A Survey on Cloud Computing Techniques for Data Integrity Checking

Dr. Rahma Teirab Abaker Haroun
Computer Science & IT Dept., Nyala Technical College, Sudan Technical University, Sudan
Dept. of Math, University College in Alkhafji, University of Hafr Al Batin, Saudi Arabia
E-mail. toha124@gmail.com

Dr. Ahmed Khameis Sharaf Eldein AlKabour
Department of ET, Nyala Technical College, Sudan Technical University, Sudan
'E-mail. kabour2006@hotmail.com

Contact number. 00966543218226

Abstract

Cloud computing is a term that used instead of internet to describe the infrastructure, software services and storage via internet. Large data centers are available in cloud for remotely store user data. The users have no data control privileges when the data transferred to the Cloud and they are not aware of any security risk. Data can be altered by unauthorized user, threats and dishonest server. Farther more, Data which are either unused for a long a time or takes large memory space can be deleted by cloud service provider. The main issue of cloud computing today is data integrity and how can be maintaining. There for, security challenges users are need to ensure that their data are integral by periodically Data integrity checking. Several integrity checking techniques have been proposed to ensure the data integrity in cloud storage. This paper provides a survey of various data integrity checking techniques for cloud data stored. Objective of this survey focusing on existing integrity check techniques for cloud data storage and presenting their characteristics, benefits, functionality and limitations.

Keyword *- Cloud computing, data integrity, integrity checking.*

1. Introduction

In data security field the way of keeping data secure are and altered only by authorized; called Data Integrity. Cloud computing is a general term for the delivery of hosted services over the internet.as well as it is a network where user can use services.

Cloud Computing enables companies to consume a compute resource, such as a virtual machine (VMs), storage or an application, as a utility rather than having to build and maintain computing infrastructures in house. So that it has gained wide acceptance for organizations as well as individuals by introducing computation, storage and services. Also it has been visualized as the second generation architecture of IT enterprise due to its matchless benefits in IT history:

- On demand self-services
- Ever-present network access
- Location independent resource pooling
- Repaid resource elasticity
- Usage-based pricing and risk
- Stop spending money on running and maintaining data centers
- Increase speed and agility
- Stop guessing capacity

Definition of cloud Computing: 1.1

According to Hewitt, C. cloud computing is defined as a next generation computing model for enabling convenient, efficient, on-demand network access to a shared pool of configurable computing resources[1]. The growing need of Technology in every field has led to the evolution of cloud computing for highly efficient usage of IT resources.

Cloud Storage is an important service of cloud computing, as it allows data owners to move their data remotely. More and more data owners start choosing to host their data in the cloud.

The concept of Cloud Computing is one of the major theories in the world of IT. Its services are now being applied to several IT scenarios. Cloud Computing is the internet based computing which provides users with a number of services. Users store their data in the cloud without the burden of local data storage [8]. Cloud Computing gained intention since 2007. It is the general term for anything that involves providing services on internet. It moves the data and computing from desktop to large datacenters. It is combination of parallel, grid and distributed computing.

Cloud Computing Service Categories: 2

Although cloud computing has changed over time, it has been divided into three broad service categories:

4.2.1 Infrastructure as a Service (IaaS)

IaaS providers supply a virtual server instance and storage, as well as application program interfaces (APIs) that let users to migrate workloads to a virtual machine. Users have an allocated storage capacity and can start, stop, access and configure the VM and storage as desired. IaaS providers offer small, medium, large, extra-large and memory- or compute-optimized instances, in addition to customized instances, for various workload needs.

2.2 Platform as a Service (PaaS)

In the PaaS model, providers host development tools on their infrastructures. Users access these tools over the internet using APIs, web portals or gateway software. PaaS is used for general software development, and many PaaS providers will host the software after it's developed.

2.3 Software as a Service (SaaS).

SaaS is a distribution model that delivers software applications over the internet; these applications are often called web services. Microsoft Office 365 is a SaaS offering for productivity software and email services. Users can access SaaS applications and services from any location using a computer or mobile device that has internet access.

3 Cloud Computing (Types) Deployment Models:

3.1 Public Cloud:

In a public cloud the computing infrastructure is used by the organization or end user through cloud service providers or vendors. Public clouds are typically offered through virtualization and distributed among various physical machines [4].

3.2 Private Cloud:

In a private cloud the computing infrastructure is dedicated to the particular organizations and not shared with other organization. Private clouds are more secure than public clouds [4] [13].

