International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2064113

Secure Data Duplication Checking with Backup Recovery in Big Data Environments

Gokulakrishnan V[1], Illakiya B[2]

[1]Assistant Professor, Department of CSE, Dhanalakshmi Srinivasan Engineering College , Perambalur, Tamil Nadu, India

[2]Department of CSE, Dhanalakshmi Srinivasan Engineering College , Perambalur, Tamil Nadu, India

Article Info
Volume 6, Issue 4
Page Number: 561-566
Publication Issue :
July-August-2020

Article History
Accepted : 25 Aug 2020
Published : 31 Aug 2020

ABSTRACT

With the rapidly increasing amounts of data produced worldwide, networked and multi- user storage systems are becoming very popular. However, concerns over data security still prevent many users from migrating data to remote storage. The conventional solution is to encrypt the data before it leaves the owner's premises. While sound from a security perspective, this approach prevents the storage provider from effectively applying storage efficiency functions, such as compression and deduplication, which would allow optimal usage of the resources and consequently lower service cost. Client-side data deduplication in particular ensures that multiple uploads of the same content only consume network bandwidth and storage space of a single upload. Deduplication is actively used by a number of backup providers as well as various data services. In this project, we present a scheme that permits the storage without duplication of multiple types of files. And also need the intuition is that outsourced data may require different levels of protection. Based on this idea, we design an encryption scheme that guarantees semantic security for unpopular data and provides weaker security and better storage and bandwidth benefits for popular data. This way, data deduplication can be effective for popular data, whilst semantically secure encryption protects unpopular content. We can use the backup recover system at the time of blocking and also analyze frequent log in access system.

Keywords : Symmetric Key Algorithm, Similarity Checking Algorithm, De-Duplication, Encryption Algorithms.

I. INTRODUCTION

Big data is a term for data sets that are so large or complex that traditional data processing application software's are inadequate to deal with them. Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that’s not the most relevant characteristic of this new data ecosystem." Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on." Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, finance, urban informatics, and business informatics.

Data sets grow rapidly - in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×1018) of data are generated. One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.

Relational database management systems and desktop statistics- and visualization-packages often have difficulty handling big data. The work may require "massively parallel software running on tens, hundreds, or even thousands of servers". What counts as "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

1.2 CHARACTERISTICS:

Big data can be described by the following characteristics:

Volume
The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

Variety
The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

Velocity
In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

Variability
Inconsistency of the data set can hamper processes to handle and manage it.

Veracity
The quality of captured data can vary greatly, affecting accurate analysis.
Factory work and Cyber-physical systems may have a 6C system:

- Cloud (computing and data on demand)
- Cyber (model and memory)
- Content/context (meaning and correlation)
- Community (sharing and collaboration)
- Customization (personalization and value)

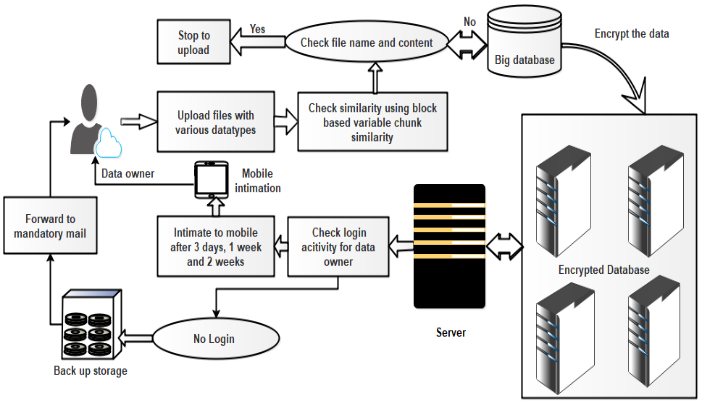
- Data must be processed with advanced tools (analytics and algorithms) to reveal meaningful information.

1.3 SYSTEM DESIGN

A Study of Practical De duplication [1] The scanner first took a consistent snapshot of fixed device (non-removable) file systems with the Volume Shadow Copy Service(VSS). VSS snapshots are both file system and application consistent1. It then recorded meta data about the file system itself, including age, capacity, and space utilization. In addition to reading the ordinary contents of files we also collected a separate set of scans where the files were read using the Win32 Back- up Read API, which includes meta data about the file and would likely be the format used to store file system backups Private Data Deduplication Protocols in Cloud Storage[2] a new notion which we call private data deduplication protocols is introduced and formalized in the context of two-party computations. A feasible result of private data de duplication protocols has been proposed and analyzed. We have shown that the proposed private data de duplication protocol is provably secure in the simulation based frame work assuming that the underlying hash function is collision-resilient, the discrete logarithm is hard and the erasure coding algorithm Ecanerasure up to fraction of the bits in the presence of malicious adversaries.

