ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

The Efficient Big Data Research Technologies, Opportunities, and Challenges

Dr. Sheik Saidhbi

Associate Professor

Department of Computer Science, Samara University

Ahemed Adem

Computer Science, Samara University

ABSTRACT

Big Data has gained much attention from the academia and the IT industry. In the digital and computing world, information is generated and collected at a rate that rapidly exceeds the boundary range. Currently, over 2 billion people worldwide are connected to the Internet, and over 5 billion individuals own mobile phones. By 2020, 50 billion devices are expected to be connected to the Internet. At this point, predicted data production will be 44 times greater than that in 2009. As information is transferred and shared at light speed on optic fiber and wireless networks, the volume of data and the speed of market growth increase. However, the fast growth rate of such large data generates numerous challenges, such as the rapid growth of data, transfer speed, diverse data, and security. Nonetheless, Big Data is still in its infancy stage, and the domain has not been reviewed in general. Hence, this study comprehensively surveys and classifies the various attributes of Big Data, including its nature, definitions, rapid growth rate, volume, management, analysis, and security. This study also proposes a data life cycle that uses the technologies and terminologies of Big Data. Future research directions in this field are determined based on opportunities and several open issues in Big Data domination. These research directions facilitate the exploration of the domain and the development of optimal techniques to address Big Data.

Introduction:

The current international population exceeds 7.2 billion [1], and over 2 billion of these people are connected to the Internet. Furthermore, 5 billion individuals are using various mobile devices, according to McKinsey (2013). As a result of this technological revolution, these millions of people are generating tremendous amounts of data through the increased use of such devices. In particular, remote sensors continuously produce much heterogeneous data that are either structured or unstructured. This data is known as Big Data [2]. Big Data is characterized by three aspects: (a) the data are numerous, (b) the data cannot be categorized into regular relational databases, and (c) data are generated, captured, and processed very quickly. Big Data is promising for business application and is rapidly increasing as a segment of the IT industry. It has generated significant interest in various fields, including the manufacture of healthcare machines, banking transactions, social media, and satellite imaging [3]. Traditionally, data is stored in a highly structured format to maximize its informational contents. However, current data volumes are driven by both unstructured and semi structured data. Therefore, end-to-end processing can be impeded by the translation between structured data in relational systems of database management and unstructured data for analytics. e staggering growth rate of the amount of collected data generates numerous critical issues and challenges described by [4], such as rapid data growth, transfer speed, diverse data, and security issues. Nonetheless, the advancements in data storage and mining technologies enable the preservation of these increased amounts of data. In this preservation process, the nature of the data generated by organizations is modified [5]. However, Big Data is still in its infancy stage and has not been reviewed in general. Hence, this study comprehensively surveys and classifies the various attributes of Big Data, including its volume, management, analysis, security, nature, definitions, and rapid growth rate. The study also proposes a data life cycle that uses the technologies and terminologies of Big Data. Future research directions in this field are determined by opportunities and several open issues in Big Data domination.

Back Ground:

The data type that increases most rapidly is unstructured data. This data type is characterized by “human information” such as high-definition videos, movies, photos, scientific simulations, financial transactions, phone records, genomic datasets, seismic images, geospatial maps, e-mail, tweets, Facebook data, call-center conversations, mobile phone calls, website clicks, documents, sensor data, telemetry, medical records and images, climatology and weather records, log files, and text [11]. According to Computer World, unstructured information may account for more than 70% to 80% of all data in organizations. These data, which mostly originate from social media, constitute 80% of the data worldwide and account for 90% of Big Data. Currently, 84% of IT managers process unstructured data, and this percentage is expected to drop by 44% in the near future [11]. Most unstructured data are not modeled, are random, and are difficult to analyze. For many organizations, appropriate strategies must be developed to manage such data. Table 1 describes the rapid production of data in various organizations further. According to Industrial Development Corporation (IDC) and EMC Corporation, the amount of data generated in 2020 will be 44 times greater [40 zettabytes (ZB)] than in 2009. This rate of increase is expected to persist at 50% to 60% annually . To store the increased amount of data, HDDs must have large storage capacities.

