Modellierung und Ausführung temporaler Datenintegrationsprozesse in Data-Warehouse-Systemen

Arne Harren

Carl von Ossietzky Universität Oldenburg
Department für Informatik, Abteilung Informationssysteme
D-26111 Oldenburg
arne.harren@informatik.uni-oldenburg.de

Zusammenfassung: Mit Data-Warehouse-Systemen hat in den letzten Jahren eine spezielle Klasse von Informationssystemen stark an Bedeutung gewonnen, die integrierte, konsolidierte und historisierte Sichten auf einen durch verschiedene Datenquellen gebildeten Gesamtdatenbestand bereitstellen. Ein zentraler Aspekt bei Aufbau und Pflege von Data-Warehouse-Systemen ist daher die Integration der Daten aus den Datenquellen unter Berücksichtigung der zeitlichen Zusammenhänge der Daten, um eine konsistente Gesamtsicht zu erreichen. Die zeitlich konsistente Integration wird jedoch aufgrund der Heterogenität der Datenquellen, der sich hieraus ergebenden Vielfalt der Zeitdarstellungen und der in bisherigen Integrationsansätzen fehlenden, flexiblen und impliziten Zeitunterstützung erschwert.

Dieser Beitrag präsentiert einen Überblick über das TEMPFLOW-Modell für die Spezifikation und Ausführung temporaler Datenintegrationsprozesse und die temporale Algebra TIA, deren Operatoren eine flexible Behandlung von Daten und Zeitangaben bereitstellen, um neben klassischen Integrationskonflikten auch zeitbezogene Konflikte aufzulösen, vorhandene Zeitangaben implizit bei der Datenverknüpfung zu nutzen und zeitlich konsistente Integrationsergebnisse zu liefern.

1 Einleitung

Unter dem Begriff des Data-Warehouse-Systems (DWS) werden Technologien zur Entscheidungsunterstützung subsumiert, die auf Basis einer historisierten, integrierten und konsolidierten Datenbasis eine schnelle, umfassende und qualitativ hochwertige Entscheidungsunterstützung ermöglichen. Kernstücke eines DWS sind dedizierte Datenbanken, z. B. Data Warehouses (DWs) oder Data Marts, die den effizienten Zugriff auf themenbezogene Sichten eines Gesamtdatenbestandes bereitstellen, der durch verschiedene Datenquellen gebildet wird. Charakterisiert durch die dauerhafte, redundante Speicherung der Daten in dedizierten Datenbanken ermöglicht ein DWS somit quellübergreifende Analysen unter Betrachtung zeitlicher Veränderungen.

Ein zentraler Aspekt bei Aufbau und Pflege von DWSen ist die Integration der Daten aus den Datenquellen und ihre Überführung in das für Analysezwecke benötigte Format.

Hierbei sind zum einen klassische Integrationskonflikte auf den Ebenen der Struktur, Beschreibung und Semantik, wie bspw. Unterschiede in Benennungen, Schemastrukturierungen und Datenkodierungen, zu beheben (vgl. z. B. [SPD92]) und fehlerhafte Daten zu bereinigen. Zum anderen muss eine konsistente zeitliche Zuordnung der Daten aus den verschiedenen Quellsystemen und eine korrekte Historisierung der Integrationsergebnisse im DW vorgenommen werden. Bedingt durch die Heterogenität der beteiligten Systeme und der verbreiteten Adaption von Konzepten temporaler Datenbanken zur Speicherung zeitbezogener Daten, indem bspw. Daten mit Zeitstempeln zur Angabe der Datengültigkeit versehen werden, müssen in der Integration in den Daten vorhandene Zeitangaben identifiziert, interpretiert, strukturell sowie semantisch vereinheitlicht und bei der Datenverknüpfung genutzt werden, um bspw. ein bzgl. der Datengültigkeit konsistentes Ergebnis zu erhalten. Ein großes Problem stellt hierbei die Darstellungsvielfalt von Zeitangaben in den Datenquellen dar, deren zeitliche Interpretation zumeist seitens der auf den Quellen operierenden operativen Unternehmensanwendungen erfolgt.

