Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 10, October - 2017

Survey on Privacy Preserving for Data Leakage in Big Data Environment

Pratiksha Patil
Computer Science Department
Rajiv Gandhi Technical University
Bhopal, (Madhya Pradesh) India

Tejalal Choudhary
Computer Science Department
Rajiv Gandhi Technical University
Bhopal (Madhya Pradesh) India

***Abstract*-The desideratum of safe sizably voluminous data storage accommodation is extra attractive than yet to date. The rudimental obligation of the accommodation is to assurance the privacy of the data. In today's, sensitive data conspicuously needs to be kept private, data owners are often incentivized, or coerced, to allocate sensitive information ever more digital world, there is often a worry between bulwark privacy and sharing information. Immensely colossal data is a word utilized for prodigiously and sizably voluminous data sets that have more diverse and composite structure. These characteristics customarily correlate with extra difficulties in storing, analyzing and applying more events or extracting results. This paper fixates on confidentiality and security concerns in immensely colossal data, distinguish between privacy and security and privacy requisites in sizably voluminous data. This paper submits privacy and security aspects of cloud storage in immensely colossal data. Comparative study between sundry recent techniques of astronomically immense data privacy is withal done as well.**

Keywords: Immensely Colossal Data, Privacy Preserving, Security, data sharing, Data Leakage

I. INTRODUCTION

Astronomically immense data betokens genuinely a sizably voluminous storage of sizably voluminous data (substantial amount of data). Immensely colossal data cannot be processed by utilizing traditional techniques. It is not a data its consummate subject. Other definition of immensely colossal data is a buzzword or catch phrase, used to describe a massive volume of both structured and unstructured data that is so immensely colossal to process utilizing traditional data base &software techniques. Astronomically Immense data is what organization does with data that matter.

A. Sources of Immensely colossal Data

There are two sources of immensely colossal data: internal sources & external sources. Internal sources provide structure or organized data that originate from enteprise. Example of internal source are CRM (customer relationship mgmt), ERP (enter), sales data etc. whereas external sources provides unstructured or unorganized data that originates from external enrollment of organization such as internet, banking, regime. Internal data are primarily used to fortify daily business organization of operation. External data sources are analyses to understand the competitor, market, enrollment& technology.

Three v of Immensely Colossal Data

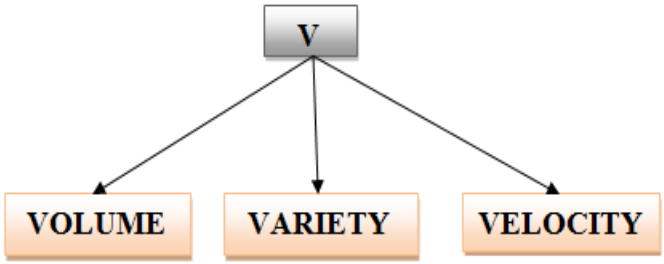

Figure 1: Immensely colossal Data Scenario

Volume: Organization accumulates data from different sources such as business transaction, convivial media &information. Volume of a data may increases day by day. Source of data have incremented significantly. for ex:typical cell phone we carry around not so long ago the only data it had was only voice call &text. With the advent of astute phones different technology sensor equipped with it the amount of data it carry are stringing coalesces with internet, digitization e commerce GPS contrivances now u can understand that not only size but additionally data volume in today's world.

Velocity: This is probably less understood probably in three V's. There is multiple dimensions in which data are expanded. There are:

1.Rate at which data is coming in. for ex: face book utilizer approx.2.7 million utilizer like a page in per day or 400 million incipient tweet are engendered by utilizer each day, that a lots of data coming in prodigiously expeditious time
2. Rate at which data needs to be analyzed. It is not a rate at which data is pouring in, it's about how expeditiously it will require processed. Batch processing is no longer adequate business manager need an authentic time data sometimes 2 minutes can be too tardy. for time sensitive processes such as to detect a fraud or catching fraud sizably voluminous data must be utilized as a streaming order to entrprenise to maximize the value.
3. Data is only subsidiary only as long as it is being processed more expeditious & it is entering in system. if the velocity of data processing is less than the velocity of the data entering the system all.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 10, October - 2017

