Operačný program
Efektívna
verejná správa

Európska únia
Európsky sociálny fond

MINISTERSTVO
INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA
A INFORMATIZÁCIE
SLOVENSKEJ REPUBLIKY

Výstup č. 1.1.5

Štandardizácia anonymizácie údajov

Zmluva o dielo č. 445/2022

Projekt:

Zlepšenie využívania údajov vo verejnej správe

ITMS kód projektu:

314011S979

Tento projekt je podporený z Európskeho sociálneho fondu.

Obsah

Zoznam skratiek	
Skratka	**Vysvetlenie**
ABC	Atribútové poverenia (*Attribute-based credentials*)
AES	Pokročilý šifrovací štandard (*Advanced Encryption Standard*)
AI	Umelá inteligencia (*Artificial Intelligence*)
CIP	Centrálna integračná platforma
CSPRNG	Kryptograficky bezpečný generátor pseudonáhodných čísel (*Cryptography Secure Pseudo-Random Number Generators*)
DNA	Deoxyribonukleová kyselina
ECC	Kryptografia eliptickou krivkou (*Elliptic Curve Cryptography*)
EÚ	Európska únia
GDPR	Všeobecné nariadenie o ochrane údajov (*General Data Protection Regulation*)
HMAC	Kľúčovaná hašovacia funkcia (*Keyed-Hash Message Authentication Code*)
IEC	Medzinárodná elektrotechnická organizácia (*International Electrotechnical Commission*)
ISO	Medzinárodná organizácia pre štandardizáciu (*International Organization for Standardization*)
KAV	Konsolidovaná analytická vrstva
MAC	Autentifikačný kód správy (*Message Authentication Code*)
NIST	Národný inštitút pre štandardy a technológiu (*National Institute of Standards and Technology*)
NZ	Nezverejniteľné údaje
OVM	Orgány verejnej moci
PID	Osobné identifikátory (*Personal Identifiers*)
PII	Osobne identifikovateľné informácie (*Personally Identifiable Information*)
PIMS	Systém na správu osobných údajov (*Personal Information Management Systems*)
RNG	Generátor náhodných čísel (*Random Number Generator*)

Zoznam skratiek	
Skratka	**Vysvetlenie**
RSA	*Rivest-Shamir-Adleman (asymetrický šifrovací algoritmus)*
SHA	Bezpečný hašovací algoritmus (*Secure Hash Algorithms*)
SSL	Secure Sockets Layer
TLS	Transport Layer Security
Z	Zverejniteľné údaje
ZKP	*Zero-Knowledge Proof*

1 Manažérske zhrnutie

Tento dokument - 1.1.5 Štandardizácia anonymizácie údajov – sa venuje nasledujúcim bodom:

- klasifikácii a kategorizácii údajov a súvisiacim politikám podľa kategórie,
- definícii metód anonymizácie a pseudonymizácie,
- výberu vhodných metód anonymizácie a pseudonymizácie pre jednotlivé scenáre a prípady použitia,
- popisu dostupných nástrojov anonymizácie a pseudonymizácie,
- popisu algoritmov anonymizácie a pseudonymizácie.

Dokument sa nevenuje aktualizácii štandardu pre anonymizáciu údajov z dôvodu, že na základe poskytnutých informácii od Ministerstva investícií, regionálneho rozvoja a informatizácie SR takýto štandard nie je definovaný a zavedený do praxe. Momentálne sa téma anonymizácie rieši v oblasti otvorených údajov, pri ktorej padlo rozhodnutie ponechať návrh a realizáciu riešenia na jednotlivých vlastníkov údajov, ktoré sa majú zverejniť. To znamená, že pre anonymizáciu otvorených údajov nevznikol štandard ani centrálna služba anonymizácie, ale anonymizácia sa realizuje v zdrojovom informačnom systéme alebo na zdrojovom datasete určenom na zverejnenie. Z toho dôvodu tento dokument predstavuje okrem návrhov štandardov pre konkrétne scenáre a prípady použitia aj teoretické východisko v téme anonymizácie a pseudonymizácie. **Okrem iného ponúka návody a modelové príklady, ako sa na tému anonymizácie a pseudonymizácie pozerať komplexne v rámci štátnej správy, keďže v čase písania dokumentu neboli k dispozícii konkrétne datasety alebo prípady použitia, ktoré by bolo treba prakticky vyriešiť z pohľadu pseudonymizácie a/alebo anonymizácie.**

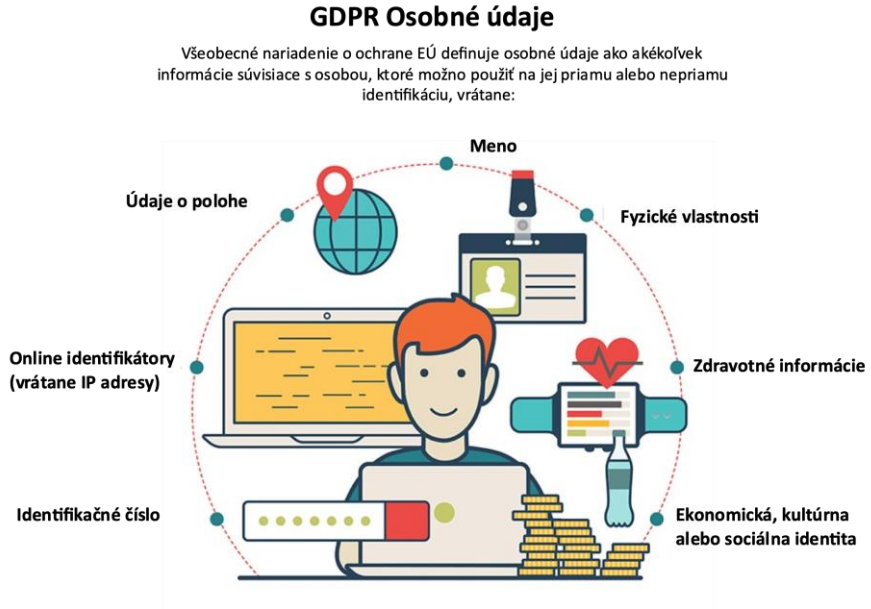
Účelom výstupu je navrhnúť štandardy, ktoré budú použiteľné pre účely anonymizácie a pseudonymizácie v rôznych scenároch a prípadoch použitia, vrátane KAV (Konsolidovaná analytická vrstva). Znamená to, že štandardy umožnia, aby bolo možné údaje analyzovať a zverejňovať s ohľadom na minimalizáciu ohrozenia ochrany osobných údajov.

Presná štandardizovaná definícia anonymizácie v dokumente vychádza z ISO/IEC 29100:2011 - Information technology — Security techniques — Privacy framework. Pseudonymizáciu definujeme v dokumente podľa slovníka pojmov Nariadenia EU 2016/679 (GDPR).

Anonymizovať údaje je nevyhnutné predovšetkým pri zverejňovaní datasetov. Pri využívaní surových dát pre dátovú vedu a dátové analýzy je nevyhnutná dôsledná pseudonymizácia, pokiaľ by anonymizácia zamedzila využiteľnosť dát pre konkrétnu dátovú analýzu. Existuje viacero prístupov k anonymizácii údajov, ktoré majú svoje výhody, nevýhody a obmedzenia. V dokumente sa nachádzajú rôzne vzorky dát s odskúšanými prístupmi, na základe ktorých sú vypracované odporúčania pre ďalšie smerovanie. V dokumente sú odporučené aj rôzne nástroje ako aj ukázané, čo dokáže samotný nástroj Talend.

2 Úvod

Potreba anonymizovať alebo pseudonymizovať osobné údaje je zakotvená v európskom Všeobecnom nariadení o ochrane údajov („General Data Protection Regulation (GDPR)“) (EU) EU 2016/679 a v Zákone č. 18/2018 Z. z. o ochrane osobných údajov. Nasledujúci obrázok znázorňuje kategórie osobných údajov tak, ako ich definuje GDPR, pričom osobné údaje môžu identifikovať osobu buď priamo alebo nepriamo.



Obrázok 1: Kategórie osobných údajov podľa GDPR[1]

Anonymizácia a pseudonymizácia dát sú procesy, ktorými sa zabezpečuje ochrana osobných údajov a tým aj súkromie jednotlivcov. **Anonymizácia** dát je proces, ktorým sa odstránia alebo zmenia osobné údaje tak, aby nebolo možné identifikovať jednotlivé osoby. Tento proces je založený na odstránení alebo zmene osobných údajov, ako sú mená, adresy, dátumy narodenia a čísla účtov. Anonymizácia je nevratný proces, ktorý nedovolí vytvorenie mechanizmu na spätnú identifikáciu. Samotná anonymizácia nie je celkom vhodná na ochranu osobných údajov. Podľa normy ISO/IEC 29100:2011 je *„**anonymizácia proces, pri ktorom sú osobne identifikovateľné údaje nezvratne zmenené tak, že dotknutú osobu z týchto údajov nemožno identifikovať či už priamo alebo nepriamo, a to ani prevádzkovateľom osobných údajov alebo ani v spolupráci so žiadnou inou stranou.“*** Nepovažuje sa za preferovanú ochranu osobných údajov, pretože ak sa nájdu nové spojitosti alebo údaje, ktoré môžu byť použité na identifikáciu jednotlivca alebo ktoré treba k nemu doplniť, už ich nebude možné vyhodnotiť a použiť. Spätná identifikácia údajov je potrebná, aby sa zabezpečilo, že údaje sú správne identifikované a správne využívané. Ďalej pomáha zabezpečiť, že údaje o jednotlivých subjektoch sa dajú postupne zhromažďovať, zdieľať a analyzovať. Okrem toho spätná identifikácia údajov môže pomôcť pri ochrane osobných údajov, pretože môže poskytnúť

[1] Zdroj: https://www.techtarget.com/searchdatamanagement/answer/What-is-included-in-the-GDPR-definition-of-personal-data. Dátum referencie: 28.02.2023

kontrolu nad tým, ako sa údaje využívajú. Napríklad ak sa používa na pseudonymizáciu šifrovanie s vytvorením kľúča pre rôznych používateľov údajov, podľa využitia konkrétneho kľúča na dešifrovanie možno odsledovať a logovať, kto a kedy s údajmi pracoval. Sú však prípady použitia, kedy je anonymizácia postačujúca alebo priam želaná.

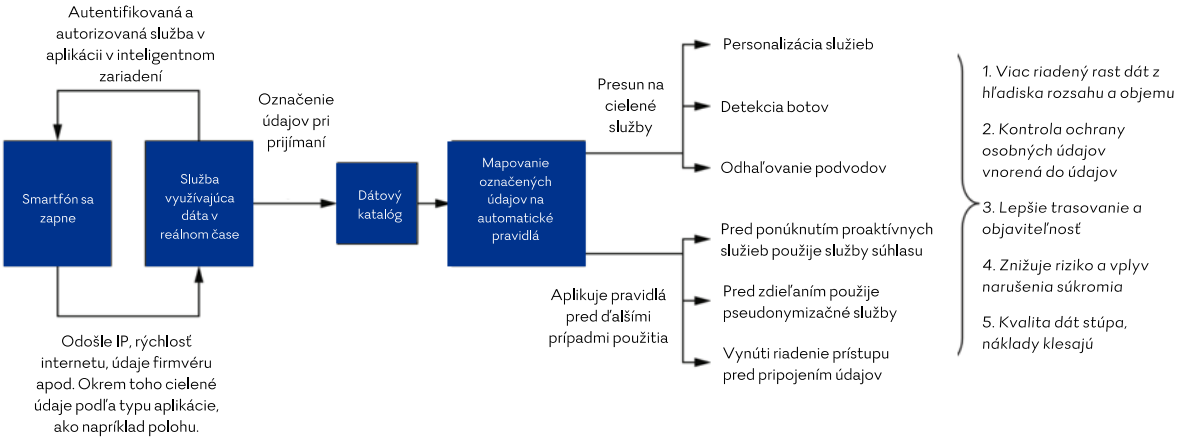
Pseudonymizácia je proces, ktorým sa zmení alebo odstráni identifikátor z údajov, pričom sa ale vytvorí náhrada, ktorá umožňuje následnú asociáciu s originálnym identifikátorom. Pseudononymizácia je reverzibilný proces, ktorý umožňuje vytvorenie mechanizmu na spätnú identifikáciu. Pseudononymizácia je vhodná na ochranu osobných údajov. Podľa slovníka pojmov Nariadenia GDPR je pseudonymizácia: ***„spracúvanie osobných údajov takým spôsobom, aby osobné údaje už nebolo možné priradiť konkrétnej dotknutej osobe bez použitia dodatočných informácií, pokiaľ sa takéto dodatočné informácie uchovávajú oddelene a vzťahujú sa na ne technické a organizačné opatrenia s cieľom zabezpečiť, aby osobné údaje neboli priradené identifikovanej alebo identifikovateľnej fyzickej osobe.“***

Tieto procesy je nevyhnutné dodržiavať predovšetkým pri nasledujúcich scenároch:

1. zverejňovaní datasetov na portáli otvorených údajov (tomuto scenáru sa venuje dokument 5.2.1),
2. využívaní surových dát pre dátovú vedu a dátové analýzy v rámci Konsolidovanej analytickej vrstvy,
3. v neposlednom rade pri využívaní osobných údajov v rôznych scenároch (napríklad pri zdieľaní údajov medzi jednotlivými ústrednými orgánmi štátnej správy alebo s tretími stranami alebo pre zabezpečenie ochrany osobných údajov pri bezpečnostnom incidente).
4. pri vypracovávaní štatistických výkazov a plnení spravodajskej povinnost[2]i, v zmysle Zákona č. 540/2001 Z. z. o štátnej štatistike a podľa súvisiacich metodík o ochrane osobných údajov.

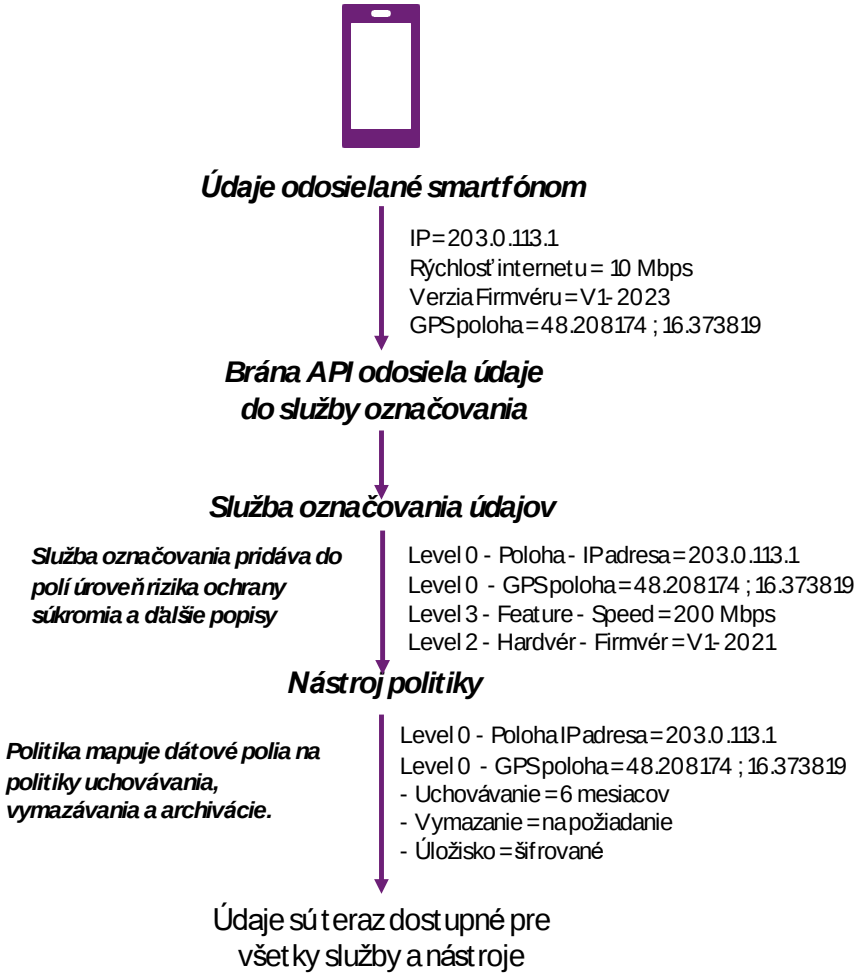
V tomto dokumente sa budeme zaoberať scenárom 2, 3 a 4. **Procesy anonymizácie a pseudonymizácie sú kľúčové pre zachovanie súkromia osobných údajov ("Data privacy“). Opatrenia na ochranu osobných údajov treba aplikovať hneď pri „vstupe“ údajov do organizácie a zakomponovať ju tak do samotného návrhu informačných systémov a súvisiacich technológií a organizačných opatrení („privacy by design“).** To vedie k oveľa efektívnejšej správe údajov, väčšej kontrole nad tým, kto má k čomu prístup, a k oveľa nižšej pravdepodobnosti porušenia ochrany osobných údajov [10.] [14.]. Cieľom tohto dokumentu ako aj nadväzujúcich štandardov v oblasti modelovania údajov, bezpečnosti a ochrany údajov, štandardizácie dôveryhodných údajov a analýzy toku údajov je, aby verejná správa dokázala fungovať pri práci s dátami a vytváraní dátami riadených služieb aj podľa príkladu na obrázku nižšie, podľa ktorého môžu fungovať dátové analýzy a zdieľanie údajov v takmer reálnom čase.

[2] Zdroj: https://intrastat.statistics.sk/Intrastat/prirucka-copy/, Dátum referencie: 06.04.2023



Obrázok 2: Príklad správnej implementácie ochrany údajov v organizácii pri implementácii personalizovaných a proaktívnych služieb s využitím dát v reálnom čase zo smartfónu

V typickom prípade sa pre každú službu vopred určia dátové polia, ktoré by mali existovať so všetkými nastavenými hodnotami a politikami v dátovom katalógu a ktoré sa budú ukladať v databáze. Databáza a katalóg sú prepojené. Čiže určenie politík, dát a ich hodnôt sa deje na základe parametricky nastavených hodnôt v katalógu. Na obrázku vyššie (Obrázok 2) je príklad služby využívajúcej **dáta v reálnom čase**, kedy niektoré štandardizované dátové prvky ako IP adresa, rýchlosť pripojenia, firmvér alebo poloha môžu byť katalogizované aj automatizovanými službami. Základom je ihneď na vstupe dát do informačných systémov verejnej správy po autentifikácii a autorizácii používateľa ich metadáta správne zaviesť do dátového katalógu (ak sa už v tomto katalógu nenachádzajú, tomu sa venuje dokument 1.1.2 Štandardizácia pre modelovanie údajov) a označiť úroveň ich citlivosti. Následne údaje treba mapovať na automatizované dátové politiky predtým, než sa dáta využívajú v ďalších službách, analýzach či sa zdieľajú tretím stranám, pričom často sa takéto dátové toky v reálnom čase ani v celom rozsahu neukladajú v databázach. V organizácii je potrebné identifikovať riziká ochrany súkromia v mieste prijímania dát do systémov a následne vytvárať správne nástroje, automatizáciu a procesy na presadzovanie ochrany súkromia. Táto postupnosť - správa údajov v katalógu a následne aplikované nástroje - je dôležitá a pomôže IT špecialistom zlepšiť ochranu osobných údajov a zároveň zvýšiť kvalitu a produktivitu údajov. Obrázok 3 ukazuje, ako má fungovať označovanie a katalogizácia . Tieto techniky budú oveľa podrobnejšie popísané neskôr v tomto dokumente, ale tento diagram ukazuje, ako sa budú meniť hodnoty jednotlivých polí metadát hneď po vstupe dát do ekosystému organizácie za predpokladu, že sú implementované automatizované služby označovania údajov. Dáta sú prijaté s ich základnými hodnotami a potom sa k nim pridá značka, ktorá označuje riziko ohrozenia súkromia používateľa. "Služba označovania údajov" zjednodušene naznačuje celú infraštruktúru inventarizácie údajov, o ktorej pojednáva štandard modelovania údajov v dokumente 1.1.2 Štandardizácia pre modelovanie údajov. Toto označovanie v počiatočnej fáze umožní priradiť k údajom vynútiteľné zásady zaobchádzania s údajmi (vymazávanie, uchovávanie atď.). Vytvára sa tak architektúra ochrany súkromia, v ktorej sú mechanizmy ochrany súkromia zakomponované už na začiatku, v samotných údajoch. Obrázok 3 poukazuje na jednoduchú vec: **neexistuje žiadny tajný recept na ochranu súkromia - len včasná identifikácia a automatizovaná orchestrácia aplikovania politík a nástrojov.**



Obrázok 3: Označovanie dát na vstupe do systémov a mapovanie riadenia ochrany súkromia = inžinierstvo ochrany súkromia v akcii

V posledných niekoľkých rokoch sa získalo mnoho informácií a poznatkov o ochrane súkromia, či už z analýzy povinne hlásených kybernetických bezpečnostných incidentov alebo z dobrej praxe dodávateľov nástrojov zameraných na ochranu súkromia. Vzhľadom na toto množstvo zdrojov majú dnešní lídri možnosť navrhnúť stratégie ochrany súkromia, ktoré dokážu zabrániť zlyhaniam, ktoré vedia organizácii zásadne uškodiť, predovšetkým stratou dôvery tých najdôležitejších – občanov a podnikateľov.

IT a dátoví špecialisti musia byť schopní uvažovať o nástrojoch na ochranu súkromia v troch tematických skupinách:

1. **Vedieť** - Vedieť, kde objavíte a lokalizujete citlivé údaje (V dokumente 1.1.2 Štandardizácia pre modelovanie údajov).
2. **Redukovať** – Redukovať objem exponovaných údajov, pričom minimalizujete „styčnú plochu“ so súkromím používateľov prostredníctvom zastretia a vymazania. (Tejto téme sa venujeme v tomto dokumente, najmä cez retenčnú politiku v kapitole 3.2.1 a techniky anonymizácie a pseudonymizácie popísané v kapitolách 6 a 7).
3. **Chrániť** - chrániť tam, kde presadzujete riadenie prístupu k údajom (V dokumente 1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov).

Keď IT špecialisti kupujú alebo vytvárajú nástroje, musia rozumieť tomu, čo riešia a ako by zvažovaný nástroj alebo prístup k tomuto riešeniu fungoval. Potom musia spraviť túto kritickú voľbu: vytvoria si nástroje na ochranu súkromia vo vlastnej réžii, alebo si kúpia hotové riešenia tretích strán, ktoré môžu siahať od komplexných platforiem na ochranu súkromia až po užšie zamerané riešenia? Týmto otázkam a výberu možných nástrojov sa budeme venovať v dokumentoch 1.1.2 Štandardizácia pre modelovanie údajov a 1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov. Tento dokument ponúkne prehľad typov nástrojov, ktoré implementujú techniky anonymizácie a pseudonymizácie.

Tabuľka 1: Rámec pre uvažovanie o nástrojoch pre ochranu údajov

Vedieť	Redukovať	Chrániť
Inventarizácia a katalóg	**Minimalizácia údajov**	**Riadenie prístupov**
Kde sa údaje nachádzajú?	Zbierajte menej údajov	Autentifikácia
O aký typ údajov ide? (Napríklad kategória citlivosti údajov)	Zastrite alebo anonymizujte údaje	Autorizácia
Kto a prečo ich zbiera alebo používa?	Nastavte retenčnú politiku (Vymažte údaje, pokiaľ ich už podľa zákona alebo z iných dôvodov netreba)	Rozsah prístupov k dátam, úrovne prístupov, logovanie prístupov

3 Kategorizácia a klasifikácia údajov z pohľadu ochrany súkromia

Na rozdiel od tradičných odvetví, ako je poľnohospodárstvo, infraštruktúra a zdravotníctvo, sú technológie vo svojej podstate odlišné, pokiaľ ide o vzťah medzi výrobou a prácou. V týchto tradičnejších odvetviach je potrebných veľa pracovníkov na dôslednú premenu plánov na produkty. To nie je prípad technologických pracovných miest, pri ktorých je veľkým lákadlom využívanie automatizácie na dosiahnutie väčšieho objemu výstupov s menším objemom práce a s menším počtom iterácií.[3]

Klasifikácia údajov je rozhodujúcim krokom, ktorého cieľom je vniesť disciplínu do vzťahu medzi technologickým sektorom a používateľmi, ktorí sú na základe týchto údajov identifikovaní. Tento proces a jeho výsledok pomôže podnikom a organizáciám vyhodnotiť svoj zber údajov z pohľadu používateľov, ktorých údaje zhromažďujú. Klasifikácia údajov by mohla podnikom a organizáciám pomôcť vyhnúť sa možným problémom s ochranou osobných údajov, preukázať všetkým zainteresovaným stranám, že podnik alebo organizácia nepovažuje svojich používateľov za tovar. Klasifikácia údajov umožňuje opatrnejšie nakladať s údajmi (alebo ich rýchlejšie vymazať) v súlade s tým, čo klasifikácia o údajoch hovorí. Hoci klasifikácia nemusí riešiť širšiu otázku ekonomickej nerovnosti spôsobenej delením bohatstva v sektore technológií, tento proces poskytuje ľudskejšiu optiku, cez ktorú sa možno pozerať na údaje a používateľov, o ktorých tieto dáta sú. **Štát ako regulátor a inovátor na poli osobných údajov by mal byť spoločnosti a súkromnému sektoru vzorom.** Mal by zdieľať svoju dobrú prax v klasifikácii údajov a v súvisiacich politikách narábania s osobnými údajmi. **Klasifikácia údajov dokonca môže pomôcť pri určovaní hodnoty osobných údajov, ktoré budú v budúcnosti občania schopní speňažiť prostredníctvom Systémov na správu osobných údajov („Personal Information Management Systems (PIMS)“), implementovaným aj v rámci národného projektu pre Manažment osobných údajov** (Kapitola 10.2).

Klasifikácia údajov odpovedá na tieto otázky pre každý typ údajov, ktoré sa zhromažďujú alebo môžu byť uložené:

- O aké údaje ide z hľadiska objemu a definície?
- Prečo ich potrebujeme zbierať?
- Čo nám to hovorí o našich zákazníkoch a našej agende?
- Čo by sa stalo, keby sa s týmito údajmi nesprávne zaobchádzalo?
- Ako dlho ich smieme alebo musíme uchovávať?

V prelomovom „white paper“ spoločnosti Microsoft "Klasifikácia údajov pre pripravenosť na cloud" sa uvádza:

„Klasifikácia údajov predstavuje pre organizácie jeden zo základných spôsobov, ako určiť a priradiť relatívne hodnoty údajom, ktoré majú k dispozícii. Proces klasifikácie údajov umožňuje

[3] NishantBhajaria,“Why isn’t the tech boom helping the economy?” LinkedIn ,5. máj 2015, http://mng.bz/v4j1, Dátum referencie: 27.01.2023

organizáciám kategorizovať uložené údaje podľa citlivosti a vplyvu na chod organizácie s cieľom určiť riziká spojené s údajmi. Po ukončení tohto procesu môžu organizácie spravovať svoje údaje spôsobom, ktorý odráža ich hodnotu, namiesto toho, aby so všetkými údajmi zaobchádzali rovnako. Klasifikácia údajov je vedomý a premyslený prístup, ktorý organizáciám umožňuje realizovať optimalizácie, ktoré by pri priradení rovnakej hodnoty všetkým údajom nemuseli byť možné.[4]

Podľa „white paper" je veľmi dôležité, aby boli vedúci pracovníci dôkladne oboznámení s klasifikáciou údajov. Patria sem riaditelia, bezpečnostní špecialisti, systémoví architekti a IT odborníci, ktorí sú zodpovední za plánovanie vývoja a nasadenie aplikácií alebo infraštruktúry v organizácii.

3.1 Kategorizácia údajov z pohľadu ochrany súkromia

Údaje, pri ktorých treba aplikovať osobitné politiky správy aktív, pretože sú zásadné pre fungovanie organizácie, ochranu jej kľúčových znalostí, zákazníkov ako aj povesti, sa nazývajú citlivé. Základná kategorizácia citlivých údajov rozlišuje dve úrovne:

1. **zverejniteľné údaje (Z)** – údaje, ktorých zverejnenie ako otvorené údaje neohrozuje fungovanie štátu, organizácie či podniku a jeho systémov, alebo neodhaľuje intelektuálne vlastníctvo, a preto ich je možné kedykoľvek zverejniť.
2. **nezverejniteľné údaje (NZ)** – údaje, ktoré nie je vhodné zverejniť v žiadnom prípade, pretože zverejnenie nesie riziko okamžitého alebo neskoršieho pokusu o narušenie informačnej bezpečnosti. Pre nezverejniteľné údaje môže existovať podmienka, za splnenia ktorej sa stanú zverejniteľnými.

Nezverejniteľným údajom sa budeme v širšom kontexte venovať v dokumente 1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov. **V tomto dokumente pod nezverejniteľnými údajmi rozumieme len osobné údaje.** Podľa Zákona č. 18/2018 Z. z. § 2: „Osobnými údajmi sú údaje týkajúce sa identifikovanej fyzickej osoby alebo identifikovateľnej fyzickej osoby, ktorú možno identifikovať priamo alebo nepriamo, najmä na základe všeobecne použiteľného identifikátora, iného identifikátora, ako je napríklad meno, priezvisko, identifikačné číslo, lokalizačné údaje, alebo online identifikátor, alebo na základe jednej alebo viacerých charakteristík alebo znakov, ktoré tvoria jej fyzickú identitu, fyziologickú identitu, genetickú identitu, psychickú identitu, mentálnu identitu, ekonomickú identitu, kultúrnu identitu alebo sociálnu identitu."

Podľa Národného inštitútu pre štandardizáciu a technológiu („National Institute of Standards and Technology"): **Osobne identifikovateľné informácie** („Personally Identifiable Information (**PII**)") predstavujú akúkoľvek reprezentáciu informácií, ktorá umožňuje prirodzene odvodiť totožnosť jednotlivca, ku ktorému sa tieto informácie vzťahujú, buď priamymi alebo nepriamymi prostriedkami. **Všetky PII sú osobné údaje, ale nie všetky osobné údaje sú PII**. PII môže byť akákoľvek informácia, ktorá umožňuje identifikovať jednotlivca a preniesť o ňom nejaké znalosti. Preto to môže byť plné meno, adresa, číslo pasu, e-mail, číslo kreditnej karty, dátum narodenia, telefónne číslo, prihlasovacie údaje a mnoho ďalších.

[4] "Klasifikácia údajov pre pripravenosť na cloud", Microsoft, http://mng.bz/Xrv1, Dátum referencie: 27.01.2023

Osobné identifikátory („Personal Identifiers (**PID**)“) sú podmnožina údajov PII, ktoré identifikujú jedinečného jednotlivca a umožňujú inej osobe "prebrať" totožnosť jednotlivca bez jeho vedomia alebo súhlasu.

Osobné údaje sú oveľa širšou kategóriou ako PII alebo PID. Podľa Metodického pokynu pre kategorizáciu citlivosti údajov[5] z dôvodu bezpečnosti má bezpečnostná dokumentácia obsahovať politiku správy aktív, v ktorej sa okrem iného nachádza aj metodika kategorizácie údajov podľa citlivosti (Z) a konkrétne zaradenie údajov podľa citlivosti (NZ). Politike správy aktív predchádza identifikácia aktív, ktorá je súčasťou bezpečnostného projektu podľa platnej metodiky a jeho bezpečnostnej dokumentácie vzhľadom na legislatívne požiadavky zákona č. 95/2019 Z.z. a jeho vykonávacích predpisov pre potreby informačného systému verejnej správy. Odporúčame, aby sa štandardom stali aj ďalšie tri podkategórie nezverejniteľných údajov (NZ) podľa nasledujúcej tabuľky. Zavedením ďalšej úrovne – podkategórie do štandardu sa táto úroveň zjednotí a umožní aplikovať jednotné a cielené politiky na ochranu súkromia na tieto podkategórie vo verejnej správe. Obdobná kategorizácia je dobrou praxou aj pri vyhodnocovaní umiestnenia údajov do verejných cloudov[6].

Tabuľka 2: Podkategórie nezverejniteľných údajov (NZ)

Kategória (Úroveň)	Definícia
Vyhradené	Osobné alebo iné údaje, ktoré podliehajú najprísnejším požiadavkám na spracovanie vzhľadom na ich citlivosť a riziko pre organizáciu a zákazníkov v prípade nesprávnej manipulácie.
Dôverné	Osobné alebo iné údaje, ktoré podliehajú prísnym požiadavkám na spracovanie vzhľadom na ich citlivosť a riziko v prípade nesprávneho zaobchádzania.
Interné	Údaje, ktoré sú zamestnancom a prípadným tretím stranám k dispozícii na základe zmluvy/dohody o (zachovaní) mlčanlivosti výlučne v dôsledku ich zamestnania v organizácii alebo prebiehajúceho projektu či poskytovania služby a ktoré nie sú kategorizované ako dôverné alebo obmedzené.

Zdroje, ktoré sa vyčlenia na ochranu údajov, by preto mali závisieť od rizika ohrozenia súkromia. V hierarchii údajov by sa mala venovať značná časť zdrojov údajom, ktoré sú veľmi citlivé – v tabuľke vyššie (Tabuľka 2) sú označené ako „Vyhradené“. Je logické, že takéto údaje môžu identifikovať občanov a ich správanie, ale môžu obsahovať aj údaje dôležité pre organizáciu.

Ďalšia úroveň údajov, ktoré treba chrániť, nemusí byť taká citlivá ako tie, ktoré sú na úrovni „Vyhradené“. Podľa toho treba kalibrovať stratégiu ochrany týchto údajov. Treba zdôrazniť, že organizácie nemôžu vyhlásiť údaje za dôverné len preto, že považujú bezpečnostné opatrenia a opatrenia na ochranu súkromia vyžadované pre úroveň „Vyhradené“

[5] Zdroj: http://www.informatizacia.sk/ext_dok-metodicky_pokyn_citlivost_udajov_podla_ib_v1-0/15139c , Dátum referencie: 09.02.2023

[6] Zdroj: https://ico.org.uk/media/for-organisations/documents/1540/cloud_computing_guidance_for_organisations.pdf, Dátum referencie: 04.04.2023

za príliš náročné. Údaje, ktoré patria do úrovne „Vyhradené“, zvyčajne spĺňajú aspoň podmnožinu nasledujúcich kritérií:

- Jednoznačne identifikujú konkrétnu osobu. Ide o subjektívne kritérium: pod menom ako "Jozef Novák" nie je možné jednoznačne identifikovať osobu, ak nie je sprevádzaná ďalšími údajmi, ako je adresa bydliska. Ale meno ako "Barack Obama" ponúka oveľa vyššiu úroveň identifikovateľnosti, keďže nejde o typické meno na Slovensku.
- Tieto údaje je možné spojiť s ďalšími ľahko dostupnými údajmi, aby bolo možné identifikovať konkrétneho jednotlivca a jeho činnosti alebo preferencie.
- Informácie o jednotlivcovi sprístupnené v týchto údajoch ho zaraďujú do jedinečnej skupiny. Predpokladajme napríklad, že sa v rámci monitoringu riešenia „eRecepty“ uchováva tabuľka obsahujúca údaje o ľuďoch, ktorí užívajú lieky na diabetes typu 2. V tabuľke sú používatelia identifikovaní prostredníctvom náhodných ID, aby sa nemenovali. Takáto tabuľka by mohla byť bezpečná z hľadiska ochrany súkromia, ak obsahuje údaje pre celé Slovensko vzhľadom na potenciálne vysoký počet ľudí. Ak by tá istá tabuľka obsahovala informácie o ľuďoch žijúcich v dedine Belejovce v okrese Svidník, mohla by predstavovať riziko ohrozenia súkromia, keďže v tejto dedine žije len 17 ľudí.

Zjednodušene povedané, údaje na úrovni „Vyhradené“ majú tendenciu byť individualizované, zatiaľ čo dôverné údaje majú tendenciu byť viac agregované. Z dôvodu súvisiacich dôsledkov na narušenie súkromia sa na údaje s obmedzeným prístupom na úrovni „Vyhradené“ vzťahujú prísnejšie kontroly prístupu a kratšie lehoty uchovávania, zatiaľ čo dôverné údaje môžu mať voľnejšie požiadavky na prístup a dlhšie lehoty uchovávania.

3.1.1 Segmentácia nezverejniteľných údajov

Klasifikácia údajov je relatívne jednoduchá, keď sa všetky dátové polia, ktoré dátoví špecialisti pre ochranu súkromia a bezpečnosť považujú za citlivé, zaradia do kategórie „Vyhradené“. Organizácie často automatizujú presadzovanie politík na základe takýchto kategórií. Napríklad všetky údaje označené ako „Vyhradene“ môžu byť šifrované pri uchovávaní („at rest“) aj pri prenose. Dátoví špecialisti pre ochranu súkromia však môžu segmentovať údaje tak, že len údaje, ktoré sú skutočne citlivé, dostanú prísnu ochranu súkromia, zatiaľ čo ostatné údaje môžu byť voľnejšie prístupné. Segmentovať údaje možno napríklad nasledovne:

- **Údaje o jednotlivcoch** - tieto údaje by opisovali konkrétne osoby, ktoré by mohli byť osobne identifikované, a teda poškodené, ak by sa narušilo ich súkromie. Tieto údaje by sa mohli ďalej segmentovať takto:
  - údaje o zamestnancoch,
  - údaje o dodávateľoch,
  - údaje o občanoch,
  - údaje o podnikateľoch.

  Údaje o jednotlivcoch by podliehali ochrane súkromia, ale organizácia môže chcieť ponúknuť rôzne úrovne ochrany rôznym druhom jednotlivcov. Napríklad občania môžu mať nárok na prísnu ochranu súkromia. Na druhej strane dodávatelia môžu podliehať sledovaniu s cieľom zmierniť riziko zneužitia dôverných informácií a krádeže informácií. Ak by sa všetky údaje patriace jednotlivcom klasifikovali rovnakým spôsobom, vznikol by prístup "jeden súbor opatrení pre všetkých", ktorý by buď príliš alebo nedostatočne chránil údaje.

- **Údaje o veciach** - organizácie musia zabezpečiť aj údaje, ktoré identifikujú objekty ako sú výrobky, návrhy, miesta atď. Tieto údaje môžu byť pre organizáciu kritické a kľúčové pre jej agendu a udržanie relevantnosti. Tieto údaje však nemusia podliehať kontrolám zameraným na ochranu súkromia, ako je vymazanie, zakrytie atď. Týmto údajom sa venujeme v dokumente 1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov. Postup kategorizácie a klasifikácie musí umožniť, aby sa tieto údaje identifikovali, označili a chránili inak, ako keby išlo o údaje jednotlivcov. Aj v tomto prípade je potrebné upozorniť, že je možné, že údaje o veciach by mohli byť spojené alebo prepojené s údajmi, ktoré identifikujú osoby. Preto by podrobná klasifikácia týchto údajov mohla pomôcť zaviesť a sledovať ochranu súkromia v budúcnosti.
- **Údaje v agregácii** - riziko ohrozenia súkromia údajov nie je nemenné a statické. Pri agregácii a zamaskovaní či zastretí údajov sa môže riziko ohrozenia súkromia citeľne znížiť, ako sa tomu budeme venovať v kapitole 6.

  Napríklad skupina záznamov o používateľoch, ktorá neobsahuje špecifické identifikačné údaje používateľov (ako sú mená), ale obsahuje adresu trvalého pobytu každého používateľa, môže byť spravidla označená ako „Vyhradené“. Riziko ochrany osobných údajov však môžete znížiť agregovaním používateľov na základe PSČ, v ktorom bývajú, a vylúčením adries trvalého pobytu zo súboru údajov. Mohli by ste si ponechať len tie záznamy používateľov, ktorí sú v skupine 100 alebo viac používateľov na PSČ. To umožní vykonávať experimenty prispôsobené agregovaným súborom údajov bez toho, aby tieto súbory údajov podliehali rovnako prísnym opatreniam na ochranu súkromia ako údaje o jednotlivcoch.

  Agregovať údaje sa tiež dajú na základe časových línií, trendov atď. Kľúčový záver je nasledovný: transformácia súborov údajov z tých, ktoré opisujú jednotlivcov, na tie, ktoré sa pozerajú na kolektív, by mohla pomôcť kategorizovať ich ako také, ktoré majú nižšie obmedzenia ochrany osobných údajov.

Tento kontext segmentácie dát slúži hlavne na to, aby sa organizácie nenechali zlákať extrémami. Buď sú príliš opatrné a kategorizujú veľké objemy údajov ako „Vyhradené“, alebo sú príliš sebavedomé a podceňujú riziká ochrany súkromia. Pohľad na údaje vo väčšom kontexte umožňuje kategorizáciu, ktorá je presnejšia a udržateľnejšia na presadenie. Takýto prístup tiež lepšie odráža fungovanie modernej dátami riadenej organizácie. Údaje, infraštruktúra a mikroslužby[7] sú prispôsobené tak, aby zodpovedali ich účelu. Je veľmi dôležité, aby dátoví špecialisti v oblasti ochrany údajov kategorizovali údaje spôsobom, ktorý vyvažuje potreby plnenia agendy a zároveň kladie ochranu údajov na vrchol priorít.

[7] Ide o spôsob realizácie služieb (service) – implementáciu nejakej funkcionality, ktorá je typicky poskytovaná servermi na internete alebo cez intranet, pomocou mikroslužieb. Celá aplikácia sa skladá z väčšieho počtu niekoľko služieb alebo niekoľko desiatok takzvaných mikroslužieb, pričom každá taká mikroslužba má presne definovanú rolu (väčšinou len jednu!), beží v samostatnom procese (procesoch) a komunikuje buď priamo či nepriamo s ostatnými mikroslužbami, typicky s využitím nejakého štandardného protokolu (REST API cez HTTP a HTTPS, poprípade nejaký protokol pre messaging). Cieľom využitia mikroslužieb je dosiahnuť vyššiu škálovateľnosť aplikácie a služieb.

3.1.2 Štruktúrované verzus neštruktúrované údaje

Neštruktúrované údaje sú údaje, ktoré sa nedajú jednoducho uložiť do tradičnej stĺpcovej/riadkovej databázy alebo tabuľky (napríklad ide o JSON blob[8]). Vo vnorených objektoch JSON sa často nachádzajú citlivé údaje ako napríklad IP adresy, ktoré sa dajú použiť na identifikáciu niektorých používateľov. Neštruktúrované údaje, hoci sa často prehliadajú, sa dajú zneužiť, a preto by sa mali spravovať s rovnakou starostlivosťou ako štruktúrované údaje (údaje uložené v tradičnej stĺpcovej/riadkovej databáze).

Naproti tomu štruktúrované údaje sú údaje, ktoré sa riadia vopred definovaným dátovým modelom a sú zamerané na prípady použitia, ktoré si vyžadujú analýzu.[9] Štruktúrované údaje zvyčajne zodpovedajú tabuľkovému formátu s definovaným vzťahom medzi jednotlivými riadkami a stĺpcami, ako napríklad databáza SQL.

Podľa časopisu Forbes rastie objem neštruktúrovaných údajov, ktoré podniky zhromažďujú a ukladajú, každoročne o 55-65 %.[10] Podľa TechRepublic je 80 % údajov, ktoré podniky spracúvajú, neštruktúrovaných.[11] Neštruktúrované údaje sa vzhľadom na svoju povahu analyzujú ťažšie ako štruktúrované údaje a nedajú sa ľahko prehľadávať, preto boli pre organizácie až do posledných rokov nepoužiteľné. Dnes však máme nástroje na analýzu neštruktúrovaných údajov poháňané umelou inteligenciou (AI), ktoré boli vytvorené špeciálne na prístup k poznatkom dostupným z neštruktúrovaných údajov. Organizácie musia pochopiť typy neštruktúrovaných údajov, ktoré zhromažďujú, a najlepšie spôsoby spracovania a ukladania týchto údajov. Platí to najmä preto, že neštruktúrované údaje sťažia implementáciu požiadaviek Zákona o ochrane osobných údajov.

Veľa neštruktúrovaných údajov končí v protokoloch alebo vnorených bloboch JSON, kde sú údaje často „zahrabané“ vnútri komplikovaných štruktúr. Zatiaľ čo na zisťovanie citlivých údajov, ako sú IP adresy, sa dali ľahko použiť nástroje ako REGEX, ak je IP adresa zahrabaná hlboko v neštruktúrovaných údajoch, môže sa stať, že ju nikdy neodhalíte, a preto sa vám ju nepodarí včas odstrániť. Vzory REGEX síce môžu zodpovedať neštruktúrovaným údajom, ale samotný objem zhromažďovaných údajov by mohol viesť k tomu, že algoritmy sa časovo oneskoria a nepodarí sa včas tieto údaje odhaliť a identifikovať ako citlivé osobné údaje.

Neštruktúrované údaje sú príkladom toho, že inovatívne metódy získavania dát, vďaka ktorým sú údaje dostupné s vysokou rýchlosťou, môžu predstavovať riziko pre súkromie. Povaha takýchto údajov poskytuje nové možnosti dátovým vedcom, ale sťažuje život IT a dátovým špecialistom zaoberajúcim sa ochranou súkromia. Z tohto vyplýva, že **politika ochrany osobných údajov zahŕňa správu štruktúrovaných aj neštruktúrovaných údajov**.

