ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

An Effective Specialization Approach for Data Anonymization Using Map Reduce on Cloud

K.Sunil Kumar
M.Tech Student,
Dept of Software Engineering,
VNR Vignana Jyothi Institute of Engineering & Technology, India.

P.Venkateswara Rao
Assistant Professor,
Dept of Computer Science & Engineering,
VNR Vignana Jyothi Institute of Engineering & Technology, India.

ABSTRACT:

Data privacy preservation is one of the most dispersed concerns on the modern business. Data sequestration concern urgency to be consigned crucially before data sets are communal on a cloud. Data anonymization refers to as concealing complicated data for heirs of data records. In this paper interrogate the complications of big data anonymization for privacy conservation from the context of scalability and time factor etc. At present, the extent of data in many cloud applications boosts exceedingly in accord with the big data trend. Here introduced a scalable Two Phase Top-Down Specialization (TPTDS) accession to anonymize substantial data sets using the Map Reduce framework on a cloud. For the data anonymization, 45,222 records of adult's report with 15 attribute aspect values were taken as the input big data.

With the assistance of multifaceted anonymization on Map Reduce framework, here implement the proposed Two-Phase Top-Down Specialization anonymization algorithm on Hadoop will boost the productivity/competence of the big data processing system. In both aspect of the access, consciously study multidimensional Map Reduce jobs to categorically achieve the specialization computation in an extremely scalable way. Data sets are generalized in a top-down demeanor and the improved results shown in multidimensional Map Reduce framework by analysing the one-dimensional Map Reduce framework anonymization job. The anonymization was implemented with specialization procedure on the taxonomy tree.

The analysis exhibit that the solutions can significantly advance the scalability and competence of big data privacy preservation compared to existing approaches. This task has enormous applications to both public and private sectors that stake data to the society.

INTRODUCTION:

Data secure concerned issues in cloud computing is very important to find the solution, and the concern aggravates in the background of cloud computing although some camouflage issues are not new. Personal data like electronic health reports and monetary transaction transcripts are usually presumed immensely sensitive although the above mentioned data can bid compelling human benefits if they are evaluated and mined by organizations such as health research centers. For instance, Microsoft HealthVault, an online cloud health service, accumulated data from users and stake the data with research institutes.

Data confidentiality can be disclosed with less effort by malevolent cloud users or providers because of the failures of some traditional privacy protection measures consignment on the cloud. This can bring ample monetary loss or severe social prominence to data owners. Hence, data privacy issues need to be addressed urgently before data sets are analyzed or shared on the cloud. Data anonymization has been extensively studied and widely adopted for data privacy preservation in no bilateral data circulating and distribution scenarios. Data anonymization refers to hiding identity and/or sensitive data for holders of data records.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Then, the solitude of an individual can be adequately retained while certain accumulated information is disclosed to data users for varied analysis and mining. A variety of anonymization algorithms with different anonymization operations has been proposed. However, the scale of data sets that need anonymizing in some cloud applications boost exceedingly in consonance with the cloud computing and Big Data trends. Data sets have become so large that anonymizing such datasets are becoming an ample challenge for conventional anonymization algorithms. The analysts have started to examine the scalability problem of large-scale data anonymization.

Large-scale data processing frameworks like MapReduce have been integrated with the cloud to provide vigorous computation proficiency for applications. So, it is auspicious to adopt such frameworks to address the scalability problem of anonymizing large-scale data for privacy preservation. In our research, we influence MapReduce, a widely ratify or take up aligned data processing framework, to address the scalability problem of the top-down specialization (TDS) approach for extensive data anonymization. The TDS approach, contributing a good pact between data expediency and data consistency, is widely applied for data anonymization.

Most TDS algorithms are centralized, emanating in their paucity in handling extensive datasets. Although some distributed algorithms are proposed, for the most part they locus on secure or stable anonymization of datasets from numerous parties, rather than the scalability facet. As the MapReduce computation paradigm is comparatively simple, it is still a challenge to design proper MapReduce jobs for TDS. Here I propose a highly scalable two-phase TDS approach for data anonymization based on MapReduce on the cloud. To make full use of the parallel capability of MapReduce on the cloud, specializations required in an anonymization process is split into two aspects. In the first one, initial datasets are segregated into an association of tinier datasets, and these datasets are anonymized in parallel, generating intermediary results. In the second one, the intermediary results are unified into one and additionally anonymize to attain persistent k-anonymous datasets. We soar MapReduce to attain the concrete computation in both phases. A collection of MapReduce jobs is determinedly devised and integrated to implement specializations collude on datasets. We assess our approach by administering experiments on real-world datasets. Experimental results demonstrate that with our approach, the scalability and efficiency of TDS can be improved significantly over existing approaches.

