Journal of Social Sciences 6 (3): 310-314, 2010
ISSN 1549-3652
© 2010 Science Publications

Circumventing Picture Archiving and Communication Systems Server with Hadoop Framework in Health Care Services

[1]Gopinath Ganapathy and [2]S. Sagayaraj
[1]Department of Computer Science, Bharathidasan University, Trichy
[2]Department of Computer Science, Sacred Heart College, Tirupattur

Abstract: Problem statement: Features and challenges of the PACS server solutions are elaborated in the context of large scale computing. **Approach:** Hadoop is a pivotal piece of the data mining renaissance offers the ability to tackle large data sets in ways that weren't previously feasible and clarifies certain functionalities such as MapReducer and Hadoop distributed file system. **Results:** The PACS server was highlighted in Health Care System with primary Functions of DICOM and basic operations of query, retrieval and routing were performed on various images. The study attempted to propose a concept called Hadoop Picture Archiving and Communication System (HPACS) same as any other PACS server except that is uses distributed storage and distributed computing on commodity of hardware. **Conclusion/Recommendations:** The features of PACS and HPACS are compared in terms of storage, backup, cost, performance, turnaround time and backup. Finally, the advantages of Hadoop solution were explained.

Key words: Hadoop, MapReducer, PACS picture archiving and communication system, DICOM

INTRODUCTION

Hadoop is one of the most salient pieces of the data mining renaissance which offers the ability to tackle large data sets in ways that weren't previously possible due to time and cost constraints It is a part of the apache software foundation and its being built by the community of contributor in all over the world. The Hadoop project promotes the development of open source software and it supplies a framework for the development of highly scalable distributed computing applications (Venner, 2009). The Hadoop framework handles the processing details, leaving developers free to focus on application logic. Hadoop handles thousands of terabytes and pitabytes on thousands of nodes. If there is a node failure the Distributed File System (DFS) will facilitate rapid data transfer rates among nodes and will allow the system to operate uninterruptedly.

The study begins with a status of the current search engine followed by the discussion of related work for semantic search. Then, Semantic search directions are presented followed by a detailed comparison of the tools in the direction based on functionality and Interface with future research avenues.

The study begins with the discussion on the Hadoop approach followed by the detailed features and challenges of the PACS Server. Then, the comparison scenario between PACS and HPACS are presented followed by the solution provided by the Hadoop is presented. Last section deals with the findings of the study.

MATERIALS AND METHODS

The Hadoop approach: Hadoop is designed in such a way to efficiently process large volumes of information. It connects many commodity computers together so that they could work in parallel. Hadoop will tie smaller and low priced machines together into a compute cluster. It is a simplified programming model which allows the user to write and test distributed systems quickly. It is an efficient, automatic distribution of data and it works across machines and in turn it utilizes the underlying parallelism of the CPU cores.

The Hadoop Distributed File System (HDFS) will break large data files into smaller parts which are managed by different nodes in the cluster. In addition to this, each part is replicated across several machines, so that a single machine failure does not lead to non-availability of any data. The monitoring system then re-replicates the data in response to system failures which can result in partial storage. Even though the file parts are replicated and distributed across several machines, they form a single namespace, so their contents are universally accessible. The replications are

Corresponding Author: S. Sagayaraj, Department of Computer Science, Sacred Heart College, Tirupattur not considered as drawbacks in a distributed environment.

MapReduce (White, 2009) is a functional abstraction which provides an easy-to-understand model for designing scalable, distributed algorithms.Hadoop limits the amount of communication which can be performed by the processes, as each individual record is processed by a task in isolation to one another. In Hadoop, program must be written to conform to a particular programming model, namely "MapReduce". Under MapReduce, Mappers process the records in isolation. The output of Mappers is provided to the second set of tasks called Reducers, where results from different mappers can be merged together.

Communication in Hadoop is performed implicitly. Pieces of data can be tagged with key names. They in turn can inform Hadoop, on how to send related bits of information to a common destination node. Individual node failures can be handled successfully by restarting tasks on other machines. User-level tasks do not communicate explicitly with one another. Therefore no messages need to be exchanged by user programs. The other workers continue to operate as though nothing went wrong because the Hadoop layer manages the restarting the program partially.

