Katedra Informatiky
Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského, Bratislava

Anonymizácia a ochrana dát

(Diplomová práca)

Bc. Peter Juhász

Evidenčné číslo : 866e5284-d7e1-4ce1-a9fb-961e6a8379b5

Vedúci: RNDr. Michal Malý

Bratislava, 2011

Katedra Informatiky
Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského, Bratislava

Anonymizácia a ochrana dát

(Diplomová práca)

Bc. Peter Juhász

Študijný program: Informatika

Študijný odbor: 9.2.1 Informatika

Školiace pracovisko: Katedra informatiky

Školiteľ: RNDr. Michal Malý

Miesto a rok predloženia: Bratislava, 2011

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Peter Juhász
Študijný program: informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.2.1. informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský

Názov: Anonymizácia a ochrana dát

Cieľ: Ciel:
Praca ma praktický a teoretický cieľ. Prakticky ciel je vytvorit uzivatelsky pristupny nastroj na anonymizaciu dat (najma s aplikaciou na databazove tabulky) a generovanie virtualnych dat. Teoreticky ciel je popisat moznosti a obmedzenia anonymizacie dat, zadefinovat a zaviest mieru anonymizacie (entropia) a na praktickych prikladoch tuto mieru spocitat a vyhodnotit vysledky.
Oba ciele sa maju okrem vseobecneho pristupu aspon scasti zaoberat podmienkami v SR.

Podrobny rozpis cielov:
Diplomant si nastuduje relevantne clanky o anonymizacii dat. Uvedie dovody na anonymizaciu dat, a to najma legislativny (sulad s predpismi) ramec a tiez trhovy ramec (dobre meno firmy). Spomenie najvaznejsie priklady uniku dat v zahranici a na Slovensku (uniknute zoznamy telefonnych operatorov prip. ine) a prakticke pouzitie anonymizovanych dat (testovanie aplikacii). Prehladovo spomenie platnu legislativu SR a to zakon 428/2002 Z.z. o ochrane osobných údajov, ktory definuje co su to osobne udaje. Zamysliet sa nad tym, co znamena identifikacia osoby z pohladu teorie informacie (zhruba: kolko bitov informacie potrebujem dodat, aby som urcil presnu konkretnu osobu na zaklade dostupnych dat).
Vytvoreny nastroj bude umoznovat rozne formy anonymizacie dat (data scrambling): nulovanie (NULL), jednoduche nahradenie ("Jozef" -> "Meno"), dynamicke nahradenie (Jozef -> "Meno1", "Fero" -> "Meno2"), permutacia, zasumenie alebo perturbacia (napr. gausovske zasumenie pre cislo x -> x*randnorm(1,c)), odrezanie ("Kovacova" -> "Kov"), hashovanie a dalsie. Nastroj tiez umozni generovanie "vierohodne" vyzerajucich, no uplne nahodnych osobnych dat v rozsahu najmenej Meno, Priezvisko, Ulica a cislo, Mesto, PSC, datum narodenia, rodne cislo, pohlavie, okres ("Michal Kovac, Liscie udolie 47/2, Bratislava 4, 84104, 10.5.1972, 720510/9373 , MUZ, Bratislava 4").
Na zaklade teoretickych clanokov a zavedenej miery sa zamysliet nad tym, ake su obmedzenia pre anonymizaciu (kedy a do akej miery je mozne rekonstruovat povodne data). Demonstrovat funkcnost aplikacie na konkretnych datach a pre tieto data spocitat zavedenu mieru.

Je tu moznost aj popisat komercny uzavrety nastroj DataMasker.

Diplomant zaroven moze prakticky preskumat aktualny stav ochrany dat (jeho osobnych udajov) pomocou zaslania ziadosti pre rozne organizacie -- banku, telefonneho operatora, a pod., a vyhodnotit, kolko dat, v akej forme a na aky ucel a z akeho zdroja je o osobe zhromazdovanych.

Literatúra: http://www.eife-l.org/publications/eportfolio/proceedings2/ep2007/papers/digital-identity-and-privacy/balancing-privacy-and-information-utility-in-data-anonymisation-1

zakon 428/2002
smernica EU 95/46/ES

Vedúci: Mgr. Michal Malý
Katedra: FMFI.KAI - Katedra aplikovanej informatiky

Dátum zadania: 18.11.2010

Dátum schválenia: 07.12.2010

prof. RNDr. Branislav Rovan, PhD.
garant študijného programu

.....
študent

.....
Vedúci

Čestne prehlasujem, že som túto diplomovú prácu vypracoval samostatne s použitím citovaných zdrojov.

.....

Chcem sa poďakovať vedúcemu mojej diplomovej práce RNDr. Michalovi Malému za cenné rady a pomoc pri písaní tejto práce.

Abstrakt

Názov práce: Anonymizácia a ochrana dát

Autor: Bc. Peter Juhász

Vedúci práce: RNDr. Michal Malý

Anonymizácia dát je jednou z metód, ktoré slúžia na ochranu osobných údajov v databázach. Táto metóda má využitie hlavne na miestach, kde nie je nutné pracovať s reálnymi dátami, pri vývoji softvéru alebo testovaní. Práca sa venuje problematike osobných údajov a ich ochrane, bližšie popisuje pojem anonymizácie, spôsoby, možnosti a dôvody využitia v praxi, definuje a zavádza mieru anonymizácie, ktorá rozhoduje, či konkrétna sada údajov postačuje na jednoznačnú identifikáciu osoby. Súčasťou práce je implementácia aplikácie na anonymizáciu databázových tabuliek, ktorá je použiteľná v praxi. V práci sme skúmali a zaoberali sa komerčnou aplikáciou DataMasker, ktorá slúží na rovnaký účel.

Kľúčové slová: Anonymizácia dát, osobné údaje, ochrana osobných údajov

Abstract

Title: Anonymizácia a ochrana dát

Author: Bc. Peter Juhász

Supervisor: RNDr. Michal Malý

Anonymisation of data is a method that is used to protect personal data in databases. This method helps mainly in places, where there is no need to work with real data, software development or testing. Thesis deals with the problems of personal data, privacy policy, describes the concept of anonymisation, methods, possibilities and reasons for use in practice, defines and establishes a level of anonymisation, which decides whether a specific set of data is sufficient to uniquely identification a person or not. The work contains an application for anonymisation database tables, which is applicable in practice. In this work, we investigated and dealt with the commercial application DataMasker, which is used for the same purpose.

Keywords: Data anonymisation, data protection, personal data, privacy policy.

Obsah

1 Úvod

S pojmom osobných údajov a ich ochranou sa stretávame stále viac a viac. Je to v súčasnosti dosť často skloňovaný pojem, či už na univerzitnej pôde, v médiách alebo aj v bežnom živote. Keď sa zamyslíme a spočítame, koľko rôznych inštitúcii eviduje o našej osobe informácie ľubovoľného charakteru, možno budeme prekvapení. Od citlivých informácií, ktoré archivujú zdravotnícke subjekty, ktoré sme za roky mnohokrát navštívili, cez mobilných operátorov, ktorí o nás majú svedomito uložené informácie kedy a s kým sme boli v kontakte, cez banky a poisťovne, až po internetové služby, ktoré v evidencii zaujímavých údajov o nás nezaostávajú. Už len samotný google, ktorý, dovolím si povedať, je neodmysliteľnou súčasťou nášho každodenného života, vie o nás viac, ako by možno mal vedieť. O problémoch, ktoré nás trápia a „googlime“ riešenia, cez plánovanie dovolenky, hľadanie darčekov, ľudí, práce, zábavy... Vie, na čo myslíme, čo potrebujeme vedieť. Rozmanité sociálne siete, na čele s Facebookom, ani nebudeme spomínať.

Okoliu sme ochotní poskytnúť len málo z týchto údajov a tak by to aj malo zostať. Stáva sa však, či už úmyselne, nedbalosťou alebo zanedbaním povinností, že tieto informácie sa dostanú do rúk, kam nepatria. Znásobené súčasnými možnosťami a silou Internetu sú tieto informácie okamžite rozdistribuované do celého sveta a cesty späť už, žiaľ, niet. Čo sa raz dostane na Internet, už z neho nikdy nezmizne. Smelo to môžeme povedať o telefónnych zoznamoch dvoch najväčších slovenských operátorov, ktorých údaje sú dodnes zálohované na diskoch mnohých zvedavcov. Ani zďaleka sa nedá všetkým takýmto únikom zabrániť, ale používaním anonymizácie, o ktorej pojednáva táto diplomová práca, sa môžeme aspoň pokúsiť znížiť počet únikov citlivých dát do sveta. Už len zavedenie jednoduchých procesov, ktoré

zamedzia brigádnikom pracujúcim na dohodu prvý deň v práci v prístupe k produkčným dátam a osobným údajom tisícok ľudí, je výrazný krok vpred. Možno na prvý pohľad smiešna požiadavka, ale rýchlym prieskumom medzi známymi sa dozvieme, že ani vo veľkých a na prvý pohľad dobre organizovaných spoločnostiach to nie je samozrejmosťou.

Moderným trendom v súčasnosti je vyvíjanie informačných systémov na mieru prostredníctvom externých spoločností. Častokrát je pri tomto vývoji potrebná databáza zadávateľa, na ktorú sa "šije" celý nový IS, skúša a testuje sa. Podľa môjho názoru nie je dôvod, prečo by externá spoločnosť mala robiť svoje testovania na reálnych citlivých dátach, keď to môžu robiť s databázou na nerozoznanie od tej reálnej, ktorá má rovnakú informačnú hodnotu, ale nenachádza sa v nej jediný skutočný citlivý údaj, ktorý by mal pravú hodnotu a dal sa zneužiť. Používanie anonymizácie má význam a táto diplomová práca ponúka nástroj, ktorý dokáže anonymizovať celú databázu údajov a aspoň v malej miere prispieva k ochrane osobných údajov na miestach, kde sa táto metóda dá využiť.

V prvej časti tejto práce sa budeme venovať teoretickému úvodu do problematiky, vysvetlíme pojmy, ako sú osobné údaje a ich ochrana, anonymizácia a dôvody anonymizácie, vysvetlíme pojem identifikácie - čo znamená byť identifikovaný a identifikovateľný.

V druhej časti si bližšie popíšeme proces anonymizácie a zavedieme mieru anonymizácie.

V tretej časti popíšeme hlavnú časť tejto práce - samotnú implementáciu problému, ktorej výstupom bude aplikácia na anonymizáciu dát.

2 Úvod do problematiky

Na úvod zadefinujeme základné pojmy, ako sú osobné údaje a anonymizácia a vysvetlíme, prečo a z akých dôvodov je dôležitá ochrana osobných údajov.

2.1 Osobné údaje

Definícia pojmu osobné údaje je zakotvená v našej legislatíve v zákone č. 428/2002 Z.z. o ochrane osobných údajov, ktorý bol v značnej miere inšpirovaný Smernica 95/46/ES Európskeho parlamentu a rady z 24. októbra 1995 o ochrane jednotlivcov, vzhľadom na spracovávanie osobných údajov a o voľnom pohybe takýchto údajov[6]. Tento zákon definuje osobné údaje nasledovne: „Osobnými údajmi sú údaje týkajúce sa určenej alebo určiteľnej fyzickej osoby, pričom takou osobou je osoba, ktorú možno určiť priamo alebo nepriamo, najmä na základe všeobecne použiteľného identifikátora alebo na základe jednej či viacerých charakteristík alebo znakov, ktoré tvoria jej fyzickú, fyziologickú, psychickú, mentálnu, ekonomickú, kultúrnu alebo sociálnu identitu.“ [2]

Podľa výkladu [3] údajom sa rozumie jednotlivý fakt a osobným údajom je súbor údajov, teda súbor jednotlivých faktov, na základe ktorých je osoba určiteľná alebo určená. Rozhodujúcim kritériom pre tvrdenie, že určitý súbor údajov je možné považovať za osobné údaje je skutočnosť, či je možné dotknutú osobu na základe týchto údajov určiť (identifikovať) priamo alebo nepriamo.

2.1.1 Relatívnosť pojmu

Takto definovaný pojem osobných údajov je teda vcelku relatívny. Ako príklad by sme mohli uviesť údaj – priezvisko – 'Kováč'. Toto priezvisko nie je samo o sebe osobný údaj, keďže existuje veľa osôb, ktoré majú dané priezvisko. Toto priezvisko teda ani priamo, ani nepriamo neurčuje jednu konkrétnu osobu, ktorú by sme mohli len na základe priezviska jednoznačne identifikovať. Osobným údajom sa však už stáva v rámci užšej skupiny osôb, akou je napríklad trieda v škole alebo menšia obec. V takejto užšej komunite je menej pravdepodobné, že existuje viac ako jedna osoba s priezviskom 'Kováč'.