Reconciling End-to-End Confidentiality and Data Reduction In Cloud Storage[3]Cloud computing has emerged as very beneficial for businesses that are looking to reduce their costs, deploy new applications rapidly or that do not want to maintain their own computational infrastructure. However, recent data breaches in prominent cloud storage providers have caused clients to be increasingly concerned about the confidentiality of their(outsourced)data. There have been cases where client data was exposed to and leaked by cloud provider employees that had physical access to the storage medium.

In this architecture diagram, implement privacy based secure compression scheme to multiple types of data files at the time data storage and retrieval and using acknowledge system to know the status of login time. Data owner can be uploading the files with various file formats. And check the similarity of data using chunk based Map Reduce algorithm. File name and file content should be analyzed. If both the contents are same means, server rejects the files. Otherwise file encrypted using symmetric encryption algorithm. Server can implement self-destruction system to recover the data from blocked account and provide alert system at the time of recovering. User can get all back up files in user alternative mail with real time mobile intimation.


Stop to upload
Yes
Check file name and content
No
Big database
Encrypt the data
Upload files with various datatypes
Check similarity using block based variable chunk similarity
Data owner
Mobile intimation
Forward to mandatory mail
Intimate to mobile after 3 days, 1 week and 2 weeks
Check login acitivity for data owner
Encrypted Database
Server
No Login
Back up storage

1.4 IMPLEMENTATION

1.4.1 CLOUD STORAGE FRAMEWORK

Cloud computing and storage solutions provide users and enterprises with various capabilities to store and process their data in either privately owned, or third-party data centers that may be located far from the user–ranging in distance from across a city to across the world. Cloud computing relies on sharing of resources to achieve coherence. Cloud computing makes computer system resources, especially storage and computing power, available on demand without direct active management by the user. The term is generally used to describe data centers available to many users over the Internet. Large clouds, predominant today, often have functions distributed over multiple locations from central servers. If the connection to the user is relatively close, it may be designated an Edge server. In this framework, can have two types of users such as data owner and data provider. The person or organization that legally owns a cloud service is called a cloud service owner. The cloud service owner can be the cloud consumer, or the cloud provider that owns the cloud within which the cloud service resides. Cloud service provider provides the storage space to the users. Storage space can be shared by multiple data owners. Data owners can be upload the files in storage system for future use.

1.4.2 FILE ENCRYPTION

Encryption is the most effective way to achieve data security. To read an encrypted file, you must have access to a secret key or password that enables you to decrypt it. Unencrypted data is called plain text encrypted data is referred to as cipher text. There are two main types of encryption; asymmetric encryption (also called public-key encryption) and symmetric encryption, then can implement symmetric encryption for encrypt the data files using single key approach. The Encrypted File System, or EFS, provides an additional level of security for files and directories. It provides cryptographic protection of individual files on NTFS file system volumes using a public-key system. Typically, the access control to file and directory objects provided by the Windows security model is sufficient to protect unauthorized access to sensitive information. However, if a laptop that contains sensitive data is lost or stolen, the security protection of that data may be compromised. Encrypting the files increases security. Symmetric key algorithms are algorithms for cryptography that use the same cryptographic keys for both encryption of plaintext and decryption of cipher text. The keys may be identical or there may be a simple transformation to go between the two keys. The keys, in practice, represent a shared secret between two or more parties that can be used to maintain a private information link. Encrypted data can be stored in cloud server.

Algorithm for Encryption

In ECC **Step1:**Key GenerationQ=d*P The key is generated for encryption and decryption purpose

Step2: EncryptionC1=k*P,C2=M+k*QEncryption is done using the above equation. Converting Plaintext into cipher text.

Step3: Decryption:M=C2-d*C1Decryptionisdone using the above equation. Converting the

Cipher text into original form or plaintext.

