A Review on Unstructured Data Processing in Hybrid Cloud Platform

Dinesh Rajassekharan

Senior Lecturer, Faculty of Computer Science, Peninsula College, Malaysia

E-mail: dinesh@peninsulacollege.edu.my

Abstract

Cloud storage systems are widely employed in many applications due to their improvement in cost, storage availability and security. Hybrid cloud platform refers to the architecture of a cloud system that combines more than one computing environments at a time. It can be either with one public and one private platform or the combination of two private or two public platforms. The hybrid cloud platform has the ability to share the information among the connected systems and that can be processed parallelly while accessing the data. The data that are stored in cloud platforms are mostly in unstructured format that could not be used for any applications like prediction, recommendation, and estimations. This paper reviews the attainments of the previous works that were used for data distribution and partitioning in a hybrid cloud platform, by ensuring the privacy and security of the stored data. The work also explores the future directions on the unstructured data processing by summarizing the research issues observed from the review analysis.

Keywords: Cloud architecture, load analysis, resource allocation, data filtering, cloud modelling

1. Introduction

Cloud computing is a system that allows a user to store, process and access data through an internet service. In some cases, the cloud systems are also utilized for accessing certain process tools, analytic software, networking facility and databases. The cloud systems are recommended in many applications due to its improved accessing speed and data accessing possibility from any location. Similarly, the user has a freedom to not worry about the security of the data stored in the cloud. The security and privacy of the data stored in the cloud is up to the responsibility of the cloud service provider. Basically, the cloud architectures are represented in two categories as, private and public cloud, where the public

DOI: https://doi.org/10.36548/jscp.2022.4.001

Received: 07.09.2022, received in revised form: 03.10.2022, accepted: 18.10.2022, published: 01.11.2022

cloud allows the general user to store the huge information in the cloud that can't be stored in the local drive. The private cloud system allows an individual client to use the allocated storage on the cloud with a nominal fee [1]. A simple architecture of a cloud platform is shown in figure 1.

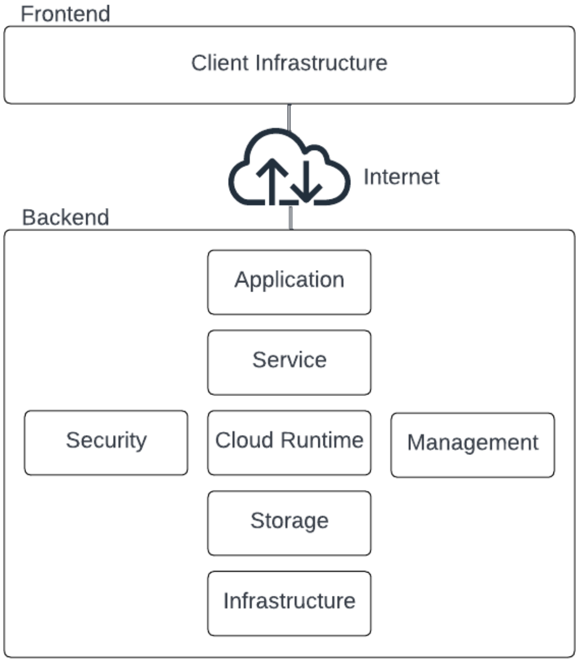

Figure 1. Architecture of cloud platform

The architecture of a cloud platform is divided into two types as frontend and backend. The frontend represents the client service model that contains infrastructure related to internet connectivity and data access. In some applications, the client infrastructure is employed with a graphical user interface that has the ability to project the status of the cloud service to the user. The backend represents the modules that are placed over the service provider location consists of storage units, control mechanisms and virtual models for machines and applications [2, 3]. Here the list of facility that are available in a general cloud architecture is explored.

- **Application:** It refers to the software and simulation frameworks that are kept for the clients in the cloud environment.
- **Service:** A cloud environment can provide three types of services to a client like software as a service, infrastructure as a service and platform as a service. The software as a service indicates the facility on software access from a remote location. The infrastructure as a service provides a storage facility to the user or client for saving their data. Platform as a service is one of the most complex applications

provided by a cloud module that contains a facility to create software modules with the inbuilt modules.

- **Cloud Runtime:** The cloud runtime provides the information related to data process and execution on a virtual machine platform. It helps a client to understand the processing speed and allows the user to take decisions accordingly.
- **Storage:** Storage is a scalable and flexible service provided to the client on data saving process based on their requirement.
- **Infrastructure:** It provides software and hardware facility to the client as servers and virtual machines for processing the stored data.
- **Security:** The backend security allows the client data to be more secure in the cloud platform from a third-party attack and intrusions.

2. Related Works

The hybrid cloud network can either be a public cloud that is combined with a private cloud or on-premises infrastructure. Figure 2 represents the architectural overview of a hybrid cloud model.

