International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal | www.ijtsrd.com
No: 2456 - 6470 | Volume - 3 | Issue – 1 | Nov – Dec 2018

Leveraging Data Duplication to Improve The Performance of Storage System with CLD and EHD Image Matching in the Cloud

Pooja M. Khandar[1], Prof. S. C. Tawalare[2], Prof. Dr. H. R. Deshmukh[3], Prof. S. Dhole[3]
[1]Student, [2]HOD, [3]Professor
Department of Computer Science & Engineering, DRGIT&R, Amravati, Maharashtra, India

ABSTRACT
With the explosive growth in data volume, the I/O bottleneck has become an increasingly daunting challenge for big data analytics in the Cloud. In existing paper, propose POD, a performance-oriented deduplication scheme, to improve the performance of primary storage systems in the Cloud by leveraging data deduplication on the I/O path to remove redundant write requests while also saving storage space. This research works aims to remove data duplication in the cloud. Improve the performance of storage system. We use concept of image processing to utilize the space. In this paper we discussed about the design and implementation of data duplication to improve the efficiency of storage in cloud. This system, implements wireless data access to servers. An alternative method for us is remove the data duplication in storage system by using web based application in which we can use two matching technic CLD (color layout descriptor) and EHD (enhance histogram descriptor). User can browse image and upload the image on web page then we apply CLD & EHD technic and then see uploaded image is already store on cloud or not, if there is matching image like uploaded image then we extract referenced of already store image then send to the receiver and receiver can receive the image. If there is no matching image then upload new image to database. By extracting reference of already store image there is no need to upload again same image to database so, we can remove data duplication, improve the storage space efficiency and utilize network bandwidth so, our system more effective than the data duplication to improve the performance of primary storage system.

Key Words: *Java JDK 6.0, Eclipse, Apache tomcat server, MY-SQL Database.*

1. INTRODUCTION
Data duplication often called intelligent compression or single instance storage. it is processes that eliminates redundant copies of data and reduce storage overhead.

Data deduplication technique insures that only one unique instance of data is retained on storage media, such as 1) disk 2) flash or tape. Data deduplication has been demonstrated to be an effective technique in Cloud backup and archiving applications to reduce the backup window, improve the storage-space efficiency and network bandwidth utilization. Recent studies reveal that moderate to high data redundancy clearly exists in virtual machine (VM) enterprise [3], [4], [5], [6], [7] and high-performance computing (HPC) storage systems [8]. CLD and EHD techniques, performance oriented deduplication scheme, to improve the performance of storage systems in the Cloud by leveraging data deduplication requests while also saving storage space. In this paper we discussed about the design and implementation of data duplication to improve the efficiency of storage in cloud.

2. Literature Review:
In existing system when we are uploading the files in to the system, if that file is already existed in that system then that file will not be uploaded and instead of that the reference will be created so that if number of times one file referenced to many files if by chance that file has deleted then we will loss the reference of the all files so for that reason we are creating the copies of that files in the multiple locations of the system memory. So if one file is deleted from the system memory other locations will maintain the copy of that file. By using Secure Hash Table Technique[1].

In another existing paper, propose POD, a performance-oriented duplication scheme, to improve the performance of primary storage systems in the Cloud by leveraging data duplication on the I/O path to remove redundant write requests while also saving storage space

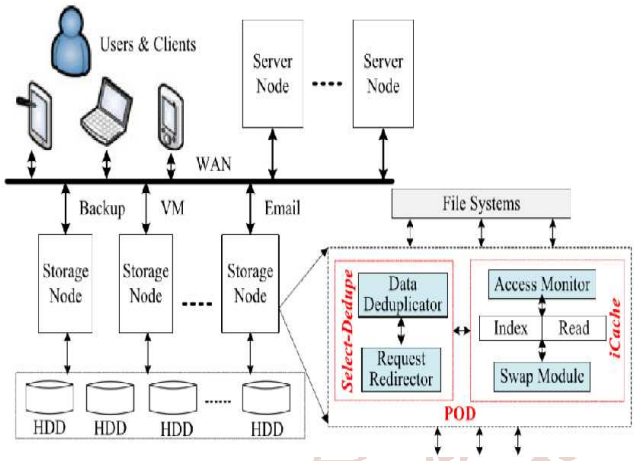


Fig 1: System architecture of POD

POD resides in the storage node and interacts with the File Systems via the standard read/write interface. Thus, POD can be easily incorporated into any HDD-based primary storage systems to accelerate their system performance. POD is independent of the upper file systems, which makes POD more flexible [5], [6].

