softtech ®
group

ETL
Challenges

Data
Sheet

ETL Challenges

ETL is an automated method that refers to three distinct functions merged into a single tool, including the collection of raw data, the extraction of information needed for analysis, the transformation into a format that fits business needs and the loading of it to the data warehouse. It usually reduces the size of the data and increases efficiency for the purpose of analysis and database management.

The data sources must be incorporated and carefully designed and checked when constructing an ETL infrastructure in order to ensure the proper transformation of the source data.

ETL Project Complexities

It seems exciting to have an ETL opportunity or a project, but it comes with its own difficulties and complexities.

➢ The initial data set and its volume can alter over time. The amount of data will increase and thus its analysis becomes complicated. Maintenance costs rise, and extra effort and time can be applied to the other relative characteristics of changing the data format.

➢ Data from multiple sources does not go together smoothly and requires a lot of effort to work together in some instances. Such attempts can be time consuming.

➢ It can be tricky to develop Architecture for an ETL project. It may cause significant problems for your team performing an ETL job to go straight to coding without putting into consideration the overall larger picture.

➢ There are many resources available for this if there is a need for real time alerts. The first thing for an ETL project to be a success is to get the basics right in place. Instruments are secondary.

➢ There are some problems that can be correlated with any information that affects the ETL process. When brought in from various sources, the method needs to be able to accommodate a number of data formats. The device needs to be fault tolerant and has the potential to recover gracefully, as issues can eventually arise, meaning that when it experiences problems, the data can be moved from one end of the pipeline to the other end.

➢ Data consistency, forms of data independence, how dynamic the data relationships are, etc. This will all affect the ETL process. Accessing data from various systems can be difficult and the accuracy of the data is not guaranteed. The data may also be inconsistent.

➢ Source and target data consistency and scalability of the ETL process are common technical challenges. Scalability can be a challenging problem that you will face because it depends on the scale of the data that you are working with. Operational changes in source data structures, and ongoing revisions to the meanings and scope of the target schema, can occur. All of that adds uncertainty to the method of ETL. Due to all these dependencies, ensuring data quality is challenging in any ETL process. Duplicate paperwork is another problem. Conflicting and duplicate records are common data issues that can have a major impact on the efficacy of an ETL process and the quality of production data.

➢ It is a challenge to incorporate data from multiple sources. This includes the programming of scripts to parse the source data. If standard drivers are not available, to complete the desired feature, coding

➢ Data accuracy is a problem in the state of transformation, as details obtained from various sources may be unpredictable.

➢ The challenge is the time taken to populate a data warehouse at the time of loading or populating the data.

Best Practices to overcome ETL challenges

Understanding the overall project and expectations and detailed analysis of the requirements will help design a robust and flexible ETL architecture accordingly. When the data tables are large then loading data incrementally is the solution as it makes sure that the last updates are brought into the ETL process. Updates should not get missed as else there will be redoing of tasks. Loading incrementally improves the overall efficiency of an ETL process. To overcome the challenges, it is best advised to focus on the bottlenecks, load data incrementally, break down large tables into small ones and use parallel processing.