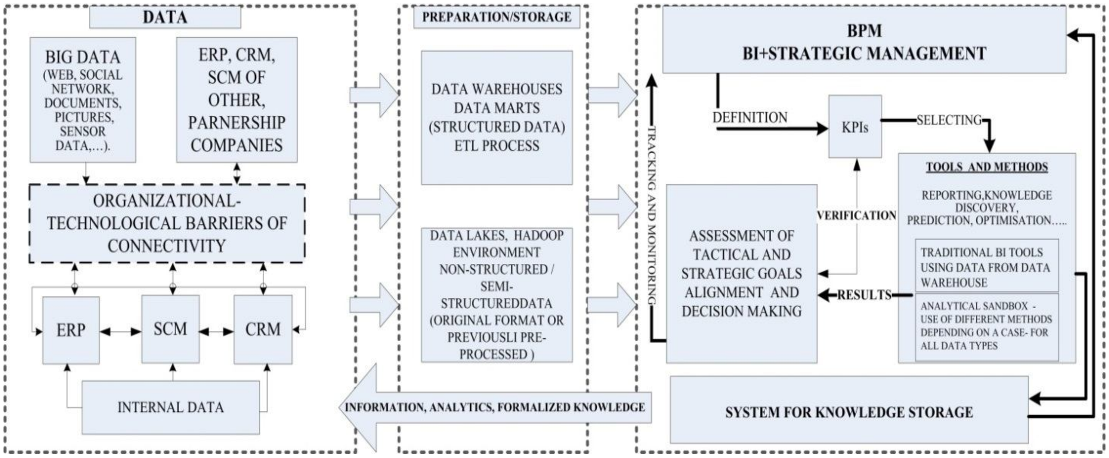

DATA
BIG DATA (WEB, SOCIAL NETWORK, DOCUMENTS, PICTURES, SENSOR DATA,...).
ERP, CRM, SCM OF OTHER, PARNERSHIP COMPANIES
ORGANIZATIONAL-TECHNOLOGICAL BARRIERS OF CONNECTIVITY
ERP
SCM
CRM
INTERNAL DATA
PREPARATION/STORAGE
DATA WAREHOUSES DATA MARTS (STRUCTURED DATA) ETL PROCESS
DATA LAKES, HADOOP ENVIRONMENT NON-STRUCTURED / SEMI-STRUCTUREDDATA (ORIGINAL FORMAT OR PREVIOUSLI PRE-PROCESSED)
INFORMATION, ANALYTICS, FORMALIZED KNOWLEDGE
BPM
BI+STRATEGIC MANAGEMENT
DEFINITION
KPIs
SELECTING
TRACKING AND MONITORING
TOOLS AND METHODS
REPORTING,KNOWLEDGE DISCOVERY, PREDICTION, OPTIMISATION.....
TRADITIONAL BI TOOLS USING DATA FROM DATA WAREHOUSE
ANALYTICAL SANDBOX - USE OF DIFFERENT METHODS DEPENDING ON A CASE- FOR ALL DATA TYPES
VERIFICATION
RESULTS
ASSESSMENT OF TACTICAL AND STRATEGIC GOALS ALIGNMENT AND DECISION MAKING
SYSTEM FOR KNOWLEDGE STORAGE

Figure 1: Business Performance Management Cycle

3.2. Synthesis of big data processing phases

Big data Processing can be seen as six dependent phases. First, data are generated in different applications and systems (internal and external data in different formats and structures).

