A BRIEF REVIEW ALONG WITH A NEW PROPOSED APPROACH OF DATA DE DUPLICATION

Suprativ Saha[1] and Avik Samanta[2]

[1]Department of Computer Science and Engineering, Global Institute of Management and Technology, Krishnagar City, West Bengal, India
reach2suprativ@yahoo.co.in
[2]Department of Computer Application, JIS College of Engineering, West Bengal, India
avik_2007engg@yahoo.co.in

ABSTRACT

Recently the incremental growth of the storage space and data is parallel. At any instant data may go beyond than storage capacity. A good RDBMS should try to reduce the redundancies as far as possible to maintain the consistencies and storage cost. Apart from that a huge database with replicated copies wastes essential spaces which can be utilized for other purposes. The first aim should be to apply some techniques of data deduplication in the field of RDBMS. It is obvious to check the accessing time complexity along with space complexity. Here different techniques of data de duplication approaches are discussed. Finally based on the drawback of those approaches a new approach involving row id, column id and domain-key constraint of RDBMS is theoretically illustrated. Though apparently this model seems to be very tedious and non-optimistic, but in reality for a large database with lot of tables containing lot of lengthy fields it can be proved that it reduces the space complexity vigorously with same accessing speed.

KEYWORDS

DeDuplication, SQL, Chunk, Genetic Algorithm, Replicated copy, Domain Key Constraint

1. INTRODUCTION

Recently the incremental growth of enterprise and digitized data volumes areexpanding rapidly along with the increase of the information and multimedia files.According to the Enterprise Strategy Group, an industry analysisfirm, 13 percent of midsized companiessurveyed in 2004 used more than 10 terabytes of data storage,but by this year, the number hasbeen increased from 40 to 50 percent.In case of *digitized data* the study by IDC [9] predicts that the information added annually to the digital universe will increase more than six fold from 161 Exabyte to 988 Exabyte between 2006 and 2010, growing by 57% annually. So it is clearthat it requires a large space for storing those data.

However, a lot of the duplicate information is stored from different sources; require more spaces to store replicas. The bitter experiences from the several events proved that data loss is common to a modern enterprise. So it is important to backup the data periodically to a reliable place for the purpose of data availability and integrity. Data backup in the external storage system is more costly and power consuming. According to the survey, although the price of the disk decreased,

Rupak Bhattacharyya et al. (Eds) : ACER 2013,
pp. 223–231, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3220

cost is the main reason to hinder the deployment of the disk-based backup solutions. In fact, there is a huge amount of duplicate or redundant data existing in current storage systems [10], especially backup systems. As a result a lot of spaces are required for data backup and replication. In case of data warehouse for improvement of data quality, data cleaning is a very important task. Data cleaning deals with the detecting and removing of inconsistencies and errors. As there are differences in conventions between the external sources and the target data warehouse as well as due to a variety of errors, data from external sources may not conform to the standards and requirements at the data warehouse. Therefore, data has to be transformed and cleaned before it is loaded into the warehouse. This is usually accomplished through an Extract-Transform-Load process.

There are numerous techniques for reducing redundancy when data is stored or sent. Recently, to cope with this, organizations are intending to usedata de duplication technology. Data de duplication, also called Intelligent Compression or Single-Instance Storage, eliminates redundancy caused by identical object which can be detected efficiently by comparing hashes of the objects' content [4]. Storage space requirements can be reduced by a factor of 10 to 20 or more when backup data is de-duplicated [8].

In earlier days, data De Duplication technique was done by several methods. Especially SQL based method; Chunk based method and model incorporated with the Genetic algorithm are used to formed Data De Duplication technique. Beside the space complexity, time complexity is also an important parameter to judge a concept. Using the previous model of de duplication technique space complexity is reduced, which is main important part in this research domain but it is not required to over look about the accessing time of data from traditional database or data warehouse. Here, a new model is proposed to establish the data de duplication technique which keeps track of the accessing speed of data from traditional database or data warehouse, availability of data besides the decreasing size of storage space.

In the next section current state of arts of de duplication technique is discussed. In section 3 a new proposed model is incorporated in this paper. Finally in section 4 conclude of this technique.

