ETL based Cleaning on Database

Arup Kumar Bhattacharjee, Partha Chatterjee, Mukesh Prasad Shaw, Manomoy Chakraborty
Department of Computer Application, RCC Institute of Information Technology

ABSTRACT
The paper analyses the problem of data cleaning and automatically identifying the "incorrect and inconsistent data" in the dataset. Extraction, Transformation and Loading (ETL) are the different steps for cleaning a data warehouse. Authors have implemented different algorithms like: cleanString, cleanNumber, hit ratio, check data dictionary, check metadata etc in addition to various existing data cleaning algorithm like PNRS. This paper tries is to improve the quality of data in the database system. This paper emphasizes on the citizen database system to make it errorless. Some of the results along with certain statistics are also provided here.

Keywords
Data warehouse, ETL, Data Dictionary, Hit Ratio, Dirty Data, Data Cleaning.

1. INTRODUCTION
Data cleaning also known as data cleansing, is the process of identifying the errors contained in a record set, table or data base and correcting them to improve the quality of data. Data quality problems are present in single database, when multiple data sources are combined then the need for data cleaning becomes very important. Data quality means that the data should be accurate, consistent and according to the required type. Data quality problems are present in single data collections due to misspelling during data entry, missing information or invalid data. When multiple sources are needed to be integrated e.g. Data Warehouse, federated database system, then the need of data cleaning increases significantly. In order to provide valid, correct and consistent data, validation of different data representation, elimination of duplicate record and fill up the missing values become necessary.

Data warehouse continuously load and refresh data from different sources. So the probability of containing "dirty data" is high. On the basis of the data present in the data warehouse, different decisions are made. In a data ware house ETL [13] i.e. Extraction, Transformation and Loading is a very important operation where data is extracted from homogeneous or heterogeneous data sources followed by transformation of the data for storing it in proper format. And finally loading the database into the target ware house. Thus data cleaning is very important in maintaining a data warehouse.

Data warehouse continuously load and refresh data from different sources. So the probability of containing "dirty data" is high. On the basis of the data present in the data warehouse, different decisions are made. Thus data cleaning is very important in maintaining a data warehouse.

For a data to be of high quality it should satisfy the following criteria [1]:

- Validity
- Accuracy
- Completeness
- Uniformity
- Consistency
- Data should match the data type
- Timeliness
- Accessibility

This paper is organized in the following way: Section 2 states the different types of reasons for which the quality of data is not up to the mark. Section 3 briefly describes the related work of data cleansing followed by section 4 that illustrates the methods on the basis of which this algorithm is made. Section 5 makes the reader aware about the steps of this algorithm which works on the basis of the statistics given in section 6. Some sample outputs of this work are shown in section 7. Section 8 tells about the significance of this work and section 9 states about the future scope of this work.

2. DATA CLEANING PROBLEMS
The major data quality problems to be solved by data cleaning are:

1. Data from single sources problems.

- Misspellings.
- Redundancy/Duplicates
- Outside domain range
- Data entry errors.

2. Data from multi source problem.

- Naming conflicts at schema level. (Homonyms- same name for different things. Synonyms- different name for same things.)
- Structural conflicts.

3. RELATED WORK
Researchers have proposed various approaches for data cleaning. Various researchers have attempted to clean data using transitive closure algorithm [5]. Many of them used the technique of using data dictionary [1]. In spite of using those techniques, in maximum cases they were not able to make the process automated. It always needed manual intervention. Also all those techniques were used in case of numeric data only. In case of string, these algorithms are not much effective.

So, the authors have tried to clean both types of data i.e. the string type as well as the numeric data. They have tried to repair [14] both the string type and the numeric type data. First the algorithm replaces any particular character by another one depending on some statistical data. Next it omits some character to make the data clean. It also modified and stored data as per the required format. Next session gives the background of the basic algorithms that have been extended to design this algorithm for data cleaning.

4. BACKGROUND OF THE WORK
In this section, a brief discussion about the processes and steps is done that the authors have implemented in this work. As mentioned previously this algorithm emphasized on citizen database. The following rules have applied to it:

4.1 PNRS Algorithm

The Personal Name Recognizing Strategy (PNRS) algorithm [5] was proposed by C. Varol, which corrects the phonetic and typographical errors present in the data, using standard dictionaries. It has two algorithms Near Miss Strategy and Phonetic algorithm.

