

QUIPU: Open Source data warehousing

Binnenkort lanceert QOSQO, het data warehouse services zusterbedrijf van het gerenommeerde Business Intelligence consultancy bedrijf Nippur, een open source data warehouse 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 daar behoefte aan heeft. Dit artikel beschrijft hoe wij tot de beslissing zijn gekomen QUIPU te ontwikkelen en waarom wij QUIPU in open source ter beschikking stellen. Daarnaast besteed ik aandacht aan de architectuur en wat onze plannen voor de toekomst zijn.

Achtergrond

Anno 2010 is het aantal hulpmiddelen dat ondersteunt bij de definitie van een integratie platform en bijbehorende laadprocessen nog beperkt. Er zijn enkele data warehouse 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 data warehouse 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 5 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, werkende voor grote ondernemingen waaronder de multinationals die ons land rijk is, zijn we gewend te werken met dergelijke hulpmiddelen en concepten. We kennen als geen ander de voordelen voor het bedrijfsleven van software die ondersteunt bij de realisatie en het beheer van een data warehouse. 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 vanzelfsprekend is. Een data warehouse draagt immers slechts indirect bij aan de 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 als opgesteld door Dan Linstedt zodat QUIPU volledig in

overeenstemming is met de eisen die hij stelt. Zie <http://www.danlinstedt.com/About/> 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 is);
- Optimalisatie voor (near) realtime 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.

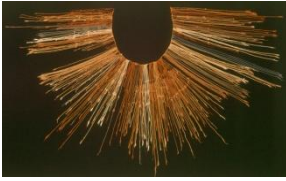
Dankzij de komst van het Data Vault modelleringconcept is een interessante nieuwe situatie ontstaan enkele jaren geleden. Dit concept is zo generiek van opzet dat het zich leent voor automatisering. Dat weten de consultants van Nippur als geen ander omdat zij al vele klanten hebben geholpen met de ontwikkeling van Data Vault georiënteerd data warehouses. Telkens ontwikkelden zij hulpmiddelen (scriptgenerators) om de repetitieve (en dus saaie) handelingen te automatiseren.

Zij zagen de voordelen hiervan omdat ze enorme versnellingen in de ontwikkeling van een data warehouse konden bewerkstelligen. In een enkel geval tot wel 60%(!) sneller dan het traditioneel handmatig bouwen van een BI architectuur. Met hetzelfde budget kunnen klanten dan meer vraagstukken onderzoeken en implementeren. En als het aan de consultants van Nippur ligt is dát het echte leuke werk. Het begrijpen van de klant, het meedenken hoe een business vraagstuk op te pakken zodat de klant echte goede business cases kan definiëren en oppakken.

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 data warehouse management systeem dat QOSQO in open source ter beschikking zal stellen aan iedereen met een data integratie behoefte. Dat project en tevens product noemen we QUIPU. Naar analogie met de quipus die archeologen vonden in en rond Cuzco (ook wel geschreven als QOSQO) in Peru.

