



Zusammenführung heterogener Datenquellen

Datenintegration fast auf Knopfdruck

(erschieden in IT Fokus 3/4-2006, S. 23-25; Autor: Christopher Hackett/Thomas Schumacher)

Die Integration wenig strukturierter oder auch gar nicht strukturierter Datenquellen und Zielsysteme zu einer einzigen Datenbasis ist ein komplexer Vorgang. Um den „Rohstoff“ Information optimal nutzen zu können, setzen viele Unternehmen daher zunehmend auf moderne Datenintegrationsplattformen, die den Austausch von Informationen zwischen den unterschiedlichsten Systemen automatisieren und vordefinierte, wieder verwendbare Schnittstellen an die Stelle aufwendiger Eigenprogrammierung setzen. Oft wird in diesem Zusammenhang der Begriff der Datendrehscheibe gebraucht.

Informationen bestimmen heute in allen Branchen die Geschäftsprozesse, ob es um komplexe Produktionssteuerung in der Automobilindustrie, die Abwicklung von Transaktionen in der Finanzbranche oder die Pflege von Kundenbeziehungen in Großunternehmen allgemein geht. Die zugrunde liegenden Daten liegen dabei meist in den unterschiedlichsten Formaten auf heterogenen Systemen. Eine schnelle und vor allem verlässliche Nutzung dieser Ressourcen in einem einheitlichen Informationskonzept wird durch Strukturen behindert, die historisch gewachsen und nicht

für eine gemeinsame Nutzung vorgesehen sind.

Die klassische Bewirtschaftung von dispositiven Datenbanken (Data Warehouses) hilft hier nicht viel weiter. Erfüllen sie doch schon lange nicht mehr die heutigen Anforderungen, die sich in Begriffen wie Konsolidierung, Synchronisation und Migration von Daten ausdrücken. Waren Integrationsprojekte lange Zeit auf Abteilungen oder einzelne Unternehmen beschränkt, finden sie mittlerweile zunehmend auch über Unternehmensgrenzen hinweg. Insbesondere Outsourcing-Projekte lassen sich nur dann effektiv betreiben, wenn ein reibungsloser Datenaustausch zwischen allen beteiligten Partner gewährleistet ist. Informatica hat diesen Trend frühzeitig erkannt und mit seiner Datenintegrationsplattform PowerCenter die derzeit technologisch führende Integrationssoftware im Angebot. Organisationen können damit Daten nutzen, die in einer Vielzahl unterschiedlicher Transaktionsanwendungen und -systeme, in Data Warehouses, Operational Data Stores (ODS), Flatfiles und Legacy-Systemen gespeichert sind.

Um zwischen unterschiedlichen Systemen ausgetauscht

werden zu können, müssen Daten oftmals neu formatiert oder umstrukturiert werden. Während früher jede einzelne Schnittstelle separat programmiert und gewartet werden musste, schrumpft dieser Vorgang mit PowerCenter zu einem simplen „drag and drop“-Vorgang. Durch ein Visualisierungstool mit einer reichhaltigen Bibliothek können sogar Gruppen von Transformationen in Mapping-Objekten kombiniert werden, unabhängig von Datenquelle und -ziel. Auf diese Weise ist die größtmögliche Flexibilität und der höchste Grad an Wiederverwendbarkeit gewährleistet, so dass auch international agierende Entwicklerteams produktiv zusammenarbeiten können. Dabei kann es sich um Standardapplikationen wie zum Beispiel SAP, Oracle oder Peoplesoft handeln, um verschiedene relationale Datenbankformate und Flatfiles aller Art, um Standards wie IBM MQSeries, TIBCO Rendezvous, webMethods, ODBC oder XML, oder um hierarchische und multidimensionale Plattformen, wie verschiedene Mainframe-Systeme, C-ISAM oder Adabas.

Durch eine Kooperation mit dem Softwareanbieter Itemfield ist Informatica in der

Lage, auch unstrukturierte oder teilstrukturierte Daten umzuwandeln. Schätzungen zufolge machen unstrukturierte Daten – etwa E-Mails, Word-Dokumente, Präsentationen, Excel-Sheets oder PDFs – bis zu 90% der in einem Unternehmen gespeicherten Informationen aus. Teilstrukturierte Daten mit Industriestandard-Formaten – etwa EDIFACT im Handel oder SWIFT im Finanzsektor – sind ebenfalls weit verbreitet und dienen in zunehmendem Maße als Basis für automatisierte Geschäftsumgebungen.

