International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Secure Data Compression Scheme for Scalable Data in Dynamic Cloud Environments

B. Raja Rajeshwari*[1], B. Sangeetha[2]
[1]PG Scholar & Computer Science and Engineering, Anna University/Sir ISSAC Newton college of Engineering and Technology, Nagapattinam, Tamilnadu, India
[2]Assistant Professor of Computer Science and Engineering, Anna University/Sir ISSAC Newton College of Engineering and Technology, Nagapattinam, Tamilnadu, India

ABSTRACT

With the rapidly growing amounts of facts produced global, networked and multi-consumer storage systems have become very famous. However, worries over information safety still prevent many customers from migrating facts to far-flung garage. The conventional solution is to encrypt the information earlier than it leaves the owner's premises. While sound from a safety angle, this method prevents the garage issuer from effectively making use of storage efficiency capabilities, which includes compression and deduplication, which could permit best utilization of the resources and consequently lower carrier fee. Client-aspect data deduplication particularly ensures that more than one uploads of the equal content handiest devour community bandwidth and garage area of a unmarried upload. A number of cloud backup providers as well as various cloud services actively use deduplication. Unfortunately, encrypted facts are pseudorandom and consequently cannot be deduplicated: therefore, cutting-edge schemes need to completely sacrifice both security and garage performance. In this paper, we present schemes that permit a greater quality-grained change-off in records chunk similarity. The instinct is that outsourced records may additionally require exceptional degrees of safety, relying on how popular it is miles content material shared through many users. Various deduplication schemes are analyze and provide experimental outcomes that suggest proposed cozy facts bite similarity provide improved effects in real time cloud environments.

Keywords: Data chunks, Similarity matching, parallel processing, Data security, Data compression.

I. INTRODUCTION

Now a day there is boom in records. With infinite storage area offer by means of cloud carrier provider customers generally tend to apply as much area as they are able to and vendors constantly search for techniques aimed to limit redundant records and maximize area financial savings. Users will get admission to statistics in step with their desires and maximum customer's access identical facts repeatedly, the fee of computation, software hosting, content material storage and transport is reduced drastically. The cloud makes it possible to be able to get admission to your information from anywhere at any time. Cloud provides advantages consisting of, flexibility, distracter, restoration, software program updates routinely, pay according to use version and cost discount. The cloud gets rid of the want in an effort to be inside the identical bodily location as the hardware that shops your records. Each issuer serves a particular characteristic, giving customers extra or much less manipulate over their cloud depending on the type. Your cloud wishes will vary relying on how you plan to use the gap and assets associated with the cloud. Cloud computing refers to the use of computers which access Internet locations for

computing power, storage and programs, without a want for the character get right of entry to factors to hold any of the infrastructure. Data deduplication is a technique for reducing the amount of garage space an organization needs to shop its statistics. In most organizations, the garage structures include reproduction copies of many pieces of statistics. For example, the identical file can be stored in numerous specific locations by way of exclusive users, or greater documents that are not same might also still consist of an awful lot of the same records. Along with low ownership prices and flexibility, users require the protection in their statistics and confidentiality ensures via encryption. To make statistics control scalable deduplication we are use Encryption for relaxed deduplication offerings. Unfortunately, deduplication and encryption are conflicting technology. While the aim of deduplication is to locate identical records segments and shop them best as soon as, the end result of encryption is to make two it records segments in distinguishable after being encrypted.A manner that if information is encrypted through customers in a fashionable way as like shared authority, the cloud storage company cannot follow deduplication for the reason that identical data segments could be one of a kind after encryption. On the other hand, if statistics are not encrypted by customers, confidentiality through cannot be guaranteed and records are not blanketed against curious cloud garage providers. Although encryption seems to be an amazing candidate to attain confidentiality and deduplication at the equal time, it sadly suffers from numerous well-known weaknesses.

The confidentiality issue can be treated through encrypting touchy information earlier than outsourcing to remote servers. Along with low possession fees and flexibility, users require the safety in their facts and confidentiality ensures via encryption. In this paper, we cope with the stated confinement problem to propose a shared authority to the files which Deduplicted primarily based privacy preserving authentication for the cloud statistics garage, which realizes authentication and authorization without compromising a person's non-public records. The basic data chunk similarity is shown in figure 1.

