Privacy-Preserving Policy for Big Data Entry Restriction Scheme

Abraham Rajan[1] Gopinath A R[2] Kiran J[3] Arun Kumar S[4]
[1,2,3,4]Assistant Professor
[1,2,3,4]Department of Computer Science Engineering
[1,2,3,4]Nagarjuna College of Engineering and Technology, Bengaluru, Karnataka, India

Abstract— In contemporary world of computers, most of the objects like computers, smart phones are connected to network through internet connection generates vast amount of information. Due to huge information corpus, it is a very challenging task to store the data in structured formats, especially when it is stored in cloud storage space. In Cloud infrastructure, preserving the privacy of the user data is one of major challenges. In order to achieve this, we have developed a technique for encryption known as Attribute-Based Encryption using Cipher text-policy (ABE-CP). This technique gives users the ability to encipher their own data using feature attribute policy values which is defined over some access policies. If the access policies of the data owners are matched with attribute values of data consumers, then such users are allowed to decipher the data. In ABE-CP, we have encrypted data in plain text formats and has access policies attached to the same, which may contain end-user's personal information. The existing methodologies only partially conceal the personal information of end users while the attribute values which contains personal information are still exposed. In this paper, we introduce a system which ensures privacy of the data consumers and data owners while maintaining the standards. And we have also introduced a bloom-filter which is used for decryption of data using attributes. These bloom-filters are used to estimate whether the end-users access policies are defined for each attribute. It can also be used to find the attributes location if it is available in the access-policy. It has been estimated by many execution performance evaluators and security analyst that our proposed technique can enable linear secret-sharing strategic access policies and prevent private information of end users without engaging much overhead. In our proposed project we have provided two login credentials, one as data consumer and the other as data owner. The owner of the data uploads a file and generates a badge number for each and every file once the encryption is done in cloud. This process will protect the file name attribute values from leaking. Since the file name contains some critical information regarding the attribute values. These information's which will be used to decrypt the files by the intruders. Therefore, by providing badge number for each file we will protect privacy of the data owners.

Keywords: Cipher Text-Policy Attribute-Based Encryption (ABE-CP), Data Owners, Data Consumers, Encryption, Decryption

I. INTRODUCTION

A. *Big Data – What is Big Data?*

Big Data is becoming the technology of the future with a lot of scope in Data world. In the modern world technology can be referred to as Data world. Each and every device such as mobile, telephone calls, internet browsing, bank transaction etc. and many more are referred to as sources of data. Each device produces data and the number of such devices which generate data are in billions as there are billions of data generating devices. It becomes a challenge how to store and process all these data. The data which is beyond storage capacity and which is beyond processing capacity can be mentioned as Big Data. There are many data generating factors such as sensors, cc camera, social networks such as Facebook, Whatsapp, Online shopping- Ecommerce, Airlines, Hospitality data. Assuming we have 100% of data in current world, 90% of it was generated from the last 6 years. In 1990's the industry standard for a system is 1gb-20gb harddisk, 64-128mb RAM, 10kbps data transfer speed. Right now, in 2018 the industry standard for a system is 1Tb-2Tb harddisk, 4gb-32gb RAM, 100mbps-1000mbps speed. In the span of about more than 25 years the harddisk capacity has increased by about 1000 times as it is same for RAM and data transfer speeds. It is clearly evident from the above scenario that data generation will keep on increasing. Every time the user wants more space to store his data, he cannot purchase a new hard disk.

Fig. 1.1: Keywords Describing Big Data

This was taken as a challenge and IBM came up with IBM data servers. Users were given storage space on the servers and were charged accordingly. It was flexible, scalable and the data could be fetched from any remote place and could be processed. Three factors contribute to Big Data, they are – Volume, Velocity, Variety. Volume may increase rapidly in gigabytes, terabytes or petabyte. Velocity matters as user needs to send, receive and process data. There are 3 different varieties of data- Structured, Unstructured and semi-structured data. RDBMS deals with Structured data which is dealing with only 20-25% of total data. Whereas Unstructured and Semi-structured data constitute to 70-80% of data. Videos, images, text which are generated through social media are Unstructured data. Log files generated through Gmail, yahoo etc are semi structured data. Processing all these data's through a single resource will be time consuming, because of this reason Hadoop concept was introduced. Hadoop has been introduced as the best solution for Big Data. Doug Cutting is the founder of Hadoop. Core concepts of Hadoop are HDFS (Hadoop Distributed File System) and MapReduce. HDFS is technique for storing huge amount of data with cluster of commodity hardware whereas MapReduce is a technique for processing data stored in the HDFS. These all techniques help in analyzing the data which can used to predict the future of data which is a study based on Data Science. Big Data is becoming the technology of the future with a lot of scope in technology. We connect daily with technologies such as sensors, smart materials which connects to internet. Which is typically equipped with sensors which sends and receives information through controllers and we use different communication control protocols to monitor and control the system. The information from the sensors can be stored onto a centralized location called cloud using Hadoop HDFS concept.

