International Journal of Education and Science Research Review

Volume-9, Issue-4 July-August- 2022

E-ISSN 2348-6457
P-ISSN 2349-1817

www.ijesrr.org Email- editor@ijesrr.org

A STUDY ON THE MANAGEMENT OF BIG DATA

Santhosh Kumar Adama
Research scholar
Department of CSE
Benguluru University

Dr. Rajneesh Kumar
Professor
Department of CSE
Benguluru University

ABSTRACT

The rapid growth of emerging applications and the evolution of cloud computing technologies have significantly enhanced the capability to generate vast amounts of data. Thus, it has become a great challenge in this big data era to manage such voluminous amount of data. The recent advancements in big data techniques and technologies have enabled many enterprises to handle big data efficiently. However, these advances in techniques and technologies have not yet been studied in detail and a comprehensive survey of this domain is still lacking. With focus on big data management, this survey aims to investigate feasible techniques of managing big data by emphasizing on storage, pre-processing, processing and security. Moreover, the critical aspects of these techniques are analysed by devising a taxonomy in order to identify the problems and proposals made to alleviate these problems. Furthermore, big data management techniques are also summarized.

KEYWORDS:

Big, Data, Management

INTRODUCTION

Big data is usually complex -- in addition to its volume and variety, it often includes streaming data and other types of data that are created and updated at a high velocity. As a result, processing and managing big data are complicated tasks. For data management teams, the biggest challenges faced on big data deployments include the following:

- **Dealing with the large amounts of data.** Sets of big data don't necessarily need to be large, but they commonly are -- and in many cases, they're massive. Also, data frequently is spread across different processing platforms and storage repositories. The scale of the data volumes that typically are involved makes it difficult to manage all of the data effectively.

- **Fixing data quality problems.** Big data environments often include raw data that hasn't been cleansed yet, including data from different source systems that may not be entered or formatted consistently. That makes data quality management a challenge for teams, which need to identify and fix data errors, variances, duplicate entries and other issues in data sets.

- **Integrating different data sets.** Similar to the challenge of managing data quality, the data integration process with big data is complicated by the need to pull together data from various sources for analytics uses. In addition, traditional extract, transform and load (ETL) integration approaches often aren't suited to big data because of its variety and processing velocity.

- **Preparing data for analytics applications.** Data preparation for advanced analytics can be a lengthy process, and big data makes it even more challenging. Raw data sets often must be consolidated, filtered, organized and validated on the fly for individual applications. The distributed nature of big data systems also complicates efforts to gather the required data.

- **Ensuring that big data systems can scale as needed.** Big data workloads require a lot of processing and storage resources. That can strain the performance of big data systems if they aren't designed to deliver the required processing capacity. It's a balancing act, though: Deploying systems with excess capacity adds unnecessary costs for businesses.

- **Governing sets of big data.** Without sufficient data governance oversight, data from different sources might not be harmonized, and sensitive data might be collected and used improperly. But governing big data environments creates new challenges because of the unstructured and semi structured data they contain, plus the frequent inclusion of external data sources.

The world is driven by data, and it’s being analysed every second, whether it’s through your phone’s Google Maps, your Netflix habits, or what you’ve reserved in your online shopping cart.

Data lakes must also be carefully managed in order not to become “data swamps”—lakes with low-quality, poorly catalogued data that can’t be easily accessed. And at some point, most unstructured data based in a data lake will need to be put in structured form in order to be analysed. Data lakes, then, require that management approaches be defined in advance to ensure quality, accessibility, and necessary data transformations.

Deloitte helped one global technology firm, for example, transition from a 600 terabyte enterprise data warehouse to a data lake platform. The data is used by 2,800 employees, so the conversion process needed to involve minimal disruption. Lake storage still uses on premise technologies, but the company now has a "consumption layer" in the cloud for easy and rapid access by users and automated processes. And instead of the time-honoured "extract, transform, and load" (ETL) process, data is only transformed when necessary for analysis. In other words, it's an ELT process.

MANAGEMENT OF BIG DATA

Most organizations establishing data modernization approaches also try not to lift and shift existing data into the new data environment. Instead, they attempt to make improvements in the data at the same time, increasing integration and quality across the enterprise. Firms are increasingly using tools like machine learning to allow probabilistic matching of data; using this approach, data that is similar but not exactly the same as other data can be matched and combined with little human intervention. This bottom-up method of data integration can sometimes be faster and more effective than more top-down approaches to integration like Master Data Management.

There's much wisdom in that saying, which has been attributed to both W. Edwards Deming and Peter Drucker, and it explains why the recent explosion of digital data is so important. Simply put, because of big data, managers can measure, and hence know, radically more about their businesses, and directly translate that knowledge into improved decision making and performance.

