FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Název:	Modulární anonymizér streamu dat
Student:	Vojtěch Skalák
Vedoucí:	Ing. David Knap
Studijní program:	Informatika
Studijní obor:	Bezpečnost a informační technologie
Katedra:	Katedra počítačových systémů
Platnost zadání:	Do konce letního semestru 2020/21

Pokyny pro vypracování

Cílem práce je navrhnout modulární architekturu a vzorovou implementaci anonymizéru, který zajistí odstranění citlivých údajů ze vstupního streamu dat na základě zadaných vzorů.

1. Porovnejte možné přístupy k anonymizérům založeným na vyhledávání dat podle vzorů.
2. Prozkoumejte vzorky telekomunikačních dat dodané vedoucím práce a navrhněte vhodnou detekci obsažených citlivých údajů.
3. Navrhněte architekturu vlastního anonymizéru s ohledem na rychlost zpracování a modularitu vstupního formátu, výstupního formátu a konfigurace předloh dat k anonymizaci.
4. Pro zvolenou architekturu připravte prototypovou implementaci s alespoň jedním modulem každého typu a anonymizér vyzkoušejte na dodaných datech.

Seznam odborné literatury

Dodá vedoucí práce.

prof. Ing. Pavel Tvrdík, CSc.
vedoucí katedry

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
děkan

V Praze dne 22. října 2019

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalářská práce

Modulární anonymizér streamu dat

Vojtěch Skalák

Katedra počítačových systémů
Vedoucí práce: Ing. David Knap

4. června 2020

Poděkování

Chtěl bych poděkovat vedoucímu své bakalářské práce Ing. Davidu Knapovi za cenné konzultace a poskytnuté ukázky dat. Děkuji také Ing. Daně Vynikarové, Ph.D. za pomoc s formální úpravou textu.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisů, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo"), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené.

V Mníšku pod Brdy dne 4. června 2020 . . . . . .

Odkaz na tuto práci

Skalák, Vojtěch. *Modulární anonymizér streamu dat*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2020.

Abstrakt

V této bakalářská práci se zabývám anonymizací streamu dat. Nejprve v teoretické části popisuji možné přístupy k anonymizaci. Analýzou možných řešení jsem došel k efektivní metodě anonymizace dat, která vyhovuje současným standardům. V rešerši se dále zabývám nároky na proces anonymizace a na výsledná data. Věnuji se i algoritmům hledání vzorů v textu, které porovnávám z hlediska efektivity hledání více vzorů v datech. V praktické části práce jsem naprogramoval algoritmus dle poznatků z teoretické části a dva moduly pro dva formáty dat. Na závěr práce jsem uvedl doporučení pro tvorbu dalších modulů a jejich propojení s vytvořeným programem.

Klíčová slova anonymizace dat, telekomunikační data, anonymizace proudu dat, modulární anonymizace, bezpečnost dat, zpracování dat

Abstract

My bachelor thesis study the problem of modular data stream anonymization. In the theoretical part, I describe possible aproaches towards data anonymization. After analyzing several solutions, I picked an effective method to anonymize data, which conforms the task given. I also compared several

pattern-searching algorithms according to their effectivity. In practical part, I implemented algorithm due to the research done in theoretical part and two modules for two data formats. In future, there could be more modules added for different data types, according to my instructions at the final part of my thesis.

Keywords data anonymization, telecomunication data, data stream anonymization, modular anonymization, data security, data processing

Obsah

Úvod

Tématem mojí bakalářské práce je Modulární anonymizér proudu dat. Téma anonymizace dat je vzhledem k narůstajícímu zájmu společnosti o osobní údaje, které sdělují firmám, čím dál tím víc aktuální. Společnosti, které data sbírají, s nimi potřebují manipulovat, což je obtížnější, pokud data nejsou anonymní. Práce s daty, která jsou anonymní, je mnohem jednodušší a nevztahují se na ni tak přísná pravidla.

Anonymizovaná data jsou tak pro firmy mnohem lepší. Lépe se uchovávají, protože na jejich uchovávání nejsou tak vysoké nároky. Zároveň ztráta takových data není takový problém, jako v případě dat neanonymních, protože nevzniká škoda lidem, od kterých byly data sebrána. Použití pro statistické účely je tak jednodušší i levnější a anonymizace je dnes pro velké firmy nutností.

K výběru tématu mě motivovala problematika sběru osobních dat a jejich bezpečnost z pohledu velkých společností. Bezpečnému nakládání s daty bych se rád věnoval i v budoucnu. V teoretické části této práce chci mimo výběru nejvhodnějšího algoritmu pro anonymizaci dat i prozkoumat data z pohledy zákona o elektronických komunikacích a GDPR, protože to je oblast, která je pro anonymizaci dat klíčová.

Ve své práci se budu zabývat nutností anonymizace dat vzhledem k GDPR a Zákonu o elektronických komunikacích. GDPR omezuje možnost nakládání s osobními daty, která nejsou anonymní a ukládá firmám, jak se k takovým datům mají chovat, včetně jejich skladování. Zákon o elektronických komunikacích zase ukládá firmám, které typy informacím mohou uchovávat a po jakou dobu.

V teoretické části se budu nejprve zabývat anonymními daty z pohledu zákona. Důležité je zejména jaká data jsou zákonem brána jako anonymní a jaké jsou nároky na bezpečný proces anonymizace. Dalším bodem je rozdíl mezi anonymními a konkrétními daty z pohledu ukládání dat.

Dále budu řešit jednotlivé možnosti anonymizace a algoritmy, které se dají

použít pro anonymizaci proudu dat. Budu se zabývat hledáním vzorů v textu i rychlým zpracováním dat a mazáním vybraných údajů. Jedním z bodů bude analýza vzorků dodaných vedoucím práce. Nakonec navrhnu řešení a konkrétní implementaci modulární anonymizéru na základě předchozích analýz.

V praktické části mojí práce naprogramuji anonymizér dat a dva moduly pro konkrétní typy dat.

Kapitola 1

Cíl práce

Cílem rešeršní části práce je získat přehled o technikách používaných při anonymizaci dat, zhodnotit a vybrat nejvhodnější algoritmus, nebo kombinaci algoritmů, pro anonymizaci proudu dat, ověřit správnost procesu anonymizace oproti GDPR a Zákonu o elektronických komunikacích a navrhnout způsob implementace vybraného algoritmu.

Dále prozkoumat druhy dat a jejich vývoj se zaměřením na citlivá a anonymní data, na základě čehož analyzuji data dodaná vedoucím práce.

Výsledný algoritmus bude mít průměrně lineární složitost, tj. jeho složitost bude při velkém počtu běhů v průměru lineární, i když nejhorší případ může být horší než lineární. Algoritmus nebude data ukládat, pouze mazat ty části dat, které budou definovány na vstupu.

Dalším cílem je analyzovat algoritmy pro vyhledávání vzorů v textu, o které by se dal algoritmus v budoucnu rozšířit.

Cílem praktické části je naprogramovat modulární anonymizér proudu dat založený na poznatcích z analýz obsažených v teoretické části. Anonymizér se bude skládat z vnitřní logiky a z alespoň dvou modulů. Program bude navržen tak, aby se daly přidávat další moduly. Ve výsledku program zvládne anonymizaci dat zaslaných v ukázce vedoucím práce.

Kapitola 2

Analýza a návrh

2.1 Body literární rešerše

2.1.1 Anonymizace dat

V této části rozeberu pojem anonymizace a její použití na různé druhy dat včetně toho, kdy je data možné brát jako anonymizovaná. Dále rozeberu, jaké přínosy má anonymizace pro bezpečnost dat. Nakonec srovnám různé přístupy k anonymizaci.

Výsledkem této části bude přehled technik anonymizace, jejich použití a důvody k využívání anonymizace. Důležitý bod této části rešerše je výběr techniky anonymizace, která je nejvhodnější pro můj program.

2.1.2 Hledání vzorů

Tato část rešerše se bude zabývat algoritmy pro hledání vzorů v textu. Proberu prostorovou a časovou složitost jednotlivých algoritmů s přihlédnutím k podobě dat. Výsledkem bude srovnání jednotlivých algoritmů a návrh optimálního řešení.

2.1.3 Obecnost anonymizace

V této části se budu zabývat definicí obecnosti a zajištění obecností při zpracovávání dat anonymizérem, stejně jako jeho modularitou. Výsledkem bude návrh implementace obecnosti a modularity pro programování anonymizéru.

2.1.4 Data

V této části se zaměřím na definování dat a jejich vývoje. Poté rozeberu jednotlivé druhy dat dle několika obecných rozdělení, včetně toho, jaká je doporučená ochrana těchto druhů dat a možnosti této ochrany ve vztahu k anonymizaci.

Nakonec analyzuji vzory dat dodané vedoucím práce vzhledem k předchozím rozdělením a z pohledu implementace jejich zpracování

Výsledkem této části bude podrobná analýza dat a jejich doporučeného zpracování se zaměřením na dodané vzory dat.

2.1.5 Osobní data z pohledu zákona

Část, ve které budu analyzovat data z pohledu GDPR a Zákona o elektronických komunikacích. Zaměřím se zejména na operace, které je možné s daty provádět a opatření, které pro různé druhy dat ukládá zákon.

Cílem je určit, jak budou data vypadat, případně co neobsahovat, aby se dala považovat za anonymní a zároveň byla použitelná pro statistické účely. Výsledkem této části bude návrh vhodného přístupu k odstraňování citlivých údajů v konkrétní sadě dat, kterou jsem obdržel od vedoucího práce.

2.2 Anonymizace dat

V této části mé bakalářské práce proberu anonymizaci dat, jak je definována, kdy jsou data považována za anonymní a jaké postupy jsou považovány za přípustné při anonymizace dat, bez ohledu na konkrétní algoritmus. Použití konkrétních algoritmů budu zkoumat v sekci Hledání vzorů v textu. Téma anonymních dat dle zákona je dále rozvedeno v sekci Osobní data z pohledu zákona.

2.2.1 Anonymizace obecně

Anonymizace dat je dle [1, vlastní překlad]: *„Proces de-identifikace citlivých dat při zachování jejich formátu a typu dat“*. To znamená, že z dat, která projdou procesem anonymizace, nejde identifikovat konkrétní osobu. K podobě anonymizovaných dat, píše [1, vlastní překlad]:

„Anonymizovaná data mohou být realistická nebo náhodná sekvence znaků. Stejně tak může být výstup anonymizace deterministický, to znamená, že bude dávat pokaždé stejný výstup. Tohle všechno záleží na technikách použitých během anonymizace.“

Přístupy a techniky anonymizování dat budu podrobně rozebírat v sekci Přístupy k anonymizaci.

2.2.1.1 Identifikovatelnost

Důležitý pojem v této oblasti je identifikovatelnost člověka jako jedince. Dle [2, str. 8, vlastní překlad]:

„Jedinec v datovém setu je identifikovatelný, pokud je rozumné očekávat, že jedinec bude identifikován, a to buď pomocí dat přímo obsažených v datovém setu, nebo v kombinaci s ostatními informacemi (externími nebo v držení útočníka).“

Z toho vyplývá, že pro zabránění identifikování člověka jako jedince, nestačí odstranit údaje, které jedince přímo identifikují, jako jsou jméno, příjmení nebo adresa, ale i informace, ze kterých by mohl útočník usoudit totožnost jedince spolu s dalšími informacemi.

Mezi tyto údaje by mohlo patřit datum a místo, kde se uživatel připojil se svým telefonem. Samy o sobě tyto dvě informace neidentifikují jedince, ale spolu se záznamy kamer nebo například knihou návštěv v hotelu by mohly vést k identifikaci jedince, protože by útočník dokázal zjistit, kde se v danou chvíli nacházel.

2.2.1.2 Soukromí

Dalším pojmem, který úzce souvisí s anonymizací, je soukromí. Dle [3] je řešení otázky soukromí velice staré, protože je možné je nalézt už ve spisech starých řeckých filozofů, kteří *„rozlišovali vnější (veřejné) a vnitřní (soukromé)“*. Jeho právní definice je podle [3] mnohem novější, a to z roku 1890, kdy ho popsali Warren a Brandeis ve své knize *„The Right to Privacy“* [4], kteří v té samé knize uvádějí, že se právo na soukromí v právní podobě ve Spojených státech amerických už vyskytovalo, jen nebylo zvlášť implementováno do právního řádu.

Bylo implementováno v zákonech, které nejsou zaměřeny přímo na soukromí člověka, ale na jiné každodenní činnosti nebo situace. Dle [5, str. 2, vlastní překlad] je soukromí velice těžko definovatelné, právě pro jeho charakteristické vlastnosti.

„Jednou z těchto vlastností je fakt, že mnohé oblasti chráněné právem na soukromí jsou také chráněné ostatními zákony. Takto se například ústavní právo na ochranu lidské důstojnosti v izraelském právním systému vztahuje i na narušení soukromí, které narušuje důstojnost jedince; trestní právo zajišťuje ochranu před tělesnou újmou nebo znásilněním; majetkové právo chrání fyzické vlastnictví, civilní právo chrání před narušováním soukromého pozemku, napadením a narušení soukromí, které způsobuje obtíže; (...) “ [5, str. 2, vlastní překlad]

Právě proto, že bylo právo na soukromí implementováno v právním řádu, jen ne jako samostatný zákon, se mnozí lidé domnívali, že není potřeba vytvářet takový zákon [5, str. 3]. V oblasti internetu a nakládání s daty tohle ve velké míře neplatilo, protože jejich výměna probíhá často na mezinárodní úrovni. I

proto bylo nutné zákony na ochranu soukromí a zejména osobních dat zavést a co nejvíce sjednotit.

Definic práva na soukromí je možno nalézt více, protože jde o složité téma, které nemá jednotnou definici, mimo jiné z důvodů uvedených výše. Definici, která dle vlastních slov autorů *„zahrnuje elementy jak přístupu tak kontroly“* [5, str. 7, vlastní překlad], definuje právo na soukromí takto:

„Právo na soukromí je naše právo udržovat osobní oblast kolem nás, která zahrnuje všechny věci, které jsou naší součástí, jako je naše tělo, domov, myšlenky, pocity, tajemství a identita. Práva na soukromí nám umožňuje vybrat vybrat si, které části naší osobní oblasti mohou být přístupné ostatním a kontrolovat rozsah, způsob a čas použití těch částí, které jsme se rozhodli zveřejnit.“

Dle této definice je soukromí souhrn všeho, co nás definuje jako jedince, zejména tedy to, čím se odlišujeme od ostatních.

S příchodem nových technologií a zejména rozšíření možnosti připojení k internetu po celém světě, bylo třeba přijít s novým pojetím soukromí a hlavně regulacemi, které by zajistili soukromí i v nepřehledném světě internetu. *„Zatímco technologický vývoj kompletně změnil pravidla hry, právní systém zůstal pozadu.“* [5, str. 8, vlastní překlad] Právní předpisy se v Evropské Unii změnily až se schválením Obecného nařízení o ochraně osobních údajů, známého pod zkratkou GDPR. Tématu GDPR a jeho dopady na nakládání s osobními daty se budu věnovat v sekci Obecné nařízení o ochraně osobních údajů.

