Design Engineering

ISSN: 0011-9342 | Year 2021
Issue: 7 | Pages: 11404- 11412

AN EFFICEINT HYBRID CLOUD STRUCTURE TO AVOID THE REDUNDENCY ISSUES IN HYBRID CLOUD COMPUTING

J.V.S. ARUNDATHI*[1] and K.V.V. SATYANARAYANA[1]

*[1]Department of Computer Science and Engineering, K L University, Guntur Dist., Andhra Pradesh, INDIA -522502
Email: jvs.arundathi@gmail.com

ABSTRACT:

Big data has large amount of data representing personal information, the biggest challenge faced by it is from safety opinion that is the defence of privacy of operator. By the growth of data, information in large scale the storage has moved from hard drive to cloud storage. Cloud storage system is basically distributed file system, which solves the problem of storage of large data whereas it faces the challenge of storing duplicate data in the storage. Most of the file systems are created to deal with storing and retrieval of data available on the storage. Among different works done on the cloud file system only little work on the De-duplication techniques at application level, especially for HDFS . In this below work represents the design of file de-duplication framework on HDFS for cloud application developer. Proposed AVOID THE REDUNDENCY ISSUES IN HYBRID CLOUD COMPUTING (ATRIHCC) which de-duplicates the data online that improves storage space usage and reduces the duplicates . At the end of the paper, testing of disk usage and file upload performance on the proposed method and the comparison of HDFS.

Keywords: HDFS, Cloud Computing, Data Deduplication, Data Integrity.

1. Introduction:

The term BIG DATA is for larger data sets which are far more complex in traditional data processing applications which are Inefficient to deal . Most of the challenges include search, store, share, and transfer and user privacy protection. Hadoop is an open-source outline to stock information then track requests on different bunches of commodity hardware. It has large storage for different types of data, high processing power and can perform multiple tasks virtually. Hadoop runs many applications with many commodity hardware nodes and can handle lots of data. HDFS is java based scalable system that stores data across many machines without former arrangement. The architecture of Hadoop is master slave, in which name node acts as master and data nodes as slaves. HDFS breaks the data file into fixed blocks and that data is stored on the data nodes. The blocks are mapped based on the name nodes, which also manages metadata (data about data) and Namespace. Most of the cloud storage accommodations apply de-duplication to

ISSN: 0011-9342 | Year 2021
Issue: 7 | Pages: 11404- 11412

reduce maintenance cost. For users point of view this raises privacy and security issues due to outsourcing.

2. Literature Survey:

"A comparative study on data de duplication techniques in cloud storage", B.Tirapathi Reddy, et.al[1], deals with various mechanisms available to eliminate redundant copies of the data, and also addresses the drawback and Advantages of all the mechanisms."CloudDedup: safe deduplication through encoded information aimed at cloud packing", Pasquale puzio, et.al [2], plans a framework which accomplishes secrecy and empowers block level deduplication in the meantime. Our framework is based on top of convergent encryption. We demonstrated that it merits performing block level deduplication rather than file level deduplication. Evades COF and LRI attacks (confirmation of record) (take in the rest of the data). "Performance Evaluation of Various Data Deduplication Schemes in Cloud Storage", B.Tirapathi Reddy, et.al [3], addresses an efficient data deduplication mechanism, that preserves the confidentiality of data and privacy of the data owners, with efficient key management. "A Hybrid cloud approach for secure authorized de-duplication", Jagadish, et.al [4], Address the issue of authorized data duplication . Deals with hybrid cloud and thus possess the benefits of both the public and private cloud. The duplicate patterned marks of forms are shaped by the secluded cloud server with isolated keys. "Dynamic Secure Deduplication in Cloud Using Genetic Programming" ,K. V. Pandu Ranga Rao, et.al [5], Data Engineering and Intelligent Computing ,In this paper, a Genetic Programming approach has been proposed to manage record deduplication that joins several bits of confirmation driven out from the data substance to find a deduplication point of confinement **"**Secure and constant cost public cloud storage auditing with Deduplication", Jiawel yuan, et.al [6], Outperforms current POR and PDP systems with deduplication. Consistent price system that accomplishes protected community information honesty. "Policy Based Data Deduplication In Cloud Storage" Sowmya, et.al [7], speaks about the Deduplication techniques like file level, block level and byte level data Deduplication mechanism and also deals with Dekey using ramp secrete key sharing scheme and dupless key scheme. "Provable ownership of file in deduplication cloud storage", Chao yang, et.al [8], proposes a plan that can produce Provable Ownership for File [POF] and keep up a high discovery likelihood of the client misbehaviour. Very proficient in lessening the weight of the client . "Secure Deduplication with Efficient and Reliable Convergent Key Management" J.Li, et.al [9], future dekey, a productive then compacted management combine aimed at secure deduplication. They perform dekey using the slope secret distribution idea then show that it incurs slight programming/decoding above distinguished by the scheme broadcast overhead in the general transmission/copy actions. "A Secure data deduplication scheme for cloud storage", J.Stanek, et.al [10], isolated operators contract out their information to cloud storing breadwinners. Late information rupture episodes branded on encryption an undeniably noticeable necessity.

