Flexibiliteit, stabiliteit, wendbaarheid en robuustheid: een fascinerend recept

Metadata-driven BI

Ivo van der Heijden en Bas Pruijn

De wereld verandert steeds sneller. Publieke en commerciële organisaties moeten dus steeds sneller schakelen om daarop in te kunnen spelen. Dat legt grote druk op de flexibiliteit van de ondersteunende IT-systemen, die immers meestal juist stabiliteit als prioriteit hebben. Dat geldt ook voor BI-omgevingen.

Hoe kunnen stabiliteit, beheerbaarheid en robuustheid van een BI-oplossing hand in hand gaan met een zeer grote mate van flexibiliteit naar de business toe? Aan de hand van een praktijkcase bij RTL Nederland wordt in dit artikel uitgelegd dat dat zeker mogelijk is, namelijk door prominent gebruik te maken van metadata.

RTL Nederland vraagt natuurlijk nauwelijks meer introductie. Van oprichter van het eerste commerciële televisiestation in Nederland is RTL inmiddels uitgegroeid tot een organisatie met vijf televisiestations, drie radiostations en ruim 150 websites met meer dan 650 medewerkers. Voor een belangrijk deel genereert RTL Nederland inkomsten door het verkopen van reclamemogelijkheden rondom de programma's die zij uitzendt. Daar komt bij dat de programmaformules steeds meer cross-mediaal worden, wat betekent dat internet en mobiele diensten er steeds vaker vaste onderdelen van zijn.

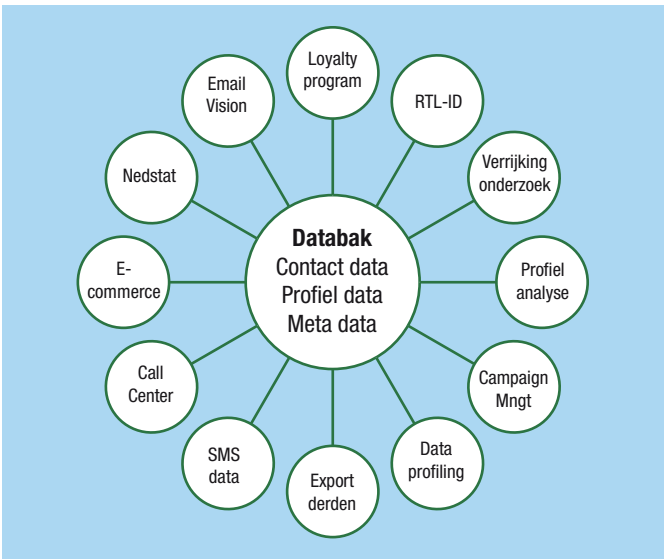
Om adverteerders in contact te brengen met relevante doelgroepen is het essentieel om zoveel mogelijk informatie te verzamelen omtrent de kijkers en luisteraars. RTL Nederland beschikt over meerdere contactmomenten met programmavolgers, zoals de aanmelding voor deelname aan een programma, het ontvangen van nieuwsbrieven en de deelname aan prijsvragen. Tijdens deze contactmomenten worden zoveel mogelijk gegevens verzameld. RTL Nederland maakt met die gegevens profielen van voornamelijk consumenten, waarmee zij adverteerders de mogelijkheid biedt om specifieke programma's te selecteren en op deze manier zo gericht mogelijk te adverteren bij een relevante doelgroep.

Echter, de dynamiek van medialand laat zich niet makkelijk voorspellen. Programma's kunnen met hetzelfde gemak een enorme flop of juist een enorm succes zijn. Positieve voorbeelden van RTL Nederland hierbij zijn programma's als Oh Oh Cherso en The Voice of Holland, twee programmaformules die ver boven verwachting succesvol bleken te zijn. Zodra deze situatie ontstaat gaan de verantwoordelijke redacties van de programma's direct schakelen en nieuwe commerciële initiatieven ontwikkelen.

Het is die dynamiek die het onmogelijk maakt om ruim van te voren, bijvoorbeeld in een projectenkalender, bij IT aan te melden welke werkzaamheden verricht moeten worden.

