Sind Ihre Daten noch von gestern?
Moderne Datenintegration für's Data Warehouse

Karsten Stöhr
ORACLE Deutschland B.V. & Co. KG
Hamburg

Schlüsselworte:

Operational Business Intelligence, Data Warehouse, Real-time Data Integration, Oracle Data Integrator, Oracle GoldenGate

Einleitung

Begnügen Sie sich in Ihrem Data Warehouse noch mit den Daten von gestern? Die Erweiterung des DWH um aktuelle, nahezu Echtzeitdaten ist die Grundlage für operationale Business Intelligence und schafft so neue Möglichkeiten, z.B. zur
- Betrugsprävention statt späterer Schadensanalyse,
- Optimierung von laufenden Kampagnen statt später Erkenntnis,
- Kundenbindung statt Zählung der verlorenen Kunden,
- Lieferkettensteuerung statt Kalkulation der Verluste durch Verzögerungen.

In diesem Vortrag wird eine Lösung vorgestellt, welche es gestattet, einen Operational Datastore oder ein Data Warehouse mit nahezu Echtzeitdaten zu füttern, ohne dabei die tägliche Arbeit mit den primären Systemen durch die zusätzlich entstehende Last merklich zu beeinträchtigen.

Das grundlegende Dilemma der immer weiter wachsenden Datenmengen bei gleichzeitig schwindenden Zeitfensterrn stellt unmißverständlich klar, daß der allabendliche Batchlauf über kurz oder lang ausgedient hat. Es heißt Abschied zu nehmen von einem ohnehin ungeliebten Kind. Die Datensynchronisation per nächtlichem Batchjob war nicht aus geschäftlichem Nutzen geboren sondern aus reiner Not, um den täglichen Betrieb nicht zu stören.

Non-invasives, heterogenes, Echtzeit-Data-Capturing mit Oracle GoldenGate ist die moderne Alternative zum Batchjob. Häufige, aktuelle E-LT Transformationen direkt auf der (Ziel-) Datenbankplattform mittels Oracle Data Integrator sind die moderne Alternative zu ETL.

Moderne Datenintegration

Verläßliche Business Intelligence (BI) Lösungen erfordern eine effektive Integration von Informationen aus einer Vielzahl von Quellen. Manche Geschäftsvorgänge erfordern einen zuverlässigen Einblick in den in dieser Minute gerade aktuellen Stand der Dinge, da ein Mitarbeiter nur mit vollständig akurater und aktueller Information geeignete Maßnahmen ergreifen kann. Um die operationale Effizienz und Effektivität zu erhöhen, suchen Unternehmen nach Wegen, eine robuste BI Infrastruktur aufzubauen, die einerseits möglichst zeitnahe operationale Informationen verwendet und andererseits den historischen Zusammenhang im Auge behält.

Aktuellere Daten

So manche Information ist ein verderbliches Gut. Mit der Zeit verlieren Daten ihren Wert, da sie sich immer weiter entfernen vom eigentlichen Ereignis, welches sie repräsentieren. Viele Anwender würden gerne von den Vorteilen von Echtzeitdaten aus geschäftskritischen Anwendungen in ihrem Data Warehouse profitieren, schrecken aber vor der möglichen Beeinträchtigung der täglichen Arbeit mit denselben Anwendungen durch die zusätzlich generierte Last zurück, die eine häufige Synchronisation mit den primären Systemen mit traditionellen Methoden mit sich bringen würde. Daher wird die Datensynchronisation zumeist auf nächtliche Stunden außerhalb der üblichen Arbeitszeit verschoben, woraus resultiert, daß das Data Warehouse nie aktuelle Daten enthält.

Immer größere Datenmengen, immer kürzere Zeitfenster

Viele Jahre waren Batchläufe zur Extrahierung der Daten aus den Quelltabellen und anschließendem Laden der gesammelten Daten in das Zielsystem das Mittel der Wahl für die tägliche Datensynchronisation. Doch mit der Globalisierung und der Ausbreitung des Internets stieg die Erwartungshaltung des Kunden, daß er jeden Tag der Woche 24 Stunden lang Geschäftsvorgänge initiieren und auf Services zugreifen kann. Die Geschäftsöffnungszeiten werden länger und in Folge die „Nachtfenster“ für die Ausführung der Batchläufe immer kürzer. Gleichzeitig wachsen die Datenmengen, die innerhalb der kürzer werdenden Zeitfenster transferiert werden müssen, weiter an.

