EEN KORT OVERZICHT VAN DATA WAREHOUSING EN OLAP

Jef WIJSEN
Professeur, Université de Mons-Hainaut (UMH) - Institut d'Informatique

▪ Veel organisaties hebben in de loop der jaren een overstelpende hoeveelheid gegevens verzameld in diverse database-systemen. De term "data warehousing" omvat de technologie om deze operationele en historische gegevens in een geïntegreerde en samengevatte vorm beschikbaar te maken voor de bedrijfsvoering van een organisatie. Het analyseren en verwerken van deze geaggregeerde informatie staat bekend onder termen als "OLAP (Online Analytical Processing)" en "data mining". Dit artikel geeft een beknopt overzicht van deze nieuwe technologieën, die de laatste jaren in een stroomversnelling zijn geraakt.

▪ Au fil des années, de nombreuses entreprises ont accumulé d'énormes quantités de données dans différents systèmes de bases de données. La notion de "data warehousing" désigne la technologie conçue en vue de rendre ces données opérationnelles et historiques disponibles pour la gestion d'une organisation dans une forme intégrée et condensée. L'analyse et l'exploitation de cette information agrégée est connue sous les appellations "OLAP (Online Analytical Processing)" et "data mining". Cet article donne un bref aperçu de ces nouvelles technologies, très en vogues ces dernières années.

Sinds de jaren zestig hebben onderzoek en ontwikkelingen in databases zich toegelegd op de automatisatie van repetitieve taken (of "transacties"), hetgeen thans aangeduid wordt met de term *online transaction processing* (OLTP). Dit onderzoek heeft onder andere geleid tot de zeer succesvolle relationele database-systemen. Veel bedrijven hebben in de loop der jaren een overstelpende hoeveelheid gegevens verzameld in dergelijke databases. Deze historische en operationele gegevens verhullen vaak een schat aan kennis over het bedrijf en de business, in de vorm van verborgen regels, trends, patronen... Een nieuwe uitdaging is om deze onbekende kennis te onthullen en bruikbaar te maken voor de bedrijfsvoering. Dit heeft geleid tot drie nieuwe technologische ontwikkelingen op het vlak van beslissingsondersteunende systemen *(decision support systems):*

	OLTP	OLAP
gebruiker	klerk.	manager.
werklast	frequente transacties:	regelmatige analyses:
toegang	lezen en schrijven, een tiental records.	vooral lezen, scan van miljoenen records.
gegevens	actueel.	actueel en historisch.
database grootte	100 MB tot GB.	100 GB tot TB.

Fig. 1: Verschillende karakteristieken van transactionele en beleidsondersteunende systemen.

Data warehousing

De OLTP gegevens zijn vaak verspreid over verschillende systemen, sterk gedetailleerd en/of van slechte kwaliteit. Data warehousing omvat het integreren, samenvatten en zuiveren van deze gegevens in een nieuwe opslagplaats, teneinde ze bruikbaar te maken voor analyse.

OLAP

Dit is een acroniem voor *online analytical processing:* het interactief analyseren van de gegevens in het data warehouse, waarbij de gegevens doorgaans voorgesteld en gemanipuleerd worden in de vorm van multidimensionale matrices of spreadheets, aangeduid met de term gegevenskubus *(data cube).*

Data mining.

Het exploreren van de gegevens op zoek naar interessante, nieuwe kennis (regels, trends, patronen...).

Er bestaat geen scherpe grens tussen OLAP en data mining. In OLAP zal de analist gewoonlijk precieze instructies geven, bijvoorbeeld over het deel van de gegevens waarop gefocust moet worden. In data mining zal het systeem vaak zelf die focus bepalen. Zo is de vraag *"Geef per maand het aantal abonnees dat ons verlaten heeft in het voorbije jaar",* een typische OLAP query[1], terwijl een data mining query eerder zou vragen *"Welke factoren beïnvloeden het verlies van abonnees?".*
Figuur 1 toont dat OLAP andere technologische eisen stelt dan het traditionele OLTP. Transactionele databases zijn afgestemd op een typische, gekende OLTP werklast. Het toevoegen van complexe OLAP queries aan deze werklast zou kun-

nen leiden tot een onacceptabele performantie van het gehele systeem. Dit is een bijkomende reden om een data warehouse te bouwen los van de bestaande transactionele databases.