Probleemstelling

Om RTL Nederland de noodzakelijke flexibiliteit te bieden is het wenselijk de verzamelde profielinformatie zo toegankelijk mogelijk op te slaan en beschikbaar te stellen. Bij de start van deze praktijkcase werden alle profieldata opgeslagen in het Content Management Systeem (CMS) waarop het merendeel van alle websites draait. Dit resulteerde in een dusdanig zware belasting van het CMS dat de data nauwelijks toegankelijk waren.



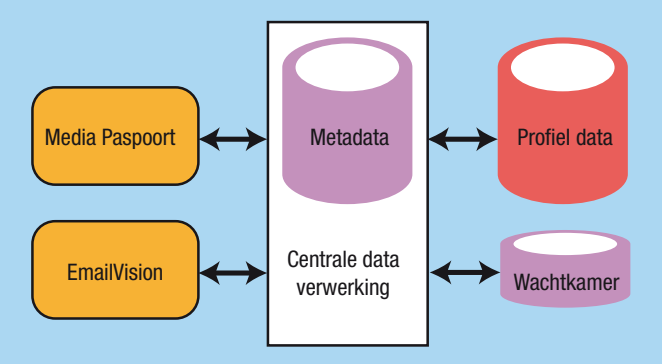
Afbeelding 1: Structuur Databak.

Opgeslagen profielinformatie krijgt pas waarde wanneer deze direct beschikbaar is voor gebruik. Met de gestelde dynamiek van medialand is wachten op wekelijkse of maandelijkse exports niet meer acceptabel.
De IT-organisatie van RTL Nederland heeft, net als elke andere, één primaire verantwoordelijkheid: zorg dat de gebruikte IT-systemen functioneren en beschikbaar zijn op het afgesproken niveau. Hierdoor lijkt het belang van de IT-organisatie, namelijk stabiliteit, strijdig met de belangen van de business, die continu werkt aan verandering. Het bestaansrecht van IT als service-organisatie is direct afhankelijk van het bestaansrecht van de business; IT kan zich dus niet veroorloven om niet tegemoet te komen aan de eisen en wensen die de business stelt. Hierdoor bevindt de IT-organisatie van RTL Nederland zich met regelmaat in de klassieke spagaat van stabiliteit aan de ene kant en dynamiek aan de andere kant.

Duidelijke wensen

Vanuit de hiervoor beschreven uitdagingen voor zowel business als IT volgt een schijnbaar onmogelijke lijst met wensen voor de te implementeren oplossing:
- Centrale, eenduidige opslag van kwalitatieve profieldata;
- Directe koppeling met alle relevante applicaties binnen en buiten RTL (o.a. EmailVision, e-mail marketingpartner van RTL);
- Real-time verwerking van inschrijvingen, uitschrijvingen, anonimiseringen;
- Volledig juridisch verantwoorde verwerking van de persoonsgegevens, rekening houdend met de Wet Bescherming Persoonsgegevens;
- Maximale onafhankelijkheid van de IT-afdeling bij het dagelijkse gebruik van het systeem;
- Hoge beschikbaarheid van het systeem tijdens en vooral ook buiten kantoortijden.

