ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

IMPROVING THE PERFORMANCE OF CLOUD STORAGE THROUGH ELIMINATION OF DUPLICATE ENCRYPTED BIG DATA

V.Mary Rajam Vandana[*1], Dr.K.Thamodaran [2]
[*] MPhil Research Scholar, Dept. of Computer Science,Maruthupandiyar College, Thanjavur, Tamilnadu, India-613 403.
Professor, Dept. of Computer Science, Marudupandiyar College, Thanjavur, Tamilnadu, India-613 403.

DOI: 10.5281/zenodo.1012533

ABSTRACT
Big data means a situation where the logistics of storing, processing, or analyzing data have surpassed traditional operational abilities of organizations. Cloud computing is purported to be the future of the IT industry. In cloud computing environment, as the infrastructure not owned by users, it is desirable that its security and integrity must be protected and verified time to time. Cloud computing marks a true paradigm shift in how the computing would happen in the future and cloud computing is likely to have the same impact on IT industry that foundries have had on the manufacturing industry. Although, providing security to cloud storage and access the information is the significant issue. In this paper, a security system is designed to eliminate the duplicate encrypted big data stored in cloud storage utilizing proxy re-encryption algorithm. The experimental results indicate the effectiveness of the proposed security system for big data deduplication in cloud storage.

KEYWORDS: Big Data, Cloud Computing, Duplicate Data, Proxy Re-Encryption, Security.

I. INTRODUCTION
Data mining is a process that uses a variety of data analysis methods to discover the unknown, unexpected, interesting and relevant patterns and relationships in data that may be used to make valid and accurate predictions. Analytical data mining algorithms automatically build models from the data. A data warehouse is a repository for long-term storage of data from multiple sources, organized so as to facilitate management decision making. The data are stored under a unified schema and are typically summarized. Data warehouse systems provide some data analysis capabilities, collectively referred to as On-Line Analytical Processing (OLAP) [5], [8], [12], [29]. The term “Big Data” was enlightened very first time by Mr. John in 1998. Big Data has to deal with large and complex datasets that can be structured, semi-structured, or unstructured and will typically not fit into memory to be processed. They have to be processed in place, which means that computation has to be done where the data resides for processing. Apache Hadoop is an open source Java framework for processing and querying vast amounts of data on large clusters of commodity hardware. Hadoop is a top level Apache project, initiated and led by Yahoo! and Doug Cutting. Apache Hadoop has become an enterprise-ready cloud computing technology for Big Data processing. Hadoop enables scalable, cost-effective, flexible, fault-tolerant solutions [21], [28], [38].

Computer Security means the protection afforded to an automated information system in order to attain the applicable objectives of preserving the integrity, availability, and confidentiality of information system resources such as hardware, software, firmware, data or information and telecommunications [1], [13]. Cloud computing is touted as the next big thing in the Information Technology (IT) industry, which is going to impact the businesses of any size and yet the security issue continues to pose a big threat on it. The security and privacy issues persisting in cloud computing have proved to be an obstacle for its widespread adoption. Cloud computing security or cloud security is an evolving sub-domain of computer security, network security and more broadly, information security. It refers to a broad set of policies, technologies, and controls deployed to protect data, applications, and the associated infrastructure of cloud computing. Cloud storage service providers such as Drop box, Mozy, and others perform deduplication to save space by only storing one copy of each file

uploaded. Should clients conventionally encrypt their files, however, savings are lost. Message-locked encryption is utilized to solve this issue. However it is naturally subject to brute-force attacks that can recover files falling into a known set [14], [26], [36]. Mihir Bellare, Sriram Keelveedhi, Thomas Ristenpart have offered Message-Locked Encryption (MLE) scheme. In this scheme, the encryption and decryption keys are derived from the message itself. MLE provides a way to achieve secure deduplication, a goal currently targeted by numerous cloud-storage providers. This scheme provides ROM security analyses of a natural family of MLE schemes that includes deployed schemes [23].

Cloud storage is a model of data storage in which the digital data is stored in logical pools, the physical storage spans multiple servers (and often locations), and the physical environment is typically owned and managed by a hosting company. These cloud storage providers are responsible for keeping the data available and accessible, and yet the physical environment protected and executing. People and organizations buy or lease storage space from the providers to store data of user, organization, or application. Data de-duplication technology means that in cloud storage environment, only one copy of the same data can be stored instead of storing multiple copies. This technology can greatly reduce the storage space and communication bandwidth [24], [33], [35]. Deduplication is an important data compression technique to save the storage cost at the cloud storage server. 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. The strategies of deduplication can be categorized to two strategies: file-level and block-level deduplication. There are two ways to identify duplications in a cloud storage system. One is *comparing blocks* or *files bit to bit*, and the other is comparing blocks by *hash values*. The advantage of comparing blocks or files bit to bit is that it is accurate, but it is also time consuming. The advantage of comparing blocks or files by hash value is that it is very fast, but there is a chance of accidental collision [18], [19], [27], [30]. The Policy-based De-duplication Mechanism is very useful one to Securing Cloud Storage. In this mechanism, security proxy (SP) and random storage, which separate storage services and security services to ensure the security of user data and improve the system efficiency at the same time [24], [37].

In order to achieve big data security in cloud the access control mechanisms are used. Therefore, it is necessary to develop appropriate security mechanism to improve the performance of cloud storage. In this paper cloud security scheme is proposed to eliminate the encrypted duplicate big data stored in cloud based on proxy re-encryption algorithm. It integrates cloud data deduplication with access control. This proposed scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric key used for data decryption.

The organization of this paper is as follows. In Section II the literature review is offered, section III having materials and methods that includes the information about Data Mining, Data Warehouses and Big Data, section IV explains about cloud computing and cloud security, section V describes about proposed cloud security system, experimental results and discussion is presented in section VI and section VII finish off this paper.

II. LITERATURE REVIEW

Many existing works related to the proposed cloud security system "Elimination of duplicate encrypted big data in cloud storage" are presented in this section. Blaze, Bleumer, and Strauss (BBS) proposed an application called atomic proxy re-encryption, in which a semi-trusted proxy converts a cipher text for Alice into a cipher text for Bob without seeing the underlying plaintext [3]. John R. Douceur, Atul Adya, William J. Bolosky, Dan Simon, Marvin Theimer have presented a mechanism to reclaim space from this incidental duplication to make it available for controlled file replication. In this mechanism includes i) convergent encryption, which enables duplicate files to coalesced into the space of a single file, even if the files are encrypted with different users' keys, and ii) SALAD, a Self- Arranging, Lossy, Associative Database for aggregating file content and location information in a decentralized, scalable, fault-tolerant manner. Large-scale simulation experiments show that the duplicate-file coalescing system is scalable, highly effective, and fault-tolerant. It provides high availability and reliability while executing on a substrate of inherently unreliable machines primarily through a high degree of replication of both file content and directory infrastructure [4].

