IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Data Mining with Big Data and Privacy Preservation

Akshaya Tupe[1], Amrit Priyadarshi[2]

Student, Computer, DGOI, FOE, Bhigwan, Pune[1]

Assistant Professor, Computer, DGOI, FOE, Bhigwan, Pune[2]

Abstract: Big data is large volume, heterogeneous, decentralized distributed data with different dimensions. In Big data applications data collection has grown continuously, due to this it is difficult to manage, capture or extract and process data using existing software tools. Performing data analysis is becoming expensive with large volume of data in data warehouse. Data privacy is one of the challenge in data mining with big data. To preserving the privacy of the user we need to use some method so that data privacy is preserve and at the same time increase the data utility. In existing centralized algorithms it assumes that the all data should be at centralized location for anonymization which is not possible for large scale dataset,and there was distributed algorithms which mainly focus on privacy preservation of large dataset rather than the scalability issue. In the proposed system we focus to maintain the privacy for distributed data, and also overcome the problems of M-privacy and secrecy approach with new anonymization and slicing technique. Our main goal is to publish an Genuine or Anonymized view of integrated data, which will be immune to attacks. We use MR-Cube approach which addresses the challenges of large scale cube computation with holistic measure.Slicing contains tuple partition, generalization, slicing and anonymization. Once slicing is done the anonymized data can freely access by user with more data availability.

Keywords: Big Data, Hadoop, Map-Reduce, HDFS, MR-Cube, Data Security, slicing.

I. INTRODUCTION

Big Data

Big data contains variety of data such as audio, video, images, text and having large volume in terms of size, velocity (from batch processing to real time processing). The data is collected from multiple sources and having different dimensions. For example patients data can be stored in textual format, x-ray is in images and videos are stored for detailed treatment. This characteristic of big data makes it difficult to process, manage, and analyse data using conventional tools within the time like Visualization packages and relational databases. The analysers consider 30-50 terabytes to multiple petabytes as big data. To explore the large volumes of data and extract useful information for future actions is the fundamental challenge for big data applications. The paper presents a HACE Theorem which stands for Heterogeneous ,Autonomous sources ,complex and Evolving Relationships among data. Big data contains huge data with Heterogeneous and Diverse Dimensionality. Volume, variety and velocity are main characteristics of big data.

- Volume: each year large amount of data created by companies, corporations, organizations through web, mobile devices, social networking sites and the sources from where data created increase exponentially. This huge amount of data needs to be processed, managed and analysed to gain knowledge or information which needs in future for decision making.

- Variety: The variety of data types ,are one of the main characteristics of big data which differentiate it from others. Structured , unstructured data and semistructured data like social media sites like Fecebook data, location-based data (data collected from sensors), and log-file data.Structured Data has some defined format and which is coming from different sources. The data has different dimensions. for example in hospital health-care datacan be stored in textual format, images, and videos.

- Velocity: The velocity of data is defined as a speed at which new data is being created and the need for real-time analysis. From batch processing to real time processing the processing of big data changes in time.

II. LITERATURE SURVEY

1. Xindong Wu, Xingquan Zhu in "Data Mining with Big Data"

The Paper describe various characteristics of big data, different challenges of big data with data mining like data processing, data security and data mining algorithm. Traditional methods cant scale out their characteristics to process, manage and handle large amount of data. Daily 2.5 qbillion of data created. Here we need some other technique or methods which have capability to handle such data. The author has explained why we need to migrate on other technology. As the big data keeps growing continuously we need a framework which process, large. Today large amount of data created daily. Traditionally different techniques and data mining algorithms were used for processing such data but due to

the nature of big data (variety, velocity) it becomes difficult to process this big data. Data security is challenge of big data. Previous work shows that the system used encryption technique for storing and retrieving data. We need to find new technique which makes it easy to process, manage and store large data securely [1].

2. Benjamin C. M. Fung in "M-Privacy for Collaborative Data Publishing" This Paper addresses the collaborative data publishing problem. When there are n data provider the new insider attack problem arises. The data provider can be the owner of records. The owner can be untrusted , he can be infer the information using his own records and some background knowledge. The data from multiple providers can first collected and then generalized that data or first it is generalized then the data are gathered to analysis purpose. To provide high utility and efficiency the author uses the data provider-aware anonymization algorithm with adaptive m privacy checking strategies . It works efficiently than baseline algorithm [2].

