

Teradata biedt ook low-end oplossing

Multi-temperature datawarehouse

Paul van der Linden

Roep Teradata en de reactie is: high-end data-warehousing. Dat is niet onjuist, maar ook zeker niet het hele verhaal. Meer dan 25 jaar kennis en ervaring in datawarehousing komt niet overeen met het beperkte beeld wat we van Teradata hebben. Wie zo lang in deze markt meeloopt, moet iets goed doen. Zo ziet Gartner Teradata als de onbetwiste leider in DBMS'en en in database servers. Hetgeen zich niet vertaalt in een dominant marktpercentage. Tijd om eens langer naar deze veteraan te kijken.

In Stephen Brobst heeft Teradata een uitstekende vertolker van Teradata anno 2006. Brobst, chief technology officer, maakt de indruk van alle hoeken en gaten van datawarehousing op de hoogte te zijn en trakteert zijn toehoorders in hoog tempo op een altijd interessant verhaal. Vaak met een (hoog) technisch gehalte, maar het kan even goed om de ontwikkelingen in de markt gaan en de bewegingen die de concurrenten maken. Dat is dan ook het nieuwe Teradata: niet langer uitsluitend een technologische club, maar nu ook met steeds meer oog voor de buitenwereld.

Lineaire schaalbaarheid

Sinds de oprichting van Teradata in 1979 heeft het bedrijf een aantal transitie meegemaakt en is nu onderdeel van NCR Corporation. Onderscheidend is de aandacht die Teradata besteedt aan de samenwerking tussen de database en de hardware. Dit is ook het punt waar de connectie met moederbedrijf NCR een rol speelt. Door software en hardware goed op elkaar te tunen pretendeert Teradata lineaire schaalbaarheid te kunnen leveren. Dat betekent dat je met een relatief eenvoudige server kunt beginnen en vervolgens CPU's en schijven kunt bijprikken al naar gelang de behoefte. Normaal gesproken is de resulterende schaalbaarheid daarbij minder dan lineair. Je verliest immers altijd een beetje prestatie vanwege de toename in communicatie die hierbij optreedt. De aandacht die Teradata besteedt aan de combinatie software/hardware zorgt ervoor dat dit prestatieverlies niet optreedt. Met andere woorden: of je nu 500 GB of 5 TB aan data hebt, wat geleverde prestaties betreft zou je het verschil niet mogen merken. De aandacht die Teradata besteedt aan hardware betekent overigens wel dat hun datawarehouseverhaal een stuk technischer is dan gebruikelijk. Termen als rotations per minute

(RPM), FSB (front-side bus) en BLI (Bynet Low Latency Interface) worden gebruikt om de superieure aanpak duidelijk te maken. Wie zich verder wil oriënteren zij hierop bedacht.

Tijdens de Technical Conference die in juni van dit jaar in Salzburg werd gehouden, werd door Scott Gnau (vice-president en general manager R&D) de NCR 5450 officieel gelanceerd. De NCR 5450 is een nieuw onderdeel van het 5400-platform en kan worden uitgerust met twee 3,8 GHz, 2MB cache processors.

PPI maakt het mogelijk om een hash-gedistribueerde tabel ook fysiek te partitioneren

Er is sprake van 64 bit extended memory en hyper-threading. De besturingssystemen die worden ondersteund zijn MP-RAS 3.03, Windows Server 2003 en SUSE Linux. Als database hebben we het over Teradata AWS 4.0 onder Windows 2003. In termen van opslag zijn de mogelijkheden: NCR Enterprise Storage 6842, EMC DMX-3, 9200 en 9202 STK Tape Library's en 9204 en 9205 Disk Backup Library's. Co-existentie met oudere hardware is mogelijk ('shared nothing'). Zoals gezegd: Teradata besteedt in hetzelfde datawarehouseverhaal ook aandacht aan de hardware die erbij hoort. Het Teradata Warehouse (huidige versie 8.1) is dan ook een suite van software, hardware en consulting services.