Ateniese et.al proposes a PDP model for auditing outsourced data based on homomorphic tags of RSA. Recently, one of the biggest concerns with cloud data is the data verification at untrusted servers, because the

service provider may decide to cheat the client for the benefit of their own. The main solution of this problem includes private verifiability and public verifiability [6]. Chow et al. have presented the traditional in-house authorization and authentication framework that were employed previously cannot be extended to the cloud environment and would probably need some modification to be compatible to the services of cloud computing[7]. Zhe Sun, Jun Shen, Jianming Yong have developed deduplication cloud storage system, named "DeDu" runs on commodity hardware which consists of two major components, a front-end deduplication application and Hadoop Distributed File System. Hadoop Distributed File System is common back-end distribution file system, which is used with a Hadoop database. At the front end, it has a deduplication application and at the back end, there are two main components, which are HDFS and HBase respectively used as a mass storage system and a fast index. Fortunately, with the rocket-like development of cloud computing, the advantages of cloud storage have become obvious, and the concept of cloud storage has become accepted by the community. Promising results were obtained from our simulation using VMware to simulate a cloud environment and execute the application on the cloud environment [15].

Alsafi, H.M., Abduallah, W.M. and Khan Pathan.A have proposed a mechanism which is detecting anomaly by differentiating between normal and abnormal activities within the cloud. This is accomplished by stating or delineating some boundaries for valid and normal activities in the cloud network. There's also an added level of focus in this technique for anomaly detection. Data mining techniques are more flexible and easily to deploy at any point. Putting data mining into effect in the cloud network makes available the opportunity to extract meaningful information from data warehouse that are integrated into the cloud, this reduces the infrastructure storage costs. Customers or users of a cloud service only have to pay for the data mining tool that's been used [16]. Pasquale Puzio, Refik Molva, Melek O nen, Sergio Loureiro have proposed ClouDedup security system to provide secure and efficient storage service which assures block-level deduplication and data confidentiality at the same time. The security of ClouDedup relies on its new architecture with metadata manager and an additional server. The server adds an additional encryption layer to prevent well-known attacks against convergent encryption and thus protect the confidentiality of the data; on the other hand, the metadata manager is responsible of the key management [27].

Wen.Z.C., J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li have developed a framework to validate the deduplication of image storage in cloud and verify the correctness of deduplication. In this framework, initially each user uploads an encrypted image, and calculate its hash value as the fingerprint. Subsequently, the fingerprint is sent to both cloud servers for checking duplicates. If the storage and verification servers both reply to the user with 'no deduplication', the user transfers his data to the servers. Otherwise, once the fingerprint is consistently found, the user gives up uploading data for deduplication. Specially, when the fingerprint is only found in one server, it implies that the results are inconsistent and at least one of servers is invalid. The security and efficiency analysis is also presented. In order to efficiently complete deduplication, the hash value of encrypted image is calculated and transferred to both two cloud servers as the fingerprint used in the deduplication process which is proved to be secure [31]. Although, there are a variety of Intrusion Detection techniques available in the cloud environment, this review paper exposes and focuses on different IDS in cloud networks through different categorizations and conducts comparative study on the security measures of Dropbox, Google Drive and iCloud, to illuminate their strength and weakness in terms of security [17], [32].

Madhuri Kavade, A.C. Lomte have represented regarding the important challenges of today's cloud storage services and the management of increasing amount of data. To create data management scalable, de-duplication has been a widely known technique to reduce space for storing and transfer information measure in cloud storage. Instead of keeping multiple data one physical copy and referring different redundant data to it copy. Convergent Encryption, additionally called content hash keying, could be a cryptosystem that produces identical cipher text from identical plaintext files. The proposed system introduces a baseline approach during which file level de-duplication is performed and it doesn't support on-line approach [34]. Julio C. S. Dos Anjosa et.al. have proposed an innovative approach for providing a security infrastructure support to Big Data Analytic in Cloud-based systems named Fast-sec. Fast-Sec handles systems with large volumes of data from heterogeneous sources, in which users may access the system by different platforms, consuming or providing data. The security infrastructure proposed in Fast-Sec provides an authentication mechanism for users, and data access control adapted to high demands from cloud based Big Data environment. The reported results show the adequacy of the proposed safety infrastructure to the cloud-based systems processing Big Data [37]. Vitor Brock and Habib Ullah Khan have proposed a model to look at the factors that associated with the usage of big data analytics, by synchronizing TAM with Organizational Learning Capabilities (OLC) framework. These models

are applied on the construct, intended usage of big data and also the mediation effect of the OLC constructs is assessed. The data for the study is collected from the students pertaining to information technology disciplines at University of Liverpool, online program. [38].

III. MATERIALS AND METHODS

Data Mining

Data mining is extracting information from meaningful data derived from the mass of figures generated every moment in every part of our life. Data mining covers a wide range of activities. It seeks to provide the answer to questions such as these: (i) What is contained in the data?, (ii)What kinds of patterns can be discerned from the maze of data?, (iii) How can all these data be used for future benefit?. Usually, data analysis is classified into two methods namely (i) supervised and (ii) unsupervised. In both cases, a sample of observed data is required. This data may be termed the training sample. The training sample is used by the data mining activities to learn the patterns in the data [12], [29]. Predictive modelling tasks, where the goal is to predict the value of one column based on the values of other columns, are called supervised tasks. These tasks are similar to the supervision of a teacher who gives you the correct answer for the question, to teach you. The goal in descriptive modelling is to discover patterns and segments of the data. These are unsupervised tasks. There is no notion of a correct answer, or any obvious agreed-upon measure of performance. Unsupervised tasks provide insight to the data as a whole by showing patterns and segments that behave similarly.

