Data Warehouse Architekturtrends

Dr. Bodo Hüsemann
Informationsfabrik GmbH
Münster

Schlüsselworte

Architektur, Business Intelligence, Data Warehouse, Realtime Data Warehouse, Operational BI, Selfservice BI,

Einleitung

Klassische Data Warehouse (DWH) Systeme sind in der Funktionalität auf die Integration von operativen Daten, der Historisierung zeitlicher Datenänderungen und auf die Bereitstellung von Data Marts mit fachlicher Themenstellung für Business Intelligence (BI) Anwendungen spezialisiert. In der Regel werden hierbei die operativen Systeme für dispositive Anfragen gekapselt, um den Betrieb nicht zu beeinträchtigen. Gleichzeitig fungiert das DWH als „Single Point of Truth" und zentralisiert Datenschnittstellen zu analytischen Systemen im Sinne einer „Hub and Spoke"-Architektur.

Vor dem Hintergrund aktueller fachlicher Anforderungen ergeben sich allerdings einige Herausforderungen und Verbesserungspotentiale:

Zeitbezug: Die Batch-Verarbeitung von ETL-Prozessen in klassischen DWH führt zu Latenzen hinsichtlich der Datenaktualität. Für heutige Anwendungen sind allerdings tagesaktuelle Analyse- und Berichtsprozesse oft nicht mehr hinreichend. Beispiele sind hier Performance Dashboards, die zeitkritische Unternehmenskennzahlen in hochaggregierter Form nicht nur zur strategischen Entscheidungsunterstützung, sondern auch für teil-automatisierte Reaktionsprozesse bereitstellen.

Stammdaten-Integration: In klassischen DWH-Szenarien wird oft die mangelnde Qualität der Stammdaten im DWH-Umfeld bereinigt. Doppelterfassungen und Konsistenzprobleme können auf DWH-Ebene nur noch eingeschränkt korrigiert werden und sind regelmäßig in der fehlenden Koordination/Abstimmung der operativen Systeme begründet. Die Erkennung und Behandlung dieser Integrationsprobleme beim Beladen des DWH erfolgt zu spät und kann zudem innerhalb der Isolation einzelner operativer Systeme nicht immer wirksam erfolgen.
Die Erfahrung zeigt, dass nur die Zentralisierung der Stammdaten und eine stärkere Kopplung an alle dispositiven und operativen Systeme eine gute Qualität der genutzten Daten gewährleisten kann.

BI-Einsatzgebiete: BI-Anwendungen für das DWH werden unlängst nicht nur zur strategischen Analyse eingesetzt. Es eröffnen sich zunehmend Anforderungen zur Nutzung von BI-Kennzahlen im operativen Umfeld (bspw. Deckungsbeiträge auf Kundenebene im CRM-System), wodurch nicht nur Daten aus dem DWH, sondern auch die für BI typische Analysefunktionalität in diesen Systemen zur Verfügung gestellt werden sollen.

BI-Flexibilität: Im Bereich ad-hoc Reporting/Analyse wird von BI Anwendern mehr Flexibilität gefordert. Beispielsweise müssen vorhandene Kennzahlen in Verbindung zu benutzerdefinierten externen Datenquellen analysiert werden. Für benutzerdefinierte Berichte ist außerdem der Anfrageraum häufig nicht hinreichend abgesichert, so dass beispielsweise bestimmte Aggregationspfade (vom BI-Tool im Hintergrund automatisiert gewählt) zu falschen

Anfrageergebnissen führen. Außerdem können bei IT-Änderungen an Data Marts keine Regressionstests/Anpassungen für benutzerdefinierte Berichte durchgeführt werden. In der Konsequenz sind aktuelle BI- und DWH-Systeme den Anforderungen für benutzerdefinierte Berichte (Self Service Reporting) nicht gewachsen.

In dem vorliegenden Papier werden die gezeigten Problemfelder in den vier Themenbereichen

- Realtime Data Warehouse
- Master Data Management
- Operational BI
- Self Service BI

diskutiert und abschließend zu einer Gesamtarchitektur vereint.

Realtime Data Warehouse

In einem Operational Data Warehouse (alternativ Operational Data Store (ODS)) wird neben dem analytischen Data Warehouse eine Integrationsschicht für zeitkritische Auswertungen etabliert (vgl. Abb. 1).