4.2 Transitive Closure Algorithm

Transitive Closure algorithm [5] for data cleaning was proposed by W.N. Li. This algorithm pre-processes the data to categorize millions and billions of records into groups of related data. The ETL tool using following algorithms processes the individual groups for data cleaning which involves following.

- Identifying and removal of redundancies.
- Filling blank cells.
- Establishment of "group" relationship between different records.

4.3 Define Data Dictionary (Base Table Formulation)

A base table for particular fields like *STATE, CITY and PINCODE* has defined which consists of actual correct data on the basis of which the algorithm performs data correction.

A table called *PINSTART* has defined which contains the state wise pincode's starting digits [Fig.1]. There is also a table *STATE* [Fig. 2] containing capital name of the states.

SNO	ST	PIN
1	DELHI	11
2	HARYANA	12
3	HARYANA	13
4	WEST BENGAL	70
5	WEST BENGAL	71
6	WEST BENGAL	72
7	WEST BENGAL	73

3RD COLUMN SHOWS 1ST 2 DIGITS OF PINCODE,2ND COLUMN SHOWS CORRESPONDING STATE NAME

Fig. 1 : A part of PINSTART table

SNO	ST	CAPITAL
2	Andhra Pradesh	Vijayawada
3	West Bengal	Kolkata
4	Assam	Dispur
5	Bihar	Patna
6	Chandigarh union territory	Chandigarh

Fig. 2 : A part of STATE table

4.4 Hit Ratio Calculation

According to the algorithm, it emphasized on providing a key to the tuples and then applying "HIT RATIO CALCULATION" which was found very effective in replacing the incorrect data by correct data and fill up the missing values.

Consider a table contains following 3 attributes:-

District name, state, pin code. First of all, categorize them into keys. Then consider District as primary key, State as secondary key and pin code as tertiary key according to their dependency on each other [3], [12].

For e.g. :- PINCODE ← DISTRICT ←STATE

Now apply the formulae of hit ratio.

Hit ratio = (no of character matches)/max length (given string from citizen table, value from this actual table)

Fig. 3 : METADATA table of CITIZEN database

SERIAL_NO	ATT_NAME	ATT_TYPE	ATT_DETAILS
1	SSN_NO	NUMBER	PRIMARY KEY,CITIGENSHIP NUMBER
2	FIRST_NAME	VARCHAR2	CITIGEN FIRST NAME
3	MIDDLE_NAME	VARCHAR2	CITIGEN MIDDLE NAME
4	LAST_NAME	VARCHAR2	CITIGEN LAST NAME
5	FATHER_NAME	VARCHAR2	CITIGEN FATHER NAME
6	MOTHER_NAME	VARCHAR2	CITIGEN MOTHER NAME
7	ADDRESS	VARCHAR2	CITIGEN PERMANENT ADDRESS
8	EMAIL_ID	VARCHAR2	CITIGEN VALID EMAIL ID
9	DISTRICT	VARCHAR2	CITIGEN DISTRICT ACCORDING TO ADDRESS
10	PINCODE	NUMBER	CITIGEN PINCODE ACCORDING TO ADDRESS
11	STATE	VARCHAR2	CITIGEN STATE ACCORDING TO ADDRESS
12	SEX	VARCHAR2	CITIGEN SEX :: ACCEPT ONLY MALE OR FEMALE
13	DOB	DATE	CITIGEN DATE OF BIRTH :: DD-MON-YY

This formula will generate a value between 0 & 1. The data corresponding to the max value of hit ratio can replace the corresponding data on citizen table.

4.5 Check METADATA

A table called METADATA has defined which contains the details of that table on which the algorithm will be applied. The algorithm has designed in such a way that it allows the program to check the MEATADATA table for that particular table. The program reads any column from the table and then it also checks the METADATA table for details about that particular column.

Here Fig. 3 shows the details of METADATA table of citizen database.