Jak jsem psal výše, právo na soukromí je implementováno v právních řádech většiny zemí.

„V dnešní době je soukromí považováno za základní právo a je podporováno mezinárodními dohodami a mnoha ústavními zákony. Například všeobecná deklarace lidských práv (1948) věnuje Článek 12 soukromí. Soukromí získalo mezinárodní uznání a je aplikováno na celou řadu situací, jako je například: zabránění vměšování se do domácností, omezení používání sledovacích prostředků, kontrolovat sběr a distribuci osobních dat, atp.“ [3, vlastní překlad]

Je to právě sběr a distribuce osobních dat, které vedly k rozšíření práva na soukromí i na prostředí internetu, nebo alespoň k jeho sjednocení v podobě GDPR.

Ačkoliv je soukromí po právní stránce těžko definovatelný pojem, který zahrnuje hodně oblastí lidské činnosti, mění se s časem a s novými technologiemi, je neodmyslitelně spjatý s dnešní civilizací. Nebýt vývoje v této oblasti a tlaku na příslušné orgány, nevznikl by požadavek, zejména na firmy, které sbírají velké množství osobních dat, aby tato data anonymizovali. Každý člověk má právo na soukromí a může se bránit, pokud se jeho osobní údaje vyskytnou ve veřejné či soukromé databázi, nebo budou použity jako předmět obchodu [6]. Proto je potřeba při anonymizaci dat postupovat vždy s přihlédnutím k sou-

časným zákonům, které definují rámec možností, co jde s danými daty dělat a co už je protiprávní.

2.2.2 Anonymizace a bezpečnost

Jeden z důvodů, proč anonymizovat data, je i zvýšená bezpečnost anonymních dat. Data, která jsou konkrétní a mohou identifikovat jedince, jsou při případném úniku mnohem větší problém, než data, která jsou anonymní. V závažných případech mohou být bezpečnostní problém jak pro společnost, které data unikla, tak pro jedince nebo subjekty, kterých se data týkají.

Bezprostřední bezpečnostní problém při úniku dat je mezera v zabezpečení firemní databáze. Tato zranitelnost, zneužitá útočníkem, může být více či méně závažná a je nutné ji opravit. Samotná oprava se může dát vyřešit poměrně rychle, pokud už je zřejmé, v čem zranitelnost spočívá, vznikají tím ale i další, závažnější problémy.

2.2.2.1 Postihy za únik dat

Jeden z nich je porušení Obecného nařízení o ochraně osobních údajů. V červenci roku 2019 byla vyměřena společnosti British Airways pokuta 204,6 milionů eur za únik informací v podobě osobních dat a platebních informací téměř půl milionu klientů [7]. Rok předtím dostala společnost Marriot International pokutu 110,3 milionů eur za únik záznamů 339 milionů hostů [7].

Uniklá data mají nejméně jeden společný jmenovatel. Ani jedna nebyla anonymizovaná a proto se na ně nařízení vztahovalo. Jsou tím ohroženy konkrétní osoby. Kdyby anonymizovaná byla, nebyl by to pro firmu takový problém, protože na anonymizovaná data se nařízení nevztahuje [8, str. 5].

Pokuta za nedodržená nařízení o ochraně osobních údajů nemusí být jediný výdaj, kterému bude firma po úniku dat čelit. Další pokutu mohou vyměřit úřady, pokud je únik dat ošetřený v jejich právním řádu a to zejména pokud existuje možnost, že uniklá data poškodí konkrétní osoby, které data společnosti svěřily. Například společnost Uber musela v Německu a Velké Británii zaplatit v přepočtu přes 25 milionů korun za únik dat jejich klientů a to včetně řidičských průkazů [9]. Je zřejmé, že útoky na firemní data probíhají neustále.

2.2.2.2 Krádež dat

Krádež dat je dle jihoafrického serveru cybercrime.org.za [10, vlastní překlad]: *„akt krádeže počítačových informací od nic netušící oběti za účelem narušení soukromí nebo získání důvěrných informací“*. K tomu dodává, že *„krádež dat je čím dál tím větší problém jak pro individuální uživatele počítačů, tak pro velké společnosti“*.

Hypotetické dokonalé zabezpečení dat proti útočníkům nemusí nutně znamenat eliminaci všech možných rizik s daty spojenými. V práci s daty je podstatný taktéž lidský faktor. Útok na data může provést zaměstnanec se znalostí

interních předpisů a bezpečnostních opatření, stejně jako zaměstnanec firmy, která byla najata na provedení prací, tzv. "outsourcing", to znamená, že firma nedělá všechny části projektu sama, ale na některé si najme specializované pracovníky. Dle [1, vlastní překlad]

„Stoupající trend outsourcingu u vyvíjení a testování softwarových aplikací do zahraničních lokací také zvýšil riziko zneužití citlivých dat a mimo jiné zapříčinil sadu regulací jako například PIPEDA (navržená Kanadskou vládou).“

PIPEDA je zákon o ochraně osobních informací a elektronických dokumentech platný v Kanadě. Zákon ukládá kanadským firmám žádat o povolení, pokud sbírají, používají nebo zveřejňují osobní informace. Stejně tak klade nároky na zabezpečení dat. Dále dává lidem právo žádat o přístup nebo změnu údajů, které firmám svěřily [11]. Jde o zákon podobný evropskému nařízení o ochraně osobních údajů a má za úkol řešit právě bezpečnost citlivých dat.

V těchto případech by anonymizace dat pomohla firmám v obraně před zveřejněním citlivých dat. V oblastech, kde je anonymizace možná, by měla být prováděna, aby nedocházelo k únikům citlivých údajů, nebo jejich krádežím interními či externími zaměstnanci.

2.2.2.3 Další rizika

Další z následků úniku neanonymních dat může být ztráta důvěryhodnosti dané společnosti. Pokud společnosti uniknou data klientů, ze kterých je možné zjistit konkrétní informace, zneužitelné třetí stranou, mohou klienti ztratit důvěru ve společnost a v to, jak nakládá s daty jí svěřenými. Pokud je to společnost, která se na práci s daty specializuje, jako například společnost, která poskytuje bezpečné datové úložiště, může ztráta důvěry znamenat zánik společnosti. To i přesto, že uniklá data by nebyla data uložená klienty do úložiště, ale metadata s nimi spojená, například jména klientů.

2.2.3 Přístupy k anonymizaci

K anonymizaci dat existují různé přístupy. V této části proberu nejčastější z nich. U každého přístupu popíšu jeho princip a zhodnotím jeho použitelnost pro implementaci v anonymizéru dat.

2.2.3.1 Generalizace

Odstranění osobních dat a ponechání hodnot, které jsou obecné. Generalizace je často používaný způsob anonymizace, protože je velice jednoduchý. Citlivá data jsou vypuštěna a po úpravě se nelze žádným způsobem vrátit zpět k původním datům. Právě jednosměrnost procesu je podmínkou pro to, aby byla data považována za anonymní [12].

Pro účely anonymizéru je tento druh anonymizace užitečný právě pro rychlost zpracování údajů. Operace s údaji, například hešování nebo nahrazování, zpomaluje výsledný program a nemusí být tak bezpečné. Při použití nevhodné hešovací funkce existuje riziko, že útočník dokáže proces hešování zvrátit a získat původní data.

2.2.3.2 Maskování dat

Dle [13] je maskování dat vytvoření záznamů se stejnou strukturou, které ovšem neodpovídají realitě. Data tím pádem zůstávají v bezpečí, ale lze s nimi i nadále provádět různé operace. Mezi tyto operace může patřit například testování, kdy je potřeba, aby nechyběla žádná data, ale už není tak podstatné, jestli data opravdu něco znamenají.

Pro můj program není důležité, aby všechny datové položky byly obsazené, pouze je potřeba zachovat formát dat, tzn. zachovat oddělovače, aby před i po anonymizaci odpovídaly sloupce. Proto by bylo používání maskování zbytečným zdržením, kdy by program u každé datové položky určené k anonymizaci nahrazoval data náhodnými, nebo předem určenými, řetězci.

2.2.3.3 Prohození dat

Prohození údajů u různých záznamů, údaje pak obsahují skutečná data, ale neodpovídají žádné konkrétní osobě. Tato operace je užitečná v případě, kdy je potřeba zachovat statistické vlastnosti dat, které by po smazání tuto vlastnost ztratily.

Pokud bych pro teoretickou úvahu měl seznam jmen a příjmení, mohl bych prohozením údajů ve sloupci jmen a údajů ve sloupci příjmení docílit toho, že výsledná jména nebudou odpovídat skutečným osobám. Četnost jmen a příjmení však zůstane stejná, protože jejich hodnoty se nezměnily.

Pokud je prohození skutečně náhodné, nelze z takových dat získat zpět původní data. Avšak útočníkovi, který by data získal, to dá informaci o původních jménech, i když konkrétní jména není schopen získat.

V případě dat dodaných vedoucím práce není tato technika nutná, citlivá data nejsou potřeba pro pozdější statistické zpracování. Stejně jako u maskování dat by prohazování dat zbytečně zpomalilo program.

2.2.3.4 Pseudonymizace

Nahrazení pravých údajů za falešné, které se generují z nějakého předem daného rozmezí. Rozdíl oproti maskování dat je v tom, že data, na které se aplikuje pseudonymizace, se dají vrátit do původního stavu. K datům existuje klíč, který, když se aplikuje na psedonymizovaná data, vytvoří data původní. [14]

Protože jedním z nároků na anonymizaci je nezvratnost procesu [12], nelze pseudonymizaci použít.

2.2.3.5 Hešování dat

Aplikace hešovací funkce na data. Tato technika není příliš vhodná, pokud jsou hodnoty z omezeného rozsahu, hrozí nebezpečí prolomení pomocí brute force útoku. Dobrá hešovací funkce je považována za jednosměrnou, ale při použití špatné hešovací funkce je možné, že se útočníkovi povede proces hešování zvrátit.

Tento postup není ze zákona považován za anonymizaci [12] a s takto upravenými daty nelze nakládat tak, jako je to možné s anonymizovanými daty.

2.2.4 Shrnutí

Pro účely anonymizéru mi přijde nejvhodnější přístup generalizace nebo maskování dat. Pomocí generalizace by se dala data znovu skládat bez citlivých údajů. Nevýhodou je odlišný formát vstupu a výstupu. Data na výstupu mají jiný formát než data na vstupu, což může být problém pro další zpracování. Na druhou stranu není žádný způsob, jak odvodit utajená data, ani určit, který typ dat je utajený.

U maskování dat je to opačně. Výhoda spočívá v tom, že data na vstupu mají stejný formát jako na výstupu, ale dá se odvozovat jaká data jsou utajená. Otázkou je, jak důležitá je informace, že se utajuje příjmení osoby, což by se dalo odvodit i bez přístupu k datům. Nahrazování znaku za znak dává informaci o délce utajených dat, takže by se muselo při nahrazování vždy slovo smazat a nahradit konstantním počtem substitučních znaků.

Z těchto důvodů jsem se pro svůj program rozhodl použít kombinaci obou metod, kdy budou citlivá data vynechávána, ale zůstanou zachovány oddělovače, aby zůstal zachován formát dat.

2.3 Hledání vzorů

2.3.1 Naivní algoritmus

Nejjednodušší algoritmus pro vyhledávání v textu. Zkusí, zda se hledaný výraz shoduje s textem od prvního znaku. Pokud ne, posune se o jeden znak doprava a zkusí to znovu, takto dokud nenalezne shodu, nebo nedojde na konec textu. Časovou složitost lze pak odvodit pro délku hledaného slova m a délku textu n jako O((n-m)*m) [15].

Algoritmus je velice neefektivní, *„(...) označuje (se) jako brute-force, tedy algoritmus s použitím hrubé síly“* [15]. Pro účely anonymizéru by v žádném případě nestačil.

2.3.2 Regex

Velice známým nástrojem pro hledání vzorců v textu je knihovna Regex v C++. Horní odhad časové složitosti se liší dle implementace. Jedna z možností je pomocí deterministického konečného automatu, která najde vzor v lineárním čase, ale stavba automatu má exponenciální složitost, časovou i prostorovou [16].

Další možností je používání nedeterministického konečného automatu, ale tím se zvedne doba běhu na na m*n, kde m je velikost vzoru a n velikost dat. Poslední možností je backtracking, který ale v některých případech vede také na exponenciální složitost. Vzhledem k tomu, že se nelze spolehnout, že Regex nespadne do exponenciální složitosti, je pro potřeby zpracování v reálném čase nevhodný [16].

2.3.3 Boyer-Moore algoritmus

Rychlý způsob hledání vzorců v textu. Používá předzpracování textu, při kterém si vytvoří dvě pole možných posunů na základě dvou různých heuristik. Tato operace zabere čas m*n v nejhorším případě, tj. pokud jsou všechny znaky stejné. Potom zpracovává text zprava doleva a posunuje se vždy o větší ze dvou připravených hodnot. Díky tomu, že vynechává kusy textu, může složitost být nižší než lineární. Tento přístup je dobrý pro dlouhé vzory, kdy algoritmus dokáže dělat velké skoky v textu.

Co se týče složitosti, je dle [15]:

„(...) časová složitost je v nejlepším případě, když se nejpravější znak nevyskytne ve vzorku nikdy, rovna O(n/m), v nejhorším případě je stejná, jako u naivního algoritmu.“

To znamená, že v nejhorším případě by měl složitost O((n-m)*m) [15]. To pro potřeby anonymizéru proudu dat nestačí, protože rychlost je lineární nebo horší.

Nejhorší případ pro tento algoritmus by byl takový, kdyby se celé hledané slovo opakovalo pořád dokola. Potom by program měl na každém znaku shodu a posunul by se právě o jeden znak doleva. Takto by prošel celý řetězec po jednom znaku a proto by jeho složitost byla stejná, jako u naivního algoritmu, který vždy prochází celý řetězec po jednom znaku [15].

Ačkoliv by v nejhorším případě algoritmus nestačil, je potřeba ho brát v úvahu v mé další práci. Nejhorší případ je velice nepravděpodobný a jeden jeho výskyt by nezpomalil běh celého programu natolik, aby to bylo poznat na čase zpracování.

Pro úvahu si vezmu jednu větu jako jeden běh programu, pravděpodobnost výskytu věty složené ze samých hledaných slov jako 1:1000 a pravděpodobnost výskytu věty, kde se nejpravější znak nevyskytne ve větě nikdy,

také jako 1:1000, potom by tyto výskyty v celkové složitosti nehrály žádnou roli a výsledná složitost by byla lepší než lineární. Výskyt takových vět bude s velkou pravděpodobností mnohem menší, jak vyplývá z podoby dat dodaných vedoucím práce, které budu rozebírat v samostatné kapitole. Pro nižší pravděpodobnost výskytu vět, které způsobí extrémní chování programu, na jednu nebo na druhou stranu, se časová složitost algoritmu nezvýší.

