Gebruiklogging op een terabyte-plus basisregistratie

De Polis Papers (4): Grote Broer in de spiegel

René Veldwijk

In het voorgaande artikel beschreven we een generiek mechanisme waarmee de afnemers van gegevens op maat kunnen worden geautoriseerd. Maar autorisatie is slechts één kant van de medaille.
Tegenover het autoriseren van toegang staat het vastleggen van het daadwerkelijke gegevensgebruik: logging. Repressieve logging is met name bij een ICT-basisvoorziening zoals de GBA, het Elektronisch Patiëntendossier of de Polisadministratie veel belangrijker dan preventieve autorisatie. En dus logt de Polisadministratie alles wat er aan gevoelige gegevens de deur uitgaat: van de meest eenvoudige opvraging van uw inkomensgegevens op een inkijkscherm tot een multigigabyte bestandslevering van 'tout' Nederland.

Voorjaar 2009. Een gewone doordeweekse dag bij de Polisadministratie. Buiten woedt de kredietcrisis en op de kantoren van het UWV wordt overgewerkt om die bij te houden. Ook de gemeentelijke sociale diensten hebben het druk. Samen vragen medewerkers van deze organisaties vandaag de persoonsgegevens, arbeidsrelaties en inkomens van ruim 40.000 personen op via een webservice. Enkele medewerkers van UWV en Belastingdienst zijn bezig met het analyseren van loonaangifte-gegevens met behulp van geavanceerde inkijkschermen. Ook wordt binnen UWV vandaag mondjesmaat gebruik gemaakt van de nog experimentele mogelijkheid om geautomatiseerd uitkeringsbedragen te berekenen. Met een andere webservice worden vandaag de gegevens van 2.000 personen opgevraagd door gerechtsdeurwaarders op zoek naar mogelijkheden om loonbeslag te leggen. Een paar dagen geleden zijn er van 30.000 personen inkomensgegevens gezonden naar het CAK dat daarmee facturen voor AWBZ bijdragen maakt. Voor morgen staat een levering van de inkomensgegevens van 250.000 personen aan het Inlichtingenbureau gepland dat op zoek is naar uitkeringsfraudeurs. Een analist/programmeur is vandaag voor de Belastingdienst bezig met een ad hoc informatieverzoek naar de verdiensten van 12.000 sexwerkers. Vrijdag draait een levering die alle mutaties van de laatste week doorgeeft aan het CBS. En komend weekend gaat er weer voor de Belastingdienst een proeflevering de deur uit met de inkomens over 2008 van miljoenen mensen die voor het eerst een vooraf ingevuld belastingaangifteformulier krijgen.

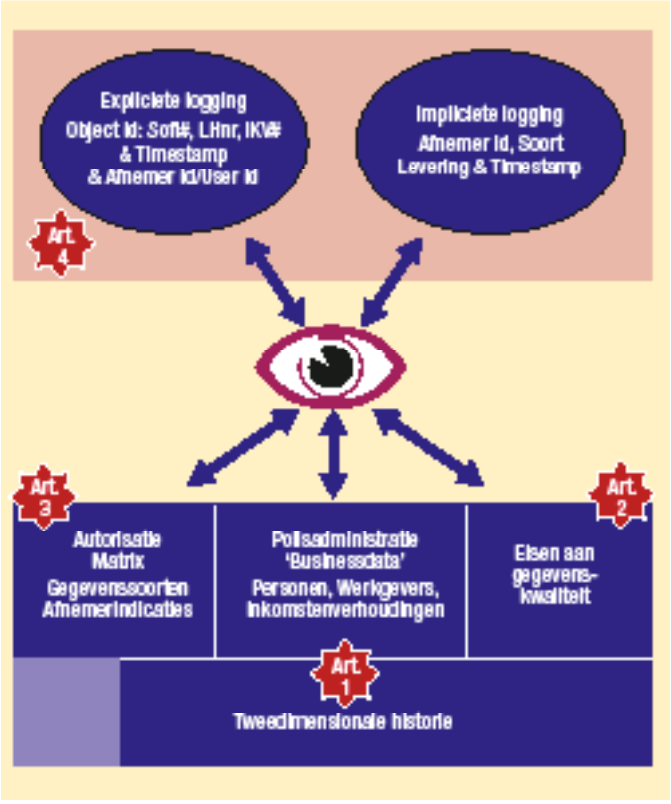
Gegevenskluis of open huis?

