Wat moeten we archiveren en op welk formaat?

Over honderd jaar nog een CD-ROM lezen

Toon Loonen

Archivering en opschonen van de gegevens in de database is een aspect dat vaak bij de architectuur of ontwerp van een systeem wordt overgeslagen. Soms wordt het punt opgepakt tijdens het technisch ontwerp en worden er nog een paar schoningsprogramma's gebouwd. Soms wordt pas actie ondernomen als na enkele jaren de database te groot dreigt te worden en er snel iets moet gebeuren om de productie niet in gevaar te brengen.

Het bouwen van de schoningsprogrammatuur kan vaak wel wachten tot enige tijd na het in gebruik nemen van het systeem, maar het meenemen van deze functionaliteit in het ontwerp is wel belangrijk. Hier kunnen namelijk de goede afwegingen gemaakt worden over wat na het schonen nog vanuit de productiedatabase gerapporteerd kan worden en wat niet meer. Vervolgens kan deze informatie gebruikt worden om de grootte van de database na enige tijd productie te berekenen.

Het meenemen van archivering in de architectuur en ontwerp is noodzakelijk om later ook een bruikbaar gegevensarchief te hebben. Daarbij komen de bekende ontwerpvragen: "Waarom, Wat, Hoe, Waarmee en Waarop?" weer terug.[1]

WAAROM ARCHIVEREN

Zoals bij elk ontwerp moet ook bij het ontwerp van het archiveren het doel goed overwogen worden. Denk hierbij aan:

- Gegevens zijn niet (nooit) meer nodig;
- Gegevens zijn niet meer nodig voor dagelijkse operaties maar alleen voor historisch onderzoek of wettelijke bewaartermijnen;
- Gegevens zijn mogelijk nog nodig maar de database wordt te groot en off-line opslaan met langere toegangstijden is acceptabel.

Als de gegevens echt niet (nooit) meer nodig zijn, dan kunnen deze uit de database worden verwijderd zonder deze te archiveren. Soms is het echter noodzakelijk om gegevens langer te bewaren:

- Er zou incidenteel nog naar gevraagd kunnen worden door de organisatie of relaties;
- Er is een wettelijke bewaartermijn, bijvoorbeeld de financiële gegevens van een bedrijf voor belastingtechnisch onderzoek;
- Er zijn wettelijke bewaartermijnen voor veel gegevens bij overheidsorganisaties;
- Sommige gegevens moeten zeer lang bewaard worden voor wetenschappelijk onderzoek, zoals weermetingen van lange tijd geleden die inzicht kunnen geven in het broeikaseffect en bijvoorbeeld kunnen historici over 100 jaar geïnteresseerd zijn in een onderzoek naar de belastingen in vroegere tijden.

Bepaalde informatie in politie-databases mag wettelijk na een zekere termijn niet meer gebruikt worden en moet uit de databases verwijderd worden.

Als het mogelijk is om de gegevens in de database te bewaren zolang dat noodzakelijk is voor raadplegen of wettelijke termijnen,

Bijzondere aspecten

In dit artikel wordt alleen gekeken naar het archiveren van gegevens uit databases. Archivering van andere elektronisch opgeslagen gegevens (Word-documenten, bouwtekeningen, etcetera) en formulieren (bankopdracht, belastingaangifte, opgave en claim bij verzekering) wordt in dit artikel niet behandeld. Maar ook bij deze archieven komen de vragen "Waarom, Wat, Hoe, Waarmee en Waarop?" weer terug. Verder spelen hier nog andere bijzondere aspecten:

- Wettelijke bepalingen: wat moet wettelijk nog op papier (bijvoorbeeld loonstrookje), mede omdat de wet vaak achterloopt bij de technische mogelijkheden;
- Handtekening: moet een formulier van een echte handtekening voorzien zijn of is een scan van het formulier ook bruikbaar?
- Beschikbaarheid hardware en software om later de oude bestanden nog te kunnen lezen;
- Opslagcapaciteit, zowel van de papieren versies als van elektronische versies;
- Het op meer plaatsen kunnen raadplegen van documenten.

dan heeft dat de voorkeur. Er is dan geen speciale programmatuur nodig om te archiveren respectievelijk om de gearchiveerde gegevens te raadplegen.

