A Comprehensive Overview on Different Cloud Storage Techniques and Security Mechanism

R.Sathiya, N.Prathipa, K.Gomathi

Abstract: - With the development of cloud computing, Data security becomes more and more important in cloud computing. This paper analyses the basic information about cloud computing and cloud computing data security issues , with the analysis of Hadoop map reduce and Merkel hash tree authentication of data elements. Finally we build a data security in real world for cloud computing Keyword-Keyword1.

Index Terms— Characteristics of Cloud storage, Security issues, Data security, Map reducing Programming Model, Avoiding Bad Hadoop and Cloud Analytics Decisions.

I. INTRODUCTION

Cloud computing means application and services over the Internet. These services are offered from data centers all over the world, which collectively are referred to as the **"cloud".** Examples of cloud computing includes online backup services, social networking services, and personal data services such as **Apple's Mobile Me**. Cloud computing also includes online applications, such as those offered through Microsoft Online Services. Hardware services, such as redundant servers, mirrored websites, and Internet-based clusters are also examples of cloud computing. These services are broadly divided into three categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). A cloud service has three distinct characteristics that differentiate it from traditional hosting. A user can have as much or as little of a service as they want at any given time and the service is fully managed by the cloud service provider.

The advantage of cloud is cost savings. The prime disadvantage is security. Cloud computing is used by many software industries now-a-days. Since the security is not provided in cloud, many companies adopt their unique security structure. For eg Amazon has its own security structure. Introducing a new and uniform security structure for all types of cloud is the problem we are going to tackle in this paper. Since the data placed in the cloud is accessible to everyone, security is not guaranteed. [1]

II. CHARACTERISTICS OF CLOUD COMPUTING

A) On demand self-service: On-demand self-service, refers to the cloud service provided by the cloud providers are controlled and monitored by themselves. This is crucial for billing, access control, resource optimization, capacity planning and other tasks. In on-demand self-service, the user accesses cloud services through an online control panel. On-demand self-service resource sourcing is a prime feature of most cloud offerings where the user can scale the required infrastructure up to a substantial level without disrupting the host operations. [2]

B) Broad network access: Broad network access refers to resources hosted in a private cloud network (operated within a company's firewall) that are available for access from a wide range of devices, such as tablets, PCs, Macs and smartphones. These resources are also accessible from a wide range of locations that offer online access. Companies that have broad network access within a cloud network need to deal with certain security issues that arise. [3]

C) Resource pooling: The cloud enables a company's employees to enter and use data within the business management software hosted in the cloud at the same time, from any location, and at any time. This is an attractive feature for multiple business offices and field service or sales teams that are usually outside the office.

D) Measured service: Going back to the affordable nature of the cloud, user only pay for what they use. User and the cloud provider can measure storage levels, processing, bandwidth, and the number of user accounts and the user is billed appropriately. The amount of resources that the user may use can be monitored and controlled from both user side and cloud provider's side which provides transparency. [4]

E) Rapid elasticity: Elasticity is defined as the ability to scale resources both up and down as needed. To the consumer, the cloud appears to be infinite, and the consumer can purchase as much or as little computing. Power as they need. Defined by the National Institute of Standards and Technology (NIST) as one of the five essential characteristics of cloud computing. [5]

III. CLOUD INFRASTRUCTURE CLASSIFICATION

A) Public Cloud: In this case the system is owned by the company that specializes in selling. Cloud services and is accessible to all individuals and businesses. Here we include all services publicly available to end users through a global public network, regardless of their location. A good example is social networks (Facebook, Twitter ...).

B) Private Cloud : A private cloud is a particular model of cloud computing that involves a distinct and secure cloud

Manuscript received October, 2013.

R.Sathiya, Department of Computer Science and Engineering, Adithya Institute of Technology, Kurumbapalayam,Coimbatore-641 107, Tamil Nadu, India.

N. Prathipa, Department of Computer Science and Engineering, Adithya Institute of Technology, Kurumbapalayam, Coimbatore-641 107, Tamil Nadu, India.

