Belang van anonieme data vereist integrale aanpak

Anonimiseren van testdata

Nico Klaassen, Maurice Siteur

Het beschikbaar stellen van data aan verschillende doelgroepen vraagt om speciale aandacht en die krijgt het meestal niet. Het verkrijgen van testdata en het anonimiseren van de testdata is een onderdeel hiervan. Zeker daar waar het gaat om data beschikbaar te stellen buiten de organisatie. Via dit artikel willen we inzicht geven in het belang, de mogelijke problemen voor projecten en een werkwijze die daarbij gehanteerd kan worden.

Gezien vanuit wettelijk kader is privacybescherming een belangrijk punt. Bij het outsourcen van projecten naar bijvoorbeeld India zijn er Europese regels waaraan voldaan moet worden. Binnen veel organisaties is hiervoor te weinig of geen aandacht en worden productiegegevens zonder voldoende voorzorgsmaatregelen naar alle uithoeken van de aarde verstuurd. Een nog groter probleem is dat 'in huis' gegevens herkenbaar in diverse instanties van (test- en ontwikkel-) systemen terecht komen en zichtbaar worden voor niet geautoriseerde (externe) medewerkers. Men hoeft niet lang te wachten voordat er claims komen.
Het anonimiseren van gegevens is belangrijk voor het beschikbaar stellen van de juiste data aan andere 'gebruikers'. Het is vaak niet wenselijk dat vertrouwelijke gegevens bij (acceptatie) testers terecht komen, maar zij willen wel graag met een zo productiegetrouwe versie werken om latere problemen te voorkomen. Dat geldt niet alleen bij testen, maar ook bij migraties en interface ontwikkeling. Het is niet altijd voldoende om via een 'geheimhoudingsverklaring' de privacy van personen of zelfs bedrijfsgeheimen te beveiligen. Hiervoor moeten andere oplossingen worden gezocht.

Een snelle 'oplossing'

Het lijkt eenvoudig; haal alle data door een 'vertaler' en het is voor niemand meer te herkennen. Maar daar draait het nu juist om. Het mag niet herkenbaar zijn, maar moet wel natuurgetrouw blijven. Een alternatief is het inzetten van een aparte omgeving waarin de gegevens worden ge-encrypt. Nu is het vaak niet meer te lezen met reguliere tools, maar de gegevens zijn via de applicaties die ermee moeten werken wel herkenbaar, en een uitdraai is vaak zo gemaakt.
Daarnaast kun je er natuurlijk voor kiezen om alleen specifieke gebruikers naar specifieke gegevens te laten kijken. Dit is een heel praktische oplossing, maar het zal zeker niet in alle gevallen een oplossing zijn. Het is heel vervelend als iemand ineens als crimineel wordt doorgegeven in een interface. En nog gevoeliger als ook zijn adresgegevens worden meegestuurd. Een test van de printstraat kan dan ineens heel veel problemen geven.
De ultieme oplossing zou kunnen zijn dat je tegen iedereen vertelt dat "de gegevens fictief zijn", maar dit werkt alleen als je heel geloofwaardig overkomt. Niet iedereen die de 'test' gegevens gebruikt heeft deze uitspraak meegekregen.

Problemen

Naast herkenbaarheid van situaties zijn er ook andere aspecten waaraan men moet denken. Het aanpassen van gegevens dient zeer gericht te gebeuren. Niet alleen om herkenbaarheid te houden, maar ook omdat systemen vaak anders reageren op specifieke waarden. Een praktisch voorbeeld is het aanpassen van het 'geslacht van vrouw naar man'. Dit is vanuit dataperspectief een technisch eenvoudige ingreep, maar de consequenties kunnen binnen systemen vaak wel verstrekkend zijn. Als het een getrouwde vrouw betrof, had ze misschien ook wel een tweede eigen naam, verschillen in rechtenopbouw. Maar het gaat ook om aanpassing van salarisgegevens, totalen op facturen zonder doorberekening naar de details, of het aanpassen naar een fictief sofinummer.
De complexiteit wordt nog groter als een belangrijk gegeven als sleutel wordt gebruikt. Een bankrekening of persoons/sofinummer wordt toch op verschillende plaatsen gebruikt als sleutel. Het anonimiseren van sleutels leidt vaak tot veel en complexer algoritmen. Als het niet consistent wordt gedaan zullen de data niet meer consistent zijn met alle gevolgen van dien.
Zeker bij het onderzoeken van functionaliteit in systemen via de data (bijvoorbeeld bij *reverse engineering* projecten) of het analyseren van datakwaliteit is het van belang dat gegevens herkenbaar blijven voor analisten. De vraag moet gesteld worden of men de gegevens in zo'n situatie ook daadwerkelijk *moet* anonimiseren en of het tot verkeerde conclusies kan leiden.
Zo komen de belangen van de klant en het project soms in het nauw. De belangrijkste vraag is dan ook of en hoe je de gegevens herkenbaar kunt houden en toch zo kunt 'verbouwen' dat deze niet te traceren zijn en onherkenbaar zijn voor mensen die er schade mee (kunnen) berokkenen.

