International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Methodological Survey on MapReduce for Identification of Duplicate Images

Amol S. Deshmukh[1], Prof. P. D. Lambhate[2]

[1]M.E (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India.
Savitribai Phule Pune University, Pune, Maharashtra, India -411007

[2]Professor (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India.
Savitribai Phule Pune University, Pune, Maharashtra, India -411007

Abstract: ***Duplicate image identification for deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization by avoiding duplicate data. With the explosive growth of digital data, deduplication schemes are widely employed to backup data and minimize network and storage overhead by detecting and eliminating redundancy among data. In this paper we propose the Duplicate image identification using MapReduce technique which improves efficiency and reliability of the System. MapReduce is simple and parallel computing techniques normally used for analyzing the huge data. Traditional deduplication schemes works if and only if the second image having the same underlying bits as first. This restricts the performance of many applications as exact images need to be there if want to succeed. In many practical applications where the storage restriction is present, users uploads the modified images varying with the quality or resolution. Experimental results demonstrate in a real dataset, the proposed approach not only effectively saves storage space, but also significantly improves the retrieval precision of duplicate images. In addition, the selection of the images can meet the requirements of people's perception.***

Keywords: Duplicate image identification, Deduplication, MapReduce technique, big data, data partitioning, Pearson Correlation

1. Introduction

The amounts of the images being uploaded daily are increasing tremendously. By the survey performed on 2010, the 2.5 billion new images are being uploaded to the Facebook daily. However the system making use of all this data re observed of having some lags as the operational data is of huge amount. Hence it becomes merely impossible to deal with such huge amount of data. Once Hadoop comes to the practice it overcomes the said drawbacks. Hadoop is an open source MapReduce platform used for the distributed processing of the data. Still Hadoop deals with many technical flaws while developing useful applications. So in order to overcome the issue Hadoop Image Processing Interface came in practice (HIPI). HIPI creates interface between the system and the MapReduce.

Deduplication is a well-known technique of reducing the size of data storage by avoiding the storage of duplicate files. Traditional deduplication schemes works if and only if the second image having the same underlying bits as first. This restricts the performance of many applications as exact images need to be there if want to succeed. In many practical applications where the storage restriction is present, users uploads the modified images varying with the quality or resolution.

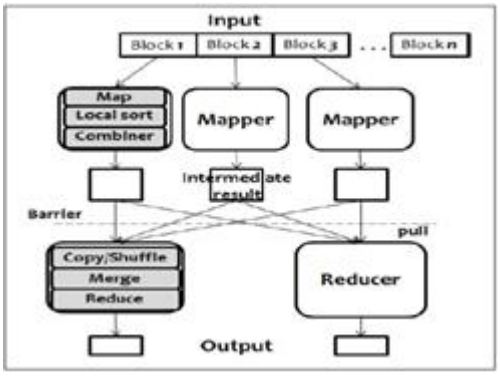

Figure: Hadoop architecture

Figure shows the inner implementation of Hadoop storage system. Generalized image deduplication is performed in five stages as discussed below.

1) Feature extraction.
2) Indexing.
3) Optimization
4) Select centroid.
5) Find duplicates.

Image deduplication and text deduplication are the two schemes normally used to save the space. Table explains the difference between these two techniques.

Table: Difference between text and image deduplication

Text Deduplication	Image Deduplication
Read text file	Read image file
Data partition	Image preprocess
Hash computation	Feature extraction
index lookup: an exact matching	index lookup: an approximate matching, D1<=T
Accuracy optimization: comparison byte by byte	Accuracy optimization: compare the number of the same elements
Storage one copy	Centroid selection and storage the centroid-image

Data duplication strategies are normally classified under three main units as shown above.

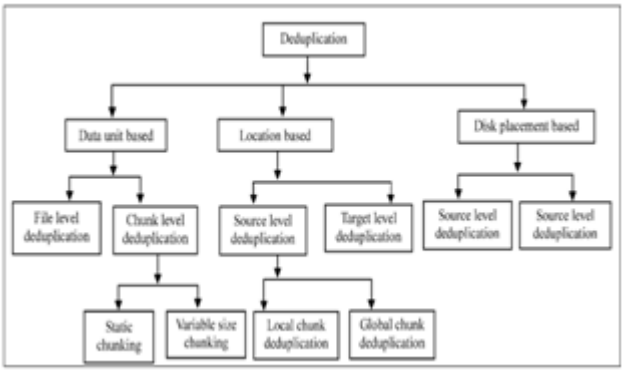

Figure: Deduplication strategies

1) Data units based deduplication
   Here only one copy of file having same content is maintained. File sharing the same hash are assumed as same files.
2) Location based deduplication
   In this scheme deduplication is done by considering the location where deduplication is performed. This method can greatly save the storage space but it fails to save the bandwidth cost.
3) Disk placement based deduplication
   Here data duplication is done by observing the way by which data is placed on the disks.