Variety: Today data comes in all types of formats & they can be primarily relegated in two types:-
Structured data: Structured data can be defined as a set of data with a defined reiterating pattern its first defined engendering a data model. Structured data is:

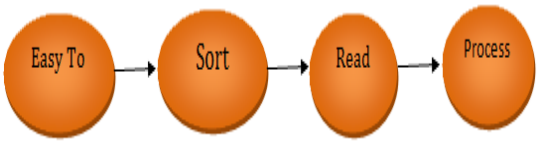

Figure 2: Structure data

Unstructured Data: Unstructured data are defined a structure that data without pre-defined pattern. Unstructured data typically text cumbersomely hefty but may additionally contain dates number as well this result in irregularities makes it arduous to decipher utilizing traditional computer programs as compared to structure data. Some sources of unstructured data include unstructured text internal an organization such as documents, logs, survey result, & email within the organization data base or data warehouse. whereas data from convivial media compromises data from gregarious media platforms including u tube, face book, twitter, LinkedIn etc.monitor hundred of live video victual from surveillance lens camera storm target point of interest or sentiments analysis of brand or products predicated on utilizer reviews on gregarious media such as fb & twitter managing merging & governing different variety of data are some of the issue .many organization did struggle with .as u visually perceive astronomically immense data is more than simply a matter of size. It is opportunity to find inside in incipient &emerging types of data &content. To make our business more limber & answer to question that will be anterior considered beyond our reach. Fig. has shown the unstructured data.

Figure3: Unstructured Data

B. Important of astronomically immense data
The consequential of big data does not revolve around how much data you have, but what you with it. You can take data from any source & analyze it to find answer that enable
1. Cost reduction
2. Time reduction
3. Incipient product development & optimized offering.
4. Smart decision making when you amalgamate big data with high powers analytics you can accomplish business cognate task such as:

- Determining root cause of failure issue & detect in near authentic time
- Generating coupon at the point of sale predicated on the customer buying habit.
- Recalculating entire risk portfolios in the minutes.
- Detecting fraudulent demeanor afore it affects your organization.

II. HADOOP

It is a framework, open source software which fortification sizably voluminous data that promulgates. Hadoop is fundamentally composed of many component, that is open source-Hadoop is fundamentally distributed storages & data processing system(DSP).Hadoop store data & processing the immensely colossal storage of the data in distributed manner. Hadoop is predicated on master slave architecture. One master & other is slave.HDFS is take care about data storage part.

A. Storage & retrieval in Hadoop
Main task of hadoop is in lieu of pooling the data from one single place than analyses data. Hadoop send a map reduce job across this data node & analyze the data node &result is return by the master. So the denomination node is spread the data across the data node & job tracker is responsible for the running the map reduce job & then it send the job across the task tracker (map reducing) on data node or the slave node. Following are the component, which involves in the procedure of storage & retrieval in hadoop:

1. Job Tracker is responsible for running the map reduce job & spread the task to each or across the each slave m/c.

2. Secondary designating node is utilized for backup of denominating node.

3. At high caliber of hadoop have five main components but two main components are hadoop are MR & HDFS.

B. Why Hadoop in sizably voluminous Data
Cost Efficacious system: Hadoop does not require sumptuous system or hardware in order to be implemented in simple hardware. This hardware component technically referred to as commodity hardware.

Astronomically immense cluster of node: Hadoop is composed of hundreds& thousand of node (node here is m/c).one of the sizably voluminous advantage having sizably voluminous cluster is, it offer sizably voluminous computing power & immensely colossal storage system to the client.

Parallel processing of data: If data can be processed simultaneously across all the nodes within the cluster & it preserving an abundance of time.

Distributed data: The hadoop framework takes care of splitting& distributing the data across all the nodes within the cluster. It additionally replicates the data with entire cluster.

Automatic failover management: In case any of the nodes in a cluster fails, the hadoop framework will supersede the failure m/c with another m/c. It additionally replicates all the configuration &syntax& data from this failed m/c on to incipient replicated m/c.