2. CURRENT STATE OF ARTS

Data De Duplication technique is generally used to reduce the storage cost and increasing accessing capability of data from traditional database or data warehouse. Data rearrangement, Data cleaning, data indexing all are incorporated in De Duplication technique. Several authors used different methods to implement De Duplication techniques. At first *SQL based model* are discussed, after that *Chunk based model* and *Genetic algorithm based model* are briefly illustrated respectively.

2.1. SQL Based Model for Data De Duplication

In a large database or data warehouse may reside many duplicate records. These duplicate records reduce the storage capacity and create more time complexity in accessing data. Apart from the storage related issues these duplicate records in the DBMS perspective may generate many inconsistencies. That's why; 'Normalization' technique is incorporated. So, to have a well formed data warehouse or data base, removing the duplicate data as far as possible is most important.
In this paper [14], many techniques related to SQL based are going to be discussed for the elimination of duplicate records.

2.1.1. De duplication by using sub queries tactically

In this technique sub queries are used to tactfully delete the duplicate records. In sub query the inner most query is evaluated only once and the resulting values are substituted into the WHERE clause of the outer query. But in this case at first the outer query is evaluated on the basis of the result obtained from inner sub Query. This approach can be used only if we have identity column on the target table or you are willing to alter the target table to add an identity column which would require ALTER TABLE permission.

2.1.2. De duplication using temporary table

In this approach distinct records from the original table are fetched out and stored in a temporary table. Hence the temporary table contains only unique entitles. Then the original table is truncated and the records from the temporary table are stored back to the original table.

```
BEGIN TRANSACTION
/*Pull out the distinct records in the temporary table*/
SELECT DISTINCT * INTO <T> FROM TABLE
/* truncate the target table*/
TRUNCATE TABLE <T>
/* Insert the distinct records from temporary table */
/* Back to target table */
INSERT INTO <T> SELECT * FROM <T>
/* Drop the temporary table*/
IF OBJECT_ID ('tempdb  ... T') IS NOT NULL
DROP TABLE <T>
COMMIT TRANSACTION
```

In this case the size of the temporary table should be at least to that extent so that it can hold those temporary records.

In SQL based technique, temporary table is used to remove the duplicate data. As a result there is a big chance of data loss. Instate of temporary table, if permanent table is used to remove the duplicate records, the ultimate aim is not satisfied. The storage cost also increases to store those permanent tables. On replace of permanent table or temporary table if view concept is implemented, the problem be solved. In this method user have to involve inserting data in to database or data warehouse. It means each and every techniques of this model is not possible to hide from the 3rd party user. Further more, the accessing complexity of the SQL based technique is high, which is not required.

2.2. Chunk Based Clustering Architecture

Data de duplication, one of the most emerging technologies of latest time reduces the storage cost and increases the storage capacity of several storage devices like magnetic disks and tapes using several techniques. In DBMS perspective, in a simpleform it can be told that it increases the efficiency and the performance of the database or a data ware house. In reality there is huge amount of duplicate data in a data warehouse.

In this paper [12], a concept about the clustering architecture consisting of a lot of nodes, by using B+ tree data structure is discussed in details.

There are multiple nodes in the storage system and the chunks of a file may be stored in different storage node. For checking a duplicate chunk each node firstly will check its own database, and then checks other nodes to detect duplicate chunk if necessary. To check the duplicity a new technique called “Fingerprint Summary” is used. The “Fingerprint Summary” keeps the compact summaries of each fingerprint of each and every node. If a chunk is new for this node, the “Fingerprint Summary” in this nodes memory should be queried to determine whether some other node has stored an identical chunk. Thus ‘inter-node duplicate data can be removed and can save more storage space.

Fingerprint Summary is a compact summary of the chunks fingerprints of every other node. As the chunks on each node are increasing dynamically, it needscontinuous refreshing the fingerprint summary to reflect the changes.

The fingerprint summary can be refreshedon every node in real time. When a new chunk is stored in some node, all other nodes fingerprint summary needs to be refreshed at the same time. Thus as the number of new chunks increases, both the communication and the processing overhead increase quadritically.

