Big Data Environment and Its Challenges: A Case Study

Dr. Ashif Ali[1], Ms. Rashika Singh[2], Ms. Antim Panghal[3]

Associate Professor[1], Assistant Professor[2,3]

hod.cse@acem.edu.in, rashika.singh@acem.edu.in, antim@acem.edu.in

Department of CSE, Aravali College of Engineering & Technology[1,2,3]

Abstract

Big Data continues to be a field that is expanding and evolving. The magnitude of big data databases and files has beyond the capacity and functionality of traditional commercial database management solutions [1]. In this case, structured data representations may impede efficient data storage and retrieval. Gartner has identified four principal issues, referred to as the four Vs: the escalating volume of data, the rising velocity of data inflow, outflow, and transformation, the broadening variety of data types and structures, and the augmenting variability of data. A fifth V, value, is often proposed, signifying the role of big data in strategic decision-making. As the impact of big data broadens across several disciplines and issue domains, the problems and opportunities for it to substantially affect business, research, and government also increase.

Keywords: Big Data, Data Integrity, Security

I. INTRODUCTION

A globally recognized definition of big data has yet to be formulated. Wikipedia defines it as "a comprehensive term for any aggregation of data sets that is so extensive and intricate that it becomes challenging to process with conventional data processing applications." [1] In this framework, the physical realm is mirrored in cyberspace via the Internet, the Internet of Things (IoT), and various information technologies, resulting in big data. Simultaneously, human civilization creates a comprehensive data-driven representation of itself in cyberspace through technologies such as human-computer interfaces and mobile Internet.

A. Definition of Big Data

Big data denotes data sets of such vast magnitude and intricacy that conventional data processing software is insufficient for their management. The related difficulties encompass the complete data lifecycle, comprising capture, storage, analysis, search, sharing, transfer, visualization, querying, updating, and safeguarding information privacy. Big data is typically characterized by five dimensions: Volume, Variety, Velocity, and the recently incorporated Veracity and Value. The phrase typically denotes the application of predictive analytics or other sophisticated techniques to get value from data, rather than indicating a particular data set size.

B. What are the 5V's of Big Data?

The 5V's provide a robust framework for understanding the multifaceted nature of big data.

These characteristics are interconnected, as visualized in Figure 1, where Volume, Velocity, Variety, and Veracity collectively converge to generate the ultimate goal .

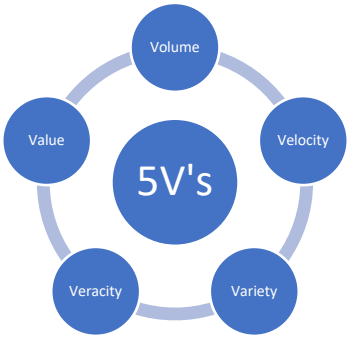

Figure 1: The 5V's of Big Data

1. Volume: This dimension is mostly linked to big data, encompassing data amounts that have attained practically unfathomable scales [3]. The everyday production of emails, social media communications, images, and sensor data is astonishing. Facebook experiences 10 billion messages and 350 million photo submissions daily. The cumulative data produced from the beginning of time until the year 2000 is currently generated every minute. This scale renders conventional database technology inadequate, necessitating big data technologies that utilize distributed systems to store and process data across various locations.

2. Velocity: The proliferation of data and social media has altered expectations around the promptness of data. Although newspapers disseminate information from the previous day, social media offers real-time updates, rendering statements mere seconds old obsolete. The near real-time transmission of data is a defining attribute of big data.

Instances include social media articles achieving virality within minutes, instantaneous fraud detection for credit card transactions, and stock trading systems that evaluate social media networks in milliseconds to guide buy/sell decisions [4]. Big data technology facilitates the analysis of data in real-time, eliminating the necessity for prior storage in databases.

3. Variety: Data is currently generated by numerous sources such as sensors and smart devices, leading to raw, semi-structured, unstructured, and rich media formats that complicate processing and storage.

Historically, the emphasis was on structured data that conformed to tabular formats. Presently, almost 80% of global data is unstructured—comprising images, video footage, and social media posts—and is not well accommodated in relational databases [5]. Big data technology facilitates the utilization of many data kinds and their amalgamation with conventional, organized data.

4. Veracity: This dimension pertains to the dependability and precision of facts. It concerns not just data quality but also its comprehensibility. Given that data governance frequently assumes a subordinate role, it is imperative to establish methods that avert the accumulation of faulty data within systems [6].

