BI gebruikt voor onderzoeksprogramma naar veiligheidsgevoel

Flexibiliteit Data Vault geschikt voor pilot project

Marco Knegt en Marcel de Wit

Is er een duidelijke relatie tussen de feitelijke veiligheid op straat en het veiligheidsgevoel van de burger? Wat voor invloed hebben berichten via kranten, radio en tv op dat gevoel? Een praktijkvoorbeeld van hoe Business Intelligence ook bij wetenschappelijke vraagstukken een rol kan spelen.

Business Intelligence-specialisten van Sogeti Nederland en VLC zijn betrokken bij het project Changing Perceptions of Security and Interventions (CPSI). CPSI is onderdeel van een Europees meerjarig onderzoeksprogramma op het gebied van veiligheid. Dit internationale project wordt onder aanvoering van TNO voor de Europese Commissie uitgevoerd. TNO werkt binnen dit project samen met acht partijen, waaronder onderzoeksinstituten, universiteiten en ICT-dienstverleners.
Doel van het onderzoek is een conceptueel model te ontwikkelen dat inzicht geeft in de factoren van het veiligheidsgevoel in relatie tot de feitelijke veiligheid in Europa. Voor het model is het belangrijk om te bepalen wat de rol van de media, publieke opinie en cultuur is. In het verlengde daarvan kan worden nagegaan welke interventies geschikt zijn om de feitelijke veiligheid en het veiligheidsgevoel van burgers te vergroten. Daarmee is dit model voor landelijke en lokale beleidsvormers en beslissers een hulpmiddel bij het effectief verhogen van de veiligheid. Ter ondersteuning van dit model is een BI-oplossing gerealiseerd. In dit artikel staat de vertaling van de onderzoeksvraag naar het informatiemodel en de uiteindelijke realisatie van een DWH centraal.

Wetenschappelijk karakter

In de praktijk worden beleidsvormers maar beperkt ondersteund met hulpmiddelen voor het verhogen van de veiligheid. De meeste onderzoeken benaderen het onderwerp eenzijdig vanuit één bepaalde invalshoek in plaats van een integrale benadering waarbij er rekening wordt gehouden met meerdere factoren. Er is dan ook behoefte aan een beter inzicht in de factoren die van invloed zijn op de feitelijke veiligheid in relatie tot de gevoelsmatige veiligheid. In dit project is gekozen voor een integrale benadering met daarnaast ook de aandachtsgebieden media, publieke opinie, cultuur en demografie.

Het project heeft een wetenschappelijk karakter. Dat blijkt ook uit de samenstelling van de projectgroep. Deze is multidisciplinair en bestaat uit onderzoekers, sociologen, psychologen en ICT'ers. Het project kent de volgende stappen:

1. De ontwikkeling van een model waarin de relaties tussen de factoren is vastgelegd;
2. De ontwikkeling van een methode om de benodigde gegevens te verzamelen, te kwantificeren, te organiseren, te analyseren en te interpreteren;
3. De ontwikkeling van een datawarehouse waarin die gegevens kunnen worden opgeslagen;
4. De validatie van het in stap 1 en 2 ontwikkelde model en methode. Hier wordt getoetst in hoeverre de aannames in het model juist waren.

We gaan alleen dieper in op stap 3; de ontwikkeling van het datawarehouse. Als BI-consultant speel je een belangrijke rol door je te verdiepen in de materie van de business en daarnaast mee te denken, kritisch te zijn en de juiste vragen te stellen. Natuurlijk doe je dat niet alleen. Om tot een succesvolle BI-oplossing te komen is veel input van de business nodig. Dit traject was bijzonder te noemen. We liepen tegen andere vragen aan dan we gewend waren, zoals: welke bronnen zijn er nodig, wat zijn de gemeenschappelijke gegevens, hoe kan je omgaan met de verschillende landen, hoe zit het met de privacy en tot slot hoe kom je tot een model dat de gevraagde analysemogelijkheden biedt?

