Forschungs- und Entwicklungsprojekte

Informatik Forsch. Entw. (2003)
Digital Object Identifier (DOI) 10.1007/s00450-003-0130-8

In dieser Rubrik erscheinen in unregelmäßiger Folge Kurzdarstellungen geplanter, laufender oder abgeschlossener Projekte. Die Darstellungen werden in der Regel von den Projektbeteiligten geliefert. Die Auswahl erfolgt durch die Herausgeber. Dabei wird die Bedeutung des Projekts für die Fortentwicklung der Informatik das Hauptkriterium sein. Bei geplanten und laufenden Projekten ist ein wichtiges Kriterium der Wunsch, Kontakte zu etablieren und die Zusammenarbeit zwischen verschiedenen Gruppen zu fördern. Bei abgeschlossenen Projekten geht es primär um die Vermittlung von Erfahrungen und Ergebnissen, die sich nicht für die Veröffentlichung in redaktionellen Beiträgen eignen.

Data Warehousing als Organisationskonzept des Datenmanagements

Eine kritische Betrachtung der Data-Warehouse-Definition von Inmon

Thomas Zeh

Merck KGaA, Information Management and Consulting, Frankfurter Str. 250, 64293 Darmstadt (E-mail: thomas.zeh@merck.de)

1 Einleitung

Es lohnt sich gelegentlich, auch etablierte Dinge infrage zu stellen. Ein derartiges Ding ist die Data-Warehouse-Definition[1] von Inmon. Kaum ein Data-Warehouse-Buch, in dem der vermeintliche Vater des Data-Warehouse-Begriffs[2] nicht mit seiner Definition zitiert wird. Nur selten findet man kritische Anmerkungen dazu, so z.B. von Bauer und Günzel [2]. Dennoch gibt es bis heute keine offizielle Standard-Definition des Data-Warehouse-Begriffs.

Im Folgenden wird dargestellt, wie Inmon seine Definition begründet, wie seine Definition in der Literatur interpretiert wird, welche einschränkende Wirkung sie auf die Praxis des Data Warehousing hat und welche Änderungen am Data-Warehouse-Begriff heute – elf Jahre nach seiner Publikation – zweckmäßig erscheinen.

Die folgenden kritischen Ausführungen sollen dazu dienen, die Diskussion über die Zweckmäßigkeit der Inmon'sche Begriffsdefinition anzustoßen und auf ein einvernehmliches Begriffsverständnis hinzuwirken.

2 Motivation

Es gibt nicht die richtige Begriffsdefinition – allenfalls eine falsche, wenn in sich widersprüchliche – sondern es gibt nur mehr oder weniger zweckmäßige Begriffsdefinitionen. In diesem Beitrag wird versucht, orientiert an den organisatorischen und technischen Möglichkeiten einer mehrschichtigen Architektur für ein Data-Warehouse-System (s. Abb. 1), zu einer pragmatischen Definition des Data-Warehouse-Begriffs hinzuführen. Der dabei entstehende neue Begriff unterscheidet sich von dem Data-Warehouse-Begriff von Inmon dadurch, dass auf einige der Einschränkungen von Inmon verzichtet wird; somit umfasst der neue Data-Warehouse-Begriff den alten.

Motiviert zu diesem Versuch wurde der Autor durch die Erfahrung, dass in seinem betrieblichen Umfeld nicht wenige IT-Projekte durchgeführt wurden, die organisatorisch und technisch mit Methoden des Data Warehousing hätten bearbeitet werden können. Daten aus verschiedenen Quellen, z.B. aus Forschungslabors und Niederlassungen, wurden in betrieblichen Anwendungen anderenorts benötigt. Keiner sprach jedoch von Data Warehousing, keiner setzte Data-Warehouse-Werkzeuge ein, und eine Datenorganisation mittels Informationsdrehscheiben (Nabe-Speiche-Architektur) war nicht in Sicht. Die Folge war, dass das Netzwerk von Datenbeständen und den direkten Datenflüssen zwischen den Dateninseln immer dichter und unüberschaubarer wurde. Hinzu kam, dass wegen fehlenden Wissens um die Existenz von bestimmten Daten diese erneut erhoben und erfasst oder abgeleitet wurden, was nicht selten zu unterschiedlichen Aussagen und damit zu Problemen führte.

Die Ursache für diese Entwicklung ist nach Auffassung des Autors mentaler Art. In den Köpfen herrschte die Vorstellung: Data Warehousing ist nur etwas für „Decision Support", es ist nur gut für das „Managementreporting", denn dies entspricht der Definition von Data Warehousing und seiner Interpretation in der Literatur.

Aufgaben mit den integrierenden Methoden des Data Warehousing stellen sich häufig im Datenmanagement; bisher wurden aber die Methoden und Techniken des Data Warehousing fast immer nur im betriebswirtschaftlich orientierten Umfeld des mittleren und oberen Managements eingesetzt. Es besteht also eine Lücke zwischen dem Potenzial der Data-Warehousing-Konzepte und der derzeitigen Praxis. Ließe sich diese Lücke nicht durch eine Erweiterung der bisher nur in einem Bereich eingesetzten Konzepte auf andere Bereiche mit ähnlich strukturierten Problemen schließen? Sollte es nicht möglich sein, das Problem ursächlich anzugehen und den

[1] Wenn im Folgenden von Data Warehouse die Rede ist, dann ist damit das zentrale Datenlager gemeint, während das Data-Warehouse-System alle für das Data Warehousing notwendigen Komponenten umfasst.

[2] Der Begriff wurde Mitte der 80er Jahre bei IBM geprägt und mit Information Warehouse bezeichnet. Der Terminus Data Warehouse wurde erstmals 1988 von Devlin verwendet.

