June 3, 2018

Course: CIS570 – Business Intelligence
Name: Robert Palumbo
Assignment: Reading Discussions – Week2-Session2
Due Date: Sunday, June 3 @ 11:59pm

Question 1. Why is ETL necessary to populate a data mart / data warehouse.

As we know, data marts and warehouses are used by organization to store myriad types of data which is then used to perform simple to very complex analytics from end-users. The goal of these analytics is to provide information and valuable insights into this vast array of data to allow the organization (or end-users) to make better business decisions – be it day-to-day management or more strategic visionary decisions.

The ETL (Extract, Transform, and Load) has a very important an integral role in the creation and management of these data repositories. Fundamentally, if the data that is used in the analytics process is not valid, then the resulting business decisions will be erroneous as well which can be catastrophic to an organization. Thus, the goal of ETL is to populate the data repositories with clean and valid data.

Extraction is the process of collecting data from a data source and making it available for processing. Transformation is the process of converting the data from its existing form into a format that is suitable for the database into which is will be loaded. Loading is the process of physically loading the data into the target database. I would argue that the transform phase is the critical component in ETL as this is the step that likely requires the most work to successfully perform.

ETL therefore is used in part to ensure the quality, integrity, and validity of the data which is used to make these decisions. However, because the data often comes from many different sources of which the type and format of specific data elements can be different, it is very important to impose a consistent format for each of these types.

For example, data may come from different sources in which currencies are not the same (e.g. US Dollar vs Yen). As a result, a consistent definition of a *currency* must be used to standardize this type of data to yield meaningful results. In a similar manner the format and storage of date-time data must be made consistent. Perhaps conversion of all date-time values to UTC would be appropriate. Similarly, ETL should account for missing and duplicated data.

Further, ETL should also attempt to aggregate, sort, trim, and other similar operations the data as much as possible to prior to loading which expedites the process and allows for more efficient use of that data after persistence in the repository. Ideally, the data should be persisted into the target database if in the most optimized form as possible.

There are certainly many other considerations similar to these that must be accounted for but the idea is to allow the data repositories to store a consistent representation of all the data types and formats.

Question 2. Is a logical (virtual) DW a viable alternative to the traditional DW? Explain.

I do believe that a logical DW is in fact a viable alternative over a traditional DW and in fact become even more viable as our technology and global networking infrastructure continues to improve. Likewise, if costs of maintaining an internal DW are of concern or an organization simply does not have the means to support their own DW then a LDW is certainly a promising option.

One of the issues with an LDW is that there is a performance hit on queries as the data must be collected or joined from remote data repositories versus having the data readily available within a local DW. However, the huge strides that have been made in networking bandwidth and data transmission speeds have, in my opinion, leveled the playing field between these two technologies.

Further, if having current data to act upon is of primary concern to a business, then an LDW is by far superior as the this is what an LDW brings to the table – always current data – since it is gathered at the time of the query. Decisions can be made knowing they are derived from using the most current data available.

Cost is always a consideration as well. An LDW is typically a less expensive approach as there are minimal costs associated with on-site hardware requirements to access the LDW. Likewise, cost is again reduced as it requires less time and resources to standup and LDW environment within and organization versus a traditional DW. For organizations operating on a tight budget an LDW is again a more than viable solution.

Additionally, analytics performed in the *cloud* can exploit the use of MPP or *massively parallel processing* to perform complex analytics even achieving results in times which are faster than those reached by traditional DW systems.

The big downside of an LDW is that if connectivity to the source systems is lost or a source system is down there is typically no backup option for performing queries. The query data resides on those systems, unless replication has been implemented which adds more complexity to the picture. While this is certainly of concern, to a large extent this issue can be mitigated thanks again to the technological achievements of today and tomorrow.

http://panoply.io