Security Ensured Big Data Mining with Public Cloud Services

S. Muruganandham[1] and S. Deepa[2]

[1]Asst. Professor, Department Of Computer Science and Applications,

[2]M.Phil Research Scholar, Department of Computer Science,

Vivekanandha College of Arts and Science for Women. (Autonomous), Tiruchengode, India.

Abstract

Big data applications are constructed under the cloud environment to process the big data values. Public cloud provides easily scaled up and scaled down computing power and storage to everyone. Private cloud services are provided to group of people only. Big data can be used in disaster management, high energy physics, genomics, connectomics, automobile simulations and medical imaging applications.

Public cloud service components and private cloud data resources are integrated to form cross cloud services. Cross cloud service composition provides a concrete approach capable for large scale big data processing. Private clouds refuse to disclose all details of their service transaction records. History record based Service optimization method (HireSome-II) is privacy aware cross cloud service composition method. QoS history records are used to estimate the cross cloud service composition plan. k-means algorithm is used as a data filtering tool to select representative history records. HireSome-II reduces the time complexity of cross cloud service composition plan for big data processing.

Big data mining operations are integrated with the History record based Service optimization method (HireSome-II). Security and privacy is provided for cross cloud service composition based big data processing environment. Privacy preserved map reduce methods are adapted to support high scalability. The HireSome-II scheme is upgraded to support mining operations on big data.

1. Introduction

Processing large datasets has become crucial in research and business environments. Practitioners demand tools to quickly process increasingly larger amounts of data and businesses demand new solutions for data warehousing and business intelligence. Big data processing engines have experienced a huge growth. One of the main challenges associated with processing large datasets is the vast infrastructure required to store and process the data. Coping with the forecast peak workloads would demand large up-front investments in infrastructure. Cloud computing presents the possibility of having a large-scale on demand infrastructure that accommodates varying workloads.

Traditionally, the main technique for data crunching was to move the data to the computational nodes, which were shared. The scale of today's datasets has reverted this trend and led to move the computation to the location where data are stored. This strategy is followed by popular MapReduce implementations. These systems assume that data is available at the machines that will process it, as data is stored in a distributed file system such as GFS, or HDFS. This situation is no longer true for big data deployments on the cloud. Newly provisioned VMs need to contain the data that will be processed.

2. Related Work

Following the concept of delegation of decryption rights introduced by Mambo and Okamoto, Blaze *et al.*

formalized the concept of PRE and proposed a seminal bidirectional PRE scheme. Afterwards, many PRE schemes have been proposed, such as [7] and [10]. Employing traditional PRE in the context of IBE, Green and Ateniese initially introduced the notion of IBPRE and proposed two unidirectional IBPRE schemes in the ROM: one is CPA secure and the other holds against CCA. Later on, two CPA-secure IBE-PRE schemes have been proposed. Afterwards, some IBPRE systems have been proposed for various requirements. In the multiple ciphertext receiver update scenario, Green and Ateniese proposed the first MH-IBPRE scheme with CPA security. Later on, a RCCA-secure MH-IBPRE scheme without random oracles was proposed by Chu and Tzeng. These schemes are not collusion-safe. To solve the problem, Shao and Cao [1] proposed a CCA-secure MH-IBPRE in the standard model with collusion-safe property.

To hide the information leaked from re-encryption key, Ateniese et al. defined the notion of key-privacy. Later on, Shao et al. [3] revised the security model introduced. To prevent a ciphertext from being traced, Emura et al. [5] proposed a unidirectional IBPRE scheme in which an adversary cannot identify the source from the destination ciphertext. To ensure the privacy of both delegator and delegatee, Shao et al. [2] proposed the first Anonymous PRE (ANO-PRE) system. The system guarantees that an adversary cannot identify the recipient of original and re-encrypted ciphertext even given the corresponding re encryption key. In 2012, Shao also proposed the first anonymous IBPRE with CCA security in the ROM. In the context of IBE/ABE well-known systems supporting anonymity that have been proposed, such as [6] and [9]. Leveraging them may partially fulfill our goals. We need to focus on the combination of anonymity and ciphertext update properties. Therefore, the aforementioned systems are not taken in comparison below [8]. Here, we compare our work with the some related systems and summarize the comparison of properties. While multiple ciphertext receiver update, conditional share, collusion resistance, anonymity and without random oracle, have all five been partially achieved by previous schemes, there is no effective CCA-secure proposal that achieves all properties simultaneously in the standard model. This paper, for the first time, fills the gap.

3. Big Data in Clouds

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reduction and reduced risk. Analysis of data sets can find new correlations, to "spot business trends, prevent diseases and combat crime and so on." Scientists, business executives, practitioners of media and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations and biological and environmental research.

