Abstract

In today’s era, different organizations and individuals tend to outsource their data to cloud storage, the security and user privacy protection attract more attention. A novel model of cloud secure storage is proposed, which combines the Hadoop distributed file system (HDFS) with symmetric and public-key cryptography. The model uses the HDFS as the storage platform and the XML format as the logical storage structure. Hadoop uses Kerberos principals and key tabs for user authentication. It also briefly describes how Hadoop uses delegation tokens to authenticate jobs at execution time, to avoid overwhelming the KDC with authentication requests for each job. Kerberos assigns tickets to Kerberos principals to enable them to access Kerberos-secured Hadoop services. Paper contains the all the survey details regarding the security issues on data isolation, intra-cloud data migration and inter-cloud data migration under the environment of a Private Storage Cloud extended with a Partner/Public Cloud. A security encryption schemes based on Hadoop which satisfy the data transmission and storage security and satisfy the server executes digital signature for client data at the same time. It is a distributed encryption system that could reduce the burden on the server, and finally achieve security, stability, efficient and effective storage.
I. INTRODUCTION

Nowadays data which is being collected and processed is extremely large. Big data is extremely large data sets that are analysed and converted into different patterns, trends, and associations. There are three types in big data structured, semi structured and unstructured. Big data technologies are important in providing accurate analysis. Big data faces many challenges like capturing data, curation storage, searching, sharing, analysis and security.

Big data analytics is often associated with cloud computing because the analysis of large data sets in real-time needs a platform like Hadoop to store large data sets across a distributed cluster and MapReduce to coordinate, combine and process data from multiple sources. Cloud computing is extremely popular and highly used computing paradigm, it provides the users massive computing, storage, and software resources on demand [1].

Hadoop is a Distributed framework for analysing big data [2]. It is a platform for structuring Big Data, which solves the problem of formatting it for subsequent analytics purposes. Hadoop has a distributed computing architecture with multiple servers, making it extremely inexpensive to scale and support extremely large data stores. DFS mainly consists of NameNode, DataNode, Job tracker and task tracker [3]. Owen O’Malley et al. designed the Kerberos protocol based on SSL to launch user identity authentication [4]; Indrajit Roy et al. designed and implemented the Airavat platform, which could ensure the vital data secure and privacy protection in the MapReduce calculation process [5]. Hadoop distributed file system is based on the Map Reduce processing technique. MapReduce is a processing technique model for distributed computing based on java [6]. There are two important tasks in it, namely Map and Reduce. Map takes a set of data and converts that into another set of data, the files are split into the lines. Reduce task, takes the output from a map as an input and combines those data rows into a smaller set of rows. As the sequence of the name Map Reduce, the reduce task is always performed after the map job. MapReduce has a major advantage that it is easy to scale data processing over multiple computing nodes. These are data processing primitives, called Mappers and reducers. The core concept is processing speed across large data sets. Breaking large data sets into small pieces, distributing them to as much storage/processing units as possible, and processing data, such that processing and data are tightly coupled with the resulting output being aggregated are the key feature to achieving the goal of speed. A framework of SecureMR has been proposed by Wei Wei et al. to guarantee the data and service integrity [7]. In fact, the research works they have done were mostly aimed at providing the users authentication to identify and to ensure the security as well as privacy protection.

In Hadoop Distributed File systems (HDFS) there is no attempt to verify the identity and group membership of users who interact with (HDFS) and logged in users store data in the Hadoop in a browser without any additional storage media, and the user can obtain the data wherever they are by computers, laptops, mobile phones. It is not having effective security for the confidential data which is not access to unauthorized user. Kerberos protocol messages are protected against eaves dropping and replay attacks. It builds on symmetric key cryptography and requires a trusted third party, and optionally may use public-key
cryptography during some phases of authentication [7]. In Hadoop, Kerberos authentication is used but authentication is not very effective for securing confidential data on Hadoop [8]. Below are the current issues on the HDFS Security

1) **Transmission security**: While transmitting data it may be intercepted, but the data transmission is not working with the strong encryption protection measures.
2) **Access control**: Authorised access control is weak the user data stored in the clouds without setting access authority.
3) **Data storage**: Data stored on the cloud is not classified hence data leakage is possible
4) **Data verification**: Data verification is not strong as it cannot be verified if the right person is accessing the data.

To solve the existing security problems, cloud disk storage based on Hadoop is proposed, the program draw lessons from a security protocols like Kerberos’ authorization process. The classic algorithms such as AES, RSA, blowfish for realizing encryption, and authentication are utilized, and time is checked to inspect if it can complete the encryption and transmission in an acceptable period.
II. SURVEY REVIEW

