Data Anonamise Of Top Down Glide Slope Using Map Reduce On Cloud Computing

Peddaboina Yamuna [1], Yata Rambabu[2]

[1,2] Assistant professor,
[1]Department of [1]computer science and engineering, [2]Department of science and humanities,
[1]Samskruti College Of Engineering And Technology, Ghatkesar, Hyderabad.
[2]SNIST, Ghatkesar, Hyderabad

Abstract:
In this paper propose a scale-up two-phase top-down specialization (TDS) attempt to anonymize large-scale data sets using the Map Reduce framework on cloud. In both phases of our approach, we deliberately design a group of sophisticated Map Reduce jobs to concretely accomplish the specialization computation in a highly scalable way. Observation pay grade results exhibit that with this approach, the scalability and efficiency of

1. INTRODUCTION

Cloud computing provides monolithic computation power and storage capacity via utilizing a large number of good computers together, enabling users to deploy applications cost effectively without heavy infrastructure investment. Cloud users can reduce huge upfront investment of IT infrastructure, and concentrate on their own core business. However, many potential customers are still hesitating to take advantage of cloud due to privacy and security concerns. Privacy is one of the most concerned issues in cloud computing, and the concern exacerbates in the context of cloud computing although some privacy issues are not new. Personal data like electronic health records and financial transaction records are usually deemed extremely sensitive although these data can offer significant human benefits if they are analyzed and mined by organizations such as disease research centers. Microsoft Health Vault is an online cloud health service, aggregates data from users and shares the data with research institutes. Data privacy can be divulged with less effort by malicious cloud users or providers because of the failures of some traditional privacy protection measures on cloud. This can bringTDS can be significantly improved over existing approaches large number of cloud services requires users to share private data like electronic health records for data analysis or mining, bringing concealment concerns. Anonymize data sets via generalization to satisfy certain privacy requirements such as k-anonymity is a widely used category of concealment conserve techniques. Considerable economic loss or severe social reputation impairment to data owners. Hence, data privacy issues need to be addressed urgently before data sets are analyzed or shared on cloud. Data anonymize has been extensively studied and widely accepted for data privacy preservation in non inter active data publishing and sharing arenas. Data anonymize refers to hiding identity and/or sensible data for owners of data records. Then, the privacy of an individual can be effectively preserved while certain total information is exposed to data -users for versatile analysis and mining. The scale of data sets that need anonym zing in some cloud applications increases tremendously in accordance with the cloud computing and Big Data trends. Data sets have become so large that anonymize such data sets is becoming a considerable take exception for traditional anonymize algorithms. Large-scale data processing frameworks like Map Reduce have been merged with cloud to provide powerful computation capability for applications.

2. LITERATURE SURVEY

The practice in data publication relies mainly on policies and guidelines as to what types of data can be published and on agreements on the use of published data. This approach alone may lead to excessive data deformation or lacking protection. Privacy-preserving data publishing provides methods and tools for publishing useful information while preserving data privacy. Privacy- preserving data publishing has received considerable tending in research communities, and many approaches have been proposed for various data publishing arenas. It systematically summarized and evaluates various approaches to Privacy-preserving data publishing provides study the challenges in practical data publishing, clarify the differences and requirements that distinguish Privacy-preserving data publishing provides methods from other related problems, and propose future research directions.A task of the outmost importance is to formulate methods and tools for publishing data in a more nonaggressive environment, so that the published data remains practically useful while individual privacy is prevented. This undertaking is called privacy-preserving data publishing (PPDP). This field is still rapidly developing, it is a good time to discuss the assumptions and desirable properties for PPDP, clarify the differences and requirements that distinguish PPDP from other related problems, and systematically summate and valuate various approaches to PPDP. The collection of digital information by governments, corporations, and individuals has created wonderful opportunities for knowledge and information-based determination making. Driven by reciprocal benefits, or by regulations that require certain data to be published, there is a demand for the exchange and publication of data among various parties. Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy. The data collection phase, the data publisher collects data from record owners. The data publishing phase, the data publisher releases the collected data to a data miner or to the public, called the data recipient, who will then conduct data mining on the published data. Data mining has a broad sense, not necessarily restricted to pattern mining or model building. Medical center is the data recipient. . Data mining has a broad sense, not necessarily restricted to pattern mining or model building. Medical center is the data recipient. The data mining conducted at the medical center could be anything from a simple count of the number of men with diabetes to a sophisticated cluster analysis.There are two models of data publishers. The un trusted model, the data publisher is not trusted and may attempt to identify sensitive information from record owners. Various cryptographic solutions anonymous communications and statistical methods were proposed to collect records anonymously from their owners without revealing the owners' identity. The trusted model, the data publisher is trustworthy and record owners are willing to provide their personal information to the data publisher; however, the trust is not transitive to the data recipient. The trusted model of data publishers and consider privacy issues in the data publishing phase. The data recipient could be an attacker. PPDP has one assumption is that the data recipient could also be an attacker.

EXISTING SYSTEM

- Lefebvre et al. addressed the expandability problem of anonymize algorithms via introducing scalable decision trees and sampling techniques.
- Iwuchukwu and Naught on proposed an R-tree index-based approach by building a spatial index over data sets, achieving high efficiency. This approaches aim at multidimensional synopsis, thereby impuissance to work in the TDS approach.
- Fung et al. proposed the 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.
- Jiang and Clifton and Mohammed et al. proposed distributed algorithms to anonymize vertically striped data from various data origins without disclosing privacy entropy from one party to another.
- Jurczyk and Xing and Mohammed et al. proposed distributed algorithms anonymize horizontally striped data sets retained by multiple holders. As to Map Reduce-relevant privacy protection.
- Roy et al. studied the data concealment problem caused by Map Reduce and presented a system named Ararat integrating mandatory access control with derivative privacy.
- Further, Zhang et al. leveraged Map Reduce to automatically partition a computing job in terms of data security levels, protecting data privacy in hybrid cloud.