4.6 Duplicate Elimination

This is a process of eliminating duplicate record in a table [7]. These types of duplicate records are basically obtained by

joining two tables badly as well as human error at the time of data entry. An example can explain the fact explicitly.

Here Fig. 4 shows duplicate rows which only differs by the primary key. This algorithm eliminates those types of rows which contains the same value except the primary key.

EID	NAME	AGE	SEX
E101	ARUP KR BHATTACHARJEE	38	M
E102	ARINDAM MONDAL	42	M
E103	ARINDAM MONDAL	42	M
E104	SHATABDI SARKAR	35	F

Fig. 4 : Duplicate rows

4.7 Integrity Constraint Enforcement

Integrity constraints represent a method of ensuring that database updates do not result in data inconsistencies. For business applications in particular, the accuracy of the data is important. Therefore, we need strong guarantees that the information stored in the database system is not tempered with. In this paper the authors have tried to maintain data integrity.

4.8 Data Cleaning for Single Table

In case of a centralized database system, it may happen the presence of duplicate data. Here duplicate data can only occur as a result of human input error. Hence the algorithm checks the data for any row if the data of that row are same to any other row for (n-2) attributes. First of all, primary key differs for each and every row. So it checks (n-2) attributes, where 'n' is total no of attributes in that table. This method is similar to "duplicate row elimination". Except this, the algorithm performs spelling check to check if there is any wrong data and replaces that wrong data using the statistics.

4.9 Merging of Multiple Tables

In distributed system, it is important to merge tables whenever it needed. [11] In those cases, sometimes wrong data are generated by joining the tables badly. Also there is a problem of type mismatch which the algorithm tried to solve. Fig. 5.1 and Fig. 5.2 illustrate this fact.

EID	NAME	AGE	SEX
#001	RANJAN JANA	40	MALE
E002	JAYANTA DUTTA	35	MALE
E003	ALOKANANDA DEY	32	FEMALE

Fig. 5.1 : Table EMP1

EID	NAME	AGE	SEX
E101	ARUP KR BHATTACHARJEE	38	M
E102	ARINDAM MONDAL	42	M
E103	SHATABDI SARKAR	35	F

Fig. 5.2 : Table EMP2

These two tables are employee table of a company in different locations. When these two tables are joined, there will be a data mismatch in the sex field. [6] So it is needed to replace them with a single value say "male" and "female".