Nicméně i s lineární složitostí je tento algoritmus vhodný spíše pro dlouhé vzory, které se v datech vyskytují minimálně, jak vyplývá z ukázky dat. Velkou nevýhodou je potřeba pro každý nový vzor připravovat nový seznam možných skoků na základě používaných heuristik. Dalším omezením algoritmu je schopnost hledat pouze jeden vzor v jednom běhu programu. Pro jiný vzor je potřeba spustit nový běh, včetně předzpracování.

2.3.4 Knuth-Morris-Pratt algoritmus

Algoritmus, který přeskakuje podřetězce, které se opakují a dosahuje tak lineární složitosti v nejhorším případě [17]. Je vhodný pro malé abecedy, kde je větší pravděpodobnost, že se budou podřetězce opakovat. Dobře by fungoval například pro hledání číselných vzorů, hůře pak pro hledání jmen a adres.

V číselných vzorech se často opakují podřetězce, vyhledávání tam je rychlejší pomocí KMP algoritmu. Obecně se dá říci, že lépe pracuje nad malými abecedami, bez ohledu na znaky. Z toho vyplývá, že je tento algoritmus užitečný pro hledání ve dvojkové soustavě, kde abeceda sestává ze dvou znaků: 0 a 1.

Pro hledání ve větších abecedách bude algoritmus pomalejší. Pravděpodobnost opakování určitého podřetězce klesá s přibývajícími znaky abecedy. Každý jeden další znak v abecedě znamená jednu další možnost při volbě dalšího znaku do řetězce. Pokud by byly řetězce náhodné, byla by pravděpodobnost, že další znak bude stejný, jako ten předchozí, pro velikost abecedy 2 znaky právě 1:2. Pro velikost abecedy 26 znaků by už taková pravděpodobnost byla 1:26. Pro opakování podřetězce o délce 2 by byla pravděpodobnost pro abecedu o velikosti 2 znaky 1:4, pro abecedu o délce 26 znaků 1:676.

Při posuzování algoritmu musím vzít v úvahu i to, že abeceda, ze které se skládají osobní data fyzických osob, je většinou větší než 26 znaků a tudíž pravděpodobnost výskytu podřetězců je menší. Na druhou stranu je třeba vzít v úvahu, že data nejsou tvořena náhodně, jsou to skutečné záznamy.

Z výše uvedeného vyplývá, že tento algoritmus není vhodný pro použití v anonymizéru. Velikost abecedy vstupních dat nevyužívá hlavní přednost tohoto algoritmu, kterou je rychlé hledání v textu s abecedou s malým počtem znaků. Zároveň nelze vyhledávat více vzorů najednou, stejně jako u Boyer-Moore algoritmu.

2.3.5 Aho-Corasick algoritmus

Algoritmus založený na stavbě konečného automatu, ve kterém pak dokáže nejhůře lineárně hledat vzory [18, str. 5]. Je to slovníkově založený algoritmus, tj. automat se staví na určitý slovník. Slovník je seznam všech vzorů, které má algoritmus v textu vyhledat. *„Je elegantním zobecněním Knuthova-Morrisova-Prattova algoritmu pro více řetězců.“* [18, str. 5]

Algoritmus si nejdříve vytvoří konečný automat na základě zadaného slovníku. Tento automat obsahuje ve svých stavech znaky řetězců, přechody jsou mezi stavy tak, jak jdou znaky v hledaném vzoru za sebou. Pokud vzor daným znakem končí, je stav označený tímto znakem koncový. Druhý typ přechodů jsou zpětné vazby, které se využijí v případě neshody dalšího znaku vzoru.

Stavba automatu i následný běh je v nejhorším případě lineární [18, str. 6]. Pro vyhledávání vzorů v datech je největší výhodou tohoto algoritmu možnost vyhledávat více vzorů najednou. Při anonymizaci proudu dat je potřeba vyhledávat různé druhy dat během jednoho běhu.

Jako slovník lze použít i regulární výrazy, které se expandují do všech možných konečných tvarů před začátkem algoritmu. Na jejich základě se pak sestaví automat, podle nějž se vyhledává.

2.3.6 Rabin-Karp algoritmus

Posunuje se v textu po jednom znaku, ale místo řetězců porovnává jejich heše. Používá se pro odhalování plagiátů.

*„Přípravná fáze algoritmu vyžaduje O(m) kroků, vlastní vyhledávání má v nejhorším případě časovou složitost O(m*n), v průměru však jen O(n).“* [15]

U Rabin-Karp algoritmu je důležitá volba hešovací funkce:

„Pokud je hešovací funkce ‘kvalitní’, málokdy se stane, že by se heše rovnaly, takže místo času Θ(J) na porovnáváním řetězců si vystačíme s porovnáním hešů v konstantním čase.“ [18, str. 9]

Stejně tak má volba hešovací funkce vliv na časovou složitost i při posunování algoritmu po znacích. Dle [18, str. 9]: *„Pořídíme si hešovací funkci, kterou lze při posunutí okénka o jednu pozici doprava v konstantním čase přepočítat.“* Potom při správné volbě hešovací funkce *„(...) proběhne aktualizace heše v konstantním čase“* [18, str. 9].

Nevýhoda algoritmu oproti výše zmíněnému Aho-Corasick algoritmu je zejména v neschopnosti vyhledávat více vzorů najednou.

2.3.7 Optimální řešení

Optimální řešení je podle mého názoru implementace Aho-Corasick algoritmu s předem připraveným konečným automatem založeným na slovníku hodnot, které je potřeba nahradit. Časová složitost takového řešení je nejmenší z porovnávaných algoritmů. Slovníků může být víc a dají se upravovat, stačí pak přegenerovat konečný automat. Konstrukce automatu se může provádět i za běhu programu, ovšem za cenu pomalejšího zpracování. Vzhledem k tomu, že se slovník dá vyplnit jako konečná expanze regulárních výrazů, má téměř takovou obecnost jako vyhledávání knihovnou regex a pro potřeby anonymizéru, který by hledal data na základě vzorů, by měl být dostačující.

Osobně vidím jako největší výhodu možnost sestavení automatu předem pro všechny hledané hodnoty, kdy vlastní hledání je pak rychlé i pro velká data. Dalo by se uvažovat o doplnění KMP algoritmem pro hledání číselných dat s omezeným rozsahem, např. data narození.

2.4 Obecnost anonymizace

V této části rozeberu pojem obecnost a navrhnu princip obecnosti anonymizéru, který budu implementovat. Tato část souvisí s modularitou anonymizéru, protože právě modularita bude zajišťovat obecnost a navržené řešení obecnosti je zároveň řešením modularity.

Jedna z definic obecnosti u právní normy zní: *„vztahuje se na neurčitý počet případů k neurčitému počtu subjektů“* [19]. Tato definice může být použita i na zpracování dat anonymizérem.

Anonymizér je obecný, pokud dokáže zpracovat neurčitý počet druhů dat a není omezen jen na některé druhy. Vzhledem k tomu, že pro každý druh dat bude existovat konkrétní modul, bude obecnost anonymizéru zajištěna.

2.4.1 Implementace obecnosti

Pro obecné použití musí centrální logika programu pracovat nezávisle na typu dat, který zpracovává. Proto bude konkrétní data nejprve zpracovávat modul, který převede data do podoby, ve které je centrální logika dokáže zpracovat. Centrální jednotka programu, která bude mazat anonymní data, tak bude pracovat pro všechny typy dat stejně. Převezme si proud dat, aby z nich odstranila předem určené sloupce. Úpravu konkrétních dat do obecné formy obstarávají moduly, které se připojují k centrální jednotce.

Modul obsahuje funkce, které centrální jednotka volá po obdržení validních dat. Každý konkrétní typ dat má svůj modul, který data zpracuje a připraví do formy, kterou dokáže zpracovat centrální jednotka. Zpracování probíhá bez ukládání dat do externího souboru, jednotky programu si je posílají po částech mezi sebou.

Po zpracování centrální jednotkou, modul převezme data zpět a upraví je do původní podoby, která je opět konkrétní pro daný modul. Úpravu dat do obecné podoby a zpět tak obstarává jediný modul.

2.5 Data

V této části proberu data z pohledu velkých firem shromažďujících data a z pohledu zákona, jednotlivé druhy dat se zaměřením na citlivá data a vývoj sběru dat a jeho regulace v posledních letech. Následně provedu analýzu vzorků, které mi dodal vedoucí mé bakalářské práce. Výstupem této analýzy bude návrh, jak s danými daty nejlépe pracovat, který později použiji a dále rozvedu v sekci Data z pohledu implementace.

2.5.1 Data obecně

Data se dají v počítačové vědě definovat jako: *„(...) označení pro čísla, text, zvuk, obraz, popř. jiné smyslové vjemy reprezentované v podobě vhodné pro zpracování počítačem“* [20]. Merriam-Webstrův slovník je definuje jako *„faktickou informaci(jako například měření nebo statistiky) použitou jako základ pro vyvozování závěrů, diskusi nebo výpočty“*, nebo jako *„informaci v digitální podobě, která může být vysílána nebo zpracována“* [21, vlastní překlad].

Z těchto definic je datům v té podobě, v jaké je bude zpracovávat můj program, blíže druhá definice. Jsou to informace, výstupy ze síťových zařízení, které jsou dále vysílány a zpracovávány.

Informace má své vlastní definice, ale pro účely mé bakalářské práce je vhodné definovat informaci jako jednotku dat, to znamená jednu položku jednoho záznamu, ze kterých jsou složeny ukázkové sady dat.

2.5.2 Druhy dat

Existuje více rozdělení dat, protože data se používají ve více oblastech lidského působení. V počítačovém světě se data dají dělit například na strukturovaná a nestrukturovaná. Další dělení dat, které souvisí s tématem mojí bakalářské práce je rozdělení dat na tzv. identifikující, pseudonymizovaná a anonymizovaná. Toto rozdělení jsem z části zmiňoval v sekci Přístupy k anonymizaci, ale v této části ho dále rozvedu. Další druh dat, který budu podrobněji rozebírat, jsou citlivá data.

2.5.2.1 Nestrukturovaná data

Nestrukturovanými daty se rozumí *„tok bytů bez dalšího rozlišení“*, jak ve své knize Data, informace, znalosti a Internet uvádí autor Vilém Sklenák [20, str. 2]. Tok bytů bez interpretace může představovat cokoliv. *„Patří sem ovšem také textové dokumenty“* [20, str. 2]. Ačkoliv, jak autor knihy dále vysvětluje,

zařazení textových dokumentů do nestrukturovaných dat se může změnit kvůli novým trendům.

Nestrukturovaný typ dat není v dodaných ukázkách použit. Existuje ještě částečně strukturovaný typ dat, kam se řadí například HTML, značkovací jazyk [22]. Ani ten není v ukázkách použit a je mimo klasické rozdělení, proto ho zmiňuji jen okrajově.

2.5.2.2 Strukturovaná data

Podstatou strukturovaných dat je struktura, kterou data dodržují. Dle [20, str. 2] strukturovaná data

„explicitně zachycují fakta, atributy, objekty apod., přičemž významným rysem je existence určitých elementů dat. (...) Díky tomuto strukturovanému uložení je potom možné snadno vybírat jen ta data, která jsou zapotřebí k řešení nějakého informačního problému, např. zjištění průměrné hodnoty určitého atributu.“

Z toho vyplývá, že strukturovaná data jsou rozdělena podle předem dané struktury na jednotlivé části.

Data, která mám k dispozici jako ukázkové, tuto definici splňují. Jsou rozdělena na jednotlivé záznamy, obsahově stejné, a každý záznam je rozdělen na sloupce, přičemž v každém sloupci se stejným pořadovým číslem v záznamu jsou stejný druhy hodnot, například IP adresa, číslo síťového zařízení nebo způsob ukončení spojení.

Strukturovaná data se využívají zejména v relačních databázových systémech, kde *„se používá hierarchie elementů od pole k záznamu, relaci až k databázi“* [22]. Databáze se hojně využívají pro ukládání dat zejména ve velkých firmách, které zpracovávají velké množství dat.

Využívání databází má své opodstatnění. Vzhledem k tomu, že se data ukládají ve strukturované formě, tak se v datech *„lépe vyhledává a také se s nimi dále snáze pracuje“* [22]. To je případ i dodané ukázky dat, anonymizace a práce s těmito daty všeobecně je jednodušší, než kdyby se jednotlivé hodnoty nacházely na různých místech v souboru a bez pravidelné struktury.

2.5.2.3 Citlivá data

Citlivá data, nebo také citlivé údaje, jsou druh dat, který náleží do soukromé sféry každého jedince. Vztahuje se na ně právo na soukromí, tak jak je popsáno v sekci Soukromí. V obecných datech jsou to právě citlivé údaje, které je třeba chránit, ať už anonymizací nebo jinými způsoby.

Dle [23] jsou citlivé osobní údaje

„speciální kategorií podle GDPR, která zahrnuje údaje o rasovém či etnickém

původu, politických názorech, náboženském nebo filozofickém vyznání, členství v odborech, o zdravotním stavu, sexuální orientaci a trestních deliktech či pravomocné odsouzení osob. Tyto údaje mohou subjekt údajů samy o sobě poškodit ve společnosti, v zaměstnání, ve škole či mohou zapříčinit jeho diskriminaci. Do kategorie citlivých údajů GDPR nově zahrnuje genetické a biometrické údaje.“

Tato definice není pro účely mojí bakalářské práce úplně přesná, protože ukázková data nejsou informace této povahy, jak budu rozebírat v sekci Analýza vzorků dodaných vedoucím práce. Ale dle mého názoru je to důležitá definice pro nakládání s daty, které se řídí evropským nařízení o ochraně osobních údajů.

Jiná definice citlivých dat pracuje s termínem *„personally identifiable information“*, což v překladu znamená informace, podle které se dá identifikovat jedinec. Nebo také dle [24, str. 3, vlastní překlad]:

„informace, u které lze rozumně předpokládat, že dokáže identifikovat jedince, a která, pokud zveřejněna, může narušit soukromí jedinců a vést ke krádeži identity nebo podvodu.“

Právě informace, která dokáže identifikovat jedince, je citlivý údaj v případě ukázkových dat, jak jsem zmiňoval v sekci Anonymizace dat.

Za citlivé údaje jsou považovány i data, která jsou tak označena interními směrnicemi, takže se může jednat o kterákoliv firemní data [24]. Jako s takovými by s nimi mělo být nakládáno s mnohem větším důrazem na jejich bezpečnost, například ukládat taková data na bezpečné úložiště, nebo, pokud je to možné, anonymizovat.

Vzhledem k povaze dat, která neobsahují žádné přímé identifikátory, jako jsou například jméno nebo příjmení, definuje citlivá data rozhodnutí vlastníka těchto dat. Taková data pak budou smazána napříč všemi záznamy.

