Deduplication in Big Data

Seema Kamble[1] Komal Ghadage[2] Dipali Devkar[3] Onkar Jadhav[4] Prof. Krushndeo Belerao[5]
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Trinity College of Engineering & Research, Pune, Maharashtra, India

Abstract— Deduplication has become a widely organize tools in cloud data centers to get better IT resources effectiveness. However, traditional techniques face a great challenge in big data deduplication to strike a rational tradeoff between the variance goals of scalable deduplication throughput and high duplicate subvert ratio. We are suggesting a scalable distributed deduplication framework in cloud environment, to meet this challenge, data similarity and locality to optimize spread deduplication with inter-node two-tiered data routing and intra-node application-aware deduplication. It first dispense data at file level, then assigns related data to the same storage node to maintain high global deduplication efficiency, meanwhile balances the workload across nodes. Our experimental evaluation of data Dedupe against state-of-the-art, driven by real-world datasets, demonstrates that Data Dedupe achieve the highest global deduplication efficiency with a higher global deduplication helpfulness than the high-overhead and badly scalable fixed scheme, but at an overhead only somewhat higher than that of the scalable but low duplicate-elimination-ratio approaches.

Key words: Hadoop, HDFS, Data Deduplication, MD5 Algorithm, Application Aware Routing Algorithm

I. INTRODUCTION

Big data is an outer term for the non-traditional strategy and technologies needed to gather, organize, process, and gather insights from big datasets. The problem of data that uses more storage of a single computer is obvious, the popularity, scale and value of this type of computing has widely expand in past few years.

Recent technological advancement in cloud computing, internet of things and social network, have led to a flood of data from unique domains over the past two decades. Cloud data centers are awash in digital data, easily amassing petabytes and even hexabytes of information, and the complexity of data management escalates in big data. However, IDC data shows that nearly 75% of our digital world is a copy.

A. Needs

Deduplication technique to manage the data deluge under the changes in storage architecture to meet the service level agreement requirements of cloud storage. It is generally in favor of source inline deduplication design, because it can immediately identify and eliminate duplicates in datasets at the source of data generation, and hence significantly reduce physical storage capacity requirements and save network bandwidth during data transfer. It performs in a typical distributed deduplication framework to satisfy scalable capacity and performance requirements in massive data. The framework includes inter-node data assignment from clients to multiple deduplication storage nodes by a data routing scheme, and independent intra-node redundancy suppression in individual storage nodes.

B. Basic Concept

Number of files created per day is increased due to rapid rise in number of users. In cloud computing environment most of the communication is done through the file processing, and hence it becomes very crucial and significant to provide efficient approach for data security. In this research we are concentrating on data deduplication to provide efficient security service in cloud computing. Data deduplication is nothing but data compression technique which is used to remove the duplicate copies of echoing data. This methodology is regularly used for dropping the storage space and save bandwidth under cloud. Also, along with deduplication for data protection and privacy the encryption methods are used. In this paper we are going to study about the data deduplication to provide the data security by keeping the information of authorized users in the cloud architecture. Different new deduplication techniques offered for authorized person to check the duplicate data.

Classify a new privacy challenge in cloud storage, and address a delicate privacy issue during a user challenging the cloud server for data distribution, in which the challenged request itself cannot expose the user's privacy no matter whether or not it can obtain the access authority. Propose an validation protocol to increase a user's access request related privacy, and the shared access authority is achieved by unsigned access request matching mechanism. Apply cipher text-policy attribute based access control to recognize that a user can dependably access its own data fields, and adopt the proxy re-encryption to provide temp authorized data sharing among multiple users.

C. Hadoop

Hadoop is an open source, which is based on Java programming framework with the purpose of supports the processing and storage of large amount of data sets in a distributed computing environment. It is element of the Apache project sponsored by the Apache Software Foundation.

Hadoop architecture has default block size in distributed system is 64 MB and if you want extend it that will be 120 MB in new version. In Hadoop there are three default replication factors that have copy of the single file in other two replication. There are basic components in Hadoop Architecture as name node, data node.

II. RELATED WORK

A. A Survey on Secure Authorized Deduplication Systems

Cloud Storage Systems are fetching progressively more popular with the constant and exponential increase of the number of users and the size of data. Data deduplication becomes more and more inevitability for cloud storage provider. Data deduplication is one of the techniques which is important for eliminating duplicate copies of repeating data. It has been widely used in the cloud storage to decrease

All rights reserved by www.ijsrd.com

the quantity of storage space and save bandwidth. The advantage of deduplication comes with high cost in security and privacy challenges. The future scheme in this paper not only reduce the cloud storage capability but also improve the speed of data deduplication. Our system is works on the authorized duplicate check to incur minimal overhead, compared to other operations and it also show the encryption for deduplicated storage, by generating the encrypted key.

