International Journal of Advancement in Engineering Technology, Management & Applied Science

Volume 3, Issue 4 April 2016 ISSN No: 2349-3224

A Methodological Study on Identification of Duplicate Images using MapReduce Technique

Geetha M.V[1], Padma Priya M.K[2]

[1] M.Tech (Department of Computer Science and Engineering), New Horizon College of Engineering, Bangalore, India.

[2]Asst Professor (Department of Computer Science and Engineering), New Horizon College of Engineering, Bangalore, India.

***Abstract*— Big data contains large quantity of data. As a result accessing and managing the data is very difficult. The central theme of this concept is to use lossless information deduplication technique which makes the big data small, thereby providing benefits for accessing and managing the data. In order to achieve this goal, big data is subjected to deduplication at block level and map reduce techniques. The deduplication technique provides non redundant data which helps in retrieving the information faster. The map reducing technique preserves the integrity of the original information.**

***Keywords*— Deduplication; MapReduce technique; Big data; Hashing function; cloud.**

I. INTRODUCTION

The big data is a technology that deals with large volume of data in petabytes and more, that includes both structured data and unstructured data. The MapReduce is a programming model which is used to split the image file into blocks and combines it while downloading the image. The MapReduce program is a combination of both Map () method and Reduce () method. In the map reduce technique the user can have a (key, value) pairs that generates a set of intermediate key and value. Also reduce function which makes use of all these intermediate keys.

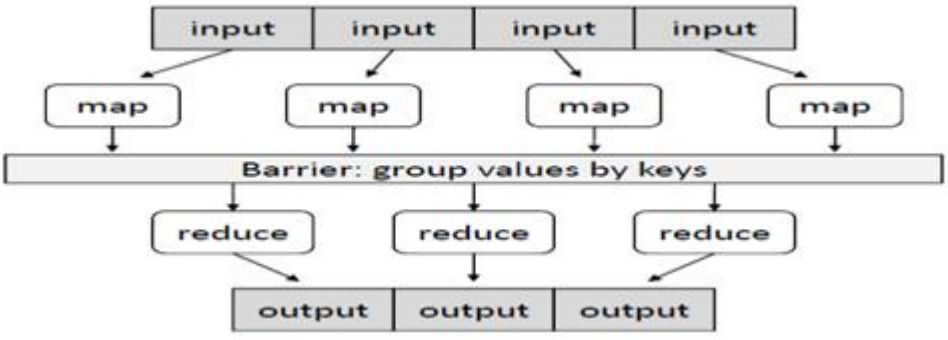

Fig 1: MapReduce Framework

In the Fig 1, it shows the working of the map reduce methods, each input is given to a map method it will generates an (key, value) pairs and then the reduce method will merge the obtained results and produces the output.

The deduplication is a technique which reduces the storage space and eliminates the extra copies of repeating blocks of image there by storing a single copy of image block. In many situations, the users upload the same copy of image many number of times in to cloud. As a result, the same copy of image is stored many times in the cloud, the capacity of memory storage increases and redundant copies of image will be stored. Due to that the retrieving the information in big data is very difficult.

II. RELATED WORK

[1] Explains the working of HIPI an image processing library on MapReduce framework. The library is designed in such way that, it hides the implementation of complex Hadoop MapReduce framework. The implementation is done by considering huge amount of data, because of this system gives higher throughput in case amount of images exceeds. MapReduce pipeline has provision of different formats for accessing the images. The types of images that can be used during MapReduce steps are filtered by providing the culling phase during the mapping phase. Float images, most important part in image processing are obtained by using the encoders and decoders phases which presents behind the scene. By adding all these features in the system it gives simplified interface to deal with the images on MapReduce.

[2] Presents a novel system of image deduplication which makes use of high precision duplication approach. The proposed system comprises of five stages as feature extraction, high-dimension indexing, accuracy optimization, and centroid selection and deduplication evaluation by evaluating the system on real datasets it had been observed that system not only gives the efficient image deduplication scheme but also greatly improves the precision of duplicate image retrieval.

[3] Elaborates the method of finding the nearby images. To accomplish the task the queries which are being popular are taken and the commercial search service to gather the images which are normally analyses as nearby images. As the removing such nearby images from the repository is practically not feasible hence the proposed work removes the nearby images from the search answers. By evaluating the system with many real world queries it had been found that the system gives the promising results compare to the traditional techniques under the same category. To bring down the idea into reality (DPF, PCA-SIFT, and HBC algorithms are being used which significantly performs better than the other.

[4] Gives a detailed survey on the various deduplication strategies being used. Various issues presents in deduplication schemes such as bandwidth, high throughput, computational overhead, deduplication efficiency, cost of transmission, usability of read and write operations are discussed here. So by observing this discussion one can choose the best technique for their application. Different deduplication schemes such as application based source deduplication scheme, a semantic attribute based source deduplication, GPU based source data deduplication, Hadoop based data duplication, hash level based deduplication, and causality based deduplication are explained in detail with their advantages and disadvantages.

[5] Illustrates the HDFS based deduplication framework by designing the techniques such as RFD-HDFS and FD-HDFS. RFD-HDFS is best suit for the application which are related with the finance where there is no chances of errors whereas FD-HDFS can be used in applications which accepts few amounts of errors. The experimental evaluation shows that space consumed by duplicate data is reduced greatly and the performance of the uploaded files are affected by the integrated schemes.

