National Conference on Emerging Trends in Computer, Electrical & Electronics (ETCEE-2015)
International Journal of Advance Engineering and Research Development (IJAERD)
e-ISSN: 2348 - 4470 , print-ISSN:2348-6406,Impact Factor:3.134

# Comparison of Distributed File System for Cloud With Respect to Healthcare Domain

[1]Ms. Khyati Joshi, [2]Prof. ChintanVarnagar
[1]*Post Graduate Student,*[2]*Assistant Professor, Computer Engineering Department, Atmiya Institute of Science and Technology, Rajkot, Gujarat, India*
[1]kpjoshi91@gmail.com, [2]chintan2287@gmail.com

***Abstract-*** **Distributed file systems are considered to be basic building blocks of cloud computing technology. Some of the file systems used to deploy cloud is Google file system, Hadoop file system and CEPH file system. Reviewing current healthcare statistics and current hospital management system, a range of benefits can be obtained if healthcare system is shifted to cloud. In cloud computing, the data storage and computing are not in the local computer and server but in the amount of computers distributed in the internet. Several distributed file systems are used over the cloud because the cloud itself includes large numbers of commodity-grade servers, harnessed to deliver highly scalable and on-demand services. But a suitable file system inside healthcare is required which is reliable, scalable, and can handle large amounts of patients records giving a higher performance. So, in this paper, we have reviewed3 distributed file system with respect to its architecture and its suitability inside healthcare cloud.The Hadoop Distributed File System and google file system is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications and both have a similar architecture with some differences.But HDFS is open source framework while gfs is not. Ceph is another distributed file system which has different architecture. Parameters such as reliability, scalability and storage requirements etc. are taken for comparison with respect to healthcare.**

***Keywords: Cloud Computing, Healthcare Cloud, GFS, HDFS, Ceph, File Systems for Cloud***

## 1. INTRODUCTION

Health and Healthcare are now pre-requisites of every people. No. of patients are increasing day by day in every hospitals. Every hospital has to maintain a lot of patient records and even technicians and IT staff to operate computers. Following are the healthcare scenarios in government hospitals.

*a. RISING HEALTHCARE EXPENDITURE AND UNSUSTAINABLE HEALTHCARE SYSTEMS [9]*

Health spending continues to rise faster than economic growth in most Developing countries. Health spending reached 9.5% of GDP on average in 2009, up from 6.6% in 1982.

*b. RISE OF CHRONIC DISEASES [9]*

Chronic diseases cost around 75% of healthcare budgets and account for 85% deaths inEurope3.

c. *MEDICATIONERRORS [10]*

Over 5 million outpatient prescription errors could be avoided yearly through the use of electronically transferred prescriptions.

*d. MEDICAL ERRORS DUE TO POOR COMMUNICATIONS [10]*

Poor communication is the causal factor in over 60% of medical errors.

As per Demography of India [1], data Sources on Health Information in India is as Follows.

TABLE I
DATA SOURCE ON HEALTH INFORMATION

| Type of Source | Description | Strengths | Limitations |
|---|---|---|---|
| Service Records | Service gen-erated data derived From various health facility and patient provider inter-actions covering morbidity | Used for service managem-ent.Yearly data on services rendered, monthly data possible and in few | Excludes those not accessing the services (in-built sele-ctionbias); Incompleten ess and data quality. Reporting problems, |

| | | | |
|---|---|---|---|
| | type and mortality by cause, care offered, Quality of care, treatment administered and service and Rendered. | cases monthly data are being compiled; Basis for disease surveillance clarity and systemsto detect out-breaks; & Useful in measuring performanc e -e of facilities and its monitoring | Data duplication and inconsistenc y; Private sector often not included; Lack of data analysis Including disaggregatio n and use of service Statistics in local area planning. |

