Wartung von Dimensionsdaten in verteilten Data Warehouse-Systemen

Moritz Schmaltz, Barbara Dinter

Institut für Wirtschaftsinformatik
Universität St.Gallen
Müller-Friedberg-Str. 8
CH-9000 St.Gallen
{moritz.schmaltz | barbara.dinter}@unisg.ch

Abstract: In großen Unternehmen finden sich häufig komplexe, organisch gewachsene Data Warehouse-Architekturen. Wenn die Daten in diesen Systemen nicht unter Verwendung einheitlicher Dimensionen gespeichert sind, ergeben sich bei der Nutzung der Daten vielfältige Probleme. Ein Weg zur Optimierung der Data Warehouse-Architektur kann in der Vereinheitlichung der Dimensionen bestehen. In diesem Beitrag zeigen wir die Anforderungen an eine zentralisierte Pflege der Dimensionsdaten in fachlicher, technischer und organisatorischer Hinsicht auf. Entlang dieser Perspektiven leiten wir Handlungsempfehlungen und mögliche Zielarchitekturen für die Umsetzung einer zentralisierten Dimensionswartung ab. Gestützt werden die Aussagen durch Erfahrungen in einem konkreten Praxisprojekt und durch Experteninterviews in Großunternehmen.

1 Einleitung

Das Data Warehouse (DWH) hat sich seit seiner Entstehung in den 90er-Jahren als fester Bestandteil des Applikationsportfolios in Unternehmen etabliert. Als Datenbasis für analytische Systeme ermöglicht es die Aggregation von Daten aus operativen Systemen und unterstützt Dispositions- und Entscheidungsprozesse. Insbesondere in großen Unternehmen besteht jedoch oft das Problem, dass im Zuge des organischen Wachstums von Applikationslandschaften und durch Zusammenschlüsse mit anderen Unternehmen mehrere, von einander unabhängige DWH-Systeme entstanden sind. Die Wartung dieser getrennten Systeme ist sehr aufwändig, zudem ist die Vergleichbarkeit der Daten aus unabhängigen Systemen erschwert. Trotz der offensichtlichen Probleme ist eine Zusammenführung der einzelnen Systeme in einem umfassenden DWH schon aus finanziellen Gründen in der Regel kaum denkbar. Die Vereinheitlichung der Dimensionsdaten über die Systemgrenzen hinweg bietet eine Möglichkeit, die Nachteile einer verteilten DWH-Architektur mit vertretbarem Aufwand zu reduzieren. Die Vergleichbarkeit der Daten aus verschiedenen Systemen nimmt zu und der Erklärungsbedarf für Diskrepanzen sinkt. Zudem verringert sich der Aufwand für die Wartung der Dimensionsdaten.

Dieser Beitrag zeigt basierend auf Projekterfahrungen, Experteninterviews und Literaturbeiträgen die Herausforderungen und möglichen Realisierungsalternativen einer solchen Zentralisierung der Dimensionsdatenpflege auf.

In Abschnitt 2 werden die grundlegenden Architekturalternativen für DWH-Systeme und das Konzept der Dimensionsdaten kurz vorgestellt sowie Probleme bei der Wartung von Dimensionsdaten in verteilten DWH-Architekturen erörtert.

Anforderungen an eine zentralisierte Wartung der Dimensionsdaten in fachlicher, technischer und organisatorischer Hinsicht stellen wir in Abschnitt 3 dar. In Abschnitt 4 leiten wir daraus eine Zielarchitektur und Hinweise für die Umsetzung ab. Der Artikel schließt mit einer Zusammenfassung und einem Ausblick (Abschnitt 5).

2 Grundlagen

2.1 Architekturen für DWH-Systeme

Das Data Warehouse dient der Schaffung einer integrierten Datenbasis für dispositive und analytische Zwecke [MSW03]. Zur Realisierung größerer DWH-Systeme existieren eine Reihe von Architekturalternativen; eine beispielhafte Klassifizierung aus [AW06] ist in Abb. dargestellt. Die Architekturansätze unterscheiden sich demnach u.a. hinsichtlich der Zentralisierung der Datenhaltung. Das eine Ende des Spektrums bildet eine Architektur mit völlig dezentralen, unabhängigen Data Marts, wie sie oft in organisch gewachsenen Systemlandschaften zu finden ist (Typ A, vgl. zur Entstehung solcher Systeme [In02]). Der weit verbreitete Ansatz von [Ki98] sieht dem gegenüber ebenfalls unabhängige Data Marts vor, die jedoch über gemeinsame, einheitliche Dimensionen verfügen (Typ B). Auf der anderen Seite propagiert z.B. Inmon eine Architektur mit einer zentralisierten Datenbasis und daraus gespeisten Data Marts (Typ D, [IIS01]). Weitere Architekturalternativen sind eine zentralisierte Datenbasis ohne separate Data Marts (Typ C) und eine föderierte Lösung, bei der Teile bestehender Datenbanken und Data Warehouses durch Integration der Daten zu einem sog. virtuellen Data Warehouse zusammengefasst werden (Typ E).

Das Verständnis von Data Marts ist in dieser Klassifikation uneinheitlich: Während es sich bei den von Kimball als Data Marts bezeichneten Systemen in der Praxis durchaus um vollständige, themenbezogene DWH-Systeme handeln kann, versteht Inmon unter Data Marts anwendungsspezifische, denormalisierte Analysedatenbanken, die aus dem DWH gespeist werden [In02].

Unter verteilten Data Warehouses verstehen wir im Kontext dieses Beitrags DWH-Architekturen ohne eine zentrale, normalisierte Datenbasis und ohne vereinheitlichte Dimensionen (Typ A). Dieser Architekturtypus bildet die Ausgangslage für den folgenden Beitrag.



Abb. 1: Architekturalternativen für Data Warehouses [AW05]

Empirische Daten zeigen, dass die genannten Architekturen unterschiedlich erfolgreich sind hinsichtlich der Faktoren Informationsqualität (Genauigkeit, Vollständigkeit, Konsistenz), Systemqualität (Skalierbarkeit, Flexibilität, Integration), individueller Nutzen (schnelle Verfügbarkeit von Daten, Analysemöglichkeiten) und Nutzen für die Organisation (Erfüllung der fachlichen Anforderungen, ROI) [AW06]. Die Architekturen mit unabhängigen Data Marts und föderierten Systemen schneiden hier durchgehend am schlechtesten ab. Architekturen mit vollständiger oder teilweiser Zentralisierung sind erfolgreicher, wobei die Architekturen von Inmon und Kimball (Typ B bzw. D) im Wesentlichen gleich erfolgreich sind, sodass keine der Architekturen als überlegen bezeichnet werden kann.

Diese Erkenntnisse stimmen mit den Erfahrungen in dem diesem Artikel zu Grunde liegenden Projekt überein, nach denen eine dezentrale DWH-Architektur insbesondere hinsichtlich Informationsqualität und Kosten problematisch ist (vgl. Abschnitt 0). Im Projekt wurde die Stammdatenverwaltung in den Data Warehouse-Systemen eines Großunternehmens vereinheitlicht, um diese Probleme zu beheben.

Da eine radikale Umstellung der Architektur von unabhängigen Systemen auf eine zentralisierte DWH-Architektur in der Regel sehr aufwändig sein dürfte, bietet sich wie in Abb. hervorgehoben eine Weiterentwicklung hin zu einer teilweise zentralisierten Architektur mit einheitlichen Dimensionen an (Typ B, vgl. Abschnitt 0).

2.2 Dimensionsdaten im Data Warehouse

Der Begriff der Dimensionsdaten wird in diesem Abschnitt kurz erläutert und von den verwandten Konzepten der Stammdaten und der Metadaten abgegrenzt.

Die Daten in Data Warehouses sind gemäß multidimensionalen Datenmodellen gespeichert. In diesem Datenmodell sind die Fakten (Kennzahlen) mit assoziierten Dimensionen gespeichert, die den Kontext der Fakten beschreiben [Ku99]. Die sog. Dimensionsdaten beschreiben die Dimensionen dieser Datenmodelle. Sie bestehen aus Basiselementen, Klassifikationshierarchien und beschreibenden Attributen.