2.3. Genetic Algorithm based Model

This approach [13] combines several different pieces of evidencesextracted from data content to produce a deduplication function that is able to identify whether two or more entries in a repository are replicas or not. Since a repository may contain a lot of duplicate data so the first aim is to develop such a function, which may be a combination of various linear or non-linear function or probabilistic function, reducing the processing time and maximizing the performance. The main advantage of Genetic Programming [2] approach is that it provides a very suitable manner to find the duplicate data very efficiently without searching the entire search space for a solution which may be very large. Till now GP has been applied in many areas of information handling like Ranking Function Discovery [3], Document Classification [5], Content Based Image Retrieval [11] and content target Advertising [6].

As a whole, the main aim of this GP based approach is to record deduplication that

1. Outperforms an existing state-of-the art machine learning based method found in the literature.
2. This process is user independent because of the fact that it actually indirectly chooses the functions appropriate for it and combines them according to the requirement.
3. It also automatically chooses the replica identification boundary value freeing the user from choosing it.
4. The suggested deduplication functions use the available evidences more efficiently.

In the development stage the approach can be classified into the following two categories

2.3.1. Ad-Hoc or Domain Knowledge Approaches

This category includes approaches that usually depend on specific domain knowledge or specific string distance unit.

2.3.2. Training-based Approaches

This category includes all approaches that depend on some sort of training –supervised or semi-supervised – in order to identify the replicas.

2.3.2.1. Domain Knowledge Approaches

In this case [7] a matching algorithm proposes that a given record in a file (or repository), will always firstly search for another record in a reference file that matches first record according to a given similarity function. The matched reference records are selected based on a user defined minimum similarity threshold. A vector space model is used for computing similarity among fields from different sources and evaluates four distinct strategies to assign weights and combining the similarity scores of each field. In this way it uses evidence extracted from individual attributes improving the results of the replica identification task

2.3.2.2. Probabilistic Approaches

At the initial phase Bayesian inference problem (a probabilistic problem) is used as the first approach to automatically handle replicas. But the implementation of this approach is obsolete in practical context since it lacks a more elaborated statistical ground. In the next phase a proposed model developed by *Fellegi and Sunter* [1] elaborated the statistical approach to deal with the problem of combining evidences.Their method relies on the definition of two boundary values that are used to classify a pair of records as being replicas or not. Most of the tools like *Febrl* [15] usually work with two boundaries as follows

1. *Positive Identification Boundary:* If the similarity value lies above this boundary, then the records will be considered as replicas.
2. *Negative Identification Boundary*: If the similarity value lies below this boundary,then the records will be considered as non-replica.

2.3.2 Genetic Operations

All the similarity functions in the GP approach are implemented using tree where the leaves are called as terminals and are inputs. The function set is the collection of operators, statements, and basic or user defined functions that can be used by the GP evolutionary process to manipulate the terminal values. These functions are placed in the internal nodes of the tree. The genetic operations use several biological genetic processes such as reproduction, crossover and mutation [2] successively spawning a new tree with the better solutions. The crossover operations make exchanges between the parents or nodes or subtrees generating a new tree with the better solution in a random way. Reproduction is nothing but the inheritance of the genetic codes of the previous generation (tree) to the next generation (tree) keeping the good features or genetic codes intact. At last the mutation operation has the role of keeping a minimum diversity level of individuals in the population avoiding premature convergence. Most of the cases the trees generated by crossover also pass through the mutation operation.

3.A NEW APPROACH OF DATA DE DUPLICATION

The topics discussed in the earlier cases suffer from time complexity though the algorithms successfully reduce the space complexity. If we concentrate on only the RDBMS where each and every data is represented in tabular form containing the rows and columns then it can be shown that any particular tuple of a particular attribute can be represented in a pair wise form like (R, C) where R denotes the rowid and C denotes serial number of the column.Now let's explain a scenario with execution flow chart shown in figure 1.