Dit artikel is gestructureerd als volgt. Sectie 2 definieert het concept "data warehouse" en legt uit welke de verschillende stappen zijn bij de constructie van een data warehouse. Het concept "data mart" komt hier eveneens ter sprake. Sectie 3 gaat dieper in op het gegevensmodel en de queries in een OLAP omgeving. De termen ROLAP en MOLAP worden verduidelijkt. Sectie 4 duidt aan wat het verschil is tussen data mining en OLAP. Het was niet ons opzet om diep in te gaan op het onderwerp data mining. Sectie 5 geeft een visie op toekomstige ontwikkelingen op het vlak van data warehousing. Sectie 6 beschrijft een praktijkervaring. Sectie 7 eindigt met een persoonlijke keuze uit de literatuur.

Data warehouse

Wat is een data warehouse?

Referentie (2) definieert een data warehouse als een subject-georiënteerde, geïntegreerde, niet-vluchtige, historische gegevensopslagplaats ter ondersteuning van de besluitvorming op het niveau van het management van een bedrijf. We lichten deze termen kort toe.

Subject-georiënteerd en geïntegreerd

OLTP gegevens zijn vaak verpreid over verschillende databases die elk een bepaalde applicatie (facturering, levering, productie, ...) ondersteunen. Deze gegevens worden in een data warehouse geïntegreerd en georganiseerd rond bepaalde onderwerpen (klant, product, leverancier, ...).

Niet-vluchtig en historisch

De term "nietvluchtig" betekent dat de gegevens, eenmaal ingevoerd in het data warehouse, niet gewijzigd worden, hoewel verwijderen eventueel wel mogelijk is. De term "historisch" duidt erop dat de gegevens een zekere tijdspanne bestrijken, nodig voor het analyseren van trends.

Wat is een data mart?

Veel bedrijven wensen één enkel data warehouse te bouwen voor de ondersteuning van het geheel van beslissingsondersteunende activiteiten. De constructie van een data warehouse dat het ganse bedrijf omspant, is evenwel een complex proces dat een uitgebreide modellering van de business vereist en verschillende jaren in beslag kan nemen. Sommige bedrijven kiezen daarom voor *data marts*, die kunnen omschreven worden als departementale data warehouses gefocust op een specifiek onderdeel van de business. Een voorbeeld is een marketing data mart met informatie rond de onderwerpen klant, product en verkoop. Deze data marts kunnen sneller gerealiseerd worden omdat ze geen consensus op het bedrijfsniveau vereisen, maar kunnen op lange termijn tot complexe integratieproblemen leiden indien geen compleet business model voorhanden is.
Zelfs als er een geïntegreerd data warehouse bestaat, kan het om redenen van flexibiliteit en performantie interessant zijn om data marts te extraheren uit het data warehouse.

Constructie van een data warehouse

De constructie en het onderhoud van een data warehouse is een complexe bezigheid, die verschillende taken omvat. Voor elk van deze taken zijn commerciële hulpmiddelen beschikbaar. Figuur 2 geeft een schematisch overzicht van de verschillende fasen en componenten betrokken bij data warehousing.

Extractie

Extractie is het proces dat gegevens onttrekt aan de transactionele databases en andere gegevensbronnen. Het betreft typisch een (nachtelijk) batch proces, vaak verschillende subprocessen in parallel om de doorlooptijd te verkorten.

