International Conference on Cyber Security and Computer Science (ICONCS'18),
Oct 18-20, 2018 Safranbolu, Turkey

Big Data Applications: A Review

Y. GÜLTEPE[1]

[1] Kastamonu University, Kastamonu/Turkey, yasemingultepe@kastamonu.edu.tr

***Abstract* – In recent years, the growing and large amounts of data, which have been associated with the widespread use of social media, smart devices and internet, define big data. With big data; The vast majority of things that were formerly never measured, stored, analyzed, or shared, were converted into processed and usable data. The big data typically describes both the type of managed data and the technology used to collect and operate it. Data can be transformed into information that can only have a value, but if without the wisdom, information can be allowed to really useful to people. Nowadays the big data attract attention with such qualities for his volume, speed and variety. With the increased use of big data, a major breakthrough in productivity, profitability and innovation in different sectors is expected. Examples are many successful applications of big data in different areas of the world; Public sector, health, insurance, banking, education, etc. Big data can help improve productivity, profitability, performance and reduce data exhaustion, etc. Education, health, banking, retail sales, government resources, defense industry, production and energy sectors, as well as facilitating human life will increase the efficiency of institutions and will constitute the infrastructure of further progress towards the future. In the study, big data was handled conceptually, relations with many concepts, big data technologies and methods used for big data processing were introduced and different examples were given about usage areas of big data in the world.**

***Keywords* - Big Data, Data Analysis, Internet, Data Extraction.**

I. INTRODUCTION

Big data is the process of producing point of view for decision making. There are a lot of data stored in one place; they will be analysed to develop insight, intuition or insight to produce decision(s) in the area of interest. The really big data process uses human resource and technology to quickly analyze various types of data from various sources. The aim is to obtain information that will lead to new activities and actions in the field of business, for example. These data may be raw or processed in the form of pictures, video, e-mail, account transactions, social media messages [1].

The big data phenomenon has two main component. The first of these is the accumulation and storage of big quantity of data in areas of interest. The second axis includes analysis of data (large stacks of data). After this; it is time to interpret, evaluate, better manage and make decisions using the right analysis methods.

"Big Data" consists of large volumes of information from internet hosts, internet, mobile and social media usage statistics, social media content, webblogs, media sensors and same devices. Big data enables businesses to take formative judgments correctly, manage their risks better, and innovate their business and products when they are interpreted by analysis methods.

The ability to analyze big data from multiple and varied sources will make deep and complex knowledge meaningful and will benefit its users in the decision-making and implementation operation. In the use of big data, "semantic" is an indispensable element in the case of a concept such as "really knowing" instead of "guessing".

The purpose of this paper is to describe big data, big data technologies and procedures used for big data analytics. It is also explained with examples of which areas the big data is used.

II. THE ANALYTICS OF THE BIG DATA

Although traditional analytics is used to find answers to predetermined questions, applying it to big data allows it to search for information to see which information can be derived from it and to identify unexpected or previously unknown connections and relationships. As a conclusion of the analytics of big data, preferable strategies and decision-making are feasible for the exercises. Analytical findings more effective marketing services can provide competitive advantages over new business occasions, better customer employments, improved transactional profitability, structured organizations, and other business benefits [2].

In order to help in the identification of big data, characteristics such as volume, speed and diversity of the data are used. It has become a tradition to use the initials of the english words that show such characteristics. Big data were originally characterized by a five-component character and were represented by 5V [3]. These were Volume, Variety, Velocity, Veracity, and Value. At present, new "V" letters such as Validity, Variability, Venue, Vocabulary and Vagueness which are defined in addition to this definition have been added [4, 5].

Volume: The dimension of the data, plays a very special role for identifying the significance of the data. At the same time, a particular data can intrinsically be considered as a big data or not, depending on the capacity of the data.

Velocity: Large capacity of continuous fluid data coming and the production of old data quickly processed and output of the data set is very dynamic causes. The flow of information is at a unique speed and the process takes place on time.

Variety: Managing data in different structures, diversity and complexity. E-mails, voice recordings, social media and

blog, as well as data that is not kept in classical database management systems.

