An Architecture for Big Data Privacy in the Hybrid Cloud

Waleed Al Shehri

Department of Computer Science, King Abdul-Aziz University Jeddah, KSA

waleed.ab2@gmail.com

ABSTRACT

The tremendous increases in the amounts of data generated and gathered by computer systems constitute a growing problem, as corporations tend to invest in hardware and software infrastructure but then use only a very limited proportion of these resources. Cloud computing was created to resolve this issue, with the vast amounts of data being termed "big data". Several tools were put in place to manage these prodigious quantities of data in efficient and effective ways, with the cloud essentially enabling sharing of resources among users to increase the throughput of the system. With this "shared resources" solution, questions over security and privacy have also arisen, resulting in the creation of different classes within the cloud, such as public clouds, private clouds and so on. However, when it comes to individual users, they only wish to secure a small subset, and being on a private cloud costs more than using the public variety. Many remedies have been offered for this dilemma, most of which are targeted at specific kinds of data or certain fields or industries. In this paper, we propose a generalised architecture to solve the privacy problem in cloud computing. Our solution includes considerations of the cost of privacy and ease of use. We will also discuss in detail the various solutions already proposed and their limitations. After providing details of our architecture, we will evaluate our work and analyse the results. The paper concludes with directions for future study in areas where more research is needed.

Keywords: *Software Architecture, Big Data, Hybrid Cloud, Privacy.*

1. INTRODUCTION

The volume of data captured by companies and organisations is increasing exponentially, what with the growth in technological usage in various fields, such as the Internet of Things and multimedia, and the expansion of social media. All these factors are causing the flow and storage of data to be enormous, with all types of data being stored, including in both structured and unstructured formats.

These data are being created by users, with the fastest-growing area being social media, which also includes multimedia data. When such a huge quantity of data is gathered, it is referred to as “big data”. The phenomenon has received much attention over the last couple of years, from industry leaders and academics alike [1]. The properties of big data include being of enormous size, consisting not only of relational data but other types as well and being continuously produced, stored and processed.

Cloud computing represents a major shift in storage and processing technology; the entire ICT industry has been affected by the development of cloud technologies. This type of computing has formed a powerful architecture for the delivery of large-scale, complex computing solutions. It provides various services to the modern ICT world, including virtual machines, parallel processors, security infrastructure and scalable storage services. These provisions are cost-effective in terms of both implementation and maintenance. A wide range of vendors, such as IBM, Microsoft Azure and Amazon AWS, provide cloud platforms, which are deployed by diverse applications. As we stated earlier, these applications gather and generate data. Virtualisation is one of the base technologies applicable to the implementation of cloud computing; the foundation for the many platform attributes that are required to access, store, analyse and manage distributed computing components in a big data environment is achieved through this technology [2]. Virtualisation is a process of resource-sharing and isolation of underlying hardware to increase the use, efficiency and scalability of computer assets.

The advent of cloud computing and shared resource utilisation has seen the concurrent emergence of concerns over privacy and security. As organisations, particularly small businesses, have the most to gain from the cloud, the protection of their business data is a major issue. Hybrid clouds offer a solution in this regard. Although private clouds were promulgated to provide enhanced security, their higher costs meant they became unfeasible for small enterprises. Hybrid clouds alleviate this problem by combining a dedicated private cloud environment for the safeguarding of security and privacy with the public cloud, which enables cheap processing and storage [3]. Together with big data, the hybrid cloud provides a basis for the development of custom architectures that support maximum flexibility in business processes in order to capitalise on the abovementioned advantages.

A range of novel approaches have arisen to secure private data in the hybrid cloud, such as security and event management applications, web firewall applications, and file integrity monitoring and intrusion prevention applications [4]. As explained by [5], although such concepts were originally developed to protect web servers against malicious attacks, they now complement and ensure the security of hybrid clouds.

The beauty of the hybrid cloud environment is that the data and applications one wishes to protect against threats, while maintaining privacy, can be placed on the private cloud component, while data that are non-sensitive in nature can be kept on the public cloud. While the privacy protection applications in the private cloud manage the transition of data, both the public and private cloud components can be extended when needed without any down time.

