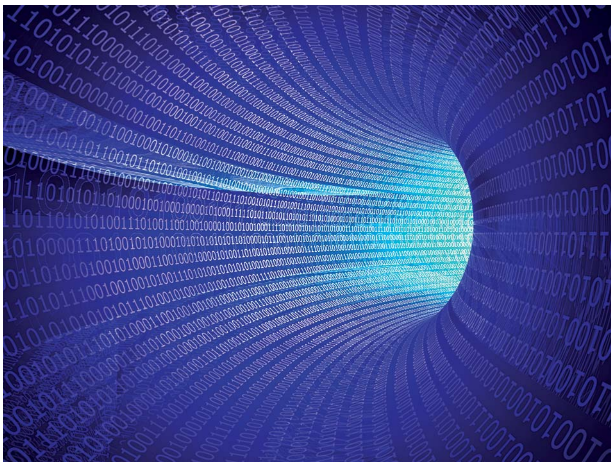
Big Data for Modern Industry: Challenges and Trends

By SHEN YIN
Research Institute of Intelligent Control and Systems, Harbin Institute of Technology, Harbin 150001, China

OKYAY KAYNAK
Department of Electrical and Electronic Engineering, Bogazici University, Istanbul 80815, Turkey



We are living in an era of data deluge and as a result, the term "big data" is appearing in many contexts, from meteorology, genomics, complex physics simulations, biological and environmental research, finance and business to healthcare. One interesting example is that a press release of SAP AG, dated 11 June 2014 reported, "SAP and the German Football Association turn big data into smart decisions to improve player performance at the World Cup in Brazil." An International Data Corporation (IDC) report [1] predicts that "from 2005 to 2020, the digital universe will grow by a factor of 300, from 130 Exabyte to 40 000 Exabyte" and that "from now until 2020 will about double every two years." As the name implies, big data literally means large collections of data sets containing abundant information. However, it has some special characteristics that distinguish it from "very large data" or "massive data" that are simply enormous collections of simple-format records, typically equivalent to enormous spreadsheets. Big data, being generally unstructured and heterogeneous, is extremely complex to deal with via traditional approaches, and requires real-time or almost real-time analysis. A short definition can therefore be that "big data" refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. For a thorough discussion on various aspects of big data and the challenges it presents, together with some potential research directions, the reader is referred to [2].

I. GROWING SIGNIFICANCE OF BIG DATA IN INDUSTRY

For modern industry, data generated by machines and devices, cloud-based solutions, business management, etc., has reached a total volume of more than 1000 Exabytes annually and is expected to increase 20-fold in the next ten years. McKinsey & Company reports that "manufacturing stores more data than any other sector—close to 2 Exabyte of new data stored in 2010" [3]. For example, in a consumer packaged goods company that produces a personal care product, a single machine alone generates 5000 data samples every 33 ms, resulting in four trillion (i.e., 4 Tera) samples per year [4]. Such facts have inaugurated the big data epoch, which

Digital Object Identifier: 10.1109/JPROC.2015.2388958

is further fueled by the recently emerging areas such as the Cloud, the Internet of Things, and the Cyber-Physical Systems.

From an industry point of view, big data is going to play an important role in the fourth industrial revolution [5]. The first industrial revolution (from the end of 18th to the start of 20th centuries) depended on water and steam power, the second (from the start of 20th century to early seventies) depended on mass production, based on division of labor and electrical energy, and the third (from early seventies to the present day) depended on electronics and on IT for further automation of production. The fuel of the fourth industrial revolution, dubbed as "Industrie 4.0" by the German Government, will be big data to be made available through Cyber-Physical Systems (CPS). The goal is to realize smart factories, in which machines and resources communicate as in a social network. Such a smart factory will produce intelligent products (smart products) that know how they have been produced, and will collect and transmit data as they are being used; these huge amounts of data (big data) will be collected and analyzed in real time. New insights will thus be generated and used to move one level up from smart factories to smart processes, and finally, to the level at which intelligent services can be provided to the customer through internet-based services.

The primary objective behind the use of big data in industrial applications is to achieve a fault-free and cost efficient running of the process, while realizing the desired performance levels, especially with respect to quality. McKinsey suggests that manufacturers could make up to a 50% decrease in product development and assembly costs, and up to a 7% reduction in working capital through the use of big data. At a higher level, the data sent by the smart devices can help the manufacturer to pinpoint the preferences of the consumers and, thus, shape future products.

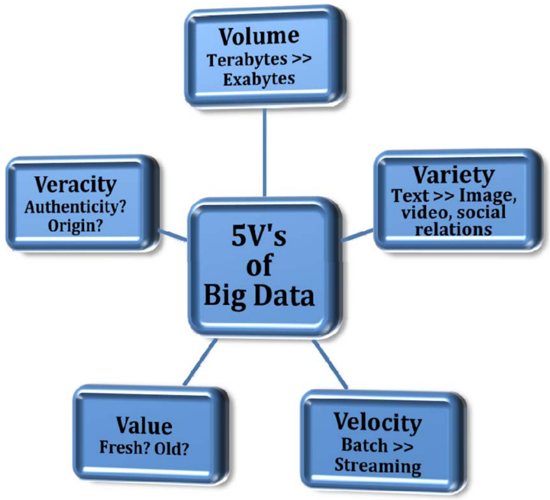


Fig. 1. Five V's of Big data.