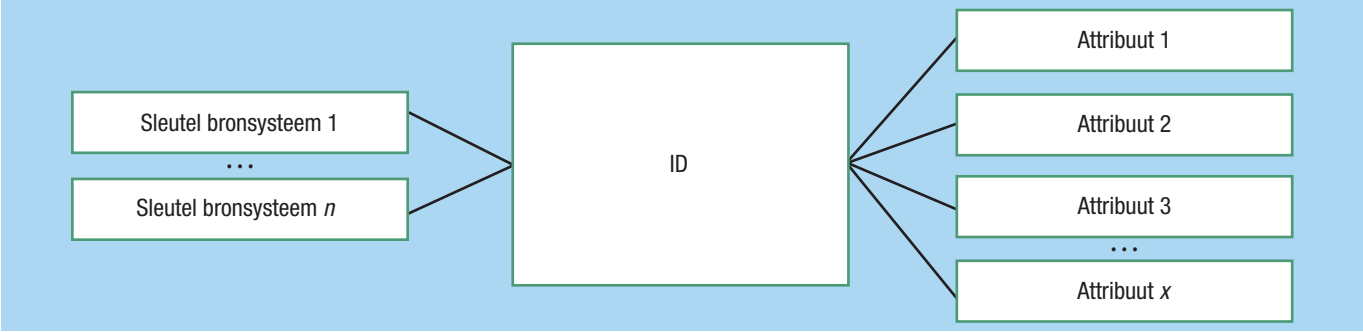
Op basis van deze punten heeft EclipseIT het concept van een Single Point Of Truth (SPOT) database geïntroduceerd als kloppend datahart in het midden van een architectuur van business functionaliteiten. Hierbij vormt de SPOT de basis voor zowel huidige als toekomstige business functionaliteiten. Omdat de taken die deze SPOT kreeg toebedeeld tamelijk eenvoudig van aard zijn, is een eenvoudige naam aan het systeem gegeven: *Databak*. Weinig glossy, maar uiterst functioneel, zie afbeelding 1.
De voorgenoemde business requirements vragen om een ultiem flexibel, robuust systeem waarin alle denkbare gegevens (tekst, datums, plaatjes, video, muziek) opgeslagen kunnen worden.
Hoe bereik je deze ultiem flexibele oplossing met een stabiel en robuust IT-systeem? Maak het metadata-gestuurd!
De kern van een goede oplossing is het verwerken van data op basis van logica, die is vastgelegd in metadata. De logica kan op die manier zeer flexibel afgestemd blijven op de te verwerken data en de wensen vanuit de business. Dit klinkt logisch en eenvoudig, maar het ontwerpen en ontwikkelen van een metadata-gestuurd systeem is al vaker een zeer complex traject gebleken. Vanuit de theorie werken metadata voor alles. In de praktijk is de ontwikkeling al snel een lang proces en worden oplossingen complex, met hoge investeringen tot gevolg. Dit is ook een duidelijke reden waarom met regelmaat 'hard' gecodeerde, minder flexibele systemen ontwikkeld worden. Dergelijke systemen zijn eenvoudiger te doorgronden, ontwerpen en ontwikkelen, waardoor de investering geringer is. Dan wordt het systeem in gebruik genomen en volgt de beheerfase. Zeker bij organisaties waarbij de functionele wijzigingen elkaar snel opvolgen, is een 'hard' gecodeerd systeem niet meer zo eenvoudig en snel. Iedere aanpassing vereist IT-inspanning en een evenredige testinspanning. Dit is nu precies het moment dat een metadata-gestuurd systeem zich gaat terugverdienen. Wanneer een dergelijk systeem eenmaal ontwikkeld en getest is, leidt een aanpassing van de metadata wel tot een aangepaste werking van het systeem, maar niet tot IT- en testinspanning. Dit voordeel wordt versterkt op het moment dat de business gebruikers zelfstandig in staat zijn de metadata van het systeem te veranderen.



Afbeelding 2: Metadata-gestuurd datamodel.

Real-time is de oplossing

Op basis hiervan werd besloten de metadata-gestuurde oplossing te ontwikkelen. Vanuit de gebruikersorganisatie is daaraan toegevoegd dat alle dataverwerking real-time moet geschieden. De Databak, zoals gezegd de naam van het systeem, gaat functioneren als SPOT voor alle business functionaliteiten, die RTL Nederland inzet voor het verzamelen, verwerken en distribueren van profieldata. Een ambitieus concept, waar nog wel wat haken en ogen aan zitten. Aangezien het onbekend is welke gegevens aangeleverd gaan worden, is een structuur nodig die met die onzekerheid uit de voeten kan. Daarbij is een datamodel nodig dat de gegevens metadata-gestuurd kan opslaan. Vervolgens zullen processen moeten worden ontwikkeld, die op basis van metadata en gebruik makend van het metadata-gestuurde datamodel, de juiste dingen doen. Het totaalplaatje van de oplossing ziet er dan uit als in afbeelding 2.
Media paspoort en EmailVision zijn de bronsystemen, zoals die in de eerste fase van het project aangesloten zijn. Dit zijn de systemen die alle voor de SPOT relevante gegevens moeten kunnen aanleveren en ontvangen. Rechts is de opslag van de gegevens van de SPOT weergegeven. Alle gegevens die aan de SPOT worden aangeleverd worden opgeslagen. Hiervoor is het metadata-