Ešte zaujímavejším údajom, ktorý zdôrazňuje túto relatívnosť, môže byť, ako uvádza zákon, kultúrna alebo sociálna identita, kde údaj o osobe v rámci jednej, hoci aj obrovskej society, nemusí hrať rolu a byť osobným údajom, pričom v inej societe môže jednotlivca jednoznačne odlišovať a priamo ho identifikovať. Ak by na Slovensku žil jediný obyvateľ, ktorý vyznáva náboženstvo džinizmus, bol by na základe príslušnosti k danému náboženstvu jednoznačne identifikovateľný len na základe tohto jediného údaju. Namiesto rodného čísla alebo štvorice údajov - meno, priezvisko, adresy trvalého pobytu a dátumu narodenia by ho takto dokázala jednoznačne identifikovať len informácia : 'Ten, ktorý vyznáva džinizmus'.

Teda niekedy, aj na základe na prvý pohľad málo významnej informácie, dokážeme, či už priamo alebo nepriamo, identifikovať jednotlivca.

2.1.2 Rôznorodosť osobných údajov

Zaužívaným názorom v súčasnej dobe je pokladať za osobné údaje len úzku množinu dát, akými sú meno, priezvisko, adresa, rodné číslo. Osobným údajom však môže byť množstvo iných údajov, ktoré by sme na prvý pohľad

osobnými nenazvali.

Jedným z údajov, nad ktorým stojí zamyslieť sa, je telefónne číslo. Telefónne číslo jednoznačne identifikuje osobu, ktorej dané číslo patrí, na ktorú je konkrétne číslo u telefónneho operátora zaregistrované. Dalo by sa teda z tohto hľadiska pokladať za osobný údaj. Na druhej strane, ak máme k dispozícii telefónne číslo, vieme pomocou neho identifikovať konkrétnu osobu? Tu narážame na problém. Museli by sme (ručne) prelistovať celý telefónny zoznam, aby sme dané číslo našli. V tom horšom prípade ešte môže byť dané číslo utajené, keď užívateľ požiadal svojho telefónneho operátora o nezverejnenie telefónneho čísla.

Táto identifikácia je však už časovo náročná, spomenuté vyhľadávanie v tlačenom zozname by zabralo veľmi veľa času. Ďalším faktorom pri rozhodovaní, či je daný údaj osobným údajom, sa teda stáva aj časová náročnosť, vykonateľnosť, resp. výpočtový výkon, ktorý máme k dispozícii.

Riešením tohto problému by bola služba, ktorá by dokázala ku konkrétnemu telefónnemu číslu vrátiť meno užívateľa. Takáto služba na Slovensku vo všeobecnosti neexistuje, jediný operátor, ktorý ponúka takúto možnosť je Slovak Telekom na čísla pevných liniek. Takéto reverzné zisťovanie v porovnaní s existujúcimi telefónnymi zoznamami, aké existuje napríklad v USA, by teda urobilo z telefónneho čísla osobný údaj.

Ďalším takýmto novodobým údajom je e-mailová adresa. Tieto adresy zväčša patria jednotlivcovi, ktorý vlastní danú adresu. Pri zmesi alfanumerických znakov, ako niektoré adresy vyzerajú, je ťažké identifikovať jednotlivca. Avšak pri firemných mailoch, kde si to firemná politika často vyžaduje alebo pri štátnej správe, majú e-mailové adresy častokrát definovaný formát. Obyčajne to býva 'meno.priezvisko@organizácia.sk'. A z takéhoto formátu adresy toho už vieme vyčítať viac než dosť. Okrem celého mena je to aj

meno zamestnávateľa. Tiež s prihliadnutím na to, že pravdepodobnosť zamestnania dvoch ľudí s rovnakým menom a priezviskom v jednej firme je dosť malá, tieto dva údaje bez väčších problémov dokážu jednoznačne identifikovať osobu. V takýchto špecifických prípadoch teda môžeme aj e-mailovú adresu považovať za osobný údaj.

Posledným príkladom, ktorý uvedieme, je číslo občianskeho preukazu. O tomto údaji by sa dalo opäť polemizovať, odpoveď nám však dáva takýto výklad zákona [13] : „Na určenie dotknutej osoby na základe čísla občianskeho preukazu by bolo potrebné získať ďalšie osobné údaje. Na ich získanie by musela fyzická osoba, ktorá nie je oprávnenou osobou, vynaložiť neúmerne vysokú námahu, nakoľko kompletnú databázu vydaných občiansky preukazov vedie Policajný zbor a databáza je prístupná výlučne oprávneným osobám prevádzkovateľa. Samotné číslo občianskeho preukazu preto nie je možné považovať za osobný údaj dotknutej osoby.“

Aj na týchto príkladoch vidieť, že vo všeobecnosti určiť, či daná množina údajov je alebo nie je osobnou, je dosť náročné. Vyžaduje si to v každom prípade dôkladné posúdenie a zamyslenie.

2.1.3 Vlastný výskum osobných údajov

Jedným z cieľov tejto práce bolo aj vykonanie prieskumu, čo považujú za osobné údaje bežné organizácie, akými sú zdravotná poisťovňa, sociálna poisťovňa, mobilný operátor alebo škola. Za týmto účelom sme vytvorili žiadosť (Príloha 1), kde sme na základe §20 ods. 1 písm. b), c) zákona 428/2002 Z.z. o ochrane osobných údajov žiadali poskytnutie výpisu všetkých údajov o mojej osobe, ktoré organizácia spracúva, ako aj presné informácie o zdroji, z ktorého boli tieto údaje získané. Túto žiadosť sme rozposlali spomenutým inštitúciám a z odpovedí, ktoré prišli, sme zistili nasledovné.

Vo všeobecnosti tieto inštitúcie za osobné pokladajú len tie údaje, ktoré sme uviedli pri zahájení spolupráce. Jednak údaje pri podpise zmluvy o pripojení pri telefónnych operátoroch (Príloha 3), údaje vyplnené na prihláške na vysokú školu a údaje z iných vyplnených registračných formulárov a žiadostí. Odpoveďou na výpis osobných údajov boli väčšinou kópie týchto zmlúv so základnými údajmi, ako sú meno, priezvisko, adresa trvalého pobytu, číslo občianskeho preukazu a rodné číslo. Súhlas so spracovaním týchto osobných údajov na špecifikované účely sme potvrdili vlastnoručným podpisom pri uzatváraní týchto zmlúv, resp. prihlášok. Podpisom sme v niektorých prípadoch takisto dali osobitný súhlas na spracovanie rodného čísla na osobitné účely jednoznačnej identifikácie v informačných systémoch. Tieto základné údaje, poskytnuté a podpisom potvrdené mojou osobou, boli teda informácie a údaje, ktoré sme viac-menej automaticky očakávali a ktoré sa v bežnom ponímaní považujú za osobné údaje.

Osobnými údajmi sú však, napríklad v prípade telefónneho operátora, aj výpisy hovorov, správ, platieb, dátových prenosov a mnoho iných informácií. Tieto informácie operátor každopádne vo svojej databáze eviduje a pri žiadosti o výpis všetkých údajov o mojej osobe ich mal na základe spomínaného zákona poskytnúť. Operátor má samozrejme právo zhromažďovať a uchovávať tieto údaje pre svoje účely, na základe zmluvy. Jedná sa o údaje, ktorých zdrojom je síce samotný operátor, ale na základe mojej činnosti - sú to teda informácie o mojej osobe. Tieto údaje nám však v odpise neboli poskytnuté.

Pri žiadosti o odpis osobných údajov od vysokej školy nám bolo podané, po vzájomnej dohode, ústne vysvetlenie a preukázanie údajov o mojej osobe. Poskytnutými údajmi boli jednak údaje z prihlášky na vysokoškolské štúdium, kde som podpisom dal súhlas na spracúvanie a evidovanie týchto

údajov na konkrétne účely, ale aj údaje zozbierané školou počas celého doterajšieho štúdia. Jednak kompletné údaje ohľadom ukončeného bakalárskeho štúdia, aktuálny výpis známok, prihlášky na štúdium, rodný list, zadania záverečných prác a iné. V tomto prípade boli teda za osobné informácie naozaj pokladané všetky zhromaždené informácie o mojej osobe, ktoré škola eviduje a uchováva. Odpoveď na túto žiadosť by sme mohli označiť za kompletnú, plnohodnotnú v rámci našich požiadaviek a v zmysle citovaného zákona.

V ďalších výpisoch a odpovediach, ktoré sme dostali, však prevažovali už len očakávané informácie, ktoré sa týkali údajov poskytnutých mojou osobou (Príloha 2). Kópie zmlúv, ktoré vznikli pri uzatváraní spolupráce s danou inštitúciou. Za osobné informácie teda organizácie vo všeobecnosti nepovažujú dáta získané, zozbierané v priebehu spolupráce, vlastnou cestou, informácie, ktorých zdroj je iný, ako moja osoba. Pri takýchto žiadostiach by mali poskytnúť všetky údaje, ktoré evidujú o žiadateľovi vo svojich informačných systémoch, čo však na základe tohto jednoduchého prieskumu na malej vzorke organizácií nebolo vo väčšine prípadov splnené.

Ospravedlnením môže byť prácnosť získavania a vypracovania takéhoto podrobného výpisu, ktorý napríklad pri mobilnom operátorovi môže zahŕňať naozaj množstvo rôznorodých dát alebo pri banke a kompletnom výpise pohybov na účte množstvo strán údajov. Pri bežnom odpise osobných údajov, aký sme žiadali, sú asi vo väčšine prípadov postačujúcou odpoveďou žiadateľovi tieto údaje a verím, že pri žiadosti o podrobnejší odpis konkrétne vymenovaných údajov o mojej osobe by sme dostali očakávanú odpoveď.

2.1.4 Rodné číslo

Často diskutovaným pojmom ohľadom osobných údajov je rodné číslo. O tomto identifikátore bližšie hovorí zákon Národnej rady Slovenskej republiky

č. 301/1995 Z. z. o rodnom čísle takto : „Rodné číslo je trvalý identifikačný osobný údaj fyzickej osoby, ktorý zabezpečuje jej jednoznačnosť v informačných systémoch. “[4] Rodné číslo je samo o sebe osobným údajom a aj preto sa na neho vzťahujú osobitné ustanovenia. „Pri spracúvaní osobných údajov možno využiť na účely určenia fyzickej osoby všeobecne použiteľný identifikátor ustanovený osobitným zákonom len vtedy, ak jeho použitie je nevyhnutné na dosiahnutie daného účelu spracúvania. Spracúvať iný identifikátor, ktorý v sebe skrýva charakteristiky dotknutej osoby alebo zverejňovať všeobecne použiteľný identifikátor sa zakazuje.“ [4]

Aj keď je rodné číslo zákonom definované ako jednoznačný identifikátor osôb v informačných systémoch, podľa Stanoviska Úradu na ochranu osobných údajov Slovenskej republiky k zverejňovaniu rodných čísel fyzických osôb z 15. marca 2006 : „Konštatovanie, že len rodné číslo zabezpečí jednoznačnosť fyzickej osoby a bez jeho uvedenia nemožno osobu v rámci daného teritória presne identifikovať, resp. že v takomto prípade nemožno bez zverejnenia rodného čísla dosiahnuť daný účel spracovania, je nenáležité. Je to paradoxné, ale podľa údajov Ministerstva vnútra SR aj v súčasnosti na Slovensku ešte stále existuje približne 30000 duplicitných rodných čísel, čo v 'preklade' znamená, že je tu minimálne 60000 dotknutých osôb, u ktorých nie je istota, že im na základe rodného čísla bude priradená správna identita.“ [5]

S touto duplicitou rodných čísel sa rátalo už pri spomínanom návrhu zákona č. 301/1995 Z. z., kde v §ods. 5 sa uvádza : „Prevádzkovateľ je povinný požiadať ministerstvo o overenie rodného čísla, ak pri zaraďovaní údajov o osobe do informačného systému alebo inak zistí, že rodné číslo nie je v súlade s §2.“ [4]

Riešením tohto stavu je, okrem stále prebiehajúcich náprav, aj globálne

riešenie MV SR, ktoré by do budúcna zabránilo takýmto stavom. Jedná sa o Jednoznačný identifikátor fyzických osôb (JIFO), ktorý bude alfanumerickou množinou znakov pre každú osobu evidovanú v registri obyvateľov. Na rozdiel od rodného čísla by tento identifikátor nemal obsahovať žiadnu informáciu, z ktorej by bolo možné odvodiť údaj charakterizujúci určenú osobu, napríklad pohlavie, vek alebo rasu.