Selbst wenn derzeit die erforderliche Datenmenge noch gerade eben in der zur Verfügung stehenden Zeit extrahiert und geladen werden kann, stellt sich die Frage, wie lange das noch gut geht. Spätestens bei einem unerwarteten Abbruch des Batchlaufes z.B. durch einen Systemfehler, kann es leicht passieren, daß nach Behebung des Problems nicht mehr genügend Zeit zur Verfügung steht, um den Batchlauf zu Ende zu führen.

Überdies kann eine ungewollte Unterbrechung des Ladevorgangs das Data Warehouse in einen inkonsistenten Zustand der Daten versetzen, was eine noch deutlich länger dauernde Wiederherstellung der Daten erfordern kann.

Abschied vom täglichen Batchjob

Das grundlegende Dilemma der immer weiter wachsenden Datenmengen bei gleichzeitig schwindenden Zeitfensterrn stellt unmißverständlich klar, daß der allabendliche Batchlauf über kurz oder lang ausgedient hat. Es heißt Abschied zu nehmen von einem ohnehin ungeliebten Kind. Die Datensynchronisation per nächtlichem Batchjob war nicht aus geschäftlichem Nutzen geboren sondern aus reiner Not, um den täglichen Betrieb nicht zu stören.

Die stärker werdende geschäftliche Forderung nach immer aktuelleren Daten verstärkt das Grundproblem der Datensynchronisation per nächtlichem Batchlauf noch weiter.

Wege zu aktuelleren Daten

Um aktuellere Daten zu erhalten, kommen drei grundsätzliche Architekturvarianten in Betracht:

- Anpassung des ETL-Systems, um mehrere große Intraday Batches zu extrahieren
- Anpassung des ETL-Systems, um nur geänderte Daten in Minibatches zu extrahieren

- Gemeinsamer Einsatz eines ELT-Werkzeugs zusammen mit einer dedizierten Lösung für Change Data Capture

Mehrere große Intraday Batchläufe

Ein Weg zu aktuelleren Daten ist, die bestehenden Batchjobs mehrmals am Tag laufen zu lassen, z.B. alle 4 – 6 Stunden. Diese Strategie hat den Vorteil, daß die bestehende ETL-Infrastruktur mit nur geringen Änderungen weiterverwendet werden kann. Doch aus den oben geschilderten Gründen ist dieser Weg kaum gangbar oder nur mit immensen Investitionen in leistungsfähigere Hardware. Wenn schon die Synchronisation einmal täglich auf die Nachtstunden verschoben wird, um die tägliche Arbeit nicht zu stören, ist eine häufigere Ausführung derselben Batchjobs mehrmals am Tag eher illusorisch.

Zudem ist eine Synchronisierung alle 4 – 6 Stunden zwar schon ein Fortschritt gegenüber tagesalten Daten aber noch sehr weit entfernt von Echtzeitdaten.

Übertragung nur geänderter Daten per Minibatches

Bei dieser Methode werden nur die Daten übertragen, die seit der letzten Synchronisation geändert wurden. Wird diese Synchronisation in entsprechend kurzen Intervallen ausgeführt, erhält man deutlich aktuellere Daten und die zu übertragende Datenmenge pro Minibatch bleibt in einem moderaten Rahmen.

Um diese Methode einsetzen zu können, sind jedoch mehr oder weniger umfangreiche Änderungen an den bestehenden Datenstrukturen und den ETL-Prozessen nötig. Oftmals müssen weitere Felder, z.B. Zeitstempel, zu den Datentabellen hinzugefügt werden, um neue Änderungen erkennen zu können. Um alle geänderten Einträge schnell selektieren zu können, sollte wohl auch ein neuer Index auf diesen Zeitstempel ergänzt werden. Allerdings sind solche Änderungen an der Datenstruktur der darauf laufenden Geschäftsanwendungen nicht immer möglich und stoßen oftmals auf wenig Akzeptanz.

