IJAER/ November-December 2021 /Volume-10/Issue-3 ISSN: 2278-9677
International Journal of Arts & Education Research

A STUDY ON THE CHALLENGES RELATED TO 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 privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data. Not only the storage of data is essential but the privacy and integrity of big data should also be ensured in Big Data Management. Only if the privacy of data will be maintained then only it can be assured that data is true and accurate. Management of Big Data does not only cover the area of managing Big Data. It also helps in organizing and retrieving of large amount of data which includes all types of structured and unstructured data. Unstructured data includes all types of data that is not arranged in any sort of manner and is irregular.

KEYWORDS:
Big, Data, Management

INTRODUCTION

Governance of Big Data is also the most important aspect under the management of Big Data. Governance not deals with the storage of data but it also deals with the different types of risks associated with the management of Big Data. Governance ensuring that corporate and governmental rules and policies are adhered to using

policies, processes, and controls. It helps us from the unauthorized access to the Big Data which needs full security and supervision. This feature of governance is provided by the Big Data management.

There's a wide variety of platforms and tools for managing big data, with both open source and commercial versions available for many of them. The list of big data technologies that can be deployed, often in combination with one another, includes distributed processing frameworks Hadoop and Spark; stream processing engines; cloud object storage services; cluster management software; NoSQL databases; data lake and data warehouse platforms; and SQL query engines.

To enable easier scalability and more flexibility on deployments, big data workloads increasingly are being run in the cloud, where businesses can set up their own systems or use managed services offerings. Prominent big data management vendors include cloud platform market leaders AWS, Google and Microsoft, plus Cloudera, Databricks and others that focus mainly on big data applications.

Mainstream data management tools are also key components for managing big data. That includes data integration software supporting multiple integration techniques, such as traditional ETL processes; an alternative ELT approach that loads data as is into big data systems so it can be transformed later as needed; and real-time integration methods, such as change data capture. Data quality tools that automate data profiling, cleansing and validation are commonly used, too.

The global big data market revenues for software and services are expected to increase from $42 billion to $103 billion by year 2027. Every day, 2.5 quintillion bytes of data are created, and it’s only in the last two years that 90% of the world’s data has been generated. If that’s any indication, there’s likely much more to come.

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 – in many ways, data is unavoidable and it’s disrupting almost every known market. The business world is looking to data for market insights and ultimately, to generate growth and revenue. Although data is becoming a game changer within the business arena, it’s important to note that data is also being utilised by small businesses, corporate and creative alike. A global survey from McKinsey revealed that when organisations use data, it benefits the customer and the business by generating new data-driven services, developing new business models and strategies, and selling data-based products and utilities. The incentive for investing and implementing data analysis tools and techniques is huge, and businesses will need to adapt, innovate, and strategies for the evolving digital marketplace.

Every day, 2.5 quintillion bytes of data are created, and it's only in the last two years that 90% of the world's data has been generated.

Data analysis, or analytics (DA) is the process of examining data sets (within the form of text, audio and video), and drawing conclusions about the information they contain, more commonly through specific systems, software, and methods. Data analytics technologies are used on an industrial scale, across commercial business industries, as they enable organisations to make calculated, informed business decisions.

Globally, enterprises are harnessing the power of various different data analysis techniques and using it to reshape their business models. As technology develops, new analysis software emerge, and as the Internet of Things (IoT) grows, the amount of data increases. Big data has evolved as a product of our increasing expansion and connection, and with it, new forms of extracting, or rather "mining", data.

CHALLENGES RELATED TO THE MANAGEMENT OF BIG DATA

Big data management is the organization, administration and governance of large volumes of both structured and unstructured data.

The goal of big data management is to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications. Corporations, government agencies and other organizations employ big data management strategies to help them contend with fast-growing pools of data, typically involving many terabytes or even petabytes stored in a variety of file formats. Effective big data management particularly helps companies locate valuable information in large sets of unstructured and semi structured data from various sources, including call detail records, system logs, sensors, images and social media sites.

