Requirement to cleanse DATA in ETL process and Why is data cleansing in Business Application?

Sweety Patel
Department of Computer Science, Fairleigh Dickinson University, NJ- 07666, USA

ABSTRACT –

How data is passed into data warehouse is very easy steps like: data transferred from one or more data bases as data source (database) to staging area and data are filtered through cleansing process before the destination as data warehouse. Data cleansing is most important part of the integration of the heterogeneous data sources. In ETL process data cleansing is most important phase of the extraction, transformation and loading cycle. We discuss different issues which required data cleansing process.

Keywords **- Data Cleansing, Data Scrubbing, Data Cleaning**

I. INTRODUCTION

We can also know data cleaning as data cleaning as well as scrubbing. Where we required a data clean?

1. Because of human error
2. Error as of the application limitation
3. Short terms of acceptance for the data input
4. Data is not dirty but not have proper technical data format as per the application rules requirements
5. Data duplication in data source as of the integration of the heterogeneous system
6. Data are entered properly but not updated as per the required time schedule lead it to the become dirty one
7. Lack of validation on the data as the time of data entering in the application by the software rules
8. Limited verification is done when data is passed through multiple filter for integration and make it one application body data from the heterogeneous system
9. Logical data mapping rules are not correct
10. Lack of training to the person who is entering a data into database as per the body of regulation to complete the input requirement

II. ETL PROCESS

ETL stands for Extraction Transformation and Loading. When operational data come to data stage and then go to Data Warehouse, in between that ETL process is done shown in Fig. 1. Finally Data goes to the data decision Maker for final analysis of the data .

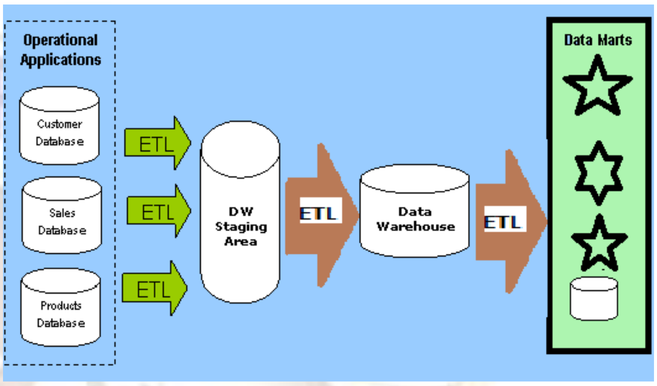

Figure-1 ETL Process

III. DATA ARE NOT CLEANED?

2.1 Data storage is for the future decision related to the data and that base on this all business decision model is made. As of any reasoning entered correct data is not usable means wastage of all energies including man power, resource allocation, time of business, timeline of workers, user of the system as applicant. Inaccurate data in terms of the application rules and regulation make a report generation failure as well also give bad decision to the next step in logical process and make a another generation of data dirty as they are produced from the originally dirty data as parents .Inaccurate data cause a false result and lead to make an analyst wrong criteria for future enhancement of the business strategy.

2.2 All application runs with data and need data to live like a live human being in their working field. Application covers all part of life surviving things. Inclusion of that is not limited to integration for the required heterogeneous data for banking operation but spread to the all money transaction filed where money is become base of the business. Secondly, data which required a time to time updating in its value make it as a dirty data without changing as per the time change as money conversion rate is not affected properly and time by time as per the change in real environment then it is an impossible guess to determine a loss in terms of tons of money in the treading market as the loss.

2.3 Calculation of one data from the parent one require a perfect conversion and also a correct input as its value with followed all required format, restricted by the application developers. One small mistake (10 to 01) in a single digit in birth date may a person’s pension late or make a person’s car insurance down to the earth in costing for getting benefit towards its loss during any accident or in damage in any incident.

2.4 In false process of mapping one to another data by accidently also make a lot of big difference in the data result as an output. As in the heterogeneous data have many fields and required too much intelligence with consciousness for mapping fields for integration.

