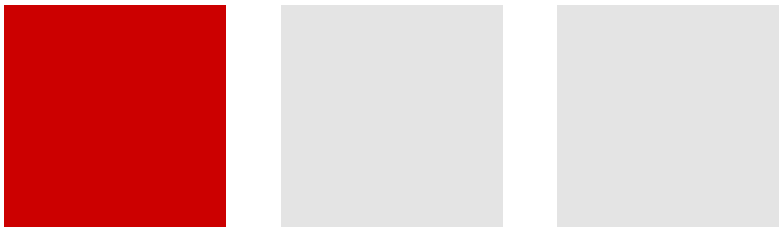

trivadis
makes IT easier.

Wenn die Fakten zu früh eintreffen

Dani Schnider
Principal Consultant
21. Dezember 2011

Eine typische Problemstellung, die beim Laden von Daten in ein Data Warehouse berücksichtigt werden muss, sind Fakten, für die zum Ladezeitpunkt noch keine zugehörigen Dimensionswerte existieren. Für diese als „Early Arriving Facts" bezeichnete Problematik gibt es verschiedene Lösungsansätze, die auf den folgenden Seiten beschrieben werden.

Ausgangslage

Das Whisky-Spezialgeschäft „My Whisky Warehouse" verkauft im Vorweihnachtsgeschäft zahlreiche Spezialitäten, wie die Verkaufszahlen vom 17. Dezember 2011 zeigen. Die folgenden Verkaufsdaten sollen am Abend des 17.12.2011 stehen in der Stage-Tabelle STG_SALES bereit und sollen nun in die Faktentabelle FCT_SALES geladen werden:

STG_SALES

SALES_DATE	*PRODUCT_CODE*	*(weitere Attribute...)*	*QUANTITY*
17.12.2011	11111-22222-33	...	12
17.12.2011	54321-98765-12	...	8
17.12.2011	44444-33333-22	...	15
17.12.2011	12345-67890-76	...	28
17.12.2011	88888-55555-44	...	14
17.12.2011	98765-43210-55	...	11

Wie beim Laden einer Faktentabelle üblich, müssen dabei die Schlüssel auf die zugehörigen Dimensionen ermittelt und in die Faktentabelle geschrieben werden. Solche „Key Lookups" können je nach eingesetzter Technologie und verwendetem ETL-Tool auf unterschiedliche Weise implementiert werden. Ebenfalls einen Einfluss auf die Implementation hat die Art der Historisierung (Slowly Changing Dimensions Typ 1 oder 2). Auf diese Aspekte wird in diesem Artikel nicht eingegangen. Was uns an dieser Stelle aber interessiert, ist die Frage, wie wir mit unbekannten Referenzen auf Dimensionen umgehen. Dazu werfen wir einen Blick auf die Produktdimension DIM_PRODUCTS des erwähnten Whiskyladens. Sie hat zu diesem Zeitpunkt folgenden Inhalt:

DIM_PRODUCTS

PRODUCT_ID	*PRODUCT_CODE*	*PRODUCT_DESC*	*REGION*	*(weitere Attribute...)*
111	12345-67890-76	Edradour 10 years	Midlands	...
112	11111-22222-33	Glenfarclas 105	Speyside	...
113	22222-44444-66	Black Bowmore 1964	Islay	...
114	44444-33333-22	Laphroaig 15 years	Islay	...
115	88888-55555-44	Macallan 25 years	Speyside	...

Bei einem Key Lookup wird nun zum Beispiel dem ersten Fakteneintrag anhand des Produkt–codes 11111-22222-33 die Produkt-ID 112 zugewiesen, die als Fremdschlüsselattribut in die Faktentabelle FCT_SALES geschrieben wird. Beim zweiten Eintrag haben wir aber bereits ein

Problem, denn ein Produkt mit dem Code 54321-98765-12 existiert nicht in der Dimensions–tabelle. Das gleiche gilt auch für den letzten Eintrag mit dem Code 98765-43210-55. Grund dafür können fehlerhafte Produktcodes sein. Meistens liegt die Ursache jedoch darin, dass die zugehörigen Produktinformationen erst zu einem späteren Zeitpunkt ins Data Warehouse geladen werden und somit zum Ladezeitpunkt der Fakten noch nicht bekannt sind. Oder mit anderen Worten: Die Fakten wurden zu früh geliefert.