Data locality optimization: It is paramount feature offered by the hadoop framework, lets us endeavor to authentically understand what it does work. In a traditional approach whenever the software program is executed, the data is transferred from the data centre on to the m/c.for example let

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 10, October - 2017

us verbally express the data required by the program is located at some data centre in the USA & the program that require this data is located at Singapore, let us surmise the data required to the program is surmise one terabyte in size.

Transferring such, immensely colossal Volume of data from USA to Singapore world consumes an abundance of bandwidth & time. Hadoop abstracts this quandary by transferring code which is of few which is few megabyte in size located at Singapore to the datacenter located at USA & then it compile & execute the code is few megabyte in size as compared to the input data which is in one megabyte in size, this preserve bandwidth & lot of time

Heterogeneous cluster: Hadoop support heterogeneous cluster. This feature is withal the one of the consequential feature offered by the hadoop framework. We ken that hadoop cluster composed of several node, In cluster fundamentally refer to the cluster with in which each node is can be from a different vendor & each node can be run of variant & it different flavor of operating system. Let us verbalize a cluster is composed of four nodes:

•IBM m/c, running if RHEL enterprises Linux

•Intel m/c, running on UBUNTU Linux.

•AMD m/c running on FEDORA Linux.

•HP m/c running on CENTOS Linux

Scability : Scability is normally refers to the integrating & abstracted of the node as well as abstraction the hardware component & integrating the hardware component to the cluster. We can facilely integrate or abstract a node from the cluster without any doubt or affecting the cluster. Even the individual hardware component such as RAM & hardware can be integrated or abstracted from a cluster.

C .An example of HDFS

Cerebrate of a file that contains the phone numbers for everyone in the Cumulated States; the people with a last name starting with A might be stored on server 1, B on server 2, and so on. In a Hadoop world, pieces of this phonebook would be stored across the cluster, and to reconstruct the entire phonebook, your program would require the blocks from every server in the cluster. To achieve availability as components fail, HDFS replicates these more minuscule pieces onto two supplemental servers by default. (This redundancy can be incremented or decremented on a per-file substructure or for a whole environment; for example, a development Hadoop cluster typically doesn't need any data redundancy) This redundancy offers multiple benefits, the most conspicuous being higher availability.

In additament, this redundancy sanctions the Hadoop cluster to break work up into chunks that are more minuscule and run those jobs on all the servers in the cluster for better scalability. Determinately, you get the benefit of data locality, which is critical when working with sizably voluminous data sets.

Other examples are:

1. Financial accommodations companies use analytics to assets risk build investment models & engender trading algorithms, Hadoop has been used to avail build & run that application.

2. Retailers utilize it to avail analyze structured & unstructured data to better understand & accommodate their customers.

3. In the assets intensive energy industry hadoop powered analytics are utilized for the predictive maintenance, with input from internal of things (IOT) contrivances alimenting data in to astronomically immense programs.

4. Telecommunication companies can acclimate all the aforementioned use cases for example they can utilize hadoop-powered analytics to execute predictive maintenance on their infrastructure.

5. There are numerous public sector programs, ranging from anticipating & obviation & disease outbreaks to crunching numbers to catch tax cheats.

Hadoop is utilized in these & other astronomically immense data programs because it is efficacious, scalable, & is well, fortified by immensely colossal vendor & utilizer communities. Hadoop is a de facto standard in immensely colossal data.

III. PRIVACY & SECURITY CONCERN OF BIG DATA

Privacy & security in terms of immensely colossal data is a consequential issue. Sizably voluminous data security model is not suggested in event if involute application due to which it get incapacitated by default. However, in its absence data can always be compromised facilely. As such, this section fixates on the privacy &security issues.

Privacy information: Privacy is the privilege to have some control over how the personal information is accumulated & used. Information privacy is the capacity of an individual or group to stop information about them from becoming kenned to people other than those they give information additionally. One earnest utilizer privacy issue is the identification of personal information during transmission over internet.

Security: Security is the practice of forfending information & information assets by technology, processes & training from unauthorized access discloser, disruption, modification, inspection, recording & eradication.

