

## **Challenges in Big Data Analytics and Cloud Computing**

#### P. Madana Mohan, B. Murali Manohar

VIT Business School, VIT University, Vellore, Tamil Nadu, India

\*Correspondence: P. Madana Mohan, Email: madanamohan.p2018@vitstudent.ac.in

**ABSTRACT:** The advancement of technology has brought up so many elements of ease for human beings that now humans cannot afford to think of a life without all these. But all is not about ease and comfort. Along with all these elements, there are many modern complexities as well as challenges too. Big data analytics and cloud computing are among the elements in which we face many challenges. This document discusses some of the challenges in big data analytics and cloud computing.

Keywords: Cloud computing, Big data analytics, Challenges, Computing, Technology

#### ARTICLE INFORMATION



### 1. INTRODUCTION

The increase in cloud computing (Cloud Computing) has been a precursor and facilitator for the emergence of Big Data. While Big Data brings many attractive opportunities, companies also face many challenges like Data collection, storage, research, sharing, analysis, and visualization. Each of these tasks requires different approaches, different levels of security, infrastructure, and trained professionals. In recent years, a large number of Big Data techniques and technologies have developed to overcome all these obstacles. And Cloud Computing has been the necessary support for the growth in the adoption of a data-driven culture that allows companies to extract from Big Data the necessary insights for more conscious decision making [1].

There are significant advantages in adopting Cloud Computing over traditional physical deployments. However, cloud platforms have multiple structures and sometimes need to be integrated with traditional architectures, as all cloud computing services work slightly differently, depending on the provider. According to the June 2017 Gartner report, AWS (Amazon Web Service) remains the world leader in providing cloud services [2]. This leads to a dilemma for decisionmakers responsible for large data projects. How and which Cloud Computing Provider is the ideal choice for computing needs, especially if the company is starting a Big Data project? These projects are sensitive, and if they are not well dimensioned, they can lead to an explosion in need of storage and processing, seriously compromising the cost involved in the project. At the same time, stakeholders and business areas expect fast, cheap, and reliable products and, of course, the results of the project [3].

### 2. CLOUD COMPUTING

In a simplified way, cloud computing (Cloud Computing) is the delivery of computing services - servers, storage, databases, networks, software, analysis - via the Internet ("the cloud"). Just as cloud platforms are growing rapidly, we also see an explosion in data generation. Humanity has never generated as much data as it does today, and the growth in the volume of data is exponential and constant. Big Data presents a great challenge for companies. How to collect, store, and analyze this data before it becomes obsolete? What is the cost of storing a volume of data that keeps growing? What are the security implications of accessing this data, once stored on the company's internal network and susceptible to attacks, information theft, and viruses? How to manage and protect the essence of this data instead of just storing it? Cloud Computing and Big Data are an ideal combination to solve many of these problems [4]. Together, they provide a solution that is scalable and adaptable for large data sets and business analysis. The advantage of the analysis would be a great benefit [5].

# **3. CHALLENGES FOR BIG DATA IN CLOUD COMPUTING**

Just as Big Data provided organizations with terabytes of data, it also posed an issue of managing that data in a traditional framework. How to analyze a large amount of data to remove only the most useful bits? Analyzing these large volumes of data often becomes a very difficult task. In the era of highspeed connectivity, moving large data sets, and providing the details needed to access them is also an issue [6]. These large sets of data often carry sensitive information, such as credit/debit card numbers, addresses, and other details, increasing data security concerns. Cloud security issues are a major concern for companies and providers. But research shows that a cloud environment tends to be more secure than a corporate network. Therefore, the main challenges in the adoption of Cloud Computing, especially for Big Data projects, include:

- \* Company culture that prefers to keep data "at home."
- \* Lack of knowledge or trained professionals to work with



### International Journal of Business and Management Research (IJBMR)

Review Article | Volume 9, Issue 2 | Pages 156-161 | e-ISSN: 2347-4696

cloud projects

\* Resistance of the IT area in "losing importance" within the organization (which is a mistake, as computing in the cloud frees IT to focus on other activities more critical to the business)

\* Problems with bandwidth for data transfer between the company's network and the Cloud Provider

In summary, Cloud Computing is part of an effective Big Data strategy, and the search for professionals who know how to work with Cloud Computing efficiently has been increasing every year [7]. Pointing out as one of the most valuable certifications today, the AWS Certified Solutions Architect - Associate certification is the gateway for anyone who wants to learn to work with the cloud environment professionally and demonstrate their skills [8].

