ISBN 2320-2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.5 , No.3, Pages : 12-18 (2016)
Special Issue of ICDCB 2016 – Held during April 14-15, 2016 in Kuala Lumpur, Malaysia
http://warse.org/IJACST/static/pdf/Issue/icdcb2016sp03.pdf

Deduplication Image Middleware Detection Comparison In Standalone Cloud Database

Nadiah bt Yusof[1] , Amirah Ismail[2] and Nazatul Aini Abd Majid[3]
[1,2,3]Faculty of Information Science and Technology
National University of Malaysia (UKM), 43600 Selangor, Malaysia
nadiahsof@siswa.ukm.edu.my, amirahismail@ukm.edu.my, nazatulaini@ukm.edu.my

Abstract: Research in image deduplication detection on internet has increasingly since 2008. Image storage in online database are too many and innumerable, including the same image uploaded by the user repeatedly and it is called as deduplication images. The big amount of deduplication images in database can lead waste of memory space for cloud database usage provided by the provider. This could make a user pays more for memory usage in cloud storage for similar image. Deduplication image will be reduced by using deduplication image detector either by plugin, middleware or software. However, there still lacks of research on deduplication image detector software or plugin for cloud storage. This is because, many researchers emphasize more on deduplication image detector in a standalone database. This paper compares standalone image deduplication detector, to identify a detail about technique using in deduplication detection and a relevant detection element of image deduplication. A new framework for deduplication detection in cloud has been proposed in this paper to begin early into the research image deduplication detection in cloud storage.

Key words: Image deduplication, cloud, database, Multimedia system, Multimedia information retrieval

INTRODUCTION

Multimedia is increasingly becoming the "biggest big data" as the most important and valuable source for insights and information. It covers from everyone's experiences to everything happening in the world. There will be a lots of multimedia big data surveillance video, entertainment and social media, medical images, consumer images, general image, voice and video [1].

To name a few, only if their volumes grow to the extent that the traditional multimedia processing and analysis systems cannot handle effectively [1]. Among them is that there is dumping a duplicate image in online either normal database or database in cloud. Database in cloud storage can be classified as an image or data storage in a database managed by the service provider for cloud storage such as Amazon Elastic Compute Cloud (EC2) has 6521 public virtual machine image [2], images or data operate independently of the cloud [3].

Some cloud platforms offer options for using a database as a service [2], without physically launching a virtual machine instance for the database. In such a configuration, application owners do not have to install and maintain the database themselves. Instead, the database service provider takes responsibility for installing and maintaining the database, and application owners pay according to their usage. For example, Amazon Web Services provides three database services as part of its cloud offering: SimpleDB, a NoSQL key-value store; Amazon Relational Database Service, a SQL-based database service with a MySQL interface; and DynamoDB. Similarly, Microsoft offers the Azure SQL Database service as part of its cloud offering [4].

Based on observation, emphasis in the use of software image or plugin detector for cloud storage deduplication still less do in image deduplication detector and the studies in image deduplication detector has been done in other research in standalone database. Hence, this study discussed the pilot test conducted on the image deduplication detector for a standalone database and the result for pilot test will be discuss in the preliminary studies section.

This paper has been divided into six part, introduction, literature review, preliminary studies, discussion, conclusion and future work

LITERATURE REVIEW

Data deduplication is a data compression techniques for removing duplicates copies of identical data and it is used in cloud storage to save bandwidth and to reduce the amount storage space. The technique is utilized to enhance the storage use and can likewise be applied to network data exchange to reduce the amount of bytes that must be sent. Keeping multiple copies with the identical content, deduplication removes redundant data by keeping only one copy and referring other identical data to that copy [5]. Deduplication image is an image that has characteristics matching in just one database differ in color, rotation, format, and action in the image and so on [6]. Thus, when there a lot of duplicate image in the database could give implication in the use of data storage memory space wasted for the same data [3].

