Christoph Schütz studierte Wirtschaftsinformatik an der Johannes Kepler Universität (JKU) Linz. Seine Diplomarbeit mit dem Titel „Extending data warehouses with hetero-homogeneous dimension hierarchies and cubes: A proof-of-concept prototype in Oracle“ belegte beim TDWI Award 2011 den ersten Platz. Derzeit arbeitet er am Institut für Wirtschaftsinformatik – Data & Knowledge Engineering der JKU als wissenschaftlicher Mitarbeiter. Seine Forschungsschwerpunkte liegen in den Bereichen Data Warehousing, Prozessmodellierung und Informationssicherheit.
E-Mail: christoph.schuetz@jku.at

Modellierung

Hetero-homogene Data Warehouses

Data Warehouses sind strategische Informationssysteme, die Daten aus mehreren operativen Quellen zusammenführen, um Entscheidungsträger bei ihrer Analysetätigkeit zu unterstützen. Die verschiedenen operativen Datenquellen sind oftmals heterogen. Die Integration der operativen Datenquellen in einem zentralen Data Warehouse bedarf zumeist der Beseitigung dieser Heterogenitäten. Dadurch geht jedoch wertvolle Information verloren, die andernfalls die Qualität der Analyse verbessern könnte. Die Berücksichtigung eventuell vorhandener Heterogenitäten im konzeptuellen Datenmodell ist deshalb wünschenswert. Bestehende Ansätze zur konzeptuellen Modellierung von Data Warehouses lösen das Problem der Einbeziehung heterogener Information jedoch nur unzureichend. Der hetero-homogene Modellierungsansatz erlaubt demgegenüber die Berücksichtigung zusätzlicher heterogener Information in einem grundsätzlich homogenen Schema. Durch Softwareunterstützung wird die Umsetzung eines hetero-homogenen Datenmodells in einem objekt-relationalen Datenbankverwaltungssystem erleichtert.

Modellierung homogener Data Warehouses

Die konzeptuelle Darstellung von Data Warehouses umfasst hierarchisch strukturierte Dimensionen und OLAP-Würfel (Cubes). Verschiedene konzeptuelle Modellierungsansätze existieren [Hah10], die verbreiteten Methoden erfordern jedoch zumeist die Beseitigung von Heterogenitäten oder bieten nur unzureichende Semantik für deren Darstellung. Das Dimensional Fact Model (DFM) [GMR98] stellt einen dieser Modellierungsansätze dar. Ausgehend von einem konzeptuellen Modell der operativen Daten erstellt der Modellierer hierarchisch strukturierte Dimensionen und Fakten. Die Fakten sind betriebswirtschaftliche Sachverhalte der Analyse, zum Beispiel Verkäufe, und werden durch Kennzahlen, zum Beispiel Umsatz, beschrieben. Eine Analyse auf unterschiedlichen Abstraktionsebenen ist durch Verdichtung (Roll-up) oder Verfeinerung (Drill-down) der Fakten entlang der Aggregationsstufen der Dimensionshierarchien möglich. Die Einbeziehung heterogener Daten im DFM, zum Beispiel optionale Aggregationsstufen und Kennzahlen, verschieden granulare Daten, erhöht die Komplexität der Modellierung und verringert die Qualität der Analyse.

Modellierung hetero-homogener Data Warehouses

Der hetero-homogene Modellierungsansatz für Data Warehouses bewahrt die Vorteile eines homogenen Schemas, ohne auf die Flexibilität einer heterogenen Modellierung zu verzichten. Die Idee des hetero-homogenen Data Warehouse wurde von Neumayr et al. [NST10] vorgeschlagen. Hetero-homogene Data Warehouses verlangen die Modellierung eines grundsätzlich homogenen Schemas für Dimensionen und Würfel, erlauben jedoch die Einbeziehung zusätzlicher Informationen für einzelne, wohldefinierte Teilbereiche des Analysegebiets, zum Beispiel bestimmte Unternehmensbereiche oder Analysezeiträume. Jeder dieser Teilbereiche, obwohl in Bezug auf das globale Schema heterogen, folgt wiederum einem homogenen Schema.

Multilevel Objects (M-Objects) sind das Kernstück des hetero-homogenen Data-Warehousing-Ansatzes. M-Objects wurden ursprünglich für die Darstellung von Konzepthierarchien entwickelt [NGS09], eignen sich aber auch hervorragend für die konzeptuelle Modellierung von Dimensionshierarchien in Data Warehouses [NST10].