MapReduce Technique

MapReduce is one of the best, simple and parallel computing techniques normally used for analyzing the huge data. The main motto of MapReduce technique is to hide the way by which partitioning takes place and thus to focus on the technique of data processing. In map reduce technique users can have a key/value pair that can generate set of intermediate key and value. Also reduce function is created which makes use of all these same intermediate keys. Number of huge real world tasks can be effectively done using this technique. Main side of the technique is the programs written using this model is automatically paralyzed which increase the speed of the execution. MapReduce makes use of Google File System (GFS) as a base layer of storage from which data can be take and store. GFS comes under chunk based data partitioning where the level of fault tolerance is reduced by using replication and data partitioning algorithms. Apache Hadoop is an open source framework of MapReduce.

Exactly like MapReduce, Hadoop consist of two implementation layers.
1) HDFS (Hadoop DFS)
2) HMF (Hadoop MapReduce)

HDFS layer comprises of data storage facility and HMF consists of data processing techniques.

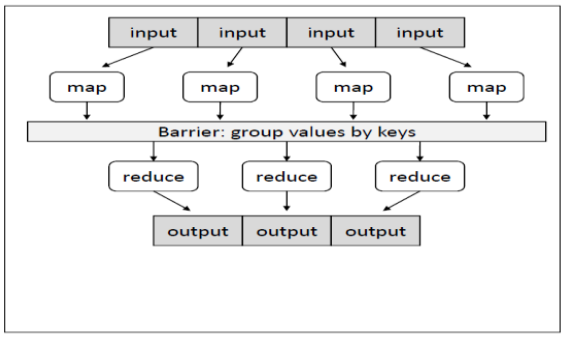

Figure: Technique of MapReduce Framework.

Data Partitioning

Partitioning techniques are emerged as a one of the best techniques to rid out of the problem time required for the execution of the system.

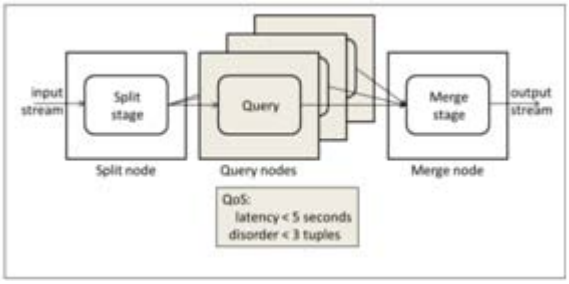

Figure: Parallel stream processing architecture overview.

In parallel partitioning systems, generally the typical split and then merge pattern is used. Here two nodes are presents, one represent split node and other represents merge node. Split node is used for splitting the things for further execution. The splitting is done in such way that node unbalancing is not happened. The intermediate part on which partitioned data is needed to feed is fixed in advance. After execution of the individual node the result is again gathered at merge node. Merge node combines the all result and then pass it as a output of the system.

The partitioning techniques are normally categorized in two techniques.
1. Batch based partitioning
2. Pane based partitioning.

In batch based partitioning system next to next tasks are grouped to form a batch and thus gives it to the common partitioning system. While in pane based partitioning exactly opposite task of batch based partitioning is done. Here instead of breaking the main task into sub task, the sub tasks are breaking down to the small tasks. Then these tasks are assigned individually to the partitioning.

2. Literature Survey

[1] Explains HIPI an image processing library on MapReduce framework. The designing of library is done in such way that it hides the implementation of complex Hadoop MapReduce framework and emphasis more on image as it is the thing about which users worrying a lot. 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 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.

