Summary of Data Integration and ETL: A Theoretical Perspective

Mike Pendleton

West Texas A@M University

The authors began discussing the topic in explaining what data integration means. This is important to cover before getting into the ETL process, the first step in utilizing a primary data set is to have the data set complete. This provides a confidence level that the data will be utilized, which best suits the business or organization. Once data sets are determined for use, then the ETL process would begin.

Data integration is the use of a subset of areas. This includes Business Intelligence, data migration, data warehouse, enterprise application with information integration, and master data management. There can be multiple data sources, once utilized in order to compile for the use of the above areas. Data sources could be raw data, databases (suggested relational or non-relational), word, excel and others. The only argument I would pose to the databases on either is non-relational databases contain an extreme hardship on data integrity. This is because only relational databases house integrity within its workflow. The use of primary, foreign and composite keys helps to retain integrity as an example.

Regarding data integrity, this is a majority of the disadvantages listed for Data Integration. The component of duplicate data would exist in both raw data and non-relational databases. Though there are less disadvantages than advantages, the disadvantages could grossly outweigh the advantages if not taken into consideration. The other example of a disadvantage would be domain data. This can be very cumbersome if integrating from two different organizations, as there would be a level of trust to take into account if the companies were not a parent and child relationship.

The advantages of Data Integration would include automation, deliver data quickly to the end user, less resources needed, and being extremely efficient. The end user could be either the customer or partner. Several applications help with the process, in order to leverage data in order to provide the data to the data warehouse. This aligns with the efficiency and low workload in order to proceed to the ETL process.

ETL is the process of extraction, transformation and loading the desired data into the main repository, which is the data warehouse. Extraction can be full or partial, dependent on the data desired. The transformation process would be where the data is cleaned. For instance, raw and non-relational databases would have to purge the non-necessary and/or duplicate data within those sources. The last phase of the ETL process is loading the now cleaned and approved data and to be stored into the Data Warehouse.

The purpose of ETL is to combine either single or multiple sources of data sets and process them to an acceptable final storage. Once complete this is only stored for historical purposes and cannot be modified. With modification of the data sources is how the Data Warehouse would be changed. The effectiveness of ETL helps organizations determine their business needs. It would be important to list a component that I think fell outside of the scope of the paper, would be that business needs change. As these changes occur, the sourced data can change in which would in turn update in the Data Warehouse. The importance of calling this reference is business needs are a constant cycle of components in order to ensure when they were incorporated, that adjustments either may or may not need to be done. However if business needs changed which can happen due to growth, expansion and other factors, then revisiting the source data and the ETL process for the changes need to occur would have to be revisited. This would ensure that the ETL process updates with the new needs for how the data is extracted, transformed and loaded into the Data Warehouse.

References:

(2021 7th International Conference on Advanced Computing & Communication Systems (ICACCS), 2021)