whitehorses

member of the human network

WHITEBOOK

Agile Business Intelligence met datavirtualisatie

Februari 2018

Auteur:
Maarten van Luijtelaar
INTEGRATIE SPECIALIST

Inleiding

Elke onderneming die zich bezighoudt met Business Intelligence kent het probleem: Hoe zorg ik ervoor dat data vanuit verschillende bronnen op een uniforme en gedocumenteerde manier beschikbaar wordt gesteld aan eindgebruikers? Hoe kan ik dit doen zonder dat er in de verschillende rapportagetools rekening gehouden moet worden met technische implementatiedetails? Hoe ontsluit ik mijn explosief gegroeide datavolumes? In dit Whitebook onderzoeken we wat datavirtualisatie is en hoe het kan helpen om sneller en goedkoper aan de informatiebehoefte te kunnen voldoen als volgende stap van het traditionele datawarehouse.

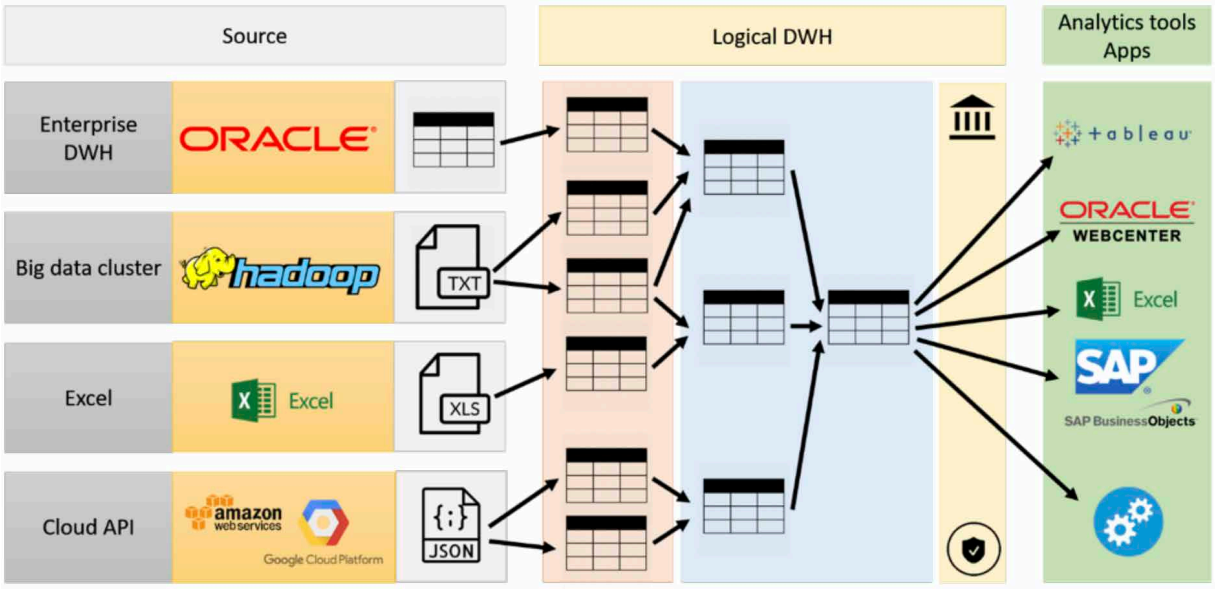
Problemen van het traditionele datawarehouse

Er was een tijd dat het enterprise datawarehouse voldoende capabel was om alle informatie binnen een organisatie te ontsluiten: data van verschillende systemen werd ingelezen, volgens strikte regels getransformeerd en gevalideerd, waarna de data beschikbaar gesteld werd via een diversiteit aan rapportage- en analysetools.
Maar in deze tijd, geregeerd door concepten als IOT, social media en cloud dataservices, ontstaat er een probleem. Door de explosieve volumegroei van data en diversiteit aan formaten of juist het gebrek aan structuur zal het op een bepaald moment zo goed als onmogelijk worden deze data nog op traditionele wijze in het datawarehouse te krijgen. Denk hierbij aan de tijd en resources die nodig zou zijn om al die data te verplaatsen vanuit de bron naar datamarts die vaak ondergebracht zijn in relationele databases. Deze relationele databases voldoen prima bij een beperkte hoeveelheid data die geanalyseerd moet worden, maar op een bepaald punt zal je het plafond bereiken waar scale-up van de infrastructuur (storage, database) erg duur gaat worden of zelfs onmogelijk is door softwarebeperkingen.

