ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 11, November 2019

Industrial Big Data Challenges and Analytical Methods for Data Acquisition- A Review

V.Sangeetha, Sr.J.Arockia Jaya M.E.,

M.E. Scholar, Idhaya Engineering College for Women, Chinnasalem, Tamil Nadu, India

Associate Professor, Idhaya Engineering College for Women, Chinnasalem, Tamil Nadu, India

ABSTRACT: While manufacturers have been generating highly distributed data from various systems, devices and applications, a number of challenges in both data management and data analysis require new approaches to support the big data era. These challenges for industrial big data analytics is real-time analysis and decision making from massive heterogeneous data sources in manufacturing space. This survey presents new concepts, methodologies, and applications scenarios of industrial big data analytics, which can provide dramatic improvements in velocity and veracity problem solving. we also provided the five typical applied case of industrial big data, which including smart factory visibility, machine fleet, energy management, proactive maintenance, and just in time supply chain. These newly methodologies and typical applied case will not only help manufacturers make more relevant and valuable decisions, but also provide capability of predictive manufacturing and service innovations. This survey also provides a comprehensive review of important technology in industrial big data analytics of related works to date, which hopefully will be a useful resource for further industrial big data analytics research.

I. INTRODUCTION

Big data analysis of industry is considered as a necessary aspect for further improvement in order to improve the profit margin of industrial production and operation, and represents the next frontier of innovation, competition and productivity [1]. Nowadays, industrial data platform is the core component of industrial data storage, computation and analysis for the management of intelligent plant. With the increasing number of intelligent equipments used in intelligent plant, however, intelligent plant can acquire a large quantity of data of Radio Frequency Identification (RFID) and intelligent equipments, thus providing rich data sets for manufacturing industry [2]–[7]. The current trend in industrial systems is to use different big data engines as a means of processing a large quantity of data that cannot be processed by ordinary infrastructures. Industrial infrastruct- ure faces a large number of problems, including challenges such as defining different efficient architecture settings for different applications and defining specific models for industrial analysis.

In recent years, Spark and Flink have been well studied and applied on the Internet. The MapReduce is a computing framework used by most industrial enterprises at present. Iterative computing is involved in many aspects of industrial data analysis, which is used to seek optimal control and management solutions. However, the MapReduce framework is ineffective in iterative computing and the performance of the Spark framework has some obvious advantages over the MapReduce [8], [9]. There have been many hot issues and difficult problems in the research of industrial big data platforms: reducing the data processing time and the size of space for big data acquisition and storage, optimizing the performance of computing framework, providing a wealth of function, improving system stability.

Our motivation for this study is to look for an efficient plan for acquiring and processing industrial data. Moreover, we need to find compression and serialization method which have good performance in the time of data processing time and the space of data storing. Furthermore, we aim at designing an industrial data platform with higher real-time performance and higher compression ratio. Based on the above considerations, we design and implement an industrial

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

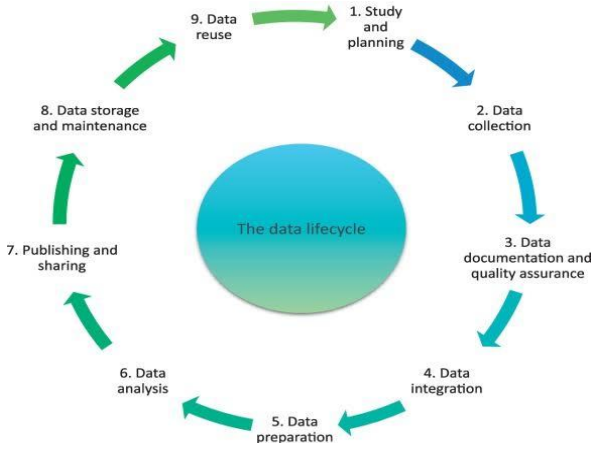
Visit: www.ijirset.com

Vol. 8, Issue 11, November 2019

data platform, which integrated both LZ4 and Protobuf method for data processing. The LZ4 is used for data compression, and the Protobuf is used for data serialization. Based Flume and Sqoop, we build up a data acquisition module in the industrial data platform. Meanwhile, the Hadoop is adopted to store data acquired by intelligent equipments in intelligent plant. In order to enhance the data analysis capabilities of the industrial data platform, both the Spark and the Flink are integrated into the computing system. In addition, the industrial data platform adopt Django as front-end framework for the sake of system stability and system maintenance management.

The structure of this paper is as follows. In Section I, we introduced the motivation of this research. In Section II, we introduce the relevant background. We build the data platform for industrial data analysis based on big data in Section III. Then, we show some results of the acquisition and processing system in Section IV. In Section V, we discuss the conclusions of this research.