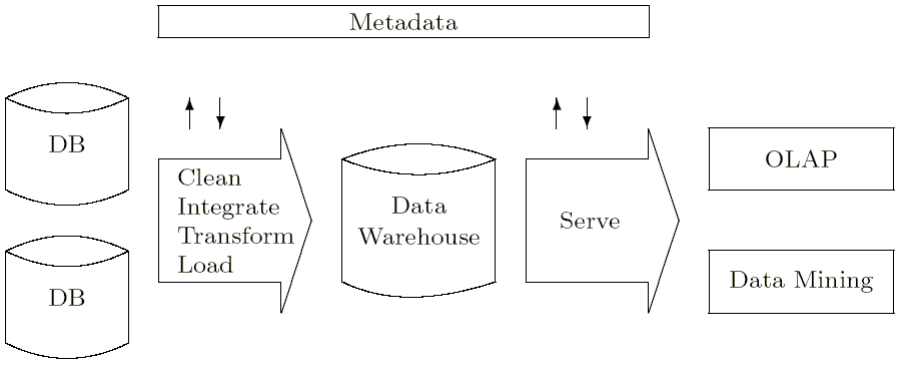

Fig. 2: Schematisch overzicht van data warehousing

Zuiveren

Vermits een data warehouse wordt gebruikt in de besluitvorming, is het belangrijk dat de gegevens in het data warehouse correct zijn. Aangezien in een data warehouse grote hoeveelheden gegevens uit verspreide, heterogene gegevensbronnen samengevat worden, bestaat er een reëel gevaar voor fouten en anomalieën. Typische zuiveringsoperaties zijn het invullen van ontbrekende waarden, het corrigeren van typografische fouten en het uniformiseren van synoniemen. Gegevens die duidelijk foutief zijn maar niet kunnen worden gecorrigeerd, worden verwijderd.

Integratie en transformatie

De gezuiverde gegevens bevinden zich doorgaans nog niet in de vorm vereist door het data warehouse. Integratie omvat het samensmelten van verschillende gegevensbronnen. Een typisch probleem hierbij is het met elkaar in overeenstemming brengen van identificaties van entiteiten: hoe kan men uitmaken dat klant-id in de ene databases naar dezelde entiteit verwijst als kl_nr in de andere? Een ander probleem betreft het doen overeenstemmen van gegevens uitgedrukt in verschillende eenheden: de ene database bevat verkoopcijfers per maand en de andere per week, waarbij een week kan gesplitst zijn over twee maanden. Integratie kan leiden tot nieuwe ongerijmdheden in de gegevens, hetgeen een bijkomende fase van zuivering kan vereisen.

Voorbeelden van gegevenstransformaties zijn aggregatie en normalisatie: individuele verkopen kunnen geaggregeerd worden tot dagelijkse verkoopcijfers; de waarde van een variabele kan genormaliseerd worden zodat hij tussen 0 en 1 komt te liggen.

Laden en actualiseren *(refresh)*

De gezuiverde, geïntegreerde en getransformeerde gegevens worden vervolgens in het data warehouse geladen. Op dit moment worden indexen gecreëerd nodig om zoekopdrachten te versnellen. Vermits de informatie in het data warehouse een (sterk bewerkte) kopie is van de gegevens in de OLTP databases, is het nodig om het data warehouse op geregelde tijdstippen te actualiseren. Het is doorgaans te duur om het data warehouse daarbij volledig te ledigen en vervolgens opnieuw te laden.

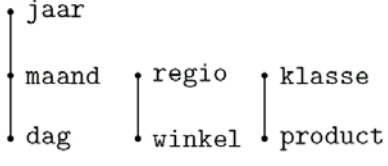

Fig. 4 : Concept hiërarchieën.

Daarom worden incrementele *update* technieken gebruikt, waarbij wijzigingen in de OLTP databases op regelmatige tijdstippen worden gepropageerd naar het data warehouse.

OLAP

Gegevenskubus

Zoals reeds aangehaald in Sectie 1 slaat OLAP *(online analytical processing)* op het interactief analyseren van de gegevens in het data warehouse. Dergelijke analyses zijn doorgaans gebaseerd op allerlei overzichtsrapporten, bijvoorbeeld de maandelijkse verkoopcijfers per regio en per product. Men kan zich zo een overzichtsrapport gemakkelijk voorstellen als een kubus: de drie dimensies van de kubus indiceren de maanden, de regio's en de producten; een cel van de kubus met coördinaat *<jan 2001, België, Lego>* bevat het aantal verkochte stuks Lego in België gedurende januari 2001.
Ter ondersteuning van OLAP worden de gegevens in het data warehouse dus doorgaans voorgesteld in de vorm van een multidimensio nale matrix of "gegevenskubus". Het zou correcter zijn om te spreken over hyperkubus in plaats van kubus, omdat het aantal dimensies verschillend van drie kan zijn. De dimensies van de kubus corresponderen met onafhankelijke variabelen en de cellen van de kubus bevatten de overeenkomstige waarden voor de afhankelijke variabele(n), ook wel aangeduid als meting(en) *(measure(s)).* Figuur 3 toont een kubus met drie dimensies dag,

