BIG DATA ANALYTICS ON CLOUD ENVIRONMENT:ADVANTAGES, LIMITATIONS AND ISSUES

[1]TRISTAN MERCIECA, [2]LALIT GARG, [3]VIJAY PRAKASH, [4]SEEMA BAWA

[1,3,4]Faculty of Information and Communication Technology, University of Malta, Msida, Malta
[2,4]Thapar Institute of Engineering & Technology, Patiala, Punjab, India
E-mail: [1]Tristan.mercieca.16@um.edu.mt, [2]lalit.garg@um.edu.mt, [3]vijay.prakash@thapar.edu, [4]seema@thapar.edu

Abstract - In Today's world, Big Data and Cloud Computing aretwo mainstream technologies that are the center of concern in IT. Fundamentally different, Big data is all about dealing with the massive scale of data whereas Cloud computing is about infrastructure. However, the simplification offered by Big data and Cloud technology is the main reason for their huge enterprise adoption. The combination of both yields beneficial outcome for the organizations. Not to mention, both the technologies are in the stage of evolution but their combination leverages scalable and cost-effective solution in big data analytics.In this paper we have mainly analysed the big data on the Cloud environment with an application to describe the challenges and limitations of this combination. Further, by analysing Big Data in the cloud can create scalable and even cost-effective solutions accessible by all types of companies.

Keywords - Cloud Computing, Big Data Analytics, and Big Data Issues.

I. INTRODUCTION

Big Data is a phenomenon that has gained much attention over recent years due to the much advancement in technology [1]. It refers to techniques used to capture, store, manage and analyse very large data-sets (Petabyte of data) with a high velocity [2]. Big Data is even seen as a way to match data, look for patterns in data and even predict new trends [3]. Big Data is centred on velocity (the speed of data frequency), volume (representing the large amount of data being produced), variety (data has changed from structured to unstructured), veracity (the quality of the produced data) and value (representing the value of Big Data) [4].

There exist two main types of Big Data such as organizational data and data that come from outside the organization [5]. Organizational data encapsulates emails, documents and all other forms of structured, semi-structured and unstructured data which is only available from within the company. On the other hand, extra-organizational data refers to data which is available publicly (and often at no cost, althoughsome of it is obtained through a form of paid subscription). This type of data includes customer complaints and hints from third parties, among others [5].

Cloud computing is another phenomenon which refers to computer resources and systems which are made available over networks [4], allowing users to access their data remotely [6]. Such resources include dynamic storage, backup and even self-synchronization [7]. Cloud Computing has even been said to have "revolutionized the way computing infrastructure isabstracted and used" [8]. Currently there exists a pressing problem with regards to Big Data since it is being created at an increasing rate. This implies that data analysis and management is becoming even more challenging for traditional data warehouses [8].Furthermore, the price of maintaining such data warehouses can be a financial problem to small-sized companies [9]. Therefore, by analysing Big Data in the cloud can create scalable and even cost-effective solutions accessible by all types of companies.

The main principle of cloud computing is to provide storage, computing power and even software (such as API's and development tools) in the form of a service[10]. Cloud computing has also been used by business to test an application before deploying and using it [1]. When talking about Cloud Computing, once can define three types, namely the public cloud, the private cloud and the hybrid cloud [1]. The public cloud is the general no-hassle, pay-as-you-go used mostly by companies who do not already have the necessary applications and infrastructure [11]. The private cloud refers to the internal data centre of a business (which is not available to the general public) which utilises a cloud structure [1]. On the other hand, the hybrid cloud uses some form of a private cloud while migrating or creating new stuff onto the public cloud [12].

The most prominent services offered by Cloud Computing include Infrastructure as a Service, Platform as a Service and Software as a Service [13]. Infrastructure as a Service (IaaS) provides the necessary infrastructure in the form of a service. Examples of this paradigm includeGoGrid and FlexiScale. Platform as a Service (PaaS) provides the necessary platform (encompassing the systems and environments) as a platform. Common examples of PaaS include Microsoft's Azure and GAE [15]. On the other hand, Software as a Service (SaaS) involves delivering applications as a service as well as the necessary hardware and software that together provide the service [16]. Common examples of this type of service include Oracle and IBM.

