Open Journal of Marine Science, 2014, 4, 239-245
Published Online October 2014 in SciRes. http://www.scirp.org/journal/ojms
http://dx.doi.org/10.4236/ojms.2014.44022

Hadoop and Its Role in Modern Image Processing

Seyyed Mojtaba Banaei[1], Hossein Kardan Moghaddam[2*]

[1]Bozorgmehr University, Ghayen, Iran
[2]Birjand University of Technology, Birjand, Iran
Email: smbanaie@buqaen.ac.ir, *h.kardanmoghaddam@birjandut.ac.ir

Received 8 July 2014; revised 13 August 2014; accepted 3 September 2014

Abstract

This paper introduces MapReduce as a distributed data processing model using open source Hadoop framework for manipulating large volume of data. The huge volume of data in the modern world, particularly multimedia data, creates new requirements for processing and storage. As an open source distributed computational framework, Hadoop allows for processing large amounts of images on an infinite set of computing nodes by providing necessary infrastructures. This paper introduces this framework, current works and its advantages and disadvantages.

Keywords

Cloud Computing, Hadoop, Big Data, Image Processing, MapReduce Model

1. Introduction

The amount of image data has grown considerably in recent years due to the growth of social networking, surveillance cameras, and satellite images. However, this growth is not limited to multimedia data. This huge volume of data in the world has created a new field in data processing which is called Big Data that nowadays positioned among top ten strategic technologies [1].

Big Data refers to the massive amounts of data collected over time that are difficult to analyze and handle using common database management tools [2]. Due to the increasing amount of data that is becoming available every day and accelerated growth of information technology, one cannot provide a clear definition of the size and scale of big data. However, a multi-terabyte data sets (each terabyte = 1000 gigabytes) to multi-petabytes (each petabyte = 1000 terabytes) are considered as Big data. While big data can yield extremely useful information, it also presents new challenges with respect to how much data to store, how much this will cost, whether

*Corresponding author.

How to cite this paper: Banaei, S.M. and Moghaddam, H.K. (2014) Hadoop and Its Role in Modern Image Processing. *Open Journal of Marine Science*, **4**, 239-245. http://dx.doi.org/10.4236/ojms.2014.44022

the data will be secure, and how long it must be maintained [3]. The first approach to solve this problem was the use of multi-processor systems with high storage capacity. The rapid growth of data volumes and real-time processing needs and the complexity of software development and algorithms that can effectively utilize all processors prevented this approach to meet the new demands of data processing [3].

The next approach for processing large volumes of data and image was distributed systems with Message Passing Interface (MPI). Along with the idea of parallel data processing in distributed computing nodes and dissemination of data in each node, this approach promised a bright future for new data processing needs. However, the problem that this technique was faced with was parallel coordination and implementation of the required algorithms that completely depended on the system programmer and developer. Therefore, it was not widely embraced due to the lack of experts and professional developers [4].

Google as one of the leading companies in the field of big data, proposed the MapReduce programming model [5] which was designed to process large amounts of distributed data. The main advantages of this model are its simple programming structure, distributed file system, and distributed management which is failure resistant.

The main problem in the prevalence of this model was the provision of computing cluster for its implementation. It requires energy, cooling systems, physical space, necessary hardware and software for setting it up. These requirements are costly for many small and mid-size companies and enterprises [4].

This barrier has been resolved now by the popularity of cloud computing that provides consumers with low cost hardware and software based on the resource use. Just rent the number of computing nodes and the required resources when needed, then run your algorithm and get the result.

One of the well-known examples in this field is the generating PDF files from scanned daily archive of the New York Times in 2007. In this case 11 million photos with a volume of about 4 terabytes were converted to PDF only in 24 hours by using 100 nodes of Amazon Cloud Computing. This task would last for many years using common systems and algorithms [6].

In this paper, we introduce the MapReduce model as the basis of the modern distributed processing, and it's open-source implementation named Hadoop, the work that has been done in this area, its advantages and disadvantages as a framework for distributed processing, especially in image processing.

In this paper, it is assumed that the readers are familiar with cloud computing. In summary, cloud computing provides online computation and computer processing for users, without being worried about the number of required computers, resources, and other considerations. Users pay the cost based on the amount of resource consumption. Refer to source number [7] to learn more about this popular topic in modern information technology.

2. Hadoop, a Distributed Framework for Data Processing

Hadoop is an open source framework for processing, storage, and analysis of large amounts of distributed and unstructured data [8]. The main origin of this processing framework returns to Internet search companies, Yahoo and Google. They needed new processing tools and models for web page indexing and searching. This framework is designed for parallel data processing at Petabyte and Exabyte scales which are distributed on the normal compute nodes, such a way that a Hadoop cluster can easily be extended horizontally. Hadoop is currently developed and expanded by The Apache Foundation.

Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Cutting, who was working at Yahoo at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project [9].

The Apache Hadoop framework is composed of the following modules [9]:

- Hadoop Common 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 YARN—a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications. this part also named MapReduce 2.0 (MRv2)
- Hadoop MapReduce—a programming model for large scale data processing.

2.1. How Does Hadoop Work?

In this system, large data files, such as transaction log files, feed reader of social networks, and other data sources are segmented and then distributed in the network.

Sharing, storing, and retrieving large files on a Hadoop cluster is undertaken by its distributed file system called HDFS [10]. To increase the reliability of the system, each part of the file is distributed among multiple compute nodes. Therefore, if a node stops working, its file can be retrieved again.

There are three types of compute nodes in HDFS [10]. Name management node is responsible for sharing the files and storing the address of each part. Periodic review of nodes and determining their being phased out are also the tasks of Hadoop file management system.

Data node that encompasses each one of Hadoop member computers contains file blocks. There is a name management node in Hadoop system for each data node set. The third type is the secondary node that there is a copy of name management node data on it. Therefore if the node stops working, the data will not be lost. **Figure 1** shows an overview of Hadoop file management.

After data distribution in Hadoop system, analysis and processing would be carried out by the MapReduce part [11]. **Figure 2** shows this process visually. In the first steps, the user sends his/her request to a node which is responsible for running the requests (job tracker). This request usually is a Java query language. At this point, job tracker checks the files to see which one are needed for answering the user's query. Then by the help of name management node, it finds the nodes containing those parts in the cluster.

After that, this request is sent to each node. These nodes, that we call them task trackers, perform data processing independently and in parallel by running Map function [12]. After the task trackers' works is finished, the results will be stored on the same node. Obviously, the intermediate results would be local and incomplete because they depend on the data available on one node. After preparation of the intermediate results, the job tracker sends the Reduce request to these nodes. Therefore, it performs the final processing on the results and the result of user's request would be saved in a final compute node. At this point, MapReduce is finished, and further processing of the results should be performed by Big Data analysts. This processing can be performed directly on the results or classical methods of data analysis can be used by transferring the resulting data into a relational databases or data warehouse [13].

2.2. Libraries and Big Data Utilities

In addition to the above mechanism which is the basis of Hadoop in distributed and parallel data processing, a number of supplemental projects are designed for it. Big data processing may be done easier and more professional with the help of these projects. In this section, we look at these libraries and Hadoop utilities.

To store and retrieve information in Hadoop in a more professional manner, NoSQL databases such as *HBase* and *Cassandra* can be used. These types of databases support MapReduce mechanism and they are specifically designed to store and retrieve large amounts of unstructured data [13].

In addition to Java, users' requests can be written in an open source language called Pig which is designed specifically for Hadoop and also is relatively simple to learn.

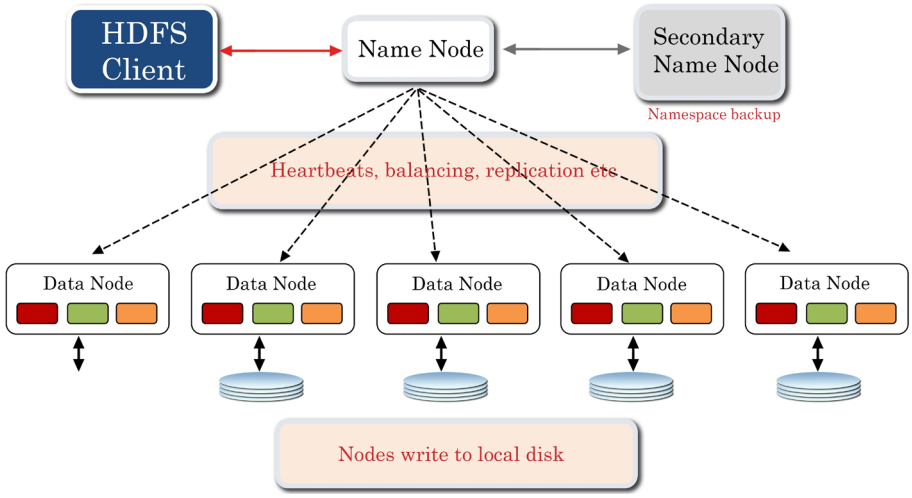

Figure 1. Structure of HDFS file system.

