IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Special Issue, 4, Feb 2019 E-ISSN: 2347-2693

Architecture for Automated Data Quality Checking in Big Data Migration Process

V. Rathika

Dept. of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamil Nadu, India

Corresponding Author: rathirajaphd2013@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—**Data are gathered from different sources that have high quality issues. Increasing volume of information is there in the digital libraries. Most of the system may be affected by the replicas. Data cleaning is the important process to remove replicas using de-duplication. It consists of process of parsing, data transformation, duplicate elimination and statistical methods. It is one of the most challenging stages to clear repeated documents. It deals with the detection and removal of errors, filling in omitted values, smoothing noisy data to improve the quality of data. De-duplication is the key function in data integration which is from various sources. It is the process of determining all categories of information contained by a data set that indicate the same real world entity. This paper is going to introduce a methodology for automated data quality checking with de-duplication algorithm.

***Keywords*—**Data Quality, Data Cleansing, De-Duplication.

I. INTRODUCTION

Data warehouses are playing vital role in day-to-day IT based environment. Many industries and systems depend on the data warehouses to process its operations. But it creates a challenging problem for data administrators when volume of information (big data) is increased [1]. Data warehouse of an enterprise integrates the data from several sources of the enterprise or organization to support enterprise wide analyzing, reporting, planning, and decision making. When combining data from different sources for implementing a data warehouse, it aware of possible systematic differences or conflicts. It depends on the accuracy of data [2].

At the time of combining data from various sources, record duplicates will occur. Due to this, many problems arise like quality loss, performance degradation and increasing operations cost. To stay away from these types of problems, data preprocessing steps are executed. They are data cleansing, data integration, data transformation and data reduction [3].

Data cleansing is the method of identifying and determining and correcting the corrupt, unwanted, faulty, and inconsistent data in exact records from a record set, table or database to enhance the data quality. It has de-duplication as an important process to find and remove duplicates which are several representations of the same real world entity. Most important objective of data cleansing is to reduce time and difficulty of the mining process and increase the accuracy.

II. RELATED WORK

Lalitha et al. [1] presented an analysis of record duplication techniques and algorithms that discover and eliminate the duplicate records. They concluded that the existing algorithms require extra memory for de-duplication. Varsha et al. [2] proposed flow diagram for applying appropriate similarity measure on appropriate data to identify the duplicates properly. Elgamal et al. [3] proposed to general framework for the data cleaning process. It has six steps namely selection of attributes, formation of tokens, Selection of the clustering algorithm, similarity computation for the selected attributes, selection of the elimination function and finally merge.

Bilal Khan et al. [4] proposed a de-duplicator algorithm. It is based on numeric conversion of entire data. Waykole et al. [5] presented a survey about genetic programming approach along with hash based similarity to eliminate the replicas and gets the optimization solution to de-duplication of records. Rohit et al. [6] developed an algorithm for eliminating duplicates in dimensional tables in a data warehouse which are usually related with hierarchies. Thilagavathi.S presented an improvement of the effectiveness of the indexing techniques for record linkage and de-duplication by executing in FEBRL format.

Bassma et al. [8] presented a survey an up to date evaluation of the existing literature in duplicate and near duplicate detection in web. Nishand et al. [9] proposed indexing techniques namely blocking index, sorting index and bigram indexing which are used with a alteration of existing techniques that reduces the difference in the quality of the blocking results. Prerna et al. [10] presented a review of the impacts of poor data quality on a typical enterprise. Sapna et al. [11] presented an summary of data cleaning problems, data quality, cleaning approaches and assessment of data cleaning tool.

Rajashree et al. [12] presented an abstract of different algorithms available to clean the data to meet the growing demand of industry and the need for more standardized data. Divya et al. [13] proposed a genetic programming from approach to record de-duplications that joins several different pieces of evidence extracted the data content to find a de-duplication task that is able to find whether two entries in a repository are replicas or not. Chitra et al. [14] presented tests to show that some parameter choices can improve the results up to 30%.

Syed et al. [15] proposed a provenance model that integrates both content based and trust based factors for ordering the results as original or near duplicates. Supriya et al.[16] proposed a two stage sampling selection (T3S) model which has two stages. In the first stage, the strategy is recommended to produce balanced subsets candidate pairs which are to be labeled. In the second stage to create smaller and more informative training sets. They proposed an Adaboostalogithm which provides better accuracy. They presented the proposed method by experiments. Jiang et al. [17] proposed an algorithm of bibliographic descriptions to reduce the problems caused by the existence of duplicate bibliographic records in a database. Kadus et al. [18] proposed architecture and windowing algorithm for election databes. It gives more efficiency and accuracy for record linkage. It is also time consuming and rapid process.