Letztlich führt die Methode der Selektierung der geänderten Daten weiterhin zu einer nicht unerheblichen zusätzlichen Belastung der Quellsysteme und damit zu einer Beeinträchtigung der Arbeit. Sie bietet zwar bereits deutlich aktuellere Daten, eignet sich aber auch nicht für die Synchronisation in nahezu Echtzeit.

ELT statt ETL zusammen mit dedizierter Change Data Capture Lösung

Eine dedizierte CDC-Lösung befreit die ETL-Prozesse vom Extrahieren der Daten. Statt die geänderten Daten selektieren zu müssen, werden die Änderungen mit kaum merkbarer Last auf den Quellsystemen permanent in nahezu Echtzeit erfasst und zum Data Warehouse übertragen. Diese Methode eliminiert die Notwendigkeit von Zeitfenstern für Batchläufe; stattdessen sorgt sie kontinuierlich für sekundenaktuelle Daten, ohne dabei die tägliche Arbeit mit den Applikationen merkbar zu beeinträchtigen.

Wie die folgende Abbildung zeigt, werden in dieser Konfiguration transaktionale Daten in nahezu Echtzeit von den diversen Produktionssystemen in den Staging-Bereich des Data Warehouses repliziert. Von dort lädt die ELT-Software die Daten nach Ausführung notwendiger Transformationen in die endgültigen Tabellen. Auf diese Weise greift die ELT-Software niemals

auf die Produktionssysteme direkt zu. Die Transformationen können daher so häufig in so kurzen Intervallen wie gewünscht ausgeführt werden, ohne die Quellsysteme zu beeinträchtigen.

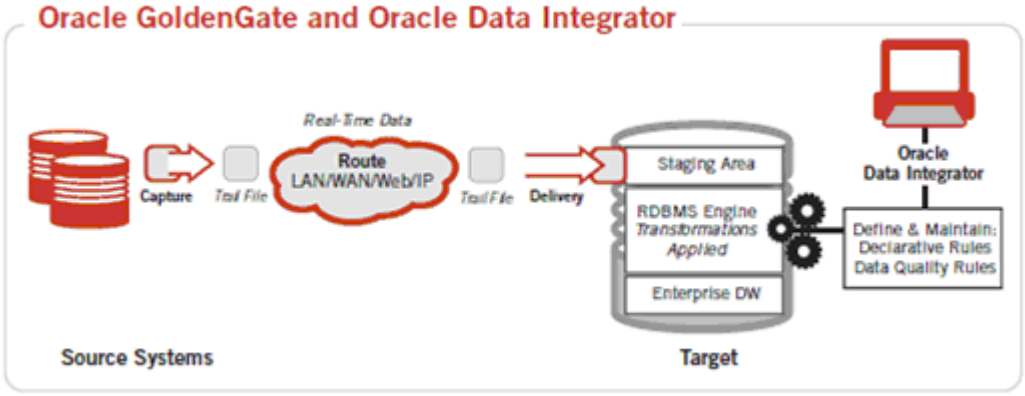

Abb. 1: Architektur einer gemeinsamen Nutzung einer CDC- und ELT-Lösung

Da die Daten unter Erhalt ihrer transaktionalen Integrität repliziert werden, können die Staging-Tabellen nach einer unerwarteten Unterbrechung bzw. nach einem Systemausfall schnell und einfach wieder auf den letzten konsistenten Stand, d.h. auf den Punkt nach der letzten vollständig übertragenen Transaktion, zurückgesetzt werden und die Replizierung an derselben Stelle fortgesetzt werden.

Eine derartige transaktionale CDC-Lösung eignet sich am besten im Zusammenspiel mit einem ELT-Werkzeug, welches die Transformationen innerhalb der Zieldatenbank ausführt, wodurch eine Bewegung der Daten außerhalb der Datenbank entfällt.

Als ein zusätzlicher Vorteil können die Staging-Tabellen parallel genutzt werden als eine Live-Kopie der operationalen Daten, um z.B. operationales Reporting hier statt auf den primären OLTP-Systemen auszuführen.

