International Journal of Trend in Scientific Research and Development (IJTSRD)

UGC Approved International Open Access Journal

ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 1 | Issue – 5

Privacy in Advertisement Services using Big Data: A Survey

Jitendra Soni1
Department of Computer Engineering
Institute of Engineering & Technology,
Devi Ahilya Vishwavidyalaya, Indore, M.P.

Imran Uddin
Department of Computer Science and Engineering
Prestige Institute of Engineering Management &
Research, Indore, M. P.

ABSTRACT

The privacy is primary requirement of growing technology. Maintaining Isolation over sensitive data in public environment is a big challenging task. It becomes more complex when data set becomes very large and number of users reaches to huge figure. Access Control principle help to classify the users according to rights and permission. Integration of Access Control model Hadoop Map Reduce is proposed to achieve privacy over sensitive data. Recommender systems have made significant utility in daily routing life. Attribute based Recommendation system help to explore targeted audience for advertisement industry. This paper has made a survey together various information about big data analytics and its importance. It also attempts to explore the concern for privacy issues in recommendation process.

Keywords: *Access Control, Privacy Protection, Sensitive Data, Hadoop*

1. INTRODUCTION

A rapid growth in society changes the responsibility and role of technology. Network plays a very important role for atomization of work. Thus it becomes the spine of system. Rapid development of applications and services raise online usage and it becomes more complex and emergences. Therefore, Network Data Analysis and Control is required to manage Network Data trends and classification. Most methods for Network Data analysis are operated on a single server environment. But if the amount of market data is increased, the existing infrastructure required increased infrastructure, memory speed and storage drives. A large quantity of data is known as BigData, required specialized methodology for efficient classification.

Big data is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, real-time data. In the recent year, a rapid hike in the large quantity of data is observed and it becomes the most useful technology tool for business.

A heavy processing and large amount of communications produces large amount of data which may be structured or unstructured from the different sources. The need of big data comes from the Big Companies like yahoo, Google, facebook etc for the reason of study of big quantity of data which is in shapeless outline. The study report conclude that Google, Facebook, tweeter, WHO etc. have large amount of data and required special technique to process them.

Here, an example of Big-Data might be in peta-bytes or exabytes. The data may be unstructured or structured some time may be incomplete or inaccessible. It can be categorized by 3Vs may be Volume, Velocity and Verity.

A brief classification of 3Vs is cited below and drawn is figure 1.1.

1. Volume
2. Velocity
3. Variety

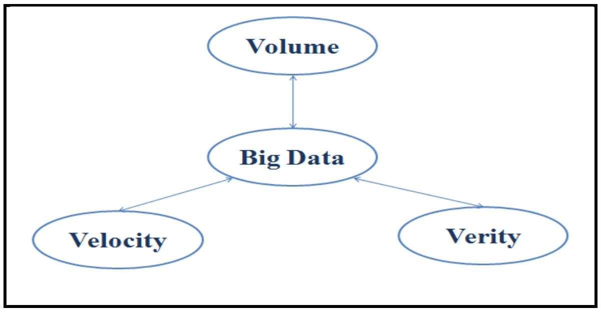

Figure 1: 3-V Relation with Big Data

Big Data term is rapidly getting a huge hike in technical aspects. A wide scale of utility it becomes centric interest of various researchers. Large size creates complex and difficult environment for processing, storage and transfer of information. Traditional algorithms and mining approach is not suitable for large data set and required a separate standards for parallel processing and distributed storage along with computation which help to reduce overhead and increase execution performance. Various tools are analyzed and discussed in this paper which is listed below;

1. Hadoop:

This is an Open source tool provides a reliable, scalable and distributed environment for creating data partitioning from inexpensive servers. Here, MapReduce framework can be used to process large scale of data with minimum overhead. This paper investigates certain situations where Hadoop can be useful.

1. Complex information processing is needed
2. Unstructured data needs to be turned into structured data
3. Queries can’t be reasonably expressed using SQL
4. Heavily recursive algorithms
5. Complex but parallelizable algorithms needed, such as geo-spatial analysis or genome sequencing
6. Machine learning
7. Data sets are too large to fit into database RAM, discs, or require too many cores (10’s of TB up to PB)
8. Data value does not justify expense of constant real-time availability, such as archives or special attention information, which can be stimulated to Hadoop and remain available at lower charge.
9. Results are not needed in real time
10. Fault tolerance is critical
11. Significant custom coding would be required to handle job scheduling

2. HDFS

The Hadoop Distributed File System (HDFS) is the file system component of the Hadoop framework. HDFS is designed to play down the storage overhead and mining on large amount of data on distributed fashion hardware.

