Service Oriented Computing and Applications (2019) 13:105–107
https://doi.org/10.1007/s11761-019-00265-x

EDITORIAL

Research challenges of big data

Muhammad Younas[1]

Received: 3 June 2019 / Accepted: 5 June 2019 / Published online: 14 June 2019

Abstract
Big data is characterised by new characteristics such as 3Vs (Volume, Velocity, Variety), and/or 5Vs (Volume, Velocity, Variety, Veracity, and Value). Due to the distinguishing characteristics of big data, it is commonly stored and processed using NoSQL (Not Only SQL) database systems. Big data has been utilised in various applications and services ranging from E-commerce through to social media to public sector and governmental organisations. The goal of this editorial note is to provide a concise summary of the big data characteristics, models, and technologies and to identify some of the crucial research challenges that are open for further research.

1 Characteristics of big data

Big data refers to the large volume of complex, (semi) structured, and unstructured data that are generated in a large size and that arrive (in a system) at a higher speed so that it can be analysed for better decision making and strategic organisation and business moves. But the process of managing (only) large volume of data is not new. For example, one of the top-ranking database conferences on Very Large Databases (VLDB) has been running for more than 40 years. The proceedings of VLDB include a number of articles that provide useful solutions for managing large volume of complex data. But the concept of big data has gained popularity with the new applications and new characteristics such as 3Vs (Volume, Velocity, Variety), and/or 5Vs (Volume, Velocity, Variety, Veracity, and Value). Figure 1 [1] shows a generic view of the 5Vs characteristics and applications of big data. These are briefly described as follows [2]:

Volume This refers to the massive amount of data which are being generated, gathered, and processed, for example, in the size of petabytes, exabytes, and zettabytes. For instance, Twitter receives/processes millions of tweets on a regular basis. Similarly, Facebook routinely handles millions of posts and images. Google receives more than a billion search queries. Further, millions of data records are gathered from sensor technologies associated with transportation, weather, environmental systems, and so on.

Velocity This refers to the speed at which data are generated, processed, and moved between different systems and devices. Examples include the speed of social media posts; online transactions and fraud checking; live transportation data received from buses, trains, aeroplanes, etc.

Variety This refers to the different types of data that can be used (together) for achieving desired information or results. Types and format of big data include structured, semi-structured, and unstructured data.

Veracity This refers to the quality of data such as correctness, consistency, trust, security, and reliability. For example, data are not stale or out of date for a given purpose. Similarly, data should be correct and consistent and it should be generated by a trusted system.

Value This refers to the different types of benefits that can be derived from processing and analysing big data. Examples include, monetary value, social value, research/education value, and so on.

2 Big data models and technologies

Classical relational models and SQL technologies do not appropriately cater for the needs of big data due to its distinguishing characteristics as illustrated above. Thus, storing and processing of big data require new data models and new technologies. The most commonly used data models for big

✉ Muhammad Younas
m.younas@brookes.ac.uk

[1] School of Engineering, Computing and Mathematics, Oxford Brookes University, Oxford, UK

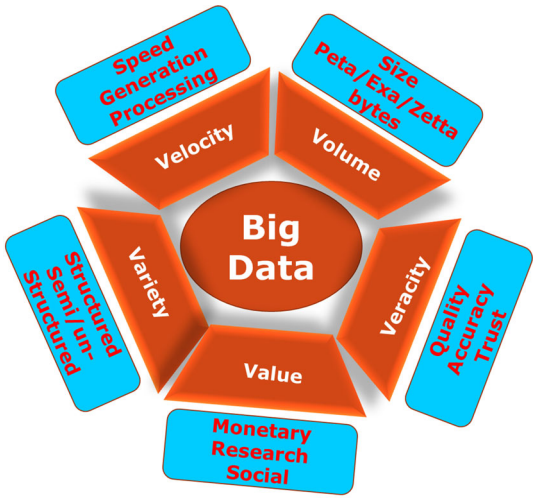

Fig. 1 Big data characteristics and applications

Fig. 2 Big data models and technologies

data are document model; key-value model; column model; and graph model.

