„Daten auf die hohe Kante legen!“
Datenbankarchivierung bei der ING-DiBa

Diana Richler
ING-DiBa AG
Nürnberg

Schlüsselworte
Datenbankenarchivierung, Archivsystem, Archivierungskonzepte, Archivierungsprojekte, Chronos.

Einleitung

Was tun, wenn die Datenbank zum Archiv wird. Löschen unmöglich. Anwenderzugriff vielleicht. Viele Datenbanken sind über Jahre gewachsen und halten Buchungen im teuren Speicher für den sofortigen Zugriff bereit. Aber muss das so sein? Ziel unseres Projekts war nicht, eine Einzelarchivierung umzusetzen, sondern einen Prozess aufzubauen, der das Archivierungskonzept in allen Bereichen der Bank etabliert.

Die Einführung einer neuen Software ist die Chance einen Prozess von klein auf zu designen und so zu steuern, dass ein Standard geprägt wird. Die Einführung eines Fragenkatalogs, der die Ziele und Eckpunkte eines Archivierungsprojekts erfasst, dient als Entscheidungsbasis der Verantwortlichen. Zudem unterstützen Verfahrensanweisungen und eine zentrale fachliche Anlaufstelle bei der Projektbeantragung, -Umsetzung und Überführung in den Regelbetrieb.

Gerne geben wir Einblicke in das erste Projekt: Die Archivierung aus der zentralen Buchungsanwendung des Accounting. Aufgrund der großen Datenmengen im System mussten Maßnahmen ergriffen werden, die Buchungsdaten revisionssicher zu archivieren, aber dennoch einen performanten Zugriff für den Fachbereich zu ermöglichen. Im Rahmen dieses Projekts wurde Chronos von der Firma CSP eingeführt und in die ING-DiBa Landschaft integriert.

Noch während der Entwicklung des ersten Projekts zeichneten sich weitere Anwendungsmöglichkeiten ab. Von einmaligen Archivierungen, um Altdatenbanken abschalten zu können, bis hin zur Archivierung von Kundenbewegungsdaten (Kontakthistorie) mit JDBC-Zugriff auf die Archive aus den bestehenden Anwendungen heraus. Wir wollen zeigen wo Archivierung mit Chronos bei uns eingesetzt werden soll und was wir damit alles bewirken können.

Sparanlagen und Datengräber

Bei der Suche nach einem Archivierungssystem wurde nicht zwingend nach einer typischen Anwendung, die im Paket der Datenbanken angeboten wird, gesucht. Sondern eher eine Archivierungslösung, die nur bei einer Datenbanktabelle ansetzt und nicht das große Ganze, die Datenbank, sieht.

Bei uns im Unternehmen geht es selten darum, die ganze Datenbank bzw. ein komplettes Schema zu archivieren. Es geht darum, Daten aus einer Hand voll Tabellen einfach, aber sicher zu archivieren, um die Datenbankgröße und somit auch Recovery Zeiten zu verkleinern. Vor allem Datenbanken mit Kunden-Stamm- oder Buchungsdaten sind erste Kandidaten für regelmäßige Archivierungen der zurückliegenden Bewegungsdaten.

Am Beispiel der Kunden-Stammdaten mit einer Datenbank-Größe von aktuell 700 GB zeigt sich schnell, wo eine Sparanlage zu finden ist. Bei der ING-DiBa werden für die unterschiedlichen Releases eine Entwicklungs- und eine Test-Datenbank mit den anonymisierten Kunden-Stammdaten aus Praxis zur Verfügung gestellt. Da nicht immer nur an einem Release gearbeitet wird, die Kunden-Stammdatenbank aber die zentrale Datenbank ist, stehen bis zu sieben Entwicklungs- und Testdatenbanken zur Verfügung. Hinzu kommen noch eine Vorproduktions- und eine Schulungs-Datenbank. Ein Abzug dauert trotz Automatisierung der Prozesse aufgrund der Datenmenge ca. 14 Stunden (8 Stunden für den Datenbank Clone und 6 Stunden für die Anonymisierung).