Vaak is dit niet mogelijk omdat de database te groot wordt met alle nadelen van dien:

- Dure opslag op schijf in plaats van op tape of CD-ROM;
- De backup (en eventueel een restore) gaat langer duren en kost ook meer opslagcapaciteit;
- De performance van bepaalde functies kan achteruit gaan.[2]

In dat geval kunnen de gegevens die slechts incidenteel nodig zijn naar een archief geschreven worden en uit de database verwijderd. Het archiveren voor historisch of wetenschappelijk onderzoek zal altijd een zeer lange termijn betreffen. Het bewaren in de productiesystemen is dan geen optie meer en er moet een speciaal archiefsysteem voor dit doel worden opgezet, mogelijk met een eigen (van het systeem afgeleide maar vereenvoudigde) gegevensstructuur.

Een databasedump zal niet vaak een bruikbaar archief opleveren

Een speciale situatie voor schonen en archiveren doet zich voor als een systeem naar een geheel nieuw systeem wordt geconverteerd. Dan moet overwogen worden welke gegevens uit het oude systeem naar het nieuwe systeem worden meegenomen en hoe de niet geconverteerde gegevens uit het oude systeem nog tijdelijk voor de gebruikers beschikbaar blijven.

WAT ARCHIVEREN

Ik maak bij dit soort afwegingen altijd onderscheid tussen enerzijds referentiegegevens, bijvoorbeeld tabellen waarvan de gegevens relatief weinig wijzigen zoals een artikeltabel, klantentabel, landentabel of BTW-tabel en anderzijds transactiegegevens, tabellen die vaak gemuteerd worden (voornamelijk inserts) waarin de gegevens echter maar een korte tijd van belang zijn, denk hierbij aan orders of facturen.

Transactiegegevens kunnen worden geschoond als de betreffende transacties geheel zijn verwerkt en ook niet meer geraadpleegd behoeven te worden via het systeem ofwel indien de wettelijke bewaartermijn is verstreken.

Vaste gegevens zullen niet worden geschoond of alleen als deze lange tijd niet meer gebruikt zijn, bijvoorbeeld klanten die twee jaar niets meer gekocht hebben. Deze termijn is een functioneel aspect waarover de gebruiker een uitspraak moet doen.

Van referentiegegevens waarvan ook de historie wordt bijgehouden, bijvoorbeeld de BTW-tabel of artikelprijzen, kan de historie worden verwijderd nadat deze niet meer in het systeem geraadpleegd wordt en er ook geen andere gegevens meer naar verwijzen.[3]

Bij het verwijderen (al of niet in combinatie met archiveren) moet de database consistent worden gehouden. Denk hierbij met name aan referentiële integriteit en correcte waarden van redundant opgenomen gegevens.

Dit betekent voor transactiegegevens dat in het gegevensmodel een entiteit (order header) en alle daar naar refererende entiteiten (orderregels, afleveringen, etcetera) worden verwijderd. Een alternatief is om alleen de details van de order te verwijderen maar de order header te laten staan. In deze header wordt een "archief-indicator" aangezet of een verwijzing opgenomen naar de CD of tape waarop de details teruggevonden kunnen worden.

Bij referentiegegevens ligt deze zaak vaak ingewikkelder: En klant kan bijvoorbeeld worden verwijderd als deze na twee jaar niets meer heeft gekocht. Maar er mag ook niets meer naar deze klant refereren, dus ook geen order header van een verder verwijderde of gearchiveerde order. Bij een tabel als de BTW-tabel, waarbij de primaire sleutel bestaat uit bijvoorbeeld BTWCODE + ingangsdatum + einddatum (mogelijk is deze laatste datum in het systeem opgenomen als de ingangsdatum van het volgende record) mogen oude records verwijderd worden als de einddatum voldoende ver in het verleden ligt en er ook geen gegevens (orderregels) meer naar deze combinatie van code en datum verwijzen.

HOE ARCHIVEREN

Voor het archiveren van databasegegevens hebben we verschillende mogelijkheden:

- databasedump;
- export van de te archiveren tabellen of subset van tabellen naar een archief database;
- export van de te archiveren gegevens naar bestand in ASCII- of XML-formaat;
- export van de gegevens in een vereenvoudigde gegevensstructuur naar een archiefdatabase, datawarehouse of bestand in ASCII- of XML-formaat;
- Niet de basisgegevens maar de uitvoer van het systeem wordt gearchiveerd, zoals de print van orders, facturen en betaling, liefst niet op papier maar in ASCII-, PDF- of Word-formaat op CD-ROM).

De hoeveelheid opslagcapaciteit kan nogmaals gereduceerd worden door de betreffende bestanden te comprimeren. Een export van databasetabellen wordt daarbij mogelijk nog met een factor vier tot tien verkleind.