Multi-temperature datawarehousing

In een door Teradata uitgevoerd onderzoek naar enterprise decision making (2004) gaf 75 procent van de ondervraagde organisaties aan dat het aantal dagelijkse beslissingen is toegenomen. Meer dan de helft van de organisaties geeft daarbij aan dat de complexiteit van de te nemen beslissingen is gestegen. Dat de hoeveelheid data is toegenomen wordt bijna unaniem (98 procent) beaamd. Een kleine 60 procent heeft het hierbij zelfs over een verdubbeling of verdrievoudiging. Dat betekent dat de behoefte aan het kunnen omgaan met grotere hoeveelheden data enorm is gestegen. Het betekent ook dat het steeds moeilijker is geworden om uit deze enorme berg data de relevante inzichten te halen. Voeg daarbij dat de frequentie van dit soort beslissingen ook alleen maar is gestegen (dagelijks in plaats van wekelijks of

maandelijks) en de omvang van het probleem wordt duidelijk. Een oplossing hiervoor zal niet alleen snel genoeg, maar ook betaalbaar moeten zijn.

Teradata schuift haar Active Data Warehouse naar voren als mogelijke oplossing voor deze problemen. Active Data Warehouse (ADW) is een relationele database met parallelle verwerking. Zowel SMP als MPP kunnen worden toegepast. De database is fouttolerant en beschikt over een cost based optimizer. Een van de kenmerken is dat automatisch wordt gezorgd voor reorganisatie van data. Hiervoor is dus geen DBA nodig. Zoals gezegd is het ontstaan van Teradata terug te brengen tot 1979. Sinds die tijd is de basisopzet van de database niet veranderd. ADW is vanaf het begin opgezet als een database die ook met enorme hoeveelheden (Terabytes aan data) moet kunnen omgaan. Op dit moment ligt de grens bij 4 Petabytes. Met een enkele I/O-actie kan 2 GB aan data worden opgehaald.

Dat de hoeveelheid data waarover organisaties beschikken alleen maar toeneemt zal genoeglijk bekend zijn. Niet alle data zijn echter even belangrijk. Sommige data worden vaak opgevraagd, terwijl andere data bijna niet worden gebruikt. Op basis van dit onderscheid in gebruik stelt Teradata voor om deze verschillende soorten data dan ook anders te behandelen. Het 'multi-temperature datawarehouse' is de metafoer die ze hiervoor hanteert. Data die vaak worden opgevraagd zijn dan de 'hot data'. Het gaat hierbij om slechts een beperkt deel van de totale hoeveelheid data. Naarmate de gebruiksfrequentie afneemt gaat het om een groter percentage van de data. Dat betekent dat het grootste deel van de data ('dormant data') bestaat uit gegevens die zo goed als niet (meer) worden opgevraagd. Zowel in de Teradata database-technologie als in de door NCR geleverde opslagtechnologie zitten elementen die het multi-temperature datawarehouse praktisch mogelijk maken.

Partitioned primary index

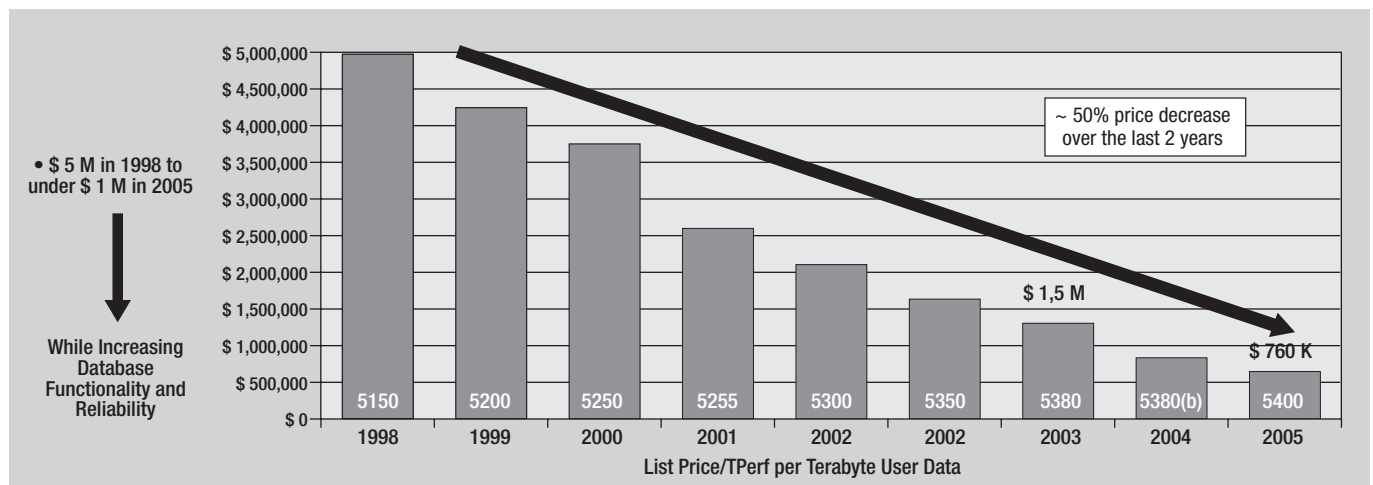
Binnen de database-technologie is een centrale rol weggelegd voor de zogeheten partitioned primary index (PPI). PPI maakt het mogelijk om een hash-gedistribueerde tabel ook fysiek te