Rest of the paper is organized as Section II discusses the literature review. Section III presents analysis of big data in Cloud environment. Section IV describes the advantage and limitations of big data on Cloud. Section V presents various issues due to analytics of big data on Cloud Environment. Section VI presents conclusion and future scope.

II. LITERATURE REVIEW

Cloud computing and Big Data combination is becoming more popular because commercial DBMSs aren't sufficient enough to store this large amount of data, mainly due to their scalability and cost [23].Big Data and Cloud Computing separately, both offer several advantages, so it is only logical to assume that together, they provide even more benefits to those who opt to use them [8]. Cloud Computing allows for the remote storage of bulk data, with the added benefit of being accessed from anywhere, through this use of the Internet [28]. On the other hand, Big Data integrates both structured and unstructured data and can give a company competitive edge over another company.

There exist a lot of research on securely storing Big Data on the cloud, mining data from social media on the cloud and using combining the use of Hadoop and Map Reduce on the cloud. This section will summarise some research papers regarding the research mentioned above while also highlighting some companies who have chosen to utilise Big Data Analytics in a cloud environment.

Aujla et al. devised a scheme known as SecSVA [28] to allow safe and secure storage, verification and auditing of cloud-based Big Data. This scheme was created to reduce and overcome some of the challenges that come with storing and processing Big Data on the cloud, namely data security, authentication, integrity, confidentiality and availability. SecSVA utilises the ABE scheme (a cryptographic technology which is used to protect data in cloud computing) [29] and a Trusted Party Auditor to securely store Big Data in a cloud environment [28].With regards to social media, Noordhuis et al. made use of Amazon's cloud infrastructure to analyse huge amounts of data on Twitter [30]. This paper sought to use cloud computing as a means of reducing the costs and allowing for large scalable, especially since such applications tend to generate a large amount of data. This was done by using the Page Rank Algorithm to determine a non-generic ranking of the user base of this social media platform [31]. As a result, a cheap solution for acquiring data and analysing it on the Amazon cloud infrastructure was successfully implemented.In 2015, Ke et al. tried to reduce the traffic when delivering data from map tasks to reduce tasks, in Map Reduce [21]. To achieve this, they proposed an aggregation scheme for Map Reduce jobs in the cloud. However, instead of analysing data fully on the cloud, this idea seeks to reduce the traffic of Map Reduce jobs by moving some of the processes on the cloud, rather than all of them. Ke et al. presented a greedy algorithm to optimize the aggregator placement problem. Furthermore, the proposedaggregation architecture was implemented with the aim of easily incorporating it with the current Map Reduce Framework [21].

Hadoop is another technology which has gained a lot of popularity in recent years. However, despite its rise in popularity, few attempts have been made to show how it can be applied to real-world applications in different fields. This led Zhang et al. to investigate the use of Hadoop to implement an application that allows scientific data to be processed in a cloud environment [32]. The main focus of this application was processing of sequences of microscopic images of living cells through MATLAB [30]. The application allowed users to upload data processing jobs on the given cloud environment. Zhang etal. concluded that this application can be reused in different domains and scenarios.Apart from research into the evolving research area, many companies have already opted to put Big Data in the Cloud. Companies include IBM [33], Alteryx [34] and even Jaspersoft [35]. IBM has implemented a Hadoop-Based distributed platform for analysing and processing data, as a cloud service. The motivation for IBM to implement such a platform is because a few weeks before, both Oracle and Microsoft showed interest in Hadoop [33]. Both Microsoft and Oracle went on to introduce software based on the Hadoop technology.On the other hand, Alteryx have implemented a cloud-based platform to analyse Big Data in the form of a user-friendly business application, which George Matthew (the President and Chief Operating Officer at Alteryx) claimed that it allows business to get the derived value of Big Data much faster [34]. Jaspersoft too created an analytical tool on Amazon Redshift (a cloud-based data warehouse service) with the added benefit of either integrating it inside other applications, or running it as a standalone application [35].Therefore, one can see that despite analysing Big Data on the cloud is already a widely used technology, research is still done to improve it further, be it in the form of security, or to try and introduce more Big Data technologies on the cloud.