Data chunk similarity

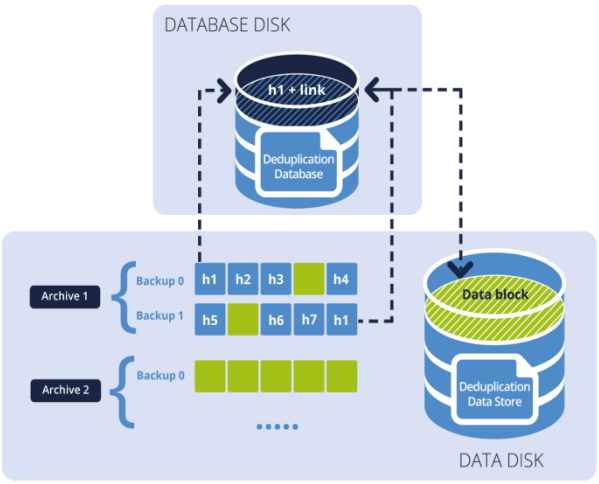

Figure 1

II. RELATED WORK

L. Wang,et.al,…[1] Proposed an progressive public cloud utilization model for small-to medium scale scientific businesses to make use of elastic sources on a public cloud internet site online while keeping their flexible gadget controls, i.e., create, activate, hunch, resume, deactivate, and destroy their excessive-diploma control entities—provider control layers without understanding the information of control. Second, we layout and implement an progressive device—Dawning Cloud, at the middle of which might be mild-weight carrier manipulate layers strolling on top of a commonplace control provider framework. The commonplace manage provider framework of Dawning Cloud no longer fine allows constructing light-weight issuer manage layers for heterogeneous workloads, but additionally makes their control responsibilities easy. Third, we examine the systems comprehensively the usage of both emulation and real experiments.

B. Li,et.al,…[2] Took a step closer to bringing the various blessings of the Map Reduce model to incremental one-pass analytics. In the brand new model, the Map Reduce system reads input information simplest as soon as, plays incremental processing as extra facts is read, and makes use of system sources effectively to attain excessive overall performance and scalability. The goal is to layout a platform to help such scalable, incremental one-bypass analytics. This platform may be used to support interactive facts analysis, which may also contain online aggregation with early approximate solutions, and, inside the future, movement query processing, which offers close to real-time insights as new data arrives. We argue that, so that it will support incremental one-skip analytics, a MapReduce system must keep away from any blocking off operations and additionally computational and I/O bottlenecks that prevent statistics from "easily" flowing thru map and reduce phases at the processing pipeline.

R. Kienzler,et.al,…[3] Propose an incremental facts get right of entry to and processing technique for statistics-intensive cloud programs that can disguise statistics switch latencies while preserving linear scalability. Similar in spirit to pipelined query evaluation in conventional database systems, information is accessed and processed in small increments, thereby propagating statistics chunks from one level of the statistics analysis venture to another as quickly as they're available rather than waiting till the complete dataset turns into available. This manner we are able to process data normally in memory (for this reason, lessen time-consuming I/O to nearby disk and cloud garage, and keep away from storage charges) as well as reaching pipelined parallelism (further to the existing partitioned parallelism), main to a reduction in average mission final touch time.

C. Olston,et.al,…[4] Proposed a device for Building and updating a search index from a movement of crawled net pages. Some of the numerous steps are deduplication, link analysis for unsolicited mail and exceptional classification, joining with click on-based totally recognition measurements, and file inversion. Processing semi-structured records feeds, e.g. Information and (micro-)blogs. Steps consist of de-duplication, geographic region decision, and named entity reputation. Processing alongside these strains is an increasing number of completed on a new technology of bendy and scalable facts management systems, inclusive of Pig/Hadoop.

Hadoop is a scalable, fault-tolerant machine for strolling character map-reduce processing operations over unstructured information documents. Pig adds higher-stage, based abstractions for statistics and processing. Despite the success of Pig/Hadoop, it's far becoming apparent that a new, higher, layer is wanted: a workow supervisor that offers with a graph of interconnected Pig Latin applications, with statistics handed among them in a non-stop style. Given that Pig itself offers with graphs of interconnected data processing steps, it's far natural to ask why one would layer any other graph abstraction on pinnacle of Pig.

K.H. Lee, et.al,…[5] Carried out The programming version is stimulated by the map and reduces primitives found in Lisp and other practical languages. Before developing the MapReduce framework, Google used masses of separate implementations to procedure and compute big datasets. Most of the computations were rather easy, but the input information turned into regularly very huge. Hence the computations needed to be distributed throughout loads of computers so as to complete calculations in an affordable time. MapReduce is noticeably efficient and scalable, and consequently may be used to system large datasets. When the MapReduce framework became delivered, Google absolutely rewrote its net search indexing gadget to apply the new programming model. The

indexing gadget produces the records structures used by Google web search.