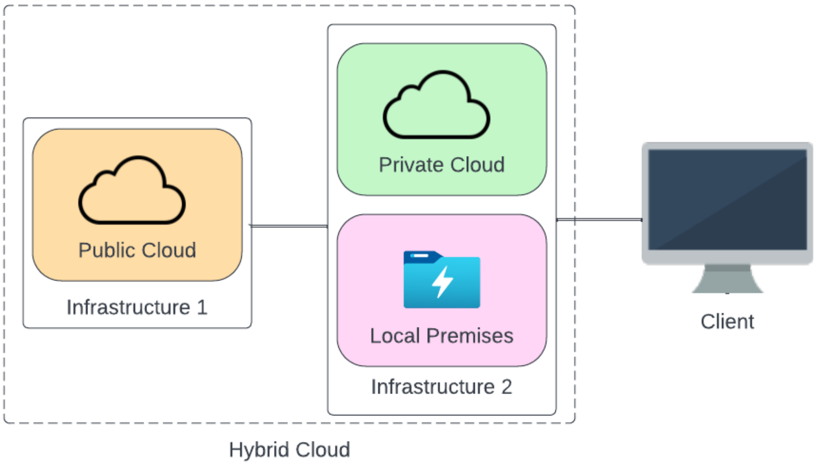

Figure 2. Architecture of hybrid cloud

The features of a hybrid cloud system are listed as follows [4, 5].

- **Data Integration:** Data integration allows a client to save the collected information either in infrastructure 1 or 2 based on the client requirement. The hybrid cloud processing system has the ability to access the data in both infrastructures at a same time for analysis.

- **Network Connections:** Network connection plays an important role in accessing the data between the infrastructure 1 and 2. It utilizes the internet facility for establishing a reliable communication between the connected networks.
- **Unified Management:** Application programming interface allows infrastructure 1 to access the information from the infrastructure 2 that is stored in different format. Similarly, service level agreement explores the facility that is given by the cloud vendor. The interfacing module and the service level agreement gives a unified management towards the hybrid network. Figure 3 explores an overview of a general workflow of an unstructured data processing approach.

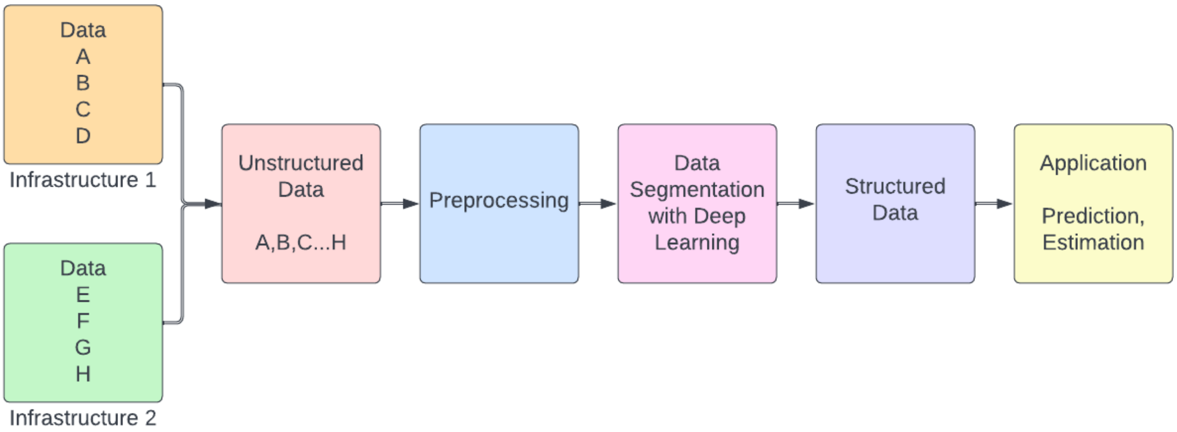

Figure 3. Unstructured data processing in hybrid cloud

The hybrid cloud networks are designed to collect different kind of data from various locations through sensors, keyboards, and automated decision-making systems. The data that are collected from the infrastructure can be either in image, data, digital value, and audio data. General hybrid architecture does not segregate those data to avoid unnecessary computational cost. The data processing applications like recommendation system, scenario prediction and value estimations require a huge amount data for taking a right decision [6]. The cloud storage systems are the one where a user store large quantity of data. However, collecting data from hybrid cloud network is always a challenging task as the data are in unstructured format. Therefore, the data processing algorithms are proposed in recent years to structure those data with preprocessing and segmentation. The preprocessing step regularizes the data that are collected with missing information and incorrect format. The data that are collected with such irregularities are removed from the database. Further a data segmentation algorithm is placed to segregate the data based upon their feature information. The structured information is forwarded towards the application that are developed for prediction and recommendation [7]. The following table 1 represents the outcome of certain recent studies that were used to process the unorganized data in the hybrid cloud platform.