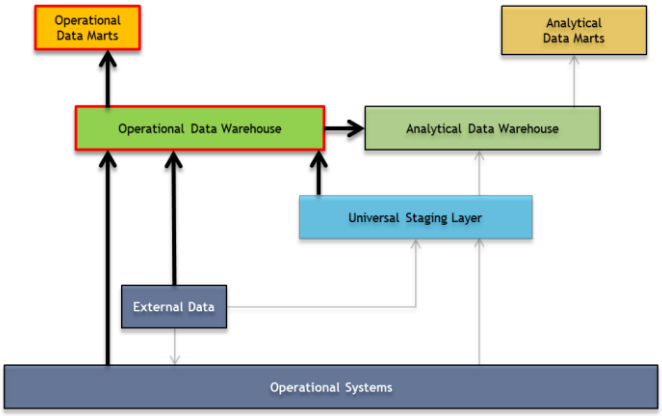

Abb. 1: Operational Data Warehouse

Dieser Datenpool wird durch neue Integrationstechnologien (z.B. Enterprise Application Integration (EAI) oder Enterprise Information Integration (EII) möglichst zeitnah gewartet, wobei auf eine Historisierung in der Regel verzichtet werden kann.

Tab. 1: Operational Data Warehouse

	ETL (Extract, Transform, Load)	EAI (Enterprise Application Integration)	EII (Enterprise Information Integration)
Informationsaustausch	Mengenorientiert	Nachrichtenbasiert	Anfrageorientiert
Zeitbezug	Scheduling	Ereignisgesteuert	Anfragegesteuert
Auslöser	Pull	Push	Pull
Steuerung	Batch Job	Workflow	Service Request
Latenz	Tag	Neartime	Realtime

In Tab. 1 werden die Kerneigenschaften aktueller Integrationsverfahren gegenübergestellt. In der Praxis werden steigenden Latenzanforderungen häufig zunächst durch die bestehenden ETL-Prozesse mit einer Erhöhung der Scheduling-Frequenz begegnet. Die Effektivität dieses Ansatzes lässt sich durch fortgeschrittene Verfahren zur Änderungserkennung weiter erhöhen, bspw. können sog. Change Data Capture (CDC)-Systeme die klassische Batch-Verarbeitung in einen Datenstrom überführen. Falls entsprechende Schnittstellen existieren, bieten EAI-Systeme in der Regel bessere Unterstützung zur Abbildung von nachrichten-basierten Geschäftstransaktionen und erleichtern dadurch die konsistente Integration in das Data Warehouse erheblich. Bei EII-Integrationssystemen erfolgt die Integration virtuell zur Anfragezeit, wodurch insbesondere der mengenmäßige Integrationsaufwand besser auf den Informationsbedarf eingeschränkt werden kann.

Master Data Management

Stammdatenintegration sollte in einer modernen DWH-Architektur in einem spezialisierten Master Data Management (MDM)-System stattfinden. Dort wird durch anwendungsübergreifende Konsistenzprüfungen ein zentrales Datenqualitätsmanagement für Stammdaten realisiert, das den gesamten Daten-Lebenszyklus überwacht.

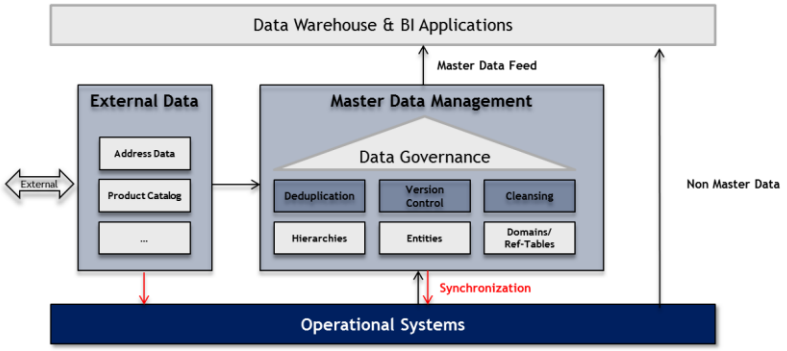

Abb. 2: Architekturalternativen Master Data Management

In der Architektur erfordert dies entsprechende Schnittstellen zu den operativen Systemen und den nachgelagerten DWH-Komponenten.