Afbeelding 3: Generiek datamodel.

gestuurde datamodel noodzakelijk. De functie van de wachtkamer wordt later in het artikel toegelicht. Centraal in het plaatje staan de metadata gestuurde processen weergegeven.

De aanlevering

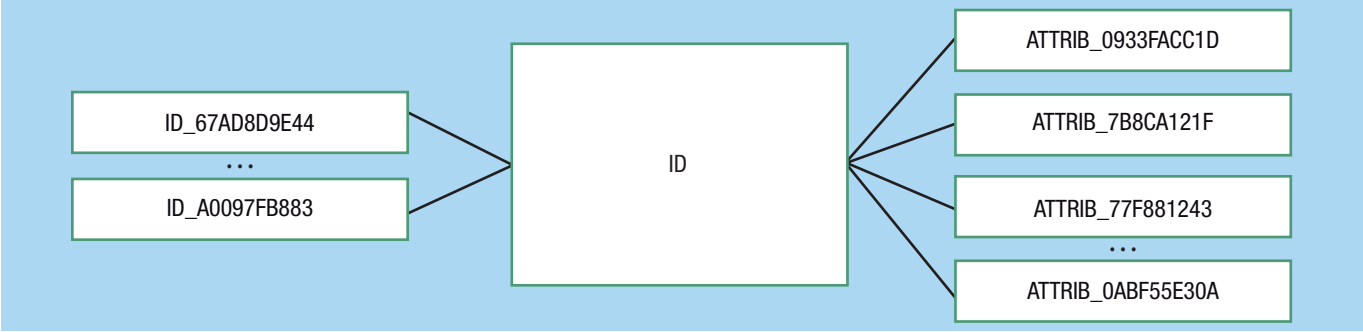
De aanlevering van gegevens vindt plaats met XML-berichten. Aangezien componenten – zoals bijvoorbeeld het formulier van een prijsvraag – allerlei vragen met mogelijke antwoorden kunnen bevatten, bevat elk XML-bericht naast generieke velden (persoon, prijsvraag en aanleverend bronsysteem) ook prijsvraagspecifieke velden (favoriete artiest, videoclip enzovoort). XML als aanleverstandaard biedt de flexibiliteit die nodig is om in te spelen op nieuwe prijsvragen met nieuwe prijsvraagspecifieke velden, zonder dat de gebruikte interfaces technisch gewijzigd hoeven te worden bij veranderende databerichten.
In de metadata van de SPOT staat geregistreerd welk bronsysteem welke attributen (XML-elementen) mag aanleveren. Ieder XML-bericht wordt tegen deze metadata gecontroleerd. Mochten er attributen worden aangeleverd die volgens de metadata niet van deze bron verwacht worden, dan worden deze tijdelijk geparkeerd in een 'wachtkamer'. Gezien de dynamiek van de organisatie kan het zomaar zijn dat er al een prijsvraag online staat op één van de websites, terwijl dat nog niet is verwerkt in de metadata van de SPOT. Door de aangeleverde gegevens vervolgens tijdelijk in de 'wachtkamer' op te slaan, kunnen deze wanneer de metadata zijn bijgewerkt alsnog verwerkt worden in de SPOT. Op dit moment kan de business de ontvangen data ook nog afkeuren, de metadata niet aanpassen en zo voorkomen dat ongewenste data in de databak terecht komen. De SPOT kan dus alle gegevens aangeleverd krijgen. De metadata bepalen welke gegevens verwacht worden, en dus verwerkt mogen worden, en welke gegevens (nog) niet verwerkt mogen worden.