3.3 Hybrid Cloud:

This is a combination of the other two types of cloud. In hybrid cloud organizations may host critical application on private clouds and applications which are having less security concerns hosted on public clouds. It is also known as cloud bursting [13].

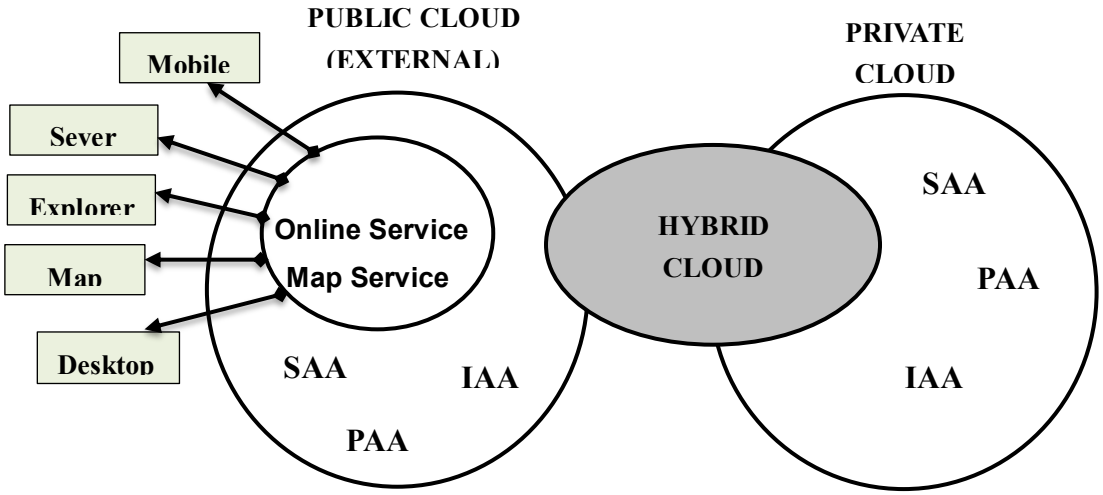

Fig 1. Cloud Types (Deployment Models)

4 Challenges in Cloud Computing:

As cloud provides many advantages but as every coin has 2 side, and cloud computing is no exception, it also has certain challenges. Every day, a fresh news item, latest publication, blog entry, highlights the cloud computing's challenges and issues. In each technology there are some security issues that affect the usage and the behavior; below some of these concerns in cloud computing [2]:

4.1 Access: When there is an unauthorized access to the data, the ability of altering on the client data arise.

4.2 Availability: The data must be available all the time for the clients without having problems that affect the storage and lead to the client data lose.

4.3 Network Load: The over load capacity on the cloud may drop the system out according to the high amount of data between the computers and the servers.

4.4 Integrity: The data correctness, legality and security is the most fields that influence on the cloud and have major lay on the service provider.

4.5 Data Location: The client does not know the actual place that the data saved or centered in because it distributed over many places that led to confusion.

5 Data Integrity

Integrity, in terms of data security is the guarantee that data can only be accessed or modified by those authorized to do so, in simple word it is process of verifying data [2]. Data Integrity is very important among the other cloud challenges. As data integrity gives the guarantee that data is of high quality, correct, unmodified. After storing data to the cloud, user depends on the cloud to provide more reliable services to them and hopes that their data and applications are in secured manner. But that hope may fail, the user's data may be altered or deleted. Sometimes, the cloud service providers may be dishonest and they may discard the data which has not been accessed or rarely accessed to save the storage space or keep fewer replicas than promised [6]. Moreover, the cloud service providers may choose to hide data loss and claim that the data are still correctly stored in the Cloud. As a result, data owners need to be convinced that their data are correctly stored in the Cloud. So, one of the biggest concerns with cloud data storage is that of data integrity verification at untrusted servers [13].

As a result data owners need to be persuaded that their data. Thus, the most major anxiety with cloud data storage is data integrity verification at untrusted server. One of the important concerns in the cloud computing that need to be addressed is to assure the customer of the integrity, accordingly in the next section will discuss data integrity Privacy Techniques. In the world of cloud computing the data integrity is most challenging and burning security issue. In order to solve the problem of data integrity checking, many researchers have proposed different systems and security models.

6 The Privacy Techniques for Data Integrity:

Storing data in the cloud has become a trend [14]. In Cloud computing the issue of data integrity is still carried out by many researchers see the figure.