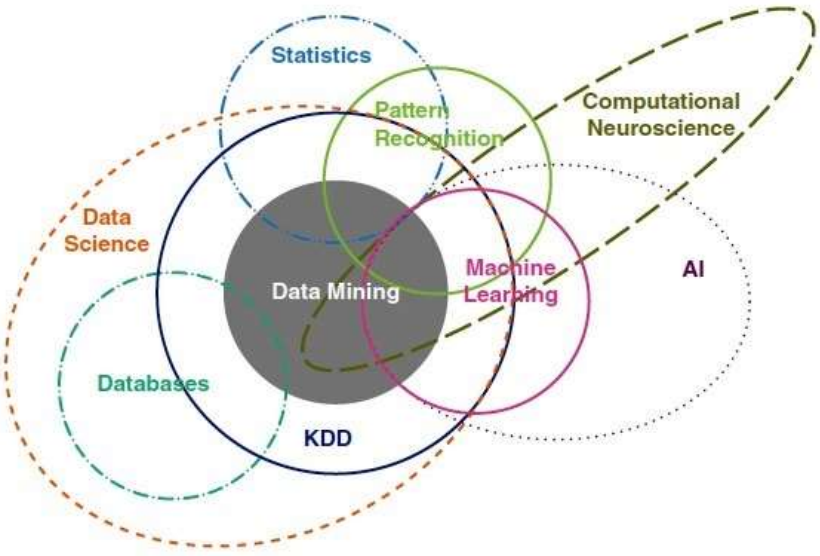

Figure 1 : Nature of Data Mining

Data Warehouses

According to William H. Inmon, a leading architect in the construction of data warehouse systems, "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process" [2]. Data warehouse is a denormalized environment. Data warehouses generalize and consolidate data in multidimensional space. The construction of data warehouses involves data cleaning, data integration, and data transformation and can be viewed as an important preprocessing step for data mining. Moreover, data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data generalization and data mining. Many other data mining functions, such as *association*, *classification*, *prediction*, and *clustering*, can be integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction. Hence, the data warehouse has become an increasingly important platform for data analysis and on-line analytical processing and will provide an effective platform for data mining. Therefore, data warehousing and OLAP form an essential step in the knowledge discovery process [12].

The major task of on-line operational database systems is to perform on-line transaction and query processing. These systems are called on-line transaction processing (OLTP) systems. They cover most of the day-to-day

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting. Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. Data warehouse systems are valuable tools in today's competitive, fast-evolving world. The four keywords, subject-oriented, integrated, time-variant, and nonvolatile, distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems [12].

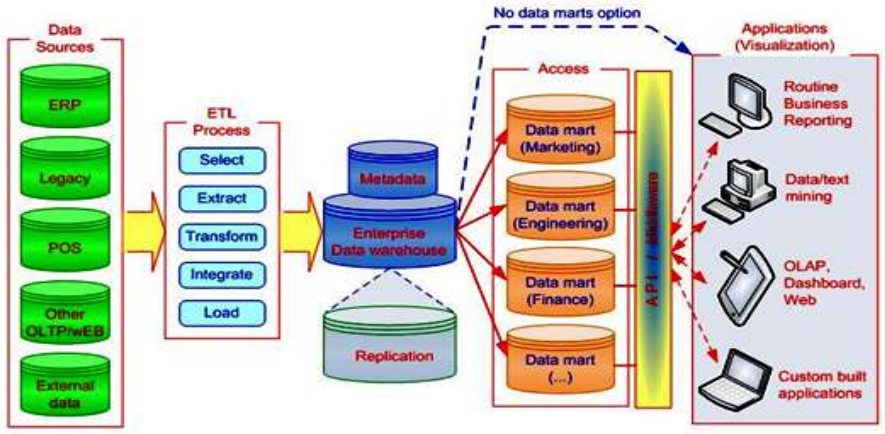

Figure 2 : Architecture of Data Warehouse

Big Data

Big data is a collection of very large and complex set of data which is remote away from the capacity of the existing database management tools or traditional data processing applications. The data includes private and sensitive information that need to be kept secure and safe. Big Data is different because it is generated on a massive scale by countless online interactions among people, transactions between people and systems, and sensor-enabled machinery. Big Data has includes many cross-disciplinary concepts which contain more data than making sense out of. Big Data is a call for us computer scientists to once again provide even better methods to crunch even more diverse, even more complex, even more dynamic, even more fine-grained, even larger data. Big Data brings new opportunities for institutions of higher education, as institutions continue to face unprecedented challenges in their environment [21]. A big data solution must address the three Vs of big data: *data velocity*, *variety*, and *complexity*, in addition to *volume*. Velocity of the data is used to define the speed with which different types of data enter the enterprise and are then analyzed. Variety addresses the unstructured nature of the data in contrast to structured data in weblogs, radio frequency ID (RFID), meter data, stock-ticker data, tweets, images, and video files on the Internet. For a data solution to be considered as big data, the volume has to be at least in the range of 30–50 terabytes (TBs). However, large volume alone is not an indicator of a big data problem. A small amount of data could have multiple sources of different types, both structured and unstructured, that would also be classified as a big data problem.

Big Data has includes many cross-disciplinary concepts which contains more data than making sense out of. Big Data is a call for us computer scientists to once again provide even better methods to crunch even more diverse, even more complex, even more dynamic, even more fine-grained, even larger data. Big Data brings new opportunities for institutions of higher education, as institutions continue to face unprecedented challenges in their environment. Big Data has denser and higher resolutions such as media, photos, and videos from sources such as social media, mobile applications, public records, and databases; the data is either in static batches or dynamically generated by machine and users by the advanced capacities of hardware, software, and network technologies. Examples include data from sensor networks or tracking user behavior. Rapidly increasing volumes of data and data objects add enormous pressure on existing IT infrastructures with scaling difficulties such as capabilities for data storage, advance analysis, and security[28]. A big data solution is differs in all aspects from a traditional Business Intelligence (BI) solution. Data is retained in a distributed file system instead of on a central server. The objectives of big data are (i)The processing functions are taken to the data rather than data being taking to the functions, (ii)Data is of different formats, both structured as well as unstructured,

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

(iii)Data is both real-time data as well as offline data, (iv) Technology relies on Massively Parallel Processing (MPP) concepts[22].

