E-BOOK

Geschikte toepassingen van datavirtualisatie

Axians
Eemnesserweg 55
1251 NB Laren
Tel: +31 35 53 94 490
axians.nl/datavirtualisatie

Een wegwijzer voor het kiezen van een data-virtualisatie-server

Inhoud

1. Leeswijzer 3

2. Van datamaatschappij naar datavirtualisatie 4

3. Toepassingen van datavirtualisatie 7

Algemene voorwaarden voor datavirtualisatie **7**

Versnellen van ad-hoc bevragingen van data **8**

Stroomlijnen van self-service BI-omgevingen **10**

Moderniseren van een bestaande datawarehouse-omgeving **11**

Vervangen van bestandsuitwisseling door direct opvragen **14**

Ondersteunen van externe bevragingen voor data **17**

Organiseren van toegang tot externe data **19**

Ondersteunen van data-analyses voor data scientists **21**

Centraliseren van specificaties voor databeveiliging **23**

Centraliseren van specificaties voor data-privacy **24**

Migreren van data naar een ander dataplatform **26**

Migreren van data en applicaties naar een cloud-platform **28**

Beschikbaar stellen van masterdata aan veel gebruikers **28**

Combineren van toepassingen **29**

4. Een vergezicht voor het toepassen van datavirtualisatie 30

5. Wegwijzer voor datavirtualisatie 33

6. Algemene adviezen voor invoering van datavirtualisatie 34

1. Leeswijzer

INLEIDING

Dit eBook helpt organisaties zelfstandig te bepalen of datavirtualisatietechnologie voor hun toepassingen geschikt en toereikend is. De volgende toepassingen worden toegelicht:

- Versnellen van ad-hoc bevragingen van data
- Stroomlijnen van self-service BI-omgevingen
- Moderniseren van een bestaande datawarehouse-omgeving
- Vervangen van bestandsuitwisseling door direct opvragen
- Ondersteunen van externe bevragingen voor data
- Organiseren van de toegang tot externe data
- Ondersteunen van data-analyses voor data scientists
- Centraliseren van specificaties voor databeveiliging
- Centraliseren van specificaties voor data-privacy
- Migreren van data naar een ander dataplatform
- Migreren van data en applicaties naar een cloud-platform
- Beschikbaar stellen van masterdata aan veel gebruikers

HET BRONDOCUMENT

Dit eBook is afgeleid van een document geschreven door Rick F. van der Lans voor en onder leiding van CIO Rijk. Ten opzichte van dit originele document zijn enkele hoofdstukken niet in het eBook opgenomen en is de tekst enigszins veralgemeniseerd, omdat het originele document alleen tot overheidsorganisaties gericht is.

DOELGROEP

Tot de doelgroep van dit document behoren IT-architecten, solutions-architecten, data-architecten, business-gerichte professionals en managers, projectleiders en geïnteresseerden.

VEREISTE VOORKENNIS

Enige voorkennis van datavirtualisatie wordt verondersteld. Als deze kennis niet aanwezig is, dan wordt de volgende literatuur geadviseerd:

- R.F. van der Lans, *Data Virtualization for Business Intelligence Systems*, Morgan Kaufmann Publishers, 2012.
- R.F. van der Lans, *Data Virtualization: Selected Writings*, Lulu.com, September 2019; zie http://www.r20.nl/DataVirtualizationBook.htm
- R.F. van der Lans, Geschikte toepassingen van datavirtualisatie, een wegwijzer voor het kiezen van een datavirtualisatie-server, 2020.

DE STRUCTUUR VAN HET DOCUMENT

Het document bestaat uit de volgende hoofdstukken:

- Leeswijzer
- Inleiding: van datamaatschappij naar datavirtualisatie
- Toepassingen van datavirtualisatie
- Wegwijzer voor datavirtualisatie
- Algemene adviezen voor invoering van datavirtualisatie

DANKWOORD

De auteur wil CIO Rijk graag bedanken voor de toestemming die zij gegeven hebben om deze versie van het originele document af te leiden en op deze wijze te mogen publiceren. Download het originele document hier.

Tevens wil de auteur de volgende personen bedanken voor hun bijdragen middels hun input en/of het reviewen van het originele document: T. Faber (Directie CIO Rijk), M.L. de Vries (Directie CIO Rijk), A. van der Helm (Axians Business Analytics Laren), J. Wisgerhof (Axians Business Analytics Laren), E. Fransen (Connected Data Group), A. Stelma (Connected Data Group), G.P. Soet (Suithouse Consulting), R. de Kruyk (Suithouse Consulting) en P. Dozy (Vektis).

2. Van datamaatschappij naar datavirtualisatie

DE DATAMAATSCHAPPIJ

We leven in een maatschappij waarin steeds meer data geproduceerd wordt door een groeiend aantal partijen en waar dataconsumptie minstens in een vergelijkbaar tempo toeneemt. Data wordt geproduceerd en geconsumeerd door allerlei partijen, zoals Albert Heijn, het Centraal Bureau voor de Statistiek, de Rabobank, het Centraal Justitieel Incassobureau, de Kamer van Koophandel en alle beheerders van basisregistraties. Data wordt door hen geproduceerd voor intern en extern gebruik. Tevens produceren individuen data door het gebruik van mobiele telefoons, televisies, auto's en social media-apps. Zowel het aantal consumenten van data als de hoeveelheid data die verwerkt wordt, neemt met rasse schreden toe. We veranderen steeds meer in een *datamaatschappij*.

De toenemende consumptie en productie van data komt overeen met fenomenen als de *data-gedreven organisatie, digitale transformatie* en de *data-economie*. Deze termen betekenen niet hetzelfde, maar ze hebben één aspect gemeen, namelijk dat organisaties 'meer met data willen doen'. Ze willen data breder, effectiever en efficiënter inzetten om zo bedrijfs- en beslissingsprocessen op een of andere wijze te verbeteren of te versnellen.

DATASTROMEN

Door toenemende dataconsumptie, neemt het aantal *datastromen* fors toe. Binnen een organisatie 'stroomt' er data van het ene naar het andere IT-systeem, tussen organisaties vindt veel data-uitwisseling plaats en steeds meer gebruikers hebben toegang tot data nodig voor een steeds breder scala aan vormen van datagebruik. Bovendien krijgen meer klanten, patiënten, toeleveranciers, burgers en externe organisaties toegang tot data. Dit zijn allemaal datastromen.

Technisch gezien worden datastromen op veel verschillende manieren ingezet. Data kan bijvoorbeeld uitgewisseld worden via een *enterprise services bus* (ESB), data kan eerst naar een bestand geschreven, daarna verstuurd en vervolgens door een dataconsument ingelezen worden, data kan met een ETL-proces *(Extract Transform Load)* van de ene naar een andere database gekopieerd worden, data kan door een BI-product *(Business Intelligence)* met behulp van een query uit een database opgevraagd worden, data kan via e-mail verstuurd worden en data kan door een sensor naar een applicatie gepusht worden.

DATAPRODUCENTEN

Organisaties draaien grote hoeveelheden applicaties en systemen die data produceren en/of consumeren. Voorbeelden van *dataproducenten* zijn de systemen bij KLM waarmee alle vluchtreserveringen worden bijgehouden, de systemen van social media-bedrijven waarmee miljoenen nieuwe berichten per dag geregistreerd worden, de vele IT-systemen bij de Belastingdienst, de sensor-gebaseerde systemen van Rijkswaterstaat die het waterpeil, de golfhoogte en de waterkwaliteit registreren en de systemen die ziekenhuizen gebruiken voor het bijhouden van patiëntgegevens en waarmee röntgenfoto's gemaakt en opgeslagen worden. Door al deze systemen wordt nieuwe data geproduceerd.

DATACONSUMENTEN

Dataconsumenten zijn systemen die de data gebruiken, zoals rapporten, dashboards, datascience-modellen, portals en websites. Veel dataproducerende systemen zijn echter ook dataconsumenten. Bijvoorbeeld, basisregistraties van de overheid consumeren data, aangeleverd door andere organisaties. Dataconsumenten kunnen door mensen bediend worden, maar ze kunnen ook volautomatisch zijn, zoals applicaties die in geavanceerde machines werken en automatisch beslissingen nemen.

DATABRONNEN

Naast de vele datastromen bestaan er ook talloze *databronnen*. Databronnen zijn systemen waar data opgeslagen ligt. Dit kan *originele data* zijn die door een

dataproducent geproduceerd is of gekopieerde en mogelijk bewerkte data. Dit wordt in dit document *afgeleide data* genoemd. Technisch gezien kan een databron allerlei vormen hebben, zoals een database, een simpel bestand of een spreadsheet. Net zoals het aantal datastromen toeneemt, neemt ook het aantal databronnen toe en daarmee ook het aantal bronnen met afgeleide data.

Wanneer de ene organisatie data naar een andere stuurt, slaat laatstgenoemde deze afgeleide data bijna altijd op. Ook binnen organisaties wordt data veelvuldig opgeslagen. In een datawarehouse-omgeving wordt bijvoorbeeld voornamelijk afgeleide data opgeslagen. Zo wordt de data die in deze omgeving binnenkomt gekopieerd vanuit productiesystemen en ontstaat er afgeleide data. Daarnaast wordt data meerdere malen opgeslagen binnen de datawarehouse-omgeving, zoals in het staging area, het centrale datawarehouse en soms meerdere keren in verschillende datamarts. Een ander voorbeeld is dat gebruikers veelvuldig met copy-paste data van het scherm naar een bestand kopiëren waarmee weer een afgeleide databron ontstaat.

DE GEBREKKIGE DATAMAATSCHAPPIJ

Als alle datastromen en databronnen van alle commerciële en overheidsorganisaties in kaart gebracht zouden (kunnen) worden, ontstaat een indrukwekkend netwerk dat een complexiteit heeft waarbij de Gordiaanse knoop in het niet valt. Dit netwerk vertoont in praktijk veel problemen. Sommige systemen zijn inflexibel, kostbaar, traag, verouderd, ontoereikend, lastig in gebruik, enzovoorts. Dit leidt bij veel organisaties tot problemen in de datahuishouding.

Enkele bekende voorbeelden:

- Binnen sommige organisaties is data beschikbaar, maar de gebruikers weten niet dat die data aanwezig is, terwijl ze die wel goed zouden kunnen gebruiken.
- Soms weten gebruikers dat data beschikbaar is, maar kunnen of mogen die om bepaalde redenen niet gebruiken.
- Het ontwikkelen van nieuwe rapporten, dashboards en analyses kan lang duren vanwege de inflexibiliteit van bestaande systemen bij de levering van data.
- Het ongeremd opslaan van afgeleide data is een voedingsbodem voor fouten en inconsistenties. Als originele data naar meerdere databronnen wordt gekopieerd, moet gezorgd worden dat al deze afgeleide data aan elkaar gelijk blijft. Zo niet, dan leidt dit tot inconsistente databronnen en vervolgens tot inconsistente rapportresultaten.
- Met elke kopieerslag verhoogt de datalatentie en veroudert de data die uiteindelijk incorrect wordt. Steeds meer gebruikers vragen om real-time of near-real-time data.
- Bij veel afgeleide data hebben de originele dataproducenten steeds minder controle over wie welke data gebruikt en dataconsumenten hebben steeds minder zicht op waar de data eigenlijk vandaan komt.
- Bij veel afgeleide databronnen wordt het complexer om aan het vergeetrecht[1] en andere aspecten van de *Algemene Verordening Gegevensbescherming* (AVG) te voldoen; zie paragraaf 3.5. Het vergeetrecht is het recht om te eisen dat persoonsgegevens worden verwijderd en dat verdere verspreiding wordt tegengegaan.
- Externe bevragingen voor data kunnen lang duren om te implementeren; soms ontbreekt de IT-architectuur hiervoor volledig.
- Er vindt veel enigszins ouderwetse bestands-uitwisseling tussen organisaties plaats, terwijl dat veel eenvoudiger met moderne technologie geregeld kan worden.

1 Justitia.nl, *Vergeetrecht - Right to be forgotten*; juni 2020 zie https://www.justitia.nl/privacy/vergeetrecht

De gewenste oplossing voor deze problemen is niet het verminderen van dataproductie en dataconsumptie. Integendeel: toegang hebben tot geschikte data kan voordelen bieden, maar de datahuishouding moet wel verbeteren. Dit is zeker belangrijk in een tijd waarin veel waarde aan data-privacy wordt gehecht en er daarom zorgvuldig en verantwoord met data omgegaan moet worden, zeker als het de AVG betreft.

DATAVIRTUALISATIE IN EEN NOTENDOP

Een technologie die kan helpen met het verbeteren van de datahuishouding is *datavirtualisatie*. Datavirtualisatietechnologie staat toe IT-systemen te ontwikkelen waarbij dataconsumenten en dataproducenten ontkoppeld worden. Deze ontkoppeling biedt de volgende voordelen:

- Datastromen kunnen efficiënter opgezet worden.
- Redundante opslag van data kan verminderd worden.
- Data kan sneller en flexibeler ter beschikking komen van dataconsumenten.
- Voldoen aan de nieuwe regelgevingen voor data-privacy en -beveiliging is eenvoudiger.

Het laatstgenoemde voordeel is belangrijk, omdat het onjuist vrijkomen van data schade kan veroorzaken bij organisaties, burgers, aan kritische bedrijfsprocessen en vitale processen voor de samenleving. Schade kan hierbij allerlei vormen aannemen, waaronder financiële-, politieke-, identiteits-, gezondheids-, imago- en materiële schade.

De technologie kent ook veel toepassingen waarvan de volgende in dit eBook beschreven zijn:

- Versnellen van ad-hoc bevragingen van data
- Stroomlijnen van self-service BI-omgevingen
- Moderniseren van een bestaande datawarehouse-omgeving
- Vervangen van bestandsuitwisseling door direct opvragen
- Ondersteunen van externe bevragingen voor data
- Organiseren van de toegang tot externe data
- Ondersteunen van data-analyses voor data scientists
- Centraliseren van specificaties voor databeveiliging
- Centraliseren van specificaties voor data-privacy
- Migreren van data naar een ander dataplatform
- Migreren van data en applicaties naar een cloud-platform
- Beschikbaar stellen van masterdata aan veel gebruikers

3. Toepassingen van datavirtualisatie

Dit hoofdstuk beschrijft geschikte toepassingen voor datavirtualisatie. Het hoofdstuk is gericht op toepassingen die momenteel binnen organisaties populair zijn. Deze staan niet op volgorde van belangrijkheid, want dat is sterk afhankelijk van de organisatie.

Naast een beschrijving worden tevens de voorwaarden van elk toepassingsgebied genoemd. Omdat sommige voorwaarden voor elke toepassing gelden, worden deze in de volgende paragraaf afzonderlijk beschreven.

3.1 ALGEMENE VOORWAARDEN VOOR DATAVIRTUALISATIE

Bij elke toepassing zijn de voorwaarden vermeld die gelden voor die specifieke toepassing. Deze paragraaf bevat de algemene voorwaarden die voor elke toepassing van datavirtualisatie gelden.

Kosten en baten

De vaste en variabele kosten van een datavirtualisatie-server moeten op een simpele wijze terug te verdienen zijn. Mogelijke baten van een datavirtualisatie-server kunnen zijn: toekomstige productiviteitswinst, vereenvoudigd onderhoud, verkorte time-to-market van aanvragen voor data, minder dataduplicatie en het wegvallen van de kosten van andere software producten. De eerste twee argumenten zijn lastig in geld uit te drukken. Een puur financiële verantwoording berust derhalve op veel aannames.

Visie

Het is niet verstandig om een datavirtualisatie-server in te zetten voor een 'klein' project om 'eventjes' data-integratie te implementeren. De organisatie behoort een visie te hebben betreffende datavirtualisatie op basis waarvan voor deze technologie gekozen wordt. In die visie moet duidelijk aangegeven worden dat data-abstractie als een groot voordeel gezien wordt om dataconsumenten te scheiden van databronnen en zo een flexibelere data-architectuur te realiseren.

Dagelijks beheer

Een datavirtualisatie-server dient net als andere serverproducten, zoals databaseservers en enterprise service bussen, beheerd te worden. Een functie moet ingericht worden voor het dagelijkse beheer en de algemene performance van de datavirtualisatie-server. Tevens is deze ook verantwoordelijk voor het installeren van nieuwe versies en upgrades en het opzetten en beheren van mogelijkerwijs een ontwikkel-, test-, acceptatie- en productieomgeving.