ISSN: 0011-9342 | Year 2021
Issue: 7 | Pages: 11404- 11412

Data deduplication container be viable for prevalent information, while semantically safe encryption ensures disagreeable content. “A reverse deduplication storage system optimized for reads to latest backups”, C.Ng , et.al [11] had present RevDedup, a deduplication framework intended for VM circle picture reinforcement in virtualization conditions. RevDedup has a few plan objectives: high stockpiling proficiency, low memory use, high reinforcement execution, and high re establish execution for the most recent reinforcements. They broadly assess our RevDedup model utilizing various responsibilities and approve our plan objectives. “Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage”, Junbeom Hur, et.al [12], a original server-side deduplication scheme for encrypted data. It permits the cloud server to control access to outsourced data even notwithstanding when the proprietorship changes progressively by misusing randomized convergent encryption and secure possession gather key distribution. This counteracts data leakage exclusively to disavowed clients despite the fact that they beforehand possessed that information, additionally to a legitimate yet inquisitive distributed storage server. “Data Deduplication In Cloud storage Using Dynamic Perfect Hash Functions”, B.Tirapathi Reddy, et.al [13],addresses the techniques to provide secure deduplication of data, by taking the popularity of data items into the count, and assuming that data items require different levels of security based on popularity. a mechanism is proposed to ensure secure data deduplication leveraging the advantages of dynamic perfect hash techniques .

“Enhanced Dynamic Whole File De-Duplication (DWFD) for Space Optimization in Private Cloud Storage Backup”, M. Shyamala Devi, et.al [14], provides dynamic space optimization in private cloud storage backup as well as increase the throughput and de-duplication efficiency.“Deduplication on Encrypted big data in cloud”, Zhen yan, et.al [15], De-duplicate encrypted data stored in cloud based on proxy re- encryption, Avoids encrypting of data while uploading thus saves bandwidth. Over came the brute force attack.

3. Implementation:

We tend to achieve incipient dispersed de-duplication schemes by advanced dependability in which the information chunks are dispersed across HDFS storage systems ATRIHCC(AVOID THE REDUNDENCY ISSUES IN HYBRID CLOUD COMPUTING)technique is employed for secure de-duplication. The proficiency which has been proposed to expunge the shortcomings of the existing deduplication concept is convergent encryption, proof of ownership and efficient key management schemes. Therefore the substrate deduplication is performed at both file level and block level and we define a HDFS Master machine to enhance the security and storage. Contributions:

- Data integrity and verification using efficient key management solution through the metadata manager.
- Preserves confidentiality and privacy against malicious storage providers by encrypting the chunks which are at random storage.

ISSN: 0011-9342 | Year 2021
Issue: 7 | Pages: 11404- 11412

- Proposed hybrid model ensures the both block-level and file-level de duplication.

Design of ATRIHCC(Avoid The Redundancy Issues In Hybrid Cloud Computing):

In applications, small errors are stood, such as mesh info removal and lexical and semantic examination. Repeated collisions are nearly impossible to occur, and the application's judgment results are not affected. A hash with the same fingerprint is considered a duplicate file. As a result, source duplication container decrease net bandwidth, but upload period, and smooth decrease the load on HBase and primary nodes. By comparing the HASH worth to the HBase file, you can rapidly know if the gratified of the site has changed. It can save a binary comparison time and retrieve the target directly from the SHA value generated at the basis. If the basis SHA does not changed, all uploaded content is saved at the same time. If the hash generation algorithm is implanted from the source host, the load on the primary node and HMaster can be alleviated, and the original web page retrieval becomes incremental update retrieval, which cannot only save time and bandwidth but also reduce the load on the server. Applications can be fault-tolerant, like video sharing or many file shares, and ATRIHCC provides from top to bottom performance and little storing ingesting answer aimed at folder deletion. ATRIHCC design as shown in Fig. 1, calculated by hash generator upload file hash value for the first time, and stored in a contrast of the hash price of Hbase, see if there exists the hash value, requests for data deduplication or returns the command, if the hash value does not exist, then perform the normal read/write command, FD-HDFS is designed to save storage space when a hash collision occurs.