Durch eine Archivierung der Bewegungsdaten, die älter als 2 Jahre sind, lassen sich in der Praxis ca. 300 GB gewinnen. Diese Ersparnis wirkt sich nicht nur auf den Plattenplatz der Praxis aus, sondern auch auf die 16 anderen Umgebungen. Um möglichst praxisnah entwickeln und testen zu können, sind

auch die anderen Umgebungen nahezu identisch zur Praxis ausgerüstet und ebenso teuer. Eine einmalige Archivierung würde aber nicht von Dauer sein, so dass es sich empfiehlt, kontinuierlich Daten zu archivieren und auch auf Strukturänderungen in der Datenbank reagieren zu können. Die Folgearchivierungen liegen dann nur noch bei wenigen GB im Monat. Ein angenehmer Nebeneffekt der Archivierung sind auch die verbesserten Suchlaufzeiten aller Applikationen, die über die nach der Archivierung verkleinerten Tabellen suchen.

Ein weiterer Ansatzpunkt sind Altdaten. Diese Daten fallen im Laufe der Zeit immer wieder bei Fusionen an, bei denen Applikationen zunächst migriert, aber die Ursprungsdatenbank nie abgeschaltet wird. Wirklich sicher lässt sich leider nie sagen, ob die Daten noch zu Recherchezwecken benötigt werden. Und so ist es in der Summe zunächst vermeintlich günstiger die Daten zumindest in der Datenbank verfügbar zu halten, auch wenn keine Applikation mehr aktiv zugreift. Der dafür notwendige Plattenplatz und die Lizenzgebühren sind ein fester Bestandteil im Budget. Solche Datengräber zeichnen sich dadurch aus, dass eine einmalige Archivierung ausreicht und keine Strukturänderungen zu erwarten sind. Schnell lassen sich so in Summe 700 GB in Datenbanken einsparen und diese nach Archivierung komplett abschalten.

Deswegen wurde nach einer Software gesucht, die folgende Kriterien erfüllt:

- Revisionssicher archivieren mit der EMC Centera
- Langfristige Unabhängigkeit bei der Wahl des Archivformats
- Export und Import von und mit verschiedenen Produktivdatenbanken
- Flexibilität bei den Tabellenstrukturen
- Sichere und automatisierte Löschung der Quell-Archivdaten
- Löschen der Archive nach Ablauf der Retention Time
- Mannigfaltige Recherchezugriffsmöglichkeiten auf die Archive
- Integration in die ING-DiBa-Systemlandschaft zum Aufruf durch UC4

Aufbau unseres Archivsystems

Bereits 2008 fand im Unternehmen eine Untersuchung für ein Archivsystem statt, welches unseren Anforderungen entspricht. Als bestens für unsere Bedürfnisse geeignet wurde die Archivierungssoftware Chronos von der Firma CSP befunden, so dass ein erstes Pilotprojekt Mitte 2011 startete. In diesem Zuge wurde ein Großteil der Infrastruktur aufgebaut, siehe Abb. 1.