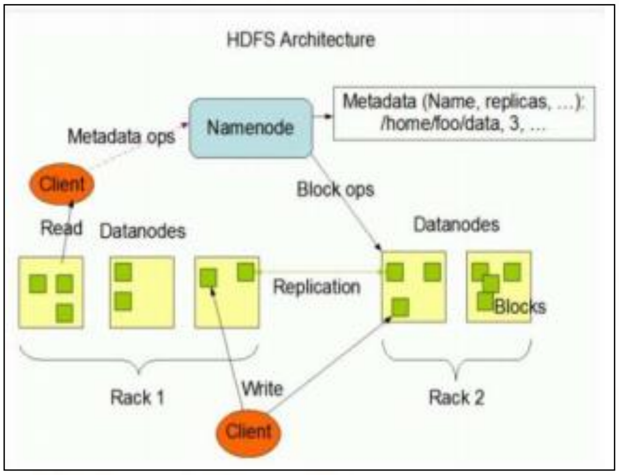

Figure 2: Block Representation of HDFS

3. PIG

PIG is the important component of Hadoop server like MapReduce and HDFS. Pig is made up of two components: Block representation of PIG is shown in figure 3.

1. The first is the language itself, which is called pig Latin (people naming various Hadoop projects for relation with naming conventions.)
2. The second is a runtime environment where Pig Latin programs are executed.

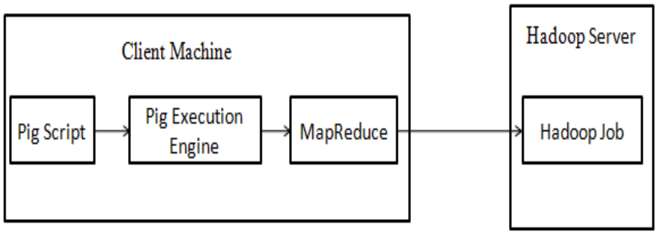

Figure 3: Block Representation of PIG

4. MapReduce

MapReduce is a framework for developing tools and source code for large data processing. It partitions the large data set into multiple parts to make processing easy and convenient. It simplifies the processing by auto making cluster according to name node based on machine.

Map-Reduce algorithm examines the dissimilar clusters and counsels the client for the common set of services used by the other users for the similar type of task. This will reduce the complexity and ambiguity of user to analysis the services provided by the cloud

The MapReduce Framework has two main function named with Map and Reduce. Here Map function is used to map large data into clusters and Reduce function is responsible to join the result into single unit. Proposed solution required that relative frequency of frequent service can be identified from different data set items. A block representation of MapReduce Framework is shown in figure 4.

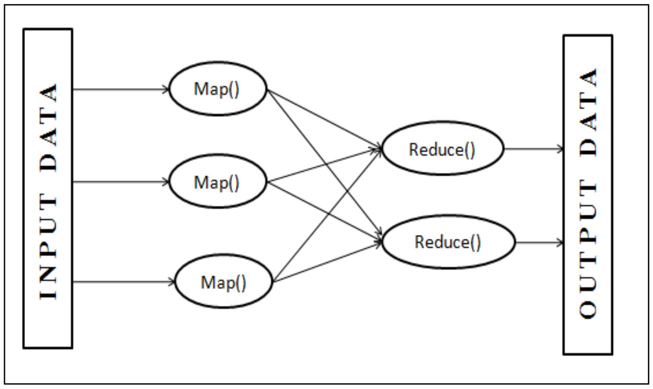

Figure 4: Block Representation of MapReduce

Analysis of Big Data through MapRecude function obtains the six phases which are;

1. Input reader
2. Map function
3. Partition function
4. Comparison function
5. Reduce function
6. Output writer

2. LITERATURE REVIEW

Mohammed Alodib **[1]** Service-oriented Architecture (SoA) is a layered architecture used to organize software resources as services that can be deployed, discovered and combined to produce new services. The interactions between services can be affected in situations where a destination service becomes unavailable. Herein, the Protocol service is introduced as a solution to coordinate interactions between services. The method is then extended to consider the automatic assignment of access control policies by the generation of a new service, called the Access Control Policies (AC Policies) service, which is linked to the Protocol service. In this context, the Protocol service manages a large amount of data. The analysis of such data sets may help improving the Protocol service performance. Dealing with such a large data sets is referred recently as “Big Data”, is a term related to large set of data that is complicated to be analyzed using traditional applications. One of the most successful implementations of Big Data is the Hadoop Framework. This research work proposes an extension to automate the integration of the Hadoop platform.