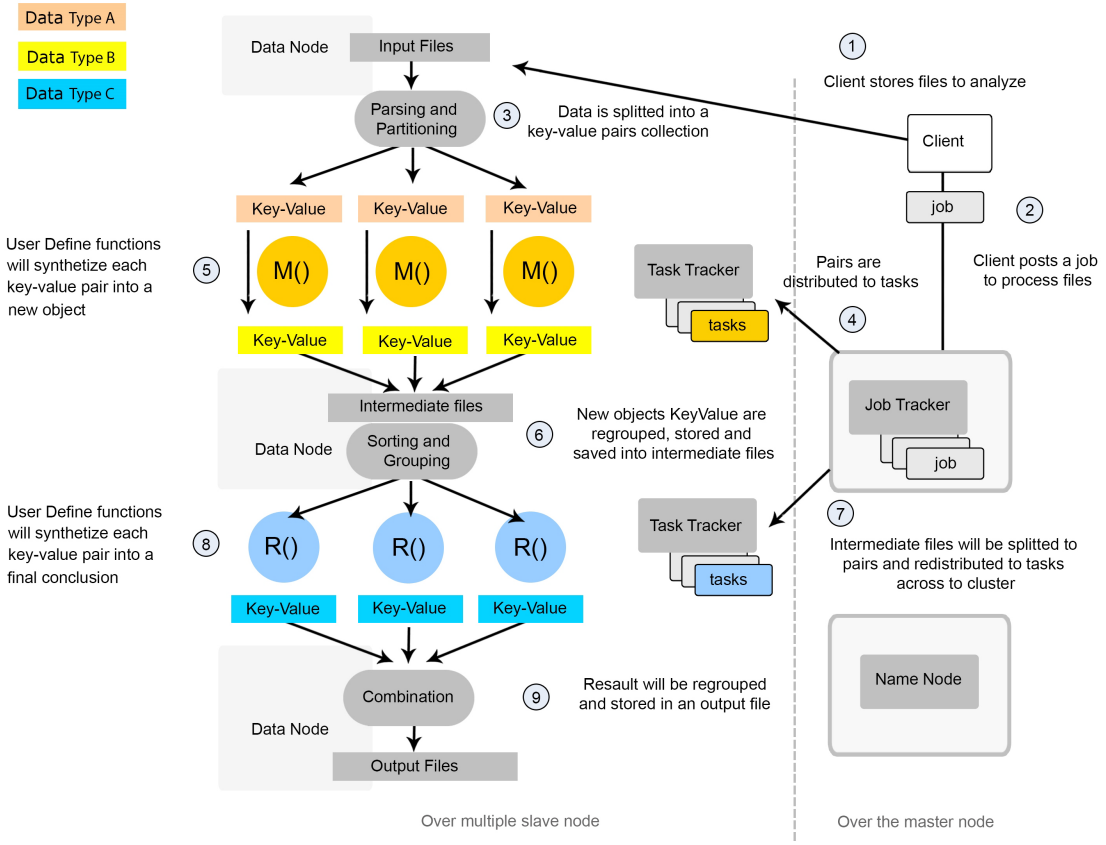

Figure 2. Hadoop operational structure and the MapReduce steps [14].

To collect data from different sources and save them in Hadoop, Flume framework was designed. Flume software agents collect data and send them instantly by placing on a web servers, mobiles, etc.

For sequential execution of the requests *Oozie* Library was considered. It allows users to run multiple requests sequentially and also each of the requests can use the output of the previous requests as an input. *Whirr* Library is recommended for running Hadoop on cloud computing systems and virtual platforms.

Mahout library has been designed for the purposes of data mining on Hadoop distributed platform. It considers the most common data mining algorithms, such as clustering and regression, as an input and converts them into proper MapReduce requests for Hadoop [13].

3. Works Carried out

Hadoop which is under the license of Apache with the support of this foundation is accessible to researchers as an open source framework. For using MapReduce models, in addition to Hadoop, we can also use Twister [15] and Phoenix [16] or other MapReduce style frameworks [17]. These two frameworks, that are both open source implementations of MapReduce model, are designed for specific purposes. Twister for iterative calculations and Phoenix for multi-processor systems with shared memory can be suitable alternatives for Hadoop. However, the emphasis in this paper is on Hadoop, because of its high popularity and its quality of being multi-purpose.

Li *et al*. [18] proposed a system for classifying ground imageries based on natural features that uses feature vectors approach and structured SVM in order to recognize natural features in any images. They used 6.5 million photos of Flickr and process them with the help of MapReduce programming model.

kennedy and the associated team [19] used visual Nearest Neighbor Method for automated labeling of images with high accuracy. Finding the visual Nearest Neighbor was performed with an image of the direct implementation of MapReduce method on 19.6 million images.