Die Dimensionsdaten enthalten als Elemente auf unterster Ebene die Basiselemente, die die gültigen Ausprägungen der Dimension beschreiben (z.B. Kunden oder Produkte). Darauf aufbauend enthalten sie Klassifikationshierarchien (Verdichtungen), anhand derer Fakten bei der Analyse (meist über mehrere Stufen) aggregiert werden können (z.B. Zuordnungen von Tagen zu Monaten und Jahren). Die Elemente der Klassifikationshierarchien sind dabei selbst auch Dimensionselemente. Die Hierarchien können einfach sein (d.h. es existiert zu jedem Element genau ein Element der übergeordneten Klassifikationsebene) oder parallel (wo innerhalb einer Dimension mehrere Arten der Gruppierung möglich sind) [BG04]. So ist z.B. bei Datumsangaben eine Hierarchie von Tag – Monat – Quartal – Jahr möglich, aber auch eine Hierarchie Tag – Kalenderwoche – Jahr (vgl. Abbildung 2). Zusätzlich können die Dimensionselemente beschreibende Attribute (sog. Dimensionsattribute) haben, die weitere Informationen über die Dimensionselemente liefern [Ku99].

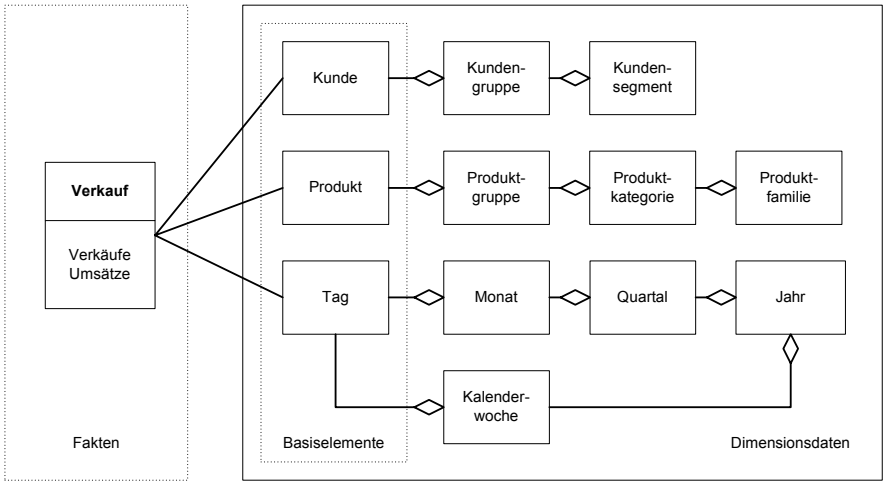


Abb. 2: Beispielhafte Dimensionen und Klassifikationshierarchien

Die Basiselemente der Dimensionen entsprechen oft den Stammdaten aus operativen Systemen. Unter Stammdaten werden in diesem Zusammenhang sich nur selten ändernde Daten zu Geschäftsobjekten wie z.B. Kunden, Lieferanten oder Produkten verstanden, etwa Adressen, Artikel-/ Kundennummern oder Abmessungen [LLG04]. Stammdaten

werden i.d.R. übergreifend von mehreren IT-Systemen verwendet, z.B. werden sie neben dem ERP-System auch von CRM-Systemen benötigt.

Ein verwandter Themenbereich ist das Metadaten-Management. Unter Metadaten im Kontext des Data Warehousing versteht man Daten, die die (Objekt-)Daten im Data Warehouse, die Prozesse, Architekturen und Organisationsstrukturen beschreiben und zusätzliche Informationen über den Kontext dieser Daten liefern, etwa im Hinblick auf Terminologie, Datenstruktur, Datentransformationen oder Datenherkunft [AM97]. Bei den Dimensionsdaten, die im Fokus dieses Beitrags stehen, handelt es sich um einen Grenzbereich von Metadaten und Daten, die Zuordnung ist in Praxis und Forschung umstritten (vgl. z.B.[Au03, 62ff.] bzw. [Ki98, 435ff.]).

2.3 Probleme bei der Wartung von Dimensionsdaten in verteilten DWH-Systemen

Die Dimensionen im Data Warehousing sind nicht statisch, sondern ändern sich teils fortlaufend (z.B. bei Produkten oder Kunden) und erfordern eine regelmäßige Pflege: Neue Elemente müssen eingefügt, nicht mehr gültige Elemente entfernt, geänderte Elemente müssen versioniert und Aggregationen aktualisiert werden.

Die Dimensionsdaten können aus verschiedenen Quellen bezogen werden. Einerseits können Dimensionsdaten von anderen Informationssystemen im Unternehmen stammen (etwa Organisationseinheiten von einem ERP-System). Ebenso ist denkbar, dass Dimensionsdaten, wie beispielsweise Währungskurse, von einem externen Informationssystem bezogen werden. Schließlich müssen Dimensionsdaten möglicherweise manuell im System selbst eingepflegt werden.

Verteilte DWH-Systeme, die organisch gewachsen sind, bilden vielfach kein logisch integriertes System, sondern sind eine Ansammlung von unabhängigen, themenspezifischen Data Warehouses. Für jedes dieser Data Warehouses werden häufig die Dimensionsdaten eigenständig gepflegt, wobei jeweils unterschiedliche Organisationseinheiten für die Pflege zuständig sind und heterogene technische Systeme zum Einsatz kommen (vgl. Abbildung 3). Eine solche Konstellation führt zu einer Reihe von Problemen in fachlicher, technischer und organisatorischer Hinsicht, die eine Vereinheitlichung der Dimensionsdaten und damit eine Entwicklung hin zu dem von Kimball postulierten Architekturmodell (unabhängige Systeme mit einheitlichen Dimensionen) erstrebenswert erscheinen lassen.

Die im Folgenden aufgeführten Probleme wurden im Projekt identifiziert und in Experteninterviews konkretisiert. Hierzu wurden zehn DWH-Experten aus deutschen und schweizerischen Großunternehmen mit umfangreichen DWH-Installationen in semistrukturierten Interviews befragt. Weitergehende Hinweise zur typischen Entstehungsgeschichte von analytischen Systemen und zu den Ursachen der sich daraus ergebenden Probleme finden sich bei [In02].

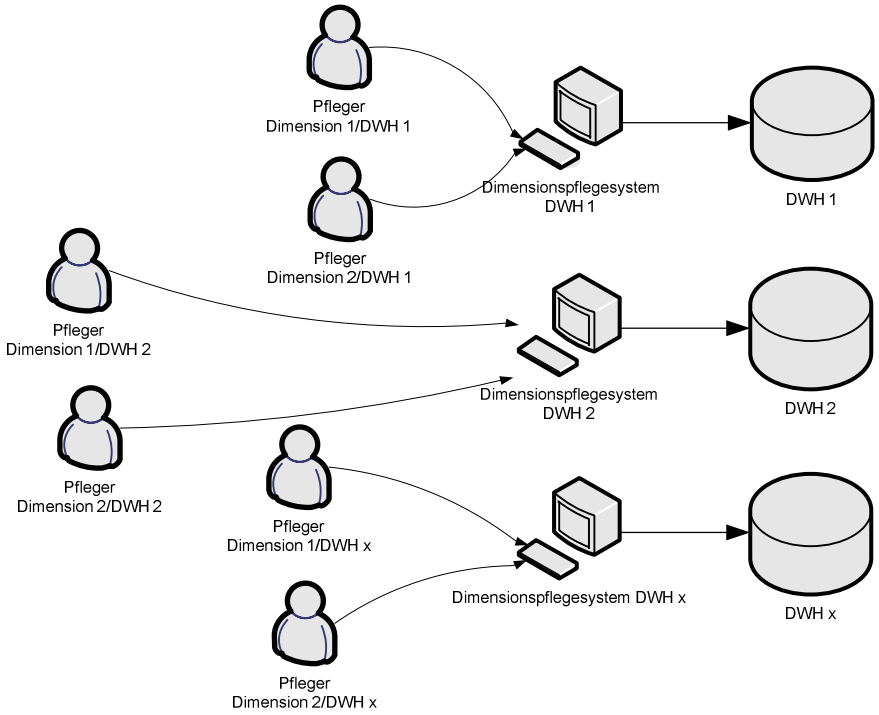


Abb. 3: Ausgangssituation bei der Dimensionswartung

2.3.1 Fachliche Aspekte

In fachlicher Hinsicht ergeben sich bei einer verteilten Wartung der Dimensionsdaten Probleme im Bereich der Vergleichbarkeit und Zusammenführung von Daten sowie bei der Konsistenzsicherung.

