VOLUME-6 | ISSUE-1 | JANUARY-2017 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 78.46

Original Research Paper

Data Cleansing Process with Real Time Scenario – A Case Study

Computer Science

KEYWORDS: Data cleansing, Data mining, Winpure, multi-sourced datasets.

Vidyalakshmi. V,

Computer Science Department, Kasturba Gandhi Degree and PG College, Osmania University, Hyderabad

ABSTRACT

Data Cleansing is a rapidly growing process which provides in removing the impurities like corrupt and inaccurate records from a record set, table or a database. Unstructured data does not follow a specified format. Around 80-90% of all potentially usable business information may originate in unstructured form.Data Cleansing provides a wide range of methods to ensure the accuracy and usability of the data. Data Cleansing is needed where data accumulate in a database through various sources from various systems in a different format. Given a rapid growth in a competitive environment there is a need for more precise and accurate data for decision making. This research paper provides detailed conceptual understanding on data cleansing processes along with a real time scenario using a literature survey.

Introduction:

The data has grown enormously from the size of kilobytes to terabytes and further to yottabytes (1000^8) (Lingma Acheson) with Global information systems becoming more accessible through World-Wide-Web. This increase in data is mainly due to data collection and data availability through various automated tools, database systems and web. The huge data is rich but is very poor for information and making analytical conclusions based on various study / parameters. Data mining is the process of automated analysis based on massive data sets and this is commonly also called as knowledge discovery from Data (KDD). The data mining process is to extract interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. For data mining to be performed on large amount of data the data needs to be cleansed and this cleansing of data is referred to Data cleansing. [2]

Need for Study

Organizations tend to get data from various sources. The data that an organization receives is typically of huge size (like BLOB's) and are of several gigabytes and more. The valuable data that an organization considers is heterogeneous data which is mainly used for decision making. But when the sources of data are multiple and huge, there are probabilities that the data may be impure which may lead to wrong conclusions in data mining process. Data Cleansing is a valuable process used by the organizations to improve the mining process which leads to correct results for decision making which in turn save time and increase efficiency.

Research Objective

The main objective of the research paper is:

- To have an understanding on data cleansing and its process
- To provide an overview of using one of the data cleaning tools for man-power consultancy (job counseling) scenario.

Data Cleansing:

Data is deemed unclean for many different reasons in any data warehouse or large data set. In the real world scenario data warehouse is loaded with huge amount of data from heterogeneous sources which requires constant cleaning. The primary reason for this unclean data is due to multiple sources of data and manual entry of data into the system. In a large system with such multiple integrated sources, the data fed into the system is unclean, inconsistent and redundant which requires the need for cleaning. [4] Data cleansing is a process of removing dirt or inconsistencies from the table or a database in order to improve the quality of data in the data warehouse in turn improving the data mining process. In order to have a quality data, the data should be accurate, complete, consistent, uniform, unique and valid. If data has an ever increasing role in data warehouse the anomalies and the impurities cause's inconveniences and averts the effective utilization of the data which in turn causes a degradation of performance in retrieving data from the database.

Data cleansing is much more than simply updating a record with good data. Serious data cleansing involves decomposing and reassembling the data. According to (Kimball, 1996) one can break down the cleansing into six steps: elementizing, standardizing, verifying, matching, house holding, and documenting. [5]

The data cleansing yields the following characteristic of data:

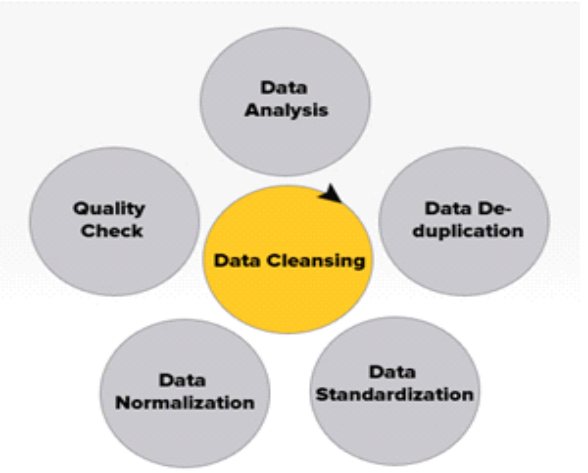

Figure 1: Characteristics of Data as part of cleansing processes

At the start of the process data rules are defined which is revised based on various scenarios. In the data warehouse, the data is analyzed for errors or unwanted characters. This is repeated at every stage of the data processing. The data is refined by removing the duplicates, removing the dirt or erroneous data and standardization of the data. The standardized data is further normalized for the appropriate KDD or data base and goes through audit for quality check of the data. [1]