Ein Beispiel für eine mögliche Speicherung zeitbezogener Daten in einer Datenquelle zeigt die in Abbildung 1 dargestellte Relation mit Zuordnungen von Produkten zu Produktgruppen. Eine Interpretation dieser Relation kann z. B. ergeben, dass die Attribute $start$ und $ende$ die zeitliche Gültigkeit der Zuordnung eines Produktes zu einer Produktgruppe angeben. Das Produkt P1 ist somit zum einen der Produktgruppe PG1 und zum anderen der Gruppe PG3 zugeordnet. Die Werte innerhalb der Attribute $start$ und $ende$ lassen darauf schließen, dass es sich um Angaben von Start- und Endmonaten handelt, für die die Zuordnung gilt. Durch Betrachtung des Produktes P1 lässt sich zudem rekonstruieren, dass der Wert im Attribut $ende$ nicht mehr zur Gültigkeit gehört; das erste Tupel der Relation besagt somit, dass die Zuordnung für Januar und Februar 2005 gilt. Enthält das Attribut $ende$ den Wert $null$, so handelt es sich um die aktuell gültige Zuordnung.

produkt	*produktgruppe*	*start*	*ende*
P1	PG1	2005-01	2005-03
P2	PG2	2005-01	*null*
P1	PG3	2005-03	*null*

Abbildung 1: Relation einer Datenquelle mit Produkt-Produktgruppen-Zuordnungen

Eine auf Basis dieser Erkenntnisse vorgenommene zeitliche Interpretation und Überführung der Relation in eine temporale Relation mit Stand April 2005 zeigt Abbildung 2. Die zuvor in den Attributen $start$ und $ende$ vorhandenen Angaben über die Gültigkeit der Zuordnungen sind in der restrukturierten Relation in eine Zeitdimension — hier in der Darstellung durch einen doppelten vertikalen Strich von den Datenattributen abgetrennt — überführt worden, so dass sie bei Anwendung der temporalen Operatoren der TIA-Algebra implizit berücksichtigt werden. Ein temporaler Verbund nutzt bspw. implizit die zeitliche Gültigkeit der jeweiligen Verbundpartner.

Ausgehend von diesem einleitenden Beispiel wird im folgenden Abschnitt 2 zunächst eine Erweiterung der klassischen Integrationsarchitektur vorgestellt, die als Infrastruktur für die temporalen Integrationskonzepte dient. Abschnitt 3 gibt einen Einblick in die temporale Integrationsalgebra TIA für Datenrestrukturierungen und -verknüpfungen in Integrati-

onsprozessen. Abschnitt 4 präsentiert das Spezifikations- und Ausführungsmodell TEMPFLOW für temporale Integrationsprozesse. Abschnitt 5 diskutiert verwandte Arbeiten. Die Zusammenfassung in Abschnitt 6 schließt den Beitrag ab.

produkt	*produktgruppe*	*Zeitobjekte (gültigkeit)*
P1	PG1	$\{\langle$2005-01$\rangle, \langle$2005-02$\rangle\}$
P1	PG3	$\{\langle$2005-03$\rangle, \langle$2005-04$\rangle\}$
P2	PG2	$\{\langle$2005-01$\rangle, \langle$2005-02$\rangle, \langle$2005-03$\rangle, \langle$2005-04$\rangle\}$

Abbildung 2: Temporale Relation mit Produkt-Produktgruppen-Zuordnungen

2 Temporale Datenintegrationsinfrastruktur

Das Ziel der Datenintegration ist die Bereitstellung einer integrierten und auf die Datenanalyse ausgerichteten, materialisierten und historisierten Sicht auf einen Gesamtdatenbestand, deren Grundlage die Datenbestände der einzelnen Datenquellen sind. Um dieses Ziel zu erreichen, dient im Rahmen des hier vorgestellten Konzeptes der temporalen Datenintegration eine Infrastruktur, durch die auf Grundlage der Transaktionszeit das zustandsorientierte und datenflussbasierte TEMPFLOW-Prozessmodell für die Spezifikation von Integrationsprozessen zum Einsatz kommen kann. Die Gesamtsicht zu einem Transaktionszeitpunkt ergibt sich hierbei direkt aus den Datenbeständen der Datenquellen zu diesem Zeitpunkt und kann im Rahmen eines Integrationsprozesses durch Datenbestände früherer Transaktionszeitpunkte angereichert werden, falls diese Daten in den Datenquellen nicht mehr verfügbar sind.

Die Integrationsinfrastruktur basiert auf der etablierten ETL-Architektur (vgl. z. B. [BG04]), die eine Unterteilung in die drei Schichten *Extraktion*, *Transformation* und *Laden* vornimmt. In der Extraktions- und Ladeschicht werden jeweils die technischen Details für den Zugriff auf die Datenquellen und -senken gekapselt. Zudem erfolgt in diesen Schichten die Abstraktion von den in den Systemen genutzten Datenmodellen, so dass der eigentlichen Datenintegration, die in der Transformationsschicht stattfindet, alle Daten auf Basis eines einheitlichen temporalen, relationalen Integrationsdatenmodells vorliegen. Des Weiteren wird durch Extraktion und Laden die möglicherweise zeitlich eingeschränkte Erreichbarkeit der jeweiligen Datenquellen und -senken vor der Integration versteckt, so dass Extraktion, Integration und Laden sowohl räumlich als auch zeitlich entkoppelt sind.