III. OBJECTIVE

1. This paper focuses about data cleansing and de-duplication to ensure data accuracy.
2. It inspires to find better architecture and algorithm to improve accuracy in the big data.

IV. AN OVERVIEW OF THE EXISTING SYSTEMS

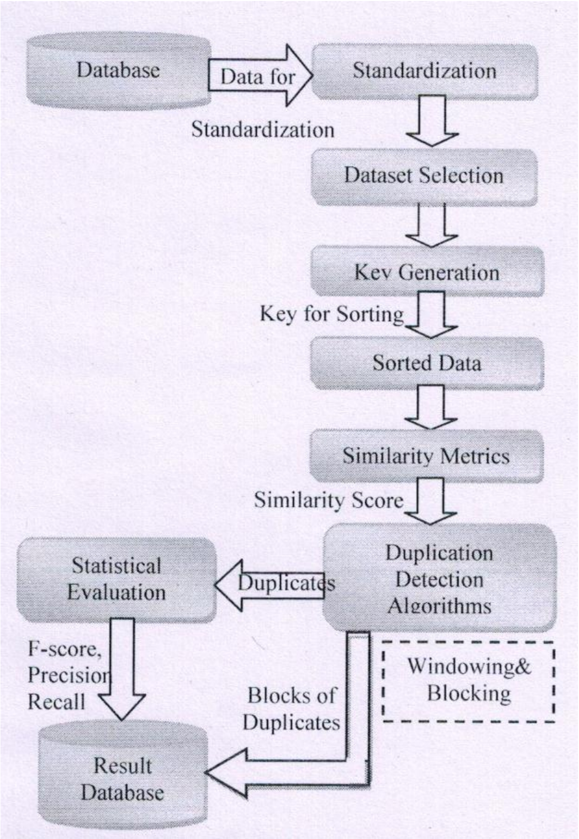

Figure1- Flow diagram of De-duplication

A.Flow diagram of de-duplication

Validation of duplicate detection is based on the similarity measure as well as windowing and blocking algorithm. It used the adaptive windowing algorithm for maintain the effectiveness.

Standardization converts the data in particular or specific standardize format. Key generation is very important and necessary task in detection of duplication. Key is selected as per categories of data set. It has blocking and windowing algorithm (see Fig.1)[2].
Limitations

1. It selected only 500 names and addresses for implementing the flow diagram.

2. It couldn't process large volume of data (big data) which are from several sources.

B.Framework for Data Cleaning

The Data cleaning framework was designed with six steps as selection of attributes, formation of tokens, selection of clustering algorithm, similarity computation for selected attributes, selection of elimination function and merge (see Fig.2)[3].

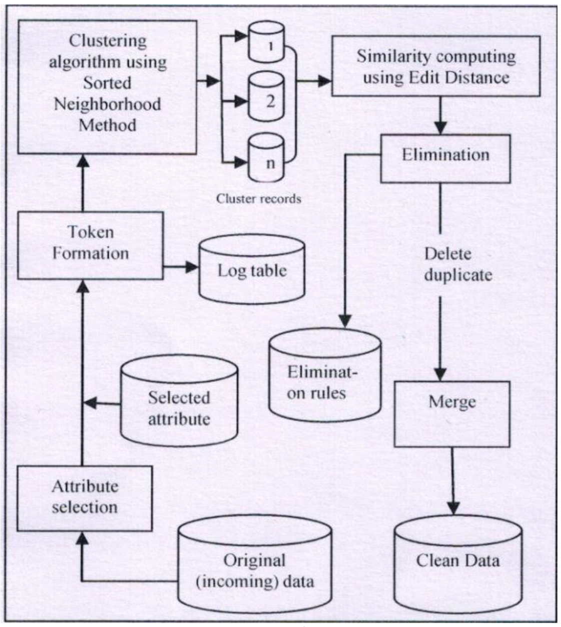

Figure2- Framework for data cleaning

1. Selection of attributes reduce the number of attributes by removing unrelated and redundant attributes, which have no significance in the classification task.

2. Formation of tokens removes typographical errors and abbreviations in data fields using Token based data cleaning algorithm.