Das mit Abstand am häufigsten genannte Problem ist die mangelnde Transparenz von Unterschieden, die Daten aus verschiedenen Data Warehouses bei gleichen Abfrageparametern aufweisen können. Diese Unterschiede können z.B. in Fällen auftreten, in denen die Dimensionsdaten in verschiedenen Systemen auf unterschiedlichen Versionsständen sind oder in Fällen, in denen fachlich äquivalente Dimensionen verschiedene Dimensionselemente und unklar definierte Aggregationen beinhalten. Dies führt zu einem Verlust an Glaubwürdigkeit für alle liefernden Systeme. Ein typischer Problemfall tritt beispielsweise dann auf, wenn zwei Systeme Daten zu äquivalenten Dimensionen aus verschiedenen Quellsystemen extrahieren. Auch wenn die Quellsysteme sich dabei untereinander Daten liefern, ist es wahrscheinlich, dass Diskrepanzen auftreten.

Ein verwandtes fachliches Problem sind Inkonsistenzen beim Zusammenführen von Daten aus verschiedenen Data Warehouses oder externen Daten während des Analyseprozesses. Ungleiche Dimensionen können zu verfälschten und inkonsistenten Ergebnissen führen. Zudem steigt der Aufwand für das Konsolidieren von Daten aus verschiede-

nen Quellen enorm. So muss z.B. sichergestellt werden, dass alle Quellen die selben Aggregationen verwenden. Wenn dies nicht der Fall ist, müssen aufwändige Korrekturen vorgenommen werden.

Zu einem weiteren Problem führen häufige Änderungen an den Dimensionen. Diese haben in einer verteilten Umgebung erhöhte Gefahr von Inkonsistenzen und steigenden Pflegeaufwand zur Folge.

2.3.2 Technische Aspekte

Technische Probleme bestehen in erster Linie bei der Fehlerbehebung, der Pflege von Aggregationen, bei Datenformaten und bei der Historisierung.

Die rückwirkende Änderung von Dimensionsdaten, etwa zur Fehlerkorrektur, ist bei verteilter Wartung problematisch. Hier müssen Korrekturen separat in den verschiedenen Systemen nachgeführt werden, was Mehraufwand zur Folge hat.

Wo nutzerspezifische Aggregationen (Zusatzverdichtungen) verwendet werden, muss ein effizienter Aufbau dieser Aggregationen aus den Basiselementen der Dimensionen gewährleistet sein. Falls hierfür keine Automatismen (z.B. Abbildungsregeln) zur Verfügung stehen, ist neben Mehraufwand die Gefahr von Inkonsistenzen gegeben. Schließlich können technische Formatprobleme auftreten, falls inhaltlich identische Felder z.B. in verschiedenen Zahlenformaten oder unterschiedlichen Längen gespeichert werden.

In technischer Hinsicht stellt die Historisierung eine weitere Herausforderung dar. Es existieren verschiedene Konzepte und technische Möglichkeiten, um die Historie von sich ändernden Daten zu speichern. Wenn diese zwischen den Systemen nicht einheitlich umgesetzt sind, wird die Vergleichbarkeit und Integrierbarkeit von Daten aus den verschiedenen Systemen eingeschränkt. Dieses Problem kann allerdings durch eine vereinheitlichte Pflege der Dimensionen nur teilweise gelöst werden, da es nicht durch die Pflege selbst, sondern durch die Abbildung der Änderungen in den Data Warehouses verursacht wird.

2.3.3 Organisatorische Aspekte

In organisatorischer Hinsicht treten in den Bereichen von Verantwortlichkeiten und Nachvollziehbarkeit sowie bei unterschiedlichen Stichtagen die schwerwiegendsten Probleme auf.

Die Verteilung der Dimensionswartung auf verschiedene Organisationseinheiten führt dazu, dass kein klares Data Ownership-Konzept mit Verantwortlichkeiten für die Aktualität und Qualität der Dimensionsdaten entwickelt werden kann. Auch das Erheben von Audit Trials zur Dokumentation von Änderungen an den Dimensionsdaten wird erschwert. Falls sich die Stichtage für das Laden zwischen den einzelnen Systemen und zwischen den Dimensionen innerhalb der Systeme unterscheiden, hat dies Einfluss auf die Vergleichbarkeit von Daten sowie auf das Bereitstellen von Daten für von den Data Warehouses belieferte Systeme. Besonders kritisch können zeitliche Abhängigkeiten in

den Fällen werden, in denen Abhängigkeiten oder Lieferbeziehungen zwischen den Data Warehouses selbst bestehen.

Ein weiteres Problem stellen die für die redundante Pflege der Dimensionsdaten erforderliche Mehrarbeit und die damit verbundenen Kosten dar. Immer wenn mehrere DWH-Systeme eine äquivalente Dimension haben, müssen Aufgaben wie Vorbereitung, Qualitätskontrolle und Aktualisierung der Dimensionsdaten mehrfach durchgeführt werden. Der Aufwand steigt daher linear mit der Zahl der Systeme. Zudem muss für jedes System ein Pflegeprozess unterhalten werden und es müssen Ressourcen mit der entsprechenden fachlichen und technischen Kompetenz zur Verfügung stehen.

Bei der Nutzung der Data Warehouses erhöht sich der Aufwand für das Erstellen von Reports und Analysen, da Zuständigkeiten und Datenverfügbarkeit geklärt und Daten aus verschiedenen Systemen integriert und abgestimmt werden müssen.

3 Anforderungen an die Zentralisierung der Dimensionswartung

Die Schaffung einer zentralen Dimensionswartung erscheint auf den ersten Blick als ein attraktiver Lösungsansatz für viele der oben vorgestellten Probleme. Allerdings fällt bei genauerer Betrachtung eine Reihe von Aspekten auf, die bei der Zentralisierung zu beachten sind und die eine erhebliche Komplexität mit sich bringen. Hieraus ergeben sich jeweils Anforderungen an das zentrale System zur Wartung der Dimensionsdaten und an die zugehörigen Prozesse. Für den folgenden Abschnitt gehen wir aus von einer Zielarchitektur mit einem einheitlichen System zur Wartung der Dimensionsdaten unter (je Dimension) einheitlicher Verantwortung (vgl. Abbildung 4). Alternativen zu dieser Architektur werden in Abschnitt 4.1 vorgestellt, konkrete technische Umsetzungsmöglichkeiten in Abschnitt 4.3. Ein solches System könnte aus einer Import-Komponente, einem Repository zur Speicherung der Dimensionsdaten, einer Benutzerschnittstelle zur Datenpflege und einer Export-Komponente bestehen.

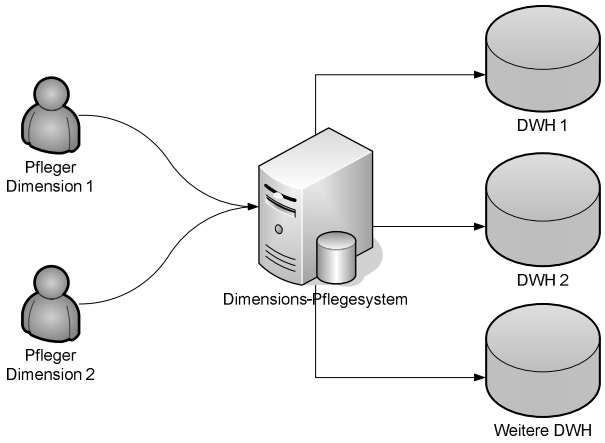


Abb. 4: Eine mögliche Zielarchitektur für die Dimensionspflege

Im Folgenden werden diese Anforderungen an ein Pflegesystem, die im Rahmen von Projekten und Experteninterviews erhoben worden sind, vorgestellt. Die jeweilige Relevanz ist für jedes Projekt individuell zu beurteilen, hier kann nur ein generischer Überblick über mögliche Anforderungen gegeben werden.

Der Übersicht halber sind sie wiederum in fachliche, technische und organisatorische Anforderungen aufgeteilt. Diese Aufteilung erhebt jedoch keinen Absolutheitsanspruch, da viele der Aspekte nicht eindeutig zuzuordnen sind.

