JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SECURED FILE SHARING ON THE CLOUD USING HADOOP CLUSTER

Dr.A.Adhiselvam

Department of Information Technology, Dr.N.G.P. Arts and Science College, Coimbatore

S.S.Sooraj, Student, Dr.N.G.P. Arts and Science College, Coimbatore

Abstract

Hadoop is an Apache open-source framework for storing and processing huge amount of information throughout clusters of computers. But processing sensitive or private data in hadoop framework requires security model. As it is realized that hadoop become designed without any safety model. The elevated extent of statistics resulting from the attack makes the contemporary detection systems inefficient to detect the hacker. In this research, a brand new kerberos technology is used to deliver the file in securely uploads and downloads. The data device that meets this assignment through a unique integration of secure cryptography strategies primarily based onhadoop cluster.

Keyword: Cloud, Hadoop cluster, MapReduce, Kerberos

1. INTRODUCTION

In 21st century everyone refers to store data on the Cloud.That data may contain account numbers,passwords and other important information that could be used and misused by a criminal or anintruder. Thisdata is retrieved, copied and achieved by Cloud Service Providers (CSPs), often without users' permission andcontrol. These problem presents challenge to protect people privacy from illegal actions. By taking thisproblem into consideration, we introduce self-destructing system to protect people privacy based onactivestorageframework.

2. LITERATUREREVIEW

Let us see some theoretical and methodological contributions to a particular topic. *Ling fang Zeng*,proposed improved Washington's Vanish system for self-destructing data under cloud computing, and it is opento "hopping attack" and "sniffer attack". In this paper working of Safe Vanish to prevent hopping attacks byway of increasing the length of the key shares to rise the attack cost did some more enhancement on the ShamirSecret Sharing algorithm implemented in the Original Vanish system. They presented an improved approach toprevent sniffing attacks by using the public key cryptography system to protect from sniffing operations. Inaddition,theyevaluatedanalyticallythefunctionalityoftheproposed SafeVanishsystem[5].In order to share the data secure cryptographically access control is necessary. Identity-based encryption is used to build data sharing system[1]. Cloud computing is a paradigm in Technology of information(IT) that provides ubiquitous access to shared pools of configurable system resources and often over the internet,Service of higher-level with minimal management effort can be rapidly provisioned[4][5][6].

Yu Zhang,presented a reconfigurable calculating solution that can provide high-performance, flexible processing capabilities for the storage nodes. The dynamic reconfiguration upturns thefunctional density; however, the configuration self-results in extra overhead, which may make the overallperformance to be downgraded [11].MrudulaVarade, VimlaJethani presented the cloud computing relies on sharing of resources to achieve coherence and economy of scale, similar to a utility.ID-based encryption, or identity-based encryption (IBE), is an important primitive of ID-based cryptography because a type of public-key encryption user of public key has some unique information about the user identity [6].

The privacy-preserving public auditing system for data storage security in Cloud Computing utilized the

homomorphic authenticator and random masking to guarantee that TPA would not learn any knowledge about the data content stored on the cloud server during the efficient auditing process, which not only eliminates the burden of cloud user from the tedious and possibly expensive auditing task, but also alleviates the users' fear of their outsourced data leakage[10].

A secret sharing scheme starts with a secret and then derives from it certain shares which are distributed to a group of users (i.e., participants). The secret may be uniquely determined only by certain predetermined subgroups of users which constitute the access structure[7]. The storage capacities increase and applications move into the cloud, cloud becomes a common concept for Internet accessible infrastructure, including the data storage and computing hardware, which is hidden from Internet users. Cloud computing makes data truly mobile and a user can simply access a chosen cloud with any internet accessible device. How to control the lifetime of sensitive data is becoming increasingly important. Data in cloud may be targets of theft or subpoena even if users cleanse their local files, because its copies may be retained for a quite a long time by backup systems, email providers, and other cloud services [8].

FU Xiao, realized emails were being watched by the government. For the advantage that big datatechnologies such as large distributed storage and user behavior analysis and so on emails became one of thehighly popular Big Data that has been targeted at as a large source of intelligence by some organizations keepseye on public accounts every hour every day. The research work was just opposite to what the NSA has did. Todesign and implement a system which can store emails securely, and terminate them clearly when they expired.In another word, a self-destructing emails system. But in this system there is no parallel processing for multiuseraccess[3]. In the existing system there are multiple disadvantages are available. In this Hacker can attack theconfidential data and gain all the information from the database. This is big disadvantage of this system because clients want to security of the data which is confidential from others. In this hacking process thesensitivedata canbemodifiedbyanyone, or ifanyonecandochangesinthisclientdata.

