Big Data Tools and Technologies

Salem Abdulali Ahmed Abdulali & Yasemin Gültepe
Department of Computer Engineering, Kastamonu University, Kastamonu, Turkey
Corresponding Author Email: salem2076@gmail.com

Abstract
The rapid increase in digital data generated around the world has induced the development of big data applications that enable storing, managing, processing and producing meaningful results of large amounts, speeds and diversity. Effective use of huge data, which has emerged thanks to big data technologies, has brought user pleasure, competitive advantage and high gainfulness with advancements in business processes, increased efficiency and increased service quality. The purpose of this study is to reveal the importance of big data and information about big data tools and technologies are given.

Key words*: Big data, big data technologies, big data tools, big data analytics.*

Introduction

When we are getting more and more digitally in line with technological developments and the data is flowing from almost everything, the world hosts faster, diverse and large amounts of digital data than ever before. Big Data technologies are widely used in our daily lives. Thanks to big data technologies, it is possible to analyze the said data at the same speed; It is provided that algorithms make more accurate inferences, make accurate decisions, discover hidden insights and automate business processes. Thus, the costs are reduced, the quality of the products and services offered is increased and the economic growth rate is increasing.

Big Data is the form of all these data that we get from different sources such as social media shares, our photo archives, and "log" files that we continuously record, and have been converted into a meaningful and machineable form.

Today, database experts categorize the concrete data associated with them in relational databases in a structural way. Company executives also make decisions thanks to reports produced through reporting systems operating in these databases. But there are many more datasets that we cannot completely classify and classify these relationships. Until now, all this information was called information dump because it was very hard to keep this data in our databases and use it in reporting systems (Bott et al., 2016).

Big data applications are used in many fields, especially in industry. In general, many examples can be given from health to marketing, from production to consumption, from public to private sector, from banking to insurance companies, from telecom to aviation sector. Health; quality trends in relation to health customs, health decision support systems, remote patient tracking, patient profiling, disease prediction, etc. Marketing; cross-selling, location-based marketing, feeling analysis, trend analysis, behavior analysis, etc. On production; sensor-based operations, supply chain, inventory management, logistics, etc. Public; basic need detection, traffic problems solution, noise, air and water pollution prevention, abuse detection, etc. In Banking and Insurance; detection of fraud, misuse or misuse, customer estimation, risk analysis, etc. In the telecom sector;

advertising with geographic target, emergency response, central planning, pricing, abuse preclusion, intrusion detection, etc. issues are used (Canbek, & Sağıroğlu, 2006).
The number of technologies used to stock, process, handle and analyze large data sets is increasing daily. Big Data Technologies has features such as processing all kinds of data, expansion according to need, backing up and accessing data and being open source projects. Big Data also focuses on the technology required to manage data in large volumes and unstructured format. Although Big Data Technologies are new, they have aroused great interest today. The most important feature of Big Data Technologies is how it can add value to organizations, reduce costs, reduce data processing time, improve new product and service quality or use new data models to make better decisions.

Big Data Components

To analyze big data, the components of the big data phenomenon must first be well known. Big data consists of five basic components. These components are; variety, velocity, volume, verification and value. This reason is known in the literature as the 5V rule.

a) *Variety:* Since the data produced is not generally structural and consists of data formats obtained from many different environments, they must be integrated and convertible.
b) *Velocity:* Big data production adds speed to each passing day and this data reaches incredible dimensions in a second. Fast growing data reveals that the number and variety of transactions that need that data increase at the same speed, and we should be able to remove this density both in software and hardware.
c) *Volume (Data Size):* Our data, which we call as big data, may be increasing day by day, and we should consider how we will deal with these data stacks in the future and make our plans accordingly.
d) *Verification:* Verification can be seen as another data component when we need to check whether the incoming data is secure during the flow of such fast growing data. This data may be visible to the right people or need to be kept private.
e) *Value:* Perhaps one of the most important layers is the "Value" layer, after our data is filtered from the above data components, the data obtained in the production and processing layers of big data should provide added value for our company.

Especially with the spread of internet technology, the amount of unstructured data is increasing rapidly. It is stated that the rate of unstructured data in the digital universe is over 90 percent. Traditional analytical platforms could not cope with different kinds of data at the same time and traditional databases could not store data in different formats (Table 1.) (Akıncı, 2019).