Aanpak

In het begin van het project was volstrekt onduidelijk welke gegevens er nodig waren, laat staan hoe de validatie moest worden uitgevoerd. De hiervoor genoemde factoren leken allemaal los van elkaar te staan en daarnaast was het onduidelijk welke

bronnen daarvoor beschikbaar waren. In een aantal gevallen moesten de brongegevens als onderdeel van het project zelfs worden gegenereerd. Een voorbeeld hiervan is het aandachtsgebied media, waarbij het internet over een periode met gerichte zoekcriteria is afgezocht. Een ander voorbeeld is de enquête naar het veiligheidsgevoel, waarvan het ontwerp en de uitvoering ook binnen het project vielen.

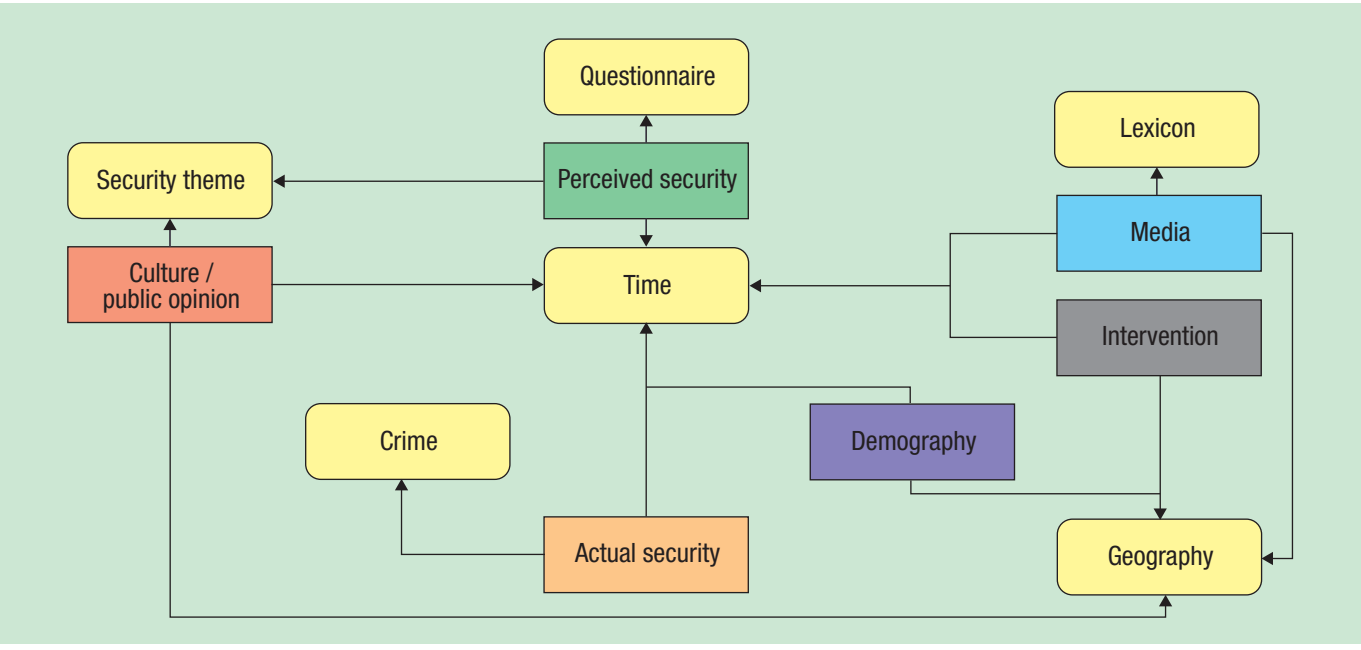
Voor het ontwikkelen van een DWH is het essentieel om feiten en dimensies te kunnen onderkennen. Normaal gesproken heb je de beschikking over onderling samenhangende bronsystemen, waaruit je deze kunt destilleren. In dit project was het echter zoeken naar de verbindende elementen. Sterker nog, deze moesten worden gedefinieerd. Het doel was om de analysemogelijkheden te vergroten door zoveel mogelijk gemeenschappelijke dimensies te gebruiken om zo de verschillende feiten en dus factoren met elkaar in verband te kunnen brengen. Uiteindelijk heeft dit geresulteerd in het datamodel in afbeelding 1. Daarbij kunnen bijvoorbeeld wel de volgende vragen worden gesteld:

- Perceived security: in welke buurten, en hoe vaak, is een bepaald antwoord op een enquêtevraag gegeven (gevoelsmatige veiligheid);
- Media: hoe vaak is in een periode een bepaald trefwoord gevonden in de media;
- Demography: wat valt er te zeggen over de demografische kenmerken van een bepaalde buurt;
- Actual security: hoe vaak, en in welke buurten, worden in een periode bepaalde misdrijven gepleegd (feitelijke veiligheid);
- Culture: welke cultuureigenschappen zijn van invloed op bepaalde veiligheidsthema's;
- Intervention: wanneer, en in welke buurt, zijn bepaalde veiligheidsverhogende maatregelen genomen?

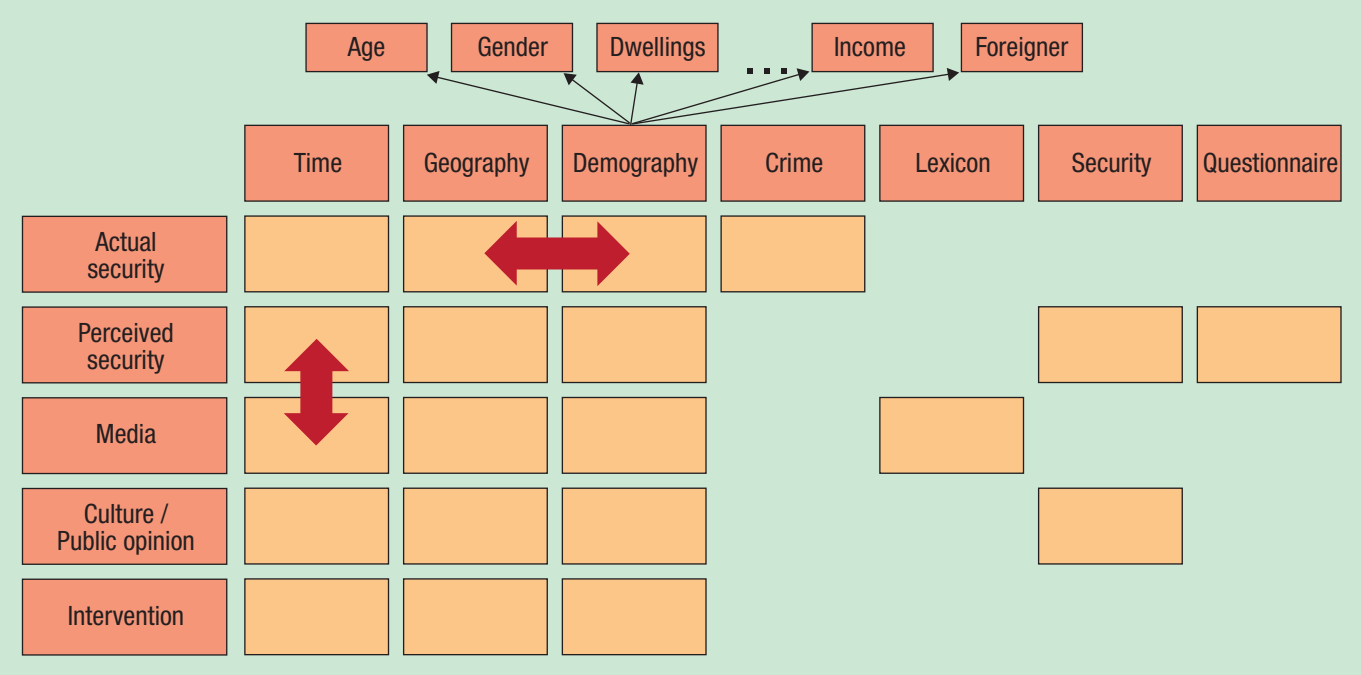
Vanwege het Europese karakter en de wens om op termijn landen te kunnen vergelijken is gekozen om zoveel mogelijk te standaardiseren. Zo is onder andere gebruik gemaakt van ISO-landencodes en postcodes voor de regionale aanduiding. Het uiteindelijke model (inclusief het onderliggende DWH) moest aan de ene kant generiek zijn om het te kunnen implementeren in verschillende Europese landen. Anderzijds moest het voldoende flexibel zijn om het te kunnen laten aansluiten op lokale eisen. Hiervoor is de ontwerpkeuze gemaakt om in het DWH het proces van gegevensaanlevering en de centrale opslag van elkaar te scheiden.
Daarnaast hadden we te maken met de dwingende eis dat vanuit het oogpunt van privacy de gegevens niet tot een individu mochten zijn te herleiden. Voor het toetsen van het model was dit echter wel noodzakelijk maar voor eindgebruikerrapportages niet. Daarom is besloten het DWH met de individuele feitelijke resultaten niet voor eindgebruik beschikbaar te stellen maar voor de eindgebruikers datamarts in te richten met daarin slechts geaggregeerde gegevens.
Een wezenlijk aspect van dit project betrof de bruikbaarheid van het informatiemodel, zie afbeelding 2. Dat komt tot uiting in het ons bekende dimensionale model. Een krachtig element daarin vormen de demografische gegevens, omdat er per regio vele demografische kenmerken beschikbaar zijn. Deze kenmerken zijn dan aan elke andere veiligheidsfactor te koppelen. Dit verruimt de analysemogelijkheden aanzienlijk.

