RDW bouwt Data Vault

Eerste gecertificeerde datawarehouse bij overheid

Hans Lamboo

De RDW is de eerste overheidsinstelling die hun Data Vault datawarehouse gecertificeerd heeft gekregen. Auditer Ronald Damhof legde zijn rapport voor aan de Genesee Academy in de VS, de organisatie van Dan Linstedt, die de certificatie toekende. Het certificeren van datawarehouses staat in de kinderschoenen, maar past in de professionalisering van het vakgebied.

Belangrijk is dat de RDW een zogenaamd Data Vault datawarehouse heeft opgezet, waarbij standaardisatie en industrialisatie voorop staan. Daarmee wordt bereikt dat er een bijzonder duurzaam, flexibel, maar vooral ook traceerbaar en auditeerbaar datawarehouse is gebouwd. De RDW gebruikt dit datawarehouse voor zowel interne als externe klanten. Het BI-team, met daarin onder anderen BI-engineer Bertus Hölscher, Data Architect Bob van der Mark en teammanager Walter Klerks, nam samen met de opdrachtgever en andere betrokkenen op 17 februari 2010 het officiële Data Vault Certificaat in ontvangst.

Drie Schema Architectuur

De RDW begon in 2008 aan de bouw van een nieuw datawarehouse. De reden daarvoor lag in een status- en toepassingsverandering: de database van de RDW kreeg van de overheid officieel de status 'Basisregister', wat impliceert dat kentekeninformatie alleen bij de RDW geregistreerd staat. Daarmee wordt de externe vraag naar deze informatie groter. Het oude datawarehouse, bestemd voor managementinformatie, was niet op deze nieuwe taak berekend, en een BI-projectgroep toog aan het werk om te komen tot een nieuw datawarehouse.

Data Architect Bob van der Mark van de RDW vertelt: "We begonnen met de constructie van een nieuw op traditionele leest geschoeid datawarehouse. Daar kwamen we echter al snel mee in de problemen. Het datawarehouse van de RDW heeft namelijk een nogal complexe relatiestructuur, hetgeen vooral te zien was aan een historie-explosie. Bij één enkele mutatie moeten immers alle afhankelijke records gedupliceerd worden. Dat liep echt de spuigaten uit. Mijn collega Bertus Hölscher, BI-engineer, was het daar volkomen mee eens. Dit ging niet werken, we moesten met een andere oplossing komen."

De RDW werkt op basis van wat zij een Drie Schema Architectuur noemen. Het Conceptuele Schema, het logisch bedrijfsdatamodel, is de bron van alle ontwerpen. Het Interne Schema bevat de implementatie van entiteiten en attributen in kolommen en tabellen van de databases. Het derde is het Externe Schema, een vastlegging van de berichten die worden uitgewisseld.

"Het gaat niet zozeer om de hoeveelheid data, die is niet zo groot, maar het is ingewikkeld door de complexe relaties. Ons bedrijfsdatamodel omvat honderden entiteiten," voegt Van der Mark daaraan toe. Het bijhouden van de historie is een vrij algemeen probleem: bij een model dat bestaat uit een groot aantal onderling afhankelijke tabellen moet een mutatie vele malen gedupliceerd worden in de onderliggende tabellen. Op zich een bekend probleem, maar naarmate er meer complexe relaties in het spel zijn, wordt het een steeds groter probleem.

Bertus Hölscher vond op internet enkele artikelen van de hand van Dan Linstedt. Zijn Data Vault model leek de oplossing te bieden voor de problemen van de RDW.

"Ik wist dat eigenlijk meteen vrij zeker. Het Data Vault model sprak me direct heel erg aan," vertelt hij. "Als je een wijziging doet in de inhoud van je OLTP-database, dan leidt dat met de Data Vault immers maar tot één wijziging in het datawarehouse."

De beslissing om halverwege het project over te stappen op een Data Vault was een belangrijke beslissing. Er zat immers al een aantal maanden werk in het project, en de keuze voor de Data Vault impliceerde dat het datawarehouse vanaf de bodem moest worden opgebouwd. "Er moest een belangrijk besluit genomen worden. Gelukkig ging dat heel soepel," herinnert Van der Mark zich. "In niet meer dan een kwartier hadden we de projectleiding overtuigd dat het een goed besluit was. Het was evident dat de Data Vault ons probleem kon oplossen. Dat maakte het feit dat het extra tijd ging kosten dubbel en dwars goed, daar waren we allemaal van overtuigd. In een kwartiertje besloten we tijdens

een staande vergadering van ons datawarehouseproject om deze nieuwe weg in te slaan."