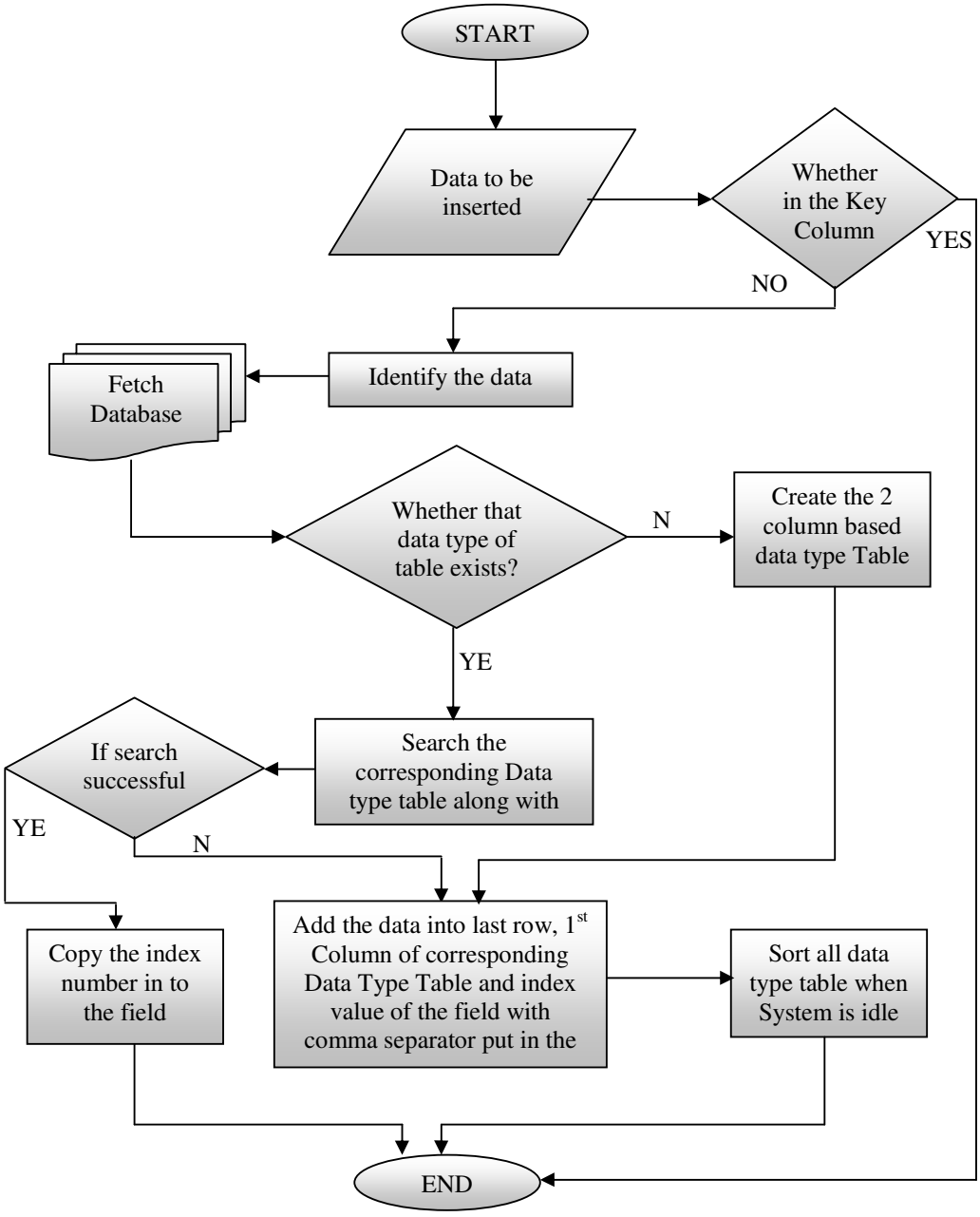

Figure 1: Flow Diagram of the Proposed Approach of Data De Duplication Technique.

In table 1 there is many duplicate data as far as the attribute 'Place of Birth' and attribute 'Country' is concerned. If a single copy of these duplicate data is stored and the replicated copies are replaced by some sort of location identifier of the original copy then the space complexity can be reduced substantially. To have this type of technique a proposed model has been developed based on the row column pair.

