Data Anonymization Approaches for Data Sets Using Map Reduce on Cloud: A Survey

Veena[1], Devidas[2]

[1]NMAM Institute of Technology, Nitte, Karnataka, India

Abstract: ***In this information age, huge amounts of data are collected and mined every day. The process of data publication is becoming larger and complex day by day. Cloud computing is the most popular model for supporting large and complex data, most organizations are moving towards to reduce their cost and elasticity features. However cloud computing has potential risk and vulnerabilities. One of major problem in moving to cloud computing is its security and privacy concerns. Cloud computing provides powerful and economical infrastructural resources for cloud users to handle ever increasing data sets in big data applications. However, processing or sharing privacy-sensitive data sets on cloud probably leads to privacy concerns because of multi-tenancy system. Data encryption and anonymization is two widely-adopted ways to combat privacy breach. The encryption is not suitable for data that are processed and shared frequently and the anonymizing big data and manages anonymized data sets are still challenges for traditional anonymization approaches. Thus, various proposals have been designed in a cloud computing for privacy preserving in data publishing. In this paper, we survey the current existing techniques, and analyze the advantage and disadvantage ofthese approaches.***

Keywords: Anonymization, Cloud computing, Hadoop, Privacy preservation, Top Down Specialization.

1. Introduction

Information sharing has become part of the routine activity of many individuals, companies, organizations, and government agencies. Such Information sharing is subject to constraints imposed by privacy of individuals or data subjects as well as data confidentiality of institutions or data providers. With wide adoption of online cloud services the privacy concern about processing and sharing of sensitive personal information is increasing. To reduce these risks various proposals have been designed for privacy preserving in data publishing. In this survey we will briefly review recent research on data privacy preservation and privacy protection in MapReduce and cloud computing environments, and survey current existing techniques, and summarize the advantage and disadvantage of these approaches.

Existing technical approaches for preserving the privacy of data sets stored in cloud mainly include encryption and anonymization. First, encrypting data sets, a straight forward and effective approach. However, processing on encrypted data efficiently is quite challenging task. Although recent progress has been made in homomorphic encryption which theoretically is valid but experimentally is very expensive due their inefficiency. Secondly partial information of data sets, e.g., aggregate information, is required to expose to data users in most cloud applications like data mining and analytics. In such cases, data sets are anonymized rather than encrypted to ensure both data utility and privacy preserving.

Cloud systems provides massive computation power and storage capacity that enable users to deploy applications without infrastructure investment. Because of its salient features, cloud is promising for users to handle the big data processing pipeline with its elastic and economical infrastructural resources. For instance, MapReduce is widely adopted large-scale data processing paradigm, which is more flexible, scalable, and cost-effective computation for big data processing. A typical example is the Amazon Elastic MapReduce service.

2. Related work

We briefly review recent research on data privacy preservation and privacy protection in Map Reduce and cloud computing environments.

LeFevre et al. [2] addressed the scalability problem of anonymization algorithms via introducing scalable decision trees and sampling techniques. Iwuchukwu et al. [14] proposed an R-tree index-based approach by building a spatial index over data sets, achieving high efficiency. However, the above approaches aim at multidimensional generalization, thereby failing to work in the Top- Down Specialization (TDS) approach. Fung et al. [15, 11] proposed the Centralized TDS approach that produces anonymous data sets without the data exploration problem. A data structure Taxonomy Indexed PartitionS (TIPS) is exploited to improve the efficiency of TDS. But the approach is centralized, leading to its inadequacy in handling large-scale data sets [1].

Several distributed algorithms are proposed to preserve privacy. Jiang et al. [17] and Mohammed et al. [3] proposed distributed algorithms to anonymize vertically partitioned data from different data sources without disclosing privacy information from one party to another. Jurczyk et al. [13] and Mohammed et al. [15] proposed distributed algorithms to anonymize horizontally partitioned data sets retained by multiple holders. However, the above distributed algorithms mainly aim at securely integrating and anonymizing multiple data sources. Our research mainly focuses on the scalability issue of TDS anonymization, and is therefore orthogonal and complementary to them.