Het metadata-gestuurde datamodel

Het feit dat alle gegevens via XML aangeleverd worden zorgt direct voor een complexiteit bij het opslaan van deze gegevens. De gemakkelijke weg – sla alles direct op als XML in de database – biedt geen oplossing om de gegevens weer eenvoudig ter beschikking van de afnemende applicaties te stellen. Het opslaan van alle informatie in XML-formaat in de database creëert namelijk een probleem om de data op snelle wijze weer op te vragen en te distribueren. Voor de SPOT hebben we gekozen een andere weg in te slaan. Op basis van de aanlever-metadata wordt dynamisch een generiek datamodel opgebouwd, zie afbeelding 3. De centrale tabel in het datamodel is de tabel ID. In deze tabel wordt voor iedere persoon een unieke identificerende sleutel geregistreerd, de SPOT_ID. Deze tabel is ook direct de enige vooraf gedefinieerde tabel van het hele datamodel.
Aangezien attributen van personen vanuit verschillende bronsystemen kunnen worden aangeleverd, wordt op basis van de metadata voor elk bronsysteem een sleuteltabel aangemaakt. In deze tabel staat geregistreerd wat de bronsleutel van een persoon is, in combinatie met de toegewezen SPOT_ID.
De metadata vertellen welke attributen er door de verschillende bronsystemen aangeleverd kunnen worden. Voor elk mogelijk aan te leveren attribuut wordt automatisch een eigen attribuuttabel aangemaakt. Wanneer een nieuw attribuut aan de metadata wordt toegevoegd, wordt automatisch de bijbehorende attribuuttabel aangemaakt. Elke attribuuttabel bevat, naast de SPOT_ID en de attribuutwaarde, ook de geldigheid van de attribuutwaarde. Bij een update van de attribuutwaarde wordt de historische waarde afgesloten, net zoals dat gebruikelijk is bij slowly changing dimensions. Omdat de betekenis van een bepaalde attribuuttabel slechts in de metadata is vastgelegd, en deze metadata gewijzigd kunnen worden, hebben alle attribuuttabellen een gegenereerde naam in de vorm: Attrib_<guid>. Hierdoor is gegarandeerd dat de tabelnamen altijd uniek zijn. Hetzelfde wordt gedaan voor de tabelnamen van de bronsleutels. Het fysieke datamodel ziet er uit zoals in afbeelding 4. Zonder bijbehorende metadata is niet te zien wat er in welke tabel staat opgeslagen.

Het metadata-gestuurde proces

Nu is een metadata gestuurd proces nodig om de gegevens vanuit de XML-berichten daadwerkelijk op te slaan in het dynamische datamodel en door te leveren aan alle systemen die de betreffende gegevens nodig hebben. Dit proces laat zich het meest eenvoudig uitleggen aan de hand van een voorbeeld. We krijgen vanuit één van de bronsystemen een bericht met een nieuw adres van een persoon.
Stap 1. Ontvang het bericht. Kijk in de metadata of het aanleverende bronsysteem gegevens mag aanleveren aan de SPOT. Zo nee: geef een foutmelding retour. Zo ja: ga verder met stap 2.



Afbeelding 4: Fysiek datamodel.