Beheer van views

Vanwege het gemak en de snelheid waarmee ontwikkeld kan worden, kan een datavirtualisatie-server snel een rommelige en inefficiënte omgeving worden, waardoor de beoogde voordelen niet behaald worden. Een belangrijke voorwaarde is dat een functie de definities van de views beheert. Dit moet ervoor zorgen dat views niet te veel overlap hebben en zoveel mogelijk hergebruikt worden.

Ontwerpregels voor views

Er moeten algemene ontwerpregels gedefinieerd worden voor de opbouw van views. Welke dataverwerkingsspecificaties dienen bijvoorbeeld in welke view-laag opgenomen te worden, hoe en waar moeten beveiligingsregels opgenomen worden en moeten de views in de transformatielaag genormaliseerd zijn of niet? Deze ontwerpregels moeten gedefinieerd, continu uitgebreid en verbeterd worden en nieuwe views moeten hierop gecontroleerd worden. Een functie moet hiervoor geïntroduceerd worden.

3.2 VERSNELLEN VAN AD-HOC BEVRAGINGEN VAN DATA

Inleiding ad-hoc bevragingen

De behoefte om onvoorzien en ongepland data op te vragen neemt toe. Deze zogenaamde *ad-hoc bevragingen* kunnen ontstaan door urgente ontwikkelingen, calamiteiten of mogelijke kansen.

De standaardoplossing voor ah-hoc bevragingen

Wat alle ad-hoc bevragingen gemeen hebben, is dat ze meestal uniek en specifiek zijn. Voor dit soort bevragingen moet mogelijkerwijs data uit meerdere databronnen samengevoegd worden, de data moet zwaar bewerkt en geaggregeerd worden en er moeten complexe opschoningsoperaties uitgevoerd worden. Soms moet de data tevens verrijkt worden met data van externe partijen.

De gebruikelijke oplossing voor ad-hoc bevragingen is om *quick-and-dirty* een compacte en *throw-away* oplossing te ontwikkelen, met de intentie om zo snel mogelijk een resultaat te leveren. Het doel is niet om een oplossing te ontwikkelen die jaren mee kan en door veel dataconsumenten herhaaldelijk gebruikt kan worden.

De ontwikkeling van de meeste quick-and-dirty oplossingen geschiedt als volgt: Aan de eigenaren van de databronnen worden extracties gevraagd van data die waarschijnlijk nodig is. Na ontvangst wordt de benodigde brondata (eventueel dus van externe partijen) verzameld en tijdelijk opgeslagen (in bestanden, spreadsheets of databases). Vervolgens wordt de data geïntegreerd en worden de benodigde verwerkingsslagen uitgevoerd. Dit wordt meestal met producten uitgevoerd die voorhanden zijn of waarmee de ontwikkelaars ervaring hebben. Het resultaat wordt dan fysiek in een database of bestand opgeslagen. Tenslotte wordt dit bestand benaderd met een product om de data te visualiseren en rapporten te produceren die naar de aanvrager teruggestuurd kunnen worden. In het ongunstigste geval moeten eerder verkregen data-extracten ververst worden. Dit leidt tot extra werk.

Nadelen van standaardoplossingen voor ah-hoc bevragingen

- Uit de databronnen wordt meestal te veel data opgehaald. De reden hiervoor kan zijn dat er selecties moeten plaatsvinden op karakteristieken die op het moment van aanvragen nog niet bekend zijn (gedreven door de vrees voor extra werk), maar er kunnen ook technische redenen bestaan of het kan zijn dat het minder werk is voor de eigenaar om alle data te geven.
- De opgehaalde data wordt redundant opgeslagen.
- Als het ontwikkelproces lang duurt, raakt de data langzaamaan verouderd. Om de data weer up-to-date te krijgen, moet deze opnieuw verzameld worden en de oplossing opnieuw uitgevoerd worden.
- De opgehaalde data dient vanaf ontvangst beveiligd te worden. Dit geldt zeker wanneer dit gevoelige persoonsgegevens betreft. In feite moeten tussenresultaten verwijderd worden en hierop moet toezicht gehouden worden. De tijdelijk opgeslagen data is meestal niet in een goed beveiligde omgeving opgeslagen, maar bijvoorbeeld op de laptop van een ontwikkelaar.
- De meeste oplossingen voor ad-hoc bevragingen worden niet hergebruikt. Vaak beginnen ontwikkelaars bij elke ad-hoc vraag weer opnieuw.
- Er is geen garantie dat alle data die voor de ontwikkeling van de ad-hoc bevraging opgeslagen wordt, vernietigd wordt. De grip op de data wordt verloren en er is geen controle op rapportresultaten.

Ad-hoc bevragingen met datavirtualisatie

Een ad-hoc vraag kan ook door middel van datavirtualisatie opgelost worden. In dit geval worden er views gedefinieerd die een connectie met de databronnen creëren. Daarna worden views gedefinieerd die de integratie en alle dataverwerkingsspecificaties uitvoeren. De ontwikkelaar richt zich daarom voornamelijk op de

ontwikkeling van dataverwerkingsspecificaties, dus hoe de data bewerkt moet worden om het gewenste resultaat te krijgen en veel minder op het lokaal opslaan van de data die vervolgens beveiligd moet worden.

Voordelen van datavirtualisatie voor ad-hoc bevragingen

- Meer specificaties kunnen hergebruikt worden. Als er bijvoorbeeld al connecties bestaan met een bepaalde databron vanwege voorgaande ad-hoc bevragingen, dan kunnen deze opnieuw gebruikt worden. Hetzelfde geldt voor allerlei dataverwerkingsspecificaties die steeds vaker hergebruikt kunnen worden. In feite ontstaat geleidelijk een hele verzameling views die toegang biedt tot diverse interne en externe databronnen, waardoor er steeds meer hergebruikt kan worden en de ontwikkelsnelheid van ad-hoc bevragingen verhoogt.
- Datavirtualisatie-servers proberen altijd de minimale hoeveelheid data uit de databronnen op te halen. In feite zijn deze producten daarvoor geoptimaliseerd.
- Als tijdens het ontwikkelen en testen bepaalde databronnen herhaaldelijk benaderd moeten worden, kan dat tot verstoringen leiden. In dit geval kunnen de views tijdelijk gecached worden. Als de oplossing goed werkt, kunnen de caches verwijderd worden om uiteindelijk met de meest actuele data te werken.
- Als de oplossing volledig virtueel uitgevoerd kan worden, wordt er helemaal geen data redundant opgeslagen waardoor deze ook niet beveiligd en later vernietigd dient te worden. Tussentijdse resultaten worden niet redundant opgeslagen.

Voorwaarden

- Er moeten voldoende ad-hoc bevragingen zijn om de investering in een datavirtualisatie-server te rechtvaardigen.
- Ontwikkelaars moeten bereid zijn bepaalde aspecten van hun oplossingen niet langer met hun eigen BI-product te ontwikkelen, maar te definiëren met een datavirtualisatie-server.
- De organisatie moet de bedrijfsvoordelen en het belang inzien van het versnellen van de ontwikkeling van ad-hoc rapporten.
- Het moet mogelijk zijn om toestemming te regelen voor het opvragen van data uit interne en externe databronnen.
- Er moeten afspraken gemaakt kunnen worden over het beveiligen, pseudonimiseren en anonimiseren van de data en de doelbinding.

3.3 STROOMLIJNEN VAN SELF-SERVICE BI-OMGEVINGEN

Inleiding self-service BI

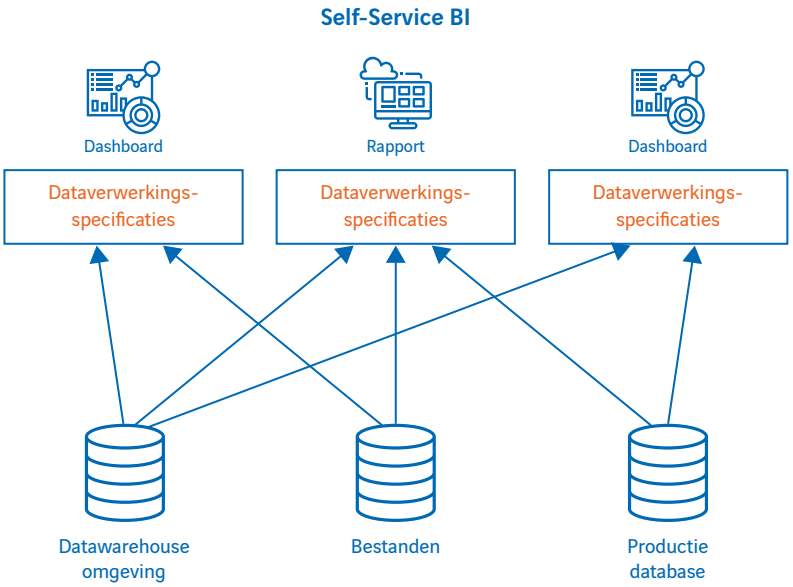
Nagenoeg elke organisatie ondersteunt tegenwoordig self-service BI. Met self-service BI kunnen gebruikers buiten de IT-afdeling om zelf rapporten en dashboards ontwikkelen en data analyseren. Populaire producten die hiervoor gebruikt worden zijn onder andere QlikSense, Microsoft PowerBI, Tableau en ook Excel. Zoals de naam al aangeeft, definiëren gebruikers bij self-service BI zelf de dataverwerkingsspecificaties; zie figuur 1. Self-service BI ontlast de IT-afdeling en vergroot de ontwikkelcapaciteit van organisaties voor rapporten en dashboards. Zeker wanneer deze gebruikers toegang hebben tot een datawarehouse-omgeving, biedt dit veel praktische voordelen. In dit geval hebben de data engineers de data al geïntegreerd en voor consumptie voorbereid, ofwel hebben zij al veel dataverwerkingsspecificaties geïmplementeerd.

Nadelen van self-service BI

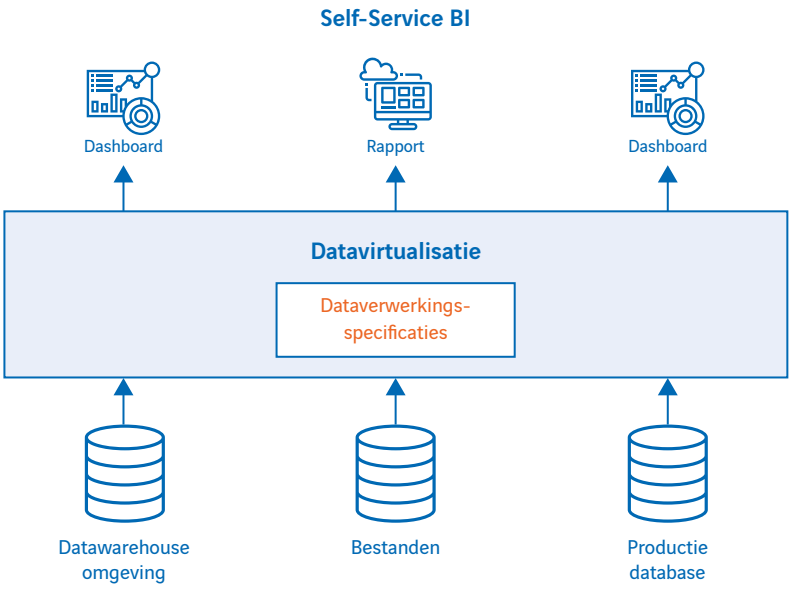
- Karakteristiek voor veel self-service BI-initiatieven is dat ze leiden tot een wildgroei aan rapportdefinities. Iedere gebruiker ontwikkelt zijn of haar eigen rapporten en daarmee ook dataverwerkingsspecificaties. Omdat veel rapporten uiteindelijk toch dezelfde data verwerken, maar alleen op een andere wijze presenteren, ontstaat er veel duplicatie van specificaties. Elke gebruiker ontwikkelt zijn eigen 'eiland met specificaties'. Het nadeel is een lage productiviteit, omdat elke gebruiker het wiel opnieuw uitvindt bij het definiëren van dataverwerkingsspecificaties en er weinig tot geen oplossingen hergebruikt worden.
- Het onderhoud van alle specificaties is complex. Een berekening voor afvalscheiding kan bijvoorbeeld in verschillende BI-rapporten opgenomen zijn, maar op verschillende manieren en met verschillende talen en producten geïmplementeerd zijn. Uitzoeken waar al die berekeningen geïmplementeerd zijn, kan op zich al moeilijk en tijdrovend zijn.
- Er kan niet gegarandeerd worden dat bepaalde dataverwerkingsspecificaties overal consistent geïmplementeerd zijn. Het gevaar hiervan is dat inconsistente implementaties tot inconsistente rapportresultaten leiden.

Self-service BI met datavirtualisatie

Door toepassing van datavirtualisatie kan een self-service BI-omgeving naar een meer *managed self-service BI-omgeving* omgezet worden. Dit betekent dat veel dataverwerkingsspecificaties die normaliter binnen de BI-producten gedefinieerd worden, nu binnen de datavirtualisatie-server gedefinieerd worden; zie figuur 2.



Figuur 1: bij self-service BI worden dataverwerkingsspecificaties gedefinieerd in de applicaties en niet centraal.



Figuur 2: met datavirtualisatie kunnen dataverwerkingsspecificaties eenmalig gedefinieerd worden en door alle self-service BI-applicaties gebruikt worden.

In een ideale situatie worden alleen nog de visualisaties binnen de BI-producten gedefinieerd.

Voordelen van datavirtualisatie voor self-service BI

- Specificaties die normaal in BI-producten herhaaldelijk gedefinieerd worden, worden nu eenmalig in de datavirtualisatie-server vastgelegd. Elk BI-product, ongeacht het merk, maakt daar vervolgens gebruik van. Deze aanpak verbetert rapportconsistentie (elk rapport hanteert dezelfde formule), productiviteit (het wiel wordt niet telkens opnieuw uitgevonden) en onderhoudbaarheid (omdat elke formule maar één keer gedefinieerd is).
- Het hergebruiken van geteste dataverwerkings-specificaties betekent dat er voor nieuwe rapporten minder nieuwe specificaties ontwikkeld kunnen worden en dat verkleint de kans op incorrecte rapportresultaten.
- Alle regels voor beveiliging en privacy kunnen centraal vastgelegd en afgedwongen worden.

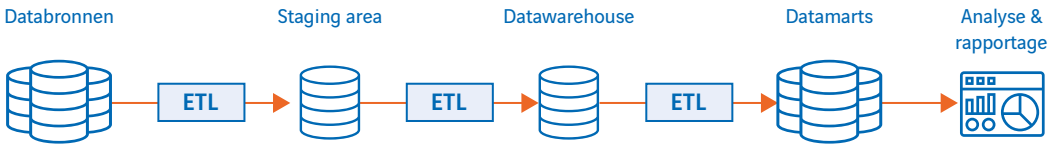
Voorwaarden

- Er moet voldoende heterogeniteit zijn van in gebruik zijnde BI-producten waarmee inderdaad dezelfde dataverwerkingsspecificaties ontwikkeld zijn en dus dezelfde databronnen benaderd worden. Als er tussen de dataverwerkingsspecificaties nagenoeg geen overlap bestaat, dan is het probleem te 'klein' om een datavirtualisatie-server te rechtvaardigen.
- Er moet een centrale functie opgezet kunnen worden die verantwoordelijk is voor het beheer van views. Deze functie werkt tevens als vraagbaak voor alle gebruikers van BI-producten.
- Gebruikers moeten bereid zijn om bepaalde aspecten van hun rapporten niet meer zelf in hun eigen BI-product te ontwikkelen, maar binnen de datavirtualisatie-server.
- Er moet reeds een achterstand zijn voor het ontwikkelen van rapporten waarvan de organisatie de pijn voelt.
- De organisatie moet de bedrijfsvoordelen van het versnellen van self-service BI-omgevingen onderkennen.

3.4 MODERNISEREN VAN EEN BESTAANDE DATAWAREHOUSE-OMGEVING

Inleiding datawarehouses

Menig organisatie heeft reeds een datawarehouse-omgeving in gebruik. Veel van deze omgevingen hebben een klassieke architectuur, bestaande uit een keten van databases, zoals een *staging area*, een centraal datawarehouse en verscheidene *datamarts*; zie figuur 3.