[6] Presents a search theory which shows how the map reduce technique is used at different works of Google. Also authors state the reason behind this. First is, it is simple to use. Even the programmer with less knowledge of parallel and distributed systems can use it effectively. It presents the work scenario in abstract way by hiding the details of load balancing, fault tolerance and parallelization. And the second is huge real word scenarios are effectively expressed using this. E.g. map reduce is effectively used at Google in web search for storing, sorting and data mining. Finally authors conclude that the map reduce can be effectively used for the keeping data without its loss.

International Journal of Advancement in Engineering Technology, Management & Applied Science
Volume 3, Issue 4 April 2016 ISSN No: 2349-3224

III. PROPOSED SYSTEM ARCHITECTURE

In this paper, we are proposing a deduplication technique at block level using MapReduce programming model for mapping and reducing the image blocks while storing in to the cloud.

The MD5 algorithm is used for generating the hash code values for corresponding blocks. The pseudo code steps for MD5 algorithm as follows

Step 1: Get the Message
Step 2: Convert message in to the bits.
Step 3: Append Padding Bits (Make the message bit length should be the exact multiple of 512 bits as well as 16 word Blocks).
Step 4: Divide total bits in to 128 bits blocks each
Step 5: Initialize MD Buffer.
A four word buffer (A,B,C,D) is used to compute the message digest , total 128 bits.
Step 6: Do AND,XOR,OR,NOT operations on A,B,C,D by giving three inputs and get one output.
Step 7: Do the Step 6 until get the 128 bits hash(16 bytes).
Step 8: Stop

In our proposed system architecture, the uploaded image file is split into blocks based on the size of image. The hashing function will be applied for every block and the resultant hash code will be generated compared against with the existing blocks hash code. If the block already exists we are mapping to that block to achieve deduplication. Finally, non redundant copies of blocks will be stored in the cloud.

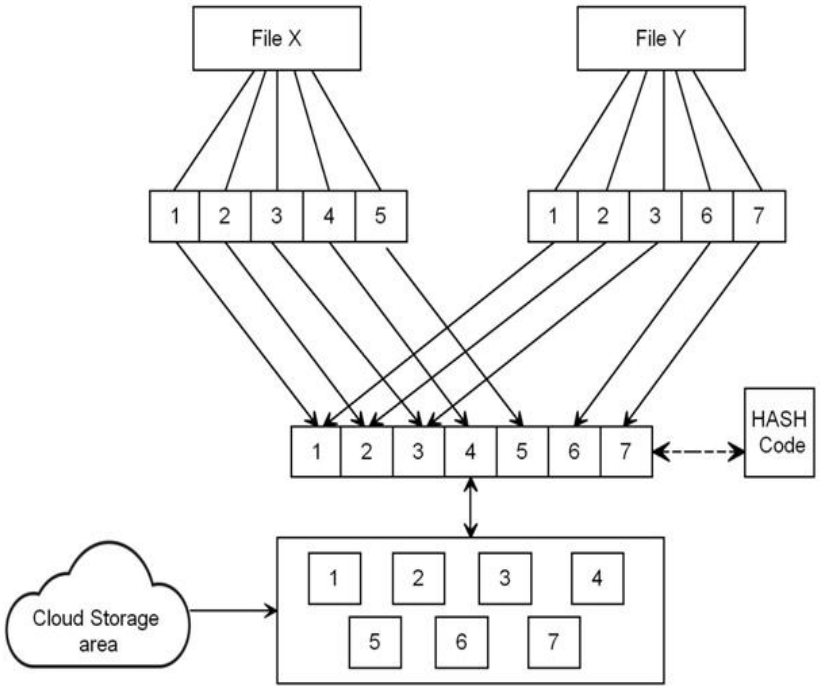

Fig 2: Proposed system architecture

In the Fig 2, two image files are uploaded by the user will split into blocks and corresponding hash codes will be generated and stored. Whenever an new image file is uploaded by the user is compared against with the hash code of already existing blocks, if the

block already exist then the mapping is done and then the non redundant blocks are stored into the cloud.

IV. CONCLUSION

In this paper, we propose image deduplication at block level using MapReduce model, the accessing and managing the data easier, it makes the information retrieval faster and reduces the access time. Thereby provides benefits in processing and management of the big data.

REFERENCES

[1] Sweeney, Chris, et al. "HIPI: a Hadoop image processing interface for image-based mapreduce tasks." Chris. University of Virginia (2011).
[2] Chen, Ming, Shupeng Wang, and Liang Tian. "A high-precision duplicate image deduplication approach." *Journal of Computers* 8.11 (2013): 2768- 2775.
[3] Foo, Jun Jie, et al. "Detection of near-duplicate images for web search."*Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007.
[4] Neelaveni, P., and M. Vijayalakshmi. "A Survey on Deduplication in cloud storage." *Asian Journal of Information Technology* 19.6 (2014): 320-330.
[5] Sheu, Ruey-Kai, et al. "Design and Implementation of File Deduplication Framework on HDFS." *International Journal of Distributed Sensor Networks*2014 (2014).
[6] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM* 51.1 (2008): 107-113.