Datenintegration in Batch, CDC und Realtime

Weil die akzeptablen Zeitfenster für das Laden schrumpfen, muss eine moderne Datenintegrationslösung für eine optimale „Operational Performance“ vor allem in den Bereichen Datendurchsatz, Skalierbarkeit und Flexibilität Maßstäbe setzen. In vielen Fällen reicht es aus, Daten über Nacht komplett im Batch-Verfahren zu übermitteln, etwa wenn Buchungssätze eines Tages an die Zentrale geschickt werden. Wenn aber Datenbank-Updates überspielt werden sollen, ist diese Vorgehensweise, bei der ja zum Großteil unveränderte Daten übertragen werden, mit enormem und überflüssigem Transfer verbunden. Für diesen Zweck wurde PowerCenter mit der so genannten Changed-Data-Capture-Funktionalität (CDC) ausgestattet, einer inkrementellen Datenintegration, bei der die Integrationslogik zu vordefinierten Zeitpunkten lediglich veränderte Datensätze verarbeitet. Die Steigerung dazu stellt die Realtime-Funktionalität dar: in Transaktionsumgebungen kann dadurch sofort reagiert werden,

wenn ein wichtiger Kunde beispielsweise seinen Orderumfang unvermittelt erhöht oder reduziert. PowerCenter setzt dieses Ereignis unverzüglich um und kann automatisch eine Anfrage an das Bestandssystem oder eine Benachrichtigung an die Produktion durchführen. Außerdem kann automatisch eine optimale Antwort, etwa in Form einer Alarmmeldung an das zuständige Salesteam über E-Mail veranlasst werden. Die Leistungsdimensionen, in denen sich moderne Datenintegrationsplattformen heute bewegen, veranschaulicht ein Benchmark, den PowerCenter im Sommer 2005 auf einem Sun-Mehrprozessorsystem erreicht hat: Ein Datenvolumen von einem Terabyte wurde dabei in nur 36,4 Minuten bewegt. Das entsprach umgerechnet 22,9 Gigabyte pro Stunde und Prozessor. Bei dem Ladevorgang von einem Terabyte in ein Oracle-basierendes Data Warehouse mit gleichzeitiger Durchführung komplexer Transformationen kam PowerCenter mit weniger als 45 Minuten aus.

Neue Techniken für optimale Performance

Um solche Leistungen zu erreichen, setzt Informatica intelligente Algorithmen ein. Mit Pipeline Partitioning werden große, sequenzielle Verarbeitungsabläufe in kleinere „Happen“ aufgeteilt, die parallel bearbeitet werden können. Beim Data Partitioning werden bei wachsenden Datenvolumina Untergruppen gebildet. So können etwa Kundendaten geographisch aufgeteilt und dann ebenfalls parallel bearbeitet werden. Auch beim Cache Partitioning, zum Beispiel für Lookups, wird Parallelbearbeitung eingesetzt. Durch die

Partitionierung des Cache können so eine Vielzahl der Suchvorgänge in parallelen Datenströmen bearbeitet werden. PowerCenter bietet eine Vielzahl an Möglichkeiten, Daten bei der Transformation zu aggregieren, prüfen oder zu ranken. Für weitere Performanceoptimierungen wurde die Methode des „Data Smart Parallelism“ implementiert. Sie sorgt für die Sortierung von Aggregationen zur Sicherstellung korrekter Ergebnisse, beschleunigt die Performance durch dynamisch aufgeteilte Caches und beseitigt die Notwendigkeit, Plattenressourcen über Partitionen hinweg durch integriertes Deadlock/Retry zu managen. Darüber hinaus erhöht man mit Block Based Data Processing den Durchsatz: PowerCenter wählt dabei automatisch die optimale Größe des benötigten Speicherblocks und ist in der Lage, Daten auch in großen Blöcken abzuarbeiten. Für nahezu grenzenlose Skalierbarkeit nutzt PowerCenter, wie viele Datenbanken und Application-Server, Threads für parallele Verarbeitungen. Mit der 64Bit-Version von PowerCenter ist der Gesamtspeicher für Thread Based Parallel Processing praktisch unbegrenzt.

