Third International Conference on "Materials, Computing and Communication Technologies"
in association with International Journal of Scientific Research in Science,
Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

Data Cleaning and Backup System in Cloud Computing Using Attribute Based Encryption

Mrs. A. Kalaiyarasi[1], Dr. E. Punarselvam[2], Mr. S. Hari Prasath[3], Mr. T. Muralitharan[3], Mr. V. Prasanth[3]

[1]Assistant Professor, Department of Information Technology, Muthayammal Engineering College (Autonomous), Rasipuram - 637 408, Tamil Nadu, India

[2]Professor and Head, Department of Information Technology, Muthayammal Engineering College (Autonomous), Rasipuram - 637 408, Tamil Nadu, India

[3]Student, Department of Information Technology, Muthayammal Engineering College (Autonomous), Rasipuram - 637 408, Tamil Nadu, India

ABSTRACT

Big data is widely considered as potentially the next dominant technology in IT industry. It offers simplified system maintenance and scalable resource management with storage systems. As a fundamental technology of cloud computing, storage has been a hot research topic in recent years. The high overhead of virtualization has been well addressed by hardware advancement in CPU industry, and by software implementation improvement in hypervisors themselves. Existingsystems have made efforts to reduce storage image storage consumption by means of de-duplication within a storage area network system It also provides a comprehensive set of storage features including instant cloning for STORAGE images, on-demand fetching through a network, and caching with local disks by copy-on-read techniques. Experiments show that SILO features perform well and introduce minor performance overhead.

I. INTRODUCTION

CLOUD COMPUTING

Cloud computing is a computing paradigm, where a large pool of systems are connected in private or public networks, to provide dynamically scalable infrastructure for application, data and file storage. The idea of cloud computing is based on a very fundamental principal of reusability of IT capabilities. The difference that cloud computing brings compared to traditional concepts of "grid computing", "distributed computing", "utility computing", or "autonomic computing" is to broaden horizons across organizational boundaries. Forrester defines cloud computing as: "A pool of abstracted, highly scalable, and managed compute infrastructure capable of hosting end customer applications and billed by consumption." Cloud Computing is a technology that uses the internet and central remote servers to maintain data and applications. A simple example of cloud computing is Yahoo email, Gmail, or Hotmail.

BIG DATA

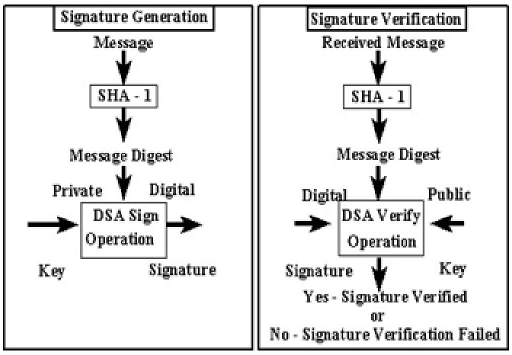
Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reduction and reduced risk. the wide variety of types of data and the velocity at which the data must be must processed. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily relational database for analysis, new approaches to storing and analyzing data have emerged that rely less on data schema and data quality. Instead, raw data with extended

STUDY OF DEDUPLICATION

In computing, data deduplication is a specialized datacompression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage. This technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenevera match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the match frequency is dependent on the chunk size), the amount of data that must be stored or transferred can be greatly reduced.

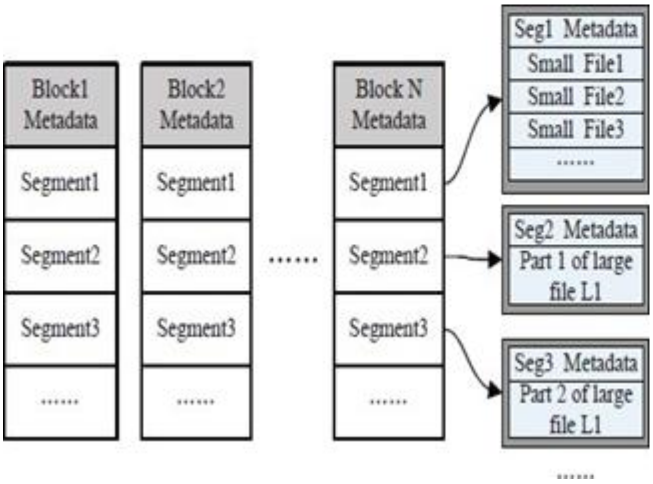
ALGORITHM

Secure Hash Algorithm



ALGORITHM

SILO similarity algorithm



Files in the backup stream are first chunked, fingerprinted, and packed into segments by grouping strongly correlated small files and segmenting large files in the File Agent. For an input segment Snew, Psuedocode for SiLo:

Step 1: Check to see if Snew is in the SHTable. If it hits in SHTable, SiLo checks if the block Bbk containing Snew's similar segment is in the cache. If it is not in the cache, SiLo will load Bbk from the disk to the Read Cache according to the referenced block ID of Snew's similar segment, where a block is replaced in the FIFO order if the cache is full.

