International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

Data Anonymization Using Map Reduce On Cloud

[1]Shivaprasad Goud & [2]K.Ganeshwar
[1]s V College Of Engineering
Assistant Professor S V College Of Engineering

ABSTRACT

A large number of cloud services require users to share private data like electronic health records for data analysis or mining, bringing privacy concerns. Anonymizing data sets via generalization to satisfy certain privacy requirements such as k-anonymity is a widely used category of privacy preserving techniques. At present, the scale of data in many cloud applications increases tremendously in accordance with the Big Data trend, thereby making it a challenge for commonly used software tools to capture, manage, and process such large-scale data within a tolerable elapsed time. As a result, it is a challenge for existing anonymization approaches to achieve privacy preservation on privacy-sensitive large-scale data sets due to their insufficiency of scalability. In this paper, we propose a scalable two-phase top-down specialization (TDS) approach 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 innovative Map Reduce jobs to concretely accomplish the specialization computation in a highly scalable way. Experimental evaluation results demonstrate that with our approach, the scalability and efficiency of TDS can be significantly improved over existing approaches.

KEYWORDS: Data anonymization; top-down specialization; MapReduce; cloud; privacy preservation

INTRODUCTION

Cloud computing is one of the most pre-dominant paradigm in recent trends for computing and storing purposes. Datasecurity and privacy of data is one of the major concern in the cloud computing.Dataanonymization has been extensivelystudied and widely adopted method for privacy preserving in data publishing and sharing methods. Data anonymizationispreventing showing up of sensitive data for owner's data record to mitigate unidentified Risk. The privacy of individualcan be adequately maintained while some aggregate information is shared to data user for data analysis and data mining.The proposed method is generalized method data anonymization using Map Reduce on cloud. Here we Two Phase TopDown specialization. In First phase, original data set is partitioned into group of smaller dataset and they are anonymizedand intermediate result is produced. In second phase, intermediate result first is further anonymized to achieve persistentdata set. And the data is presented in generalized form using Generalized Approach.A highly scalable two-phase TDS approach for data anonymization based on Map Reduce on cloud. To make use of theparallel capability of MapReduce on cloud, classification required in an anonymization process is split into two phases. Inthe first one, original datasets are partitioned into a group of small datasets, and those datasets are anonymized in parallel,creating intermediately results. In the second one, the intermediate results are aggregated into one, and further anonymized to achieve consistent k-anonymous data sets. It leveragesMapReduce to accomplish the concretecomputation in both phases. A group of MapReducejobs are deliberately designed and coordinated to performspecializations on data sets collaboratively. It evaluates the approach by

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

conducting experiments on real-world data sets.Experimental results show that with the approach, the scalability and efficiency of TDS can beimproved. It evaluates theapproach by conducting experiments on real-world data sets. Experimental results demonstrate that with the approach, thescalability and efficiency of TDS can be improved significantly over existing approaches. Themajor contributions of deliberately design a group of innovative MapReduce jobs to concretely accomplish the specializations in a highlyscalable fashion. Secondly, it propose a two-phase TDS approach to gain high scalability via allowing specializations toe conducted.

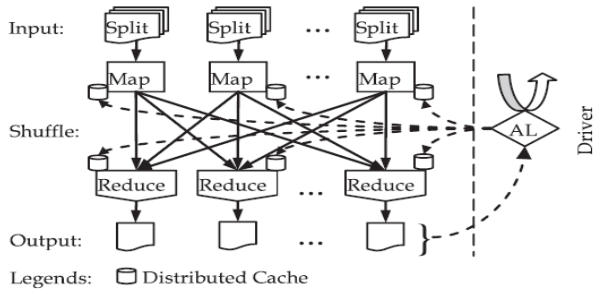

Fig 1:System architecture

In above figure software by way of a facility model provides three different functionalities such as Users, Providers and public providers so we are going to discuss about it mobile users will provide data outside to the private clouds which will store the filtered results on the public cloud. This model is supporting execution of private mechanisms and shifting storage data to the clouds which leave users with light tasks.

1.1 Cloud Computing

Capacity is picking up prevalence as of late. In big commercialsituations, we see the ascensionrequiredafterward for datasubcontracting, which assistances with the key administration of corporate information? It is additionally utilized as a center innovation behind numerous online administrations for individual applications. Nowadays, it is anything but difficult to seek free records for email, photograph collection, and record sharing and/or remote access, with capacity measure more than 25GB (or a couple of dollars for more than 1TB). Together with the present remote innovation, clients can get to the greater part of their records and messages by a cell telephone in any edge of the world. Information from distinctive customers can be facilitated on discrete virtual machines (VMs) however dwell on a solitary physical machine. Information in an objective VM could be stolen by instantiating another VM co-inhabitant with the objective one. Concerning of papers, there are a development of cryptographic tactics which go similarly as approvinganstranger reviewer to checked the convenience of records in the interest of the data proprietor starved of freeingwhatever about the information, or without trading off the information proprietor's obscurity. Correspondingly, cloud customers probably won't hold the strong conviction that the cloud server is making a good indicating similarly as classifiedness. A cryptographic arrangement, with demonstrated security depended on number-theoretic suppositions is more attractive, at whatever point the client is not consummately content with believing the security of the VM or the trustworthiness of the specialized staff. These customers are influenced to encode their data with their own specific keys before exchanging them to the server. Data sharing is a crucial helpfulness in dispersed stockpiling. For example, bloggers can let their sidekicks see a subset of their private pictures; an endeavor may give her delegates access to a fragment of fragile data. The testing issue is the means by which to viably share encoded information. Clients ought to have the

