Einsatz von Data Federation für den schnellen Aufbau eines BI-Systems

Dr. Nick Golovin
Koch Media GmbH / Eligent Data GbR
München / Leipzig

Schlüsselworte:

Data Federation, Data Warehouse, Business Intelligence, Real-time Data, Operatives Reporting

Einleitung

In diesem Beitrag werden Erfahrungen aus einem Business Intelligence Projekt bei einem internationalen mittelständischen Medienunternehmen geschildert, wo Data Federation für den schnellen Aufbau eines BI-Systems in einer stark heterogenen internationalen IT-Landschaft zum Einsatz kam. Die Highlights des Ansatzes sind die schnellen ersten Ergebnisse dank Data Federation kombiniert mit steigender Performanz ermöglicht durch den graduellen Aufbau von einem Data Warehouse im Hintergrund.

Einführung

Die zahlreichen Probleme der klassischen Data Warehousing Lösungen sind bereits seit langem gut bekannt. Dazu gehören unter anderem die langwierige Konzipierungsphase, aufwändige und teure technische Implementierung, schlechte Anpassbarkeit an die Änderungen in den Geschäftsprozessen.
Diese Herausforderungen können von den großen Konzernen dank der deren reichlichen Ressourcen in der Regel bewältigt werden, Mittelstandsfirmen haben aber ihre Probleme mit den Kosten und Aufwänden, welche für den fachgerechten Aufbau eines Data Warehouse benötigt werden. An dieser Stelle wird häufig gespart und auf Halblösungen ausgewichen – mit negativen Konsequenzen für die Zufriedenheit der internen Kunden.
Auch die Art der Projekteinführung ist bei einem klassischen Data Warehouse wenig mittelstandsfreundlich. Die häufig angewendete Wasserfall-Methode mit ausführlicher Spezifikationsphase ist für die Großkonzerne, wo die Geschäftsprozesse relativ formalisiert sind und die Key Users gewöhnt sind, die Anforderungen im Detail niederzuschreiben, besser geeignet. Für Mittelstand, wo eher die „hands-on“ Mentalität herrscht, ist die lückenlose Formalisierung der Anforderungen schwierig. Auch die Geschäftsprozesse sind beim Mittelstand viel flexibler, deshalb fällt es den Key Users schwierig, bereits heute die Anforderungen von übermorgen präzise zu definieren. Die iterative Projekteinführung, wo jeder Schritt von den Key Users kontrolliert und ggf. angepasst werden kann, wäre aus unserer Sicht eine geeignetere Methode insbesondere im Mittelstandsbereich.
Trotz dieser Schwierigkeiten wächst der Bedarf nach Datenintegrationslösungen für Business Intelligence auch im Mittelstand, es werden immer neue Anforderungen gestellt. So werden z.B. immer häufiger Echtzeit-Daten aus den Quellsystemen für Operatives Reporting gebraucht (Operatives BI, Real-Time BI). Für tiefere Analysen auf der Ebene der Fachabteilungen brauchen die Endbenutzer auch Kontrolle über die Extraktion von Daten aus den Quellsystemen (Self-Service BI, In-Memory Analyse). Diese Anforderungen werden von den klassischen Data Warehouse Lösungen in der Regel nicht erfüllt.

Als eine mögliche Antwort auf die Problematik der Datenintegration und Datenanalyse für den Mittelstand haben sich etablieren seit einiger Zeit die sogenannten Self-Service-BI-Tools und In-Memory-Analyse Tools, welche Analyse der großen Datenmengen direkt beim Endanwender ermöglichen. Diese Tools können zwar den unmittelbaren Bedarf der Fachabteilungen nach der flexiblen Datenanalyse stillen, können aber aus der strategischen Sichtweise als Rückschritt angesehen werden. Solche Lösungen weichen von dem Single-Source-of-Truth-Konzept ab und führen in gewisser Hinsicht zurück in die Zeiten vor Business Intelligence, wo jeder Miterbeiter seine Datenanalysen in den eigenen unzähligen Excel-Dateien betrieben hatte.
Aus diesen Gründen werden neue strategisch tragfähige Ansätze für die Integration der Daten aus unterschiedlichen Datenquellen erfordert, insbesondere im Hinblick auf die speziellen Anforderungen des Mittelstands bezüglich Kosten, Flexibilität. Ein solcher Ansatz wurde bei der Firma Koch Media GmbH zur Einführung des Business Intelligence Projekts erfolgreich angewendet. Dabei ist eine Business Intelligence Lösung entstanden, welche kostengünstig, flexibel und dennoch performant und strategisch tragfähig ist. Über die Details und Besonderheiten dieses Projektes berichten wir in den weiteren Abschnitten.