Es gibt verschiedene Ansätze, um dieses Problem zu lösen. Wie üblich hat jeder Ansatz seine Vor- und Nachteile. Die perfekte Lösung gibt es nicht, aber je nach spezifischen Bedürfnissen kann einer der folgenden Ansätze die geeignete Lösung sein. Alle vorgestellten Verfahren können so implementiert werden, dass die ETL-Verarbeitung ohne Abbruch durchgeführt werden kann. Das wird vorausgesetzt, denn eine manuelle Korrektur oder Ergänzung der fehlenden Daten ist in der Regel keine Option.

Erster Ansatz: Filtern von unvollständigen Fakten

Der einfachste Ansatz besteht darin, Fakten mit fehlenden oder fehlerhaften Referenzen zu ignorieren. Für unser Beispiel würde dies bedeuten, dass die Verkaufsdaten der unbekannten Produkte 54321-98765-12 und 98765-43210-55 nicht in die Faktentabelle geladen werden und somit bei späteren Auswertungen – auch in aggregierter Form – fehlen. Es mag Fälle geben, wo dies tolerierbar ist. Solange die Anforderungen der Data Marts Ungenauigkeiten innerhalb einer definierten Fehlertoleranz erlauben, kann dieser einfache Ansatz zweckmässig sein. Nach dem Laden der Fakten vom 17. Dezember enthält die Faktentabelle FCT_SALES folgende Einträge für diesen Tag:

FCT_SALES

DATE_ID	*PRODUCT_ID*	*(weitere Attribute)...*	*QUANTITY*
17.12.2011	112	...	12
17.12.2011	114	...	15
17.12.2011	111	...	28
17.12.2011	115	...	14

Die Fakteneinträge für die unbekannten Produkte wurden nicht geladen. Dies hat einerseits zur Folge, dass die einsprechenden Verkäufe nicht ausgewertet werden können, andererseits aber auch, dass die Aggregationen nicht korrekt berechnet werden können. Basierend auf den geladenen Fakten wurden am 17. Dezember insgesamt 69 Whiskyflaschen verkauft (12 + 15 + 28 + 14). In Wirklichkeit waren es aber 88 Flaschen – die 8 bzw. 11 Whiskies unbekannter Marke fehlen in der Auswertung.

Anstatt die unvollständigen Einträge einfach zu ignorieren, können sie stattdessen in eine Fehlertabelle geschrieben werden. Die Fehlertabelle hat den gleichen Aufbau wie die Stage-Tabelle und wird als eine Art „Warteliste“ für Fakten ohne gültige Dimensionsreferenzen verwendet. In jedem der nachfolgenden Ladeläufe wird der Inhalt der Fehlertabelle in die Stage-Tabelle zurückkopiert, um sie zusammen mit den neuen Einträgen in die Faktentabelle zu laden. Diese Erweiterung ist zwar einiges aufwendiger zu implementieren, ermöglicht aber, die fehlenden Fakten zu einem späteren Zeitpunkt nachzuladen. Ab diesem Zeitpunkt liefern dann die Auswertungen die vollständigen und korrekten Informationen. Ein Nachteil dieser

Variante besteht jedoch darin, dass es unter Umständen Datensätze gibt, die „ewig" zwischen Fehlertabelle und Stage-Tabelle hin- und hergeschoben werden, weil die zugehörigen Dimensionsdaten nicht geliefert werden. Das ist bei Fakten mit fehlerhaften Referenzen der Fall. Aus diesem Grund empfiehlt es sich, einen Mechanismus einzubauen, um anhand eines Zählers oder der Differenz zwischen Lieferdatum und Ladezeitpunkt zu entscheiden, wann ein Datensatz als Fehler klassifiziert und nicht mehr geladen werden soll.

Dieser Ansatz – ob mit oder ohne Nachladen der Fakten – hat einige Nachteile. In der einfachen Form werden die unvollständigen Fakten gar nicht, in der komplexeren Variante mit der Fehlertabelle erst mit Verspätung geladen. Das bedeutet in beiden Fällen, dass neben den Detaildaten auch die aggregierten Daten (z.B. Gesamtumsatz oder Anzahl verkaufter Produkte pro Tag) beeinträchtigt werden. Solange nur ein kleiner Prozentsatz der Fakten davon betroffen sind und die Auswertungen eine gewisse Fehlertoleranz erlauben, kann der einfache Ansatz jedoch problemlos verwendet werden.