Samathani, P.,et. al. [2] A cloud services require in big scale, for users to share a private data such as electronic records and health records, transactional data for analysis of data or mining of that data which brings privacy concerns. In this paper they describe k-anonymity concept for the privacy preservation. Recently data in many cloud applications increases in that accordance with the Big Data style, and it make a challenge for commonly used software tools to manage, capture and process on large-scale data within an elapsed time. So, 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. The given paper, propose and implement a scalable two-phase top-down specialization (TDS) approach to anonymize large-scale data sets using the MapReduce framework on cloud.

Dou, W. et. al. [3] Privacy Preservation in Cloud computing provides promising scalable IT infrastructure to support various processing of a variety of big data applications in sectors such as healthcare and business. Data sets like electronic health records brings about privacy concerns potentially if the information is released or shared to third-parties cloud. This paper, investigate the local-recoding problem for big data anonymization against proximity privacy breaches and attempt to identify a scalable solution to this problem. Specifically, present a proximity privacy model with allowing semantic proximity of sensitive values and multiple sensitive attributes, and model the problem of local recoding as a proximity-aware clustering problem. A scalable two-phase clustering approach consisting of a t-ancestors clustering (similar to k-means) algorithm and a proximity-aware agglomerative clustering algorithm is proposed to address the above problem. The algorithms was designed with MapReduce to gain high scalability by performing data-parallel computation in cloud

HuseyinUlusoy, Murat Kantarcioglu, Erman Pattuk, Kevin Hamlen, et al. , 2014, [4]The paper demonstrates how a broad class of safety policies, including fine-grained access control policies at the level of key value data pairs rather than files, can be elegantly enforced on MapReduce clouds with minimal overhead and without any change to the system or OS implementations. The approach realizes policy enforcement as a middleware layer that rewrites the cloud's front-end API with reference monitors. After rewriting, the jobs run on input data authorized by fine-grained access control policies, allowing them to be safely executed without additional system-level controls. Detailed empirical studies show that this more modular approach exhibits just 1% overhead compared to a less modular implementation that customizes MapReduce directly to enforce the same policies.

Neelam Memon, Grigorious Loukides, Jianhua Shao, et. al., 2014, [5]: Transaction data, such as market basket or diagnostic data contain sensitive information about individuals and are used to support analytic studies. This raises privacy concerns, as the confidentiality of individuals must be protected. RBAT is an algorithm for anonym zing transaction data that has many desirable features. These include flexible specification of privacy requirements and the ability to preserve data utility well. However, these anonymization methods, limits the applicability of RBAT in practice. To address this issue, in this paper, researcher develops a parallel version of RBAT using MapReduce. They partition the data across a cluster of computing nodes and implement the key operations of RBAT in parallel. The experimental results show that scalable anonymization of large transaction datasets can be achieved using MapReduce and proposed method can scale nearly linear to the number of processing nodes.

Weidong Shi, TaeweonSuh, et. al., IEEE 2014, PFC [6]: this paper, propose PFC, a FPGA cloud for privacy preserving computation in the public cloud environment. PFC leverages the security feature of the existing FPGAs originally designed for bit stream IP protection and proxy re-encryption for preserving user data privacy. In PFC, cloud service providers are not necessarily trusted, and during outsourced computation, user's data is protected by a data encryption key only accessible by trusted FPGA devices. As an important application of cloud computing, PFC apply to the popular MapReduce programming model and extend the FPGA based MapReduce pipeline with privacy protection capabilities. Proxy re encryption is employed to support dynamic allocations of trusted FPGA devices as mappers and reducers.

Xianfeng Yang and Liming Lian, et. al., 2014, [7]: A New Data mining algorithm based on MapReduce and Hadoop: The goal of data mining is to discover hidden useful information in large databases. Mining frequent patterns from transaction databases is an important problem in data mining. As the database size increases, the computation time and required memory also increase. Base on this, the MapReduce programming mode which has parallel processing ability to analysis the large-scale network. All the experiments were taken under Hadoop, deployed on a cluster which consists of commodity servers. Through empirical evaluations in various simulation conditions, the proposed algorithms are shown to deliver excellent performance with respect to scalability and execution time.