And if it were possible to take a 100% online and 100% Portuguese course, and learn how to set up a storage and processing environment for Big Data and Machine Learning in the cloud, with the most used Cloud environment on the planet, Amazon AWS (Amazon Web Service)? What if one could learn to publish your analytical application in the cloud? And if at the same time, one learns about AWS, you could prepare for one of the most valuable certification exams on the market, AWS Certified Solutions Architect certification [9].

With increasing connectivity speed and Web systems, the Internet systems evolute to cloud computing. It refers to a support platform that provides: management, on-demand use, fitness requirements, rational use of resources, and automation in creating the required infrastructure. In this context, systems storing data in the cloud, database scalability, data search and retrieval emerged as a new discipline, now called BIGDATA [10].

Data analysis and curation should be part of the day-to-day business.

First, companies are finding it increasingly difficult to identify the right data and determine the best way to use it for different purposes - from interpreting consumer behavior to making decisions or predictions. The construction of data-related business cases often leads to thinking outside the box and seeking revenue models that are very different from the traditional way of managing business and technology. Also, the ability to choose what will be analyzed, as well as knowing the questions that must be answered are at stake [11]. It is necessary to know what people want to achieve with the analysis and curation of the data. Generally speaking, organizations want to improve their analysis capabilities in a way that is coherent with their business strategies. The good news is that companies that get more tools and knowledge to structure, sort, and interpret data, transforming it into valuable information, will be the ones that will be more competitive and able to gain market share [12].

## 4. BIG DATA REQUIRES

The truth is that few companies have enough technology to handle Big Data. Besides, changing the mindset of both IT teams and business executives is also a major challenge. When it comes to IT infrastructure (mainly tools), you need to monitor data centers and have high-performance databases to store and foresee the data explosion [13]. The protection of this data also needs to be considered, with firewalls, access controls, and policies appropriate to this new reality.

In this sense, using solutions and services based on cloud computing (Cloud Computing) can be simpler and cheaper than acquiring tools and managing them in a traditional way - this is because, without using the virtualization strategy offered by the cloud, updating the database and its proper management could cost much more and take much more time [14].

### 5. BIG DATA GENERATES INSIGHTS BUT ALSO REQUIRES TIMELY ACTIONS AND CHANGES

Still talking about business behavior with Big Data, companies need to be quick to effect changes based on the insights generated with data analysis. That is, if it is possible to capture and mine huge amounts of data in a few seconds, generating relevant information for the business, it is also necessary to act at the same speed [15]. The challenge, therefore, is to adapt the speed of decisions to the power of information generated by structuring data analysis.

Dealing with Big Data requires collaboration and partnerships.

The IT team working side by side with business executives is not just a result of the adoption of Big Data, but the strategy is increasingly strengthening these ties. On the one hand, you have those with technical knowledge, on the other, those who need answers to make the right decisions [4]. Collaboration is also necessary when we look for partnerships formed with suppliers of tools and services that help in dealing with unstructured data, in addition to seeking consultancy and even collaboration of scientists. In other words, it is not possible to fully exploit Big Data and obtain results without collaboration.

### 6. RAPIDLY EVOLVING CLOUD COMPUTING TECHNOLOGY DISCOURAGES USERS AND INDUSTRY ALIKE FROM KEEPING PACE WITH THE CHANGES

For instance, consider the traditional method of data architecture that has long been used in most of the organizations. In this architecture, the existing data sources use integration techniques like the ELT/ETL and data capture. According to Teutsch [16] the daily transactions are transferred to the DBMS or the data warehouse [1].



## International Journal of Business and Management Research (IJBMR)

Review Article | Volume 9, Issue 2 | Pages 156-161 | e-ISSN: 2347-4696

Alternatively, the data are also transferred to the operational data storage or the data warehouse. Let us consider an example taken from the retail industry, in which various analytics capabilities are used for analyzing the data. The methods used are dashboards, reporting, or the EPM application.

The data are first stored as the master data. Then, from the master data, the data is transferred to the enterprise integration system [2]. After what the data is transferred to the data warehouse. Finally, in the data warehouse, the analysis of the data takes place. In other words, the traditional method of Big Data mainly uses the data management system, ensures the security and the good governance of the data.

