International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 3

A Review Paper on Leveraging Data Duplication to Improve the Performance of Storage System with CLD and EHD Image Matching in the Cloud

Pooja M. Khandar
Department of Computer Science & Engineering,
DRGIT&R, Amravati, Maharashtra, India

Dr. H R. Deshmukh
HOD, Department of Computer & 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 deduplication to improve the performance of primary storage system.

Keywords: *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 and high-performance computing (HPC) storage systems. 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:

1. 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.

2. K. Jinand and E. L. Miller, "The effectiveness of deduplication onvirtual machine disk images," in Proc. The Israeli Exp. Syst. Conf., May 2009.

3. D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," in Proc. 9th USENIX Conf. File StroageTechnol.,Feb. 2011.

4. 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.

5. 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.

6. S. Kiswany, M. Ripeanu, S. S. Vazhkudai, and A. Gharaibeh,"STDCHK: A checkpoint storage system for desktop gridcomputing," in Proc. 28th Int. Conf. Distrib. Comput. Syst., Jun. 2008.

7. 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.

8. X. Zhang, Z. Huo, J. Ma, and D. Meng, "Exploiting data deduplicationto accelerate live virtual machine migration," in Proc. IEEEInt. Conf. Cluster Comput., Sep. 2010.

2. PROPOSED OBJECTIVE

What is data deduplication in cloud computing?

Data duplication often called intelligent compression or single instance storage. It is process that eliminates redundant copies of data and reduces storage overhead. Data deduplication technique insure that only one unique instance of data is retained on storage media, such as 1)disk 2) flash or tape Redundant data block are replace with pointer to a unique data copy in that way data deduplication closely align with incremental backup which copies only the data that has changed since the previous backup.

For example typical e-mail system might contain 100 instances of same 1 MB file attachment if the email platform is backed up or archived. All 100 instances are saved, requiring 100 MB of storage space. With data duplication only one instance of attached stored: each subsequent instance is referenced back to the one saved copy therefore in this example 100 MB storage demands drop to 1MB.

3. IMPLEMENTATION

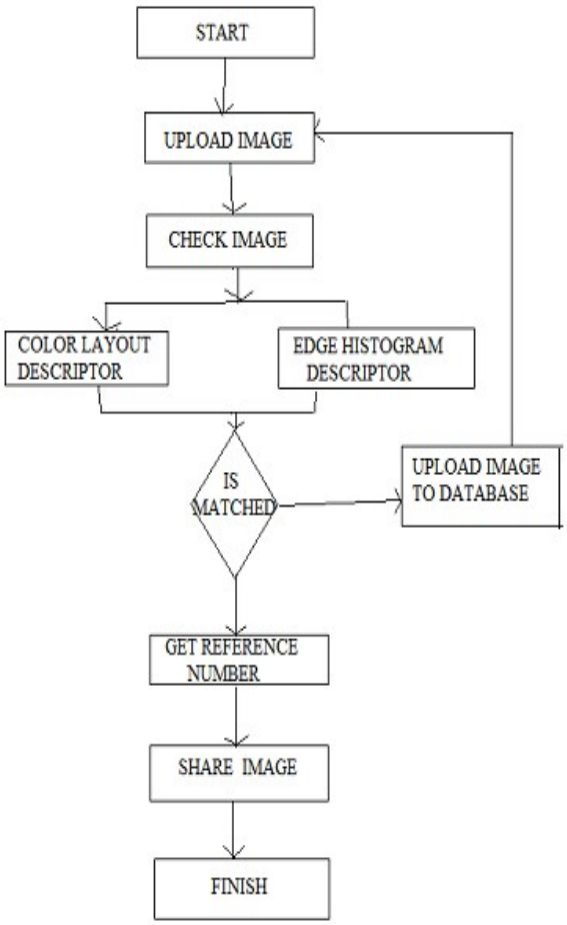

Fig 1: flowchart

The following techniques are used in data duplication to improve the performance of storage system in cloud.

