A BIRD EYE VIEW ON BIG DATA DEDUPLICATION

Md Shah Jalal Uddin, Dwaipayan Das
Netaji Subhash Engineering College.

Abstract : big data primarily based Cloud computing offers a replacement method of service provision by re-arranging varied resources over the web. The foremost necessary and standard cloud service is information storage. So as to preserve the privacy of information holders, information are typically hold on in cloud in an encrypted kind. However, encrypted information introduce new challenges for cloud information de-duplication that becomes crucial for giant information storage and process in cloud. Ancient de-duplication schemes cannot work on encrypted information. Existing solutions of encrypted information de-duplication suffer from security weakness. They cannot flexibly support information access management [13] and revocation. Therefore, few of them is pronto deployed in observe. During this paper, we propose a theme to de-duplicate encrypted information hold on in cloud supported possession challenge and proxy re-encryption. It integrates cloud information de-duplication with access management.

IndexTerms **- Big Data, Deduplication, NYSE, Social Media and Cloud Model.**

I. Introduction

The Big Data' is additionally an information however with an enormous size. 'Big Data' may be a term accustomed describe assortment of information that's large in size and however growing exponentially with time. In short, such information is thus massive and sophisticated that none of the normal data management tools are able to store it or method it with efficiency.

Big data is a term for data sets that are so large that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, Lately, the term "big data" tends to refer to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of dataset. 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, and combat crime and so on." Scientists, business executives, practitioners of medicine, advertising andgovernments alike regularly meet difficulties with large data-sets in areas including meteorology, genomics, complex physics simulations, biology and environmental research. [1].Data sets grow rapidly - in part because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial software logs, cameras, microphones, 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 Exabyte (2.5×10^{18}) of data are generated. One question for large enterprises is determining who should own big-data initiatives that affect the entire organization. [2]

The term has been in use since the 1990s, with some giving credit to John Mashey for coining or at least making it popular. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big Data philosophy encompasses unstructured, semi-structured and structured data, however the main focus is on unstructured data. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many peta bytes of data. Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale.[3]

In a 2001 analysis report and connected lectures, META cluster (now Gartner) outlined information growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), speed (speed of information in and out), and selection (range of information sorts and sources). Gartner, and currently a lot of the trade, still use this "3Vs" model for describing massive information. In 2012, Gartner updated its definition as follows: "Big information is high-volume, fast and/or high-variety data assets that demand efficient, innovative sorts of science that alter increased insight, higher cognitive process, and method automation." Gartner's definition of the 3Vs continues to be wide used, and in agreement with a accordant definition that states that "Big information represents the data assets characterised by such a High Volume, speed and selection to want specific Technology and Analytical ways for its transformation into Value". To boot, a new V "Veracity" is superimposed by some organizations to explain it, revisionism challenged by some trade authorities. The 3Vs are dilated to different complementary characteristics of huge data:

- Volume: massive information does not sample; it simply observes and tracks what happens
- Velocity: massive information is commonly obtainable in period
- Variety: massive information attracts from text, images, audio, video; and it completes missing items through information fusion
- Machine learning: massive information typically does not raise why and easily detects patterns
- Digital footprint: massive information is commonly a cost-free by product of digital interaction.

Database management systems and desktop statistics- and visualization-packages typically have problem handling huge information. The work could need "massively parallel software package running on tens, hundreds, or maybe thousands of servers". What counts as "big data" varies looking on the capabilities of the users and their tools, and increasing capabilities build huge information a moving target. "For some organizations, facing many gigabytes bandwidth for the primary time could trigger a necessity to rethink data management choices. For others, it's going to take tens or many terabytes before information size becomes a big thought."Example of big data.[4] [5]Following are some the examples of 'Big Data'-

Fig. 1. Shows the N. Y. Stock Exchange supply of massive information [6]

The newyork exchange generates regarding one computer memory unit of recent trade information per day

Fig. 2- Second major supply of massive information – Social Media [6]

Social Media Impact Statistic shows that 500+terabytes of recent information gets eaten into the databases of social media website Face book, every day. This information is especially generated in terms of pic and video uploads, message exchanges, putt comments etc.

Single reaction engine will generate 10+terabytes of information in half-hour of a flight time. With several thousand flights per day, generation of information reaches up to several Pet bytes.

