A Novel and Cost-Effective Approach for Privacy, Storage and Computation of Intermediate Data Sets in Cloud

S. L. M. DEEPTHI¹, G. YAMINI¹, P. SRINIVASA RAO¹, K. SIVARAMAKRISHNA²
¹Department of Computer Science and Engineering, MVGRCE, Vizianagaram, India
²Department of Computer Science and Engineering, Loyola Institute of Engineering and Technology
Email: slmdeeph8@gmail.com, yamini.teddy@gmail.com, psr.sri@gmail.com, sivaramkosuru@gmail.com

Abstract: Cloud computing has embarked as the succeeding generation architecture in providing a remarkable methodology in storing and accessing highly adaptable data sets on par of IT Enterprises. It swears to abolish the need of maintaining expensive computing facilities by diverse companies and institutes alike. It has become a familiar buzzword now-a-days. However, privacy and security issues always present a very strong obstruction for users to attune them in the era of Cloud computing. While processing applications in the cloud a large volume of intermediate data sets will be generated and are stored to reduce the cost of re-computation. But, safeguarding the privacy of those sets is a challenging problem and may recover privacy sensitive information by analyzing those sets. Encrypting all the existing data sets or the part of intermediate data sets is well chosen in the existing approaches to address the challenge of privacy. Here, we propose a novel upper bound privacy leakage constraint-based approach for privacy aware efficient scheduling of intermediate data sets by taking privacy as a metric together with many other metrics like storage and computation. Evaluation results would reveal that the cost would be significantly low when compared to the existing ones where either all or part of intermediate data sets are en/decrypted.

Keywords: Cloud Computing, Privacy, Intermediate Data Sets.

I. Introduction:

Cloud computing has now become universal. This refers in terms of both applications and services provided over the internet in prior relevance with the hardware and also to those services provided by the systems software in the data centers. It incorporates flexible on-demand and low-cost usage of computing resources which in turn causes security and privacy related threats as the data owned by different users are stored in cloud servers where there is always a greater chance for an unauthorized user to access it. Thus, preventing this threat has become a major issue to the world. Taking this into consideration a new technology is being implemented based on the next generation IT architecture which enables more structured and secured methods of data storage, privacy, and security to the world at a minimal cost. Cloud computing archetype carry great commitment for the performance-hungry scientific computing community thereby serving as a cheap substitute to the supercomputers [7].

The key challenge of cloud computing is to address several technological realms into reality [5]. Security, privacy and trust are the additional cross-cutting challenges in cloud computing along with many existing issues in terms of cost and resource savings, scalability and reduced time line for implementation. Privacy plays a very important role and without which data authentication cannot be achieved there by making it more fraudulent. It is yet a challenging issue to provide privacy and security for the data stored in the cloud. Privacy preserving for intermediate data sets in cloud computing is one of the predominant yet challenging research issue and hence needs thorough investigation.

In cloud computing adaptable storage and computation capabilities are obtained in the form of services and are demanded according to the utilization. Hence computation sources in the cloud are identical to the storage sources. Users are allowed to store intermediate data selectively while processing the raw data on cloud. Privacy preserving technique like the generalization reduces the attacks on single data set whereas on multiple data sets is challenging. Hence, anonymizing the entire data sets and encrypting them before storing or sharing the same in the cloud is well adopted. But, this volume might be very large so we encrypt only part of them and reduce the privacy preserving cost to a greater extent possible. This intermediate data set storage expands the attack surface preserving the original data. Intermediate dataset will be out of control of the data owner and can be retrieved and shared by various applications. Since completely eliminating the privacy leakage problem is difficult to achieve, we can try to reduce the overall cost of privacy preservation as much as possible. The first possible solution to achieve privacy would be encrypting the data and thereby securing the message [10]. En/Decrypting the whole data sets consume more time in general. We first select which intermediate data sets need to be encrypted and which need not to be thereby saving the privacy preserving cost. We use an upper bound constraint to confine the privacy disclosure. For this task to be achieved we design a heuristic algorithm to identify privacy aware efficient
scheduling for intermediate data sets in the cloud. Experimental results would highly determine that privacy preserving cost of scheduling of intermediate data sets can be significantly reduced when compared to the existing methods where all data sets are encrypted.

The reminder of this paper is organized as follows. Related work is reviewed in the next section. Approach to reduce the privacy preserving cost is discussed in section III. Motivating example behind this approach is summarized in section IV. The proposed model along with the architecture and algorithm are formulated in section V.

II. Related work:

We consider the privacy and security issues in the cloud and up to what extent these play a very important role in an emerging world. Once we select the data to be encrypted we must even select up to what extent the encryption key is to be chosen so as to provide more security and ensuring data to be error free. We review the research on privacy and consider the economical aspects adhering the pay-as-you-go feature of cloud computing.

Rafael Moreno et al. [5] explained Key Challenges in cloud computing as it plays an important role in the subsequent internet of services there by enabling on demand provision of applications, platforms and computing infrastructures. He addressed various technology challenges to transform this vision into reality. Ming Li et al. [13] proposed Privacy assured and searchable cloud data storage services as to how efficiently we have to support flexible data utilization services such as search over data in a privacy preserving manner via privacy assured search schemes like top down methodology. Cong Wand et al. [9] proposed Privacy preserving public auditing for secure cloud storage as users can have on demand high quality applications and services without burden of maintaining local data storage. Third party Auditor ensures data integrity is maintained thereby not introducing new vulnerabilities towards data privacy. Mohamed Nabeel et al. [8] in his paper Privacy Preserving Delegated access control in public clouds as two layer encryption in order to provide more security to the data. Mingyi Zhou et al. [12] in his paper Security and privacy in cloud computing explained that security and privacy is not adequate and more privacy is to be incorporated in terms of availability confidentiality, data integrity, control and audit. Secret information of individual users are stored and managed by the service providers via Infrastructure as a Service(IaaS), Software as a Service(SaaS), Platform as a Service(Paas), etc...

