Scientific Journal of Impact Factor (SJIF): 4.14
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 9, September -2016

Privacy preservation data sets on cloud in quasi-identifier method

Rayapati Venkata Sudhakar[1], Dr.T.CH.Malleswara Rao[2]

[1] *CSE dept. Research scholar JNTUH*
[2] *Professor, CSE dept, VBIT*

Abstract -- *Cloud computing is a compilation of existing techniques and technologies, packaged within a new infrastructure paradigm that offers improved scalability, elasticity, business agility, faster startup time, reduced management costs, and just-in-time availability of resources Also a massive concentration of risk expected loss from a single breach can be significantly larger concentration of "users" represents a concentration of threats Ultimately, Cloud allows to store sensitive data in which the digital data is stored in logical pools, the physical storage spans multiple server's physical environment is typically owned and managed by a hosting company. Privacy is most important sensitive data ,But the privacy requirements can be potentially violated when new data join over time Exiting methods address this problem via re-anonym zing datasets from scratch and privacy preservation over incremental data sets is still challenging in the context of cloud because most data sets are of huge volume and distributed across multiple storage anodes exiting approaches suffer from poor scalability and inefficiency because they are centralized and access all data frequently when update occurs. In this paper, we propose an efficient quasi-identifier index based approach to ensure privacy preservation and achieve high data utility over incremental and distributed data sets on cloud. Quasi-identifiers, which represent the groups of anonymized data, are indexed for efficiency.*

Key words*: Cloud, scalability, sensitive data, anonym zing, quasi-identifier.*

I. INTRODUCTION

cloud computing can be regarded as an ingenious combination of a series of developed or developing ideas and technologies, establishing a pay-as-you-go business model by offering IT services using economies of scale [1–3]. The cloud acts as a big black box, nothing inside the cloud is visible to the clients Clients have no idea or control over what happens inside a cloud Even if the cloud provider is honest, it can have malicious system admits who can tamper with the VMs and violate confidentiality and integrity Clouds are still subject to traditional data confidentiality, integrity, availability, and privacy issues, plus some additional attacks

Delivery Models

- SaaS (Software-as- a service)
- PaaS (Platform-as- a service)
- IaaS (Infrastructure -as- a service)

Deployment Models

- Private cloud
- Community cloud
- Public cloud
- Hybrid cloud
-

2. ANONYMIZATION

Participants in cloud computing business chains can benefit from this novel business model, as they can save huge IT capital investment by facilitating cloud services such as high storage and computation capabilities, and consequently can concentrate on their own core business. Cloud computing also provides attractive features for science applications in academia . Moreover, since cloud is a multi-tenant environment, it is convenient for cloud users to share data and collaborate with each other. Therefore, many companies or organizations have built up IT systems for their business in cloud computing environments. However, numerous potential cloud customers are still hesitant to take advantage of

cloud computing due to security and privacy concerns . Privacy protection is one of the concerned issues in this regard. and hospitals have deployed their health services into cloud, e.g., Microsoft Health Vault . The data sets retained in these cloud applications are highly privacy-sensitive. Once an adversary collects these data sets and menaces the privacy-sensitive information, considerable economic loss or severe social reputation impairment will be caused to corresponding individuals. Usually, these data in cloud will be shared and utilized by multiple users for value-added advantages ,rather than just for data storage. Encrypting all the data sets is a straightforward and effective privacy protection approach. However, processing effectively and efficiently on encrypted data sets on cloud can be quite a challenging task, because most existing applications run on unencrypted data sets. Recent progress has been made in homomorphism encryption research. Theoretically, computation can be performed on encrypted data sets without decrypting them, but the current techniques are rather expensive and impractical with respect to its efficiency . Worse still, encryption still fails to protect individual privacy-sensitive information to legal data users even thoughitcanen sure confidentiality against adversaries.Assuch, data anonymization techniques like generalization and anatomization have been proposed to preserve privacy when privacy-sensitive data are stored in cloud. Data sets are anonymized to satisfy certain privacy requirements such as *k*-anonymity , or *l*-diversity before they are shared with data users. The explosive growth of data sets in cloud applications poses a challenge to existing approaches of privacy preservationyou can outsource responsibility but you can't outsource accountability.A high-level discussion of the fundamental challenges and issues/characteristics of cloud computingIdentify a few security and privacy issues within this framework Propose some approaches to addressing these issues Preliminary ideas to think about Cloud computing definitely makes sense if your own security is weak, missing features, or below average. Ultimately, ifthe cloud provider's security people are "better" than yours (and leveraged at least as efficiently),the web-services interfaces don't introduce too many new vulnerabilities, and the cloud provider aims at least as high as you do, at security goals, then cloud computing has better securityUse of internet-based services to support business process Rent IT-services on a utility-like basis Rapid deployment Low startup costs/ capital investments Costs based on usage or subscription Multi-tenant sharing of services/ resources We propose one more Model: Management Models (trust and tenancy issues Self-managed 3rd party managed (e.g. public clouds and VPC)Most security problems stem from Loss of control Lack of trust (mechanisms)Multi-tenancy These problems exist mainly in 3rd party management models Self-managed clouds still have security issues, but not related to above Data, applications, resources are located with provider User identity management is handled by the cloud User access control rules, security policies and enforcement are managed by the cloud provider Consumer relies on provider to ensure