[8] https://jsonblob.com/, Dátum referencie: 27.01.2023

[9] Dátové typy: Štruktúrované vs. neštruktúrované údaje," Enterprise Big Data Framework, 9. januára 2019, http://mng.bz/yJlo , Dátum referencie: 27.01.2023

[10] Bernard Marr, "Čo sú neštruktúrované údaje a prečo sú pre podniky také dôležité? Jednoduché vysvetlenie pre každého," Forbes, 16. októbra 2019, http://mng.bz/MvND , Dátum referencie: 27.01.2023

[11] Mary Shacklett, "Neštruktúrované údaje: TechRepublic, 14. júla, http://mng.bz/aZW9 , Dátum referencie: 27.01.2023

3.2 Klasifikácia údajov z pohľadu ochrany súkromia

Aby sa proces klasifikácie údajov nestával izolovaným, **je kľúčové sa vyhnúť vytváraniu klasifikácií pre rôzne prípady použitia**, ako napríklad jeden súbor klasifikácií pre údaje, ktoré je potrebné šifrovať, a iný pre údaje, ktoré je potrebné vymazať, atď. Chcete, aby vaša klasifikácia určovala výsledky, a nie aby výsledky určovali klasifikáciu údajov.

Demokratický rozhodovací proces zdola nahor, ktorý umožňuje inovácie systémov a procesov, často nefunguje s iniciatívami na ochranu súkromia, ako je klasifikácia údajov. Je potrebný model, v ktorom majú hlas IT a dátoví špecialisti a dátoví vedci, pretože budú prijímať taktické rozhodnutia týkajúce sa údajov. Avšak vrcholové vedenie musí urobiť konečné rozhodnutie o klasifikácii a prevziať zodpovednosť za riziká s ňou spojené. Diskusie a debaty musia ustúpiť rozhodnutiam. Tieto poznatky sú dôležité, pretože klasifikácia údajov je nákladná a je základom viacerých budúcich rozhodnutí o ochrane osobných údajov. Úlohou technických lídrov, ktorí vedú IT, dátové a produktové tímy, je poskytnúť svojim tímom dostatočné politické krytie a motivovať ich k strategickému prístupu. Pomôže to zabezpečiť, aby IT a dátoví špecialisti a produktoví manažéri mysleli ďalej ako na nasledujúci týždeň alebo mesiac a aby vytvárali výsledky, ktoré pomôžu pripravenosti, kvalite údajov a ich bezpečnosti, ako aj celkovej vyspelosti organizácie v oblasti manažmentu údajov. Konkrétne sa budeme rozhodovaciemu procesu a kompetenčnému modelu venovať v kroku 2 procesu klasifikácie, popísaného neskôr v tejto kapitole.

Keďže definovanie PII je pre organizácie, ktoré v minulosti zhromažďovali obrovské množstvo údajov, výzvou, je kľúčové, aby organizácie cieľavedome zisťovali, aké údaje zhromažďujú a ako ich chránia. Preto je nevyhnutné, aby si každá organizácia, ktorá chce rozšíriť svoje úsilie v oblasti ochrany osobných údajov, vytvorila systém klasifikácie údajov podľa nasledujúcej tabuľky (Tabuľka 3), ktorá vychádza z GDPR. Organizácia, ktorá dokáže vytvoriť klasifikáciu údajov, ktorá sa vyvíja na základe meniacich sa zákonov o ochrane osobných údajov a prispôsobuje sa im, nemusí doháňať, keď regulátor začne zákony aktívne presadzovať. V tomto zmysle sa implementácia klasifikácie údajov podobá vlastnej verzii softvéru na prípravu daní, ktorý je v podstate abstrakciou daňových zákonov, ktorú bežní daňoví poplatníci nemusia nikdy pochopiť.

Tabuľka 3: Klasifikácia osobných údajov z pohľadu ochrany súkromia

Kategória	Skupina	Popis	Klasifikácia osobných údajov podľa GDPR
Vyhradené	Meno	Meno a priezvisko, rodné priezvisko	PII
	Rodinní príslušníci	Previazanie s rodičmi, potomkami, partnermi a podobne	PII
	Pobyt	Adresa trvalého alebo prechodného pobytu	PII
	Identifikačné údaje	Číslo občianskeho preukazu, číslo pasu, rodné číslo, číslo zdravotného poistenia, daňové	PID

Kategória	Skupina	Popis	Klasifikácia osobných údajov podľa GDPR
		identifikačné číslo, číslo vodičského preukazu	
	Platobné údaje	Číslo kreditnej karty (s dátumom platnosti alebo bez neho), číslo bankového účtu, informácie o platobných službách tretích strán (napr. Viamo, PayPal).	PID
	Lokalizačné údaje	Poloha mobilného telefónu či už z bázovej stanice, podľa IP adresy alebo GPS súradníc	PII
	Kontaktné údaje	Kontaktná adresa, číslo mobilného telefónu, číslo pevnej linky	PID
	Rasový pôvod alebo etnický pôvod	Patrí síce do skupiny demografických údajov, ale vzhľadom na klasifikáciu podľa GDPR ide o samostatnú skupinu.	Osobitná kategória podľa § 16 Zákona č. 18/2018 Z. z.
	Politické názory	Patrí síce do skupiny demografických údajov, ale vzhľadom na klasifikáciu podľa GDPR ide o samostatnú skupinu.	Osobitná kategória podľa § 16 Zákona č. 18/2018 Z. z.
	Náboženská viera	Patrí síce do skupiny demografických údajov, ale vzhľadom na klasifikáciu podľa GDPR ide o samostatnú skupinu.	Osobitná kategória podľa § 16 Zákona č. 18/2018 Z. z.
	Filozofické presvedčenie	Patrí síce do skupiny demografických údajov, ale vzhľadom na klasifikáciu podľa GDPR ide o samostatnú skupinu.	Osobitná kategória podľa § 16 Zákona č. 18/2018 Z. z.
	Členstvo v odborových organizáciách	Patrí síce do skupiny demografických údajov, ale vzhľadom na klasifikáciu podľa GDPR ide o samostatnú skupinu.	Osobitná kategória podľa § 16 Zákona č. 18/2018 Z. z.
	Genetické údaje	Ide o digitalizovaný záznam reťazca deoxyribonukleovej kyseliny (DNA), zmapované gény jednotlivca a jeho unikátne mutácie a podobne.	Osobitná kategória podľa § 16 Zákona č. 18/2018 Z. z.

Kategória	Skupina	Popis	Klasifikácia osobných údajov podľa GDPR
	Biometrické údaje	Ide o osobný údaj fyzickej osoby, na základe ktorého je osoba jednoznačne a nezameniteľne určiteľná. Jedná sa o biologické vlastnosti, fyziologické znaky črty alebo opakované činnosti, v prípade ktorých sú tieto vlastnosti a/alebo činnosti špecifické pre konkrétneho človeka a zároveň merateľné. Biometrickými údajmi sú napr. odtlačok prsta, odtlačok dlane, analýza deoxyribonukleovej kyseliny (DNA) a pod.	Osobitná kategória podľa § 16 Zákona č. 18/2018 Z. z.
	Údaje týkajúce sa zdravia	Zdravotná anamnéza, informácie o doplnení lekárskeho predpisu, informácie o zdravotnej starostlivosti, existujúce ochorenia atď.	Osobitná kategória podľa § 16 Zákona č. 18/2018 Z. z.
	Údaje týkajúce sa sexuálneho života alebo sexuálnej orientácie fyzickej osoby	Patrí do skupiny údajov o preferenciách a správaní sa jednotlivcov, vzhľadom na vysokú mieru intimity a klasifikáciu podľa GDPR ide o samostatnú skupinu.	Osobitná kategória podľa § 16 Zákona č. 18/2018 Z. z.
	Údaje týkajúcich sa uznania viny za spáchanie trestného činu alebo priestupku	Patrí do skupiny údajov o správaní sa jednotlivcov, vzhľadom na vysokú mieru intimity a klasifikáciu podľa GDPR ide o samostatnú skupinu.	Osobitná kategória podľa § 17 Zákona č. 18/2018 Z. z.
	Online identifikátory	Prihlasovacie meno užívateľa informačného systému a heslo, e-mailová adresa, IP adresa, PČO, certifikáty, trvalé identifikátory zariadenia	PID
Dôverné	Demografické údaje	Vek, pohlavie, rodinný stav	PII len za špeciálnych okolností
	Fyzické atribúty	Výška, váha, charakteristické črty na tvári alebo na tele (znamienka, jazvy a podobne)	PII len za špeciálnych okolností

Kategória	Skupina	Popis	Klasifikácia osobných údajov podľa GDPR
	Údaje o IKT (okrem online identifikátorov)	Logy (prihlasovanie / odhlasovanie užívateľa), nestále identifikátory zariadenia, história prehliadania internetu a vyhľadávania, rôzne súbory na pamäťových médiách, vlastné údaje databáz a registrov (t.j. používateľské údaje), nastavenie prístupových práv používateľov informačných systémov, ako aj ďalšie IKT, ktoré môžu mať určitú vypovedaciu schopnosť vedúcu k identifikácii konkrétnej fyzickej osoby. Konkrétne posúdenie citlivosti údajov možno nájsť v Metodickom pokyne pre kategorizáciu citlivosti údajov z dôvodu bezpečnosti.	PII len za špeciálnych okolností
	História transakcií	Požadované služby, poskytnuté služby, dátum a čas poskytnutia služby, účtovaná suma a mena	PII len za špeciálnych okolností
	Ekonomická situácia	Výška zárobkov, nezamestnanosť, povolanie	PII len za špeciálnych okolností
	Záznamy o osobnom vlastníctve	Vlastníctvo nehnuteľností, automobilov, podnikov a iných organizácii	PII
	Zvukové, elektronické, vizuálne, tepelné a čuchové informácie	Záznamy z kamier, senzorov, mikrofónov a podobne.	PII

3.2.1 Retenčné politiky podľa klasifikácie osobných údajov

Retenčné politiky pre ochranu osobných údajov sú dané predovšetkým GDPR a Zákonom č. 18/2018 Z.z o ochrane osobných údajov.

V rámci informačných povinností sú prevádzkovatelia povinní informovať dotknuté osoby, ktorých osobné údaje spracúvajú, o dobe uchovávania ich osobných údajov. Dotknuté osoby teda majú práva vedieť ako dlho budú ich jednotlivé osobné údaje uchovávané, respektíve kedy budú vymazané.

Doba uchovávania musí byť prevádzkovateľom stanovená v súlade s GDPR a zákonom o ochrane osobných údajov, a to najmä v súlade so zásadou minimalizácie uchovávania osobných údajov. To znamená, že údaje môžu byť spracovávané a uchovávané po dobu, ktorá je nevyhnutne potrebná na naplnenie alebo dosiahnutie konkrétneho účelu, pre ktorý boli osobné údaje získané. Spracúvanie osobných údajov na vedecké, historické a štatistické účely alebo na účely informovania verejnosti, tzv. privilegované účely spracúvania vyplývajúce z osobitných zákonov, majú špecifický režim a na tieto účely je možné osobné údaje spracovávať aj po dosiahnutí účelu, pre ktorý boli osobné údaje pôvodne získané.

Sú prípady, kedy dobu uchovávania určujú nepriamo aj niektoré zákony (napr. zákon o účtovníctve, zákonník práce a podobne), respektíve registratúrny poriadok a registratúrny plán prevádzkovateľa, ktoré reflektujú na zákonné povinnosti stanovené takýmito zákonmi.

Doba uchovávania osobných údajov môže byť určená konkrétnym spôsobom (napríklad 5 rokov od poskytnutia údajov), ale aj popisným spôsobom určenia doby uchovávania, ak nie je vopred jasné, v ktorom momente táto doba skončí (napríklad dosiahnutím účelu spracúvania – ukončením zmluvy uzatvorenej na dobu neurčitú).

V prípade kamerových záznamov z verejných priestorov sa v zmysle odporúčaní European Data Protection Board za primeranú dobu uchovávania považuje doba 3 až 5 dní, a to s prihliadnutím na príslušný účel spracúvania, okrem prípadov, kedy bude záznam použitý napríklad ako dôkaz v trestnom konaní.

Doba uchovávania sa zapíše aj do dátového katalógu, definovaného v dokumente 1.1.2 Štandardizácia pre modelovanie údajov.

Ďalším právom dotknutej osoby, ktorá má vplyv na dobu uchovávania, je právo na výmaz. Dotknutá osoba môže v zmysle GDPR a zákona o ochrane osobných údajov kedykoľvek požiadať prevádzkovateľa o výmaz svojich osobných údajov. Prevádzkovateľ je povinný tak bez zbytočného odkladu urobiť, ak:

- osobné údaje nie sú potrebné na účel, na ktorý boli získané,
- dotknutá osoba odvolala súhlas so spracúvaním svojich osobných údajov,
- dotknutá osoba namietala spracúvanie svojich osobných údajov,
- osobné údaje sa spracúvajú nezákonne,
- je dôvodom na výmaz splnenie povinnosti podľa zákona alebo medzinárodnej zmluvy,
- sa osobné údaje získavali v súvislosti s ponukou služieb informačnej spoločnosti v zmysle § 15 zákona o ochrane osobných údajov.

V prípade, že boli osobné údaje zverejnené, je prevádzkovateľ povinný údaje nielen vymazať, ale aj prijať primerané opatrenia, aby boli osobné údaje vymazané aj u ostatných prevádzkovateľov. Po skončení určenej doby uchovávania musia byť príslušné osobné údaje efektívne vymazané, v opačnom prípade prevádzkovateľom hrozia pokuty až do výšky 10 mil. EUR.

3.2.2 Proces klasifikácie osobných údajov

V moderných organizáciách je klasifikácia údajov zvyčajne výsledkom podrobného skúmania a rokovaní. Medzi kľúčové subjekty patria:

- Právne oddelenie venujúce sa ochrane osobných údajov,

- Technické oddelenie venujúce sa ochrane súkromia,
- Tím pre bezpečnosť,
- Tím IT špecialistov pre vývoj IKT,
- Produktové tímy,
- Dátoví vedci.

Hoci sa ochrana osobných údajov všeobecne považuje za právnu oblasť, bolo by veľkou chybou nechať tento proces klasifikácie riadiť len právnym tímom. Právnici môžu zaujať príliš defenzívny prístup tým, že uplatňujú zákon bez kontextu vykonávania agendy, alebo môžu dať IT a dátovým špecialistom príliš voľnú ruku a veriť vo vlastnú schopnosť vyhrať na súde. Oba prístupy nie sú optimálne. Proces klasifikácie spravidla prebieha v troch krokoch, popísaných ďalej, a je založený na dobrej zahraničnej praxi [10.] [14.], ako aj na štandarde ISO 27001:2022[12]

1. krok: Spolupráca so zainteresovanými stranami v rámci rôznych funkcií pri klasifikácii údajov,
2. krok: Formalizácia klasifikácie údajov,
3. krok: Refaktorizácia klasifikácie údajov.

1. krok: Spolupráca so zainteresovanými stranami v rámci rôznych funkcií pri klasifikácii údajov

V prvom kroku právni zástupcovia pre ochranu osobných údajov vyjadria svoju predstavu o tom, ako by klasifikovali údaje. Zároveň v spolupráci s IT a dátovými špecialistami, manažérmi projektov a produktov a dátovými vedcami treba porozumieť, aké údaje sú legitímne potrebné na prevádzkové a analytické účely. Je veľmi dôležité pochopiť, prečo v rámci klasifikácie údajov treba získať vstupy od týchto rôznorodých zainteresovaných strán. Praktický príklad pomôže objasniť niektoré skutočnosti.

Predpokladajme, že ste vedúcim oddelenia ochrany osobných údajov v organizácii, ktorá poskytuje aplikáciu s názvom "Výjazdy", ktorá pomáha vodičom optimalizovať svoje výjazdy pri plnení rôznych úloh v teréne. Pre jednoduchosť zvolíme veľmi triviálny prípad použitia, ktorý by zahŕňal začiatok výjazdu na mieste A, zadanie adresy cieľa B a aplikácia by vám potom poskytla pokyny, ktoré vám pomôžu dostať sa z miesta A do miesta B. Na backende by bola databáza, ktorá by mohla vyzerať ako Tabuľka 4.

Tabuľka 4: Backendová databáza pre aplikáciu "Výjazdy"

Meno a priezvisko používateľov aplikácie	E-mail	Východisková adresa (zemepisná šírka/dĺžka)	Koncová adresa (zemepisná šírka/dĺžka)
Jozef Novák	jnovak@gmail.com	5 desatinných miest	5 desatinných miest
Peter Malík	peter.malik@zoznam.sk	5 desatinných miest	5 desatinných miest

[12] Zdroj: https://www.itgovernance.co.uk/blog/what-is-information-classification-and-how-is-it-relevant-to-iso-27001, Dátum referencie: 06.04.2023

Meno a priezvisko používateľov aplikácie	E-mail	Východisková adresa (zemepisná šírka/dĺžka)	Koncová adresa (zemepisná šírka/dĺžka)
Jana Dúbravská	jdubrav@yahoo.com	5 desatinných miest	5 desatinných miest
Daniel Mihál	daniel@mihal.sk	5 desatinných miest	5 desatinných miest
Petra Kováčová	kovacova@gmail.com	5 desatinných miest	5 desatinných miest
Juraj Lipták	jliptak@zoznam.sk	5 desatinných miest	5 desatinných miest

Právny tím by tvrdil, že každý riadok tejto tabuľky by jednoznačne identifikoval jednotlivca a zameral by sa na dva dôvody:

1. E-mailové adresy používateľov aplikácie sú jednoznačne priradené ku konkrétnym osobám. E-mail by sa dal prepojiť s ďalšími údajmi o používateľoch na internete, čím by sa vytvoril podrobný profil používateľa. Vzhľadom na potenciál takýchto údajov by sa tým výrazne zvýšil negatívny vplyv kompromitácie databázy alebo zneužitia údajov.
2. Adresy - východiskové aj koncové - sú veľmi presné. Pri GPS lokalizácii platí, čím viac desatinných miest v adrese, tým presnejšia je poloha. Právny tím by mohol prísť s návrhom, aby sa každý záznam (každý riadok v tabuľke) označil ako „Vyhradené“ s prísnymi kontrolami toho, kto k nemu môže pristupovať, a prepojil ho s politikou na krátke obdobie uchovávania.

Tím dátovej vedy by sa mohol brániť tým, že jeho schopnosť analyzovať jednotlivé údaje, ťažiť tieto údaje pre optimalizáciu trás, aktualizáciu máp na základe využívania ulíc atď. do veľkej miery závisí od zhromažďovania týchto údajov v priebehu času a pozorovania vzorov, ktoré sa v nich objavujú. Mohli by namietať, že databázu a údaje v nej obsiahnuté treba klasifikovať tak, aby sa dali uchovávať dlhšie a s menšími obmedzeniami. Bežne sa stáva, že právny tím má pocit, že IT a dátoví špecialisti a dátoví vedci sú príliš neopatrní, pokiaľ ide o citlivosť údajov, zatiaľ čo IT a dátoví špecialisti a analytici obviňujú právnikov z neústupnosti a z toho, že prístup k údajom je bežný v celom odvetví - čo je typ argumentu „súkromný sektor to už robí, musíme držať krok s dobou".

Tím IT a dátových špecialistov môže navrhnúť riešenie, ktoré vyzerá ako Tabuľka 5, kde sú polia, ktoré by mohli jednoznačne identifikovať jednotlivca - jeho meno a e-mail - zakryté pomocou techniky nazývanej hašovanie. Možno stojí za to zdôrazniť, že zatiaľ čo e-maily jednoznačne identifikujú konkrétnu osobu, ktorá účet vytvorila, mená nie sú jedinečné. Ak však niekto má veľmi netradičné meno, je možné, že rovnaké meno má len veľmi málo osôb, ak vôbec nejaké. To znamená, že v záujme vyhnutia sa záznamom, ktoré by jednoznačne identifikovali konkrétneho používateľa, by tím IT a dátových špecialistov mohol zakryť mená aj e-mailové adresy.

Tabuľka 5: Upravená backendová databáza pre aplikáciu “Výjazdy”

Meno a priezvisko	E-mail	Východisková adresa (zemepisná šírka/dĺžka)	Koncová adresa (zemepisná šírka/dĺžka)
(hašované)	(hašované)	3 desatinné miesta	3 desatinné miesta
(hašované)	(hašované)	3 desatinné miesta	3 desatinné miesta

Meno a priezvisko	E-mail	Východisková adresa (zemepisná šírka/dĺžka)	Koncová adresa (zemepisná šírka/dĺžka)
(hašované)	(hašované)	3 desatinné miesta	3 desatinné miesta
(hašované)	(hašované)	3 desatinné miesta	3 desatinné miesta
(hašované)	(hašované)	3 desatinné miesta	3 desatinné miesta
(hašované)	(hašované)	3 desatinné miesta	3 desatinné miesta

Aby bola poloha menej presná, môže technický tím znížiť počet desatinných miest v koordinátach GPS uchovávaných na analýzu. Obmedzená presnosť by mohla znamenať, že počiatočné a koncové adresy opisujú oveľa väčšiu geografickú oblasť. To znižuje pravdepodobnosť, že by daná adresa jednoznačne identifikovala konkrétnu polohu domu alebo kancelárie.

Spoločné zmeny by mohli znamenať, že údaje ako celok sú teraz menej citlivé a že záznamy v databáze uvedené v tabuľke (Tabuľka 5) sú skôr na úrovni „Dôverné“ ako „Vyhradené“. To však otvára otázku, či existujú prípady použitia, ktoré potrebujú presné adresy a identity. Napríklad, bezpečnostný tím môže potrebovať prístup k podrobným informáciám o polohe a ku kontaktným informáciám v prípade, že sa vodič stal účastníkom nehody kvôli zlým pokynom aplikácie. Možno bude treba spustiť analytické služby, ktoré budú potrebovať históriu všetkých jázd používateľa. V tejto situácii by sa mohli zachovať obe verzie údajov, aj keď s výhradami:

1. Tabuľka 4 by bola k dispozícii malému tímu IT a dátových špecialistov s kontrolou prístupu, takže že by museli požiadať o prístup s odôvodnením a ich dátumy prístupu by sa zaznamenávali.
2. Tabuľka 5 by mala menej prísne požiadavky a bola by otvorenejšia z hľadiska obdobia prístupu a uchovávania.

Krok 1 predstavuje ideovú fázu, v ktorej sa zhromažďujú informácie z rôznych perspektív. V reálnom scenári však treba schému klasifikácie údajov formalizovať pomerne rýchlo, pretože IT a dátoví špecialisti a dátoví vedci budú pri rozhodovaní od nej závislí. Nasledujúce kroky 2 a 3 sú prispôsobené tak, aby umožnili scenár, v ktorom je potrebný **operatívny systém klasifikácie údajov** – to znamená, že sa pracuje priebežne na jeho vývoji s prichádzajúcimi novými informáciami a prípadmi použitia.

2. krok: Formalizácia klasifikácie údajov

V kroku 2 sa odporúča vytvoriť počiatočný klasifikačný systém na základe legislatívy a vstupov od právnych odborov ako aj z reálnej praxe od ostatných zainteresovaných strán. Za formálne pravidlá klasifikácie bude zodpovedná PS1 - Pracovná skupina pre dátové štandardy, štandardy názvoslovia elektronických služieb a formuláre (ďalej ako PS1), ktorá bude aj vlastníkom tohto procesu. Keďže v rámci tejto skupiny sa štandardizuje aj Centrálny model údajov verejnej správy, jeho taxonómie a ontológie, vytvára sa tam spoločné porozumenie údajom, ktoré možno využiť aj pre proces klasifikácie na zaradenie dátových prvkov do jednotlivých kategórií (Tabuľka 2 a Tabuľka 3). Táto skupina sa bude rozširovať a jej procesy sa budú nastavovať podľa metodiky v dokumente 1.1.2 Štandardizácia pre modelovanie údajov.

Tabuľka 6: Kompetenčný model

Členovia tímu	Zodpovednosť
PS1	Pracovná skupina má stály charakter a je zodpovedná za centrálnu koordináciu a následné opatrenia týkajúce sa štandardizácie. Pracovná skupina tiež monitoruje vzájomnú konzistenciu (fungovanie systému) pri uznávaní nových štandardov, monitoruje medzinárodné štandardy, ktoré majú vplyv na slovenský dátový program, a monitoruje proces všeobecného vývoja a úprav. Pracovná skupina pre dátové štandardy sa pravidelne stretáva, aby vyhodnotila prebiehajúce tematické pracovné skupiny, ak sú zriadené.
Tematické pracovné skupiny	Ide o skupinu expertov so znalosťou existujúcich dátových štandardov a implementácií, ktorá je zodpovedná za vývoj doménového modelu a za klasifikáciu dátových prvkov v tomto doménovom modeli.
Dátoví kurátori tematických pracovných skupín	Zodpovedajú za facilitáciu pracovných skupín a technické vypracovanie doménového modelu vo forme diagramov a špecifikácií.
Dátoví špecialisti za jednotlivé OVM	Vytvoria stručný analytický dokument ohľadom využívania dát v danom doménovom modeli, či už na rozhodovanie priamo v agende, sekundárny analytický účel alebo na zdieľanie tretím stranám, a navrhnú klasifikáciu dátových prvkov v tomto doménovom modeli podľa tabuliek (Tabuľka 2 a Tabuľka 3), aj s ohľadom na hodnotenie rizika podľa štandardu ISO 27001[13]. Môžu si tiež prizvať ako poradný orgán aj dátových špecialistov zo súkromného sektora na daný doménový model.
Zástupcovia právneho odboru pre jednotlivé OVM	Dajú svoje vyjadrenie ku klasifikácii navrhnutej dátovými špecialistami.
Zástupcovia odboru riadenia kybernetickej a informačnej bezpečnosti a vládnej jednotky CSIRT	Dajú svoje vyjadrenie ku klasifikácii navrhnutej dátovými špecialistami, podloženú aj prípadnými príkladmi z praxe. Zohľadňujú aj celkový súlad s politikou a metodikami informačnej bezpečnosti.
Úrad na ochranu osobných údajov SR	V prípade, že sa vyššie uvedená skupina odborníkov nedokáže zhodnúť na klasifikácii, má finálne slovo zástupca tohto úradu.
Projektové vedenie Dátovej kancelárie	Zodpovedá za organizovanie pracovných skupín a pozývanie odborníkov, ako aj za manažment dátového programu a komunikáciu s rôznymi zainteresovanými stranami. Zabezpečuje

[13] Zdroj: https://www.itgovernance.co.uk/blog/7-steps-to-a-successful-iso-27001-risk-assessment, Dátum referencie: 06.04.2023

Členovia tímu	Zodpovednosť
	tiež, že výsledná dohodnutá klasifikácia je zavedená v dátovom katalógu, definovanom v dokumente 1.1.2 Štandardizácia pre modelovanie údajov. Dohliada na proces riadenia zmien. Prijíma podnety na prijatie zmien a plánuje ich prekonzultovanie v rámci PS1.

V tejto oblasti musia dátoví špecialisti v jednotlivých OVM posilniť svoju vedúcu rolu v oblasti ochrany osobných údajov a aktívne sa zapojiť do štandardizačného procesu PS1, aby sa zabezpečilo vedomé rozhodovanie o tom, ako sa vytvárajú rôzne úrovne rizika a ako sa jednotlivé dátové prvky priraďujú k týmto úrovniam rizika. V tejto fáze je rizikom skončiť buď s polovičatou schémou klasifikácie údajov, ktorá pokrýva len najnaliehavejšie prípady použitia, alebo s niekoľkými rôznymi verziami klasifikácie údajov, ktoré sú prispôsobené rôznym tímom, prípadom použitia atď. Kým v krátkodobom horizonte môžu byť obe tieto možnosti skvelé, v rýchlo sa rozvíjajúcich organizáciách sa môže stať, že fenomén "naliehavosti nahradí dôležitosť", a v takom prípade sa organizácia rozhodne pre polovičatú klasifikáciu so záväzkom, že ju dokončí, ale nikdy sa k nej nedostane. Zároveň viaceré tímy IT a dátových špecialistov a tímy dátovej vedy nakoniec prijmú možno nezvratné rozhodnutia, čím sa táto polovičatá klasifikácia údajov upevní. Aby sa tomu dalo predísť, je potrebné sformalizovať napríklad polročný harmonogram klasifikácie údajov, v rámci ktorého sa zverejní napríklad verzia V1 ako oficiálny dokument v rámci zavedeného formálneho procesu PS1 a zavedie sa do spomínaného dátového katalógu. Zároveň sa otvorí kópia toho istého dokumentu vo formáte návrhu na zhromažďovanie pripomienok. To vedie k poslednému, tretiemu kroku.

3. krok: Refaktorizácia klasifikácie údajov

V kroku 3 sú ciele nasledovné:

1. Identifikujte zainteresované strany, ktoré sa nemuseli zapojiť do krokov 1 a 2, čím sa uistíte, že proces je skutočne inkluzívny a reprezentatívny pre rôzne silá.
2. Zabezpečte, aby sa všetky nové prípady použitia, ktoré sa objavia v súvislosti s reformami a inými zmenami, priebežne zapracovávali do klasifikácie údajov, ktorá bude skutočne živým dokumentom, podobne ako by mali byť živé samotné služby a technologický „stack“ organizácie.

Tretí krok je veľmi dôležitý, pretože pomáha odhaliť oblasti, v ktorých sa kľúčové zainteresované strany nemusia zhodnúť na tom, nakoľko je konkrétny dátový prvok kritický z hľadiska ochrany osobných údajov.

Alternatíva: Iteratívny proces

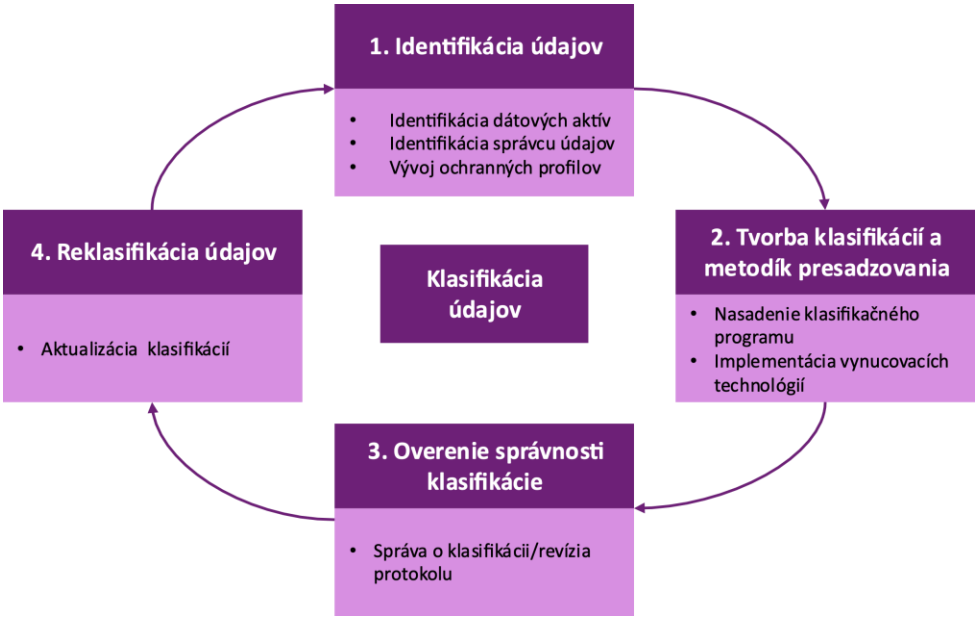
Trojkrokový proces, ktorý bol opísaný doteraz, predstavuje najefektívnejšiu a opakovanú metódu klasifikácie údajov, ktorá funguje v organizáciách rôznych veľkostí. Existuje však aj alternatíva. Spoločnosť Microsoft identifikovala model, ktorý je užitočný na replikovanie v organizáciách, ktoré sú veľké a diverzifikované, ako aj v menších organizáciách, kde nemusia existovať špecializované úlohy týkajúce sa ochrany osobných údajov.

Vo svojom „white paper“ "Klasifikácia údajov pre pripravenosť na cloud“[14] Microsoft predstavil model „Plánuj, vykonaj, skontroluj, rob nápravy“ („Plan, Do, Check, Act“) (pozri Obrázok 4):

1. **Plánuj:** Určte kľúčovú osobu z centrálneho tímu pre ochranu osobných údajov, ktorej úlohou by bolo identifikovať dátové systémy, miesta zberu údajov, systémy, cez ktoré údaje prúdia (napríklad Kafka Pipelines), systémy, v ktorých sú údaje uložené (napríklad štruktúrované databázy ako MySQL), rôzne tímy, ktoré používajú údaje, atď. Táto osoba vytvorí profil údajov, ktorý pomôže fungovaniu organizácie tým, že bude na ňom spolupracovať s multifunkčnými zainteresovanými stranami.
2. **Vykonaj**: Po odsúhlasení politík klasifikácie údajov bude táto osoba zodpovedná za zavedenie klasifikácie údajov, čo môže zahŕňať dokumenty o riadení, systémové kontroly atď.
3. **Kontroluj**: Samotná klasifikácia údajov nestačí, treba sa uistiť, že kontroly ochrany osobných údajov v organizácii, ako aj služby, na ktoré sa tieto kontroly vzťahujú, odrážajú klasifikáciu údajov. Je to veľmi dôležité, pretože účelom klasifikácie údajov je zabezpečiť, aby sa s údajmi zaobchádzalo významne odlišným spôsobom, najmä z hľadiska ochrany osobných údajov.
4. **Rob nápravy**: Klasifikácia údajov nie je jednorazová záležitosť. Organizácie sa reformujú, zákony sa neustále vyvíjajú, aktivisti za ochranu súkromia a médiá kladú otázky a tímy IT a dátových špecialistov a produktové tímy sú kreatívne pri svojich stratégiách zhromažďovania údajov. Tím pre ochranu súkromia preto bude musieť neustále klasifikovať a reklasifikovať údaje a podľa toho prispôsobovať techniky kontroly prístupu.

Obrázok 4 vysvetľuje iteratívny charakter klasifikácie údajov a to, ako tento proces zvyčajne prebieha.

[14] Microsoft, "Klasifikácia údajov pre pripravenosť na cloud", http://mng.bz/o8XD, Dátum referencie: 30.01.2023



Obrázok 4: Model klasifikácie údajov[15]

[15] Microsoft, "Klasifikácia údajov pre pripravenosť na cloud".

4 Štandardy anonymizácie a pseudonymizácie

Existujú štandardy pre anonymizáciu a pseudonymizáciu dát, ktoré je potrebné dodržiavať. Tieto štandardy sú navrhnuté tak, aby zabezpečili, že osobné údaje budú správne anonymizované alebo pseudonymizované. Štandardy pre anonymizáciu a pseudonymizáciu dát zahŕňajú:

- ISO/IEC 27001:2022 – Štandard pre riadenie bezpečnosti informácií.
- ISO/IEC 29100:2011 – Štandard pre ochranu osobných údajov – rámec ochrany súkromia.
- ISO/IEC 20889:2018: Štandard pre pseudonymizáciu a anonymizáciu: Tento štandard špecifikuje požiadavky na zabezpečenie pseudonymizácie a anonymizácie osobných údajov v súlade s princípmi ochrany súkromia, definovanými v štandarde ISO/IEC 29100:2011. Určuje tiež postupy pre odstránenie identifikátorov pre identifikáciu jednotlivca a pre všetky ďalšie spracovávanie osobných údajov.
- ISO 25237:2017 – Štandard pre pseudonymizáciu osobných údajov v zdravotníckej informatike, teda na ochranu osobných údajov o zdraví.

ISO normy pre ochranu osobných údajov poskytujú rámec pre správu, ochranu a zabezpečenie osobných údajov. Hlavné odporúčania zahŕňajú identifikáciu a hodnotenie rizík pre osobné údaje, používanie vhodných bezpečnostných opatrení na ochranu osobných údajov pred neoprávneným prístupom, zneužitím, zničením alebo stratou, kontinuálne monitorovanie a vyhodnocovanie účinnosti bezpečnostných opatrení na ochranu osobných údajov, zodpovednosť za správu, ochranu a zabezpečenie osobných údajov ktoré sa spracúvajú. V neposlednom rade je dôležitá aj transparentnosť a komunikácia s dotknutými osobami o tom, ako sa ich osobné údaje spracúvajú. Dotknuté osoby sa musia vedieť aj jednoducho dozvedieť o svojich právach.

Tieto štandardy odporúčajú nasledujúce postupy pre anonymizáciu a pseudonymizáciu osobných údajov:

- Vytvorenie bezpečnostných postupov pre spracovanie osobných údajov, vrátane vymedzenia účelu, pre ktorý sú údaje spracovávané.
- Identifikovanie údajov, ktoré musia byť premenené na pseudonymizované alebo anonymizované údaje.
- Vykonávanie pseudonymizácie a anonymizácie údajov pomocou techník ako sú šifrovanie, hashe alebo rôzne kódovacie schémy.
- Vytvorenie postupov pre zabezpečenie, aby pseudonymizované a anonymizované údaje boli uchovávané a prenášané v bezpečí.
- Pridanie mechanizmov pre detekciu a identifikáciu porušenia ochrany údajov.
- Vytvorenie postupov pre zabezpečenie, aby pseudonymizované a anonymizované údaje boli spracovávané v súlade s ostatnými právnymi predpismi.

Týmto postupom sa budeme venovať v kapitole 5 o dobrej praxi.

V norme ISO 25237:2017 sa pseudonymizácia definuje ako „osobitný typ de-identifikácie, ktorý odstraňuje spojenie s dotknutou osobou a pridáva spojenie medzi konkrétnym súborom charakteristík týkajúcich sa dotknutej osoby a jedným alebo viacerými pseudonymami“. De-

identifikácia je podľa tej istej normy „všeobecný pojem pre akýkoľvek proces redukcie asociácie medzi súborom identifikačných údajov a subjektom údajov“. Pseudonym je tiež definovaný ako „osobný identifikátor, ktorý sa líši od bežne používaného osobného identifikátora a používa sa spolu so pseudonymizovanými údajmi na zabezpečenie koherencie súboru údajov spájajúceho všetky informácie o dotknutej osobe bez toho, aby sa odhalila identita osoby v reálnom svete“. Ako poznámka k poslednej definícii sa v norme ISO 25237:2017 uvádza, že pseudonymy sa zvyčajne obmedzujú na označenie identifikátora, ktorý neumožňuje priame odvodenie bežného osobného identifikátora. Môžu byť odvodené od bežne používaného osobného identifikátora reverzibilným alebo ireverzibilným spôsobom alebo môžu byť úplne nesúvisiace.

Často dochádza k zámene pojmov pseudonymizácia a anonymizácia a ich uplatňovaniu v praxi. V norme ISO 25237:2017 sa anonymizácia definuje ako „proces, pri ktorom sa osobné údaje nezvratne zmenia takým spôsobom, že dotknutú osobu už nemožno priamo ani nepriamo identifikovať, a to buď samotným prevádzkovateľom údajov, alebo v spolupráci s akoukoľvek inou stranou“. Podobne aj NIST označuje anonymizáciu ako „proces, ktorý odstraňuje spojenie medzi identifikačným súborom údajov a subjektom údajov“. Jednoducho povedané, anonymizovaný súbor údajov neumožňuje identifikáciu žiadnej osoby, a to ani prevádzkovateľovi, ani tretej strane. Preto sa anonymizované údaje nepovažujú za osobné údaje.

Pre štandardy anonymizácie je najrelevantnejší štandard ISO/IEC 20889:2018. Tento štandard určuje terminológiu, triedenie de-identifikačných techník podľa ich charakteristík a ich uplatniteľnosť na zníženie rizika re-identifikácie. Definícia pseudonymizácie, ktorá je uvedená v tomto štandarde, uvádza, že ide o: „techniku de- identifikácie, ktorá nahrádza identifikátor (alebo identifikátory) zadávateľa údajov pseudonymom s cieľom skryť identitu tohto zadávateľa údajov“. Pseudonym je následne definovaný ako „jedinečný identifikátor vytvorený pre hlavného zodpovedného za údaje s cieľom nahradiť bežne používaný identifikátor alebo identifikátory tohto hlavného zodpovedného za údaje“.

Tento štandard ISO/IEC 20889:2018 sa vzťahuje na všetky typy a veľkosti organizácií, vrátane verejných a súkromných spoločností, štátnych orgánov a neziskových organizácií, ktoré sú prevádzkovateľmi osobných údajov alebo spracovateľmi osobných údajov a ktoré vykonávajú procesy de-identifikácie údajov za účelom zvyšovania súkromia. Podľa tohto ISO štandardu by sa mali pre **pseudonymizáciu** údajov používať nasledujúce techniky:

- Šifrovanie - Šifrovanie je metóda pre zabezpečenie osobných údajov tým, že sa premenia na nečitateľný text, ktorý je možné dešifrovať iba s použitím špeciálneho kľúča (kapitola 6.1.5).
- Hašovanie - Hašovanie je metóda pre prekódovanie osobných údajov tak, aby sa zabránilo prejaveniu identity jednotlivca (kapitola 6.1.3 a 6.1.4).
- Anonymizácia identifikátorov - Anonymizácia identifikátorov je metóda, ktorá pozostáva z odstránenia identifikátorov pre identifikáciu jednotlivca a nahradenia ich anonymnými identifikátormi (kapitola 6.1.6).
- Generovanie pseudonymov – Generovanie pseudonymov je metóda, ktorá pozostáva z vytvorenia pseudonymu, ktorý je navrhnutý tak, aby nikto nemohol identifikovať jednotlivca (kapitola 6.1.7).
- Kódovanie - Kódovanie je proces, pri ktorom sa údaje premenia na nečitateľný text pomocou kódovacej schémy (kapitola 6.2.4).
- Vytváranie pseudonymizovaných užívateľských účtov - Táto technika pozostáva z vytvorenia užívateľských účtov s pseudonymizovanými informáciami, ktoré nikto nemôže

priamo spárovať s dotknutou osobou (túto metódu neodporúčame, keďže sa nedá jednoznačne implementovať do praxe).

5 Dobrá prax pre anonymizáciu a pseudonymizáciu

V každej organizácii je v téme ochrany údajov nevyhnutné spraviť nasledujúce konkrétne rozhodnutia:

- Aké údaje zbierať.
- Ako k nim získať prístup.
- S kým ich zdieľať.
- Čo robiť s poznatkami, ktoré nám údaje prinášajú.
- Ako vhodne riadiť riziká.

Až keď máme zodpovedané tieto otázky, môžeme nastaviť očakávania v oblasti ochrany osobných údajov. Tieto očakávania sú vo všeobecnosti 4:

1. Ochrana údajov - Používatelia a regulačné orgány ako Úrad na ochranu osobných údajov očakávajú, že organizácia bude chrániť údaje o používateľoch.
2. Právo na informácie - od organizácii sa očakáva, že poskytnú kópie údajov o používateľoch na požiadanie.
3. Právo byť zabudnutý - toto právo umožňuje používateľom požiadať o vymazanie svojich údajov.
4. Zavedenie súladu s legislatívou - organizácie musia manažovať, aké údaje zhromažďujú a uchovávajú z dôvodov dodržiavania právnych predpisov.

Tieto abstraktné požiadavky sa premietajú do konkrétnych úloh v oblasti ochrany súkromia a technických mechanizmov:

- Minimalizácia údajov, keď zhromažďujete len to, čo potrebujete.
- Overovanie, ktoré zahŕňa overenie totožnosti zamestnancov a každého, kto pristupuje k údajom.
- Autorizácia, ktorá mapuje záujemcov o prístup k systému alebo údajom na oprávnenie podľa platnej politiky.
- Inventarizácia a kategorizácia údajov, ktorá si vyžaduje vytvorenie katalógu.
- Audity, ktorými sa overuje, či sú mechanizmy ochrany osobných údajov týkajúce sa vymazávania, uchovávania, atď. presadzované.