Zweiter Ansatz: Referenz auf Singleton-Einträge

Um fehlende Fakten zu vermeiden, wird oft folgender Ansatz implementiert: Jede Dimension enthält einen zusätzlichen Eintrag, der für die Zuweisung von unbekannten Referenzen verwendet werden kann. Solche künstlichen Dimensionseinträge werden auch als Singletons bezeichnet und werden oft durch negative Schlüsselwerte gekennzeichnet. Ob pro Dimension nur ein Singleton-Eintrag verwendet wird oder ob separate Einträge für unbekannte, leere oder falsche Dimensionsreferenzen verwendet werden, hängt von den jeweiligen Anforderungen ab. In unserem Beispiel verwenden wir nur einen Singleton-Eintrag mit der Produkt-ID -1. Alle beschreibenden Attribute werden mit Dummy-Werten (z.B. „Unknown", „n/a", „(leer)", etc. gefüllt). Die Dimensionstabelle DIM_PRODUCTS enthält somit neben den fünf bereits geladenen Produkten einen Singleton-Eintrag für alle unbekannten Produkte:

DIM_PRODUCTS

PRODUCT_ID	*PRODUCT_CODE*	*PRODUCT_DESC*	*REGION*	*(weitere Attribute...)*
-1	Unknown	Unknown	Unknown	
111	12345-67890-76	Edradour 10 years	Midlands	...
112	11111-22222-33	Glenfarclas 105	Speyside	...
113	22222-44444-66	Black Bowmore 1964	Islay	...
114	44444-33333-22	Laphroaig 15 years	Islay	...
115	88888-55555-44	Macallan 25 years	Speyside	...

Dank der Singleton-Einträge können nun die Fakten vollständig geladen werden, wobei bei fehlenden oder fehlerhaften Referenzen auf eine Dimension der jeweilige Singleton-Wert zugewiesen wird. Dadurch können Spezialbehandlungen bei den Abfragen, insbesondere Outer-Joins auf die Dimensionstabellen, vermieden werden.

Unsere Fakten vom 17. Dezember können nun vollständig in die Tabelle FCT_SALES geladen werden, wobei die 8 Flaschen des unbekannten Produktes dem Schlüssel -1 des Singleton-Eintrags zugewiesen werden:

FCT_SALES

DATE_ID	*PRODUCT_ID*	*(weitere Attribute...)*	*QUANTITY*
17.12.2011	112	...	12
17.12.2011	-1	...	8
17.12.2011	114	...	15
17.12.2011	111	...	28
17.12.2011	115	...	14
17.12.2011	-1	...	11

Eine Auswertung aller Verkäufe an diesem Tag ergibt, dass insgesamt 88 Whiskyflaschen verkauft wurden (12 + 8 + 15 + 28 + 14 + 11). Bei einem Drill-down auf die Regionen werden die Verkäufe für die Regionen „Midlands", „Speyside", „Islay" und „Unknown" angezeigt. Jeder Whiskyliebhaber weiss, in welchen Gegenden Schottlands diese Regionen liegen. Aber wo liegt die Region „Unknown"?

Hier liegt nun das Problem von Singleton-Einträgen. Die Fakten können zwar vollständig geladen werden, und das Gesamttotal auf der obersten Hierarchiestufe wird auch korrekt angezeigt. Doch sobald ein Drill-down auf detaillierte Daten ausgeführt wird, sind nicht alle relevanten Informationen verfügbar. Selbst wenn zu einem späteren Zeitpunkt die entsprechenden Dimensionseinträge nachgeladen werden, ist eine nachträgliche Zuordnung der Fakten nicht mehr möglich. Die unbekannten Whiskies werden in der Faktentabelle immer unbekannt bleiben.

Ein wesentlicher Vorteil des Ansatzes mit den Singleton-Einträgen besteht darin, dass er sehr einfach zu implementieren ist. Beim initialen Laden der Dimensionen müssen einmal die Singleton-Einträge eingefügt werden. In den Key Lookups beim Laden der Fakten wird der Schlüsselwert des Singleton-Eintrags zugewiesen, falls kein passender Dimensionseintrag gefunden wird. Bei Abfragen werden die Singleton-Einträge wie normale Dimensionseinträge behandelt, sodass hier keine Spezialbehandlung notwendig ist. Der Preis dafür ist, dass eine nachträgliche Zuordnung der Fakten zu den korrekten Dimensionseinträgen nicht mehr möglich ist.

Es gibt jedoch Fälle, in denen zu einem späteren Zeitpunkt die Referenzen auf die Singleton-Einträge durch die korrekten Dimensionsschlüssel ersetzt werden können. Voraussetzung dafür ist, dass die Fakten eindeutig identifizierbar sind, beispielsweise durch eine Transaktions–nummer. In diesem Fall wird der unvollständige Fakteneintrag in eine Fehlertabelle geschrieben. Sobald der fehlende Dimensionseintrag geladen wurde, wird der entsprechende Eintrag in der Faktentabelle anhand der Transaktionsnummer identifiziert und der Dimensions–schlüssel -1 mit dem korrekten Wert überschrieben.