III. ANALYSING BIG DATA ON CLOUD ENVIRONMENT

Big Data can come about from various domains and devices and can be analysed to attain value for various domains too. Cloud computing offers a number of advantages too many domains, for hosting Big Data. One of those domains is education, or more

specifically in mobile learning. Mobile learning includes both formal learning inside a classroom and informal and formal learning outside the classroom, across a variety of mobile devices [36].As previously mentioned, Shorfuzzamam et al. used the analysis of Big data in the cloud to support learning analytics on a mobile learning platform. This was done since large amounts of data are being generated during learning sessions, which lead to the use of Big Data to effectively and efficiently process this generated data to further add value to learning sessions [25]. However, there is still a lack of processing power available on such mobile devices, which has led to the idea of transferring the large computational analysis to a cloud technology (namely Microsoft's Azure) [25]. This implementation incorporated Azure as Infrastructure as a Service (IaaS). Since some data can be unstructured, Hadoop was used to effectively handle such data. Once the data is analysed the results take the form of tables,dashboards and even graphs generated by tools such as Google Data Studio and Tableau sothat the users of such an implementation can understand the generated results easily [25].Through such an implementation, it is hopedthat educators can use the generated results to monitor the performance of their students, and change their teaching approach if a negative performance is generated. Therefore, such an implementation can help to generate useful feedback and find any risks as early as possible so that they can be dealt with immediately, and thus improving education through mobile learning [25]. Big Data in the Cloud has a lot of potential to apply to many different domains, not just for companies who wish to analyse their data to help make important decisions. One such application is in education where one can use cloud-based Big Data to support learning analytics in a mobile learning environment [25]. Shorfuzzaman et al. describe a mobile learning framework on the cloud which uses Big Data analytics techniques (mainly the Map Reduce technology) to extract valuable information from the mobile learner's data [25]. This example will be further explained in a later section.

Another application of Big Data Analytics in the cloud is in healthcare as it can speed up analytics and thus getting the required data faster. However, it is worth noticing that when choosing a provider, it is important to choose a cloud which is designed specifically for healthcare data to make it as easy as possible when searching through data [26]. An example of Big Data in the cloud in healthcare is PHEMI Central in the Cloud [27].

IV. ADVANTAGES ANDLIMITATIONS

Analyzing Big Data in the Cloud offers multiple advantages, the most prominent being improved analysis. The problem with large amounts of data is how it can be processed efficiently. Thanks to the advancements made in cloud computing, analysing big data has improved while making the analysis process even easier, making it more accessible to companies of various sizes [1]. Furthermore, through the cloud, data can be integrated from various sources. Apart from being very expensive, setting up an infrastructure is also time-consuming, since installing the server and getting it to run can take a lot of time and resources [8]. Luckily, through cloud computing, an infrastructure is provided almost immediately.Furthermore, since cloud computing offers a flexible infrastructure, the cloud platform can be expanded dynamically to accommodate the ever-increasing amount of data and once the necessary insight has been attained, the storage space can be expanded or reduced based on the data[4]. Cloud-based BigData removes the need for data warehouses, since everything is stored on cloud servers. The cost responsibility is shifted to the cloud providers while the company requesting the resources only need to pay for the storage space and the power that they are consuming, while still having total control over their data.

As every other existing technology, Cloud-based analysis of Big Data does have some limitations, especially when it comes to security. Since there is a heavy use of third-party services used to store important data, there is a huge challenge to monitor and protect the stored data [23]. Furthermore, platform heterogeneity is becoming another major limitation, with regards to security since in some cases, Big Data deployments require the deployment of a new platform in the cloud which require new security tools [8].The fact that data is often stored in different locations occasionally poses a problem as it is much harder to work with data stored in different locations [24].Moreover; another pressing limitation is the fact that there is a very thin line between private clouds, public clouds and the Internet of Things [24].