PROBLEM DEFINITION:

Scalability is the main concern when we are working on the extensive datasets. The centralized TDS approaches and exploits the data structure is to improve the scalability and efficiency by indicating index on anonymous data records and maintaining analytical information. There is an assumption that all data processed should apt in memory for the centralized approaches. Unfortunately, this presumption usually fails to hold in most data-exhaustive or accelerated cloud applications nowadays. In cloud environments, computation is catered in the form of virtual machines. Usually, cloud computes services provide various flavors of VM.

As a result, the centralized approaches are difficult in handling large-scale datasets well on the cloud using just one single VM even if the VM has the highest computation and storage capability. As pointed out TDS algorithm without considering the erosion of privacy perhaps adopts a specialization that directs to an abrupt violation of anonymity requisite. Hence, the distributed algorithm declines to demonstrate anonymous datasets disclosing the same data services as centralized ones. In addition; the problems like communication protocols and fault tolerance must be kept in mind when composing such distributed algorithms. As such, it is inappropriate to influence or leverage existing distributed algorithms to deal with the scalability issue of TDS.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

SCOPE:

Mapper Task: The Name Node i.e., master node (Job Tracker) takes the input, divides it into smaller groups, and distributes them to Data Node i.e., slave nodes (Task Tracker). A Task Tracker may perform this, again and again, leading to a multi-level tree structure. The task tracker processes the smaller sets and passes the result back to its job tracker. Reducer Task: The Name Node collects the results from all the sub-groups and combines them to produce the output – the results that we got was trying to solve the given task. MR framework is designed for distributed processing of map and reduce operations.

Besides each map operation is autonomous from others, all mapper task can be performed in parallel yet in the usual procedure it is defined by the sum of independent data sources. Similarly, a set of reducer task can perform the reducer task - supply all outputs of the mapper task that provide the same key presented to same reducer task. MR framework can be applied to significantly larger datasets to sort a petabyte of data in only a few hours. The parallelism also offers some feasibility of recuperating from the limited decline of servers or storage during the operation: if any map or reduce operation aborts, the task can be revised or rescheduled – still the input data is available in the other slave nodes i.e., data nodes.

Literature Survey:

Addressing Security Issues of Cloud Computing:

The present-day development of cloud computing has exceedingly transformed everyone's conception of foundation architectures, software distribution, and progress models. Private data like electronic health records and economic transaction records are mostly regarded highly sensitive even though those data can endeavour serious human benefits if they are investigated and mined by organizations like disease examination centers. Data anonymization has been broadly studied and widely used for data privacy protection in distinct data publishing and distributions. Data anonymization refers to concealing identification and/or sensitive data for holder of data records. Then, an individual privacy can be definitely sustained even though collective information is disclosed to data users for varied analysis and mining.

Privacy Preserving Multi-Keyword Ranked Search over Encrypted (MRSE) Cloud Data:

Initially, we determine and resolve the confronting issue of privacy preserving MRSE cloud data. We stimulate a set of stringent privacy condition for such a stable cloud data usage system. Among distinct multi-keyword semantics, we select the effective parallel measure of "coordinate matching", i.e., as many trials as likely, to capture the pertinence of data scripts to the search query. Additional we use internal product relations to significantly assess such measure of correlation. We initially come up with a fundamental concept for the MRSE established on secure internal product estimation and then present two apparently enhanced MRSE schemes to accomplish numerous strict privacy concerns in two distinct risk models. Comprehensive study inspecting privacy and effectiveness guarantees the proposed schemes is given. Experiments on the actual dataset farther exhibit projected schemes indeed offers low overhead on computing and communication.

Anatomy: Simple and Effective Privacy Preservation:

Anatomy: Simple and Effective Privacy Preservation Anatomy, for briefing sensitive data. Anatomy delivers all the quasi-identifier and sensitive values precisely in two individual tables. Linked with an ordered mechanism, this way privacy is protected and bags a huge amount of equivalence in the microdata. We establish a continuous algorithm for calculating anatomized tables that comply l-diversity privacy condition and lessen the error of recreating the microdata. Vast investigations confirm that our technique allows far more productive data analysis than the traditional reporting approach placed on generalization. Particularly, anatomy charter aggregate reasoning with the moderate error below 10%, which is the error acquired from a generalized table by a line of significance.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Workload Anonymization Techniques Aware for Large Scale Datasets:

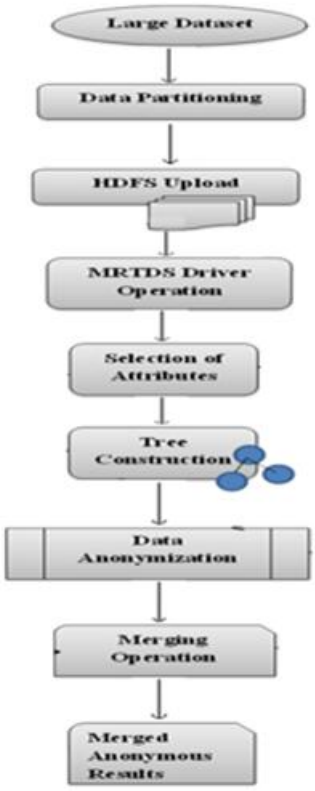
Usually, Anonymization algorithms purpose to gratify particular solitude definitions with nominal impact on the quality of the stem data. Even though much of the prior literature has steady quality over simple one-size-fits-all part, we agitate that quality is finest distinguish with regards to the workload for which the data will eventually be used. This article supports a bunch of anonymization algorithms that include an objective of workloads, exists of one or more data mining tasks as well as excerpt predicates. Thorough practical evaluation signifies that this method is much more adequate than previous techniques.

Distributed and Centralized Anonymization for High Dimensional Healthcare Data:

We normalize their data and privacy requirements to the issues of centralized, distributed anonymization and discover the major challenges that build conventional data anonymization means are not suitable. In addition to, we suggest a unique privacy model called LKC-privacy to alter the accost and accord anonymization algorithms to accomplish LKC-privacy in both the centralized and the distributed scenarios. Experiments on actual data exhibit that our anonymization algorithms can efficiently recall the vital information in anonymous data for data scrutiny and is extensible for anonymizing large datasets.

SYSTEM ARCHITECTURE:

Systems design is the process of defining the architecture, components, modules, interfaces, and data whereas system to please specific requirements. One could see it as the application of systems theory to product development.



Large Dataset:

In the very first level we have to take one large data set for Hadoop process. Here I was taken a patient large data set. Basically, in Hadoop environment, we are processing the map reducing operations on data sets.

Data Partitioning:

In Data is partitioned into Di, 1<=i<=p, it is required that the distribution of data records in Di is same as D. A data record is treated as a point in an m-dimension space, where m is the number of attributes.

HDFS Upload:

It means that Hadoop Distributed File System. In this module, whatever the dataset we are processing i.e. upload onto Hadoop environment for the purpose distribution, this distribution based on the attributes.

MRTDS Driver Operations:

Usually, a single Map-Reduce job is inadequate to accomplish a complex task in many applications. Thus, a group of Map Reduce jobs is orchestrated in a driver program to achieve such an objective.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

MRTDS subsists of MRTDS Driver operation and two types of jobs, i.e., IGPL Initialization and IGPL Update. The driver arranges the execution of jobs.

Selection of Attributes:
By selection of attributes the information complexity will decrease for further map reducing process.

Tree Construction:
Based on the attributes tree construction will happen.

Data Anonymization:
This level can apparently serve as anonymization extent of a dataset, i.e., the more specific AL a dataset has, the less extent of anonymization it coincides to. So, TDS approaches employ anonymization level to track and manage the specialization process.

Merging Operation:
Map and Reduce functions need to access current anonymization level AL, we use the distributed cache accoutrement to accede the assemblage of AL to each Mapper or Reducer node. Also, Hadoop provides the accoutrement to set plain global variables for Mappers and Reducers. The best specialization is transit into the Map function of IGPL Update job. The partition hash function in the shift phase is altered because the two jobs compel that the key-value pairs with the same key:p field rather than entire key should submit to the same Reducer.

RESULT ANALYSIS:
when handling extensive datasets on the cloud we analyze the scalability problem over existing TDS approaches. As centralized TDS to improve the scalability and efficiency by indexing anonymous data records and retaining statistical information. Centralized approaches probably undergo low efficiency and scalability while handling large-scale datasets. Our approach is established on the two levels of parallelization provisioned by Map Reduce on the cloud. Basically, Map Reduce on the cloud has two degrees of parallelization, i.e., job level and task level.

Firstly, Job level parallelization implies that multiple MR jobs can be executed simultaneously to make full use of cloud infrastructure resources. Combined with the cloud, MR becomes more powerful and elastic as the cloud offers on-demand resource infrastructure. For large dataset processing purpose, I am taking patient dataset, D is partitioned into Di, 1<=i<=p, it is required that the distribution of data records in Di is similar to D. Here a data record can be considered as a speck in an m-dimension space, where m is the number of attributes. So, the transitional anonymization levels derivate from Di, 1<=i<=p, can be more identical so that we can get a better-merged anonymization level.