There are many types of computing frameworks in the big data ecosystem, including MapReduce, Spark, Storm, Flink and so on [30]. The MapReduce is almost eliminated because of requiring constant serialization, which seriously wastes computing system resources. Moreover, many platforms for industrial big data are developed based on MapReduce computing framework. Obviously, the performance of these data platform is inefficient. The Spark, however, is the most active framework for the big data ecosystem. In addition, the Spark is based on memory, and it can automatically adjust the memory usage. When the memory is insufficient, it will automatically transfer the location of data computation from memory to disk. The speed at disk computing is also nearly 10 times faster than MapReduce [31]. Furthermore, Marhout's R&D team is no longer maintaining and perfecting for MapReduce, but the machine-learning module needs to be updated constantly in the field of industrial data analysis. Therefore, in this study, the Spark computing framework is used instead of the Hadoop computing one.



II. LITERATURE SURVEY

The magnitude of data generated and shared by businesses, public administrations numerous industrial and not-to-profit sectors, and scientific research, has increased immeasurably (Agarwal & Dhar, 2014). These data include textual content (i.e. structured, semi-structured as well as unstructured), to multimedia content (e.g. videos, images, audio) on a multiplicity of platforms (e.g. machine-to-machine communications, social media sites, sensors networks, cyber-physical systems, and Internet of Things [IoT]). Dobre and Xhafa (2014) report that every day the world produces around 2.5 quintillion bytes of data (i.e. *1 exabyte equals 1 quintillion bytes or 1 exabyte equals 1 billion gigabytes*), with 90% of these data generated in the world being unstructured. Gantz and Reinsel (2012) assert that by 2020, over 40 Zettabytes (*or 40 trillion gigabytes*) of data will have been generated, imitated, and consumed. With this

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 11, November 2019

overwhelming amount of complex and heterogeneous data pouring from any-where, any-time, and any-device, there is undeniably an era of *Big Data* – a phenomenon also referred to as the *Data Deluge*. The potential of BD is evident as it has been included in Gartner's *Top 10 Strategic Technology Trends for 2013* (Savitz, 2012a) and *Top 10 Critical Tech Trends for the Next Five Years* (Savitz, 2012b). It is as vital as nanotechnology and quantum computing in the present era. In essence, BD is the artefact of human individual as well as collective intelligence generated and shared mainly through the technological environment, where virtually anything and everything can be documented, measured, and captured digitally, and in so doing transformed into data – a process that Mayer-Schönberger and Cukier (2013) also referred to as *datafication*.

In line with the datafication concept and ever increasing technological advancements, advocates assert that in the future a majority of data will be generated and shared through machines, as machines communicate with each other over data networks (Van Dijck, 2014). Regardless of where BD is generated from and shared to, with the reality of BD comes the challenge of analysing it in a way that brings *Big Value*. With so much value residing inside, BD has been regarded as today's *Digital Oil* (Yi, Liu, Liu, & Jin, 2014) including the *New Raw Material* of the 21st century (Berners-Lee & Shadbolt, 2011). Appropriate data processing and management could expose new knowledge, and facilitate in responding to emerging opportunities and challenges in a timely manner (Chen et al., 2013). Nevertheless, the growth of data in volumes in the digital world seems to out-speed the advance of the many extant computing infrastructures. Established data processing technologies, for example database and data warehouse, are becoming inadequate given the amount of data the world is current generating. The massive amount of data needs to be analyzed in an iterative, as well as in a time sensitive manner (Jukić, Sharma, Nestorov, & Jukić, 2015). With the availability of advanced BD analysing technologies (e.g. NoSQL Databases, BigQuery, MapReduce, Hadoop, WibiData and Skytree), insights can be better attained to enable in improving business strategies and the decision-making process in critical sectors such as healthcare, economic productivity, energy futures, and predicting natural catastrophe, to name but a few (Yi et al., 2014).

As evident, much has been written on the BD phenomenon. The majority of academic research articles reviewed are analytical in nature (also evident from the findings – see Fig. 10, Fig. 11) that is either focusing on using experiments, simulations, algorithms and or mathematical modelling techniques in tackling BD. Regardless of their research approach, these articles present BD as a source that when appropriately managed, processed and analyzed, have the potential to generate new knowledge thus proposing innovative and actionable insights for businesses (Jukić et al., 2015). There is an ever-growing discourse about BD offering both *Big Opportunities* and *Big Challenges* through the plethora of sources from different domains; extending from enterprises to sciences. For instance, the opportunities include value creation (Brown, Chui, & Manyika, 2011), rich business intelligence for better-informed business decisions (Chen & Zhang, 2014), and support in enhancing the visibility and flexibility of supply chain and resource allocation (Kumar, Niu, & Ré, 2013). On the other hand, the challenges are significant such as data integration complexities (Gandomi & Haider, 2015), lack of skilled personal and sufficient resources (Kim, Trimi, & Chung, 2014), data security and privacy issues (Barnaghi, Sheth, & Henson, 2013), inadequate infrastructure and insignificant data warehouse architecture (Barbierato, Gribaudo, & Iacono, 2014), and synchronising large data (Jiang, Chen, Qiao, Weng, & Li, 2015). Advocates such as Sandhu and Sood (2014) perceive that the potential value of BD cannot be unearthed by simple statistical analysis. Zhang, Liu et al. (2015) support this perspective and state that to tackle the BD challenges, advanced BDA requires extremely efficient, scalable and flexible technologies to efficiently manage substantial amounts of data – regardless of the type of data format (e.g. textual and multimedia content).