B. Proposed Methodology

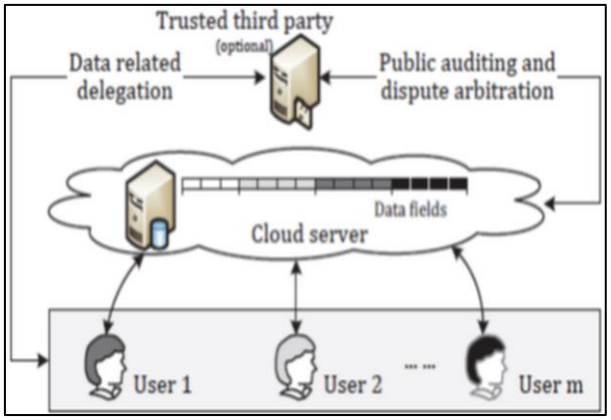

Fig. 1:

C. Architecture

S-CSP-This is an unit that provides a data storage service in communal cloud. The S-CSP provides the outsourcing service of data and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of unnecessary data via deduplication and keeps only distinctive data.

Data Users- A user is a unit that wants to out- source data storage to the S-CSP and access the data later. In a storage system sustaining deduplication, the user can only upload unique data but cannot upload any replica data, which may be owned by the same user or different users.

Private Cloud- as we are working on the cloud storage, unlike the traditional deduplication system, we are introducing new entity to facilitate the user's security in the cloud storage.

1) Technique

1) Symmetric Encryption
2) Convergent Encryption
3) Proof of ownership
4) Identification Protocol

2) Advantages

1) Authorized duplication check
2) Data privacy
3) Improve bandwidth effectiveness

3) Disadvantages

1) Custom encryption & Convergent encryption methods are not semantically secure.

III. System Architecture

A. Bigdata

The term data with high volume, high velocity data and variety of data. Bigdata is concept; the data which contain large amount of data .it help to us take very clear decision in daily. Bigdata deal with the all problem related storing data with the capacity large amount of data we can do processing in large database in bigdata.in that we can deal with both the data types structured and unstructured. Cloud storage is option to store data with advantage flexibility and scalable. Cloud storage is cloud computing from which data can accessed by remote server from the internet.

B. HDFS

Hadoop is a system which provides Distributed file system framework for storing and transforming large database. It also uses MapReduce paradigm for analysis of the data.

C. Name Node

Name node consists the information about data node. All cluster is depend on name node, if the name node is offline or switched off then there is no meaning of hadoop clusters that means hadoop is not working. With the name node secondary name node is also present for the purpose backup of name node but there is issue with updating information within limit. Name node job tracker tracking all the Incoming information and assign data nodes for that information. In that system there is one task tracker who chess the task which assigned to data node and it will track the task.

D. System Working

- User will sign in to the system.
- It will get ask for the uploading and downloading.
- For uploading it will ask for browse the file.
- After that system will read that file line by line.
- It then searches in system for replica of same file. If it is present it will show you the key generated for that file. And will not upload new file.
- If the file is not present, it generates new key for that file and store it with the file in the system.
- For downloading, user will first ask for the authentication.
- Authorized user can only access that file.

E. Basic Architecture of Deduplication Application:

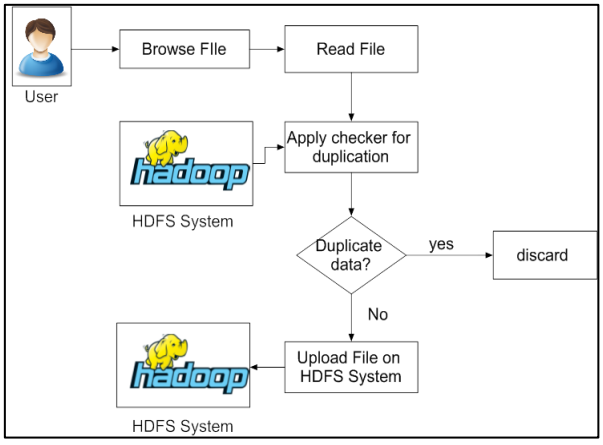

Fig. 2:

IV. Algorithm

1) Big data deduplication algorithm

Input: the full name of a file, fullname, and a list of all dedupe storage nodes {S1, S2, …, SN}

All rights reserved by *www.ijsrd.com*

Output: a ID list of application storage node, ID_list={A1, A2, … , Am}

A. Algorithm

1) Extract the filename extension as the application type from the file full name fullname, sent from client side;
2) Query the application route table in director, and find the dedupe storage node Ai that have stored the same type of application data; We get the corresponding application storage nodes ID_list={A1, A2, … , Am} ⊆ {S1, S2, …, SN};
3) Check the node list: if ID_list= ∅ or all nodes in ID_list are overloaded, then add the dedupe storage node SL with lightest workload into the list ID_list={SL};
4) Return the result ID_list to the client.

