Grenze International Journal of Engineering and Technology, June Issue

Privacy and Security of Big Data Architectures for Information Management System

Honey Jain[1], Rohan Benhal[2] and Tanmayee Parbat[3]
[1, 3]B.E IT, Maeer's MIT, Pune
Email: tanmayee.parbat01@gmail.com
[2]BBA IT, Maeer's MIT, Pune

***Abstract*—The importance of data-driven decisions and support is increasing day by day in every management area. The constant access to volume, variety, and veracity of data has made big data an integral part of management studies. New sub-management areas are emerging day by day with the support of big data to drive businesses. This study takes a systematic literature review approach to uncover the emerging management areas supported by big data in contemporary times. For this, we have analyzed the research papers published in the reputed management journals in the last ten years, fir using network analysis followed by natural language processing summarization techniques to find the emerging new management areas which are yet to get much attention. Furthermore, we ran the same exercise in each of these management areas to uncover these areas better. This research will act as a reference for future information systems (IS) scholars who want to perform analysis that is deep-dive in nature on each of these management areas, which in the coming times will get all the due attention to become dedicated research domains in the management area.**

***Index Terms*—** Big Data (BD), Big Data Analytics (BDA), Information Systems (IS)**.**

I. INTRODUCTION

The role of Big Data (BD) in driving business revenue, operation, and customer support has attracted the attention of management scholars in the information systems (IS) domain in the last decade. Several existing organizations in the business and new organizations starting fresh are investing in building Big Data Analytics (BDA) capabilities. The motive is to derive actionable insights with the volume, variety, and veracity of data and enable the individuals or organizations to take and communicate informed decisions (Constantiou & Kallinikos, 2015). With the rapid increase of the processing power available at the hand of an analyst, the motivation to use the data across the smaller day-to-day management areas has become a need for organizations. These everyday management activities supported by BDA have now significantly evolved as full-blown management domains that still tie to conventional management theories. However, there is not much research that uncovers these contemporary management emerging management areas leveraging the full potential of effective BDA techniques.

Text-based information retrieval or multimedia (image and video) based information retrieval has become an

Grenze ID: 01.GIJET.8.2.319

integral part of most research areas in the analytics domain over the last decade. The rapid and steady growth of text and multimedia data with the development and increase in internet penetration has created the need to develop the BDA frame- works to be analyzed across various management domains. Several pro- grams and tools have been designed to formulate, query, and process multiple BD vis-à-vis text and multimedia facets. This further enables three different kinds of analytics using BD. These are namely: descriptive analytics such as reporting, dash boarding, and visualizations, fol- lowed by discovery analytics which captures early signals through text summarizations, feature extraction from image and video, and third being predictive analytics primarily driven from a range of econometrics models to complex machine learning models.

To this far, BDA applications have been explored only in private sectors' management practices with the massive growth of data get- ting recorded daily (Yaqoob et al., 2016). This leaves a considerable research gap, even in the public sector management domains that could use BDA's capabilities to solve and cater to issues. This, in turn, could evolve some of the contemporary management domains even in the pub- lic sector. To attract future IS researchers' attention to this field, it is vital to know some of the problems catered to with BDA. Expanding knowledge of the data among the public sector practitioners plays a role in how BDA can reduce inefficiency and eventually increase rev- enue. Although the computational researchers have given increased attention to BDA using the public sector data ((Guenduez et al., 2020 ; Pencheva et al., 2020 ; Galetsi et al., 2020), developing the insights gen- erated from these researches to management practices still needs to be thoroughly researched upon.

Big data processing can be performed through two manners: batch processing and stream processing; see [4]. The first method is based on analyzing data over a specified period of time; it is adopted when there are no constraints regarding the response time. On the other hand, stream processing is suitable for applications requiring real-time feedback. Batch processing aims to process a high volume of data by collecting and storing batches to be analyzed in order to generate results.

Batch mode requires ingesting all data before processing it in a specified time. Mapreduce represents a widely adopted solution in the field of batch computing; see [5]; it operates by splitting data into small pieces that are distributed to multiple nodes in order to obtain intermediate results. Once data processing by nodes is terminated, outcomes will be aggregated in order to generate the final results. Seeking to optimize computational resources use, mapreduce allocates processing tasks to nodes close to data location. This model has encountered a lot of success in many applications, especially in the field of bioinformatics and healthcare. Batch processing framework has many characteristics such as the ability to access all data and to perform many complex computation operations, and its latency is measured by minutes or more.

II. Big Data Characteristic

The term big data is described by the following characteristic value, volume, velocity, variety, veracity and variability, denoted as 6 "Vs" [13], [14], shown in Figure Figure1.1. Besides these 6 "Vs", some authors have defined more than these 6 properties to describe big data characteristics [15]. The volume of health and medical data is expected to rise intensely in the years ahead, usually measured in terabytes, petabytes even yottabytes [14], [16]. Volume refers to the amount of data, while velocity refers to data in motion as well as and to the speed and frequency of data creation, processing and analysis. Complexity and heterogeneity of multiple datasets, which can be structured, semi-structured and unstructured, refer to the variety. Veracity referrers to the data quality, relevance, uncertainty, reliability and predictive value [14], while variability regards about consistency of the data over time. The value of the big data refers to their coherent analysis, which should be valuable to the patients and clinicians.