3. Madhuri Patil in "Privacy Control Methods for Anonymous And Confidential Database Using Advance Encryption Standard" This Paper propose Advance Encryption Standard (AES) for privacy consuming to achieve privacy. K-anonymization technique provide more generalized data and maintain supressed form of data which is more secure describe privacy conserving of anonymous and confidential Database using AES approach for achieving privacy. In suppression based algorithm diffie Hellman Key exchange algorithm is used to generate private secure key. Then the AES algorithm is used to encrypt and decrypt data. Diffie Hellman key exchange algorithm is used to generate key [3].

4. Noman Mohammed and Benjamin C. M. FUNG in "Centralized and Distributed Anonymization for High Dimensional Healthcare Data Description" This System Proposes the LKC privacy model and presents the two algorithm,centralized anonymization algorithm and distributed anonymization algorithm to achieve LKC privacy in both centralized and distributed environment. He focus on the sharing patient information between the Hong Kong Red Cross Blood Transfusion Service (BTS) and the public hospitals. Medical practitioners and pharmaceutical researchers need to gain information related patients healthcare data for analysis purpose. It contains confidential information about individuals, and publishing such data will breach privacy. The author considers centralized and distributed anonymization of data. The centralized anonymization algorithm works such as first integrate then generate. by using these two algorithms They can generalize their information and provide privacy[4].

III. OVERVIEW OF HADOOP

Apache Hadoop is the open source implementation of the Map-Reduce framework. It has very simple programming model. Due to simple programming model and the run-time tolerance for node failures, Map-Reduce is widely used by companies, and organizations. Facebook ,Google, the New York Times are some examples of companies which are working with Hadoop. Users need a Hadoop platform which runs on a dedicated environment like a cluster or cloud to utilize Hadoop Map-Reduce.Hadoop uses a replication to provide a fault tolerance and to avoid the data loss. It has two levels first is Node level and second is Rack level. A node level which means a node failure should not affect the data integrity of the cluster and the rack level which means the data is safe if a whole rack if nodes fail. A Hadoop cluster is composed of two systems such as Hadoop Distributed File System (HDFS) and Map-Reduce.

HDFS

A Hadoop cluster uses Hadoop Distributed File System (HDFS) to manage its data. It provides storage for the input and output data of Map-Reduce jobs. HDFS is designed as a highly fault-tolerant, high throughput, and high capacity distributed file system. It is suitable for storing terabytes, petabytes or hexabytes of data on clusters and has extensible hardware requirements, which are typically comprised of commodity hardware like personal computers. The differences between HDFS and other distributed file systems are: HDFSs write once-read-many and streaming access models that make HDFS efficient in distributing and processing data, Reliably storing large amounts of data, and Robustly incorporating heterogeneous hardware and operating system environments HDFS divides a single file into fixed size blocks and also store multiple copies of each file into small fixed-size blocks (e.g., 64 MB) and stores multiple (default is three) copies of each block on cluster node disks. The distribution of data blocks increases throughput and fault tolerance. HDFS follows the master/slave architecture,where the name node means the master node which manages the file system namespace and regulates client accesses to the data,and there are a number of worker nodes, called Data nodes, which store actual data in units of blocks. The Name node maintains a mapping table which maps data blocks to Data nodes in order to process write and read requests from HDFS clients. It is also in charge of file system namespace operations such as closing, renaming, and opening files and directories. HDFS allows a secondary Name node to periodically save a copy of the meta data stored on the Name node in case of Name node failure.Data node store the blocks of data in its local disk and run the instruction like data replacement, creation, deletion, and replication from the Name node. The Data node once in a while reports its status through a heartbeat message and asks the Name node for commands. Each and Every Data Node listens to the Name Node to perceive connectivity with its Data node. If Name Node does not receive a heartbeat from a Data node in the configured period of time ,it marks the node is down. Data blocks is stored on this node will be treated as lost one and the Name Node will replicate those block automatically onto some other Data nodes.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Map-Reduce
Hadoop Map-Reduce is the Framework build upon HDFS (Hadoop Distributed File System), which is comprised of two stages :Map and Reduce. These stages take input as a Key/Value pair and also produce a output as a Key/Value pairs. When a Map-Reduce job is submitted to the cluster then it divides it into M map tasks and R reduce tasks, where each map task will process one block of input Key/Value data. Hadoop cluster uses worker node or we can say that slave node to execute Map-Reduce tasks. There are some boundaries on the number of Map and Reduce tasks that a worker can agree to and execute this simultaneously,i.e each worker node and has a same number of map slots and same number of reduce slots. At times , worker node sends a heartbeat signal to the Master Node. Ahead receiving a heartbeat from a worker node that has empty Map-Reduce slots, The Master node invokes the Map-Reduce scheduler to allot a Map tasks to read the contents of the corresponding input data block from Hadoop Distributed File System, perhaps from a Remote Worker node. The worker node parses the input out of the block, and passes each pair to the user defined map function. Then the function generates a intermediate key value pair as a output, which are again buffered in a memory, and also once in a while written to the local disk and separate it into R regions by the partitioning function . locations of this data is passed back to the Master Node, which is responsible for forwarding these locations to reduce tasks. A R reduce tasks uses RPC to read the Intermediate data generated by the M tasks of the job. R task is responsible for partition of intermediate data with certain keys. Therefore , it has to retrieve its region of data from all worker Nodes that have executed the M map tasks this process is called as Shuffle, which involves many to many communications in the middle of Worker Nodes. The reduce task reads intermediate data and also invokes the reduce function to produce the final output data for reduce partition.