Automated robust methods for duplicate detection of images is getting more attention recently due to the exponential growth of multimedia content on the web. The large quantity of Multimedia data makes it infeasible to monitor them manually [7]. Many duplicate detection [6][9][7]and sub-image retrieval schemes have been proposed in the previous work. G. Pratim et. al. 2007 has been proposed a system that can detect duplicate image in a large scale database and focus in scalability issue. Y. Maret et. al. 2005 proposed duplicate detection based on support vector classifier. G. Pratim et. al. 2007 has been proved that scalable duplicate detection method has been demonstrated for the web and it applicable to use. L. Zhou et. al. 2014 used data deduplication method to detect image deduplication.

ISBN 2320-2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.5 , No.3, Pages : 12-18 (2016)
Special Issue of ICDCB 2016 – Held during April 14-15, 2016 in Kuala Lumpur, Malaysia
http://warse.org/IJACST/static/pdf/Issue/icdcb2016sp03.pdf

Detail research in deduplication image in cloud area are done by L. Zhou et. al. 2014**.** Their technique can significantly reduce the transmission time of image files to reduce the transmission time of image files that have already existed in storage. Also the deletion rate for image groups which have the same version of operating systems but different versions of software application is up about 58% [3]. Fig 1 show new data deduplication solution [3]. To store an image file that has been stored in the image storage server, the traditional data deduplication solution first divides the image file into image block, and then compares the fingerprint of each block with the fingerprints in database. This approach spends a lot of time on chunking and fingerprint matching. Therefore, a mechanism being able to quickly detect whether the image file is stored will effectively avoid segmentation of image storage server [3].

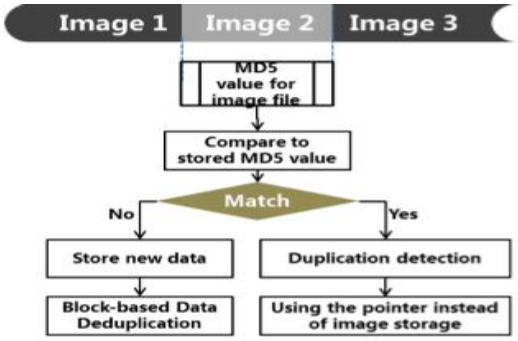

Fig 1: Flow chart of image file deduplication [3]

Based on the literature review, research and discussion about deduplication image it's still increasingly [10] and continue, in this paper we try to find current technique, method, category of image scope of data and user using in image deduplication detection because need to know the effectiveness detection deduplication image based on pilot test is only on existing software. In this paper, discussion based on deduplication image software in standalone database. Detail about deduplication image in standalone database in preliminary section.

PRELIMINARY STUDIES

A pilot test was performed to an existing system using a standalone database where the images contained in the database that is the songket motives images. Total of songket motives images in database of 326 species. The total number of songket motives images can be divide into a number of categories of images. Songket motives image is chosen as a domain for the study because there are more than 1000 images [11] and a number of the specific image is relatively large Fig 2 show some of songket motives images in standalone database. The next section describes the pilot test conducted on ten existing image deduplication software.

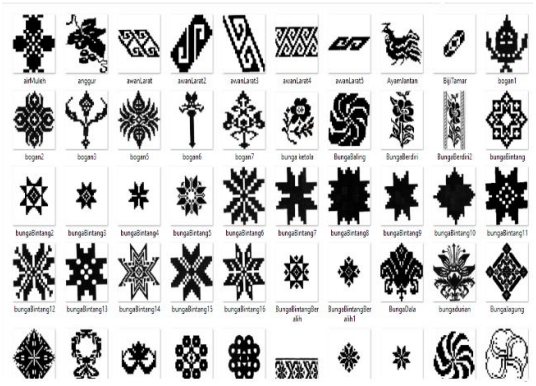

Fig 2: Some of songket image in standalone database

In this paper the aim of the pilot test is to evaluate existing software are related in data or image deduplication detection to find about advantage and disadvantage this software. Further evaluation was conducted to determine the image deduplication technique used to find similar image in the database, domain, the system founder, development goals and the effectiveness of the software in detecting duplicate images contained in the database.