Big data is commonly stored and processed using cloud-based NoSQL systems such as Riak, MongoDB, Google Cloud Big Table, and Amazon DynamoDB. As shown in Fig. 2 [1], CouchDB, MongoDB, and Azure Cosmos DB generally follow the document model. Key-value model is followed by NoSQL databases such as Riak, Amazon DynamoDB, and Cassandra. Column model is implemented in MariaDB, Apache HBASE, and Google Cloud Big Table. Graph-based models are adopted in Neo4j, TITAN, and OrientDB. Note that this is rather a broad classification and some of these NoSQL systems may belong to different (or multiple) data models.

3 Big data challenges

From the above discussion, it is observed that big data is characterised by new characteristics, new data models, and new database technologies. All such new developments have been greatly benefiting companies and organisations in various dimensions such as time and cost saving, intelligent decision making, effective product design and development, and helping in customer relationships, to name but a few. Despite significant developments in big data systems, various challenges are still open for further research. For instance, past literature has identified various challenges related to big data [3]. This editorial provides a brief description of some of the crucial research challenges of big data.

- NoSQL databases are predominantly used to store and process big data. Such databases provide key benefits such as efficiency, scalability, and availability in storing and processing big data. However, they do not provide appropriate support for transactions, data normalisation, and integrity constraints which affect the consistency of big data [4, 5]. Thus, the current models and techniques implemented in NoSQL databases should be re-examined so that they can be used in applications/services that demand strong data consistency in addition to high efficiency, scalability, and availability.
- A number of NoSQL databases have been designed and developed. Different NoSQL systems are implemented using different big data models and technologies. They also provide varying level of QoS with respect to performance, availability, and scalability. This makes the selection of a NoSQL database difficult—i.e. which NoSQL system is chosen for a particular use or application of a big data. This requires the design and development of a new benchmark which users/developers can use to select appropriate NoSQL database that effectively meets their needs.
- Data as a Service (DaaS) has emerged as a new platform in order to facilitate the provisioning of data over the Internet and cloud. DaaS is generally based on web services and service-oriented computing (SOC) technologies. DaaS aims to consolidate and organise data in a centralised place in order to enable location transparency as well as sharing of data across different systems and services. However, existing models and architectures of web services and SOC may fall short of meeting the requirements of DaaS provisioning over the Internet and cloud. Thus, new models, methods, and architectures should be developed in order to further materialise the benefits of DaaS.
- Internet of Things (IoT) is one of the major platforms (and a source) for big data given that millions of things or devices are generating and consuming a large volume of big data. However, resource scarcity is one of the major issues associated with the IoT devices as they do not have the capabilities of collecting, storing, analysing, and sharing big data in (real) time. Thus, new solutions are required to be developed in order to effectively conjoin IoT and big data.

References

1. Younas M (2018) Transactional services for NoSQL big data systems. In: Keynote talk at the 6th international conference on multimedia computing and systems (ICMCS 2018), Rabat, Morocco, 10–12 May 2018
2. Nguyen TL (2018) A framework for five big v's of big data and organizational culture in firms. In: Proceedings of the IEEE international conference on big data (Big Data 2018), Seattle, WA, USA, 10–13 Dec 2018
3. Jin X, Wah BW, Cheng X, Wang Y (2015) Significance and challenges of big data research. Int J Big Data Res 2:59–64
4. González-Aparicio MT, Younas M, Tuya J, Casado R (2018) Testing of transactional services in NoSQL key-value databases. Int J Future Gener Comput Syst 80:384–399
5. Padhye V, Tripathi A (2015) Scalable transaction management with snapshot isolation for NoSQL data storage systems. IEEE Trans Serv Comput 8:121–135