The parallelization doesn't necessarily have to be completed over many machines in a network. There are extraordinary implementations of MapReduce for parallelizing computing in specific environments. Phoenix is an implementation of MapReduce, which is aimed toward shared-reminiscence, multi-middle and multiprocessor systems, i.e. Unmarried computer systems with many processor cores.

III. DATA DUPLICATION TYPES

A. File-level de-duplication

It is usually called single-instance garage, file-level information de-duplication compares a file that has to be archived or backup that has already been saved by way of checking all its attributes against the index. The index is updated and saved only if the file is particular, if not than best a pointer to the existing file this is stored references. Only the unmarried example of record is saved within the end result and applicable copies are changed through "stub" which points to the unique record.

B. Block-level de-duplication

Block-stage statistics deduplication operates on the basis of sub-document level. As the call implies that the document is being broken into segments blocks or chunks in order to be tested for previously stored data vs. redundancy. The popular technique to decide redundant facts is by assigning identifier to chew of statistics, by the use of hash algorithm as an example it generates a unique ID to that unique block. The precise specific Id can be in comparison with the critical index. In case the ID is already present, then it represents that before best the facts is processed and stored earlier than .Due to this only a pointer reference is stored inside the vicinity of previously stored data. If the ID is new and does no longer exist, then that block is particular. After storing the unique chew the unique ID is updated into the Index. There is change in size of chew as consistent with the seller. Some will have fixed block sizes, at the same time as a few others use variable block sizes.

C. Variable block level de-duplication

It compares varying sizes of data blocks that can reduce the chances of collision, stated Data links Orlandini. The difference between deduplication schemes are shown in figure 2.

Deduplication schemes

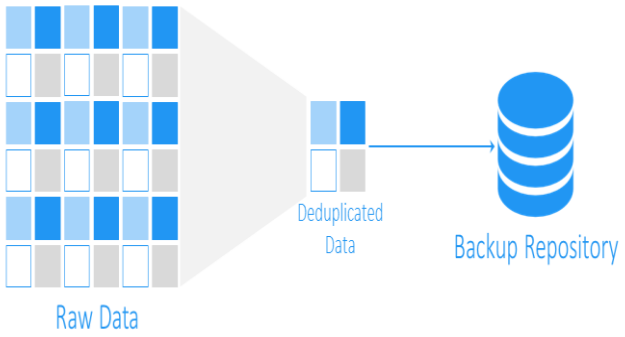

Figure 2

D. Traditional Encryption algorithm

Although it is recognized that statistics deduplication gives extra blessings, protection and confinement concerns get up due to the fact the customers touchy records is vulnerable to both the outsider in addition to insider assaults. So, whilst thinking about the traditional encryption techniques to secure the customers touchy statistics there are many issues are related. Traditional encryption offers records confidentiality but it is not well suited with deduplication. As in conventional encryption exclusive users encrypt their data with their own keys. Thus, the same data of the extraordinary customers will lead to distinct cipher text that is making the information deduplication nearly impossible on this traditional method

E. Convergent Encryption algorithm:

The convergent encryption techniques are those, which provide the data confidentiality to the users

outsourced data stored on the public clouds. These techniques while providing the confidentiality to the data are also compatible with the data deduplication process. In this algorithm, the encryption key is itself derived from the message. So it supports data deduplication also, because the same file will give the same encryption key so it will generate the same cipher text irrespective of users which makes data deduplication possible.

KeyGenCE(M) → K is the key generation algorithm that maps a data copy M to a convergent key K;

EncCE(K,M) → C is the symmetric encryption algorithm that takes both the convergent key K and the data copy M as inputs and then outputs a ciphertext C;

DecCE(K,C) → M is the decryption algorithm that takes both the ciphertext C and the convergent key K as inputs and then outputs the original data copy M; and

TagGen (M) → T (M) is the tag generation algorithm that maps the original data copy M and outputs a tag T (M).

