Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2016, 3(2):106-108**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Big Data: An Introduction for Engineers

Matthew N. O. Sadiku, Mahamadou Tembely, Sarhan M. Musa

[1]Roy G. Perry College of Engineering Prairie View A&M University

Abstract Big data refers to massive amount of data that are so large that traditional processing tools cannot cope. It is produced by emails, online transactions, medical records, etc. It is a high-volume, high-velocity, and high-variety information that require special information processing tools. This paper presents a brief introduction on big data for engineers.

Keywords

Introduction

There is data everywhere. Each day, large amounts of data are being generated. Organizations and companies now have very large data sets stored in their files, databases, and data warehouses. Advancement in sciences and technologies has collectively created a massive amount of structured and unstructured data [1]. Big data is typically unstructured (such as text, audio, video) or semi-structured (such as emails, tweets, weblogs). The ability to collect and analyze huge amounts of data is a growing problem within the engineering community.
The birth of the concept of big data is usually associated with a META Group report by Doug Laney [2].

Definition of Big Data

Big data applies to data sets of extreme size (e.g. exabytes, zettabytes) which are beyond the capability of the commonly used software tools. It involves situation where very large data sets are big in volume, velocity, veracity, and variability [3]. The data is too big, too fast, or does not fit the regular database architecture. It may require different strategies and tools for profiling, measurement, assessment, and processing.
The process of examining big data is often referred to Big Data Analytics. It is an emerging field since massive computing capabilities have been made available by e-infrastructures [4]. Analytics include statistical models and other methods that are aimed at creating empirical predictions. Data-driven organizations use analytics to guide decisions at all levels.

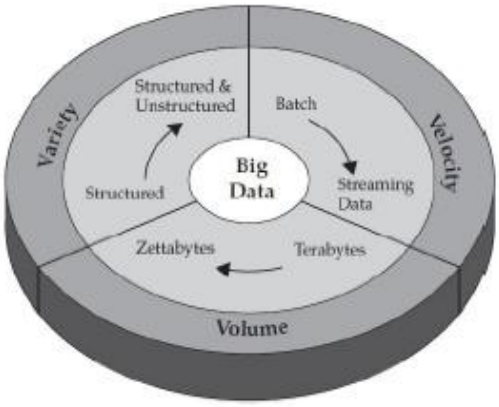

Figure 1: The volume, velocity, and variety [5].

Big data is growing rapidly and expanding in all science and engineering, including physical, biological, and medical services. Different companies use different means to maintain their big data. Big data is characterized by 3 V (volume, velocity, and variety). *Volume* captures the large amount of data that is being created. Systems nowadays are inundated with terabytes or petabytes of information. Organizations and companies are generating data at an exponential rate. *Velocity* refers to the speed of data processing or how quick the data is generated and processed. *Variety* refers to the diversity of data models and sources and lack of uniform structure in the data. Data is generated in different domains, from social media to transportation, from healthcare to wireless communication networks. Structured and unstructured data are generated in various types [5-6]. The 3V model has been extended to a 5V model, where the other 2Vs represent Value and Veracity [7].

Data Quality and Security

Quality is an important issue for all data. For any big data project to succeed, it must depend on high quality data. Quality issues include accuracy, completeness, consistency, precision, relevance, and timeliness. Big data quality varies from one type of big data to another. Data are a neutral measure of reality, but people corrupt data and use them to achieve their own ends.

The rise of big data raises fundamental challenges in privacy, security, and data ownership. Concerns are being expressed over the impact that collecting, storing, and processing large amount of data could have on security. Security of big data is a primary concern for many applications. From security point of view, big data may seriously weaken confidentiality. Security is a concern because of the variety and heterogeneity of big data; there is access to data from multiple and diverse domains. Making effective use of big data requires that access only from domain it is authorized to access. For instance, in case of smart meters, data must be protected to avoid leaking private information about consumers. Also, in outsourcing data to the cloud, the owner may lose control and privacy of the data

Application Domains

Several domains can benefit from the big data phenomena: medicine, education, manufacturing, communication, government, and industry. Big data is becoming a new technology focus in science, healthcare, business, and industry.

Science has been dealing with large volume of data in research experiments. Scientific research includes collection of data which aim to verify some scientific hypothesis. For example, a huge number of chemical structures are stored in public and private databased. Consequently, a big data problem in chemistry has appeared [8].

Health care applications of big data are helpful for improving clinical decision-making and care provision. Big data is driving the development in biomedical and healthcare informatics because big data has unlimited potential for storing, processing, and analyzing medical data. For example, big data can be leveraged to detect fraud, abuse, and errors in health insurance claims.

