Big Data Applications and Tools

Mahmoud Al-Khasawneh
Universiti Teknologi Malaysia

Chapter - 1
Big Data Applications and Tools

Author

Mahmoud Ahmad Al-Khasawneh
Faculty of Computer & Information Technology, Al-Madinah International University, Kuala Lumpur, Malaysia

Chapter - 1

Big Data Applications and Tools

Mahmoud Ahmad Al-Khasawneh

Abstract

Big data has grown into a very broad process. It refers to the huge amount of structured, semi-structured and unstructured data that is produced exponentially in many areas by high-performance applications, recently many applications of big data used, such as in Education, Healthcare and many of our daily life aspects. In this chapter the applications and tools of big data will be covered in depth.

Keywords: big data, application of big data, big data tools

1. Introduction

Gartner described Big Data as large volume information assets with fast velocity and different variety which require innovative platform in order to improve insights and decision making. Meanwhile, in describing the notion of revolution, the authors described it as a method of solving all the long unsolved problems associated with data management and handling. Big data analytics allows the hidden patterns to be unlocked and customers would be understood more comprehensively in terms of their opinions and needs. Big Data is thus considered as a revolution.

The generation of big data involves very large quantities (multi-terabyte). Big data change fast and take many forms. Hence, RDBMS or other traditional technologies do not have the capacity to manage and process them. Accordingly, through their tools, methods, and technologies, Big Data solutions allow the fast capturing, storing, searching and analysis of data. This in turn allows the discovery of the links and insights for innovation and competitive gain, and such capacities were not possible in the past.

Data today, aside from being extremely vast in amount as opposed to the amount available in the past, are mostly unstructured (about 80% of them), and evidently, the traditional technologies cannot handle such data. Furthermore, data are continuously archived just as they are analyzed. Somehow, considering their colossal amount-data generation today is in petabytes-it is rather impossible to archive and retrieve them repeatedly. Meanwhile, data scientists frequently have to use current data for various purposes including for predictive analysis unlike historical as used to be done with traditional [1-10].

2. Big data applications

The applications of Big Data can be observed in various domains, as follows:

1. Big data in retail

Competitions in the retail industry is very fierce and retailers are continuously striving to achieve a competitive edge over others, and in order to thrive, it is important that retailers understand their customers really well. Having the awareness of the needs of the customers and how to optimally satisfy them will give the company a competitive edge. Also, by performing advanced analysis on their customer's data, retailers could fully understand their customers. The data of customers can be obtained via many resources including social media, loyalty programs and so forth. For retailers, all details of the customers are of value and having understanding of all of these minute details brings the retailers to their customers as close as possible. Consequently, the retailers could provide their customers with more personalized services and also forecast their future demands. Loyal customer can thus be established. Costco, Walmart, Walgreens, and Sears and Holdings are among the retailers that heavily utilize Big Data. Relevantly, the National Retail Federation estimated that about 30% of retail annual sales come from sales made in November and December [11].

2. Big data in healthcare

Big Data greatly facilitates the healthcare industry as this industry consistently has to deal with very large amount of data. Such amount of data has made it rather impossible for the healthcare practitioners to harness them. The use of Big Data can be regarded as lifesaving as it facilitates the practitioners and researchers in this industry to detect and cure diseases such as cancer. Also through Big Data and analytics, more personalized medications can be established, and more effective treatments can be provided to the patients. Furthermore, unique patterns of certain medicines can be identified, allowing the development of more cost-effective solutions. Also, smart wearables are becoming popular among all people, as these devices can help save lives through producing real-time alerts [12].

3. Big data in education

Within the realm of education, data are generally important for future references. Hence, data are highly important in this domain. The use of Big Data greatly enhances the system of education, by specifically revitalizing the skills, both academic and non-academic ones. Also, the use of Big Data facilitates the evaluation of performances of students and teachers. Big Data has also been used in academic curriculum reformation in some leading universities. Equally, Big Data can be used in tracking the rate of dropout and then in determining the most appropriate measures to reduce it [13].

