Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.802 – 809

RESEARCH ARTICLE

Data Security using Hadoop on Cloud Computing

Miss. Pooja.D.Bardiya[1], Miss. Rutuja.A.Gulhane[2], Dr. Prof. P.P.Karde[3]

[1]CS&IT, SGBAU, INDIA
[2]CSE, SGBAU, INDIA
[3] CS&IT, SGBAU, INDIA
[1] poojabardiya95@gmail.com; [2] gulhanerutuja@gmail.com; [3] p_karde@rediffmail.com

Abstract- There is a growing trend of using cloud environments forever growing storage and data processing needs. However, adopting a cloud computing paradigm may have positive as well as negative effects on the data security of service consumers. This paper primarily aims to highlight the major security issues existing in current cloud computing environments With the development of cloud computing, Data security becomes more and more important in cloud computing. This paper analyses the basic problem of cloud computing data security. With the analysis of HDFS architecture, we get the data security requirement of cloud computing.

Key Words: data security, cloud computing, hadoop

I. DATA SECURITY

"data Security is a multidisciplinary area of study and professional activity which is concerned with the development and implementation of security mechanisms of all available types (technical, organisational, human-oriented and legal) in order to keep information in all its locations (within and outside the organisation's perimeter) and, consequently, data processing systems, where information is created, processed, stored, transmitted and destroyed, free from threats. Threats to data and data processing systems may be categorised and a corresponding security goal may be defined for each category of threats. A set of security goals, identified as a result of a threat analysis, should be revised periodically to ensure its adequacy and conformance with the evolving environment. The currently relevant set of security goals may include: confidentiality, integrity, availability, privacy, authenticity & trustworthiness, non-repudiation, accountability and auditability." *(Cherdantseva and Hilton, 2013)*

Two major aspects of data security are:

- **Security:** Sometimes referred to as computer security, Information Technology Security is information security applied to technology (most often some form of computer system). It is worthwhile to note that a computer does not necessarily mean a home desktop. A computer is any device with a processor and some memory (even a calculator). IT security specialists are almost always found in any major enterprise/establishment due to the nature and value of the data within larger businesses. They are responsible for keeping all of the technology within the company secure from malicious cyber attacks that often attempt to breach into critical private information or gain control of the internal systems.

- **Assurance:** The act of ensuring that data is not lost when critical issues arise. These issues include but are not limited to: natural disasters, computer/server malfunction, physical theft, or any other instance where data has the potential of being lost. Since most information is stored on computers in our modern era, information assurance is typically dealt with by IT security specialists. One of the most common methods of providing information assurance is to have an off-site backup of the data in case one of the mentioned issues arises.

Governments, military, corporations, financial institutions, hospitals, and private businesses a mass a great deal of confidential information about their employees, customers, products, research and financial status. Most of this information is now collected, processed and stored on electronic computers and transmitted across networks to other computers.

Should confidential information about a business' customers or finances or new product line fall into the hands of a competitor or a black hat hacker, a business and its customers could suffer widespread, irreparable financial loss, not to mention damage to the company's reputation. Protecting confidential information is a business requirement and in many cases also an ethical and legal requirement.

For the individual, information security has a significant effect on privacy, which is viewed very differently in different cultures. The field of information security has grown and evolved significantly in recent years. There are many ways of gaining entry into the field as a career. It offers many areas for specialization including securing network(s) and allied infrastructure, securing applications and databases, security testing, information systems auditing, business continuity planning and digital forensics, etc.

The rapid growth and widespread use of electronic data processing and electronic business conducted through the Internet, along with numerous occurrences of international terrorism, fuelled the need for better methods of protecting the computers and the information they store, process and transmit.

II. CLOUD COMPUTING

Cloud computing is a Kind of network where user can use services provided by Service provider on pay per use bases. It is a research area which provides a wide range of applications under different topologies where every topology computing that is expected to be adopted by government, manufacturers and academicians in the near future.

