Neem alle voor- en nadelen rationeel mee in het selectieproces

OLAP en Datawarehousing – een Siamese Tweeling?

Tom Breur

In de OLAP-markt is de laatste jaren een nieuwe trend te zien. Sommige leveranciers bevelen tegenwoordig een architectuur aan waarbij out-of-the-box oplossingen rechtstreeks op operationele systemen worden aangesloten. Groot voordeel daarvan is dat kostbare, tijdrovende en riskante datawarehouse-projecten niet meer nodig (zouden) zijn.

Dit klinkt natuurlijk aantrekkelijk, maar gaat het ook werken? De voordelen zijn evident, maar wat zijn de nadelen? Dergelijke out-of-the-box oplossingen laten zich niet vanzelfsprekend aansluiten op de reeds aanwezige metadata-oplossing. Maar behalve dit risico van alweer een Business Intelligence-eiland, kleven er nog enkele nadelen aan deze architectuur. Ik constateer dat deze architectuur – als regel – tekort schiet om de volgende redenen:

1. Operationele systemen hebben ofwel geen historie, ofwel historie die op 'de verkeerde manier' is gemodelleerd en opgeslagen voor de doeleinden van OLAP;
2. Virtuele datawarehouses zijn (doorgaans) vooralsnog geen goed alternatief voor OLAP doeleinden;
3. Datakwaliteit-issues vormen in de meeste organisaties een (onoverkomelijk) struikelblok om consistente informatievoorziening te kunnen waarborgen. Dergelijke issues kunnen in een 'managed environment' als een datawarehouse wél naar behoren worden geadresseerd;
4. Bedrijfsbrede en lange-termijn data governance wordt bemoeilijkt.

> Er dient een expliciete tijdsdimensie aanwezig te zijn in het datamodel

De conclusie is dan ook dat er nogal wat mitsen en maren kleven aan out-of-the-box oplossingen die OLAP mogelijk maken zonder beschikbaarheid van een DWH. Als alle voor- en nadelen rationeel en evenwichtig worden meegenomen in het selectieproces, kan men hier met open ogen in stappen. Maar het is een beetje naïef om te veronderstellen dat die objectiviteit van leveranciers zal komen.

Tactische versus strategische OLAP Query's

Er is een onderscheid te maken tussen tactische en strategische OLAP query's. Tactische query's dienen vooral ter ondersteuning van dagelijkse operationele taken. Een voorbeeld hiervan is bijvoorbeeld: 'welke klanten hebben op dit moment nog openstaande facturen?' Of: 'hoeveel klanten hebben deze maand een expirerende polis?' Op basis van de uitkomst kan de back-office manager zijn bezetting beter plannen als hij een rooster met werkschema's opstelt, bijvoorbeeld. Een strategische query zoekt naar trends of patronen, en hiervoor is wél historie nodig. Denk bijvoorbeeld aan vragen zoals: 'toon een lijst van onze distributiecentra, gesorteerd naar de absolute toename in overslag over de afgelopen drie jaar.' Of: 'voor hoeveel procent van de sales teams die achter blijven op hun target, geldt dat zij zes maanden later wel hun doelstellingen realiseren?'
Dus voor tactische query's is weinig of geen historie vereist, en zij dienen voornamelijk voor de ondersteuning van operationele taken. Voor de beantwoording van strategische query's is wel historie vereist. Hierbij zoekt men naar trends waarbij ook vaak enige aggregatie nodig is. Als er al voldoende historie beschikbaar is in operationele systemen, dan is deze als regel niet opgeslagen op een wijze die eindgebruikers ondersteunt bij hun OLAP vragen. In het bijzonder strategische query's vereisen dat de gegevens zodanig worden gemodelleerd dat longitudinale patronen kunnen worden opgespoord. Met andere woorden, er dient een expliciete tijdsdimensie aanwezig te zijn in het datamodel.