In order to determine the subject of images, Yan *et al*. [20] proposed the algorithm of extensibility based on MapReduce model which has been tested on 260,000 images.

For Content Based Image Retrieval (CBIR), Shi *et al*. [21] proposed a model based on MapReduce method which has been tested on 400,000 images.

Zhao, Li, and Zhou from Peking University conducted a research on the use of MapReduce model for satellite imagery documentation and management and spatial data processing [22]. They also proposed a system based on cloud computing.

Yang and the associated team [23] conducted a research concerning the use of Hadoop in the field of medical images. They designed a system named MIFAS for fast and efficient access to medical images using Hadoop and Cloud Computing.

For identification through cornea, Shelly and Raghava designed and implemented a system using Hadoop and Cloud Computing [24]. They tested it up to 13.2 GB input file and Hadoop has shown about 80% efficiency in high volume of data compared to conventional image processing methods.

Kucakulak and Temizel proposed a Hadoop-based system for pattern recognition and image processing of intercontinental missiles [25].

Almeer designed and implemented a system for remote sensing image processing systems with the help of Hadoop and cloud computing systems with 112 compute nodes [26].

4. Advantages and Disadvantages of Hadoop

The most important advantage of Hadoop is the ability to process and analyze large amounts of unstructured or semi-structured data which have been impossible to process efficiency (cost and time) so far [2].

The next advantage of Hadoop is its simple expansion and horizontal scalability. Data can easily be analyzed up to Exabyte level and there is no need for companies to work on sample data and a subset of the original data. With the help of Hadoop, the possibility of checking all types of data is provided.

Another advantage is its low set up cost, mainly because it is free and there is no need for expensive and professional hardware. In particular, with the spread of cloud computing and its reasonable prices for case processing of data as well as private clouds, it takes only a few hours to set up a Hadoop system [13].

On the other hand, Hadoop and its subsets are all in the early stages of development and they are unsteady and immature. This will lead to permanent modification of this framework that imposes costs of continuous training on organizations.

On the other hand, because of novelty of this software model, a few people have the necessary skills for establishing and working on Hadoop-based systems. Lack of expert manpower is the most important challenge of many companies in using this system.

Also the novelty of this technology causes the lack of valid standards and benchmarks for evaluating different algorithms in this area. Bajcsy *et al*. attempted to assess four different methods of Hadoop-based image processing on cluster [27]. This is one of the few efforts in this area and still we are far from establishing comprehensive benchmarks which are acceptable to academic community.

Another Hadoop's problem which has an inherent nature is lack of the ability of real-time data processing. The request tracker must wait for each compute node in the system to finish the work, and then it can deliver the final answer to the user. However, this problem will be solved to some extent by the rapid growth of NoSQL databases' technologies and its combination with Hadoop. Moreover, frameworks such as Storm [28] and Samza [29] can also be used for real-time processing of high volume data.

5. Conclusion

The huge volume of visual data in recent years and their need for efficient and effective processing stimulate the use of distributed image processing frameworks in image processing area. So that up to the coming years, many algorithms which have been introduced in the field of image processing and pattern recognition should consider the requirements for macro image processing in order to be welcomed by the outside world. This paper gives an overview of distributed processing methods and the programming models. Also some works are studied which have been done in recent years using Hadoop open source framework. Hadoop and its processing model are newly formed and like any other new technologies may have its own issues, such as lack of familiarity of the majority of IT society with it, lack of enough expert forces, and unwanted defects and problems due to its novelty. However, this processing style that uses MapReduce model and distributed file system, will be among the most useful tools for image processing and pattern recognition in the coming years due to its consistency with cloud computing structures.

References

[1] Gartner Reserch (2012) Top 10 Strategic Technologies for 2012. http://www.gartner.com/it/page.jsp?id=1826214

[2] Bakshi, K. (2012) Considerations for Big Data: Architecture and Approach. Aerospace Conference-Big Sky, MT, 3-10 March 2012.

[3] Michael, K. and Miller, K.W. (2013) Big Data: New Opportunities and New Challenges. *Journal of IEEE Computer Society*, **46**, 22-24.

[4] White, B., Tom, Y., Jimmy, L. and Davis, L.S. (2010) Web-Scale Computer Vision Using MapReduce for Multimedia Data Mining. *Proceedings of the* 10*th International Workshop on Multimedia Data Mining*, Washington DC, 25-28 July 2010, 1-10.

