ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. II, Issue X, October 2016

PROGRESSIVE DEDUPLICATION

U. Shirlyvictoria[1], Mrs.M.Parveen Taj[2]
M.phil[1], Associate Professor[2]
Department Of Computer Science
Sri Jayendra Saraswathy Maha Vidyalaya College Of Arts And Science,
Coimbatore .India

***Abstract*— Databases contains very large datasets, where various duplicate records are present. The duplicate records occur when data entries are stored in a uniform manner in the database, resolving the structural heterogeneity problem. Detection of duplicate records are difficult to find and it take more execution time. The problem is that the same data may be represented in different way in every database. While merging the databases, duplicates occur despite different schemas, writing styles or misspellings. They are called as replicas. Removing replicas from the repositories provides high quality information and saves processing time. Duplicate detection is that the strategy of distinctive multiple representations of same universe entities. Today, duplicate detection ways in which ought to methodology ever larger datasets in ever shorter time: maintaining the quality of a dataset becomes more and more hard. We have a tendency to progressive duplicate detection rule that considerably increase the potency of finding duplicates if the execution time is limited: They maximize the gain of the overall process within the time available by reporting most results much earlier than traditional approaches. Comprehensive experiments show that our progressive algorithms can double the efficiency over time of traditional duplicate detection and significantly improve upon related work**

***Index Terms*—Progressive Deduplication, replica, dataset, heterogeneity problem.**

I. INTRODUCTION

Data are among the most important assets of a company. Progressive duplicate estimate the similarity of all comparison candidates so as to check most promising record pairs initial. With the try choice techniques of the duplicate detection method, there exists a trade-off between the quantity of your time required to run a replica detection rule and therefore the completeness of the results. Progressive techniques build this trade-off a lot of useful as they deliver a lot of complete leads to shorter amounts of your time. what is more, they create it easier for the user to outline this trade-off, as a result of the detection time or result size will directly be specified rather than parameters whose influence on detection time and result size is tough to guess.

The main objective of deduplication is to spot 2 or a lot of records, that represents an equivalent object. it absolutely was antecedently known as as record matching and record linkage.

Progressive duplicate estimate the similarity of all comparison candidates therefore on check most promising record pairs initial. With the combine choice techniques of the duplicate detection method, there exists a trade-off between the number of your time required to run a reproduction detection algorithmic program and therefore the completeness of the results. Progressive techniques create this trade-off additional useful as they deliver additional complete ends up in shorter amounts of your time. moreover, they create it easier for the user to outline this trade-off, as a result of the detection time or result size will directly be such rather than parameters whose influence on detection time and result size is difficult to guess.

The main objective of deduplication is to spot 2 or additional records, that represents identical object. it absolutely was antecedently known as as record matching and record linkage. Government agencies and companies have spent large amount of money to clean the dirty data in the repositories, which results in quality data content. Replica-free repositories will improve the efficiency and save the processing time. With increase in substantial amount of data, problems like security, low response time, quality assurance and availability begin to arise.

II. LITERATURE REVIEW

Recently, lot of work has appeared in the literature on the problems of the computational grid. Some of the research work are explained as follows.

One way to address these problems is to make grid middleware incorporate the concept of autonomic systems. Such a change would involve the development of ``self-configuring" systems that are able to make decisions autonomously, and adapt themselves as the system status changes.

Alexandre et al. (2007) proposed a semantic approach to integrate selection of equivalent resources and selection of equivalent software artifacts in order to improve the schedule of resources suitable for a given set of application execution requirements. A successful grid resource allocation depends,

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. II, Issue X, October 2016

among other things, on the quality of the available information about software artifacts and grid resources. Scheduling parallel and distributed applications efficiently onto grid environments is a difficult task and a great variety of scheduling heuristics have been developed. Making use of the network in an efficient and fault tolerance manner, in the context of such existing research, leads to a significant number of research challenges. One way to handle these issues is to form grid middleware incorporate the conception of involuntary systems. Such a amendment would involve the event of ``self-configuring" systems that area unit able to build choices autonomously, and adapt themselves because the system standing changes.