Cloud Computing is the technology of building a robust data security between CSP and user. This technology is literally called Cloud Data Security. The emergence of the Cloud system has simplified the deployment of large-scale distributed systems for software vendors. The Cloud system provides a simple and unified interface between vendor and user, allowing vendors to focus more on the software itself rather than the underlying framework. Applications on the Cloud include Software as a Service system and Multi-tenant databases. The Cloud system dynamically allocates computational resources in response to customers' resource reservation requests and in accordance with customers' predesigned quality of service. Risk coming with opportunity, the problem of data security in Cloud computing become bottleneck of cloud computing. In this paper we want to set up a security model for cloud computing.

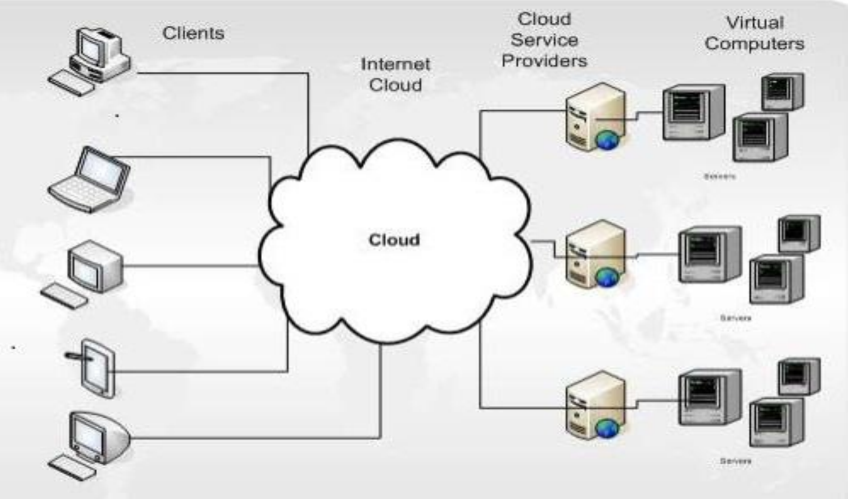

Fig 1: components in cloud computing

Cloud computing providers offer their services according to three fundamental models Infrastructure as a service (IaaS), and software as a service (SaaS) where IaaS is the most basic and each higher model abstracts from the details of the lower models.

1) Software as a Service (SaaS): The capability provided to the consumer is to use the providers applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

2) *Platform as a Service (PaaS):* The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, Operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

3) *Infrastructure as a Service (IaaS):* The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control select networking components (e.g., host firewalls).

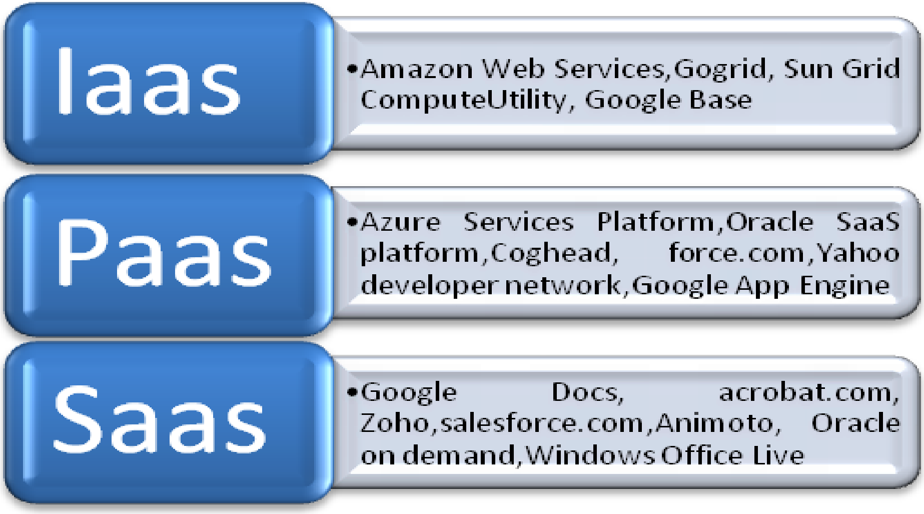
Iaas
•Amazon Web Services,Gogrid, Sun Grid ComputeUtility, Google Base
Paas
•Azure Services Platform,Oracle SaaS platform,Coghead, force.com,Yahoo developer network,Google App Engine
Saas
•Google Docs, acrobat.com, Zoho,salesforce.com,Animoto, Oracle on demand,Windows Office Live

