# PARALLEL & DISTRIBUTED APPROACH FOR CLEANING DATA IN DATA WAREHOUSE

Ayanka Ganguly[1], Ms. Pooja Mehta[2]

[1] *M.E. Student, Dept. of Computer Engineering, SAL-ITER, Gujarat, India.*
[2]*Assistant Professor, Dept. of Computer Engineering, SAL-ITER, Gujarat, India.*

**ABSTRACT**

*The quality of data can only be improved by cleaning data prior to loading into the data warehouse as correctness of data is essential for well-informed and reliable decision making. The quality of the data can only be produced by cleaning data prior to loading into data warehouse. Data Cleaning is a very important process of the data warehouse. It is not a very easy process as many different types of unclean data can be present. So correctness of data is essential for well-informed and reliable decision making. Also, whether a data is clean or dirty is highly dependent on the nature and source of the raw data. Thus, I have proposed a work on parallel & distributed approach for cleaning data gathered from external sources by combining two or more data cleaning techniques or algorithms for error correction by correcting the records from a list of suggested words mentioned in the dictionary and which will also eliminated the duplicate records that will provide a more accurate and cleaned data.*

**Keyword:** *- Data Cleaning , Modified PNRS, Modified Transitive Closure, Enhanced Technique, and Semantic Matching.*

## 1. INTRODUCTION

Data warehouse is important for storing large amount of data. The data stored in the data warehouse should be correctly entered, accurate and relevant. Data quality is an important factor in the data warehousing projects. If it consists of incorrect or inaccurate data i.e. dirty data, then it may cause various problems like taking incorrect decisions or actions based on that dirty data. Dirty data must be detected and corrected to improve the data quality. Data cleaning is an essential step in populating and maintaining data warehouses [1]. Data cleaning or scrubbing is the process of correcting or removing corrupt or inaccurate records from the database so that downstream data analysis is reliable and accurate. This is usually accomplished through an Extract-Transform-Load (ETL) process in which the data is collected from different sources, the extracted data is send to the next phase which is transformation, in transformation process series of rules are applied to the extracted data and then all the cleaning of data is performed[2]. Once the cleaned data is ready it is loaded into the data warehouse.

An attempt has been made in this paper to provide a more cleaned and an accurate data by combining modified versions of PNRS and Transitive Closure, Enhanced Technique and Semantic matching approaches all together. The remainder of the paper is as follows. **Section 2** describes the related work in the field of data cleaning and the various approaches that were proposed for data cleaning. **Section 3** provides a brief description about all the approaches that will used in our paper, **Section 4** introduces the proposed methodology for cleaning data by combining four data cleaning approaches , and **Section 5** shows some experiment results for the my proposed methodology. **Section 6** gives conclusions and recommendations for future work.

## 2. RELATED WORK

Many methods are proposed by researchers for data cleaning where cleaning based on dictionaries is a common approach. Dictionary used for the data cleansing process can be both standard dictionary as well as organization based dictionary.

**Cihan Varol1, Coskun Bayrak1, Rick Wagner and Dana Goff [8]** proposed the Personal Name Recognizing Strategy (PNRS) which was developed to provide the closest match for a misspelled name.

**W. N. Li, R. Bheemavaram, X. Zhang [9]** introduces a record grouping problem called transitive closure which reduces the number of records fed to the analysis tools by grouping related records into groups.

**Arindam Paul, Varuni Ganesan, Jagat Sesh Challa, Yashvardhan Sharma[1]** proposed a HADCLEAN algorithm which provides a hybrid approach for cleaning data which combines modified versions of PNRS and Transitive closure algorithms. The contemporary PNRS algorithm was correcting the spelling mistakes in the data on the basis of any Standard English dictionary. The modified PNR uses organization specific dictionary, along with a standard dictionary, for checking the spelling mistakes. In his Modified Transitive Closure more than one key is used for matching the records into one group.

**Russell Deaton, Thao Doan, and Tom Schweiger [7]** proposed a technique called latent semantic indexing will be applied to data matching and identifying groups of records that represent the same business entity.