B. What are the abstract ideas in Big Data?

Big Data as four major components that one as to consider before getting deep into Big Data. They are- Infrastructure security, Data Privacy, Data Management, Integrity and reactive security. In Infrastructure security the secure computations in distributed programming frameworks ensures that all computations are done without any mistakes. This becomes important when it involves bank transaction and other important procedures. We need to practice the best of security for non-relational data stores. While coming to Data Privacy we need to ensure a privacy preserving data mining, this is what we are concentrating in our project. The data privacy gives cryptographically enforced data-centric security and granular access control. To ensure privacy we encrypt the file name using some attribute and generate a tag number for a file which is 15-digit number.

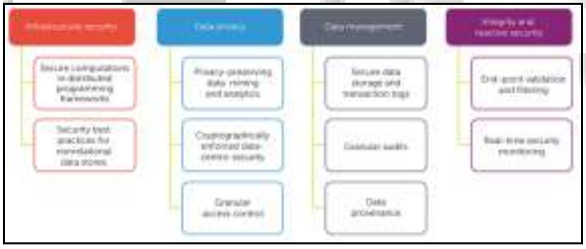
Infrastructure security
Data privacy
Data management
Integrity and reactive security
Secure computations in distributed programming frameworks
Privacy-preserving data mining and analytics
Secure data storage and transaction logs
End-point validation and filtering
Security best practices for non-relational data stores
Cryptographically enforced data-centric security
Granular audits
Real-time security monitoring
Granular access control
Data provenance

Fig. 1.2: Abstract ideas in Big Data.

In Data Management we need to take care that data storing and transaction logs are completely reliable. The Data integrity and reactive security make sure that there is End-point validation to check whether the data is accessed by valid user or not. The real time security monitoring system helps in prevented hackers from accessing the data by man in the middle attack and other techniques.

II. LITERATURE SURVEY

There are 2 types of storages as referred in the above figure, they are – 1. Analog Storage 2. Digital Storage. Analog Storage refers to Paper, film, audiotape and vinyl which contributes to 6%. Whereas Analog video tapes (VHS) contribute to 94% of Analog Storage. Analog Storage as about 19 Exabyte's Total of storage. Digital Storage as Portable media, flash drives which contribute to 2%, portable hard disks 2.4%, CD's and minidisks 6.8%, computer servers and mainframes which includes 8.9%, Digital tape 11.8%, DVD/Bluray 22.8%, PC hard disks 44.5% which as 123 billion gigabytes of storage, and another 1% includes chip cards, memory cards, floppy disks, mobile phones, PDA's, camera/camcorders, video games). Digital Storage as 280 Exabyte's memory in total. Before 2002 it was only era of Analog Storage, but 2002 was believed to be the beginning of the digital age during which period it contributed to 50% of storage. In the year 2007, 94% of storage was digital which had 280 Exabyte's of memory capacity.

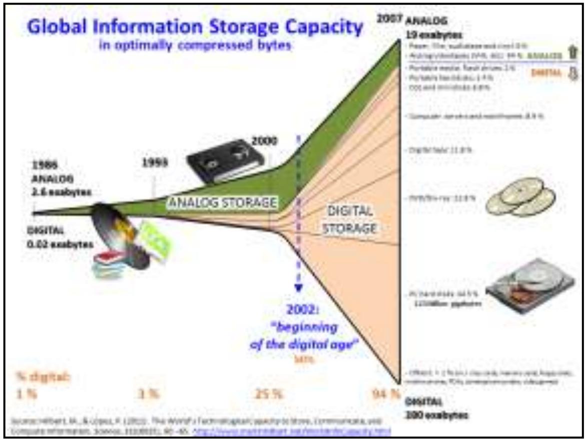

Fig. 2.1: Global Information Storage Capacity

A. Energy-efficient data replication in big data based cloud datacenters

Cloud computing is an emerging paradigm that provides computing, communication and storage resources as a service over a network. Communication resources often become a bottleneck in service provisioning for many cloud applications. Therefore, data replication which brings data (e.g., databases) closer to data consumers (e.g., cloud applications) is seen as a promising solution. This is in addition to the improved quality of service QoS obtained as a result of the reduced communication delays. The evaluation results, obtained from both mathematical model and extensive simulations, help to unveil performance and energy efficiency tradeoffs as well as guide the design of future data replication solutions. To address this gap, we propose a data replication technique for cloud computing data centers which optimizes energy consumption, network bandwidth and communication delay both between geographically distributed data centers as well as inside each datacenter. Advantages: It allows multiple virtual machines (VMs) to share the same physical server. Disadvantages: the popularity is not constant over time.