Change Data Capture mit Oracle GoldenGate

Oracle GoldenGate gestattet die Erfassung, Weiterleitung, Transformation und Übergabe von Datenänderungen zwischen diversen Applikationen und Systemumgebungen. Die Software nutzt eine lose gekoppelte Architektur, um hohe Volumina von Datenänderungen zwischen heterogenen Datenbanken in Sekundenbruchteilen zu bewegen, unter Bewahrung der Transaktionsintegrität.

Wie im nachfolgenden Diagramm gezeigt, besteht die Oracle GoldenGate Architektur aus den zwei eigenständigen Komponenten Capture und Delivery, die durch die Trail Files lose verbunden sind, so daß beide ihre Aufgaben unabhängig von der anderen ausführen können, um eine rasante, nahtlose Datenreplikation zu erreichen.

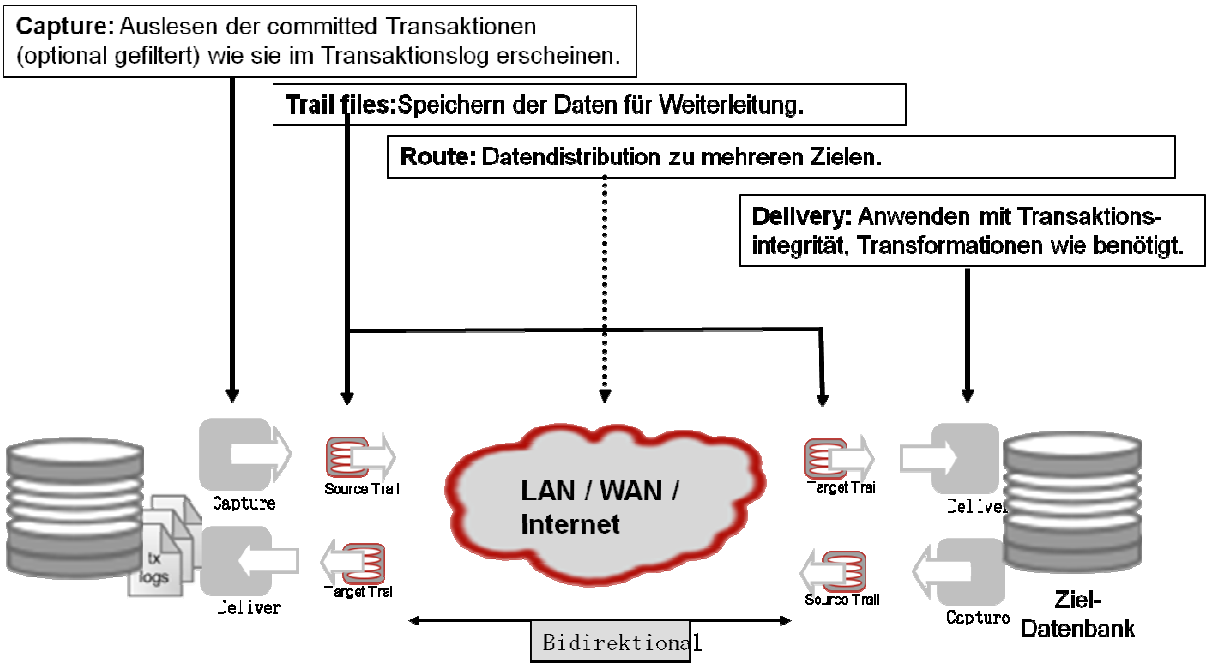

Abb. 2: Oracle GoldenGate Architektur

Oracle GoldenGate Capture

Die Oracle GoldenGate Capture Komponente befindet sich eng neben der Quelldatenbank und überwacht deren Transaktionslog (Oracle: Redo log) auf neue insert, update und delete Operationen. Jede neue Operation wird umgehend erfasst und gegebenenfalls mit einem konfigurierbaren Filter verglichen. Capture ist verfügbar für eine Vielfalt diverser Datenbanken, darunter Oracle Database, Microsoft SQL Server, IBM DB2 (auch Mainframe), Sybase, Enscribe, SQL/MP, SQL/MX und Teradata.

