Schnelle Kurzgeschichten

Dr. Andrea Kennel
InfoPunkt Kennel GmbH
CH-8600 Dübendorf

Schlüsselworte:

Data Warehousing, Dimensionen, Performance, Slowly Changing Dimensions.

Einleitung

Unsere Kundin ist im Inkassobereich tätig. Gearbeitet wird mit Fällen, die besagen, wer wem welchen Betrag schuldet. Es geht also um offene Rechnungen und um Schuldner. Eine wichtige Aufgabe des Inkassos ist es nun, mit dem Schuldner Kontakt aufzunehmen und diesen zur Bezahlung der Rechnung aufzufordern. Der Erfolg hängt also auch davon ab, ob vom Schuldner eine Telefonnummer bei Falleröffnung bekannt ist. Wenn nicht, so muss diese zuerst recherchiert werden. Ändert sich die Telefonnummer, so wird dies im produktiven System gespeichert.
Im DWH werden die Chancen des Erfolges des Inkassos berechnet. Dazu gehört auch die Telefonnummer bei Falleröffnung, respektive, ob diese bei Falleröffnung bekannt war. Somit müssen wir im DWH also die Telefonnummern der Schuldner mit ihrem zeitlichen Verlauf speichern.
Der folgende Artikel zeigt dazu Lösungsvarianten und vergleicht diese.

Geschichten im DWH

Unter Geschichten im DWH verstehen wir die Historisierung von Daten. So wird beispielsweise nicht nur die aktuelle Telefonnummer eines Schuldners gespeichert, sondern der gesamte Verlauf der Telefonnummern.
Ein kleines Beispiel:
Am 1.1.2009 wird für den Schuldner Dagobert Duck die Telefonnummer 012 345 11 11 angegeben. 10 Wochen später, am 03.03.2009 wird diese korrigiert auf 012 345 33 33. Am 1.April 2010 schliesslich bekommt Dagobert einen neuen Telefonanschluss und Nummer, die aber erst am 5. Mai 2011 ausfindig gemacht werden kann. Sie lautet 012 345 55 55.
Das ergibt folgende Telefongeschichte:
Tabelle QUELLE_SCHULDNER_TELEFON

Schuldner	Telefonnummer	Gültig ab
Dagobert Duck	012 345 11 11	01.01.2009
Dagobert Duck	012 345 33 33	03.03.2009
Dagobert Duck	unbekannt	04.04.2010
Dagobert Duck	012 345 55 55	05.05.2011

Abb. 1: Tabelle QUELLE_SCHULDNER_TELEFON

Dagobert Duck hat aber noch eine zweite Geschichte. Er hat bei Daniel Düsentrieb verschiedene Erfindungen in Auftrag gegeben, die er nicht pünktlich bezahlt hat. Daher hat Daniel Düsentrieb Inge Inkasso gebeten, bei Dagobert die offenen Rechnungen einzutreiben, was Inge teils auch geschafft hat.

Hier die Fälle (Rechnungen) mit weiteren Angaben:

Tabelle QUELLE_FALL

Fall_Nr	Datum Rechnung	Datum Falleröffnung	Schuldner	Kreditor
8	01.07.2008	08.08.2008	Dagobert Duck	Daniel Düsentrieb
9	24.12.2008	31.01.2009	Dagobert Duck	Daniel Düsentrieb
10	24.12.2009	31.01.2010	Dagobert Duck	Daniel Düsentrieb
11	24.12.2010	31.01.2011	Dagobert Duck	Daniel Düsentrieb

Abb. 2: Tabelle QUELLE_FALL

Hier noch ein Ausschnitt aus dem produktiven Datenmodell:

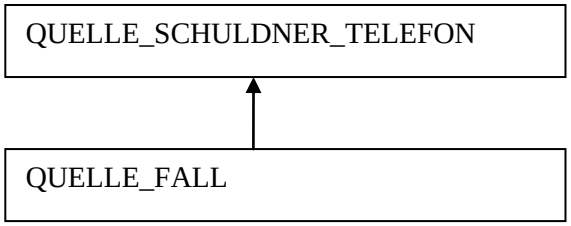

Abb.3: Ausschnitt aus dem produktiven Datenmodell

Was nun interessiert, ist die Frage, ob bei Falleröffnung Inge eine Telefonnummer von Dagobert hatte oder nicht. Wir sehen relativ rasch, dass dies für Fall 8 und 11 nicht zutrifft, für 9 und 10 aber schon.
Damit Conny Controller dies ebenfalls kontrollieren kann, werden sowohl die Fälle als auch die Angaben zum Schuldner in ein DWH geladen und dort ausgewertet.

Eine lange Geschichte

Will man in einem DWH Geschichte schreiben, so schreibt man oft eine lange Geschichte. Damit meinen wir, dass man die gesamte Geschichte im DWH festhält, indem man in unserem Beispiel alle Versionen der Telefonnummern speichert. Dies nennt sich Slowly Changing Dimension Typ II oder kurz SCDII.
Hier das Beispiel, wie dies im DWH dann aussieht:

Tabelle D_SCHULDER_LANG

ID	Schuldner	Telefonnummer	Gültig von	Gültig bis
100	Dagobert Duck	NULL	01.01.1900	31.12.2008
101	Dagobert Duck	012 345 11 11	01.01.2009	02.03.2009
102	Dagobert Duck	012 345 33 33	03.03.2009	03.04.2010
103	Dagobert Duck	NULL	04.04.2010	04.05.2011
104	Dagobert Duck	012 345 55 55	05.05.2011	31.12.9999

Abb.4: Tabelle D_SCHULDER_LANG

Weiter werden im DWH auch die Fälle gespeichert, was folgendermassen aussehen kann:
Tabelle D_FALL

ID	Fall	Eröffnet	Rechnung	Schuldner	Kreditor
201	8	08.08.2008	01.07.2008	Dagobert Duck	Daniel Düsentrieb
202	9	31.01.2009	24.12.2008	Dagobert Duck	Daniel Düsentrieb
203	10	31.01.2010	24.12.2009	Dagobert Duck	Daniel Düsentrieb
204	11	31.01.2011	24.12.2010	Dagobert Duck	Daniel Düsentrieb

Abb.5: Tabelle D_FALL

Eine Kurzgeschichte

Betrachten wir die Anforderung an die Telefonnummer etwas genauer. Damit Inge Inkasso telefonieren kann, greift sie auf das produktive System zu und liest die aktuelle Telefonnummer. Conny Controller will ab DWH eine Auswertung machen, ob bei Falleröffnung eine Telefonnummer des Schuldner Dagobert Duck bekannt war oder nicht und wie die Situation heute aussieht. So interessiert im DWH eigentlich nicht wirklich die gesamte Geschichte sondern nur Momentaufnahmen.

Die kann folgendermassen abgebildet werden:
Tabelle D_FALL_KURZGESCHICHTE

ID	Fall	Eröffnet	Rechnung	Schuldner	Kreditor	TEL_ eroeffnet	TEL_ heute
201	8	08.08.2008	01.07.2008	Dagobert Duck	Daniel Düsentrieb	N	Y
202	9	31.01.2009	24.12.2008	Dagobert Duck	Daniel Düsentrieb	Y	Y
203	10	31.01.2010	24.12.2009	Dagobert Duck	Daniel Düsentrieb	Y	Y
204	11	31.01.2011	24.12.2010	Dagobert Duck	Daniel Düsentrieb	N	Y

Abb.6: Tabelle D_FALL_KURZGESCHICHTE

Damit wird nun die Tabelle für die Schuldner einfacher, da wir dort keine Geschichte mehr brauchen. Die aktuellen Daten genügen.