Al deze vormen hebben voordelen en nadelen. Om een goede keuze te kunnen maken moet het doel van het archief weer voor ogen gehouden worden.

Gaat het alleen om het incidenteel bekijken van een oude order of factuur naar aanleiding van een vraag van een gebruiker, auditor of de belastingdienst, dan kan een CD-ROM met daarop het archief van de orders of facturen (in printvorm) van een bepaalde maand voldoende zijn. Om het archiefbestand niet te groot te maken moet de opmaak ontdaan zijn van opsmuk als plaatjes (logo's) maar de betekenis van alle velden moet wel duidelijk blijven, ook als het normale voorgedrukte papier niet meer gebruikt kan worden. Eventueel kan een pagina opnieuw worden afgedrukt.

Is het bestand niet in ASCII-formaat maar bijvoorbeeld in PDF- of WORD-formaat opgeslagen, dan moet ook de betreffende software (en de ondersteunende hardware en operating system) beschikbaar blijven zolang deze bestanden geraadpleegd moeten kunnen worden.

Voor historisch onderzoek op zeer lange termijn is export naar XML gewenst

Wordt het archief ook gebruikt voor trendanalyses binnen het bedrijf dan is opslag in een database of datawarehouse met mogelijk een vereenvoudigde structuur (sterschema structuur, eventueel enkele sterschema's vanuit verschillende gezichtspunten[4] een goede optie. Regelmatig, bijvoorbeeld maandelijks, worden de gegevens vanuit het productiesysteem naar het archief of datawarehouse overgehaald en worden de gegevens in de productie-database verwijderd.

Moeten de gegevens met de bestaande geprogrammeerde client-toepassingen kunnen worden bekeken, dan is het beter om de gegevens zolang mogelijk in de productiesystemen te laten staan. Zeker bij ingewikkelde gegevensstructuren is dit gewenst.
Als dat niet meer mogelijk is (in verband met kosten, performance of doorlooptijd van de backup) dan kan een omgeving worden ingericht die functioneel gelijk is aan een productieomgeving. Hierin kan een export geladen worden van alle facturen van een bepaalde, bijvoorbeeld maandelijkse, archiefrun. Als deze niet meer nodig is kan men deze archiefrun verwijderen en een volgende run laden. Een alternatief is om uit een archiefrun alleen de gezochte facturen in deze omgeving te laden. Bedenk dat hierbij behalve het archief (database) ook de DBMS-software en de software van de toepassing nodig is.

DE ZEER LANGE TERMIJN

Voor historisch onderzoek op zeer lange termijn is een export van de tabellen naar een ASCII- of XML-bestand gewenst. XML is in feite ook een ASCII-formaat en dit formaat is waarschijnlijk nog zeer lang leesbaar, langer dan databasedumps die afhankelijk zijn van een bepaalde versie van de software van een bepaalde leverancier. XML heeft een voordeel dat de betekenis van elk attribuut bij elk attribuut opnieuw wordt vastgelegd. Bij hele grote bestanden is dit ook weer een nadeel: het bestand wordt weer veel groter. Dan kan een ASCII-bestand met bijvoorbeeld een TAB-teken tussen elk veld een flinke ruimtebesparing opleveren. De structuur van deze bestanden (metagegevens ofwel het gegevensmodel)

moet bij elk archief mee worden opgeslagen, ook in ASCII- of XML-formaat.

Bedenk ook dat een historisch onderzoeker waarschijnlijk weinig heeft aan een export van een zeer groot aantal tabellen met een zeer ingewikkelde gegevensstructuur, zoals wordt aangetroffen bij diverse ERP-pakketten. Een vereenvoudiging naar een eenvoudige structuur is noodzakelijk, bijvoorbeeld sterschema[4] maar in elk geval een functionele gegevensstructuur waarin zoveel mogelijk details zijn weggelaten, desnoods enigszins gedenormaliseerd. Vaak is later niet elk detail nodig dat tijdens de werkelijke productie nodig is.

Moeten de gegevens zo gearchiveerd worden dat ze mogelijk later teruggeladen kunnen worden voor bewerkingen in de database, dan is een export van de tabellen naar een (DBMS-afhankelijk) export-, ASCII- of XML-formaat het beste. Denk bij het weer laden aan de integriteit van de gegevens.