Alexandre et al. (2007) projected a linguistics approach to integrate choice of equivalent resources and choice of equivalent computer code artifacts so as to enhance the schedule of resources appropriate for a given set of application execution necessities. A successful grid resource allocation depends, among different things, on the standard of the on the market info regarding computer code artifacts and grid resources. programing parallel and distributed applications expeditiously onto grid environments may be a troublesome task and an excellent form of programing heuristics are developed aiming to address this issue.
Romulo et al. (2007) focused on research methods to achieve efficient execution of parallel applications in a grid computing infrastructure. The paper presents WSPE, a grid programming atmosphere for grid-unaware applications. WSPE's runtime system employs a replacement programing mechanism, referred to as spherical Stealing, impressed on the concept of labor stealing. WSPE consists of a straightforward programming interface and a totally suburbanized runtime system following a peer-to-peer organization. they need conjointly incontestable however associate acceptable alternative for a network overlay mechanism will additional improve execution potency.

Singh and Srivastava (2007) devised a strategy to calculate Queue Length and Waiting Time utilizing entry Server info to cut back latent period variance in presence of burst traffic. the foremost widespread contemplation is performance, as a result of entry servers should supply cost-efficient and high-availability services within the elongated amount, so they need to be scaled to fulfill the expected load. Performance measurements is the bottom for performance modeling and prediction. With the assistance of performance models, the performance metrics (like buffer estimation, waiting time) is determined at the event method.

Wang et al. (2008) conferred a replacement service-oriented approach to the look and implementation of visualization systems during a grid computing atmosphere. The approach evolves the normal dataflow visualization system, supported processes human action via shared memory or sockets, into associate atmosphere within which visualization net services is coupled during a pipeline mistreatment the subscription and notification services obtainable in Globus Toolkit four. a selected aim of their style is to support cooperative visualization, permitting a geographically distributed analysis team to figure collaboratively on visual analysis of information.

III. Proposed System

The system architecture for the progressive deduplication is shown in figure 1

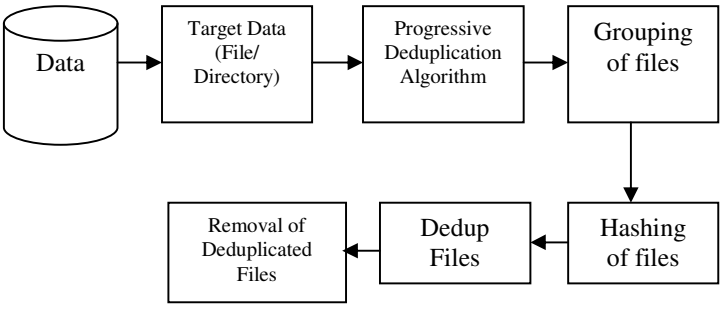


Figure 1 Progressive Deduplication Architecture Design

The deduplication process starts from a source of data. Data can be of any format or extension. The target data file or directory is chosen for the process of deduplication. By applying the deduplication algorithm using hash concept the files are scanned and grouped under same size. Then hashing of files takes place. The duplicated files are listed with the name, path location and size of the file. By using RegExp (regular expression) for pattern matching and select the files of duplicated. The duplicated files are then removed.

The progressive deduplication algorithm for fixed block based on the secure hash algorithm. Data Deduplication application creates a unique identifier for each of the chunks which are exponential small in size as against the chunk size. This can be achieved by hashing. Hashing creates a significantly smaller representation of a large data. Some of the popular hashing algorithms used are Secure Hash Algorithm (SHA1, SHA512, SHA256), Message Digest 5 (MD5) Algorithm.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. II, Issue X, October 2016

MD5 Algorithm

MD5 algorithm takes input message of arbitrary length and generates 128-bit long output hash. MD5 hash algorithm consist of 5 steps

Step 1. Append Padding Bits
Step 2. Append Length
Step 3. Initialize MD Buffer
Step 4. Process Message in 16-Word Blocks
Step 5. Output

SHA-1 Algorithm

The SHA Algorithm is a cryptography hash function. It is used in digital certificate and also data integrity. It is used for computing a compressed representation of a message or a data file. SHA is a fingerprint for use with digital signature applications. The message which is less than 264 bit in length.

SHA-256

SHA-256 operates in the manner of MD4, MD5, and SHA-1: The message to be
hashed is first
(1) padded with its length in such a way that the result is a multiple of 512 bits
long, and then
(2) parsed into 512-bit message blocks $M^{(1)}, M^{(2)}, \ldots..,M^{(N)}$.