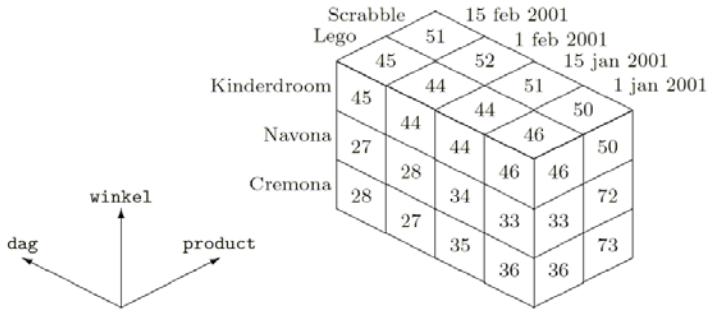

Fig. 3: Een gegevenskubus met drie dimensies.

winkel en product; de cellen van de kubus bevatten het aantal verkochte stuks.
OLAP software biedt doorgaans verschillende mogelijkheden voor de grafische visualisatie van de gegevenskubus. Elke cel kan bijvoorbeeld een blokje bevatten waarvan de grootte of de kleurintensiteit evenredig is met de numerieke celwaarde.
De dimensies zijn doorgaans georganiseerd in hiërarchieën, die de wijzen vastleggen waar-

op de gegevens logisch kunnen worden gegroepeerd. Figuur 4 toont dat dagelijkse cijfers aanleiding geven tot maandelijkse en jaarlijkse totalen; winkels zijn gegroepeerd in regio's, producten in klassen.

Rollup: een typische OLAP query

Een typische OLAP query geeft voor elke dimensie het niveau aan waarop de gegevens moeten gepresenteerd worden. Figuur 5 toont het antwoord op de query *"Geef totale verkoopcijfers per product, regio en maand".* Een dergelijke geaggregeerde kubus, berekend uit de basiskubus, wordt ook wel *kuboïde* genoemd.

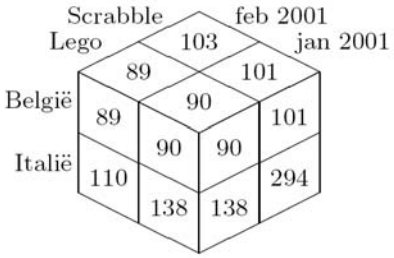

Fig. 5 : De kuboïde maand region product

Het overgaan van een gedetailleerde naar een geaggregeerde kubus wordt aangeduid met de term *rollup.*
OLAP queries kunnen ook een reductie van het aantal dimensies inhouden. De tweedimensionale kuboïde maand product in Figuur 6 geeft maandelijkse verkoopcijfers per product, over alle winkels heen.
Uiteraard kan elke kuboïde desgewenst berekend worden uit de basiskubus. Nochtans zullen OLAP systemen op bepaalde vragen anticiperen en kuboïdes reeds op voorhand berekenen en "materialiseren", teneinde antwoordtijden te verminderen. De wijze waarop dit best gebeurt, is een belangrijk onderzoeksthema.

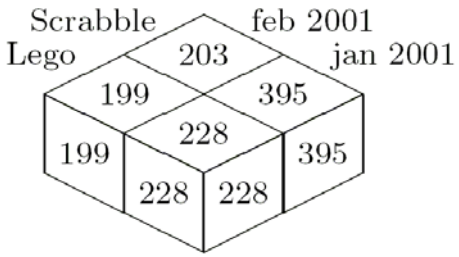

Fig. 6 : De kuboïde maand product.

Technologische keuze: ROLAP of MOLAP

De technologische uitdaging in OLAP kan ruwweg worden omschreven als het efficiënt ondersteunen van spreadsheet operaties op databases van meerdere gigabytes. Al naargelang de gebruikte technologie kan OLAP software worden bestempeld als ROLAP of MOLAP: ROLAP *(relational OLAP)* maakt gebruik van bestaande relationele database-technologie; het uitgangspunt van MOLAP *(multidimensional OLAP)* daarentegen is dat SQL databases weinig geschikt zijn om OLAP te ondersteunen.