B. A Survey on Security Issues and Vulnerabilities on Big Data

Cloud computing has gained significant traction for recent years. It is a form of distributed computing whereby resources and application platform are shared over the internet through on demand and pay on utilization basis. Several companies have already built Internet consumer services such as search engine, use of some websites to communicate with other user in websites, E-mail services, and services to purchase items online that use cloud computing infrastructure. However this technology suffers from threats and vulnerabilities that prevent the users from trusting it. The occurrence of these threats may result into damaging of confidential data in cloud environment. This survey paper aims to analyze the various unresolved security threats in cloud computing which are affecting the various stake-holders linked to it. It also describes the pros and cons of the existing security strategy and also introduces the existing issues in cloud computing such as data integrity, data segregation, and security and so on. Cloud computing is a general term for anything that involves delivering hosted services over the Internet. It is an emerging computing technology that uses the internet and central remote servers to maintain data. This system is very helpful for different users so that they can easily use the system without any external support to software and hardware. They can also access their personal files at any computer on internet. This technology allows for much more efficient computing by centralizing storage, memory, processing and bandwidth.

Limitations of Big Data are –There are hundreds of vendors in the Big Data space with each having its own limitations/ strengths. So it becomes very hard to learn multiple software's for each of the tasks. Also connecting these individual system using customized connectors becomes a big challenge. The main deterrent in the steep learning curve behind these technologies and hence no human resources can be found for implementation projects.

III. SYSTEM ANALYSIS

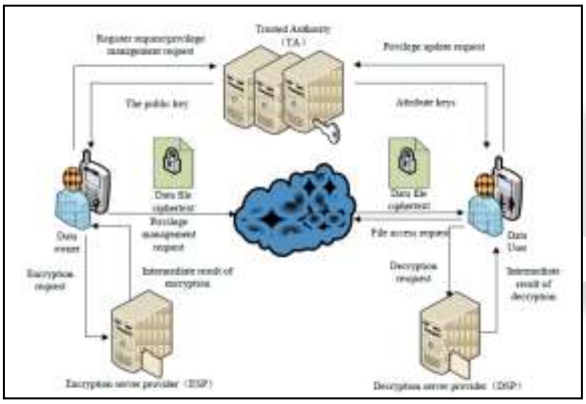
Trusted Authority (T.A.)
Register request/privilege management request
Privilege update request
The public key
Attribute keys
Data file ciphertext
Data file ciphertext
Privilege management request
Data owner
File access request
Data User
Encryption request
Intermediate result of encryption
Decryption request
Intermediate result of decryption
Encryption server provider (ESP)
Decryption server provider (DSP)

Fig. 3.1: Secure Data Sharing Scheme

In any of the data sharing schemes we have Data owner and Data consumer/user. Data owner basically uploads the file and data user download the file. The question arises whether data user is a valid user or not. Because of this we need to provide some access restriction on files. If man in middle attack or any type of attack happened to steal the data, even though he was able to steal he should not be able to view the content. For this purpose we provide ESP (Encryption Server Provide) which encrypts the data and provides the intermediate result of encryption. The user calls privilege management request and gets a public key and sends the file to the cloud in cipher text format. The data user using privilege update request gains the attribute key and uses it to decrypt the file in DSP (Decryption Server Provide).

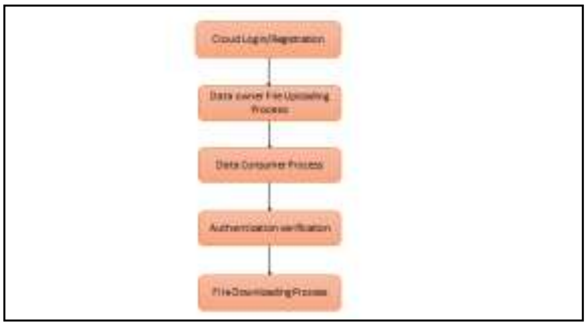

Fig. 3.1: Data flow Diagram

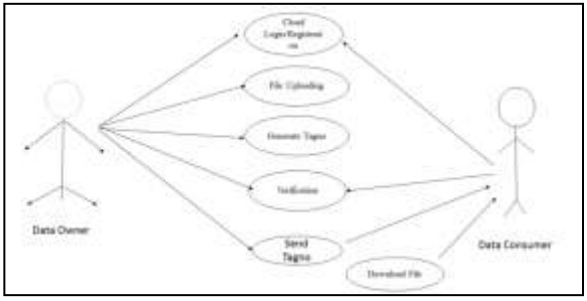

Fig. 3.2: Use case Diagram

The Use case Diagram and data flow diagram indicate the same thing as below. First, the user needs to login through the cloud registration. If he is a new user, the user should register himself for the first time. The user can be Data owner or Data consumer. Data owner as the rights to encrypt and upload the files while the data consumer as rights to download and decrypt the files. For the data consumer to download the file, he has to send a request to the data owner, the data owners sends an acknowledgment by sending the 15-digit tag number of the file. Using this attribute key the data consumer can decrypt the downloaded files.