3. A survey on privacy preserving Approaches in data publishing

The issue is how to publish the data in such a way that the privacy of individuals can be preserve. Various proposals have been designed for privacy preserving.

3.1 Data Anonymization concepts and techniques

Anonymization is a technique that can use to increase the security of the data while still allowing the data to be analyzed or used. Data anonymization is the process of changing the data that will be used or published in way that prevents the identification of key information. Data anonymization is a technique that will not take away the original field layout (position, size and data type) of the data being anonymized, so the data will still look realistic in test data environments. Anonymous technology is mainly used for database privacy, location privacy, and trajectory privacy, but we propose applying it cloud storage privacy. Using data anonymization, key pieces of confidential data are obscured in a way that maintains data privacy. The data can still be processed to gain useful information. Anonymization data can be stored in a cloud and processed without concern that other individuals may capture the data. Later, the results can be collected and mapped to the original data in a secure area. Several formal of security can help improve data anonymization including K-anonymity, L-diversity anonymous, and T-closeness anonymous.

3.2 K-anonymity

L. Sweeney [10] has proposed the concept of -anonymity. Publishing data about individuals without revealing sensitive information about them is an important problem. In recent years, a new definition of privacy called k-anonymity has gained popularity. The goal is to make each record indistinguishable from a defined number (k) other records, if attempts are made to identify the record.

K-anonymity guarantees that each sensitive attribute is hidden in the scale of k groups. This means that the probability of recognizing the individual does not exceed 1/k. The level of privacy depends on the size of k. The statistical characteristics of the data are retained as much as possible; however, k-anonymity is not only applicable to sensitive data. An attacker could mount a consistency attack or background-knowledge attack to confirm a link between sensitive data and personal data. This would constitute a breach of privacy. The extensive study resolved some shortcomings of *k*-anonymity model as listed below.

1) It can't resist a kind of attack, which is assuming that the attacker has background knowledge to rule out some possible values in a sensitive attribute for the targeted victim. That is, *k*-anonymity does not guarantee privacy against attackers using background knowledge. It is also susceptible to homogeneity attack. An attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attributes. Thus some stronger definitions of privacy are generated, such as ℓ-Diversity.
2) It protects identification information. However, it does not protect sensitive relationships in a data set.
3) Although the existing *k*-anonymity property protects against identity disclosure, it fails to protect against attribute disclosure.
4) It is suitable only for categorical sensitive attributes. However, if we apply them directly to numerical sensitive attributes (e.g., salary) may result in undesirable information leakage.
5) It does not take into account personal anonymity requirements and a *k*-anonymity table may lose considerable information from the micro data which is a valuable source of information for the allocation of public funds, medical research, and trend analysis.

3.3 L-diversity

L-diversity [11] anonymous ensures that each group's sensitive attributes have at least L different values. This means that an attack has a maximum probability of 1/L of recognizing a user's sensitive information. T-closeness anonymous is based on L-diversity anonymous. L-Diversity provides privacy preserving even when the data publisher does not know what kind of knowledge is possessed by the adversary. The main idea of L-diversity is the requirement that the values of the sensitive attributes are well-represented in each group. The *k*-anonymity algorithms can be adapted to compute L-diverse tables. L-Diversity resolved the shortcoming 1 of *k*-anonymity model.

3.4 T-closeness

T-closeness [12] anonymous, the distribution of the sensitive attribute is taken into account, and the distribution differences between sensitive properties and values in groups does not exceed T. An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