3.1 Fachliche Aspekte

Synchronisierung und Aktualität

Um tatsächlich Vergleichbarkeit zwischen den Daten aus verschiedenen Systemen herzustellen, müssen Änderungen in den Dimensionen zeitgleich in den Data Warehouses nachgeführt werden. Das führt dazu, dass das Maximum an Aktualität zum Standard für alle Data Warehouses werden muss, wenn ein Verlust an Aktualität in einzelnen Systemen nicht akzeptabel ist. Wenn Änderungen in einem DWH wöchentlich, in anderen jedoch täglich eingepflegt werden, ist also bei einer zentralen Wartung eine tägliche Pflege der Dimensionen erforderlich, um nicht an Aktualität einzubüßen. Dabei müssen aber nicht in allen Systemen täglich die Dimensionen aktualisiert werden, wenn gewisse Abstriche bei der Vergleichbarkeit akzeptabel sind. Ebenso muss für die Lieferung aktualisierter Dimensionsdaten etwa vor einem monatlichen Verarbeitungslauf der früheste aller systemspezifischen Stichtage zum gemeinsamen Standard werden. Die Prozessdurchlaufzeiten für die Aktualisierung der Dimensionsdaten können sich dadurch erheblich verkürzen.

Aggregationen und Granularität

Je nach thematischer Domäne der Data Warehouses kann sich die erforderliche Granularität der Dimensionsdaten unterscheiden. So können etwa Produktdaten in einem System auf Ebene der Einzelprodukte, in anderen Systemen nur aggregiert auf Ebene der Produktgruppen erforderlich sein. Das Pflegesystem muss daher in der Lage sein, Dimensionsdaten auf unterschiedlichen Aggregationsstufen zur Verfügung zu stellen. Zudem ist denkbar, dass die abnehmenden Data Warehouses unterschiedliche Aggregationen verwenden, weil etwa je nach Anwendungsdomäne unterschiedliche Regionenstrukturen genutzt werden. In der Praxis finden sich zusätzlich oft innerhalb einzelner Systeme auf einer Dimension mehrere nutzerspezifische Aggregationen für bestimmte Anwendungsfälle, die in den Data Warehouses gepflegt werden müssen. Das Pflegetool muss in diesem Fall mehrere Aggregationshierarchien abbilden können und idealerweise die Pflege unterstützen, etwa indem Hierarchiebestandteile zwischen Aggregationen vererbt werden können.

Historisierung und Nachvollziehbarkeit

Die primäre Aufgabe des Pflegesystems ist es, den abnehmenden Systemen jederzeit eine aktuelle Version der Dimensionsdaten zur Verfügung zu stellen. Dennoch ist eine Historisierung der Dimensionsdaten sinnvoll. Dadurch ist es möglich, definierte Versionsstände zu verwalten. Zudem ermöglicht die Historisierung das Nachvollziehen von Änderungen (Stichwort Revisionssicherheit) und bietet im Fehlerfall die Nutzung einer Vorversion. Es ist jedoch nicht Aufgabe des Pflegesystems, die Historisierung der Dimensionen in den Data Warehouses zu übernehmen, da dies tiefe Eingriffe in die Architektur der Data Warehouses mit sich bringen würde.

Qualitätsmanagement

Der Qualität der Daten im zentralen Dimensions-Pflegesystem hat unmittelbaren Einfluss auf die Datenqualität in allen belieferten Data Warehouses. Daher sind Standards und Überwachungsmechanismen für die Qualität der Dimensionen erforderlich. Die Daten sollten automatisiert auf Konsistenz und Korrektheit geprüft und Fehlerfälle ausgesteuert werden können.

3.2 Technische Aspekte

Datenmodelle

Bei mehreren belieferten Data Warehouses ist es wahrscheinlich, dass diese unterschiedliche Dimensionsattribute benötigen. Das Pflegetool muss für alle abnehmenden Systeme nutzbare Dimensionsdaten bereitstellen. Daher empfiehlt sich eine übergreifende Datenbasis, die eine Obermenge der Dimensionsattribute der Abnehmersysteme umfasst.

Um die Handhabbarkeit zu erhöhen, sollten die Nutzer systemspezifischer Dimensionsattribute erfasst werden.

Datenformate

Unterschiedliche Datenquellen können inhaltlich gleiche Daten in heterogenen Formaten anliefern, ebenso wie Zielsysteme auf die Lieferung von Daten in unterschiedlichen Formaten angewiesen sein können. Dies kann einerseits Dateiformate (wie z.B. Text oder XML), andererseits Feldformate (wie z.B. europäische und amerikanische Datumsformate) betreffen. Daher ist es wünschenswert, dass das Pflegesystem eine große Zahl von Datenformaten lesen und erzeugen kann, um den Einsatz von zusätzlichen Konvertern zu minimieren.

Schnittstellen

Ebenso wie die Datenformate können sich auch die Übertragungswege unterscheiden, über die Daten in das bzw. aus dem Pflegetool fließen. Ein ideales Pflegetool unterstützt über den Zugriff auf Dateisysteme hinausgehende Möglichkeiten des Datenzugriffs, etwa über Datenbankschnittstellen oder über Web Services.

ETL-Systeme

Vor einer Zentralisierung sind die Data Warehouses und insbesondere deren Mechanismen zur Extraktion, Transformation und zum Laden (ETL) auf die bestehenden Quellsysteme zur Wartung der Dimensionsdaten abgestimmt (z.B. erwarten sie bestimmte Dateiformate oder können Daten nur über spezifische Schnittstellen lesen). Falls die bestehenden Pflegesysteme eng mit der DWH-Architektur verzahnt und nicht etwa nur über den Austausch von Dateien angebunden sind, können Anpassungen bei den ETL-Tools und -Prozessen erforderlich sein.

Sicherheit

Im Dimensions-Pflegesystem müssen an exponierter, da von zahlreichen Seiten zugreifbarer Stelle umfangreiche vertrauliche Daten gespeichert werden. Um Vertraulichkeit, Integrität und Konsistenz der Daten sicherzustellen, ist es notwendig, den Zugriff auf die Daten nur ausgewählten Nutzern mit angemessenen Sicherheitsmaßnahmen zu gewähren.

Automatisierte Verarbeitung

Wenn als Ziel der zentralisierten Dimensionswartung eine Reduktion des Aufwands angenommen wird, ist eine Automatisierung der Verarbeitung wünschenswert. Das Sys-

tem sollte in der Lage sein, möglichst ohne manuelle Eingriffe Dimensionsdaten von einer Quelle einzulesen, zu verarbeiten und an die Abnehmersysteme auszuliefern.

Multiuser-Fähigkeit

Je nach organisatorischer Ausgestaltung der Data Warehouse-Abteilung kann die Pflege der Dimensionsdaten über mehrere Standorte verteilt sein. Dies macht es erforderlich, dass Nutzer von verschiedenen Orten, u.U. gleichzeitig auf das System zugreifen. Hierfür sind eine entsprechende Client-Technologie (z.B. ein Web-Interface) ebenso erforderlich wie Mechanismen zur Konsistenzsicherung und Sperrung von Datensätzen während des Zugriffs.

3.3 Organisatorische Aspekte

Verantwortlichkeiten

Für die Pflege der Dimensionsdaten sind klare Verantwortlichkeiten erforderlich. Es muss definiert werden, welche Organisationseinheit für die Pflege der einzelnen Dimensionen verantwortlich ist und auf welchen Wegen Änderungen an den Dimensionen gemeldet und autorisiert werden. Hierbei ist zu berücksichtigen, dass bei der Dimensionspflege teilweise Fachwissen erforderlich ist, um etwa neue Dimensionselemente korrekt in die Hierarchien einzuordnen. Eine Delegation der Dimensionspflege an die entsprechende Fachabteilung kann dabei sinnvoll sein, setzt aber ein Pflegetool voraus, das ohne tiefergehende technische Kenntnisse bedient werden kann.

Prozesse

Eine Zentralisierung der Dimensionspflege erfordert eine Umgestaltung von Prozessen auf verschiedenen Ebenen (vgl. Abschnitt 0). Um eine zeitnahe Übernahme von Dimensionsänderungen zu ermöglichen, sind definierte Pflegeprozesse erforderlich, die sich u.U. deutlich von den bisher in den Einzelsystemen praktizierten Abläufen unterscheiden. Ein ideales Pflegetool unterstützt diese Arbeitsabläufe mit einer Workflow-Komponente, die etwa Redaktion, Fehlerbehebung und Abnahme automatisch steuert.