Tabuľka 7: Pohľad rôznych zainteresovaných strán na údaje: analytika, bezpečnosť/súkromie, IT systémy a nákladová efektívnosť

Pohľad na analytiku	Pohľad na bezpečnosť/súkromie
Viac je lepšie	**Menej je bezpečnejšie**
– **Zozbierajte** čo najviac údajov – **Uchovávajte** čo najdlhšie – **Duplikujte** údaje pre efektívny prístup – (+) Lepšie predpovede → vyššia spokojnosť → viac dosiahnutých cieľov	– **Zhromažďujte** čo najmenej údajov – **Uchovávajte** čo najkratšie – (+) Ľahšie dodržiavanie → nízky rozsah porušenia pravidiel → vyššia spokojnosť → menej pokút a káuz → lepší imidž organizácie
Pohľad na IT systémy	**Pohľad na nákladovú efektívnosť**
Menej je viac	**Menej je lacnejšie**
– **Zhromažďujte** čo najmenej údajov – **Duplikujte** údaje čo najmenej – (+) Menej súborov údajov → lepšia kvalita údajov → vyššia produktivita	– **Zhromažďujte** čo najmenej údajov – **Uchovávajte** čo najkratšie – (+) Nízke náklady na skladovanie → nižšie prevádzkové náklady → zvýšená hodnota za peniaze

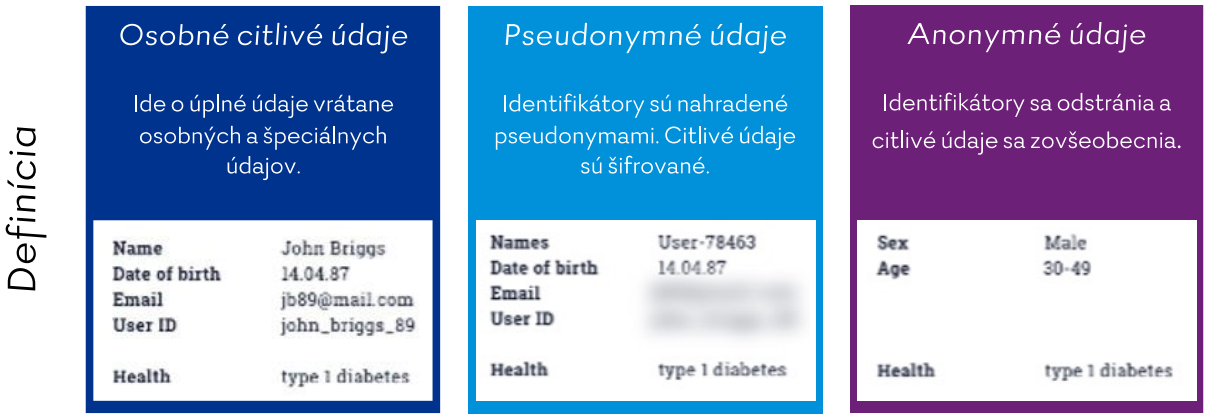
Overenou praxou, ako zabezpečiť časť týchto protichodných záujmov je aplikovať dobrú prax pseudonymizácie, ktorá je podporená aj legislatívou: „Prevádzkovateľ je povinný pred spracúvaním osobných údajov zaviesť a počas spracúvania osobných údajov mať zavedenú špecificky navrhnutú ochranu osobných údajov, ktorá spočíva v prijatí primeraných technických a organizačných opatrení, najmä vo forme **pseudonymizácie**, na účinné zavedenie primeraných záruk ochrany osobných údajov a dodržiavanie základných zásad podľa § 6 až 12.[16]“ Dobrá prax pre anonymizáciu a pseudonymizáciu dát zahŕňa:

1. Identifikáciu osobných údajov, ktoré je potrebné anonymizovať alebo pseudonymizovať.
2. Vytvorenie postupu pre anonymizáciu alebo pseudonymizáciu osobných údajov.
3. Uloženie anonymizovaných alebo pseudonymizovaných údajov v bezpečnom prostredí.
4. Monitorovanie a revízia anonymizovaných alebo pseudonymizovaných údajov.
5. Udržiavanie záznamov o anonymizácii alebo pseudonymizácii osobných údajov.

[16] § 32, (1) Zákona č. 18/2018 Z.z.

6 Anonymizačné a pseudonymizačné techniky

Častým zavádzajúcim omylom je, že pseudonymizované údaje sa vnímajú ako anonymné údaje (Obrázok 5**).** Tak to však nie je. Ak si pripomenieme príslušné definície z kapitoly 4, pseudonymizácia súvisí s existenciou spojenia medzi osobnými identifikátormi a pseudonymami, zatiaľ čo pri anonymizácii by takéto spojenie nemalo byť k dispozícii žiadnym spôsobom. Preto je pri pseudonymizácii možná (a od prevádzkovateľa údajov sa dokonca vyžaduje) spätná identifikácia, zatiaľ čo pri anonymizácii to v zásade neplatí. Inými slovami, **pseudonymizované údaje sú stále osobnými údajmi, zatiaľ čo anonymizované údaje nimi nie sú, a teda nespadajú už pod GDPR.** Takýto omyl bol aj súčasťou prípadu známeho incidentu spoločnosti AOL v roku 2006, keď bola zverejnená databáza obsahujúca dvadsať miliónov kľúčových slov pre vyhľadávanie viac ako 650 000 pseudonymov za obdobie troch mesiacov, čo následne viedlo k identifikácii viacerých používateľov[17].



Obrázok 5: Rozdiel medzi osobnými, pseudonymizovanými a anonymizovanými údajmi[18]

Napriek tomu sa pojem "anonymný" často používa v bežnom jazyku na opis prípadov, keď je totožnosť dotknutých osôb iba skrytá (ale údaje nie sú skutočne anonymizované). Existuje napríklad niekoľko takzvaných "anonymných" aplikácií sociálnych sietí, ktoré zvyčajne nevyžadujú od svojich používateľov vytvorenie profilov a zhromažďujú o nich veľmi obmedzené informácie. Pomocou týchto prostriedkov sa predpokladá, že používatelia môžu slobodne vyjadrovať svoje presvedčenie a názory bez toho, aby odhalili svoju totožnosť. Mnohé z týchto aplikácií však spracúvajú identifikátor zariadenia používateľa - napríklad na zasielanie oznámení používateľom vždy, keď sa iným "anonymným" používateľom páčia ich príspevky, alebo na poskytovanie informácií o blízkych "anonymných" používateľoch tej istej siete. Identifikátory zariadenia by sa mali v zásade považovať za osobné údaje, pretože sú spojené s používateľmi zariadenia. To platí najmä v prípade trvalých identifikátorov. Napriek tomu, aj keď sa v takýchto "anonymných" aplikáciách používa nestály identifikátor zariadenia, stále môže

[17] Pozri napríklad prípad používateľa s pseudonymom 4417749 na https://www.nytimes.com/2006/08/09/technology/09aol.html , Dátum referencie: 03.02.2023

[18] Zdroj: https://blog.isc2.org/isc2_blog/2021/06/best-practices-and-techniques-for-pseudonymization.html , Dátum referencie: 28.02.2023

existovať spojenie medzi týmto identifikátorom a zariadením, čo zase predstavuje riziko pre súkromie používateľa [1.].

Je potrebné zdôrazniť, že aj v prípade absencie osobných identifikátorov nemusia byť údaje nevyhnutne anonymné. Napríklad v predchádzajúcom prípade anonymných sociálnych sietí môže byť používateľ takejto siete identifikovaný napríklad na základe svojich príspevkov a/alebo iných aktivít bez použitia akéhokoľvek identifikátora zariadenia. Podobne, ako sa ukázalo v [2.], jednoduché histórie prehliadania by mohli byť prepojené s niektorými profilmi sociálnych sietí, ako sú účty Twitter alebo Facebook, vzhľadom na skutočnosť, že používatelia častejšie klikajú na odkazy zverejnené účtami, ktoré sledujú. V rovnakom kontexte sa v práci v [3.] ukázalo, že používateľov zariadení so systémom iOS možno vyčleniť prostredníctvom ich personalizovaných konfigurácií zariadenia napriek tomu, že aplikácie tretích strán nemali prístup k žiadnym hardvérovým identifikátorom zariadenia. Okrem toho, ako sa uvádza v [4.], osobné údaje bolo možné odvodiť z verejne dostupných informácií v skorších verziách systému Android. To poukazuje na náročnosť vytvorenia skutočne anonymných údajov a zároveň rozširuje pojem pseudonymizovaných údajov[19] .

Okrem toho vždy existuje riziko, že súbor údajov po pseudonymizácii bude stále obsahovať polia (napríklad adresu ulice) alebo kombináciu polí, ktoré by pri korelácii s inými informáciami mohli umožniť spätnú identifikáciu osôb. Napríklad voľné textové polia so správou a riadkom s pozdravom by potenciálne mohli umožniť prepojenie s konkrétnou osobou, aj keď sú údaje pseudonymizované (teda osobné identifikátory boli odstránené). Dôležitú úlohu v tomto smere zohrávajú vlastnosti súboru údajov, pretože by potenciálne mohli uľahčiť odvodzovanie identifikátorov osôb z pseudonymizovaných údajov (napríklad ak sa súbor údajov týka malej/špecializovanej skupiny osôb, určité atribúty môžu okamžite odvodiť identifikátory konkrétnych osôb v rámci tejto skupiny, aj keď boli osobné identifikátory odstránené). Toto riziko je ďalej posilnené skutočnosťou, že aj keď spätná identifikácia nie je v určitom okamihu možná, hromadenie ďalších údajov, ktoré sú spojené s pseudonymom, by mohlo umožniť spätnú identifikáciu v budúcnosti.

Treba však poznamenať, že napriek rozdielu medzi pseudonymizáciou a anonymizáciou sa prvá z nich často spolieha na techniky druhej, aby sa zvýšila jej účinnosť. V niektorých prípadoch môže byť napríklad dobrým postupom zapojiť do procesu pseudonymizácie určité techniky anonymizácie (napr. zovšeobecnenie atribútov), aby sa znížila možnosť tretích strán odvodiť osobné údaje.

Pri diskusiách o konkrétnych technikách v tejto kapitole používame nasledujúcu terminológiu odvodenú z príslušných definícií GDPR:

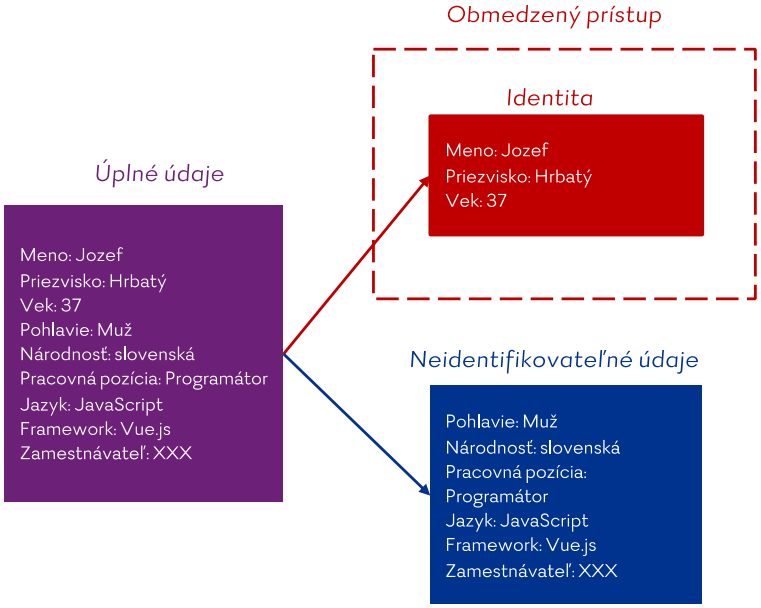
- **Prevádzkovateľ údajov** je subjekt, ktorý určuje účely a prostriedky spracovania osobných údajov (článok 4 ods. 7 GDPR). Prevádzkovateľ údajov je zodpovedný za spracovanie údajov a môže použiť pseudonymizáciu ako technické opatrenie na ochranu osobných údajov.
- **Spracovateľ údajov** je subjekt, ktorý spracúva osobné údaje v mene prevádzkovateľa (článok 4 ods. 8 GDPR). Spracovateľ môže na osobné údaje použiť techniky pseudonymizácie na základe príslušných pokynov prevádzkovateľa.

[19] Ilustratívny návod všeobecných krokov deidentifikácie je uvedený na stránke https://fpf.org/2016/04/25/a-visual-guide-to-practical-data- de-identification, Dátum referencie: 03.02.2023

- **Dotknutá osoba** je fyzická osoba, ktorej osobné údaje sa spracúvajú a môžu podliehať pseudonymizácii. V texte sa na označenie dotknutej osoby používa aj pojem fyzická osoba. Okrem toho sa v rovnakom význame používa aj pojem používateľ, najmä keď sa hovorí o online/mobilných systémoch a službách.
- **Tretia strana** je akýkoľvek subjekt iný ako dotknutá osoba, prevádzkovateľ alebo sprostredkovateľ (článok 4 ods. 10 GDPR).

6.1 Pseudonymizačné techniky

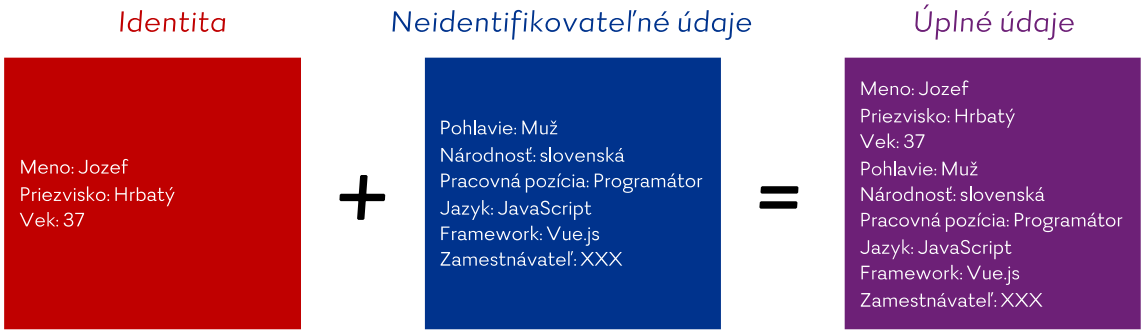
Citlivé údaje možné chrániť riadením prístupov k údajom pomocou mechanizmov riadenia prístupu (dokument 1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov). Okrem toho je možné upraviť údaje tak, aby v prípade, že sa k nim niekto dostane, nemohol spôsobiť veľké škody. Na to slúžia pseudonymizačné techniky, ktoré môžu byť doplnené aj anonymizačnými technikami, popísanými v kapitole 6.1.9. V širšom zmysle pseudonymizácia znamená proces odstránenia spojitosti medzi identitou dotknutej osoby a osobnými údajmi, ktoré sa pre ňu spracúvajú. Zvyčajne sa takýto proces môže vykonať nahradením jedného alebo viacerých osobných identifikátorov, t. j. častí informácií, ktoré môžu umožniť identifikáciu (ako napríklad meno, e-mailová adresa, rodné číslo a podobne), takzvanými pseudonymami, ako sú náhodne vygenerované hodnoty. V praxi to znamená, že pôvodné osobné údaje sa v procese pseudonymizácie rozdelia na údaje o identite (identifikátory) s veľmi obmedzeným prístupom a auditom a na neidentifikovateľné – pseudonymizované údaje, ako ukazuje Obrázok 6.



Obrázok 6: Rozdelenie osobných údajov v procese pseudonymizácie

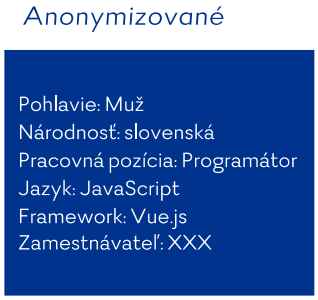
Vďaka využitiu pseudonymizačných techník sa dá zabezpečiť **spätná identifikácia údajov,** keď sa pseudonymizované údaje opäť prepoja s identitiou (Obrázok 7), a to podľa typu použitej pseudonymizačnej techniky (napríklad dešifrovaním zašifrovaného identifikátora, čím sa dajú spätne spojiť neidentifikovateľné údaje s identitou). Spätná identifikácia údajov pomáha zabezpečiť, že údaje sú správne zhromažďované, zdieľané a analyzované. Okrem toho spätná identifikácia údajov môže pomôcť pri ochrane osobných údajov, pretože môže poskytnúť kontrolu nad tým, ako sa údaje využívajú. V rámci spätnej identifikácie je potrebné:

1. **Vytváranie údajových schém,** ktoré zahŕňa vytvorenie jedinečných identifikátorov, ako je číslo sociálneho poistenia alebo dátum narodenia či iný bezvýznamový jedinečný identifikátor. Takýto identifikátor umožňuje ľahké a rýchle určenie originálu údajov. Údajové schémy môžu zahŕňať aj pravidlá na zabezpečenie, že údaje sú správne zhromažďované, zdieľané a analyzované. Údajové schémy musia byť popísané v dátovom katalógu.
2. **Vytváranie auditných protokolov** zahŕňa vytvorenie záznamov, ktoré sledujú všetky prístupy k údajom a úpravy údajov, aby sa zabezpečila ich integrita.



Obrázok 7: Spätné prepojenie osobných údajov, napríklad odmaskovaním identity

Aby sme zabránili osobnej identifikácii, musíme odstrániť (Obrázok 8) alebo nahradiť všetky identifikátory, ktoré niekoho jednoznačne identifikujú. Treba to spraviť buď pred zdieľaním údajov (scenár podľa kapitoly 10.1) alebo priamo zakomponovať do ukladania údajov v informačných systémoch (scenár podľa kapitoly 10.4).



Obrázok 8: Ak nedokážeme k dátam pridať identitu, dáta sú anonymizované

V princípe existujú dva spôsoby zastretia osobných údajov (ako sú rodné čísla, e-mailové adresy atď.):

1. Pred zdieľaním údajov identifikátory sa nahradia internými, jednoznačne vygenerovanými hodnotami.
2. Identifikátory sa nahradia hodnotami generovanými pseudonáhodnou funkciou s kľúčom.

Týmto technikám sa budeme podrobne venovať v nasledujúcich podkapitolách. Teraz sa ešte vrátime k praktickému príkladu z tabuľky (Tabuľka 4 a Tabuľka 5) pre aplikáciu „Výjazdy“. Tak, ako môžeme vymazať identifikátory, ktoré osobne identifikujú používateľov, a nahradiť ich internými identifikátormi, existujú techniky na zamaskovanie údajov o polohe:

- Čas sa zaokrúhli na najbližších 30 minút (napríklad 12: 25 by sa zaokrúhlil na 12:30).

- Prevod súradníc GPS na začiatok/stred/koniec úseku ulice.
- Skrátenie súradnice GPS na tri desatinné miesta - to je veľmi dôležité, pretože menej desatinných miest v polohe GPS znamená, že je poloha menej presná.

Predpokladajme, že máme tabuľku s dvoma záznamami:

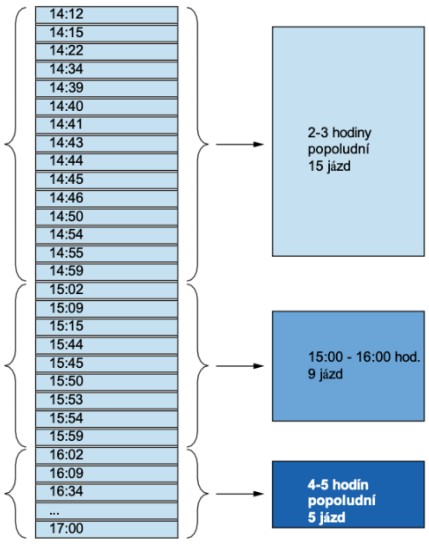
- Výjazd A: Začal sa o 12:22 a skončil o 13:09.
- Výjazd B: Začal sa o 12:24 a skončil o 13:11.

Kľúčovým príkladom narušenia súkromia je možnosť jednoznačnej identifikácie jednotlivca a/alebo činnosti. Ak by sme potrebovali zdieľať tieto údaje na účely analýzy údajov, predstavuje to riziko ohrozenia súkromia, pretože na základe času začiatku a konca a iných verejných údajov sa dá identifikovať, kto sa zúčastnil na každej ceste. **Cieľom zastretia údajov je, aby sa akékoľvek dve jedinečné činnosti viac podobali, takže sa zníži schopnosť dátového vedca alebo automatizovaného procesu identifikovať ich ako jedinečné.**

Ak by sme čas v predchádzajúcom príklade zaokrúhlili na najbližšiu polhodinu, záznamy by vyzerali takto:

- Výjazd A: Začal sa o 12:30 a skončil o 13:00.
- Výjazd B: Začal sa o 12:30 a skončil o 13:00.

Vďaka týmto zastretiam sú cesty menej unikátne, a teda osoby, ktoré ich absolvovali, menej identifikovateľné, bez toho, aby to poškodilo analýzu súhrnných údajov. Obrázok 9 to ukazuje názornejšie.



Obrázok 9: Agregované údaje pre vyššiu ochranu súkromia [10.]

Ľavý stĺpec predstavuje jednotlivé riadky údajov, pričom každý riadok predstavuje konkrétnu jazdu na základe času začiatku. V tomto jednoduchom príklade sa všetky jazdy začali v samostatných časoch, takže žiadne dve jazdy nemajú rovnaký čas začiatku. Predpokladajme, že chceme vykonať analýzu jázd z hľadiska výjazdov zamestnancov, nákladov a podobne. V tomto prípade nás nemusia zaujímať jednotlivé jazdy, ale skôr skupina jázd rozdelená podľa niektorých charakteristických znakov. Mohli by sme vytvoriť kohorty na základe hodiny (14.00,

15.00 a 16.00), ak nie je potrebné presne vedieť, aký je čas začiatku každej jazdy. Tento údaj nie je potrebný, pokiaľ vieme, do akej kohorty jazda patrila. Okrem toho, ak máme konkrétne časy začiatku, existuje riziko, že osoby, ktoré sa týchto jázd zúčastňujú, by mohli byť identifikované, najmä pomocou externých údajov.

Preto, keď ide o prípad použitia analýzy, namiesto použitia ľavého stĺpca s jednotlivými cestami identifikovanými podľa času ich začiatku, by sme mohli použiť pravý stĺpec s cestami zoskupenými na základe hodiny ich začiatku. Ľavý stĺpec predstavuje prevádzkovú verziu údajov, zatiaľ čo pravý stĺpec predstavuje archívny záznam údajov. Ľavý stĺpec možno použiť, keď potrebujeme podrobnejšie informácie (napríklad môžeme potrebovať vedieť konkrétny čas začiatku cesty v prípade, že ju potrebujeme vyhľadať z dôvodu nejakej komplikácie počas výjazdu). Tento príklad je možno trochu zjednodušený, ale kľúčovým posolstvom je vyhnúť sa slepému zdieľaniu údajov, ktoré individuálne identifikujú ľudí, o ktorých tieto údaje sú.

Pri výbere konkrétnej techniky s ohľadom na GDPR si môžu prevádzkovatelia údajov stanoviť nasledujúce ciele návrhu (vlastnosti techniky) na prijatie optimálnej techniky, pričom zohľadnia riziká konkrétnej operácie spracúvania údajov pre práva a slobody fyzických osôb:

- **V1)** pseudonymy by nemali umožňovať "ľahkú" spätnú identifikáciu akoukoľvek treťou stranou (t. j. inou ako prevádzkovateľom alebo spracovateľom) v rámci konkrétneho kontextu spracovania údajov (aby sa "skryli" pôvodné identifikátory v konkrétnom kontexte).
- **V2)** pre žiadnu tretiu stranu (t. j. inú ako prevádzkovateľ alebo spracovateľ) by nemalo byť triviálne reprodukovať pseudonymy (aby sa zabránilo používaniu rovnakých pseudonymov v rôznych oblastiach spracovania údajov - neprepojiteľnosť medzi oblasťami).
- **V3)** Treba tiež zdôrazniť, že, ako už bolo uvedené, prístup založený na pseudonymizácii môže priniesť aj ďalšie výhody v oblasti ochrany údajov, pokiaľ ide o presnosť údajov. Tým sa pridáva tretí cieľ návrhu, ktorý by mali zvážiť aj prevádzkovatelia údajov. Existujú napríklad techniky pseudonymizácie, ktoré vytvárajú pseudonymy matematicky viazané na pôvodné identifikátory, a preto tieto pseudonymy môžu postačovať na to, aby umožnili overenie totožnosti dotknutých osôb v rámci špecifických rámcov.

Uvedené ciele vychádzajú z predpokladu, že prevádzkovateľ údajov bude schopný po procese pseudonymizácie spätne identifikovať dotknuté osoby (keďže má prístup k dodatočným informáciám). Túto možnosť môže mať aj spracovateľ údajov na základe pokynov prevádzkovateľa. To však zjavne neplatí pre tretie strany, pred ktorými sú údaje skutočne chránené.

Existujú aj prípady, v ktorých nie je potrebné, aby prevádzkovateľ spájal pseudonymizované údaje s konkrétnymi počiatočnými identifikátormi. Prevádzkovateľ môže napríklad potrebovať iba sledovať jednotlivcov, t. j. byť schopný rozlíšiť každého jednotlivca od ostatných v rámci konkrétneho kontextu spracovania bez toho, aby pracoval so znalosťou skutočnej totožnosti jednotlivca alebo, všeobecnejšie, jeho pôvodných identifikačných údajov. Aj v tomto prípade môže byť prostriedkom na splnenie takejto požiadavky pseudonymizácia, a to prostredníctvom vhodného použitia techniky na zabezpečenie toho, aby bol tej istej osobe vždy pridelený rovnaký pseudonym. Ako bude uvedené ďalej, výber vhodnej techniky pseudonymizácie je silne závislý od toho, či v kontexte spracúvania údajov existuje možnosť, aby prevádzkovateľ upustil od uchovávania pôvodných identifikátorov a sledoval dotknuté osoby len na základe pseudonymov.

Okrem vyššie uvedených cieľov, ďalším dôležitým aspektom, ktorý by mali prevádzkovatelia zvážiť, je oddelenie údajov, t. j. oddelenie pseudonymizovaných údajov od dodatočných informácií (Obrázok 6 - dva rôzne výstupy pseudonymizácie). Keďže pojem pseudonymizácie skutočne predpokladá spojenie medzi pseudonymami a pôvodnými identifikátormi (dodatočnými informáciami), pravdepodobne by bolo potrebné zaviesť mapovaciu tabuľku alebo inú príslušnú štruktúru, ktorá by umožňovala toto spojenie (napr. kľúč, ako sa uvádza v kapitolách ďalej).

Pri popise pseudonymizačných techník v kapitolách 6.1.1 až 6.1.8 budeme používať nasledujúcu ukážku osobných údajov z informačného systému.

Tabuľka 8: Ukážka osobných údajov v informačnom systéme pre aplikovanie pseudonymizačných techník

Číslo zázna -mu	Meno a priezvisko používateľov aplikácie	E-mail	Rodné číslo	Výška vymeriavacieho základu v EUR
1	Peter Malík	peter.malik@zoznam.sk	5606240629	890,56
2	Jana Dúbravská	jdubrav@yahoo.com	7660240181	8593,12
3	Petra Kováčová	kovacova@gmail.com	9260129879	1235,63

6.1.1 Generátor náhodných čísel („Random number generator (RNG)“)

Generátor náhodných čísel je mechanizmus vytvárajúci hodnoty, ktoré majú rovnakú pravdepodobnosť výberu z celkovej populácie možností (môže ísť o čísla, ale aj reťazce znakov a symbolov), a preto sú nepredvídateľné. Takto vygenerované náhodné čísla sa následne používajú ako pseudonymy alebo na zastretie skutočných citlivých údajov.

Treba poznamenať, že bez náležitej opatrnosti môže dôjsť ku kolíziám. Kolízia je prípad, keď sú dva identifikátory priradené k rovnakému pseudonymu. Pravdepodobnosť, že sa objaví kolízia, súvisí so známym narodeninovým paradoxom[20]. Je intuitívne si myslieť, že výsledkom priraďovania náhodných čísiel generátora vznikne rovnomerné rozloženie pseudonymov alebo údajov v jednotlivých chlievikoch. Rovnako je prirodzené usúdiť, že 23 neznámych ľudí má pravdepodobne narodeniny rozptýlené v celom kalendárnom roku. Matematicky sa však v rámci spomínaného narodeninového paradoxu dá vypočítať, že existuje približne 50 % pravdepodobnosť, že dvaja z 23 neznámych ľudí majú spoločné narodeniny. Navyše pravdepodobnosť vyskočí takmer na 100 %, keď skupinu tvorí len 80 osôb. A preto rovnako netreba ani veľmi veľa údajov priradiť pseudonymom, kým nenastane kolízia. Tieto kolízie sa dajú riešiť buď lineárnym sondovaním alebo reťazením.

Základom tohto prístupu je mapovacia tabuľka, ktorá prelinkováva buď každý osobný dátový prvok s jeho pseudonymom, alebo vytvára jeden pseudonym pre skupinu osobných dátových

[20] Zdroj: https://towardsdatascience.com/when-birthdays-collide-6e8a17b422e7 , Dátum referencie: 10.02.2023

prvkov, čím sa ale znižuje následne granularita, s akou môžu byť originálne údaje v pseudonymizovanom datasete obnovené.

Tabuľka 9: Mapovacia tabuľka pre pseudonymizáciu nahradením osobného dátového prvku náhodným identifikátorom

Číslo záznamu	Pseudonymizovaný identifikátor osobného dátového prvku[21]	Osobný dátový prvok - Rodné číslo
1	13113	5606240629
2	10785	7660240181
3	17156	9260129879

Tabuľka 10: Mapovacia tabuľka pre pseudonymizáciu nahradením skupiny osobných dátových prvkov náhodným identifikátorom

Číslo záznamu	Pseudonymizovaný identifikátor - pseudonym[22]
1	13113
2	10785
3	17156

Tabuľka 11: Pseudonymizované údaje s využitím mapovacej tabuľky Tabuľka 9 alebo Tabuľka 10, pripravené na zdieľanie

Pseudonymizovaný identifikátor	Analytický údaj - výška vymeriavacieho základu v EUR
13113	890,56
10785	8593,12
17156	1235,63

Na základe mapovacej tabuľky (Tabuľka 9) je možné vymeniť pseudonymy za pôvodné údaje, dohľadaním ďalších údajov cez rodné číslo, a tak zrušiť pseudonymizáciu, preto prístup k tejto tabuľke musí byť striktne riadený.

RNG poskytuje silnú ochranu údajov, keďže sa na vytvorenie každého pseudonymu používa náhodné číslo, a preto je ťažké získať informácie o pôvodnom identifikátore, pokiaľ nie je

[21] Vygenerované cez https://www.random.org/sequences/, Dátum referencie: 06.04.2023

[22] Vygenerované cez https://www.random.org/sequences/, Dátum referencie: 06.04.2023

kompromitovaná mapovacia tabuľka. Problémom môžu byť kolízie, ako už bolo spomenuté, ako aj škálovateľnosť (musí sa uložiť kompletná mapovacia tabuľka pseudonymov).

6.1.2 Kryptograficky bezpečný generátor pseudonáhodných čísel („Cryptography secure pseudo-random number generators (CSPRNG)“)

Ide o techniku pseudonymizácie, ktorá tiež využíva mapovaciu tabuľku, ako bolo popísané v kapitole 6.1.1. Rozdielom je len vylepšený kryptograficky bezpečný generátor.

Kryptograficky bezpečné generátory pseudonáhodných čísel sú generátory náhodných čísel, ktoré zaručujú, že náhodné čísla z nich pochádzajúce sú absolútne nepredvídateľné. CSPRNG vyhovujú testu nasledujúceho bitu („next-bit test“) a odolávajú rozšíreniam kompromitujúcim stav generátora („state compromise extensions“). Zvyčajne sú súčasťou operačného systému alebo pochádzajú z bezpečného externého zdroja. V závislosti od požadovanej úrovne bezpečnosti môžu byť CSPRNG implementované ako softvérové komponenty alebo ako hardvérové zariadenia, prípadne ako kombinácia oboch.

Napríklad v centrách na výrobu kreditných kariet formálne bezpečnostné predpisy vyžadujú, aby sa na generovanie PIN kódov kreditných kariet, súkromných kľúčov a iných údajov, ktoré majú zostať súkromné, používali certifikované hardvérové generátory náhodných čísel. Rovnaký prípad použitia platí aj pre štátnu správu, napríklad v prípade generovania bezpečnostného osobného kódu, PIN kódu kvalifikovaného elektronického podpisu pre elektronický občiansky preukaz alebo akéhokoľvek iného krypto materiálu používateľov.

Spravidla by mal CSPRNG začínať od nepredvídateľného náhodného „seed-u“ z operačného systému, zo špecializovaného hardvéru alebo z externého zdroja. Náhodné čísla po inicializácii „seed-u“ sa zvyčajne produkujú pseudonáhodným výpočtom, čo však neohrozuje bezpečnosť. Väčšina algoritmov často "preosieva" náhodný generátor CSPRNG, keď príde nová entropia, aby bola ich práca ešte nepredvídateľnejšia.

Moderné operačné systémy zvyšujú entropiu, používanú na nastavenie počiatočného „seed-u“ generátora, pomocou okolitého šumu: kliknutia na klávesnici, pohybu myšou, sieťovej aktivity, prerušenia systémových vstupov a výstupov atď. Medzi zdroje náhodnosti z prostredia v operačnom systéme Linux patrí napríklad časovanie medzi stlačením klávesov, časovanie medzi prerušeniami niektorých rušení a iné udalosti, ktoré sú nedeterministické a pre vonkajšieho pozorovateľa ťažko merateľné.

Obvykle moderné API pre CSPRNG operačných systémov kombinujú neustále zbieranú entropiu z prostredia s vnútorným stavom svojho zabudovaného pseudonáhodného algoritmu s priebežným „re-seedingom“, aby zaručili maximálnu nepredvídateľnosť generovanej náhodnosti pri vysokej rýchlosti a zároveň neblokujúcom správaní.

CSPRNG sú aj štandardizované, napríklad v:

- NIST SP 800-90A Rev.1: Ide o pôvodný štandard NIST SP 800-90A, z ktorého sa odstránil algoritmus Dual_EC_DRBG, ktorý sa ukázal ako nie kryptograficky bezpečný.
- ANSI X9.17-1985 Appendix C,
- ANSI X9.31-1998 Appendix A.2.4,
- ANSI X9.62-1998 Annex A.4, zastaraný a nahradený ANSI X9.62-2005, Annex D (HMAC_DRBG).

Z vyššie uvedeného vyplýva, že pseudonymizácia v praxi prebieha rovnako, ako je popísané v tabuľkách v kapitole 6.1.1, len sa na generovanie pseudonymu použije CSPRNG generátor, ideálne štandardizovaný.

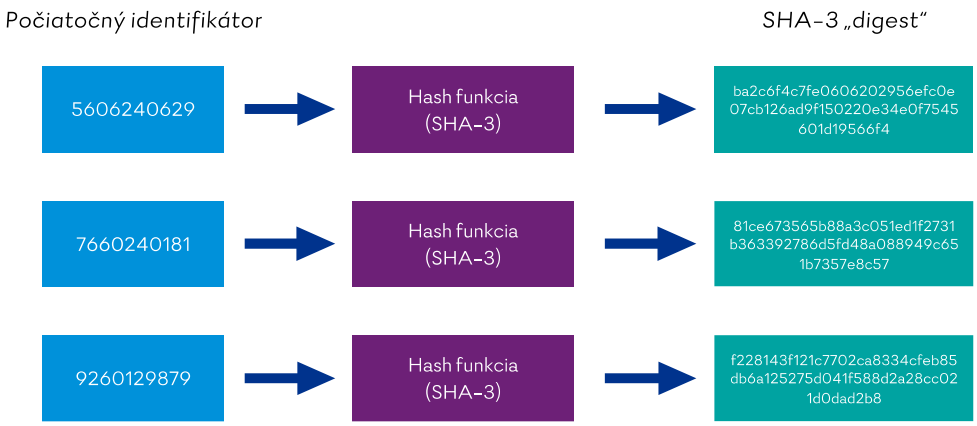
6.1.3 Hašovanie bez kľúča

Hašovanie („hashing“) je technika, ktorú možno použiť na odvodenie pseudonymov, ale ako sa ukáže neskôr v tejto časti, má niektoré závažné nedostatky vzhľadom na ciele návrhu pseudonymizácie stanovené v kapitole 6.1. Napriek tomu je východiskovým bodom pre pochopenie ďalších silnejších techník v tejto oblasti, a preto ju uvádzame ako prvú. Okrem toho hašovanie môže byť užitočným nástrojom na podporu presnosti údajov.

Kryptografická hašovacia funkcia *h* je funkcia so špecifickými vlastnosťami (ako je opísané ďalej), ktorá transformuje ľubovoľnú vstupnú správu *m* ľubovoľnej dĺžky na výstupný reťazec *h(m)* pevnej veľkosti (napr. veľkosti 256 bitov, t. j. 32 znakov), ktorý sa nazýva haš.

Prevod vstupnej správy spĺňa tieto vlastnosti [5.]:

1. Pri danom výstupe *h(m)* je výpočet neznámej *m* výpočtovo neuskutočniteľný, a to platí pre akýkoľvek výstup *h(m)* - t. j. funkcia *h* nie je inverzná.
2. Pre ľubovoľné dané *m* je výpočtovo neuskutočniteľné nájsť iné *m' ≠ m* také, že *h(m')=h(m)*.
3. Je výpočtovo neuskutočniteľné nájsť ľubovoľné dva rôzne vstupy *m, m'* také, že *h(m')=h(m)* (kolízna odolnosť).



Obrázok 10: Fungovanie kryptografickej hašovacej funkcie[23]

Inými slovami, kryptografický hašovací algoritmus je algoritmus, ktorý generuje jedinečný hašovací reťazec pevnej veľkosti (haš) pre akýkoľvek blok údajov ľubovoľnej veľkosti (napr. počiatočný identifikátor akéhokoľvek druhu). Pre akúkoľvek danú hašovaciu funkciu sa pre rovnaký vstup (rovnaký blok údajov) vždy vytvorí ten istý jedinečný haš.

[23] Zdroj: https://www.browserling.com/tools/sha3-hash pre výstup dĺžky 256 bitov, Dátum referencie: 28.02.2023

Podľa obrázka vyššie (Obrázok 10) nahradíme mapovaciu tabuľku pri RNG v kapitole 6.1.1 (Tabuľka 10) nasledujúcou tabuľkou (Tabuľka 12), v ktorej vidíme, že sa nám zásadne zvýšili kapacitné nároky na uloženie daného identifikátora, ktorý musíme prenášať aj pre pseudonymizované údaje (Tabuľka 13).

Tabuľka 12: Mapovacia tabuľka pre pseudonymizáciu hašovaním osobného dátového prvku bez kľúča

Číslo záznamu	Pseudonymizovaný identifikátor - pseudonym
1	ba2c6f4c7fe0606202956efc0e07cb126ad9f150220e34e0f7545601d19566f4
2	81ce673565b88a3c051ed1f2731b363392786d5fd48a088949c651b7357e8c57
3	f228143f121c7702ca8334cfeb85db6a125275d041f588d2a28cc021d0dad2b8

Tabuľka 13: Pseudonymizované údaje s využitím hašovacej tabuľky (Tabuľka 12), pripravené na zdieľanie

Pseudonymizovaný identifikátor	Analytický údaj - výška vymeriavacieho základu v EUR
ba2c6f4c7fe0606202956efc0e07cb126ad9f150220e34e0f7545601d19566f4	890,56
81ce673565b88a3c051ed1f2731b363392786d5fd48a088949c651b7357e8c57	8593,12
f228143f121c7702ca8334cfeb85db6a125275d041f588d2a28cc021d0dad2b8	1235,63

Je dôležité zdôrazniť, že by sa mali vyberať najmodernejšie hašovacie funkcie, v ktorých nie sú odhalené žiadne zraniteľnosti. SHA-2 , SHA-3 a BLAKE-2 sa v súčasnosti považujú aj za bezpečné z pohľadu vplyvu kvantových počítačov[24].

Pokiaľ však ide o pseudonymizáciu, napriek vyššie uvedeným vlastnostiam kryptografickej hašovacej funkcie má jednoduché hašovanie identifikátorov dotknutých osôb s cieľom poskytnúť pseudonymy veľké nevýhody. Presnejšie, pokiaľ ide o vyššie uvedené ciele návrhu V1 a V2, platí nasledovné:

- Vlastnosť V2 neplatí, pretože akákoľvek tretia strana, ktorá použije rovnakú hašovaciu funkciu na rovnaký identifikátor, dostane rovnaký pseudonym.
- V súvislosti s vyššie uvedeným zistením nemusí nevyhnutne platiť ani vlastnosť V1, pretože pre akúkoľvek tretiu stranu je triviálne overiť pre daný identifikátor, či pseudonym zodpovedá tomuto identifikátoru (t. j. prostredníctvom hašovania identifikátora). Ide o rovnaký postup, ktorý sa používa na prelomenie hesla, pretože heslá sa spravidla ukladajú v hašovanej podobe.

Preto pri tejto technike je možné zvrátiť pseudonymizáciu, keďže existencia zoznamu (možných) počiatočných identifikátorov je dostatočná na to, aby si akákoľvek tretia strana mohla tieto identifikátory priradiť k zodpovedajúcim pseudonymom, pričom neexistuje žiadne iné priradenie. V skutočnosti by sa v nadväznosti na definíciu pseudonymizácie podľa nariadenia GDPR dalo tvrdiť, že hašovanie je slabá technika pseudonymizácie, keďže ju možno zvrátiť bez použitia ďalších informácií. Relevantné príklady sú uvedené v [6.] (a v odkazoch v ňom), kde sa výskumníci odvolávajú na službu Gravatar35[25] a opisujú, ako možno odvodiť e-mailové adresy používateľov prostredníctvom ich hašu, ktorý sa zobrazuje v adrese URL a ktorý zodpovedá gravataru používateľa, bez využitia akýchkoľvek ďalších informácií. Gravatar je skratka pre globálne rozpoznávaný avatar. Možno ho vytvoriť cez webovú službu, pomocou ktorej si ľudia spravia profil a priradia obrázky avatara k ich e-mailovým adresám.

Preto sa hašovacie funkcie vo všeobecnosti neodporúčajú na pseudonymizáciu osobných údajov, hoci môžu prispieť k zvýšeniu bezpečnosti v špecifických kontextoch so zanedbateľnými rizikami pre súkromie a vtedy, keď pôvodné identifikátory nemôže tretia strana uhádnuť alebo ľahko odvodiť. V prevažnej väčšine prípadov sa takáto technika pseudonymizácie nezdá byť dostatočná ako mechanizmus ochrany údajov [6.].

6.1.4 Hašovanie pomocou kľúča alebo soli

Robustný prístup ku generovaniu pseudonymov je založený na použití kľúčovaných hašovacích funkcií, ktorých výstup závisí nielen od vstupu, ale aj od tajného kľúča. V kryptografii sa takéto primitívy nazývajú autentifikačný kód správy („message authentication code (MAC)“ - pozri napríklad v zdroji [5.]).

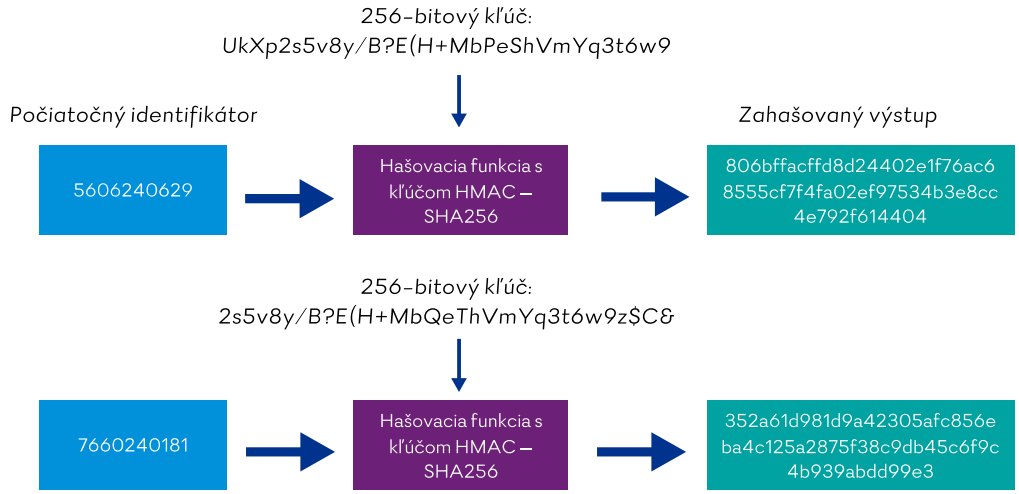
Hlavným rozdielom oproti bežným hašovacím funkciám je, že pre rovnaký vstup (identifikátor subjektu údajov) možno vytvoriť niekoľko rôznych pseudonymov podľa výberu konkrétneho

[24] Zdroj: https://cryptobook.nakov.com/quantum-safe-cryptography, Dátum referencie: 03.02.2023

[25] Pozri: https://en.gravatar.com , Dátum referencie: 06.02.2023

kľúča - a tým je zabezpečená vlastnosť V2. Okrem toho platí aj vlastnosť V1, pokiaľ akákoľvek tretia strana, t. j. iná ako prevádzkovateľ alebo spracovateľ, nepozná kľúč, nemôže overiť, či pseudonym zodpovedá konkrétnemu známemu identifikátoru. Ak prevádzkovateľ údajov potrebuje prideliť ten istý pseudonym tej istej osobe, mal by sa použiť ten istý tajný kľúč.