A user using services offered by a cloud provider can use computing capabilities when needed automatically, without needing any form of human interaction [17]. These services should also be available and accessible to every person in every location and from every device (assuming that the device supports data exchange protocols). In cloud computing, a user only pays for the services he is using as well as being able to release (and therefore not pay) for any resources or services that he is not currently using [18]. Cloud computing also allows users to use an infinite amount of computing resources on demand [10]. Furthermore, different users can use the same services differently, and for different needs.Two main technologies which utilise Big Data are Hadoop and Map Reduce. Hadoop is a java-based, open platform, programming framework which is capable of processing large amounts of data

in a distributed computing environment [2]. By making use of Hadoop, large amounts of data can be stored reliably across a number of servers very fast [4]. Hadoop makes use of two main components, HDFS(Hadoop Distributed File System) and Map Reduce [19]. An application that is running on Hadoop is divided among various nodes (or machines) and HDFS is used to store the data that will be processed [20].

Map Reduce is another technology developed specifically to process and store Big Data effectively [21]. Map Reduce is used to write applications which analyse huge amounts of data reliably and without faults [2]. This technology works by submitting a data processing request (known as a job) by providing a map and a reduce function and upon executing the request, two types of tasks are created, map and reduce. The input data is divided into splits which are processed by the map tasks. These generated inputs are then shuffled and sorted by the framework and then used by the reduce tasks to produce the output [21-22].

V. ISSUES OF BIG DATA ANALYTICS ON CLOUD ENVIRONMENT

There are various issues highlighted by the researcher [3, 8, 28, and 38] related to the analytics of big data on Cloud environment such as:

- Security: When merging theCloud computing with big data, then security is one of the major problems due theupdating and uploading of data into the cloud.This issues such as who is the real owner of the data? Whether it is actual Cloud data or big data?Where the data is stored on cloud or it is actual big data? What kind of user can access this data and what kind of permissions access they have are hard to describe.
- Data Availability, Confidentiality, Integrity and De-duplication (ACID): It is very hectic to make the data available Round-the-clock. For this purpose, Cloud Service Providers (CSP) needs to make duplicate copies of data across different clouds which cause additional computational and storage cost. It is a big challenge to maintain data confidentiality which actually refers to protect data from unauthorized disclosure. The data integrity featurechecks that the data received from a source is exactly the same as sent by storage entity.Most solutions forsecure data storage on the cloud have notexploited the data deduplication factor, whichis of utmost importance.
- Privacy: The assurance of data security and keeping the data privacy is extremely difficult due the replication of data over different platforms. Traditionally, various organizations uses encryption data to distance datafrom real identities. But it can stillbe re-identified and attributed to specific individuals.
- Heterogeneity: Big data mainly consists of large volumes of data with different velocities and varieties. Variety is one of major characteristics of big data which stores different kinds of data. Data analyst mainly suffers to deal with such a wide variety of data and with different velocity. Though, providing a consistent andgeneral way to represent and explore complex andevolving relationships from this data still poses achallenge.
- Data Governance:Data Governance can handle the total cost of ownership (TCO) problem bycreating several policies that define for durability and viability of data. The concept consists of practices andorganizational policies that describe how data should be managed through its useful economic life cycle.
- Disaster Recovery: Losing data willcertainly result in losing value. In case of natural disasters like fire, floods and earthquakes, data losses need to be minimal. To overcome this problem, data must bequickly available with minimal downtime and loss.For big corporations it is imperative to define a disaster recovery plan such as part of the data governance plan.
- Scalability and Elasticity: Most of the time, scalable systems showsthat most algorithms are reactive or proactive andfrequently explore scalability from the perspective ofbetter performance. However, a proper scalablesystem would allow both manual and automaticreactive and proactive scalability. On the other hand, elasticity ensures an even loaddistribution to each server.