Fig 2: Major Cloud Service Providers and Service Provider Names

A) Third Party Auditor

Third Party Auditor is kind of inspector. There are two categories: private auditability and public auditability. Although private auditability can achieve higher scheme efficiency, public auditability allows anyone, not just the client (data owner), to challenge the cloud server for the correctness of data storage while keeping no private information. To let off the burden of management of data of the data owner, TPA will audit the data of client. It eliminates the involvement of the client by auditing that whether his data stored in the cloud are indeed intact, which can be important in achieving economies of scale for Cloud Computing. The released audit report would help owners to evaluate the risk of their subscribed cloud data services, and it will also be beneficial to the cloud service provider to improve their cloud based service platform. Hence TPA will help data owner to make sure that his data are safe in the cloud and management of data will be easy and less burdening to data owner.

B) The Existence of TPA

For the enterprise providing cloud computing services, they have the right to carry out the management and maintenance of data, the existence of super-users to greatly simplify the data management function, but it is a serious threat to user privacy. Super-powers is a double-edged sword, it brings convenience to users and at the same time poses a threat to users. In an era of personal privacy, personal data should be really protected, and the fact that cloud computing platform to provide personal services in the confidentiality of personal privacy on the existence of defects. Not only individual users but also the organizations have similar potential threats, e.g. corporate users and trade secrets stored in the cloud computing platform may be stolen. Therefore the use of super user rights must be controlled in the cloud.

Moving data into the cloud offers great convenience to users since they don't have to care about the complexities of direct hardware management. The pioneer of Cloud Computing vendors, Amazon Simple Storage Service (S3) and Amazon Elastic Compute Cloud (EC2) are both well known examples. While these internet-based online services do provide huge amounts of storage space and customizable computing resources, this computing platform shift, however, is eliminating the responsibility of local machines for data maintenance at the same time. As a result, users are at the mercy of their cloud service providers for the availability and integrity of their data.

C) Security Disadvantages in Cloud Environments

Data Location: In general, cloud users are not aware of the exact location of the data center and also they do not have any control over the physical access mechanisms to that data.

Investigation: Investigating an illegitimate activity may be impossible in cloud environments. Cloud services are especially hard to investigate, because data for multiple customers may be co-located and may also be spread across multiple Data centers.

Data Segregation: Data in the cloud is typically in a shared environment together with data from other customers. Encryption cannot be assumed as the single solution for data segregation problems.

The concept of cloud computing is built on new architecture. The new architecture comprised of a variety of new technologies, such as Hadoop, Hbase, which enhances the performance of cloud systems but brings in risks at the same time. In the cloud environment, users create many dynamic virtual organizations, first set up in co-operation usually occurs in a relationship of trust between organizations rather than individual level. So those users based on the expression of restrictions on the basis of proof strategy is often difficult to follow; which frequently occurs in many of the interactive nodes between the virtual machine, and is dynamic, unpredictable. Cloud computing environment provides a user of the "buy" the full access to resources which has also increased security risks.

III. **HADOOP SECURITY**

It is a well-known fact that security was not a factor when Hadoop was initially developed by Doug Cutting and Mike Cafarella for the Nutch project. As the initial use cases of Hadoop revolved around managing large amounts of public web data, confidentiality was not an issue. For Hadoop's initial purposes, it was always assumed that clusters would consist of cooperating, trusted machines used by trusted users in a trusted environment.