Na zabezpečenie uvedených vlastností je potrebná bezpečná kľúčovaná hašovacia funkcia s vhodne zvolenými parametrami. Známym takýmto štandardom je HMAC („keyed-hash message authentication code“)[26], ktorého robustnosť je podmienená neprelomiteľnosťou základnej jednoduchej hašovacej funkcie (a teda začlenenie SHA-2 alebo SHA-3 do HMAC je v súčasnosti správnou voľbou). Navyše tajný kľúč musí byť nepredvídateľný a dostatočne dlhý, napr. 256 bitov, čo by sa dalo považovať za primerané aj pre postkvantovú éru. Postkvantová kryptografia je kryptografia za predpokladu, že útočník má k dispozícii veľký kvantový počítač. V súčasnosti je známe, že dĺžka kľúča 256 bitov pre symetrický kryptografický primitív - napríklad hašovaciu funkciu s kľúčom - je bezpečná veľkosť kľúča pre postkvantovú kryptografiu [7.]. Ak sa tajný kľúč prezradí tretej strane, potom sa kľúčovaná hašovacia funkcia vlastne stáva konvenčnou hašovacou funkciou z hľadiska vyhodnocovania jej pseudonymizačných schopností. Preto, ak si pripomenieme definíciu pseudonymizácie v GDPR, prevádzkovateľ údajov by mal tajný kľúč bezpečne uchovávať oddelene od ostatných údajov, pretože predstavuje dodatočnú informáciu, čo znamená, že poskytuje prostriedky na priradenie osôb - pôvodných identifikátorov - k odvodeným pseudonymom.



Obrázok 11: Fungovanie hašovacej funkcie s kľúčom[27]

V praxi tento prístup funguje obdobne, ako ukazujú tabuľky v kapitole 6.1.3 (Tabuľka 12 a Tabuľka 13), len haše sa generujú iným algoritmom podľa obrázka vyššie (Obrázok 11: Fungovanie hašovacej funkcie s kľúčom). Ak sa vyžaduje len sledovanie dotknutých osôb, prevádzkovateľ musí mať prístup ku kľúču, ale nemusí mať prístup k pôvodným identifikátorom po vykonaní pseudonymizácie. Ide o dôležitý aspekt, ktorý dodržiava zásadu minimalizácie údajov a prevádzkovateľ by ho mal zvážiť ako aspekt ochrany údajov už od návrhu.

[26] Zdroj: https://csrc.nist.gov/publications/detail/fips/198/1/final , Dátum referencie: 07.02.2023

[27] Zdroj: https://www.devglan.com/online-tools/hmac-sha256-online a https://www.allkeysgenerator.com/Random/Security-Encryption-Key-Generator.aspx , Dátum referencie: 28.02.2023

Okrem toho má kľúčovaná hašovacia funkcia aj túto vlastnosť: ak je tajný kľúč bezpečne zničený a hašovacia funkcia je kryptograficky silná, je aj pre prevádzkovateľa údajov výpočtovo ťažké vrátiť pseudonym na pôvodný identifikátor, a to aj v prípade, že prevádzkovateľ pozná pôvodné identifikátory. Použitie hašovacej funkcie s kľúčom preto môže v prípade potreby umožniť následnú anonymizáciu údajov, keďže vymazaním tajného kľúča sa vlastne vymaže akékoľvek spojenie medzi pseudonymami a pôvodnými identifikátormi. Všeobecnejšie povedané, **použitie kľúčovanej hašovacej funkcie na generovanie pseudonymov a následné vymazanie tajného kľúča je istým spôsobom ekvivalentné generovaniu náhodných pseudonymov bez akéhokoľvek spojenia s pôvodnými identifikátormi.**

Ďalším prístupom, ktorý sa často prezentuje ako alternatíva ku kľúčovanej hašovacej funkcii, je použitie nekľúčovanej (konvenčnej) hašovacej funkcie s takzvanou „soľou“ - to znamená, že vstup do hašovacej funkcie sa rozšíri o pomocný náhodne vyzerajúci reťazec, ktorý sa nazýva „soľ“. Ak sa takáto technika vhodne použije, pre ten istý identifikátor sa môže vytvoriť niekoľko rôznych pseudonymov podľa výberu „soli“ - a tým sa zabezpečí vlastnosť V2, pričom vlastnosť V1 platí aj pre tretie strany za predpokladu, že nepoznajú „soľ“. Samozrejme, tento záver platí len dovtedy, kým je „soľ“ vhodne zabezpečená a oddelená od hašu. Rovnako ako v prípade kľúčovaného hašu by mal prevádzkovateľ používať rovnakú „soľ“ v prípadoch, kedy je potrebné priradiť vždy rovnaký pseudonym tej istej osobe . Okrem toho sa hašovacie funkcie so soľou môžu využívať v prípadoch, keď prevádzkovateľ potrebuje uložiť pôvodné identifikátory, pričom je stále schopný sledovať dotknuté osoby. Ak prevádzkovateľ bezpečne zničí soľ, nie je triviálne obnoviť spojenie medzi pseudonymami a identifikátormi.

Soľ však nemá rovnaké vlastnosti z hľadiska predpovedateľnosti ako tajné kľúče (napríklad soľ môže pozostávať len z 8 znakov, teda. 64 bitov). Z kryptografického hľadiska sa kľúčovaná hašovacia funkcia považuje za účinnejší prístup ako solená hašovacia funkcia. Existuje však niekoľko kryptograficky silných techník na generovanie solených hašov, ktoré by sa zasa mohli považovať za vhodných kandidátov na generovanie pseudonymov - významným príkladom je bcrypt[28] a argon2[29].

Problém je, že „soli“ sa vo väčšine bežných scenárov spravidla ukladajú spolu s príslušnými hodnotami hašu, čím sa ochrana výrazne oslabuje. Alternatívne použitie takzvaných korení - „peperov“, ktoré sú skrytými chránenými „soľami“ a sú uložené oddelene od hašov, môže poskytnúť lepšiu alternatívu. „Pepper“ je fixný reťazec dát pridaný k dátovému prvku alebo celému záznamu či datasetu pred hašovaním. Možno ho použiť aj spolu so soľou. Z tohto dôvodu sa odporúča, aby sa „solené“ haše používali na pseudonymizáciu opatrne a v súlade s dostupnými osvedčenými postupmi v tejto oblasti.

6.1.5 Šifrovanie

Treba poznamenať, že medzi prevádzkovateľmi údajov často dochádza k nejasnostiam v súvislosti s pojmami šifrovanie a pseudonymizácia, ktoré sa v GDPR uvádzajú ako bezpečnostné opatrenia (článok 32). Napriek niektorým spoločným prvkom sú však hlavné ciele týchto techník v skutočnosti odlišné.

[28] Pozri: https://bcrypt.online/ , Dátum referencie: 07.02.2023

[29] Pozri: https://argon2.online/, Dátum referencie: 07.02.2023

Pokiaľ ide o pseudonymizáciu, z predchádzajúcej diskusie je zrejmé, že sa zameriava najmä na ochranu identity jednotlivcov (pre každého bez prístupu k ďalším informáciám). Pseudonymizované údaje však poskytujú určité čitateľné informácie, a preto tretia strana (t. j. iná ako prevádzkovateľ alebo spracovateľ) môže stále pochopiť sémantiku (štruktúru) údajov, a to napriek skutočnosti, že tieto údaje nemožno priradiť k jednotlivcovi. **Pseudonymizované údaje preto možno použiť na na analytické účely**.

Na druhej strane, cieľom šifrovania je zabezpečiť prostredníctvom vhodne využitých matematických techník to, aby celý súbor údajov, ktorý sa šifruje, bol nezrozumiteľný pre kohokoľvek okrem osobitne oprávnených používateľov, ktorí môžu túto „nezrozumiteľnosť" zvrátiť (t. j. dešifrovať). **Šifrovanie je hlavným nástrojom na dosiahnutie dôvernosti osobných údajov tým, že sa celý súbor údajov skryje a stane sa nezrozumiteľným pre akúkoľvek neoprávnenú stranu** (pokiaľ sa používajú najmodernejšie algoritmy a dĺžky kľúčov a šifrovací kľúč je vhodne chránený). Šifrovanie sa môže použiť na celý súbor údajov alebo na určité časti súboru údajov (napríklad na určité polia v databáze) v závislosti od konkrétnych cieľov ochrany. Existujú tiež špecifické kryptografické techniky, ktoré umožňujú tretej strane bez znalosti šifrovacieho kľúča vykonávať operácie so zašifrovanými hodnotami (ide o takzvané **homomorfné šifry,** ktoré sú ale stále výpočtovo veľmi náročné a v bežných systémoch nepoužiteľné) alebo získať určitý druh informácií (napr. šifrovanie zachovávajúce poradie zachováva číselné poradie počiatočných textov). Zašifrované údaje sú však stále nezrozumiteľné v tom zmysle, že nikto nemôže odhaliť pôvodné údaje.

Okrem toho, ako už bolo uvedené (Obrázok 6), pseudonymizácia spočíva v rozdelení údajov, čo znamená, že z jedného vstupu (počiatočný dataset) vzniknú dva výstupy (pseudonymizované údaje, dodatočné informácie). Zvrátenie pseudonymizácie je možné pre kohokoľvek, kto môže získať dodatočné informácie alebo kto môže prepojiť pseudonymizované údaje s pôvodnými údajmi pomocou akýchkoľvek iných informácií. Naopak, šifrovanie z jedného vstupu (počiatočné údaje), generuje opäť jediný výstup (zašifrované údaje) a jeho zvrátenie spočíva najmä v prípadnom neoprávnenom prístupe k dešifrovaciemu kľúču (pokiaľ sa používajú najmodernejšie algoritmy a dĺžky kľúčov).

Napriek uvedeným rozdielom je však dôležité uviesť, že **šifrovanie možno použiť aj ako techniku pseudonymizácie** (pričom opak nie je možný, teda pseudonymizácia neslúži na šifrovanie). Kryptografické primitívy sa vo všeobecnosti môžu použiť v technikách pseudonymizácie na generovanie pseudonymov s požadovanými vlastnosťami.

Kľúčové vlastnosti šifrovania možno zhrnúť nasledovne a porovnať s vlastnosťami tokenizácie v kapitole 6.1.6:

- **Matematicky transformuje jednoduchý text** - šifrovacie procesy používajú matematiku na transformáciu jednoduchého textu na šifrovaný text. To sa dosahuje pomocou šifrovacieho algoritmu a kľúča.
- **Škálovanie na veľké objemy údajov** - krátky šifrovací kľúč umožňuje zašifrovať a dešifrovať aj veľké objemy údajov naraz – závisí od prípadu použitia a využitého nástroja, pričom existuje veľa nástrojov optimalizovaných na nasadenie vo veľkej škále.
- **Štruktúrované polia aj neštruktúrované údaje** - šifrovanie môžete použiť na štruktúrované aj neštruktúrované údaje vrátane celých súborov.
- **Pôvodné údaje opúšťajú organizáciu** - šifrovanie umožňuje organizáciám prenášať údaje mimo organizácie v zašifrovanej podobe.
- **Šifrovanie je ideálne na výmenu citlivých informácií s tými, ktorí majú šifrovací kľúč**. Šifrovacie schémy zachovávajúce formát majú však nižšiu silu zabezpečenia.

Pri využívaní šifrovania vo všeobecnosti treba myslieť na nasledovné:

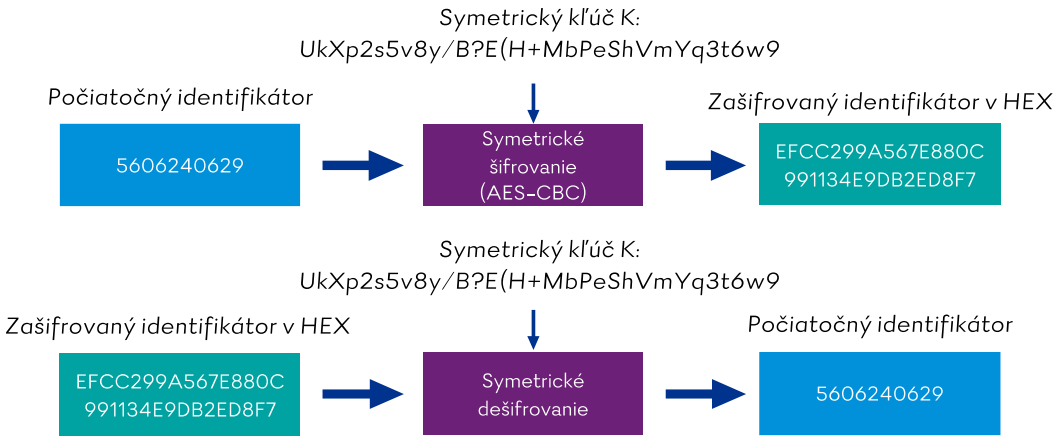
- Zabezpečte, aby šifrovanie zohľadňovalo nielen údaje, ale aj metadáta. Metaúdaje sa často môžu pre útočníka stať zlatou baňou informácií. Metaúdaje môžu poskytovať informácie o údajoch, ich vzťahoch k iným údajom a ich závislostiach od externých údajov a kreatívni útočníci by mohli využiť dátovú „lineage“ alebo odkazy na identifikáciu používateľov. Dokonca môžu byť schopní využiť výpočtový výkon na prepojenie rôznych aspektov metaúdajov s cieľom odvodiť samotné údaje bez toho, aby došlo k ich extrakcii. V tomto prípade žiadne údaje zákazníkov nikdy neopustia systémy, ale dopad by mohol byť rovnaký.
- Nezdieľajte bezstarostne údaje len preto, že sú zašifrované. Šifrovanie nie je univerzálne riešenie na zdieľanie údajov. Kryptografické proxy servery existujú a možno ich použiť na dolovanie údajov. Zdieľajte len minimum informácií potrebných na vykonanie práce. Kryptografia zvyšuje bezpečnosť samotného procesu zdieľania údajov, ale má svoje obmedzenia. Existujú techniky, spomínané v tejto kapitole, ktoré umožňujú bezpečnejšie zdieľanie samotných údajov.
- Aby ste sa uistili, že pokrývate niektoré potenciálne slabé miesta kryptografie, majte na pamäti:
  - Chyby v architektúre, politikách alebo zdrojovom kóde môžu stále odhaliť citlivé informácie.
  - Partneri nemusia byť úplne spoľahliví (kvôli nepoctivým zamestnancom alebo zmluvným partnerom alebo chybám v konfigurácii siete - úmyselným alebo neúmyselným).
  - Zabezpečte súlad so zainteresovanými stranami, pretože šifrovanie a dešifrovanie údajov si vyžaduje čas a výpočtovú kapacitu a takmer určite ovplyvní priepustnosť systémov. Používatelia často majú vzájomne nezlučiteľné očakávania úplnej bezpečnosti a zanedbateľnej latencie pri práci so systémami.
- Správa kľúčov je pri šifrovaní zásadná:
  - Šifrovacie kľúče sa musia pravidelne obmieňať, pretože kryptografické algoritmy sa neustále analyzujú a môžu sa v nich nájsť zraniteľné miesta.
  - Symetrické kľúče a súkromné kľúče musia byť starostlivo chránené.
  - Ak existuje verejný kľúč, musí byť distribuovaný v certifikáte, aby sa zabránilo útokom na zachytenie.
- Šifrovanie je nevyhnutné prispôsobiť stavu údajov:
  - Údaje na úložiskách si v ideálnom prípade vyžadujú obálkové šifrovanie - vytvorí sa hierarchia kľúčov. Nikdy nešifrujte všetky údaje v databáze rovnakým kľúčom.
  - Údaje pri prenose by si mohli vyžadovať použitie protokolu HTTPS/TLS. Veľmi citlivé údaje na úrovni „Vyhradené“ musia byť stále šifrované, a to aj cez šifrované

pripojenie. Môžete zvážiť alternatívne spôsoby ochrany údajov, napríklad protokoly SFTP[30], IPsec[31] atď.

6.1.5.1 Symetrické šifrovanie

V typickom prípade možno pôvodný identifikátor subjektu údajov zašifrovať pomocou symetrického šifrovacieho algoritmu (napríklad AES, ktorý je šifrovacím štandardom, pričom AES – 256 s veľkosťou kľúča 256 bitov by mal byť neprelomiteľný aj v kvantovej kryptografii[32]), čím sa získa šifrovaný text, ktorý sa použije ako pseudonym, na dešifrovanie je potrebný rovnaký tajný kľúč.

Spomínaný pokročilý šifrovací štandard („Advanced Encryption Standard (AES)“) šifruje bloky dát (128 bitov) za jednotku času. Pri použití 256-bitového kľúča šifruje informácie v 14 kolách. Každé kolo pozostáva z niekoľkých krokov substitúcie, miešania jednoduchého textu, transpozície a ďalších krokov. Šifrovacie štandardy AES sú dnes najrozšírenejšími metódami šifrovania údajov pri prenose a aj pri uložení („at rest“).



Obrázok 12: Fungovanie symetrického šifrovania[33]

Takýto pseudonym spĺňa vlastnosť V2, ako aj vlastnosť V1, pokiaľ k šifrovaciemu kľúču nemá prístup žiadna tretia strana – teda nikto iný ako prevádzkovateľ alebo spracovateľ, a za predpokladu, že sa používajú najmodernejšie algoritmy a dostatočná dĺžka kľúča. Hlavný rozdiel šifrovania v porovnaní s hašovacími funkciami s kľúčom - z hľadiska pseudonymizácie - spočíva v tom, že vlastník tajného kľúča (t. j. prevádzkovateľ údajov) môže vždy získať pôvodné identifikátory dotknutých osôb prostredníctvom jednoduchého procesu dešifrovania. Naopak, ako už bolo vysvetlené, hašovacie funkcie s kľúčom poskytujú prevádzkovateľom údajov možnosť sledovania osôb bez toho, aby poznali a ukladali počiatočné identifikátory. To nie je

[30] https://www.ssh.com/academy/ssh/sftp-ssh-file-transfer-protocol, Dátum referencie: 01.02.2023

[31] https://www.techtarget.com/searchsecurity/definition/IPsec-Internet-Protocol-Security, Dátum referencie: 01.02.2023

[32] Zdroj: https://thequantuminsider.com/2020/04/30/is-aes-256-quantum-resistant/, Dátum referencie: 07.02.2023

[33] Zdroj: https://www.devglan.com/online-tools/aes-encryption-decryption, Dátum referencie: 28.02.2023

prípad šifrovania (ako metódy pseudonymizácie), pri ktorom môžu byť počiatočné identifikátory prevádzkovateľovi či spracovávateľovi vždy známe vďaka dešifrovaniu.

Okrem tohto aspektu má symetrické šifrovanie ďalšie podobné vlastnosti - pokiaľ ide o pseudonymizáciu - s hašovacími funkciami s kľúčom, a to:

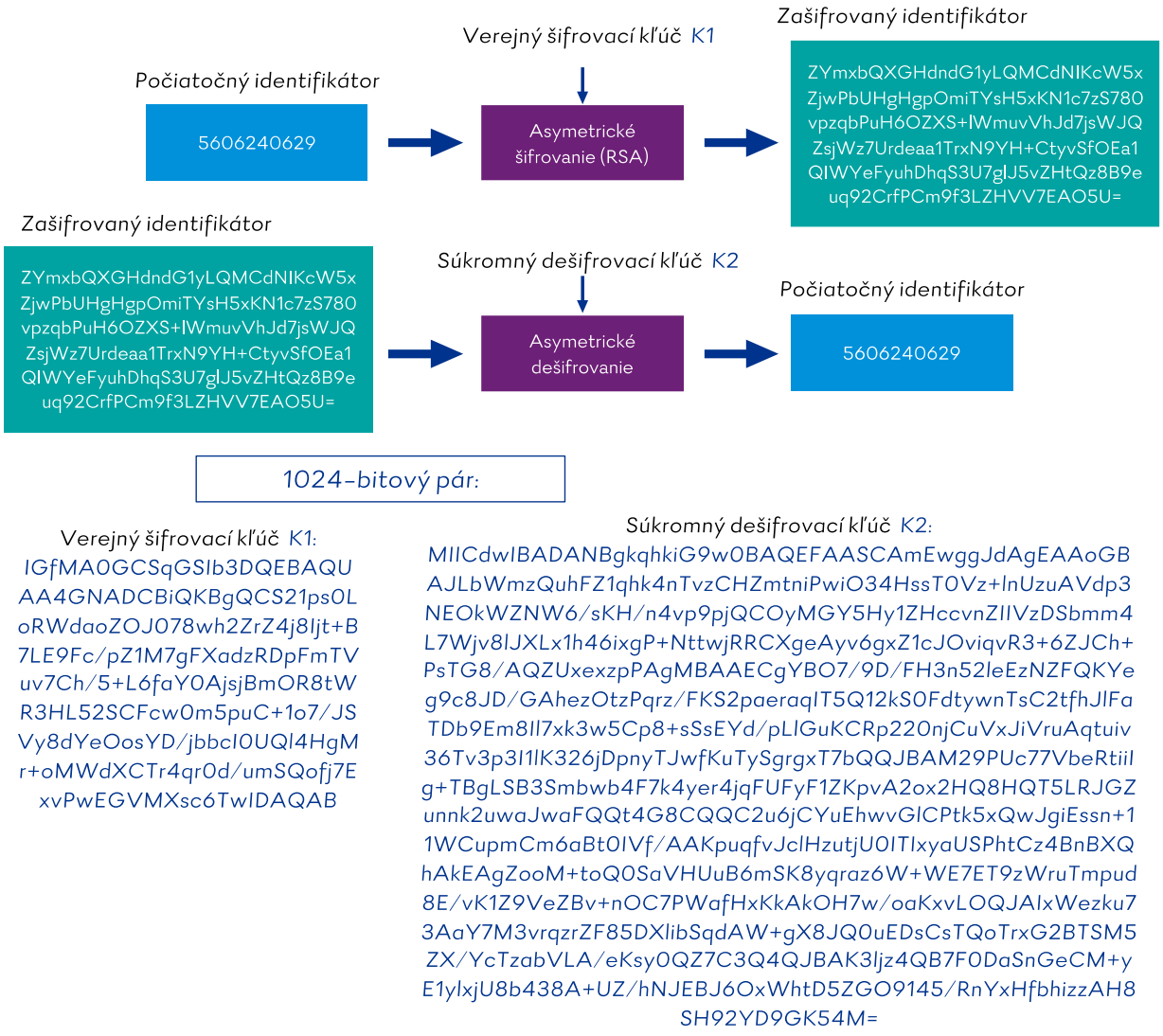
- na poskytnutie rovnakého pseudonymu pre ten istý identifikátor sa musí použiť ten istý tajný symetrický kľúč,
- ak sa symetrický kľúč zničí, nie je triviálne priradiť pseudonym k pôvodnému identifikátoru, aj keď pôvodné identifikátory uchováva prevádzkovateľ údajov.

Preto sa symetrické šifrovanie môže vo všeobecnosti použiť (ako technika pseudonymizácie) v prípadoch, keď prevádzkovateľ údajov potrebuje nielen sledovať dotknuté osoby, ale aj poznať ich pôvodné identifikátory. Sledovateľnosť je založená na deterministickej povahe metódy šifrovania, teda na šifrovaní toho istého identifikátora tým istým kľúčom, čo vždy vedie k získaniu toho istého pseudonymu. Spätná identifikovateľnosť (pôvodných identifikátorov) spočíva, ako bolo vysvetlené vyššie, v samotnej povahe symetrického šifrovania.

V praxi tento prístup funguje obdobne, ako ukazujú tabuľky v kapitole 6.1.3 (Tabuľka 12 a Tabuľka 13), len haše predstavujú zašifrované identifikátory, ktoré sa generujú algoritmom podľa obrázka vyššie (Obrázok 12).

6.1.5.2 Asymetrické šifrovanie

Okrem symetrických šifrovacích algoritmov sa na účely pseudonymizácie môžu v určitých prípadoch použiť aj asymetrické šifrovacie algoritmy (t. j. algoritmy s verejným kľúčom). Hlavnou charakteristikou šifrovania s verejným kľúčom je, že každý subjekt zúčastňujúci sa na tejto technike má dvojicu kľúčov: verejný a súkromný kľúč. Verejný kľúč subjektu môže na šifrovanie údajov použiť ktokoľvek, ale iba konkrétny subjekt môže tieto údaje dešifrovať pomocou svojho súkromného kľúča. Hoci sú tieto dva kľúče nevyhnutne matematicky prepojené, znalosť verejného kľúča neumožňuje určiť súkromný kľúč. Na zabezpečenie vlastnosti nerozlíšiteľnosti šifrovaného textu možno algoritmy verejného kľúča vhodne implementovať v pravdepodobnostnej forme zavedením náhodnosti do procesu šifrovania. To znamená, že pri každom šifrovacom cykle sa používajú náhodne vybrané hodnoty. Týmto spôsobom, ak sa tá istá správa zašifruje dvakrát s tým istým verejným kľúčom, zodpovedajúce dva šifrované texty sa budú líšiť bez toho, aby to ovplyvnilo schopnosť dešifrovania pre držiteľa dešifrovacieho kľúča. To umožňuje generovať rôzne pseudonymy pre tú istú osobu (s rovnakým verejným kľúčom). Tento prístup môže nájsť uplatnenie v prípadoch, keď prevádzkovateľ údajov potrebuje prideliť zakaždým iný pseudonym pre ten istý identifikátor (dotknutú osobu), najmä ak nie je potrebné sledovať dotknuté osoby (ale stále je možné ich spätne identifikovať). V takýchto prípadoch verejný aj súkromný kľúč zostáva u prevádzkovateľa údajov (keďže nie je potrebné, aby bol verejný kľúč prístupný iným stranám).



Obrázok 13: Fungovanie verejného (asymetrického) šifrovania[34]

„Rivest-Shamir-Adleman (RSA)“ je asymetrický šifrovací algoritmus. Je založený na faktorizácii výsledku dvoch veľkých prvočísel. Iba osoba, ktorá pozná tieto čísla, vie, ako dekódovať správu. RSA sa zvyčajne používa pri prenose údajov medzi dvoma rôznymi koncovými bodmi (napríklad webové spojenie). Pri šifrovaní veľkých objemov údajov však funguje pomaly.

Kryptografia eliptickou krivkou („Elliptic Curve Cryptography (ECC)“), ktorú uprednostňujú mnohé verejné organizácie vrátane americkej Národnej agentúry pre bezpečnosť, je rýchla a výkonná forma šifrovania údajov používaná ako súčasť protokolu SSL/TLS. Využíva úplne odlišný matematický proces, ktorý umožňuje používať kratšie dĺžky kľúčov na zvýšenie rýchlosti a zároveň ponúka vynikajúcu bezpečnosť. Napríklad 3 072-bitový kľúč RSA a 256-bitový kľúč ECC ponúkajú rovnakú úroveň zabezpečenia.

[34] Zdroj: https://www.devglan.com/online-tools/rsa-encryption-decryption, Dátum referencie: 28.02.2023

V praxi tento prístup funguje obdobne, ako ukazujú tabuľky v kapitole 6.1.3 (Tabuľka 12 a Tabuľka 13), len haše predstavujú zašifrované identifikátory, ktoré sa generujú algoritmom podľa obrázka vyššie (Obrázok 13).

Treba však zdôrazniť, že algoritmy asymetrických kľúčov si vyžadujú použitie veľmi veľkých kľúčov, čo môže viesť k obmedzeniam pri implementácii - napríklad v prípade RSA je potrebných 3072-bitový kľúč[35]. Aj keď sa berie do úvahy kryptografia s eliptickou krivkou, ktorá ponúka oveľa menšie veľkosti kľúčov ako RSA, ako aj rýchlejšie výpočty, stále je menej efektívna ako algoritmy so symetrickým kľúčom[36]. Okrem toho treba mať na pamäti, že v súčasnosti najznámejšie a najpoužívanejšie algoritmy s verejným kľúčom - vrátane algoritmov RSA a kryptografických algoritmov s eliptickou krivkou - nebudú v postkvantovej ére neprelomiteľné [7.]. Je to z dôvodu existencie rýchleho kvantového algoritmu (Shor 's algoritmus). NIST už inicioval proces hodnotenia a štandardizácie jedného alebo viacerých kvantovo odolných kryptografických algoritmov s verejným kľúčom[37].

6.1.6 Tokenizácia

Tokenizácia sa vzťahuje na proces, pri ktorom sa identifikátory dotknutých osôb nahrádzajú v úložiskách („at rest“) alebo pri prenose náhodne vygenerovanými hodnotami, známymi ako tokeny, bez toho, aby mali akýkoľvek matematický vzťah k pôvodným identifikátorom. Preto znalosť tokenu nemá pre tretiu stranu, teda pre kohokoľvek iného ako prevádzkovateľa alebo spracovateľa, žiadnu užitočnosť[38]. Tokenizácia sa bežne používa na ochranu finančných transakcií, ale neobmedzuje sa len na takéto aplikácie. Využíva princípy generovania náhodných čísiel a obdobne nahrádza osobné dátové prvky, ako bolo popísané v kapitolách 6.1.1 a 6.1.2. Akurát ide o overený systém, ktorý je aj štandardizovaný cez štandard „Payment Card Industry Data Security Standard“[39], ktorý popisuje tokenizáciu nasledovne: „Indexový token je kryptografický token, ktorý nahrádza primárne čísla účtu kreditnej karty na základe daného indexu za nepredvídateľnú hodnotu. Jednorazový blok („pad“) je systém, v ktorom sa náhodne vygenerovaný súkromný kľúč použije len raz na zašifrovanie správy, ktorá sa potom dešifruje pomocou zodpovedajúceho jednorazového bloku a kľúča.“

Je zrejmé, že systém tokenizácie by mal byť vhodne navrhnutý, aby sa zabezpečilo, že medzi pseudonymami a pôvodnými identifikátormi skutočne neexistuje žiadny matematický vzťah. Okrem toho by sa mali zohľadniť aj ďalšie obmedzenia v závislosti od kontextu celkového spracovania - napríklad ak sa tokenizácia používa na pseudonymizáciu čísla kreditnej karty v platobných systémoch, náhodne vygenerované tokeny by nemali mať žiadnu možnosť zhodovať sa so skutočnými číslami kariet (takéto riziko by mohlo existovať v prípadoch tokenizácie zachovávajúcej formát, keď majú tokeny rovnaký formát s pôvodnými údajmi). Vďaka náhodnému skrytému mapovaniu z pôvodných údajov na token je zrejmé, že tokenizácia

[35] Zdroj: https://knowledge.digicert.com/alerts/code-signing-new-minimum-rsa-keysize.html, Dátum referencie: 07.02.2023

[36] Zdroj: https://www.bearssl.org/speed.html, Dátum referencie: 07.02.2023

[37] Zdroj: https://csrc.nist.gov/Projects/post-quantum-cryptography/post-quantum-cryptography-standardization , Dátum referencie: 07.02.2023

[38] Article 29 Working Party, “Opinion 05/2014 on anonymisation techniques”, 2014.

[39] Zdroj: https://www.pcisecuritystandards.org/document_library/?category=pcidss&document=pci_dss, Dátum referencie: 06.04.2023

spĺňa vlastnosti pseudonymizácie V1 aj V2. Keďže existuje subjekt, ktorý toto skryté mapovanie uchováva (tokenový server v systéme tokenizácie), spätná identifikácia údajov subjektov zo strany prevádzkovateľa bude možná vo všetkých prípadoch. To zahŕňa aj sledovanie, pokiaľ existuje len jedno mapovanie pre každý identifikátor.

Treba však poznamenať, že napriek účinnosti tokenizácie môže byť jej nasadenie v závislosti od kontextu pomerne náročné, napríklad v niekoľkých aplikáciách môže byť potrebná synchronizácia tokenov v niekoľkých systémoch. Preto by mohli byť vhodnejšie skôr uvedené prístupy, ktoré využívajú kľúčované hašovacie funkcie (kapitola 6.1.4) alebo šifrovacie algoritmy (kapitola 6.1.5), pokiaľ ide o zníženie zložitosti a zjednodušenie ukladania.

Tu sú zhrnuté kľúčové charakteristiky tokenizácie, ktoré možno porovnať s charakteristikami šifrovania v kapitole 6.1.5:

- **Náhodne generuje hodnotu tokenu** - systémy tokenizácie generujú náhodné hodnoty tokenov, ktoré potom nahrádzajú obyčajný text. Mapovanie je uložené v databáze.
- **Je ťažké zaručiť bezpečné škálovanie** - keď sa zväčšuje veľkosť databáz, je ťažké bezpečne škálovať a udržať výkon.
- **Štruktúrované údaje** - tokenizácia sa vzťahuje na štruktúrované údaje, ako sú rodné čísla alebo informácie o platobných kartách.
- **Pôvodné údaje neopustia organizáciu** - tokenizácia pomáha splniť požiadavky na dodržiavanie predpisov, ktoré vyžadujú zachovanie pôvodných údajov.
- Tokenizácia umožňuje organizáciám zachovať formáty bez zníženia sily zabezpečenia. Výmena údajov však môže byť veľmi náročná, pretože si vyžaduje priamy prístup k trezoru tokenov, ktorý mapuje hodnoty tokenov.

6.1.7 Decentralizované riešenie pre odvodenie pseudonymov

Ďalší dôležitý kryptografický prístup na odvodenie pseudonymov spočíva vo vhodnej realizácii decentralizovaného riešenia, ktoré umožní zúčastneným používateľom generovať vlastné pseudonymy a následne im umožní uchovávať pseudonymy pod vlastnou správou [8.]. Takýto cieľ nie je triviálnou úlohou, pretože je potrebné vyriešiť niekoľko zásadných otázok - napríklad proces generovania pseudonymov by mal zabrániť duplicitám, pričom každý používateľ by mal byť schopný kedykoľvek jednoznačne dokázať, že je vlastníkom konkrétneho pseudonymu. Všetky tieto prístupy si vyžadujú vhodné použitie viacerých kryptografických primitív (v [8.] je uvedený aj konkrétny príklad, kedy decentralizovaný prístup použil veľký výrobca zdravotníckych informačných systémov so sídlom v Nemecku). Prístup v [8.] spočíva vo využití kryptografie s verejným kľúčom - presnejšie kryptografie s eliptickou krivkou - tak, že každý používateľ si vypočíta svoj vlastný pseudonym na základe tajomstva („secret“), ktoré získa.

Konkrétny príklad použitia slúži na vydávanie elektronických lístkov na dopravu, kedy si prevádzkovateľ verejnej dopravy želá ukladať o každom cestujúcom presnú stanicu nástupu a výstupu aj s presným časom. Aby sa v tomto prípade zachovalo súkromie dotknutej osoby, nemal by prevádzkovateľ poznať jej identitu. Organisation of Domestic Transport in Athens (OASA)[40] použila riešenie znázornené na obrázku nižšie (Obrázok 14).

[40] Zdroj: https://www.oasa.gr/en/tickets/products/ath-ena-ticket/, Dátum referencie: 06.04.2023

Obrázok 14: Používateľom generovaný pseudonyem pre systém elektronických lístkov verejnej dopravy v Aténach

Správca údajov (OASA), ako aj akákoľvek iná tretia strana, ktorá získa prístup k identifikátoru karty na prepravu, ho nebude môcť previesť späť na číslo sociálneho poistenia alebo na akýkoľvek iný identifikátor používateľa. Tým sa dosiahne základná vlastnosť na ochranu súkromia pri preprave, keďže každá preprava je spojená s týmto identifikátorom. Keď používateľ stratí svoju kartu, bude môcť preukázať, že tento konkrétny identifikátor karty zodpovedá jeho osobe - t. j. jeho identifikátoru, teda bude vedieť preukázať vlastníctvo ID karty (pseudonym).

Dodatočné informácie, ktoré sú potrebné na spätnú identifikáciu každého používateľa, sú teda výlučne pod kontrolou samotného používateľa, a nie prevádzkovateľa údajov, ktorého úlohou je zabezpečiť takúto decentralizovanú techniku pseudonymizácie. Tieto prístupy - hoci sú nákladné – sa zdajú byť najlepšou možnosťou v prípadoch, keď si zásada ochrany údajov už v štádiu návrhu vyžaduje zabezpečiť, aby prevádzkovateľ údajov nemal a priori vedomosť o totožnosti dotknutej osoby, pokiaľ sa dotknutá osoba nerozhodne kedykoľvek preukázať svoju totožnosť. Táto technika je veľmi vhodná pre scenár systémov na správu osobných údajov (kapitola 10.2).

Hlavnými princípmi návrhu pseudonymov vytvorených používateľmi sú[41]:

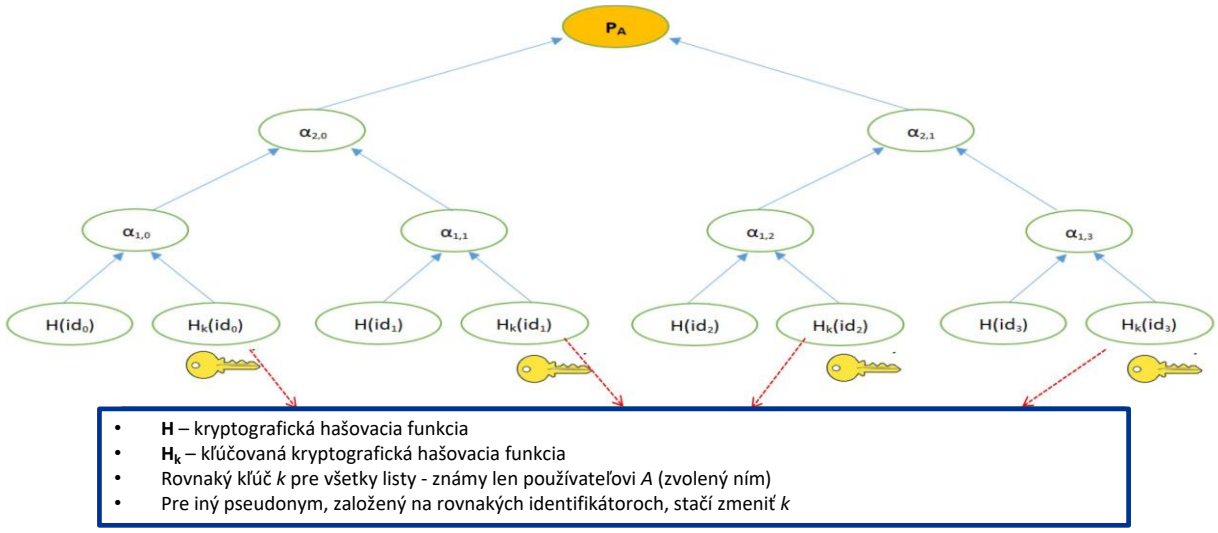
- "Skrývanie identít" - prepojenie pseudonymu s jeho vlastným používateľom by nemalo byť možné pre nikoho iného ako pre používateľa, pokiaľ to nie je výslovne povolené,
- "Neprepojiteľnosť" - V prípadoch, keď používatelia môžu mať viacero pseudonymov, by nemalo byť možné identifikovať rôzne pseudonymy ako patriace tomu istému používateľovi,
- Proces generovania pseudonymov by mal zabrániť duplicitám,
- Flexibilita - možnosť pridávať nové pseudonymy k entitám používateľov s minimálnym úsilím,
- Jednoduchosť používania.

[41] Zdroj: https://link.springer.com/chapter/10.1007/978-3-642-22890-2_10, Dátum referencie: 06.04.2023

6.1.8 Pokročilé techniky pseudonymizácie

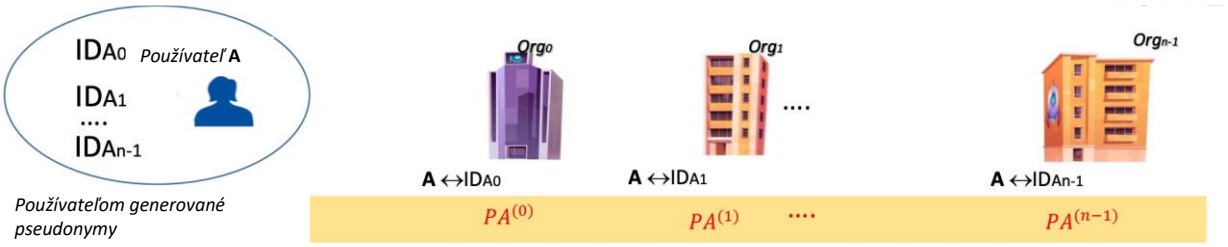
6.1.8.1 Hašovanie súborov hašov

V pokročilejších štruktúrach, ako sú Merkleho stromy[42], sa využívajú haše súborov hašov, napríklad $h_3=hash(h_1,h_2)$, na dosiahnutie štruktúrovaných pseudonymov, ktoré možno odhaliť len čiastočne, a nie úplne. Tento prístup je najviac rozvinutý a postupne sa dostáva do pripravovaných štandardov.



Obrázok 15: Odvodenie pseudonymu P_A na základe štyroch identifikátorov používateľa A

Obrázok 15 ukazuje, ako možno využiť Merkleho strom v praxi, kedy každá organizácia Org_i môže overiť, či pseudonym $PA^{(i)}$ skutočne pochádza od používateľa s identifikátorom ID_{Ai}. Používateľ A môže preukázať napríklad Org_1 , že má pseudonym $PA^{(0)}$ v Org_0, bez toho, aby odhalil akékoľvek iné informácie o iných jeho identifikátoroch (napríklad o ID_{A0}). Predpokladom je, že počiatočná registrácia každého $PA^{(i)}$ do Org_i musí byť overená.



Obrázok 16: Použitie v praxi, kedy používateľ A má rôzne doménové špecifické identifikátory

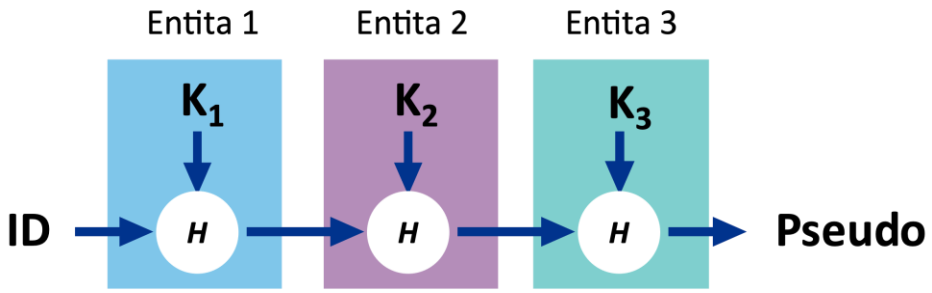
[42] Zdroj: G. Becker, "Merkle Signature Schemes, Merkle Trees and Their Cryptanalysis," Bochum, 2008

Identifikátory v Merkleho stromoch môžu byť v ľubovoľnej forme v závislosti od kontextu, ako napríklad rodné číslo, IČ DPH, e-mailová adresa či ID zariadenia.

6.1.8.2 Hašovacie reťazce

Ako bolo spomenuté v kapitole 6.1.3, nedá sa očakávať, že kryptografická hašovacia funkcia bude vhodnou technikou pseudonymizácie. Treba uprednostniť autentifikačný kód správy a kľúčované hašovacie funkcie, ktoré zahŕňajú použitie tajného kľúča. Pokročilejšie techniky však možno získať vhodným reťazením hašovacích funkcií, ako sa uvádza v tejto kapitole.

Podobne ako haše súborov hašov (6.1.8.1) aj hašovacie reťazce sa spoliehajú na opakované hašovanie hašov, napríklad *pseudo* $=h_3(h_2(h_1(x)))$, s cieľom získať hodnotu, ktorá si vyžaduje viacnásobnú inverziu hašu na spätnú identifikáciu pôvodných údajov daného pseudonymu. Pseudonymizačný reťazec zahŕňa niekoľko pseudonymizačných entít, ktoré následne preberajú pseudonymy vytvorené predchádzajúcou pseudonymizačnou entitou ako vstup na vytvorenie nových pseudonymov (napríklad použitím ďalšej vrstvy hašovania). Ako ukazuje Obrázok 17, ide o viacvrstvový prístup: (dočasne) sa generuje niekoľko nejakým spôsobom sprostredkovaných pseudonymov, aby sa nakoniec získal pseudonym, ktorý je výstupom poslednej hašovacej funkcie. Každú vrstvu vypočíta iný subjekt a každý subjekt má tajomstvo, ktoré sa používa na získanie medziproduktu pseudonymu.



Obrázok 17: Typický hašovací reťazec [11.]

Ako je znázornené na obrázku (Obrázok 17), K_1 sa používa na získanie dočasnej hodnoty $X = H_{K1}(Id)$. Hodnota X sa potom prenesie do druhého subjektu, ktorý vypočíta $Y = H_{K2}(X)$. Nakoniec posledný subjekt vypočíta $Pseudo = H_{K3}(Y)$. Takýto reťazec zmierňuje riziko porušenia ochrany údajov. Protivník musí kompromitovať tri entity, aby zvrátil pseudonymizáciu, t. j. musí poznať K_1, K_2, K_3. Teda takýto reťazec bude neprelomený aj vtedy, ak sa útočníkovi podarí odhaliť všetky pseudonymizácie, ktoré sa použili v celom reťazci, okrem jednej, čo z neho robí veľmi robustnú techniku pseudonymizácie. Ide o bežnú prax napríklad pri klinických skúškach.