SHA- 512

The SHA-512 compression function operates on a 1024-bit message block and a 512-bit intermediate hash value. It is essentially a 512-bit block cipher algorithm which encrypts the intermediate hash value using the message block as key. Hence there are two main components to describe: (1) the SHA-512 compression function, and (2) the SHA-512 message schedule.

SHA-512 is a variant of SHA-256 which operates on eight 64-bit words. The message to be hashed is first (1) padded with its length in such a way that the result is a multiple of 1024 bits long, and then
(2) parsed into 1024-bit message blocks $M^{(1)}, M^{(2)}, \ldots..,M^{(N)}$.
The message blocks are processed one at a time: Beginning with a fixed initial hash value H(0), sequentially compute
$H^{(i)} = H^{(i-1)} + C_{M(i)}(H^{(i-1)})$;
where C is the SHA-512 compression function and + means word-wise mod 264 addition. H(N) is the hash of M.

V. IMPLEMENTATION

The implementation of progressive deduplication consists of following modules:

1. Progressive Deduplication Setting
2. Scanning duplicates
3. Result
4. Install shield

Progressive Deduplication Setting

The progressive deduplication application setting consists of directory, filters of files, choosing of algorithm with limiting the hash byte and processing. A **directory** is a location for storing files on the computer. The file or folder to be deduplication is seleted using directory add functions. Filters are used for filtering exception of some directories. The progressive algorithm of hash using MD5, sha-1, sha-256, sha-512 are designed to select with their hash limits to process. Once setting is done the process starts.

Scanning Duplicates

The duplicated files in the directories are scanned based on the progressive algorithms selected and processed. The files are scanned based on the three process. First one is finding number of files based on size, extension, etc. Second one by grouping the files of same domain based on algorithm. The final one is hashing of files and total files found to be duplicated. Once scanning is done the total number of files duplicated are listed in the result form.

Result form

The result form displays the file list that are duplicated in the directory. The listed files shows name of the file, path location in the backup storage and file size. The result form also consists of selection of files both by manually or by regression of (N-1) selection files of total files.

Install Shield

The Install shield to create a software based setup file. The Install shield helps to create the process into CDROM.bak file format where setup files are located. It implements to create a software based applications as in real time applications.

VI. RESULTS

The directory files are added to our process by clicking either Add directory button or by simply drag the directory or files to the window. The files or directory to be filtered can be done in the filtering box. The figure 2 shows the directory file selection and settings. The figure 3 shows the various algorithm can be processed with the hash limit.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. II, Issue X, October 2016