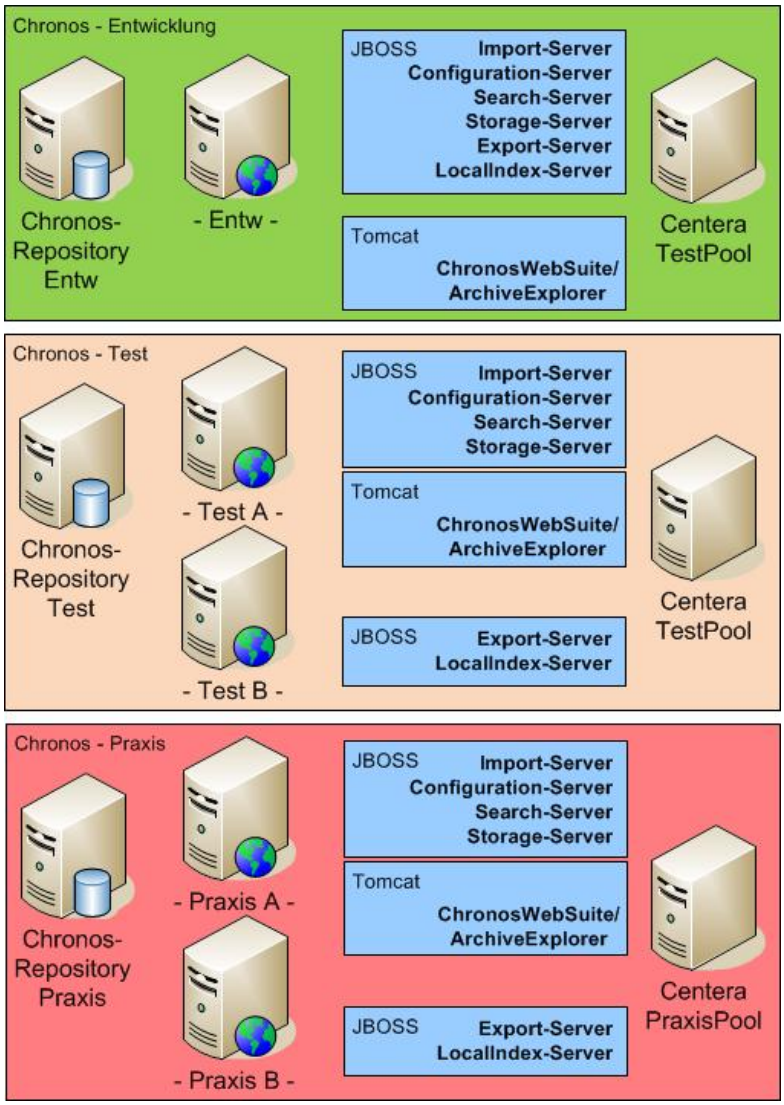

Abb. 1: Umgebungsaufbau Chronos in der ING-DiBa

Chronos besteht aus den 6 Servern für die wichtigsten Bereiche der Archivierung:

- Configuration
- Export
- Import
- Index
- Search
- Storage

Hinzu kommen die Administrationsoberfläche AdminSuite und die Recherche-Oberfläche WebSuite. Überzeugt hat Chronos, weil die revisionssichere Archivierung unter Verwendung der bereits bei uns im Einsatz befindlichen Hard- und Software möglich ist. So kann Chronos den Ex- und Import aus Oracle-Datenbanktabellen in Verbindung mit der Archivierung auf der EMC Centera leisten. Das dabei verwendete Archivformat ist mit gezippten txt-Dateien für die Daten und XML-Dateien für die Tabellenstrukturinformationen transparent und letztendlich unabhängig vom Archivierungsprodukt. Eine Löschung der Exportdaten auf der Quelldatenbank ist ebenso möglich, wie die abschließende Löschung der Archive nach Ablauf der definierten Retention Time. Zudem bietet Chronos die Möglichkeit, neben den beiden eigenen Anwendungen Admin- und WebSuite auch per JDBC-Treiber auf die im Archiv abgelegten Daten zuzugreifen.

Im Pilotprojekt wurde zunächst mit je einem Server in der jeweiligen Umgebung begonnen. Die Tests haben jedoch gezeigt, dass es bei besonders vielen und umfangreichen Suchen auf den Archiven von Vorteil ist, die Exporte von den Suchanfragen zu trennen. Aus diesem Grund sind unsere Test- und Praxismaschinen auf je zwei Maschinen pro Umgebung aufgeteilt, die zum einem den Export- und Index-Server beherbergen und zum anderen die restlichen Server. Die Entwicklungsumgebung wird nur für Update-Einspielungen und kleinere Tests verwendet, weshalb uns hier ein Sever ausreicht. Zusätzlich zum JBoss-Server benötigt Chronos einen Tomcat für die WebSuite/ArchiveExplorer und das EMC SDK für die Verbindung zur Centera. Die AdminSuite kann lokal installiert werden oder, wie bei der ING-DiBa, über Citrix angeboten werden. Alle Konfigurationen, die in der AdminSuite vorgenommen werden, werden im Chronos-Repository, einer Oracle-Datenbank, abgelegt. Diese Datenbank befindet sich in der unternehmensüblichen Sicherung, so dass für die Infrastruktur keine größeren Anschaffungen getätigt werden mussten.