The use case diagram also indicates the same thing as Data flow diagram. While login user can specify whether he is a data owner or data consumer. Once if he has login as Data owner he is allowed to upload the file content and encrypt it. He is even allowed to generate a tag number for file. The data consumer can download and decrypt the file.

IV. SYSTEM DESIGN

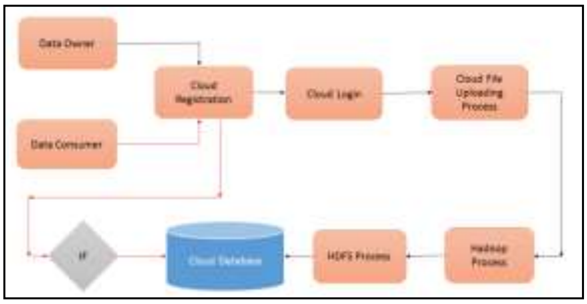

Fig. 4.1: System Architecture

The user can be Data Owner or data consumer. Different flows are determined for different type of users. The user needs to login through the cloud registration. If he is a new user, the user should register himself for the first time. The user can be Data owner or Data consumer. Data owner as the rights to encrypt and upload the files while the data consumer as rights to download and decrypt the files. The data owner encrypts data and sends to cloud Database which in our proposed system is our hard disk. For the data consumer to download the file, he has to send a request to the data owner, the data owners sends an acknowledgment by sending the 15-digit tag number of the file. Using this attribute key the data consumer can decrypt the downloaded files.

V. Future enhancement

For understanding the real life industrial development in Big Data, we have developed a prototype which resembles the original model. In the future we could integrate all of the above mentioned components of the proposed system to form a simple unit reducing much overhead on the system. This project can be taken into product level which will be cost effective, durable and user friendly in terms of providing security and preserving privacy of the user.

VI. Applications

File sharing Application: In any of the file sharing applications online or offline, the encryption of data is very important. If the valuable data regarding marketing, industries and others are in faulty hands, it can be used to damage our market and industries. In our traditional file sharing applications even though the contents are encrypted the access policy which contains attributes such as file name are in plain text format. This can be used to understand what the content is about. So, it is important to encrypt the file name also and generate Tag number of each file. When the data consumer wants to access a file he has to provide right authentication and request for the file. Once Data owner acknowledges and sends the tag number. Using that tag number Data consumer can download and decrypt the file.

VII. Conclusion

We have created a GUI (Graphical User Interface) for the user to login to cloud. The user may be a Data owner or Data consumer. The Data Owner is allowed to upload files using Hadoop and HDFS process, encrypt it and generate tag numbers used to preserve the privacy policy of user. The data Consumer can request the file and get the tag number to download the file and decrypt it.

In any of the file sharing applications online or offline, the encryption of data is very important. If the valuable data regarding marketing, industries and others are in faulty hands, it can be used to damage our market and industries. In our traditional file sharing applications even though the contents are encrypted the access policy which contains attributes such as file name are in plain text format. This can be used to understand what the content is about. So, it is important to encrypt the file name also and generate Tag number of each file. When the data consumer wants to access a file he has to provide right authentication and request for the file. Once Data owner acknowledges and sends the tag number. Using that tag number Data consumer can download and decrypt the file.

References

[1] K. Bilal, S. U. Khan, L. Zhang, H. Li, K. Hayat, S. A. Madani, N. Min-Allah, L. Wang, D. Chen, M. Iqbal, C. Z. Xu, and A. Y.Zomaya, “Quantitative comparisons of the state of the art datacenter architectures, ”Concurrency and Computation: Practice and Experience, Vol. 25, No. 12, 2013, pp. 1771-1783.

[2] K. Bilal, M. Manzano, S. U. Khan, E. Calle, K. Li, and A. Zomaya, “On the characterization of the structural robustness of data center networks, ”IEEE Transactions on Cloud Computing, Vol. 1, No. 1, 2013, pp. 64-77.

[3] D. Boru, D. Kliazovich, F. Granelli, P. Bouvry, and A. Y. Zomaya,“Energy-efficient data replication in cloud computing datacenters,” In IEEE Globecom Workshops, 2013, pp. 446-451. .

[4] Y. Deswarte, L. Blain, and J-C. Fabre, “Intrusion tolerance in distributed computing systems,” In Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy, OaklandCA, pp. 110-121, 1991.

[5] B. Grobauer, T.Walloschek, and E. Stocker, “Understanding cloud computing vulnerabilities, ”IEEE Security and Privacy, Vol.9, No. 2, 2011, pp. 50-57.