Veracity: Veracity shows how accurate or reliable the big data is. Data must be reliable enough to be used in business decisions. The high diversity of big data makes it difficult to confirm the attribute and reliability of the analyzed data.

Value: The most significant component of big data is the value creation. After the data generation and processing layers of big data, it needs to have a positive value for the organization; it needs to have immediate impact on decision-making processes; it needs to be ready to make the right decision; and it needs to be immediately at hand.

Due to the above-mentioned characteristics of big data, it is necessary to use more enriched methods instead of traditional data management systems. Because not only Volume, but also in terms of features such as Veriety, Velocity, Variability and both structural and semi-structural or non-structural form of the data continues to be produced at any moment. High computational power is needed to process, work on, and analyze these data. For this reason, instead of traditional computing approaches, computer clusters and distributed file systems; platforms with open-source software frameworks such as Hadoop, Spark, Strom, Flink, Samza are becoming popular instead of traditional programming and programming languages. Big data frameworks developed on the cloud can be listed as follows: Google Compute Engine, AWS EMR, Pure System, LexisNexis HPCC Systems and Microsoft Azure.

Big data frameworks are classified according to the following characteristics: programming model, supported programming languages, types of data sources, allowing repetitive data processing, existing learning books and framework compatibility and fault tolerance strategy [6].

Hadoop is an open source software framework written in Java that enables us to work in parallel on multiple machines with large data sets. Hadoop offers two elements: Hadoop Distributed File System (HDFS) and MapReduce. HDFS, which provides access to huge volume of data with high throughput. It makes it look like a single file system by connecting file systems on many machines. MapReduce is a system that allows easy analysis of very large data on distributed architecture. Instead of using highly equipped servers to process very big data, the same process is performed much more effectively with the help of a set of hardware from MapReduce [7].

III. Big Data and Usage Areas

After the definition of big data, the lifecycle of big data can be mentioned. There are different big data lifecyle operation according to solution needs. The Big Data Integrator (BDI) Platform [8] was developed to process big data. The Big Data Integrator Stack Lifecycle (BDI SL) methodology provides a new, simple way to create, deploy and maintain Big Data applications. BDI SL generally comprises of the following steps: development, packaging, composition, development, deployment and monitoring. Alshboul et al. (2015) are presented a security thread model for big data ve big data security threats and attacks are explained in term of big data life cycle [9]. Big data lifecycle consists of four stages: Data collection, data storage, data analytics and knowledge creation. Demchenko et al. (2014) have proposed big data lifecycle. This lifecycle includes four phases: a) Data collection and recording b) Data filter/enhance, classification c) Data analytics, modeling, prediction d) Data delivery and image [10]. Hadoop systems use a life cycle consisting of the following steps to manage big data sets: Create, capture, curation, process, transfer, store, analysis and visualization [7].

Examples are many successful applications of big data in different areas of the world; Public sector, health, insurance, banking, education, etc. Big data can help improve productivity, profitability, performance and reduce data exhaustion, etc. Here are a few examples of how big data affects different sectors:

Business: Customer personalization, determining the causes of customer loss, optimization of distribution and logistics.

Banking: With big data technologies, all banking transactions are manageable, easy and fast, advantages, efficiencies, achievements and improvements in the internal processes of the banks [11].

Technology: Decreasing process time, real-time analysis, generating rapid response during crisis periods, decision making with automatic systems to reduce risks.

Healthcare: Disease detection, follow-up and personal DNA analysis to strengthen health. The major data sources in health services are grouped as follows: Machine-generated data, biometric data, human data, process data, behavior data, epidemiological data, published data, and data on current life that can be associated with health from daily life [12].

Retail Sales: Retailers should be aware of the methods of accessing customers and the market, and the process of implementing appropriate transactions in the most effective manner. Big data is at the heart of all this.

Personal Location Data: Big data analysis platforms enable you to obtain, store and analyze both structured and unstructured data. This makes it possible to establish the basis for the strategy on which digital advertising activities can be based.

Smart Cities: Cities produce high volume, speed, and variety of data that fits the big data definition. Therefore, the role of big data and related methods and technologies in creating an effective, sustainable and intelligent city is important [13].