Een databasedump zal niet vaak een bruikbaar archief opleveren, maar kan wel gebruikt worden door bij een maandelijkse schoning eerst een dump te maken en op tape te archiveren.
Als later toch nog geschoonde gegevens bekeken moeten worden kan in een kopie van de productieomgeving deze dump worden teruggeladen. Voor het terugladen en bekijken van de gegevens is DBMS-software van de goede versie en waarschijnlijk ook de goede versie van de applicatie nodig en als het zeer oude dumps betreft mogelijk ook nog het bijbehorende operating system en de hardware.

Bij het exporteren van de inhoud van tabellen naar ASCII- of XML-bestanden moet ook rekening gehouden worden met toekomstige wijzigingen op het gegevensmodel. Het toevoegen van verplichte kolommen, het verwijderen van kolommen, het wijzigen van validatieregels die door de database worden afgedwongen, dit kan allemaal het laden van oude gegevens in gevaar brengen.
In het meest eenvoudige geval (een in het systeem verwijderde kolom mag niet meer geladen worden vanuit het archief) is hier met wat programmeren nog wel wat aan te doen. Als de gehele gegevensstructuur op de schop gaat moet mogelijk, naast de conversie van live gegevens naar het nieuwe systeem, ook een conversie op de archiefgegevens naar het nieuwe archief worden overwogen.

Zijn er na 100 jaar nog CD-ROM lezers die met de huidige CD's kunnen omgaan?

Zoals we zien zijn er veel mogelijkheden om gegevens te archiveren. Alleen duidelijkheid over het toekomstig gebruik kan de ontwerper de gewenste informatie geven voor het maken van de goede keuzes.

Hoe lang gaat een CD-ROM of tape mee?

De volgende informatie is gevonden op de website www.cdrfaq.org, vol feiten over CD-ROM en CD-RW.

Fabrikanten claimen een levensduur voor CD-ROM's van 75 tot 200 jaar, afhankelijk van het type. Ook de omstandigheden waaronder de CD's bewaard worden heeft invloed op de levensduur. Zie verder de hiervoor genoemde website en de links die daar gevonden kunnen worden.
Voor belangrijke archieven op CD-ROM is het beter om meer kopieën te maken:

- De eerste voor gewoon gebruik: raadplegen of terugladen van de gegevens.
- De tweede als backup voor de eerste en bewaard in een kluis bij de backup tapes van operationele gegevens. In geval van een probleem met de eerste CD kan van deze CD een nieuwe kopie gemaakt worden.
- En eventueel een derde als backup voor de backup-CD; deze wordt bewaard bij de off-site backup-tapes (buiten het gebouw waar de kluis met de tweede CD staat staat).

Voor magnetische media (tape, diskette) is de levensduur veel korter. Zie hiervoor de website www.dpts.co.uk/datarec.htm, waarop wordt aangeraden om een tape elk jaar over te schrijven.

WAARMEE ARCHIVEREN

Voor databasedumps en exports en imports van tabellen kunnen tools van de database gebruikt worden, eventueel in combinatie met SQL-code die de gegevensstructuur eerst vereenvoudigt tot een sterschema.[4] Denk hierbij ook aan programmatuur om de gegevens weer terug te plaatsen in de database en daarin te bekijken.

Een alternatief hulpmiddel zijn de ETL (Extract, Transfer, Load) tools uit de datawarehouse-omgeving. Voor bijzondere situaties zal eigen programmatuur geschreven moeten worden, maar er komen nu ook producten op de markt die zich specifiek op deze markt richten, zoals "Active Archiving" van Princeton Softech (zie ook het kader).
Enkele ERP-pakketten (SAP, PeopleSoft) hebben eigen tools om gegevens te archiveren.

WAAROP ARCHIVEREN

Voor het medium waarop het archief wordt opgeslagen hebben we keuze uit:

- Papier of microfiche. Voordeel: lang leesbaar (100 jaar of meer

indien goed bewaard); papier kan zonder apparatuur worden gelezen en microfiche met eenvoudige optische apparatuur; Nadeel: verouderd systeem, niet gemakkelijk zoeken, kost veel ruimte.
- Tape: Voordeel: kan vele gigabytes aan archief bevatten; Nadeel: verweert na enige jaren en moet opnieuw worden gekopieerd; meestal zullen de gegevens weer eerst van tape naar schijf gekopieerd moeten worden voordat ermee gewerkt kan worden.
- CD-ROM: Voordeel: sneller toegankelijk dan tape; er kan, als het bijvoorbeeld PDF-bestanden betreft, direct van gelezen worden, terwijl bestanden op tape eerst moeten worden gekopieerd naar een vaste schijf van de pc, fileserver of UNIX-server; zou 100 jaar leesbaar moeten blijven (zie het kader Hoe lang gaat een CD-ROM of tape mee?); Opslag op CD-ROM is mogelijk tot 700 megabyte; op DVD enkele gigabytes, dus er is minder opslag mogelijk dan op tape, maar dit is vaak geen probleem en nieuwe ontwikkelingen kunnen dit probleem mogelijk oplossen; Nadeel: zijn er na 100 jaar nog CD-ROM-lezers die met de huidige CD's en DVD's kunnen omgaan? Anders is tussentijds conversie naar een ander medium noodzakelijk.