4. Evaluation

Authors	*Concept*	*Advantage*	*Disadvantage*
Privacy-Preserving Data Publishing by BENJAMIN C. M. FUNG,KE WANG,RUI CHEN,PHILIP S. YU	It provides methods and tools for publishing useful information while preserving data privacy	PPDP has received a great deal of attention in the database and data mining research communities.	1. Degradation of data/service quality. 2. Loss of valuable information 3. Increased costs. 4. Increased complexity.
"Workload- Aware Anonymization Techniques for Large Scale Datasets," [6] .KRISTEN Lefevre, DAVID J. DeWITT , Raghu Ramakrishnan	Anonymization algorithms that incorporate a target class of workloads, consisting of one or more data mining tasks as well as selection predicates and the datasets much larger than main memory.	1. High efficiency 2. Leads to high-quality data. 3. More flexible.	1. Failing to work in the Top-Down Specialization (TDS) approach. 2. It does not address the complementary problem of reasoning about disclosure across multiple releases. 3. Fail to solve preserving privacy for

			multiple datasets.
"Privacy- Preserving Data Publishing for Cluster Analysis,"[7] B. Fung, K. Wang, L. Wang, P.C.K. Hung	Preventing the privacy threats caused by sensitive record linkage and the framework to secure data sharing for the purpose of cluster analysis.	1. Preserves both individual privacy and information usefulness for cluster analysis. 2. Avoids over-masking and improves the cluster quality. Preventing the privacy threats caused by sensitive record linkage.	1. Inadequacy in handling large-scale data sets. 2. Reconstruction process naturally leads to some loss of information
"A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-Effective Privacy Preserving of Intermediate Datasets in Cloud,"[8] R. Urgaonkar, U. Kozat, K. Igarashi, M. J. Neely	Upper bound privacy leakage constraint-based approach to identify which intermediate data sets need to be encrypted and which do not.	1. Privacy-preserving cost of intermediate data sets can be significantly reduced.	1. Highly complicated. 2. Processing on data sets efficiently will be quite a challenging task. 3. Performing general operations on encrypted data sets are still quite challenging
"Airavat: Security and Privacy for Mapreduce,"[9] Roy I, Setty STV, Kilzer A , Shmatikov V , Witchel E	Airavat enables the execution of trusted and untrusted MapReduce computations on sensitive data, while assuring comprehensive enforcement of data providers privacy policies.	1. Provides end-to-end confidentiality, integrity, and privacy using a combination of mandatory access control and differential privacy. 2. Enable large-scale computation on data items that originate from different sources and belong to different owners.	1. The results produced in this system are mixed with certain noise. 2. Airavat cannot confine every computation performed by untrusted code. 3. Does not protect sensitive data from the public cloud.
"PRISM-Privacy-Preserving Search in Mapreduce," [10] .Blass E-O, Pietro RD , Molva R, Önen M	Privacy-preserving search scheme suited for cloud computing and provides storage and query privacy while introducing only limited overhead and designed to leverage parallelism and efficiency of the MapReduce paradigm.	1. Assures data privacy confidentiality and query confidentiality 2. Meets cloud computing efficiency requirements. 3. Preserves privacy in the face of potentially malicious cloud providers.	1. Difficult to secure public clouds 2. Ccause a potential privacy breach 3. Low performance
"The Hybrex Model for Confidentiality and Privacy in Cloud Computing," [11] Ko SY , Jeon K, Morales R	The HybrEx model provides a seamless way for an organization to utilize their own infrastructure for sensitive, private data and computation, while integrating public clouds for nonsensitive, public data and computation.	1. The ability to add more computing and storage resources from public clouds to a private cloud without the concerns for confidentiality and privacy. 2. Provides the confidentiality and privacy guarantees	1. Hard to scale. 2. It do not deal with higher level query processing or optimization issues
"Sedic: Privacy-Aware Data Intensive Computing on Hybrid Clouds," [12] Zhang K, Zhou X, Chen Y, Wang X, Ruan Y	1.Sedic is designed to protect data privacy during map-reduce operations	1. Effectively protect sensitive user data 2. High privacy assurance 3. Ease to use. 4. Fully preserved the scalability	1. Lack of scalability over big data. 2. The sensitivity of data is required be labeled in advance.

References

[1] B.C.M. Fung, K. Wang, R. Chen and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Comput. Surv., vol. 42, no. 4, pp. 1-53, 2010.

[2] K. LeFevre, D.J. DeWitt and R. Ramakrishnan, "Workload- Aware Anonymization Techniques for Large-Scale Datasets," ACM Trans. Database Syst., vol. 33, no. 3, pp. 1-47, 2008.