A . Privacy Vs Security

Data privacy is fixated on the utilization and governance of individual data—things like establishing policies in place to ascertain that consumers' personal information is being amassed, shared and attacks and the misuse of purloined data for profit .While security is fundamental for forfending data, it's not adequate for addressing privacy. Table 1 fixates on supplemental distinction between privacy and security.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 10, October - 2017

Table 1: Difference between Privacy and Security

S.No.	Privacy	Security
1	Privacy is the felicitous utilization of user's information	Security is the "confidentiality, integrity and availability" of data
2	Privacy is the competency to decide what information of an individual goes.	Security offers the facility to be confident that decisions are reverenced
3	The issue of privacy is one that often applies to a consumer's right to safeguard their information from any other parties	security may provide for confidentiality. The overall goal of most security system is to for fend an enterprise or agency.
4	It is possible to have poor privacy and good security practices	However, it is arduous to have good privacy practices without a good data security program
6	For example, if utilizer make a purchase from XYZ Company & provide them payment address information in order for them to ship the product, they cannot then sell user's information to a third party without prior consent to user	The company XYZ uses sundry techniques (Encryption, Firewall) in order to avert data compromise from technology or susceptibilities in the network

B. OTHER CONCERN IN IMMENSELY COLOSSAL DATA

Afore implementation, some issue in immensely colossal data are to be noted & understood. Issue in astronomically immense data are not be arises quandary in implementation & concept, but they are paramount to ken how issue & quandary are handle. Some issues are:

- Storage & data convey
- Data processing
- Data management
- Highly adroit workforce
- Data acquisition
- Data dissemination

Storage & Data convey: The quantity of data are incremented each day, so incipient storage medium is required. Monthly data can be explosion, mainly due to gregarious media. Moreover data is being engendered by everyone & everything not just as here to fore, by professionals such as scientists, govern list, writers etc. current disk technology circumscriptions are about 4 terabytes per disk.so,1 Exabyte would require 25,000 disk. Thus transferring an immensely colossal (Exabyte) data would take 2800 hours if transferring data is sustain. But if data not consistent it would be take a longer time to transmit the data from accumulation or storage point to process the data. To handle issue in storage & convey of data should be "in place" & transmit only the resulting information. In the other words "bring the code to the data, unlike the traditional method of "Bring the data to the code".

Data processing: Surmise that Exabyte of data needs to be processed & transfer entirely. for simplicity, surmise a processor expend 100 ordinate dictation on one block at 5 giga hertz, thus efficacious processing of Exabyte of data will require extensive parallel processing 7 incipient algorithm to provides timely information.

Data management: Data mgmt in sizably voluminous data is how to manage the long term & short term issue. data management will perhaps, be the most conundrum to b addressed with sizably voluminous data. The astronomically immense data are of different variety(in terms of size by format, method of amassment. etc),Individual contribute digital data in mediums comfortable to them like documents, drawing, pictures, sound & video recordings models, software ,interfere design etc.with or without adequate meta data describing what, when, where ,who, why & how it was amassed & its provenance. In the data management, incipient approaches to data qualification & validation are needed. To summaries, there is no impeccable astronomically immense data management solution yet. This represents a consequential gap in the research literature on astronomically immense data that need to b filled.

IV. SUMMARY OF PAPERS

1. Cloud computing is igneous cumulation of a series of technologies. Many companies & organization have been migrating or building their business & work to cloud, but there is numerous of potential customers withal hesitate to store the data on cloud due to security &privacy preserving. Existing techniques approaches for preserving the privacy of dataset stored in cloud are Encryption, but it's very arduous task all to encrypt all dataset, because most of the application run on unencrypted dataset. A heuristic algorithm has designed .Cloud computing provides massive computational power & storage capacity. Which enable utilizer to deploy computational data intensive? But along with storage of data many intermediate dataset will be engendered. Encrypted pristine set are efficient but every time encrypted of intermediate data set are time consuming & withal cost efficacious.

2. With the advent of cloud computing & its model for IT accommodations cloud computing is predicated on concept of immensely colossal data. Data mining withal grows in cloud computing. Privacy preserving in data mining security is earnest issue, because here transaction can be done from server. One of the security issues is that server access to valuable data of the owner & may learn sensitive information from it.