BD and BDA as a research discipline are still evolving and not yet established, thus, a comprehensible understanding of the phenomenon, its definition and classification is yet to be fully established. The extant progress made in BD and BDA not only revealed a lack of management research in the field but a distinct lack of theoretical constructs and academic rigor – perhaps a function of an underlying methodological rather than academic challenge. At large, there has also been a lack of research studies that comprehensively addresses the key challenges of BD, or which investigates opportunities for new theories or emerging practices (e.g. George, Haas, & Pentland, 2014). Thus, there exists the need

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 11, November 2019

to culminate the BD challenges and associated BDA methods to allow signposting to take place. Following the earlier limited normative research studies conducted by Polato, Ré, Goldman, and Kon (2014) – mainly focusing on Apache Hadoop; Frehe, Kleinschmidt, and Teuteberg (2014) – BD logistics; Eembi, Ishak, Sidi, Affendey, and Mamat (2015) – on data veracity research for profiling digital news portal, and Abdellatif, Capretz, and Ho (2015) – on software analytics (a distinct branch of BDA), this paper attempts to *broaden the scope of their reviews by further investigating and assessing the different types of BD challenges and the analytical methods employed to overcome the challenges*. Although these research studies provide worthy understanding on some aspects of BD and BDA area, there seems to be a lack of comprehensive and methodical approaches to understand the phenomenon of BD – more precisely the types of BDA methods thus an aide memoir will act as a suitable frame of reference. Moreover, explicitly in respect of the conclusions offered by these existing review articles, this research specifically aims to: *analyze, synthesize and present a state-of-the-art structured analysis of the normative literature on big data and big data analytics to support the signposting of future research directions.*

Through using the abovementioned list of keywords and focusing on four subject areas that is *business and management*, *computer science*, *decision science*, and *social science*; initially 433 journal articles were identified from the Scopus database and relating to articles published during the period from 1996 to 2015. However, from period 1996 until 2002, there were no papers recorded on BD and BDA in these four subject areas. After assessing the 434 articles (from refereed journals), 206 papers were discarded, and finally 227 papers were selected and taken forward for further interrogation. As reflected in Fig. 9, contributors from across the world have made contributions to the BD and BDA area.

Nevertheless, given the limitations in the existing BD and BDA literature review studies (as reported earlier in Section 1.1), the rationale for undertaking this research is to provide a systematic state-of-the-art literature analysis of the BD and BDA area. In doing so to better understand the different types of BD challenges and associated BDA methods. Thus, the two underlying academic challenges orientate around identify the:
In this paper, the authors commenced this systematic search by using an established detailed review protocol based on the guiding principles and procedures of the SLR (Tranfield, D., et al., 2003, Kitchenham, B. and Charters, S., 2007). This protocol identifies the background review, search strategy, research questions as outlined in the abstract (i.e. Q1 and Q2), data extraction, criteria for study selection and data synthesis – based on the prescriptive three phased approach. The research questions and background of this review are described above, while the following sections provide details about other elements. As this literature review focuses on *analysing*, *synthesizing* and *presenting* a comprehensive structured analysis of the normative literature on BD and BDA, it was necessary to considered the domains for this research synthesis as both conceptual and empirical (including qualitative, quantitative and mixed method) papers. The research protocol for this literature review paper provides details on the following two points, as also followed by Delbufalo (2012) and Kamal and Irani (2014):

Types of Big Data Challenges
Among the many BD challenges large datasets (in terms of size and complexity) and the ability to process vast amount of data remains a critical challenge for outdated data processing applications and, relational database management systems (Jiang et al., 2015). According to TDWI Predictive Analytic Study (Russom, 2013), there are several BD challenges posing a peril to organizations – among these are, integrating complex and large datasets, getting started with the right BD project, developing and implementing infrastructure for managing and processing BD and a lack of skilled personnel or staff with analytics skills to make sense of BD.