PACS server: Researchers are still exploring the current performance and future potential of Picture Archiving and Communication Systems (PACS) and Radiology Information Systems (RIS). Research suggests that most, if not all, facilities' experiences with PACS and RIS are positive. This is the good omen for the future implementation and growth of PACS (Dreyer *et al.*, 2001).

Picture Archiving and Communication Systems (PACS) is a combination of hardware and software dedicated to the short and long term storage, retrieval, management, distribution and presentation of medical images. Electronic images and reports are transmitted digitally via PACS; this eliminates the need to manually file, retrieve, or transport film jackets. The universal format for PACS image storage and transfer is Digital Imaging and Communications in Medicine (DICOM).

PACS is becoming an essential tool for hospitals and imaging centers. Thus it complements digital modalities and meeting increasing productivity demands. However, PACS delivers the desired results only when fully integrated with a RIS and digital modalities conforming to Digital Imaging and Communications in Medicine (DICOM) standards.

DICOM has become the industry standard for transferring radiologic images. It has also become important in other medical information among diagnostic and therapeutic equipment and systems designed by various manufacturers. It was structured in line with the Open System Interconnection of the international Standards Organization. DICOM connectivity is useful to the technology users to provide radiology services within facilities and across geographic regions. DICOM permits workstations, Computed Tomography (CT) scanners, Magnetic Resonance (MR), imaging equipment, film digitizers, shared archives, laser printers and computers to "communicate to one another" across an open system network. This results in capturing medical images rapidly and routing quickly. Moreover physicians can make more quick diagnoses and treatment decisions.

Primary functions of DICOM: DICOM supports five primary functions at the application level: Transmission and persistence of complete objects, query and retrieval of objects, Performance of specific actions, Workflow management and Image quality and consistency for display and print.

DICOM does not define the whole structure of the system's architecture. To overcome communication barriers between dissimilar devices and systems DICOM standard is used.

PACS is an enterprise-wide information and image management system. It improves efficiency in diagnostic imaging departments. This technology integrates imaging specialties, interfaces with hospital and departmental information systems and manages the storage and distribution of images. This largely benefits radiologists, hospital physicians, remote referring physicians and specialists, clinics and imaging centers.

PACS provides real-time radiology service. Filmless imaging is a major component of this service. Real-time radiology service also consists of the IRIS, Voice Recognition Technology (VRT) and a document imaging system. With the help of PACS the radiology department, other clinical areas of the hospital and remote users can easily access images and patient data. All radiology images and patient data can be easily accessed by means of a single-tiered archive supporting a PACS-RIS network.

Query, retrieval and routing: PACS performs three basic operations namely query, retrieval and routing. Query is defined as searching a database of imaging examinations using various criteria, such as patient name or medical record number. Retrieval means "pulling" images from one storage location to another location. Image routing is a key feature of PACS-RIS integration. But the ability to quickly pull images rather than push them to various locations is a new critical demand of PACS operation. PACS allows the radiologist to rapidly select the images. By pulling the images, PACS reduces traffic on the system network and it provides greater flexibility in distributing cases among radiologists.

Data migration: The need to archive patient information may take several generations of PACS-RIS. Health care institutions are very much concern about the data migration when evaluating new technology. Migrating data requires formats or paths for data transfer-often proprietary-from existing systems to the PACS.

Moving data to new PACS platforms no longer requires a complete system overhaul. Storage middle ware can meet the challenge of keeping up with technology. It can also achieve PACS migration. These software-based solutions simplify migration and ongoing data storage. This is possible by providing well-defined interfaces between a complex storage infrastructure and the rest of the system. Middleware “virtually” stores the data away from the PACS. It also enables migration and archival management behind the scenes.

Challenges in PACS solutions:

Expensive hardware: The Hardware for PACS solution is quite expensive. There is a need to get high-end systems in order to provide the facility. As the number of images grow, it is mandatory to have high end systems to fetch the images.

Archrivals and retrievals: For hospitals with more patients, it is always difficult to maintain the Images. Mostly the PACS solutions recommend to have the current patient’s images on the server and suggest to move all other images to any Physical Removable Drives. It needs a lot of human intervention to assist the system to facilitate the images of the patients. Hence the retrieval process is not fully automated.