Figuur 3: de klassieke datawarehouse-architectuur.

Met behulp van ETL-programma's wordt data van de ene naar de andere database gekopieerd. Typerend voor deze architectuur is dat het een goed gecontroleerde omgeving is met veel procedures om te zorgen dat de data die beschikbaar gemaakt wordt voor analyses en rapportage correct, compleet en betrouwbaar is.

Nadelen van de klassieke datawarehouse-architectuur

Klassieke datawarehouse-omgevingen hebben veel organisaties al vele jaren gediend en hebben veel dataconsumenten voorzien van de juiste data. Er zijn echter nadelen aan deze architectuur, waaronder:

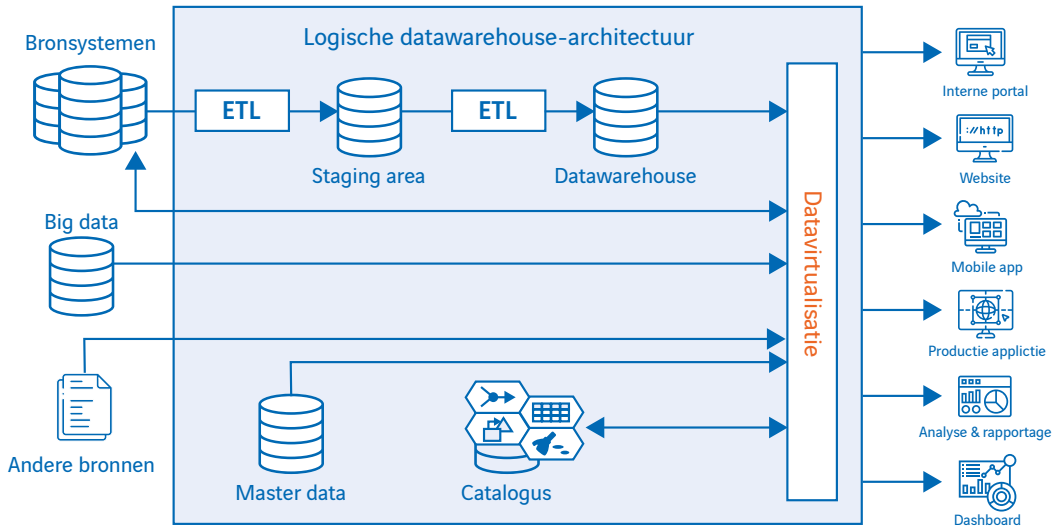
- Deze omgevingen bieden beperkte flexibiliteit. Het wijzigen of uitbreiden kan tijdrovend zijn. Een wijziging van een rapport kan bijvoorbeeld leiden tot veel wijzigingen in de rest van de architectuur en een wijziging van de definitie van een data-element kan leiden tot een waterval aan wijzigingen.
- In deze architectuur wordt veel data gedupliceerd. Dataduplicatie is een voedingsbodem voor fouten en inconsistenties in de data en maakt het lastiger om aan onder andere het vergeetrecht en andere aspecten van AVG te voldoen; zie paragraaf 3.5.
- Aangezien het gehele kopieerproces tijdrovend kan zijn, kan het lang duren voordat nieuwe data, die in de databronnen ingevoerd is, beschikbaar komt voor rapportage en analyse. De datalatentie is hoog en daardoor is er geen of slechts beperkte ondersteuning voor (near) real-time BI.
- Het is een forse uitdaging om, als de databron qua volume een big databron is, al deze data door de gehele architectuur van de ene naar de andere databases te kopiëren. Zo wordt er wel eens gesteld: "big data can be too big to move."
- Er is beperkte ondersteuning voor ad-hoc bevragingen van data. Het ontwikkelen duurt meestal te lang, omdat tijdens de ontwikkeling aan alle geldende procedures en regels voldaan moet worden, terwijl deze regels niet altijd relevant zijn voor de ad-hoc bevragingen.
- Als een databron weinig of geen ondersteuning biedt om data bulksgewijs te benaderen, zoals bij cloud-applicaties als Salesforce.com en AFAS, is het lastig die data op te nemen in een klassieke datawarehouse-omgeving.
- Deze architectuur biedt geen of weinig ondersteuning voor streaming data.

Datavirtualisatie en de logische datawarehouse-architectuur

Vanwege de beperkingen zijn veel organisaties op zoek gegaan naar nieuwe, flexibelere architecturen en naar mogelijkheden om de bestaande architectuur te verbeteren. Een architectuur die hieraan voldoet is de zogenaamde *logische datawarehouse-architectuur*. De naam voor deze architectuur is door Gartner[2] geïntroduceerd.

De logische datawarehouse-architectuur is een flexibele architectuur voor het ontwikkelen van BI-systemen, waarbij dataconsumenten en databronnen van elkaar losgekoppeld zijn; zie figuur 4. De logische datawarehouse-architectuur presenteert alle data, die in een heterogene verzameling van databronnen opgeslagen is, als één logische database. In deze architectuur is het voor dataconsumenten niet nodig te weten waar en hoe data opgeslagen wordt; alle details van databronnen zijn voor hen verborgen. Zij hoeven niet te weten of de data die ze gebruiken afkomstig is van een datamart, een datawarehouse, een big databron of misschien wel een bronsysteem, zij horen niet te zien dat data uit meerdere databronnen samengevoegd moeten

2 M. Beyer, Father of the Logical Data Warehouse, November 2011; see https://blogs.gartner.com/merv-adrian/2011/11/03/mark-beyer-father-of-the-logical-data-warehouse-guest-post/



Figuur 4: een overzicht van de minimale componenten van een logische datawarehouse-architectuur waarbij datavirtualisatie voor de ontkoppeling zorgt.

worden en ze horen ook niet te weten of ze een SQL-database, een Hadoop-cluster, een No-SQL-database, een service of gewoonweg één of meer platte bestanden benaderen. De structuur van de databron is tevens verborgen; dataconsumenten zien de data uitsluitend op een manier die voor hen nuttig is en alleen die data die relevant voor hun taak is. Dit wordt allemaal bereikt door dataconsumenten los te koppelen van dataproducenten. Kortom, data-abstractie is een belangrijke eigenschap van de logische datawarehouse-architectuur.

De technologie die het best past bij een logische datawarehouse-architectuur is uiteraard datavirtualisatie. Maar let wel: de logische datawarehouse-architectuur is niet slechts de inkapseling van de klassieke architectuur met een datavirtualisatie-server. Er zijn veel meer componenten nodig om een complete architectuur te ontwikkelen.

Een van deze componenten is een fysiek datawarehouse. Een logische datawarehouse-architectuur bevat er nagenoeg altijd een. In deze architectuur is het fysieke datawarehouse slechts één van de databronnen die door de datavirtualisatie-server beschikbaar gesteld wordt aan dataconsumenten. Het zou niet logisch zijn om die investering weg te gooien; het is goed in waar het voor bedoeld is. Dit fysieke datawarehouse is vaak nodig om bepaalde problemen op te lossen. Wanneer een databron bijvoorbeeld de datahistorie niet bijhoudt, is een fysiek datawarehouse noodzakelijk voor de opslag van de historie. Een ander probleem kan zijn dat de databron de extra werkdruk die door de dataconsumenten gegenereerd wordt niet aankan. Een fysiek datawarehouse kan dan helpen om de werklast te verlichten.

De populariteit van deze architectuur blijkt bijvoorbeeld uit dit citaat van Gartner[3]: "The logical data warehouse (LDW) is now a best practice analytics data management architecture and design that combines the strengths of traditional repository warehouses with alternative data management and access strategies. It has replaced the prior generation of Kimball/Inmon hybrid warehouse in leading organizations and is effective for cloud deployments. Position and Adoption Speed Justification: Beginning in 2018, the market adoption rate of the LDW has increased to more than 15% of all data warehouse deployments (and in one Gartner market analysis dataset has reached 17%)."

3 Gartner, Hype Cycle for Data Management, 2018, juli 2018.

Voordelen van de logische datawarehouse-architectuur

- Door het virtualiseren van bestaande fysieke onderdelen kan er sneller ontwikkeld worden en kunnen bestaande rapporten sneller gewijzigd worden. Het ontwikkelen van nieuwe rapporten verloopt sneller, omdat het alleen gaat om het definiëren van dataverwerkingsspecificaties, niet om het opbouwen en beheren van fysieke databases.
- Het is eenvoudiger om andere databronnen aan te koppelen en beschikbaar te maken voor rapportage. Dit geldt zeker voor big datasystemen die gewoonweg te veel data bevatten om naar het datawarehouse te kopiëren.
- Metadata kan in een datavirtualisatie-server gedefinieerd waarmee dataconsumenten beter kunnen begrijpen wat de betekenis van bepaalde elementen is en het geeft ze de mogelijkheid om data te zoeken.
- Externe databronnen kunnen aangekoppeld worden. Hierdoor kan het gebruik van interne data vereenvoudigd worden en kan interne en externe data eenvoudig door dataconsumenten geïntegreerd worden.
- Het migreren naar nieuwe dataplatformen is eenvoudiger. Dit wordt door de datavirtualisatie-server volledig afgeschermd zonder dat er aanpassingen nodig zijn. Een migratie kan gewenst zijn als de bestaande technologie te langzaam of te kostbaar is. Zie hiervoor ook paragraaf 3.11.
- Nieuwe vormen van dataconsumptie kunnen sneller ondersteund worden.
- De kosten van de bestaande datawarehouse-omgeving kunnen verlaagd worden, omdat bepaalde componenten weggelaten kunnen worden, zoals de datamarts. Als deze nodig zijn, kunnen ze met behulp van views virtueel gesimuleerd worden.
- Toegang tot en beveiliging van de data kan centraal vastgelegd en geregeld worden.

Voorwaarden

Er moet een klassieke datawarehouse-architectuur aanwezig zijn die één of meer van de volgende beperkingen kent:

- Bepaalde gebruiksvormen van data kunnen niet (efficiënt) ondersteund worden.
- De vereiste datalatentie kan niet geboden worden.
- De kosten van de huidige datawarehouse-omgeving zijn te hoog.
- De achterstand bij het ontwikkelen van rapporten moet verkleind worden.
- De hoeveelheid redundant opgeslagen data moet gereduceerd worden. Door middel van een logische datawarehouse-architectuur kunnen sommige databases, zoals datamarts, mogelijkerwijs verwijderd en virtueel opgezet worden.
- Externe data kan niet goed aan de datawarehouse-omgeving toegevoegd worden.
- Big data kan niet met de data van de datawarehouse-omgeving geïntegreerd worden.
- De in gebruik zijnde database-server moet vervangen worden.

3.5 VERVANGEN VAN BESTANDSUITWISSELING DOOR DATA-ON-DEMAND

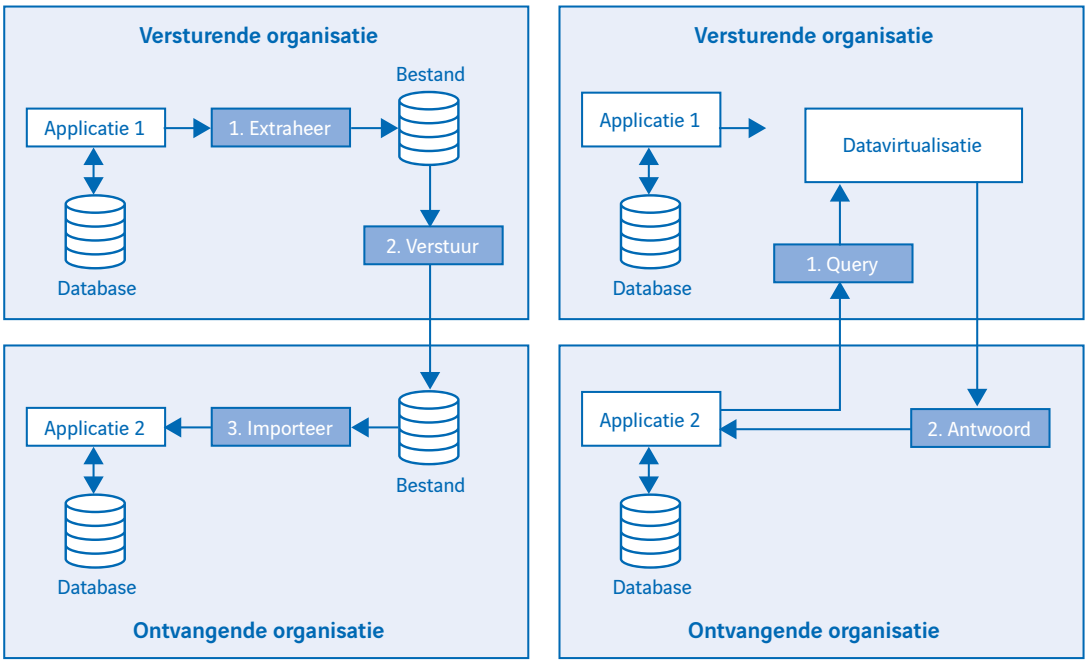
Inleiding bestandsuitwisseling

Tussen organisaties vindt veel data-uitwisseling plaats. Data wordt bijvoorbeeld uitgewisseld tussen organisaties en hun toeleveranciers, tussen commerciële organisaties en het Centraal Bureau voor de Statistiek en de Kamer van Koophandel, tussen de Belastingdienst en het UWV.

Veelal vindt deze uitwisseling plaats door middel van het versturen van bestanden. De versturende organisatie maakt een bestand aan met data die uit bepaalde systemen geëxtraheerd wordt. Dit bestand wordt naar de andere organisatie verstuurd met geavanceerde managed file transfer software of via iets simpels als e-mail. De ontvangende organisatie pakt het bestand op en importeert en verwerkt de data; zie het linker diagram in figuur 5. Meestal vindt bestandsuitwisseling batch-gewijs plaats. Dit houdt in dat het op gezette tijden plaatsvindt.

Globaal zijn er twee manieren om bestanden aan te maken. Bij de eerste manier bevat het bestand de data die al eerder gestuurd is plus alle nieuwe data, dus het complete bestand. Bij de tweede manier worden alleen de wijzigingen gestuurd. Als het om heel veel data gaat, is de tweede manier mogelijk efficiënter qua transmissiesnelheid.

Opmerking: Het uitwisselen van data tussen organisaties door middel van bestandsuitwisseling stamt nog uit de oertijd van IT toen er nog geen netwerken bestonden.



Figuur 5: links is aangegeven hoe bij bestandsuitwisseling data wordt uitgewisseld en rechts staat de oplossing met datavirtualisatie.

Nadelen van bestandsuitwisseling

In principe kan deze aanpak goed functioneren, maar kent wel de volgende nadelen en beperkingen:

- Voor de uiteindelijke dataconsumenten is de datalatentie hoog en het aanbieden van (near) real-time analyses is hierdoor niet mogelijk. Er zit altijd een vertraging tussen het moment waarop het bestand aangemaakt wordt en het moment waarop de data beschikbaar komt voor dataconsumenten.
- Om diverse redenen kan het proces van extraheren, versturen en importeren fout lopen. Bekende problemen bij bestandsuitwisseling zijn: er is iets misgegaan bij het aanmaken van het bestand, er is vergeten het bestand aan te maken waardoor er niets uitgewisseld kan worden, het bestand is wel verstuurd maar niet ontvangen en het formaat van de data in het bestand komt niet overeen met wat de ontvangende partij verwacht. De consequentie is dat als een dergelijk probleem zich voordoet, de datalatentie hoger wordt.
- Als het bestand alle data bevat, moet de ontvangende organisatie over een mechanisme beschikken dat uitzoekt wat er gewijzigd is ten opzichte van de huidige data. Als het bestand alleen de wijzigingen ten opzichte van de vorige keer bevat, moet een mechanisme ontwikkeld worden dat de reeds ontvangen data uitbreidt met nieuwe data. Er moet ook op gelet worden dat er geen bestanden met wijzigingen overgeslagen worden.
- Bestandsuitwisseling is niet geschikt voor *ad-hoc* bevragingen. Het werkt goed als de dataconsumenten voortdurend dezelfde rapporten en analyses gebruiken.
- Er moet gegarandeerd worden dat het bestand dat het aflegt niet geopend en of gelezen kan worden en dat de data in het bestand niet bewust of onbewust gewijzigd kan worden. Zeker als email gebruikt wordt, moet geregeld worden dat anderen de data niet kunnen inzien. Encryptie van de bestanden moet ongetwijfeld ingezet worden.
- Bestandsuitwisseling leidt per definitie tot opslag van veel redundante data bij zowel de versturende als de ontvangende organisatie. Eerst wordt de data uit een bronsysteem gehaald en in het bestand opgeslagen. Dit is de eerste kopie. Dan wordt het bestand verstuurd en ontstaat bij de ontvanger een tweede kopie. Vervolgens wordt de data in een systeem geïmporteerd en daarmee wordt de derde kopie aangemaakt. En als dit systeem bijvoorbeeld een datawarehouse-omgeving is, ontstaan daar nog meer kopieën. Het beheren en consistent houden van al deze dataduplicaten kan zeer complex en tijdrovend zijn. Als het bestand naar meerdere organisaties gestuurd wordt, wordt het

beheer van al deze duplicaten alleen maar complexer. Dit kan leiden tot een onbeheersbare situatie waarbij geen garantie meer gegeven kan worden dat de data correct is.