The users of cloud environments are web application administrators that host their companies‟ data on the servers. To enhance the confidence of these users that their information is secure and the privacy of data is preserved, we offer a management application that will enable these customers to manage their privacy and select which of their data items are to be stored in the public environment [6]. The users may choose the encryption settings for all their file systems, allowing them to own their own data protection mechanisms. Such considerations will bolster the users‟ confidence in the security and privacy of their data.

2. RELATED WORK

Much research has been conducted on cloud computing and big data as separate matters. That said, while some authors have restricted their purview to hybrid clouds alone, others have explored the implementation of big data in said clouds. We will be discussing the techniques and architectures that these research papers discuss in detail, as follows.

First, Huang et al. [6] discussed a technique for ensuring the privacy of image data in the cloud, which presented a one-to-one mapping function for image encryption. This technique renders the pairwise affinity among the jigsaw unreliable, and speeds up the process of substitution and diffusion. In this approach, in which random parameters of the mapping function are stored on the private cloud, more of the encrypted images are stored on the public cloud.

By reviewing the implementation of big data in the cloud, these authors investigated the status of big data tools that are supported in the cloud computing environment. They also provided definitions and characteristics of big data along with cloud computing, explained the relationship between these two technologies and presented big data storage systems and Hadoop technology. In addition, they identified relevant research challenges such as scalability, availability, data integrity, transformation, quality and so on. Ibrahim Abaker et al. [7] also discussed the privacy and legal issues as regards the usage of big data in the cloud environment.

Aisling et al. [8] also offered an overview of the cloud computing and big data technologies; in addition, by focusing on the specialised field of genomics, they furnished the reader with an idea of the scale of data generated by this field and how the cloud and big data are providing a solution to the problem. They also included an introduction to the big data technologies, such as the Apache and Hadoop projects and parallel data processing, and illustrated how the analysis of petabytes of data is done.

Meanwhile, Mianxiong et al. [2] discussed hybrid VM storage (HVSTO) and the privacy issues addressed by this framework. Their research also provided details of how the HVSTO uses the distributed structure to preserve the privacy of virtual machines, together with information on the threats posed by central systems to the privacy and security of data, and methods by which to improve the efficiency of distributed system IO tasks.

Certain of the science fields are already dealing with large amounts of sensitive data. Medical records represent one such field; mobile sensing, when used in medical monitoring of users, generates a great deal of data. Ying Zhang et al. [9], in their paper, expounded on the privacy issues inherent in using cloud and big data technologies and the perspective of patrons on their usage of the cloud. Their discussion showed that users feel more confident about using the cloud and big data when they know their data is secure and that they have control over what exactly is shared. The provision of such control, and the use of both the private and public clouds, has significantly improved the usability of the cloud and user confidence.

With the expansion of data and technologies that deal with them, security and privacy technologies have also evolved. One such technique is multi-factor authentication (MFA), which provides data security in certain internet and banking systems. Wenyi Liu et al. [10] proposed an architecture for the usage of MFA, along with the cloud, to preserve the privacy and security of users‟ data. With the wide range of research underway in the field, an impressive plethora of ideas are converging to solve the problems of public cloud environments.

3. THE PROPOSED ARCHITECTURE

The architecture we propose features several components, which work together to provide a hybrid cloud environment for users. The architecture can adopt several software and hardware configurations, to be discussed below.

3.1 Authentication

This primary component uses the ACL Manager feature via the IACL Manager interface to gain access to information. It provides an interface, iAuthenticate, which can be used by any other interface to authenticate the users of the cloud service.

3.2 Authorisation

This component uses the ACL Manager Components iACLAuthorise interface to collect the authorisation parameters from the ACL Manager and deploy them to authorise the users of the cloud to access the services that they are eligible to use.

3.3 ACL Manager

This is the component that provides authentication and authorisation data for the users. It gathers the data from the Fragmentation component and uses them to decide whether the user is authentic, which services are to be exposed to the user and which are to be hidden.