Business is the major source of big data. Service supply chain such as finance, healthcare, tourism, and telecommunications drive big data. In business, big data may be considered as cost-effective techniques for solving business problems whose resource requirements exceed the capabilities of traditional computing environment. Although companies may share third party facilities, such as clouds, they do not share data but ensure that their data is protected. Applying big data in business can enhance efficiency and competitiveness in many aspects such as marketing, supply chain, and e-commerce.

In industry, big data may involve controlling technological processes and facilities.

Today's organizations are handling increasing amounts and complexities of data. Modern computer-aided manufacturing produces huge amount of data which may need to be stored to allow effective quality control.

In addition, social networking and government systems also contain large amount of data. Big data applications are gaining momentum as more companies seek to monetize the data and move their business forward. Several techniques have been proposed for analyzing big data. These include the HACE theorem, cloud computing, Hadoop, and MapReduce [9].

Conclusions

Big data refers to large data that is generated in complex systems with the characteristics of three Vs: volume, velocity, and variety. It is also the technologies that make processing and analyzing it possible. It has attracted a growing attention from industry and academia. It has emerged as a full-fledged field. Big data may be regarded as a phenomenon since we can observe its effects like growing volume and variety. It has been regarded as the "future petroleum" and the "gold mine" to be developed [10]. It is a pretty new research area. It should not be ignored by organizations and engineers since it is here to stay. It will become intensive and diversified in the

years to come. Scholars can stay up to date on issues related to big data by consulting Institute of Electrical and Electronics Engineers (IEEE) Transactions on Big Data [e.g. 11].

References

1. Jukić, N., Sharma, A., Nestorov, S., & Jukić, B. (2015). Augmenting Data Warehouses with Big Data. *Information Systems Management*, *32*(3), 200-209.
2. WAZIRI, V. O., ALHASSAN, J. K., Morufu, O., & Ismaila, I. Big Data Analytics and the Epitome of fully Homomorphic Encryption Scheme for Cloud Computing Security. *International Journal of Developments in Big Data and Analytics Voolume*, *1*, 19-40.
3. Swan, M. (2015, March). Philosophy of Big Data: Expanding the Human-Data Relation with Big Data Science Services. In *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on* (pp. 468-477). IEEE.
4. Baumann, P., Mazzetti, P., Ungar, J., Barbera, R., Barboni, D., Beccati, A., ... & Campalani, P. (2015). Big data analytics for earth sciences: the EarthServer approach. *International Journal of Digital Earth*, 1-27.
5. Tiwari, A. K., Chaudhary, H., & Yadav, S. (2015, March). A review on Big Data and its security. In *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on* (pp. 1-5). IEEE.
6. Hoy, M. B. (2014). "Big data: an introduction for librarians," *Medical Reference Services Quarterly*, 33(3), 320-326.
7. Viceconti, M., Hunter, P., & Hose, R. (2015). Big Data, Big Knowledge: Big Data for Personalized Healthcare. *Biomedical and Health Informatics, IEEE Journal of*, *19*(4), 1209-1215.
8. Chen, Y., Chen, H., Gorkhali, A., Lu, Y., Ma, Y., & Li, L. (2016). Big data analytics and big data science: a survey. *Journal of Management Analytics*, *3*(1), 1-42.
9. Wu, X., Chen, H., Wu, G., Liu, J., Zheng, Q., He, X., ... & Li, Y. (2015). Knowledge Engineering with Big Data. *Intelligent Systems, IEEE*, *30*(5), 46-55.
10. Miao, X., & Zhang, D. (2014, September). The opportunity and challenge of Big Data's application in distribution grids. In *Electricity Distribution (CICED), 2014 China International Conference on* (pp. 962-964). IEEE.
11. Zheng, Y. (2015). Methodologies for cross-domain data fusion: an overview. *Big Data, IEEE Transactions on*, *1*(1), 16-34.

About the authors

Matthew N.O. Sadiku (sadiku@ieee.org) is a professor at Prairie View A&M University, Texas. He is the author of several books and papers. He is a fellow of IEEE.

Mahamadou Tembely (mtembely@student.pvamu.edu) is a Ph.D student at Prairie View A&M University, Texas. He received the 2014 Outstanding MS Graduated Student award for the department of electrical and computer engineering. He is the author of several papers.

Sarhan M. Musa (smmusa@pvamu.edu) is an associate professor in the Department of Engineering Technology at Prairie View A&M University, Texas. He has been the director of Prairie View Networking Academy, Texas, since 2004. He is an LTD Spring and Boeing Welliver Fellow.