Big Data Architectures for Information Management System

Pankhuri Bhatnagar[1], Rimsha Virmani[2], Pratyush Bhatnagar[3]

[1,2,3]*Dr. Akhilesh Das Gupta Institute of Technology and Management*

Abstract: - Making decisions and providing support based on data is becoming increasingly important in every aspect of management. Because of the ease with which large volumes, a wide variety, and high accuracy data can be accessed, big data has become an essential part of management studies. Big data may be able to assist businesses in forming new management divisions. A systematic review of the literature identified the new management areas supported by big data today. Because we used network analysis and natural language processing summarization techniques to examine research papers published in reputable management journals over the last decade, these emerging new management fields have received little attention until now. To better identify problem areas, the same exercise was carried out in each of these management areas. If this research holds up, it will serve as a model for future IS scholars who want to conduct detailed analyses of each management area and turn them into research domains.

Keywords: -Big Data (BD), Big Data Analytics (BDA), Information Systems (IS)

I. INTRODUCTION

Many information systems experts have spent the last ten years researching the impact of Big Data (BD) on revenue, operations, and customer service in the business world. Big Data Analytics (BDA) projects are being funded by businesses of all sizes, both established players and newcomers. The ultimate goal is to turn all data into actionable insights, allowing individuals or organizations to make well-informed decisions and communicate those decisions to others (Constantiou & Kallinikos, 2015). Because of the rapid increase in processing power available to analysts, organizations now have no choice but to use data across smaller day-to-day management areas. BDA has enabled day-to-day management activities to evolve into full-fledged management domains that are still bound by traditional management theories. However, in the current state of the art, research into these contemporary management emerging management areas that fully utilize effective BDA techniques is severely lacking.

Using text-based information retrieval as well as multimedia has become increasingly common in analytics research over the last decade (such as images and videos). As the internet's penetration increases, text and multimedia data are rapidly growing, necessitating the development of BDA frameworks for analysis in a variety of management fields. Multiple BD in terms of text and multimedia can be created, queried, and processed using a variety of programs and tools. As a result, three different types of analytics are now supported by BD. That's exactly what I'm referring to: The three types of analytics available are descriptive analytics (such as reporting), discovery analytics (such as extracting features from images and videos), and predictive analytics (such as forecasting) (which is primarily driven by a range of econometrics models to complex machine learning models). Reporting, dashboards, and visualizations are examples of descriptive analytics tools.

BDA applications have only been examined in private sector management practices until now, due to the massive increase in data being recorded every day (Yaqoob et al., 2016). There is a significant research gap even in public sector management domains, where BDA's capabilities could be used to solve and cater to issues. As a result, some current management domains in the public sector may evolve. If we can grasp some of the issues raised by BDA, future IS researchers will be drawn to this field. BDA's ability to reduce inefficiency and, ultimately, increase revenue depends on increasing public sector practitioners' data knowledge. Despite the growing interest in BDA among computational researchers working with public sector data (Guenduez et al., 2020; Pencheva et al., 2020; Galetsi et al., 2020), more research is needed to apply the findings to management practices.

[4] shows two approaches to dealing with large amounts of big data: batch processing and stream processing. The first method relies on data analysis over a predetermined period of time in the absence of response time constraints. Stream processing outperforms batch processing in real-time applications. To generate results, this technique collects and analyzes data in batches.

In order to use batch mode, all ingested and processed data must be saved for a set amount of time. To get intermediate results when batch computing with MapReduce, it's common to split data into small chunks and distribute them across multiple nodes. Take a look at [5] for more information on MapReduce. The central processing unit will collect and analyze the results of the nodes' data processing (CPU). To make the most of computational resources, MapReduce distributes processing tasks to nodes near the data source. This model has proven to be very successful, particularly in bioinformatics

and healthcare. With a latency of minutes or longer, batch processing frameworks can access all data and perform numerous complex computations.

