

Big Data Analytics and Business Intelligence in Industry

Shih-Chia Huang¹ · Suzanne McIntosh² · Stanislav Sobolevsky³ · Patrick C. K. Hung⁴

© Springer Science+Business Media, LLC 2017

1 Introduction

The pervasive nature of digital technologies as witnessed in industry, services and everyday life has given rise to an emergent, data-focused economy stemming from many aspects of human individual and business activity. The richness and vastness of these data are creating unprecedented research opportunities in several fields including urban studies, geography, economics, finance, and social science, as well as physics, biology and genetics, public health and many others.

Big data is the term for a collection of large and complex datasets from different sources that are difficult to process using traditional data management and processing applications. Big data is the description of a large amount

of either organized or unorganized data that is analyzed to make an informed decision or evaluation. The data can be taken from a large variety of sources including browsing history, geolocation, social media, purchase history and medical records. Big data consists of complex data that would overwhelm the processing power of traditional simple database systems (Hung 2016). There are three main characteristics associated with big data (Dave 2013):

- Volume is a characteristic used to describe the vast amounts of data that is utilized by big data. The amounts of data usually range from gigabytes to yottabytes. The big data should be able to handle any amount of data even as it expectedly grows exponentially.
- Variety is a characteristic used to describe the many different types of data sources that are used as part of a big data analytics system. There are multiple data storage formats that are utilized by computer devices throughout the world. There are structured data such as databases, .csv, video, Short Message Service (SMS) and excel sheets. The unstructured data could be in the form such as handwritten notes. All the data from these sources would ideally be used to a big data analytics system.
- Velocity is a characteristic used to describe the speed at which data is generated. It is also used to describe the speed at which the generated data is processed. With a click of a button, an online retailer can quickly view large data about a certain customer. Velocity is also important to ensure that data is current and updated in real-time, thus allowing the system to perform the best it can. This speed is essential as real-time data generation helps organizations speed up operations processes; which can save organizations a large amount of money.

✉ Patrick C. K. Hung
patrick.hung@uoit.ca

Shih-Chia Huang
schuang@ntut.edu.tw

Suzanne McIntosh
mcintosh@cs.nyu.edu

Stanislav Sobolevsky
sobolevsky@nyu.edu

¹ Department of Electronic Engineering, National Taipei University of Technology, Taipei City, Taiwan

² Courant Institute of Mathematical Sciences, Center for Data Science, New York University, New York, NY, USA

³ Center for Urban Science and Progress, New York University, New York, NY, USA

⁴ Faculty of Business and Information Technology, University of Ontario Institute of Technology, Oshawa, ON, Canada

Recently Firican (2017) presents seven more characteristics for the current big data scenario shown as follows:

- Variability means the number of inconsistencies in the data because of the multitude of data variable and dimensions resulting from multiple disparate data types and sources.
- Veracity refers to the provenance or reliability of the data source in the context of meaningful analysis.
- Validity refers to accuracy and correctness of the data is for its intended use.
- Vulnerability refers to the security concerns in the data source.
- Volatility refers to the rules for data currency and availability.
- Visualization means the format and layout of data analysis results.
- Value refers to the meaningful context the data source brings.

Many businesses today are increasingly interested in utilizing big data technologies for supporting their business intelligence so that it is becoming more and more important to understand the various practical issues from past experiences in business intelligence systems. Business intelligence systems today are sensing the world and harnessing those data points to recommend best possible options and to accurately predict outcomes. While non-real-time business intelligence systems continue to be built, the demand to collect, integrate, process, and visualize near real time data is on the rise. Business intelligence systems are advantaged by a rise in sensing opportunities as witnessed in both the number of sensors and the rich diversity of sensors ranging from cell phones, personal computers, and health tracking appliances to Internet of Things (IoT) technologies designed to give contextual, semantic voice to entities that previously could not contribute intelligence to key decisions.

Many companies are conducting big data analytics nowadays. Big data analytics is a process of examining diverse, large-scale data sets to uncover patterns, unknown correlations, and other useful information. Big data analytics employ software tools from advanced analytics disciplines such as data mining, predictive analytics, and machine learning. For example, big data analytics has become very popular in marketing, driving up value by understanding and engaging customers more effectively. There are many industries that have adopted the use of big data analytics and are experiencing fantastic results; the healthcare, retail, insurance, and telecommunications industries have all displayed the endless possibilities of implemented big data into their operations. Another example is Google Flu Trend; a service that predicts and locates flu outbreaks. This service uses aggregate search queries for processing which has the potential to reduce the impact of influenza (Tene and Polenetsky 2012).