ROLAP

Het multidimensionale gegevensmodel dat hoger beschreven werd, kan gemakkelijk geïmplementeerd worden in een conventionele SQL database. De basiskubus wordt gestockeerd in een zogeheten "feitentabel", en de hiërarchieën in "dimensietabellen". Het resulterende schema wordt doorgaans een "sterschema" genoemd. Figuur 7 toont dit voor het lopende voorbeeld. Merk op dat de feitentabel reeds vooraf bere-

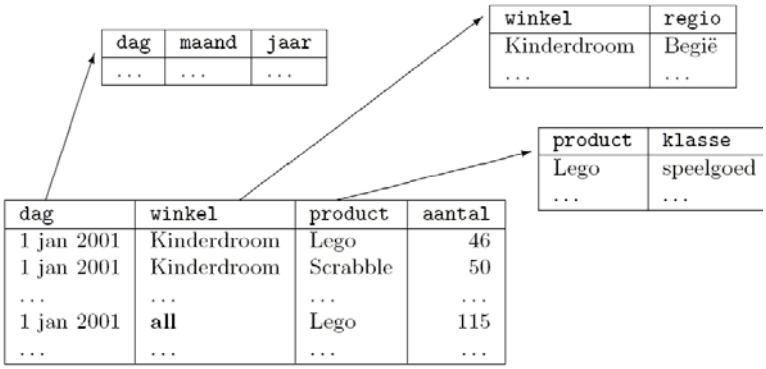

Fig. 7: Een sterschema.

kende subtotalen kan bevatten (symbool **all**).
ROLAP servers breiden relationele servers doorgaans uit met gespecialiseerde middleware om de uitvoering van OLAP queries te versnellen. Er worden ook reeds voorstellen geformuleerd om de standaard relationele querytaal SQL uit te breiden met operators ter ondersteuning van OLAP.

MOLAP

Hoewel OLAP gegevens kunnen gestockeerd worden in relationele databases, geloven velen dat MOLAP een beter alternatief biedt. MOLAP impliceert een multidimensionale database waarbij gegevenskubussen worden gestockeerd in *(sparse)* matrices. Deze opslagwijze zou efficiënter zijn dan in ROLAP; het nadeel is dat de integratie met bestaande SQL databases moeilijker is.

Data mining

In OLAP stuurt de eindgebruiker zelf de analyse: hij kiest de dimensies en afhankelijke variabelen en specificeert de queries. De analist vormt een hypothese en verifieert deze vervolgens met een reeks queries. Een probleem met deze werkwijze is dat de gegevens in het data warehouse vaak onvoldoende begrepen zijn (dit is precies de motivatie om analyses uit te voeren!), zodat het quasi onmogelijk wordt om interessante hypotheses te maken, de juiste gegevenskubus te kiezen en de goede queries te stellen. Het uitgangspunt van data mining is om de kracht van de computer te gebruiken om interessante patronen in de gegevens te onthullen, eerder dan hypothetische patronen te verifiëren.

We nemen een probleem van classificatie als voorbeeld. Veronderstel dat we het risico willen inschatten verbonden aan een nieuwe kredietaanvraag. Gebruik makend van historische data warehouse gegevens over achterstallige kredieten, zou een data mining programma de volgende regel kunnen afleiden: *"Indien inkomen ≤ 20.000 Euro en anciënniteit ≤ 5 jaar dan risico = hoog, anders risico = laag".* Het programma ontdekt zelf dat inkomen en anciënniteit het risicobedrag bepalen, eerder dan andere kenmerken zoals leeftijd, opleiding,.. .

Toekomst van data warehousing

Data warehousing, OLAP en data mining zijn snel groeiende technologieën. Toekomstige ontwikkelingen kunnen in verschillende richtingen gaan. We wagen ons aan een drietal voorspellingen.

Beslissingsondersteunende systemen zullen in de toekomst meer pro-actief worden. Ze zullen niet wachten tot de analist een OLAP of data mining query stelt, maar zullen zelf aan continue zelfanalyse doen. Wanneer het systeem een interessante nieuwe evolutie ontdekt, zal het een waarschuwing zenden naar de gebruiker.

