International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

The Wrapper Top-Down Specialization and Bottom-up Generalization Approach for Data Anonymization Using MapReduce on Hadoop

Shweta S. Bhand[1], Sonali Patil[2]

[1]P. G. Student, Department of Computer Engineering, BSIOTR, Wagholi, Pune, India

[2]Assistant Professor, Department of Computer Engineering, BSIOTR, Wagholi, Pune, India

Abstract: ***The goal of data mining is to determine hidden useful information in large databases. Mining various patterns from transaction databases is an important problem in data mining. As the database size increment, the computation time and have need memory also gain. Base on this, adopt the MapReduce programming mode which has parallel processing ability to analysis the huge-scale network. All the experiments were taking under hadoop, arrange on a cluster which consists of commodity servers. Through experimental evaluations in different simulation conditions, the planned algorithms are shown to deliver excellent performance with respect to scalability and execution time. Focused here one more data security approach, Privacy-preserving publishing of micro data has been studied extensively in recent years. Micro data have records each of which contains information about an personage entity, such as a person, a household, or an association. Several micro data anonymization techniques have been projected. some anonymization approach, such as generalization and bucketization. we present a data provider-aware anonymization algorithm with adaptive m-privacy checking strategies to ensure high utility and m-privacy of anonymized data with good organization. Experiments on real-life datasets propose that approach achieves better or comparable utility and efficiency than existing and baseline algorithms while providing m-privacy guarantee.***

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

1. Introduction

Data mining huge amounts of data collected in a wide range of domains from astronomy to healthcare has become essential for planning and performance. They are in a knowledge economy.The Data is an important asset to any organization. for e.g. Discovery of knowledge; Enabling discovery; annotation of data. They are looking at newer programming models, and Supporting algorithms and data structures.NSF refers to it as “data-intensive computing” and industry calls it “big-data” and “cloud computing”. “The big-data computing” is an essential advancement that has a potential impact .A programming model called MapReduce for processing “big-data”. A supporting file system called Hadoop Distributed File System (HDFS) . . The Map-Reduce based approach is used for data cube materialization and mining over massive datasets using holistic (non algebraic) measures like TOP-k for the top-k most frequent queries. MR-Cube approach is used for capable cube computation.

- **Data Cube**:- Data cube provide multi-dimensional views in data warehousing. If n size given in relation then there are 2^n cuboids and this cuboids need to computed in the cube materialization using algorithm[2]which is able to facilitate feature in MapReduce for efficient cube computation.
- **MapReduce:-** MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of autonomous tasks. The nature of this programming model and how it can be used to write programs which run in the Hadoop environment is explain by this model.
- **MR-Cube:-** MR-Cube is a MapReduce based algorithm introduces for efficient cube computation [5] and for identifying cube sets/groups on holistic measures. Complexity of the cubing task is depending upon two aspects: size of data and size of cube lattice.

2. Related Work

Ke Wang, Philip S. Yu, Sourav Chakraborty adapts an bottom-up generalization approach which works iteratively to generalize the data. These widespread data is useful for classification. But it is difficult to link to other sources. A hierarchical formation of generalizations specifies the generalization space. Identifying the best generalization is the key to climb up the hierarchy at each iteration [2].

Benjamin c. M. Fung, ke wang discuss that privacy-preserving technology is used to solve some problems only,But it is important to identify the nontechnical difficulties and overcome faced by decision makers when deploying a privacy-preserving technology. Their worries include the degradation of data quality, increased costs , increased complexity and loss of valuable information. They were under the impression that cross-disciplinary research is the key to remove these problems and urge scientists in the privacy protection field to conduct cross-disciplinary research with social scientists in sociology, psychology [3].

Jiuyong Li,Jixue Liu , Muzammil Baig , Raymond Chi-Wing Wong proposed two classification-aware data anonymization methods .It combines local value suppression and global attribute generalization. The attribute generalization is found by the data distribution, inspite of privacy requirement. Generalization levels are optimized by normalizing mutual information for preserving classification capability[17].