4. Big data in e-commerce

E-commerce has been regarded as a remarkable revolution in this era and it has become an integral part of life of people today. Hence, it is common for people to be thinking about E-commerce when they want to purchase something. In this regard, Amazon, Flipkart and Alibaba are among the most notable global E-commerce companies and the use of Big Data in these companies is extensive. Relevantly, Amazon as the world's biggest E-commerce company is one of the leaders *in* Big Data and analytics. Meanwhile, Flipkart which is an Indian-based company, has one of the most vigorous data platforms in the country [14].

Within the domain of Big Data, the recommendation engine of Big Data is by far the most extraordinary applications of Big Data as it provides a 360-degree view of customers to the companies. This allows the companies to make appropriate recommendations to the customers, making the services more personalized. Indeed, the experiences of online shipping of people are completely redefined through Big Data.

5. Big data in media and entertainment

Media and Entertainment industry generally involves art and the use of Big Data is regarded as part of this art. Even though discrete from one another, the combination of art and science can generate remarkable outcomes especially in this industry. The general aim of this industry is to please customers and thus, it is crucial that this industry is able to consistently present new content to customers in order to retain them. In this regard, it is vital to have recommendation engine.

Meanwhile, viewers today are inclined to choose the contents that they want, and generally, viewers prefer fairly new contents. Prior to the emergence of Big Data, companies would randomly broadcast their advertisement without performing any analysis first, and now with Big Data analytics, companies could determine the type of Ads to broadcast (i.e., those that would attract customers) and the best broadcast time to achieve the maximum attention [15].

6. Big data in finance

Financial organizations greatly rely on data in their operations, and in fact, for such organizations, data are their second most vital commodity after money. Owing to such importance, financial organizations need to assure safety of their data which is a challenging task. Financial firms were in fact among the first adopters of Big Data and Analytics, and prior to that, these firms were already mastering the technical field. Relevantly, Digital banking and payments have been among the most trending buzzwords with Big Data as their important element. In financial firms, Big Data handles the major domains including algorithmic trading**,** fraud detection**,** risk analysis**,** and customer contentment. With Big Data, the financial system becomes fluent, improved, and empowered in making available superior services to the customers [16].

7. Big data in travel industry

Within the context of travel industry, the spread of Big Data has been rather slow as opposed to other industries. Relevantly, stress-free traveling experience is desired, and the use Big Data can assist in assuring such. In this regard, travelling companies can utilize Big Data analytics to offer their customers with more personalized traveling experience. Equally, the companies can use Big Data Analytics to better understand their customer's requirements, while providing them with the best offers and suggestions in real-time. Big Data can become an excellent guide for any traveler, making it an increasingly vital part in this industry [17].

8. Big data in telecom

The telecom industry has been the main driver to the global digital revolution. Meanwhile, the increasing use of smartphones all over the world has caused the telecom industry to be flooded with data of colossal amounts. For telecom companies, such data are highly precious, and thus, it is crucial that the companies know how to appropriately exploit these data. The use of Big Data Analytics allows these telecom companies to offer customers with smooth connectivity, and this eases customers as the network barriers are eradicated. Big Data Analytics also allows the companies to track both the highest and lowest traffic areas and then decide the best actions in order to provide customers with smooth network connectivity. With proper understanding of the customers through the use of Big Data, Telecom industries can provide customers with customized services [18].

9. Big data in automobile industry

The industry of automobile is now fully and smoothly controlled by Big Data, resulting in some extraordinary and novel outcomes. It can thus be said that Big Data is now a vital part in automobile industry as it has led to the achievement of the unimaginable. With Big Data, the industry could analyze trends, understand the supply chain management, provide care to the customers, make the impossible possible, and so forth [19].