2.2 Anonymizácia dát

Anonymizácia dát (v angličtine data anonymization, data scrambling alebo data masking, v tejto práci však budeme kvôli jednotnosti používať výlučne pojem anonymizácia dát) je hlavnou témou tejto práce. Pokúsime sa bližšie vysvetliť tento pojem. Anonymizáciou dát by sa dal nazvať ľubovoľný proces, ktorý mení a pretvára vstupné dáta akýmkoľvek spôsobom s cieľom sťažiť, resp. zamedziť útočníkovi extrahovať ľubovoľné údaje o určitej osobe. Má znemožniť útočníkovi, či už priamo alebo vylučovacou metódou, z ľubovoľnej množiny dát získať hocijaké reálne informácie, ktoré by vedel zneužiť. Jedna z definícií, ktorá exaktnejšie popisuje tento pojem : „Anonymizácia je taká zmena osobných údajov, po ktorej už tieto osobné údaje nemožno prideliť určitému zistiteľnému jednotlivcovi alebo tak možno urobiť len s vynaložením nepomerne veľkého úsilia z hľadiska času, nákladov a práce.“

Tento proces anonymizácie, ktorý mení konkrétne dáta v databáze pritom nemusí byť žiaden rafinovaný algoritmus. Pri anonymizovaní dát, ktoré sú typu *String*, resp. *Varchar* sa používa už spomenuté jednoduché nahradenie konštantným reťazcom, nahradenie hodnotou NULL, dynamické nahradenie rôznymi reťazcami, ľubovoľné operácie na reťazcoch, prehadzovanie dát alebo nahradenie dát predpripravenými virtuálnymi dátami.

3 Anonymizácia dát

3.1 Metódy anonymizácie dát

V tejto časti si bližšie popíšeme konkrétne metódy ako anonymizovať dáta.

3.1.1 Jednoduché nahradenie

Už pri letmom zamyslení si uvedomíme,že pre dvojicu údajov, ako je meno a priezvisko, správne realizované triviálne nahradenie konštantnými reťazcami 'Jozef' a 'Novák' je dokonalá, útočníkom neprelomiteľná anonymizácia. Správna realizácia je však už pri tomto jednoduchom procese na mieste, keď, ako neskôr bližšie popíšeme, pri nenormalizovaných dátach v tabuľke by nás rýchlo mohol zradiť stĺpec 'CeleMeno'. Informačná hodnota, ktorú sa pri procese anonymizácie snažíme maximalizovať, síce bude veľmi nízka, naopak, ochrana dát bude na vysokej úrovni.

3.1.2 Nullovanie

Pri ešte triviálnejšom nahradení skutočných dát hodnotou 'NULL' dostávame z hľadiska kvality anonymizácie veľmi bezpečné nahradenie. Informačná hodnota je však ešte menšia ako bola minimálna v predchádzajúcom príklade. V prípade vývoja informačného systému, kde sa napríklad editujú údaje klientov, bude pri testovaní s takto anonymizovanými dátami položka 'Meno' a 'Priezvisko' prázdna. Toto určite nie sú dobré testovacie dáta, keď programátor pri vývoji aplikácie môže v takomto prípade márne hľadať chybu vo svojich kódoch, hľadajúc dôvod, prečo sa daná kolónka nevypĺňa.

3.1.3 Virtuálne dáta

Používanie virtuálnych dát pri anonymizácií by sa dalo pokladať za osobitú metódu. Táto metóda, keď skutočné dáta nahradzujeme virtuálnymi, reálne vyzerajúcimi dátami, je jednou z najefektívnejších metód anonymizácie dát. Táto metóda si však pri použití v praxi vyžaduje vopred vytvorené bázy dát, z ktorých čerpá virtuálne údaje. Pre každý údaj, ako je napr. ženské krstné meno, mužské krstné meno, ženské priezvisko, mužské priezvisko, obec, ulica, sa tak vyžaduje mať vopred vytvorený zoznam s konkrétnymi položkami. Pri anonymizácií konkrétneho údaju si užívateľ v aplikácii sám zvolí, akým zoznamom bude anonymizovať ten-ktorý stĺpec tabuľky. Vytvorenie tejto základnej pracovnej množiny je už len implementačný krok, ktorý používateľ bude už len využívať. Tento druh anonymizácie tak veľmi účinne maximalizuje aj informačnú hodnotu, aj súkromnosť dát.

3.1.4 Shuffling

Jednou z ďalších metód, prichádzajúcich do úvahy, je vyššie spomenuté prehádzanie dát, shuffling. Použitie takéhoto prehádzania dát si však vyžaduje hlbšie zamyslenie, či je v konkrétnom prípade vhodné na anonymizáciu dát. Ako príklad zlého použitia prehádzania dát môžeme uviesť anonymizáciu údaju 'Suma' v tabuľke 'Platy'. Útočník z takýchto dát môže veľmi ľahko a rýchlo zistiť napríklad plat riaditeľa spoločnosti, keď si nájde maximálne číslo v danom stĺpci 'Suma'. Pri usporiadaní celého stĺpca si z najvyšších hodnôt vie vyčítať aj platy top manažmentu, pri nájdení modusu plat úradníka juniora, ktorých býva vo firme najviac, prípadne priemerný plat a iné.

Na druhej strane, ak máme veľkú databázu osôb, anonymizácia krstných mien osôb prehádzaním jednotlivých záznamov je bezpečná a má dokonca vyššiu informačnú, štatistickú hodnotu, ako nahradenie týchto údajov virtu-

álnymi dátami, keďže sa zachováva reálna distribúcia mien daného jazyka. Pri mnohých údajoch, akými je napríklad už aj priezvisko, toto rozhodnutie o správnosti anonymizácie prehadzovaním vôbec nie je jednoduché a kvantitatívne hľadisko nie je vždy tým správnym rozhodovacím faktorom.

Proces anonymizácie môže byť ľubovoľný iný algoritmus. Či už je to vynechanie samohlások alebo spoluhlások v daných reťazcoch, aplikácia posuvnej šifry na jednotlivé znaky, vystrihovanie podreťazcov, prehadzovanie znakov, každý jeden algoritmus má svoje výhody a nebezpečné zákutia, nad ktorými sa treba zvlášť zamyslieť a zvážiť ich použitie.

3.1.5 Anonymizovanie čísel a dátumov

Pri týchto typoch údajov nie je potrebné využívať tak silný nástroj, ako sú virtuálne dáta. Nahradenie náhodnými hodnotami, ktoré si pri reťazcoch nemôžeme dovoliť, nám v tomto prípade ideálne vyhovuje. Dátumy a čísla sa tak aj po nahradení ľubovoľnou hodnotou stále javia ako reálne dáta. Tieto údaje môžeme anonymizovať ako fixnou hodnotou, tak aj náhodným posunutím v rámci určitého intervalu alebo náhodnými hodnotami z daného intervalu. Reálna informačná hodnota, ktorú daným dátam ponecháme, bude závisieť od našej voľby. Pri náhodnom posunutí v rámci percentuálne zadaného intervalu bude výpovedná hodnota vyššia, ak si zvolíme odchýlku 10 %, ako pri odchýlke 50 %. Podobne pri dátume, ak si zvolíme náhodný posun v rozmedzí 60 dní, výpovedná informačná hodnota pre napríklad štatistické účely bude vyššia, ako keby sme robili náhodný posun v rozmedzí jedného roka, 365 dní.

3.1.6 Anonymizovanie textov a poznámok

Pri anonymizovaní takýchto typov údajov, kde pracujeme s dlhšími textami, sme v literatúre nenašli žiaden odporúčaný spôsob, ktorý by dokonale spĺňal naše požiadavky. Ľubovoľné transformácie reálneho textu za účelom zachovania aspoň nejakej informačnej hodnoty vyúsťujú len do bezvýznamových zhlukov znakov, ktoré nepôsobia reálne. Jedným z riešení, ktoré z informačnej hodnoty reálnych dát zachová aspoň jeden parameter, dĺžku reťazca, je nahradenie reálnych dát zmysluplným preddefinovaným textom približne zhodnej dĺžky. Daný text by bol na rozdiel od 'Lorem ipsum' z plnej znakovej sady Slovenského jazyka, aby jednak diakritikou pomohol kontrolovať správne zobrazovanie všetkých znakov a jednak svojou pôvodnou dĺžkou testoval prípadné pretečenie, ktoré by pri testoch na reálnych dátach mohlo nastať. Takéto pretečenie by sme napríklad pri vynulovaní daných údajov na testovacej databáze nedokázali odhaliť.

3.2 Postupy pri anonymizácii dát

V tejto časti si uvedieme niektoré zaujímave postrehy a postupy, ktoré by sme mali dodržat jednak pri ľubovoľnom procese anonymizácií dát, ale hlavne aj pri implementácii aplikácie, ktorá si kladie za cieľ kvalitne anonymizovať dáta.

3.2.1 Pretečenie

Pri používaní virtuálnych dát, ktorými nahrádzame reálne dáta, by sme mali dávať pozor na typ stĺpca, do ktorého dáta vkladáme, aby nedošlo k pretečeniu. Ak teda máme stĺpec typu *Varchar(20)*, tak aby sme sa nepokúšali do neho vložiť reťazec dĺžky väčšej ako je 20 znakov. Takýto pokus by

nám spôsobil chybu.

3.2.2 Inteligentné kľúče

Niektoré údaje, ako napríklad rodné čísla alebo čísla účtov, majú istú definovanú štruktúru. Napríklad pri rodných číslach je to deliteľnosť číslom 11. Každé korektné rodné číslo spĺňa túto podmienku. Niektoré aplikácie majú implementovanú práve takúto ochranu, ktorá upozorní na náhodný preklep pri rodnom čísle, ktoré nespĺňa túto podmienku. Pri anonymizácií podobných dát preto nesmieme zabúdať aj na takéto inteligentné klúče, ktorých náhodná anonymizácia by bola problémom pri testovaní.

3.2.3 Relevantné dáta

Ako sme už viac krát spomínali, jednou z hlavných požiadaviek na anonymizované dáta je, aby vyzerali čo najdôveryhodnejšie. Kvalitné anonymizované dáta by mali byť na nerozoznanie od reálnych dát. Musíme sa preto vyvarovať prípadov, ak napríklad údaj 'Mesto' nahradíme náhodným zhlukom znakov, ktorý už na prvý pohľad pri zobrazení upúta pozornosť. Rovnako tak by sme mali číselné hodnoty nahradiť opäť len číselnými hodnotami a podobne.

3.2.4 Prípad izolovaných hodnôt

Anonymita je vo všeobecnosti založená na existencii veľkého počtu rovnakých záznamov. Spomínaný prípad s tabuľkou 'Platy' môže obsahovať údaje, ktoré sa značne odlišujú od priemeru a tým sa stávajú unikátnymi, nesú informáciu navyše, ktorá môže pomôcť pri identifikácii, priradení k určitej osobe.

3.2.5 Konzistentná anonymizácia

Pre testovacie účely je častokrát výhodou, ak vieme aj náhodne procesy simulovať viackrát po sebe s rovnakým výsledkom. Takúto požiadavku máme aj pri anonymizácií – pri viacnásobných behoch anonymizácie nad jednou databázou je vhodné, ak sú výstupy identické. Tento jav sa dá zaistiť rovnakým počiatočným nastavením generátora náhodných čísel.

3.3 Dôvody anonymizácie

Keď už máme približnú predstavu, ako prebieha proces anonymizácie, skúsime si teraz objastniť, aké sú dôvody na používanie tejto metódy.

Dôvody používania anonymizácie by sme v skratke vedeli stotožniť s dôvodmi ochrany osobných údajov. Údaje v databázach anonymizuje práve preto, aby sme ochránili osobné údaje. A dôvody, prečo chrániť osobné údaje a prístup k nim, sú zrejmé.

Súčasným trendom vo vývoji aplikácií je v dnešnej dobe vývoj informačných systémov na mieru najmä externými spoločnosťami, ktoré sa špecializujú na vývoj softvéru. Inštitúcie a spoločnosti si tak už nevytvárajú informačné systémy na svojej pôde, ale využívajú častokrát takéto objednávky. Je to elegantné riešenie, pri ktorom sa programátori zadávateľa môžu naďalej venovať prevádzkovým úlohám, zadávateľská firma nemusí na vývoj prechodne najímať a zaškoľovať nových programátorov. Pri dobrom výbere dodávateľa sa môže spoľahnúť na rýchle a profesionálne riešenie, zaručené skúsenosťami z vývoja, ktoré by mohli absentovať pri svojpomocnom riešení.

Práve z dôvodov, že nový informačný systém je šitý na mieru pre danú spoločnosť a pri nasadení do prevádzky je potrebné už len preklopiť produkčné dáta do novej databázy, externé spoločnosti už pri vývoji a hlav-

ne testovaní potrebujú reálne dáta zadávateľa. Na nich tak vedia odhaliť množstvo chýb, ktoré by sa inak určite prejavili pri ostrej prevádzke. Tu sa dostávame k dôvodom anonymizácie, keď zadávateľ nemá dôvod poskytnúť externej spoločnosti reálne dáta so všetkými citlivými údajmi, ktorými disponuje. Anonymizované dáta sú častokrát na nerozoznanie od tých reálnych a externá spoločnosť v teoretickom prípade ani nemusí vedieť, že v celej poskytnutej databáze nie je jediný reálny údaj. Je pravdou, že najhorší scenár, ktorým je únik a zneužitie citlivých dát, je zmluvne podchytený hneď na začiatku spolupráce, otázkou však zostáva, aký postih je dostatočný. Či napríklad pri citlivej zdravotnej dokumentácii desiatok tisícok ľudí, z ktorých dáta by poisťovne vedeli veľmi šikovne využívať, je hoci aj obrovská pokuta a trestnoprávna zodpovednosť konkrétnych previnilcov adekvátny postih, ktorý sa vyrovná uvedenej škode ? V čisto teoretickej rovine by sme pri hodnote týchto dát mohli za takýmito únikmi vidieť aj rýdzo obchodný motív.