Abbildung 3 skizziert auf Grundlage der Architektur die zustandsorientierte Sicht der Integration: Ausgehend von den Datenquellen und -senken werden durch Extraktion und Laden einzelne Quell- und Senkenrelationen zur Verfügung gestellt. Die Integration der Daten der Quellrelationen ist durch Anfragegraphen unter Verwendung der Operatoren der Integrationsalgebra TIA — symbolisiert als Selektion σ, Projektion π und Verbund $\bowtie$ — skizziert. Der Zustand der Senken für einen Transaktionszeitpunkt t ergibt sich hierbei bspw. unmittelbar durch die Anwendung der Algebraoperatoren auf den Quelldatenbestand zum Zeitpunkt t.

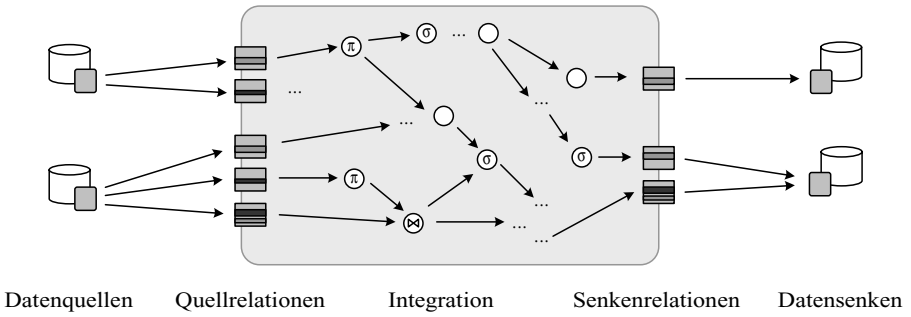

Abbildung 3: Zustandsorientierte Integration mittels Algebraoperatoren

3 Temporale Integrationsalgebra TIA

Die temporale Algebra TIA [Har03, Har04] stellt eine Menge relationaler Integrationsoperatoren bereit, die in Integrationsprozessen eingesetzt werden können, um Daten der Datenquellen in das benötigte Format der Datensenken zu überführen, Konflikte durch unterschiedliche Repräsentationsformen für Zeitangaben zu überwinden und temporale Eigenschaften der Daten zu nutzen.

Neben klassischen Anforderungen an die Algebra, wie bspw. Mengenorientiertheit, Abgeschlossenheit, Deskriptivität, Sicherheit und Vollständigkeit, existieren eine Reihe weiterer Anforderungen. Letztere ergeben sich aus der notwendigen Zeitunterstützung und der Integrationsproblematik. Aus Sicht einer nicht-temporalen Integration ist TIA zum einen relational vollständig und kann zum anderen mit Aggregationen und Tupelzerlegungen umgehen. In Bezug auf die Zeitunterstützung bildet TIA eine konsistente temporale Erweiterung einer nicht-temporalen Algebra (vgl. [Sno95]), so dass jeder nicht-temporale Ausdruck in TIA darstellbar ist [Har04]. Des Weiteren ist TIA schnappschussreduzierbar, wodurch eine temporale Anfrage, reduziert um Zeitangaben, das selbe Ergebnis liefert wie die Anfrage in einer nicht-temporalen Algebra [Sno95]. TIA unterstützt zudem eine uneingeschränkte, endliche Anzahl an Zeitdimensionen (vgl. [DBS96]). Da die Nutzung und Berechnung von Zeitangaben im Rahmen der Integration eine wichtige Rolle spielt und bspw. Daten aufgrund ihrer Gültigkeit verknüpft werden müssen, unterstützt TIA zeitliche Selektionsprädikate, die Durchführung zeitlicher Verknüpfungen, die Manipulationen der Zeitangaben und die Veränderung der Anzahl und Art von Zeitdimensionen. Zudem bietet TIA auf Basis des Typsystems zeitpunkt- und zeitintervallbasierte Interpretationen der Zeitangaben und führt hierbei keine impliziten Typumwandlungen durch [Har04].