Het voorgaande is een even kenmerkende als onvolledige greep uit de gegevensleveringen die vanuit de Polisadministratie worden gedaan. En de PA is pas net operationeel. Wellicht komen er nog duizenden medewerkers van de Belastingdienst als inkijker bij en ook uzelf kunt later dit jaar uw eigen gegevens deels inzien. Verder staan er nog circa 600 pensioenfondsen op het punt om aan te haken als afnemer. De PA is dus voor heel veel organisaties en mensen een rijke bron van gevoelige gegevens waarvan je niet wilt dat die in verkeerde handen vallen. Op papier is dat allemaal goed geregeld. Alle inkijkfaciliteiten en gegevensleveringen hebben een deugdelijke wettelijke basis die zorgvuldig is vertaald in berichtspecificaties en programmatuur. Autorisaties voor gegevensgebruik kunnen flexibel worden ingeregeld met de autorisatiekubus van het vorige artikel. Alles is dus geregeld? Nou nee. Minstens 70.000 medewerkers van allerlei organisaties beschikken over een inkijkfaciliteit direct op de Polisadministratie en al die mensen kunnen op een handjevol VIP's na iedereen bekijken die in de Polisadministratie geregistreerd staat – bijna heel Nederland dus. Hoe ze heten. Waar ze wonen. Waar ze werken. Hoeveel ze verdienen. Dat ze dat kunnen is logisch. Rijkman G. kan best zijn rekeningen niet betalen en Sophie H. kan morgen zomaar arbeidsongeschikt worden. Gegevens *kunnen* bekijken is echter niet hetzelfde als gegevens *mogen* bekijken en ook dat is in de procedures goed geregeld – in elk geval bij UWV. Maar daarmee houdt het voor wat betreft de preventieve maatregelen ongeveer op. Alle aandacht die de betrokken organisaties besteden aan procedures, bewustwording en ontwikkeling van normbesef kan niet voorkomen dat er ergens een nieuwsgierig Aagje, een Beun de Haas of een Boris Boef verboden gegevens opvraagt. Omdat preventie maar beperkt mogelijk is, is repressie noodzakelijk en de basis daarvan is het registreren van gegevensgebruik: "wie deed wanneer wat met wiens persoonsgegevens". Logging maakt het mogelijk om achteraf na te gaan of er binnen de procedures is gewerkt en misbruik te bestraffen. Logging maakt het ook mogelijk om vreemde patronen in opvragingen te controleren. Maar het belangrijkste van alles: wie weet dat hij wordt gecontroleerd houdt zich aan de regels. Repressie is preventie, vooral wanneer echte preventie door middel van autorisatie niet goed kan.

Een oude vriend: Total Recall

Logging is in veel systemen met gevoelige gegevens niet goed geregeld, vooral als die systemen veel gebruikers, veel data en zware transacties kennen. Logging en het gebruik van logging-informatie is haast overal de sluitpost. Neem als verontrustend voorbeeld informatie over banksaldi van personen en bedrijven. Stel, u hebt een zakelijk conflict met mij en besluit om beslag te leggen op mijn bankrekeningen. Uw advocaat vertelt u dan binnen een paar uur bij welke banken ik rekeningen heb en wat de saldi zijn. Uw advocaat heeft via een informatiebureau van bijbeunende bankmedewerkers deze gevoelige informatie gekregen. Ja, het is corrupt en crimineel, maar ook dagelijkse praktijk. Grote systemen zijn niet te beveiligen, zeker niet in de *access anywhere* wereld van het Internet. Jammer maar helaas, toch? Wat ons betreft kan dit niet, zeker niet bij de overheid die ons burgers dwingt om gevoelige persoonsgegevens te verstrekken. Het uitgangspunt voor de PA is daarom dat we *alle* verstrekkingen van gevoelige gegevens tot in het kleinste detail moeten kunnen reconstrueren. Het probleem is natuurlijk dat dit technisch nagenoeg ondoenlijk is, precies als bij al die andere grote systemen. We praten bij een systeem als de Polisadministratie zomaar over een paar honderd miljoen individuele gegevens-leveringen per jaar en bij de voorziene groei van het gebruik komt dat getal straks misschien wel boven de miljard. Hoeveel mag privacybescherming kosten? Gelukkig hoeven we die vraag niet te beantwoorden, want volledige logging is in de PA helemaal geen groot probleem, met dank aan het begrip *Total Recall*. Eerder in deze serie zijn we dit begrip al twee keer tegengekomen. In artikel 1 gebruikten we de term voor het kunnen terugvinden van alle mutaties op de gegevens en in artikel 2 breidden we dat begrip uit van de Polisadministratie naar de wereld van XML loonaangifteberichten. Nu zetten we de laatste stap: we willen een Polisadministratie die voor gevoelige gegevens *volledig* kan vastleggen wie welke gegevens hebben ontvangen of bekeken, wanneer dat was en hoe dat gebeurde. Driemaal is scheepsrecht.