# 7. DATA PROTECTION, PRIVACY, AND CYBERSECURITY

Two basic principles of data protection - preventing the accumulation of large amounts of data and minimizing data - are in stark contrast to the ability of big data to help track people's movements, behavior and preferences, predicting the behavior of an individual with unprecedented accuracy, often without the consent of that person [3]. For example, electronic medical records and self-recording of medical indicators in real time (people wear sensors on their bodies to control, for example, their physical fitness or sleep patterns). Good governance is essential, especially in the health sector. Using Big Data techniques is a huge step forward in streamlining the system for issuing prescriptions for medicines, diet or fitness exercises. Nevertheless, many consumers believe that such data is very sensitive [6].

Also, large data sets about calls in mobile networks, are anonymous and contain personal information. However, in combination with other data, such as geographically linked tweets, or records of visits to various places, the calls are able create "finger prints" revealing a person's identity [7]. As the volume of personal data and global digital information grows, so does the number of entities that access and use this information. Guarantees are required for the proper use of personal data, strictly restricted to the announced finality of the query and subject to strict compliance with the relevant legislation.

# 8. WHY NOT PROCESS BIG DATA IN-HOUSE

A solution for processing big data can be created in one's own data center, on physical servers, but the traditional analysis tools are gradually analyzing larger sets of data that often saturates their capabilities. These traditional tools are unable to hold and manage huge amounts of data, which makes it hard to acquire, manage, store, and process large data sets within a reasonable timeframe [8]. It happens also that traditional methods provide different results, as the size of the database increases [17].

The size of Big Data is growing on a daily basis. A while ago, the size of Big Data was measured in terabytes (TB) to those

days when several dozens of petabytes (PB) was used for a single set of data. By 2011, the cloud computing systems had solely reached about 150 Exabytes. It is proposed that by 2025, the cloud computing architecture, that is expanding every day, will have reached the zettabyte and the yottabyte scales.

The components of current cloud computing architecture involve perception which is the first capture or data generation, recording that is the physical capture of data, processing which is data transformation according to the specific needs of an organization, data storage, data retrieval and presentation which is how the information and data are reported and communicated [17].

The five basic components of a cloud computing architecture include personnel, consisting of information technology experts, like database administrators and network engineers, as well as the end-users. Another component is hardware that comprises all of the physical cloud computing hardware, such as computer parts and servers [17].



Figure 1: Cloud computing architecture Implementation Phasing.

Software is another component consisting of application software. The network component consists of networking media and related support; while the component of data comprises all the existing database knowledge, and the cloud computing architecture itself [10].

Information as a resource is important to how an organization runs its management and operations; the information should be received in a timely manner and should be readily available [17]. Industry research has shown that it is important because organizations' operations are centered on the effective performance of the functions of managers, such as organization, plan implementation, and control, and above all agility (fast and correct reaction to detected signals in the Data).



## International Journal of Business and Management Research (IJBMR)

Review Article | Volume 9, Issue 2 | Pages 156-161 | e-ISSN: 2347-4696

The sources of data in a cloud computing architecture are mainly internal. As an example, for the health sector the data cover, health records of patients in electronic form, decision supports of clinical systems, and other health records in control of doctors and caregivers [17]. External sources generate less data amounts, but they contain some of the most important aspects needed in healthcare. For instance, laboratories reports can give very detailed information on the health status of a patient. Other external sources include insurance companies and pharmacies [11]. The data often are contained in different locations, including healthcare databases and applications that process transactions. The data types utilized by the system consist of tables and also other kind of documents written in different formats; they may come from social media platforms like Twitter and Facebook. Health plan websites are other rich sources of web information [12]. The majority of the data concern biometrics in the form of fingerprints, retina scans, medical images, blood pressure measurements, pulse rate, heart beat rate, and x-ray images. Unstructured data for Big Data analytics come from notes made by physicians, paper documentation, scientific texts, doctor-patient emails, and EMRs.

According to Marx [15], the sequence of handling the data goes through different steps. First, the data from all these sources are collected and waiting for the next stage called processing. There are several options available for transforming the data. For instance, it is possible to process Big Data using an architectural approach that is serviceoriented [4]. It is a combination of web services in which the data remains in its raw form, while the processing, retrieval, and calling of data is done through services. The second approach uses the concept of data warehousing. Here, researchers combine data from different sources and transform them through the analysis [1]. The data is unavailable in real time and follows the steps of extraction as well as the transformation of large amounts of data coming from various sources under different formats. Data is classified as structured as well as unstructured data for input.