3.4 Host Manager

The Host Manager component provides services to authenticated and authorised users. It boasts an iHostManager interface that provides the functionality to read and write data to the cloud and allows the data to be searched. It uses the iWriteToPublicCloud interface to write the data to the public cloud and the iReadFromPublicCloud interface to read them. In addition, it provides search data functions through the Data Search component, as well as the Ifragmentation Interface, which both stores the data on the private cloud and reads them from it.

The iHost interface provides functionalities to those users that have already been authenticated and had their access levels set by the ACL manager. The users will invoice the component functionality to deploy the functionality of the system.

3.5 Data Search Component

This component uses the Meta Data Manager to provide the search functionality to the host manager; once thelatter is able to identify the data the user needs, they can be fetched from private storage.

3.6 Fragmentation Component

This component is used by the host manager to store and retrieve data in and from the private cloud. Data that are required to be kept in the private store, and that have higher privacy requirements, are fragmented before storage; the fragmentation component manages how the data should be fragmented, as well as how to deal with the fragmentation of structured, semi-structured and unstructured data. It uses the services from the Encrypt / Decrypt component to send the data to the private storage. The data are encrypted before storing, and upon retrieval are decrypted and de-fragmented to ensure they are as they were before fragmentation.

3.7 Meta Data Manager

This component‟s role is to maintain the search function in data that are to be fragmented; when data are fragmented, the normal search functions are limited, and accordingly the Meta Data Manager will assist in the search functions of the data that are being fragmented.

The indexing of the data on the cloud is a major function of this component, as, for the searching of data, the storage and retrieval of data and the indexing of files and their properties play important roles. The Data Downloader and Fragmentation Manager make use of the Meta Data Manager to garner information regarding the file storage on the cloud.

3.8 Encrypt / Decrypt Component

This component receives commands from the Fragmentation component. It encrypts the data sent to it, before sending them to the Update/Write component for storage. When it receives the decryption command, it reads from the Read component and decrypts the data, which are then sent to the Fragmentation component.

3.9 Update / Write

This component receives requests for storage on the private and public clouds, and sends these requests, together with the data, to the Infrastructure components.

3.10 Read

Finally, this component receives requests to read data from the private and public cloud. The component sends these requests to the Infrastructure components and forwards the data to the requesting component upon receiving it from the Infrastructure components.

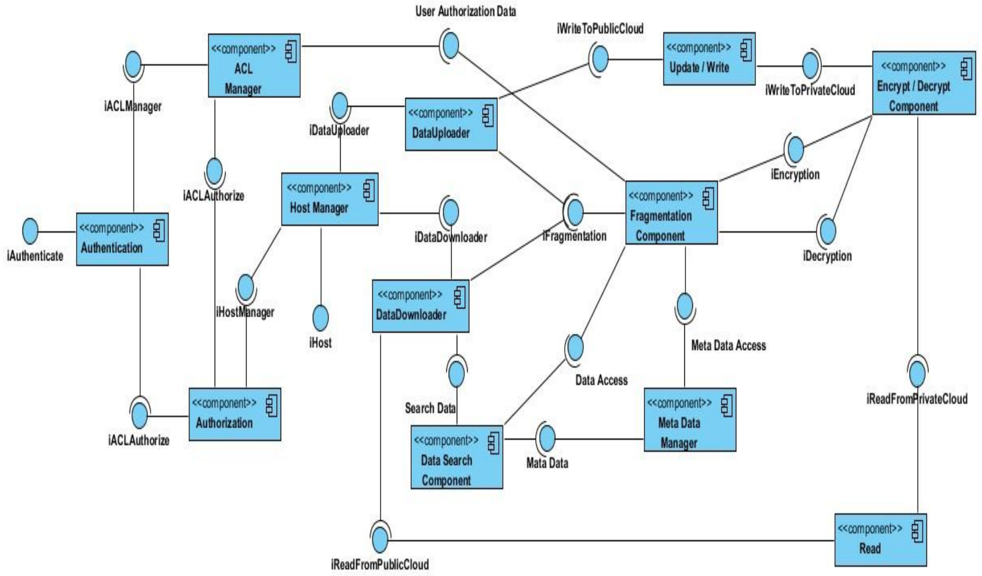
User Authorization Data
iWriteToPublicCloud
<<component>> ACL Manager
<<component>> Update / Write
iWriteToPrivateCloud
<<component>> Encrypt / Decrypt Component
iACLManager
iDataUploader
<<component>> DataUploader
iEncryption
iACLAuthorize
<<component>> Host Manager
iDataDownloader
iFragmentation
<<component>> Fragmentation Component
iDecryption
iAuthenticate
<<component>> Authentication
iHostManager
iHost
<<component>> DataDownloader
Meta Data Access
Data Access
Search Data
iACLAuthorize
<<component>> Authorization
iReadFromPrivateCloud
<<component>> Meta Data Manager
<<component>> Data Search Component
Mata Data
iReadFromPublicCloud
<<component>> Read

