Purva Mimaansa | Refereed | Peer Reviewed
ISSN: 0976-0237 | Volume 12, Issue March 2021 | Impact Factor 4.115

A REVIEW PAPER ON BIG DATA, ARCHITECTURE AND PROBLEMS

SHIKHA VERMA
Assistant Professor, Deptt of Computer Science & Applications, Sanatan Dharma College, Ambala Cantt

ABSTRACT

The term 'Big Data' describes various innovative techniques and technologies to capture data, store data, distribute data, manage data and analyze data of size peta byte or large-size datasets with high-velocity and different structures of data. Big data can be structured, unstructured or semi-structured, resulting in incapability of inflexible data management methods or techniques. Data is collected from various types of sources and can arrive in the system at various sizes. In order to process these big amounts of data in a less cost and correct way, parallelism can be used. Big Data requires new architecture, new techniques, new algorithms, and new analytics to manage the large amount of data and extract values and unseen knowledge from large data.

INTRODUCTION

BIG DATA: DEFINITION

'Big Data' is similar to 'small amount of data', but size is bigger, bigger data requires different kinds of approaches, different techniques, different tools and different architecture. Its main aim is to solve new problems or previous problems in a good manner. Big Data generates data from the storage and processing of very large amount of digital information that cannot be analyzed with previous computing techniques. Some Examples of Big Data are: The Walmart take cares of more than 1 million customer transactions every hour, Facebook takes care of 40 billion photos from its user database. Decoding the human genome originally took more than 10 years to process now it can be done in just one week.

The current generation using so many social networking sites over the internet like YouTube, Facebook, LinkedIn, Instagram and fabricating abundance of data. There is also so many another source through which a lot of heterogeneous data is generating in the day to day activities like the sensors utilize for gathering the climate information, the data being generated in sports science. All these types of activities generate a mixture of data including structured data and unstructured data.

To understand the concept of big data, it has generally three aspects:

- The data is in abundance amount.
- The data couldn't be handled easily by the traditional relational databases technique.
- The data is generated, deposited, managed and treated so rapidly.

Big data is somehow related to the data mining association rule for storing the enormous amount of data and to extract the useful information and knowledge from it. But due to its volume, velocity, and variety, it is tough to discover any association rule and frequent itemset. Many association rules of data mining are available to associate the frequent item sets but no one is suitable to handle the big data.

Big data refers to the data sets that are too big to be handled using the existing database management tools and are emerging in many important applications, such as Internet search, business informatics, social networks, social media, genomics, and meteorology. In simple words, big data can be defined as any data which challenges the currently existing techniques for handling it. Big data presents a grand challenge for database and data analytics research.

We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes). So, it can be stated that

- Big data is what happened when the cost of storing information became less than the cost of making the decision to throw it away.
- Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.

The data can be broadly classified into three categories based on the source of origin which are as under.

1. **Human-sourced information**: All information ultimately originates from people. This information is the highly subjective record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video. Human-sourced information is now almost entirely digitized and electronically stored everywhere from tweets to movies. Structuring and standardizationfor example, modellingdefines a common version of the truth that allows the business to convert human-sourced information to more reliable process-mediated data. This starts with data entry and validation in operational systems and continues with the cleansing and reconciliation processes as data moves to business intelligence.
2. **Process-mediated data:** Business processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected is highly structured and includes transactions, reference tables and relationships, as well as the metadata that sets its context. Process-mediated data has long been the vast majority of what IT managed and processed, in both operational and BI systems.
3. **Machine-generated data**: the output of sensors and machines employed to measure and record the events and situations in the physical world is machine-generated data, and from simple sensor records to complex computer logs, it is well structured and considered to be highly reliable. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is amenable to computer processing, but its size and speed is often beyond traditional approaches—such as the enterprise data warehouse—for handling process-mediated data; standalone high-performance relational and NoSQL databases are regularly used.