POD has two main components: Select-Dedupe and iCache. The request-based Select-Dedupe includes two individual modules: Data Deduplicator and Request Redirector. The Data Deduplicator module is responsible for splitting the incoming write data into data chunks, calculating the hash value of each data chunk, and identifying whether a data chunk is redundant and popular. The Request Redirector module decides whether the write request should be deduplicated, and maintains data consistency to prevent the referenced data from being overwritten and updated. The iCache module also includes two individual modules: Access Monitor and Swap Module [2]. The Access Monitor module is responsible for monitoring the intensity and hit rate of the incoming read and write requests. The Swap module dynamically adjusts the cache space partition between the index cache and read cache. Moreover, it swaps in/out the cached data from/to the back-end storage.

3. Proposed Objective

In this paper we used two techniques for finding duplication of the image. There are two techniques:

1. Color Layout Descriptor
2. Edge Histogram descriptor

Color layout descriptor:-

Is designed to capture the spatial distribution of color in an image .the feature extraction process consist of two parts;

1. Grid based representative color selection.
2. Discrete cosine transform with contization.

The functionality of CLD is basically the matching -Image to image matching

CLD is one of the most precise and fast color descriptor [8].

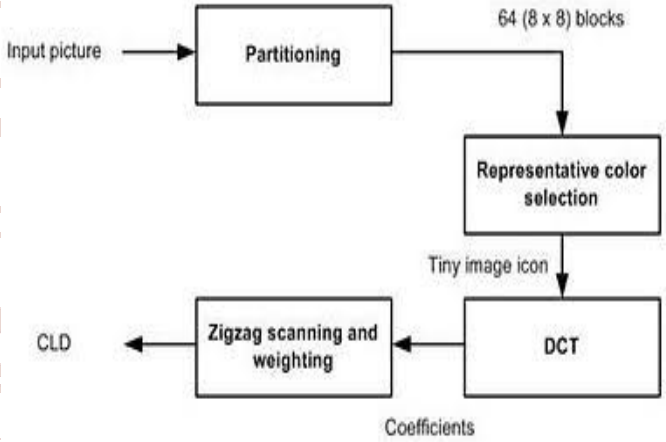


Fig 2: color layout descriptor

Edge histogram descriptor:-

The edge histogram descriptor (EHD) is one of the widely used methods for shape detection. It basically represents the relative frequency of occurrence of 5 types of edges in each local area called a sub-image or image block. The sub image is defined by partitioning the image space into 4x4 Non-overlapping blocks as shown in figure 1. So, the partition of image definitely creates 16 equal-sized blocks regardless of the size of the original image. To define the characteristics of the image block, we then generate a histogram of edge distribution for each image block. The edges of the image block are categorized into 5 types: vertical, horizontal, 45-degree diagonal, 135-degree diagonal and non-directional edges, as shown in Figure 2. Thus, the histogram for each image block represents the relative distribution of the 5 types of edges in the corresponding sub-image [8].