3. Big data tools

Big Data Analytics is now integral in nearly all organizations, and in order to achieve notable outcomes, a set of tools is required at every phase of data processing and analysis. In determining the tools to be employed, several factors to be taken into account. Among these factors include the size of the datasets, pricing of the tool, the type of analysis to be executed, and so forth. As Big Data is growing exponentially, the market has been overwhelmed by its plethora of tools that are of value in assisting big data in increasing cost efficiency, which translates into faster analysis.

1. Apache hadoop

Apache Hadoop is a very popular tool in Big Data industry and this tool stores process and analyzes Big Data. Hadoop, which is written in Java, is an open-source framework from Apache and it operates on commodity hardware. The use of Apache Hadoop allows data to be processed in parallel manner because Apache Hadoop can concurrently run on multiple machines. Apache Hadoop utilizes clustered architecture, whereby a Cluster encompasses a group of systems connected through LAN.

There are three parts of Hadoop as follows:

- Hadoop Distributed File System (HDFS) which encompasses Hadoop's storage layer
- Map-Reduce which encompasses Hadoop's layer of data processing
- YARN which encompasses Hadoop's layer of resource management

Among the shortcomings of Hadoop usage include non-support to real-time processing as Hadoop only supports batch processing and inability of performing in-memory calculations [20].

2. Apache spark

Apache Spark which encompasses a general-purpose clustering system can be regarded as the replacement of Hadoop. Apache Spark resolves the issues related to Hadoop as it supports real-time and batch processing, while also supporting in-memory calculations. Further, it has much smaller amount of read/write operations into the disk, making it much faster than Hadoop, specifically 100 times faster. Also, as opposed to Hadoop, Apache Spark runs on different data stores such as HDFS, Apache Cassandra and OpenStack, making it more flexible and versatile than Hadoop.

Apache Spark provides APIs of high-level in Python, Java, Scala and R, while also providing a weighty set of high-level tools. Among these tools are: Spark Streaming, Spark SQL for structured data processing, GraphX for graph dataset processing, and MLlib for machine learning. For efficient execution of query, Apache Spark has 80 operators of high-level [21].

3. Apache storm

Apache Storm is an open-source tool of big data and it entails a distributed real-time and fault-tolerant processing system, with the ability to process unbounded streams of data efficiently. In this context, unbounded streams relate to consistently expanding data that have a beginning but no defined end. Apache Storm is usable in any language of programming and it also supports JSON based protocols. Storm assures processing of all data and has very fast processing as high as a million tuples processed per second on each node. Storm is also fault tolerant and easily scalable, making this tool much simpler in terms of usage [22].

4. Apache Cassandra

Apache Cassandra is a distributed database and it offers high availability and scalability while the performance remains efficient. Apache Cassandra is considered as a big data tool (one of the best) with the ability to accommodate datasets of all types namely structured, semi-structured, and unstructured. This tool is considered as the most appropriate platform for mission-critical data with no failure point. Also, it tolerates faults in both commodity hardware and cloud infrastructure. Cassandra can efficiently operate under heavy loads, and since it does not follow master-slave architecture, the nodes all play the same role. Apache Cassandra supports the ACID (Atomicity, Consistency, Isolation, and Durability) properties [23].

5. MongoDB

As an open-source data analytics tool, MongoDB encompasses a NoSQL database that offers cross-platform capacities. Such capacities could cater to businesses that need fast-moving and real-time data in decision making. This tool is useful to those seeking data-driven solutions. MongoDB which is written in JavaScript, and C, C++, is reliable and cost-effective. Also, owing to its easy installation and maintenance, MongoDB is user-friendly. MongoDB expedites the management of unstructured data or the frequently changing data, making it a very popular Big Data database. As MongoDB utilizes dynamic schemas, user can quickly prepare the data, and this can decrease the overall cost. MongoDB runs on MEAN software stack, NET applications and, Java platform, and within the cloud infrastructure, MongoDB is flexible. However, for certain cases of usage, some breakdown in the speed of processing has been recognized [24].