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

capacity to delegate the entrance privileges of the sharing information to others so they can get to these information from the server specifically. Expect that Alice puts all her private photographs on Drop box, and she wouldn't like to open her photographs to everybody. Because of different information spillage plausibility Alice can't feel assuaged by simply depending on the security insurance components gave by Drop box, so she encodes all the photographs utilizing her own keys before transferring. One day, Alice's companion, Bob, requests that her share the photographs assumed control over every one of these years Which Bob showed up.Alice can then utilize the offer capacity of Drop box; however the issue now is the means by which to delegate the unscrambling rights for these photographs to Bob. A conceivable alternative Alice can pick is to safely send Bob the mystery keys included. Actually, there are two great courses for her under the customary encryption ideal model:

1. Alice scrambles all records with a solitary encryption key and gives Bob the comparing mystery key straightforwardly.

2. Alice scrambles records with particular keys and sends Bob the relating mystery keys.

1.2 Problem Definition

This task points that productivity disadvantages of the most existing ABE plans is that unscrambling is extravagant for asset restricted gadgets because of blending operations, and the quantity of matching operations needed to unscramble a figure content develops with the intricacy of the entrance arrangement.
The above perception spurs us to study ABE with unquestionable outsourced unscrambling in this theory work. Here stressed that an ABE plan with secure outsourced unscrambling does not so much ensure certainty.

2. SYSTEM ANALYSIS

2.1 Feasibility Study

Feasibility study is ansignificantstage in the software development procedure. It empowers the designer to have an appraisal of the item being produced. It alludes to the possibility investigation of the item regarding results of the item, operational utilization and specialized backing needed for executing it. Feasibility study ought to be performed on the premise of different criteria and parameters. The different feasibility studies are:

Operational feasibility

Technical feasibility

Economic feasibility

2.1.1 Operational Feasibility

The feature of learning is to check the level of receiving of the scheme by the customer. This comprises the procedure of working out the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system. Here instead of sending multiple keys we are aggregating the multiple keys into a single key. By aggregating the keys it is very easy to manage keys.

2.1.2 Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any framework

created must not have popularity on the accessible specialized assets. This will prompt levels of popularity on the accessible specialized assets. This will prompt levels of popularity being set on the customer. The created framework must have an unobtrusive necessity, as just insignificant or invalid changes are needed for executing this framework. The technology we used is java. Java is robust and it follows that write once run anywhere. Java provides security it is flat form independent.

2.1.3 Economic Feasibility

This study is done to check the monetary effect that the framework will have on the association. The measure of store that the organization can fill the innovative work of the framework is restricted. The consumptions must be advocated.

The admin uploads the files. When admin is uploading files to user he is encrypting the files. And he shared encrypted files to the users. The encrypted key is send to the user's mail and when they want to download the files and he has to enter the key.

CONCLUSION

Privacy preserving data analysis and data publishing are becoming serious problems in today's ongoing world. That's why different approaches of data anonymization techniques are proposed. To the best of our knowledge, TDS approach using MapReduce are applied on cloud to data anonymization and deliberately designed a group of innovative Map Reduce jobs to concretely accomplish the specialization computation in a highly scalable way.

REFERENCES

[1] S. Chaudhuri, "What Next?: A Half-Dozen Data Management Research Goals for Big Data and the Cloud," Proc. 31st Symp.Principles of Database Systems (PODS '12), pp. 1-4, 2012.

[2] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A.Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A View of Cloud Computing," Comm. ACM, vol. 53, no. 4, pp. 50-58, 2010.

[3] L. Wang, J. Zhan, W. Shi, and Y. Liang, "In Cloud, Can ScientificCommunities Benefit from the Economies of Scale?," IEEE Trans.Parallel and Distributed Systems, vol. 23, no. 2, pp.296-303, Feb.2012.

[4] H. Takabi, J.B.D. Joshi, and G. Ahn, "Security and Privacy Challenges in Cloud Computing Environments," IEEE Security and Privacy, vol. 8, no. 6, pp. 24-31, Nov. 2010.

[5] D. Zissis and D. Lekkas, "Addressing Cloud Computing Security Issues," Future Generation Computer Systems, vol. 28, no. 3, pp. 583-592, 2011.

[6] 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.

[7] L. Hsiao-Ying and W.G. Tzeng, "A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 6, pp. 995-1003, 2012.

[8] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc.IEEE INFOCOM, pp. 829-837, 2011.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

shivaprasad goud

S V COLLEGE OF ENGINEERING

K.GANESHWAR

Assistant Professor

S V COLLEGE OF ENGINEERING