Ein M-Object hat mehrere Ebenen (Levels). Jede Ebene wird durch Attribute beschrieben. Die Ebenen eines M-Object stellen Klassen unterschiedlicher Abstraktionsgrade dar, die untereinander in einer Aggregationsbeziehung stehen. In Zusammenhang mit der Modellierung von Data Warehouses entsprechen die Ebenen eines M-Object den Aggregationsstufen einer Dimension.

Das M-Object *Produkt* in Abbildung 1 definiert das Schema einer homogenen Produktdimension. Die Produktdimension enthält die Aggregationsstufen *kategorie* und *modell*. Jedes Modell ist genau einer Produktkategorie zugeordnet. Jede Produktkategorie hat einen verantwortlichen Manager, jedes Modell hat einen Listenpreis.

Ein M-Object definiert nicht nur mehrere Ebenen, ein M-Object ist auch Ausprägung seiner allgemeinsten Aggregationsstufe. Den Attributen dieser Aggregationsstufe

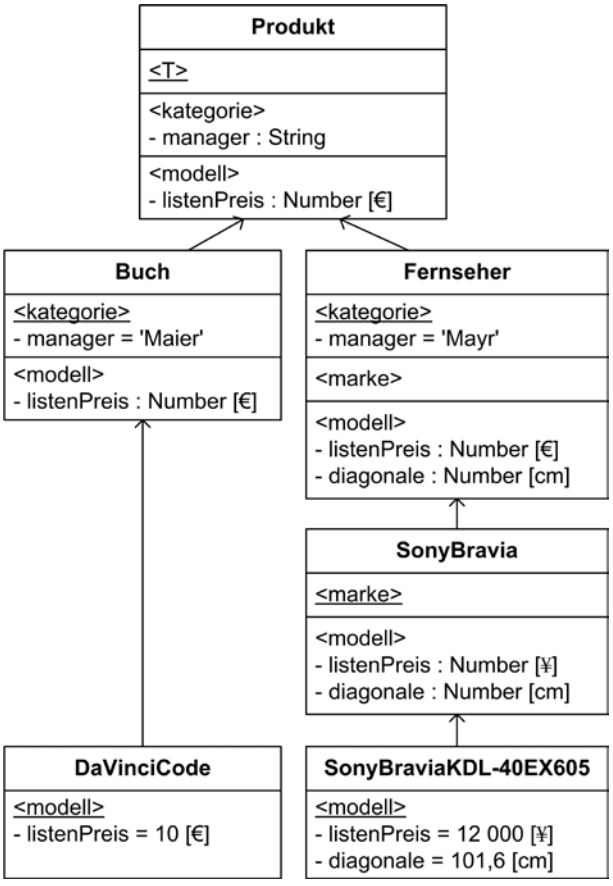


Abb. 1: *Hetero-homogene Produktdimension (in Anlehnung an [NST10])*

weist ein M-Object Werte zu. Für alle anderen Ebenen definiert das M-Object Schemainformationen. Für die Data-Warehouse-Modellierung ergibt sich daraus ein Dimension-Ebene-Dualismus: Ein M-Object ist gleichzeitig Ausprägung einer bestimmten Aggregationsstufe und (Teil-)Dimension.

Durch **Konkretisierung** von M-Objects werden Dimensionsschemata spezialisiert und Aggregationsstufen ausgeprägt. Die Konkretisierung eines M-Object ist wiederum ein M-Object, allerdings auf einer detaillierteren Aggregationsstufe als das konkretisierte M-Object, und steht mit dem konkretisierten M-Object in einer Teil-Ganzes-Beziehung. Eine Konkretisierung erbt außerdem Aggregationsstufen und Attribute, kann dieses ererbte Schema jedoch spezialisieren, indem es zusätzliche Ebenen und Attribute hinzufügt. M-Objects, die untereinander in einer Konkretisierungsbeziehung stehen, bilden eine Dimension.

Die M-Objects *Buch* und *Fernseher* in Abbildung 1 sind Konkretisierungen von *Produkt*. Sie sind Ausprägungen der Aggregationsstufe *kategorie*, sind aber auch Teildimensionen der Produktdimension. Das M-Object *Fernseher* führt eine zusätzliche Abstraktionsebene *marke* ein. Alle Fernsehermodelle sind deshalb einer Fernsehermarke zugeordnet, alle Fernsehermarken sind Teil der Produktkategorie *Fernseher*. Zu jedem Fernsehermodell wird zusätzlich zum Listenpreis die Länge der Bildschirmdiagonale erfasst.

