Comparative Study of Data Possession Techniques for Data Storage as a Service (DSaaS)

Parth D Shah
IT Department
CSPIT, CHARUSAT
Changa, INDIA

Amit P Ganatra, Ph.D
CE Department
CSPIT, CHARUSAT
Changa, INDIA

ABSTRACT

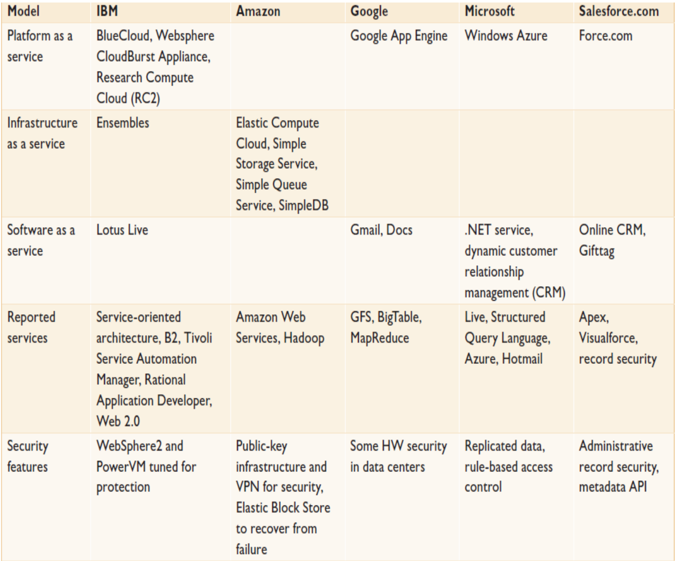
Cloud computing is a utility where software, infrastructure or storage can be purchased remotely. Users can access data from anywhere through network connectivity. In Cloud Computing resources are shared among different user so that all users can access the data from shared infrastructure via same physical hardware. However, this technique raises many concerns over the honesty of network users. Because of shared resources used in Cloud, user does not have permission to modify, delete or access of other user's data. Also users do not have control of their own data in cloud. So data integrity is one of the biggest security issues come in picture of Cloud Computing. For the Cloud service provider to provide integrity of user's data in cloud is extremely difficult task. To provide integrity of user data traditional integrity tools do not use because of dynamic nature of user data. Providing monitoring service will help data owners to ensure integrity and also provides a transparent yet cost-effective method for data owners to gain trust in the cloud. Provable Data Possession (PDP) allows data owner to periodically check their data stored in cloud storage are intact. In this method client can check without downloading the file as well does not required a local copy. Another method which could be considered as improvement over PDP is Proof of Retrivebility (POR). This method detects data integrity as well as it recovers original data. Both methods are probabilistic. As a part of this paper work data integrity techniques have been surveyed and some of the research gaps have been identified which is presented here.

Keywords

Cloud Computing, Data Storage, Secuority challenges, Integrity proofs, Remote Data Checking

1. INTRODUCTION

Many people are confused to exactly what cloud computing is, because it can be used to mean almost anything as client server application. Roughly, it can be described as a highly scalable computing resources provided as an external service via the internet on a pay-as-you-go basis. Reducing costs, accelerating processes and simplifying management are all vital to the success of an effective IT infrastructure. A federal technology agency National Institute of Standards and Technology (NIST) [1], as it covers all the essential aspects of cloud computing and gave the definition of cloud computing: *Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction*. The cloud is simply a metaphor for the internet, the symbol used to represent the worldwide network in computer network diagrams. Resources are available to be accessed from the cloud computing infrastructure at any time via the internet. Cloud users have no need to worry about how things are being managed and maintained behind the scenes – you simply purchase the IT service you require as you would any other utility. So in this concern, cloud computing has also been called utility computing, or 'IT on demand'. This new, network-based generation of computing utilizes remote servers hosted and housed in highly secure data centers for data storage management and maintenance, so organizations no longer need to purchase and look after their IT solutions in-house. The cloud model [1] is composed of five essential characteristics like On-demand self-provisioning, Resource pooling, Ubiquitous network access, Measured service, Rapid elasticity and scalability, three service models Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), Software-as-a-Service (SaaS), and the four deployment models hybrid cloud, community cloud public cloud, and private cloud,. Commercial products in the field of Cloud Computing are shown in Figure 1.



