IMPORTANCE OF DATA CLEANING TOOLS/SOFTWARES

Data cleaning is the most valuable process to the organization as it saves both time as well as increase their efficiency. There are various kinds of data cleaning tools present in the market which are been used by various companies for the purpose of removing duplicate data, fix the incorrect and incomplete files from the databases. These data cleaning tools save money by omitting bad records and deals with inaccuracies within seconds saving a lot of time and effort of the organization. According to studies organizations end up wasting upto 6 million dollars only because of no proper data cleaning tools.[1]

Data cleaning is very important to an organization which deals with data irrespective of the size of the organization. These organizations may include banks or other government organizations, small to medium enterprises dealing with data, organizations in the field of marketing. In fact every organization dealing with data needs to have a data cleaning software. These tools should be used in regular basis as duplicate and incorrect data can grow quickly leading to decrease in business efficiency.[2]

The data cleaning software help to correct incorrect data or duplicate data such as the emails, phone numbers of the customers resulting in contacting them. One of the most preferred example in this scenario is due to duplicate data the marketing organizations may send multiple marketing advertisements to the same customer due to duplicate data resulting in loosing the customer. These type of errors can be controlled by the organization within no time and save valuable resources of the organization as it would take much longer time to delete the records manually. [3]

[1] <https://www.winpure.com/blog/how-businesses-can-benefit-from-data-cleansing-software/>

[2] <https://www.winpure.com/blog/importance-of-data-cleansing/>

[3] <http://victoragina.com/data-cleansing-tools/>

COMPARISION:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| S.NO | TOOL NAME | RELESE DATE | LATESRT VERSION | LICENSE | OPERATING SYSTEMS | WEBSITE |
| 1 | OPEN REFINE | NOVEMBER 10,2010 | 2.7/  JUNE18,2017 | BSD | MICROSOFT WINDOWS, LINUX, MACOS | <http://openrefine.org/> |
| 2 | RAPID MINER | 2006 | 8.1/  FEBRUARY 6,2018 | AGPL | CROSS-PLATFORMS | <http://rapidminer.com/> |
| 3 | TRIFACTA | 2012 |  |  | WINDOWS 7 OR LATER,  OSX 10.10 OR LATER | <http://www.trifacta.com/> |

|  |  |  |  |
| --- | --- | --- | --- |
| TOOLS | OPEN REFINE | RAPID MINER | TRIFACTA |
| PROBLEMS |
| AVAILABILITY | DESKTOP | DESKTOP | DESKTOP |
| MISSING VALUES | YES | YES | YES |
| DUPLICATION | YES | YES | YES |
| ILLEGAL VALUES ELIMINATION |  | NO |  |
| VARYING VALUE REPRESENTATION |  | NO |  |
| MISSPELLING | YES | NO | YES |
| MERGE | YES | YES | YES |
| FILE FORMAT | TSV,CSV,XML,RDF TRIPLETS,JSON,GOOGLE SPREADSHEETS,GOOGLE FUSION TABLES | CSV,DATABASE,EXCEL,  ACCESS,BINARY,XML | CSV,JSON,TXT,XLS,XLSX,  TABLEAU DAT EXTRACT |
| EASE OF USE | MODERATE | MODERATE | HIGH |

RESULTS:

From the above comparisions and use of datasets into the following tools we can clearly state that open refine and trifecta are the best among the following tools. Open refine is an open source tool which is free to use by the individuals as well as the organizations where as trifecta is not an open source but is capable of dealing with data cleaning, preparation in the best possible way with various operations involved and high processing speed. Rapid miner is limited with its operations such as duplication, missing values, merges and still can be used for data cleaning for these purposes.

CONCLUSION:

Data cleaning is a very important aspect in the organizations which deals with data. From the above study we can know about the data conditioning, cleaning and readiness. Data quality problems, data cleaning processes and the steps involved in them and even the meth. This paper also presents the comparision of certain data cleaning tools and suggest the best possible tool within them with all the specific technical features and operations of the tools discussed.