www.ioirp.com
International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)
ISSN: 2394-6364, Volume – 2, Issue – 1. January 2016

Employing Attribute Relationship Methodology for Security Big Data

Arulmurugan. R[1], Manjula. R[2], Ramya. P[3], Karthikeyan.S[4]

Assistant Professor[1], Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India
Email: arulmurugan.r@ksrct.ac.in
Student[2], Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India
Email: manjulashanthi14@gmail.com
Student[3], Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India
Email: ramya1061994@gmail.com
Student[4], Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India
Email:karthis437@gmail.com

Abstract

Ever since the internet era there has been growing attention in big data and big data security due to expansion of network technology and cloud computing[1]. However, the big data is not completely new technology but a lean-to data mining. This paper, explicates the background of big data, data mining and big data features, thereby suggesting attribute selection methodology for big data for protecting the value of big data. Extracting important information is the main aim of analyzing big data which desires to be confined.

Therefore, weight between attributes of a dataset is a very important element for big data analysis. The focus is on two things. First, attribute import in big data is a key element for extracting information. In this perception, the revision on how to secure a big data though protecting valuable information is examined. Secondly, it is not feasible to protect all big data and attributes. By considering big data as a single object which has its own attributes and assuming that a attribute which have a additional relevance is more important than other attributes. Thereby considering the relevant attributes apply security mechanism for protecting those data. The AES algorithm I implemented to protect the selected attributes[15].

I.Introduction

Numerous new technologies have emerged with the expansion of network technology and distributed computing. Because of IT technology and the Internet, notable amount of digital information is produced and circulated every day. As a result, labors to pull out information from large-scale information are being accelerated. It is generally agreed today that big data and cloud are major fashion of modern computer technology. Experts usually say that "Big data" is a practice to extract the value from huge data further than the processing capabilities of offered databases. Big data is used in a variety of ways in different areas such as US health care, Europe public sector administration and global personal location data. Moreover, value is extracted from the big data in a variety ways. Among those, data mining techniques are mainly used. It is difficult to explain the big data without mentioning data mining.

By instinct, it is not possible to protect all the data inside the big data. Therefore, a methodology is used for selecting the attributes which needed to be protected.

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models[3]. It is designed to

scale up from single servers to thousands of machines, each offering local computation and storage.

The Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS is a distributed file system that provides high performance access to data across Hadoop clusters.

HDFS contains five nodes. They are,

- Name node
- Secondary Name node
- Data node
- Job Tracker
- Task Tracker

II. Existing System

From the view of data security, which has always been an crucial aspect of quality of service, Cloud Computing predictably poses a new demanding security threats for number of reasons.

Initially, established cryptographic primitives for the idea of data security protection cannot be straight adopted due to the users loss of control of data in Cloud Computing. Therefore, confirmation of correct data storage in the cloud must be conducted without a explicit knowledge of the whole data. Taking into consideration that various kinds of data for each user stored in the cloud and the claim of long term continuous declaration of their data safety, the problem of verifying exactness of data storage in the cloud becomes even more demanding.

Secondly, the Cloud Computing is not just considered as a third party data warehouse. The data stored in the cloud may be recurrently updated by the users, including deletion, insertion, modification, appending, reordering, etc[7].

The disadvantage of these techniques, which can be useful to ensure the storage correctness without having users possessing data, cannot address all the security threats in the cloud data storage, while they are all focusing on single server set-up and most of them do not reflect on dynamic data operations.

As an balancing approach, researchers have also proposed distributed protocols for ensuring the storage accuracy across multiple servers or peers. Over, none of these distributed schemes is aware of dynamic data operations. As a result, their applicability in cloud data storage can be significantly narrow.

III.Proposed System

Privacy in big data has raised serious concerns bringing into notice the need for efficient privacy preservation methods. Differential Privacy is a method enabling analysts to extract useful answers from databases containing personal information while offering well-built individual privacy protections. Access control is more difficult due to huge data scale. As mentioned earlier in the introduction, value is the key deliverable of big data. The data itself is not the subject of protection. In addition, securing the entire data is very inefficient, considering the volume of big data. There are several ways to protect the value the attribute selection methodology for big data security. So, we propose a security algorithm to protect the value of big data.

IV.Module Description

The system is divided in following modules for easier granularity

- Authorization
- Attribute Relationship
- HDFS management
- Privacy on Storage

1.Authorization

This module contains the process of user registration. The cloud user can register his profile first time and he can able to get login after this process. Once registration process completes he will get user certificate for login anywhere at any time. [4]By these credential the user can get sign into the cloud and they can perform the cloud operations like upload, download, and data sharing, provide security etc. When you log on to your machine and then try to access a resource, say a file server or database, something needs to assure that your username and password are valid. If you're logging onto a Windows machine, this authentication is performed by a component called the Local Security Authority Subsystem Service. . One of the concerns is related to authentication and authorization in the cloud in order to provide robust mechanisms to identify entities and establish their permissions and roles in the cloud,

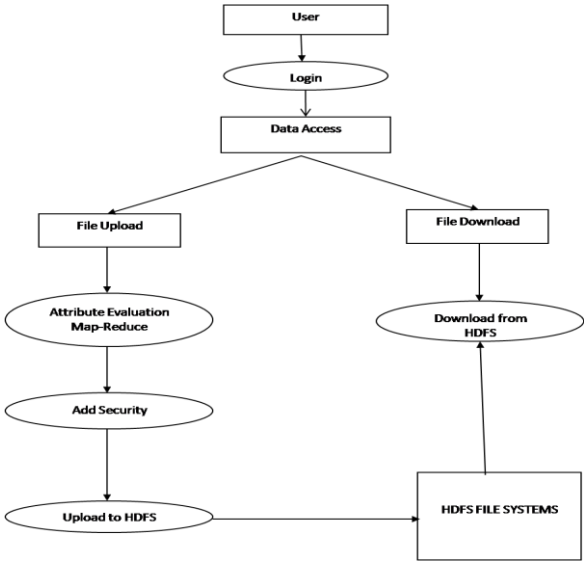
controlling resource usage and promoting accounting and isolation.

