Cloud Storage Monitoring Using File Access Pattern and De-Duplication Techniques

[1]MallegowdaM,[2]ParkaviA,[3]SiniannaAlex,[4]AnitaKanavalli,[5]Thippeswamy M N,[5]ManishV,[5]Mohammad RahilKA,[5]AchyuthKumarK
[1]Asst.Prof.DepartmentComputerScience&Engineering
[2]Assoc.Prof.DepartmentComputerScience&Engineering
[3]Assoc.Prof.DepartmentComputeScience&Engineering(AI&ML)
[4]Prof.and HeadDepartmentAIand DataScience
[5]Prof.and HeadDepartmentCSE (AIand ML),[5]UGstudents,DepartmentofComputerScienceandEngineeringMS RamaiahInstituteofTechnology,Bangalore
AffiliatedtoVTU
Email:mallegowdam@msrit.edu ,parkavi.a@msrit.edu ,sinialex@msrit.edu,anithak@msrit.edumntswamy@msrit.edu ,imvmanish19@gmail.com,mdrahil15md@gmail.comachyuthkumar773@gmail.com

Abstract: Cloud Storage Monitoring is one of the important aspects of cloud computing.Cloud Storage can be monitored by keeping track of the files that are being stored in theclouds necessary to keep track of the files that are being stored in the cloud. Also thereshould be some mechanism through which cloud storage can be optimized. There can beduplicate files in the cloud residing with the same name or a different name. The idea is tohandle the duplicate files with a technique known as De-Duplication. There are two kindsofDe-Duplicationtechniques-FileLevelDe-DuplicationandBlockLevelDe-Duplication.File Level De-Duplication is implemented in the proposed project. This technique is alsoknownassingle-instancestorage(SIS).FileLevelDe-Duplicationworksatthecoreoffilelevel where duplicate files are eliminated. Also file access patterns of the user is analyzedand determine ranking for each of his files. If the files are highly ranked, it means that thefiles are getting used very frequently. So files are ranked and make the files which aregetting accessed very frequently highly available to the users. Furthermore security for thefilesthataregettingaccessedfromthecloudisprovided.Additionallyanoptiontotheuser,where we can select if the files are of normal priority or high priority is provided. Highpriorityfilemeansthefileisverysensitive.Sotheyhavetobestoredandaccessedsecurely.SoinntheproposedprojectanextralayerofsecurityisaddedwheretheuserwillreceiveanOTPto hisemail, wehave toenterthesameto downloadthe sensitive file.

***Keywords*:**CloudStorageMonitoring,De-Duplication,Ranking,FileAccessPattern,Migration

1. Introduction

With the increase in use of Cloud Storage and huge demand of Cloud Storage, it becomesnecessary to provide efficient cloud storage to the client. Most businesses are data-drivenand are aware of the advantages of migrating data to a cloud-storage service, but cloudservicescomewiththeirowndisadvantagesalso.Ourgoalistomakeanapplicationthatisuseful,accessibleandeasy-to-useforbeginnersandexpertsalike,whereusercanstoretheirfiles in the cloud and retrieve them whenever necessary in the efficient manner. Andproposed project aim at efficiently using cloud storage so that users need not pay

forunnecessaryduplicateorlowpriorityfile.CloudStorageMonitoringisoneoftheimportantaspects of our application. Cloud Storage can be monitored by keeping track of the filesthatarebeingstoredinthecloud.Therecanbeduplicatefilesinthecloudresidingwiththesame name or a different name. The idea is to handle the duplicate files with a techniqueknownasDe-Duplication.ThereareturokindsofDe-Duplicationtechniques-
FileLevel

De-Duplication and Block Level De-Duplication. In the proposed methodology File LevelDe-Duplication is implemented. This technique is also known as single-instance storage(SIS). File Level De-Duplication works at the core of file level where duplicate files areeliminated. For example: We know that each npm project comes with a package. Son
file.Whenwecompare100'sofpackage.jsonfiles,wecaneasilyfindatleast20similarfilesinit.Sothereisnopointinstoringtheseduplicatefilesinthecloud,insteadonemainfilecanbestoredandotherscanreferencethismainfiletodownloaditscontents.Soalotofstoragespace is saved. Proposed project also analyze file access patterns of the user and determinerankingforeachofhisfiles.If
thefilesarehighlyranked,itmeansthatthefilesaregettingused very frequently. So application rank these files and make the files which are gettingaccessedvery frequently highly availableto theusers.Additionally proposed projectprovides security for the files that are getting accessed from the cloud. Proposed projectprovide an option to the user, where he can select if the files are of normal priority or highpriority. High priority file means the file is very sensitive. So they have to be stored andaccessedsecurely.Sointheproposedprojecthasaddedanextralayerofsecuritywheretheuser will receive an OTP to his email, he has to enter the same to download the sensitivefile.

2. LiteratureReview

One of the most important features of Cloud Computing is automated scaling. The authorof this paper provides some guessing algorithms for the time series to understand theworkloadandperformonitbasedonalgorithmpredictions.Asaresult,thePredictionsuitewill include a variety of guessing algorithms, each of which will be selected based on theworkload pattern. The goal was to increase the accuracy of the predictive automatedmeasurement system for the clouds IAAS layer. Thus focus was done on improving auto-scaling.Theprecisepreciseprecisionalgorithmwasabletorespondtovariedworkload,whichmettheproject'sgoal.Basedontheincomingworkloadpatternasuitabletime-
seriespredictionmethodcanbeselectedandthepredictiveauto-
scalingsystems'predictionaccuracycanbeimproved. This paper has presented formal study on auto scaling using predictive suites.This model has not been tested in cloud environments that run on large scale. This is thelimitation ofthepredictivesuites model[1].