Xiaokui Xiao Yufei Tao present a technique, named *anatomy*, for publishing sensitive datasets. Anatomy is the process of releasing all the quasi-identifier and sensitive data items directly in two separate tables. This approach protect the privacy and capture large amount of correlation in microdata by Combining with a grouping mechanism. A linear-time algorithm for calculating anatomized tables that obey the l-diversity privacy requirement is developed which minimizes the error of reconstructing micro data [13].

3. Methodology

In big data applications, data privacy is one of the most concerned issues because processing large-scale privacy-sensitive data sets often requires computation power provided by public cloud services. Sub-tree data anonymization, getting a good trade-off between data utility and distortion, is a popularly adopted scheme to anonymize data sets for privacy preservation. Top-Down Specialization (TDS) and Bottom-Up Generalization (BUG) are two ways to fulfill sub-tree anonymization. However, existing methods for sub-tree anonymization fall short of parallelization capability, thereby missing scalability in handling big data on cloud. Still, both TDS and BUG go through from poor performance for certain value of k-anonymity parameter if they are utilized individually. In this paper, propose a hybrid approach that combines TDS and BUG together for efficient sub-tree anonymization over big data. Further, design MapReduce based algorithms for two components (TDS and BUG) to gain high scalability by exploiting powerful computation capability of cloud. The hybrid approach significantly improves the scalability and efficiency of sub-tree anonymization scheme over existing approaches.

4. System Architecture

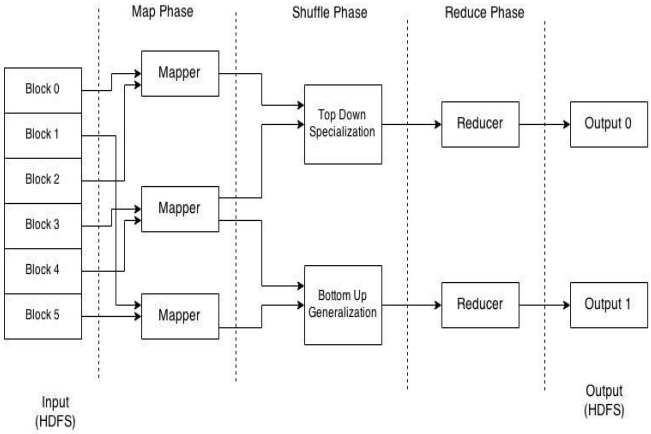

Figure 1: System Architecture

A. Top-Down Specialization

TDS is repeated process which is starting from the topmost domain values in the arrangement trees of attributes. Finding the best specialization, performing specialization and updating values of the search metric. Such a process of TDS is repeated until k-anonymity is violated, to description for the maximum data is going to utilize in that. The righteousness of a specialization is measured by a search metric The different android application permissions are fetched from android applications. These permissions are used as dataset for process. In that accept the information gain per privacy loss (IGPL), a tradeoff metric that take in mind both the privacy and information requirements. A specialization with the maximum IGPL value is regarded as best one and selected of each round. *Anytime* solution for Top-Down Specialization User may step through each specialization to determine a desired trade-off between privacy and accuracy. User may stop any time and obtain a generalized table satisfying the anonymity requirement.

Generalization: In this method, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' = 20', the value '23' by '20 < Age = 30'.

B. Bottom-Up Generalization

Bottom-Up Generalization is one of the efficient k-anonymization approach. K-Anonymity where the attributes are suppressed or generalized until each row is identical with at least k-1 other rows. Now database is said to be k-anonymous. Bottom-Up Generalization (BUG) approach of anonymization is the process of starting from the lowest anonymization level which is iteratively performed. We leverage privacy trade-off as the search metric. Bottom-Up Generalization and MR Bottom up Generalization (MRBUG) Driver are used. The following steps of the Advanced BUG are ,they are data partition, run MRBUG Driver on data set, combines all anonymization levels of the partitioned data items and then apply generalization to original data set without violating the k-anonymity.

