Middle-East Journal of Scientific Research 24 (S2): 292-295, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.24.S2.184

De-Duplication of Files in Cloud System Based on Finger Printing Methodology

R. Vinston Raja, P. Madhusudhana, K. Tarakanitesh and R. Rakesh

Department of IT, Panimalar Institute of Technology, India

Abstract: Now days, there is a continuous and exponential increase of the number of users and the size of their data in cloud storage. So, data de-duplication becomes a necessity for users. Today's Cloud system facing the problem of increasing quantity of data. To make the consistent cloud system, data duplication deduction technique will be very helpful. Proposed system will help to keep only one copy of data in cloud system by using finger printing methodology. System gives an additional option to the user that is he/she can view the already uploaded file when they try to upload the same file. Thereby, we can achieve better space and sufficient bandwidth services.

Key words: De-duplication · Finger printing methodology · IAAS · PAAS · SAAS and Content defined Chunking · CBIR · Video replica detection · Penal functions

INTRODUCTION

The rapid growth of data across multiple platforms has drastically increased the demand of Cloud storage infrastructures. Cloud Storage is a representation of data storage where the digital data is stored in virtualized logical pools, more specifically saying [1], it is stored in multiple physical storage which spans multiple servers and the physical environment is owned, administered and managed by a hosting company. More specifically saying the Cloud Storage Providers are responsible for keeping track of the data availability and accessibility.

Cloud computing involves both software and hardware which are distributed to users as a service by the service providers [2]. With the rapid growth in the domain of cloud computing, even more and much versatile services are emerging up. But the basic domains can be illustrated informs of services such as,

- IaaS - Infrastructure as a Service
- PaaS - Platform as a Service
- SaaS - Software as a Service

IAAS (Infrastructure As A Service) :

- The base layer
- Deals with Virtual Machines, Storage (Hard Disks), Servers, Network, Load Balancers etc

PAAS (Platform As A Service) :

- A layer on top of IAAS
- Environment or platform (like JAVA, .NET, PHP), Databases (like Myself, Oracle), Web Servers (Tomcat, Apache etc)

SAAS (Software As A Service):

- A layer on top on PAAS
- Applications like email (Gmail, Yahoo mail etc), Social Networking sites (Face book etc)

In our proposed system IAAS includes the common things like storages, networks etc. PAAS includes PHP as a Platform, My Sql as a Database and Apache as a Web server. SAAS includes Email as an example.

Data De-Duplication: Data de-duplication refers to methodologies that store only a single copy of redundant data and thereby provide a single copy. By eliminating redundant data both disk space and bandwidth [3]. With respect to service providers, it offers secondary cost savings in power and cooling which is achieved by reducing the number of spindles.

It ensures that only one copy of data is stored in the data enter. Therefore it clearly decreases the size of data enter. So it basically means that the number of the

Corresponding Author: R. Vinston Raja, Department of IT, Panimalar Institute of Technology, India.

replicated copies of data that were usually duplicated on the cloud server can be controlled and managed easily to shrink the physical storage space [4]. The recent statistics has found out that de-duplication is the most influential storage technology and is predicted to provide 75% of all backups in the next few years.

Types of De-Duplication

- Source De-duplication
- Target De-duplication

Source De-Duplication: Source de-duplication works through client software that communicates with the backup server to compare new blocks of data with previously stored blocks of data. If the server has previously stored a block of data, the software does not send that block and instead notes that there is a copy of that block of data at that client. If a previous version of a file has already been backed up, the software will compare files and back up any parts of the file it hasn't seen. Source de-duplication is well suited for backing up smaller remote backup sets.

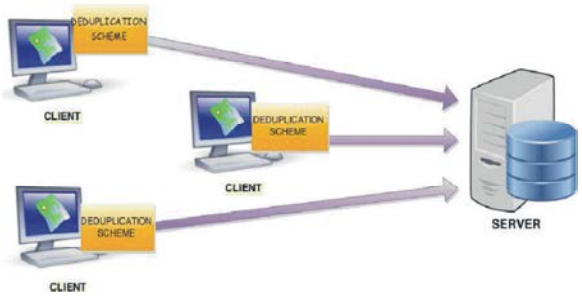

Fig. 1: Source De-duplication

Target De-Duplication: Target de-duplication requires hardware at remote sites usually called server, provides faster performance for large data sets; it is typically employed by companies with larger datasets. The main advantage of using target de-duplication is, it is enough to create single software for a server rather than creating software for multiple devices in source de-duplication.

