RESEARCH ARTICLE OPEN ACCESS

Data Security in Cloud Computing Using HDFS

Navjot Sekhon [1], Richa Mahajan [2]
Department of Computer Science and Engineering
S.S.C.E.T Badhani
Pathankot - India

ABSTRACT

Cloud Computing system provided at technology that interconnect between customers/users and providers. It contain levels are cloud user, cloud service provider and cloud provider. Cloud make privacy issues in data storage in multi-location and services. We have provide a new technique in the Cloud .So result, more users to step into cloud. In this paper we discuss about Hadoop Distributed File System to increase security services and storage services.

Keywords: — Cloud Computing Storage, HDFS, Security.

I. INTRODUCTION

Cloud computing is a model for convenient, on-demand network access to a shared pool. The configurable computing resources are servers, storage, network, applications and services .It can be rapid provision and released with minimal manage effort and service provider interaction.

The cloud computing design to establish a cloud strategy, implementation plan to get there , strategy , transformation for cloud adoption , Infrastructure strategy and design services.

The cloud computing deploy to build cloud service in the enterprise as a cloud services provider.It is used in business systems , cloud management and infrastructure.

The cloud storage is a technology that is used in most of organisations and bussiness. The software like cluster application, distributed file system and grid computing.

Hadoop is a distribute file system. Computing Power in hadoop distribute computing model for big data fast.Hadoop is general design to be use on low cost hardware to store large quantities of data.

It provide high through put access and suitable for large data sets.We can store a much data as you want and how to uses it later.Scalability easy grow your system to handle more data simply by adding nodes.

II. HDFS (HADOOP DISTRIBUTED FILE SYSTEM)

Hadoop Distributed File System is an distributed file system design to run on one or more hardware. HDFS is a high fault tolerant. HDFS provide high through put access to data and suitable for users that have large data sets. HDFS is part of the Apache Hadoop projects.

Google File System (GFS) is based on HDFS .It provides a distributed file system.The design to run a large clusters on thousands of computers of small computer machines so to provide reliable, faulttolerant way.

Job Tracker (master node) is receive user's job,how many tasks will run and where to run each mapper.

Task Tracker (slave node) is receive the task from Job Tracker , runs the task until complete and communication with the Job Tracker report progress.

A. Master Node and Slave Node

It is consist of single Namenode. A master server manages the file system namespace and regulate access files by client. There is a number of DataNode usually one per node in a cluster .The DataNodes manage storage attach to the nodes that are running.It expose a file system namespace and allow user data to be stored in files.A file is split into one or more block and set of blocks are stored in DataNodes. In DataNodes is serves read, write requests, perform block creation, deletion and replicate upon instruction from Namenode.

B. Tolerance of Faults

A HDFS consist of thousands of server machines, each storing part of the file system data. In case detection of faults , it quick and automatically recover from them is a core architectural goal of HDFS.

C. Namespace in Files System

In file system with directories and files. It create, remove, move, rename etc. Namenode maintains the file system .In

any meta information change to the file system record in the Namenode.An application can specify the number of replicas of the file needed to replication factor of the file. This information is stored in the Namenode.

D. Replication of Data

In HDFS design store very large files across machines in a large cluster.Each file is a sequence of blocks.In all block in the file except last are of the same size.The blocks are replicate for fault tolerance.Block size and replicas are configurable per file.The block report contain all blocks on a Datanode .

E. Placement Replication

The replica placement distinguishes HDFS from other distributed file system.The goal is to improve reliable, available and network bandwidth utilization.In many racks, communicate between rack are through switch.

F. Selected Replication

It tries to minimizes the bandwidth consumption and latency.If there is a replica on the reader node then that is prefer.It may span multiple data centre replica in the local data centre is preferred over the remote one.

G. Startup on Safe Mode

In start up Namenode enter safemode .The Namenode verifies that each block has accept number of replicas.A configurable percentage of safely replicate blocks check in with the Namenode , Namenode exits Safemode.

G. Metadata on File system

The creating a new file ,change replicate factor of a file , edit log is stored in the Namenode's local filesystem.The entire filesystem namespace include mapped of block to file and file system properties is stored in a file FsImage. Store in Namenode local filesystem.

H. Name Node and Data Node

A Namenode starts up it gets the FsImage and Edit log from its local file system, update FsImage with EditLog information so then stores a copy of the FsImage on the file system as a checkpoint.Periodic checkpointing is done. So that the system can recover back to the last check pointed state in case of a crash.

