© 2016 IJSRSET | Volume 2 | Issue 1 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Content Based Image Retrieval System over Hadoop Using MapReduce

Nilesh Lohar*, Dipankar Chavan, Sanjay Arade, Amol Jadhav, Deepti Chikmurge
Computer Department, MITAOE, SP Pune University, Pune, Maharashrta, India

ABSTRACT

A huge amount of image data is generated everyday due to rapid evolution of image capturing devices. Information associated to images has got diversified applications. The medical imaging systems produce more and more digitized images in all medical fields. Most of these images are stored in image databases. There is a great interest to use them for diagnostic and clinical decision such as case-based reasoning. The purpose is to retrieve desired images from a large image databases using only the numerical content of images. CBIR system (Content-Based Image Retrieval) is one of the possible solutions to effectively manage image databases. Furthermore, fast access to such a huge database requires an efficient computing model. The Hadoop framework is one of the findings based on MapReduce distributed computing model. Lately, the MapReduce framework has emerged as one of the most widely used parallel computing platforms for processing data on terabyte and petabyte scales. It can able to provide more accurate results to the users. Therefore we are introducing a new method for retrieving images called as "CBIR over Hadoop using Map Reduce."

Keywords: CBIR, Hadoop, MapReduce, HDFS, HBASE, Parallel Computing, Distributed Computing.

I. INTRODUCTION

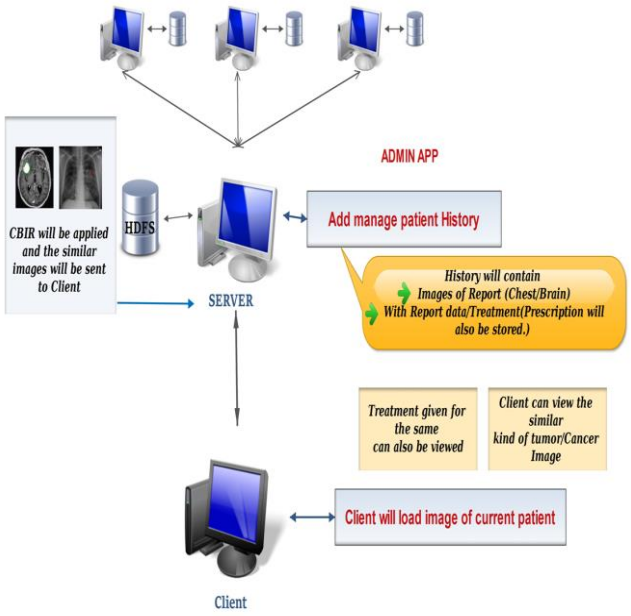
Nowadays, medical imaging systems produce more and more digitized images in all medical fields. Most of these images are stored in image databases. There is a great interest to use them for diagnostic and clinical decision such as case-based reasoning. The purpose is to retrieve desired images from a large image databases using only the numerical content of images. CBIR system (Content-Based Image Retrieval) is one of the possible solutions to effectively manage large image databases. Furthermore, fast access to such a huge database requires an efficient computing model. The Hadoop framework is one of the finding based on MapReduce distributed computing model. Lately, the MapReduce framework has emerged as one of the most widely used parallel computing platforms for processing data on terabyte and petabyte scales. Google, Amazon, and Facebook are the biggest users of the MapReduce programming model and it's been recently adopted by several universities. It allows distributed processing of data intensive computing over many machines. In CBIR systems, requests (the system inputs) are images and answers (outputs/results) are all the similar images in the database. A typical CBIR system can be decomposed in three steps: firstly, the characteristic features for each image in the database are extracted and are used to index images; secondly, the features vector of a query image is computed; and thirdly, the features vector of the query image is compared to those of each image in the database. For the definition and extraction of image characteristic features, many methods have been proposed, including image segmentation and image characterization using wavelet transform and Gabor filter bank. In this work, we used MapReduce computing model to extract features of images, then we write the features files into HBase (HBase is an open-source, distributed, versioned, column-oriented store modeled after Googles Big table).

II. ARCHITECTURE

The system architecture consists of the following important steps:

- Upload the medical images to HDFS.
- Take a medical image from HDFS and input it as Mapper.
- Extract the image features.
- Write the image and features in HBase.
- Complete the image processing in HDFS.
- Collect the output of Map Reduce phase.