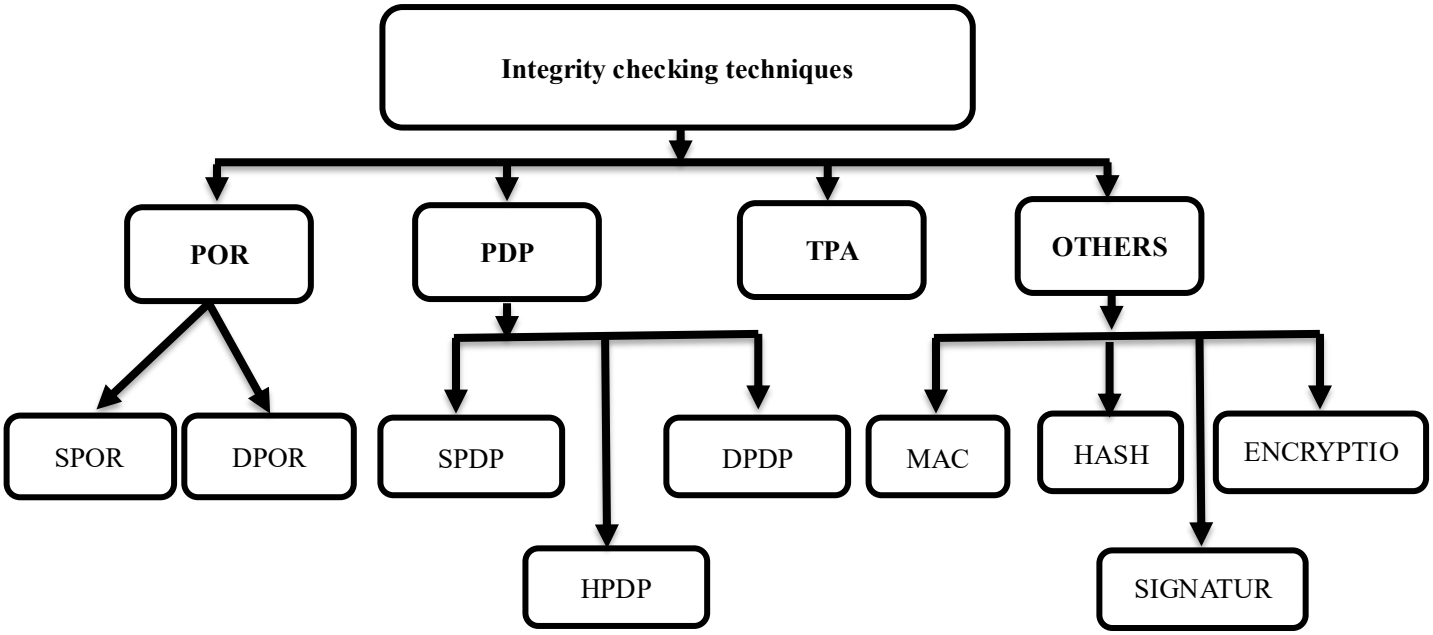

Fig 2. The Privacy Techniques for Data Integrity

6.1 Provable Data Possession (PDP)

It's a technique for assuring data integrity over remote servers. In PDP A client that has stored data at an unfaithful server can verify that the server possesses the original data without retrieving it. Ateniese et al. the first to consider public audit ability in their defined "provable data possession" model for ensuring possession of files on untrusted storages [3][7]. The client pre-computes tags for each block of a file. The client can verify that the server possesses the file by generating a random challenge against a randomly selected set of file blocks. Using the queried blocks and their corresponding tags, the server generates a proof of possession.

1) Static PDP: One of the methods is based on hashing and having a key system [2].

2) Dynamic PDP: Dynamic PDP approves dynamic operations such as modifying, deleting, insertion etc. Scalable PDP is a method which uses the symmetric encryption. Cloud tenant challenges CSP server with a set of random looking block indices [1].

6.1.1 Principal of PDP

The working principal of PDP is as shown in fig 1. It works in two stages. Set up stage and challenge stage [4].

A) Set up stage:

- The client generates pair of matching keys public & secrete key by using probabilistic key generation algorithm.
- Public key along with the file will be sent to the server for storage by client and he deletes the file from its local storage.

B) Challenge stage:

- The client challenges the server for a proof of possession for a subset of the blocks in the file.
- The client checks the response from the server.