Laten we even teruggaan naar de inleidende paragraaf van dit artikel. We zien daar dat er verschillende soorten gegevensleveringen zijn. Er zijn periodieke bestandsleveringen gebaseerd op afnemerindicaties (vorig artikel) en ad hoc bestandsleveringen, gebaseerd op een door de afnemer geleverd bestand met bijvoorbeeld Sofinummers. Er zijn ook vormen van inkijk op individuele personen, werkgevers of inkomstenverhoudingen met een request/response karakter of (in de toekomst) een spontaan, eventgedreven karakter. Het loggen van request/response en spontane leveringen is goed te doen. Bij de voorzienbare groei blijven we vermoedelijk beneden de tien miljoen opvragingen per jaar en kunnen we volstaan met het voor elke actie wegschrijven van een record in een loggingtabel. We loggen het id van het geleverde object, het verwerkingstijdstip en de identificatie van de afnemende organisatie en, indien mogelijk, de identificatie van de opvragende gebruiker. Daarmee zijn we er ook, want *welke* gegevenssoorten en gegevenswaarden we precies hebben geleverd weten we. De polisadministratie is immers Total Recall: we weten wat we voor het moment van opvraging geregistreerd hadden, wat de kwaliteitseigenschappen waren en voor welke objecten en gegevenssoorten de afnemer geautoriseerd was op het moment van opvraging. Zo combineren we alles dat we in de drie voorgaande artikelen hebben besproken in één eenvoudige maar extreem krachtige loggingvoorziening. Die voorziening is nog generiek ook, dat wil zeggen dat deze automatisch wordt aangetrapt voor elke webservice of spontane queue levering die wordt ontwikkeld. Het enige dat voor een nieuwe levering moet gebeuren is wat parameters aangeven: welke afnemende organisatie, welke service id, welk logging-object? Easy does it. Bij bulk-bestandleveringen doen we iets vergelijkbaars. We leveren bijvoorbeeld een aantal keer per jaar de inkomensgegevens van honderdduizenden mensen aan pensioenfonds XYZ en we gaan echt niet jaarlijks voor die ene middelgrote afnemer miljoenen loggingrecords wegschrijven. Dat kost gewoon teveel opslagcapaciteit en beïnvloedt de doorlooptijd van de leveringen teveel. Maar gelukkig hoeft dat ook helemaal niet. De PA kent een submodel waarin de afnemers en de (soorten) leveringen worden geregistreerd. In dat model loggen we de momenten waarop bulkleveringen hebben plaatsgevonden, en daarmee leggen we *impliciet* vast welke personen, bedrijven of inkomstenverhoudingen we hebben uitgeleverd. Als ik wil nagaan of, en zo ja welke, gegevens van Pietje Puk zijn geleverd aan pensioenfonds XYZ, dan kijken we of het fonds op



Afbeelding 1: Total Recall logging in de Polisadministratie.

het moment van levering een indicatie op Puk of op zijn werkgever had uitstaan. Zo ja dan is de rest net als bij een request/response opvraging.

Eye in the sky...?

