Gegevenslogistiek in de praktijk en in theorie

De Polis Papers (9): Tussen Utopia en Dystopia

René Veldwijk

In artikel 8 bespraken we een mechanisme om vanuit de Polisadministratie allerlei soorten gegevensleveringen te verzorgen op basis van dezelfde code: standen of mutaties, FTP bestandstransfer of webservices, flat file of XML, request/response of event-based, kilobyte of multigigabyte, enzovoort. In dit artikel bespreken we de logische vervolgvraag: "moet alles dat kan ook gebeuren?" Ons antwoord daarop is "Neen". En dat "Neen" confronteren we natuurlijk met de ICT-praktijk bij de vele afnemers van gegevens uit de PA.

Twee artikelen geleden keken we naar de verdeling van de gegevenshuishouding van de BV Nederland over een aantal (toekomstige) basisregistraties. We waren toen heel positief over de manier waarop dit nationale informatieskelet was uitgedacht. Hier gaan we het hebben over de spieren en de aderen die aan dat skelet zijn opgehangen, ofwel de gegevensstromen die nodig zijn voor de procesuitvoering van allerlei organisaties die afnemer zijn van de PA.
Laten we beginnen bij de visie die in deze artikelen wordt uitgedragen over hoe een basisadministratie als de PA moet worden opgezet. In artikel 2 waren we daar heel uitgesproken over: gesleep met data van systeem naar systeem is bij de enorme gegevensstromen in de loonaangifteketen een garantie voor enorme ICT-ellende. Die ellende is daadwerkelijk opgetreden en men heeft er van geleerd. Dit jaar verdwijnen er enkele overbodige en kostbare databases en als alles goed gaat komt daarna het saneren van de gegevenslogistiek tussen de werkgevers en de PA aan de orde. Ondertussen werkt de loonaangifteketen heel redelijk, maar van het kostenplaatje en de permanente staat van 'verhoogde dijkbewaking' wordt behalve een paar grote ICT-leveranciers niemand vrolijk.

Utopia versus Dystopia?

De lessen die bij het transporteren van gegevens *naar* de PA zijn geleerd, gaan natuurlijk net zo hard op voor het leveren van informatie aan afnemers *vanuit* de PA. Uitgangspunt zou moeten zijn dat de gegevens in de PA zo min mogelijk worden overgenomen in de databases van afnemende systemen maar worden geraadpleegd op het moment waarop de interne of externe afnemer daaraan behoefte heeft. Zoals we in artikel 7 stelden dat de PA geen behoefte heeft aan redundante opslag van 13 miljoen personen uit de GBA, zo zouden afnemers van PA-gegevens zich verre moeten houden van opslag van 'polisgegevens' in hun databases. De enige proactieve service van de PA zou moeten bestaan uit het versturen van mutatiesignalen aan die afnemers voor wie een mutatie op de PA een bedrijfsproces *triggert*. Zo werkt het al jaren bij de GBA en zo kan het ook voor de PA. Voor *total recall* basisregistraties zoals de GBA en de PA (zie artikel 1) is dit geen probleem en de voordelen van deze 'opvraag-wanneer-nodig' benadering zijn legio. Om te beginnen verdwijnen de problemen die voortkomen uit het raadplegen van verouderde gegevens. Ten tweede is de greep op het gebruik van de behoorlijk privacygevoelige gegevens in de PA veel groter en kan op één plek verantwoording worden afgelegd over het gegevensgebruik. En tenslotte vervallen nagenoeg alle inspanningen die nodig zijn om de consistentie en volledigheid van gegevens over de diverse databases in de hand te houden. Kortom, afnemers van PA-gegevens kunnen hoge kosten en ICT FAIL's voorkomen wanneer ze de gegevens uit de PA halen als ze die nodig hebben en verder volstaan met mutatiesignalen op basis van abonnementen. Welkom in Utopia!