Die Konkretisierung von M-Objects ermöglicht die Einführung von Heterogenitäten in einer wohldefinierten Teildimension. Obwohl einzelne Teildimensionen zusätzliche Informationen erfassen können, garantiert eine hetero-homogene Dimension dennoch ein allen Teildimensionen gemeinsames, minimales Schema. Dieses Prinzip lässt sich rekursiv auf die Teildimensionen der Teildimensionen anwenden.

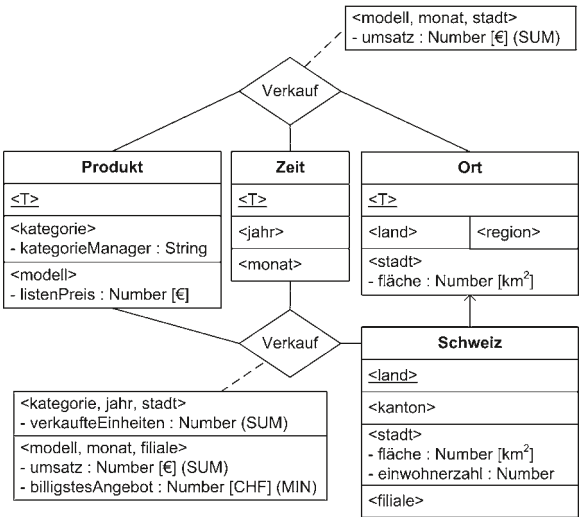
Multilevel Relationships (M-Relationships) verbinden M-Objects verschiedener Dimensionen und stellen so die Fakten eines OLAP-Würfels dar. Diese Fakten werden durch Kennzahlen quantifiziert. Daher werden jeder M-Relationship Kennzahlen zugeordnet. Jede Kennzahl hat eine Aggregatfunktion und eine Maßeinheit und wird auf einer bestimmten Granularität erfasst. In Zusammenhang mit Data Warehouses werden M-Relationships auch als **Multilevel Facts (M-Facts)** bezeichnet. Ein **Multilevel Cube (M-Cube)** ist eine Sammlung von M-Facts über vorgegebenen Dimensionen.

M-Relationships stehen in einer impliziten Konkretisierungsbeziehung, abgeleitet aus den verbundenen M-Objects. Konkretisierungen von M-Relationships können zusätzliche Kennzahlen einführen oder bestehende Kennzahlen auf einer feineren Granularität erfassen. Dadurch ist es möglich, heterogene Information in einem bestimmten, wohldefinierten Teilbereich eines OLAP-Würfels zu erfassen und gleichzeitig einem gemeinsamen, minimalen Schema zu folgen.

Abbildung 2a zeigt einen hetero-homogenen dreidimensionalen OLAP-Würfel für Produktverkäufe. Erfasst werden die Umsätze in Euro je Produktmodell, Monat und Stadt. Dieses minimale Schema gilt für den gesamten Würfel, kann jedoch in einzelnen Teilgebieten verfeinert werden. Für Verkäufe in der Schweiz wird zusätzlich je Produktkategorie, Jahr und Stadt die Anzahl der verkauften Einheiten erfasst. Umsätze werden auf einer feineren Granularität, je Filiale, erfasst. Außerdem wird zu jedem Produktmodell das billigste Angebot in Schweizer Franken in einem Monat je Filiale erfasst.

Prototyp-Implementierung in Oracle

Die bisherigen Betrachtungen zu hetero-homogenen Dimensionen und OLAP-Würfeln bezogen sich auf die konzeptuelle Data-Warehouse-Modellierung. Es stellt sich jedoch die Frage nach der Übertragung des konzeptuellen, hetero-homogenen Entwurfs in ein geeignetes logisches Datenmodell. Zur Unterstützung dieser Aufgabe wurde eine Prototyp-Implementierung für die Verwaltung hetero-homogener Data Warehouses entwickelt. Sie dient als Machbarkeitsstudie für den hetero-homogenen Data-Warehousing-Ansatz. Erste Überlegungen zu dieser Implementierung wurden bereits durch Neumayr