3. Selection of clustering algorithm splits the data sets into blocks.

4. Similarity computation for selected attributes identify tuples where closeness is evaluated using a variety of similarity functions chosen to suit the domain applications.

5. Selection of elimination function removes the duplicate records from one cluster or many clusters.

6. Merge collects records as a single cluster. The user must maintain the merged record and the prime representative as a separate file in the data warehouses.

Limitations

1. It selected only two attributes for implementing the framework.

2. It couldn't handle large volume of data (big data) which are from different sources.

IV. PROPOSED SYSTEM

A. Architecture for Automated Data Quality Checking

Proposed system is divided into Sources, Stages of Migration and Target Data Warehouse. Sources are the different origins of data. Stages of migration are having five major steps to migrate data from sources to target.

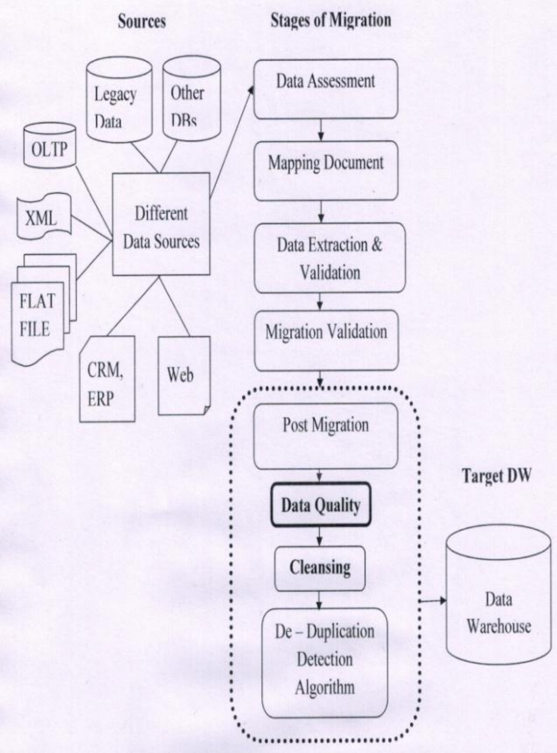

Figure3- Architecture for automated data quality checking

Data assessment stage is identified data sources. System extracts and queries are processed to conduct user interviews on the data migration process. Migration scope and validation strategy are examined here to develop work plan.

Mapping document stage is based on the work plan document of previous stage. This creates the environment of source and destination. Validation and transformation rules are developed by this stage only.

Data element mappings, tables, scripts jobs to automate the extraction are created by Data extraction and validation stage. After the creation it checks them also. Data are collected from the source system and mock migration will be conducted to validate the environment of the migration process.

In Migration validation stage only data is moving (pilot migrations) from sources to destination data warehouse. This performs specific customizations on target database and application. This makes data validation and prepares migration validation reports and data movement metrics which are reviewed here.

Post Migration stage is the important stage of the proposed system. Data quality has six important steps that are profiling, cleansing, standardization, matching enrichment and monitoring. Cleansing can do de-duplication to ensure accuracy using de-duplication detection algorithm.

Finally Target (data warehouse) is having data without duplicates. This ensures accuracy which is one of the parameters of data quality. So, the proposed system will handle large volume of big data.

B. De-Duplication detection algorithm

```
Input: Databases with duplicates
Output: Databases without duplicates
Begin:
    For database i=1 to d (last database)
    For table j=1 to t (last table)
    For row k =1 to n (total number of records)
        1. Apply wrinkler similarity measure
        2. Compare row k and k+1
        3. If wrinkler value is 1for all attributes
eliminate duplicates
        4. Repeat steps 2 & 3 until all duplicates
will be eliminated.
End
```

V. Conclusion

Data Cleansing is really a key success factor and an enabler of big data migration projects without duplicates. This paper described architecture for automated data quality checking using de-duplication detection algorithm. Many frameworks are available to handle data quality problems, but they couldn't process the large volume. Volume is one of the characteristics of big data which refers above 1 TB of data. So, the proposed architecture is handled big data in efficient manner.

References

[1] Lalitha.L, Maheswari.B, Dr.Karthik.S, *"A Detailed Survey on Various Record Deduplication Methods"*, International Journal of Advanced Research in Computer Engineering and Technology, Volume **1**, No.**8**, October **2012**, ISSN: **2278-1323.**

[2] VarshaWandhekar, ArtiMohanpurkar, *"Validation Of Deduplication In Data Using Similarity Measure"*, International Journal of Computer Applications, Volume **116**, No.**21**, April **2015**, ISSN: **0975-8887**.