Data-Warehouse-Begriff zu überdenken und ihn im Hinblick auf die Schließung der Lücke zu erweitern?

3 Definition des Data-Warehouse-Begriffs von Inmon

Inmon hat 1992 den Data-Warehouse-Begriff wie folgt definiert und publiziert:

"A data warehouse is a

- subject -oriented,
- integrated,
- time-variant,
- nonvolatile
- collection of data
- in support of management's decision-making process" [7].

Inmon nennt bei seiner Definition sechs den Data-Warehouse-Begriff charakterisierende Eigenschaften.

. . . *subject-oriented* . . .

Mit diesem Charakteristikum der Themenorientierung soll die Auswahl der relevanten Daten und somit der Inhalt des Data Warehouse grob umrissen werden.

Inmon schreibt dazu. "The orientation around the major subject areas of the corporation causes the data warehousing design to be "data driven". The data-driven subject organization of the data warehouse is in contrast to the more classical process / functional organization of applications. . . " [7]. Und er führt als Beispiele für solche Themen(gebiete) an: customer, vendor, product und activity. Im Gegensatz dazu sieht er die folgenden Beispiele aus dem Bankenbereich als anwendungsorientiert: loans, savings, bank card und trust.

Stock interpretiert die Themenorientierung folgendermaßen. „Die Informationseinheiten sind auf die inhaltlichen Kernbereiche einer Organisation ("subject oriented") wie z.B. Produkt und Kunden fokussiert. Im Gegensatz hierzu finden operative Daten, die lediglich auf die Prozessdurchführung oder die Organisationseinheiten eines Unternehmens ausgerichtet sind, keinen Weg in das Data Warehouse" [15].

Die Themenorientierung – also die Fokussierung auf die inhaltlichen Kernbereiche – sowie die Abgrenzung der themenorientierten zu den anwendungsorientierten Informationseinheiten erscheint willkürlich und ist kaum nachvollziehbar, denn das Kriterium für die Abgrenzung bleibt im Verborgenen. Die Mehrfachverwendung der Daten kann Inmon als Kriterium nicht gemeint haben, denn auch die Objekttypen in den obigen Beispielen wie loans und trust können für mehrere Auswertungen – so auch für solche wie den management support - genutzt werden. Analog zu einem Warenlager sollte sich der Inhalt des Datenlagers ständig der Nachfrage anpassen. Es ist kein Grund ersichtlich, warum der Inhalt von vornherein im obigen Sinne beschränkt wird. Sollte es nicht vielmehr im Ermessen der Nutznießer des Data Warehouse liegen, welche Prozesse sie mittels Data Warehousing unterstützen wollen? Die hierzu erforderlichen Daten wären dann zu integrieren; sie würden den Inhalt des Data Warehouse bestimmen. Dieser Aspekt wird durch das Zweck-Charakteristikum im Folgenden noch näher beschrieben.

Sollte daher auf die Themenorientierung als Restriktion für die Data-Warehouse-Definition nicht ganz verzichtet werden?

. . . *integrated* . . .

Inmon erklärt den Integrationsaspekt: "The integration of data warehouse data shows up in many different ways – in consistent naming conventions, consistent measurement of variables, consistent encoding structures, consistent physical attributes of data, and so forth" [7]. Kürzer formuliert bei Stock: „Die Daten werden in einem einheitlichen und konsistenten Datenbestand zusammengefasst" [15] und in der Länge nicht zu unterbieten bei Jung und Winter: „. . . definitorisch und inhaltlich konsolidiert. . . " [9].

Der Integrationsaspekt umfasst die Schemaintegration (Metadaten) wie auch die Datenintegration, das heißt die Abbildung bzw. Transformation der Originaldaten aus den zumeist heterogenen Datenquellen in die nach dem gemeinsamen Schema konstruierte Zieldatenbank. Durch das „Integrieren durch Kopieren" [11] unterscheidet sich das Data-Warehouse-System von dem föderierten Datenbanksystem[3].

Dieses Charakteristikum der Data-Warehouse-Definition wird hier wie auch andernorts einvernehmlich gesehen.

Für Inmon ist die Integration das wichtigste Charakteristikum. Die einheitliche Sicht auf ursprünglich heterogene Daten, die zu diesem Zweck an einem anderen Ort zusammengeführt und entsprechend neu beschrieben werden, ist der grundlegende Gedanke, der hinter dem Data-Warehouse-Begriff steht. Dadurch unterscheidet sich das Data Warehouse von der allgemeinen Datenbank. Dabei gibt es implizit die Forderung der Konsistenz, d.h. die im Data Warehouse zusammengeführten Daten sollen untereinander wie auch zu ihrer Beschreibung widerspruchsfrei sein.

Die Forderung von Inmon nach der Integration der Daten sollte somit bestehen bleiben.

. . . *time-variant* . . .

Inmon verfolgt mit der Zeitorientierung den Schnappschussgedanken, d.h. für ihn ist das Data Warehouse eine Folge von Momentaufnahmen über einen 5- bis 10-jährigen Zeitraum. Konsequenterweise hat bei ihm jeder Satzschlüssel eines Datensatzes im Data Warehouse einen Zeitbezug. Stock interpretiert und motiviert die Zeitorientierung folgendermaßen: „. . . im Data Warehouse werden die Daten über längere Zeiträume ("time variant") vorgehalten. Somit ist es möglich, Analysen von Zeitreihen über längere und mittlere Zeiträume . . . vorzunehmen. . . Jeder Import bietet somit einen Schnappschuss des Unternehmensgeschehens" [15].