3. PROPOSED METHOD

The most often problem while using Cloud and mobile computing is security of personal data stored onthe cloud and handling the multiple client node efficiently without affecting the speed of data transferring fromserver. In case of security, one will alwaysprefer to cloud forstoring his personal data. That data may containaccount numbers, passwords and other personal information. The personal information may get misused byintruder, dark side hackers, etc. While handling multiple clients, the server may slow down and results into lessthroughput.Somain motiveis tohandlemultipleclientalongwith maximumthroughput.so we are using the hadoop cluster and Kerberos.

3.1 MethodologyUsedinProposedSystem

Hadoop

Hadoop is an Apache open source framework written in java that allows distributed processing of largedatasets across clusters of computers using simple programming models. The Hadoop framework applicationworksinanenvironmentthatprovidesdistributedstorageandcomputationacrossclustersofcomputers.Hadoop is designed to scaleup from single server to thousands of machines, each offering local computationand storage.

Atitscore, Hadoop hastwomajorlayersnamely Processing/Computationlayer(MapReduce) and Storagelayer(HadoopDistributedFileSystem).

MapReduce

From past few years, there is an exponential growth and availability of data, both structured andunstructured. Structured data of traditional database to unstructured data of social networking sites, simple datalike text data to complex data like video data are increasing at high rate. More is the data higher is thecomplexityofanalyzingit.MapReduceisaprogrammingframeworkdevelopedbyGooglein2004forprocessingoflargedatasetsacrossdistributedsystems.Basically,MapReduceisusedtosimplifydata,processingacrossmassivedatasets. Itisanabstractiontoorganize parallelizabletasks.

WorkflowofMapReducemainlyoperatesonkey:valuepairsthatisinputissubmittedasajobofkey:valuepair andproducesset ofkey:value pairsasa outputofjob.GenerallythreemainphasesareinvolvedinMapReduce:

I. Map Phase: It take some input data from user and map it to key:value pairs as per specifications provided bytheuser.

II. Shuffle Phase: At this intermediate stage, key: value pairs emitted from mapper are collected. Pairs with

samekeyare grouped togetherandpassed toreducer forfurtherprocessing.
III. Reduce Phase: At this phase, intermediate list ofkey: value pairs are reduced according to userspecified reduce function and produce output of multiset key: value pairs with same key are generated. This is aphasewhereusergetshis/her expectedoutput.

HDFS

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides adistributed file system that is designed to run on commodity hardware. It has many similarities with existingdistributed file systems. However, the differences from other distributed file systems are significant. It is highlyfault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access toapplicationdata andissuitablefor applicationshavinglargedatasets.Apartfromtheabove-mentionedtwocorecomponents,Hadoopframeworkalsoincludesthefollowingtwomodules:

HadoopCommon: ThesareJavalibrariesand utilitiesrequiredbyotherHadoopmodules.

HadoopYARN: Thisisaframeworkforjobschedulingandclusterresourcemanagement.

RSA(Rivest–Shamir–Adleman**) Algorithm**

Using RSA algorithm we divide our generated session key and share it among the nodes and one share isprovided to client. To perform any operation on database client need to provide that share. If all the neededsharesmustbeprovidedcorrectlythenpermissionisgrantedtoclienttoperformoperationondatabase.Ifanyofshar esislostthenoperationhasbeendiscardedbymetadata server.

Kerberos

Kerberos is the standard and most widely used way of implementing the user authentication in the Hadoop cluster. It is the network authentication protocol developed at MIT. Kerberos is designed to provide authentication for client-server application and for that it uses secret key-cryptography.

Kerberos in Hadoop

To implement kerberos security and authentication in Hadoop one need to configure Hadoop to work with Kerberos.

The following steps are used to create a Key Distribution System
1.To start with, a key distribution center (KDC) is creaated for the Hadoop cluster. It is advisable to use a separate KDC for Hadoop which will be exclusive for Hadoop and should not be used by any other application.
2.The second step is to create service principals. We will create separate service principals for each of the Hadoop services i.e. mapreduce, yarn and hdfs.
3.The third step is to create Encrypted Kerberos Keys (Keytabs) for each service principal.
4.The fourth step is to distribute keytabs for service principals to each of the cluster nodes.
5.The fifth is to configure all services to rely on kerberos authentication.

