Auditing authorization of big data on cloud

[1]Prasad Patel, [2]Tushar Mali, [3]Chaitanya Patil, [4]Ritesh Patil, [5]Vinita Patil
[6] Prof. Vinay T. Patil
[1, 2, 3, 4, 5] Information Technology, Department,
[6]Assistant Professor in Information Technology Department,
D. N. Patel College of Engineering, Shahada,
North Maharashtra University, Jalgaon, Maharashtra, India.

Abstract—— Cloud computing is widely scattering era. It includes it companies, business line , all online shopping sites including cell phone service providers etc… but in other hand storage capacity and security are increasing issues. Cloud user have no more longer direct control over their data, which makes data security one of the major worries of using cloud. Earlier research work already allows data integrity to be verified without possession of the actual data file. The trusted third party known as auditor. And verification done by this auditor is known as authorized auditing. The Earlier system has many disadvantages regarding third party like any one can challenge to the cloud service provider for proof of data integrity.

Also in it includes research in BLSS signature algorithm to supporting fully dynamic data updates. This algorithm is used to update an only fixed-sized block known as coarse-grained updates. Though this system takes more time for updating data. In our paper, we are providing a system which support authorized auditing and fine-grained update request. Thus, our system dose not only increases security and flexibility but also providing a new big data application to all cloud service providers for large data many small updates.

Keywords: Cloud computing, big data, data security, authorized auditing, fine-grained dynamic data update.

I. INTRODUCTION

Cloud computing is unbearable relate to one of the most commanding novelties in information technology in recent years. A cloud is a type distributed system consisting of a collection of interrelated and virtualized computers that are dynamically possibility and presented as one or more unified computing resources based on service-level agreements established through compromise between the service provider and consumers. Cloud service is categorized into Infrastructure-as-a service(Iaas),Software-as-a-service(Saas),Platform-as-a-service(PaaS).IaaS is the way of providing on-demand computing resources like Server, storage array, virtualized data centers etc. PaaS is providing a advanced level Software application which can be comfortable to the different user's requirement. SaaS is the way of providing some particular applications as fully or partly remote services. It may include web based application or network interactions. In today's world of digitalization cloud computing has occurred as a concept of handling big data. This paper focus on the nature, origin and security related issues of the Big Data. Many single enterprises all over the world for example Amazon AWS, IBM Smart Cloud, Microsoft Azure. Which offers power-full public cloud services to users. Here Cloud based infrastructure, storage, and network, high computing performance helps to manage the features of big data. The above services provided by CSS makes the cloud user to be relaxed from burden of storing, managing and providing on -demand service to the client. The overhead incurred by implementing these entire infrastructure by own is reduced somehow. So now days Cloud computing is in great demand .Data privacy/security is the major worries in adoption of cloud computing. User will lose direct control over their data by comparing conservative systems. We will investigate the problem of integrity verification for big data storage in cloud .The issues related to security, integrity and availability of data. There is no direct control of user on cloud. But data integrity can be verified without possession of actual data. Verification done by a trusted third party (TPA) called data auditing. TPA can be anyone challenging the integrity of data stored in CSS. Our research work aim to add modification that can dramatically reduce communication overheads for verification of small updates. It not only enhance security and flexibility, but also significantly lower overheads for Big Data application with large no. of many small update such as application in social media and business transactions.

II. LITERATURE SURVEY

Ralph C. Merkle proposed new scheme "Digital signature based on a conventional encryption function such as DES is described which is more secure as the basic function. Existing Research like "New direction in cryptography" in 1976 and "Making the digital signature legal and safeguard" by S.M.Lipton, S.M.Matyas in Feb. 1978. Earlier work of Ralph C.Merkle in 1982 "Secrecy,Authentication, and public key systems these allrely on conventional encryption functions like one-way-function.But not single among them are much succeed in providing the convenience of system based on more complex mathematical problem.Ralph C. Merkleprovides advantage to reduce computational cost as compared with system which require modular arithmetic.The data encryption standard software implementation which runs faster than exponentiation modulo N, because of this digital signature system which is based on use of DES would get benefit from it.DES chips are already available at low cost of different manufacturer. New digital signature system is very fast when Retro fitted to a system that already has a DES chip. In this paper, they describe how new one time signature system can be used in a new way to provide a digital signature system that overcomes the limitation of Earlier . The general idea in this new system is to use an infinite tree of one time signature.