Ein Vorteil des Data-Warehouse-Systems gegenüber dem föderierten Datenbanksystem besteht darin, bestimmte Daten

[3] Nach Conrad ist ein föderiertes Datenbanksystem ein Multidatenbanksystem mit einem alle Komponentensysteme umfassenden (globalen) konzeptuellen Schema. Alle Komponentensysteme müssen dabei ihre Autonomie und ihr lokales konzeptuelles Schema bewahren, d.h. sie bleiben im Hinblick auf Entwurf, Ausführung und Kommunikation selbstständig [3].

auch über ihre Lebensdauer in den Quellsystemen hinaus halten zu können, um beispielsweise Analysen über die Zeit anzustellen sowie Auswertungen zu reproduzieren.

Inmon fordert hier allerdings eine spezielle Zeitbetrachtung – die Momentaufnahmen zu vorgegebenen Zeitpunkten – und eine besondere Form der Datenhaltung, die mit einer schubweisen Beschickung des Data Warehouse einhergeht. Hinzu kommt, dass mit größeren Zeitintervallen zwischen den Beschickungen die Daten im Data Warehouse weniger aktuell sind.

Neben der Historisierung sind drei weitere Zeitaspekte aus Praxissicht zu berücksichtigen.

Es gibt im zentralen Datenlager auch Bedarf nach

- zeitnahen Daten,
- effizienter Verwaltung von Daten, die sich im Betrachtungszeitraum nicht ändern (konstante Daten),
- Daten, von denen „nur“ die gegenwärtigen nicht jedoch die historischen Werte von Interesse sind.

Zum Bedarf nach integrierten Daten und Sichten, die relativ zeitnah oder sogar fast „real-time“ sind:

Dies ist immer dann der Fall, wenn kurze Reaktionszeiten gefordert sind wie beispielsweise bei der Analyse von Aktienkursen und im Kundenmanagement. So ist von dem Handelsunternehmen Wal-Mart bekannt, dass es Teile seines Data Warehouse im Minutentakt aktualisiert. Ist es in solchen Fällen nicht sinnvoll, von den i.d.R. langsam getakteten, schnappschussorientierten Datenübernahmen zu einem nahezu steten Datenstrom zu gelangen, der den bisherigen Datenbestand laufend ergänzt?

Zur effizienten Handhabung von konstanten Daten:

Wie verhält es sich mit Daten, die zeitlebens unverändert bleiben? Warum sollen zeitlich stabile Datenobjekte wie z.B. chemisch-physikalische Produkteigenschaften oder fertiggestellte Dokumente in eine Zeitstruktur hineingepresst werden?

Zur gegenwartsbetonten Sicht auf einige Daten:

Aus dem GI-Arbeitskreis „Konzepte und Techniken des Data Warehousing“ berichten Lehner und Bauer über die potenziellen Entwicklungen: „Interessanterweise tritt der explizite Historienbezug ... zum Teil in den Hintergrund, da auch integrierte, analyseorientierte Informationssysteme ohne explizite Historisierung gefragt sind“ [11]. Derartige Datenbestände könnten gemäß der Inmon'schen Forderung kein Bestandteil eines Data Warehouse sein.

Festzuhalten bleibt, dass die schnappschussorientierte Beschickung des zentralen Datenlagers eine z.T. erhebliche redundante Datenspeicherung bewirken kann und nicht alle Anforderungen an zeitnahe Daten abdeckt. Diese Schwierigkeiten ließen sich begriffskonform umgehen, wenn man in der Data-Warehouse-Definition die Zeit als einen der zuvor behandelten Integrationsaspekte verstünde, anstatt „time-variant“ im Sinne von Inmon zu fordern. Interessanterweise ist die Zeit der einzige Integrationsaspekt, den Inmon für wert hält, in seine Data-Warehouse-Definition explizit aufzunehmen.

Es stellt sich somit die Frage, ob das Charakteristikum time-variant als conditio sine qua non in die Definition eingehen soll oder ob der Zeitaspekt nicht bereits hinreichend durch die Integrationsforderung mit abgedeckt ist. Dies hieße dann auch, dass die schnappschussorientierte Ablage im Data Warehouse kein Muss sondern eine von mehreren Möglichkeiten ist, zeitabhängige Daten zu halten.

... nonvolatile ...

Inmon beschreibt die Dauerhaftigkeit wie folgt: “There are only two kinds of operations in the data warehouse: the initial loading of the data, and the access of the data ... but once the snapshot is made, the data in the data warehouse does not change. It is nonvolatile. ... There is no update of data (in the general sense of update)...” [7]. Und aufgrund seiner schnappschussorientierten Interpretation der Zeitorientierung folgert er: “Since data in the data warehouse are snapshot data it cannot be updated” [8].

Das heißt: keine inhaltliche Veränderung der Daten im Data Warehouse nach der Übernahme bis auf Korrekturen; diese gesteht Inmon noch zu.

Inmon begründet sein Verbot zum einen damit, dass dann keine Änderungsanomalien auftreten. Zum anderen sieht er vor allem eine wesentlich einfachere Technologie: Backup und Recovery seien nicht notwendig, ebensowenig Vorkehrungen zur Transaktions- und Datenintegrität.