1) The user sends a query image to system, then the image will be stored temporarily in HDFS.
2) Run a map-reduce job to extract features from query Image.
3) Store image features in HDFS.
4) The similarity/distance between the features vectors of the query image in HDFS and the target images in the HBASE are computed.
5) A reducer collects and combines all the result from all the map function.
6) The reducer stores the result into HDFS.
7) Send the result to the user.



III. HADOOP

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework. The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce). Hadoop splits large files into large blocks and distributes them amongst the nodes in the cluster. To process the data, Hadoop MapReduce transfers packaged code for nodes to process in parallel, based on the data each node needs to process. This approach takes advantage of data locality nodes manipulating the data that they have on hand-to allow the data to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel system where computation and data are connected via high-speed networking. The base Apache Hadoop framework is composed of the following modules:

Hadoop Commonly contains libraries and utilities needed by other Hadoop modules.

Hadoop Distributed File System (HDFS) - a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;

Hadoop HBase - HBase is a column-oriented database management system that runs on top of HDFS.

Hadoop MapReduce- A programming model for large scale data processing.

MAPREDUCE MODEL

The MapReduce framework that consists of subcategories as follows:

1) Map Processing: HDFS splits the large input data set into smaller data blocks (64 MB by default) controlled by the property dfs.block.size
2) Spill: When the buffer size reaches a threshold size controlled by io.sort.spill.percent (default 0.80, or 80%), a background thread starts to spill the contents to disk
3) Partitioning : Before writing to the disk the background thread divides the data into partitions

4) Sorting: memory sort is performed on key (based on the method of key class).
5) Merging: Before the map task is finished, the spill files are merged into a single partitioned and sorted output file
6) Compression: The map output can be compressed before writing to the disk for faster disk writing, lesser disk space, and reducing the amount of data to transfer to the reducer.
7) Reduce Operations: the reducer has three phases as following: Copy, Sort and Reduce.

IV. CONCLUSION

In CBMIR medical image retrieval is a data-intensive computing process, the traditional B/S single-node retrieval system has the defects of low efficiency and poor reliability and so on. Thus, a kind of Hadoop medical image retrieval system is put forward. The results of the simulation test show that the Hadoop medical image retrieval system improves the efficiency of the image storage and image retrieval, obtains a better retrieval result, and can satisfy the real-time requirements of the medical image retrieval. Especially when deals with the massive medical images, it has the advantages the traditional B/S single-node system cannot compared with. Therefore, the working focuses in the future are improving the transmission speed of data between the Map task and the Reduce task, reducing more time consumption which is due to the transfer of information, to further improve the execution efficiency of the existing image retrieval system

V. REFERENCES

[1] YAO Qing-An , ZHENG Hong , XU Zhong-Yu , WU Qiong , LI Zi-Wei , and Yun Lifen , "Massive Medical Images Retrieval System Based on Hadoop", JOURNAL OF MULTIMEDIA, VOL. 9, NO. 2, FEBRUARY 2014.

[2] Said Jai-Andaloussi, Abdeljalil Elabdouli, Abdelmajid Chaai, Nabil Madrane, Abderrahim Sekkaki, "Medical Content Based Image Retrieval by Using the HADOOP Framework", ICTEL. JANUARY 2013.

[3] Prof.DeeptiChikmurge," Implementation of CBIR Using MapReduce Over HADOOP", International Journal of Computer, Information Technology Bioinformatics (IJCITB) June 2014.

[4] WichianPremchaiswadi, AnuchaTungkatsathan, SarayutIntarasema, NuchareePremchaiswadi, "Improving Performance of Content-Based Image Retrieval Schemes using Hadoop MapReduce." (IJCITB) June 2014. 2008.

[5] Hinge Smita, Gaikwad Monika, Chincholkar Shraddha, "Retrieval of Images Using Map Reduce" International Journal of Advanced Research in Computer Science and Software Engineering, December 2014. Journal of Huazhong University of Science and Technology 2011.

[6] Byung Kwan Lee, EunHeeJeong ,"A Design of a Patient-customized Healthcare System based on the Hadoop with Text Mining (PHSHT) for an efficient Disease Management and Prediction" International Journal of Software Engineering and Its Applications Aug 2014.

[7] Sh. Akbarpour, A REVIEW ON CONTENT BASED IMAGE RETRIEVAL IN MEDICAL DIAGNOSIS, International Journal on Technical and Physical Problems of Engineering (IJTPE) June2013.