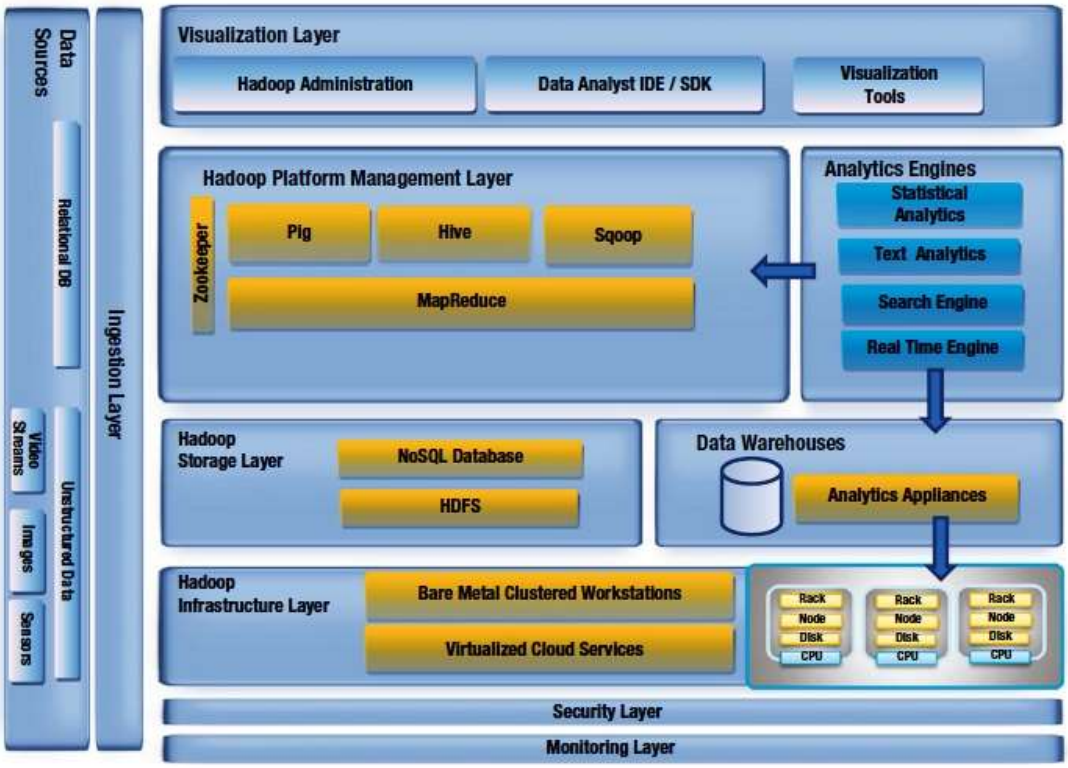

Figure 3: Components of Big Data Architecture

Cloud Enabled Big Data

The big data is affected by cloud-based virtualized environments. The inexpensive option of storage that big data and Hadoop deliver is very well aligned to the "everything as a service" option that cloud-computing offers. Infrastructure as a Service (IaaS) allows the CIO a "pay as you go" option to handle big data analysis. This virtualized option provides the efficiency needed to process and manage large volumes of structured and unstructured data in a cluster of expensive virtual machines. This distributed environment gives enterprises access to very flexible and elastic resources to analyze structured and unstructured data. Map reduce works well in a virtualized environment with respect to storage and computing.

Amazon *Elastic MapReduce* (EMR) is a public cloud option that provides better scaling functionality and performance for MapReduce. Each one of the Map and Reduce tasks needs to be executed discreetly, where the tasks are parallelized and configured to run in a virtual environment. EMR encapsulates the MapReduce engine in a virtual container so that the user can split their tasks across a host of virtual machine (VM) instances. Cloud computing and virtualization have brought the power of big data to both small and large enterprises [25].

An organizational big data value classified in to two, (i) *Analytical use* and (ii) *Enabling new products*. Big data analytics can reveal insights hidden previously by data too costly to process, such as peer influence among customers, revealed by analyzing shoppers transactions, social and geographical data. Big data is increasingly becoming a factor in production, market competitiveness and, therefore, growth. Cutting-edge analysis technologies are making inroads into all areas of life and changing our day-to-day existence. Sensor technology, biometric identification and the general trend towards a convergence of information and communication technologies are driving the big data movement [20].

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IV. CLOUD COMPUTING

Cloud computing and delivery of content stored on a cloud are feasible only due to the interconnectivity supported by a continually evolving internet and by the access to remote resources provided by the World Wide Web. The cloud itself is built around a high-performance interconnect; the servers in a cloud communicate through high-bandwidth and low-latency specialized networks. A packet-switched network transports data units called packets through a network of switches where packets are queued and routed towards their destination thus, subject to variable delay, loss, and possibly arriving at their final destination out of order. A packet-switched network has a network core consisting of routers and control systems interconnected by very high bandwidth communication channels and a network edge where the end-user systems reside [15].

As cloud computing offers exciting new opportunities to the companies to expand their Infrastructure, some companies took it to the next level and started providing cloud services. The big names in cloud service providing industry are Amazon, Google and of late, IBM and Microsoft. Oracle or Sun and HP are also not far behind. Google has built the largest Cloud Computing infrastructure with Data Centers existing in Taiwan, Singapore, Finland, Belgium and Ireland apart from various US states. Amazon, besides being a huge online shopping site, is also a big mover in cloud computing revolution. With Microsoft Azure, Microsoft has also entered the Cloud Computing industry. Oracle or Sun, IBM and Rack Space have also tied their future to Cloud Computing. However, the security issues existing in cloud computing also reflects upon the security breaches and attacks to the Data Centers of these companies [10].

Cloud Data Storage (CDS) is composed of thousands of cloud storage devices clustered by network, distributed file systems and other storage middleware to provide cloud storage service for users. The typical structure of CDS includes storage resource pool, distributed file system, service level agreements (SLAs), and service interfaces, etc. Globally, they can be divided by physical and logical functions boundaries and relationships to provide more compatibilities and interactions. CDS is tending to combine with CDS security, which will provide more robust security [9]. Cloud computing is a completely internet dependent technology where client data are stored and maintained in the data center of a cloud provider like Google, Amazon, Apple Inc., Microsoft etc. The Anomaly Detection System is one of the intrusion detection techniques. It's an area in the cloud environment that is been developed in the detection of unusual activities in the cloud networks [32].

National Institute of Standards and Technology (NIST) defines cloud computing as a computing model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services). These services can be rapidly provisioned and released with minimal management effort or service provider interaction. NIST also defines that the cloud computing can be achieved through three service models: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Cloud computing can be implemented by the four deploymentmodels: Private Cloud, Community Cloud, Public Cloud and Hybrid Cloud. This emerging paradigm allows an organization to reduce costs and develops highly scalable solutions. Cloud promises customers with the benefits of a more convenient way of provisioning IT resources at a faster speed and with a lower cost, compared to traditional IT processes and systems [11].