Education: Big data in the education sector; improvement of student performance, planning of curriculum of education, improvement of inefficient administrative processes in education, restructuring of course contents, follow-up of student performance of instructors and administrators, etc. it is used for many purposes [14].

With the increased use of big data, a major breakthrough in productivity, profitability and innovation in different sectors is expected. This breakthrough will be developed with the help

of large data sources in customer analysis, supply chain management, quality control management, risk management, performance management and corruption. In these areas, the demand for the use of big data will increase and the demand for competent staff for data processing, analysis, interpretation and communication will increase.

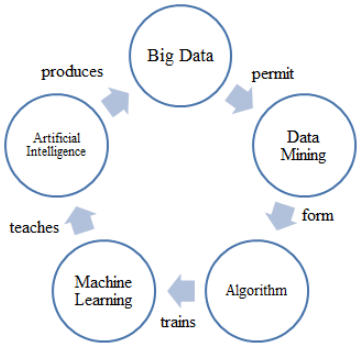

Figure 1: Big Data's Technology Cycle [15].

Big data sets a complex cycle with data mining, algorithms, machine learning and artificial intelligence. Big data are contributing to a technology cycle. The cycle is shown in Figure 1. Big data applications can be grouped about is done at what stage of the technology cycle.

Big data reveals a great change for life sciences. Life scientific provides a range of solutions for semantic web technologies to address the heterogeneous diversity of big data. RDF (Resource Description Framework), SPARQL (an RDF query language), RDF store and ontology facilitate the integration and analysis of heterogeneous multidisciplinary data. Linked data turns the web into a large global database. The RDF store in the cloud takes full advantage of cloud services to address the exponential growth of biological data. Cloud-based analytical applications for big data storage provide companies with significant cost savings. The entire scientific community is trying to develop new technologies and tools to ensure that Big Data's life sciences area is accessible, analytical, and feasible [16].

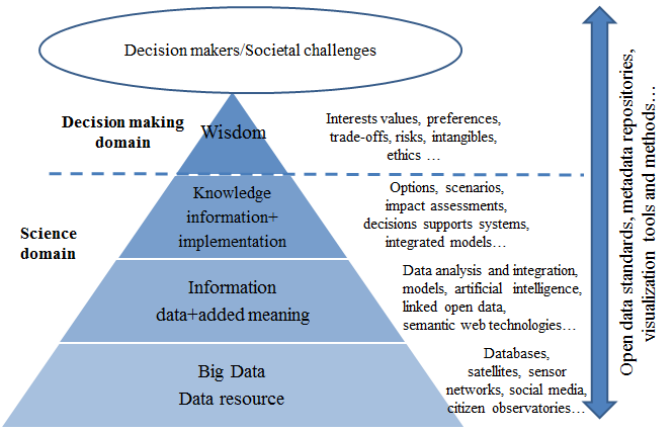

Figure 2: DIKW From Big Data to decision making for societal challenges [17].

Data can be transformed into information that can only have a value, but if without the wisdom, information can be allowed to really useful to people. Nowadays the big data attract attention with such qualities for his volume, speed and variety. But in practice; the values plays a more important role. In some cases, small data may also be of great value to query the value of data that is not required for big data. Therefore, data science is valued more than the internal value of the data. Rowley (2007) is shown DIKW (Data-Information-Knowledge-Wisdom) model in Figure 2. Model is used to contextualize data, information, knowledge and wisdom according to each other and define the processes involved in the ralization of a transformation of an entity at a lower level in the hierarchy to an entity at a highest entity in the hierarchy.

IV. Big Data Solutions

It is inevitable to have difficulties in addition to the advantages provided by the big deal. As can be understood from the definition and properties of the big data, this data differs from the data types commonly used in analyzes. The collection, storage, sharing, transferring, visualization and analysis of such data are confronted as one of the most important difficulties due to the characteristics of the big data.

The most frequent variability with big data use is being developed with a new approach that will be successful in the face of many problems such as noise accumulation, false regression, and internalization. In short, researchers are suggested that new statistical considerations and computer methods be used together with machine learning [18] to obtain desired information and to make predictions.