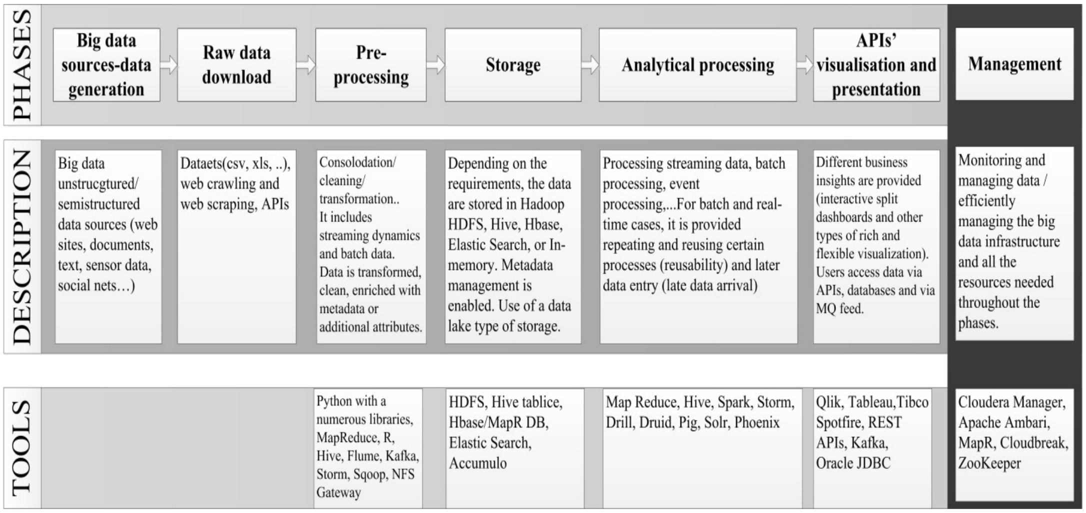
The second phase includes all steps to download generated data from different sources (web scraping, web crawling, APIs ...).

The third phase involves cleaning and converting data from multiple sources processing. Today, data management systems are expected to be able to process data in real-time (streaming data) and batch-aggregate processing. This means that the dynamics must also adjust the processes of preparation, cleaning, and other data transformation actions.

The fourth phase is storage - usually, all data types are permanently stored in some type of the file system or database, or they combine local storage with cloud storage services.

In the fifth phase, different tools, methods and techniques, for analysing and using data are used to obtain information important for supporting business activities and making decisions. While the data warehouse is used to process and analyse structured data, the Hadoop cluster is used for processing and transforming unstructured and semi-structured into structured data. For analytical processes, the Hadoop ecosystem has multiple extensions for queries, data processing, storage in NoSQL databases (e.g. HBase), data warehouses (e.g. Hive) and advanced data mining and machine learning algorithms.

The Sixth Phase - the information and results obtained from analysis phase should be visually presented, assigned, distributed and presented to its users in the final phase [30]. All the phases are also synthesized visually (based on data from [31], [32] and [33]) on a Figure 2.



PHASES	Big data sources-data generation	Raw data download	Pre-processing	Storage	Analytical processing	APIs' visualisation and presentation	Management
DESCRIPTION	Big data unstrucgtured/ semistructured data sources (web sites, documents, text, sensor data, social nets…)	Dataets(csv, xls, ..), web crawling and web scraping, APIs	Consolodation/ cleaning/ transformation.. It includes streaming dynamics and batch data. Data is transformed, clean, enriched with metadata or additional attributes.	Depending on the requirements, the data are stored in Hadoop HDFS, Hive, Hbase, Elastic Search, or In-memory. Metadata management is enabled. Use of a data lake type of storage.	Processing streaming data, batch processing, event processing,...For batch and real-time cases, it is provided repeating and reusing certain processes (reusability) and later data entry (late data arrival)	Different business insights are provided (interactive split dashboards and other types of rich and flexible visualization). Users access data via APIs, databases and via MQ feed.	Monitoring and managing data / efficiently managing the big data infrastructure and all the resources needed throughout the phases.
TOOLS			Python with a numerous libraries, MapReduce, R, Hive, Flume, Kafka, Storm, Sqoop, NFS Gateway	HDFS, Hive tablice, Hbase/MapR DB, Elastic Search, Accumulo	Map Reduce, Hive, Spark, Storm, Drill, Druid, Pig, Solr, Phoenix	Qlik, Tableau,Tibco Spotfire, REST APIs, Kafka, Oracle JDBC	Cloudera Manager, Apache Ambari, MapR, Cloudbreak, ZooKeeper

Figure 2: Synthesis of the big data processing phases

4. CONCLUSION

Traditional data management architectures cannot meet the current needs of companies for integrating and analysing a wide range of data types generated from a variety of sources. The modern data platform enables analytical processing of both historical data and in real time, and for the structured, semi-structured and unstructured data, clouded or locally stored. The new big data technologies are complementary to existing data management technologies and serve to manage, process, and analyse new types and forms of data that are not supported in standard BI / DW systems. Therefore, new data management platforms complement and optimize the existing ones. This paper presents the context of data management in big data era, provides a review of new technologies, concepts, platforms in modern data management architecture.

