

QUIPU vanaf 1 juli beschikbaar

Open Source datawarehousing van eigen bodem

Pieter Rambags

Binnenkort lanceert QOSQO, het datawarehouse services zusterbedrijf van BI consultancy bedrijf Nippur, een Open Source datawarehouse management systeem onder de naam QUIPU. Aan QUIPU is de afgelopen jaren gewerkt door een team van ontwikkelaars met als doel data-integratie mogelijk te maken voor iedere organisatie die daaraan behoefte heeft.

In dit artikel wordt beschreven wat de beslissing om QUIPU te ontwikkelen was en waarom QUIPU als Open Source ter beschikking wordt gesteld. Daarnaast komen de architectuur en de toekomstplannen aan bod.

Anno 2010 is het aantal hulpmiddelen dat ondersteunt bij de definitie van een integratieplatform en bijbehorende laadprocessen nog beperkt. Er zijn enkele datawarehouse management systemen zoals Kalido, SAP BW en BIReady; allemaal hulpmiddelen gebaseerd op een eigen specifieke, gesloten architectuur. Hulpmiddelen die een forse investering vereisen maar die daarvoor wél een aantal voordelen opleveren waar het de kwaliteit, onderhoudbaarheid en de snelheid van implementatie van een datawarehouse betreft. De acceptatie door het bedrijfsleven van deze producten valt echter tegen. SAP BW lijkt voorbehouden aan en wordt het meest aangetroffen in grote ondernemingen die flink hebben geïnvesteerd in SAP R/3. Kalido komt al veel minder voor en lijkt alleen door de echt hele grote ondernemingen te worden gebruikt. Dat heeft ongetwijfeld met de prijs ervan te maken. Daarnaast heeft Kalido de afgelopen jaren de aandacht verschoven naar een specifieke toepassing van de technologie als Master Data Management (MDM) systeem. BIReady wordt nu, ongeveer vijf jaar na introductie, nog steeds nauwelijks toegepast. Vermoedelijk speelt naast prijs de omvang van het bedrijf en het uitblijvende succes hen parten bij de acquisitie van nieuwe klanten.

In onze praktijk, werkend voor grote ondernemingen waaronder multinationals, zijn we gewend te werken met dergelijke hulpmiddelen en concepten. We kennen de voordelen voor het bedrijfsleven van software die ondersteunt bij de realisatie en het beheer van een datawarehouse. Hierdoor zijn wij ervan overtuigd geraakt dat automatisering van data-integratie beschikbaar zou moeten zijn voor iedereen! Ook voor de minder grote organisaties voor wie de investering in software-oplossingen niet van-

zelfsprekend is. Een datawarehouse draagt immers slechts indirect bij aan Business Intelligence oplossingen: het wordt door eindgebruikers niet gezien en niet begrepen. Dan wordt het moeilijk om (initiële) kosten te verdedigen. De genoemde producten zijn ook nog eens zo kostbaar dat de licentiekosten zelfs de business case ongedaan kunnen maken.

Data Vault

QUIPU is gebaseerd op het Data Vault concept als omschreven door Dan Linstedt. Wij hebben er voor gekozen ons te houden aan de criteria zoals ze opgesteld zijn door Dan Linstedt, zodat QUIPU volledig in overeenstemming is met de door hem gestelde eisen. Kijk op www.danlinstedt.com voor meer informatie over de Data Vault. De belangrijkste voordelen van de Data Vault worden hieronder nog eens opgesomd:

- verzamelen van historie van willekeurig welke data (of dat nu transactie- of masterdata betreft);
- optimalisatie voor (near) real-time updates (dat parallelle laadprocessen mogelijk maakt en zogenaamde 'late arriving transactions' toestaat);
- isolatie van veranderingen in het model en dus verworven flexibiliteit waar het aanbrengen van aanpassingen in het model betreft. Dit voorkomt kapitaalvernietiging waar het eerder ontwikkelde logica betreft.

Dankzij de komst van het Data Vault modelleringconcept is een interessante situatie ontstaan. Dit concept is zo generiek van opzet dat het zich leent voor automatisering. Consultants van Nippur hebben al vele klanten geholpen met de ontwikkeling van Data Vault georiënteerde datawarehouses. Daarbij zijn hulpmiddelen (scriptgenerators) ontwikkeld om de repetitieve (en dus saaie) handelingen te automatiseren. Hierdoor zijn enorme versnellingen in de ontwikkeling van een datawarehouse bewerkstelligd; in een enkel geval tot wel 60 procent sneller dan het traditioneel handmatig bouwen van een BI-architectuur. Met hetzelfde budget kunnen klanten dan meer vraagstukken onderzoeken en implementeren.

QOSQO

QOSQO is in 2008 gestart, volledig gericht op het bieden van services rondom de Data Vault. QOSQO beheert onder andere een in-house ontwikkeld datawarehouse management systeem voor een grote bank. De belangrijkste activiteit van QOSQO is

het ontwikkelen van een datawarehouse management systeem dat QOSQO in Open Source ter beschikking zal stellen aan iedereen met behoefte aan data-integratie. Dat project en tevens product noemen we QUIPU, naar analogie met de quipu's die archeologen vonden in en rond Cuzco (ook wel geschreven als QOSQO) in Peru.