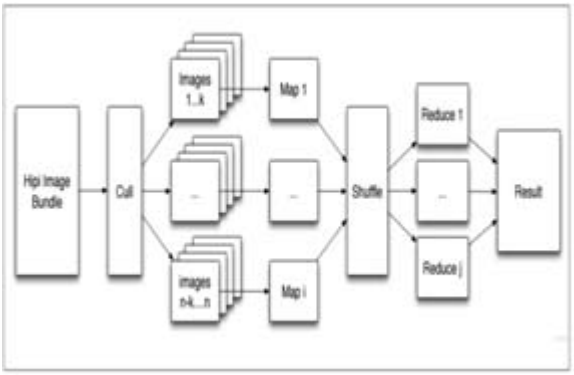

Figure: MapReduce pipeline architecture

As discussed above traditional deduplication systems are performed well if and only if the images to be compared are having same underlying bit codes. But this scenario reduces the usability of applications. So to overcome this [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. The performance of the system is putted in following table.

Indicator	Real dataset
Deduplication rate	9.7%
Deduplication accuracy	98.8%

As the huge amount of duplicate images are available on web, web search for a particular images tends to give the number of nearby images also which degrades the performance of the system. [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.

In order to shows the performance of different searching and sorting task on the system having different configurations a useful study is presented by the [7]. To bring down this idea into reality Hadoop and map reduce technique for distributed data processing technique is used. Here the machine learning problems classes are distinguished within the map reduce framework to improve the implementation of Hadoop. At the final part of the system they makes statement that the map reduce technique is a best option for the simple operations but still it has many flaws for the complex operations over large database.

[8] Elaborates the cost cutting solution by making use of MapReduce technique in place of model building algorithms in statistical machine translation. Without MapReduce technique same task can be accomplished by parallelizing the task but it increases the cost of hardware which in turn increases the software cost. On 20-machine cluster system gives excellent performance and also it does not require the cost burden of hardware.

K-Nearest neighbor processing on large datasets have great impact on the performance of system. [9] Considers above problem as a base for their work and proposed a technique which combines the map reduce and locality sensitive hashing (LSH). These combination gives good performances as the mapping phases of the map reduce technique have provision hashing principle of LSH. Apart from this different problems of map reduce and LSH are briefly explained by the author. To evaluate the performance of the

system both flickr.com dataset and synthetic datasets are considered. Real compute cluster is kept as a future work by the writer as they currently working on the same.

[10] Discussed the techniques of removing near duplicate images from the image dataset. To use the traditional deduplication technique proposed system presents the image in visual word. As the visual word representation of the image loses all the geometric features, result may have higher false positive rate when the size of the dataset increases. To increase the discriminability of the visual word images local image features are used which groups the visual word images. Difference of Gaussian is taken for the purpose of feature point detection.

[11] Presents a scalable data partitioning techniques for the purpose of data streams processing. Traditional schemes used for the splitting of data are failed to achieve high degree of scalability which degrades the performance of the system and thus increase time and cost complexity of the system. So to overcome the problem, [11] finds a good technique. Here to alternative partitioning techniques are proposed: partitioning based on batch and pane based partitioning. Out of the number of techniques discussed above, pane based partitioning gives good result. To show the experimental performance the system is tested against the linear load benchmark. Also it gives fewer loads on the load which is splitting the data. The current work does not bother about the fault tolerance of the system. This issue is kept as a future work by the writers.

As the world is facing the problem of managing the huge amount of data, number of techniques are proposing to get rid of these. Recent survey conducted by IBM shows that approximately 2.5 quintillion bytes of data are being daily generated. This data comprises of many formats such as images, videos, social media site opinions, sensor data, transactional data etc. it is practically infeasible to deal with this data. Hence from last decade MapReduce is evolved as promising framework to deal with this Huge amount of data. [12] Gives a detail survey on the family of MapReduce framework. As the main advantage of the MapReduce is to give scalable applications, it has been used in many levels from academics to the industry. Authors try to presents the complete theories behind the map reduce: the reason behind use, the way by which it can be used, databases supporting MapReduce etc.

3. Problem Definition

The main purpose of this proposed system is to reduce the time required to identify the duplicate image in storage server using map reducing technique that is been powered with correlation technique.