Table 1: Difference between Traditional data and Big data

S.No.	TRADITIONAL DATA	BIG DATA
01.	Traditional data is generated in enterprise level.	Big data is generated in outside and enterprise level.
02.	Its volume ranges from Gigabytes to Terabytes.	Its volume ranges from Petabytes to Zettabytes or Exabytes.
03.	Traditional database system deals with structured data.	Big data system deals with structured, semi structured and unstructured data.
04.	Traditional data is generated per hour or per day or more.	But big data is generated more frequently mainly per seconds.
05.	Traditional data source is centralized and it is managed in centralized form.	Big data source is distributed and it is managed in distributed form.
06.	Data integration is very easy.	Data integration is very difficult.

II. BIG DATA CHARACTERISTIC

The six "Vs" of big data (volume, velocity, variety, truthfulness, and variability) are depicted in Figure 1.1 [13, 14]. Other authors have defined additional "Vs" to describe big data characteristics in addition to these six [14], [16]. Terabytes, petabytes, and even yottabytes of health and medical data will be generated in the future. When it comes to the amount of data created, processed, and analyzed while in motion, there is a distinction between volume and velocity. Based on their level of complexity and heterogeneity, multiple datasets can be classified as structured, semi-structured, or unstructured. Reliability refers to the data's consistency over time, whereas veracity refers to its relevance. The value of big data stems from its ability to be analyzed in a way that benefits both patients and clinicians.

The Apache Hadoop MapReduce [1] distributed data processing platform, which is based on data-intensive computing techniques and NoSQL data modeling [18], is a promising and appropriate software platform for developing applications that can handle large amounts of big data in medicine and healthcare.

2.1 **Big Data Analytics**

Big data applications can help improve patient-centered services by detecting disease outbreaks earlier, generating new insights into disease mechanisms, and monitoring the quality of medical and healthcare institutions. EHRs, the web, and social media data can all be mined for useful information, such as hospital best practices and association rules. Combining and analyzing various types of information, such as social and scientific data, is another method of learning and intelligence discovery [14].

Table II: Traditional Data Analysis Vs Big Data Analysis characteristics

Determining Factor	Traditional BI	Big Data
Data volume	Typically Terabytes	Tens to Hundreds of Terabytes, to Petabytes
Velocity of change in scope	Slower	Faster. Can adapt to frequent change of analytics needs
Total Cost of Ownership	TOC tends to be expensive	TOC tends to be lower due to lower cost storage and Open source tools
Source data diversity, variety	Lower	Higher
Analysis driven	Typically supports known analytics and required reporting	Inherently supports the data analysis and data discovery process by certain users
Requirements driven	Most of the time	Rarely
Exploration & discovery	Some of the time	Most of the time
Structure of queries	Robust	Un-structured
Accuracy of Results	Deterministic	Approximated
Availability of results (SLA)	Slower (longer batch cycles)	Faster
Stored data	Schema is required to write data	No pre-defined schema is required