Somit gibt es inhaltliche und technische Gründe für Inmon; beide treffen zu. Nur welche Relevanz haben diese beiden Gründe und wie stark sind sie heute zu gewichten? Der inhaltliche Grund, die Änderungsanomalie, wird hinfällig, wenn der Inhalt des Data Warehouse normalisiert strukturiert wird und somit frei von Redundanz ist. Die wegen der Zugriffsoptimierung ggf. notwendige Denormalisierung erfolgt dann in der Datenbasis für die Auswertungen, den Data Marts (s. Abb. 1). Die von Inmon aufgeführten technischen Gründe greifen allenfalls bei Massendaten mit möglichen Performanzproblemen. Durch geschickte Datenorganisation (Beschränkung auf die Aktualisierung geänderter Daten in normalisiert strukturierten Datenablagen anstatt der schnappschussorientierten Speicherung) und durch den Einsatz von leistungsfähigen Datenverwaltungssystemen lassen sich diese Probleme aber grundsätzlich lösen.

Einen weiteren Grund für das Aktualisierungsverbot nennt Holthuis mit der dann möglichen Reproduktion von Auswertungen [5]. Die Reproduktion lässt sich jedoch auch ohne Aktualisierungsverbot erreichen. Denn mit der vierten Schicht der Data-Warehouse-Architektur, bestehend aus den Data Marts, lassen sich in Verbindung mit einer Historisierung der geänderten Daten im Data Warehouse Teilsichten auf das Data Warehouse erzeugen. Wenn auf diesem Datenbestand die Auswertungen erfolgen, können sie jederzeit mit gleichen Ergebnissen wiederholt werden.

Bastian interpretiert die Dauerhaftigkeit anders: “... data is stored persistently. Access is only by reading the data; analysis does not change the data” [1]. Wenn “change” bei Bastian auch “insert” umfasst, dann schließt Bastian aus, dass im zentralen Datenlager Analyseergebnisse (z.B. Ergebnisse aus komplexen Data-Mining-Prozessen basierend auf Daten des zentralen Datenlagers) in dieses zurückfließen. Können Analyseergebnisse aus dem Data Warehouse bzw. den Data Marts nicht als weitere Datenquelle betrachtet werden? Dann könnten auch derartige Daten ihren Weg über den Extract-Transform-Load-Process (ETL-Process) in das Data Warehouse finden. Ist nicht die Konsistenz des zentralen Datenlagers das Entscheidende? Wenn dem so ist, welchen Grund gibt es dann noch – solange die Konsistenz gewahrt bleibt – das Aktualisierungsverbot – ob von Inmon oder von Bastian – aufrechtzuerhalten?

Hinzu kommt, dass die dauerhafte Speicherung der Daten wegen des Platzes und der damit verbundenen längeren Zugriffszeiten ihre Grenzen hat. Daher wird eine gelegentlich durchzuführende teilweise Auslagerung der Daten (Archivierung) erforderlich, um diese im Data Warehouse löschen zu können.

Es ist somit kein hinreichender Grund ersichtlich, das Aktualisierungsverbot als Charakteristikum des Data Warehouse bestehen zu lassen. Daher sollte auch diese Eigenschaft ersatzlos aus der Definition gestrichen werden.

. . . collection of data . . .

Mit der Sammlung von Daten wird zum einen der zu integrierende Gegenstand, die Daten, festgelegt. Zum anderen wird mit der Sammlung von Daten assoziiert, dass die zu integrierenden Objekte nach unterschiedlichen Aspekten zusammengetragen werden können. So kann neben dem bereits aufgeführten inhaltlichen und dem zeitlichen Aspekt z.B. auch nach dem organisatorischen und vor allem dem räumlichen Aspekt integriert werden. Die Datenquellen können regional verteilt sein. Besonders in einer Zeit, wo für die Unternehmen Globalisierung als immer wichtiger erachtet wird, ist es notwendig, Daten aus unterschiedlichen Lokalitäten zusammenzuführen.

. . . in support of management's decision-making process

Die Entscheidungsunterstützung des Managements ist der einzige Zweck, den Inmon für das Data Warehousing vorsieht. Das heißt, die Anwender sind im wesentlichen Manager und – was in der heutigen Praxis eher der Fall sein dürfte – diejenigen, die dem Manager zuarbeiten wie z.B. Controller und Mitarbeiter von Planungsstäben. Warum strapaziert Inmon mit dieser Beschränkung auf eine bestimmte Nutzergruppe hier eine Klassengesellschaft? Oder anders gefragt: Warum soll das Data Warehousing nicht auch den Bedarf von anderen Werktätigen an integrierten Daten befriedigen?

Der andernorts, z.B. bei Bauer und Günzel [2], auftauchende Analyse-Zweck ist zwar schon weiter gefasst als der Management-Unterstützungs-Zweck, aber können nicht auch das Bereitstellen von einzelnen Fakten bei Auskunftssystemen sowie Synthese-Zwecke geltend gemacht werden, wie z.B. bei der Erstellung eines Kataloges über alle Vertriebsprodukte eines international tätigen Chemieunternehmens mit mehreren Produktionsstätten?

Warum beschränkt Inmon den Zweck eines Data Warehouse auf die Entscheidungsunterstützung des Managements? Sollte nicht vielmehr ein Data Warehouse immer dann erstellt werden, wenn eine integrierte Sicht auf bestimmte Datenobjekte erforderlich wird, und sollte dies nicht unabhängig davon sein, wozu die integrierte Sicht dient? Somit wäre es wegen der grundsätzlichen Beliebigkeit des Zwecks unnötig, den Zweck in die Data-Warehouse-Definition mit aufzunehmen, in Analogie zu dem föderierten Datenbanksystem, bei dessen Definition auf jegliche Zwecknennung verzichtet wird.

4 Weitere restriktive Aspekte beim Data-Warehouse-Begriff

Es gibt in der Literatur noch weitere einschneidende Beschränkungen für das Data Warehousing.

