Populate the Stage
1001 Möglichkeiten eine Staging Area zu füllen

Sven Bosinger
ist-people
Frankfurt

Schlüsselworte
Data Warehouse, Business Intelligence, Staging Area, ETL, Datenbewirtschaftung

Einleitung

Mit Oracle-Bordmitteln gibt es mindestens genauso viele Möglichkeiten eine Tabelle in einer Staging Area zu füllen wie es Datenbank Releases gibt. Vom Flatfile über einen Database-Link bis hin zum Change Data Capture stellt die Datenbank mannigfaltige technologische Möglichkeiten zur Verfügung, um Daten effizient von A nach B zu transportieren.

Der Vortrag stellt die wichtigsten Verfahren vor und ordnet sie nach den Gesichtspunkten Geschwindigkeit, Stabilität, Komplexität und Wartbarkeit ein. Dabei werden vor allem die technologischen Aspekte von Flat-Files, Datenbank-Links, Queuing, Replikationsmechanismen und Transportable Tablespaces beleuchtet.

Ausgangslage

In einer ERP-Datenbank gibt es eine Tabelle, welche in die Staging Area eines Data-Warehouse übernommen werden soll. Dazu werden im Weiteren unterschiedliche Lösungen betrachtet. Als Ausgangsbasis dient die Tabelle Umsatz. Diese wird initial mit 90.000 Datensätzen gefüllt. In einem zweiten Schritt werden weitere 10.000 Inserts und 5.000 Updates auf dieser Tabelle durchgeführt. Sowohl die initiale Befüllung als auch die Inkrementellen Inserts und Updates sollen in eine gleichnamige Stage-Tabelle übernommen werden.

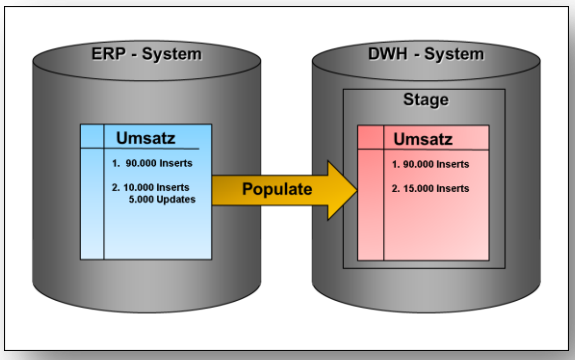

Abb. 1: Füllen der Stage-Tabelle

Dies geschieht in zwei Transaktionen: Insert der initialen Sätze und danach Insert und Update der inkrementellen Sätze. In der Stage- Tabelle sollen sowohl die Inserts als auch die Updates als einzelner Datensatz gespeichert, und über ein Flag (IUD) gekennzeichnet werden. Die Weiterverarbeitung der Stage-Daten ist nicht Bestandteil der weiteren Betrachtung.

Lösungsalternative
Um eine Vergleichbarkeit der Lösungsalternativen zu gewährleisten werden die Tabellen in der ERP und DWH-Datenbank jeweils leer erzeugt und dann die ERP-Tabelle zuerst initial und darauf folgend inkrementell gefüllt. Die dargestellten Verfahren werden dabei zweimal durchlaufen, einmal für die initialen Daten (90.000 Inserts) und danach für die inkrementellen Daten (10.000 Inserts und 5.000 Updates). Der Transport der Daten aus der ERP- in die DWH- Datenbank erfolgt über folgende Lösungsalternativen:

1. Flat-File Upload mit dem SQL-Loader:
    a. Die Daten werden mit einer PL/SQL-Procedure eingefügt/geändert
    b. Die Daten werden mit einer PL/SQL-Procedure in ein Comma Separated File (.csv) geschrieben
    c. Das File mit den initialen Daten wird in das Filesystem der DWH- Datenbank kopiert
    d. Das File wird mittels SQL-Loader in die Stage-Tabelle geladen

2. Flat-File Upload mit External Table:
    a. Die Daten werden mit einer PL/ SQL-Procedure eingefügt/geändert
    b. Die Daten werden mit einer PL/ SQL-Procedure in ein Comma Separated File (.csv) geschrieben
    c. Das File mit den initialen Daten wird in das Filesystem der DWH- Datenbank kopiert
    d. Das File wird mittels einer External Table in die Stage-Tabelle geladen

3. Datapump Upload mit External Table:
    a. Die Daten werden mit einer PL/SQL-Procedure eingefügt/geändert
    b. Die Daten werden mit Datapump in ein File geschrieben
    c. Das File mit den initialen Daten wird in das Filesystem der DWH- Datenbank kopiert
    d. Das File wird mittels einer External Table in die Stage-Tabelle geladen

4. Insert as Select (IAS) über einen Database-Link:
    a. Die Daten werden mit einer PL/SQL-Procedure eingefügt/geändert
    b. Die Daten werden mit einem Insert as Select über einen Datenbank-Link aus der ERP-Datenbank gelesen und in die Stage-Tabelle der DWH-Datenbank eingefügt

5. Create Table as Select (CTAS) über einen Database-Link:
    a. Die Daten werden mit einer PL/SQL-Procedure eingefügt/geändert
    b. Die Daten werden mit einem Create Table as Select über einen Datenbank-Link aus der ERP-Datenbank gelesen und in die dabei angelegte Stage-Tabelle der DWH-Datenbank eingefügt

6. Datapump:
    a. Die Daten werden mit einer PL/SQL-Procedure eingefügt/geändert
    b. Die Daten werden mit Datapump über einen Datenbank-Link aus der ERP-Datenbank gelesen und in die Stage-Tabelle der DWH-Datenbank eingefügt.