K.Gomathi, Department of Computer Science and Engineering, Adithya Institute of Technology, Kurumbapalayam, Coimbatore-641 107, TamilNadu, India.

based environment in which only the specified client can operate. As with other cloud models, private clouds will provide computing power as a service within a virtualized environment using an underlying pool of physical computing resource. However, under the private cloud model, the cloud (the pool of resource) is only accessible by a single organization providing that organization with greater control and privacy.

C) Community Cloud: **Community cloud** computing refers to a shared cloud computing service environment that is targeted to a limited set of organizations or employees (such as banks or heads of trading firms). The organizing principle for the community will vary, but the members of the community generally share similar security, privacy, performance and compliance requirements. Community members may wish to invoke a mechanism that is often run by themselves (not just the provider) to review those seeking entry into the community

D) Hybrid Cloud: Hybrid cloud is an environment that employs both private and public cloud services. Companies are realizing that they need many different types of cloud services in order to meet a variety of customer needs.

The growing importance of hybrid cloud environments is transforming the entire computing industry as well as the way businesses are able to leverage technology to innovate. Economics and speed are the two greatest issues driving this market change.

E) A cloud is hybrid: If a company uses a public development platform that sends data to a private cloud or a data center–based application.

- When a company leverages a number of SaaS (Software as a Service) applications and moves data between private or data center resources.
- When a business process is designed as a service so that it can connect with environments as though they were a single environment.

F) A cloud is not hybrid:

- If a few developers in a company use a public cloud service to prototype a new application that is completely disconnected from the private cloud or the data center.
- If a company is using a SaaS application for a project but there is no movement of data from that application into the company’s data center.

IV. DIFFERENT SERVICES OFFERED BY CLOUD

Platform **as a service:** Platform as a Service (PaaS) is a way to rent hardware, operating systems, storage and network capacity over the Internet. The service delivery model allows the customer to rent virtualized servers and associated services for running existing applications or developing and testing new ones. Platform as a Service (PaaS) is an outgrowth of Software as a Service (SaaS), a software distribution model in which hosted software applications are made available to customers over the Internet. PaaS has several advantages for developers. With PaaS, operating system features can be changed and upgraded frequently. Geographically distributed development teams can work together on software development projects. Services can be obtained from diverse sources that cross international boundaries. Initial and ongoing costs can be reduced by the use of infrastructure services from a single vendor rather than maintaining multiple hardware facilities that often perform duplicate functions or suffer from incompatibility problems. Overall expenses can also be minimized by unification of programming development effort. On the downside, PaaS involves some risk of "lock-in" if offerings require proprietary service interfaces or development languages. Another potential pitfall is that the flexibility of offerings may not meet the needs of some users whose requirements rapidly evolve.

Infrastructure **as a service:** Infrastructure as a Service is a provision model in which an organization outsources the equipment used to support operations, including storage, hardware, servers and networking components. The service provider owns the equipment and is responsible for housing, running and maintaining it. The client typically pays on a per-use basis.

Characteristics and components of IaaS include:

- Utility computing service and billing model.
- Automation of administrative tasks.
- Dynamic scaling.
- Desktop virtualization.
- Policy-based services.
- Internet connectivity.

Software as a service: Software as a Service (SaaS) is a software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network, typically the Internet.

SaaS is becoming an increasingly prevalent delivery model as underlying technologies that support Web services and service-oriented architecture (SOA) mature and new developmental approaches, such as Ajax, become popular. Meanwhile, broadband service has become increasingly available to support user access from more areas around the world. SaaS is closely related to the ASP (application service provider) and on demand computing software delivery models. IDC identifies two slightly different delivery models for SaaS. The hosted application management (hosted AM) model is similar to ASP: a provider hosts commercially available software for customers and delivers it over the Web. In the software on demand model, the provider gives customers network-based access to a single copy of an application created specifically for SaaS distribution.