> Een aandachtspunt is het woud van formaten dat gebruikt wordt om gegevens te leveren

Jammer genoeg leven we niet in de ideale wereld maar in de reële. Die wereld wordt bijvoorbeeld nog steeds gekenmerkt door beperkte mogelijkheden van ICT-hulpmiddelen. Een multigigabyte *videostream* over een netwerk mag tegenwoordig normaal zijn, een multigigabyte gegevenslevering gebaseerd op een multiterabyte database is dat zeker nog niet. Maar het echte probleem is het gangbare denken bij overheidsorganisaties. Dat denken is vanouds niet *gegevens*gericht maar *proces*gericht. En die processen lopen steeds vaker door meerdere organisaties heen. Men spreekt dan van *ketens*, zoals de strafrechtketen, de jeugdzorgketen en natuurlijk de loonaangifteketen. (Overheids)organisaties vormen de schakels in die procesketens en wat er

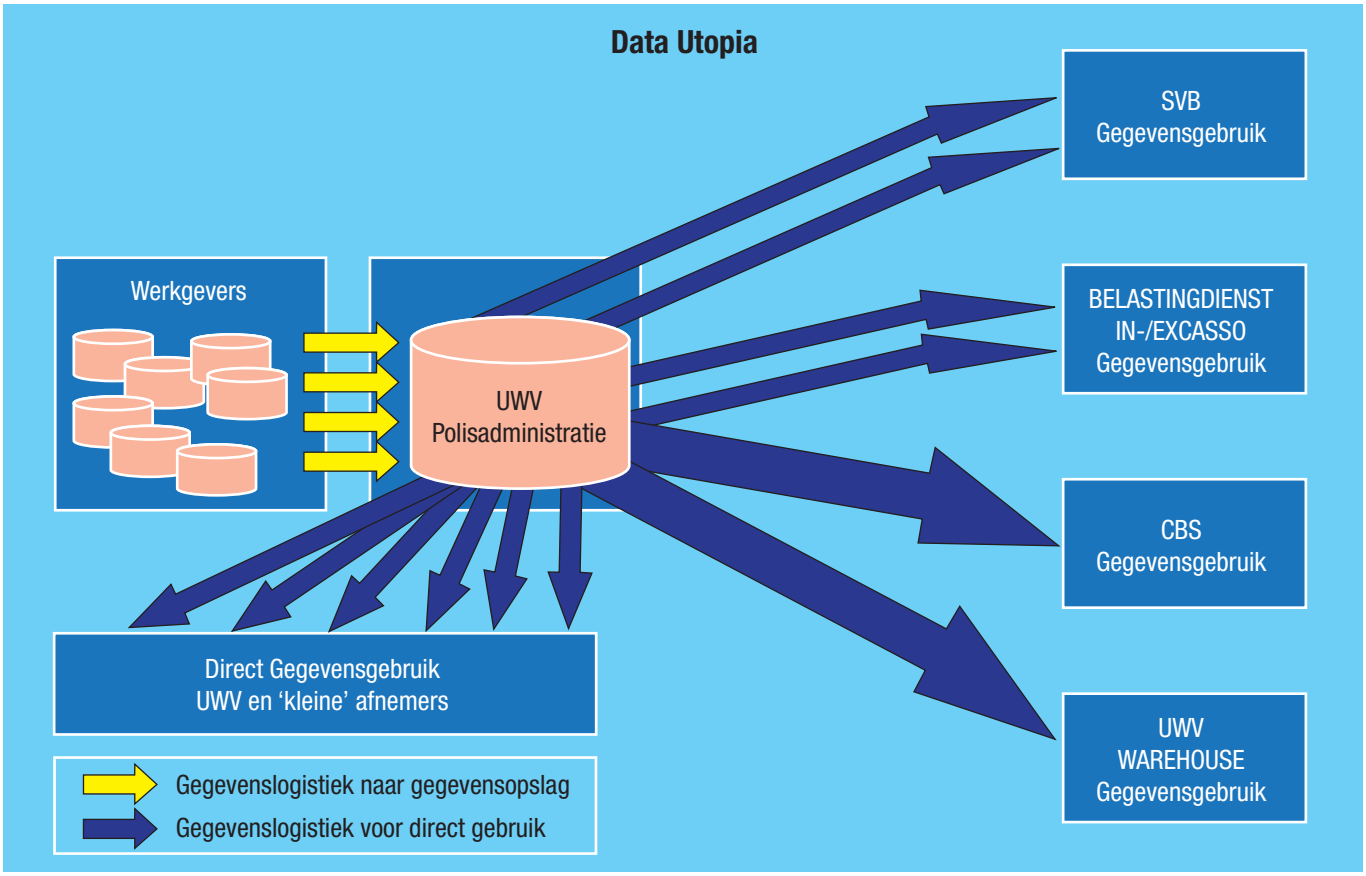
door die ketens stroomt is informatie, vaak zelfs niets anders dan informatie. Het is dan een kleine stap om een gegevensstroom te zien als analoog aan een goederenstroom die van magazijn naar magazijn wordt vervoerd. Elk van die 'gegevensmagazijnen' is natuurlijk een database, minimaal één per organisatie want iedereen wil volledige controle op zijn 'eigen' gegevens. Jammer genoeg is gegevenslogistiek iets heel anders dan goederenlogistiek. Bij gegevenslogistiek worden gegevens niet verplaatst maar gekopieerd en is elke beweging tussen gegevensverzamelingen synoniem met kostbare en soms onbeheersbare redundantie.
Die simpele waarheid die iedereen die wel eens iets *download* kent wordt, uitzonderingen daargelaten, genegeerd door ICT-architecten en ICT-beleidsmakers bij overheidsorganisaties. Een frequent voorkomende houding van de klanten van de PA laat zich samenvatten als *"Lever ons gegevens(standen) uit de PA die wij vervolgens opslaan in een of meer van onze eigen databases."* Daarna blijkt natuurlijk dat gegevens wel eens kunnen wijzigen of gewoon te laat binnenkomen en volgt de vraag *"Lever ons de gegevensmutaties"*. De eindsituatie is dan op zijn best een situatie waarin de afnemer van PA-gegevens de gegevens in een eigen database heeft zitten die vertraagd synchroon loopt met de PA. Bedenk daarbij ook dat die afnemer vaak de gegevens weer voor diverse doeleinden gebruikt en er bij de afnemer ook intern vaak wordt gewerkt met het rondpompen van gegevens tussen databases en het wordt aannemelijk dat al het jonassen met gegevens en het controleren of dat wel goed gaat veel meer geld kost dan het daadwerkelijk gebruiken van die gegevens. Kortom, de vroegere logistieke ellende tussen werkgevers en PA kan zomaar weer terugkomen tussen de PA en de afnemers maar dan in het meervoud. Welkom in het tegengestelde van Utopia: Dystopia!