Fig. 1. The proposed architecture.

4. SYSTEM SEQUENCE DIAGRAMS

The system sequence diagrams depict the activation of system objects when the users interact with the hybrid cloud system. The sequence diagram explains the behavioural aspects of the architecture and the interactions between components.

4.1 Login

The login diagram shows the authentication, authorisation and access control in the system. The interactions between the user and the components are arranged in a timely manner.

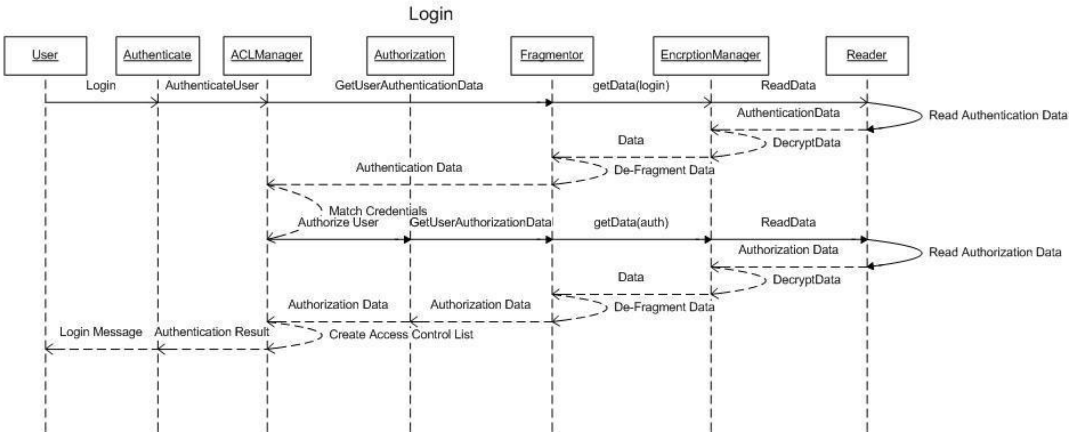

Fig. 2. Login sequence diagram.

4.2 Save File

The Save File sequence diagram displays the interaction of the authenticated and authorised user with the system; the host manager provides the functionality to the user and invokes the data uploader and fragmenting components, which in turn call on the encryption manager and writer/updater. The diagram shows the mechanism for the storage of files on the private and public servers in the hybrid cloud.

