IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-4, Issue-5 E-ISSN: 2347-2693

Survey on De-Duplication Techniques at Public Cloud

Rajani Sajjan[1], Gayatri Chavan[2], Vijay R. Ghorpade[3]

[1]Pursuing Ph.D. in Computer Science & Engineering Shivaji University, Kolhapur.
[2]Pursuing M.E in Computer Science & Engineering Department VVPIET, Solapur University
[3]Completed Ph.D. from STRM University, Nanded

Available online at: www.ijcseonline.org

Received: Apr/21/2016 Revised: May/04/2016 Accepted: May/18/2016 Published: May/31/2016

Abstract— **Since the demand for data storage is increasing day by day and by the industry analysis we can say that digital data is increasing gradually, but the storage of redundant data is excess which results in most of the storage used unnecessary to keep identical copies. So this survey paper introduces various de-duplication techniques to efficiently utilize the cloud storage system.**

Keywords- cloud computing, de-duplication, cloud storage, data availability, data Integrity, confidentiality, authorization, cloud service provider

I. INTRODUCTION

Cloud computing delivers massively scalable computing resource's as service with Internet based technologies. As digital data is growing tremendously, cloud storage service is gaining popularity since they provide convenient and efficient storage service that can be accessed anytime from anywhere. Cloud computing integrates the computing storage, networking and other computing resources and leases to users; the cloud storage is designed in the form of virtualized computing environment. According to the definition by NIST [1] (National Institute of standards and Technology), *"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."* This cloud model is composed of four deployment models.

(i) "Private cloud" used only in one organization.
(ii) "Community cloud" used in multiple organizations sharing concern.
(iii) "Public cloud "used by general public.
(iv) "Hybrid cloud" composed of two or more distinct deployment model.

Cloud Computing provides several means of interaction between cloud servers and users through the service layer provided in cloud architecture such as:

(i) Software as service (SaaS): This provides complete application as service.
(ii) Platform as a service (PaaS): This provides business clients with independently maintained platform for developing other application on top of it.
(iii) Infrastructure as a service (IaaS): This provides a complete environment for deploying, running and managing virtual machine and storage.

Despite the significant advantages that cloud computing has there are still many security obstacles, factors on security of cloud computing are: data confidentiality, integrity, and availability (CIA). Data confidentiality means that only authorized persons can use the data. Data integrity refers to information that has not been modified or remains untouched. Data availability refers to use of data in time whenever needed and also refer to the availability of cloud service provider (CSP) on demand. Authentication refers to the process of verifying whether the incoming user is authorized or not. As cloud computing becomes prevalent, information are made available by virtualized resources to user as service across the whole Internet by hiding the platform and implementation details. There are some entities that are commonly used in cloud environment those are Cloud Client, Third Party Auditor (TPA), Cloud Service Provider (CSP) and Cloud Server (CS).

(i) Cloud Client: Client is that entity who is using of cloud services and who has to store data on cloud. Multiple clients can use cloud storage services.
(ii) TPA: TPA is an optional entity. It has expertise and capability to expose dummy client. E.g. authentication of client.
(iii) CSP: CSP is an entity which provides cloud services. E.g. client want to upload file then CSP give call to CS.
(iv) CS: CS is an entity which allow client to perform operation on data stored on it.

Recently cloud based storage service such as Drop-box, Google drive, Apple icloud, Mozy, Microsoft SkyDrive competitively offer easy to access, secure, reliable and low cost remote storage space for file-sharing, document suites and online backup services for their users. As they enable easy data access from anywhere anytime, the main quality characteristics of such services are how efficiently they can handle the large amount of network bandwidth requirement from user to cloud storage and how effectively they can reduce the storage space usages. However storage of high redundant data makes inefficient use of cloud storage resource and upload bandwidth due to which the volume of data stored in cloud increase quickly. To solve this problem, the data de-duplication method is a specialized technique used for eliminating the redundant data and the objective is to improve the storage efficiency.

The de-duplication can be categorized into two strategies : file level and block level de-duplication .The file level de-duplication eliminates duplicates data copy at the file granularity if two files have the same hash value and are identified as identical, this methods needs low computational overhead but has low duplicate elimination effectives. The block level de-duplication is also popular technique, which first divides each input file and then use hash value of each block to eliminate the block already stored in cloud, typical block size is 4KB to 8KB. However benefit of data de-duplication in terms of storage space and storage cost is also associated with some security concerns from the user's perspective and the user that outsource their data to cloud service provider.