Neben der Infrastruktur ist der Aufbau des Archivierungsprozesses ein wesentlicher Bestandteil des ersten Projekts gewesen. Ziel war und ist es, nicht nur einmalig Daten gewinnbringend zu archivieren, sondern langfristig einen Standardprozess im Unternehmen zu etablieren, der die Folgeprojekte einfach für den Projektinitiator und alle Projektbeteiligten macht.

Deswegen wurde ein Fragenkatalog erstellt, anhand dessen es möglich ist, die Vorteile für eine Archivierung mit Chronos herauszustellen. Ebenso soll aber auch schnell aufgezeigt werden können, wenn Chronos ungeeignet ist. So ist eine grundlegende Frage eines jeden Projekts, ob Daten aus einer Datenbanktabelle archiviert werden sollen, oder ob als Gegenbeispiel aus einem Dateisystem eine Archivierung gewünscht wird. Auch wenn es sicherlich mit Chronos Möglichkeiten gibt, diese Archivierungen zu unterstützen, wollen wir uns vorerst auf die Datenbanktabellen beschränken.

Für die richtige Chronos-Lizenz ist entscheidend, ob es sich um Einmalarchivierung oder kontinuierliche Archivierung eines Einzelschemas oder Multischemas handelt. Des Weiteren gibt es das „Continuus Archiving“, dass in die bereits bestehenden Archivpakete weitere Archive ergänzen kann. Für uns sind hier bisher Single-Schema- und Einmalarchivierungen relevant.

Sollen High-Risk-Daten archiviert werden, so ist die revisionssichere Archivierung auf der EMC Centera Pflicht. Muss das nicht sein, kann auch ein kostengünstigerer Speicherplatz gewählt werden. Dann ist aber auch hier auf die ausreichende Sicherung der Daten zu achten.
Die Menge der zu archivierenden Daten wird benötigt, um den Speicherbedarf der Archivablage abschätzen und bereitstellen zu können. Er ist außerdem wichtig, um überhaupt über den Nutzen der Archivierung entscheiden zu können. Wenn die Datenmenge zu gering ist, muss der Aufwand in Relation gesetzt werden.

Weitere Kriterien für die Abschätzung des Projektaufwands sind die Anzahl der Archivtabellen und deren Verknüpfungen, da diese bei der eigentlichen Anlage im Chronos zwar wenig Arbeit verursachen, aber die Laufzeit einer Archivierung (Ex- und Import, aber auch das Löschen) beeinflussen. Ob und wann die Daten nach dem Export aus der Quell-Datenbank gelöscht oder nur markiert werden, erhöht ebenfalls die Joblaufzeiten eines Archivierungslaufs.
Je nach Gewichtung spielen noch Auswertungen bzw. Recherchezugriffe auf den Archiven eine Rolle. Hier müssen ggf. Reports für den Fachbereich erstellt werden, die genauen Vorgaben unterliegen. Je nach Anzahl und Vorgaben sollte der Posten in der Aufwandsschätzung beachtet werden. Wenn der Fachbereich mit den zur Verfügung stehenden Mitteln Admin- oder WebSuite nicht zufrieden zu stellen ist, kann auch in eigenen Applikationen ein Zugriff auf die Archive per JDBC-Treiber angebunden werden.

Diese wichtigen Eckpunkte dienen als Basis zur Erstellung einer Entscheidungsvorlage für unser Management, damit nicht jeder Archivierungstrend aufgegriffen wird und bedacht mit dem Centera-Plattenplatz umgegangen wird. Ist die Entscheidung für ein Projekt gefallen und liegt ein Auftrag vor,

kann gestartet werden. Dabei akzeptieren alle Beteiligten die zugrundeliegende Verfahrensanweisung, in der alle Rollen, Schnittstellen und Prozesse definiert sind. Diese abteilungsübergreifende Vereinbarung sichert einen reibungslosen Projektablauf und auch spätere Linientätigkeiten ab.