Results of Preliminary Studies

Based on observations, software duplicate cleaner for detect duplicate data through the same keywords as shown in Fig 3. In addition, the results of keyword observations duplicate cleaner software provides results duplicate the same (90%) and only another result (10%) of duplicate data acquisition inaccurate via keywords and this is a question in the survey. The characteristics of which viewed in matching duplicate image is the content or format of the duplicate image detection. Duplicate cleaner software using CRC (Cyclic Redundancy Check) to detect duplicate image or data and using MD5 Hash technique to extract the similar content of data deduplication [12].

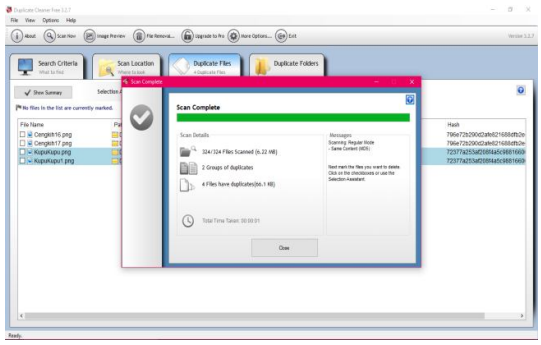

Fig 3: Result of data deduplication by duplicate cleaner software [12]

The next pilot test is using of deduplication image or data detector software is AntiTwin. AntiTwin software gives users the option to delete duplicate data to be retained

ISBN 2320-2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.5 , No.3, Pages : 12-18 (2016)
Special Issue of ICDCB 2016 – Held during April 14-15, 2016 in Kuala Lumpur, Malaysia
http://warse.org/IJACST/static/pdf/Issue/icdcb2016sp03.pdf

Table 1: Comparison Image deduplication detection standalone software using multi technique

Bil	Founder Cloud/OS	Year	Software	Content (Image duplication detection based on?)				Technique							Duration
				Text	Sketch	Color	Image	Hash	Map reduce	SIFT/ GIFT	CRC	Pixel Based	Spatial Layout	Visual Similarity	
1	James (DigitalVolcano Software) http://www.duplicatecleaner.com/ (Standalone)	2015	Duplicate Cleaner (Support 17 different language)	√				√			√				0.1 S/ 16.6 MB 4/326
2	Jorg Rosenthal (http://www.anti-twin.com/) (Standalone)	2010	Anti-Twin (Support 15 different language)	√								√			2 S/ 16.6 MB 4/326 Image deduplication
3	Nirsof (http://www.nirsoft.net/utils/search_my_files.html) (Standalone)	2008	SearchMyFiles (Support 30 different language)	√				√							5 S/ 16.6 MB 0/326
4	Tago Software http://www.similarimagefinder.com/ (Standalone)	2012	Similar Image Finder				√						√		2 S / 16.6 MB 44 /549 Deduplication image
5	Bolide Software (http://www.bolidesoft.com/imagecomparer.html?dvs/) (Standalone)	2011	Image Search Pony				√							√	8 S / 16.6 MB 4/326
6	*Alexander Nikolaev* http://www.duplicate-finder.com/photo.html (Standalone)	2010	Awesome Duplicate Photo Finder				√							√	18 S/ 16.6 MB 14/326 deduplication image detect
7	UngSoft Developer Group http://www.ungsoft.com/ (Standalone)	2011	Similar Picture Find				√						√		3 S/ 16.6 MB 0/326 Deduplication image finder
8	Nils Maier https://tn123.org/about/ (Standalone)	2006	Similar Image				√							√	2 S/ 6.16 MB 14/326 deduplication image found in this software
9	Indeep Software http://indeepsoft.blogspot.my/p/exact-duplicate-finder.html (Standalone)	2015	Exact Duplicate Finder	√											2 S/ 6.16 MB 2/326 Deduplication image found
10	MindGems (http://www.mindgems.com/products/VS-Duplicate-Image-Finder/VSDIF-About.htm) (Standalone)	2009	Visual Similarity Duplicate Image Finder	√			√	√							2 S/ 16.6 MB 24 Duplicate image found
11	Phash http://phash.org/ (Cloud)	2008	Phash				√	√							Comparison similarity based one by one image

ISBN 2320-2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.5 , No.3, Pages : 12-18 (2016)
Special Issue of ICDCB 2016 – Held during April 14-15, 2016 in Kuala Lumpur, Malaysia
http://warse.org/IJACST/static/pdf/Issue/icdcb2016sp03.pdf

or deleted in addition, the software also provides the option button use to help erase the image automatically.