Kundenszenario und verwendete Technologien

Die Firma Koch Media GmbH wurde im 1994 gegründet und spezialisiert auf Distribution, Publishing und Producing von Entertainment-Produkten (Computerspiele, Videospiele, Online-Spiele, DVD/Blueray Filme, Consumer-Software). Unter anderem befindet sich das internationale Spiele-Label „Deep Silver“ im Besitz der Firma Koch Media GmbH. Die Firma zählt in 2011 ca. 350 Mitarbeiter weltweit (Europa und USA) und macht über €300 Mio Umsatz jährlich.
Die Entwicklung der Firma Koch Media GmbH in der Zeit vor der Einführung des Business Intelligence wurde durch starkes Wachstum und Expansion ins Ausland gekennzeichnet. Dabei wurden einige ausländische Unternehmen mit vorhandener IT-Infrastruktur übernommen. Des Weiteren waren einige ausländischen Niederlassungen historisch bedingt im IT Bereich relativ eigenständig. Dadurch ist eine stark heterogene Systemlandschaft mit 4 großen ERP-Systemen und zahlreichen anderen kleineren Systemen und Datenbanken entstanden. Es wurde notwendig, der Konzern-Geschäftsführung und der Leitung der zentralen Fachabteilungen wie Controlling, Finanzbuchhaltung und internationales Marketing einen Überblick und Einblick in die Daten in allen diesen Systemen zu verschaffen. Unter diesen Systemen und Datenquellen befinden sich neben der 4 ERP-Systeme noch mehrere Internet-Shops, Produktdatenbanken, Projektmanagement-Systeme, Groupware-Systeme, externe Partner-Daten und Marktforschungsdaten z.B. von GfK, Massive Multiplayer Online-Spiele-Plattformen etc. Die hinter diesen Systemen liegende Datenbanksysteme umfassen Oracle, Informix, IBM DB2/AS400, MS SQL, MySQL, PostgreSQL, MS Access. Die Anzahl der Datensätze in den größten Tabellen liegt bei mehreren zehn Millionen Datensätze. Die gesamten Datenbestände betragen mehrere zehn Terabyte. Die Dokumentation der Daten und Strukturen war leider in großen Teilen lückenhaft.
Vor der Einführung von Business Intelligence wurden unterschiedlichste Werkzeuge verwendet, um Daten aus den Systemen zu extrahieren und Berichte zu generieren (RPG, Access, Excel, SQL Server, Lotus, AS400-Data Transfer, ASP, JSP/Jasper). Einen Überblick über die Ausgangslage gibt die Abbildung 1.