VI. CONCLUSIONS AND FUTURE SCOPE

Big Data Analytics in the cloud has created a paradigm shift in the way that one thinks of the term “data management”. The reason for this is now; the Cloud is serving as the data platform. The Senior Vice President of technology at Oracle stated that Big Data analytics will allow all sizes of companies to be able to use Big Data, and explains the competitive advantage between companies will emerge from the way that the data is analysed [37].

Future research into this topic may include searching for the “next-big-thing” as to the way we look at and process Big Data, focusing on whether the current cloud technologies will be enough for the ever-growing amount of data [38]. Furthermore, another possible future research could focus on whether all types of Big Data workloads and application frameworks are suited for the cloud, or if only some parts of Big Data should be stored, processed and analysed on the cloud [38].

REFERENCES

[1] Purcell, B. M. (2014). Big data using cloud computing. Journal of Technology Research, 5, 1.

[2] Yadav, S. L., &Sohal, A. Review Paper on Big Data Analytics in Cloud Computing.

[3] Neves, P. C., Schmerl, B. R., Cámara, J., & Bernardino, J. (2016, April). Big Data in Cloud Computing: Features and Issues. In IoTBD (pp. 307-314).

[4] Zanoon, N., Al-Haj, A., &Khwaldeh, S. M. (2017). Cloud computing and big data is there a relation between the two: a study. Int. J. Appl. Eng. Res, 12(17), 6970-6982.

[5] Zhu, W. D., Gupta, M., Kumar, V., Perepa, S., Sathi, A., &Statchuk, C. (2014). Building Big Data and Analytics Solutions in the Cloud. IBM Redbooks.

[6] Wang, L., Von Laszewski, G., Younge, A., He, X., Kunze, M., Tao, J., & Fu, C. (2010). Cloud computing: a perspective study. New Generation Computing, 28(2), 137-146.

[7] Kshetri, N. (2010). Cloud computing in developing economies. Computer, 43(10), 47-55.

[8] El-Seoud, S. A., El-Sofany, H. F., Abdelfattah, M., & Mohamed, R. (2017). Big Data and Cloud Computing: Trends and Challenges. International Journal of Interactive Mobile Technologies, 11(2).

[9] White, C. (2015). Data communications and computer networks: A business user's approach. Cengage Learning.

[10] Buyya, R., Broberg, J., &Goscinski, A. M. (Eds.). (2010). Cloud computing: Principles and paradigms (Vol. 87). John Wiley & Sons.

[11] P. Central, "The great climb toward the public cloud," 2011.

[12] Xu, J., Xi, W., Guo, G., BJORNSON, E., Yu, J., & WEI, L. C. (2019). U.S. Patent Application No. 10/289,310.

[13] Reviewer-Gvero, I. (2014). Cloud computing concepts, technology and architecture by Thomas Erl, Zaigham Mahmood and Ricardo Puttini. ACM SIGSOFT Software Engineering Notes, 39(4), 37-38.

[14] Jadeja, Y., & Modi, K. (2012, March). Cloud computing-concepts, architecture and challenges. In 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET) (pp. 877-880). IEEE.

[15] Rimal, B. P., Choi, E., &Lumb, I. (2009, August). A taxonomy and survey of cloud computing systems. In 2009 Fifth International Joint Conference on INC, IMS and IDC (pp. 44-51). Ieee.

[16] JoSEP, A. D., KAtz, R., KonWinSKi, A., Gunho, L. E. E., PAttERSon, D., &RABKin, A. (2010). A view of cloud computing. Communications of the ACM, 53(4).

[17] Mell, P., & Grance, T. (2011). The NIST definition of cloud computing.

[18] Stipravietis, P., Žeiris, E., &Ziema, M. (2013). Electronic service architecture model assessment of conformity to cloud computing key features.

[19] Velinov, A., &Zdravev, Z. (2018). Analysis of Apache Logs Using Hadoop and Hive. Tem Journal, 7(3), 645-650.

