International Journal of Scientific Research in Computer Science and Engineering
Vol.7, Issue.4, pp.18-21, August (2019)
DOI: https://doi.org/10.26438/ijsrcse/v7i4.1821

Research Paper

E-ISSN: 2320-7639

DDDA: Development of a Distributed De-Duplication Approach using Big Data Analysis in Hybrid Cloud Environment

Aarfa Khan[1*], Snehlata Kothari[2]

[1,2] Dept. of Computer Science Dept. Pacific University, Udaipur, Indian

[*]*Corresponding Author: skothariudr@gmail.com*

Available online at: www.isroset.org

Received: 18/Aug/2019, Accepted: 25/Aug/2019, Online: 31/Aug/2019

Abstract— A renewed interest in cloud computing adoption has occurred in academic and industry settings because emerging technologies have strong links to cloud computing and Big Data technology. Big Data technology is driving cloud-computing adoption in large business organizations. For cloud computing adoption to increase, cloud computing must transition from low-level technology to high-level business solutions.

Security, privacy and elimination of repetitive copies of data is of primary concern for many applications of Big Data (BD). Data of the consumers must be protected else private information can be leaked. Cloud should let the owners or a trusted third party to check for the integrity of their data storage without demanding a local copy of the data. For this reason, this paper covered: Issues in big data management (BDM), secure data processing (DP) and access control (AC's) of data in cloud by data owner, data integrity verification in cloud. On performance basis, proposed approach is tested and simulated on different raw data and their processing compared with few of existing algorithm based on security, accessibility and integrity parameters. Results obtained are satisfactory to achieve all in single approach.

Keywords—DDDA, Big Data, Security, Cloud Computing, BDM etc.

I. INTRODUCTION

Cloud computing and Big Data technologies have the potential to disrupt industries and businesses through predictive analytics, machine learning, and artificial intelligence (AI) applications, bringing massive competitive advantages for businesses (Hashim, Hassan, & Hashim, 2015). Therefore, when scholar-practitioner Liu (2013) claimed that Big Data technology was the driver of cloud computing adoption in large business organizations, academic research was needed to accept or reject that claim. To examine Liu's recommendation that cloud computing adoption can be increased by increasing usefulness of the cloud computing through cloud-powered business enhancing service offerings such as Big Data analytics and business intelligence, a predictive model for cloud computing adoption, was created.

Security and privacy of big data is of primary concern for many applications. Data of the consumers must be protected else private information can be leaked. Similarly, due to the cost-efficiency, reduced overhead management and dynamic resource needs, content owners are outsourcing their data to the cloud who can act as a service provider on their behalf. Therefore, data integrity is of critical importance. Cloud should let the owners or a trusted third party to check for the integrity of their data storage without demanding a local copy of the data. Owners often replicate their data on the cloud servers across multiple data centers to provide a higher level of scalability, availability, and durability.

However, the data owners need to be strongly convinced that the cloud is storing data copies agreed on in the service level contract, and data-updates have been correctly executed on all the remotely stored copies. For this reason, this paper covered: Security and privacy issues in big data management, secure data processing and access control of big data in cloud, data integrity verification of big data in cloud, and security and privacy of sensing data for big data applications. Even though the novel computer hardware and network technologies can handle the demand of generating a big volume of data, effective file reduplication can save storage space in either the private computing environment or the public cloud system.

II. RELATED WORK

[1] Towards the Design of a System and a Workflow Model for Medical Big Data Processing in the Hybrid

© 2019, IJSRCSE All Rights Reserved

Cloud :2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology
Congress(DASC/PiCom/DataCom/CyberSciTech)
Globally, a big data analytics technology is being issued in various business areas including the medical field. Because the technologies demand a large number of resources, studies on a distributed cloud including a hybrid cloud technology are needed. However, there are limitations of performance and cost since the most of hospitals use the private cloud of low computing resource. Thus, a research on the hybrid cloud is required to resolve the limitations. In this paper, we propose and evaluate a system called BigPros for medical big data processing with the proposed workflow model in the hybrid cloud resolving above issues. Based on this study, the BigPros can provide an efficient data processing method for a medical big data analytics in the hybrid cloud environment.