Source	Production
Youtube	(i) Users upload 100 hours of new videos per minute (ii) Each month, more than 1 billion unique users access YouTube (iii) Over 6 billion hours of video are watched each month, which corresponds to almost an hour for every person on Earth. This figure is 50% higher than that generated in the previous year
Facebook	(i) Every minute, 34,722 Likes are registered (ii) 100 terabytes (TB) of data are uploaded daily (iii) Currently, the site has 1.4 billion users (iv) The site has been translated into 70 languages
Twitter	(i) The site has over 645 million users (ii) The site generates 175 million tweets per day
Google	The site gets over 2 million search queries per minute Every day, 25 petabytes (PB) are processed
Apple	Approximately 47,000 applications are downloaded per minute
Instagram [20]	Users share 40 million photos per day
LinkedIn [20]	2.1 million groups have been created

Table 1: Rapid growth of unstructured data

Big Data Management:

The architecture of Big Data must be synchronized with the support infrastructure of the organization. To date, all of the data used by organizations are stagnant. Data is increasingly sourced from various fields that are disorganized and messy, such as information from machines or sensors and large sources of public and private data. Previously, most companies were unable to either capture or store these data, and available tools could not manage the data in a reasonable amount of time. However, the new Big Data technology improves performance, facilitates innovation in the products and services of business models, and provides decisionmaking support [8]. Big Data technology aims to minimize hardware and processing costs and to verify the value of Big Data before committing significant company resources. Properly managed Big Data are accessible, reliable, secure, and manageable. Hence, Big Data applications can be applied in various complex scientific disciplines (either single or interdisciplinary), including atmospheric science, astronomy, medicine, biology, genomics, and biogeochemistry. In the following section, we briefly discuss data management tools and propose a new data life cycle that uses the technologies and terminologies of Big Data. 3.1. Management Tools. With the evolution of computing technology, immense volumes can be managed without requiring supercomputers and high cost. Many tools and techniques are available for data management, including Google BigTable, Simple DB, Not Only SQL (NoSQL), Data Stream Management System (DSMS), MemcacheDB, and Voldemort [3]. However, companies must develop special tools and technologies that can store, access, and analyze large amounts of data in near-real time because Big Data differs from the traditional data and cannot be stored in a single machine. Furthermore, Big Data lacks the structure of traditional data. For Big Data, some of the most commonly used tools and techniques are Hadoop, MapReduce, and Big Table. These innovations have redefined data management because they effectively process large amounts of data efficiently, costeffectively, and in a timely manner. The following section describes Hadoop and MapReduce in further detail, as well as the various projects/frameworks that are related to and suitable for the management and analysis of Big Data. 3.2. Hadoop. Hadoop is written in Java and is a top-level Apache project that started in 2006. It emphasizes discovery from the perspective of scalability and analysis to realize near-impossible feats. Doug Cutting developed Hadoop as a collection of open-source projects on which the Google MapReduce programming environment could be applied in a distributed system. Presently, it is used on large amounts of data. With Hadoop, enterprises can harness data that was previously difficult to manage and analyze. Hadoop is used by approximately 63% of organizations to manage huge number of unstructured logs and events (Sys.con Media, 2011). In particular, Hadoop can process extremely large volumes of data with varying structures (or no structure at all). The following section details various Hadoop projects and their links according to [12]. Hadoop is composed of HBase, HCatalog, Pig, Hive, Oozie, Zookeeper, and Kafka; however, the most common components and well-known paradigms are Hadoop Distributed File System (HDFS) and MapReduce for Big Data. Figure 3 illustrates the Hadoop ecosystem, as well as the relation of various components to one another.

HDFS. This paradigm is applied when the amount of data is too much for a single machine. HDFS is more complex than other file systems given the complexities and uncertainties of networks. Cluster contains two types of nodes. The first node is a name-node that acts as a master node. The second node type is a data node that acts as slave node. This type of node comes in multiples. Aside from these two types of nodes, HDFS can also have secondary name-node. HDFS stores files in blocks, the default block size of which is 64 MB. All HDFS files are replicated in multiples to facilitate the parallel processing of large amounts of data.