The results of the pilot test early as in Fig 4 finds duplicate data obtained more detailed listed by the matching software AntiTwin investigated by the equation name (text) file position and date of all the matches and count as duplicate data by software. In addition, the time for detection of image deduplication taking longer than Duplicate cleaner, it is based on a comparison of the size of the memory and the time taken to perform the scan as shown in Table 1.

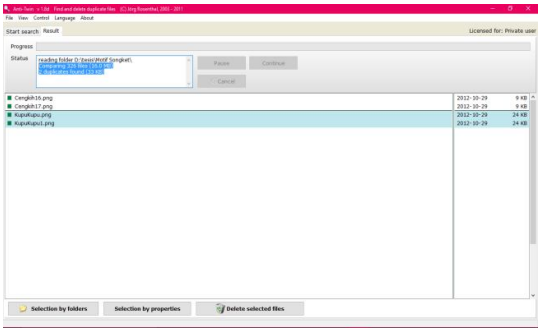

Fig 4: AntiTwin Software detection interface and result [13]

The third pilot test deduplication image detection software is SearchMyFiles. Result image deduplication detection for songket motives image state in Table 1 and this software use a short times to gives a result it is 5 second for 16.6 MB image files sizes. Based on the preliminary pilot test the result is 0. Result and interface for this software shows in Fig 5.

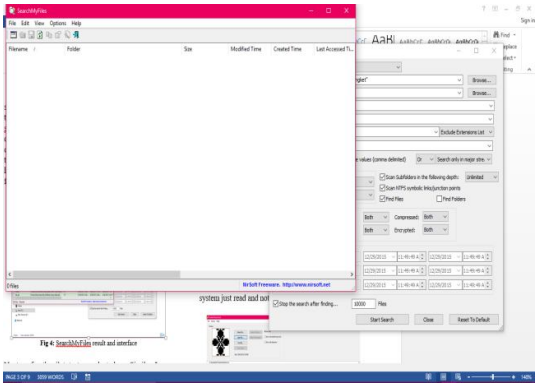

Fig 5: SearchMyFiles result and interface [14]

Next, a fourth pilot tests conducted on Similar Image Finder software, based on researcher observation software detects in deduplication image through the common position of the object contained in the image, this can be seen in Fig 6, two image are detect as a deduplication image but in songket motives category this two image are two different image. In this software, image detection calculation maybe based on foundation shape S that why this system assume this songket motives as a relevant deduplication image.

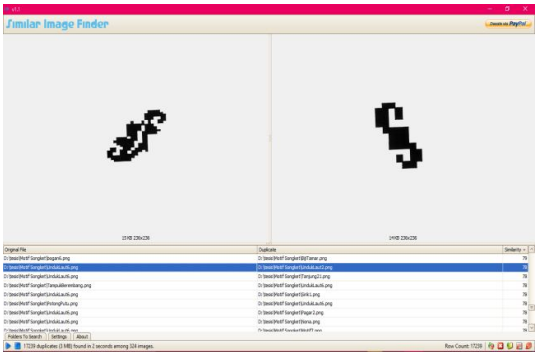

Fig 6: Similar Image Finder image deduplication result [15]

Image Search Pony software is also a fifth tested through a pilot test to see the usability of the software. This software is tested with 16.6 MB sized files and the time taken to perform the scan is 8 seconds. Thus, this software takes a longer time compared to some other software. The scan also showed the results of duplicate data is not found, this is also the question of how to duplicate the image scanning software works because other software tested to produce a duplicate image that is considered relevant as it makes it difficult for researchers duplicate the image to understand briefly how recognition techniques implemented by software this. Fig 7 shows the Interface Software Image Search Pony after pilot tests conducted.