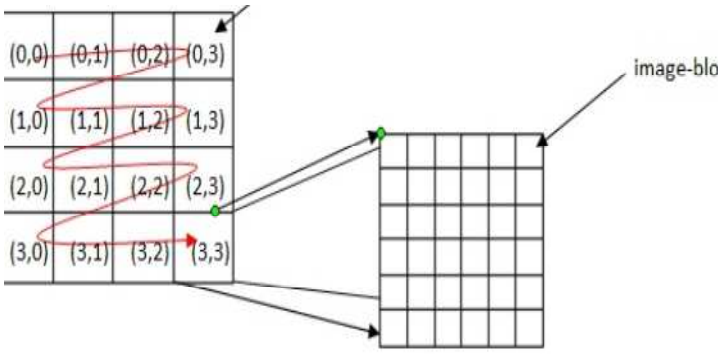


Fig 3: Definition of Sub-image and Image-block in the EHD

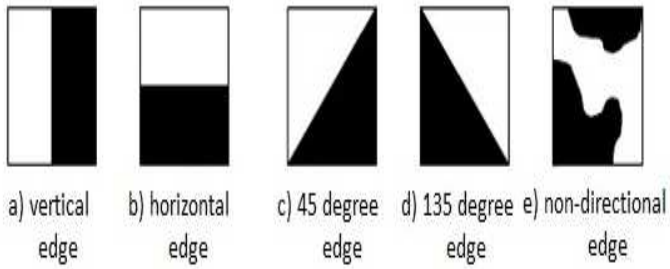


Fig2. Five Types of Edges in EHD

4. Implementation:-

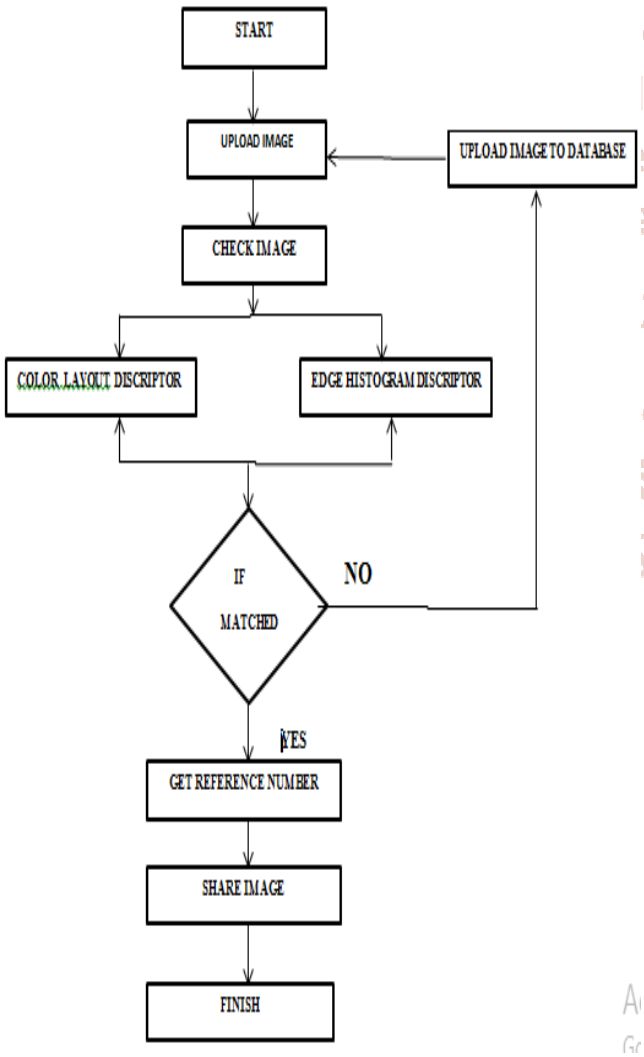
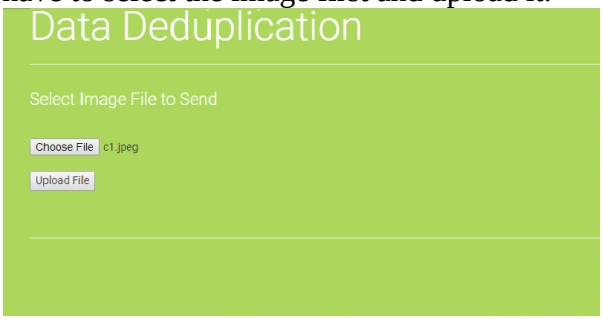


Fig4: flowchart

Module 1:
We have to select the image first and upload it.



Screenshot 1: Select image

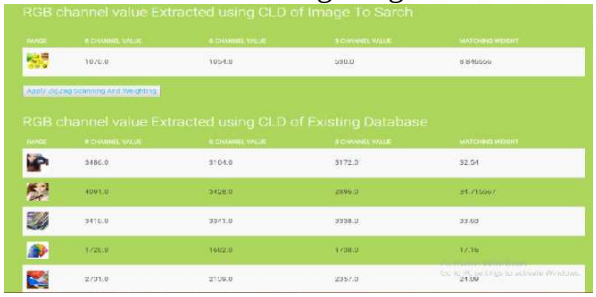
Module 2:
Then apply Color Layout Descriptor on image.