Proof:M=C2-d*C1M can be represented asC2-d*C1

C2-d*C1=(M+K*Q)-d*(K*Q)(C2=M+K*Qand

C1=K*P)=M+K*d*P-d*K*PAES Analysis Flexibility of key length for a levelof future-fixing The segments of AES are According to the accompanying

- Symmetric key symmetricpiece figure
- 128-piecedata,128/192/256,piecekeys

1.4.3 SIMILARITY CHECKING

In computing, data compression is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage. This technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. In the compression process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. In this module, can check the files using file name with file contents. Encrypted files are splited into chunks. Service provider checks the chunks at the time of uploading files. Data owner only upload original file so save storage space in cloud system. Then can compress all types of files such as text file, document file, image file and also video files.

1.4.4 ALERT SYSTEM

It can design application for alert system for every week. After four weeks completed, if there is no access means the files are automatically sent to alternate mail and mobile which are stored at the time of registration. Server can save huge amount of storage and provide to other users.

1.4.5 BACKUP RECOVERY APPROACH

Admin can check access time for each user login. If user login to the system means, activity is registered in storage, and also monitor each user access. If the user access is paused more than 3 days means, admin automatically send alert to user based on registered mobile numbers. Finally if there is no access in storage system means, backup is generated. And flush the storage space and save storage for server for future use.

1.5 SYSTEM TESTING

Testing is a set activity that can be planned and conducted systematically. Testing begins at the module level and work towards the integration of entire computers based system.

The most common types of testing involved in the development process are:

- Unit Test
- System Test
- Integration Test
- Functional Test

II. CONCLUSION

This system proposed the distributed compression systems to improve the reliability of data while achieving the confidentiality of the users and also shared authority outsourced data with an encryption mechanism. Then implemented the compression systems using the secret sharing scheme and demonstrated that it incurs small encoding/decoding overhead. In this work, have identified a new privacy challenge during data accessing in the cloud computing to achieve privacy-preserving access authority sharing for similarity files. Authentication is established to guarantee data confidentiality and data integrity. User privacy is enhanced by access requests to privately inform the cloud server about the users access desires. The backup recovery scheme is to improve the recovered scheme to avoid the blockages and also refund the amount to unused spaces in cloud system.

III. REFERENCES

[1]. D. T. Meyer, and W. J. Bolosky, “A study of practical deduplication,” Proc. USENIX Conference on File and Storage Technologies 2011

[2]. W. K. Ng, W. Wen, and H. Zhu, “Private data deduplication protocols in cloud storage,” Proc. ACM SAC’12, 2012.

[3]. N. Baracaldo, E. Androulaki, J. Glider, A. Sorniotti, “Reconciling end-to-end confidentiality and data reduction in cloud storage,” Proc. ACM Workshop on Cloud Computing Security, pp. 21–32, 2014.

[4]. P. Anderson, L. Zhang, “Fast and secure laptop backups with encrypted de-duplication,” Proc. USENIX LISA, 2010.

[5]. J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, “Secure deduplication with efficient and reliable convergent key management,” IEEE Transactions on Parallel and Distributed Sytems, Vol. 25, No. 6, 2014.

[6]. J. Li, Y. K. Li, X. Chen, P. Lee, and W. Lou, “A hybrid cloud approach for secure authorized deduplication,” IEEE Transactions on Parallel and Distributed Systems, Vol. 26, No. 5, pp. 1206–1216, 2015.

[7]. M. Bellare, S. Keelveedhi, T. Ristenpart, “DupLESS: Server aided encryption for deduplicated storage,” Proc. USENIX Security Symposium, 2013.

[8]. M. Bellare, S. Keelveedhi, “Interactive message-locked encryption and secure deduplication,” Proc. PKC 2015, pp. 516–538, 2015.

[9]. L. Mingqiang, C. Qin, P.P.C. Lee, and J. Li, “Convergent Dispersal: Toward Storage-Efficient Security in a Cloud-of- Clouds,” Proc. USENIX Conference on Hot Topics in Storage and File Systems, 2014.

[10]. J. Li, X. Chen, X. Huang, S. Tang, Y. Xiang, M. Hassan, and A. Alelaiwi, “Secure Distributed Deduplication Systems with Improved Reliability,” IEEE Transactions on Computer, Vol. 64, No. 2, pp. 3569–3579,201

Cite this article as :

Gokulakrishnan V, Illakiya B, "Secure Data Duplication Checking with Backup Recovery in Big Data Environments", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 561-566, July-August 2020. Available at doi : https://doi.org/10.32628/CSEIT2064113
Journal URL : http://ijsrcseit.com/CSEIT2064113