Companies should realize the benefits and problems of using big data technology, but above all, should understand their needs to successfully implement those new technologies that help them achieve successful business goals. Whereas much of the big data is stored using the Hadoop File System (HDFS) and on distributed computing platforms that support Hadoop clusters, most companies need to be thoroughly informed about the technologies within the Hadoop ecosystems related to all the above-mentioned big data processing phases. In terms of assistance and aid during that process, an overview of some of the possible solutions is provided in Figure 2 and it provides a general and broad overview of the big data technology. In order to create an optimal strategy and appropriate software selection, a skilled team is needed with a spectrum of knowledge and competencies in data management, a modern data architecture, big data technologies, and also business domain experts. The following research activities are directed toward identifying new employees' profiles and knowledge, competencies and capabilities requirements. Subsequent research activities will aim at adjusting the study programs in higher education structures in order to meet the companies' demands for management and decision making in large data and IoT era.

REFERENCES

1. I. A. Atoum and N. A. Al-Jarallah, **Big data analytics for value-based care: Challenges and opportunities**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 6, pp. 3012–3016, 2019. https://doi.org/10.30534/ijatcse/2019/55862019
2. I. P. Popchev and D. A. Orozova, **Towards Big Data Analytics in the e-Learning Space**, *Cybern. Inf. Technol.*, vol. 19, no. 3, pp. 16–24, 2019.
3. J. Campos, P. Sharma, U. Gorostegui Gabiria, E. Jantunen, and D. Baglee, **A big data analytical architecture for the Asset Management**, *Procedia CIRP 64*, pp. 369 – 374, 2017. https://doi.org/10.1016/j.procir.2017.03.019
4. J. Morris, **Top 10 categories for Big Data sources and mining technologies**, 2012. [Online]. Available: https://www.zdnet.com/article/top-10-categories-for-big-data-sources-and-mining-technologies/.
5. V. Mayer-Schönberger and K. Cukier, *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt, 2013.
6. B. Butler, **Cloud Cronicles**, 2015. [Online]. Available: http://www.networkworld.com/article/2973963/big-data-business-intelligence/5-problems-with-big-data.html.
7. D. Ahamad, M. Akhtar, and S. A. Hameed, **A review and analysis of big data and mapreduce**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1, pp. 1–3, 2019.
8. C. Ji, Y. Li, W. Qiu, U. Awada, and K. Li, **Big data processing in cloud computing environments**, in *International Symposium on Pervasive Systems, Algorithms and Networks*, 2012.