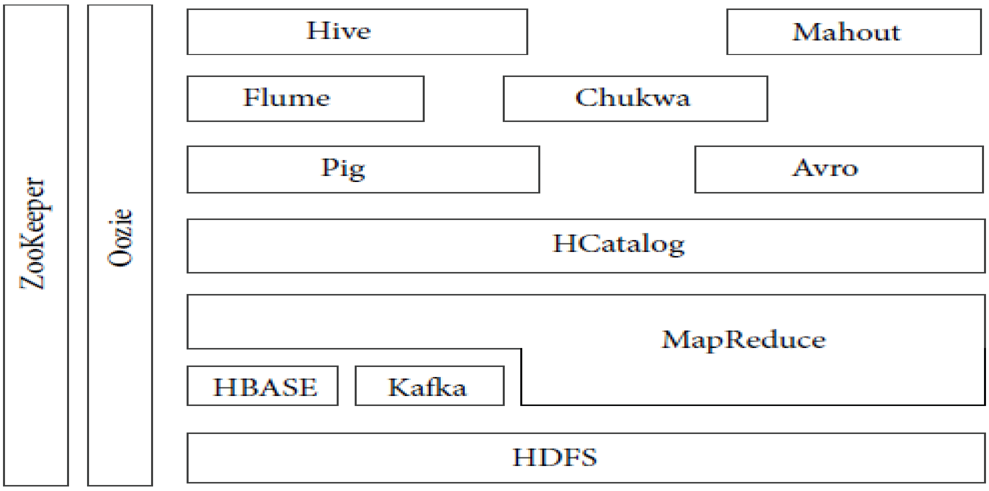

FIGURE 1: Hadoop ecosystem

Life Cycle and Management of Data Using Technologies and Terminologies of Big Data:

During each stage of the data life cycle, the management of Big Data is the most demanding issue. This problem was first raised in the initiatives of UK e-Science a decade ago. In this case, data were geographically distributed, managed, and owned by multiple entities [4]. The new approach to data management and handling required in e-Science is reflected in the scientific data life cycle management (SDLM) model. In this model, existing practices are analyzed in different scientific communities. The generic life cycle of scientific data is composed of sequential stages, including experiment planning (research project), data collection and processing, discussion, feedback, and archiving. The following section presents a general data life cycle that uses the technology and terminology of Big Data. The proposed data life cycle consists of the following stages: collection, filtering & classification, data analysis, storing, sharing & publishing, and data retrieval & discovery. Raw Data. Researchers, agencies, and organizations integrate the collected raw data and increase their value through input from individual program offices and scientific research projects. The data are transformed from their initial state and are stored in a value-added state, including web services. Neither a benchmark nor a globally accepted standard has been set with respect to storing raw data and minimizing data. The code generates the data along with selected parameters. 4.2. Collection/Filtering/Classification. Data collection or generation is generally the first stage of any data life cycle. Large amounts of data are created in the forms of log file data and data from sensors, mobile equipment, satellites, laboratories, supercomputers, searching entries, chat records, posts on Internet forums, and microblog messages. In data collection, special techniques are utilized to acquire raw data from a specific environment. A significant factor in the management of scientific data is the capture of data with respect to the transition of raw to published data processes. Data generation is closely associated with the daily lives of people. These data are also similarly of low density and high value. Normally, Internet data may not have value; however, users can exploit accumulated Big Data through useful information, including user habits and hobbies. Thus, behavior and emotions can be forecasted. The problem of scientific data is one that must be considered by Scientific Data Infrastructure (SDI) providers. In the following paragraphs, we explain five common methods of data collection, along with their technologies and techniques.

(i) Log Files. This method is commonly used to collect data by automatically recording files through a data source system.