4 Lösungsansätze für die Integration der Dimensionswartung

Im Rahmen eines Zentralisierungsprojekts für die Pflege der Dimensionsdaten sollten die oben genannten Anforderungen möglichst weitgehend umgesetzt werden. In diesem Abschnitt zeigen wir mögliche Zielarchitekturen auf und geben basierend auf Projekterfahrungen Hinweise auf Möglichkeiten zur fachlichen Konsolidierung der Dimensionsdaten, zur Unterstützung durch Software-Werkzeuge und zur Umgestaltung der Ablauforganisation.

4.1 Zentralisierungsdimensionen für die Dimensionswartung

Die Wartung der Dimensionsdaten ist in der Ausgangssituation einer verteilten DWH-Architektur in zwei Dimensionen dezentral organisiert. Einerseits ist die Wartung organisatorisch dezentral, d.h. über verschiedene Organisationseinheiten verteilt, andererseits ist sie auch technisch dezentral, sie wird also über verschiedene Systeme abgewickelt.

Um die Dimensionswartung in einer verteilten DWH-Umgebung zu zentralisieren, gibt es grundsätzlich verschiedene Lösungsmöglichkeiten. Wir nehmen dabei an, dass die datenliefernden operativen Systeme bzw. externen Organisationseinheiten sowie die DWH-Systeme selbst nicht grundlegend verändert werden sollen. Dann ist eine Zentralisierung in zwei Dimensionen denkbar: in organisatorischer und in technischer Hinsicht.

Hieraus ergeben sich vier mögliche Architekturen, die in Abbildung 5 dargestellt sind. Diese Architekturen stellen Idealtypen dar, Mischformen sind durchaus denkbar. Die erste Alternative entspricht der eingangs dargestellten Ausgangssituation, hier findet keine Zentralisierung statt. Alternative 2 ist eine Zentralisierung nur in organisatorischer Hinsicht. Hier wird die Dimensionswartung für jede Dimension von einer verantwortlichen Organisationseinheit für alle Systeme mit getrennten technischen Pflegesystemen durchgeführt. Alternative 3 ist eine lediglich technische Zentralisierung. Die Pflege wird weiterhin für jedes DWH und jede Dimension von einer eigenständigen Organisationseinheit durchgeführt, wobei jedoch auf ein zentrales Pflegesystem zurückgegriffen wird. Alternative 4 schließlich entspricht einer Zentralisierung in organisatorischer wie in technischer Hinsicht. In dieser Situation werden die Dimensionsdaten jeweils für alle Abnehmersysteme von einer Organisationseinheit mit einem zentralen Pflegesystem gewartet.

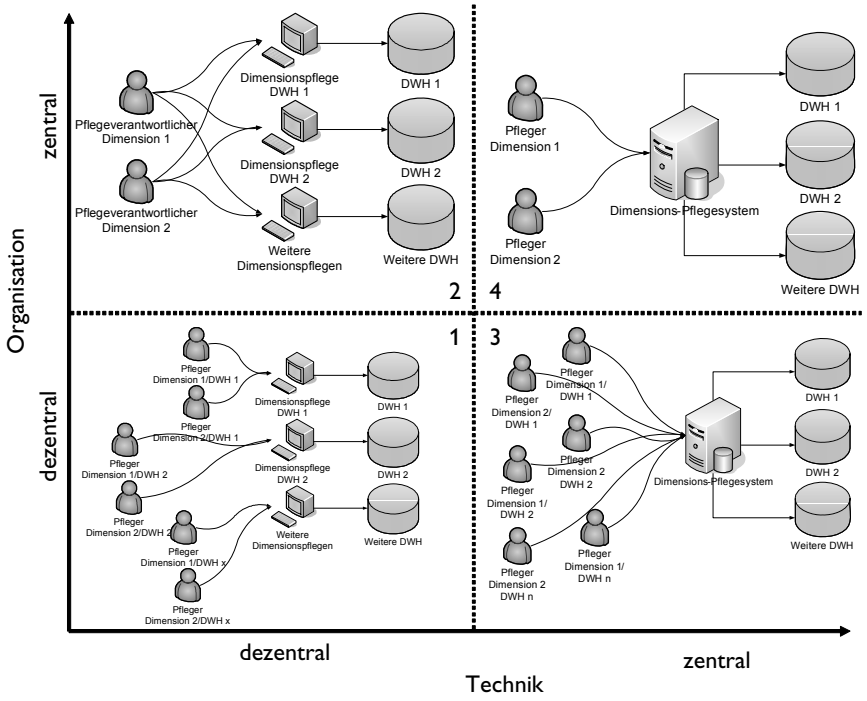


Abb. 5: Architekturalternativen für die Dimensionswartung

Eine in technischer und organisatorischer Hinsicht zentralisierte Wartung der Dimensionsdaten (Alternative 4 in Abb. 5) verspricht eine Reihe von Vorteilen. Das Problem der mangelnden Vergleichbarkeit von Daten aus verschiedenen Systemen kann bei Nutzung einheitlicher Dimensionen deutlich gemindert werden und es können die Grenzen zwischen den Systemen transparent gemacht werden. Die Einrichtung eines zentralen Systems vereinfacht die Implementierung von Konsistenz- und Korrektheitsprüfungen, um die Fehlerwahrscheinlichkeit zu mindern. Zudem wird die Wartung und Weiterentwicklung des Pflegesystems vereinfacht, wenn nur noch ein System existiert. In organisatorischer Hinsicht können definierte Verantwortlichkeiten geschaffen und eindeutige Ansprechpartner benannt werden. Der Aufwand für die Dimensionspflege sinkt, da Pflegearbeiten nur noch einmalig ausgeführt werden müssen. Schließlich wird der Aufbau des für die Pflege nötigen Know-Hows vereinfacht, da für die Datenpflege nötiges Fachwissen nur noch an einer Stelle vorgehalten werden muss.

Allerdings ist die Zentralisierung auch nicht völlig ohne Risiken. Falls das zentrale Wartungstool nicht den gesamten Funktionsumfang bestehender Wartungslösungen der einzelnen Data Warehouses unterstützt, kann es zu Akzeptanzproblemen und zum Parallelbetrieb von Altlösungen kommen. Die Verteilung der Verantwortlichkeiten für die einzelnen Dimensionen auf einzelne Abteilungen kann ebenfalls problematisch sein. Bei der initialen Generierung der übergreifenden Dimensionsdaten ist mit erheblichem Mapping-Aufwand zu rechnen. Im laufenden Betrieb schließlich können verstärkte Maßnah-

men zur Qualitätssicherung erforderlich sein, da sich Fehler in den Dimensionsdaten sofort auf alle belieferten Systeme auswirken.

Bei einer nur teilweisen Zentralisierung (Alternativen 2 und 3) können nicht alle Vorteile realisiert werden, die Nachteile einer dezentralen Organisation bzw. Technik bleiben bestehen (vgl. Abschnitt 4.4).

Um die Vorteilhaftigkeit der einzelnen Alternativen zu beurteilen, wurden die befragten Experten aufgefordert, eine Einschätzung des Kosten-Nutzen-Verhältnisses der einzelnen Alternativen abzugeben. Die einzelnen Einschätzungen sind in Abbildung 6 zusammengefasst. Der vollständigen Zentralisierung (Alternative 4) wurde der höchste Nutzen beigemessen, allerdings auch die höchsten Kosten. Weniger nützlich scheinen die Alternativen 2 und 3, wobei das Aufwand/Nutzen-Verhältnis bei einer lediglich organisatorischen Zentralisierung als tendenziell schlechter eingeschätzt wurde als bei einer technischen Zentralisierung. Ein Beibehalten der dezentralen Wartung (Alternative 1) wurde generell als wenig vorteilhaft eingeschätzt, auch wenn der Aufwand für ein Zentralisierungsprojekt entfällt. Weitgehend übereinstimmend waren die befragten Experten der Meinung, dass einer umfassenden Zentralisierung gegenüber den Teillösungen der Vorzug zu geben sei. In den folgenden Abschnitten geben wir Hinweise auf fachliche, technische und organisatorische Aspekte zur Umsetzung eines Zentralisierungsprojekts mit der Zielsetzung einer vollständigen Zentralisierung.