Quipu's (of Khipu's) waren bij de Inca's een soort koorden die gebruikt werden om getallen weer te geven door middel van knopen. Er werden meerdere koorden aan een koord geregen. Het tellen van getallen gebeurde door het leggen van knopen, waarbij het decimale stelsel werd toegepast en de soort van te tellen eenheden werd door kleur aangegeven.

QUIPU 1.0

QUIPU 1.0 zal 1 juli 2010 worden vrijgegeven. QUIPU heeft tot doel het invullen van de kerncomponenten van een datawarehouse architectuur. Dit zijn het genereren, onderhouden en vullen van (database) structuren die voorbereid zijn op het bijhouden van historie. Zowel van transactie- als van referentiedata. QUIPU 1.0 is in staat bronmodellen in te lezen door middel van 'reverse engineering'. Dit kan een bronapplicatie zijn maar ook een business model dat is opgeslagen als ERD-schema. Zodra het model in de repository aanwezig is zijn de volgende functies beschikbaar:

- Afleiding van een Data Vault model. Een vertaling wordt uitgevoerd door analyse van het bronmodel, de sleutels en de relaties. Voor dat doel hebben we algoritmes opgenomen in QUIPU die bij voortschrijdend inzicht zullen worden aangepast c.q. aangescherpt. De afleidingsregels worden in de gebruikersinterface in natuurlijke taal getoond, waardoor de architect kan controleren of de getrokken conclusies juist zijn;
- In de gebruikersinterface kan de architect vervolgens, onder andere door middel van 'drag and drop' functionaliteit, aanpassingen maken in de beoogde 'mappings' tussen source en target Data Vault model. Daarbij kan hij ook aanpassingen maken in de gegenereerde Data Vault.
- Generatie van laadfuncties. Doordat het bronmodel en het target Data Vault bekend zijn in de repository kunnen de laadfuncties worden gegenereerd (in standaard ANSI SQL) en ook worden opgeslagen in de repository. Deze laadfuncties kunnen net als het Data Vault model eenvoudig worden overgebracht naar het RDBMS om daar te worden uitgevoerd;
- Executie en logging van laadfuncties. QUIPU zal – aanvankelijk nog eenvoudige – functies bieden om de gegenereerde laadcode uit te voeren op het datawarehouse platform en de resultaten te loggen in de repository;
- Als eerste stap in de richting van datamartgeneratie zal QUIPU views genereren die de Data Vault structuur weer vertalen naar herkenbare structuren zoals onderkend in de bron.

Open Source

Een belangrijke reden om voor Open Source te gaan is gelegen in het feit dat wij vinden dat data-integratie voor iedereen be-

schikbaar moet zijn: bedrijfsleven, overheid, gezondheidszorg, onderwijsinstellingen. We willen de blokkade die er ligt bij het kunnen toepassen van innovaties op dit gebied doorbreken. Een andere reden is gelegen in het feit dat op deze wijze de toepassing van QUIPU gemakkelijker kan worden gestimuleerd. Ten tijde van het verschijnen van dit artikel zijn er al twee grote ondernemingen die (een prerelease versie van) QUIPU inmiddels hebben toegepast. Andere hebben de afgelopen maanden aangegeven te willen wachten met data-integratie tot de komst van QUIPU, overtuigd als ze nu al zijn van de mogelijkheden. Als licentiemodel voor de community editie van QUIPU is gekozen voor de GPL v3 licentie.

Community editie versus enterprise editie

Op dit moment wordt een community editie (CE) van QUIPU gelanceerd. Hiermee kunnen bedrijven een volwaardige datawarehouse-omgeving opzetten. Aan deze editie zijn en zullen geen licentiekosten worden verbonden zoals boven omschreven en deze kent geen beperkingen ten aanzien van omvang of complexiteit van het datamodel, het aantal datamodellen en het datavolume dat het gegenereerde datawarehouse kan verwerken. De broncode is beschikbaar volgens het genoemde GPL model, waarmee het iedereen vrij staat de code naar eigen behoefte aan te passen. Het is en blijft onze bedoeling data-integratie mogelijk te maken voor iedereen.

De broncode is beschikbaar volgens het GPL model

In de nabije toekomst zal ook een enterprise editie (EE) verschijnen, een volledig losstaand en separaat product. Deze zal zich vooral richten op:

- het managen van wijzigingen in verschillende datawarehousemodellen door toepassing van versiebeheer over de modellen in de repository;
- het genereren van deltascripts die alléén de wijzigingen doorvoeren in het datawarehouse;
- het ondersteunen van projecten waarin meerdere ontwikkelaars tegelijk op eenzelfde model werken;
- het ondersteunen van toepassing van QUIPU in een ontwikkel-, test-, acceptatie- en productiestraat.

Voor de enterprise editie zal een vergoeding worden gevraagd. De community editie zal zich als client gaan gedragen voor de enterprise editie van QUIPU.

QUIPU is vanaf 1 juli ter beschikking op de website www.datawarehousemanagement.org

Drs. Pieter Rambags is managing partner van Nippur.