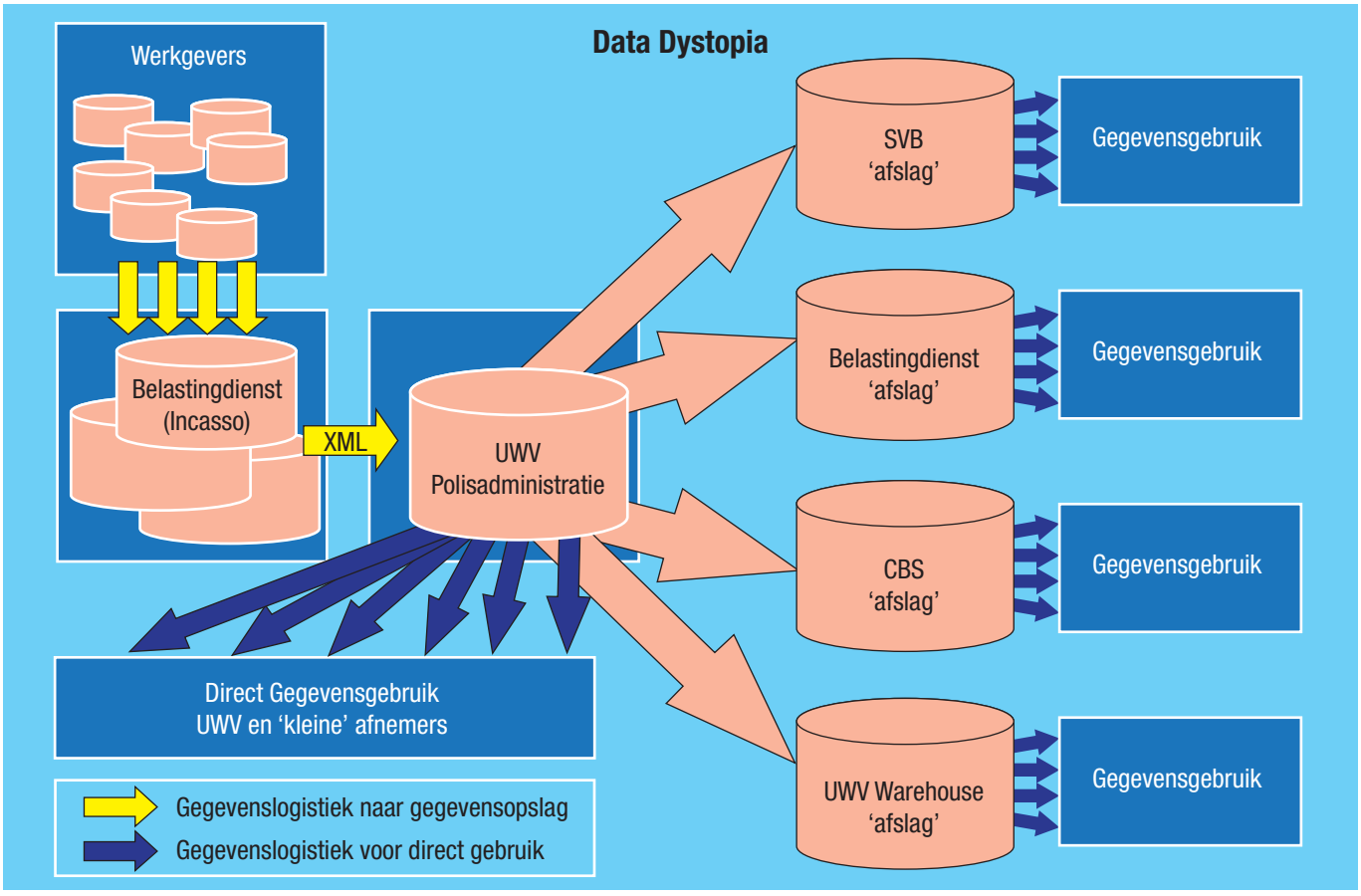
De PA en de 'Big Three-to-Five'