Table 1: Example Database before execution of Data De Duplication Technique

Name	Date Of Birth	Place Of Birth	Country	Serial No
Avik Samanta	12/03/1975	KOLKATA	INDIA	101
Suprativ Saha	12/03/1975	KOLKATA	INDIA	102
Robert Pattinson	23/06/1985	NEW YORK	USA	103
James Thomson	17/10/1990	CHICAGO	USA	101
Rajeev Malhotra	25/02/1970	MUMBAI	INDIA	105

At the first stage the several data types and keys of the attributes will be identified based on the domain key constraints. Then an individual table [table2, table 3, table 4, table 5] should be created for each data type without considering the data type of the key attribute because the key attribute can never hold any duplicate data. These tables consists of two attributes namely category and position. According to the above example the tables created are

Table 2: C(For character type data)

Data	Index
Avik Samanta	1,1
Suprativ Saha	2,1
Robert Pattinson	3,1
James Thomson	4,1
Rajeev Malhotra	5,1

Table 3: D (For Date type Data)

Data	Index
12/03/1975	1,2
23/06/1985	3,2
17/10/1990	4,2
25/02/1970	5,2

Table 4: V (For Alphanumeric Type Data)

Data	Index
KOLKATA	1,3
NEW YORK	3,3
CHICAGO	4,3
MUMBAI	5,3
INDIA	1,4
USA	3,4

Table 5: N (For Number type Data)

Data	Index
101	1,5
102	2,5
103	3,5
105	5,5

The first column holds a single copy of each data from the original table satisfying the domain key constraints and the second column is a pair wise form of the location of the data in the original table, where the most significant digit represent the rowid and the remaining the column represent serial number in a (R,C) form. An extra comma separator is used between the R and C to classify the rowid and serial number separately. All the tablesare created to hold different data categorically based on data type and should be sorted at the system idle time so that at the very next time when need to access the individual table, a binary search can easily be used to find out the original location of the master record in the corresponding table. Hence the searching complexity is going to be reduced to $O(\log_2 n)$ in compare to the sequential search of complexity of O(n).

After the execution of the algorithm the resultant table will be like as table 6.

All the bold and highlighted data in the above table are actually the replicated copies which are deduplicated by reducing the data length in the two digit format.In this way the space complexity can be reduced. The C table will not be created in reality as it is a key column and there doesn't

exist any other character type of column, so it can be omitted. This algorithm will be beneficial as long as the data length be bigger.

Table 6: Example Database after execute of Data De Duplication Technique

Name	Date Of Birth	Place Of Birth	Country	Serial No
Avik Samanta	1,2	1,3	1,4	1,5
Suprativ Saha	**1,2**	**1,3**	**1,4**	2,5
Robert Pattinson	3,2	3,3	3,4	3,5
James Thomson	4,2	4,3	3,4	**1,5**
Rajeev Malhotra	5,2	5,3	**1,4**	5,5

Let explain how storage space is reduced. Take a data from the previous example, in 'place of birth' attribute the data value 'KOLKATA' used twice. 'KOLKATA' is an alphabetic as well as alphanumeric character. Let's think it takes 1 byte to store every character; in this case, KOLKATA takes 7 bytes to store. In case of thrice takes 21 bytes. According this proposed approach to store this scenario it takes only 19 bytes (In original table takes 3 bytes to store rowid, column serial number and comma separator, so total 3*3 = 9 bytes are required. In the Index table KOLKATA takes 7 bytes and next column takes 3 bytes. In summation 9 + 7 + 3 = 19 bytes are required). For this small scenario a fruitful result is provided by this new proposed approach so it is obvious it will provide better result to store more number of replicas.

4. CONCLUSION

The duplicate information is stored from different sources, which require more spaces to store replicas. The bitter experience of the several events proved that data loss is common to a modern enterprise.Different types of approaches are involved to solve those problems. SQL based data deduplication, Chunk based clustering architecture and Genetic approach models all are very tedious in their respective field of implementation because of their huge dependence on file organization or file system. Moreover all of these have indecent time complexity which is a matter of great concern. But in the case of this new proposed model as huge as be the relational database with huge number of tables (which is the most anticipated consequence) this proposed model will perform more optimistically than a tiny database. The domain identifier, key identifier, rowid, column id are the key role players in this model. Apart from that the binary search and the table index position are the main optimisation factor for this model. Hence the length of a data of a particular field of a particular row and column are not the matter of concern regarding space wastage, rather it proves our model more practically, in that case. Since the entire algorithm is basically based on the row column approach that is why this can only be implemented in RDBMS initially, though our goal will be to renovate this concept in a broader sense so that it can cover up the all the branches of DBMS.