2a: Konzeptuelles Modell

<modell, monat, stadt>

Produkt	Zeit	Ort	Umsatz
DaVinciCode	Jan2011	Wien	€ 30 000
MobyDick	Jan2011	Wien	€ 45 000
DaVinciCode	Jan2011	Zürich	€ 5 000
MobyDick	Jan2011	Zürich	€ 20 000

abgeleitet

<kategorie, jahr, stadt> (Schweiz)

Produkt	Zeit	Ort	Einheiten
Buch	Jahr2011	Zürich	3 000

<modell, monat, filiale> (Schweiz)

Produkt	Zeit	Ort	Umsatz	Billigstes
DaVinciCode	Jan2011	ZürStore1	€ 2 000	CHF 10
DaVinciCode	Jan2011	ZürStore2	€ 3 000	CHF 12
MobyDick	Jan2011	ZürStore1	€ 10 000	CHF 18
MobyDick	Jan2011	ZürStore2	€ 10 000	CHF 16

2b: Faktentabellen

Abb. 2: *Hetero-homogener OLAP-Würfel für Verkäufe (in Anlehnung an [NST10])*

[Neu10] angestellt. Die Weiterentwicklung erfolgt im Rahmen eines internen Projekts des Instituts für Wirtschaftsinformatik – Data & Knowledge Engineering der Johannes Kepler Universität Linz.

Die Prototyp-Implementierung besteht aus mehreren PL/SQL-Programmpaketen für Oracle 11g, die Definitions- und Abfrageoperationen für hetero-homogene Data Warehouses zur Verfügung stellen. Die Programmpakete werden direkt in der Datenbank gespeichert (Stored Procedures), eine lokale Installation von Software ist nicht notwendig.

M-Objects, M-Relationships und M-Cubes werden als Objekttypen dargestellt. Die Prototyp-Implementierung nutzt dafür die objekt-relationalen Fähigkeiten der Oracle-Datenbank. Das logische Datenmodell basiert auf dem Fact-Constellation-Schema, einer Variante des Star-Schemas mit normalisierten Dimensionstabellen und mehreren Faktentabellen unterschiedlicher Granularität [Hah10]. Die Aufteilung in Dimensions- und Faktentabellen erfolgt dabei so, dass jede Tabelle für sich homogen ist (Abbildung 2b). Neben den eigentlichen Analysedaten in den Dimensions- und Faktentabellen werden M-Objects, M-Relationships und M-Cubes in Objekttabellen abgelegt, wodurch zusätzliche Schemainformation zur Verfügung steht.

Durch die Verwendung einer Star-Schema-Variante können bestehende Algorithmen und Tools einfacher an hetero-homogene Data Warehouses angepasst werden, was die Umstellungskosten für Unternehmen verringert. Daneben erleichtern speziell für M-Cubes angepasste Varianten von Standard-Operationen (Slice, Dice, Projektion [NST10]) dem Benutzer die Abfrage von Informationen. Außerdem besteht die Möglichkeit, hetero-homogene Daten in eine homogene Faktentabelle zu transformieren, um mit SQL und deren OLAP-Erweiterungen darauf zuzugreifen.

Literatur

[GMR98] Golfarelli, M. / Maio, D. / Rizzi, S.: The Dimensional Fact Model: A Conceptual Model for Data Warehouses. International Journal of Cooperative Information Systems, Vol. 7, No. 2 & 3, 1998, S. 215–247

[Hah10] Hahne, M.: Mehrdimensionale Datenmodellierung für analyseorientierte Informationssysteme. In: Chamoni, P. / Gluchowski, P. (Hrsg.): Analytische Informationssysteme. Business-Intelligence-Technologien und -Anwendungen. 4. Aufl., Springer, 2010, S. 229–258

[Neu10] Neumayr, B.: Multi-Level Modeling with M-Objects and M-Relationships. Dissertation, Institut für Wirtschaftsinformatik – Data & Knowledge Engineering, Johannes Kepler Universität Linz, 2010

[NGS09] Neumayr, B. / Grün, K. / Schrefl, M.: Multi-Level Domain Modeling with M-Objects and M-Relationships. 6th Asia-Pacific Conference on Conceptual Modelling, 2009

[NST10] Neumayr, B. / Schrefl, M. / Thalheim, B.: Hetero-Homogeneous Hierarchies in Data Warehouses. 7th Asia-Pacific Conference on Conceptual Modelling, 2010

Informationsmaterial

Hetero-Homogeneous Data Warehouses, Projekt-Homepage, http://hh-dw.dke.uni-linz.ac.at/