Initially, there was no security model – Hadoop didn't authenticate users or services, and there was no data privacy. As Hadoop was designed to execute code over a distributed cluster of machines, anyone could submit code and it would be executed. Although auditing and authorization controls (HDFS file permissions) were implemented in earlier distributions, such access control was easily circumvented because any user could impersonate any other user with a command line switch. Because impersonation was prevalent and done by most users, the security controls that exist were not really effective.

Back then, organizations concerned about security segregated Hadoop clusters onto private networks, and restricted access to authorized users. However, because there were few security controls within Hadoop, many accidents and security incidents happened in such environments. Well-intended users can make mistakes (e.g. deleting massive amounts of data within seconds with a distributed delete). All users and programmers had the same level of access to all of the data in the cluster, any job could access any data in the cluster, and any user could potentially read any data set. Because Map Reduce had no concept of authentication or authorization, a mischievous user could lower the priorities of other Hadoop jobs in order to make his job complete faster – or worse, kill the other jobs.

All the data security technic is built on confidentiality, integrity and availability of these three basic principles. Confidentiality refers to the so-called hidden the actual data or information, especially in the military and other sensitive areas, the confidentiality of data on the more stringent requirements. For cloud computing, the data are stored in "data center", the security and confidentiality of user data is even more important. The so-called integrity of data in any state is not subject to the need to guarantee unauthorized deletion, modification or damage. The availability of data means that users can have the expectations of the use of data by the use of capacity.

A) Security goals

(1) *Storage correctness*: to ensure users that their data are indeed stored appropriately and kept intact all the time in the cloud.

(2)*Fast localization of data error*: to effectively locate the malfunctioning server when data corruption has been detected.

(3) *Dynamic data support*: to maintain the same level of storage correctness assurance even if users modify, delete or append their data files in the cloud.

(4) *Dependability*: to enhance data availability against Byzantine failures, malicious data modification and server colluding attacks, i.e. minimizing the effect brought by data errors or server failures.

(5) *Lightweight*: to enable users to perform storage correctness checks with minimum overhead.

As Hadoop became a more popular platform for data analytics and processing, security professionals began to express concerns about the insider threat of malicious users in a Hadoop cluster. A malicious developer could easily write code to impersonate other users' Hadoop services (e.g. writing a new Task Tracker and registering itself as a Hadoop service, or impersonating the hdfs or map red users, deleting everything in HDFS, etc.). Because Data Nodes enforced no access control, a malicious user could read arbitrary data blocks from Data Nodes, bypassing access control restrictions, or writing garbage data to Data Nodes, undermining the integrity of the data to be analyzed. Anyone could submit a job to a Job Tracker and it could be arbitrarily executed.

B) Hadoop architecture

Hadoop is a distributed computing framework which can scale up to thousands of computing nodes and large amount of data. If Hadoop users need to scale computation capacity or storage capacity, they just need to add commodity servers to the Hadoop cluster. The preferred operating system is GNU/LINUX with some experimental support for Windows. The name of Hadoop comes from a toy elephant which belongs to the child of the Hadoop's creator, Doug Cutting.

Hadoop is made up of two primary components, the Hadoop Distributed File System and the map reduce engine. HDFS follows master/slave architecture. A HDFS cluster usually contains a single Name Node and a bunch Data Nodes. Name Node is responsible monitoring and distributing data to Data Nodes. Data Nodes within the same cluster would communicate over the network to balance data blocks, and ensure data is replicated throughout the cluster.

Hadoop's Map Reduce was originally based on Google's Map Reduce. This type of paradigm processes application by breaking input into small parts, and these parts can be run on any node in a cluster.

The Map Reduce engine is made up of two main components, Job Tracker and Task Tracker. Users submit jobs to a Job Tracker which distributes the task to Task Trackers for data processing .HDFS has many features which fits data-intensive computing, such as high scalability, reliability and throughput. As Hadoop is increasingly being used on the grid and cloud environment, more companies will invest their research and development in Hadoop.