Jedinou nevýhodou reťazenia je, že riešenie pseudonymov vyžaduje spoluprácu troch subjektov. Na druhej strane sa tým však zabezpečuje ďalšia vlastnosť, ktorú nemožno dosiahnuť pomocou hašovacej funkcie s jedným kľúčom: žiadna entita, ktorá dostane pseudonym v rámci procesu, ho nemôže zvrátiť, zatiaľ čo prvá entita (ktorá samozrejme pozná pôvodné identifikátory) nie je schopná porovnať konečné pseudonymy s identifikátormi (samozrejme, tieto vlastnosti platia za predpokladu, že sa medzi entitami pseudonymizácie nevymieňajú tajné kľúče). Napríklad príjemca konečného (alebo dokonca akéhokoľvek medziproduktu) pseudonymu môže vykonávať štatistickú/vedeckú analýzu pseudonymizovaných údajov bez toho, aby bol schopný priradiť pseudonymy k pôvodným identifikátorom používateľov. Hašovací reťazec možno ďalej zovšeobecniť do zložitejších štruktúr.

Pojem reťazenia pseudonymizačných mechanizmov by sa zrejme dal použiť aj všeobecnejšie, teda nielen pre kryptografické hašovacie funkcie, ale aj pre iné techniky (napríklad pre typické symetrické kryptografické algoritmy). V skutočnosti môže v závislosti od scenára každý subjekt v takomto reťazovom prístupe použiť inú techniku pseudonymizácie, čím sa umožní väčšia flexibilita, ktorá môže viesť k vzniku sofistikovanejších schém pseudonymizácie.

6.1.8.3 Bloom filtre

Ak sa vstupná doména nachádza vo viacerých dimenziách (príklad prípadu použitia je uvedený v kapitole 10.4.5), bloom filtre[43] sa okrem použitia ako anonymizačnej techniky dajú využiť aj na efektívne vykonanie výpočtovo uskutočniteľnej pseudonymizácie vo všetkých možných kombináciách vstupných hodnôt v rôznych doménach, a to napriek problému explózie stavov (všetky možné kombinácie vstupných hodnôt vytvárajú obrovské množstvo rôznych možných stavov, v ktorom bloom filter môže byť).

6.1.9 Porovnanie jednotlivých pseudonymizačných techník

Nasledujúca Tabuľka 14 poskytuje prehľadné porovnanie výhod, nevýhod a plnenia cieľov jednotlivých pseudonymizačných techník.

Tabuľka 14: Porovnanie pseudonymizačných techník

Pseudonymizačná technika	Výhody	Nevýhody	Plnenie cieľov
RNG (kapitola 6.1.1)	■ Jednoduché na implementáciu ■ Silná ochrana údajov	■ Potreba chrániť mapovaciu tabuľku ■ Problém so škálovateľnosťou ■ Možná kolízia	Splnenie cieľov V1 a V2 za predpokladu, že nenastane kompromitácia mapovacej tabuľky alebo kolízia
CSPRNG (kapitola 6.1.2)	■ Jednoduché na implementáciu ■ Existujú štandardizované generátory, ■ Odstránený problém s kolíziou, čím je ešte posilnená ochrana údajpv oproti RNG	■ Potreba chrániť mapovaciu tabuľku, ■ Problém so škálovateľnosťou	Splnenie cieľov V1 a V2 za predpokladu, že nenastane kompromitácia mapovacej tabuľky
Hašovanie bez kľúča (kapitola 6.1.3)	■ Slúži na porozumenie komplexnejším technikám pseudonymizácie ■ Vlastnosť nekorelovania – napr. pre SHA3-512	■ Slabá technika pseudonymizácie, neodporúča sa, pretože pomerne jednoducho sa dá vykonať útok pomocou hrubej sily (8.3.1)	Vlastnosť V2 neplatí, pretože akákoľvek tretia strana, ktorá použije rovnakú hašovaciu funkciu na rovnaký identifikátor, dostane rovnaký pseudonym.

[43] Zdroj: B. H. Bloom, “Space/time trade-offs in hash coding with allowable errors,” Communications of the ACM, pp. 422-426, júl 1970.

Pseudonymizačná technika	Výhody	Nevýhody	Plnenie cieľov
	0,5% zmena vo vstupnej hodnote vedie k 85 až 95% zmene vo výstupnom haši[44] ■ Možno použiť na kontrolu, či sa údaje prenosom nepozmenili	■ Vyššie nároky na úložisko ako pôvodné údaje	Nemusí nevyhnutne platiť ani vlastnosť V1, pretože pre akúkoľvek tretiu stranu je triviálne overiť pre daný identifikátor, či pseudonym zodpovedá tomuto identifikátoru.
Hašovanie pomocou kľúča alebo soli (kapitola 6.1.4)	■ Zachováva vlastnosť nekorelovania ■ Pridáva náhodnosť do procesu, čím zásadne sťažuje útok pomocou hrubej sily (potreba vedieť algoritmus pre RNG na generovanie soli, štruktúru hašovaných osobných údajov, hodnotu korenia ("pepper"), ak sa používa) ■ Možno takto hašovať nielen jeden dátový prvok, ale aj celý záznam či dataset, čím sa ešte viac zvýši náročnosť identifikovať štruktúru vstupných osobných údajov ■ Odporúčané pre dátové sklady, a teda aj KAV ■ Použitie kľúčovanej hašovacej funkcie na generovanie pseudonymov a následné vymazanie tajného kľúča je istým spôsobom ekvivalentné generovaniu náhodných pseudonymov bez akéhokoľvek spojenia s pôvodnými identifikátormi	■ Vyššie nároky na úložisko ako pôvodné údaje ■ Ak prevádzkovateľ údajov potrebuje prideliť ten istý pseudonym tej istej osobe, musí použiť ten istý tajný kľúč – tajné kľúče treba ukladať a chrániť	Pre rovnaký vstup možno vytvoriť niekoľko rôznych pseudonymov podľa výberu konkrétneho kľúča alebo soli - a tým je zabezpečená vlastnosť V2. Okrem toho platí aj vlastnosť V1, pokiaľ akákoľvek tretia strana nepozná kľúč alebo soľ a korenie, nemôže overiť, či pseudonym zodpovedá konkrétnemu známemu identifikátoru.

[44] Zdroj: https://www.slideshare.net/alanmcsweeney/data-privatisation-data-anonymisation-data-pseudonymisation-and-differential-privacy-250972293, Dátum referencie: 06.04.2023

Pseudonymizačná technika	Výhody	Nevýhody	Plnenie cieľov
Šifrovanie (kapitola 6.1.5)	■ Možno šifrovať nielen jeden dátový prvok (štruktúrované polia aj neštruktúrované údaje), ale aj celý záznam či dataset, alebo aj celú databázu ■ Škálovateľnosť - krátky šifrovací kľúč umožňuje zašifrovať a dešifrovať aj veľké objemy údajov naraz ■ Hlavný rozdiel v porovnaní s hašovacími funkciami s kľúčom - z hľadiska pseudonymizácie - spočíva v tom, že vlastník tajného kľúča môže vždy získať pôvodné identifikátory prostredníctvom jednoduchého procesu dešifrovania ■ Šifrovanie je ideálne na výmenu citlivých informácií s tými, ktorí majú šifrovací kľúč.	■ Potreba starostlivo chrániť tajné symetrické a súkromné kľúče, aby niekto neoprávnený nezískal pôvodné identifikátory procesom dešifrovania ■ Ak existuje verejný kľúč, musí byť distribuovaný v certifikáte, aby sa zabránilo útokom na zachytenie. ■ Šifrovacie schémy zachovávajúce formát majú však nižšiu silu zabezpečenia ■ Šifrovacie kľúče sa musia pravidelne obmieňať, pretože kryptografické algoritmy sa neustále analyzujú a môžu sa v nich nájsť zraniteľné miesta.	Pseudonym vytvorený šifrovaním spĺňa vlastnosť V2, ako aj vlastnosť V1, pokiaľ k šifrovaciemu kľúču nemá prístup žiadna tretia strana a za predpokladu, že sa používajú najmodernejšie algoritmy a dostatočná dĺžka kľúča.
Tokenizácia (kapitola 6.1.6)	■ Systémy tokenizácie generujú náhodné hodnoty tokenov bez prepojenia s pôvodným identifikátorom ■ Systémy sú certifikované a štandardizované ■ Keď sa zväčšuje veľkosť databáz, je ťažké bezpečne škálovať a udržať výkon ■ Tokenizácia umožňuje organizáciám zachovať formáty bez zníženia sily zabezpečenia	■ Mapovanie tokenov je uložené v databáze, ktorú treba starostlivo chrániť ■ Výmena údajov môže byť veľmi náročná, pretože si vyžaduje priamy prístup k trezoru tokenov, ktorý mapuje hodnoty tokenov ■ Len štruktúrované údaje ako sú rodné čísla alebo informácie o platobných kartách. ■ Napriek účinnosti tokenizácie môže byť jej nasadenie v závislosti od kontextu pomerne náročné. Napríklad v niekoľkých IS môže byť potrebná synchronizácia tokenov Potom sú vhodnejšie skôr uvedené prístupy, ktoré využívajú kľúčované hašovacie	Vďaka náhodnému skrytému mapovaniu z pôvodných údajov na token je zrejmé, že tokenizácia spĺňa vlastnosti pseudonymizácie V1 aj V2. Keďže existuje subjekt, ktorý toto skryté mapovanie uchováva (tokenový server v systéme tokenizácie), spätná identifikácia údajov subjektov zo strany prevádzkovateľa bude možná vo všetkých prípadoch.

Pseudonymizačná technika	Výhody	Nevýhody	Plnenie cieľov
		funkcie (kapitola 6.1.4) alebo šifrovacie algoritmy (kapitola 6.1.5).	
Decentralizácia (kapitola 6.1.7)	■ Umožňuje zúčastneným používateľom generovať vlastné pseudonymy a uchovávať ich pod vlastnou správou ■ Prevádzkovateľ tak nepozná identitu používateľa – dotknutej osoby, pokiaľ svoju identitu dobrovoľne neodhalí ■ Vhodné pri implementácii osobných systémov na manažment osobných údajov	■ Riziko kolízie pseudonymov ■ Náročné na implementáciu, techniky vo výskume a vývoji	Takto vytvorený pseudonym spĺňa vlastnosti V1 aj V2, keďže pravú identitu má plne pod kontrolou používateľ – dotknutá osoba, a nikomu ju nemusí zazdieľať.
Pokročilé techniky (kapitola 6.1.8)	■ Uľahčenie výmeny informácií medzi prevádzkovateľmi údajov, pokiaľ ide o minimalizáciu údajov ■ Výmena údajov prebieha na žiadosť používateľa	■ Náročné na implementáciu, techniky vo výskume a vývoji, odporúčame zamerať sa v prípade potreby len na Merkleho stromy	Okrem cieľov V1 a V2, vedia spĺňať aj cieľ V3 - vytvárajú pseudonymy matematicky viazané na pôvodné identifikátory, a preto tieto pseudonymy môžu postačovať na to, aby umožnili overenie totožnosti dotknutých osôb bez odhalenia úplného súboru osobných údajov identity.

6.2 Anonymizačné techniky

Techniky uvedené v tejto kapitole predstavujú predovšetkým doplnkové techniky k pseudonymizácii na účel jej posilnenia. To znamená, že ak by došlo k prelomeniu techniky pseudonymizácie alebo šifrovania mimo účelu pseudonymizácie, budú údaje v istej miere ešte aj anonymizované (na určenie miery anonymizácie slúžia techniky popísané v kapitole 6.3).

Tiež anonymizačné techniky v tomto dokumente navrhujeme predovšetkým pre tieto prípady použitia:

- Maskovanie údajov na:
  - školenie používateľov,
  - vývoj softvérových riešení,
  - testovanie softvérových riešení,
  - alebo na niektoré štatistické analýzy.
- Generalizácia údajov:

- Analytické spracovanie údajov,
- Dolovanie veľkých údajov (“Big Data”).

6.2.1 Maskovanie údajov

Maskovanie údajov zahŕňa umožnenie prístupu k upravenej verzii citlivých údajov – teda údajov na úrovni „Vyhradené“ a za určitých okolností aj na úrovni „Dôverné“. Ide o techniku na vytvorenie verzie údajov, ktorá sa štruktúrou podobá pôvodným údajom, ale skrýva (maskuje) citlivé informácie. **Verzia so zamaskovanými informáciami sa potom môže použiť na rôzne účely, napríklad na školenie používateľov, vývoj a testovanie softvéru alebo na niektoré štatistické analýzy.** Hlavným cieľom maskovania údajov je vytvoriť funkčnú náhradu, ktorá neodhalí skutočné údaje, teda ich možno aj zdieľať s tretími stranami. To možno dosiahnuť:

- **Statickým maskovaním údajov**, teda vytvorením zrkadlovej verzie databázy s anonymizovanými údajmi, ktorá obsahuje úplne alebo čiastočne maskované údaje. Fiktívna databáza sa udržiava oddelene od produkčnej databázy.
- **Dynamickým maskovaním údajov**, teda úpravou údajov v reálnom čase, keď sa k nim pristupuje: táto technika sa používa priamo na produkčné datasety. Zabezpečuje, aby pôvodné údaje videli len autorizovaní používatelia, používatelia bez oprávnení vidia len maskované údaje.
- **Maskovaním údajov „za pochodu“** - modifikuje citlivé informácie pri prenose medzi prostrediami, čím zabezpečuje, že citlivé údaje sú maskované skôr, ako sa dostanú do cieľového prostredia. Táto technika je ideálna pre organizácie, ktoré migrujú údaje medzi systémami alebo udržiavajú kontinuálnu integráciu či synchronizáciu rozdielnych súborov údajov.

Anonymizácia sa môže vykonávať pomocou rôznych techník vrátane:

- **Šifrovania** (kapitola 6.1.5), kedy ale samozrejme neoprávnení používatelia nedokážu zašifrované údaje dešifrovať, respektíve sa kľúč na dešifrovanie zahodí.
- Ako aj pomocou rôznych ďalších techník **pseudonymizácie** v kapitole 6.1 za rovnakého predpokladu, že neoprávnení používatelia vidia vždy len bezvýznamový reťazec, napríklad výstup hašovacej funkcie (kapitola 6.1.4) alebo token (kapitola 6.1.6).
- **Nahradením pomocou slovníka**: Produkčnú databázu možno maskovať vytvoreným slovníkom, ktorý poskytuje alternatívne hodnoty k pôvodným citlivým údajom. To umožní napríklad používať reálne údaje v testovacom prostredí bez toho, aby sa odhalil originál.
- **Priemerovaním**: Ak treba reflektovať citlivé údaje z hľadiska priemerov alebo agregátov, ale nie na individuálnom základe, možno všetky hodnoty v tabuľke nahradiť priemernou hodnotou. Ak napríklad tabuľka obsahuje zoznam platov zamestnancov, môžeme skutočné individuálne platy zamaskovať tak, že ich všetky nahradíme priemerným platom, takže celkový stĺpec bude zodpovedať skutočnej priemernej hodnote všetkých platov.
- **Zamiešania termínov alebo znakov („Scrambling“)**: Ak treba pri maskovaní hodnôt zachovať jedinečnosť, možno údaje chrániť poprehadzovaním tak, že skutočné hodnoty zostanú zachované, ale budú priradené iným jednotlivcom. V príklade tabuľky platov by boli uvedené všetky skutočné platy, ale nebolo by odhalené, ktorý plat patrí jednotlivým zamestnancom, pretože by boli náhodne premiešané. Táto metóda je najvhodnejšia pre väčšie súbory údajov. Problémom je, že tento proces je niekedy zvratný, napríklad „Obama“ sa môže stať „Baoam.“

Podmienkou je, že pri maskovaní údajov musia byť hodnoty vždy zmenené takým spôsobom, ktorý znemožňuje spätnú identifikáciu. S maskovaním údajov sú spojené nasledujúce výzvy:

- **Zachovanie formátu** - riešenie na maskovanie údajov musí rozumieť údajom (t. j. tomu, čo predstavujú). Keď systém maskovania nahradí pôvodné údaje neautentickými údajmi, mal by zachovať pôvodný formát. To je dôležité najmä v prípade dátových reťazcov, ktoré si vyžadujú špecifické poradie alebo formát, napríklad dátumy.
- **Referenčná integrita** - tabuľky v relačnej databáze sú prepojené prostredníctvom primárnych kľúčov. Keď maskovacie riešenie zakryje alebo nahradí hodnoty primárneho kľúča tabuľky, tieto hodnoty sa musia konzistentne modifikovať v celej databáze.
- **Zachovanie pohlavia** - systém maskovania by mal mať pri nahrádzaní mena osoby v databáze prehľad o pohlaví a mal by byť schopný zistiť, či ide o mužské alebo ženské meno. Rozloženie pohlaví v tabuľke sa zmení, ak maskovací systém náhodne mení mená.
- **Sémantická integrita** - databázy zvyčajne presadzujú pravidlá, ktoré obmedzujú rozsah povolených hodnôt (napríklad rozsah platov). Všetky maskované údaje musia patriť do špecifikovaného rozsahu, aby sa zachovala sémantika (význam) údajov.
- **Jedinečnosť údajov** - pri maskovaní jedinečných údajov by mal maskovací systém použiť jedinečné hodnoty pre každý prvok údajov. Ak napríklad príslušná tabuľka uchováva rodné čísla zamestnancov, každý zamestnanec by mal dostať unikátne číslo aj po maskovaní. Pravdepodobnostné rozdelenie maskovaných dát by malo byť zhodné s rozdelením pôvodných údajov, obzvlášť ak má význam, napríklad pri geografickom rozložení populácie. Každý stĺpec tabuľky by mal mať v priemere rovnaké maskované hodnoty k originálu.

Z uvedených výziev vyplýva, že nemožno aplikovať rovnaký prístup k maskovaniu plošne, ale treba ho prispôsobiť danému datasetu, prípadu použitia a stavu technológií v informačných systémoch.

6.2.2 Generalizácia

Generalizácia je postup, pri ktorom sa špecifická hodnota nahrádza všeobecnejšou hodnotou. Napríklad poštové smerovacie čísla v súbore údajov možno zovšeobecniť na okresy alebo obce (to znamená. zmeniť 85101 na 851**). Vek sa môže zovšeobecniť na vekovú kategóriu (teda zoskupiť vek 22 a 25 do kategórie [20-30]).

Táto technika odstraňuje rozpoznávanie informácií, ktoré možno získať z datasetu znížením špecifickosti atribútu. Používa sa predovšetkým pri dolovaní údajov („Data Mining“) pri vytváraní analytických výstupov. Dolovanie údajov sa bude vykonávať predovšetkým v Konsolidovanej analytickej vrstve. Ide o techniku na získavanie vzorov a iných užitočných informácií z obrovských datasetov. V dôsledku pokroku v technológiách dátových skladov a nárastu veľkého objemu údajov sa v posledných desaťročiach prudko rozšírilo používanie techník dolovania údajov pri premene nespracovaných údajov na cenné znalosti.

V dolovaní údajov existujú dve hlavné formy generalizácie údajov:

1. **Automatizovaná generalizácia** skresľuje hodnoty, kým sa nedosiahne daná hodnota *k pre* k-anonymitu (kapitola 6.3.1). Keďže možno použiť algoritmus na aplikovanie čo najmenšieho skreslenia potrebného na dosiahnutie uvedenej hodnoty *k*, táto metóda môže ponúknuť optimálnu rovnováhu medzi súkromím a presnosťou.

2. **Deklaratívna generalizácia** naopak umožňuje vopred nastaviť veľkosti „košov“ („bins“), napríklad dátum vždy zaokrúhliť na celé mesiace. Pri tomto postupe sa niekedy vylučujú odľahlé hodnoty, ktoré môžu skresľovať údaje a pridávať skreslenie. Treba si ale uvedomiť, že deklaratívna generalizácia nemusí vždy viesť ku k-anonymite.

Hoci deklaratívna generalizácia nemusí pomôcť dosiahnuť k-anonymitu, je dobré ju používať ako predvolenú. Tretia strana, ktorá prijíma de-identifikované údaje, teda vidí len takú úroveň detailov, akú potrebuje.

Ako už bolo spomenuté, existujú dva hlavné typy identifikátorov: priame identifikátory a kváziidentifikátory, ktoré sa tiež používajú pri generalizácii údajov v dolovaní údajov:

1. **Priame identifikátory** sú dátové body, ktoré môžu identifikovať jednotlivca a zároveň umožňujú prepojenie iných údajov s touto osobou. Aj keď v údajoch existuje viacero rovnakých dátových bodov, dátový bod môže byť priamym identifikátorom. Napríklad aj keď sa dve osoby volajú "Mária", meno je stále priamym identifikátorom.
2. Na druhej strane, **kvázi-identifikátory** neumožňujú samé o sebe identifikovať osobu. Napriek tomu ich možno na tento účel použiť v spojení s ďalšími informáciami. Kváziidentifikátory môžu byť v rámci datasetu jedinečné. Napriek tomu sa očakáva, že sa čoskoro objavia aj v iných datasetoch alebo sa v súčasnosti už vyskytujú v iných jedinečných datasetoch.

Predpokladajme, že máme dataset, ktorý obsahuje pohlavie a poštové smerovacie číslo osoby. V danom poštovom smerovacom čísle žije dostatok ľudí tohto pohlavia, aby sa žiadna osoba nedala identifikovať len na základe týchto dvoch dátových premenných. Predpokladajme však, že istá osoba z tohto datasetu sa vyskytuje aj v inom datasete, ktorý obsahuje jej pohlavie, poštové smerovacie číslo a adresu. Ak by na tejto adrese bol len jeden dom, v ktorom bývajú dve osoby, jedna žena a druhý muž, tak už presne vieme, o koho sa jedná.

Kedy je teda generalizácia údajov dôležitá? Generalizácia údajov pri dolovaní údajov umožňuje abstrahovať osobné údaje odstránením identifikačných znakov. Toto zovšeobecnenie umožňuje preskúmať zozbierané údaje bez toho, aby sa ohrozilo súkromie ľudí v danom datasete. Je veľmi dôležité si uvedomiť, že existuje niekoľko metód generalizácie údajov a treba si vybrať tú, ktorá má pre daný prípad použitia najväčší zmysel.

Agregácia údajov je pojem, ktorý súvisí generalizáciou údajov v dolovaní údajov a často sa s ním zamieňa. Základným rozdielom medzi generalizáciou údajov a agregáciou údajov je to, že agregácia vytvára všeobecnú triedu z mnohých tried, do ktorej akumuluje jednotlivé pozorovania o jednotlivcoch. Naproti tomu generalizácia je proces vytvárania špecifickej všeobecnej triedy z mnohých tried pre citlivé premenné, ktoré sú identifikátormi alebo kváziidentifikátormi. Príkladom dátovej agregácie je dátová kocka v databázových systémoch typu OLAP[45]. Každá dimenzia dátovej kocky reflektuje iný aspekt databázy, napríklad denné, mesačné alebo ročné výdavky. Údaje v dátovej kocke umožňujú analyzovať takmer všetky metriky pre všetkých dodávateľov, produkty, služby a podobne. Vďaka tomu môže dátová kocka pomôcť pri identifikácii trendov a analýze výkonnosti.

[45] Zdroj: https://www.youtube.com/watch?v=yoE6bgJv08E, Dáturm referencie: 13.02.2023

6.2.3 Potlačenie

Potlačenie je proces úplného odstránenia hodnoty atribútu z datasetu. V príklade informácií o veku by potlačenie znamenalo úplné odstránenie údajov o veku z každej kohorty.

Majte na pamäti, že potlačenie by sa malo používať len v prípade údajov, ktoré nie sú relevantné pre účel zberu alebo analýzy údajov. Ak napríklad zhromažďujete údaje na určenie toho, v akom veku majú jednotlivci najväčšiu šancu na vznik konkrétnej choroby alebo stavu, potlačením údajov o veku by sa samotné údaje stali nepoužiteľnými.

Potlačenie by sa malo uplatňovať na väčšinou irelevantné dátové body, prípad po prípade, namiesto používania súboru všeobecných pravidiel, ktoré sa uplatňujú univerzálne.

6.2.4 Globálne prekódovanie

Pri tejto metóde možno spojité alebo diskrétne číselné premenné zoskupiť do preddefinovanej triedy. Znamená to, že daná špecifická hodnota sa nahradí všeobecnejšou hodnotou, ktorú možno vybrať odkiaľkoľvek z celého datasetu.

Napríklad pri globálnom prekódovaní sa môže v datasete zovšeobecniť poštové smerovacie číslo bez ohľadu na pohlavie alebo inú opisnú premennú. Proces prekódovania môže byť jednorozmerný alebo viacrozmerný:

- Pri **jednorozmernom prekódovaní** sa každý atribút mapuje samostatne (napríklad poštové smerovacie číslo).
- Pri **viacrozmernom prekódovaní** sa mapovanie môže vykonávať na základe funkcie viacerých atribútov spoločne, ako napríklad pri kvázi-identifikátoroch (napríklad poštové smerovacie číslo, pohlavie a dátum narodenia).

6.3 Techniky merania vplyvu techník ochrany súkromia

6.3.1 K-anonymita

Podľa profesora Sweeneyho kombinácia dátumu narodenia, pohlavia a poštového smerovacieho čísla stačí na jednoznačnú identifikáciu najmenej 87 % obyvateľov USA vo verejne prístupných databázach. Zabezpečenie skutočnej záruky súkromia sa musí dokázať a stanoviť matematicky, a tu pomáha K-anonymita.

Intuitívne vysvetlenie tejto metódy pochádza z Carnegie Mellon University[46]: „Pri k-anonymite sa atribúty potláčajú dovtedy, kým každý riadok nie je identický s aspoň *k- 1* inými riadkami. Vtedy sa hovorí, že databáza je k-anonymná. K-anonymita teda zabraňuje definitívnym prepojeniam databázy s jednotlivcami. V najhoršom prípade sa zverejnené údaje týkajú skupiny *k* osôb a nie jednotlivcov.“

[46] K-Anonymity," www.cs.cmu.edu/~jblocki/Slides/K-Anonymity.pdf , Dátum referencie: 03.02.2023

K-anonymita sa implementuje intuitívne, používa ju napríklad spoločnosť Google vo svojom rozhraní API pre reklamu a poskytuje záruku, že ste jedným z minimálnej skupiny, a nie že vás možno jednoznačne identifikovať.

Na prvý pohľad anonymita znamená mať databázu s údajmi bez mena a jednoznačného identifikátora (PID). Keď však začneme anonymitu študovať bližšie, rýchlo pochopíme, že na dosiahnutie skutočnej anonymizácie nestačí len odstránenie mien a jednoznačného identifikátora z databáz či datasetov. Anonymizované údaje je možné spätne identifikovať spojením s iným datasetom. Údaje môžu obsahovať časti informácií, ktoré samy o sebe nie sú jedinečnými identifikátormi, ale po prepojení s inými datasetmi ich možno identifikovať.

K-anonymita zabraňuje definitívnemu prepojeniu databáz. Definuje atribúty, ktoré nepriamo poukazujú na identitu osoby, ako kvázi-identifikátory, a spracováva údaje tak, že aspoň *k* osôb má rovnakú kombináciu hodnôt kvázi-identifikátorov. Výsledkom je, ako už bolo spomenuté, že v najhoršom prípade zverejnené údaje zužujú individuálny záznam na skupinu *k* jednotlivcov.

Najbežnejšie implementácie K-anonymity využívajú anonymizačné transformačné techniky, ako je generalizácia (kapitola 6.2.2), potlačenie (kapitola 6.2.3) a globálne prekódovanie (kapitola 6.2.4).

Nasledujúci príklad načrtne, ako sa identifikovateľnosť používateľa mení na základe presnosti údajov. Budeme sa zaoberať tisíckami jázd, kde každá jazda má miesto vyzdvihnutia a cieľa. Budeme meniť počet desatinných miest v súradniciach GPS miesta, aby sme mohli poskytnúť rôzne stupne presnosti. Ak je poloha GPS veľmi presná, môže opisovať konkrétnu adresu, napríklad niečí dom. Ak je menej presná, môže opisovať blok alebo kilometre štvorcové, v takom prípade by sa mohlo zoskupiť veľa rôznych jázd do jednej.

K-anonymita s nepresnými údajmi

Tabuľka 15 ukazuje, ako funguje k-anonymita pri rôznych úrovniach presnosti údajov o polohe. V tabuľke riadky predstavujú počet desatinných miest v údajoch o polohe (od 0 po 5 desatinných miest), zatiaľ čo stĺpce predstavujú hodnotu *k*-anonymity od 2 po 1 000. Na vysvetlenie hodnoty *k* sa zamerajme na predposledný a posledný riadok s označením 4 a 5 a tretí stĺpec s označením 5 (vrátane stĺpca, kde sú v riadkoch hodnoty desatinných miest). Keď je v polohe GPS 5 desatinných miest, iba 35,5 % záznamov o polohe malo 5 ďalších hodnôt, priradených k ceste, ktoré boli podobné. To znamená, že ostatných 64,5 % záznamov malo menej ako 5 hodnôt, ktoré boli podobné, čím sa stali identifikovateľnými. Odstránenie jednej desatinnej čiarky - umožnenie len 4 desatinných miest v GPS - znamená, že v tejto vzorke 93,2 % súborov údajov malo iné podobné údaje, čím sa znížila pravdepodobnosť identifikácie v porovnaní s použitím 5 desatinných miest. To znamená, že čím sú údaje presnejšie, tým je menšia pravdepodobnosť, že sa vyskytnú ďalšie podobné údaje, čím je každý záznam oveľa lepšie identifikovateľný. **Čím je hodnota k-anonymity vyššia, tým údaje zabezpečujú viac súkromia**.

Tabuľka 15: K-anonymita s 0 desatinnými miestami v údajoch o polohe

		Hodnota K-anonymity					
		2	**5**	**10**	**50**	**100**	**1 000**
Počet desatinných	**0**	100%	100%	100%	100%	100%	100%
	1	100%	100%	100%	100%	100%	100%

		Hodnota K-anonymity					
		2	**5**	**10**	**50**	**100**	**1 000**
miest v údajoch o polohe	**2**	100%	100%	100%	99.9%	99.9%	99.9%
	3	99.9%	99.8%	99.5%	97.6%	95.3%	87.9%
	4	97.4%	93.2%	89.3%	73.1%	59.3%	17.3%
	5	68.4%	35.5%	18.3%	2.5%	1.5%	0.9%

Dve cesty s rovnakou hodnotou by museli mať rovnaké miesta nástupu a výstupu. V hornom riadku tabuľky s označením 0 (Tabuľka 15) na základe údajov o polohe GPS s 0 desatinnými miestami vidíme, že pre všetkých používateľov (100 %) môžeme nájsť aspoň 1 ďalšiu cestu (čo nám dáva k-anonymitu 2), 4 ďalšie cesty (čo nám dáva k-anonymitu 5), až po 999 ďalších ciest (čo nám dáva k-anonymitu 1 000) používateľov s rovnakou hodnotou cesty. Keď zdieľame údaje o polohe GPS s 0 desatinnými miestami, používatelia sú menej identifikovateľní a majú vysokú k-anonymitu. Dôvodom, prečo sa to deje, je, že veľa rozdielnych miest vyzdvihnutia a vysadenia sa z dôvodu zaokrúhľovania GPS stalo rovnakými. Tu je cieľ k-anonymity splnený, ale za obety kvality údajov.

K-anonymita s presnými údajmi

Tabuľka 16 ukazuje, ako funguje k-anonymita, keď potrebujeme mať veľmi presné údaje o polohe. Čím viac desatinných miest sa nachádza v súradniciach GPS, tým presnejšia je poloha používateľa a tým viac je používateľ identifikovateľný.

Tabuľka 16: K-anonymita s 4 až 5 desatinnými miestami v údajoch o polohe

		Hodnota K-anonymity					
		2	**5**	**10**	**50**	**100**	**1 000**
Počet desatinných miest v údajoch o polohe	**0**	100%	100%	100%	100%	100%	100%
	1	100%	100%	100%	100%	100%	100%
	2	100%	100%	100%	99.9%	99.9%	99.9%
	3	99.9%	99.8%	99.5%	97.6%	95.3%	87.9%
	4	97.4%	93.2%	89.3%	73.1%	59.3%	17.3%
	5	68.4%	35.5%	18.3%	2.5%	1.5%	0.9%

Keď sa pozrieme na spodné dva riadky v tabuľke, ktoré majú 4 alebo 5 desatinných miest pre polohu GPS pri vyzdvihnutí a vysadení, tak je k-anonymita nižšia, pretože počet používateľov, ktorí spĺňajú presnú polohu GPS, je nižší. Preto čím nižšie v tabuľke ideme, tým nižšie je percento kohorty, ktorá spĺňa príslušný prah k-anonymity, a tiež klesá, keď ideme zľava doprava. Napríklad pri zobrazení 5 desatinných miest má 68,4 % používateľov k-anonymitu 2, čo znamená, že pre 68,4 % používateľov môžeme nájsť jedného ďalšieho používateľa s rovnakými hodnotami cesty. Ak obetujeme jedno desatinné miesto a poloha GPS bude mať 4 desatinné miesta, vidíme, že pre 97,4 % používateľov môžeme nájsť podobnú jazdu, takže majú

hodnotu k-anonymity 2. Rovnako ako predtým, čím menej presné údaje, tým väčšiu anonymitu môžeme používateľom poskytnúť.

Úplne vpravo, ak má GPS poloha 5 desatinných miest a chceme zabezpečiť k- anonymitu 1 000, čo znamená, že môžeme nájsť 999 ďalších podobných jázd, môžeme tak urobiť len pre 0,9 % jázd, takže vzhľadom na presnosť údajov o polohe je anonymita veľmi malá. Čísla sa trochu zlepšia, ak odstránime jedno desatinné miesto a zaokrúhlime polohu na 4, ale stále môžeme poskytnúť k-anonymitu 1 000 len pre 17,4 %, teda približne pre 1 zo 6 jázd. **Ak potrebujeme vyššie percento používateľov na splnenie konkrétnej hodnoty k- anonymity, budeme musieť poskytnúť menej desatinných miest v údajoch o polohe.**

K-anonymita s osvedčenými postupmi v odvetví

Na záver sa zameriame na **hodnotu k-anonymity 5, pretože sa považuje za najlepší postup v odvetví** (zelený stĺpec v tabuľke - Tabuľka 16). V tomto prípade sú údaje zastrené tak, že pre každý záznam budú existovať aspoň 4 ďalšie, ktoré sú od neho nerozoznateľné, čím sa tento záznam stane viac chráneným z hľadiska ochrany súkromia a menej individuálne identifikovateľným. V našom príklade teda pre konkrétnu cestu existuje 5 ďalších podobných ciest, čím sa jednotlivé cesty v tejto skupine stávajú menej identifikovateľnými. Keď prechádzame po percentách v tomto stĺpci zhora nadol, môžeme vidieť, čo sa stane s k-anonymitou, keď pridáme viac desatinných miest do našich údajov o polohe. Ako sme už zistili, čím viac desatinných miest pridáme, tým sú naše údaje presnejšie a tým viac bude identifikovateľný používateľ, ktorého súradnice GPS sledujeme. Počet používateľov, ktorí majú k-anonymitu 5, sa teda zníži.

Keď máme 0 desatinných miest (veľmi nepresné umiestnenie), môžeme nájsť 4 ďalšie s rovnakými hodnotami pre všetkých používateľov (t. j. 100 % používateľov má k- anonymitu 5). To isté platí pre 1 a 2 desatinné miesta - môžeme mať teda súradnice GPS s maximálne 2 desatinnými miestami a stále budeme mať k-anonymitu 5 pre všetky záznamy. Keď pridáme tretiu desatinnú čiarku, narazíme na bod zlomu - je to prvýkrát, čo nie každý používateľ má k-anonymitu 5. Existujú aspoň niektorí používatelia, pre ktorých nebudú existovať aspoň 4 ďalší s rovnakými hodnotami. Ukazuje sa však, že aj keď pridáme tretiu desatinnú čiarku a zároveň sa snažíme o k-anonymitu 5, stále zahrnieme 99,8 % používateľov. Ak je teda naším cieľom dosiahnuť k-anonymitu 5, potrebujeme potlačiť iba 0,2 % údajov.

6.3.2 L-diverzita

L-diverzita je ďalšou formou skupinovej anonymizácie ako k-anonymita, ktorá sa používa na zachovanie súkromia v súboroch údajov znížením granularity modelu reprezentácie údajov prostredníctvom metód ako generalizácia (kapitola 6.2.2) a potlačenie (kapitola 6.2.3).

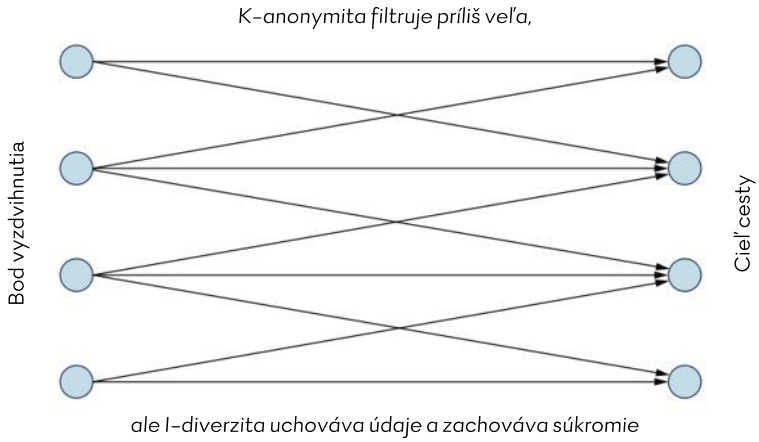
Model l-diverzity sa často používa aj ako technika na posúdenie, či úsilie o k-anonymitu zašlo dostatočne ďaleko, aby sa zabránilo spätnej identifikácii. **O datasete sa hovorí, že spĺňa l-diverzitu, ak existuje aspoň *l* dobre reprezentovaných hodnôt pre každý citlivý atribút v každej skupine záznamov, ktoré majú spoločné kľúčové atribúty.** Inými slovami, každý atribút, ktorý sa považuje za citlivý, ako napríklad zdravotný stav osoby alebo to, či študent zložil alebo nezložil skúšku, nadobúda v rámci každej podmnožiny *k* aspoň *l* rôznych hodnôt.

Najlepšia prax v odvetví, k-anonymita 5, poskytuje zmysluplnú rovnováhu medzi súkromím a použiteľnosťou údajov. K-anonymita má však svoje obmedzenia, a preto je k dispozícii ďalší nástroj, ktorý pomáha anonymizovať údaje pred ich zdieľaním: l- diverzita.

Uvažujme ďalej o prípade použitia z tabuľky (Tabuľka 16) v kapitole 6.3.1, ktorý ukazuje obmedzenia k-anonymity a to, ako môže pomôcť l-diverzita.

Predpokladajme, že máme dosiahnutú k-anonymitu 5, ale pritom existuje aspoň jeden taký bod vyzdvihnutia, že každá cesta z tohto bodu vyzdvihnutia smeruje do rovnakého cieľa. V tomto prípade môžeme pomocou externých údajov zistiť, kam ide každý cestujúci z tohto zdroja. Tu môže pomôcť l-diverzita. L-diverzita pomôže zabezpečiť rozmanitosť potenciálnych zdrojov alebo cieľov. Teda pre každú cestu, ktorá sa vykazuje v danom časovom okne, musí mať vyzdvihnutie aspoň l rôznych potenciálnych vysadení a každé výstupné miesto musí mať l potenciálnych vyzdvihnutí.

Môžu nastať situácie, keď k-anonymita odfiltruje príliš veľa údajov, a v týchto prípadoch môže byť l-diverzita oveľa lepším nástrojom. Obrázok 18 pomáha túto skutočnosť objasniť. Na obrázku predstavujú body na ľavej strane vyzdvihnutie jazdy a body na pravej strane ukončenie jazdy. Ak použijeme hodnotu k-anonymity 2, odfiltrujeme všetky jazdy, pretože žiadne dve jazdy nemajú rovnaký nástup a výstup. Na druhej strane, ak od seba oddelíme vyzdvihnutia a vysadenia a použijeme l -diverzitu 2, môžeme zachovať celý súbor údajov a zároveň zachovať aj súkromie.



Obrázok 18: L-diverzita v praxi

V tomto konkrétnom prípade predpokladajme, že sa snažíme preskúmať hustotu vyzdvihnutí. Ak sa na určitom mieste zvyšuje počet vyzdvihnutí, možno tam bude treba poslať viac vodičov, aby sa skrátil čas čakania. V tomto prípade má ešte väčší zmysel oddeliť vyzdvihnutia a vysadenia a neukladať ich ako jednu cestu. Dokonca by sa mohla vymazať aj časť údajov o vysadení. Tak zostane menej údajov, čo znamená nižšie náklady na ukladanie. Údaje, ktoré ostanú, sa viac vzťahujú na tento prípad použitia, takže je k dispozícii lepšia kvalitu údajov. A tým, že sa neukladá celá cesta, súkromie je lepšie ochránené. L-diverzita je v tomto konkrétnom prípade výhodná pre všetky strany. Možno ju použiť pri internom ukladaní údajov a pri ich externom zdieľaní.

6.4 Nástroje implementujúce anonymizačné a pseudonymizačné techniky

6.4.1 Knižnice dostupné v rôznych programovacích jazykoch

Techniky merania ochrany súkromia, spomínané v kapitole 6.3, sú implementované napríklad v Python knižnici pyCANON, popísanej v [9.]. Okrem k-anonymity a l-diverzity sa v tejto knižnici overuje anonymita datasetu aj pomocou ďalších techník: (α,k)-anonymity, ℓ-diverzity s entropiou, rekurzívnej (c,ℓ)-diverzity, t-blízkosti, základnej β-podobnosti, vylepšenej β-podobnosti a δ-súkromia v zdieľaní („disclosure privacy“).

Pre implementáciu hašovacích a šifrovacích algoritmov existuje veľa knižníc v rôznych programovacích jazykoch, napríklad:

- Blake 2 v node.js (Javascript)[47],
- SHA-3 v jazyku C[48],
- SHA-3 v node.js (Javascript) [49],
- AES v node.js (Javascript) [50].

Pri týchto knižniciach je však vždy dôležité mať overené, že implementácia je správna podľa štandardu a bezpečná. Napríklad NIST publikoval manuál na systém validácie pre SHA-3[51].

NIST tiež validuje implementáciu najrôznejších kryptografických algoritmov vrátane AES a SHA-3, ktoré sú implementované v tých najrôznejších IT systémoch IT dodávateľov[52].

6.4.2 Dostupné techniky v nástroji Talend

Nástroj Talend je momentálne základnou platformou pre Centrálnu integračnú platformu verejnej správy, pomocou ktorej si OVM vymieňajú údaje. V prvom rade nástroj Talend podporuje ochranu osobných údajov už tým, že umožňuje katalogizovať údaje, v rámci čoho je možné identifikovať citlivé datasety a dátové prvky.

Prostredníctvom nástroja Talend Data Quality možno maskovanie údajov a premiešavanie údajov vynútiť v ktoromkoľvek kroku dátovej pipeline (pozri Obrázok 19). Premiešavanie údajov je typ maskovania údajov, ktorý zahŕňa stĺpec (alebo zložitejší dataset ako napríklad skupinu stĺpcov), a ten náhodne premieša tak, aby sa skryla jeho identita, ale príslušné hodnoty zostali

[47] Zdroj: https://www.npmjs.com/package/blake2, Dátum referencie: 13.02.2023

[48] Zdroj: https://github.com/brainhub/SHA3IUF, Dátum referencie: 13.02.2023

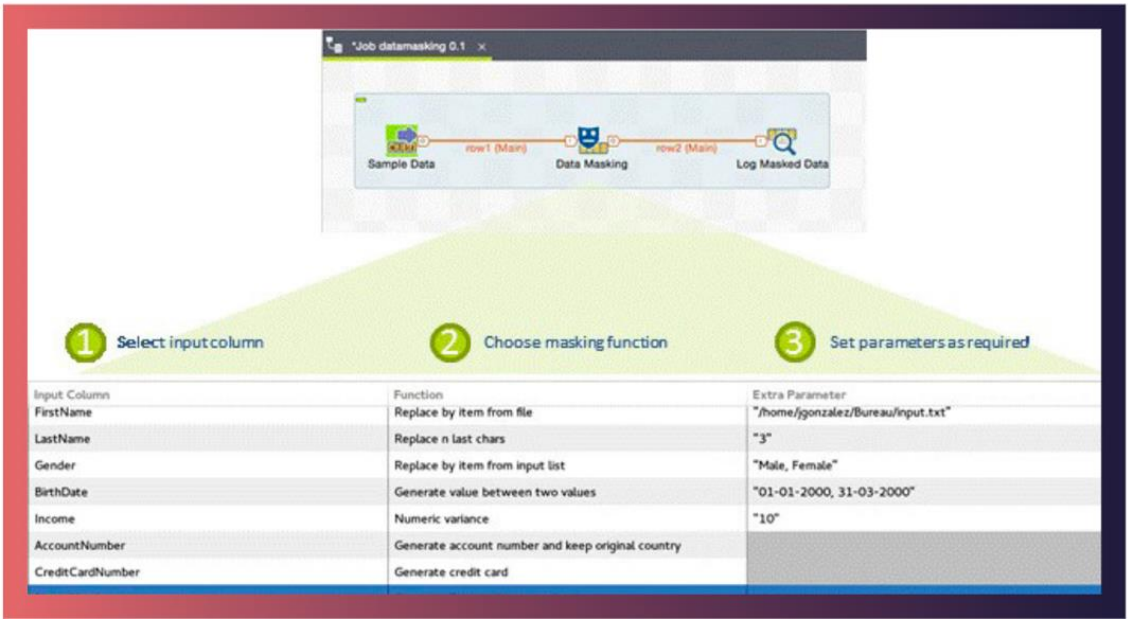
[49] Zdroj: https://www.npmjs.com/package/sha3, Dátum referencie: 13.02.2023

[50] Zdroj: https://www.npmjs.com/package/aes-js, Dátum referencie: 13.02.2023

[51] Zdroj: https://csrc.nist.gov/CSRC/media/Projects/Cryptographic-Algorithm-Validation-Program/documents/sha3/sha3vs.pdf, Dátum referencie: 13.02.2023

[52] Zdroj: https://csrc.nist.gov/projects/cryptographic-algorithm-validation-program/validation-search, Dátum referencie: 13.02.2023

na svojom mieste. Týmto spôsobom sa zachováva súkromie, ale analýzy a testovanie údajov sa môžu naďalej vykonávať s použitím pôvodných hodnôt údajov.