Hochverfügbarkeit – Service 24/7

Um den wachsenden Anforderungen gerecht werden zu können, müssen Systeme heute rund um die Uhr verfügbar sein. Dieser Anforderung trägt Informatica mit seiner neuesten Version PowerCenter 8 Rechnung. Mit dieser Version steht eine Datenintegrationsplattform zur Verfügung, die auch den Anforderungen eines automatisierten 24x7 Betriebs gewachsen ist. Mit dem dort umgesetzten Domänen- und

Servicekonzept können Primary- und Backup-Services einfach eingerichtet und betrieben werden. Konkret heißt dies, fällt ein Primary-Service aus, wird automatisch der Backup-Service die anstehenden Aufgaben übernehmen und so die Verarbeitung sicherstellen. Damit ist gewährleistet, dass benötigte Daten auch für unternehmenskritische Bereiche „in Time“ zur Verfügung gestellt werden können.

Metadaten – die DNA der Datenintegration

Im Zentrum der Datenintegration stehen immer Metadaten, die Informationen in einen Zusammenhang stellen. Die Zahl „10“ etwa hat für sich genommen keine Bedeutung, solange sie nicht mit weiteren Informationsteilen oder Metadaten verknüpft wird, wie z.B. „10 ist der Preis einer bestimmten Aktie zu einem bestimmten Zeitpunkt“. Man unterscheidet dabei zwischen passiven und aktiven Metadaten. Passive Metadaten benötigen Interventionen durch Menschen, um eine direkte Verbindung zwischen der Beschreibung und dem Beschriebenen herzustellen. Architekturen mit passiver Nutzung von Metadaten können also bestenfalls eine Momentaufnahme von Datenbeziehungen liefern.

Eine Besonderheit der Informatica-Welt ist die ausschließliche Verwendung von aktiven Metadaten, die zu jedem Zeitpunkt so aktuell wie möglich sind. Realisiert wird das Konzept der „Active Metadata“ durch einen Repository-Server in Verbindung mit einem Metadaten-Repository. Im Zusammenspiel dieser beiden Datenspeicher wird eine durchgängige Datenintegrität in jeder Phase eines Projekts sichergestellt. Active Metadata

sind quasi die DNA der Datenintegration und lassen sich unbegrenzt weiter verwenden, was Wartungsaufwand und Fehler auf ein Minimum reduziert. Auf Active Metadata baut auch die SOA-Architektur von Informatica auf, die als Universal Data Services (UDS) bezeichnet wird. Mit der Fähigkeit zur Verwaltung von Batchverarbeitung und zur Optimierung der Echtzeitintegration, kann die UDS-Architektur sowohl transaktionale Daten aus EAI Message Queues erfassen als auch Informationen aus einem Data Warehouse oder ODS über EAI-Warteschlangen an andere Unternehmensanwendungen und -systeme übermitteln. In Verbindung mit der objektorientierten Entwicklungsumgebung von Informatica können Designer Transformationslogik einmal erzeugen, debuggen und testen und dann verschiedenen Teams, Projekten und Betriebsumgebungen bereitstellen. Dieser „einmal definieren, überall nutzen“-Ansatz ermöglicht wiederholbare Integrationsprozesse. Fehleranfälliges Handcoding gehört damit der Vergangenheit an.

Integration wird zur Softwarefunktion

In einer anlässlich der „Winter 2005 Conference“ durchgeführten Studie des Data Warehousing Instituts gaben 55 Prozent der Befragten an, dass sie in den nächsten 18 Monaten eine Zunahme des Datenvolumens in ihrem Unternehmen um mehr als 25 Prozent erwarten. Immerhin 67 Prozent nannten Leistung und Skalierbarkeit als Top-Prioritäten bei der Auswahl einer Datenintegrationsplattform – noch vor Faktoren wie dem Preis, einer einheitlichen Plattform, einer Lösung ohne Code und einfa-

cher Bedienung. Diese Zahlen machen deutlich: wer das Potenzial seiner ständig wachsenden Daten realisieren will, muss seinen Fokus auf moderne Datenintegration richten. PowerCenter ist in diesem Umfeld derzeit die weltweit führende Lösung, die von über 2.300 Kunden eingesetzt wird. Mit der Entscheidung pro Informatica können Unternehmen langfristig den optimalen Nutzen aus ihren Daten ziehen – mit einer Lösung, die absolut hersteller- und plattformunabhängig ist. Der Austausch von Daten zwischen heterogenen Systemen wird dadurch von einer komplexen Herausforderung zu einer einfachen Softwarefunktion.