Table 1: Comparison of Big Data and Traditional Data

	Traditional Data	Big Data
Data Type	Structured	Structured, semi-structured, unstructured
Data Volume	Terabytes	Petabytes, and exabytes
Data Structure	Centralized	Distributed
The relationship of data	Uncertain	Complex

Therefore, before big data technologies, unstructured data could be ignored or used with very low efficiency. Data in an unknown form or structure is classified as unstructured data. Data that can be stored, accessed and processed in fixed format is called structured data. If the data is defined but not structured, it can be classified as semi-structured data. Today, unstructured data in databases designed using the NoSQL structure can be managed, processed and analyzed by data mining methods and new techniques such as Hadoop and MapReduce. NoSQL databases are specially designed for specific data models and have flexible schemes for creating modern applications (Qussous et al., 2015). NoSQL databases have been widely accepted with ease of development, functionality and performance at the appropriate scale.

Data analytics is the main purpose and key characteristic of Big Data. With the amounts of data emitted from various digital sources, the importance of data analysis has increased enormously over all these years. Big data analytics is the use of innovative analytical techniques against large, different datasets containing different sizes of structural, semi-structural and non-structural data from different sources and terabytes to zetabytes.

Big Data Technologies

It is stated that two big technological developments have contributed to the solution of problems related to big data. The first of these technological developments is that with the emergence of cloud-based solutions, data storage costs have decreased significantly and the use of commercial databases has become widespread. Cloud computing is a modern approach that provides hardware and software resources as a service, enabling a much higher scalability than traditional client server architectures (Ullah et al., 2018). The transition from services managed as open source or virtual file systems of certain companies to service-based management has been accelerated by meeting information technologies needs (Zheng et al., 2013).

The second is the creation of new technological solutions consisting of simple hardware combined with distributed file systems for the analysis of large volumes of data. At the beginning of these solutions; Developed by Google to quickly process problems by dividing them into different units, MapReduce, the Hadoop cluster used by Facebook, Storm, which provides real-time data processing on Twitter, and Hana, developed by SAP, which enables faster processing in the main memory instead of storing the data on disk, comes (Bharti et al., 2019). Among these technologies, Hadoop, Spark and NoSQL (Not only SQL) are the most widely used today.

Hadoop: Hadoop is an open source software framework that allows data to be stored, processed and analyzed in a distributed environment across your computer. Hadoop is designed to distribute data to different machines instead of using a single server where each machine offers local computing and storage. Hadoop uses a simple programming model to perform the necessary operations between these clusters and runs the application using the MapReduce algorithm, where data is processed in parallel on distinctive CPU nodes. In other words, Hadoop is used to perform statistical analysis on big data stored on a computer. Hadoop consists of several components that work together to process Bulk Data: Hadoop Distributed File System (HDFS), Yet Another Resource Negotiator (YARN), and MapReduce (Cumbane & Gidófalvi, 2019).

HSFS: HDFS is the basic file system that enables disks in a distributed environment to function as a single virtual disk. There are many tools and technologies on Hadoop for processing, interpreting, querying, and resource management (Shvachko et al., 2010).
YARN: YARN is known as Hadoop 2.0 and is widely used for processing and managing Big Data distributed today (Dezyre, 2017). YARN performs business planning and resource management tasks that do not have to use Hadoop MapReduce on Hadoop systems. Hadoop YARN has an improved architecture different from the original features of Hadoop 1.0, so systems can scale to new levels and responsibilities can easily be assigned to various components in Hadoop HDFS.
MapReduce: MapReduce is a Java-based system created by Google that enables effective processing of real data in HDFS (Maitrey & Jha2015). MapReduce does a large data processing job by dividing it into small tasks. MapReduce analyzes the data in parallel before shrinking data to find results. In the Hadoop ecosystem, Hadoop MapReduce is a framework based on YARN architecture. The YARN-based Hadoop architecture promotes parallel processing of large data sets, and MapReduce ensures a framework for easily writing applications on thousands of nodes, taking into account defect and error management.
Spark: Spark is an open source framework designed for faster analysis. Spark is drew up to promote and facilitate a wide range of data analysis tasks, from graphics processing to machine learning, as well as big data processing (Lathar, 2015). For example, if you load data using an SQL query and then appraise a machine learning model using Spark's ML library, Spark can associate these steps in a single scan. Spark's libraries provide users with a wide variety of functions. Spark uses its own clustering system. Resilient Distributed Data Set (RDD) is Spark's basic data structure. It is a literal, distributed object collection. Flexible distributed datasets can contain Python, Java, or Scala objects.