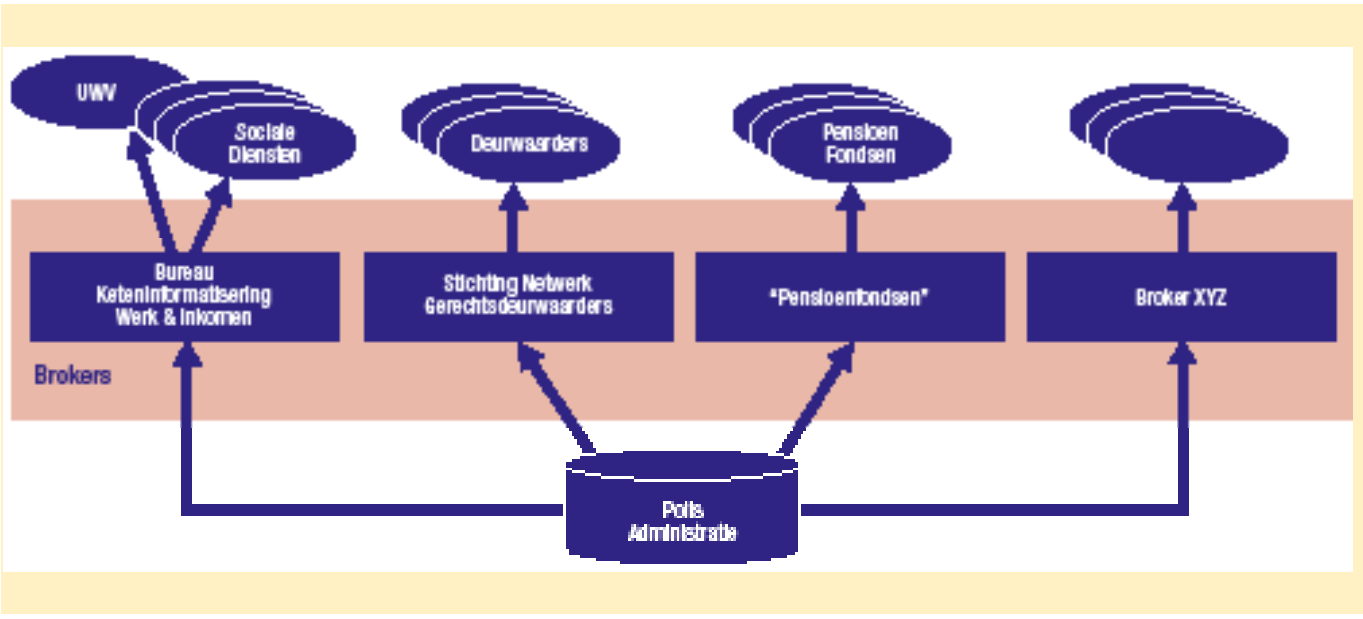
Net als in het voorgaande artikel zien we ook hier weer hoe eenvoudig het is om schijnbaar *high tech* kunstjes uit te halen als je de opzet van je systeem in vrij combineerbare componenten opdeelt. Hier zitten we echter zomaar op vijf of zes dimensies en dat tekent wat lastig uit op tweedimensionaal papier. Afbeelding 1 doet desondanks een poging om de essentie van Total Recall logging te vangen.

Het principe van Total Recall gegevensopslag dat we hebben besproken in het eerste artikel en natuurlijk ook in het 'Tijd in de Database' boekje (download van www.dbm.nl) is de basis van het 'alziend oog' waarmee we de PA hebben uitgerust. De dubbele tijdlijn ligt onder de hele PA, met één vervelende uitzondering: de datadictionary waarop de hele PA is gebaseerd kent nu net géén tijdsdimensie. Dat betekent dat we een klein probleem hebben wanneer we willen weten welke gegevens*soorten* we hebben geleverd. Als aan een service een additionele gegevenssoort wordt toegevoegd, dan weten we niet dat dit gegeven eerder niet kon worden uitgeleverd. Hoewel we dat in de praktijk gemakkelijk kunnen oplossen door elke wijziging op een service te definiëren als een nieuwe service, is dat niet de te volgen weg. Wat nodig is, is dat ook de datadictionary tijdbewust wordt, niet alleen voor de transactietijdlijn ("wanneer werd de dictionary gemuteerd") maar ook voor de geldige tijdslijn ("voor welk tijdinterval geldt de informatie in de dictionary"). In een omgeving als die van de PA waar definities voortdurend wijzigen en structuurwijzigingen schering en inslag zijn, is dit geen overbodige luxe. Totdat we deze tekortkoming hebben weggenomen blijft ons alziend oog een beetje troebel zien.