Het enterprise datawarehouse biedt een centrale bibliotheek van goed gestructureerde data. Doorgaans wordt er erg veel aandacht besteed aan de kwaliteit en betrouwbaarheid ervan waardoor implementaties vrij lang kunnen duren. Dit strookt niet altijd met de werkelijkheid waarin we nu leven. Door de enorme snelheid en het volume van beschikbare data en de behoefte om deze data te kunnen gebruiken om als bedrijf de juiste beslissingen te kunnen maken kunnen we simpelweg niet lang wachten om de competitie nog voor te zijn. Het zal niet de eerste keer zijn dat wanneer een rapportage eindelijk is opgeleverd de waarde ervan al niet meer relevant is.

Een logisch vervolg: het logisch datawarehouse

Door al het “Big Data” geweld wat op ons is afgekomen is een nieuwe BI-architectuur een noodzaak geworden. Het logisch datawarehouse is een architectuur waarbij het traditionele, batch georiënteerde enterprise datawarehouse gecombineerd wordt met functies die (near) real-time toegang tot zowel gestructureerde als ongestructureerde data en transformaties mogelijk maken. Het biedt mechanismes die abstracte “views” op verschillende databronnen mogelijk maakt, ongeacht of die data zich binnen of buiten de organisatie bevindt. Naast overkoepelende metadata en repository management is datavirtualisatie een belangrijke eigenschap van het logisch datawarehouse.



Wat is datavirtualisatie?

Datavirtualisatie geeft het logisch datawarehouse de mogelijkheden om geïntegreerde views te maken op data vanuit verschillende bronnen zonder die data te repliceren.
Let wel: virtualisatie heeft in deze context weinig te maken met de technische abstractie van hardware of applicaties, des te meer met abstractie van de data zelf. Datavirtualisatie is een term die gebruikt wordt om data aan te bieden aan eindgebruikers of applicaties zonder daarbij geconfronteerd te worden met technische implementatiedetails als gevolg van verschillende typen en dataformaten van de aangesproken bronsystemen.

Kenmerken van een datavirtualisatieplatform

In de huidige markt zijn er een aantal spelers die een datavirtualisatieplatform aanbieden zoals Denodo, Tibco en jBoss. Maar wat moet zo'n platform kunnen bieden?

- **Uniforme toegankelijkheid van data**
  Het verschaffen van één toegangspunt voor gestructureerde en ongestructureerde data. Een bibliotheek met data die op een eenduidige manier geconsumeerd kan worden door analysetools zoals Tableau en Excel of door andere applicaties voor verdere verwerking.

- **Datasecurity**
  De beveiliging van gegevens kan centraal geregeld worden in plaats van op het bronsysteem zelf.
  Denk hierbij niet alleen aan gebruikers en rollen, maar ook aan de mogelijkheden voor het definiëren van regels binnen de datasets zelf zoals het beperken van rijen of kolommen en het maskeren van gevoelige data.

- **Multi-source**
  Het vormt een brug tussen data vanuit een diversiteit aan onderliggende opslagtechnologieën zoals relationele databases, Hadoop clusters en (cloud) services op mogelijk verschillende fysieke locaties. Het platform moet een optimizer bevatten die vendor-specifieke optimalisaties kan benutten zoals query-pushdowns.

- **Multi-format**
  Het kunnen interpreteren van een verscheidenheid aan dataformaten zoals XML, JSON, CSV om bijvoorbeeld de response van een cloud dataservice om te kunnen vormen naar een waardevolle en gestructureerde dataset die waarde voor de business heeft.