Steht keine eindeutige Identifikation der Fakteneinträge zur Verfügung, besteht stattdessen auch die Möglichkeit, den fachlichen schlüssel der Dimension – in unserem Beispiel das Attribut PRODUCT_CODE – als zusätzliches Attribut in der Faktentabelle zu speichern. Da dies für jede Dimension gemacht werden muss, wird die Faktentabelle viel grösser, was sich bei grossen Datenmengen auf die Datenbankgrösse und bei Data Marts auch auf die Performance der Abfragen auswirkt. Anhand dieser redundanten Informationen können dann die fehlenden

Dimensionsschlüssel nachträglich aktualisiert werden, sobald die entsprechenden Produkt–informationen verfügbar sind.

Ein nachträgliches Aktualisieren der Fakten sollte je nach eingesetzter Technologie möglichst vermieden werden. So kann es problematisch sein, wenn Updates auf Fakten in einer kompri–mierten Faktentabelle durchgeführt werden. Solche Aspekte müssen bei der Realisierung dieser komplexen Ansätze berücksichtigt werden.

Wird auf das nachträgliche Aktualisieren der Fakten verzichtet, ist der Ansatz mit den Singleton-Einträgen sehr einfach zu implementieren und hat den Vorteil, dass die Fakten sofort geladen werden können. Auswertungen auf die höchste Aggregationsstufe liefern korrekte Resultate. Beim Drill-down auf untergeordnete Hierarchiestufen stehen aber nicht alle Detaildaten zur Verfügung. Wenn diese Einschränkung aufgrund der Anforderungen vertretbar ist, sind Singleton-Einträge der richtige Ansatz.

Dritter Ansatz: Generieren von Embryo-Einträgen

Falls die Anforderung besteht, dass nachträglich geladene Dimensionseinträge den bereits vorhandenen Fakten zugeordnet werden müssen, um auch Detaildaten korrekt auswerten zu können, sind komplexere Lademechanismen notwendig. Der nächste hier vorgestellte Ansatz kann zur Lösung dieser Anforderung verwendet werden.

Bevor die Fakten geladen werden, wird für die neuste Datenlieferung geprüft, ob sie Referenzen auf nicht vorhandene Dimensionseinträge enthält. Wenn dies der Fall ist, werden zusätzliche Einträge in den Dimensionstabellen erstellt. Da die beschreibenden Attribute zu diesem Zeitpunkt noch fehlen, werden stattdessen Dummy-Werte eingefügt.

Für unser Beispiel heisst dies, dass vor dem Laden der Fakten vom 17. Dezember zwei neue Dimensionseinträge für die unbekannten Produkte 54321-98765-12 und 98765-43210-55 in die Produktdimension geschrieben wird. Diese Einträge werden später durch die echten Produktinformationen ersetzt.

DIM_PRODUCTS

PRODUCT_ID	*PRODUCT_CODE*	*PRODUCT_DESC*	*REGION*	*(weit. Attr...)*	*EMBRYO*
111	12345-67890-76	Edradour 10 years	Midlands	...	No
112	11111-22222-33	Glenfarclas 105	Speyside	...	No
113	22222-44444-66	Black Bowmore 1964	Islay	...	No
114	44444-33333-22	Laphroaig 15 years	Islay	...	No
115	88888-55555-44	Macallan 25 years	Speyside	...	No
116	54321-98765-12	Unknown	Unknown	...	Yes
117	98765-43210-55	Unknown	Unknown	...	Yes

Da die Datensätze bereits in die Dimensionstabelle geschrieben werden, bevor die zugehörigen Informationen vom Quellsystem geliefert werden, also vor dem „Geburtszeitpunkt" des Dimensionseintrages, werden solche vorgängig erstellen Datensätze als „Embryo-Einträge" bezeichnet. Um sie von den bereits gelieferten Dimensionseinträgen zu unterscheiden, empfiehlt es sich, sie speziell zu kennzeichnen, beispielsweise mit einem Attribut Embryo-Flag.

