Analysis of File Retrieval Technique in Cloud Scenario: A Review

Mithilesh Sagar [1], Dr. Avinash Sharma[2]

[1]M.Tech Student, [2]Head & Associate Professor

Department of Computer Science & Engineering, MIT, Bhopal, India

Abstract - This paper presents the survey on data storage and retrieval in cloud computing. In this paper the study on scope and security issues related to data storage and information retrieval in cloud computing is done. Data storage and retrieval with data security is typical issue in today's scenario. The main objective of cloud computing is to enables users with limited computational resources to outsource their large computation workloads to the cloud. The integrity and confidentiality of the data uploaded by the user is ensured doubly by not only encrypting it but also providing access to the data only on successful authentication. In this paper the different security problems existing in the cloud was identified and solutions for the same have suggested..

Keywords- Cloud computing; Data security; cloud security; Information Retrieval; security issues;

1. INTRODUCTION

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". Cloud Computing is a completely new Information Technology and it is known as the third revolution after PC and Internet in IT. The basic principle of Cloud Computing is collecting large quantities of information and resources stored in personal computers, mobile phones and other equipment, Cloud Computing is capable of integrating them and putting them on the public cloud for serving users. Cloud computing technology redefines the advances in information technology. Cloud computing is an emerging technology which is being used widely these days. Due to the cost-effectiveness, flexibility and scalability of cloud, more and more organizations are now using cloud to outsource their data for sharing. In a cloud computing environment, an organization subscribes the cloud services and gives access to its staff to share files in the cloud. Each file is identified with some keywords, and only authorized users can retrieve files, so that they can send query with those keywords to cloud and retrieve interested files. In such a scenario, protection of user privacy from the cloud, which is outside the security boundary of the organization, becomes a key problem. In this paper it was observed from a users point of view, relying upon data retrieving which he needs by performing an effective query optimization. In addition, providing reliability, availability are crucial and equally important to query optimization. Query Optimization can be achieved by implementing the optimization techniques for effective retrieval of data.

2. WHAT IS CLOUD COMPUTING?

Cloud Computing is a computer model that provides services in the form of on-demand services, it's accessible for everyone, everywhere and every time, including clouds referring to the internet and the web. "A cloud is a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumer." Cloud computing seems to offer some incredible benefits for communicators: the availability of software applications, quick processing, unlimited storage, and the ability to easily share and process information. All of this is available through the browser any time you can access the Internet. Typical kinds of cloud computing services are:

COMMUNICATION AS A SERVICE (CAAS): Allow for certain messaging tools viz voice over IP (VOIP), Instant Messages (IM) and Video Conferencing.

INFORMATION AS A SERVICE (IAAS): Clouds that provide large and affordable computing resources that run enterprise programs, which is known as Infrastructure as a service (IaaS), Allows user to maintain their application while loading data on the Iaas infrastructure.

MONITORING AS A SERVICE (MAAS):

Outsourcing of security service to a third party security team.

PLATFORM AS A SERVICE (PAAS): Local Storage Clouds also known as a Platform as a Service (PaaS), meant for web-based development infrastructure.

SOFTWARE AS A SERVICE (SAAS): Application Clouds also called Software as a Service (SaaS), when a software vendor supplies software over a network as opposed to the typical distribution of installation of individual computers.

Cloud computing is becoming popular. The popularity of this technology is based on the benefits to the users.

However, like any other technology, cloud computing has its own limitations.

Information Retrieval (IR) is the discipline that deals with retrieval of unstructured data, especially textual documents, in response to a query or topic statement, which may itself be unstructured, e.g., a sentence or even another document, or which may be structured, e.g., a Boolean expression. A search engine is performing information retrieval task. A search engine presents the retrieved document set as a ranked list of document titles. The documents in the list are ordered by the probability of being relevant to the user's request. The highest ranked document is considered to be the most likely relevant document; the next one is slightly less likely and so on.

3. TECHNIQUES FOR INFORMATION RETRIEVAL

There are different models, techniques being used for information retrieval. Some are summarized as follows. Vector Space Model: This model is used to represent the text by a vector of functions. The terms are the words and phrases. If words are considered as terms, every word becomes an independent dimension in a very high dimension vector space. If term represents a text, it gets a non- zero value in the text-vector along the dimension corresponding to the term. Text vectors are very space and no term is assigned a negative value.