Scalability: As the number of images grow, it is always difficult to manage with the existing hardware infrastructure. It is not easy to expand the hard wares for the PACS Solution. Mostly the images are moved to any removable external devices

HL7 standard: HL7 standard always recommends to have replicated copies of patients records. PACS server needs to have replicated copies of the images of the Patients and it is not an easy task. For every image of a patient, at least one backup is needed within the setup and another back-up is required to take the images outside the building.

RESULTS AND DISCUSSION

HPACS and PACS scenario: Hadoop Picture Archiving and Communication System (HPACS) is the same as any other PACS server except that is uses distributed storage and distributed computing on commodity of hardware. Functional vise it works the same as OPACS, but in implementation vise it uses distributed mechanism. Hence the performance is greater and the cost is reduced compared to Enterprise PACS servers.

In a hospital consider every day there are 100 patience to visit and there are at least 20 DICOM images generated. Each images average size is 50 MB. So the total average size for the storage needed per day is 20×50 MB (1000 MB).

Table 1 illustrates the required storage capacity for one year (the following HPACS solution discussed during the comparison uses a Hadoop cluster consists of 1 server and 4 nodes).

System configuration: The PACS uses only one high configured server:

HDD	I TB
RAM	4 GB
Processor	XEON (server), 3.40 Ghz speed
Mother board	Intel

The HPACS systems required are:

- Single server
- 4 nodes to establish a cluster

The configurations required are:

Server:

HDD	320 GB
RAM	4 GB
Processor	XEON (server), 3.40 Ghz speed
Mother board	Intel

Table 1: PACS storage capacity for one year

Duration	No. of images	Size
1 day	20	1000 MB
1 Month	600	29.3 GB
1 year	7200	351.6 GB
5 years	21600	1.7 TB

Nodes:

HDD	320 GB
RAM	2 GB
Processor	Intel dual core, 2.93 Ghz
Mother board	Intel

Storage analysis: A typical PCAS server can have average of I TB storage. In HPACS it's the number Nodes multiplied by the hard disk size. As in this case the HPACS storage capacity is 4×320 GB.

The total number of images that can be stored on these servers can be obtained using the following formula:

S = Average size of a DICOM image
P = PCAS server storage capacity
M = Nodes storage capacity
N = Number of Nodes
H = HPCAS storage capacity
= (M×N)

So, the number of images that can be stored on a PACS server are No. of images on PACS = (P/S).

The number of images that can be stored on HPACS serve r are No. of images on HPACS = (H/S).

As per the scenario mentioned above P = 1 TB = 1048576 MB. H = (320 GB * 4) = 1310720 MB. No of images can be stored on PACS (D) = (P/S) = (1048576/50) = 20972. No of Images cab be stored on HPACS (F) = (H/S) = (1310720/50) = 26214.

Cost analysis: A high configured PACS server cost around Rs. 10,000,000. For the HPACS implementation it requires a ordinary server and 4 more ordinary systems. The cost involved for implementing this would come around 1,000,000 Rs. The ratio is 10:1.

PACS per-image cost formula derivation: C = PACS server cost. No of images can be stored on PACS (D) = (P/S). Per image PACS cost = C/D.

HPACS per-image cost formula derivation: E = HPACS Server Cost. No of images can be stored on HPACS (F) = (H/S). Per Image HPACS Cost = E/F.

As per the scenario the per-image in both implementations are per Image PACS Cost = 10,000,000/20972 = Rs. 477. Per Image HPACS Cost = 1,000,000/26214 = Rs. 38.

Scalability analysis: To scale up PACS server to accommodate another 1 TB of data and to increase its performance relatively, one needs to purchase PACS blade which cost around Rs. 6,000,000. In case of Hadoop to scale up HPACS to accommodate more data and to increase its performance, one needs to add one or two nodes to the existing cluster. This is an ordinary system with the cost of 2 lacks per system.

Scalability cost for PACS = Rs. 6,000,000. Scalability cost for HPACS = Node Cost (Rs. 200,000) * No. of new Nodes.

From the above after 3 years the PACS needs to be extended, where as HPACS with 4 nodes, needs a scale up only after 3.6 years. As per the above scenario extension of one new node spending 2 lacks of rupees with server up to another 9 months.