Temporales Datenmodell Um eine eindeutige Semantik für zeitbezogene Daten zu erhalten, nimmt das TIA-Datenmodell eine strikte Trennung zwischen dem nicht-temporalen Datenanteil und den zugehörigen Zeitinformationen vor. Des Weiteren legt es keine feste Struktur und Interpretation für Zeitinformationen fest, so z. B. auf eine Menge zweidimensionaler Zeitpunkte, wie dies beim BCDM der Fall ist, sondern nutzt Mengen mehrdi-

mensionaler Zeitobjekte auf Basis einer erweiterbaren Menge zeitlicher Datentypen. Die Interpretation von Zeitinformationen wird somit eindeutig durch die zur Bildung der Dimensionen der Zeitobjekte genutzten Datentypen festgelegt: Wird eine zeitpunktbasierte Interpretation benötigt, so kann bspw. ein Zeitpunkt-Datentyp eingesetzt werden.

Eine temporale Relation r hat daher den Aufbau[1] $r :: L \rightarrow \{\!\{Z\}\!\}$, wobei L ein Tupeltyp mit $L = \langle\!\langle a_1 :: D_1, \ldots, a_n :: D_n \rangle\!\rangle$ und $n \in \mathbb{N}$ ist und Z ein mehrdimensionaler Zeittyp für die Speicherung von Zeitangaben mit $Z = \langle\!\langle b_1 :: E_1, \ldots, b_m :: E_m \rangle\!\rangle$ und $m \in \mathbb{N}$ ist. D_1 bis D_n sind normale Datentypen und E_1 bis E_m granularitätsbasierte temporale Datentypen wie $Instant$ (Zeitpunkt), $Interval$ (Zeitintervall) oder $Span$ (Zeitspanne). Die Namen b_1 bis b_m der Zeitdimensionen sind wie die Namen a_1 bis a_n der Datenattribute frei wählbar und unterliegen keiner speziellen Interpretation durch die Algebra.

Exemplarische Datenrestrukturierungen Zur Verdeutlichung der Restrukturierung von Daten und der Interpretation von Daten als Zeitangaben, so dass diese anschließend von den Operatoren der Algebra implizit genutzt werden, wird im Folgenden die Überführung der Quellrelation aus Abbildung 1 in die in Abbildung 2 gezeigte Form beschrieben. Ein Überblick der TIA-Operatoren und ihre formalen Definitionen finden sich in [Har04].

Den Ausgangspunkt der Restrukturierung bildet die von der Datenextraktion auf Basis des Integrationsdatenmodells gelieferte Relation, die in Abbildung 4 dargestellt ist. Da diese Relation keine Zeitdimensionen besitzt, enthält die Menge der Zeitobjekte jeweils das 0-dimensionale Zeitobjekt $\langle\rangle$. Da die Angabe im Attribut $start$ den Startmonat einer Zuordnung angibt und als Endmonat $null$ eingetragen ist, kann zunächst eine Überführung des Wertes $null$ in den aktuellen Monat $+\,1$ erfolgen. Der aktuelle Monat lässt sich durch die Konstantenfunktion now ermitteln, die den aktuellen Transaktionszeitpunkt zurückliefert. Exemplarisch liefere now den Wert 29. April 2005. Eine anschließende Konvertierung dieses Zeitpunktes auf die Monatsgranularität liefert den Monat 2005-04, so dass das $ende$-Attribut den Wert 2005-05 zugewiesen bekommt.

produkt	*produktgruppe*	*start*	*ende*	*Zeitobjekte* ()
P1	PG1	2005-01	2005-03	$\{\langle\rangle\}$
P2	PG2	2005-01	*null*	$\{\langle\rangle\}$
P1	PG3	2005-03	*null*	$\{\langle\rangle\}$

Abbildung 4: Temporale Quellrelation mit Produkt-Produktgruppen-Zuordnungen

Im nächsten Schritt können die Monatsangaben in den Attributen $start$ und $ende$ in ein Zeitintervall überführt werden, wobei von dem Ende des Intervalls ein Monat abzuziehen ist. Das Attribut, welches nun das Zeitintervall enthält, kann anschließend in eine Zeitdimension überführt werden, so dass wir die in Abbildung 5 dargestellte Relation erhalten.

Um schließlich die Relation aus Abbildung 2 zu erhalten, kann abschließend eine Transformation der intervallbasierten Zeitobjekte in Zeitpunkte erfolgen.