2.5 Conversion of one filed from the another make if any shrinking of data it also required to make conscious point to be noted for the bigger changes in the larger calculations. As conversion from the .335$ to .33$ make a lot of difference in the money transfer rate as we go for unexpected result for the very big equitation dealing with money.

2.6 Duplication of the data because of the lack of the mapping of the fields to be integrated in database make a database a lot of inconsistent way to retrieve back data. Failures will produced in each of the result make a lot of difference in actual output as a result while integrating data.

2.7 Typo in the form of the input may block result while filtering and capturing a data into the analyzed report may need lot of correction to make it to it proper and lead to reach as a perfect outcome required a lot of efforts on the scale of smaller mistakes which a normal mistake done by applicant user as a data insertion process.

2.8 Validation is most important part of the any software application development procedure. Missing any criteria into the validation on single value of field may lead in dirty data insertion in to the database and lead to generate a non expected outcome in the report generation and also many other process in integration of data like: while mapping of fields many data are filtered out which are also required to produce a 100% finial analytical report decision criteria for business growth our make in a one step ahead into the profit line as comparing with previous one.

2.9 Missing value of one field causes as in some records which are negligible with compared in billions and in trillion data may required a lot of hard work to find out a particular fault in the particular data base and especially in some or in unexpectedly only in single record make it a secured and correct as a outcome of the integration phase which lies on the bases of cleansing a data into database.

2.10 Sometime data case sensation retrieval methodology may produce a result out of expectation as this mistake is done as of human error in building a query while writing a procedure to produce a desire result in different place of query writing as not proper field.

2.11 Incompleteness as in value in data field is also became a result in filtering as a dirty data once if it is not getting matched with compared with formatted full text. A’bad and Ahmadabad make a lot of difference in the data filtering process and make it as a dirty data once it is not filtered and getting as an output.

2.12 Mentioned all reason is make a lot of difference in integration of data in heterogeneous system and lots of efforts make it possible to put a 100% result in an output. Cost concern with human efforts, moneywise, timeline to complete a task as a report completion and also lot other in terms. Minute mistake make a data filling process as an input to the any application software make a lot of efforts to correct it as a result wise proven records.

2.13 Business stand always with the analysis of the past data. Analytical model makes a business strategy perfect to go with the profit direction or say in the direction of growth of the business. History make future in data storage life as without stand point of the previous or relevant data make a decision power looser for the efforts to be done.

IV. CAN BUSINESS STAND WITH DIRTY DATA?

Business runs with data transaction cycles. And cycle include from beginning to current and up to future projection data. So any beginning become previous stand point of the current one and current become future stand point for data analytical projection. Error in any of this phase lead data to be dirty and make a result insufficient as a efficient proof for the building a false report generation and getting bad result with dirty data. Any Business can’t stand with this dirty data for the progress projection in business line. Efforts to make application software data cleanse are not lesser equivalent as to make software and develop it as successful applicable to all kind of user input followed by application software rules and regulation.

V. CONCLUSION

Data Cleansing make a dirty data to be clean for further work on the data as in report generation or also in the data maintenance for creating further data to be created which can be used as an intermediate stage of final data production. Bearing with dirty data must lead organization to the wrong direction regard than in progress. So making out each and every effort in data cleansing is a main fundamental part of the ETL process and that make a ETL system more effective and efficient. Many tools also available for data cleansing but still more human effort is required to make it is a simpler and easier for the developer to work on this fundamental problem in ETL process.

References

Books

[1] Nong Ye, *The Handbook of Data Mining* (Lawrence Erlbaum Associates, Mahwah, NJ. Publication, 2003).

[2] Jiawei Han and Micheline Kamber, *Data Mining:Concepts and Techniques* (Morgan Kaufmann Publishers, University of Illinois at Urbana-Champaign).

[3] Bharat Bhushan Agarwal and Sumit Prakash Taval, *Data Mining and Data Warehousing* (Laxmi Publications, New Delhi - 110002, India).