Analytical and Perspective Approach of Big Data in Cloud Computing

Rekha Pal, Tanvi Anand and Sanjay Kumar Dubey

Abstract Cloud computing is a term which involves delivering services over the Internet at a cheaper cost. Big data is a massive collection of data sets having huge and complicated structures which are difficult to store, analyze and visualize. Cloud computing is the commonly used technology over which big data are managed and stored. The research in this domain has increased over past few years. In order to investigate the usage, issues and challenges by combining the big data in cloud computing, a systematic literature review is conducted. The review included various publications from 2004 to 2014 as a primary study. With the use of search techniques considered, 96 research papers were recognized out of which 23 were identified as relevant papers. The paper presents the various research progresses related to big data in cloud computing. It will also help the researchers to figure out the current and future scenario of research in big data using cloud computing technology.

Keywords Big data · Cloud computing · Hadoop · Data analytics · Distributed data base

1 Introduction

Cloud computing is emerged to be a new paradigm that relies on sharing of computer resources instead of locally handling by personal devices. It is a term which involves delivering services over the internet at a cheaper cost. Cloud

R. Pal · T. Anand (✉) · S.K. Dubey
Computer Science and Engineering Department, Amity University,
Sector-125, Expressway, Noida, Uttar Pradesh, India
e-mail: tanvi29anand@gmail.com

R. Pal
e-mail: palrekha106@gmail.com

S.K. Dubey
e-mail: skdubey1@amity.edu

V.B. Aggarwal et al. (eds.), *Big Data Analytics*, Advances in Intelligent Systems and Computing 654, https://doi.org/10.1007/978-981-10-6620-7_23

computing is not about relying on the hard drive for accessing all sort of data but it provides everything that is not physically close to you but can be easily available on the local network. With the development in today's business environment cloud computing serves us many benefits some are like scalability, collaboration efficiency, access to automatic updates, cost efficiency, easy to use and much more. Today suppliers, including AWS, working hard so as to boost public cloud adoption, while IT enterprise is working over hybrid and private cloud architectures so as to manage and control.

Big data is a term that outlines about the bulk amount of organized, semi-organized and unorganized data which mined from the web based application. Big data analytics is linked with cloud computing so as to the analyze the large data sets in real-time which is possible on a platform like Hadoop for storing large volume and variety of data among different distributed cluster and Map Reduce organize, collaborate and compute data from multiple sources. The reason for combine big data with cloud computing is to get the benefit from both the technology, which provides an advantage to an organization to think about how to operate necessary analysis that responds to their actual business requisite rather than still involve in finding ways to accumulate large data. While combining these two technologies user will get benefit with usability, cost-saving, accessibility and disaster management etc. The importance of this review paper is to focus on various issues and challenges that came across while merging these two technologies.

2 Literature Review

This section includes the research done related to big data in cloud computing and what new technologies and issues and challenges that are faced by the researcher during their work within 2004–14. A technique of micro array is used for measuring and reading the mRNA level which accesses a large number of DNA sequences which is done by using cluster and classification [1]. Also by focusing on various techniques of data mining and its future trends, we can have a simplified representation of useful information which will be very effective for coming future which will improve the usability [2]. Various data mining algorithms like k-mean algorithm can be used, which is a Gaussian mixture model for clustering and divide the customer or given set of data into categories which help in the formulation of market strategic planning and guidelines for future [3]. With the help of data mining algorithms, the stored data can be secured without transferring the complete data by using metadata, data segregation, and storage methodology that provide a way to access segregated data [4]. Gene based clustering which is very useful mechanism to extract useful information from noisy and redundant data, unsupervised gene selection can also apply for future selection [5, 6]. According to Satoshi Tsuchiya various technologies for big data processing in cloud computing environment by using key value store technology which distributes our data and offers high performance and high resistance to failure so by combining various functions together

that run in cloud will improve complex problems [7]. There are also various data mining privacy and threats on cloud, so to avoid them distributed architecture came which prevent data mining based attacks on cloud [8]. Big data is the term for data sets so large and complicated that it becomes difficult to process using traditional data management tools or processing applications [9]. Big data itself is a large data and varying as well as complex which leads to difficulty while analyzing and storing along with visualizing data for using it to form result [10]. Big data analytics which means analyzing of a large amount of data to select required information, unfold the hidden patterns [11, 12]. Big data analyses service generated big data, the relationship between service logs and also studied big data as a service and discusses its business aspects. We can use tape based service model and combine a large scale tape infrastructure with the cloud [13, 14]. There are various big data applications along with its usefulness in the developing world and its increasing work in the various sector of the organization, there is an increasing need of Hadoop for storing data, but various challenges faced are also there while adapting big data technology in the Hadoop tool [15]. By using integrity verified mechanism will continue to progress then it can meet security challenges faced in the cloud. ICT technology provides improved services of data being received from the cloud [16, 17]. A cloud computing framework can be used which helps in scheduling various distributed application of data mining so as to reduce overall execution time and improve the quality of data provided. With this complexity of cloud is hidden from cloud end users [18–20]. So with the enhancement in cloud computing, the internet or web technologies the data is also increasing and their efficient utilization of that vast data is a big concern [21]. RDBMS is not much capable of handling a large amount of data that's why HDFS comes in the picture because it is fast, secure, consistent and scalable to manage a large amount of unstructured data [22]. Cloud computing is becoming an increasingly popular enterprise model in which computing resources are made available on-demand to the user as needed. The unique value proposition of cloud computing creates new opportunities to align IT and business goals [23].