6. Apache Flink

Apache Flink which is written in Java and Scala entails an Open-source data analytics tool distributed processing framework for data streams, bounded and unbounded. Even for late-arriving data, Apache Flink provides highly accurate results. Apache Flink can easily recover from faults and therefore it is fault tolerant. At a large scale, Flink offers high-performance efficiency on thousands of nodes. Other advantages of Apache Flink include low-latency, high throughput streaming engine and supports on event time and state management [25].

7. Kafka

Introduced by LinkedIn in 2011, Apache Kafka entails a distributed event processing or streaming open-source platform providing high throughput to the systems. This tool has the capacity to handle trillions of events on a daily basis, aside from being fault tolerant. As a streaming platform, Apache is very scalable, and among the processes of streaming are publishing and subscribing to streams of records that are akin to the messaging systems. These records are permanently stored in groups called topics and then processed. Apache Kafka offers high-speed streaming with no downtime [26].

8. Tableau

Tableau is regarded as among the best tools of data visualization and software solution within the industry of Business Intelligence. Tableau turns raw data into important insights while also improving the decision-making process of the businesses. The process of data analysis provided by Tableau is fast, resulting in visualizations in the form of interactive dashboards and worksheets. Also, Tableau allows the best data blending in the market, and also an efficient real-time analysis. As a vital part in certain industries, Tableau is bound to the industry of technology as well. Furthermore, running Tableau does not require technical or programming skills [27].

9. Rapid miner

Rapid miner which is an open-source tool written in java, is a cross-platform tool. It offers a strong environment for Machine Learning, Data Science and Data Analytics procedures. RapidMiner is a unified platform for the full Data Science lifecycle, and it begins from data prep to machine learning to the employment of predictive model. Many licenses are provided by RapidMiner for small, medium, and large proprietary editions. A free edition is also provided by RapidMiner, and this edition allows only 1 logical processor and up to 10,000 data rows. When combined with APIs and cloud services, Rapid Miner is very efficient. This tool offers some robust Data Science tools and algorithms [28].

10. R Programming

As an open-source programming language, R is among the most inclusive languages of statistical analysis. R, which is written in C and Fortran, encompasses a multi-paradigm language of programming offering a dynamic environment of development, and being an open-source project has allowed countless of individuals to partake in its development. A massive package ecosystem is offered by R, making it among the most popular tools of statistical analysis. R aids the efficient performance of various statistical operations while also easing the production of data analysis results both in text and graphic format. R offers remarkable graphics and charting [29].

2. Advantages of using big data in business

The use of Big Data in the competitive business world today is highly beneficial, as Big Data provide the current market with endless services. Accordingly, proper utilization of Big Data can lead to remarkable results. Hence, owing to various reasons, nearly all companies today are shifting towards Big Data Analytics. As noted, the use of Big Data has facilitated these companies in improving their general growth. Accordingly, the advantages of big data to businesses are as discussed below:

1. Better decision making

The application of Big Data Analytics has been found to greatly enhance the process of decision making. Hence, companies have begun to consider using Big Data Analytics in making decisions, as opposed to making decisions anonymously. Big Data Analytics allows companies to consider various factors associated with customers such as what customers want and what can be done to resolve problems. Also, Big Data Analytics allows companies to analyze the needs of customers in accordance with the market trends. All of these factors help companies in improving their decision-making process.

As indicated, the use of big data improves the process of decision making. The use of Data Flair as part of big data is one such example. The details are as follows:

2. Big data in greater innovations

Organizations need to innovate to achieve success, and for the purpose, data are needed. In this regard, more data is better, and the use of Big Data allows the organizations to accomplish what was once deemed impossible or even unthinkable. Many firms make use of Big Data Analytics in creating new products and services, and using Big Data, firms could analyze opinions of various customers in regards to their products and how these customers perceive these products.

The tools provide firms with useful information about the products, for instance, the information on the strengths and weaknesses of the products. Accordingly, firms will take into account such information during the development of new products in order that the newly created products will cater to the needs and wants of the customers. In essence, Big Data Analytics allows companies to 'think' beyond the ordinary.