As per above information in this table, various problems occurs related to patient's records. So, for recovering all these problems, a proper manage-ment of patient's data is required. Nowadays most of the hospitals are managing their operations and medical information through Hospital Information System (HIS). A typical HIS system can have the following sub systems or modules and can have more: Information security has always been a big issue as the number of malware/malicious attacks is increasing regardless the improvement in the area of information security. The network connection (internet) is never thought of and realized as an effective and most safe connection. Due to the lack of proper technology (safe internet) government hospitals has established independent and stand-alone HIS systems in some districts / small towns. These independent standalone HIS systems require high construction and maintenance cost; hence the information sharing is a big issue. 2nd issue is that currently, most of the hospitals (either district hospital or town hospital) are maintaining patients' medical information and some staff management operations manually (on papers); however some of the hospitals have their own independent stand-alone HIS system. This creates a high risk for data loss and the problematic issues in data sharing. The main problematic issue is that the patients could not seek proper treatment in any hospital of the other district or town because of the existing independent information systems due to lack of data sharing. A whole information system setup cost usually takes the hardware (equipments including servers and client computers) cost, software development, testing and maintenance cost in account. Current independent standalone HIS systems are very much expensive to setup and just to facilitate specific hospitals of specific districts/towns. Setting up an information system has never been a big issue; however managing, upgrading and maintaining the information system have always been a big concern. Existing independent standalone HIS systems are hard to manage, upgrade and maintain; as for different hospitals in different districts/ towns, separate management and maintenance is required. However, management and maintenance process requires continuous investment including technical skills, proper users and experts. Up-grading the current information system has also been a big issue as the user needs/requirements are changing with time to time. This is same in hospitals as different hospitals produce different individual needs/requirements for upgrading the current HIS system.

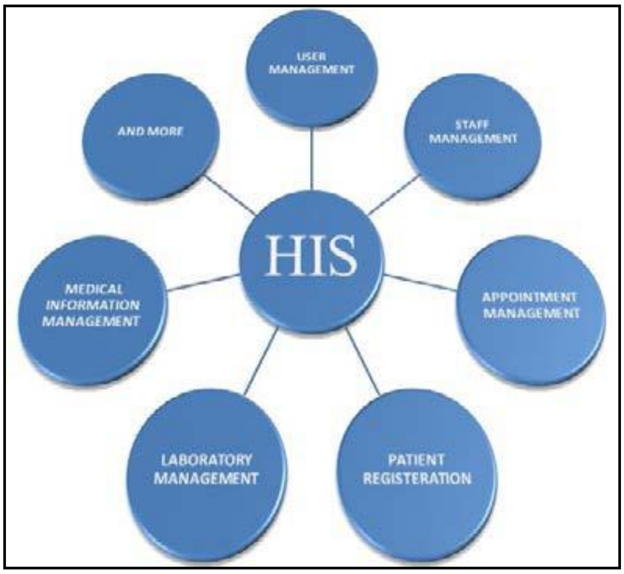


FIGURE 1: HIS SYSTEM

## A. Changes Required In Government Hospitals related to Patient Records

As per above analysis, to recover from present scenarios in government hospitals, a more concerned and a reliable approach should be implemented for maintaining patient records. Because every patient are critical and centre factor in a hospital. If patient's clinical data and records stored safely and retrieval also is safe and faster, it creates a major benefit to doctors, patients and other hospital staffs. Although

Technology is erecting day by day, Our Government Hospitals are still not utilizing it in a concise manner.

Following changes need to be made with relation of patient records.

a. reliable access of records
b. Long-term Storage and Management of records
c. Easy up gradation of records
d. Security of records

## 2. CLOUD COMPUTING AND HEALTHCARE CLOUD

[4] Cloud computing is the delivery of computing and storage capacity as a service to users. Cloud Computing, the long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service and shaping the way IT hardware is designed and purchased.

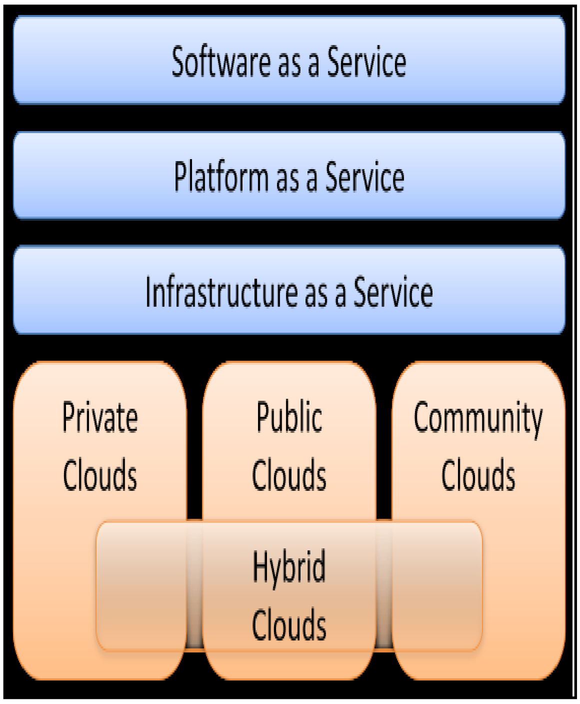


FIGURE-2: CLOUD COMPUTING FRAMEWORK

### A. Characteristics of cloud framework

1. *On-demand self-service.* A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