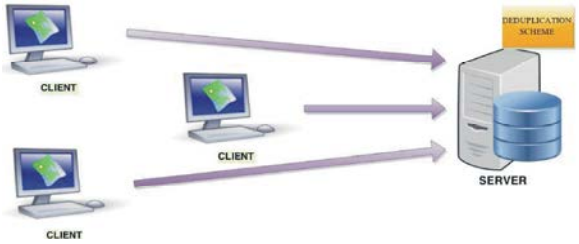

Fig. 2: Target De-Duplication

Finger Printing Methodology: Finger printing involves the following methodology like the sequence of chunking, duplicate detection and storage. The first methodology known as “Chunking? involves the splitting of data into non-overlapping blocks called “chunks”. The processing of one chunk is independent of the other chunk. The following are the two commonly used strategies with respect to chunking- Static Chunking and Content-defined Chunking.

The first approach i.e. Static Chunking involves splitting of data into chunks which are always similar in size. The size of Chunk is always a multiple of the disk sector or the block size of the file system. This is also known as Fixed –size Chunking or Fixed block Chunking or Fixed Chunking.

The next approach i.e. Content defined Chunking does not involve splitting of data into chunks of similar size. But this is usually preferred more in backup systems because it is not prone to the “boundary- shifting effect”, which reduces the redundancy found by the data de-duplication systems. This mainly occurs due to the fact that the data gets slightly shifted i.e. other data was inserted into the main data stream.

As a result, the conventional approach of static chunking is unable to identify the duplication because the chunks are not similar. The content-defined chunking overcomes this situation by realigning the chunking methodology with the content and the similar chunks are created as before and thus duplication can be identified.

Content-defined chunking is found to be give high de-duplication ratio with respect to the backup workloads. Hence, it is the most widely used methodology in most de-duplication systems for backup workloads.

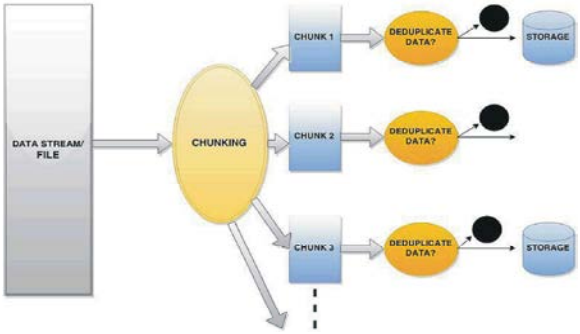

Fig. 3: De-Duplication Methodology for Videos

De-Duplication Methodology for Videos: This presents a video replica detection and non- replicated video Retrieval system which evolves and uses different video similarity measures for different users query image using

patio-temporal pattern index. Specifically, a user supplied query video allows the system to determine which subset of a set of objective features approximates more efficiently the subjective image similarity of a specific users query. The following algorithm describes the steps involved in the proposed system, which includes the RIFT based feature extraction, this explores the use of rotation-invariant feature transform (RIFT) features to model the subjective perception of similarity between two frames that have been extracted from a video database.

The following algorithm describes the steps involved in the proposed system, which includes the RIFT based feature extraction.

Algorithm:

Input: Input files (F) Key Frame (K)
Output: Pattern descriptor (P)

Step 1: Segment the video frames and extract features of the key frames. The first step performs the temporal video segmentation method to segment the video sequences into frames.

Frames [K] =Split (FT)

Step 2: Then extracts RIFT features of the key frames. Dense RIFT technique has been applied in this step. This Collects more features at each location and scale in a frame, this helps at increasing recognition accuracy accordingly.