partitioneren. Een tabel kan daarbij tot wel 65.000 partities hebben. Als er gepartitioneerd wordt op basis van tijd kunnen vele partities worden gedefinieerd waardoor sneller de beoogde resultaatset kan worden opgehaald. De optimizer ondersteunt partition elimination waardoor niet-relevante partities direct overgeslagen worden. Ook dit resulteert in tijdsinstaat. Met de Dynamic Query Manager is het mogelijk om de query concurrency te regelen. Zo zou je ervoor kunnen kiezen om in het geval van 'hot data' een hogere query concurrency toe te staan dan bij 'cool data'. Natuurlijk kunnen ook de gebruikersrol, periode van de dag of kosten van de query als bepalende variabele worden meegenomen. Het gebruik van sparse indexen, multiple value compressie en het gebruik van velden met variabele lengten zijn enkele andere aspecten van de database waarmee multi-temperature kan worden ingevuld.

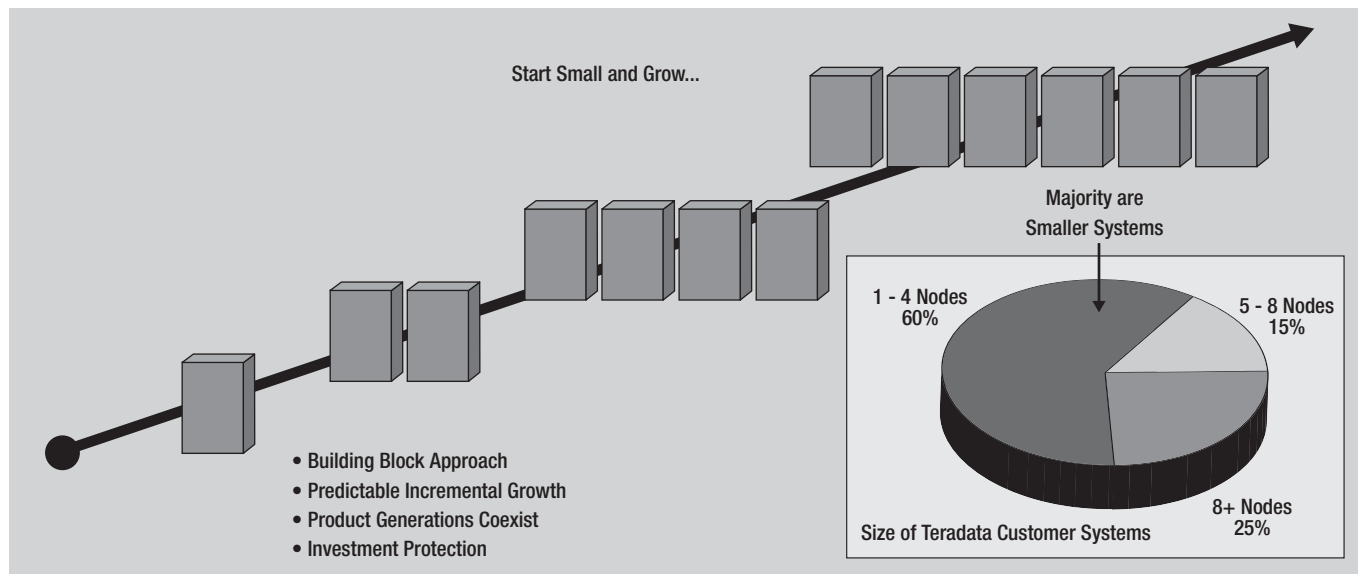
Maar ook de opslagtechnologie van NCR draagt een steentje bij. Wie alleen beschikt over 'hot' en 'warm' data kan het beste kiezen voor een 73 GB 15K RPM on-line disk. Heb je daarnaast ook 'cool' data in je datawarehouse (wat dus betekent; veel meer data die relatief weinig worden opgevraagd) dan ligt de 146 GB variant meer voor de hand. Wie ook zijn 'dormant data' in het datawarehouse wil opslaan kiest het beste voor de 300 GB 15K RPM on-line disk. Deze laatste is overigens nog niet beschikbaar, maar wordt binnen afzienbare tijd verwacht. Ook in storage cabinets laat NCR zich niet onbetuigd. Zowel van de 73 GB-schijven als van de 146 GB-variant kunnen er maar liefst 144 in een cabinet. Dat betekent dat de raw capacity per cabinet neerkomt op 10,5 TB in het eerste geval en 21 TB als wordt gekozen voor 146 GB-schijven.