Table 1. Literature on unorganized data processing in hybrid cloud

Reference	Methodology	Application	Attainments
Singh et al. [8]	Dynamic analysis with Halstead complexity	Student scholarship analysis	Reduced complexity through dynamic study
Daniel et al. [9]	Hadoop technique on Oracle	Organization risk management	Improved data analytic quality with security
Shehab et al. [10]	Feature selection processing with KNN	Unstructured complex data	Reduced 30% of processing time
Thenmozhi et al. [11]	Hybrid ML approach	Biological data	Improved data accessibility
Subramanian et al. [12]	Hybrid crypto system	Multi-cloud storage	Improved privacy and reduced insider attack
Azad et al. [13]	Classification on SQL, graph and non-SQL	IoT data	Betterment in data volume handling
Sun et al. [14]	IPv6 smart gateway	Ocean cloud data	Reduced the computational speed
Lang et al. [15]	Cloud native and middle platform	Building automation	Improved productivity
Franca et al. [16]	Differential Quadrature Phase Shift Keying	Big data	Local data accessing speed improved
Zhang et al. [17]	CNN with hybrid features	Cloud classification	84.37% of classification accuracy
Zhifeng et al. [18]	Distributed cloud structure	Oil and gas industry data	Better quality on data sharing
Shafqat et al. [19]	Parallel processing on data analytics	Electronic healthcare data	Improved sequential process
Sheik et al. [20]	Stored data analytics though classification	Cloud IoT data	Better storage space occupancy
Rashid et al. [21]	Density-based spatial clustering	Medical IoT data	Improved data retrieval speed
Tsung et al. [22]	Flow based update	Micro manufacturing	Optimal utilization in cache size

3. Discussion

The literature study indicates that most of the existing methods take the security and privacy concern into account on processing the unstructured data. The data processing algorithms have a capability to store their feature on its cache memory [22] and that are easily vulnerable to the hacker's algorithm. Therefore, a separate security algorithm is also fed into the workflow of a data processing system. Similarly certain algorithms are found to be implemented with a complexity maintenance algorithm that allows the work to structure the data with minimum process time [8, 14]. Only very few approaches are found in the literature that consider the process sequence into their account on data processing [13, 19]. The data processing is an extra module that is placed in the hybrid cloud system in between the memory and data receiving block. The following statements are the research gaps identified from the literature study and that need to be considered in future research.

- The data processing algorithms are found to be taking additional memory space from the cloud infrastructure and that may drag the performances of the regular process. Therefore, light weight architectures on data processing algorithms are in expectation.
- Minimum utilization on cache memory is required. It restricts the hackers to collect certain data related to the nature of the organized data.
- In some cases, the data retrieval speed drags to certain limit as the processing algorithm takes a huge space in the virtual Random Access Memory on the cloud.
- Analytical quality of structured hybrid systems is quite good, but the accessibility towards the stored data must also be taken into account for analysis.

4. Conclusion

Hybrid cloud platforms are mostly preferred by many clients due to its simplicity on installation and cost saving. The private cloud systems are safer than the hybrid systems, and they need the clients to install a security firewall policy for saving their information. Also, the client needs to provide a complete maintenance for ensuring the smooth operation of the private cloud. Considering such factors, the hybrid clouds are better, but they don't save the transferred data in an organized manner for reducing the complexity on data storage and retrieval. In recent years, certain algorithms have been developed to organize the data in the hybrid cloud environment by maintaining the data security and processing speed. This work has reviewed the achievements and limitations of the previous models, and suggests the future directions of hybrid cloud data processing to consider resource to be utilized in the cache memory and to design a lightweight algorithm for ensuring the processing speed.

References

[1] Sunyaev, Ali. "Cloud computing." In Internet computing, pp. 195-236. Springer, Cham, 2020.

[2] Tabrizchi, Hamed, and Marjan Kuchaki Rafsanjani. "A survey on security challenges in cloud computing: issues, threats, and solutions." The journal of supercomputing 76, no. 12 (2020): 9493-9532.

[3] Kumar, Mohit, Subhash Chander Sharma, Anubhav Goel, and Santar Pal Singh. "A comprehensive survey for scheduling techniques in cloud computing." Journal of Network and Computer Applications 143 (2019): 1-33.

[4] Aktas, Mehmet S. "Hybrid cloud computing monitoring software architecture." Concurrency and Computation: Practice and Experience 30, no. 21 (2018): e4694.

[5] Awotunde, Joseph Bamidele, Akash Kumar Bhoi, and Paolo Barsocchi. "Hybrid cloud/Fog environment for healthcare: an exploratory study, opportunities, challenges, and future prospects." Hybrid Artificial Intelligence and IoT in Healthcare (2021): 1-20.