Priznávame, že k úniku citlivých dát môže prísť na ktoromkoľvek mieste, nielen na strane externej firmy, faktom však zostáva, že čím menší je počet kópií databáz a prístupov k nim, tým je menšia pravdepodobnosť, že k úniku dôjde. Zvlášť to platí pri šikovnosti dnešných hackerov a nedbalosti niektorých zamestnancov. Toto je jedno z miest, kde sa citlivé dáta dajú chrániť a anonymizácia je jedným z riešení.

Ďalším dôvodom, ktorý stojí za spomenutie, je používanie anonymizácie v rámci samotných spoločností. Súvisí to opäť s pozorovaním, že v niektorých prípadoch nie je pri práci s údajmi nutné využívať reálne dáta. Ako príklad by sme mohli uviesť programátora v poisťovni alebo banke, ktorý robí štatistiku o počtoch volaní na Call Centre svojej spoločnosti. Tento programátor nemusí mať prístup k reálnym menám, číslam a dátam k tomu, aby vedel plnohodnotne vykonávať svoju prácu. Pri svojej práci tak nemá nutkanie

zisťovať finančný stav klientov svojho zamestnávateľa a vo veľkých spoločnostiach s desiatkami zamestnancov opäť odpadá ďalšia významná hrozba úniku dát, ktorej sa dá týmto spôsobom ľahko zabrániť.

Anonymizácia nie je univerzálnou metódou, ako chrániť dáta pred ich únikom a zneužitím a toto boli len dva prípady, kde je táto ochrana použiteľná. Stále však zostáva množstvo prípadov a miest, kde anonymizácia nedokáže chrániť dáta. Preto aspoň na miestach, kde to anonymizácia zvláda, by sme mali rozhodne zhodnotiť jej používanie.

3.4 Informačná hodnota vs. ochrana údajov

Pri anonymizácií dát v prvom rade hovoríme o pojmoch informačná hodnota a ochrana údajov. Prirodzene sa snažíme anonymizovať dáta takým spôsobom, aby mali čo najväčšiu výpovednú hodnotu z hľadiska informácií, na druhej strane sa zas snažíme, aby obsahovali minimum reálnych osobných údajov. Tieto dve veličiny, informačnú hodnotu a úroveň ochrany údajov, sa snažíme maximalizovať. Jednoduchým „vyNULLovaním“ všetkých dát dosiahneme najvyššiu úroveň ochrany, takéto dáta nám však dajú len minimálnu výpovednú hodnotu, ktorou je schéma údajov. Na druhej strane pôvodné dáta, ktoré sme vôbec neanonymizovali, majú maximálnu informačnú hodnotu, keď obsahujú kompletné nezmenené citlivé dáta. Riešením, na ktorú stranu sa prikloniť, je stredná cesta, ale je to častokrát náročné. Ide o hľadanie rovnováhy, balansu, keď sa snažíme maximalizovať obe veličiny.

3.5 Príklad

Celý proces anonymizácie sa pokúsime ukázať na jednoduchom príklade. Vezmime si súkromnú strednú školu, ktorá svoj existujúci informačný systém potrebuje nahradiť novším. Na jeho vývoj si najala externú spoločnosť, kto-

rej, z dôvodov ochrany súkromia, chce poskytnúť iba testovacie dáta. Svoju existujúcu databázu s reálnymi dátami teda potrebuje anonymizovať.

Databáza tejto školy pozostáva z viacerých tabuliek, spomenieme niektoré. Tabuľka 'Osoby' obsahuje údaje o menách, priezviskách, dátumoch narodenia a pohlaví študentov. Tabuľka 'Adresy' obsahuje adresy trvalého pobytu a kontaktné adresy všetkých osôb. Tabuľka 'Predmety' obsahuje zoznam všetkých predmetov, ktoré sa vyučujú, tabuľka 'Hodnotenia' obsahuje hodnotenia žiakov ku týmto predmetom.

Tabuľkou, ktorá si najviac vyžaduje našu pozornosť, je tabuľka 'Osoby', keďže sa v nej nachádzajú tie najcitlivejšie údaje. Na anonymizáciu údajov, ako je meno a priezvisko, je v tomto prípade vhodné použiť virtuálne dáta. Podľa pohlavia vieme nahradiť krstné mená a priezviská príslušnými fiktívnymi ženskými a mužskými menami, priezviskami. V prípade potreby je možné anonymizovať aj stĺpec pohlavie, kde by sa náhodne vygenerovalo úplne nové, náhodné rozdelenie pohlaví. Takúto operáciu je však každopádne potrebné vykonať ešte pred anonymizáciou mena a priezviska, z dôvodu zachovania konzistencie údajov. Pri dátume narodenia by sme mohli použiť techniku náhodného posunu v rozmedzí napríklad 150 dní, ktoré je dostatočné na jeho anonymizáciu.

Tabuľku 'Adresy' by sme podobne anonymizovali pomocou virtuálnych dát, keďže prehádzanie údajov by síce určite sťažilo potencionálnemu útočníkovi možnosť obnoviť pôvodné dáta a zneužiť ich, adresa je však jeden zo štandardných údajov, na ktoré sú v dobrej aplikácii anonymizéra predpripravené virtuálne dáta a táto metóda je omnoho účinnejšia a bezpečnejšia. Fiktívnymi dátami by sme teda osobitne nahradili názov ulice, mesta, čísla, smerovacie čísla.

Pri tabuľke 'Predmety' a rýchlom zamyslení sa môžme dospieť k záveru

zbytočnosti anonymizácie týchto dát. V tomto prípade - školy - je tento konkrétny údaj jednak ľahko predvídateľný, zistiteľný a jednak nesie minimálnu informačnú hodnotu, ktorá by sa dala zneužiť. Je teda na mieste zamyslieť sa pri niektorých stĺpcoch, či anonymizácia daného údaju je vôbec potrebná.

Pri tabuľke 'Hodnotenia' prichádza do úvahy viacero možnosti. Jednak nie často používaná metóda prehadzovania, ktorá v tomto prípade môže byť bez problémov použitá a len rozhádže reálne známky medzi všetkých žiakov. Druhou možnosťou môže byť nahradenie náhodnými dátami, v tomto prípade náhodnými celými číslami v intervale <1,5>, resp. inom intervale pri snahe zachovania dobrého mena školy. Treťou možnosťou je vôbec otázka použitia anonymizácie na takýto údaj.

Tento proces anonymizácie, ako sme mali možnosť vidieť, nie je teda žiaden náročný algoritmus. Pri konkrétnych dátach sa stačí zamyslieť a zvážiť, aké dôverné sú tieto dáta a možno očami útočníka, ktorý by sa dostal práve k tejto množine údajov zvážiť, ako by mohli byť zneužité. Na základe tohto pozorovania si tak vybrať techniku anonymizácie - buď použitie najbezpečnejšej techniky virtuálnych dát, v prípade absencie príslušného typu predgenerovaných zoznamov jednoduché nahradenie konštantou alebo vynullovanie. Na druhej strane použitie jednoduchého prehádzania dát, keď dáta nie sú príliš citlivého charakteru.

3.6 Existujúce riešenia

Anonymizácia nie je úplnou novinkou vo svete a preto samozrejme existuje niekoľko aplikácií, ktoré sa viac či menej pokúšajú o komplexné riešenia anonymizácie dát. Niektoré z týchto dostupných aplikácií sú len drobné projekty, na ktorých sa autori snažili demonštrovať nový nápad, zaujímavé riešenie. Iné sú prepracované aplikácie s množstvom funkcií, za ktorých vývojom sto-

ja väčšie spoločnosti, ktoré na vývoj investovali určité prostriedky. Takéto aplikácie sú tak prístupné iba s časovým, resp. funkčne obmedzeným použitím, samozrejme s možnosťou zakúpenia platenej licencie pre neobmedzené použitie. Ďalšou z veľkých nevýhod takýchto riešení je absencia generovania slovenských virtuálnych dát, ktoré tieto aplikácie neponúkajú, resp. nemožnosť rozširovať existujúce množiny virtuálnych dát užívateľom. Ako sme už spomínali, pri procese anonymizácie je jedným z najväčších kladov práve čo najreálnejší vzhľad virtuálnych dát, ak sú tieto dáta vizuálne na nerozoznanie od reálnych dát.

Analýzy takýchto existujúcich aplikácií v tejto práci poslúžili ako vhodné námety, inšpirácie na našu vlastnú aplikáciu, ktorá čerpá námety z kladných pozorovaní, zistení a snaží sa ich ešte vylepšiť, inšpiruje sa zaujímavými riešeniami. Zároveň sa vďaka týmto pozorovaniam môžeme vyvarovať chýb a nedostatkov, ktoré sme objavili, poučiť sa z nich a priniesť nové, lepšie riešenia.

Jednou zo skúmaných aplikácií, ktoré sme mali možnosť testovať, bola 30 dňová verzia aplikácie Data Masker od firmy Net 2000 Ltd [14].

3.6.1 Data Masker

Data Masker je aplikácia, ktorá slúži na odstránenie citlivých dát z testovacích databáz a nahrádza ich reálne vyzerajúcimi nepravdivými informáciami. Dôvodmi na takúto anonymizáciu sú právne a praktické podnikové dôvody, keď únik citlivých údajov (zákazníka) z testovacích databáz môže spôsobiť devastujúcu ujmu verejnej dôveryhodnosti, prípadne môže zahŕňať právnu zodpovednosť. Aplikácia eliminuje riziko nenáležitej viditeľnosti dát – testovacie dáta vyzerajú reálne a vývojový tím nikdy nespozná rozdiel, pričom citlivé informácie zostanú úplne chránené.

Zámer tejto aplikácie je teda zhodný s nami popísanou očakávanou funkcionalitou aplikácie na anonymizáciu dát. Data Masker je prepracovanou, komplexnou a najdokonalejšou implementáciou, ktorú sme mali možnosť skúmať a testovať. Zahŕňa rozmanitú funkcionalitu, ktorá má pomôcť pri definovaní pravidiel anonymizácie. Techniky anonymizácie, ktoré ponúka táto aplikácia sú rozmanité.

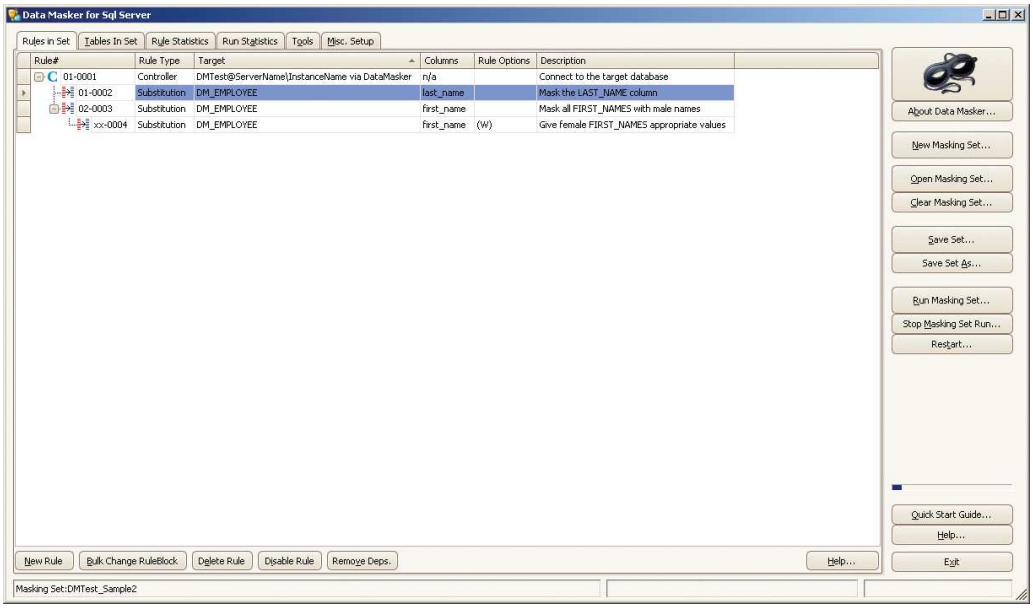
V prvom rade je to technika substitúcie s využívaním virtuálnych dát, ako jedna z najúčinnejších metód, ktorá pozostáva z náhodného nahradzovania obsahov stĺpcov z virtuálnych pripravených zoznamov. Aplikácia obsahuje pripravené zoznamy najčastejšie používaných dát, ako zaujímavé zoznamy čísel kreditných kariet, TCP/IP adries, poznávacích značiek vozidiel alebo bankových účtov.