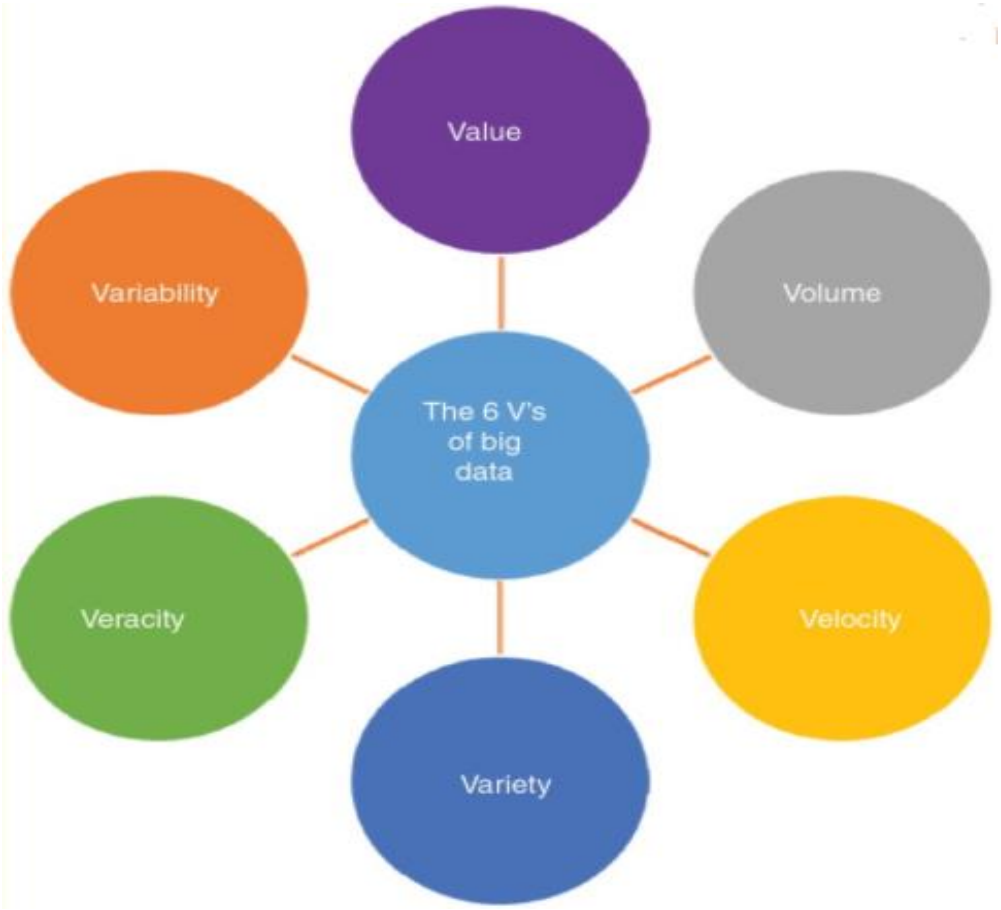

Figure 1: The 6 V's of big data

Patients can now receive messages tailored to their specific needs, such as suggestions for healthy lifestyle changes, thanks to this new technology. Smartphones are a fantastic medium for this type of communication. Patients can receive medical and motivational advice via text messages instead of traditional methods like letters and visits. [14].

2.2 Challenges in Big Data Analytics

When collecting a large amount of data, some complex issues must be considered. Only after the researchers have paid for the experiments can they obtain high-throughput omics data. It's critical to consider noise, as well as the various experimental techniques used, environmental conditions, and biological nature when working with experimental omics data. On these heterogeneous biomedical data sets, data mining techniques such as anomaly detection and visualization can be applied.

They can result in untrustworthy data points such as missing values or outliers. Human error, misinterpretation, or misunderstanding of the patient's records can result in incorrect data being entered into electronic health records (EHRs) [5, 6]. Standardizing laboratory protocols and integrating data from multiple databases is still a difficult problem to solve [10].

To put it another way, omics data has many more dimensions or features than samples, making data mining individual/patient-specific EHR data more difficult.

Data pre-processing is the next step, which involves dealing with noisy data, outliers, and missing values, as well as data transformation and normalization. This data pre-processing enables the application of statistical and data mining techniques, resulting in improved big data analytics quality and results, as well as the discovery of new knowledge. With the new information derived from the integration of omics data with EHR records, health care providers should be better equipped to make informed policy decisions.

2.3 Big data Privacy and Security

When using big data in healthcare and medicine, medical professionals are concerned about the safety and privacy of patients and employees. Every country considers patient medical data to be their legal property [2]. To address these security and privacy concerns, big data analytics software solutions should use advanced encryption algorithms and pseudo-anonymization of personal data. These software solutions should ensure privacy and security, as well as establish good governance standards and practices, in addition to providing network security and authentication for all users.

III. METHODOLOGY

A Workable Solution for Data Quality and Integrity in Medical Data Streams A web-based unified system can accept Unicode as well as image, audio, and voice formats. The integration bucket for the landing zone is a Level 1 solution. This focuses on Data Filtering, which removes all redundant data from the system while maintaining high-quality data. The landing zone, where data gathered from various sources is directed, causes data repetition in the cloud.