Obrázok 19: Maskovanie a premiešanie údajov možno použiť na dávkové spracovanie a toky údajov v reálnom čase prostredníctvom predkonfigurovaných alebo prispôsobených funkcií, ktoré sú vhodné pre bežné osobné údaje

Prostredníctvom nástroja Talend Data Preparation možno maskovanie údajov vynútiť aj ad-hoc spôsobom, čo umožňuje používateľom chrániť citlivé údaje pred zdieľaním s kolegami. Vezmite si príklad manažéra PR kampane, ktorý chce podať správu o úspešnosti kampane s partnerom. Môže zdieľať súbor údajov na analýzu po anonymizácii údajov, ktoré by mohli nevhodne odhaliť informácie súvisiace so súkromím.

6.4.3 Prehľad nástrojov pre Konsolidovanú analytickú vrstvu

Projekt „Konsolidovaná analytická vrstva - využitie dát pre zlepšenie fungovania inštitúcií verejnej správy (ďalej len KAV)“ prináša nové spôsoby využitia údajov, ktoré má Slovenská republika k dispozícii. Pomôže modernému riadeniu štátu, ktoré si vyžaduje výrazné zlepšenie využívania a spracovania údajov na analytické účely inštitúciami verejnej správy. Množstvo analytikov naráža na častý problém, ktorým je získať potrebné údaje na plánované dátové analýzy. V prípade získania požadovaných údajov, trvá veľa času, kým získané údaje dostanú do požadovanej podoby. Následné čistenie, úprava, prípadne prepojenie rôznych databáz tiež veľmi náročné. Práve s týmito ťažkosťami analytikov má pomôcť KAV.

Moderné dátové stacky, ktoré možno implementovať vo verejných cloudoch, sú navrhované ako jednou z alternatív rozvoja Konsolidovanej analytickej vrstvy v dokumente 4.2.1 Koncept pre zavádzanie analytického spracovania údajov do praxe. Ďalšou výhodou takýchto dátových

stackov alebo dátových cloudov ako je napríklad Snowflake[53] je, že majú priamo implementované anonymizačné techniky, napríklad dynamické maskovanie údajov podľa kapitoly 6.2.1. Momentálne sa Konsolidovaná analytická vrstva uvažuje len pre analytikov, ale keďže pôjde o moderný dátový sklad, ktorý môže mať v budúcnosti rôzne integrácie na systémy v takmer reálnom čase, možno ju využívať aj na modernizáciu mnohých ďalších agend, ktoré potrebujú pracovať s konsolidovanými údajmi z viacerých rezortov.

Ak chceme použiť dynamické maskovanie údajov, ktoré poskytne napríklad tímu audítorov plný prístup na čítanie, tímu analytikov zaheslované údaje na štatistické účely a ostatným redigované údaje, dá sa jednoducho vytvoriť abstrakt zobrazenia, napríklad:

```
CREATE SECURE VIEW v_customers AS
SELECT id, (
  CASE
    WHEN CURRENT_ROLE ( ) IN ( 'ACCOUNTING' ) THEN name
    WHEN CURRENT_ROLE ( ) IN ( 'ANALYST' ) THEN sha2 ( name )
    ELSE '[REDACTED]'
  END
) AS name FROM customers;
```

Obrázok 20: Zadefinovanie pravidiel pre zobrazovanie v Snowflake

Zrušením prístupu k podkladovému datasetu („customers“) a udelením prístupu k zobrazeniu (v_customers) budú mať používatelia teraz zavedené dynamické maskovanie údajov podľa svojich rolí a môžu získavať údaje len na základe zavedených pravidiel:

```
USE ROLE ACCOUNTING;
SELECT * FROM v_customers;
// returns Ben, Karl
USE ROLE ANALYST;
SELECT * FROM v_customers;
// returns hashed values
USE ROLE OTHER;
SELECT * FROM v_customers;
// returns redacted values
```

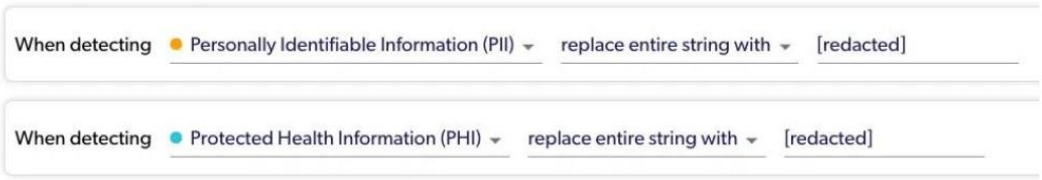
Obrázok 21: Dostupné zobrazenie údajov dynamickým maskovaním pre rôzne role používateľov

Dátový cloud Snowflake ako aj ďalšie alternatívy od Amazonu, Google či iné možno ďalej rozširovať o rôzne nástroje na bezpečnosť údajov, napríklad Satori[54]. Tieto nástroje umožňujú nastaviť kompletné maskovacie profily na jednom centrálnom mieste a rôzne ich používať v globálnom kontexte. Profily môžu byť tak všeobecné ako "maskujte všetko, čo je kategória

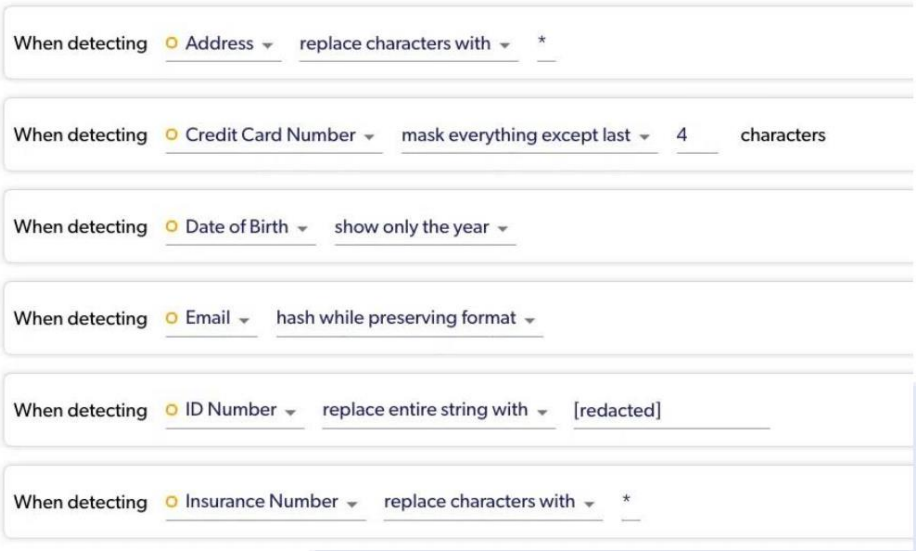
[53] Zdroj: https://www.snowflake.com/en/, Dátum referencie: 08.02.2023

[54] Zdroj: https://satoricyber.com/docs/getting-started/, Dátum referencie: 08.02.2023

PII", alebo tak podrobné ako nastavenie špecifickej akcie pre každý typ údajov (Obrázok 22 a Obrázok 23).



Obrázok 22: Ukážka nastavovania všeobecných maskovacích profilov



Obrázok 23: Ukážka nastavovania podrobných maskovacích profilov

K dispozícii sú ale aj rôzne open source nástroje, ako napríklad ARX[55], ktorý anonymizuje citlivé osobné údaje. Podporuje širokú škálu modelov ochrany súkromia a rizík, techník na transformáciu údajov a techník na analýzu užitočnosti výstupných údajov. Tento nástroj sa používa v rôznych kontextoch vrátane výskumných projektov, komerčných platforiem na analýzu veľkých objemov údajov a zdieľania informácií o klinických skúškach. ARX dokáže spracovať veľké súbory údajov na komoditnom hardvéri a má intuitívne multiplatformné grafické používateľské rozhranie. Je kompatibilný s SQL databázami a súbormi vo formáte Microsoft Excel a CSV.

[55] Zdroj: https://arx.deidentifier.org , Dátum referencie: 13.02.2023

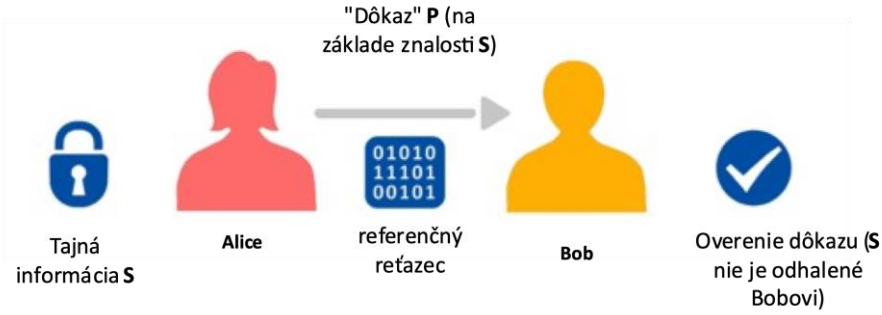
6.5 Nové trendy v anonymizácii a pseudonymizácii

6.5.1 Zero-knowledge proof

Známym kryptografickým primitívom je takzvaný Zero-Knowledge Proof (ZKP), čo je vlastne v typickom scenári termín opisujúci akýkoľvek protokol, ktorý dosahuje nasledovné: overovateľ („prover“) je schopný dokázať inej strane (verifikátor), že vlastní tajomstvo bez toho, aby prezradil akúkoľvek informáciu o samotnom tajomstve. ZKP bolo prvýkrát zavedené na overovanie totožnosti (Feige, Fiat a Shamir, 1988) tým, že sa poskytujú prostriedky na preukázanie totožnosti bez odhalenia autentifikačných informácií (ale dokazujú len to, že správnu autentifikačnú informáciu má prover). Všeobecnejšie povedané, dôkazy s nulovou znalosťou zahŕňajú dokazovanie, že výrok je pravdivý, bez odhalenia podrobností o výroku. ZKP by mal spĺňať tieto vlastnosti:

- **Úplnosť**: V prípade, že je výrok správny, poctivý overovateľ („prover“) presvedčí poctivého verifikátora, že fakt zodpovedajúci výroku je správny.
- **Správnosť**: V prípade, že výrok je nepravdivý, útočný overovateľ nemôže presvedčiť poctivého overovateľa, že výrok je správny, s výnimkou prípadov, keď je pravdepodobnosť zanedbateľná.
- **Nulová znalosť**: V prípade, že je výrok správny, overovateľ nezistí nič viac ako to, že výrok je správny.

Na dosiahnutie tohto cieľa sa zavádza model spoločného referenčného reťazca, čo znamená, že referenčný reťazec zdieľaný medzi overovateľom a verifikátorom by mal byť bezpečne vytvorený, pretože prístup k nemu by mali mať len overovateľ a verifikátor (Obrázok 24, kde Alice a Bob sú prover, resp. verifikátor).



Obrázok 24: Dôkaz nulovej znalosti pre pseudonymizáciu [11.]

V kontexte pseudonymizácie, ak osoba spojená s pseudonymom potrebuje preukázať, že je vlastníkom tohto pseudonymu bez toho, aby odhalila svoju presnú identitu, môže byť riešením ZKP. Ako konkrétny príklad takéhoto scenára uvádzame použitie ZKP na anonymné transakcie v kryptomenách. V týchto prípadoch sa používajú dôkazy nulovej znalosti, ktoré umožňujú overenie transakcií bez toho, aby overovatelia („miners“) vedeli čokoľvek o obsahu transakcií (a týmito prostriedkami sú odosielatelia a príjemcovia transakcií utajení). Tak je to napr. v

systéme Zcash[56], v ktorom odosielateľ transakcie (ktorý je tienený) zostavuje dôkaz, ktorý má preukázať, že:

1. vstupné hodnoty sa pri každom tienenom prevode rovnajú výstupným hodnotám,
2. má príslušné súkromné kľúče, ktoré mu dávajú oprávnenie uskutočniť transakciu,
3. súkromné kľúče na uskutočnenie transakcie sú kryptograficky spojené s podpisom nad celou transakciou takým spôsobom, že transakciu nemôže modifikovať strana, ktorá tieto súkromné kľúče nepoznala.

6.5.2 Atribútové poverenia („Attribute-based credentials (ABC)“)

Atribútové poverenia sú formou autentifikačného mechanizmu, ktorý umožňuje flexibilne a selektívne autentifikovať rôzne atribúty o entite bez toho, aby sa odhalili ďalšie informácie o entite (majú vlastnosť nulovej znalosti ako bolo uvedené aj v kapitole 6.5.1).

Overovanie atribútov si klasicky vyžaduje úplné a jedinečné overenie entity. Napríklad atribúty (ako je vek) by sa mohli vložiť do certifikátu spolu s menom používateľa, e-mailovou adresou, verejným kľúčom a ďalšími údajmi o danej entite. Na potvrdenie atribútu (napríklad, že používateľ je dospelý) je potrebné predložiť certifikát a odhaliť všetky informácie. Toto riešenie sa nepovažuje za riešenie zachovávajúce súkromie. Existuje viacero schém na realizáciu ABC a k dispozícii sú aj implementácie. Všetky zvyčajne zahŕňajú riadiaci subjekt, ktorý oprávňuje vydavateľov vydávať poverenia subjektom, ktoré by potom mohli pôsobiť ako overovatelia určitých skutočností o povereniach smerom k overovateľom. Cieľom je umožniť používateľovi selektívne preukázať špecifické atribúty, ako napríklad vek > 18 rokov, overujúcemu subjektu bez toho, aby odhalil akékoľvek ďalšie informácie.

Napríklad treba vydať taký občiansky preukaz v mobile, ktorý obsahuje dátum narodenia používateľa *DN* a môže sa použiť na preukázanie, že držiteľ karty má dostačujúci vek na to, aby mohol v kine sledovať filmy s vekovým obmedzením. V závislosti od hodnotenia filmu (minimálny vek *x*) môže držiteľ karty cez overovateľa dokázať, že:

"dnes - *DN* > *x*"

Viacnásobné použitie karty v tom istom kine by nemalo byť možné zaznamenávať.

Systémy ABC si vyžadujú značný výpočtový výkon alebo zásadnú optimalizáciu, takže ich implementácia nemusí byť jednoduchá. Medzi najobľúbenejšie implementácie patria:

- IDEMIX spoločnosti IBM vyvinutý v rámci projektu PRIME/PRIMELIFE[57].
- U-Prove spoločnosti Microsoft[58].
- Projekt IRMA univerzity Radboud v Nijmegene[59].

[56] Zdroj: https://z.cash/technology/zksnarks/ , Dátum referencie: 23.02.2023

[57] Zdroj: https://github.com/IBM/idemix, Dátum referencie: 13.02.2023

[58] Zdroj: https://www.microsoft.com/en-us/research/project/u-prove/, Dátum referencie: 13.02.2023

[59] Zdroj: https://irma.app/?lang=en, Dátum referencie: 13.02.2023

6.5.3 Diferenciálne súkromie („Differential privacy“)

Diferenciálne súkromie predstavuje prísnu matematickú definíciu súkromia. V najjednoduchšom prípade uvažujme algoritmus, ktorý analyzuje dataset a vypočíta o ňom štatistické údaje (napríklad priemer, rozptyl, medián, modus atď.). O takomto algoritme sa hovorí, že je diferenciálne súkromný, ak sa pri pohľade na výstup nedá zistiť, či údaje nejakého jednotlivca boli zahrnuté do pôvodného datasetu alebo nie. Inými slovami, zárukou diferenciálne súkromného algoritmu je, že jeho správanie sa takmer nezmení, keď sa k datasetu pripojí alebo ho opustí jeden jednotlivec. To znamená, že čokoľvek, čo by algoritmus mohol vyprodukovať v databáze obsahujúcej informácie o nejakom jednotlivcovi, takmer rovnako pravdepodobne pochádza z databázy bez informácií o tomto jednotlivcovi. Najpozoruhodnejšie je, že táto záruka platí pre akéhokoľvek jednotlivca a akýkoľvek dataset. Preto záruka diferenciálneho súkromia stále platí bez ohľadu na to, aké jedinečné sú údaje o jednotlivcovi, a bez ohľadu na údaje o komkoľvek inom v databáze,. To dáva formálnu záruku, že informácie o jednotlivcoch v databáze na individuálnej úrovni neuniknú.

Kľúčovou črtou diferenciálneho súkromia je, že nedefinuje súkromie ako binárny fakt, či „boli údaje jednotlivca vystavené tretej strane alebo nie“, ale skôr ako záležitosť kumulatívneho rizika. To znamená, že pri každom spracovaní údajov osoby sa zvyšuje riziko jej odhalenia. Na tento účel je definícia diferenciálneho súkromia pokrytá parametrami ("epsilon a delta"), ktoré kvantifikujú "stratu súkromia" - dodatočné riziko pre jednotlivca, ktoré vyplýva z použitia jeho údajov. Bez ohľadu na akékoľvek pomocné znalosti použité pri útoku na spätnú identifikáciu (kapitola 8.1.2) bude riziko ohrozenia súkromia spôsobené diferenciálne súkromným algoritmom navždy ohraničené touto stratou súkromia.

Tento prístup k anonymizácii a analýze miery ochrany súkromia nabral na dôležitosti práve kvôli rozsiahlemu využívaniu dolovania veľkých údajov a využívaniu systémov umelej inteligencie. Len donedávna bol aj predmetom aktívneho výskumu na popredných svetových univerzitách ako Harvard[60].

V praxi diferenciálne súkromie zahŕňa zanesenie malého množstva šumu do údajov pred ich vložením do systému umelej inteligencie, čím sa sťaží extrakcia pôvodných údajov zo systému. Niekto, kto vidí predpovede diferenciálne súkromného systému umelej inteligencie, nemôže zistiť, či boli na vývoj systému použité informácie konkrétnej osoby.

Knižnica pre diferenciálne súkromie od spoločnosti Google[61] ponúka sadu stavebných blokov, ktoré umožňujú vývojárom vytvárať rôzne súkromné aplikácie v jazykoch Go, Java a C++, Java a Go. Poskytuje vývojárom jednoduchý prístup k metrikám súvisiacim s tým, ako úspešne ich aplikácie zapájajú svojich používateľov, ako sú napríklad denní aktívni používatelia a príjmy na jedného aktívneho používateľa, spôsobom, ktorý pomáha zabezpečiť, aby nebolo možné identifikovať jednotlivých používateľov alebo ich spätne identifikovať. Spoločnosť Google túto knižnicu ďalej rozširuje o programovací jazyk Python v spolupráci s OpenMined, komunitou open source zameranou na technológie na ochranu súkromia. Spoločnosť tiež vydala nový nástroj na diferenciálne súkromie, ktorý umožňuje odborníkom z praxe vizualizovať a lepšie

[60] Zdroj: https://privacytools.seas.harvard.edu/differential-privacy, Dátum referencie: 13.02.2023

[61] Zdroj: https://github.com/google/differential-privacy, Dátum referencie: 23.02.2023

vyladiť parametre použité na vytvorenie diferenciálne súkromných údajov, ako aj manuál, v ktorom sa uvádzajú techniky na škálovanie diferenciálneho súkromia na veľké súbory údajov.

7 Pseudonymizačné politiky

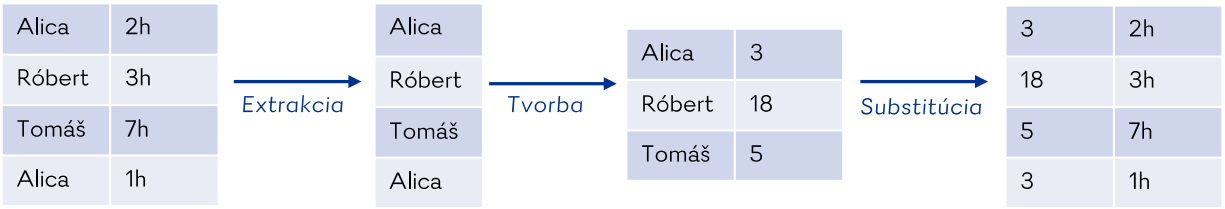
Zatiaľ čo výber techniky pseudonymizácie je zásadný, politika implementácie pseudonymizácie je rovnako dôležitá pre jej praktické použitie. Politiky pseudonymizácie predstavujú rôzne prístupy k nahrádzaniu skutočných údajov inými údajmi v databázach systémov, v dokumentoch a datasetoch. Každá politika môže mať vplyv na jednoduchosť implementácie a prísnosť ochrany údajov.

Táto kapitola sa zaoberá všeobecnejším problémom pseudonymizácie databáz systémov alebo akéhokoľvek dokumentu či datasetu, ktorý obsahuje identifikátory. Uvažujme identifikátor Id, ktorý sa niekoľkokrát vyskytuje v dvoch datasetoch A a B. Po pseudonymizácii sa identifikátor Id nahradí s ohľadom na vybranú politiku pseudonymizácie.

7.1 Deterministická pseudonymizácia

Táto politika vyžaduje nahradenie pôvodných údajov identickou náhradou vo všetkých databázach a vždy, keď sa objavia. Tým sa zabezpečí konzistentnosť zámeny v rámci databázy a medzi viacerými databázami. Pri implementácii tejto politiky je potrebné najprv z databázy vyextrahovať zoznam jedinečných identifikátorov. Potom sa musí tento zoznam namapovať na substitúcie a nakoniec treba nahradiť pôvodné údaje v databáze.

To v praxi znamená, že vo všetkých databázach a pri každom výskyte je identifátor Id vždy nahradený rovnakým pseudonymom $pseudo$. Je konzistentný v rámci databázy a medzi rôznymi databázami. Prvým krokom pri implementácii tejto politiky je extrahovanie zoznamu jedinečných identifikátorov obsiahnutých v databáze. Potom sa tento zoznam namapuje na pseudonymy a nakoniec sa identifikátory nahradia pseudonymami v databáze (pozri Obrázok 25).

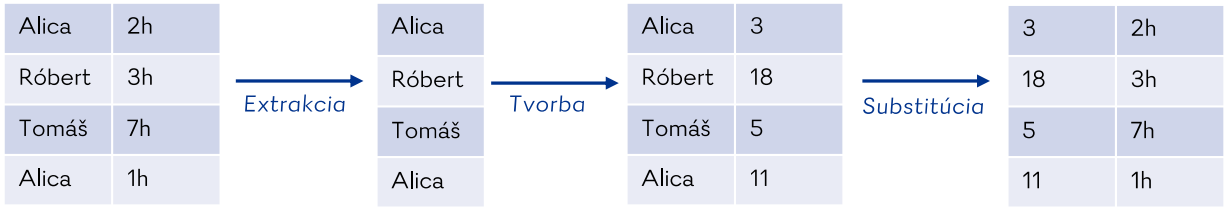


Obrázok 25: Fungovanie deterministickej pseudonymizácie

Všetky techniky uvedené v kapitolách 6.1.1 až 6.1.6 možno priamo použiť na implementáciu deterministickej pseudonymizácie. **Táto politika zachováva plnú mieru užitočnosti údajov tým, že umožňuje prepojenie medzi dotknutými osobami aj v rámci datasetu alebo databázy ako aj medzi datasetmi a databázami. Z tohto dôvodu ide ale aj o politiku s najnižšou úrovňou ochrany pri pseudonymizácii.**

7.2 Pseudonymizácia randomizovaná v rámci dokumentu ("document-randomized pseudonymisation")

Zakaždým, keď sa identifikátor Id objaví v databáze, datasete alebo v dokumente, nahradí sa iným náhodným pseudonymom ($pseudo_1$, $pseudo_2$,...). Id sa však vždy mapuje na ten istý súbor pseudonymov ($pseudo1$, $pseudo2$) v datasetoch A a B.



Obrázok 26: Fungovanie pseudonymizácie randomizovanej v rámci dokumentu

Pseudonymizácia je v tomto prípade konzistentná len medzi rôznymi databázami. Mapovacia tabuľka sa vytvára s použitím všetkých identifikátorov obsiahnutých v databáze. S každým výskytom daného identifikátora (napríklad Alica na obrázku - Obrázok 26) sa zaobchádza samostatne. **Táto politika zachováva istú mieru užitočnosti údajov tým, že umožňuje prepojenie medzi dotknutými osobami naprieč rôznymi databázami či datasetmi. V rámci databázy, dokumentu či datasetu však poskytuje úplnú ochranu pred prepojením.**

7.3 Plne randomizovaná pseudonymizácia ("fully randomized pseudonymisation")

Pri každom výskyte Id v rámci databázy A alebo B sa Id nahradí iným náhodným pseudonymom ($pseudo_1$, $pseudo_2$, ...). V tomto prípade ide o plne náhodnú pseudonymizáciu. Túto politiku možno považovať za ďalšie rozšírenie pseudonymizácie randomizovanej v rámci dokumentu. V skutočnosti majú tieto dve politiky rovnaké správanie, keď sa uplatňujú na jeden dokument - dataset - databázu. Ak sa však ten istý dokument pseudonymizuje dvakrát pomocou plne náhodnej pseudonymizácie, získajú sa dva rôzne výstupy. Pri pseudonymizácii randomizovanej v rámci dokumentu by sa dvakrát získal rovnaký výstup. Inými slovami, pri pseudonymizácii randomizovanej v rámci dokumentu je náhodnosť selektívna (napr. len pre Alicu), zatiaľ čo pri plne randomizovanej pseudonymizácii je náhodnosť globálna (platí pre akýkoľvek záznam). **Plne randomizovaná pseudonymizácia ponúka najvyššiu úroveň ochrany, ale znemožňuje akékoľvek porovnávanie a prepájanie tých istých dotknutých osôb medzi databázami – datasetmi - dokumentmi.**

8 Techniky útokov na pseudonymizáciu

8.1 Ciele útoku na pseudonymizáciu

V závislosti od kontextu a techniky pseudonymizácie môže mať útočník rôzne ciele, ktoré chce dosiahnuť voči pseudonymizovaným údajom:

- získanie pseudonymizačného tajomstva („secret“),
- úplná spätná identifikácia dotknutej osoby,
- čiastočné rozoznanie dotknutej osoby či skupiny („discrimination“).

Hoci väčšina príkladov opísaných v nasledujúcich kapitolách sa zameriava na odhalenie „skutočnej“ identity dotknutých osôb, treba poznamenať, že úspešný útok nie je (len) otázkou spätnej identifikácie, ale skôr schopnosti vyčleniť konkrétneho jednotlivca zo skupiny (aj keď „skutočná“ identita nie je odhalená).

8.1.1 Získanie pseudonymizačného tajomstva („secret“)

V tomto prípade sa útočník zameriava na odhalenie pseudonymizačného tajomstva (v prípade, že sa pseudonymizačné tajomstvo používa). Tento útok je najzávažnejší, pretože s použitím pseudonymizačného tajomstva je útočník schopný spätne identifikovať akýkoľvek pseudonym v datasete (úplná spätná identifikácia alebo diskriminácia), ako aj vykonávať ďalšie procesy pseudonymizácie datasete.

8.1.2 Úplná spätná identifikácia

Ak je cieľom útoku úplná spätná identifikácia, útočník chce dosiahnuť spätné prepojenie jedného alebo viacerých pseudonymov s identitou ich držiteľov. Najzávažnejší útok úplnej spätnej identifikácie spočíva v spätnej identifikácii všetkých pseudonymov. Útočník môže na dosiahnutie tohto cieľa použiť dve stratégie:

1. obnovenie každého identifikátora z príslušného pseudonymu samostatne,
2. alebo obnovenie pseudonymizačného tajomstva (kapitola 8.1.1).

Najmenej závažná forma útokov na úplnú spätnú identifikáciu zahŕňa útočníka, ktorý môže spätne identifikovať len podmnožinu pseudonymov v datasete. Uvažujme napríklad pseudonymizovaný dataset o známkach študentov univerzitného štúdia. Každá položka súboru údajov obsahuje pseudonym zodpovedajúci identite študenta (meno a priezvisko) a druhý pseudonym na pohlavie študenta (napr. priradením študentov mužského pohlavia k nepárnym číslam a študentiek k párnym číslam). Útočník uspeje v úplnom útoku na spätnú identifikáciu, ak získa meno, priezvisko a pohlavie študenta.

8.1.3 Čiastočné rozoznanie dotknutej osoby či skupiny („discrimination“)

Cieľom diskriminačného útoku je identifikovať vlastnosti držiteľa pseudonymu (aspoň jedného). Tieto vlastnosti nemusia priamo viesť k odhaleniu identity držiteľa pseudonymu, ale môžu postačovať na jeho rozoznanie určitým spôsobom.

Ak vezmeme do úvahy príklad známok študentov, ktorý bol uvedený v kapitole 8.1.2, dataset o známkach študentov môže obsahovať dve párne čísla medzi mnohými nepárnymi číslami ako pseudonymy. Párne čísla zodpovedajú študentkám, zatiaľ čo nepárne čísla zodpovedajú študentom (táto skutočnosť je útočníkovi známa). Obe párne čísla dosiahli na záverečnej skúške výsledok 100 %. Ďalej predpokladajme, že v pseudonymizovanom datasete nie sú žiadni ďalší študenti, ktorí by dosiahli 100 %. Ak útočník získa dodatočnú informáciu, že určitý študent dosiahol v tomto predmete 100 %, okamžite sa dozvie, že tento študent bol ženského pohlavia. A naopak, ak sa útočník dozvie, že študent tohto predmetu bola žena, útočník sa okamžite dozvie, že tento študent získal 100 %. Je dôležité si uvedomiť, že útočník sa tu nedozvie identitu držiteľa pseudonymu, ale dozvie sa len nejakú vlastnosť (t. j. pohlavie alebo hodnotu známky) držiteľa. Vzhľadom na to, že niekoľko študentov má rovnakú kombináciu hodnôt vlastností, útočník nie je schopný presne určiť záznam údajov konkrétneho držiteľa pseudonymu. Tieto získané dodatočné informácie však už môžu stačiť na účely istej formy rozoznania, ktorú má útočník v úmysle vykonať, alebo môže byť využitá pri následnom útoku na základe znalostí o pozadí na odhalenie identity, ktorá sa skrýva za pseudonymom.

8.2 Riziká spojené s anonymizáciou a pseudonymizáciou

Aby útočníci nedokázali naplniť ciele uvedené v kapitole 8.1, musia použité techniky pseudonymizácie a anonymizácie adresovať tieto tri riziká:

1. **Vyčlenenie** („singling out“): ide o možnosť izolovať niektoré alebo všetky záznamy, ktoré identifikujú jednotlivca v datasete.
2. **Možnosť prepojenia** („linkability“): ide o možnosť prepojenia viacerých záznamov týkajúcich sa tej istej dotknutej osoby alebo skupiny dotknutých osôb (buď v tom istom datasete, alebo vo viacerých rôznych datasetoch). Spočíva v kombinácii najmenej dvoch anonymizovaných datasetov s cieľom odhaliť identitu niektorých osôb prítomných v oboch. Ak útočník môže zistiť (napr. pomocou korelačnej analýzy), že dva záznamy sú priradené k tej istej skupine osôb, ale nemôže vyčleniť osoby v tejto skupine, zvolená technika poskytuje odolnosť proti vyčleneniu, ale nie proti prepojiteľnosti.
3. **Inferencia** („inference“): ide o možnosť odvodiť so značnou pravdepodobnosťou správnosti hodnotu atribútu z hodnôt súboru iných atribútov. Používajú sa pri tom techniky dolovania údajov za účelom extrakcie informácií z údajov.
4. **Homogenita** („homogeneity“): môže nastať, keď sú všetky hodnoty citlivého atribútu v triede ekvivalencie rovnaké.
5. **Znalosť pozadia** („background“): v tomto prípade má útočník určité predbežné znalosti o cieli útoku (napríklad pozná niektoré pomocné informácie o jednotlivcovi v databáze).
6. **Skreslenosť** („skewness“): môže nastať vtedy, keď v celej databáze existuje zriedkavá hodnota citlivého atribútu, ktorá je mimoriadne častá v triede ekvivalencie.
7. **Podobnosť** („similarity“): môže sa vyskytnúť, keď sú hodnoty citlivého atribútu v triede ekvivalencie sémanticky podobné (hoci sa líšia).

Tabuľka 17 popisuje, akými technikami z kapitoly 6.3 a zdroja [9.] možno vyššie uvedené riziká eliminovať.

Tabuľka 17: Techniky merania vplyvu techník na ochranu súkromia a riziko, ktoré eliminujú

Technika (kapitola 6.3 a [9.])	Hlavné riziko, ktoré eliminuje (zo zoznamu nad tabuľkou)						
	1.	2.	3.	4.	5.	6.	7.
k-anonymita	✓	✓					
ℓ -diverzita				✓	✓		
(α,k)-anonymita	✓	✓		✓			
ℓ-diverzita s entropiou				✓	✓		
rekurzívna (c,ℓ)-diverzita				✓	✓		
t-blízkosť						✓	✓
základná β-podobnosť						✓	
vylepšená β-podobnosť						✓	
δ-súkromie v zdieľaní			✓			✓	

8.3 Hlavné útočné techniky

Existujú tri hlavné všeobecné techniky na prelomenie pseudonymizačnej funkcie:

1. útoky pomocou hrubej sily („brute force attack“ alebo aj vyčerpávajúce vyhľadávanie),
2. vyhľadávanie v slovníku („Dictionary Search“)
3. a hádanie („Guesswork“).

Okrem týchto útočných techník možno pri útoku využiť aj existenciu rizík, spomenutých v kapitole 8.2, vtedy možno typické útoky označiť ako:

1. Útok na vyčlenenie („singling out“),
2. Útok cez prepájanie („linkage attack“),
3. Útok cez inferencie („inference attack“),
4. Útok cez znalosť pozadia („background“) (napríklad štatistický distribučný útok).

Účinnosť týchto útokov závisí od niekoľkých parametrov vrátane:

– Množstvo informácií o držiteľovi pseudonymu (dotknutej osoby) obsiahnutých v pseudonyme.

- Poznatky o pozadí, ktoré útočník má k dispozícii.
- Veľkosť domény identifikátora.
- Veľkosť domény pseudonymu.
- Výber a konfigurácia použitej funkcie pseudonymizácie (to zahŕňa aj veľkosť pseudonymizačného tajomstva).

Tieto techniky útoku sú stručne opísané v ďalších kapitolách.

8.3.1 Útok pomocou hrubej sily („brute force attack“)

Praktickosť tejto techniky útoku je podmienená schopnosťou útočníka vypočítať pseudonymizačnú funkciu (to znamená, že neexistuje pseudonymizačné tajomstvo) alebo jeho prístupom k implementácii pseudonymizačnej funkcie, ktorá predstavuje „čiernu skrinku“. V závislosti od cieľa útoku sa môžu uplatniť ďalšie podmienky. Ak sa útok pomocou hrubej sily používa na dosiahnutie úplnej spätnej identifikácie (t. j. obnovenie pôvodnej identity), doména identifikátora musí byť konečná a relatívne malá. Pre každý pseudonym, na ktorý útočník narazí, sa môže pokúsiť obnoviť pôvodný identifikátor tak, že použije pseudonymizačnú funkciu na každú hodnotu domény identifikátorov, kým nenájde zhodu.

Tabuľka 18: Pseudonymizácia mesiaca narodenia

Mesiac narodenia	Pseudonym	Mesiac narodenia	Pseudonym
Január	281	Júl	299
Február	269	August	285
Marec	288	September	296
Apríl	291	Október	294
Máj	295	November	307
Jún	301	December	268

Uvažujme o pseudonymizácii mesiaca narodenia v datasete. Veľkosť oblasti identifikátorov je 12, takže útočník môže rýchlo vymenovať všetky možnosti. Pseudonymy priradené ku každému mesiacu sa v tomto prípade vypočítajú ako súčet ASCII kódu prvých troch písmen mesiaca narodenia (pričom prvé je veľké písmeno). Uvažujme, že útočník narazil na pseudonym 301. Na každý mesiac narodenia môže použiť funkciu pseudonymizácie, kým nenájde mesiac, ktorý zodpovedá hodnote 301. V tabuľke - Tabuľka 18 sú uvedené výsledky výpočtov, ktoré vykoná útočník na spätnú identifikáciu pseudonymu 301, ktorých výsledkom je mapovacia tabuľka pseudonymizačnej funkcie.

Je zrejmé, že na úspešné uskutočnenie tohto útoku je rozhodujúca veľkosť domény identifikátorov. V prípade malých domén identifikátorov, ako je to v uvedenom príklade, je útok hrubou silou jednoducho uskutočniteľný. Ak je veľkosť domény identifikátorov nekonečná, útok hrubou silou sa zvyčajne stáva neuskutočniteľným. Ak je veľkosť domény identifikátorov príliš veľká, úplná spätná identifikácia je mimoriadne náročná, čo ponecháva útočníkom možnosť diskriminačného útoku.

V takom prípade môže útočník uvažovať o časti domény identifikátorov, pre ktorú môže vypočítať všetky pseudonymy. Vráťme sa k príkladu z tabuľky - Tabuľka 18, pričom predpokladajme, že doména je malá. Predpokladajme, že protivník chce odlíšiť ľudí s mesiacom narodenia začínajúcim na písmeno J od ľudí začínajúcich na iné písmeno. Táto subdoména obsahuje január, jún a júl. Protivník môže vykonať vyčerpávajúce vyhľadávanie v tejto subdoméne tak, že vypočíta pseudonymy zodpovedajúce januáru, júnu a júlu. Ak nájde pseudonym odlišný od 281, 301 a 299, potom vie, že mesiac narodenia nezačína písmenom J.

V prípade, že sa používa pseudonymizačné tajomstvo, ani malá doména identifikátorov nemusí umožniť uskutočnenie takéhoto útoku (keďže útočník nie je schopný vypočítať pseudonymizačnú funkciu a za predpokladu, že nemá prístup k implementácii tejto funkcie predstavujúcej „čiernu skrinku“). V takom prípade je možné vykonať útok hrubou silou na celý priestor pseudonymizačných tajomstiev - útočník totiž vyčerpávajúco skontroluje všetky možné tajomstvá a pre každé z nich vypočíta funkciu obnovy. Tento útok bude úspešný, ak útočník správne odhadne pseudonymizačné tajomstvo bez ohľadu na veľkosť domény identifikátorov. Preto, aby sa takýto útok zmaril, počet možných pseudonymizačných tajomstiev by mal byť dostatočne veľký, aby bol útok prakticky nemožný.

8.3.2 Vyhľadávanie v slovníku („Dictionary Search“)

Vyhľadávanie v slovníku je optimalizáciou útoku pomocou hrubej sily, ktorá môže ušetriť výpočtové náklady. Útočník sa musí vysporiadať s veľkým množstvom pseudonymov, aby vykonal úplnú spätnú identifikáciu alebo diskrimináciu. Preto vopred vypočíta (obrovský) súbor pseudonymov a výsledok uloží do slovníka. Každý záznam v slovníku obsahuje pseudonym a príslušný identifikátor alebo informáciu. Zakaždým, keď útočník potrebuje spätne identifikovať pseudonym, vyhľadá ho v slovníku. Toto vyhľadávanie má náklady na vyčerpávajúce vyhľadávanie a výsledok ukladá do veľkej pamäte. Spätná identifikácia pseudonymu má len náklady na vyhľadávanie v slovníku. Vyhľadávanie v slovníku je v podstate výpočet a uloženie mapovacej tabuľky. Robiť kompromisy medzi časovou náročnosťou a využívanou pamäťou možno pomocou Hellmanových tabuliek[62] alebo dúhových tabuliek („rainbow tables“)[63] na ďalšie rozšírenie rozsahu. Existujú však špecifické varianty tohto útoku, ktoré využívajú ďalšie znalosti o spôsobe fungovania pseudonymizačnej funkcie. Takéto útoky môžu fungovať aj pre nekonečné vstupné domény.

Tento typ útoku si trochu podrobnejšie ukážeme na prípade pseudonymizácie IP adries. Hlavnou charakteristikou problému pseudonymizácie IP adries je veľkosť vstupného priestoru (domény identifikátora): existuje len 4 294 967 296 možných IP adries verzie 4. Vďaka tomu má útočník k dispozícii vyčerpávajúce aj slovníkové vyhľadávanie, aby mohol uskutočniť útoky na spätnú identifikáciu alebo diskrimináciu, ak pseudonymizačná funkcia nie je správne zvolená.

Vzhľadom na vyššie uvedenú charakteristiku sú kryptografické hašovacie funkcie v tomto prípade použitia obzvlášť zraniteľné. Ako príklad uvažujme IP adresu pseudonymizovanú pomocou hašovacej funkcie SHA-256. Útočník s pseudonymom/digestom môže použiť existujúce nástroje (napríklad softvér na prelomenie hesla, ako je "John The Ripper" alebo iný) na vyčerpávajúce vyhľadávanie. V tabuľke (Tabuľka 19) je uvedené trvanie tohto vyhľadávania

[62] M. Hellman, “A cryptanalytic time-memory trade-off,” IEEE transactions on Information Theory, vol. 26, číslo 4, strany 401-406, 1980

[63] P. Oechslin, “Making a Faster Cryptanalytic Time-Memory Trade-off,” in CRYPTO 2003, 2003.

na jednom bežnom notebooku s procesorom Intel(R) Core(TM) i7-8650U @ 1,90 GHz (8 jadier) a veľkosť potrebného slovníka. Aj v najhoršom prípade trvá obnovenie IP adresy patriacej danému pseudonymu len približne 2 minúty.

Tabuľka 19: Praktická náročnosť útokov na pseudonymizáciu pomocou hašovacích funkcií v prípade IP adresy

Trieda IP adresy	Počet možných IP adries	Čas vyčerpávajúceho vyhľadávania	Veľkosť slovníka
145.254.160.X	256	200 ms	8 KB
145.254.X.X	65 536	200 ms	2 MB
145.X.X.X	16 777 216	2 s	512 MB
X.X.X.X	4 294 967 296	2 min. 16s	128 GB

Vyššie uvedený jednoduchý prípad ukazuje, že pseudonymizácia IP adries len pomocou kryptografických hašovacích funkcií zlyháva. Preto sa na ochranu údajov musia uprednostniť iné pseudonymizačné funkcie, ako napríklad kódy na overovanie správ, šifrovanie s tajným ad hoc generovaným kľúčom alebo generátory náhodných čísel. Vtedy útočník nemôže uskutočniť rovnaké útoky, pretože tieto metódy používajú tajný kľúč (MAC a šifrovanie) alebo zdroj náhodnosti (v prípade CPRNG).

8.3.3 Hádanie („Guesswork“)

Tento typ útoku využíva určité znalosti o pozadí (napríklad rozdelenie pravdepodobnosti alebo iné vedľajšie informácie), ktoré môže mať útočník o niektorých (alebo všetkých) držiteľoch pseudonymov, pseudonymizačnej funkcii alebo datasete. Vyčerpávajúce vyhľadávanie a vyhľadávanie v slovníku implicitne predpokladajú, že všetky identifikátory majú rovnakú pravdepodobnosť alebo frekvenciu výskytu. Niektoré identifikátory však môžu byť častejšie ako iné. Využívanie štatistických charakteristík identifikátorov je známe ako hádanie[64] [65] a je široko aplikované v komunite, ktorá sa zaoberá prelamovaním hesiel. Je dôležité si všimnúť, že hádanie sa dá použiť aj vtedy, keď je doména identifikátorov obrovská. Útočník nemusí mať nevyhnutne prístup k pseudonymizačnej funkcii (keďže diskriminácia je možná jednoduchým vykonaním frekvenčnej analýzy pozorovaných pseudonymov).

Uvažujme o prípade, ktorý sa týka pseudonymov zodpovedajúcich "krstným menám". Doménu „krstných mien“ je ťažké preskúmať v celej jej šírke. Útočník však vie, ktoré krstné mená sú najobľúbenejšie. Protivník môže vykonať vyčerpávajúce vyhľadávanie alebo slovníkové vyhľadávanie v oblasti najpopulárnejších krstných mien a dosiahnuť istú mieru rozoznania.

[64] Y. Yona and S. Diggavi, “The effect of bias on the guesswork of hash functions,” in 2017 IEEE International Symposium on Information Theory (ISIT), 2017.