Gut gesichertes Datenkapital

Das erste Projekt mit dem Namen Buchungsdaten umfasst fünf Datenbanktabellen und eine Materialized View, die Chronos analog zu den Tabellen behandelt. Die Objekte werden für jeden Archivierungslauf mit den bereits vorselektierten Daten gefüllt und dann von Chronos komplett archiviert, sowie nach erfolgreicher Ablage auf der Centera in der Quell-Datenbank gelöscht. Der Weg über die Archive-Tabellen wurde gewählt, weil auf Tabellenstrukturen eines Drittanbieters, dem Anbieter der Buchungsanwendung, zugegriffen wird. Ziel ist es auch für weitere Projekte, auf separaten Tabellen zu archivieren, um unabhängig von den Tabellen anderer zu sein und auch eine klare Abgrenzung der von Chronos abgegriffenen Tabellen zu den Live-Tabellen vorweisen zu können.

Eine wichtige Anforderung für die Archivierung ist die revisionssichere Ablage mit einer Aufbewahrungsfrist von 10 Jahren ab Einstellung in das Archiv. Hier gewährleistet zum einen Chronos die Sicherheit, dass die Quell-Daten erst gelöscht werden, wenn die Centera die erfolgreiche Ablage zurückgemeldet hat und zum anderen garantiert die Centera die unveränderliche Archivierung für die kommenden 10 Jahre. Nach Ablauf der Retention Time müssen die Daten von der Centera gelöscht werden, das kann ebenfalls über Chronos initiiert werden, muss aber aktuell noch manuell (Anstoß über die AdminSuite) erfolgen und kann noch nicht per Chronos-Job konfiguriert werden.

Neben dem Export sollte auch die Möglichkeit eines Imports zurück in die Quell-Datenbank offen gehalten werden. Das Projekt wollte für eine evtl. Revisionsprüfung die Daten auch zurückführen können. Als technische Rahmenbedingung ist der Aufruf der Archivierung per UC4, einer Prozess-Automatisierungssoftware, die bei uns eingesetzt wird, Pflicht. Außerdem sollen alle Datenfelder in der Archivierung analog zur Praxis verfügbar sein. Chronos überträgt die Export-Daten eins zu eins ins Archiv und behält hier die Tabellen, Spaltenbezeichnungen sowie die Datentypen bei, so dass dies kein Problem darstellt.

Die fachlichen Anforderungen hingegen waren in diesem Projekt die größte Herausforderung. Die Archivierung aller Buchungen älter 25 Monate und ein Archivierungsrhythmus alle ein bis drei Monate sind mit Chronos schnell abzubilden. Gegen den Anspruch des Fachbereichs, die Daten im Archiv genauso schnell selektieren zu können, wie in Praxis, musste erst einige Aufklärungsarbeit geleistet werden.

Mit Chronos und dort definierten lokalen Indizes lässt sich die Suche auf den Archivdaten beschleunigen. Auch der bereits vordefinierte globale Index von Chronos spielt hier eine wichtige Rolle. Die Vorgabe bei einer Archivsuche über mehrere Jahre, mit zum Teil optionalen Parametern, in einer Suchzeit unter 30 Sekunden zu bleiben, lässt sich aber trotzdem nicht immer halten. Auch wenn die Archivpakete von monatlich auf täglich verkleinert wurden, da so zwar mehr Pakete geschrieben werden, diese aber auch schneller bei Suchen ausgeschlossen werden können. Ein Archiv ist letzten Endes keine Datenbank und so muss bei großen Suchanfragen eine gewisse Laufzeit einkalkuliert werden. Bei besonders aufwendigen Suchen nutzt der Fachbereich deswegen die E-Mail-Option, bei der das Ergebnis zugesendet wird.

Der Aufbau der Suchmasken bzw. des Ergebnisses sollte sich an den bereits verwendeten Masken und Excels des Fachbereichs aus der Praxis-Anwendung orientieren. Für die Fachbereichs-Recherchen wurde die WebSuite vorgestellt. Die eigentlichen Suchmasken können in der AdminSuite über Reports angelegt werden. Diese werden dann in der WebSuite angeboten. Da speziell die Buchungen für Auswertungen des German und International Accounting verwendet werden, war eine Ergebnisausgabe und Aufbereitung in Excel eine feste Vorgabe. Letztendlich wurden zwei Reports

umgesetzt, die für die beiden Fachbereiche die Daten auf Basis der archivierten MView aufbereiten. Chronos arbeitet hier mit JasperReports und deren Framework, so dass eine Bearbeitung der Reports mit iReports erfolgen kann. Zu kleineren Herausforderungen führten hier noch Formatvorgaben für Datums- und Zahlenfelder, weil diese nicht beim Aufbau der MView beachtet wurden. Chronos in Kombination mit den Reports beim Selektieren der Daten nicht alle SQL-typischen Befehle zulässt. Trotzdem können die Excels vergleichbar zu den Praxis-Excels der Buchungsanwendung erstellt werden.