A Datanode store data in files in its local file system. Datanode has no knowledge about HDFS filesystem.It stores each block of HDFS data in a separate file.When the filesystem starts up, it generates a list of all HDFS blocks and send report to Namenode.

J. Staging

When data reach HDFS block size then the client contacts the Namenode. Namenode inserts the filename into its hierarchy and allocate a data block for it.In Namenode responds to the client with the identity of the Datanode and the destination of the replicas Datanode for the block,then the client flushes it from its local memory.The client sends a message that the file is closed. The Namenode proceeds to commit the file for creation operation into the persistent store.In case, the Namenode dies before file is closed and the file is lost.This client side caching is requires to avoid network congestion and also it has precedence is Andrew file system.

H. Pipelining Replication

The client receive response from Namenode,it flushes its block in small pieces (4K) to the first replica, so turn copies to the next replica.The data is pipeline from Datanode to the next.

I. Reclamation of Space

In case a file is deleted by a client, then HDFS renames file to a file in be the trash directory for a within amount of time.A client can request for an undelete in this allow time.In the specified time the file is deleted and the space is reclaimed.

III. SECURITY

The protect from unforeseen actions that cause it to stop function. Unforeseen action can be either intentional or unintentional.

The issues like identity, access rights, safeguarding data developing and enforcing security policies.The potential to be more secure.

Security Test ensure system and applications in an organization. They are free from any loopholes so may not be cause a big loss. Security test system is about finding all possible loopholes and weakness of the system that may result into loss of important information at the hand of the employees or outsiders of the organization

The tester goal is to identify the threat in the system and measure its potential vulnerabilities. It helps in detect all possible security risk in the systems so it help developer in fixing these problem through coding.

In security test is important testing for user . To check whether confidential data stays confidential. In this ,tester plays a role of the attacker and play around the system. So to know security related errors and find the solution to them .

Security in test plan is test data related to security test, test tools required for security testing, analysis on various test outputs from different security tools. The password should be in encrypted format and System should not allow invalid users.This is the most important to make security issues to be resolved.

IV. IMPLEMENTATION

Cloud based services are very effective in terms of delivering hardware independence and infrastructure maintenance for clients . The hardware and softwareupgrade helps in providing a great deal of freedom for clients to proceed with their work without worry about maintenance and downtime for their websites . Also , cloud based storage system helps in unlimited storage without having to bear maintenance and hardware addition overhead .

There are many facilities but there are few cons of the cloud based services . Data of multiple organisations and individuals reside with same hardware infra . In this scenario security of data relies only on software validation , authentication and authorisation . Security of data is a big concern in cloud system as sensitives information stored by the organisation and individuals is able to break into the security of the system he can get access to the data within the organization or other organizations as well . There are a number of security issues that should be resolved for making cloud secure.

Cloud based services have a great scope in current times. Cloud systems allows potential of scaling hardware infrastructure without actually shutting down the servers and adding hardware. Infrastructure as a service opens a way to be free from cumbersome process of hardware and server management but as cloud systems allow data from multiple users / organizations reside side by side it poses a threat to confidentiality and integrity of data.

Thus, everyone has to be sure the data is securely transferred and only authorized users have access to the service. We propose authorization as a technique to avoid unauthorized data access. Authorization ensure that private or public cloud is not accessed from outside the system without authentication. Authorization process uses a user / password combination to ensure data integrity.

We will accomplish objectives in following steps:

First Step - Implement cloud system.
Second Step - Implement Hadoop in cloud system.
Third Step - Evaluate Hadoop services for data security.
Fourth Step - Evaluate Hadoop service reliability using authorization
Fifth Step - To compare the reliability of existing and enhanced system
Sixth Step - Present result analysis

Session test is a software test method.The combination of accountable and exploratory test to provide rapid defect discovery, control management and metrics report.It contain the following parameters are number of session complete,number of problem found,function areas cover, percentage of session time spent setting up for test, percentage of session time spent test,percentage of session time spent investigated problem.