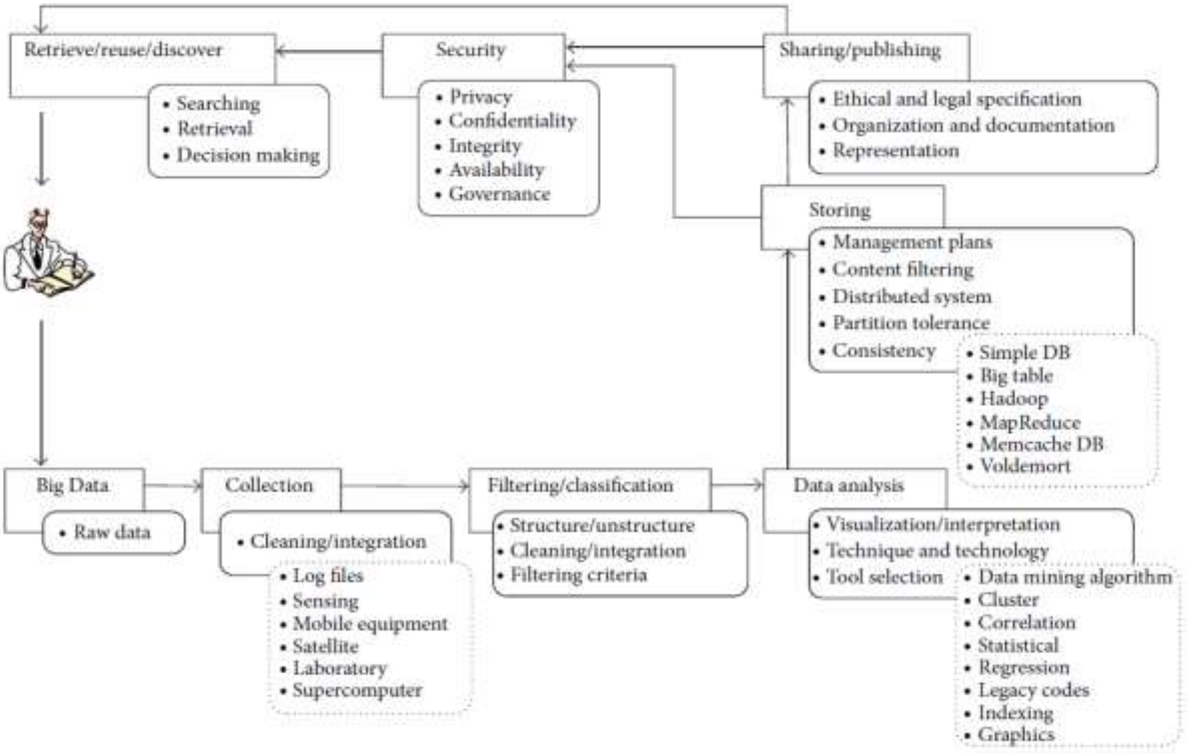

Figure 2: Proposed data life cycle using the technologies and terminologies of Big Data.

5. Opportunities, Open Issues, and Challenges:

According to McKinsey [8] the effective use of Big Data benefits 180 transform economies and ushers in a new wave of productive growth. Capitalizing on valuable knowledge beyond Big Data is the basic competitive strategy of current enterprises. New competitors must be able to attract employees who possess critical skills in handling Big Data. By harnessing Big Data, businesses gain many advantages, including increased operational efficiency, informed strategic

direction, improved customer service, newproducts, and new customers and markets.

With Big Data, users not only face numerous attractive opportunities but also encounter challenges. Such difficulties lie in data capture, storage, searching, sharing, analysis, and visualization. These challenges must be overcome to maximize Big Data, however, because the amount of information surpasses our harnessing capabilities. For several decades, computer architecture has been CPU-heavy but I/O poor. This system imbalance limits the exploration of Big Data. CPU performance doubles every 18 months according to Moore’s Law], and the performance of disk drives doubles at the same rate. However, the rotational speed of the diskshas improved only slightly over the last decade. As a result of this imbalance, random I/O speeds have improved moderately, whereas sequential I/O speeds have increased gradually with density. Information is simultaneously increasing at an exponential rate, but information processing methods are improving relatively slowly. Currently, a limited number of tools are available to completely address the issues in BigData analysis.

The state-of-the-art techniques and technologies in many important Big Data applications (i.e., Hadoop, Hbase, and Cassandra) cannot solve the real problems of storage, searching, sharing, visualization, and real-time analysis ideally. Moreover, Hadoop and MapReduce lack query processing strategies and possess low-level infrastructures with respect to data processing and its management. For large-scale data analysis, SAS, R, and Matlab are unsuitable. Graph lab provides a framework that calculates graph-based algorithms related to machine learning; however, it does not manage data effectively. Therefore, proper tools to adequately exploit Big Data are still lacking.