Suthan and Kesavaraja proposed scheme " Granule based File Storage System with Secure Transparent Availability". In this system, they focus on many security algorithms to provide security to the data that we store. Thefile is to be Splitted into n number of different particles and this particles be distributed within the system providing clearness to the user. The GFMS makes sure that the file is splitted and spread inside the system in such a way that, even if some part of the file is Retrieved no data can be recognized.Three major issues: Spyware, Encryption, File splitting so for Security purpose-The file is Encrypted with AES algorithm. and the secure key is generated for AES File security by using merkledamagard hash construction. This hash key is used for AES Encryption process.Cong Wang, Qian Wang, Kui Ren and Wenjing Louproposed system in which they focus on cloud data storage security, which has always been an very much significant aspect of quality of service. To make sure the accuracy of users data in the cloud. And also an effective and flexible distributed scheme with two salient structures, opposing to its ancestors. By utilizing the homomorphic token with distributed verification of erasure-coded data, our scheme achieves the integration of storage correctness insurance and data error localization. Algorithms used-token Pre-calculation, correctness verification & error localization, error recovery. Advantages are providing dynamic actions support and security strength against weak advisory.Suganya .S, Mrs. Sumathi they proposed system in the recent years, the Network-Attached Storage (NAS) and the Network File System (NFS) provide storage devices over the network so that user can contact the storage devices through network connection.

Disadvantages:

Encryption schemes supports confidentiality of the data, but the functionality of the storage system is limited because only certain operations are supported Data robustness is the important need for storage systems. Participating parties in the auditing scheme Rank-based Merkle hash tree.Ari Juels and Burton S. Kaliski JrProofs of retriev ability for Large Files some facility provided to user like archive and back up service provide to the user. User can retrieve the target file into the storage/server. POR can specially design to handle large file cryptographic techniques help users ensure the privacy and integrity of files they retrieve. In this system for users those want to verify that archives do not delete or modify files prior to retrieval. The main advantages of a POR are to accomplish these checks without users having to download the target files themselves. POR can be efficient enough to provide regular checks of file retrieve ability. They introduce a POR protocol in which the verifier stores only a single cryptographic key—irrespective of the size and number of the files whose retrieve ability it seeks to verify as well as a small amount of dynamic state (some tens of bits) for each file. it is worth considering a straightforward design involving a keyed hash function. Data-integrity protection is one of the fundamental goals of cryptography. Primitives such as digital signatures and message-authentication codes (MACs), are used in POR(Proofs of Retrieve ability). POR protocol encrypts file and randomly embeds a set of randomly-valued check blocks called sentinels. The verifier challenges the prove by specifying the positions of a collection of sentinels and ask the prove to return the associated sentinel values. Some challenges to accept in this system First, they offer a formal, concrete security definition of PORs that we believe to be of general interest and applicability in practical settings. Second, they introduce a sentinel-based POR scheme with several interesting properties, such as its uses function key generates secret key and encoded the file which is store in server, extraction, response and verify function used in this system to check the file will secure or not when we retrieve the file.

III.PROBLEM IDENTIFICATION

The test/confirmation procedure of our plan, we attempt to secure the plan against a malignant CSS who tries to cheat the verifier TPA about the honesty status of the customer's information, which is the same as past work on both PDP and por. In this progression, beside the new approval handle (which will be talked about in detail later in this section), the just distinction contrasted with is the and variable-sectored pieces. Along these lines, the security of this stage can be demonstrated through a procedure exceedingly comparable with utilizing the same system, ill-disposed model and intelligent amusements characterized in. A point by point security confirmation for this stage is thusly discarded here. We Are Having Three Main Components Viz. 1. Client 2. Cloud Service Provider (CSP) 3. Third Party Auditor (TPA) Functions or Authorities of Components: 1. Client Can create account Can select a file Can upload a file to CSS Can do updates in file 2. Cloud Service Provider (CSP) Can get file Can store file Can convert it in blocks 3. Third Party Authenticator (TPA) can get a file request can verity file integrity can challenge to CSS.