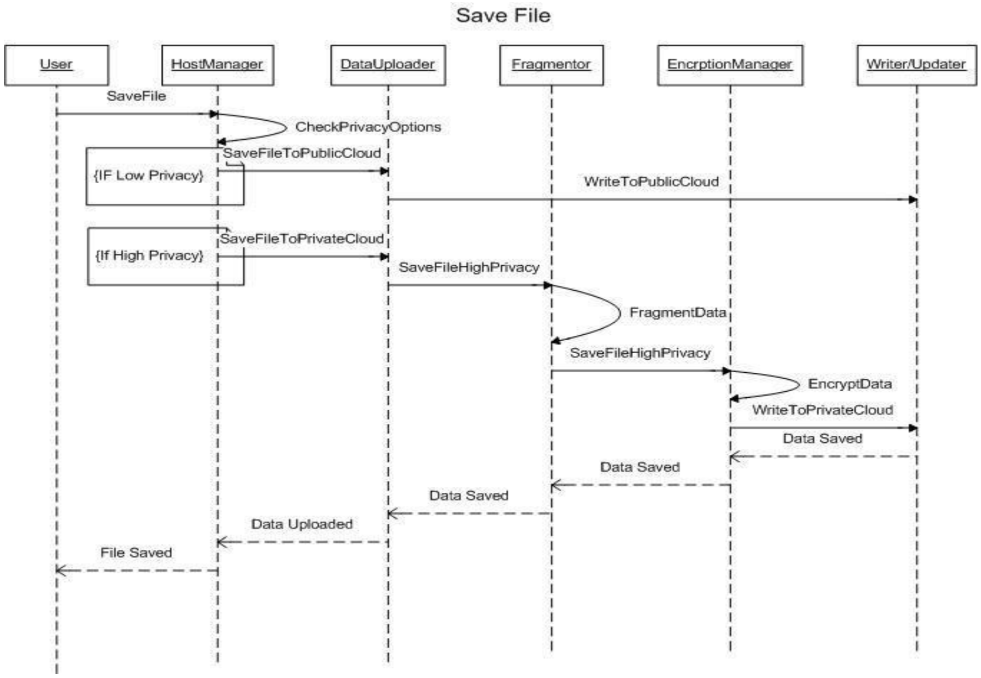

Fig. 3. Save File sequence diagram.

4.3 Search File

The Search diagram shows the interaction of the user with the system. To search the files, the host manager calls on the data searcher, which checks the meta data manager to get the file information; then, the meta data manager calls on the fragmenting component, which in turn calls the encryption manager for the data. The encryption manager calls on the reader component, which reads the data and sends them back to the encryption manager; The data are decrypted here and returned to the fragmentation manager, which de-fragments the data and returns them to the data searcher; finally, the file record is sent back to the host manager and then the user.

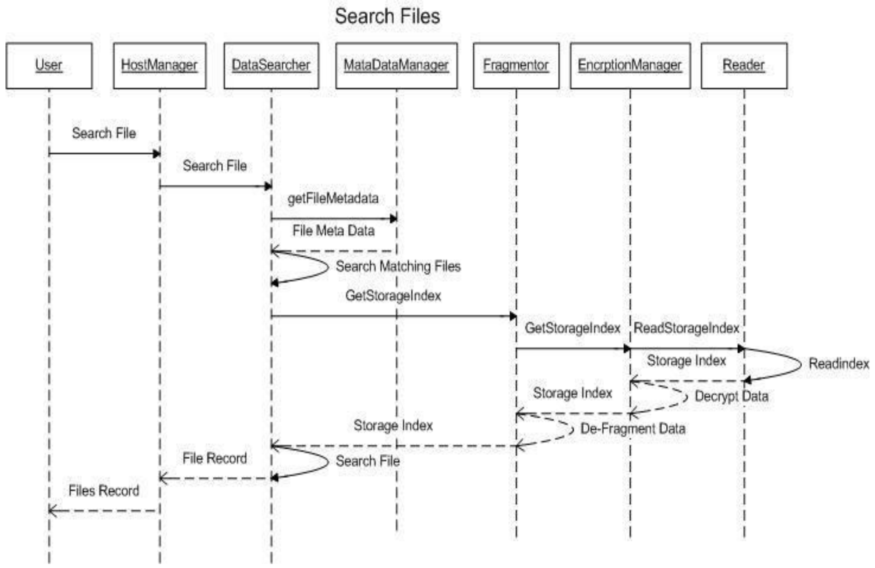

Fig. 4. Search file sequence diagram

4.4 Download File

The download file component calls on the host manager to download a file; the host manager then searches for the file location using the storage index. The storage index is read, decrypted and de-fragmented before the file record is finally searched for; if the file is found in the low privacy settings, then it is saved in the public cloud and the host manager will send a request to the reader to read the file. If the file has high privacy settings, the host manager will ask for the defragmenting component, which calls the encryption manager. The latter further calls the reader; then, the data are read, the encryption manager decrypts them and the fragmentation manager de-fragments them and returns the file to the host manager, which finally sends it back to the user.