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

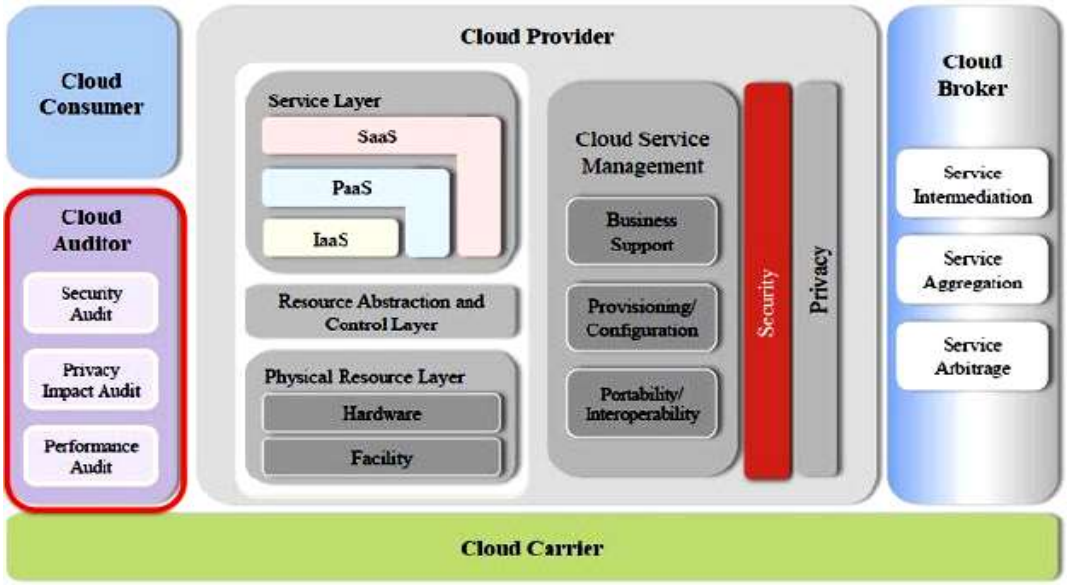

Figure 4: Cloud Computing Architecture

V. PROPOSED CLOUD DATA SECURITY SYSTEM

Asymmetric Key Encryption and Elliptic Curve Cryptography

Public key cryptography or asymmetrical cryptography is using pair of keys for security purpose. The public key may be distributed widely and private key kept as secret by the owner. This accomplishes two functions: authentication, which is when the public key is used to verify that a holder of the paired private key sent the message, and encryption, whereby only the holder of the paired private key can decrypt the message encrypted with the public key. In a public key encryption system, any person can encrypt a message using the public key of the receiver, but such a message can be decrypted only with the receiver's private key. The Elliptic curve cryptography works with points on curve. The security of this type of public key cryptography depends on the elliptic curve discrete logarithm problem. The main advantage of elliptic curve cryptography is that the keys can be much smaller. Elliptic curve cryptography provides a methodology for obtaining high speed, efficient, scalable implementation of networks security protocols.

Cryptographic Hash Function

A hash function is any function that can be used to map data of arbitrary size to data of fixed size. The values returned by a hash function are called hash values, hash codes, hash sums, or simply hashes. One use is a data structure called a hash table, widely used in computer software for rapid data lookup. Hash functions accelerate table or database lookup by detecting duplicated records in a large file. An example is finding similar stretches in DNA sequences. They are also useful in cryptography. A cryptographic hash function allows one to easily verify that some input data maps to a given hash value, but if the input data is unknown, it is deliberately difficult to reconstruct it (or equivalent alternatives) by knowing the stored hash value. This is used for assuring integrity of transmitted data, and is the building block for HMACs, which provide message authentication.

Elimination of Duplicate Encrypted Big Data in Cloud Storage

In order to improve the performance of cloud storage data deduplication technique is applied on encrypted big data stored in cloud. The data deduplication techniqueis a specialized data compression technique for eliminating duplicate copies of repeating data in storage area. 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.

According to the data granularity, deduplication strategies can be categorized into two main categories: file-level deduplication and block-level deduplication , which is nowadays the most common strategy. In block-based deduplication, the block size can either be fixed or variable. Another categorization criteria is the location at which deduplication is performed: if data are deduplicated at the client, then it is called source-based

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

deduplication, otherwise target-based. In source-based deduplication, the client first hashes each data segment he wishes to upload and sends these results to the storage provider to check whether such data are already stored: thus only ”undeduplicated” data segments will be actually uploaded by the user. While deduplication at the client side can achieve bandwidth savings, it unfortunately can make the system vulnerable to side-channel attacks whereby attackers can immediately discover whether a certain data is stored or not. On the other hand, by deduplicating data at the storage provider, the system is protected against side-channel attacks but such solution does not decrease the communication overhead.

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 whenever a 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.

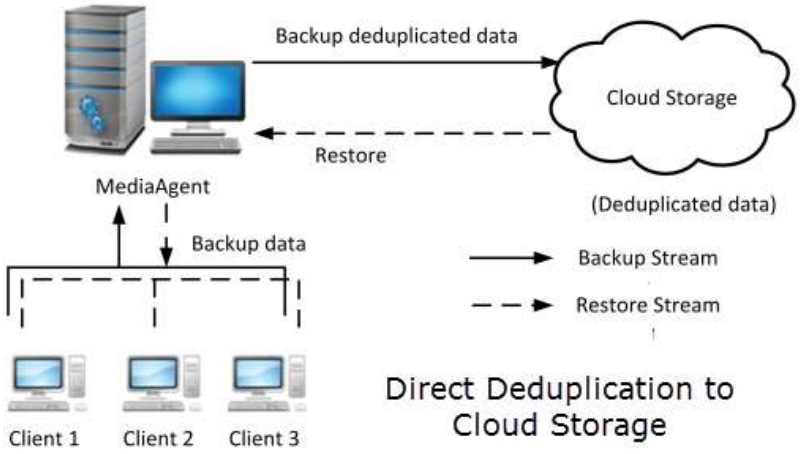

Figure 5: Direct Deduplication to Cloud Storage

Advantages of Deduplication

i.) The network data deduplication is employed to reduce the number of bytes that must be transferred between endpoints, which can reduce the amount of bandwidth required.

ii.) In storage based data deduplication reduces the quantity of storage required for a given set of files. It is most effective in applications where many copies of very similar or even identical data are stored on a single disk.

iii.) Generally the data backup process regularly performs to protect against data loss, most data in a given backup remain unchanged from the previous backup. Common backup systems try to exploit this by omitting files that haven't changed or storing differences between files.

iv.) Virtual servers and virtual desktops benefit from deduplication because it allows nominally separate system files for each virtual machine to be coalesced into a single storage space. At the same time, if a given virtual machine customizes a file, deduplication will not change the files on the other virtual machines.

Record Linkage

Record linkage (RL) is the task of finding records in a data set that refer to the same entity across different data sources (e.g., data files, books, websites, and databases). Record linkage is necessary when joining data sets based on entities that may or may not share a common identifier (e.g., database key, URI, National identification number), as may be the case due to differences in record shape, storage location, or curator style or preference. A data set that has undergone RL-oriented reconciliation may be referred to as being cross-linked. Record linkage is called data linkage in many jurisdictions, but is the same process.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Data Preprocessing

Record linkage is highly sensitive to the quality of the data being linked, so all data sets under consideration (particularly their key identifier fields) should ideally undergo a data quality assessment prior to record linkage. Many key identifiers for the same entity can be presented quite differently between (and even within) data sets, which can greatly complicate record linkage unless understood ahead of time.

Proposed Big Data Security in Cloud

The proposed security scheme is developed based on data ownership challenge and Proxy Re-Encryption (PRE) to manage encrypted data storage with deduplication in cloud environment. The main aspire is to solve the issue of deduplication in the situation where the data holder is not available or difficult to get involved. Meanwhile, the performance of data deduplication in our scheme. It is not influenced by the size of data, thus applicable for big data. In the proposed a cryptographically secure and efficient scheme to check the ownership of a file, in which a client proves to the server that it indeed possesses the entire file without uploading the file. Anonym zing the data is also important to ensure that privacy concerns are addressed. It should be ensured that all sensitive information is removed from the set of records collected. The data deduplication mechanism provides a practical solution with partial semantic security. This solution supports deduplication on plaintext and cipher text. It works based on the assumption that CSP knows the encryption key of data. Thus it cannot be used in the situation that the CSP cannot be fully trusted by the data holders or owners. The proposed a hybrid data deduplication mechanism having the merits such as

(i) Save Storage Space
(ii) High in Privacy
(iii) Multiple Encryption Process for protecting the data.