A. PROBLEM DEFINITION

In Earlier research it is shown that cloud environment provide various advantages by providing infrastructure as a service and maintenance as a service. It relieves the burden of user's task but security became a major concern in all time. User hire a TPA to check the integrity of data stored in cloud server. But again the problem arises whether user should trust or not on TPA. Another concern is related to the utilization of resources in cloud environment. There are number of resources as well as requests. There is no better way to serve the requests within a particular time and with available resource. There are also an increasing range of Information Communication Technology (ICT) vulnerabilities and threats that have to be effectively and efficiently managed. As a consequence, the confidentiality, integrity, availability and reliability of computerized data and of the systems that process, maintain and report these data are a major concern to audit. Earlier ly scheduling algorithms were performed in grid but reduces the performance by requiring advance reservation of resources. In cloud environment due to scalability of resources, manually allocate resources to task is not possible. Scheduling should be done in such a way that it will utilize the resources efficiently and also adopt the changes in environment configurations.

B. EXISTING SYSTEM

The description of the existing scheme in the aim of supporting variable

-sized data blocks, authorized third party auditing and fine - grained dynamic data updates.

The scheme is described in three parts:

- Setup: the client will generate keying materials via KeyGen and FileProc, and then upload the data to CSS. Different from Earlier schemes, the client will store a rank based Merkle Hash Tree (MHT)as metadata. Moreover, the client will authorize the TPA by sharing a value sigAUTH.
- Verifiable DataUpdating: the CSS performs the client‘s fine-grained update requests via PerformUpdate, then the client runs VerifyUpdate to check whether CSS has performed the updates on both the data blocks and their corresponding authenticators (used for auditing) honestly.
- Challenge, Proof Generation and Verification: Describes how the integrity of the data stored on CSS is verified by TPA via GenChallenge, GenProof and Verify.

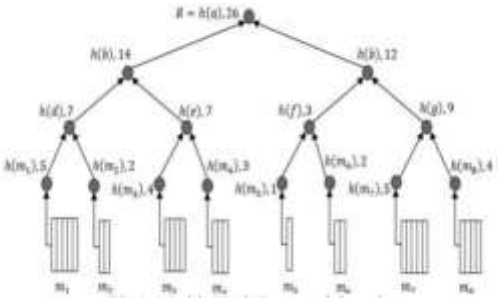

Figure 1.1 : MerkleHash Tree-rank based

In the rank based Merkle Hash Tree (Fig.1.1) each node N will have a maximum of 2 child nodes. In fact, according to the update algorithm, every non-leaf node will constantly have 2 child nodes.Each child nodes have varied in their block size. The time for retrieving the data from the block may vary according to the size of the block. If any user wants to update their file in the block, the server will return the block which is unstayed for the long time. Existing system requires lot of paper work. Moreover any unnatural cause (such as fire in the organization) can destroy all data of the organization. Loss of even a single paper led to difficult situation because all the papers are interrelated.

d) Poor identification of patient: In exiting system patient tries to hide information such as name, address, mobile number etc, because in our society this disease is seen from some different point of view so to escape from this they provide fake details.

e) Difficulty in reports generating: Reports generating in a current system are generated with great difficulty. It take time to generate report in the current system.

C. Project Objective

The objectives are as follows:

(i) Identify various security threats in cloud computing.

(ii) Enhance the security of the cloud through data mining techniques by making use of a single cache system.

(iii) Provide valuable suggestions to enhance the security of the cloud through data mining techniques.

For providing more security we are using TPA (third party authenticator).Which is able to verify our Data from cloud and check our data’s integrity. We are providing authenticity to the TPA using md5 hashing algorithm which is going to perform main function in our system. It will allow achieving us the security of our data from TPA also.Md5 has hinge algorithm gives 128bit hash key which is allocate to every TPA which should be given at the time of verifying data at cloud. To add modification that can dramatically reduce communication overheads for verifications of small updates. To not only enhance security and flexibility, but also significantly lower overheads for big data applications with a large number of many small updates such as applications in social media and business transactions.

D. Need For The New System

For marketing and research, many of the businesses uses big data, but may not have the essential assets particularly from a security perspective. If a security break occurs to big data, it would result in even more severe legal consequences and reputational damage than at present.In this new era, many companies are using the technology to store and analyze peta bytes of data about the company, business and the customers. As a result, information sorting becomes even more critical. For making big data protected, techniques such as encryption, logging, and honey pot detection must be necessary. The challenge of detecting and preventing advanced threats and malicious intruders must be solved using big data style analysis. These methods help in detecting the threats in the premature stages using more sophisticated pattern analysis and analyzing multiple data sources. There should be stability between data privacy and national security.