Fig. 3. Different ways of Big Data [6]

II. ARCHITECTURE OF BIG DATA AND CLOUD MODEL

Big information repositories have existed in several forms, typically designed by firms with a special want. Business vendors traditionally offered parallel direction systems for giant information starting within the Nineteen Nineties. For several years, water crop revealed a largest information report. Teradata Corporation in 1984 marketed the data processing DBC 1012 system. Teradata systems were the primary to store and analyse one TB of information in within the era of ninety's. . Disk drives were 2.5GB in 1991 that the definition of massive information incessantly evolves in step with Kryder's Law. Teradata put in the primary computer memory unit category RDBMS primarily based system. As of current era, there are a number of dozen computer memory unit category Teradata relative databases put in, the most important of that exceeds fifty lead. Systems up till 2008 were 100 percent structured relative information. Since then, Teradata has added unstructured information varieties together with XML, JSON, and Avro.

LexisNexis cluster developed a C++-based distributed file-sharing framework for information storage and question. The system stores and distributes structured, semi-structured, and unstructured information across multiple servers. Users will build queries in a very C++ non-standard speech known as ECL. ECL uses an "apply schema on read" methodology to infer the structure of keep information once it's queried, rather than once its keep. In 2004, LexisNexis non-inheritable sensing opposition and non-inheritable selection purpose, Inc. and their high-speed data processing platform. The 2 platforms were integrated into HPCC (or superior Computing Cluster) Systems. HPCC was open-sourced beneath the Apache v2.0 License. Quantcast classification system was on the market concerning identical time. [7] [8]

Google revealed a paper on a method known as Map cut back that uses an identical design. The Map cut back thought provides a data processing model, and an associated implementation was discharged to method Brobdingnagian amounts of information. With Map cut back, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the cut back step). The framework was terribly triple-crown, therefore others wished to duplicate the formula. Therefore, an implementation of the Map cut back framework was adopted by an Apache ASCII text file project named Hadoop.

MIKE2.0 is associate open approach to info management that acinformations the requirement for revisions as a result of huge information implications known in a writing titled "Big information answer Offering" The methodology addresses handling huge information in terms of helpful permutations of information sources, quality in interrelationships, and issue in deleting (or modifying) individual records. Studies showed that a multiple-layer design is one choice to address the problems that huge information presents. A distributed parallel design distributes information across multiple servers; these parallel execution environments will dramatically improve processing speeds. This sort of design inserts information into a parallel software package, which implements the employment of Map-Reduce and Hadoop frameworks. This sort of framework appearance to form the process power clear to the top user by employing a front-end application server.

III. BIG DATA WITH CLOUD

Cloud public clouds, services (i.e. applications and storage) are on the market for general use over the net. A non-public cloud could be a virtualized information centre that operates among a firewall. During this analysis introduce mixture of public and personal cloud, hybrid cloud. Cloud computing provides computation and storage resources on the net. Increasing quantity of information is being hold on within the cloud and it's shared by users with such privileges that defines special rights to access hold on information. Managing the exponential growth of ever-increasing volume of information has become an important challenge. Consistent with IDC cloud report 2014, corporations in Asian nation are creating a gradual move from on premise bequest to totally different sorts of cloud. Whereas the method is gradual, it's started by migrating bound application workloads to cloud. To create scalable management of hold on information in cloud computing, de-duplication has been renowned technique that becomes additional standard recently. De-duplication could be a specialised information compression technique that scale back space for storing and transfer information measure in cloud storage. In de-duplication, only 1 distinctive instance of the information is truly on the server and redundant information is replaced with a pointer to the distinctive data copy. Deduplication will occur either at file level or block level. From the user perspective, security and privacy issues are arise as information are prone to each corporate executive and outsider attack. We should properly enforce confidentiality, integrity checking, and access management mechanisms each attacks. De-duplication doesn't work with ancient secret writing. User encrypts their files with their individual secret writing key, totally different cipher text would emerge even for identical files. Thus, ancient secret writing is incompatible with information Delaware duplication. Confluent secret writing could be a wide used technique to mix the storage saving of de-duplication to enforce confidentiality. In confluent secret writing, the information copy is encrypted below a key derived by hashing the information itself. This confluent secret's used for write in code and decode a information copy. Once key generation and encoding, users retain the keys and send the cipher text to the cloud. Since secret writing is settled, identical information copies can generate an equivalent confluent key and also the same cipher text. This permits the cloud to perform Delaware duplication on the cipher texts. The cipher texts will solely be decrypted by the corresponding information homeowners with their confluent keys. [9] [10]