[2] Geospatial Big Data Processing in Hybrid Cloud Environments:IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium
The importance of big geospatial data hosted on cloud environments is constantly growing. Main reasons are the rapid increase in volume of remote sensing data, the trend to persistently store and share more in-situ data at higher sampling rates, and the reduced management overhead of data hosted on commercial cloud platforms compared to in-house solutions. At the same time, cloud computing has the advantage of high scalability (and often reliability) and the capability to match the increasing computational requirements entailed by Big Data processing. This paper discusses interoperability and portability issues of cloud computing architectures and introduces a standards-based architecture to facilitate geospatial big data processing in hybrid cloud environments by leveraging and extending standards released by the Open Geospatial Consortium, OGC.

[3] Data security in cloud: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)
Cloud computing is moving increasingly to a destination with no return: the consolidation is an essential tool for the future existence of the internet world. Cloud computing is used as private, hybrid cloud, big data analytics and backup. Today the trend followed is to store the databases and software's in cloud environment. The services are taken whenever required. Without installing software's in systems users can go to cloud where already software is available and make use of it similarly for storing data also. Security is the main aspect now for the confidential data what we store in cloud.

[4] SMART: An Application Framework for Real Time Big Data Analysis on Heterogeneous Cloud Environments:2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing
The amount of data that human activities generate poses a challenge to current computer systems. Big data processing techniques are evolving to address this challenge, with analysis increasingly being performed using cloud-based systems. Emerging services, however, require additional enhancements in order to ensure their applicability to highly dynamic and heterogeneous environments and facilitate their use by Small & Medium-sized Enterprises (SMEs).

[5] Economical and efficient big data sharing with i-Cloud:2014 International Conference on Big Data and Smart Computing (BIGCOMP)
Big data can be hosted on cloud and being shared distributed through cloud services in an unprecedented volume, variety and velocity. These causes not only cloud network congestions and delayed cloud services but also increase in public cloud data-out charges. Client-side cloud cache alleviates these problems. Furthermore, cloud cache must be aware of no uniform data-out costs when big data is stored in hybrid clouds built with different public cloud providers. Deploying i-Cloud approach as the core mechanism of cloud cache could save data-out cost up to 14.78% or 4,425 USD saved per annum based on our representative scenario, and delivered 17.24% byte-hit, 17.96% delay-saving and 29.33% cache hit outperforming LRU, GDSF and LFU-DA approaches. A main finding is that i-Cloud, learning uniform cost patterns, could perform well against nonuniform cost environment.

III. METHODOLOGY

The researcher used a quantitative, nonexperimental, correlational research design to investigate whether perceived usefulness of cloud computing, perceived ease of use of cloud computing, security effectiveness of cloud computing, the cost-effectiveness of cloud computing, intention to use Big Data technology and the need for Big Data technology, could predict adoption of cloud computing. A quantitative approach was selected, which helps to facilitate an examination of the strength of association between quantified constructs (Howell, 2013). The variables of interest for the study were gathered in a numerical format using the validated survey instruments. Within this study, the researcher intended to assess the presence and magnitude of the relationship between the independent variables and the dependent variable

© 2019, IJSRCSE All Rights Reserved

Basically our system is consisting of clients (Data owners), a metadata server and main data servers (Cloud Service Providers), access and authentication policy. A distributed file system is built on those data servers and the metadata server. Clients interact to those servers via network connection to perform data de-duplication by using an access interface.

IV. RESEARCH PARAMETERS

Following are the research parameters on which the proposed study is going to be evaluated some of them are summed up as follows:

- Ease of Access
- System Complexity
- Internal Security
- External Security
- System Efficiency

V. RESEARCH TOOLS

There are many tools available to implement the proposed study, following are some illustrations:
The source code of widely used Hadoop distributed file system will be modifying to adapt new optimizing procedures into it. Scheme's performance will be test in a

- SHA-1
- AES
- Online Algorithmic Complexity Calculator: OACC.