Data replication is required to make systems fault-tolerant, but placement and replicaselection are significant challenges in Cloud Computing. The suggested research aims tocreate an algorithm for identifying and deploying appropriate ideal duplicates in the cloudto increase data availability. The systems presented were developed on the Eucalyptus.When compared to previous techniques, the authors' experimental results revealed
thatreplicaselectionandplacementtransparentlyplacesdataingeographicregionsandimproves

bandwidthconsumptionanddataaccessperformance.Oneoftheproject'slimitations is the requirement for the replica placement strategy to be extended in a cloudenvironment which shouldbe integrated[2].

Author focuseson Cloud Managementproblems.To address these issues,this studypresents new management strategies. This article discusses failure management, virtualserverissues,autonomicsandcloudresourcemonitoringgroupsetc.Theauthorfocusesonthe issues that Cloud Computing faces and provides theoretical solutions using a variety ofmethodologies.Theauthorhastackledissuessuchasscale,numerouslevelsofabstraction,sustainability,federation,dynamism,andresourcetypes,amongothers.Thisresearch

purely focuses on theory rather than practical. This is the limitation of this research. Butthispaper givesanideatobeginnersabout cloudcomputinganditschallenges [3].

Replication of data, as we all know, is critical for making a system fault tolerant andavoiding network delays. There are other techniques for replicating data, but this studyfocuses on the system's energy efficiency and bandwidth use. The goal is to improve QoS(Quality of Service) while lowering service costs. Power consumption is minimized bydropping the voltage or frequency. The adjustments are intended to reduce on powerutilization. Green Cloud, a simulator focusing on optimized energy and communicationprotocols in cloud computing data centers, is used to evaluate performance, and variousexperiments are undertaken for energy efficiency and reduced power usage. The conceptsmentioned in this research paper are simulated using a Green Cloud simulator. This is nottested in a real world cloud environment and there is no guarantee that it will produce thesameresultsinarealworldcloudenvironment[4].

A physical section is organized up into multiple logical systems when a Cloud Computingsystem is virtualized. Isolation and therefore security are delivered via virtualization.
WemustintegratecertainsecurityfeaturesinMultiUsersystems.Dikeauthorizationarchitecture is proposed in this study. Native access control,, and object-based file systemcompatibility, tenant namespace isolation are included. The Dike system architecture wasput into effect with the help of Ceph, a production-grade file system. They empiricallydemonstrate that multi tenancy overhead can be limited up to 16 percent in configurationswith thousands of cloud occupiers. Also focused and elaborated are various methods ofvirtualization specific requirements, authentication, and file management. The experimentwas a success for CPU intensive applications. But in deep I / O systems to a large extentover object-basedfilesystems, considerationoftrustassumptionswereweak[5].

In Cloud Computing, Cloud Resource Management is crucial. The system is IO dense inthe case of an information system. Cloudlets are a notion akin to servlets that operate asvirtualizedsystemsthatconsistofresourcesbundledtogethertofacilitatecloudmanagement.
The compound framework integrates hierarchical structure and peer-to-peerways of making a structure that not only efficiently regulates but also efficiently sharesinformationwithinthesystem_RCL-basedresourcemanagementcanorganizetask-responding nodes, lowering network traffic and enhancing system efficiency. The systemis evaluated in local cloud settings or simulators that are open source and available on theinternet. In cloud computing, the challenge/limitation is to test it in a multi-cloud context[6].

Intrusion is minimized via Secret Sharing Algorithms in Cloud Platforms. Some data

mustbe shared in secret since it includes critical information and hashes. As a result, It isnecessary to optimally placement of secure data items on the internet. They divide theproblemofsharereplicadistributionintotwoparts:theresidentsettlementproblem,whichassigns a small set of shares to the collections, and the intra-cluster distribution problem,whichdeterminesthenumberandplacementofsharereplicas.Themathematicalcomplexityoftheproblem-solvingalgorithmsetresidentO.(n2).Theintra-clusterdistribution problem is solved using the O (n3) method. These built-in systems have theability to efficiently send data objects across the Internet. The resident set and intra-clusterallocation difficulties were used to tackle the intrusion problem. However, there are nosecure access protocols in use to solve this issue. As a result, future research on this areawillconcentrateonbuildinga safeaccessprotocol[7].

CloudResourceManagementisveryimportantinCloudComputing.Whentherearemultiagents that are involved then we have to do dynamic cloud resource management. MultiAgents will handle each part in resource management in a divide and conquer manner.Cloud Computing Resource Management Model with Multi Agent Module Integrationgives a practical solution to manage cloud resources. Multiple agents and users can behandled by the system. This method has been tried in a cloud setting with a large numberofusers.Thismethodhasyettobeputtothetestwithagrowingnumberofcloudusers[8].

Cloud hook formation is an excellent analog of cloud computing, where the most pressingissueswithexternalservicesaresecurityandprivacyconcerns(i.e.,cloudhook).Thisstudyanalyzes key issues in cloud computing security and privacy that are considered long-termvalue based on known challenges and errors. Investigations look at both quality and pricefactors.Indicatesareasofpublicconcernthatrequirespecialattention,aswellasbasicdataneeded to make informed security decisions. Although the cost of cloud and performancebenefits are emphasized, some basic security issues have been overlooked. A few keytechnological components, such as a federated trust solution, are still in operation, puttingthem at risk of successful deployment. Determining the security of complex computersystemshaslongbeena securityissueforcomputersingeneral[9].

A contemporary System model is used here .Ownership-based encryption, multi-level keymanagement, and data re-encryption are among the most important programs. Each modelisanalyzedandcomparedaccordingtostorageandcommunicationrequirements,assessing the impact of the resource limits on mobile device contexts, including expensive wirelessdatausage,limitedcellularprocessingcapacity,andlimitedbatterylife.Removingexpensive key reproduction and redistribution for all users whenever group membershipchanges is a major advantage of this model. Protects client data confidentially; data in thecloud is always encrypted and cannot be read in its original form by a cloud provider. Nodatacanbeindependentlyspecifiedbyapersonleavingthegroupandtheiraccessterminated, which ensures confidentiality. The administrator should secretly rewrite eachnew data block or attempt to retrieve the record using an asymmetric key, contrary to thisstrategy. Scalar recurrence costs several times more than the bilinear mating process basedonWeilandTate pairing[10].