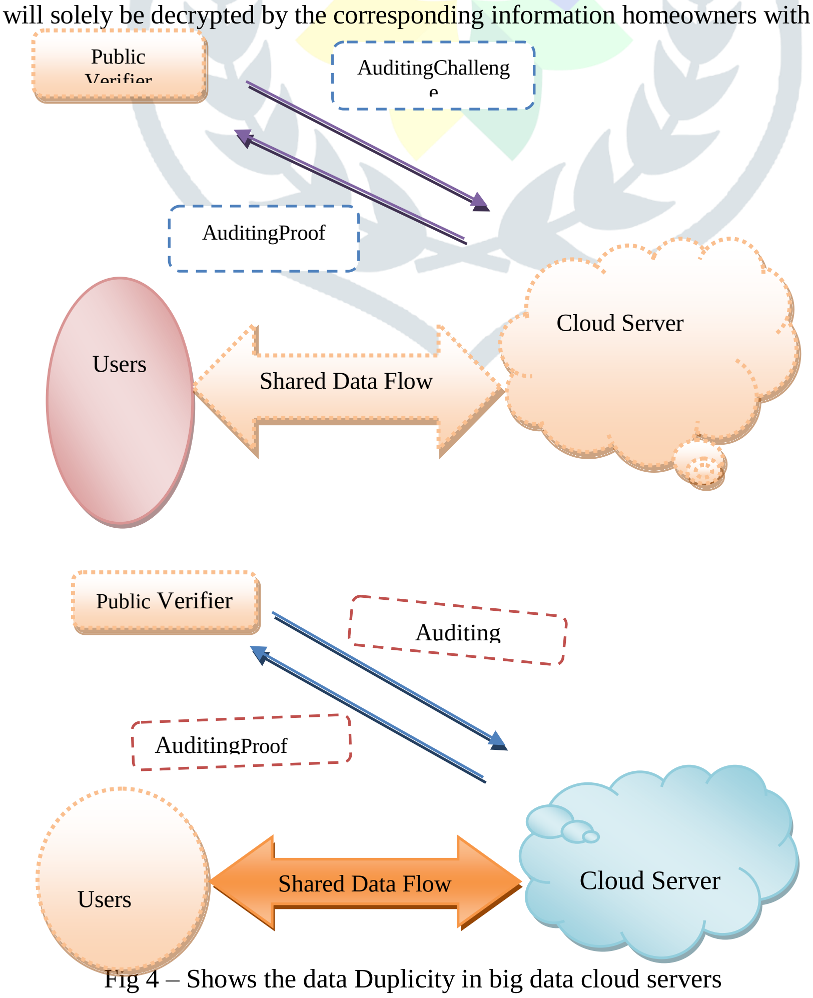

Fig 4 – Shows the data Duplicity in big data cloud servers

Data deduplication could be a technique to cut back cupboard space. By characteristic redundant information exploitation hash values to match information chunks, storing only 1 copy, and making logical tips to different copies rather than storing different actual copies of the redundant information Deduplication scale backs information volume thus disc space and network information measure is reduced that reduce prices and energy consumption for running storage systems [11] [12]

Duplicity could be a piece of code which will perform encrypted backups to remote storage over the network. It uses the various algorithmic program to implement progressive backups, therefore minimising the number of information that must be transferred over the network and keep remotely. The wildebeest Privacy Guard is employed to produce sturdy coding, creating it safe to stay your backups in one in every of the numerous public cloud storage solutions.

IV. DEDUPLICATION IN CLOUD STORAGE

Data deduplication may be applied at nearly each purpose wherever information is hold on or transmitted in cloud storage. Several cloud suppliers supply disaster recovery and deduplication may be accustomed create disaster recovery more practical by replicating information when deduplication for rushing up replication time and information measure value savings. Backup and deposit storage in clouds may also apply information deduplication so as to cut back physical capability and network traffic [13], [14]. Moreover, in live migration method, we want to transfer an outsized volume of duplicated image information [15]. There are three major performance metrics of migration to consider: total information transferred, total migration time and repair period of time. Longer migration time and period of time would be cause service failure. Thus, deduplication will assist in migration. Deduplication may be accustomed cut back storage of active information like virtual machine pictures. Factors to contemplate once victimization deduplication in primary storage is the way to balance the trade-offs between space for storing saving and performance impact. in addition, Mandagere, state that deduplication algorithms mirror the performance of de-duplicated storage in terms of fold issue, reconstruction information measure, information overhead, and resource usage.