Virtuele datawarehouses

Aanvankelijk hielden we datawarehousing en operationele systemen strikt gescheiden. De nieuwe trend is echter om deze systemen steeds meer te consolideren. Dit vereist echter wel aanzienlijk krachtiger hardware. En inderdaad wordt hardware ook steeds betaalbaarder, niet in het minst door de concurrentiedruk van nieuwe spelers op deze markt. Maar daarnaast stelt nauwere integratie tussen operationele systemen en BI/DWH/oplossingen veel hogere eisen aan netwerk en bandbreedte. Indien deze problemen technisch al overwonnen kunnen worden, lijken contracten in deze telecommunicatiemarkt nog goeddeels binnen een oligopolie beschermd. Maar hoe lang nog? Feit blijft dat de ontwikkeling van 'virtuele datawarehousing' gestoeld is op twee aannames. Er is de aanname dat er voldoen-

de rekenkracht beschikbaar is, dan wel dat eventueel beslag dat BI-systemen doen op hardware die (primair) ter beschikking staat voor operationele processen hier geen noemenswaardige ontregeling veroorzaakt. De tweede aanname is dat er voldoende bandbreedte beschikbaar is om zowel operationele processen alsook BI-applicaties naar behoren te kunnen laten functioneren. Maar aan deze aannames wordt (lang) niet altijd voldaan. Wanneer achterliggende operationele systemen de fysieke repository voor het virtuele DWH vormen, zal daar voldoende historie beschikbaar moeten zijn – voldoende voor zowel tactische als strategische query's. Let wel, als meerdere operationele systemen gegevens aanleveren, vormt het systeem met de *minste* historie de bottleneck.

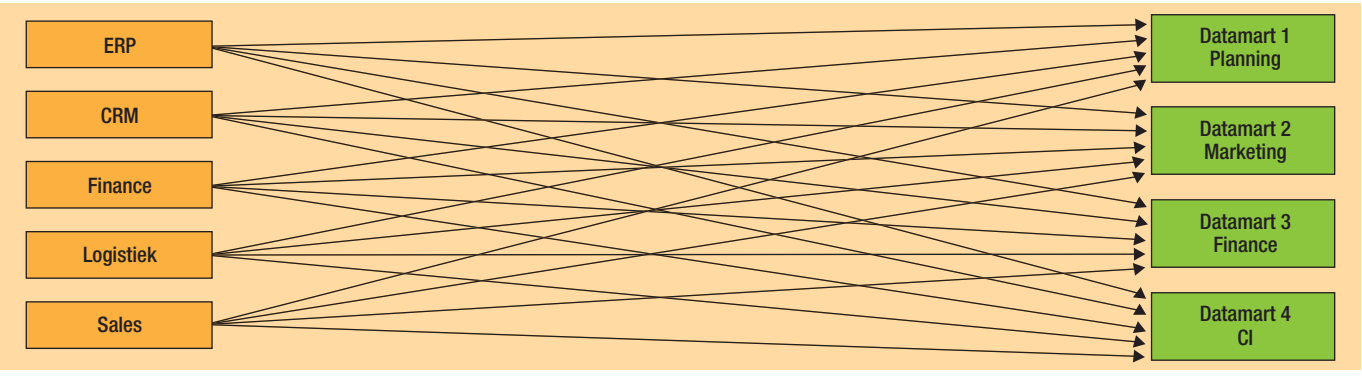
De hoeksteen van OLAP is dat snelle responstijden voor multidimensionele query's de eindgebruiker ondersteunen bij zijn (interactieve) denkproces. De praktijk is dat acceptabele responstijden alleen gerealiseerd kunnen worden door aggregaten op zijn minst ten dele voor te berekenen. Ofwel dit moet in het operationele systeem gebeuren en vormt daar een (hele) vreemde eend in de bijt, óf dit gebeurt in het virtuele DWH – een oxymoron!
Om een goede performance te realiseren vereist een OLAP datamart ruime, en zorgvuldig geadministreerde indexering. Hier moet gebalanceerd worden tussen zo veel mogelijk indexen aanleggen, en de bijbehorende (extra) overhead die dat met zich mee brengt. Wanneer werkelijk *alle* mogelijke waarden worden voorberekend, zal de kubus dermate groot worden dat de besparing in CPU's teniet wordt gedaan door een hogere I/O als gevolg van de (veel) grotere database (MOLAP). De zoektocht naar een 'optimale' indexering (ROLAP) is complex, aangezien je voor het gros aan gebruikers goede performance wilt aanbieden, en voor de kleine groep *power users* (die typisch veel zwaardere query's afvuren) ook een acceptabele responstijd wilt bewerkstelligen. Je probeert zowel average, best-case als worst-case respons te minimaliseren. Deze balans wordt het beste bewaakt in een daartoe geoptimaliseerde (fysieke) omgeving.
Doorgaans huizen de data die nodig zijn voor het complete spectrum aan OLAP query's (zowel qua onderwerp als gebruiksdoel) in een verzameling heterogene databases. Hier moet heel wat aan 'gemasseerd' worden, voor de gegevens geschikt zijn om te integreren in één omgeving. Attribuutwaarden moeten op elkaar worden afgestemd. Dit kan eenvoudig zijn als 'M' in het ene systeem overeen komt met '1' in het andere systeem en wordt afgebeeld als 'man'. Maar veel belangrijker en lastiger: gebruik van de attribuuthiërarchieën waaraan OLAP voor een belangrijk deel zijn unieke waarde ontleent is gebaat bij een 'managed' omgeving. Vooral *ragged* hiërarchieën zijn veranderlijk, *leveled* hiërarchieën veel minder. Deze uitdaging is nog groter als je bedenkt dat dit consistent door de tijd heen moet gebeuren.