Model	IBM	Amazon	Google	Microsoft	Salesforce.com
Platform as a service	BlueCloud, Websphere CloudBurst Appliance, Research Compute Cloud (RC2)		Google App Engine	Windows Azure	Force.com
Infrastructure as a service	Ensembles	Elastic Compute Cloud, Simple Storage Service, Simple Queue Service, SimpleDB			
Software as a service	Lotus Live		Gmail, Docs	.NET service, dynamic customer relationship management (CRM)	Online CRM, Gifttag
Reported services	Service-oriented architecture, B2, Tivoli Service Automation Manager, Rational Application Developer, Web 2.0	Amazon Web Services, Hadoop	GFS, BigTable, MapReduce	Live, Structured Query Language, Azure, Hotmail	Apex, Visualforce, record security
Security features	WebSphere2 and PowerVM tuned for protection	Public-key infrastructure and VPN for security, Elastic Block Store to recover from failure	Some HW security in data centers	Replicated data, rule-based access control	Administrative record security, metadata API

Fig 1: Commercial products [3] in the field of Cloud Computing

2. DIFFERCENCE WITH THE EXISTING TECHNOLOGIES

In the world of IT one may get confuse, especially from a "business" perspective that how it differs from other related technologies. Often this confusion is brought about by IT professionals, mixing what is technology with a business concept / way in which to implement IT technology. Cloud-based software is similar to virtualization, web application, utility computing and grid computing in many ways but there are certain differences.

Virtualization is the big thing that always gets associated with the cloud or as the same thing. Virtualization (hardware or software) technology is being confused with a business solution / concept / way of implementing IT. The lines of difference is Virtualization – a technology whereas Cloud Computing – A way of working. Cloud computing is not a technology; rather it is a way in which we can use technology to decrease IT overheads (cost wise in theory). Virtualization – not just for the cloud but also used to Reducing administration cost, reducing hardware cost, Reducing electricity bills etc [3].

Cloud computing is lending a service as per the requirement of the user and web application is the platform through which the service is being used. Web application is just a small part of cloud computing that processing on server. Cloud computing is computer resource providing type of utility computing that is pay as you use. User can Buy or Get Software, Platform or even processing from cloud.

This difference between cloud computing and utility computing is substantial; since it reflects a difference in the way computing is approached. The utility computing relies on standard computing practices utilizes traditional programming styles in a well-established business context. Whereas Cloud computing involves creating an entirely distinctive virtual computing environment that empowers programmers and developers in new ways [4].

3. CHALLENGES IN CLOUD COMPUTING FROM BUSINESS PERSPECTIVE

From the business perspective there are some challenges in implementing cloud computing. The complexity of finding an optimum resource allocation is exponential, in systems like big clusters, Grids or data centers, is growing very quickly. Multiple, remote users can access the systems, and everyone has its own preferences, which can be in conflict with the preferences of the other users. The idea of business is being introduced: some providers will sell their resources to the users, which are willing to pay for accessing them. This introduces new high-level metrics: Quality of Experience, Quality of Business. It is very difficult to manage resources having into account these metrics because they can be different for every provider and client, and the central resource manager does not have to know what good Quality of Experience of a user is. If the central resource manager breaks, the whole system gets useless. This is a big waste of resources in large systems.

4. ISSUES IN CLOUD COMPUTING [5]

Security – It is the biggest concern matter for cloud data storage. Many people concern about the weakness of remote data to such hackers, dissatisfied or curious employees. Solution providers should extremely sensitive to this issue and apply substantial resources to mitigating concern.