IV. How Does It Works

An Anti-Money Laundering Big Data engine collects the raw, external data from various sources such as Know Your Customer (KYC) information, real time transaction data, regulatory data etc. The input data undergoes enrichment, transformation, and vectorization, post which it is evaluated and scored for fraud checks. Event data often needs to be combined with data from other sources such as location, account details, or transaction data from other systems prior to being evaluated for norms such as security intelligence. AML engine uses high volume data inputs, click stream data, combination of rule based models, dynamic profiling analytics, intelligent scoring algorithms and Dynamic Anomaly Detection rules for fraud scoring & investigation. From AML engine, Risk management systems & Regulatory reporting identify the Transaction risks and compliance risks.

Figure 1.2: AML Engine

V. *WHAT IS AN IDEAL BIG DATA PLATFORM, BEST SUITED FOR ANTI MONEY LAUNDERING ?*

The Hadoop Big Data platform possesses the critical attributes ideal for AML activities. Some of these include pulling data; structured as well as unstructured, from various sources with ease, preparing the data followed by cleansing (this is critical for accuracy of insights), building a data model for analysing the data and compliance checks. The AML architecture is fully integrated with an organization's data hub. Staging Data will be complex since it contains multiple source data and this staging data provides runtimes for the predictive models to perform fraud detection.

Machine Learning (Neural Networks, Decision Trees, Bayesian Analysis etc.) is another key lever to predict and prevent Money Laundering patterns in Financial Institutions, by analysing an identified set of illicit operations. The system can also be taught to differentiate fraudulent transactions from legitimate ones by analysing data base(s) having records of only legitimate transactions.

VI. PROPOSED SYSTEM & METHODOLOGY

This project will investigate the problem of integrity verification for big data storage in server and focus on better support for minor dynamic updates, which welfares the scalability and efficiency of a server. This scheme only focuses on big data. To achieve this, this scheme utilizes a flexible data segmentation strategy and a data auditing protocol .Meanwhile, it address a potential security problem in supporting public verifiability to make the scheme more protected and robust, which is achieved by adding an additional authorization process among the three participating parties of client, server and a Manager.

A. System Architecture

According to architecture we are having three main components viz.:

1. Client
2. Cloud Service Provider (CSP)
3. Third Party Auditor (TPA)

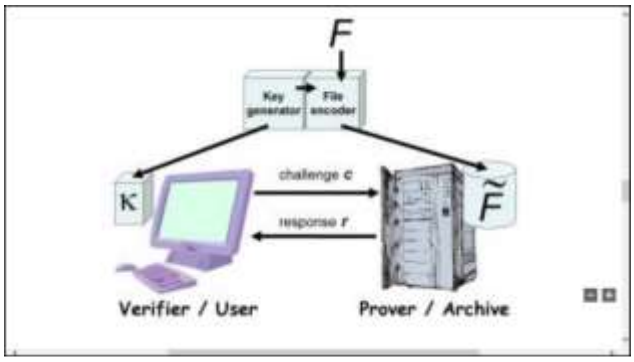

Figure 1.3: System Architecture

Functions or Authorities of Components:

1. Client
- Can create account
- Can select a file
- Can upload a file to CSS
- Can do updates in file

2. Cloud Service Provider (CSP)
- Can get file
- Can store file
- Can convert it in blocks

3. Third Party Authenticator (TPA)
- Can get a file request
- Can verity file integrity
- Can challenge to CSS

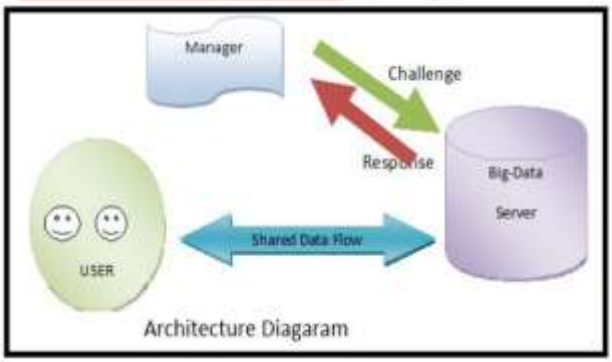

Figure 1.4: Architecture Diagram

The manager is responsible for the big data server to preserve the data integrity in the server. Fig.2. The manager can challenge the big data for the proof of data integrity. The big data server will produce response to the manager. The authorized user can share the data from the big data server and
can do data updates operation like insertion, updation and deletion. The manager will audit the server for every data update operations.