III. CONCLUSION

BD and BDA as a research discipline are still evolving and not yet established, thus, a comprehensible understanding of the phenomenon, its definition and classification is yet to be fully established. The extant progress made in BD and BDA not only revealed a lack of management research in the field but a distinct lack of theoretical constructs and

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 11, November 2019

academic rigor – perhaps a function of an underlying methodological rather than academic challenge. At large, there has also been a lack of research studies that comprehensively addresses the key challenges of BD, or which investigates opportunities for new theories or emerging practices (e.g. George, Haas, & Pentland, 2014). Thus, there exists the need to culminate the BD challenges and associated BDA methods to allow signposting to take place.

REFERENCES

[1] GE. The rise of industrial big data. [online]. Available: http://www.geautomation.com/download/rise-industrial-big-data-2012.
[2] J. Lee. Industrial big data. China: Mechanical Industry Press, 2015.
[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers. Big data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, 2012.
[4] GE. The case for an industrial big data platform. [online]. Available: http://www.ge.com/digital/sites/default/files,2013.
[5] Brian Courtney. Industrial big data analytics: the present and future. [online]. Available: https://www.isa.org/intech/20140801/, 2014.
[6] ABB. Big data and decision-making in industrial plants.
[7] [online]. Available: http://new.abb.com/cpm/industry-software/ MOMmanufacturing-operations-management/big-dataanalyticsdecisionmaking, 2015.
[8] J. Lee, H. D. Ardakani, S. H Yang, B. Bagheri. Industrial big data analytics and cyber-physical systems for future maintenance and service innovation. In Procedia CIRP, vol.38, pp. 3-7, 2015.
[9] Ghemawat S, Gobioff H, Leung S-T. The google file system. In ACM SIGOPS Operating Systems Review, vol.37, pp. 29-43, 2003.
[10] Dean J, Ghemawat S. Mapreduce: simplified data processing on large clusters. In Commun ACM, vol.51,no.1, pp. 107-113, 2008.
[11] L. D. Xu, W. He, and S. Li. Internet of things in industries: A survey. In IEEE Trans. Ind. Informat., vol.11,no.3, pp. 841-852, Nov. 2014.
[12] J. Lee , E. Lapira, B. Bagheri, H. A. Kao. Recent advances and trends in predictive manufacturing systems in big data environment. In Manufacturing Letters, vol.1, pp. 38-41, 2013.
[13] Storm. [online]. Available: http://storm-project.net/, 2012.
[14] V. Gulisano, R. Jimenez-Peris, M. P. Martinez, C. Soriente, P. Valduriez. Streamcloud: an elastic and scalable data streaming system. In IEEE Trans. Parallel Distrib. Syst., vol.23,no.12, pp. 2351-2365, 2012.
[15] L. Neumeyer, B. Robbins, A. Nair, A. Kesari. Streamcloud: S4: distributed stream computing platform. In Proceedings of the IEEE Data Mining Workshops (ICDMW), Sydney, Australia, 2010, pp. 170-177.
[16] Sqlstream. [online]. Available: http://www.sqlstream.com/products/ server/, 2012.
[17] Splunk Storm Brings Log Management to the Cloud. [online]. Available: http://www.infoworld.com/t/managed-services/splunkstormbringslogmanagement- the-cloud-201098?source=footer, 2012.
[18] S. Kraft, G. Casale, A. Jula, P. Kilpatrick, D. Greer. Wiq: workintensive query scheduling for in-memory database systems. In IEEE 5th International Conference on Cloud Computing , 2012, pp. 33-40.
[19] The big data research and development plan. [online]. Available: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal/, 2012.
[20] Industry 4.0 in Big Data environment. In German Harting Magazine, vol.26, pp. 8-10, 2013. [21] D. Guinard, V. Trifa, S. Karnouskos, P. Spiess. Interacting with the SOAbased Internet of Things: Discovery, Query, Selection, and OnDemand Provisioning of Web Services. In IEEE TRANSACTIONS SERVICES COMPUTING, vol.3,No.3, pp. 223-235, July 2010.
[22] J. Cho, H. Garcia-Molina. Parallel crawlers. In Proceedings of the 11th international conference on World Wide Web, 2002, pp. 124-135.
[22] M. J. Cafarella, A. Halevy, N. Khoussainova. Data integration for the relational web. In Proceedings of the VLDB Endowment, 2009, pp. 10901101.
[23] A. Schultz, A. Matteini, R. Isele, P. Mendes, C. Bizer, C. Becker. LDIFA framework for large-scale linked data integration. In Proceedings of 21st International World Wide Web Conference (WWW2012), Developers Track. Lyon, France, April 2012, pp. 1097-1110.
[24] G. Demartini, D. E. Difallah, C. M. Philippe. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. In The VLDB Journal, 2013, vol.22, pp. 665-687.
[25] K. B. Liu, and S. Huang. Integration of Data Fusion Methodology and Degradation Modeling Process to Improve Prognostics. In IEEE Transactions on Automation Science and Engineering, vol.13, no.1, pp. 344-354, Jan 2016.