3.2 SystemArchitecture

The system architecture is given in Figure 1. Initially, the client has to register at Kerberos server. After registration, client has to perform loginoperation. For performing operations, valid user has to enter into database with session. At the metadata server,MapReduce framework accepts multiple client requests to register them on server in which, clients' requests aredivided by MapReduce to decrease the load of server. To check the validation of user, divided part of sessionkey for each client will be forward to client as well as to the storage node. To validate user, there is need toconquer this parts of session keys at storage node and kerberos server. If entered user is valid then kerberosserver provides access to the database for file operations such as encryption and decryption. As we usingRSAalgorithm, securityisalso increased.

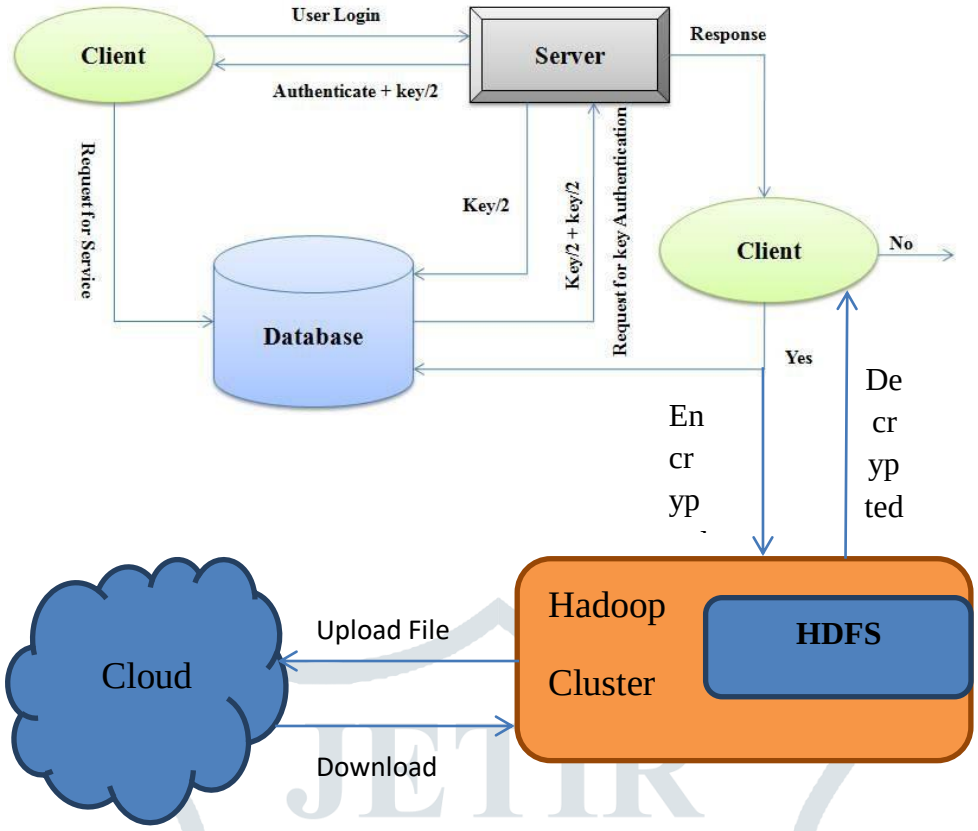

Figure 1: System Architecture

3.3Implementation of Proposed System

The proposed system's implementation involves the following phases.

1. Registration: In registration phase, user details are taken. If user was registered already by usingattributes specified in the registration phase, then that user is discarded from registration. If user was notregistered,thentheuserregistrationisprocessedanddatabaseisupdatedwithgeneration ofsecretekey.
2. Login: In login phase, the user login details are taken from admin. After taking user login details check the validation of user.
3. Split: When user enters key, then this key is divided into "n" shares from which one share is given to clientand anotherisgiventodatabases.
4. Encrypt: Before uploading file, it is convert from plaintext to cipher text using public key cryptographytechnique.
5. Upload:In this phase,useruploadstheencryptedfile.
6. Combine: Asthe keysharesaredistributedamongallstoragenodesandone shareisdistributedtoclient.Toauthenticateanyfileoperation,thereisrequirementtogatherallrequiredsharestoreconstructthekey.
7. Decrypt**:**When downloadingfile,itisconvert fromciphertexttoplaintextusingpublickeycryptographytechnique.

4. RESULT AND DISCUSSION

Hence, proposed system provides security to our personal data with help of hadoop. Hadoop has beenefficientsolutionforcompaniesdealingwiththedatainPetabytes.Accordingtheabovesectionsonecansaythathadoopisone ofthebestwaysto provide the securitytosensitive data and kerbouscan be used for authentication in hadoop cluster. Multiple secret keys, third-party authorization, and cryptography make Kerberos a secure verification protocol. Passwords are not sent over the networks, and secret keys are encrypted, making it difficult for attackers to impersonate users or services. Upload /Download Time are calculated using the following formula. The results are tabulated in Table 1 and Table 2.