Data Vault

Voor de realisatie van het DWH is gekozen voor de 'nieuwkomer' in datawarehousingland. Dit betreft de Data Vault methodiek van Dan Linstedt. Waarom is er bij het CPSI-project gekozen voor deze nieuwe modelleringstechniek en niet voor de meer gangba-

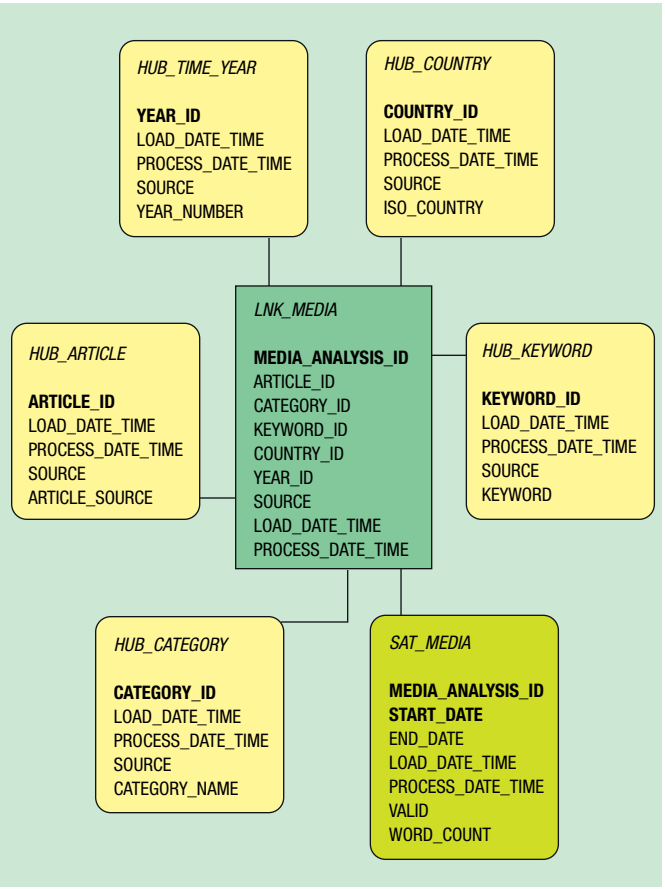


Afbeelding 1: Datamodel.



Afbeelding 2: Informatiemodel.

re technieken van Inmon en Kimball? Beide genoemde gangbare technieken zijn minder toepasbaar in omgevingen die onderhevig zijn aan veranderingen van bedrijfsprocessen en operationele systemen. Data Vault kent een aantal voordelen. Data Vault neemt de data als uitgangspunt, aangezien dit minder onderhevig zal zijn aan wijzigingen dan processen en systemen. Data Vault staat voor grotere flexibiliteit, uitbreidbaarheid en schaalbaarheid van het informatiemodel door de manier waarop relaties worden gemodelleerd. Relationele data worden gescheiden van de identificerende en de beschrijvende attributen. Deze flexibiliteit is nodig aangezien CPSI een groeimodel betreft. Er is op dit moment een aantal gegevenselementen gedefinieerd waaruit de feitelijke en gevoelsmatige veiligheid afgeleid kan worden. In de toekomst kan dit inzicht uiteraard wijzigen waardoor gegevenselementen niet meer relevant zijn en er nieuwe benodigd zijn. Data Vault is uitermate geschikt voor het integreren van nieuwe gegevensverzamelingen uit nieuwe of gewijzigde bronsystemen.