Process engineers, instead of working with a physical model of the system (which may be very difficult, if not impossible to drive), prefer a model-free approach and use advanced techniques to monitor, control, and optimize the performance of the process, based only on the huge amount of measurements. Efficient capture and analysis of big data have the potential to enhance productivity with a resulting competitiveness in a wide range of industrial sectors. From the point of view of manufacturing engineers, supply chain management can be improved via big data solutions [6]. Additionally, by the proper interpretation of big data, more efficient risk management systems can be created to help company management to make better-informed decisions and improve corporate governance [7]. Based on these observations, it is apparent that fundamental research focusing on big data solution is imperative and critical for our life and especially for the future industrial applications.

II. CHARATERISTICS OF AND CHALLENGES POSED BY BIG DATA

IBM data scientists break big data into four dimensions: volume, variety, velocity, and veracity. This is inspired by the original discussion of E-commerce in three dimensions, namely, volume, velocity, and veracity by Doug Laney, a META Group (now Gartner) analyst [8]. It is now reflected in definition of big data by Gartner on their web page as: "Big data is high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, and decision making." A further "V": value (sometimes changed to validity or verification), can be added to the 4 V's of IBM, as shown in Fig. 1. The ever-increasing volume (amount of data), the velocity (speed of data in and out), and the variety (range of data types and sources) of big data constitute the basis of its challenges [8]. Inspired by the comprehensive discussion and the relevant comments available on IBM website of The Big Data & Analytics Hub, we discuss the challenges of big data in respect of 5 Vs, namely, volume, variety, veracity, velocity, and value and grouping them into three as below.

i) Volume and variety: These translate into requirements of hardware and software. Both the huge volume and the limitless variety of big data (from simple structured data samples to e-mail messages, tweets, facebook contents, etc.), bring challenges in the requirements of hardware and software to deal with data. Although the Cloud-related technologies (e.g., Cloud computing, virtualization, and storage) have proven to be successful in the Internet-related sectors, they still need improvements to cope with the requirements of real-time applications in the complex industries such as smart manufacturing and power systems. Predictably, programming models like MapReduce (that has proven to be successful in group-aggregation tasks) and large-scale data processing frameworks like Hadoop will become efficient tools in the coming years in many big-data-based projects.

ii) Veracity and velocity: These translate into urgency of online detection/processing ability. The data measured or collected from practical processes or systems should be detected timely (before any possible corruption or manipulations or becoming out of date), to ensure adequate trustworthiness. Although the raw data obtained from original sources are "washed," i.e., preprocessed or filtered before real application in most cases to avoid obvious irrelevant information, its large volume can intolerably influence the veracity of big data because the larger the volume is, the more difficult it becomes to "wash" it adequately. Given the validity of the washed data, the next key problem is not necessarily the large volume, but the online processing ability. It is the limiting factor in the velocity of industrial applications. The advantages brought by advanced network technologies of recent years have somewhat alleviated the velocity problem.

iii) Value: This translates into the necessity of interdisciplinary cooperation. It is proportional to veracity and they together raise the most difficult issues for industrial use of big data (it is to be noted that in a recent study by Deloitte [9], "viability" is named as the fifth V, and the sum of the 5V's is equated to the "value."). How do we validate the data that is useful, reliable, and accurate from a variety of huge data sets online? And once we have done that, how do we extract the value from it? Besides, big data is an interdisciplinary issue requiring the cooperation of academia, industry, and enterprise. Today, most approaches to big data inevitably contain a cavity between the ideal and the actual. Methods and algorithms focusing on data-based aspects like statistical analyses in business, management, and biomedicine [10–12], data-driven process monitoring/prognostics [13], [14], and system control and optimization [15], [16] have been widely investigated in recent years but still on the starting line. Many enterprises, however, regard their big data as confidential information, which hinders the development of novel approaches by academic research. Therefore, enterprises should consider a greater cooperation with researchers and engineers by making their specially processed data public for improvements of the existing techniques and for fostering new ideas. Presently, the challenges of industrial big data are still associated with their properties related to data measurement, detection, and processing. Notwithstanding this, acquiring the highest value from the available big data through appropriate analysis, utilization, and management should be the most important aspiration for modern industry.

III. EXPECTED BENEFITS OF BIG DATA FOR INDUSTRY

The potential benefits to be gained from the use of big data, as well as the challenges it will pose, will naturally differ from sector to sector. It is expected that computer and electronic products and information sectors, government sectors as well as finance and insurance, and government will gain substantially from the use of big data [18]. In general terms, the use of big data can unlock significant value in such areas as product and market development, operational efficiency, market demand predictions, decision making, and customer experience and loyalty [19]. In fact, in a recent study [20], the results of a survey carried out indicated that the functional objectives of the use of big data by the respondents were found to be as follows:

- Customer-centric outcomes 49%.
- Operational optimization 18%.
- Risk/financial management 15%.
- New business model 14%.
- Employee collaboration 4%.