... of een blinde muur?

Er is nog een probleempje met ons streven naar een alziend oog. Tussen de PA en de gebruikers zit in een aantal gevallen waarin gewerkt wordt met webservices een muur in de vorm van een gegevens*broker*, zie afbeelding 2.

Op dit moment kennen we twee brokers, het Bureau Ketenintormatisering Werk en Inkomen (BKWI) en de Stichting Netwerk Gerechtsdeurwaarders (SNG), maar daar zal het niet bij blijven. Elke categorie afnemers heeft zijn eigen brokerorganisatie en dat is goed uit te leggen. UWV is bijvoorbeeld verplicht om de gerechtsdeurwaarders informatie te verstrekken die noodzakelijk is voor het leggen van loonbeslag, maar de uitvoering daarvan is belegd bij de SNG. Niet UWV maar SNG is verantwoordelijk voor het inregelen en bewaken van het gebruik van de PA door haar leden. SNG deelt autorisaties uit en SNG controleert of laat controleren dat de PA alleen voor de wettelijke taak wordt gebruikt. Alles is netjes geregeld in een contract met UWV waarin onder meer wordt geregeld dat er jaarlijks extern wordt gecontroleerd door een EDP auditor. Tot zover is alles goed. Het vervelende is alleen dat een broker niet alles *kan* bewaken. Stel dat een niet integere Sociale Dienstmedewerker de gegevens van allerlei mensen in zijn buurt gaat uitvragen, of dat een deurwaarder stelselmatig alle Iraniërs in Nederland gaat opvragen. Een broker is niet in staat om dit soort zaken boven water te halen, tenzij hij een complete database met doorgeleverde gegevens opzet en dat is wel het laatste wat we willen. Nee, de broker hoort voor zijn controletaken ook gebruik te maken van de gegevens in de PA. Dat is geen probleem, ware het niet dat de meeste brokers niet de informatie over de opvragende gebruiker meeleveren in de webservice request. De PA weet dus alleen dat de gegevens van Pietje Puk op tijdstip X zijn opgevraagd door een deurwaarder maar niet door welke. De SNG weet dat wel en door de gegevens van de PA en die van SNG te combine-



Afbeelding 2: Gegevensbrokers rond de Polisadministratie.

ren zijn we toch weer Total Recall, maar dat is een schrale troost. Het probleem is overigens eenvoudig op te lossen en vermoedelijk zal dat ook gebeuren, maar tot die tijd kijken we tegen een blinde muur aan. Zonder in details te treden kunnen we in elk geval melden dat het combineren van logginggegevens van brokers met die van de PA werkt. Iedere gebruiker van de PA die zijn boekje te buiten gaat is gewaarschuwd, ook omdat op een later moment kan worden teruggekeken.

Overigens hebben die brokers in ieder geval een belang om goed met de PA en elkaar samen te werken. Kijk nog eens naar afbeelding 2 en trek het plaatje door naar de toekomst. Moeten we naar een architectuur met tientallen brokers die allemaal ongeveer hetzelfde doen? Het lijkt ons dat brokerfunctionaliteit deels moet worden gebundeld over de individuele brokers heen, ook al is dat niet het probleem of de verantwoordelijkheid van UWV. Wanneer dat gebeurt zal ongetwijfeld ook de logging onderdeel worden van een generieke afnemersvoorziening.

En wat doen de Goden?