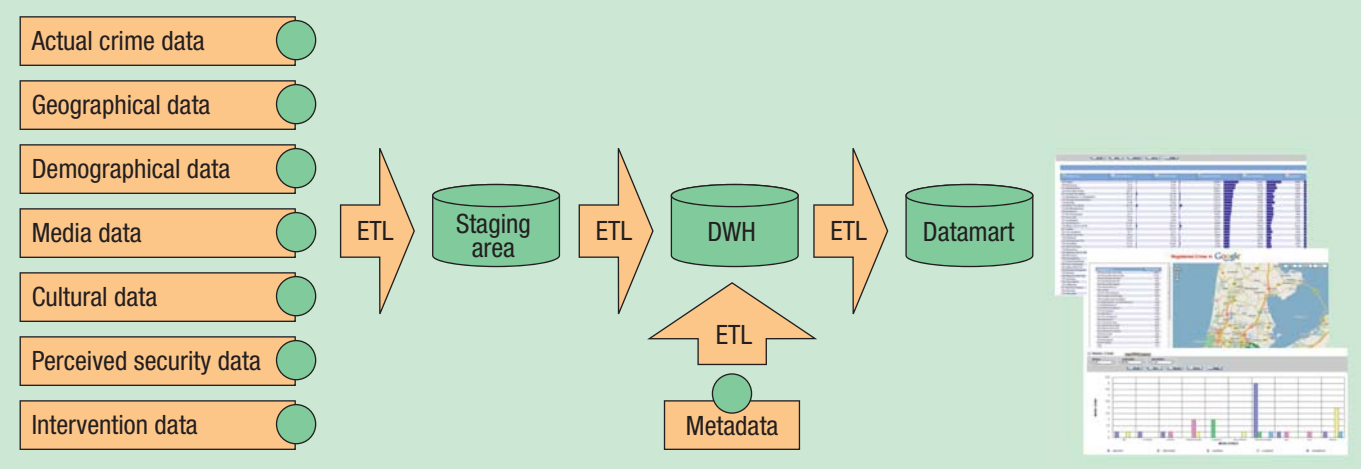


Afbeelding 3: Primaire sleutels.

CPSI betreft op dit moment een pilot project met een beperkt onderzoeksgebied. Bij positieve resultaten kan de CPSI-systematiek op grotere schaal worden uitgerold, bijvoorbeeld op gemeentelijk, landelijk of op Europees niveau. Data Vault is uitermate geschikt voor het schaalbaar en uitbreidbaar maken van gegevensverzamelingen. Kenmerkend voor deze datamodelleringsmethodiek is verder dat alle historische wijzigingen op efficiënte wijze in het model worden opgeslagen. Hierdoor is het mogelijk de exacte historische situatie uit de database te rapporteren. Een laatste voordeel is tenslotte dat alle data altijd geladen worden. Hierdoor wordt een compleet overzicht verkregen en gaat er geen informatie verloren.

Uitwerking

Om de structuur van het Data Vault model toe te lichten wordt de component 'Media' uit het CPSI-ontwerp hier verder uitge-



Afbeelding 4: Drie specifiek ingerichte onafhankelijke omgevingen.

