

2 Der ETL-Prozess

2.1 Grundlagen

ETL steht für »Extraktion, Transformation und Laden«. Hierunter versteht man den Vorgang der regelmäßigen Aktualisierung der Daten eines Data Warehouses bzw. einer OLAP-Applikation. Dabei müssen die Daten zunächst aus den Quellsystemen extrahiert, dann auf Konsistenz überprüft, gegebenenfalls angepasst und schließlich in die Datenbank des Data Warehouses geladen sowie in die OLAP-Würfel eingearbeitet werden. Die Erstellung eines geeigneten ETL-Prozesses ist häufig der aufwendigste Schritt bei der Data-Warehouse-Entwicklung. Er sollte nicht vernachlässigt werden, denn ein solide aufgebautes Data Warehouse ist nur mit einer qualitativ hochwertigen Datenbasis möglich.

2.1.1 Extraktion

In diesem ersten Schritt müssen die Quelldaten selektiert und für den folgenden Transformationsvorgang zur Verfügung gestellt werden. Dabei trifft man im Allgemeinen auf ein sehr hohes Maß an Heterogenität, denn die Ausgangsdaten werden durch die unterschiedlichsten in einem Unternehmen anzutreffenden Datenverarbeitungssysteme erzeugt.

Diese unter Umständen sehr großen Datenvolumina sind günstigstenfalls in relationalen Datenbanken abgelegt. Oftmals trifft man jedoch auch auf hierarchische Datenbanken oder Textdateien oder aber die Daten sind ausschließlich über Programmierschnittstellen zugänglich, die das jeweilige System zur Verfügung stellt.

Die Aufgabe eines ETL-Tools besteht nun darin, den Zugriff auf diese vielfältigen Datenformate zu ermöglichen. Dazu kann durch das eingesetzte ETL-Tool direkt auf die Datenbanken zugegriffen werden, sofern entsprechende native Treiber zur Verfügung stehen. Falls man Microsoft Windows als Betriebssystem verwendet, bietet sich der Einsatz von ODBC und des neueren OLE DB an. Diese durch Microsoft entwickelten Standards definieren einheitliche Schnittstellen, die durch nahezu alle Datenbanken und ETL-Tools unterstützt werden.

2.1.2 Transformation

Die Datentransformation ist die zentrale Aufgabe des ETL-Prozesses, bei dem die Ausgangsdaten an das geforderte Zielschema angepasst werden müssen. Weiterhin sollte die vorhandene Datenqualität dabei analysiert und automatisch mittels ausgewählter Algorithmen aus dem Bereich des Data Cleansing (siehe Abschnitt 2.2.3) erhöht werden.

Zur Durchführung der Konsistenzprüfungen und einer gegebenenfalls notwendigen Korrektur müssen die heterogenen Ausgangsdaten miteinander in Verbindung gesetzt werden. Dieser Vorgang kann auch bei mittelgroßen Datenmengen bereits sehr zeitaufwendig sein, da zum Beispiel zur Prüfung der referenziellen Integrität sehr viele so genannte Lookups nötig sind.

Ein Lookup liegt vor, wenn während einer Transformation eine weitere Abfrage durchgeführt wird, um zusätzliche Informationen zu dem momentan behandelten Datensatz zu erhalten. Zum Beispiel kann nachgesehen werden, ob zu allen in einer Bestellung aufgeführten Artikelnummern auch entsprechende Datensätze in der Artikeltabelle vorhanden sind, falls dies nicht durch eine entsprechende Fremdschlüsselbedingung abgesichert ist.

2.1.3 Laden

Nachdem die Daten in geprüfter Form zur Verfügung stehen, erfolgt die Integration in das Data Warehouse. Dazu müssen sie physisch in die Datenbank des Data Warehouses verschoben und die darauf aufbauenden Würfel mit ihren Aggregationen aktualisiert werden. In diesem Zusammenhang bezeichnet man den Arbeitsbereich, in dem sich die Daten bis zu diesem letzten Schritt befinden, als »Staging Area«.

Ab Beginn des Ladevorgangs bis zum Ende des ETL-Prozesses ist in Abhängigkeit des eingesetzten Produkts unter Umständen kein Zugriff der Endanwender auf das Data Warehouse und die OLAP-Dienste möglich. Deshalb ist es wichtig, sicherzustellen, dass dieser Vorgang stets im Rahmen des zur Verfügung stehenden Zeitfensters abgeschlossen werden kann. Im Allgemeinen ist dieses Zeitfenster in der Nacht gelegen, da dann die OLAP-Dienste nicht benötigt werden. Bei global vertretenen Unternehmen führt dieses Vorgehen jedoch zu Schwierigkeiten, so dass der ETL-Prozess so kurz wie möglich gehalten werden muss.

Sicherlich macht es keinen Sinn, die kompletten Daten bei jedem Aktualisierungsvorgang neu zu laden. Stattdessen ist es sehr viel zeit- und ressourcensparender, nur die veränderten und neu hinzugekommenen Daten zu laden. Dazu ist es erforderlich, schon in der Phase der Extraktion nur die veränderten Daten zu selektieren. Dies lässt sich relativ komfortabel realisieren, falls in jedem zu ladenden Ausgangsdatensatz der Zeitpunkt der letzten Änderung als so genannter