Probabilistic Model: The principle of probabilistic model is that the documents in a collection should be ranked by decreased probability to query relevance. This principle is called as the probabilistic ranking principle. The ranking criteria are monotonic under log-odd transformations. Each probabilistic model that is proposed is based on a different probabilistic estimation technique.

Inference Network Model: A model that is used for a document to instantiate a term. The credit from multiple terms is accumulated given to compute the equivalent of a numeric score for data.

Term Weighting: Term weighting is techniques that rely upon the better estimation of various probabilities. The main three factors plays in term weight formulation are term Frequency, document Frequency and document Length. Term frequency is calculated on the basis of words that repeat multiple times in a document. Document frequency calculated on the basis of words that appear in many documents are considered common and document length is calculated on the basis of appearance of the word in the sentences. When collections have documents of varying lengths, longer documents tend to score higher since they contain more words and more repetition.

Query Modification: Query modification is a technique which found a solution when it was hard for users to formulate effective search requests. It is done by adding synonyms of query words to query would improve search effectiveness. The synonyms relied on a thesaurus. So they developed a technique called query modification to automatically generate thesauri for query modification. Natural Language Processing: NLP is a tool that is used to enhance retrieval effectiveness but had only very limited success. The advantages are easier and faster information retrieval and information discovery.

4. SECURITY AND AVAILABILITY IN THE CLOUD

Security problems can be major issue in information retrieval and storage. Security requires guaranteeing proper protection of sensitive information stored or processed in the cloud. There is always a risk that digital information may become corrupted in the event of network breaks, service disruptions or network congestion. If this happens, it can be difficult or impossible to access and use it.

Many of the risks associated with cloud computing are not new. Some examples of cloud computing risks for the enterprise that need to be properly managed include:

- Choice of trusted cloud service provider based on history and sustainability. Sustainability is of particular importance to ensure that services will be available and data can be tracked.
- Responsibility of information handling is transferred to the cloud provider, which is a critical part of the business. Failure to perform to agreed-upon service levels can impact not only confidentiality but also availability, severely affecting business operations.
- Dynamic nature of cloud computing may result in confusion as to where information actually resides. When information retrieval is required, this may create delays.
- High risk of compromise to confidential information since sensitive information is kept under the control of a third-party.
- Technical risks such as data leakage, distributed denial of service attacks, loss of encryption keys, and conflicts between customer hardening procedures and cloud platforms.
- Legal risks such as data protection and software licensing risks. Risks not specific to the cloud such as network problems, unauthorized access to data centres, and natural disasters.

Cloud computing raises a range of important policy issues, which include issues of privacy, security, anonymity, telecommunications capacity, government surveillance, reliability, and liability, among others. The basic problem on

cloud is to guarantee protection to the stored data themselves. With current solutions, users typically need to completely trust the cloud providers. In fact, although cloud providers apply security measures to the services they offer, such measures allow them to have full access to the data. For instance, Google Docs or Sales force support encryption of the data both in transit and in storage but they also manage the encryption keys, and therefore users do not have direct control on who can access their data. Whenever data confidentiality needs to be guaranteed even to the provider's views, other solutions have to be considered. Solutions for protecting confidentiality in this scenario typically require encrypting data before releasing them to the cloud providers. Encryption guarantees both confidentiality as well as integrity. For performance reasons, symmetric encryption is usually adopted. While encryption can be effective in many environments, it brings in several complications in scenarios where error retrieval of data needs to be supported. For this reason, recent approaches have put forward the idea of using fragmentation, instead of encryption, when what need to be maintained confidential are the associations among data values, in contrast to the values themselves.

5. IDEA FOR SECURITY

First prevent the privacy constraints. Here maintain the privacy for the database clients and the database owner. Also define information retrieval privacy and outsourcing privacy for them respectively below. Information retrieval privacy starts with a database client retrieving a record with the help of query vector. Assuming that the number of colluding servers in transaction with the client does not exceed the privacy threshold, none of the servers learns anything about query vector through the transaction. The database owner also learns nothing about query vector. Outsourcing privacy refer to the database owner updates the database over time. Neither the database clients nor the untrusted servers learn anything about the update pattern for records they are not entitled to access. There is no communication between different clients in the protocol, so we do not model a privacy notion among different clients. It is important that the cloud service provider undertakes regular backups and give the facility for recovery of information.