Ďalej je to jednoduchá technika prehadzovania dát, shuffling, ktorá si nevyžaduje žiadne ďalšie parametre a nastavenia. Jej použitie ale treba podľa autorov dobre zvážiť, hlavne pri tabuľkách s malým počtom riadkov.

Pre číselné hodnoty a dátumy tento nástroj ponúka široký výber techník:

- nullovanie hodnoty
- nahradenie konštantnou hodnotou
- nahradenie náhodnou hodnotou zo zadaného intervalu
- náhodný posun hodnoty v rámci percentuálne zadaného rozsahu
- možnosť zakázania nulových hodnôt

V príručke sa autori zamýšľajú aj nad technikou šifrovania. Táto technika sa na prvý pohľad javí ako zaujímavé riešenie, keď ponecháme reálne dáta v tabuľkách, avšak zašifrované kľúčom. Podľa autorov sa ale jedná



Obrázok 1: Hlavné okno aplikácie Data Masker

o jednu z najmenej používaných techník. Výhoda prístupu k reálnym dátam pre ľubovoľnú osobu vlastniacu kľúč je v tomto prípade anonymizácie naopak obrovskou nevýhodou pri testovacích databázach. Pri úniku kľúča sa tak útočník vie dostať k celej databáze reálnych údajov. Šifrovanie zároveň nezachováva formátovanie a z hodnôt údajov sa stávajú nezmyselné alfanumerické reťazce. Otázkou zostáva tak isto aj použitie konkrétnych šifrovacích algoritmov, ktoré sú rozdielne silné a pri niektorých je iba otázkou času a úsilia, kým sa ho podarí prelomiť. Nikdy netreba používať jednoduché algoritmy, napríklad posun písmen o určitú hodnotu, ktoré sú triviálne rozbité použitím jednoduchých techník frekvenčnej pravdepodobnosti.

Samotná aplikácia pozostáva zo základného okna, kde si v záložkách môžeme editovať :

- Rules in Set – zoznam všetkých zatiaľ nastavených pravidiel pre ano-

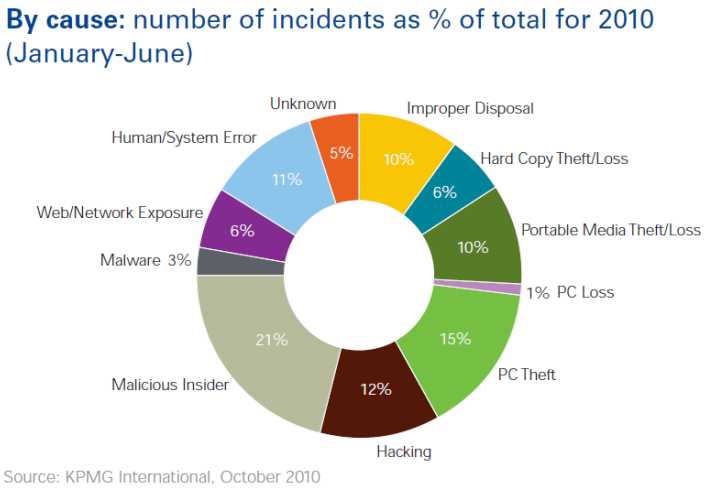
nymizáciu konkrétnych stĺpcov - typ, popis a parametre danej techniky

- Tables in Set – množinu tabuliek so stĺpcami, pre ktoré samostatne definujeme konkrétne pravidlá
- Rule Statistics – štatistiku zatiaľ použitých pravidiel na anonymizáciu, ktorá je len prehľadová
- Run Statistics – štatistiku jednotlivých behov programu pri samotnom procese anonymizácie - dĺžky behov konkrétnych fáz prepisu dát, počty upravených riadkov a podobne
- Tools a Misc. Setup – panel nástrojov, ktorý slúži na nastavovanie širokého spektra doplnkových funkcií

3.7 Prípady únikov

Pri súčasnej sile informatizácie, pokrytí Internetu, množstve počítačov a databáz, prakticky o každom a o všetkom, pri súčasnej šikovnosti IT znalých a nedbalosti menej IT znalých ľudí, dochádza k množstvu únikov dát. Ťažko možno hovoriť o nejakých číslach, ale ľubovoľné straty a úniky, možno nie celých databáz, ale len menšieho množstva údajov, sa dejú denne. Či sa už jedná o oficiálne vyhlásenia poškodených strán, s priznaním, že došlo k úniku dát, či už sú to množstvá únikov, ktoré sa pre ochranu dobrého mena spoločnosti nikdy verejnosť nedozvie, alebo sú to individuálne trofeje hackerov, ktorí sa o takýchto úspechoch, spolu s nejakým dôkazom, veľmi radi pochvália na svojich stránkach. O únikoch typu „Predstav si, že tento a tento má u nás v banke na účte toľko a toľko peňazí“, a podobných, nehovoriac.

Ako sme už spomínali, ale aj na tomto mieste chceme opäť zdôrazniť, anonymizácia vo väčšine prípadov nedokáže ani zďaleka zabrániť úniku cit-

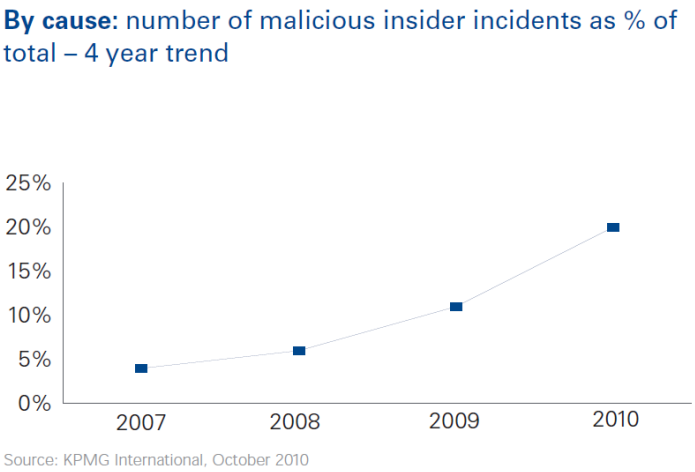


Obrázok 2: Percentuálne rozdelenie príčin únikov dát v prvom polroku 2010

livých dát, ktoré sa stávajú. Väčšina prístupov k databázovým údajom si práve vyžaduje prístup k reálnym dátam a o anonymizácii nemôže byť ani reč. Existujú však prípady, o ktorých sme písali - vývoj aplikácií, testovanie, značná časť práce programátora, kde si to dovoliť môžeme. A aspoň na tých miestach, kde touto metódou vieme veľmi elegantne zabrániť únikom , by sme používanie anonymizácie mali zvážiť.

Čo sa všeobecne týka spomínaných únikov dát, výskum spracovaný firmou KPMG z roku 2010 konštatuje nasledovné fakty.

Ako ukazuje graf na obrázku číslo 2, v prvom polroku 2010 bola viac ako jedna pätina zistených incidentov spáchaná záškodníctvom zvnútra organizácie. Boli to teda zamestnanci vlastných firiem, ktorí sa nejakým spôsobom pokúsili nelegálne vyniesť firemné dáta. Ako ukazuje graf na obrázku číslo 3, tento stav má za posledné 4 roky značne stúpajúci charakter. Kým v roku 2007 bolo podľa grafu takýchto prípadov okolo 5 %, v roku 2010 to už bolo 21 %. Problémovými v týchto prípadoch sa častokrát stávajú hlavne



Obrázok 3: Počet únikov dát spôsobený záškodníctvom zvnútra firmy v priebehu 4 rokov

zamestnanci, ktorí nedobrovoľne opúšťajú svoje pracovné pozície.

Takýto trend je na vzostupe aj podľa [9] bezpečnostného špecialistu technologickej spoločnosti S&T CZ, ktorý uvádza : „Krádeží cenných dat se snaží zaměstnanci zvýšit svoji hodnotu na trhu práce. I jinak loajální zaměstnanci mají tendenci si z firmy odnést informace, které by mohly být pro potenciálního zaměstnavatele zajímavé.“

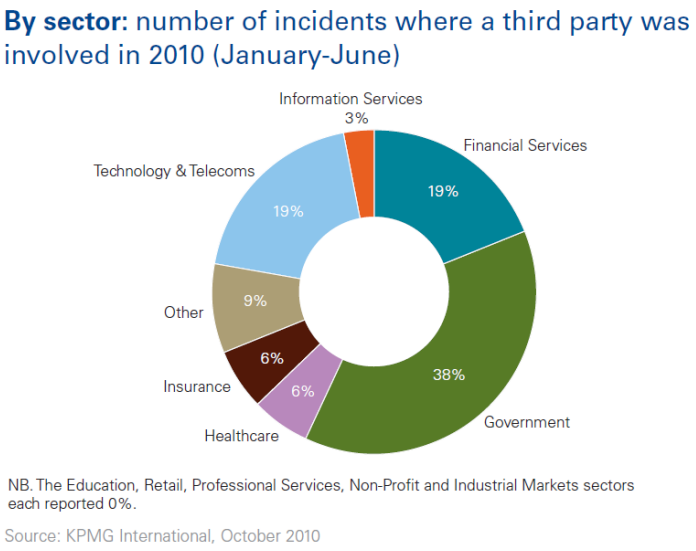
Podobne podľa [9] faktom ostáva, že stále väčšie množstvo zamestnancov pristupuje k firemným dátam s cieľom ukradnúť ich. Potvrzuje to i prieskum spoločnosti McAfee. Podľa prieskumu 42 percent respondentov pokladá odchádzajúceho zamestnanca za najväčšie ohrozenie citlivých firemných dát. Riziko naviac predstavujú aj zamestnanci, ktorým reálne prepustenie nehrozí. „Ti si vytvářejí často pouze 'pojistku' pro případ, že by si museli hledat nové místo. Kromě zaměstnanců mají mnohdy k citlivým firemním datům přístup i třetí strany, které mohou mít stejnou motivaci či eminentní zájem o krádež dat„ uvádza[9] riaditeľ českej a slovenskej pobočky spoločnosti

McAfee.

Za takýmito únikmi väčšieho množstva dát, databáz, musia ale zákonite stáť ľudia z pozícií, kde majú prístup k celým databázam. Nie sú to teda koncoví pracovníci alebo operátori, ktorí len využívajú informačný systém na podávanie informácií klientom. Na tejto pozícií by museli tieto dáta len dlhodobo spisovať a pomaly si svojpomocne vytvárať databázu. Domnievame sa teda, že za takýmito únikmi databáz zvnútra firmy, musia byť pravdepodobne ľudia z IT. Ak je to tak, tak práve anonymizácia a už vyššie spomenuté dôvody, že bežný programátor, ktorý vytvára reporty, nepotrebuje svoje skripty testovať na ostrých dátach.

Správa rovnakej firmy z roku 2009 poukazuje aj na problematiku tretích strán. Už len samotný nadpis správy o tretích stranách – 'Trusted third parties – an oxymoron?', voľne preložený ako 'Dôveryhodné tretie strany – oxymoron ?', hovorí dosť o charaktere uvedených zistení. Správa popisuje niektoré významné a rozsiahle úniky citlivých dát a dodáva, že relatívne nízke percento únikov dát zavinených tretími stranami – 11 % v roku 2009, je len špičkou ľadovca, keď väčšina prípadov ostáva nezverejnených. Správa ďalej uvádza niekoľko spôsobov, ako sa chrániť a vyvarovať problémom pri takejto spolupráci. Pre ilustráciu, graf na obrázku číslo 3 popisuje rozdelenie incidentov, do ktorých boli zahrnuté tretie strany v prvej polovici roku 2010 podľa sektorov.

Čo sa týka známych únikov citlivých dát na Slovensku, hádam najznámejšími sú úniky dát telefónnych operátorov EuroTel a Orange. Tieto databázy pochádzali z roku 2002, počet záznamov bol okolo 3 miliónov a medzi uniknutými dátami sa vyskytovali aj utajené čísla členov tajnej služby a policajtov. Únik databáz zo svojich systémov vylúčil mobilný operátor Orange. Podľa [11] oficiálne stanovisko spoločnosti uvádza, že štruktúra dát bola totožná



Obrázok 4: Počty incidentov, kde bola zahrnutá tretia strana v prvom polroku 2010

s tou, ktorú v uvedenom čase poskytla spoločnosť Orange tretím stranám. Operátori totiž v tom čase museli svoje databázy poskytovať Slovenskej informačnej službe a Ministerstvu vnútra SR.

Známe sú ešte dva prípady [10] významnejších únikov citlivých dát na Slovensku, ktoré spomenieme.