Reliability – Some people worry also about whether a cloud service provider is financially stable and their data storage system is reliable. Most cloud providers attempt to appease this concern by using redundant storage techniques like backup, but it is still possible that a service could crash or go out of business, leaving users with limited or no access to their data.

Ownership – Once data has been transferred to the cloud, some people think that they could lose some or all of their privileges or be unable to guard the privileges of their customers. Many service providers are addressing this issue with well-crafted service level agreements.

Data Backup – Cloud providers employ backup servers and routine data backup processes, but some client worry about being able to regulate their own backups.

Data Portability and Conversion – Some people are concerned that, should they wish to change providers, they may have difficulty transferring data. Porting and converting data is highly reliant on the nature of the cloud provider’s data retrieval format, especially in cases where the format cannot be easily discovered. As competition grows and open standards become established, the portability issue will ease, and processes of conversion will become available supporting the more popular cloud providers. Worst case, a client will have to pay for some custom data conversion.

Multiplatform Support – More an issue for IT departments using managed services is how the cloud-based service integrates across different platforms and operating systems. Usually, some tailored adjustment of the service takes care of any problem.

Intellectual Property – A company invents something new and it uses cloud services as part of the discovery. Is the discovery still patentable? Does the solution provider have any claim on the discovery? Can they provide similar services to competitors?

5. SECURITY ISSUES IN CLOUD COMPUTING

Network-based application, storage and communication platforms have certain vulnerabilities in several broad areas shared with cloud computing are:

Web application vulnerabilities, such as SQL injection, cross-site scripting, field input validation, buffer overflow; as well as default configurations or mis-configured applications.

Accessibility vulnerabilities, which are inherent to the TCP/IP stack and the operating systems, like denial of service and distributed denial of services.

Authentication of the respondent device or devices, IP spoofing, ARP poisoning (spoofing), RIP attacks, and DNS poisoning are all too common on the Internet.

Data Verification, tampering, loss and theft, is a crucial thing to be concerned because cloud storages use multiple remote machines for storage.

Physical access issues, organization’s staff should not have physical access to the machines storing and processing a data.

Privacy and control issues halting from third parties having physical control of a data is an issue for all outsourced networked applications and storage, but cloud architectures have some specific issues that are distinct from the usual issues.

6. NEED OF SECURITY IN CLOUD COMPUTING [6]

Today Small and Medium Business (SMB) companies are increasingly realizing that simply by tapping into the cloud they can gain fast access to best business applications or drastically boost their infrastructure resources, all at negligible cost. Gartner defines cloud computing as ‘‘a style of computing where massively scalable IT- enabled capabilities are delivered ‘as a service’ to external customers using Internet technologies’’. The providers must ensure they are the ones who will shoulder the responsibility if things go wrong. The cloud offers numerous benefits like pay-for-use, fast deployment, lower costs, scalability, rapid elasticity, rapid provisioning, higher resiliency, hypervisor protection against network attacks, pervasive network access, low-cost disaster recovery and data storage solutions, real time detection of system tampering and rapid re-constitution of services. While the cloud offers these advantages, until some of the jeopardies are better understood, many of the major players will be desirous to hold back. According to a recent IDCI survey, 74% of IT executives and CIO's cited security as the top challenge preventing their adoption of the cloud services model. Analysts' estimate that within the next five years, the worldwide market for cloud computing will grow to $95 billion and that 12% of the worldwide software market will move to the cloud in that period. To realize this marvelous potential, corporate must address the privacy questions raised by this new computing model. Cloud computing moves the application software and databases to the large data centers, where the administration of the data and services are not trustworthy. This sole characteristic, however, stances many new security challenges. These challenges contain but not limited to virtualization vulnerabilities, physical access issues, web application vulnerabilities like SQL (Structured Query Language) injection and cross-site scripting, issues related to identity and credential management, privacy and control issues, problems related to data verification, data loss, integrity, interfering, confidentiality, and theft, issues related to authentication of the respondent device or devices and IP spoofing.