Disadvantages: The scalability problem of existing TDS approaches when handling large scale data sets on cloud is very complicated.

ISSUES IN ANONYMIZATION

- No real Dataset Data owner won't publish confidential dataset. Inconsistent Quasi Identifiers
- No standard metrics for quantifying risk complicated models. Risk depends on many Factors, e.g. dataset, technical skill,
- Availability of background data. Utility depends on use case (but

- Which is unknown when collecting data?)
- No standard model of Adversary "Mildly motivated adversary y" vs. "highly motivated Adversary"

PROPOSED SYSTEM

In this paper propose a scalable two-phase top-down specialization (TDS) approach to anonymize large-scale entropy sets using the Map Reduce fabric on cloud.

- .In both phases of this approach, it deliberately design a group of innovative Map Reduce jobs to concretely accomplish the specialization computation in a highly scalable way.To make full use of the analogue effectiveness of Map Reduce on cloud, specializations needed in an anonymization process are divide into two phases.

○ In the first one, original data sets are partitioned into a group of smaller data sets, and these datasets are anonym zed in parallel, producing intermediate results.

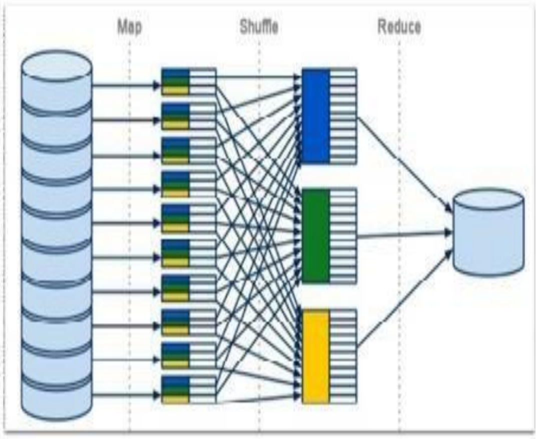
○ In the second one, the average results are merged into one, and further anonym zed to achieve consistent k-anonymous datasets. It leverage Map Reduce to execute the concrete computation in both phases.

○ A group of Map Reduce jobs is deliberately designed and coordinated to perform specializations on data sets collaboratively. It evaluates this approach by conducting experiments on real-world datasets. Experimental results demonstrate that with this approach, the scalability and efficiency of TDS can be improved significantly over existing approaches.

○ **Advantages:** The scalability and efficiency of TDS can be improved significantly over existing approaches.

MAPREDUCE

Large-scale data processing frameworks like Map Reduce is used to provide powerful computation capability for applications. So it is promising to adopt such frameworks to address the scalability problem of Anonym zing large-scale data for privacy preservation. In our research, we leverage Map Reduce, a widely adopted parallel data processing modeling, to address the expandability problem of Top-Down Specialization approach for large scale data anonymization. The prime open source implementation of Map Reduce is Apache's Hardtop. The core concept of Map Reduce in Hardtop is that input may be split into logical chunks, and each chunk may initially processed independently, by a map task as shown in Fig.1. The result of these individualprocessing chunks can be physically partitioned into distinct sets, which are then sorted.



CONCLUSION

In this paper, we have studid the expandability problem of large-scale data anonymization by TDS, and proposed a highly expandable two-phase TDS approach using Map Reduce on cloud. Data sets are divided and anonym zed in parallel in the first phase, producing average results. Then, the average results are merged and further anonym zed to produce consistent k-anonymous data sets in the second phase. We have creatively applied Map Reduce on concretely accomplished the specialization A highly scalable way. Experimental results on real world data sets have demonstrated that with our approach, the scalability and efficiency of TDS are improved significantly over existing approaches and anonym zed in parallel in the first phase, producing intermediate results.

REFERENCES:

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

2) X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen, "A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud," IEEE Trans. Parallel and Distributed Systems, to be published, 2012.

3) K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan, "Sedic: Privacy-Aware Data Intensive Computing on Hybrid Clouds," Proc. 18th ACM Conf. Computer and Comm. Security (CCS '11), pp. 515-526, 2011.

4) Roy, S.T.V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel, "Airavat: Security and Privacy for Map reduce," Proc. Seventh USENIX Conf. Networked Systems Design and Implementation (NSDI'10), pp. 297-312, 2010.

5) N. Mohammed, B. Fung, P.C.K. Hung, and C.K. Lee, "Centralized and Distributed Anonymization for High-Dimensional Healthcare Data," ACM Trans. Knowledge Discovery from Data, vol. 4, no. 4, Article 18, 2010.

6) N. Mohammed, B.C. Fung, and M. Debbabi, "Anonymity Meets Game Theory: Secure Data Integration with Malicious Participants," VLDB J., vol. 20, no. 4, pp. 567-588, 2011.

Author Profile

PEDDABOINA YAMUNA,have complted M.Tech in CSE from samskruti College of engineering and Technology ,Affiliated to JNTU ,Hyderabad. Presently working as Assistant Professor in CSE dept. in Samskruti College of engineering and Technology,Ghatkesar ,Hyderabad.

YATA RAMBABU,have completed M.TECH in CSE from holymary institute of technology and science,affiliated to JNTU Hyderabad.presently working assistant professor in sience and humanities dept in SNIST,Ghatkesar,Hyder abad