Data sets grow in size in part because they are increasingly being gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data were created; The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization. Work with big data is necessarily uncommon; most analysis is of "PC size" data, on a desktop PC or notebook that can handle the available data set. Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the users and their tools and expanding capabilities make Big Data a moving target. Thus, what is considered "big" one year becomes ordinary later. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

4. Cloud Services and Security

In recent years, Cloud Computing and big data receives enormous attention internationally due to various business-driven promises and expectations such as lower upfront IT costs, a faster time to market, and opportunities for creating value-add business. As the latest computing paradigm, cloud is characterized by delivering hardware and software resources as virtualized services by which users are free from the burden of acquiring the low level system administration details [4]. Cloud computing promises a scalable infrastructure for processing big data applications such as the analysis of huge amount of medical data. Cloud providers including Amazon Web Services (AWS), Salesforce. com, or Google App Engine, give users the options to deploy their application over a network of a nearly infinite resource pool. By leveraging Cloud services to host Web, big data applications can benefit from cloud advantages such as elasticity, pay-per-use and abundance of resources with practically no capital investment and modest operating cost proportional to actual use.

In practice, to satisfy different security and privacy requirements, cloud environments usually consist of public clouds, private clouds and hybrid clouds, which lead a rich ecosystem in big data applications. Generally, current implementations of public clouds mainly focus on providing easily scaled up and scaled-down computing power and storage. If data centers or domain specific services center tend to avoid or delay migrations of themselves to the public cloud due to multiple hurdles, from risks and costs to security issues and service level expectations, they often provide their services in the form of private cloud or local service host. For a complex web-based application, it probably covers some public clouds, private clouds or some local service host. For instance, the healthcare cloud service, a big data application illustrated, involves many participants like governments, hospitals, pharmaceutical research centers and end users. As a result, a healthcare application often covers a series of services respectively derived from public cloud, private cloud and local host.

In practice, some big data centers or software services cannot be migrated into a public cloud due to some security and privacy issues. If a web based application covers some public cloud services, private cloud services and local web services in a hybrid way, cross-cloud collaboration is an ambition for promoting complex web based applications in the form of dynamic alliance for value-add applications. It needs a unique distributed computing model in a network-aware business context.

Cross-cloud service composition provides a concrete approach capable for large-scale big data processing. Existing (global) analysis techniques for service composition, often mandate every participant service provider to unveil the details of services for network-aware service composition, especially the QoS information of the services. Unfortunately, such an analysis is infeasible when a private cloud or a local host refuses to disclose all its service in detail for privacy or business reasons. In such a scenario, it is a challenge to integrate services from a private cloud or local host with public cloud services such as Amazon EC2 and SQS for building scalable and secure systems in the form of mashups. As the diversity of Cloud services is highly available today, the complexity of potential cross-cloud compositions requires new composition and aggregation models.

As a cloud often hosts a lot of individual services, cross cloud and on-line service composition is heavily time-consuming for big data applications. It always challenges the efficiency of service composition development on Internet. Besides, for a web service which is not a cloud service and its bandwidth probably fails to match to the cloud, it is a challenge to trade off the bandwidth between the web service and the cloud in a scaled-up or scaled-down way for a cross-cloud composition application. Here, the time cost is heavy for cross-platform service composition. With these observations, it is a challenge to tradeoff the privacy and the time cost in cross cloud service composition for processing big data applications. In view of this challenge, an enhanced History record-based Service optimization method named HireSome-II, is presented in this paper for privacy-aware cross-cloud service composition for big data applications. In our previous work, a similar method, named HireSome has been investigated, which aims at enhancing the credibility of service composition. HireSome-I is incapable of dealing with the privacy issue in cross-cloud service composition. Compared to HireSome-I, HireSome-II greatly speeds up the process of selecting a optimal service composition plan and protects the privacy of a cloud service for cross-cloud service composition.

5. Problem Statement

Cloud computing environment provides scalable infrastructure for big data applications. Cross clouds are formed with the private cloud data resources and public cloud service components. Cross cloud service composition provides a concrete approach capable for large scale big data processing. Private clouds refuse to disclose all details of their service transaction records. History record based Service optimization method

(HireSome-II) is privacy aware cross cloud service composition method. QoS history records are used to estimate the cross cloud service composition plan. k-means algorithm is used as a data filtering tool to select representative history records. HireSome-II reduces the time complexity of cross cloud service composition plan for big data processing. The following drawbacks are identified from the existing system. The following issues are identified from the current cross cloud service composition methods.

- Big data processing is not integrated with the system
- Security and privacy for big data is not provided
- Limited scalability in big data process
- Mining operations are not integrated with the system

6. Security Ensured Big Data Mining with Public Cloud Services

History record based Service optimization method (HireSome-II) is enhanced to process big data values. Security and privacy is provided for cross cloud service composition based big data processing environment. Privacy preserved map reduce methods are adapted to support high scalability. The HireSome-II scheme is upgraded to support mining operations on big data.