Different with other computing models, there are no obvious users boundaries or perimeters in cloud computing. Nevertheless, some existed data protection mechanisms are invalid because the exact data location in the cloud is uncertain, data may migrate to diverse servers according to performance or scalability needs. When performing the cryptography algorithm, such as CPU utilization, it often consumes a lot of system resources, and stronger algorithm that generates more significant impact to the system performance.

7. CURRENT ISSUES IN CLOUD DATA SECURITY

Securing virtual machines in the cloud [7]

Choosing protection for a virtual infrastructure is a lot like buying an antivirus product for the Mac OS: most people would wonder why you bothered. However, as more IT shops migrate their servers to virtual machines and cloud-based environments, it is only a matter of time before protecting these resources becomes considerably more important.

Intrusion detection in a cloud computing environment [8]

Attacks on systems and data are a reality in the world we live in. Noticing and replying to those attacks has become the norm and is considered due diligence when it comes to security. As a matter of fact, many of the standards and regulations applied in the technology space today have explicit instructions regarding the need for monitoring and alerting.

Securing data in the cloud [9]

- Protect your data in a real world environment.
- Meet compliance requirements.

Cloud Aggregation [10]

- Novel architectural models for aggregation of cloud providers
- Brokering algorithms for high availability, performance, proximity, legal domains, price, or energy efficiency
- Sharing of resources between cloud providers
- Networking in the deployment of services across multiple cloud providers
- SLA negotiation and management between cloud providers
- Additional privacy, security and trust management layers atop providers
- Support of context-aware applications
- Automatic management of service elasticity

Cloud Management [10]

- Scalable management of network, computing and storage capacity
- Scalable orchestration of virtualized resources and data
- Placement optimization algorithms for energy efficiency, load balancing, high availability and QoS
- Accounting, billing, monitoring and pricing models
- Security, privacy and trust issues in the cloud
- Energy efficiency models, metrics and tools at system and datacenter levels

Cloud Enablement [10]

- Technologies for virtualization of infrastructure resources
- Virtualization of high performance infrastructure components
- Autonomic and intelligent management of resources
- Implications of Cloud paradigm on networking and storage systems
- Support for vertical elasticity
- Provision of service related metrics

Cloud Interoperability [10]

- Common and standard interfaces for cloud computing
- Portability of virtual appliances across diverse clouds providers

8. NEED OF DATA INTEGRITY PROOFS

One problem is to verify that the server continually and faithfully stores the entire file entrusted to it by the client. If the server is untrusted in terms of both security and reliability: it might maliciously or accidentally erase the data or place it onto temporarily unavailable storage media. This could occur for many reasons including cost-savings or external pressures (e.g., government censure). The client's limited resources and the limited bandwidth between the client and server are the factors that exacerbating the problem. When users store their data in the external service providers, they mostly concern about whether the data is intact; whether it can be recovered when there is a failure. So a critical issue in storing data on untrusted servers is that verifying the storage server keeps on holding their data completely and correctly.

9. CHALLENGES IN DATA INTEGRITY PROOFS

- How to incorporate wireless moveable devices, especially lightweight devices such as cell phones and sensors, into the cloud system.
- How to share encrypted data with a huge number of users, in which the data sharing group can be changed frequently.
- How to update and upload/download encrypted data stored in the cloud system.
- How Misbehavior can be easily detected and punished by the customers.

Generally, to design a Remote Data Possession Checking (RDPC) scheme, the following factors must be considered.

- Computation complexity, which refers to the initialization and verification overheads in the client and the proof generating overheads on the server. It means that the system should be efficient in terms of computation.
- Communication complexity, which refers to the amount of communication between client and server required by the scheme. It means that the amount of communication should be low.
- Storage cost, which refers to the additional storage of client and server required by the scheme. It means that the extra storage should be as low as possible.
- Data updating, including modifying, inserting, adding and deleting etc. It can only be used for static data if it doesn't support data update, such as data archive.
- The number of verification. It ought to run the verification an unlimited number of times.
- Public verification. It ought to support public verification.
- Data recovery, which means that the scheme can recover the data in case of a failure. It can be achieved by introducing error correcting code or erasure code.
- Provable security. Generally, it is necessary to prove that the scheme is secure.
- Data blocks access, which refers to that how much data blocks the scheme needs to access.

