

## Erfahrungen zu ETL aus der Praxis

*ETL ist im Data Warehouse ein zentraler Prozess, der für den Endbenutzer jedoch nicht sichtbar ist. Wird ETL nicht richtig angepackt, kostet dies viel Entwicklungsarbeit, lange Ladeprozesse und im schlimmsten Fall schlechte Datenqualität. Es ist daher wichtig, dass in diesem Bereich professionell gearbeitet wird.*

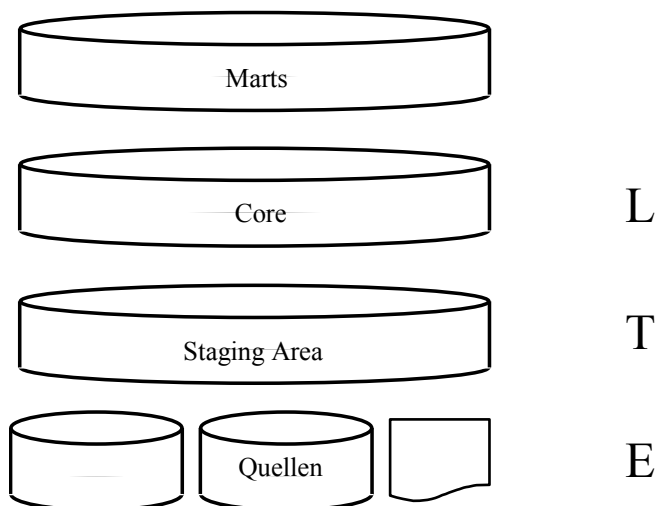
Dr. Andrea Kennel, Trivadis AG

In einem Data Warehouse werden die Daten üblicherweise schrittweise umgewandelt, bis sie in einer Form sind, die für die Abfragen geeignet ist. Konkret werden die Daten aus den unterschiedlichen Quellen in einen vorgelagerten Bereich, die Staging Area, geladen. Hier werden die Daten soweit bereinigt, dass sie einfach in den Datenspeicher, das Core, geladen werden können. In diesem Schritt werden die Daten historisiert. Da die Daten im Core möglichst vollständig und historisiert sein sollen, sind direkte Abfragen auf das Core zu komplex und langsam. Daher werden die Daten für die unterschiedlichen Anforderungen und Abfragebedürfnisse in Marts bereitgestellt. Dabei stellt ein Mart immer ein Ausschnitt der Daten dar.

### **Was ist ETL (Extraction, Transformation, Load)**

ETL steht für Extraktion, Transformation und Laden. Dieser Prozess ist vor allem beim Laden des Cores aus den Quellen via Staging wichtig. Die drei Elemente werden hier kurz allgemein beschrieben.

### **Data Warehouse Referenzarchitektur**



## **Extraktion**

Die Extraktion beschreibt wie die Daten aus der Quelle gelesen werden. Dabei gibt es die Möglichkeit, dass die Daten in der Quelle zusammengestellt und in einem File geliefert werden oder dass sie via DB-Link direkt in die Staging Area gelesen werden.

Weiter muss festgelegt werden, wie oft Daten extrahiert werden sollen und wer den Extrakt auslöst. Normalerweise wird ein Extrakt automatisch ausgelöst, entweder jede Nacht, jede Woche oder jeden Monat.

## **Transformation**

Kommen Daten aus unterschiedlichen Quellen, so müssen oft Transformationen vorgenommen werden. Beispielsweise müssen Einheiten und Währungen abgeglichen werden.

## **Laden**

Sind die Daten transformiert in der Staging Area, so muss geprüft werden, welche Daten schon im Core vorhanden, welche Daten neu sind oder geändert wurden. Man sollte meinen, dass dies bereits beim Extrakt geschehen sollte. Soweit dies über Zeitstempel möglich ist, werden nur Daten extrahiert, die nicht schon geladen wurden. Doch kann es vorkommen, dass in der Quelle ein Attribut geändert wird, das im DWH nicht interessiert. Dann kommt der entsprechende Datensatz als geändert in die Staging Area.

## ***Wo braucht es ETL***

ETL wird primär für das Laden des Cores, also des eigentlichen Warehouses eingesetzt. Dies wird oft mit entsprechenden ETL-Tools wie Informatica oder Oracle Warehouse Builder (OWB) implementiert.

Diese Tools und das Verfahren des ETL kann auch beim Laden der Marts aus dem Core eingesetzt werden. Dieser Ladeprozess ist aber oft einfacher, da keine Versionierung gemacht wird.