Figure No. 1 gives Layered Architecture of Big Data System. It can be decomposed into three layers, including Infrastructure Layer, Computing Layer, and Application Layer from top to bottom.

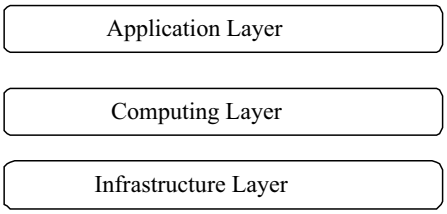
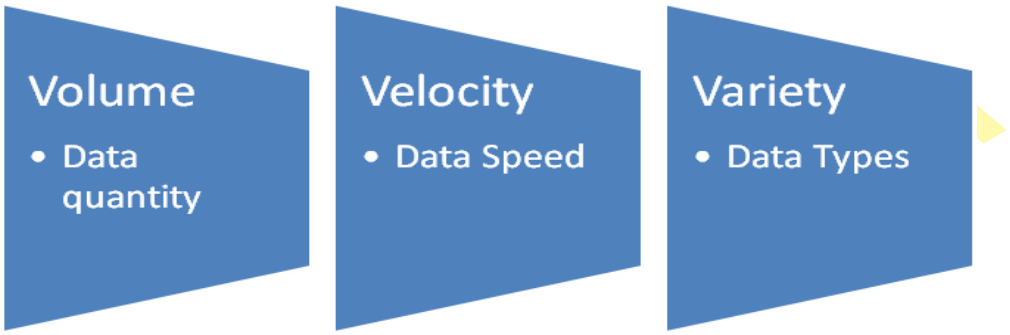


Figure 1: Layered Architecture of Big Data System

3 Vs OF BIG DATA



Volume of data: A typical PC might have had 10 gigabytes of storage in 2000, now this data increases as Facebook ingests 500 terabytes of new data every day. Another Example is, Boeing 737 will generate 240 Tb of flight data at a single flight across the US. The smart mobile phones, the data they use sensors embedded into everyday objects will generates billions of new data, constantly-updated data feeds containing environmental, location, and other information, including video.

Variety of data: Big Data includes different types of data like numbers, dates, and strings. Big Data is also includes geospatial data, 3D data, audio and video, and unstructured data; it includes log files and social media files. Old database systems were designed to address smaller amount of structured data, less updates or a predictable data , consistent data structure. Big Data analysis includes different varieties of data.

Velocity of data: Velocity includes speed of data processing. For time-based processes such as catching fraud, big data is used because it streams into your enterprise in order to increase its value, online gaming support millions of users, each providing multiple inputs per second.

PROBLEM WITH BIG DATA PROCESSING

i. Heterogeneity and Incompleteness

Big Data deals with large amount of data. Big data seeks to explore complex and evolving relationships among large amount of data. It contains heterogeneous and autonomous data with

distributed and decentralized control. Hence, it is an extreme challenge for finding useful knowledge from the Big Data

ii. Scale

Yes, the first thing we can think of with Big Data is its size of data. As, its name includes the word "big" .To Manage large and continuously increasing amount of data has been a main concern for all users. Previously, this concern was done by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing amount of data. Because, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are fix.

iii. Timeliness

The flip side of size includes speed. The time required for analyzing the data is proportional to the size of the data set to be processed. The system design that deals with size of the amount of data will also result in a system that can process a given size of data set more quickly. In many situations result of the analysis is required immediately. We have to develop partial results in advance so that a small amount of incremental computation with new data can be used at a quick determination. In a large data set, it is often necessary to find elements in it that follows a specified criterion. It is not easy to scan the entire data set. With new analysis on Big Data, new types of criteria are used, and new index structures to support such criteria are needed to be devised. Designing such structures becomes particularly challenging as, the way the amount of data is growing rapidly; to design such structures is not an easy task.

iv. Privacy

The privacy of large amount of data is another big concern, and one that increases in the terms of Big Data. There are strict laws governing what can and cannot be done for electronic health records. There is public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy in big data is effectively both a technical and a sociological problem.

v. Human Collaboration

In spite of the immense advances made in computational analysis, there are many patterns or techniques that humans can easily find but computer algorithms have a hard time find that patterns. In today's complex world, it often takes multiple experts from different domains or technologies to really understand what is going on. A Big Data analysis system must support input from multiple human experts from various domains, and shared exploration of results. These multiple experts may be separated in space and time when it is too low cost to assemble an entire team together in one room.