- In de AVG is het vergeetrecht expliciet opgenomen. Daarnaast is het *recht van rectificatie* opgenomen voor onjuiste of incomplete persoonsgegevens. Zoals eerder aangegeven, worden bij bestandsuitwisseling veel duplicaten gemaakt. De eigenaar van de data heeft de uitdaging om ervoor te zorgen dat als een bepaalde hoeveelheid data verwijderd of gewijzigd wordt, dit ook aangepast wordt in de afgeleide bestanden bij alle ontvangers.
- Veel van deze oplossingen zijn standalone en delen nagenoeg geen dataverwerkingsspecificaties.

Bestandsuitwisseling vervangen door datavirtualisatie

De genoemde nadelen kunnen geheel of gedeeltelijk weggenomen worden door middel van datavirtualisatie: zie het rechter diagram in figuur 5. In plaats van te werken met bestanden wordt op de databron een view gedefinieerd die virtueel dezelfde inhoud heeft als het te versturen bestand. Op deze view worden alle benodigde databeveiligingsregels gedefinieerd. De ontvangende organisatie krijgt toegang tot de view via een gewenste interface (bijvoorbeeld SQL, REST of SOAP/XML). Deze kan de data vervolgens ophalen wanneer dat nodig is. In feite wordt data niet meer periodiek van de ene naar de andere organisatie verstuurd, maar kan de ontvanger de data op elk moment bij de eigenaar opvragen. Vergelijk dit met een ziekenhuis, dat aan het einde van de dag onderzoeksresultaten naar een patiënt toestuurt, met dat een patiënt zelf op elk gewenst moment de resultaten kan opvragen via bijvoorbeeld een medisch portaal.

Als het direct ophalen van data geen optie is, kan datavirtualisatie nog steeds nuttig ingezet worden voor bestandsuitwisseling. De virtuele inhoud van een view kan namelijk door een datavirtualisatie-server naar een bestand omgezet worden waarbij uit verscheidene formaten gekozen kan worden, zoals CSV, XLS, HTML en TDE (voor Tableau). Dit vereist geen aanpassingen van de dataverwerkingsspecificaties van de view. Dit lost niet alle nadelen op, maar maakt het voor de versturende organisatie wel eenvoudiger om bestanden aan te maken.

Voordelen van bestandsuitwisseling vervangen door datavirtualisatie

- Data wordt minder vaak gedupliceerd, waardoor het eenvoudiger is om te voldoen aan aspecten van AVG.
- Als de view niet gecached is, ziet de ontvangende partij altijd volledig bijgewerkte data. De datalatentie kan dus tot nul teruggebracht worden en dit biedt de ontvangende organisatie de mogelijkheid voor real-time analyse.
- Het is in de meeste gevallen niet noodzakelijk een mechanisme te ontwikkelen dat versiebeheer van de data uitvoert, aangezien dit door de eigenaar van de databronnen uitgevoerd wordt.
- De ontvanger kan nu zelf bepalen of telkens het gehele bestand opgevraagd wordt of slechts een deel ervan of misschien zelfs maar één of enkele records. Doordat er continu toegang is tot de data, wordt de noodzaak om telkens alle data op te vragen waarschijnlijk kleiner.
- Oplossingen voor de ene organisatie kunnen hergebruikt worden voor een andere organisatie.
- Als een dataconsument voor een bepaalde periode een stabiele data-toestand nodig heeft, kan de view voor die periode gecached worden. Caching kan ook ingezet worden als de verstoring op de databronnen te groot geacht wordt.
- Deze oplossing kan ook gebruikt worden voor ad-hoc bevragingen.

Voorwaarden

- De databron die door een datavirtualisatie-server benaderd wordt, moet wel krachtig genoeg zijn om de workload van ontvangende organisaties te kunnen verwerken. Deze workload kan groter zijn dan het bestaande alternatief waarbij bestanden aangemaakt en verstuurd worden. Caching door de datavirtualisatie-server zou een performance probleem kunnen oplossen.
- Er wordt mogelijk een meer dan gemiddeld aantal views gecached, dus hiervoor moet een oplossing ingericht en beheerd worden.
- Omdat er nu data tussen organisaties uitgewisseld wordt, moet de datavirtualisatie-server alle beveiligingseisen ondersteunen. Dit moet grondig worden bestudeerd en in een Proof-of-Concept uitgebreid getest worden. Een PEN-test (Penetration Testing) wordt tevens sterk aanbevolen om eventuele zwakke plekken in de beveiliging op te sporen.

- Een ontvangende organisatie moet wel accepteren dat deze de data zelf moet opvragen, omdat deze niet meer automatisch toegestuurd wordt. Anderzijds heeft de ontvanger toegang tot meer actuele data.
- Als de ontvanger met grotere regelmaat data zal opvragen, ontstaan er ook minder dataduplicaten wat de ondersteuning voor het vergeetrecht vereenvoudigt.

3.6 ONDERSTEUNEN VAN EXTERNE BEVRAGINGEN VOOR DATA

Inleiding externe bevragingen

Het komt steeds vaker voor dat organisaties externe partijen de mogelijkheid (moeten) bieden om hun data op te vragen. Dit kunnen allerlei partijen zijn, waaronder het Centraal Bureau voor de Statistiek, de Kamer van Koophandel, maar ook commerciële organisaties. Een ander voorbeeld dat steeds vaker terug te zien is, is dat ziekenhuizen patiënten toestaan in te loggen op een portaal waar ze medische data over zichzelf kunnen opvragen. Over het algemeen zullen steeds meer burgers data live kunnen opvragen via onder meer mobiele apps, portalen en smart televisies.

In het verlengde van deze behoefte, waarbij de vraag voor externe data vaststaat, moet tevens de mogelijkheid bestaan dat externe partijen de data zelf kunnen analyseren. Ondernemingen krijgen bijvoorbeeld de mogelijkheid om het UWV-vacaturebestand live en vrij te analyseren. Het CBS biedt al vele jaren de mogelijkheid om data te analyseren. Het verschil met de vorige toepassing is dat nu niet telkens dezelfde vraag gesteld wordt (Bijvoorbeeld: geef de geboorteplaats behorend bij een burgerservicenummer), maar dat deze elke keer anders kan zijn, aangezien de analysevraag telkens weer anders kan zijn.

Voor beide vormen van externe bevraging zit er tijd tussen het insturen van de vraag en het ontvangen van de data. Dit kan uren tot dagen duren en is afhankelijk van hoe de organisatie dit proces geïmplementeerd heeft.

Nadelen van het verzenden van data

- Het kost tijd om de opgevraagde data te extraheren, in een bestand op te slaan en vervolgens te versturen. Hierdoor kan er veel tijd verloren gaan tussen de aanvraag en het moment waarop de dataconsument de data ontvangt en kan gebruiken. De datalatentie kan dus hoog zijn.
- Als het verzendproces niet volledig automatisch verloopt, kunnen er fouten optreden en kan mogelijk de verkeerde data gestuurd worden of de goede data naar de verkeerde dataconsument.
- Wanneer bestanden voor analyse aangevraagd en verstuurd zijn en bij nader inzien blijkt dat er meer data voor de analyse nodig is, moet het gehele proces opnieuw doorlopen worden.
- Nadat de dataconsument een bestand voor analyse ontvangen heeft, slaat de dataconsument deze op. Het is onduidelijk hoe gegarandeerd kan worden dat deze data goed beveiligd wordt.
- Met elke minuut die verstrijkt, veroudert de data.
- Eventueel noodzakelijke tracering van data is niet mogelijk.

Externe bevragingen met datavirtualisatie

In figuur 6 is weergegeven hoe externe bevragingen door een datavirtualisatie-server verwerkt zouden kunnen worden. In deze oplossing is de data opvraagbaar via een service-interface (voor de eerste toepassing) en een SQL(-achtige) interface (voor de tweede toepassing). Bij een service-interface roept de dataconsument een service (onderdeel van de datavirtualisatie-server) aan die op een of andere wijze data uit de databron ophaalt. Dit soort services kunnen door andere applicaties en systemen aangeroepen worden. De services kunnen tevens gebruikt worden om data te integreren, aggregeren, transformeren en beveiligen. Bij een SQL-interface verstuurt een dataconsument een SQL-instructie naar de datavirtualisatie-server. Wellicht ten overvloede: bij beide vormen van bevragen van data spelen databeveiliging en data-privacy een belangrijke rol.

Het essentiële verschil tussen de traditionele oplossing en de oplossing met een datavirtualisatie-server is dat bij de eerstgenoemde het lang kan duren voordat de data toegestuurd wordt, terwijl bij de laatstgenoemde de data direct na aanvraag wordt gestuurd. Bovendien worden bij de traditionele oplossing kopieën aangemaakt (met alle nadelen van dien) en bij de oplossing met een datavirtualisatie-server is het niet noodzakelijk om dataduplicaten op te slaan.

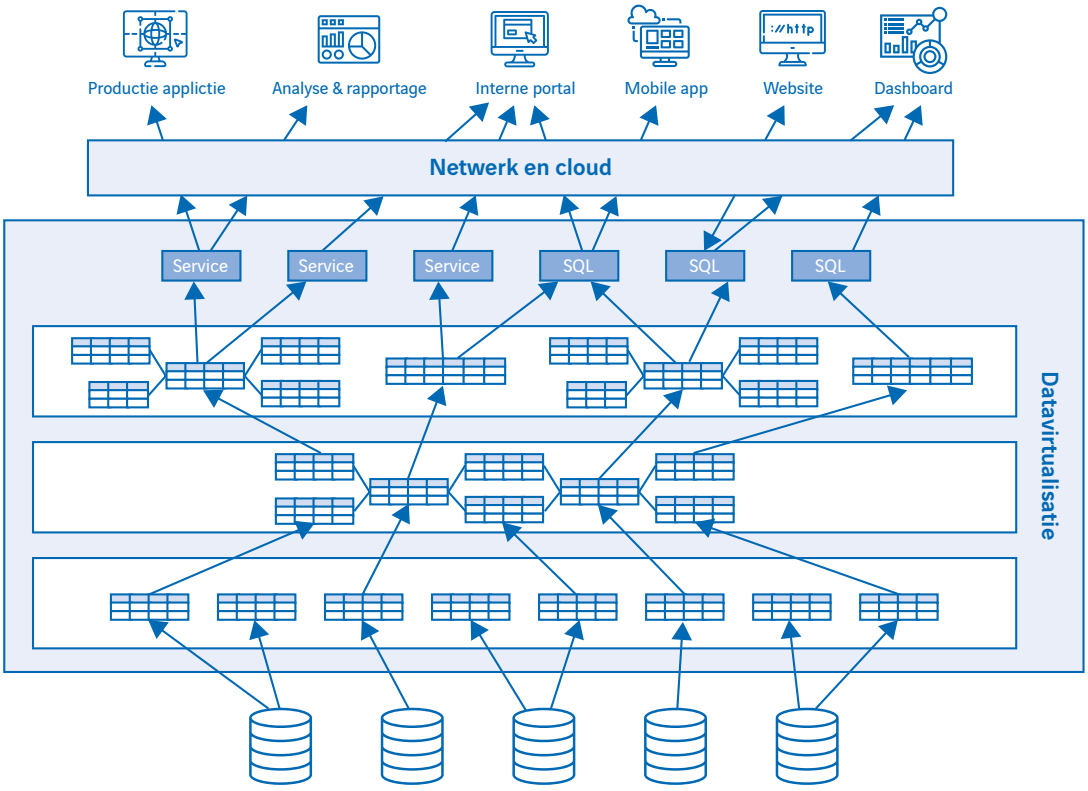
Dit verschil kan enigszins vergeleken worden met een YouTube-oplossing tegenover een eReader-oplossing. Bij

een eReader wordt een elektronische kopie van een boek naar elke eReader gestuurd waarna het boek gelezen kan worden. Dit heeft tot gevolg dat er duizenden exemplaren rondzwerven. Bij YouTube worden de films niet naar de apparaten van de kijkers gekopieerd, maar wordt de film live gestreamd vanuit een systeem. De film ligt dus maar één keer opgeslagen en wordt vele malen opgevraagd.

De oplossing met een datavirtualisatie-server kan bijvoorbeeld toegepast worden op het elektronisch patiëntendossier. In dit geval worden alle patiëntengegevens centraal in één database opgeslagen. Alle toegang tot deze data verloopt via één datavirtualisatie-server. Dus elke persoon die toegang nodig heeft tot de data, doet dit via de datavirtualisatie-server. Hier wordt aangegeven welke personen welk deel van een patiëntendossier mogen inzien. Als het dossier met meer data uitgebreid moet worden of gewijzigd moet worden, dan hoeft dit maar op één plek te gebeuren en alle dataconsumenten zien in principe de actuele data. Een mechanisme waar een patiënt zelf kan aangeven wie welke data mag zien, kan hier ook mee gebouwd worden.

Voordelen van bevragingen door externe partijen met datavirtualisatie

- Elke binnenkomende vraag wordt direct verwerkt, waardoor de datalatentie zeer laag is.
- Aangezien het verzendproces volledig automatisch verloopt, kunnen er geen fouten optreden en er wordt geen verkeerde data gestuurd of de gevraagde data gaat niet naar de verkeerde dataconsument.
- Als een analist ontdekt dat de data die met een query opgevraagd is niet toereikend is, hoeft er slechts een nieuwe query gestart te worden. Die extra data is vrijwel direct beschikbaar voor analyse.
- Data hoeft niet meer zelf door analisten opgeslagen te worden waardoor deze ook niet apart beveiligd hoeft te worden. Let wel, ze kunnen de data nog wel opslaan, maar dan is de vraag wie er verantwoordelijk voor is.
- De data die door de datavirtualisatie-server geretourneerd wordt, is even actueel als de databron waaruit de data opgehaald wordt.
- Door de datavirtualisatie-server centraal data te laten loggen en analyseren, is er inzicht in onrechtmatig gebruik.



Figuur 6: door externe partijen de data via een datavirtualisatie-server te laten bevragen, wordt met meer actuele data gewerkt, waardoor er minder dataduplicatennodig zijn.

- Toegang tot data kan centraal geregeld worden, waardoor de kans op datalekken verkleind en het beheer van data vereenvoudigd wordt.
- Als interne data met data van externe partijen verrijkt moet worden en als deze externe partij deze verrijkte data zelf weer afneemt, is de verrijkte data real-time ter beschikking, omdat de verrijking in de datavirtualisatie-server plaatsvindt.

Voorwaarden

- Als het aantal bevragingen voor data groot is, moet de datavirtualisatie-server wel werken met een dataplatform dat de workload aankan en waarbij de datavirtualisatie-server query pushdown vol kan inzetten.
- Als het aantal bevragingen voor data groot is, moet de datavirtualisatie-server zeer schaalbaar zijn en query's parallel kunnen verwerken.
- Er moet veel aandacht besteed worden aan de beveiliging van data tegen incorrect gebruik, hacks van buitenaf en tegen DDoS aanvallen.
- Het is belangrijk dat veel aandacht besteed wordt aan de kwaliteit van de data in de databronnen die bij deze oplossing geraadpleegd worden. Het opschonen van de data bij een virtuele oplossing is moeilijker dan bij een klassieke oplossing.