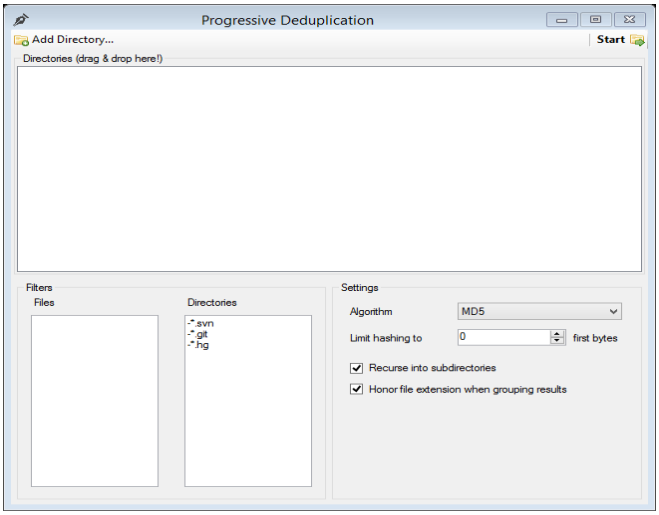


Figure 2 Progressive Deduplication

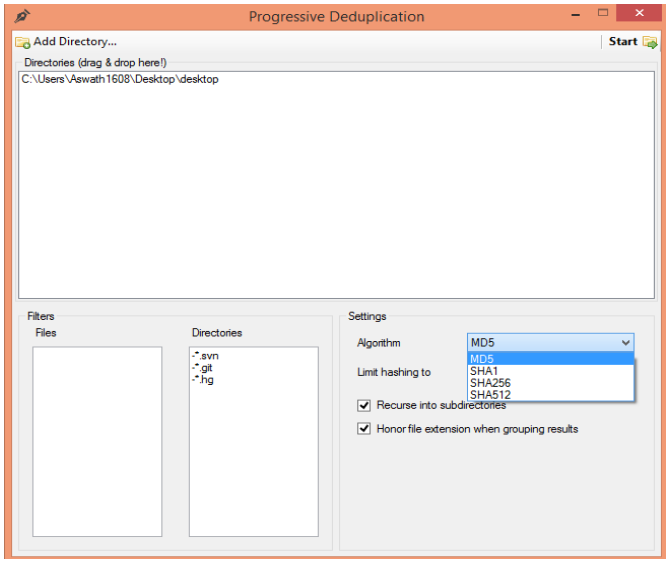
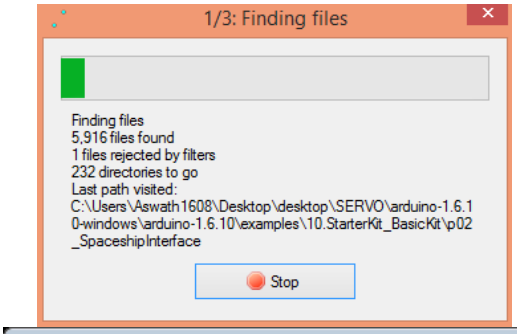
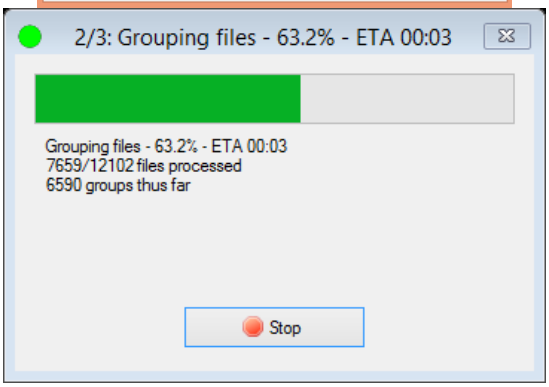
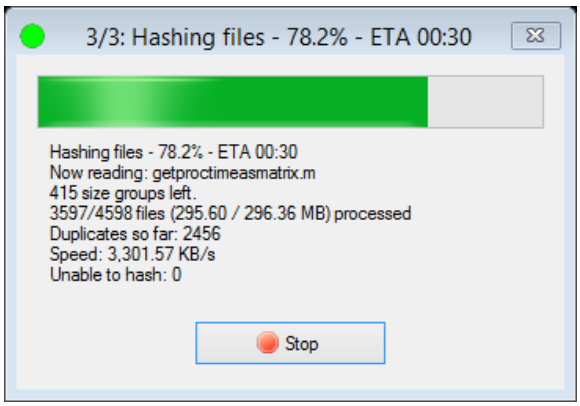


Figure 3 Algorithm Selection

The directories are started to scan by finding files as shown in figure 4. After finding the file sources, the grouping of files are processed as shown in figure 5. The hashing of files for the grouped files are processed as shown in figure 6.



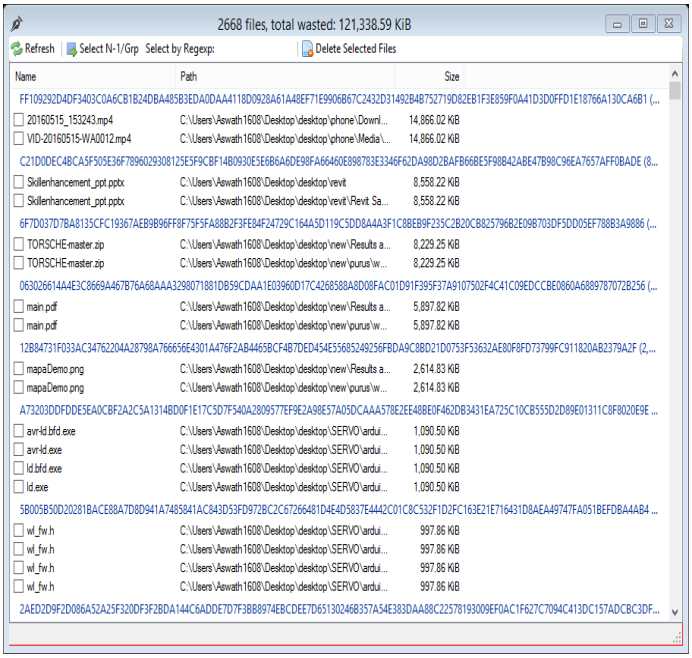




The list of duplciated files are listed with the name,path location and size which can be removed as shown in figure 7