> Beheerkosten nemen als regel een groot deel van de totale BI-kosten voor hun rekening

Consistente informatievoorziening vereist dat hetzelfde 'feit' bij alle afdelingen ook gebaseerd is op dezelfde definitie. Maar wanneer ik bij klanten navraag wat de definitie is van zoiets eenvoudigs als een 'klant', ben ik elke keer weer verbaasd over de uiteenlopende antwoorden. Virtuele datawarehousing veronderstelt impliciet dat metadata over systemen heen op elkaar zijn aangesloten. De praktijk is echter weerbarstig.

Datakwaliteit

Gegevens in operationele systemen worden dikwijls gekenmerkt door een hoog percentage *missings*. Soms zijn gegevens terecht missing, en soms ook niet. Die ontbrekende waarden moeten ieder op hun eigen manier worden geïmputeerd. Een (longitudinaal) consistente vervanging van deze missings door de gepaste waardes vereist een managed omgeving als een DWH. Het is een illusie dat dergelijke complexe operaties 'on the fly' in een virtueel DWH kunnen worden bewerkstelligd.
In sommige gevallen kan men inconsistent of foutief gevulde waarden tegenkomen, en ook hiervoor zal een adequate mapping moeten worden gekozen. Foutief gevulde attributen moeten door middel van datakwaliteitsregels worden opgespoord. Indien er voldoende redundantie is, of een 'system of record', kunnen

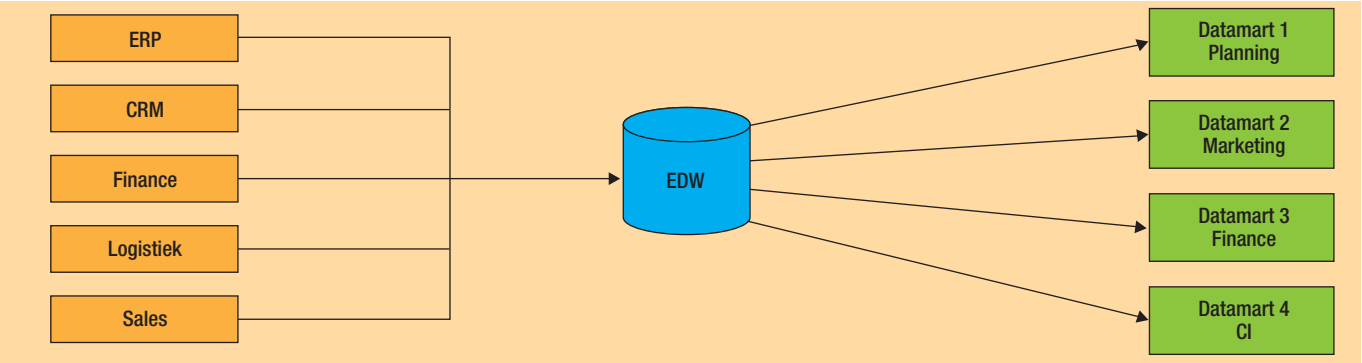


Afbeelding 1: Onafhankelijke datamarts.