De hedendaagse systemen voor data warehousing en data mining zijn vaak ontworpen om te werken voor gelijk welke business. Toekomstige systemen zullen meer gericht zijn op een welbepaalde business, bijvoorbeeld de petroleumsector. Dergelijke business-specifieke systemen zullen meer domeinafhankelijke logica bevatten die de kwaliteit van de analyses zal verhogen.

Het data warehouse zal in toenemende mate aangevuld worden met achtergrondinformatie afkomstig uit externe gegevensbronnen, beschikbaar via bijvoorbeeld het Web. Een onderneming heeft geen directe controle over de toegang tot deze externe gegevensbronnen, dit in tegenstelling tot de eigen databases. Standaarden voor gegevensuitwisseling, zoals XML, spelen hierbij een belangrijke rol.

Getuigenis

De eindverhandeling (3) is illustratief voor de problemen die kunnen gepaard gaan met de invoering van data warehousing en data mining in een onderneming. De onderneming in kwestie beschikt over een tiental belangrijke applicaties, die gebruik maken van database management systemen van een viertal verschillende leveranciers. Het oorspronkelijke opzet van het onderzoek was om door middel van data mining technieken een antwoord te vinden op vragen als: *"Wie zijn onze klanten?"* en *"Met welke service zijn de klanten meest gebaat?".* Al spoedig bleek dat data mining onmogelijk zou zijn zonder een grondige "preparatie" (zie sectie 2.3) van de gegevens. Enkele netelige kwaliteitsproblemen waren de volgende:

- Dubbele registratie van eenzelfde entiteit. Bijvoorbeeld, <*RAYTEC, Rue du Commerce 2,...>* en *<S.A. RAYTEC,* 2 *Rue du Commerce,...>.*
- Veelvuldig gebruik van de "fourre-tout" code *"andere"* voor attributen zoals opleiding of beroep. Een ander probleem is dat de codes in verschillende toepassingen niet op elkaar afgestemd zijn.
- Ontbrekende, onmogelijke of achterhaalde attribuutwaarden.

Verdere informatie

Mijn persoonlijke ervaring i.v.m. de literatuur over OLAP en data warehousing is dat sommige documenten op het Web interessanter en actueler zijn dan heel wat boeken. De Web site <http://www.daniel-lemire.com/OLAP/> bevat een interessant overzicht van wetenschappelijk onderzoek in data warehouses en OLAP; de site bevat ook links naar *white papers* geschreven door commerciële software leveranciers. Referentie (1) geeft een degelijk overzicht van data warehousing en data mining. Referentie (4) is een goed boek over data mining.

In tegenstelling tot de database-industrie, is de OLAP-markt vandaag de dag sterk gefragmenteerd, zonder dominante spelers. De Web site <http://www.olapreport.com> bevat interes-

sante informatie over de OLAP-markt. De Web site van KDnuggets <http://www.kdnuggets.com> geeft een overzicht van data mining software.

Jef WIJSEN
Université de Mons-Hainaut
Bâtiment «Le Pentagone»
Avenue du Champ de Mars, 6
7000 Mons
Jef.Wijsen@umh.ac.be

17 januari 2006

Bibliografie

1. Han, J. ; Kamber, M. *Data Mining: Concepts and Techniques, 2nd ed.* San Francisco : Elsevier/Morgan Kaufmann, 2005, 800 p.

2. Inmon, W. *Building the Data Warehouse 2nd ed,* New-York : Wiley, 1996.

3. Van Puyvelde, H. *De l'information opérationnelle à l'intelligence décisionnelle par le data mining - Étude de faisabilité appliquée au cas d'un service public*. Master's thesis, Université de Mons-Hainaut, 2000.

4. Witten, I. ; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed.,* San Francisco : Elsevier/Morgan Kaufmann, 2005, 560 p.
   Deze studie bevestigt wat verschillende auteurs vermelden: dat vaak 80% of meer van de inspanningen in een data mining project besteed wordt aan het prepareren van de gegevens.

Nota's

1. Een query is een vraag gesteld aan een databasesysteem in een bepaalde gegevensopvraagtaal.