Consider retailing. Booksellers in physical stores could always track which books sold and which did not. If they had a loyalty program, they could tie some of those purchases to individual customers. And that was about it. Once shopping moved online, though, the understanding of customers increased dramatically. Online retailers could track not only what customers bought, but also what else they looked at; how they navigated through the site; how much they were influenced by promotions, reviews, and page layouts; and similarities across individuals and groups. Before long, they developed algorithms to predict what books individual customers would like to read next—algorithms that performed better every time the customer responded to or ignored a recommendation. Traditional retailers simply couldn't access this kind of information, let alone act on it in a timely manner. It's no wonder that Amazon has put so many brick-and-mortar bookstores out of business.

The familiarity of the Amazon story almost masks its power. We expect companies that were born digital to accomplish things that business executives could only dream of a generation ago. But in fact the use of big data has the potential to transform traditional businesses as well. It may offer them even greater opportunities for competitive advantage (online businesses have always known that they were competing on how well they understood their data). As we'll discuss in more detail, the big data of this revolution is far more powerful than the analytics that were used in the past. We can measure and therefore manage more precisely than ever before. We can make better predictions and smarter decisions. We can target more-effective interventions, and can do so in areas that so far have been dominated by gut and intuition rather than by data and rigor.

As the tools and philosophies of big data spread, they will change long-standing ideas about the value of experience, the nature of expertise, and the practice of management. Smart leaders across industries will see using big data for what it is: a management revolution. But as with any other major change in business, the challenges of becoming a big data–enabled organization can be enormous and require hands-on—or in some cases hands-off—leadership. Nevertheless, it's a transition that executives need to engage with today.

Contemporary big data initiatives in health care will benefit from greater integration with nursing science and nursing practice; in turn, nursing science and nursing practice has much to gain from the data science initiatives. Big data arises secondary to scholarly inquiry (e.g., -omics) and everyday observations like cardiac flow sensors or Twitter feeds. Data science methods that are emerging ensure that these data be leveraged to improve patient care. Big data encompasses data that exceed human comprehension, that exist at a volume unmanageable by standard computer systems, that arrive at a velocity not under the control of the investigator and possess a level of imprecision not found in traditional inquiry. Data science methods are emerging to manage and gain insights from big data. The primary methods included investigation of emerging federal big data initiatives, and exploration of exemplars from nursing informatics research to benchmark where nursing is already poised to participate in the big data revolution. We provide observations and reflections on experiences in the emerging big data initiatives. Existing approaches to large data set analysis provide a necessary but not sufficient foundation for nursing to participate in the big data revolution. Nursing's Social Policy Statement guides a principled, ethical perspective on big data and data science. There are implications for basic and advanced practice clinical nurses in practice, for the nurse scientist who collaborates with data scientists, and for the nurse data scientist. Big data and data science has the potential to provide greater richness in understanding patient phenomena and in tailoring interventional strategies that are personalized to the patient.

Reliability field data such as that obtained from warranty claims and maintenance records have been used traditionally for such purposes as generating predictions for warranty costs and optimizing the cost of system operation and maintenance. In the current (and future) generation of many products, the nature of field reliability data is changing dramatically. In particular, products can be outfitted with sensors that can be used to capture information about how and when and under what environmental and operating conditions products are being used. Today some of that information is being used to monitor system health and interest is building to develop prognostic information systems. There are, however, many other potential applications for using such data. In this article we review some applications where field reliability data are used and explore some of the opportunities to use modern reliability data to provide stronger statistical methods to operate and predict the performance of systems in the field. We also provide some examples of recent technical developments designed to be used in such applications and outline remaining challenges.

There is a trend that, virtually everyone, ranging from big Web companies to traditional enterprisers to physical science researchers to social scientists, is either already experiencing or anticipating unprecedented growth in the amount of data available in their world, as well as new opportunities and great untapped value. This paper reviews big data challenges from a data management respective. In particular, we discuss big data diversity, big data reduction, big data integration and cleaning, big data indexing and query, and finally big data analysis and mining.

The rapid growth of emerging applications and the evolution of cloud computing technologies have significantly enhanced the capability to generate vast amounts of data. Thus, it has become a great challenge in this big data era to manage such voluminous amount of data. The recent advancements in big data techniques and technologies have enabled many enterprises to handle big data efficiently. However, these advances in techniques and technologies have not yet been studied in detail and a comprehensive survey of this domain is still lacking. With focus on big data management, this survey aims to investigate feasible techniques of managing big data by emphasizing on storage, pre-processing, processing and security. Moreover, the critical aspects of these techniques are analyzed by devising a taxonomy in order to identify the problems and proposals made to alleviate these problems.

DISCUSSION

Big data with its vast volume and complexity is increasingly concerned, developed and used for all professions and trades. Remote sensing, as one of the sources for big data, is generating earth-observation data and analysis results daily from the platforms of satellites, manned/unmanned aircrafts, and ground-based structures. Agricultural remote sensing is one of the backbone technologies for precision agriculture, which considers within-field variability for site-specific management instead of uniform management as in traditional agriculture. The key of agricultural remote sensing is, with global positioning data and geographic information, to produce spatially-varied data for subsequent precision agricultural operations. Agricultural remote sensing data, as general remote sensing data, have all characteristics of big data. The acquisition, processing, storage, analysis and visualization of agricultural remote sensing big data are critical to the success of precision agriculture. This paper overviews available remote sensing data resources, recent development of technologies for remote sensing big data management, and remote sensing data processing and management for precision agriculture. A five-layer-fifteen-level (FLFL) satellite remote sensing data management structure is described and adapted to create a more appropriate four-layer-twelve-level (FLTL) remote sensing data management structure for management and applications of agricultural remote sensing big data for precision agriculture where the sensors are typically on high-resolution satellites, manned aircrafts, unmanned aerial vehicles and ground-based structures. The FLTL structure is the management and application framework of agricultural remote sensing big data for precision agriculture and local farm studies, which outlooks the future coordination of remote sensing big data management and applications at local regional and farm scale.