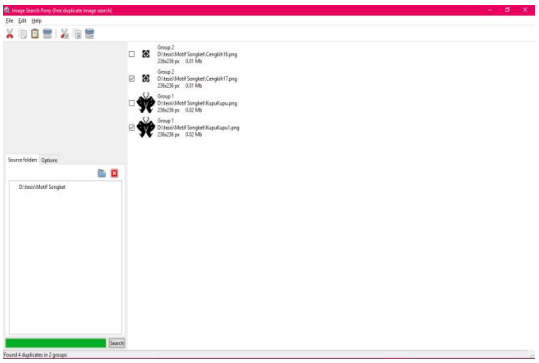

Fig 7: Pilot test result for Image Search Pony software [16].

Pilot test in similar picture find in Fig 8 shows that this system is does not function correctly, it is because we try to use the same database songket motives images but this system just read and not appear the deduplication result.

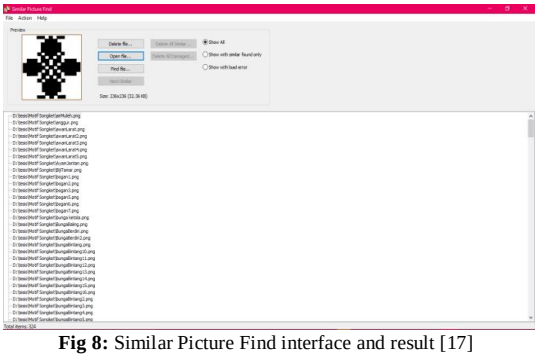

Fig 8: Similar Picture Find interface and result [17]

ISBN 2320-2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.5 , No.3, Pages : 12-18 (2016)
Special Issue of ICDCB 2016 – Held during April 14-15, 2016 in Kuala Lumpur, Malaysia
http://warse.org/IJACST/static/pdf/Issue/icdcb2016sp03.pdf

Awesome Duplicate Photo Finder software is software that is run sixth pilot test aimed at identifying the software's ability to scan and duplicate images contained in the operating system. Based on Fig 9 duplicate image detected by the software shows a high similarity, the image is only distinguish the movements of the hand and shows the system is effective in obtaining a duplicate image contained in the operating system. However this system faces a disadvantage in terms of interface that uses buttons fungi that are not in accordance with international standards.

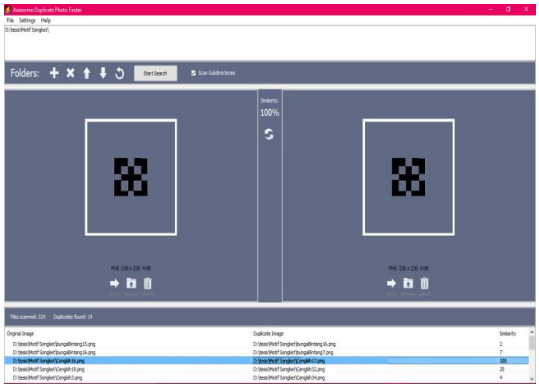

Fig 9: Result of deduplication image detection by Awesome Duplicate Photo Finder. [18]

SimilarImage is another one of deduplication image detector software that we have using test on image songket motives database that to know about software functionality and get an idea how software read and match the similarity deduplication image in the same database or database. Fig 10, show the result and interface SimilarImage software. Based on review this comparison in this software, show the percentage of the similarity image deduplication detection, file size, image format and dimension image to firm the similarity of image.