Tabelle D_SCHULDNER_KURZ

ID	Schuldner	Telefonnummer
104	Dagobert Duck	012 345 55 55

Abb.7: Tabelle D_SCHULDNER_KURZ

Geschichtsschreibung

Fragestellung
Es stellt sich die Frage, wie die Daten in das DWH geladen werden können. Dies sieht für die lange Geschichte sicher anders aus als für die Kurzgeschichte. Was ist hier komplizierter und wie sieht es mit der Performance aus?

Lange Geschichte
Zuerst laden wir die Tabelle D_SCHULDER_LANG basierend auf der Tabelle aus der Quelle berechnet. Die Quelltabelle ist QUELLE_SCHULDNER_TELEFON.
Betrachten wir die Quelle genauer, so sehen wir, dass wir für die Gültigkeit nur ein Gültig_von haben. Damit wir aber möglichst einfach die Daten abfragen können, brauchen wir in der DWH-Tabelle auch ein Gültig_bis. Dieses errechnet sich aus dem Gültig von des nächsten Records. Hier wie das zu verstehen ist:

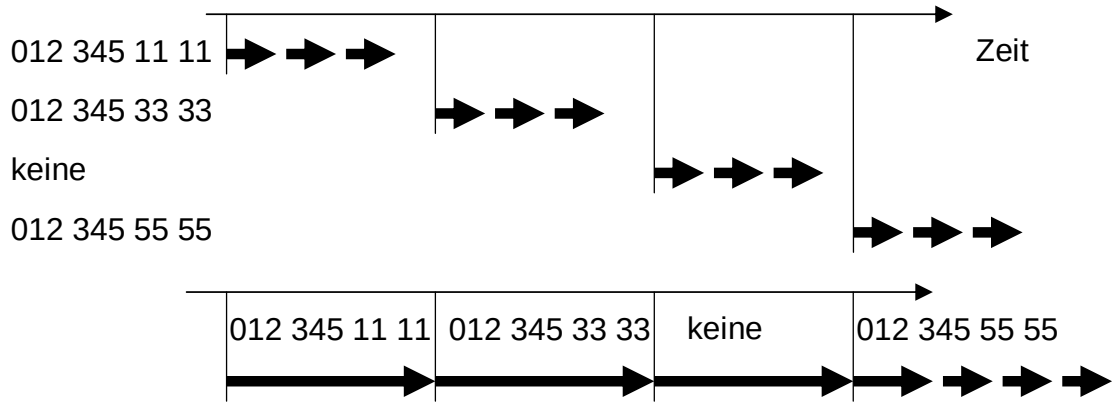

Abb.8: Gültigkeit berechnen

In der Grafik sehen wir, dass wir so für die Zeit bis zur ersten Telefonnummer gar keine Version des Schuldners gespeichert haben. So müssen wir noch eine Version des Schuldners ohne Telefonnummer mit Gültig_von gleich einem Minimaldatum und Gültig_bis direkt vor dem kleinsten Gültig_von der existierenden Einträge einfügen.

Weiter muss jeder Record einen neuen künstlichen Schlüssen bekommen, was am einfachsten mit einer Sequenz gemacht wird.

Die Tabelle D_FALL ist einfacher zu laden. Die Daten der Quelle können direkt übernommen werden und müssen mit einem künstlichen Schlüssel ergänzt werden.

Hier nun ein Ausschnitt aus dem DWH-Datenmodell:

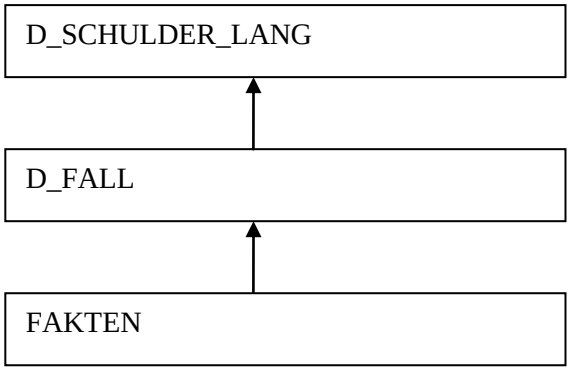

Abb.9: Datenmodell lange Geschichte

Wenn wir es genau betrachten, so ist der Fall eine Dimension, zu der es Fakten gibt. Die Tabelle SCHULDNER_LANG ist eine weitere Angabe zum Fall. Nun haben wir das Problem, dass wir über die Zeit gesehen eine m-zu-m Beziehung zwischen Schuldner und Fall haben. Ein Schuldner kann mehrere Fälle haben und ein Fall mehrere Versionen eines Schuldners, aber immer desselben Schuldners.
Korrekterweise müssten wir also vom Fall auf die zum Fall passende Version des Schuldners verweisen. Das wäre dann sinnvollerweise ein Verweis auf die Schuldnerangaben aktuell oder die Schuldnerangaben zum Zeitpunkt der Falleröffnung.
Weiter sehen wir, dass wir es eigentlich nicht mit einem STAR Modell sondern mit einem SNOWFLAKE Modell zu tun haben. So müsste man sich also überlegen, ob nicht die

Schuldnerangaben als Hierarchie auf dem Fall hinterlegt werden sollten. Dann haben wir in unserem Beispiel für jeden Fall von Dagobert Duck 4 Versionen des Falles, die gespeichert werden müssen.
Für unser Beispiel genügt es zu sehen, dass es hier noch ein Architekturproblem gibt. Wir lösen es momentan so, dass wir in beiden Tabellen den Namen des Schuldners als Business-Key ablegen, damit also den Schlüssel der Quelle, was zugegebenermassen keine saubere Lösung ist.

Kurzgeschichte
Bei der Kurzgeschichte ist liegt der Knackpunkt beim Berechnen der Felder TEL-eröffnet und TEL-heute in der Tabelle D_FALL_KURZGESCHICHTE. Damit diese einfach gefüllt werden kann, brauchen wir eine Hilfstabelle, die gleich aussieht wie die Tabelle D_SCHULDNER_LANG. Wir müssen also auch die Gültigkeiten berechnen. Zum Bestimmen, ob zur Eröffnung und heute eine Telefonnummer bekannt ist, kann dann auf der Hilfstabelle mit „between gültig_von and gültig_bis“ gearbeitet werden.

Hier das Ganze grafisch:

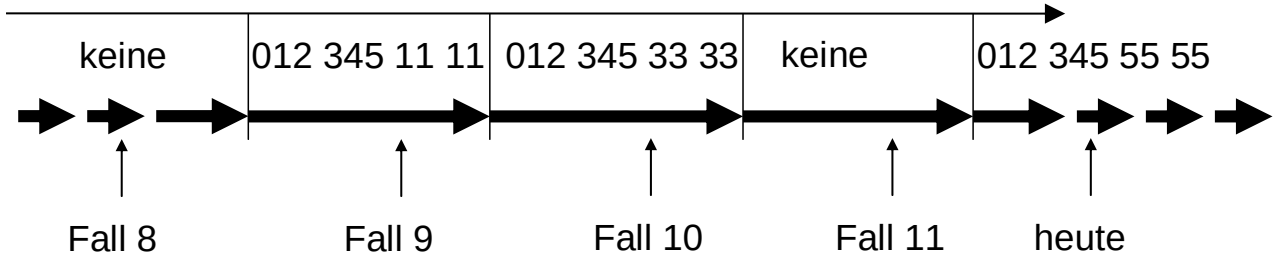

Abb.10: Zeitabhängige Werte lesen

Das Füllen der Tabelle D_SCHULDNER_KURZ ist nun so einfach, dass dies hier nicht beschrieben wird.