Step 2: The duplicate chunks in Snew are detected and eliminated by checking the fingerprint sets of Snew with LHTable (fingerprints index) of Bbk in the cache.

Step 3: If Snew misses in SHTable, it is then checked against recently accessed blocks in the read cache for potentially similar segment (i.e., locality-enhanced similarity detection).

Step 4: Then SiLo will construct input segments into blocks to retain access locality of the input backup stream. For an input block Bnew, SiLo does following:

Step 5: The representative fingerprint of Bnew will be examined to determine the stored backup nodes of data block Bnew.

Step 6: Silo checks if the Write Buffer is full. If the Write Buffer is full, a block there is replaced in the FIFO order by Bnew and then written to the disk.

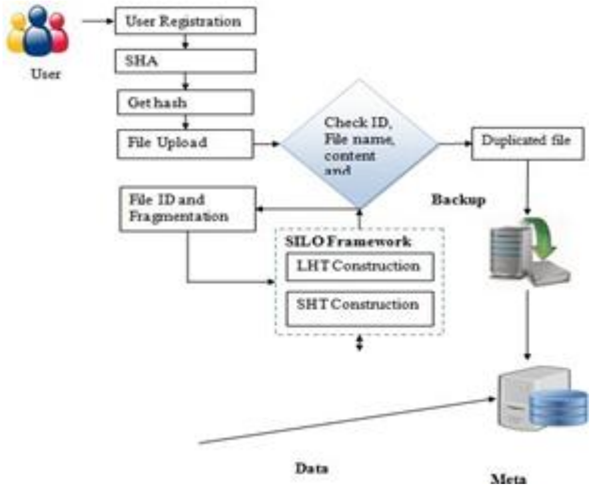
II. PROPOSED SYSTEM

In de-duplication framework, propose system implement block level de-duplication system and named as similarity and locality based de-duplication framework that is a scalable and short overhead near-exact de¬duplication system, to defeat the aforementioned shortcomings of existing schemes. Data de-duplication not only reduces the storage space overheads, but also minimizes the network transmission of redundant data in the network storage system.

ADVANTAGES

- Silo is able to remove large amounts of redundant data, dramatically reduce the numbers of accesses to on-diskindex.
- Maintain a very high de-duplication throughput.

ARCHITECTURE



III. MODULES

- Cloud resource allocation
- De-duplication scheme
- File system analysis
- Data sharing components Evaluation criteria

MODULES DESCRIPTION

CLOUD RESOURCE ALLOCATION

The virtualization is being used to provide ever-increasing number of servers on virtual machines (STORAGEs), reducing the number of physical machines required while preserving isolation between machine instances. This approach better utilizes server resources, allowing many different operating system instances to run on a small number of servers, saving both hardware acquisition costs and operational costs such as energy, management, and cooling.

DATA SHARING COMPONENTS

In this module, we can analyze data sharing components and Meta server in SILO responsible for managing all data servers. It contains SHT and LHT table for indexing each files details for improving search mechanisms. A dedicated background daemon thread will immediately send a heartbeat message to the problematic data server and determines if it is alive. This mechanism ensures that failures are detected and handled at an early stage. The stateless routing algorithm can be implemented since it could detect duplicate data servers even if no one is communicating with them.

DEDUPLICATION SCHEME

De-duplication is a technology that can be used to reduce the amount of storage required for a set of files by identifying duplicate "chunks" of data in a set of files and storing only one copy of each chunk. Subsequent requests to store a chunk that already exists in the chunk store are done by simply recording the identity of the chunk in the file's blocklist; by not storing the chunk a second time, the system stores less data, thus reducing cost.

FILE SYSTEM ANALYSIS

In this module, we first broke STORAGE disk images into chunks, and then analyzed different sets of chunks to determine both the amount of de-duplication possible and the source of chunk similarity. We use the term disk image to denote the logical abstraction containing all of the data in a STORAGE, while image files refers to the actual files that make up a disk image. A disk image is always associated with a single STORAGE; a monolithic disk image consists of a single image file, and a spanning disk image has one or more image files, each limited to a particular size. Files are storedin data server with block id and this can be monitored by Data servers. Data servers are mapped by using Meta servers.

EVALUATION CRITERIA

We showed that data localization have little impact on deduplication ratio. However, factors such as the base operating system or even the Linux distribution can have a major impact on deduplication effectiveness. Thus, we recommend that hosting centers suggest "preferred" operating system distributions for their users to ensure maximal space savings. If this preference is followed subsequent user activity will have little impact on de-duplication effectiveness.

