DATA WAREHOUSING

Wagenradstrategie.

Immer mehr Geschäftsprozesse hängen von agilen Data Warehouses ab. Je mehr Anwendungen und Nutzer es gibt, desto komplizierter wird die Informationsversorgung. Doch es gibt Auswege.

› Autor: Hilmar Buchta*

Eine fast euphorische Stimmung begleitet häufig die Anfangsphase eines Data Warehouses. Denn diese ist in der Regel geprägt von geringer Komplexität, hoher Agilität und großer Dynamik bei allen Beteiligten.

Wenige Anwender, kaum gegenläufige Anforderungen und eine konsequente Ausrichtung auf schnelle Erfolge und Analyseergebnisse erlauben es, neue Anfragen eher in Stunden als in Tagen oder Wochen umzusetzen.

Auf der Welle dieses Erfolgs indes steigt die Komplexität schnell. Mehr Prozesse werden jetzt durch das Data Warehouse abgebildet. Auch die Anzahl der direkten oder indirekten Anwender wächst rasant. So kann aus der anfänglichen Agilität schnell ein Chaos entstehen. Änderungen am Datenmodell müssen immer mehr Abhängigkeiten beachten und können damit nicht «mal eben» umgesetzt werden. Auch die Entwicklungs- und Bereitstellungsprozesse müssen sich entsprechend mitentwickeln.

Schnell werden auch geschäftsrelevante Systeme, wie zum Beispiel die Abrechnung oder der Webshop, mit Daten aus dem Data Warehouse versorgt. Weitere operative Systeme folgen und werden nun auch mit Daten versorgt. Die Stabilität wird immer wichtiger. Verglichen mit der dynamischen Anfangsphase müssen Änderungen nun sehr viel genauer konzeptionell geprüft und getestet werden und benötigen dadurch in der Umsetzung viel mehr Zeit. Die Unternehmens-IT einerseits, die in der Regel für den Betrieb des Data Warehouse zuständig ist, und die Fachbereiche andererseits, die sich weiterhin die Dynamik der Anfangszeit wünschen, driften immer mehr auseinander.

Was tun? Agilität und Stabilität, beide Faktoren sind wichtig, wenn ein Data Warehouse nachhaltigen Nutzen bringen soll. Die Fachabteilungen benötigen dringend die Agilität, um bei sich ändernden Marktbedingungen schnell verlässliche Informationen zu erhalten. Ein Beispiel aus dem Mobilfunk ist der Trend von starren Tarifen hin zu Optionsmodellen. War der Kunde früher über einen vergleichsweise langen Zeitraum gut einzuschätzen, müssen nun völlig andere Prognoseverfahren für den Deckungsbeitrag und die Kundenbindung eingesetzt werden.

Dies zieht oft deutliche Änderungen am Datenmodell nach sich. In der Regel kommt es bei vielen dieser fachlichen Anforderungen dabei weniger auf eine hundertprozentige Exaktheit der Daten an. Auch eine Ausfallsicherheit steht meist nicht an der obersten Stelle der Prioritätenliste. Viel wichtiger ist es, dass die Anforderung schnell umgesetzt werden kann und die Auswertungsergebnisse kurzfristig zur Verfügung stehen.

Ganz anders sieht es aus der Perspektive geschäftsrelevanter Prozesse aus: Zum Beispiel benötigen Abrechnungssysteme sehr exakte Daten. Eine Verschiebung des Rechnungslaufs zieht Kosten nach sich, Pannen beim Rechnungslauf können sehr negative Effekte auf das Image des Unternehmens haben. Operative Systeme, die ›

durch das Warehouse versorgt werden, benötigen zuverlässige, aktuelle Daten.

Als Lösung bietet sich hier eine Trennung beider Aspekte in unterschiedliche Schichten an: Die Basis bildet das zentrale Data Warehouse, häufig als Core oder Hub bezeichnet, mit hoher Verfügbarkeit, qua-

Karussellkabinen: Ein Sinnbild für agile und stabile Data Warehouses.

litätsgesicherten Entwicklungs- und Bereitstellungsprozessen und hoher Datenqualität. Auf diesem können dann unterschiedliche fachliche Auswertungsdatenbanken, meist als Data Marts oder Spokes bezeichnet, aufsetzen und bei Bedarf sehr agil die Daten des Cores transformieren, zusammensetzen und somit die fachlich erforderlichen Auswertungen sehr schnell bereitstellen. Operative und geschäftsrelevante Anwendungen werden in diesem Szenario meist aus dem Core heraus mit Daten versorgt. Diese Architektur wird als Hub & Spoke bezeichnet (in Anlehnung an Nabe und Speichen im Wagenrad). Für den Core ist es hierbei wichtig, dass die Daten aus den Vorsystemen vollständig, verlustfrei und korrekt erhalten bleiben. Meistens laufen dort nur wenige Transformationen. So entstehen zum Beispiel bei der Erzeugung von Snapshot-Daten oder bei der konkreten Auflösung historischer Bezüge schnell Verluste, die später Auswertungen in den Data Marts stark einschränken können. Da die angeschlossenen Data Marts nicht aktueller sein können als das Core, ist meist eine hohe Aktualität der Daten, bis hin zur Echtzeit erforderlich. Für die Modellierung im Core bieten sich unterschiedliche Verfahren an. Die von Inmon vorgeschlagene Modellierung in einem