Our approach can ensure the degree of data privacy preservation, as Two Phase TDS produces k-anonymous datasets. Level 1 ensures that the first phase produces consistent anonymous datasets that gratify higher degree of privacy preservation than user's requirement. Accordingly, MRTDS can more anonymize the entire datasets to produce final k anonymous datasets in the second phase. An original dataset D is explicitly stated for anonymization in a Map Reduce job. After obtaining the merged transitional anonymization level ALI, we run MRTDS (D, k, ALI) on the entire D, and get the final ALI. Details of Map and Reduce functions of the data specialization MR job. The Map function emits anonymous records and its count. The Reduce function plainly accumulate these anonymous records and counts their index number.

Anonymous records and its count represent a Quasi Identifier-group which constitute the final anonymous datasets. MRTDS consists of MRTDS Driver operation and other types of jobs, i.e., IGPL Initialization and IGPL Update. The driver organizes the execution of jobs. Initializes the values of information gain per privacy loss for all requirements, which can be achieved by the IGPL Initialization job. the best specialization is selected from valid specializations in current anonymization level information gain of the newly added specializations

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

and privacy loss of all requirements need to be recomputed, which are achieved by IGPL Update job. The iteration continues until all requirements become invalid, achieving the maximum utilization of data. The IGPL Update job is quite identical to IGPL Initialization, other than that it requires minor computation and consumes less network bandwidth. MRTDS exploits Map Reduce on the cloud to make the computation of IGPL parallel and scalable. We bestow IGPL Initialization and IGPL Update accordingly.

We conduct three groups of experiments in this section to evaluate the effectiveness and efficiency of our avenue. In the first one, we compare Two Phase TDS with Cent TDS from the aspect of scalability and efficiency. The other two, we investigate on the trade-off between scalability and data utility via adjusting configurations. All the experimental results demonstrate that our approach significantly improves the efficiency and scalability of TDS over existent TDS approach.

CONCLUSION:

Privacy preserving data investigation and data broadcasting are becoming serious difficulties in today's fragmentary world. In this study, we have detected the scalability problematic of large-scale data anonymization and found some problems regarding privacy preservation and data gain. To provide these functions a new system is to be proposed and is identical to extensive data anonymization by TDS approach with privacy preservation and information Gain.

The model is stimulated by the map and reduce functions commonly used in programming, even though their resolution in the Map Reduce structure is not the matching as in their original forms. The main influences of the Map Reduce framework are not the authentic map and reduce functions, but the extensibility and fault-tolerance extended for a variety of presentations by enhancing the execution engine once.

FUTURE SCOPE:

There are possible ways of data anonymization in which the current situation may be improved and next generation solutions may be developed. As future work a combination of top-down and bottom-up approach generalization is contributed for data anonymization in which data Generalization hierarchy is utilized for anonymization.

REFERENCE:

[1]S. Chaudhuri, "What Next?: A Half-Dozen Data Management Research Goals for Big Data and the Cloud," in 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," Commun. ACM, vol. 53, no. 4, pp. 50-58, 2010.

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

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

[5]D. Zissis and D. Lekkas, "Addressing Cloud Computing Security Issues," Fut. Gener. Comput.Syst., vol. 28, no. 3, pp. 583-592, 2011.

[6]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.

[7]L. Hsiao-Ying and W.G. Tzeng, "A Secure Erasure Code-Based Cloud Storage System with Secure Data

Forwarding," IEEETrans. Parallel Distrib.Syst., 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. 31st Annual IEEE Int'l Conf. Computer Communications (INFOCOM'11), pp. 829-837, 2011.

[9]P. Mohan, A. Thakurta, E. Shi, D. Song and D. Culler, "Gupt: Privacy Preserving Data Analysis Made Easy," Proc. 2012 ACMSIGMOD Int'l Conf. Management of Data (SIGMOD'12), pp. 349- 360, 2012.

[10] Microsoft HealthVault, http://www.microsoft.com/health/ww/products/Pages/ healthvault .aspx, accessed on: Jan. 05, 2013.

[11] 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.

[12] B.C.M. Fung, K. Wang and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans.Knowl..Data Eng., vol. 19, no. 5, pp. 711-725, 2007.

[13] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. 32nd Int'l Conf. Very Large Data Bases(VLDB'06), pp. 139-150, 2006.

[14] K. LeFevre, D.J. DeWitt and R. Ramakrishnan, "Incognito: EfficientFull-Domain K Anonymity," Proc. 2005 ACM SIGMODInt'lConf' Management of Data (SIGMOD '05), pp. 49-60, 2005.

[15] K. LeFevre, D.J. DeWitt and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," Proc. 22nd Int'l Conf. Data Engineering (ICDE '06), artical 25, 2006.