2. *Broad network access.* Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

3. *Resource pooling.* The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or data centres). Examples of resources include storage, processing, memory, and network bandwidth.

4. *Rapid elasticity.* Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

5. *Measured service.* Cloud systems automatically control and optimize resource use by leveraging a metering capability1 at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

*B. Healthcare cloud*

Healthcare cloud consists of Cloud which can be private or public or hybrid. Patient records are stored and accessible through cloud. Users who can be patients, doctors, nurses or Data operator can view patient records inside cloud and anywhere at any time. Although security parameters have to be taken into consideration of patient records and access capability of each type of user, the main benefit for moving patient record to cloud is anywhere accessibility by patients as well as doctors. They just need internet connection and they can use health information application stored in cloud .All other storage and maintenance of information handled by cloud providers.
There will be private or public clouds in each state maintaining patient records of all its Government

e-ISSN: 2348 - 4470 , print-ISSN:2348-6406,Impact Factor:3.134

hospitals. The reason behind connecting states cloud is that if a patient for any reason either visits or permanently migrates to out states hospitals, then all his records can be shared or migrated to corresponding state cloud and connection between all state clouds kept on demand basis.

## 3. STUDY OF VARIOUS FILE SYSTEMS TO IMPLEMENT CLOUD SERVICES

### A: Google File System

Google File System, a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients. It is widely deployed within Google as the storage platform for the generation and processing of data used by Google service as well as research and development efforts that require large data sets[8].

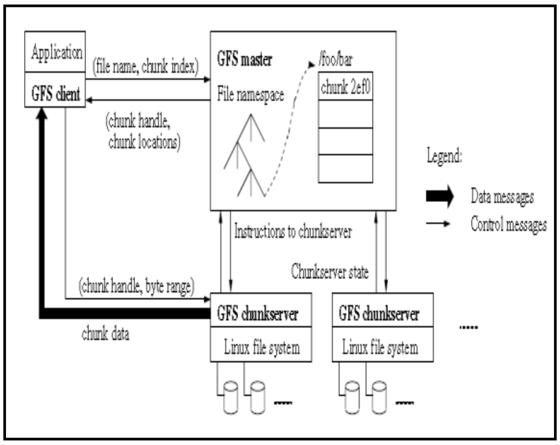


FIGURE 3: ARCHITECTURE OF GOOGLE FILE SYSTEM [8]

A GFS cluster consists of a single *master* and multiple *chunk servers* and is accessed by multiple *clients*, as shown in Figure. Each of these is typically a commodity Linux machine running a user-level server process. It is easy to run both a chunk server and a client on the same machine, as long as machine resources permit and the lower reliability caused by running possibly flaky application code is acceptable. Files are divided into fixed-size *chunks*. Each chunk is identified by an immutable and globally unique 64 bit *chunk handle* assigned by the master at the time of chunkcreation.Chunkservers store chunks on local disks as Linux files and read or write chunk data specified by a chunk handle and byte range. For reliability, each chunk is replicated on multiple chunk servers. By default, we store three replicas, though users can designate different replication levels for different regions of the file namespace. The master maintains all file system metadata. This includes the namespace, access control information, the mapping from files to chunks, and the current locations of chunks. It also controls system-wide activities such as chunk lease management, garbage collection of orphaned chunks, and chunk migration between chunk servers.