normalisierten Modell gilt als sehr flexibel, allerdings auch als sehr aufwendig. Eine reine Speicherung der Quelldaten (persistierte Staging-Schicht) erfordert eventuell sehr viele Transformationen in den Data Marts. Da letztere untereinander möglichst unabhängig sein sollen, sind diese Übergange dann auch häufig noch redundant erforderlich. Neuere Modellierungsideen, wie zum Beispiel das Data Vault, können hier Abhilfe schaffen, aber letztlich muss im Einzelfall ein guter Kompromiss gefunden werden: soviel Vorverarbeitung wie möglich, ohne dass dabei Informationen verloren gehen. Zur Speicherung von Daten, deren Bedeutung für die Analyse noch nicht klar ist und die gegebenenfalls später genutzt werden sollen, bietet sich zum Beispiel auch ein Hadoop Cluster an.

Ein solcher Core lässt sich in der Regel leicht erweitern. Die Änderungen sind jedoch aufwendiger und müssen ausführlich getestet werden. Denn alle Data Marts verlassen sich auf den Core, auf dessen Datenqualität und Aktualität. Organisatorisch wird er meist von der IT betrieben. Die Datenbereitstellung für die Data Marts und operativen Systeme kann über individuelle Service Level Agreements (SLA) sichergestellt werden.

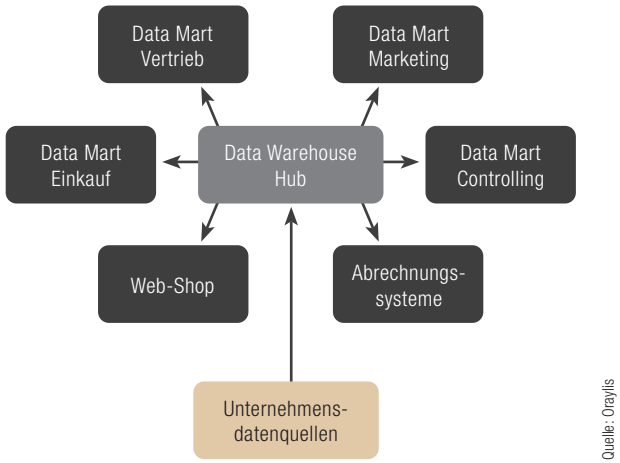
Bei den Data Marts können in der Praxis unterschiedliche Ansätze und Technologien zum Einsatz kommen. Die Verantwortung muss heute nicht mehr bei der IT liegen, sondern kann von der jeweiligen Fachabteilung übernommen werden. Hier finden sich häufig leicht bedienbare Werkzeuge aus den Bereichen Visual- und Self-Service-BI, Berichtserstellung und -verteilung, OLAP (Online Analytical Processing) und Data Mining.

Aber wie kann auf der Seite der Data Marts eine hohe Flexibilität erreicht werden, wenn schon die Datengrundlage im Core einen eher starren Charakter hat? Tatsächlich ist es die Transformation der Daten aus dem Core, die die Flexibilität ausmacht. In der Praxis finden sich Beispiele, bei denen der Core über Jahre hinweg quasi unverändert geblieben ist und alleine in der Kombination von Daten aus dem Core ständig wechselnde Anforderungen dynamisch umgesetzt werden konnten. Häufig werden in den Data Marts auch weitere Daten, zum Beispiel Marktforschungsdaten, verwendet, um die Daten aus dem Core anzureichern.

Die konkrete Trennung von Core und Data Marts kann dabei in der Praxis sehr unterschiedlich aussehen. Von einer einfachen Trennung im Datenmodell, zum Beispiel durch unterschiedliche Zonen, Datenbanken und Schemata, bis hin zu einer strikten Trennung über unterschiedliche Systeme und organisatorische Verantwortungen. Im Einzelfall sollte die Lösung gewählt werden, die den Anforderungen am besten Rechnung trägt und dabei keinen großen Verwaltungsaufwand verursacht. Denn wie bei jedem Schichtenmodell kann auch hier eine übertriebene Trennung zu einem deutlich höheren Mehraufwand führen.

Fazit: Agilität und Stabilität müssen sich im Data Warehouse nicht widersprechen. Der Hub & Spoke - Ansatz, bei dem stabile und flexible Teile getrennt werden, bietet ein gutes und in der Praxis bewährtes Modell, um beiden Aspekten gerecht zu werden. Allerdings müssen bei der Ausgestaltung die Anforderungen im Mittelpunkt stehen. Eine Trennung mit hohen organisatorischen Hürden ist hier wenig erfolgversprechend. ■

Die Hub & Spoke-Architektur



Quelle: Oraylis

*** Der Autor**

Hilmar Buchta, Geschäftsführer und Projektmanager beim Düsseldorfer Data Warehouse- und Business Intelligence-Experten Oraylis GmbH, verfügt über mehr als 15 Jahre Erfahrung in BI- und IT-Projekten. Darüber hinaus zählt er zu den weltweit renommiertesten Experten für die Datenbanksprache MDX.