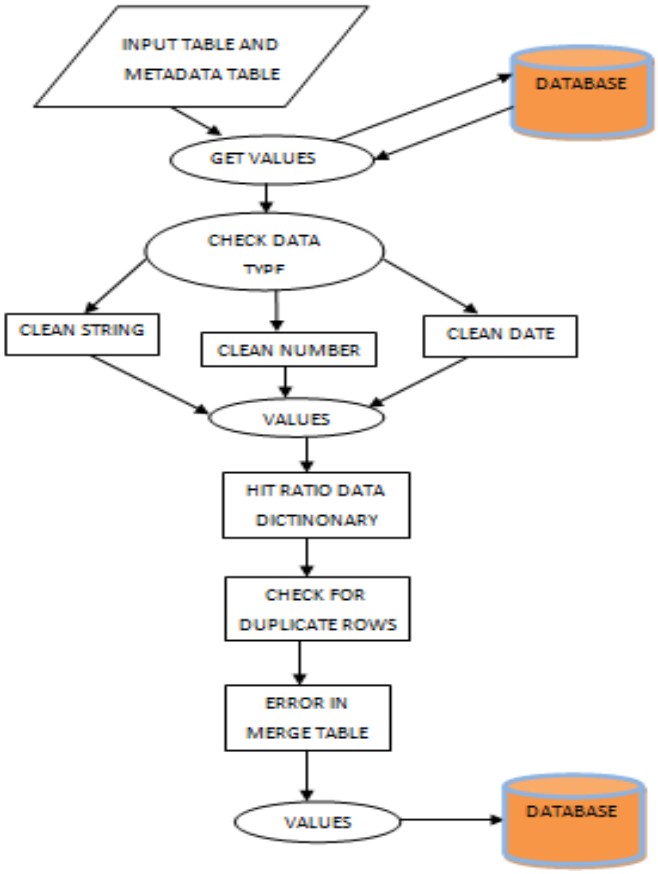

Fig. 6 : Steps of the algorithm

5. RULES AND STEPS

1. Input the name of METADATA table and read the content.
2. Check the column named ATT_NAME, ATT_TYPE and stores them into 2 different arrays.
3. Input the name of the table on which the algorithm will be applied.
4. According to the data types that it received from the METADATA table, arrange the ATT_NAME in different arrays.
5. If the data type is STRING type i.e. VARCHAR or VARCHAR2, then call the function named CleanString.
    a. Check the METADATA table if there is any specific rule for that particular field i.e. length of that field or starts with or ends with etc.
    b. Check for some special character which can be replaced with specific character.
    c. Check for the remaining special character, if exists.
    d. Load the corrected data on the database table.
6. If the data type is number type i.e. NUMBER or LONG, then call the function named CleanNumber.
    a. Check the METADATA table if there is any specific rule for that particular field i.e. length of that field or starts with or ends with etc.
    b. According to that specific rule, clean data.
    c. Load the corrected data on the database table.
7. If the data type is DATE type, then call the function CleanDate [2]
    a. Check the METADATA table for the specification of date type.
    b. There may be different types of format as follows….
        i. yyyyMMdd format e.g."20110914"
        ii. ddMMyyyy format e.g. "14092011"
        iii. dd-MM-yyyy format e.g. "14-09-2011"
        iv. dd/MM/yyyy format for example "14/09/2011"
        v. dd-MMM-yy format e.g. "14-Sep-11"
8. Apply table specific rules.

    For e.g. consider the CITIZEN table.

    Apply the method of HIT RATIO using DATA DICTIONARY and correct data and also fill those missing values.
9. Count no of attributes (n) in the table. If (n-1) values of any two rows are same, then eliminate any of the two rows.
10. Check for the attributes like “SEX” that may be represented as “MALE/FEMALE” or “M/F” that we discussed before.
11. Load data into the table.

6. STATISTICS USED

After studying different cases the statistics are made to clean data. This is an important step. Accurate statistics is needed to make the data clean process efficient. Those statistics are used to replace any character with another character.

Here is some sample statistics of data that is used to clean data.

1. Any word starts with “Y” or “Z” should be followed by a vowel.
2. If any vowel is followed by “H”, then it is incorrect. Hence drop “H” in that case.
3. Drop “W” if it is not followed by a vowel.
4. “TH” should be replaced by “O”.
5. “TCH” should be replaced by “CH”.

Table. 1 : Some of those statistics that is used to replace data

SR. NO.	CHARACTERS TO BE REPLACED	CHARACTERS REPLACED BY	OTHER POSSIBILITY
1.	@	A	
2.	5	S	S
3.	0	O	O
4.	1	I	i , l
5.	8	B	6
6.	)	P	O
7.	3	N	8
8.	7	Y	Z(if next char is vowel)
9.	!	I	1
10.	#	E	W

7. SAMPLE OUTPUT

Here is some sample output.