**Prerna S. Kulkarni, J.W. Bakal [2]** proposes a combined approach of Semantic Data Matching algorithm along with the HADCLEAN algorithms to get better results in data corrections. By using Semantic Data Matching, it can taken care by keeping a unique consistent name based on the semantic similarity between the attribute values in different documents.

**Dr. Mortadha M. Hamad, Alaa Abdulkhar Jihad proposed [4]** the enhanced algorithm for cleaning data that allows user interaction by selecting the rules and any sources and the desired targets. It works well on the quantitative data and any data that have limited values.

**Ashwini M.Save,Seema Kolkur [3]** makes an attempt to clean the data by combining different approaches such as Improved PNRS, Enhance Technique and Transitive closure which will correct mis-spellings on textual data, errors in quantitative attributes and removal of duplicities respectively.

## 3. BACKGROUND

In this section, a brief description on Modified PNRS and Transitive Closure, Enhanced Technique and Semantic matching is given.

**3.1 Modified PNRS-** The PNRS algorithm, proposed by C. Varol et al.[8] corrects the typographical errors present in the raw data, using standard dictionaries. It consists of the 'Near Miss Strategy where two words are considered "near" if they can be made identical – by interchanging 2 letters, by changing/adding/deleting a letter or by inserting or removing blank spaces.

The contemporary PNRS algorithm was correcting the spelling mistakes in the data on the basis of any Standard English dictionary.

The modified PNRS uses organization specific dictionary, along with a standard dictionary, for checking the spelling mistakes. This is important because most of the verbal data present in data warehouses are official data and contain organizational jargons, sometimes even limited to a particular organization [1].

**3.2 Modified Transitive Closure-** In this technique the records are matched on the basis of matching of the keys(attributes). Transitive Closure algorithm matches two or more records into one group when one of the key (attribute) matches between the two records[9]. It also needs manual intervention to check whether any duplication or correction in the records have been done or not.

But in Modified Transitive Closure more than one key is used for matching the records into one group. It does not even require any manual intervention to check the corrected records. It is done in 2 levels[1]:

At the first level, it divides keys in three categories i.e. into primary(either one-to-one or one-to-many), secondary(relatively unique), and territory(not so unique).Then at second level the ordering of keys is done on their decreasing priority of Uniqueness/importance.

**3.3 Enhanced Technique-** The enhanced algorithm for cleaning data allows user interaction by selecting the rules and any sources and the desired targets. It works on the quantitative data and any data that have limited values. It solves all the quantitative errors and problems such as Lexical Error, Domain Format Error, Irregularities, Integrity Constraint Violation, Duplicates and Missing Values[4].

**3.4 Semantic Matching-** In Semantic Data Matching, a unique consistent name is given based on the semantic similarity between the attribute values in different documents. It gets some reference sets from the data based on the

key values[2,7]. These keys will match that key and will replace the words which has the similar meaning with the standard words.

## 4. THE PROPOSED METHODOLOGY

My proposed data cleaning methodology used here is by combining the modified PNRS and Transitive Closure, Enhanced Technique and Semantic Matching approaches all together. The modified PNRS algorithm will correct the mis-spelled words in the textual data then the enhanced technique will be used to detect and correct errors on quantitative data. Thus, data cleaning system can solve errors on both 'Textual' as well as 'Quantitative' data. After applying these algorithms i.e. Modified PNRS and Enhanced technique; The Transitive closure algorithm can be applied on the data which will remove duplicate set of data and fill missing values. And lastly applying the Semantic Matching approach, records having similar meaning or abbreviations will be replaced with the standard names specified in the dictionary in order to get more precise data..

Hence rather than using these algorithms separately i.e. only to correct text data or quantitative data or to avoid duplicate records, this proposed methodology covers all areas of the data i.e. text fields, quantitative fields, then removal of duplication and replacing similar names with standard names. So by combining the four approaches all together, this methodology will provide more accurate and cleaned data.