7. Trigger auf Quelltabelle (direkt):
    a. Die Daten werden mit einer PL/SQL-Procedure eingefügt/geändert
    b. Die Daten werden durch einen Insert/Update-Trigger auf der ERP-Tabelle über einen Datenbank-Link aus der ERP-Datenbank gelesen und beim Commit in die Stage-Tabelle der DWH-Datenbank eingefügt

8. Trigger auf Quelltabelle (indirekt):
    a. Die Daten werden mit einer PL/SQL-Procedure eingefügt/geändert

   b. Die Daten werden durch einen Insert/Update-Trigger auf der ERP-Tabelle in eine temporäre Tabelle in der ERP-Datenbank geschrieben
   c. Die Daten werden mit einem Insert as Select über einen Datenbank-Link aus der temporären ERP-Tabelle gelesen und in die Stage-Tabelle der DWH-Datenbank eingefügt

9. Trigger auf Quelltabelle mit Advanced Queuing (AQ):
   a. Die Daten werden mit einer PL/SQL-Procedure eingefügt/geändert
   b. Die Daten werden durch einen Insert/Update-Trigger auf der ERP-Tabelle in eine Advanced Queue in der ERP-Datenbank geschrieben
   c. Diese wird über einen Database-Link vom DWH gelesen und die Stage-Tabelle befüllt

10. Change Data Capture:
    a. Die Daten werden mit einer PL/SQL-Procedure eingefügt/geändert
    b. Die Daten werden durch einen Change Data Capture Asynchronous Distributed HotLog Mechanismus von der ERP-Datenbank auf die DWH-Datenbank repliziert

11. Transportable Tablespace:
    a. Die Daten werden mit einer PL/SQL-Procedure eingefügt/geändert
    b. Der Tablespace der ERP-Tabelle wird Offline gesetzt
    c. Die Datenfiles des Tablespaces werden in das Filesystem der DWH-Datenbank kopiert
    d. Der Tablespace der ERP-Tabelle wird wieder Online gesetzt
    e. Der kopierte Tablespace wird in die DWH-Datenbank eingefügt

12. ODI und Golden Gate:
    a. Der Oracle Data Integrator (ODI) generiert eine eigene Change Data Capture Funktionalität. Dabei kann auch das Produkt Oracle Golden Gate genutzt werden.

Bewertung
Eine Bewertung der einzelnen Szenarien wird in verschiedenen Kategorien vorgenommen:

1. Stabilität: Wie Stabil ist die eingesetzte Technologie, was passiert bei Abbrüchen, gibt es Wiederaufsetzpunkte, wie fehlertolerant ist das Verfahren?

2. Geschwindigkeit: Wie viele Datensätze können pro Zeiteinheit verarbeitet werden?

3. Komplexität/Wartbarkeit: Wie komplex ist der Aufbau und Betrieb der Lösung, ist spezielles Know How erforderlich, werden zusätzliche Produkte benötigt?

4. Funktionalität: Welchen Funktionsumfang hat die Lösung, wird real-/neartime unterstützt, wird eine automatische Protokollierung vorgenommen, werden Fehler reportet?

5. Aufwand/Kosten: Welche Lizenzkosten entstehen, welcher Programmieraufwand steckt dahinter, wie Aufwendig sind Änderungen?

Führt man die Bewertung der einzelnen Szenarien durch so kommt man zu folgender Bewertungsmatrix:

	Flat-File mit SQL-Loader	Flat-File mit External Table	Datapump mit External Table	Insert as Select	Create Table as Select	Datapump	Trigger (direkt)	Trigger (indirekt)	Advanced Queuing	Change Data Capture	Transportable Tablespace	ODI und Golden Gate
Stabilität	++	++	++	+	+	+	-	+	+	o	+	o
Geschwindigkeit	+	+	++	++	++	++	o	o	-	o	++	o
Komplexität/Wartbarkeit	+	+	+	++	+	o	+	+	-	--	+	o
Funktionalität	+	+	+	o	-	o	++	++	++	++	-	++
Aufwand/Kosten	+	+	+	++	+	+	o	o	o	o	+	--

Empfehlung
Eine allgemeine Empfehlung für ein bestimmtes Szenario kann nicht gegeben werden! Es kommt hierbei auf das genaue Anforderungsprofil der DWH-Lösung an. Wird z.B. großer Wert auf real/neartime gelegt, so sind die klassischen Batch-Verfahren (z.B. IAS oder CTAS) nicht zu empfehlen. Kommt es aber auf Geschwindigkeit in einem klassischen Daily-Load an, so sind diese Verfahren sicher in der engeren Auswahl.
Um zu einer auf das Anforderungsprofil ausgerichteten Empfehlung zu kommen, ist es notwendig die einzelnen Kriterien zu gewichten (z.B. Geschwindigkeit geht zu 50%, aber Funktionalität nur zu 10% in das Ergebnis ein). Diese Gewichtung ist DWH-spezifisch. Daher wurde hier bewusst auf eine Empfehlung verzichtet.

Kontaktadresse:
Sven Bosinger
its-people
Lyoner Straße 44-48
D-60528 Frankfurt am Main

Telefon: +49 (0) 69-247521-00
Fax: +49 (0) 69-247521-021
E-Mail sven.bosinger@its-people.de
Internet: www.its-people.de