Auch hier ein Ausschnitt aus dem DWH-Datenmodell:

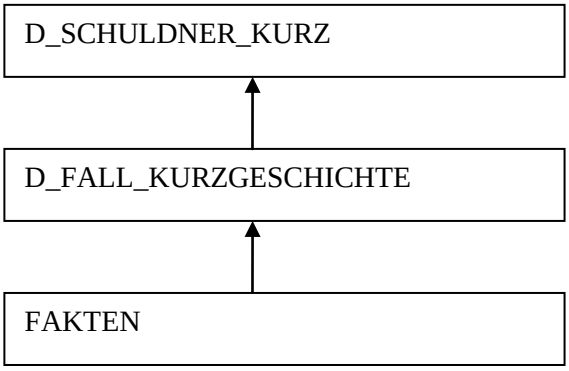

Abb.11: Datenmodell Kurzgeschichte

Wir sehen, dass wir hier auch ein SNOWFLAKE Modell haben. Doch haben wir zwischen den beiden Dimensionen eine klare Hierarchie, da ein Fall immer genau einen Schuldner hat. So können wir gut in D_FALL_KURZGESCHICHTE einen Fremdschlüssel auf D_SCHULDNER_KURZ definieren. Oder wir können die zusätzlichen Felder von D_SCHULDNER_KURZ direkt in D_FALL_KURZGESCHICHTE übernehmen. Je nachdem wie viele Attribute es sind und wie gross

die Tabellen sind. In unserem Fall speichern wir für den Schuldner viele Attribute, so dass wir tatsächlich eine eigene Tabelle haben und bewusst das SNOWFLAKE Modell einsetzen.

Geschichtsforschung

Fragestellung
Bei der Geschichtsforschung stellt sich die Frage, wie nun die gewünschten Auswertungen gemacht werden können und welche Geschichtsschreibung die Geschichtsforschung erleichtert.

Erste Auswertung
Conny Controller möchte in einem ersten Schritt eine allgemeine Auswertung über die Datenqualität zu Falleröffnung haben. Dazu möchte sie wissen, für wie viele Fälle bei Eröffnung eine Telefonnummer bekannt war und für wie viele Fälle nicht. Conny Controller ist der Meinung, dass das relativ einfach machbar sein sollte und möchte bald ein Resultat sehen.
Nun prüfen wir, wie einfach diese Auswertung in welcher Geschichtsschreibung ist.

Lange Geschichte
Die Fälle mit Eröffnungsdatum sind in der Tabelle D_FALL zu finden. Die Telefonnummer in D_SCHULDER_LANG. So müssen wir die beiden Tabellen zuerst joinen, um für jeden Fall zu prüfen, ob es bei Falleröffnung eine Telefonnummer gab.
In einem zweiten Schritt können dann die Fälle mit und ohne Telefonnummer gezählt werden.

Kurzgeschichte
Bei der ersten Fragestellung ist die Kurzgeschichte wirklich eine Kurzgeschichte. Wir könne ganz einfach die Fälle aus D_FALL_KURZGESCHICHTE zählen, die TEL_eroeffnet = ‚Y' respektive = ‚N' ist.

Zweite Auswertung
Für Conny Controller war die erste Auswertung schon eine grosse Hilfe. Nun möchte sie die Fälle ohne Telefonnummer bei Falleröffnung weiter analysieren. Sie möchte wissen, wie viele davon heute eine Telefonnummer haben und wie viele nicht.

Lange Geschichte
Die Basisdaten sind weiterhin die beiden Tabellen D_FALL und D_SCHULDER_LANG. Diesmal müssen wir aber D_SCHULDER_LANG zweimal zu D_FALL joinen. Einmal mit der Eröffnungsdatum und einmal mit dem aktuellen Datum.
Nach diesem Join können dann die Fälle selektiert werden, die bei Eröffnung keine Telefonnummer hatten und danach kann gezählt werden, wie viele davon heute eine resp. keine Telefonnummer haben.

Kurzgeschichte
Bei der zweiten Fragestellung ist die Kurzgeschichte wieder kürzer, einfacher und auch schneller. Wir könne ganz einfach die Fälle aus D_FALL_KURZGESCHICHTE nach TEL_eroeffnet = ‚N' einschränken und zählen dann die TEL_heute = ‚Y' respektive = ‚N' ist.

