Auteurs: Geoffrey Van den Bergh is Privacy en Data Management Consultant bij CRANIUM Nederland. Hij ondersteunt organisaties in het oplossen van complexe problemen op het gebied van gegevensverwerking. Geoffrey is te bereiken op geoffrey.vandenbergh@cranium.eu. Marloes de Bruin is Privacy, Security en Data Management Consultant bij CRANIUM Nederland. Ze begeleidt organisaties en heeft programma's voor privacy, security en data management gecreëerd en geïmplementeerd. Marloes is te bereiken via marloes.debruin@cranium.eu.

Méér met data.
Maar wát eigenlijk?

De mogelijkheden van data zijn eindeloos. Vooral het analyseren en inzetten van gestructureerde en ongestructureerde data voor bepaalde doeleinden kan een schat aan waarde opleveren voor organisaties. Data in bezit van organisaties leidt naar inzichten en inzichten leiden weer naar kennis. Welke manieren zijn er om data te analyseren onder de Algemene Verordening Gegevensbescherming (AVG) en hoe haal je de meeste waarde uit data?

Het internet der dingen gaat in essentie over 'dingen' of objecten die via het internet met elkaar verbonden zijn. In feite wordt door deze koppeling een model van onze realistische wereld in het internet afgebeeld. Een model waarvan we de elementen kunnen 'voelen' en 'beïnvloeden', onafhankelijk waar wij ons bevinden. In het algemeen zie je dat IoT wordt gebruikt om slimmer om te gaan met alles wat al bestaat, door de eigenschappen te meten en deze waarden naar een platform op het internet te sturen. Dit platform kan vervolgens besluiten hier iets slims mee te doen, bijvoorbeeld de omgeving aan te passen. Vandaar dat de term 'smart' vaak valt in combinatie met IoT. Er gaat geen dag voorbij of je hoort over Smart City, Smart Industry, Smart Home, Smart Energy en ga zo maar door. IoT maakt het mogelijk om steeds meer privacygevoelige data te generen en te combineren. Door de toename van dit soort toepassingen, vormen privacy en security een steeds grotere uitdaging binnen dergelijke ontwikkelingen.

De verwerkingsverantwoordelijke bepaalt de doeleinden waarvoor en de middelen waarmee persoonsgegevens worden verwerkt. Als een organisatie dus beslist 'waarom' en 'hoe' persoonsgegevens moeten worden verwerkt, is zij de verwerkingsverantwoordelijke. De verwerker verwerkt persoonsgegevens uitsluitend namens de verwerkingsverantwoordelijke, en is vaak een derde partij buiten de organisatie.

Organisaties worden niet zozeer beperkt in het doen van data-analyses, maar de manier waarop is wel afhankelijk van de verplichtingen uit de wetgeving. Er zijn inmiddels goede oplossingen beschikbaar om analyses te kunnen uitvoeren zonder dat de gegevens herleidbaar zijn naar individuen en de privacy van de betrokkenen wordt gewaarborgd. Zo kun je gegevens aggregeren waardoor je alleen nog totalen zichtbaar maakt, of je kunt een hele database anonimiseren of synthetiseren. Op deze manier kan een organisatie meer inzicht krijgen in gegevens en verbanden maar kunnen zij ook nieuwe diensten aanbieden of hun huidige dienstverlening verbeteren.

Data is in bits z'n goud waard

Vaak is data-analyse binnen organisaties gericht op het opleveren van managementinformatie zoals omzet- of verzuimcijfers. Maar sommige organisaties willen meer inzicht verkrijgen om nieuwe producten te bouwen en de huidige dienstverlening te verbeteren. Neem als voorbeeld een financiële dienstverlener die o.a. salarissen betaalt en ook verzekeringen aanbiedt. De organisatie die al deze salarissen en verlofaanvragen verwerkt en bijhoudt, heeft een schat aan data opgeslagen. Zo wordt bijvoorbeeld de functie bijgehouden, het salaris geregistreerd en leeftijd bijgewerkt. Maar deze organisaties worden in het kader van de AVG vaak gekwalificeerd als verwerker. Het doen van data-analyses op alle salarisgegevens van de medewerkers van de klant is voor eigen doeleinden vaak niet toegestaan, mits dit contractueel wordt geregeld of de analyses aantoonbaar kunnen helpen bij het onderhoud en verbetering van het systeem.