File Size In Megabytes / (upload /Download Speed In Megabits / 8) = Time In Seconds

Table 1: Update time for different file size

Upload File Size	Connection Speed	Time
5MB	5Mbps	8 seconds
	8Mbps	5 seconds
	10 Mbps	4 seconds
10MB	5Mbps	16 seconds
	8Mbps	10 seconds
	10 Mbps	8 seconds
25MB	5Mbps	41 seconds
	8Mbps	26 seconds
	10 Mbps	20 seconds
1GB	5Mbps	28m/seconds
	8Mbps	17m/ seconds
	10 Mbps	14m/seconds

Table 2: Download Time for different file size

Download File Size	Connection Speed	Time
25MB	5Mbps	5 seconds
	8Mbps	3.125 seconds
	10 Mbps	2.5 seconds
110MB	5Mbps	22 seconds
	8Mbps	13.75 seconds
	10 Mbps	2.5 seconds
250MB	5Mbps	50 seconds
	8Mbps	31.25 seconds
	10 Mbps	25 seconds
1GB	5Mbps	200 seconds
	8Mbps	125 seconds
	10 Mbps	100 seconds

5. CONCLUSION

In current intervals, Hadoop is the most famous platform for processing big-data, because it consists of the advantages like rapid velocity, low costs and easy comfort. Nowadays, Hadoop is significantly applied in private and government sectors, wherein its security is considered to be a first-rate difficulty. In this research, a brand new authentication system turned into proposed for the clients in order to analyze the information security problems inside the Hadoopsystem. The proposed system applied Shamir's encryption together with Kerberos authentication device for protective the data that stored in HDFS from replay and data attacks. In future, this system can be compared with different protection systems to measure effective performance.

6. REFERENCES

[1]
WenfenLiu,XuexianHu,"*SecureDatasharinginCloudComputingusingRevocableStorageIdentityBasedEncryption*",IEEETransactionsonCloud Computing,VOL.*14*,NO.*8*,AUGUST2015.
[2].
ShaofengZou,StudentMemberIEEE,YingbinLiang,"*AnInformationTheoreticApproachtoSecretSharing*",IEEETransactionsOnInformationTheory,VOL.*61*,NO. *6*,JUNE 2015.
[3].FUXiao,WANGZhi-jian,WUHao,YANGJia-qi,WANGZi-zhao,"How*tosendaSelf-destructingEmail*",IEEEInternationalCongressonBigData, 978-1-4799-5057-7/142014.
[4].
R.C.Dharmik,HemlataDakhore,VaishaliJadhao,"*Sedas:ASelfDestructiveActiveStorageFrameworkforDataPrivacy*",InternationalJournalofScientificEngineeringandResearch,Volume*2*,Issue*3*,March 2014.

[5]. LingfangZeng,ShibinChen,QingsongWei,"SeDas:*ASelf-DestructingDataSystemBasedonActiveStorageFramework*",IEEETransactionsOn Magnetics,VOL.*49*, NO.*6*,JUNE 2013.
[6].
MrudulaVarade,VimlaJethani,"*DistributedMetadataManagementSchemeinHDFS*",InternationalJournalofScientificandResearchPublications,VOL.*3*, NO.*5*,May2013.
[7]. XukaiZou,FabioMaino,ElisaBertino,YanSui,KaiWangandFengLi,"*ANewApproachtoWeightedMulti-SecretSharing*",IEEE,978-1-4577-0638-7/11,2011.
[8]. L.Zeng,Z.Shi,S.Xu,andD.Feng,"*Safevanish:Animproveddataself-destructionforprotectingdataprivacy*",IEEE,978-0-7695-4302-4/10,2010.
[9]. CongWang,QianWang,KuiRen,"*Privacy-PreservingPublicAuditingforDataStorageSecurityinCloudComputing*", IEEE,978-1-4244-5837-0/10,2010.
[10].SeungWooSon,SamuelLang,PhilipCarns,RobertRoss,RajeevThakur,"*EnablingActiveStorageonParallelI/OSoftwareStacks*", IEEE,978-1-4244-7153-9/10,2010.
[11]. YuZHANG,DanFENG,"*AnActiveStorageSystemforHighPerformanceComputing*",IEEE22ndInternationalConferenceonAdvancedInformationNetworkingandApplications,1550-445X/08,2008.