[1] $\{\!\{\}\!\}$ stellt hierbei einen Mengentyp-Konstruktor und $\langle\!\langle\rangle\!\rangle$ einen Tupeltyp-Konstruktor dar.

produkt	*produktgruppe*	*Zeitobjekte (gültigkeit)*
P1	PG1	$\{\langle[2005\text{-}01, 2005\text{-}02]\rangle\}$
P1	PG3	$\{\langle[2005\text{-}03, 2005\text{-}04]\rangle\}$
P2	PG2	$\{\langle[2005\text{-}01, 2005\text{-}04]\rangle\}$

Abbildung 5: Restrukturierte Relation mit Zeitobjekten auf Intervallbasis

Unerwartete Fehler Bei der Datenintegration ist stets mit unerwarteten datenbezogenen Fehlersituationen zu rechnen, so dass es in diesen Situationen wichtig ist zu wissen, welche Ursache zum Fehler geführt hat. In der TIA-Algebra liefern daher zum einen Funktionen in Fehlersituationen *Exceptions*, die einen Hinweis auf das Problem enthalten. Zum anderen erfolgt bei einem aufgetretenen Fehler kein Abbruch der Operatorausführung, sondern statt dessen eine Aussortierung fehlerhafter Eingabetupel, die um die aufgetretene Exception erweitert sind. Ein Operator besitzt daher neben der eigentlichen Ergebnisrelation zusätzlich eine Ausgaberelation mit Fehlertupeln, auf die mit Hilfe eines speziellen Operators zugegriffen werden kann. In Verbindung mit dem TEMPFLOW-Prozessmodell ist eine Analyse der Fehler, eine Korrektur der Eingabedaten, eine Anpassung der Integrationsprozesse und anschließend eine konsistente (partielle) Neuausführung des Prozesses für die Transaktionszeitpunkte möglich, in denen Fehler aufgetreten sind.

4 Temporales Prozessmodell TEMPFLOW

Während die TIA-Algebra „nur“ dazu dient, temporale Relationen zu restrukturieren und zu verknüpfen und diese in neue Relationen zu überführen, ohne dass hierbei bereits ein Bezug zwischen Datenquellen und -senken eines DWS existiert, dient das formale, datenflussbasierte TEMPFLOW-Prozessmodel der Modellierung und Ausführung von Integrationsprozessen. Die Prozessausführung erfolgt hierbei inkrementell für eine voranschreitende Transaktionszeit, indem für jeden Zeitpunkt der Zustand der Senkenrelationen anhand einer Prozessspezifikation aus den Zuständen der Quellrelationen berechnet wird. Das Datenmodell der Algebra wird daher durch das Prozessmodell um eine implizite, vom Integrationssystem kontrollierte, nicht manipulierbare Transaktionszeit-Dimension ergänzt, für deren Behandlung spezielle Operatoren im Prozessmodell existieren. Im Folgenden wird überblicksartig auf die Prozessspezifikation eingegangen. Eine vertiefte, formale Betrachtung des Prozessmodells findet sich in [Har04].

Spezifikationsmittel Ausgehend von der Menge der Quellrelationen eines DWS können mit Hilfe der TIA-Algebra Ausdrücke formuliert werden, die Quellrelationen restruktieren und wie benötigt verknüpfen. Seien bspw. qr_1 und qr_2 zwei Quellrelationen aus Datenquellen. So definiert der (temporale) Algebraausdruck $\cup(\pi_{\{a_1,a_2,a_3\}}(qr_1), \pi_{\{a_1,a_2,a_3\}}(qr_2))$ eine Vereinigung der beiden Relationen, die zuvor auf die Attribute a_1 bis a_3 eingeschränkt wurden. Um die Verbindung zwischen Algebraausdrücken und Relationen der Datensenken herzustellen, werden Algebraausdrücke an Senkenrelationen, z. B. sr_1, zugewiesen: $sr_1 := \cup(\pi_{\{a_1,a_2,a_3\}}(qr_1), \pi_{\{a_1,a_2,a_3\}}(qr_2))$.

Aufgrund der transaktionszeitbasierten Zustandsorientierung ergibt sich der Zustand der Senkenrelation sr_1 für den Zeitpunkt t durch die Zustände der Quellrelationen qr_1 und qr_2 zum Zeitpunkt t. Durch die Kapselung technischer Details durch die Integrationsinfrastruktur kann sich die Prozessbeschreibung auf den reinen Datenfluss zwischen Quellen und Senken beschränken. Ausgehend von der Menge der Quell- und Senkenrelationen in einem DWS ist eine TEMPFLOW-Prozessspezifikation vereinfacht eine Menge von Zuweisungen von Algebraausdrücken über den Quellrelationen an die Senkenrelationen. Hierauf aufbauend existieren eine Reihe von Strukturierungsmitteln wie Ausdrucksvariablen und parametrisierbare Ausdrucksfunktionen, um gemeinsam genutzte Teilausdrücke nur einmal in der Spezifikation zu definieren und wiederkehrende Ausdrücke in Form von Funktionen als „neue" Operatoren bereitzustellen. Das Funktionskonzept bietet neben der Parametrisierung, die z. B. genutzt werden kann, um ein konkretes Selektionsprädikat beim Funktionsaufruf zu übergeben, auch die (partielle) Festlegung von Typen für Parameter und die als Argumente übergebenen Algebraausdrücke, so dass Funktionen möglichst allgemein definiert werden können. Bei der Typangabe können auch Typvariablen genutzt werden. Eine Typprüfung erfolgt zum Zeitpunkt der Prozessübersetzung.