Big Data is an emerging paradigm in almost all industries. Finance big data (FBD) is becoming one of the most promising areas of management and governance in the financial sector. It is significantly changing business models in financial companies. Many researchers argue that Big Data is fuelling the transformation of finance and business at-large in the ways that we cannot as yet assess. A new research area is evolving to study quantitative models and econometric approaches for financial studies that can bridge the gap between empirical finance research and data science. In this fascinating area, experts and scientists can propose novel finance business models by using the Big Data methods, present sophisticated methods for risk control with machine learning tools, provide visualization tools for financial markets analysis, create new finance sentiment indexes by mining public feelings from the massive textual data from social networks, and deploy the information-based tools in other creative ways.

Due to the 4V characteristics of Big Data—volume (large data scale), velocity (real-time data streaming), variety (different data formats), and veracity (data uncertainty)—a long list of challenges for FBD management, analytics, and applications exists. These challenges include (1) to organize and manage FBD in effective and efficient ways; (2) to find novel business models from FBD analytics; (3) to handle traditional finance issues like high-frequency trading, sentiments, credit risk, financial analysis, risk management and regulation, and others, in creative Big Data–driven ways; (4) to integrate the variety of heterogeneous data from different sources; and (5) to ensure the security and safety of finance systems and to protect the individual privacy in view of the availability of Big Data. To meet these challenges, we need fundamental research on both data analytics technology and finance business.

Big Data is relatively a new concept which refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. The accumulated huge amount of data that previously of no significant importance or value have been put into maximum use due to the availability of newly designed Big Data tools that surpass earlier available data mining tools. Big Data is now of tremendous importance to organizations and data mining researchers because better results are gotten from larger volume of data. Predictions and Analysis of business are becoming more accurate and interesting with the advent of Big Data Tools. The scale and scope of changes that Big Data are bringing about are at an inflection point, set to expand greatly, as a series of technology trends accelerate and courage. In this paper, we introduced readers to the concept of Big Data, the various sources of data for Big Data. Some of the advantages and applications that have been successfully implemented using Big Data tools.

CONCLUSION

Big Data encompasses collection, management, processing and analysis of the huge amount of data that varies in types and changes with high frequency. Often data component of Big Data has a positional component as an important part of it in various forms, such as postal address, Internet Protocol (IP) address and geographical location. If the positional components in Big Data extensively used in storage, retrieval, analysis, processing, visualization and knowledge discovery (geospatial Big Data) the Big Data systems need certain type of techniques and algorithms for management, analytics and sharing.

REFERENCES

1. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. In: McKinsey Global Institute Reports, pp. 1–156 (2011)

2. Mouthami, K., Devi, K.N., Bhaskaran, V.M.: Sentiment Analysis and Classification Based on Textual Reviews. In: International Conference on Information Communication and Embedded Systems (ICICES), pp. 271–276 (2013)

3. Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011)

4. Russom, P.: Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40 (2011)

5. Sanchez, D., Martin-Bautista, M.J., Blanco, I., Torre, C.: Text Knowledge Mining: An Alternative to Text Data Mining. In: IEEE International Conference on Data Mining Work- shops, pp. 664–672 (2008)

6. Serrat, O.: Social Network Analysis. Knowledge Network Solutions 28, 1–4 (2009)

7. Shen, Z., Wei, J., Sundaresan, N., Ma, K.L.: Visual Analysis of Massive Web Session Data. In: Large Data Analysis and Visualization (LDAV), pp. 65–72 (2012)

8. Song, Z., Kusiak, A.: Optimizing Product Configurations with a Data Mining Approach. International Journal of Production Research 47(7), 1733–1751 (2009)

9. TechAmerica: Demystifying Big Data: A Practical Guide to Transforming the Business of Government. In: TechAmerica Reports, pp. 1–40 (2012)

10. Van der Valk, T., Gijsbers, G.: The Use of Social Network Analysis in Innovation Studies: Mapping Actors and Technologies. Innovation: Management, Policy & Practice 12(1), 5–17 (2010)

11. Zeng, D., Hsinchun, C., Lusch, R., Li, S.H.: Social Media Analytics and Intelligence. IEEE Intelligent Systems 25(6), 13–16 (2010)

12. Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., Keim, D.: Visual Analytics for the Big Data Era—A Comparative Review of State-of-the-Art Commercial Systems. In: IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173–182 (2012)