Reporting Werkzeug
Nun könnte der berechtigte Einwand kommen, dass solche Auswertungen kaum ad-hoc, sondern mit einem Reporting Werkzeug gemacht werden. Nun, in diesem Fall wäre es wohl für einen Endanwender zu komplex, mit den beiden Tabellen D_FALL und D_SCHULDER_LANG zu arbeiten. Also müsste eine View zur Verfügung gestellt werden, die ähnlich aussieht wie die Tabelle D_SCHULDER_LANG. Damit wird die Abfrage dann einiges einfacher, aber nicht schneller. So

kommt dann die nächste Idee, die View als Matrialized View zur Verfügung zu stellen. Damit können wir die lange Geschichte tatsächlich verkürzen. Konkret werden dann zuerst die ganzen Tabellen mit allen Versionen geschrieben und danach werden nur zwei Momentaufnahmen in die Materialized View geschrieben. Doch stellt sich dann die Frage, wieso zuerst eine lange Geschichte geschrieben werden soll, wenn dann doch nur die Kurzgeschichte gebraucht wird.
So kommen wir zum Schluss, dass auch mit Einsatz von Reporting Werkzeugen die Kurzgeschichte sinnvoller und einfacher ist.

Unsere Geschichte

Im DWH unseres Kunden haben wir nicht nur die Telefonnummer, die nur zu den zwei Zeitpunkten Falleröffnung und heute interessiert. Es sind aktuell 15 Attribute. Diese Attribute speichern wir als sogenannte SCDIII direkt auf dem Fall. Dazu speichern wir jeweils den Wert zum Zeitpunkt der Falleröffnung und den aktuellen Wert. Dafür brauchen wir bei anderen Dimensionen keine Versionierung. So können einige versionierte Einträge eingespart werden und das Datenmodell wird einfacher. Damit werden auch Abfragen einfacher und schneller.
Der für uns entscheidende Vorteil der Lösung mit SCDIII liegt aber nicht in der Performance, sondern in der Einfachheit. Die Daten werden so gespeichert, wie sie ausgewertet werden und in der Tabelle, in der sie fachlich erwartet werden. Will ein Controller Fälle auswerten und dabei prüfen, bei wie vielen bei Falleröffnung ein Telefonnummer bekannt war, so ist es für ihn viel einfacher, wenn er diese Information direkt auf dem Fall hat und nicht über eine versionierte Tabelle einschränken muss.

Fazit

Stellt sich in einem DWH eine Anforderung, die Geschichtsschreibung verlangt, so lohnt es sich zu prüfen, ob eine Kurzgeschichte genügt oder ob eine lange Geschichte nötig ist. Normalerweise werden versionierte Dimensionen gebraucht, um Fakten einzuschränken. Dann ist klar, dass alle Versionen gespeichert werden müssen, da im Normalfall Fakten zu unterschiedlichen Zeitpunkten anfallen. Doch wenn historische Daten für Auswertungen auf Dimensionen gebraucht werden, so wird es wohl häufig vorkommen, dass nur Momentaufnahmen gefragt sind. Dann lohnt es sich nicht, viele Versionen zu speichern und dann nur zwei davon wirklich zu brauchen.
Unser Beispiel zeigt, dass wenn Geschichtsschreibung zur Einschränkung von Dimensionen benötigt wird, Kurzgeschichten klar den langen Geschichten vorzuziehen sind. Das heisst, dass in diesem Fall SCDIII geeigneter ist als SCDII.
Die Vorteile sind:

- Einfachere, saubere Architektur
- Einfachere, schnellere Abfragen

Kontaktadresse:

Dr. Andrea Kennel
InfoPunkt Kennel GmbH
Bahnhofstr. 48
CH-8600 Dübendorf

Telefon: +41 (0) 44 820 71 40
E-Mail andrea@infokennel.ch
Internet: www.infokennel.ch