3.7 ORGANISEREN VAN TOEGANG TOT EXTERNE DATA

Inleiding

Voor veel analyses en rapportages kan het toevoegen van externe data een enorme verrijking zijn. Voor een gemeente kan het bijvoorbeeld interessant zijn om bij bepaalde analyses data over de te verwachten instroom en uitstroom van inwoners van het Centraal Bureau van de Statistiek te gebruiken, voor het Ministerie van Infrastructuur en Waterstaat kan het nuttig zijn om te bestuderen of bepaalde epidemieën invloed hebben op verkeersstromen en voor een investeringsbedrijf kan financiële informatie van Bloomberg erg belangrijk zijn. De data over de verkeersstromen zijn in eigen beheer, maar die over epidemieën moeten van een externe partij betrokken worden. Ook basisregistraties[4], zoals BGT (Basisregistratie Grootschalige Topografie) en BRK (Basisregistratie Kadaster) worden door steeds meer organisaties geraadpleegd. Hetzelfde geldt voor *open databronnen*[5] zoals Open Archieven en Openspending

Nadelen van separate oplossingen

Koppelingen met externe databronnen worden meestal tot stand gebracht door voor elk rapport een aparte koppeling te ontwikkelen; zie figuur 7. Dit betekent dat binnen een organisatie talloze, separate oplossingen ontstaan. Deze aanpak kent de volgende nadelen:

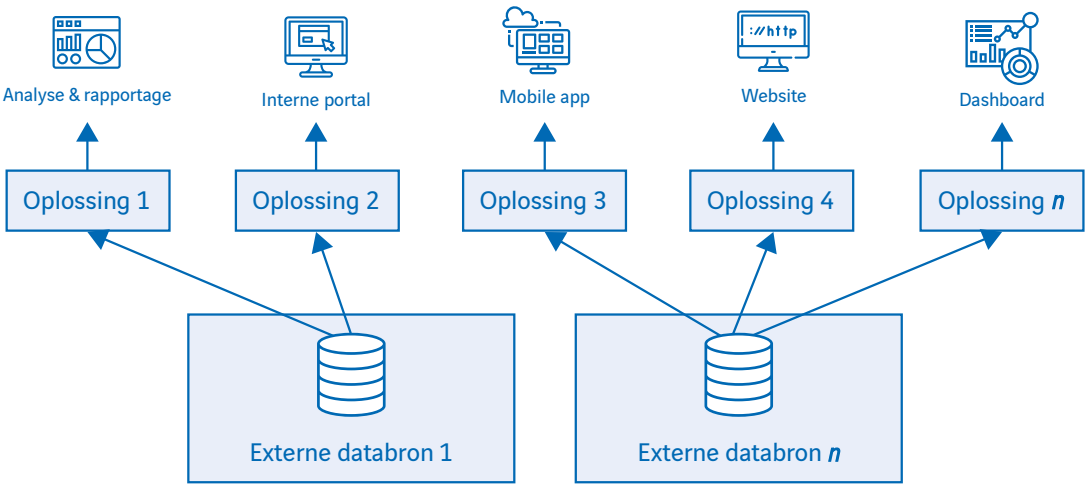
- Het is een verspilling van tijd en geld als elke gebruiker van eenzelfde organisatie een aparte koppeling met een specifieke externe databron ontwikkelt. Onafhankelijk van elkaar wordt uitgezocht hoe dat technisch bewerkstelligd moet worden, wat alle data betekent, wat de betekenis van alle codes is, hoe die data gekoppeld kan worden aan de eigen interne data, enzovoort. Het wiel wordt telkens opnieuw uitgevonden.
- De kans bestaat dat sommigen externe data verkeerd interpreteren en de data foutief in hun analyse verwerken, waardoor foute conclusies getrokken kunnen worden.
- Als de eigenaar van een externe databron iets verandert aan de wijze waarop de data aangeleverd wordt, hoe de data in elkaar zit of wat het betekent, dan moet iedere ontwikkelaar aan de slag om deze verandering correct door te voeren.

Externe data beschikbaar maken met datavirtualisatie

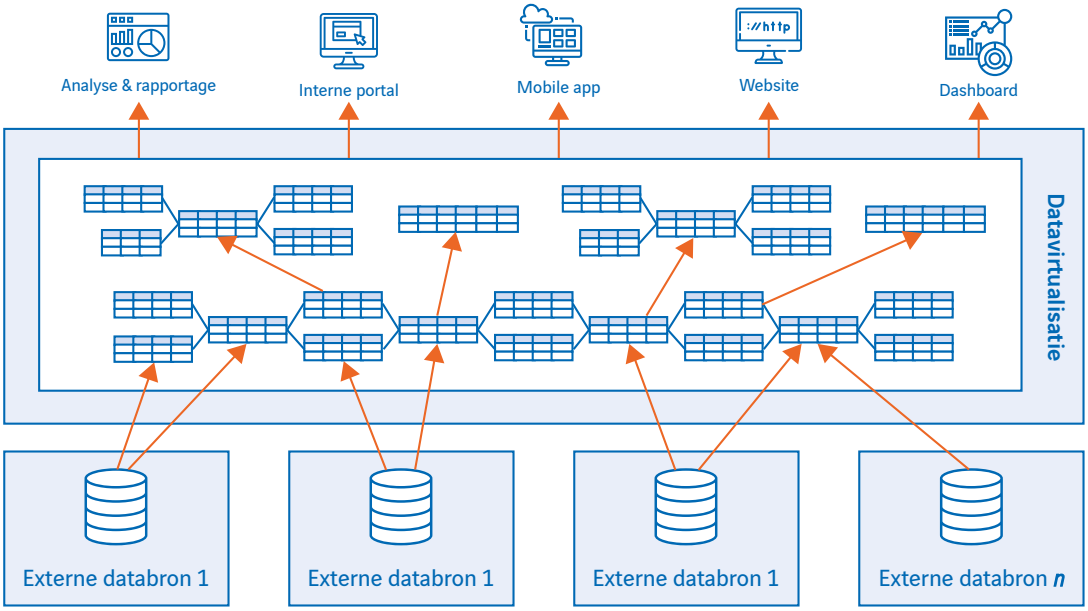
Datavirtualisatie-servers ondersteunen veel mogelijkheden om views te definiëren op allerlei databronnen, dus ook op externe databronnen met complexe interfaces. Door datavirtualisatie-servers op externe databronnen te definiëren, kan externe data op een eenvoudige manier aan alle mogelijke dataconsumenten binnen de organisatie ter beschikking gesteld worden; zie figuur 8. Door views te stapelen kan externe data ook getransformeerd worden tot de vorm die de dataconsument wenst.

4 DigitaleOverheid.nl; www.digitaleoverheid.nl/dossiers/basisregistraties/

5 Open Overheid; zie https://www.open-overheid.nl/open-data/



Figuur 7: als de toegang tot externe databronnen via aparte oplossingen verloopt die elk ontwikkeld is voor één applicatie, ontstaat een wildgroei aan oplossingen.



Figuur 8: met datavirtualisatie kan de toegang tot externe data beter georganiseerd worden waardoor meerdere dataconsumenten eenvoudiger en sneller toegang hebben.

Voordelen van externe data beschikbaar maken met datavirtualisatie

- Een koppeling die binnen een datavirtualisatie-server gedefinieerd is, kan door veel dataconsumenten hergebruikt worden, ook wanneer zij de data via een andere interface willen benaderen en wanneer de data in een andere vorm aangeleverd moet worden. In dit geval wordt eenmalig uitgezocht hoe dit technisch bewerkstelligd moet worden, wat alle data betekent, wat de betekenis van de codes is, hoe die data gekoppeld kan worden aan interne data, enzovoort.
- Binnen een datavirtualisatie-server kan metadata aan data toegevoegd worden waarin beschreven wordt wat de data betekent en hoe deze geïnterpreteerd moet worden. Dit vergroot de kans dat externe data correct geïnterpreteerd wordt.
- Als de eigenaar van een externe databron iets verandert aan de wijze waarop de data aangeleverd wordt, hoe de data in elkaar zit of wat het betekent, dan hoeft dit alleen

binnen de datavirtualisatie-server eenmalig aangepast te worden. Voor alle dataconsumenten wordt deze wijziging afgeschermd (data-abstractie) waardoor zij niets hoeven aan te passen. Eventueel kunnen de oude en de nieuwe versie tegelijk aangeboden worden zodat dataconsumenten overgangstijd hebben om de nodige aanpassingen door te voeren.

- Een vereenvoudigde bereikbaarheid van externe data kan het gebruik ervan vergroten.
- De snelheid verbetert waarmee externe data beschikbaar komt. Externe data kan dan ook voor ad-hoc bevragingen ingezet worden. In feite kan een gehele catalogus met externe databronnen opgebouwd worden.

Voorwaarden

- Er moet bij voldoende dataconsumenten de wens bestaan om externe data te gebruiken voor analyses en rapportage.
- Er moet een centrale functie geïntroduceerd worden die verantwoordelijk is voor het beheer van de views die toegang tot de externe data geven. Deze functie werkt tevens als vraagbaak voor alle ontwikkelaars en heeft de verantwoordelijkheid om de beschikbaarheid van nieuwe externe databronnen binnen de organisatie bekend te maken.
- Het moet binnen de organisatie bekend gemaakt worden dat toegang tot externe databronnen via de datavirtualisatie-server verloopt.
- Ontwikkelaars moeten bereid zijn om bepaalde aspecten van hun rapporten niet meer zelf in hun eigen BI-product te ontwikkelen.
- Er moet al een achterstand bestaan wat betreft het ontwikkelen van rapporten die het gebruik van externe data vereisen.
- De organisatie moet de business voordelen inzien van het versnellen van de ontwikkeling van rapporten die gebruik maken van externe databronnen.

3.8 ONDERSTEUNEN VAN DATA-ANALYSES VOOR DATA SCIENTISTS

Inleiding

Datascience is een populaire vorm van datagebruik bij organisaties geworden. Bij datascience worden *descriptive* (beschrijvende) of *prescriptive* (normatieve) modellen ontwikkeld waarmee bedrijfs- of beslissingsprocessen verbeterd of versneld kunnen worden. Datascience wordt bijvoorbeeld ingezet bij het opsporen van cyberspionage, fraudedetectie van creditcardbetalingen en het automatisch beheren van verkeer. Wel kleven er mogelijk enkele sociaal-maatschappelijke problemen aan, waardoor datascience zorgvuldig ingezet moet worden.

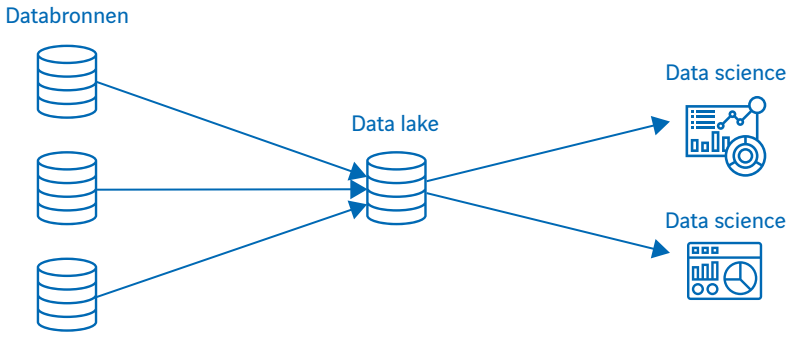
Studies melden dat datascientists slechts 20% van hun tijd aan analytisch werk besteden en maar liefst 80% aan taken die te maken hebben met het voorbereiden van data, ofwel *data preparation*[6]. Deze taken omvatten het bepalen welke data-elementen en datawaarden vereist zijn, het identificeren van de systemen waaruit data geëxtraheerd moet worden, het ontwikkelen van programma's om de data uit die systemen te extraheren, het omzetten van de data naar zinvolle data, het laden van de geëxtraheerde data in een bepaald systeem en de data gemakkelijk beschikbaar maken voor analyse.

Een van de verschillen tussen data scientists en de meeste andere dataconsumenten is dat de eerstgenoemde groep vaak toegang wil hebben tot de originele data, die in het geheel nog niet bewerkt is. Voor data scientists is bijvoorbeeld het benaderen van datamarts niet ideaal, want deze data is meestal sterk bewerkt, wat de primaire gebruikers uiteraard juist nodig hebben. Data scientists geven vaak aan dat het benaderen van bewerkte brondata het resultaat van hun analyses beïnvloedt.

Het data lake en datascience

Om data scientists sneller en eenvoudiger toegang tot brondata te geven, wordt steeds vaker een data-architectuur ontwikkeld, genaamd het *data lake*; zie figuur 9. Deze architectuur wordt meestal onafhankelijk van het data warehouse ontwikkeld. Het data lake bevat alle onbewerkte data uit de bronsystemen die data scientists nodig hebben. Het grote voordeel van een data lake is dat data scientists niet veel tijd meer hoeven te besteden aan het verzamelen van alle benodigde data. Met het data lake hebben ze

6 A. Ruiz (InfoWorld), The 80/20 *Data Science Dilemma*, September 2017; zie https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html



Figuur 9: het data lake.

één centraal systeem tot hun beschikking. In principe is het centraliseren van alle data voor data scientists een goede manier om data snel en eenvoudig voor hen beschikbaar te maken. Met andere woorden verkort het de data preparation fase.

Veelal wordt een data lake ontwikkeld op een cloud platform, zoals Amazon, Google of Microsoft Azure.

Nadelen van het data lake

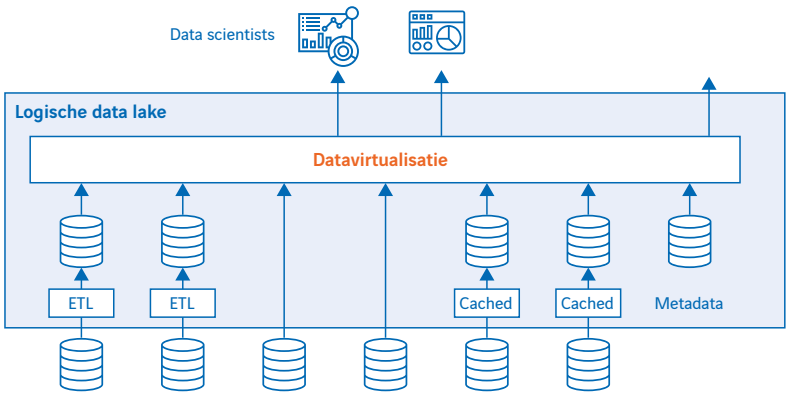
Het inloggen op één systeem en daarmee toegang hebben tot alle gewenste data is ontegenzeggelijk een groot voordeel voor data scientists. Toch kent het data lake enkele praktische nadelen:

- In een datawarehouse-omgeving besteden ETL-programmeurs het grootste deel van hun tijd aan het ontwikkelen van de "T" (de specificaties om de data te transformeren) en niet zozeer aan de "E" of de "L". Omdat de in het data lake opgeslagen data nog onbewerkt is, moeten data scientists zelf de "T" ontwikkelen en dat kost hen net zoveel tijd als de ETL-programmeurs.
- In sommige situaties kan de enorme hoeveelheid data afkomstig van de databronnen te veel zijn om naar het data lake te versturen en te veel om fysiek te kopiëren en te laden. Bandbreedte beperkingen en langdurige laadprocessen van data kunnen het onmogelijk maken om een big databron naar het data lake te kopiëren.
- Om verschillende redenen, van politiek tot beveiliging, zijn niet alle onderdelen van een organisatie altijd bereid hun data te delen met een gecentraliseerde omgeving te delen. Dit kan ertoe leiden dat hun data niet beschikbaar gesteld wordt.
- Steeds meer wetten, regels en voorschriften beperken hoe data ingezet mag worden. In sommige omgevingen is bijvoorbeeld het samen opslaan van bepaalde data niet toegestaan. Soms mag data een land niet verlaten.
- De data in sommige bronsystemen is zwaar beveiligd tegen onjuist en frauduleus gebruik. Eigenaren staan daarom niet altijd toe dat hun data geëxtraheerd wordt en naar een ander systeem, zoals een data lake, gekopieerd wordt.
- Niet alle brondata wordt met beschrijvende metadata geleverd, waardoor data scientists moeilijk kunnen bepalen wat specifieke data-elementen betekenen. Een verkeerde interpretatie kan leiden tot onjuiste bedrijfsinzichten.
- Het data lake is een systeem waarin data opgeslagen wordt en beheerd dient te worden. Eigenaarschap, metadata en data life cycle management moeten dus ingeregeld worden.