The architecture of Big Data can be mainly analyzed with the help of the traditional framework. In most cases, there are various methods that can be used so that the data that are stored are in the form of unstructured data. However, they also receive data that are structured. Thus, in that master data warehouse of the storage, the data that are stored are in the form of both the structured and the unstructured data.

According to Hashem et al. [18], the question of data formats remains not very clear: On the one hand, we have a huge number of different types of stored data, on the other hand, even the same type of data can be stored in incompatible formats. Here we already come to the formulation of the problem of a certain free data format, when at the time of data collection, it is still not known in which applications and for what purposes they will subsequently be used. But at the same time, low efficiency or strong redundancy of stored data cannot be allowed. Until the Russian pioneers in the development of new technologies voice the economic effects of their use, all the rest will wait, take a closer look. Speaking about this, Marx [17] draws a parallel with RFID: Everyone understands the advantages, there are a lot of pilot zones, but the technology has not been widely used, because, at the current price of equipment, its scope is limited [2]. But all the polls in the West, show that many companies consider unstructured data and the opportunities provided by working with them to be very important for a business. Since approaches to working with big data (MapReduce, Hadoop) are developing rapidly, up to the time these projects are ready for the market, we will use already proven, mature technologies [3].

Hashem, et al. [18] noted one more point: working with big data is impossible without the corresponding sharpening of both the program part and the hardware part ("one hand does not clap"). Therefore, the companies that deal at the same time with hardware and software quickly picked up the idea of providing a single product, although there are both disadvantages and advantages for the customer in this approach.

Marx [17] said that the most important factor for the success of big data is the creation of a flexible infrastructure that will provide the right combination of various aspects affecting this technology. First of all, it should be based on the business goals, and business requirements of the organization. It is important to take into account such important components as access to all data sources in near real-time and even real-time mode, support for various types of devices, data management, integrated analytics, etc. [6].

Trovati, et al. [19] pointed out the presence of five factors that determine corporate analytics: 1) the growth of data volume, the need for large amounts of memory; 2) an increase in the number of users; 3) lack of boundaries for unstructured data data diversity becomes the standard; 4) the speed of data entering the system; 5) data quality. He further explained: "The expansion of the traditional infrastructure in the areas of processing data in random access memory and storing data by columns will allow companies to analyze both structured and unstructured data in a single consolidated environment, as well as process them in real time and respond with the least delay to events" [4]. This is especially true for environments with arbitrary queries and various user-profiles and queries. Besides, investing in mobile analytics will enable the business community to maximize the value of data and increase the productivity of their employees [7]. The introduction of new software models, such as the MapReduce framework, as well as the support of the analytic server for paradigms that provide massively parallel and distributed computing, such as Hadoop, will create a more manageable, integrated and accessible analytical environment.

### 9. CONCLUSION

All in all, the value of systems based on Big Data technologies lies in the quality and speed of the provided analytics. Noting



this, Trovati [19] noted: "Since such analytical tasks that go beyond the scope of data warehouses were previously solved only for individual unique projects, for launching projects on a new set of technologies, it is necessary to develop new industrial models, indicators for specific types of customers. In addition to the industrial focus, a creative approach to data mining is also required - the possibilities for Big Data research are huge".

Buyya [20] are sure that the main problem is the shortage of qualified IT personnel, of which a shortage exists presently both in the Russian market and in the world. In companies facing the need to process large amounts of data, it is important to ensure the possibility of transparent scaling without interrupting their work. -In such an environment, it is necessary to ensure the ability to pay for resources as they grow with single and simple management of all elements of the infrastructure. Therefore, we believe that flexible work with big data is impossible without cloud storage and cloud computing, and in the form of complete solutions. However, cloud technologies alone do not work well with large blocks of information. The requirements for scaling storage systems, analytical applications, and related hardware and software systems are becoming higher. The challenge is to ensure maximum performance without dramatically expanding servers or disk arrays capacities. This is now possible with the latest technological advances offered by leading companies.

Bi and David [21] drew attention to the fact that the main task of the client when solving Big Data problems is to select the technology that suits him: If the organization is not ready to wait for its software to work with Hadoop or does not want to work with open-source software, then most likely the said organization will select an option integrating software and hardware. If we are talking about solutions based on Hadoop, then at this stage, there are three main problems. The first is that these systems are not a self-contained product, like a new server or array. Thus, the option consisting in "installing a new, more powerful server and solving this way the problem" does not work. The second is the correct positioning of the systems. These systems, of course, cannot be promoted as a universal replacement for large databases, but they are competitive in their specific field of application (analysis and processing of large volumes of heterogeneous data).