Kromě základního dělení dat se dle [24, str. 5] dají citlivá data dělit na různé kategorie dle stupně jejich ochrany. Tři kategorie jsou tajné, interní a veřejné. Při rozhodování, do jaké kategorie data zařadit, je třeba brát v potaz zejména následující faktory.

Prvním faktorem je dle [24, str. 5] dopad na firmy, pokud by data měla uniknout, z hlediska peněz, reputace nebo kontraktů s firemními partnery. To znamená, že firma, které by unikla data, by tím utrpěla finanční ztrátu, protože by data například obsahovala informace o konkurenční výhodě, kterou firma připravovala. Dále by utrpěla její reputace, kterýžto problém jsem již nastínil v sekci Anonymizace a bezpečnost. Posledním hlediskem je ztráta kontraktů s firemními partnery. Tento problém z části souvisí se ztrátou reputace, protože s firmou, která není důvěryhodná, spíše nebudou chtít spolupracovat nový partneři.

Druhým faktorem jsou právní důsledky neoprávněného zveřejnění nebo šíření dat. Ty mohou zahrnovat policejní vyšetřování, právní postihy, včetně

pokut, případně i zákazu činnosti. Právní postihy budou detailně popsány v sekci Osobní data z pohledu zákona.

Třetím faktorem, který je důležitý při rozhodování se, do které kategorie data patří, je *„důvěrnost a přesnost dat s ohledem na obchodní funkce a potřeby“* [24, str. 5]. To znamená určit, jakým způsobem data ovlivňují obchodní činnost firmy a co by jejich zveřejnění pro tuto činnost znamenalo.

Po zhodnocení všech faktorů by se vedení firmy mělo rozhodnout, do které kategorie data, o kterých se rozhoduje, patří. V případě, že by se jednalo o data citlivá, která by proto patřila do kategorie tajné, případně interní, je vhodné podniknout kroky k tomu, aby data neunikla na veřejnost.

2.5.2.4 Big data

Jako big data, neboli velká data, se souhrnně označuje velké množství informací, které velké společnosti sbírají a dále zpracovávají. Jsou to právě tato data, u nichž je největší riziko úniku, protože databáze velkých firem představují zajímavý cíl pro útočníky, kteří chtějí při jednom útoku získat co největší množství dat.

Jak uvádí [25, str. 6, vlastní překlad]:

„Neexistuje žádná rigorózní definice big data. Původní myšlenka byla taková, že množství informací se zvětšilo natolik, že se při zpracování nevešlo do paměti počítače, takže technici museli vylepšit nástroje, kterými je zpracovávali.“

To znamená, že dat je příliš velké množství, než aby se daly zpracovávat v rozumné době.

Big data se dají také definovat pomocí tří V. Jsou to volume, velocity a variety. Další termíny, které jsou s nimi spojované, jsou veracity, value, validity a volatility.

Pojem volume, neboli obsah, znamená, že data mají příliš velký objem. Zároveň přibývají příliš velkou rychlostí, velocity. A také obsahují více druhů dat, jak strukturovaná, tak nestrukturovaná, variety. Zbylé pojmy udávají další vlastnosti spojované s big data: nemusí být věrohodná, jsou velmi drahá, mají omezenou dobu platnosti a nutnost najít úložiště pro velké množství dat [26, str. 20]. Kombinace těchto vlastností neumožňuje zpracovávat data tak, jako tomu bylo s menším množstvím dat.

Jak jsem zmiňoval výše, big data potřebují velké množství úložného prostoru, kde se data uloží než jsou zpracována. Potřebná velikost se může měnit a jednorázově dosahovat vysokých hodnot, zatímco v některých chvílích nemusí být využita téměř vůbec. Proto je vhodné řešení ukládat data tzv. v cloudu, tj. v pronajatém úložišti mimo pozemek firmy [27].

Cloudové služby poskytuje množství firem, které na svých serverech zdarma nebo za poplatek ukládají informace svých zákazníků. Toto řešení je vhodné pro big data díky své škálovatelnosti. Cloudové úložiště obecně zvládne po-

jmout velké množství dat, které se denně vytváří. Například dle [26, str. 20] uvedla společnost IBM odhad pro rok 2020, ve kterém tvrdí, že:

„6 miliard lidí na zemi bude vlastnit mobilní telefon, denně vznikne 2,8 kvintilionů (2,8 x 10 na 18) bajtů dat, a celkově bude na discích uloženo 40 zettabajtů dat (kde 1 zettabajt je 10 na 21 bajtů, tedy ekvivalent miliardy pevných disků o velikosti 1 TB).“

Dle [28] je dnes počet lidí s mobilním zařízením 5,28 miliardy a mobilních připojení je 9,82 miliardy. Množství dat je velice vysoké a není důvod předpokládat, že by mělo do budoucna klesat, spíše naopak. Proto jsou cloudová úložiště pro firmy i koncové uživatele dobrým řešením, jak zvládat velké objemy vytvářených dat.

Zároveň ale mohou představovat bezpečnostní riziko, protože data opouštějí pozemek firmy, která je vlastní, a jsou uložena na úložišti jiné firmy. Pro neanonymizovaná data je v tomhle směru limitující Obecné nařízení o ochraně osobních údajů. Dle tohoto nařízení lze skladovat data, která byla sesbírána na území Evropské unie, jinde než na území Evropské unie pouze za předpokladu, že budou dodržena všechna nařízení, tak jako by data byla uložena na území Evropské unie, což nemusí být snadné zajistit [8, str. 19]. To může představovat problém, pokud firmy, která poskytuje datové úložiště, má některé ze svých úložišť i v jiné zemi, například ve Spojených státech amerických.

Při používání těchto úložišť tak musí být zřejmé, kde se skutečně data nacházejí. Bezpečnostní riziko však může představovat i zpracování na straně firmy, poskytující úložiště, o jejíž infrastruktuře obecně nemusí být nic známo.

2.5.2.5 Identifikující data

Rozdělení dat, které nepatří do klasického dělení, ale které je podstatné pro moji bakalářskou práci, je rozdělení dat na identifikující, pseudonymizovaná a anonymizovaná. Toto rozdělení pomáhá určit, jaký přístup je k datům vhodný, vzhledem k jejich zabezpečení, ať už z hlediska nařízení firmy, tak z hlediska zákona.

Identifikující data jsou taková, která v sobě obsahují informace, které mohou přímo identifikovat jedince. Tyto informace mohou zahrnovat jeho jméno, příjmení, nebo všeobecně uznávaný identifikátor, jako je číslo občanského průkazu nebo pasu [2, str. 8].

Definice identifikujících dat se z velké části překrývá s definicí citlivých dat, tak jak jsem ji použil v této bakalářské práci. Vztah mezi těmito termíny je takový, že všechna identifikující data jsou zároveň citlivá data, ale nikoliv obráceně. Citlivá data mohou zahrnovat i informace, které nespadají do identifikujících dat, jako je třeba náboženské vyznání jedince, které je uvedeno v definici citlivých dat dle [23]. Proto je potřeba oddělovat citlivá data a identifikující data, i když se z části překrývají.

Podle [2, str. 8] je nutné rozlišovat dva podobné pojmy *„identifiable“* a *„identified“*, které by se daly přeložit jako identifikovatelný a identifikovaný.

„Jedinec je v sadě dat identifikovatelný, pokud je rozumné se domnívat, že v nich jedinec může být identifikován buď za pomoci dat obsažených v sadě, nebo s pomocí externích dat veřejně dostupných nebo známých útočníkovi.“

Pojem identifikovaný se pak vztahuje na data, ve kterých se identita jedince přímo vyskytuje. Je v nich například přímo uvedeno jméno jedince. Taková data se mohou vyskytovat například na návštěvních knihách hotelů nebo čekáren lékaře.

Ačkoliv je mezi oběma pojmy malý rozdíl po fonetické stránce, rozdíl po stránce významové je zásadní. Pod data, které jsou identifikovatelné, může spadat velké množství dat a nelze vždy přesně určit, která to jsou, vzhledem k tomu, že identifikovatelnost jedince záleží nejen na datech, ale i na informacích ve vlastnictví útočníka, o kterých zpravidla nemusí být nic známo.

Naopak identifikující data jsou vždy jasně rozpoznatelná a jejich získáním by útočník okamžitě získal identitu osob, které se v datech vyskytují. Proto by taková data měla být dobře chráněna a pokud možno uchovávána jen po dobu, po kterou je to skutečně nutné [2, str. 8].

Identifikující i identifikovatelná data by pro účely uchovávání nebo dalšího zpracování měla být vždy anonymizována. V případě, že neanonymizovaná data uniknou, vznikají pro firmu mnohem větší sankce, než v případě dat anonymizovaných, jak jsem rozebral v sekci Anonymizace a bezpečnost.

2.5.2.6 Pseudonymizovaná data

Dalším druhem dat, který rozeberu v této části jsou pseudonymizovaná data. Dle [2, str. 8] byl tento termín popularizován až s příchodem GDPR. Dále popisuje pseudonymizaci jako *„odstranění přímých identifikátorů“*, ať už se jedná o jejich nahrazení, zašifrování, nebo smazání a *„informace potřebná k jejich opětovné identifikaci je uložena zvlášť a je předmětem (…) kontroly“*.

Z toho vyplývá, že u pseudonymizovaných dat existuje možnost získat zpět původní data, pokud by se útočníkovi podařilo získat jak data tak klíč, který k nim náleží. Zisk samotného klíče nebo samotných dat by útočníkovi neumožnil si pseudonymizovaná data přečíst.

Z této definice [2, str. 8] dále vyvozuje fakt, že pseudonymizovaná data jsou data identifikovatelná, a to zejména pokud byla předtím identifikující. V takovém případě útočník se znalostí klíče dokáže po získání dat identifikovat osoby, které jsou v nich uvedené.

Z těchto důvodů jsou pseudonymizovaná data méně bezpečná než data anonymizovaná. Zároveň se na ně vztahují odlišné právní nároky, které budu rozebírat v sekci Osobní data z pohledu zákona. Uplatnění pro tento typ dat může být skladování identifikujících dat v případě, že bude v budoucnu po-

třeba opětovný přístup k nezměněným datům. V takovém případě by anonymizace dat nevyhovovala a zajištění bezpečnosti dat a klíče by bylo třeba zajistit jinými prostředky.

2.5.2.7 Anonymizovaná data

Největší stupeň bezpečnosti pro firmy poskytují anonymizovaná data. Taková data splňují následující podmínky. Nesmí obsahovat přímé ani nepřímé identifikátory [2, str. 8] a zároveň nesmí být možné z anonymizovaných dat získat data původní [12]. V druhé podmínce se liší od dat pseudonymizovaných.

Ačkoliv by získání původních dat nemělo být možné ani z dat pseudonymizovaných, čehož je docíleno skrze bezpečnostní opatření, je právě tento rozdíl mezi pseudonymizací a anonymizací po právní stránce velice významný, jak budu rozebírat v sekci Osobní data z pohledu zákona.

Při úniku tohoto typu dat může vzniknout škoda firmě, které unikla, pouze z důvodů, jako jsou konkurenční boj, nebo cena za předchozí zpracování, ale nikoliv z důvodu pokuty. Tato data neobsahují žádné osobní údaje a nemohou tak poškodit konkrétní jedince nebo právní subjekty.

2.5.3 Vývoj dat

Pojem data nevznikl až s vývojem počítačů. Data v nějaké podobě existují již od počátků lidské civilizace. První písemné záznamy se datují do doby 4000 let před naším letopočtem, kdy v Egyptě a Mezopotámii vznikly první druhy písma, hieroglyfy a klínové písmo [29]. S prvním písmem nastal přechod od prehistorie k historii.

2.5.3.1 Počátky dat

Data jsou ale mnohem starší, objevila se ještě před příchodem písma. Dle [30, str. 2, vlastní překlad]:

„Byly nalezeny zářezy na klaccích, kamenech a kostech datované do doby Mladého paleolitu. Předpokládá se, že tyto zářezy měly reprezentovat data představující číselný systém, ačkoliv je to stále předmětem akademické debaty. Asi nejznámější příkladem je Ishango Bone, nalezená v Demokratické republice Kongo v roce 1950, jejíž stáří se odhaduje na 20 000 let.“

Vzhledem k tomu, že Mladý paleolit je období, které se datuje mezi lety 40 000 a 11 000 před naším letopočtem [31], je možné datovat počátky dat do tohoto období.

Opravdu potvrzené a nezpochybnitelné použití dat je datováno právě k sumerským a egyptským civilizacím, které se objevily kolem roku 4000 před naším letopočtem [32] [33].

2.5.3.2 Využití dat

Z toho vyplývá, že data jsou v lidské civilizaci velmi dlouhou dobu a byla velmi důležitá v celém jejím vývoji. Dle [30, str. 1, vlastní překlad] užívali výhody sběru a analýzy dat lidé i před naším letopočtem. Jako příklad uvádí válku mezi Spartou a Athénami, kdy Athénský generál z vícero nepřesných pozorování použil jejich průměr a podařilo se mu tak se svými muži uniknout. *„Tato epizoda může být považována za nejvíce působivé využití sběru dat a jejich následné analýzy.“*

Data se sbírala a využívala i nadále, například v roce 1854 anglický lékař John Snow vyvrátil domněnku, že cholera je šířena vzduchem, díky nasbírání velkého množství dat během epidemie cholery v Londýně. Zároveň tím potvrdil svoji hypotézu, že se nákaza šíří skrze kontaminovanou vodu [30, str. 3].

2.5.3.3 První počítače

Největší změna dat nastala s příchodem počítačů. Povaha dat se nezměnila, ale zmnohonásobil se jejich objem a rychlost zpracování. Právě rychlost zpracování umožnila velký průlom ve vnímání dat. Bylo možné sbírat velké množství dat, které následně místo ručního vyhodnocování analyzoval počítač.

Jeden z prvních nejznámějších příkladů, kdy stroj urychlil zpracování a analýzu dat, je britský elektromechanický stroj Bombe. Sestrojili ho Alan Turing spolu s Gordonem Welchmanem v roce 1940. Nejednalo se o počítač v dnešním smyslu slova, protože byl sestrojen a fungoval pouze pro jeden účel, a to prolomit německý kód Enigma. Jeho úkolem bylo analýzou velkého množství dat, zachycených německých zpráv, prolomit denní klíč tohoto kódu [34]. Zpracováním dat rychlostí, které by člověk manuálním zpracováním nemohl dosáhnout, dokázal prolomit mnohé klíče a panuje všeobecný názor, že urychlil konec druhé světové války.

První počítač představil Konrad Zuse v roce 1941 v nacistickém Německu. Z3, jak svůj počítač nazval, pomáhal s návrhem designu německých letadel, když vypočítával jejich aerodynamické vlastnosti. Zuseho počítač byl první programovatelný automatický počítač. V roce 1943 byl zničen při bombardování Berlína [35].