In addition to big data-inspired research, businesses and policy makers have seized on big data technologies to support their decisions and propel growing business intelligence needs. As businesses build out big data hardware and software infrastructure, it becomes increasingly important to anticipate technical and practical challenges and to identify best practices learned through experience. Companies that use big data analytics outperform in their industry. This data allows companies to identify their most profitable customers and enables them to reach out to them on a personal level, enabling a better customer experience. This technology also allows organizations to improve the speed of processes, better understand customer needs, and improve innovation. By properly understanding customer needs, companies know exactly what type of innovations will be successful; where they should invest in with their Research and Development (R&D) department (IBM 2014).

The goal of this special issue is to present different use cases in big data applications and share the related practical experiences with the readers. This special issue contains extended versions of the best papers from the IEEE International Congress on Big Data sponsored by the Technical Committee on Services Computing, IEEE Computer Society, as well as high-quality papers from an open call for papers. There are eight papers, each of which is concerned with a specific and industrial aspect of big data analytics and business intelligence and discussed in the following paragraphs.

2 Papers in the Special Issue

In this special issue, most of the papers assume that (1) the data sources are stored in Cloud (Ren et al. 2015); and (2) the data sources should be encrypted before outsourcing to Cloud storage for privacy requirements, which makes the traditional and efficient plain text keyword search technique useless (Fu et al. 2015; Xia et al. 2016). Referring to the first paper “Revealing Determinant Factors for Early Breast Cancer Recurrence by Decision Tree,” by Guo et al. (2017) present a Decision Tree algorithm to the clinical information of a cohort of non-metastatic invasive breast cancer patients, to establish a classifier that categorizes patients based on whether they develop early recurrence and on similarities of their clinical and pathological diagnoses. As a result, this paper identified pathological nodal stage, the percentage of intra-tumor stroma and components of TGF β -Smad signaling pathway as highly relevant factors for early breast cancer recurrence.

Next, the second paper “Grammatical facial expression recognition in sign language discourse: a study at the syntax level,” by Freitas et al. (2017) present a study which applies inductive reasoning to recognize patterns, to study the problem involving the automated recognition of Grammatical Facial Expressions (GFEs) at the discourse syntactic level in

the Libras (Brazilian) Sign Language. In general, GFEs stand out in automated recognition processes for sign languages, as they help to remove ambiguity among signals, and they also contribute to compose the semantic meaning of discourse.

Referring to the third paper “Research on the Influence of Inter-user Relationships on Social Commerce Platforms on the Value Proposition of Electronic Commerce Enterprises,” by Zheng et al. (2017) discuss the impact of social commerce on the participants from the individual consumers to electronic commerce enterprises, reveals the influence of inter-user relationships on social commerce platforms on the value proposition of electronic commerce enterprises theoretically and is good for electronic commerce enterprises in social network marketing practically.

Then, Huang et al. (2017) present the fourth paper “Reliable and Efficient Big Service Selection” that describes a big service selection approach based on the coefficient of variation and mixed integer programming that improves the solution in two senses: (1) minimizing the time cost and (2) maximizing the reliability. The selection approach was tested on real-world datasets, and the experimental results indicated that the approach is more effective than others.

Further, Chu et al. (2017) present the fifth paper “The Design of a Cloud-based Tracker Platform based on System-of-systems Service Architecture.” Chu et al. present the design of a generic, programmable position tracking platform, namely CQtracker. This platform is incorporated with a cloud-based engine of advanced analytics. CQtracker is constructed based on a concept of system-of-systems service architecture to deliver data-system-as-a-service.

Then, the sixth paper “Analyzing Social Choice and Group Ranking of Online Games for Product Mix Innovation,” by Tung and Lan (2017) discuss an integrative analysis from many data-users’ preference sequences. Firstly, this research utilizes a sequence recommendation technology that can be used to analyze their preference sequences of online game’s types based on other similar users’ sequence preference data. Secondly, this research also utilizes data mining methodology to explore the possibilities of different product mix of game types from numerous user data.

Referring to the seventh paper “Crowdsourcing Social Network Service for Social Enterprise Innovation,” Tung and Jordann (2017) present an integrative platform of social media and crowdsourcing service to support business innovation of Social Enterprise (SE). Such an SE-specific crowdsourcing social network service (SNS) facilitate people can obtain the shared resources to further achieve the projects or even start up an innovative SE.