Bepaal noodzaak voor anonimiseren

Stap één is het bepalen van de redenen om gegevens te anonimiseren. Is het wel nodig en wat als het niet gebeurt? In veel gevallen is een beschermde omgeving afdoende om gegevens goed te beveiligen. Tenslotte zul je het ook in een productieomgeving goed moeten afschermen voor oneigenlijk gebruik; een goede testcase dus. Een belangrijke toetsing hierbij is de eventuele juridische noodzaak.

Bepaal welke gegevens het betreft

Als het inzetten van een beschermde omgeving niet afdoende is en data daadwerkelijk geanonimiseerd moeten worden: bepaal eerst welke gegevens geanonimiseerd moeten worden. Vaak blijkt dat specifieke gegevens niet hoeven te worden geanonimiseerd, omdat deze door de context te veranderen geen betekenis meer hebben. Voor acceptatietesten is het van belang dat de gebruikers zich nog wel kunnen herkennen in het systeem.
Het loont om de te anonimiseren gegevens eerst goed in kaart te brengen en vervolgens specifieke algoritmen af te spreken.

> De complexiteit wordt groter als een belangrijk gegeven als sleutel wordt gebruikt

Voorbeelden van algoritmen zijn:

- Het vervangen van characterreeksen (a wordt z, b wordt a, c wordt z, d wordt a enzovoort, met als resultaat een serie characters aa.azzazaa.azz). Bijzondere tekens zoals ö of ç blijven echter wel in stand en kunnen van groot belang zijn in testwerkzaamheden;
- Het vervangen van de cijfers 1 tot en met 5 door 1, 6 tot en met 8 door 5, 0 en 9 blijven gelijk. De nummers zijn bijna niet meer te traceren. Maar voldoet het banknummer nog wel aan de 11-proef?;
- Alle keys door elkaar husselen en vervolgens via een vertaaltabel alle sleutelvelden in de gegevens te vervangen. Daarna kan men de vertaaltabel verwijderen zodat niemand meer weet hoe het door elkaar is gehaald.

Bepaal de definitieve scope en algoritmen

Er zijn diverse oplossingen, alle met eigen voor- en nadelen. De ene oplossing is eenvoudiger dan de andere. Maar wees bewust van de consequenties. Gegevens zijn niet meer te traceren naar een individu en bij specifieke problemen dus ook niet meer naar de originele waarden waar zich mogelijk een probleem heeft voorgedaan.

Een klant stelde voor om alle gegevens (inclusief alle coderingen) te anonimiseren. Dit werd afgeraden omdat ook alle hulptabellen daarmee consistent moesten worden geanonimiseerd. Niet alleen zou dat veel werk zijn, maar vooral een onherkenbaar systeem opleveren waarin alle coderingen onherkenbaar zouden worden: 'woonadres' zou dan weleens 'azzazazzaa' kunnen worden. Of de codering '16' zou gaan verwijzen naar '19', dat niet het 'woonadres' maar 'werkadres' is (als dit al voorkomt!). Dus dat levert een hoop extra werk op, veel risico's en veel extra tijd voor bijvoorbeeld testen.

Realiseren van de anonimisering

Als de anonimiseringskenmerken in kaart zijn gebracht en de impact duidelijk is, kan het in werking worden gezet. Het proces moet herhaalbaar zijn en geen onverwachte resultaten geven. Als er in velden specifieke tekens worden gebruikt is het verstandig om deze te negeren waardoor specifieke kenmerken toch intact blijven; wat van groot belang is voor bijvoorbeeld testen en migraties. Een randomizer kan een getal ernstig beschadigen en geeft daarnaast bij (bijna) elke iteratie een ander resultaat. Dat maakt het werken met de gegevens ook steeds anders. Een tweede testcyclus met een persoon met een specifiek kenmerk bestaat mogelijk ineens niet meer.

Conclusie

Anonimiseren is vaker noodzakelijk dan men denkt en wordt door de globalisering van projecten steeds belangrijker. Maar ook als de gegevens binnen het bedrijf en op locatie blijven, is het van belang om hier goed over na te denken. Creatieve oplossingen kunnen helpen, maar een goede aanpak en goede afwegingen en hulpmiddelen maken het proces stabieler en gecontroleerd.

Denk niet alleen maar aan dat 'ene' project maar maak tijdig goede afspraken binnen de organisatie over hoe en welke gegevens en in welke situaties moeten worden geanonimiseerd. Dit voorkomt dat tijdens projecten steeds opnieuw afspraken met gegevenseigenaren, auditors, security-afdelingen, gebruikers, management, ontwikkelaars enzovoort moeten worden gemaakt. Ga maar eens na hoeveel projecten de gegevens op dagbasis nodig hebben. Het maken én vastleggen van goede afspraken over gebruik van (échte of anonieme) gegevens loont snel en voorkomt veel irritaties, zoals gegevens die op straat komen te liggen als gevolg van een test, of twee keer een rekening met hoge 'testbedragen', of belgedrag dat inzichtelijk wordt. Een anonimiseringsproces gaat niet alleen om de eigen gegevens, maar over het omgaan met vertrouwelijke gegevens van klanten. Voorzorgen zijn dus op haar plaats.

Nico Klaassen is werkzaam bij Capgemini als principal data consultant. **Maurice Siteur** is werkzaam bij Capgemini als senior test consultant.