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. II, Issue X, October 2016



VII. Conclusion And Future Work

Data deduplication is a scalable and efficient redundant data reduction technique for large-scale storage systems, which addresses the challenges imposed by the explosive growth in demand for data storage capacity. The concept of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. Similarly open several new deduplication constructions supporting authorized duplicate check Security analysis proves that our schemes are secure in terms of insider and outsider attacks specified in the proposed security ideal. As a resistant of idea, implemented a prototype of authorized duplicate check scheme and conduct tested experiments on
our prototype.

The file systems built into modern Operating Systems do not provide adequate support for managing file duplication. File duplication can be identified in detail with initial comparison of files, followed by MD5 algorithm in earlier and by SHA-3 algorithm in later. Based on MD5 and SHA-3, hash values for files have been generated. One of the redundant files is removed, if hash values of similar type files are same. The time taken to compute hash value by SHA-3 is much lesser than MD5, SHA-2, and SHA-1. SHA-1, SHA-2, MD5 consumes more memory than SHA-3 algorithm. The performance of SHA- 1, SHA-2, MD5 hash function is severely compromised in terms of memory consumption and time compared with SHA-3 algorithm. SHA-3 helps in retrieving valuable disk space and in improving the efficiency. SHA-3 is the best in identifying the redundant files and removing it. The environment we studied, despite being homogeneous, shows a large diversity in file system and file sizes.

References

[1] Khan B, Rauf A et al. (2011). Identification and removal of duplicated records, World Applied Sciences Journal, vol 13(5), 1178–1184.

[2] Christen P (2011). A survey of indexing techniques for scalable record linkage and deduplication, IEEE Transactions on Knowledge and Data Engineering, vol 24(9), 1537–1555.

[3] DeCarvalho M G, Alberto H F et al. (2013). A genetic programming approach to record deduplication, IEEE Transactions on Knowledge and Data Engineering, vol 24(3), 399–412.

[4] Deepa K, and Rangarajan R (2012). Record deduplication using particle swarm optimization, European Journal of Scientific Research, vol 80(3), 366–378.

[5] Subramaniya swamy V, and Pandian S C (2012). A complete survey of duplicate record detection using data mining techniques, Information Technology Journal, vol 11(8), 941–945.

[6] Banu A F, and Chandrasekar C (2012). A survey on deduplication methods, International Journal of Computer Trends and Technology, vol 3(3), 364–368.

[7] Shanmugavadivu P, and Baskar N (2012). An improved genetic programming based approach to deduplication using KFINDMR, International Journal of Computer Trends and Technology, vol 3(5), 694–701.

[8] S. E. Whang, D. Marmaros, and H. Garcia-Molina, "Pay-as-you-go entity resolution," IEEE Trans. Knowl. Data Eng., vol. 25, no. 5,pp. 1111–1124, May 2012.

[9] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proc. 32nd Annu. Int. Conf. Theory Appl. Cryptographic Techn., 2013, pp. 296–312.

[10] Meera K. Krishna Sankar P. and Shriram Kumar N K(2015), "Redundant file finder, remover in mobile environment through SHA-3 algorithm", Electronics and Communication Systems (ICECS), pp.1440-1447.

[11] Suresh Subramanian, Sivaprakasam (2014), "Efficient Algorithm for Removing Duplicate Documents", International Journal of Soft Computing and Engineering (IJSCE), Vol-3, Iss-6.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. II, Issue X, October 2016

BIBLIOGRAPHY

U. Shirlyvictoria parsuing my M.phil in stream of Computer Science from Sri Jayendra College of Arts and Science, I have completed my M.C.A in Sasurie Engineering College with aggregate of 82% and I have done my B.Sc Computer Science in Sri Jayendra College of Arts and Science with aggregate of 73%.My area of interest are Data Mining ,java Programming,Photoshop.

Mrs.M.Parveen Taj working as Associate Professor in the Department of Computer Science, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Singanallur, Coimbatore. She has 11.5 years of teaching experience. Her area of interest Networking.