Om het verhaal nog complexer te maken is er ook daadwerkelijk een *business case* voor grootschalig gesleep met gegevens tussen systemen. Het feit wil namelijk dat we in Nederland het verschijnsel 'uitkering' over diverse organisaties hebben verdeeld. Al deze organisaties hebben potentieel of actueel behoefte aan massale gegevensstromen vanuit de PA. Soms is die behoefte zelfs zo grootschalig dat de mogelijke eindsituatie uitdraait op een kopie van de PA. Tel mee: we hebben het UWV voor werknemersverzekeringen, de Sociale Verzekeringsbank (SVB) voor een aantal volksverzekeringen zoals de AOW, de Belastingdienst voor allerlei toeslagen en de Gemeenten voor de Bijstand. UWV en Belastingdienstafdeling Toeslagen zijn al grootafnemers van de PA en de SVB gaat dat worden zodra de AOW afhankelijk wordt van het arbeidsverleden of de kinderbijslag inkomensafhankelijk wordt. Daarnaast moeten we ook nog rekening houden met het CBS en UWV zelf als mega-afnemer voor statistieken. Afbeelding 1 op pagina 26 laat zien hoe Utopia en Dystopia kunnen uitpakken.
De vraag is nu natuurlijk of een data Utopia haalbaar is en, zoniet, of een data Dystopia vermijdbaar is. Puur vanuit de ICT gezien is data Utopia vermoedelijk wel te realiseren. Natuurlijk is in data Utopia de belasting van de PA door afnemers groter dan in data Dystopia met al zijn lokale afslagen. Hoe groot die extra belasting uitvalt wordt natuurlijk bepaald door de ratio tussen gegevensraadplegingen en gegevenswijzigingen. Elke mutatie leidt in Dystopia tot gegevensstromen naar de databases van alle grote en een aantal kleine afnemers, terwijl in Utopia kan worden volstaan met gerichte mutatiesignalen voor sommige afnemers. Dat nadeel wordt pas goed gemaakt wanneer de gegevens bij de lokale afnemer veelvuldig worden geraadpleegd. Vermoedelijk zal het aantal raadplegingen per gegevensitem meestal beperkt zijn ten opzichte van het aantal mutaties.