Illustration: Twitter tweets with hashtags, abbreviations, and typographical errors. Big data analytics enables the manipulation of vast datasets, where the substantial volume can frequently offset individual inaccuracies [6].

5. Value: This is the paramount V, signifying the ultimate business advantage obtained from data. The enterprise must delineate the analytical use of the data and its prospective value, which is crucial for substantiating projects and formulating a strategic roadmap [7].

Big data provides benefit across multiple sectors:

Comprehending Consumers: Tailored suggestions from Amazon or Netflix.

Enhancing Operations: Uber's demand forecasting and variable pricing strategy.

Enhancing Healthcare: Government agencies forecasting influenza epidemics.

Augmenting Security: Preventing terrorist strikes and identifying cybercrime.

Enhancing Athletic Performance: Evaluating data from sensors and GPS trackers to optimize athlete efficacy [7].

II. BIG DATA ANALYTICS

Big data analytics entails the collection, organization, and analysis of extensive data sets to identify trends and extract valuable insights. It allows organizations to utilize extensive data to discern trends and derive essential insights for future decisions, mostly aimed at enhancing decision-making. Data scientists employ these insights on transactional data that may be overlooked by traditional Business Intelligence applications.

The software technologies employed belong to advanced analytics fields such as predictive analytics and data mining. Nevertheless, the Volume and Velocity of big data render conventional data warehouses incapable of managing real-time, continually updated data streams from sources such as social media. Recent technologies for this purpose encompass Hadoop and its associated tools (YARN, MapReduce, Spark, Hive, Pig), in addition to NoSQL databases [8].

A typical architecture for handling these tasks is depicted in Figure 2. It illustrates the logical flow from diverse data sources through various stages of ingestion, storage, processing, and finally to analysis and visualization, forming a complete analytics pipeline.

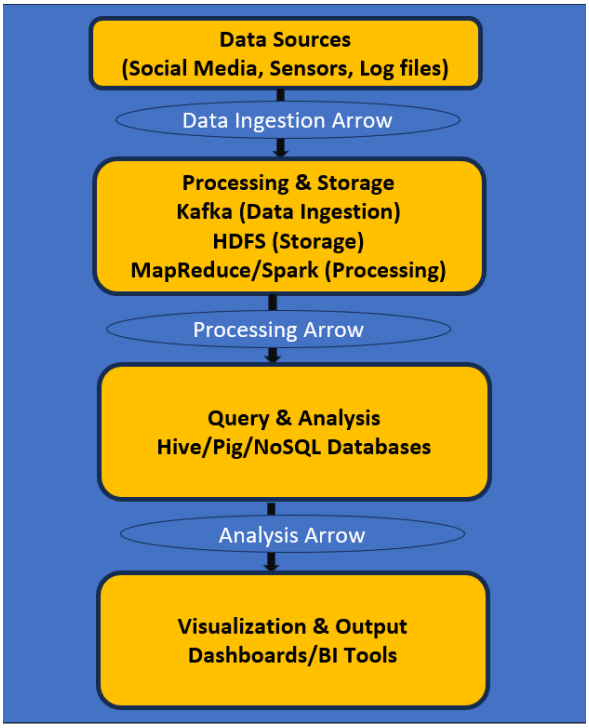

Fig. 2 Big Data Analytics Architecture - A layered model showing the data pipeline from ingestion to insights.

A. Distinctions and Challenges between Big Data and Real-Time Big Data:

Big data is defined by multi-source heterogeneity, extensive dispersion, dynamic expansion, and a "schema-after-the-data" methodology [8, 9]. Real-time big data possesses these characteristics but necessitates more rigorous requirements for data gathering equipment, analytical tools, and data security, especially for data integration, analysis, security, management, and benchmarking.

The fundamental difference in their processing paradigms is highlighted in Figure 3. Traditional

batch processing handles data in large, scheduled chunks, resulting in higher latency. In contrast, real-time processing deals with continuous data streams, enabling immediate analysis and action.