In Abb. 2 werden verschiedene Achitekturalternativen zur Realisierung von MDM-Systemen abgebildet. Die Varianten unterscheiden sich hinsichtlich der Verteilung der Stammdatenerfassung (dezentraler oder zentraler Master) und dem Datenaustausch mit den operationalen Systemen.

Die virtuelle Integration in einem föderierten MDM eignet sich insbesondere bei existierenden Applikationsschnittstellen der operativen Systeme und möglichst geringen gemeinsamen Datenbeständen. Dem gegenüber wird in einem integrierenden MDM ein materialisierter Stammdatenpool verwaltet, jedoch Datenänderungen weiterhin vom operationalen System getätigt und per Synchronisationsmechanismen weitergeleitet. Bei einem operationalen MDM wird schließlich die Stammdatenpflege und Datenhaltung auf das MDM übertragen, so dass von operativen Systen nur noch lesend zugegriffen wird.

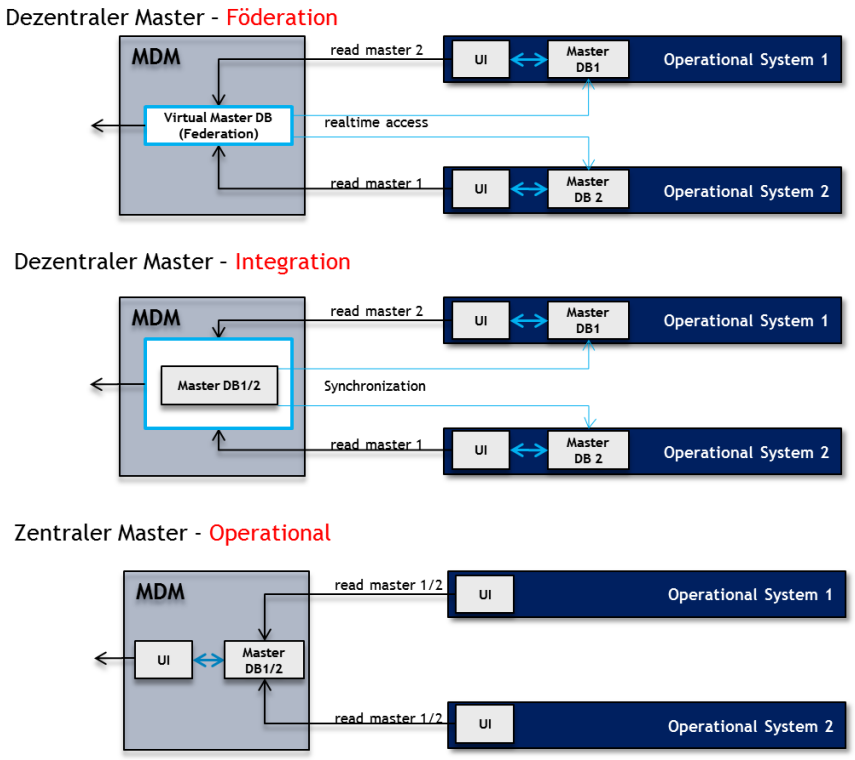

Abb. 3: Architekturalternativen Master Data Management

Operational BI

Beim Operational BI (OBI) werden im Sinne einer Feedback-Schleife aus den verschiedenen Data Warehouse-Komponenten Daten zurück an die operativen Systeme geliefert. Alternativ werden aus den operativen Systemen Anfragen auf die DWH-Daten ausgeführt oder über Applikationsschnittstellen BI-Komponenten (z.B. Portlets) in operative Systeme eingehängt.