Další počítače dále urychlovaly zpracování a analýzu dat. Mezi ně patří ENIAC, dokončený v roce 1946 [36], nebo UNIVAC, dokončený v roce 1951 [37].

2.5.3.4 Data v moderní době

V dnešní době jsou počítače schopné zpracovávat obrovské množství dat, jak jsem psal v sekci Big data. Významný průlom v šíření dat v moderní době přišel s počátkem World Wide Webu, který vytvořil sir Tim Berners-Lee z Velké Británie v roce 1990 ve švýcarském CERNU [38]. Tam také zprovoznil první

webový server a prohlížeč s první webovou adresou info.cern.ch, to vše běžící na počítači NeXT [39].

Díky World Wide Webu se změnil přístup k datům. S jeho rozšířením, spolu se stále vyšší dostupností počítačů, se stala data přístupnější a jejich šíření jednodušší. Z těchto důvodů narůstal objem dat, jak produkovaných, tak zpracovávaných.

Tento trend pokračuje dodnes, data stále přibývají s novými technologiemi a zařízeními, které zjednodušují přístup k datům a jejich šíření. S těmito trendy se objevují problémy, které dříve neexistovaly, a jeden z nich je právě přístup k datům a jejich šíření ve velkém množství. I proto jsou dnes data a zejména jejich potencionálně citlivý obsah regulovány zákony a nařízeními po celém světě.

Z uvedených údajů lze vyvodit závěr, že objem dat se bude zvyšovat, stejně jako jejich přístupnost. Je proto důležité chránit citlivá data, která by mohla při nesprávném zpracování uškodit firmám, kterým patří, nebo jednotlivcům, o kterých obsahují citlivé informace.

2.5.4 Analýza vzorků dodaných vedoucím práce

V této části budu analyzovat data, která jsem obdržel od vedoucího své bakalářské práce, a která budou zkušebními daty při implementaci anonymizéru a modulů. Proto bude výsledkem této části návrh implementace vstupu a úpravy dat.

2.5.4.1 Formát dat

Dodané vzorky dat jsou CSV formátu, neboli čárkou oddělené hodnoty.

„Jedná se o triviální formát ukládání dat, která mají podobu tabulky. každá řádka textového souboru obsahuje jeden záznam s několika položkami, které jsou odděleny nějaým oddělovačem - typicky středníkem nebo čárkou. První řádka souboru může obsahovat názvy sloupců(položek).“ [26, str. 43]

Tento formát je velice jednoduchý, data pouze řadí za sebe a přidává oddělovače. Problémem s tímto druhem dat může být, že *„neexistuje jeho standardizovaná podoba“* a *„používají se různé oddělovače položek, různě se řeší situace, kdy položka obsahuje jako svoji hodnotu znak oddělovače“* [26, str. 43].

Tyto problémy v mém programu řeší moduly, které jsou naprogramované na zpracování určitého druhu dat, včetně použitých oddělovačů. Toto řešení je nutné, protože *„neexistuje ani způsob, jak do souboru uložit informace o použitém kódování“* [26, str. 43].

Dodaná data jsou s ohledem na předchozí analýzu strukturovaná a interní. Nejsou identifikující, ale data, která bude anonymizér zpracovávat, mohou být identifikovatelná.

2.5.4.2 Obsah dat

Vzorky dat neobsahují žádná jména a příjmení, emailové adresy nebo jiné přímé identifikátory, které by šlo vyhledávat pomocí algoritmu pro hledání vzorů. Jde o výpisy síťových zařízení a jako u takových je potřeba nejdříve určit, které sloupce, nebo kombinace sloupců, jsou citlivé údaje.

Dodaná data proto nebudu analyzovat z pohledu hledání vzorů, ale z pohledu zpracování jednotlivých řádků a sloupců. Každý řádek výpisu má stejný počet sloupců, ačkoliv některé položky mohou chybět. Ve sloupcích jsou tak pod sebou data stejného typu. Pro zpracování citlivých údajů tak bude nejlepší zadat programu sloupce, které z něj má odstranit.

Data mají formát seznamu s pevným oddělovačem. Program podle oddělovače určí pořadí sloupce, ve kterém se daný údaj nachází a pokud se sloupec bude nacházet v seznamu sloupců určených k vymazání, odstraní ho, jinak ho ponechá. Výsledná data budou obsahovat všechny oddělovače, aby byl zachován formát dat.

Speciálním případem je seznam hodnot, které jsou brány jako samostatný údaj. Takovýto seznam je z pohledu programu samostatný sloupec a v dodaných datech je označen uvozovkami. Algoritmus odstraní ze vstupních dat celý seznam, ale ponechá jen oddělovač na konci seznamu, vnitřní oddělovače odstraní.

Různé styly dat, které se liší v oddělovačích nebo rozdělením sloupců, budou zpracovávat moduly.V datech dodaných vedoucím práce se nachází dva různé styly zápisu dat.

2.5.4.3 Ukázky

První ukázka dat k anonymizaci obsahuje vždy stejný počet položek, 58 sloupců datových položek oddělených čárkami. Seznam položek v jednom sloupci je ohraničen uvozovkami. Není zajištěno, že všechny záznamy jsou celé na jednom řádku. Jeden záznam může být přes více řádků a řádkový zlom se může vyskytovat i uvnitř seznamu hodnot. V takové situaci, kdy nastane řádkový zlom v seznamu hodnot, zůstane na novém řádku osamocený znak uvozovek. Ne všechny hodnoty záznamu jsou vyplněné, některé položky jsou prázdné, ale obsahují oddělovače, aby zůstal zachován vždy stejný počet sloupců

Druhá ukázka dat má, stejně jako první ukázka, fixní počet položek. Má 58 datových položek. Rozdíl oproti první ukázce je v tom, že ve druhé jsou všechny záznamy na přesně jedné řádce. Každý záznam je oddělen řádkovým zlomem, což dělá zpracování mnohem jednodušší.

Třetí ukázka dat je velmi podobná té první. Také obsahuje 58 položek a řádkové zlomy mohou nastat i uvnitř záznamu nebo seznamu hodnot. Liší se tím, že oproti první ukázce dat může chybět poslední položka v záznamu, což by mohlo dělat potíže při implementaci.

2.5.4.4 Shrnutí

Z mé analýzy vyplývá, že budu implementovat dva moduly pro dva druhy dat. Jeden modul pro první a třetí ukázku dat a druhý pro druhou ukázku dat.

První modul bude zpracovávat data podle sloupců takovým způsobem, aby jejich počet odpovídal počtu položek v jednom záznamu. Teprve po složení celého záznamu předá data centrální jednotce k úpravě.

Druhý modul bude zpracovávat data po řádcích, protože záznam v druhé ukázce je vždy na jednom řádku. Data pro tento modul by mohl zpracovávat i první modul, ale vzhledem k rychlosti je lepší implementovat zvlášť modul pro tyto data, protože program bude pracovat rychleji, když bude načítat celé řádky.

2.5.5 Data z pohledu implementace

V rámci mé bakalářské budu mimo jiné programovat přijímání vstupu ve formě CSV dat a následné skládání dat zpět do tohoto formátu. Tento formát dat je jeden z jednodušších formátů dat, jak jsem zmínil výše.

Proto bude můj program zpracovávat data po řádcích a oddělovat jednotlivé hodnoty pomocí oddělovačů. Tuto funkcionalitu zajistí moduly, které data předají centrální logice.

Následně bude data opět skládat a přidávat oddělovače, dokud nesloží celou řádku. Tu pak uloží do výstupního souboru. Takto budou mít data na vstupu stejný formát, jako data na výstupu.

2.6 Osobní data z pohledu zákona

V této části se věnuji problematice sbírání a ukládání dat z pohledu zákona. Budu se soustředit zejména na Zákon o elektronických komunikacích a Obecné nařízení o ochraně osobních údajů. Tato problematika je pro mou práci důležitá z několika důvodů. Pro správné zpracování dat potřebuji určit, která data by mohla vést k identifikaci jedince a která je tím pádem třeba odstranit. Dále budu rozebírat, jaké nároky klade zákon na zpracování různých druhů dat a jak to ovlivní můj program. Poté určím, jaké operace je možné s daty provádět a jestli jsou ze zákona některé operace s citlivými daty zakázané.

Nakonec se zaměřím na ochranu dat v české legislativě. Součástí této části bude i sekce zaměřená na Zákon o elektronických komunikacích. Další zákony, které rozeberu podrobněji, budou Zákon o zpracování osobních údajů a Občanský zákoník.

Výsledkem této části bude souhrn právních nároků na různé druhy dat a na zpracování těchto dat mým programem.

2.6.1 Obecné nařízení o ochraně osobních údajů

V této části se budu zabývat Obecným nařízenímm o ochraně osobních údajů, neboli Nařízení Evropského parlamentu a rady (EU) 2016/679 ze dne 27. dubna 2016 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů, zkráceně GDPR.

Je to nařízení Evropské unie, které

„představuje právní rámec ochrany osobních údajů platný na celém území EU, který hájí práva jejích občanů proti neoprávněnému zacházení s jejich daty a osobními údaji“ [40]

a zároveň

„přebírá všechny dosavadní zásady ochrany a zpracování údajů, na nichž unijní systém ochrany osobních údajů stojí a potvrzuje, že ochrana cestuje přes hranice současně s osobními údaji.“ [40]

Je to důsledek snahy zpřehlednit a sjednotit ochranu osobních údajů, aby při jejich zpracovávání v různých zemích Evropské unie nedocházelo k operacím s daty, které jsou v nezákonné v zemi původu dat, nebo mohou poškodit subjekt údajů.

Subjektem údajů, nebo subjektem dat, se v rámci GDPR rozumí:

„identifikovatelná fyzická osoba, kterou lze přímo či nepřímo identifikovat, zejména odkazem na určitý identifikátor, např. jméno, identifikační číslo, lokační údaje, síťový identifikátor nebo na jeden či více zvláštních prvků fyzické, fyziologické, genetické, psychické, ekonomické, kulturní nebo společenské identity této fyzické osoby“ [41, str. 9]

S tímto pojmem souvisí pojem správce dat, nebo také jen správce, který se definuje jako *„každý subjekt, který určuje účel a prostředky zpracování osobních údajů, provádí zpracování a odpovídá za něj“* [41, str. 9]. Je to tedy fyzická či právnická osoba, která nese odpovědnost za data, která se týkají jednotlivých subjektů údajů.

Zpracování a uchovávání osobních údajů, tak jak je upraveno v GDPR, je zapracováno do České legislativy v podobě Zákona o zpracování osobních údajů [42].

2.6.1.1 Zásady zpracování dat

Mezi hlavní zásady zpracování dat dle GDPR patří zákonnost, korektnost, transparentnost, omezení účelu, minimalizace údajů, přesnost, omezení uložení, integrita a důvěrnost [43].

Za zákonnost a korektnost se považuje zpracovávání údajů pouze pokud k tomu má správce dat alespoň jeden právní důvod [43].

Transparentní zpracování dat obnáší zejména informování subjektu údajů o probíhající operaci a jejích účelech [41, str. 49]. Zároveň by měl správce údajů

„poskytnout veškeré další informace nezbytné pro zajištění spravedlivého a transparentního zpracování, s přihlédnutím ke konkrétním okolnostem a kontextu, v němž jsou osobní údaje zpracovávány.“ [41, str. 49]

Omezením účelu se rozumí, že *„ osobní údaje musí být shromažďovány pro určité a legitimní účely a nesmějí být zpracovávány neslučitelným způsobem s těmito účely“* [43].

Stejným způsobem je omezen i sběr dat, při kterém je správce oprávněn vyžadovat jen ty údaje, které jsou relevantní vzhledem k účelu sběru a zpracování dat. Zároveň musí údaje, které sbírá nebo zpracovává, přesně zaznamenat, tj. nesmí v nich být chybné údaje. Tyto zásady jsou nazývány minimalizace údajů a přesnost.

Zásada omezení uložení znamená ukládání dat, obsahujících citlivé údaje, jen na nezbytně nutnou dobu, tzn. po dobu zpracování dat k potřebným účelům a zásada integrity a důvěrnosti zajišťuje *„technické a organizační zabezpečení osobních údajů“*. [43]

Zásadu integrity a důvěrnosti jsem již probíral v sekci Anonymizace a bezpečnost, ve které jsem nastínil problematiku neanonymních dat včetně postihů, které hrozí při úniku citlivých údajů. Jedním z navrhovaných řešení je anonymizace dat, pokud to povaha dat a jejich účel povolují.

Pokuty správci údajů hrozí také za porušení ostatních zásad zpracování dle GDPR. Mezi největší pokuty, které v tomto ohledu byly uděleny, patří pokuta společnosti Google, *„který za porušení GDPR dostal pokutu 50 milionů eur od francouzského Úřadu na ochranu osobních údajů CNIL“* a to v souvislosti se zásadou transparentnosti, kdy firma dle soudu *„uživatelům ve svých podmínkách neposkytla dostatek informací o tom, jak s jejich daty nakládá“* [7].

Pro zpracování osobních je nutné udělit souhlas, který je definován jako:

„svobodný, konkrétní, informovaný a jednoznačný projev vůle, kterým subjekt údajů dává prohlášením či jiným zjevným potvrzením své svolení ke zpracování svých osobních údajů“. [43]

Při zpracování dat se souhlas chápe jako právní důvod zpracování dat, ale je poskytován vždy k určitému účelu zpracování [43].

2.6.1.2 Osobní údaje

V této sekci rozeberu pojem osobní údaje, který je podstatný pro zpracovávání dat, při kterém je třeba zpracovávat osobní údaje dle jiných kritérií, než ostatní data. Osobní údaje vyžadují vyšší stupeň ochrany, protože jejich únik může znamenat pro správce dat mnohem větší ztráty nebo právní postihy, než je tomu u ostatních dat, jak jsem rozebíral v sekci Anonymizace a bezpečnost.

Osobními údaji se dle GDPR rozumí:

„veškeré informace o identifikované nebo identifikovatelné fyzické osobě (dále jen 'subjekt údajů'); identifikovatelnou fyzickou osobou je fyzická osoba, kterou lze přímo či nepřímo identifikovat, zejména odkazem na určitý identifikátor, například jméno, identifikační číslo, údaje o lokaci, síťový identifikátor nebo na jeden či více zvláštních prvků fyzické, fyziologické, genetické, psychické, ekonomické, kulturní nebo společenské identity této fyzické osoby“ [8, str. 33].

Protože se dle této definice počítá mezi osobní údaje i síťový identifikátor, je nutné považovat za osobní údaj IP adresu fyzické osoby.