4. Architectural View

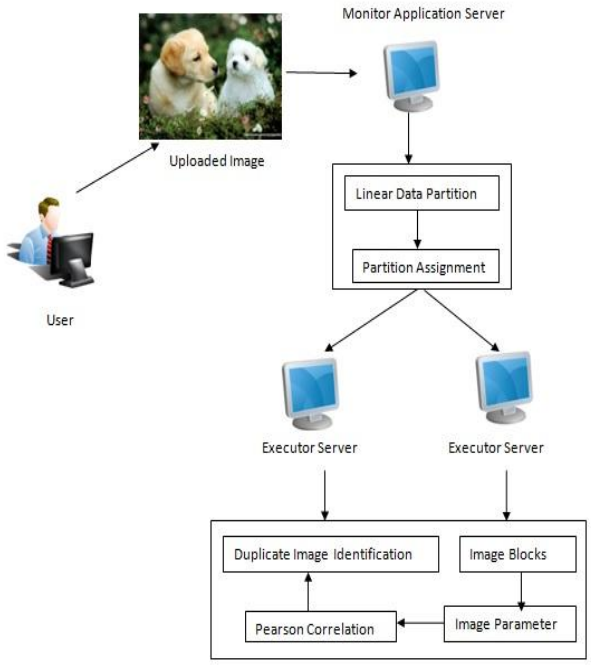

Figure: System Architecture

Step 1: First user creates his profile by entering his/ her personal data like first name, last name, DOB, Address, Email id, Contact number and username etc. and user enters his username and password to login into the account for uploading and downloading images on/from the web server.

Step 2: User uploads the image file which implicitly avoiding for the duplications by our system. The uploaded image will be accessed by the monitor server.

Step 3: Monitor Application Server is responsible for access and store image from user to Executor Server by duplicate image identification.

Step 4: The data partition has used for Storage data partition to empower multi threads.

Step 5: The image parameter is use for Calculation of mean and standard deviation based on Image RGB Values for correlation calculation.

Step 6: The Pearson Correlation is use for calculate the correlation between two images (i.e. user uploading image and stored image) based on image parameter by comparing image block.

Step 7: The Duplicate image identification is done based on Pearson correlation technique by comparing image blocks for avoiding the duplicate image.

5. Conclusion

As this complete paper narrates different methodology for duplicate image identification, but none of the methodology are seems to be perfect. So In this paper, we proposed the Duplicate image identification systems for eliminating duplicate images for improve storage utilization. Duplicate image identification System avoids the duplicate image to store on the storage server, where we stores huge amount of

data. MapReduce technique is used for fasten duplicate image detection process. Duplicate image detection process is done based on Pearson correlation technique through the computation of mean and standard deviation. This System reduce the time required to identify the duplicate image in storage server using map reducing technique that is been powered with correlation technique.

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.
[7] Gillick, Dan, Arlo Faria, and John DeNero. "Mapreduce: Distributed computing for machine learning." *Berkley, Dec* 18 (2006).
[8] Dyer, Christopher, et al. "Fast, easy, and cheap: Construction of statistical machine translation models with MapReduce." *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2008.
[9] Stupar, Aleksandar, Sebastian Michel, and Ralf Schenkel. "RankReduce processing k-nearest neighbor queries on top of MapReduce."*Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval*. 2010.
[10] Wen, Tzay-Yeu. "Large Scale Image Deduplication."
[11] Balkesen, Cagri, and Nesime Tatbul. "Scalable data partitioning techniques for parallel sliding window processing over data streams." *International Workshop on Data Management for Sensor Networks (DMSN)*. 2011.
[12] Sakr, Sherif, Anna Liu, and Ayman G. Fayoumi. "The family of MapReduce and large-scale data processing systems." *ACM Computing Surveys (CSUR)*46.1 (2013): 11.

Author Profile

Amol S. Deshmukh, is currently pursuing M.E (Computer) from Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India. Savitribai Phule, Pune University, Pune, Maharashtra, India -411007. He received his B.E. (Information Technology) Degree from BIGCE, College of Engg. Solapur, Solapur University, Solapur, Maharashtra, India -413255 in 2014. His area of interest is Cloud Computing and Parallel Computing.

Prof. P. D. Lambhate, received her Degree from WIT, Solapur, ME (Comp) from BVCOE Pune, Pursing PhD. In computer Engineering. She is currently working as Professor at Department of Computer and IT , JayawantraoSawant College of Engineering, Hadapsar, Pune, India 411028,affiliated to Savitribai Phule Pune University, Pune, Maharashtra, India -411007. Her area of interest is Data mining, search engine.