Prvý je ešte z roku 1999, keď pri modernizácii systému Slovenskej poisťovne zmizla časť databázy s poistenými motorovými vozidlami, ktorá sa následne objavila na internetovej stránke hackerskej skupiny 'F'. Poisťovňa podala trestné oznámenie na zamestnanca dodávateľskej firmy, ktorého podozrievala, že dáta ukradol. Prípad sa dostal až pred súd, ktorý však obvineného pre nedostatok dôkazov oslobodil.

Druhý medializovaný hackerský kúsok prebehol začiatkom júna 2003. Na stránke Hysteria.sk sa objavil článok, ktorý podrobne opisoval prienik do podnikovej siete Slovenských telekomunikácií. Ako v reakcii uviedol vicep-

rezident spoločnosti pre komunikáciu, hackeri zverejnili dva roky starý telefónny zoznam a obsah schránok elektronickej pošty administrátorov siete. Okrem toho však útočníci v on-line článku zverejnili aj čiastočné informácie o nastavení dôležitých sieťových zariadení, ako aj prihlasovacie mená zákazníkov ST, ktorí mohli byť v konkrétnom čase pripojení na internet.

3.8 Miera anonymizácie

V nasledujúcej časti sa zamyslíme, aké údaje sú dostatočné na jednoznačnú identifikáciu osoby a pokúsime sa tento neurčitý pojem identifikovateľnosti popísať konkrétnym algoritmickým kritériom. Vďaka takémuto 'matematickému' popisu by sme tak o jednotlivých množinách údajov vedeli rozhodnúť, či postačujú na identifikáciu osoby alebo nie. Takéto rozhodnutia v súčasnosti poskytujú hlavne výklady príslušného zákona o osobných údajoch, ktoré vznikajú priebežne na základe podnetov, nejasností.

Vo všeobecnosti teda chceme zistiť, či daná sada údajov je schopná identifikovať osobu vo 'väčšine' prípadov. Údaj o náboženstve (džinizme), ktorý sme uvádzali, je síce občas schopný identifikovať danú osobu, vo väčšine prípadov k tomu však nedochádza. Takýmito známymi množinami údajov je zatiaľ rodné číslo a trojica údajov meno, priezvisko, adresa. Možným riešením, ako tento jav merať, je spočítanie strednej hodnoty osôb, ktoré sú danou množinou údajov určené. Metódami na výpočet strednej hodnoty, nad ktorými budeme uvažovať sú aritmetický priemer, modus a medián. Tieto tri charakteristiky spočítame pre jednotlivé množiny údajov.

Rodné číslo. Ak vezmeme do úvahy spomínaných zhruba 30000 prípadov duplicitných rodných čísel a 5 429 763 obyvateľov Slovenskej republiky (stav k 30. júnu 2010), dostávame priemer počtu osôb, identifikovaných týmto údajom 1,0055. Modus, teda najčastejšie sa vyskytujúca hodnota a medián,

ktorý delí sledované spektrum na polovicu, sú v tomto prípade 1.

Trojica údajov meno, priezvisko, adresa. V tomto prípade nemáme také presné čísla ako pri rodnom čísle. Nápomocný by bol počet osôb s rovnakým menom, ktoré majú totožnú adresu. Z neskôr uvedených dôvodov si priemer v tomto prípade dovolíme iba odhadnúť na 1,05, hodnota modusu a mediánu bude v tomto prípade opäť 1. Možnými kritériami nášho problému identifikácie by takto mohli byť :

(a) Množina údajov má byt považovaná za identifikujúcu, ak aritmetický priemer počtu osôb, ktoré sa v tejto množine údajov zhodujú, je $1 + \epsilon$,

(b) Množina údajov má byť považovaná za identifikujúcu, ak modus počtu osôb, ktoré sa v tejto množine údajov zhodujú, je 1,

(c) Množina údajov má byť považovaná za identifikujúcu, ak medián počtu osôb, ktoré sa v tejto množine údajov zhodujú, je 1.

Prvá možnosť je nejednoznačná, kvôli parametru ϵ, kde by bolo potrebné bližšie dodefinovať, aké epsilon je dosť malé na to, aby bolo ešte vyhovujúce. Tretia možnosť sa javí ako dobrá charakteristika avšak podľa nášho názoru, druhá možnosť najlepšie vyjadruje intuíciu, že vo "väčšine prípadov„ dokáže izolovať konkrétnu osobu.

Nechávame však aj na zamyslenie čitateľa, ktorá z uvedených troch charakteristík je z jeho pohľadu pre tento problém vyhovujúca.

Ďalej budeme experimentovať s vybranou charakteristikou a zrátame ju pre niektoré množiny údajov. Zaujímavou množinou z pohľadu identifikovateľnosti sa nám javí množina údajov meno, priezvisko, mesto. Ako testovacie dáta sme pri tomto pokuse používali verejne prístupnú databázu na zisťovanie telefónnych čísel na stránke 'http://telefonny.zoznam.sk/'.

Na náhodnej vzorke priezvisk a miest sme si nechali vyhľadávať celé mená a presné adresy a na základe týchto údajov sme zrátali, že modus týchto hodnôt bol aj v tomto prípade 1. Na rovnakej adrese samozrejme môžu bývať otec a syn, resp. slobodná dcéra a matka s rovnakými krstnými menami. Takýchto prípadov pri našom testovaní bolo viacero, avšak nevyvrátilo to našu domnienku, že aj meno, priezvisko a mesto identifikuje osobu vo väčšine prípadov. Podľa našej miery teda táto množina údajov je schopná identifikovať osobu.

Takto definovaná miera a výsledky sú v súlade s definíciou osobných údajov [12], podľa ktorej : "Definícia osobných údajov nevyžaduje, aby išlo o konkrétnu identitu fyzickej osoby, ale postačuje, aby za splnenia daných podmienok bola osoba určiteľná. Z vecnej stránky sa pojem osobných údajov vzťahuje na také údaje, ktoré sa týkajú identifikovanej alebo identifikovateľnej fyzickej osoby, t.j. určenej alebo určiteľnej, či už priamo alebo nepriamo. Z pohľadu ochrany osobných údajov sa „určiteľnosťou" rozumie taký stav, keď na základe jedného alebo viacerých údajov možno osobu identifikovať. Identifikácia sa teda realizuje prostredníctvom konkrétnych charakteristických znakov, tzv. „identifikátorov", ktoré možno priradiť ku konkrétnej fyzickej osobe. Do akej miery sú určité identifikátory dostačujúce pre dosiahnutie identifikácie konkrétnej fyzickej osoby závisí od komplexného posúdenia dostupných údajov v ich vzájomnej súvislosti a zároveň aj situácie ako celku.

Ďalšími sadami údajov, ktoré by bolo zaujímavé preskúmať, by boli množiny údajov, ktoré by obsahovali údaj o dátume narodenia, resp. veku osôb. Pri takýchto množinách údajov, ako napríklad meno, priezvisko a vek, by však už na testovanie bola potrebná databáza, ktorá by obsahovala takéto údaje. Takéto databázy však nie sú bežne prístupné a tak ostáva zatiaľ na zamyslenie, či takáto trojica údajov by dokázala jednoznačne identifikovať

osobu. Výsledky pre takúto sadu údajov by boli zaujímavé aj z pohľadu počtu obyvateľov konkrétnej krajiny, ktorý je v tomto prípade rozhodujúci, keďže počty mien a priezvisk rôznych krajín môžeme rádovo považovať za približne rovnaké. V niektorej krajine by teda podľa nášho kritéria modusu mohla byť trojica údajov meno, priezvisko, vek považovaná za osobný údaj, v ľudnatejšej krajine však už asi nie.

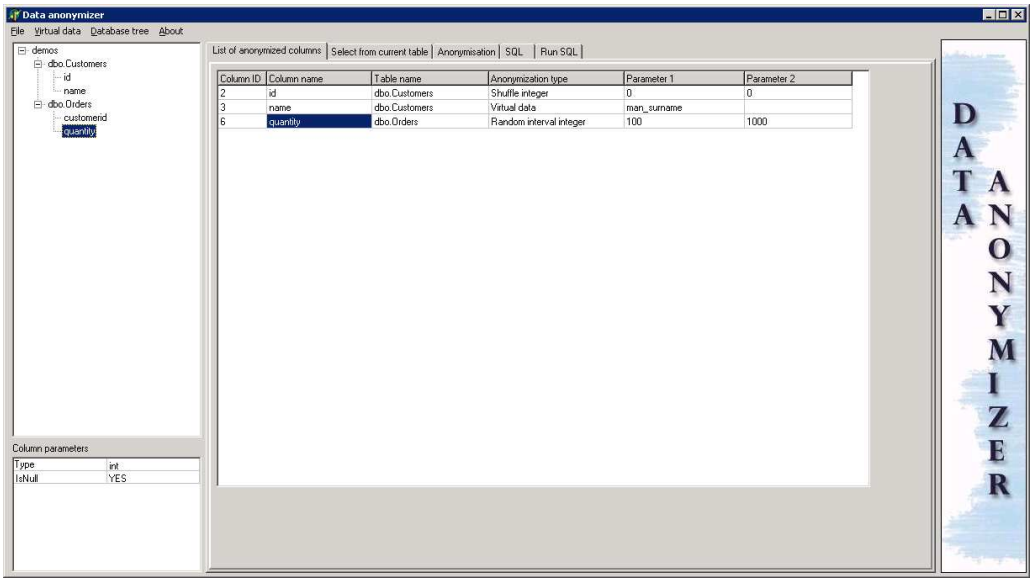
4 Implementácia

V ďalšej časti práce sa budeme venovať implementácii aplikácie na anonymizáciu dát v databázových tabuľkách, ktorú sme vytvorili. Túto aplikáciu sme pomenovali `Data anonymizer`. Vyvíjali sme ju vo vývojovom prostredí `Delphi` na platforme `SQL Server`. Táto aplikácia nie je taká robustná a rozsiahla ako spomínaná aplikácia `Data Masker`, veľmi jednoduchým a účinným spôsobom však plní účel anonymizácie dát a prináša aj niekoľko inovácií a nových nápadov.

5 Triedy

Samotná implementácia projektu `DataAnonymizer` pozostáva z niekoľkých samostatných tried :

- `TLogin`
- `TMain`
- `TList`
- `TSelect`
- `TAnonymization`
- `TSQL`
- `TRunSQL`
- `TSetSeed`
- `TVirtualData`



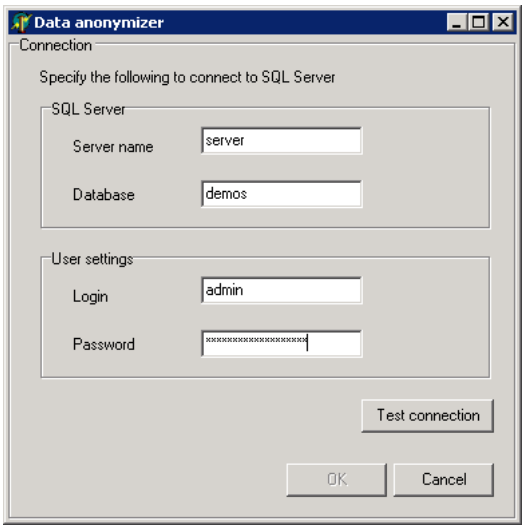
Obrázok 5: Ukážka naimplementovanej aplikácie Data Anonymizer

- `TLoadSave`

ktoré zabezpečujú chod aplikácie. Bližšie si popíšeme jednotlivé triedy.

5.1 Trieda `TLoginForm`

Trieda `TLoginForm` zabezpečuje pripojenie k SQL serveru a prihlásenie užívateľa k databáze. Všetky údaje – názov SQL servera, názov databázy, meno užívateľa a jeho heslo sú povinné údaje, ktoré sú potrebné na prihlásenie užívateľa k databáze. Na prihlásenie využívame komponent `TADOConnection`, ktorý zabezpečuje požadovaný proces prihlasovania. Metóda `TestConnection` slúži na otestovanie spojenia a pripojenia s aktuálnymi parametrami. Pripojenie k databáze je povolené len v prípade úspešného testu spojenia. V prípade odchytenej chyby trieda upozorní na nesprávne parametre pripojenia a nedovolí ďalej pokračovať.



Obrázok 6: Prihlasovací formulár aplikácie Data Anonymizer

Behom používania aplikácie je možné kedykoľvek zmeniť ľubovoľné parametre pripojenia v menu hlavného okna aplikácie v záložke `File Login`.

5.2 Trieda `TMain`

Trieda `TMain` je kľúčová trieda projektu, ktorá obsluhuje hlavné okno aplikácie a postupne dynamicky vytvára ďalšie dopytované triedy, formuláre. Na úvod táto trieda vytvorí spomínané pripojenie k databáze. SQL dotazy na databázu sú vykonávané pomocou komponentu `TADOQuery`. Trieda `TMain` po prihlásení vykoná hlavný dotaz na databázu, ktorý zo systémových tabuliek informačnej schémy `INFORMATION_SCHEMA`, ktoré existujú na každej databáze, získa všetky potrebné údaje, s ktorými aplikácia pracuje.