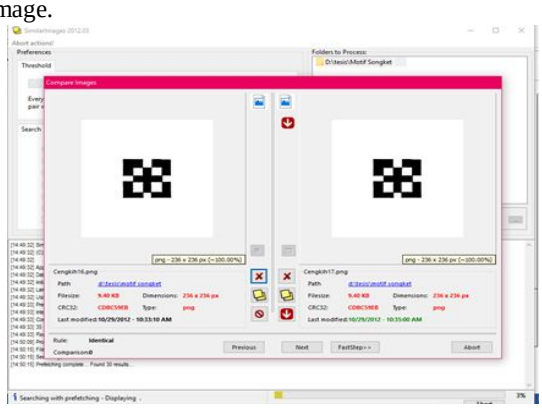

Fig 10: Pilot test result and Interface of SimilarImage Software [19]

Fig 11 shows the result from exact duplicate finder software. In this pilot test this software has detect 2 deduplication image result but the result is a different image not the deduplication image, maybe this software using partial region as a technique to calculate the similarity of the image in the database.

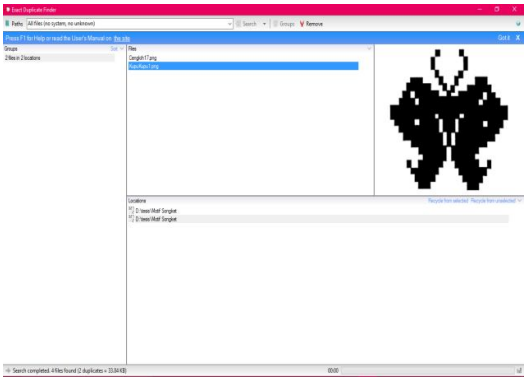

Fig 11: Exact duplicate finder interface and result [20]

Next, Visual Similarity Duplicate Image Finder is the last pilot test testing in this paper. The test results revealed the recovery of the data and duplicate images in more detail because the software calculates the percentage equation image equation that has 95% and above are taken into the result field as a duplicate image. Based on the review, this software check in the detail like dimension, file sizes. But the problem with the software is similarity of image songket motives has image near similarity but not the same image and this is the problem with the system and Fig 12 result and interface for Visual Similarity Duplicate Image Finder.

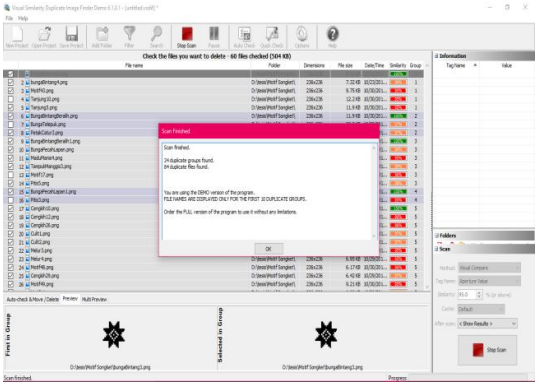

Fig 12: Visual Similarity Duplicate Image Finder [21]

As is shown in Fig 13 pHash system show how using hash technique will be detect the similar image. This middleware using photo clearance certificate (PCC) as similarity accuracy of deduplication image and image categorize as similar image are above than 85% similarity. Based on observation, detection deduplication very detail because this system detect same image as a duplicate image and near duplicate image as a different image. Fig 13 and 14 shows the result this system using near deduplication image and deduplication image. Result for Fig 13 shows image similarity 72%, Fig 14 show the similarity of image deduplication is 100%. Based on this two comparison this system detection for similarity image very detail.

ISBN 2320-2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.5 , No.3, Pages : 12-18 (2016)
Special Issue of ICDCB 2016 – Held during April 14-15, 2016 in Kuala Lumpur, Malaysia
http://warse.org/IJACST/static/pdf/Issue/icdcb2016sp03.pdf

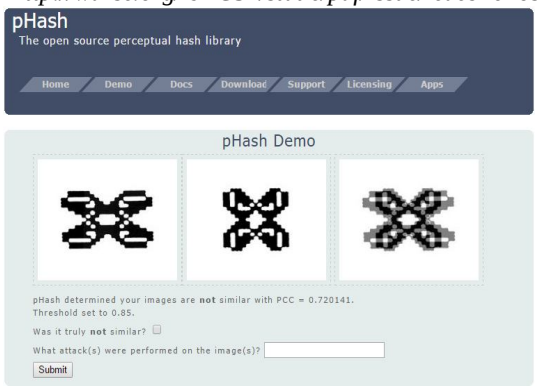

Fig 13: Detection result by pHash cloud software [22]