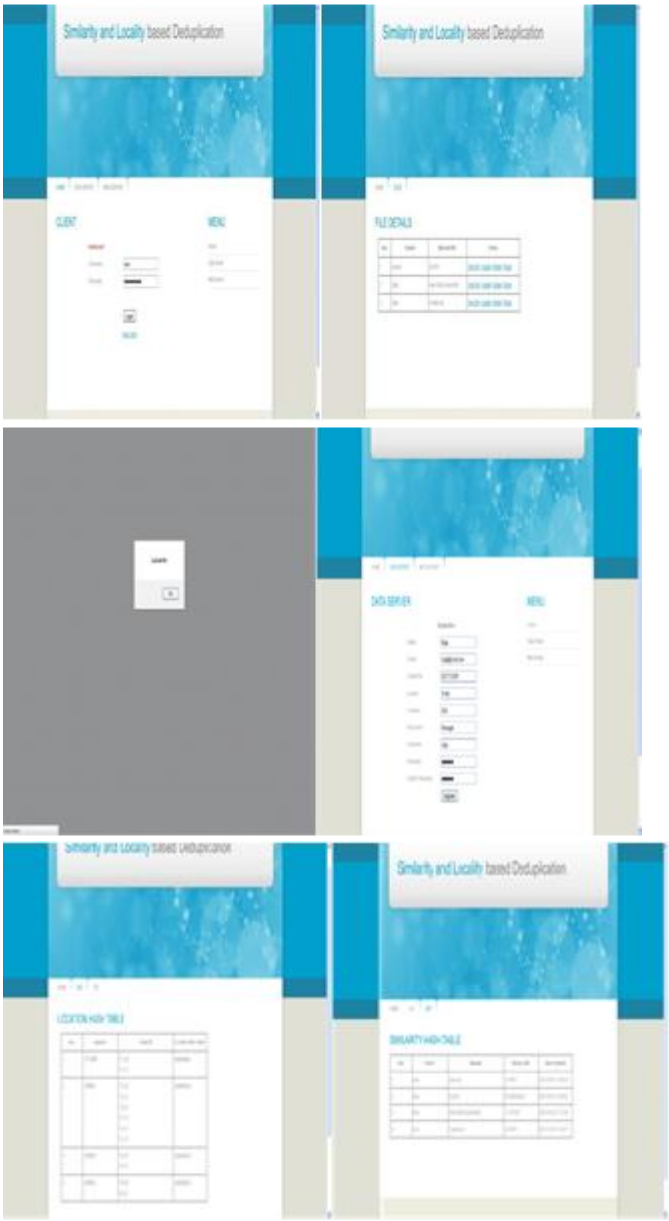
IV. CONCLUSION

In cloud many data are stored again and again by user. So the user need more spaces store another data. That will reduce the memory space of the cloud for the users. To overcome this problem uses the de¬duplication concept. Data de-duplication is a method for sinking the amount of storage space an organization wants to save its data. In many associations, the storage systems surround duplicate copies of many sections of data. For instance, the similar file might be keep in several dissimilar places by dissimilar users, two or extra files that aren't the same may still include much of the similar data Experimental metrics are proved that our proposed approach provide improved results in de-duplication process.

V. FUTURE ENHANCEMENT

In future we can extend our work to handle multimedia data for de-duplication storage. The multimedia data includes audio, image and videos. And also implement heart beat protocol recover each data server and increase scalability process of system.

VI. RESULTS



VII. REFERENCES

[1]. D. Meyer and W. Bolosky, "A study of practical deduplication," in Proceedings of the 9th USENIX Conference on File and Storage Technologies, 2011.

[2]. B. Debnath, S. Sengupta, and J. Li, "Chunkstash: speeding up deduplication using flash inline storage memory," in Proceedings of the 2010 USENIX conference on USENIX annual technical conference. USENIX Association, 2010.

[3]. W. Dong, F. Douglis, K. Li, H. Patterson, S. Reddy, and P. Shilane, “Tradeoffs in scalable data routing for deduplication clusters,” in Proceedings of the 9th USENIX conference on File and storage technologies. USENIX Association, 2011.

[4]. E. Kruus, C. Ungureanu, and C. Dubnicki, “Bimodal content defined chunking for backup streams,” in Proceedings of the 8th USENIX conference on File and storage technologies. USENIX Association, 2010.

[5]. G.Wallace, F. Douglis, H. Qian, P. Shilane, S. Smaldone, M. Chamness, and W. Hsu, “Characteristics of backup workloads in production systems,” in Proceedings of the Tenth USENIX Conference on File and Storage Technologies, 2012.

[6]. A. Broder, “On the resemblance and containment of documents,” in Compression and Complexity of Sequences 1997.

[7]. D. Bhagwat, K. Eshghi, and P. Mehra, “Content-based document routing and index partitioning for scalable similarity-based searches in a large corpus,” in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007, pp. 105–112.

[8]. Y. Tan, H. Jiang, D. Feng, L. Tian, Z. Yan, and G. Zhou, “SAM: A Semantic-Aware Multi-Tiered Source De- duplication Framework for Cloud Backup,” in IEEE 39th International Conference on Parallel Processing. IEEE, 2010, pp. 614–623.

[9]. M. Lillibridge, K. Eshghi, D. Bhagwat, V. Deolalikar, G. Trezise, and P. Camble, “Sparse 7th conference on File and storage technologies, 2009, pp. 111–123.

[10].D. Bhagwat, K. Eshghi, D. Long, and M. Lillibridge, “Extreme binning: Scalable, parallel deduplication for chunk-based file backup,” in IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems. IEEE, 2009, pp. 1–9.