REDUCE DATA REDUDANCY IN BIGDATA USING DATA PROFILING TECHNIQUE WITH CROSS TABLE RELATIONSHIP SCHEMA

K. Makesh Babu[#], Dr. K. Mohan Kumar[*]
[#]Research Scholar, [*]Research Supervisor & Head
PG and Research Department of Computer Science
Rajah Serfoji Government College (Autonomous), Thanjavur 613005
Affiliated to Bharathidasan University, Trichirappalli, Tamil Nadu, India.
Corresponding Author: [*]tnjmohankumar@gmail.com

Abstract — Data profiling is so important for all private and government organizations, because they store huge volume of data in various formats. Data profiling focus heavily on data with various formats created from various data sources. Data profiling plays an important role to maintain the data quality in data governance. This paper mainly focuses on data redundancy in large volumes of data generated by various data sources. Cross-table profiling is a data profiling technique which is used to reduce the data redundancy. In this work normalization is added into the cross tabling technique. This new method called Cross Table Relationship Schema (CTRS) is used to identify and eliminate inter-table transitive dependency. This technique uses foreign key analysis to identify the orphaned records and determination of semantic and syntactic differences in different tables. In this new model the redundant data is eliminated in better way compared with existing models.

Key words: Big data, Data Redundancy, Data profiling, Cross Table analysis, relational data schema, Normalization

I- INTRODUCTION

In e-Governance applications the transaction volume per day is growing constantly due to more transaction in all sectors by the common citizens. This abnormal growth of data in e-Governance is referred to as BIG data. The challenges with big data are data quality, data discovery, data storage and the security. The traditional data processing applications and database management systems face difficulty to store and process this huge data. So, a big data analysis tool with efficient technologies is needed in this circumstance. Only then these applications can able to generate faster analytical report from huge amount of historical data. Normally the huge data generated through various data sources contain the redundant data. While generating reports for future action, there should not be redundant data in the database. It means that the redundant data affect the data quality. To eliminate redundant data from the organization's data base, there is a need of data profiling [1].

Data redundancy occurs when the same set of data is stored in two or more separate storage places. This redundancy is happen in all the organization. Every organization has a challenge of identifying duplicate data entries in a data bases. They are finding a way to reduce data redundancy efficiently, so it can help for future analysis. This is done by only Cross table profiling technique. The Table 1 given below shows the advantages and disadvantages of data redundancy occur in a data base [2].

Table 1: Advantages and Disadvantages of Data Redundancy

S.No.	Advantages	Disadvantages
1	Alternative data backup method	Possible data inconsistency
2	Better data security	Increase in data corruption
3	Faster data access and updates	Increase in database size
4	Improved data reliability	Increase in cost

A. Data profiling

Data profiling is the process of fetching source data and collects statistics information about that data for future planning and decision making. Data profiling helps to discover, understand and organize the data in an efficient way. The significant benefits derived from data profiling are given below [3].

- Discovering business knowledge
- Validation with metadata
- Pattern matching and the use of basic statistics
- Frequency counts and outliers
- Redundancy and similarity discovery
- Business rule validation

B. Data Profiling Discovery Analysis

Data profiling encompasses three major techniques in discovery analysis. The following Figure 1 shows the different types of discovery analysis and their techniques. This work mainly deals with Cross table analysis. Cross-table analysis examines overlapping value sets across different tables such as foreign key analysis and identification of orphaned records.

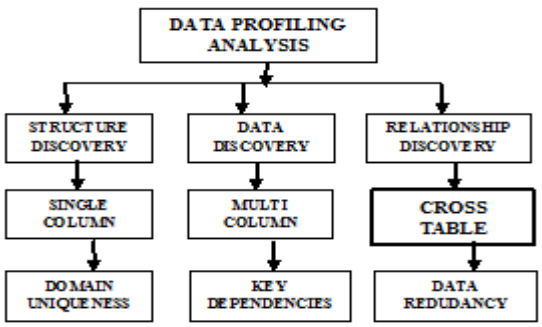

Figure 1. Data Profiling Discovery Analysis

C. Structure discovery Analysis

Structure analysis uses some rules that define how columns relate to other columns to form tables. Structure analysis deals with primary keys, foreign key, redundant data columns, column synonyms and other referential constraints to maintain the correctness of data. The structure cannot be defined or known or not reflected accurately in source system definitions. Knowing the columns structure, knowing the connections between object tables and knowing the rules on how tables are relate to each other. Identifying these inaccuracies in the data is very important if the data is to be moved to a well-defined target database system. Structure analysis is helpful for how data is structured and useful for correcting the metadata. In addition structure rule involves more than one column or more than one row of a single table.