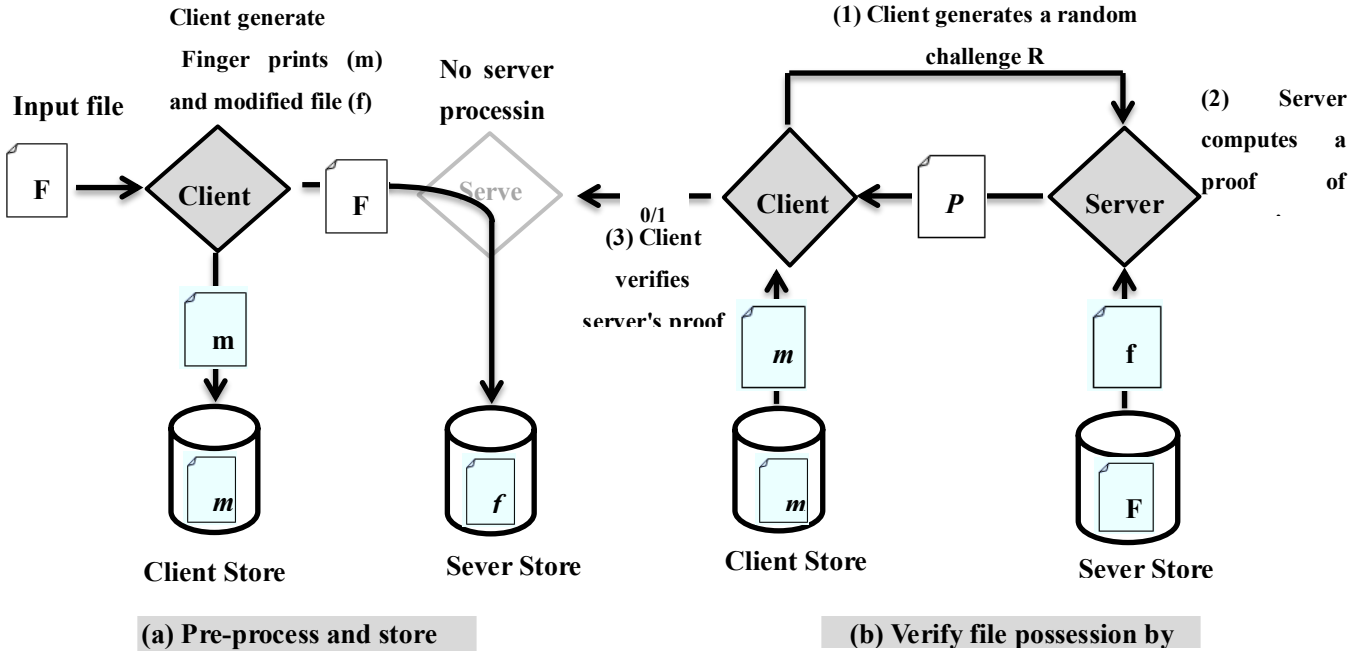

Fig 3. Protocol for Provable Data Possession

Advantages:

- The server does not actually have to access the file blocks.
- support public verification.

Limitations:

- Absence of privacy and dynamic support
- number of queries are limitless

6.2 PDP Based on MAC (Message Authentication Code)

In this technique the client (data owner) computes a MAC of the whole file with a set of secret keys and store them locally before outsourcing it to cloud service provider (CSP) [3]. The client keeps only the computed MAC on local storage then sends the file to the CSP and deletes the local copy of the file. For verifying the data integrity of the file, a request should be sent so as to retrieve the file from CSP, detect

a secret key to cloud server and asks to re-compute the MAC of the whole file and compare with previously stored value.

Limitations:

- Number of verification is limited by number of secret keys.
- The owner has to retrieve the entire file to compute new MACs, which is not available for large file.
- Public auditability is not supported as private keys are required for verification.
-

6.3 Scalable PDP:

Scalable PDP its improved version of original PDP and uses symmetric encryption also support dynamic operation on remote data. It has all challenges and answers are pre-computed as well as limited numbers of updates. Scalable PDP does not require bulk encryption. It relies on the symmetric-Key which is more efficient than public-Key encryption. So it does not offer public verifiability.

Limitations:

- A client can perform limited number of updates and challenges.
- It does not perform block insertions; only append-type insertions are possible.
- This scheme is problematic for large files as each update requires re-creating all the remaining challenges.