2. SHA-1 –

SHA-1 is a Secure Hash Algorithm. It produces a 160-bit (20-byte) hash value known as a message digests from the taken input string, typically a hexadecimal number 40 digit long.
This is examples of SHA-1 message digests in hexadecimal text encoding.
SHA1 (“The quick brown fox jumps over the lazy dog”)
Gives hexadecimal:
2fd4e1c67a2d28fced849ee1bb76e7391b93eb12
Even if you make a small change in the message, result in many bits changes due to the avalanche effect. For example, changing dog to cog produces a hash with different values for 81 of the 160 bits:
SHA1 (“The quick brown fox jumps over the lazy cog”)
Gives hexadecimal:
de9f2c7fd25e1b3afad3e85a0bd17d9b100db4b3

V. Conclusion

Hence we succeed in recognize duplicate file in big data using hadoop and able to free up the memory space and reduced duplication of data in the cloud environment, we are able to upload and download the file with authorized entry of users in the system, by using distributed framework.

Acknowlegement

We take this opportunity to thank our project guide Prof. Krushnadeo Belerao and Head of the Department Prof. Pawan Kulkarni for their valuable guidance and for providing all the necessary facilities, which were indispensable in the completion of this project report. We are also thankful to all the staff members of the Department of Computer of ”Trinity College of Engineering and Research, Pune” for their valuable time, support, comments, suggestions and persuasion. We would also like to thank the institute for providing the necessary facilities, Internet access and important books.

References

[1] H. Biggar, White Paper, the Enterprise Strategy Group, Feb. 2007 “Experiencing Data De-Duplication: Improving Efficiency and Reducing Capacity Requirements,”
[2] J. Gantz, D. Reinsel, White Paper, IDC, May 2010, “The Digital Universe Decade-Are You Ready?” .
[3] K.R. Jayaram, C. Peng, Z. Zhang, M. Kim, H. Chen, H. Lei. ,” Proc. Of the ACM/IFIP/USENIX Middleware Industry Track Workshop (Middleware’11), Dec. 2011. , “An Empirical Analysis of Similarity in Virtual Machine Images.
[4] K. Srinivasan, T. Bisson, G. Goodson, and K. Voruganti. Proc. of the 10th USENIX Conference on File and Storage Technologies (FAST’12). Feb. 2012, “iDedup: Latency-aware, inline data deduplication for prima-ry storage,”.
[5] P. Shilane, M. Huang, G. Wallace, and W. Hsu. ACM communication on Storage (TOS), 8(4): 915-921, Nov. 2012. “WAN opti-mized replication of backup datasets using stream-informed delta compression,”
[6] D. Bhagwat, D.D. Long, K. Eshghi , M. Lillibridge, Proc. of the 17th IEEE International Symposium on Modeling, Analysis and recreation of Computer and Tele-communication Systems (MASCOTS’09), pp.1-9, Sep. 2009. “Extreme Binning: Scalable, Parallel Deduplication for Chunk-based File Backup”.
[7] F. Douglis, H. Qian, D. BhardwajP. Shilane, Proc. of the 25th USENIX Conf. on Large Installation System Administration (LISA’11), pp.151-168, Dec. 2012, “Content-aware Load Balancing for Distributed Backup,”.
[8] H. Patterson, W. Dong, F. Douglis, K. Li, S. Reddy, P. Shilane, Proc. of the 9th USENIX Conf. on File and Storage Tech-nologies (FAST’11), pp. 15-29, Feb. 2011,“Tradeoffs in Scalable Data Routing for Deduplication Clus-ters,” .
[9] C. Dubnicki, L. Gryz, L. Heldt, M. Kaczmarczyk, W. Kilian, P. Strzelczak, J. Szczepko-wski, C. Ungureanu, M. Welnicki, Proc. of the 7th USENIX Conf. on File and storage space Technologies (FAST‘09), pp. 197-210, Feb. 2009,“HYDRAstor: a Scalable Secondary Storage,”.
[10] J. Wei, H. Jiang, K. Zhou, D. Feng, Proc. of the 26th IEEE Conf. on Mass Storage Systems and Technologies (MSST’10), pp. 1-14, May 2010,“MAD2: A Scalable High Throughput Exact Deduplication Approach for Network Backup Services,” .
[11] T. Yang, H. Jiang, D. Feng, Z. Niu, K. Zhou, Y. Wan, Proc. of the 24th IEEE Internation-al Parallel and Distributed Processing Symposium (IPDPS’10), pp. 1-12, Apr. 2010. “DEBAR: a Scalable High-Performance Deduplication Storage System for Backup and Archiving,”.
[12] D. Meister ,H. Kaiser, A. Brinkmann, S. Effert, Proc. of the 28th IEEE conference on Mass Storage Systems and Technologies (MSST’12), pp. 1-12, Apr. 2012 “Design of an Exact Data Deduplication Cluster,”.

All rights reserved by *www.ijsrd.com*