Thesuggestedstudy'sauthorsexaminetheadvantagesofcontemporaryDCNs.Theauthors'main contributions are: (a) a multi-layer graph model for various DCNs; (b) a comparativeanalysisofancientstrengthmetricstakingvariousfailuresintoaccount;(c)theintroductionof old metrics for network strength metrics to properly assess DCN strength; and (d)

newproceduresformeasuringDCNstrength.Therearepresentlynoextensivestudieson DCN'seffectiveness. The authors of this study assess the topological properties and robustness ofthree current DCNs: (a) ThreeTier (b) FatTree, and (c) DCell. Under various failuresituations,therobustnessofthefollowingnetworksisdiscussed:(a)30Knetworks(VII-B)and (b) 2K networks (VII-C).Despite the fact that the study included nodes that wereaffected by failures in the range of 0.1 percent to 10%, the results suggest that just about6% of the nodes were affected. Because of the higher rates of targeted and network-onlyfailures, several networks were completely shut down. Experiments indicated that, ascompared to the FatTree and ThreeTier architectures, the DCell design degrades smoothlyunder all failure situations. Based on the hierarchical and pattern-based connectivity ofDCN designs new robustness measurements are required. In addition, network traffic andperformancestudyinvariousDCNcomponentfailuresituationsmustbecarriedout.The costofanetworkwithahighlevelofresiliencyishigher.Furthermore,networkperformance characteristics such as bisection bandwidth and bottleneck degree must betakenintoaccountwhencalculatingDCNperformance[11].

Author proposes ATOM, an efficient and effective framework for automatically tracking,monitoring, and organizing resource utilization in a cloud architecture Infrastructure as aService (IaaS) system, in this paper. Author develop a Principal Component Analysis(PCA)-based strategy to constantly monitor and automatically detect anomalies based onapproximated tracking data, and we use an unique tracking method to follow essentialsystemutilizationindicatorswithminimaloverhead.ATOMframeworkinIAASallows VMIto be triggered only when a potential attack is detected, and it also aids in locating therelevant memory region. Author show how to change the tracking algorithm to ensure itsmaximum performance under changing circumstances. Author has employed introspectiontoolstodomemoryforensicsonVMs whenprobableanomaliesarediscovered.Thisisnotan "active detection and reaction" system. Continued development on ATOM could includeupgradingittosupportcomplexorchestrationofresourcesandcombiningresistanceagainstevenmore complex threats [12].

Tuba was built using Microsoft Azure Storage (MAS), which provides the same API forreadingandwritingblobs.Tubaimprovesread-onlystatisticaldataby63percent,accordingtoauthors'surveys.Accordingtotheresultsofaclient-widesurveyconducteddondatacentersaroundtheworld,redevelopmenteverytwohoursincreasestheshareofstudiesthat confirm the characterization of character units from 33% to 54%. This section showsthat the default setting will provide accurate and reliable results. Making these moments ofsyncing or adding a copy to the site will not affect customers who are not close to the site.Theauthorsconductedexperimentswithclientsavailableatvariousdatacentersaroundtheworld to prove that redesigning Tuba over a two-hour period improved statistical dataconsistency by 63 percent and improved intermediate usage by 18 percent. These findingssuggest that flexible restructuring can provide appropriate benefits that can be achieved inthe real world. Most systems will accept these results as the list of cloud storage sites (i.e.,datacenters where data can be stored) appears to be limited and redesign is

possible.Algorithmstoimprovepruningforsearchspacecangetimportantresultsforprogramswi tha large number of clients and a large list of potential management sites, as well as otherstrategiessuchas ILPorconstrainingprogrammingcan be used[13].

Authorsintroduceasimpleandeffectivesingleserverapproachtoqueryencryptedwebsites on reliable servers for this task. Author's method is based on the use of theinformation index benefit encrypted on the encrypted website, which is used by the serverto select relevant data that will be retrieved from the query result without exposing thecontenttotheinternalwebsite.Authors'guideswillachieveabalancebetweentheeffectivenes s of the query and some security to avoid the assumptions that may causeconflict when identifying the data. The main advantage of this approach is that the datawithin the nodes of the B+-tree is not visible to the unauthorized DBMS users. Thedisadvantage of this approach, is that the B+-tree traversal can now only be done by theauthorizedusers [14].

TheBBOtechniqueisimplementedtoallocatedataduringtheimplementationofdistributed database systems. The obtained output of this optimized technique for the dataallocation are decided along with the results of a normal genetic technique to run theperformanceofaparticularmethod.Fromthesetechniquesitisobtainedthattheminimumcost used by appropriate optimized BBO algorithm for allocation of data is decreased thanthatbyleastcostachievedbyGenetictechnique.Inalltheaboveprocesses,BBOtechniquefor data replicas allocation is providing data allocation rules which is good than Genetictechniquebasedonthedatacopiesallocation.Inthesecasestheaveragecostofdatareplicas allocation using BBO technique is greater than Genetic technique which shows in thecompressingthedataofGAandBBO.Inhugenumberofthecasesmeringingrateof BBOalgorit hmis largerthanthatof GAalgorithm[15].

Inthispaperauthorshavedealtwithalltheaspectsofcloudcomputingsystem.Whatitis?How does it works? How it is developed? How it is implemented and many more. In thispaperdevelopershavediscussedthecloudcomputingsystemandalsoitsfeatures,assurance and its optimizations. It also includes with what regarding the cloud computingis?Howdoesitimplemented?Therepresentationofcloudstoragemanagementisproc essed by a huge variables. It limits the cloud processing to a number of parameters.There is always the consequence of leakage of the data produced on it. Hence, developersareconstantlyworkingforother ways togivetheprotectiontotheirdata[16].