> Een van de gevolgen van de XML hype is dat het gegevensbestand met allerlei hiërarchische structuren weer helemaal terug is

De gegevens in de PA zijn immers heel anders dan, zeg, de gegevens in de GBA: waar de persoonsgegevens in de GBA hoogfrequent worden geraadpleegd en laagfrequent wijzigen is dat bij de PA omgekeerd. We worden in een leven maar één keer geboren, trouwen en scheiden soms, verhuizen een handvol keren en krijgen 2,1 kinderen. De kerngegevens in de PA komen elke maand weer opnieuw binnen en ook verzekeringsgegevens en dienstverbanden wijzigen dat het een lust is. Gegevens die ouder zijn dan één of twee jaar worden nauwelijks meer geraadpleegd. Voeg daarbij het feit dat de PA binnenkort drie jaar zonder veel optimalisaties draait (zie artikel 6) en eind 2010 zijn maximale omvang bereikt en de conclusie is dat Utopia zelfs voor een monsterlijk grote en volatiele database als de PA technisch haalbaar is voor afnemers als SVB en de Belastingdienst. Voor de statistiekafnemers ligt dat wat anders want die willen de gegevens uit de PA grootschalig en veelvuldig vergelijken met gegevens uit andere bronnen. Daarmee wil je de PA dataserver niet belasten, ook omdat optimalisatie van query's over databases heen nog steeds een onopgelost probleem is. (Vervang maar eens een join over twee tabellen met een *table scan* plus een webservice-aanroep en klok het verschil.) Heel erg is dat niet omdat de twee 'statistiekafnemers' wekelijks een platte dump krijgen van de mutaties op de PA, wat zo ongeveer de meest efficiënte vorm van gegevens leveren is. Er wordt feitelijk één levering gedaan die naar beide afnemers wordt ge-FTPt. Simpeler en minder belastend kan niet.
Het echte probleem is dus niet technisch maar bestuurlijk. Naast de al genoemde ongelukkige visie op gegevenslogistiek die data Dystopia ziet als data Utopia, zijn bij data Utopia de grote afnemers vergaand afhankelijk van de beheerder van de PA. Wil je als bestuurder van de Belastingdienst of SVB afhankelijk zijn van de beschikbaarheid en de prioriteit van een systeem dat bij een andere organisatie staat? Natuurlijk is het antwoord op die vraag





Afbeelding 1: Twee visies op gegevenslogistiek. Data Utopia (boven) en Data Dystopia (onder) (sterk versimpeld).

negatief en dat is alleszins begrijpelijk. De ervaringen met de loonaangifteketen in 2006/2007 hebben laten zien dat de schuld van 'ketenfalen' vaak eenzijdig wordt gelegd bij de organisatie aan het einde van de keten. Als ik morgen de ICT-baas word van de belastingdienst of van SVB dan wil ik ook mijn eigen data Dystopia. Samengevat: bestuurders zijn misschien wel goed maar niet gek.

