ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

An Efficient Searching Scheme with Data Integrity and Anonymization on Mobile Cloud

Anju Chandran[1], P.K.K.Thampi [2]

Final Year M Tech Student, Dept. of CSE., Sree Narayana Gurukulam College of Engineering, Kerala, India[1]

Associate Professor, Dept. of CSE., Sree Narayana Gurukulam College of Engineering, Kerala, India[2]

ABSTRACT: Cloud storage provides a convenient, massive, and scalable storage at low cost, but data privacy is a major concern that prevents users from storing files on the cloud trustingly. One way of enhancing privacy from data owner point of view is to encrypt the files before outsourcing them onto the cloud and decrypt the files after downloading them. However, data encryption is a heavy overhead for the mobile devices, and data retrieval process incurs a complicated communication between the data user and cloud. Normally with limited bandwidth capacity and limited battery life, these issues introduce heavy overhead to computing and communication as well as a higher power consumption for mobile device users, which makes the encrypted search over mobile cloud very challenging. The proposed scheme, An Efficient Searching Scheme with Data Integrity and Anonymization on Mobile Cloud), a bandwidth and energy efficient encrypted search architecture over mobile cloud. This architecture offloads the computation from mobile devices to the cloud, and optimizes the communication between the mobile clients and the cloud. It is demonstrated that the data privacy does not degrade when the performance enhancement methods are applied.

KEYWORDS:Data Integrity, Anonymization.

I. INTRODUCTION

Cloud storage system is a service model in which data are maintained, managed and backup remotely on the cloud side, and meanwhile data keeps available to the users over a network. Mobile Cloud Storage (MCS) denotes a family of increasingly popular on-line services, and even acts as the primary file storage for the mobile devices. MCS enables the mobile device users to store and retrieve files or data on the cloud through wireless communication, which improves the data availability and facilitates the file sharing process without draining the local mobile device resources.

The data privacy issue is paramount in cloud storage system, so the sensitive data is encrypted by the owner before outsourcing onto the cloud, and data users retrieve the interested data by encrypted search scheme. In MCS, the modern mobile devices are confronted with many of the same security threats as PCs, and various traditional data encryption methods are imported in MCS. However, mobile cloud storage system incurs new challenges over the traditional encrypted search schemes, in consideration of the limited computing and battery capacities of mobile device, as well as data sharing and accessing approaches through wireless communication. Therefore, a suitable and efficient encrypted search scheme is necessary for MCS.

The mobile cloud storage is in great need of the bandwidth and energy efficiency for data encrypted search scheme, due to the limited battery life and payable traffic fee. Therefore, focus on the design of a mobile cloud scheme that is efficient in terms of both energy consumption and the network traffic, while keep meeting the data security requirements through wireless communication channels.

To this end, introduce TEES (Traffic and Energy saving Encrypted Search) architecture for mobile cloud storage applications. TEES achieves the efficiencies through employing and modifying the ranked keyword search as the encrypted search platform basis, which has been widely employed in cloud storage systems. Traditionally, two categories of encrypted search methods exist, that can enable the cloud server to perform the search over the encrypted data: Ranked keyword search and Boolean keyword search. The ranked keyword search adopts the relevance scores to

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

represent the relevance of a file to the searched keyword and sends the top-k relevant files to the client. It is more suitable for cloud storage than the Boolean keyword search approaches, since Boolean keyword search approaches need to send all the matching files to the clients, and therefore incur a larger amount of network traffic and a heavier post-processing overhead for the mobile devices.

Cloud computing is defined as a type of computing that relies on sharing computing resources rather than having local servers or personal devices to handle applications. The goal of cloud computing is to apply traditional supercomputing, or high-performance computing power, normally used by military and research facilities, to perform tens of trillions of computations per second, in consumer-oriented applications such as financial portfolios, to deliver personalized information, to provide data storage or to power large, immersive online computer games. To do this, cloud computing uses networks of large groups of servers typically running low-cost consumer PC technology with specialized connections to spread data-processing chores across them. This shared IT infrastructure contains large pools of systems that are linked together. Often, virtualization techniques are used to maximize the power of cloud computing. Cloud computing has started to obtain mass appeal in corporate data centers as it enables the data center to operate like the Internet through the process of enabling computing resources to be accessed and shared as virtual resources in a secure and scalable manner.