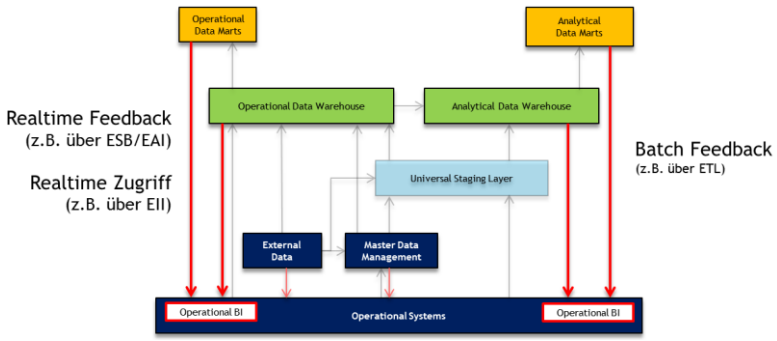

Abb. 4: Operational Business Intelligence

Im operativen Kontext können Daten sowohl in der Tagesendverarbeitung übertragen (vgl. Abb. 4 Batch Feedback) oder über EAI/EII-Schnittstellen im Tagesverlauf aktualisiert oder auf sie zugegriffen werden. Alternativ werden in aktuellen Architekturen operative Applikationsplattformen in Form eines Enterprise Service Bus (ESB) zusammengeführt und über ereignisorientierte Schnittstellen mit dem Data Warehouse bidirektional verbunden. Für den Anwender ist in der operativen Anwendung der Übergang von der Kernanwendung zu BI-Komponenten und Daten fließend.

Self Service Reporting
Im klassischen DWH werden analytische Datenpools für technisch versierte Anwender zur Erstellung benutzerdefinierter Berichte und Analysen freigegeben. Die meist komplexen Anfrageräume (z.B. hochdimensionale Cubes mit >20 Dimensionen) führen bei Endanwendern häufig zur Überforderung und entsprechend schlechter Endergebnisqualität. Zusätzlich wird regelmäßig ein kurzfristiger ad-hoc-Zugriff auf benutzerdefinierte (evtl. externe) Datenquellen gefordert, der durch toolseitige oder infrastrukturelle Einschränkungen oft nicht realisiert werden kann. In diesen Fällen werden durch die Fachabteilungen, oft auch über längere Zeiträume hinweg, Schattenkopien von (Teil-)bereichen der analytischen/operativen Systeme erstellt, die außerhalb des IT-Betriebs liegen und somit mit betrieblichen Risiken behaftet sind.

Für den BI Poweruser mit großen Flexibilitätsanforderungen wird daher ein personalisierter Importbereich mit entsprechenden Tools für den regulierten Zugriff/Import auch von semistrukturierten (bspw. XML) und unstrukturierten Dokumenten (Office Dokumente oder PDF) benötigt (vgl. Abb. 5).

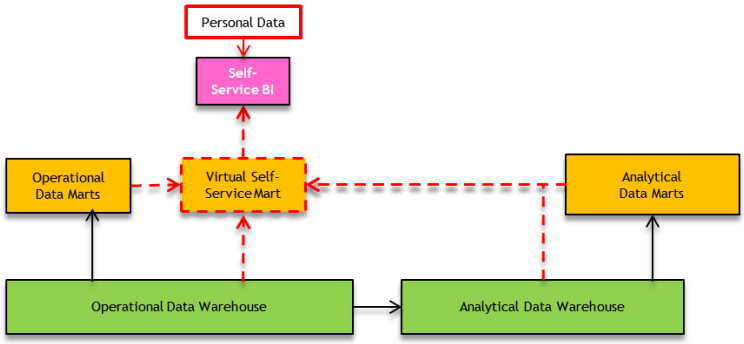

Abb. 5: Self Service Business Intelligence

Zusätzlich werden für Poweruser virtuelle Anfragebereiche geschaffen, die hinsichtlich Inhalt und Anfragemöglichkeiten vereinfacht präsentiert werden und deren Anfragefunktionalität sich auf Endbenutzeranforderungen beschränken. Die virtuellen Strukturen bilden eine konstante, fachliche Sicht auf die vorhandenen Data Marts und DWHs mit dem Ziel, eine weitgehende Entkopplung von dort stattfindenden inhaltlichen und strukturellen Änderungen durchzusetzen.