Een haalbaar perspectief

Als Utopia zoals gewoonlijk niet haalbaar is en Dystopia moet worden vermeden, wat is dan wijsheid? Welnu, allereerst dienen de grote afnemers alleen mutaties geleverd te krijgen en geen standen. De SVB die binnenkort voor het eerst wordt aangesloten op de PA krijgt niet elke maand de kleine 20 miljoen actuele inkomstenverhoudingen maar alleen de nieuwe, de gewijzigde en de verwijderde. Niet alleen wordt de gegevensstroom daarmee een factor 100 kleiner, maar de afname wordt ook schaalbaar: als SVB later de gegevens per week, per dag of per seconde geleverd wil krijgen dan wordt de gegevensstroom navenant kleiner. Overigens geldt dat voordeel niet zo sterk voor alle gegevens in de PA. Als SVB in tweede instantie ook de maandelijkse inkomensgegevens gaat afnemen dan worden de gegevensstromen alsnog enorm. Ook dan is logistiek op basis van mutaties echter nog steeds verre te prefereren boven logistiek op basis van standen.

Misschien nog wel belangrijker is dat de gegevens zo min mogelijk transformaties ondergaan bij transport van bron naar afnemer. Een van de gevolgen van de XML hype is dat het gegevensbestand met allerlei hiërarchische structuren weer helemaal terug is. Alle gegevensleveringen die al op de oude PA waren ontwikkeld kennen dergelijke complexiteiten, met als gevolg dat zowel aan de kant van de PA als aan de kant van de afnemer allerlei fijne in- en uitpakoperaties moeten worden doorgevoerd. Al die complexiteit is kostbaar in bouw, test en uitvoering en levert niet veel nuttigs op. Het alternatief is de levering aan het CBS en de UWV warehouse-afdeling die één-op-één lopen met de structuur van de PA: elke tabel in de PA levert één bestand met mutaties op voor de afnemer. Zo'n 'slimme dump' is op een *total recall* database als de PA zeer eenvoudig te bouwen en te onderhouden en belast de PA server minimaal. Het nadeel is natuurlijk dat deze afnemers de kleine wijzigingen in de structuur van de PA die er elk jaar zijn moeten volgen. De praktijk van gegevensleveringen aan de Belastingdienst, een andere mega-afnemer, leert dat dit nadeel vele malen kleiner is dan alle complexiteit en kosten die verbonden zijn aan gegevenstransformaties bij de bron en bij de afnemer.

