Abstract

Timely and accurate decision-making provides a competitive advantage for each organization that needs to store and speedy access to the large volume of everyday transactional data. In the Data Warehouse (DW) environment, Extract, Transform, and Loading (ETL) plays a key technology that refines and integrates a large stream of heterogeneous operational and external data of any organization. The value of organizational data is significantly enhanced when content migration from various sources is done significantly by using the ETL process.

In the last few years, the use of ETL for constructing and managing Data warehouses has been gaining popularity in various real-life applications like e-commerce, banking, e-governance, etc. Moreover, many industry applications (Fraud detection, payment processing, IoT edge analytics, etc.) require real-time integration and reporting over data acquired from heterogeneous data sources. Many types of ETL solutions are coming to resolve these issues, like Batch versus Real-time and On-Premise versus Cloud.

ETL is a significant area of research for a well-established Data Warehouse environment. In this Thesis, I have discussed the main motivation behind the Ph.D. research work along with a brief literature survey and my research work accomplished in this domain. In this work, I have focused on planning and implementing a standard ETL process incorporating real-time data integration features in a cloud environment to handle Big data for performing data analytics efficiently. In this Thesis, I have worked towards the modeling, simulation, and empirical analysis of traditional and real-time ETL processes and advanced proposal of ETL workflows management by use of Machine learning and shifting the ETL workload in the Cloud environment.