IV. PROPOSED SYSTEM

In the propose system we focus to maintain the privacy for distributed data, and also overcome the problem of M-privacy and secrecy approach with new anonymization and slicing technique. We improve the privacy and utility of data with the help of slicing technique which fulfils privacy verification with better performance using new secrecy view algorithm. MR-Cube approach is used in proposed system for efficiently computing cubes. MR-Cube first generate the annotated lattice and then it is used to perform MR-Cube Map-Reduce. On this data we can perform the anonymization using slicing and verification is done through updated provider aware algorithm .

Slicing
Slicing is used to beat the limitation of generalization and bucketization. Slicing is used to Fig. 1. Cube Lattice handle high dimensional data.

1. Generalization: generalization is used to show less specific value.for example we can show the age of patients in generalize form i.e. in range.K-anonymity is very popular for generalization.

2. Anonymization: Anonymization is depend on vertical partitioning (attribute partition) and horizontal partitioning (tuple partition). zip code, address, disease , name-age partition together because they are highly correlated items. disease is sensitive attribute.

Map-Reduce
We will first introduce some preliminaries, Cube Lattice nd Dimension attributes: High Dimensional Data Attribute Security And Utility Using Hadoop A user want to analyse the dimension attributes. Cube lattice is formed by resenting all possible groupings of attributes that user want to analyse. Given n dimensions there are 2n-cuboids. Each cuboid captures the aggregated data over one combination of dimensions. These cuboids are stored in databases as view to speed up query processing. After forming cube lattice it is observe that not all regions (node in lattice) are represent valid aggregation condition. Invalid cube regions are eliminated and more compact hierarchical lattice is formed.The Cube is computed by computing measures which is given for all valid Cube group, aggregation function computes the measure which is based on all the tuples in the cube group.MR-cube accepts the challenges of large cube computation with holistic measures. Cube computation complexity is depend on size of data and size of cube lattice.

Proposed Algorithm

1) First MR-Cube approach is used for efficiently computing Cubes.
2) MR-Cube first generate Annotated Lattice and then it is used to perform MR-Cube Map-Reduce on Input Key-Value.
3) Then Slicing Technique is used to handle the High Dimentional data ,in this Slicing technique two techniques are used Generalization and Anonymization Technique.
4) Finally the Anonymization Technique is used for to generate anonymous data for user.

Proposed System Architecture
In the proposed system architecture contains user module, cube query, Map-Reduce, Anonymization modules. In first module the user can be an administrator, guest user or authorized one which already have an account. Administrator have all writes to accept request for create account, delete account and give rights. User can access the data on role based. If the user is guest user he/she can access the anonymized data for privacy preservation. Here we are going to use Hadoop frame-work for managing large scale distributed data and processing it. User pass a query as input in hadoop framework where data node is master node which receive this query. Cube query is generated annotated lattice which is further given to process main MR-Cube Map-Reduce Process. The materialized data is distributed for Map-reduce after generating lattice.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Finally the anonymization technique is used on the resultant data to generate anonymous data for user.