3. Big data in education sector

The sector of education can benefit from the use of Big Data as such usage allows the management of student related data which is otherwise difficult to execute using the traditional methods. In general, student data are of large size and teacher often find it difficult to properly exploit them. Hence, Big Data Analytics is valuable to the education sector as it has facilitated the much-needed transformation in the system of this sector. Using Big Data Analytics, teacher could analyze the capacities of students, and based on the outcomes, teacher could help nurture the future of these students. The tools of Big Data Analytics allow teachers to discover both the strengths and weaknesses of students and appropriately guide them.

4. Big data in product price optimization

Companies can utilize Big Data for price optimization purposes. In this regard, the goal is to establish prices that will maximize profits, and Big Data can be used in determining the prices which generate the maximum profits under many historic market circumstances. Big Data analytics allows companies to establish prices based on the willingness of customers to pay under different conditions. The primary purpose is to have customers feel that they get good value for money, and such view can help the company to continue growing. However, the company should appropriately improve their product based on the trend to assure consistent satisfaction of customers and Big Data can facilitate the company in doing so.

5. Big data in recommendation engines

Online platform users today are provided with recommendations according to both their past and present choices on the numerous online platforms. Being provided with choices of favored things will ease the life or users. Also, such availability has transformed the way online platform is viewed and such platform provides more comfort use to users.

Big Data recommendation engine is available in many platforms of online shopping. The data of customers are analyzed and appropriate recommendations are then made. In general, the recommendations relate to what the customer did during the last visit and in the customer’s real-time activities. Comparisons are made between the customers who searched or purchased familiar items, and suggestions are then made based on these comparisons. This eradicates the physical barriers between online platforms and the customers, and changed the experience of online shopping.

6. Big data in life-saving application in healthcare industry

The Healthcare Industry has been greatly benefitting from the introduction of Big Data Analytics, and in fact, Big Data Analytics can be considered as a progressing Revolution. In fact, experts of Big Data at QUANTZIG, which is a provider of Global Analytics Solutions, *regard* Big Data and Advanced Analytics as the potential solution to the most challenging hurdles of Healthcare.

The application of Big Data in healthcare facilitates practitioners in making available to their patients the cutting-edge and quality healthcare utilizing the patients’ electronic health records. Through the use of Big Data, the entire operational efficiency of the healthcare companies is improved. Also, the healthcare companies could make changes as required. Furthermore, the use of Big Data Analytics allows the healthcare companies to detect the unidentified connections and concealed patterns in a given disease such as cancer, and then find a better cure.

Summary

Big Data applications worldwide denote its significance and its benefits are enormous. This is evidenced by the proclivity of all industries worldwide towards the application of Big Data Analytics. Big Data Analytics is increasingly becoming an integral part in businesses of all types globally that soon it cannot be ignored.

References

1. Rodríguez-Mazahua L, Rodríguez-Enríquez CA, Sánchez-Cervantes JL, Cervantes J, García-Alcaraz JL, Alor-Hernández G. A general perspective of Big Data: applications, tools, challenges and trends. The Journal of Supercomputing. 2016; 72(8):3073-3113.

2. Ali Shah SA, Uddin I, Aziz F, Ahmad S, Al-Khasawneh MA, Sharaf M. An enhanced deep neural network for predicting workplace absenteeism. Complexity, 2020.

3. Khan FA, Abubakar A, Mahmoud M, Al-Khasawneh MA, Alarood AA. Cotton crop cultivation oriented semantic framework based on IoT smart farming application. International Journal of Engineering and Advanced Technology. 2019; 8(3):480-484.

4. Al-Khasawneh MA, Shamsuddin SM, Hasan S, Bakar AA. Map Reduce a comprehensive review. In International Conference on Smart Computing and Electronic Enterprise (ICSCEE). IEEE, 2018, 1-6.