Das Gesamtergebnis der Archivierung zeigt, dass sich der Aufwand gelohnt hat. Ziel war es, die Praxis-Anwendung durch die Auslagerung aller Buchungen älter als 25 Monate für die Jahresendläufe zu beschleunigen und für die kommenden Jahre zu stabilisieren. Dieses Ziel wurde erreicht, indem 10 GB aus der Praxis-Datenbank auf die Centera archiviert wurden. Hier macht der Speicherbedarf nach der Komprimierung nur noch 3,9 GB aus.

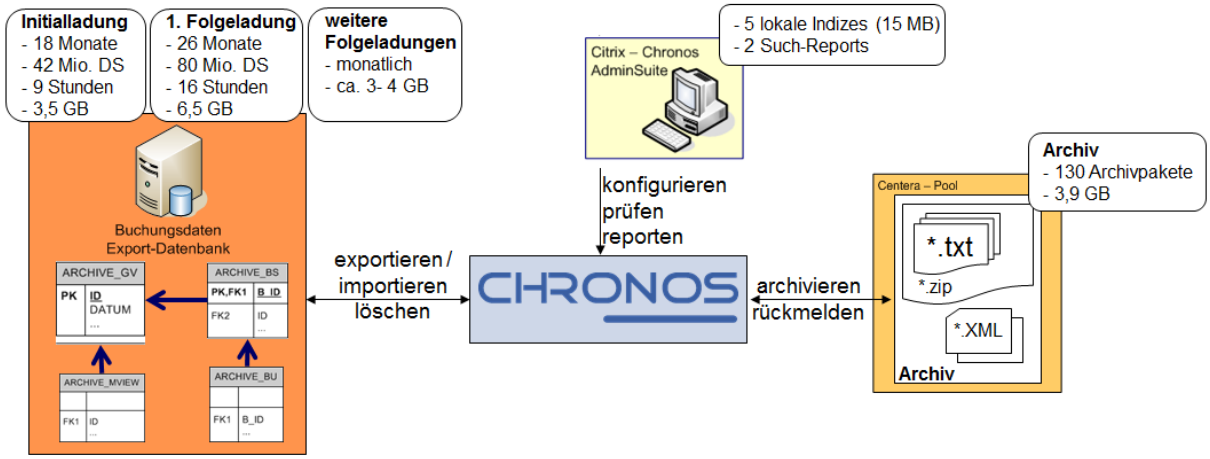

Abb. 2: Ergebnis des ersten Projekts Buchungsdaten

Abb. 2 zeigt, in welchen Schritten die Archivierung vorgenommen wird. Hierbei durfte das Tagesgeschäft und vor allem die Ultimoläufe nicht beeinflusst werden. Darum wurden Wochenenden für die ersten beiden Ladungen und die Nacht für die monatlichen Folgeladungen gewählt.

Als weiteres geeignetes Projekt für Chronos wurde die Archivierung der Homebanking Computer Interface Protokolldaten auf genommen. Diese Daten entstehen, wenn ein Kunde über die HBCI-Schnittstelle auf seine ING-DiBa-Konten zugreift. Diese Protokolldaten sind in den letzten Jahren stark gewachsen, so dass sich eine Archivierung aller älteren Daten anbietet, um die Datenbank zu verkleinern und dort Speicherplatz einzusparen.