B. ALGORITHM USED

1. Merkle Hash Tree for Block Tag Authentication (MHTBA)

A common form of hash trees is the Merkle hash tree, hence the name. The root hash along with the total size of the file set and the piece size are now the only information in the system that needs to come from a confidential source. A client that has only the root hash of a file set can check any piece as follows. It calculates the hash of the piece it received.

Specifically, the server

- Replaces the block.
- Replaces outputs and
- Replaces the hash functions.

KeyGen (1k): This probabilistic algorithm is run by the client. It takes as input security parameter 1k, and returns public key pk and private key sk. (Ω, sig sk (H(R))). SigGen(sk; F) Φ← This algorithm is run by the client. It takes as input private key sk and a file F which is an ordered collection of blocks{ i m }and outputs the signature set Ω, which is an ordered collection of signatures { Ω } on{ m i} iσon m .It also outputs metadata the signature sig sk (H(R)) sk of the root R of a Merkle hash tree.

2. Elliptic curve cryptography (ECC)

Elliptic curve cryptography (ECC) is an approach to public-key cryptography based on the algebraic structure of elliptic curves over finite fields. Elliptic curves are also used in several integer factorization algorithms that have applications in cryptography, such as Lenstra elliptic curve factorization. ECC can yield a level of security with 164 bit key. Elliptic curves are believed to provide good security with smaller key sizes, something that is very useful in many applications. Smaller key sizes may result in faster execution timings. It establishes low computing power and battery resource usage. Using this algorithm the files are retrieved efficiently and securely in cryptographic manner.

3. Message Digestion (MD5):

a. It Is Designed To Run Effectively On 32-Bit Processor.

b. Generate Unique Hash Value For Each Input.

c. It Produce Fixed Length 128-Bit Hash Value with No Limit Of Input Message.

d. Advantage Is Fast Computing And Uniqueness.

e. Also Known As Hashing Function.

4. Advanced Encryption Standards (AES):

a. Secrete Key Generation Algorithm.

b. AES Work By Repeating The Same Defined Steps Multiple Times For Encryption & Decryption.

c. It Operates On Fixed Number Of Bytes.

d. Block Size: 128-Bit.

e. Key Length: 128,192,256-Bits.

f. Encryption Primitives: Substitution, Shift, Bit Mixing.

C. MODULES

System Model

User: users, who have data to be stored in the cloud and rely on the cloud for data computation, consist of both individual consumers and organizations.

Cloud Service Provider (CSP): a CSP, who has significant resources and expertise in building and managing distributed cloud storage servers, owns and operates live Cloud Computing systems.

Third Party Auditor (TPA): an optional TPA, who has expertise and capabilities that users may not have, is trusted to assess and expose risk of cloud storage services on behalf of the users upon request.

1. Block-Level Operations in Fine- Grained Updates

Block-level operations in fine-grained dynamic data updates may contain the following 6 types of operations: partial modification PM- consecutive part of a certain block needs to be updated; whole-block modification M - whole block needs to be replaced by a new set of data; block deletion D - whole block needs to be deleted from the tree structure; block insertion J - whole block needs to be created on the tree structure to contain newly inserted data; and block splitting SP - part of data in a block needs to be taken out to form a new block to be inserted next to it.

2. Our Scheme

a) Update Operation

In cloud data storage, sometimes the user may need to modify some data block(s) stored in the cloud, we refer this operation as data update. In other words, for all the unused tokens, the user needs to exclude every occurrence of the old data block and replace it with the new one.

b) Delete Operation

Sometimes, after being stored in the cloud, certain data blocks may need to be deleted. The delete operation we are considering is a general one, in which user replaces the data block with zero or some special reserved data symbol. From this point of view, the delete operation is actually a special case of the data update operation, where the original data blocks can be replaced with zeros or some predetermined special blocks.

c) Append Operation

In some cases, the user may want to increase the size of his stored data by adding blocks at the end of the data file, which we refer as data append. We anticipate that the most many append operation in cloud data storage is bulk append, in which the user needs to upload a large number of blocks (not a single block) at one time.