V. SECURITY ISSUES

Internal clouds are not inherently secure: Many organizations have foregone using public clouds choosing instead to build private clouds behind their firewalls. This may be the best solution for risk-averse groups. These teams, though, need to understand that just because they've built a cloud inside their firewall doesn't mean that their solution is safe. It still takes just one bad apple to spoil the barrel—a single department, user or application that is not behaving as it should.

An organization that is risk-averse enough to avoid the public cloud should be building a secure cloud—possibly the company should be building its dream cloud, which contains all the security controls that it thinks are missing from a public environment. Since the company physically owns the private cloud, incident response can be very swift. Detection capabilities need to be cloud-specific (for example, sensors need to monitor inside the cloud, not just at its perimeter) and operational capabilities such as patch management must be sharp. A vulnerable service that's in a cloud might have greater exposure and risk than the same service in a standard server farm thanks to the shared nature of cloud resources. [7]

Companies lack security visibility and risk awareness: The paucity of security visibility that most providers offer their customers is itself getting plenty of visibility. Obviously, when using a public cloud service, companies must balance the competing factors of control, visibility and cost. This can be a significant issue—reduced visibility results in diminished situational awareness and a questionable understanding of risk. When planning a move to the cloud, an organization needs to recognize this lack of visibility and determine how to best leverage what insight they can get their hands on. Really, this means designing mitigating controls. [7]

Sensitive information needs safer storage: Safely storing sensitive information is one of the toughest problems in cloud computing. The solution is to encrypt data, but the critical questions are where to encrypt, and how. The first requirement of successful encryption in the cloud, which some providers do not yet understand (or at least don't practice), is: Do not store the encryption key with the encrypted data. Doing so more or less negates any value gained from encrypting the data. [7]

Apps aren't secure: Application security has been getting attention for years. In my mind, its importance increases when an application is deployed to a cloud environment, as the application is more exposed. One of the biggest mistakes an organization can make is to take an existing application and simply deploy it to a cloud without first considering what new attack vectors this move opens up. Never trust user input, and always encode output back to the user. Getting those two things right will remove about 80 percent of application security issues. After input and output are taken care of, next up is proper authentication and authorization. [7]

Authentication and authorization must be more robust: Of all the problems covered in this article, cloud authentication and authorization has the greatest number of commercial solutions available. This does not mean the issue is easily solved, however. Every organization has its own way to manage authentication and authorization. First, it must determine if its current authentication system could also work in a secure and reliable way for users in a cloud environment. If the answer is yes, the follow-up question is whether that is also the best way to authenticate cloud services.

Security Problem in VM: The virtual server from the logical server group brings a lot of security problems. The traditional data center security measures on the edge of the hardware platform, while cloud computing may be a server in a number of virtual servers, the virtual server may belong to different logical server group, virtual server, therefore there is the possibility of attacking each other ,which brings virtual servers a lot of security threats. Virtual machine extending the edge of clouds makes the disappearance of the network boundary, thereby affecting almost all aspects of security, the traditional physical isolation and hardware-based security infrastructure cannot stop the clouds computer environment of mutual attacks between the virtual machine. [8]

The existence of super user: The existence of super-users to greatly simplify the data management function, but it is a serious threat to user privacy. Super-powers is a double-edged sword, it brings convenience to users and at the same time poses a threat to users. In an era of personal privacy, personal data should be really protected, and the fact that cloud computing platform to provide personal services in the confidentiality of personal privacy on the existence of defects. Not only individual users but also the organizations have similar potential threats. [8]

Consistency of Data:

Cloud environment is a dynamic environment, where the user's data transmits from the data center to the user's client. For the system, the user's data is changing all the time. Read and write data relating to the identity of the user authentication and permission issues. In a virtual machine, there may be different users' data which must be strict managed. [8]

Principle of Data Security:

All the data security technic is built on confidentiality, integrity and availability of these three basic principles. Confidentiality refers to the so-called hidden the actual data or information, especially in the

Military and other sensitive areas, the confidentiality of data on the more stringent requirements. For cloud computing, the data are stored in "data center", the security and confidentiality of user data is even more important. The so-called integrity of data in any state is not subject to the need to guarantee unauthorized deletion, modification or damage. The availability of data means that users can have the expectations of the use of data by the use of capacity. [9]