Transaktionszeit, Rekursion und Prozessevolution Neben den Strukturierungsmitteln existieren spezielle Operatoren für die Behandlung der Transaktionszeit, um bspw. durch eine Transaktionszeitverschiebung auf einen alten Zustand einer Quellrelationen zuzugreifen oder die Daten anhand eines temporalen Prädikats bzgl. der Transaktionszeit zu diskretisieren. Die folgende Prozessspezifikation definiert bspw. die Zustände zweier Senkenrelationen sr_1 und sr_2 auf Basis der Quellrelation qr_1, wobei der Zustand der Relation sr_2 stets auf das letzte Chronon eines Tages diskretisiert ist und somit die Zustände für andere Transaktionszeitpunkte vernächlässigt werden (v_1 ist hier eine Ausdrucksvariable):

$$\begin{aligned} p = \{ \;& v_1 = \pi_{\{a_1,a_2,a_3\}}(qr_1), \\ & sr_1 := v_1, \\ & sr_2 := \vec{\delta}_{now \; equals \; last_{Chronons}(cast_{Days}(now)), \; event}(v_1) \; \}. \end{aligned} \tag{1}$$

Für die eigentliche Diskretisierung auf das letzte Chronon eines Tages ist das Selektionsprädikat $now \; equals \; last_{Chronons}(cast_{Days}(now))$ der Anwendung des $\vec{\delta}$-Operators zuständig, welches den aktuell betrachteten Transaktionszeitpunkt now mit dem Zeitpunkt vergleicht, der sich aus dem Skalieren des Zeitpunktes now auf die Tagesgranularität ($cast_{Days}(now)$) und anschließendes Ermitteln des letzten Chronons dieses Tages ergibt ($last_{Chronons}$).

Weitere Aspekte, die durch das TEMPFLOW-Prozessmodell unterstützt werden, sind zum einen „zeitliche" Rekursionen, um für frühere Transaktionszeitpunkte auf Ergebnisse von Ausdrücken zuzugreifen und somit Integrationsergebnisse von früheren Zeitpunkten wiederzuverwenden. Zum anderen unterstützt das Modell eine feingranulare zeitliche Versionierung der Prozessspezifikation, um Prozessanpassungen vornehmen zu können, die bspw. durch geänderte Schemata der Datenquellen oder -senken bedingt sind bzw. aufgrund bisher nicht berücksichtigter Integrationskonflikte erforderlich sind (vgl. [Har04]).

Prozessausführung Für die Ausführung einer Prozessspezifikation p wird in [Har04] eine denotationelle Ausführungssemantik definiert, die den Zustand einer Senkenrelation sr aufgrund des in der Spezifikation p hinterlegten Algebraausdruckes $p(sr)$ und der Zustände benötigter Quellrelationen definiert. Vereinfacht ergibt sich aufgrund der in Formel 1 dargestellten Spezifikation für die Senkenrelation sr_1 folgende Zustandsdefinition in Abhängigkeit des Transaktionszeitpunktes t, die durch ein nachgestelltes $[t]$ gekennzeichnet ist:

$$p(sr_1)[t] = \pi_{\{a_1,a_2,a_3\}}(qr_1[t]), \tag{2}$$

Der Zustand von $sr_1[t]$ zum Transaktionszeitpunkt t ist somit wie gewünscht unmittelbar an den Zustand der Quellrelation qr_1 für den selben Zeitpunkt t gebunden, welches durch $qr_1[t]$ dargestellt ist. Die Ausführung der Datenintegration erfolgt inkrementell für eine voranstreitende Transaktionszeit, so dass das Datenintegrationssystem bereits berechnete Zwischenergebnisse wiederverwenden kann. Liegen die für die Berechnung des Zustandes einer Senkenrelation für einen Zeitpunkt t benötigten Zustände einer Quellrelation noch nicht vor, so wird die Prozessausführung zunächst blockiert und bei Verfügbarkeit der Daten ab t fortgesetzt.