2.Attribute Relationship

Relevance between attributes of a dataset is a very important element for big data analysis. Attribute relevance in big data is a key element for extracting information[1]. In this project we make solution to secure a big data through protecting valuable information inside. For that we will use the map-reduce framework for analyzing relationship for each attribute in content. First, extracting all attribute of the target big data. Second, nodes are regarded as attributes and arranging in a circular. Third, if there is a relationship between two nodes, between the two nodes connected by an edge. Setting relationship based on the universal criteria and domain specific criteria. Universal criteria mean that two nodes have a hierarchical or interaction relationship. Additionally, in case of primary key exist, it include the relation on between primary key and other attributes. Domain specific is determined by own policies, depending on the type of big data. Next, select the protecting nodes by considering the number of edges. Finally, determine how to protect the selected nodes by policies.

3.HDFS management

The Hadoop Distributed File System(HDFS) is designed to store very large data sets dependably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks. In our projects we use HDFS for storing user data files. The Hadoop Distributed File System (HDFS) implements a permissions model for files and directories that shares much of the POSIX model. Each file and directory is associated with an owner and a group[13]. The file or directory has detach permissions for the user that is the owner, for other users that are members of the group, and for all other users. For files, the r permission is required to read the file, and the w permission is required to write or append to the file. For directories, the r permission is required to list the contents of the directory, the w permission is required to create or delete files or directories, and the x permission is required to access a child of the directory.



4.Privacy on Storage

The field of privacy in big data which contains a bunch of challenges involves interaction with individuals, re-identification attacks, probable and provable results, and economic effects[14]. In this module data are given by security constraints[2]. After analyzing attributes the data is analyzed that whether it is to be secured or not. If any file is to be secure the basic encryption methods will be used and encrypted then it will be stored into HDFS. Otherwise stored as a plain content. The key will be maintained by user for making security. As a result, potentially business sensitive and classified data is at risk from insider attacks. According to a recent Cloud Security coalition Report, insider attacks are the third prevalent threat in cloud computing. Therefore, Cloud Service providers must ensure that thorough background checks are conducted for employees who have physical access to the servers in the data center. Additionally, data centers must be frequently monitored for suspicious activity.

V.Conclusion

The proposed methodology for selecting attributes that should be protected on the object type of big data. This security methodology is appropriate for handling big data with multiple attributes. Especially, it is useful for cases that group of attributes and the value on multiple big data are not known. In the future, the improvement in matching the attribute generalization and then try to appraise the proposed method through big data environment and tools.

References

[1] D. Ritchey, "Attribute Relationship Evaluation Methodology for Big Data Security" Security, vol. 49, no. 7, pp. 28-30

[2] C. Tankard, "Big data security," Network Security, vol. 2012, no. 7

[3] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," pp. 1–13..

[4] M. Chen, J. Han and P.S. Yu, "Data mining: An overview from a database perspective,"knowledge and and data Engineering, IEEE Transactions on, vol. 8, no. 6, pp. 866-883 1996.

[5] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Commun ACM, vol. 51, no. 1, pp. 107-113 2008.

[6] C. Strauch, U.S. Sites and w, Kriha, "NoSQL databases," URL: http://www.christof-strauch.de/nosqldbs. 2011

[7] J. Bughin, M. Chui and J. Manyika, "Clouds, big data, and smart assets: Ten tech-enabled business trends to watch," McKinsey Quarterly, vol. 56, 2010.

[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers, "Big data: The next frontier for innovation, competition and productivity," McKinsey Global Institute, pp. 1-137, 2011.

[9] R. Gupta, H. Gupta and M. Mohania, "Cloud Computing and Big Data Analytics: What Is New from Databases Perspective?" in Big Data Analytics, Anonymous : Springer, 2012, pp. 42-61. ,2012.

[10] D. Boyd and K. Crawford, "Six provocations for big data," 2011.

[11] F. Gorunescu, Data Mining: Concepts, models and techniques,Springer, 2011.

[12] J. Han, M. Kamber and J. Pei, Data mining: concepts and techniques, Morgan kaufmann, 2006.

[13] J. Cohen, "Graph twiddling in a MapReduce world," Computing in Science & Engineering, vol. 11, no. 4, pp. 29-41 2009.

[14] Matturdi Bardi, "Big Data security and Privacy," Big data, cloud & mobile computing,2014.

[15] S. M¨ uller, S. Katzenbeisser, and C. Eckert, "Distributed attributebased encryption," in Information Security and Cryptology–ICISC2008. Springer, 2009, pp. 20–36.

[16] A. Sathi. Big Data Analytics:Disrupting Technologies for Changing Game. MC Press, 2012.

[17] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[18] IBM Big Data Analytics. IBM, 2o13 [Online]. Available:http://www.01.ibm.com/software/data/infosphere/bigdata-analytics.html