Considering the big data characteristics, data searching, storage and analysis, a very appropriate and promising software platform for development of applications that can handle big data in medicine and healthcare is the open-source distributed data processing platform Apache Hadoop MapReduce [1], [17] that is based on data-intensive computing and NoSQL data modeling techniques [18].

A. Big Data Analytics

Applications of big data analytics can improve the patient-based service, to detect spreading diseases earlier, generate new insights into disease mechanisms, monitor the quality of the medical and healthcare institutions as well as provide better treatment methods [7].

Data mining techniques employed on EHRs, web and social media data enable identifying the optimal practical guidelines in the hospitals, identifying the association rules in the EHRs [8] and revealing the disease monitoring and health-based trends. Moreover, integration and analysis of the data with different nature, such as social and

scientific, can lead to new knowledge and intelligence, exploring new hypothesis, identifying hidden patterns [14].

Figure 1: The 6 V's of big data

Nowadays, smart phones are excellent platforms to deliver personal messages to patients to involve them in behavioral changes to improve their wellbeing and health conditions. The mobile phone messages can substitute delivering of medical and motivational advices to the patients [14].

B. Challenges in Big Data Analytics

Regarding collection of large amount data, some challenging issues should be considered. Obtaining high-throughput – omics data is tied to the cost of experimental measurements. Concerning heterogeneity of the data sources, the noise of the experimental – omics data and the variety of the experimental techniques, environmental conditions, biological nature should be considered, before integration of these heterogeneous data and before employing of the data mining methods. Different data mining techniques can be applied on these heterogeneous biomedical data sets, such as: anomaly detection, clustering, classification, association rules as well as summarization and visualization of those big data sets.

These shortcomings might lead to the unreliability of some of the data points, such as missing values or outliers. Despite of these drawbacks of the – omics data, EHRs data are very influenced by the staff who entered the patient's data, which can lead to entering missing values, incorrect data as a result of mistakes, misunderstanding or wrong interpretation of the original data [5]. Integration of data from various databases and standardization for laboratory protocols and values still remain challenging issues [10].

High dimensionality of the – omics data means, that there have many more dimensions or features than the number of samples, and on the other side the EHRs data which regard to the individuals/patients, makes data mining techniques to be more challenging task.

The subsequent stage is the pre-processing of the data, which usually envelop handling noisy data, outliers, missing values, data transformation and normalization. This data pre-processing enables to be applied statistical techniques and data mining methods and thus the big data analytics quality and outcomes can improve and can result with discovering of novel knowledge. This novel knowledge obtained by integration of the – omics and EHRs data should results with improving of the implemented healthcare to the patients as well to advanced decision making by the healthcare decision policy makers.

C. Big data Privacy and Security

Two important issues towards big data in healthcare and medicine are security and privacy of the individuals/patients [4], [3]. All medical data are very sensitive and different countries consider these data as legally possessed by the patients [2]. To address these security and privacy challenges, the big data analytics

software solutions should use advanced encryption algorithms and pseudo-anonymization of the personal data. These software solutions should provide security on the network level and authentication for all involved users, guarantee privacy and security, as well as set up good governance standards and practices.

III. Methodology

Solution for Integrating Medical Data Streams and Data quality: A unified web based system can be developed having the features of not just in taking Unicode but also image, audio and voice format. Level 1 Solution: Inclusion of "Integration Bucket" functionality in landing zone. This concentrates on Data filtering which maintains high quality by removing all redundant data. Data collected from various sources are directed to landing zone causing repetition of data in cloud.