Initially the processing of password authentication in shadow files is considered as a"hybrid" model where the client memorises the server's public key and also to store apassword into the database. In this setting all the 4 persons they are: Gong, Lomas,Needham, and Seltzer are the starting persons initiated to represent authentication rulesalong with heuristic security to take offline dictionary attacks. Henceforth in huge part
ofthepracticalsignificanceandimportanceofpasswordauthentication,activatestheresearchcon tinues in this area. We only represented those outputs most related to what is displayedhere.The significant limitation of this technique is particular variants of different rules,wherepasswordswillbesharedamongthehugenumberofdeveloperstogivetheappropriate resultsagainstappropriateserver[17].

The entrepreneurs invest lot of money in order to protect the data somewhere else. Hence,an efficient method is needed for handling such a large amount of data. There are twotechniquesforreducingtheduplicatecopyinthestoragesystemtheyaredatadeduplicationan

d data reduction. It is observed and concluded that inconsistent sized data reduplicationis done and good when decided along with other techniques by differentiating the hash ofevery data replicas. Data warehousing reaches up huge amount of memory in tera, petabytes of data. Huge amount of the data memory consists the data got from parent data, sothere is a paradigm, a huge of redundant data shall be contained within the data storagemanagement system [18].

Cloud system is a new modern technique jardon shift in modern technology. Data copieswill reduce the memory storage and decreases the amount of wavelength for files transferfromonesideofmemorytoothersideofmemory.Hugetechniqueshavebeenimplemente dto files duplication based on the file level and block with print or hashing methods. In thispaper as per results, we are tracking the secure replication on cloud system data on non-variable block and data level replication with a new technique for sending the data tostorage. Client data transfer technique for data replication technique will be analyzed forthe new techniques files transfer. Large design of techniques in the present situation arebased on the commonly used hashing techniques, cloud files transfer along with compress& security with increasing execution time which effects the rate of throughput in moderndatabackup[19].

Inreallife,thecloudsystemisanintegralpartofthecloudbuilder.TheentirecurrentCloudMemory systemworksadmirablyonbigdata,butseemstobeunabletomanagethestorageof small files. Large and small files are processed as the deployment of cloud services iscompleted.

In fact, large-scale science applications are cloud-based applications. Cloud memory is animportantpartofthecloudsystem.Smallfiles,ontheotherhand,arerarelyusedincurrentclouda pplications.Thisdocumentsplitssmallfilesintofourtypes,eachofwhichispackedintoalargerfile basedonitstype.Asaresult,ithasahighereffectthantheothertwowaysofcomparingdata.Thereco mmended approachhasnotbeenemployedinausedinalargeamount cloudsystembecause ofthelimitedobtainedresults [20].

The most prominent topic in data processing is data duplication. As always, launchingprojects in the cloud system results in lower efficiency and profitability. It will work
besttoincreasedataperformance,improvesystemdataifthenumberofcopiesisincreasedanddistr ibuted in various locations. This study uses the Cloud Simulator method to performtests on various replica collections and cloud locations. The results obtained for the MO-PSO and MO-ACO methods are compared with most known methods, according to thecomparativeresultsobtainedfortheMO-PSOandMO-ACOmethods.Thereal-timecloudlocation will be used to test the cloud formation created in the latest phases. Now we lookat two concerns about copies of files and their location. The consistency and energythroughputofcomplicateddatareplicationwereonceaconcern.Inanycase,themostrecent swarmtechniquefordata replicationincloudsystems willbe used.[21].

The amount of data obtained by computers and memory devices for two users is growingrapidly in everyday life. It is much simpler, and will always have a data security test,reducing the number of important data providers. As a result, the end result is that the datathat has already been exchanged with the client is secure. Many cloud developers, as wellas cloud services such as Dropbox, Google Drive, and other Google products, use dataextraction. It is not possible to duplicate encrypted data because it is generated randomly.Duplicatingdataisawayofcompressingdataintoasinglefile.Asmentionedearlier,usi

ngmethods to remove duplicate of source data will take a lot of time, but it is very importantas a great way to get resources and optimize data from a specific data destination toeliminate recycling data, as it can provideimproved results when dealing with largeamounts of data. As always, it requires the installation of computer hardware in existingenvironments,whichisaduplicatecomparedtoduplicationremovalnotes.Anotherdisadvantage is that although data must be transmitted across the network, there is no suchthingasbandwidthinthis technology [22].

Toaddresstheabove-mentionedissues,wepresentthemostappropriateandconsistentdatamanagement methods in this study. Access limits and strategies embedded in the proposedprotocol are analyzed. It also has a flexible cloud management system that allows the datamanager or a specific third-party user to set file duplication and access restrictions. Thepapersuggestsawaytomanagethevariousfilesinthisproject,providinganadequatewayto monetize the cloud system and access control. As data stored in memory is encrypted,our approach provides data security for cloud users. Another first option to use privacy isto use random in the Key Generation (KGC) system, where your real identity is linked torandomvariables, whichare processed byKGCand verified [23].

Huge research papers and programming paradigm deals on the processor scheduling,relatingtoserviceprocessing,aspertheresourcemanagementifitdidnotsucceedin

specific handling of provision of service and communicating between services in-terms ofVirtualMachineprocessinganditsallocation.
Theobtainedenhancements&itsprocessing'sof these calls for more knowledge , manages the scheduler that would have data about thecapacity of required host and obtained a huge number of Virtual Machines based on theservice request that can be created. In this paper, we have noticed all the appropriate waysofenhancingalltheallotmentofservicesincloudsystem.Manyresearchservicesprovidesenhancedtechniqueforobtainingtheresourcesincloudstoragemanagement [24].