One of the foremost common varieties of information deduplication implementations works by scrutiny chunks of information to discover duplicates. For that to happen, every chunk of information is appointed an identification, calculated by the software system, generally exploitation science hash functions. In several implementations, the idea is created that if the identification is identical, the info is identical, although this can't be true all told cases because of the pigeonhole principle; alternative implementations don't assume that 2 blocks of information with identical symbol are identical, however truly verify that information with identical identification is identical. If the software system either assumes that a given identification already exists within the deduplication namespace or truly verifies the identity of the 2 blocks of information, betting on the implementation, then it'll replace that duplicate chunk with a link. [16]

Once the information has been duplicated, upon scan back of the file, where a link is found, the system merely replaces that link with the documented information chunk. The deduplication method is meant to be transparent to finish users and applications. [17]

Commercial deduplication implementations disagree by their unitisation strategies and architectures.

Chunking. In some systems, chunks are outlined by physical layer constraints (e.g. 4KB block size in WAFL). In some systems solely complete files are compared, that is named single-instance storage or SIS. The foremost intelligent (but computer hardware intensive) methodology to unitisation is usually thought of to be sliding-block. In slippery block, a window is passed on the file stream to hunt out additional present internal file boundaries. [18]

Client backup deduplication. This can be the method wherever the deduplication hash calculations are created on the supply (client) machines. Files that have identical hashes to files already within the target device don't seem to be sent, the target device simply creates applicable internal links to reference the duplicated information. The advantage of this can be that it avoids information being unnecessarily sent across the network thereby reducing traffic load. [19]

Primary storage and storage device. By definition, primary storage systems are designed for best performance, instead of lowest doable value. The look criteria for these systems is to extend performance, at the expense of alternative issues. Moreover, primary storage systems are abundant less tolerant of any operation which will negatively impact performance. Conjointly by definition, storage device systems contain primarily duplicate, or secondary copies of information. These copies of information are generally not used for actual production operations and as a result are additional tolerant of some performance degradation, in exchange for enhanced potency. [20]

To date, information de-duplication has preponderantly been used with external storage systems. The explanations for this are two-fold. First, information de-duplication needs overhead to find and take away the duplicate information. In primary storage systems, this overhead could impact performance. The second reason why de-duplication is applied to secondary information, is that secondary information tends to own a lot of duplicate information. Backup applications especially usually generate vital parts of duplicate information over time. Data de-duplication has been deployed with success with primary storage in some cases wherever the system style doesn't need vital overhead, or impact performance. [21] [22][23]

V. LITERATURE SURVEY

"**Cho, Ei Mon** et.al ., [24] "Big knowledge Cloud Deduplication supported Verifiable Hash oblique cluster Signcryption" during this paper, we've designed a theme that supports secure deduplication wherever many teams are sharing knowledge by exploitation VHCGS. this can be an endeavor to undertake out cross-group user deduplication in an exceedingly real huge knowledge management. In doing therefore, we are taking the utility of existing schemes instead of proposing a wholly new one. we introduce a framework for a gaggle signcryption theme which might shield against duplication for the cloud suppliers and defend against unpredictable knowledge attacks. VHCGS fits the initial framework of settled hash oblique secret writing whereas satisfying a clusterfeature by adding the cloud server verifiable group signcrypto. VHCGS consists of three protocols: a setup protocol, an transfer protocol, and a transfer protocol. VHCGS ensures each message security and tag consistency likewise because the information measure potency of the cluster user and cloud storage server. VHCGS supports the extended demands that arise in realistic and secure situations. In future work, we arrange to investigate the support for a complete cross-group deduplication system for multiple access teams for giant knowledge cloud computing and extend our style to a full deduplication system.