It is seen that for almost half the respondents, the most important outcomes expected from the use of big data are the customer-centric ones. They would like to use information gathered in various ways and forms for customer analytics; to understand customer needs and anticipate future behaviors and thus provide better service to them. For example, sensors embedded in smart products are expected to, through the use of cyber-physical systems, send niche customer insights back, such as information on how they are being used by the customers, what types of functionality are being preferred and what newer functions would be welcome, etc. Furthermore, innovative after-sales services such as proactive maintenance can be offered by means of which preventive measures take place before a failure occurs. In this way, big data can be used to improve the development of the next generation of products and services.

IV. POSSIBLE FUTURE RESEARCH AND APPLICATION TRENDS OF INDUSTRIAL BIG DATA

Big data systems can be decomposed into four sequential modules; namely, data generation, data acquisition, data storage, and data analytics [1]. Each component of this value chain presents various challenges that require deep research into, mostly because of the heterogeneous and complex character of the data involved. For a deep analysis of the issues involved, the reader is referred to [1] and [21].

As an interdisciplinary or boundary-pushing topic, big data has the potential to attract an ever-growing attention of both the industrial communities and the industry related management and finance sectors. Some possible future trends of big data for modern industry include, but are not limited to:

- Novel techniques and improvements for big data analysis and mining.
- Cloud based solution related to big data storage and transmission.
- Big data solution focused on control and monitoring.
- Big-data based plant-wide optimization and prognosis.
- Big data solution for supply chain and risk management systems.
- Big data theory for modern industrial applications.
- Big data solution for smart grids and clean power systems.

Finally, it is reasonable to expect that the notion of big data will extend to other relevant fields (hand in hand with Cyber-Physical Systems and smart products) and create amazing additional possibilities and/or surprises. ■

REFERENCES

[1] J. Gantz and D. Remsel, "The digital universe in 2020: Big data, bigger digital shadows and biggest growth in the far east," in *Proc. IDC iView, IDC Anal. Future*, 2012.

[2] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Towards scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[3] M. Bailly and J. Manyika, "Is manufacturing 'cool' again?" McKinsey Global Inst., Jan. 21, 2013.

[4] General Electric Intelligent Platforms, "The rise of industrial big data," 2012, White Paper.

[5] H. Kagermann and W. Wahlster, "Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative INDUSTRIE 4.0," Working Group, Acatech—National Academy of Science and Engineering, Germany, 2013, Final Report of the Industrie 4.0.

[6] P. Reichert, "Comarch EDI platform case study: The advanced electronic data interchange hub as a supply-chain performance booster," in *Logistics Operations, Supply Chain Management and Sustainability*. Berlin, Germany: Springer International, pp. 143–155, 2014.

[7] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, Dec. 2012.

[8] B. Mark, "Gartner says solving 'big data' challenge involves more than just managing volumes of data," Gartner. Archived from the original on Jul. 10, 2011, Retrieved Jul. 13, 2011.

[9] Deloitte, "Big data: Challenges and success factors," 2013.

[10] A. McAfee *et al.*, "Big data, he management revolution," *Harvard Bus. Rev.*, vol. 90, no. 10, pp. 61–67, 2012.

[11] S. LaValle *et al.*, "Big data, analytics and the path from insights to value," *MIT Sloan Manage. Rev.*, vol. 21, 2013.

[12] E. E. Schadt *et al.*, "Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology," *Nature Rev. Genet.*, vol. 12, no. 3, pp. 224–224, Mar. 2011.

[13] S. Yin *et al.*, "Data-driven monitoring for stochastic systems and its application on batch process," *Int. J. Syst. Sci.*, vol. 44, no. 7, pp. 1366–1376, Jul. 2013.

[14] X. Si *et al.*, "Remaining useful life estimation-A review on the statistical data driven approaches," *Eur. J. Oper. Res.*, vol. 213, no. 1, pp. 1–14, Aug. 2011.

[15] S. Yin, H. Luo, and S. X. Ding, "Real-time implementation of fault-tolerant control systems with performance optimization," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 2402–2411, May 2014.

[16] Z. Hou and Z. Wang, "From model-based control to data-driven control: Survey, classification and perspective," *Inf. Sci.*, vol. 235, pp. 3–35, Jun. 2013.

[17] S. Yin, X. Li, H. Gao, and O. Kaynak, "Data-based techniques focused on modern industry: An overview," *IEEE Trans. Ind. Electron.*, 2014, DOI: 10.1109/TIE.2014.2308133.

[18] McKinsey, "Big data: The next frontier for innovation, competition, productivity," in McKinsey Global Institute Report, 2011.

[19] Information Systems Audit and Control Association (ISACA), "Big data: Impacts and benefits," Mar. 2013, White Paper.

[20] IBM and Said Business School, "Analytics: The real-world use of big data: How innovative enterprises extract value from uncertain data," IBM Institute for Business Value and Said Business School Executive Report, Oct. 2012.

[21] H. Moed *et al.*, "Special issue on big data," *Res. Trends*, Sep. 30, 2012.