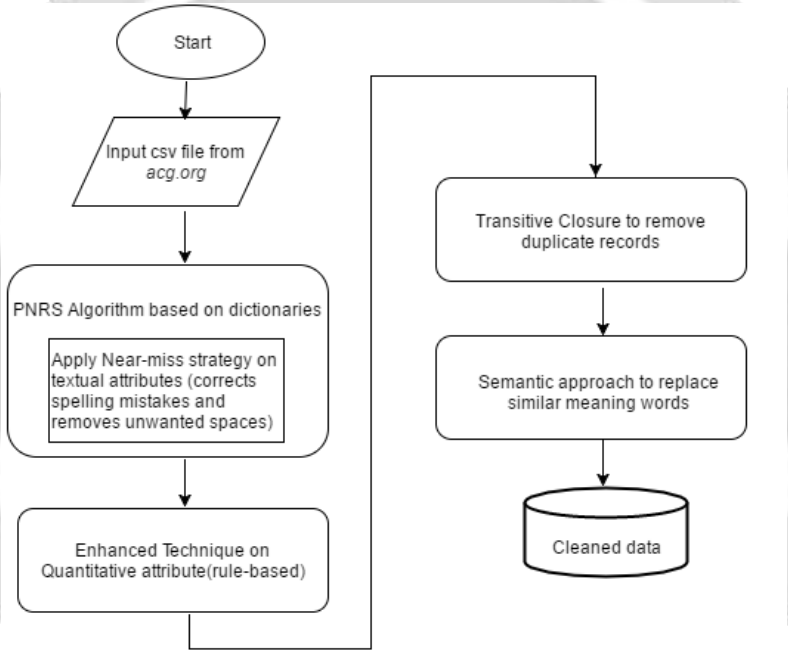


**Fig -1**: Proposed Flowchart

## 5. EXPERIMENT ANALYSIS

After performing the experiments, we have analyzed that the number of records cleaned during the PNRS phase to Enhanced Technique phase is less than number of records cleaned from PNRS to Semantic phase. We get a higher number of corrected records if we apply all the four approaches simultaneously. Chart-1 shows the bar chart of the number of records corrected when we applied the PNRS to Enhanced approach and then from PNRS to Semantic matching approach. It is also observed that while performing from PNRS to Enhanced Technique approach 62.98% of noise(uncleaned data) is removed and from Transitive Closure to Semantic Matching 69.92% of noise is removed. Thus, more 7% of noise will be removed if we perform all the approaches ,i.e, from PNRS to Semantic Data Matching approach as shown in Chart-2.