- Data security and privacy
- Resource availability
- Monitoring and repairing of services/resources

3. PRIVACY PRESERVATION

Privacy issues raised via massive data mining Cloud now stores data from a lot of clients, and can run data mining algorithms to get large amounts of information on clients Increased attack surface Entity outside the organization now stores and computes data, and so Attackers can now target the communication link between cloud provider and clien Cloud provider employees can be phished

Large-volume datasets are partitioned in to avarietyofrelatively small data sets which are then stored in cloud data nodes. We partition original generalized data sets according to QI-groups. Indexed by domain values in their quasiidentifiers in the current generalization level. Statistical information about QI-groups is computed to facilitate data updates when new data arrive. After initialization of original anonymized data sets, we are able to handle data updates when new data are added. The new data are generalized to the current generalization level and added to corresponding QI-groups. Statistical information affected by newly added data is updated accordingly. Then we check whether *k*-anonymous statusis violated and whether anonymized data sets are over -generalized. In the form ercasegeneralization will be performed onthe data sets, while in the later case specialization will be performed. Finally, statistical information is updated again forthe next update. We propose an algorithm named Quasi-identifier Indexed based Privacy Preservation algorithm (QuIPP)to realize our approach described above.Algorithm1brieflydescribes the algorithm. Details for this algorithm will be formulated in the following sections. Data breaches have a cascading effect Full reliance on a third party to protect personal data? In-depth understanding of responsible dataset warship Organizations can transfer liability, but not accountability Risk assessment and mitigation throughout the data life cycle is critical. Many new risks and unknowns

the overall complexity of privacy protection in the cloud represents a bigger challenge Problem Statement: authorizing data access scopes (relations, attributes, tuples) to users of DBMS Discretionary access control Authorization administration policies, ie, granting and revoking authorization (centralized, ownership, etc) Content-based using views and rewriting for fine-grained access control Role-based access control: a function with a set of actions, consisting of users members Mandatory access control Object and subject classification (eg, top secret, secret, unclassified, etc).Problem: protecting Personally Identifiable Information (PII) and their sensitive attributes

Quasi-identifier			Sensitive
DOB	*Gender*	*Zipcode*	*Disease*
1/21/76	Male	53715	Heart Disease
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Flu
2/28/76	Female	53706	Hang Nail

Table 1

Quasi-identifiers indistinguishable among k individuals Implemented by building generalization hierarchy or partitioning multi-dimensional data space

At least l values for sensitive attributes in each equivalence class

Zip code	*Age*	*Salary*	*Disease*
476**	2*	20K	Gastric Ulcer
476**	2*	25K	Gastritis
476**	2*	30K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Table 2

Problems: hide sensitive rules or private individual data in data mining [Verykios et al. SIGMOD'04]

1. sanitize sensitive item sets or sensitive rules2. build data mining model without access to precise data, e.g. privacy-preserving classification, clustering 3. private parties compute together on their private inputs, e.g. distributed association rule mining, collaborative filtering

1. Data perturbation, blocking → rule confusion
2. Data perturbation → Distribution reconstruction [Agrawal et al. SIGMOD'00, PODS'01]
3. Secure Multi-party Computation (SMC) [Clifton et al. KDD'02]

This need strong privacy notion that is independent of arbitrary external information guarantee little risk for an individual joing a database a randomized function K gives €-differential privacy if for all databases D and D differing in at most