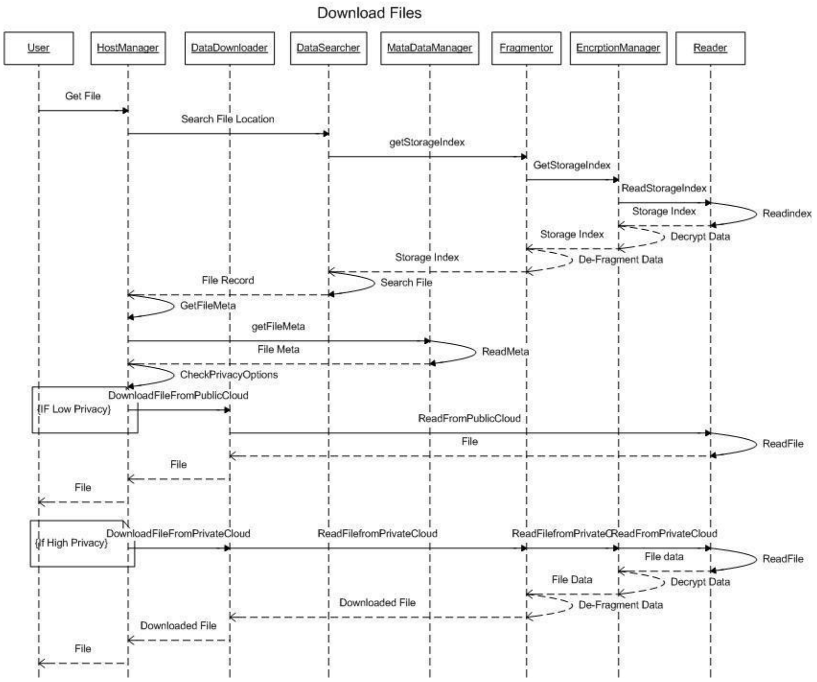

Fig. 5. Download File sequence diagram

5. TESTING AND EVALUATION

There are many ways in which we can apply our architecture to test the effectiveness and efficiency of the hybrid cloud we are offering [11]. The effectiveness of such a cloud can be measured by the security and privacy it provides, while its efficiency can be measured by its response time, scalability and certain other factors.

5.1 Security

Security is a major feature in the creation of any private cloud, as it is important for any supported applications. In the hybrid cloud, we are able to manage the security of sensitive data through the private cloud [12], while in our architecture we provide indexing that is accessible only by the cloud management app, as well as encryption of any kind of data that is present in the cloud; both of these measures, along with the private cloud, provide multiple layers of security to the data.

5.2 Privacy

The privacy of the content is our major concern; therefore, in the architecture we provide security that offers flexibility and confidence to the users, leaving them free to manage the privacy of their content. They can use the architecture‟s provided tools to maintain their privacy needs.

5.3 Efficacy

The hybrid cloud provides a means of managing data by storing them either in private or public, depending on their type and properties; if the data files are of large size, the management and checking of their properties while the storing and retrieving the data is straightforward. However, if the file size is small, the amount of files increases and the management of the file properties, searching of metadata and operational costs become complicated, reducing the efficiency of the hybrid mechanism.

5.4 Response Time

As with all other features, response time is also a concern, especially in a hybrid cloud, as there are at least two different networks involved; the communication latency has been reported in many studies. Our architecture provides indexing to solve this problem; by using this, we are able to easily locate the data and preload them in case we are able to design applications to use the architecture.

5.5 Scalability

The public and private clouds both provide scalability to the architecture, in which there is nothing to hinder the scalability of the underlying cloud architectures [13].