10. RELATED WORK

The Remote integrity checking [11] uses RSA based hash functions to hash the entire file. That requires the prover to exponentiate over the entire file F and accesses the entire file's blocks. Other method Provable data possession at untrusted stores [12] authors have formally define protocols for PDP and present two provably secure PDP schemes. The protocol uses homomorphic verifiable tags which is computationally intensive and doesn't consider data updation. Scalable And Efficient Provable Data Possession [13] presents a provably secure PDP scheme based on symmetric key cryptography. And the scheme provisions some runtime operations, including updation, deletion and appending. In this method Number of modifications and challenges is limited and fixed a priori and it is unsuitable for public verifiability. Multiple-replica provable data possession [14] provided a provably secure multiple-replica PDP (MR-PDP) scheme. The scheme is also based on RSA and it doesn't consider data updating. A PDP Scheme for Networked Archival Storage [15] provides a scheme based on symmetric key cryptography which is called data possession checking (DPC). This scheme proposes a challenge renewal mechanism based on verification block circular queue to allow the dynamic increase of the number of effective challenges.

Some of the work is done for the integrity checks for dynamic data. The Dynamic provable data possession [16] presented a framework and a construction for dynamic provable data possession (DPDP). The scheme support data updating and can be extended to construct complete file systems and version control systems at untrusted servers. This is also based on RSA. PoRs: proofs of retrievability for large files [17] authors have introduced the notion of proof of retrievability (POR) and proposed a formal POR protocol definition and accompanying security definitions. Compared with PDP, POR offers uses Error Correction/Erasure Codes to tolerate the damage to portions of the outsourced data this is an extra property that the client can actually "recover" the data outsourced to the cloud. The Schemes don't consider data updation. [18] have tackle security and efficiency problem by proposing a "2-in-1" notion we call Proof of Data Storage with Deduplication.

11. OPEN ISSUES

The following aspects are considered to be future directions of RDPC.

(1) The design of efficient RDPC schemes.

On the one hand, it is to improve computation, communication and storage efficiency. On the other hand, it is to improve detection efficiency, to detect the error and to recover the data with high probability and accuracy.

(2) Supporting more extensive application environments.

On the one hand, any networking devices can be used, such as PDA, mobile phone, wireless phone and net book etc. It means that these schemes are suitable for wireless environments. As the computation and storage capability is limited, the requirement of these schemes is much data set.

(3) Effectively supporting data updating.

The main data updating operations include modifying, inserting, adding and deleting. It is an important feature for storage service, which will determine whether the users choose the service.

(4) Providing quality of service (QoS).

On the one hand, it is to provide different QoSs. On the other hand, to guarantee to achieve declared performance and QoS, users can assess the QoS of SSPs by utilizing performance tracking tool and MR-PDP protocol. For example, when the declared bandwidth is 100KB/s, it is 100KB/s in fact. And when SSPs declared that there are n copies, there are n copies in fact.

(5) Providing security proving.

If a scheme is not proved to be secure, users won't select to use it. As the most schemes utilize modern cryptography, the security proving methods can be categorized into two types, namely standard model and random oracle model. But for different threat model and application environments, for example, there may be mobile attacks in some environments, it needs to consider the dynamic of the threats. Thus, different schemes require appropriate security proving methods.

In addition, introducing a third-party auditor to remove the burden of verification from the users and provide a method for independent arbitration of data retention contracts is also a hot issue. But it will cause the problem of privacy-preserving, which means that it must not leak users' data to the third-party auditor.

12. CONCLUSION

Remote data possession checking is a topic that focuses on how to efficiently and securely verify that a storage server is faithfully storing its client's (potentially very large) original data without retrieving it. There are two types of schemes, namely provable data possession (PDP) and proof of retrievability (POR). The difference between PDP and POR is that POR checks the possession of data and it can recover data in case of a failure. In this report, a simplified architecture of Storage as a Service model is presented and the security requirements are discussed.