Most big data environments go beyond relational databases and traditional data warehouse platforms to incorporate technologies that are suited to processing and storing no transactional forms of data. The increasing focus on collecting and analysing big data is shaping new data platforms and architectures that often combine data warehouses with big data systems.

As part of the big data management process, companies must decide what data must be kept for compliance reasons, what data can be disposed of and what data should be analysed in order to improve current business

processes or provide a competitive advantage. This process requires careful data classification so that, ultimately, smaller sets of data can be analysed quickly and productively.

In the IDG 2016 Data and Analytics Survey, 90 percent of those surveyed said they had experienced challenges related to big data management. Several different factors make big data management more challenging than managing smaller repositories of data. Common issues include the following:

1. **Data silos:** Within most organizations, different departments and business units use different applications and store information in separate databases. These separate databases may include similar information, but the data isn't always consistent from one database to another. For example, a retailer may store customer addresses in a marketing database, a customer service database, an accounting database and an ecommerce website database. If just one of those databases has slightly different information for a particular customer — such as listing a customer's street address as an "avenue" when the other databases list it as a "street" — it could lead to problems like duplicate mailings, losing track of customer service records, double billing or inaccurate reporting. In addition, storing the same piece of information in several different locations eats up storage space — particularly when the problem is multiplied across an entire customer base.

And unfortunately, siloed data is a very common problem for enterprises. The 2016 Enterprise Data Management Survey from Unisphere Research found that 59 percent of those surveyed had very little or only a few of their data systems integrated while most data still resided in siloes.

2. **Growing data stores:** Managing big data is also difficult because of the sheer size of the data involved, compounded by the fact that it keeps getting bigger. In the customer address example above, fixing the customer records would be fairly easy for a very small company with only a hundred customers. In that case, someone could just look at the records involved and fix them. But for a national retailer with millions of customers and multiple petabytes worth of data, a different solution is necessary.

In some cases, simply moving data around, say from a database into an analytics solution, can take a long time because of the large quantities involved. And performing any sort of operations on that data can slow performance to a crawl.

3. **Data and architectural complexity:** Not only is enterprise data stored in disparate siloes and constantly growing, today's data can be extremely complex. Enterprises often have both structured data (data that resides

in a database) and unstructured data (data contained in text documents, images, video, sound files, presentations, etc.), and that data resides in a wide variety of different formats. A single enterprise may have thousands of applications on its systems, and each of those applications may read from and write to many different databases. As a result, simply cataloguing what kinds of data an organization has in its storage systems can be a very difficult job.

4. **Ensuring data quality:** All of these challenges make it very difficult for enterprises to ensure that their data is reliable and accurate. As already mentioned above, the lack of synchronization across data silos can make it difficult for managers to know which piece of data is correct. But data quality is also affected by another big problem — human error.

In the Experian study, 56 percent of those surveyed said that human error was the biggest challenge that affected their data accuracy. Everyone makes mistakes when typing. But when a data-driven organization is using information typed by humans as the basis for major business decisions, simple typos could have potentially disastrous consequences.

5. **Inadequate staffing:** Another big issue complicating big data management is a lack of trained staff. There simply aren't enough data scientists and other big data professionals to fill all the available positions. As a result, salaries tend to be quite high. According to Indeed.com, the current average salary for a data scientist is $130,235, while a data warehouse engineer typically makes $112,607. By comparison, software engineers, which are generally some of the best paid employees in an IT department, earn an average of $100,512. Clearly, the demand for big data skills exceeds the available supply.

6. **Lack of executive support:** Another potential challenge for big data management efforts is senior managers who do not understand the importance and value of good data management. Flashier technologies like predictive analytics and artificial intelligence may get a lot of attention — and budget — while the mundane processes of moving and cleaning data don't generate as much excitement.

However, this problem appears to be diminishing somewhat. In the Experian study, the number of respondents citing inadequate senior management support as a big challenge to data management diminished from 21 percent in 2016 to 19 percent in 2017. And in the New Vantage Partners Big Data Executive Survey 2017, 52.5 percent of executives said that data governance was critically important to big data business adoption.