Fu, Yinjin, et.al ., [25] "Application-Aware massive information Deduplication in Cloud Environment" during this paper, we describe AppDedupe, an application-aware ascendable inline distributed deduplication frame-work for large information management, that achieves a trade-off between ascendable performance and distributed deduplication effectiveness by exploiting application awareness, information similarity and neighborhood. It adopts a two-tiered information routing theme to route information at the super-chunk granularity to scale back cross-node information redundancy with sensible load balance and low communication overhead, and employs application-aware similarity index primarily based optimization to boost deduplication potency in every node with terribly low RAM usage. Our real-world trace-driven analysis clearly demonstrates AppDedupe's vital advantages over the progressive distributed deduplication schemes for big clusters within the following necessary 2 ways that.

Yang, Xue, et.al., [26] "Achieving economical and Privacy-Preserving Cross-Domain massive knowledge Deduplication in Cloud. Cloud storage adoption, notably by organizations, is probably going to stay a trend within the predictable future. This is, unsurprising , because of the conversion of our society. One associated analysis challenge is the way to effectively scale back cloud storage prices because of knowledge duplication. during this paper, we projected an economical and privacy-preserving massive knowledge deduplication in cloud storage for a three-tier cross domain design. we then analyzed the safety of our projected theme and incontestable that it achieves improved privacy protective, answerableness and knowledge handiness, whereas resisting brute-force attacks. we conjointly incontestable that the projected theme outperforms existing progressive schemes, in terms of computation, communication and storage overheads. additionally, the time complexness of duplicate search in our theme is an economical index time. Future analysis includes extending the projected theme to totally shield the duplicate data from revealing, even by a malicious CSP, while not touching the potential to perform knowledge deduplication. Future analysis agenda also will embody extending the theme to be resilient against a wider vary of security threat by external attackers, similarly as rising the time complexness of duplicate search.

Karthika et.al .,[27] "Perlustration on techno level classification of deduplication techniques in cloud for giant knowledge storage" a brand new trend has set in storage wherever knowledge de-duplication plays, a completely unique role in compression and alternative knowledge reduction in sophisticated as a bottom component of the answer. knowledge deduplication technique improves knowledge protection, that will increase the speed of service, and reduces the prices and therefore the use of information measure. De-duplication is integrated with cloud space for storing, this helps within the simple maintenance of knowledge and eradicating the replica's within the cloud server. Cloud computing, storage resources is with efficiency used this permits each organization to create their own non-public cloud and hybrid cloud in step with their functions and desires. to satisfy the rising performance would like, improvement is developed to satisfy the growing would like for SSSD. The key downside with solid state storage (SSSD) is that it can't meet volume wants, wherever a cluster of knowledge reduction tools comes into play. There are many de-duplication techniques are used deduplication tools are on the market wherever there are several execs and cons. In our future work are going to be concentrating on deduplication of compressed knowledge wherever several benchmark providers are endeavor for his or her best.

Kumar, Naresh et.al., [28] "Bucket based mostly information deduplication technique for information storage system" In big information storage information is just too large and expeditiously store information is tough task. to unravel this downside Hadoop tool provides HDFS that manages information by maintain duplication of information however this inflated duplication. To expeditiously stores information and de-duplication the info this paper presents a bucket based mostly technique. In planned technique totally different buckets are wont to store information and once same information is accessed by map cut back i.e. already keep in bucket then that information are going to be discarded therefore this method positively will increase potency of big data storage. Results shows that in planned mechanism deduplication quantitative relation is high, information size reduction is high hash time and chunk time is low as compare to existing fastened size unitization technique. In future continue acting on it and refine results with low computation time conjointly we propose new mechanism within which all modules are combined like unitization, deduplication and hashing which will notice additional duplicate content and take away them in correct manner with less time period.

Xia, Wen et.al., [29] "DARE: A deduplication-aware likeness detection and elimination theme for knowledge reduction with low overheads" during this paper, we gift DARE, a deduplication-aware, low-overhead likeness detection and elimination theme for knowledge reduction in backup/archiving storage systems. DARE uses a completely unique approach, DupAdj, that exploits the duplicate-adjacency info for economical likeness detection in existing deduplication systems, and employs an improved super-feature approach to any detective work likeness once the duplicate adjacency info is lacking or restricted. Results from experiments driven by real-world and artificial backup data recommend that DARE will be a robust and economical tool for maximizing knowledge reduction by any detective work resembling data with low overheads. Specifically, DARE solely consumes concerning ¼ and 1/2 severally of the computation and classification overheads needed by the standard super-feature approaches whereas detective work 2-10% a lot of redundancy and achieving a better turnout. what is more, the DAR Enhanced knowledge reduction approach is shown to be capable of rising the data-restore performance, dashing up the deduplication-only approach by an element of 2(2X) by using delta compression to any eliminate redundancy and effectively enlarge the logical house of the restoration cache.