[6] Mishra, Smita Prava, Sukant Kumar Sahoo, and Biswaranjan Jena. "Migrating on-premise application workloads to a hybrid cloud architecture." Journal of Information and Optimization Sciences 43, no. 5 (2022): 1099-1108.

[7] Tang, Qingqing, Zesong Fei, Bin Li, and Zhu Han. "Computation offloading in leo satellite networks with hybrid cloud and edge computing." IEEE Internet of Things Journal 8, no. 11 (2021): 9164-9176.

[8] Singh, Harinder Pal, Harpreet Singh, and A. K. Paul. "Dynamic ICT Modeling for Handling Student Data Using Big Data Technology and Hybrid Cloud Computing." In Computer Communication, Networking and IoT, pp. 9-21. Springer, Singapore, 2021.

[9] Daniel, R. Sheela, S. Raja, P. Ebby Darney, and Y. Harold Robinson. "Hybrid Cloud Computing Model for Big Data Analytics in Organization." In Further Advances in Internet of Things in Biomedical and Cyber Physical Systems, pp. 19-31. Springer, Cham, 2021.

[10] Shehab, Noha, Mahmoud Badawy, and H. Arafat Ali. "Toward feature selection in big data preprocessing based on hybrid cloud-based model." The Journal of Supercomputing 78, no. 3 (2022): 3226-3265.

[11] Thenmozhi, K., M. Pyingkodi, and K. Ramesh. "Hybrid Machine Learning Models for Distributed Biological Data in Multi-Cloud Environment." In Operationalizing Multi-Cloud Environments, pp. 19-29. Springer, Cham, 2022.

[12] Subramanian, K., F. Leo John, and F. L. John. "Dynamic and secure unstructured data sharing in multi-cloud storage using the hybrid crypto-system." International Journal of Advanced and Applied Sciences 5, no. 1 (2018): 15-23.

[13] Azad, Poopak, Nima Jafari Navimipour, Amir Masoud Rahmani, and Arash Sharifi. "The role of structured and unstructured data managing mechanisms in the Internet of things." Cluster computing 23, no. 2 (2020): 1185-1198.

[14] Sun, Wenjie, Zhiqiang Wei, Bowei Hong, and Yongquan Yang. "A Digital Ocean Cloud Platform Architecture Based on IPv6 Smart Gateway." In 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), pp. 438-442. IEEE, 2019.

[15] Lang, Hongbo, Hua Tian, Daping Li, Ziyang Niu, and Lijun Wen. "Design of A Cloud Native-Based Integrated Management Platform for Smart Operation of Multi-Business Buildings." In 2022 14th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), pp. 169-173. IEEE, 2022.

[16] Franca, Reinaldo Padilha, Yuzo Iano, Ana Carolina Borges Monteiro, Rangel Arthur, and Vania V. Estrela. "A proposal based on discrete events for improvement of the transmission channels in cloud environments and big data." In Big Data, IoT, and Machine Learning, pp. 185-204. CRC Press, 2020.

[17] Zhang, Xinliang, Chenlin Fu, Yunji Zhao, and Xiaozhuo Xu. "Hybrid feature CNN model for point cloud classification and segmentation." IET Image Processing 14, no. 16 (2020): 4086-4091.

[18] Zhifeng, Yang, Feng Xuehui, Han Fei, Yuan Qi, Cao Zhen, and Zhang Yidan. "Cloud Computing and Big Data for Oil and Gas Industry Application in China." Journal of Computers 1 (2019).

[19] Shafqat, Farzana, Muhammad Naeem A. Khan, and Sarah Shafqat. "SmartHealth: IoT-Enabled Context-Aware 5G Ambient Cloud Platform." In IoT in Healthcare and Ambient Assisted Living, pp. 43-67. Springer, Singapore, 2021.

[20] Sheikh, Anjum, Sunil Kumar, and Asha Ambhaikar. "IoT Data Analytics Using Cloud Computing." Big Data Analytics for Internet of Things (2021): 115-141.

[21] Rashid, Mamoon, Harjeet Singh, Vishal Goyal, Shabir Ahmad Parah, and Aabid Rashid Wani. "Big data based hybrid machine learning model for improving performance of

medical Internet of Things data in healthcare systems." In Healthcare Paradigms in the Internet of Things Ecosystem, pp. 47-62. Academic Press, 2021.

[22] Tsung, Chen-Kun, Chun-Tai Yen, and Wen-Fang Wu. "A software defined-based hybrid cloud for the design of smart micro-manufacturing system." Microsystem Technologies 24, no. 10 (2018): 4329-4340.

Author's biography

Dinesh Rajassekharan is currently working as a Senior Lecturer in the Faculty of Computer Science, Peninsula College, Malaysia. His area of research includes cloud computing and data analytics.