Ook voor het medium moet weer gekeken worden naar het doel van het archief. Teruglezen van papier of microfiche is niet mogelijk of anders zeer bewerkelijk. Tape, en vooral een tape-robot waarin de tapes automatisch worden opgehaald, is een goede optie als geautomatiseerd archiefbestanden teruggelezen moeten kunnen worden naar een database. CD-ROM's zijn weer handiger als een gebruiker er zelf een factuur in ASCII-, XML-, PDF- of Word-formaat in wil opzoeken en bekijken.

Fabrikanten claimen een levensduur voor CD-ROM's van 75 tot 200 jaar

HET BEKIJKEN OF RESTOREN VAN GEARCHIVEERDE GEGEVENS

Bij het ontwerpen van de archivering moet ook bekeken worden hoe de gearchiveerde gegevens weer geraadpleegd kunnen worden:
- Als gegevens die bij elkaar horen, bijvoorbeeld een factuur met detailregels en betalingen, ook bij elkaar staan in rapport vorm in ASCII-, XML-, PDF- of Word-formaat, dan kunnen deze gegevens gemakkelijk bekeken en opnieuw afgedrukt worden; ook zoeken is hierin eenvoudig met de bestaande hulpmiddelen.
- Als het archief ook een relationele database is kunnen de gegevens met SQL-query's (stored procedures), mogelijk in combinatie met de gebruikerstoepassing of een report writer, worden bekeken.
- Als het archief wordt gevormd door exports van tabellen, dan moet eerst het archief in de originele of eventueel in een kopie database worden teruggeplaatst. Hierbij kan men hele archieven (alle orders of facturen van een archiefrun) terugplaatsen of (met wat meer programmeerwerk) alleen de gewenste facturen.
- Als het archief een databasedump is zal deze dump (geheel) teruggeladen moeten worden, waarna met de gebruikerstoepassing of met SQL de gegevens kunnen worden bekeken.

MANIPULATIE VAN GEARCHIVEERDE GEGEVENS

De gegevens in een PDF-bestand kunnen niet (al of niet met opzet) gewijzigd worden. Hetzelfde geldt voor gegevens op een CD-ROM, maar een Word-document dat op een CD-ROM staat kan wel op de pc worden gewijzigd voordat de gevraagde pagina's worden afgedrukt. Als manipulatie door de gebruiker, of het per ongeluk wijzigen en weer opslaan van documenten, voorkomen moet worden, dan moet daar in de keuze van medium en hulpmiddelen rekening worden gehouden.

CONCLUSIE

Het opnemen van schoning en/of archivering in de architectuur en het ontwerp van een systeem is noodzakelijk om:
- later ook een bruikbaar gegevensarchief te hebben;
- te voorkomen dat er technische problemen ontstaan in de productiesystemen door te grote (= onnodig grote) databases, bijvoorbeeld vollopen van een database of lange doorlooptijd van databasedumps en gebruikersfuncties;
- te voorkomen dat de kosten van opslag (op dure UNIX-schijfsystemen in plaats van op tape of CD-ROM) uit de hand lopen.

Daarbij komen de bekende ontwerpvragen: "Waarom, Wat, Hoe, Waarmee en Waarop?" weer terug. Hierbij moet ook bekeken worden wat er nodig is om later de gegevens in het archief te kunnen raadplegen.

LITERATUUR

1. Loonen, Ontwikkelstraat, hergebruik door inzet van architectuur. Software Release 2000/7-8.
2. Loonen, Performance. Database Magazine 2002/1-3.
3. Loonen, Mutatierapportage, de tijdgeest van de database. Database Magazine 1999/3.
4. Van der Lek, Wanneer een ster de job doet. Database Magazine 1998/2.

Toon Loonen (toon.loonen@cgey.nl, toon.loonen@inter.nl.net) is als consultant werkzaam bij Cap Gemini Ernst & Young.