Deduplication incurs a significant overhead in storing metadata, such as the fingerprintindexandfilerecipes.
Asthismetadataislarge,itmustbesavedondisc,causingsignificantreaddelayinDeduplicatedCloudStorage(DCS).Whencomparedtoearlierinvestigations,thesuggestedprefetchingstrategyimprovesthecachehitrateby140percentandincreasesthe reading rate by 88 percent. Prefetching the fingerprints associated with the files andstoring them in a cache for future access is advantageous. As a result,theFrequent-Patterns-Based Prefetching Technique (FPBPT) was suggested with the goal of improving readperformancefornon-backupworkloadswithalargenumberofdissimilarfiles. Becausethemaximum number of expected fingerprints changes over time. In order to maximize thepotential for read speed, it is necessary to build a cache repository of the right size that canaccommodateacompleteset ofpredictivefingerprintsor asmall setofthem [25].

3. ProposedSystem/Methodology

System Design: TheapplicationisdesignedusingClientServerArchitecture.

Abrowserwhich acts as a client. A nodeJS server acts as the server. The client makes requests to theserver and gets the response accordingly. We have chosen Client-Server Architecture andCall and Return Architecture for our project. This Project uses Client-Server Architecturesince Web works on Client-Server Architecture. We have a HTTP Client (Browser) andHTTPServer(WebServer)whicharecommunicatingtogether. Wecanrequestanyresource and get it from the Web Server using this protocol. Inside the Server we are usingCall and Return architecture since the CSM calls the services it needs whenever required.Proposed application has a main backend server which handles all other entities in theprogram

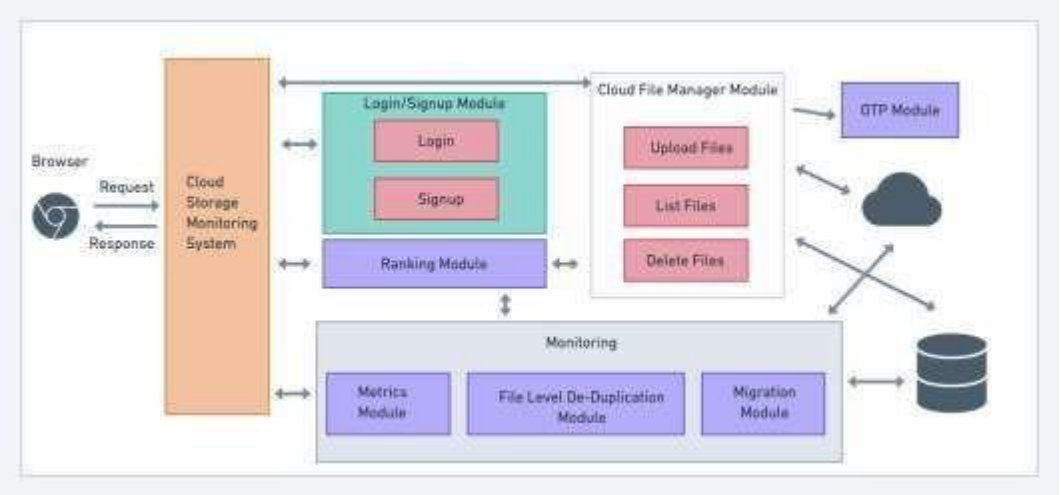

Fig1:ArchitectureDiagram

LowLevelDesign:

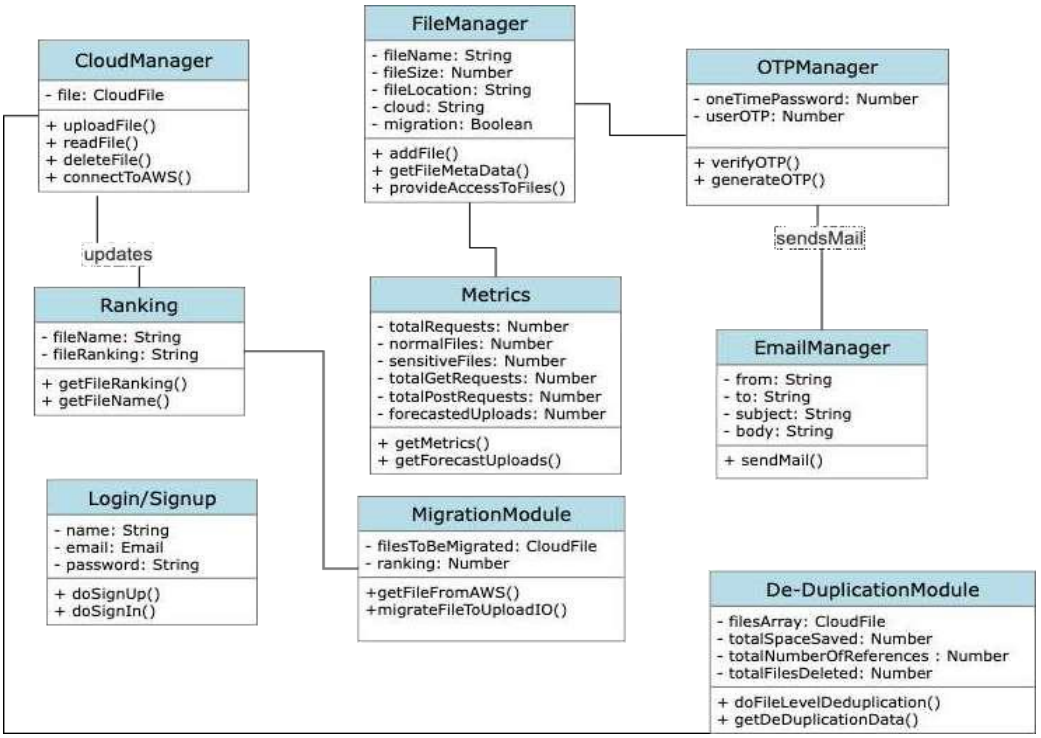

Fig2:Class Diagram

Algorithms:

- ### FileLevelDe-Duplication:

  File Level De-Duplication workson optimizing the storage space.Instead ofstoring various copies of the same file. The idea is to store only one copy of thesamefile and makeothersreference to it. So like thisstoragespacecan beoptimized. We have optimized the File Level De-Duplication algorithm by usingfile'smetadata.