In graph of figure 1 and figure 2 shows the enhanced work .In figure 1 indicate the performance comparison and figure 2 indicate the cloud data transfer

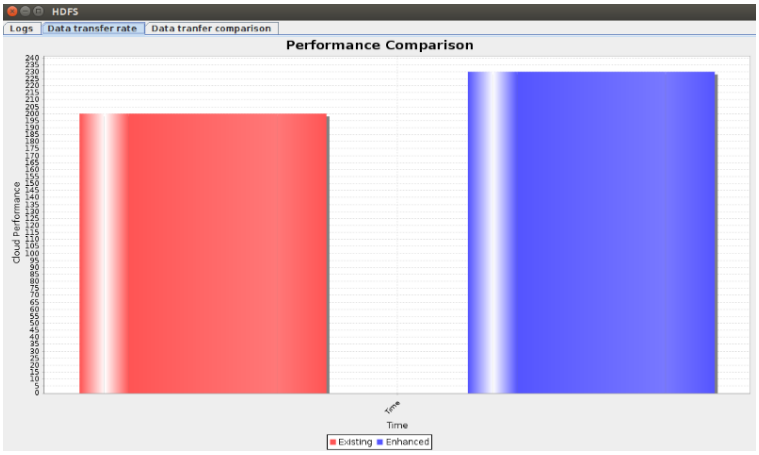

Figure 1 - Performance Comparison

Figure 2- Cloud Data Transfer

V. CONCLUSION

The security solution on the HDFS with master node and slave node architecture.The performance analysis proves the more secure and reliable for the security. In future, we want to make our security services flexible with other cloud storage software systems.Many integrated company uses open source for the massive computing power. Hadoop is truly access easy to work with application and incorporate in to data flows.Hadoop applications to be truly accessible to your organization in overall data flows.Hadoop provide parallelism provides data process across many nodes in a compute cluster and speed up large computing. The Hadoop is used for large data processing tasks.The distributed file system to cheaper, reliable replicate chunks of data to nodes in the cluster.The make availability of data locally on the machine processing it.

FUTURE WORK

Cloud computing is emerging technology of information and communication.The paper goal is to find out security issues and privacy concerns (from both the user and provider side) that may arise with this new approach.The survey results,there are several security and privacy related issues that must be resolved in order to get the real benefits of this new approach.

REFERENCES

[1] Minqi Zhou†, Rong Zhang,et. "Security and Privacy in Cloud Computing: A Survey", 2010 Sixth International Conference on Semantics, Knowledge and Grids

[2] Ei Ei Mon, Thinn Thu Naing"THE PRIVACY-AWARE ACCESS CONTROL SYSTEM USING ATTRIBUTE-AND ROLE-BASEDACCESS CONTROL IN PRIVATE CLOUD"978-1-61284-159-5/11/$26.00 ©2011IEEE

[3] Wen-Hwa Liao,Shuo-Chun Su"A Dynamic VPN Architecture for Private Cloud Computing", 2011 Fourth IEEE International Confer ence on Utility and Cloud Computing

[4] Susmita Horrow, Sanchika Gupta,et. "Secure Private Cloud Architecture for Mobile Infrastructure as a Service",2012 IEEE Eighth World Congress on Services

[5] Chenguang Wang, Huaizhi Yan "Study of Cloud Computing Security Based on Private Face Recognition",978-1-4244-5392-4/10/$26.00 ©2010 IEEE

[6] Ashish Tanwer, Abhishek Tayal,et. "THIN APPS STORE FOR SMART PHONES BASED ON PRIVATE CLOUD INFRASTRUCTURE",2010 ITU-T Kaleidoscope Academic Confer ence

[7] Daniel W.K. TSE"Challenges on Privacy and Reliability in Cloud Computing Security"978-1-4799-3197-2/14/$31.00 ©2014 IEEE

[8] Ramgovind S, Eloff MM, Smith E"The Management of Security in Cloud Computing",978-1-4244-5495-2/10/$26.00 ©2010 IEEE

[9] Gurudatt Kulkarni, Rani Waghmare ,et. "THIN APPS STORE FOR SMART PHONES BASED ON PRIVATE CLOUD INFRASTRUCTURE", Cloud Storage Architecture", 2012 7th International Conference on Telecommunication Systems, Services, and Applications (TSSA)

[10] http://whatiscloud.com/basic_concepts_and_terminology/cloud

[11] Mohammed A. AlZain, Ben Soh,et. "MCDB: Using Multi-Clouds to Ensure Security in Cloud Computing", 2011 INEiEntEh NIEinEtEh International Conf

[12] Saakshi Narula,Arushi Jain,et. "CLOUD COMPUTING SECURITY:AMAZON WEB SERVICE", 2015 Fifth Internationational Conference on Advanced Computing & Communication Technologies

[13] Nahid Bohlol,Zohreh Safari"Systematic Parameters vs. SLAs for Security in Cloud Computing"978-1-4799-8654-5/15/$31.00 ©2015 IEEE