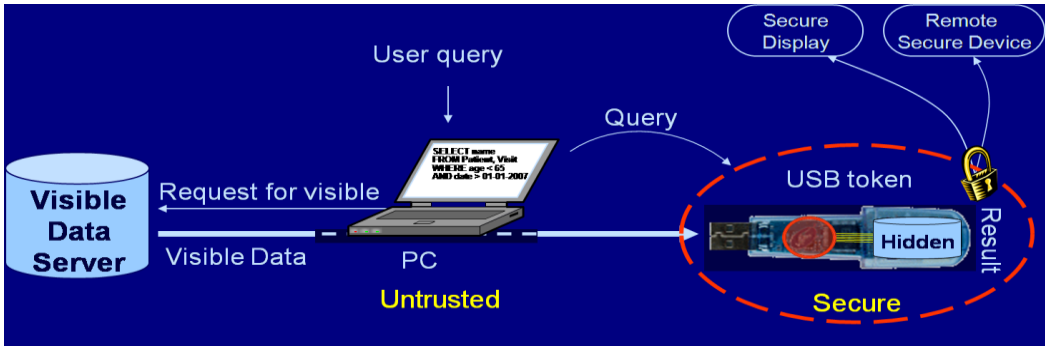

Fig 1

Protects data from steeling but plaintext data can still be seen on the server Write – encrypt before storing insert into lineitem (discount) values (encrypt(10,key)) Read – decrypt before access select decrypt(discount, key) from lineitem where custid = 300 Encryption alternatives Software level v.s. Hardware level (cryptographic coprocessor) encryption Granularity: field, row, page Directly search on encrypted data without decryption on server side Encrypt word by word. For word W_i Block_ciphertext $X_i = E_k(W_i)$, Word key $k_i = f_k(X_i)$, Pseudorandom sequence $T_i = <S_i, F_{ki}(S_i)>$ Searchable_ciphertext $C_i = X_i \quad T_i$ Search for a word W Block_ciphertext $X = E_k(W)$, Word key $k_i = f_k(X)$ Check cipher texts one by one to see if C $X = (X_i \quad T_i)$ X is of the form $<s, F_{ki}(s)>$ for some random value s

4. CONCLUSION

In this paper, we have investigated the challenge about how to efficiently update huge-volume incremental data sets to ensure privacy requirements of data owners and simultaneously achieve high data utility to data users. We have proposed and efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud. In our approach, I-groups (QI: quasi-identifier) are indexed by the domain values in the current generalization level, which makes it possible to access only a part of records in a data set in the presence of data updates rather than access all data records as required by existing approaches. To further improve the performanceofquasi-identifierindexing,locality-sensitivehashingmethodis incorporated to place similar QI-groups on the same data storage nodes. Thus, the number of data nodes that a QI-group link across will be reduced considerably with high probability. Based on the established indexes of an anonymized data set, we have designed an efficient quasi-identifier index based privacy preservation algorithm (QuIPP) for our approach .Evaluation results on real-world data sets have demonstrated that with our approach, the efficiency of privacy preservation on large-volume incremental data sets can be improved significantly over existing approaches. In accordance with various data and computation intensive applications on cloud, processing of huge-volume incremental data sets is becoming an important research area. Privacy preservation for such data set Silone of important yet challenging research issues, and needs thorough investigation. With the contributions of this paper, we plan to investigate privacy-aware efficient scheduling of anonymized data sets in cloud by taking privacy preservation a same trice together with other metrics.

REFERENCES

1. NIST (Authors: P. Mell and T. Grance), "The NIST Definition of Cloud Computing (ver. 15)," National Institute of Standards and Technology, Information Technology Laboratory (October 7 2009).

2. J. McDermott, (2009) "Security Requirements for Virtualization in Cloud Computing," presented at the ACSAC Cloud Security Workshop, Honolulu, Hawaii, USA, 2009.

3. J. Camp. (2001), "Trust and Risk in Internet Commerce," MIT Press4.T. Ristenpart et al. (2009) "Hey You Get Off My Cloud," Proceedings of the 16th ACM conference on Computer and communications security, Chicago, Illinois, USA

4. Security and Privacy in Cloud Computing, Dept. of CS at Johns Hopkins University. www.cs.jhu.edu/~ragib/sp10/cs412

5. Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance by Tim Mather and Subra Kumaraswamy

6. Afraid of outside cloud attacks? You're missing the real threat. http://www.infoworld.com/d/cloud-computing/afraid-outside-cloud-attacks-youre-missing-real-threat-894

7. Targeted Attacks Possible in the Cloud, Researchers Warn.http://www.cio.com/aricle/506136/Targeted_Attacks_Possible_in_the_Cloud_Researchers_Warn

8. VulnerabilitySeenin Amazon's Cloud-Computing by David Talbot. http://www.cs.sunysb.edu/~sion/research/sion2009mitTR.pdf

9. Cloud Computing Security Considerations by Roger Halbheer and Doug Cavit. January 2010. http://blogs.technet.com/b/rhalbheer/archive/2010/01/30/cloud-security-paper-looking-for-feedback.aspx

10. Security in Cloud Computing Overview.http://www.halbheer.info/security/2010/01/30/cloud-security-paper-looking-for-feedbackHey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds by T. Ristenpart, E. Tromer, H. Shacham and Stefan Savage. CCS'09

11. Cloud Computing Security. http://www.exforsys.com/tutorials/cloud-computing/cloud-computing-security.html

12. Update From Amazon Regarding Friday's S3 Downtime by Allen Stern. Feb. 16, 2008. http://www.centernetworks.com/amazon-s3-downtime-update

13. R. Ranchal, B. Bhargava, L.B. Othmane, L. Lilien, A. Kim, M. Kang, "Protection of Identity Information in Cloud Computing without Trusted Third Party," Third International Workshop on Dependable Network Computing and Mobile Systems (DNCMS) in conjunction with 29th IEEE Symposium on Reliable Distributed System (SRDS) 2010

14. P. Angin, B. Bhargava, R. Ranchal, N. Singh, L. Lilien, L.B. Othmane, "A User-Centric Approach for Privacy and Identity Management in Cloud Computing," 29th IEEE Symposium on Reliable Distributed System (SRDS) 2010

15. H. Khandelwal, *et al.*, "Cloud Monitoring Framework," Purdue University. Dec 2010.