#### REFERENCES

- [1] Mehta, N., & Pandit, A. (2018) Concurrence of big data analytics and healthcare: A systematic review. International journal of medical informatics, 114: 57-65.
- [2] Darwish, T. S., & Bakar, K. A. (2018) Fog based intelligent transportation big data analytics in the internet of vehicles environment: motivations, architecture, challenges, and critical issues. IEEE Access, 6: 15679-15701.
- [3] Yassine, A., Singh, S., Hossain, M. S., & Muhammad, G. (2019) IoT big data analytics for smart homes with fog and cloud computing. Future Generation Computer Systems, 91: 563-573.
- [4] Kobusińska, A., Leung, C., Hsu, C. H., Raghavendra, S., & Chang, V. (2018) Emerging trends, issues and challenges in Internet of Things, Big Data and cloud computing, 416-419.

- [5] Jin, Xiaolong, Benjamin W. Wah, Xueqi Cheng, and Yuanzhuo Wang. (2015) Significance and challenges of big data research. Big Data Research 2(2): 59-64.
- [6] Elshawi, R., Sakr, S., Talia, D., & Trunfio, P. (2018) Big data systems meet machine learning challenges: towards big data science as a service. Big data research, 14: 1-11.
- [7] Arunachalam, D., Kumar, N., & Kawalek, J. P. (2018) Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. Transportation Research Part E: Logistics and Transportation Review, 114: 416-436.
- [8] Choi, T. M., Wallace, S. W., & Wang, Y. (2018) Big data analytics in operations management. Production and Operations Management, 27(10): 1868-1883.
- [9] Saggi, M. K., & Jain, S. (2018). A survey towards an integration of big data analytics to big insights for value-creation. Information Processing & Management, 54(5): 758-790.
- [10] Govindan, K., Cheng, T. E., Mishra, N., & Shukla, N. (2018) Big data analytics and application for logistics and supply chain management.
- [11] Manogaran, G., Thota, C., & Lopez, D. (2018) Human-computer interaction with big data analytics. In: HCI challenges and privacy preservation in big data security (pp. 1-22). IGI Global.
- [12] Nguyen, T., Li, Z. H. O. U., Spiegler, V., Ieromonachou, P., & Lin, Y. (2018). Big data analytics in supply chain management: A state-of-the-art literature review. Computers & Operations Research, 98: 254-264.
- [13] Grover, V., Chiang, R. H., Liang, T. P., & Zhang, D. (2018) Creating strategic business value from big data analytics: A research framework. Journal of Management Information Systems, 35(2): 388-423.
- [14] Wang, L., Ma, Y., Yan, J., Chang, V., & Zomaya, A. Y. (2018) pipsCloud: High performance cloud computing for remote sensing big data management and processing. Future Generation Computer Systems, 78: 353-368.
- [15] Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A. A., et al. (2018) The 10 Vs, issues and challenges of big data. ACM International Conference Proceeding Series.
- [16] Teutsch SM, Churchill RE (2000) Principles and practice of public health surveillance (2nd edn). New York, USA: Oxford University Press.
- [17] Marx, Vivien. (2013) Biology: The big challenges of big data. p. 255.
- [18] Hashem, Ibrahim Abaker Targio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, et al. (2015) The rise of "big data" on cloud computing: Review and open research issues. Information systems 47: 98-115.
- [19] Trovati, Marcello, Richard Hill, S. Y. Zhu and L. Liu. (2015) Big-data analytics and cloud computing. Springer Berlin Heidelberg.
- [20] Buyya, Rajkumar, Kotagiri Ramamohanarao, Chris Leckie, Rodrigo N. Calheiros, Amir Vahid Dastjerdi, et al. (2015) Big data analytics-enhanced cloud computing: Challenges, architectural elements, and future directions. 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS), pp. 75-84.
- [21] Bi, Zhuming, and David Cochran. (2014) Big data analytics with applications. Journal of Management Analytics 1(4): 249-265.



### International Journal of Business and Management Research (IJBMR) Review Article | Volume 9, Issue 2 | Pages 156-161 | e-ISSN: 2347-4696

CC I

© 2021 by the P. Madana Mohan, B. Murali Manohar. Submitted for possible open access publication under the terms and conditions of

the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).