Een vergelijkende studie tussen Data Vault 2.0 en het dimensioneel model bij het modelleren van een data warehouse.

Spiesschaert Jorik, Lievens Stijn, Stroobants Jochen
Hogeschool Gent, Valentin Vaerwyckweg 1, 9000 Gent
jorik.spiesschaert.y9481@hogent.be

Abstract

Bij het modelleren van een data warehouse moet rekening gehouden worden met de noden van een organisatie. Afhankelijk van die noden, kan er beslist worden om een bepaald data model te kiezen.

Introductie

DHL Pharma Logistics wil een data warehouse aanmaken om hun rapporteringen te automatiseren. De data warehouse wordt gemodelleerd via de Data Vault 2.0 methodologie. In deze paper wordt onderzocht of modelleren via deze manier wel degelijk de juiste keuze was ten opzichte van het dimensioneel model.

Deze paper kan nuttig zijn voor informatici die actief zijn of interesse hebben in Business Intelligence. Er is geen specifieke voorkennis rond Data Vault of het dimensioneel modelleren nodig om door deze paper te gaan. Dit vergelijkend onderzoek kan informatici helpen bij het beslissen van een data model methodologie.

Experimenten

Voor er kan gestart worden met het onderzoek, moet er een data warehouse opgesteld worden zowel via de Data Vault methodologie als de methodologie voor het dimensioneel modelleren. In deze bachelorproef wordt de opbouw van van beide data warehouses uitgebreid uitgeschreven (ETL-processen van verschillende lagen). Bovendien wordt er ook een overzicht gegeven over de rapporteringsnood voor DHL Pharma Logistics, wat de betekenissen zijn van de gebruikte data in dit onderzoek en hoe een remote connectie moet opgezet worden voor SAP HANA.

Op basis van de opgebouwde data warehouses worden dan vergelijkingen gemaakt op basis van performantie en schaalbaarheid door het uitvoeren van enkele queries verwerkt in procedures.

Overzicht data modellen

Dimensioneel model

Wie consultant Business Intelligence is, heeft ongetwijfeld al gehoord van dimensioneel modelleren. Over de jaren heen is het als het ware een standaard geworden wanneer men een data warehouse wil ontwerpen. Dimensioneel modelleren heeft als voordeel dat het model niet complex is, dus gemakkelijk te begrijpen, zelf voor niet IT-opgeleide personen. Ook is er vaak een snel resultaat beschikbaar. In deze paper zal het dimensioneel modelleren dieper bekeken worden.

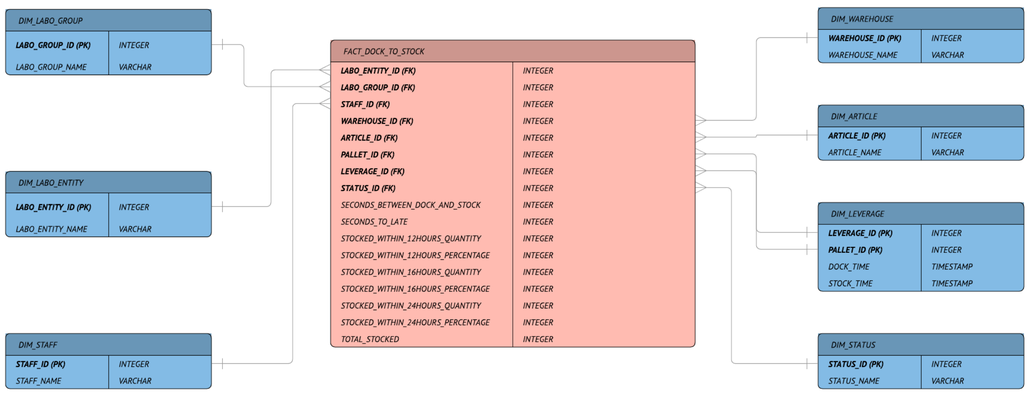

Figure 1: Voorstelling van een dimensioneel model (gemaakt via Lucidchart.com).

In figuur 1 worden de dimensions voorgesteld als de blauwe entiteiten. Deze bevatten de beschrijvende data die iets meer vertellen over de ”facts”. De business key wordt gebruikt als attribuut die de relatie legt naar de facts-table (die wordt voorgesteld in het rood). In dit model worden geen gegevens opgeslagen die meer vertellen over de oorsprong van de data, tevens wordt de historiek van de data niet bijgehouden.

Data Vault

Data Vault 2.0 is een modelleertechniek die ontworpen is door Daniel Linstedt. Het model zorgt ervoor dat dimensies gemakkelijk uitgebreid kunnen worden en dat databronnen toevoegen vlot moet gaan. Linstedt is van mening dat business requirements vaak veranderen, dus moet het model waarin de data warehouse ontworpen is ook flexibel zijn(D. Linstedt & Olschimke, 2016).