Ook voor een verwerkingsverantwoordelijke is het interessant om te weten hoe behoeftes van klanten aan elkaar gekoppeld kunnen worden. Zo kan bol.com bijvoorbeeld uit de data van gekochte producten een database opbouwen, en als deze groot genoeg is, op basis van de resultaten behoeften voorspellen van klanten die soortgelijke aankopen doen, en deze met een korting aanbieden.

Wanneer er analyses worden uitgevoerd, wordt data al snel geanonimiseerd, omdat op geanonimiseerde data de AVG niet van toepassing is. Maar het anonimiseren van persoonsgegevens zelf wordt gezien als een verwerking van persoonsgegevens in het kader van de AVG. Derhalve moet deze verwerking worden uitgevoerd op een wijze die in overeenstemming is met de AVG en zich houdt aan de beginselen van gegevensbescherming. Er zijn verschillende methodes om gegevens te anonimiseren. De methode die je kiest, is afhankelijk van het type gegevens dat je wilt anonimiseren.

Analyse op traditionele wijze

Bij klassiek anonimiseren wordt de originele dataset gemaskeerd om te waarborgen dat individuen niet kunnen worden herleid. Een veelgebruikte methode is om gegevens zoals voornaam en achternaam in willekeurige volgorde te 'husselen', zodat je nieuwe voornaam/achternaam combinaties krijgt. Een andere methode om data te maskeren is om een kolom, die je niet nodig hebt voor je test, leeg te maken. Op die manier worden privacygevoelige gegevens en alle risico's ervan letterlijk uit het veld geruimd. Het 'scramblen' van data is een derde methode die data onherkenbaar maakt: het vervangt tekens door x en cijfers door 1. De data kan via vooraf gedefinieerde regels dus worden vervangen.

Een nadeel hiervan is dat de datakwaliteit verslechtert doordat de data wordt bewerkt. Dit komt doordat in het procedé van anonimiseren bepaalde velden ofwel weggehaald worden ofwel gemaskeerd, omdat ze op basis hiervan niet meer herleidbaar zijn tot een individu. Het nadeel hiervan is dat al deze data aan elkaar gekoppeld is. Wanneer een computer deze data niet meer kan lezen, komen er omissies in het profiel van de dataset. Zo kan er in de praktijk geen koppeling meer plaatsvinden tussen historische en toekomstige datasets. Daarnaast blijft de data een bewerkte versie van het origineel waarbij er vaak nog relaties blijven bestaan tussen de oorspronkelijke en geanonimiseerde data. In de praktijk is het dus onvoldoende om identificeerbare gegevens te verwijderen uit bepaalde datasets. Het loskoppelen van (in)direct identificerende persoonsgegevens en de overige gegevens moet onomkeerbaar zijn. Je mag deze gegevens ook in een later stadium niet alsnog aan elkaar kunnen koppelen. Bijvoorbeeld door andere (bijkomende of nieuwe) gegevens of technieken te gebruiken, waardoor je personen toch nog zou kunnen identificeren. Daarom is anonieme data vaak niet echt anoniem. Binnen eenzelfde organisatie is het namelijk relatief eenvoudig te achterhalen op wie de geanonimiseerde gegevens betrekking hebben door intern beschikbare verschillende datasets naast elkaar te leggen. Om gegevens daadwerkelijk volledig te anonimiseren moeten deze datasets geaggregeerd worden over een voldoende grote groep personen.