D. Data discovery Analysis

This process assesses the quality of individual pieces of data. Once the invalid data values identified within columns and the entire structure rule violations, it is time to get more specific rules that require for multiple columns. These rules are defined as data rules. A data rule is a rule that specifies a condition that must hold true across one or more columns at any point in time. Because data rules involve more than one value, it can be identified which value within the error set is wrong or the combination of values is wrong.

E. Relationship data discovery Analysis

Complex data rules require values to execute multiple table data. This detects connections, similarities, differences and associations between data sources. It identified the inaccurate data or redundant data that is hidden within a much larger set of data rows. This rule capture all rows with inaccurate data, store it in the inaccurate data fact table. Finally it is identified and it is transferred to accurate fact table. This redundancy analysis is done completely by Cross table analysis. [4]

The main Objectives of this research article are:

- To identify orphaned records in the related table.
- To identify overlapping records across multiple tables.
- To avoid possible data inconsistency
- Decrease in data corruption
- Decrease in database size
- Decrease in cost

II-LITERATURE REVIEW

Ziawasch Abedjan et.al. (2015) provided a comprehensive survey of data profiling and the set of activities and processes to determine metadata about a given database. The single-column profiling tasks such as identifying data types, value distributions and patterns are discussed. Also, it dealt with a multi column task to detect various kinds of dependencies. Finally, it concluded that the future of data profiling tasks with relational databases is essential [5].

Kim Nguyen (2017) supported some transactions to demonstrate many-to-many relationship between the two tables. And also it determines the relationships among the tables.

The design Process figures out the Primary Key for each table and Reduces redundancy. Finally it provided the information carefully among tables to eliminate data redundancy. Duplicate data wastes space and can lead to inconsistency [6].

M Misbachul Huda et.al. (2015) dealt with two different databases such as RDBMS and non-RDBMS. Also it dealt with two key issues such as growing size of the databases and increasing of data complexity. The map reduce model is designed to process large volume of data in to smaller tasks. Finally it concluded that the relational databases have high popularity. Non RDBMS have more attributes to resolve the problem of big data [7].

M. Nalini (2016), mainly focused on the redundant data that is fetched by the Quick Search Bad Character (QSBC). QSBC function compares the entire data with patterns taken from index table created for all the data persisted in the DBMS to easy comparison of duplicate data in the database. It examined the database for the performance evaluated in terms of time and accuracy [8].

III-METHODOLOGY

Cross-table analysis converge the value sets across different tables. The redundancy analysis is provided by cross-table profiling. In that the values in each pair of columns respectively selected from different tables are evaluated for set intersection. The key capability for cross-table profiling includes foreign key analysis, identification of orphaned records and determination of semantic and syntactic differences. Also the Cross-Table Anomalies includes Referential consistency, Syntactic inconsistency and Semantic inconsistency

Cross-table profiling uniquely reviews the values within column sets in different tables potentially intersects and overlaps. Column datasets overlap across tables suggests dependencies, relationships, and redundant storage that are mapped together. This is done only by modifying the relationship data schema and normalization. The Figure 2 shows the process of identifying redundant data in a big data. Redundant data occurs in Big data is processed and identified using Cross table analysis. This analysis, design the new relational data schema process and normalization.

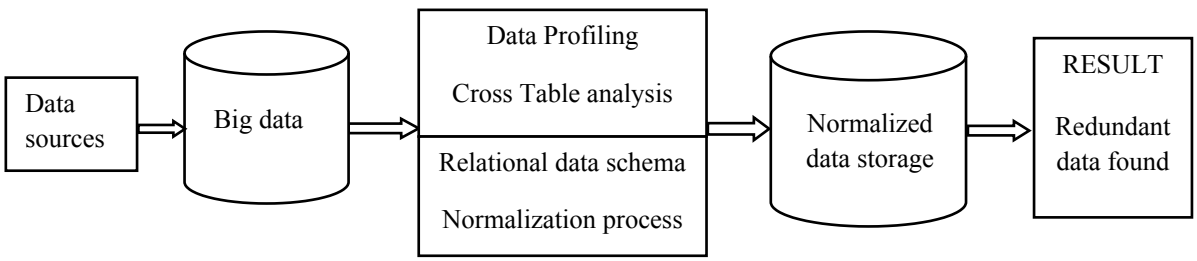

Figure 2. Cross Table Relationship Schema profiling process

A. Relational data schema

Data redundancy is a data organization issue that allows the unnecessary duplication of data within datasets. A schema is easy to understand and navigate, with dimensions joined only through the fact table. These joins are more significant to represent the fundamental relationship between the data process.