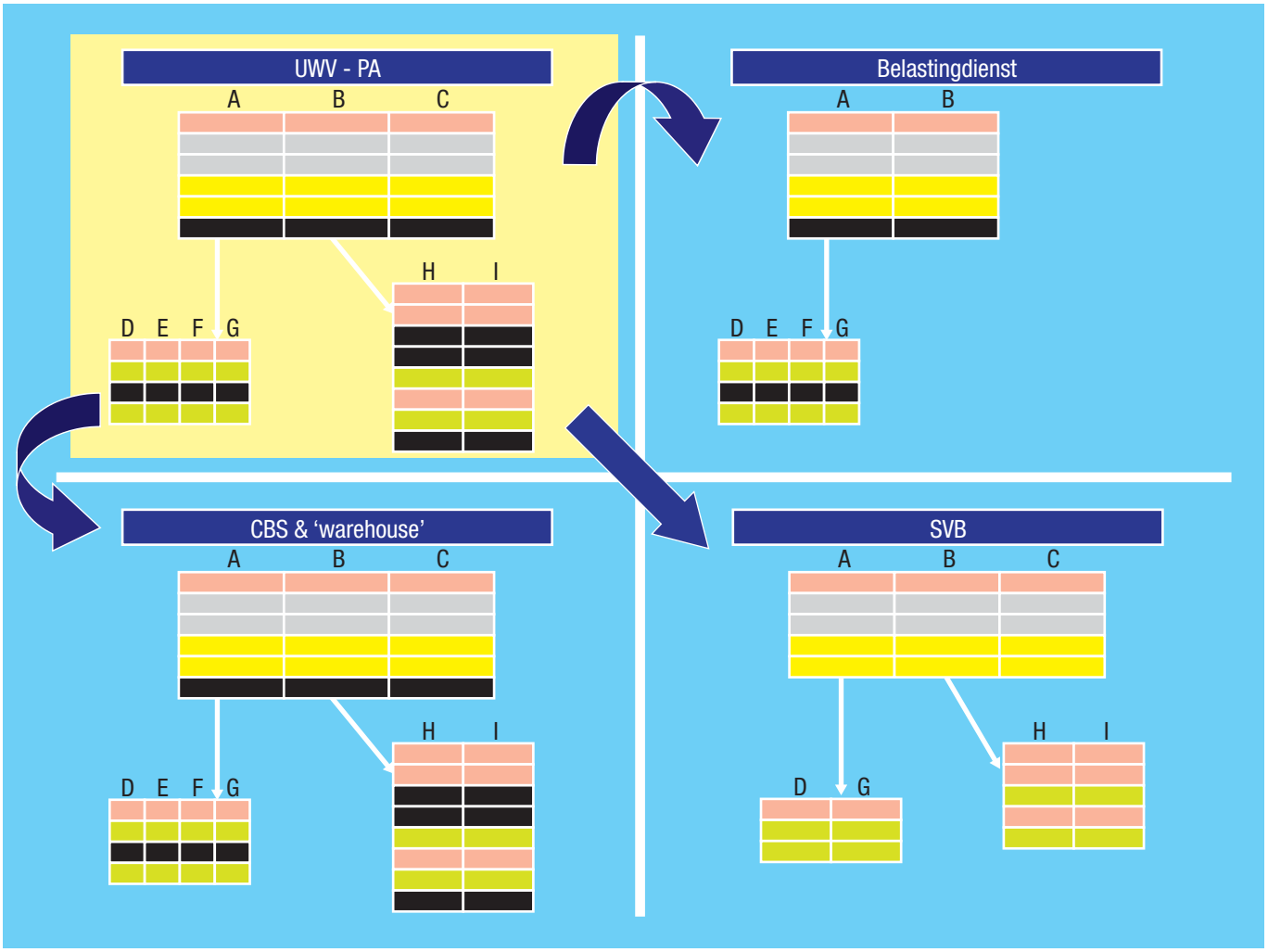
Hier dringt zich een idee op: als er nu elke week een totaal mutatiebestand wordt gemaakt dat aan CBS en het UWV warehouse wordt geleverd, waarom dit dan niet gewoon doorgestuurd naar de Belastingdienst, de SVB en andere potentiële mega-afnemers? Inderdaad zou daarmee de PA enorm worden ontlast en zouden de kosten van de exploitatie van de PA navenant worden verlaagd, maar het gevolg zou dan wel zijn dat de mega-afnemers ook gegevens krijgen die ze niet nodig hebben. Een afnemer als de Belastingdienst is natuurlijk geen incassobureautje maar dat moeten we toch niet zomaar willen. Een eenvoudige tussenvorm zou dan kunnen zijn om de mega-afnemers de PA-structuur op te leggen maar op maat projecties en selecties over de PA-tabellen uit te voeren. In afbeelding 2 is te zien hoe dat werkt.

We zien dat vanuit de PA een complete kopie wordt geleverd aan CBS en het UWV 'warehouse'. De belastingdienst krijgt in dit hypothetische voorbeeld wel alle objecten maar niet alle gegevenselementen (projectie) en de SVB krijgt noch alle objecten noch alle gegevenselementen (projectie en selectie). Het valt eenvoudig in te zien dat op deze wijze alle grote afnemers de voor hen relevante gegevens (nagenoeg) compleet op maat aangeleverd kunnen krijgen. Uiteraard kunnen dergelijke leveringen op basis van één betrekkelijk eenvoudig programma worden aangemaakt op basis van de informatie in de datadictionary (zie artikel 3). En als we tenslotte afspreken dat alle grote afnemers 'hun' mutaties met dezelfde frequentie geleverd krijgen dan kan de PA met behulp van één run worden uitgelezen voor de mega-afnemers. We zitten nu voor die afnemers toch weer vrij dichtbij Utopia en misschien wel dichterbij dan we denken. Want stel eens dat op een dag besloten wordt om het geheel van uitkeringen te herverkavelen over bestaande of nieuwe uitkeringsorganen dan wordt het herverkavelen van de gegevensverwerking een fluitje van een cent in plaats van het normale ICT-hoofdpijndossier.