Van links naar rechts: Bob van der Mark, Jacky van Hogen, Bertus Hölscher, Andre van Luijn, Ronald Damhof, Gert Jan Holland en Petra Smid.

Voordelen

Naast het feit dat het historieprobleem werd opgelost, waarborgt de Data Vault een gestandaardiseerde manier van werken en blijft het ook op termijn overzichtelijk en gemakkelijk te onderhouden. "De voordelen van die gestructureerde manier van werken zag je terug bij onze junior engineer," vertelt Hölscher. "Die was al op de tweede dag aan het bouwen."
Drie leden van het BI-team volgden de meerdaagse Data Vault training, onder meer gegeven door Dan Linstedt zelf. Eén dagdeel werd lesgegeven door Ronald Damhof van Prudenza, waarna de mannen van RDW met hem in gesprek raakten. "We hadden wel wat aanloopproblemen die we met hem hebben doorgesproken," zegt Hölscher. "We hebben hem uitgenodigd om bij de RDW een keer een presentatie te komen geven. Naar aanleiding van Damhofs bezoek hebben we wat redesign toegepast; we hadden wat aannames gedaan die achteraf niet zo gelukkig waren." Als ander bijkomend voordeel van de Data Vault noemt hij dat het mogelijk is om zelf het genereren van de mappings te ontwikkelen.
Het nieuwe datawarehouse werd dus in twee stappen gebouwd. Het projectteam stond onder grote druk om iets op te leveren dat werkte; het besluit om het halverwege het project principieel anders aan te gaan pakken kostte veel tijd. Als eerste werd het al gebouwde datawarehouse op een geautomatiseerde manier omgezet naar een Data Vault. Van der Mark: "We splitsten alle tabellen in een deel met een sleutel en een deel met afhankelijke kolommen. De eerste werd een 'hub', de tweede een 'satellite', alle foreign keys hebben we vertaald naar 'links' – alle drie termen uit Dan Linstedts model. "Zo ontstond een Data Vault die voor bijna 100 procent gegenereerd was. Bijna 100 procent, want niet alles laat zich gemakkelijk zo manipuleren als het meest gewenst is. Groot voordeel van deze aanpak voor het projectteam was dat er al na enkele weken een werkende Data Vault was.

Audit

Inmiddels is het gedeelte Basisregistratie Kentekens operationeel. Het totale datawarehouse bevat ongeveer 100 Gigabyte aan data in 56 hub-tabellen met bijbehorende satellieten en 25 referentietabellen, het volledige bedrijfsdatamodel is aanzienlijk groter. "Een van de eerste externe vragen betrof de bedrijfsvoorraad van garagebedrijven. Zo'n dealer of garage heeft een aantal voertuigen op naam staan, tot ze verkocht worden. Ze vragen dus bij ons op hoeveel kentekens er volgens de RDW op hun naam staan en vergelijken dat met de situatie in hun bedrijf. Aan die vraag konden we heel snel voldoen," aldus Hölscher. "We zijn al bezig om de APK-keuringen in de Data Vault te zetten en daarna een groot gedeelte van de rijbewijzenadministratie. Alles wat in de Centrale Database komt, komt uiteindelijk ook in het datawarehouse."
Er wordt geprobeerd zoveel mogelijk data gegenereerd in de Data Vault te krijgen, om zo snel te kunnen voldoen aan toekomstige BI-vragen, zonder eerst nieuwe bronsystemen te moeten ontsluiten. Van der Mark: "We zijn nu bezig dat helemaal geautomatiseerd te maken op basis van metadata, met Informatica's Power Center. De metadata komen in onze Data Dictionary. Het idee is om daarvan XML te genereren en dat in te lezen in Informatica."
Vanwaar de vraag om certificering? Teamleider Walter Klerks verklaart: "Bij de RDW loopt al een aantal interne kwaliteitsprojecten. In het verlengde daarvan wilden we dat de managers die bij oplevering het datawarehouse moesten accepteren graag een gecertificeerd model aanbieden. En ik kan niet anders zeggen dan dat de managers ontzettend blij zijn met de kwaliteit die we ze geleverd hebben."

In december 2009 deed Ronald Damhof een officiële audit en legde zijn rapport voor aan de Genesee Academy, die de aanvraag honoreerde. Op 17 februari 2010 ontving het ICT-bedrijf van de RDW het officiële certificaat uit handen van Ronald Damhof. Daarmee is de RDW de eerste Nederlandse overheidsinstelling met een gecertificeerd datawarehouse.

Hans Lamboo is hoofdredacteur van Database Magazine.
Met dank aan Ronald Damhof van Prudenza.