[65] D. G. Malone and W. Sullivan, “Guesswork and entropy,” IEEE Transactions on Information Theory, vol. 50, číslo 3, strany 525--526, 2004

Predpokladajme podobný prípad, ale s nekonečnou veľkosťou domény identifikátorov. Možno definovať konečnú subdoménu identifikátorov, ktoré sú zahrnuté v datasete. Ak útočník dokáže odhadnúť túto subdoménu, môže uskutočniť vyčerpávajúce vyhľadávanie. V závislosti od množstva podkladových informácií alebo metadát, ktoré má útočník k dispozícii, a množstva informácií, ktoré sa dajú prepojiť a ktoré sa nachádzajú v pseudonymizovanom dataste, môže tento typ útoku viesť k odhaleniu identity jedného držiteľa pseudonymu, ale aj k časti identít alebo identít celého datasetu. Najmä v prípade malých datasetov môžu byť takéto útoky uskutočniteľné s vysokou mierou úspešnosti.

9 Súčasný stav anonymizácie a pseudonymizácie v rámci Dátového programu MIRRI

9.1 Zdrojové systémy

Zdrojové systémy by mali anonymizovať údaje pomocou techník opísaných v kapitole 6.1.9 a 6.3 pri zdieľaní údajov a datasetov s tretími stranami a pri publikovaní otvorených údajov. Pri mapovaní konkrétnych prípadov použitia pre jednotlivé scenáre v praxi sa môžu tieto štandardy spresňovať.

9.2 Centrálna integračná platforma

Centrálna integračná platforma je kľúčovým riešením pre efektívne a bezpečné zdieľanie údajov medzi OVM, a preto predstavuje jadro Dátového programu MIRRI. Z tohto dôvodu je dôležité implementovať pseudonymizačné politiky (kapitola 7) aj v rámci CIP. Bude v prvej fáze implementovať služby pre anonymizáciu s využitím techník opísaných v kapitole 6.2.1 a pseudonymizáciu s využitím techniky hašovania s kľúčom podľa kapitoly 6.1.4. Pripravuje sa DNR pre centrálnu pseudonymizačnú funkcionalitu, ktorej prvým používateľom by mala byť platforma KAV. Táto DNR musí nasledovať Prípad použitia 1: Štatistická analýza anonymizovaných a pseudonymizovaných údajov podľa kapitoly 10.3.1 a musí obsahovať minimálne:

1. Výber a odôvodnenie politiky podľa kapitoly 7,
2. Aká kombinácia metód pseudonymizácie a anonymizácie sa bude implementovať pre aké kategórie údajov podľa citlivosti,
3. Aké techniky na meraniu vplyvu techník na ochranu súkromia podľa kapitoly 6.3 budú implementované,
4. S akými rizikami podľa kapitoly 8.2 sa počíta a aké opatrenia budú implementované na ich minimalizáciu, aké možné útoky sa adresujú podľa kapitoly 8.3,
5. Či sú pre túto funkcionalitu relevantné aj niektoré ďalšie prípady použitia podľa kapitoly 10.1,
6. Musí prejsť a adresovať všetky kontrolné položky zoznamu štandardov anonymizácie a pseudonymizácie (Tabuľka 23).

Je nevyhnutné zdôrazniť, že pre platformu KAV nemusí byť dostačujúce využívať len takúto centrálnu pseudonymizačnú funkcionalitu, ale podľa vybraných technológií pre platformu bude potrebné nasadiť aj obdobné nástroje popísané v kapitole 6.4.3.

10 Scenáre a prípady použitia pre anonymizáciu a pseudonymizáciu a výber techník a politík

Daný scenár a konkrétny prípad použitia a ďalšie parametre ovplyvňujú výber techniky alebo politiky anonymizácie a pseudonymizácie v praxi.

10.1 Scenár ochrany údajov pri prenose

Ako bolo spomenuté v kapitole 3.1.2, čoraz viac údajov, ktoré organizácie používajú a spracúvajú, má tendenciu byť neštruktúrovaných. To predstavuje výzvu, pokiaľ ide o schopnosť identifikovať, zisťovať a chrániť údaje s vplyvom na súkromie. Preto je veľmi dôležité mať nastavenú jasnú ochranu súkromia pre údaje pri prenose. To znamená vtedy, kedy údaje opúšťajú pôvodné agendové systémy a predtým, ako sa dostanú do systémov alebo rúk príjemcov. Obdobne musí byť chránený aj opačný smer, kedy dotknuté osoby alebo spravodajské jednotky si plnia svoju povinnosť pri vypracovávaní štatistických výkazov a svoju spravodajskú povinnosť. V tejto oblasti dochádza k zaujímavému prekrývaniu bezpečnosti údajov a súkromia. Ak by niekto zachytil údaje počas ich prenosu medzi dvoma systémami, toto porušenie bezpečnosti by takmer nevyhnutne viedlo k narušeniu súkromia. Znížiť vplyv tohto narušenia súkromia možno použitím obfuskačných techník na údaje pred ich prenosom, popísaných v kapitole 6.1, ale treba tiež znížiť pravdepodobnosť zachytenia týchto údajov aj použitím špecifických kontrolných mechanizmov. Bezpečnostné stratégie treba navrhnúť v spolupráci s bezpečnostným tímom a sú podrobnejšie popísané v dokumente 1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov. Tu dávame k dispozícii základný kontrolný zoznam, ktorý je východiskovým bodom.

V prípade najcitlivejších údajov (úroveň „Vyhradené“ a „Dôverné“) zasielaných elektronickým prenosom by ste mali zvážiť nasledujúce skutočnosti:

- Údaje musia byť v ideálnom prípade predmetom „end-to-end“ šifrovania pri tranzite na servisnej vrstve so vzájomným TLS, ak je to technicky možné. Ak vzájomné TLS nie je možné, treba použiť bežné SSL/TLS, pretože nešifrované HTTP by mohlo predstavovať bezpečnostné riziko.
- Zabráňte distribúcii prostredníctvom e-mailu, pretože e-mail je zo svojej podstaty nezabezpečený. Vytvára sa tak viacero kópií údajov, čím sa zväčšuje priestor na útoky, a je oveľa ťažšie kontrolovať a sledovať, kde sa tieto kópie nachádzajú, a vytvárať kontroly prístupu k nim. Okrem toho aj skúsení IT špecialisti často používajú e-mail na odosielanie citlivých údajov, keď sa ponáhľajú, mysliac si, že to urobia "len raz". Časom sa z týchto inštinktov stanú návyky. Takto sa darí pokusom o phishing - externí útočníci môžu používať e-maily ako vstupný bod do agendových IT systémov a extrahovať citlivé údaje.
- Sledovanie a prispôsobenie kontroly súkromia na základe dvoch bodov prenosu pre údaje: napríklad pri spracovaní a analýze sa údaje môžu presúvať:
  - Z interného dátového centra do iného interného dátového centra (napríklad do Konsolidovanej analytickej vrstvy),
  - Z interného dátového centra do externého dátového centra,

- Z interného dátového centra do verejnej cloudovej inštancie, ktorú vlastníte (napríklad do Konsolidovanej analytickej vrstvy, za predpokladu, že minimálne jej časť bude aj vo verejnom cloude),
- Z verejnej cloudovej inštancie, ktorú vlastníte, do inej verejnej cloudovej inštancie, ktorú vlastníte (napríklad v rámci Konsolidovanej analytickej vrstvy, ak využíva „multi-vendor“ cloud).

Treba použiť techniky šifrovania prispôsobené každej z týchto ciest prenosu.

Predchádzajúci kontrolný zoznam je teda len východiskový a treba ho ďalej rozširovať a prispôsobovať. Mal by sa pritom využiť kontext z vykonanej klasifikácie údajov a inventarizácie na vytvorenie procesu, ktorý skutočne zachytáva existujúce rizikové prípady a potreby. Základom tohto procesu sú nasledujúce osvedčené postupy:

- Klasifikujte údaje pred prenosom, ako sme už uviedli.
- Spolupracujte so svojím bezpečnostným tímom, aby ste odhadli pravdepodobnosť zachytenia pri prenose údajov a pri ich ukladaní treťou stranou.
- Spolupracujte s právnym tímom, aby ste pochopili všetky zmluvné alebo právne požiadavky vyplývajúce z prenosu údajov, najmä ak je niektorý z prenosov údajov cezhraničnej povahy.
- Uistite sa, že ste pochopili, ako replikácia údajov ovplyvní prenos, pretože každá kópia údajov môže zvýšiť rizikové skóre.
- Zabezpečte, aby všetky produktové a technické tímy mali miesto pri stole, pretože budú mať väčšie povedomie o pohybe údajov ako tímy pre ochranu súkromia a bezpečnosť, ktorých úlohou je ich monitorovanie a vytváranie kontrolných mechanizmov na zmiernenie narušenia súkromia.
- Nezabudnite, že tento zoznam je len východiskový. Mali by ste využiť kontext z vašej klasifikácie údajov a inventarizácie na vytvorenie procesu, ktorý skutočne zachytáva vaše rizikové prípady a potreby.

Zdieľanie údajov často predstavuje zaujímavé dilemy. Môžu sa vyskytnúť prípady použitia, keď chcete niekoho interne identifikovať, ale chcete zdieľať jeho údaje tak, aby ho tretia strana alebo iný informačný systém, ktorý jeho údaje prijíma, nemohol identifikovať. Takéto prípady použitia si ukážeme na príklade nasledujúcej tabuľky (Tabuľka 20), ktorá prepojí externé identifikátory (napríklad číslo pasu) s vlastnými internými identifikátormi. Ak to urobíte, budete chcieť starostlivo spravovať prístup k tejto prepojovacej tabuľke, aby ste predišli problémom s narušením súkromia. Ak nebudete riadiť prístup k tejto tabuľke, interní a externí aktéri budú môcť jednoducho prepojiť externé údaje s internými údajmi prostredníctvom tejto tabuľky. Tabuľka internej aplikácie pre „Výjazdy“ uchováva údaje o cestách priradené k identifikačným číslam pasov cestujúcich.

Tabuľka 20: Údaje o cestách priradené k číslam pasov v systéme “Výjazdy”

Číslo pasu	Začiatok cesty	Koniec cesty
BA7678987	13.00 hod.	14.00 hod.
BI9892821	14.00 hod.	16.00 hod.
BH8753116	12.00 hod.	16.00 hod.

Číslo pasu	Začiatok cesty	Koniec cesty
BA3736111	11.00 hod.	11.30 hod.

Predpokladajme, že interný tím za aplikáciou „Výjazdy“ chce tieto údaje zdieľať s externým analytickým tímom, aby im pomohol zistiť, v ktorých denných hodinách je dopyt vyšší, vďaka čomu sa môžu prispôsobiť kapacity a trasy. Samozrejme by bolo riskantné až priam neprípustné zdieľať čísla pasov s týmto externým analytickým tímom, takže aby sme znížili riziko ohrozenia súkromia, mohli by sme najprv vytvoriť internú mapovaciu tabuľku - niečo ako Tabuľka 21.

Tabuľka 21: Čísla pasov priradené k interným identifikačným číslam

Číslo pasu	Interné ID
BA7678987	ghsvfydvbdv
BI9892821	hgavdchgdfe
BH8753116	dhbchchvhge
BA3736111	wdjhpdjdiehf

V tabuľke (Tabuľka 21) sme vytvorili interné identifikátory, ktoré sa vzťahujú na čísla pasov. Tieto interné identifikátory možno vytvoriť technikami opísanými v kapitolách 6.1.1, 6.1.2, 6.1.4, 6.1.5.1, 6.1.6, a 6.1.8. Ako to pomáha chrániť súkromie, je zrejmé z tabuľky nižšie (Tabuľka 22).

Tabuľka 22: Spoločné údaje s internými identifikátormi

Interné ID	Začiatok cesty	Koniec cesty
ghsvfydvbdv	13.00 hod.	14.00 hod.
hgavdchgdfe	14.00 hod.	16.00 hod.
dhbchchvhge	12.00 hod.	16.00 hod.
wdjhpdjdiehf	11.00 hod.	11.30 hod.

Ak by interný tím poskytol nespracované a osobne identifikovateľné údaje (Tabuľka 20 s číslami pasov), predstavovalo by to vysoké riziko ohrozenia súkromia. Zdieľanie tých istých údajov s internými identifikačnými číslami priradenými k údajom o cestách (Tabuľka 22) namiesto čísel pasov však umožňuje požadovaný druh analýzy a plánovania bez toho, aby sa presne odhalilo, kto tieto cesty absolvoval. Týmto spôsobom si interný tím môže ponechať údaje na audit a spätné cielenie jednotlivcov zúčastnených na cestách bez toho, aby tieto údaje zdieľala s dodávateľom, ktorého analýza tieto údaje nevyžaduje.

Je to mimoriadne dôležité z niekoľkých dôvodov:

- Vždy, keď sa údaje zdieľajú, je potrebné ich chrániť pomocou šifrovania alebo iných prostriedkov kontroly prístupu počas ich prenosu, aby sa znížilo riziko odchytenia údajov.
- Keď sa údaje dostanú k externému partnerovi, ste zraniteľní voči akýmkoľvek bezpečnostným chybám, ktoré sa uňho odohrajú. Zníženie identifikovateľnosti používateľov, o ktorých údaje ide, môže pomôcť riadiť riziko ohrozenia súkromia.

- Napokon, pri každom zdieľaní údajov v podstate vytvárate kópiu údajov. **Čím viac kópií osobných údajov, tým viac prostriedkov musíte vynaložiť na ich ochranu.** Ako vám povie každý skúsený profesionál v oblasti ochrany osobných údajov, najlepšou ochranou údajov je tieto údaje vôbec nemať. Vytváranie nepotrebných kópií údajov nie je strategické, najmä keď mapovacie tabuľky, ako bolo ukázané, umožňujú vykonať analýzu bez zdieľania údajov umožňujúcich identifikáciu osôb.

Avšak ani takéto zdieľanie údajov s tretími stranami nie je bez rizika. Externé zverejnenie interných identifikátorov používateľov by mohlo spôsobiť riziko identifikácie, aj keď tieto identifikátory nie sú spojené s údajmi, ktoré osobne identifikujú používateľa. Keďže niektoré z týchto identifikátorov sú dlhodobé a počas svojej životnosti sa nikdy nemenia, ich sprístupnenie externým stranám by mohlo umožniť sledovanie tých istých používateľov vo viacerých súboroch údajov (buď tej istej strane, alebo, čo je horšie, viacerým stranám, ak sa tieto súbory údajov zdieľajú alebo uniknú) počas dlhšieho časového obdobia.

Vo všeobecnosti platí zásada, že pri zdieľaní s externými stranami by sa v súbore údajov nikdy nemali odhaľovať interné identifikátory - je potrebné ich pseudonymizovať takým spôsobom, aby sa pôvodný identifikátor nedal zrekonštruovať z pseudonymizovaného identifikátora, pričom sa zachová konzistentnosť medzi hodnotami súboru údajov. Okrem toho, ak sú zaheslované interné identifikátory zverejnené externe a následne sú predmetom prelomenia, zaheslovanie pred zdieľaním môže umožniť určiť, ktorá externá strana sa na porušení podieľala. Stručne povedané, nezdieľajte bezstarostne interné identifikátory len preto, že sú menej identifikovateľné ako napríklad číslo pasu. V súlade so zásadou minimalizácie údajov, ktorá stanovuje, že by ste mali používať len toľko údajov, koľko je potrebné na dosiahnutie stanoveného účelu, by mal mať interný identifikátor danú pseudonymizovanú hodnotu len v rámci plánovanej „session“. V závislosti od zamýšľaného ohraničenia „session“, môže mať používateľ viacero „sessions“ v rámci daného súboru údajov alebo „session“, ktorá zahŕňa viacero súborov údajov.

10.1.1 Prípad použitia 1: Minimalistická „session“ (nie je potrebné prepojenie aktivít používateľa)

Tento prístup by sa mal použiť, ak sa zdieľa s treťou stranou aj ten najdetailnejší dátový prvok v rámci datasetu. Predpokladajme napríklad, že existuje interné webové riešenie na rozvoz materiálov a pomôcok, ktoré potrebujú pri svojej práci jednotliví zamestnanci. Tieto údaje sa zdieľajú s doručovateľskou spoločnosťou, aby sa uľahčili jednotlivé transakcie. Keďže neexistuje explicitný zámer korelovať „sessions“ pre každého používateľa tohto webového riešenia v rámci datasetu, interný identifikátor by mal mať v rôznych „sessions“ rôzne pseudonymizované hodnoty. Keďže súbor údajov, ktorý sa zdieľa, je veľmi podrobný, je potrebné identifikátory zodpovedajúcim spôsobom zastrieť.

Navrhované techniky pseudonymizácie

Na generovanie 128-bitového náhodného čísla ako pseudonymizovaného identifikátora sa odporúča použiť kryptograficky bezpečný generátor pseudonáhodných čísel (CSPRNG v kapitole 6.1.2). Toto číslo má požadovanú vysokú kvalitu náhodnosti a je dĺžkovo kompatibilné s interným identifikátorom (128-bitovým).

Ak nie je k dispozícii CSPRNG, môže sa namiesto neho použiť generický generátor náhodných čísel (RNG v kapitole 6.1.1). Aj keď nie je taký robustný ako CSPRNG, na tento účel je stále

prijateľný. Toto použitie je len poslednou možnosťou a malo by si vyžadovať dodatočné preskúmanie a schválenie.

10.1.2 Prípad použitia 2: Jedna „session“ na dataset (prepojenie aktivít toho istého používateľa v rámci datasetu)

Na rozdiel od prípadu 1, v ktorom sme zdieľali samostatnú činnosť, tento prípad použitia zahŕňa zdieľanie „session“, ktorá môže mať viacero činností. Ide napríklad aj o plnenie štatistickej povinnosti pri sčítaní obyvateľstva za daný rok dotknutými osobami, ktorá môže byť splnená v rámci viacerých „sessions“. Predpokladajme ale opäť prípad internej platformy na zabezpečenie materiálu a pomôcok. Jej používateľ si vybavil na nej akútnu potrebu a následne si uložil nejakú ďalšiu vec na svoj zoznam potrieb v budúcnosti. Tieto údaje potom treba zdieľať s externým analytickým tímom, ktorý vie poradiť, ako zlepšiť používateľské rozhranie s cieľom rýchlejšie nachádzať správne položky.

V tomto prípade sa každý dataset považuje za jednu nezávislú „session“. Všetky dátové prvky spojené s rovnakým interným identifikátorom by mali mať v rámci súboru údajov deterministickú a konzistentnú pseudonymizovanú hodnotu. Zvyčajne sa to dá dosiahnuť použitím kryptografickej alebo bezpečnej hashovacej funkcie alebo použitím vyhľadávacej tabuľky na udržiavanie konzistentných hodnôt pseudonymizovaného vnútorného identifikátora v rámci súboru údajov.

Navrhované techniky pseudonymizácie

Aplikujte kryptografickú funkciu HMAC-SHA256 na interný identifikátor s jedinečným, náhodne vygenerovaným 256-bitovým kryptografickým kľúčom pre súbor údajov (kapitola 6.1.4). Výstupnú 256-bitovú hashovaciu hodnotu potom možno použiť ako pseudonymizovaný identifikátor. Ak je žiaduce zachovať 128-bitovú dĺžku ako interný identifikátor, haš sa môže skrátiť na 128 bitov. Kľúč HMAC by sa mal použiť len raz a hneď potom zlikvidovať.

Prípadne použite na interný identifikátor zabezpečenú hašovaciu funkciu SHA-256 s jedinečnou 256-bitovou náhodne vygenerovanou „soľou“ („salt“ – ide o pridanie náhodne vygenerovaného reťazca znakov do pôvodného reťazca vytvoreného hašovaním ako je popísané v kapitole 6.1.4) pre dataset. Táto metóda tiež poskytuje 256-bitovú výstupnú hodnotu haš a v prípade potreby ju možno skrátiť na 128 bitov.

Ďalšou možnosťou je generovať náhodnú pseudonymizovanú hodnotu pre každý interný identifikátor ako v prípade použitia 1, ale každý interný identifikátor a jeho pseudonymizovanú hodnotu uložiť do vyhľadávacej tabuľky, aby sa zachovala konzistentnosť. Táto možnosť by sa mala zvážiť, ak nie je k dispozícii kryptografická alebo hašovacia funkcia. Vyhľadávacia tabuľka sa musí zabezpečiť a po použití zlikvidovať. Táto možnosť má dodatočné náklady, buď v podobe spotreby pamäte, alebo závislosti od externej služby.

Keďže rôzne datasety sa považujú za samostatné „sessions“, kryptografický kľúč, soľ alebo mapovacia tabuľka by sa nemali zdieľať medzi súbormi údajov.

10.1.3 Prípad použitia 3: „Session“ presahujúca viaceré datasety (prepojenie medzi datasetmi)

Tento prípad platí aj pri vypracovávaní štatistických výkazov a plnení spravodajskej povinnosti, kedy sa údaje od spravodajských jednotiek a právnických osôb zdieľajú so Štatistickým úradom

SR. Pokračujme ale v našom príklade s interným webovým riešením na objednanie materiálu a pomôcok a predpokladajme, že chcete zdieľať viacero príkladov úspešného naplnenia potrieb a opustenia portálu bez objednávky s analytickou spoločnosťou, aby ste mohli predpovedať výdavky na jednotlivé položky. V tomto prípade si zakrytie identifikátorov vyžaduje určitú pozornosť. Keď je potrebné konzistentne udržiavať pseudonymizovanú identitu vo viacerých datasetoch pre tú istú tretiu stranu, ide o rozšírenie prípadu použitia 2.

Navrhované techniky pseudonymizácie

Pri pseudonymizácii týchto súborov údajov by sa mal použiť rovnaký kryptografický kľúč (pre HMAC-SHA256) alebo „soľ“ (pre SHA-256), ako je popísané v kapitole 6.1.4. Je potrebné venovať náležitú pozornosť tomu, aby tieto kľúče alebo „soli“ boli riadne interne chránené, a každý kľúč alebo „soľ“ by sa mali používať len pre jednu tretiu stranu. HMAC-SHA256 vykonáva dve kolá hašovania, takže má približne dvojnásobné náklady na výpočet v porovnaní so SHA256. Jeho výhodou je, že používa štandardné kryptografické primitívy, ktoré sa starajú o dĺžku kľúča a nevyžadujú napojenie interného identifikátora a „soli“.

Pre každú tretiu stranu by ste mali viesť jednu vyhľadávaciu tabuľku. Každá vyhľadávacia tabuľka obsahuje verziu identifikátora prepojenú s interným datasetom, takže v prípade prelomenia bezpečnosti tretej strany budete presne vedieť, ktorá z nich bola zasiahnutá na základe interného identifikátora, ktorý sa nakoniec extrahuje. Potom môžete informovať zamestnancov, ktorých sa toto konkrétne narušenie súkromia týka, namiesto toho, aby ste museli kontaktovať všetkých. Tabuľka musí byť tiež zašifrovaná podľa osvedčených postupov s minimálne potrebnými prístupovými právami (kapitola 6.1.5.1).

10.2 Špecifický scenár ochrany údajov pri zdieľaní pomocou Systémov na správu osobných údajov v rámci Manažmentu osobných údajov

Koncepcia systému správy osobných informácií („Personal Information Management Systems (PIMS)“) je charakterizovaná ako "nové technológie a ekosystémy, ktorých cieľom je umožniť jednotlivcom kontrolovať zhromažďovanie a zdieľanie ich osobných údajov" (EDPS, 2016). Koncepcia PIMS je implementovaná v rámci Manažmentu osobných údajov. Aplikuje sa na takýto scenár: Všetky relevantné údaje, ktoré štátna správa eviduje o jednotlivcovi, sú uložené v dedikovanom úložisku občana, ktoré má plne pod kontrolou (Obrázok 27). V tomto scenári všetku kontrolu prístupu k údajom (vrátane pseudonymizovaných údajov) riadia samotní občania: ak poskytnú prístup prostredníctvom svojho PIMS, poskytujú prístup k údajom. Ak to technická architektúra umožňuje, prístup k týmto osobným údajom môže byť dokonca obmedzený na časti celkových záznamov údajov alebo len na súbory pseudonymizovaných údajov.

Obrázok 27: Scenár zdieľanie údajov pomocou PIMS

Občan poskytuje tretím stranám prístup k pseudonymizovaným údajom alebo pôvodným údajom na základe súhlasu na daný účel. Občana možno osloviť aj proaktívne prostredníctvom jeho WebID[66] a požiadať ho o súhlas s poskytnutím prístupu k (pseudonymizovaným) údajom na daný účel. PIMS je implementovaný aj ako aplikácia v smartfóne na udelenie (alebo odvolanie) súhlasu a uľahčenie prístupu k údajom uloženým v dôveryhodnom osobnom úložisku občana. Momentálne sa v Manažmente osobných údajov používa šifrovanie a podpisovanie osobných údajov len pre ich bezpečnosť, integritu a dôveryhodnosť. Ak by sa v budúcnosti uvažovalo o zdieľaní pseudonymizovaných datasetov, bude potrebné použiť techniky popísané v kapitole 6.1.7 alebo v kapitole 6.1.8.1 pre potlačenie zdieľania identity, ako aj techniky v kapitole 6.1.4 alebo 6.1.5 pre pseudonymizáciu samotného datasetu. Ďalšou možnosťou pri zdieľaní s ohľadom na ochranu súkromia je využívať techniky ZKP (kapitola 6.5.1) a ABC (kapitola 6.5.2).

10.3 Scenár analýzy osobných údajov pre Konsolidovanú analytickú vrstvu

V nasledujúcich prípadoch použitia sa berie do úvahy aktuálny stav rozvoja Konsolidovanej analytickej vrstvy, do ktorej budú mať prístup len interní zamestnanci štátnej správy a nie aj tretie strany. Ak by v budúcnosti vznikla potreba prístupu tretích strán, údaje v analytickej vrstve by sa museli dôslednejšie anonymizovať takými technikami, aby sa eliminovali v maximálnej možnej miere všetky riziká spomenuté v kapitole 8.2 (pozri aj tabuľku - Tabuľka 17). **Tiež by sa v tomto prípade neodporúčalo použitie najmenej bezpečnej deterministickej politiky pseudonymizácie podľa kapitoly 7.1.**

10.3.1 Prípad použitia 1: Štatistická analýza anonymizovaných a pseudonymizovaných údajov

V Konsolidovanej analytickej vrstve by sa mali pre všetky účely spracovania prioritizovať anonymizované údaje podľa techník popísaných v kapitole 6.1.9, pričom výber sa urobí na základe vhodnosti danej techniky pre daný dátový prvok a pre zachovanie jeho užitočnosti v datasete pre daný účel spracovania. Dôvodom je, že v KAV sa budú zhromažďovať rôzne datasety a informácie o pozadí, čo by umožňovalo mnohé útoky na spätnú identifikáciu

[66] Zdroj: https://www.w3.org/wiki/WebID, Dátum referencie: 14.02.2023

dotknutých osôb. Na druhej strane, odstránenie možnosti prepájať medzi sebou datasety od rôznych orgánov verejnej moci pre danú dotknutú osobu by zásadne degradovalo užitočnosť údajov, preto sa uvažuje aj o práci so pseudonymizovanými údajmi na účel prepájania datasetov, čo je vlastne mierna obmena prípadu použitia v kapitole 10.1.3. **Ide o aplikovanie najmenej bezpečnej deterministickej politiky pseudonymizácie podľa kapitoly 7.1, preto treba veľmi dbať na aplikovanie ďalších techník bezpečnosti údajov, ako napríklad dôsledné riadenie prístupov.**

Na pseudonymizáciu údajov, ktoré sa do KAV dostanú cez CIP, sa budú používať techniky implementované v CIPe podľa kapitoly 9.2. Ak by KAV neskôr využívala aj verejný cloud a v ňom vytvorený dátový „stack“, bude využívať nástroje uvedené v kapitole 6.4.3. Na pseudonymizované údaje budú aplikované aj ďalšie techniky bezpečnosti údajov ako vyhradené, chránené úložiská s dôsledným riadením prístupu, logovania a auditu. V typickej obmene prípadu použitia pre štatistické spracovanie sa uvažuje s kombináciou pseudonymizačných a anonymizačných techník, čo znamená, že:

- Údaje na úrovni „Vyhradené“ by mali byť prioritne vyňaté z datasetu (technikou potlačenia v kapitole 6.2.3) s ponechaním len jedného pseudonymizovaného identifikátora pre účel prepájania jednotlivcov medzi datasetmi. Na techniku pseudonymizácie sa odporúča použiť hašovanie HMAC s 256-bitovým kľúčom (kapitola 6.1.4).
- Ak niektoré údaje na úrovni „Vyhradené“ nemôžu byť vyňaté z datasetu, musia byť anonymizované technikou primeranou datasetu pomocou:
  - Generalizácie (kapitola 6.2.2),
  - Globálneho prekódovania (kapitola 6.2.4).

  Hoci podľa kapitoly 6.3.1 je štandardom k-anonymita 5 pre 100 percent údajov, odporúčame ju dosiahnuť čo najvyššiu pri zachovaní užitočnosti údajov.
- Údaje na úrovni „Dôverné“ musia byť anonymizované podľa kapitoly 6.3.1 na k-anonymitu 5 pre 100 percent údajov, pričom sa aplikuje podľa typu dátových prvkov a datasetu niektoré z týchto techník:
  - Maskovanie (kapitola 6.2.1)
  - Generalizácia (kapitola 6.2.2),
  - Globálne prekódovanie (kapitola 6.2.4).

10.3.2 Prípad použitia 2: Obohacovanie pseudonymizovaných údajov

V určitých scenároch môže byť potrebné, aby sa pseudonymizované identifikátory po zdieľaní s treťou stranou interne obnovili (ide o rozšírenie scenára podľa kapitoly 10.1). Tretia strana alebo aj analytický tím v rámci KAV môže napríklad vrátiť niektoré údaje s pridanými vlastnými metadátami alebo s pridanými výstupmi analýzy. Potom by bolo potrebné obnoviť pôvodné interné identifikátory z pseudonymizovaných hodnôt. **Tento scenár sa uvádza nateraz len pre úplnosť, pretože aktuálne nie je uvažovaný pre Konsolidovanú analytickú vrstvu.** V budúcnosti môžu dáta v rámci Konsolidovanej analytickej vrstvy putovať do externých systémov vo verejných cloudoch, alebo ich môžu analyzovať externé tímy dátových vedcov a analytikov. Existujú dve možnosti ako dosiahnuť internú obnovu identifikátorov: použitie mapovacej tabuľky na krížové porovnávanie medzi nespracovanými internými identifikátormi a ich pseudonymizovanými hodnotami alebo použitie obojsmernej kryptografickej funkcie na

zašifrovanie interných identifikátorov na pseudonymizované hodnoty a ich obnovenie prostredníctvom dešifrovania.

Mapovacia tabuľka

Použitie generickej obojsmernej mapovacej tabuľky na uloženie každého vygenerovaného pseudonymizovaného identifikátora s jeho pôvodným interným identifikátorom umožní pohodlné vyhľadávanie na úkor úložného priestoru a režijných nákladov na údržbu. Táto technika je výhodná pri spracovaní údajov veľkého rozsahu, ako sú operácie dátového skladu a analýzy, pretože hodnoty interných identifikátorov možno obnoviť jednoduchým spojením tabuliek. Tabuľka musí byť zašifrovaná podľa osvedčených postupov s minimálne potrebnými prístupovými právami.

Obojsmerná kryptografická funkcia

Namiesto jednosmernej hašovacej funkcie sa na generovanie pseudonymizovaného identifikátora s podobnými vlastnosťami môže použiť vhodná obojsmerná šifrovacia/dešifrovacia funkcia. Konkrétne namiesto použitia HMAC-SHA256 na interné identifikátory na generovanie pseudonymizovaných hodnôt možno interné identifikátory šifrovať pomocou AES s 256-bitovým kľúčom (v režime CBC s nulovým IV podľa kapitoly 6.1.5.1). Pôvodné interné identifikátory sa potom dajú obnoviť jednoduchým dešifrovaním pseudonymizovaných hodnôt tým istým kľúčom. Na rozdiel od jednosmerného hashovania by sa vygenerovaná hodnota nikdy nemala orezávať. Táto metóda nespôsobuje žiadne dodatočné režijné náklady na ukladanie, ale pri každom použití budete musieť vykonať operáciu dešifrovania.

Poznámka: Použitie funkcie obojsmerného šifrovania/dešifrovania je určené výlučne na účely generovania pseudonymizovaných hodnôt, nie na všeobecné šifrovanie údajov. Niektoré z robustnejších techník, ako je AES-GCM, sú ako také vynechané, pretože buď neprinášajú pridanú hodnotu (napr. režim spätnej väzby nezískava žiadnu výhodu, pretože dĺžka kľúča je väčšia alebo rovná bitovej dĺžke vnútorného identifikátora), alebo by výrazne zvýšili dĺžku výstupu bez zjavných výhod (napr. zahrnutie autentizačnej značky).

10.4 Scenár pre interné používanie pseudonymizovaných osobných údajov vo verejnej správe

10.4.1 Prípad použitia 1: Sledovanie bez uloženia počiatočných identifikátorov

Predstavme si, že by napríklad Ministerstvo hospodárstva vytvorilo portál na zber spätnej väzby a diskusie podnikateľov o existujúcich a pripravovaných reguláciách, pričom by diskusiu moderovali zamestnanci ministerstva, ktorí by aj zároveň odpovedali na rôzne nejasnosti. Aby sa podnikatelia cítili komfortne pri zdieľaní svojich podnikateľských skúseností a ťažkostí, bola by sociálna sieť anonymná. To znamená, že by používatelia mohli jednoducho sledovať príspevky iných používateľov a komentovať ich alebo vytvárať nové príspevky bez nutnosti prihlásenia. Ako sa však popisuje aj v úvode kapitoly 6, napriek tomu, že neexistuje postup prihlásenia, poskytovateľ aplikácie musí stále sledovať zariadenie používateľa (napr. na základe identifikátora zariadenia), aby mu mohol posielať notifikácie, keď niekto lajkuje a/alebo komentuje jeho príspevky. Hoci je sledovanie potrebné, poskytovateľ aplikácie v skutočnosti nepotrebuje poznať konkrétny identifikátor zariadenia (pokiaľ ho možno odlíšiť od všetkých ostatných identifikátorov). Tiež nie je potrebné, aby táto aplikácia zdieľala rovnaký identifikátor

používateľa/zariadenia s inými aplikáciami (treba však zdôrazniť, že to závisí aj od použitej platformy operačného systému).

V tomto prípade je zrejmé, že jednoduché použitie trvalého identifikátora zariadenia na sledovanie používateľa môže potenciálne viesť k identifikácii používateľa prostredníctvom identifikácie jeho zariadenia. Situácia sa mierne zlepší, ak sa použije nestály identifikátor (napr. GAID[67] v zariadeniach so systémom Android), ale opäť je možná identifikácia v rámci určitých časových limitov. Jednoduché hašovanie takéhoto identifikátora by neposkytlo významnú ochranu, pretože ktokoľvek so znalosťou identifikátora zariadenia bude môcť jednoducho spätne identifikovať zariadenie (a teda možno aj používateľa). Okrem toho by poskytovateľ aplikácie musel vo všetkých prípadoch uchovávať uvedené identifikátory zariadenia, hoci to nie je potrebné na účely konkrétnej operácie spracovania.

Na tento účel môže pseudonymizácia výrazne podporiť ochranu údajov v tomto prípade použitia, ak je správne implementovaná v návrhu. Možným prístupom by bolo použitie kľúčovanej hašovacej funkcie nad nestálym identifikátorom na vytvorenie pseudonymov, ktoré sa môžu použiť namiesto pôvodných identifikátorov (kapitola 6.1.4). Týmto spôsobom by poskytovateľ aplikácie nemusel uchovávať ani počiatočné identifikátory, zatiaľ čo príslušný tajný kľúč na hašovanie by mal byť bezpečne uchovávaný v inej databáze, ako je tá, v ktorej sú uložené pseudonymy. Okrem toho by sa prenos identifikátora na server aplikácie mal uskutočniť cez zabezpečený kanál - napríklad prostredníctvom protokolu TLS (Transport Layer Security) - aby sa zabezpečilo, že pri odpočúvaní siete nemožno zachytiť identifikátory pri prenose, a teda sa ich nedá žiadnym spôsobom spojiť s príslušnými pseudonymami. Protokol TLS však zároveň zabezpečuje, že zariadenie je skutočne pripojené k legitímnemu aplikačnému serveru, čo je potrebné na účely ochrany súkromia aj bezpečnosti. Ide o vnútornú vlastnosť protokolu TLS založenú na používaní digitálnych certifikátov. Existujú však útoky zamerané na kompromitáciu infraštruktúry certifikátov TLS. Na zmarenie takýchto útokov by aplikácie mali používať pripojenie certifikátu alebo predinštalované kľúče.

10.4.2 Prípad použitia 2: Ochrana prístupových údajov v databáze pre mobilné aplikácie

Zoberme si príklad, že by Úrad verejného zdravotníctva chcel vytvoriť preventívny program, ktorý by zahŕňal vytvorenie aplikácie na sledovanie počtu krokov občanov, ktorý by sa do neho zapojili. Informácie o krokoch používateľov by sa ukladali na serveri aplikácie, aby k nim používateľ mohol pristupovať cez internet z akéhokoľvek zariadenia. Pre jednoduchosť predpokladáme, že aplikácia počíta počet krokov používateľa bez toho, aby tieto údaje kombinovala s inými údajmi o používateľovi (napr. z iných aplikácií) alebo ich posielala iným príjemcom. Napriek tomu aj v tomto jednoduchom prípade poskytovateľ aplikácie vytvára profil používateľa s ohľadom na jeho denné návyky pri chôdzi. Používateľ sa autentifikuje na serveri aplikácie, aby mal prístup k svojim údajom, kombináciou e-mailovej adresy a hesla. Poskytovateľ aplikácie tak môže jasne identifikovať používateľa, pretože každý registrovaný používateľ by mal mať explicitný prístup k svojmu špecifickému používateľskému profilu.

V tomto prípade použitia budeme skúmať možnosť použitia pseudonymizácie na ochranu prihlasovacích údajov používateľov na serveri (v databáze) aplikácie. Jednoduchá hašovacia funkcia na meno používateľa alebo e-mailovú adresu zjavne nie je správnym prístupom k pseudonymizácii. Naopak, mohla by sa použiť kľúčovaná alebo solená hash funkcia (kapitola

[67] Zdroj: https://www.appsflyer.com/glossary/gaid/, Dátum referencie: 07.02.2023

6.1.4). Príslušný kľúč/soľ, ako aj pôvodné identifikátory by mali byť bezpečne uložené a oddelené od databázy s pseudonymizovanými údajmi, napríklad na dôveryhodnom autentifikačnom serveri.

Alternatívne možno pseudonymy vytvoriť použitím deterministickej symetrickej šifry, ako je napríklad AES, šifrovací kľúč, ktorý sa v tomto prípade zhoduje s dešifrovacím kľúčom, by mal byť opäť bezpečne uchovávaný oddelene (kapitola 6.1.5.1).

Po uplatnení takéhoto procesu pseudonymizácie a v závislosti od rozsahu a špecifických vlastností databázy pseudonymizovaných údajov (údajov o životnom štýle) by sa takáto databáza mohla používať na štatistické účely, a to aj pri analýzach od tretej strany. Pokiaľ totiž táto strana nemá prístup k tajnému kľúču/soli, nie je triviálne identifikovať používateľov. Na určenie tejto možnosti je však dôležitý celkový kontext, pretože najmä v malých/špecializovaných súboroch údajov sa zvyšujú možnosti odvodenia osobných údajov (ako bolo vysvetlené v úvode kapitoly 6). Ak dôjde k narušeniu tejto pseudonymizovanej databázy, spätná identifikácia bude výpočtovo náročná.

Pseudonymizovaná databáza by v žiadnom prípade nemala byť korelovaná so žiadnym iným identifikátorom zariadenia, ktorý vývojár aplikácie prípadne spracúva - napríklad na poskytovanie personalizovaných konfigurácií aplikácie, ktoré si vyberie používateľ. Vtedy treba použiť ďalší proces pseudonymizácie na pseudonymizáciu akéhokoľvek takéhoto identifikátora (pozri Prípad použitia 1 v kapitole 10.4.1).

10.4.3 Prípad použitia 3: Viacero pseudonymov pre rovnaké údaje

Zoberme si príklad inteligentného elektromeru pre riešenie politiky optimalizácie využívania zdrojov energie. Ide o elektromer, ktorý zaznamenáva údaje o spotrebe v domácnosti a odosiela ich príslušnému dodávateľovi elektriny a regulátorovi. Tieto stopy používa dodávateľ (prevádzkovateľ údajov) na účely fakturácie, aj regulátor (spracovávateľ údajov) na vyhodnocovanie účinnosti politík šetrenia energie. Používatelia (odberatelia elektriny) môžu prostredníctvom príslušnej mobilnej aplikácie kontrolovať informácie o svojej spotrebe energie v reálnom čase.

Jednou z možností zmierniť riziká ohrozenia súkromia v súvislosti s profilovaním zvyklostí domácností (ktoré možno získať prostredníctvom prevádzky inteligentného merača) je, že dodávateľ bude uchovávať údaje o spotrebe v pseudonymizovanej forme tak, že každému jednotlivému meraniu pochádzajúcemu z tej istej domácnosti (spotrebiteľa) budú pridelené rôzne pseudonymy. Pre daného spotrebiteľa sa teda jeho stopy v jednom časovom intervale ukladajú pod pseudonymom A, v ďalšom časovom intervale pod iným pseudonymom B atď. Na splnenie takejto vlastnosti by možným prístupom k pseudonymizácii mohla byť pravdepodobnostná šifrovacia schéma (opísaná v kapitole 6.1.5). Pritom ide o aplikovanie politiky podľa kapitoly 7.3.

10.4.4 Prípad použitia 4: Lokálne generovanie pseudonymov

V tejto časti uvedieme príklad, kedy by regulátor (Ministerstvo dopravy) chcel inteligentnou aplikáciou monitorovať správanie vodičov služieb ako Bolt s cieľom ochrániť bezpečnosť prepravovaných osôb. Kedykoľvek by vodič (používateľ aplikácie) udržiaval aplikáciu aktívnu, vytváral by sa profil jeho jazdných návykov (a ukladal sa do databázy poskytovateľa aplikácie). Poskytovateľ aplikácie v predvolenom nastavení nespája údaje so žiadnymi inými údajmi zariadenia používateľa a automaticky neodosiela údaje žiadnemu príjemcovi. Voliteľná funkcia aplikácie umožňuje používateľovi povoliť prenos údajov poskytovateľom aplikácie regulátorovi

(za čo by mohol získať isté výhody ako napríklad daňové bonusy) alebo do zapojených poisťovní (napríklad aby používateľ získal zľavu na poistení). Hoci poskytovateľ aplikácie musí byť schopný sledovať používateľa (vodiča), aby mu mohol doručiť údaje, ktoré sú pre neho relevantné v jeho konkrétnom zariadení, nie je potrebné, aby poskytovateľ poznal skutočnú identitu používateľa. Takýto typ identifikácie bude potrebný len vtedy, keď používateľ výslovne povolí poskytovateľovi zasielať jeho údaje regulátorovi alebo poisťovni.

Pseudonymizácia môže jednoznačne podporiť aj tento scenár. Možné riešenie ochrany údajov už v štádiu návrhu spočíva v tom, že sa používateľovi umožní vytvoriť pseudonym v jeho zariadení tak, aby ho nikto iný nemohol spätne identifikovať, pokiaľ to používateľ nepovolí. Napríklad to môže spraviť vhodným zašifrovaním identifikátorov používateľa tak, aby mal prístup k dešifrovaciemu kľúču len používateľ. Samozrejme, pri tomto prístupe by sa mali zaviesť vhodné bezpečnostné mechanizmy; napríklad tajný kľúč/heslo by sa nemali dodávať hneď s aplikáciou. Okrem toho by sa pseudonym vygenerovaný aplikáciou v zariadení používateľov mal prenášať na server aplikácie zašifrovaný - napríklad prostredníctvom protokolu TLS - a nemal by byť korelovaný so žiadnym iným identifikátorom zariadenia. Existujú tiež špecializované kryptografické techniky (pozri kapitolu 6.1.7), ktoré umožňujú používateľovi generovať pseudonym lokálne v jeho prostredí bez nutnosti výmeny informácií s vydávajúcimi stranami tak, aby mohol kedykoľvek dokázať, že je vlastníkom pseudonymu.

10.4.5 Komplexný prípad použitia 5: Kontinuálne vytváranie profilu používateľa

Pokiaľ ide o reálne systémy, najväčšie problémy často nespôsobuje výber techniky pseudonymizácie použitej pre jeden alebo dva konkrétne identifikátory, ale skôr implicitná prepojiteľnosť medzi súborom pseudonymov a iných hodnôt údajov, ktoré sú spojené do zložitejšej štruktúry údajov. Najbežnejším príkladom je systém, ktorý vytvára profily používateľov a obohacuje tieto profily o osobné údaje o používateľovi vždy, keď sú k dispozícii nové údaje, spravidla aj z iného rezortu verejnej správy. V tomto prípade, aj keď sú identifikátory ako rodné číslo a e-mailová adresa používateľa prísne pseudonymizované, stále existuje veľká hrozba spätnej identifikácie alebo bližšieho určenia, ktoré sú možné aj na samotnej pseudonymizovanej štruktúre údajov.