Meer dan een database

De technisch/technologische invalshoek heeft ertoe geleid dat Teradata vaak werd gezien als louter leverancier van een database. Zoals al aangegeven hoort hardware eveneens tot het aanbod (de NCR-connectie). Maar daarnaast beschikt Teradata ook over andere software dan alleen een database. Allereerst is er TAP; het Teradata Application Platform. TAP biedt



Afbeelding 1: Prijsontwikkeling Teradata.



Afbeelding 2: Groeimodel Teradata.

een collectie van herbruikbare analytische componenten. Een aantal applicaties staat al klaar, zoals customer value, response model, active inventory model, targeted campaign, product sales forecast, propensity model en sales opportunity. Ook is het mogelijk om nieuwe componenten in TAP te definiëren. TAP is Teradata's poging om meer business value te genereren. Het platform is met name interessant voor organisaties die hun eigen applicaties bouwen (in SQL en/of Java), die J2EE gebruiken als enterprise applicatiestandaard, applicaties als componenten opzetten, SOA willen toepassen en open industriestandaarden als XML, UDDI en SOAP toepassen.

Teradata beschikt over een indrukwekkende klantenlijst met vooral veel klanten in Finance en Telecom

De analytische applicaties die Teradata biedt worden door Brobst omschreven als 'deep analytics'. Teradata Profiler wordt gebruikt voor data-exploratie, statistische analyses en het vaststellen van de datakwaliteit. Warehouse Miner is een tool waarmee analytische modellen kunnen worden gebouwd en gebruikt. Daarnaast is er ook software beschikbaar die de samenwerking tussen de Teradata database en bijvoorbeeld SAS software faciliteert. Zo is Teradata ADS bedoeld voor het bouwen en 'refreshen' van analytische datasets. Hiermee is het een tool die wordt ingezet voordat SAS en soortgelijke analytische software kan worden toegepast. Teradata PMML Consumer maakt het direct toepassen van SAS-modellen tegen de Teradata database mogelijk. Naast SAS kan ook software van SPSS, KXEN en andere partijen op deze wijze worden ingezet. Onder de naam Offer Management wordt CRM aangeboden. Offer Management

maakt gebruik van de Teradata database en het Logical Data Model (LDM). Offer Management bestaat uit Teradata Customer Management software, de Offer Optimization Module en The Delivery Channel van Copient. Copient is ook een NCR company. De Teradata Offer Optimizer Module is een webbased applicatie en werkt op de TAP Value Analyzer en de Analytic Calculator zijn andere voorbeelden van applicaties die Teradata in huis heeft.