II. RELATED WORK

K. Deepa et al. [5] proposed a heuristic global optimization method called particle swarm optimization algorithm for record de-duplication. They considered the fitness function of the PSO algorithm and it is based on swarm of data. Here the proposed approach has two phases such as training phase and duplicate detection phase. First they find the similarity between the all attributes of record pairs using levenshtein distance and cosine similarity. Then they formed the feature vectors for representing the set of elements which required detection of duplicates from this feature vectors, they found the duplicate records by using the PSO algorithm.

Sunita Sarawagi and Anuradha Bhamidipaty [6] proposed an interactive learning based de-duplication system called Active Learning led Interactive Alias Suppression (ALIAS). This technique automatically constructs the de-duplication function by interactively finding the challenging training pairs. An active learner actively picks the subset of instances. It easing the de-duplication task by limiting the manual effort for inputting simple, domain specific attributes similarity functions. It interactively model a small number of record pairs. First they took the small subset of pair of records. Then they find the similarity between records and this initial set of modelling data creates the training data for the preliminary classifier. To improve the accuracy of classifier they selected only n instances from the pool of unlabeled data.

Bilal Khan et al. [7] suggested an approach for duplicate record detection and removal. In this approach, they first convert the attributes of data into numeric form. Then, this numeric form is used to create clusters by using K-Means clustering algorithm. The use of clustering reduces the number of comparisons. After that the divide and conquer technique is used in parallel with these clusters for identification and removal of duplicated records.

Peter Christen [8] surveyed various indexing techniques for record linkage and de-duplication. Record linkage refers to the task of identifying records in a data set that refers to the same entity across different data sources [9]. Blocking technique is used in traditional record linkage approach. Blocking key values are used to place the records into different blocks. According to this BKV, the matched records are placed in same block and non-matching records into different blocks. The record linkage process has divided into two phases: Build and Retrieve. In build phase, at the time of linking two data bases, a new data structure is formed: i) Separate index data structures ii) Single data structures with common key values. The hash table data structure is also used for indexing. In retrieve phase, the retrieval of records from block and it will be paired with other records which having same index value.

Weifeng Su et al. [10] proposed an unsupervised, online record matching method called Unsupervised Duplicate Detection (UDD) algorithm. There are two classifiers in UDD for iteratively identify the duplicate records. The duplicate records from the same source are removed using the exact matching method. In this method relative distance of each field of the records are calculated and according to this value, field's weight will be assigned. After that Weighted Component Similarity Summing (WCSS) Classifier utilizes this weight set for matching the records from various data sources. It places the duplicate records in the positive set and non duplicate records in the negative set. The SVM classifier again identifies duplicates from the positive set. These two classifiers iteratively working together and identify the duplicate records in efficient manner. The iteration stops when new duplicates cannot be found.. So it solves the online duplicate detection problem where the query results are generated on- the-fly.

Hamid Haidarian Shahri and Saied Haidarian Shahri [11] invented an adaptive and extensible framework for eliminating the duplicates. In this framework there are six

steps of workflow for duplicate elimination. In all these steps, user can select appropriate items based on that step. In first step, a clustering algorithm is selected for grouping the records (duplicates). In second step, attributes of records are selected for comparing a pair of tuples. In the next step, similarity functions are selected for measuring attribute similarity. In the fourth step, fuzzy rules are used in the fuzzy inference engine to detect the duplicates. For that, it uses the rule viewer, logging membership functions and machine learning capabilities. Neuro-fuzzy modelling is used for applying the learning technique. By using the rule viewer the user can fine-tune the system's rules and membership functions. For this tuning Adaptive Network-based Fuzzy Inference system (ANFIS) is used in this framework. In the fifth step, membership functions selection is done. At last step, the selection of merging technique is done for choosing which tuple will be the prime representative of the duplicates. In this way the duplicate elimination process is done in this framework.

Peter Christen [12] provided the overview about the Febrl system. Febrl system (Freely Extensible Biomedical Record Linkage) is an open-source data cleaning toolkit. It has two components: first one manages the data standardization using Hidden-Markov Models (HMMs) and second one performs the actual duplicate detection. Febrl requires the training to correctly parse the database entries. It implements the variety of string similarity metrics. Febrl uses the phonetic encoding to detect similar names [12].