Network essentials and applications are proposed by Stallings [9]. Hadoop eco system for big data security is proposed by pradeep adluru [10]. David Nunez, Isaac Agudo proposed Cryptographically Enforced accessed control system, Encryption and decryption on data at job tracker which faces performance issue [11]. There is Implementation of SEHadoop model which has delegation token limited performance impact [12]. Certificate and timestamp are checked before login and TSA and directory to verify on server by secure cloud management method [13]. There is a solution proposed by yan Wen for Hadoop but it faces security in terms of efficiency [14]. Data at rest security is proposed for Hadoop security threats this system uses tuning for hdfs [15]. Auth hacker and perimeter eyesdropper is used by Sharif, Sarah Cooney [16]. Spark uses HDFS security system and analysis for spark cloud computing security is done using static security analysis [17]. One solution provided in combination with Kerberos for authentication and single sign on (SSO) to make the system secure [18]. Many security and privacy issues are discussed by Ibrahim Lahmer, Ning Zhang [19]. The RSA digital signature algorithm and its mathematics are used and proved [20]. Large and integrated integer library is built by using C++ and the implementations of Miller-Rabin. Hong Bo Zhou explained cloud computing its applications and standards [21]. There is Hatman intra cloud trust management system in which the first full-scale, reputation-based, trust management system is implemented. By comparing job replica outputs for consistency, Hatman dynamically assesses node integrity [22]. BANLOGIC is implemented by distributing encryption system to reduce burden on the system to achieve security and stability [23].

XU Guang-hui, in his research mentioned, Hadoop as an Apache open source project consisting of several projects like HDFS, MapReduce, HBase, Hive, ZooKeeper and others [24]. Further to his findings it was concluded that HDFS and MapReduce being vital parts of the open source. According to ZHUO Tang, in his findings aimed at Algorithm for MapReduce; MapReduce had aimed at paralleling and dealing with tasks on a much larger scale. Paralleling & Distributed Processes had eventually made MapReduce scheduler become particularly more important [25].

Ghemawat S. also stated that Hadoop lacked serious safety measures making it vulnerable to data leakages. To overcome such vulnerabilities Ghemawat S. integrated Kerberos in Hadoop in the year 2009 with the help of Yahoo. Researchers of Kerberos suggested its implementation by only authorizing the access to users. According to the researchers users of Kerberos had to obtain access from third party centers which in turn will produce an authorization certificate only after which the users will be allowed access. After the authorised certificate, Hadoop Cluster will sort key issues first depending upon the issue minimizing the risk of user data theft. Many other researchers put their theories in different methods to reduce data leakages caused identification and to increase the security of cloud storage. [27] ZHANG Da-Wei, 2009, a researcher on Hadoop -based enterprise file cloudstorage system proposed that the use of HDFS can be used to build a private enterprise.
cloud which can combine Hadoop’s fault tolerance and can also be suitable for the Big Data feature.

In addition to building a secure and private enterprise [28] HouQinhua, Wu Yongwei, Zheng Weimin, researched on Protection of User Data Privacy in Cloud Storage Platform. Hou, further evaluated that a System Security Layer (SSL) secure connection and secure virtual machine monitor could enhance the security of user data in cloud storages. Furthermore, YU Shu-cheng [29], encoded the existing cloud data using the attribute encryption scheme (ABE), the work was simple but effective by simply using a public key to decode the same data before even it was uploaded, the use of public key limited the data access the user had. The user had to ensure to use the “K attributes to decrypt the data, in which the K is the number of threshold to decrypt the data”. The use of the attribute scheme had ensured the safety and security of the data storage but eventually eliminated the use of public key for each individual user at the server side. YU Shu-cheng[30] scheme of attributes enabled the user attributes to be used as user’s public key which also came along with some cons, the data could be decrypted completely when different user attributes would hold their attributes together and get access to all other attributes.

## III. SCHEME DESIGN

As per the design view the software module, the system can be divided into client module, server call module, secret key production and distribution module, data encryption module, data signature module, the data transmission module, data authentication module, and data storage module, in which the client module includes the data encryption module, data transmission module, data signature module while the server call module includes the secret key production and distribution module, data authentication module and data storage module. The three times handshake generated between client and server.

![Figure 3: Three Times Handshake](image-url)
If a user wants to upload file to the server, he should initiate a request first. And if the file is confidential which needs to be encrypted, the client needs a secret key to encrypt the file. Client terminal cannot generate a secret key. The secret key is generated by server and then sent to the client, after which the client will finish data’s symmetric encryption and RSA data signature. The transportation and encryption of data need a certain delay. They use asynchronous request to ensure the fluency of user interface when other operations can be carried out normally. Encryption can ensure the security of the data, but at the same time, it will increase the burden on the system. Considering that, not all the files need to be encrypted, such as common archive files, music files, etc. the client must make sure that the file uploaded needs to be encrypted.

IV. CONCLUSION

In this review paper survey is related to security issues on data isolation, intra-cloud data migration and inter-cloud data migration under the environment of a Private Storage Cloud extended with a Partner/Public Cloud. The security solutions based on the HDFS layer, with master/slave architecture, for the PSC are proposed. The performance analysis of them proves the efficiency of the security design.

A security encryption schemes based on Hadoop which satisfy the data transmission and storage security and satisfy the server executes digital signature for client data at the same time. It is a distributed encryption system that could reduce the burden on the server, and finally achieve security, stability, efficient and effective storage. The next research direction is to optimize the I/O efficiency of HDFS and ensure the scheduling security when executing the Map Reduce tasks in Hadoop. Hadoop uses Kerberos principals and key tabs for user authentication.

V. REFERENCES

Review on Providing Security to Data Stored on HDFS Using Security Protocol

[14] Qiaoyan Wen, A New Solution of Data Security Accessing for Hadoop Based on CP-ABE 2014

FOR CITATION