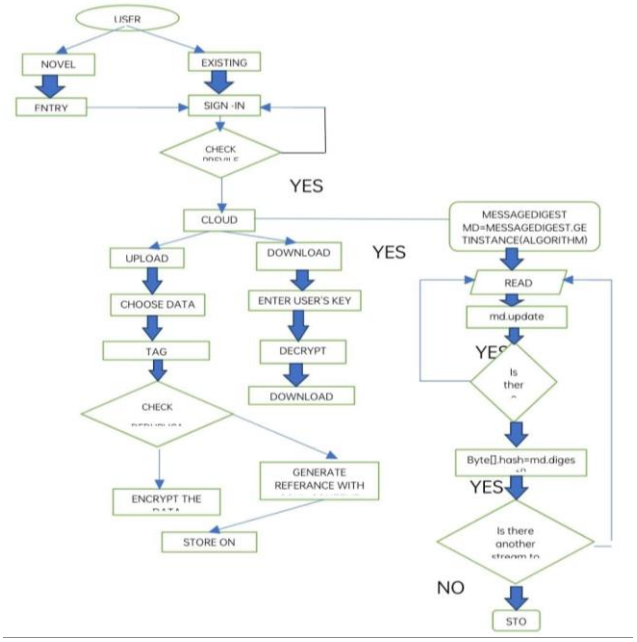

Fig1: The Proposed Design (ATRIHCC)

ISSN: 0011-9342 | Year 2021
Issue: 7 | Pages: 11404- 11412

Similarity Detection:

This is the interaction done in the backend and is liable for getting comparative information by the proposed strategy.. For proposed-HDFS compose guidelines, if a comparable record exists at HBase, it implies that in any event one document has similar substance put away in HDFS. On the off chance that the current document has a similar way as the given record, there is no compelling reason to compose the record to HDFS. All things considered, no more document moves are needed for HDFS from the customer to the HDFS worker on the grounds that HDFS can recover the record substance from itself. On the off chance that the current document way is unique in relation to the one given, square should make another record at HBase and store the document in a transitory document pool to forestall hash crashes and guarantee the dependability of additional document content recovery. Additionally, for document read guidelines, the information similitude indicator can help improve record download execution if the hash esteem is in the HBase table. Work on the ATRIHCC execution.

Data Verification and Indexing:

Data verification and indexing Strategy using Semantics for distributed deduplication systems. File semantic information includes the information about a file, file ownership, document visitors, file size and other information. This information reflects the similarity degree of files. The higher the similarity degree , the higher the probability of duplicate data. In this module strategy of the distributed deduplication system is designed by using the file name as the semantic information of the file. When the primary node receives the superblock generated by the backup server, the superblock data needs to be distributed to different heavy deletion points at the back end based on the file semantic information. The algorithm needs to establish and maintain a semantic data routing index table, which holds the semantic routing information related to the file name. The master node includes the semantic routing index and the container index. The semantic index includes the mapping between the file semantics and the target container. The reason why the algorithm chooses to migrate the smallest container instead of the largest container is that migrating the largest container is more likely to overload other nodes and cause a load balancing process for a series of nodes. After the data migration process is complete, the storage node needs to return the metadata of the migrated files to the master node.