VI. DATA SECURITY OF CLOUD LAYERS

- **The first layer:** responsible for user authentication, the user of digital certificates issued by the appropriate, manage user permissions.
- **The second layer:** responsible for user's data encryption, and protect the privacy of users through a certain way.
- **The third layer:** The user data for fast recovery, system protection is the last layer of user data. R[9]

Different Types of Levels

- **Single-server level** means to store every pieces of data block in only one server, which means that as long as any of the data block that stores the file server failure will lead to file corruption, we've calculated that the probability is great, so for sub-block storage file generally will not use this level of security policy, but for small files is not implementation of block storage can be considered such a security level.
- **Cross-server level** is to copy data to more than one backup server. When one server fails, the access requests are automatically redirected to backup server.
- **Across the data centre level** is the data block and copy backup strategy layer will store a copy of data block data centres in different regions, between the two centres may be thousands of miles away, this level of document security in the face of major disasters and

major accidents when they can ensure data integrity, data storage cost of this security level higher, generally apply to banks, telecommunications and other core data, and data storage are critical applications. [10]

Hadoop:

Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop Map Reduce and the Hadoop Distributed File System. Map Reduce is the parallel-processing engine that allows Hadoop to churn through large data sets in relatively short order. HDFS is the distributed file system that lets Hadoop scale across commodity servers and, importantly, store data on the compute nodes in order to boost performance (and potentially save money). These are the two must-have components for any Hadoop distribution. [17]

A wide variety of companies and organizations use Hadoop for both research and production. Users are encouraged to add themselves to the Hadoop

The project includes these modules:

* Hadoop Common
* Hadoop Distributed File System (HDFS™)
* Hadoop YARN
* Hadoop Map Reduce

Hadoop Map Reduce:

Hadoop includes a Java implementation of the Map Reduce framework, its underlying components and the necessary large scale data storage solutions. Although application programming is mostly done in Java, it provides APIs in different languages such as Ruby and Python, allowing developers to integrate Hadoop to diverse existing applications. It was first inspired by Google's implementation of Map Reduce and the GFS distributed file system, absorbing new features as the community proposed new specific sub projects and improvements. Currently, Yahoo is one of the main contributors to this project, making public the modifications carried out by their internal developers. The basis of Hadoop and its several sub projects is the Core, which provides components and interfaces for distributed I/O and file systems. The Avro data serialization system is also an important building block, providing cross-language RPC and persistent data storage. [18]

Map Reduce Programming Model :

1.Very simple programming model

- a=(Key, value)
- Map(anything) ------> a
- Sort, partition on key
- Reduce(Key, value) ---------> a

2. No parallel processing/message passing semantic

3. Programmable in java or any other language (streaming)

The Web has become the repository of most the world's pubic knowledge. Almost all of it is still bound up in text, images, audio and video, which are easy for people to understand but less accessible for machines. While the computer interpretation of visual and audio information is still challenging, text is within reach. The Web's infrastructure makes access to all this information trivial, opening up tremendous opportunities to mine text to extract information that can be republished in a more structured representation (e.g., RDF, databases) or used by machine learning systems to discover new knowledge. Current technologies for human language understanding are far from perfect, but can harvest the low hanging fruit and are constantly improving. All that's needed is an Internet connection and cycles — lots of them.

The latest approach to focusing lots of computing cycles on a problem is cloud computing, inspired in part by Google's successful architecture and Map Reduce software infrastructure. [19]

Avoiding Bad Hadoop and Cloud Analytics Decisions:

1. Big Data is purely about volume—NOT TRUE
2. Traditional SQL doesn't work with Hadoop—NOT TRUE
3. Kill the Mainframe! Hadoop is the only the new IT data platform—NOT TRUE
4. Virtualized Hadoop takes a performance hit—NOT TRUE
5. Hadoop only works in your data center—NOT TRUE
6. Hadoop doesn't make financial sense to virtualize—NOT TRUE
7. Hadoop doesn't work on SAN or NAS—NOT TRUE [20].