A small Hadoop cluster includes a single master and multiple worker nodes. The master node consists of a Job Tracker, Task Tracker, Name Node and Data Node. A slave or *worker node* acts as both a Data Node and Task Tracker, though it is possible to have data-only worker nodes and compute-only worker nodes.

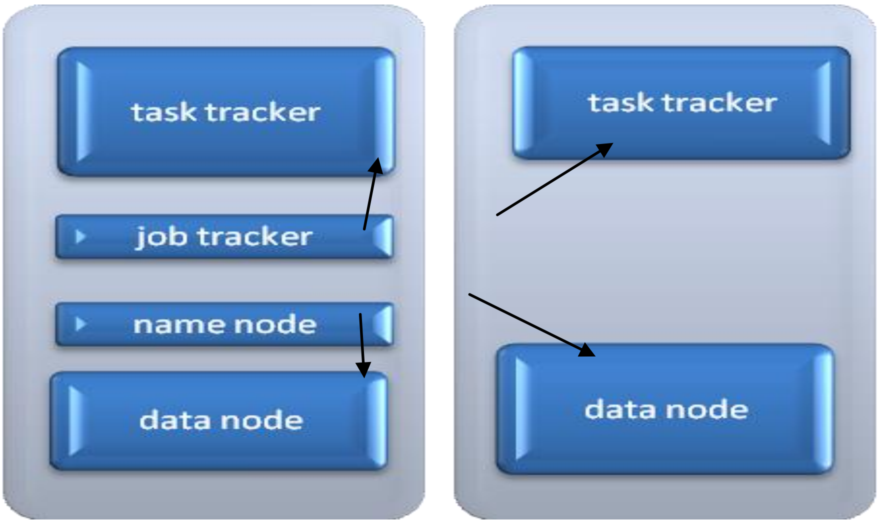

Fig3: Smaller cluster

Name Node

The HDFS namespace is a hierarchy of files and directories. Files and directories are represented on the Name Node by *inodes*, which record attributes like permissions, modification and access times, namespace and disk space quotas. The file content is split into large blocks (typically 128 megabytes, but user selectable file-by-file) and each block of the file is independently replicated at multiple Data Nodes (typically three, but user selectable file-by-file). The Name Node maintains the namespace tree and the mapping of file blocks to Data Nodes (the physical location of file data).

An HDFS client wanting to read a file first contacts the Name Node for the locations of data blocks comprising the file and then reads block contents from the Data Node closest to the client. When writing data, the client requests the Name Node to nominate a suite of three Data Nodes to host the block replicas. The client then writes data to the Data Nodes in a pipeline fashion. The current design has a single Name Node for each cluster. The cluster can have thousands of Data Nodes and tens of thousands of HDFS clients per cluster, as each Data Node may execute multiple application tasks concurrently. HDFS keeps the entire namespace in RAM. The inode data and the list of blocks belonging to each file comprise the metadata of the name system called the *image*.

The persistent record of the image stored in the local host's native files system is called a *checkpoint.* The Name Node also stores the modification log of the image called the *journal* in the local host's native file system. For improved durability, redundant copies of the checkpoint and journal can be made at other servers. During restarts the Name Node restores the namespace by reading the namespace and replaying the journal. The locations of block replicas may change over time and are not part of the persistent checkpoint.

Data Nodes

Each block replica on a Data Node is represented by two files in the local host's native file system. The first file contains the data itself and the second file is block's metadata including checksums for the block data and the block's *generation stamp*. The size of the data file equals the actual length of the block and does not require extra space to round it up to the nominal block size as in traditional file systems. Thus, if a block is half full it needs only half of the space of the full block on the local drive.