Data center refers to on premise hardware that stores data within an organization's local network. Data centers may exist in physical environment or virtually and can be organized as a public data center for large scale usage or a private data center specific to an organization. While cloud services are outsourced to third-party cloud providers who perform all updates and ongoing maintenance, data centers are typically run by an in-house IT department. Cloud storage moves the user's data to large data centers, which are remotely located. However, this unique feature of the cloud poses problems like data integrity, data security etc.

Data integrity in simple terms can be understood as the maintenance of intactness of any data during transactions like transfer, retrieval or storage. Data integrity can be defined as ensuring that the data is unaltered, correct and consistent. The data may change if and only if an authorized operation is valid on the data. Integrity of the data can be hampered at any level of storage, any type of factor being the reason. Therefore, for the same reason, integrity surveillance in cloud storage is the most critical issue for any data center. Ensuring data integrity requires a bond of trust between the client and the provider. The conspicuous factor in ensuring the integrity of the generated data is having trust on the process generating that data.

II. RELATED WORK

D. Song et al [1]; proposed Practical Techniques for Search on Encrypted Data, it is desirable to store data on data storage servers, but this implies data security. This paper proposes cryptographic schemes for data security. In this technique, it supports searches on encrypted data. It uses primitives from classical symmetric-key cryptography to define security. The proposed scheme encrypts each word in the document separately. This encrypted search improves the confidentiality. But one who can able to use statistical technique to learn important information. To overcome the problem, periodically change the key and re-encrypt the document. This will be a burden for mobile devices and this file encryption method is not compatible with the existing scheme and cannot deal with data. The techniques provide provable secrecy for encryption, in the sense that the untrusted server cannot learn anything about the plaintext given only the cipher text. The techniques provide controlled searching, so that the untrusted server cannot search for a word without the user's authorization. The techniques support hidden queries, so that the user may ask the untrusted server to search for a secret word without revealing the word to the server. The techniques also support query isolation, meaning that the untrusted server learns nothing more than the search result about the plaintext. The algorithms in this technique are simple and fast. But the algorithm for encryption is not efficient, not compatible and cannot deal with data compression.

J. Ramos [2] proposed TF-IDF to Determine Word Relevance in Document Queries, examine the result of applying Term Frequency to determine what words in a corpus of documents might be more favorable to use in a query. Word with high TF-IDF numbers imply a strong relationship with the document they appear in. This simple algorithm efficiently categorizes relevant words that can enhance query retrieval. The main advantage of TF-IDF is an

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

efficient and simple algorithm for matching words in a query to documents that are relevant to that query. From the data collected, TF-IDF returns documents that are highly relevant to a particular query. If a user were to input a query for a particular topic, TF-IDF can find documents that contain relevant information on the query. Furthermore, encoding TF-IDF is straightforward, making it ideal for forming the basis for more complicated algorithms and query retrieval systems. Despite its strength, TF-IDF has its limitations. In terms of synonyms, notice that TF-IDF does not make the jump to the relationship between words. For large document collections, this could present an escalating problem.

D. Boneh et al [3]; proposed Public Key Encryption with Keyword Search. This paper studies the problem of searching on data that is encrypted using a public key system with secure searching scheme. This system known as non-interactive public key encryption with keyword search or searchable public key encryption. This scheme implies identity based encryption through PEKs seems to be harder to construct. But the converse is currently an open problem. Based on recent IBE construction, it provides security by exploiting extra properties.