Cloud Security in the real world: Securing cloud computing environments will be a major focus of vendor efforts over the next year, says Jonathan Penn, an analyst at Forrester Research. In the short term, he sees users having to do a lot of the legwork, but over time, "cloud providers themselves will see the opportunity to differentiate themselves by integrating security," he says. Security vendors accustomed to selling directly to the enterprise will find that they need these cloud providers as a way to reach the market, Penn says, and as the market matures, customers will want this stuff baked into the services they're buying. "That will be quite a radical change and a disruption," he adds.

In the meantime, organizations such as the Cloud Security Alliance (CSA) are working to put some shape around the security issues and the ways to address them. The CSA recently released a summary of the strategic and tactical security pain points within a cloud environment, along with recommendations on how to address them.

VII. CONCLUSION

As the development of cloud computing, security issue has become a top priority. This paper discusses the cloud computing environment and we conclude safety issues through analyzing hadoop map reduce and security model. Finally we conclude a cloud security in real world.

ACKNOWLEDGMENT

The authors are thankful to **Mr.M.R.CHRISTHU RAJ** Of Adithya Institute of Technology, Faculty of Computer Science and Information Technology for providing the necessary facilities for the preparation of the paper.

REFERENCES

[1] Cloud Computing: http://simple.wikipedia.org/wiki/Cloud_computing
[2] Characteristics of cloud computing : http://cloudglossary.com/home/id.On-demand-self-service/i.html
[3] Characteristics of cloud computing : http://www.techopedia.com/definition/28785/broad-network-access
[4] Characteristics of cloud computing : http://www.cloudcommons.com/about-smi
[5] Characteristics of cloud computing : http://cloudglossary.com/home/id.Rapid-Elasticity/i.html

[6]Cloudclassificationsandbenefits:http://en.wikipedia.org/wiki/Cloud_computing, http://searchcloudcomputing.techtarget.com

[7] Cloud computing security issues : http://www.csoonline.com/article/507974/the-cloud-security-survival-guide , http://www.csoonline.com/article/596819/cloud-security-the-basics-----cloud

[8] Cloud Computing Security: making Virtual Machines Cloud-Ready, www.cloudreadysecurity.com 2008

[9] Greg Boss, Cloud Computing□IBM 2007.10

[10] Ensure Data Security in Cloud Storage Xiao Zhang, Hong-tao Du, Jian-quan Chen, Yi Lin, Lei-jie Zeng

[11] H. Shacham and B. Waters, “Compact proofs of retrievability,” in Proc. of ASIACRYPT’ 08. Springer-Verlag, 2008, pp. 90–107.

[12] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, “Provable data possession at untrusted stores,” in Proc. of CCS’07. New York, NY, USA: ACM, 2007, pp. 598–609.

[13] K. D. Bowers, A. Juels, and A. Oprea, “Proofs of retrievability: Theory and implementation,” Cryptology ePrint Archive, Report 2008/175, 2008.

[14] A. Juels and B. S. Kaliski, Jr., “Pors: proofs of retrievability for large files,” in Proc. of CCS’07. New York, NY, USA: ACM, 2007, pp. 584–597.

[15] D. Boneh, B. Lynn, and H. Shacham, “Short signatures from the weil pairing,” in Proc. of ASIACRYPT’01. London, UK: Springer-Verlag, 2001, pp. 514–532.

[16] R. C. Merkle, “Protocols for public key cryptosystems,” Proc. of IEEE Symposium on Security and Privacy’80, pp. 122–133, 1980.

[17] Hadoop definition: http://hadoop.apache.org/

[18] Hadoop map reduce: http://www.linuxjournal.com/content/open-source-cloud-computing-hadoop

[19] Hadoop map reduce programming model: http://ebiquity.umbc.edu/blogger/2007/12/26/cloud-computing-with-hadoop/

[20] Avoiding of badhadoop and cloud analytics: http://blogs.vmware.com/vfabric/2013/04/myths-about-running-hadoop-in-a-virtualized-environment.html