5. Uddin MI, Zada N, Aziz F, Saeed Y, Zeb A, Ali Shah SA *et al*. Prediction of Future Terrorist Activities Using Deep Neural Networks. Complexity, 2020.

6. Al-Khasawneh MA, Shamsuddin SM, Hasan S, Bakar AA An improved chaotic image encryption algorithm. In International Conference on Smart Computing and Electronic Enterprise (ICSCEE). IEEE, 2018, 1-8.

7. Ullah Z, Zeb A, Ullah I, Awan KM, Saeed Y, Uddin MI *et al*. Certificateless Proxy Reencryption Scheme (CPRES) Based on Hyperelliptic Curve for Access Control in Content-Centric Network (CCN). Mobile Information Systems, 2020.

8. Uddin MI, Shah SAA, Al-Khasawneh MA. A Novel Deep Convolutional Neural Network Model to Monitor People following Guidelines to Avoid COVID-19. Journal of Sensors, 2020.

9. Khan FA, Abubakar A, Mahmoud M, Al-Khasawneh MA, Alarood AA. Rift: A High-Performance Consensus Algorithm for Consortium Blockchain. International Journal of Recent Technology and Engineering (IJRTE), 2019.

10. Khan FA, Abubakar A, Mahmoud M, Al-Khasawneh MA, Alarood AA. BSCL: Blockchain-Oriented SDN Controlled Cloud Based Li-Fi Communication Architecture for Smart City Network. International Journal of Engineering & Technology, 2018.

11. Santoro G, Fiano F, Bertoldi B, Ciampi F. Big data for business management in the retail industry. Management Decision, 2018.
12. Baro E, Degoul S, Beuscart R, Chazard E. Toward a literature-driven definition of big data in healthcare. BioMed research international, 2015.
13. Williamson B. Big data in education: The digital future of learning, policy and practice. Sage, 2017.
14. Akter S, Wamba SF. Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets. 2016; 26(2):173-194.
15. Lippell H. Big Data in the Media and Entertainment Sectors. In New Horizons for a Data-Driven Economy. Springer, Cham, 2016, 245-259.
16. Begenau J, Farboodi M, Veldkamp L. Big data in finance and the growth of large firms. Journal of Monetary Economics. 2018; 97:71-87.
17. Davenport T. Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press, 2014.
18. Cheng X, Xu L, Zhang T, Jia Y, Yuan M, Chao K. A novel big data-based telecom operation architecture. In 1st International Conference on Signal and Information Processing, Networking and Computers, 2016, 385-396.
19. Martin KE. Ethical issues in the big data industry. MIS Quarterly Executive. 2015; 14:2.
20. Murthy AC, Vavilapalli VK, Eadline D. Apache Hadoop YARN: moving beyond MapReduce and batch processing with Apache Hadoop 2. Pearson Education, 2014.
21. Spark A. Apache spark. Retrieved January, 17, 2018.
22. Evans R. Apache storm, a hands-on tutorial. In 2015 IEEE International Conference on Cloud Engineering. IEEE, 2015, 2-2.
23. Cassandra A. Apache cassandra. Website, 2014. Available online at http://planetcassandra. org/what-is-apache-cassandra, 13.
24. Banker K. MongoDB in action. Manning Publications Co, 2011.
25. Carbone P, Katsifodimos A, Ewen S, Markl V, Haridi S, Tzoumas K. Apache Flink: Stream and batch processing in a single engine. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2015, 36(4).
26. Garg N. Apache Kafka. Packt Publishing Ltd., 2013.

27. Nair L, Shetty S, Shetty S. Interactive visual analytics on Big Data: Tableau vs D3. js. Journal of e-Learning and Knowledge Society, 2016, 12(4).

28. Hofmann M, Klinkenberg R. (Eds.). RapidMiner: Data mining use cases and business analytics applications. CRC Press, 2016.

29. Jones O, Maillardet R, Robinson A. Introduction to scientific programming and simulation using R. CRC Press, 2014.