During startup each Data Node connects to the Name Node and performs a *handshake*. The purpose of the handshake is to verify the *namespace ID* and the *software version* of the Data Node. If either does not match that of the Name Node the Data Node automatically shuts down. The namespace ID is assigned to the file system instance when it is formatted. The namespace ID is persistently stored on all nodes of the cluster. Nodes with a different namespace ID will not be able to join the cluster, thus preserving the integrity of the file system.

The consistency of software versions is important because incompatible version may cause data corruption or loss, and on large clusters of thousands of machines it is easy to overlook nodes that did not shut down properly prior to the software upgrade or were not available during the upgrade.

A Data Node that is newly initialized and without any namespace ID is permitted to join the cluster and receive the cluster's namespace ID. After the handshake the Data Node *registers* with the Name Node. Data Nodes persistently store their unique *storage IDs*. The storage ID is an internal identifier of the Data Node, which makes it recognizable even if it is restarted with a different IP address or port. The storage ID is assigned to the Data Node when it registers with the Name Node for the first time and never changes after that. A Data Node identifies block replicas in its possession to the Name Node by sending a *block report*. A block report contains the *block id*, the *generation stamp* and the *length* for each block replica the server hosts.

The first block report is sent immediately after the Data Node registration. Subsequent block reports are sent every hour and provide the Name Node with an up-to date view of where block replicas are located on the cluster. During normal operation Data Nodes send *heartbeats* to the Name Node to confirm that the Data Node is operating and the block replicas it hosts are available. The default heartbeat interval is three seconds. If the Name Node does not receive a heartbeat from a Data Node in ten minutes the Name Node considers the Data Node to be out of service and the block replicas hosted by that Data Node to be unavailable.

The Name Node then schedules creation of new replicas of those blocks on other Data Nodes. Heartbeats from a Data Node also carry information about total storage capacity, fraction of storage in use, and the number of data transfers currently in progress. These statistics are used for the Name Node's space allocation and load balancing decisions. The Name Node does not directly call Data Nodes. It uses replies to heartbeats to send instructions to the Data Nodes.

The Instructions include commands to:

- Replicate blocks to other nodes;
- remove local block replicas;

- re-register or to shut down the node;
- send an immediate block report.

These commands are important for maintaining the overall system integrity and therefore it is critical to keep heartbeats frequent even on big clusters. The Name Node can process thousands of heartbeats per second without affecting other Name Node operations.

HDFS Client

User applications access the file system using the HDFS client, a code library that exports the HDFS file system interface. Similar to most conventional file systems, HDFS supports operations to read, write and delete files, and operations to create and delete directories. The user references files and directories by paths in the namespace. The user application generally does not need to know that file system metadata and storage are on different servers, or that blocks have multiple replicas.

When an application reads a file, the HDFS client first asks the Name Node for the list of Data Nodes that host replicas of the blocks of the file. It then contacts a Data Node directly and requests the transfer of the desired block. When a client writes, it first asks the Name Node to choose Data Nodes to host replicas of the first block of the file. The client organizes a pipeline from node-to-node and sends the data. When the first block is filled, the client requests new Data Nodes to be chosen to host replicas of the next block. A new pipeline is organized, and the Client sends the further bytes of the file.

C) HDFS Design Features

Detection and recovery

Hardware system failure is very common among HDFS clusters, As HDFS cluster consists of thousands of server machines, and each of them stores part of the file system's data. So failure detection and automatic recovery of nodes are important features of HDFS.

Throughput

HDFS emphasizes on high throughput of data access. It is designed for batch processing and requires streaming access to its data sets.

Handling large data

HDFS cluster provides high bandwidth and scalability. It normally runs applications with terabytes file in size.

Write once read many

HDFS applications requires access model that means a file cannot be changed once created and written by any particular user to keep data coherency and enable high data throughput.

Computation nears Data

It is inefficient to move the data to where the application is running. When the dataset is huge, computation is more efficient if it is processed near the data. It reduces network congestion and increases the throughput.