Step 3: Store the frame in the database and construct a PI. For each pattern p from pattern set

Step 4: Perform the step 2 for the query video or frame.

Step 5: Match the query video and target video by ST_PI method.

This performs the following steps:

Spatio-temporal pattern indexing and matching.
Weight assignment based on the feature.
Performs visual and textual similarity matching

Step 6: Retrieve the set of frames matched for the query frame.

Step 7: Perform the motion matching alignment scheme for video making from the retrieved frames

Step 8: Use RIFT_PI descriptor and Patterns for alignment.

De-Duplication Methodology for Audio: As said previously, our proposed system will be developed in PHP. There are some pre-defined functions to implement the audio file de-duplication. They are listed below.

- openal_buffer_create — Generate OpenAL buffer
- openal_buffer_data — Load a buffer with data
- openal_buffer_destroy — Destroys an OpenAL buffer
- openal_buffer_get — Retrieve an OpenAL buffer property
- openal_buffer_loadwav — Load a .wav file into a buffer
- openal_context_create — Create an audio processing context
- openal_context_current — Make the specified context current
- openal_context_destroy — Destroys a context
- openal_context_process — Process the specified context
- openal_context_suspend — Suspend the specified context
- openal_device_close — Close an OpenAL device
- openal_device_open — Initialize the OpenAL audio layer
- openal_listener_get — Retrieve a listener property
- openal_listener_set — Set a listener property
- openal_source_create — Generate a source resource
- openal_source_destroy — Destroy a source resource
- openal_source_get — Retrieve an OpenAL source property
- openal_source_pause — Pause the source
- openal_source_play — Start playing the source
- openal_source_rewind — Rewind the source
- openal_source_set — Set source property
- openal_source_stop — Stop playing the source
- openal_stream — Begin streaming on a source

Fig. 4: De-Duplication Methodology for Audio

De-Duplication Methodology for Images: Content based image retrieval (CBIR) was introduced in the early1980s. CBIR uses visual contents to search images from large scale image database as per users? query. “Content based” means that the search will analyzes the actual contents either colour, texture, shapes or spatial locations of the images. Chang published a pioneering work in 1984, in which a picture indexing and abstraction approach for pictorial database retrieval is used.

Low level Image Features: CBIR system is based on feature extraction, image features can be extracted globally for the entire image or locally for regions. There are two types of features.

- General features
- Domain specific features

General features which are application independent which can be easily retrieved from the image database by using query and domain specific features which are application dependent such as human face, fingerprints and conceptual features.

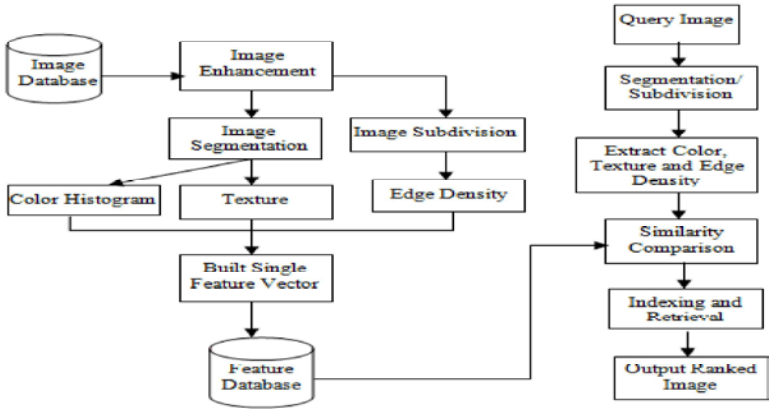

Fig. 5: Flowchart of Content Based Image Retrieval

REFERENCES

1. A Survey on Data Reduplication in Cloud Storage Environment - Manikantan U.V, Prof.Mahesh G.
2. Review on Content Based Duplicate Image Detection - Jag tap Ankita K, Tidke B.A.
3. Video Replica Detection and Localization Using Sequence Pattern Analysis Algorithm with Spatio-Temporal Pattern Index - T.Sivakumar, Rinju.M.
4. Audio files De-duplication [Online] Available: http://php.net/manual/en/ref.openal.php.