Gesamtarchitektur
In der folgenden Gesamtarchitektur einer zukunftsorientierten DWH- und BI-Systemlandschaft werden die gezeigten Ansätze zu einem Gesamtbild zusammengefügt. Die aufgeführten Bestandteile und Vielzahl der Schnittstellen sind als „Maximalausprägung“ zu verstehen, wobei im konkreten Unternehmenskontext eine Beschränkung, insbesondere der Schnittstellen, sinnvoll sein kann. Im Rahmen dieser Arbeit wurde nicht explizit auf die beiden unscheinbar grau gefärbten Komponenten Metadata und Data Quality Management eingegangen. Beide Themen sind als Querschnittsthemen zu verstehen und bilden umfangreiche eigene Themenbereiche, die notwendige Voraussetzungen und Basis für aktuelle DWH- und BI-Szenarien darstellen.

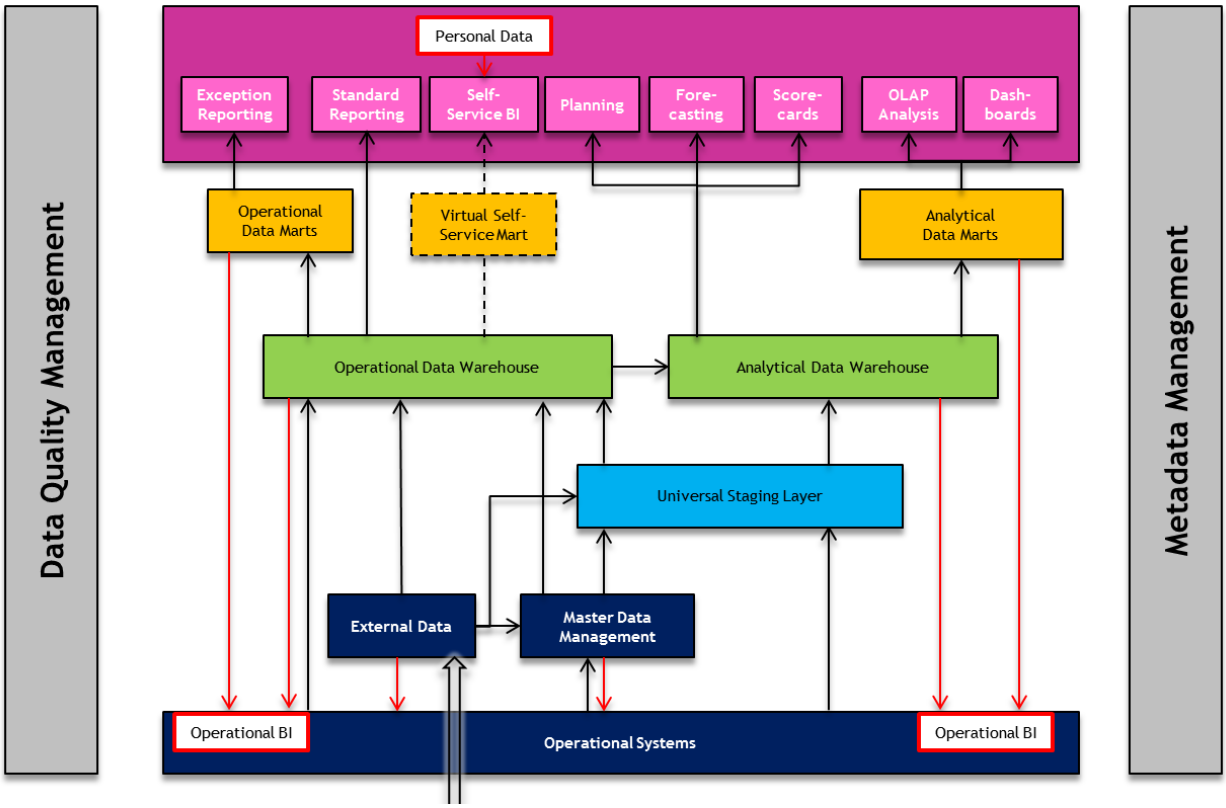

Abb. 6: Gesamtarchitektur eines zukunftsorientierten DWH- und BI-System

Kontaktadresse:

Dr. Bodo Hüsemann
Informationsfabrik GmbH
Scheibenstraße 117
48153 Münster

Telefon: +49 251 91997961
Fax: +49 251 91997971
E-Mail bhuesemann@informationsfabrik.de
Internet: www.informationsfabrik.de