Artificial Intelligence in Big Data Analysis

The concept of big data emphasizes not only the extraordinary size of the data set, but also the high data generation speed and data diversity. With the use of big data interpreted as the beginning of a new era, problems such as storage, security and privacy, processing and analysis of data and decision making based on data arise (Chen & Zhang, 2014). The increase in the amount of data can be valuable when it is possible to develop algorithms that can make sense of this data, understand what it is related to, find the classes to which they belong. Data that cannot be processed and converted into information cannot be benefited from. These processes require data mining, computer science, machine learning, database management, mathematical algorithms, and statistics to work together.
Artificial intelligence techniques, which mimic the smart behavior of live data, think like a human and aim to create models that aim to make decisions with the approaches brought by technology and applications, are preferred because of their advantages in big data studies.
The difference of artificial neural networks from traditional analysis methods; parallel processing, that is, the work of independent calculation resources on the same task at the same time. Through artificial neural network models, data is separated into independent processors and each processor operates independently. The most common parallel processing models used in big data analysis are MPI (Message Passing Interface), MapReduce and Dryad models (John & Thomas, 2014). Artificial neural network applications are effectively used in different areas such as face

recognition, credit decisions, handwriting recognition, grading of financial status of businesses and fraud detection.

In the fight against global and national epidemics, big data has reached the point where it can provide the system and technology that has the potential to monitor, capture and even stop the spread of the disease. For example, in the epidemic of Coronavirus (COVID-19) occurring in Wuhan city of China and spreading all over the world, it is known that big data and artificial intelligence studies are gaining speed and being used effectively. The processing of digital data has made the outbreak pre-identifiable and predictable (Toker, 2020).

Conclusion

Big data can give businesses more intellect and foresight by using artificial neural networks, artificial intelligence and machine learning applications in big data analysis, deep learning, natural language processing, image recognition and forward personalization technologies. On the other hand, artificial intelligence techniques, which aim to create models that mimic the smart behavior of living creatures in the nature, and think and make decisions as people, are also preferred with the advantages they provide in studies on big data.

In today's technology, many tools and infrastructures have been developed for big data analysis, operation and management. In order to use big data tools efficiently, your system infrastructure must be strong. In this article, the importance of big data is revealed and the most applied big data analysis tools are explained.

References

Akıncı, A.N. (2019). Big data, personal data privacy in applications. Specialization Thesis. General Directorate of Sectors and Public Investments.

Bharti, U., Bajaj, D., Goel, A. & Gupta, S. C. (2019). Identifying requirements for big data analytics and mapping to hadoop tools. International Journal of Recent Technology and Engineering, 8(3): 4384-4392.

Botta, A., de Donato, W., Persico, V., & Pescapé, A. (2016). Integration of cloud computing and internet of things: A survey. Future Generation Computer Systems. 56, 684-700.

Canbek, G., & Sağıroğlu, Ş. (2006). Bilgi, bilgi güvenliği ve süreçleri üzerine bir inceleme. Politeknik Dergisi, 9(3).

Chen, C.P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. Information Sciences. 275, 314-347.

Cumbane, S.P., & Gidófalvi, G. (2019). Review of big data and processing frameworks for disaster response applications. International Journal of Geo-Information. 8(387).

Dezyre (2017). https://www.dezyre.com/article/hadoop-2-0-yarn-framework-the-gateway-to-easier-programming-for-hadoop-users/84#:~:text=Hadoop%202.0%20popularly%20known%20as,and%20managing%20distributed%20big%20data.&text=Hadoop%20YARN%20comes%20along%20with,are%20shipped%20by%20Hadoop%20distributors.

John, I. A., & Thomas, T. A. (2014). A survey on big data mining challenges. International Journal of Modern Trends in Engineering and Research. 2(01): 253-256.

Lathar, P. (2015). Big data analysis: Apache spark perspective. International Journal of Technical Innovation in Modern Engineering & Science. 4(5): 849-857.

Maitrey, S., & Jha, C.K. (2015). MapReduce: Simplified data analysis of big data. Procedia computer science. 57(2015): 563-571.

Oussous, A., Benjelloun, F.-Z., Lahcen, A.A., & Belfkih, S. (2015). Comparison and classification of NoSQL databases for big data. International Conference on Big Data, Cloud and Applications.

Shavachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). 26th Symposium on Mass Storage Systems and Technologies.

Toker, S. (2020). Great data and artificial intelligence studies in the fight against the coronavirus outlook. Seta Perspective.

Ullah, S., Awan, M.D., & Khiyal, M.S.H. (2018). Big data in cloud computing: A resource management perspective. HPC sSoftware and Programming Environments for Big Data Applications.

Zheng, Z., Zhu, J., & Lyu, M. R. (2013). Service-generated big data and big data-as-a-service: An overview. IEEE International Congree on Big Data. 403-410.