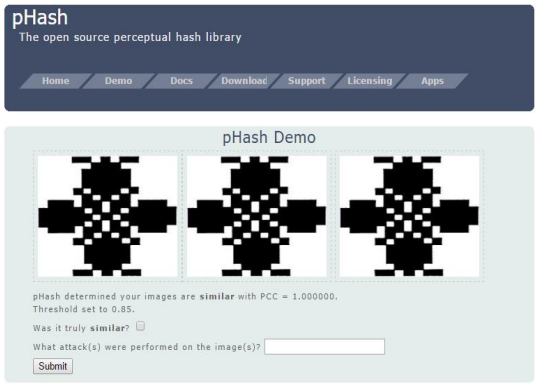

Fig 14: Detection result for the same image

DISCUSSION

Result of preliminary studies show, all the 11 software that has doing a pilot test is using standalone and cloud database, as a place to detect deduplication image. Then software or plugin to detect image deduplication at cloud database is still new and increasingly in further. But for research in mobile cloud deduplication image detector it's has existing since 2008 [23].

After that, result of this pilot test shows, all the elevan software use four different technique in image deduplication detection. The four of that technique is; hash (3 standalone software 1 cloud), visual similarity (3 software), pixel based (1 software) and spatial layout (1 software), another left one software (exact duplicate finder) still didn't know which technique the developer use in image deduplication detection. Deduplication image detection in cloud software using hash technique to detect similarity of image.

Based on the comparison recall of deduplication image detection using songket motives image as a domain, shows three from ten of that software get four deduplication image from 326 image. The software is duplicate cleaner, anti-twin and Image Search Pony. After that, result for similar image finder software, similar image and SearchMyFiles is 0 deduplication image detection. Awesome duplicate photo finder and exact duplicate finder software is 14 deduplication image detection. Lastly is visual similarity duplicate image finder get recall result is 24 deduplication image.

However, this pilot test is to see the results of the technical precision of duplicate images that are applied in software. Then 2 of 10 techniques that yield high precision combined with the technique used and the assumptions map reduce early techniques help reduce map calculation in terms of image indexing calculation. Two techniques that yield high-precision image deduplication detector is hash and visual similarity and pilot test on cloud system has used hash as a techniques in image deduplication detection.

CONCLUSION

Lastly, this research to support multimedia cloud computing concept to provide a better quality of services for user choose to using software and hardware as cloud and to achieve a better quality of experience using cloud storage [23].

Image deduplication detection middleware in cloud is very important because, the middleware will be help user realize the existing deduplication image or data in their cloud database storage. Another that, the middleware will be help to reduce space for deduplication image in the cloud database.

FUTURE WORK

Research in deduplication image in cloud it's still increasingly and need a detail research in this area. Development for software, middleware or plugin using these three technique hash, visual similarity and map reduce will be help us to see the real result about deduplication image detection in cloud database.

ACKNOWLEDGEMENT

This research was supported by research university grant (GUP-2015-008). Alhamdulillah and Thank to Ministry of Higher Education under MyPhd program for supporting my financial education at PHD level education. National University of Malaysia, especially Faculty of Information Science and Technology for support and allowing us to study at here. Thanks to Hj Yusof bin Ismail to support from the beginning and thanks to my supervisor Dr. Amirah and Dr. Nazatul to guide me writing this article from the scratch and all the related person.

REFERENCES

[1] C. C. Shue, J. Ramesh, T. Yonghong, W. Hohong. Special Issue of IEEE Transaction on Multimedia "Multimedia: The Biggest Big Data". Pp- 1-2. Available:http://www.bigmm2015.org/doc/TMM_BigMMSICfP.pdf. 2015.

[2] Amazon Cloud Drive. Unlimited Cloud Storage from Amazon. aws.amazon.com. 2015.