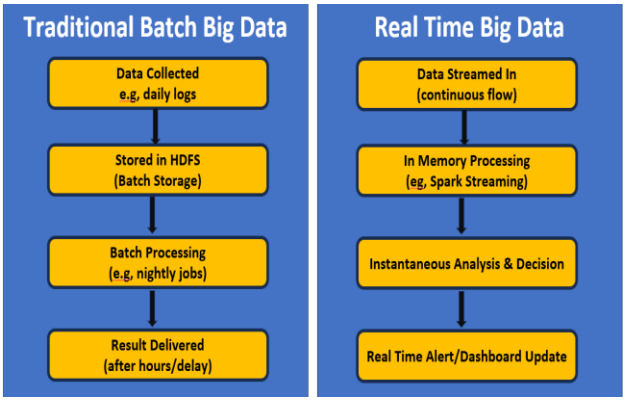

Fig. 3 Comparative Workflow: Traditional vs. Real-time Big Data Processing - Contrasting the high-latency batch approach with the low-latency streaming model.

B. Data Acquisition

The progression of the Internet of Things and Cyber-Physical Systems has resulted in elevated demands for data processing velocity. In a big data context, the extensive array of distributed sensors and mobile devices renders data collection a considerable issue. Real-Time Data Processing (RTDP) systems face significant hurdles in data integration due to the necessity for real-time data collecting.

C. Significant Variability

In conventional big data systems, data from mobile terminals and various devices is frequently retained in cache. Nonetheless, RTDP systems necessitate real-time data synchronization, imposing significant demands on wireless network transmission [12]. Although big data systems can employ NoSQL and storage technologies such as Hadoop HDFS to manage heterogeneity, these are optimized for "write-once, read-many" scenarios with minimal real-time requirements and are insufficient for the synchronization demands of real-time big data. The significant heterogeneity requires data transformation during integration, a procedure for which conventional data warehouses are too sluggish and constrained in capacity [12, 13, 14].

D. Data Quality Assurance

A prevalent occurrence in the big data era is the valuable information being obscured by an abundance of irrelevant data [15]. Ensuring data quality encompasses two primary challenges: the management of extensive datasets and their effective cleansing. The data cleaning procedure requires careful calibration; excessive granularity may eliminate valuable information, while insufficient granularity renders the cleaning inefficient. The trade-off between quantity and quality is especially evident in real-time big data systems, which must synchronize and respond to data nearly instantaneously. The performance criteria for data transmission and analytical velocity are exceptionally stringent. Furthermore, data excluded at one point may subsequently prove essential, posing a significant barrier in appropriately assessing data relevance and connection [15].

III. SIGNIFICANT OBSTACLES IN BIG DATA

Leveraging the capabilities of big data entails various hurdles, including the design of foundational processing systems and the formulation of advanced analytical techniques, as well as a multitude of unresolved research issues. These grand challenges can be broadly categorized into three areas of complexity, as summarized in Figure 4, which are often direct consequences of the 5V's.

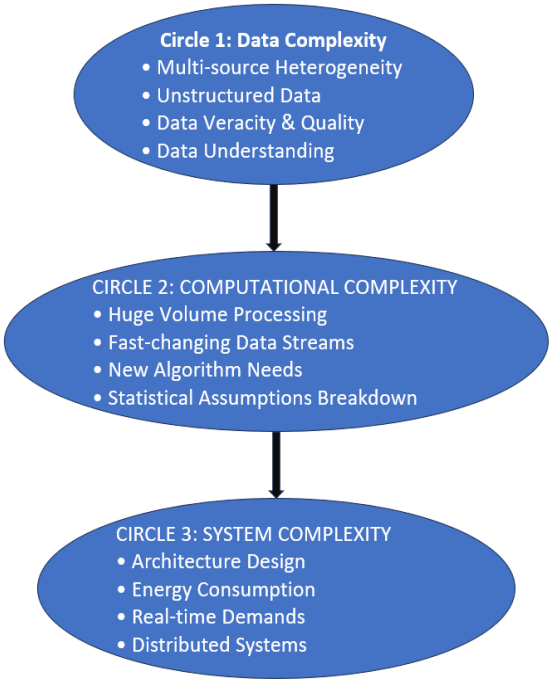

Figure 4: The Grand Challenges of Big Data Management - A Venn diagram representing the three intertwined domains of complexity that pose significant hurdles.

A. Complexity of Data

The examination of data complexity metrics is a burgeoning field in data mining, concentrating on the analysis of diverse data set attributes to facilitate the selection of the best suitable algorithms for information extraction.