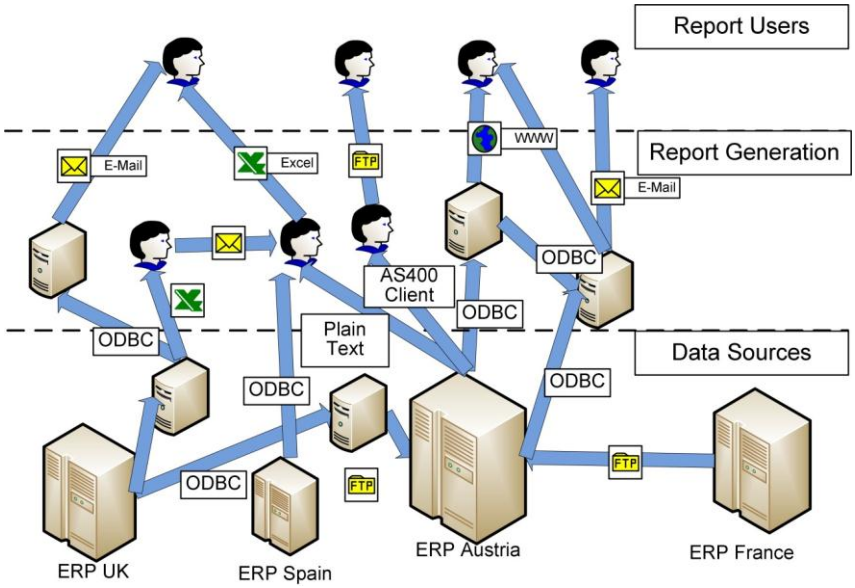

Abbildung 1. Reporting-Situation beim Kunden vor der Einführung von Business Intelligence

Solches dezentrale Berichtswesen war natürlich schwer zu unterstützen, es gab viele redundante Auswertungen, die Diskrepanzen in den Auswertungen waren sehr schwer zu verfolgen. Es konnte auch keine Rechteverwaltung eingesetzt werden.
Durch die Verwendung von unterschiedlichen Technologien war der Aufbau von Know-How sehr schwierig.
Bei den Anforderungen für das zu implementierende Business Intelligence Architektur wurden zwei große Aufgabenbereiche definiert:

- operatives Berichtswesen für jedes Land: Berichte sowohl mit Echtzeit-Daten (Verkaufsdaten, Finanzdaten, Produktdaten) als auch mit historischen Daten.
- analytisches Berichtswesen, vor allem länderübergreifend: Überblick über alle Koch-Firmen, interactive Analyse mit Drill-Down, ad-hoc Berichte, Analyse von historischen Daten. Die Endkunden für solches analytische Berichtswesen sind Top-management, Controlling, Internationales Marketing, internationales Producing/Publishing (Produktionsplanung)

Bei der Analyse der Anforderungen wurde schnell klar, dass ein klassisches Data Warehouse keine passende Lösung bietet. Ausschlaggebend waren dabei die folgenden Kriterien:

- In vielen Bereichen wurden Echtzeitdaten aus mehreren Systemen benötigt, z.T. in Kombination mit historischen Daten.
- Einige der sehr wichtigen Berichte aus mehreren Datenquellen waren von der Geschäftsleitung schnellstmöglich benötigt, das Abwarten bis das Data Warehouse fertig ist stand nicht zur Diskussion
- Es war abzusehen, dass die Geschäftsprozesse sich in der absehbaren Zukunft stark ändern werden, so dass keine zuverlässige Konzipierung möglich war.

Im Hinblick auf diese speziellen Anforderungen haben wir uns dazu entschlossen, die Data Federation Technologie einzusetzen. Der Einsatz von Data Federation lässt alle an das Data Federation Tool angeschlossene Datenquellen wie eine einzige große relationale Datenquelle aussehen. Die Datenquellen können über ODBC, JDBC oder WebServices angeschlossen werden, das Data Federation Tool selbst kann ebenfalls über JDBC, ODBC oder WebService angesprochen werden. Das Data Federation Tool wandelt die empfangene SQL-Abfrage in mehrere Teil-Abfragen, welche er an die angeschlossenen Datenquellen schickt, bekommt die Ergebnisse zurück und führt die Teilergebnisse zum endgültigen Ergebnis zusammen. Das passiert on-the-fly, ohne dass der Entwickler sich um die Speicherung der Zwischenergebnisse kümmern muss. Ein Data Federation

Tool ist in der Lage, auch große Datenmengen mit mehreren Millionen Datensätzen problemlos zu bewältigen, es versucht jedoch, einen möglichst großen Anteil der Abfragenbearbeitung an die Datenquellen auszulagern, so dass die Menge der von den Datenquellen zu übertragenden Daten im Vorfeld minimiert wird. Beispiele von Data Federation Tools sind Oracle Data Service Integrator, Business Objects Data Services von SAP und InfoSphere Federation Server von IBM.
Durch Einsatz von Data Federation können die Datenquellen sehr schnell angebunden sein und die ersten Ergebnisse viel schneller präsentiert werden – bereits am ersten Projekttag -- als es mit einem klassischen Data Warehouse möglich wäre. Allerdings gibt es auch Nachteile, welche gegen den alleinigen Einsatz von Data Federation bei den Business Intelligence Projekten sprechen. So werden dabei z.B. keine historischen Daten gespeichert. Da Data Federation direkt auf die Datenquellen geht, werden diese unter Umständen stärker belastet. Auch die Geschwindigkeit der Ausführung der Abfragen, welche über das Data Federation Tool an die Datenquellen geschickt ist langsamer als bei einem Data Warehouse. Das liegt jedoch nicht an dem Data Federation Tool selbst, sondern daran, dass die dahinterliegenden Datenquellen evtl. eine höhere Netzwerklatenz haben und für die Ausführung von analytischen Abfragen nicht ausgelegt sind.
Aus diesen Gründen haben wir uns für eine hybride Architektur bei der Durchführung des Projekts entschieden. Die Architektur des BI-Systems ist in der Abbildung 2 präsentiert.