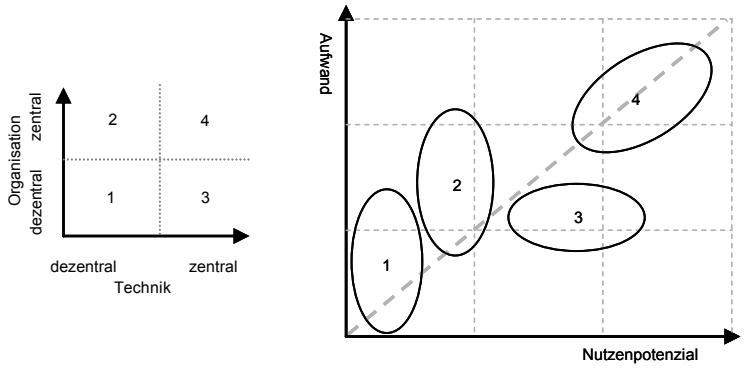


Abb. 6: Aufwand/Nutzen-Verhältnis der Architekturalternativen

4.2 Fachliche Aspekte

Aus fachlicher Sicht ist die Zentralisierung von Dimensionsdaten vergleichbar mit der Modellierung von Dimensionen in einem zentralen Data Warehouse. In beiden Fällen sollen und müssen unterschiedliche fachliche Sichten, beispielsweise in der Dimension „Kunde“, zu einer maximalen Überdeckung gebracht werden.

In den meisten Fällen werden die Basiselemente (Dimensionselemente auf der niedrigsten Hierachiestufe, wie Kunden und Produkte) aus Stammdaten in den operativen Systemen befüllt. Falls dabei Basiselemente aus mehreren „konkurrierenden“ Quellsystemen zur Verfügung stehen, sollte jeweils ein führendes System festgelegt werden. Dazu

werden – wie bei einem zentralen Data Warehouse – die operativen Systeme mit Stammdaten (die dann in die Basiselemente münden) auf Querverbindungen untersucht, d.h. ob es operative Systeme gibt, die Stammdaten aus anderen operativen Systemen beziehen. Dann bietet sich in vielen Fällen an, das originäre Stammdatensystem zum führenden System zu erklären. Ergänzt werden sollte diese Festlegung um die Etablierung der Data Ownership für die Basiselemente, um Qualität und Aktualität dieser Daten zu gewährleisten.

Die Konsolidierung von Dimensionshierarchien (dazu gehören der Aufbau der Hierarchie, die konkrete Zuordnung von Dimensionselementen und die Aggregationsfunktionen entlang der Hierarchiepfade) gestaltet sich in der Praxis schwieriger: Die Koexistenz mehrerer Hierarchien (z.B. zur Klassifizierung von Kunden oder Produkten) wird in der Regel mit vermeintlicher oder tatsächlicher fachlicher Notwendigkeit begründet. Diese ist jedoch durchaus kritisch zu hinterfragen. Zwar sollte eine möglichst weitgehende Konsolidierung angestrebt werden, auf der anderen Seite müssen ggf. fachlich begründete Redundanzen auch zugelassen werden, wie etwa unterschiedliche Regionalstrukturen im Vertrieb für Geschäfts- und Privatkunden. Wichtig ist dann aber eine transparente und kommunizierte Terminologie und die Vermeidung von Homonymen, also gleichen Begriffen für unterschiedliche Dimensionen oder Hierarchiestufen (Beispiel: „Kundensegment“). Damit sollen Fehlinterpretationen bei Auswertungen vermieden werden. Aufgrund der ähnlichen Anforderungen bietet es sich an, ggf. im Unternehmen bereits bestehende Metadaten-Management-Systeme einzubinden. So kann Transparenz für die nutzenden Systeme geschaffen werden hinsichtlich Bedeutung, Herkunft und Aktualität der Dimensionsdaten – was letztlich eine notwendige Voraussetzung für die korrekte Interpretation von Analysedaten ist. Es bewährt sich, ein vordefiniertes „Standard-Set“ von Verdichtungen (in der Zeitdimension, in den Organisationsstrukturen, etc.) anzubieten; bei der Anforderung nach Umsetzung neuer Aggregationen sollte nach Möglichkeit auf diese Verdichtungen zurückgegriffen werden. Damit wird der „Wildwuchs“ nutzerspezifischer Verdichtungen mit den bereits genannten Nachteilen eingedämmt. Der Konsistenz zuträglich ist zudem die Umsetzung des Vererbungsprinzips. Vererbung heißt in diesem Kontext, dass Teilhierarchien von mehreren DWH-Systemen genutzt werden. Ändert sich an der „Master“-Hierarchie etwas, so wird diese Änderung automatisch an die mitnutzenden Systeme vererbt, was divergierende Hierarchien und Dimensionen vermeidet.

Ein bewährtes Steuerungsinstrumentarium zur Eindämmung beliebiger Vielfalt in den nutzerspezifischen Sichten stellt ein entsprechendes Verrechnungsprinzip dar. Alternativ kann die Bereitstellung individueller Hierarchien mit der Verpflichtung verbunden werden, diese Hierarchien auch anderen Nutzern bzw. Systemen zugänglich zu machen – dies wäre insbesondere bei dezentraler Organisation sinnvoll. Allerdings ist dann auch dafür Sorge zu tragen, dass Änderungen an den Hierarchien mit später hinzugekommenen Benutzern abgestimmt werden.

4.3 Technische Umsetzung

Die zentralisierte Wartung der Dimensionsdaten erfordert die Unterstützung durch ein Softwaresystem (vgl. Abschnitt 3.2). Neben der Erstellung einer Individualsoftware

bietet sich die Nutzung eines kommerziellen Tools an, das die Anforderungen möglichst vollständig abdecken sollte. Das Angebot an Software zur Verwaltung von Stammdaten ist auf den ersten Blick umfangreich, allerdings zeigt sich, dass hier unter dem Schlagwort „Master Data Management“ sehr unterschiedliche Produkte zur Verwaltung von Dimensionsdaten und Stammdaten von verschiedensten Herstellern auf den Markt gebracht (bzw. umetikettiert) worden sind [Wa06].

Der Grund hierfür ist darin zu sehen, dass die Verwaltung und Vereinheitlichung von Stammdaten im operativen Bereich in den letzten Jahren erhebliche Aufmerksamkeit seitens Software-Herstellern (insbesondere SAP) und Anwendern erfahren hat. Das Problem der Dimensionsdatenwartung in verteilten analytischen Systemen ist von den Herstellern ebenfalls erkannt und adressiert worden. Hier stehen jedoch mit den Dimensionsdaten andere Datenarten im Vordergrund. Schließlich sind in Data Warehousing-Software oftmals Komponenten für das Metadaten-Management integriert, zudem existieren generische Tools für das Metadaten-Management.

Bei den Softwareprodukten sind dementsprechend drei (sich teils überschneidende und konvergente) Kategorien erkennbar. Einerseits gibt es Produkte, die auf die Verwaltung und Vereinheitlichung von Kunden- und Lieferantendaten spezialisiert sind (z.B. SAP NetWeaver Master Data Management). Diese sind für die Dimensionsdatenverwaltung im DWH-Umfeld eher weniger geeignet, da der funktionale Fokus auf der Verwaltung und Vereinheitlichung von Stammdatensätzen (z.B. Adressen) liegt. DWH-spezifische Anforderungen werden weniger umgesetzt. Zudem ist die Schnittstellenausstattung eher auf ERP-Systeme und operative Applikationen z.B. zum CRM fokussiert.

Andererseits existieren Produkte, die aus Spezialapplikationen für die Wartung von DWH-Dimensionen hervorgegangen sind (z.B. Hyperion Master Data Management oder Kalido 8). Bei diesen Produkten liegt der Fokus auf DWH-spezifischen Fragestellungen, auch die Ausstattung mit Schnittstellen zu DWH-spezifischer Software ist umfassender. Ein Produkt aus dieser Gruppe kann bei entsprechendem Funktionsumfang für die Dimensionsdatenverwaltung genutzt werden.

Die DWH-spezifischen Tools unterstützen Aufgaben wie Import und Integration von Dimensionsdaten aus verschiedenen Quellen, Änderung, Versionierung und Dokumentation der Dimensionsdaten, Bildung von Zusatzverdichtungen, Validierung und Verifikation, Zugriffsverwaltung und Export in verschiedene Systeme. Dabei sind die Produkte i.d.R. nicht integraler Teil der DWH-Applikationen einzelner Hersteller, sondern kooperieren mit Produkten von verschiedensten Anbietern bzw. stellen generische Import- und Exportschnittstellen bereit. Ein Trend zu engerer Verzahnung mit der Software bestimmter Anbieter und zur Konsolidierung des Marktes ist jedoch zu beobachten. Die Kompatibilität mit der gesamten DWH-Architektur des Unternehmens muss detailliert überprüft werden, insbesondere wenn heterogene DWH-Softwarelösungen eingesetzt werden. Zudem ist zu prüfen, in wie weit die Funktionalität der verfügbaren Lösungen über die Anforderungen hinausgeht und so unnötige Komplexität und Kosten mit sich bringt. Die in Abschnitt 0 vorgestellten Anforderungen können individuell gewichtet und u.U. ergänzt als Kriterienkatalog für eine Toolauswahl dienen.