- **Decoupling**
  Het kunnen loskoppelen van applicaties en analyses van de fysieke datastructuren om zo een flexiblele architectuur te waarborgen waarbij implementaties (andere technologieën) over tijd kunnen veranderen zonder al te veel impact voor gebruikers en systemen die data afnemen.

- **Governance**
  Het moet inzichtelijk zijn waar de data die beschikbaar wordt gesteld vandaan komt. Hoe zijn de attributen van de businessview uiteindelijk gemapped op de data in het bronsysteem? Welke kwaliteit heeft de data en welke normen hanteer je voordat deze data beschikbaar wordt gesteld? Hoe is de data geproduceerd? Het gebruik en management van metadata speelt hier een sleutelrol.

Aandachtspunten

Als je gaat starten met datavirtualisatie zijn er een aantal aspecten waar je rekening mee moet houden. Misschien zal een organisatie heel snel geneigd zijn om alle data die zich lokaal op de verschillende afdelingen bevindt virtueel beschikbaar te stellen. Dit leidt natuurlijk tot een bredere en snelle beschikbaarheid van die data, maar pas op dat er geen twijfel gaat ontstaan binnen de hele organisatie over de kwaliteit hiervan. Daarom moet een andere aanpak gekozen worden: over datavirtualisatie moet vanuit een bedrijfsbreed perspectief nagedacht worden. Waarschijnlijk gebeurt dit al voor het enterprise datawarehouse dus de gebruikte principes en methodieken kunnen ook nu gebruikt worden.

Waar eerder het datawarehouse het alleenrecht had op het publiceren van data zal het verantwoordelijke team tenminste kennis moeten hebben van de concepten en technische aspecten van het virtualiseren. Er is immers een verandering gaande: Het datawarehouse of datamarts worden niet meer rechtstreeks uitgevraagd door eindgebruikers maar door de virtualisatielaag!

Een ander aspect waar uiteraard rekening mee gehouden moet worden is performance. Het hele idee achter virtualisatie is dat er geen data meer verplaatst gaat worden vanuit de gebruikte bronnen. Dit kan voor bijvoorbeeld transactionele systemen (ERP) een probleem opleveren. Immers, één van de redenen om het enterprise datawarehouse op te zetten is juist het ontlasten van operationele of transactionele systemen en een geoptimaliseerde omgeving te creëren waar gebruik gemaakt kan worden van parallelle verwerking en bulktransformaties. Het lijkt mij dan ook niet zinvol of zelfs haalbaar om te proberen de datamodellen van operationele systemen op detailniveau te virtualiseren. De meeste datavirtualisatieoplossingen hebben wel een vorm van caching, maar dit moet met de nodige voorzichtigheid worden toegepast op data binnen het eigen domein vanwege mogelijke actualiteitsproblematiek. Daarnaast ben je met caching alsnog data aan het verplaatsen wat eigenlijk indruist tegen de pricipes van datavirtualisatie.

Conclusie

Om het logisch datawarehouse als architectuur te kunnen bereiken zal de ondersteuning van een datavirtualisatieplatform nodig zijn. Het grote verschil met een enterprise datawarehouse is dat bij een logisch warehouse niet meer wordt nagestreefd om alle data te consolideren in een enkele fysieke database voordat het gebruikt kan worden. Dat wil echter niet zeggen dat het enterprise datawarehouse in één keer overbodig gemaakt is. Sterker nog, data waarvan de kwaliteit en integriteit gewaarborgd moet worden of dusdanig complex is dat het niet real time te berekenen valt (bv. financiële rapportage, handelsposities of benchmarking) zal nog steeds geproduceerd worden door middel van het enterprise datawarehouse. Datzelfde geldt voor historische of tijdsconsistente datamarts, hoewel je wel zou kunnen overwegen deze data te verplaatsen van het datawarehouse naar bijvoorbeeld goedkopere opslag in een Hadoop cluster en daarna te combineren met actuele data in de virtualisatielaag. In mijn optiek is het kunnen virtualiseren van data juist een oplossing om het enterprise datawarehouse te ontlasten van allerlei technische integratievraagstukken. Zo kun je je meer richten op de toegankelijkheid van die data samen met data uit (externe) bronnen in de virtualisatielaag.