## ***Probleme und Lösungsansätze***

### **Zuständigkeit für Datenqualität**

Ein grosses Problem beim ETL ist die oft mangelhafte Datenqualität. Das Warehouse verlangt von allen Datenquellen dieselbe Qualität um verlässliche Aussagen zu ermöglichen.

Ein Beispiel: Jede Fluggesellschaft will wissen, welcher Flug sich lohnt. Aus Sicherheitsvorschriften kann genau festgestellt werden, wie viele und welche Passagiere in einem Flugzeug sassen. Anders verhält es sich mit den Postsäcken. Da ist es weniger wichtig, ob der Postsack einen Flug früher oder später mitfliegt. Will man nun je einzelnen Flug eine präzise Kostenrechnung erstellen, so muss man einen Postsack eindeutig einem Flug zuordnen, auch wenn die gelieferten Angaben unvollständig oder falsch sind.

Wer soll nun falsche Daten korrigieren? Offensichtlich gibt es zwei Möglichkeiten: die Daten werden in der Quelle korrigiert und neu geliefert oder die Daten werden im Warehouse korrigiert. Dies ist normalerweise kein technisches, sondern ein administratives Problem.

### **Diverse Quellen mit unterschiedlichen Codierungen**

Das Problem der unterschiedlichen Codierungen wird in der Quellanalyse oft unterschätzt und taucht im schlimmsten Fall erst beim ersten produktiven Ladeprozess auf. Die Lösungen können unterschiedlich aussehen und reichen von einfachen Mappingtabellen bis hin zur manuellen Datenbereinigung. Wieso manuell? Daten zu Personen sind oft unpräzise. So ist es nicht einfach zu bestimmen, ob der Peter Muster aus Quelle 1 dieselbe Person ist, wie Peter Muster aus Quelle 2. Da bleibt oft nur die Akzeptanz von Duplikaten oder eine manuelle Bereinigung basierend auf weniger strukturierter Zusatzinformation.

### **Historisierungsmöglichkeiten (Snapshot, SCD I und SCD II)**

Bewegungsdaten oder Fakten ändern sich häufig und werden entsprechend mit einem Zeitstempel als sogenannte Snapshots gespeichert. Doch auch Dimensionsdaten können sich ändern. So kann ein Kunde von Zürich nach Bern und später nach Basel umziehen. Nun stellt sich die Frage, ob diese Geschichte des Kunden interessiert.

Interessiert die Geschichte nicht, so kann der Wohnort einfach überschrieben werden. Dies wird nach Kimball Slowly Changing Dimension Typ I (SCD I) benannt.

Interessiert die Geschichte des Kunden, so werden vom Kundendatensatz mehrere Versionen mit einer Gültigkeitsperiode gespeichert. Nach Kimball bezeichnet man dies als Changing Dimension Typ II (SCD II).

Leider werden noch nicht alle möglichen Historisierungsarten von allen Tools optimal unterstützt, so dass oft eigene Prozeduren geschrieben werden müssen.

### ***Best Practice bei ETL***

Sobald vom Analyseprozess her bekannt ist, welche Daten aus welchen Quellen geladen werden müssen, kann mit der Spezifikation des ETL begonnen werden. Unabhängig davon, welches ETL-Werkzeug eingesetzt wird, muss zuerst festgelegt werden, wie welche Daten geladen werden. Dazu gehört:

- Snapshot
- SCD I und SCD II
- Verdichtungen
- Komplexe Datenbereinigung

Je nach vorhandenen Tools, Vorkenntnissen und Komplexität des Ladevorgangs gibt es verschiedene Implementationsmöglichkeiten:

- Grafisch mit Tools
- Generatoren (selber geschrieben)
- Manuelle Programmierung

Teilweise können die erwähnten Möglichkeiten auch kombiniert werden. Umso wichtiger wird dann eine saubere Dokumentation. Denn der Ladeprozess wird erfahrungsgemäss mit neuen Quellen erweitert werden. Für eine Automatisierung des Ladeprozesses ist ein einheitliches und übersichtliches Scheduling zentral. Dies kann am besten mit einem Werkzeug erzielt werden.

### ***Fazit***

Wird ETL nicht richtig angepackt, kostet dies viel Entwicklungsarbeit, lange Ladeprozesse und schlechte Datenqualität. Dies bedeutet nicht zwingend, dass ein teures Werkzeug eingekauft wird. Es braucht vor allem Fachwissen und Erfahrung, wobei ein gutes Werkzeug natürlich die Dokumentation und Umsetzung erleichtert.

Andrea Kennel ist Consultant bei Trivadis AG

[Andrea.kennel@trivadis.com](mailto:Andrea.kennel@trivadis.com)