werkt. Waar gaat het om bij de component Media? Media is als een van de bepalende factoren onderkend die bijdragen aan een subjectief veiligheidsgevoel. Hoe dit nu te kwantificeren en op te nemen in het model? Hiertoe is op basis van een vooraf opgestelde lijst met categorieën en trefwoorden gedurende een aantal maanden in bepaalde online media gezocht naar het aantal hits op deze trefwoorden. Deze hits zijn opgenomen in het DWH.
Een Data Vault model bestaat uit de componenten Hub, Link en Satellite. In het ontwerp bevat de Hub primaire sleutels van bedrijfsobjecten. In het geval van 'Media' betreffen dit de entiteiten 'Article', 'Category', 'Keyword', 'Country' en 'Time', zie afbeelding 3. Een Link zorgt voor het onderhouden van de relaties tussen de door de Hub's beschreven entiteiten. In het model is de Link 'Media' opgenomen. Dit is het verbindende element tussen de hiervoor genoemde Hub's. Een Satellite bevat de aan de tijdsbeelden onderhevige attributen die bij een primaire sleutel in een Hub of Link horen. In afbeelding 3 is alleen de Satellite van 'Media' uitgewerkt. In deze Satellite wordt hier het aantal gevonden hits van een bepaald keyword geregistreerd.
In de Satellite wordt door middel van de attributen valid, start-date en end-date aangegeven of een record valide is en gedurende welke tijdsperiode het record geldig is geweest. Verder worden er in zowel Hub, Satellite als Link timestamps meegegeven, zodat er verschillende tijdsbeelden gegenereerd kunnen worden. Een proces date timestamp is opgenomen indien gegevens meerdere keren per dag worden geladen. Ook wordt de herkomst van gegevens opgenomen.
Het Data Vault model is echter minder geschikt voor rapportagedoeleinden. Hiervoor kan beter gebruik gemaakt worden van een stermodel. Maar hoe te komen van een Data Vault model naar een stermodel? In een stermodel gaat het om feiten en dimensies, terwijl een Data Vault model is opgebouwd uit Hub-, Satellite- en Link-tabellen. Op hoofdlijnen kan gesteld worden dat de basis van een dimensie altijd wordt gevormd door een combinatie van een Hub-tabel met één of meerdere Satellite-tabellen. De basis van een feitentabel is altijd een Link-tabel eventueel in combinatie met één of meerdere Satellite-tabellen.

Projectarchitectuur

Het CPSI DWH bestaat uit drie specifiek ingerichte onafhankelijke omgevingen. De eerste betreft de Staging Area, de tweede het eigenlijke DWH en tenslotte is er een Datamart-omgeving, zie afbeelding 4. De Staging Area is ingericht met een tabelstructuur die correspondeert met de gewenste structuur volgens het datamodel zoals eerder getoond. Hiertoe is een interfacelaag nodig die aangeleverde bronbestanden converteert naar deze structuur. Het DWH is zoals vermeld ingericht volgens de Data Vault methodiek. De Datamart is gerealiseerd met een dimensionaal model als grondslag. Dit vanwege gebruikerspecifieke vragen en performance eisen.
Voor de realisatie is gebruik gemaakt van MySQL als databaseplatform. De producten Data Migrator van iWay en WebFOCUS van Information Builders zijn ingezet voor respectievelijk het vormgeven van de ETL-stromen en rapportagedoeleinden.

Bereikte resultaten

Het bijzondere aan dit project is dat BI gebruikt is voor een wetenschappelijke toepassing. De inzet van het DWH was enerzijds ter toetsing van het model en anderzijds voor inzage in de samenhang tussen de verschillende factoren. Daarbij is gebruik gemaakt van vaste overzichten, ad hoc bevraging met OLAP-functionaliteit, trendanalyses en geografische rapportages.
Binnen het project is duidelijk waardering voor de 'kracht' van het DWH als instrument om de verschillende soorten gegevens uit diverse bronnen integraal op te slaan en ook weer gemakkelijk daaruit te kunnen ophalen. Het DWH heeft daarmee zeker een toegevoegde waarde geleverd. Waar wij als specialisten al snel roepen van "ach, dat is toch standaard", werd het binnen het project als een prestatie ervaren en daar zijn we erg trots op.

Marco Knegt en Marcel de Wit
Marco Knegt (marco.knegt@vlc.nl) is projectleider bij VLC en specialist op het gebied van Enterprise Information Management.
Marcel de Wit (marcel.de.wit@sogeti.nl) is senior consultant Business Intelligence bij Sogeti Nederland.