Zvláštní kategorií osobních údajů jsou citlivé údaje. Tuto kategorii jsem rozebíral v sekci Citlivá data. Pro účely zpracování těchto údajů dle GDPR je důležité, že se jedná o

„údaje o rasovém či etnickém původu, politických názorech, náboženském nebo filozofickém vyznání, členství v odborech, o zdravotním stavu, sexuální orientaci a trestních deliktech či pravomocném odsouzení osob“ [23]

přičemž se nově za citlivý údaj považují i genetické a biometrické údaje. Dále je podstatné, že *„Zpracování citlivých osobních údajů podléhá mnohem přísnějšímu režimu, než je tomu u obecných údajů“*. [23]

Jak by měl tento přísnější režim vypadat, nechává Obecné nařízení o ochraně osobních údajů na bližší určení členským státům. *„Toto nařízení rovněž poskytuje členským státům určitý prostor ke stanovení vlastních pravidel, včetně pravidel pro zpracování zvláštních kategorií osobních údajů ('citlivé osobní údaje').“* [8, str. 2]

Osobní údaje se dále dělí do dvou skupin. První skupina jsou přímé identifikátory a druhá skupina jsou identifikátory nepřímé. Obě skupiny jsou brány jako osobní údaje, ale je v nich podstatný rozdíl s ohledem identifikování jedince na základě jedné nebo druhé skupiny dat.

Mezi přímé identifikátory patří jméno, příjmení, emailová adresa obsahující jméno nebo příjmení, telefonní číslo, nebo číselný identifikátor, jako je číslo občanského průkazu, pasu nebo zdravotního pojištění. Na základě tohoto typu identifikátoru lze přímo určit identitu jedince.

Nepřímé identifikátory jsou datum narození, pohlaví, poštovní směrovací číslo, nebo číslo poznávací značky. Tyto identifikátory přímo neurčují identitu

jedince a nelze je bez dalších údajů použít k jednoznačné identifikaci jedince.

Pojmy přímých a nepřímých identifikátorů souvisí s už dříve zmíněnými pojmy identifikovaný a identifikovatelný. Vztah mezi těmito pojmy je takový, že jedinec je v sadě dat identifikovaný, pokud sada dat obsahuje přímé identifikátory. Pokud sada dat obsahuje pouze nepřímé identifikátory, je jedinec v sadě dat identifikovatelný.

2.6.1.3 Anonymizovaná data

Zvláštní kategorií v GDPR jsou anonymizovaná data. Vztahují se na ně jiné požadavky, mnohem méně přísné, jak budu dále rozebírat, než na ostatní druhy dat. Anonymizovaná data jsou dle GDPR taková, na základě kterých nelze z dat přímo či nepřímo vyčlenit informace o jednotlivcích:

„Při určování, zda je fyzická osoba identifikovatelná, by se mělo přihlédnout ke všem prostředkům, jako je například výběr vyčleněním, o nichž lze rozumně předpokládat, že je správce nebo jiná osoba použijí pro přímou či nepřímou identifikaci dané fyzické osoby. Ke stanovení toho, zda lze rozumně předpokládat použití prostředků k identifikaci fyzické osoby, by měly být vzaty v úvahu všechny objektivní faktory, jako jsou náklady a čas, které si identifikace vyžádá, s přihlédnutím k technologii dostupné v době zpracování i k technologickému rozvoji.“ [8, str. 5]

To znamená, že anonymizace dat musí být provedena takovým způsobem, aby identifikace jedince z dat byla nemožná nebo výpočetně složitá a nákladná s použitím současných prostředků i těch, které je možné očekávat v nejbližších pár letech. Teprve po zpracování dat takovýmto způsobem je možné data považovat za anonymní.

Pokud jsou data anonymní, neklade na jejich zpracování a uložení GDPR žádné nároky, protože jsou dle recitálu 26 GDPR taková data vyňata z nároků na zpracování ostatních typů dat tímto nařízením.

„Zásady ochrany osobních údajů by se proto neměly vztahovat na anonymní informace, totiž informace, které se netýkají identifikované či identifikovatelné fyzické osoby, ani na osobní údaje anonymizované tak, že subjekt údajů není nebo již přestal být identifikovatelným. Toto nařízení se tedy netýká zpracování těchto anonymních informací, včetně zpracování pro statistické nebo výzkumné účely.“ [8, str. 5]

Operace s takovými daty, které zahrnují například ukládání, statistické zpracování, nebo publikování, je mnohem jednodušší, protože se na ně nevztahují zásady, které jsem uvedl v sekci Zásady zpracování dat.

Nároky na zpracování dat se tím pádem vztahují na data, která ještě neprošla programem na anonymizaci. Jinak řečeno, můj program bude mít na vstupu data, na které se vztahují jiné nároky, než na data na výstupu.

Odlišné nároky má GDPR na data, která prošla procesem pseudomizace. Taková data nejsou považována za anonymní a vztahují se na něj zásady pro zpracování dat.

„Osobní údaje, na něž byla uplatněna pseudonymizace a jež by mohly být přiřazeny fyzické osobě na základě dodatečných informací, by měly být považovány za informace o identifikovatelné fyzické osobě.“ [8, str. 5]

Taková data tedy stále patří ke konkrétní fyzické osobě, protože existuje způsob, jak zvrátit proces pseudomizace a získat původní data, jak jsem popisoval v sekci Pseudomizace.

Za bezpečná pro zpracovávání lze považovat i pseudomizovaná data, pokud správce údajů splní několik podmínek, které rozvedu v následující sekci.

2.6.1.4 Zpracování dat

Pro účely GDPR se zpracováním rozumí:

„jakákoliv operace nebo soubor operací s osobními údaji nebo soubory osobních údajů, který je prováděn pomocí či bez pomoci automatizovaných postupů, jako je shromáždění, zaznamenání, uspořádání, strukturování, uložení, přizpůsobení nebo pozměnění, vyhledání, nahlédnutí, použití, zpřístupnění přenosem, šíření nebo jakékoliv jiné zpřístupnění, seřazení či zkombinování, omezení, výmaz nebo zničení; “ [8, str. 33]

Z toho vyplývá, že proces anonymizace, jak ho provádí můj soubor, spadá do definice zpracování dle GDPR a musí dodržovat nároky dané tímto nařízením na práci s daty.

O zabezpečení zpracování se Obecné nařízení o ochraně osobních údajů zmiňuje i dále [8, str. 51]:

„S přihlédnutím ke stavu techniky, nákladům na provedení, povaze, rozsahu, kontextu a účelům zpracování i k různě pravděpodobným a různě závažným rizikům pro práva a svobody fyzických osob, provedou správce a zpracovatel vhodná technická a organizační opatření, aby zajistili úroveň zabezpečení odpovídající danému riziku, případně včetně:

a) pseudonymizace a šifrování osobních údajů; (...)“

Z tohoto vyplývá, že vhodná technická a organizační opatření je třeba zajistit i pro proces anonymizace dat, ale konkrétní kroky ponechává nařízení na správci dat, případně na detailnější úpravu členskými státy ve své právním systému. Zabezpečení dat pomocí procesu pseudonymizace je v rámci GDPR povoleno, pokud splní daná kritéria.

„S cílem vytvořit pobídky pro uplatňování pseudonymizace při zpracování osobních údajů by opatření pseudonymizace při současném umožnění obecné analýzy měla být možná v rámci téhož správce, pokud tento správce přijal technická a organizační opatření nezbytná k zajištění toho, aby bylo v případě daného zpracování provedeno toto nařízení a aby doplňkové informace pro přiřazení osobních údajů konkrétnímu subjektu údajů byly uchovány samostatně.“ [8, str. 5]

Za těchto podmínek je možné data zpracovávat a uchovávat a stále jsou považována za zabezpečená, což je důležité v případě úniku dat, kdy se přihlíží i k použitým technickým a organizačním opatřením. Tato část nařízení je implementována do českého právního systému v Zákoně o zpracování osobních údajů, jak budu rozebírat dále.

2.6.2 Ochrana dat v české legislativě

Ochrana osobních údajů má ústavní základ, jak vyplývá z následujících příkladů. Dle Listiny základních práv a svobod Evropské unie [44] hlavy II, článek 8: *„každý má právo na ochranu osobních údajů, které se ho týkají“*. Dále dle Listiny základních práv a svobod České republiky [45]: *„Každý má právo na ochranu před neoprávněným shromažďováním, zveřejňováním nebo jiným zneužíváním údajů o své osobě“*.

Obecnou úpravu ochrany dat v České republice zajišťují zákon č. 110/2019 Sb., Zákon o zpracování osobních údajů [42], a zákon č. 89/2012 Sb., Občanský zákoník [6].

Dalším důležitým zákonem pro nakládání s daty je pro účely mé bakalářské práce Zákon o elektronických komunikacích, protože data v ukázce jsou data telekomunikační. Tento zákon rozeberu v samostatné sekci.

2.6.2.1 Zákon o zpracování osobních údajů

Zákon č. 110/2019 Sb., neboli zákon o zpracování osobních údajů, který vešel v účinnost ke dni 24. 4. 2019 vyhlášením ve sbírce zákonů. Vyšel jako implementace Obecného nařízení o ochraně osobních údajů, vydaného Evropskou unií, které jsem rozebíral v sekci Obecné nařízení o ochraně osobních údajů.

Zároveň nahradil starší zákon, zákon č. 101/2000 Sb., o ochraně osobních údajů, který nebyl v souladu s evropským nařízením. Proto se nový zákon dá chápat jako aktualizace toho starého, tak aby vyhovoval evropskému nařízení, které mělo za cíl sjednotit ochranu dat v zemích Evropské unie. [46]

Nová verze zákona pouze neimplementuje nařízení GDPR tak, jak byl schválen Evropskou komisí, ale upravuje a detailněji rozvádí některé části tohoto nařízení.

„(...) dikce Nařízení GDPR umožňuje členským státům úpravu některých právních institutů vnitrostátní legislativou, přičemž dává rovněž možnost upřesnit či jinak rozvést již v Nařízení GDPR zakotvená ustanovení, čímž umožňuje zákonodárci odchýlit se od obecné úpravy a přijmout určité národní výjimky, které jsou lépe uzpůsobeny konkrétnímu právnímu prostředí.“ [46]

Proto se v některých ohledech, z nichž jsem některé zmiňoval v sekci Obecné nařízení o ochraně osobních údajů, můžou jednotlivé státy rozhodnout, jakým způsobem budou nařízení do svého právního systému implementovat.

Z tohoto důvodu je pro zpracování dat na území České republiky důležitější místní úprava evropského nařízení ve formě Zákona o ochraně osobních údajů a dalších zákonů, které toto téma ošetřují. Konkrétní úprava má přednost při sporech o osobní data.

Jedním z případů, kdy se česká úprava liší od obecného evropského nařízení, je věk nabytí způsobilosti k udělení souhlasu se zpracováním osobních údajů. Dle [42]

„dítě nabývá způsobilosti k udělení souhlasu se zpracováním osobních údajů v souvislosti s nabídkou služeb informační společnosti přímo jemu dovršením patnáctého roku věku“,

zatímco dle [8, str. 37] *„v souvislosti s nabídkou služeb informační společnosti přímo dítěti, je zpracování osobních údajů dítěte zákonné, je-li dítě ve věku nejméně 16 let“*. Český zákon takto posunuje spodní věkovou hranici k udělení souhlasu o jeden rok.

V otázce bezpečnosti osobních údajů při zpracování a rizik s tím spojených, neukládá česká úprava povinnost správci dat vypracovat posouzení tohoto zpracování na ochranu osobních údajů, jak vyplývá z paragrafu 10 Zákona o zpracování osobních údajů:

„Správce nemusí provádět posouzení vlivu zpracování na ochranu osobních údajů před jeho zahájením, pokud mu právní předpis stanoví povinnost takové zpracování osobních údajů provést.“ [42]

Oproti tomu, GDPR v recitálu 84 stanoví:

„S cílem přispět k zajištění souladu s tímto nařízením v případech, kdy je pravděpodobné, že operace zpracování budou představovat vysoké riziko pro práva a svobody fyzických osob, by měl být správce odpovědný za provedení posouzení vlivu na ochranu osobních údajů, aby vyhodnotil zejména původ, povahu, zvláštnost a závažnost tohoto rizika.“ [8, str. 16]

Cílem tohoto posouzení by měla být analýza rizik a jejich pravděpodobnosti při plánovém zpracování dat. Při zjištění vysoké pravděpodobnosti úniku dat,

ukládá nařízení kontaktovat příslušný úřad [8, str. 16].

V ostatních případech česká úprava spíše konkretizuje, například konkrétní hodnoty pokuty za únik osobních údajů. Dle [42] se právnická osoba dopustí přestupku mimo jiné tím, že *„nepřijme opatření zajišťující, aby osobní údaje byly přesné ve vztahu k povaze a účelu jejich zpracování“*, nebo *„ uchovává osobní údaje po dobu delší než nezbytnou k dosažení účelu jejich zpracování“*. Za tyto a další přestupky může být uložena pokuta do 10 000 000 Kč [42].

Pokuty dle GDPR mohou být mnohem vyšší, protože horní hranice u stejného druhu přestupku činí 20 000 000 eur nebo 4% obratu podniku, podle toho, která hodnota je vyšší [8, str. 83].

Zákon dále stanoví, že

„pokud to umožňuje dosáhnout účelu uvedeného v odstavci 1, osobní údaje uvedené v čl. 9 odst. 1 nařízení Evropského parlamentu a Rady (EU) 2016/679 správce nebo zpracovatel dále zpracovává v podobě, která neumožňuje identifikaci subjektu údajů, ledaže tomu brání oprávněné zájmy subjektu údajů“.

To znamená, že i v české úpravě je preferovaná možnost skladování dat v anonymní podobě, tedy takových, které prošly procesem anonymizace, pokud to umožňuje jejich povaha a účel, za kterým jsou zpracovávány. Tato část zákona je přímo implementována z GDPR, dle kterého se na taková data nevztahují pravidla v tomto nařízení uvedená [8, str. 5].

Jak vyplývá z [42], správce dat má mimo jiné následující možnost pro zabezpečení zpracování dat:

„Správce nebo zpracovatel při zpracování osobních údajů za účelem vědeckého nebo historického výzkumu nebo pro statistické účely zajistí dodržování konkrétních opatření k ochraně zájmů subjektu údajů, která odpovídají stavu techniky, nákladům na provedení, povaze, rozsahu, kontextu a účelům zpracování i různě pravděpodobným a závažným rizikům pro práva a svobody fyzických osob.“

Tato část Zákona o zpracování osobních údajů je implementací GDPR do českého právního systému, konkrétně nařízení, které umožňuje využití pseudomizace jako nástroje na zabezpečení citlivých dat během jejich skladování [8, str. 5].

Ačkoliv pseudomizace není považování v rámci GDPR za úpravu dat, po které už není třeba data zpracovávat v souladu s tímto nařízením, jako je tomu u anonymizace, stanoví, že je to za určitých podmínek bráno jako dostatečná ochrana citlivých dat. Tato část Zákona o zpracování osobních údajů tuto skutečnost implementuje do českého právního systému.