deze foutief gevulde waarden worden overschreven. Wederom is een 'echt' DWH de beste plaats hiervoor, aangezien deze correctie in redelijkheid alleen tijdens een overslag van bron naar DWH of van DWH naar datamart kan plaatsvinden.

Architectuur

Er is een belangrijk onderscheid tussen afhankelijke en onafhankelijke datamarts. Een onafhankelijke datamart wordt rechtstreeks gevuld vanuit achterliggende operationele systemen.
Een afhankelijke datamart wordt (alleen) gevuld vanuit een DWH. De architectuur die out-of-the-box OLAP vendors aanbevelen, soms refererend naar een virtueel DWH waar data-integratie plaats vindt, is dus een onafhankelijke datamart.
Het zal menigmaal gebeuren dat een en dezelfde bron in meerdere datamarts wordt geraadpleegd. Of wellicht gaat een al ontsloten bron nog gebruikt worden in nieuw te bouwen datamarts. Wat is het 'juiste' niveau van granulariteit dat daarbij moet worden opgevraagd?
Een minder evident probleem is dat gegevensbeheer in dienst van de organisatie als geheel niet altijd in lijn hoeft te liggen met de belangen van datamart eigenaars (op BU niveau). In deze gevallen zal goede Data Governance er toe leiden dat er 'overall' een rationele afweging van belangen gemaakt zal worden, iets dat per definitie ondenkbaar is vanuit het perspectief van de (onafhankelijke) datamart.
Wat op de korte termijn een snelle, efficiënte en flexibele oplossing lijkt, kan op de lange termijn rampzalig uitpakken – op een moment dat de weg terug langer en langer is geworden!
Bill Inmon heeft de term 'informatie-ecologie' geïntroduceerd. Indianen zeggen dat wij de aarde niet bezitten, maar in bruikleen hebben van onze kinderen. Analoog hieraan dragen we als IT-professionals verantwoordelijkheid voor de keuzes die we maken, dus moeten we in ogenschouw nemen wat de impact op onze 'informatie-omgeving' en toekomstige ontwikkeling zal zijn. Een model kan dit verduidelijken.
Wanneer we kiezen voor een topologie met onafhankelijke datamarts, en we stellen ons een omgeving voor met M bronnen (5) en N datamarts (4), dan ontstaat het model zoals te zien in afbeelding 1. In een hub-and-spoke topologie, waarbij dus wél gebruik gemaakt wordt van een DWH, ziet ditzelfde model er uit zoals in afbeelding 2.
Beheerkosten nemen als regel een groot deel van de totale BI-kosten voor hun rekening, een zorgpunt voor menige CIO. De totale overhead aan interfaces die gemanaged moet worden in het model uit afbeelding 1 (onafhankelijke datamarts) is gelijk aan M x N, in dit geval 20. Uiteraard is dit een maximum omdat mogelijk niet alle bronnen in alle datamarts worden ontsloten.
Bij het hub-and-spoke model is het aantal interfaces M + N (9). Wanneer de getallen M en N (nog) laag zijn, liggen deze vrij dicht bij elkaar. Maar als M en N groeien wordt duidelijk waarom de CIO op den duur gewurgd wordt door excessieve beheerkosten.



Afbeelding 2: Afhankelijke datamarts, hub-and-spoke topologie.