Screenshot 2: Apply CLD Technique

Module 3:
In this partitioning the original image into 4*4 matrixes

Screenshot 3: Image Partitioning

Module 4:
It generates RGB value of the targeted image and check values with the existing images in the database.



Screenshot 4: Check RGB Values

Module 5:

It gives us result of topmost matching images using RGB values and apply EHD.

Screenshot 5: CLD result and Apply EHD

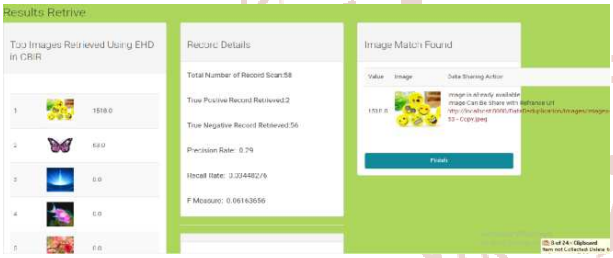
Module 6:

After applying edge histogram descriptor it generates the edges values of targeted images and match with existing image in database.

Screenshot 6: Check edges value

Module 7:

It gives us result of topmost matching images using edges values.



Screenshot 7: EHD result

5. **Technical Specifications And Result Analysis**:

The technologies which are used to implement the system are:

- Java jdk.6.0
- Eclipse: In computer programming, Eclipse is an integrated development environment (IDE). It contains a base workspace and an extensible plug-in system for customizing the environment. Written mostly in Java, Eclipse can be used to develop applications. By means of various plug-ins, Eclipse may also be used to develop applications in other programming languages: Ada, ABAP, C, C++, COBOL, Fortran, Haskell, JavaScript, Lasso, Lua, Natural, Perl, PHP, Prolog, Python, R, Ruby (including Ruby on Rails framework), Scala, Clojure, Groovy, Scheme, and Erlang.
- MySQL is open source relational database system. It is static. Database size is unlimited in MySQL. MySQL support Java. MySQL does not support except & intersect operation. MySQL does not have resource limit. MySQL is available under GPL proprietary license. The MySQL development project has made its source code available under the term of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL is a popular choice of database for used in web application. MySQL is written in C and C++.

6. Advantages:

- It require less storage as it is data duplication application.
- It saves time.
- Efficient and fast access.

7. Disadvantages:

- Required Internet:

For the total execution of this project required the internet.

8. Conclusion:

In this paper, we propose CLD and EHD techniques, a performance oriented duplication scheme, to improve the performance of storage systems in the Cloud by leveraging data duplication requests while also saving storage space. In this paper we discussed about the design and implementation of data duplication to improve the efficiency of storage in cloud. This system, implements wireless data access to servers. An alternative method for us is remove the data duplication in storage system by using web based application.

9. References:

1. k. Lavanya, Dr. A. Sureshbabu, “Data Reduction using A Dedplication Aware Resemblance Detection & Elimination Scheme” International Journal of Advance Research in Computer Science and Management Volume 5, Issue 8, August 2017.
2. Bo Mao, Hong Jiang, Suzhen Wu and Lei Tian , “Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud ” IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016.

3. A. T. Clements, I. Ahmad, M. Vilayannur, and J. Li, “Decentralized deduplication in SAN cluster file systems,” in Proc. Conf. USENIX Annu. Tech. Conf., Jun. 2009.
4. K. Jinand and E. L. Miller, “The effectiveness of deduplication on virtual machine disk images,” in Proc. The Israeli Exp. Syst. Conf., May 2009.
5. D. T. Meyer and W. J. Bolosky, “A study of practical deduplication,” in Proc. 9th USENIX Conf. File Stroage Technol., Feb. 2011.
6. K. Srinivasan, T. Bisson, G. Goodson, and K. Voruganti, “iDedup: Latency-aware, inline data deduplication for primary storage, ”in Proc. 10th USENIX Conf. File Storage Technol., Feb. 2012.
7. A. El-Shimi, R. Kalach, A. Kumar, A. Oltean, J. Li, and S. Sengupta, “Primary data deduplication-large scale study and system design,” in Proc. USENIX Conf. Annu. Tech. Conf., Jun. 2012.
8. D. Meister, J. Kaiser, A. Brinkmann, T. Cortes, M. Kuhn, and J. Kunkel, “A study on data deduplication in HPC storage systems, ”in Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal., Nov. 2012.
9. www.wikipedia.com