The relational database schema is the primary element of the relational database. This allows for database management based on entity relationships, making them easy to organize according to volume of data. This work aims to avoid redundancy by changing the multiple fields of a database and their relationship. The database relational model is responsible for to port any changes to the data field across the database. Normally redundant data makes wastes valuable space and creates troubling the database.

A transitive dependency in a relational database is an indirect relationship between values in the same table that may lead to functional dependency. The Third Normal Form (3NF) to eliminate any transitive dependency and improves data integrity

B. Relational Database normalization

To eliminate redundant data from a big data, must take intensive care to organize the data in a database tables. Normalization is a method of organizing the data to prevent redundancy. Normalization involves establishing and maintaining the integrity of data tables as well as eliminating inconsistent data dependencies. Normalization is the process of efficiently organizing data in a relational database so that redundant data is eliminated. The 3NF plays an important part in this model and eliminate fields that do not depend on the key.

C. Proposed CTRS profiling algorithm

Step 1: Develop a logical data model for any user interface application with normalization principles to represent entities.

Step 2: Combine normalized data requirements into one consolidated logical relational database model. Normalize the relations using 3NF.

Step 3: Translate the conceptual E-R data model for the application into the normalized data model. Then Merge the relations after representing entities and relationships.

Step 4: Eliminate inter-table transitive dependency. If a transitive dependency occurs, then remove the transitively dependent attributes. This is achieved by placing the attributes in a new relation along with a copy of the determinant.

Step 5: Finally consolidate the logical database design with the translated E-R model and produce a final logical database model.

D. Data Collection

Two different Dataset collected from government organization through online to execute this proposed CTRS algorithm. This data is related towards with TWAD Board. The Datasets related with the purchase of material from various dealers and the materials distributed to various contractors for pipe lining job. Two different datasets collected for analyzing these algorithms. The Table 1 shows the number of records and redundant data occurs in a datasets.

Table 1. Total number of records

	Total records	Actual number of redundant data
Dataset 1	1700	270
Dataset 2	2350	385

IV-RESULTS AND DISCUSSION

The various methods used at present for Cross-table analysis are Cross-Database Table Relationships, AWS data relationship schema and Relational data base schema. The Table 2 shows the number of redundant data found by each of these algorithms along with the proposed CTRS algorithm. These algorithms tested with two different dataset consist of 1700 and 2350 records for analyzing. The actual redundant records in the datasets are 270 and 385.

Table 2. Comparison of various schema methods

		Redundant data found	
S.No	**Schema design methods**	**Dataset 1**	**Dataset 2**
1	Cross-Database Table Relationships	130	240
2	AWS data relationship schema	145	265
3	Relational data base schema design	125	258
4	Cross Table Relationship Schema	210	296

The following Figure 3 shows the number of redundant data found using various algorithms. This graph shows CTRS algorithm found maximum redundant data occurred in both data sets when compared to other algorithms.

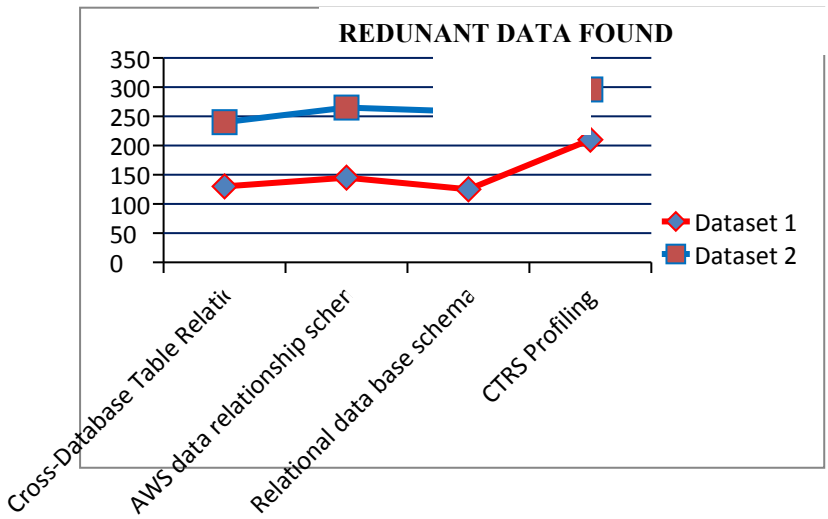

Figure 3. Redundant data found

The Table 3 shows the percentage of redundant data found by each algorithm. In both the dataset the proposed algorithm found maximum percentage than any other algorithms.