OLAP bzw. Multidimensionalität

Erst nachdem Inmon seine Data-Warehouse-Definition publiziert hat, ist der OLAP-Begriff von Codd geschaffen worden. Um so erstaunlicher ist es, dass nicht wenige Theoretiker wie Praktiker im OLAP – also i.W. dem Halten und Auswerten von multidimensional strukturierten Daten – die entscheidende oder sogar einzige Funktion des Data Warehousing sehen. Diese eingeschränkte Sicht vertritt z.B. Kurz: „Ein Data Warehouse wird immer multidimensional modelliert“ [10]. Ist es stattdessen nicht zweckmäßiger, in OLAP einen Teilbereich – wenn auch nicht den unwichtigsten – des Data Warehousing zu sehen? Daten lassen sich nicht immer in multidimensionale Strukturen – also Matrizen höherer Ordnung – pressen. Hinzu kommt, dass die OLAP-Operatoren nur auf genau eine Matrix wirken und Verknüpfungen von zwei oder mehr Matrizen durch Verbundoperationen bisher nur ansatzweise unterstützt werden. Zieht man sich hier nicht unnötig auf bestimmte Daten wie auch auf bestimmte Funktionen zurück und bringt man sich nicht um etliche weitere Möglichkeiten, die integrierte Sichten und Verknüpfungsoperatoren bieten?

Dispositive Daten

In enger Verbindung mit der Zweckbeschränkung auf die Erfüllung der Managementbedürfnisse wird von Inmon wie auch von fast allen anderen Autoren die Typisierung der Daten in einerseits „operativ“ (“is used to run the business and is related to short term actions or decision”) und andererseits „dispositiv“ bzw. „informational“ (“is used to manage the business in the longer term” [4]) gesehen.

Wird heute im Kontext Data Warehousing die Unterscheidung in operative und dispositive Daten überhaupt noch benötigt? Ist die Beschränkung auf die dispositiven Daten nicht sogar kontraproduktiv? Ob ein Datum operativ oder dispositiv ist, hängt davon ab, wie es genutzt wird. Es ist somit keine Eigenschaft des Datums per se, sondern des Umgangs mit dem Datum. Und je besser – genauer gesagt vielseitiger – die Anwendungen erstellt werden, um so weniger wird diese Unterscheidung zumindest im Kontext Data Warehousing relevant. So schreibt Imping: „Das Data Warehouse schreit geradezu danach, als Datenplattform für Business-Applikationen genutzt zu werden“ [6]. Was hindert daran, gemeinsam mit Imping danach zu rufen?

Vollständigkeit

Gelegentlich taucht in der Literatur die Forderung nach Vollständigkeit der Daten auf. Das Bezugsobjekt ist bei einigen Autoren das „subject“, für andere das für das Management notwendige Spektrum der Daten und nicht selten sogar der gesamte Datenbestand des Unternehmens. So schreibt z.B.

Holthuis: „ . . . Datenbank, die als unternehmensweite Datenbasis für das gesamte Spektrum managementunterstützender Informationssysteme dient“ [5]; und auch Schrempf sieht im Data Warehouse „eine zentrale Datenbank, in der alle Daten, die im Unternehmen anfallen, gesammelt und archiviert werden“ [13]. Selbst Devlin entzieht sich nicht dem Universalgedanken: “. . . data warehouse with his enterprise-wide scope defines the size of the data warehouse. He is big – very big” [4].

Dieser universelle Anspruch scheint überzogen und nach Kenntnis des Autors erfüllt ihn auch kein Unternehmen. Wenn in den meisten Firmen noch nicht einmal das unternehmensweite Datenmodell erstellt werden konnte, wie sollte dann ein unternehmensweiter Datenbestand als Data Warehouse existieren können?

Inmon meint hierzu: “Only data that is needed for DSS processing finds its way into the data warehouse environment” [7]. Da der von Inmon postulierte Zweck die Erfüllung des Managementbedarfs ist, beschränkt er sich auf die dafür notwendigen Daten. Dieser Gedanke ist konsequent. Hinzu kommt, dass Inmon die Themenorientierung fordert. Diese würde sich erübrigen, wenn er den Vollständigkeitsanspruch erhöbe.

Sollte nicht vielmehr allein der Zweck, der mit dem an den Unternehmenszielen orientierten betrieblichen Bedarf einhergeht, Inhalt und Umfang des zentralen Datenlagers bestimmen? Dann müsste die Devise lauten: “Not for management needs only – but for business needs in general!”

Einmaligkeit

Idealtypisch darf es in einem Unternehmen nur ein einziges Data Warehouse geben, das die integrierte Sicht auf alle relevanten Daten des Unternehmens liefern soll. Dies mag in kleineren und mittleren Unternehmen auch machbar sein; in Großunternehmen, bei denen die einzelnen Sparten das Sagen haben oder bei denen Tochterunternehmen relativ selbstständig agieren können, scheint heute das eine Data Warehouse mit seiner globalen Sicht eine Illusion zu sein. Wider alle Integrationsbemühungen des Datenmanagements entstehen – abhängig von der Firmenkultur – in Großunternehmen häufig mehrere Data Warehouses, die an Organisationsbereichen (z.B. Sparten, Tochterunternehmen) und/oder Funktionsbereichen (z.B. Forschung, Vertrieb) ausgerichtet sind. Solange es hierbei keine Konsistenzprobleme gibt, die Anwender ihre Anforderungen erfüllt sehen und die mehrfache Data-Warehouse-Infrastruktur wirtschaftlich vertretbar ist, kann gegen diese Praxis nur schwer argumentiert werden.