1. Color layout descriptor(CLD)
2. Edge histogram descriptor(EHD)

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

-Video clip to video clip matching

Remark that the CLD is one of the most precise and fast color descriptor

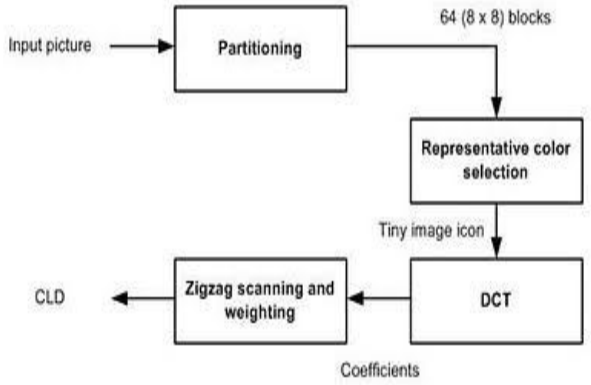

Fig 2: color layout descriptor

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 4x4Non-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.

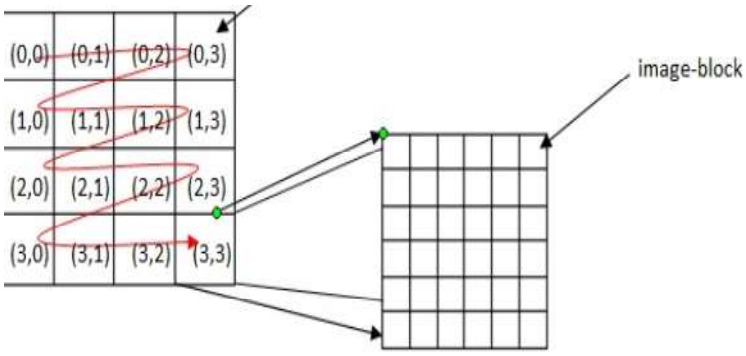

Fig 3: Definition of Sub-image and Image-block in the EHD

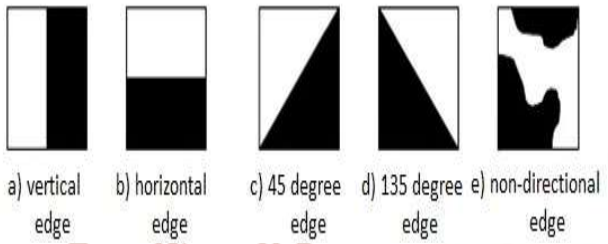

Fig 2: Five Types of Edges in the EHD

4. 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 requires less storage as it is data duplication application.
- Efficient and fast access.

7. Conclusion:

In this paper, we propose CLD and EHD techniques, a 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. 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.

REFERENCES:

1. A. T. Clements, I. Ahmad, M. Vilayannur, and J. Li,"Decentralized deduplication in SAN cluster file systems," inProc. Conf. USENIX Annu. Tech. Conf., Jun. 2009.
2. K. Jinand and E. L. Miller, "The effectiveness of deduplication onvirtual machine disk images," in Proc. The Israeli Exp. Syst. Conf.,May 2009.
3. D. T. Meyer and W. J. Bolosky, "A study of practicaldeduplication," in Proc. 9th USENIX Conf. File Stroage Technol.,Feb. 2011.
4. 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.
5. 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.
6. S. Kiswany, M. Ripeanu, S. S. Vazhkudai, and A. Gharaibeh,"STDCHK: A checkpoint storage system for desktop gridcomputing," in Proc. 28th Int. Conf. Distrib. Comput. Syst., Jun.2008.
7. D. Meister, J. Kaiser, A. Brinkmann, T. Cortes, M. Kuhn, andJ. Kunkel, "A study on data deduplication in HPC storage systems,"in Proc. Int. Conf. High Perform.Comput., Netw., StorageAnal., Nov. 2012.
8. X. Zhang, Z. Huo, J. Ma, and D. Meng, "Exploiting data deduplicationto accelerate live virtual machine migration," in Proc. IEEEInt. Conf. Cluster Comput., Sep. 2010.
9. www.wikipedia.org/history.
10. http://en.wikipedia.org/wiki