3. For Example-visually examining any transaction, the server (intruder who can access the server) can learn which items are always co-purchased. The transaction & mined pattern are the personal information of utilizer. It should be safe from the server & this property of forfending private information of organization/companies is referred to co-operate privacy. During transaction personal information of utilizer or forfending personal info of organization or companies are

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 10, October - 2017

referred as corporate privacy. To forefend corporate privacy, the data owner transforms data to the server (mining query to server) recuperate the authentic pattern from server.

4. Data analytics concept in term of privacy preserving denotes in cities number of elder person which work in an industrial sector is growing rapidly & for this purport health is paramount concern. AIP (aging-in-place) is concept to elongate traditional healthcare accommodations to residential home utilizing sensor networks fortified by data analytics. In this concept, amassed sensor data from keenly intellective homes represent sensitive & personal information & consummate living deportment of individual. In this paper, Privacy preserving in framework is consist of three module & two storage units. First module is data collector its is present in keenly intellective home, which amass the data from sensor, second module is data receiver, its received the data emanate from data collector. Storage unit is de-indentified storage which store genuine data, third module is result provider this modules control end utilizer access to data processing result. In this paper requisite of access control technique to amend the privacy & security

5. There is many way to amass the human information from the sensor such as from mobile, laptops,etc accumulate all kind of data, variants of data from human convivial life,. Aggregate such a data by processed, analyzed& mined to extract the subsidiary information from the sizably voluminous volume of information. Abaft this concept the trust management(TM) is paramount concept.IOT is utilized for reliable data fusion mining opportunely, extract congruous information of utilizer with utilizer privacy & information security. In IOT number of quandaries arises in terms of trust. In IOT three layers works Physical layer, network layer, application layer. This layer is connected to each other. If the astronomically immense volume of data from the physical layer are not trust worthy than it engender the issue in data damaged, due to the malevolent input of some sensor. The next layer withal effected & quality will be great to influenced & hard to utilizer even through the network layer trust & application layer trust. In this paper main prospective is discover the properties of trust, propose objective of IOT trust management.

V. CONCLUSION

The astronomically immense data technology is utilized in those cases where a consequential amount of data needs to be processes. An opulent number of applications where per second a substantial amount of data appeared is processed and analyzed utilizing immensely colossal data such as Face book, yahoo and others are usages this technology. In these applications, the provision is made to communicate with each other and additionally these applications offers to apportion the data for their contacts or by public. But one the other hand the client is always worried about the privacy and sensitivity of data during the data sharing. In this presented work an ABE predicated cryptographic security for preserving user's privacy is proposed for design and implementation. That security scheme not only offers the cryptographic technique for securing the data that additionally simulate the community predicated data sharing techniques utilizing ABE technique.

After prosperously implementation of the proposed privacy preserving technique, the following outcomes are expected from the system.

[1] Enhancing security in data sharing applications by demonstrating the one to one, one to many and many to one schemes
[2] Design of efficient and secure cryptographic technique predicated utilize attributes
[3] Learning of gregarious networking and sizably voluminous data

REFERENCES

[1] AntorweepChakravorty, Tomasz Wlodarczyk, ChunmingRong, "Privacy Preserving Data Analytics for Smart Homes", 2013 IEEE Security and Privacy Workshops, 2013, Under license to IEEE.
[2] FoscaGiannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases", IEEE Systems Journal, Vol. 7, No. 3, September 2013
[3] Zheng Yan, Peng Zhang, Athanasios V. Vasilakos, "A survey on trust management for Internet of Things", & 2014 Elsevier Ltd. All rights reserved.Journal of Network and Computer Applications 42 (2014) 120–134
[4] Xuyun Zhang, Chang Liu, Surya Nepal, SurajPandey, and Jinjun Chen, "A Privacy Leakage Upper BoundConstraint-Based Approach forCost-Effective Privacy Preservingof Intermediate Data Sets in Cloud", IEEE Transactions on Parallel and Distributed Systems, Vol. 24, NO. 6, JUNE 2013.