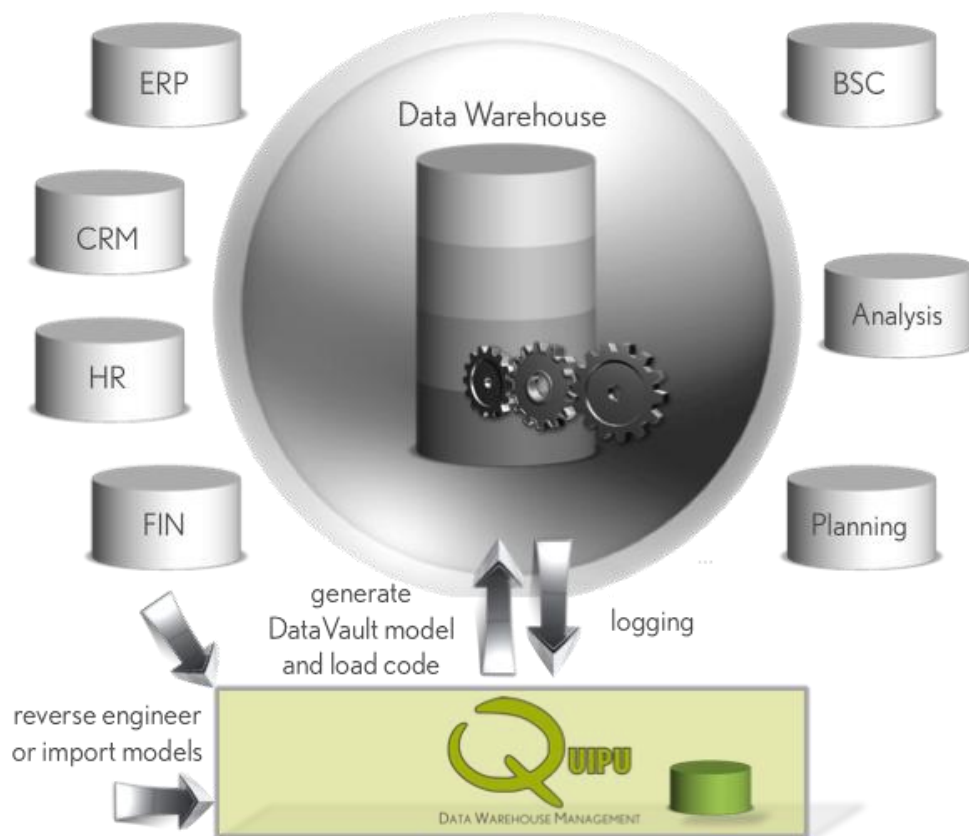
Quipus (of Khipus) waren bij de Inca's een soort koorden die gebruikt werden om getallen weer te geven. Dit deden ze door middel van knopen. Er werden meerdere koorden aan een koord geregen. Het tellen van getallen gebeurde door het leggen van knopen, waarbij het decimale stelsel werd toegepast. De onderste knopen gaven de eenheden aan, daarboven de tientallen. De soort van het te tellen eenheden werd door de kleur aangegeven. [bron: wikipedia]



Figuur 1 - Voorbeeld van een quipu

QUIPU

QUIPU 1.0 zal 1 juli 2010 worden vrijgegeven. QUIPU heeft tot doel het invullen van de kern componenten van een data warehouse 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.



Figuur 2 - QUIPU architectuur

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 natuurlijke taal getoond in de gebruikers interface waardoor de architect kan controleren of de getrokken conclusies juist zijn;
- In de gebruikers interface 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 laad functies. Doordat het bron model 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 laad functies. QUIPU zal –aanvankelijk nog eenvoudige- functies bieden om de gegenereerde laadcode uit te voeren op het data warehouse platform en de resultaten te loggen in de repository;
- Als eerste stap in de richting van Data Mart generatie zal QUIPU views generen 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 beschikbaar moet zijn voor iedereen. Bedrijfsleven, overheid, gezondheidszorg, onderwijsinstellingen etcetera. We willen de blokkade die er nu ligt op 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 makkelijker kan worden gestimuleerd. Ten tijde van het verschijnen van dit artikel zijn er al twee grotere ondernemingen die (een pre-release versie van) QUIPU inmiddels hebben toegepast. Anderen hebben in de loop van de afgelopen maanden aangegeven te willen wachten met data integratie tot de komst van QUIPU! Overtuigd als ze nu al zijn van de mogelijkheden.

We hebben als licentie model voor de community editie van QUIPU 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 data warehouse omgeving opzetten. Aan deze editie zijn en zullen geen licentiekosten worden verbonden zoals boven omschreven en kent geen beperkingen ten aanzien van :

- omvang of complexiteit van het datamodel;
- het aantal datamodellen;
- het datavolume dat het gegenereerde data warehouse 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.

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 data warehouse modellen door toepassing van versiebeheer over de modellen in de repository;
- het genereren van delta scripts die alléén de wijzigingen doorvoert in het data warehouse;
- het ondersteunen van projecten waarin meerdere ontwikkelaars tegelijk op eenzelfde model werken;
- het ondersteunen van toepassing van QUIPU in een ontwikkel, test, acceptatie en productie straat .

Voor deze versie zal een vergoeding worden gevraagd.

De community editie zal zich als client gaan gedragen voor de enterprise editie van QUIPU.



Figuur 3 - QUIPU Enterprise Edition

QUIPU zal vanaf 1 juli ter beschikking worden gesteld op de website

http://www.data_warehousemanagement.org

De auteur

Drs. Pieter Rambags is managing partner van Nippur. Hij heeft een achtergrond in business intelligence en data warehousing sinds 1992 en studeerde Bestuurlijke Informatie Kunde aan de Universiteit van Tilburg. Samen met Peter Kurstjens heeft hij Nippur en QOSQO opgericht in respectievelijk 2002 en 2008.