Data cleansing process can be broadly classified into the following three phases:

- To define and determine the type of errors to be eliminated from the big data sets
- Search and identify the error instances
- Correct the identified error instances

Each of these phases constitutes a complex problem in itself, and a wide variety of specialized methods and technologies can be applied to each. While data integrity analysis can uncover a number of possible errors in a data set, it does not address more complex errors. Errors involving relationships between one or more fields are often very difficult to uncover. These types of errors require deeper inspection and analysis. [6]

VOLUME-6 | ISSUE-1 | JANUARY-2017 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 78.46

While the above phases are defined for data cleansing process, the overall process can be seen as a simple 5 steps process as illustrated in the image below:

Figure 2: Cleansing process

1. Planning: As a starting step, from the overall large data set that needs to be cleansed, a small set of data having maximum priority is taken first for cleaning process. In this step, the cleansing rules and type of errors / redundant data / dirt data which needs to be removed are defined. For example, making sure your postal codes and state codes agree, making sure the addresses are all standardized the same way, etc.

2. Analyze to cleanse: After you have an idea of the priority data your company desires, it's important to go through the data you already have in order to see what is missing, what can be thrown out, and what, if any, are gaps between them. This is the step in which the resources who would actually do the cleansing through the defined rules are identified. The amount of manual intervention is directly correlated to the amount of acceptable levels of data quality you have.

3. Implement Automation: Once you've begun to cleanse, you should begin to standardize and cleanse the flow of new data as it enters the system by creating scripts or workflows. These can be run in real-time or in batch (daily, weekly, monthly) depending on how much data you're working with. These routines can be applied to new data, or to previously keyed-in data.

4. Append missing data: This is a very crucial step to correct those records which cannot be corrected automatically. Examples for such cases would be email address, phone numbers etc. It is very important to identify the correct way of getting a hold of the missing data, whether it's from 3rd party append sites, reaching out to the contacts or just via good old-fashioned Google.

5. Monitor: After cleansing the data, constant review of dataset needs to be performed periodically to monitor issues before the issue gaining priority. You should also be aware of bounce rates, and keep track of bounced emails as well as response rates.

The end of this cycle, or step six if you will, is to bring the whole process full circle. Revisit your plans from the first step and reevaluate. Also, need to revisit if the priority data and data rules defined in step 1 is still valid. The complete process of 5 steps needs to be performed frequently to ensure cleansed data in the warehouse. [3]

Materials for Study and Methodology:

For the data cleansing activity to be performed the raw data is maintained in an excel file. The excel file has few columns and data related to those columns in rows. A study was conducted in a reputed job oriented consulting organization to understand the data cleansing process followed. The organization receives the raw data from many companies regarding the different kind of opening that they have and the criteria associated with those job opening. The methodology used for this paper is using a data cleansing tool called "Winpure clean and match". The tool has varied features to clean the redundant / duplicate data and also to update the data as per business needs.

Analysis and Findings:

A snippet of the raw data maintained in an excel sheet is shown below:

Company	Designation	Job Description	Location	Required Experience	offered CTC Range
TCS	SSE	Senior Developer for Java	Hyderabad, Chennai	2-4yrs	6.0 to 6.5
TCS	Test Lead	Testing lead for all web based app	Gurgaon	7-10yrs	14.0 to 15.2
TCS	Tech Arch	Architect for Microsoft applications	Chennai	10-12 yrs	16.0 to 19.0
TCS	Lead consultant	Lead for SAP various modules	Hyderabad, Pune	10-12yrs	18-20.0
TCS	Project Manager	Manage large scale projects for AMS, PMP added advantage	Blr	10-12 yrs	19-21.0
Infosys	Technology Analyst	Developer with Java and spring frame work	Chandigarh, Mangalore	4-6yrs	6.0-6.5
Infosys	Test Engineer	To test web based applications through automation	All locations	2-4yrs	5.0 to 6.5
Infosys	Sr Project Manager	To manage large scale projects, PMP preferred	Mangalore, Trivandrum	13-16yrs	20 to 22.5
Infosys	Lead consultant	Lead for SAP various modules	Hyderabad, Pune	10-12yrs	18-20.0
Infosys	Project Manager	Manage large scale projects for AMS, PMP added advantage	Blr	10-12 yrs	19-21.0
Wipro	Associate consultant	Design and build Java based applications with Struts frame work	Pune	4-6yrs	7.0 to 8.5
Wipro	Test lead	To test Enterprise applications, knowledge on SAP / Oracle	Chennai	2-4yrs	9.0 to 12.0
Wipro	Associate Manager	To manage a teams of 10-15 people, good understanding of the technology like Java / .Net	Cochin	8-10 yrs	10.0 to 12.0

Wipro	Project Leader	Lead for SAP various modules. Having good exp working on SD, CRM and MM	Hyderabad	10-14yrs	9.0 to 13.5
Wipro	Program Manager	Manage large scale projects for AMS. Having exp of handlilng multiple projects is a must	Chennai	13-16yrs	21.0 to 23.0
CTS	SSE	Senior Developer for Java	Hyderabad, Chennai	2-4yrs	6.0-6.5
CTS	Test Engineer	To test web based applications through automation	All locations	2-4yrs	5.0 to 6.5
CTS	Sr Lead	To manage a teams of 10-15 people, good understanding of the technology like Java / .Net	Cochin	8-10 yrs	10.0 to 12.0
CTS	Principal consultant	Manage large scale projects for AMS, PMP added advantage	Blr	10-12 yrs	19-21.0
CTS	Tech Arch	Architect for Oracle applications	Chennai	10-12 yrs	16.0 to 19.0

Table 1: Raw dataset extracted from job consultancy firm

There are many situations in which data are collected from different ways for example data are collected in random in a shopping mall, theater complex etc. The data that are collected could be for various categories like, educational counseling, shopping interests, career counseling and many more. The tool provides a very user friendly wizard consisting of 6 steps to clean, match and update the data in the excel sheet. The updated data is retrieved and used for all consulting purposes, here sending the job notifications to suitable candidates.

The process of data cleansing is explained below with screenshots:
1. As per the first step, we need to create / open an existing project in which the cleansing needs to be performed.

Figure 3: Project creation screen

In the second step, the type of the file is selected and the raw data file is imported to the tool.

Figure 4: File upload / selection screen

The supported types of files in this case are text/csv/excel/access/dbase/outlook/sql server/MySQL.

3. In step 3, the tool provides varied features as listed below:

- Helps remove rows with empty or specified values
- Replace values having empty or specified value to a new value as desired by the business.
- Helps in removing unwanted space in the data.
- Helps in removing unwanted characters like .,/:- etc.
- Helps to convert data to either UPPER / lower / Proper case

These options would be available for every individual column of the imported excel.

The user needs to choose the appropriate action from the listed above options and cleanse the data.

Figure 5: Screen to depict the options as part of cleaning process

Step 3 is mainly to clean the data as per the business need.
4. This is the step to identify duplicate records and the tool gives us varied options like to delete / to merge / to upload to master record.

VOLUME-6 | ISSUE-1 | JANUARY-2017 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 78.46

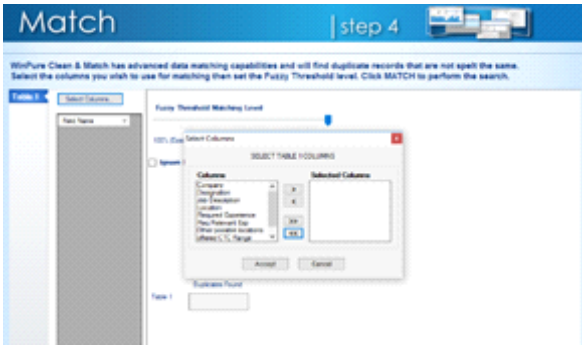

Figure 6: Screen showing the different options as part of match process

Figure 7: Screen showing final results after identifying the duplicates

5. In the above step, the business would need to decide and take appropriate action on the duplicate records and perform the action.
6. On completion of step 5, the project is saved and the table data is exported back to excel.