REFERENCES

[1] I.P.Fellegi and A.B.Sunter,"A theory for record linkage, (1969)" *Journal American Statistical Association*, vol, 66, no.1, pp. 1183-1210.

[2] J.R.Koza, (1992)"*Genetic Programming: on the Programming of Computers by Means of Natural Selection"*. MIT Press.

[3] W.Banzhaf, P.Nordin, R.E.Kellar and F.D.Francone, (1998)"*Genetic Programming-An Introduction: on the Automatic Evolution of Computer Programs and its Applications"*, Morgan Kaufmann Publishers.

[4] W. J. Bolosky, S. Corbin, D. Goebel, and J. R. Douceur, (2000), "*Single instance storage in Windows 2000*", in Proceedings of the 4th USENIX Windows Systems Symposium, Seattle, WA, pp. 13-24.
[5] B.Zhang, Y.Chen, W.Fan, E.A.Fox, M.Gonclaves, M.Cristo and P.Calado, (2005) "*Intelligent GP fusion from from multiple sources for text classification*", in CIKM'05: Proceedings of the 14th ACM international conference on Information and knowledge management. New York, NY, USA: ACM, pp. 477-484.
[6] A.Lacerda, M.Cristo, M.A.Goncalves, W.Fan, N.Ziviani, and B.Ribiero-Neto, (2006) "*Learning to advertise*", in SIGIR'06: proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, pp, 549-556.
[7] A.K.Elmagarmid, P.G. Ipeirotis, and V.S. Verykios, (2007) "*Duplicate record detection: A survey*", IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp 1-16.
[8] H. Biggar, (2007) "*Experiencing data de-duplication: Improving efficiency and reducing capacity requirements*" White Paper, the Enterprise Strategy Group.
[9] J F Gantz, (2007) "*The expanding Degital universe: A forecast of worldwide growth through 2010*", IDC.
[10] Dave Reinsel, (2008) "*Our expanding digital world: Can we contain it? Can we manage it?*", Intelligent Storage Workshop (ISW 2008), UMN, MN, pp. 13-14.
[11] R.d.s. Torres, A.X. Falcao, M.A. Goncalves, J.P.Papa, B.Zhang, W.Fan, and E.A.Fox. (2009)"*A genetic programming framework for content based image retrieval*", Pattern Recogn. vol. 42, no. 2, pp 283-292.
[12] Guohua Wang, Yuelong Zhao, Xiaoling Xie, and Lin Liu, (2010), "*Research on a clustering data de-duplication mechanism based on Bloom Filter*", IEEE.
[13] Mois´es G. de Carvalho, UFMG Alberto H. F. Laender, UFMG Marcos Andr´e Goncalves, UFMG Altigran S. daSilva, UFAM, (2010), *A Genetic Programming Approach to Record De duplication*", IEEE TRANSACTIONS on Knowledge and Data Engineering.
[14] Srivatsa Maddodi, Girija V. Attigeri, Dr. Karunakar A. K, (2010) "*Data Deduplication Techniques and Analysis*", Third International Conference on Emerging Trends in Engineering and Technology, IEEE Computer Society, pp. 664-668.

AUTHORS

Suprativ Saha associate with Global Institute of Management and Technology, Krishnagar City, West Bengal,Indiaas an assistant professor since 2012. He received his M.E degree from West Bengal University of Technology in year 2012 and B-Tech degree from JIS college of Engineering in 2009. His research interests include the field of Database Management System, DataMining, Distributed Database and Optical Network.Mr. Saha has about 2 referred international publications toher credit.

Avik Samantareceived his M.C.A degree from JIS college of Engineering in 2010 and B.S.C degree from The University of Calcutta in year 2007. His research interests include the field of Database Management System, Data Mining, and Distributed Database.