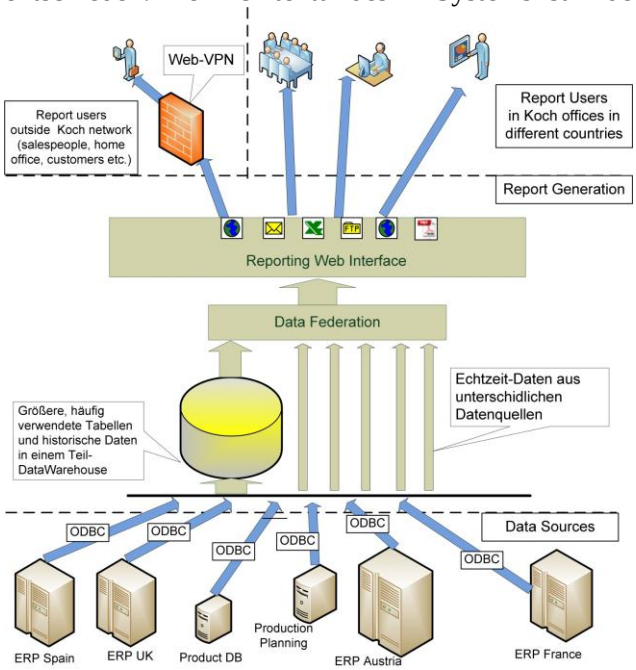

Abbildung 2. Business Intelligence mit Data Federation und (Teil)Data Warehouse

Wie in der Abbildung 2 dargestellt, werden bei der eingesetzten Architektur alle analytische Abfragen grundsätzlich durch den Data Federation Layer bearbeitet. Als eine der Datenquellen, welche an das Data Federation Layer angeschlossen sind, fungiert ein klassisches Data Warehouse, welches mithilfe von ETL-Tools erstellt und befüllt wird. Wichtig dabei ist jedoch, dass nicht alle für die Analyse benötigten Daten in diesem Data Warehouse gespeichert werden müssen. Nur in folgenden Fällen macht die Unterbringung von Tabellen in einem Data Warehouse Sinn:

- Die Tabelle ist besonders groß oder kommt von einem System welches über langsames Netzwerk angebunden ist
- Es ist eine spezielle Verarbeitung/Bereinigung von Daten gewünscht, welche mit SQL nicht möglich ist oder besonders langsam ist
- Es muss eine Historie für die Daten aufbewahrt werden, welche aus den Quellsystemen regelmäßig gelöscht werden

In unserem Fall mussten wir insgesamt nur ca 10% der Gesamtanzahl der Tabellen, welche für die Analysen benötigt werden, auch tatsächlich in unserem Data Warehouse physisch abspeichern. Deshalb sprechen wir hier eher von einem Teil-Data-Warehouse. Durch den Wegfall der Notwendigkeit, alle Daten im Data Warehouse zu haben, schätzen wir eine 90-prozentige Ersparnis bei der Entwicklung von ETL-Prozessen und Datenstrukturen im Data Warehouse erzielt zu haben.
Eine wichtige Eigenschaft der vorgestellten Architektur ist, dass die Nutzerberichte nicht wissen müssen, ob bestimmte Daten direkt aus dem Quellsystem oder aus dem Teil-Data-Warehouse kommen. Dadurch wird eine iterative Entwicklung des BI-Systems in engem Kontakt mit den Endnutzern ermöglicht. Diese Architektur erlaubt uns, die ersten Berichtprototype ausschließlich basierend auf Data Federation zu entwickeln. Ein solcher Prototyp kann extrem schnell entstehen und schnell in Rücksprache mit dem Endbenutzer nachgebessert und finalisiert werden. Erst wenn der Bericht fertig ist, kann man dessen Performanz analysieren und ggf. entsprechende Datenstrukturen im Teil-Data-Warehouse vorbereiten. Danach kann das Data Federation Layer so umkonfiguriert werden, dass die relevanten Abfragen aus dem Teil-Data-Warehouse bedient werden. Dabei muss der Bericht und dessen SQL-Abfragen nicht geändert werden.