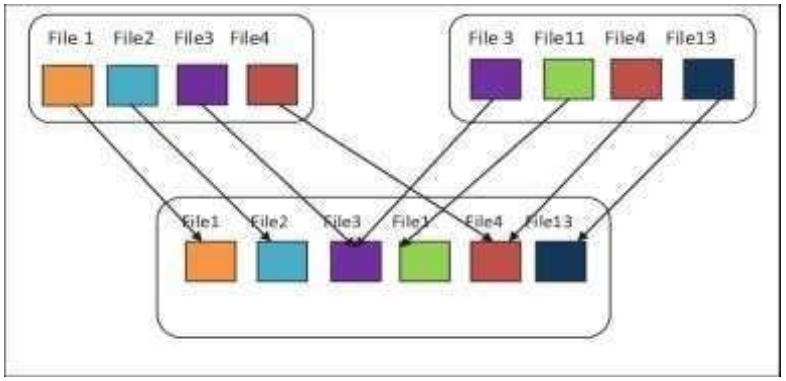

Fig3:FileLevel Deduplication

- **LinearRegression:**A variable's value can be predicted using linearregression analysis based on the value of another variable. The dependentvariableistheone youwanttobeabletoforecast.Theindependentvariableis the one you're using to make a prediction about the value of the othervariable. As we train the provided model: input training data (one inputvariable,orparameter),x,Labelstodata,y(supervisedlearning).Themodelf its the best line to predict the value of y for a given value of x duringtraining. By locating the best 1 and 2 values, the model produces the bestregression fit line. The best fit line is obtained after determining the best 1and 2 values. Therefore, when we ultimately use our model to make aprediction, it willforecastthevalueofybased onthevalue oftheinput x.

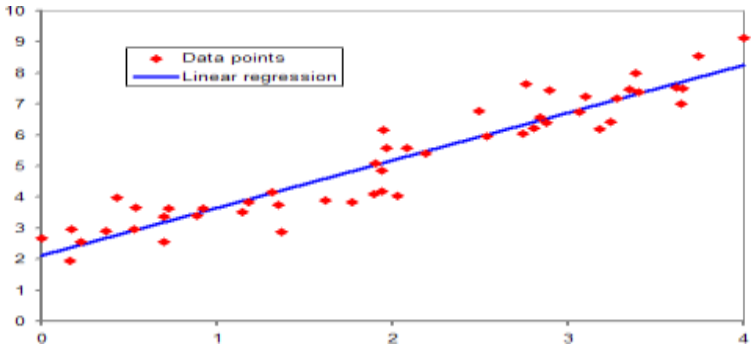

Fig4:LinearRegression

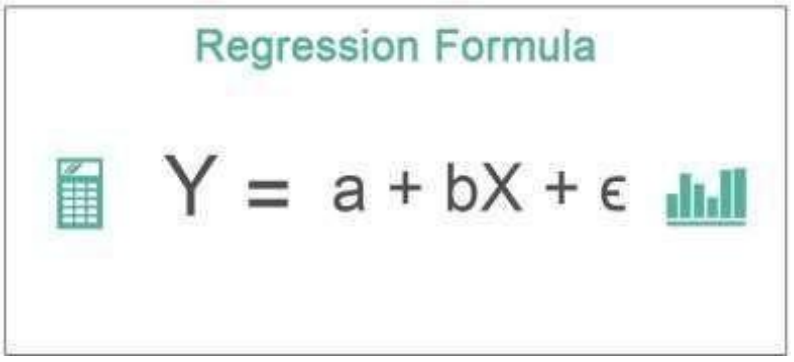

Fig5:Linear RegressionEquation

- **Frequency/Counter Based Ranking Algorithm**: To compute the rankingof algorithms we have used the frequency of the files. We keep track of thenumber of times each file is accessed and handle. Increase the frequency ofthe file whenever it is accessed. This frequency ranking is used in theMigration module also where we have set a threshold value of ranking onlythenthe migration can happen.

4. Results and

DiscussionThe total time complexity of for de-duplication is as follows,(1)Worstcase:T(n)=O(nlogn) +O(x)[ComparisonTime]

(2) AverageCase:T(n)=O(nlogn)+O(x)[AverageComparisonTime]

(3) BestCase: T(n)=O(nlogn)