### B: Hadoop Distributed File System

HDFS is the file system which is used in Hadoop based distributed file system. The Hadoop is an open-source distributed computing framework and provided by Apache. Many network stations use it to create systems such as Amazon, Facebook. The Hadoop cores are Map reduce [11] and HDFS [7]. The map reduce can make the decomposition of tasks and integration of results. The HDFS is a distributed file system and provide the base support for the storage of file in the storage node.[2]

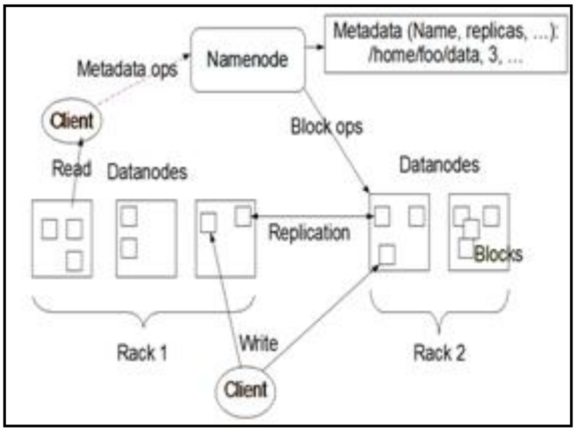


FIGURE-4: ARCHITECTURE OF HADOOP DISTRIBUTED FILE SYSTEM [7]

HDFS stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. A HDFS installation consists of single name node as the master node and a number of data nodes as the slave nodes. The name node manages the file system namespace and regulates access to files by clients. The data nodes are distributed, one data node per machine in the cluster, which manage data blocks attached to the machines where they run. The name node executes the operations on file system namespace and maps data blocks to data nodes. The data nodes are responsible for serving read and write requests from clients and perform block operations upon instructions from name node. HDFS distributes data chunks and replicas across the server for higher performance, load-balancing and resiliency.

With data distributed across all servers, any server may be participating in the reading, writing, or computation of a data-block at any time. HDFS replicates file blocks for fault tolerance. An application can specify the number of replicas of a file at the time it is created, and this number can be changed any time after that. The name node makes all decisions concerning block replication. For a large cluster, it may not be practical to connect all nodes in a flat topology. The common practice is to spread the nodes across multiple racks. Nodes of a rack share a switch, and rack switches are connected by one or more core switches.

### C: Ceph Distributed File System

Ceph is an open source model and it is based on object based parallel file system. [5]Ceph maximizes the separation between data and metadata management by replacing allocation tables with a pseudo-random data distribution function (CRUSH) designed for heterogeneous and dynamic clusters of unreliable object storage devices (OSDs).

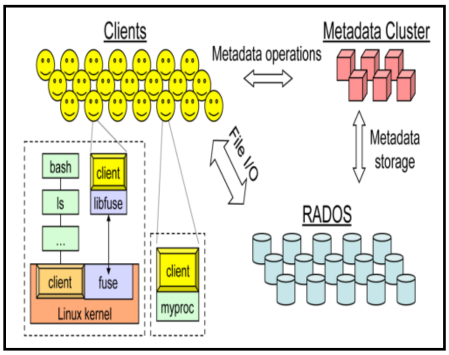


FIGURE 5: ARCHITECTURE OF CEPH DISTRIBUTED FILE SYSTEM.[5]

Ceph is an *object-based parallel file system* whose design is based on two key ideas. The first key idea is object-based storage, which splits the traditional file system architecture into a client component and a storage component.[6] The storage component manages disk scheduling and layout locally, relieving clients and servers from low-level per-disk details and increasing scalability. This design allows clients to communicate with storage nodes via a high-level interface and manage data in terms of objects, which are chunks of data much larger than 512-byte blocks. The T10 standard of the SCSI Object Storage Device (OSD) command set is an example of an object interface specification.

Ceph uses and significantly extends the concept of OSDs. For all practical purposes, think of a Ceph OSD as a process that runs on a cluster node and uses a local file system to store data objects. The second key idea in the Ceph design is the separation of data and metadata. Management of data differs fundamentally from management of metadata: file data storage is trivially parallelizable and is limited primarily by the network infrastructure. Metadata management is much more complex, because hierarchical directory structures impose interdependencies (e.g., POSIX access permissions depend on parent directories) and the metadata server must maintain file system consistency. Metadata servers have to withstand heavy workloads: 30–80% of all file system operations involve metadata, so there are lots of transactions on lots of small metadata items following a variety of usage patterns. The three unique aspects of Ceph's design are: distributed metadata management in a separate metadata server (MDS) cluster that uses dynamic sub tree partitioning to avoid metadata access hot spots and that is robust against non-byzantine failures; calculated pseudo-random data placement that allows for very compact state that can be shared easily throughout the system (CRUSH); and distributed object storage using a cluster of intelligent OSDs which forms a reliable object store that can act autonomously and intelligently (RADOS).