Rashid, Fatema et.al ., [30] "Proof of storage for video deduplication within the cloud "We have projected proof of retrieval and proof of possession protocols for secure deduplication of video assumptive that the CSP is semi-honest so as to handle the large knowledge drawback within the cloud storages. Through experiments victimization 6 completely different video sequences, we've determined what proportion volume of extra knowledge needs to be keep within the cloud storage as a results of the information deduplication method, so as to make sure the protection of the video file victimization the POR protocol. we've got conjointly shown that the POW protocol is economical in terms of process overhead since it doesn't need any calculation of the information one by one, however uses the information already generated for the POR theme. As future work, the projected POR and POW protocols may be changed to incorporate the support for dynamic knowledge and third party auditing within the context of information deduplication involving a semi-honest CSP.

Yujuan Tan et. al.[31] In this analysis article propose a relation based mostly deduplication performance booster for each cloud backup and restore operations known as CABdedupe. In which, they establish that file and knowledge chunk are modified or stay unchanged, so facultative the removal of the unadapted knowledge from transmission for backup/restore operation to enhance performance. They illustrate the role of CABdedupe in abackup system by mistreatment backup consumer and backup server to represent the functionalities of original consumer and server in module in existing backup system.

Kan Yang et. al. [32] In this analysis work, they 1st style an auditing framework for cloud storage system and projected an economical and privacy protective auditing protocol. Then they extend the auditing protocol to support the information dynamic operations. Then they additional extend the auditing protocol to support batch auditing for each multiple homeowners and multiple clouds. The projected secure dynamic auditing protocol solve the information privacy issues by generating an encrypted proof with the challenge stamp by victimisation the metallic element dimensionality property of the linear pairing, specified the auditor willnot rewrite it however can verify the correctness of the proof. The auditing theme during this paper incurs less computation value and communication value of the auditor by moving the computing numerous the auditing to the server, that greatly improves the auditing performance and may be applied to massive scale cloud storage system of these things.

Yang Zhang et. al. [33]In this analysis work They strips the computer file stream onto multiple storage nodes, so limits range of hold on information segments on every node and make sure the fingerprint index may be fitted into memory. The distributed deduplication storage system during this paper aims for top output and measurability. It encompass 3 parts one Meta server, multiple procedure servers and multiple storage nodes. It accomplish most deduplication output by storing whole fingerprint index within RAM, utterly avoid access once doing fingerprint search. They used Cuckoo hash to index the fingerprint index.

N. Mandagereetet al. [34]In this analysis work This paper aims to develop a taxanomy to characterize and classify the increasing variety of accessible deduplication technologies, and experimental analysis of fold issue, resource necessities, reconstruction time of deduplication algorithmic rule on information|a knowledge|an information} from planet backup servers employed by enterprise users to backip official email and different business data. during this paper, they apply completely different deduplication algorithmic rule i.e. VBH, FBH, WFH to entire dataset to know distinction between them. the higher fold issue usually return at the expense of enlarged resource overhead, however converse isn't continuously true.

VI.CONCLUSION

In the survey paper discuss regarding large data and secure data de-duplication in big knowledge. Conjointly discuss the matter of massive knowledge primarily based de-duplication in cloud. There are completely different ways are used for cryptography of massive knowledge, like AES, RSA and code methodology within the projected methodology in future implement HECC for improvement secure and time in huge knowledge. In similar manner privacy conserving is additionally a main concern in huge knowledge, for security improvement used third party evidence (TPA) services construct in any implementation. In future work implementation SHA-2 formula for Hash worth calculation. Hash vales provide higher result as compare to SHA-1 formula. There are major considerations of this pre thesis report. In future try to implement data de-duplication on image processing. Image data de-duplication will be preserved with the help of impulse noise. [35] [36]. Use TPA concept for next generation communication system for secure data transfer in next generation communication system. [37]