As per the final results, the updated excel record would like as shown below:

Company	Designation	Job Description	Location	Require	offered CTC Range
TCS	SSE	Senior Developer for Java	Hyderabad, Chennai	2-4yrs	6.0-6.5
TCS	Test Lead	Testing lead for all web based app	Gurgoan	7-10yrs	14.0 to 15.2
TCS	Tech Arch	Architect for Microsoft applications	Chennai	10-12 yrs	16.0 to 19.0
TCS	Lead consultant	Lead for SAP various modules	Hyderabad, Pune	10-12yrs	18-20.0
TCS	Project Manager	Manage large scale projects for AMS, PMP added advantage	Blr	10-12 yrs	19-21.0
Infosys	Technology Analyst	Developer with Java and spring frame work exp	Chandigarh, Mangalore	4-6yrs	6.0-6.5
Infosys	Test Engineer	To test web based applications through automation	All locations	2-4yrs	5.0 to 6.5
Infosys	Sr Project Manager	To manage large scale projects, PMP preferred	Mangalore	13-16yrs	20 to 22.5
Infosys	Lead consultant	Lead for SAP various modules	Hyderabad, Pune	10-12yrs	18-20.0
Infosys	Project Manager	Manage large scale projects for AMS, PMP added advantage	Blr	10-12 yrs	19-21.0
Wipro	Associate consultant	Design and build Java based applications with Struts frame work	Pune	4-6yrs	7.0 to 8.5
Wipro	Test lead	To test Enterprise applications, knowledge on SAP / Oracle	Chennai	2-4yrs	9.0 to 12.0
Wipro	Associate Manager	To manage a teams of 10-15 people, good understanding of the technology like Java / .Net	Cochin	8-10 yrs	10.0 to 12.0
Wipro	Project Leader	Lead for SAP various modules. Having good exp working on SD, CRM and MM	Hyderabad	10-14yrs	9.0 to 13.5
Wipro	Program Manager	Manage large scale projects for AMS. Having exp of handlilng multiple projects is a must	Chennai	13-16yrs	21.0 to 23.0
CTS	SSE	Senior Developer for Java	Hyderabad, Chennai	2-4yrs	6.0-6.5
CTS	Test Engineer	To test web based applications through automation	All locations	2-4yrs	5.0 to 6.5
CTS	Sr Lead	To manage a teams of 10-15 people, good understanding of the technology like Java / .Net	Cochin	8-10 yrs	10.0 to 12.0
CTS	Principal consultant	Manage large scale projects for AMS, PMP added advantage	Blr	10-12 yrs	19-21.0
CTS	Tech Arch	Architect for Oracle applications	Chennai	10-12 yrs	16.0 to 19.0

Table 2: Final results dataset after the cleansing processes

The above steps can be performed to any file having large number of records and the data is cleansed within minutes saving enormous time to the users.

Conclusion:

In the observance the author has provided an understanding about the data cleansing process which is defined as a sequence of operations proposing to enhance the overall data quality. Furthermore an overview of using one of the data cleansing tools is

provided. Many data cleansing approaches mostly focus on the transformation of data and the elimination of duplicates. But due to the enormous inflow of data there remain a lot of open problems and challenges. So far a little research has appeared, although a large number of tools are used in the cleansing process. Data cleaning is not only for data mining but also for heterogeneous data sets. This paper is further explorative.

References

1. http://www.artechsoft.com/material_management.html/
2. http://cs.iupui.edu/~linglu/class/481/chap1.ppt-slide 3, 5
3. http://www.salesify.com/keeping-it-clean-the-five-step-data-cleansing-process/
4. http://www.springer.com/cda/content/document/cda_downloaddocument/9780387098227-c1.pdf - page 3
5. http://www.springer.com/cda/content/document/cda_downloaddocument/9780387098227-c1.pdf - page 4
6. http://www.springer.com/cda/content/document/cda_downloaddocument/9780387098227-c1.pdf - page 6
7. http://www.winpure.com/cleanmatch.html