D. APPLICABILITY

The scope of 'Financial-crime' has considerably widened over time, due to which Governments and Organizations alike need to be extremely cautious; anti money laundering, anti-fraud, anti-corruption, sanctions and embargoes are they key. Anti-Financial crime and Money Laundering is a big area of spending, especially for financial institutions. As per a leading analyst firm, Risk & Compliance spending will grow from USD 79 to USD 97 Billion globally, with a significant amount being spent specifically on compliance domains. From a business perspective, the major areas of Regulatory & Compliance focus include SEC, FINRA, FINCEN, FCA, OCC, APRA, FINMA, RBI / SEBI and MSA to name a few. How do Financial Institutions successfully identify the legitimacy of the millions (and in some cases billion!) of financial transactions occurring every day? The answer- A robust Anti Money Laundering Strategy powered by Big Data.

VII. EXPECTED RESULT

As a result, every small update will cause re-computation and updating of the authenticator for an entire le block, which in turn causes higher storage and communication overheads. In this project, we provide a formal analysis for possible types of ne-grained data updates and propose a scheme that can fully support authorized auditing and ne-grained update requests. Based on our scheme, we also propose an enhancement that can dramatically reduce communication overheads for verifying small updates.

VIII. CONCLUSION

Thus in our project we are providing a formal analysis and ne-grained data updating. Purpose of our scheme is that fully sup-port authorized auditing and ne-grained data updating as per request. Based on our scheme we have also proposed modification that are dramatically reduce communication overheads for verification of small updates. We also plan that for further investigate on the next step how to improve server side protection methods for data security. Hence, in our project data security, storage and computation, efficient security plays important role under cloud computing context.

REFERENCES

1. JUELS AND B.S. KALISKI JR., ''PORS: PROOFS OF RETRIEVABILITY FOR LARGE FILES,'' IN PRO. 14TH ACM CONF. ON COMPUT.AND COMMUN. SECURITY (CCS), 2007, PP. 584-597.

2. H. SHACHAM AND B. WATERS, ''COMPACT PROOFS OF RETRIEVABILITY,''IN PROC. 14TH INT'L CONF. ON THEORY AND APPL. OF CRYPTOL. AND INF.SECURITY (ASIACRYPT), 2008, PP. 90-107.

3. R.C. MERKLE, ''A DIGITAL SIGNATURE BASED ON A CONVENTIONAL ENCRYPTION FUNCTION,'' IN PROC. INT'L CRYPTOL. CONF. ON ADV. IN CRYPTOL. (CRYPTO), 1987, PP. 369-378.

4. HADOOP MAPREDUCE. [ONLINE]. AVAILABLE: HTTP://HADOOP.APACHE.ORG

5. OPENSTACK OPEN SOURCE CLOUD SOFTWARE, ACCESSED ON: MARCH 25,2013. [ONLINE].
AVAILABLE: HTTP://OPENSTACK.ORG/

6. ARMBRUST, A.FOX, R.GRIFFITH, A.D.JOSEPH, R.KATZ, A.KONWINSKI, G.LEE,D.PATTERSON,
A.RABKIN,i.STOCIA, AND M ZAHARIA "A VIEW OF CLOUD COMPUTING ."COMMUM,ACM,
VOL.53,NO.4,PP.50-58,APR.2010

7. MR.VINAY TILA PATIL & GAJENDRA SINGH CHANDEL, PROF. (2014). IMPLEMENTATION OF TPA AND DATA INTEGRITY IN CLOUD COMPUTING USING RSA ALGORITHM. INTERNATIONAL JOURNAL OF ENGINEERING TRENDS AND TECHNOLOGY. 12. 85-93. 10.14445/22315381/IJETT-V12P215.

8. VINAY TILA PATIL, PROF. GAJENDRA SINGH CHANDEL, "APPLYING PUBLIC AUDITABILITY FOR SECURING CLOUD DATA FROM MODIFICATION ATTACK", IJSRD - INTERNATIONAL JOURNAL FOR SCIENTIFIC RESEARCH & DEVELOPMENT, VOL. 2, ISSUE 04, 2014, ISSN (ONLINE): 2321-0613