Datascience en datavirtualisatie

Het *logische data lake* (ofwel het virtuele data lake) is een alternatieve data-architectuur om data scientists toegang te geven tot dezelfde data waartoe ze toegang hebben bij een fysiek data lake; zie figuur 10.

Bij het logische data lake lijkt het voor data scientists of alle data centraal in één systeem opgeslagen is. Dit is echter niet het geval, ook al geeft een datavirtualisatie-server die indruk. Een deel van de data wordt centraal gekopieerd en opgeslagen (de twee databronnen aan de linkerkant), sommige databronnen worden op afstand benaderd (de twee databronnen in het midden) en sommige worden lokaal in de cache opgeslagen (de twee aan de rechterkant). Afhankelijk van wat vereist, mogelijk en haalbaar is, kan een van de drie oplossingen worden gebruikt om een bepaalde databron voor data scientists toegankelijk te maken. Daar waar het



Figuur 10: het logische data lake.

kopiëren en centraal opslaan van data slechts een van de opties is in de oorspronkelijke data lake-architectuur, is dit zeker niet de enige optie in het logische data lake. Welke van de opties gekozen wordt, blijft voor data scientists verborgen. Dit lost veel van de in deze paragraaf beschreven problemen op.

Voordelen van datavirtualisatie voor datascience

Het toegankelijk maken van data via een logisch data lake met een datavirtualisatie-server biedt voor data scientists de volgende voordelen:

- Als data scientists zelf databronnen benaderen, moeten ze ook werken met de verschillende interfaces, talen en APIs van die systemen. Dit kost veel tijd. Door middel van datavirtualisatie kunnen data scientists alle benodigde data met één taal benaderen, ongeacht de taal die de onderliggende databron ondersteunt.
- Als de data in een databron met een zogenaamd schema-on-read formaat opgeslagen is, kan met een datavirtualisatie-server het schema, ofwel de structuur van de data, gedefinieerd worden. Zo hebben data scientists niet te maken met dit soort technische details. Let wel: zo wordt de data niet gewijzigd, maar uitsluitend de structuur ervan vastgelegd.
- Bij veel databronnen hebben data scientists geen toegang tot beschrijvende metadata (of deze ontbreekt volledig). De metadata kan aan de datavirtualisatie-server toegevoegd worden en voor data scientists toegankelijk gemaakt worden.
- In sommige databronnen wordt data zeer cryptisch opgeslagen. Views kunnen gedefinieerd worden die de data ontsleutelen en transformeren naar betekenisvollere data en deze zo eenvoudiger, toegankelijk en begrijpelijker maken.

Voorwaarden

- Data scientists moeten bereid zijn om met SQL te werken en data via een datavirtualisatie-server uit de databronnen te betrekken. Zij moeten de voordelen hiervan inzien. Data scientists kunnen grofweg in twee groepen verdeeld worden: zij die liever hun eigen modellen ontwikkelen met talen zoals R en Python en zij die zo snel mogelijk modellen proberen te creëren en elk product zullen inzetten om hierbij te helpen. De datavirtualisatie-aanpak past het beste bij de tweede groep. De ervaring leert dat de eerste groep moeilijk van de voordelen te overtuigen is.
- Binnen de organisatie bestaan strenge regels betreffende de ontwikkeling en het gebruik van datascience-modellen. Zo moet er bijvoorbeeld versiebeheer van de modellen plaatsvinden en modellen moeten *immutable* (onveranderbaar) en transparant geoperationaliseerd kunnen worden. Dit soort regels vereist dat het gehele ontwikkelproces van datascience-modellen geprofessionaliseerd wordt. Datavirtualisatie past bij een dergelijke toepassing van datascience.

3.9 CENTRALISEREN VAN SPECIFICATIES VOOR DATABEVEILIGING

Inleiding databeveiliging

Veel bestaande systemen ondersteunen voornamelijk hun eigen databeveiligingsfuncties en sommige ondersteunen geen of slechts beperkte mogelijkheden hiervoor. Dit leidt

tot twee problemen. Ten eerste moeten dataconsumenten, die meerdere databronnen benaderen, omgaan met al die verschillende beveiligingssystemen en ten tweede zijn beveiligingsspecificaties, zoals wie mag welke data zien, gedupliceerd in deze systemen. Als een werknemer ontslag neemt, dan kan het verwijderen van deze persoon uit alle databronnen een complexe exercitie zijn.

Databeveiliging met datavirtualisatie

Met datavirtualisatie kunnen databeveiligings-specificaties centraal gedefinieerd worden; zie figuur 11. Dataconsumenten hebben dan te maken met slechts één beveiligingssysteem en de beveiligingsspecificaties zijn niet gedupliceerd. In dit geval fungeert de datavirtualisatie-server als een centraal punt waar alle data gerelateerde vragen binnenkomen en verwerkt worden.

In deze tijd, waarin organisaties door de media aan de schandpaal genageld kunnen worden voor fouten met databeveiliging, is het toevoegen van één centraal beveiligingssysteem een groot voordeel.

Voorwaarden

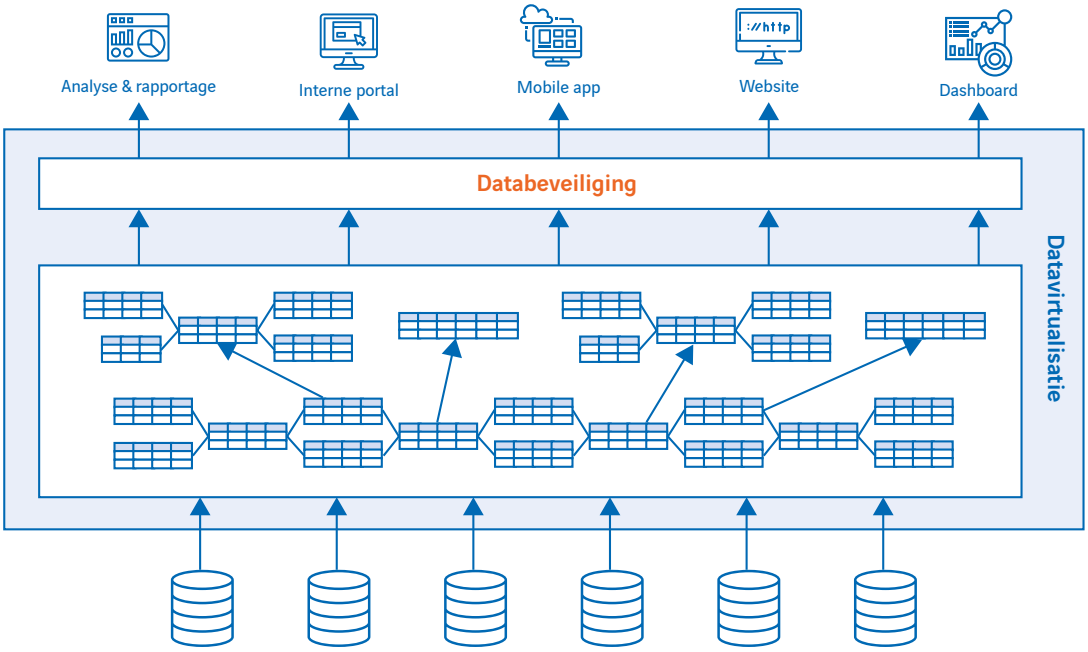
- Veel dataconsumenten benaderen meerdere databronnen en hebben zo te maken met meerdere beveiligingssystemen.
- De organisatie is zich bewust van het belang van databeveiliging en onderschrijft de huidige gebreken op dit gebied.

3.10 CENTRALISEREN VAN SPECIFICATIES VOOR DATA-PRIVACY

Inleiding data-privacy

Veel organisaties registreren persoonsgegevens. Zij moeten aan de AVG-regels voldoen. In de IT-systemen moeten deze regels voor data-privacy ingebouwd worden om ervoor te zorgen dat bepaalde data voor bepaalde gebruikers onbereikbaar is. Oftewel: sommige data-elementen moeten voor sommige gebruikers afgeschermd worden en andere moeten op één of andere manier gepseudonimiseerd of geanonimiseerd worden. Veel BI- en andere soorten producten ondersteunen dit soort mogelijkheden echter nog niet of slechts minimaal. Daarom moet aan de aanleverende databronnen een mechanisme toegevoegd worden. Dit betekent dat elke databron moet bijhouden wie welke data mag zien en ook het gebruik moet loggen.

Bij een veelvuldig gekozen oplossing wordt data gekopieerd naar een andere database of tabel en tijdens het kopieerproces gepseudonimiseerd of geanonimiseerd. Dit kan alleen niet gebruikersafhankelijk, behalve wanneer



Figuur 11: het centraal definiëren van regels voor databeveiliging.

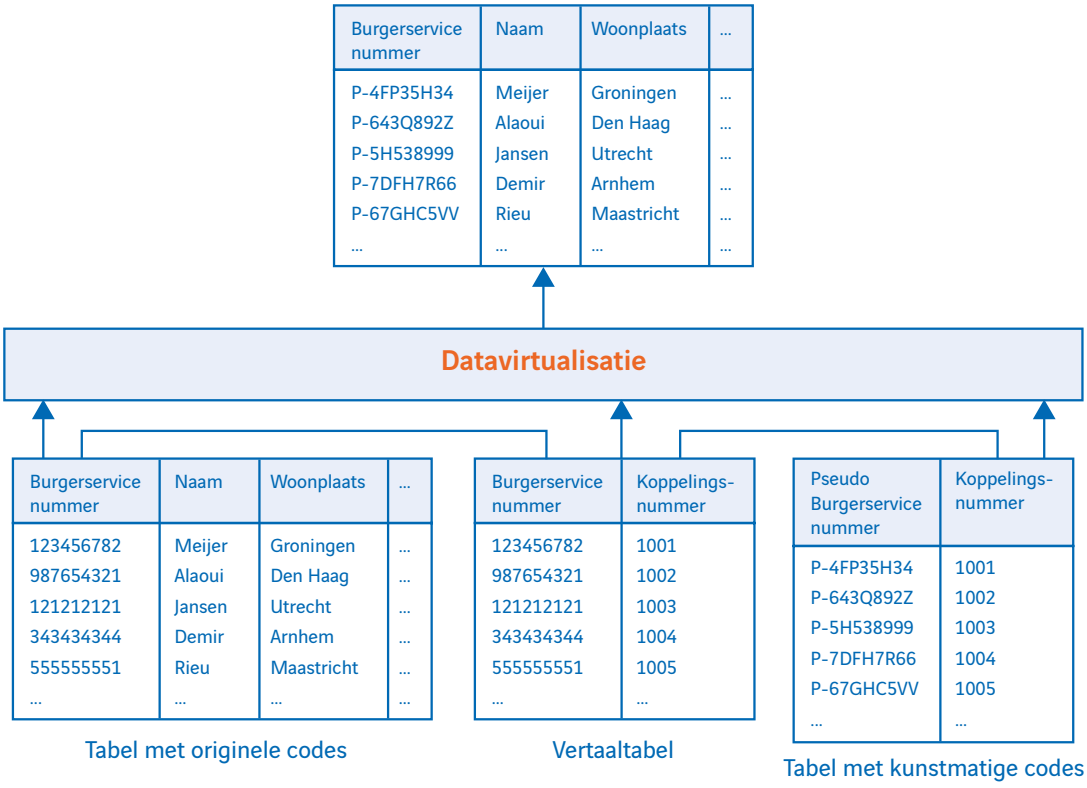
er twee kolommen voor het burgerservicenummer opgenomen worden: een met de originele data en een met de beveiligde data. Het nadeel van deze aanpak is dat de datalatentie weer omhooggaat en bovendien moet de data, die nogmaals opgeslagen wordt, ook weer beveiligd worden.

Data-privacy met datavirtualisatie

Het is eenvoudiger om de specificaties voor het afdwingen van data-privacy-regels centraal in een datavirtualisatie-server te definiëren. Wanneer bijvoorbeeld een bepaald rapport burgerservicenummers (BSN) bevat die gebruikers niet mogen zien, maar dat de laatste vier tekens wel zichtbaar moeten zijn om bepaalde controles te kunnen uitvoeren, kan binnen een datavirtualisatie-server een view gedefinieerd worden die alle benodigde data uit de databron haalt, maar wel zo dat de eerste tekens van het nummer door sterretjes vervangen worden. Deze masking-specificatie kan gebruikersafhankelijk gemaakt worden; voor sommigen wordt het BSN volledig vervangen, bij anderen gedeeltelijk en weer anderen krijgen de volledige waarde te zien. Op deze manier kan veel data gepseudonimiseerd worden. Het voordeel van deze aanpak is dat voor elke dataconsument die de data via de datavirtualisatie-server benadert, deze gepseudonimiseerd wordt zonder dat de data nogmaals opgeslagen hoeft te worden. Pseudonimisatie vindt dan tijdens het bevragen van de data plaats.

Met een datavirtualisatie-server kunnen persoonsgegevens ook volledig gepseudonimiseerd worden. Hierbij wordt bijvoorbeeld een unieke code van een persoon vervangen door een kunstmatige code die op geen enkele wijze naar deze persoon verwijst. Deze unieke codes kunnen volledig willekeurig zijn.

Als de te genereren kunstmatige codes gedurende lange tijd voor alle tabellen gelijk moeten zijn, is deze oplossing mogelijk niet geschikt. Mogelijker moeten de gepseudonimiseerde codes onthouden en dus opgeslagen worden. In een bepaald systeem wordt dan een vertaaltabel aangemaakt; zie figuur 12. Het voordeel van deze aanpak is dat de kunstmatige codes altijd dezelfde zijn en dus altijd dezelfde gepseudonimiseerd code voor dezelfde persoon opgehaald kan worden.



Figuur 12: het pseudonimiseren van persoonsgegevens door middel van vertaaltabellen.

Opmerking: als voor deze aanpak gekozen wordt, moeten de drie tabellen in figuur 12 wel in aparte systemen opgeslagen worden. Zeker de vertaaltabel mag niet bij de andere tabellen opgeslagen worden. Alleen zo is gewaarborgd dat identificatie voorkomen wordt. Als de vertaaltabel vernietigd wordt of als identificatie anderszins onmogelijk is, is er sprake van anonieme data.

Pas bovendien op met het ter beschikking stellen van data-elementen die afzonderlijk niet een persoon identificeren, maar dat samen wel doen. Een klassiek voorbeeld is dat de kolom 'woonplaats' absoluut niet uniek is, dit geldt ook voor straatnaam (in Nederland wonen bijvoorbeeld veel mensen in een straat genaamd Herenstraat) en huisnummers (de meeste straten hebben een huisnummer 1). Samen identificeren ze echter wel een huis en dus misschien een persoon. Dergelijke data moet daarom mogelijk ook weggelaten of gepseudonimiseerd worden.

Voordelen van data-privacy met datavirtualisatie

- Specificaties voor data-privacy hoeven maar één keer voor een databron gedefinieerd te worden en zijn dan voor elke dataconsument actief. Dit vereenvoudigt de implementatie van deze regels en vergroot de bereidheid ze te definiëren.
- De specificaties voor data-privacy kunnen centraal beheerd worden.
- Indien noodzakelijk kunnen verschillende data-privacy specificaties voor dezelfde data voor verschillende dataconsumenten gedefinieerd worden.
- Afgezien van eventuele vertaaltabellen vereisen datavirtualisatie-servers niet dat geanonimiseerde en/of gepseudonimiseerde data opnieuw opgeslagen wordt. Anonimisatie en pseudonimisatie vinden on-demand plaats op het moment dat de data opgevraagd wordt.
- Veel aspecten van data-privacy kunnen met een datavirtualisatie-server gedefinieerd worden.
- Gebruik van data kan door de datavirtualisatie-server gelogd worden en hiermee kan misbruik in een vroeg stadium opgespoord worden.

Voorwaarden

- De organisatie is verantwoordelijk voor het registeren en beheren van persoonsgegevens die wat betreft opslag en gebruik moeten voldoen aan AVG en vergelijkbare wet- en regelgeving.
- De bestaande systemen bieden momenteel geen of niet voldoende mogelijkheden om specificaties voor data-privacy te definiëren.