REFERENCES

1. T. Y. Wu, J. S. Pan, and C. F. Lin, “Improving accessing efficiency of cloud storage using de-duplication and feedback schemes,” IEEE System Journals, vol. 8, no. 1, pp. 208–218, Mar. 2014, doi:10.1109/JSYST.2013.2256715.
2. C. Fan, S. Y. Huang, and W. C. Hsu, “Hybrid data deduplication in cloud environment,” in Processing International Conference Inf. Security Intelligent. Control 2012, pp. 174–177, doi:10.1109/ISIC.2012.6449734.
3. J. W. Yuan and S. C. Yu, “Secure and constant cost public cloud storage auditing with deduplication,” in Proc. IEEE International Conference Communication Network Security 2013, pp. 145–153, doi: 10.1109/ CNS.2013. 6682702.
4. N. Kaaniche and M. Laurent, “A secure client side deduplication scheme in cloud storage environments,” in Proc. 6th Int. Conference .New Technol. Mobility Security, 2014, pp. 1–7, doi: 10.1109/NTMS.2014.6814002.
5. Z. Yan, W. X. Ding, and H. Q. Zhu, “A scheme to manage encrypted data storage with deduplication in cloud,” in Proc.ICA3PP2015,Theoretical Computer Science and General Issue, SpringerNov. 2015, pp. 547–561
6. https://cdn.guru99.com/images/Big_Data/061114_0759_WhatIsBigDa3.jpg
7. Z. C. Wen, J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li, “A verifiable data deduplication scheme in cloud computing,” in Proc. International Conference Intelligent Network Collaborative Syst., 2014, pp. 85–90,doi: 10.1109/INCoS.2014.111.
8. J. Li, Y. K. Li, X. F. Chen, P. P. C. Lee, and W. J. Lou, “A hybrid cloud approach for secure authorized deduplication,” IEEE Trans. Parallel Distribution System vol. 26, no. 5, pp. 1206–1216, May 2015, doi:10.1109/ TPDS. 2014. 2318320.
9. T. Y. Wu, J. S. Pan, and C. F. Lin, “Improving accessing efficiency of cloud storage using de-duplication and feedback schemes,” IEEE System Journals, vol. 8, no. 1, pp. 208–218, Mar. 2014, doi:10.1109/JSYST.2013.2256715.
10. C. Fan, S. Y. Huang, and W. C. Hsu, “Hybrid data deduplication in cloud environment,” in Processing International Conference Inf. Security Intelligent. Control 2012, pp. 174–177, doi:10.1109/ISIC.2012.6449734.
11. T. Y. Wu, J. S. Pan, and C. F. Lin, “Improving accessing efficiency of cloud storage using de-duplication and feedback schemes,” IEEE System Journals, vol. 8, no. 1, pp. 208–218, Mar. 2014, doi:10.1109/JSYST.2013.2256715.