Enthält eine Spezifikation Operatoren, die sich auf die Transaktionszeit beziehen, wie bspw. eine transaktionszeitbasierte Diskretisierung oder eine zeitliche Rekursion, so ist für die Ermittlung des Algebraausdrucks, der den Zustand einer Senkenrelation zum Zeitpunkt t beschreibt, eine zeitliche Umformung der Prädikate, Funktionen etc. notwendig. Diese Umformung gewährleistet, dass ein Algebraausdruck für den alten Zeitpunkt ausgewertet bzw. der alte Zustand einer Datenquelle verwendet wird. Exemplarisch sei hierfür die Definition der Senkenrelation sr_2 zu einem Zeitpunkt t betrachtet:

$$p(sr_2)[t] = \vec{\delta}_{now\ equals\ last_{Chronons}(cast_{Days}(now)),\ event}(\pi_{\{a_1,a_2,a_3\}}(qr_1))[t]. \tag{3}$$

Im Rahmen der Umformung ist der Zeitpunkt t sukzessiv durch die Operatoranwendungen zu propagieren, bis sich schließlich der Ergebnisausdruck

$$p(sr_2)[t] = \pi_{\{a_1,a_2,a_3\}}(qr_1[t']), \tag{4}$$

ergibt, bei dem t' dem letzten Zeitpunkt entspricht, für den das Diskretisierungsprädikat $now\ equals\ last_{Chronons}(cast_{Days}(now))$ erfüllt ist, wenn now durch t ersetzt wird. Es wird somit für die Berechnung des Zustandes von $sr_2[t]$ der Zustand der Quellrelation qr_1 des früheren Transaktionszeitpunktes t' benötigt.

Durch die Zustandsorientierung der Prozessspezifikation und -ausführung wirken sich Änderungen in den Datenquellen direkt auf das Integrationsergebnis aus. Um dies bei Bedarf vermeiden zu können, bietet das Prozessmodell die zeitliche Rekursion und ermöglicht so in Verbindung mit der zeitlichen Diskretisierung das „Fixieren“ alter Daten in den Senken. Werden bspw. in den Datenquellen Daten, deren Gültigkeit älter als drei Monate zurückliegt, gelöscht, so kann im Integrationsprozess definiert werden, dass Aktualisierungen des DW inkl. Datenveränderungen aufgrund rückwirkender Änderungen jeweils nur für die letzten drei Monate erfolgen.

5 Verwandte Arbeiten

Eine Vielzahl der existierenden Datenintegrationsansätze, z. B. [Vav02, VVS+01] und auch kommerzielle Werkzeuge, beschränken sich bisher auf direkte Abbildungen zwischen Quellsystemen und einem DW oder stützen sich für die Beschreibung und Ausführung auf herkömmliche Datenbanktechnologie wie bspw. ein relationales DBMS. In Analogie zur Entwicklung von Applikationen, die mit temporalen Daten in nicht-temporalen DBMSen arbeiten müssen, ist auch bei diesen Ansätzen die Behandlung von Zeitaspekten durch SQL bzw. vorhandene Transformationssprachen nachzubilden, so dass insbesondere die Verständlichkeit und Anpassbarkeit der Prozesse beeinträchtigt wird. Andere Ansätze, z. B. [Huy97, ZWG97], die auf Sichten basieren, bieten zwar Unterstützung für eine inkrementelle und selbstwartbare Aktualisierung materialisierter Sichten, berücksichtigen jedoch keine zeitbezogenen Daten oder vernachlässigen die im DWS-Kontext notwendige Datenintegration und -transformation. Zudem können Sichten nur den Datenbestand widerspiegeln, der sich in den Quellsystemen befindet, so dass eine Historisierung bei reinen Sichtenansätzen nicht unterstützt wird.

Auf Basis einer eingeschränkten Form des in [Sno95] vorgestellten BCDM ermöglicht der Integrationsansatz von Yang et al. [Yan01] die Definition materialisierter, temporaler Sichten und folgt somit dem DWS-Gedanken mit einem redundant vorgehaltenen, integrierten Datenbestand. Der Schwerpunkt der Arbeit liegt in der Entwicklung von Algorithmen für die Selbstwartbarkeit temporaler Sichten unter Verwendung einer Operatormenge, die einer Variante der konzeptionellen TSQL2-Algebra (vgl. [Sno95]) entspricht. Entsprechend werden keine Operatoren für die Auflösung von Integrationskonflikten bereitgestellt. Zudem unterstützt das Datenmodell nur eine Zeitdimension und nutzt stets eine Menge von Zeitpunkten für Zeitangaben.