EID	NAME	AGE	SEX
#001	RANJAN JANA	40	MALE
E002	JAYANTA DUTTA	35	MALE
E003	ALOKANANDA DEY	32	FEMALE

ERROR

Fig. 7.1 : Before clean

EID	NAME	AGE	SEX
E001	RANJAN JANA	40	MALE
E002	JAYANTA DUTTA	35	MALE
E003	ALOKANANDA DEY	32	FEMALE

CORRECTED

Fig. 7.2 : After clean

DIST	PIN
HOWRH	711109
HOWRAH	7111092

HERE SIZE OF PINCODE 7

WRONG NAME

Fig. 8.1 : Before clean

DIST	PIN
HOWRAH	711109
HOWRAH	711109

corrected corrected

Fig. 8.2 : After clean

ID	TITLE	HOURS
1	)@rTH.A	6
2	manoMoy	8
3	MukesH	10

ERROR

Fig. 9.1 : Before clean

ID	TITLE	HOURS
1	PArTHA	6
2	manoMoy	8
3	MukesH	10

CORRECTED DATA

Fig. 9.2 : After clean

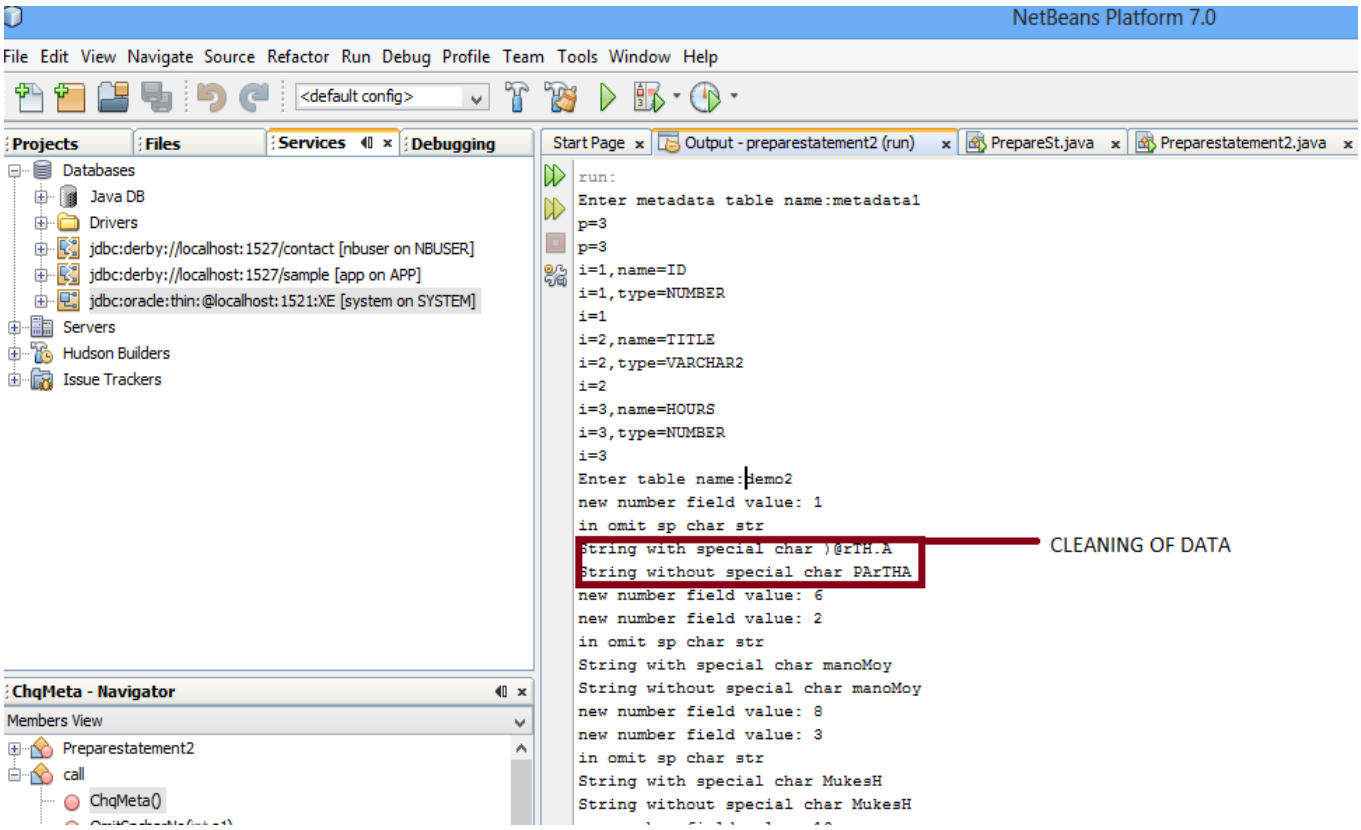

Fig. 10 : Sample output to show steps of data cleaning

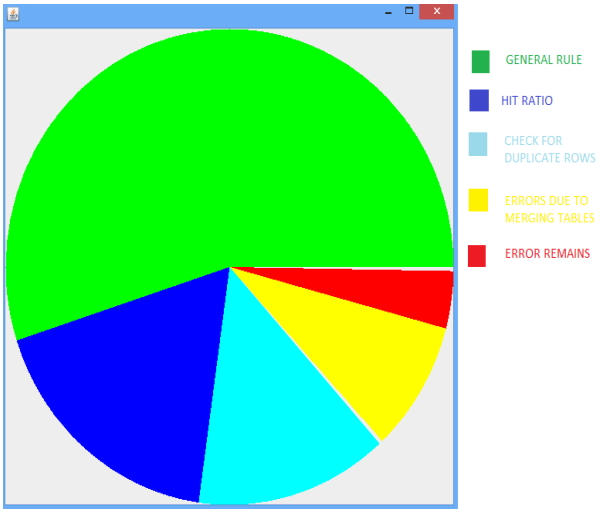

Fig. 11 : Output Chart

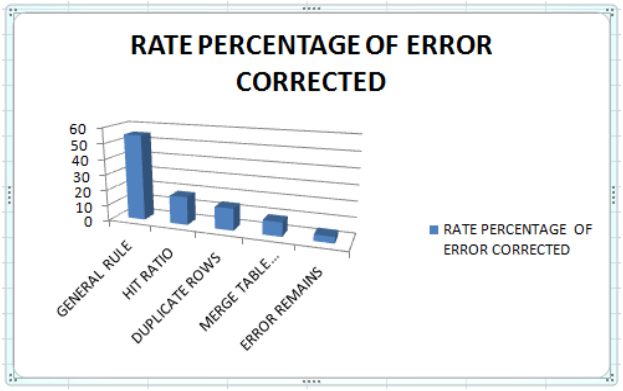

Fig. 12 : Rate Percentage Chart

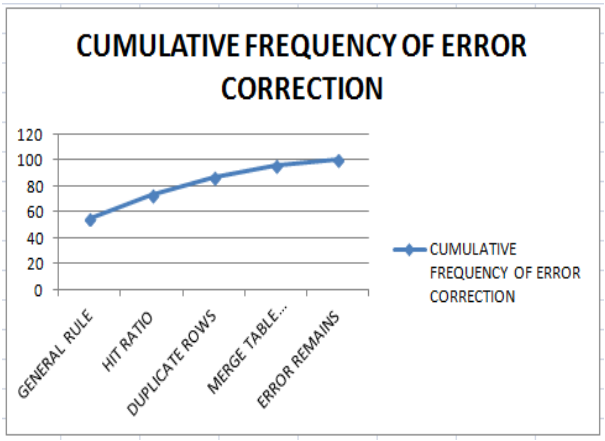

Fig. 13 : Cumulative Frequency Chart

Fig. 11, the chart is automatically generated after executing the program shows the output based on the different algorithm.

In Fig. 12, the output chart shows the percentage of data corrected through the program step by step. After applying the general rule the rate percentage of data corrected is approximately 55%. Then by applying hit ratio, 18% of data were corrected. Those steps were followed by duplicate elimination and merge table error cleaning. After all those steps were done, approximately 3% error remained in the table.

This following Fig. 13 shows the cumulative frequency of data correction through the work.

8. SIGNIFICANCE

Any database system can be used to clean different types of error that can occur in the existing database. The system will help to clean the following types of error

- Missing values
- Duplicate rows
- Out of range
- Format check
- Incorrect value
- Blank field
- Data integrity
- Data transformation

9. FUTURE WORKS AND CONCLUSION

Different issues to be considered for the future are -

- Semantic data matching algorithm can be applied to the data along with the processes discussed here.
- Authors have applied different techniques only on STRING, NUMBER and DATE data type. Data can be cleaned for other data types like raw, long raw, pictures etc.
- These techniques are tested on self- made database, so they may be tested on any enterprise database with huge data.
- Although these functions are generic in nature, but they are not fully automated. In some cases, the authors had to introduce some special function according as the database table to get more improvement in result.

Now a day's data cleaning is very important, usually the real world data is incomplete (i.e. lacking attributes values, containing aggregate data), noisy, inconsistent. Data cleaning is used to improve the quality of data. Data warehouse are used for decision making so data quality is very essential.

These paper tries is to improve the citizen database system that is used to clean data which can be applied to other databases also. The system should be able to produce results with greater accuracy with larger sets of data. A lot of methods or algorithms have been developed but data cleaning is still an active area of research.