Daný SQL dotaz :

```
SELECT
  co.column\_name,
```

```
  co.table\_schema+'.'+ta.table\_name,
  Fkeys.FK\_Column ,
  Fkeys.PK\_Table ,
  Fkeys.PK\_Column,
  Fkeys.Constraint\_Name,
  Pkeys.column\_name,
  co.is\_NULLAble,
  co.data\_type,
  co.character\_maximum\_length
FROM
  INFORMATION\_SCHEMA.TABLES ta
  INNER JOIN INFORMATION\_SCHEMA.COLUMNS co on ta.table\_name=
    co.table\_name and ta.table\_type='BASE TABLE'
LEFT JOIN
(
  SELECT
    K\_Table = FK.TABLE\_NAME,
    FK\_Column = CU.COLUMN\_NAME,
    PK\_Table = PK.TABLE\_NAME,
    PK\_Column = PT.COLUMN\_NAME,
    Constraint\_Name = C.CONSTRAINT\_NAME
  FROM
    INFORMATION\_SCHEMA.REFERENTIAL\_CONSTRAINTS C
    INNER JOIN INFORMATION\_SCHEMA.TABLE\_CONSTRAINTS FK ON
      C.CONSTRAINT\_NAME = FK.CONSTRAINT\_NAME
    INNER JOIN INFORMATION\_SCHEMA.TABLE\_CONSTRAINTS PK ON
      C.UNIQUE\_CONSTRAINT\_NAME = PK.CONSTRAINT\_NAME
```

```
    INNER JOIN INFORMATION\_SCHEMA.KEY\_COLUMN\_USAGE CU ON
      C.CONSTRAINT\_NAME = CU.CONSTRAINT\_NAME
    INNER JOIN
    (
      SELECT
        i1.TABLE\_NAME,
        i2.COLUMN\_NAME
      FROM
        INFORMATION\_SCHEMA.TABLE\_CONSTRAINTS i1
        INNER JOIN INFORMATION\_SCHEMA.KEY\_COLUMN\_USAGE i2 ON
          i1.CONSTRAINT\_NAME = i2.CONSTRAINT\_NAME
      WHERE
         i1.CONSTRAINT\_TYPE = 'PRIMARY KEY'
    ) PT ON PT.TABLE\_NAME = PK.TABLE\_NAME
) Fkeys on Fkeys.fk\_column=co.column\_name
    and Fkeys.k\_table=ta.table\_name
LEFT JOIN
(
  SELECT
    i1.TABLE\_NAME,
    i2.COLUMN\_NAME
  FROM
    INFORMATION\_SCHEMA.TABLE\_CONSTRAINTS i1
    INNER JOIN INFORMATION\_SCHEMA.KEY\_COLUMN\_USAGE i2 ON
      i1.CONSTRAINT\_NAME = i2.CONSTRAINT\_NAME
  WHERE
    i1.CONSTRAINT\_TYPE = 'PRIMARY KEY'
```

```
) Pkeys on Pkeys.column\_name=co.column\_name
    and Pkeys.table\_name=ta.table\_name
ORDER BY
  co.column\_name,
  co.ordinal\_position
```

Sú to údaje o všetkých stĺpcoch tabuliek danej databázy, podľa poradia v danom selecte : názov stĺpca, názov tabuľky spolu so schémou, informácia, či daný stĺpec je cudzím kľúčom - v tom prípade názov tabuľky a stĺpca, kde je tento údaj primárnym kľúčom, informácia, či je daný stĺpec primárnym kľúčom, informácia, či má daný stĺpec povolené ukladať hodnoty *NULL*, typ stĺpca a v prípade typu *Varchar*, maximálnu dĺžku.

Tieto údaje trieda uchováva v poli typu `TData`, ktoré obsahuje záznamy typu `TColumn`. Do tejto štruktúry sa okrem uvedených dát následne ukladajú aj všetky informácie o budúcom nastavení anonymizácie konkrétneho stĺpca, teda typ anonymizácie s príslušnými parametrami, ktoré zadá užívateľ.

Štruktúra databázy je zobrazená v ľavej časti hlavného okna pomocou komponentu `TTreeView`, ktorý stromovito v postupnosti *Databáza*
Tabuľka
Stĺpec zobrazuje formou rozklikávania všetky stĺpce databázy.

Pod týmto komponentom v ľavej časti okna sú použitím komponentu `TValueListEditor` dynamicky zobrazované ostatné dôležité údaje o stĺpcoch.

V hlavnej časti okna táto trieda do záložiek (`TTabSheet`)

- `TsList`
- `TsSelect`
- `TsAnonymization`

- `TsSQL`
- `TsRunSQL`

komponentu `TPageControl` vytvára príslušné triedy a formuláre.

V hornej časti okna je štandardné menu typu `TMainMenu`, ktoré zabezpečuje prístup k pomocnej funkcionalite aplikácie, akou je nastavovanie seedu náhodného generátora, ukladanie a načítavanie projektov, prácu s virtuálnymi dátami a podobne.

5.3 Trieda `TList`

Trieda `TList` zobrazuje (`TStringGrid`) zoznam stĺpcov, pre ktoré bol užívateľom nastavený ľubovoľný typ anonymizácie. Táto záložka sa automaticky obnovuje pri vytváraní nových pravidiel anonymizácie a má výlučne informatívny charakter o stĺpcoch, type a parametroch anonymizácie.

5.4 Trieda `TSelect`

Trieda `TSelect` zobrazuje (`TStringGrid`) obsah ľubovoľne zvolenej tabuľky. Dynamicky sa kanonickým selectom dotazuje na vyžiadanú tabuľku a informačne zobrazuje obsah celej tabuľky. Táto funkcionalita len dopĺňa kompletnosť aplikácie, umožňuje užívateľovi bez potreby využívania inej aplikácie prezerať obsah tabuliek.

5.5 Trieda `TAnonymization`

Trieda `TAnonymization` zabezpečuje výber typu anonymizácie a jej nastavenie pre jednotlivé stĺpce. Formou záložiek, kde sú odlíšené 4 rôzne typy stĺpcov

- `TsInteger`
- `TsVachar`
- `TsDateTime`
- `TsFloat`

užívateľ vyberá konkrétny typ anonymizácie v rámci príslušnej záložky. Výber aktívnej záložky sa pre štandardné typy stĺpcov vykoná automaticky metódou `SetColumnIndex` podľa typu stĺpca. Je však možná aj manuálna zmena záložky užívateľom podľa potreby. Konkrétne typy anonymizácie implementované v tejto aplikácií sú napríklad :

- *No mask* – defaultná hodnota, daný stĺpec sa nebude anonymizovať
- *NULL* – v prípade, že daný stĺpec povoľuje ukladanie hodnoty *NULL*, tak sa celému stĺpcu nastaví táto hodnota
- *Constant value* – podľa typu stĺpca sa celý stĺpec anonymizuje jednou zadanou hodnotou
- *Noise* – zašumenie, ktoré zmení reálny údaj posunutím hodnoty v rámci zadaného rozsahu
- *Random* – nahradí pôvodnú hodnotu náhodnou hodnotou zo zadaného intervalu
- *Virtual data* – nahradí dáta virtuálnymi dátami z vybraného predpripraveného zoznamu
- *Shuffle* – náhodne prehádže existujúce dáta

Pri stĺpoch typu *Varchar* je možné nastaviť aj iné typy anonymizácie :

- *User defined text with same length* – zamení textové hodnoty v danom stĺpci za zreplikované reťazce z užívateľom zadaného textu, pričom zachová pôvodnú dĺžku reťazca (výhodne pri poznámkach a dlhších textových poliach, kde sa vďaka zachovaniu pôvodnej dĺžky dosahuje opticky lepší dojem reálnych dát)

- *Lorem ipsum random length* – nahradí pôvodné reťazce známym bezvýznamovým textom Lorem ipsum v náhodnej dĺžke z intervalu zadaného užívateľom.

V prípade, že sa jedná o cudzí kľúč, tak už v tomto bode táto trieda nedovolí vložiť napríklad pri *Constant value* ľubovoľnú hodnotu, ale upozorní užívateľa a ponúkne mu (`TComboBox`) na výber len také hodnoty, ktoré obsahuje príslušný stĺpec primárnej tabuľky.

Pri *Random* alebo *Noise* metóde sa tento test vykonáva až pri behu konkrétneho SQL príkazu, keďže na tomto mieste to ešte nie je realizovateľné. Update riadku je v tomto prípade vykonaný len vtedy, ak sa v danom prípustnom intervale nachádza aspoň jedna hodnota zo stĺpca primárnej tabuľky. Inak sa update nevykoná a postupuje sa na ďalší riadok. Túto funkcionalitu má na starosti trieda TSQL.

5.6 Trieda `TSQL`

Trieda `TSQL` zabezpečuje vytváranie SQL kódov, ktoré vykonávajú samotný proces anonymizácie. Metóda `SQLGenerator` prechádza dátovú štruktúru a pre každý stĺpec, ktorý má nastavený nejaký typ anonymizácie (nie `No mask`), vygeneruje príslušný kód. Každý typ anonymizácie je interne reprezentovaný prirodzeným číslom a parametrami, ktoré sú v danom prípade potrebné. Predpripravené SQL kódy, ktoré využívajú dynamické vytváranie

selectov a prácu s kurzormi, sú v tomto kroku prispôsobené pre konkrétnu tabuľku a stĺpec. Pri virtuálnych dátach načíta vybraný textový súbor a vloží ho do pomocnej tabuľky, s ktorou neskôr pracuje. Jednou z možností tejto triedy je pre potreby užívateľa aj vygenerovanie kompletného SQL kódu do súboru. Manuálne vykonanie tohto kódu je ekvivalentné so spustením tohto procesu v rámci aplikácie v nasledujúcej triede.

5.7 Trieda `TRunSQL`

Trieda `TRunSQL` slúži na spustenie, vykonanie vygenerovaných SQL kódov. Tieto kódy spúšťa pomocou komponentu `TADOQuery` metódou `ExecSQL`. Ak z ľubovoľného dôvodu nastane chyba pri update niektorého zo stĺpcov, táto chyba je odchytená a proces pokračuje ďalej. Priebežný beh je zobrazovaný komponentom `TMemo` a pre každý stĺpec informuje o úspešnosti procesu.

Menšie pomocné triedy, ktoré v rýchlosti spomenieme :

5.8 Trieda `TsetSeed`

Trieda `TSetSeed` zabezpečuje nastavovanie seedu pre generátor náhodných čísel. Vďaka tejto funkcionalite, keď sa na začiatku behu SQL skriptov iniciuje náhodný generátor stále rovnakou hodnotou dosiahneme stav, keď pri opakovanom behu rovnakého SQL kódu viackrát po sebe dostávame na rovnakom vstupe rovnaký výstup. Je to jedna z požadovaných vlastností, ktorá pomáha pri testovaní.

5.9 Trieda `TVirtualData`

Trieda `TVirtualData` zabezpečuje funkcionalitu pri vytváraní nových virtuálnych dát, ktoré dokáže užívateľ v implementovanom prostredí jednodu-

cho pridávať, meniť a následne využívať. Okrem množiny základných súborov najčastejšie využívaných virtuálnych dát, akými sú mužské a ženské mená a priezviská, ulice, mestá, telefónne čísla, rodné čísla, štáty si tak užívateľ môže vytvárať a pridávať ľubovoľné iné. Toto dodáva aplikácií značnú rozšíriteľnosť a silu, keď si užívateľ môže vytvoriť špecifické zoznamy z vlastnej domény, v ktorej pracuje.

5.10 Trieda `TLoadSave`

Trieda `TLoadSave` slúži na uloženie resp. načítanie vytvoreného projektu.

Táto aplikácia spĺňa základné požiadavky, ktoré by mala aplikácia na anonymizovanie dát obsahovať. Je jednoduchá na ovládanie a vďaka možnosti vytvárania vlastných virtuálnych súborov v značnej miere využiteľná. Dala by sa samozrejme na mnohých miestach ešte vylepšiť a implementovaním ďalších zaujímavých častí rozšíriť, veríme však, že aj napriek tomu bude svoju primárnu úlohu ochrany dát naimplementovanými postupmi plniť spoľahlivo.

6 Záver

Táto práca mala za cieľ oboznámiť čitateľa s problematikou ochrany osobných údajov, ktorá je v dnešnej dobe veľmi dôležitá. Práca popisuje jednu z používaných metód ochrany osobných údajov - anonymizáciu dát, jej možnosti a dôvody použitia v praxi, prípady, kedy jej jednoduché nasadenie môže priniesť požadované výsledky. Podarilo sa nám realizovať hlavný zámer, implementovať aplikáciu, ktorá dokáže anonymizovať databázové tabuľky, jednak pomocou metódy virtuálnych dát a množstvom iných, naimplementovaných voliteľných metód. Zadefinovali sme mieru anonymizácie, ktorá matematickou metódou dokáže posúdiť, či daná sada údajov dokáže jednoznačne identifikovať osobu.