F. Block cipher algorithm

In cryptography, a block cipher is a deterministic set of hints walking on everyday-length businesses of bits, referred to as blocks, with an unvarying transformation that is particular thru way of a symmetric key. Block ciphers perform as vital smooth additives in the format of many cryptographic protocols, and are drastically used to location into effect encryption of bulk records. Iterated product ciphers carry out encryption in a couple of rounds, every of which uses a considered one in each of a kind sub key derived from the real key. One huge implementation of such ciphers is drastically completed within the DES cipher. Many particular realizations of block ciphers, collectively with the AES, are labelled as substitution-permutation networks. The guide of the DES cipher modified into critical in the public records of cutting-edge block cipher format. It moreover brought on the instructional improvement of cryptanalytic attacks. Both differential and linear cryptanalysis arose out of research on the DES layout. There is a palette of assault techniques inside the direction of which a block cipher want to be secure, in addition to being strong closer to brute strain assaults. Even a at ease block cipher is suitable great for the encryption of a unmarried block under a tough and fast key. A multitude of modes of operation was designed to allow their repeated use in a secure manner, commonly to gain the protection dreams of confidentiality and authenticity. However, block ciphers can also function as constructing-blocks specifically cryptographic protocols, which incorporate time-commemorated hash skills and pseudo-random variety turbines.

One vital sort of iterated block cipher known as a substitution-permutation network (SPN) takes a block of the plaintext and the crucial detail as inputs, and applies numerous alternating rounds, which incorporate a substitution degree located thru a permutation degree—to offer every block of cipher textual content output. The non-linear substitution degree mixes the vital thing bits with those of the plaintext, growing Shannon's confusion. The linear permutation diploma then dissipates redundancies, developing diffusion. A substitution field (S-situation) substitutes a small block of enter bits with every other block of output bits. This substitution wants to be one-to-one, to make sure inevitability (because of this decryption). A secure S-field may additionally moreover furthermore have the assets that converting one enter bit will trade about half of the output bits on commonplace, displaying what is known as the avalanche impact—i.e. It has the assets that every output bit will depend upon every enter bit.

G. Variable chunk similarity

It calls for greater processing energy than the file deduplication, seeing that the amount of identifiers that want to be processed increases considerably.

Correspondingly, its index for tracking the person iterations gets additionally a whole lot massive. Using of variable period blocks is even more source-big. Moreover, occasionally the equal hash variety may be generated for 2 exquisite records fragments that are referred to as hash collisions. If that occurs, the system will now not keep the contemporary statistics as it sees that the hash wide variety already exists inside the index. The algorithm steps as follows

BlockTag(FileBlock) - It computes hash of the File block as file block Tag;
DupCheckReq(Token) - It requests the Storage Server for Duplicate Check of the file block.
FileUploadReq(FileBlockID, FileBlock, Token) – It uploads the File Data to the Storage Server if the file block is Unique and updates the file block Token stored.
FileBlockEncrypt(Fileblock) - It encrypts the file block with Convergent Encryption, where the convergent key is from SHA Hashing of the file block;
TokenGen(File Block, UserID) – the process loads the associated privilege keys of the user and generate token.
FileBlockStore(FileBlockID, FileBlock, Token) - It stores the FileBlock on Disk and updates the Mapping. The variable chunk similarity level deduplication is shown in figure 3.

Variable chunk similarity backup server

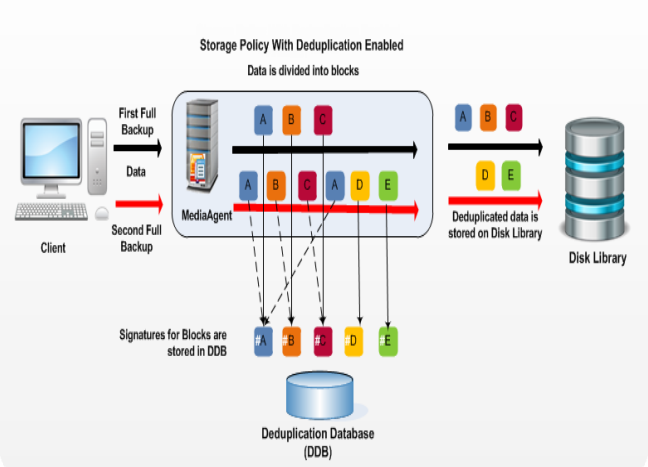

Figure 3

The mathematical model as follows

Let S be the system object.
It consist of following S={U,F,CSP}
U= no of users U={u1,u2,u3,…..un}
F= no of files F={f1,f2,f3,…..fn}
B=no of blocks. B{B1,B2,…,Bn}
CSP={C,PF,V,POW}
C=challenge
PF =proof by CSP
V= verification by TPA
POW= proof of ownership
CSP= Cloud Service provider
CSP={PF,F} PF=proof F=files

The proposed architecture is shown in figure 4.