6 Zusammenfassung

Ausgehend von einem kurzen Einblick in die Zeitproblematik im DWS-Kontext wurde in diesem Beitrag ein Überblick über einen neuen Ansatz für die temporale Datenintegration, bestehend aus einer temporal erweiterten ETL-Infrastruktur, der temporalen Algebra TIA und dem Prozessmodell TEMPFLOW, vorgestellt. Die Algebra ermöglicht hierbei die für die Datenintegration benötigte Restrukturierung von Schemata und Daten, die flexible zeitliche Interpretation der Quelldaten und die Abbildung des Integrationsergebnisses auf das im Zielsystem benötigte (temporale) Format. Durch die temporalen Algebraoperatoren ist zudem eine kompakte Beschreibung temporaler Datenverknüpfungen und -selektionen gewährleistet. Aufbauend auf der Algebra dient das Prozessmodell für die eigentliche Spezifikation von Integrationsprozessen und ergänzt hierbei das Datenmodell der Algebra um die implizit vorhandene und durch das Integrationssystem kontrollierte Transaktionszeit-Zeitdimension. Mit Hilfe spezieller Transaktionszeit-Operatoren ermöglicht das Prozessmodell den Zugriff auf alte Zustände der Datenquellen und alte Integrationsergebnisse und ermöglicht hierdurch u. a. eine sichtenorientierte Spezifikation der Datenhistorisierung im DW.

Auf Basis von TEMPFLOW und TIA wurde ein Softwaresystem entwickelt, das die präsentierten Konzepte in Form einer Laufzeitumgebung für die Prozessausführung umsetzt. Im Rahmen einer Fallstudie aus dem Bereich der Telekommunikation konnte zudem die Praxistauglichkeit der Konzepte gezeigt werden. Im Einzelnen wurden in der Evaluation die in den Quelldaten vorliegenden Zeitinformationen interpretiert, bei der Datenverknüpfung genutzt und in die im DW benötigte Repräsentationsform überführt. Zudem wurden auf Basis eines Aktualisierungszeitfensters rückwirkende Datenänderungen ermöglicht, um Korrekturen an den Quelldaten in das DW aufzunehmen. Für ältere Daten wurden sowohl Fakten als auch Dimensionen als nicht mehr veränderbar angesehen, um die Konsistenz dieser Daten zu erhalten.

Literatur

[BG04] A. Bauer und H. Günzel, Hrsg. *Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung*. dpunkt-Verlag, 2. Auflage, 2004.

[DBS96] D. Dey, T. M. Barron und V. C. Storey. A complete temporal relational algebra. *The VLDB Journal*, 1996(5):167–180, 1996.

[Har03] A. Harren. Towards Semantic Temporal Support in Data Integration. In R. Jardim-Gonçalves, J. Cha und A. Steiger-Garçao, Hrsg., *Proceedings of the 10th ISPE International Conference on Concurrent Engineering: Research and Applications*. A. A. Balkema Publishers, 2003.

[Har04] A. Harren. *Temporale Datenintegration in Data-Warehouse-Systemen*. Dissertation, Carl v. Ossietzky Universität Oldenburg, Department für Informatik. dissertation.de - Verlag im Internet, 2004.

[Huy97] N. Huyn. Multiple-View Self-Maintenance in Data Warehousing Environments. In M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos und M. A. Jeusfeld, Hrsg., *Proceedings of the 23rd VLDB Conference, Athen, Griechenland*. Morgan Kaufmann, 1997.

[Sno95] R. T. Snodgrass, Hrsg. *The TSQL2 Temporal Query Language*. Kluwer Academic Publishers, 1995.

[SPD92] S. Spaccapietra, C. Parent und Y. Dupont. Model Independent Assertions for Integration of Heterogeneous Schemas. *The VLDB Journal*, 1(1):81–126, 1992.

[Vav02] A. Vavouras. *A Metadata-Driven Approach for Data Warehouse Refreshment*. Dissertation, Universität Zürich, Wirtschaftswissenschaftliche Fakultät, 2002.

[VVS+01] P. Vassiliadis, Z. Vagena, S. Skiadopoulos, N. Karayannidis und T. K. Sellis. ARKTOS: towards the modeling, design, control and execution of ETL processes. *Information Systems*, 26(8), 2001.

[Yan01] J. Yang. *Temporal Data Warehousing*. Dissertation, Stanford University, Department of Computer Science, 2001.

[ZWG97] Y. Zhuge, J. L. Wiener und H. Garcia-Molina. Multiple View Consistency for Data Warehousing. In *Proceedings of the 13th International Conference on Data Engineering, Birmingham, Großbritannien*, 1997.