6. COMPARISON WITH OTHER MODELS

In the available literature, there are many studies that have focused on similar problems relating to big data privacy on cloud storage; some of them also proposed the use of a hybrid cloud, such as the study by Huang et al. [6], who presented a hybrid cloud abstraction to oversee the privacy management. Mianxiong et al. [2] and Qadri and Shaik [5] also provided details on managing a hybrid cloud for privacy preservation. The big data aspect poses a challenge for the management of privacy as the data is generated at enormous rate; therefore, the aim is to preserve the privacy of the ever-increasing amounts of data in the hybrid cloud.

7. FUTURE STUDIES

There are many aspects of the architecture that require further investigation, with many researchers working on features such as the encryption mechanisms, which require less processing, while others are focused on the I/O Latency timing [14]. Some of the features that I feel will affect our architecture and require further investigation are as follows:

7.1 Encryption / Decryption

Encryption is conducted in many ways, and for our architecture, we should investigate more concrete methods of performing the process. There are many encryption techniques that are only applicable to certain types of data, and it is necessary to look further into which of these techniques can be used with our generalised framework.

7.2 Public and Private Cloud I/O Latency

The public and private cloud send and receive data, with many researchers focusing on the input and output latency of such transmissions. This latency is a major factor in the effectiveness and efficiency of the hybrid cloud; as we are using public storage in our architecture, we should also further explore the possibilities of deploying reliable I/O methods [14].

8. CONCLUSION

The hybrid cloud provides a method that is both achievable and beneficial in terms of offering cost-effective solutions for handling the privacy of cloud-based big data. The data are generated continuously and accumulated in large quantities over the cloud, creating a "big data" opportunity; that said, the data belong to several users who are sometimes engaged in many different applications and fields, leading to a privacy and security issue. To address this problem, we have devised an architecture that enables users to select their privacy options, with the hybrid cloud using their preferences to manage their data security. This makes it possible to achieve privacy and security on the cloud while keeping it cost-effective for big data implementation.

Although there are many aspects of the architecture that require further investigation, we may conclude that the possibility of creating a generalised architecture for such a solution, which caters to all kinds of data, is possible and can yield promising results.

REFERENCES

[1] B. M. Purcell, “Big data using cloud computing,” Holy Family Univ., 2013.

[2] H. L. Mianxiong, et al., “Efficient privacy preserving hybrid storage in cloud data center,” 2014

[3] S. Goyal, “Public vs private vs hybrid vs community – Cloud computing: A critical review,” I. J. Computer Network and Information Security, vol. 3, pp. 20-29, 2014.

[4] E. Sayed Ali Ahmed and R. A. Saeed, “A survey of big data cloud computing security,” International Journal of Computer Science and Software Engineering (IJCSSE), vol. 3, no. 1, pp. 78-85, December 2014.

[5] S. J. Qadri and M. A. Shaik, “Acquiring big data privacy using hybrid cloud,” Aurora’s International Journal of Computing, vol. 2, no. 1, 2015.

[6] X. Huang, et al., “Achieving big data privacy via hybrid cloud,” in IEEE INFOCOM – 2014, Temple University, 2014.

[7] I. Abaker, et al., “The rise of „big data‟ on cloud computing: Review and open research issues,” Information Systems, no.47, pp. 98-115, 2015.

[8] Aisling, et al., “„Big data‟, Hadoop and cloud computing in genomics,” Journal of Biomedical Informatics, 2013. [9] J. Ying Zhang, et al., “Privacy-preserved mobile sensing through hybrid cloud trust framework.”

[9] A. Wenyi Liu, et al., MACA: A Privacy-Preserving Multi-Factor Cloud Authentication System Utilizing Big Data. Atlanta: Georgia Institute of Technology, 2014.

[10] C. Liu and C. Yang, “External integrity verification for outsourced big data in cloud and IoT: A big picture,” Future Generation Computer Systems, August 2014.

[11] S. Nepal, et al., “Trustworthy processing of healthcare big data in hybrid clouds,” in IEEE Computer Society, 2015. [13] D. Agrawal, et al., “Big data and cloud computing: Current state and future opportunities,” in 2011 ACM, 2011.

[12] C. Ji, et al., “Big data processing in cloud computing environments,” in International Symposium on Pervasive Systems, Algorithms and Networks, 2012.