[20] Parmar, R. R., Roy, S., Bhattacharyya, D., Bandyopadhyay, S. K., & Kim, T. H. (2017). Large-scale encryption in the Hadoop environment: Challenges and solutions. IEEE Access, 5, 7156-7163.

[21] Ke, H., Li, P., Guo, S., &Stojmenovic, I. (2015). Aggregation on the fly: Reducing traffic for big data in the cloud. IEEE Network, 29(5), 17-23.

[22] Yu, S., & Guo, S. (Eds.). (2016). Big data concepts, theories, and applications. Springer.

[23] Noraziah, A., Fakherldin, M. A. I., Adam, K., & Majid, M. A. (2017). Big Data Processing in Cloud Computing Environments. Advanced Science Letters, 23(11), 11092-11095.

[24] Chang, V., Ramachandran, M., Wills, G., Walters, R. J., Li, C. S., & Watters, P. (2016). Editorial for FGCS special issue: Big Data in the cloud.

[25] Shorfuzzaman, M., Hossain, M. S., Nazir, A., Muhammad, G., &Alamri, A. (2019). Harnessing the power of big data analytics in the cloud to support learning analytics in mobile learning environment. Computers in Human Behavior, 92, 578-588.

[26] Archenaa, J., & Anita, E. M. (2015). A survey of big data analytics in healthcare and government. Procedia Computer Science, 50, 408-413.

[27] Hassan, M. K., El Desouky, A. I., Elghamrawy, S. M., &Sarhan, A. M. (2019). Big Data Challenges and Opportunities in Healthcare Informatics and Smart Hospitals. In Security in Smart Cities: Models, Applications, and Challenges (pp. 3-26). Springer, Cham.

[28] Aujla, G. S., Chaudhary, R., Kumar, N., Das, A. K., & Rodrigues, J. J. (2018). SecSVA: secure storage, verification, and auditing of big data in the cloud environment. IEEE Communications Magazine, 56(1), 78-85.

[29] Liao, Y., He, Y., Li, F., Jiang, S., & Zhou, S. (2018). Analysis of an ABE scheme with verifiable outsourced decryption. Sensors, 18(1), 176.

[30] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information systems, 47, 98-115.

[31] Noordhuis, P., Heijkoop, M., &Lazovik, A. (2010, July). Mining twitter in the cloud: A case study. In 2010 IEEE 3rd International Conference on Cloud Computing (pp. 107-114). IEEE.

[32] Zhang, C., De Sterck, H., Aboulnaga, A., Djambazian, H., &Sladek, R. (2009, June). Case study of scientific data processing on a cloud using hadoop. In International Symposium on High Performance Computing Systems and Applications (pp. 400-415). Springer, Berlin, Heidelberg.

[33] Mosco, V. (2015). To the cloud: Big data in a turbulent world. Routledge.

[34] C., Irvine, Alteryx Consumerizes Big Data Analytics in the Cloud [online] Available at https://www.alteryx.com/press-releases/2012-10-08-alteryx-consumerizes-big-data-analytics-cloud (accessed on May 31th, 2019).

[35] Jaspersoft Delivers Big Data Analytics in the Cloud with Amazon Redshift [online]. https://www.jaspersoft.com/press/jaspersoft-delivers-big-data-analytics-in-the-cloud-with-amazon-redshift. (accessed on May 31th, 2019).

[36] Hockly, N. (2012). Mobile learning. Elt Journal, 67(1), 80-84.

[37] Seth, B., Dalal, S., & Kumar, R. (2019). Securing Bioinformatics Cloud for Big Data: Budding Buzzword or a Glance of the Future. In Recent Advances in Computational Intelligence (pp. 121-147). Springer, Cham.

[38] Nepal, S., & Pandey, S. (2015). Guest Editorial: Cloud Computing and Scientific Applications (CCSA)—Big Data Analysis in the Cloud. The Computer Journal, 59(3), 285-286.

★★★