Een nieuwe speler op de markt: fictieve data

Het nadeel van geaggregeerde gegevens is dat deze gegevens vaak niet geschikt zijn om te gebruiken voor verdere ontwikkeling en testen. Om dit wel mogelijk te maken, kunnen gegevens gesynthetiseerd worden. Dit houdt in dat gegevens worden gegenereerd op basis van fictieve gegevens. Er kan kunstmatige intelligentie worden toegepast om de kenmerken, structuur en waarde van originele data te behouden. Het gevolg is volledig nieuwe, kunstmatig gegenereerde data met een dusdanig hoge kwaliteit dat deze data gebruikt kan worden alsof het originele data is, maar dan zonder te hoeven voldoen aan privacywetgeving omdat er geen relatie meer is tussen de originele data en er dus ook geen sprake meer is van persoonsgegevens. Het gebruik van persoonsgegevens wordt daarnaast geminimaliseerd, hierdoor wordt het risico op datalekken ook drastisch gereduceerd. Daarnaast kunnen organisaties of afdelingen binnen de organisatie, toegang krijgen tot datasets die eerder niet mochten worden gebruikt wegens privacywetgeving.

De datakwaliteit is dusdanig hoog dat zelfs het ontwikkelen van complexe algoritmes en machine-learning modellen mogelijk is op basis van synthetische data. AI gegenereerde synthetische gegevens kunnen worden gebruikt alsof het originele gegevens zijn. Synthetische data biedt het best mogelijke alternatief omdat het de kenmerken, relaties en statistische patronen behoudt, zoals ook in de originele data. Hiermee kan men dus echt datagedreven innovaties realiseren.

De oorsprong van het gebruik van synthetische data komt uit testmanagement, omdat er door het synthetiseren geen productiedata wordt gebruikt in de testomgeving. Het nadeel is dus dat synthetische data nog niet wordt gezien als een algemene geaccepteerde manier van werken. Andere nadelen zijn dat een kopie van de productieomgeving overzetten naar een ontwikkel- en testomgeving relatief eenvoudig is, hiertegenover is het een stuk ingewikkelder om een goede synthetische dataset te ontwikkelen. Want goede synthetische datasets bevatten weliswaar fictieve gegevens, toch dienen ze bepaalde overeenkomsten te hebben met de originele data. Een voorbeeld hiervan zijn de verhoudingen tussen verschillende data-objecten, deze dienen ongeveer gelijk te blijven aan de originele data. In het geval van de salarisadministrateur, is dit ook van belang. De organisatie maakt namelijk gebruik van meerdere bronsystemen. Wil je dat de verbanden tussen verschillende ID-nummers uit meerdere systemen blijft bestaan, dan dient het gesynthetiseerde ID-nummer ook hetzelfde zijn.

Fictie leidt tot werkelijke waarde

Afhankelijk van wat voor organisatie je bent en hoe datagedreven jouw organisatie wil en kan zijn, is de keuze voor synthetische data ingewikkeld. Wanneer de financiële dienstverlener namelijk de databases in de verschillende systemen synthetiseert, kunnen zij nog steeds de juiste analyses doen en nieuwe diensten aanbieden. Denk bijvoorbeeld aan een salarisvergelijker. De salarisveranderingen die normaal doorkomen van de medewerkers van klanten, worden ook gebruikt om te analyseren wat een marktconform salaris is. Het kan voordelig zijn voor zowel werknemer als werkgever, want een werkgever kan bepalen wie te veel of juist te weinig betaald krijgt, terwijl een werknemer zichzelf kan vergelijken ten opzichte van anderen met gelijke kenmerken.

Met de huidige trends en ontwikkelingen in techniek is het de vraag of klassiek anonimiseren nog echt anoniem is en de kwaliteit van data goed genoeg om analyses te kunnen doen die inzicht geven. Door het gebruik van nieuwere technieken zoals synthetiseren kunnen organisaties analyses doen die niet alleen AVG-compliant zijn, maar ook inzicht geven en waarde creëren voor de organisatie.