Table 3. Percentage comparison of redundant data found

	Percentage of Redundant data in Data set1	**Percentage of Redundant data in Data set2**
Cross-Database Table Relationships	48	62
AWS data relationship schema	54	69
Relational data base schema	46	67
Cross Table Relationship Schema profiling	78	77

The following Figure 4 shows the percentage of redundant data found using various algorithms. This graph shows CTRS algorithms found average of 77.5 percentage redundancy in both datasets. This shows the proposed algorithm performs better than the other algorithms. The major objective of this research is to improve the quality of datasets by reducing data redundancy.

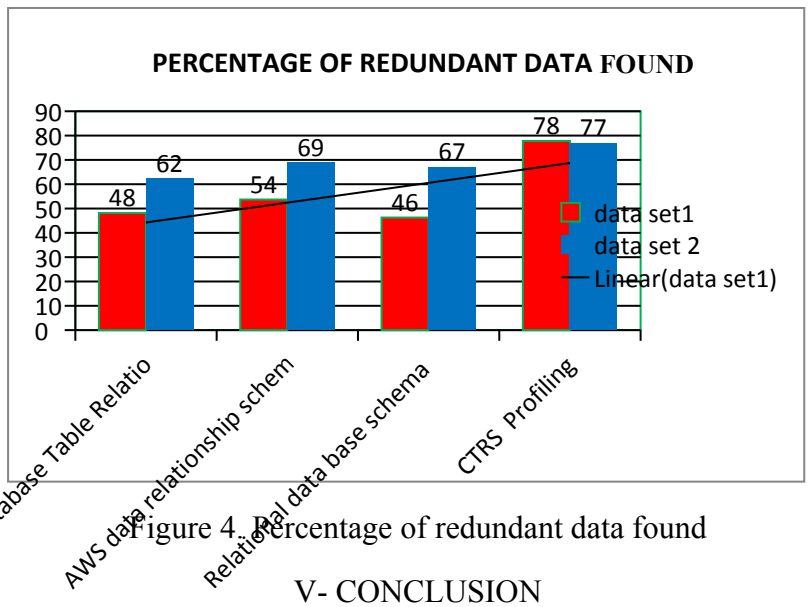

Figure 4. Percentage of redundant data found

V- CONCLUSION

Big data is mainly used for strategic decision making. The data quality plays a major role in big data analysis. Maintaining the quality on big data is a difficult task because of data redundancy. Data profiling take part an important role in maintaining the data quality. The proposed Cross Table Relationship Schema (CTRS) profiling technique is used to reduce the data redundancy at some extent by adding normalization rules. This research work gives an idea to design the relational database to the database designer to improve the data quality in big data.

REFERENCES

1. Preet Navdeep, Dr. Manish Arora and Dr. Neeraj Sharma(2016), "Role of Big Data Analytics in Analyzing e-Governance Projects", Gian Jyoti e-journal, Vol. 6, Issue 2.

2. Muhammad Saleem Vighio, Taooz J. Khanzada and Mukesh Kumar(2017), " Analysis of the effects of redundancy on the performance of relational database systems" , IEEE 3rd International Conference on Engineering Technologies and Social Sciences**.**

3. Brett Dorr and Pat Herbert (2005), "Data profiling: Designing the blueprint for improved data quality", SAS SUGI Procedings 30, Paper 102-30, April 10-13.

4. Suraj Juddoo(2015), "Overview of data quality challenges in the context of Big Data", IEEE, International Conference on Computing, Communication and Security.

5. Ziawasch Abedjan , Lukasz Golab and Felix Naumann(2015), "Profiling Relational Data A Survey", The VLDB Journal 24, 557–581.

6. Kim Nguyen (2017) , "Relational Database Schema Design Overview", medium.com.

7. M Misbachul Huda, Dian Rahma Latifa Hayun and Zhin Martun(2015), "Data Modeling for Big Data. Jurnal ULTIMA InfoSys", Volume1, Issue 11.

8. M. Nalini (2016), "Elimination of Data Redundancy before Persisting into DBMS using SVM Classification" International Journal of Engineering Research & Technology, Vol. 5 Issue 03.