## 4. COMPARISON OF GFS, HADOOP AND CEPH WITH RESPECT TO HEALTHCARE CLOUD

| CRITERIA OF HEALTHCARE CLOUD FILE SYSTEM | GFS | HDFS | CEPH |
|---|---|---|---|

| | | | |
|---|---|---|---|
| AVAILABLE OPENLY | No. Not an open source framework. But new similar file system can be made by implementing all its features. | Available as an open source code. Based on map reduce [11] framework. | Available as an open source code. Implemented in c++ language. |
| RELIABLE | Yes | Yes | Yes |
| SCALABLE | Yes | Yes | More than other two file system.[3] |
| METADATA SERVER AND DATA SERVER DISTRIBUTION | Metadata and data distribution management done by one centralized server. More than 1 metadata server can be placed in a cluster for more performance. | Metadata and data distribution management done by one centralized server. More than 1 metadata server can be placed in a cluster for more performance. | Metadata and data distribution management done separately by metadata server and data server respectively. |
| METASERVER STORAGE | 1 or many depends on requirement | 1 or many depends on requirement. | Dynamically distributed metadata. Requires less no. of servers compared to other two file system. |
| DATA STORAGE DEVICES | Data servers | Data servers. | Object storage devices[10] |
| FAULT TOLERANCE | YES | YES | YES |

## 5. CONCLUSION

As far as the requirement of healthcare domain is concerned distributed file systems, Hadoop and ceph file system are more suitable to achieve the entire necessary performance requirement in healthcare cloud. But for extending more scalability, ceph file system remains more suitable choice because Hadoop and ceph both are reliable but ceph gives more scalability than hadoop.So, for required reliability, scalability and performance and less number of metaservers, ceph File system is more suitable file system in the domain of healthcare cloud.

## 6. REFERENCES

[1]. ArvindPandey, Nandini Roy, Rahul Bhawsar and R.M.Mishra, **"***Health Information System in India: Issues of Data Availability and Quality"***,** Demography India Vol. 39, No. 1, 2010, pg. 111-128.

[2]. S.ChandraMouliswaran and ShyamSathyan, et. al, *"Study On Replica Management And High Availability In Hadoop Distributed File System (HDFS) "*, Journal of Science Vol 2. , 2012 , pg. 65-70.

[3]. R.Vijayakumari, R.Kirankumar, K.GangadharaRao, *"Comparative Analysis of Google File System and Hadoop Distributed File System"*, ICETS- International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3, No.1, 2014, Pg. 553– 558.

[4]. P. Mell and T. Grance, *"The NIST Definition of Cloud Computing"*, NIST Special publication, September 2011, pg. 145-800.

[5]. Carlos Maltzhan, Esteban Molina-Estolona, et.al. ,*"Ceph as a scalable alternative to the Hadoop Distributed File System"*, Login: Vol. 35, No. 4, August 2010.

[6].Sage A. Weil, Scott A. Brandt, Ethan L. Miller, Darrell D.E. Long, CarlosMaltzahn, *"Ceph: A Scalable, High-Performance Distributed File System"*, OSDI, 7th USENIX.

[7].J.Hanson. (Feb 2011). *An Introduction to Hadoop Distributed File System.* [Online]Available: http://www.ibm.com/developerworks/web/library/wa-introhdfs

[8].SanjayGhemawat, Howard Gobioff, and Shun-TakLeung,"*The Google File System"*, SOSP, ACM, Oct. 2003.

[9]. Atiya Parveen1, Sobia Habib2, Waseem Ahmad3, "*The Cloud Changing The Indian Healthcare System"*, IJCSMC Vol. 2, Issue. 5, May 2013, pg. 238 – 243.

[10]. *"Advancing healthcare delivery with cloud computing"*, Cocir eHealth Toolkit Part-2, 2012.

[11]. *Map Reduce tutorial*[Online].Available: http://hadoop.apache.org/common/docs/r0.20.2/mapred_tutorial.html.