# Global Staging Area

# Implementierung einer zentralen Datendrehscheibe

**Sven Bosinger**
**ist-people**
**Frankfurt**

**Schlüsselworte**
Data Warehouse, Business Intelligence, Staging-Area, ETL, Datenbewirtschaftung, ETL-Metadaten

**Einleitung**
Die Versorgung eines Data Warehouse (DWH) mit frischen Daten kann zuweilen eine große Herausforderung sein. Es muss in der Regel nicht nur ein DWH-System mit Daten versorgt werden sondern zumeist mehrere Systeme, z.B. die Entwicklungs-, Test-, Integrations-, Wartungs- und Produktions-Umgebungen. Jede dieser Umgebungen hat spezielle Anforderungen. Gleichzeitig soll auf der Datenlieferanten-Seite die Anzahl der Schnittstellen, über die Daten abgegeben werden, überschaubar bleiben. Zusätzlich spielen regulatorische und rechtliche Vorgaben eine Rolle. Entwickler dürfen immer häufiger keinen Zugang mehr zu personalisierten Daten erhalten, sondern müssen auf maskierten und verfremdeten Daten entwickeln.

In der hier dargelegten Lösung geht es um die Implementierung einer zentralen Datendrehscheibe genannt Global Staging Area (GSA), die aus verschiedensten Quellsystemen mit Daten bestückt wird. Die GSA gibt wiederum die gepufferten Daten an die diversen DWH-Systeme gezielt weiter. Dadurch wird in jedem Quellsystem nur noch eine Schnittstelle benötigt, die damit Datenlieferant für alle nachgelagerten DWH-Systeme ist. Die GSA entscheidet dann nach einem vorgegebenen Regelwerk, wann, wie und in welcher Form die Daten an die DWH-Systeme weitergegeben werden. So kann ein permanenter Datenstrom mit allen Echtdaten an die Produktionsumgebung eingerichtet werden, wohingegen die Entwicklungsumgebung mit einem reduzierten und verfremdeten Datenbestand versorgt wird. Neu Quellsysteme können einfach über Oracle Standardtechnologien (Streams, CDC, AQ, Trigger, ...) an die GSA angebunden werden.

**Ausgangslage**
Viele Anwender einer DWH-Lösung haben sich für einen klassischen Aufbau ihres DWHs entschieden. Dabei werden die Daten der Quellsysteme in einer Staging Area gesammelt, durch einen Batch-Lauf in ein zentrales Enterprise Modell integriert und abschließend Business Area spezifische Data Marts aufgebaut.

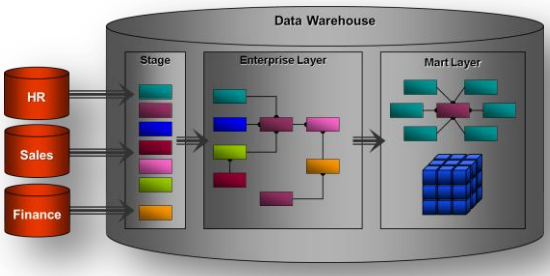


*Abb. 1: klassisches DWH*

In der Regel betreibt ein Anwender aber nicht nur eine DWH-Instanz sondern mehrere. Je nach Vorgehen werden neben der Produktionsinstanz noch Instanzen für Entwicklung, Test, Abnahme und Wartung benötigt. Dies bedeutet, dass nicht nur eine Instanz permanent mit Daten aus den Quellsystemen versorgt werden muss, sondern beliebig viele. Bei einer DWH-Entwicklung besteht zudem die Besonderheit, dass eine erfolgreiche Entwicklung nur auf produktionsnahen Echtdaten und nicht auf Testdaten möglich ist. Jegliche DWH-Entwicklung ist eine Daten getriebene Entwicklung.