The proposed security scheme is developed based on data ownership challenge and Proxy Re-Encryption (PRE) algorithm to manage encrypted data storage with deduplication. This scheme is used to solve the issue of deduplication in the situation where the data holder is not available or difficult to get involved. In addition to that the performance of data deduplication scheme is appropriate for big data. Specifically, the objectives of this security scheme are summarized as below:

i.) Motivate to save cloud storage and preserve the privacy of data holders by proposing a scheme to manage encrypted data storage with deduplication. Our scheme can flexibly support data sharing with deduplication even when the data holder is offline, and it does not intrude the privacy of data holders.
ii.) Propose an effective approach to verify data ownership and check duplicate storage with secure challenge and big data support.
iii.) This scheme integrates cloud data deduplication with data access control in a simple way, thus reconciling data deduplication and encryption.
iv.) This scheme proves the security and assesses the performance of the proposed scheme through analysis and simulation. The results show its efficiency, effectiveness and applicability.

Diffie–Hellman Key Exchange

Diffie–Hellman key exchange (D–H) is a specific method of securely exchanging cryptographic keys over a public channel and was one of the first public-key protocols as originally conceptualized by Ralph Merkle and named after Whitfield Diffie and Martin Hellman. D–H is one of the earliest practical examples of public key exchange implemented within the field of cryptography. Traditionally, secure encrypted communication between two parties required that they first exchange keys by some secure physical channel, such as paper key lists transported by a trusted courier. The Diffie–Hellman key exchange method allows two parties that have no prior knowledge of each other to jointly establish a shared secret key over an insecure channel. This key can then be used to encrypt subsequent communications using a symmetric key cipher. Diffie–Hellman is used to secure a variety of internet services.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

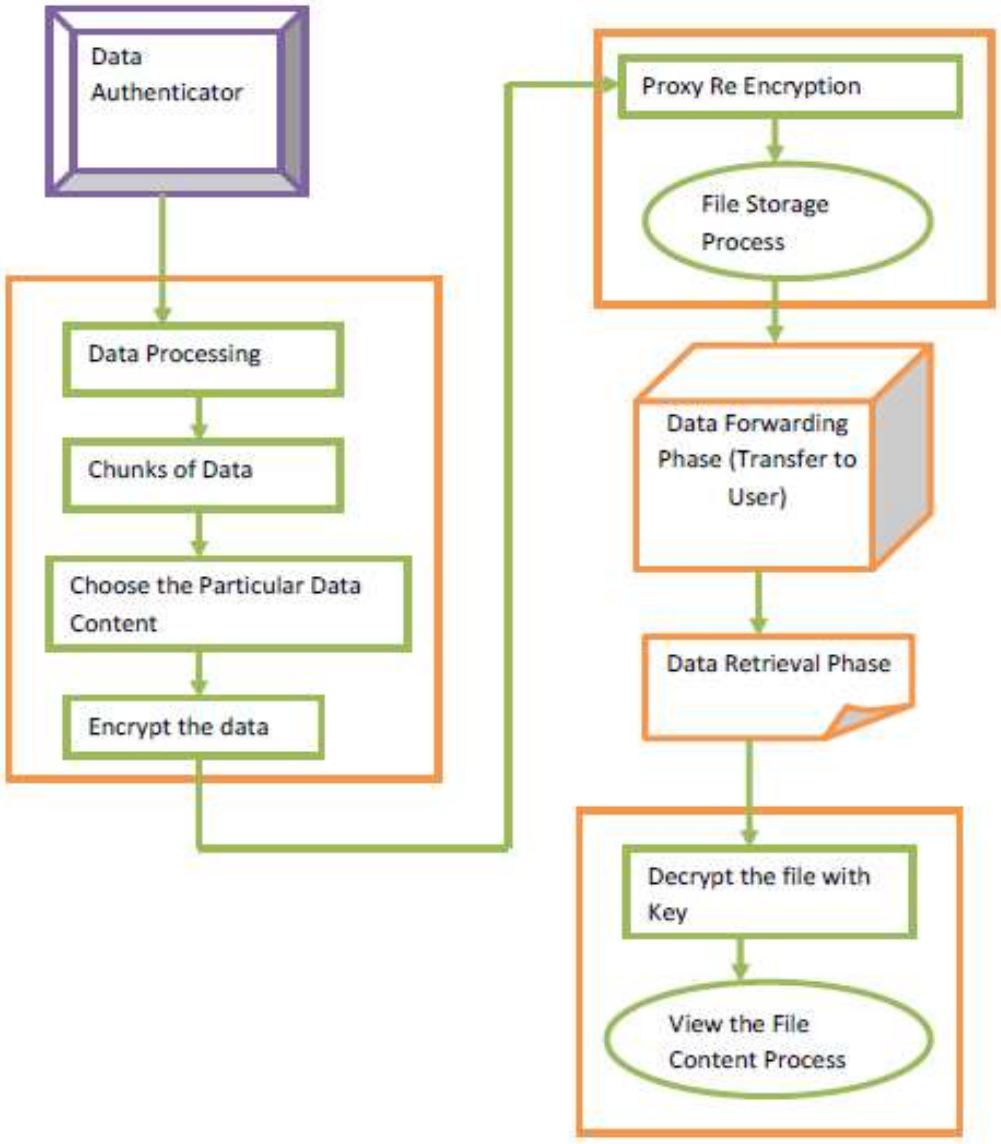

Figure 6: Architecture of Elimination of Duplicate Encrypted Big Data in the Cloud

Proxy Re-encryption

Proxy re-encryption schemes are proposed by Mambo and Okamoto. The proxy re-encryption scheme provides security comparing to previous existed methods. Proxy server is maintained separately for storing public keys and performs actions like encoding, decoding and encryption, decryption. Whenever user wants to send messages, initially he sends the re-encryption key to the proxy server which is also called as storage server along with the message. Now the storage server encrypts that user message by using key and stored in system. Here no third party involved, all keys are maintained by authorized server only. And one more interesting thing is, during transformation server does not know the plaintext. If any another user wants to take those messages, then he connects with storage server and sending key which is used by sender. If the key was matched to the original key, then storage sever gives the permission for accessing data, and send it to the receiver. The Proxy re-encryption algorithm is given below.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

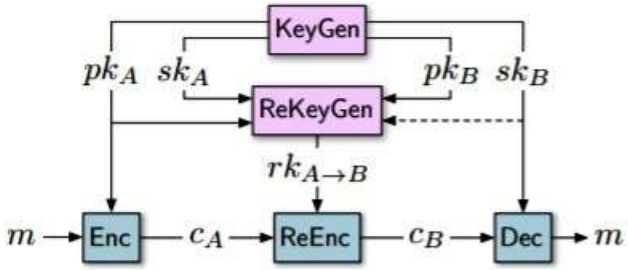

Figure 7: Architecture of Proxy Re-encryption system

Algorithm for Proxy Re-encryption

A proxy re-encryption algorithm includes the terms such as *KeyGen, ReKeyGen, Enc, ReEnc, Dec.*.