Security and privacy preserved big data processing is performed under the cross cloud environment. Big data classification is carried out with the support of map reduce mechanism. Service composition methods are used to assign resources. The system is divided into six major modules. They are Cross Cloud Construction, Big Data Management, History Analysis, Map Reduce Process, Service Composition and Big Data Classification. Public and private clouds integrated in the cross cloud construction process. Big data management module is designed to provide big data for the cloud users. Resource sharing logs are analyzed under the history analysis. Task partition operations are performed under the map reduce process. Service provider selection is carried out service composition module. Classification process is carried out under the cross cloud environment.

6.1. Cross Cloud Construction

Private and public cloud resources are used in the cross cloud construction process. Big data values are provided under the data centers in private cloud environment. Service components are provided from public cloud environment. Public cloud services utilize the private cloud data values.

6.2. Big Data Management

Larger and complex data collections are referred as big data. Medical data values are represented in big data form. Anonymization techniques are used to protect sensitive attributes. Big data values are distributed with reference to the user request.

6.3. History Analysis

Service provider manages the access details in the history files. User name, data name, quantity and requested time details are maintained under the data center. History data values are released with privacy protection. Data aggregation is applied on the history data values.

6.4. Map Reduce Process

Map reduce techniques are applied to break the tasks. Map reduce operations are partitioned with security and privacy features. Redundancy and fault tolerance are controlled in the system. The data values are also summarized in the map reduce process.

6.5. Service Composition

HireSome-II scheme is adapted for the service composition process. History records are analyzed with K-means clustering algorithm. Privacy preserved data communication is employed in the system. Public cloud service components are provided to the big data process.

6.6. Big Data Classification

Medical data analysis is carried out on the cross cloud environment. Privacy preserved data classification is applied on the medical data values. Public cloud resources are allocated for the classification process. Bayesian algorithm is tuned to perform data classification on parallel and distributed environment.

7. Conclusion

Service composition methods are used to provide resources for big data process. History record based Service optimization method (HireSome-II) is used as privacy ensured service composition method. HireSome-II scheme is enhanced with privacy preserved big data process mechanism. Map reduce techniques are also integrated with the HireSome-II scheme to support high scalability.Security and privacy are provided for the big data and history data values under the cloud environment. Map reduce techniques reduces the computational complexity in big data processing. Data classification is performed on sensitive big data values with cloud resources. Efficient resource sharing is performed under cross cloud environment

References

[1] J. Shao and Z. Cao, "Multi-use unidirectional identity-based proxy re encryption from hierarchical identity-based encryption," *Inf. Sci.*, vol. 206, pp. 83–95, Nov. 2012. [Online]. Available: http://dx.doi org/10.1016/j.ins.2012.04.013

[2] J. Shao, P. Liu, G. Wei, and Y. Ling, "Anonymous proxy re-encryption," *Secur. Commun. Netw.*, vol. 5, no. 5, pp. 439–449, May 2012.

[3] J. Shao, P. Liu and Y. Zhou, "Achieving key privacy without losing CCA security in proxy re-encryption," *J. Syst. Softw.*, vol. 85, no. 3, pp. 655–665, 2011. [Online]. Available: http://doi:10.1016/j.jss.2011.09. 034

[4] Kaitai Liang, Willy Susilo and Joseph K. Liu, "Privacy Preserving Ciphertext Multi-Sharing Control for Big Data Storage", IEEE Transactions On Information Forensics And Security, Vol. 10, No. 8, August 2015

[5] K. Emura, A. Miyaji and K. Omote, "An identity-based proxy re-encryption scheme with source hiding property, and its application to a mailing-list system," in *Public Key Infrastructures, Services and Applications* (Lecture Notes in Computer Science), vol. 6711. Berlin, Germany: Springer-Verlag, 2011, pp. 77 92.

[6] C.-I. Fan, L.-Y. Huang and P.-H. Ho, "Anonymous multireceiver identity based encryption," *IEEE Trans. Comput.*, vol. 59, no. 9, pp. 1239–1249, Sep. 2010.

[7] K. Liang, J. K. Liu, D. S. Wong and W. Susilo, "An efficient cloudbased revocable identity-based proxy re-encryption scheme for public clouds data sharing," in *Computer Security–ESORICS* (Lecture Notes in Computer Science), vol. 8712. Berlin, Germany: Springer-Verlag, Sep. 2014, pp. 257–272.

[8] T. Mizuno and H. Doi, "Secure and efficient IBE-PKE proxy re-encryption," *IEICE Trans. Fundam. Electron., Commun., Comput. Sci.*, vol. E94-A, no. 1, pp. 36–44, 2011.

[9] Y. S. Rao and R. Dutta, "Recipient anonymous ciphertext-policy attribute based encryption," in *Information Systems Security* (Lecture Notes in Computer Science), vol. 8303. Berlin, Germany: Springer-Verlag, 2013, pp. 329–344.

[10] R. Lu, X. Lin, J. Shao and K. Liang, "RCCA-secure multi-use bidirectional proxy re-encryption with master secret security," in *Provable Security* (Lecture Notes in Computer Science), vol. 8782. Berlin, Germany: Springer-Verlag, 2014, pp. 194–205.