Na moji práci a vyvíjený program se vztahují nároky kladené na zpracovatele osobních údajů pro statistické účely. Dle [42] jsou to zejména:

„pořizování záznamů alespoň o všech operacích shromáždění, vložení, pozmě-

nění a výmazu osobních údajů, které umožní určit a ověřit totožnost osoby provádějící operaci, a uchovávání těchto záznamů nejméně po dobu 2 let od provedení operace“

a také:

„technická a organizační opatření zaměřená na důsledné uplatnění povinnosti podle čl. 5 odst. 1 písm. c) nařízení Evropského parlamentu a Rady (EU) 2016/679“,

které se vztahuje k GDPR a které rozeberu v samostatné kapitole.

Z těchto důvodů je nutné zaznamenávat informaci o tom, že byla některá data smazána, nebo že byla data upravena tak, aby byla anonymní. Tato informace nebude obsahovat konkrétní data, jen informaci o tom, že data byla zpracována a anonymizována. Informace může být například ve formě textového souboru s datem zpracování. Tento soubor je určen pro následné uchování, jak je to vyžadováno v [42].

2.6.2.2 Občanský zákoník

Zákon č. 89/2012 Sb., neboli Občanský zákoník, je zákon, který vstoupil v platnost 22. 3. 2012 a obsahuje pět částí: Obecnou část, Rodinné právo, Absolutní majetková práva, Relativní majetková práva a Ustanovení přechodná, společná a závěrečná. Otázce osobnosti a osobních údajů se věnuje Obecná část, Hlava II, konkrétně šestý oddíl. [6]

Pro definování práv jedince na ochranu osobních údajů dle Občanského zákoníku je podstatný pojem právní osobnost. Právní osobnost je dle [6] *„způsobilost mít v mezích právního řádu práva a povinnosti“*. Právní osobnosti se nelze za žádných okolností vzdát a to ani z části. Trvá od narození člověka až do jeho smrti. Zároveň je zakázáno narušovat právo na ochranu osobnosti, které se vztahuje i na ochranu osobních údajů. [6]

Konkrétní narušení soukromí specifikuje zákon například jako zachycení podoby člověka, aniž by k tomu tento člověk dal výslovný souhlas, pokud je možné na základě tohoto údaje určit jeho totožnost. Stejně tak je možné šířit zachycenou podobu člověka pouze s jeho svolením. [6]

Sběr dat o určitém člověku je v tomto zákoně definován jako zasažení do soukromí jiného, pokud k tomu správce dat nemá zákonný důvod.

„Zejména nelze bez svolení člověka narušit jeho soukromé prostory, sledovat jeho soukromý život nebo pořizovat o tom zvukový nebo obrazový záznam, využívat takové či jiné záznamy pořízené o soukromém životě člověka třetí osobou, nebo takové záznamy o jeho soukromém životě šířit. Ve stejném rozsahu jsou chráněny i soukromé písemnosti osobní povahy.“

K těmto úkonům, tedy sběru dat a následném zpracování, je opět nutné udělit souhlas, který může být kdykoliv subjektem dat odvolán. [6]

2.6.3 Zákon o elektronických komunikacích

Zákon č. 127/2005 Sb., neboli Zákon o elektrických komunikacích, vešel v platnost 31. 3. 2005 a je to zákon, jehož předmět úpravy je oblast elektronických komunikací, včetně podmínek podnikání a regulací ze strany státu. Pro moji bakalářskou práci je důležitý, protože data dodaná vedoucím práce mají povahu telekomunikačních dat a jako takové je nutné je zpracovávat v souladu s tímto zákonem. [47]

Porušení ochrany osobních údajů, jak je definováno pro potřeby Zákona o elektronických komunikacích, znamená:

„porušení bezpečnosti, které vede k neoprávněnému přístupu nebo k neoprávněné nebo nahodilé změně, zničení, vyzrazení či ztrátě osobních údajů zpracovávaných v souvislosti s poskytováním veřejně dostupné služby elektronických komunikací“ [47]

To znamená, že jako narušení bezpečnosti je brán nejen únik dat, ale i zničení svěřených údajů.

Některé z těchto údajů není možné anonymizovat, protože je podnikatel, který provozuje veřejně dostupnou telefonní služby, povinen předat poskytovatelům univerzální služby. Jde zejména o osobní a identifikační údaje fyzických a právnických osob. Osobními údaji se rozumí jméno, příjmení, adresa trvalého pobytu a elektronické pošty a telefonní číslo. Identifikačními údaji se rozumí *„obchodní firma nebo název nepodnikající právnické osoby, adresa sídla, popřípadě adresa sídla organizační složky, adresa a telefonní číslo provozovny a adresa elektronické pošty.“* [47]

Univerzální službou se v rámci tohoto zákona rozumí *„soubor služeb stanovený v paragrafu 38, které jsou dostupné ve stanovené kvalitě všem koncovým uživatelům na celém území státu za dostupnou cenu“*. [47]

Všechny tyto údaje je možné univerzální službě poskytnout pouze se souhlasem uživatelů, tj. právnických a fyzických osob, subjektů dat, kteří o tomto postoupení údajů musejí být informováni, stejně jako o účelu univerzální služby. [47]

Kromě univerzální služby je podnikatel poskytující veřejně dostupnou službu elektronických komunikací povinen poskytnout aktuální osobní a identifikační údaje všech fyzických a právnických osob, kteří jsou jeho uživateli, podnikateli, který zajišťuje *„připojení k veřejné pevné komunikační síti subjektu, který provozuje pracoviště pro příjem volání na čísla tísňového volání“*. Údaje slouží pro lokalizaci a identifikaci a je nutné je pravidelně aktualizovat. [47]

O tomto předání informací nemusí být uživatel telekomunikační společnosti informován. Podnikatel má však povinnost informovat uživatele o způ-

sobu používání tísňové linky. Data, určená k tomuto účelu nesmí být použita pro žádný jiný účel. [47]

2.6.3.1 Ochrana osobních údajů

Pro provozovatele telekomunikačních služeb vyplývají ze Zákona o elektronických komunikacích povinnosti na zajištění ochrany osobních údajů a citlivých dat, ke kterým má přístup, a to včetně důvěrnosti komunikací. Dozor nad dodržováním těchto nařízení má, stejně jako v případě Zákona o ochraně osobních údajů, Úřad pro ochranu osobních údajů. [47]

Dle Zákona o elektronických komunikacích je jejich provozovatel mimo jiné povinen:

„zajistit technicky a organizačně bezpečnost poskytované služby s ohledem na ochranu osobních údajů fyzických osob v souladu se zvláštním právním předpisem, ochranu provozních a lokalizačních údajů a důvěrnost komunikací fyzických a právnických osob při poskytování této služby“

To znamená, že během provozu služby musí být data, která poskytují uživatelé, a která jsou nezbytná pro provoz služby, a dle kterých je možné určit identitu nebo polohu uživatele služby, zabezpečeny po celou dobu služby. Stejně tak je provozovatel povinen zajistit, že neunikne komunikace uživatelů, neboli zpráva.

Zpráva je definována jako:

„jakákoli informace, která se vyměňuje nebo přenáší mezi konečným počtem účastníků nebo uživatelů prostřednictvím veřejně dostupné služby elektronických komunikací, s výjimkou informace přenášené jako součást veřejného rozhlasového nebo televizního vysílání sítí elektronických komunikací, nelze-li ji přiřadit k určitelnému účastníkovi nebo uživateli, který tuto informaci přijímá“ [47]

Anonymizace těchto údajů by zajistila dostatečnou bezpečnost, ale neumožnila by provozování služby, protože data, jako jsou polohy uživatelů, nebo jejich komunikace, jsou pro službu podstatné.

Zákon dále stanoví, že provozovatel *„ochranu údajů a důvěrnost komunikací zajistí s ohledem na stávající technické možnosti a na náklady potřebné k zajištění ochrany na úrovni odpovídající existujícímu riziku porušení ochrany“* [47]. Provozovatel proto musí brát vždy v úvahu technologický vývoj, aby zabezpečení dat odpovídalo sofistikovanosti pokusů o získání těchto dat.

Ochranu osobních údajů a citlivých dat, zejména pak obsah zpráv a komunikace jednotlivých uživatelů, je provozovatel povinen zajistit nejen před vnějším útokem, ale také vnitřním. Z těchto důvodů je ze zákona nepřípustné odposlouchávat nebo ukládat soukromé zprávy uživatelů, až na výjimky dané

zákonem. Do těchto výjimek spadají například soudně nařízené odposlechy. Jinou povahu mají technická data pro přenos služby, která jsou potřebná pro přenos zpráv, protože zásada důvěrnosti zpráv *„nebrání technickému ukládání údajů, které je nezbytné pro přenos zpráv, aniž by byla dotčena zásada důvěrnosti“*. [47]

Pokud je bezpečnost dat narušena, je o tom provozovatel povinen informovat uživatele [47]. To může znamenat ztrátu důvěry stávajících i potenciálních uživatelů, jak jsem rozebíral v sekci Postihy za únik dat.

Práva a povinnosti neupravené v tomto zákoně se řídí dle Zákona o zpracování osobních údajů, který jsem rozebíral v sekci Zákon o zpracování osobních údajů [47].

2.6.3.2 Provozní údaje

Technickými údaji, neboli provozními údaji, *„se rozumí jakékoli údaje zpracovávané pro potřeby přenosu zprávy sítí elektronických komunikací nebo pro její účtování“*. Tyto údaje nejsou zahrnuty do zásady důvěrnosti zpráv uživatele, nicméně podléhají regulacím ze strany tohoto zákona. [47]

Provozovatel telekomunikačních služeb, který

„zpracovává a ukládá provozní údaje, včetně příslušných lokalizačních údajů, vztahujících se k uživateli nebo účastníku, je musí smazat nebo učinit anonymními, jakmile již nejsou potřebné pro přenos zprávy“. [47]

Tyto údaje tedy nelze poté, co již nejsou potřeba k provozu služby, uchovávat jinak než jako anonymní.

Tato část zákona se týká ukázek dodaných vedoucím práce. Data, která se uchovávají pro další zpracování, například pro statistické vyhodnocení, proto musí projít procesem anonymizace.

Kapitola 3

Realizace

3.1 Programování anonymizéru

Můj program se dělí na dvě části: centrální logiku anonymizéru a modul, který data připraví ke zpracování. Modul je určen pro konkrétní typ dat. Tímto způsobem může centrální logika fungovat bez ohledu na vstupní data. Pro svoji bakalářskou práci jsem naprogramoval dva moduly pro dva typy dat.

3.1.1 Centrální jednotka

Při programování anonymizéru jsem se nejprve zaměřil na centrální část programu, která má za úkol z dat, které dostane, odstranit sloupce, které dostane zadané. Tato část programu neřeší, jak data vypadají, to mají na starost moduly, které data upraví do podoby, ve které je centrální část zpracuje, a následně je opět převezme a pokud je třeba, upraví je do formátu, ve kterém byly před úpravou modulem.

Při tvorbě této části jsem se nejprve zaměřil na správné čtení vstupu. Jednotka dostane v parametrech jméno souboru, ve kterém jsou sloupce k vymazání a jméno souboru s daty. Tyto soubory otevře a zkontroluje, jestli se je podařilo správně otevřít. Zároveň s tím si otevře soubor pro výstup, jehož jméno je zadáno v programu. Všechny soubory musí existovat.

Potom pokračuje načítáním vstupu a kontroluje, jestli se číslo sloupce shoduje s čísly sloupců určených k vymazání. Pokud se neshoduje, překopíruje data do nového souboru. V opačném případě data nekopíruje.

Výstupem této části programu jsou anonymizovaná data, která si převezme konkrétní modul určený pro daný typ dat. Data se v programu nikde neukládají, program je překopíruje přímo do předem určeného výstupního souboru.

3.1.2 Moduly pro různé druhy dat

V mé bakalářské práci jsem programoval dva moduly pro dva různé druhy dat. Modul má za úkol upravit data na vstupu do obecné podoby, ve které je zpracovává centrální jednotka.

Modul dostane od centrální jednotky jeden řádek ze vstupu a vektor z knihovny vector.h. Potom parsuje řádek podle předem daných pravidel, která jsou jiná pro každý modul a závisí na vstupních datech. Každou položku zvlášť umístí do vektoru předaného v parametru centrální částí. Každá položka ve vektoru odpovídá jednomu sloupci v datech.

Vektor obsahuje pouze čistá data, bez oddělovačů. Jedinou výjimkou jsou znaky uvozovek, které zůstávají zachovány, aby bylo možné zachovat zdrojové formátování. I přesto jsou data v obecné podobě a centrální část je zpracuje, protože pracuje s každou položkou vektoru zvlášť.

Po zpracování dat centrální část vrátí zbylá data modulu, který je upraví zpět do zdrojového formátování. Postupně vybírá data z vektoru a přidává k nim oddělovače, které z nich předtím odstranila a vytvoří nový řádek. Oddělovače jsou i na místech, kde předtím byla data, aby zůstalo zachováno pořadí sloupců. Hotový řádek předá centrální jednotce, která ho uloží do předem definovaného výstupního souboru.

Zpracování po řádcích je podle mého názoru dobrou možností pro zpracování proudu dat. Centrální část si vybírá řádky ze vstupního proudu a po jednom je předává modulu, který je hned vrátí upravené. Po osekání si centrální jednotka opět nechá data upravit v modulu a předá je do výstupního souboru, poté začne zpracovávat další řádek. Program tak může pracovat nepřetržitě, dokud má nějaká data na vstupu.

3.1.2.1 První modul

První modul jsem navrhl tak, aby zpracovával data z první a třetí ukázky dodané vedoucím mé práce. Data jsou formátována jako jednotlivé záznamy, kde každý záznam má 58 položek oddělených čárkou. U těchto ukázek není vždy jeden záznam na jedné řádce, ale může být rozdělen na více řádků.

Modul proto načítá data po řádcích a počítá jednotlivé položky pomocí oddělovačů. Pokud je položek správný počet, tj. 58, vrátí centrální části informaci o tom, že záznam je kompletní a předá mu data k anonymizaci.

V opačném případě předá centrální části pouze informaci o tom, že záznam není kompletní. Centrální část pak modulu pošle další část dat. Toto se opakuje, dokud záznam není kompletní.

Po předání dat centrální části a následné anonymizaci, pošle centrální část anonymizovaná data modulu, které je opět skládá do podoby položek oddělených čárkou. Celý záznam pak předá centrální části, která ho uloží do výstupního souboru.

3.1.2.2 Druhý modul

Druhý modul jsem navrhl tak, aby zpracovával data z druhé ukázky dodané vedoucím mé práce. Data mají podobu čárkou oddělených datových položek, sloupců, které tvoří jeden záznam. Každý záznam má 58 položek a je na samostatné řádce.

Modul dostává od centrální jednotky jednotlivé záznamy, které rozděluje na položky. Tyto položky pak předá centrální jednotce k anonymizaci. Centrální jednotka je anonymizuje a předá zpět modulu, který je složí do původní podoby. Poté předá data zpět centrální jednotce, která je uloží do výstupního souboru.