Maar er zijn meer problemen:

- Wanneer bronnen onafhankelijk gegevens extraheren, moet er herhaaldelijk een beroep gedaan worden op processorcapaciteit van elk achterliggend systeem: ERP, CRM, enzovoort;
- Als een nieuwe datamart gebouwd wordt, is de impact op het model in afbeelding 1 groot, en in afbeelding 2 relatief klein;
- Wanneer een nieuwe bron beschikbaar komt, heeft dit in het eerste model grote consequenties, in het tweede geval slechts beperkt;
- Transformatie en aggregatie zijn dikwijls 'dure' processen voor de server, het kan niet efficiënt zijn die voor afzonderlijke datamarts te herhalen;
- Een centrale metadata repository is op zichzelf al moeilijk te realiseren, maar als die ook nog versnipperd wordt, en er moeizame interfaces tussen afzonderlijke metadata-oplossingen nodig worden is de hoop op goede en betrouwbare metadata snel vervlogen.

Al met al komt er een overduidelijk beeld naar voren van waar veel organisaties mee worstelen: IT is 'te duur', de uitgaven voor beheer zijn te groot in verhouding tot budget voor ontwikkeling, veranderingen gaan te langzaam en kosten te veel, er is 'nooit iets mogelijk', etcetera. Het is natuurlijk aan IT-professionals om hierin hun verantwoordelijkheid te nemen en de afwegingen met betrekking tot deze keuzes helder en transparant aan hun business partners voor te leggen.

Conclusie

De keuze tussen wel of geen DWH voor de implementatie van OLAP is ten dele een keuze tussen korte termijn en lange termijn. Het belang van snelle, waardevolle resultaten wordt door iedereen onderschreven. De vraag is of nadelen op de lange termijn bij een architectuur met onafhankelijke datamarts niet teveel problemen veroorzaken zoals:

- Redundantie tussen (onafhankelijke) datamarts;
- Inconsistente, onverenigbare resultaten over datamarts heen;
- Onevenredige beheerkosten (duplicatie, minder flexibiliteit) voor de interfaces tussen datamarts en bronsystemen.

Voor strategische OLAP query's is historie nodig die dikwijls niet voorhanden is. Het is ook niet functioneel die historie op te slaan in operationele systemen. Een hatchback kan uitermate handig zijn om iets te vervoeren, maar als je gaat verhuizen huur je toch ook een busje of boedelbak?

Aangezien het systeem met de minste historie de bottleneck is voor een virtuele OLAP datamart, is het belangrijk om niet alleen de onmiddellijke, maar ook de te verwachten toekomstige requirements in overweging te nemen bij het maken van een tool-keuze.

Virtuele DWH's kennen vooralsnog een aantal beperkingen. Op de eerste plaats qua hardware en bandbreedte. Dit conflict wordt ten dele veroorzaakt door het feit dat primaire processen voorrang verdienen boven BI query's als het gaat om de toewijzing van CPU's. Maar tegelijkertijd is bij OLAP interactieve ondersteuning van het denkproces van analisten een requirement waar je niet (graag) op wilt inleveren.

Datakwaliteit is in veel organisaties een onderschat probleem. Gartner stelt dat inaccurate of incomplete data een van de belangrijkste redenen is voor het falen van BI- en CRM-projecten. Al gauw tweederde van het budget voor een DWH-project wordt opgesoupeerd door ETL, en dit wordt voor een groot gedeelte veroorzaakt door problemen met datakwaliteit. Imputeren van ontbrekende waarden, en correctie van inconsistenties vereisen een managed omgeving. Integratie van metadata-oplossingen verloopt moeizaam door het ontbreken van dominante standaarden, waardoor het haast ondenkbaar is dat al deze problemen virtueel, dus 'on the fly' kunnen worden gemanaged.

Alhoewel een snelle implementatie van 'out-of-the-box' OLAP-oplossingen aantrekkelijk lijkt, zitten er voor de lange termijn zwaarwegende nadelen aan vast. De beheerkosten voor de bedrijfsbrede Data Governance dreigen over de volledige levenscyclus van een DWH onacceptabel hoog te worden. En belangrijker, de toekomstige wendbaarheid komt onder druk te staan. En als er een ding zeker is voor een succesvol DWH, dan is het dat er veranderingen gevraagd zullen worden. En zoals Peter Drucker al zei: als er geen business case is om het goed te doen, hoe kan het dan de moeite zijn om er überhaupt aan te beginnen?

Tom Breur is eigenaar van XLNT Consulting.