Die Oracle GoldenGate Capture-Komponente erfasst ausschließlich abgeschlossene (committed) Transaktionen, keine Zwischenaktivitäten oder zurückgerollte Transaktionen. Auf diese Weise wird nicht nur die zu transportierende Datenmenge reduziert sondern auch potenziellen Dateninkonsistenzen vorgebeugt.

Oracle GoldenGate Trail Files

Trail Files, ein integraler Bestandteil des Oracle GoldenGate eigenen Warteschlangenmechanismus, speichern die Datenänderungen in einem transportablen, plattformunabhängigen, universalen Datenformat. Trail Files werden idealerweise auf Quell- und Zielserver angelegt, existieren aber außerhalb der Datenbanken, um Heterogenität zu gewährleisten, die Zuverlässigkeit zu erhöhen und den Datenverlust zu minimieren. Diese Architektur reduziert eine etwaige Beeinträchtigung des Quellsystems, da keine zusätzlichen Tabellen oder viele Abfragen der Datenbank benötigt werden. Die Capture-Komponente liest und schreibt die erfassten Änderungen immer nur einmal, unabhängig davon, ob die Daten aus den Trail Files zu einem oder mehreren Zielen weitergeleitet werden sollen.

Oracle GoldenGate Delivery

Die Oracle GoldenGate Delivery-Komponente liest die Daten aus dem jüngsten Trail File und übergibt sie der Zieldatenbank unter Verwendung des nativen SQL für das gewünschte RDBMS. Jede Transaktion wird in derselben Reihenfolge und innerhalb derselben Transaktionsgrenzen, wie sie auf der Quelle entstanden, an das Ziel abgeliefert. Dabei verwendet die Delivery-Komponente gewisse Optimierungstechniken für die Datenübergabe, z.B. kann durch Gruppierung der Transaktionen die Commit Rate reduziert werden. Auch die Delivery-Komponente ist verfügbar für eine Vielzahl diverser Datenbanken. Zusätzlich können die Datenänderungen als Flat Files geschrieben werden oder in verschiedenen Formaten, z.B. XML, zu Enterprise Messaging Systemen publiziert werden.

Operationale Data Warehouse-Lösungen mit Oracle GoldenGate und Oracle Data Integrator

Oracle Data Integrator eignet sich ideal im Zusammespiel mit Oracle GoldenGate als hoch performante Plattform für operationale Business Intelligence-Lösungen. Gemeinsam eliminieren sie den traditionell von anderen ETL-Lösungen erforderlichen separaten ETL-Server und implementieren eine einfachere und effizientere Architektur für Real-Time Enterprise Data Warehouse-Lösungen bei verringerten Gesamtkosten (TCO).

Die Kombination von Oracle Data Integrator und Oracle GoldenGate bietet diverse Vorteile:

Immer aktuelle Daten ohne Wartezeit
In dieser Konfiguration werden transaktionale Daten in nahezu Echtzeit von den diversen Produktionssystemen in den Staging-Bereich des Data Warehouses repliziert, d.h. die Business Intelligence-Lösung hat permanent Zugriff auf den neuesten Stand der Informationen. Es ist *kein* separater Server erforderlich. Da die oben erläuterte CDC-Technologie keine merkbare Last auf die Quellsysteme legt, wird die tägliche Arbeit nicht beeinträchtigt.

Schnellere Datentransformationen
Anstatt von einem separaten konventionellen ETL-Transformationsserver abhängig zu sein, generiert Oracle Data Integrator's ELT-Architektur nativen Code für die verwendete Zieldatenbank, z.B. SQL- oder bulk loader-Skripte. Dieser Ansatz nutzt die Leistungsfähigkeit der Zieldatenbank und optimiert so die Performanz und Skalierbarkeit der Lösung bei gleichzeitiger Senkung der Gesamtkosten. Die Ausführung der Datentransformationen direkt in der Datenbank, welche auch die Zieltabellen enthält, reduziert die Netzwerklast und gewährleistet die höchstmögliche Transformationsgeschwindigkeit.