Challenges in Big Data analysis include data inconsistency and incompleteness, scalability, timeliness, and security. Prior to data analysis, data must be well constructed. However, considering the variety of datasets in Big Data, the efficient representation, access, and analysis of unstructured or semi structured data are still challenging. Understanding the method by which data can be preprocessed is important to improve data quality and the analysis results. Datasets are often very large at several GB or more, and they originate from heterogeneous sources. Hence, current real-world databases are highly susceptible to inconsistent, incomplete, and noisy data. Therefore, numerous data preprocessing techniques, including data cleaning, integration, transformation, and reduction, should be applied to remove noise and correct inconsistencies. Each sub process faces a different challenge with respect to data-driven applications. Thus, future research must address the remaining issues related to confidentiality. These issues include encrypting large amounts of data, reducing the computation power of encryption algorithms, and applying different encryption algorithms to heterogeneous data. Privacy is major concern in outsourced data. Recently, some controversies have revealed how some security agencies are using data generated by individuals for their own benefits without permission. Therefore, policies that cover all user privacy concerns should be developed. Furthermore, rule violators should be identified and user data should not be misused or leaked.

Cloud platforms contain large amounts of data. However, the customers cannot physically assess the data because of data outsourcing. Thus, data integrity is jeopardized. The major challenges in integrity are that previously developed hashing schemes are no longer applicable to such large amounts of data. Integrity checking is also difficult because of the lack of support given remote data access and the lack of information regarding internal storage.

Big Data has developed such that it cannot be harnessed individually. Big Data is characterized by large systems, profits, and challenges. Thus, additional research is needed to address these issues and improve the efficient display, analysis, and storage of Big Data. To enhance such research, capital investments, human resources, and innovative ideas are the basic requirements.

Conclusion

This research the fundamental concepts of Big Data. These concepts include the increase in data, the progressive demand for HDDs, and the role of Big Data in the current environment of enterprise and technology. To enhance the efficiency of data management, we have devised a data-life cycle that uses the technologies and terminologies of Big Data. The stages in this life cycle include collection, filtering, analysis, storage, publication, retrieval, and discovery. All these stages (collectively) convert raw data to published data as a significant aspect in the management of scientific data. Organizations often face teething troubles with respect to creating, managing, and manipulating the rapid influx of information in large datasets. Given the increase in data volume, data sources have increased in terms of size and variety. Data are also generated in different formats (unstructured and/or semi structured), which adversely affect data analysis, management, and storage. This variation in data is accompanied by complexity and the development of additional means of data acquisition.

Reference

[1] Worldometers, "Real time world statistics," 2014, http://www .worldometers.info/world-population/.

[2] D. Che, M. Safran, and Z. Peng, "From Big Data to Big Data Mining: challenges, issues, and opportunities," in *Database Systems for Advanced Applications*, pp. 1–15, Springer, Berlin, Germany, 2013.

[3] M. Chen, S. Mao, and Y. Liu, "Big data: a survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[4] S. Kaisler, F. Armour, J. A. Espinosa, andW.Money, "Big data: issues and challenges moving forward," in *Proceedings of the IEEE 46th Annual Hawaii International Conference on System Sciences (HICSS '13)*, pp. 995–1004, January 2013.

[5] R. Cumbley and P. Church, "Is "Big Data" creepy?" *Computer Law and Security Review*, vol. 29, no. 5, pp. 601–609, 2013

[6] S. Hendrickson, *Getting Started with Hadoop with Amazon's Elastic MapReduce*, EMR, 2010.

[7] M. Hilbert and P. L´opez, "The world's technological capacity to store, communicate, and compute information," *Science*, vol. 332, no. 6025, pp. 60–65, 2011.

[8] J. Manyika, C. Michael, B. Brown et al., "Big data: The next frontier for innovation, competition, and productivity," Tech.Rep., Mc Kinsey, May 2011.

[9] J. M. Wing, "Computational thinking and thinking about computing," *Philosophical Transactions of the Royal Society of London A:Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1881, pp. 3717–3725, 2008.

[10] J.Mervis, "Agencies rally to tackle big data," *Science*, vol. 336, no.6077, p. 22, 2012.

[11] K. Douglas, "Infographic: big data brings marketing big numbers," 2012, http://www.marketingtechblog.com/ibm-big-datamarketing/.

[12] S. Sagiroglu andD. Sinanc, "Big data: a review," in *Proceedings of the International Conference on Collaboration Technologies and Systems (CTS '13)*, pp. 42–47, IEEE, San Diego, Calif, USA,May 2013.