Performance analysis: In PACS servers the performance gradually goes down due to the growth of data. To optimize the performance the cache needs to be cleared in a regular interval. And increase of RAM size is also needed. If the performance is reduced after 500 GB of data growth, one needs to double to size of RAM to optimize the performance.

Where as in HPACS, the performance is the same. The growth of the data does not affect the performance; the load is shared among systems and processed on local machines.

Turnaround time analysis: The Turnaround time is calculated using the following formula:

Turnaround Time (TT) = Seek Time (ST) + Retrieval Time (RT)

Seek time: Seek time refers to the time taken by the head of a hard disk to locate a particular address. A Disk's seek time is directly Propositional to its Rotation Speed (RPM). The average RPM of a HDD calculated to 5,400 RPM. The average seek time of a HDD is about 0.1 Ms.

Retrieval time: Retrieval time is the time taken to read the data from hard disk. It varies for each process depending on the size of the data to be retrieved. Retrieval time is calculated based on capacity of the disk, how much data it can read at a time. The average value of HDD is 50 MB sec^{-1}.

PACS: The Table 2 describes the turnaround time required in an ordinary system. Turnaround time is a approximate value. In real time scenario there would be hardware signal delay, buffering delay. All these factors are not considered.

HPACS: Since the process is distributed the time is much reduced in HPACS solution. The Table 3 shows the turnaround time taken for HPACS solution.

Table 2: PACS seek, retrieval and turnaround times

No. of images	Size of data (MB)	Seek time	Retrieval time	Turnaround time (sec)
1	50	0.1 Sec	1 Sec	1.1
10	500	0.1 * 10 = 1 Sec	(1/50)*500 = 10	11
100	5000	10 Sec	(1/50) * 5000 = 100	110

Table 3: HPACS seek, retrieval and turnaround times

No. of images	Size of data (MB)	No. of machines	Seek time (sec)	Retrieval time	Turnaround time (sec)
1	50	5	0.1	1 Sec	1.1
10	500	5	0.1	(500/5)/50 = 2 Sec	2.1
100	5000	5	10	(5000/5)/50 = 20	30

Backup: In PACS Solutions the backup's are maintained manually, System does not handle it automatically. However the backup options can be provided using Redundant Array of Independent Disks (RAID) architecture. Other than that the backups are done by copying the DICOM images to a DVD or to some external Storage devices. These costs are very less when compared to the PACS server. But on the long run it incur a significant amount.

Where as in HPACS solution the backups are set to be default. The number of replicas set for the cluster configuration ensures the same numbers of copies are available for each file. This makes the system reliable. And there is no external cost involved for backups. In case if a single system fails, there is always a copy available in the other system which can be replicated automatically.

Proposed solution: The proposed solution is Hadoop based Imaging Solution. This will replace the PACS Server in the above mentioned Architecture and other components remain the same in the Architecture. Hadoop framework provides the following benefits in the field of imaging solution.

Scalability: Always extend the nodes as per the requirement. Adding a node to the network is as simple as hook a Linux box to the network and copy few configuration files. Also, Hadoop provides details about the available space in the Cluster. So, by just looking at the report one can decide whether one can add a node or not.

Cost effective: Since the Linux nodes are always cheap, there is no need for investing much on the hardware as well as OS.

Replication: Since Hadoop automatically makes three replications, it meets the requirements of HL7 standard. It replicates at-least 3 copies of each file and by default it meets the requirements of HL7.

CONCLUSION

Apache Hadoop is a framework for running applications on large clusters built of commodity hardware. The Hadoop framework transparently provides both reliability and data motion. Hadoop implements a computational paradigm named Map/Reduce, where the application is divided into many small fragments of work. Each of which may be executed or re-executed on any node in the cluster. So, by just replacing the PACS Server with Hadoop Framework can lead to good, scalable and cost effective tool for the Imaging solution for Health Care System.

REFERENCES

Dreyer, K.J., A. Mehta and J.H. Thrall, 2001. PACS: A Guide to the Digital Revolution. 1st Edn., Springer, ISBN: 10: 0387952918, pp: 435.

Venner, J., 2009. Pro Hadoop. 1st Edn., Apress, ISBN: 13: 9781430219439, pp: 440.

White, T., 2009. Hadoop: The Definitive Guide. 1st Edn., O'Reilly Media, Inc., ISBN: 10: 0596521979 pp: 528.