An application security framework provides a systematic means to manage access to privileged operations in an application. The entity requesting access is typically a human user that is logged in to the application; the requesting entity does not, however, always need to be a human user: for example, it could also be another system interacting with the application. The requesting entity is called the subject in some security frameworks; in OACC, it is called the **accessor**. The object or asset that is accessed in a security relationship is called the **accessed** in OACC.

VI. CONCLUSION

Researchers in academia and practicing communities are standing at a crossroad of technology transformation. Some companies have had successes with their products and services in transforming industries with the help of disruptive, emerging, and innovative technologies. However, many companies with limited resources are struggling with this surge of the technology revolution. In This research attempted to close the knowledge gap between academia and practitioners by studying two of the most prominent technologies of the 21st century: cloud computing and Big Data technology.

As the technological world is moving at a rapid pace and AI, deep learning, machine learning, Big Data technology, and cloud computing are all inter-connected, it is crucial to have a solid understanding of how to adopt cloud computing. Cloud computing is directly linked to infrastructure and an organization's IT strategy.

REFERENCES

[1] Towards the Design of a System and a Workflow Model for Medical Big Data Processing in the Hybrid Cloud :2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)

[2] Geospatial Big Data Processing in Hybrid Cloud Environments: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium

[3] Data security in cloud: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)

[4] SMART: An Application Framework for Real Time Big Data Analysis on Heterogeneous Cloud Environments:2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing

[5] Economical and efficient big data sharing with i-Cloud:2014 International Conference on Big Data and Smart Computing (BIGCOMP)

[6] Rao, R. V., & Selvamani, K. (2015). Data security challenges and its solutions in cloud computing. *Procedia Computer Science*, *48*, 204–209.

[7] Saad, S. A., Adam, A., & Abdelateef, A. H. (2016). Binary logistic regression to estimate household income efficiency (South Darfur rural areas-Sudan). *International Journal of Advanced Statistics and Probability*, *4*(1), 31–35. doi: 10.14419/ijasp.v4i1.5657

[8] Schroedl, J. (2016). Hyperconverged infrastructure meets Big-Data-as-a-service.*Bluedata*. Retrieved from https://www.bluedata.com/blog/2016/05/hyperconverged-infrastructure-meets-big-data/

[9] Shahrivari, S. (2014). Beyond batch processing: towards real-time and streaming Big Data. *Computers*, *3*(4), 117-129.Siegel, E. (2013a). *Predictive analytics*. Hoboken, NJ: John Wiley & Sons.

[10] Siegel, E. (2013b, July/August). Predictive analytics: Harnessing the power of big data. Retrieved from http://analytics-magazine.org/predictive-analytics-2/Singh, J. (2017). Study on challenges, opportunities and predictions in cloud computing. *International Journal of Modern Education and Computer Science*, *9*(3), 17.

© 2019, IJSRCSE All Rights Reserved

Authors Profile

Ms. Aarfa Khan pursed Bachelor of Computer Science and Enggineering from R.G.P.V. University in 2011 and Master of Computer Science and Engineering from L.N.C.T. Group of College in year 2013. She is currently pursuing Ph.D. and currently working as Foreign faculty in Department of Computer Sciences, at Hainan Normal University, Haikou China since 2019. She has published 4 research papers in reputed international journals and conferences including IEEE and it's also available online. Her main research work focuses on Data Deduplication Algorithms, Networking, Cloud Security and Privacy, Big Data Analytics based education. She has 7 years of teaching experience and one year of Research Experience.

Dr.Snehlata Kothari. She is currently working as Assistant Professor in Department of Computer Science and Engineering, Vidhya Bhawan institute,Udaipur Rajasthan since 2016. She was member of IEEE in 2018. She has published Approx. 20 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. She has 18 years of teaching experience and 5 years of Research Experience.

© 2019, IJSRCSE All Rights Reserved