Fragen der Performance, statistische Auswertungen und Variationsvielfalt sind durch eingeschränkte Testdaten in der Regel nicht zu beantworten. Es ist also häufig notwendig schon in den Entwicklungs-Instanzen mit Produktionsdaten zu arbeiten. Spätestens in der Abnahmeumgebung muss auf Produktionsdaten gearbeitet werden, um abnahmefähige Testfälle generieren zu können. Daher ergibt sich die Problematik, dass die produktiven Quellsysteme nicht nur mit dem Produktions-DWH über Schnittstellen verbunden werden müssen, sondern auch mit den übrigen DWH-Systemen. Dies führt zu einem überproportionalen Anwachsen der Anzahl der Schnittstellen.

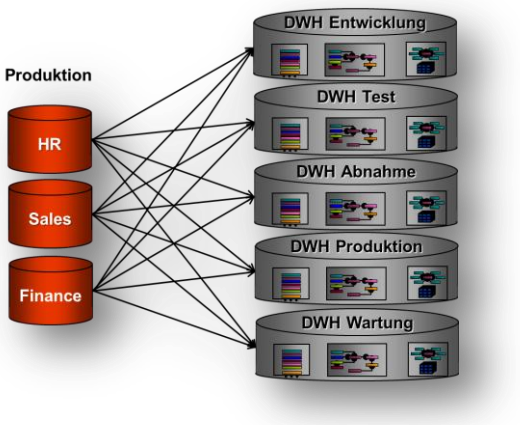



*Abb. 2: Schnittstellen-Explosion*

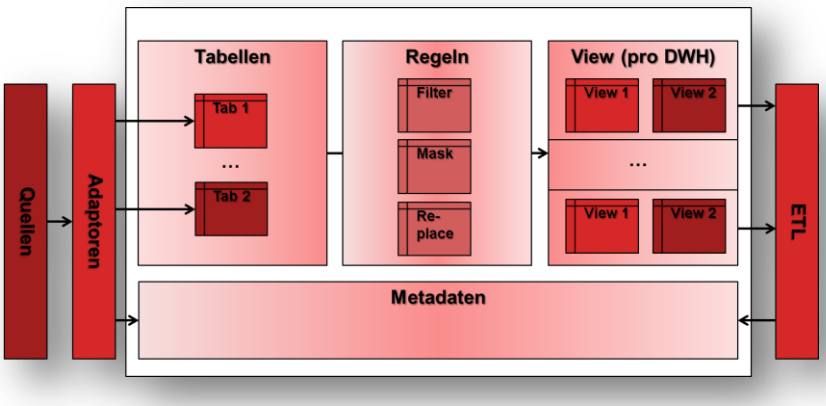
Aufgrund von Datensicherheitsaspekten dürfen häufig sensible Produktionsdaten, wie z.B. Kreditkarten Informationen oder Bankdaten, nicht in eine ungeschützte Entwicklungsumgebung gelangen. Vor allem nicht, wenn bei der Entwicklung Near- oder Offshore Kräfte eingesetzt werden sollen. Hier müssen Daten gegebenenfalls verfremdet oder ausgeblendet werden.

**Global Staging Area**


*Abb. 3: Global Staging Area*

Das Verfahren der Global Staging Area (GSA) ersetzt alle lokalen Staging Areas in den einzelnen DWH-Instanzen. Alle DWH-Instanzen verarbeiten die Stage Daten weiterhin im Rahmen eines klassischen ETL-Prozesses. Insofern wird auf die GSA zugegriffen, als ob es sich um eine klassische lokale Staging Area handeln würde. Die Quellsysteme werden ausschließlich über Schnittstellen an die GSA angebunden. Daher muss für jedes Quellsystem nur noch eine Schnittstelle definiert werden, egal wie viele DWH-Instanzen bedient werden müssen. Dabei wird ausnahmslos ein Push-Verfahren angewandt. D.h. die Quellsysteme stellen die Datenlieferungen zusammen und übertragen diese direkt in die GSA. Dabei werden die technischen Verfahren Change Data Capture, Advanced Replication, Advanced Queuing, Streams und Trigger (via Database Link) unterstützt.