Interview

"Proprietary chips are dead"; Stephen Brobst is heel stellig in zijn uitspraak dat alleen Intel-chips zullen overleven. Met Intel-chips bedoelt hij dan chips met een Intel-instructieset zoals die worden geleverd door onder andere Intel en AMD. Wie Teradata wil gebruiken kan gaan voor SMP dan wel voor MPP. Teradata geeft zelf de voorkeur aan MPP. Teradata noemt hun versie daarvan Bynet. Het is een gepatenteerde selectie van hardware en software. Terwijl Bynet 1000 nodes kan hebben gaat het meestal om minder dan 12 nodes. Storage is daarbij volgens Brobst niet zo interessant, die kun je bijvoorbeeld bij LSI of EMC halen. Door de prestaties die geleverd kunnen worden zijn datamarts en kubussen niet nodig. Teradata gebruikt virtuele kubussen en een HOLAP-aanpak. Brobst geeft aan dat meer dan de helft van de organisaties beschikt over een datawarehouse waarin minder dan 1 Terabyte aan data zit. Zit er minder dan 100 GB in het datawarehouse dan is Teradata voor de betreffende organisatie niet interessant. Als Teradata met Warehouse 8.1 zo'n onverslaanbare combinatie in huis heeft, waarom is dit dan niet algemeen bekend? "Marketing moet via klanten plaatsvinden", is het oordeel van Brobst. Organisaties die tevreden zijn met Teradata zullen dit verder vertellen. Dat het dan langer duurt voordat iedereen de boodschap heeft begrepen neemt hij voor lief. Het ondersteunen van open marktstandaarden vindt Brobst zeer zeker belangrijk. Maar niet alle standaarden zijn even belangrijk. Zo worden MDX, Xquery en CWM allemaal door Teradata

ondersteund. Maar dat geldt niet voor XML/A. Volgens Brobst heeft XML voor Analysis nog onvoldoende kritische massa om dit initiatief te ondersteunen.

TCO

De combinatie van lineaire schaalbaarheid en multi-temperature datawarehousing zorgt volgens Brobst voor een lage total cost of ownership (TCO). Het commitment dat hij neerlegt is dat de prijs die een organisatie moet betalen voor een datawarehousesysteem van Teradata, lager dan wel hetzelfde is als die van de concurrentie. Daarbij gaat het om een systeem van dezelfde grootte, met dezelfde prestaties en hetzelfde service level. Het gaat hierbij om de hardware, opslag, database, maintenance & support en software subscription.

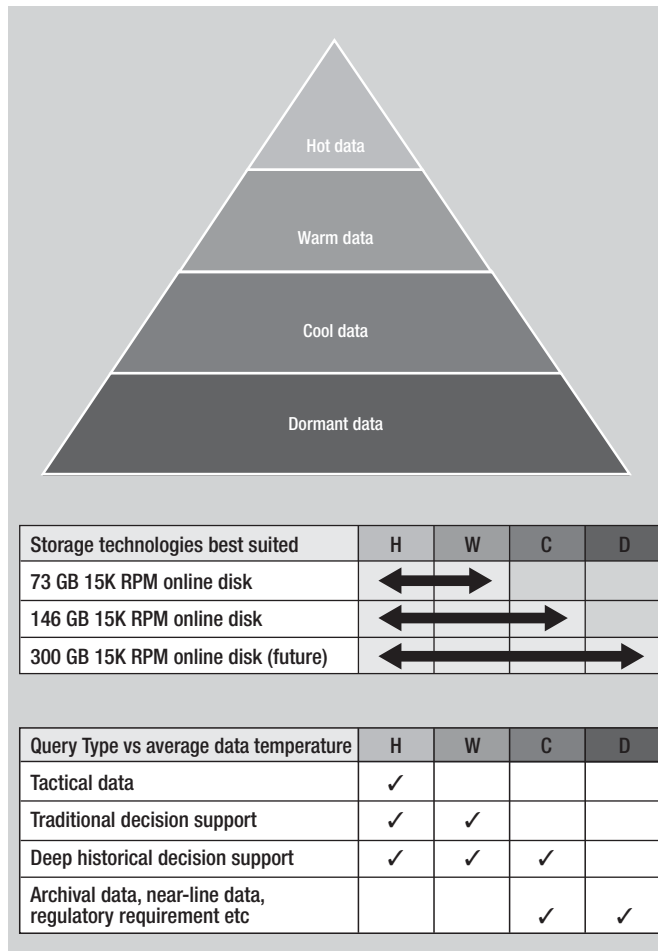
Door de building block-aanpak en een voorspelbare incrementele groei geldt dit commitment niet alleen voor de allergrootste implementaties, maar ook voor kleinere organisaties. Hiermee weerlegt Brobst de in de markt gangbare opvatting dat Teradata alleen geschikt is voor de meest complexe en omvangrijke omgevingen. Jammer alleen dat er geen cijfers beschikbaar zijn gesteld die aantonen dat de TCO van de geschetste Teradata-omgeving inderdaad gunstig afsteekt tegenover de TCO-cijfers van de concurrentie. Die concurrentie wordt overigens gezien als alleen IBM en Oracle. Daarbij is Oracle volgens Brobst meer gefocust op OLTP en op enterprise-applicaties. En heeft IBM zijn ogen gericht op de markt voor infrastructuur.