Als dritte Werkzeugkategorie existiert zwar für das Metadaten-Management eine große Auswahl an ausgereiften Softwarewerkzeugen. Diese sind jedoch für die hier behandelte

Fragestellung eher ungeeignet, falls sie nicht bereits im Unternehmen zum Einsatz kommen. Repositories sind i.d.R. äußerst generisch angelegt. Während sie grundsätzlich zur Wartung von verteilten DWH-Dimensionen einsetzbar sind, ist hier mit einem erheblichen Anpassungsaufwand zu rechnen, der eine spezialisiertere Softwarelösung attraktiver erscheinen lässt, die größere Teile der benötigten Funktionalitäten bereits zur Verfügung stellt.

4.4 Organisatorische Aspekte

Um die Konsequenzen einer zentralen Dimensionswartung für die Organisation des Data Warehousing herauszuarbeiten, geben wir zunächst einen Überblick über die von der Zentralisierung betroffenen Prozesse. In Abschnitt 0 wird dann exemplarisch ein Sollprozess für die zentralisierte Dimensionswartung entworfen.

4.4.1 Beeinflusste DWH-Prozesse

Eine Zentralisierung der Wartung der Dimensionsdaten kann sich an vielen Stellen auf die Organisation des Data Warehousing auswirken, nicht nur auf den Prozess der Dimensionspflege selbst. Um den Einfluss der Zentralisierung auf die Organisation zu analysieren, wurden die Prozesse des Data Warehousing im Rahmen von Experteninterviews analysiert. Diese Prozesse lassen sich in vier Bereiche gliedern: DWH-Nutzung, Leistungsprozesse (mit den Untergruppen Entwicklung und Betrieb des DWH-Systems), Zulieferungsprozesse und Managementprozesse (mit den Untergruppen DWH-Steuerung und Steuerung von Unterstützungsprozessen).

Die befragten Experten wurden aufgefordert, den Einfluss der Zentralisierung auf die Prozesse einzuschätzen. Die Prozesse und die Stärke des Einflusses sind in Tabelle 1 dargestellt. Bei denjenigen Prozessen, die durch die Zentralisierung stark beeinflusst werden, kann ein Prozess-Reengineering erforderlich werden.

Der Bereich Nutzung des DWH mit strategischen Entscheidungsprozessen und operativen Geschäftsprozessen ist von einer Zentralisierung der Dimensionswartung kaum betroffen. Dies ist einleuchtend, da sich die im Frontend des DWH verfügbaren Daten nicht ändern (außer im Hinblick auf verbesserte Datenqualität). Lediglich die Konsolidierung von Daten aus verschiedenen Quellen für strategische Entscheidungen und nachgelagerte qualitätssichernde Maßnahmen in den Nutzungsprozessen werden vereinfacht.

Der am stärksten beeinflusste Prozesscluster ist der der Leistungsprozesse zur Entwicklung des DWH. Dieser Einfluss entsteht hauptsächlich aus der Umsetzung des Zentralisierungsprojekts selbst. Allerdings ist auch nach der Zentralisierung ein Einfluss auf die Entwicklungsprozesse gegeben, insbesondere dann, wenn zusätzliche Liefersysteme oder zusätzliche Data Warehouses angeschlossen werden sollen.

Im Bereich der Prozesse zum Betrieb des DWH sind teils erhebliche Änderungen notwendig, da durch die Nutzung eines neuen Systems zur Wartung der Dimensionen die Ladeprozesse angepasst werden müssen.

Die Prozesse zur Zulieferung und Steuerung sind nur gering beeinflusst, allenfalls muss das Architekturmanagement die geänderte Architektur der Dimensionswartung und die damit einhergehende größere Integration der Teilsysteme berücksichtigen.

Beeinflusste Prozesse		
Bereich	Prozess	Stärke
Nutzungsprozesse	Strategische Entscheidungsprozesse	-
	Operative Geschäftsprozesse	--
Leistungsprozesse Entwicklung	Design	++
	Realisierung	++
	Analyse	○
	Test	○
	Einführung	○
Leistungsprozesse Betrieb	Transformation und Laden des DWH	+
	Transformation und Laden der Data Marts	○
	Betrieb des DWH	-
	Betrieb der Data Marts	-
	Support	-
Zulieferungsprozesse	Datenbereitstellung	-
	Softwarebereitstellung	-
	Hardwarebereitstellung	-
Steuerung DWH	Architekturmanagement	○
	Finanzmanagement	-
	Strategieentwicklung	-
Steuerung Unterstützungsprozesse	Metadaten-Management	+
	Wartung der Dimensionen	++
	Datenqualitätsmanagement	+
	Datensicherheitsmanagement	○
	Datenschutz	-

Tab. 1: Einfluss auf die Prozesse des Data Warehousing

Anzupassende Unterstützungsprozesse sind neben der Dimensionswartung u.U. das Metadaten-Management und das Sicherheitsmanagement, wo eine neue zentrale DWH-Komponente zu verwalten ist.

Die Zentralisierung der Dimensionspflege hat also vielfache Auswirkungen auf die Prozesse des Data Warehousing. Daher muss der Prozessperspektive bei der Umsetzungsplanung und beim Change Management besondere Aufmerksamkeit geschenkt werden.

4.4.2 Prozessmodell für die zentralisierte Dimensionspflege

Der Prozess der Dimensionspflege muss im Rahmen einer Zentralisierung an die geänderte Systemarchitektur angepasst werden. Im Folgenden wird die Lösung für einen vereinheitlichten Pflegeprozess aus einem Projekt vorgestellt. Der Prozess ist in einen systemübergreifenden Teil (vgl. Abbildung 7) und in einen systemspezifischen Teil (vgl. Abbildung 8) aufgeteilt. Der systemübergreifende Teil des Prozesses wird über das zentrale Wartungssystem abgewickelt, der systemspezifische Teil über die Data Warehouses bzw. deren ETL-Tools. Die wichtigsten Änderungen zum bestehenden Prozess ergeben sich im systemübergreifenden Teil, in dem die früher separat ausgeführten Pflegeprozesse zusammengefasst werden.

Sobald die Quelldaten für die Dimensionen bereitstehen, kann der Pflegeprozess gestartet werden. Zuerst wird eine neue Version der Dimensionsdaten angelegt und in den Fällen, in denen nur Änderungen bereitgestellt werden, mit den Daten der Vorversion initialisiert. Dann werden die Quelldaten in das Pflegetool importiert und soweit möglich von den Dimensionselementen abhängige Dimensionsattibute automatisch generiert. Anschließend werden gemeinsam genutzte und systemspezifische Attribute (z.B. eindeutige Schlüssel) ergänzt. Sämtliche redaktionelle Bearbeitung der Daten findet in diesen Arbeitsschritten statt. Daraufhin werden die Daten auf Konsistenz und Plausibilität (z.B. auf Einhaltung erlaubter Wertebereiche) kontrolliert. Falls Fehler auftreten, wird in die vorhergehenden Bearbeitungsschritte zurückgesprungen. Ansonsten kann die gemeinsame Verarbeitung mit der Freigabe der neuen Version abgeschlossen werden.