Finally, in the eighth paper “A Study on the Author Collaboration Network in Big Data,” by Peng et al. (2017) investigates the author collaboration groups and the core author collaboration groups as well as the collaboration trends in big data by combining bibliometric analysis and social

network analysis. The results of this paper can be useful for researchers interested in finding suitable partners in the big data field.

In summary, the papers in this special issue illustrate some of the current research areas pertinent to big data and business intelligence; while, in many ways, also amplifying the many challenges, which remain to be addressed. It is expected that new topics will emerge while existing research will shift concentration into this information systems frontier in the coming years.

Acknowledgements We would like to acknowledge the program committee members who provided high-quality reviews and constructive feedback to paper authors.

This work was supported by the Ministry of Science and Technology (MOST), Taiwan, under MOST Grants: NTUT-KMITL-106-03, 105-2811-E-027 -001, and 104-2221-E-027 -020; and the Natural Sciences and Engineering Research Council of Canada (NSERC), under Discovery Grants Program.

References

- Chu, V. W., Wong, R. K., Chi, C.-H., Zhou, W., & Ho, I. (2017). The Design of a Cloud-Based Tracker Platform based on system-of-systems service architecture. *Information Systems Frontiers*, 19(6). <https://doi.org/10.1007/s10796-017-9768-9>.
- Dave, P. (2013). What is big data - 3 vs of big data. Retrieved from SQL Authority Blog: <http://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/>.
- Firican, G. (2017). The 10 vs of big data, UPSIDE where DATA means BUSINESS. Retrieved from: <https://upside.tdwi.org/Articles/2017/02/08/10-Vs-of-Big-Data.aspx?Page=1>.
- Freitas, F. A., Peres, S. M., Lima, C. A. M., & Barbosa, F. V. (2017). Grammatical facial expression recognition in sign language discourse: A study at the syntax level. *Information Systems Frontiers*, 19(6). <https://doi.org/10.1007/s10796-017-9765-z>.
- Fu, Z., Sun, X., Liu, Q., Zhou, L., & Shu, J. (2015). Achieving efficient cloud search services: Multi-keyword ranked search over encrypted cloud data supporting parallel computing. *IEICE Transactions on Communications*, E98-B(1), 190–200.
- Guo, J., Fung, B. C. M., Iqbal, F., Kuppen, P. J. K., Tollenaar, R. A. E. M., Mesker, W. E., & Lebrun, J.-J. (2017). Revealing determinant factors for early breast cancer recurrence by decision tree. *Information Systems Frontiers*, 19(6). <https://doi.org/10.1007/s10796-017-9764-0>.
- Huang, L., Zhao, Q., Li, Y., Wang, S., Lei, S., & Chou, W. (2017). Reliable and efficient big service selection. *Information Systems Frontiers*, 19(6). <https://doi.org/10.1007/s10796-017-9767-x>.
- Hung, P. C. K. (2016). *Big data applications and use cases, the springer international series on applications and trends in computer science*. Switzerland: Springer International Publishing AG.
- IBM. (2014). Better business outcomes with IBM Big Data & Analytics. Retrieved from IBM: http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/59898_Better%20Business%20Outcomes_White%20Paper_Final_NIW03048-USEN-00_Final_Jan21_14.pdf.
- Peng, Y., Shi, J., Fantinato, M., & Chen, J. (2017). A study on the author collaboration network in big data. *Information Systems Frontiers*, 19(6). <https://doi.org/10.1007/s10796-017-9771-1>.

- Ren, Y., Shen, J., Wang, J., Han, J., & Lee, S. (2015). Mutual verifiable provable data auditing in public cloud storage. *Journal of Internet Technology*, 16(2), 317–323.
- Tene, O., & Polenetsky, J. (2012). Data, privacy in the age of big data. Retrieved from Stanford Law Review: <https://www.stanfordlawreview.org/online/privacy-paradox/big-data>.
- Tung, W.-F., & Jordann, G. (2017). Crowdsourcing social network Service for Social Enterprise Innovation. *Information Systems Frontiers*, 19(6). <https://doi.org/10.1007/s10796-017-9770-2>.
- Tung, W.-F., & Lan, Y.-J. (2017). Analyzing social choice and group ranking of online games for product mix innovation. *Information Systems Frontiers*, 19(6). <https://doi.org/10.1007/s10796-017-9769-8>.
- Xia, Z., Wang, X., Sun, X., & Wang, Q. (2016). A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data. *IEEE Transactions on Parallel and Distributed Systems*, 27(2), 340–352.
- Zheng, C., Yu, X., & Jin, Q. (2017). How user relationships affect user perceived value propositions of enterprises on social commerce platforms. *Information Systems Frontiers*, 19(6). <https://doi.org/10.1007/s10796-017-9766-y>.