Hohe IT-Flexibilität durch Unterstützung heterogener Umgebungen
Oracle Data Integrator unterstützt alle führenden Data Warehouse-Plattformen inklusive Oracle Database und Exadata, Teradata, Netezza und IBM DB2. Dies wird komplettiert durch die Oracle GoldenGate-Architektur, welche Quell- und Zielsysteme voneinander entkoppelt und auf diese Weise eine große Vielfalt heterogener Umgebungen aus verschiedenen Datenbanken, Betriebssystemen und Hardwareplattformen unterstützt. Administratoren können jederzeit schnell und einfach neue oder andere Datenbanken als Quelle oder Ziel zur bestehenden Konfigration hinzufügen.

Deklaratives Design für erhöhte Produktivität
Oracle Data Integrator's deklaratives Design verkürzt die Implementierungszeit. Designer spezifizieren, was mit den Daten gemacht werden soll, und die Software generiert die Schritte, wie die Aufgabe umgesetzt wird. Sowohl Entwickler als auch Analysten können die Regeln eines Integrationsprozesses spezifizieren. Die Software generiert dann automatisch die Datenflüsse und die korrekten Instruktionen für die verschiedenen Quell- und Zielsysteme. Mit einem deklarativen Design werden Anzahl und Komplexität der Schritte reduziert und somit die Implementierungszeit verkürzt. Dies unterstützt besonders auch Mitarbeiter, die keine IT-Profis sind, bei der Definition ihrer gewünschten Integrationsprozesse und Datenformate.

Hohe Zuverlässigkeit
Beim Konzept der kontinuierlichen Datenreplizierung von den Quellen zum Data Warehouse spielt die Robustheit der Lösung gegen Unterbrechungen oder Ausfälle an der Quelle, dem Ziel oder dem Netzwerk eine besondere Rolle. Oracle GoldenGate ist für hohe Ausfallsicherheit konzipiert worden und stellt sicher, daß keine Daten verloren gehen oder beschädigt werden. Wie oben beschrieben sind Capture und Delivery lose über die Trail Files gekoppelt, welche auf Quelle und Ziel liegen können. Sobald die Systeme nach einem Ausfall wieder gestartet sind, werden die zuletzt erfassten Datenänderungen umgehend zum Ziel übertragen.

Zusammenfassung

Begnügen Sie sich in Ihrem DWH noch mit den Daten von gestern? Die Erweiterung des DWH um aktuelle, nahezu Echtzeitdaten als Grundlage für operationale Business Intelligence ist keine Hexerei. Mit der Kombination von Oracle GoldenGate und Oracle Data Integrator steht eine leistungsfähige und dabei einfach zu bedienende Plattform für die Implementierung von Operational BI-Lösungen zur Verfügung und wird bereits erfolgreich für diesen Zweck von diversen Kunden aus aller Welt eingesetzt.

Das grundlegende Dilemma der immer weiter wachsenden Datenmengen bei gleichzeitig schwindenden Zeitfensterrn stellt unmißverständlich klar, daß der allabendliche Batchlauf über kurz oder lang ausgedient hat. Wir sollten Abschied nehmen sowohl von der Datensynchronisation per Batchlauf als auch von separaten ETL-Servern. Beides limitiert die Performanz der Integrationsprozesse und die Zuverlässigkeit der BI-Lösung. Und beides ist nicht geeignet die steigende geschäftliche Forderung nach aktuelleren Daten zu unterstützen.

Mit Hilfe der flexiblen Architektur von Oracle GoldenGate können zertifizierte Lösungen implementiert werden, welche es gestatten, Daten aus diversen Quellen in nahezu Echtzeit mit äußerst geringer Belastung in ein Data Warehosue zu replizieren, wo sie hochfrequent mittels Oracle Data Integrator direkt in der Datenbank transformiert und in den Zieltabellen abgelegt werden. Dabei wird die tägliche Arbeit mit den Applikationen nicht durch Abfragen beeinträchtigt.

Kontaktadresse:

Karsten Stöhr
ORACLE Deutschland B.V. & Co. KG
Kühnehöfe 5
D-22761 Hamburg

Telefon: +49 (0) 40-89091 117
Fax: +49 (0) 40-89091 250
E-Mail karsten.stohr@oracle.com
Internet: www.oracle.com/de/products/middleware/data-integration/index.html