Veríme, že cieľ tejto práce sa nám podarilo naplniť a naimplementovaná aplikácia bude nápomocná, či už v reálnom nasadení alebo na študijné účely.

Referencie

[1] HILL, David G., Data protection, CRC - Press, 2009

[2] Zákon Národnej rady Slovenskej republiky č. 428/2002 Z. z. o ochrane osobných údajov, http://www.dataprotection.gov.sk, 2002

[3] Správa o stave ochrany osobných údajov 2005 - 2007, Úrad na ochranu osobných údajov SR, http://www.dataprotection.gov.sk/buxus/docs/31082007.pdf, 2007

[4] Zákon Národnej rady Slovenskej republiky č. 301/1995 Z. z. o rodnom čísle, http://www.zbierka.sk/zz/predpisy, 1995

[5] Stanovisko Úradu na ochranu osobných údajov Slovenskej republiky k zverejňovaniu rodných čísel fyzikých osôb, http://www.dataprotection.gov.sk/buxus/docs/MSSRst150306v2.pdf, 2006

[6] Smernicou 95/46/ES Európskeho parlamentu a rady z 24. októbra 1995 o ochrane jednotlivcov vzhľadom na spracovávanie osobných údajov a o voľnom pohybe takýchto údajov, http://www.telecom.gov.sk/externe/legeu/priemys/95-0046.pdf, 1995

[7] Balancing privacy and information utility in data anonymisation, http://www.eifel.org/publications/eportfolio/proceedings2/ep2007/papers/digital-identity-and-privacy/balancing-privacy-and-information-utility-in-data-anonymisation-1, 2007

[8] KPMG - Data loss barometer, http://www.datalossbarometer.com/docs/KPMG_Data_Loss_Barometer_-_Issue_3_-_November_2010.pdf, 2010

[9] Krádeže firemních dat rostou v dobách ekonomické nejistoty, http://businessworld.cz/business-rizeni-podniku/Kradeze-firemnich-dat-rostou-v-dobach-ekonomicke-nejistoty-4675, 2009

[10] Ak Firma hackerskému útoku nezabráni, mala by ho aspoň vysvetliť, http://technologie.etrend.sk/technologie/ak-firma-hackerskemu-utoku-nezabrani-mala-by-ho-aspon-vysvetlit.html, 2003

[11] Unikli osobné údaje troch miliónov ľudí, http://www.sme.sk/c/1291222/unikli-osobne-udaje-troch-milionov-ludi.html#ixzz1J7bQS3ll, 2004

[12] Vyjadrenie k povinnému zverejňovaniu informácií http://www.dataprotection.gov.sk/buxus/docs/Vyjadrenie02032011.pdf, 2011

[13] Uplatňovanie zákona č.52/1998 Z.z. v praxi http://www.dataprotection.gov.sk/buxus/docs/vs_2002_kap_4.pdf, 2002

[14] Domovská stránka produktu Data Masker. Dostupné na internete : http://www.datamasker.com

Prílohy

Príloha č. 1

Meno Priezvisko, Adresa Trvalého Pobytu

Organizácia, a.s.
Ulica 47
PSC Bratislava

1. 12. 2010

Vec: **Žiadosť o odpis osobných údajov**

Dobrý deň,

na základe §20 ods. 1 písm. b), c) zákona 428/2002 Z.z. o ochrane osobných údajov žiadam o poskytnutie výpisu všetkých údajov o mojej osobe, ktoré Organizácia spracúva, ako aj presné informácie o zdroji (zdrojoch), z ktorého tieto údaje boli získané.

identifikácia žiadateľa:

meno a priezvisko:	**Meno Priezvisko**
dátum narodenia:	**1.1.1970**
trvalý pobyt:	**Adresa Trvalého Pobytu**

Úhradu vo výške nepresahujúcej výšku materiálnych nákladov spojených so zhotovením tohoto výpisu a jeho odoslaním som ochotný uhradiť prostredníctvom bankového prevodu alebo iným vhodným spôsobom.

Ďakujem,

Meno Priezvisko

Príloha č. 2

DÔVERA+
ZDRAVOTNÁ POISŤOVŇA

Vážený pán
PETER JUHÁSZ
JÁNA. A. KOMENSKÉHO 14
071 01 MICHALOVCE

V Nitre, dňa 24.03.2011

Naša značka: 2400BK7203 Vybavuje: Zuzana Matejkovičová tel:037/ 64 13 359

Vec: Zaslanie odpisu osobných údajov

Vážený pán Juhász,

na základe Vašej žiadosti Vám zasielam odpis osobných údajov, ktoré evidujeme a spracovávame v našom informačnom systéme:

Osobné údaje získané z prihlášky poistenca na verejné zdravotné poistenie k 01.01.2009:
Meno a priezvisko: PETER JUHÁSZ
dátum narodenia: 11.01.1985
rodné číslo:
adresa poistenca: JÁNA. A. KOMENSKÉHO 14, 07101 MICHALOVCE

Informácie o platiteľovi poistného boli zaevidované na základe Oznámenia poistenca/platiteľa poistného, ktoré bolo odovzdané v Dôvere a. s.

S pozdravom

Zuzana Matejkovičová
špecialista kontaktného centra
Krajská pobočka Nitra
DÔVERA zdravotná poisťovňa, a. s.

Príloha č. 3

Zmluva o pripojení

Podľa § 43 zákona č. 610/2003 Z. z. o elektronických komunikáciách v znení neskorších predpisov (ďalej len "Zákon") medzi spoločnosťou Telefónica O2 Slovakia, s r o ("ďalej len poskytovateľ") a účastníkom

e01#1-51JVP6

1. ÚČASTNÍK - FYZICKÁ OSOBA NEPODNIKATEĽ

Oslovenie	Vážený pán		
Meno	Peter		
Priezvisko	Juhász		
Ulica č	Jána A. Komenského 1332/14		
Mesto	Michalovce	PSČ	07101
Štát	Slovensko		
Štátna príslušnosť	SLOVENSKÁ		
Rodné číslo		OP	
Druhý doklad	Vodičský preukaz	Číslo	
Platnosť dokladu	21.08.2016		

2. FAKTURAČNÁ ADRESA

Meno	Peter		
Priezvisko	Juhász		
Ulica (obec)	Jána A. Komenského		
Súpisné číslo	1332/14	PSČ	07101
Mesto	Michalovce		
Štát	Slovensko		

3. KONTAKT PRE FAKTURÁCIU

Telefón +421918132325

4. SPÔSOB PLATBY

☐ Bankový prevod ☐ Poštová poukážka ☒ Inkaso

Číslo bankového účtu 0000002457132053

Špec. Symbol Kód banky 0200

Limit 50

5. SPLNOMOCNENÁ OSOBA / ZÁKONNÝ ZÁSTUPCA

Titul pred menom	
Meno	
Priezvisko	
Titul za menom	
Rodné číslo	
Platnosť dokladu	Štátna príslušnosť
Telefón	
Email	

6. SLUŽBY

Cena za aktiváciu (s DPH)	0,01 EUR / 0,30 Sk
Referenčné číslo zákazníka	80562177
Účastnícke číslo	
Číslo SIM karty	
Program služieb	O2 Fér - fakturovaný
Dátum aktivácie	21.04.2010

Konverzný kurz 1 EUR = 30,1260 Sk

7. ZASIELANIE ELEKTRONICKEJ FAKTÚRY

Email

Telefón

Uvedením emailovej adresy Účastník súhlasí so zasielaním elektronickej faktúry

8. VYHLÁSENIE ZÁSTUPCU ÚČASTNÍKA

Vyhlasujem, že som oprávnený konať v mene a na účet Účastníka na základe písomného splnomocnenia zo dňa a úradne overeným podpisom Účastníka alebo na základe zákona alebo na základe rozhodnutia štátneho orgánu. Dokumenty preukazujúce moje oprávnenie prikladám k tejto žiadosti.

9. SPOLOČNÉ USTANOVENIA

1. Zmluva nadobúda platnosť aktiváciou SIM karty okrem čl. 5, 6, 7 Spoločných ustanovení tejto Zmluvy, ktoré nadobúdajú platnosť dňom podpisu Zmluvy oboma zmluvnými stranami. Poskytovateľa sa zaväzuje vykonať aktiváciu SIM karty do 30 dní od podpisu Zmluvy. Ak v tejto lehote nastane niektorá zo skutočností v zmysle § 42 ods. 1 písm. c) Zákona, Poskytovateľ si vyhradzuje právo neuzavrieť zmluvu.

2. Účastník v zmysle osobitných podmienok pre prenositeľnosť čísla spoločnosti Telefónica O2 Slovakia, s.r.o. berie na vedomie a súhlasí s tým, že ak sa neuskutoční prenesenie Prenášaného čísla budú práva a povinnosti Poskytovateľa a účastníka zo zmluvy nadne trvať vo vzťahu k telefónnemu číslu pridelenému pôvodne poskytovateľom z vlastnej číselnej množiny, ak je takéto telefónne číslo pridelené. Realizáciu technického prenesenia Prenášaného čísla Poskytovateľ oznámi Záujemcovi prostredníctvom SMS správy najneskôr v deň tejto realizácii.

3. Miestom poskytovania služieb je územie Slovenskej republiky podľa Špecifikácie obsiahnutej vo všeobecných podmienkach. Cena za poskytované služby je stanovená platným cenníkom. Zmluva sa uzatvára na dobu neurčitú. Podpisom potvrdzujem, že som sa oboznámil/a so Všeobecnými podmienkami a platným Cenníkom, ktoré su neoddeliteľnou súčasťou zmluvy.

4. Zmluvné strany sa v zmysle § 262 ods. 1 zákona č. 513/1991 Zb. Obchodný zákonník v znení neskorších predpisov (ďalej len "Obchodný zákonník") dohodli, že
a) Zmluva
b) Vzťahy ktoré vznikli na základe Zmluvy a ktoré nie sú výslovne upravené,
Sa spravujú Obchodným zákonníkom okrem práv a povinností výslovne upravených Zákonom.
Sudom, ktorý má právomoc na súdne konanie vedené proti Účastníkovi, ktory nie je občanom Slovenskej republiky je vždy súd Slovenskej republiky miestne príslušný podľa sídla Poskytovateľa. Obdobne to platí ja pre právomoc Telekomunikačného úradu Slovenskej republiky pre mimosúdne vyrovnanie sporu medzi Poskytovateľom a zahraničným Účastníkom v zmysle Zákona. Pokiaľ v tejto zmluve nie je dohodnuté inak, príslušnosť a pravomoc súdu sa spravuje zákonom č. 99/1963 Zb., Občiansky súdny poriadok v znení neskorších predpisov.

5. Účastník svojim podpisom potvrdzuje, že súhlasí s tým, aby jeho osobné údaje v rozsahu uvedenom v tejto Zmluve a prevádzkové a lokalizačné údaje vzťahujúce sa na telefónne číslo pridelené Poskytovateľom na základe tejto žiadosti, Poskytovateľ spracúval na marketing služieb (vrátane spracúvania prevádzkových a lokalizačných údajov podľa telefónnych čísel volaných z prístupu k sérii znadeného na základe tejto zmluvy o pripojení). Tento súhlas mám právo kedykoľvek odvolať.
V prípade, že nesúhlasíte, označte

☒ NIE

6. V zmysle ustanovení § 10 ods. 6 zák. č. 428/2002 Z.z. o ochrane osobných údajov v znení neskorších predpisov Účastník súhlasí s kopírovaním, skenovaním a archivovaním predložených dokladov na účely v zmysle Všeobecných podmienok.
V prípade, že nesúhlasíte, označte

☐ NIE

7. Účastník berie na vedomie, že jeho osobné, prevádzkové a lokalizačné údaje môže poskytovateľ spracúvať prostredníctvom tretích osôb ako sprostredkovateľov Poskytovateľa

8. Účastník žiada o nezverejnenie telefónneho čísla v zozname účastníkov.

☒

10. PODPIS ÚČASTNÍKA

Podpisom potvrdzujem prevzatie SIM karty

Vlastnoručný podpis Účastníka / oprávneného zástupcu Pečiatka

Bratislava, dňa 21.04.2010

11. ZA POSKYTOVATEĽA

Podpis Pečiatka

Zmluvu prevzal Ludmila Tkáčová

Kód predajcu PPZ05.00001.00001

Bratislava, dňa 21.04.2010

Telefónica O2 Slovakia, s.r.o.
Značková predajňa Aupark
Einsteinova 18
851 01 Bratislava
IČ: 35 848 863
DIČ: 2020216748
227

Zmluva sa vyhotovuje v 3 originálnoch.