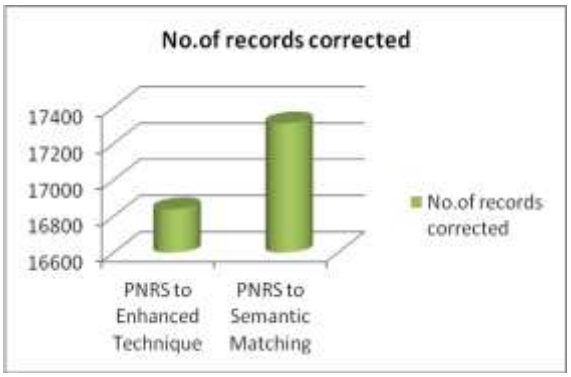


**Chart -1**: Analysis of corrected records

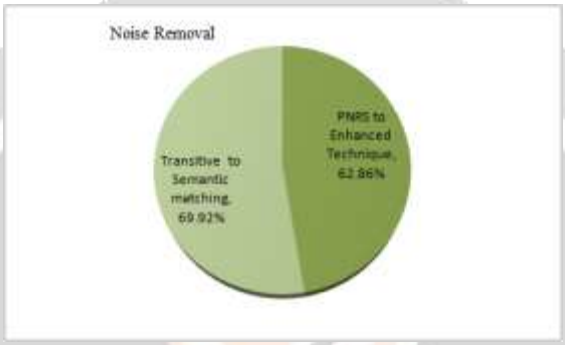


**Chart -2**: Analysis of noise removal

## 6. CONCLUSION & FUTURE WORK

Data Cleaning is a very important part of the data warehouse management process. It is not a very easy process as many different types of unclean data (bad data, incomplete data, typos, etc) can be present. Many attempts have been made to clean the data such as by using the PNRS technique, Enhanced technique, Transitive Closure approach and Semantic Data matching approach. The hybrid approach for data cleaning uses the modified versions of PNRS and Transitive closure. In my work, by combining all the approaches together more number of records are deleted. While applying the PNRS to Enhanced technique 62.86% of unwanted data were cleaned whereas applying from PNRS to Semantic data matching approach 62.92% of unwanted data were cleaned. Thus, the data was cleaned by 7% more accuracy if we combine all the approaches together resulting in to provide more cleaned and accurate data which can be loaded into the data warehouse.

As of now, my approach has been tested on data with 26773 records. This could be tested on huge Enterprise Data that can give us better knowledge of performance and efficiency of this approach for data cleaning.

## 7. REFERENCES

[1] Arindam Paul, Varuni Ganesan, Jagat Sesh Challa, Yashvardhan Sharma, "***HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses***", Information Retrieval & Knowledge Management (CAMP), International Conference, 13-15 March 2012.

[2] Prerna S. Kulkarni, J.W. Bakal ,"***Hybrid Approaches for Data Cleaning in Data Warehouse***", International Journal of Computer Applications (0975 – 8887) Volume 88 – No.18, February 2014.

[3] Ashwini M.Save,Seema Kolkur, "***Hybrid Technique for Data Cleaning***", International Journal of Computer Applications, 2014.

[4] Dr. Mortadha M. Hamad, Alaa Abdulkhar Jihad ,"***An Enhanced Technique to Clean Data in the Data Warehouse***", Developments in E-systems Engineering (DeSE), 2011.

[5] Erhard Rahm , Hong Hai Do ,"***Data Cleaning: Problems and Current Approaches***", IEEE Computer Society Technical Committee on Data Engineering, 2001.

[6] Wing Ning Li, Johnson Zhang, Roopa Bheemavaram,"***A Parallel and Distributed Approach for Finding Transitive Closures of Data Records: A Proposal***", ALAR Conference on Applied Research in Information Technology, 2006

[7] Russell Deaton, Thao Doan, and Tom Schweiger, "***Semantic Data Matching: Principles and Performance***", Springer Science, 2010.
[8] Cihan Varol, Coskun Bayrak, Rick Wagner, and Dana Goff, "***Application of the Near Miss Strategy and Edit Distance to Handle Dirty Data***", Springer Science, 2010.
[9] W. N. Li, R. Bheemavaram, X. Zhang, "***Transitive Closure of Data Records: Application and Computation***", Data Engineering - International Series in Operations Research & Management Science, Springer US, vol. 132, pp. 39-75, 2010.
[10] Heiko Müller, Johann-Christoph Freytag, "***Problems, Methods, and Challenges in Comprehensive Data Cleansing***", Humboldt-Universität, Berlin, Germany
[11] Manjunath T.N, Ravindra S Hegadi, Ravikumar G.K, "***Analysis of Data Quality Aspects in Data Warehouse Systems***", International Journal of Computer Science and Information Technologies, Vol. 2 (1) , 2010.
[12] Arthur D. Chapman, "***Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data***", version 1.0, Global Biodiversity Information Facility, Copenhagen, 2005.
[13] Nidhi Choudhary, "***A Study over Problems and Approaches of Data Cleansing/Cleaning***", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014.
[14] Jonathan I. Maletic, Andrian Marcus, "***Data Cleansing: A Prelude to Knowledge Discovery***", Springer US, 2010.
[15] Surajit Chaudhuri, Umeshwar Dayal, "***An Overview of Data Warehousing and OLAP Technology***", ACM SIGMOD Record, Volume 26 Issue 1, March 1997.