B. Computational Complexity

The principal characteristics of big data—diverse sources, substantial volume, and rapid evolution—render standard computing techniques (such as machine learning and data mining) inadequate for effective processing and analysis [8]. These methods, predicated on assumptions such as independent and identically distributed data and sufficient sampling, frequently prove inadequate. Novel methodologies must reassess computability, computational complexity, and algorithms pertaining to huge data. These creative computing paradigms must exhibit great efficiency, offer novel approaches for processing and analyzing large datasets, and facilitate value-driven applications, while simultaneously tackling issues such as inadequate samples and ambiguous data linkages.

C. System Complexity

Creating processing systems that can manage various data kinds and applications is essential for facilitating big data research. The processing of large volumes of data with intricate structures and sparse values faces significant computing complexity, extended duty cycles, and real-time demands. These demands present novel problems for system architecture, computing frameworks, and processing system design, while simultaneously enforcing stringent limitations on operational efficiency and energy usage. The primary concern is the creation of highly energy-efficient systems, necessitating the development of effective system designs, computational frameworks, processing methodologies, and benchmarks. Addressing these issues is crucial for formulating guidelines for the design, implementation, and optimization of big data processing systems.

Section IV: Conclusion

Big data has significantly impacted almost every business and industry. Several variables are essential for the success of a big data project. Initially, there must be unequivocal requirements, whether technological, social, or economic. Secondly, it is essential to investigate and delineate the "kernel" data or structure—a sufficiently tiny subset that remains manageable while effectively representing the behaviors and attributes of the underlying big data, a task that is extremely particular to the area. Third, a top-down management style should be implemented, as a bottom-up strategy may produce fragmented answers that cannot be synthesized into a cohesive

whole. The objective should be to address the overall issue with a cohesive solution, rather than pursuing individual achievements in select areas.

V. REFERENCES

[1] N. Komninos, ―Intelligent cities: variable geometries of spatial intelligence, ‖Intelligent Buildings International, vol. 3, no. 3, pp. 172–188,2011.

[2] T. Kalil, "Big data is a big deal," The White House, 2012. [Online]. Available: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

[3] d. boyd and K. Crawford, "Six Provocations for Big Data," *Social Science Research Network*, SSRN Scholarly Paper ID 1926431, Sep. 2011. doi: 10.2139/ssrn.1926431.

[4] D. Gewirtz, "Big Data Analytics," DIY-IT, Apr. 20, 2016.

[5] P. Dave, "Big Data, SQL, SQL Server, SQL Tips and Tricks," Oct. 02, 2013.

[6] D. Maney, "3D data management: controlling data volume, velocity and variety," *Appl. Deliv. Strateg.*, vol. 949, p. 1, 2001.

[7] ImpactRadius, "The 7 V's of Big Data," [Online]. Available: https://www.impactradius.com/blog/7-vs-big-data

[8] Sage SE, "Volume, Velocity, Value," [Online]. Available: http://www.sagese.com/dosage/volume-velocity-value

[9] X.-P. Qin, H.-J. Wang, X.-Y. Du, and S. Wang, "Big Data Analysis—Competition and Symbiosis of RDBMS and MapReduce," *Journal of Software*, vol. 23, no. 1, pp. 32–45, 2012.

[10] X. Tan, H. Wang, F. Li et al., "New Landscape of Data Management Technologies," *Journal of Software*, vol. 24, no. 2, pp. 175–197, 2013.

[11] H.-M. Chen, L. Cui, and K.-B. Xie, "A Comparative Study on Architectures and Implementation Methodologies of Internet of Things," *Chinese Journal of Computers*, vol. 36, no. 1, pp. 168–188, 2013.

[12] E. A. Lee and S. A. Seshia, *Introduction to Embedded Systems: A Cyber-Physical Systems Approach*, 1st ed. Lee & Seshia, 2011.

[13] A. Thusoo, J. S. Sarma, N. Jain et al., "Hive—A Petabyte Scale Data Warehouse Using Hadoop," in *Proceedings of the 2010 IEEE 26th International Conference on Data Engineering (ICDE)*, 2010, pp. 996–1005.

[14] A. Abouzied, K. Bajda-Pawlikowski, J. Huang et al., "HadoopDB in Action: Building Real World Applications," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, 2010, pp. 1111–1114.

[15] S. Chen, "Cheetah: A High Performance, Custom Data Warehouse on Top of MapReduce," *Proc. VLDB Endow.*, vol. 3, no. 1–2, pp. 1459–1468, Sep. 2010.

[16] J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling, "Climate Data Challenges in the 21st Century," *Science*, vol. 331, no. 6018, pp. 700–702, Feb. 2011.