Betrachtet man die Ursachen (Selbstorganisation der Bereiche), die evolutionäre Entwicklung (Dynamik der Prozesse) und ihre Ausprägungen (Selbstähnlichkeit), so drängt sich die Analogie zum fraktalen Unternehmen auf. Struktur und Aufgabe im Kleinen ähneln denen im Großen. Folgt man diesem Gedanken, so kann man bei der zuvor geschilderten Praxis vom „fraktalen Datenlager“ (“fractal data warehouse”) sprechen.

5 Vorschlag für einen modifizierten Data-Warehouse-Begriff

Der restriktiven Definition von Inmon wird eine breit gefasste Definition des Data-Warehouse-Begriffs gegenübergestellt:

> *Ein Data Warehouse ist ein physischer Datenbestand, der eine integrierte Sicht auf die zugrunde liegenden Datenquellen ermöglicht.*

Zur Erläuterung dieser Definition wird angeführt:

- Die Einschränkung „physisch“ ist notwendig, um das Data Warehouse von dem „logischen“ föderierten Datenbanksystem abzugrenzen.
- Mit der Forderung nach „integrierter Sicht“ auf zugrunde liegende Datenquellen unterscheidet sich das Data Warehouse von der allgemeinen Datenbank wie auch von einem „bunt zusammengetragenen Datenhaufen“.
- Das Data Warehouse dient nicht nur dem Management. Der Zweck des Data Warehouse ist beliebig. Daher wird der Aspekt Zweck in die Definition nicht aufgenommen. Die möglicherweise unterschiedlichen Zwecke sind für ein konkretes Data Warehouse allerdings zu definieren und sie können jederzeit erweitert werden.
- Welche Daten (von betrieblicher Kennzahl bis hin zur Videosequenz) in welchem Umfang (Zeithorizont und räumliche Herkunft) im Data Warehouse gehalten werden, wird durch den Bedarf der Anwender bestimmt. Daher werden weder inhaltliche, zeitliche noch räumliche Aspekte mit in die Definition aufgenommen.
- Die Strukturierung des Datenbestands ist nicht notwendigerweise schnappschussorientiert; das Schema des zentralen Datenlagers sollte vielmehr anwendungsneutral sein, wenn die Daten mehrfach genutzt werden.
- Der Anspruch an die Beschaffenheit der Daten (Datenqualität) – insbesondere an die Zeitnähe (Aktualität) der Daten und ihre Verfügbarkeit – ergibt sich ebenfalls aus dem Bedarf.
- Die „integrierte Sicht“ impliziert die Forderung nach konsistenten Daten, also der Widerspruchsfreiheit der Daten untereinander wie auch zu ihrer Beschreibung. Daher ist es nicht nötig, in der Definition Aktualisierungsverbote aus Konsistenzgründen zu fordern.
- Der Datenbestand im Data Warehouse kann physisch verteilt sein; der Anwender sieht ihn aber als Einheit (logisch zentriert gemäß dem gemeinsamen Schema).
- Das Data Warehouse kann ganz oder teilweise als “single point of truth” (SPOT) definiert werden.
- Das Data Warehouse ist die Ausgangsbasis für die Beschickung von Informations- und Analysesystemen wie auch mögliche Datenquelle für prinzipiell beliebige Anwendungssysteme.
- Die jeweilige Nutzung des Data Warehouse bestimmt die Datenhaltungsarchitektur und in Verbindung mit den das Data Warehouse nutzenden Anwendungssystemen werden die Lastverteilung und damit die Netz- und Rechnerarchitektur (z.B. als Client-Server-Architektur) festgelegt.
- Das Data Warehouse muss nicht auf einem elektronischen Träger gehalten werden; auch wird nicht gefordert, dass es von einem Datenbankmanagementsystem verwaltet wird. So kann nach der obigen Definition das Manuskript der

Gebrüder Grimm – also die zusammengetragene und aufbereitete Sammlung von Märchen – als ein Data Warehouse auf dem Träger Papier betrachtet werden.

- Das Data Warehouse kann auch direkt (über Massenänderungen hinaus) punktuell aktualisiert bzw. korrigiert und vervollständigt werden, wenn die Konsistenz gewahrt bleibt. Dies gilt insbesondere dann, wenn die Änderungen über den ETL-Process erfolgen und das zugrundeliegende Datenverwaltungssystem über ein Transaktionskonzept und über Möglichkeiten von Restart und Recovery verfügt. Entsprechendes gilt für die Datenentsorgung; d.h. der Datenbestand kann um nicht mehr benötigte Daten bereinigt werden. Die Reproduktion von Auswertungen muss dabei gewährleistet bleiben. Hierfür sind im Bedarfsfall Vorkehrungen zu treffen, so z.B. bei der Datenorganisation der Data Marts und für die Reaktivierung archivierter Bestände.
- Der modifizierte Data-Warehouse-Begriff umfasst auch den Begriff des Operational-Data-Store (ODS). Dies ist nach Schwinn ein Datenspeicher mit sowohl (fast) aktuellen sowie historischen Daten [14]. Der Data-Warehouse-Begriff von Inmon hingegen erlaubt wegen der Schnappschussausrichtung keinen steten und somit zeitnahen Datenstrom bei der Beschickung des Data Warehouse. Um zeitnahe Daten in der Data-Warehouse-Umgebung verfügbar zu machen, gibt es bei Inmon verschiedene Architekturansätze für einen vom Data Warehouse getrennten Datenspeicher, den ODS.
- Die „integrierte Sicht“ schließt mit ein, dass Inhalt, Struktur und Herkunft der Daten im Data Warehouse für den Benutzer (Entwickler wie Endanwender) offen liegen müssen. Daher kommt dem Repositorium besondere Bedeutung zu.