Im Laufe des Projektes bei der Firma Koch Media GmbH ist uns dank unserer Architektur folgendes gelungen:
- Alle Daten im Unternehmen zu integrieren und für die Berichte zur Verfügung zu stellen, ohne sie komplett in einem Data Warehouse abzuspeichern
- Bei der Projekteinführung eine sehr signifikante Kostenersparnis zu erreichen, dank dem verringerten Aufwand für den Aufbau des Data Warehouse
- Der Geschäftsleitung die notwendigsten Berichte schnellstmöglich zu präsentieren, was mit einem klassischen Data Warehouse unmögliche wäre
- Eine iterative Vorgehensweise bei der Projekteinführung anzuwenden. Dabei waren wir imstande, den Endnutzern die jeweils nächsten Versionen der Berichtprototype in kurzen Zeitabständen präsentieren zu können, so dass der Fortschritt sichtbar wird und das Feedback besser berücksichtigt werden kann. Dadurch wird die Zufriedenheit der internen Kunden aufrechterhalten.

Ursprung und Ausblick
Die bei dem beschriebenen Projekt angewendeten Ideen und Ansätze basieren auf den Forschungsergebnissen, welche am Lehrstuhl für Datenbanken, Universität Leipzig entstanden sind (Lehrstuhlinhaber Herr Prof. Dr. E. Rahm). Im Rahmen des Projekts Eligent Data an der Universität Leipzig wird diese Forschung fortgeführt. Dabei entsteht ein Datenintegrationsprodukt, welches den hier beschriebenen Ansatz auf die nächste Stufe bringt. Dank der vorgelagerten Data-Federation-Schicht ist das neue Produkt imstande, die für die Berichte verwendeten Daten und deren Nutzungsmuster zu sammeln und zu analysieren. Basierend auf dieser Analyse kann unsere Datenintegrationssoftware die Struktur von dem Teil-Data-Warehouse vorschlagen und sie per Knopfdruck samt der nötigen ETL-Prozesse automatisch implementiert, was zur weiteren Senkung der Kosten und des Aufwands für den Aufbau von BI-Projekten führt. Das Projekt Eligent Data an der Universität Leipzig sucht momentan nach Kooperationspartner und potenziellen Kunden für unser innovatives Datenintegrationsprodukt.

Zusammenfassung
In diesem Artikel wurde ein innovativer Ansatz für den schnellen und kostengünstigen Aufbau eines BI-Systems unter Verwendung von Data Federation beleuchtet. Der geschilderte Ansatz stammt aus den Forschungsergebnissen des Lehrstuhls für Datenbanken an der Universität Leipzig und wurde bei einem international aufgestellten mittelständischen Medienunternehmen zur Einführung eines Business Intelligence Projekts erfolgreich angewendet. Die resultierende Business Intelligence Lösung ist durch geringe Kosten, kurze Einführungszeit, hohe Flexibilität und Performanz sowie den

problemlosen Zugriff auf Echtzeitdaten aus den Quellsystemen gekennzeichnet. Dieser Ansatz wird im Rahmen des Projekts Eligent Data an der Universität Leipzig erweitert und weitergeführt, so dass eine nahtlose Integration von Data Warehousing und Data Federation ermöglicht wird.

Kontaktadresse:

Dr. Nick Golovin
Koch Media GmbH / Eligent Data GbR

E-Mail n.golovin@kochmedia.com; nick.golovin@eligent.com
Internet: www.kochmedia.com; www.eligent.com