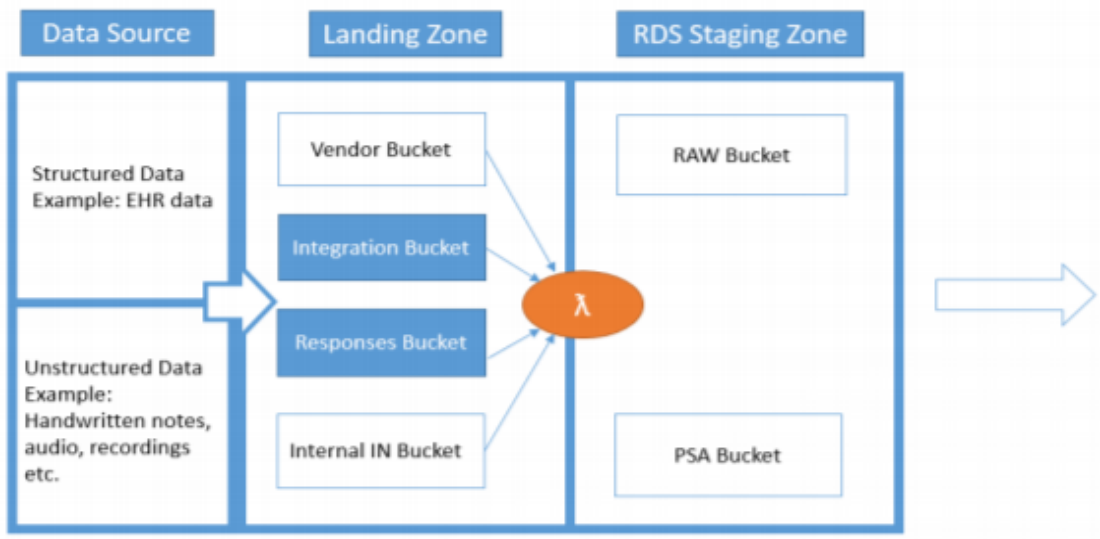

Figure 2: Components in Cloud Computing

This problem can be avoided by adding an "Integration" function that removes redundant and repeated data. When you need to do on-the-spot analysis, it's a lifesaver. (See Illustration No. 1 for more information.) Here's a real-world example of a real-time integration system in action: Most existing solutions use Hadoop-MapReduce technologies to efficiently deal with large amounts of data. Hadoop is slowed by using an ostentatious sorting algorithm for basic tasks. Cloud computing, according to our findings, may be the solution to this problem.

As a Level 2 fix, add a "Responses Bucket" function to the landing zone. Real-time interactions only produce real-time responses. You can record repeated interactions by using a pointer to call a specific real-time response function in the landing stage. There is no such thing as a one-of-a-kind encounter (Figure 2). Finally, this implementation will result in a significant reduction in processing time. We sort through the responses to avoid duplication and provide the best solution. This function can be used to record important interactions. Everything that has survived the filtering process is then deposited in the LAMBDA container. LAMBDA is more efficient because it is written in Python. To obtain the data, a SQL query is used.

IV. CONCLUSION

More companies are attempting to implement BDA to benefit from the data's hidden trends and signals and assist decision-makers in making data-driven decisions. However, in order to appreciate the true value of BDA, these companies must understand the lessons learned from successful use-cases, success criteria, and appropriate metrics to measure real incremental business value. Future IS scholars and management practitioners can benefit from BDA's experiences in each of these areas by using this manuscript and the research findings reported as EMDs. Because BDA hasn't yet become commoditized like IT products, more research into resource management and implementation is needed. According to the findings, previous research articles examined in this study reported a lack of theoretically-driven studies identifying how BDA solutions can be implemented to gain a business and competitive edge. A recent study looked at several business value-calculation frameworks to see if they can be used to determine the true, unbiased value of BDA and concepts from

the operations and crisis management perspectives. Based on the prior research used in this study, BDA implementation could become a commodity for businesses, allowing them to earn higher returns on their 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. InfManag. 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. ActaNaturae. 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.