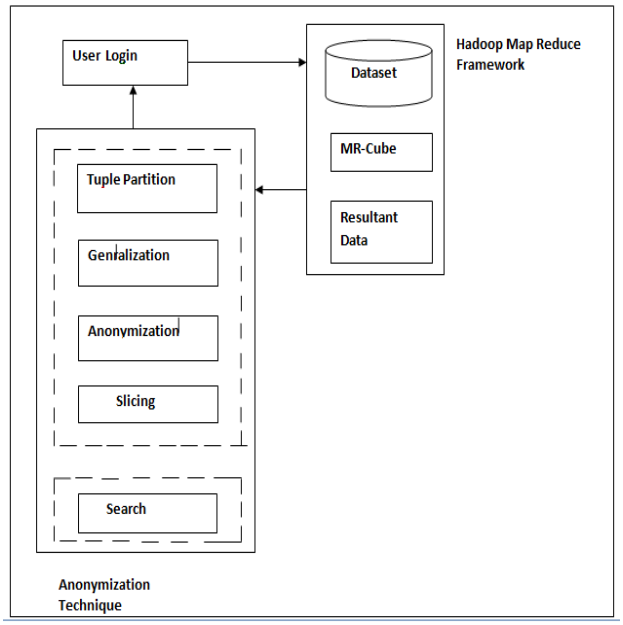

Fig. 1 System Architecture.

V. MATHEMATICAL MODEL

M = (Q, , , q0, F)

Where, Q is the set of States.
q0 is the initial State.
F is the final State.
Q = P0,P1,P2,Pf .

P0= Initial State
P1= Create Cube Query
P2= Map Reduce
Pf= Anonymization.

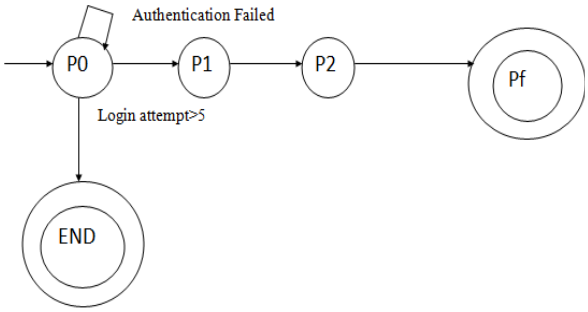

Fig.2 Mathematical Model.

VI. EXPECTED RESULTS

1) After our Experimentation we expect anonymized view of integrated data, which will be protected from attack.
2) Our Slicing technique will improve the security and privacy.
3) With the help of slicing technique which fulfill privacy verification with better performance than Provider Aware (Base Algorithm) and Encryption algorithm.

VII. CONCLUSION

In this paper, we have discussed that how Big Data is mined using the Hadoop MapReduce Framework.Hadoop Framework provides a Privacy over unauthorized access.In this paper, The proposed architecture is defined which includes different techniques to analyse the Big data such as, Map-Reduce,Slicing technique , by using this techniques final output is generated.

ACKNOWLEDGMENT

I would like to thank **Prof.Priyadarshi A.** for helping me in this work and huge support of Pune University and all the other people who have encouraged me for this research.

REFERENCES

[1] Xindong Wu,Fellow, IEEE Xingquan Zhu,Senior Member, IEEE, Gong-Qing Wu,and wei Ding,Senior Member,IEEE "Data Mining with Big Data" IEEE Trans. Knowledge and Data Eng., vol. 26, no. 1,January 2014.
[2] SBenjamin C. M. Fung in "m-Privacy for Collaborative Data Publishing" .
[3] Madhuri Patil, "Privacy Control Methods for Anonymous And Confidential Database Using Advance Encryption Standard".
[4] Noman Mohammed and Benjamin C. M. Fung, "Centralized and Distributed Anonymization for High Dimensional Healthcare Data Description".
[5] Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy in Slicing, "New Approach for Privacy Preserving Data Publishing".
[6] D. Mohanapriya, Dr.T.Meyyappan, "High Dimensional Data Handling Technique Using Overlapping SlicingMethod for Privacy Preservation"

BIOGRAPHIES

Akshaya Tupe is Currently pursuing Master Of Engineering in the field of Computer Engineering, from DGOI's FOE Swami-Chincholi,Bhigwan, and I Completed my Bachelor Of Engineering in the field of Computer Engineering , from SVPM's COE,Malegaon(BK),and currently working on final Year project with subject "DATA MINING WITH BIG DATA AND PRIVACY PRESERVATION" under the guidance of Prof.Priydarshi A.

Prof.Priyadarshi A. is Currently working as a Assistant Professor in Dattakala Group of Institution. He has Done B.Tech and M.Tech in Computer Science and Engineering from VTU ,Belgaon.