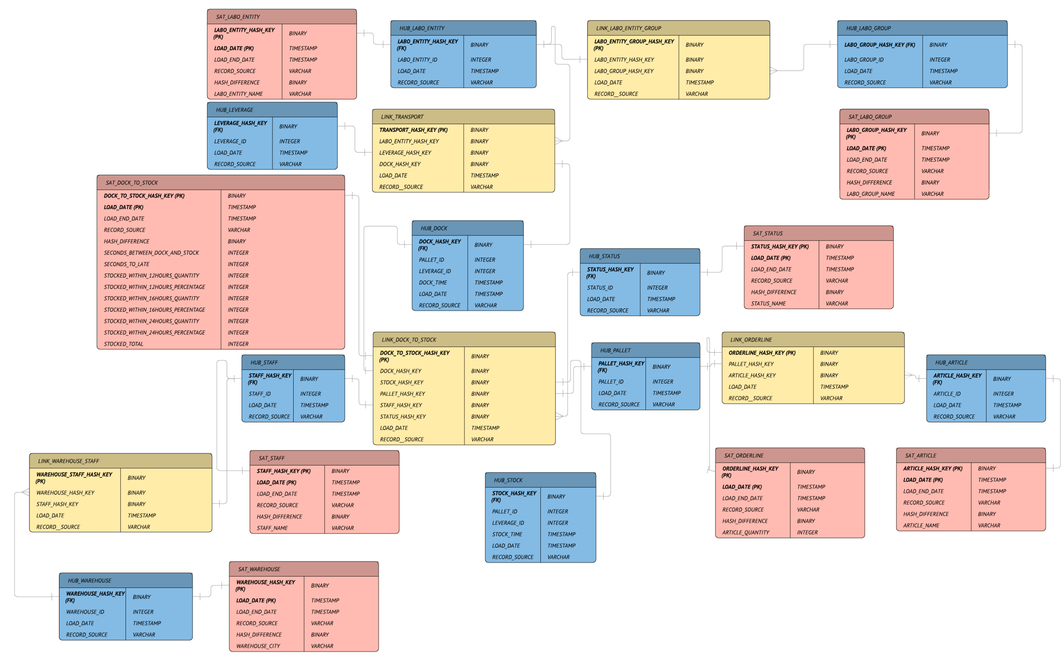

Figure 2: Voorstelling van een Data Vault 2.0 model (gemaakt via Lucidchart.com).

Het model in figuur 2 is opgebouwd uit drie soorten tabellen: de rode entiteiten worden voorgesteld als sattelites, de blauwe entiteiten als hubs en de gele entiteiten vormen de links tussen de verschillende hubs. In de SAT_DOCK_TO_STOCK worden de berekeningen opgeslagen die nodig zijn voor het berekenen van de KPI. Hash keys worden gebruikt als primary en foreign keys in dit model en worden opgeslagen onder het type ”binary”.

Vergelijkende onderzoek

In deze bachelorproef wordt Data Vault 2.0 vergeleken met het dimensioneel model. In dit vergelijkend onderzoek wordt voornamelijk gefocust op deze 5 pijlers:

- **Performantie:** is er een significant verschil in het uitvoeren van leesopdrachten?
- **Complexiteit:** zijn beide modellen makkelijk interpreteerbaar door IT & business?
- **Flexibiliteit:** hoe flexibel zijn beide modellen wanneer een business requirement gewijzigd wordt?
- **Schaalbaarheid:** hoe gaan beide modellen overweg met het inladen van enorme hoeveelheid data?
- **Audit:** is er metadata beschikbaar over de werkelijke data? Kunnen problemen makkelijk opgespoord worden?

Conclusies

Het onderzoek wijst uit dat het dimensioneel model een betere keuze was voor het modelleren van de data voor DHL Pharma Logistics. Dit is bovendien ook het resultaat dat ik verwacht had.

Het dimensioneel model voert sneller leesresultaten uit in vergelijking met het Data Vault, dit omdat het Data Vault model meer relaties heeft. Hierdoor zullen veel meer joins moeten gebeuren wanneer alle data moet opgehaald worden.

Bij de Data Vault methodologie wordt er meer informatie (Hash keys, informatie over extractie, ..) en tabellen opgeslagen in een databank. Dit zorgt ervoor dat de volumes van data enorm stijgen. Bijgevolg zal er dus een hogere kostprijs zijn om deze data te stockeren. Aangezien dit geen requirement is voor DHL Pharma Logistics, zou dit leiden naar een onnodige meerkost voor dit project.

De KPI’s waarvoor een model moet opgesteld worden zijn gestandaardiseerd, en dienen niet flexibel te zijn. Indien de berekening voor de KPI’s zouden gewijzigd worden, hoeven enkel sommige parameters uit het ETL-proces aangepast te worden en niet het datamodel zelf.

Toekomstig onderzoek

Het onderzoek wijst uit dat Data Vault 2.0 een voordeel heeft bij het inladen van data ten opzichte van het dimensioneel model door gebruik te maken van hash keys en doordat tabellen in een Data Vault 2.0 niet afhankelijk zijn van elkaar, kunnen deze allemaal tegelijk ingeladen worden. Naar de toekomst toe kan er onderzocht worden of Data Vault 2.0 een juiste keuze kan zijn bij het opstellen van big data modellen aangezien het toevoegen van enorme volumes data snel en efficiënt kan gebeuren.