Shih-Chia Huang is a Professor with the Department of Electronic Engineering at National Taipei University of Technology, Taiwan, and an International adjunct professor with the Faculty of Business and Information Technology, University of Ontario Institute of Technology, Canada. Professor Huang has published more than 80 journal and conference papers and holds more than 60 patents in the United States, Europe, Taiwan, and China. In 2009, he received a doctorate degree in Electrical Engineering from National Taiwan University, Taiwan. Dr. Huang was presented with the Kwoh-Ting Li Young Researcher Award in 2011 by the Taipei Chapter of the Association for Computing Machinery, as well as the Dr. Shechtman Young Researcher Award in 2012 by National Taipei University of Technology. In addition, Professor Huang is an associate editor of the Journal of Artificial Intelligence. His research interests include image and video coding, wireless video transmission, video surveillance, error resilience and concealment techniques, digital signal processing, cloud computing, mobile applications and systems, embedded processor design, and embedded software and hardware co-design.

Suzanne McIntosh is a Visiting Professor at New York University Courant Institute of Mathematical Sciences and an Affiliated Professor with NYU Center for Data Science. Professor McIntosh has led cross-disciplinary research teams in virtualization, data center energy optimization analytics, and security efforts which yielded innovative architectures, patents, and publications. She is an inventor with patents in security and virtualization technologies for which she was awarded two IBM Research Invention Plateau awards and two

IBM Research Technical Accomplishment Awards for Open Source Contributions to Virtualization Security, and for the development of the Caernarvon High Assurance Operating System. Professor McIntosh is an editor and chair for IEEE and ACM cloud, Big Data, and security conferences and maintains research interests in distributed computing and supercomputing architectures through long-standing service to the annual ACM Supercomputing conference, the ACM SIGHPC executive committee, and the ACM SIGHPC-BigData chapter, which she founded. She earned an Engineer Degree in Computer Engineering from Stevens Institute of Technology (2000), following an MS and BS in Computer Science. Professor McIntosh is an ACM and IEEE Senior Member and serves as a STEM mentor through SWE and the New York Academy of Sciences.

Stanislav Sobolevsky received his Ph.D. degree in 1999 and Doctor of Science (Habilitation degree) in 2009 in mathematics in Belarus. He has been an Associate Professor of Practice at the Center for Urban Science and Progress at New York University since 2015. His former work experience includes research, faculty and administrative positions at MIT, Belarusian State University and Academy of Science of Belarus. Dr. Sobolevsky applies his fundamental quantitative background to studying human behavior in urban context and a city as a complex system through its digital traces—spatiotemporal big data created by various aspects of human activity. He authored over a 100 of research papers in the top journals like PNAS, Scientific Reports, Physical Review E, PLoS ONE, Royal Society Open Science, EPJ Data Science, Applied Geography, Environment and Planning B, International Journal of GIS, Studies In Applied Mathematics and others. Dr. Sobolevsky's research is conducted in close cooperation with city agencies and industrial partners from banking, telecom, defense, insurance and other areas.

Patrick C. K. Hung is a Professor at the Faculty of Business and Information Technology in University of Ontario Institute of Technology, Canada. Patrick has been working with Boeing Research and Technology at Seattle on aviation services-related research with two U.S. patents on mobile network dynamic workflow system. He currently works with the College of Technological Innovation at Zayed University on several smart cities and cyber security research projects in the United Arab Emirates. He is also a Visiting Researcher at University of São Paulo, Brazil and National Technological University (UTN)-Santa Fe, Argentina. He is an Honorary International Chair Professor at National Taipei University of Technology in Taiwan and an Adjunct Professor at Nanjing University of Information Science & Technology in China. Before that, he was a Research Scientist with Commonwealth Scientific and Industrial Research Organization in Australia as well as he worked as a software engineer in industry in North America. He is a founding member of the IEEE Technical Committee on Services Computing, the IEEE International Congress of Services and the IEEE Transactions on Services Computing. He is a Coordinating Editor of the Information Systems Frontiers.