Y. Chang et al [4]; proposed Privacy Preserving Keyword Searches on Remote Encrypted Data. This paper proposes single keyword search remotely. The users save files in an encrypted form on remote file server. Then efficiently retrieve some of encrypted files containing specific keywords. The keywords kept secret and search using keyword. This search will not affect the security of remotely stored files. This scheme proposes patterns matching techniques instead of keyword matching. The user can submit new files, are secure against previous queries but still searchable against future queries. It does not deal with multiple keywords. This technique increases storage overhead on server side and it deals with occurrence queries in a less efficient ways.

A Swaminathan et al [5]; proposed Confidentiality Preserving Rank Ordered Search. This technique introduces a new framework for confidentiality preserving rank ordered search and retrieval over large document collections. Cryptographic encryption protects data from compromise due to theft or intrusion. In addition to outsider attacks, security measures should also be taken against potential insider attacks. To accomplish goals, collect term frequency information for each document in the collection to build indices, as in traditional retrieval systems for plaintext. Further secure these indices that would otherwise reveal important statistical information about the collection to protect against statistical attacks. During the search process, the query terms are encrypted to prevent the exposure of information to the data center and other intruders, and to confine the searching entity to only make queries within an authorized scope. Utilizing term frequencies and other document information, apply cryptographic techniques such as order-preserving encryption to develop schemes that can securely compute relevance scores for each document, identify the most relevant documents, and reserve the right to screen and release the full content of relevant documents. The proposed framework not only protects document/query confidentiality against an outside intruder, but also prevents an un-trusted data center from learning information about the query and the document collection. They present techniques for integration of relevance scoring methods and cryptographic techniques such as order preserving encryption, to protect data collection and indices and provide efficient and accurate search capabilities to securely rank order documents in response to a query.

J. Oberheide et al [6]; proposed Virtualized In-Cloud Security Services for Mobile Devices. This paper proposes a model whereby mobile antivirus functionality is moved to an off device network service employing multiple virtualized malware detection engines. They propose an architecture that consists of two primary components, light weight host agent and network services. The host agent just like existing antivirus software, the host agent is a lightweight process that runs on each device and inspects file activity on the system. Access to each file is trapped and diverted to a handling routine which begins by generating a unique identifier (such as a hash) of the file and comparing that identifier against a cache of previously analyzed files. If a file identifier is not present in the cache, then the file is sent to the in-cloud network service for analysis. The network service is responsible for file analysis. The task of the network service is to determine whether a file is malicious or unwanted. The use of virtualization allows the network service to scale to large numbers of engines and users. If demand for a particular engine increases, more instances of that container can be spun up to service analysis requests. The proposed architecture could be deployed by a mobile services provides or third party vendor. This is an extension of the existing cloud AV platform. The result of this approach demonstrates the current model of on device antivirus software is not scalable. As the number and complexity

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

of mobile threats increases, on-device engines and their signature database will require more processing, storage and power.

A Boldyreva et al [7]; proposed Order Preserving Symmetric Encryption describes deterministic encryption scheme. They propose a security notion in the spirit of pseudorandom functions (PRFs) and order preserving constraint. Then design the efficient OPE scheme and prove its security based on pseudo randomness of an underlying block cipher. Encryption function preserves numerical ordering of the plaintext. OPE allows indexing and query processing to be done exactly and as efficiently for unencrypted data. This technique permits efficient range queries and proposes security.

K. Kumar et al [8]; proposed Cloud Computing for Mobile Users can Offloading Computation save Energy?, define 4 approaches to saving energy and extending battery life time in mobile devices,

1. Adopting new generation of semiconductor technology.
2. Avoiding energy wastage.
3. Executes programs slowly.
4. Eliminating computation all together.

Offloading of computation saves energy. The disadvantage of offloading is the high risk of privacy and security. The computation offloading depends on the wireless network communication, which will cause reliability concern. The wireless communication prevented by limited connectivity and low energy efficiency. Data storage is another reliability problem. The services should consider the energy overhead for privacy security, reliability and data communication before offloading. If the data processed are in smaller size, sending processed data to the server reduces the wireless transmission energy. One possible privacy and security solution when offloading is to use a technique called steganography and also use encryption.