Amazon Web Services (AWS) offers a wide range of services to help you quickly and easily build and deploy big data analytics applications [19]. AWS enables you to quickly scale almost all major data applications, including data warehouse, clickstream analytics, fraud detection, suggestion infrastructures, event-oriented ETL, serverless processing and internet processing of objects by providing fast access to flexible and cost-effective it resources.

This newly formulation of big data applications involving both interior and exterior big data necessitates new models and methods to implement cognitive modeling steps. Merino et al. (2016) proposed the 3C model [20]. To interpret the quality of use of large data sets, it consists of three sub data quality dimensions. These; Combinatory consistency, functional consistency, and chronic consistency. The quality of the data is not the model on the quality of use of big data. The intersection of conceptual modeling and big data terms appears to be one of the difficulties in big data. Volume for data is quite large, diversity is quite high, speed is very fast and veracity is quite uncertain [1].

Bukhari et al. (2018) presented several examples of the use of semantic web and big data technologies for insurance industry and social network analysis. Interfaces and access protocols have recently been used for distributed processing. Steep learning curves, the proliferation of non-standard resources, scarcity of specialized researchers, a few reasons for

the wider acceptance of the semantic web for big data. Other technical challenges for the wider acceptance of semantic web for big data include reasoning on performance optimization of large-scale data and logical data-driven systems. Nevertheless, new trends such as FAIR data and Blockchain technologies make the general big data and the semantic web interesting and challenging at the same time [21].

Gil and Song (2016) presented big data challenges. These are collected under the following headings: Data capture, cleaning and storage, data consolidation, collection, query processing, data modeling, analysis and intepretation, and envision. As a solution, People's Ontology have created. This ontology is used as a database of search mechanisms and classes, some mechanisms such as web semantics, rescue binary associations, attribute correlations and synonyms [22].

IT managers face different barriers to implementing big data solutions. Alharthi et al. (2017) stated that the infrastructure preparation, confusion, lack of skills, confidentiality, cultural obstacles as five different sub-barriers [23]. In Sivarajah et al. (2017) that data challenges (volume, velocity, variety, variability, veracity, visualization and value), the process challenges (data acquisition&warehousing, data mining&cleaning) and management challenges (privacy, security, data governance, data&information sharing, cost/operational expenditure, data ownership) as three sub-headings are grouped under [24]. On the other hand, Ostrowski et al. (2016) proposed challenges in big data integration as follows: incomplete data, scalability of semantic web tools, lack of industrial ontologies, new applications, incompatible data, support for realtime streaming, parallelization of big data tools [25].

V. Conclusion

Big data has an important role in analyzing, agreement, defining certain patterns and trends, making strategic plans for companies future, solving problems effectively and improving their products and services according to their needs and preferences.

With the increased use of big data, a major breakthrough in productivity, profitability and innovation in different sectors is expected. This breakthrough will be developed with the help of big data sources in customer analysis, supply chain management, quality control management, risk management, performance management and corruption. In these areas, the demand for the use of big data will increase and the demand for competent staff for data processing, analysis, interpretation and communication will increase. Technology and methods associated with big data such as internet of objects, cloud computing, machine learning and data mining; AR-GE and innovativeness are triggered, and new products and services are produced. Therefore, big data will be of considerable benefit in integrating it into different information systems as a whole with its methods and technologies.

References

[1] S. V. Mayer, and K. Cukier, *Büyük Veri – Yaşama, Çalışma ve Düşünme Şeklimizi Dönüştürecek Bir Devrim*. Çev. Banu Erol. İstanbul Paloma, 2013.

[2] K. Das, and P. M. Kumar, "Big data analytics: A framework for unstructured data analysis", *International Journal of Engineering and Technology*, vol. 5, no. 1, pp.153-156, 2013.

[3] A. F. Mohammed, V. T. Humbe, and S. S. Chowhan, "A review of big data environment and its related Technologies", *ICICES 2016*, pp. 1-5.

[4] M. A. Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value". *2014 IEEE Conference of the American Society for Engineering Education.*

[5] G. Firican. (2017). The 10 Vs of Big Data. Available: https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

[6] W. Inoubli, S. Aridhi, H. Mezni, and A. Jung, "An Experimental Survey on Big Data Frameworks", *Future Generation Computer Systems*, vol. 86, pp. 546-564, September 2018.