10. REFERENCES

[1] Arup Kumar Bhattacharjee, Atanu Mallick, Arnab Dey and Sananda Bandyopadhyay, "Data Cleaning in Text File", Dept. of MCA, RCC Institute of Information Technology, India.

[2] R. Cody, "Data cleaning 101," Proceedings for the Twenty-Seventh SAS User Group International Conference. Cary, NC: SAS Institute Inc, 2000.

[3] Dr. Mortadha M. Hamad and Alaa Abdulkhar Jihad, "An Enhanced Technique to Clean Data in the Data Warehouse". Computer Science Department. University of Anbar, Ramadi, Iraq.

[4] Hasimah Hj Mohamed, Tee Leong Kheng, Chee Collin and Ong Siong Lee, "E-Clean: A Data Cleaning Framework for Patient Data". School of Computer Sciences. University Sains Malaysia Penang, Malaysia.

[5] Arindam Paul, Varuni Ganesan, Jagat Sesh Challa and Yashvardhan Sharma, "HADCLEAN: A Hybrid

Approach to Data Cleaning in Data Warehouses". Department of Computer Science & Information Systems . Birla Institute of Technology & Science, Pilani, Rajasthan, India – 333031.

[6] Erhard Rahm and Hong Hai Do. "Data Cleaning Problems and Current Approaches". University of Leipzig, Germany.

[7] Srivatsa Maddodi, Girija V. Attigeri and Dr. Karunakar A. K, "Data Deduplication Techniques and Analysis". Manipal Institute of Technology, Manipal, India.

[8] R. Kimball and J. Caserta, "The Data Warehouse ETL Toolkit". Wiley, 2004.

[9] Cleaning the Spurious Links in Data -Mong Li Lee, Wynne Hsu, and Vijay Kothari NationalUniversity of Singapore.

[10] An Important Issue in Data Mining-Data Cleaning-Qi Xiao Yang Institute of High Performance of Computing Sung Sam Yuan, LuChun School of Computing National University of Singapore, Jay Rajasekera Graduate School of International Management International University of Japan.

[11] Generic and Declarative Approaches to Data Cleaning : Some Recent Developments – Leopoldo Bertossi and Loreto Bravo.

[12] Conditional Functional Dependencies for Data Cleaning – Philip Bohannon from Yahoo! Research, Wenfei Fan from Bell Laboratories,Floris Geerts from University of Edinburgh,Xibei Jia from University of Edinburgh,Anastasios Kementsietsidis from Hasselt University/Transnational university Limburg.

[13] A Study over Problems and Approaches of Data Cleansing/Cleaning by Nidhi Chowdhury, dept. of CS,UPTU,India.

[14] NADEEF: A Commodity Data Cleansing System Michele Dallachiesa,Amr Ebaid,Ahmed Eldawy,Ahmed Elmagarmid,Ihab F. Llyas,Mourad Ouzzani,Nan Tang, OCRI,University of Trento,Purdue University,University of Minnesota.

11. AUTHOR'S PROFILE

Arup Kumar Bhattacharjee[1] is working as an Assistant Professor at RCC Institute Of Information Technology, Beliaghata, Kolkata since January 2006. He completed his MCA from Kalyani University and M.Tech from West Bengal University of Technology. His area of interest is Software Engineering, Database Management, Fuzzy and Rough set.

Partha Chatterjee[2] is currently pursuing MCA from RCC Institute Of Information Technology, Beliaghata, Kolkata under West Bengal University Of Technology. He has done his B.Sc. in Mathematics from Narasinha Dutta College, Howrah under University Of Calcutta. His areas of interest are coding, database design.

Mukesh Prasad Shaw[3] is currently pursuing MCA from RCC Institute of Information Technology, Beliaghata, Kolkata under West Bengal University Of Technology. He has done his BCA from Asansol Engineering College under West Bengal University Of Technology. His areas of interest are software engineering.

Manomoy Chakraborty[4] is currently pursuing MCA from RCC Institute Of Information Technology, Beliaghata, Kolkata under West Bengal University Of Technology. He has done his BCA from Sikim Manipal University. His areas of interest are object oriented programming.