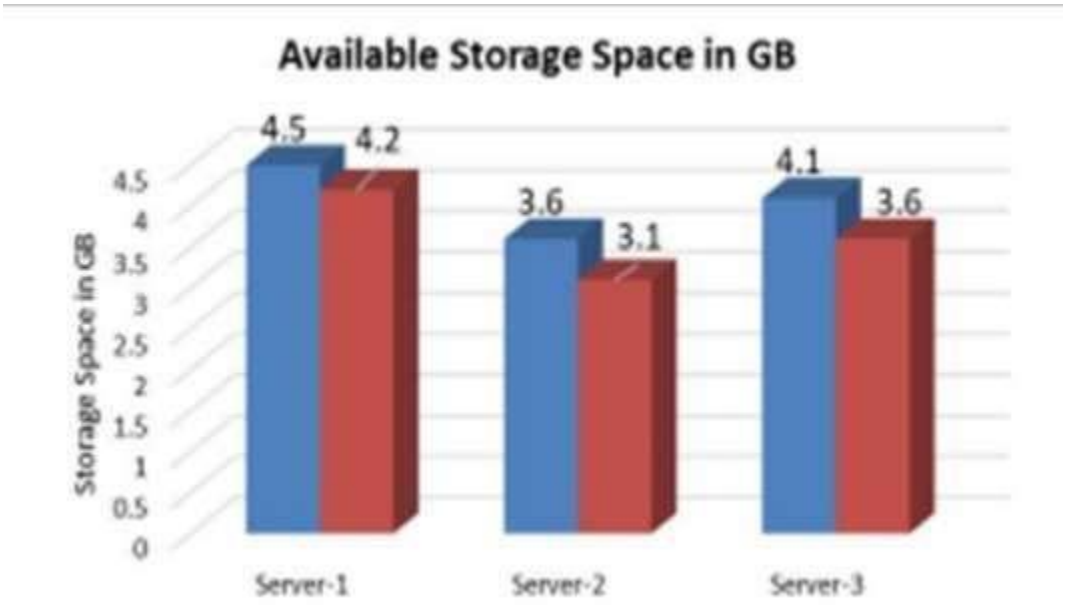

Fig6:Comparisonofstorage WithandwithoutDe-Duplication

Thefig6referstostoragespacesavedusingDe-Duplication.AfteruploadingcoupleofduplicatefilesandtestingthededuplicationfeaturemanuallyitwasobservedthatDeDuplicationfeaturesavesconsiderableamountofspace.Whereredcolourdenotesthestoragespacesaved.

Here n - number of files which are being compared, basically files are comparedwhen both files size are same and then only "x" factor will be considered. x - extratimetaken byalgorithm to comparethecontent offiles.

Thetotaltimecomplexityofcounterbasedrankingalgorithmusedintheapplicationis as follows:

O(1)Perfile, Ifthere arenfilesthis algorithmgivesthecomplexityofO(n).

For the migration of high ranked files into the CDN's,their is an rank set which isreferredas thresholdrank.

Authentication of the priority files would give an complexity of O(1),which can beignored.

The Linear Regression algorithm was used to forecast the uploads, the results ofsomeuploads areshownbelow,

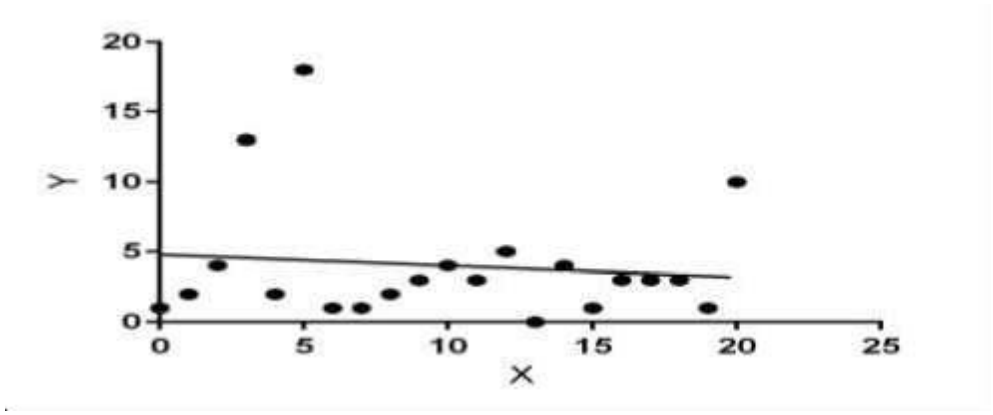

Fig7:LinearRegressionExperiment Graph

3. ConclusionandFutureScope

It's important to understand various techniques that can be used to monitor cloud storage.We have used File-Level De-Duplication and Ranking algorithms to monitor the cloudstorage. We did a lot of research and analysis to choose those algorithms. The goal of theresearch is to understand, choose and improve the techniques used for cloud storagemonitoring. The choice of Linear Regression to predict next day uploads was also part oftheresearch.Thetwomostimportantattributesofcloudstoragearepriceandspeed.Cloudstorage providers will need to focus on driving down costs and making storage faster toremaincompetitive,allwhilebeingclosertotheircustomers.Vendorscanhelpdrivedowncosts and increase agility for their customers by removing large egress fees, implementingdatasecurityframeworkandeliminatingstoragetiersforfastandeasyaccess.Overthenextcouple of years, we believe we'll also see more cloud storage moving toward the edge ofthe network so they can be closer to their customers, regardless of their location. Also acritical problem of Vendor Lock-in can be solved by further appropriate research onMigration.

4. References

[1] A. Y. Nikravesh, S. A. Ajila and C. -H. Lung, "Towards an Autonomic Auto-scaling PredictionSystem for Cloud Resource Provisioning," 2015 IEEE/ACM 10th International Symposium onSoftwareEngineeringforAdaptiveandSelf-ManagingSystems,2015,pp.35-45,doi:10.1109/SEAMS.2015.22.

[2] A. Rajalakshmi, D. Vijayakumar and K. G. Srinivasagan, "An improved dynamic data replicaselection and placement in cloud," 2014 International Conference on Recent Trends in

InformationTechnology, pp. 1-6, doi: 10.1109/ICRTIT.2014.6996180.

[3] T.Forell,D.MilojicicandV.Talwar,"CloudManagement:ChallengesandOpportunities,"IEEEIntern ational Symposium on Parallel and Distributed Processing Workshops and Phd Forum, 2011,pp.881-889,doi:10.1109/IPDPS.2011.233.

[4] D. Boru, D. Kliazovich, F. Granelli, P. Bouvry and A. Y. Zomaya, "Energy-efficient datareplicationincloudcomputingdatacenters,"2013IEEEGlobecomWorkshops(GCWkshps),2013,p p.446-451,doi:10.1109/GLOCOMW.2013.6825028.