III. CONCLUSION

An analysis of the existing data de-duplication techniques is done here. From this survey, it is possible to conclude that there are many existing algorithms such as PSO, UDD, ALIAS any many others discussed in related work provide efficiency in storing unique data by using various coding techniques to make it feasible to store data at public cloud environment

REFERENCES

[1]NIST Cloud Computing Standards Roadmap Working Group NIST Cloud Computing Program Information Technology Laboratory.

[2] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios, "Duplicate Record Detection: A Survey", IEEE Trans. Knowledge and Data Eng., vol. 19, no. 1, pp. 1-16, Jan. 2007.

[3] V. Subramaniyaswamy, S. Chenthur Pandian, "A Complete Survey of Duplicate Record Detection Using Data Mining Techniques", Information Technology Journal 11(8)., ISSN 1812-5638, pp.941-945, 2012.

[4]K. Deepa, R. Rangarajan, "Record De-duplication using Particle Swarm Optimization", European Journal of Scientific Research ISSN 1450-216X.,vol.80,no. 3, pp. 366-378, 2012.

[5]Qinghai Bai, "Analysis of Particle Swarm Optimization Algorithm", Computer and Information Science, vol.3, no.1, pp. 180-184, Feb. 2010. www.ccsenet.org/cis.

[6]S. Sarawagi and A. Bhamidipaty, "Interactive De-duplication Using Active Learning", Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining(KDD'02), pp.269-278, 2002.

[7] Bilal Khan, Azhar Rauf, Sajid H. Shah and Shah Khusro, "Identification and Removal of Duplicated Records", World Applied Sciences Journal 13(5): ISSN 1818-4952, pp.1178-1184, 2011.

[8] Peter Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and De-duplication", IEEE Trans. Knowledge and Data Eng., vol. 24, no. 9, pp. 1537-1555, Sept.2012.

[9] Weifengsu, Jiying Wang, Frederick H. Lochovsky, " Record Matching over Query Results from Multiple Web Databases", IEEE Trans. Knowledge and Data Eng., vol. 22, no. 4, pp.578-588, April. 2010.

[10] A.FarithaBanu, C.Chandrasekar,"A Survey on De-duplication Methods", International Journal of Computer Trends and Technology, ISSN: 2231-2803,vol.3,Issue.3,pp.364368,2012,http://www.internationaljournalssrg.org.

[11] Hamid HaidarianShahri, Saied HaidarianShahri, "Eliminating Duplicates in information Integration: An Adaptive, Extensible Framework", IEEE Computer Society 1541-1672, pp. 63-71, September/October 2006.

[12] Peter Christen, Development and User Experiences of an Open Source Data Cleaning, De-duplication and Record Linkage System", SIGKDD Explorations., vol. 11, Issue 1, pp. 39-48.

[13]V.P.Arunachalam,S.Karthik, "A Novel approach for mining inter-transaction itemsets", European Scientific Journal, 8(14).

[14] Nick Larusso." A Survey of Uncertain Data Algorithms and Applications". IEEE Transaction On Knowledge And Data Engineering, 2009 .

[15] Elliott, Chip. "Quantum Cryptography", IEEE Security & Privacy, 2004.

[16]T. Rubya, N. Prema Latha ,B. Sangeetha "A Survey on Recent Security Trends using Quantum Cryptography ".

[17] P.Shanthi Bala "Intensification of educational cloud computing And crisis of data security in public clouds"

[18] S.SATHAPPAN, Dr.D.C.TOMAR "A study on Cluster Uncertain Data based on Probability Distribution"

Authors Profile

Ms. Rajani S. Sajjan completed B.E. (Computer Science & Engineering) from Walchand College of Engineering from Sangli in 1999. Completed M.Tech(Computer Science & Engineering)from PDA College of Engineering, Gulbarga. Pursuing Ph.D. in Computer Science & Engineering from Shivaji University, Kolhapur.

Ms. Gayatri .S.Chavan completed B.E. (Computer Science & Engineering) from Solapur University in 2012 and pursuing M.E.(Computer Science & Engineering) from Solapur University.

Dr. Vijay R. Ghorpade completed Ph.D. from STRM University, Nanded. Specialized in Mobile Ad-hoc networks, Data Mining & Cloud Computing.