IV. **DESIGN FOR CLOUD COMPUTING WITH HADOOP**

Hadoop is a framework built to support large-scale data-intensive processing. It runs on commodity hardware and used by some of the world's most prominent companies, including IBM, HP, Apple, and Microsoft. In addition to the high-tech companies, Hadoop is also being used for science and academic research.

Once we have all the data available, we come up with a more secured HDFS system with faster file transfer protocol. First, client needs to login to the server and calls a bash script. Then, HDFS authenticates the user by matching the user with his password. If the password is correct, the requested file will be copied out of HDFS to local disk and transferred from server to client with GridFTP. After file transfer completes, the local copy is deleted. At the end, a secure file system and faster file transfer can be achieved by the proposed solution.

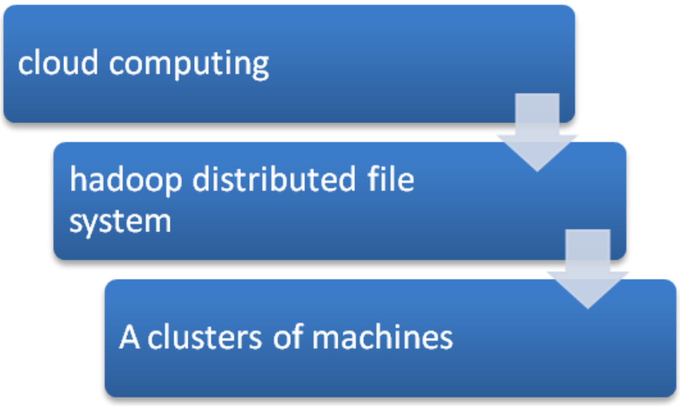

Fig 3: HDFS is a great object store solution for cloud system

V. CONCLUSION

In this paper firstly we discussed the data security with its related terms .as the development of cloud computing is increasing day–by-day. There are various providers that provide their respective services to the user .but every coins had two sides there are some security issues in cloud computing and the presence of the third party is the major threat and many more are discussed. Then we have studied secured the data model hadoop that can be the option for security in cloud. We have studied the hadoop model with its security goals and architecture. Hence a research paper is made that has the close focus on security model “hadoop” and fast running technology “cloud computing”.

REFERENCES

[1]Hadoop. (n.d.). Retrieved from http://hadoop.apache.org

[2]Mladen A. Vouk Cloud Computing – Issues, Research and Implementations Journal of Computing and Information Technology - CIT 16, 2008, 4, 235–246

[3]Cloud computing security, http: //en. wikipedia.org/ wiki/ Cloud_computing_security.

[4]S. Subashini, V.Kavitha. A survey on security issues in service delivery models of cloud computing. Journal of Network and Computer Applications 34(2011)1-11

[5]Mohamed Al Morsy, John Grundy, Ingo Müller, “An Analysis of The Cloud Computing Security Problem,” in Proceedings of APSEC 2010 Cloud Workshop, Sydney, Australia, 30th Nov 2010

[6]Sun Cloud Architecture Introduction White Paper (in Chinese).
http://developers.sun.com.cn/blog/functionalca/resource/sun_353cloudc omputing_chinese.pdf

[7]Voorsluys, William; Broberg, James; Buyya, Rajkumar (February 2011). "Introduction to Cloud Computing". In R. Buyya, J. Broberg, A.Goscinski. Cloud Computing: Principles and Paradigms. New York, USA.

[8]http://dwachira.hubpages.com/hub/Data-Security-Risks-In-Cloud-Computing.

[9]Abhishek Mohta,Ravi Kant Sahu and LK Awasthi, “Robust Data Security for Cloud while using Third Party Auditor” in International Journal of Advanced Research in Computer Science and Software Engineering, Vol No. 2, Issue 2,Feb 2012

[10]http://www.wikipedia.com

[11]“Enhanced Data Security in Cloud Computing with Third Party Auditor”byAuditor”Bhavna Makhija, VinitKumar Gupt, Indrajit Rajput of International Journal of Advanced Research in Computer Science and Software Engineering.