[3] A.F.Elgamal, N.A.Mosa, N.A.Amasha, *"Application Of Framework For Data Cleaning To Handle Noisy Data In Data Warehouse"*, International Journal of Soft Computing and Engineering, Volume **3**, No.**6**, January **2014**, ISSN: **2231-2307.**

[4] Bilal Khan, AzharRauf, HumaJaved, Shah Khusro, *"Removing Fully And Partially Duplicated Records Through K-Means Clustering"*, International Journal of Engineering and Technology, Volume **4**, No.**6**, December 2012.

[5] J.R.Waykole, S.M.Shinde, *"A Survey Paper On Deduplication By Using Genetic Algorithm Alongwith Hash Based Algorithm"*, International Journal of Engineering Research and Applications, Volume **4**, Issue **1**, January, **2014**, ISSN: **2248 -9622**.

[6] Rohitananthakrishna, SurajChaudhari, VenkateshGanthi, *"Eliminating Fuzzy Duplicates In Data Warehouses"*, Proceedings of the 28th VLDB Conference, Hong Kong, **China**, 2002.

[7] Thilagavathi.S, *"Record Linkage And Deduplication Using FEBRL Frameworl And Block, Sorting, Bigram Indexing Techniques"*, International Journal of Innovative Trends and Emerging Technologies", Volume **1**, No.**1**, March **2014**, ISSN: **2349-9842**.

[8] BassmaS.Alsulami, MaysoonF.Abulkhir, FathyE.Eassa, *"Near Duplicate Document Detection Survey"*, International Journal of Computer Science and Communication Networks, Volume **2(2)**, **2012, 147-151**, ISSN: **2249-5789**.

[9] Nishand.K, Ramasami.S, T.Rajendran, *"An Efficient Way Of Record Linkage System And Deduplication Using Indexing Techniques, Classification And FEBRL Framework"*, International journal of Emerging Science and Engineering, Volume **01**, Issue **07**, May-**2013**, ISSN: **2319-6378**.

[10] PrernaS.Kulkarni, Dr.J.W.Bakal, *"Survey On Data Cleaning", International Journal of Engineering Science and Innovative Technology"*, Volume **3**, Issue **4**, No. **2**, July -**2014**, ISSN: **2319 – 5967**.

[11] Sapna Devi, Dr.ArvindKalia, *"Study Of Data Cleaning & Comparison Of Data Cleaning Tools*", International Journal of Computer Science and Mobile Computing, Volume **4(3),** pp. **360–370**, March **2015**.

[12] RajashreeY.Patil, Dr.R.V.Kulkarni, *"A Review Of Data Cleaning Algorithms For Data Warehouse Systems"*, International Journal of Computer Science and Information Technologies, Volume **3**, Number **5**, **2012**. ISSN: **5212 -5214**.

[13] seetalamDivyaManusha, ValivetiKarthik, PrathipatiRatna Kumar, *"De-Duplication Of Citation Data By Genetic Programming Approach"*, International journal of Recent Advances in Engineering & Technology, Volume **1**, Issue **3**, **2013**, eISSN:**2374-2812.**

[14] L.Chitra Devi, S.M.Hansa, Dr.G.N.SureshBabu, *"A Genetic Programming Approach For Record Deduplication"*, International Journal of Innovative Research in Computer and Communication Engineering, Volume **1**, No**.4**, June **2013**, ISSN: **2320-9798.**

[15] Y.SyedMudhasir, J.Deepika, S.Senthilkumar, and G.S.Mahalakshmi, *"Near Duplicates Detection And Elimination Based On Web Provenance For Effective Web Search"*, International Journal on Internet and Distributed Computing Systems, Volume **1**, No.**1**, August **2011**.

[16] SupriyaAllampallewar, J.Ratnaraja Kumar, *"A Survey Study ForDeduplication In Large Scale Data"*, International Journal of Advanced Research in Computer and Communication Engineering, Volume **5,** No.**2**, February **2016**.

[17] AnestisSitas, SarantosKapidakis, *"Duplicate detection algorithms of bibliographic descriptions"*, Library Hi Tech., Volume **26**, No.**2**, **2008**, ISSN:**0737-8831.**

[18] S.B.Kadus, H.A.Sawant, S.S.Tilekar and H.D.Zendage, *"Data deduplication of election database using windowing algorithm"*, International Journal of Current Research in Science and Technology, Volume **1**, No.**4**, **2015**.