[3] Z. Lei, Z. Li, Y. Lei, Y. Bi, L. Hu, and W. Shen, "An improved image file storage method using data deduplication," *Proc. - 2014 IEEE 13th Int. Conf. Trust. Secur. Priv. Comput. Commun. Trust. 2014*, pp. 638–643, 2015.

[4]Microsoft Azure. https://azure.microsoft.com/en-us/documentation/services/sql-database/. 2016.

[5] B. Choudhary and A. Dravid, "A Study on Authorized Deduplication Techniques in Cloud Computing," vol. 3, no. 12, pp. 4191–4194, 2014.

[6] Y. Maret, F. Dufaux, and T. Ebrahimi, "Image replica detection based on support vector classifier," *Proc. SPIE - Int. Soc. Opt. Eng.*, vol. 5909, pp. 1–9, 2005.

[7] P. Ghosh, E. D. Gelasca, K. R. Ramakrishnan, and B. S. Manjunath, "Chapter 1 Duplicate Image Detection in Large Scale Databases," *Adv. Intell. Inf. Process. Tools Appl. Eds. B. Chandra CA Murthy*, vol. 1, pp. 149–169, 2007.

[8] E. Ave, A. A. Mi, and Z. Wang, "High-Confidence Near-Duplicate Image Detection," *Proc. 2Nd ACM Int. Conf. Multimed. Retr.*, pp. 1–8, 2012.

[9] S. Roy, "A Unified Framework for Resolving Ambiguity in Copy Detection Categories and Subject Descriptors," pp. 648–655, 2005.

[10] S. Kim, X. J. Wang, L. Zhang, and S. Choi, "Near duplicate image discovery on one billion images," *Proc. - 2015 IEEE Winter Conf. Appl. Comput. Vision, WACV 2015*, pp. 943–950, 2015.

[11] A. A. Arba'iah, Y. Othman. 2009. Falsafah di sebalik motif-motif songket Melayu Terengganu. Seminar Antarabangsa Tenunan Nusantara: Kesinambungan tradisi dan budaya. http://elib.uum.edu.my/kip/Record/um782898 [12/1/2013].

[12] James DigitalVolcano Software. Available: http://www.duplicatecleaner.com/. 2015.

[13] J. Rosenthal. Available: http://www.anti-twin.com/. 2010

[14] Nirsof. Available:
http://www.nirsoft.net/utils/search _my_files.html. 2008

[15] Tago software. Available: http://www.similar imagefinder.com. 2008.

[16] Bolide Software. http://www.bolidesoft.com/image comparer.html?dvs/. 2011.

[17]UngSoft Developer Group. http://www.ungsoft.com. 2011.

[18]Alexander Nikolaev. http://www.duplicate-finder.com/photo.html. 2010.

[19] Nils Maier. https://tn123.org/about/. 2006.

[20]Indeep Software. http://indeepsoft.blogspot.my/ p/exact-duplicate-finder.html. 2015

[21] MindGems. http://www.mindgems.com/products/ VS-Duplicate-Image-Finder/VSDIF-About.htm.2009

[22] W. Zhu, C. Luo, J. Wang, and S. Li, "Multimedia cloud computing," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 59–69, 2011.

[23] X. Song, X. Peng, J. Xu, and F. Wu, "Cloud-based distributed image coding," *2014 IEEE Int. Conf. Image Process. ICIP 2014*, pp. 4802–4806, 2014.

[24] S. Kesavan, "Network performance analysis of cloud based multimedia streaming service," vol. 4, no. 3, pp. 156–166, 2014.

[25] J. G. Hansen and E. Jul, "Lithium: Virtual Machine Storage for the Cloud," *Proc. 1st ACM Symp. Cloud Comput. - SoCC '10*, p. 15, 2010.

[26] K. Jin and E. L. Miller, "The Effectiveness of Deduplication on Virtual Machine Disk Images," *Proc. SYSTOR 2009 Isr. Exp. Syst. Conf.*, no. May, pp. 1–12, 2009.

[27] Tago Software.
http://www.similarimagefinder.com. 2015