Algorithm ATRIHCC:

Input: filename, chunking Results, Num of Blocks

Output: Node

- Container = Index. Find (filename);
- if Container == NULL then
- BlockChunckIndex = Container Index (chunkingResults)

ISSN: 0011-9342 | Year 2021
Issue: 7 | Pages: 11404- 11412

- totalmeetcontainer = BlockChunckIndex % NumOfContainers;
- insert (filename, Container);
- end if
- Node = CIndex.find (container);
- while sizeofNode (Node) > (Node Size()*(1+Threshold)) do
- Snoed = FSNode ();
- SContainer = FSContainer (ANode);
- MContainer (SContainer, MNode);
- MChunkIndex (SContainer, MNode);
- end while
- return ANode;

4. Results &Discussion:

In this paper, we have taken the consideration of the file level deduplication for simplicity. For this, we refer a data copy to be a whole file and file-level deduplication which eliminates the storage of the redundant files. Operationally, to upload a file, a user first performs the file-level duplicate check by sending a request to the CSP. If the file is a duplicate, then all its blocks must be duplicates as well; otherwise, the user further performs the block-level duplicate check and identifies the unique blocks to be uploaded. Each data copy (i.e., a file or a block) is associated with a token for the duplicate check.

• S-CSP. This is an entity that provides a data storage service in public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data. In this paper, we assume that S-CSP is always online and has abundant storage capacity and computation power in Fig. 2

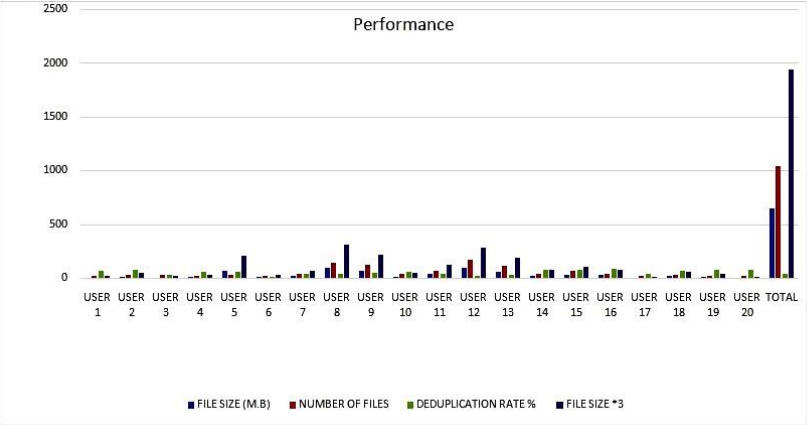

Fig.2 Storage capacity and computation power

ISSN: 0011-9342 | Year 2021
Issue: 7 | Pages: 11404- 11412

Data Users. A user is an entity that wants to outsource data storage to the S-CSP and access thedata later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. In the authorized deduplication system, each user is issued a set of privileges in the setup of the system. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

• **Private Cloud.** Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating user's secure usage of cloud service. Specifically, since the computing resources at data user/owner side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud. The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users. The interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively. This is a novel architecture for data deduplication in cloud computing, which consists of a twin clouds (i.e., the public cloud and the private cloud). Actually, this hybrid cloud concept has attracted more attention from the industry recently. Alternatively, the trusted private cloud could be a cluster of virtual cryptographic co-processors, which are offered as a service by a third party and provide the necessary hardware based security features to implements a remote execution environment trusted by the users given in Table 1.

Table 1: Performance Analysis

Host name	Filesize(m.b)	Number of files	Deduplicate rate%	File size*3
USER1	7.2	18	66	21.6
USER2	14.1	25	76	42.3
USER3	6.3	30	30	18.9
USER4	8.3	22	59	24.9
USER5	68.9	24	58	206.7
USER6	10	22	13	30
USER7	23.2	38	42	69.6
USER8	101.6	144	41	304.8
USER9	70.3	120	50	210.9
USER10	15.6	33	57	46.8
USER11	41.9	62	38	125.7
USER12	94.3	164	17	282.9
USER13	61.1	108	26	183.3
USER14	25.3	38	73	75.9
USER15	34.7	62	74	104.1

ISSN: 0011-9342 | Year 2021
Issue: 7 | Pages: 11404- 11412

USER16	34.4	42	83	73.2
USER17	3.6	15	33	10.8
USER18	20.1	31	61	60.3
USER19	12.1	20	75	36.3
USER20	3.7	18	72	11.1
TOTAL	646.7	1036	40.9	1940.1

5. CONCLUSION

In this paper, the proposed method of de duplication with file level and block level, provide high excellence in which the data lumps are properly inclined over HDFS storage. The security then privacy is well managed with reliable key administration by enforcing secure de duplication and security of labels consistency. We also presented several new deduplication mechanisms that support authorized duplicate check in hybrid cloud architecture, in which the duplicate check tokens of files are generated by the private cloud server with private keys. The security analysis of proposed system demonstrates that our schemes are secure in terms of insider and outsider attacks which are more susceptible in most of the existing systems. As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme and conduct test-bed experiments on our prototype with the help of Drop-Box API and Cloud HQ. We conclude that our authorized duplicate check scheme incurs minimal overhead as observed over convergent encryption and network transfer.