3.11 MIGREREN VAN DATA NAAR EEN ANDER DATAPLATFORM

Inleiding

Aangezien er de laatste jaren veel nieuwe, krachtige en relatief goedkope dataplatformen beschikbaar gekomen zijn, overwegen organisaties hun data te migreren naar een van die nieuwe dataplatformen. Kostenbesparing (minder administratie en lagere licentiekosten) en performance-verbetering kunnen de redenen hiervoor zijn.

Het migreren van data naar een ander dataplatform is relatief eenvoudig. De uitdaging en vaak de showstopper is dat het nieuwe dataplatform een andere taal of dialect spreekt. Dat betekent dat de code van de applicaties die het dataplatform benaderen mogelijker aangepast moet worden. Hier wordt de meeste tijd aan besteed en het brengt grote risico's met zich mee.

Belangrijk is dat de code van de applicaties die de database benaderen met de hand geschreven is of dat deze gegenereerd is. Als deze gegenereerd is, is de vraag in hoeverre rekening gehouden wordt met de mogelijkheden van het dataplatform. Er zijn drie opties voor de code die op een dataplatform afgevuurd wordt:

- De code is met de hand geschreven.
- De code wordt gegenereerd, maar is generiek (dat wil zeggen, altijd hetzelfde ongeacht het dataplatform), er wordt bijvoorbeeld standaard SQL-code gegenereerd.
- De code wordt gegenereerd en geoptimaliseerd voor het dataplatform. Er worden dus speciale functies van het dataplatform gebruikt.

Als data naar een ander dataplatform gemigreerd wordt, zijn de grootste problemen te verwachten met handgeschreven code, want dan moet de code instructie voor instructie bestudeerd, waarschijnlijk aangepast en getest worden. Zeker wanneer er honderden rapporten met handgeschreven code ontwikkeld zijn met behulp van tientallen producten, dan is dit een uitermate tijdrovende en risicovolle exercitie. Dit weerhoudt veel organisaties ervan te migreren. De tweede optie leidt tot minder werk, maar de mogelijkheid bestaat dat de gegenereerde code niet werkt, omdat het nieuwe dataplatform de gegenereerde code niet ondersteunt of dat deze niet efficiënt genoeg is voor het nieuwe dataplatform. Dit laatste kan performanceproblemen veroorzaken. De ideale situatie is optie 3. In dit geval kan een migratie tamelijk snel en eenvoudig verlopen.

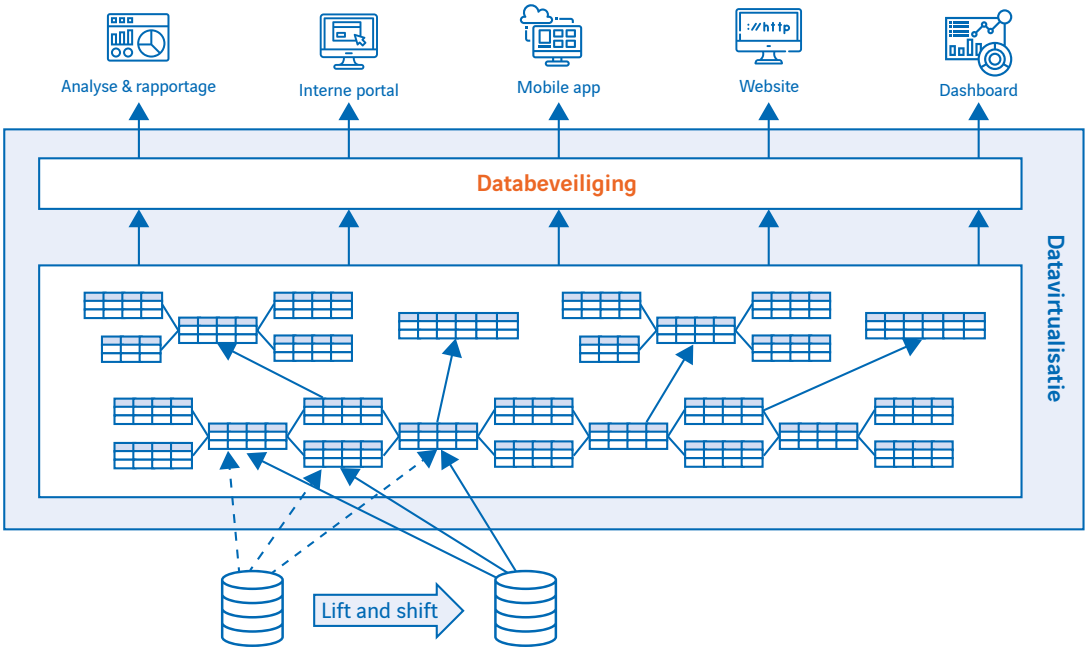
Migratie en datavirtualisatie

Een migratie, waarbij er eerst tussen de dataconsumenten en het dataplatform een datavirtualisatie-server geïmplementeerd wordt, kan soepel verlopen; zie figuur 13. Als de code handgeschreven is, is de code hopelijk te verwerken door de datavirtualisatie-server. Als de ontwikkelaars bijvoorbeeld standaard SQL gebruikt hebben, zal dit goed verlopen. Bij de andere twee opties vereenvoudigt datavirtualisatie de migratie. Een datavirtualisatie-server geeft de binnenkomende instructies van dataconsumenten niet simpelweg door, maar rafelt deze eerst uiteen en geeft de instructies daarna geoptimaliseerd aan de databron door. Dit vereenvoudigt een migratie. Het kan zelfs tot performancewinst leiden, aangezien de datavirtualisatie-server de instructies optimaliseert. Hiermee wordt ook de volledige kracht van het nieuwe dataplatform benut, zonder dat hiervoor applicatiecode aangepast moet worden.

Als data van een SQL- naar een niet-SQL-platform gemigreerd wordt, zoals Amazon S3 of Microsoft Data Lake, dan verricht de datavirtualisatie-server de benodigde vertalingen. Een handmatige migratie bij opties 1 en 2 kan veel werk met zich meebrengen. Als deze exercitie zonder datavirtualisatie uitgevoerd wordt, is het de vraag of alle rapporten wel correct blijven functioneren of dat daarvoor code aangepast moet worden. Deze migratie kan met datavirtualisatie zo uitgevoerd worden dat het merendeel van de rapporten ongewijzigd blijft doorwerken.

Voordelen van migratie met datavirtualisatie

- Een migratie naar een ander dataplatform is eenvoudiger en leidt tot minder aanpassingen bij dataconsumenten wanneer datavirtualisatie ingezet



Figuur 13: migreren naar een andere database-server met het gebruik van datavirtualisatie.

wordt. In feite wordt hier de waarde van ontkoppeling van dataconsumenten en -bronnen volledig benut.
- Bij migratie met behulp van een datavirtualisatie-server hoeft niet alle data in één keer gekopieerd te worden. Er kan voor gekozen worden om deze stapsgewijs uit te voeren waarbij groepen met tabellen gemigreerd worden. Dit betekent dat telkens na een bepaalde periode enkele tabellen worden overgeheveld. Dit vereenvoudigt mogelijk het migratieproject.

Voorwaarden

- Het in gebruik zijnde dataplatform voldoet niet meer aan de nieuwe eisen op het gebied van performance, schaalbaarheid en/of databeveiliging.
- Er bestaat interesse voor een meer zelf-managing dataplatform waardoor minder FTE's nodig zijn om deze te beheren.
- De datavirtualisatie-server moet efficiënt met het nieuwe dataplatform kunnen samenwerken. Een Proof-of-Concept is sterk aan te raden.
- De taal die door de dataconsumenten gebruikt wordt om het huidige dataplatform te benaderen moet wel door de datavirtualisatie-server ondersteund worden.
- De dataconsumenten moeten zodanig gebouwd zijn dat ze veel dataverwerking aan het dataplatform overlaten en dus niet het platform gebruiken als een 'dom' bestandssysteem.

3.12 MIGREREN VAN DATA EN APPLICATIES NAAR EEN CLOUD-PLATFORM

Inleiding

Als een migratie, zoals beschreven in de vorige paragraaf, tevens een migratie naar een cloud-platform behelst, kan ook deze met een datavirtualisatie-server afgeschermd worden. De datavirtualisatie-server zoekt zelf uit waar de data zich op dat moment bevindt. Als dit op de één of andere manier tot performanceverlies vanwege te veel netwerkvertraging, kan gekeken worden of caching van bepaalde views de snelheid kan verhogen.

Voordelen van migratie naar de cloud met datavirtualisatie

- Alle voordelen genoemd in de vorige paragraaf zijn ook hier van toepassing.

Voorwaarden

- Alle voorwaarden genoemd in de vorige paragraaf zijn ook hier van toepassing.
- De netwerkvertraging moet acceptabel zijn of op een bepaalde manier gecompenseerd worden.

3.13 BESCHIKBAAR STELLEN VAN MASTERDATA AAN VEEL GEBRUIKERS

Inleiding masterdata

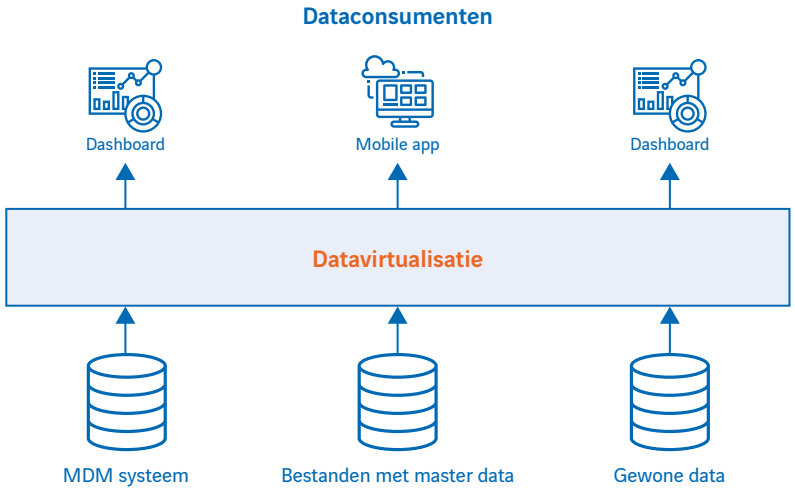
Masterdata (ook wel *stamdata*) is data die van cruciaal belang is voor een bepaald proces of een bepaalde organisatie. Masterdata komt voor in allerlei situaties en disciplines. Het gaat echter altijd om belangrijke data die nodig is voor de uitvoering van een bepaald proces. De benodigde masterdata verschilt bij een logistiek proces, een boekhoudkundig- of medisch proces. In de logistiek gaat het bijvoorbeeld vooral om kritische productgegevens zoals lengte, breedte, hoogte en gewicht, eventueel aangevuld met kleur, barcode en verpakkingsspecificatie, maar deze data is minder of niet relevant voor een medisch proces. Data opgeslagen in basisregistraties kan ook als masterdata gezien worden.

Masterdata wordt als de waarheid beschouwd, ook wel aangeduid met de term 'single version of the truth'. *Masterdata-management* heeft als doel ervoor te zorgen dat deze data door organisaties op een consistente manier gebruikt wordt. Referentiedata, zoals landcodes en grootboekcodes, wordt in dit eBook als een vorm van masterdata beschouwd.

Toegang tot masterdata is vaak beperkt tot IT-specialisten. Zij verwerken die data in interne systemen. Steeds meer gebruikers en applicaties willen ook toegang tot masterdata. Veel masterdata-management systemen hebben technische interfaces en zijn dus lastig te gebruiken. Daarnaast ligt niet alle masterdata uitsluitend in een gespecialiseerd masterdata-management systeem opgeslagen, maar ook in andere databronnen.

Masterdata en datavirtualisatie

Datavirtualisatie kan hetzelfde voor masterdata doen als het voor 'gewone' data kan: alle masterdata eenvoudig toegankelijk maken voor een grote groep toepassingen en dataconsumenten ongeacht waar en hoe het opgeslagen is; zie figuur 14.



Figuur 14: met datavirtualisatie kan de toegang tot alle masterdata vereenvoudigd en gestandaardiseerd worden.

Als externe dataconsumenten toegang krijgen tot de data van een organisatie (zie paragrafen 3.5 en 3.6), is het belangrijk dat correcte data getoond wordt. Het kan een optie zijn om ze toegang te geven tot de masterdata in plaats van de data in een databron.

Voordelen van masterdata en datavirtualisatie

- Vereenvoudiging van de toegang tot alle masterdata voor alle dataconsumenten.
- Vereenvoudiging van de integratie van masterdata en andere data voor alle dataconsumenten.
- Vereenvoudiging van het beschikbaar maken van masterdata aan externe partijen.
- Een federatieve vorm van masterdata-management kan opgezet worden waarbij de werkelijke masterdata in meerdere systemen ligt opgeslagen, maar door de datavirtualisatie-server geïntegreerd geleverd wordt.

Voorwaarden

- Er moet reeds een masterdata-management systeem aanwezig zijn waarin de masterdata opgeslagen en beheerd wordt. De datavirtualisatie-server is een toevoeging om de toegang voor een grote groep dataconsumenten te vereenvoudigen.
- Er moet voldoende vraag zijn naar masterdata om toegang via een datavirtualisatie-server te rechtvaardigen.

3.14 COMBINEREN VAN TOEPASSINGEN

In de vorige paragrafen zijn alle populaire toepassingen van datavirtualisatie afzonderlijk beschreven. Uiteraard is het goed mogelijk en bovendien sterk aan te raden om een datavirtualisatie-server voor meerdere toepassingen in te zetten. Zo kan een datavirtualisatie-server door een organisatie ingezet worden om centrale pseudonimisatie van persoonsgegevens te verwezenlijken en tevens een migratie naar een cloud platform te vereenvoudigen, of een datavirtualisatie-server kan ingezet worden om een logische data lake te ontwikkelen en tegelijkertijd te verrijken met masterdata en de self-service BI-ontwikkeling te stroomlijnen.

4. Een vergezicht voor het toepassen van datavirtualisatie

De in het vorige hoofdstuk beschreven toepassingen van datavirtualisatie zijn alle actuele toepassingen. Dit hoofdstuk beschrijft twee toepassingen die een vergezicht vormen. Beide zijn sterk gericht op toepassingen ten behoeve van de overheid.

VERGEZICHT 1

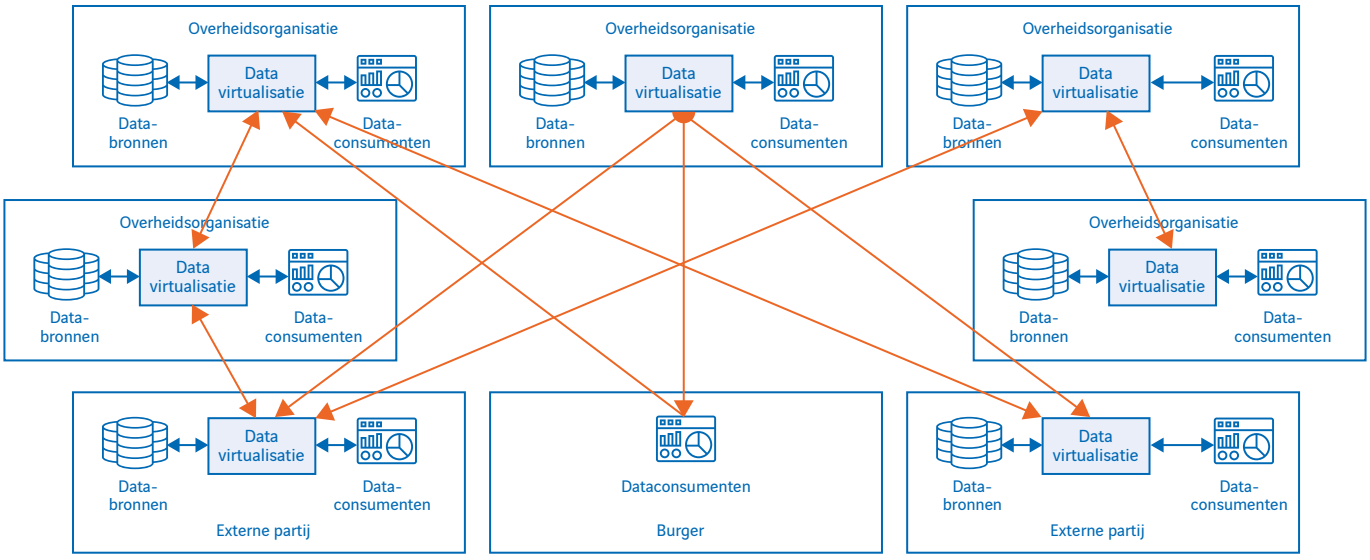
EEN NETWERK VAN DATAVIRTUALISATIE-SERVERS

In een ideale situatie wordt alle overheidsdata slechts één keer bij de aangewezen eigenaar opgeslagen. Deze stelt de data ter beschikking aan andere partijen waarbij de toegang door een datavirtualisatie-server beheerd wordt; zie figuur 15. Deze partijen gebruiken hun eigen datavirtualisatie-server om data uit de andere datavirtualisatie-servers op te halen en te integreren. Indien nodig, kan er lokaal gecachet worden. Voor alle data is er dan één eigenaar en één loket waar die data opgehaald kan worden.