Proposed framework

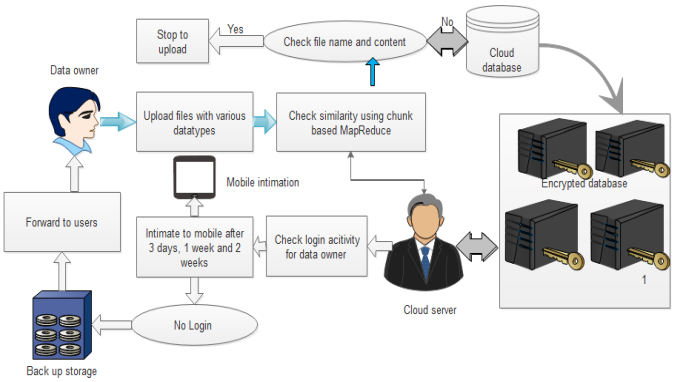

Figure 4

The block cipher algorithm steps as follows

def pf3(block,key)

d0 = ZZ(key.index(block[0])).digits(3,padto=3)
d1 = ZZ(key.index(block[1])).digits(3,padto=3)
d2 = ZZ(key.index(block[2])).digits(3,padto=3)
f0 = [d1[1],d2[1],d0[1]]
f1 = [d1[2],d2[2],d0[2]]
f2 = [d1[0],d2[0],d0[0]]

return key[ZZ(f0,3)]+key[ZZ(f1,3)]+key[ZZ(f2,3)]

IV. CONCLUSION

In this paper analyzed disbursed deduplication systems to enhance the reliability of records whilst attaining the confidentiality of the customers and

additionally shared authority outsourced records with an encryption mechanism. The structures have been proposed to help report-degree and block-stage data deduplication. The security of tag consistency and integrity has been finished. We carried out our deduplication systems the usage of the block cipher scheme with variable bite similarity and tested that it incurs small encoding/decoding overhead as compared to the network transmission overhead in ordinary add/down load operations. Data anonymity is accomplished for the reason that wrapped values are exchanged throughout transmission. User confinement is superior with the aid of get entry to requests to privately tell the cloud server about the customer's access dreams.

V. REFERENCES

[1]. Prof. L. Wang, J. Zhan, W. Shi and Y. Liang, "In cloud, can scientific communities benefit from the economies of scale?" IEEE Transactions on Parallel and Distributed Systems 23(2): 296-303, 2012.

[2]. B. Li, E. Mazur, Y. Diao, A. McGregor and P. Shenoy, "A platform for scalable one-pass analytics using mapreduce," in: Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'11), 2011, pp. 985-996.

[3]. R. Kienzler, R. Bruggmann, A. Ranganathan and N. Tatbul, "Stream as you go: The case for incremental data access and processing in the cloud," IEEE ICDE International Workshop on Data Management in the Cloud (DMC'12), 2012

[4]. C. Olston, G. Chiou, L. Chitnis, F. Liu, Y. Han, M. Larsson, A. Neumann, V.B.N. Rao, V. Sankarasubramanian, S. Seth, C. Tian, T. ZiCornell and X. Wang, "Nova: Continuous pig/hadoop workflows," Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'11), pp. 1081-1090, 2011.

[5]. K.H. Lee, Y.J. Lee, H. Choi, Y.D. Chung and B. Moon, "Parallel data processing with mapreduce: A survey," ACM SIGMOD Record 40(4): 11-20, 2012.

[6]. X. Zhang, C. Liu, S. Nepal and J. Chen, "An Efficient Quasiidentifier Index based Approach for Privacy Preservation over Incremental Data Sets on Cloud," Journal of Computer and System Sciences (JCSS), 79(5): 542-555, 2013.

[7]. X. Zhang, T. Yang, C. Liu and J. Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization using Systems, in MapReduce on Cloud," IEEE Transactions on Parallel and Distributed, 25(2): 363-373, 2014.

[8]. N. Laptev, K. Zeng and C. Zaniolo, "Very fast estimation for result and accuracy of big data analytics: The EARL system," Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE), pp. 1296-1299, 2013.

[9]. T. Condie, P. Mineiro, N. Polyzotis and M. Weimer, "Machine learning on Big Data," Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE), pp. 1242-1244, 2013.

[10]. Aboulnaga and S. Babu, "Workload management for Big Data analytics," Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE), pp. 1249, 2013