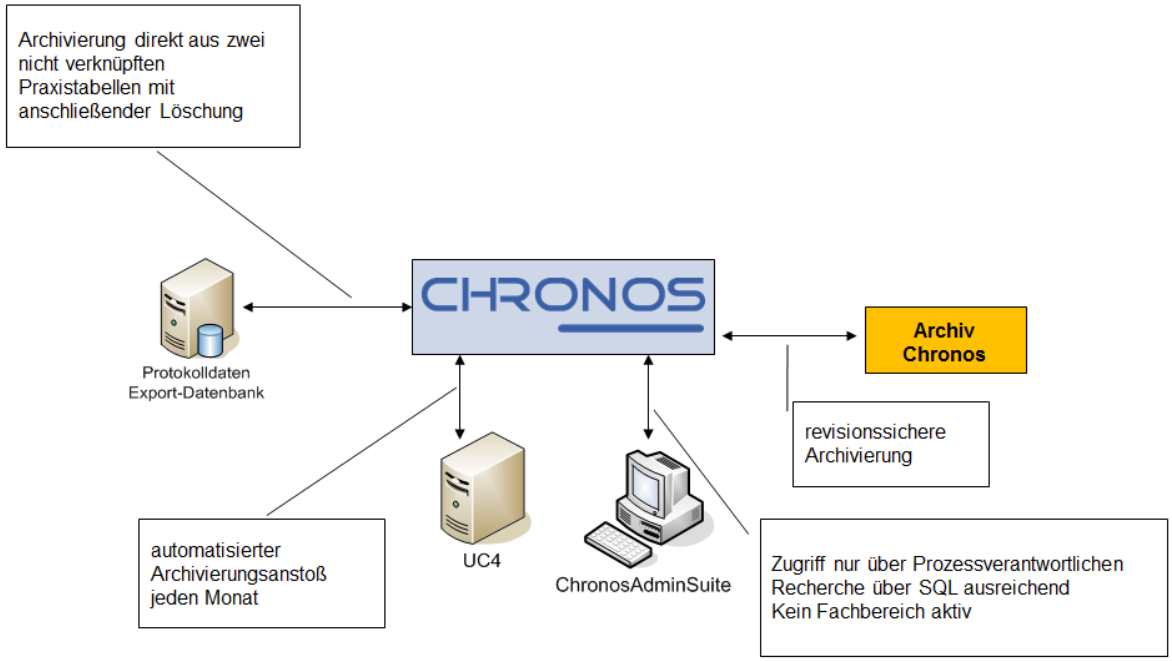

Abb. 3: Anforderungen aus dem Projekt HBCI Protokolldaten

Da in diesem Projekt nur zwei Datenbanktabellen archiviert werden und ein Recherchezugriff nur für den Prozessverantwortlichen gewährt werden muss, ist der Aufwand überschaubar. Wichtigster Teil ist die revisionssichere Archivierung. Die aktuelle Größe der Praxis-Logdaten beträgt 380 GB. Hiervon müssten ca. 90% in die Archive ausgelagert werden können. Da das Projekt aktuell noch läuft, kann zu dem wirklich benötigten Centera-Plattenplatz noch keine Aussage getroffen werden. Aufgrund der einfachen Struktur der Daten kann aber mit einer Plattenersparnis über 50% gerechnet werden.

Lust auf mehr Daten

Die bisherigen Erfahrungen zeigen, dass mit Chronos die richtige Entscheidung getroffen wurde, um die Daten-Sparanlagen richtig zu nutzen und die Datengräber platzsparend zu archivieren. Für das Pilotprojekt hatten wir uns Zeit erbeten, um auch den gesamten Prozess richtig aufbauen zu können. Mit Ende des Jahres sind alle Aufbauarbeiten abgeschlossen, so dass wir uns den ca. 700 GB für Einmalarchivierungen und vielen 100GB Dauerarchivierungen widmen können. In diesem Zuge werden wir an dem Konzept der temporären Archive-Tabellen, wie wir sie im ersten Projekt verwendet haben, festhalten, um eine klare Trennung von Live-Tabellen vorweisen zu können. Außerdem bietet dieser Ansatz den Vorteil, dass auch bei nicht taggleicher Archivierung die in den Archiv-Tabellen gelagerten Daten revisionssicher abgelegt sind. Die Laufzeiteinsparungen nach Archivierungen in den Praxis-Anwendungen und die Plattenplatzersparnis macht definitiv Lust auf mehr Projekte.

Kontaktadresse:

Diana Richler
ING-DiBa AG
Südwestpark 97
D-90449 Nürnberg

Telefon: +49 (0) 911-149 22 583
Fax:
E-Mail d.richler@ing-diba.de
Internet: www.ing-diba.de