REFERENCES

A. Gandomi, M. Haider. (2015). "Beyond the hype: big data concepts, methods, and analytics", International Journal of Information Management, 35(2), pp. 137-144.

Albert Bifet. (2013). "Mining Big Data In Real Time"Informatica, 37, 15–20 DEC 2012.

Bernice Purcell. "The emergence of "big data"technology and analytics" Journal of Technology Research 2013.

D. Che, M. Safran, Z. Peng. (2013). "From big data to big data mining: challenges, issues, and

opportunities", International Conference on database systems for Advanced Applications. 7827, pp. 1-15.

J. Dean, S. Ghemawat. (2004). "MapReduce: Simplified data processing on large clusters", Communications of the ACM, 53(1), pp. 1-13.

Jimmy Lin "Map Reduce Is Good Enough?" The control project. IEEE Computer 32 (2013).

Kumara Reddi & Dnvsl Indira "DifferentTechnique to Transfer Big Data: Survey" IEEETransactions on 52(8) (Aug.2013) 2348.

L. Greeshma, G. Pradeepini. (2016). "Big data analytics with apache Hadoop MapReduce framework", International Journal of Science and Technology, 9(26), pp. 1-6.

N. Khan, I. Yaqoob, I. A. T. Hashem, Z. Inayat, W. K. M. Ali, M. Alam, M. Shiraz, A. Gani. (2014). "Big data: Survey, technologies, opportunities, and challenges", The Scientific World Journal, pp. 1-18.

N. Lounes, H. Oudghiri, R. Chahal, W.K. Hidouci. (2018). "From KDD to KUBD: big data characteristics within the KDD process steps", Springer World conference on information systems and technologies, 746, pp. 931-937.

R. Dhaka, A. Kumar. (2018). "Need and application of data mining", International Journal of Innovations & Advancement in Computer Science, 7(4), pp. 166-169.

R. Rein, D. Memmert. (2016). "Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science",SpringerPlus, 5,pp. 1-13.

S. Agarwal, L. Yadav, S. Mehta. (2017). "Cricket team prediction with Hadoop: statistical modelingapproach", Procedia Computer Science, 122, pp. 525-532.

S. Anjali, V. Aswini, M. Abirami. (2015). "Predictive analysis with cricket tweets using big data", International Journal of Scientific & Engineering Research, 6(10), pp. 78-83.

S. Mukherjee, R. Shaw. (2016). "Big data- concepts, applications, challenges and future scope", International Journal of Advance Research in Computer and Communication Engineering, 5(2), pp. 66-73.

S.G Manikandan, S. Ravi. (2014). "Big data analysis using apache Hadoop", IEEE International Conference on IT Convergence and Security, pp. 1-4.

S.V. Phaneendra, E.M. Reddy. (2013). "Big data- solutions for RDBMS problems- A survey", International Journal of Advance Research in Computer and Communication Engineering, 2(9), pp. 3686-3691.

S.Vikram Phaneendra & E.Madhusudhan Reddy. "Big Data- solutions for RDBMS problems- A survey" In 12thIEEE/IFIP Network Operations &Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19(23 2013).

Shweta, K. Garg. (2013). "Mining efficient association rules through Apriori algorithm using attributes and comparative analysis of various association rule algorithms", International Journal of Advance Research in Computer Science and Technology, 3(6), pp. 306-312.

T. Ahlawat, Dr. R. K. Rambola. (2016). "Literature review on big data", International Journal of Advancement in Engineering Technology, Management & Applied Science, 3(5), pp. 21-30.