Das Charakteristische des Data-Warehouse-Begriffs ist die Datenintegration auf Schema- wie auch auf Datenebene. Diese wird durch eine Nabe-Speiche-Architektur mit dem Data Warehouse, dem zentralen Datenlager als Nabe, und den Datenquellen wie auch den funktional oder bereichsspezifisch orientierten Extrakten, den Data Marts, als Speichen unterstützt. Die Staging Area hat dabei Aufbereitungs- und Pufferfunktion. Aus informationslogistischer Betrachtung sorgt somit das Data Warehouse in Verbindung mit dem Repositorium für die Datenbereitstellung, während die einzelnen Data Marts der Datennutzung dienen.

Aufgabe des Datenmanagers ist es, die Datenlandschaft zu organisieren, das heißt hier vor allem, den Bebauungsplan zu gestalten und für einen sicheren und reibungslosen Betriebsablauf zu sorgen, so dass die Benutzeranforderungen funktional, qualitativ und wirtschaftlich erfüllt werden. Dazu muss der Datenmanager von dem heute oft chaotischen Beziehungsgeflecht mit den vielen mehr oder weniger komplexen direkten Schnittstellen zwischen den einzelnen Datenquellen und den für die Auswertungen notwendigen Datenbeständen wegkommen. Er muss auf eine geordnete Datenhaltung und einen Datenfluss hinarbeiten, der im Falle integrierter Sichten grundsätzlich über das Data Warehouse als Informationsdrehscheibe geht.

Für diesen Architekturansatz gibt es mindestens drei Gründe:

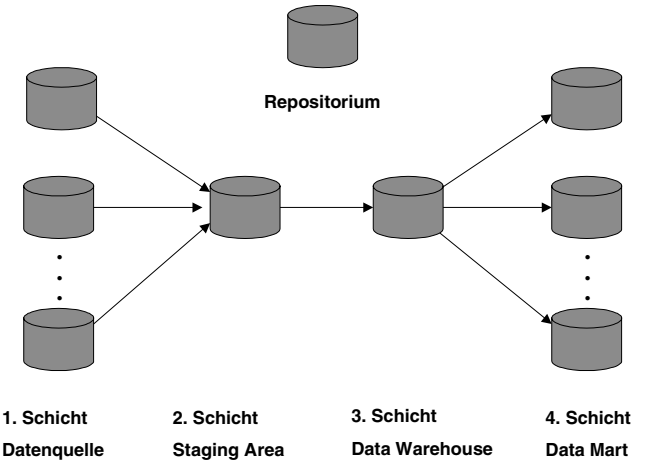

Abb. 1. Grobdarstellung der vierschichtigen Datenarchitektur eines Data-Warehouse-Systems angelehnt an Bauer, Günzel [2]

- Die Anzahl der Schnittstellen zum und vom Data Warehouse ist geringer als die Anzahl der Direktverbindungen zwischen den Datenquellen und den Datenbeständen für die Auswertungssysteme (Data Marts). Damit wird die Datenlandschaft weniger komplex und eher beherrschbar und letztendlich vor allem infolge der mehrfachen Nutzung von Daten im Data Warehouse durch unterschiedliche Anwendungen i.d.R. wirtschaftlicher.
- Da die Daten in einem einheitlichen und konsistenten Datenbestand, dem Data Warehouse, über die Staging Area mit ihrer Datenbereinigungsfunktion zusammengefasst werden, können die qualitativen Anforderungen an die Daten besser erfüllt werden.
- Mit den Data Marts als Extrakte des Data Warehouse lassen sich Datenbestände schaffen, die auf die funktionalen Anforderungen einzelner Anwender oder Bereiche zugeschnitten sind.

Das Data Warehousing kann somit vom Datenmanagement als ein Organisationskonzept für zeitlich, räumlich und organisatorisch verteilt anfallende Daten verstanden werden, die an anderen Stellen gesamtheitlich benötigt werden. Die Hauptfunktionen des Data Warehouse sind somit das Sammeln, Aufbereiten und Zusammenführen/Vorhalten von Daten und ihren Beschreibungen an einem definierten Platz und das Verteilen auf die Data Marts und damit Verfügbarmachen für möglichst viele Nutznießer. Diesem Organisationskonzept verspricht die neue Data-Warehouse-Definition eher gerecht zu werden als die Definition gemäß Inmon.

Die Bedeutung des Data Warehouse gemäß der modifizierten Definition dürfte nach Einschätzung des Autors auch im Kontext von Portalen wachsen. Das Data Warehouse kann eine wesentliche Komponente eines Portals werden, wenn neben der Integration von Anwendungen (EAI) die Integration der Daten eine Rolle spielt. Ein Vergleich zur Objektorientierung drängt sich auf: So wie die objektorientierte Modellierung nach Verständnis des Autors die systematische Erweiterung der Datenmodellierung ist, so kann die Applikationsintegration (wie im Portal) als Erweiterung der Datenintegration (wie im Data Warehouse) angesehen werden. Somit wäre dann in der Evolution der IT nach dem Data Warehouse das Portal eine neue Spezies.

6 Erweiterung des Anwendungsspektrums – Beispielsammlung

Befreit von den Inmon'schen Restriktionen ist das Spektrum der möglichen Data-Warehouse-gestützten Anwendungen erheblich breiter. Um die Breite zu skizzieren, werden bewusst einige Anwendungen am Rande des Spektrums aufgeführt. Alle Beispiele verletzen mindestens eine der Inmon'schen Forderungen; somit ist keines der Beispiele nach dem Verständnis von Inmon ein Data Warehouse. Auch hat keines der Beispiele etwas mit OLAP zu tun wie auch kaum ein Datum dispositiv verwendet wird.