De PA en de 'kleine' afnemers

Naast de hiervoor besproken buitencategorie afnemers is er natuurlijk nog een grote en gemêleerde groep van kleinere afnemers waarvan de interesse in PA-gegevens zich beperkt tot tienduizenden of honderdduizenden inkomstenverhoudingen. Denk aan gemeenten, pensioenfondsen, deurwaarders, zorgverzekeraars, de uitkeringsdivisie van UWV zelf, de burgers (met DigiD), enzovoort. Een aantal van die afnemers benadert de PA alleen of voornamelijk door middel van webservices waarna de gegevens worden gepresenteerd op een inkijkscherm. Geen probleem daar, al wordt het wellicht druk op de dag dat iedereen zijn inkomensgegevens en uitkeringsrechten in detail kan bewonderen. (Weet u hoeveel uw WW-uitkering uitpakt als u morgen wordt ontslagen?)

Een ander aandachtspunt is het woud van formaten en levermechanismen dat gebruikt wordt om gegevens te leveren en dat is overgenomen van de eerdere PA. Voor nerds is het allemaal een feest maar daarmee houdt het ook op. Zo wordt er soms gewerkt met *message queuing* (MQ Server van IBM). Volgens de aanhangers levert dat *guaranteed data delivery*, mits natuurlijk de applicatieprogrammatuur die de gegevens verder verwerkt *bugfree* is en er geen handmatige exploitatiestappen zijn – NOT. Dus ontdekt iedereen dat er niets zo beheersbaar en performant is als een batchprogramma. En omdat de loonaangifteketen



Afbeelding 2: Projecties en selecties als basis voor bulksynchronisatie.

natuurlijk niet vraagt om gegevensverwerking in *real-time* kun je met één batchrun per dag nagenoeg altijd uit de voeten.
Ook een aandachtspunt is het bundelen van veel leveringen aan 'kleine' afnemers tot één levering. Zo moet de PA dit jaar honderden pensioenfondsen en pensioenverzekeraars van gegevens gaan voorzien. De oplossing is natuurlijk een batchrun met een slim mechanisme om de uitvoer uit te splitsen naar de belanghebbende afnemer. In artikel 6 zagen we al dat zo'n benadering een geweldig positieve invloed heeft op de doorlooptijden en systeembelasting. Alle maatregelen tezamen bieden perspectief op leveringen met de omvang van vele, vele Terabytes aan netto gegevens (dus exclusief XML en andere overhead) per jaar bij een min of meer constante inkomende gegevensstroom van minder dan een Terabyte per jaar.
Het meest problematisch zijn ook hier tenslotte natuurlijk de gevallen waarin gegevens uit de PA worden overgetankt in eigen databases van afnemers. Alle problemen die we noemden met betrekking tot gegevenskwaliteit, autorisatie en logging spelen bij dergelijke afnemers volop. Het zal de BV Nederland en UWV als houdster van de PA de kop niet kosten, maar er zijn nog aardig wat kleine Dystopia'tjes die kunnen worden opgeruimd of voorkomen. In combinatie met de slimme standaardisatie van leveringen die we in het voorgaande artikel bespraken valt er veel gegevenskwaliteit te winnen en geld te besparen. Met de sinds kort sterk toenemende belangstelling voor privacybescherming en de nog grotere behoefte aan kostenbesparingen is er een wereld te winnen door afnemers aan te sporen om de PA-gegevens op te halen wanneer ze die nodig hebben en ze niet op voorraad te gaan aanhouden. Te vrezen valt echter dat het nog wel een paar mislukte projecten en bezuinigingsrondes zal vergen voordat dit inzicht is doorgedrongen bij alle afnemers van de PA.

In het volgende artikel sluiten we deze serie af met een melancholieke blik op wat er is bereikt en een kritische blik op het vele dat open is blijven staan. Daarnaast stellen we ons de vraag of in een tijd waarin bijna alle grote overheids ICT-projecten mislukken en er 35 miljard moet worden bezuinigd er eigenlijk wel een business case is voor een hyperdatabase als de PA.

René Veldwijk is partner bij FAA Partners, onderdeel van de Ockham Groep.