10.4.5.1 Modelový príklad

V rámci detailnejšej diskusie si zoberieme hypoptetický príklad použitia pre online portál s elektronickými službami. Prevádzkovateľom je Organizácia A (ďalej len ORG_A), ktorá vystupuje ako prevádzkovateľ údajov a umožňuje svojim používateľom (predpokladáme, že sú to len fyzické osoby) zaregistrovať sa na účet, ktorý je uložený v dátovom centre ORG_A. Prostredníctvom tohto účtu môžu používatelia využívať súbor funkcií, ktoré napríklad umožňujú prepojenie sa na iných používateľov s cieľom spoločne vybaviť životnú situáciu (napríklad majiteľ nehnuteľnosti pri nahlásení trvalého pobytu v nej). Pri registrácii sa využíva elektronický občiansky preukaz, z ktorého sa vyčíta skutočné meno a priezvisko, dátum narodenia a pohlavie, rodné číslo. Používateľ môže ďalej zadať súbor nepovinných osobných údajov ako miesto aktuálneho pobytu, záujmy ako aj platnú e-mailovú adresu. Vždy, keď používatelia pristupujú k niektorej zo služieb ORG_A, ich interakcia sa zaznamená a pridá do ich používateľského profilu - vrátane časovej pečiatky a IP adresy prístupu.

S cieľom zlepšiť súlad s GDPR sa vedenie ORG_A rozhodlo pseudonymizovať IP adresy v protokoloch o prístupe podľa techník uvedených v kapitole 8.3.2, ktoré predchádzajú možnému útoku. Zostávajúce informácie sa uchovávajú v jednoduchom texte, pretože sú potrebné na to, aby sa používateľovi v prípade potreby prezentovali na webových stránkach ORG_A alebo aby sa vykonali kontroly a validácie (napríklad dátum narodenia je potrebný na výpočet veku a

overenie, či je používateľ starší ako 16 rokov pri prístupe k špeciálnym službám). Pseudonymizácia e-mailovej adresy tu nie je možná, pretože ORG_A musí byť schopná posielať používateľom e-maily s oznámeniami (a iným obsahom).

Predpokladajme druhú organizáciu, Organizáciu B (ďalej len ORG_B), ktorá pôsobí ako spracovateľ údajov v mene ORG_A a ktorej úlohou je udržiavať úložiskové a bezpečnostné služby pre časti databázy používateľov ORG_A. V tejto pozícii má ORG_B prístup k pseudonymizovaným súborom logov ORG_A, teda k pseudonymizovaným IP adresám a časovým pečiatkam všetkých prístupov na webové stránky, ale nie k samotným pôvodným IP adresám. V takomto nastavení ORG_B nemôže spätne identifikovať používateľov patriacich k IP adrese, pretože tieto údaje sú uložené v inej databáze ORG_A, ku ktorej ORG_B nemá prístup. V súvislosti s pseudonymizáciou sa teda objavuje scenár podľa kapitoly 10.1, pričom ORG_A je prevádzkovateľom údajov a ORG_B je následným spracovateľom údajov.

10.4.5.2 Informácie obsiahnuté v údajoch

Na prvý pohľad nie je ORG_B schopná prelomiť pseudonymizáciu IP adries vykonanú ORG_A, ak sa predpokladá, že ORG_A využíva dostatočne silnú pseudonymizačnú funkciu. V závislosti od pseudonymizačnej funkcie a najmä od politiky pseudonymizácie (pozri kapitolu 7) môže byť ORG_B stále schopná odvodiť, či sa určitý pseudonym vyskytuje v databáze často, zriedkavo, len raz alebo vôbec. To by samo o sebe nemuselo stačiť na odhalenie identity, ale už sa to dá využiť na identifikáciu často pristupujúcich používateľov. Ak záznam o prístupe obsahuje pseudonym s vysokou frekvenciou výskytu, ORG_B môže vyvodiť, že ide pravdepodobne o intenzívneho používateľa portálu. A naopak, ak sa pseudonym vyskytuje v súbore údajov prvýkrát, s najväčšou pravdepodobnosťou sa tento používateľ práve zaregistroval na portál a prvýkrát pristupoval k svojmu používateľskému účtu, alebo sa zmenila IP adresa registrovaného používateľa (čo sa môže stať často, čím sa všetky tieto pozorovania stávajú pravdepodobnostnými).

Tento druh informácií obsiahnutých v údajoch už môže byť pre ORG_B užitočný, napríklad na zistenie, koľko používateľov ORG_A je trvalých používateľov a koľko z nich sa raz zaregistruje a druhýkrát sa už nevráti (s určitou pravdepodobnostnou mierou chyby na základe zmeny IP adries). Tieto informácie už môžu byť rozhodujúce v nastavení zmluvného vzťahu medzi ORG_A a ORG_B.

Okrem týchto informácií obsiahnutých v údajoch, skutočnosť, že ORG_B má nepretržitý prístup k databáze ORG_A, umožňuje ORG_B ďalší typ zhromažďovania informácií: nepretržitým monitorovaním datasetu uloženého pre ORG_A sa ORG_B dozvedá o zmene daného datasetu. To zahŕňa celkový počet prístupov na webovú stránku ORG_A, čo sa môže využiť aj napríklad na určovanie počtu nových registrácií používateľov (prvýkrát sa vyskytujúcich pseudonymov) za deň alebo mesiac. Aj keď tieto informácie majú väčšinou štatistický charakter, môžu sa už teraz využiť na zinscenovanie skutočných útokov na rozoznanie používateľov („discrimination attacks“) ako aj určiť rôzne vplyvy na rôzne skupiny používateľov (pozri kapitolu 8.1.3). A to preto, lebo ORG_B sa dozvie, ktorý pseudonym nového používateľa sa v ktorý deň objavil ako prvý, čo umožní ORG_B sledovať množstvo interakcií tohto konkrétneho používateľa s ORG_A. Tieto informácie sa môžu ľahko stať problémom ochrany dotknutých osôb, ako sa ukáže neskôr.

10.4.5.3 Prepojené údaje

V modelovom príklade prípadu použitia poskytujú údaje prístupné pre ORG_B viac informácií ako len adresy IP: každý záznam logu ukladá aj časovú pečiatku prístupu. Preto namiesto častého monitorovania zmien v databáze ORG_A sa ORG_B môže jednoducho spoliehať na

časové pečiatky pripojené ku každému pseudonymu a vykonávať rovnaký typ rozoznávania používateľov ako predtým. Časové pečiatky sú uložené spolu so pseudonymizovanými IP adresami, a preto sú priamo prepojené s týmito informáciami jedna k jednej. Na základe týchto prepojených údajov môže ORG_B ďaleko rozšíriť svoje znalosti o konkrétnych používateľoch ORG_A: Pristupuje konkrétny používateľ k portálu viac ráno, počas obedňajšej prestávky alebo večer? Len alebo väčšinou v nedeľu? Len počas náboženských sviatkov podľa pravoslávneho kalendára? Len v čase školských prázdnin v Košickom kraji?

Každý takýto dodatočný typ charakteristiky umožňuje ORG_B priblížiť sa k porušeniu pseudonymizácie, a to len na základe uložených časových značiek a možnosti prepojiť rôzne záznamy údajov s rovnakými pseudonymami. Ako vidno, tento druh informácií začína poskytovať ORG_B určitú charakteristiku používateľov ORG_A, ktorú možno považovať za osobné údaje. Takéto prepojenia si však vyžadujú prepojenie ďalších informácií so samotnými štruktúrovanými datasetmi, ako je napríklad pravoslávny kalendár alebo prázdniny podľa krajov. Takýto druh získaných informácií má štatistickú povahu, teda nie je stopercentne spoľahlivý, ale len s určitou pravdepodobnosťou. V tomto prípade platí, že čím viac záznamov údajov sa nachádza v databáze, tým spoľahlivejšia (alebo falzifikovateľnejšia) je hypotéza o prepojení. Teda čím väčšia je sieť používateľov ORG_A, tým ľahšie sa pre ORG_B vykonáva takýto diskriminačný alebo dokonca reidentifikačný útok.

Tento príklad obsahoval len pseudonymizovanú IP adresu a časovú pečiatku. Platí to aj pre pseudonymizovanú e-mailovú adresu namiesto pseudonymizovanej IP adresy, pretože tá sa mení menej často, a preto je pre človeka identifikátorom, ktorý je viac jedinečný.

10.4.5.4 Zodpovedajúce pravdepodobnostné rozdelenia výskytov

Dátové štruktúry uvedeného modelového príkladu sú pomerne malé a zjednodušené: len IP adresa a časová pečiatka. Napriek tomu môžu stačiť na diskriminačné alebo dokonca reidentifikačné útoky, ak je k dispozícii dostatok znalostí pozadia („background“). Okrem toho záznamy reálnych údajov zvyčajne uchovávajú viac informácií ako len tieto dve hodnoty, preto záznamy údajov obsahujú viac podrobností, ktoré sa dajú využiť na odhalenie pseudonymov.

Zoberme do úvahy, že ORG_A v každom dátovom zázname ukladá viac než len časovú pečiatku a pseudonymizovanú adresu IP, napríklad ukladá aj typ a verziu prehliadača používaného daným používateľom (treba poznamenať, že ide o predvolené správanie protokolu, napríklad webového servera Apache), súbor a preferencie jazykov, ktoré používateľ ovláda (ako sú definované v nastaveniach prehliadača), verziu operačného systému počítača používateľa atď. Ako zistila nadácia Electronic Frontier Foundation v projekte Panopticlick[68], už len táto kombinácia nastavení prehliadača môže stačiť na jednoznačnú identifikáciu určitého prehliadača - a teda aj používateľa - online webovej stránky. Ak ORG_A teraz ukladá všetky tieto informácie pre každý prístup na svoju webovú stránku, ORG_B k nim môže mať prístup.

Aj keď ORG_A vykonáva určitý druh pseudonymizácie každej z týchto konfigurácií (napríklad ukladaním iba kľúčovaného hašu reťazca s verziou prehliadača prijatého z prehliadača používateľa podľa kapitoly 6.1.4), ORG_B môže aj napriek správnemu využitiu funkcie pseudonymizácie vidieť všetky tieto pseudonymizované reťazce verzie prehliadača, vypočítať štatistiku o tom, ktorá hodnota hašu sa ako často objavuje v celkovej databáze ORG_A, a

[68] Zdroj: https://panopticlick.eff.org/, Dátum referencie: 08.03.2023

porovnať toto rozdelenie rôznych existujúcich hodnôt s verejne dostupnými štatistikami zhromaždenými na spomínanej webovej stránke Panopticlick, aby odhalila skutočný reťazec verzie prehliadača za každou hodnotou hašu. Na odhalenie týchto pseudonymov môže s vysokou pravdepodobnosťou úspechu stačiť už len to, že štatistické rozdelenie rôznych pseudonymov sa zhoduje so štatistickým rozdelením ich predpokladaných jednoduchých textov.

To samozrejme do veľkej miery závisí od zvoleného prístupu k pseudonymizácii. Ak sa použije vhodný prístup, napríklad pridanie metadát k argumentu pseudonymizačnej funkcie, môže sa zabezpečiť väčšia ochrana proti spätnej identifikácii.

10.4.5.5 Ďalšie znalosti pozadia

Ak má ORG_B dodatočné znalosti o charakteristikách určitého používateľa a snaží sa odhaliť záznamy o tomto používateľovi z pseudonymizovanej databázy, ktorú získava od ORG_A, každá dodatočná informácia sa môže stať kritickou. Ak ORG_B vie, že konkrétny cieľový používateľ je muž a používa prehliadač Chrome na zariadení iPad, už len táto informácia výrazne zužuje súbor možností používateľských profilov, ktoré ORG_B vidí. Každá z týchto hodnôt údajov, aj keď je pseudonymizovaná, zmenšuje množinu možností, t. j. množinu používateľských profilov obsiahnutých v databáze ORG_A, ktoré môžu patriť konkrétnemu cieľovému používateľovi, ktorého hľadá ORG_B. Informácie o prehliadači možno riešiť útokom pomocou pravdepodobnostného rozdelenia načrtnutým v kapitole 10.4.5.4, pričom sa odstráni veľká časť profilov používateľov, ktorí majú pseudonymy prehliadačov s príliš veľkým alebo príliš malým počtom výskytov, aby sa zhodovali s pravdepodobnosťou konfigurácie konkrétneho "Chrome na iPade".

Zo zostávajúcich profilov sa triviálnym útokom pomocou hrubej sily (kapitola 8.3.1) alebo štatistickým distribučným útokom zistí (ide o útok s využitím znalosti pozadia („background“), stručne spomenutý v kapitole 8.3), ktorý pseudonym zodpovedá ktorému pohlaviu, čím sa vylúči približne polovica zostávajúcich profilov používateľov. Ak majú teraz všetky zostávajúce profily používateľov spoločné to, že prvý prístup k ORG_A bol medzi májom a júlom 2020, ORG_B sa už o tomto konkrétnom používateľovi niečo dozvedela: v tomto období sa zaregistroval na portál ORG_A. Ide o úspešný inferenčný útok (stručne spomenutý v kapitole 8.3). Pri ďalšej analýze zvyšných používateľských profilov sa ORG_B môže dozvedieť o špecifickom vzore časových pečiatok využívania riešenia ORG_A zistenom pri dvoch z týchto používateľských profilov, a to tak, že sa zhodujú s predpokladaným vzorom využívania cieľovej osoby (ktorý ORG_B mohla pozorovať pri niektorých príležitostiach v minulosti). Cieľová množina na vyhľadávanie sa teda zredukuje len na dva používateľské profily.

Všetky informácie, ktoré majú oba tieto profily spoločné, musia platiť aj pre konkrétnu cieľovú osobu, čo pravdepodobne ORG_B o ich cieli vyhľadávania už veľa napovedá. Na vylúčenie zostávajúceho falošného kandidáta môže ORG_B jednoducho monitorovať využívanie portálu týmito dvoma profilmi a pri ďalšom prístupe overiť, či tento prístup mohol pochádzať od cieľovej osoby alebo nie (na základe dodatočných znalostí získaných z tých faktov, ktoré sa ORG_B už o svojom cieli dozvedela). Nakoniec je ORG_B schopná prepojiť profil používateľa s cieľovou identitou. Tým je ORG_B tiež schopná odhaliť všetky pseudonymizácie vykonané aj na hodnotách údajov tejto osoby, čo potenciálne umožňuje odhaliť alebo rozoznať aj iné používateľské profily.

Napriek tomu je potrebné poznamenať, že problém dodatočných dostupných informácií je "ortogonálny" k pseudonymizácii, pričom ide v prvom rade o problém ochrany údajov priamo v návrhu. Preto sa odporúča okrem pseudonymizácie uvažovať aj o vnášaní šumu do argumentov

pseudonymizačnej funkcie alebo o použití generalizácie s cieľom znížiť účinnosť útokov pomocou hrubej sily (pozri aj kapitoly 6.1.8 a 6.1.9). Tieto dodatočné techniky predstavujú stratégiu, ako ďalej posilniť pseudonymizáciu a chrániť sa pred príslušnými útokmi.

10.4.5.6 Prepojenie viacerých zdrojov údajov

Okrem vyššie uvedeného prípadu použitia ORG_A a ORG_B vzniká ešte náročnejší scenár pseudonymizácie, keď sa neuvažuje len o účasti dvoch organizácií (ORG_A a ORG_B), ale keď sa predpokladá vznik rozsiahleho zdieľania pseudonymizovaných údajov. V takýchto scenároch viaceré organizácie zdieľajú pseudonymizované datasety osobných údajov so zámerom umožniť určitú pridanú hodnotu (napríklad vytváranie profilov na komplexné proaktívne služby), avšak zároveň chrániť identitu samotných dotknutých osôb. Často počúvaným argumentom v takýchto scenároch je, že pseudonymizácia zabraňuje spätnej identifikácii dotknutých osôb, čím sa legalizuje takéto zdieľanie údajov. Nejdeme teraz argumentovať v prospech alebo proti legitímnosti zdieľania pseudonymizovaných datasetov, ale skôr hľadať správne uplatňovanie pseudonymizácie v takýchto scenároch.

Predpokladajme súbor organizácii C až G, ktoré všetky zhromažďujú osobné údaje o svojich používateľoch, ako napríklad údaje zhromaždené organizáciou ORG_A v predchádzajúcom modelovom príklade. Prepojenie profilov používateľov rôznych organizácií by sa mohlo vykonať porovnaním e-mailových adries, ktoré využívajú príslušní používatelia. Ak sa dva používateľské profily nájdené napríklad v organizáciách C a E zaregistrovali s presne rovnakou e-mailovou adresou, s najväčšou pravdepodobnosťou patria tomu istému subjektu údajov. Samotná e-mailová adresa je však samozrejme osobným údajom, ako bolo uvedené v kapitole 3.2. Preto je potrebné použiť pseudonymizáciu e-mailových adries v datasetoch organizácií C a E pred ich zdieľaním medzi spoločnosťami C, D, E, F a G.

Výzvou je, že všetci účastníci chcú zachovať užitočnosť pseudonymizovaných údajov na prepojenie profilov patriacich tej istej osobe bez toho, aby sa znížila ochrana identity daného používateľa. Preto musí všetkých päť organizácií uplatňovať úplne rovnakú pseudonymizáciu s využitím rovnakej pseudonymizačnej funkcie a pseudonymizačného tajomstva, aby mohli navzájom porovnávať a spájať záznamy údajov z rôznych datasetov (ide o aplikovanie najmenej bezpečnej politiky podľa kapitoly 7.1). V tomto prípade existuje jasný rozpor medzi užitočnosťou (prepojenie pseudonymizovaných e-mailových adries) a ochranou (používateľov týchto e-mailových adries). Inými slovami, organizácie C a E by mali mať možnosť a povolenie dozvedieť sa, že ich konkrétne dátové záznamy majú rovnakú e-mailovú adresu, a teda patria k tomu istému používateľovi, ale nemali by byť schopné zistiť, o akú e-mailovú adresu - a teda dotknutú osobu - ide.

V takýchto scenároch použitie slabých pseudonymizačných techník (ako je obyčajné hašovanie) umožňuje útoky pomocou hrubej sily, hádaním alebo použitím pravdepodobnostného rozdelenia, ako sa uvádza vyššie. Po obohatení o dodatočné (neosobné) údaje obsiahnuté v záznamoch zdieľaných údajov a možno aj s určitými dodatočnými znalosťami o pozadí sa tieto útoky dajú považovať za praktické a v mnohých scenároch sú aj do značnej miery úspešné. Ešte horšie je, že čím viac organizácií zdieľa informácie o atribútoch konkrétnej dotknutej osoby, tým viac informácií má k dispozícii útočník na použitú pseudonymizáciu, a teda tým pravdepodobnejší je úspech takýchto útokov.

Riziká ochrany súkromia sa môžu vyskytnúť aj vo všeobecnejšom scenári, keď organizácie používajú rôzne (a dokonca silné) techniky pseudonymizácie identifikátorov svojich používateľov (napr. e-mail alebo IP adresa). Predpokladajme, že uvedená skupina organizácií C až G poskytuje takéto pseudonymizované údaje ORG_B, aby im poskytla napríklad analytické

služby. Ak sú k poskytnutým pseudonymom pripojené informácie o prehliadači/zariadení používateľa, ako je opísané v kapitole 10.4.5.4 (nastavenia prehliadača, operačný systém atď.), a pripomeňme si, že sa očakáva, že všetky takéto informácie o zariadení sú pre každé zariadenie jedinečné, potom ORG_B môže triviálne prepojiť rôzne pseudonymy poskytnuté rôznymi organizáciami, ktoré zodpovedajú tomu istému používateľovi.

10.4.5.7 Protiopatrenia

Ako sa uvádza v kapitole 7, politiky (dokumentovej alebo úplne) randomizovanej pseudonymizácie znižujú prepojenie medzi rôznymi pseudonymami z rôznych datasetov, a preto môžu zmierniť alebo dokonca odstrániť štatistické charakteristiky pseudonymizovaných databáz. Zároveň obmedzujú možnosť prepojenia rôznych záznamov údajov (potenciálne rozmiestnených v mnohých organizáciách) s jedným profilom používateľa. Preto aj v prípade použitia náhodnej pseudonymizácie môže byť ORG_B stále schopná vykonať vyššie uvedené útoky, ak dokáže odhaliť, či dva rôzne pseudonymy patria tomu istému identifikátoru. Podobne aj organizácie C a E môžu úspešne spätne identifikovať dotknutú osobu, ktorá sa skrýva za spoločnými používateľskými profilmi. Tu sa opäť prejavuje kompromis medzi ochranou a užitočnosťou údajov.

Pre príklad prípadu použitia v tejto kapitole môže ORG_A použiť pseudonymizačnú politiku, ktorá pseudonymizuje nielen samotné IP adresy, ale aj všetky možné kombinácie IP adries a časových pečiatok. Potom sa prepojenie časovej pečiatky s akýmkoľvek externým zdrojom údajov stáva neuskutočniteľným, pretože tieto informácie už nie sú pre ORG_B k dispozícii. Na úspešnú spätnú identifikáciu by ORG_B musela poznať (alebo odhadnúť) presnú kombináciu IP adresy a časovej pečiatky. Vo všeobecnosti sa pseudonymizácia kombinácie vstupných údajov nedá rozumne odhaliť bez toho, aby ste poznali (alebo uhádli) všetky vstupné údaje v jednoduchom texte. Pri tomto nastavení by takáto pseudonymizácia zablokovala akýkoľvek pokus ORG_B odhaliť daný pseudonym oveľa robustnejším spôsobom.

Na rozšírenie základných pseudonymizačných techník na štruktúrované dátové záznamy často stačí považovať celý dátový záznam za vstup a použiť prispôsobenú kombináciu kľúčovaných hašovacích funkcií a techník spoločných pre anonymizáciu vo všeobecnosti.

11 Zhrnutie: Odporúčania pre štandardy anonymizácie

V moderných organizáciách je zdieľanie údajov kľúčovým faktorom rastu, angažovanosti, personalizácie služieb, získavania znalostí z údajov, a takmer všetkých aspektov inovácií. Na zdieľanie údajov s kontrolou súkromia je k dispozícii niekoľko techník. Niektoré z týchto techník sú súčasťou štandardných požiadaviek na bezpečnosť údajov (dokument 1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov) v tom, že riadia prístup k údajom, keď opúšťajú systémy organizácie. Iné techniky zahŕňajú anonymizáciu a pseudonymizáciu - zahmlievanie údajov a ich spracovanie tak, aby sa dalo vyhnúť narušeniu súkromia. **Pseudonymizácia** je zavedený a akceptovaný proces deidentifikácie, ktorý získal ďalšiu pozornosť po prijatí GDPR, kde sa uvádza ako **bezpečnostný mechanizmus a mechanizmus ochrany údajov už od návrhu**. Existujú aj odvetvím uznávané techniky ako k-anonymita a l-diverzita na meranie vplyvu techník ochrany súkromia, aby ste mohli posúdiť, či môžete bezpečne zdieľať údaje. Zdieľanie údajov je jedným z rozhodnutí, ktoré už prakticky nemožno zvrátiť, a pritom môže predstavovať značný vplyv na používateľov, o ktorých tieto údaje sú. Preto organizácie by mali údaje zdieľať veľmi opatrne.

Klasifikácia údajov je dôležitou súčasťou celkovej stratégie manažmentu a ochrany údajov. Ide o proces, ktorý pomáha identifikovať dátové aktíva a vplyv ich vlastníctva na celkové riziko ochrany osobných údajov pre organizáciu a jej klientov. Keď klasifikujete údaje a prehodnocujete svoje klasifikácie, môžete inteligentnejšie zavádzať techniky, nástroje a politiky na riešenie rôznych prípadov použitia bez zbytočnej byrokracie a procesov. Zložitosť zákonov o ochrane osobných údajov možno zvládnuť pomocou dôkladnej klasifikácie údajov a poznatkov, ktoré z nej vyplývajú. Klasifikácia údajov je strategickou investíciou do budovania skutočne medzirezortného systému, keďže ide spravidla o medzirezortný proces, ktorý pomôže pri ochrane súkromia, pretože ku konečnému výsledku bude musieť prispieť niekoľko tímov. Klasifikácia údajov musí byť zohľadnená a zaznamenaná aj v dátovom katalógu, ktorému sa podrobne venujeme v dokumente 1.1.2 Štandardizácia pre modelovanie údajov.

Rôzne techniky ochrany súkromia treba vnímať v širšom kontexte kvality a bezpečnosti údajov. **Je to skôr umenie a veda než hľadanie jedného univerzálneho riešenia.** Používajú sa rôzne kombinácie týchto techník s rôznym stupňom zahmlievania údajov, ktoré sa následne iterujú s rôznymi datasetmi. V niektorých prípadoch je ľahké rozoznať vplyv na súkromie, zatiaľ čo v iných prípadoch treba urobiť kvalifikované rozhodnutie. Hlavným záverom je, že sú k dispozícii techniky a nástroje na zlepšenie ochrany súkromia a kvantifikáciu vplyvu. Treba ich použiť pred zdieľaním údajov a budúcnosť sa zavďačí.

Ak je cieľom pseudonymizácia**, jedným z hlavných ponaučení je, že neexistuje jediné ľahké riešenie pre pseudonymizáciu, ktoré funguje pre všetky prípady použitia všetkých možných scenárov** (kapitola 10)**.** Naopak, vytvorenie robustného procesu pseudonymizácie si vyžaduje vysokú úroveň kompetencie na to, aby sa dala znížiť hrozba rozlíšiteľnosti údajov alebo útokov na re-identifikáciu a aby sa zároveň udržiaval stupeň užitočnosti údajov nevyhnutný pre spracovanie pseudonymizovaných dát.

11.1 Odporúčané postupy podľa ISO štandardov, GDPR a dobrej praxe

Nasledujúcich päť kľúčových postupov pomáha riadiť riziko narušenia súkromia osobných údajov na každom kroku:

1. Klasifikácia (kapitola 3.1 a 3.2),

2. Posúdenie príslušných rizík pre ochranu údajov pre každý konkrétny prípad spracúvania údajov (kapitola 8.2),
3. Inventarizácia (v dátovom katalógu popísanom v dokumente 1.1.2 Štandardizácia pre modelovanie údajov),
4. Utajenie prostredníctvom pseudonymizácie na základe dôsledne vybranej techniky a politiky, v prípade potreby doplnenie o techniky anonymizácie (kapitoly 6.1, 7, 6.3 a 6.1.9).
5. Ochrana pseudonymizačného tajomstva („secret“) na základe zásad popísaných nižšie (1 až 3).

Aby bola pseudonymizácia účinná, musí ten, kto pseudonymizáciu vykonáva, vždy chrániť pseudonymizačné tajomstvo pomocou vhodných technických a organizačných opatrení. Tieto opatrenia jednoznačne závisia od konkrétneho scenára pseudonymizácie (pozri kapitolu 10), avšak napriek tomu vždy platí nasledovné:

1. Po prvé, pseudonymizačné tajomstvo musí byť izolované od datasetu, t. j. pseudonymizačné tajomstvo a dataset sa nikdy nesmú spracúvať v tom istom súbore alebo databáze (inak by bolo pre útočníka príliš jednoduché obnoviť identifikátory).
2. Po druhé, pseudonymizačné tajomstvo sa musí bezpečne vymazať zo všetkých nezabezpečených úložísk a systémov.
3. Po tretie, prísne politiky riadenia prístupu musia zabezpečiť, aby k tomuto tajomstvu mali prístup len oprávnené subjekty. Bezpečný systém logovania a auditu musí sledovať všetky žiadosti o prístup k tajomstvu. V neposlednom rade pseudonymizačné tajomstvo musí byť zašifrované, ak je uložené na úložisku, čo si zasa vyžaduje správnu správu a uchovávanie kľúčov na toto šifrovanie.

Je zrejmé, že pseudonymizácia nie je nevyhnutnou podmienkou pre všetky prípady spracúvania osobných údajov, preto je posúdenie príslušných rizík pre ochranu údajov (pre každý konkrétny prípad spracúvania údajov) neoddeliteľnou súčasťou rozhodnutia o tom, či a ako možno pseudonymizáciu zaviesť. **Vymedzenie cieľov a zámerov pseudonymizácie v každom konkrétnom prípade použitia je v tomto procese kľúčové.** Na tento účel môžu byť pre prevádzkovateľov (ako aj dodávateľov produktov, služieb a aplikácií) veľmi cenné príslušné osvedčené postupy a príklady pseudonymizácie v kontexte GDPR. Odporúča sa vždy odkazovať na akúkoľvek úspešnú implementáciu v súkromnom alebo verejnom sektore, analyzovať jej kľúčové vlastnosti, ako aj možnosti prevádzkovateľov údajov využiť rovnaký model v budúcnosti.

11.2 Odporúčané techniky a nástroje

Technická realizácia pseudonymizácie a anonymizácie a výber nástrojov do veľkej miery závisí od súčasného stavu implementácie IT systémov. Okrem toho závisí od:

- Kontextu, v ktorom sa má pseudonymizácia a anonymizácia použiť,
- Zapojených subjektov,
- Typov údajov,
- Scenára, prípadu použitia, účelu spracovania a súvisiacej potrebnej užitočnosti údajov,

- Cieľov pseudonymizácie a anonymizácie pre konkrétny prípad použitia (komu je potrebné skryť identity, aká je požadovaná užitočnosť odvodených pseudonymov, aká je požadovaná prepojiteľnosť datasetov pre dátovú analýzu, atď.),
- Identifikovaných hrozieb útokov a súvisiacich rizík, ak by boli tieto útoky úspešné, pričom cieľom nemusí byť eliminovať všetky riziká.

Aby údaje v rámci verejnej správy ostali užitočné a použiteľné, neodporúča sa používať techniky anonymizácie, až na dve výnimky:

1. Zdieľanie otvorených údajov, pričom anonymizácii v tomto prípade sa venuje dokument 5.2.1 Pravidlá pre anonymizáciu osobných údajov v otvorených údajoch,
2. V rámci konsolidovanej analytickej vrstvy, kedy sa zozbiera veľký počet pseudonymizovaných datasetov na granulárnej úrovni a nebudú splnené požadované podmienky podľa techník popísaných v kapitole 6.3. Anonymizácii v dátovej vede sa podrobnejšie venuje dokument 4.2.1 Koncept pre zavádzanie analytického spracovania údajov.

Nie všetky techniky pseudonymizácie a nástroje, ktoré ich implementujú, sú rovnako účinné, v súvislosti s každou technikou môžu existovať určité problémy pri implementácii alebo obmedzenia, ako je uvedené v kapitole 6.1.9. Týka sa to nielen výberu samotnej techniky, ale aj celkového návrhu procesu pseudonymizácie, najmä ochrany dodatočných informácií (t. j. informácií, ktoré umožňujú spojenie pseudonymov s pôvodnými identifikátormi). Kombinácia pseudonymizácie s anonymizáciou s ďalšími technikami na zvýšenie ochrany súkromia je tiež veľmi dôležitá na zvýšenie celkovej účinnosti vo vybraných prípadoch použitia, kedy ide o obzvlášť citlivé údaje, ako napríklad identitu a zdravotnícke údaje.

Na výpočet pseudonymu nie potrebné použiť všetky identifikačné údaje. Stačí vykonať výber identifikačných údajov tak, aby osoba mohla byť identifikovaná pri zbere dodatočných údajov, ktoré sa majú pseudonymizovať. Pri zbere údajov sa údaje o identite môžu nahradiť niekoľkými pseudonymami, ktoré sú vypočítané z rôznych atribútov údajov o identite. Proces pseudonymizácie poskytuje prepojiteľné pseudonymy („linkable pseudonyms“), ak sú rovnaké alebo podobné pseudonymy generované pre osoby s rovnakými alebo podobnými identifikačnými údajmi. V tomto prípade sa záznamy údajov môžu zlúčiť pomocou takýchto pseudonymov. Spájateľné metódy sú dôležité pre dlhodobé štúdie, napríklad ak súbory údajov pochádzajú z rôznych zdrojov a majú sa zlúčiť pre účely jednej štúdie. Proces zlučovania pomocou prepojiteľných pseudonymov sa v odbornej literatúre označuje ako spájanie záznamov („record linkage“). Príkladom sú nemecké epidemiologické onkologické registre, ktoré zhromažďujú pseudonymizované datasety o pacientoch s rakovinou s cieľom skúmať úspešnosť rôznych liečebných metód. Údaje pochádzajú od lekárov, z nemocničných informačných systémov a z registrov úmrtí. Niektoré údaje sa vzťahujú na dlhé časové obdobia a môžu dokonca pochádzať z rôznych spolkových krajín, keďže pacienti mohli zmeniť svoje miesto bydliska. Užitočné štúdie v rámci takého registra možno vytvoriť len na základe prepojiteľných pseudonymov.

Za základné techniky pseudonymizácie vzhľadom na ich výhody a nevýhody popísané v kapitole 6.1.9 sa odporúča šifrovanie a hašovacie funkcie s kľúčom alebo so soľou. Odporúčajú sa používať kryptografické nástroje alebo moduly, ktoré prešli nejakou formou

validácie alebo certifikácie, napríklad cez NIST[69] alebo cez FIPS 140-3[70]. **Spoločnou výzvou pre tieto kryptografické techniky je správa kľúčov, ktorá zvyčajne nie je triviálna**, a to aj v závislosti od celkového rozsahu aplikácie, ako aj od konkrétnej zvolenej techniky. Na riešenie tejto výzvy treba okrem overených postupov používať aj overené systémy na manažment kľúčov („Key Management Systems“). Akýkoľvek prístup ku kryptografickému kľúču pre spätnú identifikáciu údajov o identite musí dodržiavať zásadu štyroch očí. To sa dá vyriešiť technicky alebo organizačne. Okrem toho žiadna zo zapojených osôb by nemala mať prístupové práva aj ku kryptografickému kľúču, aj k pseudonymu ako aj k súvisiacim údajom. Ak princíp štyroch očí nie je možné dodržať, aspoň prístup ku kryptografickému kľúču musí byť zaznamenaný jednotlivo.

11.3 Odporúčané politiky

Spoľahlivá stratégia manažmentu a bezpečnosti údajov dokáže poskytnúť flexibilitu, ktorá umožní zohľadniť ľudský rozmer, ktorý je neodmysliteľnou súčasťou niečoho tak kontextuálneho a osobného, ako je súkromie.

Politika správy aktív vrátane správy neštruktúrovaných údajov musí byť súčasťou bezpečnostnej dokumentácie:

- Kategorizácia údajov z pohľadu citlivosti,
- Konkrétne zaradenie údajov podľa citlivosti,
- Klasifikácia údajov z pohľadu ochrany súkromia,
- Konkrétne zaradenie údajov do skupín z pohľadu ochrany súkromia.

Pri zdieľaní údajov s tretími stranami platia dve zásady:

1. Znížiť množstvo údajov,
2. Znížiť presnosť údajov, ku ktorým majú dodávatelia, partneri a ďalšie tretie strany prístup.

V prípade interných zainteresovaných strán môžete mať väčšiu istotu, že kontroly prístupu a kontroly auditu dokážu zmierniť následky narušenia súkromia. V prípade externých partnerov je možné, že majú prístup k iným údajom, ktoré by mohli jednoznačne identifikovať používateľa, a že ich schopnosť kontrolovať, kto má prístup k údajom a ako ich používa, môže byť neoptimálna. Preto namiesto toho, aby ste sa spoliehali výlučne na kontrolu prístupu, treba údaje zastrieť tak, aby aj v prípade, že by k nim niekto získal prístup, a aj keby ich skombinoval s inými údajmi, bolo možné obmedziť narušenie súkromia.

Najlepším prístupom k pseudonymizácii, pomocou ktorého sa možno spoľahlivo brániť proti typom útokov na pseudonymizáciu, spomínaných v kapitole 8.3, je:

- Zvážte celý dostupný dataset.
- Posúďte informácie o veľkosti vstupnej domény jednotlivých hodnôt údajov.

[69] Zdroj: https://csrc.nist.gov/Projects/cryptographic-module-validation-program/validated-modules, Dátum referencie: 06.04.2023

[70] Zdroj: https://www.wolfssl.com/license/fips/, Dátum referencie: 06.04.2023

- Pseudonymizáciu aplikujte na všetky hodnoty údajov takým spôsobom, aby sa útoky pomocou hrubej sily a slovníkové útoky stali neuskutočniteľnými.
- Odstráňte akúkoľvek možnosť útokov na základe znalostí o pozadí alebo pravdepodobnostného rozdelenia.
- Navrhnite výslednú funkciu pseudonymizácie veľkého rozsahu tak, aby sa v pseudonymizovanom datasete zachoval len typ užitočnosti údajov potrebný na účel spracovania, ale aby sa z pseudonymizovaného datasetu odstránila všetka ostatná užitočnosť. Snažte sa vybrať vždy tú najbezpečnejšiu politiku podľa kapitoly 7.3, ak plní účel spracovania.

Nasledujúca tabuľka sumarizuje v kontrolnom zozname, čo všetko je potrebné zohľadniť v komplexnej politike ochrany súkromia údajov.

Tabuľka 23: Kontrolný zoznam pre štandardy anonymizácie a pseudonymizácie

Zoznam dobrej praxe	Nevyhnutné implementovať	Poznámka
Umožňuje prístup založený na riziku, ktorý zohľadňuje požadovanú ochranu a užitočnosť/škálovateľnosť	✓	
Rozvíja súčasný stav poznania („state of the art“)	✓	
Je v súlade s definíciou pseudonymizácie podľa GDPR	✓	
Využíva jednu alebo viacero funkcií pseudonymizácie	✓	
Využíva pseudonymizačné tajomstvo, ktoré dôsledne chráni	✓	
Má funkciu obnovy pre funkcie pseudonymizácie		Môžu nastať scenáre a prípady použitia, kedy sa dodatočné informácie ako kľúče, bezpečne zničia.
Používa tabuľku mapovania pseudonymizácie		Nemusí platiť pre všetky scenáre a prípady použitia.
Maximalizácia užitočnosti a ochrany údajov	✓	
Osobné identifikátory nahradené pseudonymami	✓	

Zoznam dobrej praxe	Nevyhnutné implementovať	Poznámka
Pseudonymy neumožňujú priame odvodenie osobných identifikátorov	✓	
Osobné údaje už nie je možné priradiť ku konkrétnej dotknutej osobe bez použitia dodatočných informácií	✓	
Zrušenie pseudonymizácie je netriviálne pri absencii dodatočných informácií	✓	
Dodatočné informácie sa uchovávajú oddelene s použitím technických a organizačných kontrol na obmedzenie prístupu	✓	
Pseudonymy použité na priame a nepriame identifikátory	✓	
Odolnosť voči útokom na odhalenie pseudonymizačného tajomstva	✓	
Odolnosť voči útokom hrubou silou	✓	
Odolnosť voči vyhľadávaniu v slovníku	✓	
Odolnosť voči spätnej identifikácii prostredníctvom vyčlenenia	✓	
Odolnosť voči spätnej identifikácii prostredníctvom útokov na prepojenie		Závisí od konkrétneho scenára a prípadu použitia.
Odolnosť voči spätnej identifikácii prostredníctvom útokov na inferenciu		Závisí od konkrétneho scenára a prípadu použitia.
Anonymizačné techniky používané na ďalšie zníženie možnosti odvodenia identity tretími stranami	✓	
Výsledkom jedného vstupu je oddelená dvojica výstupov: pseudonymizované údaje a dodatočné informácie potrebné na spätnú identifikáciu		Nemusí platiť v prípade šifrovania (kapitola 6.1.5).
Identifikácia dotknutej osoby skrytá v kontexte konkrétnej operácie spracovania údajov	✓	

Zoznam dobrej praxe	Nevyhnutné implementovať	Poznámka
Žiadny príjemca alebo tretia strana, ktorá má prístup k pseudonymizovaným údajom, nemôže triviálne odvodiť pôvodný dataset a totožnosť dotknutých osôb	✓	
Podpora neprepojiteľnosti v rôznych oblastiach spracovania údajov		Existujú výnimky v rámci analytického spracovania v Konsolidovanej analytickej vrstve (kapitola 10.3.1)
Podpora presnosti zachovaním prístupu k pseudonymizovaným výstupom aj k dodatočným informáciám potrebným na spätnú identifikáciu		Môžu nastať scenáre a prípady použitia, kedy sa dodatočné informácie ako kľúče, bezpečne zničia.
Na generovanie pseudonymov sa nepoužívajú počítadla ani hašovanie bez kľúča alebo soli	✓	
Najčastejšie sa využíva hašovacia funkcia s kľúčom (HMAC, SHA2/3, 256+ bitové kľúče) na generovanie pseudonymov	✓	
Alternatívne sú k dispozícii kryptografický generátor pseudonáhodných čísel a tokeny (náhodne generované hodnoty) ako pseudonymy		Pre osobitné scenáre a prípady použitia (kapitola 6.1.2 a 6.1.6)
Správny výber politiky pseudonymizácie podľa kapitoly 7, ktorá je čo najbezpečnejšia	✓	
V prípade potreby sú dostupné pokročilé techniky pseudonymizácie (kapitola 6.1.8)		
Riadená prepojiteľnosť pseudonymov	✓	
Dostupné techniky anonymizácie podľa kapitoly 6.1.9	✓	
Dostupné techniky merania podľa kapitoly 6.3	✓	
Pseudonymizácia zachovávajúca prefixy a sufixy		

Zoznam dobrej praxe	Nevyhnutné implementovať	Poznámka
Pseudonymizácia zachovávajúca formát		

12 Použitá literatúra

[1.] Chatzistefanou and K. Limniotis, "On the (non-)anonymity of anonymous social networks", E-Democracy – Privacy-Preserving, Secure, Intelligent E-Government Services, Communications in Computer and Information Science, Springer, vol. 792, strany 153-168, 2017

[2.] Su, J., Shukla, A., Goel, S. and Narayanan, A. (2017) 'De-anonymizing web browsing data with social networks', WWW '17, strany 1261–1269, 2017.

[3.] A. Kurtz, H. Gascon, T. Becker, K. Rieck and F. C. Freiling, "Fingerprinting Mobile Devices Using Personalized Configurations", PoPETs 2016 (1), strany 4-19, 2016.

[4.] X. Zhou, S. Demetriou, D. He, M. Naveed, X. Pan, X. Wang, C. A. Gunter, and K. Nahrstedt, "Identity, location, disease and more: Inferring your secrets from Android public resources", ACM CCS 2013.

[5.] A. J. Menezes, S. A. Vanstone, and P. C. V. Oorschot, Handbook of Applied Cryptography, CRC Press, 1996.

[6.] L. Demir, A. Kumar, M. Cunche and C. Lauradoux, "The pitfalls of hashing for privacy", IEEE Communications Surveys and Tutorials, vol. 20, č. 1. strany 551-565, 2018.

[7.] D. J. Bernstein and T. Lange, „Post-quantum cryptography – dealing with the fallout of physics success", Cryptology ePrint Archive, 2017.

[8.] J. Lehnhardt and A. Spalka, "Decentralized Generation of Multiple, Uncorrelatable Pseudonyms without Trusted Third Parties", In: Furnell S., Lambrinoudakis C., Pernul G. (eds.), Trust, Privacy and Security in Digital Business (TrustBus) 2011, Lecture Notes in Computer Science, vol. 6863, strany. 113-124, Springer, Berlin, Heidelberg, 2011.

[9.] Judith Sáinz-Pardo Díaz, Álvaro López García, „A Python library to check the level of anonymity of a dataset", Nature, Scientific Data vol. 9, číslo článku: 785, 2022, link

[10.] Nishant Bhajaria,"Data Privacy", Manning Publications, 2021

[11.] ENISA report: „DATA PSEUDONYMISATION: ADVANCED TECHNIQUES & USE CASES", Január 2021

[12.] ENISA: „Recommendations on shaping technology according to GDPR provisions: An overview on data pseudonymisation", November 2018

[13.] ENISA: „Pseudonymisation techniques and best practices: Recommendations on shaping technology according to data protection and privacy provisions", November 2019

[14.] European Data Protection Board: „Guidelines 4/2019 on Article 25 Data Protection by Design and by Default Version 2.0", prijaté 20. októbra 2020, link

[15.] Rolf Schwartmann / Steffen Weiß (Ed.):"Requirements for the use of pseudonymisation solutions in compliance with data protection regulations," A working paper of the Data Protection Focus Group of the Platform Security, Protection and Trust for Society and Business at the Digital Summit 2018, link

[16.] Rolf Schwartmann / Steffen Weiß (Ed.): " White Paper on Pseudonymization Drafted by the Data Protection Focus Group for the Safety, Protection, and Trust Platform for Society and Businesses in Connection with the 2017 Digital Summit, link