Alles wat we tot nu toe hebben gezegd over logging heeft betrekking op het werk van gewone stervelingen: eindgebruikers die ergens in het land achter hun scherm zitten of UWV-medewerkers die de exploitatie doen. Ieder systeem kent echter mensen die toegang hebben met programmeerhulpmiddelen tot de data. Bij de PA zijn dat SQL*Plus en vooral Toad. Die mensen kunnen geen gegevens muteren maar ze zijn wel in staat om gegevens te raadplegen en er in het ergste geval met complete bestanden vandoor te gaan. Dat is in elke omgeving een groot risico, maar bij de PA is dit probleem extra groot, omdat er tot nog toe veel snelle ad hoc leveringen worden gedaan en we niet goed uit de voeten konden met testdata vanwege de kwaliteitsproblemen in de loonaangiftegegevens (zie artikel 2 in DB/M 2). Binnenkort zal de situatie vermoedelijk zijn genormaliseerd, maar ook dan zijn er net als overal enkele mensen die met SQL tools bij de data kunnen komen. Naast allerlei organisatorische maatregelen van UWV en de systeembeheerder (IBM) voorziet de PA daarom in een mechanisme dat alle individuele SQL statements logt die iets doen met gegevens van personen, werkgevers of inkomsten(verhoudingen). Voor elke tabel of view die een domein Sofinummer, Loonheffingnummer of Inkomstenverhouding bevat, wordt op basis van de datadictionary een loggingprocedure gegenereerd die per gebruiker alle SQL query's registreert die zo'n tabel of view benadert. Het is natuurlijk niet mogelijk om daaruit automatisch af te leiden welke personen precies door een beheerder zijn opgevraagd, maar die informatie is wel te achterhalen. Mocht er in de toekomst iets fout gaan dan ligt alles klaar voor de forensisch onderzoeker. Niet alleen de stervelingen maar ook de Goden blijven in het zicht van het alziend oog van de PA. En als binnenkort het aantal mensen met toegang tot productiegegevens verder is teruggebracht resteert een PA-systeem waarin op de oppergod, de DBA, na iedereen op de vingers kan worden gekeken.

De autorisatie/loggingspiraal

Een van de leveringen die we in de inleiding van dit artikel beschreven was een ad hoc levering op verzoek van de Belastingdienst van de gegevens van 12.000 sexwerkers. Die levering wordt gelogd. Omdat het een ad hoc levering is, leggen we een eenmalige service vast en loggen we maximaal 12.000 sofinummers. De mensen achter die nummers zijn vanaf dat moment dus herkenbaar als sexwerkers, want ook de logging behoort tot het PA-systeem. En stel dat de Belastingdienst deze mensen in beeld wil houden, dan is de kans groot er een afnemerindicatie bij ze wordt geplaatst en ben je als sexwerker nog eenvoudiger herkenbaar in de PA. Er is hier sprake van een soort spiraal: hoe beter er wordt gelogd, des te belangrijker is het om het gebruik van autorisatie- en logginggegevens zelf te autoriseren en te loggen. Er is wat dat betreft geen middenweg. Aan de positieve kant van de balans staat natuurlijk dat de PA verslag kan doen van alles dat met de gegevens is gebeurd, want ook het gebruik van de logging wordt gelogd. En als bonus wordt de PA over een aantal jaar tot veel meer dan een registratie van inkomsten(verhoudingen). De logging van de PA is te beschouwen als een soort feitentabel in een dynamisch en uitdijend sterschema. *Gefundenes Fressen* voor alle Business Intelligence specialisten die dit blad lezen.

We besluiten dit artikel met de observatie dat we het hier vrijwel alleen over techniek en technische concepten hebben gehad en niet over het enorme belang dat we als burgers allemaal hebben bij een sluitende registratie en verantwoording van het gebruik van onze privé-gegevens. Er is alle reden om ons zorgen te maken over de manier waarop overheden en bedrijven omgaan met persoonsgegevens en elke stap richting verdere integratie van gegevenshuishoudingen maakt die risico's groter. Voor de overheidshuishouding is de PA wat dat betreft een voorlopig hoogtepunt. Mogelijk komt eens een dag dat privacybescherming weer echt belangrijk wordt en dan betalen we de prijs voor het behandelen van dit onderwerp als sluitpost van ICT-projecten. Dat de PA wel klaar is om uw en mijn privacy te beschermen is geen gevolg van een speciaal daarop gericht ontwerp en uitgebreide inspanningen. Daarvoor ontbrak de tijd en zeker ook het gevoel van urgentie. Dat de PA alles ziet wat er met de gegevens gebeurt is gewoon één van de vele cadeautjes die je krijgt als je met een robuuste en consequent doorgevoerde applicatie architectuur werkt.

In het volgende artikel werpen we een blik op de gebruikersinterface van de Polisadministratie die binnen UWV wordt gebruikt. Die gebruikersinterface is het resultaat van een aantal jaren van succesvol R&D werk gericht op het elimineren van het onderscheid tussen administratieve frontend producten en OLAP en BI tools. De Polisadministratie kent één een dezelfde gebruikersinterface voor datadummy's, data experts en alles daar tussenin.

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