Teradata beschikt over een indrukwekkende klantenlijst met vooral veel klanten in de Finance- en Telecom-sectoren. Dat is niet vreemd, aangezien daar enorme hoeveelheden data aanwezig zijn en snelle beschikbaarheid daarvan een basisvereiste is. Tot de klanten in Nederland behoren onder andere SNS Bank, ABN/AMRO en Vodafone. Andere EMEA-klanten zijn onder andere Axa België, Belgacom, Deutsche Telecom, France Telecom, Telefonica (Spanje), Metro, Tesco, Casino, BA, Lufthansa en Reuters.

Conclusies

Teradata wordt vaak afgedaan als 'high-end datawarehousing'. Een database die met enorm veel data om kan gaan en een bijbehorend prijskaartje. De werkelijkheid is complexer. Teradata Warehouse 8.1 is een combinatie van hardware, software en diensten. Software is slechts een onderdeel van de oplossing. Met name de onderlinge afstemming tussen software en hardware is wat Teradata onderscheidend maakt. En ook binnen het onderdeel software is er meer dan alleen de database. Een hele reeks analytische applicaties en een heus applicatieplatform (TAP) maken eveneens onderdeel uit van hetgeen Teradata softwarematig te bieden heeft.

Stephen Brobst, chieft technology officer, geeft het commitment af dat een datawarehousing-systeem van Teradata niet duurder is dan dat van de concurrentie. Dat heeft dan betrekking op de hardware, software, opslag en onderhoud & beheer. Lineaire schaalbaarheid en een uitgekende manier om met historische



Afbeelding 3: Het multi-temperature datawarehouse.

data om te gaan ('multi-temperature datawarehouse') maken dit alles mogelijk. Teradata is al 25 jaar lang een door technologie gedreven club. Langzaam maken ze zich op om ook wat meer en beter te communiceren over wat ze nu eigenlijk te bieden hebben. Stel dat de TCO van Teradata over al deze componenten inderdaad lager is dan die van de concurrentie. Welke organisaties doen dit soort (zinnige) berekeningen?

Het ontbreken van ondersteunende TCO-informatie helpt dan niet. Voor Teradata is het belangrijk dat ze weet duidelijk te maken dat het TCO-verhaal dat ze houdt inderdaad klopt. En dat Teradata niet alleen interessant is voor de allergrootste organisaties. Deze boodschap helder overbrengen is de grootste uitdaging voor deze technologische club. Lukt dat, dan zal de technologische superioriteit van Teradata de datawarehouse-landkaart herschrijven. Lukt het ze niet, dan blijft de aanwezigheid beperkt tot de huidige categorie van high-end organisaties. Dat zou jammer zijn, maar is dan volledig toe te schrijven aan de eigen prestaties – en niet die van de concurrentie.

Paul van der Linden (Paul.PFH.vanderLinden@AtosOrigin.com) is senior consultant Data Warehousing/BI bij Atos Origin en geeft leiding aan Data Warehousing Cost & Lifecycle Management (CLM).