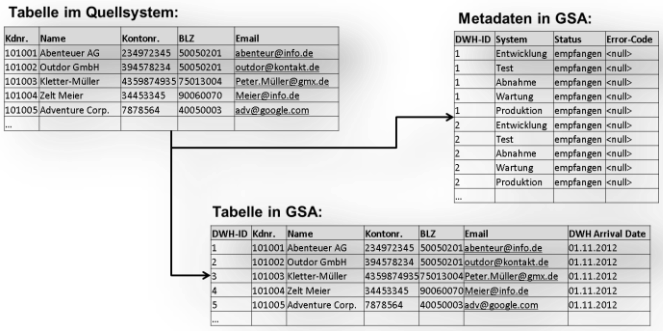
Die Quellsysteme liefern die Produktionsdaten umfänglich, d.h. es werden die Daten so geliefert, wie sie im Produktions-DWH benötigt werden. Etwaige Filterungen, Datenmaskierungen oder Verfremdungen werden innerhalb der GSA durchgeführt. Diese Aktivitäten sind DWH-Instanz spezifisch, d.h. die Produktionsumgebung



bekommt z.B. die Daten ungefiltert, wohingegen die Daten für die Test-Umgebung gefiltert und Konto-Informationen maskiert werden können. Die Bereitstellung der Daten für die nachgelagerten ETL-Prozesse wird von der GSA durch entsprechende Instanz spezifischen Sichten auf die Stage Daten gewährleistet. Während des Aufbaus dieser Sichten werden dabei metadatengesteuert die entsprechenden Filterregeln, Maskierungen und Verfremdungen angewandt. Jede DWH-Instanz erhält nur Zugriff auf seine Sichten.

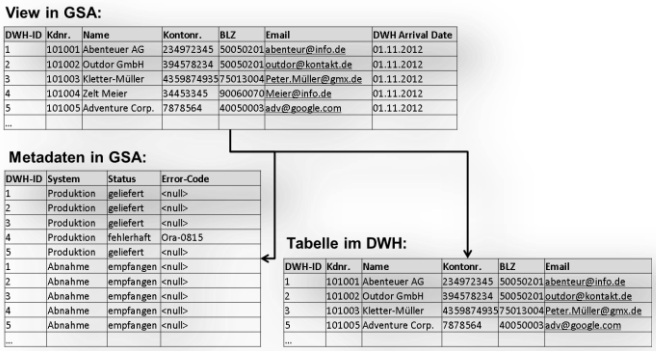
In der GSA werden drei (vier) Prozesse betrieben:

1. Push in die GSA:



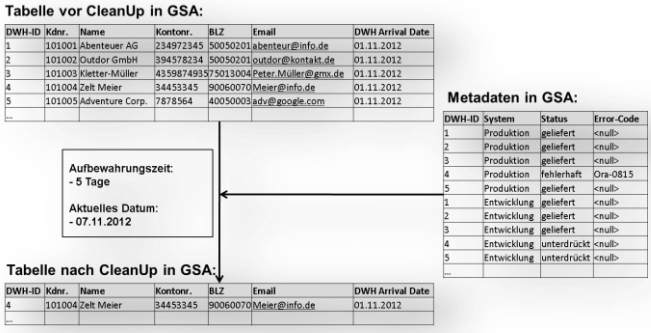

Die Quellsysteme liefern die Daten Real-/Neartime in die GSA. Dort werden sie ungeprüft in die entsprechenden Stage-Tabellen eingefügt. Jeder Datensatz wird mit einer systemweiten, eindeutigen DWH-ID und einem Liefer-Datum versehen. Zusätzlich wird im Metadatenkatalog der Status jeden einzelnen Datensatzes DWH-Instanz bezogen festgehalten. Das Einfügen der Daten in die GSA ist transaktionsgesichert.