Nun steht in der Produktdimension für alle Fakten ein Dimensionseintrag zur Verfügung – entweder ein bereits gelieferter oder ein Embryo-Eintrag. Somit können nun die Verkaufsdaten vollständig in die Faktentabelle geladen werden:

FCT_SALES

DATE_ID	*PRODUCT_ID*	*(weitere Attribute)...*	*QUANTITY*
17.12.2011	112	...	12
17.12.2011	116	...	8
17.12.2011	114	...	15
17.12.2011	111	...	28
17.12.2011	115	...	14
17.12.2011	117	...	11

Der Unterschied zu den Singleton-Einträgen ist zu diesem Zeitpunkt noch nicht ersichtlich, denn auch hier wird nun bei Auswertungen ein unbekanntes Produkt angezeigt. Der Vorteil dieses Verfahrens besteht aber darin, dass der neue Dimensionseintrag zu einem späteren Zeitpunkt mit den richtigen Bezeichnungen ersetzt werden kann, ohne dass die Fakten neu zugeordnet werden müssen. Mit dem Laden der vollständigen Produktinformationen in die Dimensionstabelle wird der Embryo-Eintrag in einen echten Dimensionseintrag übergeführt, der nun sozusagen das Licht der (DWH-)Welt erblickt. Zu diesem Zeitpunkt wird das Embryo-Flag auf „No“ gesetzt.

Sobald die vollständigen Informationen in der Dimensionstabelle verfügbar sind, erscheinen Sie auch in den Auswertungen, und zwar ohne Nachladen oder nachträgliches Aktualisieren der Fakten. Hier liegt der grosse Vorteil dieses Ansatzes.

Im Whiskygeschäft „My Whisky Warehouse“ wurde inzwischen fleissig gearbeitet, und am Abend des 19. Dezember stehen die vollständigen Produktinformationen für die neuen Spezialitäten zur Verfügung und können ins Data Warehouse geladen werden. In der Produktdimension werden nun die Embryo-Einträge durch echte Produkteinträge ersetzt. Der künstliche Schlüssel PRODUCT_ID sowie der fachliche Schlüssel PRODUCT_CODE bleiben erhalten, alle weiteren Attribute werden aktualisiert und das Embryo-Flag auf „No“ gesetzt:

DIM_PRODUCTS

PRODUCT_ID	*PRODUCT_CODE*	*PRODUCT_DESC*	*REGION*	*(weit. Attr...)*	*EMBRYO*
111	12345-67890-76	Edradour 10 years	Midlands	...	No
112	11111-22222-33	Glenfarclas 105	Speyside	...	No
113	22222-44444-66	Black Bowmore 1964	Islay	...	No
114	44444-33333-22	Laphroaig 15 years	Islay	...	No
115	88888-55555-44	Macallan 25 years	Speyside	...	No
116	54321-98765-12	BenRiach Curiositas	Speyside	...	No
117	98765-43210-55	Bruichladdich Octomore	Islay	...	No

Fazit

Die Behandlung von zu früh gelieferten Fakten beziehungsweise zu spät gelieferten Dimensionseinträgen kann auf unterschiedliche Weise implementiert werden. Entscheidend ist dabei nicht primär die eingesetzte Technologie, denn alle hier beschriebenen Ansätze können mit unterschiedlichen ETL-Tools und Datenbanksystemen realisiert werden. Wichtiger für die Wahl der richtigen Variante sind die konkreten Bedürfnisse: Wie relevant ist es für die fachlichen Anforderungen, dass die Fakten vollständig geladen werden? Was sind die Konsequenzen, wenn Dimensionsdaten bei einem Drill-down unvollständig angezeigt werden? Ist es notwendig oder wünschenswert, dass die fehlenden Informationen nachgeladen werden, sobald sie verfügbar sind, oder genügt es, wenn nur die oberste Hierarchiestufe vollständig ist? Nach dem Klären dieser Fragen mit den Fachabteilungen und den Anwendern der Data Marts kann der technische Entscheid gefällt werden, welche der hier beschriebenen Ansätze realisiert werden soll.

Anhand eines konkreten Sachverhalts, wie er bei ETL-Prozessen in Data Warehouses häufig auftritt, wurde in diesem Artikel aufgezeigt, wie eine typische Problemstellung im DWH-Umfeld mit unterschiedlichen Lösungsansätzen behandelt werden kann. Solche und weitere Praxistipps werden auch im Trivadis-Kurs „Data Warehousing mit Oracle“ (O-DWH) sowie im Buch „Data Warehousing mit Oracle – Business Intelligence in der Praxis“ (Hanser Verlag, ISBN 978-3-446-42562-0) vermittelt.

Viel Erfolg beim Einsatz von Trivadis-Know-how wünscht Ihnen

Dani Schnider
Trivadis AG
Europa-Strasse 5
CH-8152 Glattbrugg
Internet: www.trivadis.com

Tel: +41(0)44-808 70 20
Fax: +41(0)44-808 70 21
Mail: info@trivadis.com