REFERENCES

[1] B.Tirapathi Reddy1, U.Ramya 2, Dr.M.V.P Chandra Sekhar3: “A Comparative Study On Data Deduplication Techniques In Cloud Storage”, Published in IJPT, 2016.

[2] Pasquale puzio1, Refik Molva2, Melek onen3: “Cloud Dedup - Secure Deduplication with Encrypted Data for Cloud Storage”, In IEEE 5th International Conference, Dec 2013.

[3] B.Tirapathi Reddy1, Dr.M.V.P.Chandra Sekhara Rao2:“Performance Evaluation Of Various Data Deduplication Schemes In Cloud Storage”, Published in International Journal of Pure and Applied MathematicsVolume 116, 2017.

[4] Jagadish1, Dr.Suvurna Nandyal2: “A Hybrid Cloud Approach For Secure Authorized De-Duplication”. Published in International Journal of Science and Research (IJSR), 2013.

[5] K. V. Pandu Ranga Rao1,V. Krishna Reddy2,S. K. Yakoob3, “Dynamic Secure Deduplication in Cloud Using Genetic Programming”, Published in Data Engineering and Intelligent Computing ,2017.

[6] Jiawel Yuan1, Shucheng yu2 : “Secure And Constant Cost Public Cloud Storage Auditing With Deduplication” In IEEE Conference, published in communication and network security, 2013.

[7] Sowmya1 , Revathi2 , Dr. Thirumala Rao3, “Policy Based Data Deduplication In Cloud Storage”,Publishd in International Journal Of Pure And Applied Mathematics ,Volume 116,2017.

[8] Chao Yang1, Jianren2, Jianfengma3: “Provable Ownership Of File In Deduplication Cloud Storage”. Published in Global Communication Conference, 2013.

[9] J.Li1 , X.Chen2 , M.Li3 , J.Li4 , P.Lee5 , and W.Lou6 : “Secure Deduplication with Efficient and Reliable Convergent Key Management”. In IEEE Transactions on parallel and Distributed systems, 2013.

ISSN: 0011-9342 | Year 2021
Issue: 7 | Pages: 11404- 11412

[10] J.Stanek1, A.Sorniotti2, E.Androulaki3, and L.Kenel4: "A Secure Data Deduplication Scheme for Cloud Storage". In Technical Report, 2013.

[11] C.Ng and P.Lee. Revdedup: "A Reverse Deduplication Storage System Optimized For Reads To Latest Backups". In Proc of APSYS, Apr 2013.

[12] Junbeom Hur1, Dongyoung Koo2, Youngjoo Shin3, and Kyungtae Kang4: "Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage", In IEEE Transactions on Knowledge and Data Engineering, June 2016.

[13] B.Tirapathi Reddy1,M.V.P. Chandra Sekhara Rao2, "Data Deduplication In Cloud Storage Using Dynamic Perfect Hash Functions",In Journal of Advanced Research inDynamical and Control Systems, vol-9,2017.

[14] M. Shyamala Devi1, V. Vimal Khanna2, A. Naveen Bhalaji3: "Enhanced Dynamic Whole File De-Duplication (DWFD) for Space Optimization in Private Cloud Storage Backup", in International Journal of Machine Learning and Computing, April 2014

[15] Zhen Yan1, Wenxiu Ding2, Robert.H.Deng3 : "De-Duplication On Encrypted Big Data In Cloud ". In IEEE transaction on big data Vol.2, No.2, April – June 2016.

AUTHORS PROFILE

Ms. J V S ARUNDATHI is currently a PhD scholar in the Department of Computer Science and Engineering, KL Deemed To Be University, Guntur, India. Her research interest includes Information Security and Cloud Computing Security, IOT.

Dr. K.V.V. SATYANARAYANA is a Professor and M.Tech Coordinator in the Department of Computer Science and Engineering, KL Deemed To B e University, Vaddeswaram, Guntur District, India. His research interest includes Bioinformatics and Cloud Computing. He is the Senior Life Member of Computer Society of India.