[5] Dean, J. and Ghemawat, S. (2008) MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, **51**, 107-114.

[6] The New York Times Blog (2007) http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun/

[7] Kalagiakos, P. (2011) Cloud Computing Learning. Application of Information and Communication Technologies (AICT). 5*th International Conference*, 12-14 October 2011.

[8] Hadoop. http://hadoop.apache.org/

[9] http://en.wikipedia.org/wiki/Apache_Hadoop

[10] HDFS. http://hadoop.apache.org/hdfs/

[11] MapReduce. http://en.wikipedia.org/wiki/MapReduce

[12] Bhandarkar, M. (2010) MapReduce Programming with Apache Hadoop. Parallel &Distributed Processing (IPDPS) IEEE, 19-23 April 2010.

[13] Kelly, J. (2012) Big Data: Hadoop, Business Analytics and Beyond. Wikibon Whitepaper, 27 August 2012. http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond

[14] http://hadooper.blogspot.com/

[15] Twister. http://www.iterativemapreduce.org/

[16] The Phoenix System for MapReduce Programming. http://mapreduce.stanford.edu/

[17] Qura Question and Answer Website. http://www.quora.com/What-are-some-promising-open-source-alternatives-to-Hadoop-MapReduce-for-map-reduce

[18] Li, Y., Crandall, D.J. and Huttenlocher, D.P. (2009) Landmark Classification in Large-Scale Image Collections. *ICCV*, 1957-1964.

[19] Kennedy, L., Slaney, M. and Weinberger, K. (2009) Reliable Tags Using Image Similarity: Mining Specificity and Expertise from Large-Scale Multimedia Databases. *Proceedings of the* 1*st Workshop on Web-Scale Multimedia Corpus*, Beijing, 23-23 October 2009, 17-24.

[20] Yan, R., Fleury, M.-O., Merler, M., Natsev, A. and Smith, J.R. (2009) Large-Scale Multimedia Semantic Concept Modeling Using Robust Subspace Bagging and MapReduce. *Proceedings of the* 1*st ACM Workshop on Large-Scale Multimedia Retrieval and Mining*, Beijing, 23-23 October 2009, 35-42.

[21] Shi, L.L., Wu, B., Wang, B. and Yan, X.G. (2011) Map/Reduce in CBIR Application. 2011 *International Conference on Computer Science and Network Technology* (*ICCSNT*), Vol. 4, Harbin, 24-26 December 2011, 2465-2468.

[22] Zhao, J.Y., Li, Q. and Zhou, H.W. (2011) A Cloud-Based System for Spatial Analysis Service. 2011 *International Conference on Remote Sensing*, *Environment and Transportation Engineering* (*RSETE*), Nanjing, 24-26 June 2011, 1-4. http://dx.doi.org/10.1109/RSETE.2011.5964031

[23] Yang, C.-T. and Chen, L.-T., Chou, W.-L. and Wang, K.-C. (2010) Implementation of a Medical Image File Accessing System on Cloud Computing. 2010 *IEEE* 13*th International Conference on Computational Science and Engineering* (*CSE*), Hong Kong, 11-13 December 2010, 321-326. http://dx.doi.org/10.1109/CSE.2010.48

[24] Shelly and Raghava, N.S. (2011) Iris Recognition on Hadoop: A Biometrics System Implementation on Cloud Computing. 2011 *IEEE International Conference on Cloud Computing and Intelligence Systems* (*CCIS*), Beijing, 15-17 September 2011, 482-485. http://dx.doi.org/10.1109/CCIS.2011.6045114

[25] Kocakulak, H. and Temizel, T.T. (2011) A Hadoop Solution for Ballistic Image Analysis and Recognition. 2011 *International Conference on High Performance Computing and Simulation* (*HPCS*), Istanbul, 4-8 July 2011, 836-842. http://dx.doi.org/10.1109/HPCSim.2011.5999917

[26] Almeer, M.H. (2012) Cloud Hadoop MapReduce For Remote Sensing Image Analysis. *Journal of Emerging Trends in Computing and Information Sciences*, **3**, 637-644.

[27] Bajcsy, P., Vandecreme, A., Amelot, J., Nguyen, P., Chalfoun, J. and Brady, M. (2013) Terabyte-Sized Image Computations on Hadoop Cluster Platforms. 2013 *IEEE International Conference on Big Data*, Silicon Valley, 6-9 October 2013, 729-737.

[28] Storm Project. http://storm.incubator.apache.org/

[29] Samza Project. http://samza.incubator.apache.org/

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.