2. Pull aus den DWHs:




Die Verarbeitung der Daten aus der GSA in der jeweiligen DWH-Instanz erfolgt mittels eines klassischen ETLs. Die zu verarbeitenden Sätze werden aus Instanz spezifischen Views gelesen und im Metadatenkatalog die erfolgreiche Verarbeitung protokolliert. Fehler werden ebenfalls im Metadatenkatalog festgehalten. Die Views werden pro Instanz unter Berücksichtigung der Instanz spezifischen Filterregeln, Maskierungen und Verfremdungen erzeugt. Der Pull pro Instanz kann völlig unabhängig von allen anderen Instanzen betrieben werden. Unterschiedliche Ladezeitpunkte und Frequenzen sind problemlos realisierbar.

3. CleanUp der SGA:




Alle erfolgreich in die DWH-Instanzen geladenen Daten, erkenntlich über den Status im Metadatenkatalog, werden regelmäßig, asynchron gelöscht. Dazu kann eine Vorhaltezeit definiert werden,

d.h. dass zu Nachverfolgungszwecken die Daten einen definierten Zeitraum in der GSA verbleiben.

4. Real-/Neartime-Auswertungen (optional):
   Solange die Daten in der GSA stehen, können sie zusätzlich noch zu Auswertungszwecken genutzt werden. Sobald sie aus der GSA gelöscht wurden, stehen sie im DWH bereit. Real-/Neartime Auswertungen können somit auf der SGA einfach realisiert werden.

**Vorteile**
Wie oben dargelegt bietet der Einsatz einer GSA eine Reihe von Vorteilen:

- Die Anzahl der Schnittstellen in den Quellsystemen wird massiv reduziert. Pro Quellsystem ist lediglich eine Schnittstelle notwendig, um beliebig viele DWH-Instanzen mit Daten zu versorgen. Dadurch sinkt die Komplexität der Schnittstellen.

- Alle DWH-Instanzen werden permanent über ein Push-Verfahren mit aktuellen Echtdaten versorgt. Dadurch kann bei der Entwicklung schon auf realistischen Datenmengen gearbeitet werden. Darüber hinaus können alle vorkommenden Datenkonstellationen berücksichtigt werden.

- Die Datenmenge in der GSA ist deutlich geringer als die Datenmenge aller lokalen Staging Areas zusammen. Jeder Datensatz wird nur einmal gespeichert und kann an beliebig viele DWH-Instanzen verteilt werden. Die sichten sind lediglich logischer Natur und sind als Datenbank Views aufgeteilt auf ein Schema pro DWH-Instanz realisiert.

- Die Datenmengen können, wenn gewollt, für einzelne Instanzen einfach durch konfigurierbare Filterregeln reduziert werden.

- Sicherheitsrichtlinien werden durch Metadatenkonfiguration einfach und transparent umgesetzt, so können Maskierung und Verfremdung der Daten Instanz abhängig eingestellt werden.

- Das komplette Verfahren ist Transaktionsgesichert, d.h. es können keine Datensätze verloren gehen. Kommt es zu Abbrüchen oder Lieferausfällen so werden keine unvollständigen Lieferungen gespeichert und damit auch nicht im ETL verarbeitet.

- Durch das Push-Verfahren werden Daten in Real-/Neartime in die GSA gespeichert, dadurch ist ein Real-/Neartime Reporting einfach zu realisieren.

- Zusätzliche DWH-Instanzen, die eventuell sogar nur temporär benötigt werden, sind einfach und schnell mit Daten zu versorgen. Lediglich die Instanz bezogenen Views müssen einmalig erzeugt werden.

Größere Nachteile konnten beim Einsatz einer GSA bisher nicht festgestellt werden. Die Implementierung einer GSA kann bei Vorliegen der oben beschriebenen Vorbedingungen (mehrere DWH-Instanzen greifen auf die gleichen Quellesysteme zu) uneingeschränkt empfohlen werden.

**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