A Carrol et al [9]; proposed An Analysis of power consumption in a smart phone, to measure the power consumption and energy efficiency of mobile devices. This paper presents the power consumption rate of different usage scenarios and analyses the contribution of different components in the mobile devices to overall power consumption. For mobile phones, power from batteries which are limited in size therefore capacity. Good energy management requires a good understanding of where and how the energy is used. They measure not only overall system power, but the exact breakdown of power consumption by the device's main hardware components and present the power breakdown for micro benchmarks. They develop a power model of the free runner device and analyze the energy usage and battery lifetime under a number of usage patterns, the significance of the power drawn by various components and identify the most promising areas to focus on for further improvements of power management.

S. Kamara et al [10]; proposed Cryptographic Cloud Storage to address the issues like confidentiality and integrity and increases the adoption of cloud storage, they support a virtual private storage service based on cryptographic technique. The architecture of cryptographic storage service consist of three components:- Data processor, data verifier and token generator. The core properties of cryptographic storage services are, control of the data is maintained by the customer and the security properties are derived from cryptography. To implement the core component, uses asymmetric searchable encryption (AES). Efficient AES schemes are appropriate in any setting, where the party that searches over the data is different from the party that generates it and where the keywords are hard to guess. The main advantage is efficiency and disadvantage is vulnerable to dictionary attacks.

III. PROPOSED ALGORITHM

The proposed system ensure data integrity, data anonymization and multi keyword search.

A. DATA INTEGRITY CHECKING

The integrity verifier has two options: either check integrity of original data in secure enclave or check integrity of published data in cloud server. To verify integrity of original data in secure enclave, it uses a private key Pk and secret key Sk. The public key known to everyone and secret key kept by client. For each record computes verification tag and makes it publicly known to everyone. The verifier sends a challenge to the secure enclave to check integrity of original

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

data. When receives challenge then it computes the responds. When receives the response it checks the validity then compare with the responds, the output either success or failure.

B. DATA ANONYMIZATION

To anonymize data verifier uses private key and secret key. For each anonymized data the function computes a verification tag and make it publicly known to everyone. The verifier sends a challenge to server to check integrity of published data. When server receives the challenge, the server computes the value for each anonymized data and sends to the secure enclave. The enclave receives values from server and generates the responds. When receives the response it checks the validity then compare with the responds, the output either success or failure.

C. MULTI KEYWORD SEARCH

Multi-keyword searching is done by using apriori algorithm. It proceed by identifying the frequent individual items in the database. This algorithm generates frequent itemset and which helps to reduce the searching time for a frequently searching item.

IV. IMPLEMENTATION

The project has been implemented in cloudera. The project divided into 3 modules – Data owner, data user and Cloud server.

Cloudera provides enterprise support and professional training for Hadoop.Cloudera is revolutionizing enterprise data management by offering the first unified Platform for big data, an enterprise data hub built on Apache Hadoop. Cloudera offers enterprises one place to store, access, process, secure, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Cloudera's open source big data platform is the most widely adopted in the world, and Cloudera is the most prolific contributor to the open source Hadoop ecosystem. Cloudera provides its own take on Hadoop's powerful open-source data management software and couples it with IT support and management. Using Hadoop, enterprises can store and process huge amounts of unstructured data. But Hadoop can often be unwieldy and difficult to manage, so Cloudera helps make it possible to manage that data effectively. Cloudera is regarded by many industry experts as the benchmark for Hadoop adoption and commercial success. CDH (Cloudera distribution hadoop) is Cloudera's 100% open source Hadoop distribution, built specifically to meet enterprise demands. Comprised of Apache Hadoop and a number of other leading open source projects, CDH combines storage and computation into a single, scalable system and delivers the flexibility and economics required to perform operations on Big Data that are not possible with traditional solutions due to time or cost. CDH is the world's most complete, tested, and popular distribution of Apache Hadoop and is designed for the enterprise. By integrating Hadoop with more than a dozen other critical open source projects, Cloudera delivers a functionally advanced system that helps to gain value from all data.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Fig.1 Cloudera

Modules:

The main purpose of the project is to implement traffic and energy efficient multi-keyword searching scheme with data integrity checking and anonymization. In order to achieve multi-keyword searching, data integrity and anonymization, implemented different modules.