[7] R. Posa, Introduction to Hadoop, Big Data Life-Cycle Management. Available: https://www.journaldev.com/8795/introduction-to-hadoop

[8] I. Ermilov, A. N. Ngomo, A. Versteden, H. Jabeen, G. Sejdiu, G. Argyriou, L. Selmi, J. Jakobitsch, and J. Lehmann, "Managing Lifecycle of Big Data Applications", *Knowledge Engineering and Semantic Web 2017*, pp. 263-276.

[9] Y. Alshboul, Y. Wang, and R. Nebali, "Big Data LifeCycle: Thretas and Security Model", *21st Americas Conference on Information Systems, 2015*.

[10] Y. Demchenko, C. De-Laat, and P. Membrey, "Defining architecture components of the Big Data Ecosystem", *2014 International Conference on Collaboration Technologies and Systems*, pp. 104-112.

[11] M. C. K. Khalilov, and M. Gündebahar, "Bankacılıkta Büyük Veri Uygulamaları: Bir İnceleme, *Akademik Bilişim Konferansı 2013*, pp. 165-172.

[12] S. Altındiş, and İ. K. Morkoç, "Sağlık Hizmetlerinde Büyük Veri", Ömer Halisdemir Üniversitesi, *İktisadi ve İdari Bilimler Fakültesi Dergisi*, vol. 11, no. 2, pp. 257-271, Nisan 2018.

[13] E. Akdamar, "Akıllı Kent İdealine Ulaşmda Büyük Verinin Rolü", *Rewieved Journal of Urban Culture and Management*, vol. 10, no. 2, pp. 200-215.

[14] Z. Özen, E. Kartal, and İ.E. Emre, *Eğitimde Büyük Veri*, Eğitim Teknolojileri Okumaları, Edition: 1, Chapter: 11, Publisher: Pegem akademi Yayıncılık, pp.106-118, 2017.

[15] T. M. Scjolz, *Big Data in Organizations and the Role of Human Resource Management: a complex systems theory-based conceptualization*, Chapter: 2, 2016.

[16] H. Wu and, A. Yamaguchi, "Semantic Web Technologies for the Big Data in Life Sciences", *Bioscience Trends*, vol. 8, no. 4, pp.192-201, August 2014.

[17] J. Rowley, "The wisdom hierarchy: representations of the DIKW hierarchy", *Journal of Information Science 2007*, vol. 33, no. 2, pp. 163-180.

[18] Ç. E. Akay, "Ekonometride Yeni Bir Ufuk: Büyük Veri ve Makine Öğrenmesi", *Social Sciences Research Journal*, vol. 7, no. 2, pp.41-53, June 2018.

[19] Cognizant Technology solutions Talend, Inc., "Data Lake on the AWS Cloud with Talend big Data Platform, AWS Services and Cognizant Best Practices", *Quick start Deployment Guide*, 2017.

[20] J. Merino, I. Caballero, B. Rivas, M. Serrano, and M. Piattini, "A Data Quality in Use model for Big Data", *Future Generation Computer Systems 2016*, vol. 63, pp. 123–130.

[21] S. A. C. Bukhari, A. K. Bashir, and K. M. Malik, "Semantic Web in the Age of Big Data: A Perspective", OSF Storage (United states), 2018,

[22] D. Gil, and I. Song, "Modeling and Management of Big Data: Challenges and Opportunities", *Future Generation Computer Systems 2016*, vol. 63, pp. 96-99.

[23] A. Alharthi, V. Krotov, and M. Bowman, "Addressing barriers to big data", *Business Horizons 2017*, vol. 60, pp.285-292.

[24] U. Sivarajah, M.M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods", *Journal of Business Research 2017*, vol. 70, pp.263-286.

[25] D. Ostrowski, N. Rychtyckyj, P. MacNeille, and M. Kim, "Integration of Big Data Using Semantic Web Technologies, *IEEE Tenth International Conference on Semantic Computing 2016.*