Step 1. KeyGen(n) → (pkA, skA). On input security parameter n, the key generation algorithm KeyGen outputs a pair of public and secret keys (pkA, skA) for user A.

Step 2. ReKeyGen(pkA, skA, pkB, skB) → rkA→B. On input the pair of public and secret keys (pkA, skA) for user A and the pair of public and secret 1 keys (pkB, skB) for user B, the re-encryption key generation algorithm ReKeyGen outputs a re-encryption key rkA→B.

Step 3. Enc(pkA, m) → cA. On input the public key pkA and a message m ∈ M, the encryption algorithm Enc outputs a ciphertext cA ∈ C.

Step 4. ReEnc(rkA→B, cA) → cB. On input a re-encryption key rkA→B and a ciphertext cA ∈ C, the re-encryption algorithm ReEnc outputs a second ciphertext cB ∈ C or the error symbol ⊥ indicating cA is invalid.

Step 5. Dec(skA, cA) → m. On input the secret key skA and a ciphertext cA ∈ C, the decryption algorithm Dec outputs a message m ∈ M or the error symbol ⊥ indicating cA is invalid. The plaintext and ciphertext spaces are denoted by M and C, respectively.

VI. RESULTS AND DISCUSSION

In public key encryption, an attacker has gained access to the storage can perpetrate the so called “dictionary attacks” by comparing the cipher texts resulting from the encryption of well-known plaintext values from a dictionary with the stored cipher texts. In reality, even if encryption keys are encrypted with users’ private keys and stored somewhere else, the potentially malicious cloud provider, who has no access to the encryption key but has access to the encrypted chunks (blocks), can easily perform offline dictionary attacks and discover predictable files. This issue arises in where chunks are stored at the storage provider after being encrypted with convergent encryption. Since this has direct access to server data are less secure. In attribute based encryption the complexity of implementation is high and it is not scalable and flexible for large volume of data but this supports large number of duplicate copies so it is much suitable for small data that has much copies. Since it is implemented in symmetric key policy the key distribution and computation is the bigger problem and could damage when compressed.

The data that needs to be stored is first preprocessed if necessary. In certain cases the pre-processing takes up a lot of time based on the type of processing that is implemented. Data cleaning, Normalization, Data hiding, Structuring of data, etc are some of the pre-processing steps available. The next step is the chunking. The process of splitting the given data into any blocks or chunks of data is called as chunking. This is the crucial step in the deduplication process. This is because, based on the size of each chunk the number of duplicate data changes. Based on the initial data, the chunk size should be fixed in such a way that the obtained chunks have large number of duplicates and thus the storage size will be reduced as much as possible.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Table 1: Results of Various Functions of Proposed Security System

	Public Key Encryption	Proxy Re- encryption	Attribute Based Encryption
Key Process	1.3	2.4	3.6
Chunking Process	2.5	4.4	3
Encryption Process	1.8	2.8	3
Decryption Process	1.3	2.8	5

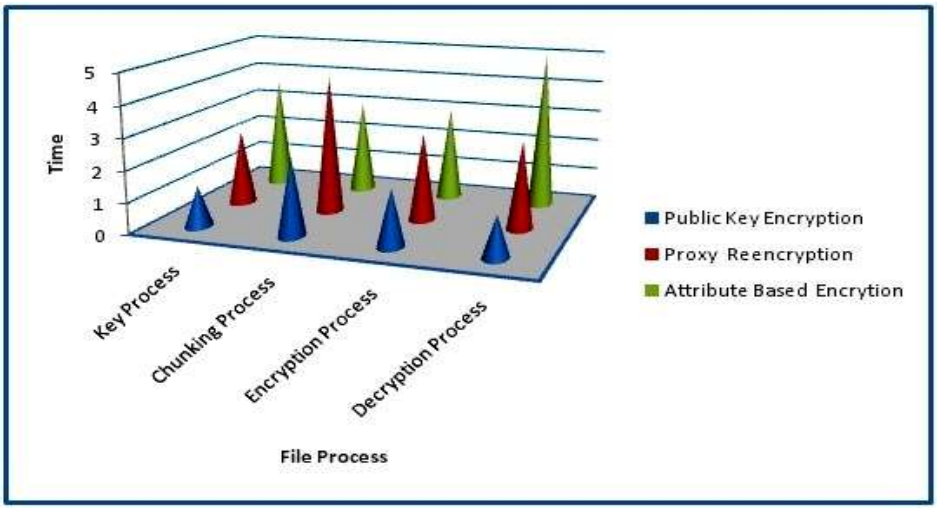

Figure 8: Functions of Proposed Cloud Security System

The goal of the proposed system is to increase security to the data and avoid duplication. Here their uses a message digest technique to avoid the duplication of the file. The message digest technique generates a hash function by using that hash function the data checker can easily find the duplicate data because for the same file the generated hash function will also be same. So it is easy to find the duplication. By finding the duplication can increase the storage and efficiency. We use a data encryption standard for encrypting, decrypting and key generation. The encryption and decryption process uses a 64 bit key so if the attacker gets the file the key cannot be identified and the data will be kept safe and for the key generation it performs the binary rotation operation. This binary rotation keeps on moving the key so the key cannot be hacked. The data owner provides will provide the one time password to the other user. So the other users are requested to view the file only once and they don't have authorization to share that particular file to third party applications. So therefore our project is more efficient and secure.

Table 2 : Competence of Proposed Security Algorithm

PROCESS	PROPOSED ALGORITHM
Secret Key Process	Generating 256-bit Secret Key
Chunking Process	Chunking Process For avoid the Duplication
Encryption Process	Multiple Encryption Process using Proxy Encryption
Pairing Key Process	Using Partially Key Process find the Match Keys
Decryption Process	Without Paring can't able to Decrypt

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

VII. CONCLUSION

Cloud computing is a new computational paradigm that offers an innovative business model for organization to adopt. The security is an important aspect of quality of service. Cloud storage is much more beneficial and advantageous than the earlier traditional storage systems especially in terms of scalability, cost reduction, portability and functionality requirements. Managing encrypted data with deduplication is important and significant in practice for achieving a successful cloud storage service, especially for big data storage. The proposed scheme manage the encrypted big data in cloud with deduplication based on ownership challenge and PRE. This scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption. Extensive performance analysis and test showed that our scheme is secure and efficient under the described security model and very suitable for big data deduplication. The results of computer simulations further showed the practicability of this scheme. Future work includes optimizing our design and implementation for practical deployment and studying verifiable computation to ensure that CSP behaves as expected in deduplication management.