1. Searching

An authenticated user stems the keyword to be queried, encrypts it with the keys and hashes it to get its entry in the index. Then the encrypted keyword is sent to the cloud server. On receiving the encrypted keyword, the cloud server first searches for it in the index. Then the index related to this keyword is sent back to the data user. The data user calculates the relevance scores with the selected index to find the top-k relevant files and sends a follow-up request to the cloud server in order to retrieve the files. The position of these files is selected and they are sent back to the data user from the cloud server. The data user decrypts the files and recovers the original data. This scheme incur Two Round trip Search (TRS). The searching is implemented using the apriori algorithm. Which helps multi-keyword search more efficient and fast.

2. Data Integrity

An authenticated user upload files to the cloud as public or private. At the time of file upload, checks the integrity value. At the time of downloading again checks the integrity of files and then compare with the former, if the value doesn't matches the file integrated by intruders.

3. Data anonymization

The key for encryption is converted to another format for more security.

V. IMPLEMENTATION RESULTS

The proposed system has been implemented using cloudera stimulator and the programs are written in java language. The system uses an encrypted multi keyword search overcloud environment. The searching is done by using apriori algorithm(see fig 2). Data integration helps to ensure the data is unaltered, correct and consistent. Data

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

anonymization prevents the identification of the key information used for the encryption. The computation cost of the client, enclave and server are analyzed. At the client side computation cost include integrity checking. At secure enclave computation cost include the data encryption, decryption, anonymization and de anonymization. At the server computation cost include the proof generation. From the analysis, at the client and server the computation cost is very less than compared with the enclave. The low cost in client and server is applicable for the thin devices like mobile.

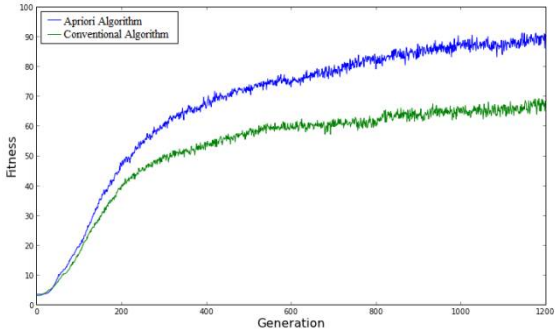

Fig 2. Comparison

VI. CONCLUSION AND FUTURE WORK

The proposed system TEES as an initial attempt to create a traffic and energy efficient encrypted keyword search tool over mobile cloud storages. The security study of TEES showed that it is secure enough for mobile cloud computing, while a series of experiments highlighted its efficiency. TEES is slightly more time and energy consuming than keyword search over plain-text, but at the same time it saves significant energy compared to traditional strategies featuring a similar security level. Based on TEES, this work can be extended to more other novel implementations.

For secure cloud computing, clients have to ensure that their data stored in cloud servers are not lost or corrupted. When encryption is used as a means of ensuring privacy, statistical computation on published data becomes a difficult task. Even though homomorphic encryption is a strong encryption technique which supports computations without decryption, it is computationally too intensive for practical use. Another technique called data anonymization can be used in cloud computing which makes processing of cloud data possible. Using data anonymization, the key pieces of confidential data are obscured but still data can be processed to get useful information. A better data anonymization scheme is proposed in cloud computing where anonymization and deanonymization is performed by the secure enclave which in turn saves the computational power at the client. Anonymization techniques used in the experiment ate generalization, suppression, k-anonymity and l-diversity which enhances the privacy of data to a larger extent. Also, a remote integrity checking protocol to check the integrity of anonymized data in cloud environment is explored.