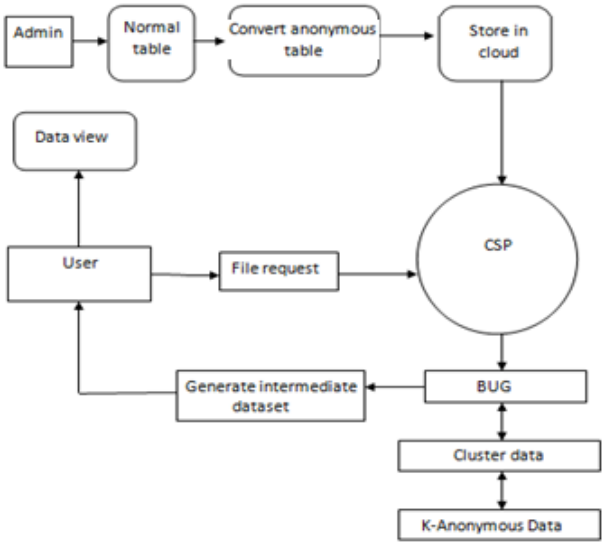

Figure 2: System architecture of bottom up approach

Here an highly developed Bottom-Up Generalization approach which improves the scalability and performance of BUG. Two levels of parallelization which is done by mapreduce(MR) on cloud environment. Mapreduce on cloud has two levels of parallelization. First is job level parallelization which means multiple MR jobs can be executed simultaneously that makes full use of cloud infrastructure. Second one is task level parallelization which means that multiple mapper or reducer tasks in a MR job are executed simultaneously on data partitions. The following steps are performed in our approach,

Bottom Up generalization algorithm
1: **while** *R* that does not gratify anonymity requirement **do**

2: **for all** generalizations *G* **do**
3: compute the *IP*(*G*);
4: **end for**;
5: find best generalization *Gbest*;
6: generalize *R* through *Gbest*;
7: **end while**;
8: output *R*;

First the datasets are split up into smaller datasets by using several job level map reduce, and then the partitioned data sets are anonymized Bottom up Generalization Driver. Then the obtained intermediate anonymization levels are merged into one. Ensure that all merged intermediate level never violates K- anonmity property. Getting then the merged intermediate anonymized dataset Then the driver is executed on original data set, and create the resultant anonymization level. The Algorithm for Advanced Bottom Up Generalization[15] is given below,The above algorithm describes bottom-up generalization. In ith iteration, generalize R by the best generalization G best .

C. MapReduce

The Map framework which is classified into map and reduce functions.Map is a function which parcels out task to other different nodes in distributed cluster. Reduce is a function that consolidate the task and resolves results into single value.

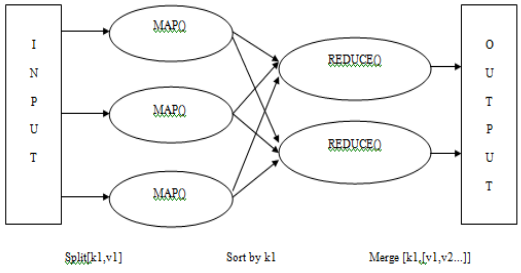

Figure 3: MapReduce Framework

The MR framework is fault-tolerant since each node in cluster had to report back with status updates and completed work periodically. Example if a node remains static for longer interval than the expected ,then a master node notes it and re-assigns that task to other nodes. A multiple MR jobs are required to accomplish task. So, a group of MR jobs are orchestrated in one MR driver to achieve the task. MR framework consists of MR Driver and two types of jobs. One is IGPL Initialization and second one is IGPL Update. The MR driver arranges the execution of jobs. Hadoop which provides the mechanism to set global variables for the Mappers and the Reducers. The best Specialization which is passed into Map function of IGPL Update job. In Bottom-Up Approach, the data is initialized first to its current state. Then the generalizations process are followed so that *k* - anonymity is not violated. That is, to climb the Taxonomy Tree of the attribute till required Anonymity is achieved.