VIII. REFERENCES

[1] Computer Security Handbook, NIST 1995.
[2] W. H. Inmon. Building the Data Warehouse. JohnWiley & Sons, 1996.
[3] Matt Blaze, G. Bleumer, and M. Strauss, "Divertible protocols and atomic proxy cryptography. In Proceedings of Eurocrypt" , volume 1403, pp 127–144, 1998.
[4] John R. Douceur, Atul Adya, William J. Bolosky, Dan Simon, Marvin Theimer, Reclaiming Space from Duplicate Files in a Serverless Distributed File System, pp 1-14, 2002.
[5] Ian H. Witten and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition, Morgan Kaufmann, 2005.
[6] Ateniese.G, K. Fu, M. Green, and S. Hohenberger, "Improved proxy re-encryption schemes with applications to secure distributed storage," ACM Trans. Inform. Syst. Secur., vol. 9, no. 1, pp. 1– 30, 2006.
[7] Chow, R., Golle, P., Jakobsson, M., Shi, E., Staddon, J., Masuoka, R. and Molina, J. Controlling Data in the Cloud: Outsourcing Computation without Outsourcing Control. Proceedings of the 2009 ACM Workshop on Cloud Computing Security, Chicago, pp 85-90, 2009.
[8] Gupta.G.K., Introduction to Data Mining with case studies by, PHI pvt. Ltd., Third printing, 2009.
[9] Talib, A.M., Atan, R., Abdullah, R. and Murad, M.A.A., Security Framework of Cloud Data Storage Based on Multi Agent System Architecture: Semantic Literature Review. Computer and Information Science, 3, 175, 2010.
[10] Danny Harnik, Benny Pinkas, and Alexandra Shulman-Peleg. Side channels in cloud services: Deduplication in cloud storage. Security & Privacy, IEEE, 8: 6, 40–47, 2010.
[11] Mell, P. and Grance., "The NIST Definition of Cloud Computing", NIST Special Publication 800-145, National Institute of Standards and Technology, Gaithersburg, 2011.
[12] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques", 3rd ed., Morgan Kaufmann, 2011.
[13] William Stallings, "Cryptography and Network Security Principles and Practice", Fifth Edition, PHI Publication, 2011.
[14] Wilcox.Z.O., "Convergent Encryption Reconsidered", 2011.
[15] Sun.Z, J. Shen, and J. M. Yong, "DeDu: Building a deduplication storage system over cloud computing," in Proc. IEEE Int. Conf. Comput. Supported Cooperative Work Des., pp. 348–355, 2011.
[16] Alsafi, H.M., Abduallah, W.M. and Khan Pathan, A. IDPS: An Integrated Intrusion Handling Model for Cloud Computing Environment. International Journal of Computing and Information Technology , pp 1-16, 2012.
[17] Grant Wallace, Fred Douglis, Hangwei Qian, Alsafi, H.M., Abduallah, W.M. and Khan Pathan, A. IDPS: An Integrated Intrusion Handling Model for Cloud Computing Environment. International Journal of Computing and Information Technology , pp 1-16, 2012.
[18] Meyer.D.T. and W.J.Bolosky, "A study of practical deduplication", ACM Trans.Storage, vol. 7, no.4, pp.1–20, 2012.
[19] Wallace.G, et al., "Characteristics of backup workloads in production systems", in Proc. USENIX Conf. File Storage Technol., pp. 1–16, 2012.

[20] Alvero A. Cardenas, Pratyusa K. Manadhata, Sreeranga P.Rajan, "Big Data Analytics for Security", IEEE Security & Privacy, vol.11 no.6, pp. 74-76 Nov.-Dec. 2013.
[21] Hurwitz.J, "Big Data for Dummies," Wiley, 2013.
[22] Michael Minelli, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley, 2013.
[23] Bellare.M., S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proc. Cryptology, pp. 296–312, 2013.
[24] Liu.C.Y., X. J. Liu, and L. Wan, "Policy-based deduplication in secure cloud storage," in Proc. Trustworthy Comput. Serv., pp. 250–262, 2013.
[25] Nitin Sawant and Himanshu Shah, " Big Data Application Architecture Q & A", Apress, 2013.
[26] Oltsik.O, Defining big data security analytics. Network world, April 2013.
[27] Puzio.P., R. Molva, M. Onen, and S. Loureiro, "ClouDedup: Secure deduplication with encrypted data for cloud storage," in Proc. IEEE Int. Cof. Cloud Comput. Technol. Sci., pp. 363–370, 2013.
[28] Vignesh Prajapati, "Big Data Analytics with R and Hadoop", Packt Publishing, 2013.
[29] Andrea Ahlemeyer-Stubbe and Shirley Coleman,"A Practical Guide to Data Mining for Business and Industry", John Wiley & Sons, Ltd, 2014.
[30] Meye.P, P.Raipin, F.Tronel, and E. Anceaume, "A secure two phase data deduplication scheme," in Proc. HPCC/CSS/ICESS, pp. 802–809, 2014.
[31] Wen.Z.C., J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li, "A verifiable data deduplication scheme in cloud computing," in Proc. Int. Conf. Intell. Netw. Collaborative Syst., pp. 85–90, 2014.
[32] Arif Sari , A Review of Anomaly Detection Systems in Cloud Networks and Survey of Cloud Security Measures in Cloud Storage Applications. Journal of Information Security, vol 6, pp 142-154, 2015.
[33] Li.J, 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 Distrib. Syst., vol. 26, no. 5, pp. 1206–1216, May 2015.
[34] Madhuri Kavade, A.C. Lomte, "Secure De-Duplication using Convergent Keys for Cloud Storage", 2015.
[35] Zhen-Yu Wang, Yang Lu, Guo-Zi Sun, "A Policy-based De-duplication Mechanism for Securing Cloud Storage", 2015.
[36] Perttula.D, B. Warner, and Z. Wilcox-O'Hearn, "Attacks on convergent encryption", 2016.
[37] Julio C. S. Dos Anjosa et.al. , "Fast-Sec: an approach to secure Big Data processing in the cloud", International Journal of Parallel, Emergent and Distributed Systems, pp 1-18, 2017.
[38] Vitor Brock and Habib Ullah Khan, "Big data analytics: does organizational factor matters impact technology acceptance?", Journal of Big Data, vol 4:21,pp1-28, 2017.

CITE AN ARTICLE

Vandana, V. R., & Thamodaran, K., Dr. (2017). IMPROVING THE PERFORMANCE OF CLOUD STORAGE THROUGH ELIMINATION OF DUPLICATE ENCRYPTED BIG DATA. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, *6*(10), 314-328.