6.4 Dynamic PDP:

Dynamic PDP is a collection of seven polynomial-time algorithms (KeyGen DPDP, PrepareUpdate DPDP, PerformUpdate DPDP, VerifyUpdate DPDP, GenChallenge DPDP, Prove DPDP, and Verify DPDP). It supports full dynamic operations like insert, update, modify, delete etc. Here in this technique uses rank-based authenticated directories and along with a skip list for inserting and deleting functions. .It has some computational complexity, it is still efficient. For example, for verifying the proof for 500MB file, DPDP only produces 208KB proof data and 15ms computational overhead.

Limitations:

- It has some computational complexity.
- Not suitable for thin client.
- DPDP does not include provisions for robustness.

6.5 Basic Proof of Retrievability (POR)

POR it's a cryptographic method for remotely verifying of integrity of files stored in the cloud without keeping a copy of original files in local storage. The users are checks integrity using authentication key without retrieving back the data file [3]. In this method, user backup's data file together with the authentication data to a potentially dishonest cloud storage server. User can check the data for its integrity

stored with CSP using the authentication key, without retrieving back the data file from cloud possession [5]. Everything in cloud computing provided as services; Even the storage provided as service; for example amazon storage services.

6.5.1 Principles of POR:

A POR works on two phases first is setup phase and another is sequence of verification phases.

A) Setup phase:

- User generate authentication code by using private key
- Sends data with authentication code to the clod and remove them locale
- User has his private key locally and CSP has both data + authentication code.

B) Sequence of verification phase:

- User generate a random challenge query and CSP produce response or proof for challenge query base on user's data file and authentication code .
- Users verify the response using his private key and decide to accept or reject this response.

Limitations:

- Support limit numbers of queries, since it deals with a finite number of check blocks.
- A POR does not provide in prevention to the file stored on CSP.

6.6 POR based on keyed hash function hk(F)

It is very simple and effortlessly implementable technique. The user pre-computes the cryptographic hash of F using hk (f) before sending data file to cloud storage and store secret key K with computed hash. The user releases the secret key K to the CSP to check the integrity of the file F and asks it to compute and return the value of hk (F)[4] [8]. If the user want to check the integrity of the file F for multiple times he has store multiple hash values for different keys.

Limitation:

- For every check the user should store the key as well as hash value.
- It requires higher resource cost for implementation at every time hashing to perform on entire file.
- Hash value computation for large file can be burdensome for thin client

4.6.7 Proof of Retrievability for Large file

POR for large files using guards and only a single key can be used regardless to the file size or number of files. Here, the user needs to access small part of the file F "this part in fact separated from the original

length of F", also special guard's blocks which are hidden among other blocks in data file are random attached among the data blocks [4]. To check the integrity of data file F, user as CSP during verification phase by specifying the position of a collection sentinels and ask CSP to return the associated sentinels values; if the CSP has modified or delete some part of F then the position of sentinels should be changed. The encryption is performed on whole modified file to in distinguish the sentinels from the data blocks, and stored in the CSP [9].

Limitations:

- Work only on static data
- Computational over head for large file should be on whole file
- Storage overhead on the server due to new inserted sentinels
- Download of the whole file increase input/output and transmission cost

6.8 Hail:

Hail is a high availability and integrity layer for cloud storage which user can be able to store data on multiple servers. To ensure data integrity hail uses Message Authentication Code (MAC), pseudo random function and Hash function independent of file size.

Limitation:

- File corruption may occurred by threats attack.
- Applicable for static data and not suitable for thin client
- Require more computational power

6.9 POR based on Selection Random Bits in data Blocks:

In this technique just few data should be encrypted instead of the whole file so as to reducing the computational load, storage on client and bandwidth. A high probability of security can be achieved by encrypting fewer bits instead of whole data.

Advantages:

- Well for thin client
- Store only encrypted key and two random sequence functions
- No data store locally
- Users can append some metadata before sending a file to CSP
- This metadata can be used to verify data integrity

Limitation:

- Only for static data
- No data mechanism included

7 Comparison of integrity checking methods:

Technique	Advantages	Disadvantages	Method Used
Provable Data Possession	Support public verification and strong proof of data integrity	· Absence of dynamic support and privacy · Number of queries are limitless.	Algorithm for Key Generation
PDP based on MAC	· Secure and simple technique · Gives strong proof Integrity of Data.	· Limit number of verification · Not allow retrieve of large files	Message Authentication Code
Scalable PDP	· Support dynamic operation and not require bulk encryption · Does not offer public verifiability	· Limit number of update performed by client · Problematic for large files	Cryptographic Hash function and symmetric key encryption
Dynamic PDP	· Support full dynamic operations	· Some computational complexity. · Not suitable for thin client · Does not include provisions for robustness.	rank-based authenticated directories
Basic Proof of Retrievability (POR)	· Data file not retrieving back	· Not support dynamic data, only for static · Support limit numbers of queries	Encryption
POR based on keyed hash function	Effortlessly implementable technique	· Only for static data · Large numbers of keys · II. Requires high cost for computation	Hash Function
POR for Large file	· Ensures both possession and Retrievability of files on CSP.	· Only on static data · Storage overhead on the server · Download of the whole file increase transmission cost	sentinel-based scheme
POR based on Selection Random Bits in data Blocks	· No data store locally · Better for thin client · Store only encrypted key and two random sequence functions	· Only for static data · No data mechanism included	Meta Data Generation
Hail	· Allow clients to store data in multiple cloud	· No Data Prevention mechanism · For static data · Require more computational power	MAC, pseudo random function and Hash function

8 Conclusion

The main issues of any technology are Security and privacy. As the cloud is mainly used for data storage, data integrity and integrity checking are the major concern for the client.

After storing data to the cloud, clients are hope that their data and applications are in secured manner. But that hope may fail sometimes data may be changed or deleted. Also, dishonest CSP may discard the data which has been rarely accessed. In this paper several data integrity techniques and their methods for checking integrity of remotely stored data on cloud and their merits and demerits are explained. The analytical study briefly compares all this techniques. From this survey it is conclude that there is need to design secure, dynamic, efficient data integrity technique which is still wide area of research.

9 Acknowledgment

I would like to express my sincere thanks to my supervisor Prof. Saih.E.Fatoh, for his great efforts and encouraging for this work. I would like to thanks my friend Dr. Mokhtar.M.M. for his continuous guidance & support. I would like to thanks all my friends and family members for their support.

10 References:

[1] Kamile Nur Seviş , el [ ] , Survey on Data Integrity in Cloud, 2015

[2] Haichun Zhao,el[ ],A Survey on the Integrity Checking of Outsourced Data In Cloud Computing 2015.

[3] P.Parvathi, T.Meyyappan, Survey on Data Integrity Checking Protocols in Cloud Computing, International Journal of Computer Applications (0975 – 8887) International Conference on Computing and information Technology (IC2IT-2013)

[4] Charmee V. Desai, Prof. Gordhan B. Jethava, Survey on Data Integrity Checking Techniques in Cloud Data Storage, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 12, December 2014, ISSN: 2277 128X.

[5] Ms. R. K. Pandya, 2 Prof. K. K. Sutaria, Data Integrity Techniques in Cloud: An Analysis, Journal Of Information, Knowledge And Research In Computer Engineering Vol.02 Issue 02, (2013), ISSN: 0975 – 6760

[6] A. Methew, Security and Privacy Issues of Cloud computing: solution and secure framework, International Journal of Multidisciplinary Research Vol.2 Issue 4, (2012), ISSN 2231 5780

[7] Mahesh S.Giri and Bhupesh Gaur, Ph.D, Deepak Tomar, A Survey on Data Integrity Techniques in Cloud Computing, International Journal of Computer Applications (0975 – 8887) Volume 122 – No.2, July 2015

[8] B. Priyadharshini and P. Parvathi, "Data Integrity in Cloud Storage", in Proceedings of IEEE 2012.

[9] P. Metri and G. Sarote, Privacy Issues and Challenges in Cloud Computing, International journal of Advanced Engineering Sciences and Technologies, Vol. No. 5, Issue No. 1,001-006.

[10] Sanchika Gupta,Anjani Srdan, Padam kumar,Ajit Abraham,"A Secure and Lightweight Approach for Critical Data Security in Cloud" in roc.CASoN-IEEE Jan 2012

[11] Wenjun Luo, Guojing Bai,"Ensuring the Data Integrity in Cloud Data Storage" in Proc. CCIS-IEEE Jan 2011

[12] R. Sravan kumar and Saxena ,"Data integrity proofs in cloud storage" in Proceedings of IEEE 2011.

[13] MS. R. K. Pandya, Rof. K. K. Sutaria, Data Integrity Techniques in Cloud: An Analysis, Journal Of Information, Knowledge And Research Incomputer Engineering, Volume – 02, Issue – 02 Nov 12 To Oct 13, Issn: 0975 – 6760

[14] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599 – 616, 2009.