3 Review Methodology

Present review methodology is related to the literature review that targets various research questions to identify, evaluate, select and incorporate all the high-quality research evidence necessary for the research. Its aim is to present a fair evaluation of research topic by using a trust worthy, rigorous and auditable methodology. Among the available methodologies, paper chooses the journals and research paper methodologies. In the planning the review stage, the need for the review is identified, the research questions are specified, and the review protocol is defined. Finally, in the reporting the review stage, the dissemination mechanisms are specified, and the review report is presented. Papers before 2004 are not considered due to not much work was there in this area.

3.1 Research Questions

The goal of our review is to find the issues and challenges faced while storing a large amount of data in the cloud. During the study following research questions are framed:

RQ1. As per the future aspect how big data is dependent on cloud?

RQ2. How often is the Hadoop technology implemented as a tool?

These review questions allow identifying research gaps in a related area.

3.2 Identification of Primary Studies and Data Inclusion and Analysis Criteria

The main authors were chosen for studies are from IEEE Explore, Springer digital libraries and reputed conferences. We have limited our research with the papers that available online. After completing the search of papers from the year 2004–2014 in the first stage, we made notes (doc file) where we kept the records of all the important keywords that we extracted during the study. In analysis around 100 paper are fetched for study, from this 50 are evaluated and 24 are basically related to analysis.

4 Analysis

4.1 Significance of Review

Present review paper discusses various research progress, issues and challenges related to big data in cloud computing. Security and privacy are one of the main issues found in big data in cloud computing. Various technologies can be used for big data processing in cloud computing environment by using key value store technology which distributes data and offers high performance and high resistance to failure. So by combining various functions together that run in cloud will improve the complex problems

5 Overall Evaluation

The main goal of this paper is to analyze the importance of combine big data in cloud computing as well as research papers in this area. This section produces overall evaluation about the frame research question.

RQ1. As per the future aspect how big data is dependent on the cloud?

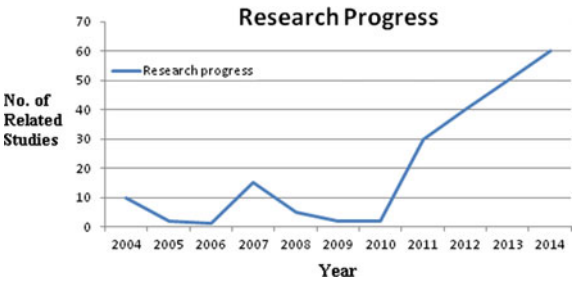

Fig. 1 Research progress (2004–14) for big data in cloud

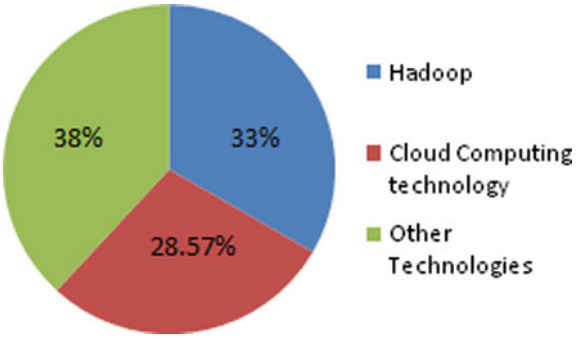

Fig. 2 Mathematical representation of HADOOP

After undergoing much research it is observed that for big data and cloud computing are more productive for building an application with faster and better understanding without worrying about underlying infrastructure. From the study, it is noticed that there is a quite variation in researches in this area as shown in Fig. 1. The graph indicates the growing research in the related area in past 10 years.

RQ2. How often the Hadoop technology is implemented as a tool?

While studying many researches during these 10 years it is observed that there is a quite high increase in the implementation and study of Hadoop technology. From the study conducted it is observed that 1/3rd of researchers used Hadoop as their tool.

From Fig. 2 it is observed 7 papers out of 21 papers reviewed uses specific Hadoop technology, 6 other Cloud Computing technologies and remaining other techniques

6 Conclusion

In this paper, different cloud based big data techniques, approaches are discussed and their usability in the organization is also taken into account. For this purpose, an analytical review is done and also frames research questions in this regard. Selection of primary study and data inclusion/ analysis criteria is also identified.