Hamid Baniroostam et al. [1] proposed A Trust based approach for increasing security in cloud computing infrastructure as confidentiality, data accuracy and data integrity is lagging in un trusted cloud and that User Trusted Entity (UTA) would resolve these problems. Lei Wang et al. [4] proposed Scientific communities benefit from the economies of scale as research infrastructure is almost restricted to hardware component and software infrastructure is neglected, hardware must be provided by experts to achieve efficiency and to control the expenses thereby proposing cloud usage models and build the systems that enable the scientific community to benefit from the economy of scale of the public clouds. Patrick P. Tsang et al. [6] in his paper has analyzed that, administrators of various websites depend on the IP-address to block misbehaving users but it is not practical when abusers route from an anonymizing network and hence administrators block all the known exit nodes to deny the access for misbehaving users.

Michaela Gotz et al. [11], discussed “Publishing Search Logs: A Comparative Study of Privacy Guarantees” as anonymity in search logs are insufficient in the light of attackers who can actively effect the search logs and is impossible to achieve good utility with different privacy.

Qian Wang et al. [2] Proposed “Enabling Public Auditability and Data Dynamics for Storage Security in Cloud computing” that we allow a third party auditor(TPA) on behalf of the cloud client to verify the integrity of the data stored in the cloud thereby reducing the burden of the client to verify the data stored is indeed intact or not. Salvatore J. Stolfo et al. [3] proposed “Fog Computing: Mitigating Insider Data Theft Attacks in the Cloud” by introducing new offensive decay technology can be used to secure the data stored in the cloud.

Encrypting all the available data sets a straight forward way is widely adopted in the current research [15]. But processing only encrypted form of data sets is a difficult task since most applications run on unencrypted data sets. Generally, intermediate data sets are available to access and process by multiple users but are rarely under the control of the original holders Ciriani et al. [16] in his paper proposed an approach to combines encryption and the data fragmentation for achieving privacy protection in distributed data storage via encrypting only the part of the data sets.

Taking into account all the above issues, this paper is going to address privacy via efficient scheduling of intermediate data sets in the cloud by taking privacy as a metric along with many other metrics such as storage and computation. Optimized scheduling strategies would be developed for high efficient privacy aware data scheduling.

III. Approach:

Here, we identify intermediate data sets that need to be encrypted first via an approach of scheduling the data sets so as to provide more privacy with less storage and computation time when compared to the encryption of the complete data sets.
IV. Motivating Example:
The motivating model behind this approach is shown below explaining how the data sets are encrypted and stored in the cloud for confidentiality.

Figure 1: A typical scenario depicting need for privacy using intermediate data sets.

The above scenario provides an overview regarding online health service providers. The health centre has moved its data storage into the cloud for economical benefits. Original data sets are being encrypted to preserve confidentiality. In general there always raises issues regarding what data is to be made specific depending on time aspect. In this case as depicted in the figure 1., we have taken two primary cases where a lady with zip code 214 effected by diabetics and HIV, and at the same stage been reported to have been effected by HIV and flu. It is an advisory who claims of which medical aspect to be made confidential to avoid data breach. Let us consider that the frequency of accessing data from 1.a is 10 and that of 1.b is 50, now it is in the hands of the corresponding advisory body to encrypt needy data for high confidentiality. In the same manner, we come across many situations where we need certain data to be made confidential to prevent breaches to the system, which is always a challenging issue. Here a steady work has to be done in increasing the level of security by proposing large intermediate data sets in promising confidentiality at lower cost.

V. Proposed system model:
A. Architecture:

Figure 2 illustrates how intermediate data set is being scheduled and encrypted in cloud. Client sends the intermediate data sets for calculation of storage and computation time. The result is given to the master node from which the data set with less storage and computation time is selected for encryption and thereby obtaining the scheduled encrypted intermediate data set.

B. Algorithm:
Step1: Client sends the data set.
Step2: Storage and computational time is calculated.
Step3: Data obtained is sent to master node.
Step4: Scheduling is performed on the intermediate data sets obtained from Step3.
Step5: Encryption is then performed.
Step6: Privacy aware scheduling of the intermediate data sets is obtained in encrypted format.

VI. Methodology:
Client being at a terminal submits data sets to the user file from which the storage and the computational time is calculated and stored in the master node, then the data set with least storage space and computation time is extracted and is encrypted first and then the encrypted file is obtained at the end. All together all the intermediate data sets selected are encrypted thereby following the above fashion. Encryption can be done can be done using various algorithms.

VII. Conclusion and future work:
In this paper an approach is proposed to describe the privacy and security framework of scheduling the intermediate data sets in cloud. Hence, by using above technique we could achieve less storage and less computation time. With the contribution of this paper, we are planning to further probe privacy preserving for scheduling of the intermediate data sets by considering the semantics of the data sets in Cloud.

References:
[4] Lei Wng, Jianfeng Zhan, Weisong Shi, Yi Liang, “In Cloud, can Scientific Communities