Let wel: de gebruikte datavirtualisatie-servers hoeven niet van hetzelfde merk te zijn. Dit zou de gehele overheid afhankelijk maken van één leverancier.

Voordelen van een netwerk van datavirtualisatie-servers

Deze netwerkarchitectuur biedt uiteraard alle voordelen die vermeld zijn in het vorige hoofdstuk, maar die voordelen gelden voornamelijk voor één organisatie. Deze netwerkarchitectuur biedt eveneens voordelen over de breedte van de gehele overheid:



Figuur 15: met een netwerk van datavirtualisatie-servers zal de noodzaak tot dataduplicatie verminderen.

- De hoeveelheid dataduplicatie wordt geminimaliseerd.
- Het wordt eenvoudiger om zaken als het vergeetrecht te implementeren.
- Externe data is veel meer up-to-date.
- Het aantal datastromen tussen en binnen organisaties zal afnemen.
- Veel dataconsumenten zullen met correcte data werken, omdat er minder verouderde kopieën gebruikt worden.
- Het beantwoorden van ad-hoc vragen waarbij data vanuit verscheidene organisaties geïntegreerd moet worden, wordt vereenvoudigd.
- Oplossingen zoals de personal-health-train[7] kunnen op een eenvoudige manier gerealiseerd worden.

Youtuben van data

Sommige IT-specialisten zullen uiteraard beargumenteren dat deze architectuur technisch en qua performance onhaalbaar is. Hun grootste zorg zal zijn dat elke datavirtualisatie-server in dit geval met veel capaciteit ondersteund moet worden om alle vragen van alle dataconsumenten snel te kunnen beantwoorden. Er zal misschien zelfs gesuggereerd worden dat deze technologie momenteel nog niet bestaat.

Als bedrijven als YouTube erin slagen om miljoenen klanten on-demand te voorzien van films, oftewel megabytes aan data per seconde kunnen versturen, dan moet de Nederlandse overheid dit kunnen realiseren met traditionele, administratieve data. Om een idee te geven van hoeveel data YouTube verwerkt, volgen hier enkele statistieken van begin 2020[8]:

- Per dag worden er ongeveer 300.000 nieuwe video's toegevoegd.
- Per dag worden bijna 5 miljard video's bekeken.
- Per dag heeft YouTube meer dan 30 miljoen bezoekers.
- Per maand wordt er in totaal 3,25 miljard uur video bekeken.

Met een netwerk van datavirtualisatie-servers kan overheidsdata *geyoutubed* worden en zullen praktische problemen die beschreven zijn in de inleiding van dit eBook deels of geheel verdwijnen, zullen burgers en bedrijven op een eenvoudige en eenduidige manier toegang krijgen tot veel meer overheidsdata, zal data minder vaak gekopieerd worden en zal iedereen meer met up-to-date data en consistente data werken.

Wat wel ontbreekt bij de huidige versies van de datavirtualisatie-servers, is een *mission impossible*-achtige faciliteit, waarbij data die door een dataconsument opgehaald is, zichzelf na een bepaalde tijd vernietigt. In de Mission Impossible films vernietigt de opdracht zichzelf ook automatisch na gebruik.

VERGEZICHT 2

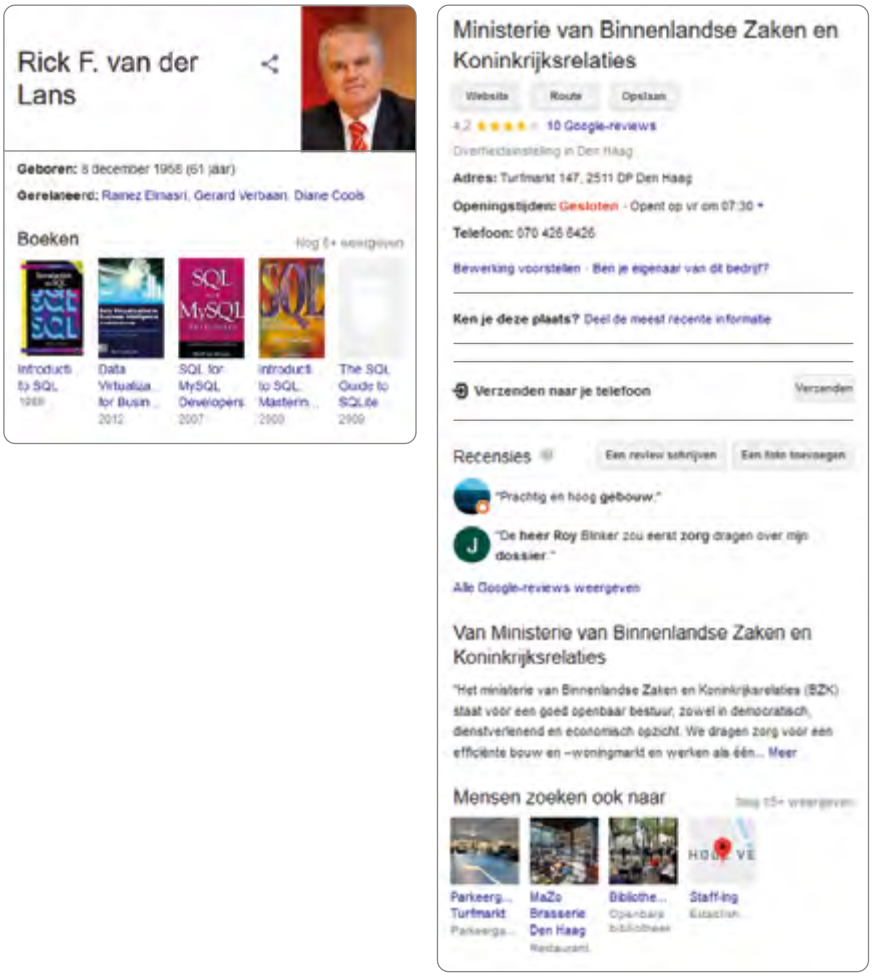
Extreme data distributie

Bij de bestaande netwerkarchitectuur wordt data bij de eigenaar opgeslagen, vergelijkbaar met de huidige basisregistraties. Het is echter verdedigbaar om te stellen dat burgers en bedrijven zelf de eigenaren van hun persoonlijke data zijn. Is een burger bijvoorbeeld niet de eigenaar van zijn of haar adres? Momenteel wordt de data over burgers in centrale databases opgeslagen. Waarom krijgt niet iedere burger een eigen plek toegekend waarin persoonlijke data opgeslagen wordt? Alle burgergegevens worden dus over 17 miljoen miniatuur databases verspreid. Deze data wordt dan ook door de eigenaar - de burger - beheerd. Met datavirtualisatie worden al deze miniatuur databases virtueel samengevoegd tot één database. Alle dataconsumenten van deze data zullen het nog als één geïntegreerde database ervaren, maar eigenlijk is de data extreem gedistribueerd.

Dit klinkt als een futuristische architectuur, maar deze is enigszins vergelijkbaar met wat Google al jaren doet. Als men bijvoorbeeld bij Google zoekt op de naam van de auteur van dit document, toont Google naast alle hits ook gestructureerde data; zie de linkerkant van figuur 16. Het lijkt alsof deze data uit één database komt, maar in feite sprokkelt Google deze data vanuit allerlei websites bij elkaar. De rechterkant van figuur 16 bevat het resultaat als er naar het Ministerie van Binnenlandse Zaken gezocht wordt.

7 Health RI; zie https://www.health-ri.nl/initiatives/personal-health-train

8 MerchDope, 37 *Mind Blowing YouTube Facts*, Figures and Statistics – 2020, februari 2020; zie https://merchdope.com/youtube-stats/



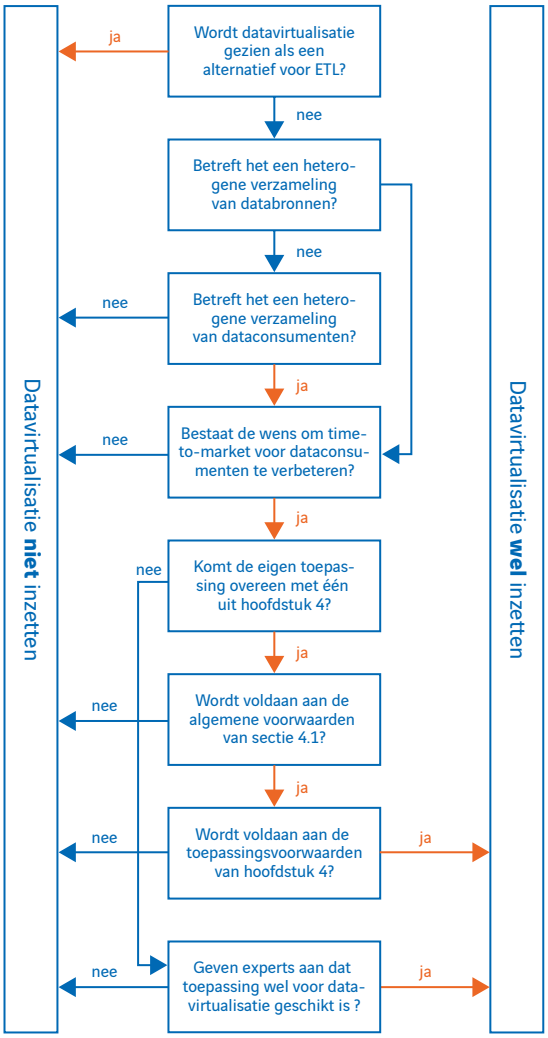
Figuur 16: gestructureerde data, door Google verzameld over de auteur en het Ministerie van Binnenlandse Zaken.

Een burger kan dan zelf, bij bijvoorbeeld een verhuizing, het adres aanpassen. Daardoor is het direct voor elke dataconsument beschikbaar. Elke burger hoort dan zelf te beheren welke data beschikbaar gemaakt wordt voor welke dataconsument. Dat zou ook veel beter passen bij AVG op het gebied van data-privacy. Een burger is dan eigenaar van zijn of haar eigen data. Als iemand niet langer wil dat een ander zijn of haar data kan inzien, worden de inzichtrechten verwijderd. Uiteraard bestaat er wet- en regelgeving die ervoor zorgt dat bepaalde data voor bepaalde instanties beschikbaar moet zijn en blijven.

5. Wegwijzer voor datavirtualisatie

Met de wegwijzer in figuur 17 kunnen organisaties bepalen of datavirtualisatie een technologie is waarmee ze hun toepassingen kunnen ontwikkelen. Er wordt op enkele plaatsen verwezen naar hoofdstuk 3 en paragraaf 3.1. Voor een volledig begrip van deze wegwijzer is het aan te raden tenminste dat hoofdstuk door te nemen.

De eerste vraag 'Wordt datavirtualisatie gezien als een alternatief voor ETL?' is toegevoegd om duidelijk te maken dat datavirtualisatie dat niet is. Er zijn veel overeenkomsten die maken dat de twee technologieën vaak vergeleken worden, maar beide zijn verschillend en worden voor andere toepassingen ingezet; zie hiervoor hoofdstuk 6.



Figuur 17: de algemene wegwijzer voor organisaties om te helpen bepalen of datavirtualisatie geschikt is voor hun toepassingen.

6. Algemene adviezen voor invoering van datavirtualisatie

Dit hoofdstuk beschrijft enkele algemene adviezen voor de invoering en het gebruik van datavirtualisatie. Ze zijn gebaseerd op de ervaringen van organisaties die reeds datavirtualisatie toepassen.

GEBRUIK DATAVIRTUALISATIE NIET OP EEN ONEIGENLIJKE MANIER, BIJVOORBEELD ALS VERKAPT ETL-PRODUCT.

Als nieuwe technologie ingevoerd wordt, ontstaat weleens de neiging om deze op dezelfde wijze in te zetten als oude technologie. Oneigenlijk gebruik van een technologie betekent altijd dat niet het optimale uit die technologie gehaald wordt. Dit geldt voor zowel functionaliteit als performance. Bij elke technologie hoort een bepaalde, passende werkwijze. Zo zijn er organisaties die datavirtualisatie toch op een verkapte ETL-manier zijn gaan gebruiken. Een passende werkwijze dient bestudeerd te worden wanneer een organisatie voor het eerst met een datavirtualisatie-server gaat werken.

ZORG VOOR EEN GEDEGEN PROOF-OF-CONCEPT.

Er zijn veel verschillende datavirtualisatie-servers op de markt beschikbaar. Aangezien het een redelijk homogene markt is, zien organisaties op het eerste gezicht geen significante verschillen. Toch zijn de producten onder de motorkap wel degelijk anders opgebouwd. Dit kan betekenen dat het ene product meer geschikt is voor een bepaalde workload dan een ander. Het kan ook zijn dat een bepaald product efficiënter samenwerkt met andere producten die een organisatie reeds gebruikt dan een ander. Het algemene advies is om te zorgen voor een gedegen *Proof-Of-Concept* (of *Pilot*) als onderdeel van het gehele keuzeproces. Een gedegen Proof-Of-Concept houdt in ieder geval in dat een realistische en representatieve omgeving gesimuleerd wordt.

SELECTEER DE GESCHIKTE BEMANNING.

Wanneer men met datavirtualisatie begint en een team van ontwikkelaars samengesteld moet worden, kies dan ontwikkelaars die graag met nieuwe software werken en met alle plezier tot diep in de nacht willen doorgaan om iets te laten werken. Het moeten ontwikkelaars zijn die van nature bereid zijn zaken tot op de bodem uit te willen zoeken. Er is aan het begin van dit soort projecten geen ruimte voor cynische ontwikkelaars of personen die het liefst blijven werken met de producten waarmee ze al jaren werken. Kies ook ontwikkelaars die vanuit modellen willen werken en overzicht hebben.

DEFINIEER EEN VERZAMELING ONTWERPREGELS VOOR DE VIEWS.

Bestudeer beschikbare literatuur over het georganiseerd opzetten van alle views. Veel organisaties die datavirtualisatie inzetten, ontwikkelen bijvoorbeeld meerdere lagen met views. Deze ontwerpaanpak wordt breed geadviseerd, maar de vraag blijft welke dataverwerkingsspecificaties in welke laag opgenomen moeten worden. Maak ontwerpkeuzes en leg deze vast als ontwerpregels. Maak deze bij alle ontwikkelaars bekend en controleer of iedereen zich eraan houdt.

HUUR SPECIALISTEN IN VOOR DE INITIËLE INSTALLATIE.

Een datavirtualisatie-server moet goed geïnstalleerd zijn om de specifieke workload van een organisatie efficiënt te verwerken. Installaties ter ondersteuning van data scientists of installaties die door externe partijen gebruikt worden, kunnen verschillen. Sommige parameters zullen bijvoorbeeld anders ingesteld moeten worden. Een 'verkeerde' installatie kan leiden tot onnodig slechte performance of schaalbaarheid. Het is raadzaam om voor de initiële installatie een specialist van het product in te huren om de datavirtualisatie-server gelijk goed in te stellen.

Over de auteur

Rick van der Lans
Onafhankelijk analist R20/Consultancy BV

rick@r20.nl

Rick van der Lans is een internationale autoriteit op het gebied van data-atchitectuur, datamanagement, datavirtualisatie en business intelligence. Hij is als adviseur, docent, spreker en auteur actief over de hele wereld. Van zijn boeken zijn internationaal meer dan 100.000 exemplaren verkocht. Rick is sinds 2017 ambassadeur van Axians Business Analytics Laren.

The best of ICT with a human touch

axians

Eemnesserweg 55, 1251 NB Laren
Tel: +31 35 53 94 490

www.axians.nl/data-architectuur