12. C. Fan, S. Y. Huang, and W. C. Hsu, “Hybrid data deduplication in cloud environment,” in Processing International Conference Inf. Security Intelligent. Control 2012, pp. 174–177, doi:10.1109/ISIC.2012.6449734.
13. Y.-K. Li, M. Xu, C.-H. Ng, and P. P. C. Lee, “Efficient hybrid inline and out-of-line deduplication for backup storage,” ACM Transaction Storage, vol. 11, no. 1, pp. 2:1-2:21, 2014, doi: 10.1145/2641572.
14. M. Fu, et al., “Accelerating restore and garbage collection in deduplication-based backup systems via exploiting historical information, “in Proc. USENIX Annual Technology Conference, 2014, pp. 181–192.
15. P. Meye, P. Raipin, F. Tronel, and E. Anceaume, “A secure two phase data deduplication scheme,” in Proc. HPCC/CSS/ICESS, 2014, pp. 802–809, doi:10.1109/HPCC.2014.134.
16. M. Kaczmarczyk, M. Barczynski, W. Kilian, and C. Dubnicki, “Reducing impact of data fragmentation caused by inline deduplication,” in Proc. 5th. International System Storage Conference, 2012, pp. 15:1–15:12, doi: 10.1145/ 2367589.2367600.
17. M. Lillibridge, K. Eshghi, and D. Bhagwat, “Improving restore speed for backup systems that use inline chunk-based deduplication,” in Proceeding. USENIX Conf. File Storage Technol., Springer Book Chapter, 2013, pp. 183–198.
18. Z. Yan, X. Y. Li, M. J. Wang, and A. V. Vasilakos, “Flexible data access control based on trust and reputation in cloud computing,” IEEE Trans. Cloud Computing, vol. PP, no. 99, Aug. 2015, doi:10.1109/TCC.2015.2469662,
19. C. Yang, J. Ren, and J. F. Ma, “Provable ownership of file in deduplication cloud storage,” in Proc. IEEE Global Communication Conference 2013, pp. 695–700, doi:10.1109/GLOCOM.2013.6831153.
20. C. Y. Liu, X. J. Liu, and L. Wan, “Policy-based deduplication insecure cloud storage,” in Proceeding Trustworthy Computing Services, 2013,pp. 250–262, doi: 10.1007/978-3-642-35795-4_32.
21. P. Puzio, R. Molva, M. Onen, and S. Loureiro, “Clouded up: Secure deduplication with encrypted data for cloud storage,” in Proc. IEEE International Conference Cloud Computing Technology and Science 2013, pp. 363–370.
22. Z. Sun, J. Shen, and J. M. Yong, “De Du: Building a deduplication storage system over cloud computing,” in Proc. IEEE Int. Conference Computing Supported Cooperative Work Des., 2011, pp. 348–355, doi: 10.1109/CSCWD. 2011.5960097.
23. J. Paulo and J. Pereira, “A survey and classification of storage deduplication systems,” ACM Computing Surveys vol. 47, no. 1, pp. 1–30, 2014, doi: 10.1109/HPCC.2014.134.
24. Cho, Ei Mon, and Takeshi Koshiba. "Big Data Cloud Deduplication based on Verifiable Hash Convergent Group Signcryption."2017 IEEE Third International Conference on Big Data Computing Service and Applications
25. Fu, Yinjin, Nong Xiao, Hong Jiang, Guyu Hu, and Weiwei Chen. "Application-Aware Big Data Deduplication in Cloud Environment." IEEE Transactions on Cloud Computing (2017).
26. Yang, Xue, Rongxing Lu, Kim-Kwang Raymond Choo, Fan Yin, and Xiaohu Tang. "Achieving Efficient and Privacy-Preserving Cross-Domain Big Data Deduplication in Cloud." *IEEE Transactions on Big Data* (2017).
27. Karthika, R. N., C. Valliyammai, and D. Abisha. "Perlustration on techno level classification of deduplication techniques in cloud for big data storage." In *Advanced Computing (ICoAC), 2016 Eighth International Conference on*, pp. 206-211. IEEE, 2017.
28. Kumar, Naresh, Rahul Rawat, and S. C. Jain. "Bucket based data deduplication technique for big data storage system." In *Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 2016 5th International Conference on*, pp. 267-271. IEEE, 2016.
29. Xia, Wen, Hong Jiang, Dan Feng, and Lei Tian. "DARE: A deduplication-aware resemblance detection and elimination scheme for data reduction with low overheads." *IEEE Transactions on Computers* 65, no. 6 (2016): 1692-1705.
30. Rashid, Fatema, Ali Miri, and Isaac Woungang. "Proof of storage for video deduplication in the cloud." In *Big Data (BigData Congress), 2015 IEEE International Congress on*, pp. 499-505. IEEE, 2015..
31. T.Yujuan, J.Hong, F.Dan, T.Lei, and Y.Zhichao, "CAB dedupe: A Causality- Based Deduplication Performance Booster for Cloud Backup Services" in Parallel & Distributed Processing Symposium, IEEE International, 2011, pp.1266-1277.
32. K. Yang and X. Jia, "An Efficient and Secure Dynamic Auditing Protocol for Data Storage in Cloud Computing " Parallel and Distributed Systems, IEEE Transactions on, vol., pp.1-1,2012
33. Z. Yang, W. Yongwei, and Y. Guangwen, "Droplet: A Distributed Solution of Data Deduplication " in Grid Computing (GRID), 2012 ACM/IEEE 13th International Conference on, 2012, pp.114-121.
34. N. Mandagere, P. Zhou, M.A. Smith, and S. Uttamchandani, "Demystifying data deduplication" presented at the Proceedings of the ACM/IFIP/USENIX Middleware'08 Conference Companion, Leuven, Belgium, 2008.