3.1.3 Testování

Program jsem testoval pomocí dat dodaných vedoucím mé bakalářské práce. Testoval jsem načítání dat a mazání sloupců, včetně sloupců hraničních a tvořených seznamem hodnot. Po zpracování programem jsem data porovnával pomocí [48]. Z tohoto testování mi vyšlo, že program pro dané ukázky dat funguje správně. Centrální logika by takto měla fungovat i pro jiné druhy dat, pokud pro ně bude přidaný modul.

3.1.4 Shrnutí

Program funguje správně pro vzorky dodané vedoucím práce, citlivá data nejsou nikam vypisována ani ukládána. Zůstávají pouze ve vstupním souboru, který lze následně smazat nebo uchovat k dalšímu zpracování, podle konkrétní situace. Jeho běh je lineární vzhledem k datům, s každou řádkou je prováděn konstantní počet operací.

Pro změnu modulů je nutné program znovu zkompilovat. To je možné provést v linuxovém prostředí příkazem make [49] a přidáním klíčového slova module1 nebo module2.

Další rozšíření je možné skrze moduly, které zpracovávají data. Centrální jednotka obsahuje metody, které volá pro zpracování dat. Jsou to metody processInputData a processOutputData, které upravují vstupní, resp. výstupní, data. Tyto metody je nutné implementovat v nových modulech.

Závěr

V rešeršní části práce jsem měl za cíl získat přehled o technikách používaných při anonymizaci data o technikách používaných při hledání vzorů v datech, zhodnotit a vybrat nejvhodnější algoritmus pro anonymizaci dat a ověřit správnost procesu anonymizace oproti Obecnému nařízení o ochraně dat a Zákonu o elektronických komunikacích. Dále jsem měl za cíl analyzovat data dodaná vedoucím mé bakalářské práce a navrhnout optimální postup pro implementaci modulárního anonymizéru.

Tuto část práce jsem splnil. V rámci rešeršní práce jsem dospěl k algoritmu, který splňuje požadavky na anonymizaci dat, jak v rámci nároků kladených na fungování algoritmu, tak v rámci nároků kladených ze strany české legislativy. Výsledný algoritmus má lineární složitost. Data jsem analyzoval z obecného pohledu, i z pohledu různých kategorií a dělení dat včetně citlivých, pseudonymních a anonymních dat.

Cíl praktické části byl naprogramovat modulární anonymizér, skládající se z centrální části a modulů pro dodané druhy dat, přičemž centrální část měla fungovat nezávisle na modulech.

V rámci této části práce jsem navrhl a implementoval modulární anonymizér dat. V praktické části jsem vyzkoušel, že navržený program funguje správně pro vzorová data. Navrhl jsem dva moduly, do budoucna je možné rozšířit program o další moduly, které se napojí na centrální logiku programu.

Bibliografie

1. RAGHUNATHAN, B. *The Complete Book of Data Anonymization: From Planning to Implementation*. CRC Press, 2013. Infosys Press. ISBN 9781482218565. Dostupné také z: https://books.google.co.uk/books?id=yfEnAAAAQBAJ.
2. ARBUCKLE, L.; EMAM, K.E. *Building an Anonymization Pipeline: Creating Safe Data*. O'Reilly Media, 2020. ISBN 9781492053408. Dostupné také z: https://books.google.co.uk/books?id=f5bcDwAAQBAJ.
3. DOMINGO-FERRER, J.; SÁNCHEZ, D.; SORIA-COMAS, J. *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool Publishers, 2016. Synthesis Lectures on Information Security, Privacy, and Trust. ISBN 9781681731988. Dostupné také z: https://books.google.co.uk/books?id=ylk7DwAAQBAJ.
4. WARREN, S. D.; BRANDEIS, L. D. *The Right to Privacy* [online]. 1890 [cit. 2020-05-15]. Dostupné z: https://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html.
5. WORKSHOP, S.I.L.T. *Privacy in the Digital Environment*. Haifa Center of Law & Technology. ISBN 9789659092413. Dostupné také z: https://books.google.co.uk/books?id=yeVRrrJw-zAC.
6. ČESKO. *Zákon č. 89/2012 Sb. Občanský zákoník* [online]. 2012 [cit. 2020-04-30]. Dostupné z: https://www.zakonyprolidi.cz/cs/2012-89.
7. COMPUTER WORLD. *Za porušení GDPR letos padly pokuty ve výši stovek milionů eur* [online]. 2019 [cit. 2020-05-24]. Dostupné z: https://computerworld.cz/securityworld/za-poruseni-gdpr-letos-padly-pokuty-ve-vysi-stovek-milionu-eur-55733.
8. EVROPSKÁ UNIE. *Nařízení Evropského parlamentu a Rady (EU) 2016/679 ze dne 27. dubna 2016 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů a o zrušení směrnice 95/46/ES (obecné nařízení o ochraně osobních údajů)* [online]. 2016

[cit. 2020-04-21]. Dostupné z: https://eur-lex.europa.eu/legal-content/CS/ALL/?uri=CELEX%3A32016R0679.

9. ČTK. *Za dva roky starý únik dat pokuta. Uber musí v Británii a Nizozemsku zaplatit celkem přes 26 milionů* [online]. 2018 [cit. 2020-05-24]. Dostupné z: https://www.irozhlas.cz/ekonomika/uber-velka-britanie-nizozemi-unik-dat-pokuta_1811271222_pj.
10. ISC AFRICA. *Data Theft Definition* [online]. 2020 [cit. 2020-05-24]. Dostupné z: http://cybercrime.org.za/data-theft/.
11. OFFICE OF THE PRIVACY COMMISSIONER OF CANADA. *PIPEDA in brief* [online]. 2019 [cit. 2020-05-24]. Dostupné z: https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/.
12. EVROPSKÁ KOMISE. *Co jsou to osobní údaje?* [online]. 2019 [cit. 2020-04-25]. Dostupné z: https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data.
13. ROUSE, M. *What is data masking? - Definition from WhatIs.com. Information Security information, news and tips - SearchSecurity* [online]. 2009 [cit. 2020-04-21]. Dostupné z: https://searchsecurity.techtarget.com/definition/data-masking.
14. PECOVATELSKA.CZ [online] [cit. 2020-05-28]. Dostupné z: https://www.pecovatelska.cz/help/gdpr.html.
15. PROKOP, J. *Algoritmy v jazyku C a C++*. 3. vydání. Praha: Grada Publishing, a.s., 2015. ISBN 978-80-247-5467-3.
16. GOYVAERTS, J. *Runaway Regular Expressions: Catastrophic Backtracking* [online]. 2019 [cit. 2020-05-04]. Dostupné z: https://www.regular-expressions.info/catastrophic.html.
17. EPPSTEIN, D. *Design and Analysis of Algorithms, Lecture notes for February 27, 1996* [online]. 1996 [cit. 2020-05-04]. Dostupné z: https://www.ics.uci.edu/~eppstein/161/960227.html.
18. MAREŠ, M. *Přednáška z algoritmů a datových struktur II, Vyhledávání v textu* [online]. 2012 [cit. 2020-05-04]. Dostupné z: http://mj.ucw.cz/vyuka/1112/ads2/1-kmp.pdf.
19. VYSOKESKOLY.CZ. *Teorie práva* [online] [cit. 2020-05-28]. Dostupné z: https://www.vysokeskoly.cz/maturitniotazky/zaklady-spolecenskych-ved/teorie-prava.
20. SKLENÁK, V. *Data, informace, znalosti a Internet*. C.H. Beck, 2001. C.H. Beck pro praxi. ISBN 9788071794097. Dostupné také z: https://books.google.co.uk/books?id=UJh-gLdTH8IC.

21. MERRIAM, WEBSTER. *Data, Merriam-Webster Dictionary* [online] [cit. 2020-05-24]. Dostupné z: https://www.merriam-webster.com/dictionary/data.

22. WIKISOFIA.CZ. *Data* [online]. 2013 [cit. 2020-05-24]. ISSN 2336-5897. Dostupné z: https://wikisofia.cz/wiki/Data.

23. ŠKORNIČKOVÁ, E. *Citlivé osobní údaje* [online] [cit. 2020-05-24]. Dostupné z: https://www.gdpr.cz/gdpr/heslo/citlive-osobni-udaje/.

24. PHOTOPOULOS, C. *Managing Catastrophic Loss of Sensitive Data: A Guide for IT and Security Professionals*. Elsevier Science, 2011. ISBN 9780080558714. Dostupné také z: https://books.google.cz/books?id=u6Bm9NmzzxsC.

25. MAYER-SCHÖNBERGER, V.; CUKIER, K. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt, 2013. An Eamon Dolan book. ISBN 9780544002692. Dostupné také z: https://books.google.co.uk/books?id=uy4lh-WEhhIC.

26. HOLUBOVÁ, I.; KOSEK, J.; MINAŘÍK, K.; NOVÁK, D. *Big Data a NoSQL databáze*. Grada, 2015. ISBN 9788024754666. Dostupné také z: https://books.google.cz/books?id=x93yCgAAQBAJ.

27. BUYYA, B.; CALHEIROS, R. N.; DASTJERDI, A. V. *Big Data: Principles and Paradigms*. Elsevier Inc., 2016. ISBN 978-0-12-805394-2. Dostupné také z: https://www.sciencedirect.com/topics/computer-science/big-data-processing.

28. BANKMYCELL.COM. *HOW MANY SMARTPHONES ARE IN THE WORLD?* [online] [cit. 2020-05-26]. Dostupné z: https://www.bankmycell.com/blog/how-many-phones-are-in-the-world.

29. LD.JOHANESVILLE.NET. *Historie písma* [online] [cit. 2020-05-27]. Dostupné z: http://ld.johanesville.net/historie/03-historie-pisma.

30. HOLMES, D.E. *Big Data: A Very Short Introduction*. Oxford University Press, 2017. Very short introductions. ISBN 9780198779575. Dostupné také z: https://books.google.cz/books?id=NXw7DwAAQBAJ.

31. REGIONÁLNÍ MUZEUM K. A. POLÁNKA. *Mladší paleolit* [online]. 2012 [cit. 2020-05-27]. Dostupné z: https://www.muzeumzatec.cz/mladsi-paleolit.html.

32. MARK, J. J. *Sumerians* [online]. 2019 [cit. 2020-05-27]. Dostupné z: https://www.ancient.eu/Sumerians/.

33. HISTORY.COM. *Ancient Egypt* [online]. 2009 [cit. 2020-05-28]. Dostupné z: https://www.history.com/topics/ancient-history/ancient-egypt.

34. THE NATIONAL MUSEUM OF COMPUTING. *The Turing-Welchman Bombe* [online] [cit. 2020-05-28]. Dostupné z: https://www.tnmoc.org/bombe.

35. BROWN, M. *KONRAD ZUSE'S Z3, THE WORLD'S FIRST PROGRAMMABLE COMPUTER, WAS UNVEILED 75 YEARS AGO* [online]. 2016 [cit. 2020-05-28]. Dostupné z: https://www.inverse.com/article/15542-konrad-zuse-s-z3-the-world-s-first-programmable-computer-was-unveiled-75-years-ago.

36. FREIBERGER, P. A.; SWAINE, M. R. *ENIAC* [online] [cit. 2020-05-28]. Dostupné z: https://www.britannica.com/technology/ENIAC.

37. FREIBERGER, P. A.; SWAINE, M. R. *UNIVAC* [online] [cit. 2020-05-28]. Dostupné z: https://www.britannica.com/technology/UNIVAC.

38. WORLD WIDE WEB FOUNDATION. *History of the Web* [online] [cit. 2020-05-28]. Dostupné z: https://webfoundation.org/about/vision/history-of-the-web/.

39. CERN. *A short history of the Web* [online] [cit. 2020-05-28]. Dostupné z: https://home.cern/science/computing/birth-web/short-history-web.

40. ÚŘAD PRO OCHRANU OSOBNÍCH ÚDAJŮ. *GDPR (obecné nařízení)* [online] [cit. 2020-05-28]. Dostupné z: https://www.uoou.cz/gdpr/ds-3938/p1=3938.

41. JANEČKOVÁ, E. *GDPR - Řešení problémů v praxi obcí*. Grada Publishing, 2019. Právo pro praxi. ISBN 9788024729251. Dostupné také z: https://books.google.co.uk/books?id=PsaSDwAAQBAJ.

42. ČESKO. Zákon č.110/2019 Sb. Zákon o zpracování osobních údajů. In: *Sbírka zákonů České republiky* [online]. 2019 [cit. 2020-04-30]. Dostupné z: https://www.zakonyprolidi.cz/cs/2019-110.

43. MINISTERSTVO VNITRA ČESKÉ REPUBLIKY. *ORIENTACE V GDPR* [online] [cit. 2020-05-28]. Dostupné z: https://www.mvcr.cz/gdpr/clanek/zasady-zpracovani-osobnich-udaju.aspx.

44. EVROPSKÝ PARLAMENT. *Listina základních práv Evropské unie* [online]. 2012 [cit. 2020-04-30]. Dostupné z: http://data.europa.eu/eli/treaty/char_2012/oj.

45. ČESKO. *Listina základních práv Evropské unie* [online]. 1993 [cit. 2020-04-30]. Dostupné z: https://www.psp.cz/docs/laws/listina.html.

46. CHLEBUS, T.; DOSTÁL, J. *Nový zákon o zpracování osobních údajů* [online]. 2019 [cit. 2020-05-30]. Dostupné z: https://www.epravo.cz/top/clanky/novy-zakon-o-zpracovani-osobnich-udaju-109312.html.

47. ČESKO. *Zákon č. 127/2005 Sb. Zákon o elektronických komunikacích* [online]. 2005 [cit. 2020-05-30]. Dostupné z: `https://www.zakonyprolidi.cz/cs/2005-127`.

48. TEXT-COMPARE.COM. *Text Compare!* [online] [cit. 2020-05-30]. Dostupné z: `https://text-compare.com/`.

49. FREE SOFTWARE FOUNDATION, INC. *make(1) - Linux man page* [online] [cit. 2020-05-30]. Dostupné z: `https://linux.die.net/man/1/make`.

Příloha **A**

Seznam použitých zkratek

CSV Comma separated values

GDPR General data protection regulation

PIPEDA The personal information protection and electronic documents act

KMP Knuth-Morris-Pratt

HTML Hypertext Markup Language

ENIAC Electronic Numeric Integrator and Computer

UNIVAC Universal Automatic Computer

Příloha **B**

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `exe` ..... adresář se spustitelnou formou implementace
- `src`
  - `impl` ..... zdrojové kódy implementace
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `data` ..... adresář s ukázkami dat
  - `data1` ..... první ukázka dat
  - `data2` ..... druhá ukázka dat
  - `data3` ..... třetí ukázka dat
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF
- `doc` ..... adresář dokumentace programu