[3] B. Fung, K. Wang, L. Wang and P.C.K. Hung, "Privacy-Preserving Data Publishing for Cluster Analysis," Data Knowl.Eng., Vol.68, no.6, pp. 552-575, 2009.

[4] X. Zhang, Chang Liu, S. Nepal, S. Pandey and J. Chen, "A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-Effective Privacy Preserving of Intermediate Datasets in Cloud," IEEE Trans. Parallel Distrib. Syst., In Press, 2012.

[5] Roy I, Setty STV, Kilzer A, Shmatikov V, Witchel E, "Airavat: Security and Privacy for Mapreduce," Proceedings of 7th USENIX Conference on Networked Systems Design and Implementation (NSDI'10), 2010; 297–312.

[6] Ko SY, Jeon K, Morales R, "The Hybrex Model for Confidentiality and Privacy in Cloud Computing," Proceedings of the 3rd USENIX Conference on Hot Topics in Cloud Computing (HotCloud'11), 2011; Article 8.

[7] Zhang K, Zhou X, Chen Y, Wang X, Ruan Y. Sedic: "Privacy-Aware Data Intensive Computing on Hybrid Clouds," Proceedings of 18th ACM Conference on Computer and Communications Security (CCS'11), 2011; 515–526.

[8] Wei W, Juan D, Ting Y, Xiaohui G, "Securemr A Service Integrity Assurance Framework for Mapreduce," Proceedings of Annual Computer Security Applications Conference (ACSAC '09), 2009; 73–82.
[9] Blass E-O, Pietro RD, Molva R, Önen M, "PRISM-Privacy-Preserving Search in Mapreduce," Proceedings of the 12th International Conference on Privacy Enhancing Technologies (PETS'12), 2012; 180–200.
[10] L. Sweeney, "k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledgebased Systems, 2002, pp. 557-570.
[11] A.Machanavajjhala, J.Gehrke, and D.Kifer, et al, "ℓ-diversity: Privacy beyond k-anonymity", In Proc. of ICDE, Apr.2006.
[12] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-anonymity and l-Diversity", In Proc. of ICDE, 2007, pp. 106-115.
[13] Dean J, Ghemawat S. Mapreduce: a flexible data processing tool. Communications of the ACM 2010; 53(1):72–77. DOI: 10.1145/1629175.1629198.
[14] P.JurczykandL.Xiong,"DistributedAnonymization:AchievingPrivacyforBothData Subjects and Data Providers, "*Data and Applications Security XXIII(DBSec'09)*,pp. 191-207,2009.
[15] T. IwuchukwuandJ.F. Naughton, "K-Anonymization as SpatialIndexing: Toward ScalableandIncremental Anonymization," *Proc. 33rdInt'lConf. VeryLarge DataBases (VLDB'07)*, pp.746-757, 2007.
[16] N.Mohammed, B.Fung,P.C. K.Hungand C.K.Lee, "Centralized and Distributed Anonymization for High-Dimensional Healthcare Data," *ACM Trans. Knowl.Discov.Data*,vol.4,no.4,article18,2010.
[17] N. Mohammed, B.C. Fungand M. Debbabi, "Anonymity Meets Game Theory: Secure Data Integration with MaliciousParticipants,"*VLDBJ.*,vol.20,no.4,pp.567-588,20
[18] W.JiangandC.Clifton,"ASecureDistributedFrameworkforAchieving*k*-anonymity,"*VLDBJ.*,vol.15,no.4,pp.316-333,2006.

Author Profile

Veena has received the B.E degree in Computer Science and Engineering from Visvesvaraya Technological University in 2011. She is currently pursuing her M.Tech degree in Computer Networks under the same University.

Devidas has received the B.E degree in Information Science and Engineering from Visvesvaraya Technological University in 2006 & M.Tech in Computer Science in 2010 under same university. He is currently working as Assistant Professor in Information Science and Engineering Department in NMAM Institute of Technology, Nitte.