7. **Establishing a data-friendly culture:** For any organization, moving from a culture where people made decisions based on their gut instincts, opinions or experience to a data-driven culture marks a huge transition. The New Vantage study found that 52.5 percent of executives pointed to "organizational impediments" as a reason they had failed to achieve the goals of their big data projects, and only 27.9 percent of those surveyed said that they had been successful in their efforts to establish a data-driven culture. Changing the mindset of employees and managers takes time, but most experts agree that it is necessary for big data management to be effective.

DISCUSSION

Today, data is flowing into various organizations at an unprecedented scale. The ability to scale out for processing an enhanced workload has become an important factor for the proliferation and popularization of database systems. Big data applications demand and consequently lead to the developments of diverse large-scale data management systems in different organizations, ranging from traditional database vendors to new emerging Internet-based enterprises.

In today's 21st century as technology is getting so much advanced, Apache Kafka emanate as one of the finest technology in the present world. Its fast, scalable, distributed stream processing platform and fault tolerant messaging system has made this technology to roar in the field of data processing and analysis. Apache Kafka is a distributed streaming platform mainly designed for low latency and high throughput. It is publish-subscribing messaging reassess as a constituent to each of number of legatee of commit log. The key notion of Apache Kafka is that it is used as a cluster on any number of servers. Server of Kafka stores record streams in classes known as topics. Every record contains a key, a value and a time stamp. It has two classes of application. Firstly, for building pipelines of real time data streams which is reliable to get the data between the systems or between the applications. Secondly, build applications streaming for real time that reacts to the record streams. A single Kafka mediator can handle hundreds of megabytes of reads and writes per second from thousands of clients.

CONCLUSION

Big data is characterised by the three V's: the major volume of data, the velocity at which it's processed, and the wide variety of data. It's because of the second descriptor, velocity, that data analytics has expanded into the technological fields of machine learning and artificial intelligence. Alongside the evolving computer-based

analysis techniques data harnesses, analysis also relies on the traditional statistical methods. Ultimately, how data analysis techniques function within an organisation is twofold; big data analysis is processed through the streaming of data as it emerges, and then performing batch analysis' of data as it builds – to look for behavioural patterns and trends. As the generation of data increases, so will the various techniques that manage it. As data becomes more insightful in its speed, scale, and depth, the more it fuels innovation.

REFERENCES

1. Mircea Raducu TRIFU, Mihaela laura IVAN "Big Data: present and future", Database Systems Journal, Vol. 1, no.1, 2014.
2. Hiba Jasim, Ammar Hameed Shnain, Sarah Hadishaheed, Azizahbt Haji Ahmad "Big Data and Five V'S Characteristics", International Journal of Advances in Electronics and Computer Science, ISSN:2393-2835, Vol. 2, Issue-1, 2015.
3. Ishwarappa, Anuradha J, "A Brief Introduction on 5Vs Characteristics and Hadoop Technology, Science Direct, Procedia Computer Science 48, p319-324, 2015.
4. Ishan Verma, Lokendra Singh, "Multi-structured Data Analytics using Interactive Visualization to aid Business Decision Making", ACM COMPUTE, Bhopal, India, November 16–17,2017.
5. Owen O'Malley, Kan Zhang, Sanjay Radia, Ram Marti, and Christopher Harrell, "Hadoop Security Design" ,Yahoo, Inc., Technical Report, October,2009.
6. Dr Satyam Priyadarshy, "The 7 pillars of Big Data", Petrolium Review, 14 january, 2015.
7. George Firican, "The 10 V'S BIG DATA", Work Paper, February 8, 2017.
8. Dr Kirk Borne, " Top 10 Big Data challenges-A serious look at 10 big Data V's", blog post , April 11, 2014.
9. William Vorhies,View Blog "How Many V'S in Big Data", Work Paper, October 31, 2014.
10. Abhinav Rai, "What is Big Data – Characteristics, Types, Benefits & Examples", Work paper, June 2019.