[5] G. Kappes, A. Hatzieleftheriou and S. V. Anastasiadis, "Virtualization-aware access control formultitenantfilesystems,"201430thSymposiumonMassStorageSystemsandTechnologies(MSST),2 014,pp.1-6, doi:10.1109/MSST.2014.6855543.

[6] H. Xingye,L.XinmingandL. Yinpeng,"Researchon ResourceManagementforCloudComputing Based Information System,"2010 International Conference on Computational andInformationSciences, 2010,pp.491-494,doi:10.1109/ICCIS.2010.127.

[7] ManghuiTu, Peng Li, Qingkai Ma, I-Ling Yen and F. B. Bastani, "On the optimal placement ofsecure data objects over Internet," 19th IEEE International Parallel and Distributed ProcessingSymposium, 2005, pp. 10 pp.-, doi:10.1109/IPDPS.2005.306.

[8] Q. Chen, "Research on Cloud Computing Resource Management Model Based on Multi-AgentSystem," 2016 12th International Conference on Computational Intelligence and Security (CIS),2016,pp.378-381,doi: 10.1109/CIS.2016.0092.

[9] W. A. Jansen, "Cloud Hooks: Security and Privacy Issues in Cloud Computing," 2011 44thHawaiiInternationalConferenceonSystemSciences,2011,pp.1-10,doi:10.1109/HICSS.2011.103.

[10] TowardsSecureCommunicationforHighlyScalableMobileApplicationsinCloudComputingSyste ms Piotr K. Tysowski and M. Anwarul Hasan Dept. of Electrical & Computer EngineeringUniversityofWaterlooWaterloo, Ontario, Canada.

[11] K.Bilal,M.Manzano,S.U.Khan,E.Calle,K.LiandA.Y.Zomaya,"OntheCharacterizationoftheStru cturalRobustnessofDataCenterNetworks,"inIEEETransactionsonCloudComputing,vol.1, no. 1,pp.1-1,Jan.-June 2013, doi:10.1109/TCC.2013.6.

[12] M.DuandF.Li,"ATOM:EfficientTracking,Monitoring,andOrchestrationofCloudResources,"inI EEETransactionsonParallelandDistributedSystems,vol.28,no.8,pp.2172-2189,1 Aug.2017, doi:10.1109/TPDS.2017.2652467.

[13] MasoudSaeidaArdekaniandDouglasB.Terry.2014.Aself-configurablegeo-replicatedcloudstorage system. In Proceedings of the 11th USENIX conference on Operating Systems Design andImplementation(OSDI'14). USENIX Association, USA, 367–381.

[14] ErnestoDamiani,S.DeCapitaniVimercati,SushilJajodia,StefanoParaboschi,andPierangelaSamar ati.2003.Balancingconfidentialityandefficiencyinuntrustedrelational DBMSs.InProceedings of the 10th ACM conference on Computer and communications security (CCS '03).Associationfor ComputingMachinery, NewYork, NY,USA, 93–102.

[15] Singh,Arjan&Kahlon,Karanjeet.(2009).Non-replicatedDynamicDataAllocationinDistributed Database Systems.

[16] A. S. Al-Ahmad and H. Kahtan, "Cloud Computing Review: Features And Issues," 2018International Conference on Smart Computing and Electronic Enterprise (ICSCEE), 2018, pp. 1-5,doi:10.1109/ICSCEE.2018.8538387.

[17] Katz, Jonathan &Ostrovsky, Rafail& Yung, Moti. (2009). Efficient and secure authenticatedkeyexchange usingweakpasswords. J.ACM. 57.10.1145/1613676.1613679.

[18] E. Manogar and S. Abirami, "A study on data deduplication techniques for optimized storage,"2014 Sixth International Conference on Advanced Computing (ICoAC), 2014, pp. 161-166, doi:10.1109/ICoAC.2014.7229702.

[19] Suresh, L. &Bharathi, M.. (2019). Analysis of Block-Level Data Deduplication on CloudStorage. 10.1007/978-981-13-5934-7_36.

[20] Fu,Xiong&Liu,Wenjie&Cang,Yeliang&Gong,Xiaojie&Deng,Song.(2016). OptimizedDataReplicationforSmallFilesinCloudStorageSystems.MathematicalProblemsinEngineering.2016.1-8. 10.1155/2016/4837894.

[21] A. Awad, R. Salem, H. Abdelkader and M. A. Salam, "A Novel Intelligent Approach forDynamic Data Replication in Cloud Environment," in IEEE Access, vol. 9, pp. 40240-40254, 2021,doi:10.1109/ACCESS.2021.3064917.

[22] K. A. Tayade and G. S. Malande, "Survey paper on a secure and authorized de-duplicationschemeusinghybridcloudapproachformultimediadata,"2017InternationalConferenceonEnergy,Communication,DataAnalyticsandSoftComputing(ICECDS),2017,pp.2966-2969,doi:10.1109/ICECDS.2017.8389999.

[23] Z. Yan, L. Zhang, W. DING and Q. Zheng, "Heterogeneous Data Storage Management withDeduplication in Cloud Computing," in IEEE Transactions on Big Data, vol. 5, no. 3, pp. 393-407,1 Sept. 2019,doi: 10.1109/TBDATA.2017.2701352.

[24] E. Oppong, S. Khaddaj and H. E. Elasriss, "Cloud Computing: Resource Management andServiceAllocation,"201312thInternationalSymposiumonDistributedComputingandApplicationstoBusiness,Engineering&Science,2013,pp.142-145,doi:10.1109/DCABES.2013.33.

[25] Balasundaram, Prabavathy&Babu, Chitra& M, Subha. (2016). Improving Read ThroughputofDeduplicatedCloudStorageusingFrequentPattern-BasedPrefetchingTechnique.TheComputerJournal. 60.10.1093/comjnl/bxw013.