Stap 2. Bepaal op basis van het aanleverende bronsysteem in welke fysieke tabel de bronID's worden opgeslagen door dit in de metadata op te zoeken.
Stap 3. Bepaal het bronID uit het XML-bericht. Dit bronID geeft aan over welke persoon het bericht gaat. Kijk of de betreffende persoon al voorkomt in de bronID tabel. Mocht dit nog niet het geval zijn, voeg deze persoon dan toe aan de bronID tabel.
Stap 4. Bepaal van deze persoon de SPOT_ID. Dit is de basis voor alle verdere verwerking binnen de SPOT.
Stap 5. Kijk in de XML welke attributen er worden aangeleverd. In dit geval worden voor het adres de attributen "straat", "huisnummer" en "woonplaats" aangeleverd.
Stap 6. Pak het eerste attribuut uit de XML, in dit geval "straat". Kijk in de metadata in welke attribuuttabel van de SPOT het attribuut "straat" van deze bron opgeslagen moet worden. Mocht het attribuut "straat" van deze bron nog niet bekend zijn in de metadata, sla dan de waarde van "straat" op in de wachtkamer, zodat deze later alsnog verwerkt kan worden.
Stap 7. Mogelijk hebben we al een attribuut "straat" van deze persoon opgeslagen. Als dat het geval mocht zijn, en de nieuw aangeleverde attribuutwaarde is anders dan de reeds geregistreerde attribuutwaarde, dan wordt de bestaande attribuutwaarde afgesloten met een einddatum.
Stap 8. Als de attribuutwaarde veranderd is, of als er voor deze persoon nog geen "straat" bekend was in de SPOT, wordt de nieuw aangeleverde attribuutwaarde opgeslagen.
Stap 9. Nu wordt in de metadata gekeken welke afnemende systemen geïnteresseerd zijn in het attribuut "straat". Voor elk afnemend systeem wordt geregistreerd dat er een (nieuwe) waarde voor "straat" bekend is voor deze persoon.
Stap 10. Het attribuut "straat" is nu verwerkt. Vervolgens worden de stappen 6, 7, 8 & 9 herhaald voor de attributen "huisnummer" en "woonplaats".
Stap 11. Nadat alle attributen zijn opgeslagen wordt er aan het aanleverende systeem gemeld dat het XML-bericht ontvangen en verwerkt is.
Stap 12. De in stap 9 geregistreerde gegevens worden aan de verschillende afnemende systemen doorgegeven. Afhankelijk van de wens van het afnemende systeem kan dit een real-time of een batchgeoriënteerde melding zijn. Dit wordt op basis van de metadata vastgesteld.

In de praktijk bevatten de meeste berichten veel meer dan drie attributen. Als we kijken hoe vaak er informatie in de metadata opgezocht moet worden, voordat de eigenlijke data opgeslagen kunnen worden, is direct duidelijk dat het initieel opzetten van dit proces behoorlijk complex is.

Conclusie

Het ontwikkelen van een real-time metadata-gestuurd systeem levert veel hoofdbrekens op. Het opzetten van het systeem is een stuk complexer dan het ontwikkelen van een conventioneel 'hard' gecodeerd systeem. Het maken van een ETL-stroom waar zowel data als metadata bij elkaar komen, vraagt enorm veel inzicht en overzicht van de ontwikkelaars en beheerders.
Bij RTL Nederland is gekozen voor Informatica Powercenter Real Time Edition voor het realiseren van de data-integratie voor de SPOT. Hierdoor is het beschikbaar stellen van een webservice voor het real-time vervangen van gegevens zeer eenvoudig. Ook het aanroepen van webservices is voor kenners van Powercenter niet erg ingewikkeld.

Zoals bij elke software-oplossing die werkt met een generiek datamodel, is ook bij deze oplossing gebleken dat het bereiken van een goede performance een uitdaging is. Pas op het moment dat de data verwerkt worden door de Powercenter-engine wordt duidelijk voor welke data de metadata bevraagd moeten worden. Om al deze gegevens op tijd in het proces beschikbaar te hebben, is de inrichting van de onderliggende database van essentieel belang. De juiste inrichting van de onderliggende Microsoft SQL Server 2008 database heeft veel aandacht gekost.
De SPOT die EclipseIT voor RTL Nederland heeft ontwikkeld komt tegemoet aan de wensen van de business om snel, zonder de IT-organisatie te hoeven inschakelen, in te kunnen spelen op nieuwe ontwikkelingen. De SPOT is tegelijk een zeer stabiel systeem, zodat de IT-organisatie zich kan richten op haar taken: het op niveau houden van de dienstverlening, zonder continu geconfronteerd te worden met wijzigingen in het systeem. Dankzij de metadata-gestuurde opzet zijn flexibiliteit en robuustheid op een goede manier samengebracht.

Ivo van der Heijden en **Bas Pruijn** zijn beiden werkzaam als BI-consultant bij EclipseIT.