9. P. Amarendra Reddy and O. Ramesh, **Security mechanisms leveraged to overcome the effects of big data characteristics**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 2, pp. 312–316, 2019.
10. A. Al-Drees, R. Bin-Hezam, R. Al-Muwayshir, and W. Haddoush, **Unified Retrieval Model of Big Data**, in *Advances in Big Data Proceedings of the 2nd INNS Conference on Big Data, October 23–25,* 2016.
11. D. Zhu, Y. Zhang, X. Wang, and E. Al., **Research on the methodology of technology innovation management with big data**, *Sci. Sci. Manag. S. T.*, vol. 4, pp. 172–180, 2013.
12. P. Russom, **Data Lakes Purposes, Practices, Patterns, and Platforms**, 2017.
13. B. Inmon, **Big Data Implementation vs. Data Warehousing**, 2013. [Online]. Available: http://www.b-eye-network.com/view/17017.
14. C. Russell, **Database Development with IBM Hybrid Data Architecture**, 2017. [Online]. Available: http://www.ibm.com/developerworks/.
15. D. Marjanović, **Hadoop i analitika u realnom vremenu**, 2017. [Online]. Available: http://www.datascience.rs/Hadoop-i-analitika-u-realno-vremenu/.
16. V. Dagade, M. Lagali, S. Avadhani, and P. Kalekar, **Big Data Weather Analytics Using Hadoop**, *Int. J. Emerg. Technol. Comput. Sci. Electron.*, vol. 14, no. 2, 2015.
17. N. Garg, S. Singla, and S. Jangra, **Challenges and Techniques for Testing of Big Data**, *Procedia Comput. Sci.*, vol. 85, pp. 940–948, 2016.
18. D. Stodder, **New Strategies for Visual Big Data Analytics, How organizations can apply modern data platform technologies and practices to support analytics innovation**, 2017.
19. S. Alapati, *Expert Hadoop Administration Managing Tuning, and Securing Spark,Yarn and HDFS*. Addison-Wesley, pp 24-25, 2017.
20. **MapReduce**, 2019. [Online]. Available: https://www.ibm.com/analytics/hadoop/mapreduce.
21. R. Alapati, S., *Expert hadoop administration managing, tuning, and securing Spark, Yarn,and HDFS*. Addison-Wesley, pp 149-151,2017.
22. A. Bekker, **Spark vs. Hadoop MapReduce: Which big data framework to choose**, 2017. [Online]. Available: https://www.scnsoft.com/blog/spark-vs-hadoop-mapreduce.
23. T. Oktay and A. Sayar, **Analyzing Big Security Logs in Cluster with Apache Spark**, in *Advances in Big Data, Advances in Intelligent Systems and Computing*, Angelov et al., Ed. 2016. https://doi.org/10.1007/978-3-319-47898-2_14
24. J. Caserta and E. Cordo, **Data Warehousing in the Era of Big Data**, 2016. [Online]. Available: http://www.dbta.com/BigDataQuarterly/Articles/Data-Warehousing-in-the-Era-of-Big-Data-108590.aspx.
25. B. Sharma, *Architecting Data Lakes: Data Management Architectures for Advanced Business Use Cases*. O'Reilly Media, 2018.
26. M. Knight, **Data Warehouse vs. Data Lake Technology: Different Approaches to Managing Data**, 2017. [Online]. Available: https://www.dataversity.net/data-warehouse-vs-data-lake-technology-different-approaches-managing-data/.
27. J. Watson,H., **Data Lakes, Data Labs, and Sandboxes**, *Bus. Intell. J.*, vol. 20, no. 1, 2015.
28. S. Šuman and I. Pogarčić, **Development of ERP and other large business systems in the context of new trends and Technologies**, in *27th Daaam International Symposium On Intelligent Manufacturing And Automation*, 2016, pp. 319–327.
29. S. Šuman, *Sustavi poslovne inteligencije - teorija i riješeni primjeri*. Rijeka: Veleučilište U Rijeci, 2017.
30. Heilig,L. and S. Voß, **Managing Cloud-Based Big Data Platforms: A Reference Architecture and Cost Perspective**, in *Big Data Management*, B. García Márquez,F.,P., Lev, Ed. Springer International Publishing AG, p.29, 2017.
31. W. El Kaim, **Big Data Architecture**, 2016. [Online]. Available:https://www.slideshare.net/welkaim/big-data-architecture-part-2.
32. K. Singh, **Top 10 Big Data Tools in 2019**, 2019. [Online]. Available: https://dimensionless.in/top-10-big-data-tools-in-2019/.
33. G. Ginde, R. Aedula, S. Saha, A. Mathur, S. Roy Dey, S. Sampatrao, G., and D. Sagar, **Big Data Acquisition, Preparation, and Analysis Using Apache Software Foundation Tools**, in *Big Data Analytics Tools and Technology for Effective Planning*, D. Somani, A.K., Ganesh, C., Ed. Boca Raton: CRC Press Taylor & Francis Group, 2018. https://doi.org/10.1201/b21822-9