REFERENCES

[1] D. Song, D. Wagner, and A. Perrig, “Practical techniques for searches on encrypted data,” in Security and Privacy, 2000. S&P 2000. Proceedings. 2000 IEEE Symposium on. IEEE, 2000, pp. 44– 55.
[2] J. Ramos, “Using tf-idf to determine word relevance in document queries,” Technical report,Department of Computer Science, Rutgers University, 2003.
[3] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano, “Public key encryption with keyword search,” in Advances in Cryptology- Eurocrypt 2004. Springer, 2004, pp. 506–522.
[4] Y. Chang and M. Mitzenmacher, “Privacy preserving keyword searches on remote encrypted data,” in Applied Cryptography and Network Security. Springer, 2005, pp. 391–421.
[5] A. Swaminathan, Y. Mao, G. Su, H. Gou, A. Varna, S. He, M. Wu, and D. Oard, “Confidentiality-preserving rank-ordered search,” in Proceedings of the 2007 ACM workshop on Storage security and survivability. ACM, 2007, pp. 7–12.
[6] J. Oberheide, K. Veeraraghavan, E. Cooke, J. Flinn, and F. Jahanian, “Virtualized in-cloud security services for mobile devices,” in Proceedings of the First Workshop on Virtualization in Mobile Computing. ACM, 2008, pp. 31–35.
[7] A. Boldyreva, N. Chenette, Y. Lee, and A. O´ neill, “Orderpreserving symmetric encryption,” Advances in Cryptology- EUROCRYPT 2009, pp. 224–241, 2009.
[8] K. Kumar and Y. Lu, “Cloud computing for mobile users: Can offloading computation save energy?” Computer, vol. 43, no. 4, pp. 51–56, 2010.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

[9] A. Carroll and G. Heiser, "An analysis of power consumption in a smartphone," in Proceedings of the 2010 USENIX conference on USENIX annual technical conference. USENIX Association, 2010, pp. 271–284.
[10] S. Kamara and K. Lauter, "Cryptographic cloud storage," in Financial Cryptography and Data Security. Springer, 2010, pp. 136– 149.
[11] A.Miettinen and J. Nurminen, "Energy efficiency of mobile clients in cloud computing," in Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, 2010, pp. 21–28.
[12] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure ranked keyword search over encrypted cloud data," in Distributed Computing Systems (ICDCS), 2010 IEEE 30th International Conference on. IEEE, 2010, pp. 253–262.
[13] C. Wang, N. Cao, K. Ren, and W. Lou, "Enabling secure and efficient ranked keyword search over outsourced cloud data," Parallel and Distributed Systems, IEEE Transactions on, vol. 23, no. 8, pp. 1467–1479, 2012.
[14] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," Parallel and Distributed Systems, IEEE Transactions on, vol. 25, no. 1, pp. 222–233, 2014.
[15] J. Li, R. Ma, H. Guan, "TEES: An Efficient Search Scheme over Encrypted Data on Mobile Cloud," IEEE Transactions on Cloud Computing, 2015.
[16] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z.Peterson and D. Song, "Provable data possession at untrusted stores", Proceedings of the 14th ACM conference on Computer and communications security, CCS'07 ,New York, USA, ACM, 2007, pp. 598-609.
[17] A. Juels and B.S. Kaliski, Jr., "Pors: proofs of retrievability for large files," in CCS˝07: Proceedings of the 14th ACM conference on Computer and communications security.
[18] A.M. Talib, R.Atan, R.Abdullah and M.A. Azmi Murad, "CloudZone: Towards an integrity layer of cloud data storage based on Multi Agent System Architecture", 2011 IEEE Conference on open systems (ICOS 2011), September 25-28 2011.
[19] Salah H. Abbdal, Hai Jin, Deqing Zou, Ali. A. Yassen ,"Secure Third Party Auditor for Ensuring Data Integrity in Cloud Storage ", 14th Intl Conf on Scalable Computing and Communications and Associated Symposia/Workshops 2014.
[20] Reenu Sara George, Sabitha. S "Data Anonymization and Integrity Checking in Cloud Computing" , 4th ICCCNT 2013