5. Result

The system is developed by using JAVA (Version JDK). All systems are connected via 10Gb Ethernet and are running Ubuntu Linux 12.04. Each slave node runs our proprietary DataNode implementation for the DNN tests, and the standard Hadoop DataNode for the Hadoop 1.1.1 NameNode. A 14-drive RAID-6 volume is used for data storage at each of the slave nodes.

I. Macro MapReduce Benchmark Performance:-

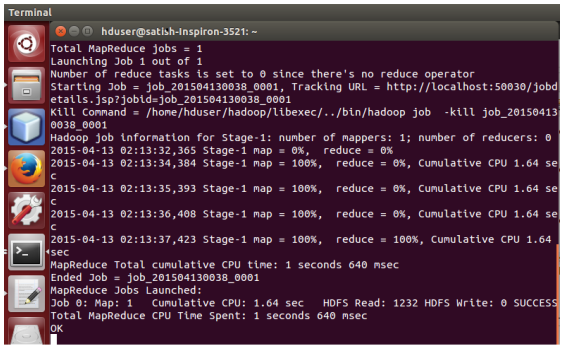

Figure 4: Hive Time for Execution

II. Health care high dimension data:-

MapReduce execution framework with health care high dimension data; they are included to demonstrate that DNN supports standard MapReduce.

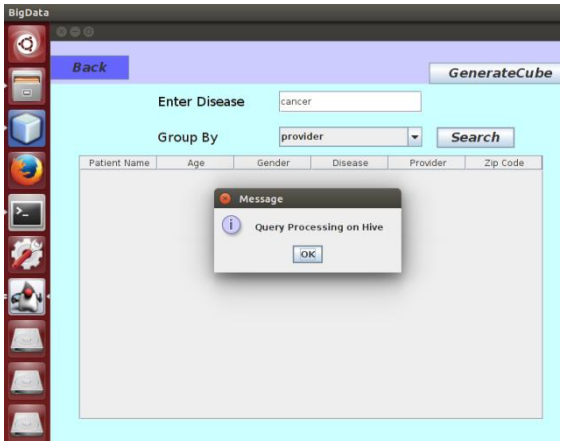

Figure 5: Hive Query

III. Cube Query

First in the first phase of application we generate the cube query it will directly execute on database.