A series of evaluation factors considered in designing an RDPC scheme are listed. Then of the art research works on RDPC are reviewed. At last, according to these factors listed above, the existing schemes are compared. From the results of comparisons, the drawbacks of the current schemes are pointed out, which help to define the future directions for improving the existing schemes. Although all of the schemes are not perfect, they have their appropriate application areas, which are also discussed.

13. REFERENCES

[1] The NIST Definition of Cloud Computing (v15) Mell, Peter and Grance, Tim (2009) http://csrc.nist.gov/groups/SNS/cloud-computing/

[2] Trusted Cloud Computing with Secure Resources and Data Coloring by: Kai Hwang, Deyi Li IEEE Internet Computing, Vol. 14, No. 5. (September 2010), pp. 14-22, oi:10.1109/MIC.2010.86 Key: citeulike:8983216

[3] Andrew OneDegree Blog, Andrew Smith, http://andrewonedegree.wordpress.com/2010/01/20/virtualisation-its-not-a-cloud/20012010

[4] Difference between Cloud computing and Utitlity computing, Uchit Vyas, http://cloudbyuchit.blogspot.in/2012/03/difference-between-cloud-computing-and.html

[5] Cloud Computing Issues, Dynaroll Corp., http://www.dataplex.com/blog/index.php/2010/01/07/cloud-computing-issues/

[6] A survey on security issues in service delivery models of cloud computing, Subashini S, Kavitha V. J Network Comput Appl (2010), doi:10.1016/j.jnca.2010.07.006

[7] Securing virtual machines in the cloud, David Strom, http://searchcloudcomputing.techtarget.com/tip/Securing-virtual-machines-in-the-cloud

[8] Intrusion detection in a cloud computing environment, Phil Cox, http://searchcloudcomputing.techtarget.com/tip/Intrusion-detection-in-a-cloud-computing-environment

[9] Securing data in the cloud, Phil Cox, http://searchcloudcomputing.techtarget.com/tip/Securing-data-in-the-cloud

[10] http://blog.cloudplan.org/2011/05/key-research-challenges-in-cloud.html

[11] “Remote integrity checking”, Yves Deswarte, Jean-Jacques Quisquater, Ayda Saïdane, In: Proc. Of IICIS '03, pp.1–11, 2003

[12] “Provable data possession at untrusted stores”, Giuseppe Ateniese, Randal Burns, Reza Curtmola, Joseph Herring, Lea Kissner, Zachary Peterson,Dawn Song, In: Proc. of ACM-CCS '07, pp.598–609, 2007.

[13] “Scalable and efficient provable data possession”, Giuseppe Ateniese, Roberto Di Pietro, Luigi V. Mancini, Gene Tsudik, In: Proc. of SecureComm '08, pp.1-10, 2008.

[14] “MR-PDP: Multiple-replica provable data possession”, Reza Curtmola, Osama Khan, Randal Burns, Giuseppe Ateniese, In: Proc. of ICDCS '08, pp.411-420, 2008.

[15] “A Practical Data Possession Checking Scheme for Networked Archival Storage”, Xiao Da, Shu Jiwu, Chen Kang, Zheng Weimin, Journal of Computer Research and Development, 46(10):1660-1668, 2009.

[16] “Dynamic provable data possession”, C. Chris Erway, Alptekin Kupcu, Charalampos Papamanthou, Roberto Tamassia, In: Proc. of ACM-CCS '09, pp.213-222, 2009.

[17] “Pors: proofs of retrievability for large files”, Ari Juels, Burton S. Kaliski Jr., In: Proc. of ACMCCS '07, pp.584-597, 2007.

[18] Qingji Zheng and Shouhuai Xu. Secure and efficient proof of storage with deduplication. In CODASPY ’12: ACM conference on Data and Application Security and Privacy, pages 1–12, 2012.