6. CONCLUSION

This paper describes the problem of retrieving files from cloud storage and keeping the files secure from untrusted Cloud Service Providers and third parties. For security users can encrypt their files along with keywords using some algorithm. It provides confidentiality for user querying patterns and privacy for user data. On that type of storage the Cloud Service Providers are unaware of the files and the keywords. Ranking scheme improves the retrieval of files. It reduces the time taken to retrieve the files. As a technology, cloud storage and computing has its own properties, some of which are highly attractive to the users. In general, users are interested in this service if all their access and retrieval actions through the cloud providers can be guaranteed by service provider.

7. RECOMMENDATIONS

Based on the findings from the study, it is recommended that multi keyword searching facility provided for data retrieval may be used for retrieval a desired file from a database. When a user wants to retrieve a file, he can enter multiple keywords. The Cloud Server could determine those all files, which contains the specified keyword, without revealing anything about the contents of document.

REFERENCES

[1] Haifeng Lu, Chuan Heng Foh, Yonggang Wen and Jianfei Cai, (2017). Delay-Optimized File Retrieval under LT-Based Cloud Storage. IEEE TRANSACTIONS ON CLOUD COMPUTING.

[2] US Department of Commerce. May 2011. Available online at: http://csrc.nist.gov/ publications /drafts/800-146/Draft-NIST-SP800-146.pdf (Accessed on: November 20, 2012).

[3] D. Catteddu and G. Hogben Cloud computing: Benefits, risks and recommendations for information security. Technical Report, European Network and Information Security Agency, 2009.

[4] Cong Wang, Qian Wang, Kui Ren, Ning Cao, Wenjing Lou, "Towards Secure and Dependable Storage Services in Cloud Computing", International Conference on Services Computing, Vol 5, Issue 2, 06 May 2011, PP 220-232, ISSN :1939-1374, DOI: 10.1109/TSC.2011.24.

[5] Bhabendu Kumar Mohanta, Debasis Gountia, "Fully homomorphic encryption equating to cloud security: An Approach", IOSR Journal of Computer Engineering (IOSRJCE), Vol 9, Jan-Feb 2013, PP 46-50, ISSN: 2278-8727.

[6] Wen, X., Gu, G., Li, Q., Gao, Y., & Zhang, X. (2012, May). Comparison of open-source cloud management platforms: OpenStack and OpenNebula. In Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on (pp. 2457-2461). IEEE.

[7] López, V. F., de la Prieta, F., Ogihara, M., & Wong, D. D. (2012). A model for multi-label classification and ranking of learning objects. Expert Systems with Applications, 39(10), 8878-8884.

[8] N. Padia and M. Parekh, "Cloud Computing Security Issues, in Enterprise Architecture and Its Solutions," International Journal of Computer Application, vol.2, issue 1, pp. 149-155, December 2011.

[9] R. Bhadauria, R. Chaki , N. Chaki, and S. Sanya, "A Survey on Security Issues in Cloud Computing,"IEEE Communications Surveys and Tutorials, pp. 1-15, 2011.

[10] S. Sajithabanu and E. G. P. Raj, "Data Storage Security in Cloud,"International Journal of Computer Science and Technology, vol. 2, issue 4,pp.437-440,Oct. –December 2011.

[11] S. H. Shin, K. Kobara, “Towards secure cloud storage”, Demo forCloudCom2010, Dec 2010.

[12] M. Jensen, J. Schwenk, N. Gruschka, L.L. Iacono, “On Technical Security Issues in Cloud Computing”, IEEE International Conference on Cloud Computing, (CLOUD II 2009), Banglore, India, September 2009, 109-116.

[13] R. Gellman, “Privacy in the clouds: Risks to privacy and confidentiality from cloud computing”, Prepared for the World Privacy Forum, online at http://www.worldprivacyforum.org/pdf/WPF Cloud Privacy Report.pdf, Feb 2009.