6.1 Qualitätssicherung Produktion: Referenzsystem für Dokumente

Mehr als 100 Datenhaltungssysteme (z.B. Lotus Notes Files) mit Dokumenten aus dem Bereich Qualitätssicherung in der Produktion eines Automobilkonzerns mussten für das Wiederauffinden und Anzeigen einheitlich verfügbar gemacht werden. Um dem Endanwender diese integrierte Sicht zu ermöglichen, wurde für die Suche ein Data Warehouse mit allen Werten zu den Suchattributen sowie den Verweisen auf die zugehörigen Dokumente beschickt. Für die Anzeige der gefundenen Dokumente wurde eine Zugriffsschicht auf die verteilten Datenhaltungssysteme geschaffen. Über diese Schicht wird auch die Aktualisierung für die Indizierung im Data Warehouse gesteuert.

6.2 IT-Infrastruktur: Zentrales Repositorium als Informationsdrehscheibe

Nußdorfer schlägt ein Data Warehouse als zentrale Ablage von Begriffsdefinitionen, konzeptionellen / logischen Datenmodellen, realisierten / physischen Datenbankschemata und Anwendersichten vor [12]. Dabei wird das Data Warehouse im Wesentlichen von den anwendungsspezifischen Dictionaries bzw. Repositorien in einer heterogenen Anwendungslandschaft mit ihren jeweiligen Begriffswelten gespeist. Es verfolgt den Zweck, einen geordneten Datenhaushalt zu erreichen und allen Mitarbeitern des Unternehmens als Auskunftssystem zu dienen.

6.3 Forschung: Integrierte Untersuchungsergebnisse in der Pharmaforschung

In der Präklinischen Forschung eines weltweit agierenden Pharmakonzerns wurden die in den lokalen Forschungslabors anfallenden Untersuchungsergebnisse einschließlich ihrer jeweiligen Untersuchungsbedingungen in einem Data Warehouse unmittelbar nach ihrer Validierung zusammengeführt, um sie in allen Pharmaforschungsstätten verfügbar zu machen. Mit diesen vollständigen und zeitnahen Daten können alle Pharmaforscher des Konzerns gesamtheitliche Betrachtungen anstellen.

6.4 Querschnitt: Mitarbeiterdaten weltweit

Der gleiche Konzern nutzt ein so genanntes Metadirectory, also eine besonders performante Datenbank mit standardisierter Zugriffsmethode LDAP, um sowohl bestimmte Personaldaten als auch administrative Daten über alle Niederlassungen strukturiert und standardisiert verfügbar zu haben. Dieses zentrale Datenlager dient als Informationsdrehscheibe der Bereitstellung aktueller Daten für übergreifende Auswertungen und verbessert im Umfeld der Administration die Übersicht über Zugriffslegitimation und Lizenzen. Weiter ist es durch die Synchronisierung von Zugriffsberechtigungen der Nutzer auf verschiedene Anwendungen Basis für ein zukünftiges Single-Sign-On.

6.5 Querschnitt: Produktdaten weltweit

Ein anderer Chemie- und Pharmakonzern setzt ein separates SAP R/3 ausschließlich für die Integration von Materialdaten (i.W. Konvertierung von Material-Identifikationen und Berücksichtigung lokal unterschiedlicher Stücklisten) aus den diversen lokalen Produktions- und Lagerstätten ein. Damit gewinnen Mitarbeiter aus Entwicklung, Produktion und Logistik einen Überblick über alle Produkte weltweit.

References

1. Bastian, M.: Analytical Information System, RWTH Aachen, Vorlesungsunterlagen WS 2002/2003
2. Bauer, A.; Günzel, H.: Data-Warehouse-Systeme. Heidelberg: dpunkt 2000
3. Conrad, S.: Föderierte Datebanksysteme – Konzepte der Datenintegration. Berlin: Springer 1997
4. Devlin, B.: Data Warehouse: from Architecture to Implementation. Massachusetts: Addison-Wesley 1997
5. Holthuis, J.: Der Aufbau von Data-Warehouse-Systemen: Konzeption – Datenmodellierung – Vorgehen.Wiesbaden: Gabler 1999
6. Imping, K.: ROI-Gebot: Data Warehousing nicht nur für Reporting nutzen!, IT-Director, 6-6 (06 2002)
7. Inmon, W.H., Hackethorn, R.D.: Using the Data Warehouse. New York: John Wiley & Sons 1994
8. Inmon, W.H.: Shnapshots in the Data Warehouse. http://63.170.41.42/library/whiteprs/earlywp/ttsnap.pdf: 2000
9. Jung, R.; Winter, R.: Data Warehousing: Nutzungsaspekte, Referenzarchitektur und Vorgehensmodell in Data Warehousing Strategie. Berlin, Heidelberg: Springer, 2000
10. Kurz, A.: Data Warehousing Enabling Technology. Bonn: MITP-Verlag 1999
11. Lehner, W.; Bauer, A.: Data-Warehouse-Systeme – derzeitiger Stand und aktuelle Entwicklungen, Datenbank-Spektrum, Heft 4, 76-78 (2002)
12. Nußdorfer, R.: MetaData-Strategy; it FOKUS, 8-16 (11 1999)
13. Schrempf, M.: Shopping im Data Warehouse – Alter Wein in neuen Schläuchen? it Management, 28-33 (07/08 1995)
14. Schwinn, K.; Dippold, R.; Ringgenberg, A.; Schnider, W.: Unternehmensweites Datenmanagement. Wiesbaden: Gabler 1999
15. Stock, S.: Modellierung zeitbezogener Daten im Data Warehouse. Wiesbaden: Gabler 2001