The present paper also tries to find the solution of framed research questions. By combining various functions together that run in a cloud environment by improving technologies, complex problems can be solved. It is still a big issue in providing security in cloud data. Even there is data, tools are available but there are other issues are also required to improve for efficient usage. Some improvement is required in HADOOP tool so that in future big data using technology will be highly skilled. The correlated technology of big data that are present today is not ideal to deal with challenges so require more examining and analysis.

References

1. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: A survey. IEEE Trans. Knowl. Data Eng. **16**(11), 1370–1386 (2004)
2. Kriegel, H.P., Borgwardt, K., Kroger, P., Pryakhin, A., Schubert, M., Zimek, A.: Future Trends in Data Mining. Springer Science+Business Media, LLC (2007)
3. Chen, X., Gao, L., Zhejiang, Wang, X., Zhang, Z., Wei, Z., Liao Z.: The research on the data mining technology in the active demand management. In: 2011 international conference on internet computing and information services (2011)
4. Subashini, S., Kavihta, V.: A meta data based storage model for securing data in cloud environment. In: International conference on cyber enabled distributed computing and knowledge discovery (2011)
5. Chandrasekhar, T., Thangavel, K., Elayaraja, E.: Gene expression data clustering using unsupervised methods. IEEE, pp. 146–150 (2011)
6. Agrawal, D., Das, S., Abbadi, A.E.: Big data and cloud computing: current state and future opportunities. ACM (2011)
7. Tsuchiya, S., Yoshinori, Tsuchimoto, Y., lee, V.: Big Data processing in cloud environments. Fujitsu Sci. Technol. **48**, 159–168
8. Dev, H., Sen, T., Basak, M., Ali, M.E.: An approach to protect the privacy of cloud from data mining based attacks, sc companion: High performance computing, networking storage and analysis (2012)
9. Li, B., Jain, R.: Survey of recent research progress and issues in big data, pp. 1–13. (2013). Available at: http://www.cse.wustl.edu/∼jain/cse570-13/index.html
10. Sagiroglu, S., Sinanc, D.: Big data; a review, pp. 42–47. IEEE (2013)
11. Shilpa, Kaur, M.: Big data and methodology. Intern. J. Adv. Res. Comput. Sci. Softw. Eng. **3** (10), 991–995 (2013)
12. Demchenko, Y., Grosso, P., Laat, L., Membrey, P.: Addressing big data issues in scientific data infrastructure. IEEE pp. 48–55 (2013)
13. Zheng, Z., Zhu, J., Lyu, M.R.: Service generated big data and big data as a service. In: IEEE conference on big data, pp. 403–410 (2013)
14. Prakash, V.S., Wen, Y., Weidong, S.: Tape cloud: Scalable and cost efficient big data infrastructure for cloud computing. In: IEEE sixth conference on cloud computing, pp. 541–548 (2013)
15. Katal, A., Wazid, M., Goudar, R.H.: Big data: Issues, challenges, tools and good practices. IEEE pp. 404–409 (2013)
16. Liu, C., Ranjan, R., Zhang, X., Yang, C., Georgakopoulos, D., Chen, J.: Public auditing for big data storage in cloud computing—A survey 2013. IEEE, pp. 463–468 (2013)
17. Lu, C.-W., Hsieh, C.-M., Chang, C.-H., Yang, C.-T.: An improvement to data services in cloud computing with content sensitive transaction analysis and adaption. In: IEEE 37th annual computer software and application conference workshops, pp. 463–468 (2013)

18. Ismail, L., Masud, M.M., Khan, L.: FSBD: A framework for scheduling of big data mining in cloud computing. In: IEEE international congress on big data , pp. 514–521 (2014)
19. Kadu, P.S., Deshmukh, H.R., Angaitkar, P.G., Karale, S.A.: A review on big data management and its security. Int. J. Pure Appl. Res. Eng. Technol. **2**(9), 1011–1017 (2014)
20. Wang, Y., Zhao, Y.: Transplantation of data mining algorithms to cloud computing platform when dealing big data. In: International conference on cyber-enabled distributed computing and knowledge discovery, pp. 175–178 (2014)
21. Mu, L., Lei, Z.: Big data processing technology research and application prospects. In: Fourth international conference on instrumentation and measurement, computer, communication and control, pp. 269–273 (2014)
22. Dwivedi, K., Dubey, S.K.: Analytical review on Hadoop distributed file system. IEEEXplore, In proceeding of 5th international conference—The next generation information technology summit, confluence-2014, pp. 174–181. Noida, India, 25–26 Sept 2014
23. Insfran, E., Fernandez, A.: A systematic review of usability evaluation in web development. In: Proceedings of 2nd international workshop on web usability and accessibility (IWWUA'08), New Zealand, LNCS, vol. 5176, pp. 81–91. Springer, Berlin (2008)