Chun-Yu Wang, Tzu-Li Tai, et. al., IEEE 2014,[8]: The paper propose Federated MapReduce (Fed-MR), a framework aimed at analyzing geometrically distributed data among independent organizations while avoiding data movement. Fed-MR also integrates multiple clusters in different locations to form hierarchical Top-Region relationships. Experiments, compared to a single cluster with the same number of worker nodes, had shown that the computation time was only increased by an average of 30% in Word Count and 10% in Grep. Therefore, Fed-MR has reasonable overheads in performance for analyzing data across Internet-connected clusters while no additional Global Reduce function was required as in traditional hierarchical MapReduce frameworks Christos Doulkeridis.

3. PROBLEM DOMAIN

"Privacy is a state in which one is not observed or disturbed by other people" Privacy protection policy is an approach to isolate the sensitive information from unauthorized access. The complete work concludes that MapReduce Framework does not consist proposed security policy and suffering with data leakage problem.

Subsequently, Security threat attack is also possible and malicious framework may give open system access to unauthorized user. The complete phenomena generate a problem to implement security policy with MapReduce Algorithm.

The complete study observes that large scale of data requires advance level of processing and support to achieve desire level of concert. The Hadoop framework is an open source tool which not only supports for large data processing but also gives several components like MapReduce, HBase, Pig, Hive, and HDFS for distributed data processing. The rate at which large data generates and process, certain restriction and shortcomings may occur. Security is one of the big challenges to maintain the originality and preserve the privacy of information throughout the processing.

The privacy is a primary requirement of growing technology. To maintain isolation over sensitive (such as transactional data, medical diagnosis, and customer personal information in market dataset) is a big challenging task. The MapReduce Hadoop framework come with several advantages and disadvantages, have certain privacy issue as it discloses private and sensitive information. To prevent the information leaks, and to balance the goal of permissive model, the untrusted code should be confined. The traditional approach to data privacy is based on cryptography, which alone cannot enforces privacy demanded by bigdata services. The conventional system proposed for privacy problem in MapReduce is also unsuitable to many application that need data sets without noise, e.g., advertisement experiment data mining and analysis**.**

4. SOLUTION DOMAIN

Access Control is primary principle of information security and specifies “Who Can Access What”. Implementation of Access Control mechanism will filter out complete user access to framework data and avoid data leakage. It will also help to classify the request and response according to user rights.

To implement the access control, a list of services and users are expected. Access Control Matrix will give relation between users and services. It will classify the all user into categories and also services as same. The complete phenomena will help develop structured security plan to implement privacy protection mechanism using Access Control.

The complete solution will implement into MapReduce Framework to avoid security attack.

5. CONCLUSION

The complete work concludes that there is need to implement privacy issues during data mining for advertisement services and analytics process.

REFERENCES

[1] Mohammed Alodib, Zaki Malik “*A Big Data approach to enhance the integration of Access Control Policies for Web Services*” IEEE ICIS 2015, June 28-July 1 2015, Las Vegas, USA

[2] Dou, W. et. al. “Privacy Preservation in Cloud computing “Computer Science Munster University of Applied Sciences ¨ Steinfurt.

[3] HuseyinUlusoy, Murat Kantarcioglu, ErmanPattuk, Kevin Hamlen, et al. , 2014”Fine grained Access Control in Mapreduce ”.

[4] NeelamMemon,GrigoriousLoukides, Jianhua Shao, et. al., 2014, “A Parallel Method for Scalable Anonymization of Transaction Data” School of Computer Science & Informatics Cardiff University, UK.

[5] Weidong Shi, TaeweonSuh, et. al., IEEE 2014,”A FPGA cloud for Privacy Preservation computation ”

[6] Xianfeng Yang and Liming Lian,et. al., 2014, “A New Data Mining Algorithm based on MapReduce and Hadoop,” Xinxiang University, Xinxiang Henan, P.R.CHINA

[7] B.C.M. Fung, K. Wang and P.S. Yu, “Anonymizing Classification Data for Privacy Preservation,” IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 5, pp. 711-725, 2007.

[8] Christos Doulkeridis, Kjetil Nørv˚ag “A Survey of Large-Scale Analytical Query Processing in MapReduce”.

[9] ManolisTerrovitis, Nikos Mamoulis, Panos Kalnis, et. al. 2008, “Privacy-preserving Anonymization of set-valued data” Proc. VLDB Endow, vol 1. No. 1