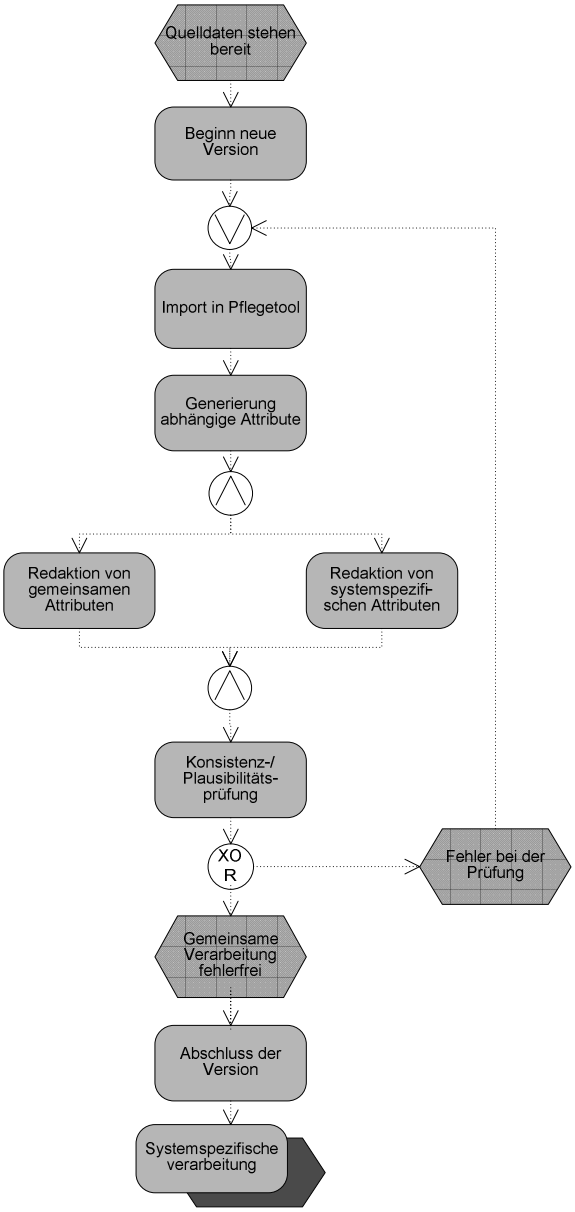


Abb. 7: Pflegeprozess - gemeinsamer Teil[1]

[1] Darstellung als vereinfachte ereignisgesteuerte Prozesskette (EPK) [BKR02].

Im zweiten Teil des Prozesses finden systemspezifische Verarbeitungsschritte für jedes belieferte DWH statt. Dazu werden die Dimensionsdaten aus dem Pflegetool in die Data Warehouses importiert. Eine systemspezifische Weiterverarbeitung kann etwa für den Aufbau multidimensionaler Datenstrukturen erforderlich sein. Nach dem Laden in die Data Warehouses müssen zudem eventuell während des Ladeprozesses aufgetretene Fehler zur Korrektur an die für die Dimensionspflege zuständige Organisationseinheit weitergeleitet werden. Damit ist der Verarbeitungsprozess abgeschlossen.

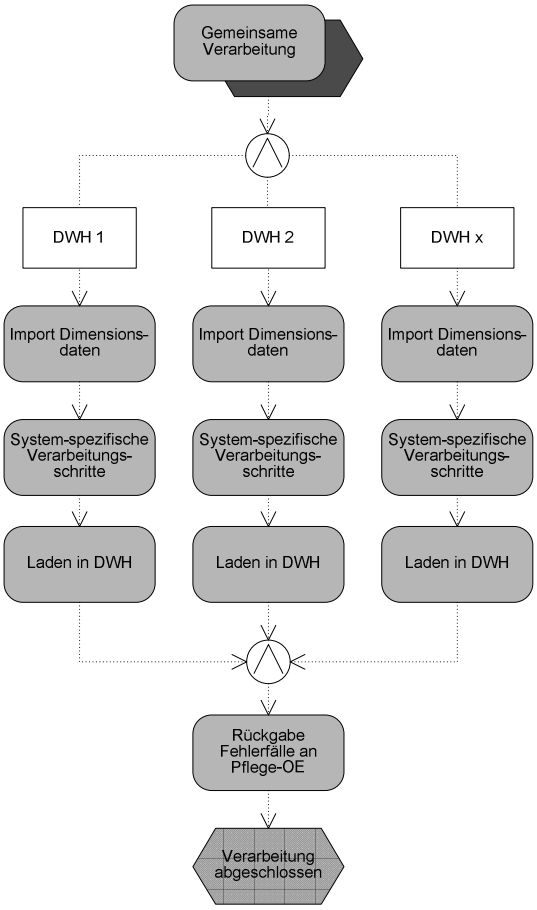


Abb. 8: Pflegeprozess - systemspezifischer Teil

5 Zusammenfassung und Ausblick

Verteilte Data Warehouse-Architekturen ohne gemeinsame Dimensionen stellen eine in vielerlei Hinsicht unbefriedigende Lösung dar und verursachen diverse Probleme. An eine mögliche Lösung – ein System zur zentralen Wartung von Dimensionsdaten – wer-

den zahlreiche Anforderungen in fachlicher, technischer und organisatorischer Hinsicht gestellt, die eine Umsetzung erschweren. Dieser Artikel untersucht und diskutiert Realisierungsoptionen für die organisatorische und / oder technische Zentralisierung. Für die Umsetzung eines solchen zentralen Pflegesystems existieren Softwarelösungen von verschiedenen Herstellern. Sie unterscheiden sich jedoch erheblich im Funktionsumfang und in ihrer Eignung für DWH-Projekte. Eine sorgfältige Evaluation ist daher unumgänglich. Zudem bedingt eine zentrale Lösung die Anpassung der Organisation. Zahlreiche Prozesse des Data Warehousing müssen adaptiert werden. Insbesondere die Prozesse für die Wartung der Dimensionsdaten bedürfen der Vereinheitlichung. Der vorliegende Artikel berichtet von Erfahrungen in einem solchen Projekt und gibt Empfehlungen für die Realisierung.

Die Fragestellungen zentralisierter Dimensionswartung wurden bisher in der Forschung noch kaum explizit betrachtet. So gibt es zwar Arbeiten zum Stammdatenmanagement in operativen Systemen und zahlreiche Beiträge zur multidimensionalen Modellierung, die den Aufbau (einzelner) Dimensionen adressieren. Jedoch finden sich unserer Beobachtung nach noch keine Ansätze, die die fachliche, technische und organisatorische Konsolidierung von Dimensionen im Detail darstellen. Eine methodische Unterstützung wäre hier sehr wünschenswert und würde der hohen Praxisrelevanz dieses Themas Rechnung tragen.

6 Literaturverzeichnis

[AM97] Anahory, S.; Murray, D.: Data Warehouse. Planung, Implementierung und Administration. Addison-Wesley, Harlow, 1997.

[AW05] Ariyachandra, T.; Watson, H.: Key Factors in Selecting a Data Warehouse Architecture, in: Business Intelligence Journal, 10(2005)2; S. 19-26.

[AW06] Ariyachandra, T.; Watson, H.: Which Data Warehouse Architecture Is Most Successful?, in: Business Intelligence Journal, 11(2006)1; S. 4-6.

[Au03] Auth, G.: Prozessorientierte Organisation des Metadatenmanagements für Data Warehouse-Systeme, St. Gallen, Bamberg, 2003.

[BG04] Bauer, A.; Günzel, H.: Data Warehouse Systeme – Architektur, Entwicklung, Anwendung, 2. Auflage Aufl., dpunkt.verlag, Heidelberg, 2004.

[BKR02] Becker, J.; Kugeler, M.; Rosemann, M.: Prozessmanagement – Ein Leitfaden zur prozessorientierten Organisationsgestaltung, 3 Auflage, Springer, Berlin, Heidelberg, 2002.

[In02] Inmon, W.H.: Building the Data Warehouse, 3 Auflage, Wiley Computer Publishing, New York, 2002.

[IIS01] Inmon, W.H.; Imhoff, C.; Sousa, R.: Corporate Information Factory, Wiley, New York, 2001.

[Ki98] Kimball, R.; Reeves, L.; Ross, M.; Thornthwaite, W.: The Data Warehouse Lifecycle Toolkit. Expert Methods for Designing, Developing, and Deploying Data Warehouses. Wiley, New York, 1998.

[Ku99] Kurz, A.: Data Warehousing Enabling Technology, MITP, Bonn, 1999.

[LLG04] Loser, C.; Legner, C.; Gizanis, D.: Master Data Management for Collaborative Service Processes, International Conference on Service Systems and Service Management, Research Center for Contemporary Management, Tsinghua University, forthcoming, 2004.

[MSW03] von Maur, E.; Schelp, J.; Winter, R.: Integrierte Informationslogistik – Stand und Entwicklungstendenzen, In (von Maur, E.; Winter, R.) (Hrsg.): Data Warehouse Management. Das St.Galler Konzept zur ganzheitlichen Gestaltung der Informationslogistik, Springer, Berlin, 2003; S. 3-23.

[Wa06] Waddington, D.: MDM for Operations and Analytics: Intelligent Enterprise, 2006.