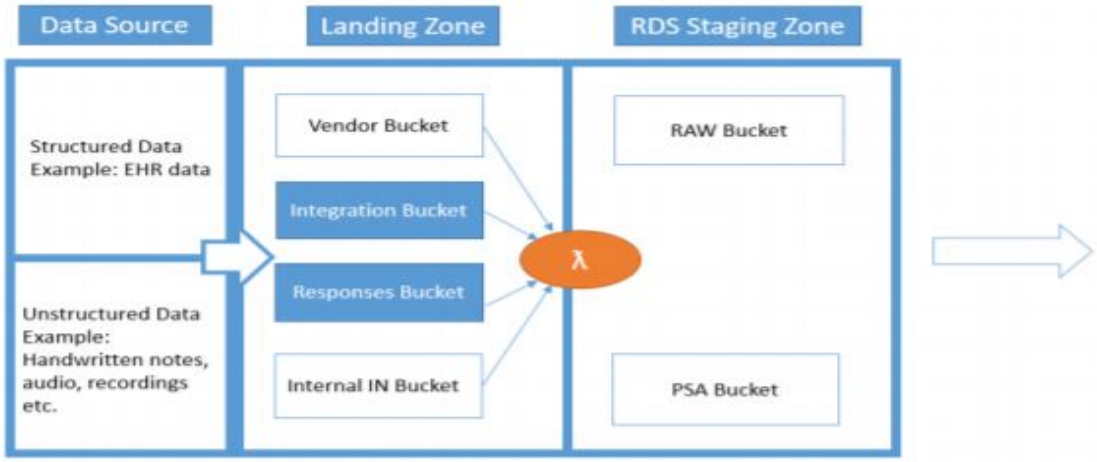

Figure 2: Components in Cloud Computing

To avoid this, the adding a functionality of "Integration" to remove repeated and redundant data can save time and complexity. Such a solution is tremendously useful for performing a critical analysis on the fly. (Figure 1) Solution for integration with real-time responses: Most of the existing solutions are built on the Hadoop-MapReduce framework, which mostly solved the data volume challenge. However, due to the extravagant sorting algorithm that Hadoop relies heavily on for performing causes the performance to bottleneck. Based on our study we concluded that this problem can be resolved by incorporating it with cloud computing.
Level 2 Solution: Inclusion of "Responses Bucket" function in landing zone. Real time responses occur only when there are real time interactions. Interactions are not unique; repeated interactions can be captured in landing stage with the use of a pointer to call particular function of real time response (Figure 2). This will help save tremendous amount of processing time. Responses are processed and to give out a relevant solution, relevant interactions can be captured by this function to avoid duplication. All the filtered solutions are then provided to LAMBDA bucket. Since LAMBDA is python based programming, it reduces the number of operations. The data is retrieved using SQL query.

IV. Conclusion

An ever-increasing number of companies attempt to adopt BDA to use the data's hidden trends and signals and help decision manager's aid data-driven decisions. While these companies must utilize the real value of BDA, it is also equally crucial for them to understand the learnings from the successful use-cases, success criteria, and the right metrics to measure real incremental business value. Hence, the current manuscript and the findings reported as EMDs as an outcome of the study under- taken will act as a reference for future IS scholars and become an essen- tial note for management practitioners to learn from the issues catered by BDA in each of these domains. With BDA yet to be commoditized like IT products, some research needs to be done on managing resources and implementation phases of BDA. While the earlier research articles an- alyzed in the current study have well reported how BDA have helped various real-time and looking at the issue retrospectively, there is a lack of theoretically driven research identifying how these BDA solutions can be well implemented to gain business and competitive advantage. To this end, the current study has analyzed various business value measuring frameworks that can measure a real unbiased value of BDA and concepts from

operations and crisis management perspective. The synthesis of the prior research work in the current study acts as a ref- erence for broader implementation of BDA to become a commodity for business houses, which will allow business organizations to get higher returns on investments.

REFERENCES

[1] Yang C, Li C, Wang Q, Chung D, Zhao H. Implications of pleiotropy: challenges and opportunities for mining big data in biomedicine. *Front Genet.* 2015;6:229.

[2] Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Health Inform.* 2015;19:1209–15.

[3] Kankanhalli A, Hahn J, Tan S, Gao G. Big data and analytics in healthcare: introduction to the special section. *Inform Syst Front.* 2016;18:233–5.

[4] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inform Sci Syst.* 2014;2:3.

[5] Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. –Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Trans Biomed Eng.* 2017;64:263–73.

[6] Wang Y, Kung LA, Wang WY, Cegielski CG. An integrated big data analytics-enabled transformation model: application to health care. *Inf Manag.* 2017;55:64–79.

[7] El-Gayar O, Timsina P. Opportunities for business intelligence and big data analytics in evidence based medicine. System Sciences (HICSS); 47th Hawaii international conference on 2014.2014. pp. 749–57.

[8] Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. *Int J Med Inform.* 2017;98:22–32.

[9] Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. *Proteomics.* 2016;16:741–58.

[10] Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. *Biomed Inform Insights.* 2016;8:1.

[11] Gaitanou P, Garoufallou E, Balatsoukas P. The effectiveness of big data in health care: a systematic review. *In: Metadata and semantics research.* 2014:141–53.

[12] Lillo-Castellano JM, Mora-Jimenez I, Santiago-Mozos R, Chavarria-Asso F, Cano-González A, García-Alberola A. et al. Symmetrical compression distance for arrhythmia discrimination in cloud-based big-data services. *IEEE J Biomed Health Inform.* 2015;19:1253–63.

[13] Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. *IEEE J Biomed Health Inform.* 2015;19:1193–1208.

[14] Archenaa J, Anita EM. A survey of big data analytics in healthcare and government. *Procedia Comput Sci.* 2015;50:408–13.

[15] Borne K. *Top 10 big data challenges – a serious look at 10 big data V's*. MAPR; 2014. NO4, 80.

[16] Hermon R, Williams PA. Big data in healthcare: what is it used for?; Australian Ehealth Informatics and Security Conference; 2014. pp. 40–9.

[17] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM*. 2008;51:107–13.

[18] Trifonova OP, Il'in VA, Kolker EV, Lisitsa AV. Big data in biology and medicine. *Acta Naturae.* 2013;5:13–6.

[19] Agarwal M, Adhil M, Talukder AK. *International Conference on Big Data Analytics*. Cham, Switzerland: Springer International Publishing; 2015. Multi-omics multi-scale big data analytics for cancer genomics; pp. 228–43.