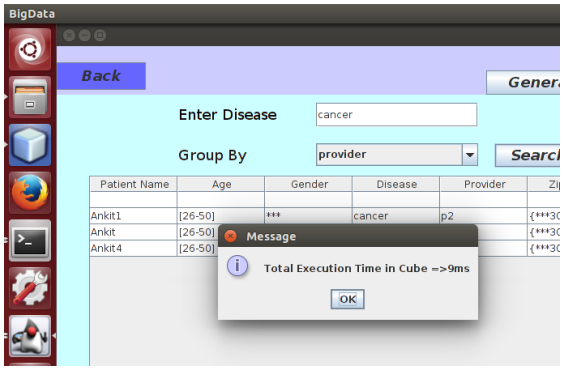

Figure 6: Cube time for complex query

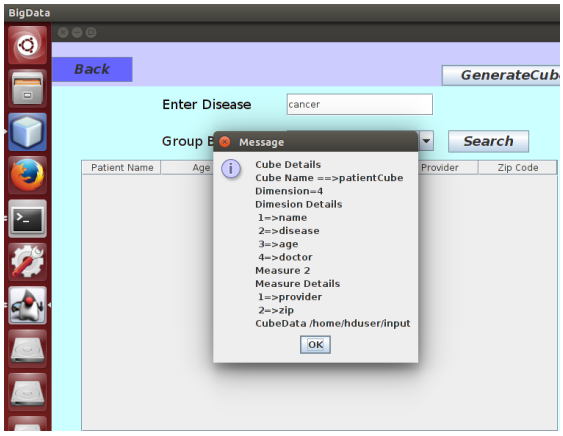

Figure 7: Cube Generation

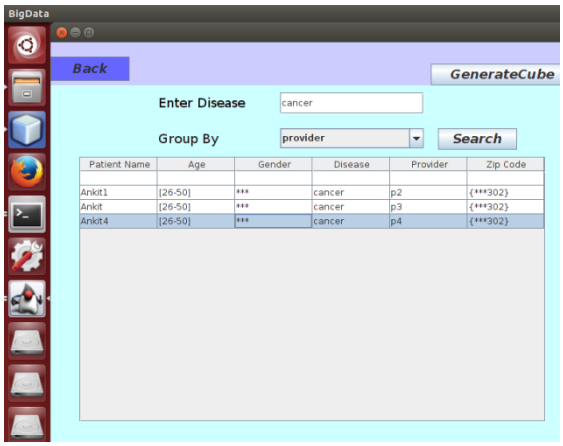

Figure 8: Search Result

6. Conclusion

Here the scalability problem for anonymizing the data on cloud for big data applications by using Bottom Up Generalization and proposes a scalable Bottom Up Generalization. The BUG approach performed as follows, first Data partitioning ,executing of driver that produce a intermediate result. After that, these results are merged into one and apply a generalization approach. This creates the anonymized data. The data anonymization is done using MR Framework on cloud. This indcate that scalability and efficiency are improved significantly over existing approaches.

References

[1] Xuyun Zhang, Laurence T. Yang, Chang Liu, and Jinjun Chen,"A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud", vol. 25, no. 2, february 2014.
[2] Ke Wang, Yu, P.S,Chakraborty, S, " Bottom-up generalization: a data mining solution to privacy protection".
[3] B.C.M. Fung, K. Wang, R. Chen and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACMComput. Surv., vol. 42, no. 4, pp.1-53, 2010.
[4] K. LeFevre, D.J. DeWitt and R. Ramakrishnan, "Workload- Aware Anonymization Techniques for Large-Scale Datasets," ACMTrans. Database Syst., vol. 33, no. 3, pp. 1-47, 2008.
[5] 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.
[6] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.
[7] Hassan Takabi, James B.D. Joshi and Gail-Joon Ahn, "Security and Privacy Challenges in Cloud Computing Environments".
[8] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain K-Anonymity," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '05), pp. 49-60, 2005.
[9] T. IwuchukwuandJ.F. Naughton, "K-Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization," Proc. 33rdInt'lConf. VeryLarge DataBases (VLDB'07), pp.746-757, 2007
[10] J. Dean and S. Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters," Comm. ACM, vol. 51, no. 1, pp. 107-113,2008.
[11] Dean J, Ghemawat S. "Mapreduce: a flexible data processing tool," Communications of the ACM 2010;53(1):72–77. DOI:10.1145/1629175.1629198.
[12] Jiuyong Li, Jixue Liu , Muzammil Baig , Raymond Chi-Wing Wong, "Information based data anonymization for classification utility"
[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] Raj H, Nathuji R, Singh A, England P. "Resource management for isolation enhanced cloud services," In: Proceedings of the 2009ACM workshop on cloud computing security, Chicago, Illinois, USA, 2009, p.77–84.
[15] K.R.Pandilakshmi, G.Rashitha Banu. "An Advanced Bottom up Generalization Approach for Big Data on Cloud" , Volume: 03, June 2014, Pages: 1054-1059..
[16] Intel "Big Data in the Cloud: Converging Technologies".
[17] Jiuyong Li, Jixue Liu Muzammil Baig, Raymond Chi-Wing Wong, "Information based data anonymization for classification utility".